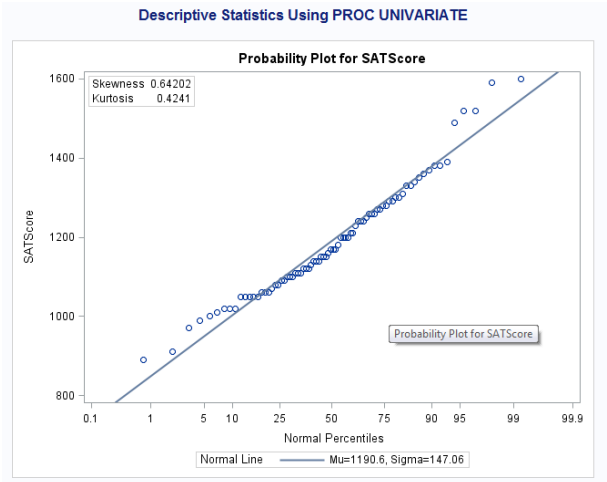
STATISTIKA I

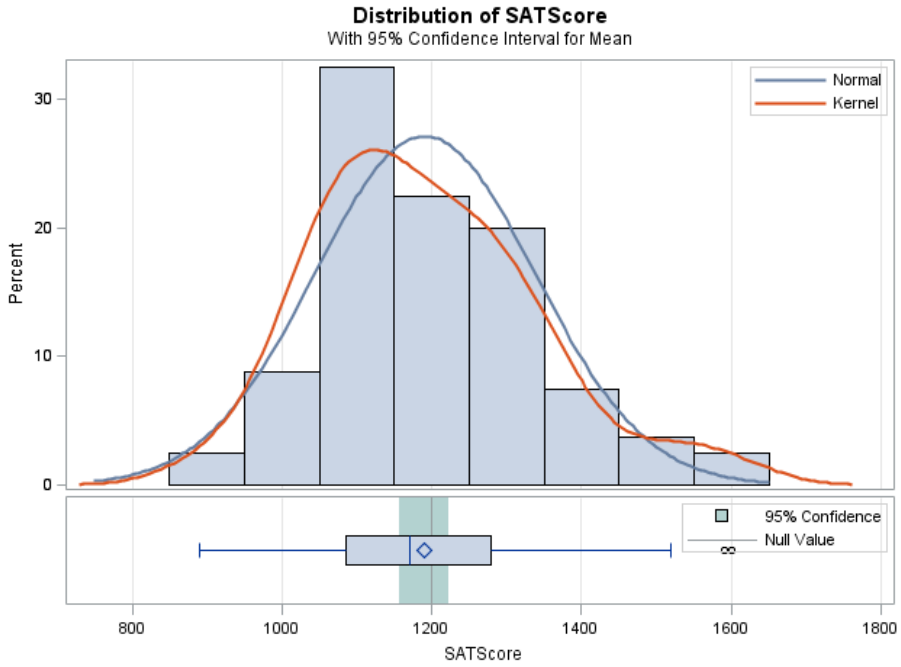
Martin Řezáč

2014

Obsah

0. Motivační příklad

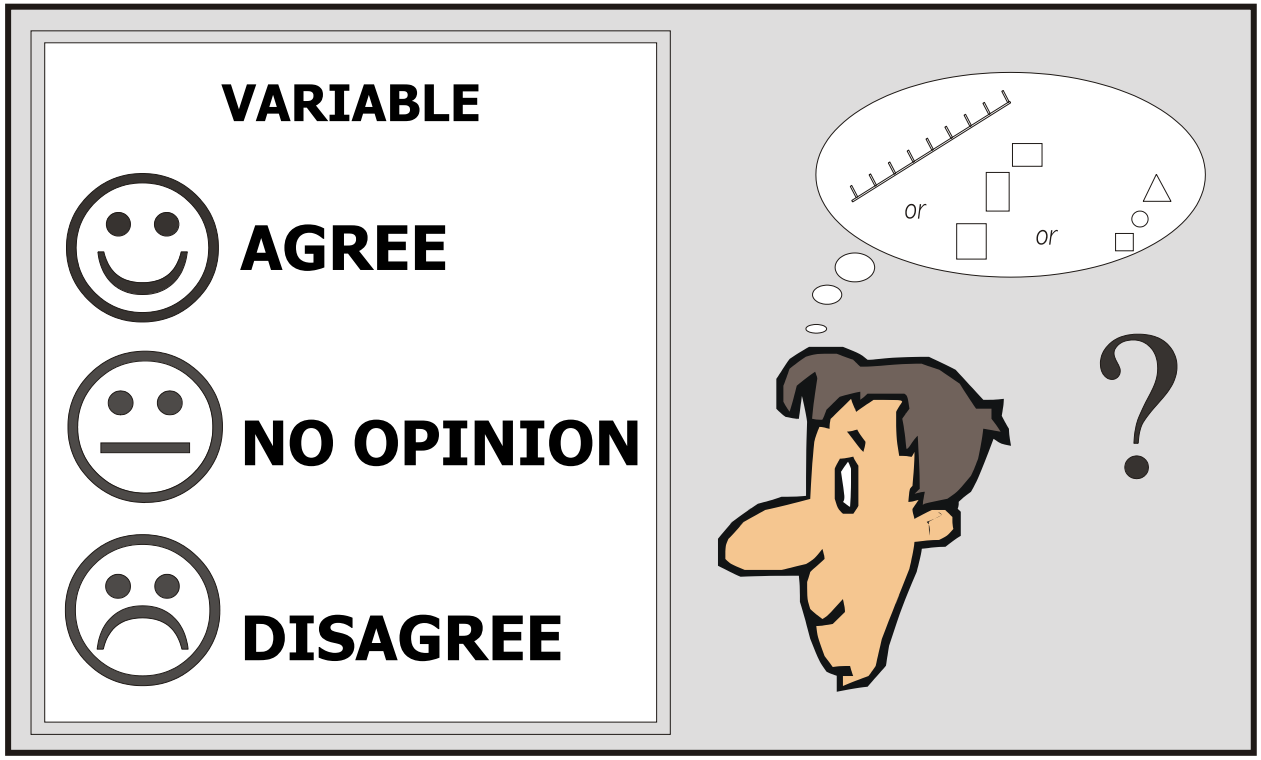




Statistika - motivační problém

- Vedení SŠ v Horní dolní chce potvrdit/vyvrátit hypotézu (předpoklad), že průměrné skóre studentů z testu matematických a verbálních schopností (SAT) je rovno 1200. Mimo to chce porozumět tomu, jakých výsledků v tomto testu studenti dosahují.

Variable Type and Level of Measurement



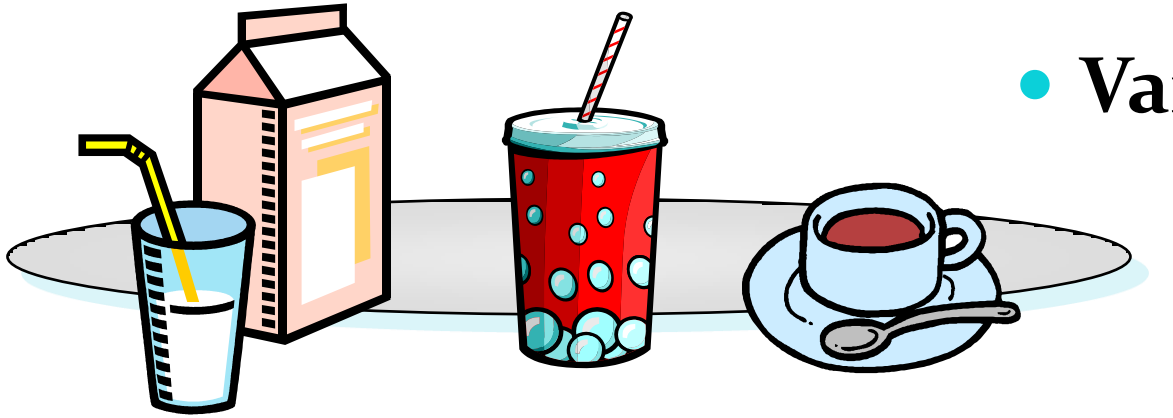
- Before analyzing, identify the **variable type** (continuous or categorical) and **level of measurement** (nominal or ordinal).

Continuous versus Categorical Variables

Variable: Temperature of Beverage (teplota nápoje)

Variable: Gender (pohlaví)

Levels of Measurement: Nominal



1 2 3

- **Variable: Type of Beverage**

1 2 3

Levels of Measurement: Ordinal

Variable: Size of Beverage

Small Medium Large

Overview of Statistical Models

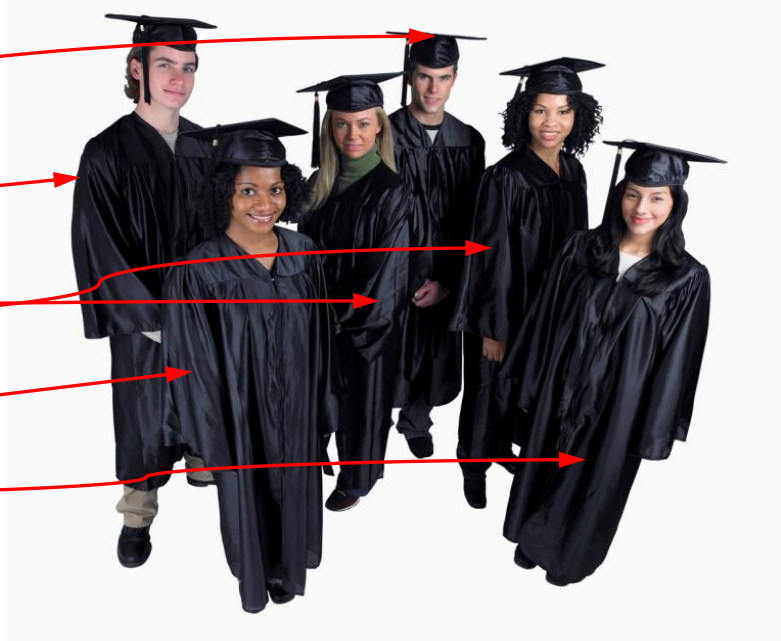
Type of Response \ Type of Predictors	Categorical	Continuous	Continuous and Categorical
Continuous	Analysis of Variance (ANOVA)	Ordinary Least Squares (OLS) Regression	Analysis of Covariance (ANCOVA)
Categorical	Contingency Table Analysis or Logistic Regression	Logistic Regression	Logistic Regression

Pro různé typy proměnných je třeba použít různé statistické metody.

Populations and Samples

Population – the entire collection of individual members of a group of interest.

Sample – a subset of a population drawn to **enable inferences** to the population.



- **Assumption– The sample that is drawn is *representative* of the population.**

Parameters and Statistics

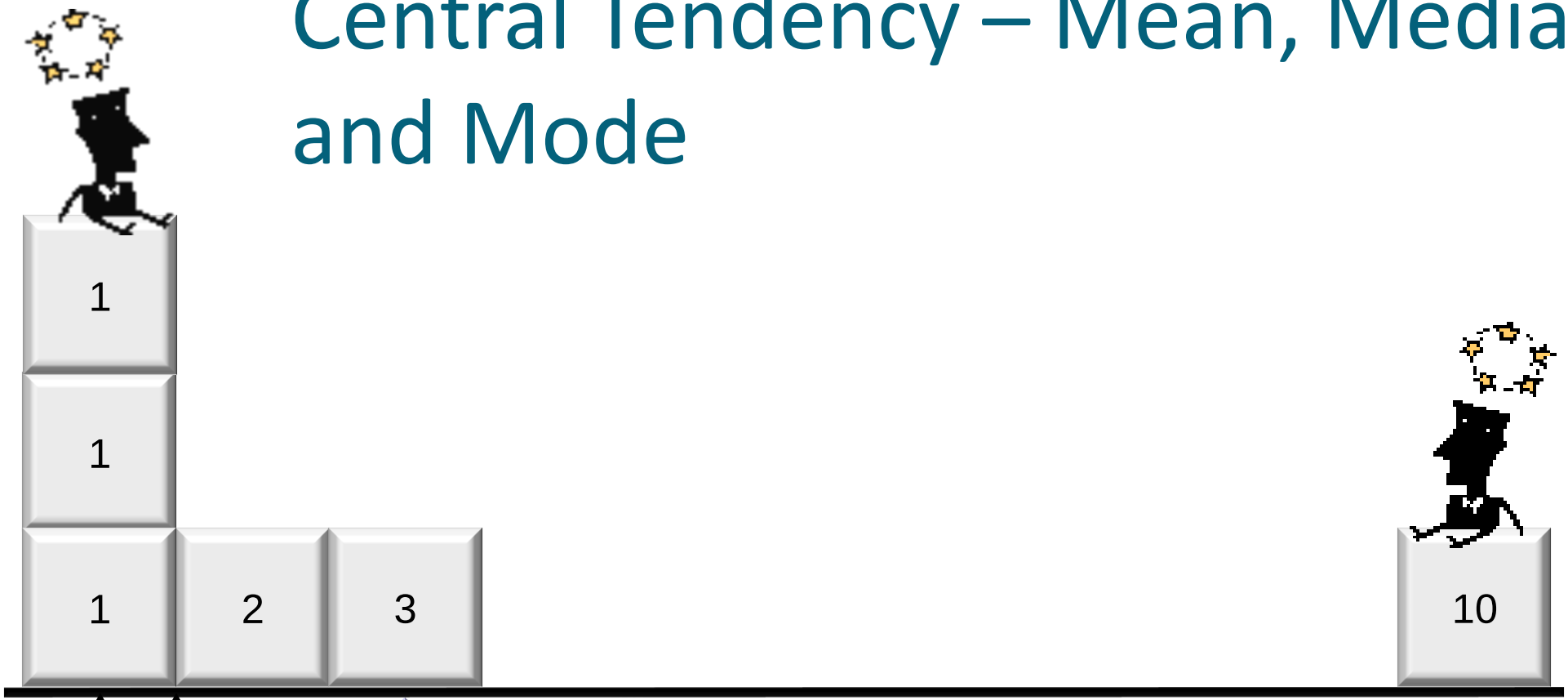
- Statistics are used to approximate population parameters.

	Population Parameters	**Sample Statistics**
Mean	μ	$\overline{X}$
Variance	σ^2	s^2
Standard Deviation	σ	s

Distributions

- When you examine the distribution of values for the variable **SATScore**, you can determine the following:
  - the range of possible data values
  - the frequency of data values
  - whether the data values accumulate in the middle of the distribution or at one end

Central Tendency – Mean, Median, and Mode

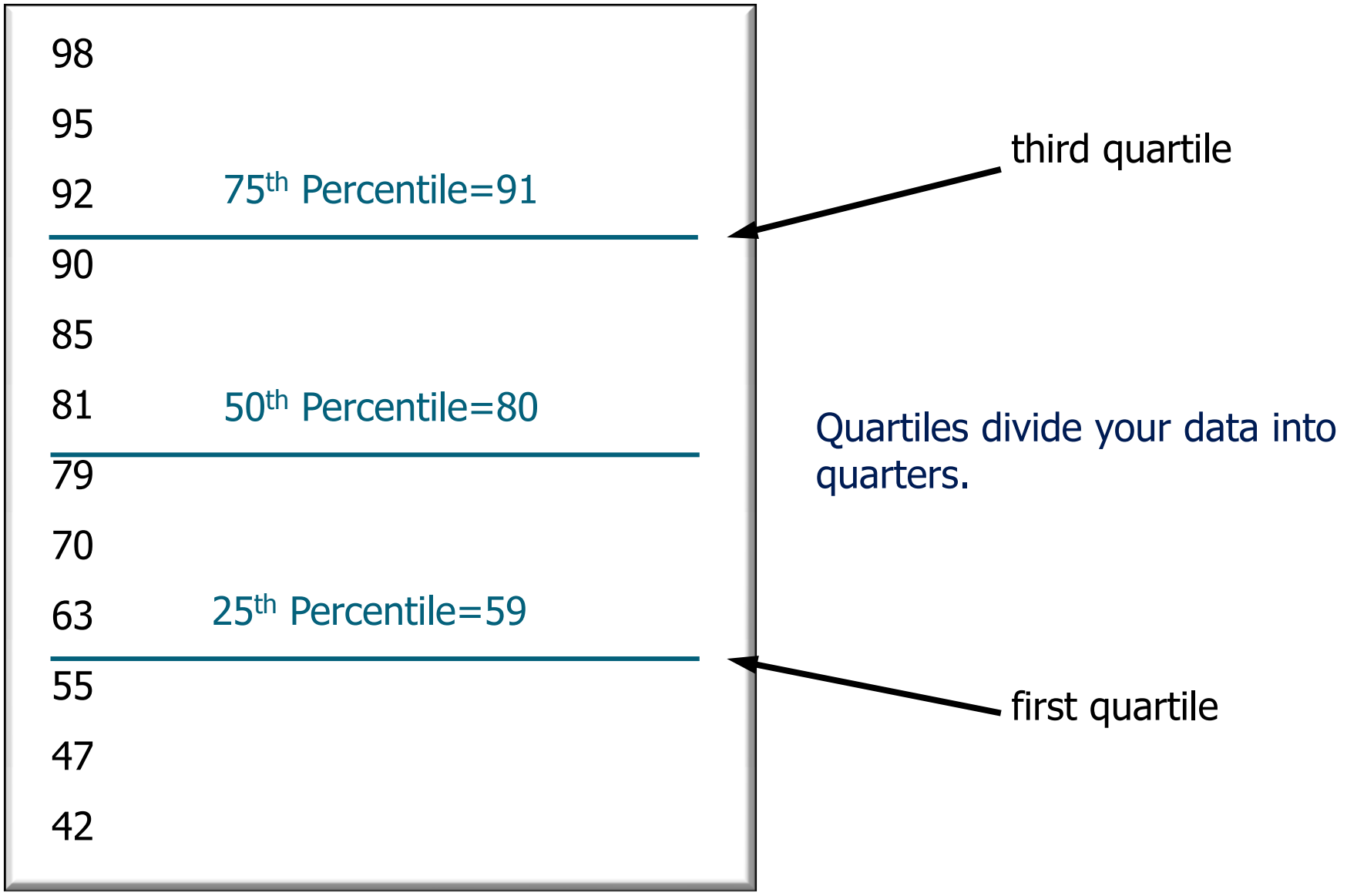


the sum of all the values in the data set divided by the number of values $\frac{\sum_{i=1}^{n} x_i}{n}$

Median=1.5 the middle value (also known as the 50th percentile)

Mode=1 the most common or frequent data value

Percentiles



The Spread of a Distribution: Dispersion

Measure	Definition
Range	the difference between the maximum and minimum data values
Interquartile Range	the difference between the 25th and 75th percentiles
Variance	a measure of dispersion of the data around the mean
Standard Deviation	a measure of dispersion expressed in the same units of measurement as your data (the square root of the variance)

SATscore descriptive statistics

Basic Statistical Measures			
Location		Variability	
Mean	1190.625	Std Deviation	147.05845
Median	1170.000	Variance	21626
Mode	1050.000	Range	710.00000
		Interquartile Range	195.00000

Quantiles (Definition 5)	
Quantile	Estimate
100% Max	1600
99%	1600
95%	1505
90%	1375
75% Q3	1280
50% Median	1170
25% Q1	1085
10%	1020
5%	995
1%	890
0% Min	890

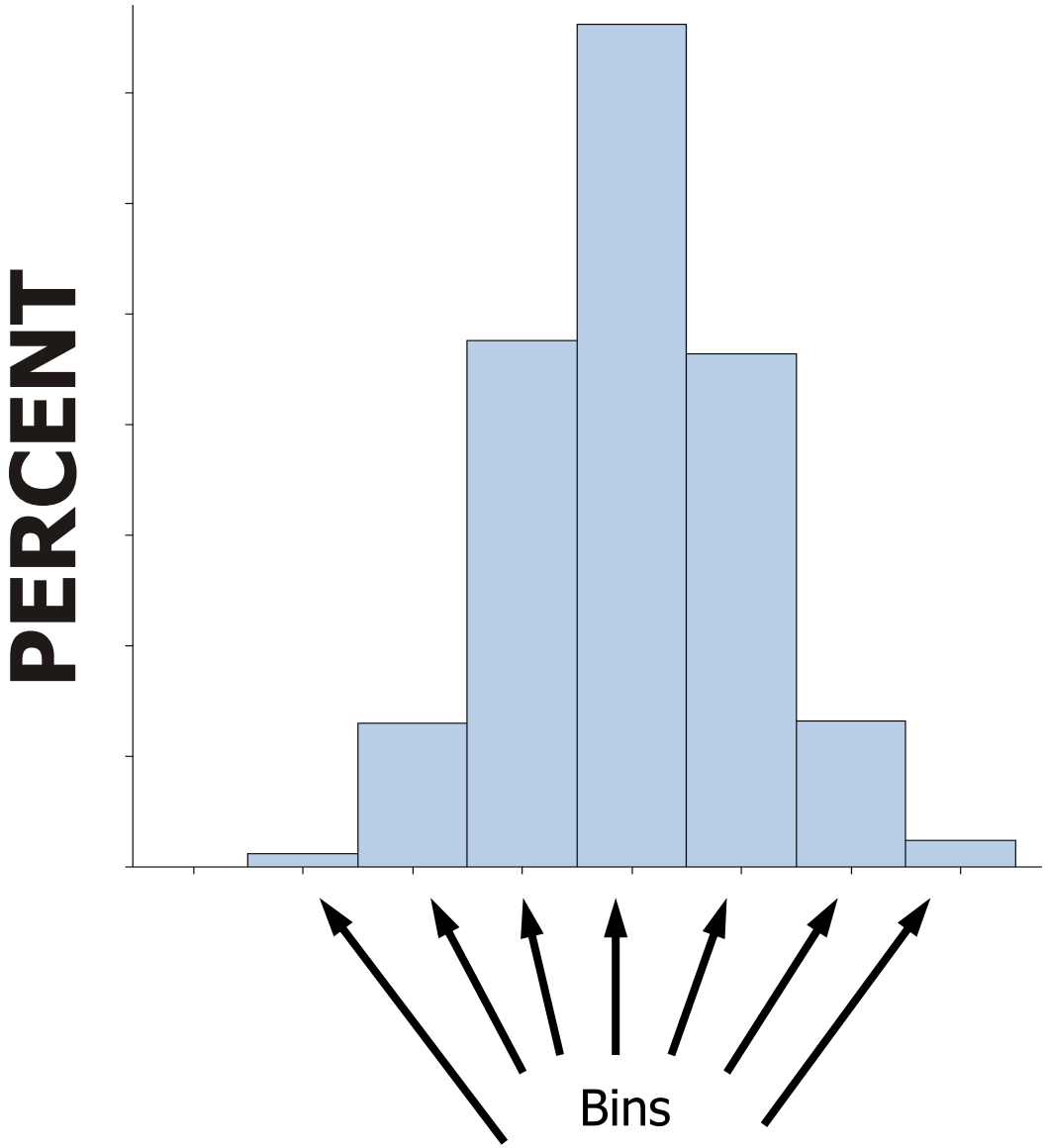
Selected Descriptive Statistics for SAT Scores

Analysis Variable : SATScore								
N	Mean	Median	Std Dev	Minimum	Maximum	Lower Quartile	Upper Quartile	Quartile Range
80	1190.63	1170.00	147.06	890.00	1600.00	1085.00	1280.00	195.00

Graphical Displays of Distributions

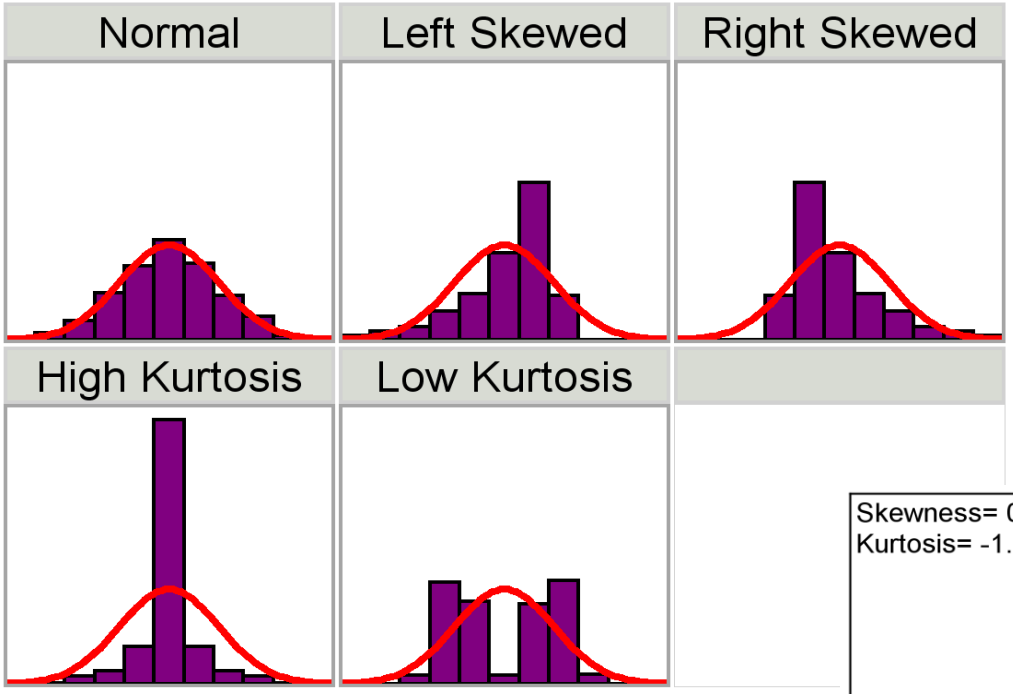
- You can produce three types of plots for examining the distribution of your data values:
  - histograms
  - normal probability plots
  - box plots

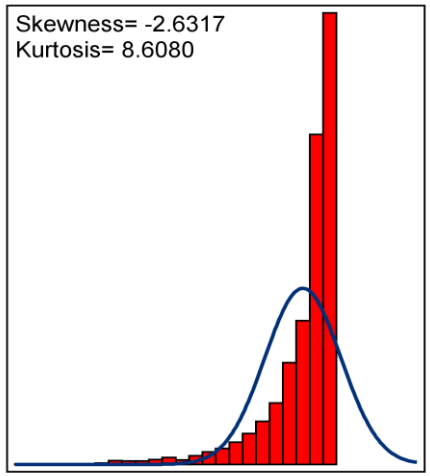
Picturing Distributions: Histogram



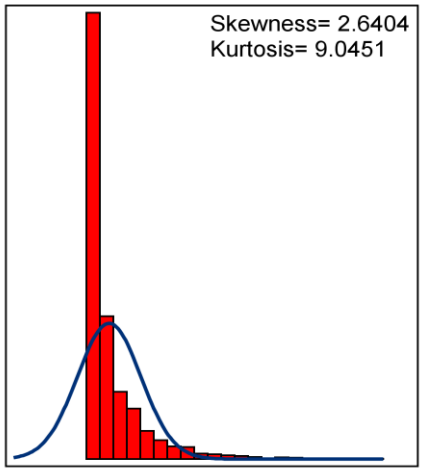
- Each bar in the histogram represents a group of values (a *bin*).
- The height of the bar represents the frequency or percent of values in the bin.
- SAS determines the width and number of bins automatically, or you can specify them.

Data Distributions

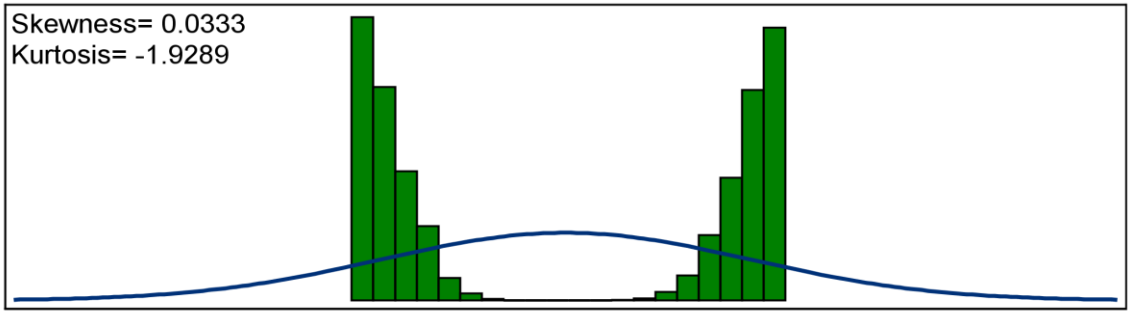




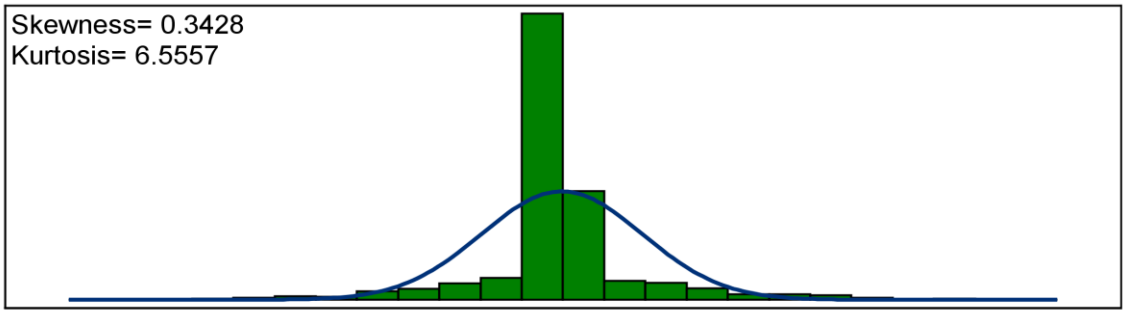
A Left Skewed Distribution



A Right Skewed Distribution



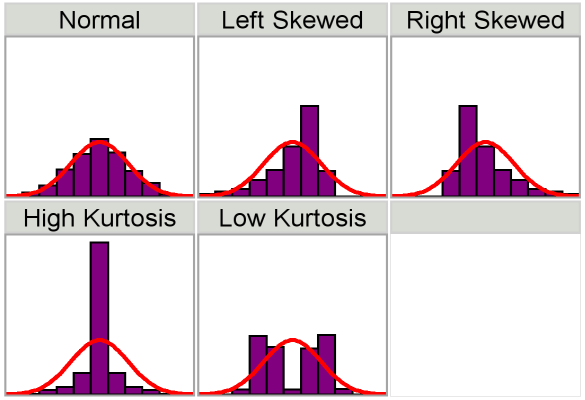
A Platykurtotic Distribution

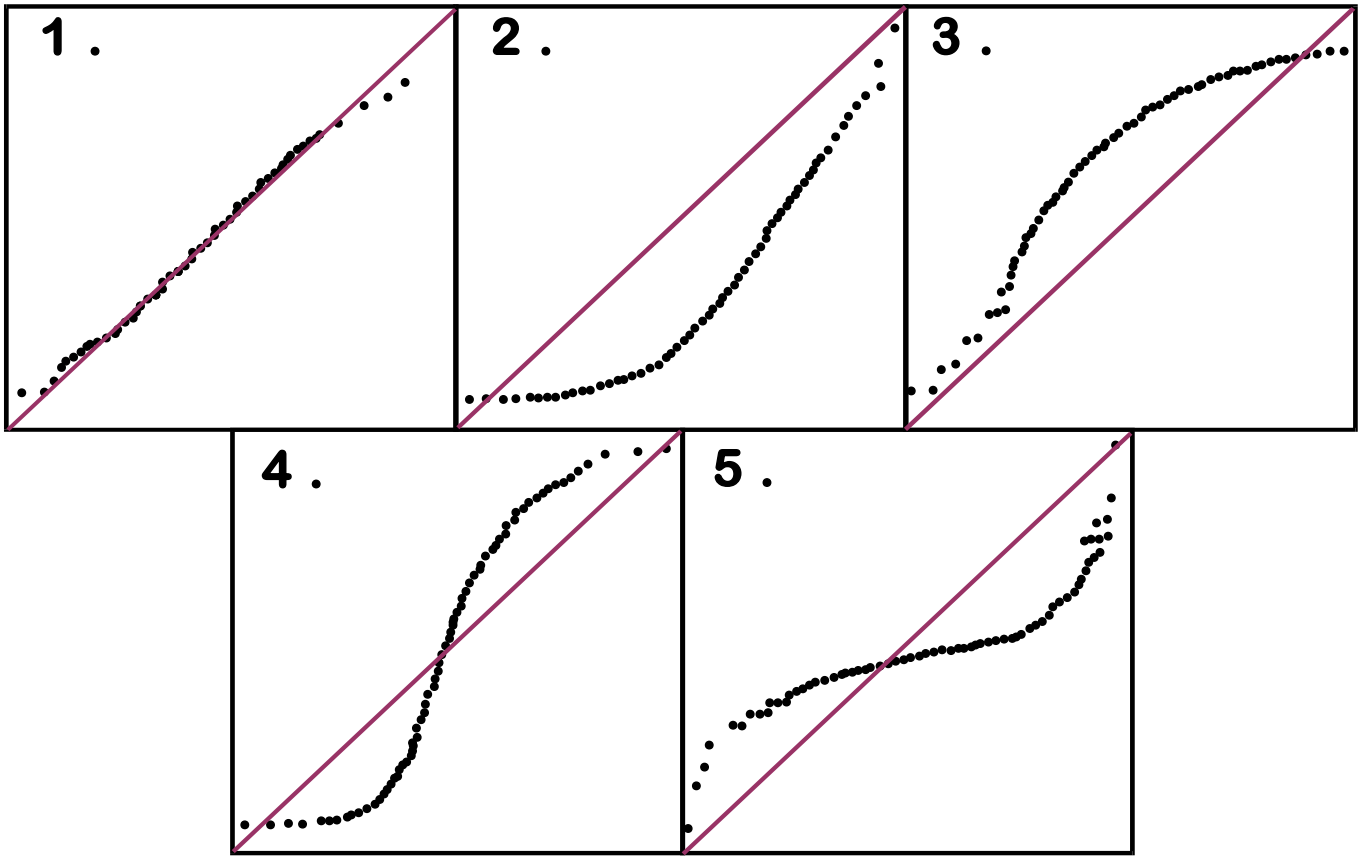


A Leptokurtotic Distribution

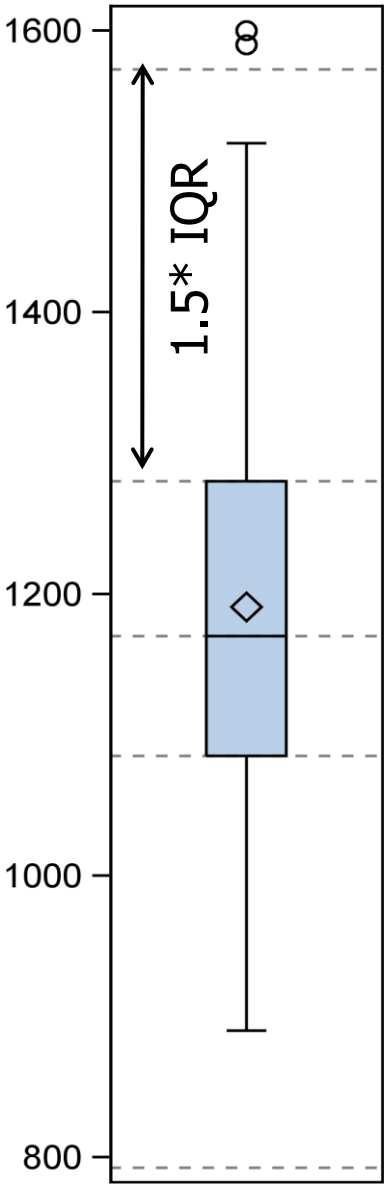
Skewness ... šikmost
Kurtosis ... špičatost

Normal Probability Plots





Box Plots



outliers > 1.5 IQR from the box

largest point <= 1.5 IQR from the box

the 75th percentile

the 50th percentile (median)

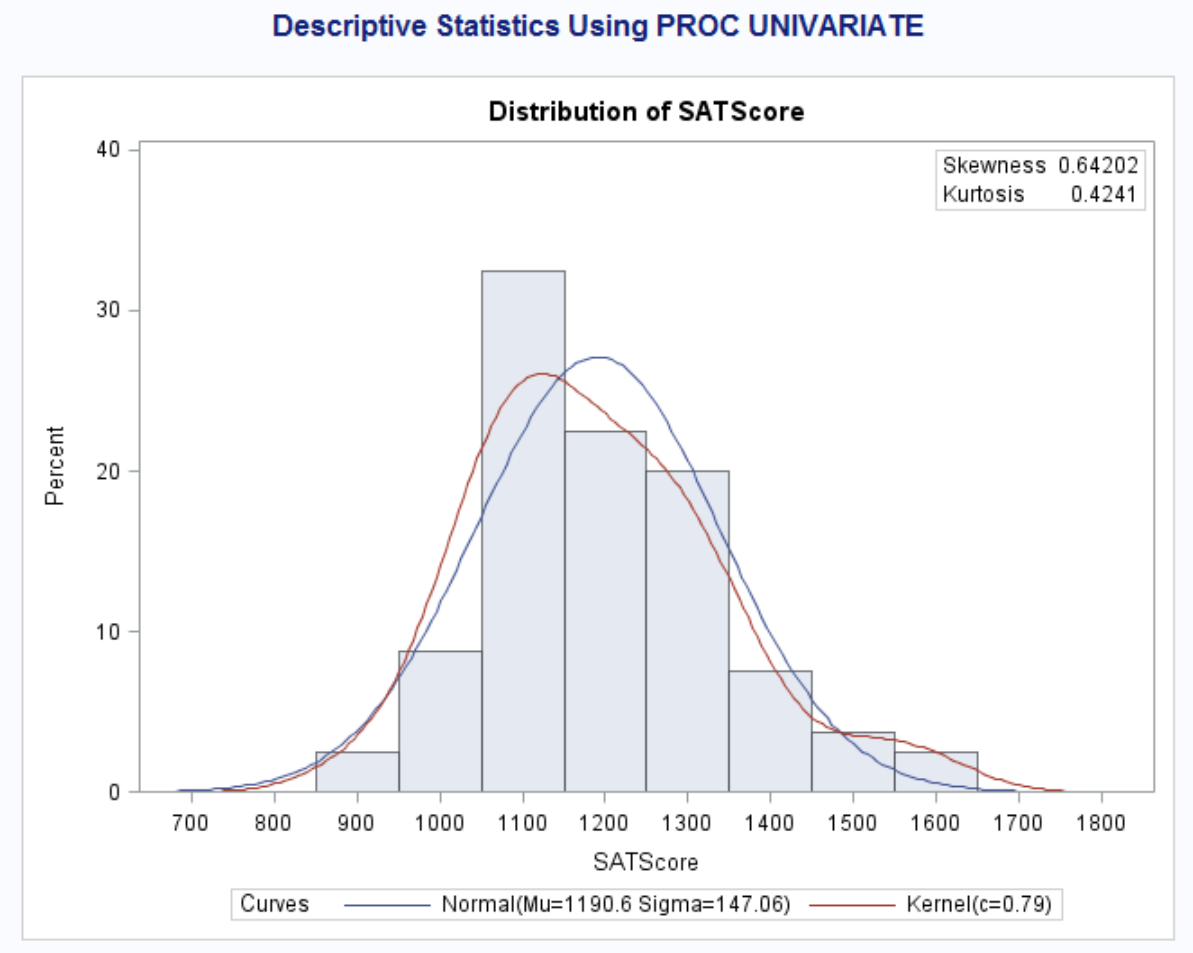
the 25th percentile

smallest point <= 1.5 IQR from the box

The mean is denoted by a ◇.

SATscore distribution

Moments			
N	80	Sum Weights	80
Mean	1190.625	Sum Observations	95250
Std Deviation	147.058447	Variance	21626.1867
Skewness	0.64202018	Kurtosis	0.42409987
Uncorrected SS	115115500	Corrected SS	1708468.75
Coeff Variation	12.3513656	Std Error Mean	16.4416342



SATscore distribution

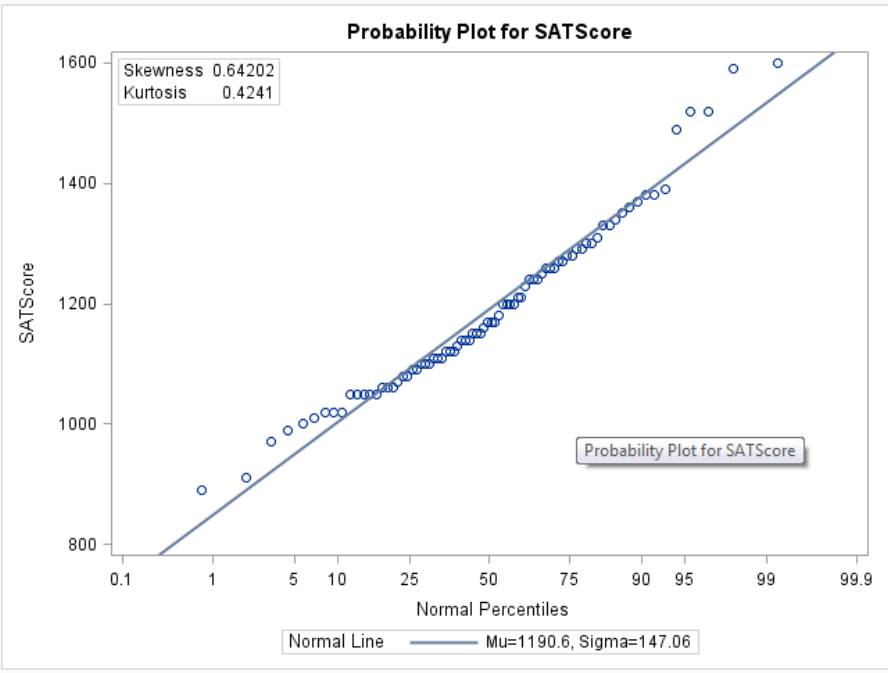
Descriptive Statistics Using PROC UNIVARIATE

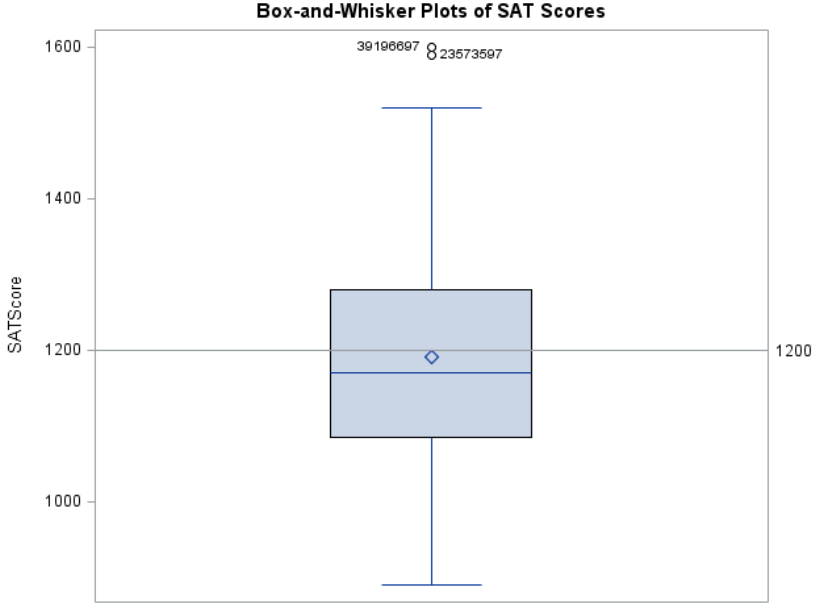
Fitted Normal Distribution for SATScore

Parameters for Normal Distribution		
Parameter	Symbol	Estimate
Mean	Mu	1190.625
Std Dev	Sigma	147.0584

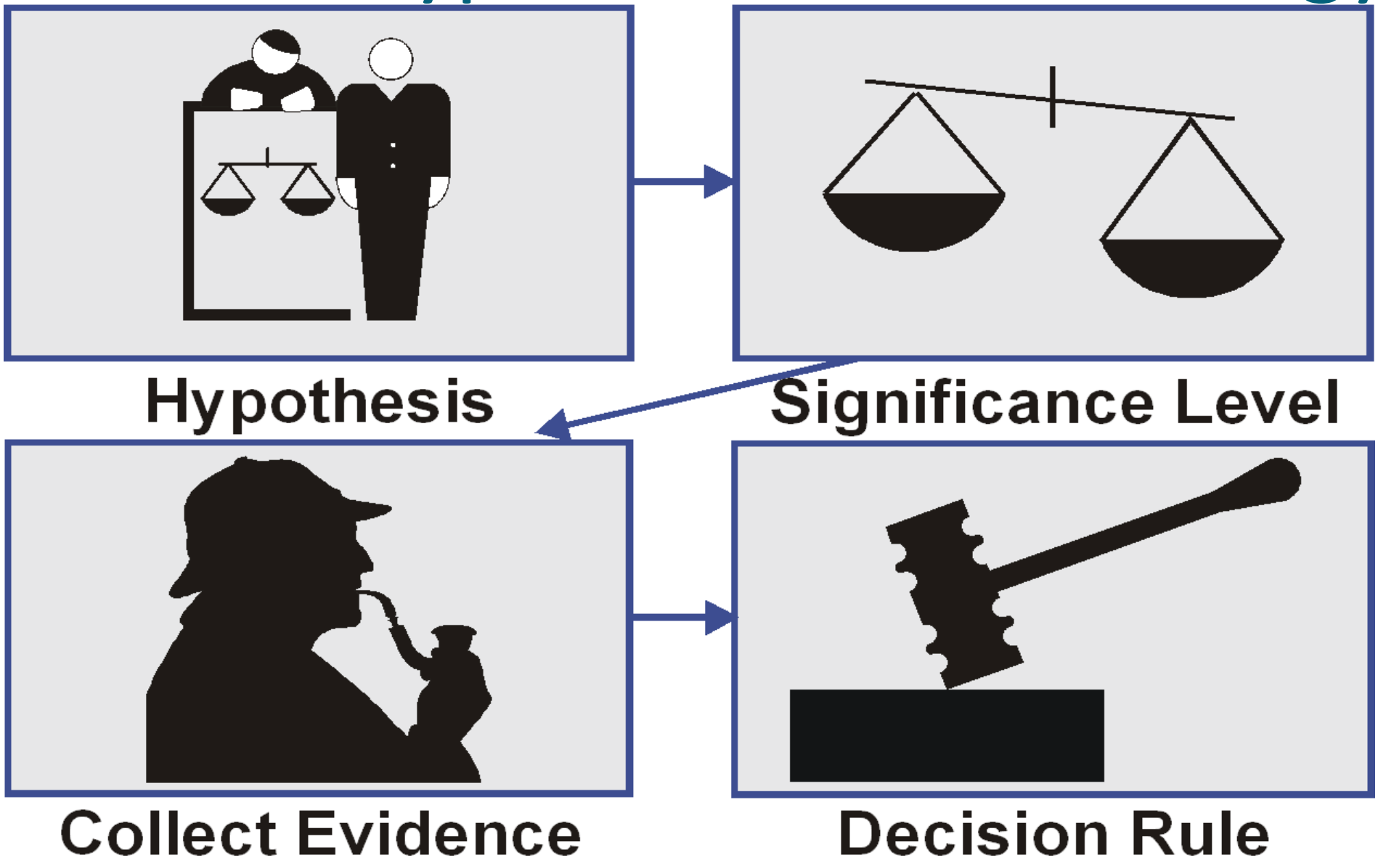
Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.08382224	Pr > D	>0.150
Cramer-von Mises	W-Sq	0.09964577	Pr > W-Sq	0.114
Anderson-Darling	A-Sq	0.70124822	Pr > A-Sq	0.068

Descriptive Statistics Using PROC UNIVARIATE





Testování hypotéz - Judicial Analogy



Coin Example

If you have a fair coin and flip it 100 times, is it possible for it to land on heads 100 times?

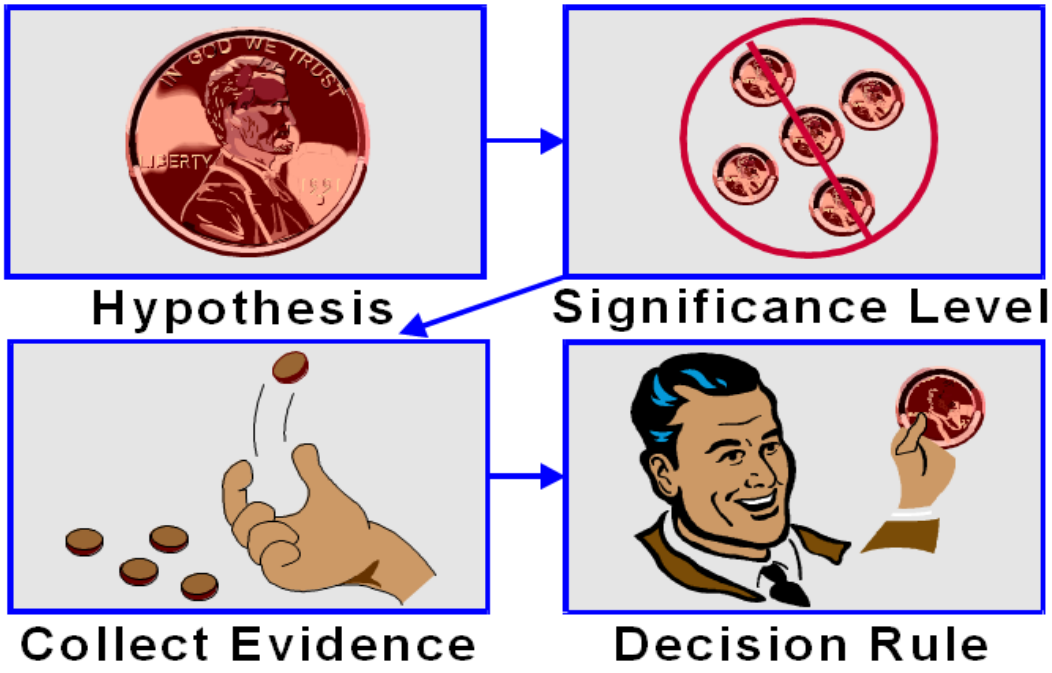
- ❍ Yes
- ❍ No

Poll – Correct Answer

If you have a fair coin and flip it 100 times, is it possible for it to land on heads 100 times?

- Yes
- No

Coin Analogy



Je třeba připustit, že s nějakou, byť malou, pravděpodobností se může stát něco nečekaného – co povede k zamítnutí hypotézy (férová mince) přestože hypotéza platí.

Types of Errors

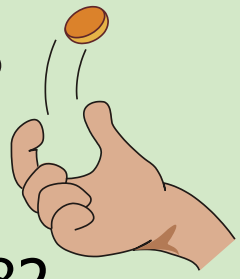
- You used a decision rule to make a decision, but was the decision correct?

ACTUAL / DECISION	H_0 Is True	H_0 Is False
Fail to Reject Null	Correct	Type II Error
Reject Null	Type I Error	Correct

Coin Experiment – Effect Size Influence

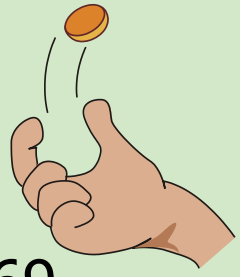
- Flip a coin 100 times and decide whether it is fair.

55 Heads
45 Tails



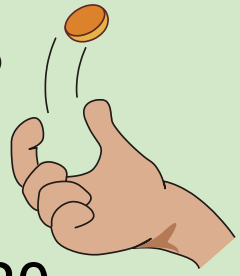
p-value=.3682

40 Heads
60 Tails



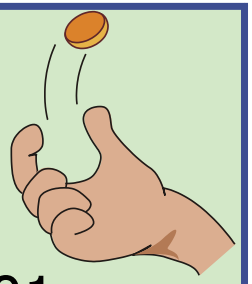
p-value=.0569

37 Heads
63 Tails



p-value=.0120

15 Heads
85 Tails



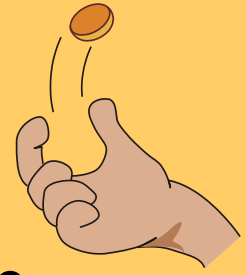
p-value<.0001

Čím vyšší p-hodnota, tím více máme důvod se domnívat, že je naše hypotéza správná.

Coin Experiment – Sample Size Influence

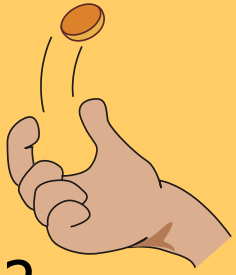
- Flip a coin and get 40% heads and decide whether it is fair.

4 Heads
6 Tails



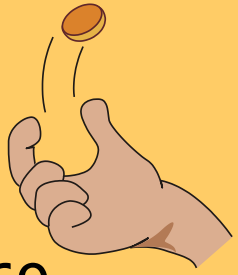
p-value=.7539

16 Heads
24 Tails



p-value=.2682

40 Heads
60 Tails

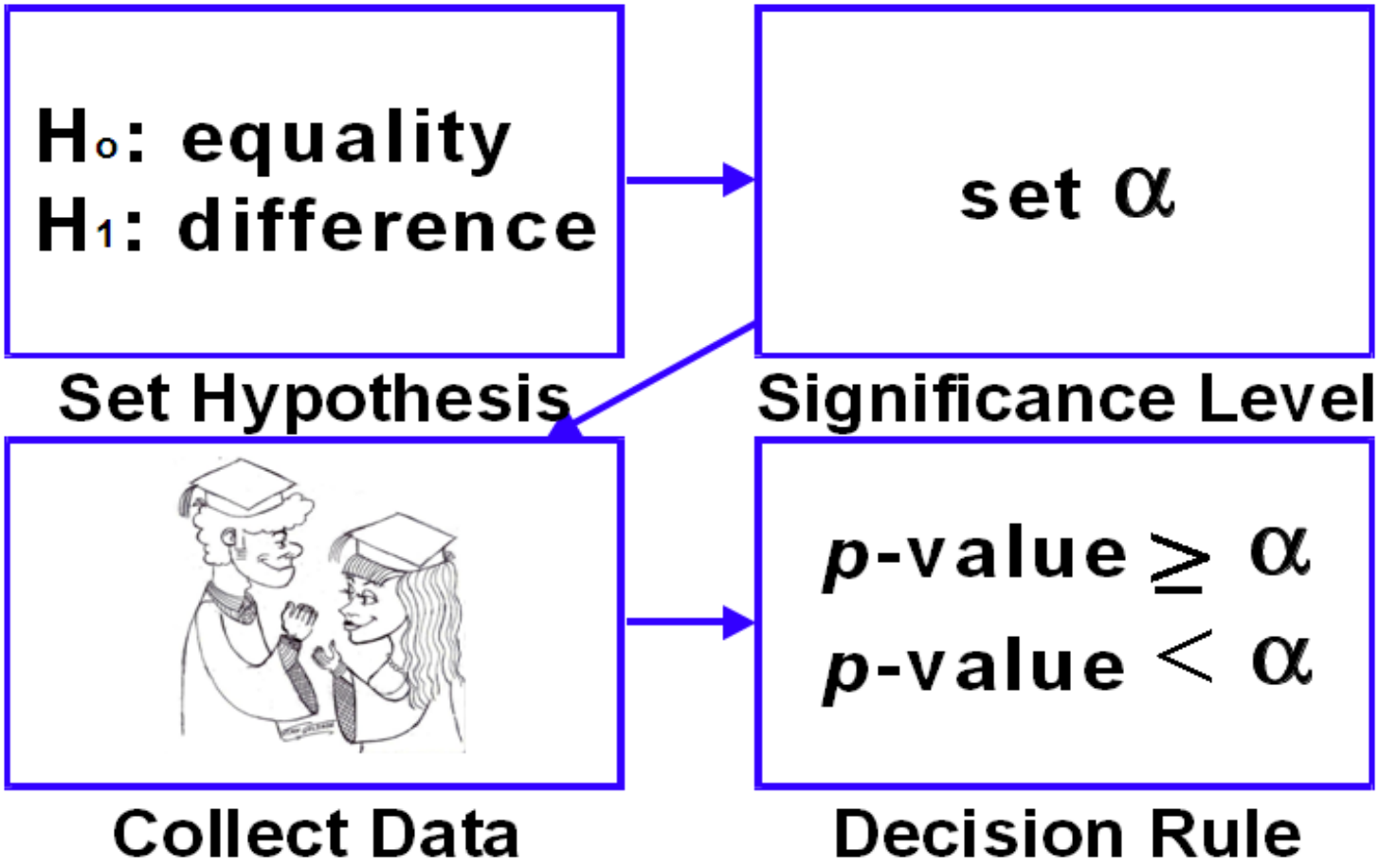


p-value=.0569

160 Heads
240 Tails

p-value<.0001

Statistical Hypothesis Test



- In general, you do one of the following:
  - reject the null hypothesis if *p*-value<α
  - fail to reject the null hypothesis if *p*-value≥α.

Proc Univariate for SATscore

Testing Whether the Mean of SAT Scores = 1200

Variable: SATScore

Moments			
N	80	Sum Weights	80
Mean	1190.625	Sum Observations	95250
Std Deviation	147.058447	Variance	21626.1867
Skewness	0.64202018	Kurtosis	0.42409987
Uncorrected SS	115115500	Corrected SS	1708468.75
Coeff Variation	12.3513656	Std Error Mean	16.4416342

Basic Statistical Measures			
Location		Variability	
Mean	1190.625	Std Deviation	147.05845
Median	1170.000	Variance	21626
Mode	1050.000	Range	710.00000
		Interquartile Range	195.00000

Tests for Location: Mu0=1200				
Test	Statistic		p Value	
Student's t	t	-0.5702	Pr > |t|	0.5702
Sign	M	-5	Pr >= |M|	0.3019
Signed Rank	S	-207	Pr >= |S|	0.2866

p-hodnota je vetší než 0.05, tudíž bych hypotézu nezamítal.

TTEST for SATscore

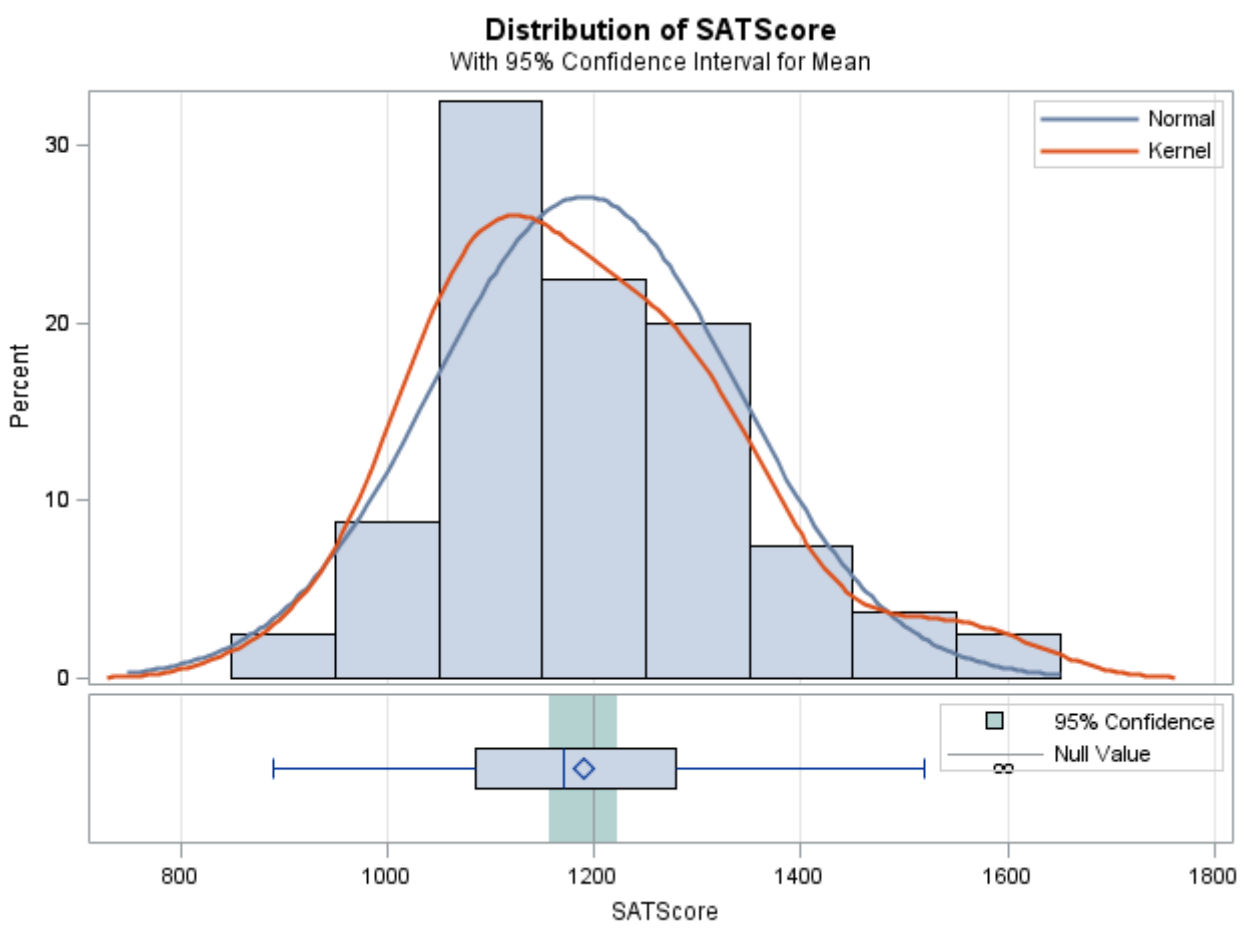
Testing Whether the Mean of SAT Scores = 1200 Using PROC TTEST

Variable: SATScore

N	Mean	Std Dev	Std Err	Minimum	Maximum
80	1190.6	147.1	16.4416	890.0	1600.0

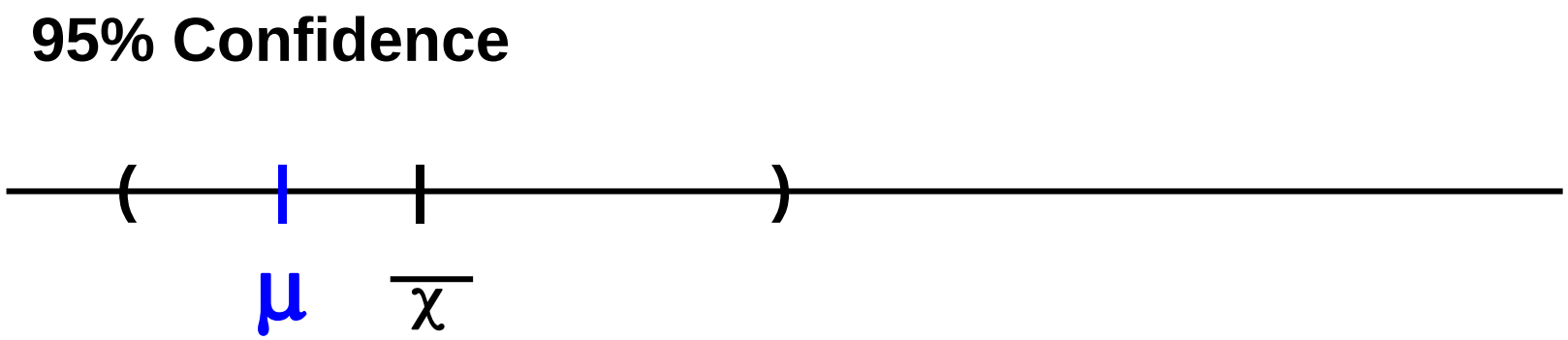
Mean	95% CL Mean		Std Dev	95% CL Std Dev	
1190.6	1157.9	1223.4	147.1	127.3	174.2

DF	t Value	Pr > |t|
79	-0.57	0.5702



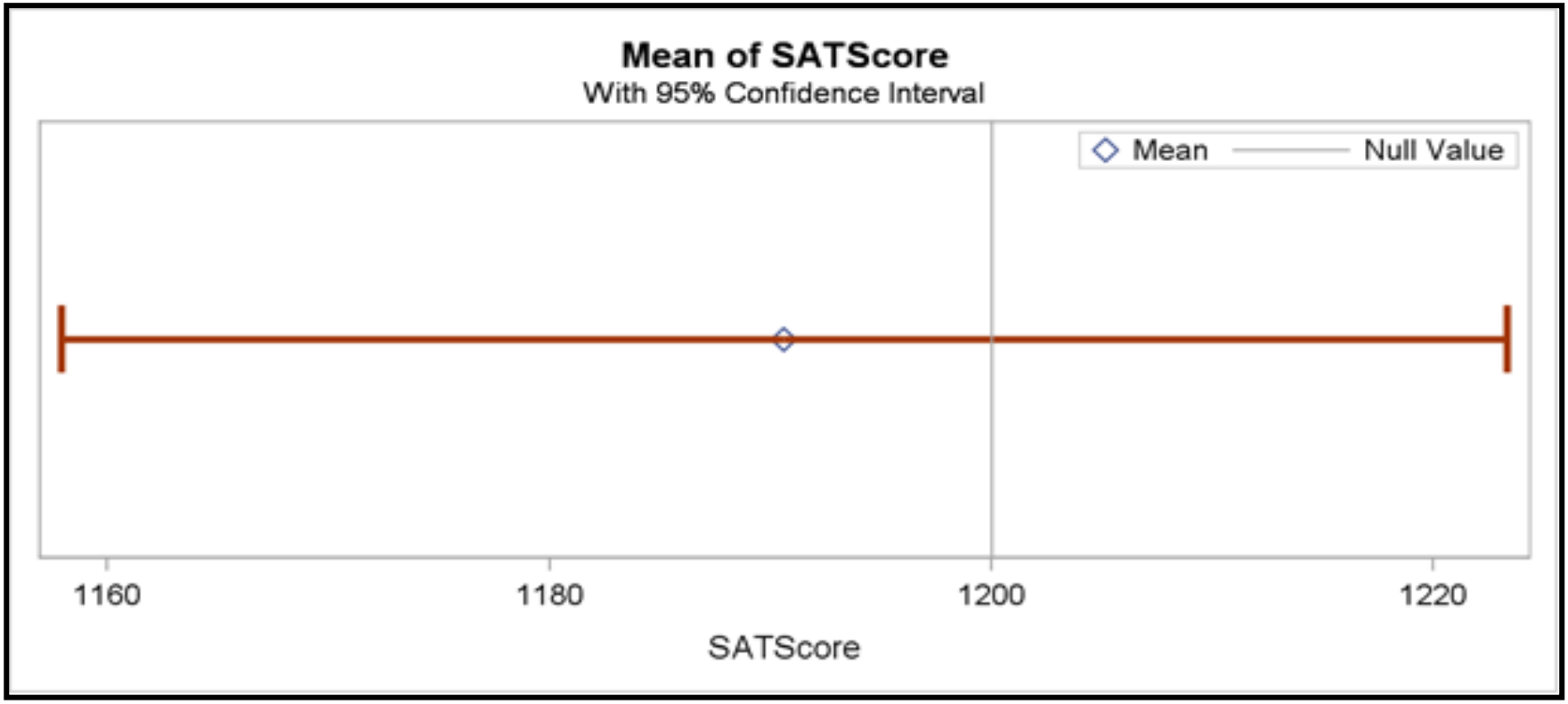
p-hodnota je vetší než 0.05, tudíž bych hypotézu nezamítal.

Confidence Intervals



- A 95% confidence interval represents a range of values within which you are 95% certain that the true population mean exists.
  - One interpretation is that if 100 different samples were drawn from the same population and 100 intervals were calculated, approximately 95 of them would contain the population mean.

Confidence Interval Plots



Mean	95% CL Mean	
1190.6	1157.9	1223.4

Multiple Choice Poll

- A 95% confidence interval for SAT scores is (1157.90, 1223.35). From this, what can you conclude, at alpha=0.05?

  a. The true average SAT score is significantly different from 1200.

  b. The true average SAT score is not significantly different from 1200.

  c. The true average SAT score is less than 1200.

  d. None of the above – You cannot determine statistical significance from confidence intervals.

Multiple Choice Poll – Correct Answer

- A 95% confidence interval for SAT scores is (1157.90, 1223.35). From this, what can you conclude, at alpha=0.05?

  a. The true average SAT score is significantly different from 1200.

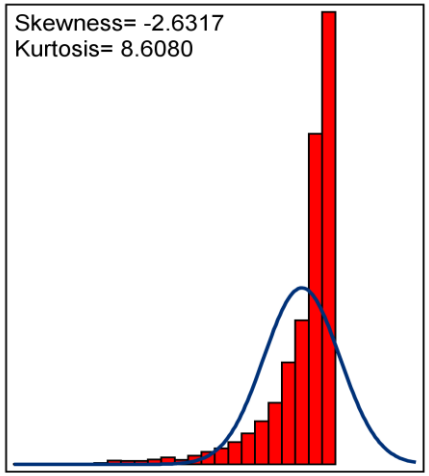
  b. The true average SAT score is not significantly different from 1200.

  c. The true average SAT score is less than 1200.

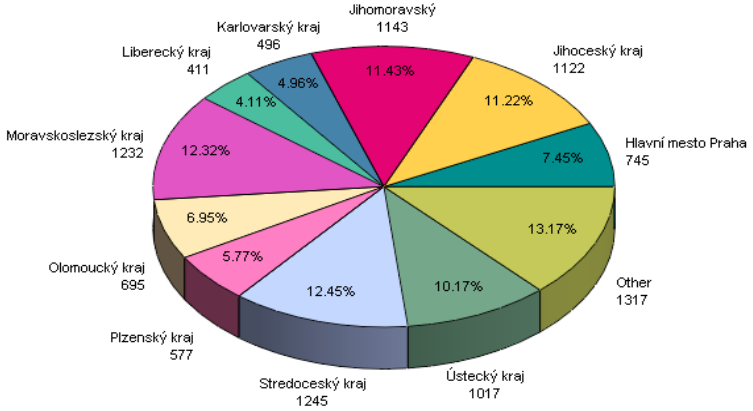
  d. None of the above – You cannot determine statistical significance from confidence intervals.

1. Tabulkové a grafické zpracování datových souborů.

$x_{[j]}$	n_j	p_j	N_j	F_j
$x_{[1]}$	n_1	p_1	N_1	F_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x_{[r]}$	n_r	p_r	N_r	F_r



A Left Skewed Distribution



Popisná statistika

Popisná statistika je disciplína, která popisuje a sumarizuje informace obsažené ve velkém množství dat pomocí tabulek, grafů, funkcionálních a číselných charakteristik. Činí tak pomocí základních matematických operací. Cílem popisné statistiky je zpřehlednit informace „ukryté´" v datových souborech.

Popisná statistika je velmi důležitá minimálně ze dvou důvodů:

- v praxi se často používá (všichni znají takové pojmy jako je průměr, směrodatná odchylka, tabulka rozložení četností, výsečový graf apod.)
- motivuje pojmy, se kterými pak pracuje počet pravděpodobnosti (např. relativní četnost motivuje pravděpodobnost, hustota četnosti motivuje hustotu pravděpodobnosti, průměr motivuje střední hodnotu apod.)

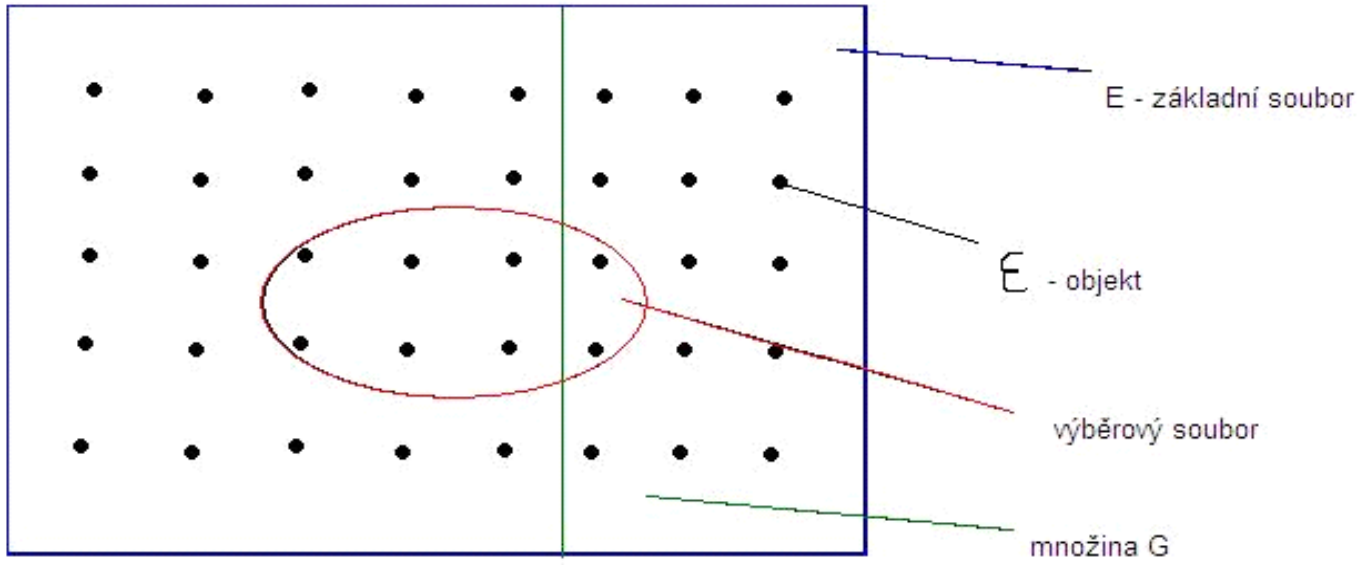
Dobré pochopení pojmů popisné statistiky tedy velmi usnadní studium počtu pravděpodobnosti.

Základní, výběrový a datový soubor

Základním souborem rozumíme libovolnou neprázdnou množinu E. Prvky množiny E značíme ε a nazýváme je objekty. Libovolnou neprázdnou podmnožinu $\{\varepsilon_1, \ldots, \varepsilon_n\}$ základního souboru E nazýváme výběrový soubor rozsahu n. Je-li množina $G \subseteq E$, pak symbolem N(G) rozumíme absolutní četnost množiny G ve výběrovém souboru, tj. počet těch objektů množiny G, které patří do výběrového souboru. Relativní četnost množiny G ve výběrovém souboru zavedeme vztahem

$$p(G) = \frac{N(G)}{n}.$$

Ilustrace



Příklad

Příklad: Základním souborem E je množina všech ekonomicky zaměřených studentů 1. ročníku českých vysokých škol. Množina G_1 je tvořena těmi studenty, kteří uspěli v prvním zkušebním termínu z matematiky a množina G_2 obsahuje ty studenty, kteří uspěli v prvním zkušebním termínu z angličtiny. Ze základního souboru bylo náhodně vybráno 20 studentů, kteří tvoří výběrový soubor $\{\varepsilon_1, ..., \varepsilon_{20}\}$. Z těchto 20 studentů 12 uspělo v matematice, 15 v angličtině a 11 v obou předmětech. Zapište absolutní a relativní četnosti úspěšných matematiků, angličtinářů a oboustranně úspěšných studentů.

Řešení:

$$N(G_1) = 12, N(G_2) = 15, N(G_1 \cap G_2) = 11, n = 20, p(G_1) = \frac{12}{20} = 0{,}6, p(G_2) = \frac{15}{20} = 0{,}75,$$

$$p(G_1 \cap G_2) = \frac{11}{20} = 0{,}55$$

Vidíme, že úspěšných matematiků je 60%, angličtinářů 75% a oboustranně úspěšných studentů jen 55%.

Relativní četnost

Vlastnosti relativní četnosti: Relativní četnost má následujících 12 vlastností, které jsou obdobné vlastnostem procent.

- $p(\varnothing) = 0$
- $p(G) \geq 0$ (nezápornost)
- $p(G) \leq 1$
- $p(G_1 \cup G_2) + p(G_1 \cap G_2) = p(G_1) + p(G_2)$
- $1 + p(G_1 \cap G_2) \geq p(G_1) + p(G_2)$
- $p(G_1 \cup G_2) + 0 \leq p(G_1) + p(G_2)$ (subaditivita)
- $G_1 \cap G_2 = \varnothing \Rightarrow p(G_1 \cup G_2) = p(G_1) + p(G_2)$ (aditivita)
- $p(G_2 \setminus G_1) = p(G_2) - p(G_1 \cap G_2)$
- $G_1 \subseteq G_2 \Rightarrow p(G_2 \setminus G_1) = p(G_2) - p(G_1)$ (subtraktivita)
- $G_1 \subseteq G_2 \Rightarrow p(G_1) \leq p(G_2)$ (monotonie)
- $p(E) = 1$ (normovanost)
- $p(G) + p(\overline{G}) = 1$ (komplementarita)

Podmíněná relativní četnost

Pokud se v daném základním souboru zajímáme o dvě podmnožiny, můžeme zavést pojem podmíněné relativní četnosti jedné podmnožiny v daném výběrovém souboru za předpokladu, že objekt pochází z druhé množiny.

Nechť E je základní soubor, G_1 , G_2 jeho podmnožiny, $\{\varepsilon_1, \ldots, \varepsilon_n\}$ výběrový soubor. Definujeme

podmíněnou relativní četnost množiny G_1 ve výběrovém souboru za předpokladu G_2 :

$$p(G_1/G_2) = \frac{N(G_1 \cap G_2)}{N(G_2)} = \frac{p(G_1 \cap G_2)}{p(G_2)}$$ a

Podmíněnou relativní četnost G_2 ve výběrovém souboru za předpokladu G_1 :

$$p(G_2/G_1) = \frac{N(G_1 \cap G_2)}{N(G_1)} = \frac{p(G_1 \cap G_2)}{p(G_1)}.$$

Příklad

Příklad: Pro údaje z příkladu o studentech vypočtěte podmíněnou relativní četnost úspěšných matematiků mezi úspěšnými angličtináři a podmíněnou relativní četnost úspěšných angličtinářů mezi úspěšnými matematiky.
(Připomínáme, že z 20 studentů 12 uspělo v matematice, 15 v angličtině a 11 v obou předmětech.)

Řešení:

$N(G_1) = 12, N(G_2) = 15, N(G_1 \cap G_2) = 11, n = 20,$

$p(G_1/G_2) = \frac{N(G_1 \cap G_2)}{N(G_2)} = \frac{11}{15} = 0{,}73$ (tzn., že 73% těch studentů, kteří byli úspěšní v angličtině, uspělo i v matematice)

$p(G_2/G_1) = \frac{N(G_1 \cap G_2)}{N(G_1)} = \frac{11}{12} = 0{,}92$ (tzn., že 92% těch studentů, kteří byli úspěšní v matematice, uspělo i v angličtině)

Četnostní nezávislost

Pojem četnostní nezávislosti dvou množin: O četnostní nezávislosti dvou množin v daném výběrovém souboru hovoříme tehdy, když informace o původu objektu z jedné množiny nijak nemění šance, s nimiž soudíme na jeho původ i z druhé množiny.

V příkladě se studenty by množiny úspěšných matematiků a úspěšných angličtinářů byly četnostně nezávislé, pokud podíl úspěšných matematiků mezi úspěšnými angličtináři by byl stejný jako podíl úspěšných matematiků mezi všemi zkoušenými studenty a stejně tak podíl úspěšných angličtinářů mezi úspěšnými matematiky by byl stejný jako podíl úspěšných angličtinářů mezi všemi zkoušenými studenty, tj.

$$\frac{N(G_1 \cap G_2)}{N(G_2)} = \frac{N(G_1)}{n} \wedge \frac{N(G_1 \cap G_2)}{N(G_1)} = \frac{N(G_2)}{n}.$$

Po snadné úpravě dostaneme multiplikativní vztah

$$\frac{N(G_1 \cap G_2)}{n} = \frac{N(G_1)}{n} \cdot \frac{N(G_2)}{n}, \text{ tj. } p(G_1 \cap G_2) = p(G_1)p(G_2)$$

Řekneme tedy, že množiny G_1, G_2 jsou četnostně nezávislé v daném výběrovém souboru, jestliže $p(G_1 \cap G_2) = p(G_1)p(G_2)$.

(V praxi jen zřídka dojde k tomu, že uvedený vztah platí přesně. Většinou je jen naznačena určitá tendence četnostní nezávislosti.)

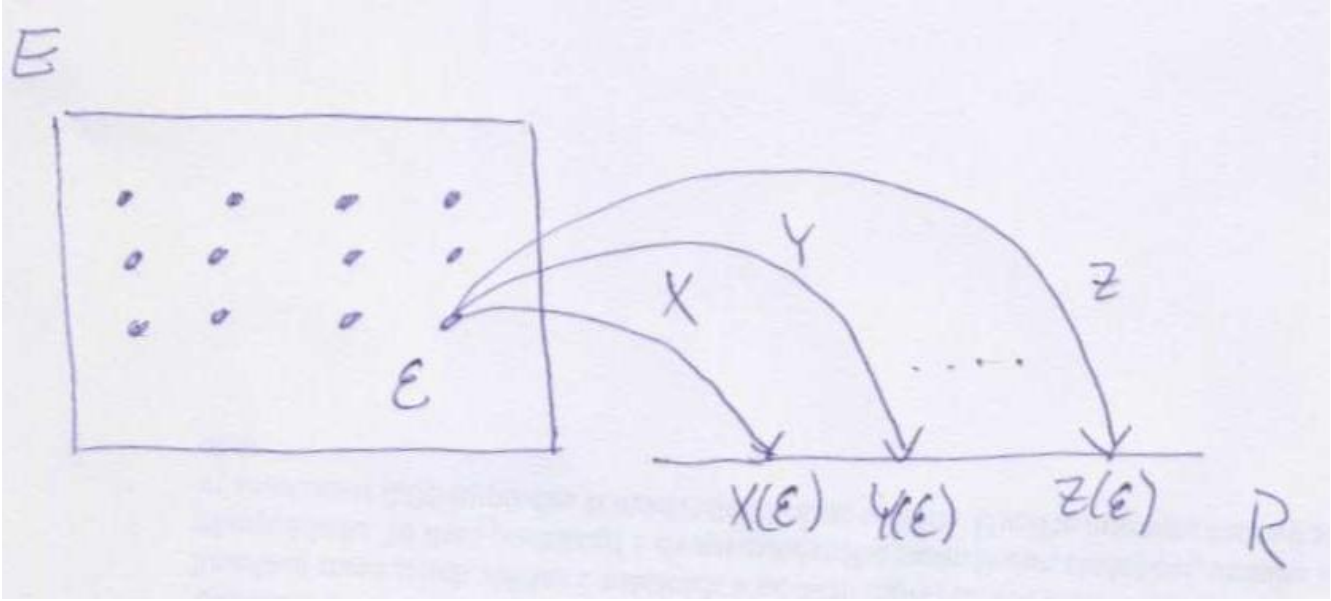
Příklad

Příklad: Pro údaje z příkladu o studentech zjistěte, zda úspěchy v matematice a angličtině jsou v daném výběrovém souboru četnostně nezávislé.
(Připomínáme, že oboustranně úspěšných studentů bylo 55%, úspěšných matematiků 60% a úspěšných angličtinářů 75%.)

Řešení:
$p(G_1 \cap G_2) = 0{,}55$, $p(G_1)p(G_2) = 0{,}6 \times 0{,}75 = 0{,}45$, tedy skutečná relativní četnost oboustranně úspěšných studentů je větší než by odpovídalo četnostní nezávislosti množin G_1, G_2 v daném výběrovém souboru. Znamená to, že úspěch v matematice se zpravidla sdružuje s úspěchem v angličtině a naopak.

Skalární a vektorový znak

Pojem skalárního a vektorového znaku: Vlastnosti objektů vyjadřujeme číselně pomocí znaků.
Nechť E je základní soubor. Funkce
X: E → R, Y: E → R, ..., Z: E → R,
které každému objektu přiřazují číslo, se nazývají (skalární) znaky.
Uspořádaná p-tice (X, Y, ..., Z) se nazývá vektorový znak.



Označení: Nechť je dán výběrový soubor $\{\varepsilon_1, ..., \varepsilon_n\} \subseteq E$. Hodnoty znaků X, Y, ..., Z pro i-tý objekt označíme $x_i = X(\varepsilon_i)$, $y_i = Y(\varepsilon_i)$, ..., $z_i = Z(\varepsilon_i)$, i = 1, ..., n.

Datový soubor

Pojem datového souboru:

Matice $\begin{pmatrix} x_1 & y_1 & \cdots & z_1 \\ x_2 & y_2 & \cdots & z_2 \\ \cdots & \cdots & \cdots & \cdots \\ x_n & y_n & \cdots & z_n \end{pmatrix}$ typu n × p se nazývá datový soubor. Její řádky odpovídají jednotlivým objektům, sloupce znakům.

Libovolný sloupec této matice nazýváme jednorozměrným datovým souborem. Jestliže uspořádáme hodnoty některého znaku (např. znaku X) v jednorozměrném datovém souboru vzestupně podle velikosti, dostaneme uspořádaný datový soubor

$\begin{pmatrix} x_{(1)} \\ \vdots \\ x_{(n)} \end{pmatrix}$, kde $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$.

Vektor $\begin{pmatrix} x_{[1]} \\ \vdots \\ x_{[r]} \end{pmatrix}$, kde $x_{[1]} < \ldots < x_{[r]}$ jsou navzájem různé hodnoty znaku X, se nazývá vektor variant.

Příklad

Příklad: Pro studenty z výběrového souboru uvedeného výše byly zjišťovány hodnoty znaků X – známka z matematiky v prvním zkušebním termínu, Y – známka z angličtiny v prvním zkušebním termínu, Z – pohlaví studenta (0 ... žena, 1 ... muž). Byl získán datový soubor

$$\begin{pmatrix} 2 & 2 & 0 \\ 1 & 3 & 1 \\ 4 & 3 & 1 \\ 1 & 1 & 0 \\ 1 & 2 & 1 \\ 4 & 4 & 1 \\ 3 & 3 & 1 \\ 3 & 4 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 4 & 2 & 1 \\ 4 & 4 & 0 \\ 2 & 2 & 0 \\ 4 & 3 & 1 \\ 2 & 3 & 1 \\ 4 & 4 & 0 \\ 1 & 1 & 0 \\ 4 & 3 & 1 \\ 4 & 4 & 1 \\ 1 & 3 & 0 \end{pmatrix}$$

Utvořte jednorozměrný uspořádaný i neuspořádaný datový soubor pro známky z matematiky a vektor variant pro známky z matematiky.

Řešení:

$$\begin{pmatrix} 2 \\ 1 \\ 4 \\ 1 \\ 1 \\ 4 \\ 3 \\ 3 \\ 1 \\ 1 \\ 4 \\ 4 \\ 2 \\ 4 \\ 2 \\ 4 \\ 1 \\ 4 \\ 4 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 2 \\ 2 \\ 2 \\ 3 \\ 3 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \end{pmatrix}, \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}$$

Jev

Pojem jevu: Nechť $\{\varepsilon_1, ..., \varepsilon_n\}$ je výběrový soubor, X, Y, ..., Z jsou znaky, B, B_1, …, B_p jsou číselné množiny.
Zápis $\{X \in B\}$ znamená jev „znak X nabyl hodnoty z množiny B“ .
Zápis $\{X \in B_1 \wedge Y \in B_2 \wedge ... \wedge Z \in B_p\}$ znamená jev „znak X nabyl hodnoty z množiny B_1 a současně znak Y nabyl hodnoty z množiny B_2 atd. až znak Z nabyl hodnoty z množiny B_p“.
Symbol $N(X \in B)$ značí absolutní četnost jevu $\{X \in B\}$ ve výběrovém souboru, tj. počet těch objektů ve výběrovém souboru, pro něž $x_i \in B$.
Symbol $p(X \in B)$ znamená relativní četnost jevu $\{X \in B\}$ ve výběrovém souboru, tj. $p(X \in B) = \frac{N(X \in B)}{n}$.
Analogicky $N(X \in B_1 \wedge Y \in B_2 \wedge ... \wedge Z \in B_p)$ resp. $p(X \in B_1 \wedge Y \in B_2 \wedge ... \wedge Z \in B_p)$ znamená absolutní resp. relativní četnost jevu $\{X \in B_1 \wedge Y \in B_2 \wedge ... \wedge Z \in B_p\}$ ve výběrovém souboru.

Příklad

Příklad : Pro datový soubor s údaji o známkách najděte relativní četnost

a) matematických jedničkářů
b) úspěšných matematiků
c) oboustranně neúspěšných studentů.

Datový soubor má tvar:

$$\begin{pmatrix} 2 & 2 & 0 \\ 1 & 3 & 1 \\ 4 & 3 & 1 \\ 1 & 1 & 0 \\ 1 & 2 & 1 \\ 4 & 4 & 1 \\ 3 & 3 & 1 \\ 3 & 4 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 4 & 2 & 1 \\ 4 & 4 & 0 \\ 2 & 2 & 0 \\ 4 & 3 & 1 \\ 2 & 3 & 1 \\ 4 & 4 & 0 \\ 1 & 1 & 0 \\ 4 & 3 & 1 \\ 4 & 4 & 1 \\ 1 & 3 & 0 \end{pmatrix}$$

Řešení:

ad a) $p(X = 1) = \frac{7}{20} = 0{,}35;$

ad b) $p(X \leq 3) = \frac{12}{20} = 0{,}60;$

ad c) $p(X = 4 \wedge Y = 4) = \frac{4}{20} = 0{,}20.$

Zjistili jsme, že jedničku z matematiky mělo 35% studentů, zkoušku z matematiky úspěšně složilo 60% studentů a oboustranně Neúspěšných bylo 20% studentů.

Jednorozměrné bodové rozložení četností

Jestliže počet variant znaku X v jednorozměrném datovém souboru není příliš velký, pak přiřazujeme četnosti jednotlivým variantám a hovoříme o bodovém rozložení četností .

Nechť je dán jednorozměrný datový soubor $\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$, v němž znak X nabývá r variant .

Pro j = 1, ..., r definujeme:

$n_j = N(X = x_{[j]})$ – absolutní četnost varianty $x_{[j]}$ ve výběrovém souboru

$p_j = \frac{n_j}{n}$ - relativní četnost varianty $x_{[j]}$ ve výběrovém souboru

$N_j = N(X \leq x_{[j]}) = n_1 + ... + n_j$ – absolutní kumulativní četnost prvních j variant ve výběrovém souboru

$F_j = \frac{N_j}{n} = p_1 + ... + p_j$ – relativní kumulativní četnost prvních j variant ve výběrovém souboru

Tabulka typu

$x_{[j]}$	n_j	p_j	N_j	F_j
$x_{[1]}$	n_1	p_1	N_1	F_1
⋮	⋮	⋮	⋮	⋮
$x_{[r]}$	n_r	p_r	N_r	F_r

se nazývá variační řada (nebo též tabulka rozložení četností).

Příklad

Příklad: Máme jednorozměrný datový soubor, který obsahuje údaje o známkách z matematiky (znak X) u 20 studentů.

$$\begin{pmatrix} 2 \\ 1 \\ 4 \\ 1 \\ 1 \\ 4 \\ 3 \\ 3 \\ 1 \\ 1 \\ 4 \\ 4 \\ 2 \\ 4 \\ 2 \\ 4 \\ 1 \\ 4 \\ 4 \\ 1 \end{pmatrix}$$

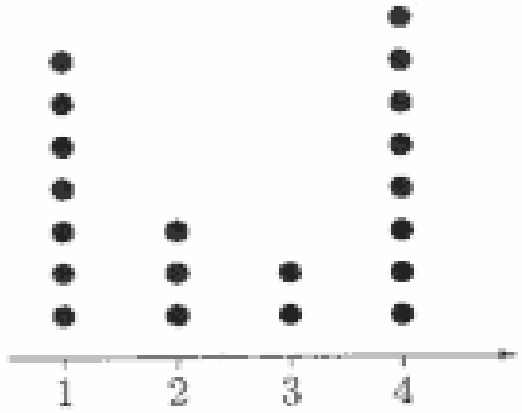
Sestavte tabulku rozložení četností.

Řešení:

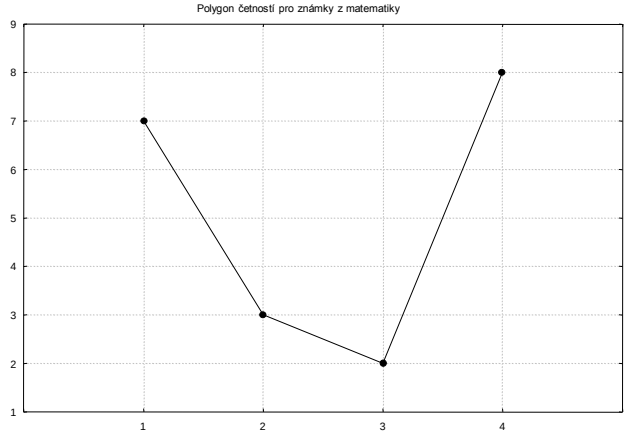
$x_{[j]}$	n_j	p_j	N_j	F_j
1	7	7/20=0,35	7	7/20=0,35
2	3	3/20=0,15	10	10/20=0,50
3	2	2/20=0,10	12	12/20=0,60
4	8	8/20=0,40	20	20/20=1,00
∑	20	1,00	-	-

Grafické znázornění jednorozměrného bodového rozdělení četností

Tečkový diagram : na číselné ose vyznačíme jednotlivé varianty znaku X a nad každou variantu nakreslíme tolik teček, jaká je její absolutní četnost.

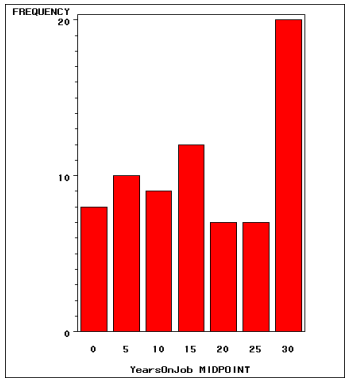


Polygon četnosti : je lomená čára spojující body, jejichž x -ová souřadnice je varianta znaku X a y -ová souřadnice je absolutní či relativní četnost této varianty.

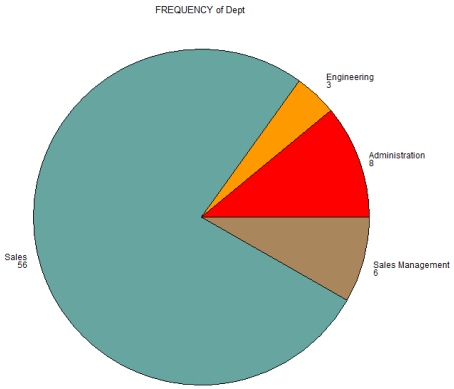


Grafické znázornění jednorozměrného bodového rozdělení četností

Sloupkový diagram : je soustava na sebe nenavazujících obdélníků, kde střed základny je varianta znaku X a výška je absolutní či relativní četnost této varianty.



Výsečový graf : je kruh rozdělený na výseče, jejichž vnější obvod odpovídá absolutním (relativním) četnostem variant znaku X.



Dvourozměrné bodové rozložení četností

Nechť je dán dvourozměrný datový soubor $\begin{pmatrix} x_1 & y_1 \\ \dots & \dots \\ x_n & y_n \end{pmatrix}$, kde znak X má r variant a znak Y má s variant. Pak definujeme:

$n_{jk} = N(X = x_{[j]} \wedge Y = y_{[k]})$ – simultánní absolutní četnost dvojice $(x_{[j]}, y_{[k]})$ ve výběrovém souboru

$p_{jk} = \frac{n_{jk}}{n}$ – simultánní relativní četnost dvojice $(x_{[j]}, y_{[k]})$ ve výběrovém souboru

$n_{j.} = N(X = x_{[j]}) = n_{j1} + \dots + n_{js}$ – marginální absolutní četnost varianty $x_{[j]}$

$p_{j.} = \frac{n_{j.}}{n} = p_{j1} + \dots + p_{js}$ – marginální relativní četnost varianty $x_{[j]}$

$n_{.k} = N(Y = y_{[k]}) = n_{1k} + \dots + n_{rk}$ – marginální absolutní četnost variantyn $y_{[k]}$

$p_{.k} = \frac{n_{.k}}{n} = p_{1k} + \dots + p_{rk}$ – marginální relativní četnost varianty $y_{[k]}$

$N_{jk} = N(X \le x_{[j]} \wedge Y \le y_{[k]}) = \sum_{u \le j} \sum_{v \le k} n_{uv}$ Absolutní kumulativní četnost dvojice $(x_{[j]}, y_{[k]})$

Simultánní četností zapisujeme do kontingenční tabulky.

Kontingenční tabulka simultánních absolutních četností má tvar:

	y	$y_{[1]}$	…	$y_{[s]}$	$n_{j.}$
x	n_{jk}				
$x_{[1]}$		n_{11}	…	n_{1s}	$n_{1.}$
⋮		…	…	…	…
$x_{[r]}$		n_{r1}	…	n_{rs}	$n_{r.}$
$n_{.k}$		$n_{.1}$	…	$n_{.s}$	n

Příklad

Příklad: Máme datový soubor, který obsahuje údaje o známkách z matematiky (znak X), z angličtiny (znak Y) a pohlaví studenta (znak Z, 0 – žena , 1 – muž) u 20 studentů:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
X	2	1	4	1	1	4	3	3	1	1	4	4	2	4	2	4	1	4	4	1
Y	2	3	3	1	2	4	3	4	1	1	2	4	2	3	3	4	1	3	4	3
Z	0	1	1	0	1	1	1	0	0	0	1	0	0	1	1	0	0	1	1	0

Vytvořte kontingenční tabulku simultánních absolutních a relativních četností pro známky z matematiky a angličtiny.

Řešení:

Kontingenční tabulka simultánních absolutních četností

x \ y	1	2	3	4	$n_{j\cdot}$
1	4	1	2	0	7
2	0	2	1	0	3
3	0	0	1	1	2
4	0	1	3	4	8
$n_{\cdot k}$	4	4	7	5	$n = 20$

Kontingenční tabulka simultánních relativních četností

x \ y	1	2	3	4	$p_{j\cdot}$
1	0,20	0,05	0,10	0,00	0,35
2	0,00	0,10	0,05	0,00	0,15
3	0,00	0,00	0,05	0,05	0,10
4	0,00	0,05	0,15	0,20	0,40
$p_{\cdot k}$	0,20	0,20	0,35	0,25	1,00

Řádkově a sloupcově podmíněné relativní četnosti

Sloupcově podmíněná relativní četnost varianty $x_{[j]}$ za předpokladu $y_{[k]}$

$$p_{j(k)} = \frac{n_{jk}}{n_{\cdot k}}$$

Řádkově podmíněná relativní četnost varianty $y_{[k]}$ za předpokladu $x_{[j]}$

$$p_{(j)k} = \frac{n_{jk}}{n_{j\cdot}}$$

Příklad

Příklad: Pro datový soubor známek z matematiky a angličtiny sestavte kontingenční tabulku sloupcově a poté řádkově podmíněných relativních četností.

Řešení:

Nejprve se budeme zabývat sloupcově podmíněnými relativními četnostmi. Použijeme vzorec $p_{j(k)} = \frac{n_{jk}}{n_{.k}}$.

Vyjdeme z kontingenční tabulky simultánních absolutních četností.

x \ y	1	2	3	4	$n_{j.}$
1	4	1	2	0	7
2	0	2	1	0	3
3	0	0	1	1	2
4	0	1	3	4	8
$n_{.k}$	4	4	7	5	$n = 20$

x \ y	1	2	3	4
1	1,00	0,25	0,29	0,00
2	0,00	0,50	0,14	0,00
3	0,00	0,00	0,14	0,20
4	0,00	0,25	0,43	0,80
$\sum$	1,00	1,00	1,00	1,00

Interpretujeme např. třetí sloupec: z těch studentů, kteří měli trojku z angličtiny, mělo 2/7 = 29% jedničku z matematiky, 1/7 = 14% dvojku z matematiky, 1/7 = 14% trojku z matematiky a 3/7 = 43% čtyřku z matematiky.

Příklad

Dále se budeme zabývat řádkově podmíněnými relativními četnostmi.

Použijeme vzorec $p_{(j)k} = \frac{\mathbf{n_{jk}}}{\mathbf{n_{j.}}}$.

Opět nám poslouží kontingenční tabulka absolutních četností.

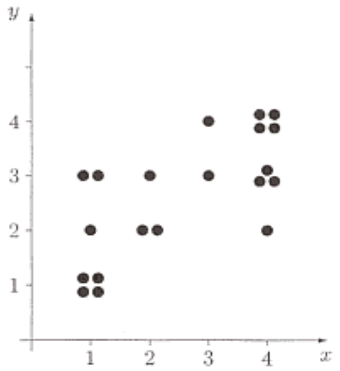
x \ y	1	2	3	4	$n_{j\cdot}$
1	4	1	2	0	7
2	0	2	1	0	3
3	0	0	1	1	2
4	0	1	3	4	8
$n_{\cdot k}$	4	4	7	5	$n = 20$

x \ y	1	2	3	4	$\sum$
1	0,57	0,14	0,29	0,00	1,00
2	0,00	0,67	0,33	0,00	1,00
3	0,00	0,00	0,50	0,50	1,00
4	0,00	0,12	0,38	0,50	1,00

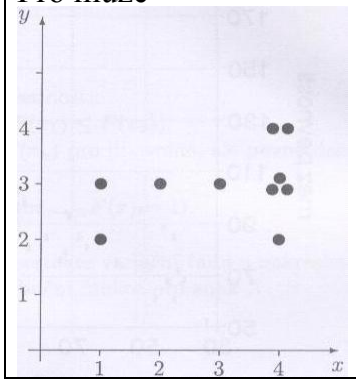
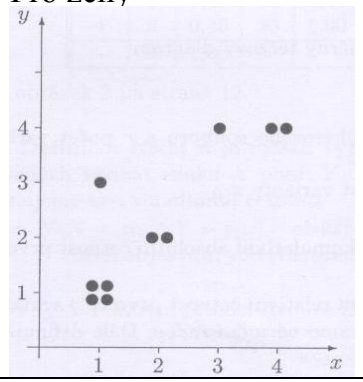
Interpretujeme např. první řádek: z těch studentů, kteří měli jedničku z matematiky, mělo 4/7 = 57% jedničku z angličtiny, 1/7 = 14% dvojku z angličtiny a 2/7 = 29% trojku z angličtiny.

Dvourozměrný tečkový diagram

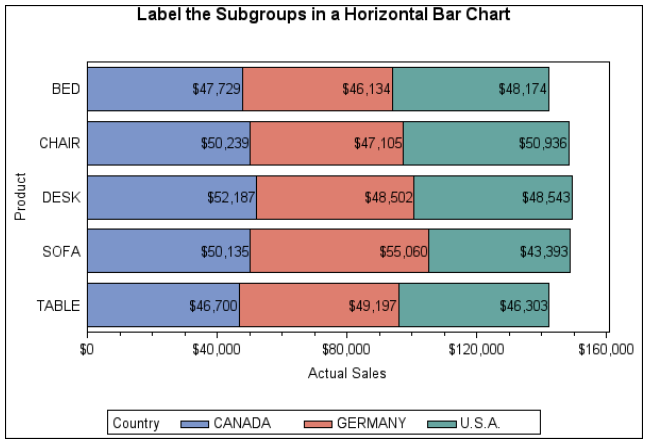
Dvourozměrné rozložení četností lze znázornit pomocí dvourozměrného tečkovéhodiagramu . Na vodorovnou osu vyneseme varianty znaku X, na svislou varianty znaku Y a do příslušných průsečíků nakreslíme tolik teček, jaká je absolutní četnost dané dvojice.
V našem příkladě se studenty dostaneme tento diagram:

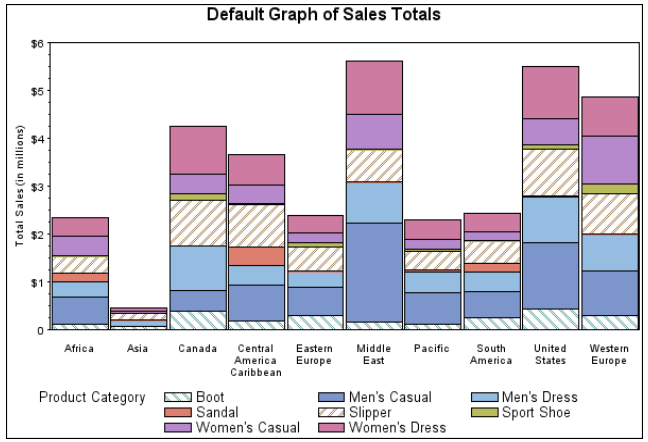


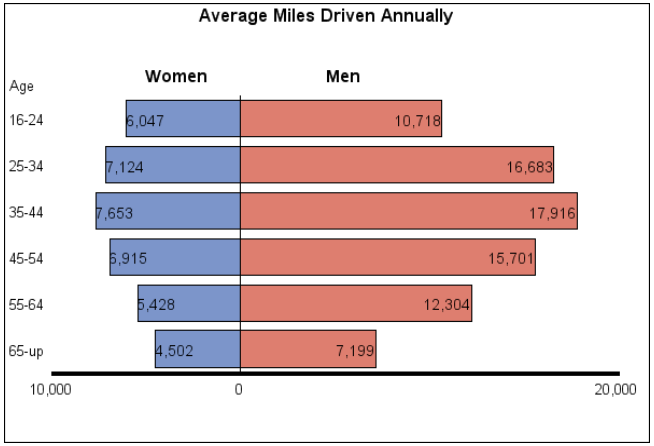
Dvourozměrný tečkový diagram svědčí o nepříliš výrazné tendenci k podobné klasifikaci v obou předmětech.
Zcela odlišný vzhled má diagram pro muže a pro ženy:

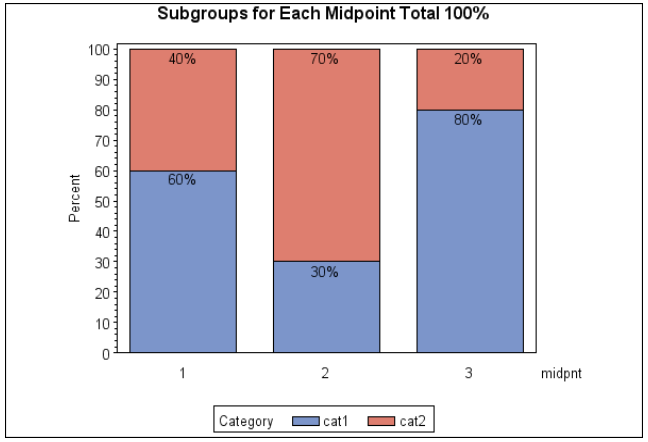
Pro muže	Pro ženy
	

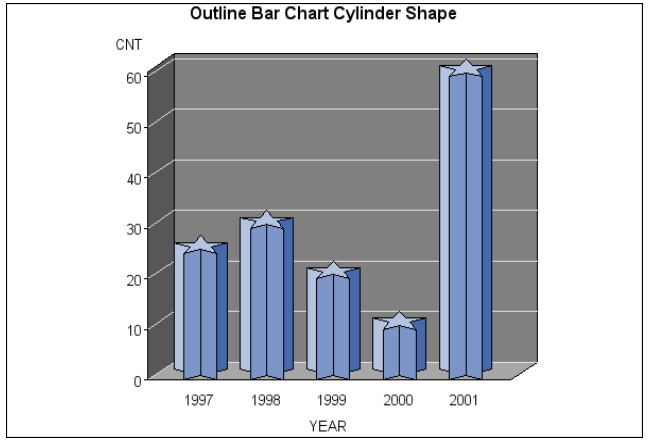
Další možnosti grafického znázornění

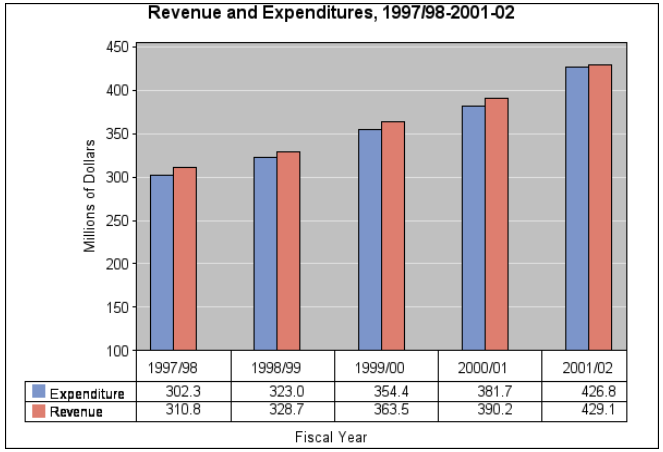




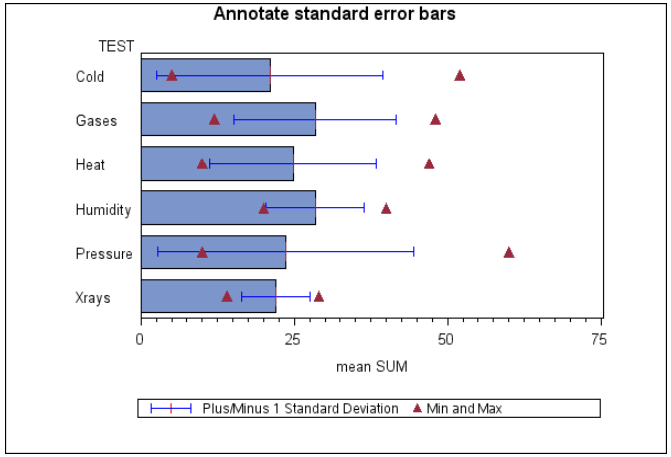


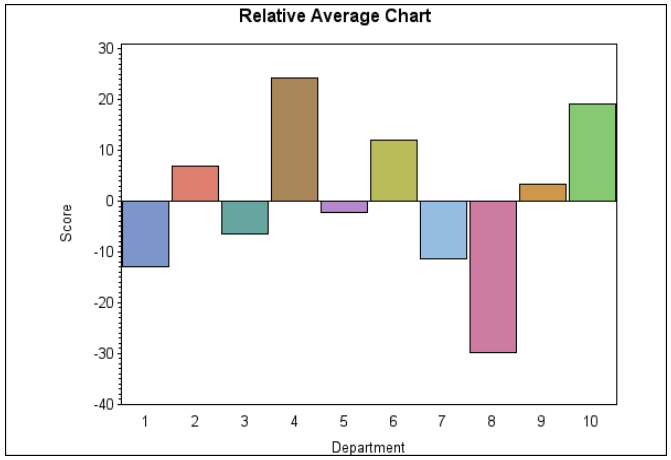


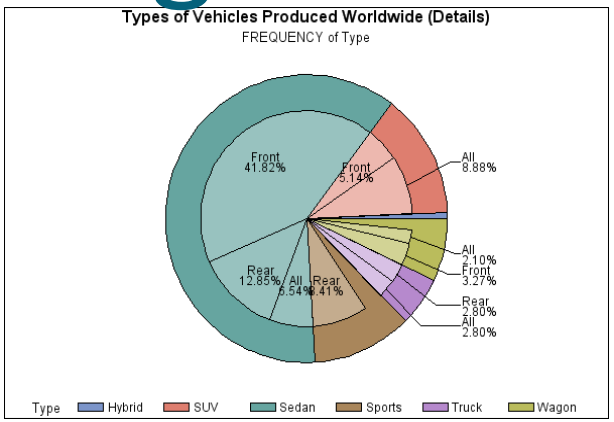


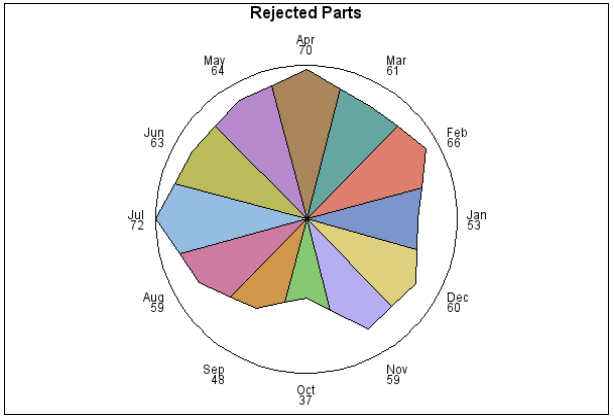


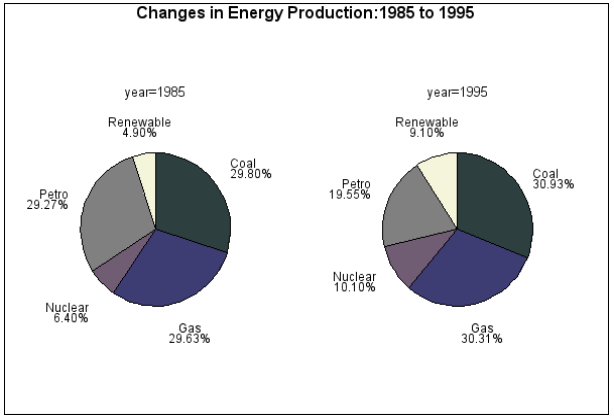
Další možnosti grafického znázornění

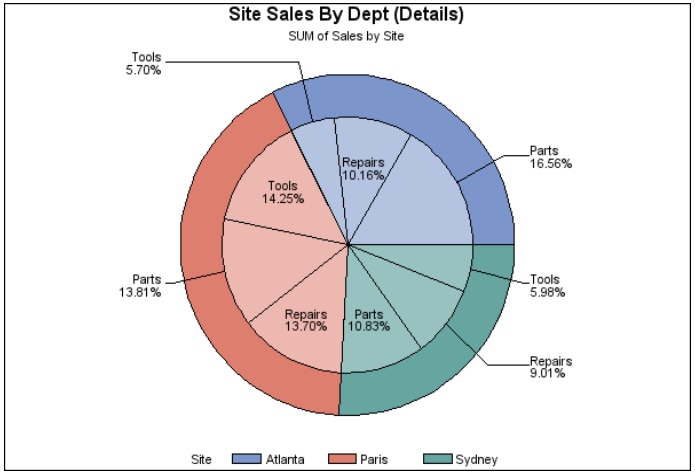


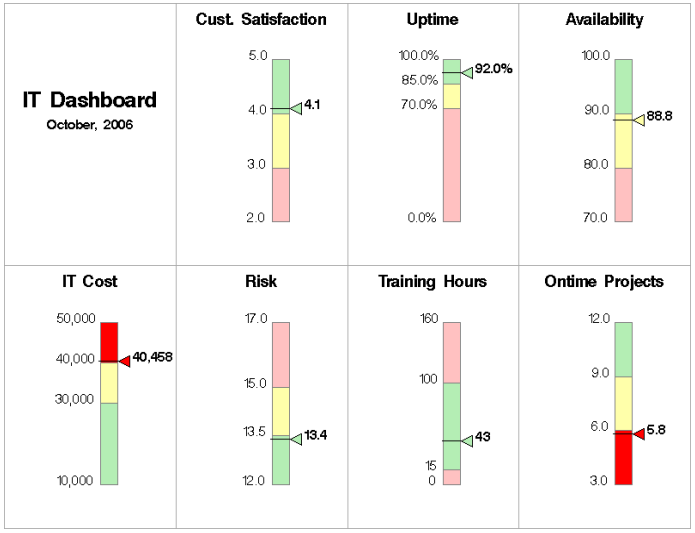




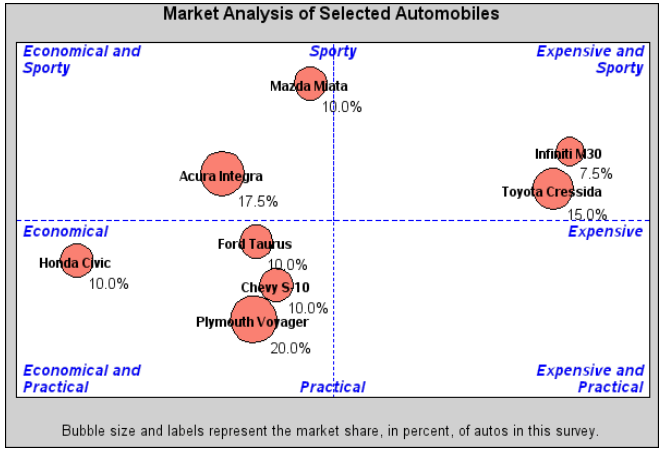


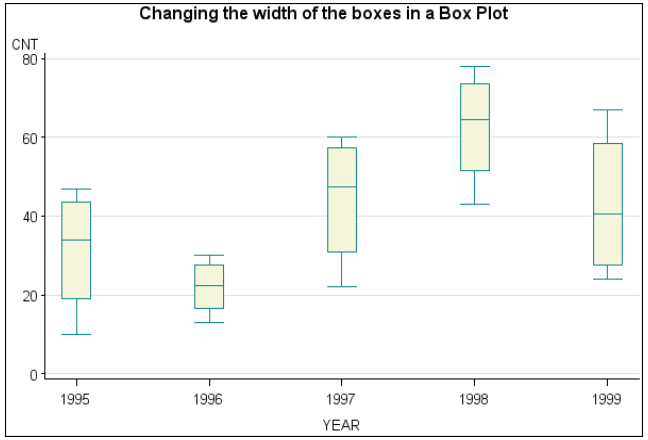


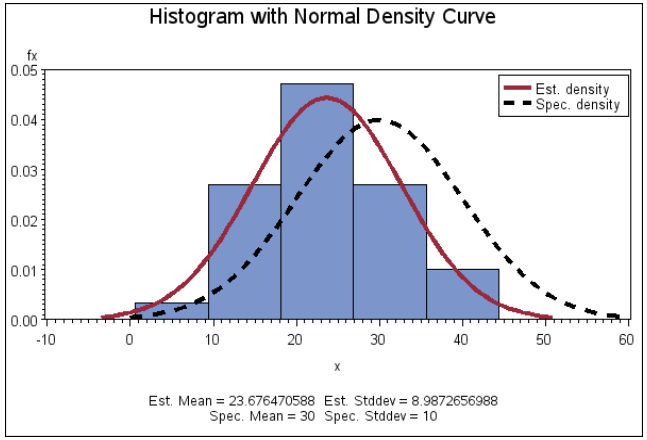


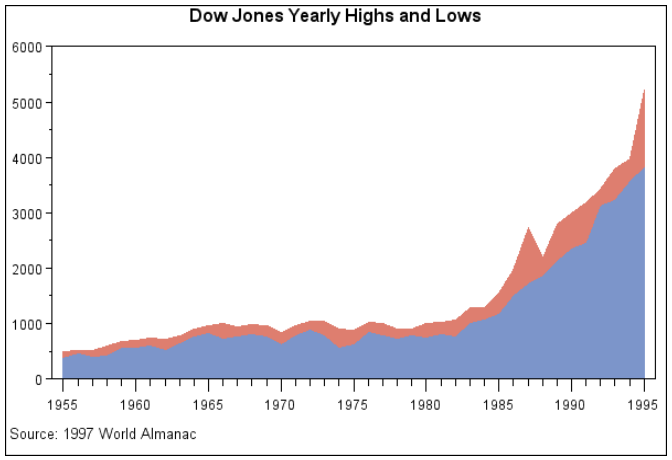


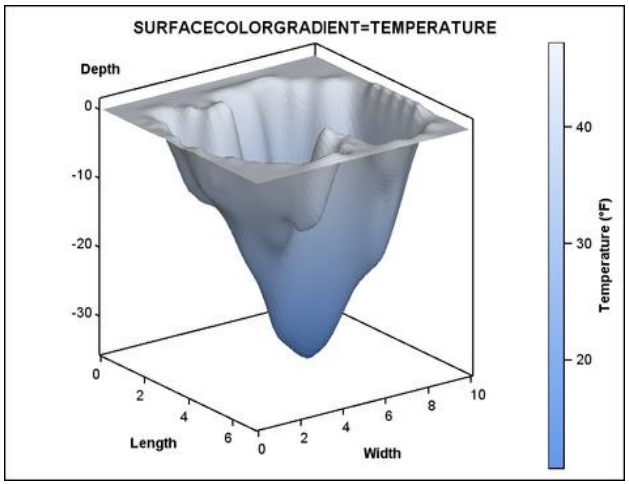
Další možnosti grafického znázornění

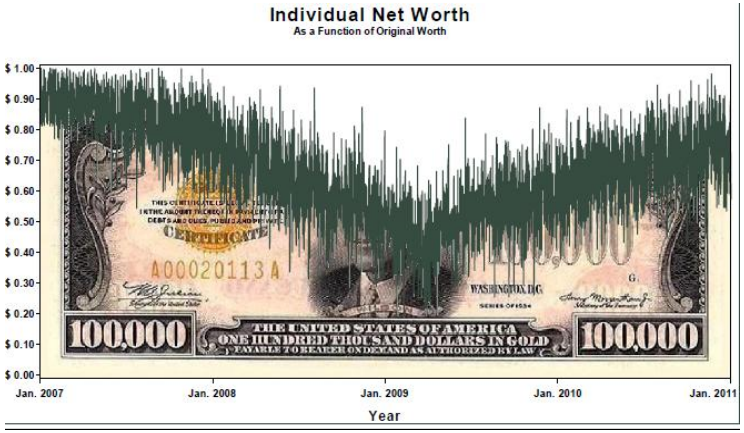












Intervalové rozložení četností

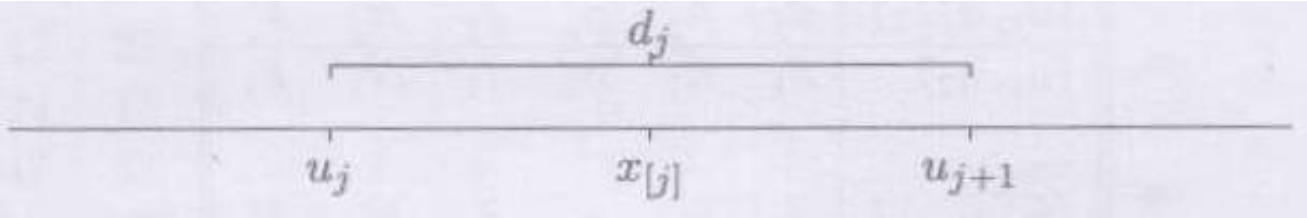
Nechť je dán jednorozměrný datový soubor. Jestliže počet variant znaku X je blízký rozsahu souboru, pak četnosti přiřazujeme nikoliv jednotlivým variantám, ale celým intervalům hodnot. Hovoříme pak o intervalovém rozložení četnosti.

Číselnou osu rozložíme na intervaly typu $(-\infty, u_1\rangle$, $(u_1, u_2\rangle$, ..., $(u_r, u_{r+1}\rangle$, (u_{r+1}, ∞) tak, aby okrajové intervaly neobsahovaly žádnou pozorovanou hodnotu znaku X. Užíváme označení:

$(u_j, u_{j+1}\rangle$ – j-tý třídicí interval znaku X, j = 1, ..., r.

$d_j = u_{j+1} - u_j$ – délka j-tého třídicího intervalu znaku X

$x_{[j]} = \frac{u_j + u_{j+1}}{2}$ – střed j-tého třídicího intervalu znaku X



Intervalové rozložení četností – stanovení počtu tříd

Třídicí intervaly volíme nejčastěji stejně dlouhé. Jejich počet určíme např. pomocí Sturgesova pravidla: r = 1 + 3,3 log n, kde n je rozsah souboru.

- počet tříd (*r*):
  - do 100 prvků……6 až 9 tříd
  - do 500 prvků……10 až 15 tříd
  - nad 500 prvků……..Sturgesovo pravidlo

$$r \approx 1 + 3{,}3 \log n$$

log…dekadický logaritmus!!!

Sestavení tabulky rozložení četností

Hodnoty znaku X roztřídíme do r třídicích intervalů. Pro j = 1, ..., r definujeme:

$n_j = N(u_j < X \leq u_{j+1})$ – absolutní četnost j-tého třídicího intervalu ve výběrovém souboru

$p_j = \frac{n_j}{n}$ – relativní četnost j-tého třídicího intervalu ve výběrovém souboru

$f_j = \frac{p_j}{d_j}$ – četnostní hustota j-tého třídicího intervalu ve výběrovém souboru

$N_j = N(X \leq u_{j+1}) = n_1 + ... + n_j$ – absolutní kumulativní četnost prvních j třídicích intervalů ve výběrovém souboru

$F_j = \frac{N_j}{n} = p_1 + ... + p_j$ – relativní kumulativní četnost prvních j třídicích intervalů ve výběrovém souboru.

Tabulka typu

(u_j, u_{j+1})	d_j	n_j	p_j	f_j	N_j	F_j
(u_1, u_2)	d_1	n_1	p_1	f_1	N_1	F_1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
(u_r, u_{r+1})	d_r	n_r	p_r	f_r	N_r	F_r
Součet		n	1			

se nazývá tabulka rozložení četností.

Příklad

Příklad: Do laboratoře bylo dodáno 60 vzorků a byly zjištěny hodnoty znaku X – mez plasticity (v kp/cm^2) a Y – mez pevnosti (v kp/cm^2). Datový soubor má tvar:

X	Y	X	Y	X	Y
154	178	83	98	73	76
133	164	106	111	77	86
58	75	92	104	47	61
145	161	85	103	68	85
94	107	112	118	137	142
113	141	98	102	44	68
86	97	103	108	92	116
121	127	99	119	141	157
119	138	104	128	155	189
112	125	107	118	136	155
85	97	98	140	82	81
41	72	97	115	136	163
96	113	105	101	72	79
45	89	71	93	66	81
99	109	39	69	42	61
51	95	122	147	113	123
101	114	33	52	42	85
160	169	78	117	133	147
87	101	114	137	153	179
88	139	125	149	85	91

a) Pro znak X stanovte optimální počet třídicích intervalů dle Sturgersova pravidla.
b) Sestavte tabulku rozložení četností.

Příklad

Řešení:

ad a) Rozsah souboru je 60. Podle Sturgersova pravidla je optimální počet třídicích intervalů r = 7. Budeme tedy volit 7 intervalů stejné délky tak, aby v nich byly obsaženy všechny pozorované hodnoty znaku X, z nichž nejmenší je 33, největší 160; volba $u_1 = 30, \ldots, u_8 = 170$ splňuje požadavky.

ad b)

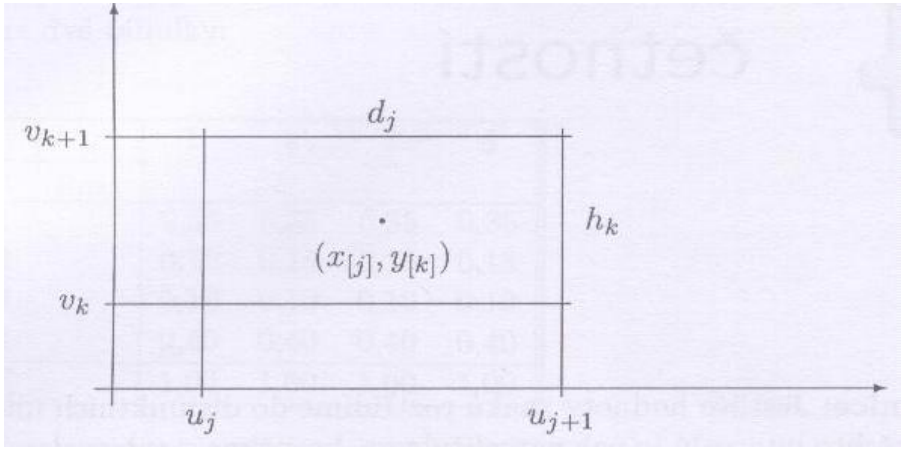
$\langle u_j, u_{j+1}\rangle$	d_j	$x_{[j]}$	n_j	p_j	N_j	F_j	f_j
$\langle 30, 50\rangle$	20	40	8	$8/60 = 0{,}1\overline{3}$	8	$8/60 = 0{,}1\overline{3}$	$8/(60\cdot 20) = 0{,}00\overline{6}$
$\langle 50, 70\rangle$	20	60	4	$4/60 = 0{,}0\overline{6}$	12	$12/60 = 0{,}2$	$4/(60\cdot 20) = 0{,}00\overline{3}$
$\langle 70, 90\rangle$	20	80	13	$13/60 = 0{,}21\overline{6}$	25	$25/60 = 0{,}41\overline{6}$	$13/(60\cdot 20) = 0{,}01083$
$\langle 90, 110\rangle$	20	100	15	$15/60 = 0{,}25$	40	$40/60 = 0{,}\overline{6}$	$15/(60\cdot 20) = 0{,}0125$
$\langle 110, 130\rangle$	20	120	9	$9/60 = 0{,}15$	49	$49/60 = 0{,}81\overline{6}$	$9/(60\cdot 20) = 0{,}0075$
$\langle 130, 150\rangle$	20	140	7	$7/60 = 0{,}11\overline{6}$	56	$56/60 = 0{,}9\overline{3}$	$7/(60\cdot 20) = 0{,}00583$
$\langle 150, 170\rangle$	20	160	4	$4/60 = 0{,}0\overline{6}$	60	$60/60 = 1$	$4/(60\cdot 20) = 0{,}00\overline{3}$
Součty			60	1			

Dvourozměrné intervalové rozložení četností

Dále se budeme věnovat dvourozměrnému intervalovému rozložení četností, tj. budeme pracovat s dvourozměrným datovým souborem. Zavedeme podobné pojmy jako u dvourozměrného bodového rozložení četností

Nechť je dán dvourozměrný datový soubor $\begin{pmatrix} x_1 & y_1 \\ \dots & \dots \\ x_n & y_n \end{pmatrix}$, kde hodnoty znaku X roztřídíme do r třídicích intervalů $\langle u_j, u_{j+1}\rangle$, j = 1, ..., r s délkami $d_1, \dots, d_r$ a hodnoty znaku Y roztřídíme do *s* třídicích intervalů $\langle v_k, v_{k+1}\rangle$, k = 1, ..., s s délkami $h_1, \dots, h_s$.

Obdélník $\langle u_j, u_{j+1}\rangle \times \langle v_k, v_{k+1}\rangle$ se nazývá (j,k) - tý dvourozměrný třídicí interval.



Simultánní a marginální četnosti

$n_{jk} = N(u_j < X \leq u_{j+1} \wedge v_k < Y \leq v_{k+1})$ – simultánní absolutní četnost (j, k)-tého třídicího intervalu.

$p_{jk} = \frac{n_{jk}}{n}$ – simultánní relativní četnost (j, k)-tého třídicího intervalu.

$n_{j.} = n_{j1} + \ldots + n_{js}$ – marginální absolutní četnost j-tého třídicího intervalu pro znak X.

$p_{j.} = \frac{n_{j.}}{n}$ – marginální relativní četnost j-tého třídicího intervalu pro znak X.

$n_{.k} = n_{1k} + \ldots + n_{rk}$ – marginální absolutní četnost k-tého třídicího intervalu pro znak Y.

$p_{.k} = \frac{n_{.k}}{n}$ – marginální relativní četnost k-tého třídicího intervalu pro znak Y.

$f_{jk} = \frac{p_{jk}}{d_j h_k}$ – simultánní četnostní hustota v (j, k)-tém třídicím intervalu.

$f_{j.} = \frac{p_{j.}}{d_j}$ – marginální četnostní hustota v j-tém třídicím intervalu pro znak X.

$f_{.k} = \frac{p_{.k}}{h_k}$ – marginální četnostní hustota v k-tém třídicím intervalu pro znak Y.

Kteroukoliv ze simultánních četností zapisujeme do kontingenční tabulky.

Kontingenční tabulka simultánních absolutních četností:

	(v_k, v_{k+1})	(v_1, v_2)	$\ldots$	(v_s, v_{s+1})	
(u_j, u_{j+1})	n_{jk}				$n_{j.}$
(u_1, u_2)		n_{11}	$\ldots$	n_{1s}	$n_{1.}$
$\vdots$					$\vdots$
(u_r, u_{r+1})		n_{r1}	$\ldots$	n_{rs}	$n_{r.}$
$n_{.k}$		$n_{.1}$	$\ldots$	$n_{.s}$	n

Příklad

Příklad: Pro datový soubor obsahující údaje o mezi plasticity (znak X) a mezi pevnosti (znak Y) oceli

a) stanovte dle Sturgersova pravidla optimální počet třídicích intervalů pro znak Y

b) sestavte kontingenční tabulku simultánních absolutních četností.

Řešení:

ad a) Rozsah datového souboru je 60. Podle Sturgersova pravidla je tedy optimální počet třídicích intervalů 7. Nejmenší hodnota je 52 a největší 189. Volíme $v_1 = 50$, $v_2 = 70$, ..., $v_8 = 190$.

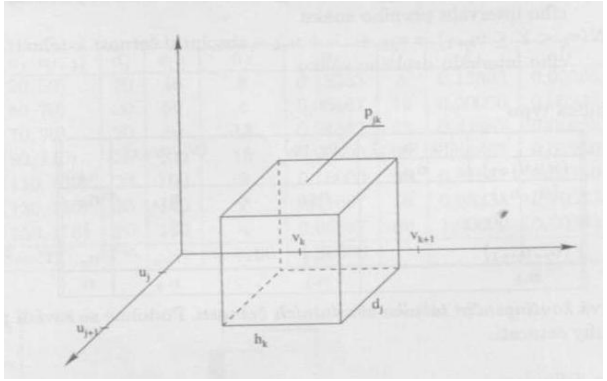
ad b)

$\langle u_j, u_{j+1}\rangle$ \ $\langle v_k, v_{k+1}\rangle$	n_{jk}	$\langle 50, 70\rangle$	$\langle 70, 90\rangle$	$\langle 90, 110\rangle$	$\langle 110, 130\rangle$	$\langle 130, 150\rangle$	$\langle 150, 170\rangle$	$\langle 170, 190\rangle$	$n_{j.}$
$\langle 30, 50\rangle$		5	3	0	0	0	0	0	8
$\langle 50, 70\rangle$		0	3	1	0	0	0	0	4
$\langle 70, 90\rangle$		0	4	7	1	1	0	0	13
$\langle 90, 110\rangle$		0	0	6	8	1	0	0	15
$\langle 110, 130\rangle$		0	0	0	4	5	0	0	9
$\langle 130, 150\rangle$		0	0	0	0	2	5	0	7
$\langle 150, 170\rangle$		0	0	0	0	0	1	3	4
$n_{.k}$		5	10	14	13	9	6	3	$n = 60$

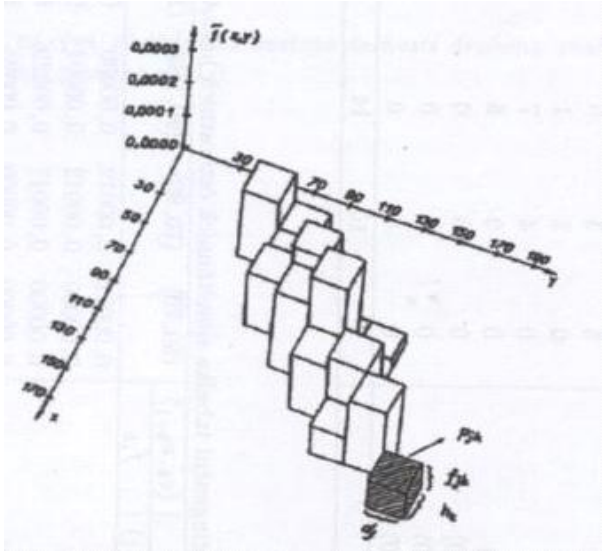
Stereogram

Dvourozměrné intervalové rozložení četností graficky znázorňujeme pomocí stereogramu. Je to graf skládající se z r * s kvádrů, sestrojených nad dvourozměrnými třídicími intervaly, přičemž objem (j, k) - tého kvádru je roven relativní četnosti p_{jk} (j, k) -tého třídicího intervalu,

j = 1, ..., r, k = 1, …, s. Výška kvádru tedy vyjadřuje simultánní četnostní hustotu.

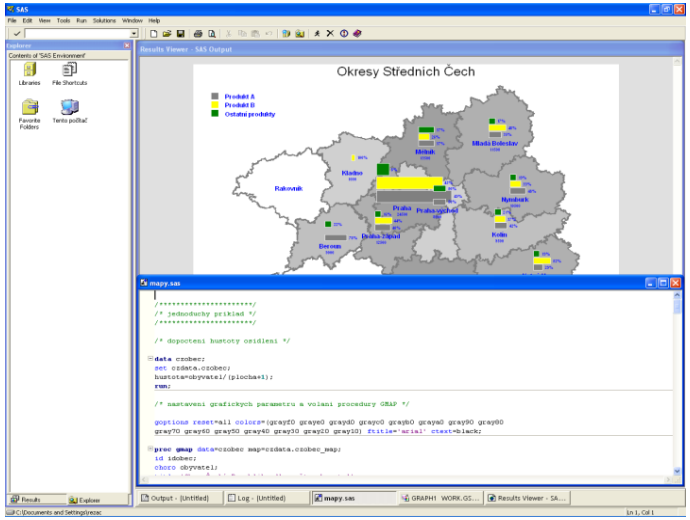


V našem příkladě s mezí plasticity
a mezí pevnosti oceli bude mít stereogram tvar:

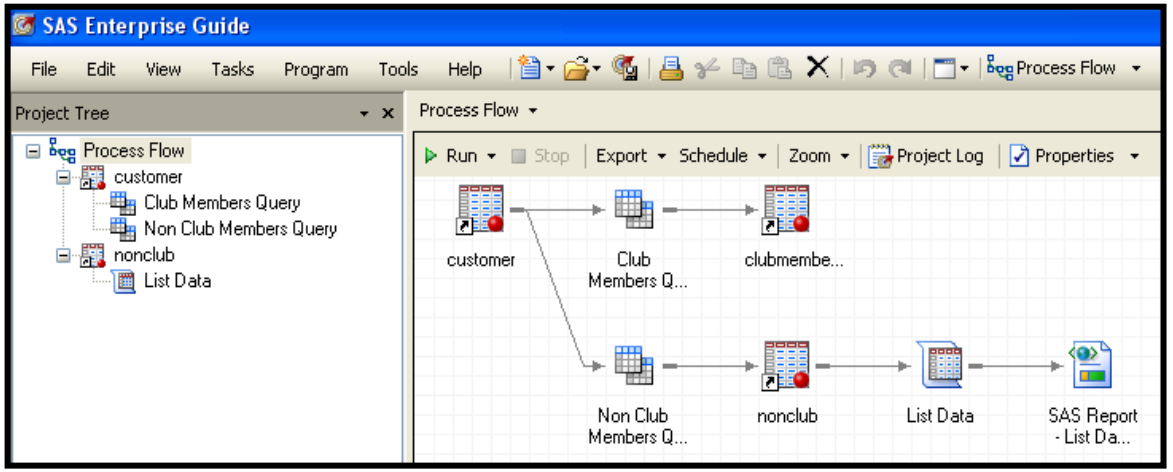


SAS - stručné seznámení

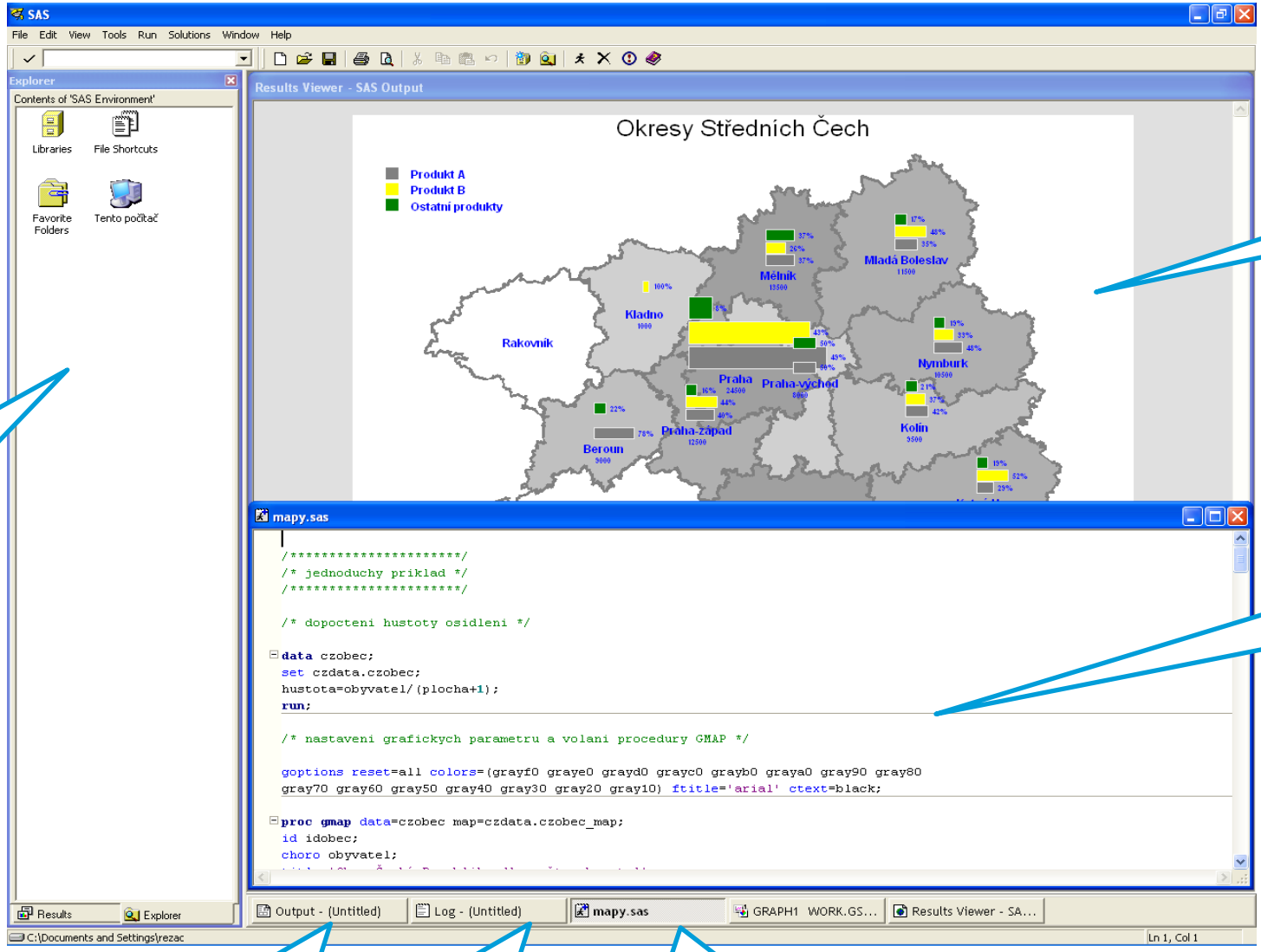
- 2 základní SAS rozhraní:
  - SAS windowing environment



  - SAS Enterprise Guide (GUI)



SAS - stručné seznámení



SAS Output

SAS Explorer window

Program editor window

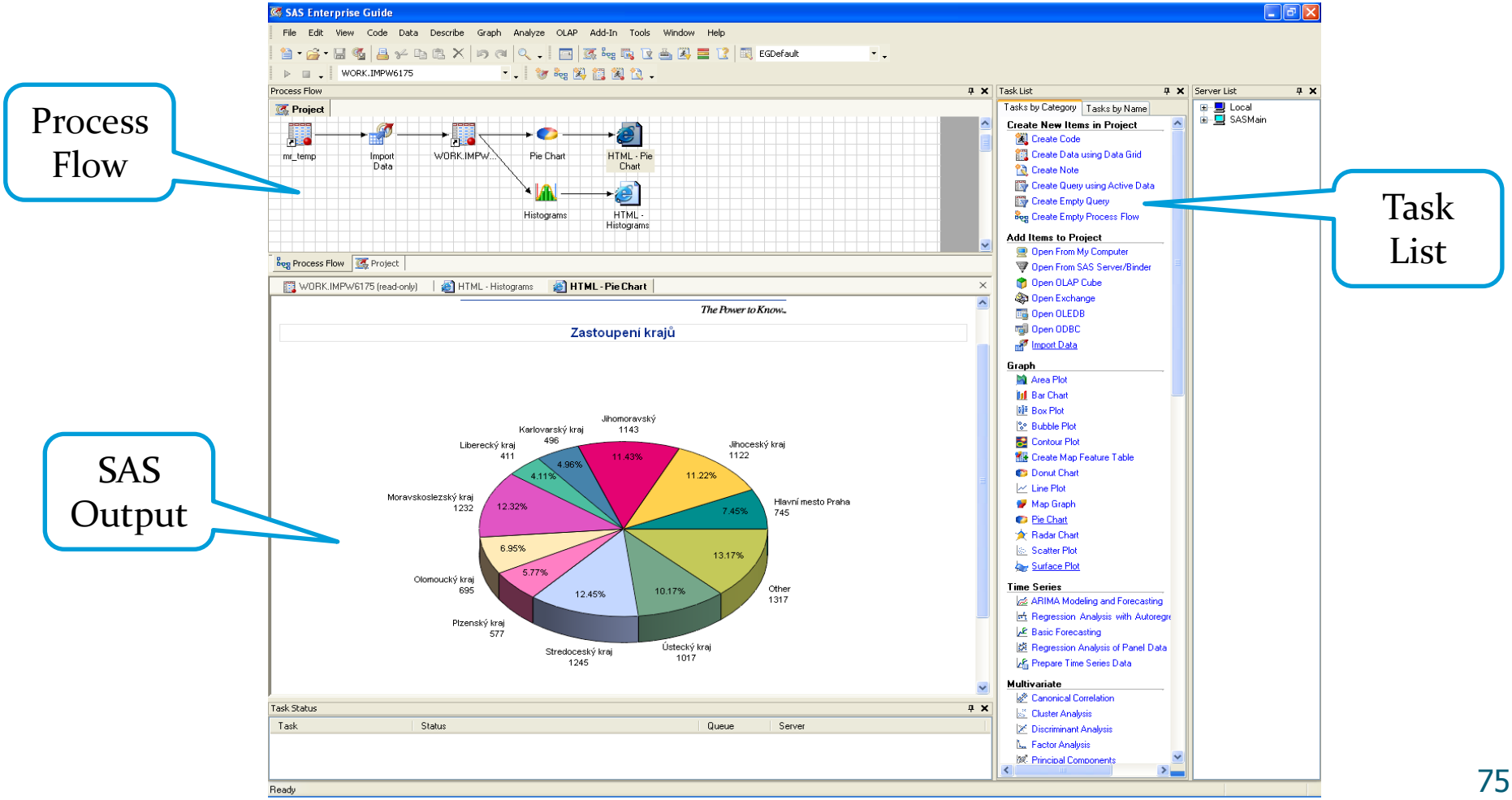
Output tab

Log tab

Editor tab

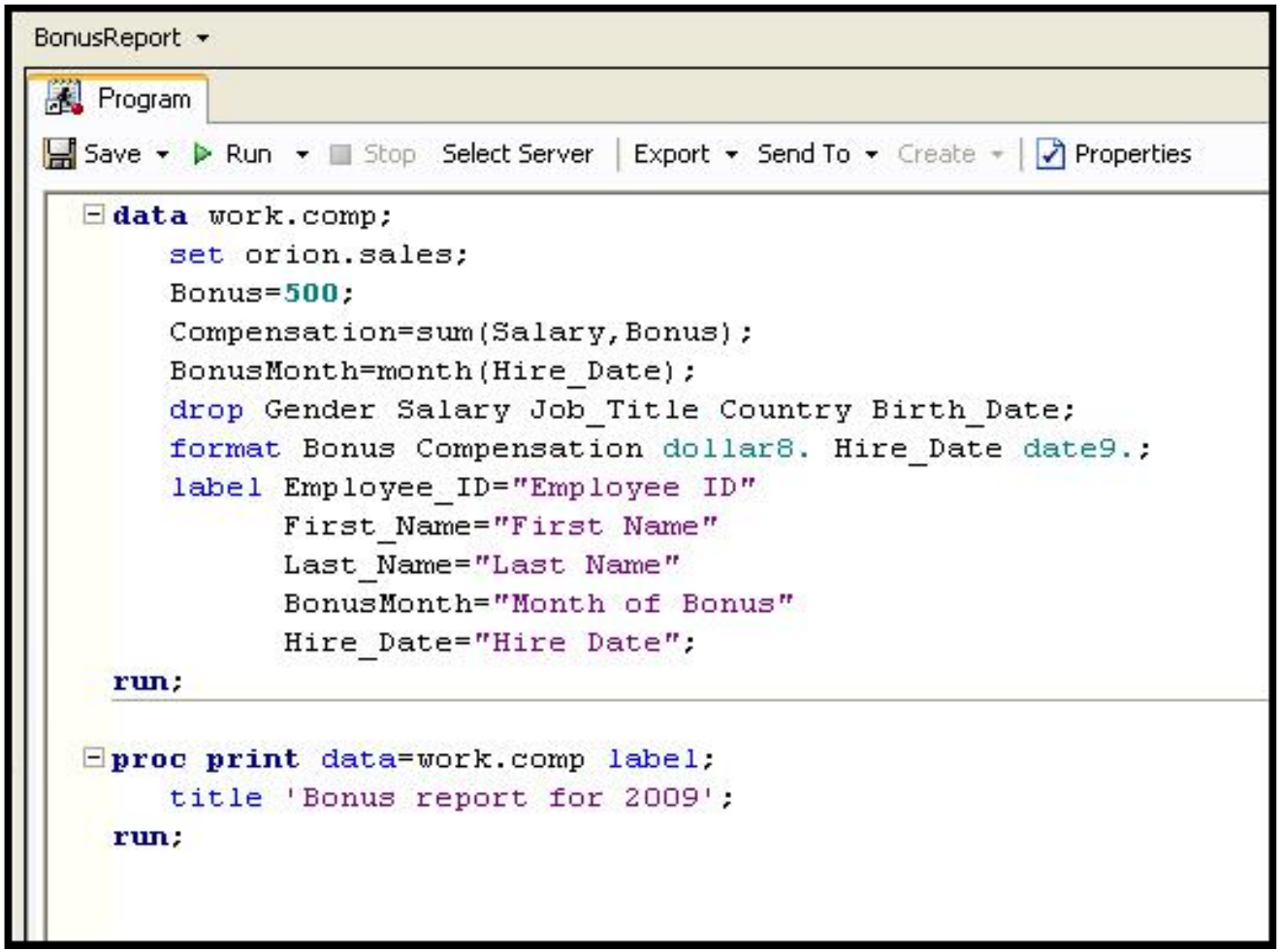
SAS - stručné seznámení

- Pomocí klikání a přetahování myší je budován procesní tok.



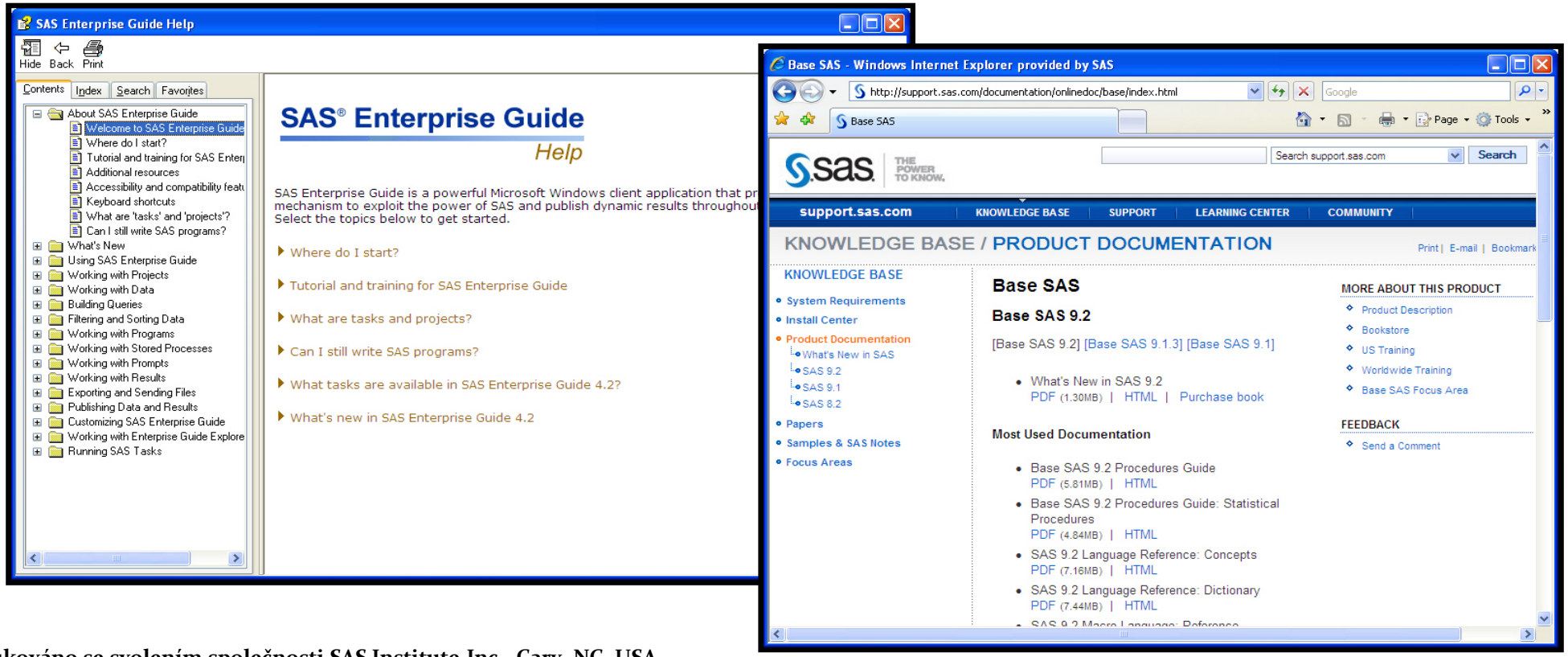
SAS Enterprise Guide (EG) Interface

- EG automaticky generuje kód, který možné dále editovat



SAS Help

- Use the SAS Enterprise Guide Help facility or SAS OnlineDoc for additional direction on SAS Enterprise Guide or the SAS programming language. Go to support.sas.com and select
- **Product Documentation** ⇨ **Base SAS**.



- **SAS používají např.:**



Více na http://www.sas.com/offices/europe/czech/reference/

SAS na webu

Michal Kulich: *Malý manuál uživatele SASu*
http://www.karlin.mff.cuni.cz/~kulich/sas/SASMain.html

Phil Spector: *An Introduction to the SAS System*
http://www.stat.berkeley.edu/classes/s100/sas.pdf

Patric McLeod : *Introduction to SAS 9*
http://www.unt.edu/rss/class/sas1/

http://en.wikipedia.org/wiki/SAS_%28software%29

Sofware SAS

Aktuálně k dispozici:

- SAS 9.3 TS1M2, Rev. 930_12w41 for
  - Microsoft® Windows® Workstation & Server 32-bit
  - Microsoft® Windows® Server & Workstation for x64
  - Linux® for X64
    - SAS EAS
    - Credit Scoring for SAS Enterprise Miner
    - SAS Enterprise Guide
    - SAS Enterprise Miner Personal Client
    - SAS Enterprise Miner Server, including the products:
      - SAS Enterprise Guide
    - SAS Forecast Server
    - SAS Metadata Server
    - SAS Text Analytics for Czech
    - SAS Text Miner Server
- JMP Pro (Microsoft® Windows® for x64, JMP 10.0.1 TS1M2, Rev. 930_12w41)

Sofware SAS

- **SAS EAS:**

Education **A**nalytical **S**uite = Base SAS®, SAS/ACCESS® rozhraní (pro všechny databáze), SAS/AF®, SAS/ASSIST®, SAS® Bridge for ESRI, SAS/CONNECT®, SAS/EIS®, SAS® Enterprise Guide®, SAS/ETS®, SAS/FSP®, SAS/GRAPH®, SAS/IML®, SAS/INSIGHT®, SAS/Integration Technologies®, SAS/LAB®, SAS/OR®, SAS/QC®, SAS/SECURE®, SAS/SHARE®, SAS/STAT®

Instalační soubory, licenční podmínky

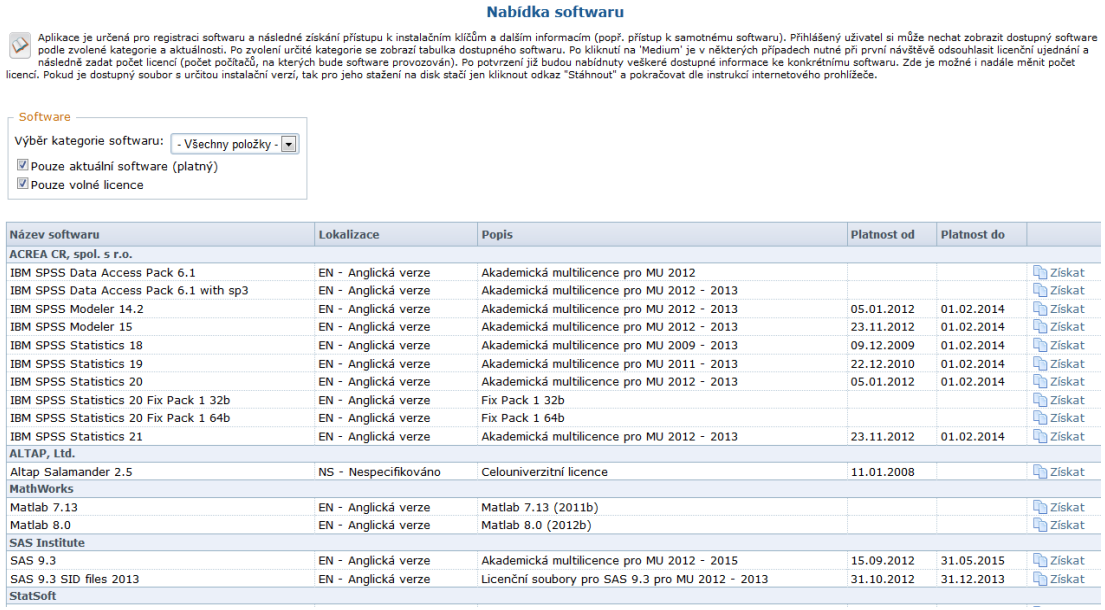
- Instalační soubory SASu (v.9.3) jsou k dispozici všem studentům a učitelům MU na adrese

https://inet.muni.cz/app/soft/licence

.

- Před vlastním zobrazením stránky s inst. soubory je nutné odsouhlasit licenční podmínky.

- Plný instalační depot 23 GB!



Nabídka softwaru

Aplikace je určená pro registraci softwaru a následné získání přístupu k instalačním klíčům a dalším informacím (popř. přístup k samotnému softwaru). Přihlášený uživatel si může nechat zobrazit dostupný software podle zvolené kategorie a aktuálnosti. Po zvolení určité kategorie se zobrazí tabulka dostupného softwaru. Po kliknutí na 'Medium' je v některých případech nutné při první návštěvě odsouhlasit licenční ujednání a následně zadat počet licencí (počet počítačů, na kterých bude software provozován). Po potvrzení již budou nabídnuty veškeré dostupné informace ke konkrétnímu softwaru. Zde je možné i nadále měnit počet licencí. Pokud je dostupný soubor s určitou instalační verzí, tak pro jeho stažení na disk stačí jen kliknout odkaz "Stáhnout" a pokračovat dle instrukcí internetového prohlížeče.

Software

Výběr kategorie softwaru: - Všechny položky -

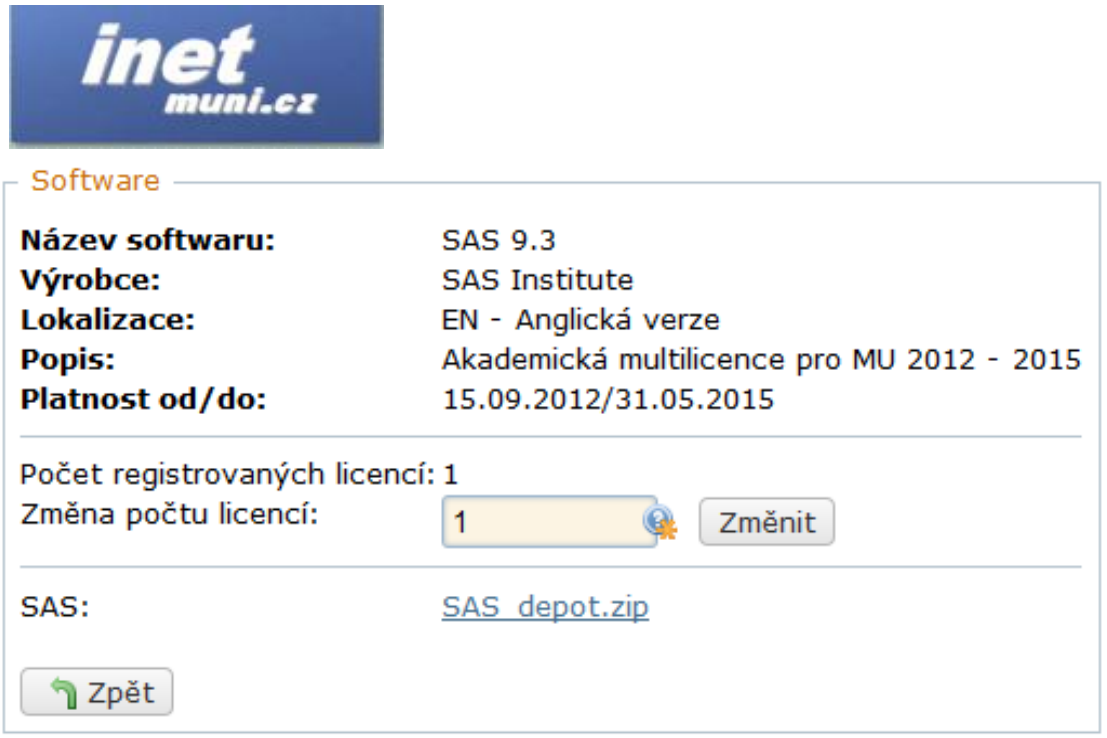
☑ Pouze aktuální software (platný)

☑ Pouze volné licence

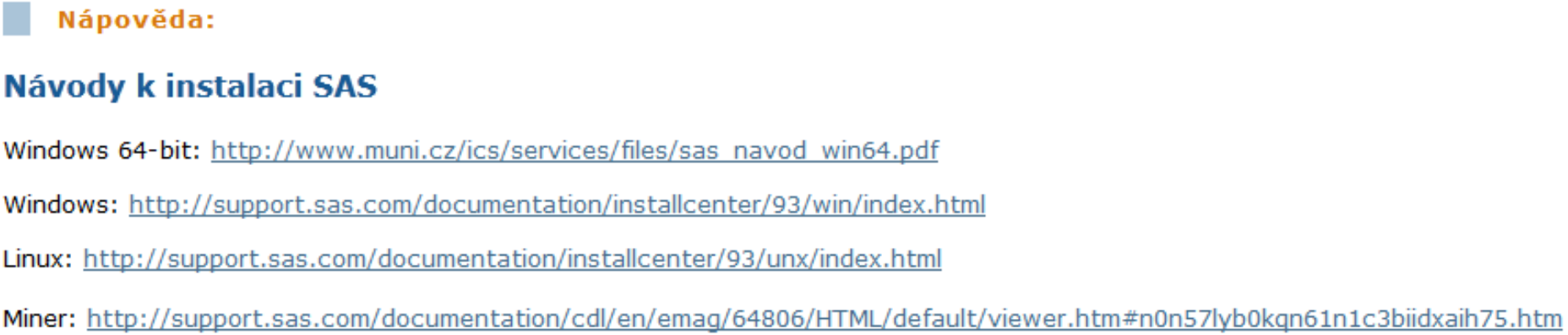
Název softwaru	Lokalizace	Popis	Platnost od	Platnost do	
ACREA CR, spol. s r.o.					
IBM SPSS Data Access Pack 6.1	EN - Anglická verze	Akademická multilicence pro MU 2012			Získat
IBM SPSS Data Access Pack 6.1 with sp3	EN - Anglická verze	Akademická multilicence pro MU 2012 - 2013			Získat
IBM SPSS Modeler 14.2	EN - Anglická verze	Akademická multilicence pro MU 2012 - 2013	05.01.2012	01.02.2014	Získat
IBM SPSS Modeler 15	EN - Anglická verze	Akademická multilicence pro MU 2012 - 2013	23.11.2012	01.02.2014	Získat
IBM SPSS Statistics 18	EN - Anglická verze	Akademická multilicence pro MU 2009 - 2013	09.12.2009	01.02.2014	Získat
IBM SPSS Statistics 19	EN - Anglická verze	Akademická multilicence pro MU 2011 - 2013	22.12.2010	01.02.2014	Získat
IBM SPSS Statistics 20	EN - Anglická verze	Akademická multilicence pro MU 2012 - 2013	05.01.2012	01.02.2014	Získat
IBM SPSS Statistics 20 Fix Pack 1 32b	EN - Anglická verze	Fix Pack 1 32b			Získat
IBM SPSS Statistics 20 Fix Pack 1 64b	EN - Anglická verze	Fix Pack 1 64b			Získat
IBM SPSS Statistics 21	EN - Anglická verze	Akademická multilicence pro MU 2012 - 2013	23.11.2012	01.02.2014	Získat
ALTAP, Ltd.					
Altap Salamander 2.5	NS - Nespecifikováno	Celouniverzitní licence	11.01.2008		Získat
MathWorks					
Matlab 7.13	EN - Anglická verze	Matlab 7.13 (2011b)			Získat
Matlab 8.0	EN - Anglická verze	Matlab 8.0 (2012b)			Získat
SAS Institute					
SAS 9.3	EN - Anglická verze	Akademická multilicence pro MU 2012 - 2015	15.09.2012	31.05.2015	Získat
SAS 9.3 SID files 2013	EN - Anglická verze	Licenční soubory pro SAS 9.3 pro MU 2012 - 2013	31.10.2012	31.12.2013	Získat
StatSoft					

SAS Institute					
SAS 9.3	EN - Anglická verze	Akademická multilicence pro MU 2012 - 2015	15.09.2012	31.05.2015	Získat
SAS 9.3 SID files 2013	EN - Anglická verze	Licenční soubory pro SAS 9.3 pro MU 2012 - 2013	31.10.2012	31.12.2013	Získat

Instalační soubory, licenční podmínky



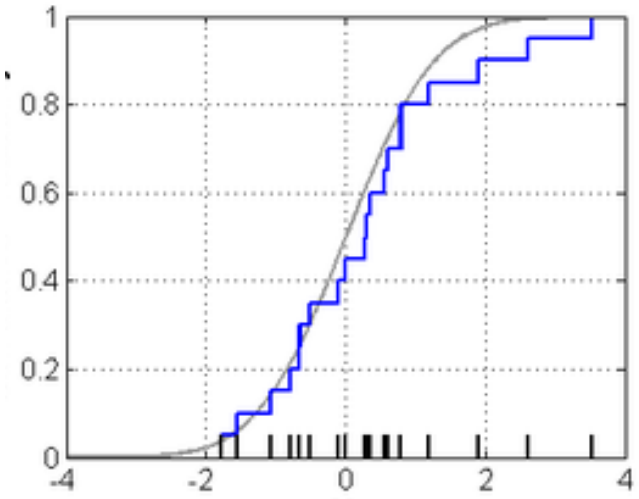
- K dispozici i návody pro instalaci

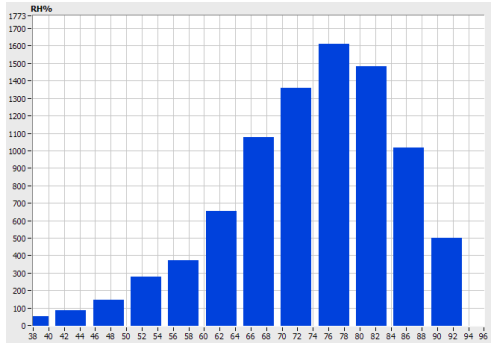


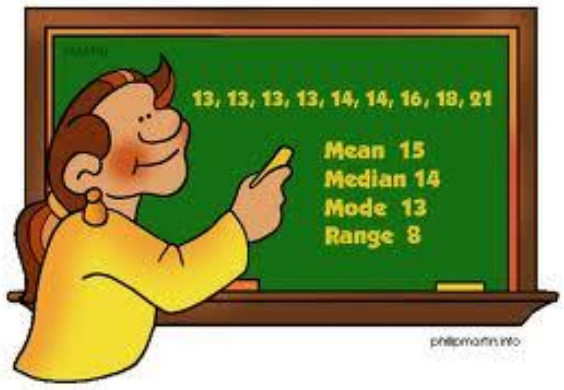
Instalační soubory, licenční podmínky

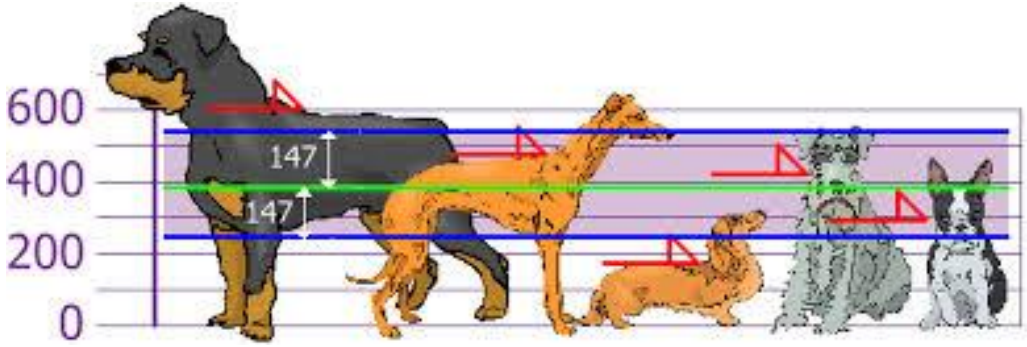
- Dále je třeba stáhnout SID files, ve kterých je uložena informace o platnosti licence a umožní fungování SASu. Instrukce, jak tyto soubory použít, je součástí stahovaného souboru.

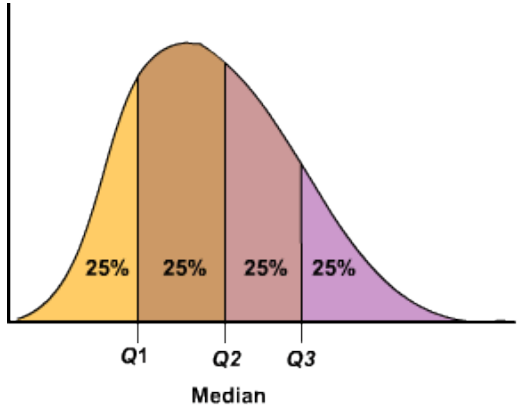
3. Funkcionální a číselné charakteristiky znaků.











Četnostní funkce, empirická distribuční funkce

Pomocí relativních četností zavedeme četnostní funkci .

Funkce $p(x) = \begin{cases} p_j \text{ pro } x = x_{[j]},\ j = 1, \ldots, r \\ 0 \text{ jinak} \end{cases}$ se nazývá četnostní funkce.

Četnostní funkce je
nezáporná ($\forall x \in R: p(x) \geq 0$)
a normovaná ($\sum_{x=-\infty}^{\infty} p(x) = 1$).

Pomocí kumulativních relativních četností zavedeme empirickou distribuční funkci .

Funkce $F(x) = \begin{cases} 0 \text{ pro } x < x_{[1]} \\ F_j \text{ pro } x_{[j]} \leq x < x_{[j+1]},\ j = 1, \ldots, r-1 \\ 1 \text{ pro } x \geq x_{[r]} \end{cases}$ se nazývá empirická distribuční funkce.

Empirická distribuční funkce je
neklesající ($\forall x_1, x_2 \in R,\ x_1 < x_2: F(x_1) \leq F(x_2)$),
zprava spojitá ($\forall x_0 \in R$ libovolné, ale pevně dané: $\lim_{x \to x_{0+}} F(x) = F(x_0)$)
a normovaná ($\lim_{x \to -\infty} F(x) = 0$, $\lim_{x \to \infty} F(x) = 1$).

Platí $\forall x \in R : F(x) = \sum_{t \leq x} p(t)$.

Příklad

Příklad: Pro známky z matematiky nakreslete graf četnostní funkce a empirické distribuční funkce.

Řešení:

Variační řada

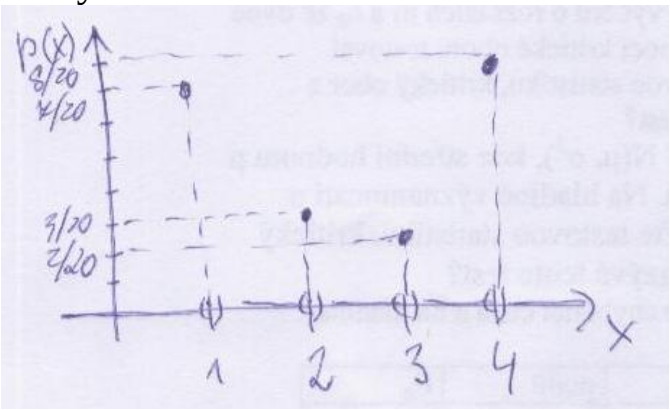
$x_{[j]}$	n_j	p_j	N_j	F_j
1	7	7/20=0,35	7	7/20=0,35
2	3	3/20=0,15	10	10/20=0,50
3	2	2/20=0,10	12	12/20=0,60
4	8	8/20=0,40	20	20/20=1,00
∑	20	1,00	-	-

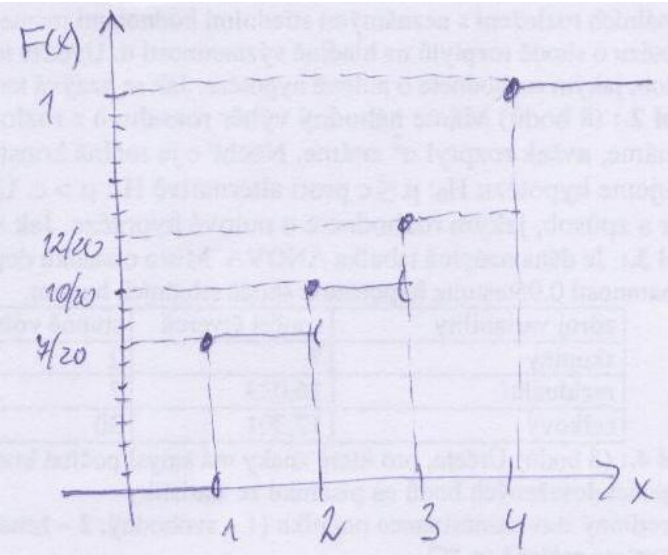
Vzorce

$$p(x)=\begin{cases} p_j \text{ pro } x = x_{[j]},\ j=1,\ \ldots, r \\ 0 \text{ jinak} \end{cases}$$

$$F(x)=\begin{cases} 0 \text{ pro } x < x_{[1]} \\ F_j \text{ pro } x_{[j]} \le x < x_{[j+1]},\ j=1,\ldots, r-1 \\ 1 \text{ pro } x \ge x_{[r]} \end{cases}$$

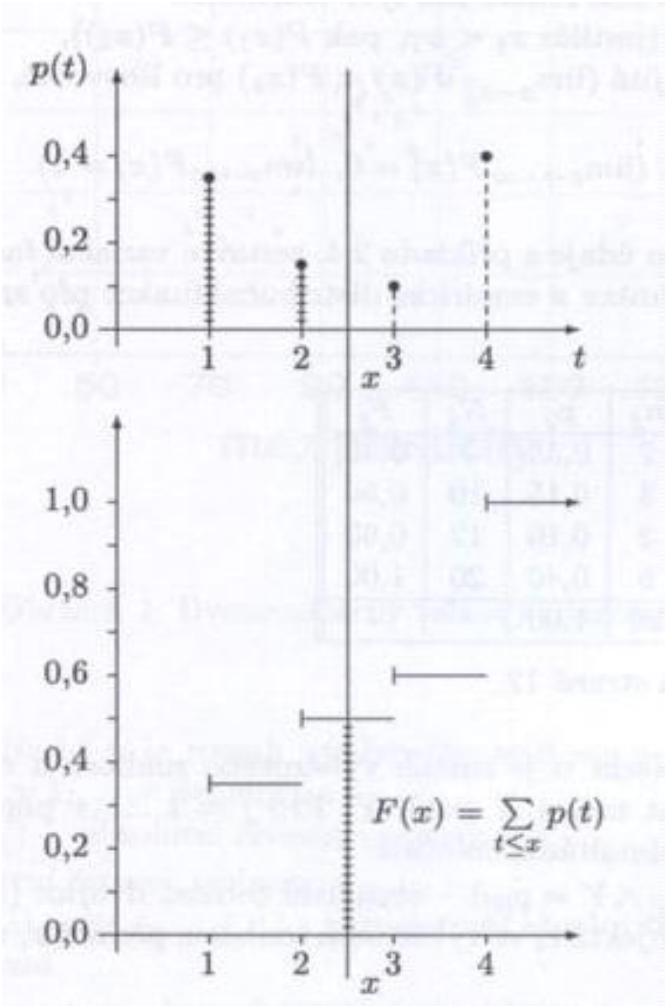
Grafy





Vztah mezi četnostní funkcí a empirickou distribuční funkcí

$$\forall x \in R : F(x) = \sum_{t \le x} p(t)$$



Simultánní a marginální četnostní funkce

Pomocí simultánních relativních četností zavedeme simultánní četnostní funkci :

Funkce

$$p(x, y) = \begin{cases} p_{jk} \text{ pro } x = x_{[j]}, y = y_{[k]}, j = 1, \ldots, r, k = 1, \ldots, s \\ 0 \text{ jinak} \end{cases}$$

se nazývá simultánní četnostní funkce.

Pomocí marginálních relativních četností zavedeme marginální četnostní funkce pro znaky X a Y . Odlišíme je indexem takto:

$$p_1(x) = \begin{cases} p_{j.} \text{ pro } x = x_{[j]}, j = 1, \ldots, r \\ 0 \text{ jinak} \end{cases},$$

$$p_2(y) = \begin{cases} p_{.k} \text{ pro } y = y_{[k]}, k = 1, \ldots, s \\ 0 \text{ jinak} \end{cases}.$$

Mezi simultánní četnostní funkcí a marginálními četnostními funkcemi platí vztahy:

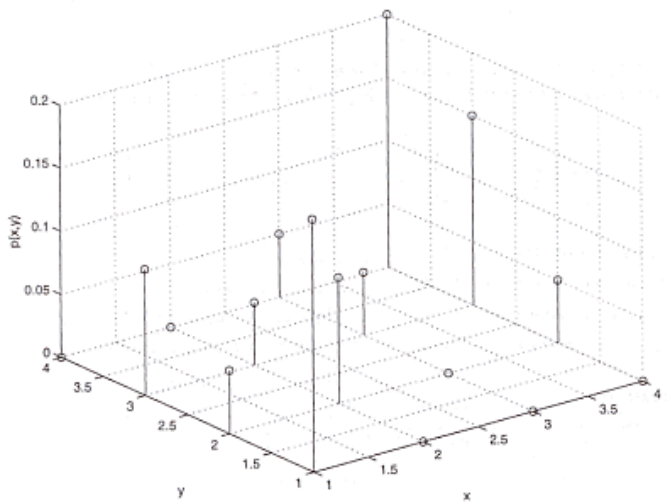
$$p_1(x) = \sum_{y=-\infty}^{\infty} p(x, y), \quad p_2(y) = \sum_{x=-\infty}^{\infty} p(x, y).$$

Příklad

Příklad: Sestrojte graf simultánní četnostní funkce pro známky z matematiky a angličtiny.

Řešení: Vyjdeme z kontingenční tabulky simultánních relativních četností.

x \ y	1	2	3	4	$p_{j\cdot}$
1	0,20	0,05	0,10	0,00	0,35
2	0,00	0,10	0,05	0,00	0,15
3	0,00	0,00	0,05	0,05	0,10
4	0,00	0,05	0,15	0,20	0,40
$p_{\cdot k}$	0,20	0,20	0,35	0,25	1,00



Četností nezávislost znaků v daném výběrovém souboru

Řekneme, že znaky X, Y jsou v daném výběrovém souboru četnostně nezávislé , právě když pro všechna j = 1, ..., r a všechna k = 1, ..., s platí multiplikativní vztah:
$p_{jk} = p_{j.}\, p_{.k}$ neboli pro $\forall (x, y) \in R^2$: $p(x, y) = p_1(x)\, p_2(y)$.

Příklad: Ověřte, zda v našem datovém souboru jsou známky z matematiky a angličtiny četnostně nezávislé.

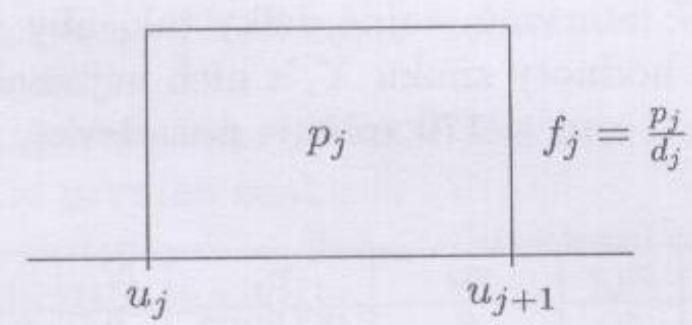
Řešení: Vyjdeme z kontingenční tabulky relativních četností.

x \ y	p_{jk}	1	2	3	4	$p_{j\cdot}$
1		0,20	0,05	0,10	0,00	0,35
2		0,00	0,10	0,05	0,00	0,15
3		0,00	0,00	0,05	0,05	0,10
4		0,00	0,05	0,15	0,20	0,40
$p_{\cdot k}$		0,20	0,20	0,35	0,25	1,00

Známky z matematiky a angličtiny nejsou četnostně nezávislé, protože už pro j = 1, k = 1 je multiplikativní vztah porušen:
$p_{11} = 0{,}20$, $p_{1.} = 0{,}35$, $p_{.1} = 0{,}20$, tudíž $0{,}20 \neq 0{,}35.0{,}20$

Histogram, hustota četnosti, intervalová empirická distribuční funkce

Intervalové rozložení četností graficky znázorňujeme pomocí histogramu. Je to graf skládající se z r obdélníků, sestrojených nad třídicími intervaly, přičemž obsah j-tého obdélníku je roven relativní četnosti p_j j-tého třídicího intervalu, j = 1, ..., r.



Histogram je shora omezen schodovitou čarou, která je grafem funkce zvané hustota četnosti:

$$f(x) = \begin{cases} f_j \text{ pro } u_j < x \le u_{j+1},\ j = 1, \cdots, r \\ 0 \text{ jinak} \end{cases}$$

Pomocí hustoty četnosti zavedeme intervalovou empirickou distribuční funkci:

$$F(x) = \int_{-\infty}^{x} f(t)dt .$$

Hustota četnosti je nezáporná ($\forall x \in R : f(x) \ge 0$) a normovaná ($\int_{-\infty}^{\infty} f(x)dx = 1$). Intervalová empirická distribuční funkce je neklesající, spojitá a normovaná ($\lim_{x \to -\infty} F(x) = 0$, $\lim_{x \to \infty} F(x) = 1$).

Příklad

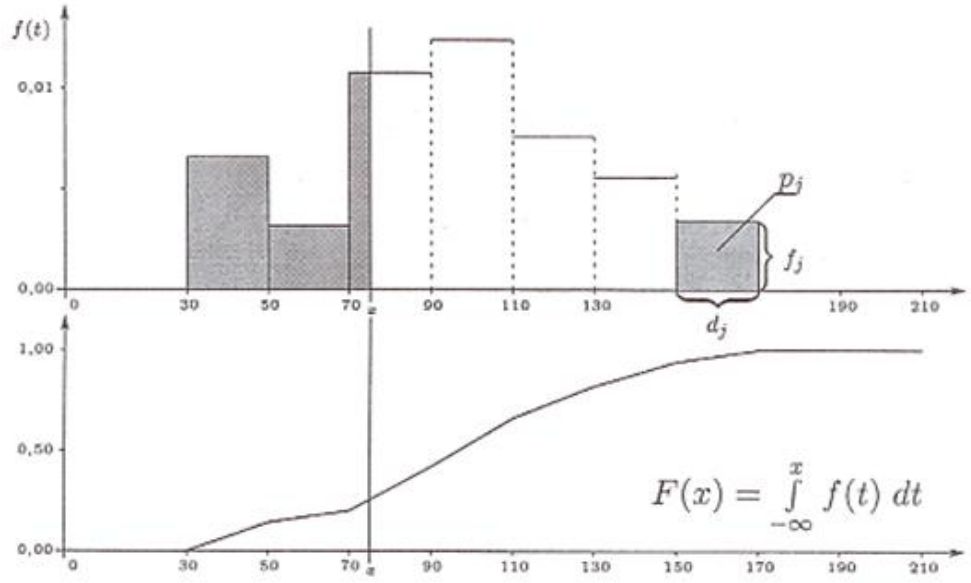
Příklad: Pro mez plasticity oceli nakreslete histogram a pod histogram graf intervalové empirické distribuční funkce.

Řešení: Vyjdeme z tabulky rozložení četností.

$(u_j, u_{j+1}\rangle$	d_j	$x_{[j]}$	n_j	p_j	N_j	F_j	f_j
$(30,50\rangle$	20	40	8	$8/60 = 0{,}1\overline{3}$	8	$8/60 = 0{,}1\overline{3}$	$8/(60\cdot 20) = 0{,}00\overline{6}$
$(50,70\rangle$	20	60	4	$4/60 = 0{,}0\overline{6}$	12	$12/60 = 0{,}2$	$4/(60\cdot 20) = 0{,}00\overline{3}$
$(70,90\rangle$	20	80	13	$13/60 = 0{,}21\overline{6}$	25	$25/60 = 0{,}41\overline{6}$	$13/(60\cdot 20) = 0{,}01083$
$(90,110\rangle$	20	100	15	$15/60 = 0{,}25$	40	$40/60 = 0{,}\overline{6}$	$15/(60\cdot 20) = 0{,}0125$
$(110,130\rangle$	20	120	9	$9/60 = 0{,}15$	49	$49/60 = 0{,}81\overline{6}$	$9/(60\cdot 20) = 0{,}0075$
$(130,150\rangle$	20	140	7	$7/60 = 0{,}11\overline{6}$	56	$56/60 = 0{,}9\overline{3}$	$7/(60\cdot 20) = 0{,}00583$
$(150,170\rangle$	20	160	4	$4/60 = 0{,}0\overline{6}$	60	$60/60 = 1$	$4/(60\cdot 20) = 0{,}00\overline{3}$
Součty			60	1			

Příklad

$\langle u_j, u_{j+1}\rangle$	d_j	$x_{[j]}$	n_j	p_j	N_j	F_j	f_j
$\langle 30,50\rangle$	20	40	8	$8/60 = 0{,}1\overline{3}$	8	$8/60 = 0{,}1\overline{3}$	$8/(60\cdot 20) = 0{,}00\overline{6}$
$\langle 50,70\rangle$	20	60	4	$4/60 = 0{,}0\overline{6}$	12	$12/60 = 0{,}2$	$4/(60\cdot 20) = 0{,}00\overline{3}$
$\langle 70,90\rangle$	20	80	13	$13/60 = 0{,}21\overline{6}$	25	$25/60 = 0{,}41\overline{6}$	$13/(60\cdot 20) = 0{,}01083$
$\langle 90,110\rangle$	20	100	15	$15/60 = 0{,}25$	40	$40/60 = 0{,}\overline{6}$	$15/(60\cdot 20) = 0{,}0125$
$\langle 110,130\rangle$	20	120	9	$9/60 = 0{,}15$	49	$49/60 = 0{,}81\overline{6}$	$9/(60\cdot 20) = 0{,}0075$
$\langle 130,150\rangle$	20	140	7	$7/60 = 0{,}11\overline{6}$	56	$56/60 = 0{,}9\overline{3}$	$7/(60\cdot 20) = 0{,}00583$
$\langle 150,170\rangle$	20	160	4	$4/60 = 0{,}0\overline{6}$	60	$60/60 = 1$	$4/(60\cdot 20) = 0{,}00\overline{3}$
Součty			60	1			



Simultánní a marginální hustota četnosti

Pomocí simultánních četnostních hustot zavedeme simultánní hustotu četnosti:

Funkce $f(x,y)=\begin{cases} f_{jk} \text{ pro } u_j < x \le u_{j+1},\ v_k < y \le v_{k+1},\ j=1,\cdots,r,\ k=1,\cdots,s \\ 0 \text{ jinak} \end{cases}$ se nazývá simultánní hustota četnosti. Jejím grafem je schodovitá plocha shora omezující stereogram.

Hustoty četnosti pro znaky X a Y odlišíme indexem takto:

$$f_1(x)=\begin{cases} f_{j.} \text{ pro } u_j < x \le u_{j+1},\ j=1,\cdots,r \\ 0 \text{ jinak} \end{cases},$$

$$f_2(y)=\begin{cases} f_{.k} \text{ pro } v_k < y \le v_{k+1},\ k=1,\cdots,s \\ 0 \text{ jinak} \end{cases}.$$

Mezi simultánní hustotou četnosti a marginálními hustotami četnosti platí vztahy:

$$f_1(x)=\int_{-\infty}^{\infty} f(x,y)\,dy,\quad f_2(y)=\int_{-\infty}^{\infty} f(x,y)\,dx.$$

Četnostní nezávislost znaků v daném výběrovém souboru při intervalovém rozložení četností

Pomocí simultánních a marginálních četnostní zavedeme pojem četnostní nezávislosti znaků v daném výběrovém souboru při intervalovém rozložení četností:
Řekneme, že znaky X, Y jsou v daném výběrovém souboru četnostně nezávislé při intervalovém rozložení četností, jestliže pro všechna j = 1, ..., r a všechna k = 1, ...,s platí multiplikativní vztah: $f_{jk} = f_{j.}\, f_{.k}$ neboli pro $\forall (x, y) \in R^2$: $f(x, y) = f_1(x)\, f_2(y)$.

V našem příkladě nejsou mez pevnosti a mez plasticity četnostně nezávislé, protože už pro j = 1, k = 1 je multiplikativní vztah porušen:

$\langle u_j, u_{j+1}\rangle$ \ $\langle v_k, v_{k+1}\rangle$	n_{jk}	$\langle 50, 70\rangle$	$\langle 70, 90\rangle$	$\langle 90, 110\rangle$	$\langle 110, 130\rangle$	$\langle 130, 150\rangle$	$\langle 150, 170\rangle$	$\langle 170, 190\rangle$	$n_{j.}$
$\langle 30, 50\rangle$		5	3	0	0	0	0	0	8
$\langle 50, 70\rangle$		0	3	1	0	0	0	0	4
$\langle 70, 90\rangle$		0	4	7	1	1	0	0	13
$\langle 90, 110\rangle$		0	0	6	8	1	0	0	15
$\langle 110, 130\rangle$		0	0	0	4	5	0	0	9
$\langle 130, 150\rangle$		0	0	0	0	2	5	0	7
$\langle 150, 170\rangle$		0	0	0	0	0	1	3	4
$n_{.k}$		5	10	14	13	9	6	3	$n = 60$

$f_{11} = \frac{5}{60 \cdot 20 \cdot 20} = 0{,}000208$, $f_{1.} = \frac{8}{60 \cdot 20} = 0{,}006667$, $f_{.1} = \frac{5}{60 \cdot 20} = 0{,}004167$, tudíž

$0{,}000208 \neq 0{,}006667.0{,}004167 = 0{,}000028$

Číselné charakteristiky znaků

Doposud jsme se zabývali funkcionálními charakteristikami znaků, jako jsou:
- empirická distribuční funkce $F(x)$,
- simultánní četnostní funkce $p(x,y)$,
- marginální četnostní funkce $p_1(x)$, $p_2(y)$,
- simultánní hustota četnosti $f(x,y)$,
- marginální hustoty četnosti $f_1(x)$, $f_2(y)$,

které nesou úplnou informaci o rozložení četností.
Nyní zavedeme číselné charakteristiky, které nás informují o některých rysech tohoto rozložení četností:
- o poloze (úrovni) hodnot znaku,
- o jejich variabilitě (rozptýlení),
- o těsnosti závislosti dvou znaků
- a pod.

Pro různé typy znaků se používají různé číselné charakteristiky, proto se nejdřív seznámíme s jednotlivými typy znaků.

Typy znaků

Nominální znak: připouští obsahovou interpretaci pouze u relace rovnosti =. O dvou variantách nominálního znaku lze pouze konstatovat, že jsou buď stejné nebo různé. Čísla, která přiřadíme jednotlivým variantám znaku, nereprezentují skutečnou hodnotu použitých čísel, ale jsou pouhým označením variant znaku.
Příklady nominálních znaků: lékařská diagnóza, typ profese, barva očí, rodinný stav, národnost, …

Ordinální znak: připouští obsahovou interpretaci nejen u relace rovnosti =, ale též u relace uspořádání <. Můžeme tedy konstatovat, že varianta $x_{[j]}$ je větší (dokonalejší, silnější, vhodnější) než varianta $x_{[k]}$.
Příklad ordinálního znaku: školní klasifikace vyjadřuje menší nebo větší znalosti zkoušených žáků – jedničkář je lepší než dvojkař, ale intervaly mezi známkami nemají obsahovou interpretaci. Nelze tvrdit, že rozdíl ve znalostech mezi jedničkářem a dvojkařem je stejný jako mezi trojkařem a čtyřkařem.
Další příklady: Různá bodování ve sportovních a uměleckých soutěžích, posuzování různých rysů sociálního chování, posuzování stavu pacientů, hodnocení postojů respondentů k různým otázkám, …

Typy znaků

Intervalový znak: kromě relací rovnosti = a uspořádání < umožňuje obsahovou interpretaci také u operace rozdílu -, tj. stejný interval mezi jednou dvojicí hodnot a jinou dvojicí hodnot vyjadřuje i stejný rozdíl v extenzitě zkoumané vlastnosti.
Příklad intervalového znaku: teplota měřená ve stupních Celsia. Např. naměříme-li ve čtyřech po sobě jdoucích dnech polední teploty 0, 2, 4, 6 ºC, znamená to, že každým dnem stouply teploty o 2 ºC. Nelze však říci, že z druhého na třetí den vzrostla teplota dvojnásobně, kdežto ze třetího na čtvrtý den pouze jeden a půl krát.
Další příklady: kalendářní systémy, směr větru, inteligenční kvocient, …
Společný znak intervalových znaků: nula byla stanovena uměle, pouhou konvencí.

Poměrový znak: kromě relací rovnosti = a uspořádání < umožňuje obsahovou interpretaci také u operací rozdílu - a podílu /, tj. stejný poměr mezi jednou dvojicí hodnot a jinou dvojicí hodnot vyjadřuje i stejný podíl v extenzitě zkoumané vlastnosti.
Příklad poměrového znaku: délka předmětu měřená v cm. Má-li jeden předmět délku 8 cm a druhý 16 cm, má smysl prohlásit, že druhý předmět je dvakrát delší než první předmět.
Další příklady: počet dětí v rodině, výška kapesného v Kč, hmotnost osoby, …
Společný znak poměrových znaků: Poměrový znak má přirozený počátek, ke kterému jsou vztahovány všechny další hodnoty znaku.

Mimo uvedenou klasifikaci stojí alternativní znaky, které nabývají jen dvou hodnot, např. 0,1, což znamená absenci a prezenci nějakého jevu. Například 0 bude znamenat neúspěch, 1 úspěch při řešení určité úlohy. Alternativní znaky mohou být ztotožněny s kterýmkoliv z předcházejících typů.

Číselné charakteristiky nominálních znaků

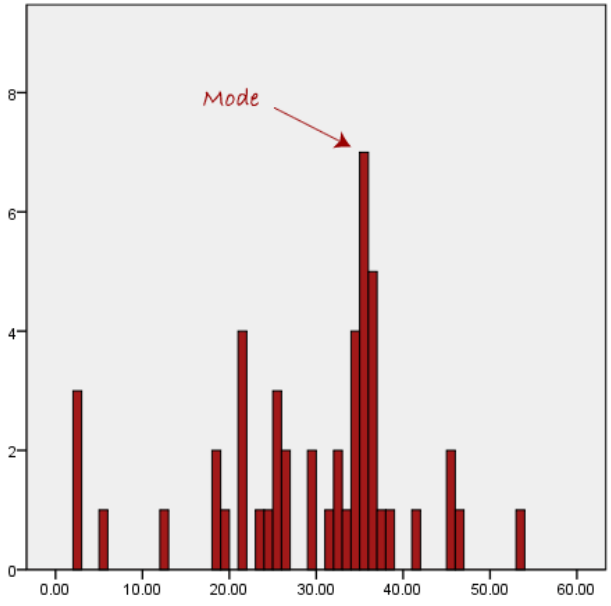
Charakteristika polohy: modus – nejčetnější varianta resp. střed nejčetnějšího třídicího intervalu.

Příklad na stanovení modu
20 náhodně vybraných osob mělo odpovědět na otázku, který z pěti výrobků (označíme je A, B, C, D, E) preferují. Výsledky máme v tabulce:

Výrobek	A	B	C	D	E
Četnost odpovědí	3	5	3	6	3

Stanovte modus.

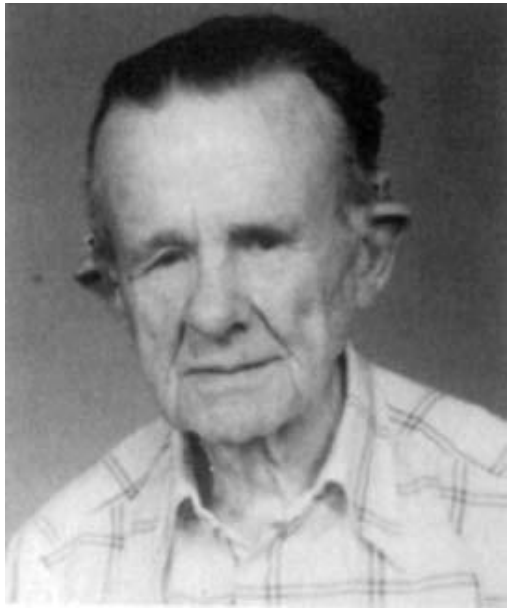
Řešení:
Modus = D



Označení: $\hat{x}$

Cramérův koeficient

Charakteristika těsnosti závislosti dvou nominálních znaků: Cramérův koeficient kontingence.



Carl Harald Cramér (1893 – 1985): Švédský matematik

Cramérův koeficient

Nechť znak X nabývá variant $x_{[1]}, ..., x_{[r]}$ a znak Y nabývá variant $y_{[1]}, ..., y_{[s]}$. Máme dvourozměrný datový soubor $\begin{pmatrix} x_1 & y_1 \\ \dots & \dots \\ x_n & y_n \end{pmatrix}$. Zjistíme absolutní četnosti n_{jk} dvojice variant $(x_{[j]},y_{[k]})$, j = 1, …, r, k = 1, …, s a uspořádáme je do kontingenční tabulky:

	y	$y_{[1]}$	...	$y_{[s]}$	$n_{j.}$
x	n_{jk}				
$x_{[1]}$		n_{11}	...	n_{1s}	$n_{1.}$
⋮		...	...	...	...
$x_{[r]}$		n_{r1}	...	n_{rs}	$n_{r.}$
$n_{.k}$		$n_{.1}$	...	$n_{.s}$	n

Vypočteme tzv. teoretické četnosti $\frac{n_{j.}n_{.k}}{n}$ a s jejich pomocí pak statistiku

$K = \sum_{j=1}^{r}\sum_{k=1}^{s} \frac{\left(n_{jk} - \frac{n_{j.}n_{.k}}{n}\right)^2}{\frac{n_{j.}n_{.k}}{n}}$. Cramérův koeficient: $V = \sqrt{\frac{K}{n(m-1)}}$, kde m = min{r,s}. Tento koeficient nabývá hodnot mezi 0 a 1. Čím blíže je 1, tím je těsnější závislost mezi X a Y, čím blíže je 0, tím je tato závislost volnější.

Cramérův koeficient

Význam hodnot Cramérova koeficientu:

mezi 0 až 0,1 … zanedbatelná závislost,
mezi 0,1 až 0,3 … slabá závislost,
mezi 0,3 až 0,7 … střední závislost,
mezi 0,7 až 1 … silná závislost.

Příklad

Příklad na výpočet Cramérova koeficientu:
686 náhodně vybraných osob bylo dotázáno, zda vlastní auto (znak X, varianty 1 – ano, 2 – ne) a zda jsou ochotny používat MHD (znak Y, varianty 1 – ano, 2 – ne). Výsledky průzkumu jsou uvedeny v kontingenční tabulce

X	Y		$n_{j.}$
	ano	ne	
ano	56	312	368
ne	283	35	318
$n_{.k}$	339	347	686

Vypočtěte a interpretujte Cramérův koeficient.

Řešení: Nejprve vypočteme teoretické četnosti:

$$\frac{n_{1.}n_{.1}}{n} = \frac{368 \cdot 339}{686} = 181{,}8542, \quad \frac{n_{1.}n_{.2}}{n} = \frac{368 \cdot 347}{686} = 186{,}1458,$$

$$\frac{n_{2.}n_{.1}}{n} = \frac{318 \cdot 339}{686} = 157{,}1458, \quad \frac{n_{2.}n_{.2}}{n} = \frac{318 \cdot 347}{686} = 160{,}8542$$

Nyní dosadíme do vzorce pro výpočet statistiky K:

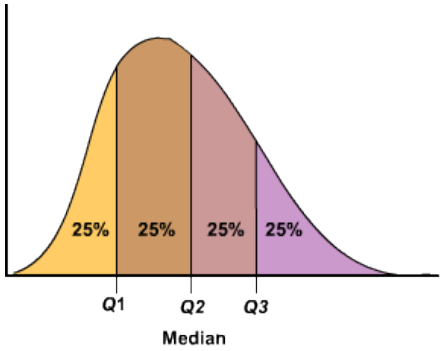
$$K = \frac{(56 - 181{,}8542)^2}{181{,}8542} + \frac{(312 - 186{,}1458)^2}{186{,}1485} + \frac{(283 - 157{,}1458)^2}{157{,}1458} + \frac{(35 - 160{,}8542)^2}{160{,}8542} = 371{,}456$$

Nakonec vypočteme Cramérův koeficient:

$$V = \sqrt{\frac{371{,}456}{686 \cdot 1}} = 0{,}7358$$

Hodnota Cramérova koeficientu svědčí o tom, že mezi znaky X a Y existuje silná závislost.

Číselné charakteristiky ordinálních znaků

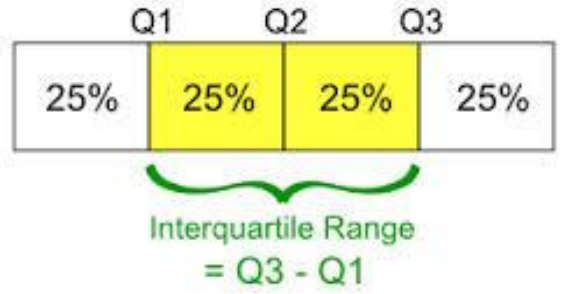


Charakteristika polohy: α-kvantil. Je-li $\alpha \in (0;1)$, pak α-kvantil x_α je číslo, které rozděluje uspořádaný datový soubor na dolní úsek, obsahující aspoň podíl α všech dat a na horní úsek obsahující aspoň podíl 1 – α všech dat. Pro výpočet α-kvantilu slouží algoritmus:

$$n\alpha = \begin{cases} \text{celé číslo } c \Rightarrow x_\alpha = \dfrac{x_{(c)} + x_{(c+1)}}{2} \\ \text{necelé číslo} \Rightarrow \text{zaokrouhlíme nahoru na nejbližší celé číslo } c \Rightarrow x_\alpha = x_{(c)} \end{cases}$$

Pro speciálně zvolená α užíváme názvů: $x_{0,50}$ – medián, $x_{0,25}$ – dolní kvartil, $x_{0,75}$ – horní kvartil, $x_{0,1}$, ..., $x_{0,9}$ – decily, $x_{0,01}$, ..., $x_{0,99}$ – percentily.

Charakteristika variability: kvartilová odchylka: $q = x_{0,75} - x_{0,25}$.



Příklad

Příklad na výpočet kvantilů:
U 50 žáků 7. ročníku jedné základní školy byly na pololetním vysvědčení zjištěny známky z matematiky:

známka	1	2	3	4	5
četnost známky	9	15	20	4	2

Určete medián, 1. a 9. decil a kvartilovou odchylku.

Řešení:
Pro snadnější výpočet tabulku doplníme ještě o absolutní kumulativní četnosti:

známka	1	2	3	4	5
n_j	9	15	20	4	2
N_j	9	24	44	48	50

Rozsah souboru n = 50

α	nα	c	x_α
0,50	50.0,5=25	25	$\frac{x_{(25)}+x_{(26)}}{2}=\frac{3+3}{2}=3$
0,10	50.0,1 = 5	5	$\frac{x_{(5)}+x_{(6)}}{2}=\frac{1+1}{2}=1$
0,90	50.0,9 = 45	45	$\frac{x_{(45)}+x_{(46)}}{2}=\frac{4+4}{2}=4$
0,25	50.0,25 = 12,5	13	$x_{(13)} = 2$
0,75	50.0,75 = 37,5	38	$x_{(38)} = 3$

Kvartilová odchylka: q = 3 – 2 = 1.

Interpretace např. Dolního kvartilu: V souboru žáků je aspoň čtvrtina takových, kteří mají z matematiky jedničku nebo dvojku (neboli v souboru 50 žáků jsou aspoň tři čtvrtiny takových, kteří mají z matematiky dvojku či horší známku).

Příklad

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	4	4	5	6	8	8	12	12	13	14	14	14	18	19

$x_{0,25}$ $x_{0,5}$ $x_{0,75}$

	n_j	N_j	p_j	F_j
1	1	1	0,07	0,07
4	2	3	0,13	0,20
5	1	4	0,07	0,27
6	1	5	0,07	0,33
8	2	7	0,13	0,47
12	2	9	0,13	0,60
13	1	10	0,07	0,67
14	3	13	0,20	0,87
18	1	14	0,07	0,93
19	1	15	0,07	1,00
Součet	15	x	1,00	x

$x_{0,25} = 5$

$x_{0,5} = \tilde{x} = 12$

$x_{0,75} = 14$

$\hat{x} = 14$

$q = x_{0,75} - x_{0,25} = 14 - 5 = 9$

$X_{0,25}$ je tedy hodnota, u které F_j poprvé překročí 0,25.

!!! Pokud ale $F_j = \alpha$ pro nějaké $x_{[j]}$, $x_\alpha = (x_{[j]} + x_{[j+1]})/2$

$x_{0,2} = (4+5)/2 = 4,5$

Modus a kvantily pro intervalově tříděná data

$$\hat{x} = d_m + \frac{n_m - n_{m-1}}{2n_m - n_{m-1} - n_{m+1}} \cdot h$$

d_m je dolní mez modální třídy,
n_m, n_{m-1}, n_{m+1} je četnost modální, předcházející a následující třídy,
h je šířka třídy

$$x_P = d_P + \frac{P - F_{P-1}}{p_P} \cdot h$$

d_P je dolní mez třídy obsahující příslušný P-kvantil,
p_P je relativní četnost této třídy,
F_{P-1} je kumulativní relativní četnost předcházející třídy,
h je šířka třídy

Příklad

Určete modus a medián.

x_i	N_i
méně než 15>	22
(15;20>	34
(20;25>	72
(25;30>	102
(30;35>	127
více než 35	135

Příklad

x_i	N_i
méně než 15>	22
(15;20>	34
(20;25>	72
(25;30>	102
(30;35>	127
více než 35	135

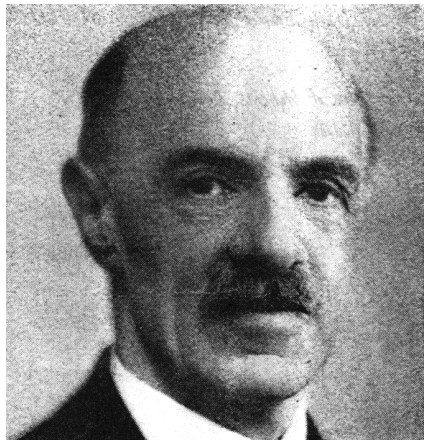
x_i	n_i	p_i	N_i	F_i
12,5	22	0,16	22	0,16
17,5	12	0,09	34	0,25
22,5	38	0,28	72	0,53
27,5	30	0,22	102	0,76
32,5	25	0,19	127	0,94
37,5	8	0,06	135	1,00
Součet	135	1,00	x	x

$$\hat{x} = 20 + \frac{38-12}{2\cdot 38 - 12 - 30}\cdot 5 = 23{,}82$$

$$\tilde{x} = 20 + \frac{0{,}5-0{,}25}{0{,}28}\cdot 5 = 24{,}46$$

Spearmanův koeficient

Charakteristika těsnosti závislosti dvou ordinálních znaků: Spearmanův koeficient pořadové korelace



Charles Edward Spearman (1863 – 1945): Britský psycholog a statistik

Nejprve je nutné vysvětlit pojem pořadí čísla v posloupnosti čísel.
Nechť $x_1, \ldots, x_n$ je posloupnost reálných čísel.
a) Jsou-li čísla navzájem různá, pak pořadím R_i čísla x_i rozumíme počet těch čísel $x_1, \ldots, x_n$, která jsou menší nebo rovna číslu x_i.
b) Vyskytují-li se mezi danými čísly skupinky stejných čísel, pak každé takové skupince přiřadíme průměrné pořadí.

Příklad

Příklad na stanovení pořadí
a) Jsou dána čísla 9, 4, 5, 7, 3, 1.
b) Jsou dána čísla 6, 7, 7, 9, 6, 10, 8, 6, 6, 9.
Stanovte pořadí těchto čísel.

Řešení
ad a)

usp. čísla	1	3	4	5	7	9
pořadí	1	2	3	4	5	6

ad b)

usp. čísla	6	6	6	6	7	7	8	9	9	10
pořadí	1	2	3	4	5	6	7	8	9	10
prům. pořadí	2,5	2,5	2,5	2,5	5,5	5,5	7	8,5	8,5	10

Spearmanův koeficient

Vzorec pro výpočet Spearmanova koeficientu:

Předpokládejme, že máme dvourozměrný datový soubor $\begin{pmatrix} x_1 & y_1 \\ \cdots & \cdots \\ x_n & y_n \end{pmatrix}$. Označíme R_i pořadí hodnoty x_i a Q_i pořadí hodnoty y_i, $i = 1, \ldots, n$.

Spearmanův koeficient pořadové korelace: $r_S = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (R_i - Q_i)^2$.

Vlastnosti Spearmanova koeficientu pořadové korelace:

Koeficient nabývá hodnot mezi -1 a 1. Čím je bližší 1, tím je silnější přímá pořadová závislost mezi znaky X a Y, čím je bližší -1, tím je silnější nepřímá pořadová závislost mezi znaky X a Y.

Je-li $r_S = 1$ resp. $r_S = -1$, pak dvojice (x_i, y_i) leží na nějaké vzestupné resp. klesající funkci.

Hodnoty r_S se nezmění, když provedeme vzestupnou transformaci původních dat.

Hodnoty r_S se vynásobí -1, když provedeme sestupnou transformaci původních dat.

Koeficient je symetrický.

Koeficient je rezistentní vůči odlehlým hodnotám.

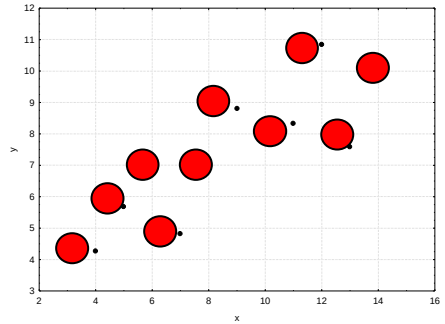
Spearmanův koeficient

Význam absolutní hodnoty Spearmanova koeficientu:

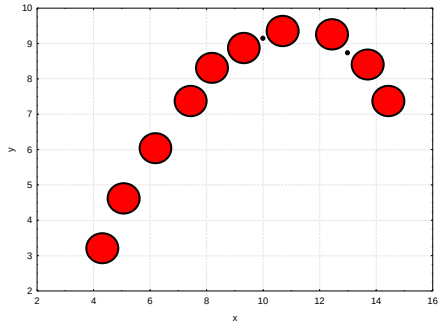
mezi 0 až 0,1	… zanedbatelná pořadová závislost,
mezi 0,1 a ž 0,3	… slabá pořadová závislost,
mezi 0,3 až 0,7	… střední pořadová závislost,
mezi 0,7 až 1	… silná pořadová závislost.

Ilustrace významu Spearmanova koeficientu pořadové korelace

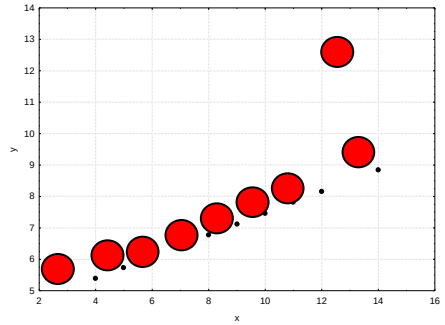
$r_S = 0{,}82$



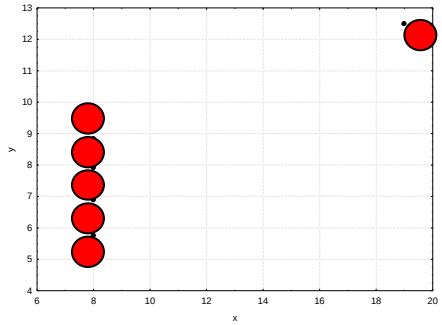
$r_S = 0{,}69$



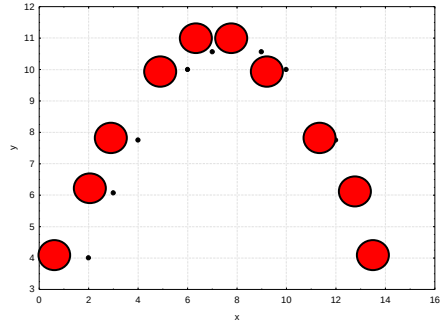
$r_S = 0{,}99$



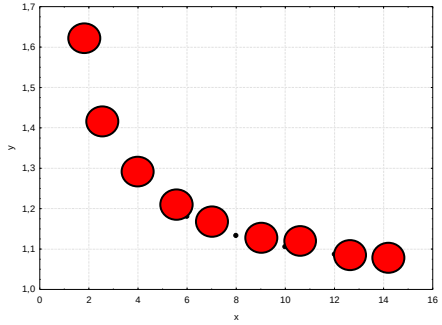
$r_S = 0{,}5$



$r_S = 0$



$r_S = -1$



Příklad

Příklad na výpočet Spearmanova koeficientu pořadové korelace:
Je dán dvourozměrný datový soubor

$$\begin{pmatrix} 2,5 & 13,4 \\ 3,4 & 15,2 \\ 1,3 & 11,8 \\ 5,8 & 13,1 \\ 3,6 & 14,5 \end{pmatrix}$$

Vypočtěte Spearmanův koeficient pořadové korelace.

Řešení:

x_i	2,5	3,4	1,3	5,8	3,6
y_i	13,4	15,2	11,8	13,1	14,5
R_i	2	3	1	5	4
Q_i	3	5	1	2	4
$(R_i-Q_i)^2$	1	4	0	9	0

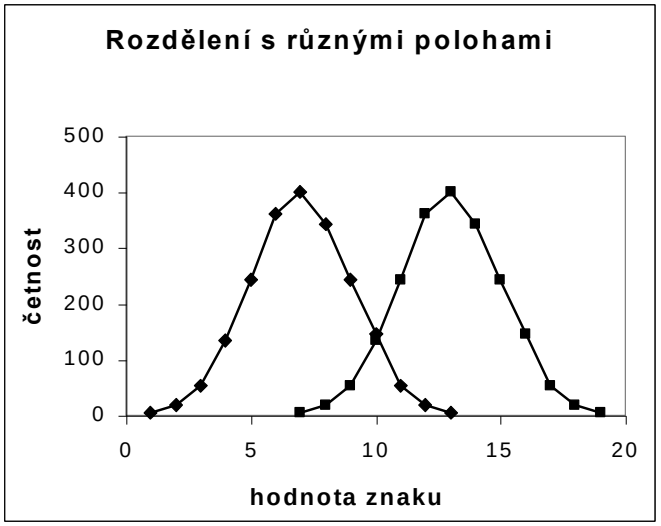
$$r_S = 1 - \frac{6}{n\left(n^2-1\right)} \sum_{i=1}^{n} \left(R_i - Q_i\right)^2 = 1 - \frac{6}{5 \cdot 24}\left(1+4+0+9+0\right) = 1 - \frac{6 \cdot 14}{5 \cdot 24} = 0,3$$

Znamená to, že mezi znaky X a Y existuje slabá přímá pořadová závislost.

Číselné charakteristiky intervalových znaků

Charakteristika polohy: aritmetický průměr je součet hodnot dělený jejich počtem: $m = \frac{1}{n}\sum_{i=1}^{n} x_i$. Pomocí průměru zavedeme i-tou centrovanou hodnotu $x_i - m$ (podle znaménka poznáme, zda i-tá hodnota je podprůměrná či nadprůměrná).

Znázornění rozložení četností dvou datových souborů, které se liší aritmetickým průměrem



Často se aritmetický průměr označuje : $\bar{x}$

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

Příklad

Příklad: Pro datový soubor obsahující údaje o mezi plasticity (znak X) a mezi pevnosti oceli (znak Y) vypočtěte aritmetické průměry znaků X, Y.

X	Y	X	Y	X	Y
154	178	83	98	73	76
133	164	106	111	77	86
58	75	92	104	47	61
145	161	85	103	68	85
94	107	112	118	137	142
113	141	98	102	44	68
86	97	103	108	92	116
121	127	99	119	141	157
119	138	104	128	155	189
112	125	107	118	136	155
85	97	98	140	82	81
41	72	97	115	136	163
96	113	105	101	72	79
45	89	71	93	66	81
99	109	39	69	42	61
51	95	122	147	113	123
101	114	33	52	42	85
160	169	78	117	133	147
87	101	114	137	153	179
88	139	125	149	85	91

Řešení:

$$m_1 = \frac{154+133+\ldots+85}{60} = 95{,}9, \quad m_2 = \frac{178+164+\ldots+91}{60} = 114{,}4$$

Aritmetický průměr

Vlastnosti aritmetického průměru

- Aritmetický průměr si lze představit jako těžiště dat – součet podprůměrných hodnot je stejný jako součet nadprůměrných hodnot – oba součty jsou v rovnováze.

- Průměr centrovaných hodnot je nulový, protože $\frac{1}{n}\sum_{i=1}^{n}(x_i - m) = \frac{1}{n}\sum_{i=1}^{n}x_i - \frac{1}{n}\sum_{i=1}^{n}m = m - \frac{1}{n}\cdot n \cdot m = 0 = 0.$

- Výraz $\sum_{i=1}^{n}(x_i - a)^2$ (tzv. kvadratická odchylka) nabývá svého minima pro a = m. Uvedený výraz charakterizuje celkovou chybu, které se dopustíme, když datový soubor nahradíme jedinou hodnotou a. Tato chyba je tedy nejmenší, když datový soubor nahradíme aritmetickým průměrem, přičemž za míru chyby považujeme kvadratickou odchylku.

- Aritmetický průměr je silně ovlivněn extrémními hodnotami.

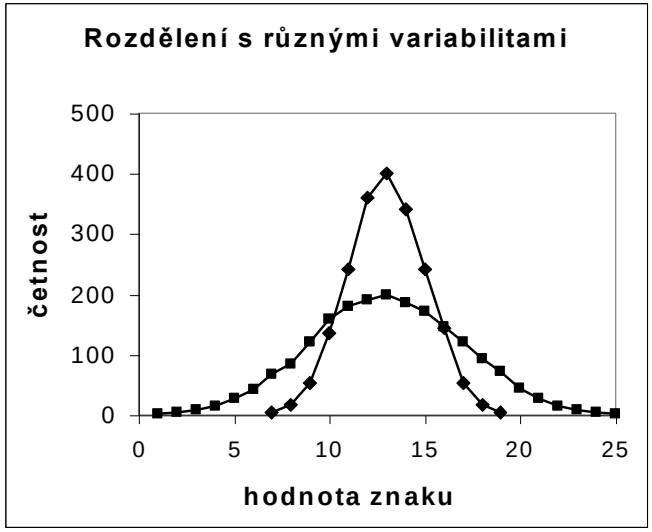
- Aritmetický průměr je vhodné použít, pokud je rozložení dat přibližně symetrické.

Rozptyl, směrodatná odchylka

Charakteristika variability: rozptyl je průměrná kvadratická odchylka hodnot od jejich aritmetického průměru $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2$. Kladná odmocnina z rozptylu se nazývá směrodatná odchylka $s = \sqrt{s^2}$. Pomocí směrodatné odchylky zavedeme i-tou standardizovanou hodnotu $\frac{x_i - m}{s}$ (vyjadřuje, o kolik směrodatných odchylek se i-tá hodnota odchýlila od průměru).

Výpočetní tvar vzorce pro rozptyl: $s^2 = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - m^2$

Znázornění rozložení četností dvou datových souborů, které se liší rozptylem:



Příklad

Příklad: Pro datový soubor obsahující údaje o mezi plasticity (znak X) a mezi pevnosti oceli (znak Y) vypočtěte rozptyly a směrodatné odchylky znaků X, Y. Přitom již víme, že $m_1 = 95,5$ a $m_2 = 114,4$.

X	Y	X	Y	X	Y
154	178	83	98	73	76
133	164	106	111	77	86
58	75	92	104	47	61
145	161	85	103	68	85
94	107	112	118	137	142
113	141	98	102	44	68
86	97	103	108	92	116
121	127	99	119	141	157
119	138	104	128	155	189
112	125	107	118	136	155
85	97	98	140	82	81
41	72	97	115	136	163
96	113	105	101	72	79
45	89	71	93	66	81
99	109	39	69	42	61
51	95	122	147	113	123
101	114	33	52	42	85
160	169	78	117	133	147
87	101	114	137	153	179
88	139	125	149	85	91

Řešení:

$$s_1^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - m_1^2 = \frac{1}{60}\left(154^2 + 133^2 + \ldots + 85^2\right) - 95,5^2 = 1052,40,\ s_1 = \sqrt{1052,40} = 32,4$$

$$s_2^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - m_2^2 = \frac{1}{60}\left(178^2 + 164^2 + \ldots + 91^2\right) - 114,4^2 = 1057,21,\ s_1 = \sqrt{1057,21} = 32,5$$

Rozptyl, směrodatná odchylka - vlastnosti

Vlastnosti rozptylu a směrodatné odchylky:

- Směrodatná odchylka je nulová pouze tehdy, když jsou všechny hodnoty stejné, jinak je kladná.

- Rozptyl centrovaných hodnot je roven původnímu rozptylu, neboť $\frac{1}{n}\sum_{i=1}^{n}[(x_i - m) - 0]^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2 = s^2$

- Rozptyl standardizovaných hodnot je 1, protože $\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - m}{s} - 0\right)^2 = \frac{1}{s^2}\cdot\frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2 = \frac{s^2}{s^2} = 1$

- Rozptyl či směrodatná odchylka jsou stejně jako průměr silně ovlivněny extrémními hodnotami.

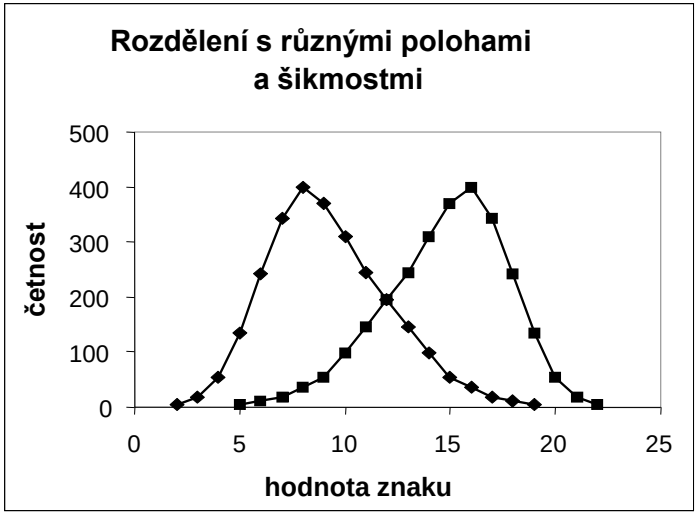
- Rozptyl či směrodatná odchylka se nehodí jako charakteristiky variability, je-li rozložení dat nesymetrické.

Šikmost

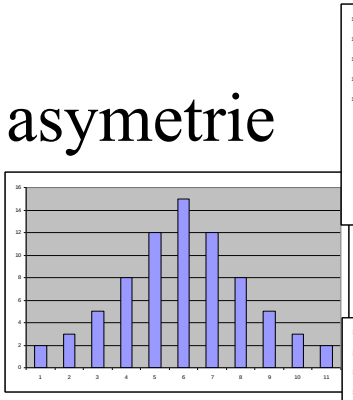
Charakteristika nesymetrie dat: **šikmost** $\alpha_3 = \dfrac{\frac{1}{n}\sum_{i=1}^{n}(x_i - m)^3}{s^3}$

Je-li rozložení dat symetrické kolem aritmetického průměru, pak $\alpha_3 = 0$.
Má-li rozložení dat prodloužený pravý konec, jde o kladně zešikmené rozložení $\alpha_3 > 0$.
Má-li rozložení dat prodloužený levý konec, jde o záporně zešikmené rozložení $\alpha_3 < 0$.

Znázornění rozložení četností dvou datových souborů, které se liší aritmetickým průměrem a šikmostí

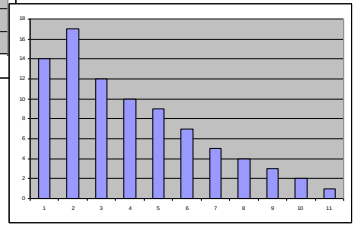


$\alpha_3 < 0$: Pravostranná asymetrie



$\alpha_3 = 0$: Symetrie

$\alpha_3 > 0$: Levostranná asymetrie

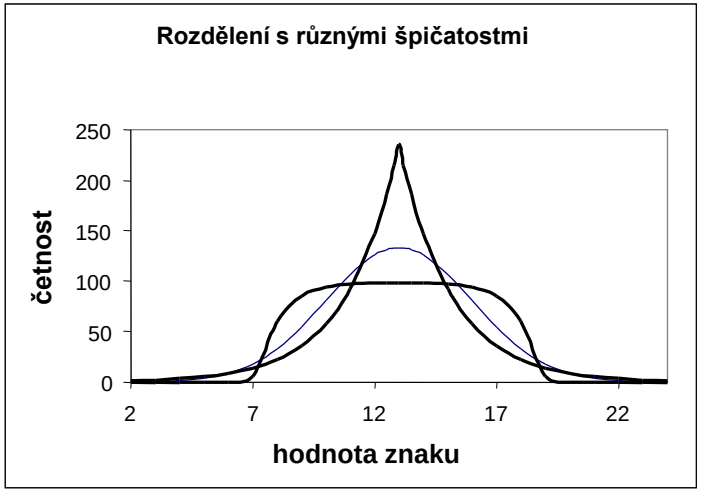


Špičatost

Charakteristika koncentrace dat kolem průměru : špičatost $\alpha_4 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - m)^4}{s^4} - 3$

Je-li rozložení dat normální (Gaussovo), pak $\alpha_4 = 0$.
Je-li rozložení dat strmé, pak $\alpha_4 > 0$.
Je-li rozložení dat ploché, pak $\alpha_4 < 0$.

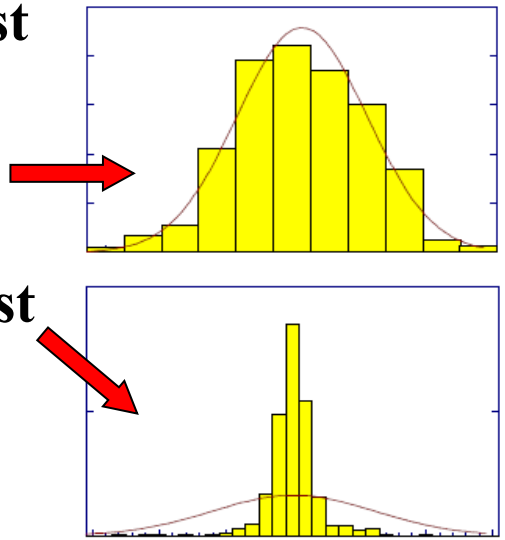
Znázornění rozložení četností dvou datových souborů, které se liší špičatostí



$\alpha_4 < 0$: **Podnormální špičatost**

$\alpha_4 = 0$: **Normální špičatost**

$\alpha_4 > 0$: **Nadnormální špičatost**



Kovariance

Charakteristika společné variability dvou intervalových znaků: kovariance

Předpokládejme, že máme dvourozměrný datový soubor $\begin{pmatrix} x_1 & y_1 \\ \cdots & \cdots \\ x_n & y_n \end{pmatrix}$. Označme m_1, m_2 průměry znaků X, Y a s_1, s_2 směrodatné odchylky znaků X, Y. Zavedeme kovarianci jako charakteristiku společné variability znaků X, Y kolem jejich průměrů

$$s_{12} = \frac{1}{n}\sum_{i=1}^{n}(x_i - m_1)(y_i - m_2).$$

Kovariance je průměrem součinů centrovaných hodnot.
Pokud se nadprůměrné (podprůměrné) hodnoty znaku X sdružují s nadprůměrnými (podprůměrnými) hodnotami znaku Y, budou součiny centrovaných hodnot $x_i - m_1$ a $y_i - m_2$ vesměs kladné a jejich průměr (tj. kovariance) rovněž. Znamená to, že mezi znaky X, Y existuje určitý stupeň přímé lineární závislosti. Říkáme, že znaky X, Y jsou kladně korelované.

Pokud se nadprůměrné (podprůměrné) hodnoty znaku X sdružují s podprůměrnými (nadprůměrnými) hodnotami znaku Y, budou součiny centrovaných hodnot vesměs záporné a jejich průměr rovněž. Znamená to, že mezi znaky X a Y existuje určitý stupeň nepřímé lineární závislosti. Říkáme, že znaky X, Y jsou záporně korelované.

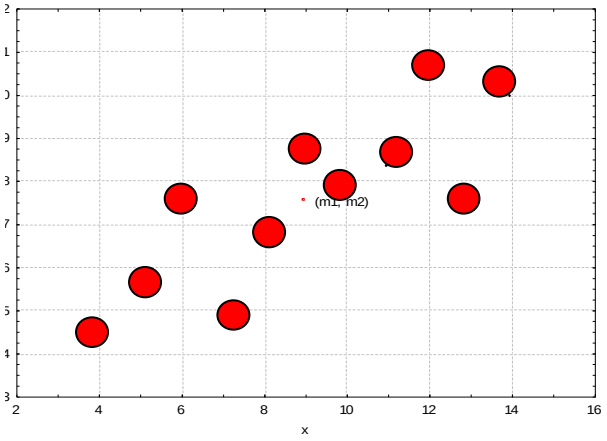
Je-li kovariance nulová, pak řekneme, že znaky X, Y jsou nekorelované a znamená to, že mezi nimi neexistuje žádná lineární závislost.

Pro výpočet kovariance používáme vzorec: $$s_{12} = \left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right) - m_1 m_2 .$$

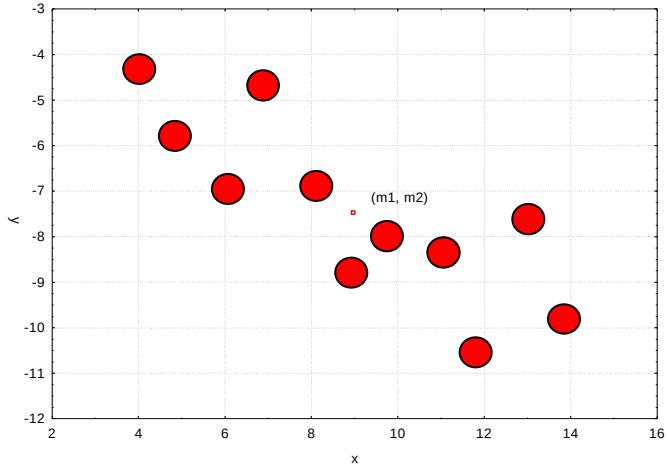
Kovariance

Znázornění významu kovariance

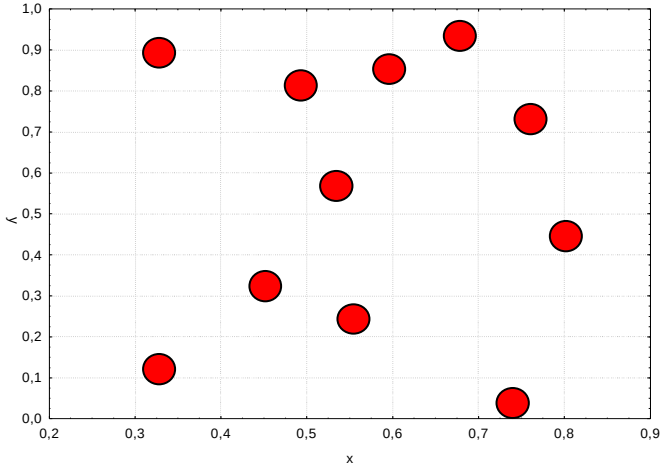
= 5,5



s_{12} = -5,5



s_{12} = 0



Příklad

Příklad: Pro datový soubor obsahující údaje o mezi plasticity (znak X) a mezi pevnosti oceli (znak Y) vypočtěte kovarianci znaků X, Y. Přitom již víme, že $m_1 = 95,5$, $m_2 = 114,4$, $s_1 = 32,4$, $s_2 = 32,5$

X	Y	X	Y	X	Y
154	178	83	98	73	76
133	164	106	111	77	86
58	75	92	104	47	61
145	161	85	103	68	85
94	107	112	118	137	142
113	141	98	102	44	68
86	97	103	108	92	116
121	127	99	119	141	157
119	138	104	128	155	189
112	125	107	118	136	155
85	97	98	140	82	81
41	72	97	115	136	163
96	113	105	101	72	79
45	89	71	93	66	81
99	109	39	69	42	61
51	95	122	147	113	123
101	114	33	52	42	85
160	169	78	117	133	147
87	101	114	137	153	179
88	139	125	149	85	91

Řešení:

$$s_{12} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - m_1 m_2 = \frac{1}{60}(154 \cdot 178 + 133 \cdot 164 + \ldots + 85 \cdot 91) - 95,5 \cdot 114,4 = 985,76$$

Pearsonův koeficient korelace

Charakteristika těsnosti závislosti dvou intervalových znaků: Pearsonův koeficient korelace

Jsou-li směrodatné odchylky s_1, s_2 nenulové, pak definujeme Pearsonův koeficient korelace znaků X, Y vzorcem: $r_{12} = \frac{1}{n}\sum_{i=1}^{n}\frac{x_i - m_1}{s_1}\frac{y_i - m_2}{s_2}$. Je to průměr součinů standardizovaných hodnot. Počítá se podle vzorce $r_{12} = \frac{s_{12}}{s_1 s_2}$.

Příklad: Pro datový soubor obsahující údaje o mezi plasticity (znak X) a mezi pevnosti oceli (znak Y) vypočtěte koeficient korelace znaků X, Y. Přitom již víme, že $m_1 = 95,5$, $m_2 = 114,4$, $s_1 = 32,4$, $s_2 = 32,5$, $s_{12} = 985,76$.

Řešení:

$$r_{12} = \frac{s_{12}}{s_1 s_2} = \frac{985,76}{32,4.32,5} = 0,936$$

Koeficient korelace svědčí o tom, že mezi oběma znaky existuje velmi silná přímá lineární závislost – čím je vyšší mez plasticity, tím je vyšší mez pevnosti a čím je nižší mez plasticity, tím je nižší mez pevnosti.

Vlastnosti Pearsonova koeficientu korelace:

Pro koeficient korelace platí $-1 \le r_{12} \le 1$ a rovnosti je dosaženo právě když mezi hodnotami $x_1, ..., x_n$ a $y_1, ..., y_n$ existuje úplná lineární závislost, tj. existují konstanty a, b tak, že $y_i = a + bx_i$, $i = 1, ..., n$, přičemž znaménko + platí pro $b > 0$, znaménko – pro $b < 0$. (Uvedená nerovnost se nazývá Cauchyova – Schwarzova – Buňakovského nerovnost.)

Tedy čím je r_{12} bližší 1, tím je silnější přímá lineární závislost mezi znaky X a Y, čím je bližší –1, tím je silnější nepřímá lineární závislost mezi X a Y.

Je-li $r_{12} = 1$ resp. $r_{12} = -1$, pak dvojice (x_i, y_i) leží na nějaké rostoucí resp. klesající přímce.

Hodnoty r_{12} se nezmění, když u x-ových a y-ových hodnot současně provedeme vzestupnou resp. sestupnou lineární transformaci.

Hodnoty r_{12} se vynásobí -1, když u x-ových hodnot provedeme vzestupnou (resp. sestupnou) a u y-ových hodnot sestupnou (resp. vzestupnou) lineární transformaci.

Koeficient je symetrický, tj. $r_{12} = r_{21}$.

Početní pravidla pro číselné charakteristiky

Početní pravidla pro číselné charakteristiky

Nechť m_1 je aritmetický průměr a s_1^2 rozptyl znaku X. Pak znak **Y = a + bX** má:

aritmetický průměr $m_2 = a + bm_1$, rozptyl $s_2^2 = b^2 s_1^2$

Nechť m_1 , m_2 jsou aritmetické průměry, s_1^2 , s_2^2 rozptyly a s_{12} kovariance znaků X, Y. Pak znak **U = X + Y** má

aritmetický průměr $m_3 = m_1 + m_2$, rozptyl $s_3^2 = s_1^2 + s_2^2 + 2s_{12}$

Nechť s_{12} je kovariance znaků X, Y a m_1 , m_2 jsou aritmetické průměry znaků X, Y. Pak znaky **U = a + bX, V = c + dY** mají kovarianci

$s_{34} = bds_{12}$

Příklad

Příklad:
a) Znak X má aritmetický průměr 2 a rozptyl 3. Najděte aritmetický průměr a rozptyl znaku Y = -1 + 3X.

b) Znaky X a Y mají aritmetické průměry 3 a 2, rozptyly 2 a 3, kovarianci 1,5. Vypočtěte aritmetický průměr a rozptyl znaku Z = 5X – 4Y.

c) Součet rozptylů dvou znaků je 120, součin 1000 a rozptyl jejich součtů je 100. Vypočtěte koeficient korelace těchto znaků.

Řešení:
ad a) $m_2 = -1 + 3m_1 = -1 + 3 \times 2 = 5$, $s_2^2 = 3^2 \times s_1^2 = 9 \times 3 = 27$.

ad b) $m_3 = 5m_1 - 4m_2 = 5 \times 3 - 4 \times 2 = 7$, $s_3^2 = 5^2 \times s_1^2 + (-4)^2 \times s_2^2 + 2 \times 5 \times (-4) \times s_{12} = 25 \times 2 + 16 \times 3 - 40 \times 1{,}5 = 38$.

ad c) $s_1^2 + s_2^2 = 120$, $s_1^2 \times s_2^2 = 1000$, $s_{1+2}^2 = 100 = s_1^2 + s_2^2 + 2s_{12} \Rightarrow s_{12} = \frac{1}{2}\left(s_{1+2}^2 - s_1^2 - s_2^2\right) = \frac{1}{2}(100 - 120) = -10$

$$r_{12} = \frac{s_{12}}{s_1 \times s_2} = \frac{-10}{\sqrt{1000}} = -0{,}3162\,.$$

Vážené číselné charakteristiky

Pokud nemáme k dispozici původní datový soubor, ale jenom tabulku rozložení četností (resp. kontingenční tabulku), můžeme vypočítat tzv. vážené číselné charakteristiky.

Vážený aritmetický průměr: $m = \frac{1}{n}\sum_{j=1}^{r} n_j x_{[j]}$

Vážený rozptyl: $s^2 = \frac{1}{n}\sum_{j=1}^{r} n_j \left(x_{[j]} - m\right)^2 = \frac{1}{n}\sum_{j=1}^{r} n_j x_{[j]}^{\ 2} - m^2$

Vážená kovariance: $s_{12} = \frac{1}{n}\sum_{j=1}^{r}\sum_{k=1}^{s} n_{jk}\left(x_{[j]} - m_1\right)\left(y_{[k]} - m_2\right) = \frac{1}{n}\sum_{j=1}^{r}\sum_{k=1}^{s} n_{jk} x_{[j]} y_{[k]} - m_1 m_2$

Příklad

Příklad na výpočet vážených číselných charakteristik

Z dvourozměrné ho datového souboru rozsahu 27, v němž znak X má varianty 1, 2, 3 a znak Y má rovněž varianty 1, 2, 3, byly určeny simultánní absolutní četnosti: $n_{11} = 5$, $n_{12} = 1$, $n_{13} = 3$, $n_{21} = 4$, $n_{22} = 3$, $n_{23} = 4$, $n_{31} = 2$, $n_{32} = 3$, $n_{33} = 2$.

a) Vypočtěte průměry a směrodatné odchylky znaků X a Y.

b) Vypočtěte a interpretujte koeficient korelace znaků X a Y.

x	y			$n_{j.}$
	1	2	3	
1	5	1	3	9
2	4	3	4	11
3	2	3	2	7
$n_{.k}$	11	7	9	27

Řešení:

Kontingenční tabulka simultánních absolutních četností:

ad a) $m_1 = \frac{1}{27}(1\cdot 9 + 2\cdot 11 + 3\cdot 7) = \frac{52}{27} = 1{,}926$, $m_2 = \frac{1}{27}(1\cdot 11 + 2\cdot 7 + 3\cdot 9) = \frac{52}{27} = 1{,}926$

$s_1^2 = \frac{1}{27}(1^2\cdot 9 + 2^2 11 + 3^2\cdot 7) - \left(\frac{52}{27}\right)^2 = \frac{116}{27} - \frac{2704}{729} = \frac{428}{729}$, $s_1 = 0{,}766$

$s_2^2 = \frac{1}{27}(1^2\cdot 11 + 2^2\cdot 7 + 3^2\cdot 9) - \left(\frac{52}{27}\right)^2 = \frac{120}{27} - \frac{2704}{729} = \frac{536}{729}$, $s_2 = 0{,}857$

ad b)

$s_{12} = \frac{1}{27}(1\cdot 1\cdot 5 + 1\cdot 2\cdot 1 + 1\cdot 3\cdot 3 + 2\cdot 1\cdot 4 + 2\cdot 2\cdot 3 + 2\cdot 3\cdot 4 + 3\cdot 1\cdot 2 + 3\cdot 2\cdot 3 + 3\cdot 3\cdot 2) - \frac{52}{27}\cdot\frac{52}{27}$

$= \frac{102}{27} - \frac{2704}{729} = \frac{2754 - 2704}{729} = \frac{50}{729} = 0{,}0685871$

$r_{12} = \frac{\frac{50}{729}}{\sqrt{\frac{428}{729}\cdot\frac{536}{729}}} = 0{,}10439$.

Mezi znaky X a Y existuje velmi slabá přímá lineární závislost.

Koeficient variace, geometrický průměr

Pro poměrové znaky používáme jako charakteristiku variability koeficient variace $\frac{s}{m}$. Je to bezrozměrné číslo, které se často vyjadřuje v procentech. Umožňuje porovnat variabilitu několika znaků.

Jsou-li všechny hodnoty poměrového znaku kladné, pak jako charakteristiku polohy lze užít geometrický průměr $\sqrt[n]{x_1 \cdot \ldots \cdot x_n}$.

Příklad: Pro datový soubor obsahující údaje o mezi plasticity (znak X) a mezi pevnosti oceli (znak Y) vypočtěte koeficienty variace znaků X, Y. Přitom již víme, že $m_1 = 95{,}5$, $m_2 = 114{,}4$, $s_1 = 32{,}4$, $s_2 = 32{,}5$

Řešení:

$$cv_1 = \frac{s_1}{m_1} = \frac{32{,}4}{95{,}5} = 0{,}339, cv_2 = \frac{s_2}{m_2} = \frac{32{,}5}{114{,}4} = 0{,}284$$

Výpočty zavedením pomocné proměnné

- pomocná proměnná $\Rightarrow$ $$v_i = \frac{x_i - a}{h}$$

- konstanty:
  - $a \rightarrow$ střed třídy s nejvyšší četností
  - $h \rightarrow$ šířka třídy

Výpočty zavedením pomocné proměnné

$$v = \frac{\bar{x} - a}{h} \Rightarrow \bar{x} = vh + a$$

$$s_v^2 = \frac{s_x^2}{h^2} \Rightarrow s_x^2 = h^2 s_v^2$$

Příklad

x_i	n_i
<30 – 40)	10
<40 – 50)	31
<50 – 60)	27
<60 – 70)	19
<70 – 80)	13
Celkem	**100**

Vypočítejte:

- aritmetický průměr, rozptyl, směrodatnou odchylku a variační koeficient zavedením pomocné proměnné

Příklad

$a = 45$ $h = 10$

x_i	n_i	v_i
35	10	-1
45	31	0
55	27	1
65	19	2
75	13	3
Součet	**100**	**x**

Příklad

x_i	n_i	v_i	$v_i n_i$
35	10	-1	-10
45	31	0	0
55	27	1	27
65	19	2	38
75	13	3	39
Součet	**100**	**x**	**94**

$$\bar{v} = 0{,}94 \Rightarrow \bar{x} = \bar{v}h + a =$$

$$= 0{,}94 \cdot 10 + 45 = 54{,}4$$

Příklad

x_i	n_i	v_i	$v_i^2 n_i$
35	10	-1	10
45	31	0	0
55	27	1	27
65	19	2	76
75	13	3	117
Součet	**100**	**x**	**230**

$$s_v^2 = \frac{1}{n}\sum_{i=1}^{n}(v_i^2 n_i) - \bar{v}^2 = 2{,}3 - 0{,}8836 = 1{,}4164$$

$$\Rightarrow \quad s_x^2 = h^2 \cdot s_v^2 = 10^2 \cdot 1{,}4164 = 141{,}64$$

Příklad

$$s_x = \sqrt{s_x^2} = \sqrt{141{,}64} = 11{,}9$$

nebo $$s_x = h \cdot s_v^2 = 10 \cdot \sqrt{1{,}4164} = 11{,}9$$

$$cv_x = \frac{s_x}{\bar{x}} = \frac{11{,}9}{55{,}4} = 0{,}2188$$

nebo $$cv_x = \frac{h \cdot s_v}{\bar{v} \cdot h + a} = \frac{10 \cdot 1{,}19}{0{,}94 \cdot 10 + 45} = 0{,}2188$$

Společný rozptyl

$$s^2 = \overline{s}^2 + s_{\overline{x}}^2$$

$\overline{s}^2$ **……vnitroskupinová variabilita** ($s_{\#}^2$)[1]

$s_{\overline{x}}^2$ **……meziskupinová variabilita** (s_*^2)

[1] Značení ze skript „Popisná statistika“

Společný rozptyl

- **vnitroskupinová variabilita**

$$\bar{s}^2 = \frac{1}{n}\sum_{i=1}^{k} s_i^2 \cdot n_i$$

- **meziskupinová variabilita**

$$s_{\bar{x}}^2 = \frac{1}{n}\sum_{i=1}^{k} (\bar{x}_i - \bar{x})^2 \cdot n_i$$

Příklad

D1:	**104**	**108**	**79**	**155**
D2:	**93**	**65**	**76**	**111**

Vypočítejte:

- **dílčí průměry,**
- **společný průměr,**
- **dílčí rozptyly,**
- **společný rozptyl.**

Příklad

$$\bar{x}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i^1 =$$

$$= \frac{1}{4}\cdot(104+108+79+155) = \underline{\underline{111{,}5}}$$

$$\bar{x}_2 = \frac{1}{4}\cdot(93+65+76+111) = \underline{\underline{86{,}25}}$$

$$\bar{x} = \frac{1}{8}\cdot(111{,}5\cdot 4+86{,}25\cdot 4) = \underline{\underline{98{,}875}}$$

Příklad

$$s_{x1}^2 = \frac{1}{n}\sum_{i=}^{n_1}\left(x_i^1 - \bar{x}_1\right)^2 =$$

$$= \frac{7{,}5^2 + 3{,}5^2 + 32{,}5^2 + 43{,}5^2}{4} = \underline{\underline{754{,}25}}$$

$$s_{x2}^2 = \frac{6{,}75^2 + 21{,}25^2 + 10{,}25^2 + 24{,}75^2}{4} = \underline{\underline{303{,}69}}$$

Příklad

$$\bar{s}^2 = \frac{1}{n}\sum_{i=1}^{k} s_i^2 \cdot n_i = \frac{1}{8}\left(s_{x1}^2 \cdot 4 + s_{x2}^2 \cdot 4\right) =$$

$$= \frac{1}{8} \cdot (754{,}25 \cdot 4 + 303{,}69 \cdot 4) = \underline{\underline{528{,}97}}$$

$$s_{\bar{x}}^2 = \frac{1}{n}\sum_{i=1}^{k} \left(\bar{x}_i - \bar{\bar{x}}\right)^2 \cdot n_i =$$

$$= \frac{(111{,}5 - 98{,}875)^2 \cdot 4) + (86{,}25 - 98{,}875)^2 \cdot 4}{8} = \underline{\underline{159{,}39}}$$

Příklad

$$s^2 = \bar{s}^2 + s_{\bar{x}}^2 =$$

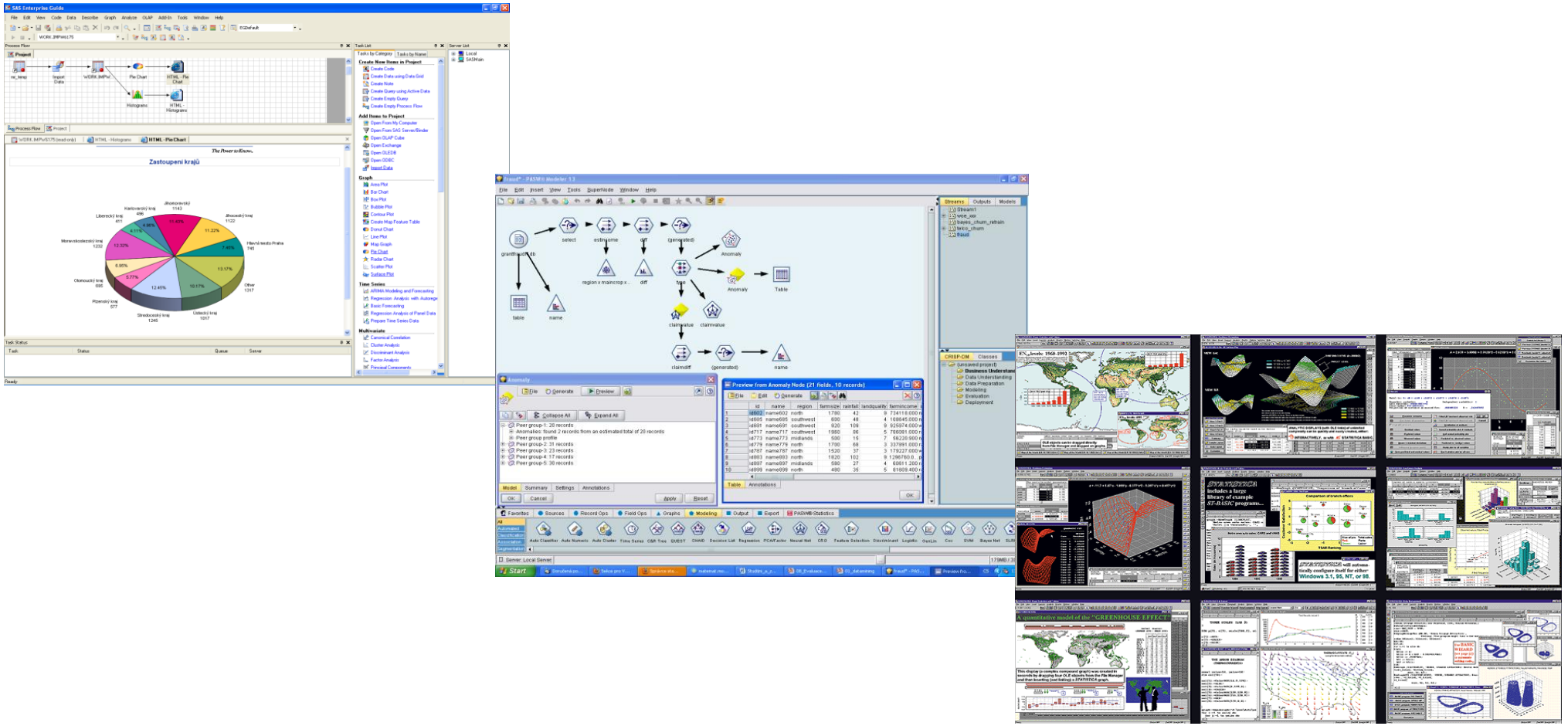
$$= 528{,}97 + 159{,}39 = \underline{\underline{688{,}36}}$$

Pro kontrolu ještě spočteme rozptyl přímo:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = \frac{1}{8} \cdot 83717 - 98{,}875^2$$

$$= 10464{,}63 - 9776{,}27 = \underline{\underline{688{,}36}}$$

3. Statistický software, základy práce v SAS



Software

AcaStat	GAUSS	MRDCL	RATS	StatsDirect
ADaMSoft	GAUSS	NCSS	RKWard[4]	Statistix
Analyse-it	GenStat	OpenEpi	SalStat	SYSTAT
ASReml	Golden Helix	Origin	SAS	The Unscrambler
Auguri	gretl	Ox programming language	SOCR	UNISTAT
BioStat	JMP	OxMetrics	Stata	VisualStat
BrightStat	MacAnova	Origin	Statgraphics	Winpepi
Dataplot	Mathematica	Partek	STATISTICA	WinSPC
EasyReg	Matlab	Primer	StatIt	XLStat
Epi Info	MedCalc	PSPP	StatPlus	XploRe
EViews	modelQED	R	SPlus	
Excel	Minitab	R Commander[4]	SPSS	

Some Available Statistical Packages

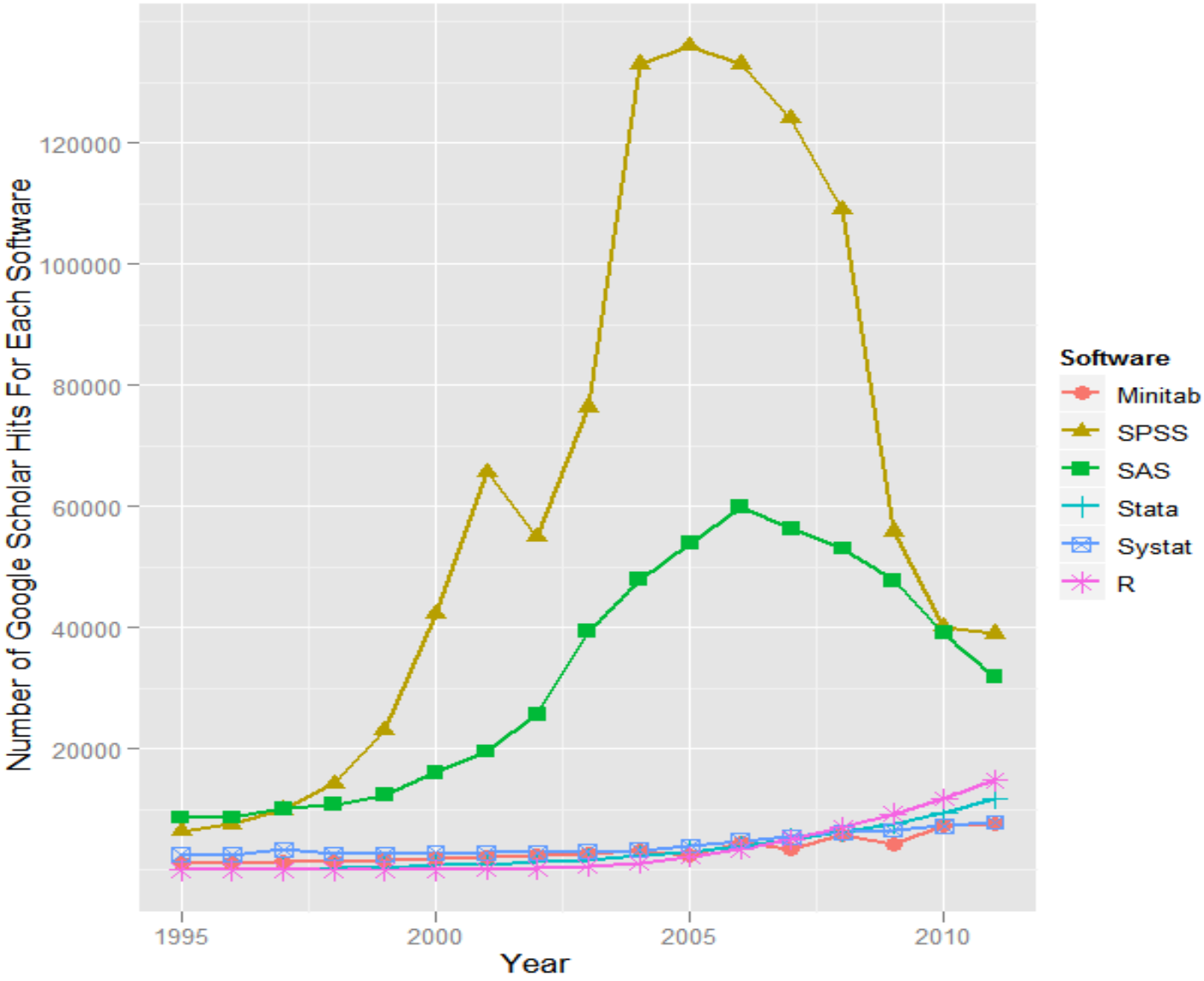
Proprietary

- Excel
- SPSS
- MINITAB
- Matlab
- Statistica
- SAS

Free Software

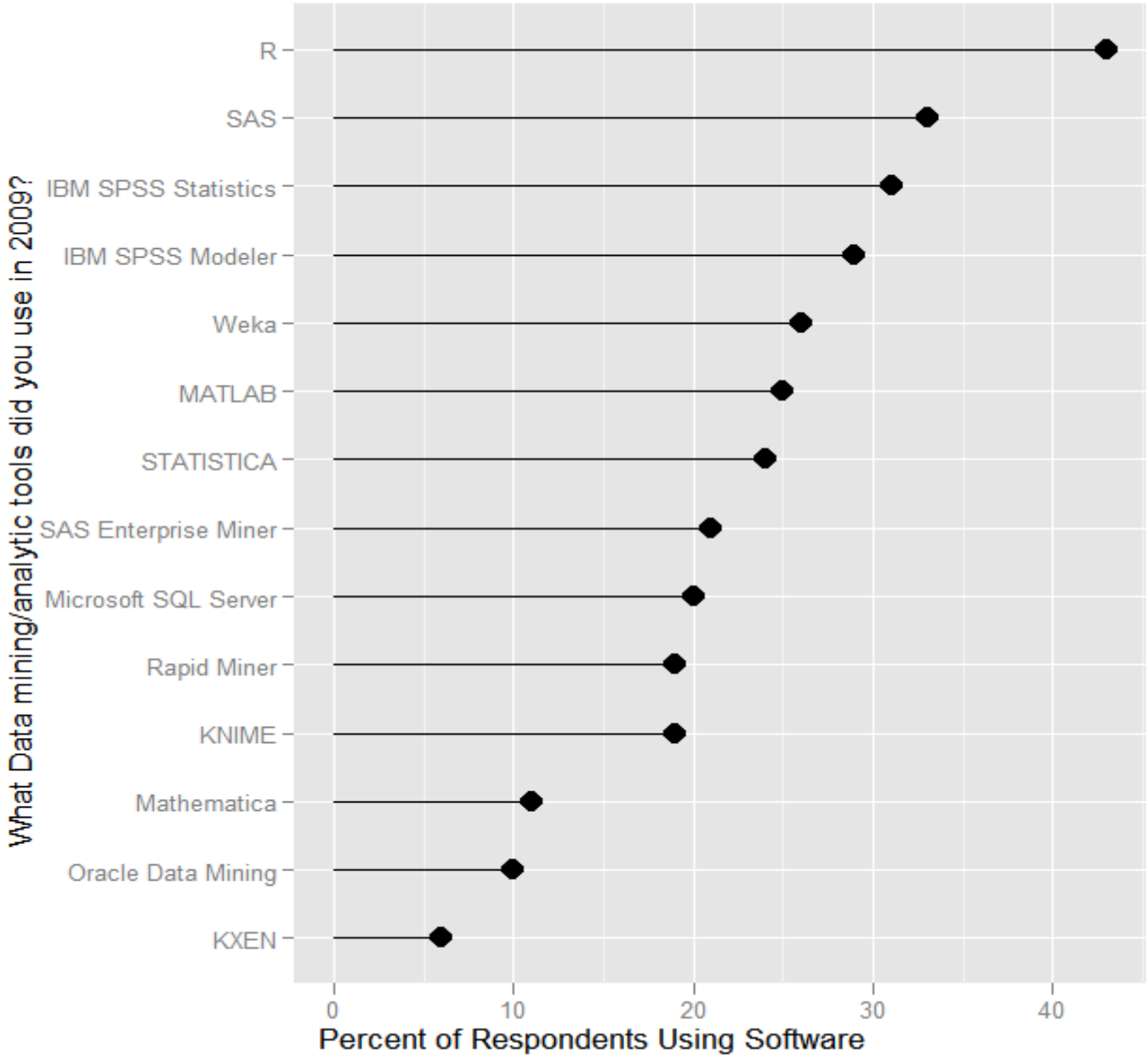
- LibreOffice Calc
- PSPP
- EpiInfo
- R
- ...

What is Used? (Academia)

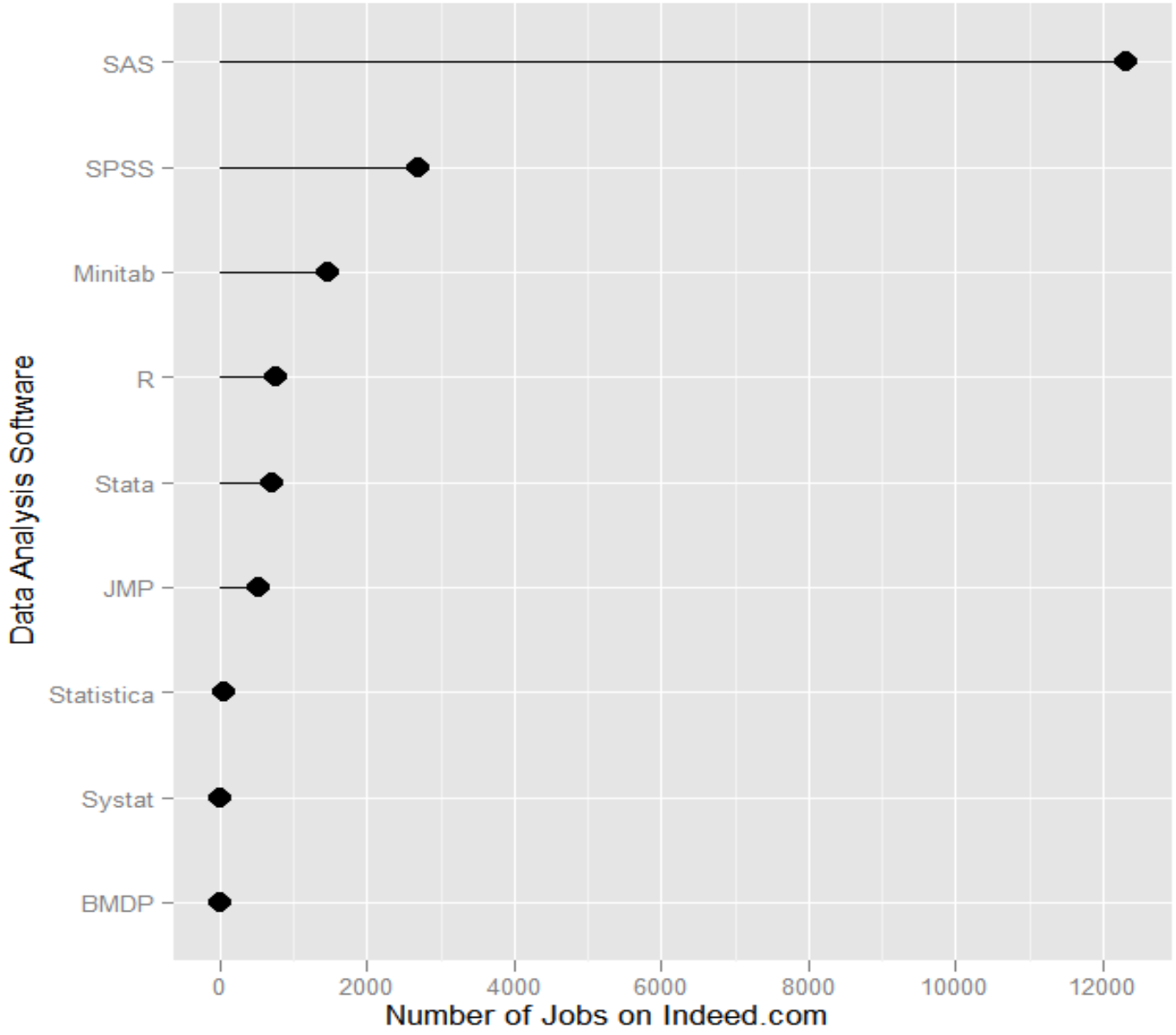


Use of data analysis software in academic publications as measured by hits on Google Scholar.

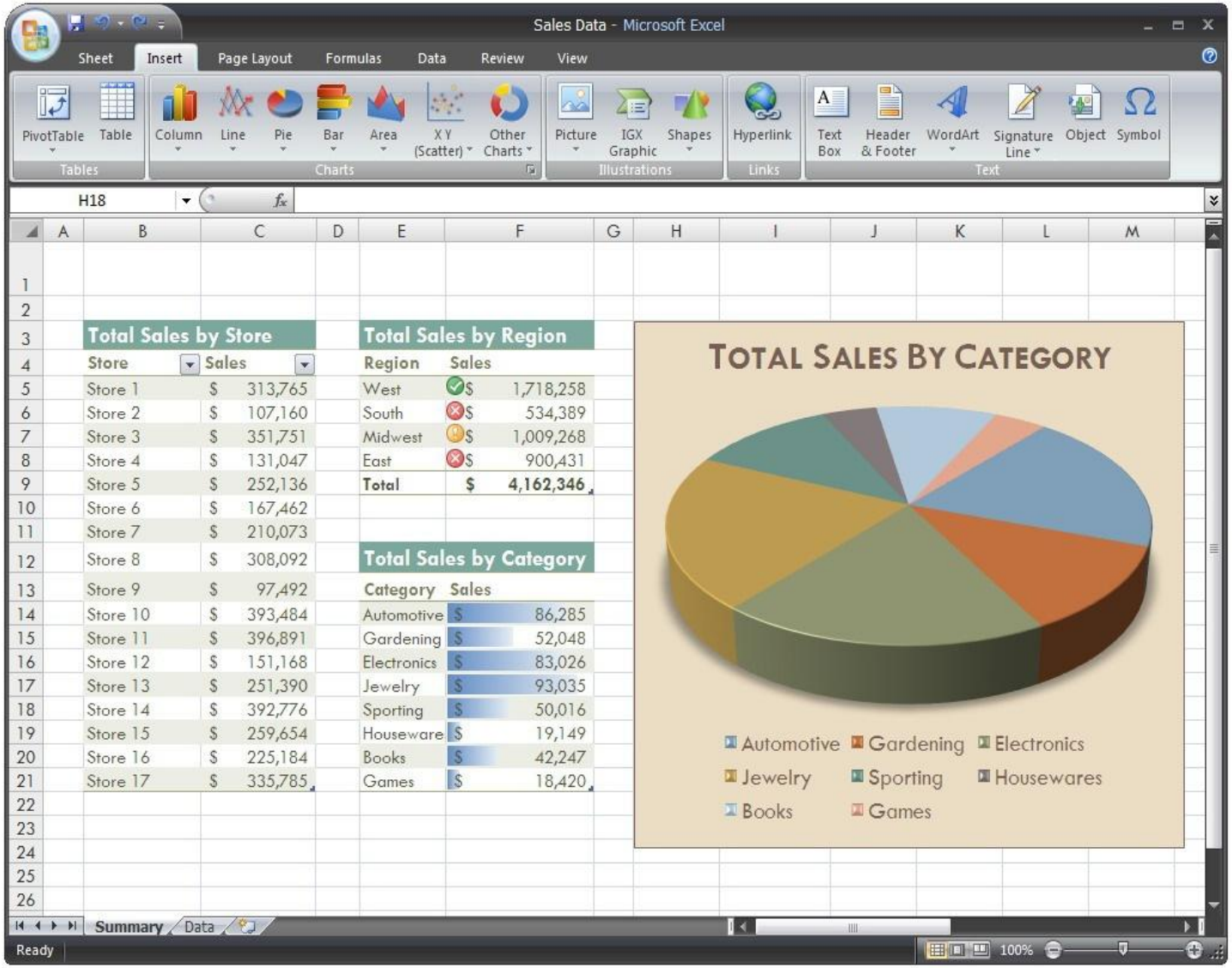
What is Used? (Survey)



What is Used? (Job Market)



Microsoft Excel



Microsoft Excel

COST

- Individual License for Microsoft Office Professional $350
- Microsoft Office University Student License: $99
- Volume Discounts available for large organizations and universities
- Free Starter Version available on some new PCs

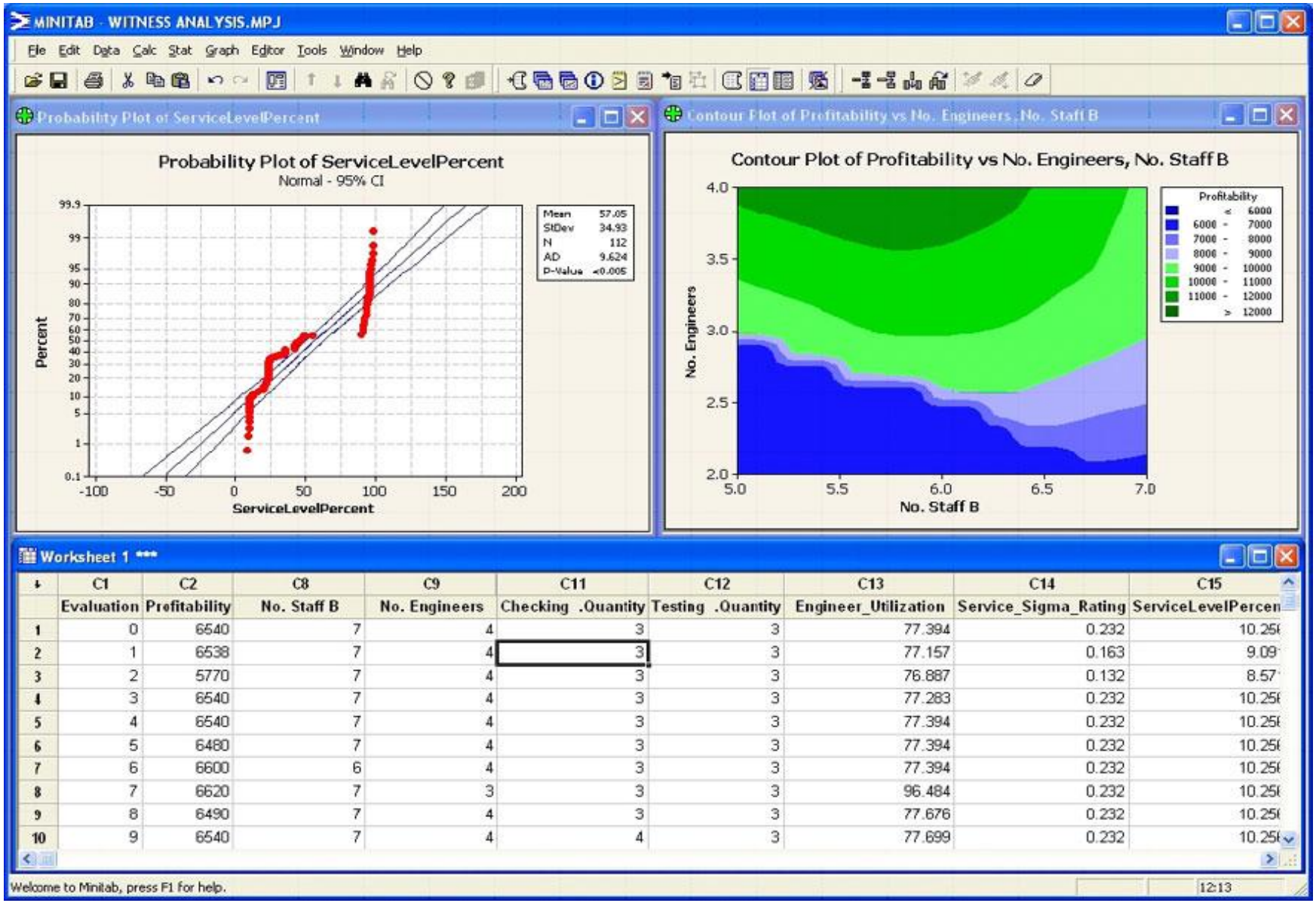
PRO

- Nearly ubiquitous and is often pre-installed on new computers
- User friendly
- Very good for basic descriptive statistics, charts and plots

CON

- Costs money
- Not sufficient for anything beyound the most basic statistical analysis

Minitab



	C1	C2	C8	C9	C11	C12	C13	C14	C15
	Evaluation	Profitability	No. Staff B	No. Engineers	Checking .Quantity	Testing .Quantity	Engineer_Utilization	Service_Sigma_Rating	ServiceLevelPercen
1	0	6540	7	4	3	3	77.394	0.232	10.25
2	1	6538	7	4	3	3	77.157	0.163	9.09
3	2	5770	7	4	3	3	76.887	0.132	8.57
4	3	6540	7	4	3	3	77.283	0.232	10.25
5	4	6540	7	4	3	3	77.394	0.232	10.25
6	5	6480	7	4	3	3	77.394	0.232	10.25
7	6	6600	6	4	3	3	77.394	0.232	10.25
8	7	6620	7	3	3	3	96.484	0.232	10.25
9	8	6490	7	4	3	3	77.676	0.232	10.25
10	9	6540	7	4	4	3	77.699	0.232	10.25

Minitab

COST

- $1,395.00 per single user license

CON

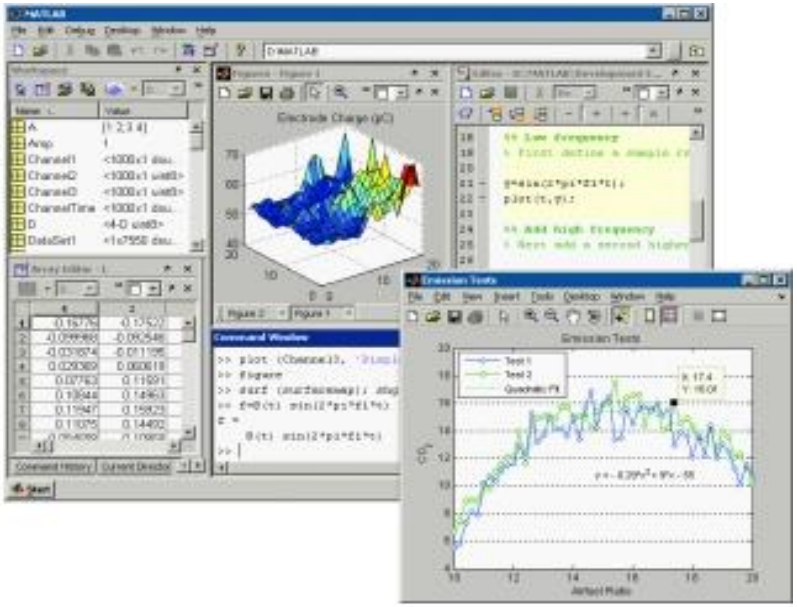
- Costs Money
- Not suitable for very complicated statistical computation and analysis
- Not often used in academic research

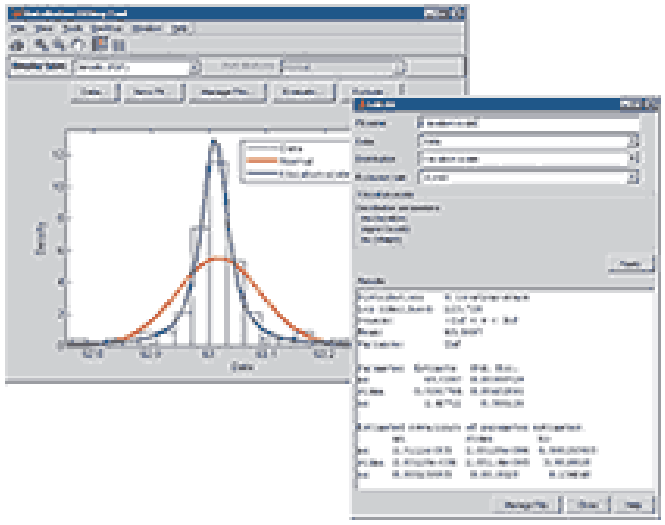
PRO

- Easy to learn and use
- Often taught in schools in introductory statistics courses
- Widely used in engineering for process improvement

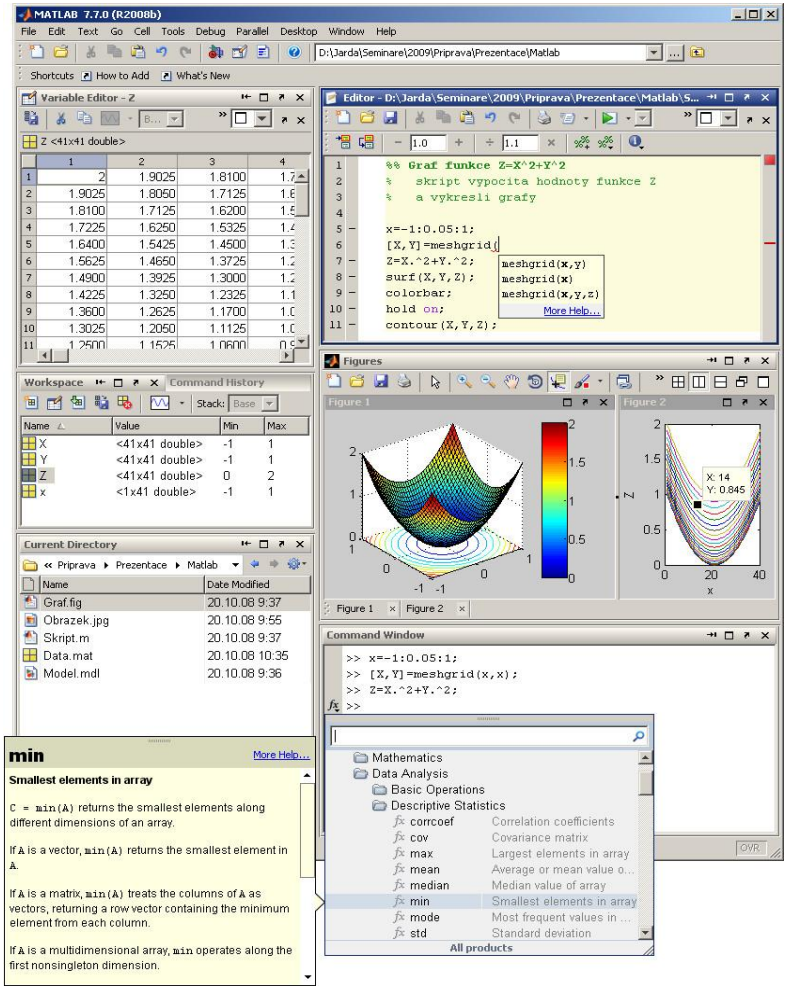
Matlab

- Matlab www.mathworks.com,
  www.humusoft.cz

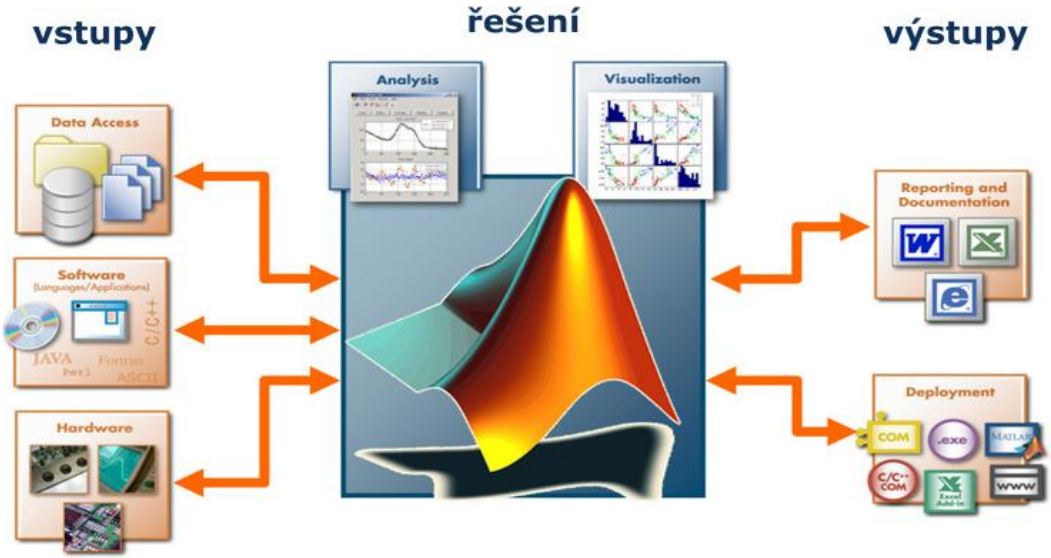




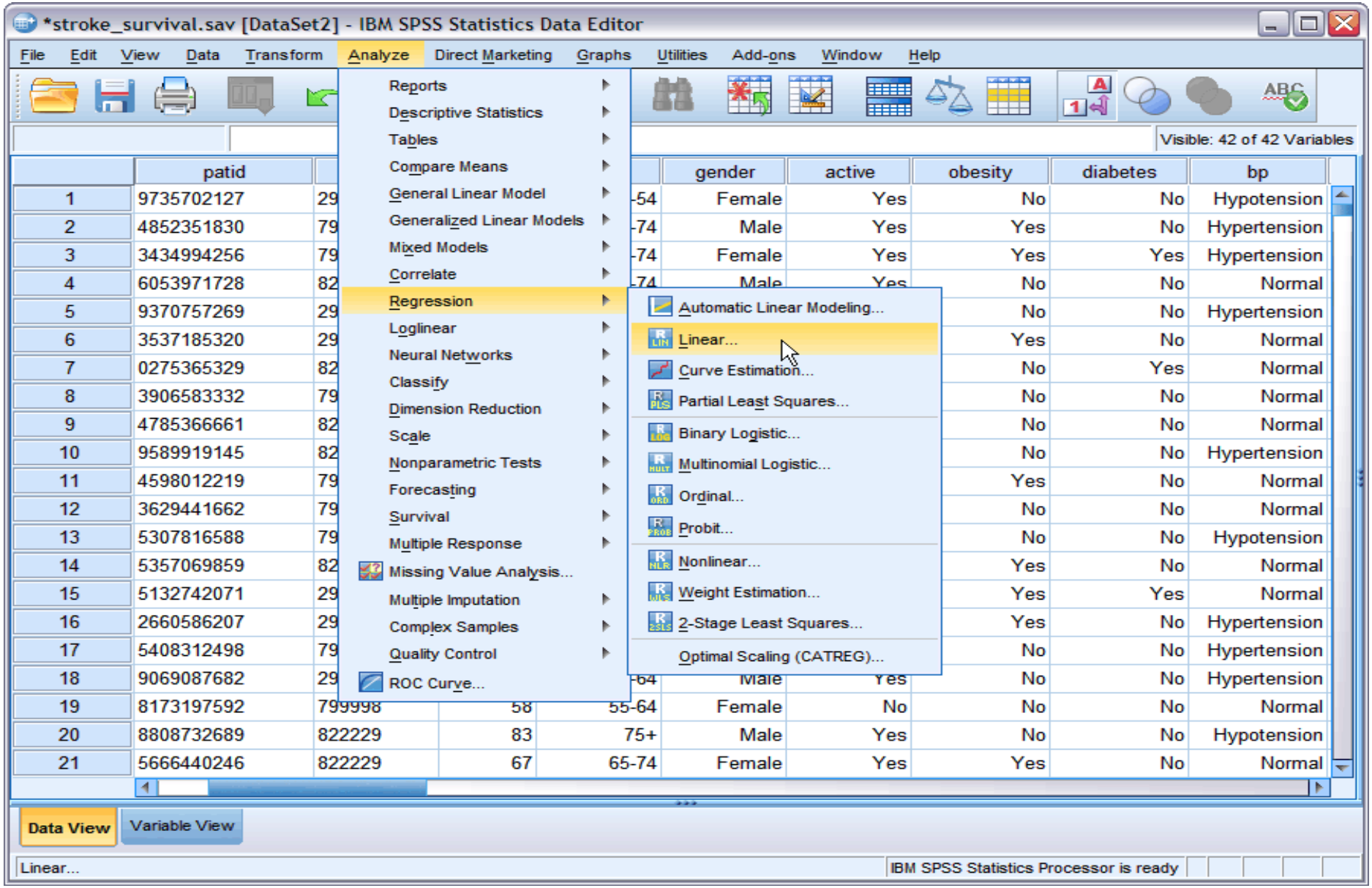
Matlab



http://www.humusoft.cz/produkty/matlab/matlab/

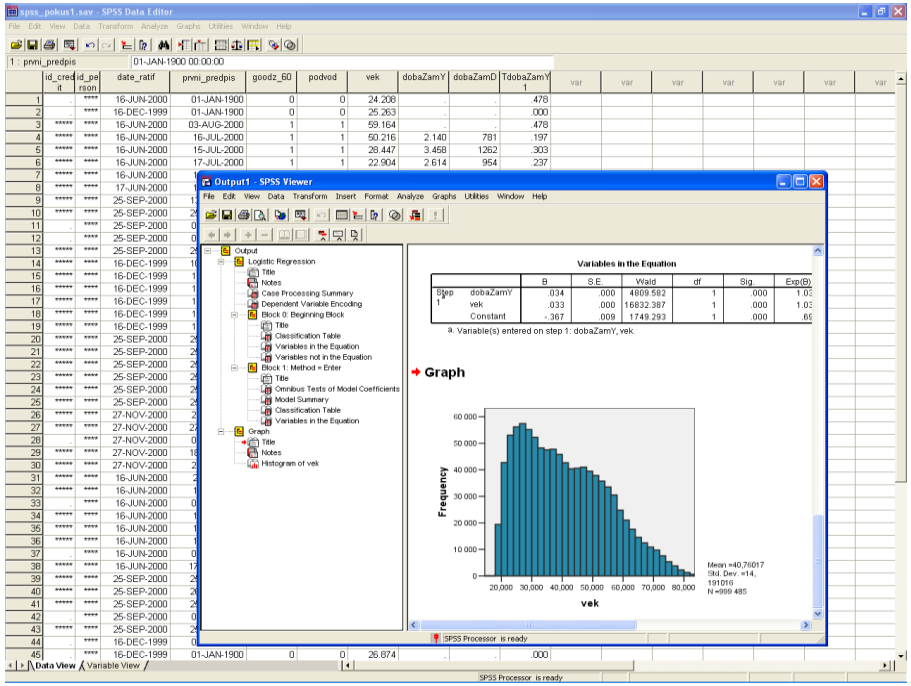


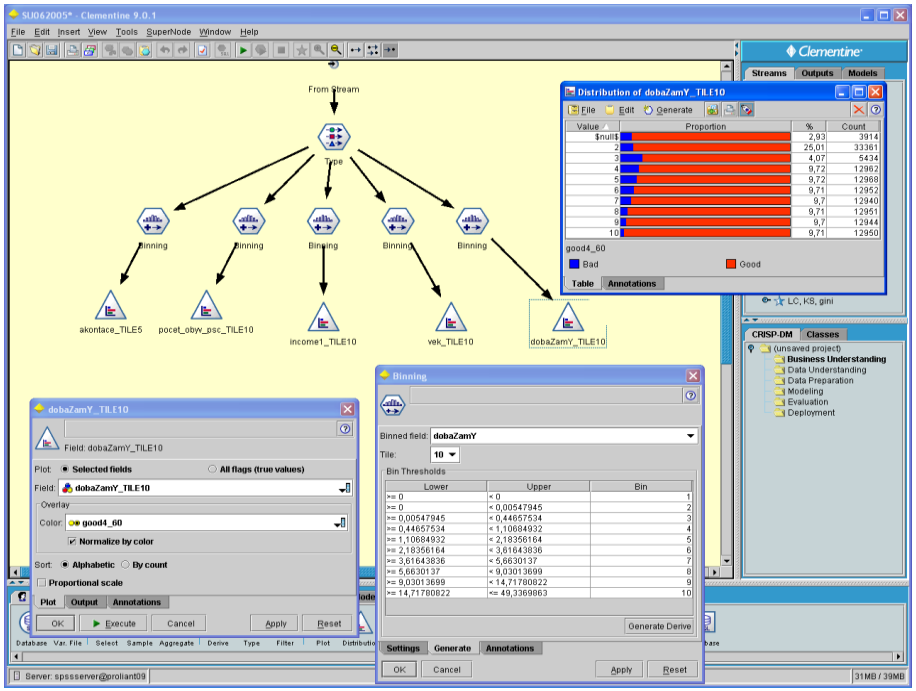
SPSS


*stroke_survival.sav [DataSet2] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Neural Networks
Classify
Dimension Reduction
Scale
Nonparametric Tests
Forecasting
Survival
Multiple Response
Missing Value Analysis...
Multiple Imputation
Complex Samples
Quality Control
ROC Curve...
Automatic Linear Modeling...
Linear...
Curve Estimation...
Partial Least Squares...
Binary Logistic...
Multinomial Logistic...
Ordinal...
Probit...
Nonlinear...
Weight Estimation...
2-Stage Least Squares...
Optimal Scaling (CATREG)...
Visible: 42 of 42 Variables
patid gender active obesity diabetes bp
1 9735702127 Female Yes No No Hypotension
2 4852351830 Male Yes Yes No Hypertension
3 3434994256 Female Yes Yes Yes Hypertension
4 6053971728 No No Normal
5 9370757269 No No Hypertension
6 3537185320 Yes No Normal
7 0275365329 No Yes Normal
8 3906583332 No No Normal
9 4785366661 No No Normal
10 9589919145 No No Hypertension
11 4598012219 Yes No Normal
12 3629441662 No No Normal
13 5307816588 No No Hypotension
14 5357069859 Yes No Normal
15 5132742071 Yes Yes Normal
16 2660586207 Yes No Hypertension
17 5408312498 No No Hypertension
18 9069087682 No No Hypertension
19 8173197592 799998 58 55-64 Female No No No Normal
20 8808732689 822229 83 75+ Male Yes No No Hypotension
21 5666440246 822229 67 65-74 Female Yes Yes No Normal
Data View
Variable View
Linear...
IBM SPSS Statistics Processor is ready

Software -SPSS

- : www.spss.cz

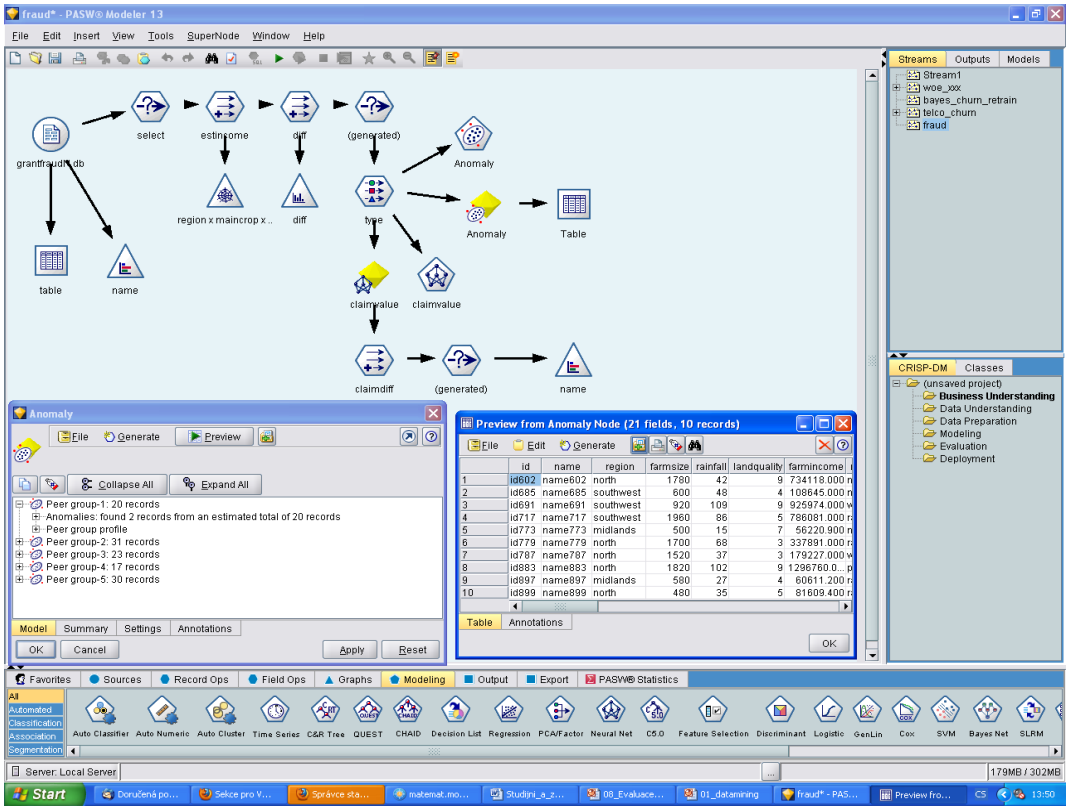


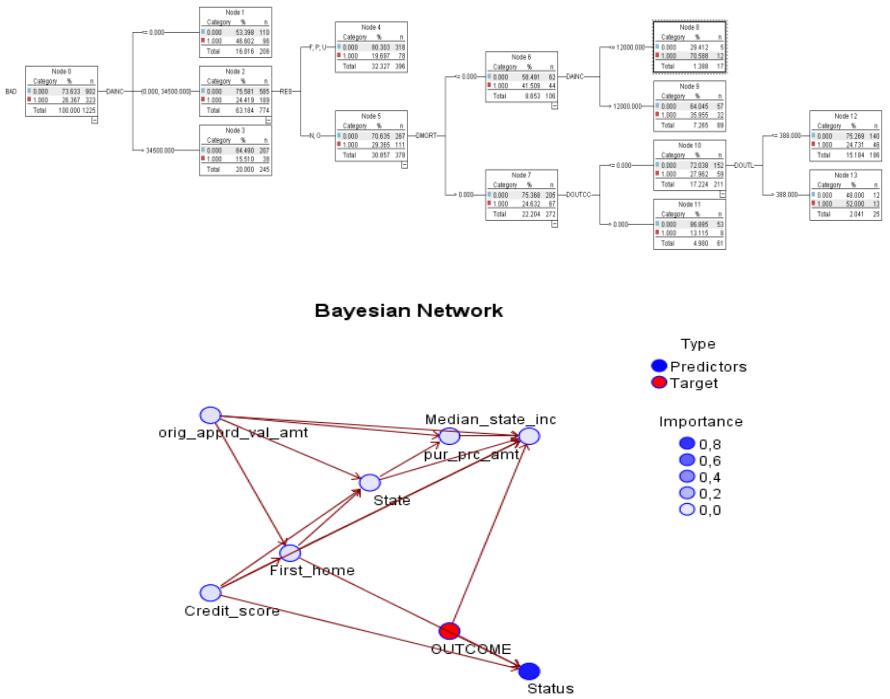


SPSS

- IBM SPSS/ PASW Modeler 13 (dříve Clementine)

  http://www.spss.cz/ibmspss_modeler.htm





SPSS

- Více o IBM SPSS Modeler 13 (dříve Clementine): http://www.spss.cz/ibmspss_modeler.htm
- (neúplný) seznam zákazníků: http://www.spss.cz/zakaznici.htm
- Akademický program: http://www.spss.com/academic/

SPSS

COST

- From $1000 to $12000 per license depending on license type.

CON

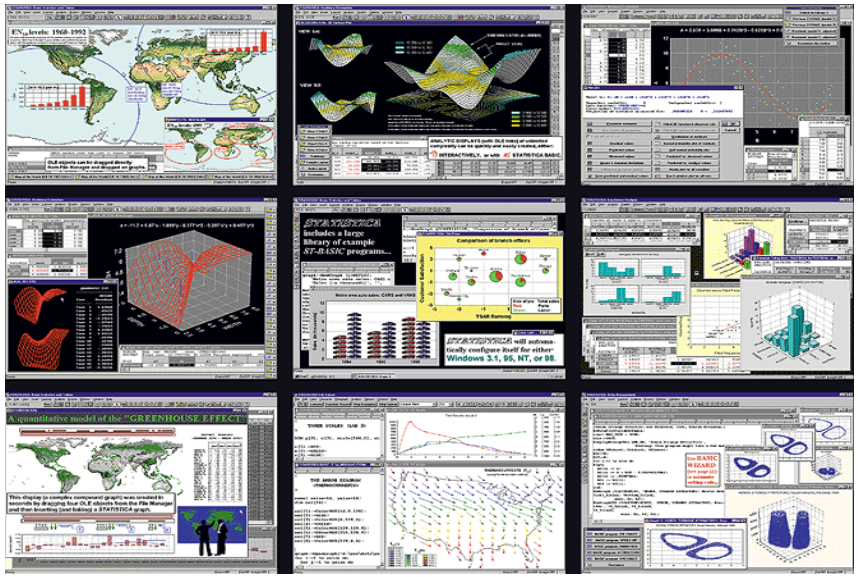
- Very expensive
- Not adequate for modeling and cutting edge statistical analysis

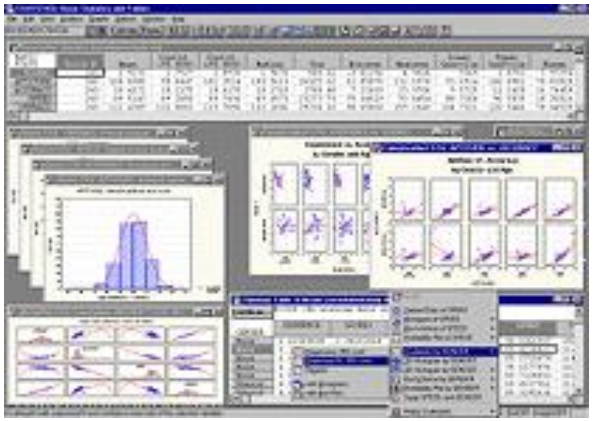
PRO

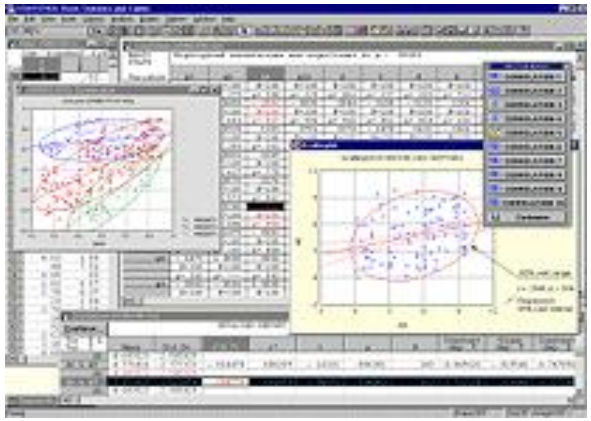
- Easy to learn and use
- More powerful then Minitab
- One of the most widely used statistical packages in academia and industry
- Has a command line interface in addition to menu driven user intefrace
- One of the most powerful statistical package that is also easy to use.

Software -Statistica

- StatSoft www.statistica.cz

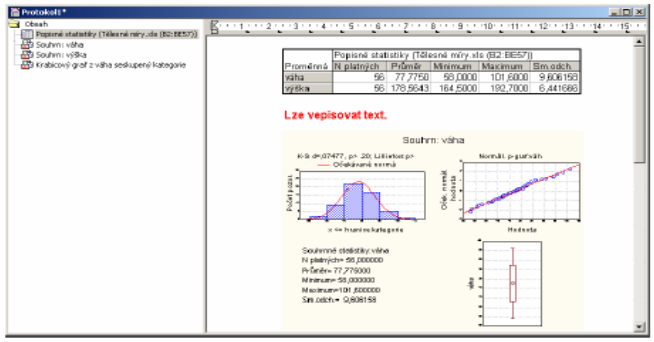




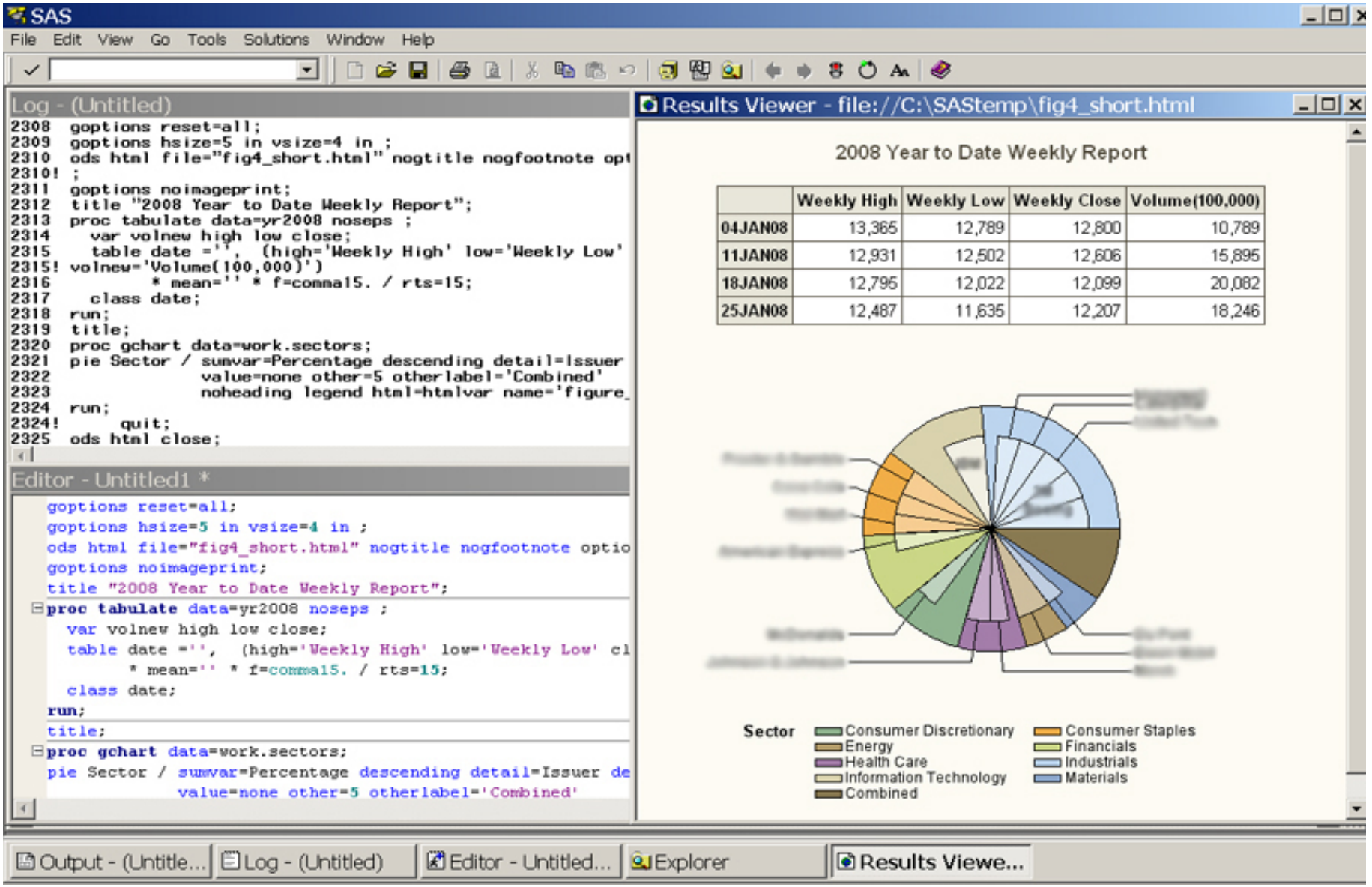


Statistica

- Více o Statistica Data Miner: http://www.statistica.cz/produkty/5-dataminingove-nastroje/21-statistica-data-miner/detail/
- (neúplný) seznam zákazníků: http://www.statsoft.com/customers/
- Akademický program: http://www.statsoft.com/academic/
- Petra Beranová – stručný manuál k ovládání programu STATISTICA: http://www.statsoft.cz/download/soubory/STATISTICA_manual.pdf

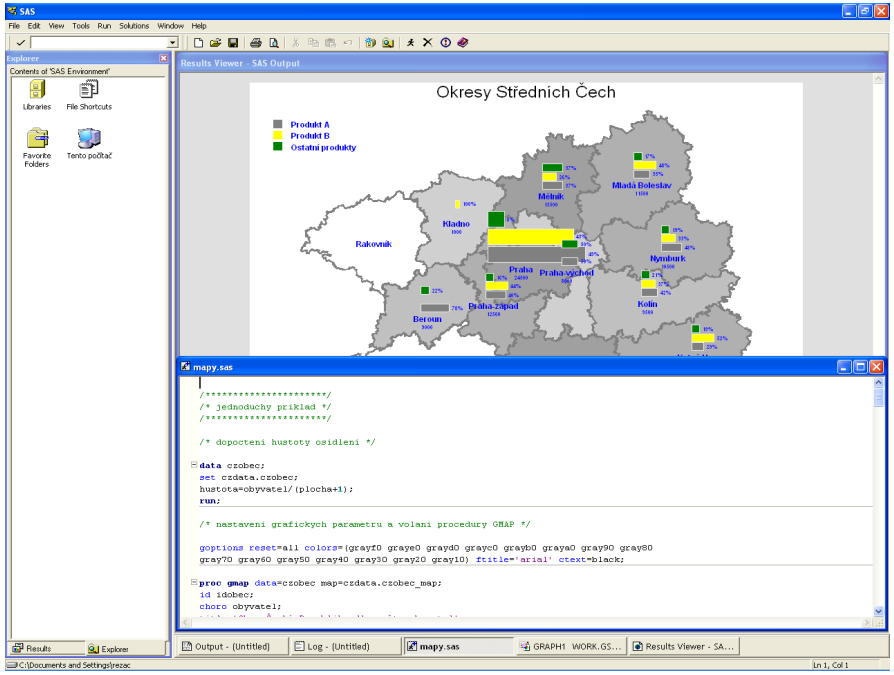


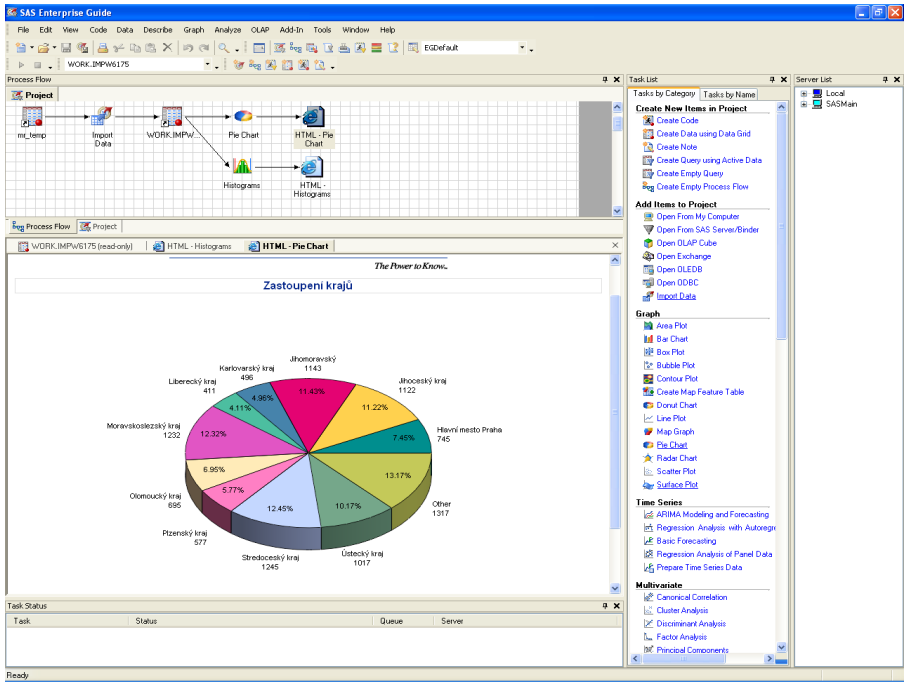
SAS


SAS
File Edit View Go Tools Solutions Window Help
Log - (Untitled)
2308 goptions reset=all;
2309 goptions hsize=5 in vsize=4 in ;
2310 ods html file="fig4_short.html" nogtitle nogfootnote opt
2310! ;
2311 goptions noimageprint;
2312 title "2008 Year to Date Weekly Report";
2313 proc tabulate data=yr2008 noseps ;
2314 var volnew high low close;
2315 table date ='', (high='Weekly High' low='Weekly Low'
2315! volnew='Volume(100,000)')
2316 * mean='' * f=comma15. / rts=15;
2317 class date;
2318 run;
2319 title;
2320 proc gchart data=work.sectors;
2321 pie Sector / sumvar=Percentage descending detail=Issuer
2322 value=none other=5 otherlabel='Combined'
2323 noheading legend html=htmlvar name='figure_
2324 run;
2324! quit;
2325 ods html close;
Editor - Untitled1 *
goptions reset=all;
goptions hsize=5 in vsize=4 in ;
ods html file="fig4_short.html" nogtitle nogfootnote optio
goptions noimageprint;
title "2008 Year to Date Weekly Report";
proc tabulate data=yr2008 noseps ;
var volnew high low close;
table date ='', (high='Weekly High' low='Weekly Low' cl
* mean='' * f=comma15. / rts=15;
class date;
run;
title;
proc gchart data=work.sectors;
pie Sector / sumvar=Percentage descending detail=Issuer de
value=none other=5 otherlabel='Combined'
Results Viewer - file://C:\SAStemp\fig4_short.html
2008 Year to Date Weekly Report
Weekly High Weekly Low Weekly Close Volume(100,000)
04JAN08 13,365 12,789 12,800 10,789
11JAN08 12,931 12,502 12,606 15,895
18JAN08 12,795 12,022 12,099 20,082
25JAN08 12,487 11,635 12,207 18,246
Sector
Consumer Discretionary
Consumer Staples
Energy
Financials
Health Care
Industrials
Information Technology
Materials
Combined
Output - (Untitle...
Log - (Untitled)
Editor - Untitled...
Explorer
Results Viewe...

Software - SAS

- : www.sas.com





SAS

- Společnost SAS Institute
  - Vznik 1976 v univerzitním prostředí
    - Dnes:největší soukromá softwarová společnost na světě (více než 11.000 zaměstnanců)
    - přes 45.000 instalací
    - cca 9 milionů uživatelů ve 118 zemích
    - v USA okolo 1.000 akademických zákazníků (SAS používá většina vyšších a vysokých škol a výzkumných pracovišť)

SAS

Soutěž o nejlepší studentskou práci

- lze přihlásit bakalářskou, diplomovou, dizertační, semestrální nebo ročníkovou práci využívající SAS.
- **1. místo** – letenky dle vlastního výběru v hodnotě 15.000 Kč.

SAS GLOBAL FORUM

2012 Orlando Florida
April 22-25, 2012

Ročník 2010:

- **1. místo - Účast na SAS Global Forum v Las Vegas.** Výherce měl hrazenou letenku, ubytování a účastnický poplatek.

http://www.sas.com/offices/europe/czech/academic/soutez.html
http://www.sas.com/offices/europe/czech/academic/poster.html

SAS

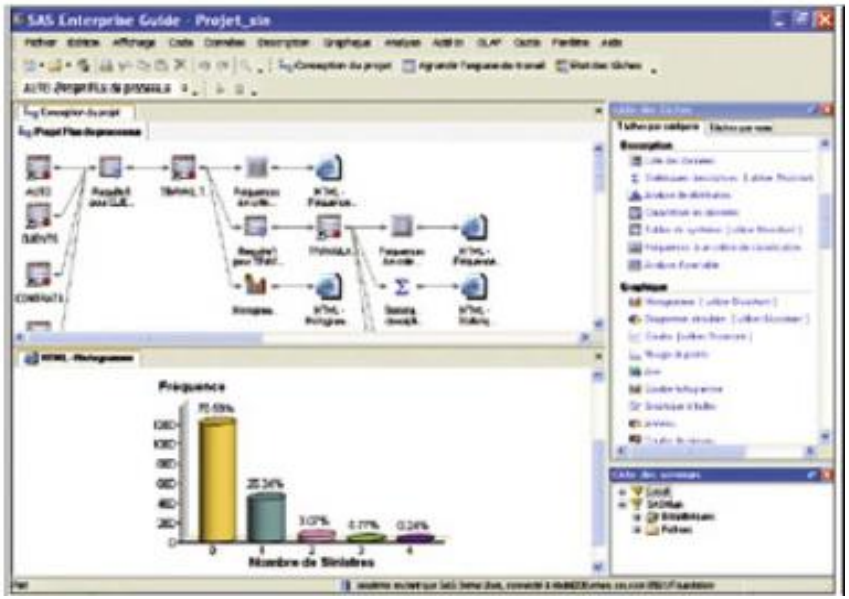
Podpora studentů

- Možnost rozšíření licence na domácí instalace pro studenty
- SAS Fellowship Program – software zdarma pro diplomku či dizertaci
- Zadávání a vedení diplomových prací
- Sdílení informací, zkušeností či příkladů v uživatelských skupinách

- Interaktivní moduly nebo programovací prostředí
  - Statistická analýza
  - Matice
  - Časové řady
  - Operační výzkum
  - Kontrola kvality

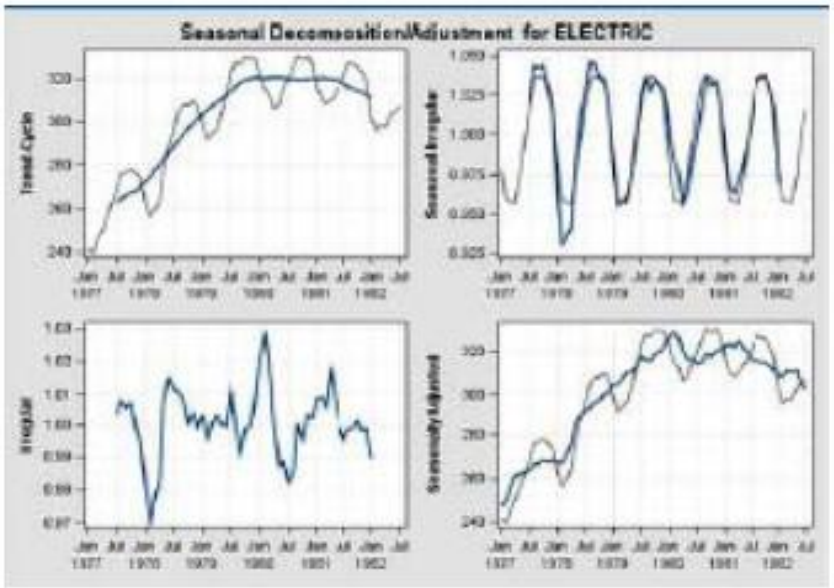
SAS

- Statistická analýza:
  - Popisná statistika
  - Analýza kontingenčních (frekvenčních) tabulek
  - Regresní, korelační, kovarianční analýza
  - Logistická regrese
  - Analýza rozptylu
  - Testování hypotéz
  - Diskriminační analýza
  - Shluková analýza
  - Analýza přežití
  - …



SAS

- Analýza časových řad:
  - Regresní modely
  - Modely se sezónními faktory
  - Autoregresní modely
  - ARIMA
  - Metody exponenciálního vyrovnání
  - …



SAS

- Více o SASu: http://www.sas.com/offices/europe/czech/

- (neúplný) seznam komerčních společností využívající SAS: http://www.sas.com/offices/europe/czech/reference/list.html

- o akademickém programu: http://www.sas.com/offices/europe/czech/academic/index.html

- o konferenci SAS forum: http://www.sas.com/reg/offer/cz/2010_sas_forum_2010 http://www.sas.com/reg/offer/cz/2011_sasforum

SAS

COST

- Complicated pricing model
- $8,500 first year license fee

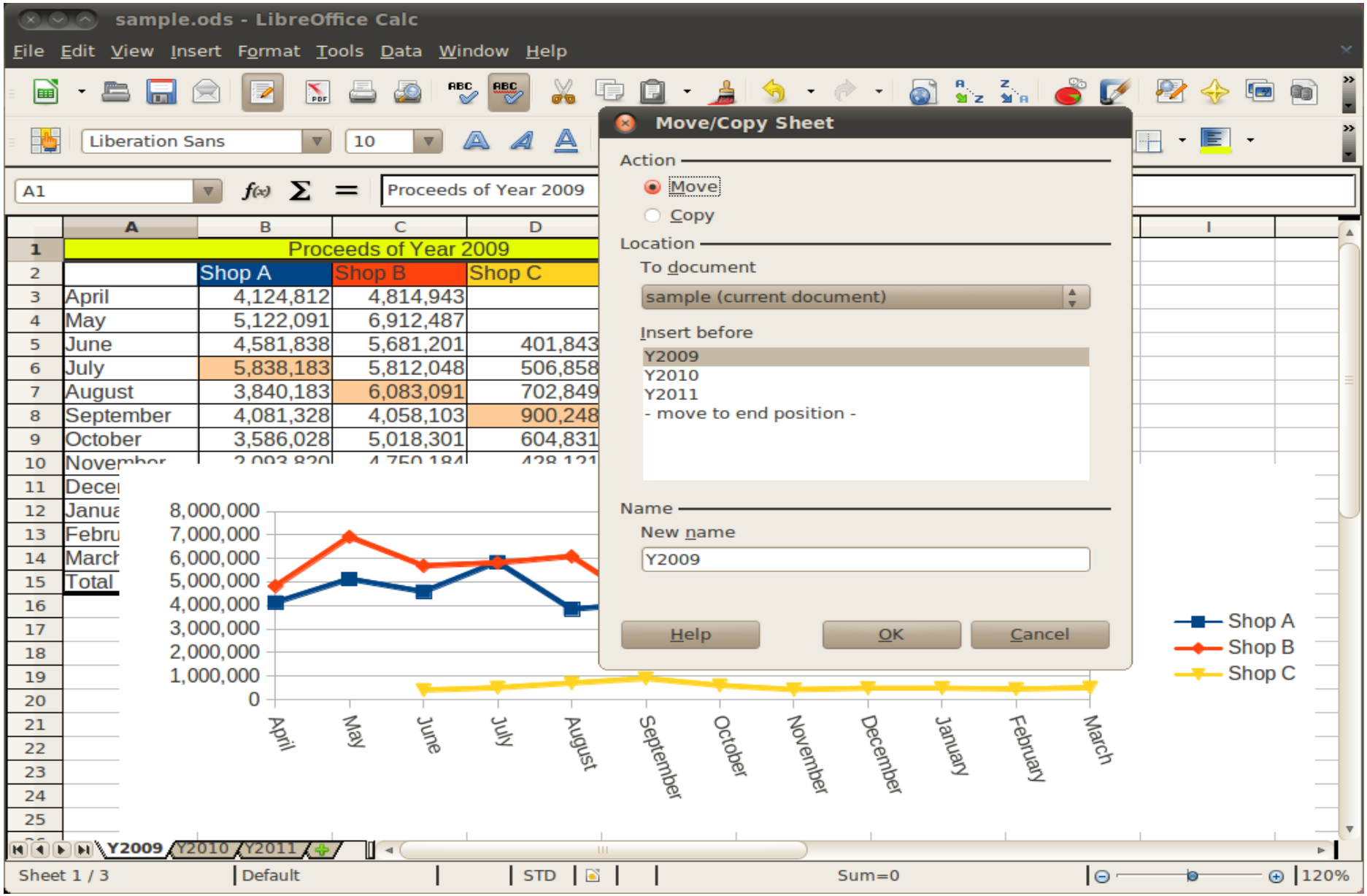
CON

- Very very expensive
- Not user friendly
- Steap learning curve
- Relatively poor graphics capabilities

PRO

- Widely accepted as the leader in statistical analysis and modeling
- Widely used in the industry and academia
- Very flexible and very powerful.

LibreOffice Calc



	A	B	C	D
1	Proceeds of Year 2009			
2		Shop A	Shop B	Shop C
3	April	4,124,812	4,814,943	
4	May	5,122,091	6,912,487	
5	June	4,581,838	5,681,201	401,843
6	July	5,838,183	5,812,048	506,858
7	August	3,840,183	6,083,091	702,849
8	September	4,081,328	4,058,103	900,248
9	October	3,586,028	5,018,301	604,831
10	Novembe	2,093,820	4,750,184	428,121
11	Decei			
12	Janua			
13	Febru			
14	March			
15	Total			

LibreOffice Calc

LibreOffice is a free and open source office suite, developed by The Document Foundation. It is descended from OpenOffice.org, from which it was forked in 2010

- OpenOffice vs LibreOffice
- Star → Sun → Oracle → Apache, Document Foundation
- OpenOffice
  http://www.openoffice.org/download
- LibreOffice
  http://www.libreoffice.org/download/

LibreOffice Calc

PRO

- Very similar to Microsoft Excel in functionality and look and feel (earlier versions)
- User friendly
- Very good for basic descriptive statistics, charts and plots
- Inter-operable with Microsoft Office

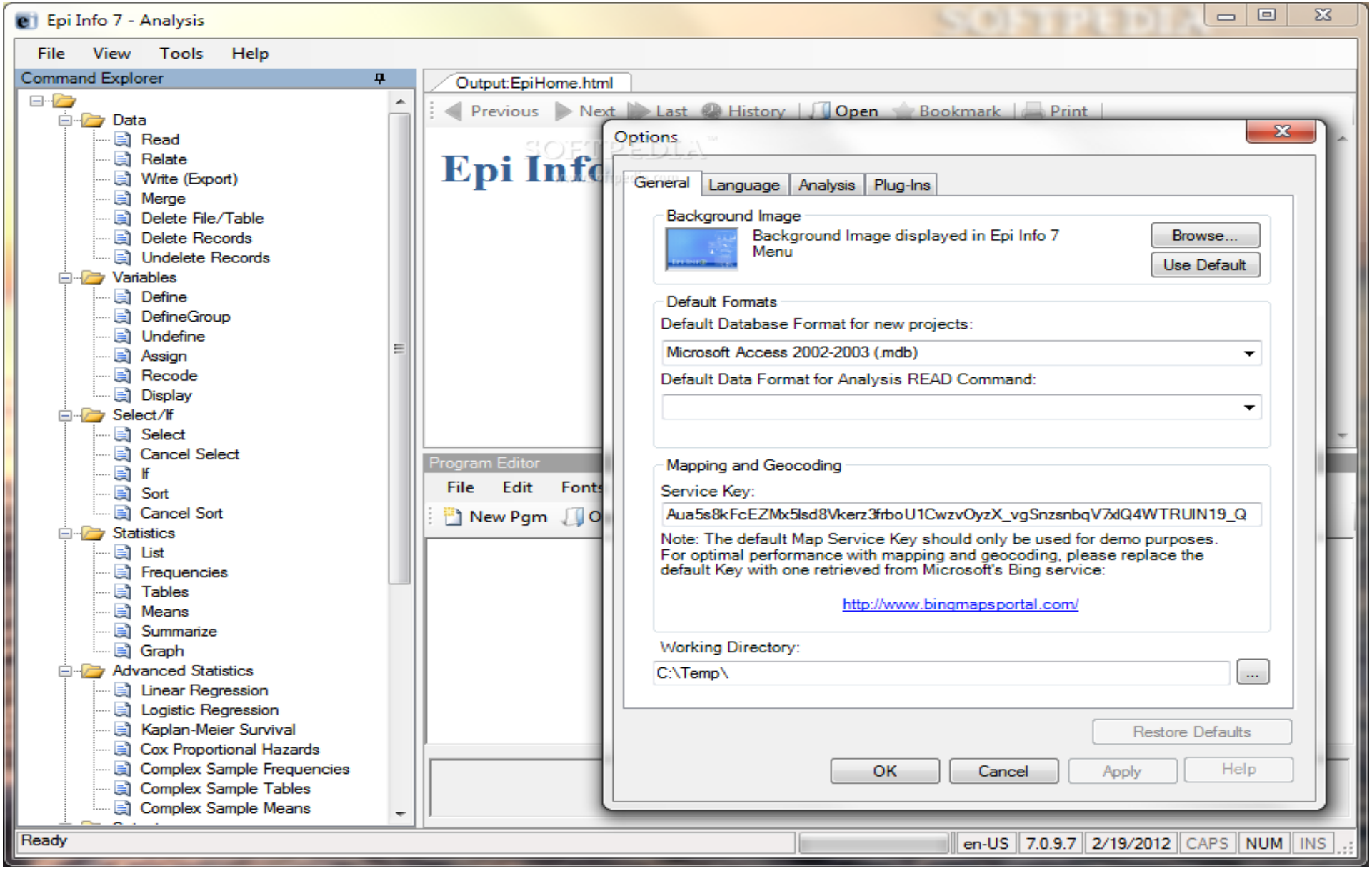
COST

- Free

CON

- Not sufficient for anything beyound the most basic statistical analysis

EpiInfo


Epi Info 7 - Analysis
File View Tools Help
Command Explorer
Data
Read
Relate
Write (Export)
Merge
Delete File/Table
Delete Records
Undelete Records
Variables
Define
DefineGroup
Undefine
Assign
Recode
Display
Select/If
Select
Cancel Select
If
Sort
Cancel Sort
Statistics
List
Frequencies
Tables
Means
Summarize
Graph
Advanced Statistics
Linear Regression
Logistic Regression
Kaplan-Meier Survival
Cox Proportional Hazards
Complex Sample Frequencies
Complex Sample Tables
Complex Sample Means
Output:EpiHome.html
Previous Next Last History Open Bookmark Print
Epi Info
Program Editor
File Edit Fonts
New Pgm
Options
General Language Analysis Plug-Ins
Background Image
Background Image displayed in Epi Info 7 Menu
Browse...
Use Default
Default Formats
Default Database Format for new projects:
Microsoft Access 2002-2003 (.mdb)
Default Data Format for Analysis READ Command:
Mapping and Geocoding
Service Key:
Aua5s8kFcEZMx5lsd8Vkerz3frboU1CwzvOyzX_vgSnzsnbqV7xlQ4WTRUIN19_Q
Note: The default Map Service Key should only be used for demo purposes. For optimal performance with mapping and geocoding, please replace the default Key with one retrieved from Microsoft's Bing service:
http://www.bingmapsportal.com/
Working Directory:
C:\Temp\
Restore Defaults
OK Cancel Apply Help
Ready
en-US 7.0.9.7 2/19/2012 CAPS NUM INS

EpiInfo

Epi Info is public domain statistical software for epidemiology developed by Centers for Disease Control and Prevention (CDC)

Epi Info has been in existence for over 20 years and is currently available for Microsoft Windows. The program allows for electronic survey creation, data entry, and analysis. Within the analysis module, analytic routines include t-tests, ANOVA, nonparametric statistics, cross tabulations and stratification with estimates of odds ratios, risk ratios, and risk differences, logistic regression (conditional and unconditional), survival analysis (Kaplan Meier and Cox proportional hazard), and analysis of complex survey data. The software is in the public domain, free, and can be downloaded from http://www.cdc.gov/epiinfo. Limited support is available

EpiInfo

PRO

- Consists of multiple modules to accomplish various tasks beyond just statistical analysis.
- ability to rapidly develop a questionnaire
- customize the data entry process
- quickly enter data into that questionnaire
- analyze the data

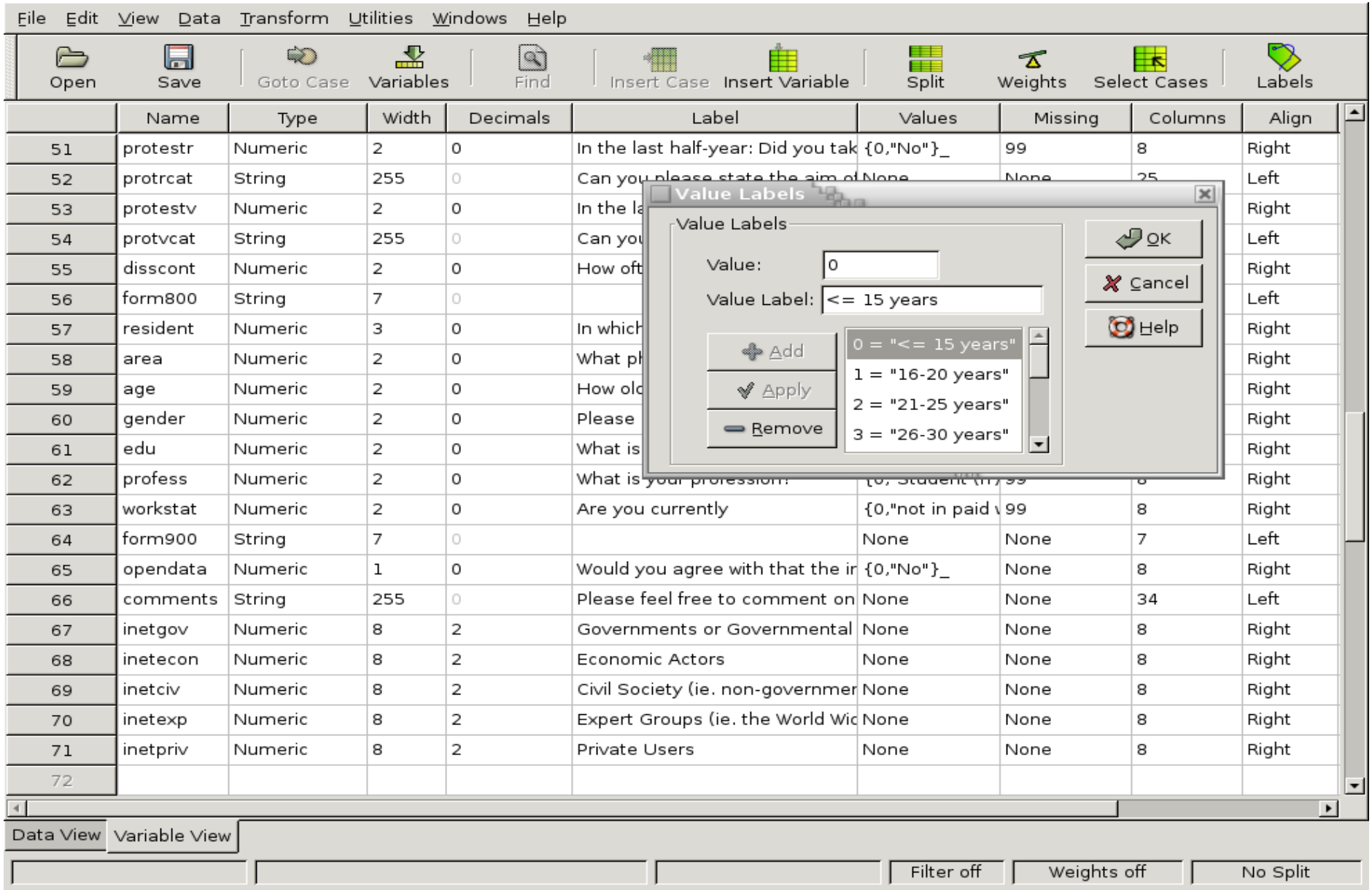
COST

- Free

CON

- Not a dedicated statistical package
- Not as powerful as commercial alternative for performing advanced analysis and modeling

PSPP



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align
51	protestr	Numeric	2	0	In the last half-year: Did you tak	{0,"No"}_	99	8	Right
52	protrcat	String	255	0	Can you please state the aim of	None	None	25	Left
53	protestv	Numeric	2	0	In the la				Right
54	protvcat	String	255	0	Can you				Left
55	disscont	Numeric	2	0	How oft				Right
56	form800	String	7	0					Left
57	resident	Numeric	3	0	In which				Right
58	area	Numeric	2	0	What pl				Right
59	age	Numeric	2	0	How old				Right
60	gender	Numeric	2	0	Please				Right
61	edu	Numeric	2	0	What is				Right
62	profess	Numeric	2	0	What is your profession?	{0, Student (n	99		Right
63	workstat	Numeric	2	0	Are you currently	{0,"not in paid v	99	8	Right
64	form900	String	7	0		None	None	7	Left
65	opendata	Numeric	1	0	Would you agree with that the ir	{0,"No"}_	None	8	Right
66	comments	String	255	0	Please feel free to comment on	None	None	34	Left
67	inetgov	Numeric	8	2	Governments or Governmental	None	None	8	Right
68	inetecon	Numeric	8	2	Economic Actors	None	None	8	Right
69	inetciv	Numeric	8	2	Civil Society (ie. non-governmer	None	None	8	Right
70	inetexp	Numeric	8	2	Expert Groups (ie. the World Wic	None	None	8	Right
71	inetpriv	Numeric	8	2	Private Users	None	None	8	Right
72									

PSPP

COST

- Free

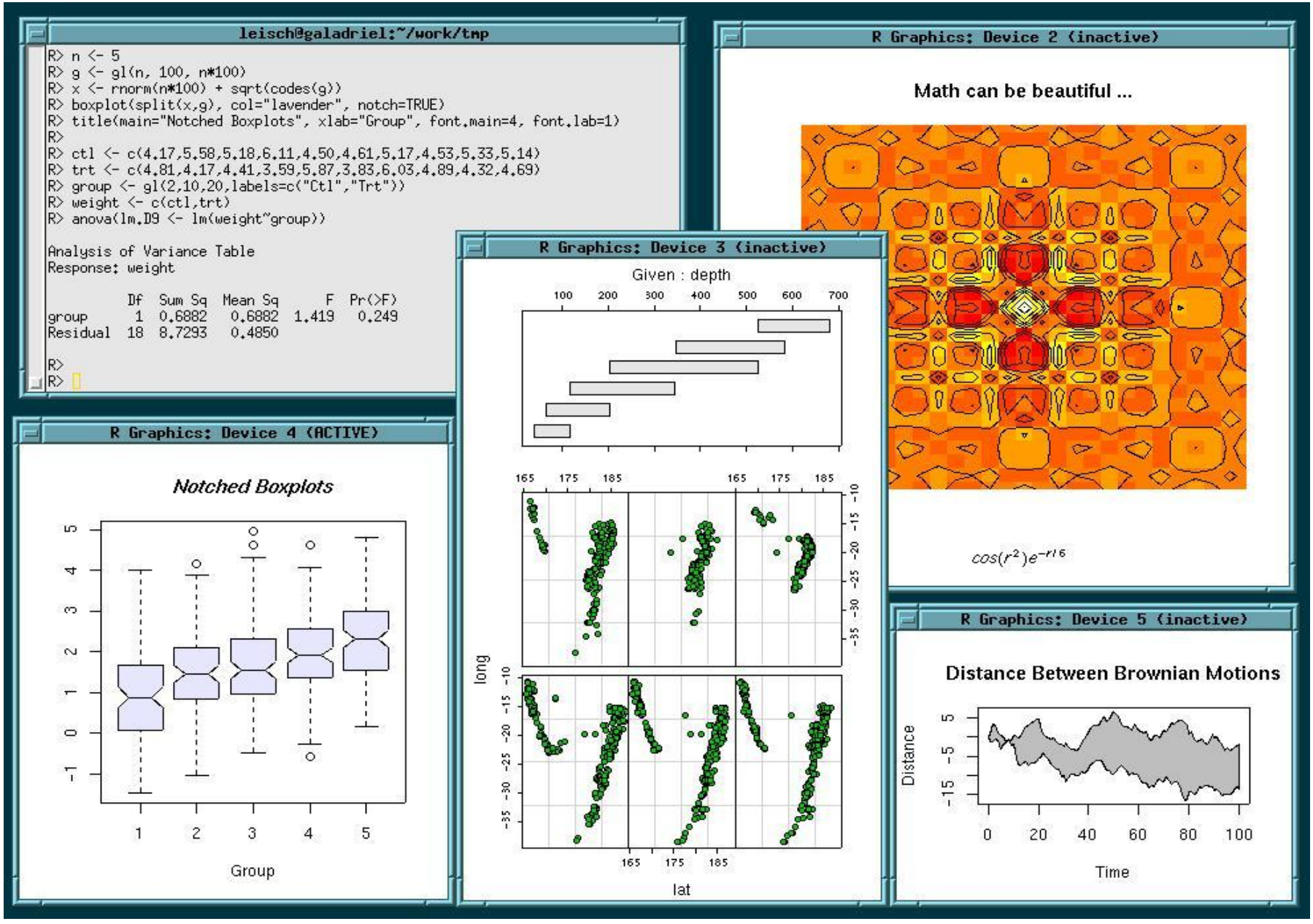
PRO

- Aims as a free SPSS alternative with an interface that closely resembles SPSS
- User friendly
- Good enough for basic statistical analysis

CON

- Lacks many advanced statistical tests and features that are present in SPSS
- Last version released in 2010
- Not very well known nor widely used

R


leisch@galadriel:~/work/tmp
R> n <- 5
R> g <- gl(n, 100, n*100)
R> x <- rnorm(n*100) + sqrt(codes(g))
R> boxplot(split(x,g), col="lavender", notch=TRUE)
R> title(main="Notched Boxplots", xlab="Group", font.main=4, font.lab=1)
R>
R> ctl <- c(4.17,5.58,5.18,6.11,4.50,4.61,5.17,4.53,5.33,5.14)
R> trt <- c(4.81,4.17,4.41,3.59,5.87,3.83,6.03,4.89,4.32,4.69)
R> group <- gl(2,10,20,labels=c("Ctl","Trt"))
R> weight <- c(ctl,trt)
R> anova(lm.D9 <- lm(weight~group))
Analysis of Variance Table
Response: weight
Df Sum Sq Mean Sq F Pr(>F)
group 1 0.6882 0.6882 1.419 0.249
Residual 18 8.7293 0.4850
R>
R>
R Graphics: Device 2 (inactive)
Math can be beautiful ...
$cos(r^2)e^{-r/6}$
R Graphics: Device 3 (inactive)
Given : depth
long
lat
R Graphics: Device 4 (ACTIVE)
Notched Boxplots
Group
R Graphics: Device 5 (inactive)
Distance Between Brownian Motions
Distance
Time

R

R provides a wide variety of statistical and graphical techniques, including linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, and others. R is easily extensible through functions and extensions, and the R community is noted for its active contributions in terms of packages. There are some important differences, but much code written for S runs unaltered. Many of R's standard functions are written in R itself, which makes it easy for users to follow the algorithmic choices made.

R is highly extensible through the use of user-submitted packages for specific functions or specific areas of study. Due to its S heritage, R has stronger object-oriented programming facilities than most statistical computing languages. Extending R is also eased by its permissive lexical scoping rules.[10]

According to Rexer's Annual Data Miner Survey in 2010, R has become the data mining tool used by more data miners (43%) than any other.[11]

Another strength of R is static graphics, which can produce publication-quality graphs, including mathematical symbols. Dynamic and interactive graphics are available through additional packages.[12]

R

PRO

- Widely used and accepted in industry and academia
- Very powerful and flexible
- Very large user base
- Lots of books and manuals
- Several User Interface Shells available

COST

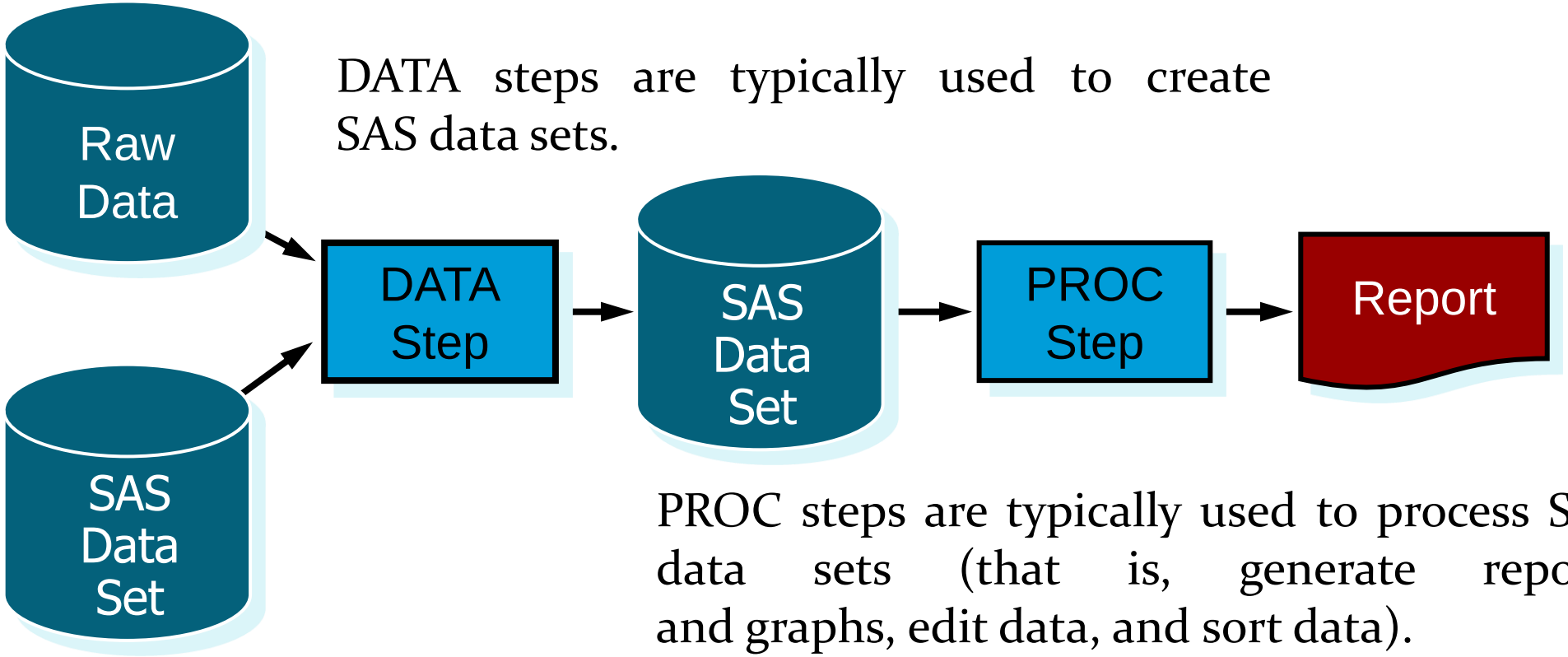
- Free / Open Source

CON

- Not user friendly
- Requires steep learning curve

SAS Programs

- A *SAS program* is a sequence of steps that the user submits for execution.



DATA steps are typically used to create SAS data sets.

PROC steps are typically used to process SAS data sets (that is, generate reports and graphs, edit data, and sort data).

SAS Programs

```
data work.clubmembers work.nonclub;
   set orion.customer;
   if Customer_Type_ID = 3010
      then output work.nonclub;
   else output work.clubmembers;
run;

proc print data=work.nonclub;
   title "Non Club Members";
   var Country Gender Customer_Name;
run;
```

DATA Step

PROC Step

Step Boundaries

SAS steps begin with either of the following:

- DATA statement
- PROC statement

SAS detects the end of a step when it encounters one of the following:

- a RUN statement (for most steps)
- a QUIT statement (for some procedures)
- the beginning of another step (DATA statement or PROC statement)

Step Boundaries

```
data work.clubmembers work.nonclub;
   set orion.customer;
   if Customer_Type_ID = 3010
      then output work.nonclub;
   else output work.clubmembers;
run;

proc print data=work.clubmembers;

proc print data=work.nonclub;
   title "Non Club Members";
   var Country Gender Customer_Name;
run;
```

Submitting a SAS Program

- When you execute a SAS program, the results generated by SAS are divided into two major parts:

SAS log contains information about the processing of the SAS program, including any warning and error messages.

SAS output contains reports generated by SAS procedures and DATA steps.

- The Workspace includes tabs containing both the log and output, while the Process Flow, by default, displays icons only for the output.

SAS Log

```
18          data work.clubmembers work.nonclub;
19             set orion.customer;
20             if Customer_Type_ID = 3010
21                then output work.nonclub;
22             else output work.clubmembers;
23          run;

NOTE: There were 77 observations read from the data set ORION.CUSTOMER.
NOTE: The data set WORK.CLUBMEMBERS has 69 observations and 12 variables.
NOTE: The data set WORK.NONCLUB has 8 observations and 12 variables.
NOTE: DATA statement used (Total process time):
      real time           0.06 seconds
      cpu time            0.00 seconds


24
25          proc print data=work.nonclub noobs;
26             title "Non Club Members";
27             var Country Gender Customer_Name;
28          run;

NOTE: There were 8 observations read from the data set WORK.NONCLUB.
NOTE: PROCEDURE PRINT used (Total process time):
      real time           0.09 seconds
      cpu time            0.00 seconds
```

PROC PRINT Output

sas | Enterprise Guide

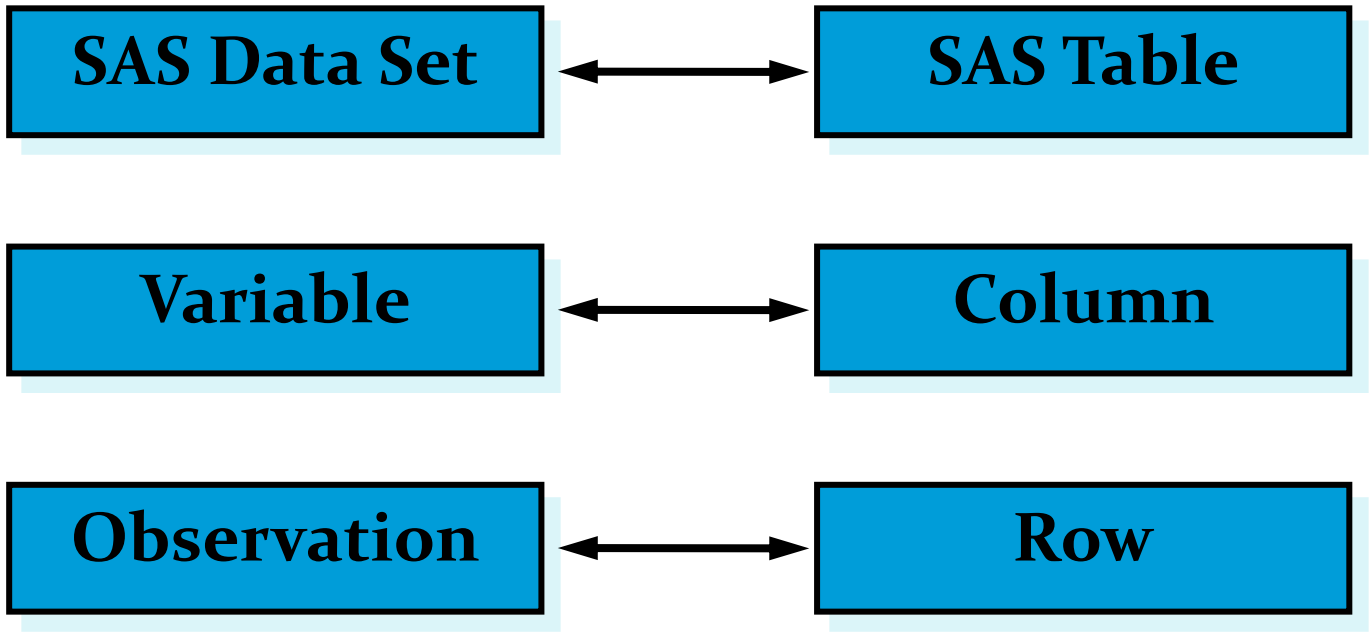
The Power to Know.

Non Club Members

Obs	Country	Gender	Customer_Name
1	DE	M	Ulrich Heyde
2	US	M	Tulio Devereaux
3	US	F	Robyn Klem
4	US	F	Cynthia Mccluney
5	AU	F	Candy Kinsey
6	US	M	Phenix Hill
7	IL	M	Avinoam Zweig
8	CA	F	Lauren Marx

SAS Terminology

• SAS documentation and text in the SAS windowing environment use the following terms interchangeably:



SAS Syntax Rules

SAS statements have these characteristics:

- usually begin with an **identifying keyword**
- always end with a **semicolon**

```
data work.clubmembers work.nonclub;
   set orion.customer;
   if Customer_Type_ID = 3010
      then output work.nonclub;
   else output work.clubmembers;
run;

proc print data=work.nonclub;
   title "Non Club Members";
   var Country Gender Customer_Name;
run;
```

SAS Syntax Rules

SAS statements are free-format.

- One or more blanks or special characters can be used to separate words.
- Statements can begin and end in any column.
- A single statement can span multiple lines.
- Several statements can be on the same line.

Unconventional Spacing

```
data work.clubmembers work.nonclub;
set orion.customer;
if Customer_Type_ID =      3010
           then output work.nonclub;
      else output work.clubmembers;run;
 proc print data=work.nonclub;              run;
```

SAS Syntax Rules

SAS statements are free-format.

- One or more blanks or special characters can be used to separate words.
- Statements can begin and end in any column.
- A single statement can span multiple lines.
- Several statements can be on the same line.

Unconventional Spacing

```
data work.clubmembers work.nonclub;
set orion.customer;
if Customer_Type_ID =       3010
           then output work.nonclub;
     else output work.clubmembers;run;
 proc print data=work.nonclub;          run;
```

SAS Syntax Rules

SAS statements are free-format.

- One or more blanks or special characters can be used to separate words.
- Statements can begin and end in any column.
- A single statement can span multiple lines.
- Several statements can be on the same line.

Unconventional Spacing

```
data work.clubmembers work.nonclub;
set orion.customer;
if Customer_Type_ID =      3010
           then output work.nonclub;
      else output work.clubmembers;run;
 proc print data=work.nonclub;          run;
```

SAS Syntax Rules

SAS statements are free-format.

- One or more blanks or special characters can be used to separate words.
- Statements can begin and end in any column.
- A single statement can span multiple lines.
- Several statements can be on the same line.

Unconventional Spacing

```
data work.clubmembers work.nonclub;
set orion.customer;
if Customer_Type_ID =      3010
          then output work.nonclub;
     else output work.clubmembers;run;
 proc print data=work.nonclub;              run;
```

SAS Syntax Rules

SAS statements are free-format.

- One or more blanks or special characters can be used to separate words.
- Statements can begin and end in any column.
- A single statement can span multiple lines.
- Several statements can be on the same line.

Unconventional Spacing

```
data work.clubmembers work.nonclub;
set orion.customer;
if Customer_Type_ID =      3010
           then output work.nonclub;
      else output work.clubmembers;run;
 proc print data=work.nonclub;              run;
```

SAS Comments

SAS comments consist of text that SAS ignores during processing. You can use comments anywhere in a SAS program to

- document the purpose of the program
- explain segments of the program
- mark SAS code as non-executing text.

Two methods of commenting are shown below:

```
/* comment */
```

```
* comment ;
```

SAS Comments: Examples

```
/* Split data based on membership */
data work.clubmembers work.nonclub;
   set orion.customer;
   if Customer_Type_ID = 3010
      then output work.nonclub;
   else output work.clubmembers;
run;
```

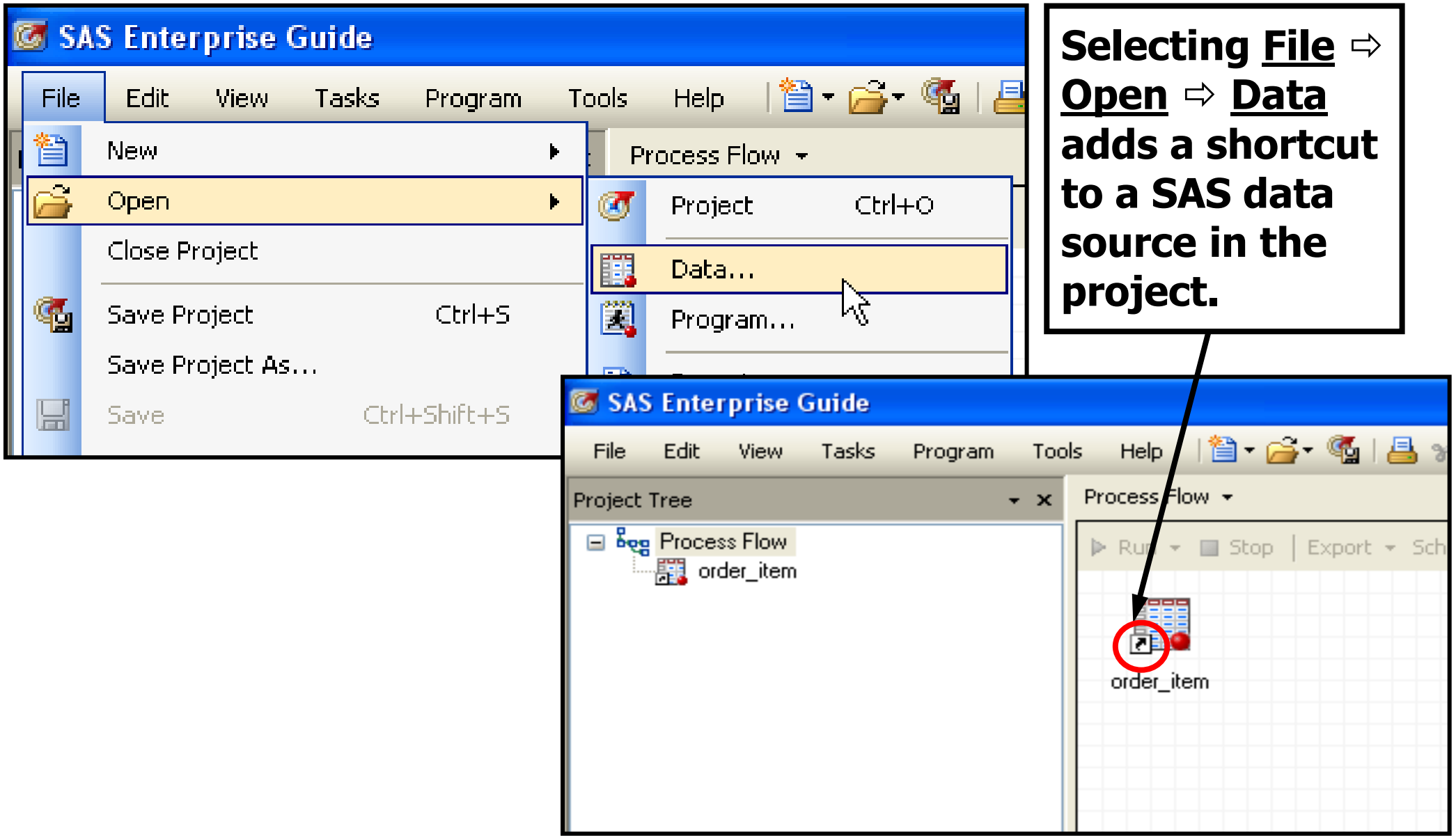
```
proc print data=work.nonclub;
   title "Non Club Members";
   *var Country Gender Customer_Name;
run;
```

Syntax Errors

- Syntax errors occur when program statements do not conform to the rules of the SAS language.
- Examples of syntax errors:
  - misspelled keywords
  - unmatched quotation marks
  - missing semicolons
  - invalid options
- When SAS encounters a syntax error, SAS prints a warning or an error message to the log.

```
ERROR 22-322: Syntax error, expecting one of the following:
              a name, a quoted string, (, /, ;, _DATA_, _LAST_,
              _NULL_.
```

How Do You Include Data in a Project?



How Do You Include Data in a Program?

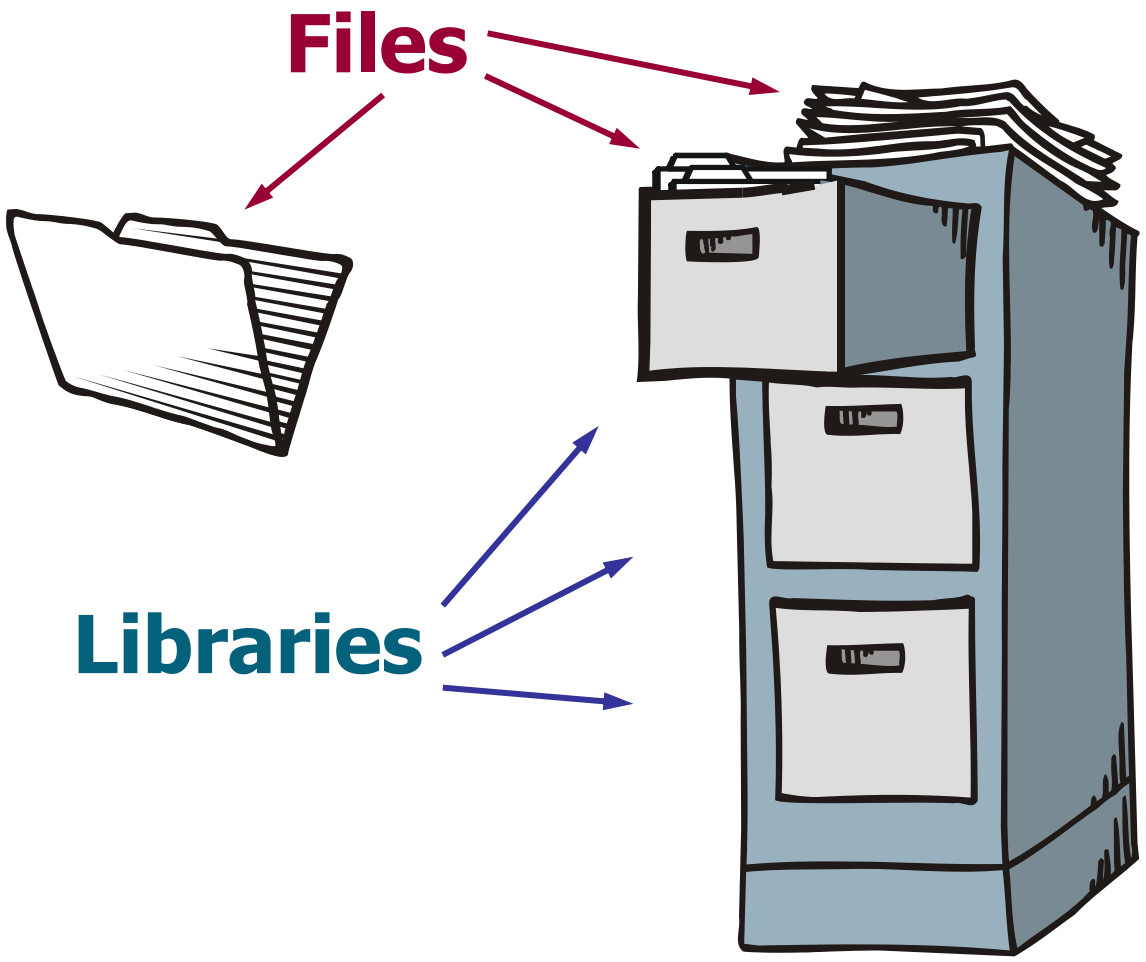
- One possibility is to include the full path and filename each time that a SAS data set is referenced.

```
data "s:\workshop\cust_age.sas7bdat";
   set "s:\workshop\customer.sas7bdat";
   /*Calculate each customer's age*/
   Age=int(yrdif(Birth_Date,today(),"actual"));
run;

proc print data="s:\workshop\cust_age.sas7bdat";
   var Customer_Name Gender Country Age;
   title "Customer Listing";
run;
```

ep02d03.sas

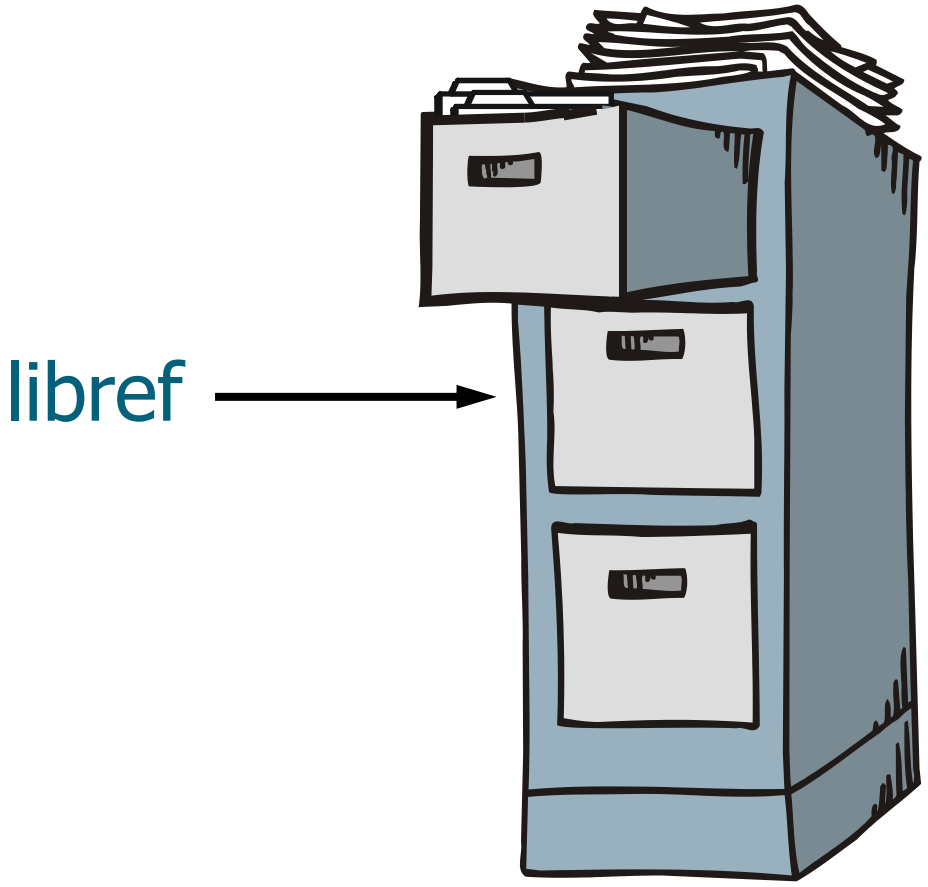
SAS Libraries

You can think of a SAS library as a drawer in a filing cabinet and a SAS data set as one of the file folders in the drawer.



Assigning a Libref

- Regardless of which host operating system you use, you identify SAS libraries by assigning a *library reference name (libref)* to each library.
- This libref can serve as a shortcut in SAS programs in place of the full path or filename.

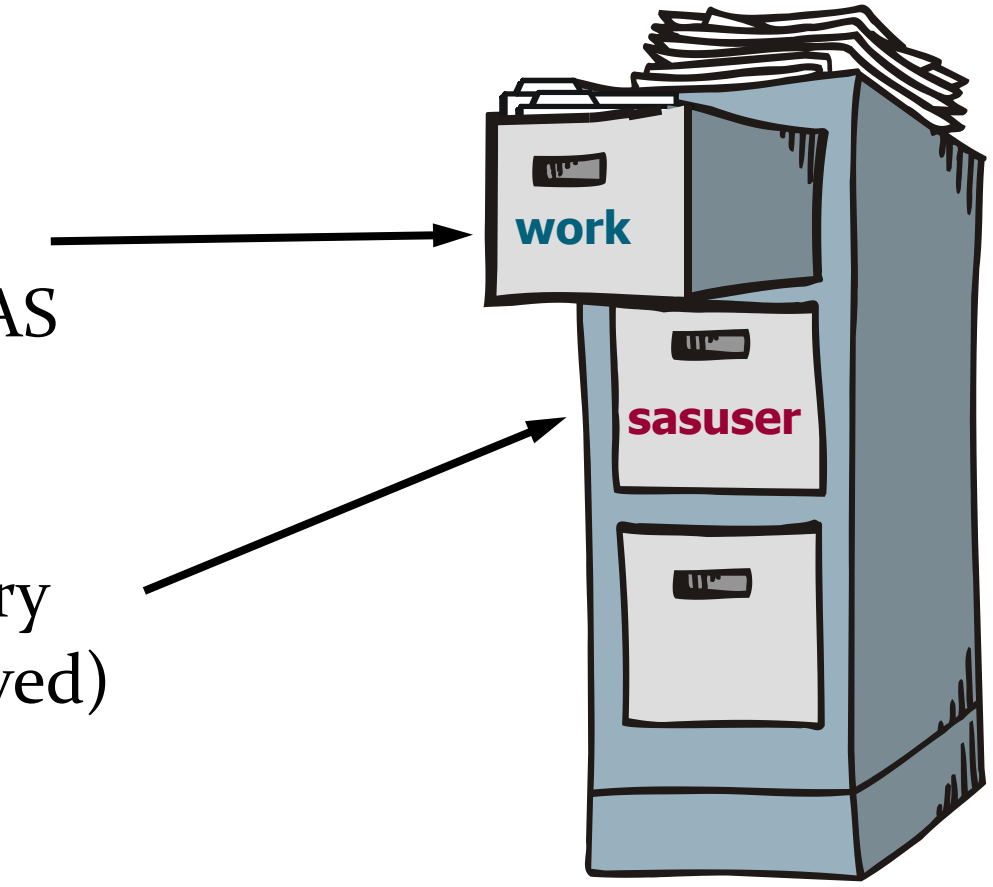


SAS Libraries

When a SAS session starts, SAS automatically creates one temporary and at least one permanent SAS library that you can access.

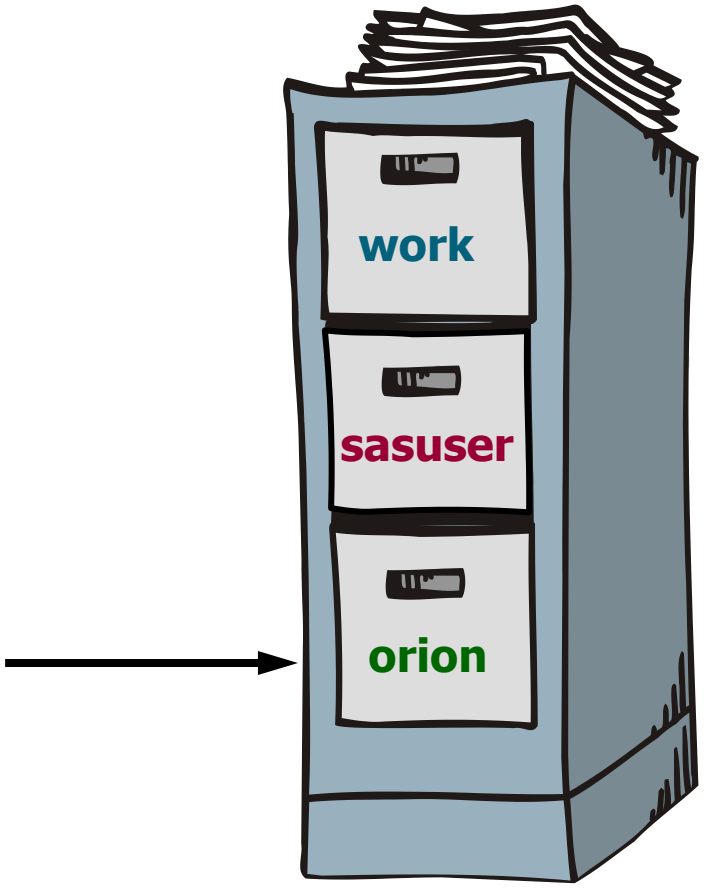
- **`work`** - temporary library (contents are deleted when SAS closes)

- **`sasuser`** - permanent library (contents are permanently saved)



SAS Libraries

- You can also create and access your own permanent libraries.



orion – permanent library

Assigning a Libref

- You can use the LIBNAME statement to assign a libref to a SAS library. The LIBNAME statement is a global statement.
- General form of the LIBNAME statement:

LIBNAME *libref* '*SAS-data-library*' <*options*>;

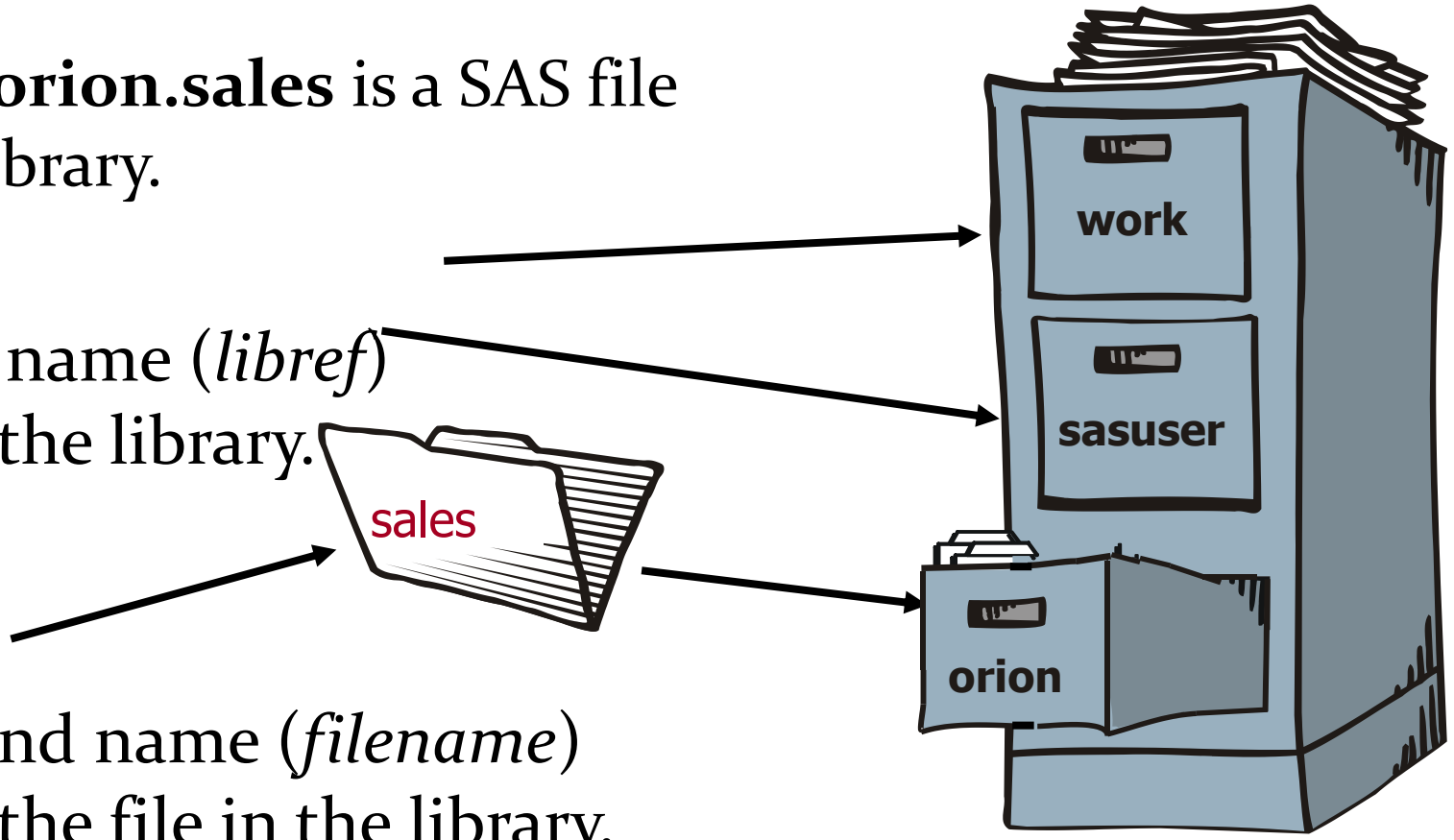
- The rules for naming a libref are as follows:
  - must be 8 or fewer characters
  - must begin with a letter or underscore
  - remaining characters are letters, numbers, or underscores

Two-Level SAS Filenames

- Every SAS file has a two-level name: *libref.filename*

- The data set **orion.sales** is a SAS file in the **orion** library.

  - The first name (*libref*) refers to the library.
  - The second name (*filename*) refers to the file in the library.



How Do You Include Data in a Program?

- využijeme knihovny (libraries)

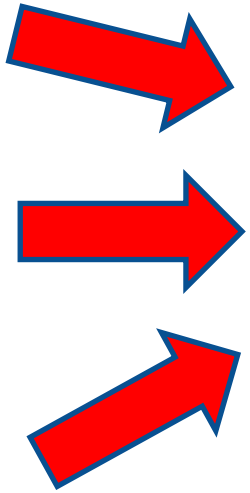
```
libname orion "s:\workshop";
data work.cust_age;
   set orion.customer;
   /*Calculate each customer's age*/
   Age=int(yrdif(Birth_Date,today(),"actual"));
run;

proc print data=work.cust_age;
   var Customer_Name Gender Country Age;
   title "Customer Listing";
run;
```

Temporary SAS Filename

- The default libref is **work** if the libref is omitted.


```
libname orion "s:\workshop";
data work.cust_age;
   set orion.customer;
   /*Calculate each customer's age*/
   Age=int(yrdif(Birth_Date,today(),"actual"));
run;

proc print data=cust_age;
   var Customer_Name Gender Country Age;
   title "Customer Listing";
run;
```

Import dat



***.sas7bdat**

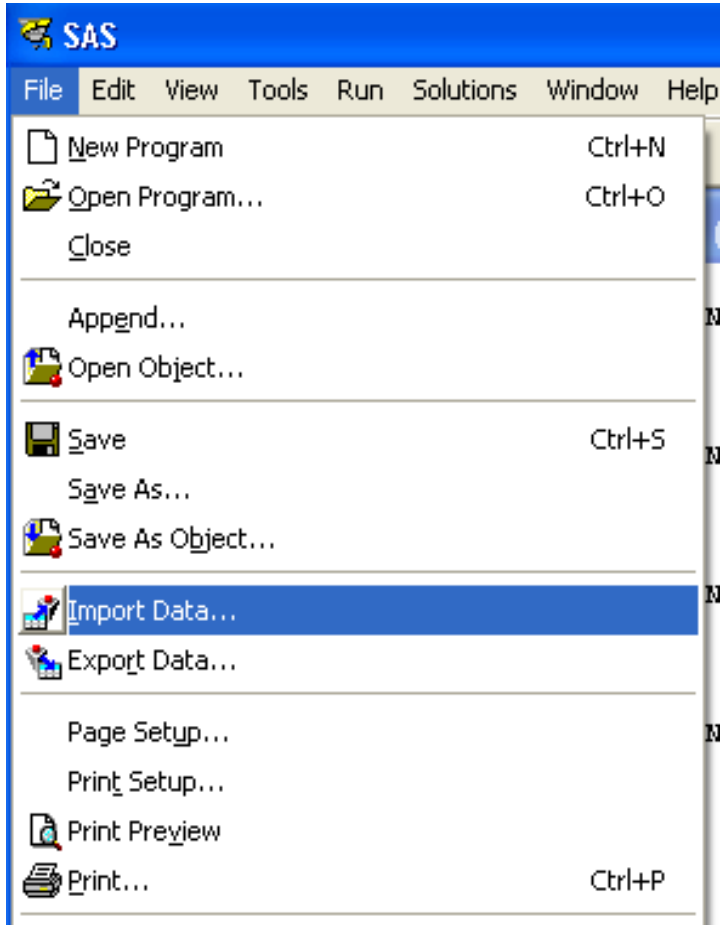
Základních pět možností importu dat:

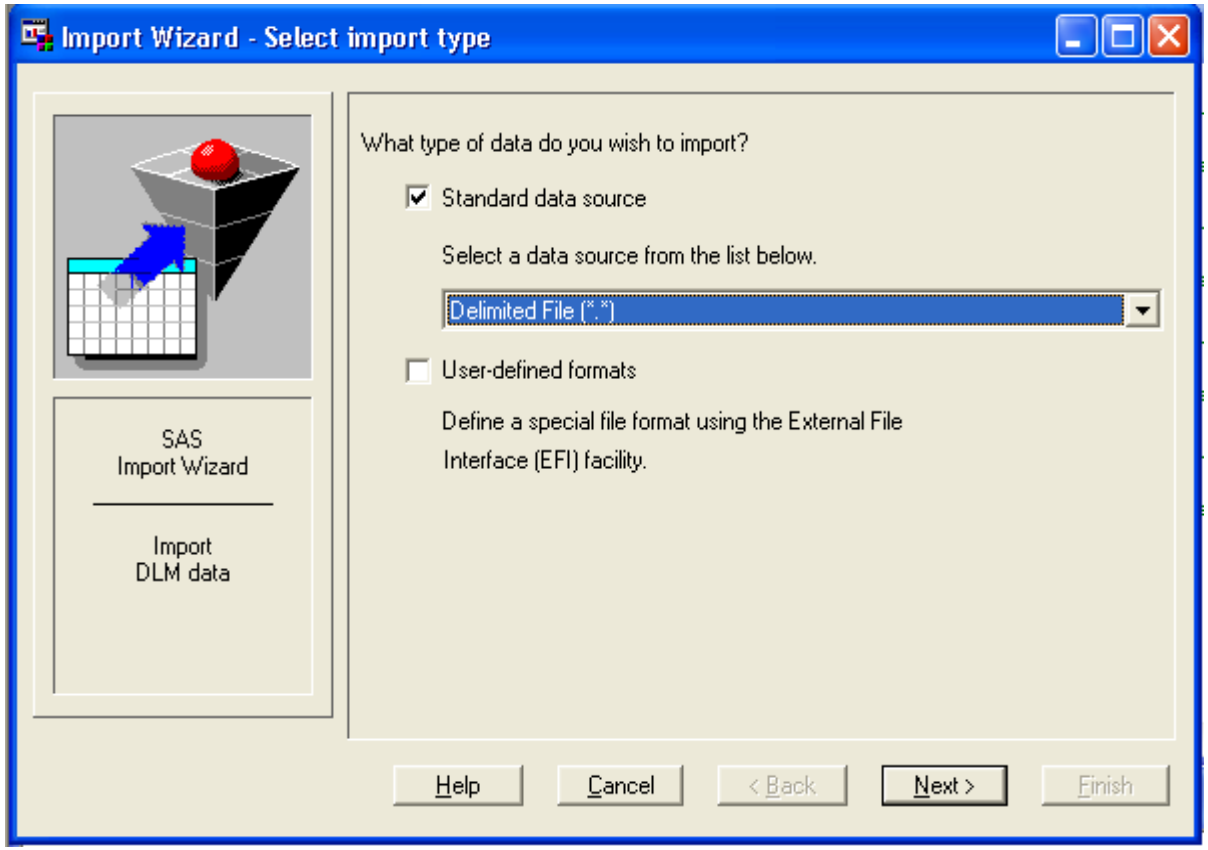
1. Import v SAS EG
2. Import wizard
3. Proc import
4. Data step
5. Proc SQL

Import Wizard

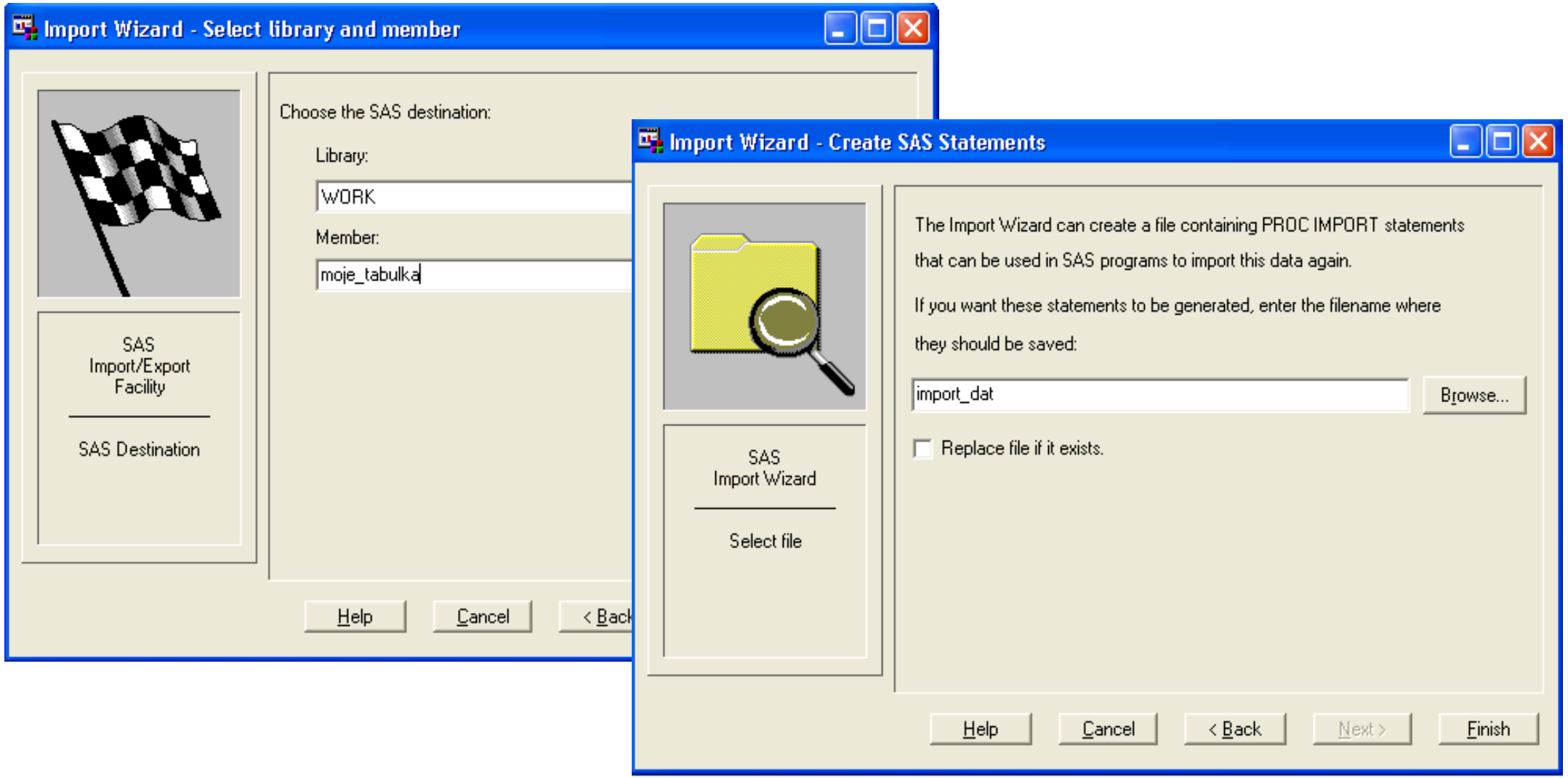
- The *Import Wizard* is a point-and-click graphical interface that enables you to create a SAS data set from several types of external files including the following:
  - dBASE files (*.DBF)
  - Excel spreadsheets (*.XLS)
  - Microsoft Access tables (.MDB)
  - delimited files (*.*)
  - comma-separated values (*.CSV)
  - …

Import Wizard





Import Wizard



PROC IMPORT

```
PROC IMPORT OUT= WORK.sales
            DATAFILE= "S:\Workshop\sales.xls"
            DBMS=EXCEL REPLACE;
     RANGE="Australia$";
     GETNAMES=YES;
     MIXED=NO;
     SCANTEXT=YES;
     USEDATE=YES;
     SCANTIME=YES;
RUN;
```

PROC IMPORT

GETNAMES=YES | NO

- determines whether SAS will use the first row of data in a Microsoft Excel worksheet or range as column names.

  YES specifies to use the first row of data in an Excel worksheet or range as column names.

  NO specifies **not** to use the first row of data in an Excel worksheet or range as column names. SAS generates and uses the variable names F1, F2, F3, and so on.

- The default is YES.

PROC IMPORT

MIXED=YES | NO

- specifies whether to import data with both character and numeric values and convert all data to character.

  YES specifies that all data values will be converted to character.

  NO specifies that numeric data will be missing when a character type is assigned. Character data will be missing when a numeric data type is assigned.

- The default is NO.

PROC IMPORT

SCANTEXT=YES | NO

specifies whether to read the entire data column and use the length of the longest string found as the SAS column width.

YES	scans the entire data column and uses the longest string value to determine the SAS column width.
NO	does not scan the column and defaults to a width of 255.

- The default is YES.

PROC IMPORT

SCANTIME=YES | NO

specifies whether to scan all row values in a date/time column and automatically determine the TIME data type if **only** time values exist.

YES specifies that a column with only time values be assigned the TIME8. format.

NO specifies that a column with only time values be assigned the DATE9. format.

- The default is NO.

PROC IMPORT

USEDATE=YES | NO

- specifies whether to use the DATE9. format for date/time values in Excel workbooks.

  YES specifies that date/time values be assigned the DATE9. format.

  NO specifies that date/time values be assigned the DATETIME16. format.

- The default is YES.

Proc import vs. Data step

```
PROC IMPORT OUT= WORK.MDATA1
            DATAFILE=
"G:\dukumenty\diplomka-data.txt"
            DBMS=CSV REPLACE;
     GETNAMES=YES;
     DATAROW=2;
RUN;
```

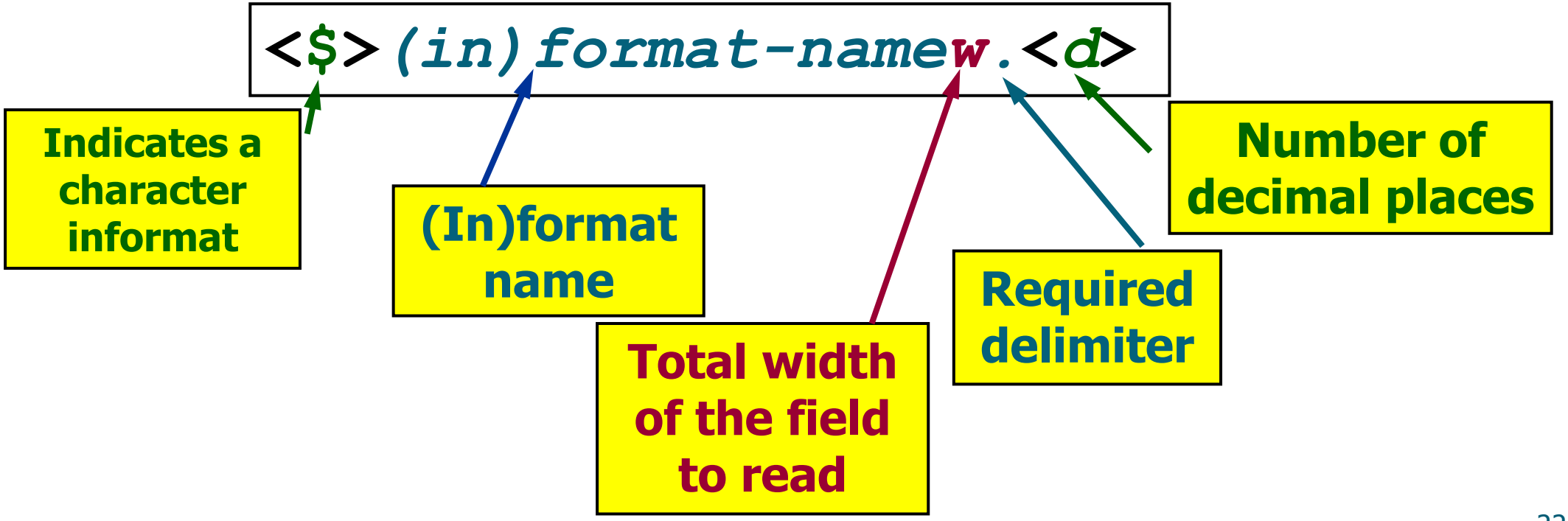
```
data work.mdata2;
length
BIRTHPLACE $ 25
AGE $ 25
.
.
.
EDUCATION $ 25
;
infile  'G:\dukumenty\diplomka-data.csv' delimiter = ';'
DSD lrecl=3276 firstobs=2 ;
input
BIRTHPLACE
AGE
.
.
.
EDUCATION
;
run;
```

Import z SQL databáze

```
libname my_data 'C:\Scoring\SASdata\';

proc sql;
connect to odbc as mssql (complete="DRIVER=SQL Server;
SERVER=sqlserv;Trusted_connection=Yes ");
create view my_data.wset_of_segments as select * from connection to mssql
(select * from db1.rezac.segmenty);
disconnect from mssql;
quit;

proc sql;
create table my_data.set_segments as
select
*
from my_data.wset_of_segments
;
quit;
```

Formats (Informats)

- An *informat* is an instruction that SAS uses to **read** data values.
- A *format* is an instruction that SAS uses to **write** data values.
- SAS (in)formats have the following form:



Formats (Informats)

InFormats by Category:

Category	***Description***
Character	instructs SAS to read character data values into character variables.
Column Binary	instructs SAS to read data stored in column-binary or multipunched form into character and numeric variables.
Date and Time	instructs SAS to read date values into variables that represent dates, times, and datetimes.
ISO 8601	instructs SAS to read date, time, and datetime values that are written in the ISO 8601 standard into either numeric or character variables.
Numeric	instructs SAS to read numeric data values into numeric variables.

http://support.sas.com/documentation/cdl/en/lrdict/64316/HTML/default/viewer.htm#a001239776.htm

Formats (Informats)

Formats by Category:

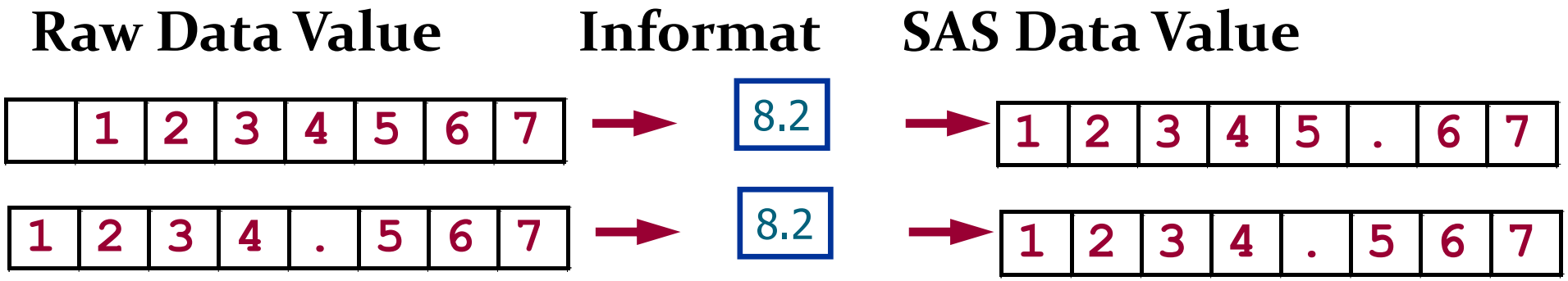
Category	*Description*
Character	instructs SAS to write character data values from character variables.
Date and Time	instructs SAS to write data values from variables that represent dates, times, and datetimes.
ISO 8601	instructs SAS to write date, time, and datetime values using the ISO 8601 standard.
Numeric	instructs SAS to write numeric data values from numeric variables.

http://support.sas.com/documentation/cdl/en/lrdict/64316/HTML/default/viewer.htm#a001263753.htm

Selected Informats

8. or 8.0 reads eight columns of numeric data.

8.2 reads eight columns of numeric data and **may** insert a decimal point in the value.



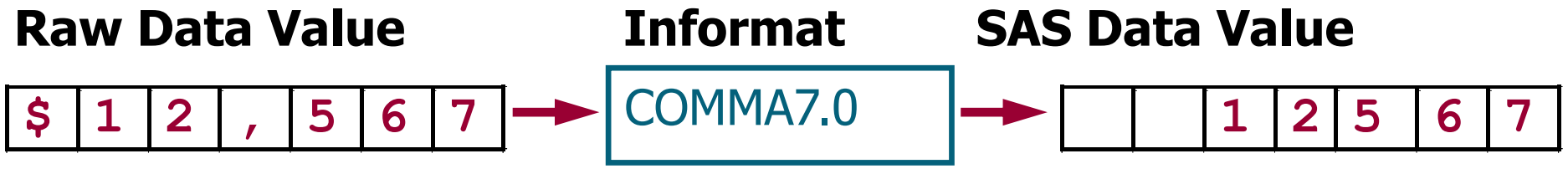
Selected Informats

$8.	reads eight columns of character data and removes leading blanks.

$CHAR8.	reads eight columns of character data and preserves leading blanks.

Selected Informats

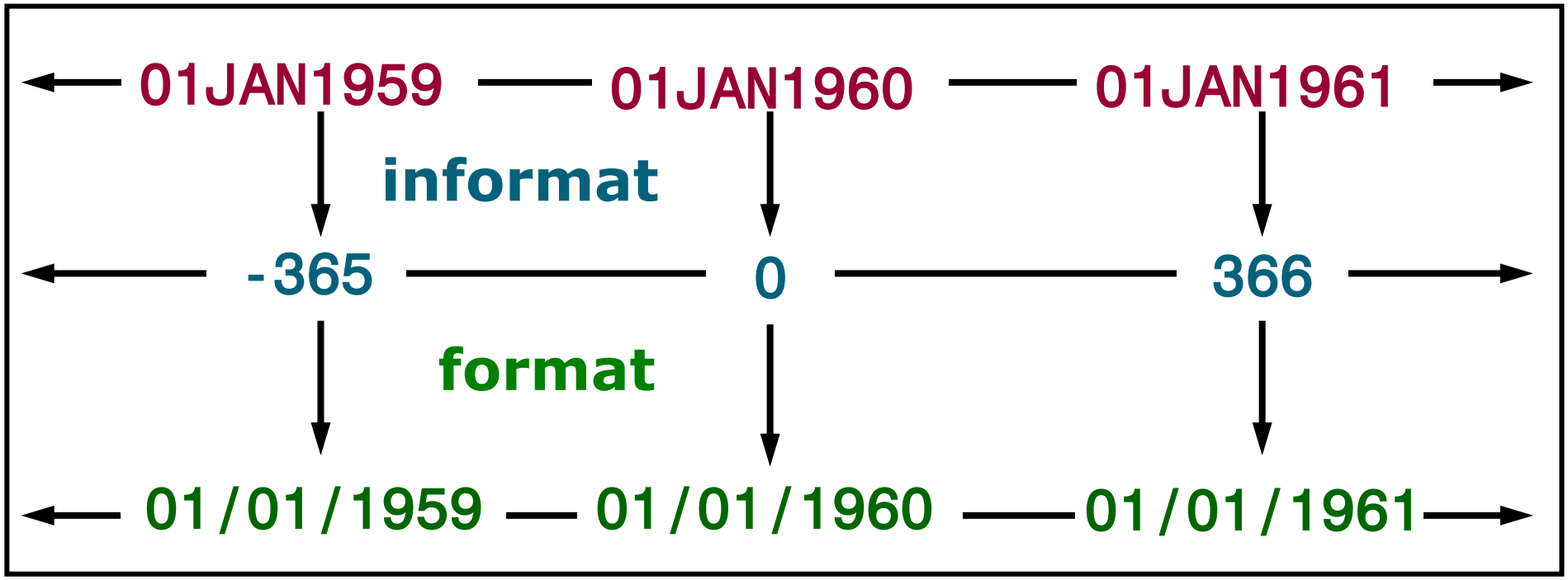
COMMA7.	reads seven columns of numeric data and removes selected nonnumeric characters such as dollar signs and commas.



MMDDYY8.	reads dates of the form 10/29/01.

Datumové formáty

- Date values that are stored as SAS dates are special numeric values.
- A *SAS date value* is interpreted as the number of days between January 1, 1960, and a specific date.



Datumové formáty

- SAS uses date **informats** to **read** and **convert** dates to SAS date values.

Examples:

Raw Data Value	Informat	Converted Value
10/29/2001	MMDDYY10.	15277
10/29/01	MMDDYY8.	15277
29OCT2001	DATE9.	15277
29/10/2001	DDMMYY10.	15277

Number of days between 01JAN1960 and 29OCT2001

Optimalizace práce s daty v SAS

- Pro (velmi) velké datové soubory je vhodné použití **komprese** a **indexování** SASovských tabulek. Více na:

http://www2.sas.com/proceedings/sugi27/p023-27.pdf

http://www2.sas.com/proceedings/sugi28/003-28.pdf

http://support.sas.com/documentation/cdl/en/lrdict/64316/HTML/default/viewer.htm#a001288760.htm

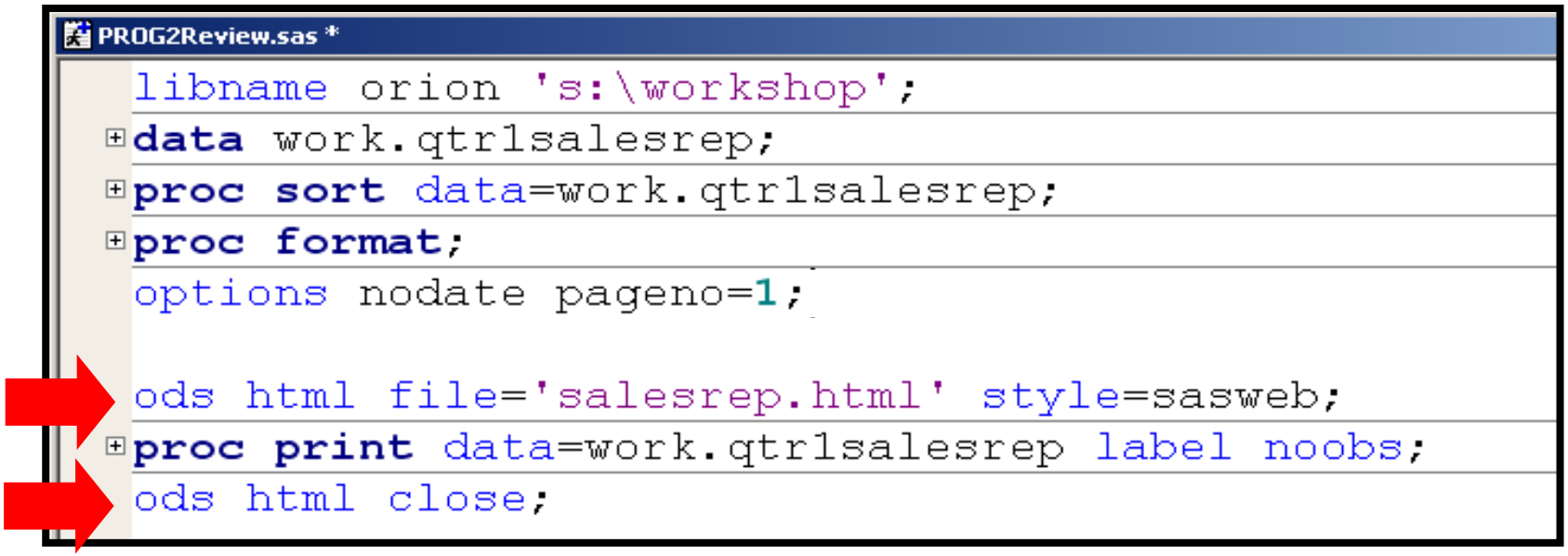
http://support.sas.com/documentation/cdl/en/lrdict/64316/HTML/default/viewer.htm#a000131138.htm

Příklad:

```
data lib1.tab2 (compress=binary index=(var1 var2));
set lib1.tab1;
...
run;
```

ODS – The Output Delivery System

- The Output Delivery System (ODS) enables you to produce output in a variety of formats, including HTML, RTF, PDF, and the default SAS listing.



- The ODS statements above create an HTML file, salesrep.html, using the output produced by the PROC PRINT step.

The PRINT Procedure

- The PRINT procedure prints the observations in a SAS data set and uses all or some of the variables.

PROG2Review.sas *

```
ods html file='salesrep.html' style=sasweb;
proc print data=work.qtr1salesrep label noobs;
   var Last_Name First_Name BonusMonth Bonus;
   title1 'Quarter 1 Orion Sales Reps';
   title2 'Males Only';
   footnote 'Confidential';
   format Bonus dollar8.;
   where Gender='M';
   by Country;
run;
ods html close;
```

- The PRINT procedure above includes TITLE and FOOTNOTE statements, which are global statements and do not need to be enclosed in a DATA or PROC step.

Program Output

Partial PROC PRINT Output
(SAS Output window)

```
                Quarter 1 Orion Sales Reps
                        Males Only

---------------------- Country=AU -----------------
------

                    First       Month of
     Last Name      Name          Bonus          Bonus

     Wills          Matsuoka          1          $300
     Surawski       Marinus           1          $300
     Shannan        Sian              1          $300
     Scordia        Randal            2          $300
     Pretorius      Tadashi           3          $300
     Nowd           Fadi              1          $300
     Magrath        Brett             1          $300
```

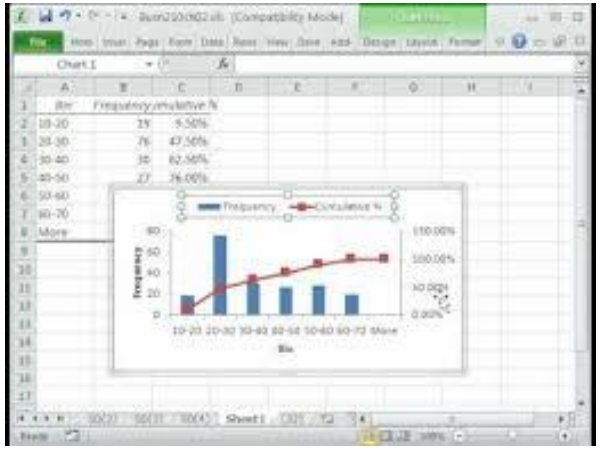
Partial PROC PRINT Output
(HTML format)

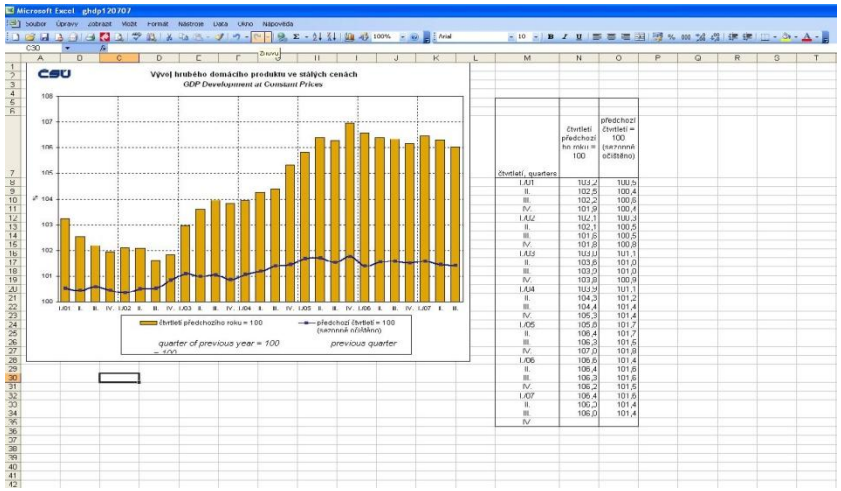
Quarter 1 Orion Sales Reps
Males Only

Country=AU

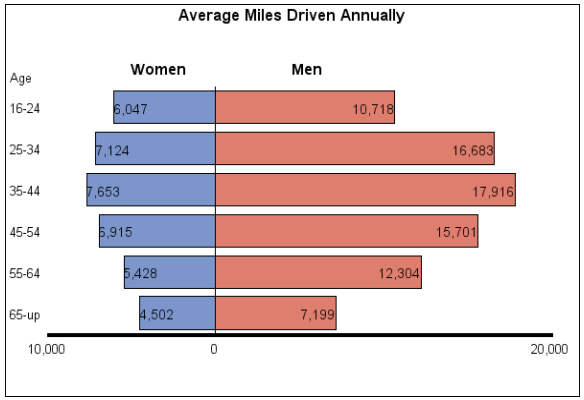
Last Name	First Name	Month of Bonus	Bonus
Wills	Matsuoka	1	$300
Surawski	Marinus	1	$300
Shannan	Sian	1	$300
Scordia	Randal	2	$300
Pretorius	Tadashi	3	$300
Nowd	Fadi	1	$300
Magrath	Brett	1	$300

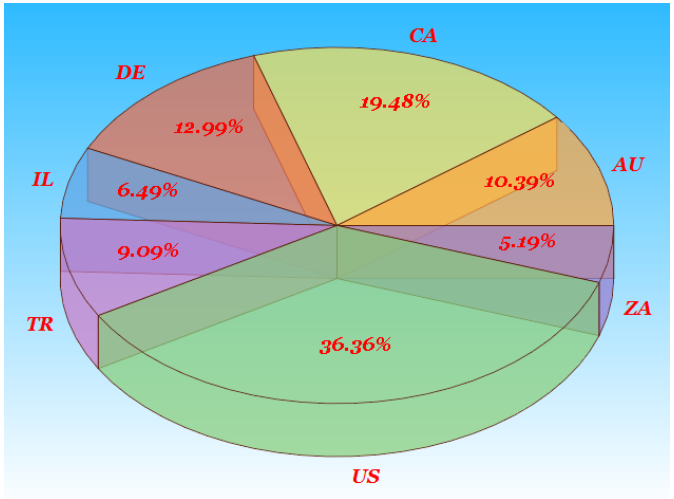
4. Popisná statistika v MS Excel a SAS





	Customer Gender							
	F				M			
	N	PctN	RowPctN	ColPctN	N	PctN	RowPctN	ColPctN
Customer Country								
AU	3	3.90	37.50	10.00	5	6.49	62.50	10.64
CA	8	10.39	53.33	26.67	7	9.09	46.67	14.89
DE	3	3.90	30.00	10.00	7	9.09	70.00	14.89
IL	.	.	.	.	5	6.49	100.00	10.64
TR	.	.	.	.	7	9.09	100.00	14.89
US	13	16.88	46.43	43.33	15	19.48	53.57	31.91
ZA	3	3.90	75.00	10.00	1	1.30	25.00	2.13





The FREQ Procedure

- The FREQ procedure can do the following:
  - produce one-way to *n*-way frequency and crosstabulation (contingency) tables
  - compute chi-square tests for one-way to *n*-way tables and measures of association and agreement for contingency tables
  - automatically display the output in a report and save the output in a SAS data set
- General form of the FREQ procedure:

```
PROC FREQ DATA=SAS-data-set <option(s)>;
      TABLES variable(s) </ option(s)>;
RUN;
```

- A FREQ procedure with **no TABLES statement** generates one-way frequency tables for **all data set variables**.

The TABLES Statement

A one-way frequency table produces frequencies, cumulative frequencies, percentages, and cumulative percentages.

```
proc freq data=orion.sales;
   tables Gender Country;
run;
```

one-way frequency tables

The FREQ Procedure

Gender	Frequency	Percent	Cumulative Frequency	Cumulative Percent
F	68	41.21	68	41.21
M	97	58.79	165	100.00

Country	Frequency	Percent	Cumulative Frequency	Cumulative Percent
AU	63	38.18	63	38.18
US	102	61.82	165	100.00

The TABLES Statement

An *n*-way frequency table produces cell frequencies, cell percentages, cell percentages of row frequencies, and cell percentages of column frequencies, plus total frequency and percent.

```
proc freq data=orion.sales;
   tables Gender*Country;
run;
```

rows

columns

two-way frequency table

The FREQ Procedure

Table of Gender by Country

Gender / Country Frequency Percent Row Pct Col Pct	AU	US	Total
F	27 16.36 39.71 42.86	41 24.85 60.29 40.20	68 41.21
M	36 21.82 37.11 57.14	61 36.97 62.89 59.80	97 58.79
Total	63 38.18	102 61.82	165 100.00

Additional SAS Statements

- Additional statements can be added to enhance the report.

```
proc format;
   value $ctryfmt 'AU'='Australia'
                  'US'='United
States';
run;

options nodate pageno=1;

ods html file='p112d01.html';
proc freq data=orion.sales;
   tables Gender*Country;
   where Job_Title contains 'Rep';
   format Country $ctryfmt.;
   title 'Sales Rep Frequency Report';
run;
ods html close;
```

Sales Rep Frequency Report

The FREQ Procedure

Frequency Percent Row Pct Col Pct	Table of Gender by Country		
	Country		
Gender	Australia	United States	Total
F	27 16.98 40.30 44.26	40 25.16 59.70 40.82	67 42.14
M	34 21.38 36.96 55.74	58 36.48 63.04 59.18	92 57.86
Total	61 38.36	98 61.64	159 100.00

Options to Suppress Display of Statistics

- Options can be placed in the TABLES statement after a forward slash to suppress the display of the default statistics.

Option	Description
NOCUM	suppresses the display of cumulative frequency and cumulative percentage.
NOPERCENT	suppresses the display of percentage, cumulative percentage, and total percentage.
NOFREQ	suppresses the display of the cell frequency and total frequency.
NOROW	suppresses the display of the row percentage.
NOCOL	suppresses the display of the column percentage.

Option	Description
LIST	displays *n*-way tables in list format.
CROSSLIST	displays *n*-way tables in column format.
FORMAT=	formats the frequencies in *n*-way tables.

LIST and CROSSLIST Options

Gender	Country	Frequency	Percent	Cumulative Frequency	Cumulative Percent
F	Australia	27	16.36	27	16.36
F	United States	41	24.85	68	41.21
M	Australia	36	21.82	104	63.03
M	United States	61	36.97	165	100.00

```
tables Gender*Country / list;
```

Table of Gender by Country

Gender	Country	Frequency	Percent	Row Percent	Column Percent
F	Australia	27	16.36	39.71	42.86
	United States	41	24.85	60.29	40.20
	Total	68	41.21	100.00	
M	Australia	36	21.82	37.11	57.14
	United States	61	36.97	62.89	59.80
	Total	97	58.79	100.00	
Total	Australia	63	38.18		100.00
	United States	102	61.82		100.00
	Total	165	100.00		

```
tables Gender*Country / crosslist;
```

PROC FREQ Statement Options

- Options can also be placed in the PROC FREQ statement.

Option	Description
NLEVELS	displays a table that provides the number of levels for each variable named in the TABLES statement.
PAGE	displays only one table per page.
COMPRESS	begins the display of the next one-way frequency table on the same page as the preceding one-way table if there is enough space to begin the table.

```
proc freq data=orion.sales nlevels;
   tables Gender Country Employee_ID;
run;
```

```
                The FREQ Procedure

        Number of Variable Levels

 Variable                   Levels
 ---------------------------------
 Gender                          2
 Country                         2
 Employee_ID                   165
```

Output Data Sets

- PROC FREQ produces output data sets using two different methods.
  - The TABLES statement with an OUT= option is used to create a data set with **frequencies and percentages**.

TABLES *variables* **/ OUT=***SAS-data-set* <*options*>**;**

  - The OUTPUT statement with an OUT= option is used to create a data set with **specified statistics** such as the chi-square statistic.

OUTPUT OUT=*SAS-data-set* <*options*>**;**

The MEANS Procedure

- The *MEANS procedure* provides data summarization tools to compute descriptive statistics for variables across all observations and within groups of observations.
- General form of the MEANS procedure:

```
PROC MEANS DATA=SAS-data-set <statistic(s)> <option(s)>;
      VAR analysis-variable(s);
      CLASS classification-variable(s);
RUN;
```

- By default, the MEANS procedure reports the number of nonmissing observations, the mean, the standard deviation, the minimum value, and the maximum value of all numeric variables.

```
proc means
data=orion.sales;
run;
```

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
Employee_ID	165	120713.90	450.0866939	120102.00	121145.00
Salary	165	31160.12	20082.67	22710.00	243190.00
Birth_Date	165	3622.58	5456.29	-5842.00	10490.00
Hire_Date	165	12054.28	4619.94	5114.00	17167.00

The VAR Statement

The *VAR statement* identifies the analysis variables and their order in the results.

```
proc means data=orion.sales;
   var Salary;
run;
```

The MEANS Procedure

Analysis Variable : Salary

N	Mean	Std Dev	Minimum	Maximum
165	31160.12	20082.67	22710.00	243190.00

The CLASS Statement

- The *CLASS statement* identifies variables whose values define subgroups for the analysis.

```
proc means data=orion.sales;
   var Salary;
   class Gender Country;
run;
```

The MEANS Procedure

Analysis Variable : Salary

Gender	Country	N Obs	N	Mean	Std Dev	Minimum	Maximum
F	AU	27	27	27702.41	1728.23	25185.00	30890.00
	US	41	41	29460.98	8847.03	25390.00	83505.00
M	AU	36	36	32001.39	16592.45	25745.00	108255.00
	US	61	61	33336.15	29592.69	22710.00	243190.00

The CLASS Statement

```
proc means data=orion.sales;
   var Salary;
   class Gender Country;
run;
```

classification variables

The MEANS Procedure

Analysis Variable : Salary

analysis variable

Gender	Country	N Obs	N	Mean	Std Dev	Minimum	Maximum
F	AU	27	27	27702.41	1728.23	25185.00	30890.00
	US	41	41	29460.98	8847.03	25390.00	83505.00
M	AU	36	36	32001.39			00
	US	61	61	33336.15	29592.69	22710.00	243190.00

statistics for analysis variable

The CLASS statement adds the N Obs column, which is the number of observations for each unique combination of the class variables.

PROC MEANS Statistics

- The statistics to compute and the order to display them can be specified in the PROC MEANS statement.

```
proc means data=orion.sales sum mean range;
   var Salary;
   class Country;
run;
```

The MEANS Procedure

Analysis Variable : Salary

Country	N Obs	Sum	Mean	Range
AU	63	1900015.00	30158.97	83070.00
US	102	3241405.00	31778.48	220480.00

- další dostupné statistiky:

Descriptive Statistic Keywords				
CLM	CSS	CV	LCLM	MAX
MEAN	MIN	MODE	N	NMISS
KURTOSIS	RANGE	SKEWNESS	STDDEV	STDERR
SUM	SUMWGT	UCLM	USS	VAR
Quantile Statistic Keywords				
MEDIAN | P50	P1	P5	P10	Q1 | P25
Q3 | P75	P90	P95	P99	QRANGE
Hypothesis Testing Keywords				
PROBT	T			

PROC MEANS Statement Options

- Options can also be placed in the PROC MEANS

Option	Description
MAXDEC=	specifies the number of decimal places to use in printing the statistics.
FW=	specifies the field width to use in displaying the statistics.
NONOBS	suppresses reporting the total number of observations for each unique combination of the class variables.

```
proc means data=orion.sales maxdec=0;
```

Analysis Variable : Salary

Country	N Obs	N	Mean	Std Dev	Minimum	Maximum
AU	63	63	30159	12699	25185	108255
US	102	102	31778	23556	22710	243190

```
proc means data=orion.sales maxdec=1;
```

Analysis Variable : Salary

Country	N Obs	N	Mean	Std Dev	Minimum	Maximum
AU	63	63	30159.0	12699.1	25185.0	108255.0
US	102	102	31778.5	23555.8	22710.0	243190.0

Output Data Sets

- PROC MEANS produces output data sets using the following method:

```
OUTPUT OUT=SAS-data-set <options>;
```

- The output data set contains the following variables:
  - BY variables
  - class variables
  - the automatic variables **_TYPE_** and **_FREQ_**
  - the variables requested in the OUTPUT statement

OUTPUT Statement OUT= Option

```
proc means data=orion.sales sum mean range;
   var Salary;
   class Gender Country;
   output out=work.means1;
run;

proc print data=work.means1;
run;
```

The statistics in the PROC statement impact only the MEANS report, not the data set.

Obs	Gender	Country	_TYPE_	_FREQ_	_STAT_	Salary
1			0	165	N	165.00
2			0	165	MIN	22710.00
3			0	165	MAX	243190.00
4			0	165	MEAN	31160.12
5			0	165	STD	20082.67
6		AU	1	63	N	63.00
7					MIN	25185.00
8					MAX	108255.00
9					MEAN	30158.97
10		AU	1	63	STD	12699.14
11		US	1	102	N	102.00
12		US	1	102	MIN	22710.00
13		US	1	102	MAX	243190.00
14		US	1	102	MEAN	31778.48
15		US	1	102	STD	23555.84
16	F		2	68	N	68.00
17	F		2	68	MIN	25185.00
18	F		2	68	MAX	83505.00
19	F		2	68	MEAN	28762.72
20	F		2	68	STD	6974.15

default statistics

OUTPUT Statement OUT= Option

- The OUTPUT statement can also do the following:
  - specify the statistics for the output data set
  - select and name variables

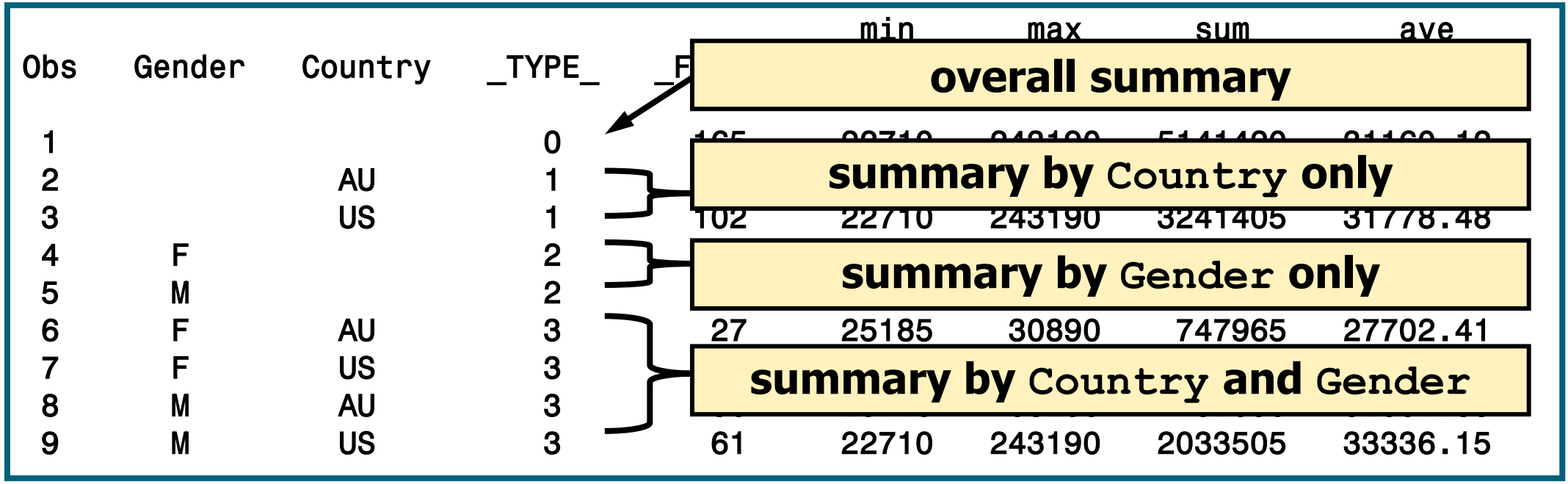
```
proc means data=orion.sales noprint;
   var Salary;
   class Gender Country;
   output out=work.means2
          min=minSalary max=maxSalary
          sum=sumSalary mean=aveSalary;
run;
proc print data=work.means2;run;
```

- The NOPRINT option suppresses the display of all output.

Obs	Gender	Country	_TYPE_	_FREQ_	min Salary	max Salary	sum Salary	ave Salary
1			0	165	22710	243190	5141420	31160.12
2		AU	1	63	25185	108255	1900015	30158.97
3		US	1	102	22710	243190	3241405	31778.48
4	F		2	68	25185	83505	1955865	28762.72
5	M		2	97	22710	243190	3185555	32840.77
6	F	AU	3	27	25185	30890	747965	27702.41
7	F	US	3	41	25390	83505	1207900	29460.98
8	M	AU	3	36	25745	108255	1152050	32001.39
9	M	US	3	61	22710	243190	2033505	33336.15

OUTPUT Statement OUT= Option

- **_TYPE_** is a numeric variable that shows which combination of class variables produced the summary statistics in that observation.



OUTPUT Statement OUT= Option

Obs	Gender	Country	_TYPE_	_FREQ_	min Salary	max Salary	sum Salary	ave Salary
1			0	165	22710	243190	5141420	31160.12
2		AU	1	63	25185	108255	1900015	30158.97
3		US	1	102	22710	243190	3241405	31778.48
4	F		2	68	25185	83505	1955865	28762.72
5	M		2	97	22710	243190	3185555	32840.77
6	F	AU	3	27	25185	30890	747965	27702.41
7	F	US	3	41	25390	83505	1207900	29460.98
8	M	AU	3	36	25745	108255	1152050	32001.39
9	M	US	3	61	22710	243190	2033505	33336.15

TYPE	Type of Summary	_FREQ_
0	overall summary	165
1	summary by **Country** only	63 AU + 102 AU = 165
2	summary by **Gender** only	68 F + 97 M = 165
3	summary by **Country** and **Gender**	27 F AU + 41 F US + 36 M AU + 61 M US = 165

OUTPUT Statement OUT= Option

- Options can be added to the PROC MEANS statement to control the output data set.

Option	Description
NWAY	specifies that the output data set contain only statistics for the observations with the highest **_TYPE_** value.
DESCENDTYPES	orders the output data set by descending **_TYPE_** value.
CHARTYPE	specifies that the **_TYPE_** variable in the output data set is a character representation of the binary value of **_TYPE_**.

without options

Obs	Gender	Country	_TYPE_	_FREQ_	min Salary	max Salary	sum Salary	ave Salary
1			0	165	22710	243190	5141420	31160.12
2		AU	1	63	25185	108255	1900015	30158.97
3		US	1	102	22710	243190	3241405	31778.48
4	F		2	68	25185	83505	1955865	28762.72
5	M		2	97	22710	243190	3185555	32840.77
6	F	AU	3	27	25185	30890	747965	27702.41
7	F	US	3	41	25390	83505	1207900	29460.98
8	M	AU	3	36	25745	108255	1152050	32001.39
9	M	US	3	61	22710	243190	2033505	33336.15

OUTPUT Statement OUT= Option

with NWAY

Obs	Gender	Country	_TYPE_	_FREQ_	min Salary	max Salary	sum Salary	ave Salary
1	F	AU	3	27	25185	30890	747965	27702.41
2	F	US	3	41	25390	83505	1207900	29460.98
3	M	AU	3	36	25745	108255	1152050	32001.39
4	M	US	3	61	22710	243190	2033505	33336.15

with DESCENDTYPES

Obs	Gender	Country	_TYPE_	_FREQ_	min Salary	max Salary	sum Salary	ave Salary
1	F	AU	3	27	25185	30890	747965	27702.41
2	F	US	3	41	25390	83505	1207900	29460.98
3	M	AU	3	36	25745	108255	1152050	32001.39
4	M	US	3	61	22710	243190	2033505	33336.15
5	F		2	68	25185	83505	1955865	28762.72
6	M		2	97	22710	243190	3185555	32840.77
7		AU	1	63	25185	108255	1900015	30158.97
8		US	1	102	22710	243190	3241405	31778.48
9			0	165	22710	243190	5141420	31160.12

with CHARTYPE

Obs	Gender	Country	_TYPE_	_FREQ_	min Salary	max Salary	sum Salary	ave Salary
1			00	165	22710	243190	5141420	31160.12
2		AU	01	63	25185	108255	1900015	30158.97
3		US	01	102	22710	243190	3241405	31778.48
4	F		10	68	25185	83505	1955865	28762.72
5	M		10	97	22710	243190	3185555	32840.77
6	F	AU	11	27	25185	30890	747965	27702.41
7	F	US	11	41	25390	83505	1207900	29460.98
8	M	AU	11	36	25745	108255	1152050	32001.39
9	M	US	11	61	22710	243190	2033505	33336.15

The SUMMARY Procedure

- The SUMMARY procedure provides data summarization tools to compute descriptive statistics for variables across all observations and within groups of observations.

General form of the SUMMARY procedure:

```
PROC SUMMARY DATA=SAS-data-set <statistic(s)>
                                <option(s)>;
     VAR analysis-variable(s);
     CLASS classification-variable(s);
RUN;
```

The SUMMARY Procedure

- The SUMMARY procedure uses the same syntax as the MEANS procedure.
- The only differences to the two procedures are the following:

PROC MEANS	PROC SUMMARY
The PRINT option is set by default, which displays output.	The NOPRINT option is set by default, which displays no output.
Omitting the VAR statement analyzes all the numeric variables.	Omitting the VAR statement produces a simple count of observations.

The TABULATE Procedure

- The TABULATE procedure displays descriptive statistics in tabular format.

General form of the TABULATE procedure:

```
PROC TABULATE DATA=SAS-data-set <options>;
      CLASS classification-variable(s);
      VAR analysis-variable(s);
      TABLE  page-expression,
                  row-expression,
                  column-expression </ option(s)>;
RUN;
```

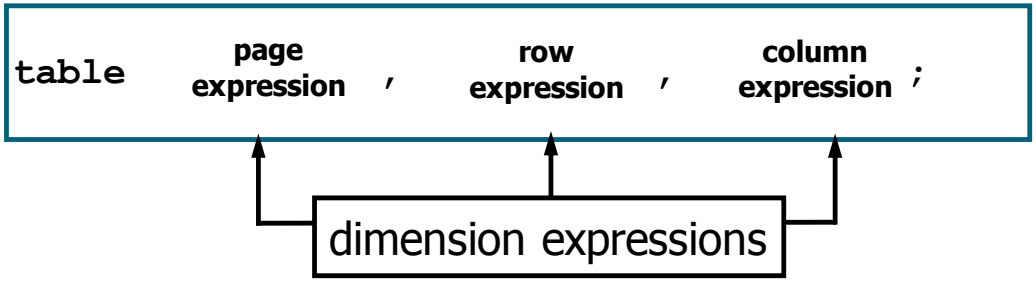
Dimensional Tables

- The TABULATE procedure produces one-, two-, or three-dimensional tables.

	page dimension	row dimension	column dimension
one-dimensional			✓
two-dimensional		✓	✓
three-dimensional	✓	✓	✓

The TABLE Statement

- The TABLE statement describes the structure of the table.



  - Commas separate the dimension expressions.
  - Every variable that is part of a dimension expression must be specified as a classification variable (CLASS statement) or an analysis variable (VAR statement).

```
table   page expression , row expression , column expression ;
```

- Příklady:

```
table Country;
```

```
table Gender , Country;
```

```
table Job_Title , Gender , Country;
```

The CLASS Statement

- The CLASS statement identifies variables to be used as classification, or grouping, variables.
- General form of the CLASS statement:

CLASS *classification-variable(s)*;

- N, the number of nonmissing values, is the default statistic for classification variables.
- Examples of classification variables:

  `Job_Title`, **`Gender`**, and **`Country`**

The VAR Statement

- The VAR statement identifies the numeric variables for which statistics are calculated.
- General form of the VAR statement:

VAR *analysis-variable(s)*;

- SUM is the default statistic for analysis variables.
- Examples of analysis variables:

  `Salary` and **`Bonus`**

One/two-Dimensional Table

```
proc tabulate data=orion.sales;
   class Country;
   table Country;
run;
```

Country	
AU	US
N	N
63.00	102.00

```
proc tabulate data=orion.sales;
   class Gender Country;
   table Gender, Country;
run;
```

	Country	
	AU	US
	N	N
Gender		
F	27.00	41.00
M	36.00	61.00

Three-Dimensional Table

```
proc tabulate data=orion.sales;
   class Job_Title Gender Country;
   table Job_Title, Gender, Country;
run;
```

Job_Title Sales Rep. I

	Country	
	AU	US
Gender		
F		
M		

Job_Title Sales Rep. II

	Country	
	AU	US
	N	N
Gender		
F	10.00	14.00
M	8.00	14.00

Dimension Expression

- Elements that can be used in a dimension expression:
  - classification variables
  - analysis variables
  - the universal class variable ALL
  - keywords for statistics

- Operators that can be used in a dimension expression:
  - blank, which concatenates table information
  - asterisk *, which crosses table information
  - parentheses (), which group elements

Dimension Expression

```
proc tabulate data=orion.sales;
   class Gender Country;
   var Salary;
   table Gender all, Country*Salary;
run;
```

	Country	
	AU	US
	Salary	Salary
	Sum	Sum
Gender		
F	747965.00	1207900.00
M	1152050.00	2033505.00
All	1900015.00	3241405.00

PROC TABULATE Statistics

Descriptive Statistic Keywords				
	CSS	CV	LCLM	MAX
MEAN	MIN	MODE	N	NMISS
KURTOSIS	RANGE	SKEWNESS	STDDEV	STDERR
SUM	SUMWGT	UCLM	USS	VAR
PCTN	REPPCTN	PAGEPCTN	ROWPCTN	COLPCTN
PCTSUM	REPPCTSUM	PAGEPCTSUM	ROWPCTSUM	COLPCTSUM
Quantile Statistic Keywords				
MEDIAN | P50	P1	P5	P10	Q1 | P25
Q3 | P75	P90	P95	P99	QRANGE
Hypothesis Testing Keywords				
PROBT	T			

PROC TABULATE Statistics

```
proc tabulate data=orion.sales;
    class Gender Country;
    var Salary;
    table Gender all, Country*Salary*(min max);
run;
```

	Country			
	AU		US	
	Salary		Salary	
	Min	Max	Min	Max
Gender				
F	25185.00	30890.00	25390.00	83505.00
M	25745.00	108255.00	22710.00	243190.00
All	25185.00	108255.00	22710.00	243190.00

Additional SAS Statements

- Additional statements can be added to enhance the

```
proc format;
   value $ctryfmt 'AU'='Australia'
                  'US'='United States';
run;

options nodate pageno=1;

ods html file='p112d08.html';
proc tabulate data=orion.sales;
   class Gender Country;
   var Salary;
   table Gender all, Country*Salary*(min
max);
   where Job_Title contains 'Rep';
   label Salary='Annual Salary';
   format Country $ctryfmt.;
   title 'Sales Rep Tabular Report';
run;
ods html close;
```

Sales Rep Tabular Report

	Country			
	Australia		United States	
	Annual Salary		Annual Salary	
	Min	Max	Min	Max
Gender				
F	25185.00	30890.00	25390.00	32985.00
M	25745.00	36605.00	22710.00	35990.00
All	25185.00	36605.00	22710.00	35990.00

Output Data Sets

- PROC TABULATE produces output data sets using the following method:

```
PROC TABULATE DATA=SAS-data-set
                  OUT=SAS-data-set <options>;
```

- The output data set contains the following variables:
  - BY variables
  - class variables
  - the automatic variables **_TYPE_**, **_PAGE_**, and **_TABLE_**
  - calculated statistics

PROC Statement OUT= Option

```
proc tabulate data=orion.sales
              out=work.tabulate;
   where Job_Title contains 'Rep';
   class Job_Title Gender Country;
   table Country;
   table Gender, Country;
   table Job_Title, Gender, Country;
run;

proc print data=work.tabulate;
run;
```

Obs	Job_Title	Gender	Country	_TYPE_	_PAGE_	_TABLE_	N
1			AU	001	1	1	61
2			US	001	1	1	98
3		F	AU	011	1	2	27
4		F	US	011	1	2	40
5		M	AU	011	1	2	34
6		M	US	011	1	2	58
7	Sales Rep. I	F	AU	111	1	3	8
8	Sales Rep. I	F	US	111	1	3	13
9	Sales Rep. I	M	AU	111	1	3	13
10	Sales Rep. I	M	US	111	1	3	29
11	Sales Rep. II	F	AU	111	2	3	10
12	Sales Rep. II	F	US	111	2	3	14
13	Sales Rep. II	M	AU	111	2	3	8
14	Sales Rep. II	M	US	111	2	3	14
15	Sales Rep. III	F	AU	111	3	3	7
16	Sales Rep. III	F	US	111	3	3	8
17	Sales Rep. III	M	AU	111	3	3	10
18	Sales Rep. III	M	US	111	3	3	9

PROC Statement OUT= Option

- **_TYPE_** is a character variable that shows which combination of class variables produced the summary statistics in that observation.

- Partial PROC PRINT Output

Obs	Job_Title	Gender	Country	_TYPE_	_PAGE_	_TABLE_	N
1			AU	001	1	1	61
2			US	001	1	1	98
3		F	AU	011	1	2	27
4		F	US	011	1		
5		M	AU	011	1		
6		M	US	011	1		

0 for **Job_Title**,
1 for **Gender**, and
1 for **Country**

PROC Statement OUT= Option

- **_PAGE_** is a numeric variable that shows the logical page number that contains that observation.
- Partial PROC PRINT Output

Obs	Job_Title	Gender	Country	_TYPE_	_PAGE_	_TABLE_	N
7	Sales Rep. I	F	AU	111	1	3	
8	Sales Rep. I	F	US	111	1	3	
9	Sales Rep. I	M	AU	111	1	3	
10	Sales Rep. I	M	US	111	1	3	29
11	Sales Rep. II	F	AU	111	2	3	
12	Sales Rep. II	F	US	111	2	3	
13	Sales Rep. II	M	AU	111	2	3	
14	Sales Rep. II	M	US	111	2	3	14
15	Sales Rep. III	F	AU	111	3	3	
16	Sales Rep. III	F	US	111	3	3	
17	Sales Rep. III	M	AU	111	3	3	
18	Sales Rep. III	M	US	111	3	3	9

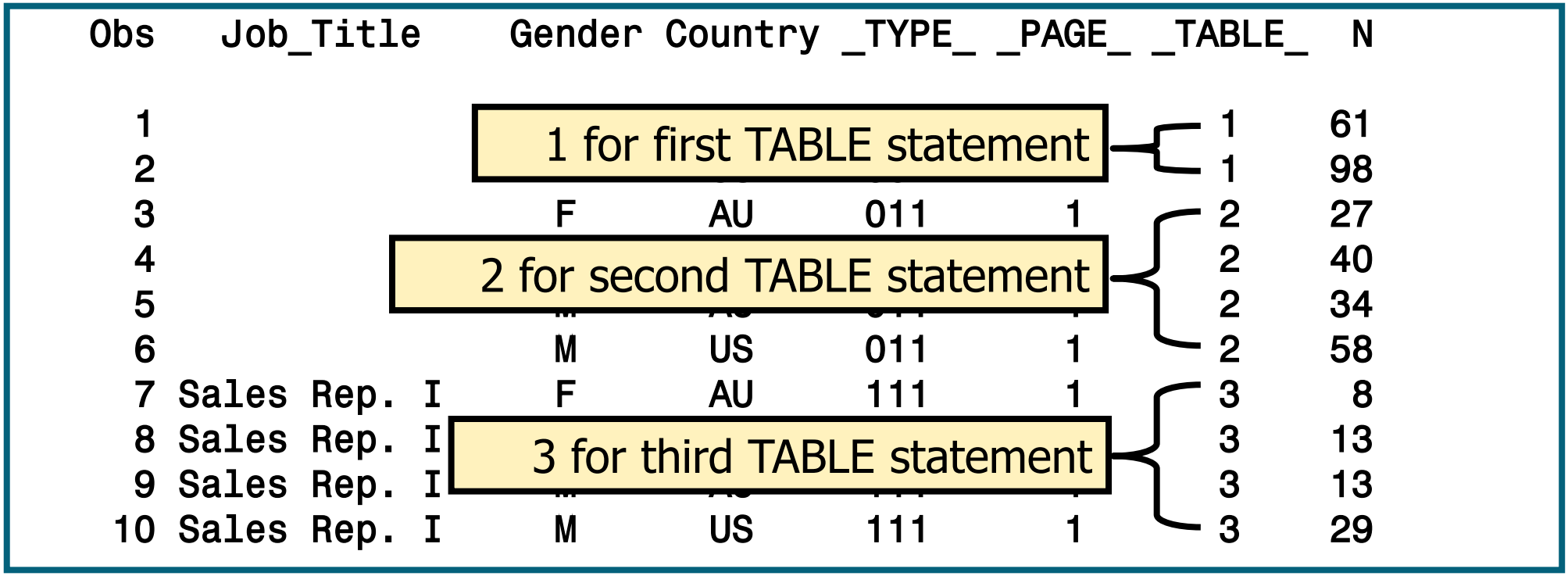
Page 1 for Sales Rep. I

Page 2 for Sales Rep. II

Page 3 for Sales Rep. III

PROC Statement OUT= Option

- **_TABLE_** is a numeric variable that shows the number of the TABLE statement that contains that observation.
- Partial PROC PRINT Output



Více o PROC TABULATE:

- In the SUGI 28 proceedings:
  - *"The Simplicity and Power of the TABULATE Procedure"*, by Dan Bruns

  http://www2.sas.com/proceedings/sugi28/197-28.pdf
- Online (from the SUGI 27 proceedings):
  - *"Anyone Can Learn PROC TABULATE"*, by Lauren Haworth, http://www2.sas.com/proceedings/sugi27/p060-27.pdf

The UNIVARIATE Procedure

- The UNIVARIATE procedure produces summary reports that display descriptive statistics.

- General form of the UNIVARIATE procedure:

```
PROC UNIVARIATE DATA=SAS-data-set;
     VAR variable(s);
RUN;
```

- The VAR statement specifies the analysis variables and their order in the results.

The UNIVARIATE Procedure

The following PROC UNIVARIATE step shows default descriptive statistics for **Salary**.

```
proc univariate data=orion.nonsales;
   var Salary;
run;
```

- Without the VAR statement, SAS will analyze all numeric variables.

The UNIVARIATE Procedure

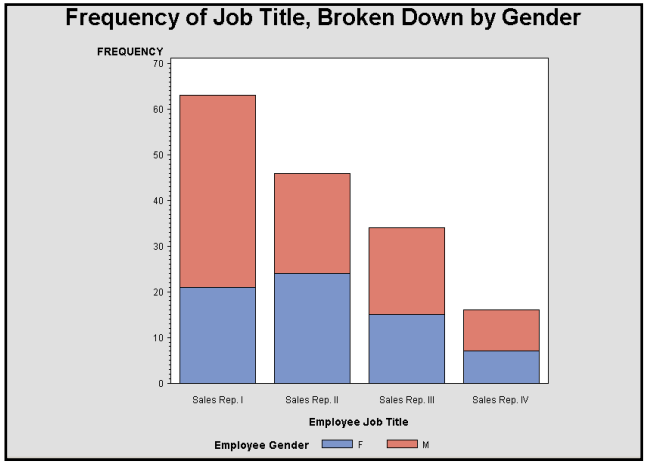
- The UNIVARIATE procedure can produce the following sections of output:
  - Moments
  - Basic Statistical Measures
  - Tests for Locations
  - Quantiles
  - Extreme Observations
  - Missing Values

What Is SAS/GRAPH Software?

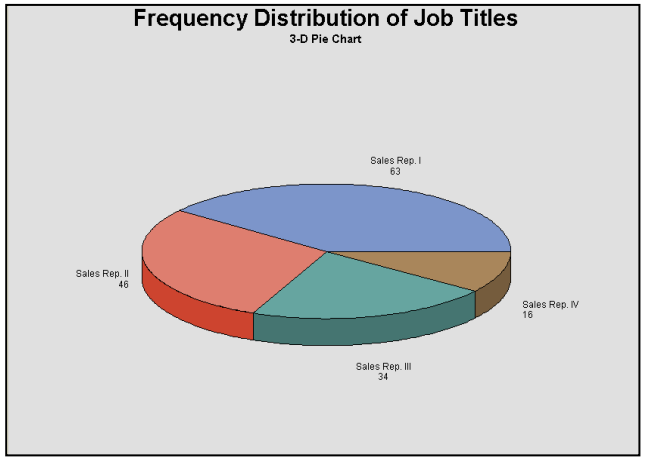
- *SAS/GRAPH software* is a component of SAS software that enables you to create the following types of graphs:
  - bar, block, and pie charts
  - two-dimensional scatter plots and line plots
  - three-dimensional scatter and surface plots
  - contour plots
  - maps
  - text slides
  - custom graphs

Základní typy grafů

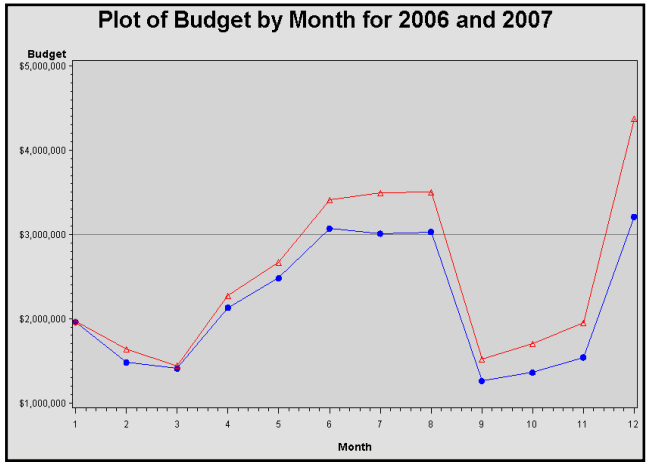
- Bar Charts (GCHART Procedure)



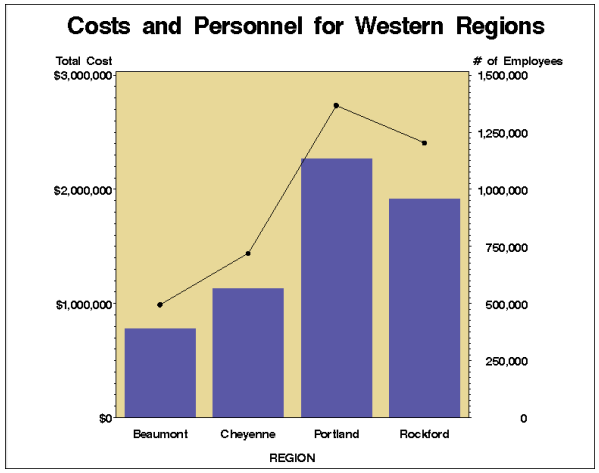
- Pie Charts (GCHART Procedure)



- Scatter and Line Plots (GPLOT Procedure)

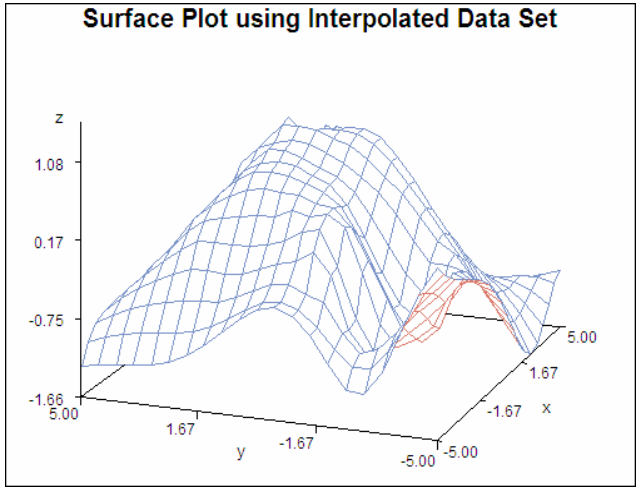


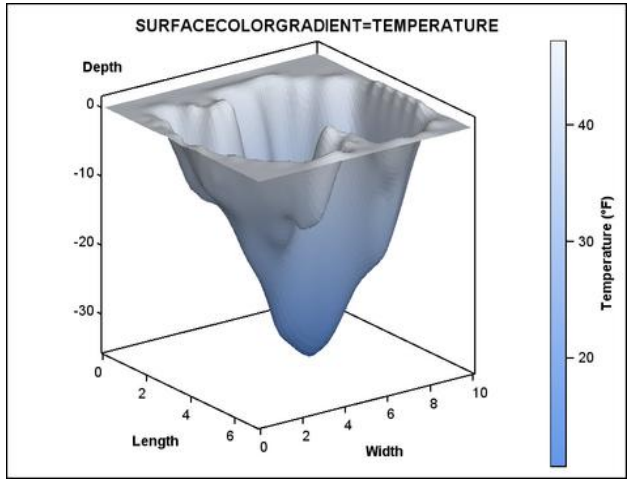
- Bar Charts with Line Plot Overlay (GBARLINE Procedure)



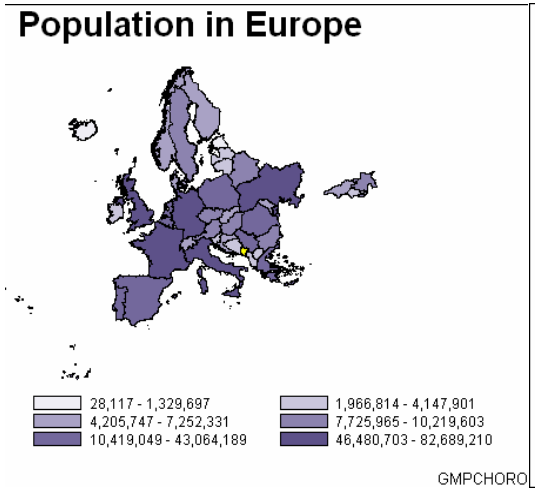
Three-Dimensional Surface and Scatter Plots, Maps

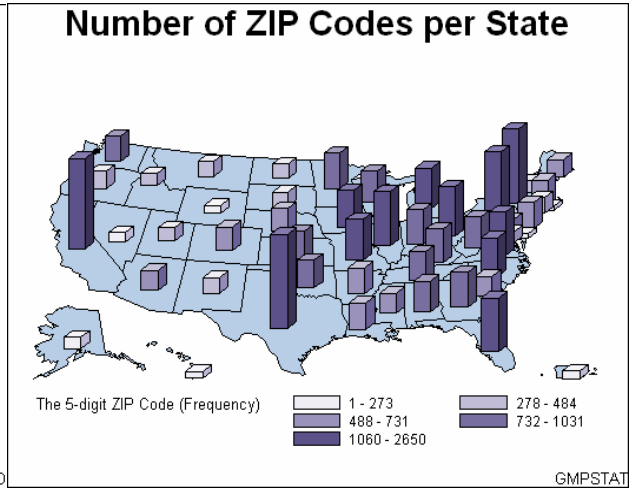
- Procedury G3D, G3GRID, SGRENDER …více na support.sas.com



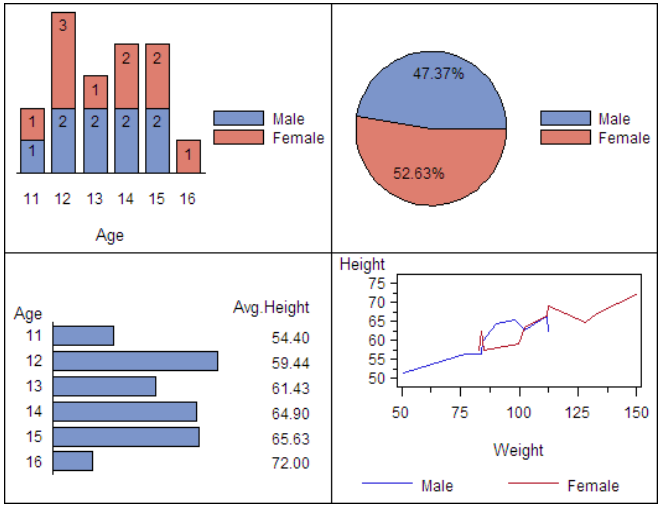


- Maps (GMAP Procedure)





- Multiple graphs on a page (GREPLAY Procedure)



Producing Bar and Pie Charts with the GCHART Procedure

- General form of the PROC GCHART statement:

PROC GCHART DATA=*SAS-data-set*;

- Use one of these statements to specify the chart type:

HBAR *chart-variable* . . . *</ options>*;
HBAR3D *chart-variable* . . . *</ options>*;

VBAR *chart-variable* . . . *</ options>*;
VBAR3D *chart-variable* . . . *</ options>*;

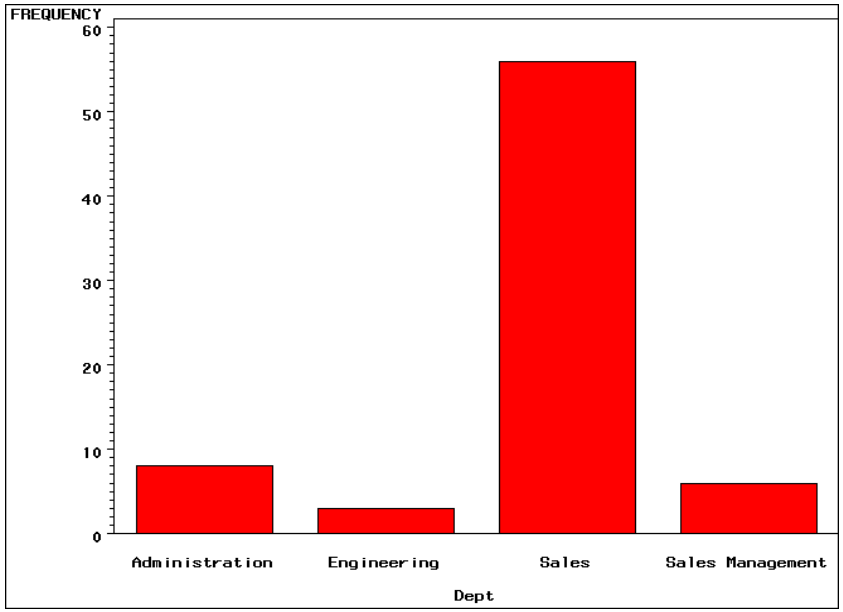
PIE *chart-variable* . . . *</ options>*;
PIE3D *chart-variable* . . . *</ options>*;

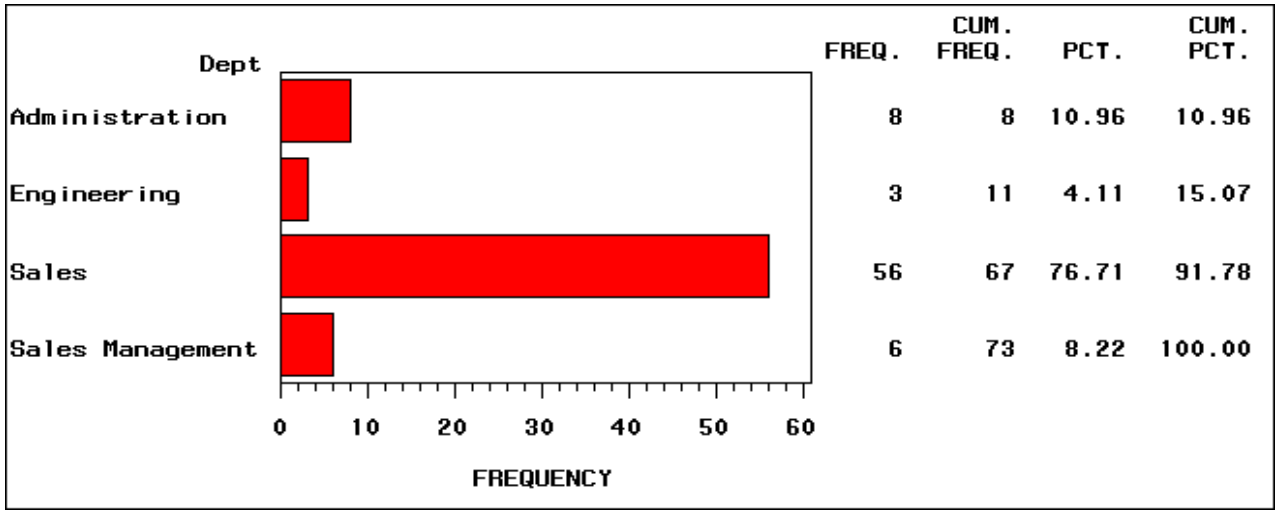
Vertical/horizontal Bar Chart

- Produce a vertical/horizontal bar chart that displays the number of employees in each department.

```
proc gchart
 data=univ.employees;
   vbar dept;
run;
```

```
proc gchart
 data=univ.employees;
   hbar dept;
run;
```





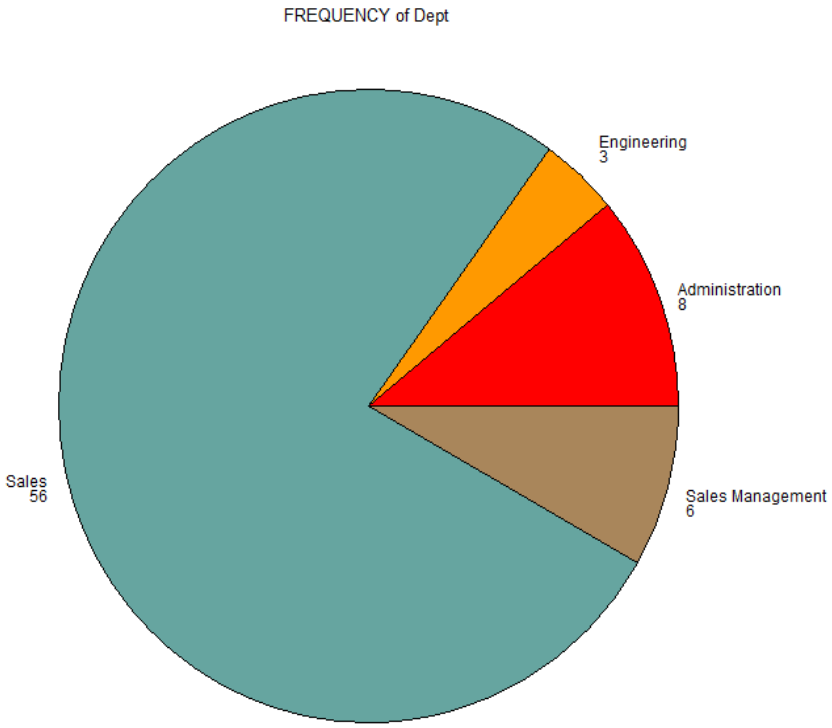
dept is the chart variable

Pie Chart

- Produce a pie chart that displays the number of employees in each department.

```
proc gchart data=univ.employees;
   pie dept;
run;
```

dept is the chart variable



Character/Numeric Chart Variable

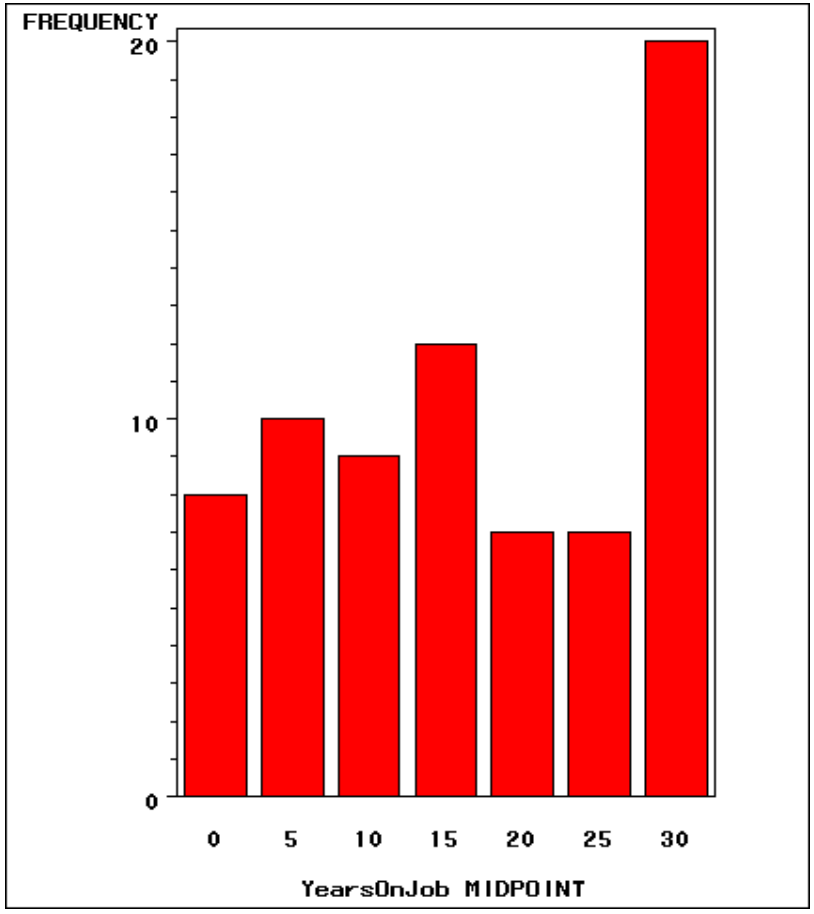
- If the chart variable is **character**, then a bar or slice is created for each unique variable value.
- For **numeric** chart variables, the variables are assumed to be continuous unless otherwise specified.
- The GCHART procedure creates the equivalent of a histogram from the data.
  - Intervals are automatically calculated and identified by midpoints.
  - One bar or slice is constructed for each midpoint.

Numeric Chart Variable

- Produce a vertical bar chart on the numeric variable **YearsOnJob**.

```
proc gchart data=univ.employees;
   vbar YearsOnJob;
run;
```

YearsOnJob is the chart variable

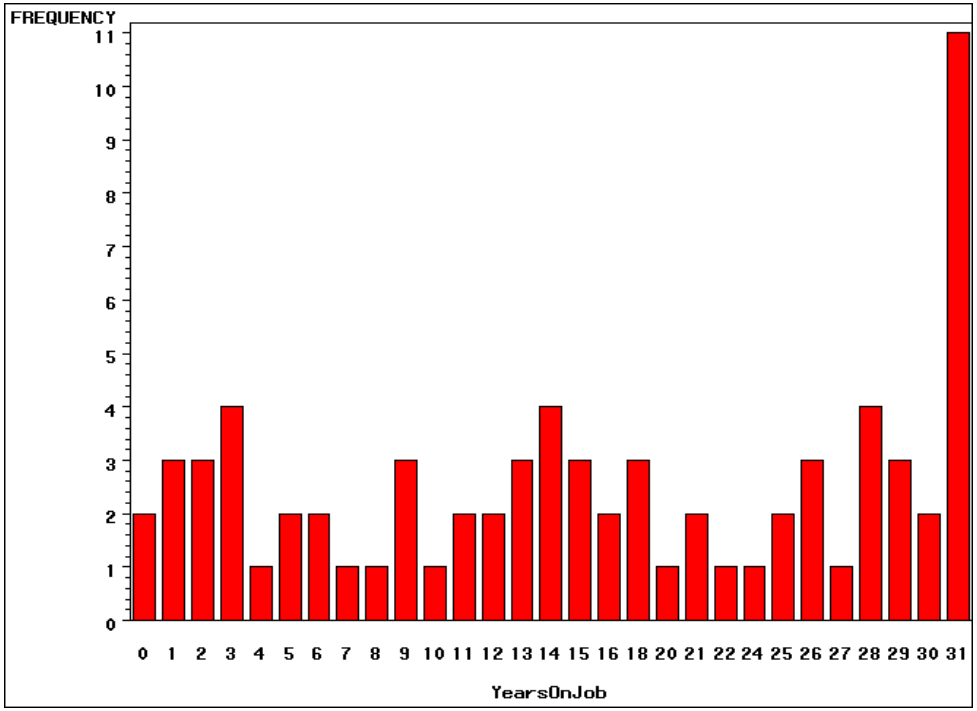


The DISCRETE Option

- To override the default behavior for numeric chart variables, use the DISCRETE option in the HBAR, VBAR, or PIE statement.
- The DISCRETE option produces a bar or slice for each unique numeric variable value; the values are no longer treated as intervals.

```
proc gchart data=univ.employees;
   vbar YearsOnJob / discrete;
run;
```

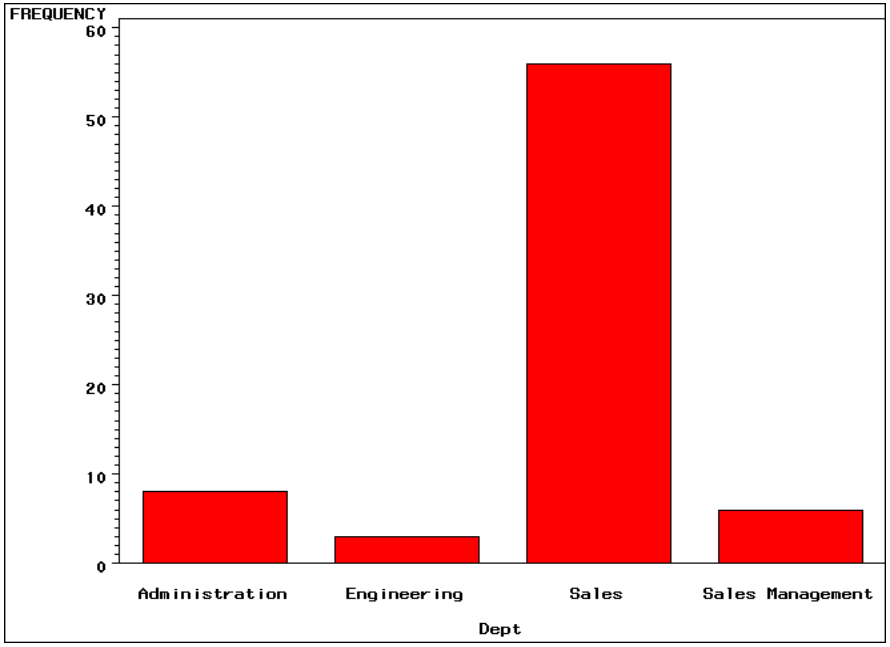
YearsOnJob is the chart variable, but the DISCRETE option modifies how SAS displays the values.

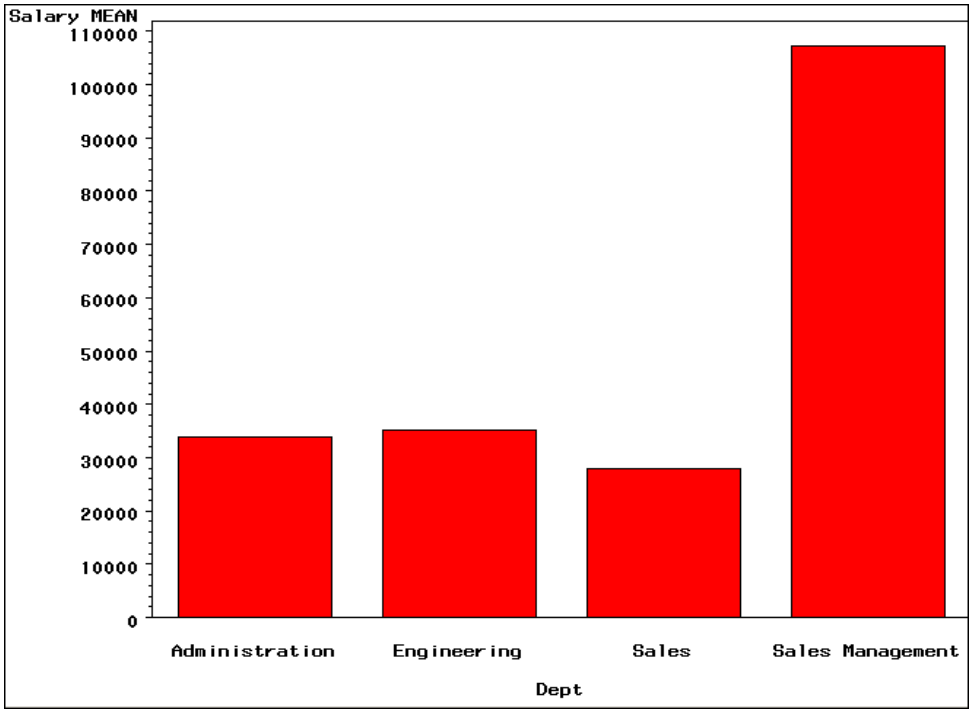


Summary Statistic

- By default, the statistic that determines the length or height of each bar or size of pie slice is a frequency count (N).

```
proc gchart data=univ.employees;
   vbar dept / sumvar=salary type=mean;
run;
```





Analysis Variable

- To override the default frequency count, you can use the following HBAR, VBAR, or PIE statement options:

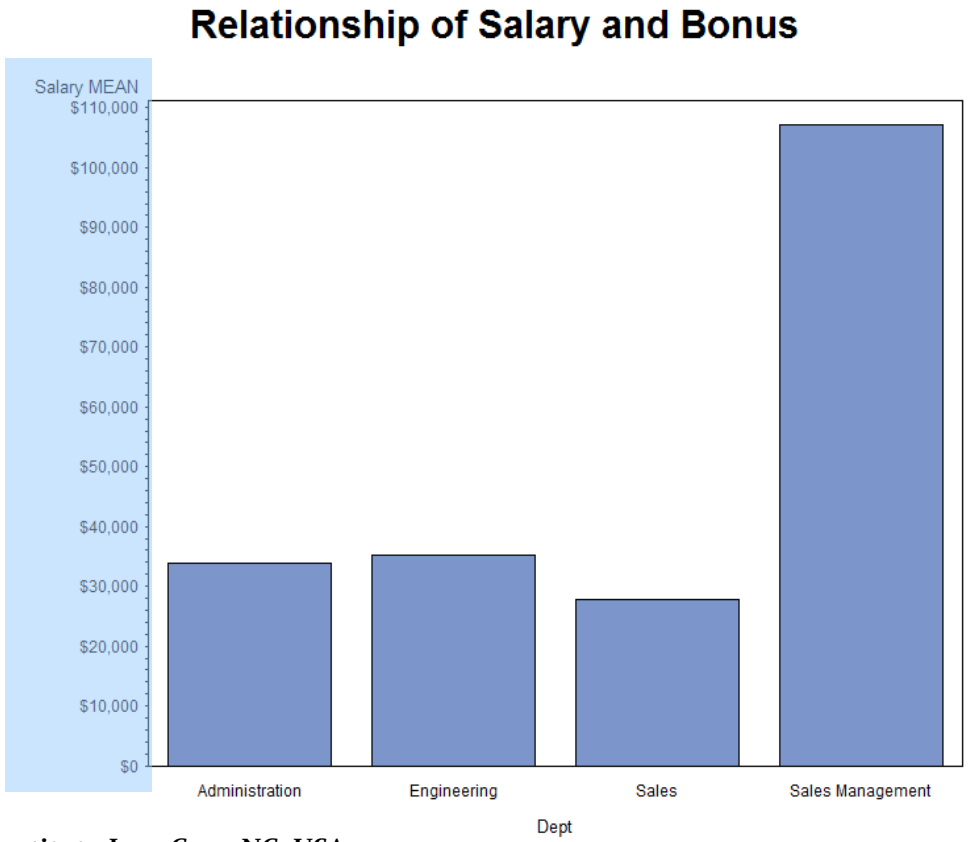
SUMVAR=	identifies the analysis variable to use for the sum or mean calculation.
TYPE=	specifies that the height or length of the bar or size of the slice represents a mean or sum of the *analysis-variable* values.

- If an analysis variable is
  - specified, the default value of TYPE is SUM
  - not specified, the default value of TYPE is FREQ.

Bar Chart Using Formats

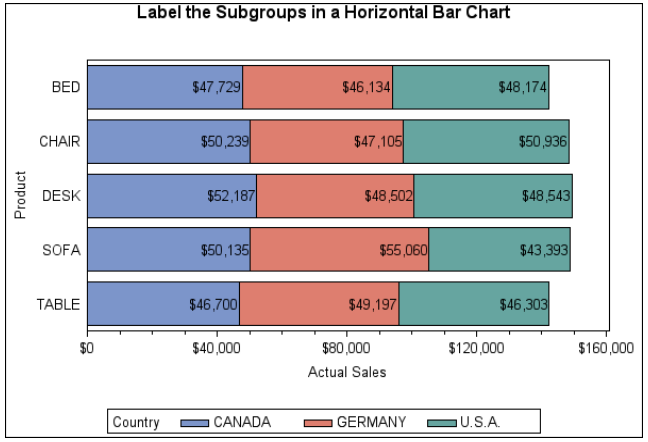
- Produce a bar chart that displays the average salary of employees in each department.

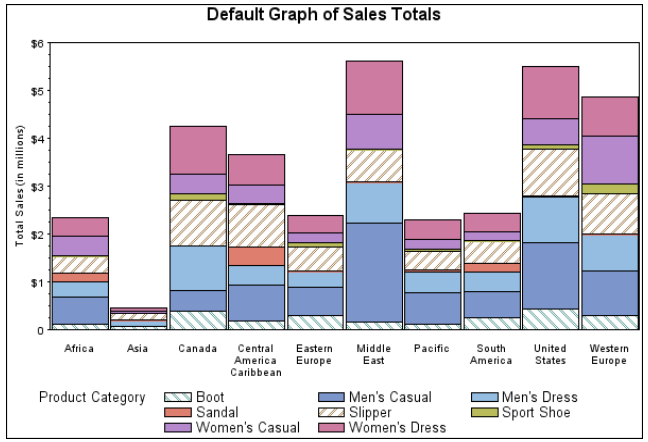
```
proc gchart data=univ.employees;
    vbar dept / sumvar=Salary type=mean;
    format Salary dollar8.;
run;
```

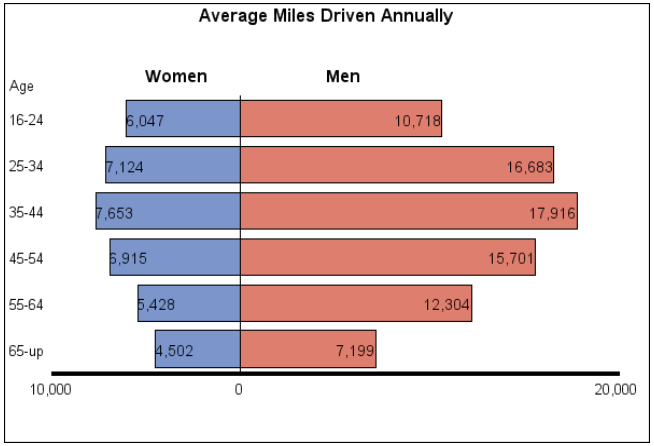


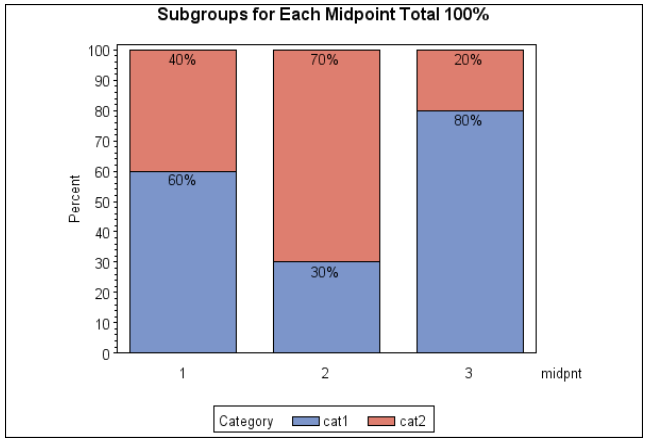
Další možnosti proc gchart

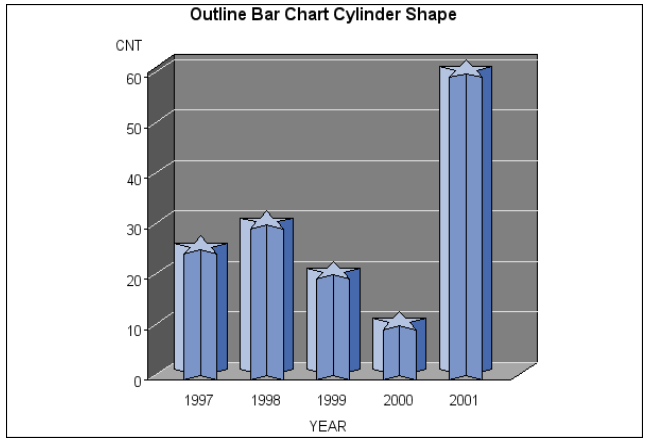
- Na adrese http://support.sas.com/sassamples/graphgallery/PROC_GCHART.html lze nalézt galerii možných typů grafů (**včetně kódů**!).

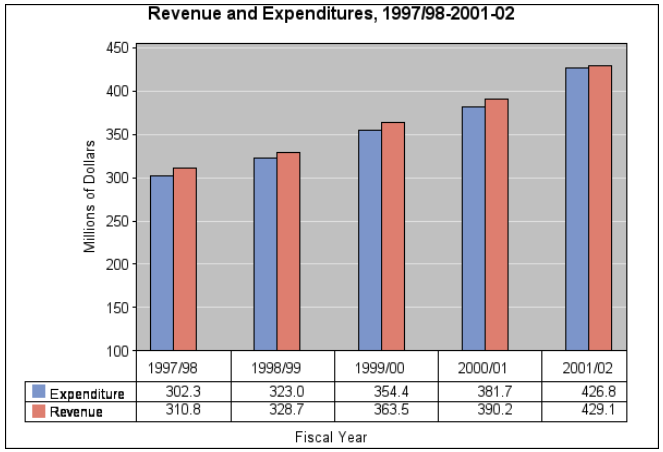






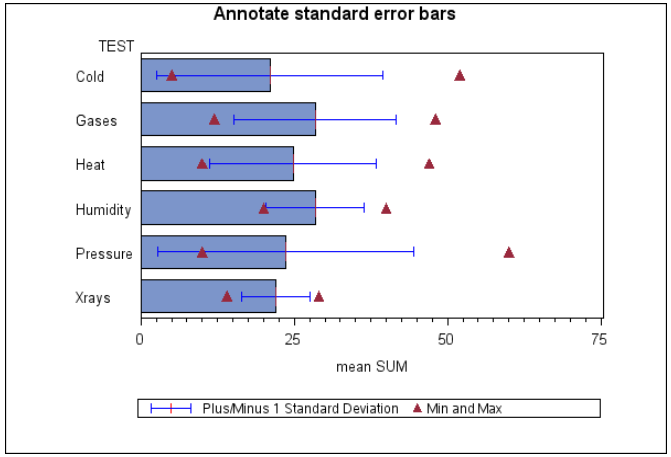


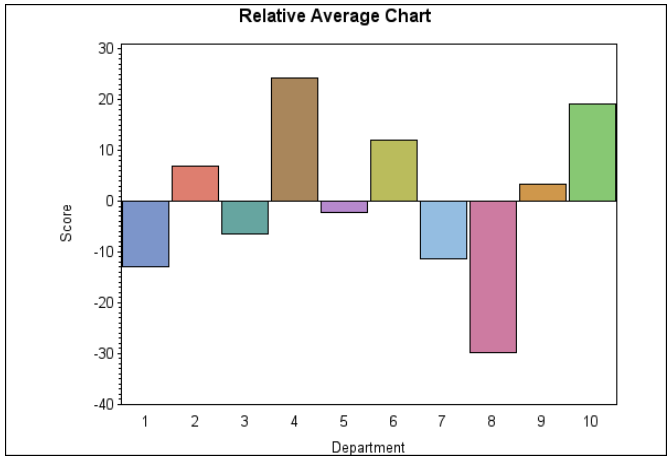


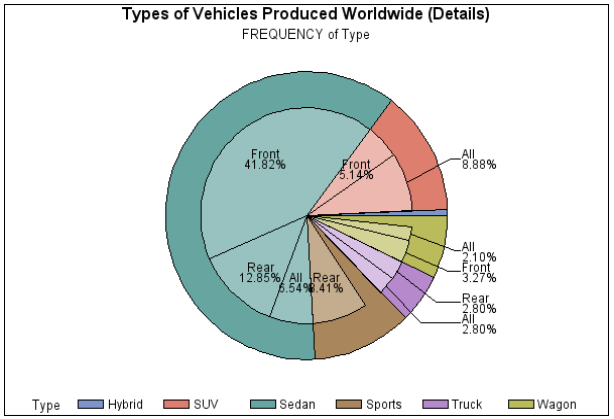


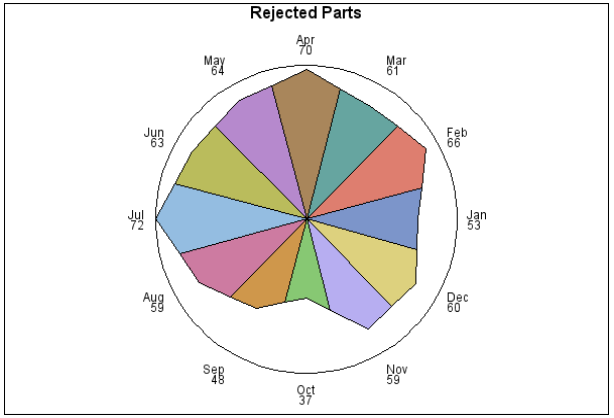
Další možnosti proc gchart

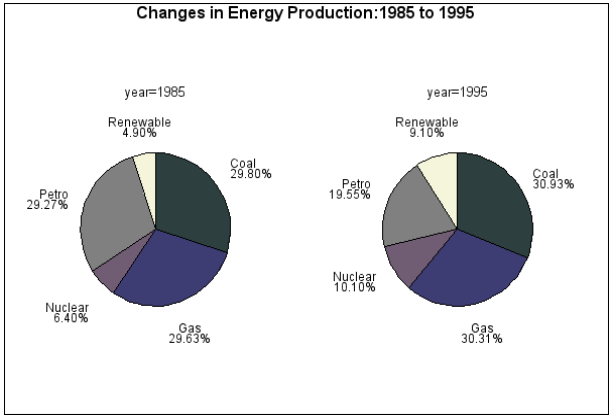
- A ještě několik typů…

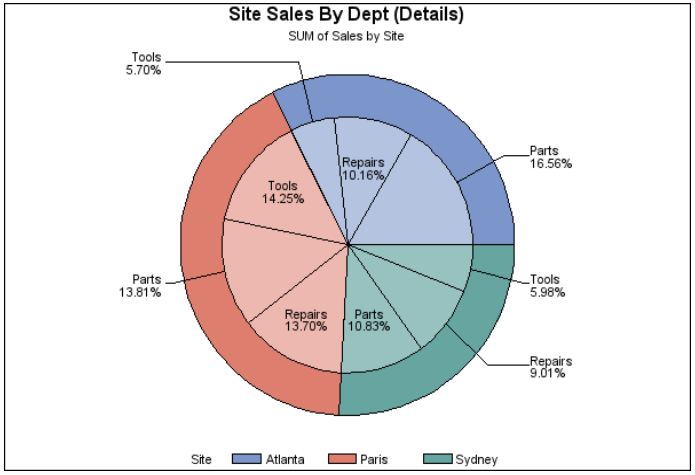


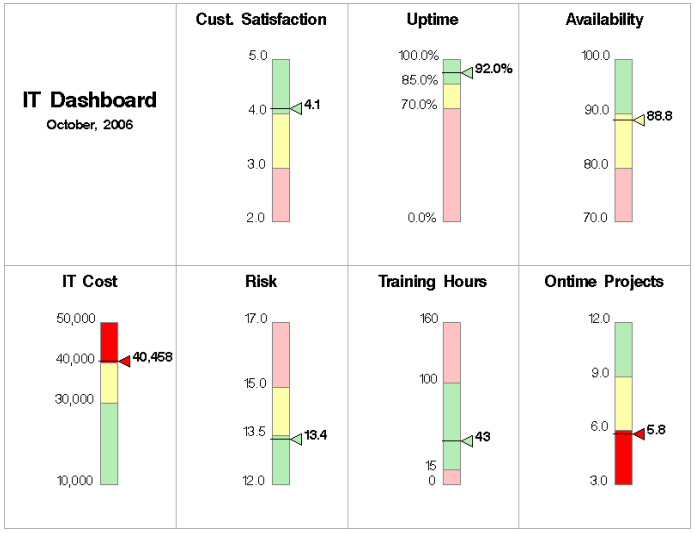












Producing Plots with the GPLOT Procedure

- You can use the GPLOT procedure to plot one variable against another within a set of coordinate axes.

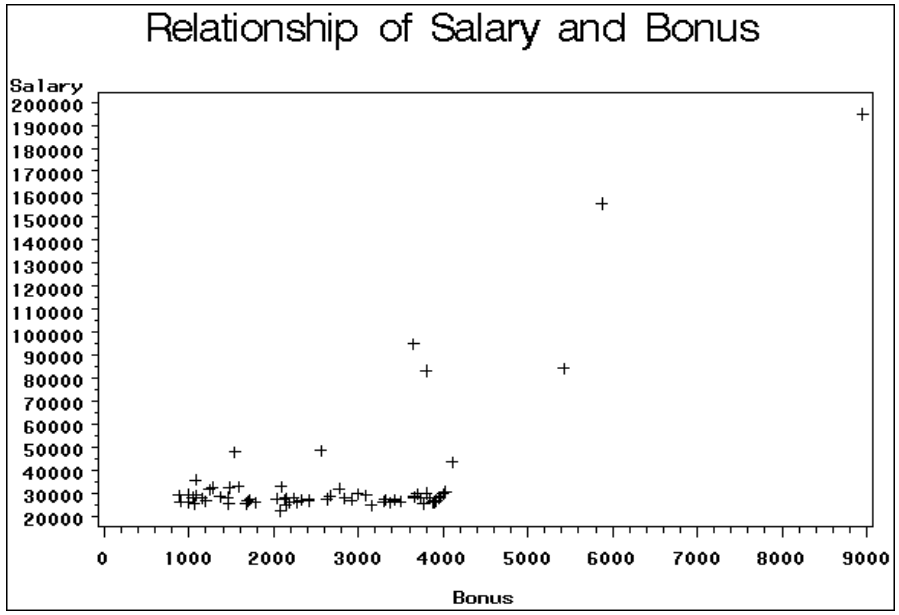
- General form of a PROC GPLOT step:

```
PROC GPLOT DATA=SAS-data-set;
        PLOT vertical-variable*horizontal-variable </ options>;
RUN;
QUIT;
```

The GPLOT Procedure

Produce a plot of salary versus bonus for each employee.

```
proc gplot data=univ.employees;
   plot Salary*Bonus;
   title 'Relationship of Salary and Bonus';
run;
```



SYMBOL Statement

- You can use the SYMBOL statement to do the following:
  - define plotting symbols
  - draw lines through the data points
  - specify the color of the plotting symbols and lines

- General form of the SYMBOL statement:

SYMBOL*n* options;

- The value of *n* can range from 1 to 255.
- If *n* is omitted, the default is 1.
- Symbol statement is global and additive:

global	After being defined, the statements remain in effect until changed or until the end of the SAS session.
additive	Specifying the value of one option does not affect the values of other options.

SYMBOL Statement Options

- You can specify the plotting symbol you want with the VALUE= option in the SYMBOL statement:

VALUE=*symbol* | **V=***symbol*

- Selected *symbol* values are shown below:

PLUS (default)	DIAMOND
STAR	TRIANGLE
SQUARE	NONE (no plotting symbol)

- You can use the I= option in the SYMBOL statement to draw lines between the data points.
- Selected *interpolation* values:

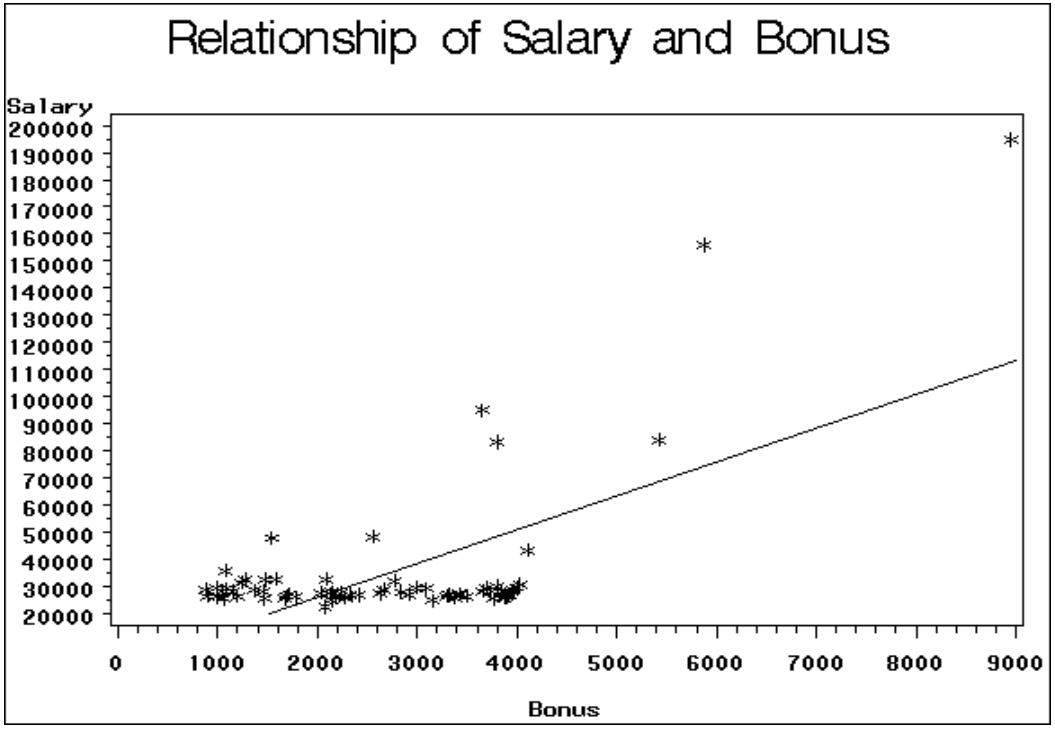
I=*interpolation*

JOIN	joins the points with straight lines.
SPLINE	joins the points with a smooth line.
NEEDLE	draws vertical lines from the points to the horizontal axes.
R	overlays a simple linear regression line on the plot.

SYMBOL Statement Options

- Use a star as the plotting symbol and superimpose a regression line on the plot.

```
plot Salary*Bonus;
 symbol value=star i=r;
run;
```

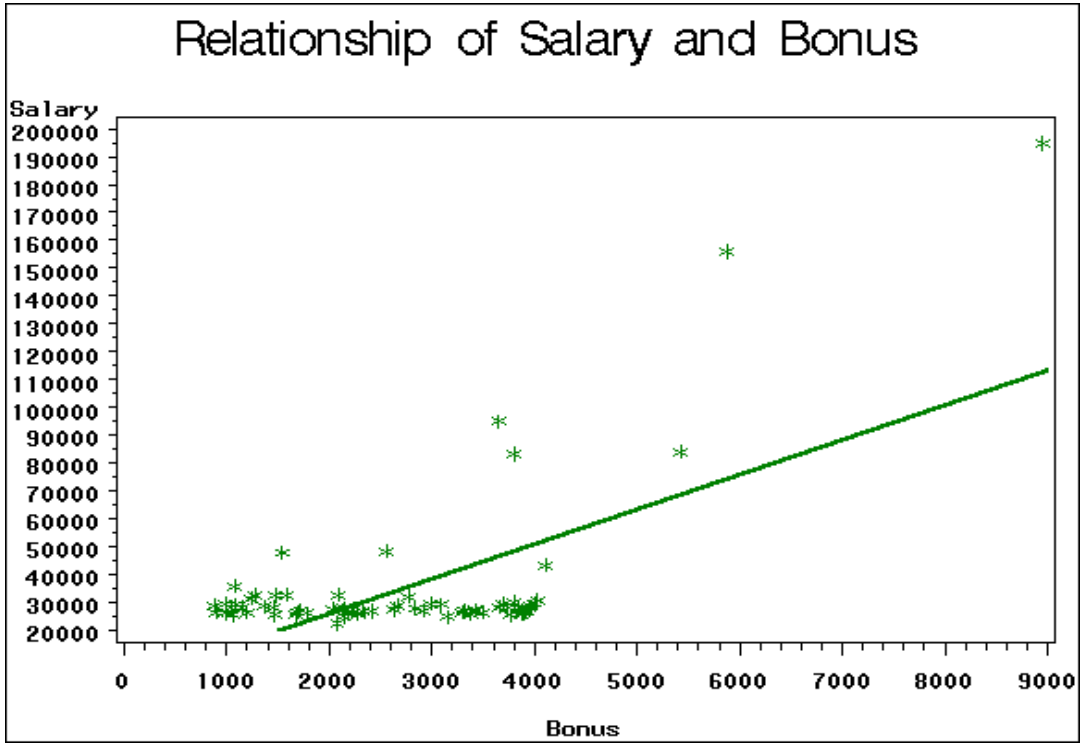


Additional SYMBOL Statement Options

- You can enhance the appearance of the plots with the following selected options:

WIDTH=*width* **W**=*width*	specifies the thickness of the line.
COLOR=*color* **C**=*color*	specifies the color of the line and plot symbols.

```
plot Salary*Bonus;
symbol c=green w=3;
run;
```



Canceling SYMBOL Statements

- You can cancel a SYMBOL statement by submitting a null SYMBOL statement.

```
symbol1;
```

- To cancel all SYMBOL statements, submit the following statement:

```
goptions reset=symbol;
```

- Zrušení všech předchozích voleb (návrat k defaultnímu nastavaní)

```
goptions reset=global;
```

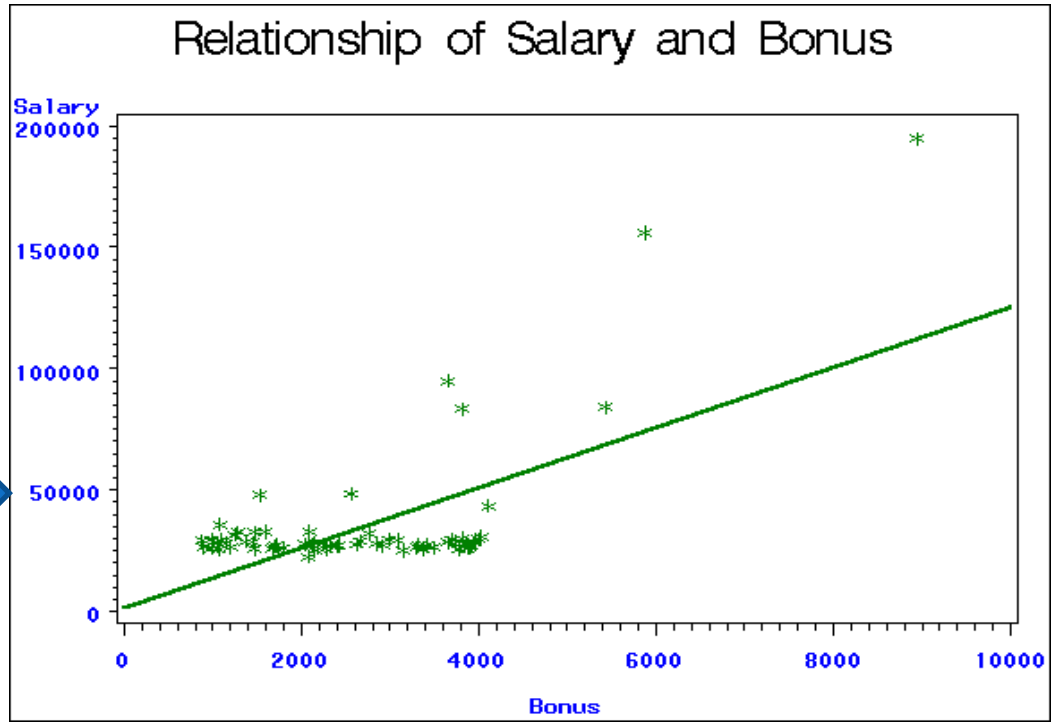
Controlling the Axis Appearance

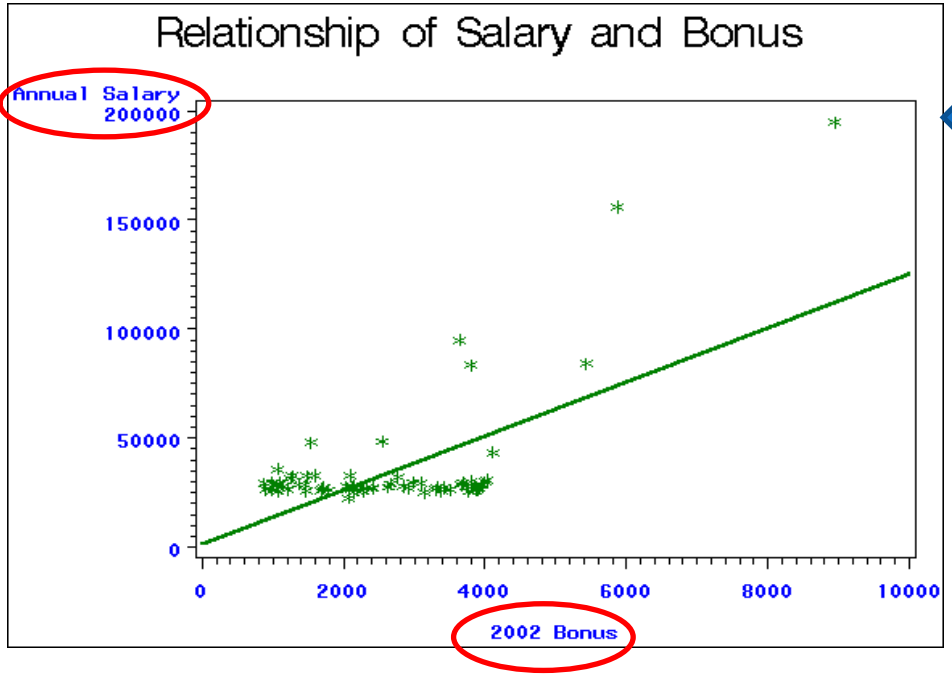
- You can modify the appearance of the axes that PROC GPLOT produces with the following:
  - PLOT statement options
  - the LABEL statement
  - the FORMAT statement
- You can use PLOT statement options to control the scaling and color of the axes, and the color of the axis text.
- Selected PLOT statement options for axis control:

HAXIS=*values*	scales the horizontal axis.
VAXIS=*values*	scales the vertical axis.
CAXIS=*color*	specifies the color of both axes.
CTEXT=*color*	specifies the color of the text on both axes.

PLOT Statement Options, Label statement

```
plot Salary*Bonus
    / vaxis=0 to 200000 by 50000
      haxis=0 to 10000 by 2000
      ctext=blue;
run;
```






```
plot Salary*Bonus /
    vaxis=0 to 200000 by 50000
    haxis=0 to 10000 by 2000
ctext=blue;
    label Salary='Annual Salary'
       Bonus='2002 Bonus';
run;
```

Gplot options – další možnosti

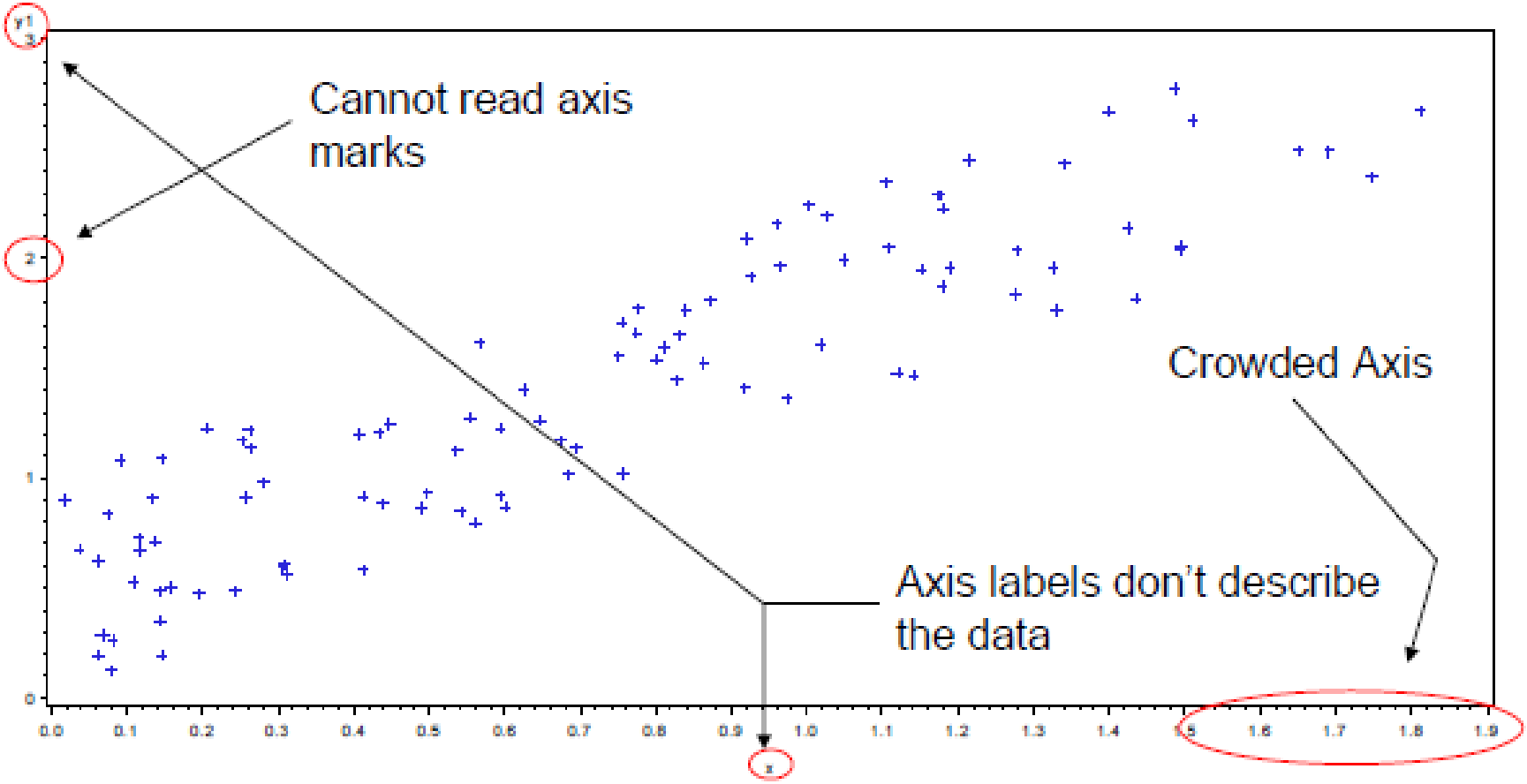
Plot *<Y Variable>*<X Variable> / <options>;*

- Options for plotting
  - Plot options
    - •Legend= or nolegend: specifies figure legend options
    - •Overlay: allows overlay of more than one Y variable
    - •Skipmiss: breaks the plotting line where Y values are missing
  - •Appearance option
    - •Axis: Specifies axis label and value options
    - •Symbol: Specified symbol options
    - •href, vref: Draws vertical or horizontal reference lines on plot
    - •frame/fr or noframe/nofr: specifies whether or not to frame the plot
    - •caxis/ca, cframe/cfr, chref/ch, cvref/cv, ctext/c: specifies colors used for axis, frame, text or reference lines.

Gplot options – další možnosti

```
proc gplot data=twovar;
      plot y1*x;
run;
```

Very basic plot, below we get all of the default options. Not very exciting. Definitely not publication quality.



Gplot options – další možnosti

- AXIS<1..99> <options>;
  - Label Option;
    - Angle/a=degrees (0-359)
    - Color/c=text color
    - Font/f=font
    - Height/h=text height (default=1)
    - Justify=(left/center/right)
    - Label="text string"
  - Order Option;
    - Order=(a to b by c): major tick marks will show up at intervals based on c.
      - Example order=(0 to 3 by 1);
  - Value Option;
    - value=("" "" ""): applies text label to each major tick.
      - Example Value=("Start" "Middle" "End")

- axis1 label=(a=90 c=black f="arial" h=1.2 "time" a=90 c=black f="arial" h=1.0 "hours");

Gplot options – další možnosti

Resets previous options

```
goptions reset=global ;
```

Horizontal axis (X Variable)

```
axis1 label=(f='arial/bo' h=1.9 "Dose" justify=c
             f='arial/bo' h=1.3 "mg/24 Hrs" );
```

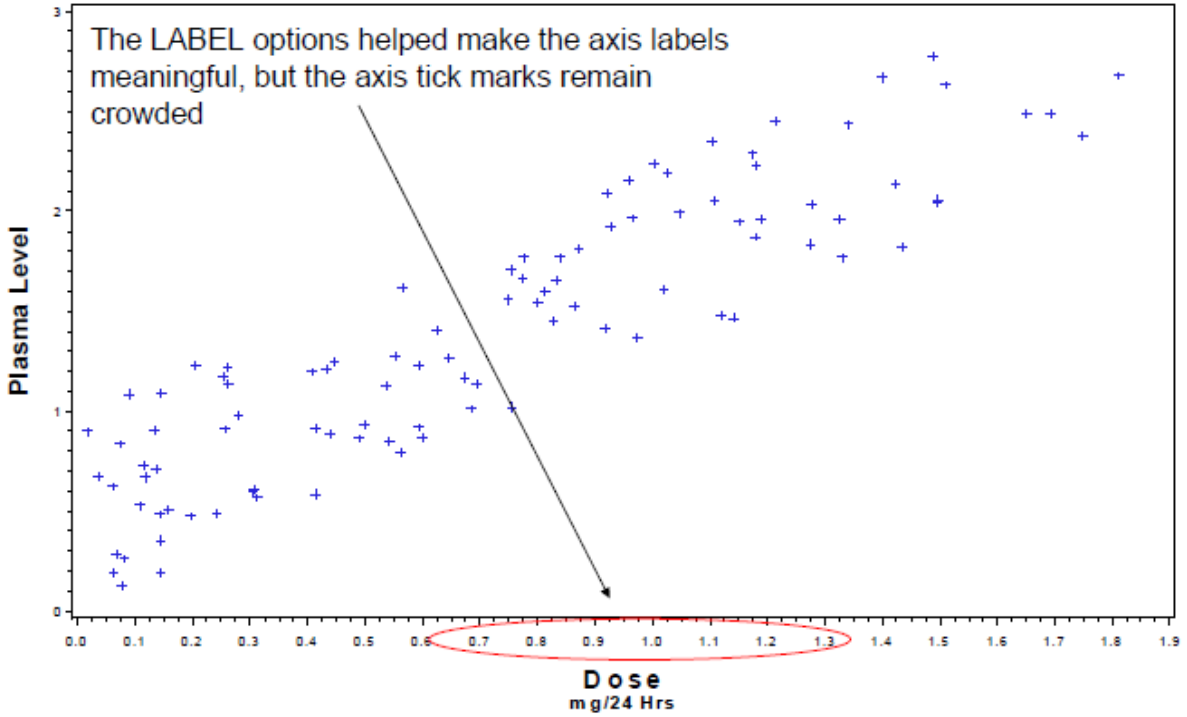
Vertical axis (Y Variable)

```
axis2 label=(a=90  f='arial/bo' h=1.9 "Plasma Level");
```

```
proc gplot data=twovar;
     plot y1*x / haxis=axis1 vaxis=axis2;
run;
```

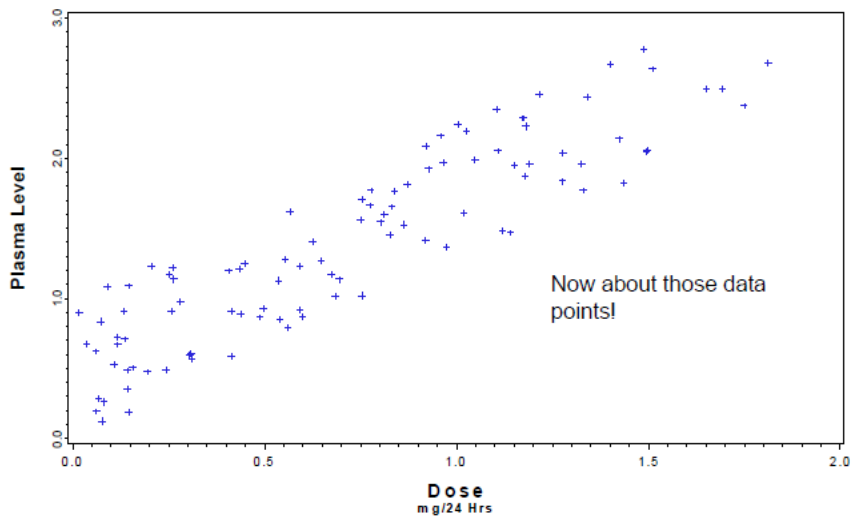
Call Axis statements

NOTE: you can also place the AXIS statements within the gplot proc



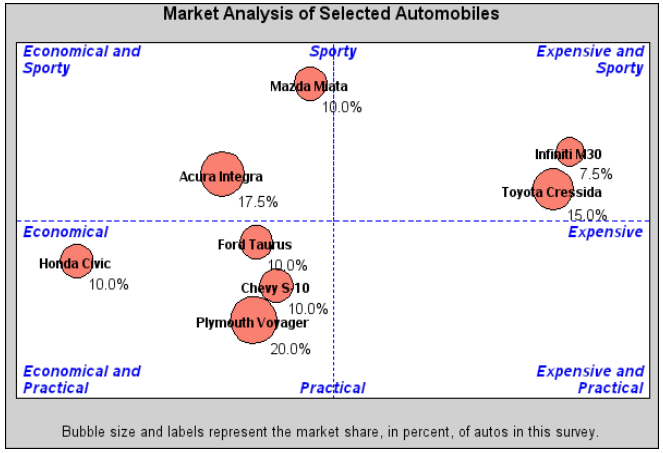
Added VALUE option to Axis statement

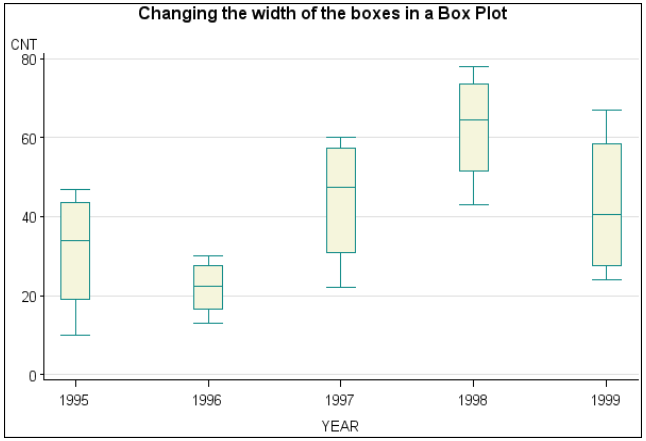
```
axis1 label=(f='arial/bo' h=1.9 "Dose" justify=c
             f='arial/bo' h=1.3 "mg/24 Hrs")
      order=(0 to 2 by 0.5)
      value=(f='arial' h=1.3 "0.0" "0.5" "1.0" "1.5" "2.0");

axis2 label=(a=90  f='arial/bo' h=1.9 "Plasma Level")
      order=(0 to 3 by 1)
      value=(a=90 f='arial' h=1.3 "0.0" "1.0" "2.0" "3.0");


proc gplot data=twovar;
     plot y1*x / haxis=axis1 vaxis=axis2;
run;
```

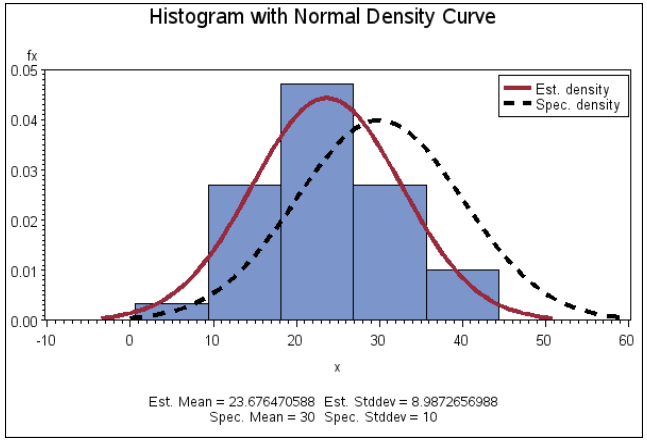


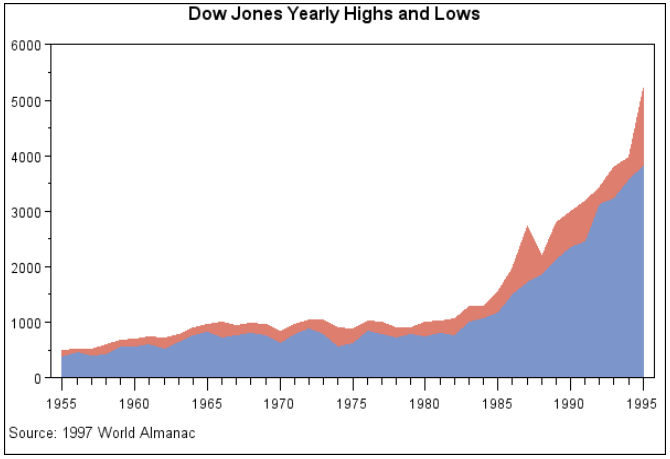
Další možnosti proc gplot

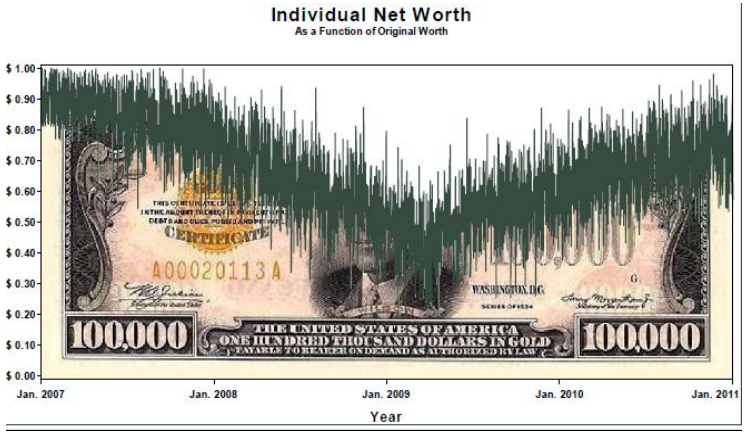
Na adrese http://support.sas.com/sassamples/graphgallery/PROC_GPLOT.html lze nalézt galerii možných typů grafů (**včetně kódů**!). Na adrese http://ebookbrowse.com/sas-gplot-slides-1-26-2011-ppt-d138883835 lze najít další návody a ukázky včetně kódů.









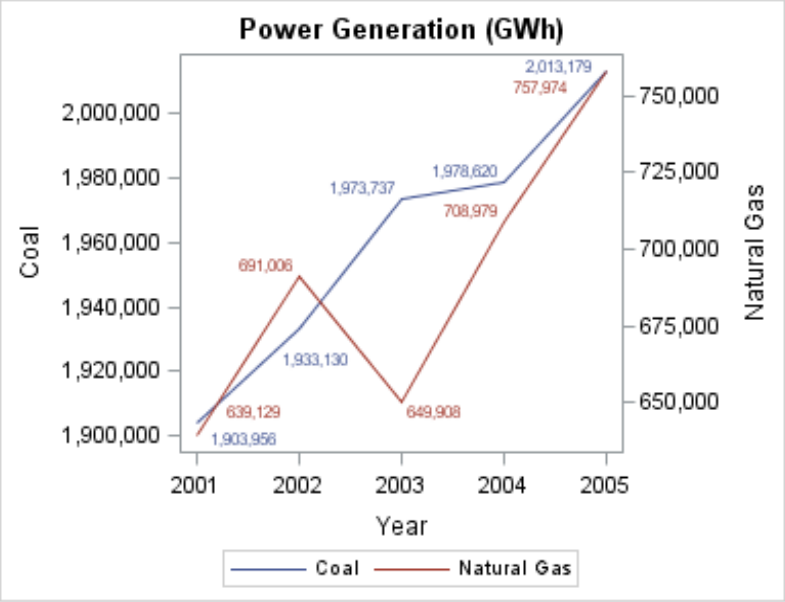


The SGPLOT Procedure

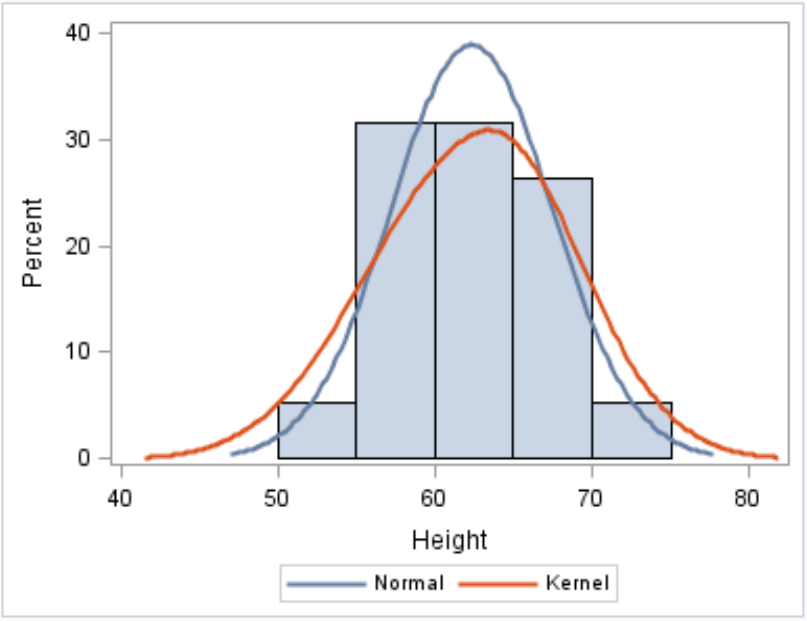
- General form of the SGPLOT procedure:

```
PROC SGPLOT <option(s)>;
      DOT category-variable </option(s)>;
      HBAR category-variable < /option(s) >;
      HBOX response-variable </option(s)>;
      HISTOGRAM response-variable < /option(s)>;
      NEEDLE X= variable Y= numeric-variable </option(s)>;
      REG X= numeric-variable Y= numeric-variable
                  </option(s)>;
      SCATTER X= variable Y= variable </option(s)>;
      VBAR category-variable < /option(s)>;
      VBOX response-variable </option(s)>;
RUN;
```

Proc Sgplot

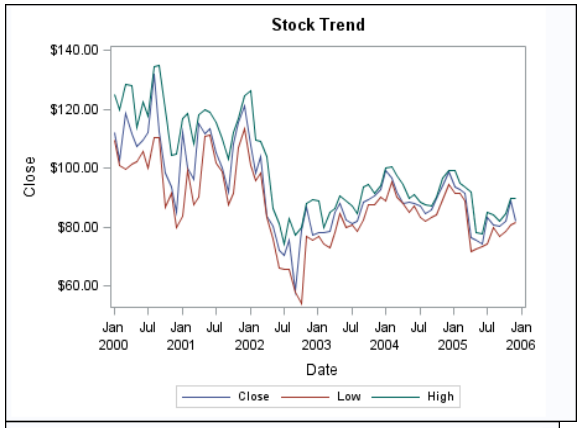



```
title "Power Generation (GWh)";
proc sgplot data=sashelp.electric(where=
  (year >= 2001 and customer="Residential"));
  xaxis type=discrete;
  series x=year y=coal / datalabel;
  series x=year y=naturalgas /
      datalabel y2axis;
run;
title;
```

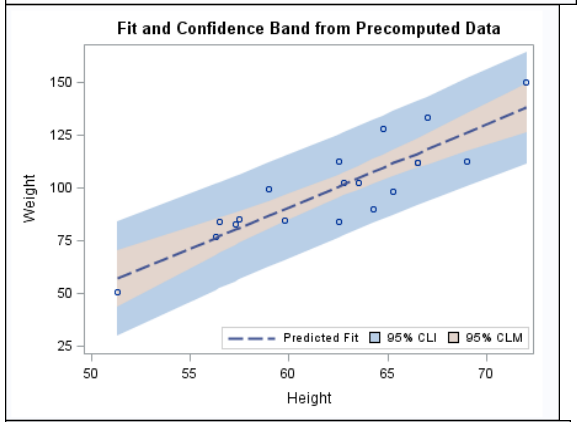



```
proc sgplot data=sashelp.class;
  histogram height;
  density height;
  density height / type=kernel;
run;
```

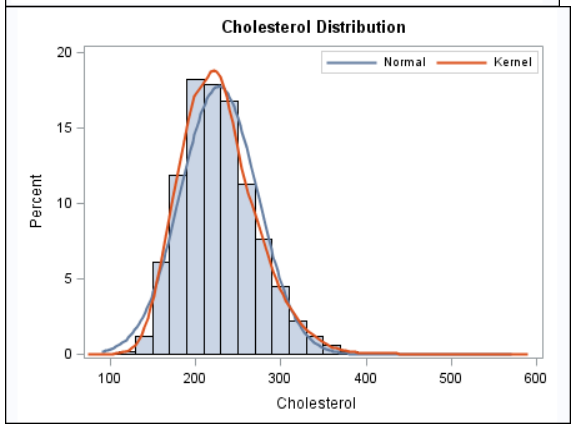
Proc Sgplot



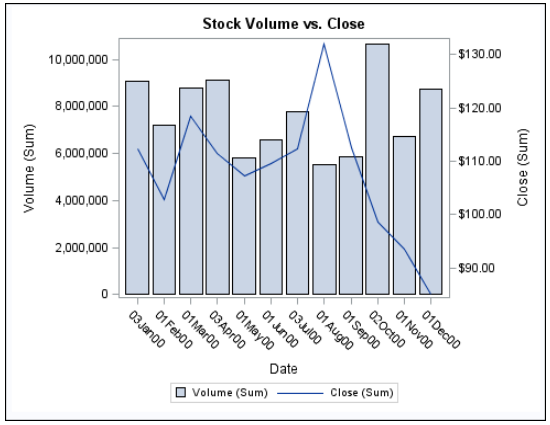

```
proc sgplot data=sashelp.stocks (where=(date
>= "01jan2000"d and stock = "IBM"));
 title "Stock Trend";
series x=date y=close;
series x=date y=low;
series x=date y=high;
run;
```




```
proc sgplot data=sashelp.classfit;
title "Fit and Confidence Band from Precomputed Data";
 band x=height lower=lower upper=upper / legendlabel="95% CLI"
name="band1";
 band x=height lower=lowermean upper=uppermean /
fillattrs=GraphConfidence2 legendlabel="95% CLM" name="band2";
scatter x=height y=weight;
series x=height y=predict / lineattrs=GraphPrediction legendlabel="Predicted Fit"
name="series";
keylegend "series" "band1" "band2" / location=inside
position=bottomright;
run;
```

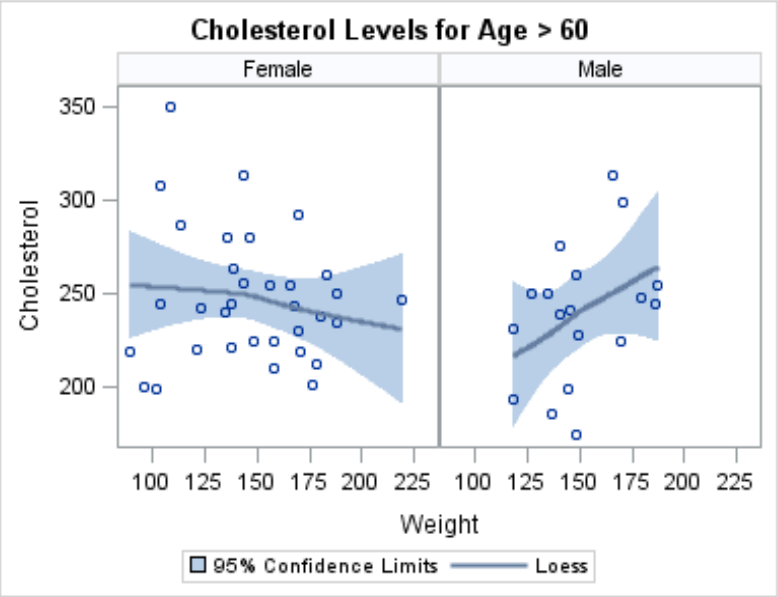



```
proc sgplot data=sashelp.heart;
title "Cholesterol Distribution";
histogram cholesterol;
density cholesterol;
density cholesterol /
type=kernel;
keylegend / location=inside
position=topright;
run;
```

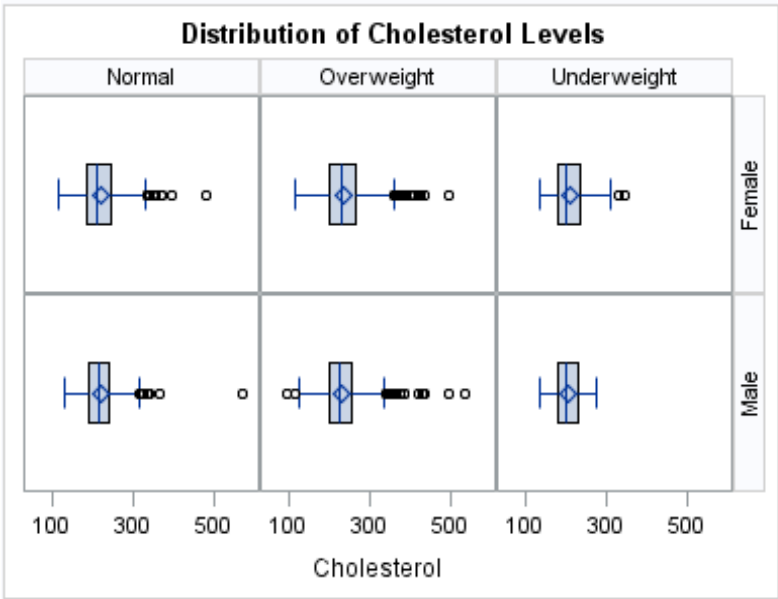



```
proc sgplot data=sashelp.stocks
(where=(date >= "01jan2000"d and
date <= "01jan2001"d and stock =
"IBM")); title "Stock Volume vs. Close";
vbar date / response=volume;
vline date / response=close y2axis;
run;
```

Proc Sgpanel




```
title1 "Cholesterol Levels for Age > 60";
proc sgpanel data=sashelp.heart(
    where=(AgeAtStart > 60)) ;
  panelby sex / novarname;
  loess x=weight y=cholesterol / clm;
run;
title1;
```



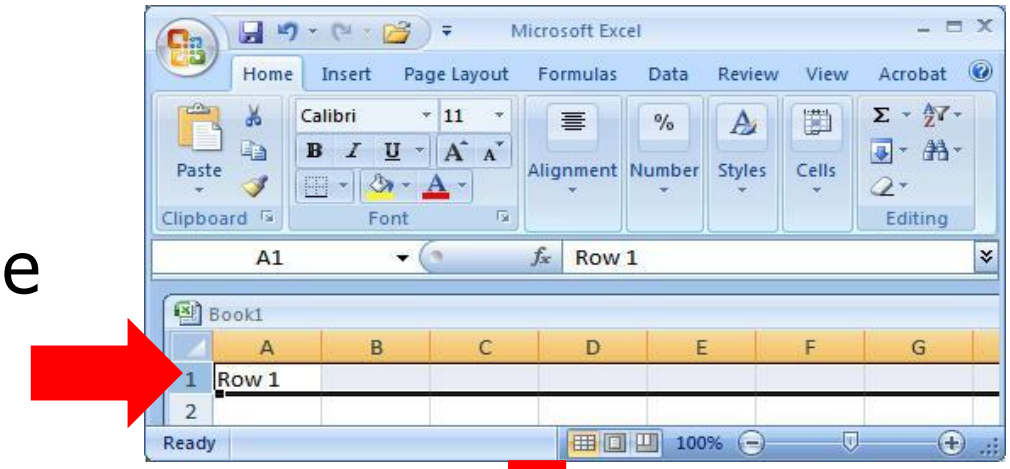

```
title1 "Distribution of Cholesterol Levels";
proc sgpanel data=sashelp.heart;
  panelby weight_status sex / layout=lattice
                              novarname;
  hbox cholesterol;
run;
title1;
```

Další viz :
http://support.sas.com/documentation/cdl/en/grstatproc/65235/HTML/default/viewer.htm#poomgdlxbij4v3nozewfb9cpfxu1.htm

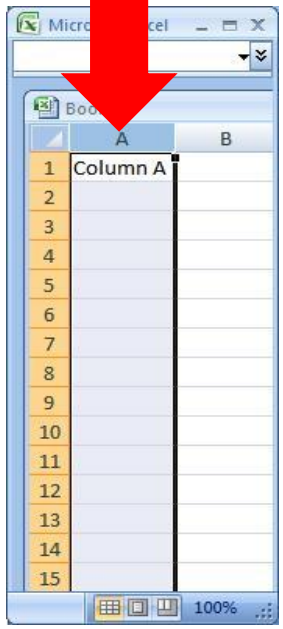
Excel Basics

Excel spreadsheets organize information (text and numbers) by rows and columns:

This is a **row**.
Rows are represented by **numbers** along the side of the sheet.



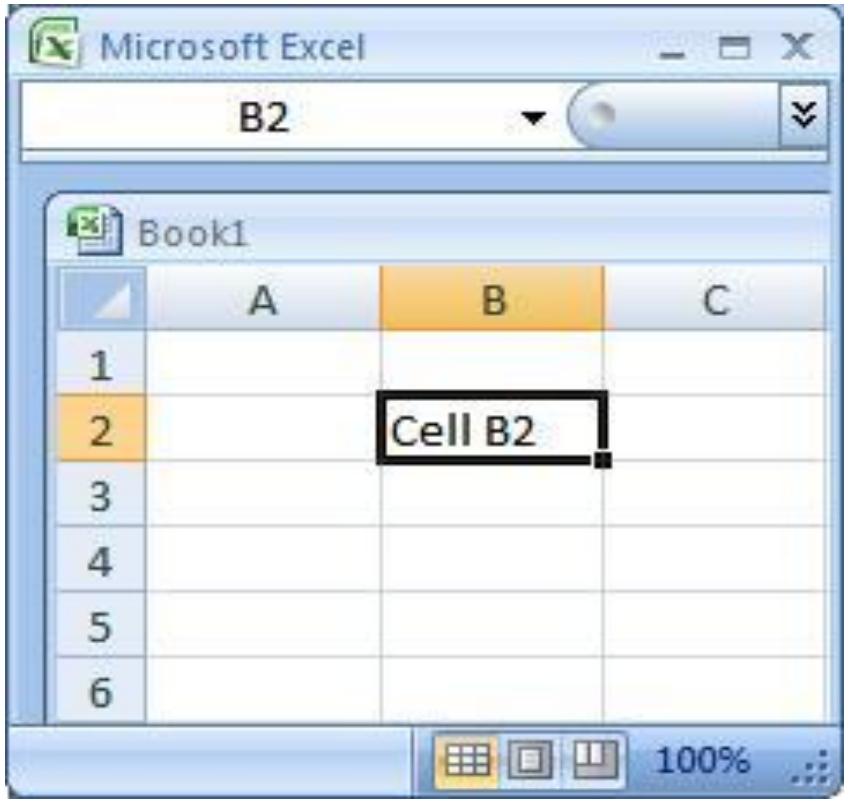
This is a **column**.
Columns are represented by **letters** across the top of the sheet.



Excel Basics

A **cell** is the intersection between a column and a row.

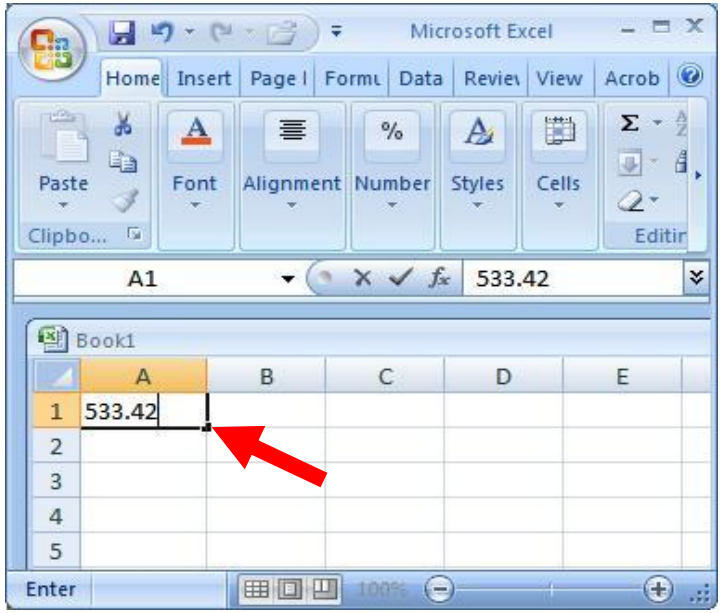
Each cell is named for the column letter and row number that intersect to make it.



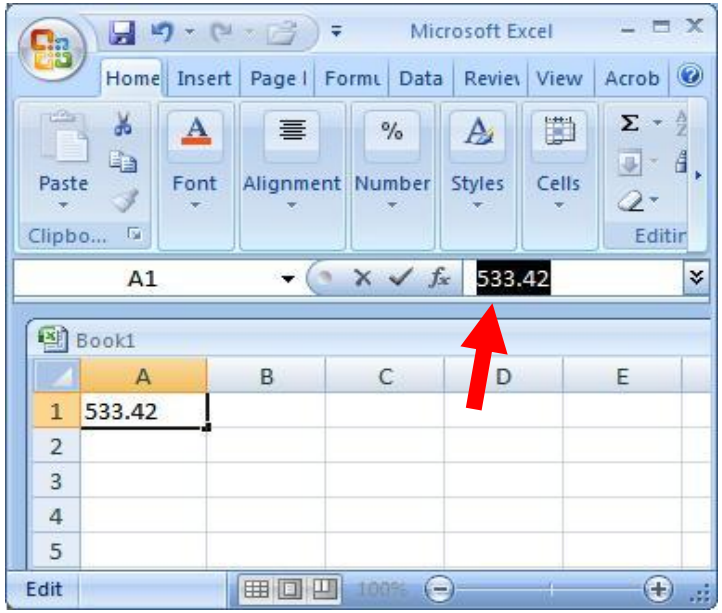
Data Entry

There are <u>two</u> ways to enter information into a cell:

1. Type directly into the cell.
Click on a cell, and type in the data (numbers or text) and press Enter.



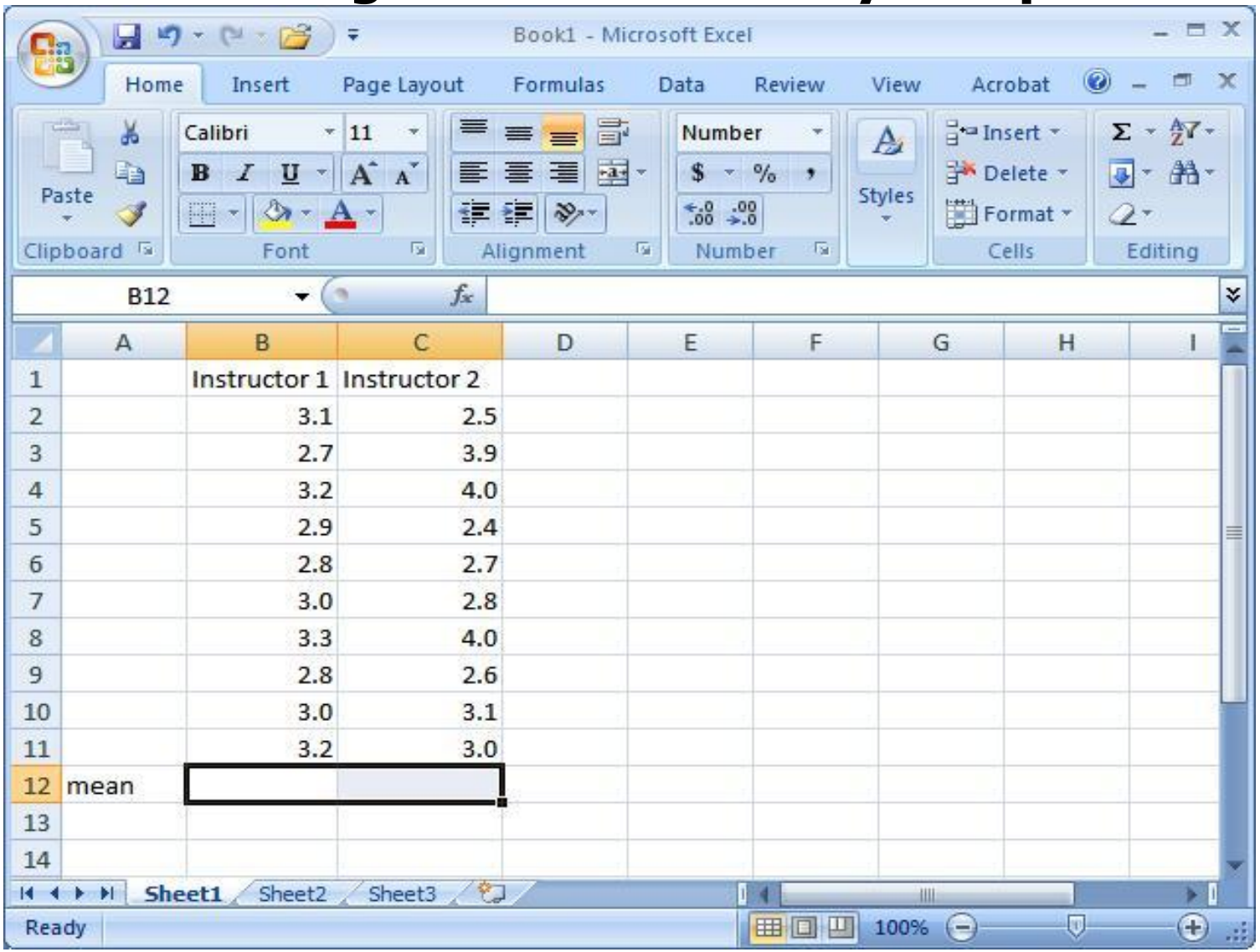
2. Type into the formula bar.
Click on a cell, and then click in the formula bar (the space next to the f_x). Now type the data into the bar and press Enter.



Data Entry

1. Open Excel (Start → All Programs → MS Office → Excel**).**

2. Enter the following information into your spreadsheet:



Formulas and Functions

- Formulas are equations that perform calculations in your spreadsheet. Formulas always begin with an equals sign (=). When you enter an equals sign into a cell, you are basically telling Excel to “calculate this.”

- Functions are Excel-defined formulas. They take data you select and enter, perform calculations on them, and return value(s).

More on Functions

- All functions have a common format – the equals sign followed by the function name followed by the input in parentheses.
- The input for a function can be either:
  - A set of numbers (e.g., "=AVERAGE(2, 3, 4, 5)")
    - This tells Excel to calculate the average of these numbers.
  - A reference to cell(s) (e.g., "=AVERAGE(B1:B18) or "=AVERAGE (B1, B2, B3, B4, B5, B6, B7, B8)"
    - This tells Excel to calculate the average of the data that appear in all the cells from B1 to B8.
    - You can either type these cell references in by hand or by clicking and dragging with your mouse to select the cells.

Functions for Descriptive Statistics

Below are several functions you will need to learn for this class. Try them out with the practice data set.

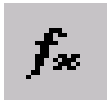
=AVERAGE(first cell:last cell): calculates the mean

=MEDIAN(first cell:last cell): calculates the median

=MODE(first cell:last cell): calculates the mode

=VAR**P**(first cell:last cell): calculates the variance

=STDEV**P**(first cell:last cell): calculates the standard deviation

- You may directly write the functions for these statistics into cells or the formula bar, OR
- You may use the function wizard ( in the toolbar)

Measures of Central Tendency in Excel

- **Average (Mean)**

Write the formula into a cell …

	A	B	C	E
1	23	2	11	34
2	11		4	12
3	5			8
4	7			13

=AVERAGE(B2:E5) = **=AVERAGE(A1:A4;B2;C1:C2;E1:E4)**

A colon (**:**) stands between the upper left corner and the lower right corner of an array

Empty cells don't effect on the value of mean

The reference to an array is made by painting the array

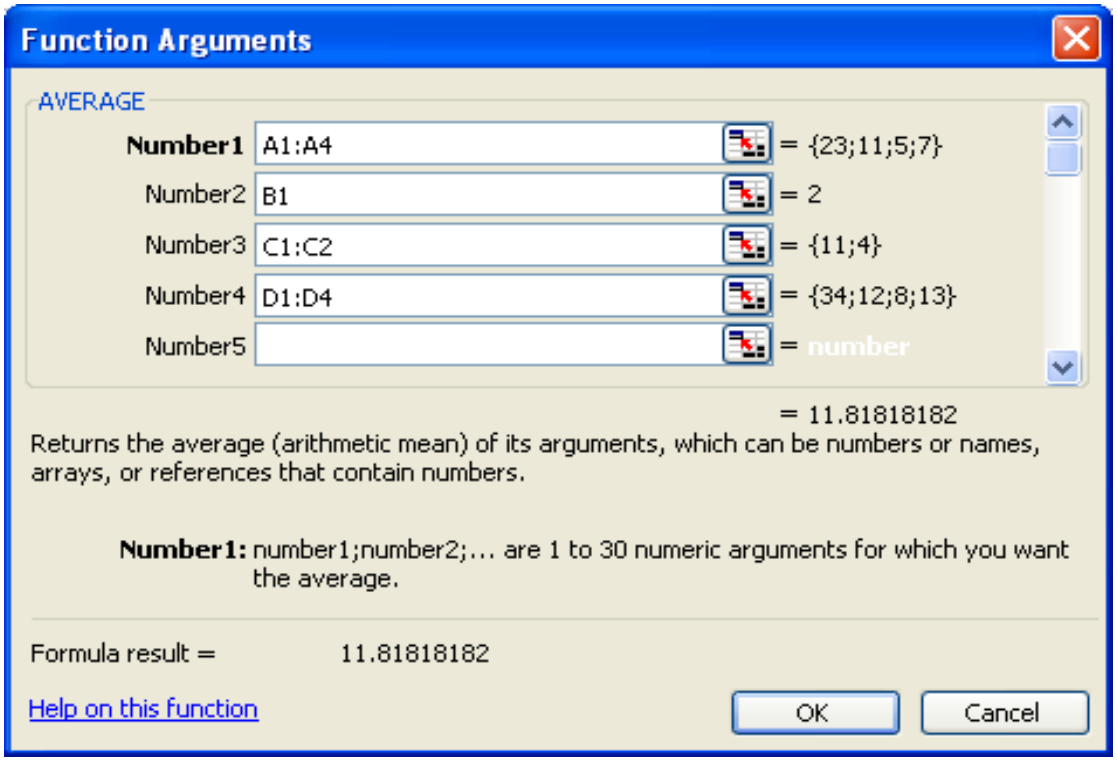
The semicolon (**;**) connects separate arrays

Measures of Central Tendency in Excel

Or, use the wizard: **Insert function…**

Select
Category: *Statistics*
Function: *Average*

	A	B	C	D
1	23	2	11	34
2	11		4	12
3	5			8
4	7			13



Activate the command line in the box and paint an array in Excel sheet

Measures of Central Tendency in Excel

- Mode

```
=MODE(B2:B5;D2:E4)
```

- Median

```
=MEDIAN(B2:B5;D2:E4)
```

- Quartiles

```
=QUARTILE((B2:B5;D2:E4);1)
```

1., 2. or 3. quartile

(the 2. = Median)

- Percentiles

```
=PERCENTILE((B2:B5;D2:E4);0.35))
```

Measures of Dispersion in Excel

- Average deviation **=AVEDEV(B2:B5;D2:E4)**
- Variance **=VAR(B2:B5;D2:E4)**
- Standard Deviation **=STDEV(B2:B5;D2:E4)**
- Skewness **=SKEW(B2:B5;D2:E4)**

Classification (Grouping) of Data

In classification we arrange a large sample of data into classes

26.6	22.9	25.8	23.1	23.8	20.9	25.2	26.9	22.6	27.9
28	23.7	23.9	20.5	25.7	27.1	24.7	17.8	23.9	22.8
22	23.6	27.5	30.6	21.6	19	22.7	26.9	25.5	27.6
27.5	22.1	26.7	27.5	28.3	31.1	32.1	28.8	21.8	23.3

There are some rules usually followed when arranging classes

- The classes should be of equal size (if possible)
- All data values from the original table need to be included in one and only in one class
- The number of classes should be between 5 and 15.

class	frequency
< 19	2
19 - 21.5	3
21.5 - 24	15
24 - 26.5	5
26.5 - 29	13
> 29	3

Classification in Excel

class	upper limits (bins)	f
< 19	19	2
19 - 21.5	21.5	2
21.5 - 24	24	15
24 - 26.5	26.5	5
26.5 - 29	29	13
> 29		3

The frequencies indicate the number of observations in the data array that are more than the upper limit in the previous row but less than or equal to the upper limit in this row

Activate the (whole) frequency column and write the formula

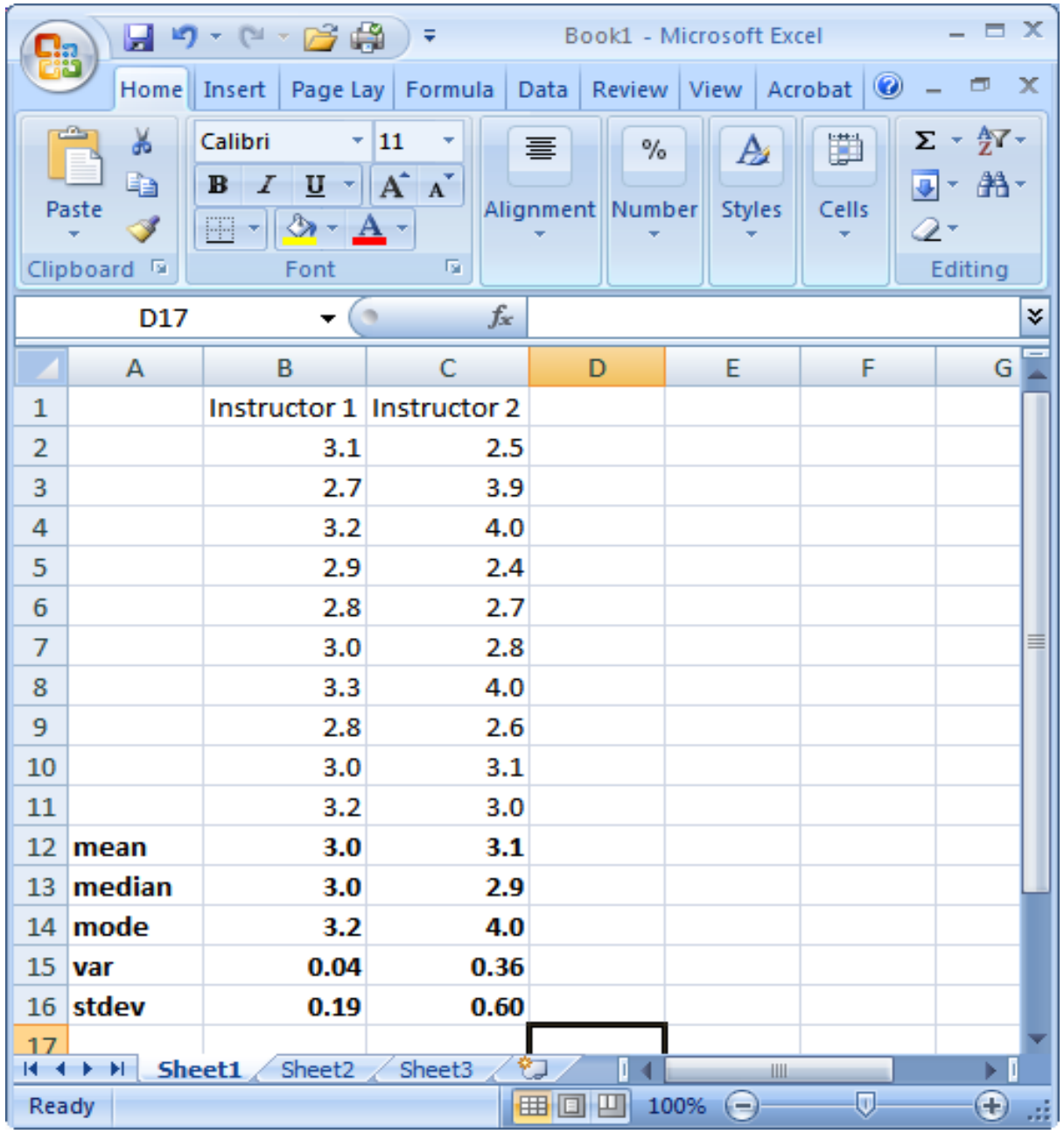
=FREQUENCY(data;bins)

into the first cell.
Remark! This is an **array formula**, which means that we have to accept the formula by pressing:

shift + ctrl + enter

Functions for Descriptive Statistics

- Your Excel spreadsheet should now look like this:



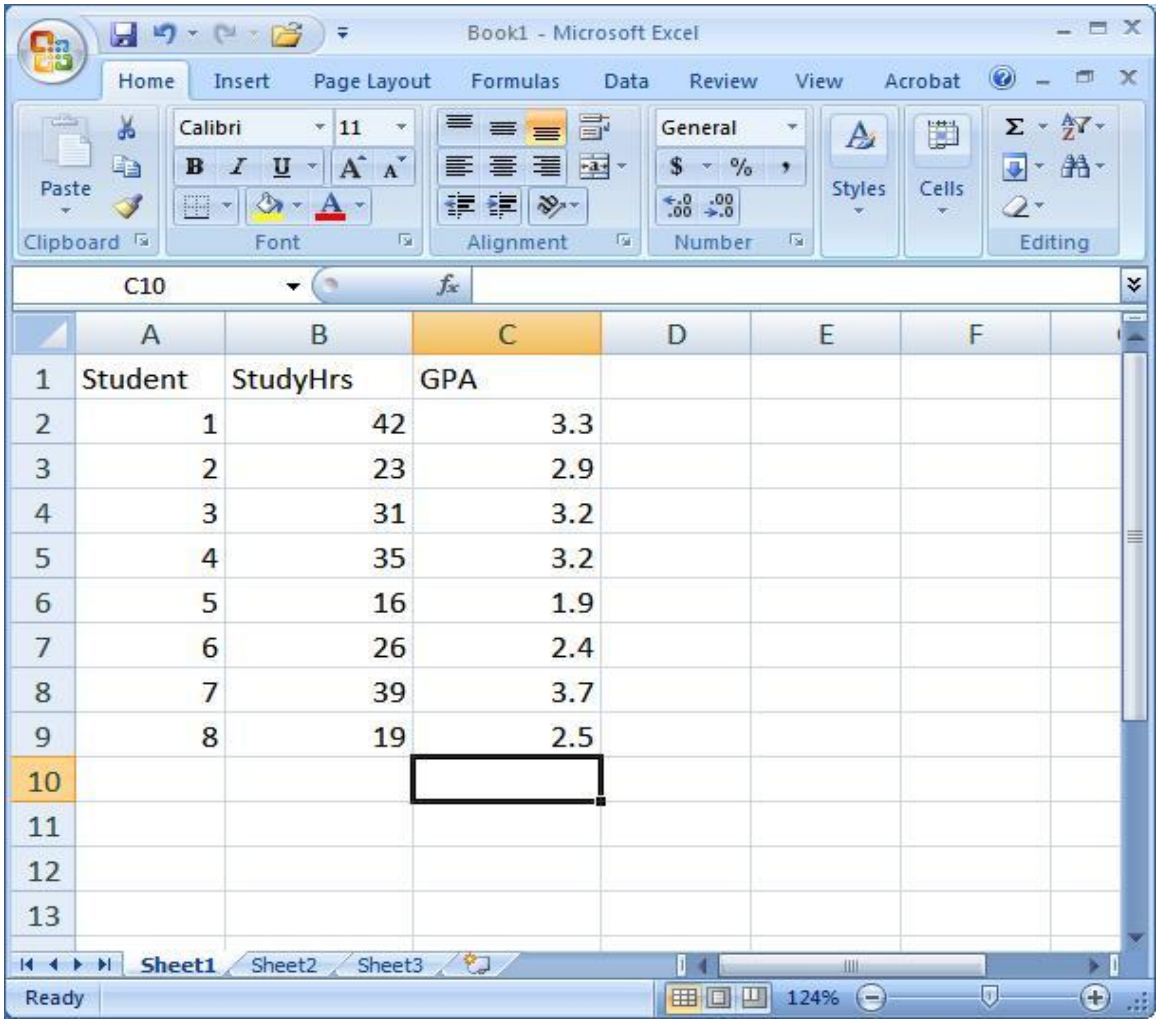
	A	B	C	D	E	F	G
1		Instructor 1	Instructor 2				
2		3.1	2.5				
3		2.7	3.9				
4		3.2	4.0				
5		2.9	2.4				
6		2.8	2.7				
7		3.0	2.8				
8		3.3	4.0				
9		2.8	2.6				
10		3.0	3.1				
11		3.2	3.0				
12	**mean**	**3.0**	**3.1**				
13	**median**	**3.0**	**2.9**				
14	**mode**	**3.2**	**4.0**				
15	**var**	**0.04**	**0.36**				
16	**stdev**	**0.19**	**0.60**				
17							

Calculating Pearson's *r*

- Correlations are described using the Pearson Product-Moment correlation statistic, or *r* value.
- In Excel, there are many functions that can calculate a correlation statistic, however, we will only use =PEARSON in this class.

Let's say we want to determine if there is a relationship between number of hours spent per week studying for Statistics and GPA (score) earned in the class at the end of the quarter. To do so, we can calculate Pearson's *r* for our two variables.

Enter the following data into Excel:



	A	B	C	D	E	F
1	Student	StudyHrs	GPA			
2	1	42	3.3			
3	2	23	2.9			
4	3	31	3.2			
5	4	35	3.2			
6	5	16	1.9			
7	6	26	2.4			
8	7	39	3.7			
9	8	19	2.5			
10						
11						
12						
13						

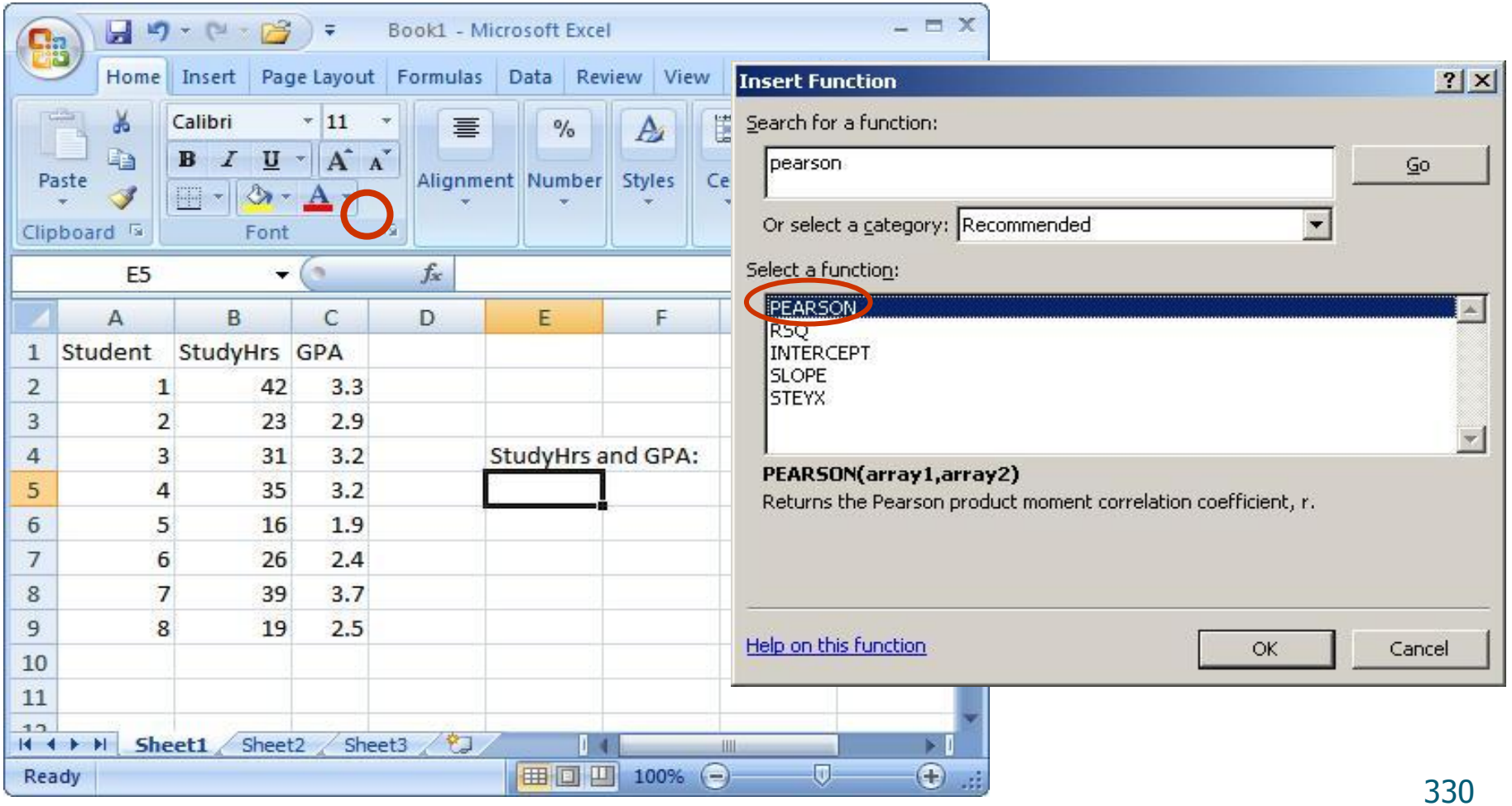
StudyHrs = average number of hours spent per week studying for 209

GPA = grade-point average earned in 209 at the end of the quarter

Step 1: Select the cell where you want your *r* value to appear (you might want to label it).

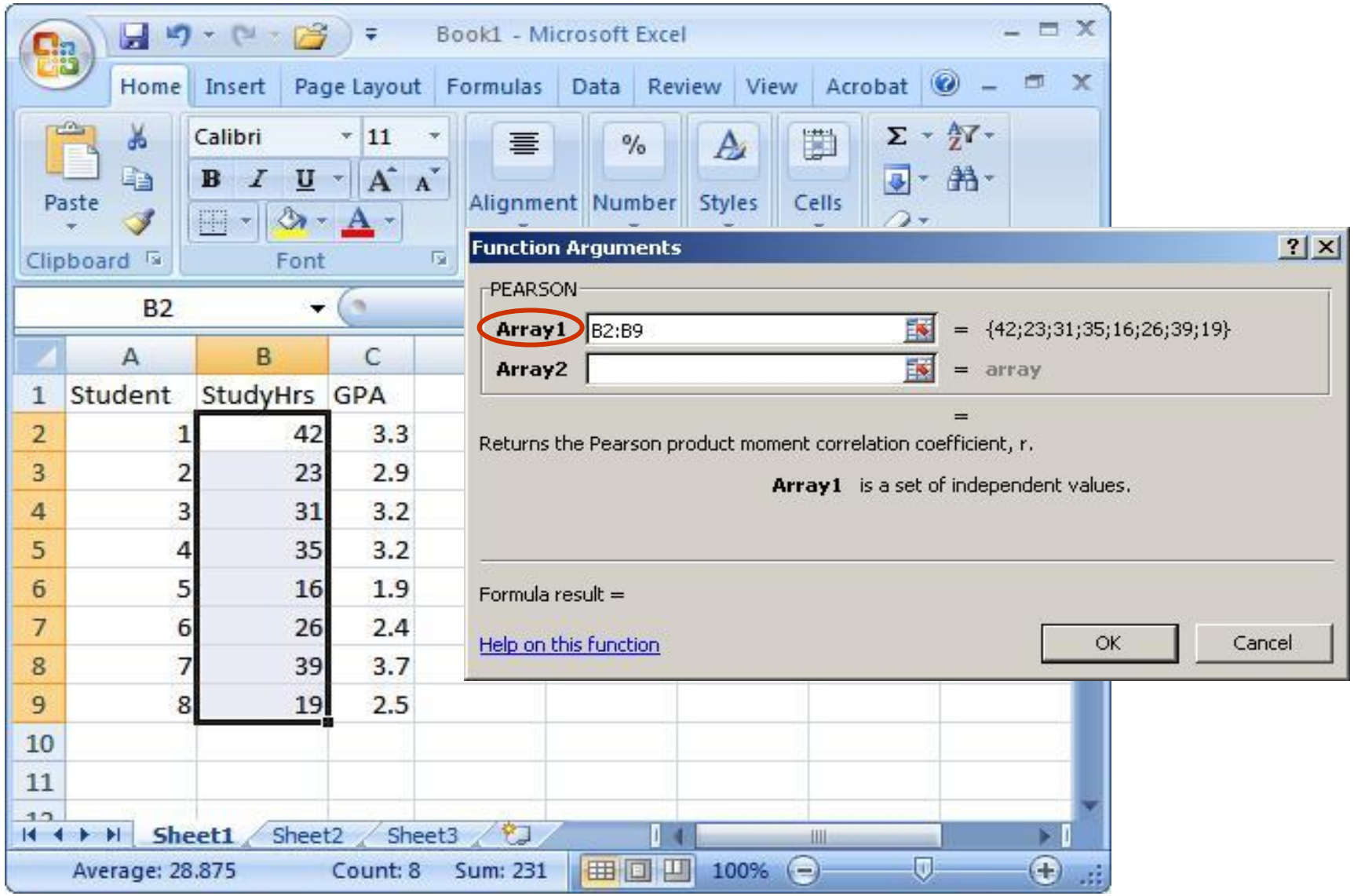
Step 2: Click on the function wizard f_x button.

Step 3: Search for and select PEARSON.

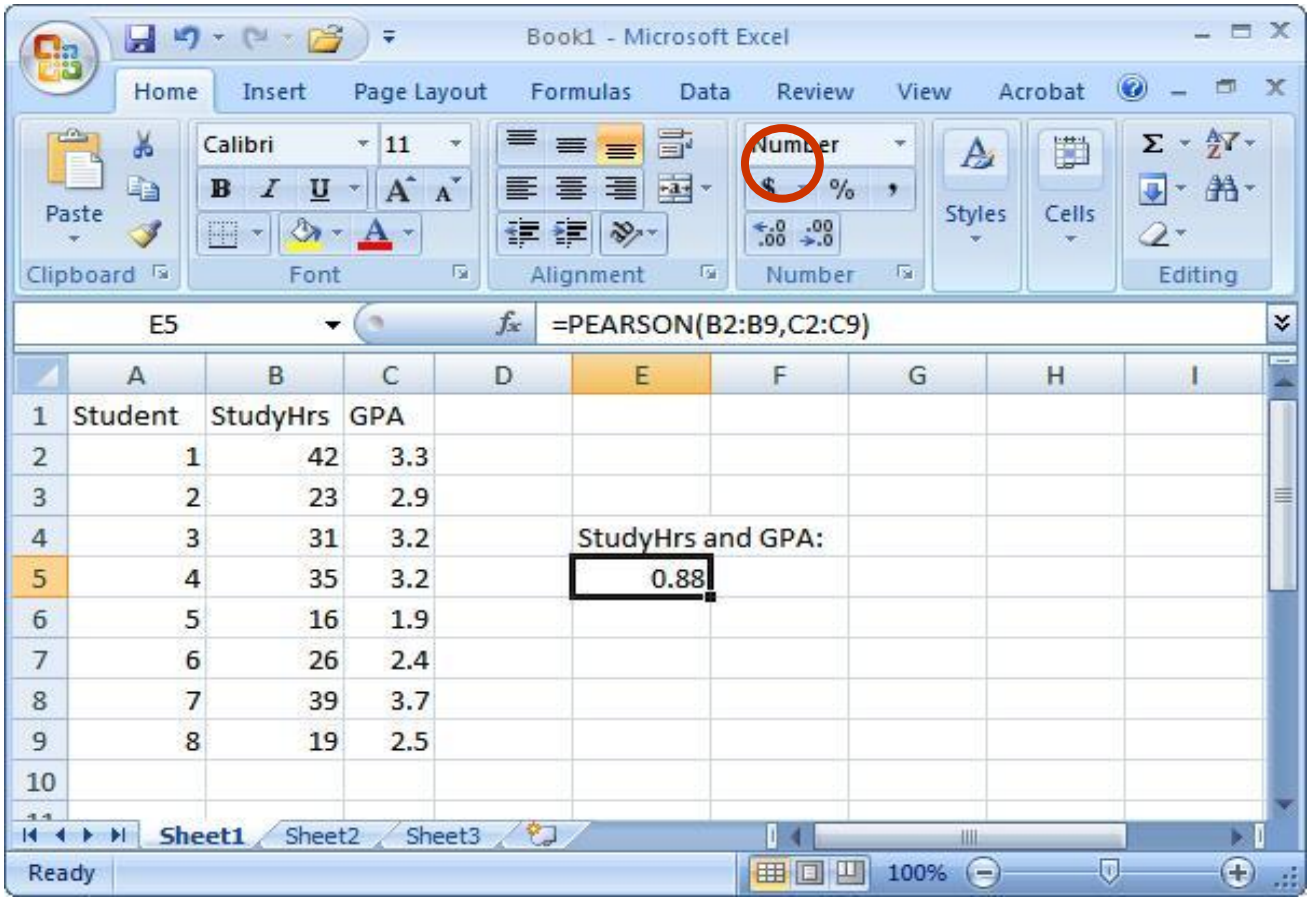


Step 4: For Array1, select all the values under StudyHrs.

For Array2, select all the values under GPA.



Step 5: That's it! Once you have your *r* value, don't forget to round to 2 decimal places.



Knowledge check: What does the *r* value of 0.88 tell you about the strength and direction of the correlation between StudyHrs and GPA?

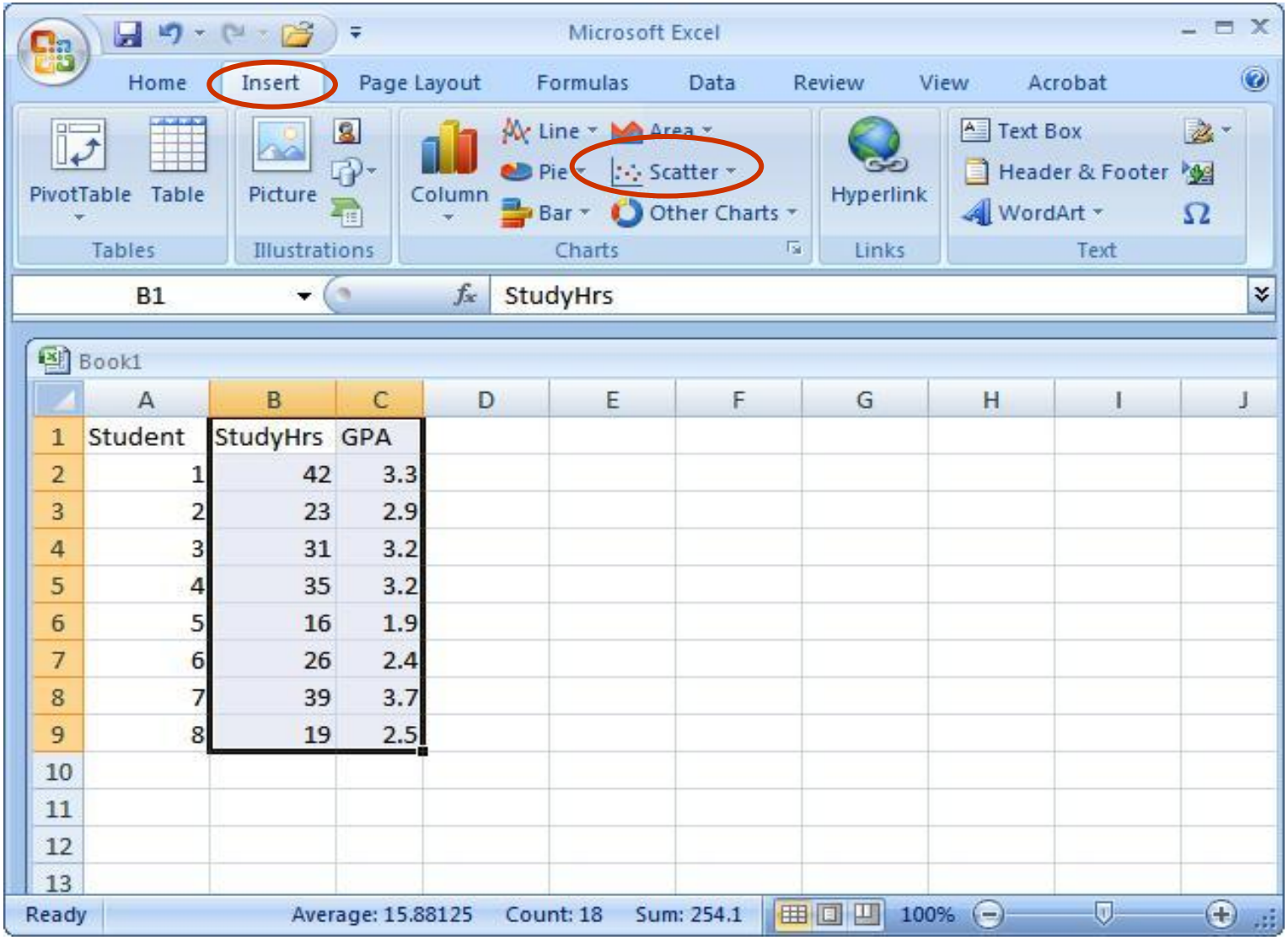
Scatterplots

- A scatterplot is an excellent way to visually display the relationship (correlation) between two variables.
- Each point on the scatterplot represents an individual's data on the two variables.

- We will now create a scatterplot for StudyHrs and GPA.

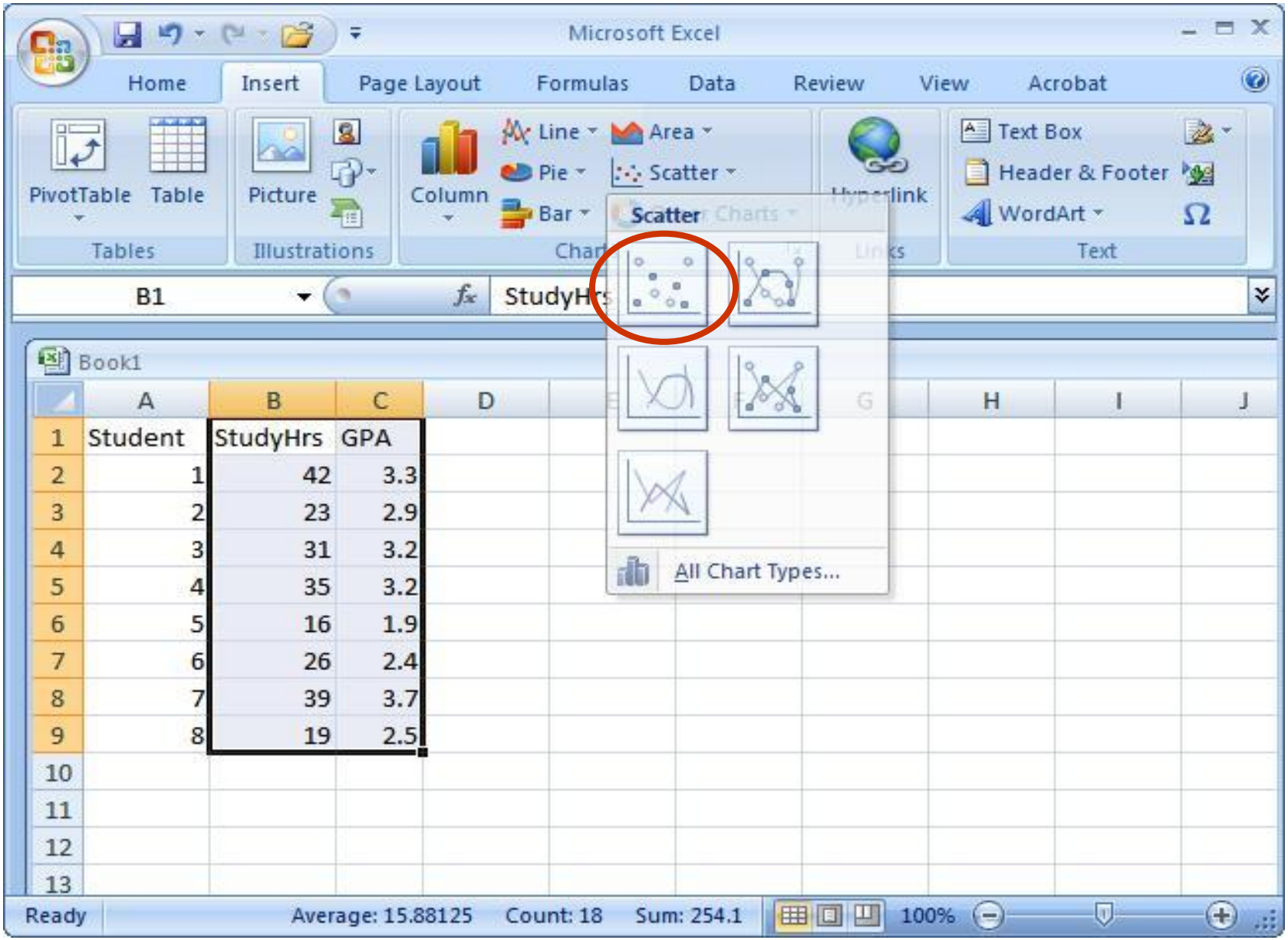
Step 1: Select <u>both</u> columns of variables you wish to plot (StudyHrs and GPA).

Step 2: Click on the tab labeled 'Insert', and then select 'Scatter' in the 'Charts' menu.

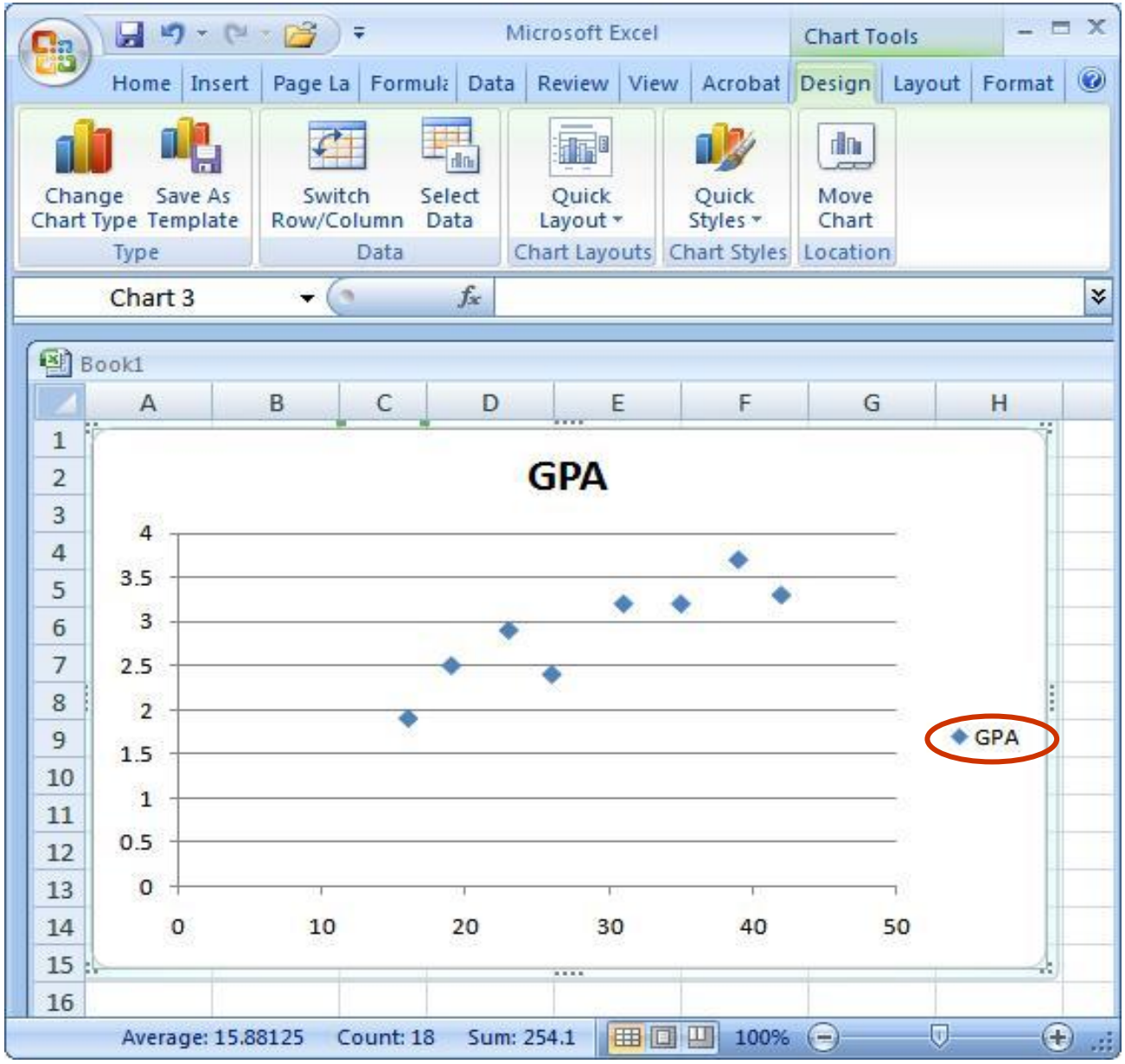


	A	B	C	D	E	F	G	H	I	J
1	Student	StudyHrs	GPA							
2	1	42	3.3							
3	2	23	2.9							
4	3	31	3.2							
5	4	35	3.2							
6	5	16	1.9							
7	6	26	2.4							
8	7	39	3.7							
9	8	19	2.5							
10										
11										
12										
13										

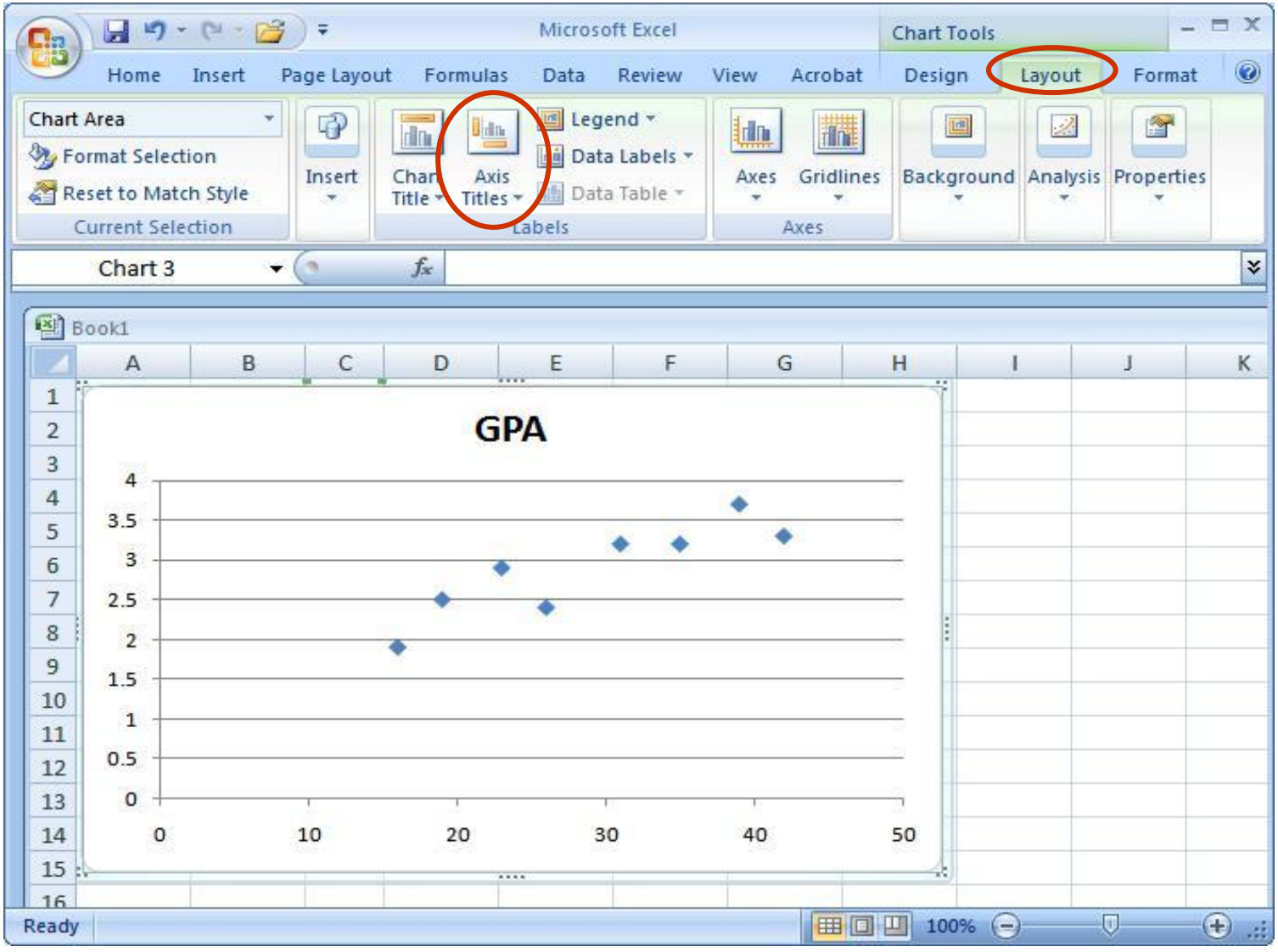
Step 3: Select the first plot in the drop-down menu.



Step 4: Remove the legend by clicking on it and pressing Delete.



Step 5: Add axis titles by selecting the 'Layout' tab and clicking on 'Axis Titles.' For the horizontal title, you want it below the x-axis. For the vertical title, you want the 'Rotated Title' option.

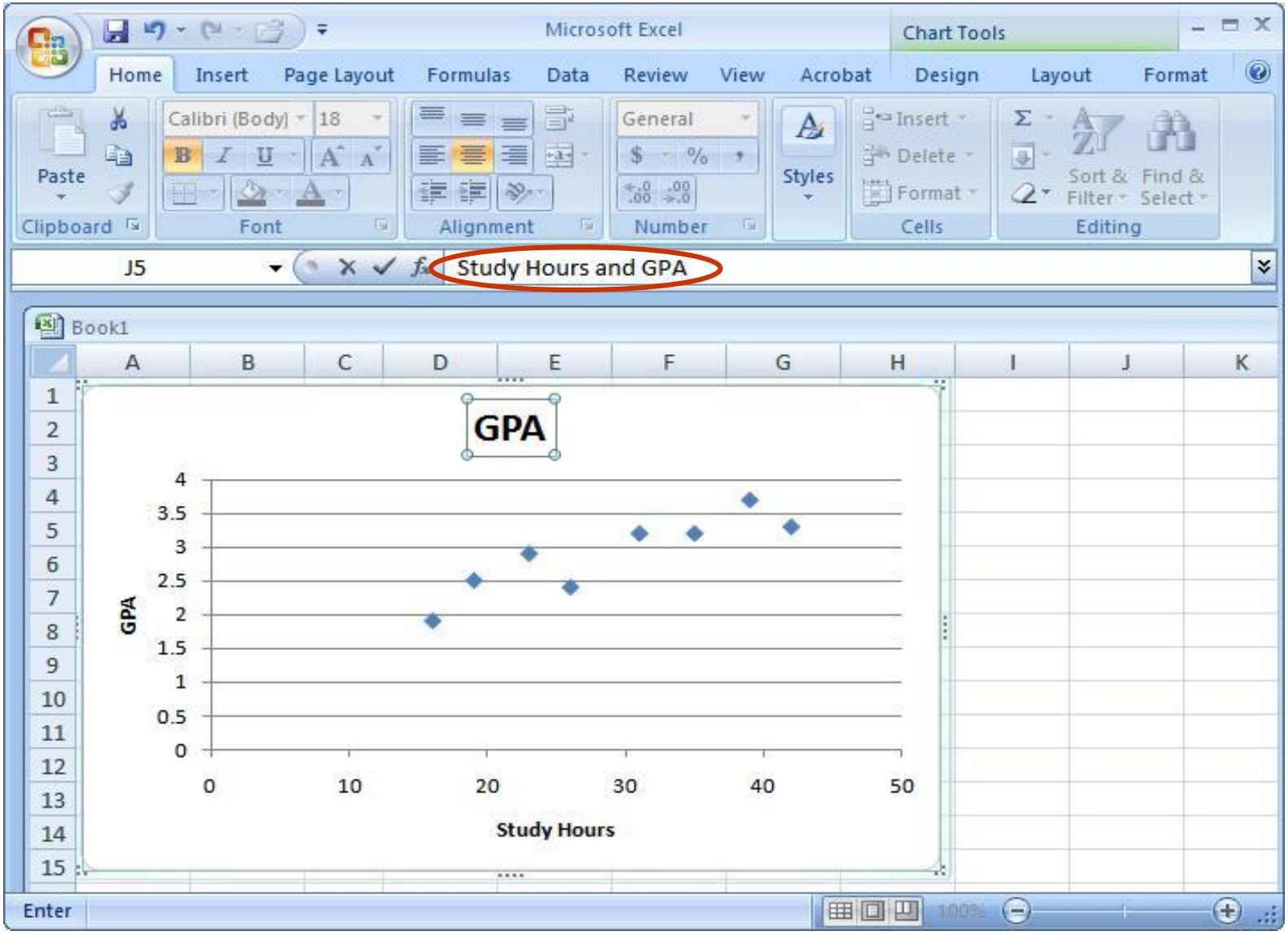


NOTE: Your chart must be highlighted for the 'Layout' tab to appear under 'Chart Tools.'

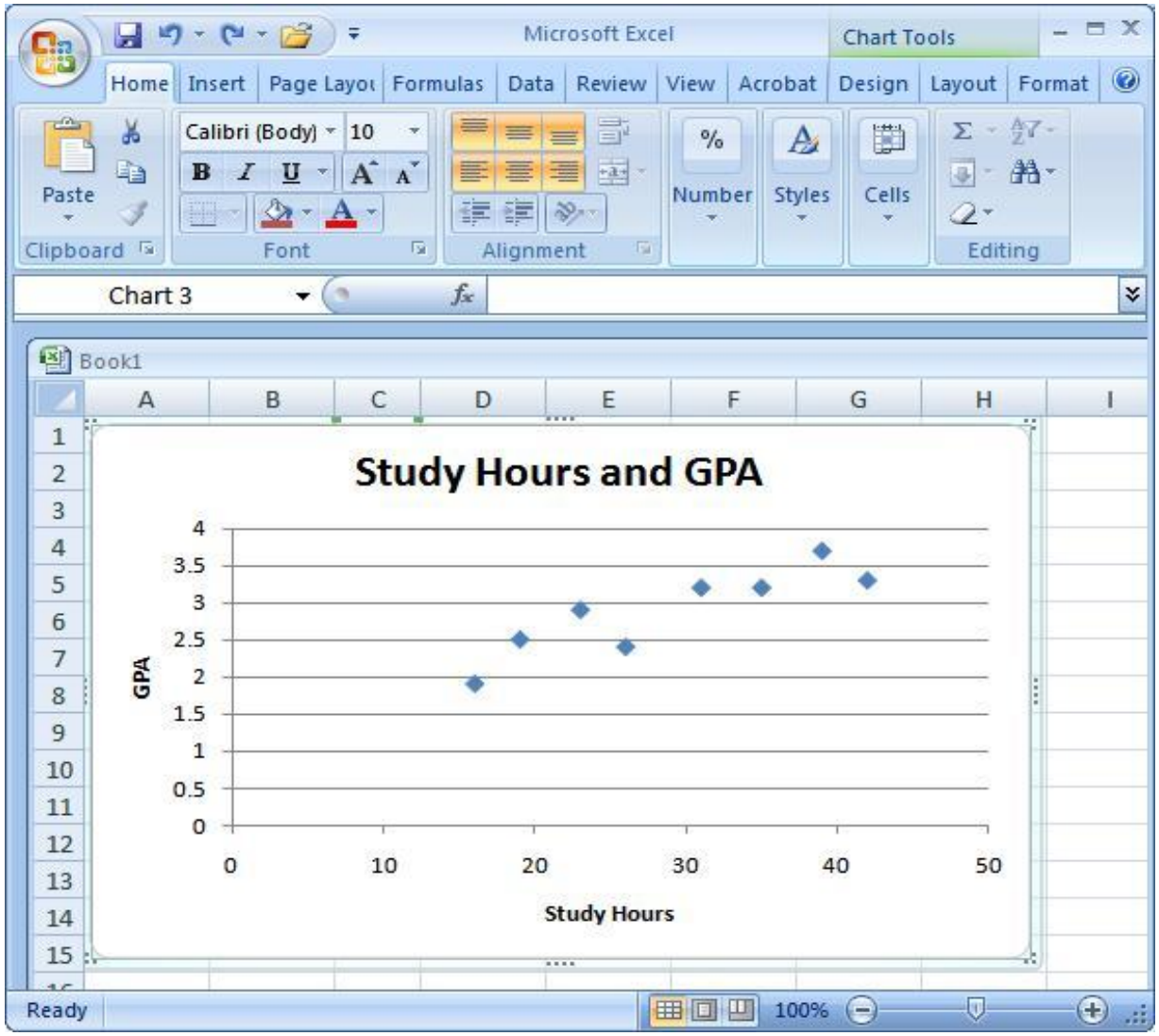
A note about x- and y-axes:

- For scatterplots, it does not matter which variable goes on each axis (this is NOT true for other types of charts).
- However, you need to make sure you label your axes with the proper variable name.
- In this example, GPA is on the y-axis and Study Hours is on the x-axis (we can tell this based on their different ranges of values).
- As a helpful hint, Excel will automatically put the first variable (left-hand column) on the x-axis, and the second variable (right-hand column) on the y-axis.

Step 6: Change the chart title by selecting it, typing a new one, and pressing Enter. Chart and axis titles may be altered by right-clicking on them.



Your scatterplot is now finished!



Remember: Each point in the scatterplot represents an individual's data.

Knowledge check: Identify Student 8 in the scatterplot.

Describing Correlations and Scatterplots

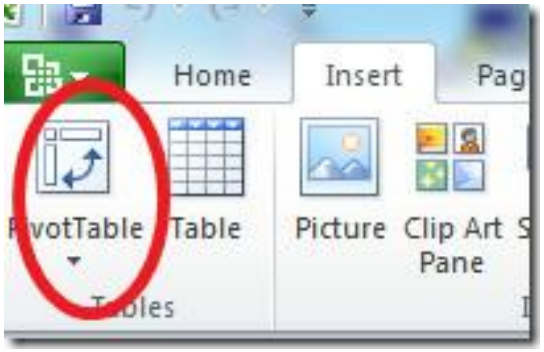
- Scatterplots and correlations are described:
  - As positive or negative.
  - As weak, moderate, or strong.
  - Using the *r* value.
  - Sentence 1: There is a strong, positive correlation ($r = 0.88$) between the number of hours studied and GPA.
- Then you want to describe the general relationship between the two variables:
  - Sentence 2: More hours of studying for Statistics was associated with a higher GPA earned in the class at the end of the quarter.
- NOTE: We cannot say "More studying led to a higher GPA" – this implies *causation*, which **cannot** be determined using correlational research.

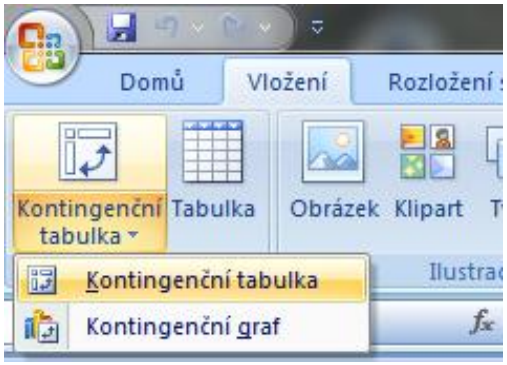
Pivot Tables (kontingeční tabulky)

Step 1: First of all, please make sure to select the data range for which you want to make the pivot table.

	55178	0.0999	
	55179	0.0999	
	55186	0.0999	
	55187	0.0999	
	55188	0.0999	
	55189	0.0999	
	55226	0.0999	
	55227	0.0999	
	55228	0.0999	
	55229	0.0999	
	55246	0.0999	
	55247	0.0999	
	55248	0.0999	
	55249	0.0999	
	55286	0.0999	
	55287	0.0999	
	55288	0.0999	
	55289	0.0999	

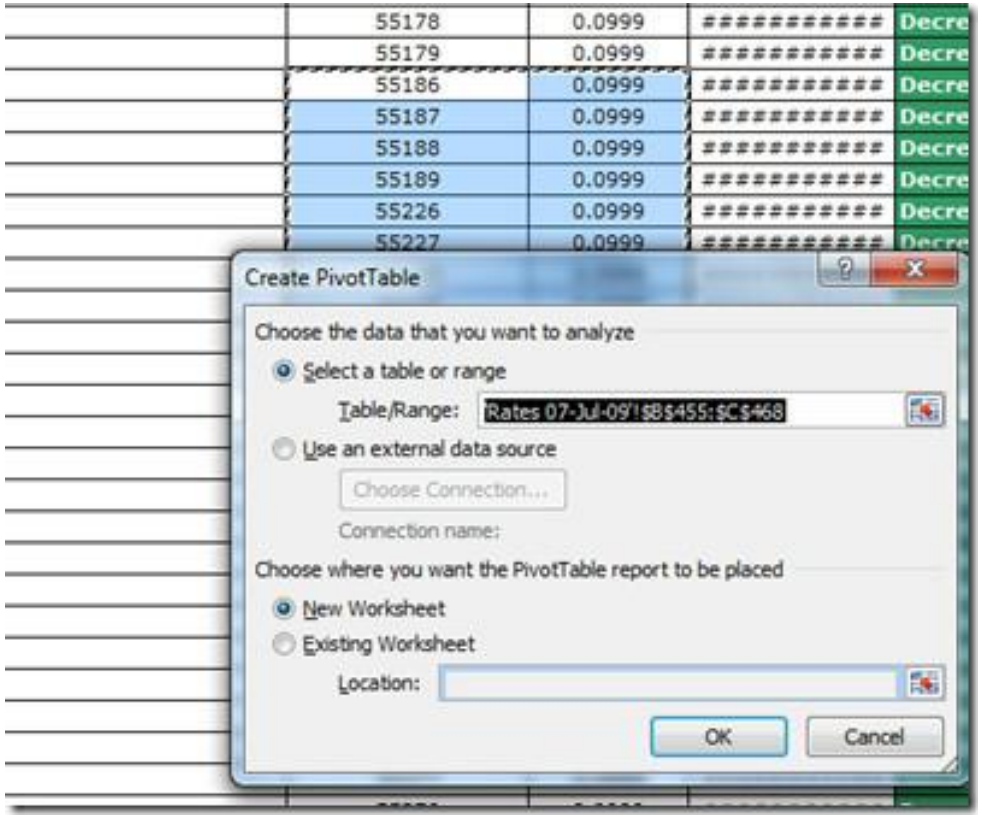
Step 2: Insert the Pivot Table by going to the *Insert* tab and then clicking the *Pivot Table* icon.



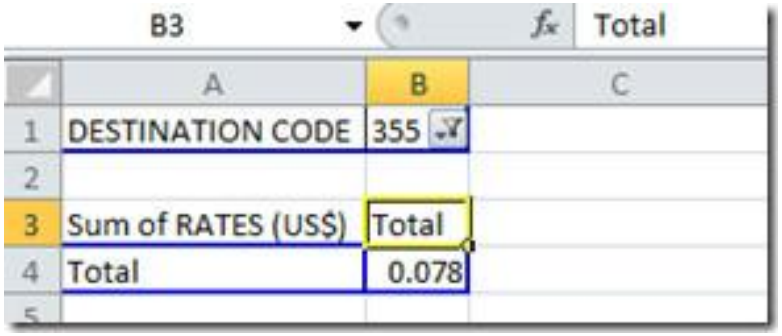


Pivot Tables (kontingeční tabulky)

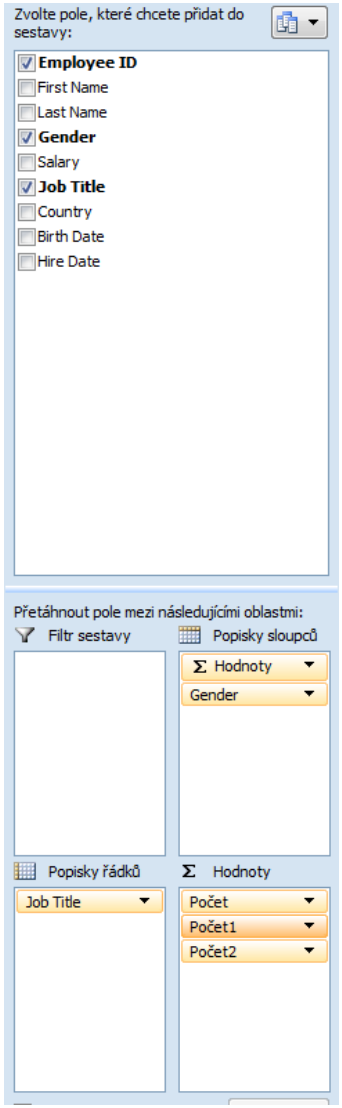
Step 3: Select the target cells where you want to place the pivot table. For starters, select the *New Worksheet* option.

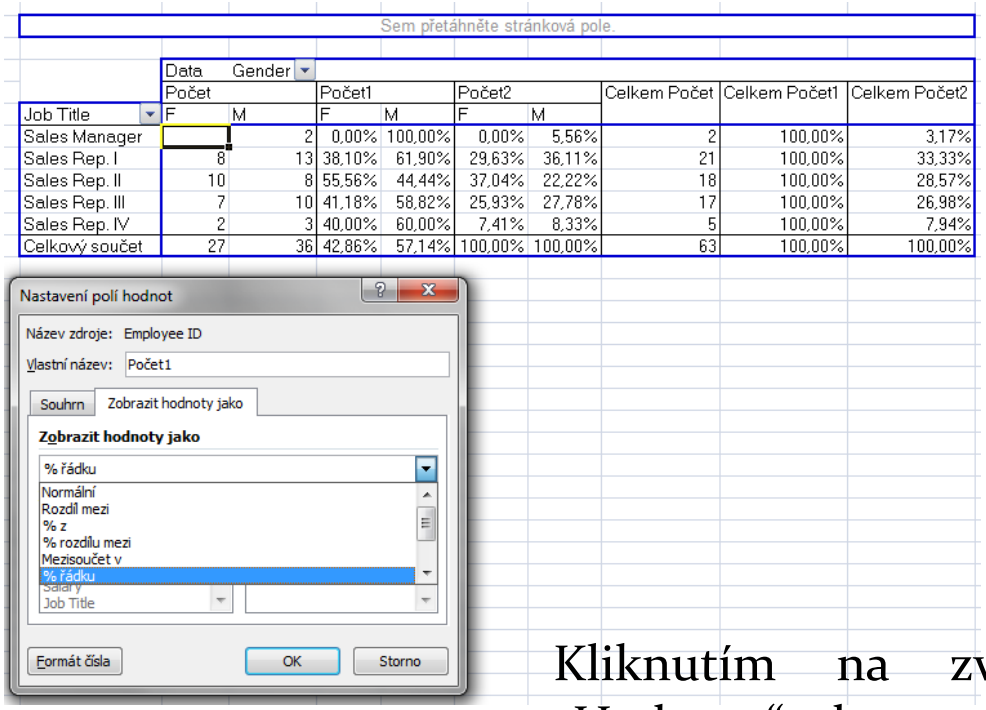


Step 4: The new worksheet will open and you will be able to see the pivot table that you just created, you can now generate the report from this table and can perform various operations on this table for better visualization and presentation of data. Just for example I calculated the sum of all of the selected cells.



Pivot Tables (kontingeční tabulky)



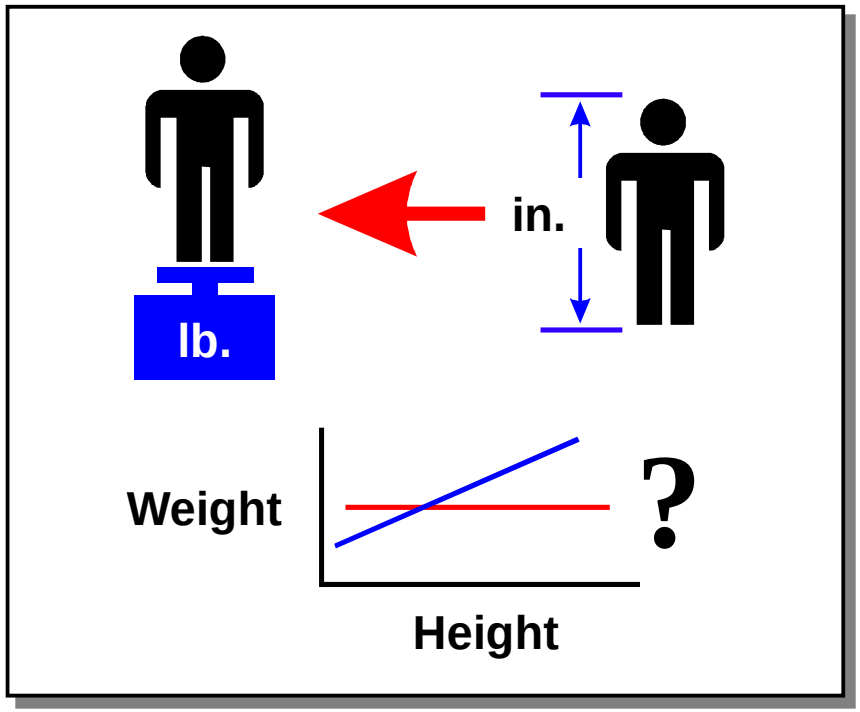


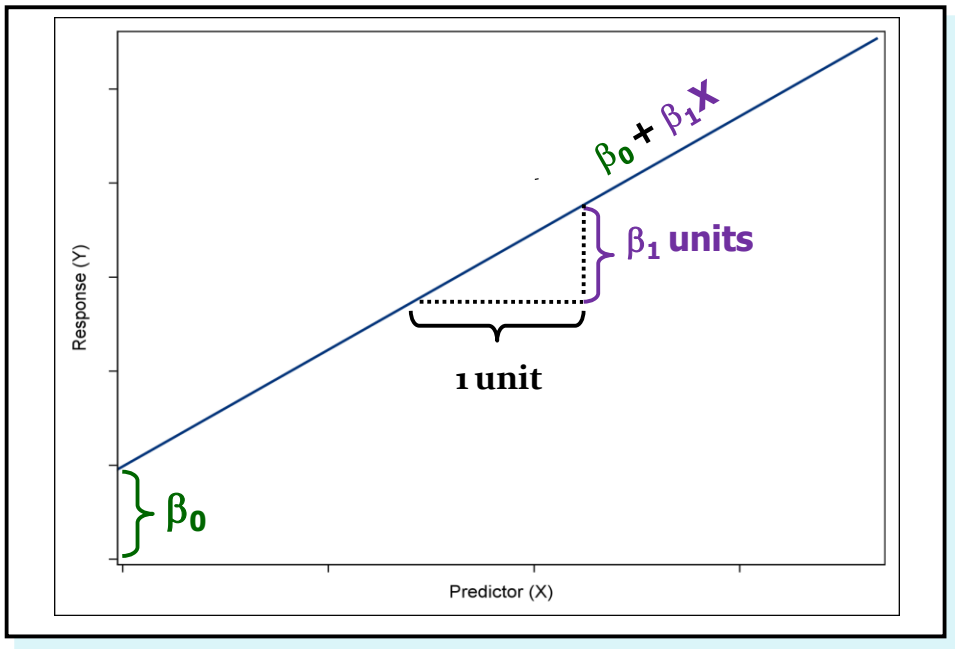
Sem přetáhněte stránková pole.

	Data Gender								
	Počet		Počet1		Počet2		Celkem Počet	Celkem Počet1	Celkem Počet2
Job Title	F	M	F	M	F	M			
Sales Manager		2	0,00%	100,00%	0,00%	5,56%	2	100,00%	3,17%
Sales Rep. I	8	13	38,10%	61,90%	29,63%	36,11%	21	100,00%	33,33%
Sales Rep. II	10	8	55,56%	44,44%	37,04%	22,22%	18	100,00%	28,57%
Sales Rep. III	7	10	41,18%	58,82%	25,93%	27,78%	17	100,00%	26,98%
Sales Rep. IV	2	3	40,00%	60,00%	7,41%	8,33%	5	100,00%	7,94%
Celkový součet	27	36	42,86%	57,14%	100,00%	100,00%	63	100,00%	100,00%

Kliknutím na zvolené pole v části „Hodnoty“ lze vyvolat „Nastavení polí hodnot“, kde v „Zobrazit hodnoty jako“ lze vybrat např. „%řádku“ (v tabulce označeno jako Počet1) …tím získáme řádkově podmíněné relativní četnosti. Pro sloupcově podmíněné volíme %sloupce.

5. Regresní analýza v MS Excel a SAS





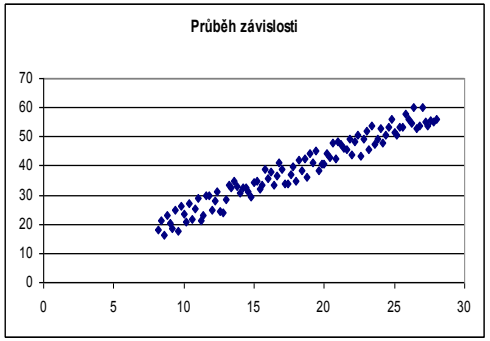
Regresní analýza

Cíl regresní analýzy: vystižení závislosti hodnot znaku Y na hodnotách znaku X.

Při tom je nutné vyřešit dva problémy:

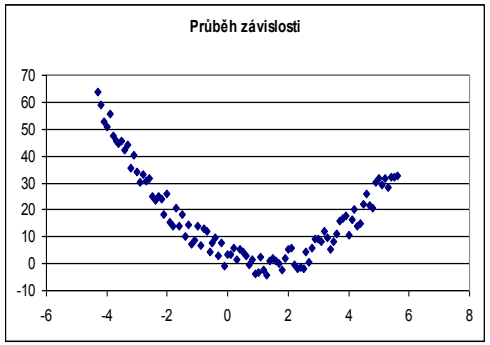
- jaký typ funkce použít k vystižení dané závislosti
- jak stanovit konkrétní parametry zvoleného typu funkce?

Typ funkce určíme buď logickým rozborem zkoumané závislosti nebo se snažíme ho odhadnout pomocí dvourozměrného tečkového diagramu.



přímka

$$y = \beta_0 + \beta_1 x$$

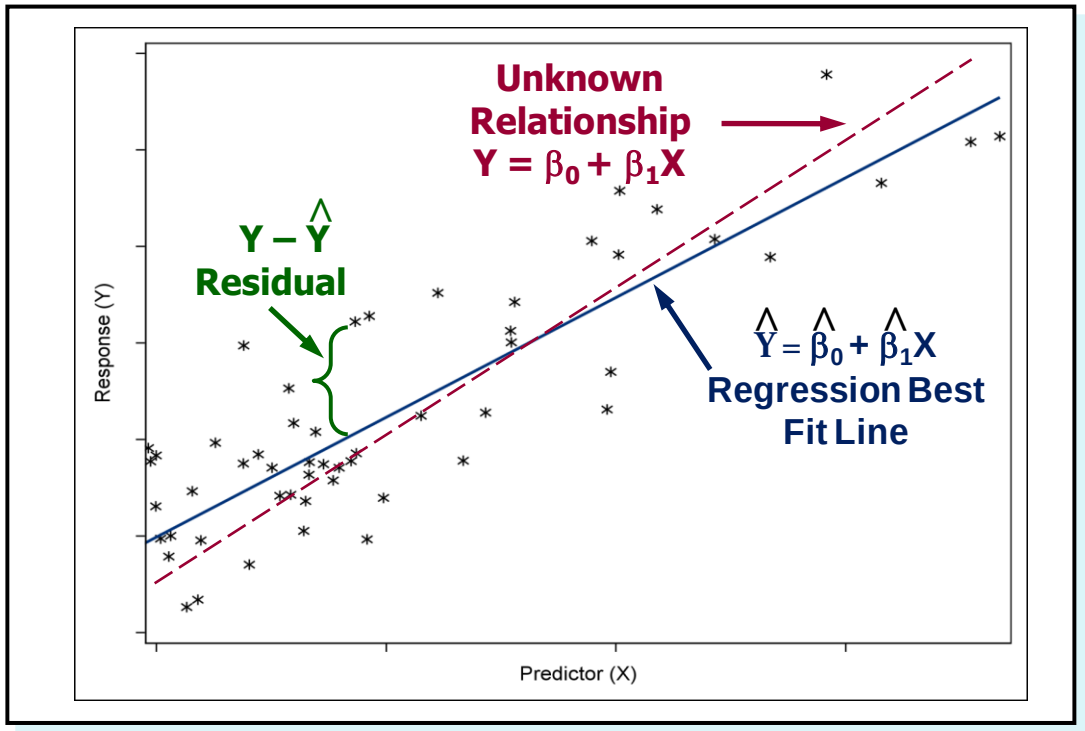


parabola

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

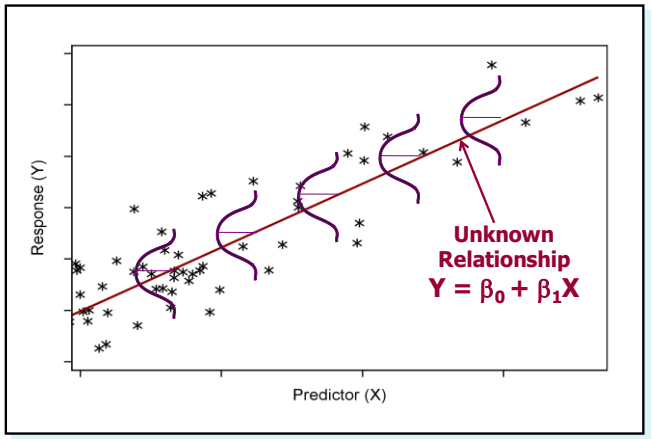
Simple Linear Regression Model

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$



Assumptions:

- **The mean of the Ys is accurately modeled by a linear function of the Xs.**
- **The random error term, ε, is assumed to have a normal distribution with a mean of zero.**
- **The random error term, ε, is assumed to have a constant variance, σ^2.**
- **The errors are independent.**



Violation of Model Assumptions

- Normality – does not affect the parameter estimates, but it affects the test results.
- Constant Variance – does not affect the parameter estimates, but the standard errors are compromised.
- Independent observations – does not affect the parameter estimates, but the standard errors are compromised.
- Linear in the parameters – indicates a misspecified model, and therefore the results are not meaningful.

Regresní přímka

Zde se omezíme na lineární závislost $y = \beta_0 + \beta_1 x$.

Odhady b_0 a b_1 neznámých regresních parametrů β_0, β_1 získáme na základě datového souboru $\begin{pmatrix} x_1 & y_1 \\ \dots & \dots \\ x_n & y_n \end{pmatrix}$ metodou nejmenších čtverců. Požadujeme, aby výraz $\frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$ nabýval svého minima vzhledem k β_0 a β_1. Tento výraz je minimální, jsou-li jeho první derivace podle β_0 a β_1 nulové. Stačí tyto derivace spočítat, položit je rovny 0 a řešit systém dvou rovnic o dvou neznámých, tzv. systém normálních rovnic.

Nechť je dán dvourozměrný datový soubor $\begin{pmatrix} x_1 & y_1 \\ \dots & \dots \\ x_n & y_n \end{pmatrix}$ a přímka $y = \beta_0 + \beta_1 x$.

Výraz $q(\beta_0, \beta_1) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$ se nazývá rozptyl hodnot znaku Y kolem přímky $y = \beta_0 + \beta_1 x$,

přímka $y = b_0 + b_1 x$, jejíž parametry minimalizují rozptyl $q(\beta_0, \beta_1)$ v celém dvourozměrném prostoru, se nazývá regresní přímka znaku Y na znak X ,

$\hat{y}_i = b_0 + b_1 x_i$, $i = 1, ..., n$ … regresní odhad i-té hodnoty znaku Y,

$r_{12}^2 = ID^2$ … index determinace (Index determinace udává, jakou část variability hodnot znaku Y vystihuje regresní přímka. Nabývá hodnot z intervalu $\langle 0, 1 \rangle$. Čím je bližší 1, tím lépe vystihuje regresní přímka závislost Y na X.)

Index determinace se definuje složitěji, pouze v případě přímky platí uvedený vztah.

Odvození odhadů regresních parametrů

Systém normálních rovnic získáme derivováním výrazu

$q(\beta_0,\beta_1)=\frac{1}{n}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2$ parciálně podle β_0 a β_1:

$$\frac{\partial q(\beta_0,\beta_1)}{\partial \beta_0}=\frac{2}{n}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)(-1)=0$$

$$\frac{\partial q(\beta_0,\beta_1)}{\partial \beta_1}=\frac{2}{n}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)(-x_i)=0$$

Systém normálních rovnic:

$$\sum y_i = nb_0 + b_1\sum x_i$$
$$\sum y_i x_i = b_0\sum x_i + b_1\sum x_i^2$$

Řešením tohoto systému získáme odhady

$$b_0=\frac{\sum_{i=1}^{n}x_i^2\sum_{i=1}^{n}y_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}x_i y_i}{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2},\quad b_1=\frac{n\sum_{i=1}^{n}x_i y_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2}$$

$$y=m_2+\frac{s_{12}}{s_1^2}(x-m_1)$$

Po jednoduchých úpravách dospějeme ke tvaru $b_1=\frac{s_{12}}{s_1^2}$, kde s_{12} je kovariance znaků X, Y a s_1^2 je rozptyl znaku X. Dále dostáváme $b_0=m_2-b_1m_1$, tedy regresní přímku můžeme vyjádřit ve tvaru $y=m_2+\frac{s_{12}}{s_1^2}(x-m_1)$.

Úsek b_0 regresní přímky udává, jaký je regresní odhad hodnoty znaku Y, nabývá-li znak X hodnoty 0. Směrnice b_1 udává, o kolik jednotek se změní hodnota znaku Y, změní-li se hodnota znaku X o jednotku. Je-li $b_1 > 0$, dochází s růstem X k růstu Y a hovoříme o přímé závislosti hodnot znaku Y na hodnotách znaku X. Je-li $b_1 < 0$, dochází s růstem X k poklesu Y a hovoříme o nepřímé závislosti hodnot znaku Y na hodnotách znaku X.

Příklad

Příklad: Pro datový soubor obsahující údaje o mezi plasticity (znak X) a mezi pevnosti oceli (znak Y)

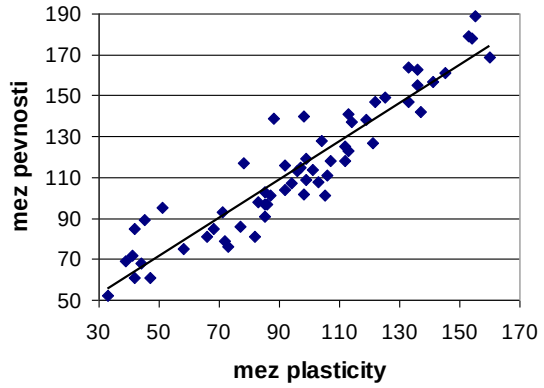
a) Určete regresní přímku meze pevnosti na mez plasticity.
b) Zakreslete regresní přímku do dvourozměrného tečkového diagramu.
c) Jak se změní mez pevnosti, vzroste-li mez plasticity o jednotku?
d) Najděte regresní odhad meze pevnosti pro mez plasticity = 60.
e) Vypočtěte index determinace a interpretujte ho.

Přitom již víme, že $m_1 = 95{,}5$, $m_2 = 114{,}4$, $s_1 = 32{,}4$, $s_2 = 32{,}5$, $s_{12} = 985{,}76$, $r_{12} = 0{,}936$.

Řešení:

ad a) $b_1 = \frac{s_{12}}{s_1^2} = \frac{985{,}76}{1052{,}4} = 0{,}937$, $b_0 = m_2 - b_1 m_1 = 114{,}4 - 0{,}937 \cdot 95{,}9 = 24{,}5$, $y = 24{,}5 + 0{,}937x$.

ad b)



ad c) Mez pevnosti vzroste o 0,937 kpcm^{-2} – viz parametr b_1 vypočtený v bodě (a)

ad d) $\hat{y} = 24{,}5 + 0{,}937 \times 60 = 80{,}72$.

ad e) $ID^2 = r_{12}^2 = 0{,}936^2 = 0{,}876$. Znamená to, že 87,6% variability hodnot meze pevnosti je vysvětleno regresní přímkou.

Příklad

x_i	y_i	$x_i * y_i$	x_i^2
100	120	12 000	10 000
90	105	9 450	8 100
86	95	8 170	7 396
94	100	9 400	8 836
120	135	16 200	14 400
135	140	18 900	18 225
79	102	8 058	6 241
62	98	6 076	3 844
110	125	13 750	12 100
125	134	16 750	15 625
1 001	**1 154**	**118 754**	**104 767**

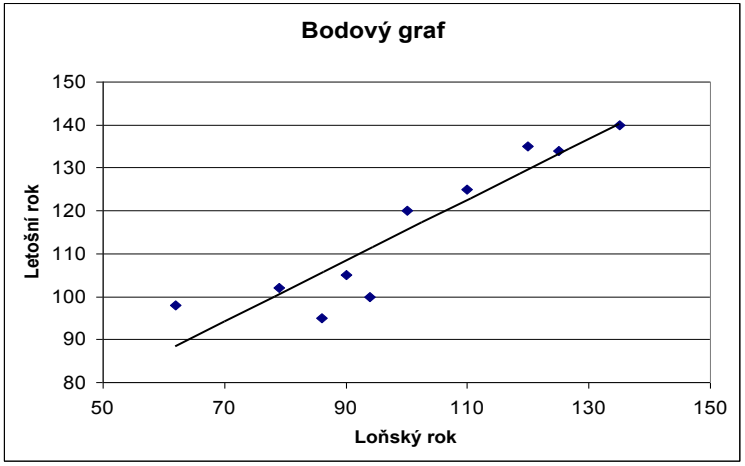
$$\sum y_i = nb_0 + b_1 \sum x_i$$

$$\sum y_i x_i = b_0 \sum x_i + b_1 \sum x_i^2$$

$$1154 = 10 \cdot b_0 + 1001 \cdot b_1$$

$$118754 = 1001 \cdot b_0 + 104767 \cdot b_1$$

$$y = 44{,}41 + 0{,}709 \cdot x$$



Příklad

Index determinace lze vyjádřit ve tvaru:

$$ID^2 = \frac{\sum \hat{y}_i^2 - \frac{1}{n} \cdot \left(\sum y_i\right)^2}{\sum y_i^2 - \frac{1}{n} \cdot \left(\sum y_i\right)^2}$$

x_i	y_i	y_i	y_i^2	$\hat{y}_i^2$
100	120	115	14 400	13 301
90	105	108	11 025	11 715
86	95	105	9 025	11 109
94	100	111	10 000	12 337
120	135	130	18 225	16 773
135	140	140	19 600	19 642
79	102	100	10 404	10 088
62	98	88	9 604	7 811
110	125	122	15 625	14 987
125	134	133	17 956	17 704
1 001	**1 154**	**1 154**	**135 864**	**135 468**

$$ID^2 = \frac{135468 - \frac{1}{10} \cdot 1154^2}{135864 - \frac{1}{10} \cdot 1154^2} = 0{,}853$$

Maticové vyjádření MNČ

$$b = \left(X^T \cdot X\right)^{-1} \cdot X^T \cdot y$$

$$b = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \qquad y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

b	sloupcový vektor 2 neznámých parametrů regresní funkce,
X	matice rozměru $n \times 2$, tvořená konstantou 1 a hodnotami znaku X
y	sloupcový vektor n hodnot znaku Y

Příklad

Nalezněte koeficienty regresní přímky:

$$\mathbf{y} = \begin{bmatrix} 120 \\ 105 \\ 95 \\ 100 \\ 135 \\ 140 \\ 102 \\ 98 \\ 125 \\ 134 \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & 100 \\ 1 & 90 \\ 1 & 86 \\ 1 & 94 \\ 1 & 120 \\ 1 & 135 \\ 1 & 79 \\ 1 & 62 \\ 1 & 110 \\ 1 & 125 \end{bmatrix}$$

$$\mathbf{X}^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 100 & 90 & 86 & 94 & 120 & 135 & 79 & 62 & 110 & 125 \end{bmatrix}$$

Příklad

$$\mathbf{g} = \mathbf{X}^T \cdot \mathbf{y} = \begin{bmatrix} 1154 \\ 118754 \end{bmatrix} \qquad \mathbf{A} = \mathbf{X}^T \cdot \mathbf{X} = \begin{bmatrix} 10 & 1001 \\ 1001 & 104767 \end{bmatrix}$$

$$\mathbf{A}^{-1} = \begin{bmatrix} 2{,}2941 & -0{,}0219 \\ -0{,}0219 & 0{,}0002 \end{bmatrix}$$

$$\Rightarrow \quad \mathbf{b} = \mathbf{A}^{-1} \cdot \mathbf{g} = \begin{bmatrix} 44{,}414 \\ 0{,}709 \end{bmatrix}$$

Sdružené regresní přímky

V některých situacích má smysl zkoumat nejenom závislost znaku Y na znaku X, ale též závislost X na Y. V takovém případě hledáme druhou regresní přímku a souhrnně hovoříme o sdružených regresních přímkách.
Regresní přímkou znaku X na znak Y nazveme tu přímku $x = \bar{b}_0 + \bar{b}_1 y$, jejíž parametry minimalizují rozptyl $q(\bar{\beta}_0, \bar{\beta}_1) = \sum_{i=1}^{n} (x_i - \bar{\beta}_0 - \bar{\beta}_1 y_i)^2$ v celé rovině. Nazývá se též druhá regresní přímka. Regresní přímka znaku Y na znak X a regresní přímka znaku X na znak Y se nazývají sdružené regresní přímky.

Rovnice regresní přímky znaku X na znak Y má tvar:

$$x = m_1 + \frac{s_{12}}{s_2^2}(y - m_2)$$

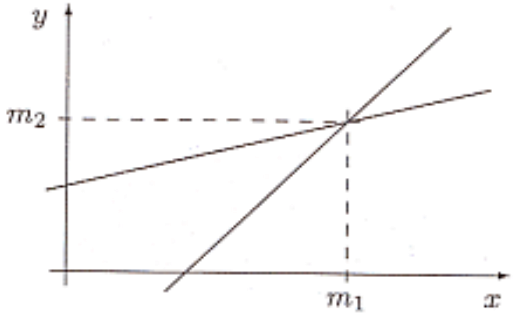
Vlastnosti sdružených regresních přímek

Sdružené regresní přímky se protínají v bodě (m_1, m_2).
Pro regresní parametry $b_1, \bar{b}_1$ platí: $b_1 \bar{b}_1 = r_{12}^2$.
Rovnice sdružených regresních přímek můžeme psát ve tvaru

$$y = m_2 + r_{12}\frac{s_2}{s_1}(x - m_1), \quad y = m_2 + \frac{1}{r_{12}}\frac{s_2}{s_1}(x - m_1) \text{ (je-li } r_{12} \neq 0).$$

Regresní přímky svírají tím menší úhel, čím méně se od sebe liší r_{12} a $\frac{1}{r_{12}}$.

Regresní přímky splynou, je-li $r_{12}^2 = 1$. K tomu dojde právě tehdy, existuje-li mezi X a Y úplná lineární závislost. Všechny body (x_i, y_i), $i = 1, ..., n$ leží na jedné přímce, tedy ze znalosti x_i můžeme přesně vypočítat y_i, $i = 1, ..., n$.
Jsou-li znaky X, Y nekorelované, pak mají sdružené regresní přímky rovnice $y = m_2$, $x = m_1$ a jsou na sebe kolmé.
Označíme-li α úhel, který svírají sdružené regresní přímky, pak platí:
$\cos \alpha = 0$, právě když mezi X a Y neexistuje žádná lineární závislost,
$\cos \alpha = 1$, právě když mezi X a Y existuje úplná přímá lineární závislost,
$\cos \alpha = -1$, právě když mezi X a Y existuje úplná nepřímá lineární závislost.



Příklad

Příklad: Pro datový soubor obsahující údaje o mezi plasticity (znak X) a mezi pevnosti oceli (znak Y):

a) Určete regresní přímku meze plasticity na mez pevnosti.

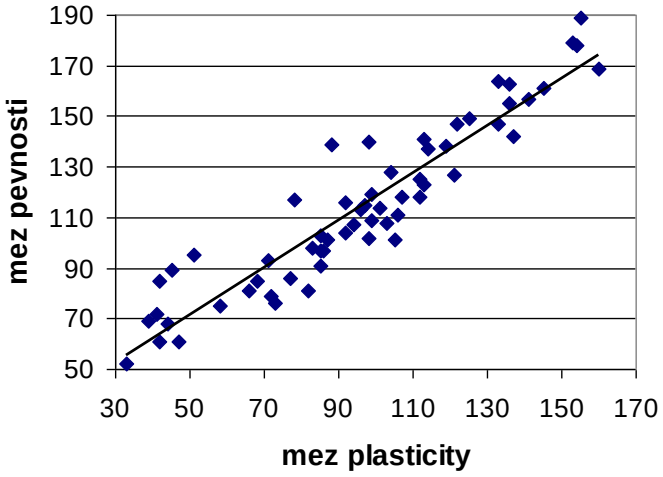
b) Zakreslete tuto druhou regresní přímku do dvourozměrného tečkového diagramu.

Přitom již víme, že $m_1 = 95{,}5$, $m_2 = 114{,}4$, $s_1 = 32{,}4$, $s_2 = 32{,}5$, $s_{12} = 985{,}76$, $r_{12} = 0{,}936$.

Řešení:

ad a) $\bar{b}_1 = \frac{s_{12}}{s_2^2} = \frac{985{,}76}{1057{,}21} = 0{,}932$, $\bar{b}_0 = m_1 - \bar{b}_1 m_2 = 95{,}9 - 0{,}932 \times 114{,}4 = -10{,}7$, tedy x = -10,7 + 0,932y.

ad b)



Příklad

Poptávka po vepřovém mase	154	164	123	181	193	105	143	167	158	62
Poptávka po hovězím mase	103	116	98	175	165	90	103	140	113	49

- Sestrojte sdružené regresní přímky.
- Vypočtěte koeficient korelace.

Příklad

x_i	y_i	$x_i * y_i$	x_i^2	y_i^2
154	103	15 862	23 716	10 609
164	116	19 024	26 896	13 456
123	98	12 054	15 129	9 604
181	175	31 675	32 761	30 625
193	165	31 845	37 249	27 225
105	90	9 450	11 025	8 100
143	103	14 729	20 449	10 609
167	140	23 380	27 889	19 600
158	113	17 854	24 964	12 769
62	49	3 038	3 844	2 401
1 450	**1 152**	**178 911**	**223 922**	**144 998**

$$m_1 = \frac{1}{10}\cdot 1450 = 145 \qquad m_2 = \frac{1}{10}\cdot 1152 = 115{,}2$$

$$s_1^2 = \frac{1}{10}\cdot 223922 - 145^2 = 1367{,}2$$

$$s_2^2 = \frac{1}{10}\cdot 144998 - 115{,}2^2 = 1228{,}76$$

$$s_{12} = \frac{178911}{10} - 145\cdot 115{,}2 = 1187{,}1$$

$$b_1 = \frac{1187{,}1}{1367{,}2} = 0{,}868$$

$$\bar{b}_1 = \frac{1187{,}1}{1228{,}76} = 0{,}966$$

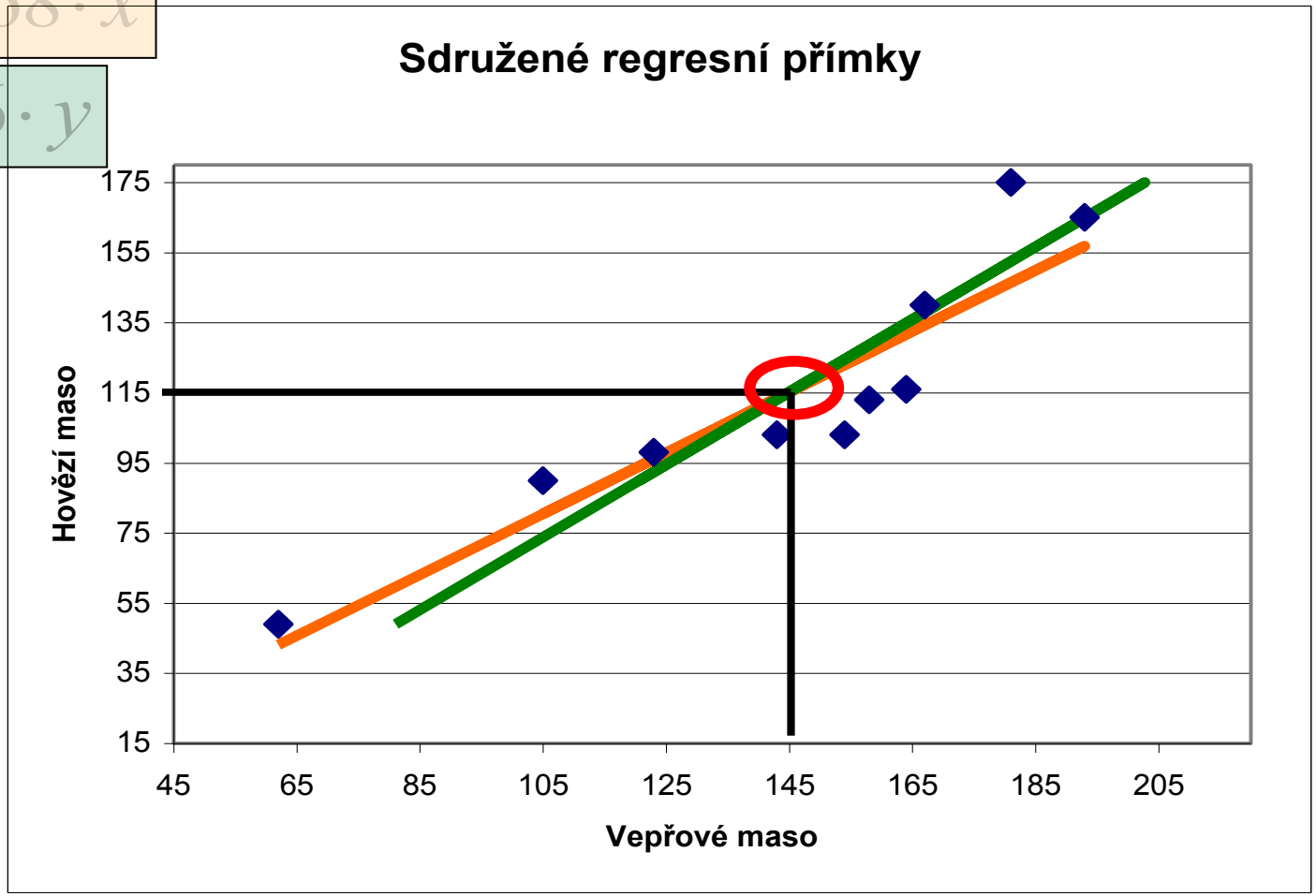
$$b_0 = 115{,}2 - 0{,}868\cdot 145 = -10{,}66$$

$$\bar{b}_0 = 145 - 0{,}966\cdot 115{,}2 = 33{,}72$$

Příklad

$$y = -10{,}66 + 0{,}868 \cdot x$$

$$x = 33{,}72 + 0{,}966 \cdot y$$



Příklad

$$r_{12} = \frac{\sum x_i y_i - n m_1 m_2}{\sqrt{\left[\sum x_i^2 - n m_1^2\right] \cdot \left[\sum y_i^2 - n m_2^2\right]}}$$

$$r_{12} = \frac{s_{12}}{s_1 \cdot s_2}$$

$$r_{12} = \mathrm{sgn}(b_1) \sqrt{b_1 \cdot \overline{b_1}}$$

Příklad

$$r_{12} = \frac{178911 - 10 \cdot 145 \cdot 115,2}{\sqrt{\left[223922 - 10 \cdot 145^2\right] \cdot \left[144998 - 10 \cdot 115,2^2\right]}} = 0,916$$

$$r_{12} = \frac{1187,1}{36,976 \cdot 35,054} = 0,916$$

$$r_{12} = \sqrt{0,868 \cdot 0,966} = 0,916$$

Příklad

x_i	y_i
154	103
164	116
123	98
52	175
193	165
105	90
143	103
167	140
158	113
191	49

- Sestrojte sdružené regresní přímky.
- Vypočtěte koeficient korelace.
- Porovnejte výsledky s výsledky předchozího příkladu.

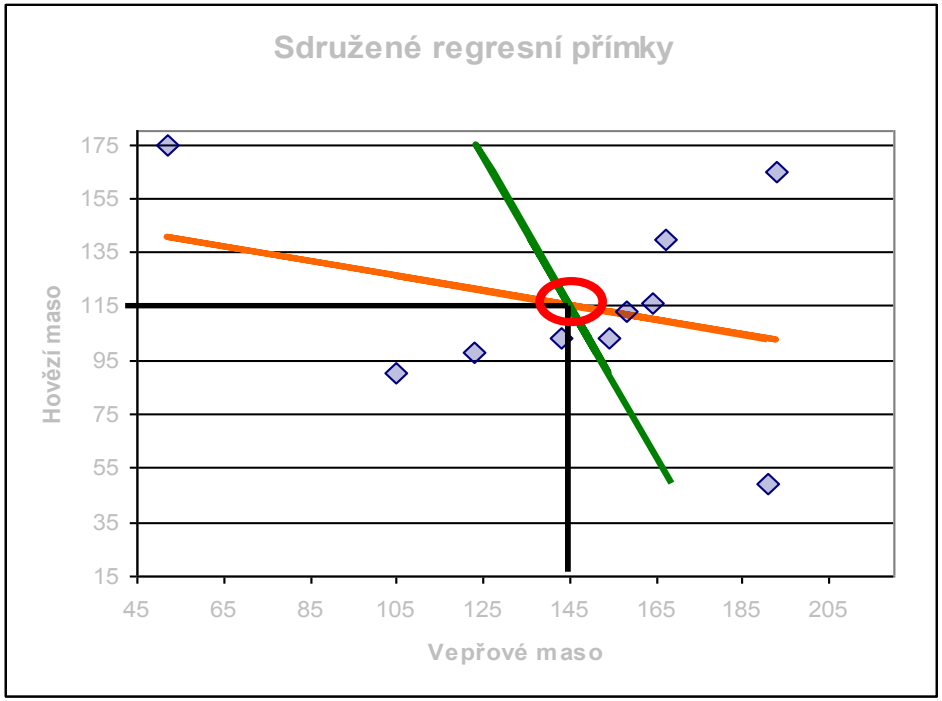
Příklad

x_i	y_i	x_i*y_i	x_i^2	y_i^2
154	103	15 862	23 716	10 609
164	116	19 024	26 896	13 456
123	98	12 054	15 129	9 604
52	175	9 100	2 704	30 625
193	165	31 845	37 249	27 225
105	90	9 450	11 025	8 100
143	103	14 729	20 449	10 609
167	140	23 380	27 889	19 600
158	113	17 854	24 964	12 769
191	49	9 359	36 481	2 401
1 450	**1 152**	**162 657**	**226 502**	**144 998**

průměry	145,0
	115,2
rozptyly	1625,2
	1228,76
směrodatné odchylky	40,3138
	35,0537

$$y = 154{,}31 - 0{,}270 \cdot x$$

$$x = 186{,}09 - 0{,}357 \cdot y$$



$$r = -\sqrt{0{,}270 \cdot 0{,}357} = -0{,}310$$

Příklad

Rozhodněte zda následující dvojice přímek mohou být sdruženými regresními přímkami:

$A)\ y = 13 - 2x$
$x = 2{,}5$

$B)\ y = 13 - 2x$
$x = 0{,}4y$

$C)\ y = 13 - 2x$
$x = 8 - y$

$D)\ y = 13 - 2x$
$x = 6{,}5 - 0{,}5y$

$E)\ y = 13 - 2x$
$x = -2 - 0{,}4y$

$F)\ y = 13 - 2x$
$x = -0{,}5y$

Příklad

$A)\ y = 13 - 2x$ $\quad x = 2{,}5$

$B)\ y = 13 - 2x$ $\quad x = 0{,}4y$

$C)\ y = 13 - 2x$ $\quad x = 8 - y$

$D)\ y = 13 - 2x$ $\quad x = 6{,}5 - 0{,}5y$

$E)\ y = 13 - 2x$ $\quad x = -2 - 0{,}4y$

$F)\ y = 13 - 2x$ $\quad x = -0{,}5y$

1. b_1 i $\bar{b}_1$ mají stejná znaménka
2. je-li jeden roven nule, pak je 0-vý i druhý
3. $r \in [-1, 1]$,tj. $b_1 \cdot \bar{b}_1 \in [0, 1]$
4. pro r = 1 (r = -1) platí $\bar{b}_0 = -\frac{b_0}{b_1}$

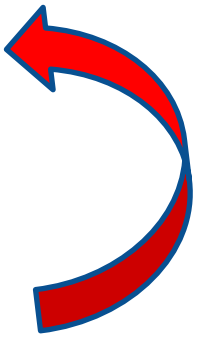
A) NE(2)
B) NE(1)
C) NE(3)

D) ANO
E) ANO
F) NE(4)

Přehled procedur SASu pro regresi

- SAS/STAT:

  CATMOD, GAM, GENMOD, GLIMMIX, GLM, LIFEREG, LOESS, LOGISTIC, MIXED, NLIN, NLMIXED, ORTHOREG, PHREG, PLS, PROBIT, REG, ROBUSTREG, RSREG, SURVEYLOGISTIC, SURVEYPHREG, SURVEYREG, TRANSREG.



„klasická“ lineární regrese

- SAS/ETS:

  AUTOREG, COUNTREG, MODEL, PANEL, PDLREG, SYSLIN.

The CORR Procedure

- S regresní analýzou souvisí analýza korelační.
- Když pro nic jiného, tak alespoň v souvislosti s explorační analýzou je vhodné prozkoumat data pomoví procedury CORR.
- General form of the CORR procedure:

```
PROC CORR DATA=SAS-data-set <options>;
        VAR variables;
        WITH variables;
        ID variables;
RUN;
```

The CORR Procedure

- Scatter plots and scatter plot matrices are available through ODS Graphics.
- ID statement enables you to specify additional variables to identify observations in scatter plots and scatter plot matrices.
- Selected options:
  - **PLOTS** <**(ONLY)**> <= *plot-request*>
  - **PLOTS** <**(ONLY)**> <= (*plot-request* < *plot-request* >) >
    - **ALL**
    - **MATRIX** <*(matrix-options)*>
    - **SCATTER** <*(scatter-options)*>
    - **HIST | HISTOGRAM**
    - **NVAR=ALL** | *n*
    - **ELLIPSE=PREDICTION | CONFIDENCE | NO**

PROC CORR –příklad výstupu

Correlations and Scatter Plots with Oxygen_Consumption

The CORR Procedure

1 With Variables:	Oxygen_Consumption	
7 Variables:	RunTime Performance	Age Weight Run_Pulse Rest_Pulse Maximum_Pulse

Simple Statistics

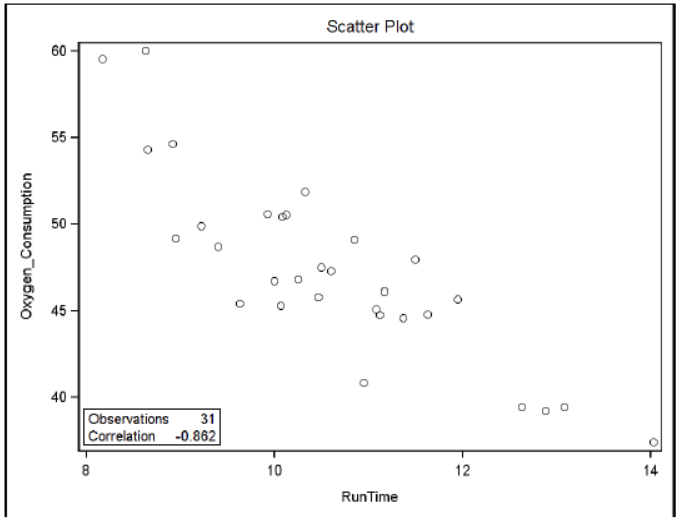
Variable	*N*	*Mean*	*Std Dev*	*Sum*	*Minimum*	*Maximum*
Oxygen_Consumption	31	47.37581	5.32777	1469	37.39000	60.06000
RunTime	31	10.58613	1.38741	328.17000	8.17000	14.03000
Age	31	47.67742	5.26236	1478	38.00000	57.00000
Weight	31	77.44452	8.32857	2401	59.08000	91.63000
Run_Pulse	31	169.64516	10.25199	5259	146.00000	186.00000
Rest_Pulse	31	53.45161	7.61944	1657	40.00000	70.00000
Maximum_Pulse	31	173.77419	9.16410	5387	155.00000	192.00000
Performance	31	56.64516	18.32584	1756	20.00000	94.00000

Pearson Correlation Coefficients, N = 31
Prob > |r| under H0: Rho=0

	RunTime	Performance	Rest_Pulse	Run_Pulse	Age	Maximum_Pulse	Weight
Oxygen_Consumption	-0.86219	0.77890	-0.39935	-0.39808	-0.31162	-0.23677	-0.16289
	<.0001	<.0001	0.0260	0.0266	0.0879	0.1997	0.3813

Correlations and Scatter Plots with Oxygen_Consumption

The CORR Procedure



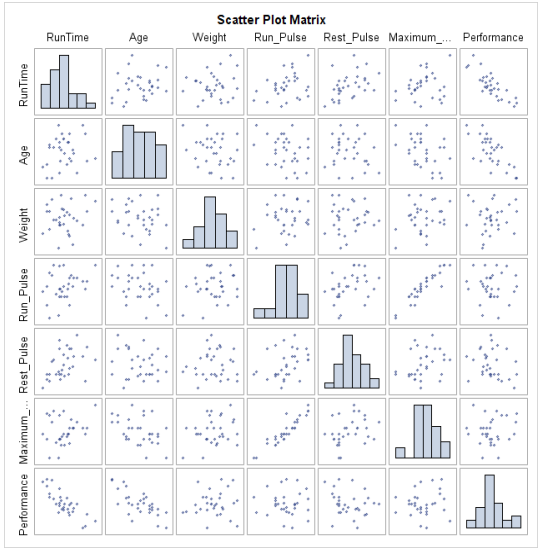
Correlations and Scatter Plot Matrix of Fitness Predictors

The CORR Procedure

7 Variables: RunTime Age Weight Run_Pulse Rest_Pulse Maximum_Pulse Performance

Pearson Correlation Coefficients, N = 31
Prob > |r| under H0: Rho=0

	RunTime	Age	Weight	Run_Pulse	Rest_Pulse	Maximum_Pulse	Performance
RunTime	1.00000	0.19523	0.14351	0.31365	0.45038	0.22610	-0.82049
		0.2926	0.4412	0.0858	0.0110	0.2213	<.0001
Age	0.19523	1.00000	-0.24050	-0.31607	-0.15087	-0.41490	-0.71257
	0.2926		0.1925	0.0832	0.4178	0.0203	<.0001
Weight	0.14351	-0.24050	1.00000	0.18152	0.04397	0.24938	0.08974
	0.4412	0.1925		0.3284	0.8143	0.1761	0.6312
Run_Pulse	0.31365	-0.31607	0.18152	1.00000	0.35246	0.92975	-0.02943
	0.0858	0.0832	0.3284		0.0518	<.0001	0.8751
Rest_Pulse	0.45038	-0.15087	0.04397	0.35246	1.00000	0.30512	-0.22560
	0.0110	0.4178	0.8143	0.0518		0.0951	0.2224
Maximum_Pulse	0.22610	-0.41490	0.24938	0.92975	0.30512	1.00000	0.09002
	0.2213	0.0203	0.1761	<.0001	0.0951		0.6301
Performance	-0.82049	-0.71257	0.08974	-0.02943	-0.22560	0.09002	1.00000
	<.0001	<.0001	0.6312	0.8751	0.2224	0.6301	



Multiple Linear Regression with Two Variables

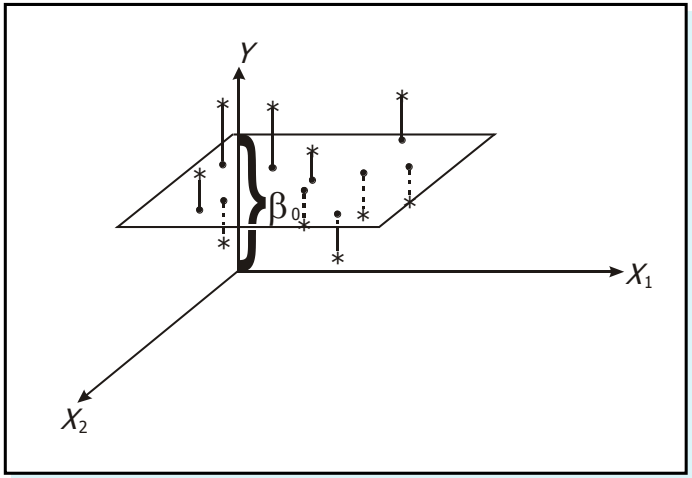
- Consider the two-variable model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

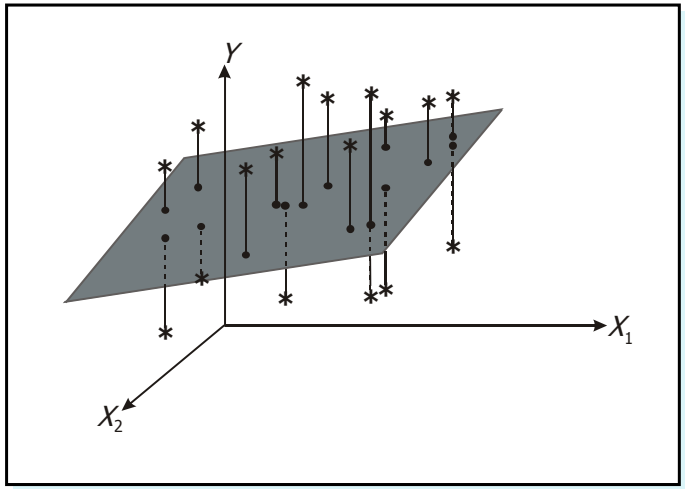
where

Y	is the dependent variable.
X_1 and X_2	are the independent or predictor variables.
ε	is the error term.
β_0, β_1, and β_2	are unknown parameters.

No relationship:



A relationship:



The Multiple Linear Regression Model

- In general, you model the dependent variable Y as a linear function of *k* independent variables, (the Xs) as

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon$$

Model Hypothesis test:

- **Null Hypothesis:**
  - The regression model does not fit the data better than the baseline model.
  - $\beta_1 = \beta_2 = \ldots = \beta_k = 0$

- **Alternative Hypothesis:**
  - The regression model does fit the data better than the baseline model.
  - Not all β_is equal zero.

Analytical Analysis vs. Prediction

Analytical Analysis:

- The focus is on understanding the relationship between the dependent variable and the independent variables.
- Consequently, the statistical significance of the coefficients is important as well as the **magnitudes** and **signs** of the coefficients.

$$\hat{Y} = \underline{\hat{\beta}_0} + \underline{\hat{\beta}_1} X_1 + \ldots + \underline{\hat{\beta}_k} X_k$$

Prediction:

- The terms in the model, the values of their coefficients, and their statistical significance are of secondary importance.
- The focus is on producing a model that is the best at predicting future values of Y as a function of the Xs. The predicted value of Y is given by $\underline{\hat{Y}} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_k X_k$

Model Selection Options

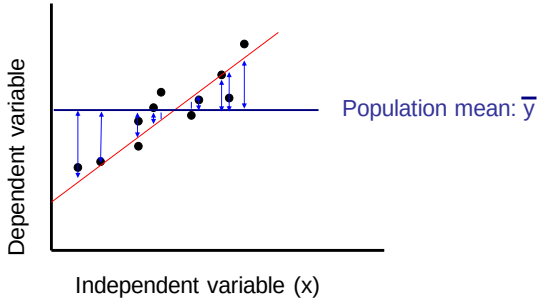
- The SELECTION= option in the MODEL statement of PROC REG supports these model selection techniques:
- **All-possible regressions ranked using**
  - RSQUARE, ADJRSQ or CP
- **Stepwise selection methods**
  - STEPWISE, FORWARD, or BACKWARD
-

SELECTION=NONE is the default.

Model Selection Statistics

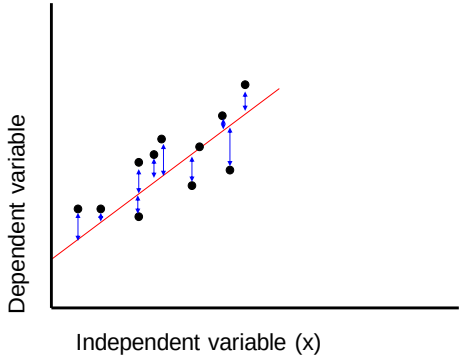
- Coefficient of determination (R^2)
- Adjusted coefficient of determination (adjusted R^2)
- Mallows' C_p statistic
- Akaike's information criteria (AIC)
- Schwarz's Bayesian criteria (SBC)

SSR = $\Sigma (\hat{y} - \bar{y})^2$



SSE = $\Sigma (y - \hat{y})$

SST = SSR+SSE



$$R^2 = 1 - \frac{SSE}{SST}$$

$$AdjR^2 = 1 - \frac{SSE / df_E}{SST / df_T} = 1 - \frac{SSE /(n-p)}{SST /(n-1)}$$

$$= 1 - \frac{(n-1)}{(n-p)}(1-R^2)$$

Information Criteria

- Akaike's information criteria (AIC)

$$AIC = (n)\ln(\frac{SSE}{n}) + 2p$$

- Schwarz's Bayesian criteria (SBC)

$$SBC = (n)\ln(\frac{SSE}{n}) + p\ln(n)$$

Smaller values indicate a better model.

Select Candidate Models

- Candidate models can be identified by using
  - your subject-matter knowledge
  - information gathered from data exploration
  - automatic selection criteria available in the REG procedure
    - all possible models ranked by
      - R^2, adjusted R^2, or Mallows' C_p
    - stepwise selection
      - forward, backward , stepwise, MAXR, or MINR
    - other statistics such as AIC and SBC
  - residual plots to evaluate model fit and model assumptions.

The REG Procedure

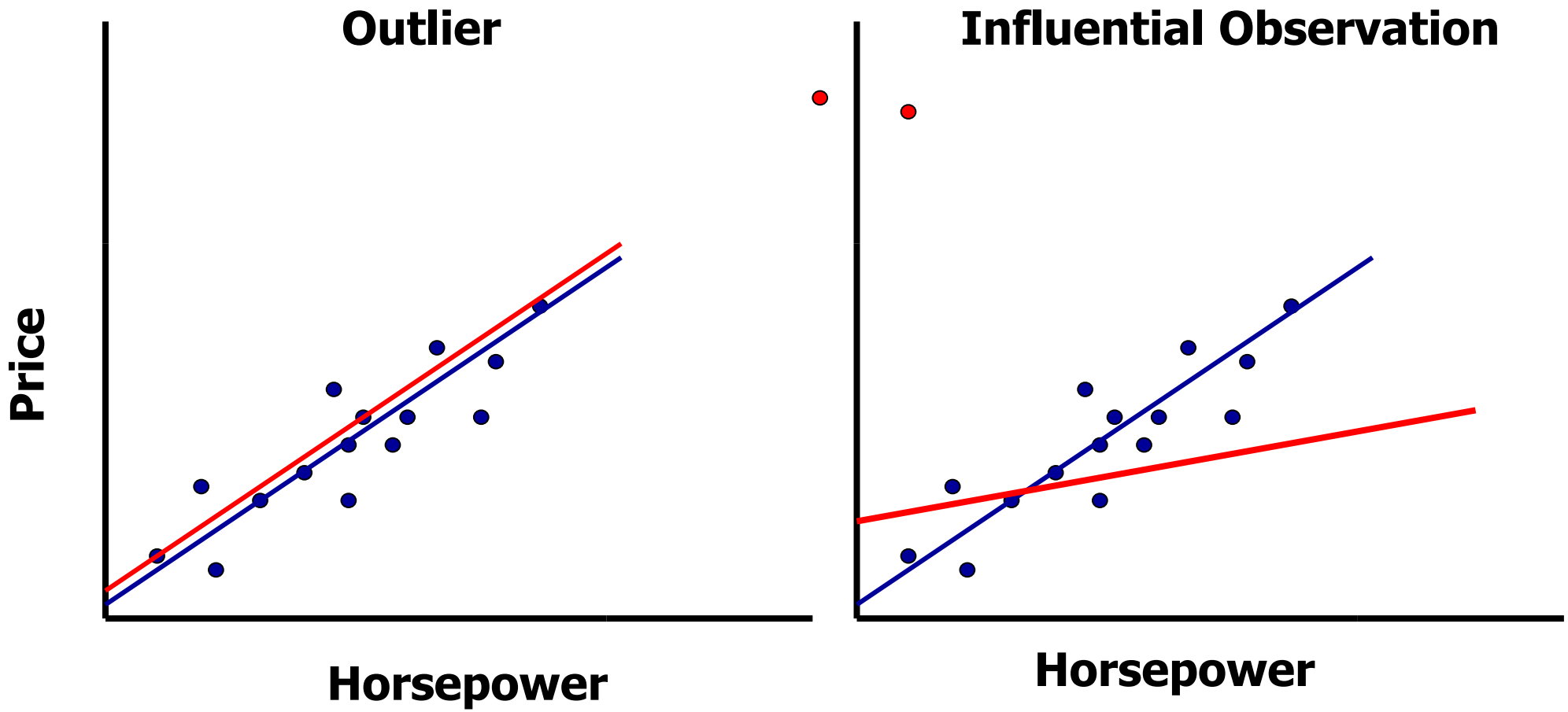
- General form of the REG procedure:

```
PROC REG DATA=SAS-data-set <options>;
      MODEL dependent(s)=regressor(s) </ options>;
RUN;
```

Popis + jednoduchý příklad:

http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_reg_sect003.htm

Influential Observations versus Outliers



Studentized Residual

- Studentized residuals (SR) are obtained by dividing the residuals by their standard errors.
- Suggested cutoffs are as follows:
  - |SR| > 2 for data sets with a relatively small number of observations
  - |SR| > 3 for data sets with a relatively large number of observations

Cook's D Statistic

- Cook's D statistic is a measure of the simultaneous change in the parameter estimates when an observation is deleted from the analysis.

- A suggested cutoff is $D_i > \frac{4}{n}$, where n is the sample size.

- If the above condition is true, then the observation might have an adverse effect on the analysis.

DFFITS

- $DFFITS_i$ measures the impact that the i^{th} observation has on the predicted value.
-

$$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_{(i)}}{s(\hat{Y}_i)}$$

$\hat{Y}_i$ is the i^{th} predicted value.

$\hat{Y}_{(i)}$ is the i^{th} predicted value when the i^{th} observation is deleted.

$s(\hat{Y}_i)$ is the standard error of the i^{th} predicted value.

Identifying Influential Observations – DFBETAs

$$DFBETA_{j(i)} = \frac{b_j - b_{j(i)}}{\hat{\sigma}(b_j)}$$

measures the change in each parameter estimate when an observation is deleted from the model.

- b_j is the parameter estimate for the j^{th} independent variable
- $b_{j(i)}$ is the parameter estimate for the j^{th} independent variable with the i^{th} observation deleted from the analysis
- $\hat{\sigma}(b_j)$ is the standard error of the j^{th} parameter estimate when all observations are included in the analysis

Identifying Influential Observations – The Covariance Ratio

$$COVRATIO_i = \frac{\left| s_i^2 \left(X_i' X_i \right)^{-1} \right|}{\left| s^2 \left(X'X \right)^{-1} \right|}$$

measures the change in the precision of the parameter estimates when an observation is deleted from the model.

Identifying Influential Observations – Summary of Suggested Cutoffs

Influential Statistics	Cutoff Values
RSTUDENT Residuals	|RSTUDENT| > 2
LEVERAGE	LEVERAGE > $\frac{2p}{n}$
Cook's D	CooksD > $\frac{4}{n}$
DFFITS	|DFFITS| > $2\sqrt{\frac{p}{n}}$
DFBETAS	|DFBETAS| > $\frac{2}{\sqrt{n}}$
COVRATIO	COVRATIO < $1-\frac{3p}{n}$ or COVRATIO > $1+\frac{3p}{n}$

Lineární regrese – PROC REG

PROC REG <options> ;
<label:>MODEL dependents=<regressors> </ options> ;
BY variables ;
FREQ variable ;
ID variables ;
VAR variables ;
WEIGHT variable ;
ADD variables ;
DELETE variables ;
<label:>MTEST <equation, ...,equation> </ options> ;
OUTPUT <OUT=SAS-data-set>< keyword=names> <...keyword=names> ;
PAINT <condition | ALLOBS> </ options > | < STATUS | UNDO> ;
RESTRICT equation, ...,equation ;
REWEIGHT <condition | ALLOBS> </ options > | < STATUS | UNDO> ;
PLOT <yvariable*xvariable> <=symbol> <...yvariable*xvariable> <=symbol> </ options> ;
PRINT <options> <ANOVA> <MODELDATA> ;
REFIT ;
RESTRICT equation, ...,equation ;
REWEIGHT <condition | ALLOBS> </ options > | < STATUS | UNDO> ;
<label:>TEST equation,<,...,equation> </ option> ;

Více na: http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_reg_sect001.htm

The SGPLOT Procedure

- General form of the SGPLOT procedure:

```
PROC SGPLOT <option(s)>;
      DOT category-variable </option(s)>;
      HBAR category-variable < /option(s) >;
       HBOX response-variable </option(s)>;
       HISTOGRAM response-variable < /option(s)>;
      NEEDLE X= variable Y= numeric-variable </option(s)>;
      REG X= numeric-variable Y= numeric-variable
                  </option(s)>;
       SCATTER X= variable Y= variable </option(s)>;
      VBAR category-variable < /option(s)>;
       VBOX response-variable </option(s)>;
RUN;
```

1.03 Quiz

- Suppose the regression model that you fit is

$$\hat{y} = 3 + 5x$$

- How do you interpret the slope for x, which is 5?

1.03 Quiz – Correct Answer

- Suppose the regression model that you fit is

$$\hat{y} = 3 + 5x$$

- How do you interpret the slope for x, which is 5?

- **For every 1-unit increase in x, the predicted value for y increases by 5.**

Regression Using Excel

- Tools / Data Analysis / Regression

Microsoft Excel - 13data.xls

File Edit View Insert Format Tools Data Window Help Acrobat

Chart 1

	A	B
1	House Price	Square Feet
2	245	1400
3	312	1600
4	279	1700
5	308	1875
6	199	1100
7	219	1550
8	405	2350
9	324	2450
10	319	1425
11	255	1700
12		
13		
14		
15		

Regression

Input
- Input Y Range: $A$1:$A$11
- Input X Range: $B$1:$B$11
- Labels
- Constant is Zero
- Confidence Level: 95 %

Output options
- Output Range:
- New Worksheet Ply:
- New Workbook

Residuals
- Residuals
- Residual Plots
- Standardized Residuals
- Line Fit Plots

Normal Probability
- Normal Probability Plots

OK
Cancel
Help

Excel Output

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

The regression equation is:

$$\widehat{\text{houseprice}} = 98.24833 + 0.10977(\text{squarefeet})$$

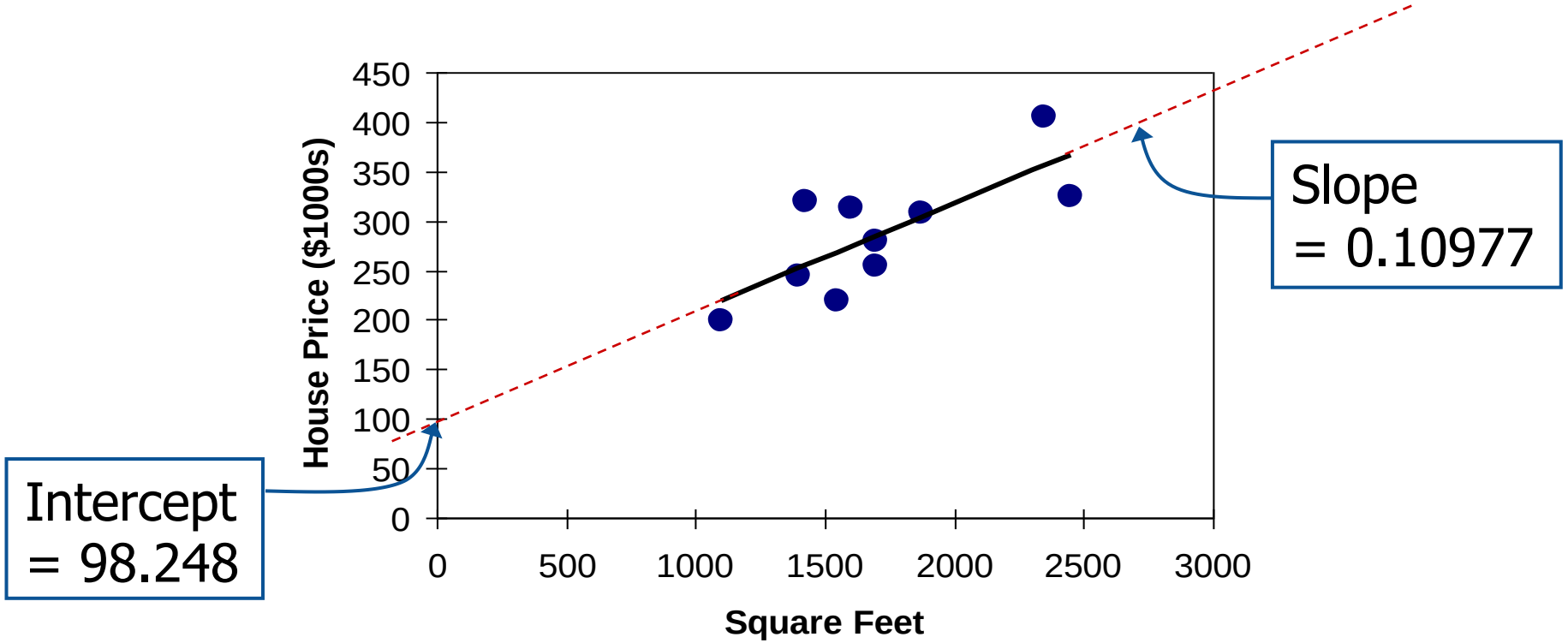
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Graphical Presentation

- House price model: scatter plot and regression line



$$\widehat{\text{houseprice}} = 98.24833 + 0.10977(\text{squarefeet})$$

Interpretation of the Intercept, b_0

$$\widehat{\text{houseprice}} = 98.24833 + 0.10977(\text{squarefeet})$$

- b_0 is the estimated average value of Y when the value of X is zero (if x = 0 is in the range of observed x values)
  - Here, no houses had 0 square feet, so b_0 = 98.24833 just indicates that, for houses within the range of sizes observed, $98,248.33 is the portion of the house price not explained by square feet

Interpretation of the Slope Coefficient, b_1

$$\widehat{\text{houseprice}} = 98.24833 + 0.10977(\text{squarefeet})$$

- b_1 measures the estimated change in the average value of Y as a result of a one-unit change in X
  - Here, $b_1 = .10977$ tells us that the average value of a house increases by .10977($1000) = $109.77, on average, for each additional one square foot of size

Explained and Unexplained Variation

- Total variation is made up of two parts:

$$SST = SSE + SSR$$

Total sum of Squares	Sum of Squares Error	Sum of Squares Regression
$SST = \sum(y - \bar{y})^2$	$SSE = \sum(y - \hat{y})^2$	$SSR = \sum(\hat{y} - \bar{y})^2$

where:

$\bar{y}$ = Average value of the dependent variable

y = Observed values of the dependent variable

$\hat{y}$ = Estimated value of y for the given x value

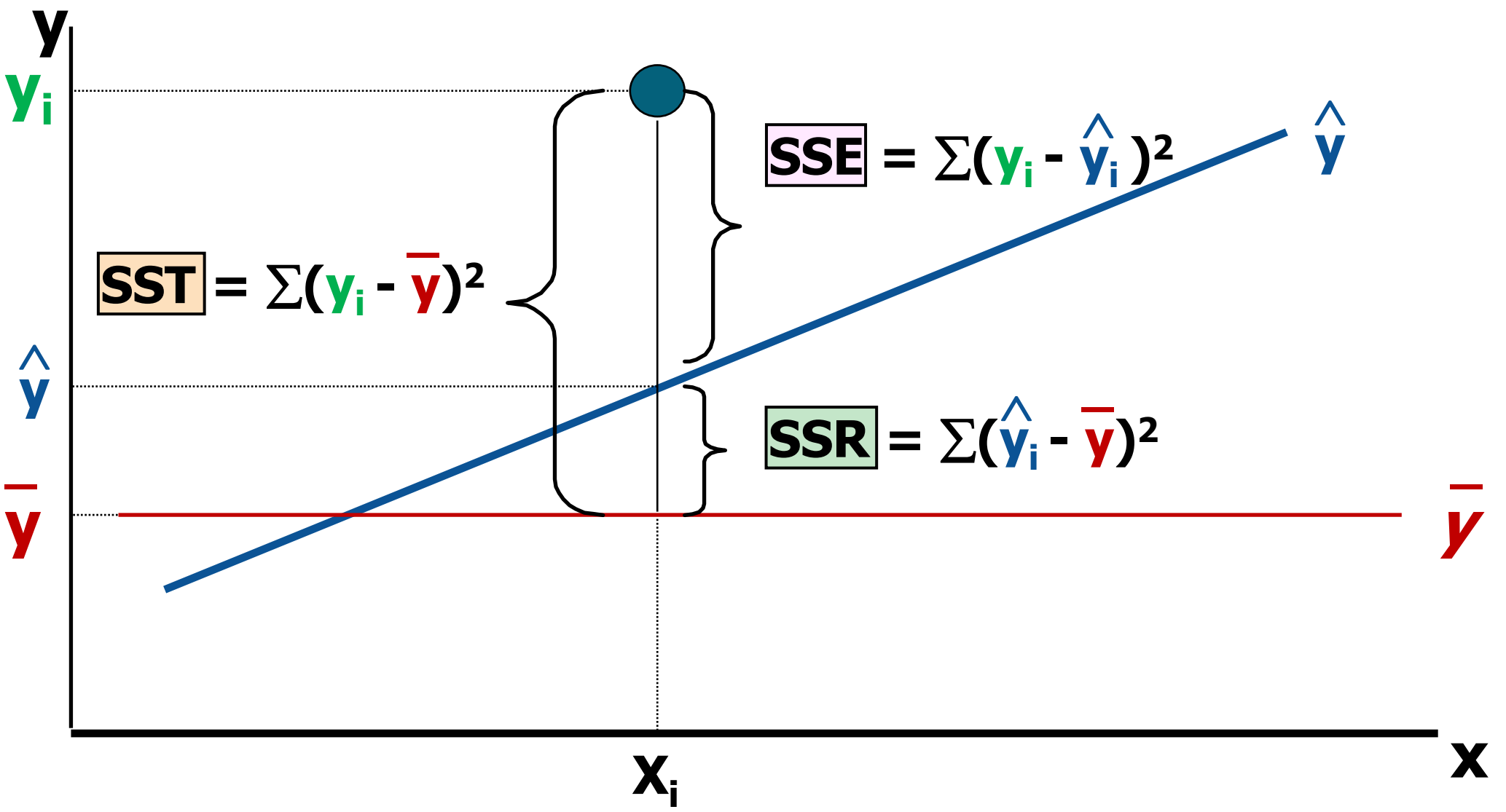
Explained and Unexplained Variation

(continued)

- SST = total sum of squares
  - Measures the variation of the y_i values around their mean y
- SSE = error sum of squares
  - Variation attributable to factors other than the relationship between x and y
- SSR = regression sum of squares
  - Explained variation attributable to the relationship between x and y

Explained and Unexplained Variation

(continued)



Coefficient of Determination, R^2

- The coefficient of determination is the portion of the total variation in the dependent variable that is explained by variation in the independent variable

- The coefficient of determination is also called R-squared and is denoted as R^2

$$R^2 = \frac{SSR}{SST}$$

where $0 \le R^2 \le 1$

Coefficient of Determination, R^2

(continued)

Coefficient of determination

$$R^2 = \frac{SSR}{SST} = \frac{\text{sum of squares explained by regression}}{\text{total sum of squares}}$$

Note: In the single independent variable case, the coefficient of determination is

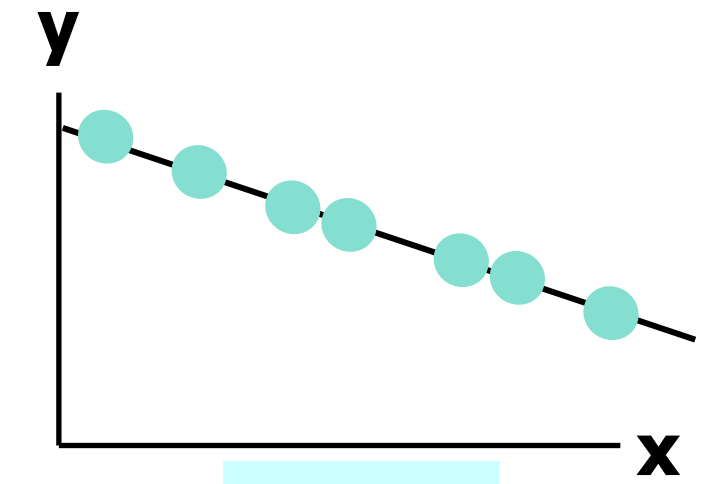
$$R^2 = r^2$$

where:

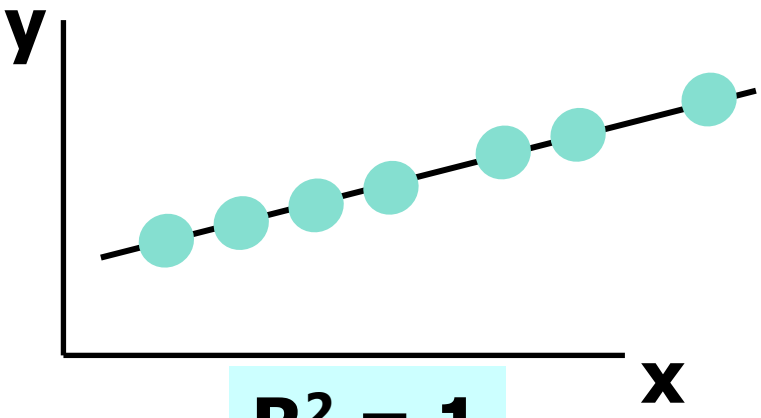
R^2 = Coefficient of determination

r = Simple correlation coefficient

Examples of Approximate R^2 Values



$R^2 = 1$



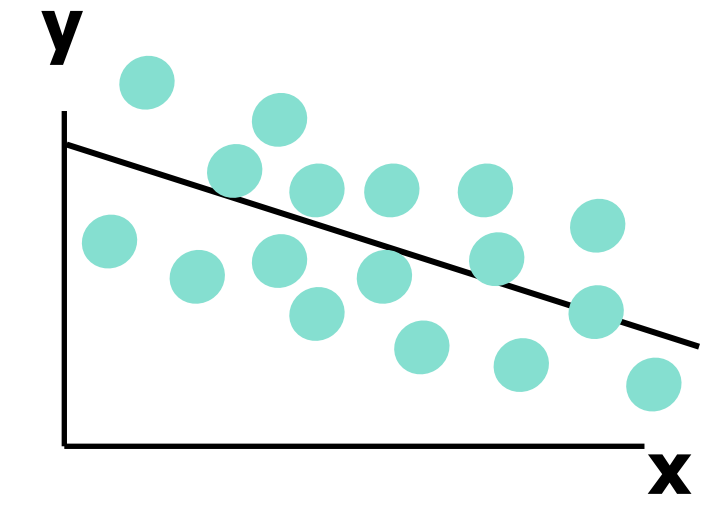
$R^2 = 1$

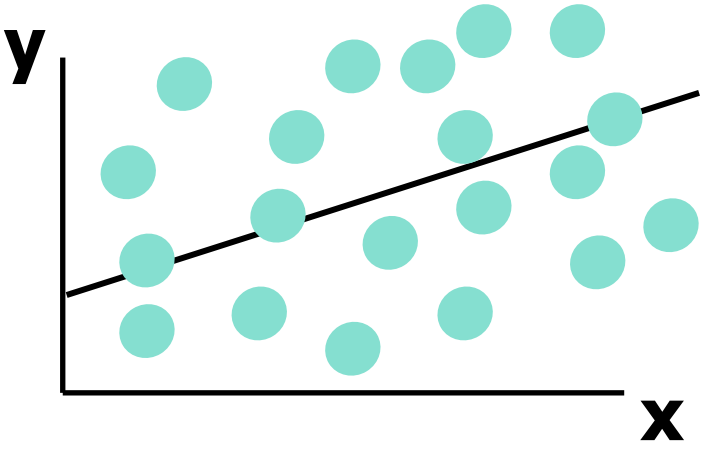
$R^2 = 1$

Perfect linear relationship between x and y:

100% of the variation in y is explained by variation in x

Examples of Approximate R^2 Values



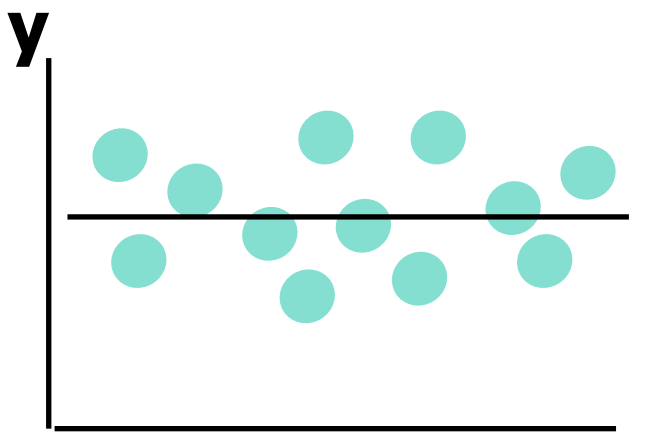


$0 < R^2 < 1$

Weaker linear relationship between x and y:

Some but not all of the variation in y is explained by variation in x

Examples of Approximate R^2 Values



$R^2 = 0$

$R^2 = 0$

No linear relationship between x and y:

The value of Y does not depend on x. (None of the variation in y is explained by variation in x)

Excel Output

Regression Statistics

Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

$$R^2 = \frac{SSR}{SST} = \frac{18934.9348}{32600.5000} = 0.58082$$

58.08% of the variation in house prices is explained by variation in square feet

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Standard Error of Estimate

- The standard deviation of the variation of observations around the regression line is estimated by

$$s_{\varepsilon} = \sqrt{\frac{SSE}{n-k-1}}$$

Where

SSE = Sum of squares error

n = Sample size

k = number of independent variables in the model

The Standard Deviation of the Regression Slope

- The standard error of the regression slope coefficient (b_1) is estimated by

$$s_{b_1} = \frac{s_\varepsilon}{\sqrt{\sum (x - \bar{x})^2}} = \frac{s_\varepsilon}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}}$$

where:

s_{b_1} = Estimate of the standard error of the least squares slope

$s_\varepsilon = \sqrt{\frac{SSE}{n-2}}$ = Sample standard error of the estimate

Excel Output

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

$s_\varepsilon = 41.33032$

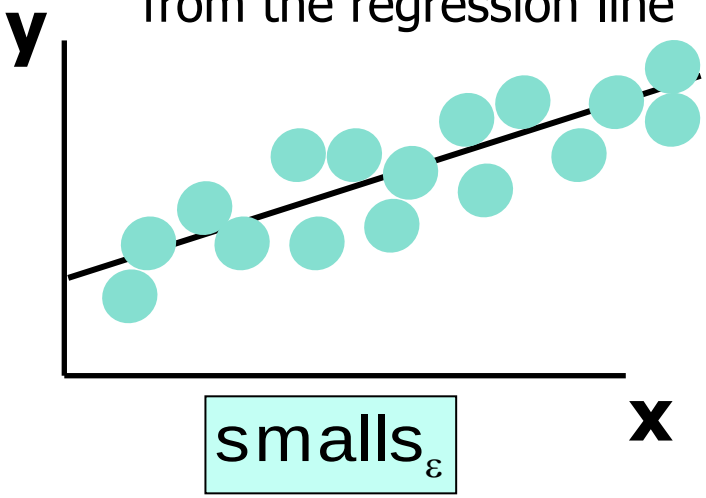
$s_{b_1} = 0.03297$

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

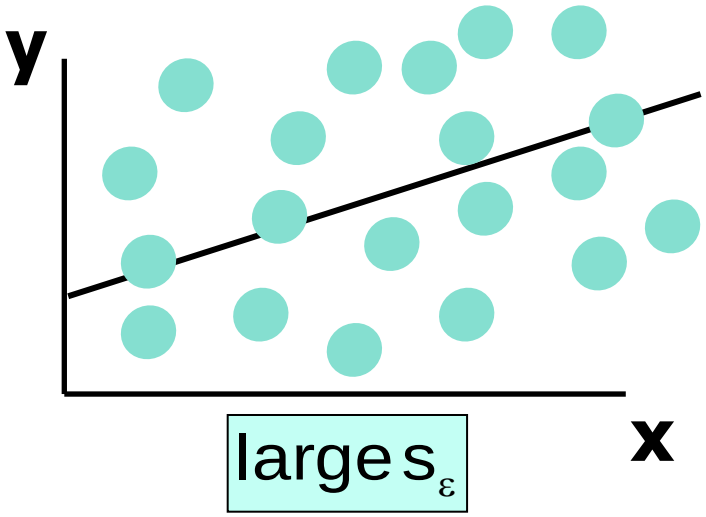
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Comparing Standard Errors

Variation of observed y values from the regression line

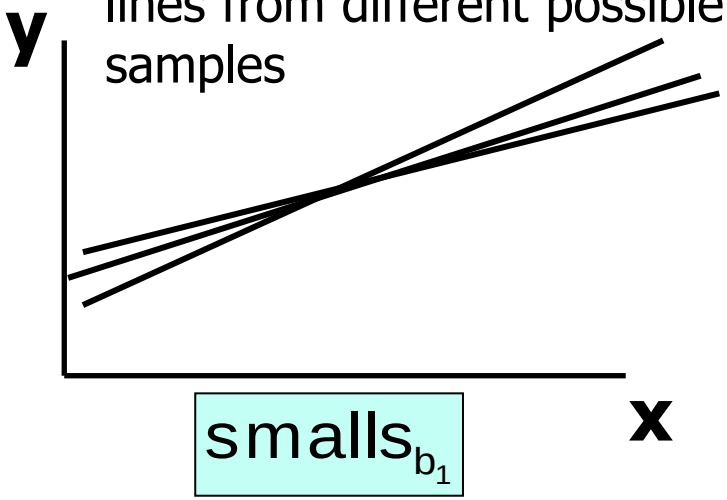


$small\ s_{\varepsilon}$

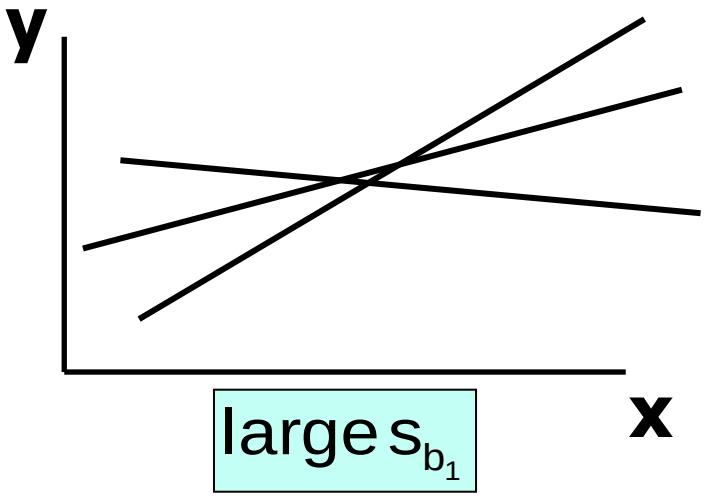


$large\ s_{\varepsilon}$

Variation in the slope of regression lines from different possible samples



$small\ s_{b_1}$

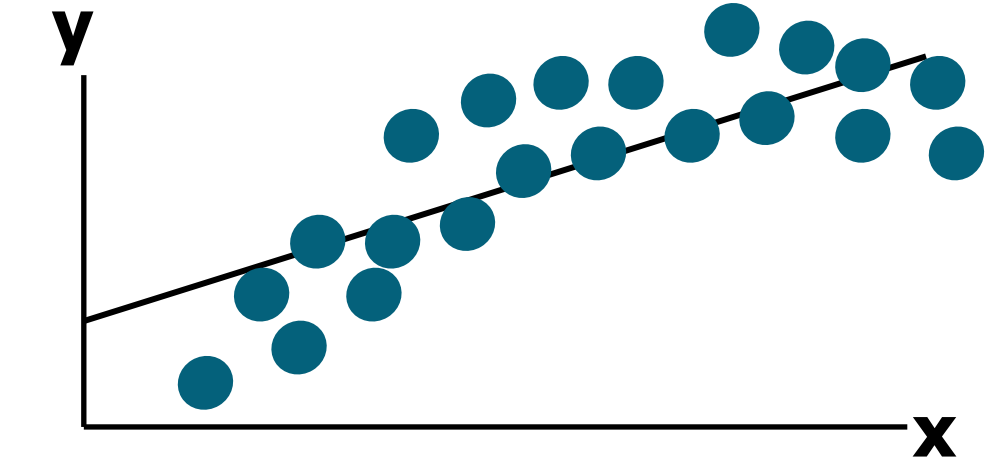


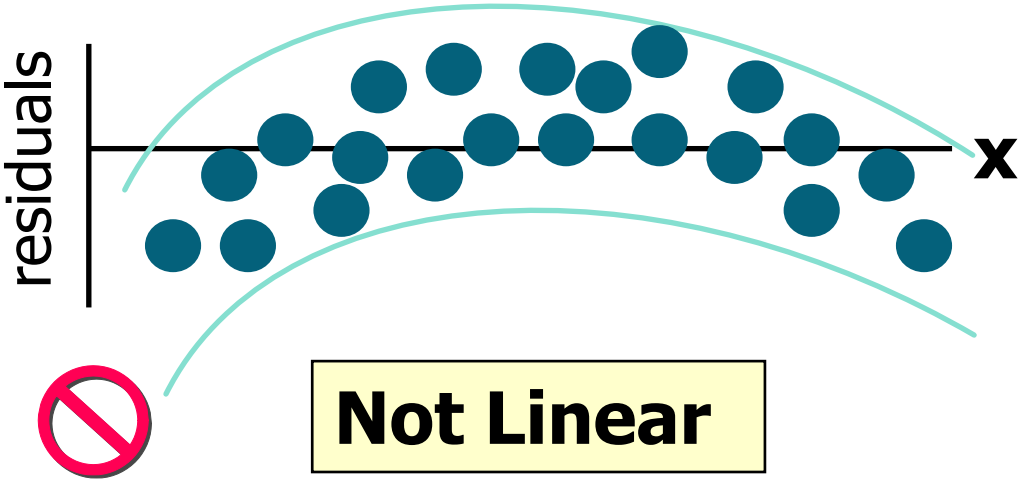
$large\ s_{b_1}$

Residual Analysis

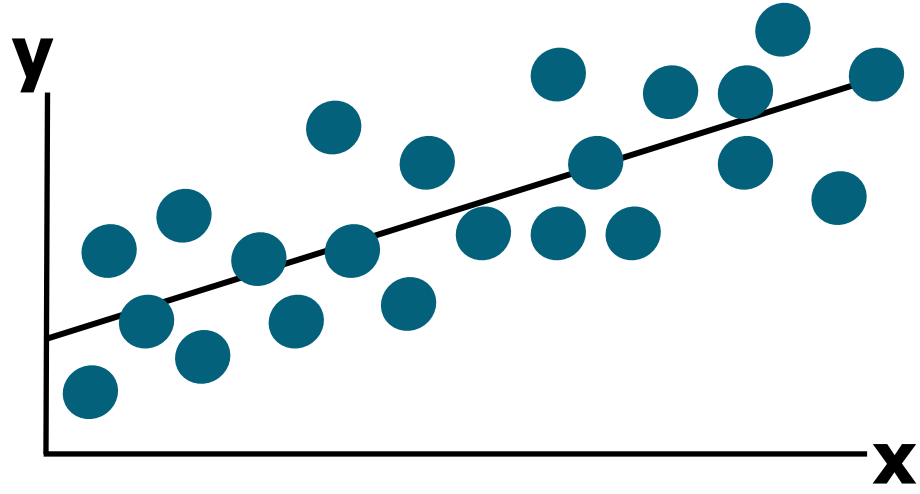
- Purposes
  - Examine for linearity assumption
  - Examine for constant variance for all levels of x
  - Evaluate normal distribution assumption
- Graphical Analysis of Residuals
  - Can plot residuals vs. x
  - Can create histogram of residuals to check for normality

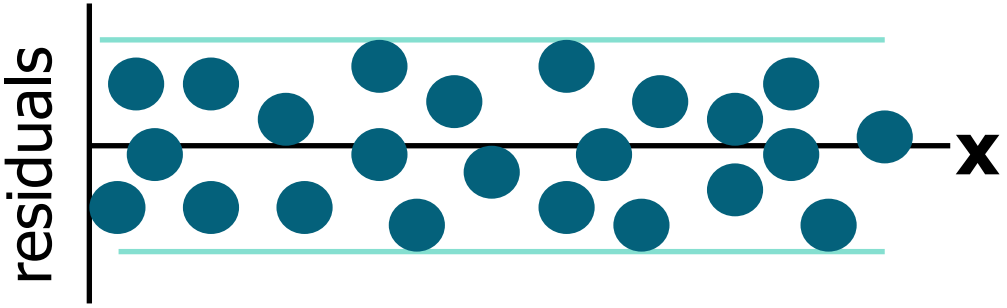
Residual Analysis for Linearity





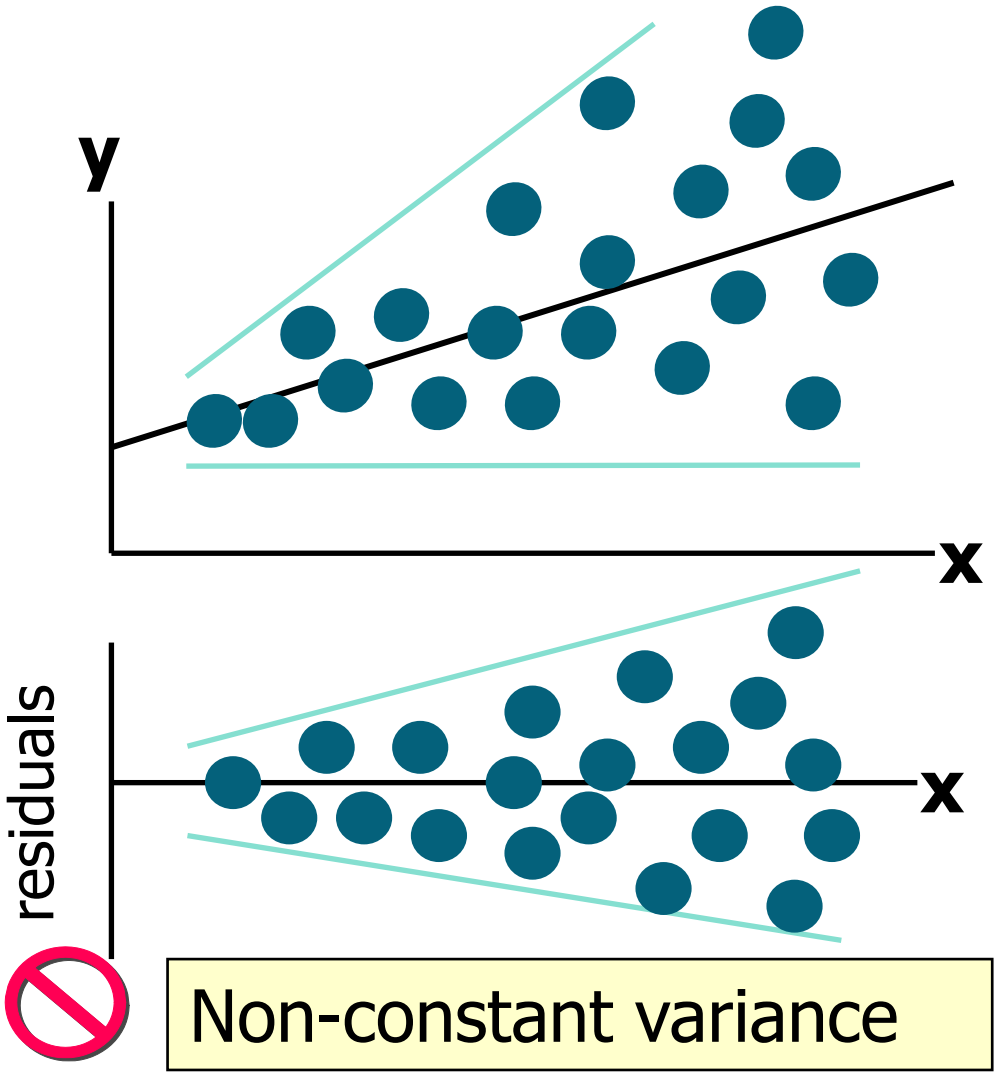
Not Linear

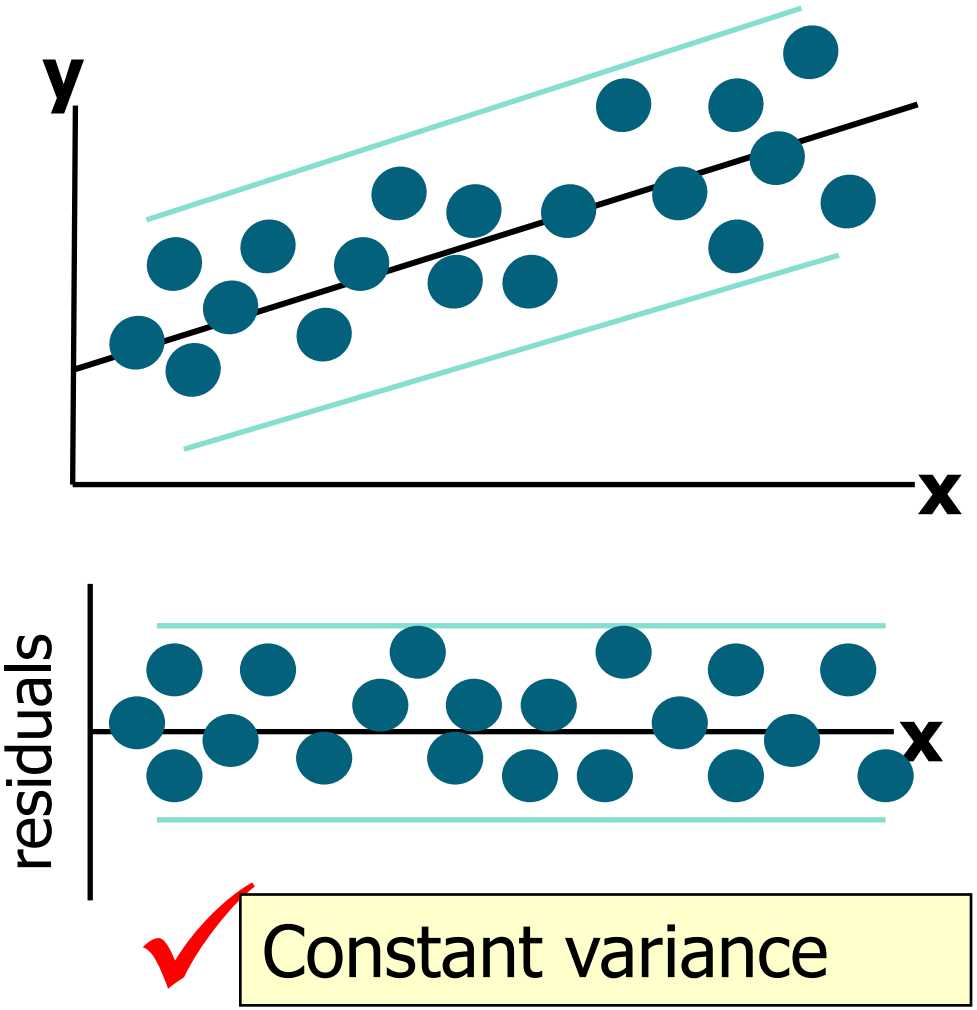




Linear

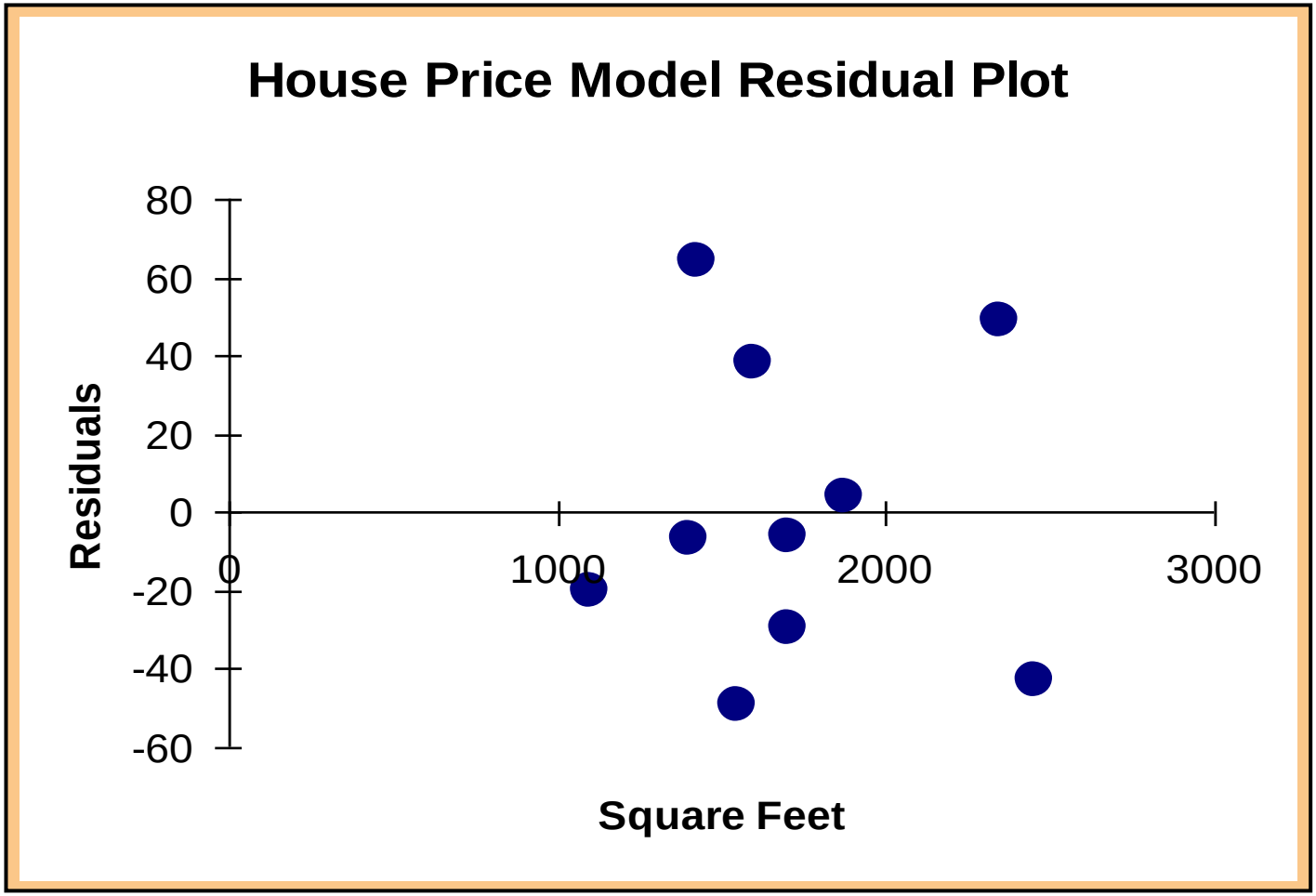
Residual Analysis for Constant Variance



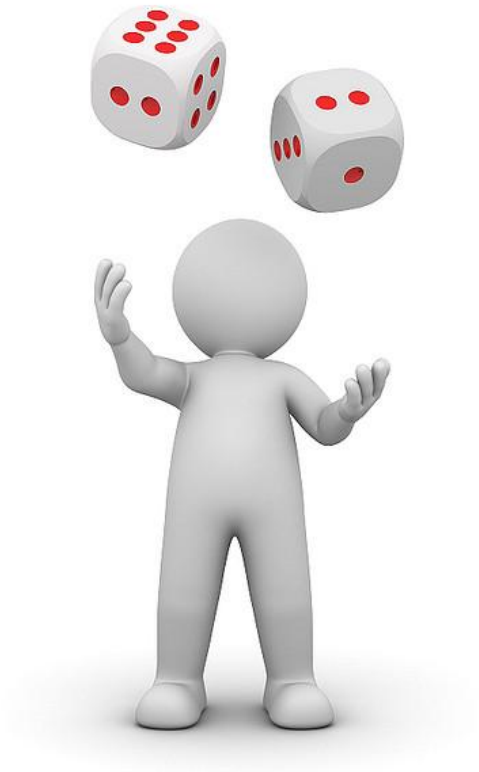


Excel Output

RESIDUAL OUTPUT		
	Predicted House Price	*Residuals*
1	251.92316	-6.923162
2	273.87671	38.12329
3	284.85348	-5.853484
4	304.06284	3.937162
5	218.99284	-19.99284
6	268.38832	-49.38832
7	356.20251	48.79749
8	367.17929	-43.17929
9	254.6674	64.33264
10	284.85348	-29.85348



6. Úvod do teorie pravděpodobnosti



Počet pravděpodobnosti -úvod

Počet pravděpodobnosti se zabývá studiem zákonitostí v náhodných pokusech. Matematickými prostředky modeluje situace, v nichž hraje roli náhoda. Pod pojmem náhoda rozumíme působení faktorů, které se živelně mění při různých provedeních téhož pokusu a nepodléhají naší kontrole.
Počet pravděpodobnosti jako vědecká disciplína se začal vytvářet **v 17. století** a jeho počátky jsou spjaty se jmény Blaise Pascala, Pierra de Fermata, Christiana Huygense (studovali hazardní hry, zformulovali takové pojmy, jako je pravděpodobnost a střední hodnota, odvodili jejich vlastnosti) a především Jakoba Bernoulliho (dokázal zákon velkých čísel).
V 18. století: Abraham de Moivre a Pierre Simeon Laplace – formulace jedné z forem centrální limitní věty, Georges Buffon odvodil binomickou větu, zavedl diferenciální a integrální počet do teorie pravděpodobnosti, Thomas Bayes odvodil způsob výpočtu aposteriorních pravděpodobností pomocí apriorních pravděpodobností (Bayesův vzorec).
V 19. století: Petrohradská matematická škola – dala teorii pravděpodobnosti pevný logický a matematický základ (Viktor Jakovlevič Buňakovskij, Pafnutij Lvovič Čebyšev, Andrej Andrejevič Markov, Alexandr Michailovič Ljapunov), Karl Fridirich Gauss (mj. vyvinul netodu zpracování experimentálních údajů známou pod názvem metoda nejmenších čtverců), Siméon Denis Poisson (zobecnil Bernoulliho zákon velkých čísel a odvodil speciální zákon rozložení pravděpodobností - Poissonův zákon rozložení).
Ve 20. století: Andrej Nikolajevič Kolmogorov (axiomatická teorie pravděpodobnosti), Norbert Wiener , William Feller (rozvoj teorie stochastických procesů).
Odkaz na zajímavou webovou stránku:
http://www-groups.dcs.st-and.ac.uk/~history

http://turnbull.mcs.st-and.ac.uk/~history/HistTopics/Statistics.html

Základní prostor

Definice (definice pokusu): Pokusem rozumíme jednorázové uskutečnění konstantně vymezeného souboru definičních podmínek. Předpokládáme, že pokus můžeme mnohonásobně nezávisle opakovat za dodržení definičních podmínek (ostatní podmínky se mohou měnit, proto různá opakování pokusu mohou vést k různým výsledkům). Dále předpokládáme, že opakováním pokusu vzniká opět pokus.
Deterministickým pokusem nazýváme takový pokus, jehož každé opakování vede k jedinému možnému výsledku.
Náhodným pokusem nazýváme takový pokus, jehož každé opakování vede k právě jednomu z více možných výsledků, které jsou vzájemně neslučitelné.
Příklad deterministického pokusu: při tlaku 1015 hPa zahříváme vodu na 100 °C. Jediným možným výsledkem je var vody.
Příklady náhodných pokusů: hod hrací kostkou, hod mincí, vylosování čísla z osudí apod.

Definice (definice základního prostoru): Neprázdnou množinu možných výsledků náhodného pokusu značíme Ω a nazýváme ji základní prostor. Možné výsledky značíme ω_t, kde $t \in T$, T je indexová množina.

Příklad

Příklad

a) Náhodný pokus spočívá v hodu kostkou. Možný výsledek ω_i znamená polohu kostky číslem i nahoru, i = 1, …, 6. Základní prostor $\Omega = \{\omega_1, \ldots, \omega_6\}$, počet možných výsledků $m(\Omega) = 6$.

b) Náhodný pokus spočívá v hodu dvěma kostkami. Možný výsledek je uspořádaná dvojice $[\omega_i, \omega_j]$, i, j = 1, …, 6. Základní prostor $\Omega = \{[\omega_1, \omega_1], [\omega_1, \omega_2], \ldots, [\omega_1, \omega_6], \ldots, [\omega_6, \omega_6]\}$, počet možných výsledků $m(\Omega) = 6^2 = 36$.

c) Náhodný pokus spočívá v opakovaném házení mincí tak dlouho, dokud nepadne první líc. Potom základní prostor $\Omega = \{\omega_1, \omega_2, \omega_3 \ldots\}$, kde ω_1 znamená, že hned v prvním hodu padl líc, ω_2 znamená, že až ve druhém hodu padl líc, ω_3 znamená, že až ve třetím hodu padl líc atd. Sybmolicky lze zapsat $\omega_1 = [\mathrm{L}]$, $\omega_2 = [\mathrm{R}, \mathrm{L}]$, $\omega_3 = [\mathrm{R}, \mathrm{R}, \mathrm{L}]$, … Tedy základní prostor Ω má nekonečně spočetně mnoho možných výsledků.

Jevové pole

Definice (definice jevového pole): Systém podmnožin $\mathcal{A}$ základního prostoru Ω, který splňuje následující tři axiómy:

J5: $A_1, A_2 \in \mathcal{A} \Rightarrow A_1 - A_2 \in \mathcal{A}$,

J6: $\Omega \in \mathcal{A}$,

J8: $A_1, A_2, \ldots \in \mathcal{A} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$

se nazývá jevové pole. Jestliže $A \in \mathcal{A}$, pak řekneme, že A je jev. Dvojice $(\Omega, \mathcal{A})$ se nazývá měřitelný prostor.

(Axióm J5 nám říká, že jevové pole obsahuje s každými dvěma množinami i jejich množinový rozdíl. Axióm J6 říká, že jevové pole obsahuje celý základní prostor a konečně axióm J8 říká, že když jevové pole obsahuje každou ze spočetné posloupnosti množin, obsahuje i jejich spočetné sjednocení. Znamená to, že systém $\mathcal{A}$ je uzavřený vzhledem k množinovým operacím.

Protože jevy jsou množiny, pro operace s nimi platí stejné zákony jako pro operace s množinami - komutativní zákon, asociativní zákon, de Morganova pravidla.)

Množinové a pravděpodobnostní pojmy

Poznámka (slovník množinových a pravděpodobnostních pojmů)

Ω se nazývá jistý jev, $\varnothing$ se nazývá nemožný jev

$\omega \in A$ znamená, že možný výsledek ω je příznivý nastoupení jevu A

$A \subseteq B$ znamená, že jev A má za důsledek jev B

$A \cup B$ znamená nastoupení aspoň jednoho z jevů A, B

$A \cap B$ znamená společné nastoupení jevů A, B

$A - B$ znamená nastoupení jevu A za nenastoupení jevu B

$\overline{A} = \Omega - A$ znamená jev opačný k jevu A

$A \cap B = \varnothing$ znamená, že jevy A, B jsou neslučitelné.

Příklad

Příklad: Je dán systém složený ze dvou bloků, který jednorázově použijeme. Nechť jev A_i znamená bezporuchovou funkci i -tého bloku, i = 1, 2. Pomocí jevů A_1, A_2 vyjádřete jevy:

a) bezporuchová funkce aspoň jednoho bloku: $A_1 \cup A_2$

b) bezporuchová funkce obou bloků: $A_1 \cap A_2$

c) porucha aspoň jednoho bloku: $\overline{A_1} \cup \overline{A_2}$

d) porucha obou bloků: $\overline{A_1} \cap \overline{A_2}$

e) porucha právě jednoho bloku: $(\overline{A_1} \cap A_2) \cup (A_1 \cap \overline{A_2})$

Jevové pole - poznámky

Poznámka: Systém axiómů jevového pole je bezesporný (tj. na každém základním prostoru lze sestrojit aspoň jedno jevové pole) a neúplný (tzn., že na každém aspoň dvouprvkovém základním prostoru lze vytvořit jevových polí více).
Neúplnost systému axiómů jevového pole je výhodná, protože umožňuje rozlišovat výsledky náhodného pokusu s různým stupněm podrobnosti.
Např. jevové pole $\mathcal{A}_{\min} = \{\Omega, \varnothing\}$ se nazývá minimální jevové pole a charakterizuje krajně „tupozrakého“ pozorovatele, který rozliší pouze jev jistý a jev nemožný.
Jevové pole $\mathcal{A}_1 = \{\Omega, \varnothing, A, \overline{A}\}$ již dovolí rozeznat, zda nastal jev A nebo jev opačný $\overline{A}$.
Tak můžeme konstruovat stále bohatší jevová pole, až dostaneme maximální jevové pole $\mathcal{A}_{\max} = \{A; A \subseteq \Omega\}$.
To charakterizuje krajně „bystrozrakého“ pozorovatele, který rozliší jevy do všech podrobností. Pro libovolné jevové pole $\mathcal{A}$ ovšem platí: $\mathcal{A}_{\min} \subseteq \mathcal{A} \subseteq \mathcal{A}_{\max}$.

Příklad

Příklad: Sestrojte všechna možná jevová pole na základním prostoru $\Omega = \{\omega_1, \omega_2, \omega_3\}$.

Řešení:

$\mathcal{A}_1 = \{\Omega, \varnothing\} (= \mathcal{A}_{\min})$

$\mathcal{A}_2 = \{\Omega, \varnothing, \{\omega_1\}, \{\omega_2, \omega_3\}\}$

$\mathcal{A}_3 = \{\Omega, \varnothing, \{\omega_2\}, \{\omega_1, \omega_3\}\}$

$\mathcal{A}_4 = \{\Omega, \varnothing, \{\omega_3\}, \{\omega_1, \omega_2\}\}$

$\mathcal{A}_5 = \{\Omega, \varnothing, \{\omega_1\}, \{\omega_2\}, \{\omega_1, \omega_2\}, \{\omega_3\}, \{\omega_1, \omega_3\}, \{\omega_2, \omega_3\}\} (= \mathcal{A}_{\max})$

Jevové pole - vlastnosti

Věta (vlastnosti jevového pole): Nechť $(\Omega, \mathcal{A})$ je měřitelný prostor. Pak jevové pole $\mathcal{A}$ má následujících 9 vlastností:

J1: $\mathcal{A} \neq \emptyset$,

J2: $\emptyset \in \mathcal{A}$,

J3: $A_1, A_2 \in \mathcal{A} \Rightarrow A_1 \cup A_2 \in \mathcal{A}$,

J4: $A_1, A_2 \in \mathcal{A} \Rightarrow A_1 \cap A_2 \in \mathcal{A}$,

J5: $A_1, A_2 \in \mathcal{A} \Rightarrow A_1 - A_2 \in \mathcal{A}$ (axióm),

J6: $\Omega \in \mathcal{A}$ (axióm),

J7: $A \in \mathcal{A} \Rightarrow \overline{A} \in \mathcal{A}$,

J8: $A_1, A_2, \ldots \in \mathcal{A} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$ (axióm),

J9: $A_1, A_2, \ldots \in \mathcal{A} \Rightarrow \bigcap_{i=1}^{\infty} A_i \in \mathcal{A}$.

Důkaz: J1 plyne z J6.

J2 plyne z J5 a J6, protože $\Omega - \Omega = \emptyset$.

J3 plyne z J2 J8 speciální volbou $A_3 = \emptyset, A_4 = \emptyset, \ldots$ Pak $\bigcup_{i=1}^{\infty} A_i = A_1 \cup A_2$.

J7 plyne J5 a J6, protože $\overline{A} = \Omega - A$.

J9 odvodíme z J7 a J8 užitím de Morganových pravidel $\bigcap_{i=1}^{\infty} A_i = \overline{\bigcup_{i=1}^{\infty} \overline{A_i}}$:

$A_1, A_2, \ldots \in \mathcal{A} \Rightarrow \overline{A_1}, \overline{A_2}, \ldots \in \mathcal{A} \Rightarrow \bigcup_{i=1}^{\infty} \overline{A_i} \in \mathcal{A} \Rightarrow \overline{\bigcup_{i=1}^{\infty} \overline{A_i}} \in \mathcal{A}$, ovšem $\overline{\bigcup_{i=1}^{\infty} \overline{A_i}} = \bigcap_{i=1}^{\infty} A_i$.

J4 plyne z J9 speciální volbou $A_3 = \Omega, A_4 = \Omega, \ldots$

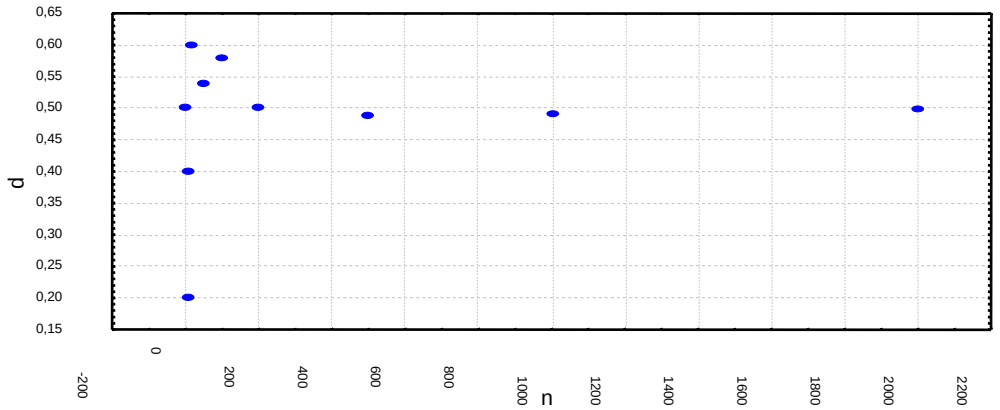
Pravděpodobnostní prostor

Motivace: Provádíme opakovaně nezávisle týž náhodný pokus a v každém pokusu sledujeme nastoupení jevu A, kterému říkáme úspěch. Označme n celkový počet pokusů a N(A) počet těch pokusů, kdy nastal úspěch. S rostoucím n pozorujeme, že relativní četnost úspěchu $\frac{N(A)}{n}$ se blíží číslu P(A), které považujeme za pravděpodobnost úspěchu. (Tento poznatek je znám jako **empirický zákon velkých čísel**).

Ilustrace empirického zákona velkých čísel

Provádíme n nezávislých hodů mincí. Padnutí líce považujeme za úspěch. Budeme sledovat závislost relativní četnosti úspěchu na počtu pokusů. (Počet pokusů volíme 2, 5, 10, 20, 50, 100, 200, 500,1000, 2000.)

n	2	5	10	20	50	100	200	500	1000	2000
p	0,5	0,2	0,4	0,6	0,54	0,58	0,5	0,488	0,49	0,4975



Axiomatická teorie pravděpodobnosti

Vzniká otázka, jak zavést pravděpodobnost, aby byla „zidealizovaným“ protějškem relativní četnosti. Zdálo by se vhodné zavést pravděpodobnost takto:

$$P(A) = \lim_{n\to\infty} \frac{N(A)}{n}.$$

Jde o tzv. **statistickou definici pravděpodobnosti** . Z matematického hlediska tato definice není v pořádku, protože počet pokusů je vždy konečný a nelze se přesvědčit o existenci uvedené limity. Proto ve 30. letech 20. století ruský matematik A. A. Kolmogorov (1903 – 1987) vybudoval **axiomatickou teorii pravděpodobnosti.**

Axiomatická teorie pravděpodobnosti zavádí pravděpodobnost jako funkci, která každému jevu přiřazuje číslo mezi 0 a 1 a přitom je zidealizovaným protějškem relativní četnosti. Má tedy všechny vlastnosti relativní četnosti a kromě toho některé další vlastnosti, které vyplývají z vnitřních potřeb matematické teorie.

Pravděpodobnost - definice

Definice: Nechť $(\Omega, \mathcal{A})$ je měřitelný prostor. Reálná množinová funkce P: $\mathcal{A} \to R$ se nazývá pravděpodobnost, když splňuje následující 3 axiómy:

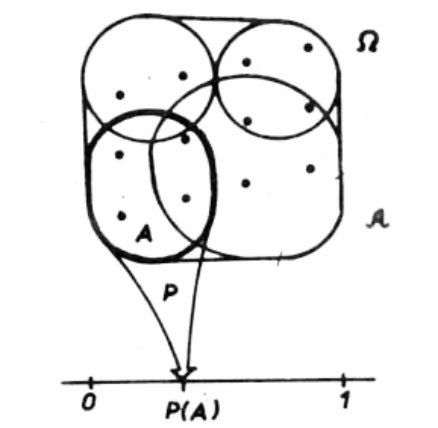
P2: $\forall A \in \mathcal{A}$: $P(A) \geq 0$ (nezápornost)

P10: $P(\Omega) = 1$ (normovanost)

P15: $A_1, A_2, \ldots \in \mathcal{A}$ jsou neslučitelné $\Rightarrow P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$ (spočetná aditivita)

Trojice $(\Omega, \mathcal{A}, P)$ se nazývá pravděpodobnostní prostor. (Je to matematický model jednorázového provedení náhodného pokusu.)

Ilustrace pravděpodobnostního prostoru



Poznámka: Systém axiómů pravděpodobnosti je bezesporný (tj. na každém měřitelném prostoru lze sestrojit pravděpodobnost) a neúplný (tj. na každém měřitelném prostoru, jehož jevové pole není minimální, lze sestrojit pravděpodobností více).

Pravděpodobnost - vlastnosti

Věta (vlastnosti pravděpodobnosti): Nechť $(\Omega, \mathcal{A}, P)$ je pravděpodobnostní prostor, $A, A_1, A_2, \ldots \in \mathcal{A}$ libovolné jevy. Pak pravděpodobnost P má následujících 17 vlastností:

P1: $P(\varnothing) = 0$

P2: $P(A) \geq 0$ (nezápornost – axióm)

P3: $P(A_1 \cup A_2) + P(A_1 \cap A_2) = P(A_1) + P(A_2)$

P4: $1 + P(A_1 \cap A_2) \geq P(A_1) + P(A_2)$

P5: $P(A_1 \cup A_2) \leq P(A_1) + P(A_2)$ (subaditivita)

P6: $A_1 \cap A_2 = \varnothing \Rightarrow P(A_1 \cup A_2) = P(A_1) + P(A_2)$ (aditivita)

P7: $P(A_2 - A_1) = P(A_2) - P(A_1 \cap A_2)$

P8: $A_1 \subseteq A_2 \Rightarrow P(A_2 - A_1) = P(A_2) - P(A_1)$ (subtraktivita)

P9: $A_1 \subseteq A_2 \Rightarrow P(A_1) \leq P(A_2)$ (monotonie)

P10: $P(\Omega) = 1$ (normovanost – axióm)

P11: $P(A) + P(\bar{A}) = 1$ (komplementarita)

P12: $P(A) \leq 1$

P13: $P\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} P(A_i)$ (spočetná subaditivita)

P14: $A_1, A_2, \ldots \in \mathcal{A}$ jsou neslučitelné $\Rightarrow \sum_{i=1}^{\infty} P(A_i) < \infty$ (absolutní konvergence)

P15: $A_1, A_2, \ldots \in \mathcal{A}$ jsou neslučitelné $\Rightarrow P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$ (spočetná aditivita – axióm)

P16: $A_1 \subseteq A_2 \subseteq \ldots \in \mathcal{A} \Rightarrow P\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim_{i \to \infty} P(A_i)$ (spojitost pravděpodobnosti zdola)

P17: $A_1 \supseteq A_2 \supseteq \ldots \in \mathcal{A} \Rightarrow P\left(\bigcap_{i=1}^{\infty} A_i\right) = \lim_{i \to \infty} P(A_i)$ (spojitost pravděpodobnosti shora)

Pravděpodobnost - vlastnosti

P14 Položme $A_0 = \overline{\bigcup_{i=1}^{\infty} A_i}$. Pak jevy $A_0, A_1, A_2, \ldots$ jsou neslučitelné a jejich sjednocením je celý základní prostor, tedy podle axiómu P10 dostáváme: $1 = P(\Omega) = P(\bigcup_{i=0}^{\infty} A_i) = \sum_{i=0}^{\infty} P(A_i)$, přičemž poslední rovnost vyplývá z axiómu P15. $\sum_{i=0}^{\infty} P(A_i)$ tedy absolutně konverguje, tudíž bude konvergovat také $\sum_{i=1}^{\infty} P(A_i)$, kde jsme vynechali první člen.

P1 Položme $A_1 = \emptyset, A_2 = \emptyset, \ldots$ Pak $\bigcup_{i=1}^{\infty} A_i = \emptyset$, tedy podle axiómu P15 $0 = P(\emptyset) = P(\bigcup_{i=1}^{\infty} \emptyset) = \sum_{i=1}^{\infty} P(\emptyset)$, což je možné jen tak, že $P(\emptyset) = 0$.

P6 V axiómu P15 položíme $A_3 = \emptyset, A_4 = \emptyset, \ldots$, tedy $P(\bigcup_{i=1}^{\infty} A_i) = P(A_1 \cup A_2) = \sum_{i=1}^{\infty} P(A_i) = P(A_1) + P(A_2)$.

P11 Plyne z vlastnosti P6 a axiómu P10: $P(A \cup \overline{A}) = P(\Omega) = 1 = P(A) + P(\overline{A})$.

P12 Plyne okamžitě z axiómu P2 a vlastnosti P11.

Pravděpodobnost - vlastnosti

Pro důkaz vlastností P3, P4 a P5 jevy $A_1 \cup A_2$, A_1 a A_2 rozložíme na součet disjunktních sčítanců:
$A_1 \cup A_2 = (A_1 \setminus A_2) \cup (A_1 \cap A_2) \cup (A_2 \setminus A_1)$
$A_1 = (A_1 \setminus A_2) \cup (A_1 \cap A_2)$
$A_2 = (A_2 \setminus A_1) \cup (A_1 \cap A_2)$

P3 Podle P6 dostáváme: $P(A_1 \cup A_2) + P(A_1 \cap A_2) = P(A_1 \setminus A_2) + P(A_1 \cap A_2) + P(A_2 \setminus A_1) + P(A_1 \cap A_2) = P(A_1) + P(A_2)$.

Protože podle P12 je $P(A_1 \cup A_2) \leq 1$ a podle P2 je $P(A_1 \cap A_2) \geq 0$, dostáváme z P3 okamžitě P4 a P5.

P7 Opět vyjádříme A_2 jako sjednocení neslučitelných jevů: $A_2 = (A_2 \setminus A_1) \cup (A_1 \cap A_2)$. Podle P3 pak dostaneme: $P(A_2) = P(A_2 \setminus A_1) + P(A_1 \cap A_2)$, tedy $P(A_2 \setminus A_1) = P(A_2) - P(A_1 \cap A_2)$.

P8 Jelikož $A1 \subseteq A_2$, platí $A_1 \cap A_2 = A_1$ a P8 plyne z P7.

P9 Plyne z P8, protože podle P2 je $P(A_2 \setminus A_1) \geq 0$, tudíž $P(A_2) - P(A_1) \geq 0$, tj. $P(A_1) \leq P(A_2)$.

Pravděpodobnost - vlastnosti

P13 Položíme $\bigcup_{i=1}^{\infty} A_i = A_1 \cup (A_2 \setminus A_1) \cup (A_3 \setminus (A_1 \cup A_2)) \cup \ldots$ Tím jsme dostali sjednocení posloupnosti neslučitelných jevů a aplikujeme axióm P15 a vlastnost P7: $P(\bigcup_{i=1}^{\infty} A_i) = P(A_1) + P(A_2 \setminus A_1) + P(A_3 \setminus (A_1 \cup A_2)) + \ldots \leq P(A_1) + P(A_2) + P(A_3) + \ldots = \sum_{i=1}^{\infty} P(A_i)$.

P16 Jev $\bigcup_{i=1}^{\infty} A_i$ vyjádříme jako sjednocení neslučitelných jevů. Z předpokladu $A_1 \subseteq A_2 \subseteq \ldots$ plyne $\bigcup_{i=1}^{\infty} A_i = A_1 \cup (A_2 \setminus A_1) \cup (A_3 \setminus A_2) \cup \ldots \cup (A_i \setminus A_{i-1}) \cup \ldots$, tedy podle axiómu P15 a vlastnosti P8 dostáváme: $P(\bigcup_{i=1}^{\infty} A_i) = P(A_1) + P(A_2 \setminus A_1) + P(A_3 \setminus A_2) + \ldots + P(A_i \setminus A_{i-1}) + \ldots = P(A_1) + [P(A_2) - P(A_1)] + [P(A_3) + P(A_2)] + \ldots + [P(A_i) + P(A_{i-1})] + \ldots = \lim_{i \to \infty} P(A_i)$.

P17 Podle vlastnosti P16 dostáváme $P(\overline{\bigcup_{i=1}^{\infty} A_i}) = \lim_{i \to \infty} P(\overline{A_i})$. Z de Morganových pravidel plyne $P(\bigcap_{i=1}^{\infty} A_i) = P(\overline{\bigcup_{i=1}^{\infty} \overline{A_i}}) = 1 - P(\bigcup_{i=1}^{\infty} \overline{A_i}) = 1 - \lim_{i \to \infty} P(\overline{A_i}) = 1 - \lim_{i \to \infty} [1 - P(A_i)] = \lim_{i \to \infty} P(A_i)$.

Pravděpodobnost - vlastnosti

Věta ♠ (další vlastnosti pravděpodobnosti): Nechť $(\Omega, \mathcal{A}, P)$ je pravděpodobnostní prostor, $A_1, A_2, \ldots, A_n \in \mathcal{A}$ libovolné jevy. Pak platí:

a) $$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) - \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} P(A_i \cap A_j) + \sum_{i=1}^{n-2}\sum_{j=i+1}^{n-1}\sum_{k=j+1}^{n} P(A_i \cap A_j \cap A_k) - \ldots + (-1)^{n-1} P(A_1 \cap \ldots \cap A_n)$$

(Pro neslučitelné jevy $A_1, \ldots, A_n$ dostáváme $P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i)$.) (Věta o sčítání pravděpodobností)

b) $$\max_{1\le i\le n} P(A_i) \le P\left(\bigcup_{i=1}^{n} A_i\right) \le \sum_{i=1}^{n} P(A_i)$$

c) $$1 - n + \sum_{i=1}^{n} P(A_i) \le P\left(\bigcap_{i=1}^{n} A_i\right) \le \min_{1\le i\le n} P(A_i)$$ (nerovnost vlevo se nazývá Bonferroniho nerovnost)

Pravděpodobnost - vlastnosti

Důkaz:

ad a) Vlastnost vyjadřuje princip inkluze a exkluze. Tvrzení o neslučitelných jevech plyne z axiómu P15, kde položíme $A_{n+1} = \emptyset, A_{n+2} = \emptyset, \ldots$

ad b) Levá strana: Plyne z monotonie P9. Pro $\forall i \in \{1, \ldots, n\}$ je $A_i \subseteq \bigcup_{j=1}^{n} A_j$, tedy pro $\forall i \in \{1, \ldots, n\}$ platí $P(A_i) \leq P\left(\bigcup_{j=1}^{n} A_j\right)$.

Tvrzení musí platit i pro ten index *i*, pro který je $P(A_i)$ maximální.

Pravá strana: Plyne ze spočetné subaditivity P13, kde položíme $A_{n+1} = \emptyset, A_{n+2} = \emptyset, \ldots$

ad c) Levá strana: $P\left(\bigcap_{i=1}^{n} A_i\right) = P\left(\overline{\bigcup_{i=1}^{n} \overline{A_i}}\right) = 1 - P\left(\bigcup_{i=1}^{n} \overline{A_i}\right) \geq 1 - \sum_{i=1}^{n} P\left(\overline{A_i}\right) = 1 - \sum_{i=1}^{n} [1 - P(A_i)] = 1 - n + \sum_{i=1}^{n} P(A_i)$

Pravá strana: Plyne z monotonie P 9. Pro $\forall i \in \{1, \ldots, n\}$ je $A_i \supseteq \bigcap_{j=1}^{n} A_j$, tedy pro $\forall i \in \{1, \ldots, n\}$ platí

$P(A_i) \geq P\left(\bigcap_{j=1}^{n} A_j\right)$. Tvrzení musí platit i pro ten index i, pro který je $P(A_i)$ minimální.

Příklad

Příklad: Je dán systém složený ze dvou bloků. Jev A_i značí bezporuchovou funkci i-tého bloku, i = 1, 2. Je známo, že $P(A_i) = \vartheta_i$, i = 1, 2.

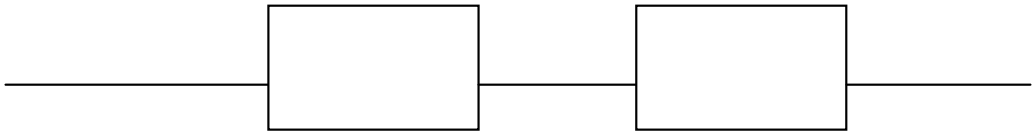
a) Odhadněte pravděpodobnost správné funkce celého systému, jsou-li bloky zapojeny
α) sériově, β) paralelně.

b) Předpokládejme navíc, že $P(A_1 \cap A_2) = \vartheta_{12}$. Vypočtěte nyní pravděpodobnost správné funkce celého systému, jsou-li bloky zapojeny
α) sériově, β) paralelně.

Řešení:

ad a)

Případ sériového zapojení



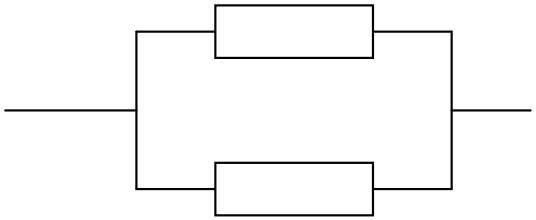
$P(A_1 \cap A_2)$ lze shora i zdola odhadnout pomocí věty ♠ (c), kde n = 2:

$$1 - 2 + P(A_1) + P(A_2) \le P(A_1 \cap A_2) \le \min\{P(A_1), P(A_2)\}$$

$$\vartheta_1 + \vartheta_2 - 1 \le P(A_1 \cap A_2) \le \min\{\vartheta_1, \vartheta_2\}$$

Příklad

Případ paralelního zapojení



$P(A_1 \cup A_2)$ lze shora i zdola odhadnout pomocí věty ♠ (b), kde n = 2:

$$\max\{P(A_1), P(A_2)\} \le P(A_1 \cup A_2) \le P(A_1) + P(A_2)$$

$$\max\{\vartheta_1, \vartheta_2\} \le P(A_1 \cup A_2) \le \vartheta_1 + \vartheta_2$$

ad b)

Případ sériového zapojení: $P(A_1 \cap A_2) = \vartheta_{12}$

Případ paralelního zapojení: Podle vlastnosti P3 dostáváme: $P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2) = \vartheta_1 + \vartheta_2 - \vartheta_{12}$

Stochasticky nezávislé jevy

Motivace: Při provádění pokusu se může stát, že z informace o nastoupení či nenastoupení jednoho jevu jsme schopni odvodit, zda jiný jev nastoupí či nenastoupí, tzn., že platí jedna z inkluzí $A \subseteq B, \overline{A} \subseteq B, A \subseteq \overline{B}, \overline{A} \subseteq \overline{B}$. V takovém případě hovoříme o deterministicky závislých jevech. Jejich protipólem jsou jevy stochasticky nezávislé – informace o nastoupení či nenastoupení jednoho jevu nijak neovlivní šance, s nimiž očekáváme nastoupení jiného jevu.
V popisné statistice jsme zavedli četnostní nezávislost dvou množin G_1, G_2 v daném výběrovém souboru pomocí multiplikativního vztahu: $p(G_1 \cap G_2) = p(G_1)p(G_2)$. V počtu pravděpodobnosti požadujeme pro stochasticky nezávislé jevy A_1, A_2 splnění multiplikativního vztahu: $P(A_1 \cap A_2) = P(A_1)P(A_2)$. Pro tři jevy budeme požadovat, aby i jevy $A_1 \cap A_2$ a A_3 byly stochasticky nezávislé, což vede ke vztahu $P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3)$. Tak můžeme pokračovat pro libovolný počet jevů.

Definice: Nechť $(\Omega, \mathcal{A}, P)$ je pravděpodobnostní prostor. Řekneme, že jevy $A, B \in \mathcal{A}$ jsou stochasticky nezávislé (vzhledem k P), jestliže $P(A \cap B) = P(A)\,P(B)$.

Stochasticky nezávislé jevy

Věta: Pro libovolné jevy $A, B \in \mathcal{A}$ platí:

a) $\varnothing$ a A jsou stochasticky nezávislé jevy.

b) Ω a A jsou stochasticky nezávislé jevy.

c) Jsou-li A, B stochasticky nezávislé jevy, pak jsou stochasticky nezávislé též jevy $\overline{A}, B$ a $A, \overline{B}$ a $\overline{A}, \overline{B}$.

Důkaz:

ad a) $P(\varnothing \cap A) = P(\varnothing)P(A) : 0 = 0 \cdot P(A) = 0$

ad b) $P(\Omega \cap A) = P(\Omega)P(A) : P(A) = 1 \cdot P(A) = P(A)$

ad c) $P(\overline{A} \cap B) = P(B - (A \cap B)) = P(B) - P(A \cap B) = P(B) - P(A)P(B) = P(B)[1 - P(A)] = P(\overline{A})P(B)$

Tvrzení pro jevy $A, \overline{B}$ se dokáže analogicky.

$$P(\overline{A} \cap \overline{B}) = P(\overline{A \cup B}) = 1 - P(A \cup B) = 1 - [P(A) + P(B) - P(A \cap B)] = 1 - P(A) - P(B) + P(A)P(B) = 1 - P(A) - P(B)[1 - P(A)] =$$
$$= [1 - P(A)][1 - P(B)] = P(\overline{A})P(\overline{B})$$

Stochasticky nezávislé jevy

Definice: Nechť $(\Omega, \mathcal{A}, P)$ je pravděpodobnostní prostor. Jevy $A_1, ..., A_n \in \mathcal{A}$ jsou stochasticky nezávislé (vzhledem k P), jestliže platí systém multiplikativních vztahů:

$\forall 1 \le i < j \le n:$ $P(A_i \cap A_j) = P(A_i)\,P(A_j)$ (dvojmístný multiplikativní vztah)

$\forall 1 \le i < j < k \le n:$ $P(A_i \cap A_j \cap A_k) = P(A_i)\,P(A_j)\,P(A_k)$ (trojmístný multiplikativní vztah)

$\vdots$

$P(A_1 \cap \ldots \cap A_n) = P(A_1) \ldots P(A_n)$ (n-místný multiplikativní vztah)

Jevy $A_1, A_2, \ldots \in \mathcal{A}$ jsou stochasticky nezávislé (vzhledem k P), jestliže pro všechna přirozená n jsou stochasticky nezávislé jevy $A_1, ..., A_n \in \mathcal{A}$.

Příklad: V osudí jsou 4 lístky s číslicemi 000, 011, 110 a 101. Označme A_i jev, že na náhodně vytaženém lístku je 1 na i-tém místě. Zjistěte, zda jevy A_1, A_2, A_3 jsou stochasticky nezávislé.

Řešení: $P(A_1)=\frac{1}{2}$, $P(A_2)=\frac{1}{2}$, $P(A_3)=\frac{1}{2}$, $P(A_1 \cap A_2)=\frac{1}{4}$, $P(A_1 \cap A_3)=\frac{1}{4}$, $P(A_2 \cap A_3)=\frac{1}{4}$. Vidíme, že dvoumístné multiplikativní vztahy jsou splněny, avšak trojmístný vztah nikoli, neboť $P(A_1 \cap A_2 \cap A_3)=0$ a $P(A_1)P(A_2)P(A_3)=\frac{1}{8}$. Jevy A_1, A_2, A_3 nejsou stochasticky nezávislé.

Příklad

Ukázka příkladu, kdy jsou jevy po dvou nezávislé, ale jsou celkově závislé. Uvažujme náhodný pokus „hod dvěmi mincemi“, kdy sledujeme zda na mincích padl líc (L) nebo (R). Množina všech možných výsledků (elementárních jevů) je tedy $\Omega = \{LL, LR, RL, RR\}$ a všechny elementární jevy jsou stejně pravděpodobné, tj. mají pravděpodobnost $\frac{1}{4}$.

Najděte pravděpodobnost a zjistěte zda jsou nezávislé a po dvou nezávislé jevy

(a) A_1 na první mince padne líc;

(b) A_2 na druhé minci padne líc;

(c) A_3 na obou mincích padne totéž.

Řešení: $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$, $\mathcal{P}(\omega_i) = 1/4$ $A_1 = \{\omega_1, \omega_2\}$, $\mathcal{P}(A_1) = 1/2$
$A_2 = \{\omega_1, \omega_3\}$, $\mathcal{P}(A_2) = 1/2$
$A_3 = \{\omega_1, \omega_4\}$, $\mathcal{P}(A_3) = 1/2$
jevy A_1 a A_2 jsou nezávislé, protože $A_1 \cap A_2 = \{\omega_1\}$ a

$$\mathcal{P}(A_1 \cap A_2) = 1/4 = 1/2 \cdot 1/2$$

jevy A_1 a A_3 jsou nezávislé, protože $A_1 \cap A_3 = \{\omega_1\}$ a

$$\mathcal{P}(A_1 \cap A_3) = 1/4 = 1/2 \cdot 1/2$$

jevy A_2 a A_3 jsou nezávislé, protože $A_2 \cap A_3 = \{\omega_1\}$ a

$$\mathcal{P}(A_2 \cap A_3) = 1/4 = 1/2 \cdot 1/2$$

jevy A_1,A_2 a A_3 jsou závislé, protože $A_1 \cap A_2 \cap A_3 = \{\omega_1\}$ a

$$\mathcal{P}(A_1 \cap A_2 \cap A_3) = 1/4 \neq 1/2 \cdot 1/2 \cdot 1/2$$

Příklad

Mohou být neslučitelné (disjunktní) jevy A a B nezávislé?

Řešení: $0 = P(\varnothing) = P(A \cap B) = P(A) \cdot P(B)$

tedy disjunktní jevy mohou být nezávislé, jen když alespoň jeden z nich má nulovou pravděpodobnost.

Stochasticky nezávislé jevy

Příklad: Zjistěte, zda existuje jev, který je stochasticky nezávislý sám se sebou.
Řešení: $P(A \cap A) = P(A)P(A)$, tedy $P(A) = P(A)^2$. To je možné jen tak, že $P(A) = 0$ nebo $P(A) = 1$.

Věta:
a) Jestliže z třídy n stochasticky nezávislých jevů vybereme libovolnou podtřídu r jevů ($2 \le r \le n$), dostaneme opět třídu stochasticky nezávislých jevů.
b) Stochastická nezávislost se neporuší, jestliže některé (nebo i všechny) jevy nahradíme jevy opačnými.
c) Jestliže z třídy n stochasticky nezávislých jevů vybereme r disjunktních podtříd jevů ($2 \le r \le n$) a členy těchto podtříd libovolně sjednotíme nebo pronikneme, pak vzniklá sjednocení a průniky jsou opět stochasticky nezávislé jevy.
d) Neslučitelné jevy nemohou být stochasticky nezávislé (pokud nemají všechny nulovou pravděpodobnost).
e) Nemožný jev je stochasticky nezávislý s každým jevem.
f) Jistý jev je stochasticky nezávislý s každým jevem.

Příklad

Příklad: Firma investovala do tří nezávislých projektů. Pravděpodobnost zisku z těchto projektů je 0,4, 0,5 a 0,7. Jaká je pravděpodobnost, že firma bude mít zisk

a) právě jedenkrát (jev A)
b) alespoň jedenkrát (jev B)
c) práv ě dvakrát (jev C)
d) aspoň dvakrát (jev D)
e) ze všech tří projektů (jev E)
f) ze žádného projektu? (jev F)

Řešení :

Označme A_i jev, že firma bude mít zisk z i-tého projektu, i = 1, 2, 3.

ad a)

$$P(A) = P(A_1)\cdot P(\overline{A}_2)\cdot P(\overline{A}_3) + P(\overline{A}_1)\cdot P(A_2)\cdot P(\overline{A}_3) + P(\overline{A}_1)\cdot P(\overline{A}_2)\cdot P(A_3) =$$

$$= 0{,}4\cdot 0{,}5\cdot 0{,}3 + 0{,}6\cdot 0{,}5\cdot 0{,}3 + 0{,}6\cdot 0{,}5\cdot 0{,}7 = 10^{-3}(60 + 90 + 210) = 0{,}36$$

ad b)

$$P(B) = 1 - P(\overline{A_1}\cap\overline{A_2}\cap\overline{A_3}) = 1 - P(\overline{A_1})\,P(\overline{A_2})P(\overline{A_3}) = 1 - 0{,}6\cdot 0{,}5\cdot 0{,}3 = 1 - 0{,}09 = 0{,}91$$

ad c)

$$P(C) = P(A_1)\cdot P(A_2)\cdot P(\overline{A}_3) + P(A_1)\cdot P(\overline{A_2})\cdot P(A_3) + P(\overline{A}_1)\cdot P(A_2)\cdot P(A_3) =$$

$$= 0{,}4\cdot 0{,}5\cdot 0{,}3 + 0{,}4\cdot 0{,}5\cdot 0{,}7 + 0{,}6\cdot 0{,}5\cdot 0{,}7 = 10^{-3}(60 + 140 + 210) = 0{,}41$$

ad d)

$$P(D) = P(C) + P(A_1\cap A_2\cap A_3) = P(C) + P(A_1)\cdot P(A_2)\cdot P(A_3) = 0{,}41 + 0{,}4\cdot 0{,}5\cdot 0{,}7 =$$

$$= 0{,}41 + 0{,}14 = 0{,}55$$

ad e)

$$P(E) = P(A_1\cap A_2\cap A_3) = P(A_1)\cdot P(A_2)\cdot P(A_3) = 0{,}4\cdot 0{,}5\cdot 0{,}7 = 0{,}14$$

ad f)

$$P(F) = P(\overline{A_1}\cap\overline{A_2}\cap\overline{A_3}) = P(\overline{A_1})\,P(\overline{A_2})\,P(\overline{A_3}) = 0{,}6\cdot 0{,}5\cdot 0{,}3 = 0{,}09$$

Podmíněná pravděpodobnost. Geometrická pravděpodobnost.

Motivace: Opakovaně nezávisle provádíme týž náhodný pokus a sledujeme nastoupení jevu A v těch pokusech, v nichž nastoupil jev H. Podmíněnou relativní četnost A za podmínky H jsme v popisné statistice zavedli vztahem $p(A/H) = \frac{p(A \cap H)}{p(H)}$. Tato podmíněná relativní četnost se s rostoucím počtem pokusů ustaluje kolem konstanty P(A/H), kterou považujeme za podmíněnou pravděpodobnost jevu A za podmínky H.

Definice: Nechť $(\Omega, \mathcal{A}, P)$ je pravděpodobnostní prostor, $H \in \mathcal{A}$. jev s nenulovou pravděpodobností. Podmíněnou pravděpodobností za podmínky H rozumíme funkci
$P(./H): \mathcal{A} \rightarrow R$ danou vzorcem: $\forall A \in \mathcal{A}: P(A/H) = \frac{P(A \cap H)}{P(H)}$.

Podmíněná pravděpodobnost

Věta: Podmíněná pravděpodobnost je pravděpodobnost ve smyslu axiomatické definice a kromě toho pro ni platí:

a) $P(A_1 \cap A_2) = P(A_1)\,P(A_2/A_1)$ pro $P(A_1) \neq 0$.

b) $P(A_1 \cap A_2) = P(A_2)\,P(A_1/A_2)$ pro $P(A_2) \neq 0$.

c) Jevy A_1, A_2 jsou stochasticky nezávislé, právě když $P(A_1/A_2) = P(A_1)$ nebo $P(A_2) = 0$ a právě když $P(A_2/A_1) = P(A_2)$ nebo $P(A_1) = 0$.

Důkaz:

Stačí ověřit platnost axiómů P2, P10, P15.

ad a), ad b) Plyne přímo z definičního vzorce.

ad c) Nechť A_1, A_2 jsou stochasticky nezávislé $\Rightarrow P(A_1/A_2) = \frac{P(A_1 \cap A_2)}{P(A_2)} = \frac{P(A_1)P(A_2)}{P(A_2)} = P(A_1)$.

Nechť naopak $P(A_1/A_2) = P(A_1)$. Z definice: $P(A_1/A_2) = \frac{P(A_1 \cap A_2)}{P(A_2)} = P(A_1) \Rightarrow P(A_1 \cap A_2) = P(A_1)P(A_2)$, tedy A_1, A_2 jsou stochasticky nezávislé.

Příklad

Příklad: Jaká je pravděpodobnost, že při hodu kostkou padlo sudé číslo, je -li známo, že padlo číslo menší než 5?

Řešení: $\Omega = \{\omega_1, \ldots, \omega_6\}$, A … padlo sudé číslo, $A = \{\omega_2, \omega_4, \omega_6\}$, H … padlo číslo menší než 5, $H = \{\omega_1, \omega_2, \omega_3, \omega_4\}$, $A \cap H = \{\omega_2, \omega_4\}$

$$P(A/H) = \frac{P(A \cap H)}{P(H)} = \frac{\frac{2}{6}}{\frac{4}{6}} = \frac{1}{2}$$

Příklad: Dvakrát hodíme kostkou. Jaká je pravděpodobnost, že součet přesáhne 10, víme-li, že padla (aspoň jedna) šestka?

Řešení:

$$P(A|H) = \frac{\frac{|\{[6,5],[5,6],[6,6]\}|}{6 \cdot 6}}{\frac{2 \cdot 5 + 1}{6 \cdot 6}} = \frac{3}{11}$$

Věta o násobení pravděpodobností

Věta: (Věta o násobení pravděpodobností)
Nechť $(\Omega, \mathcal{A}, P)$ je pravděpodobnostní prostor, $A_1, A_2, ..., A_n$ takové jevy, že $P(A_1 \cap ... \cap A_{n-1}) \neq 0$.
Pak $P(A_1 \cap A_2 \cap ... \cap A_n) = P(A_1)\, P(A_2/A_1)\, P(A_3/A_1 \cap A_2) ... P(A_n/A_1 \cap ... \cap A_{n-1})$.

Důkaz: Matematickou indukcí. Předpokládáme, že vztah platí pro libovolné přirozené $n \geq 2$ a dokážeme jeho platnost pro n+1:

$$P(A_1 \cap \ldots \cap A_n \cap A_{n+1}) = P\left(\bigcap_{i=1}^{n} A_i \cap A_{n+1}\right) = P\left(\bigcap_{i=1}^{n} A_i\right) P\left(A_{n+1} / \bigcap_{i=1}^{n} A_i\right) = P(A_1) P(A_2 / A_1) \ldots P(A_{n+1} / A_1 \cap \ldots \cap A_n)$$

Příklad: Ze skupiny 100 výrobků, která obsahuje 10 zmetků, vybereme náhodně bez vracení 3 výrobky. Vypočtěte pravděpodobnost jevu, že první dva výrobky budou kvalitní a třetí bude zmetek.

Řešení:
Jev A_i znamená, že i-tý vybraný výrobek je kvalitní, i = 1, 2, 3.
Počítáme $P(A_1 \cap A_2 \cap \overline{A_3}) = P(A_1)\, P(A_2/A_1)\, P(\overline{A_3}/A_1 \cap A_2) = \frac{90}{100}\frac{89}{99}\frac{10}{98} = 0{,}083.$

Věta o úplné pravděpodobnosti, Bayesův vzorec

Věta (vzorec pro výpočet úplné pravděpodobnosti a Bayesův vzorec)
Nechť $(\Omega, \mathcal{A}, P)$ je pravděpodobnostní prostor, $H_i \in \mathcal{A}$, $i \in I$ (I je nejvýše spočetná indexová množina) takové jevy, že $P(H_i) > 0$, $\bigcup_{i\in I} H_i = \Omega$, $H_i \cap H_j = \varnothing$ pro $i \neq j$ (říkáme, že jevy H_i, $i \in I$ tvoří úplný systém hypotéz).

a) Pro libovolný jev $A \in \mathcal{A}$ platí vzorec úplné pravděpodobnosti: $P(A) = \sum_{i\in I} P(H_i)P(A/H_i)$

b) Pro libovolnou hypotézu H_k, $k \in I$ a jev $A \in \mathcal{A}$ s nenulovou pravděpodobností platí Bayesův vzorec:

$$P(H_k/A) = \frac{P(H_k)P(A/H_k)}{P(A)}.$$

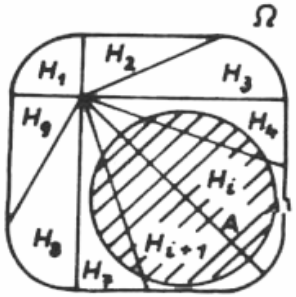
($P(H_k/A)$ se nazývá aposteriorní pravděpodobnost hypotézy H_k, $P(H_k)$ je apriorní pravděpodobnost.)

Důkaz:
ad a) Jev A vyjádříme jako sjednocení neslučitelných jevů: $A = \bigcup_{i\in I} (A \cap H_i)$.

Pak $P(A) = P\left(\bigcup_{i\in I} (A \cap H_i)\right) = \sum_{i\in I} P(A \cap H_i) = \sum_{i\in I} P(H_i)P(A/H_i)$

ad b) $P(H_k/A) = \frac{P(H_k \cap A)}{P(A)} = \frac{P(H_k)P(A/H_k)}{P(A)}$

Ilustrace vzorce pro úplnou pravděpodobnost



Příklad

1) Bez vracení taháme z urny s *a* černými a *b* bílými koulemi. Jaká je pravděpodobnost, že ve druhém tahu vytáhneme černou kouli, jestliže v prvním tahu jsme vytáhli kouli bílou?

Řešení: $$P(A\,|\,H)=\frac{\frac{b}{a+b}\cdot\frac{a}{a+b-1}}{\frac{b}{a+b}\cdot\frac{a+b-1}{a+b-1}}=\frac{a}{a+b-1}$$

2) V dostihu zvítězí kůň A (B) s pravděpodobností 0,5 (0,3). Kůň A ztratil na startu příliš a je jisté, že nezvítězí. Jaká je nyní pravděpodobnost, že zvítězí B?

Řešení:

$$P(A\,|\,\overline{H})=\frac{P(A\cap\overline{H})}{P(\overline{H})}=\frac{P(A)}{1-P(H)}=\frac{0{,}3}{0{,}5}=0{,}6$$

Příklad

V první urně je 6 bílých a 2 černé koule, ve druhé jsou 4 bílé a 2 černé koule. Náhodně zvolíme urnu a vytáhneme jednu kouli. Jaká je pravděpodobnost, že bude bílá?

Řešení:

Pravděpodobnost tahu z první (resp. druhé) urny, je 1/2. Označíme-li B =[tah bílé koule], U_i = [tah z urny i], je podle věty o celkové pravděpodobnosti

$$P(B) = P(B \mid U_1) \cdot P(U_1) + P(B \mid U_2) \cdot P(U_2) = \frac{6}{6+2} \cdot \frac{1}{2} + \frac{4}{4+2} \cdot \frac{1}{2} = \frac{17}{24} = 0{,}708$$

Příklad

Automat X vyrobí za směnu dvakrát více výrobku než automat Y. Pravděpodobnost vzniku zmetku je u automatu X 0,02, u Y 0,05. Po skončení směny se výrobky ukládají do jedné bedny. Jaká je pravděpodobnost, že výrobek náhodně vybraný z této bedny není zmetek?

Řešení:

Podle věty o celkové pravděpodobnosti (poměr výrobků v bedně je 2 : 1 ve prospěch automatu X, tj. 2/3 výrobků pochází od X a 1/3 od Y)

$$P(A) = \frac{2}{3} \cdot 0{,}98 + \frac{1}{3} \cdot 0{,}95 = \frac{2{,}91}{3} = 0{,}97$$

Příklad

Mezi 20 střelci jsou 4 výborní, 10 dobrých a 6 průměrných s pravděpodobnostmi zásahu 0,9, 0,7 a 0,5. Jaká je pravděpodobnost, že dva náhodně vybraní střelci oba zasáhnou cíl?

Řešení:

Podle toho, která dvojice bude vybrána

$$P(A) = (0{,}9 \cdot 0{,}9) \cdot \frac{4 \cdot 3}{20 \cdot 19} + (0{,}9 \cdot 0{,}7) \cdot \frac{4 \cdot 10}{20 \cdot 19} + \ldots + (0{,}5 \cdot 0{,}5) \cdot \frac{10 \cdot 9}{20 \cdot 19} = 0{,}46$$

Věta o úplné pravděpodobnosti, Bayesův vzorec

Thomas Bayes (1702 – 1761) : Presbytariánský duchovní

Poznámka (Návod na použití vzorce pro výpočet úplné pravděpodobnosti a Bayesova vzorce)
Nejprve podle textu úlohy stanovíme úplný systém hypotéz, tj,. jevy, které se navzájem vylučují a přitom vyčerpávají všechny možnosti.
V úlohách vedoucích na vzorec pro výpočet úplné pravděpodobnosti se zajímáme o pravděpodobnost jevu, který s hypotézami nesouvisí, zatímco v úlohách vedoucích na Bayesův vzorec nás zajímá pravděpodobnost některé hypotézy za podmínky, že nastal jev, který s hypotézami nesouvisí.

Příklad

Příklad: Test obsahuje 100 otázek. Zkoušený si nejprve vylosuje otázku a pak si jeho postup zjednodušeně představíme takto: zná-li správnou odpověď, zatrhne ji. Nezná-li správnou odpověď, zvolí se stejnou pravděpodobností kteroukoliv ze čtyř možných odpovědí. Předpokládejme, že ve skutečnosti zná zkoušený právě k správných odpovědí.

a) S jakou pravděpodobností správně odpoví?

b) S jakou pravděpodobností je při správné odpovědi pravdivé tvrzení, že zkoušený ve skutečnosti jenom hádal?

Řešení: H_1 … zkoušený zná správnou odpověď, H_2 … zkoušený nezná správnou odpověď, A … zkoušený správně odpoví

$$P(H_1)=\frac{k}{100}, P(H_{21})=\frac{100-k}{100}, P(A/H_1)=1, P(A/H_2)=\frac{1}{4}$$

ad a) $$P(A)=P(H_1)P(A/H_1)+P(H_2)P(A/H_2)=\frac{k}{100}\cdot 1+\frac{100-k}{100}\cdot\frac{1}{4}=\frac{3k+100}{400}$$

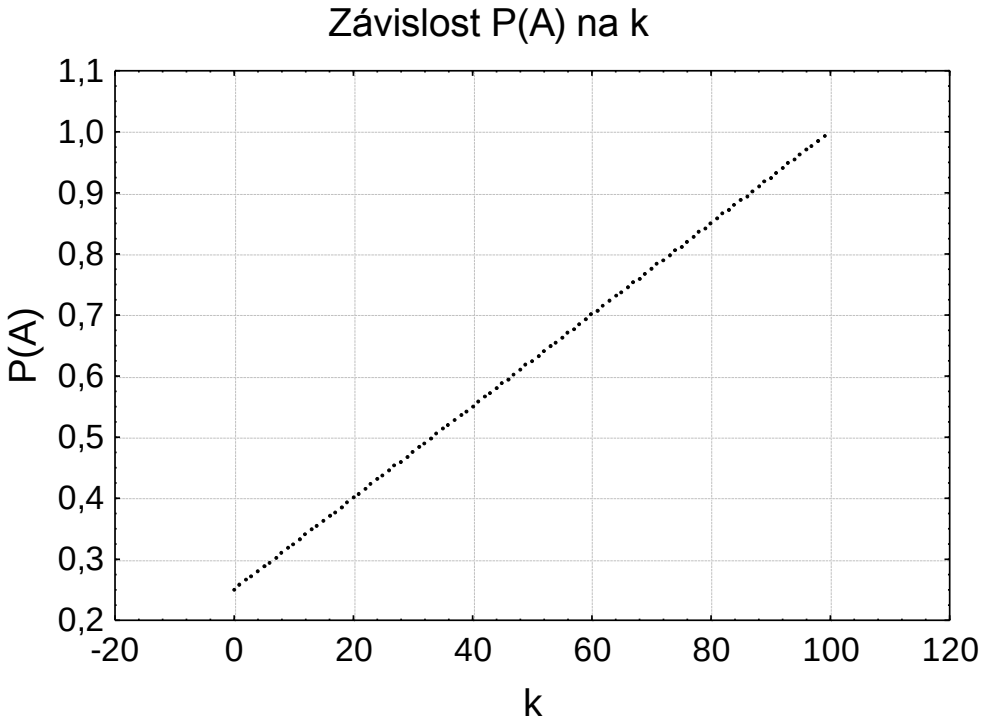
ad b) $$P(H_2/A)=\frac{P(H_2)P(A/H_2)}{P(A)}=\frac{\frac{100-k}{100}\cdot\frac{1}{4}}{\frac{3k+100}{400}}=\frac{100-k}{3k+100}$$

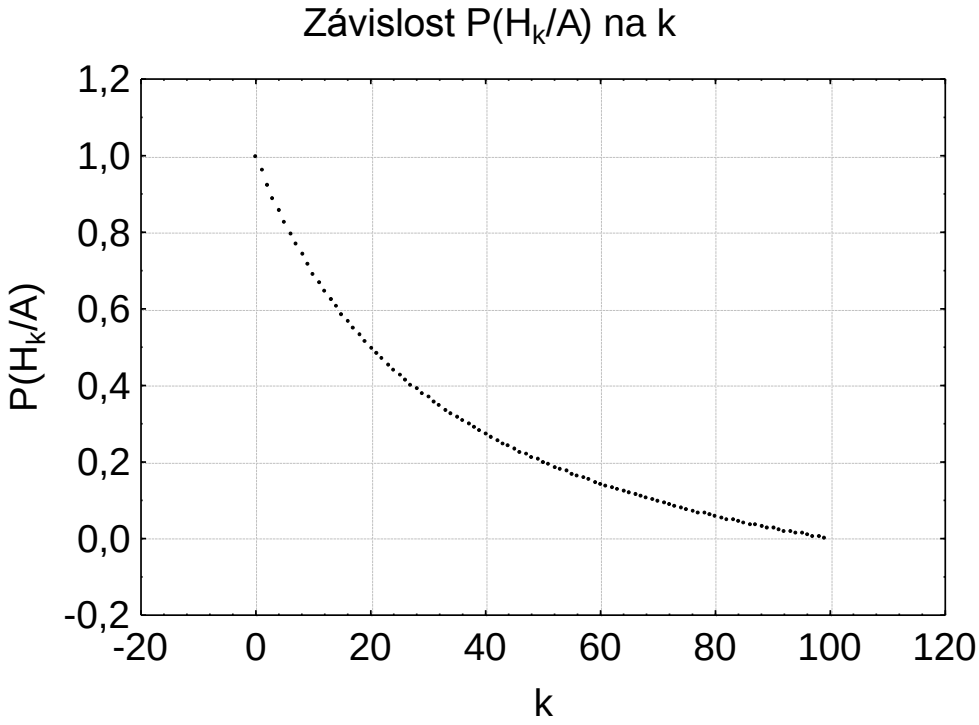
k	0	10	50	90
P(A)	0,25	0,325	0,625	0,925
$P(H_2/A)$	1	0,692	0,2	0,027

Příklad

$$P(A) = \frac{3k + 100}{400}$$

$$P(H_2 / A) = \frac{100 - k}{3k + 100}$$





Příklad

Příklad: K osevu byly vybrány dvě odrůdy pšenice, a to 20% první odrůdy a 80% druhé odrůdy. Pravděpodobnost, že ze zrna vyroste klas, je pro první odrůdu 0,95 a pro druhou odrůdu 0,98. Jaká je pravděpodobnost, že

a) z náhodně vybraného zrna vyroste klas?
b) náhodně vybrané zrno, z něhož vyrostl klas, pocházelo z první odrůdy pšenice?
c) náhodně vybrané zrno, z něhož vyrostl klas, pocházelo z druhé odrůdy pšenice?
d) náhodně vybrané zrno, z něhož nevyrostl klas, pocházelo z první odrůdy pšenice?
e) náhodně vybrané zrno, z něhož nevyrostl klas, pocházelo z druhé odrůdy pšenice?

Řešení:

Jev A … z náhodně vybraného zrna vyroste klas
Jev H_1 … zrno pochází z první odrůdy pšenice
Jev H_2 … zrno pochází z druhé odrůdy pšenice

$$P(H_1)=0{,}2,\ P(A|H_1)=0{,}95,\ P(H_2)=0{,}8,\ P(A|H_2)=0{,}98$$

ad a) $$P(A)=P(H_1)P(A/H_1)+P(H_2)P(A/H_2)=0{,}2\cdot 0{,}95+0{,}8\cdot 0{,}98=0{,}19+0{,}784=0{,}974$$

ad b) $$P(H_1/A)=\frac{P(H_1)P(A/H_1)}{P(A)}=\frac{0{,}2\cdot 0{,}95}{0{,}974}=0{,}1951$$

ad c) $$P(H_2/A)=\frac{P(H_2)P(A/H_2)}{P(A)}=\frac{0{,}8\cdot 0{,}98}{0{,}974}=0{,}8049$$

ad d) $$P(H_1/\overline{A})=\frac{P(H_1)P(\overline{A}/H_1)}{P(\overline{A})}=\frac{0{,}2\cdot 0{,}05}{1-0{,}974}=\frac{0{,}01}{0{,}026}=0{,}3846$$

ad e) $$P(H_2/\overline{A})=\frac{P(H_2)P(\overline{A}/H_2)}{P(\overline{A})}=\frac{0{,}8\cdot 0{,}02}{1-0{,}974}=\frac{0{,}016}{0{,}026}=0{,}6154$$

Příklad

1) Jeden ze 3 střelců s pravděpodobnostmi zásahu 0,3, 0,5, 0,8 vystřelil a zasáhl. Jaká je pravděpodobnost, že střílel druhý střelec?

Řešení:

$$P(A) = \frac{0{,}5 \cdot \frac{1}{3}}{0{,}3 \cdot \frac{1}{3} + 0{,}5 \cdot \frac{1}{3} + 0{,}8 \cdot \frac{1}{3}} = \frac{5}{16} = 0{,}3125$$

2) Mezi 20 střelci je 5 výborných, 9 dobrých a 6 průměrných s pravděpodobnostmi zásahu 0,9, 0,8 a 0,7. Náhodně vybraný střelec ze 2 ran trefil jednou. Jaká je pravděpodobnost, že šlo o výborného (dobrého, průměrného) střelce?

Řešení:

$$P(\text{byl to výborný}) = \frac{2 \cdot 0{,}9 \cdot 0{,}1 \cdot \frac{5}{20}}{2 \cdot 0{,}9 \cdot 0{,}1 \cdot \frac{5}{20} + 2 \cdot 0{,}8 \cdot 0{,}2 \cdot \frac{9}{20} + 2 \cdot 0{,}7 \cdot 0{,}3 \cdot \frac{6}{20}} = 0{,}143$$

$$P(\text{byl to dobrý}) = 0{,}457$$

$$P(\text{byl to průměrný}) = 0{,}4$$

Příklad

Víme-li, že pravděpodobnost odhalení AIDS při testu je 0,999, že pravděpodobnost správného otestování zdravého jedince je 0,99 a že AIDS se vyskytuje u 0,006 lidí, jaká je pravděpodobnost, že člověk, u kterého byl test pozitivní, AIDS skutečně má?

Řešení:

Označíme-li A = [má AIDS], T = [test říká AIDS], známe P(T |A) = 0, 999, $P(\overline{T} \mid \overline{A}) = 0{,}99$, P(A) = 0, 006. Bayesova věta nám dá

$$P(A \mid T) = \frac{P(T \mid A)P(A)}{P(T \mid A)P(A) + P(T \mid \overline{A})P(\overline{A})}$$

$$= \frac{0{,}999 \cdot 0{,}006}{0{,}999 \cdot 0{,}006 + (1 - 0{,}99) \cdot (1 - 0{,}006)} = 0{,}376$$

Geometrická pravděpodobnost

Motivace: V některých situacích je vhodné zvolit za základní prostor nikoliv obecnou množinu Ω, ale n-rozměrný prostor $\mathbf{R}^n$ a za možné výsledky reálné vektory $(x_1,\ldots,x_n)$. Za jevové pole však nevezmeme systém všech podmnožin prostoru $\mathbf{R}^n$ (ten totiž obsahuje i tzv. neměřitelné množiny), ale méně podrobné borelovské pole $\mathcal{B}^n$.

Émile Borel (1871 – 1956) – francouzský matematik a politik. Zabýval se teorií míry, teorií pravděpodobnosti A teorií her. Byl poslancem francouzského parlamentu a ministrem námořnictva.

Na borelovském poli pak speciálním způsobem zavedeme geometrickou pravděpodobnost a dostaneme pravděpodobnostní prostor $(\mathbf{R}^n, \mathcal{B}^n, Q)$.

Borelovské pole, Borelovské množiny

Definice

Nechť n je přirozené číslo. Množinu $R^n = (-\infty, \infty) \times \ldots \times (-\infty, \infty) = (-\infty, \infty)^n$ nazýváme **n-rozměrným prostorem**. Minimální jevové pole na R^n obsahující třídu všech polouzavřených intervalů typu $(-\infty, x_1\rangle \times \ldots \times (-\infty, x_n\rangle$ pro $(x_1, \ldots x_n) \in R^n$ nazýváme **n-rozměrným borelovským polem** $\mathcal{B}^n$ a prvky tohoto pole nazýváme **(n-rozměrnými) borelovskými množinami**. Dvojice $(R^n, \mathcal{B}^n)$ je tedy měřitelný prostor.

(Není podstatné, že borelovské pole je generováno právě intervaly typu $(-\infty, x_1\rangle \times \ldots \times (-\infty, x_n\rangle$. Mohlo by být generováno i jinými typy intervalů.)

Věta: Borelovské pole je jevové pole, tzn., že splňuje axiómy J2, J6, J8.

Věta: Mezi borelovské množiny náleží zejména prázdná množina, celý základní prostor, všechny jednobodové, konečné a spočetné množiny, intervaly všech typů, všechny uzavřené a otevřené oblasti a všechna konečná a spočetná sjednocení a průniky těchto množin. Rovněž kartézský součin borelovských množin je borelovská množina, ovšem vyšší dimenze.

Borelovsky měřitelná zobrazení, Borelovské funkce

Definice

Nechť $(\Omega, \mathcal{A}), (R^n, \mathcal{B}^n)$ jsou měřitelné prostory. Zobrazení $\mathbf{X} : \Omega \mapsto R^n$ se nazývá **borelovsky měřitelné** (vzhledem k $\mathcal{A}$), právě když úplný vzor každé n-rozměrné borelovské množiny je jev, tj.

$\forall B \in \mathcal{B}^n : \mathbf{X}^{inv}(B) = \{\omega \in \Omega; X(\omega) \in B\} \in \mathcal{A}$.

Ve speciálním případě, kdy $\Omega = R^m$ a $\mathcal{A} = \mathcal{B}^m, \mathbf{X} = \mathbf{g} = (g_1, \ldots, g_n)$, tj.

$\forall B \in \mathcal{B}^n : \mathbf{g}^{inv}(B) =$

$\{(x_1, \ldots, x_m) \in R^m; (g_1(x_1, \ldots, x_m), \ldots, g_n(x_1, \ldots, x_m)) \in B\} \in \mathcal{B}^m$, hovoříme o **borelovské funkci**.

Věta: Mezi borelovské funkce náleží zejména všechny spojité a po částech spojité funkce. Rovněž limita všude konvergentní posloupnosti borelovských funkcí je borelovská funkce.

Definice:

Nechť $(R^n, \mathcal{B}^n)$ je měřitelný prostor a $G \in \mathcal{B}^n$ je borelovská množina. **Objemem** borelovské množiny G rozumíme číslo

$mes(G) = \int_G \ldots \int dx_1 \ldots dx_n$, pokud Riemannův integrál vpravo existuje.

Geometrická pravděpodobnost

Definice:

Nechť objem $mes(G)$ borelovské množiny G je nenulový a konečný. **Geometrickou pravděpodobností** soustředěnou na množině G rozumíme funkci $Q : \mathcal{B}^n \mapsto R$ danou vzorcem

$\forall B \in \mathcal{B}^n, B \subseteq G : Q(B) = \frac{mes(B)}{mes(G)}$, pokud $mes(B)$ existuje.

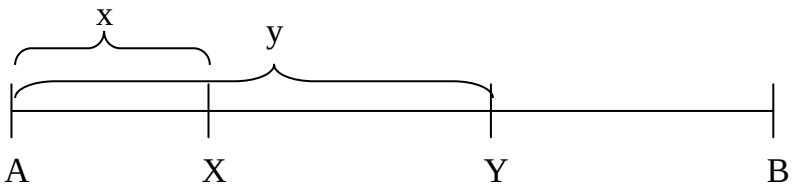
Věta: Geometrická pravděpodobnost je pravděpodobnost ve smyslu axiomatické definice, tj. splňuje axiómy P2, P10, P15. Trojice ($\mathbf{R}^n$, $\mathcal{B}^n$, Q) je tedy pravděpodobnostní prostor.

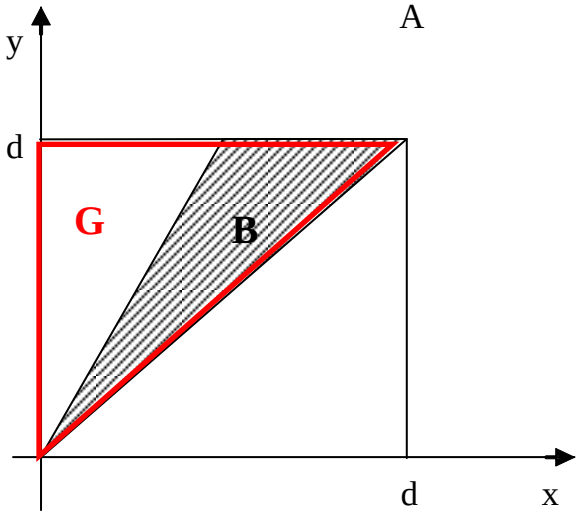
Příklad

Příklad: Na úsečce AB délky d jsou náhodně zvoleny body X a Y, přičemž vzdálenost bodu X od bodu A je menší než vzdálenost bodu Y od bodu A. Jaká je pravděpodobnost, že délka úsečky AX je větší než délka úsečky XY?

Řešení :

$G = \{(x, y) \in R^2; 0 \le x \le d, 0 \le y \le d, x \le y\}$ $\quad B = \{(x, y) \in G; x > y - x\}$





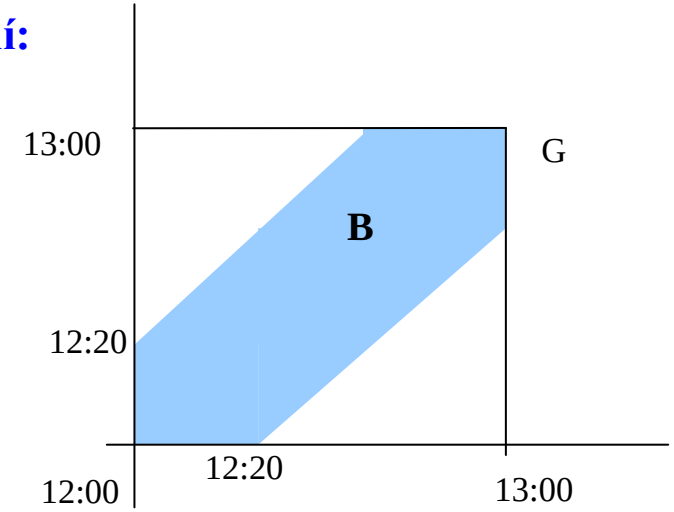
$$mes(G) = \frac{d^2}{2}, \; mes(B) = \frac{d^2}{2} - \frac{\frac{d}{2} \cdot d}{2} = \frac{d^2}{4}, \; Q(B) = \frac{mes(B)}{mes(G)} = \frac{1}{2}$$

Délka úsečky AX je větší než délka úsečky XY s pravděpodobností 0,5.

Příklad

Dívka a chlapec si smluvili schůzku mezi 12:00 a 13:00. Přijdou náhodně v tomto rozmezí a čekají na sebe 20 minut, nejdéle však do 13:00. Jaká je pravděpodobnost, že se setkají?

Řešení:



$$mes(G) = 1$$

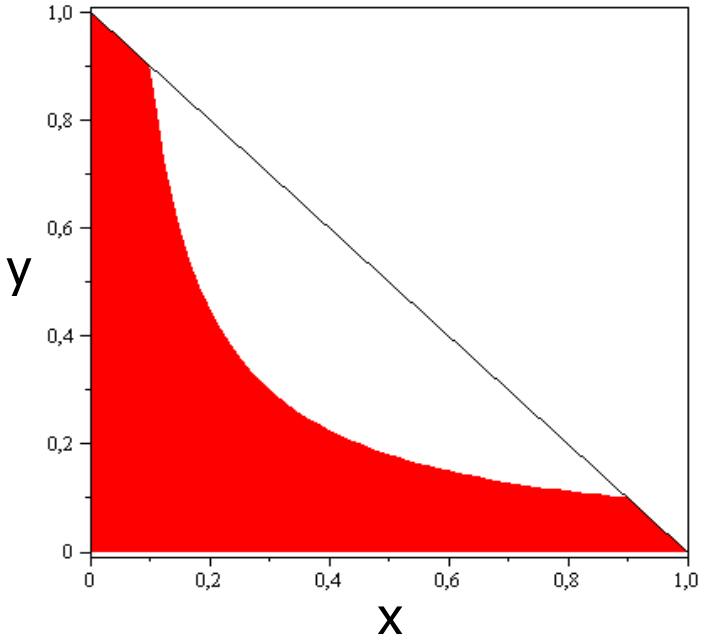
$$mes(B) = \frac{5}{9}$$

$$Q(B) = \frac{5}{9}$$

Příklad

Volíme náhodně dvě čísla z intervalu (0,1). Jaká je pravděpodobnost, že jejich součet je menší než jedna a současně jejich součin menší než 0,09?

Řešení



$$1-x=\frac{0{,}09}{x}$$

$$x^2-x+0{,}09=0$$

$$x_1=0{,}1,\ x_2=0{,}9$$

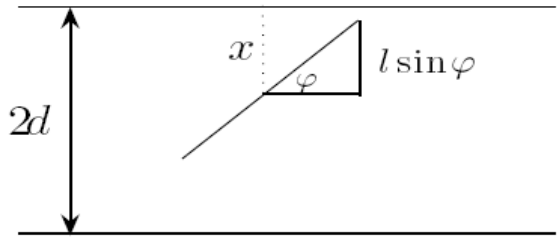
$$Q(B)=\frac{1}{2}-\int_{0{,}1}^{0{,}9}\left(1-x-\frac{0{,}09}{x}\right)dx=0{,}29775$$

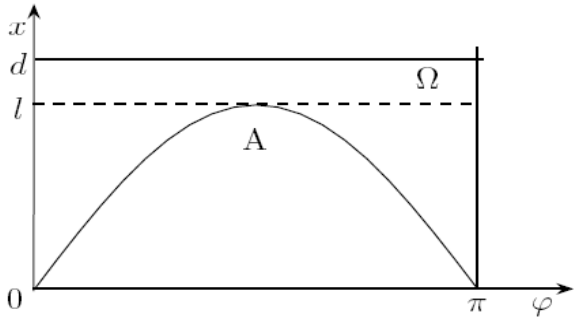
Příklad

Buffonova úloha. V rovině jsou rozmístěny rovnoběžky ve vzdálenosti d > 0. Na rovinu hodíme náhodně jehlu délky 0 < l < d. Jaká je pravděpodobnost, že jehla protne některou rovnoběžku?

Řešení

Předpokládejme, že náhodně znamená, že každá poloha (středu) a každá orientace jehly je stejně pravděpodobná a že tyto dvě nahodile proměnné jsou na sobě nezávislé. Nechť *x* je vzdálenost středu jehly od nejbližší rovnoběžky a φ je úhel, který jehla svírá s rovnoběžkami.





$$\Omega = \{0 \le \varphi < \pi, 0 \le x \le d\}$$

$$\mathrm{A} = \{(\varphi, x) \in \Omega : x \le l \sin\varphi\}$$

$$Q(A) = \frac{mes(A)}{mes(\Omega)} = \frac{\int_0^{\pi} l \sin\varphi \, d\varphi}{\pi d} = \frac{2l}{\pi d}$$

7. Náhodné veličiny

Motivace: Výsledky náhodného pokusu lze popsat reálnými čísly (resp. reálnými vektory) pomocí nějakého zobrazení $X : \Omega \to R$ ($\mathbf{X} = (X_1, \ldots, X_n) : \Omega \to \mathbf{R}^n$). Pokud bude toto zobrazení splňovat určité podmínky, nazveme ho náhodnou veličinou. Příklady náhodných veličin: počet členů náhodně vybrané domácnosti, počet chyb, jichž se dopustí nějaké zařízení za určitou dobu, doba do poruchy nějakého zařízení, hmotnost náhodně vybraného výrobku apod.

Vztah mezi znakem a náhodnou veličinou

Pojem „znak“, který jsme zavedli v popisné statistice, je sice blízký pojmu „náhodná veličina“, ale není s ním totožný. Znak může být považován za náhodnou veličinu, jestliže jeho hodnoty zjišťujeme na objektech, které byly vybrány ze základního souboru náhodně.

Definice:

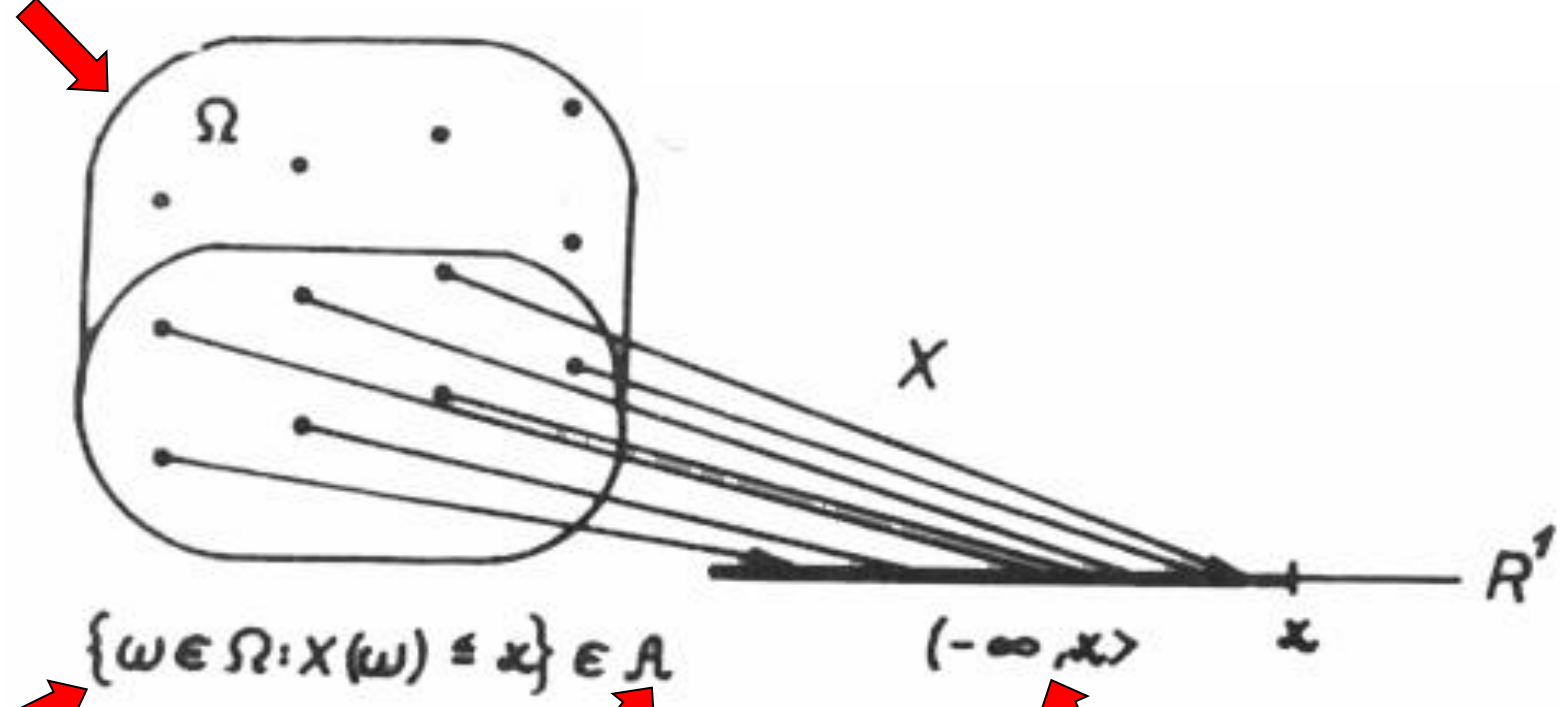
Nechť $(\Omega, \mathcal{A}), (R^n, \mathcal{B}^n)$ jsou měřitelné prostory. Zobrazení $\mathbf{X} : \Omega \mapsto R^n$ se nazývá **náhodná veličina** (vzhledem k $\mathcal{A}$), právě když je borelovsky měřitelné (vzhledem k $\mathcal{A}$). Pro $n = 1$ hovoříme o **skalární náhodné veličině**, pro $n \geq 2$ o **náhodném vektoru**. Přitom zobrazení $X_1 : \Omega \to R, \ldots, X_n : \Omega \mapsto R$ se nazývají **složky náhodného vektoru**. Obraz $\mathbf{X}(\omega) = (X_1(\omega), \ldots, X_n(\omega))$ se nazývá **číselná realizace** náhodné veličiny $\mathbf{X}$ příslušná možnému výsledku ω.

Ilustrace náhodné veličiny

Nechť $(\Omega, \mathcal{A}), (R^n, \mathcal{B}^n)$ jsou měřitelné prostory. Zobrazení $\mathbf{X} : \Omega \mapsto R^n$ se nazývá **borelovsky měřitelné** (vzhledem k $\mathcal{A}$), právě když úplný vzor každé n-rozměrné borelovské množiny je jev, tj.

$\forall B \in \mathcal{B}^n : \mathbf{X}^{inv}(B) = \{\omega \in \Omega; X(\omega) \in B\} \in \mathcal{A}.$

Základní prostor



Jev

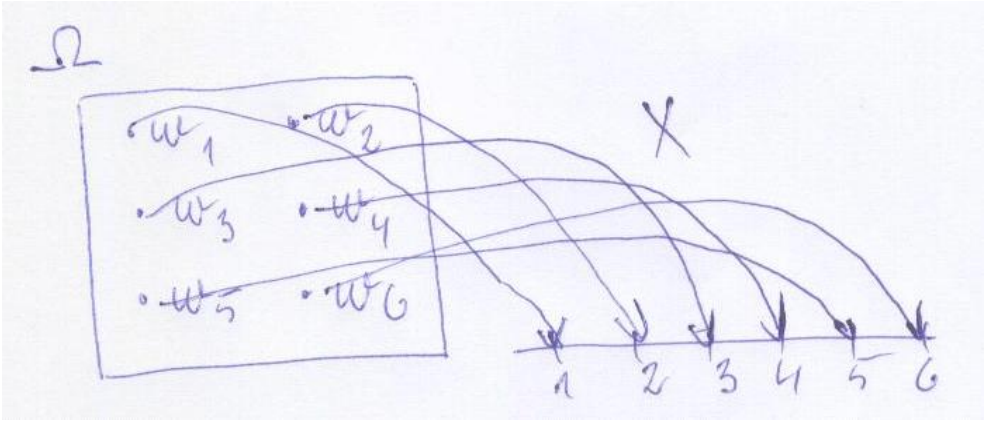
Jevové pole

Borelovská množina

Příklad

Příklad: Náhodný pokus spočívá v hodu kostkou. Základní prostor $\Omega = \{\omega_1, \ldots, \omega_6\}$. Uvážíme dvě jevová pole, a to $\mathcal{A}_{max} = \{A; A \subseteq \Omega\}$ a $\mathcal{A} = \{\Omega, \varnothing, \{\omega_1, \omega_3, \omega_5\}, \{\omega_2, \omega_4, \omega_6\}\}$. Zjistěte, zda zobrazení $X : \Omega \to R$, které poloze kostky číslem i nahoru přiřazuje číslo i, i = 1 , …, 6, je náhodná veličina vzhledem k $\mathcal{A}_{max}$ a vzhledem k $\mathcal{A}$.

Řešení:



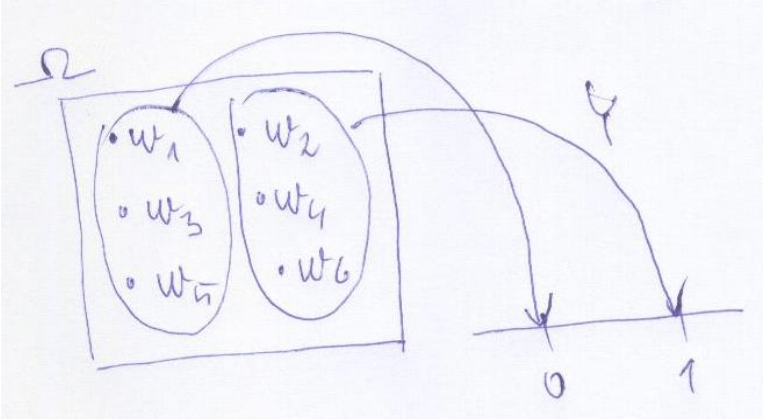
Zobrazení $X : \Omega \to R$ je náhodná veličina vzhledem k $\mathcal{A}_{max}$, protože úplný vzor každé borelovské množiny je jev vzhledem k $\mathcal{A}_{max}$. Vzhledem k $\mathcal{A}$ však X není náhodná veličina:

Úplný vzor množiny $(-\infty, 4\rangle$ je $\{\omega_1, \omega_2, \omega_3, \omega_4\} \notin A$.

Příklad

Zavedeme zobrazení $\Omega \to \mathrm{R}$, které poloze kostky lichým číslem nahoru přiřazuje 0 a sudým 1.

$\mathcal{A} = \{\Omega, \varnothing, \{\omega_1, \omega_3, \omega_5\}, \{\omega_2, \omega_4, \omega_6\}\}$



Toto zobrazení je náhodná veličina vzhledem k $\mathcal{A}$ a nazývá se ukazatel parity.

Náhodná veličina

Označení

a) Jestliže nehrozí nebezpečí nedorozumění, zapisujeme náhodnou veličinu i její číselnou realizaci týmž symbolem $\mathbf{X}$.

b) Množinu $\{\omega \in \Omega; \mathbf{X}(\omega) \in B\}$ zkráceně zapisujeme $\{\mathbf{X} \in B\}$ a čteme: náhodná veličina $\mathbf{X}$ se realizovala v borelovské množině B. Ve speciálním případě, kdy $B = \{\mathbf{x}\}$ resp. $B = (-\infty, \mathbf{x})$, píšeme $\{\mathbf{X} = \mathbf{x}\}$ resp. $\{\mathbf{X} \leq \mathbf{x}\}$.

c) Zápis pravděpodobnosti zkrátíme takto:

$P(\{\omega \in \Omega; \mathbf{X}(\omega) \in B\}) = P(\mathbf{X} \in B)$

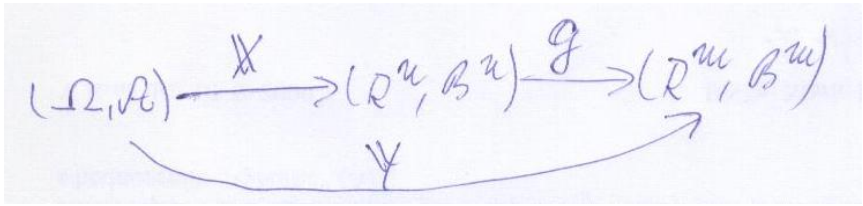
$P(\{\omega \in \Omega; \mathbf{X}(\omega) \in B\}/\{\omega \in \Omega; \mathbf{Y}(\omega) \in C\}) = P(\mathbf{X} \in B/\mathbf{Y} \in C)$.

Transformovaná náhodná veličina

Věta:

Nechť $(\Omega, \mathcal{A}), (R^n, \mathcal{B}^n), (R^m, \mathcal{B}^m)$ jsou měřitelné prostory. Nechť $\mathbf{X} : \Omega \mapsto R^n$ je náhodná veličina a $\mathbf{g} : R^n \mapsto R^m$ je borelovská funkce. Pak složené zobrazení $\mathbf{Y} : \Omega \mapsto R^m$ dané vzorcem $\forall \omega \in \Omega : \mathbf{Y}(\omega) = \mathbf{g}(\mathbf{X}(\omega))$ je náhodná veličina. Nazývá se **transformovaná náhodná veličina**, pro $m = 1$ skalární, pro $m \geq 2$ vektorová.

Důkaz:



Aby zobrazení $\mathbf{Y}: \Omega \to \mathrm{R}^m$ bylo náhodnou veličinou vzhledem k $\mathcal{A}$ musí platit:
$\forall B \in \mathcal{B}^m$: $\mathbf{Y}^{inv}(B) = \{\omega \in \Omega; \mathbf{Y}(\omega) \in B\} \in \mathcal{A}$. Nechť tedy $B \in \mathcal{B}^m$. Protože $\mathbf{g}$ je borelovská funkce, je $\mathbf{g}^{inv}(B) \in \mathcal{B}^n$. Protože $\mathbf{X}$ je náhodná veličina, je $\mathbf{X}^{inv}(\mathbf{g}^{inv}(B)) \in \mathcal{A}$. Ovšem $\mathbf{X}^{inv}(\mathbf{g}^{inv}(B)) = \mathbf{Y}^{inv}(B)$.

Transformovaná náhodná veličina

Poznámka: (Příklady transformovaných náhodných veličin) Nechť $\mathbf{X} = (X_1, \ldots, X_n)$ je náhodný vektor.
a) Nechť $\{i, \ldots, j\} = \{1, \ldots, n\} - \{k, \ldots, l\}$. Náhodný vektor $(X_i, \ldots, X_j)$ se nazývá vybraný marginální vektor, $(X_k, \ldots, X_l)$ se nazývá zbylý marginální vektor. Původní náhodný vektor $(X_1, \ldots, X_n)$ se v této souvislosti nazývá zbylý marginální vektor.
b) $\sum_{i=1}^{n} X_i, \max\{X_1, \ldots, X_n\}, \sin(X_i), \ldots$ jsou transformované náhodné veličiny.

Definice: Posloupnost $\{X_n\}_{n=1}^{\infty}$ spočetně mnoha náhodných veličin definovaných na témž měřitelném prostoru $(\Omega, \mathcal{A})$ se nazývá náhodná posloupnost.

Distribuční funkce náhodné veličiny

Motivace: Při pozorování realizací náhodné veličiny si povšimneme, že některé její hodnoty se vyskytují s větší pravděpodobností, jiné s menší. Pravděpodobnostní chování náhodné veličiny X budeme popisovat pomocí distribuční funkce , která udává pravděpodobnost jevu, že náhodná veličina X se realizuje hodnotou nejvýše x:

$$\forall x \in R : \Phi(x) = P(X \leq x)$$

Je to zidealizovaný protějšek empirické distribuční funkce zavedené v popisné statistice:

$$\forall x \in R : F(x) = \frac{N(X \leq x)}{n}$$

Lze očekávat, že s rostoucím rozsahem výběrového souboru se budou hodnoty empirické distribuční funkce F(x) ustalovat kolem hodnot distribuční funkce Φ(x). Vlastnosti empirické distribuční funkce se přenášejí i na distribuční funkci.

Distribuční funkce náhodné veličiny

Definice:

a) Nechť $(\Omega, \mathcal{A}, P)$ je pravděpodobnostní prostor, $X : \Omega \mapsto R$ je skalární náhodná veličina. Funkce $\Phi : R \mapsto R$ daná vzorcem:

$\forall x \in R : \Phi(x) = P(X \leq x)$

se nazývá **distribuční funkce** náhodné veličiny X.

b) Nechť $(\Omega, \mathcal{A}, P)$ je pravděpodobnostní prostor, $\mathbf{X} = (X_1, \ldots, X_n) : \Omega \mapsto R^n$ je náhodný vektor. Funkce $\Phi : R^n \mapsto R$ daná vzorcem:

$\forall (x_1, \ldots, x_n) \in R^n : \Phi(x_1, \ldots, x_n) = P(X_1 \leq x_1 \wedge \ldots \wedge X_n \leq x_n)$

se nazývá **distribuční funkce** náhodného vektoru **X**.

Příklad

Příklad: Najděte distribuční funkci náhodné veličiny X, která udává, jaké číslo padlo při hodu kostkou a nakreslete graf této distribuční funkce.

Řešení:

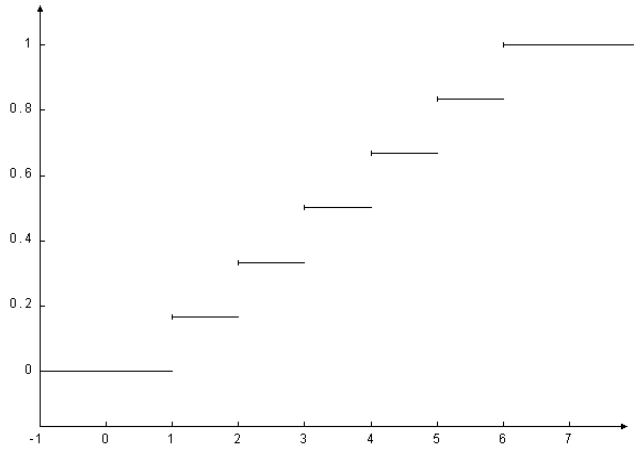
Náhodná veličina X může nabývat hodnot 1, 2, 3, 4, 5, 6. Číselnou osu tedy rozdělíme na 7 intervalů.

$x \in (-\infty, 1)$: $\Phi(x) = P(X \leq x) = 0$, $x \in \langle 1,2)$: $\Phi(x) = P(X \leq x) = \frac{1}{6}$

$x \in \langle 2,3)$: $\Phi(x) = P(X \leq x) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6}$, $x \in \langle 3,4)$: $\Phi(x) = P(X \leq x) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6}$

$x \in \langle 4,5)$: $\Phi(x) = P(X \leq x) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{4}{6}$, $x \in \langle 5,6)$: $\Phi(x) = P(X \leq x) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{5}{6}$

$x \in \langle 6,\infty)$: $\Phi(x) = P(X \leq x) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{6}{6} = 1$



Vlastnosti distribuční funkce NV

Věta ☼: :

Nechť $\Phi(x)$ je distribuční funkce skalární náhodné veličiny X. Pak $\Phi(x)$ má následující vlastnosti:

a) $\Phi(x)$ je neklesající, tj. $\forall x_1 < x_2 : \Phi(x_1) \leq \Phi(x_2)$.

b) $\Phi(x)$ je zprava spojitá, tj. pro libovolné, ale pevně dané $x_0 \in R$ je $\lim_{x \to x_{0+}} \Phi(x) = \Phi(x_0)$.

c) $\Phi(x)$ je normovaná, tj. $\lim_{x \to \infty} \Phi(x) = 1$, $\lim_{x \to -\infty} \Phi(x) = 0$.

d) $\forall a, b \in R, a < b \Rightarrow P(a < X \leq b) = \Phi(b) - \Phi(a)$.

e) Pro libovolné, ale pevně dané $x_0 \in R : P(X = x_0) = \Phi(x_0) - \lim_{x \to x_{0-}} \Phi(x)$.

Důkaz: Jenom náznakem.

ad a) Plyne z monotonie pravděpodobnosti P9.

ad b) Plyne ze spojitosti pravděpodobnosti shora P17.

ad c) $\lim_{x \to -\infty} \Phi(x) = \lim_{x \to -\infty} P(X \leq x) = P(\varnothing) = 0$, $\lim_{x \to \infty} \Phi(x) = \lim_{x \to \infty} P(X \leq x) = P(\Omega) = 1$

ad d) Plyne ze subtraktivity pravděpodobnosti P8.

ad e) Plyne ze spojitosti pravděpodobnosti zdola P16.

P9: $A_1 \subseteq A_2 \Rightarrow P(A_2) \leq P(A_1)$

P17: $A_1 \supseteq A_2 \supseteq \ldots \in A \Rightarrow P\left(\bigcap_{i=1}^{\infty} A_i\right) = \lim_{i \to \infty} P(A_i)$

P8: $A_1 \subseteq A_2 \Rightarrow P(A_2 - A_1) = P(A_2) - P(A_1)$

P16: $A_1 \subseteq A_2 \subseteq \ldots \in A \Rightarrow P\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim_{i \to \infty} P(A_i)$

Příklad

Příklad: Náhodná veličina X udává denní počet obsazených pokojů v určitém penziónu. Známe její distribuční funkci, tj. pravděpodobnost, že bude obsazeno nejvýše x pokojů:

$$\Phi(x)=\begin{cases}0 \ pro \ x<7\\ 0{,}02 \ pro \ 7\le x<8\\ 0{,}05 \ pro \ 8\le x<9\\ 0{,}12 \ pro \ 9\le x<10\\ 1 \ pro \ x\ge 10\end{cases}$$

a) Určete pravděpodobnost, že v náhodně zvolený den bude obsazeno právě 7, 8, 9, 10 pokojů.
b) Jaká je pravděpodobnost, že bude obsazeno nejvýše 10 a nejméně 8 pokojů?

Řešení:

ad a) Využijeme vlastnost (e) z věty ☼.

$$P(X=7)=\Phi(7)-\lim_{x\to 7_-}\Phi(x)=0{,}02-0=0{,}02$$

$$P(X=8)=\Phi(8)-\lim_{x\to 8_-}\Phi(x)=0{,}05-0{,}02=0{,}03$$

$$P(X=9)=\Phi(9)-\lim_{x\to 9_-}\Phi(x)=0{,}12-0{,}05=0{,}07$$

$$P(X=10)=\Phi(10)-\lim_{x\to 10_-}\Phi(x)=1-0{,}12=0{,}88$$

ad b) Využijeme vlastnost (d) z věty ☼.

$$P(8\le X\le 10)=P(7<X\le 10)=\Phi(10)-\Phi(7)=1-0{,}02=0{,}98$$

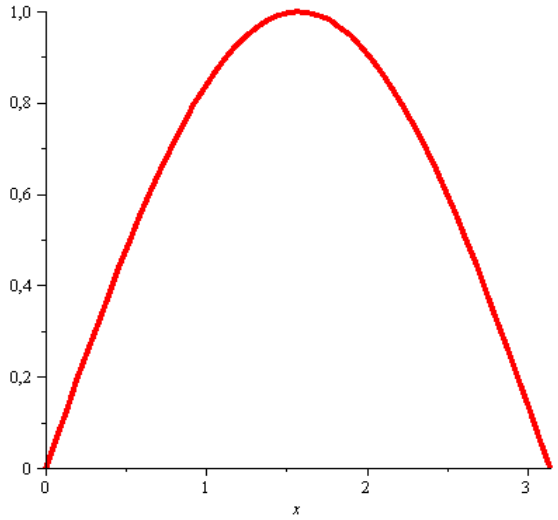
Příklad

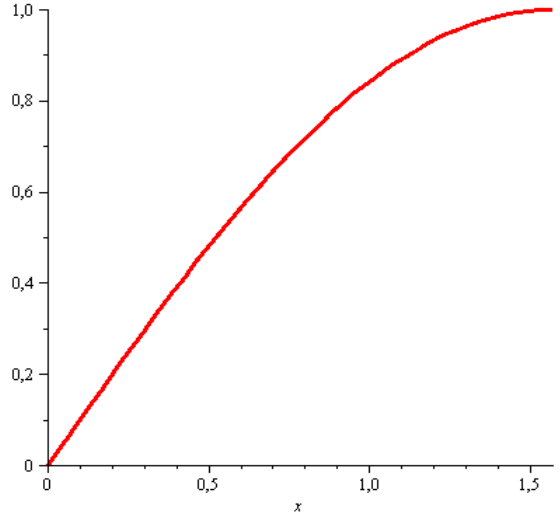
Je funkce $\Phi(x) = \sin x$ distribuční funkcí náhodné veličiny X v intervalu

a) $\langle 0, \pi \rangle$,

b) $\langle 0, \pi/2 \rangle$?

Řešení: a) NE b) ANO





Příklad

Určete

a) konstanty *A*, *B* tak, aby funkce $\Phi(x) = A + Be^{-x}$ byla distribuční funkcí náhodné veličiny pro $x \in (0, \infty)$,

b) pravděpodobnost $P(1 < X \leq 4)$

Řešení: a) $$0 = \lim_{x \to 0} A + Be^{-x} = A + B \lim_{x \to 0} e^{-x} = A + B$$

$$1 = \lim_{x \to \infty} A + Be^{-x} = A + \lim_{x \to \infty} Be^{-x} = A$$

$$\Rightarrow \quad A = 1, \quad B = -1$$

$$\Rightarrow \quad \Phi(x) = 1 - e^{-x}$$

b) $$P(1 < X \leq 4) = \Phi(4) - \Phi(1) = \frac{e^3 - 1}{e^4} = 0{,}3496$$

Vlastnosti distribuční funkce náhodného vektoru

Věta ☼☼: Nechť $\Phi(x_1,\ldots,x_n)$ je distribuční funkce náhodného vektoru $\mathbf{X}$. Pak $\Phi(x_1,\ldots,x_n)$ ma následující vlastnosti:

a) $\Phi(x_1,\ldots,x_n)$ je neklesající vzhledem ke každé jednotlivé proměnné.

b) $\Phi(x_1,\ldots,x_n)$ je zprava spojitá vzhledem ke každé jednotlivé proměnné.

c) $\lim_{x_1\to\infty} \Phi(x_1,\ldots,x_n) = 1$

$\vdots$

$x_n\to\infty$

$\forall i \in \{1,\ldots,n\} : \lim_{x_i\to-\infty} \Phi(x_1,\ldots,x_n) = 0$

d) $\forall (x_1,\ldots,x_n) \in R^n, \forall (h_1,\ldots,h_n) \in R^n_+ :$

$P(x_1 < X_1 \le x_1+h_1 \wedge \ldots \wedge x_n < X_n \le x_n+h_n) = \Phi(x_1+h_1,\ldots,x_n+h_n) -$

$\sum_{i=1}^{n} \Phi(x_1+h_1,\ldots,x_i,\ldots,x_n+h_n) + \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \Phi(x_1+h_1,\ldots,x_i,\ldots,x_j,\ldots,x_n+h_n) -$

$\ldots + (-1)^n \Phi(x_1,\ldots,x_n)$

e) $\forall i \in \{1,\ldots,n\} : \lim_{x_1\to\infty} \Phi(x_1,\ldots,x_n) = \Phi_i(x_i).$

$\vdots$

$x_{i-1}\to\infty$

$x_{i+1}\to\infty$

$\vdots$

$x_n\to\infty$

Vlastnosti distribuční funkce náhodného vektoru

Důkaz: Jenom náznakem.
ad a), ad b) Podobně jako ve skalárním případě.

ad c)

$$\lim_{\substack{x_1\to\infty\\ \vdots \\ x_n\to\infty}} \Phi(x_1,\ldots,x_n) = \lim_{\substack{x_1\to\infty\\ \vdots \\ x_n\to\infty}} P(X_1 \le x_1 \wedge \ldots \wedge X_n \le x_n) = P(X_1 \in R \wedge \ldots \wedge X_n \in R) = P(\Omega) = 1$$

$$\forall i \in \{1,\ldots,n\}: \lim_{x_i\to-\infty} \Phi(x_1,\ldots,x_n) = \lim_{x_i\to-\infty} P(X_1 \le x_1 \wedge \ldots \wedge X_i \le x_i \wedge \ldots \wedge X_n \le x_n) = P(X_1 \le x_1 \wedge \ldots \wedge \varnothing_i \wedge \ldots \wedge X_n \le x_n) = P(\varnothing) = 0$$

ad d) Vlastnost vyjadřuje princip inkluze a exkluze.

ad e)

$$\lim_{\substack{x_1\to\infty\\ \vdots \\ x_{l-1}\to\infty\\ x_{l+1}\to\infty\\ \vdots \\ x_n\to\infty}} \Phi(x_1,\ldots,x_n) = \lim_{\substack{x_1\to\infty\\ \vdots \\ x_{l-1}\to\infty\\ x_{l+1}\to\infty\\ \vdots \\ x_n\to\infty}} P(X_1 \le x_1 \wedge \ldots \wedge X_n \le x_n) = P(X_1 \in R \wedge \ldots \wedge X_i \le x_i \wedge \ldots \wedge X_n \in R) = P(\Omega \wedge \ldots \wedge X_i \le x_i \wedge \ldots \wedge \Omega) =$$

$$= P(X_i \le x_i) = \Phi_i(x_i)$$

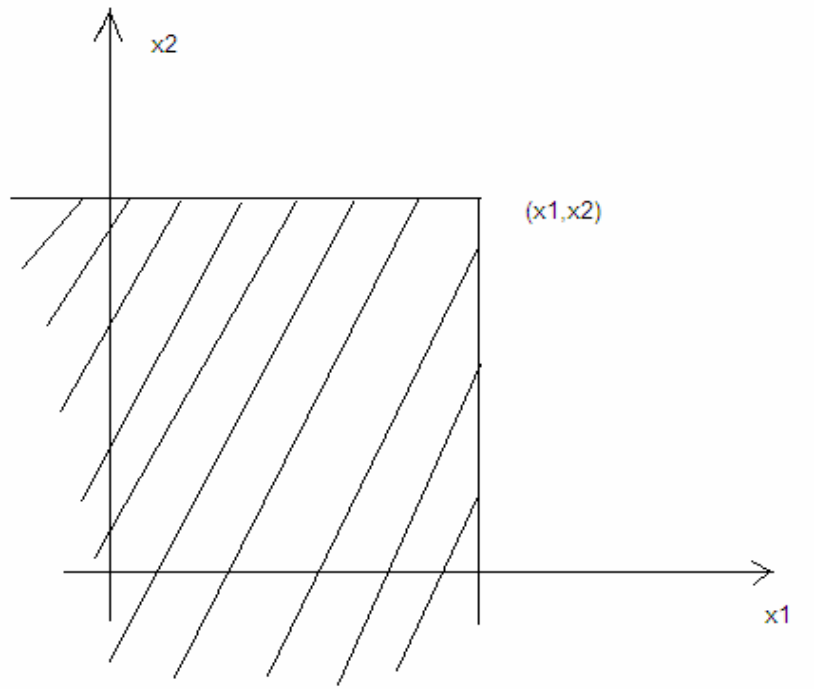
Marginální/simultánní distribuční funkce

Funkce $\Phi_i(x_i)$ je distribuční funkce náhodné veličiny X. Nazývá se **marginální distribuční funkce** a $\Phi(x_1, \ldots, x_n)$ se v této souvislosti nazývá **simultánní distribuční funkce**. Analogicky lze zavést marginální distribuční funkce k proměnných, $k \in \{2, 3, \ldots, n-1\}$.

Vlastnosti distribuční funkce náhodného vektoru

Poznámka: (Ilustrace vlastnosti (d) z věty ☼☼ pro n = 2)

Pro libovolné $(x_1, x_2) \in R^2$ udává $\Phi(x_1, x_2)$ pravděpodobnost, že náhodný vektor (X_1, X_2) se bude realizovat v oblasti $(-\infty, x_1\rangle \times (-\infty, x_2\rangle$:

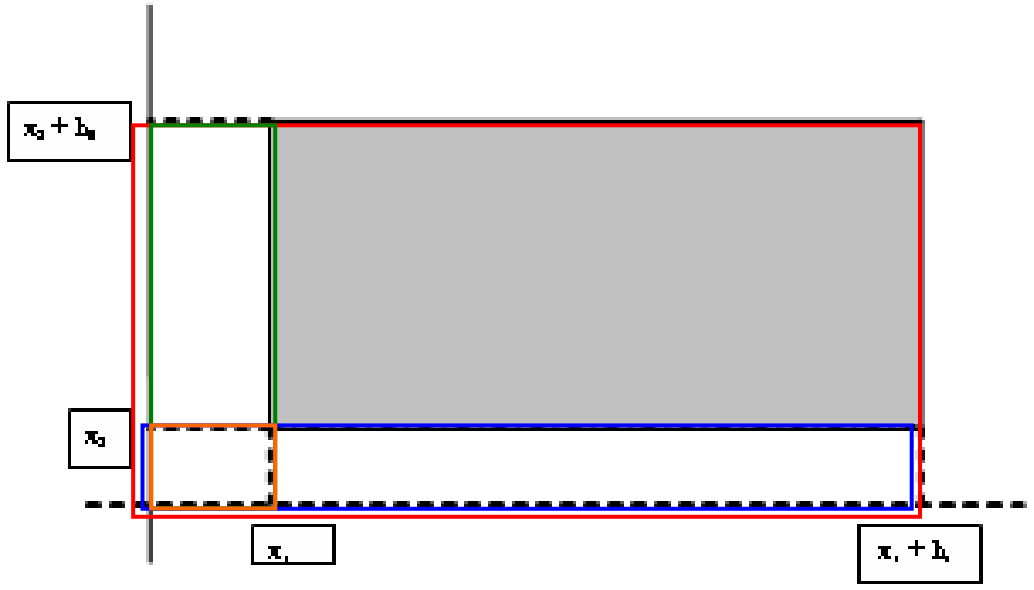


Vlastnosti distribuční funkce náhodného vektoru

Pro libovolné $h_1 > 0, \quad h_2 > 0$ nás zajímá pravděpodobnost , že náhodný vektor (X_1, X_2) se bude realizovat v obdélníku $(x_1, x_1 + h_1\rangle \times (x_2, x_2 + h_2\rangle$:

$$P\left(x_1 < X_1 \le x_1 + h_1 \wedge x_2 < X_2 \le x_2 + h_2\right) =$$

$$= \Phi\left(x_1 + h_1, x_2 + h_2\right) - \Phi\left(x_1 + h_1, x_2\right) - \Phi\left(x_1, x_2 + h_2\right) + \Phi\left(x_1, x_2\right)$$



Příklad

Příklad: Náhodný vektor (X_1, X_2) má distribuční funkci $\Phi(x_1, x_2) = \frac{1}{\pi^2}(\text{arctg } x_1 + \frac{\pi}{2})(\text{arctg } x_2 + \frac{\pi}{2})$.

Vypočtěte pravděpodobnost, že náhodný vektor (X_1, X_2) se bude realizovat v jednotkovém čtverci $(0,1\rangle \times (0,1\rangle$.

Najděte obě marginální distribuční funkce $\Phi_1(x_1)$, $\Phi_2(x_2)$.

Řešení:

$$P(0 < X_1 \leq 1 \wedge 0 < X_2 \leq 1) = \Phi(1,1) - \Phi(1,0) - \Phi(0,1) + \Phi(0,0) =$$

$$= \frac{1}{\pi^2}(\frac{\pi}{4}+\frac{\pi}{2})(\frac{\pi}{4}+\frac{\pi}{2}) - \frac{1}{\pi^2}(\frac{\pi}{4}+\frac{\pi}{2})(0 + \frac{\pi}{2}) - \frac{1}{\pi^2}(0 + \frac{\pi}{2})(\frac{\pi}{4}+\frac{\pi}{2}) + \frac{1}{\pi^2}(0 + \frac{\pi}{2})(0 + \frac{\pi}{2}) = \frac{1}{16}.$$

$$\Phi_1(x_1) = \lim_{x_2 \to \infty} \frac{1}{\pi^2}(\text{arctg } x_1 + \frac{\pi}{2})(\text{arctg } x_2 + \frac{\pi}{2}) = \frac{1}{\pi}(\text{arctg } x_1 + \frac{\pi}{2})$$

$$\Phi_2(x_2) = \lim_{x_1 \to \infty} \frac{1}{\pi^2}(\text{arctg } x_1 + \frac{\pi}{2})(\text{arctg } x_2 + \frac{\pi}{2}) = \frac{1}{\pi}(\text{arctg } x_2 + \frac{\pi}{2})$$

Existence distribuční funkce

Věta: (existenční věta)

a) Skalární případ: Jestliže funkce $\Phi(x)$ má vlastnosti (a), (b), (c) z věty o vlastnostech distribuční funkce skalární náhodné veličiny, pak existuje pravděpodobnostní prostor $(\Omega, \mathcal{A}, P)$ a na něm definovaná skalární náhodná veličina X tak, že $\Phi(x)$ je její distribuční funkce.

b) Vektorový případ: Jestliže funkce $\Phi(x_1, \ldots, x_n)$ má vlastnosti (a), (b), (c) z věty o vlastnostech distribuční funkce náhodného vektoru, pak existuje pravděpodobnostní prostor $(\Omega, \mathcal{A}, P)$ a na něm definovaný náhodný vektor $\mathbf{X} = (X_1, \ldots, X_n)$ tak, že $\Phi(x_1, \ldots, x_n)$ je jeho distribuční funkce.

8. Diskrétní a spojité NV, vybraná rozložení NV

Motivace: Distribuční funkce popisuje pravděpodobnostní chování jakékoliv náhodné veličiny. V praxi však mají význam dva speciální typy náhodných veličin, a to diskrétní a spojité náhodné veličiny.

Diskrétní náhodná veličina nabývá nejvýše spočetně mnoha izolovaných hodnot. Je to např. počet zásahů do terče při střelbě, počet chyb, jichž se dopustí nějaké zařízení za určitou dobu, počet zákazníků ve frontě apod.

Pravděpodobnostní chování diskrétní náhodné veličiny popisujeme pravděpodobnostní funkcí:

$$\forall x \in R : \pi(x) = P(X = x).$$

Je to zidealizovaný protějšek četnostní funkce zavedené v popisné statistice v souvislosti s bodovým rozložením četností:

$$\forall x \in R : p(x) = \frac{N(X = x)}{n}.$$

S rostoucím rozsahem výběrového souboru se budou hodnoty četnostní funkce ustalovat kolem hodnot pravděpodobnostní funkce. Vlastnosti četnostní funkce se přenášejí i na pravděpodobnostní funkci, tedy pravděpodobnostní funkce

je nezáporná $\forall x \in R : \pi(x) \geq 0$,

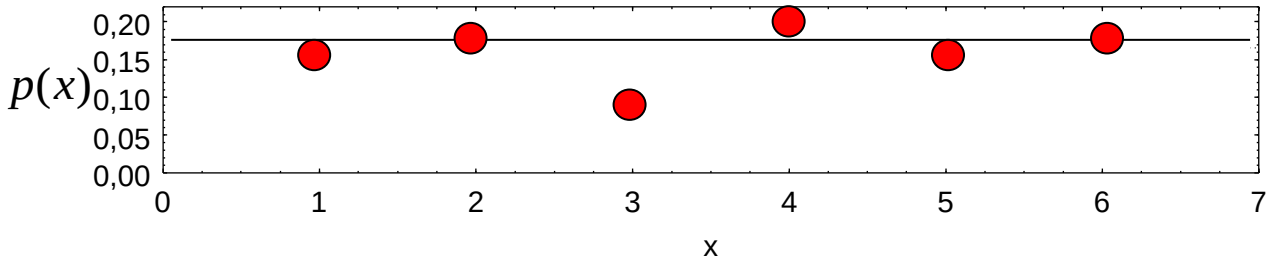
je normovaná $\sum_{x=-\infty}^{\infty} \pi(x) = 1$,

s distribuční funkcí je spjata součtovým vztahem $\forall x \in R : \Phi(x) = \sum_{t \leq x} \pi(t)$

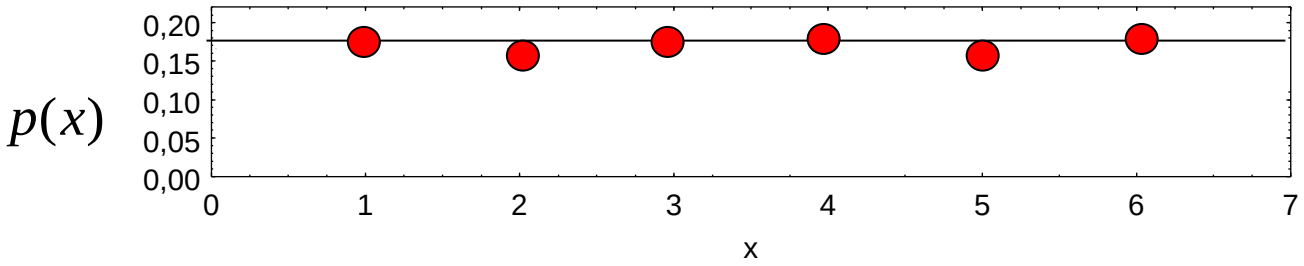
Ilustrace vztahu mezi četnostní funkcí a pravděpodobnostní funkcí

Provedeme n hodů kostkou. Zajímáme se o četnostní funkci počtu ok.

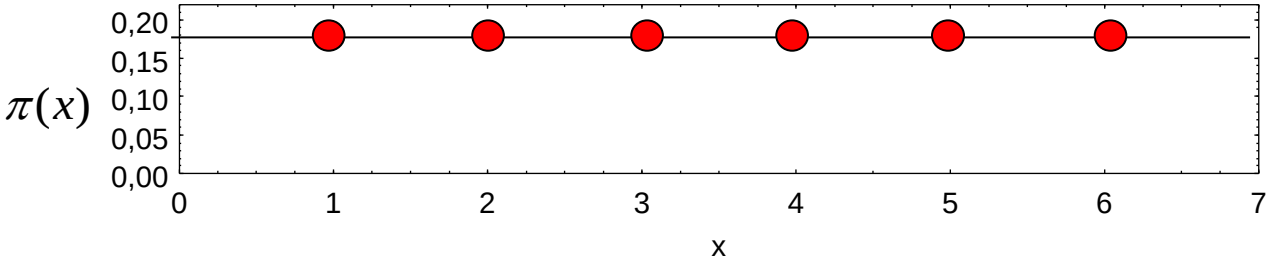
n = 60:



n = 600:



$n \to \infty$:



Spojitá náhodná veličina - motivace

Spojitá náhodná veličina nabývá všech hodnot z nějakého intervalu. Je to např. :

- výsledek nějakého fyzikálního či chemického měření,
- hektarový výnos pšenice,
- hmotnost sériově vyráběného výrobku apod.

Pravděpodobnostní chování spojité náhodné veličiny popisujeme hustotou pravděpodobnosti $\varphi(x)$, což je zidealizovaný protějšek hustoty četnosti $f(x)$ zavedené v popisné statistice v souvislosti s intervalovým rozložením četností. S rostoucím rozsahem výběrového souboru a klesajícími šířkami třídicích intervalů se budou hodnoty hustoty četnosti ustalovat kolem hodnot hustoty pravděpodobnosti.

Vlastnosti hustoty četnosti se přenášejí i na hustotu pravděpodobnosti, tedy hustota pravděpodobnosti

je nezáporná $\forall x \in R : \varphi(x) \geq 0$,

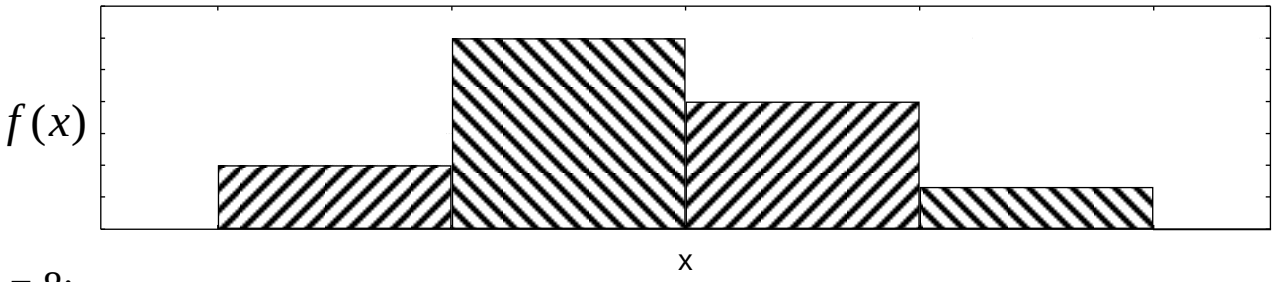
je normovaná $\int_{-\infty}^{\infty} \varphi(x)dx = 1$,

s distribuční funkcí je spjata integrálním vztahem $\forall x \in R : \Phi(x) = \int_{-\infty}^{x} \varphi(t)dt$

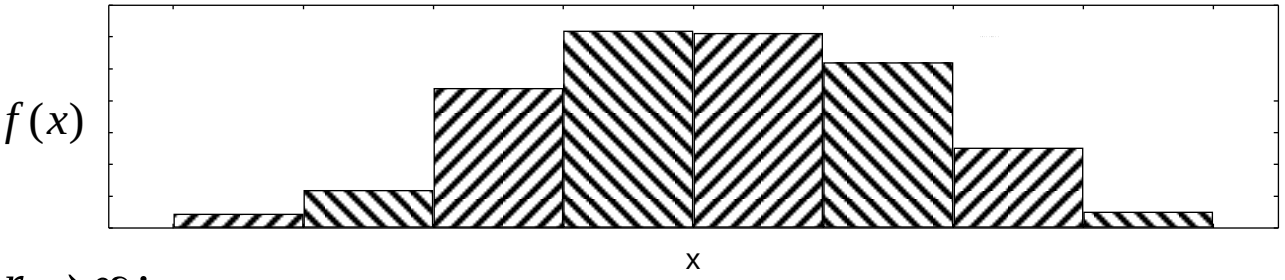
Ilustrace vztahu mezi hustotou četnosti a hustotou pravděpodobnosti

Náhodně vybereme n sériově vyráběných součástek, změříme jejich délku a budeme se zajímat o hustotu četnosti odchylek těchto měření od deklarované délky součástky.

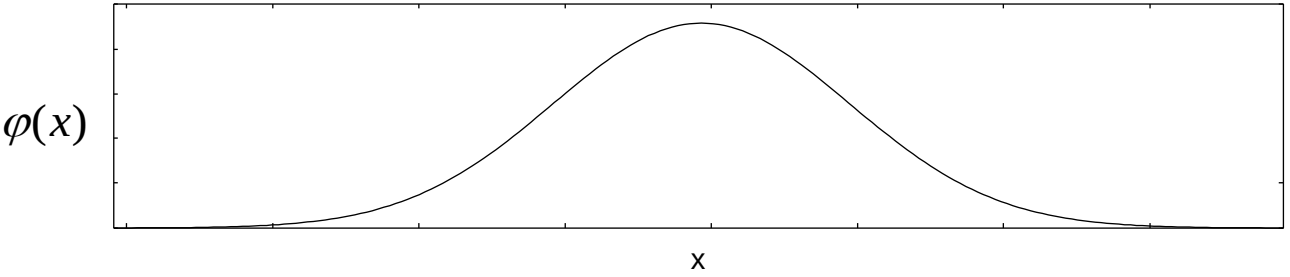
n = 40, r = 4:



n= 400, r = 8:



$n \to \infty, r \to \infty$:



Diskrétní náhodná veličina

Definice:

Nechť $(\Omega, \mathcal{A}, P)$ je pravděpodobnostní prostor, X náhodná veličina definovaná na měřitelném prostoru $(\Omega, \mathcal{A})$, která má distribuční funkci $\Phi(x)$. Řekneme, že náhodná veličina X je **diskrétní** (vzhledem k P), právě když existuje reálná funkce $\pi(x)$, která je nulová v R s výjimkou nejméně jednoho a nejvýše spočetně mnoha bodů, kde je kladná a platí pro ni: $\forall x \in R : \Phi(x) = \sum_{t \leq x} \pi(t)$. Tato funkce se nazývá **pravděpodobnostní funkce** diskrétní náhodné veličiny X.

Vlastnosti pravděpodobnostní funkce

Věta:
Nechť $\pi(x)$ je pravděpodobnostní funkce diskrétní náhodné veličiny X. Pak platí:

a) $\forall x \in R : \pi(x) \geq 0$ (vlastnost D1 - nezápornost)
b) $\sum_{x=-\infty}^{\infty} \pi(x) = 1$ (vlastnost D2 - normovanost)
c) $\forall x \in R : \pi(x) = P(X = x)$
d) $\forall B \in \mathcal{B} : P(X \in B) = \sum_{x \in B} \pi(x)$.

Důkaz:

ad a) Vlastnost D1 je součást definice.

ad b) $\sum_{x=-\infty}^{\infty} \pi(x) = \lim_{t\to\infty} \sum_{x=-\infty}^{t} \pi(x) = \lim_{t\to\infty} \Phi(t) = 1$

ad c) $P(X = x_0) = \Phi(x_0) - \lim_{x\to x_{0-}} \Phi(x) = \sum_{t\leq x_0} \pi(t) - \lim_{x\to x_{0-}} \sum_{t\leq x} \pi(t) = \lim_{x\to x_{0-}} \sum_{x<t\leq x} \pi(t) = \pi(x_0)$

ad d) Označme $G \subseteq R$ tu nejvýše spočetnou množinu, na níž $\pi(x)$ nabývá kladných hodnot. Pak pro libovolnou borelovskou množinu B platí:

$$P(X \in B) = P(X \in B \cap G) + P(X \in B \cap \overline{G}) = P\left(X \in \bigcup_{x\in B\cap G} \{x\}\right) + 0 = \sum_{x\in B\cap G} P(X = x) = \sum_{x\in B} \pi(x)$$

Příklad

Příklad: Náhodná veličina X udává počet líců při hodu dvěma mincemi. Určete její pravděpodobnostní a distribuční funkci a nakreslete jejich grafy.

Řešení:

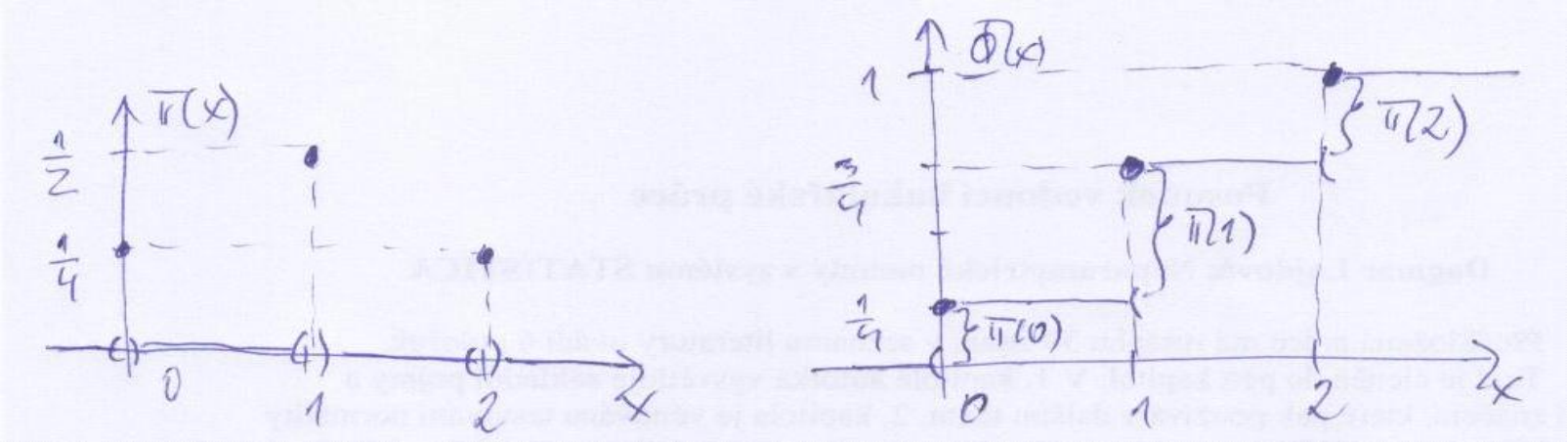
Základní prostor: $\Omega = \{[L,L],[L,R],[R,L],[R,R]\}$, jevové pole: maximální, pravděpodobnost: klasická , náhodná veličina X nabývá hodnot z množiny $\{0, 1, 2\}$.

$$\pi(0) = P(X=0) = \frac{1}{4} \qquad x \in (-\infty,0): \Phi(x) = 0$$

$$\pi(1) = P(X=1) = \frac{2}{4} \qquad x \in \langle 0,1): \Phi(x) = \pi(0) = \frac{1}{4}$$

$$\pi(2) = P(X=2) = \frac{1}{4} \qquad x \in \langle 1,2): \Phi(x) = \pi(0) + \pi(1) = \frac{3}{4}$$

$$\pi(x) = 0 \text{ jinak} \qquad x \in \langle 2,\infty): \Phi(x) = \pi(0) + \pi(1) + \pi(2) = 1$$



Příklad

Dva střelci (s pravděpodobnostmi zásahu p_1 a p_2) se střídají ve střelbě, dokud někdo nezasáhne. Určete pravděpodobnostní funkci počtu výstřelů.

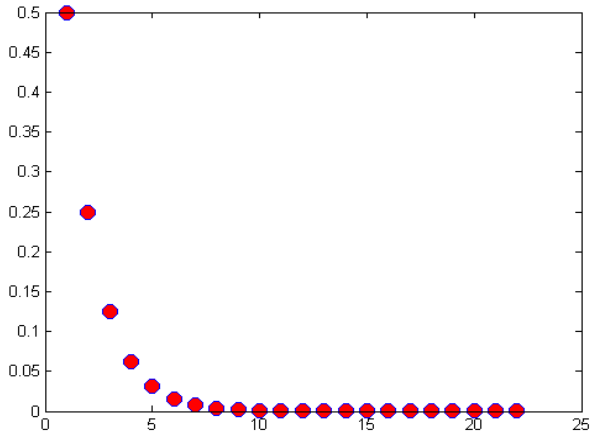
Řešení:

$$\pi(2n+1) = (1-p_1)^n (1-p_2)^n p_1$$

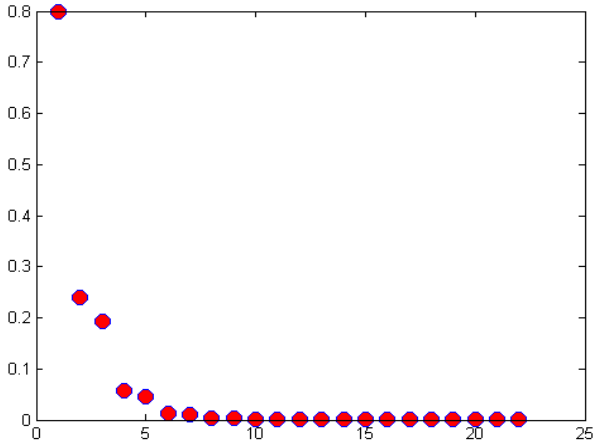
$$\pi(2n+2) = (1-p_1)^{n+1} (1-p_2)^n p_2 \quad n = 0,1,\ldots$$

$$\pi(x) = 0 \text{ jinak}$$

Pro $p_1 = p_2 = 0{,}5$:



Pro $p_1 = 0{,}8$ a $p_2 = 0{,}3$:



Příklad

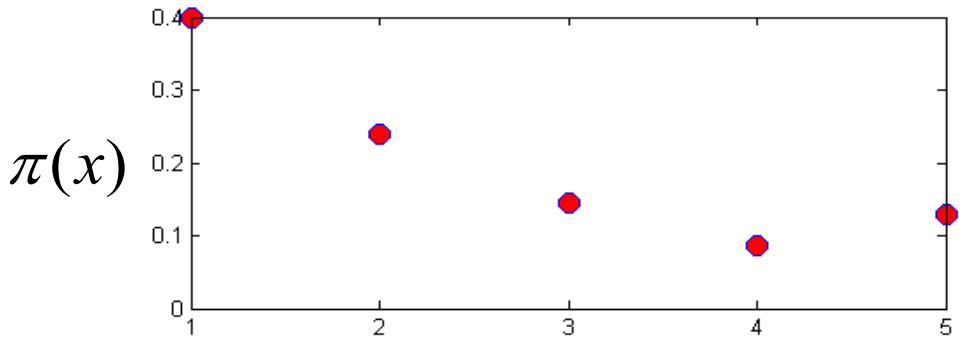
Lovec má 5 patron a pravděpodobnost zásahu 0,4. Střílí, dokud netrefí (a dokud má čím). Určete pravděpodobnostní funkci.

Řešení:

$$\pi(k) = 0{,}6^{k-1} \cdot 0{,}4 \quad k = 1,\ldots,4$$

$$\pi(5) = 0{,}6^4$$

$$\pi(x) = 0 \text{ jinak}$$



Diskrétní náhodný vektor

Poznámka: Distribuční funkce diskrétní náhodné veličiny má schodovitý průběh. Pravděpodobnostní funkce je distribuční funkcí určena jednoznačně.

Definice:

Nechť $(\Omega, \mathcal{A}, P)$ je pravděpodobnostní prostor, $\mathbf{X} = (X_1, \ldots, X_n)$ náhodný vektor definovaný na měřitelném prostoru $(\Omega, \mathcal{A})$. Nechť $\Phi(x_1, \ldots, x_n)$ je jeho distribuční funkce. Řekneme, že náhodný vektor $\mathbf{X}$ je **diskrétní** (vzhledem k P), právě když existuje reálná funkce $\pi(x_1, \ldots, x_n)$, která je nulová v R^n s výjimkou nejméně jednoho a nejvýše spočetně mnoha bodů, kde je kladná a platí pro ni: $\forall (x_1, \ldots, x_n) \in R^n : \Phi(x_1, \ldots, x_n) = \sum_{t_1 \leq x_1} \ldots \sum_{t_n \leq x_n} \pi(t_1, \ldots, t_n)$. Tato funkce se nazývá **pravděpodobnostní funkce** diskrétního náhodného vektoru $\mathbf{X}$.

Vlastnosti pravděpodobnostní funkce

Věta:

Nechť $\pi(x_1, \ldots, x_n)$ je pravděpodobnostní funkce diskrétního náhodného vektoru $\mathbf{X}$. Pak platí:

a) $\forall (x_1, \ldots, x_n) \in R^n : \pi(x_1, \ldots, x_n) \geq 0$ (vlastnost D1 - nezápornost)

b) $\sum_{x_1=-\infty}^{\infty} \ldots \sum_{x_n=-\infty}^{\infty} \pi(x_1, \ldots, x_n) = 1$ (vlastnost D2 - normovanost)

c) $\forall (x_1, \ldots, x_n) \in R^n : \pi(x_1, \ldots, x_n) = P(X_1 = x_1 \wedge \ldots \wedge X_n = x_n)$

d) $\forall B \in \mathcal{B}^n : P(\mathbf{X} \in B) = \sum \ldots \sum_{(x_1,\ldots,x_n)\in B} \pi(x_1, \ldots, x_n)$

e) $\forall i \in \{1, \ldots, n\} : \sum_{x_1=-\infty}^{\infty} \ldots \sum_{x_{i-1}=-\infty}^{\infty} \sum_{x_{i+1}=-\infty}^{\infty} \ldots \sum_{x_n=-\infty}^{\infty} \pi(x_1, \ldots, x_n) = \pi_i(x_i)$.

Funkce $\pi_i(x_i)$ je pravděpodobnostní funkce náhodné veličiny X_i. Nazývá se **marginální pravděpodobnostní funkce**. Funkce $\pi(x_1, \ldots, x_n)$ se v této souvislosti nazývá **simultánní pravděpodobnostní funkce**. Podobně lze zavést marginální pravděpodobnostní funkce k proměnných, kde $k \in \{2, 3, \ldots, n-1\}$.

Příklad

Příklad: Je dán systém složený ze dvou bloků. Pravděpodobnost, že i-tý blok správně funguje, je υ_i, i = 1, 2 a pravděpodobnost, že správně fungují oba bloky, je υ_{12}. Nechť náhodná veličina X_i je ukazatel fungování i-tého bloku, tj. $X_i = \begin{cases} 1, \text{pokud i - tý blok funguje} \\ 0, \text{pokud i - tý blok nefunguje} \end{cases}$, i = 1, 2. Najděte simultánní pravděpodobnostní funkci $\pi(x_1, x_2)$ náhodného vektoru (X_1, X_2) a obě marginální pravděpodobnostní funkce $\pi_1(x_1)$ a $\pi_2(x_2)$.

Řešení:
Hodnoty pravděpodobnostních funkcí zapíšeme do kontingenční tabulky.

x_1	x_2		$\pi_1(x_1)$
	0	1	
0	$1-\vartheta_1-\vartheta_2+\vartheta_{12}$	$\vartheta_2-\vartheta_{12}$	$1-\vartheta_1$
1	$\vartheta_1-\vartheta_{12}$	ϑ_{12}	ϑ_1
$\pi_2(x_2)$	$1-\vartheta_2$	ϑ_2	1

$\pi(0,0) = P(X_1=0 \wedge X_2=0) = 1 - P(X_1=1 \vee X_2=1) = 1 - (\upsilon_1 + \upsilon_2 - \upsilon_{12}) = 1 - \upsilon_1 - \upsilon_2 + \upsilon_{12}$
$\pi(0,1) = P(X_1=0 \wedge X_2=1) = P(X_2=1) - P(X_1=1 \wedge X_2=1) = \upsilon_2 - \upsilon_{12}$
$\pi(1,0) = P(X_1=1 \wedge X_2=0) = P(X_1=1) - P(X_1=1 \wedge X_2=1) = \upsilon_1 - \upsilon_{12}$
$\pi(1,1) = P(X_1=1 \wedge X_2=1) = \upsilon_{12}$
$\pi(x_1,x_2) = 0$ jinak

Existenční věta

Věta (existenční věta)

a) Skalární případ: Jestliže funkce $\pi(x)$ má vlastnosti D1, D2 z věty o vlastnostech pravděpodobnostní funkce skalární náhodné veličiny, pak existuje pravděpodobnostní prostor $(\Omega, \mathcal{A}, P)$ a na něm definovaná skalární diskrétní náhodná veličina X tak, že $\pi(x)$ je její pravděpodobnostní funkce.

b) Vektorový případ: Jestliže funkce $\pi(x_1, \ldots, x_n)$ má vlastnosti D1, D2 z věty o vlastnostech pravděpodobnostní funkce náhodného vektoru, pak existuje pravděpodobnostní prostor $(\Omega, \mathcal{A}, P)$ a na něm definovaný diskrétní náhodný vektor $\mathbf{X} = (X_1, \ldots, X_n)$ tak, že $\pi(x_1, \ldots, x_n)$ je jeho pravděpodobnostní funkce.

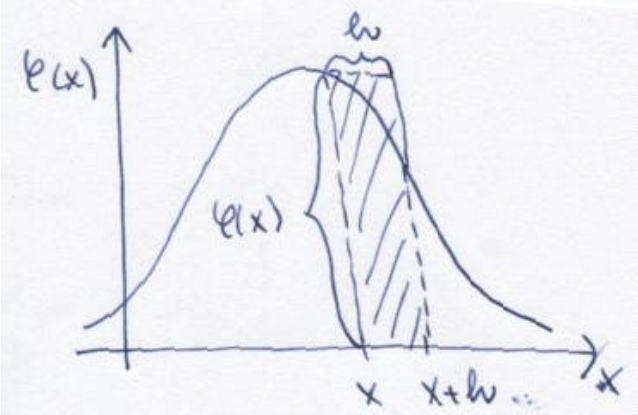
Spojitá náhodná veličina

Definice:

Nechť $(\Omega, \mathcal{A}, P)$ je pravděpodobnostní prostor, X náhodná veličina definovaná na měřitelném prostoru $(\Omega, \mathcal{A})$, která má distribuční funkci $\Phi(x)$. Řekneme, že náhodná veličina X je **spojitá** (vzhledem k P), právě když existuje po částech spojitá nezáporná reálná funkce $\varphi(x)$ tak, že pro $\forall x \in R : \Phi(x) = \int_{-\infty}^{x} \varphi(t)dt$. Tato funkce se nazývá **hustota pravděpodobnosti** spojité náhodné veličiny X.

Spojitá náhodná veličina - poznámka

Poznámka: Na rozdíl od pravděpodobnostní funkce diskrétní náhodné veličiny nemá hustota pravděpodobnosti spojité náhodné veličiny význam pravděpodobnosti. Její význam lze odvodit z integrálního vztahu mezi distribuční funkcí a hustotou pravděpodobnosti.



Pravděpodobnost, že náhodná veličina se bude realizovat v intervalu $\langle x, x+h \rangle$, je:

$$P(x < X \le x+h) = \Phi(x+h) - \Phi(x) = \int_{-\infty}^{x+h} \varphi(t)\,dt - \int_{-\infty}^{x} \varphi(t)\,dt = \int_{x}^{x+h} \varphi(t)\,dt$$

Bude -li h dostatečně malé číslo, lze plochu pod grafem hustoty nahradit obsahem obdélníka o stranách φ(x) a h, tj.

$$P(x < X \le x+h) \approx \varphi(x) \cdot h$$

Vlastnosti hustoty pravděpodobnosti

Věta:

Nechť $\varphi(x)$ je hustota spojité náhodné veličiny X. Pak platí:

a) $\forall x \in R : \varphi(x) \geq 0$ (vlastnost S1 - nezápornost)

b) $\int_{-\infty}^{\infty} \varphi(x)dx = 1$ (vlastnost S2 - normovanost)

c) $\forall x \in R, \forall h > 0 : P(x < X \leq x + h) = \int_{x}^{x+h} \varphi(t)dt$

d) Pro libovolné, ale pevně dané $x \in R : P(X = x) = 0$.

e) $\varphi(x) = \frac{d\Phi(x)}{dx}$ ve všech bodech spojitosti funkce $\varphi(x)$.

Důkaz:

ad a) Vlastnost S1 je součást definice.

ad b) $\int_{-\infty}^{\infty} \varphi(x)dx = \lim_{x\to\infty} \int_{-\infty}^{x} \varphi(t)dt = \lim_{x\to\infty} \Phi(x) = 1$

ad c) $\int_{x}^{x+h} \varphi(t)dt = \int_{-\infty}^{x+h} \varphi(t)dt - \int_{-\infty}^{x} \varphi(t)dt = \Phi(x+h) - \Phi(x) = P(x < X \leq x+h)$

ad d) $P(X = x) = \int_{x}^{x} \varphi(t)dt = 0$

ad e) $\frac{d\Phi(x)}{dx} = \frac{d}{dx} \int_{-\infty}^{x} \varphi(t)dt = \varphi(x)$ ve všech bodech spojitosti funkce $\varphi(x)$.

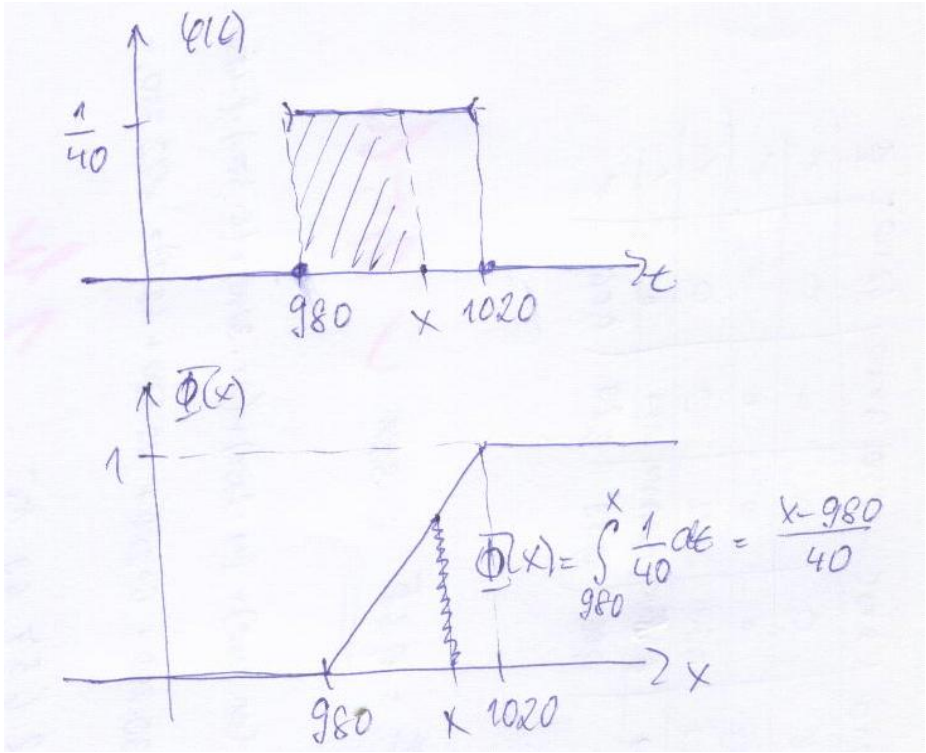
Příklad

Příklad: Na automatické lince se plní láhve mlékem. Každá láhev má obsahovat přesně 1000 ml mléka, ale v důsledku působení náhodných vlivů množství mléka kolísá v intervalu (980 ml, 1020 ml). Každé množství mléka v tomto intervalu považujeme za stejně možné. Náhodná veličina X udává množství mléka v náhodně vybrané lahvi. Najděte její hustotu pravděpodobnosti $\varphi(x)$ a distribuční funkci $\Phi(x)$. Jaká je pravděpodobnost, že v náhodně vybrané láhvi bude aspoň 1000 ml mléka?

Řešení: $\varphi(x) = \begin{cases} k \text{ pro } x \in (980, 1020) \\ 0 \text{ jinak} \end{cases}$

Z normovanosti hustoty plyne: $1 = \int_{980}^{1020} k\,dx = 40\,k$, tedy $k = \frac{1}{40}$.

Pro distribuční funkci platí: $\Phi(x) = \begin{cases} 0 \text{ pro } x \le 980 \\ \int_{980}^{x} \frac{1}{40}\,dt = \frac{x-980}{40} \text{ pro } 980 < x < 1020 \\ 1 \text{ pro } x \ge 1020 \end{cases}$



$$P(X \ge 1000) = \int_{1000}^{1020} \frac{1}{40}\,dx = \frac{1}{40}[x]_{1000}^{1020} = \frac{20}{40} = 0{,}5$$

Příklad

Napište distribuční funkci rozdělení daného hustotou f(x) = x/2 na (0, 1), 1/2 na (1, 2), (3 − x)/2 na (2, 3).

Řešení:

$$\text{Na}\,(0,1):$$

$$F(x)=\int_0^x f(t)\,dt=\int_0^x \frac{t}{2}\,dt=\frac{1}{2}\left[\frac{t^2}{2}\right]_0^x=\frac{x^2}{4},$$

$$\text{Na}\,(1,2):$$

$$F(x)=\frac{1}{4}+\frac{1}{2}(x-1),$$

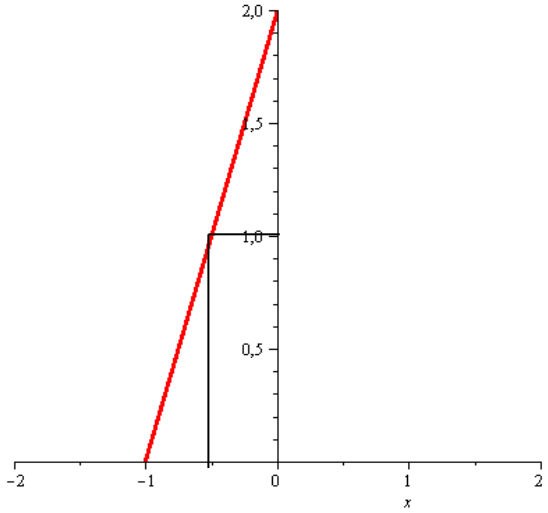
$$\text{Na}\,(2,3):$$

$$F(x)=\frac{1}{4}+\frac{1}{2}+\int_2^x \frac{3-t}{2}\,dt=\frac{3}{4}-\frac{1}{2}\left[\frac{(3-t)^2}{2}\right]_2^x=1-\frac{(3-x)^2}{4}.$$

Příklad

Rozdělení náhodné veličiny X je dáno hustotou f(x) = 2x+2, na (−1, 0) a nulovou jinde. Najděte $P(-2 \le X \le -0{,}5)$.

Řešení:



$$P(-2 \le X \le -0{,}5) = P(-1 \le X \le -0{,}5) =$$

$$= \int_{-1}^{-0{,}5} (2x+2)\,dx = \frac{1}{2} \cdot \frac{1}{2} \cdot 1 = \frac{1}{4}$$

Příklad

Náhodná veličina X má hustotu $f(x)=\frac{a}{1+x^2}$ na R.

Určete *a*, distribuční funkci, $P\left(X>\sqrt{3}\right)$

Řešení:

$$1=\int_{-\infty}^{\infty} f(x)\,dx = a\left(\lim_{x\to\infty} arctg(x) - \lim_{x\to-\infty} arctg(x)\right) =$$

$$= a\left(\frac{\pi}{2}+\frac{\pi}{2}\right) \qquad \Rightarrow a=\frac{1}{\pi}$$

$$F(x)=\int_{-\infty}^{x} f(x)\,dx = \frac{1}{\pi}\left(arctg(x)+\frac{\pi}{2}\right)$$

$$P\left(X>\sqrt{3}\right)=1-F\left(\sqrt{3}\right)=1-\frac{1}{\pi}\left(\frac{\pi}{3}+\frac{\pi}{2}\right)=\frac{1}{6}$$

Spojitý náhodný vektor

Definice:

Nechť $(\Omega, \mathcal{A}, P)$ je pravděpodobnostní prostor, $\mathbf{X} = (X_1, \ldots, X_n)$ náhodný vektor definovaný na měřitelném prostoru $(\Omega, \mathcal{A})$. Nechť $\Phi(x_1, \ldots, x_n)$ je jeho distribuční funkce. Řekneme, že náhodný vektor $\mathbf{X}$ je **spojitý** (vzhledem k P), právě když existuje po částech spojitá nezáporná reálná funkce $\varphi(x_1, \ldots, x_n)$ tak, že pro

$\forall (x_1, \ldots, x_n) \in R^n : \Phi(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \ldots \int_{-\infty}^{x_n} \varphi(t_1, \ldots, t_n) dt_1 \ldots dt_n$. Tato funkce se nazývá **hustota pravděpodobnosti** spojitého náhodného vektoru $\mathbf{X}$.

Vlastnosti hustoty pravděpodobnosti

Věta: Nechť $\varphi(x_1,\ldots,x_n)$ je hustota pravděpodobnosti spojitého náhodného vektoru $\mathbf{X} = (X_1, \ldots, X_n)$. Pak platí:

a) $\forall (x_1,\ldots,x_n) \in R^n : \varphi(x_1,\ldots,x_n) \geq 0$ (vlastnost S1 - nezápornost)

b) $\int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \varphi(x_1,\ldots,x_n)dx_1 \ldots dx_n = 1$ (vlastnost S2 - normovanost)

c) $\forall B \in \mathcal{B}^n : P(\mathbf{X} \in B) = \int_B \ldots \int \varphi(x_1,\ldots,x_n)dx_1 \ldots dx_n$

d) $\varphi(x_1,\ldots,x_n) = \frac{\partial^n \Phi(x_1,\ldots,x_n)}{\partial x_1 \ldots \partial x_n}$ ve všech bodech spojitosti funkce $\varphi(x_1,\ldots,x_n)$.

e) $\forall i \in \{1,\ldots,n\} : \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \varphi(x_1,\ldots,x_n)dx_1 \ldots dx_{i-1}dx_{i+1}\ldots dx_n = \varphi_i(x_i)$.

Funkce $\varphi_i(x_i)$ je hustota náhodné veličiny X_i. Nazývá se **marginální hustota**. Funkce $\varphi(x_1,\ldots,x_n)$ se v této souvislosti nazývá **simultánní hustota**. Podobně lze zavést marginální hustoty k proměnných, kde $k \in \{2,3,\ldots,n-1\}$.

Existenční věta

Věta: (existenční věta)

a) Skalární případ: Jestliže funkce $\varphi(x)$ má vlastnosti S1, S2 z věty o vlastnostech hustoty skalární náhodné veličiny, pak existuje pravděpodobnostní prostor $(\Omega, \mathcal{A}, P)$ a na něm definovaná skalární spojitá náhodná veličina X tak, že $\varphi(x)$ je její hustota.

b) Vektorový případ: Jestliže funkce $\varphi(x_1, \ldots, x_n)$ má vlastnosti S1, S2 z věty o vlastnostech hustoty náhodného vektoru, pak existuje pravděpodobnostní prostor $(\Omega, \mathcal{A}, P)$ a na něm definovaný spojitý náhodný vektor $\mathbf{X} = (X_1, \ldots, X_n)$ tak, že $\varphi(x_1, \ldots, x_n)$ je jeho hustota.

Příklad

Příklad: Spojitý náhodný vektor (X_1, X_2) má simultánní hustotu pravděpodobnosti $\varphi(x_1, x_2) = \frac{1}{\pi^2(1+x_1^2)(1+x_2^2)}$.

Najděte obě marginální hustoty $\varphi_1(x_1), \varphi_2(x_2)$.

Řešení :

$$\varphi_1(x_1) = \int_{-\infty}^{\infty} \frac{1}{\pi^2(1+x_1^2)(1+x_2^2)}\,dx_2 = \frac{1}{\pi^2(1+x_1^2)} \int_{-\infty}^{\infty} \frac{1}{1+x_2^2}\,dx_2 =$$

$$= \frac{1}{\pi^2(1+x_1^2)} \left[\operatorname{arctg} x_2\right]_{-\infty}^{\infty} = \frac{1}{\pi^2(1+x_1^2)} \left(\frac{\pi}{2} - (-\frac{\pi}{2})\right) = \frac{1}{\pi(1+x_1^2)}.$$

Analogicky dostáváme $\varphi_2(x_2) = \frac{1}{\pi(1+x_2^2)}$.

Vybraná rozložení diskrétních a spojitých náhodných veličin

Motivace

Nyní se seznámíme s přehledem důležitých pravděpodobnostních funkcí a hustot pravděpodobnosti. Uvedeme nejenom analytické vyjádření těchto funkcí, ale též jejich grafy. Vysvětlíme rovněž, v jakých situacích se lze s uvedenými rozloženími pravděpodobností setkat. Zvláštní pozornost budeme věnovat normálnímu rozložení, které hraje velkou roli v celé řadě praktických aplikací počtu pravděpodobnosti i v matematické statistice.

Označení

Známe -li distribuční funkci $\Phi(x)$ náhodné veličiny X (resp. pravděpodobnostní funkci $\pi(x)$ v diskrétním případě resp. hustotu pravděpodobnosti $\varphi(x)$ ve spojitém případě), pak řekneme, že známe rozložení pravděpodobností (zkráceně rozložení) náhodné veličiny X. Toto rozložení závisí na nějakém parametru ϑ, což je nejčastěji reálné číslo nebo reálný vektor.

Zápis $X \sim L(\vartheta)$ čteme: náhodná veličina X má rozložení L s parametrem ϑ.

Na webu:

http://en.wikipedia.org/wiki/List_of_probability_distributions

Vybraná rozložení diskrétních náhodných veličin

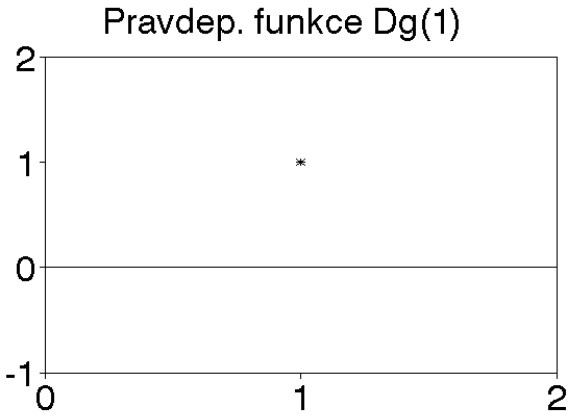
Důležitá diskrétní rozdělení:

- Degenerované rozložení
- Alternativní (Bernoulliho) rozdělení
- Binomické rozdělení
- Multinomické rozdělení
- Poissonovo rozdělení
- Negativně binomické (Pascalovo) rozdělení
- Geometrické rozdělení (zvláštní případ negativně binomického rozdělení)
- Hypergeometrické rozdělení
- Rovnoměrné rozdělení

Degenerované rozložení

Degenerované rozložení: Náhodná veličina X nabývá pouze konstantní hodnoty μ, píšeme $X \sim Dg(\mu)$.

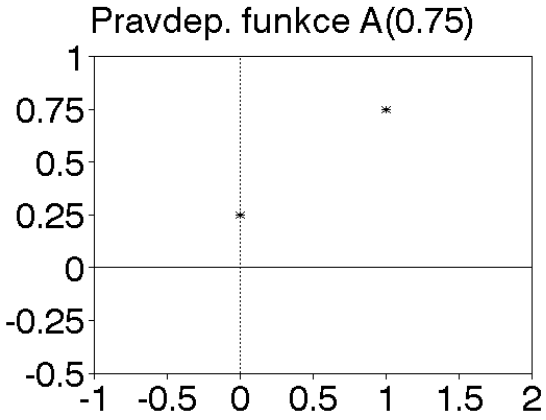
$$\pi(x) = \begin{cases} 1 \text{ pro } x = \mu \\ 0 \text{ jinak} \end{cases}$$



Alternativní rozložení

Alternativní rozložení: Náhodná veličina X udává počet úspěchů v jednom pokusu, přičemž pravděpodobnost úspěchu je ϑ. Píšeme $X \sim A(\vartheta)$.

$$\pi(x) = \begin{cases} 1-\vartheta \text{ pro } x = 0 \\ \vartheta \text{ pro } x = 1 \\ 0 \text{ jinak} \end{cases} \quad \text{neboli} \quad \pi(x) = \begin{cases} \vartheta^x (1-\vartheta)^{1-x} \text{ pro } x = 0,1 \\ 0 \text{ jinak} \end{cases}$$



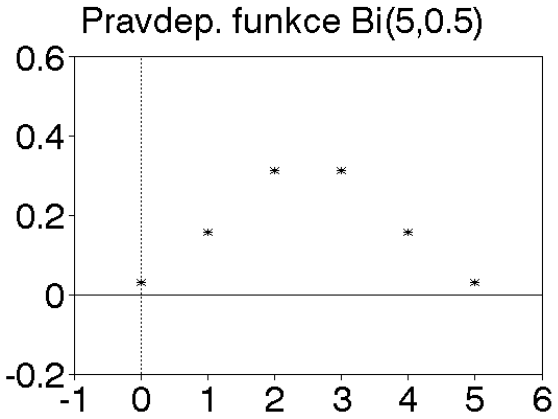
Binomické rozložení

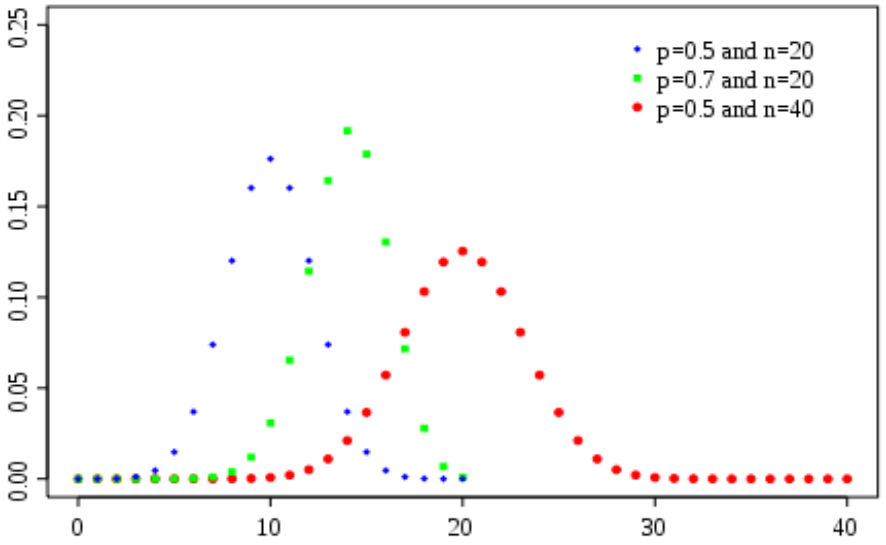
Binomické rozložení: Náhodná veličina X udává počet úspěchů v posloupnosti n nezávislých opakovaných pokusů, přičemž pravděpodobnost úspěchu je v každém pokusu ϑ. Píšeme $X \sim Bi(n, \vartheta)$.

$$\pi(x) = \begin{cases} \binom{n}{x} \vartheta^x (1-\vartheta)^{n-x} & \text{pro } x = 0, \ldots, n \\ 0 & \text{jinak} \end{cases}$$

(Alternativní rozložení je speciálním případem binomického rozložení pro n = 1.

Jsou-li $X_1, \ldots, X_n$ stochasticky nezávislé náhodné veličiny, $X_i \sim A(\vartheta)$, i = 1, ..., n, pak $X = \sum_{i=1}^{n} X_i \sim Bi(n, \vartheta)$.)





Příklad

Příklad na binomické rozložení pravděpodobností: Firma se účastní čtyř nezávislých výběrových řízení. Pravděpodobnost, že uspěje v kterémkoliv z nich, je pro všechny konkurzy stejná a je rovna 0,7. Jaká je pravděpodobnost, že firma uspěje

a) právě 2x
b) aspoň 2x
c) nejvýše 2x?

Řešení: X … počet úspěšných konkurzů, X ~ Bi(4; 0,7)

ad a) $P(X=2)=\pi(2)=\binom{4}{2}0{,}7^2 0{,}3^2=0{,}2646$

ad b) $P(X\geq 2)=\pi(2)+\pi(3)+\pi(4)=\binom{4}{2}0{,}7^2 0{,}3^2+\binom{4}{3}0{,}7^3 0{,}3+\binom{4}{4}0{,}7^4=0{,}9163$

ad c) $P(X\leq 2)=\Phi(2)=\pi(0)+\pi(1)+\pi(2)=\binom{4}{0}0{,}3^4+\binom{4}{1}0{,}7\cdot 0{,}3^3+\binom{4}{2}0{,}7^2 0{,}3^2=0{,}3483$

Příklad

Pravděpodobnost narození chlapce je 0,515. Určete takový počet dětí, aby pravděpodobnost, že mezi nimi bude aspoň jeden chlapec, byla vetší než 0,99.

Řešení:

Označme jako X veličinu udávající počet chlapců mezi n dětmi, je X~Bi(n,0,515). Hledáme takové n, aby P(X > 0) > 0,99, přitom platí

$$P(X>0)=1-P(X\le 0)=1-P(X=0)=1-\binom{n}{0}\cdot 0{,}515^0\cdot(1-0{,}515)^{n-0}$$

$$\Rightarrow \quad 1-(0{,}485)^n > 0{,}99$$

$$\Rightarrow \quad n > \frac{\ln 0{,}01}{\ln 0{,}485} \cong 6{,}36$$

$$\Rightarrow \quad n \ge 7$$

Multinomické rozložení

Multinomické rozložení: Zobecnění binomického rozložení. Složky náhodného vektoru $(X_1,\ldots X_k)$ udávají počty úspěchů (nastane jev $A_1,\ldots A_k$) v posloupnosti n nezávislých opakovaných pokusů, přičemž pravděpodobnosti úspěchů jsou $\vartheta_1,\ldots,\vartheta_k$. Předpokládáme, že při každém pokusu nastane právě jeden z jevů $A_1,\ldots A_k$, přičemž platí $\vartheta_1+,\ldots,+\vartheta_k=1$. Píšeme

$$X \sim Mu(n,\vartheta_1,\ldots,\vartheta_k)$$

$$\pi(x_1,\ldots,x_k)=\frac{n!}{x_1!\ldots x_k!}\vartheta_1^{x_1}\cdot\ldots\cdot\vartheta_k^{x_k}, \quad x_1,\ldots,x_k\in\{1,\ldots,n\}, \sum_{i=1}^{k}x_i=n$$

$$=0 \qquad\qquad jinak$$

Platí: $X_j \sim Bi(n,\vartheta_j)$

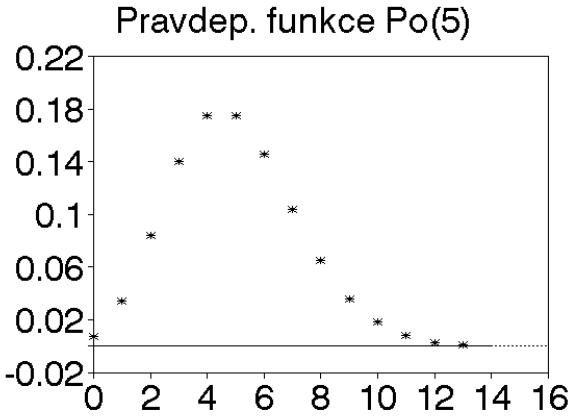
Multinomické rozložení – příklady využití

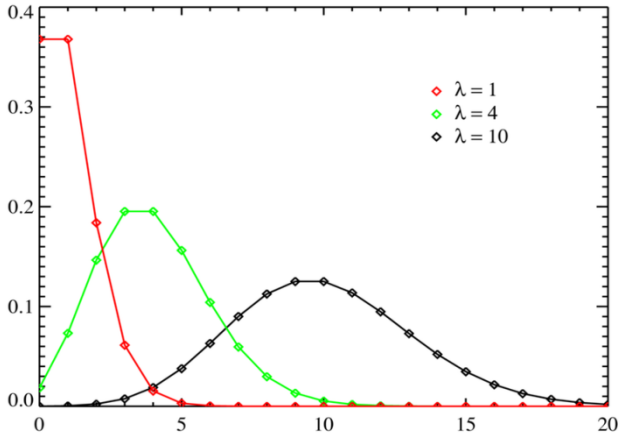
- Předvolební průzkum:
  - n – počet tázaných
  - ϑ_j – skutečný podíl voličů j-té strany v populaci
  - X_j – počet (četnost) voličů j-té strany ve výběru
- Hody hrací kostkou:
  - n – počet hodů
  - $\vartheta_1,\ldots,\vartheta_6$ – pravděpodobnost jednotlivých stran kostky
  - $X_1,\ldots X_6$ – absolutní četnosti jednotlivých stran kostky
- Krevní skupiny:
  - n=4 (skupiny 0,A,B,AB)
  - $\vartheta_0,\vartheta_A,\vartheta_B,\vartheta_{AB}$ – pravděpodobnosti skupin 0, A, B, AB
  - X_0, X_A, X_B, X_{AB} – počty osob se skupinami 0, A, B, AB

Poissonovo rozložení

Poissonovo rozložení: Náhodná veličina X udává počet událostí, které nastanou v jednotkovém časovém intervalu (resp. jednotkové oblasti), přičemž události nastávají náhodně, jednotlivě a vzájemně nezávisle. Parametr $\lambda > 0$ je střední počet těchto událostí. Píšeme $X \sim Po(\lambda)$.
(Poissonovým rozložením se řídí např. počet výzev, které dojdou na telefonní ústřednu během určitého časového intervalu nebo počet mikroorganizmů v zorném poli mikroskopu. Jde o tzv. řídce se vyskytující jevy.)

$$\pi(x) = \begin{cases} \dfrac{\lambda^x}{x!} e^{-\lambda} & \text{pro } x = 0, 1, \ldots \\ 0 & \text{jinak} \end{cases}$$





Příklad

Vztah mezi pravděpodobnostní funkcí binomického a Poissonova rozložení:
Nechť náhodná veličina X ~ Po(λ) a náhodná veličina Y ~ Bi(n, ϑ_n). Nechť $\vartheta_n \to 0$ pro $n \to \infty$ a přitom $n\vartheta_n \to \lambda$. Pak pravděpodobnostní funkce náhodné veličiny Y konverguje k pravděpodobnostní funkci náhodné veličiny X, tj.

$$\lim_{n\to\infty}\binom{n}{y}\vartheta_n^{\,y}(1-\vartheta_n)^{n-y} = \frac{\lambda^y}{y!}e^{-\lambda}$$

(Aproximace binomického rozložení pomocí Poissonova rozložení je vyhovující, když n > 30 a ϑ < 0,1.)

Příklad na Poissonovo rozložení: Dělnice v přádelně obsluhuje 800 vřeten. Pravděpodobnost toho, že se příze přetrhne během časového intervalu délky t, je pro všechna vřetena stejná a je rovna 0,005. Určete pravděpodobnost, že během intervalu délky t dojde k nejvýše 10 přetržením.
Řešení: Y – počet přetržení v časovém intervalu délky t, Y ~ Bi(800;0,005).

Přesný výpočet: $P(Y \le 10) = \sum_{y=0}^{10}\binom{800}{y}0{,}005^y(1-0{,}005)^{800-y} = 0{,}997239$

00).

Aproximativní výpočet: podmínky dobré aproximace jsou splněny, parametr

$\lambda = n\vartheta = 800.0{,}005 = 4,\ P(Y \le 10) = \sum_{y=0}^{10}\frac{4^y}{y!}e^{-4} = 0{,}9971602$

Příklad

1) Průměrný telefonní hovor trvá 1,5 min. Má-li ústředna 10 linek a dochází-li průměrně k 120 hovorům za hodinu, jaká je pravděpodobnost ztráty volání?

Řešení: X udává počet volajících, X~Po(2*1,5). Ke ztrátě volání dojde, pokud chce současně volat více než 10 volajících (tj. není volná linka). Tedy

$$P(X>10)=1-P(X\le 10)=1-\sum_{x=0}^{10}\frac{3^x}{x!}e^{-3}\cong 0{,}001.$$

2) Průměrný telefonní hovor trvá 1,5 min. Kolik linek musí ústředna mít, dochází-li průměrně k 240 hovorům za hodinu a pravděpodobnost ztráty volání nemá překročit a) 0,01, b) 0,001?

Řešení: X udává počet volajících, X~Po(240/60*1,5). Hledáme n tak aby $P(X>n)\le 0{,}01$

tj. $P(X\le n)\ge 0{,}99 \Rightarrow \sum_{x=0}^{n}\frac{6^x}{x!}e^{-6}\ge 0{,}99 \Rightarrow n=12.$

Pro případ b) chceme $\sum_{x=0}^{n}\frac{6^x}{x!}e^{-6}\ge 0{,}999 \Rightarrow n=15.$

Negativní binomické (Pascalovo) rozložení

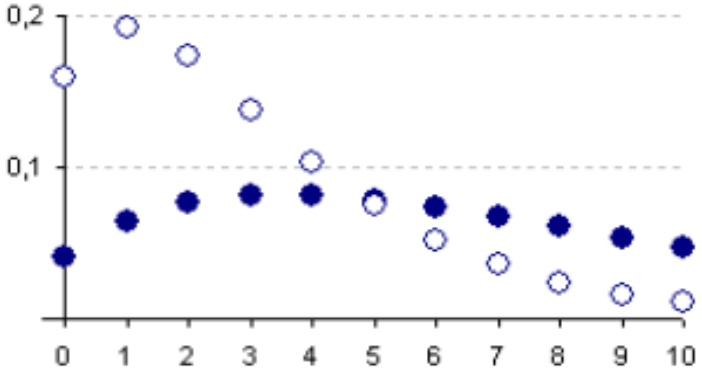
Negativní binomické rozložení (Pascalovo):

Náhodná veličina X udává počet neúspěchů před n-tým úspěchem v posloupnosti n nezávislých opakovaných pokusů, přičemž pravděpodobnost úspěchu je v každém pokusu ϑ. Píšeme X ~ NB(n, ϑ).

$$\pi(x) = \binom{n+x-1}{x} \vartheta^n (1-\vartheta)^x, \quad x = 0,1,\ldots, \quad 0 < \vartheta < 1$$

$$= 0 \qquad\qquad jinak$$

$NB(2,0.2)$ *a* $NB(2,0.4)$

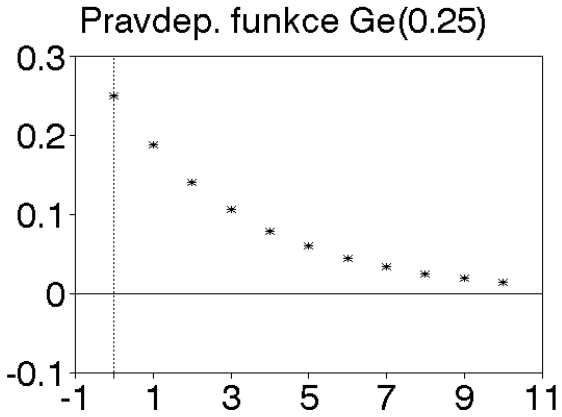


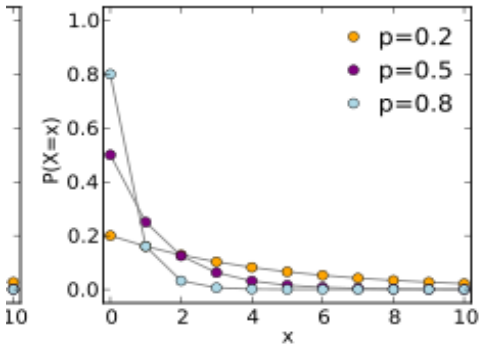
➢ Negativně binomické rozdělení lze definovat obecněji. Tak jak je zde uvedeno jde o rozdělení Pascalovo.

Geometrické rozložení

Geometrické rozložení: Náhodná veličina X udává počet neúspěchů v posloupnosti opakovaných nezávislých pokusů předcházejících prvnímu úspěchu, přičemž pravděpodobnost úspěchu je v každém pokusu rovna ϑ. Píšeme X ~ Ge(ϑ)

$$\pi(x) = \begin{cases} (1-\vartheta)^x \vartheta \text{ pro } x = 0, 1, \ldots \\ 0 \text{ jinak} \end{cases}$$





Příklad

Dva hráči střídavě házejí kostkou. Vyhrává ten, kdo první hodí šestku. Jaká je pravděpodobnost, že vyhraje ten, který začínal?

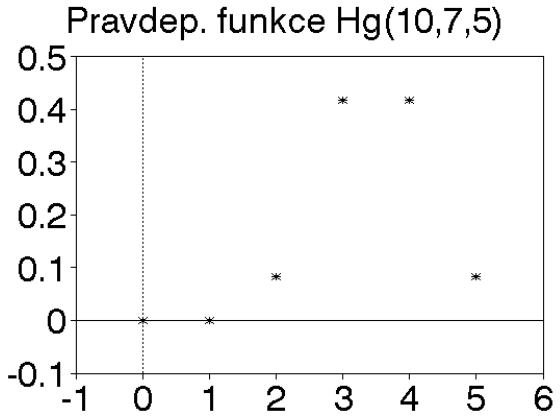
Řešení: X udává počet nehození šestky (neúspěch) před prvním hozením šestky (úspěch), X ~ Ge(1/6). Hledáme tedy pravděpodobnost jevu
A: 1. úspěch po sudém počtu neúspěchů.

$$P(A)=\sum_{k=0}^{\infty}\frac{1}{6}\left(\frac{5}{6}\right)^{2k}=\frac{1}{6}\frac{1}{1-\frac{25}{36}}=\frac{6}{11}=0{,}545$$

Hypergeometrické rozložení

Hypergeometrické rozložení: V souboru N prvků je M prvků označeno. Náhodně vybereme n prvků bez vracení. Náhodná veličina X udává počet vybraných označených prvků. Píšeme X ~ Hg(N, M, n)

$$\pi(x) = \begin{cases} \dfrac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} & \text{pro } x = \max\{0, M - N + n\}, \ldots, \min\{M, n\} \\ 0 & \text{jinak} \end{cases}$$



Příklad

V klobouku jsou 3 černé a 4 bílé koule. Určete pravděpodobnost, že při vytažení 3 koulí budou aspoň 2 černé.

Řešení: X udává počet vytažených černých koulí, X ~ HG(7,3,3). Hledaná pravděpodobnost je

$$P(X \geq 2) = 1 - P(X \leq 1) = 1 - \frac{\binom{3}{0}\binom{4}{3} + \binom{3}{1}\binom{4}{2}}{\binom{7}{3}} =$$

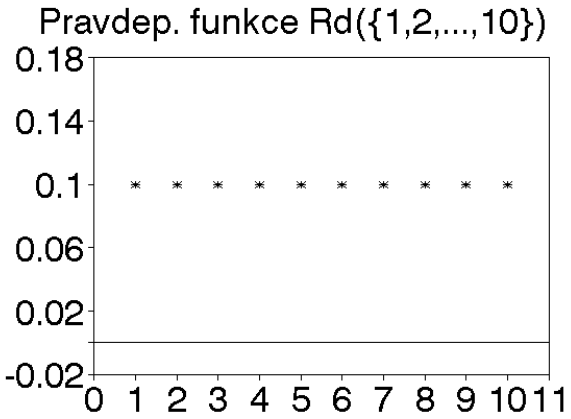
$$1 - \frac{1 \cdot 4 + 3 \cdot 6}{35} = 1 - \frac{22}{35} = \frac{13}{35} = 0{,}371.$$

Rovnoměrné diskrétní rozložení

Rovnoměrné diskrétní rozložení: Nechť G je konečná množina o n prvcích. Náhodná veličina X nabývá se stejnou pravděpodobností každé hodnoty z množiny G. Píšeme X ~ Rd(G)

$$\pi(x) = \begin{cases} \frac{1}{n} \text{ pro } x \in G \\ 0 \text{ jinak} \end{cases}$$

(Typickým příkladem je náhodná veličina udávající počet ok při hodu kostkou.)



Vybraná rozložení spojitých náhodných veličin

Důležitá spojitá rozdělení:

- Rovnoměrné rozdělení
- Normální rozdělení (označované také jako Gaussovo rozdělení)
- Logaritmicko-normální rozdělení (také log-normální rozdělení)
- Studentovo rozdělení
- Fischerovo-Snedecorovo rozdělení
- χ^2 rozdělení (Chí-kvadrát)
- Cauchyho rozdělení
- Exponenciální rozdělení
- Laplaceovo rozdělení (nebo také dvojitě exponenciální rozdělení)
- Weibullovo rozdělení

Rovnoměrné spojité rozložení

Rovnoměrné spojité rozložení: Předpokládejme, že veličina X

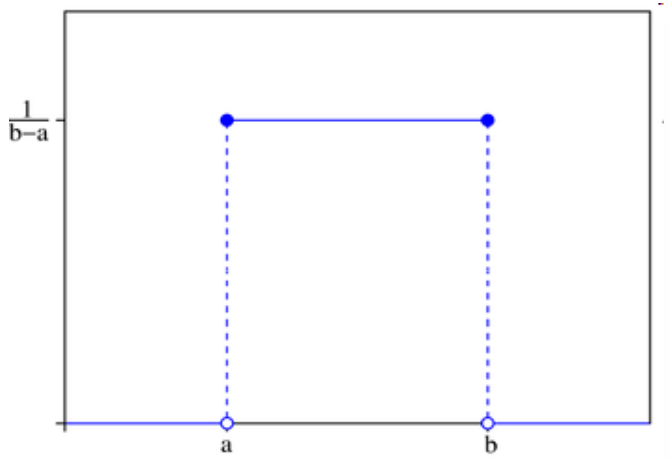
- může nabýt jakékoliv hodnoty mezi čísly a, b
- pravděpodobnost, že nabude hodnoty z jakéhokoliv intervalu v tomto rozmezí je stejná jako pravděpodobnost, že nabude hodnoty z jakéhokoliv jiného intervalu stejné délky.

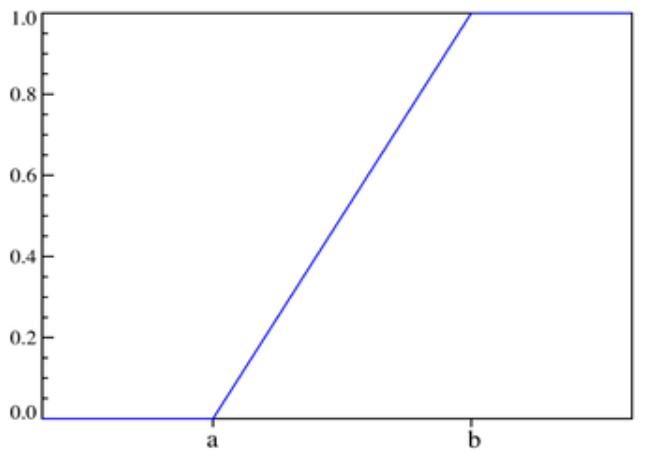
Jsou-li tyto podmínky splněny, pak X má rovnoměrné spojité rozložení na intervalu (a, b). Hustota pravděpodobnosti náhodné veličiny X je konstantní na intervalu (a, b) a plocha pod křivkou hustoty tvoří obdélník.

Píšeme X ~ Rs(a, b).

$$\varphi(x) = \begin{cases} \dfrac{1}{b-a} & \text{pro } x \in (a,b) \\ 0 & \text{jinak} \end{cases}$$

$$\Phi(x) = \begin{cases} 0 & x \le a \\ \dfrac{x-a}{b-a} & x \in (a,b) \\ 1 & x \ge b \end{cases}$$

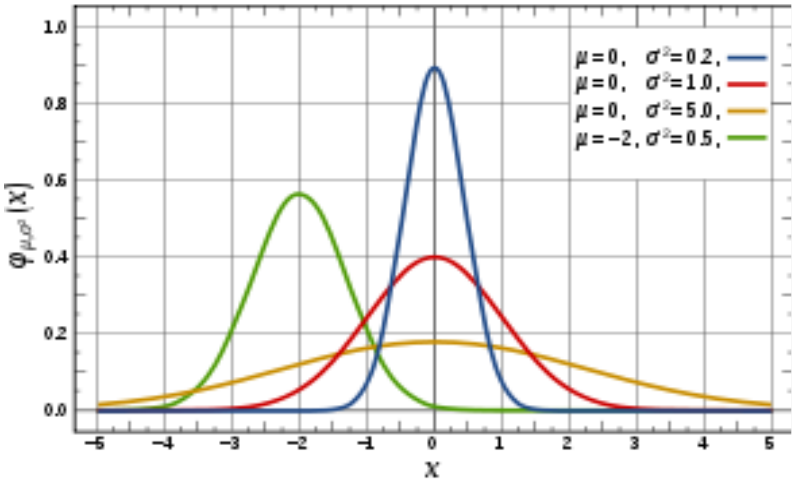


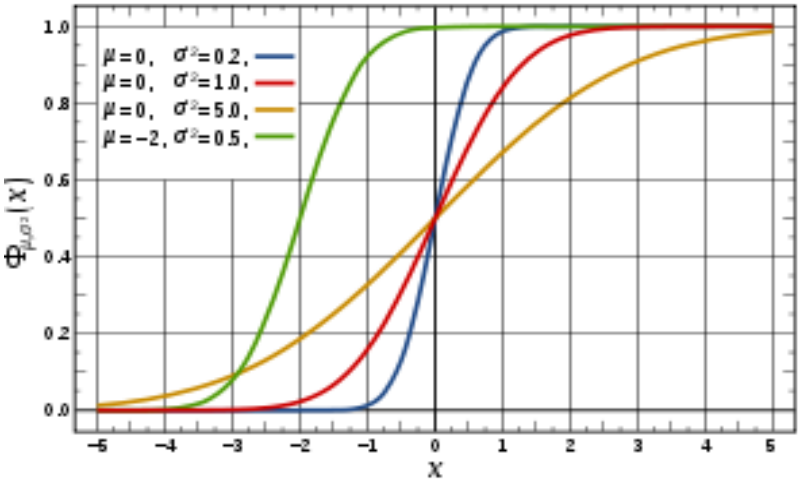


Normální rozložení

Normální rozložení: Tato náhodná veličina vzniká např. tak, že ke konstantě μ se přičítá velké množství nezávislých náhodných vlivů mírně kolísajících kolem nuly. Proměnlivost těchto vlivů je vyjádřena konstantou $\sigma > 0$.

Píšeme $X \sim N(\mu, \sigma^2)$, hustota $\varphi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. Grafem této hustoty je tzv. Gaussova křivka.

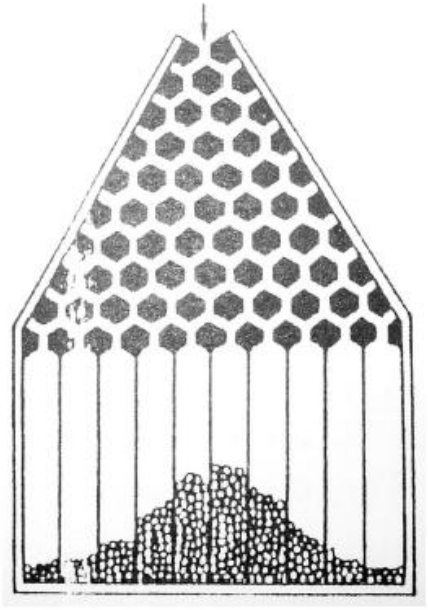




Galtonova deska

Ilustrace vzniku normálního rozložení pomocí Galtonovy desky:
Deska obsahuje n řad pravidelně uspořádaných klínů, a to tak, že v k-té řadě je právě k klínů. Do otvoru nahoře padají kuličky, které jsou v každé řadě se stejnou pravděpodobností 1/2 vychylovány vlevo nebo vpravo. Pod poslední radou je n - 1 přihrádek, ve kterých se kuličky shromažďují. Nasypeme-li do tohoto systému velké množství kuliček, vytvoří v přihrádkách jakýsi ”kopec”, jehož tvar je velmi podobný tvaru grafu hustoty náhodné veličiny s normálním rozložením.
Náhodné vychylování kuliček jednotlivými řadami překážek je možno chápat jako speciální případ velkého množství chybových faktorů, náhodně působících na nějaký proces, jako působení mnoha blíže nespecifikovatelných vlivů, které ovlivňují zcela náhodně rozložení jeho výsledku.

Obrázek

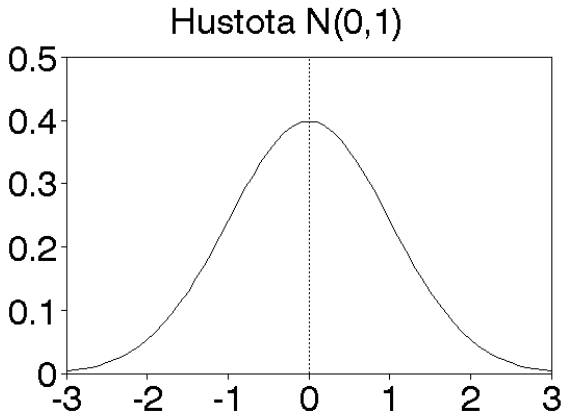


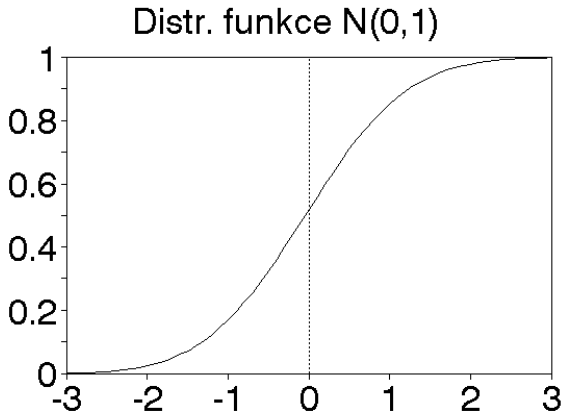
Standardizované normální rozložení

Standardizované normální rozložení:

Pro $\mu = 0$, $\sigma^2 = 1$ se jedná o standardizované normální rozložení, píšeme

$U \sim N(0, 1)$. Hustota pravděpodobnosti má v tomto případě tvar $\varphi(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$.





$\Phi(u) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$ je tabelována pro $u \geq 0$, pro $u < 0$ se používá přepočtový vzorec $\Phi(-u) = 1 - \Phi(u)$.

Příklad

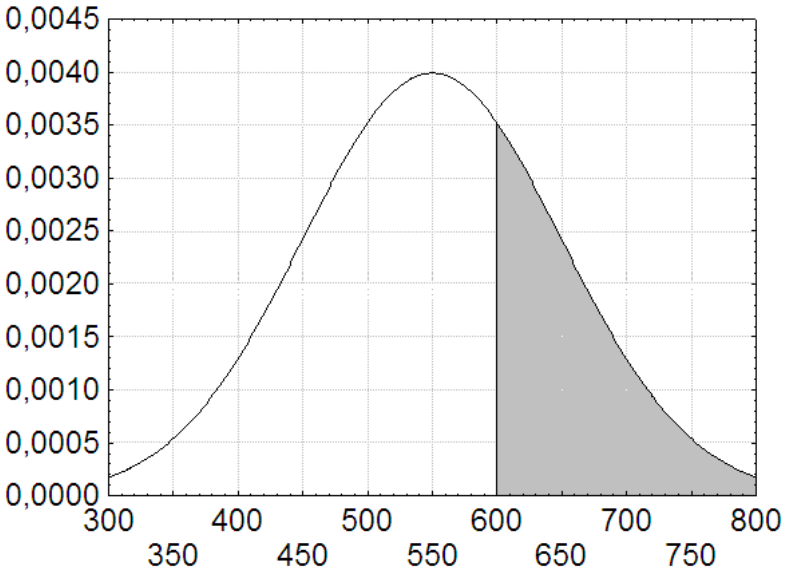
Příklad na normální rozložení: Výsledky u přijímacích zkoušek na jistou VŠ jsou normálně rozloženy s parametry $\mu = 550$ bodů, $\sigma = 100$ bodů. S jakou pravděpodobností bude mít náhodně vybraný uchazeč aspoň 600 bodů?

Řešení:

X – výsledek náhodně vybraného uchazeče, $X \sim N(550, 100^2)$,

$P(X \geq 600) = 1 - P(X \leq 600) + P(X = 600) = 1 - P(X \leq 600) =$

$$=1 - P\left(\frac{X-\mu}{\sigma} \leq \frac{600-\mu}{\sigma}\right) = 1 - P\left(U \leq \frac{600-550}{100}\right) = 1 - \Phi(0{,}5) = 1 - 0{,}69146 = 0{,}30854.$$



Normální rozložení - vlastnosti

Některé vlastnosti normálního rozložení:

Jestliže $X \sim N(\mu, \sigma^2)$, pak $U = \frac{X-\mu}{\sigma} \sim N(0,1)$.

Jestliže $X \sim N(\mu, \sigma^2)$, a $Y = a + bX$, pak $Y \sim N(a + b\mu, b^2\sigma^2)$.

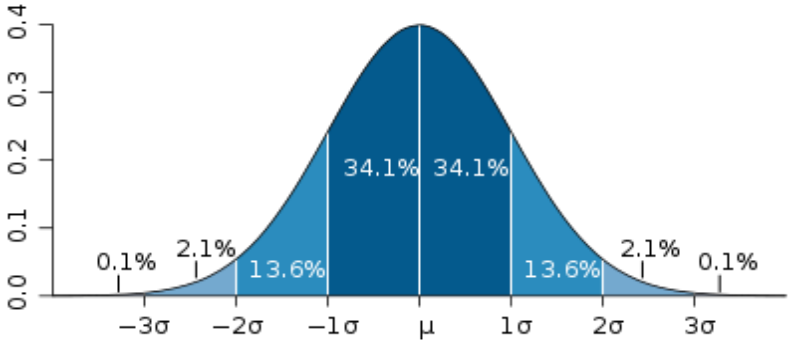
Jestliže $X_1, \ldots, X_n$ jsou stochasticky nezávislé náhodné veličiny, $X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, \ldots, n$, $Y = \sum_{i=1}^{n} X_i$, pak $Y \sim N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$.

Význam normálního rozložení:

Normální rozložení hraje ústřední roli v počtu pravděpodobnosti i matematické statistice. Jeho význam spočívá jednak v tom, že normálním rozložením se řídí pravděpodobnostní chování mnoha náhodných veličin a jednak v tom, že za určitých podmínek konverguje k normálnímu rozložení součet nezávislých náhodných veličin s týmž rozložením (viz centrální limitní věta).

„koncentrace hodnot“ normální NV:

Přes 68% hodnot „leží“ v intervalu (μ-σ, μ+σ).
Přes 95% hodnot „leží“ v intervalu (μ-2σ, μ+2σ).
Přes 99% hodnot „leží“ v intervalu (μ-3σ, μ+3σ).



Dvojrozměrné normální rozložení

Definice:

O spojitém náhodném vektoru $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ říkáme, že má dvojrozměrné normální rozložení s parametry $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ a $\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$, když jeho hustota je dána vzorcem

$$\varphi(\mathbf{x}) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\frac{x_1-\mu_1}{\sigma_1}\cdot\frac{x_2-\mu_2}{\sigma_2} + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]}, \quad \mathbf{x} \in \mathbf{R}^2.$$

Zkráceně píšeme $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2(\mu, \mathbf{\Sigma})$.

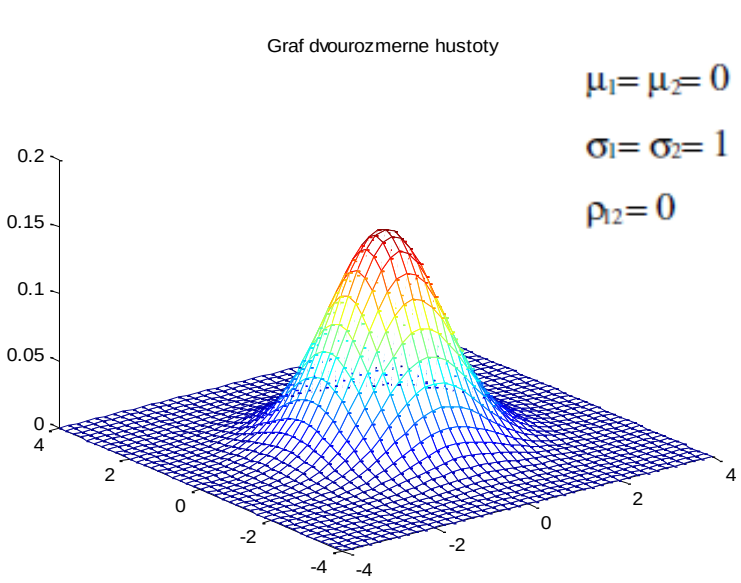
Pro $\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ a $\mathbf{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ mluvíme o standardizovaném dvojrozměrném normálním rozložení.

Poznámka:

Význam parametrů je následující:

$\mu_1 = E(X_1),\ \mu_2 = E(X_2),\ \sigma_1^2 = D(X_1),\ \sigma_2^2 = D(X_2),\ \rho = R(X_1, X_2)$

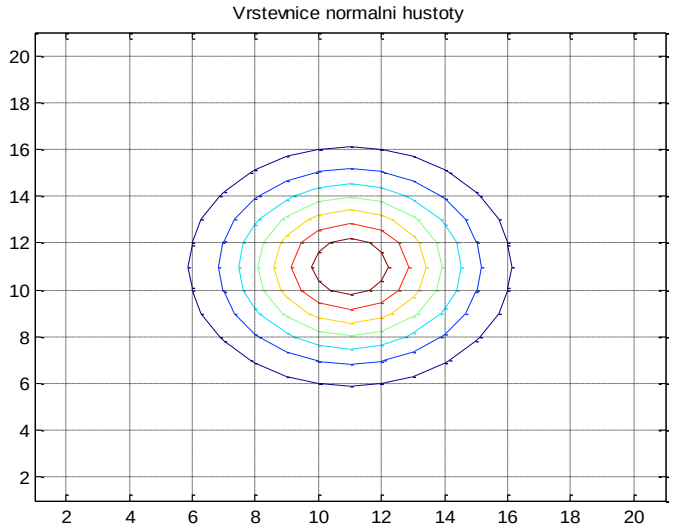
Dvojrozměrné normální rozložení

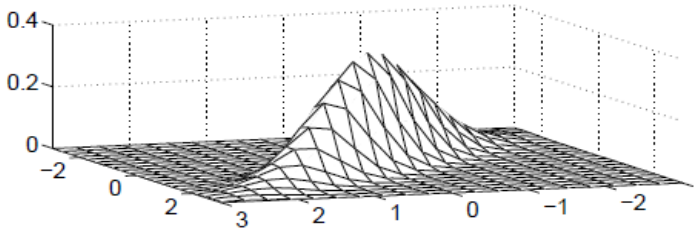


$\mu_1 = \mu_2 = 0$

$\sigma_1 = \sigma_2 = 1$

$\rho_{12} = 0$

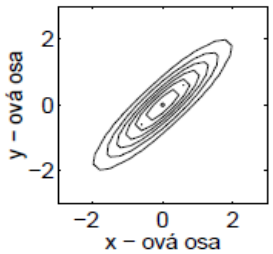


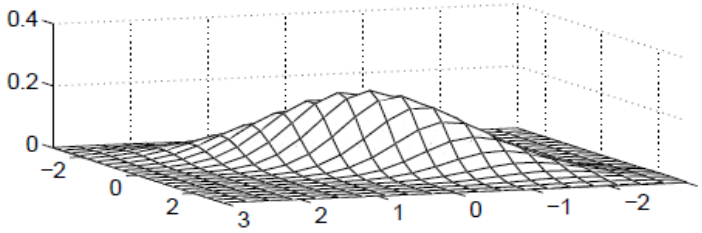


$\mu_1 = \mu_2 = 0$

$\sigma_1 = \sigma_2 = 1$

$\rho_{12} = 0.9$

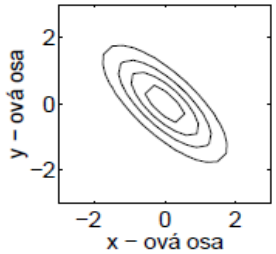




$\mu_1 = \mu_2 = 0$

$\sigma_1 = \sigma_2 = 1$

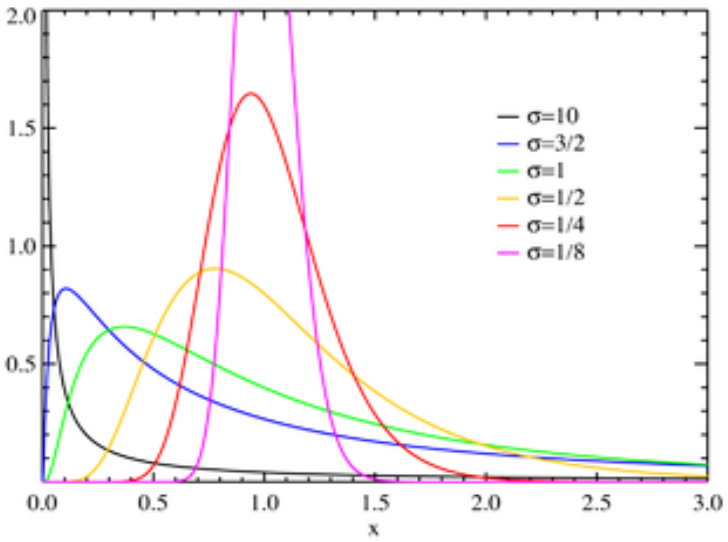
$\rho_{12} = -0.75$

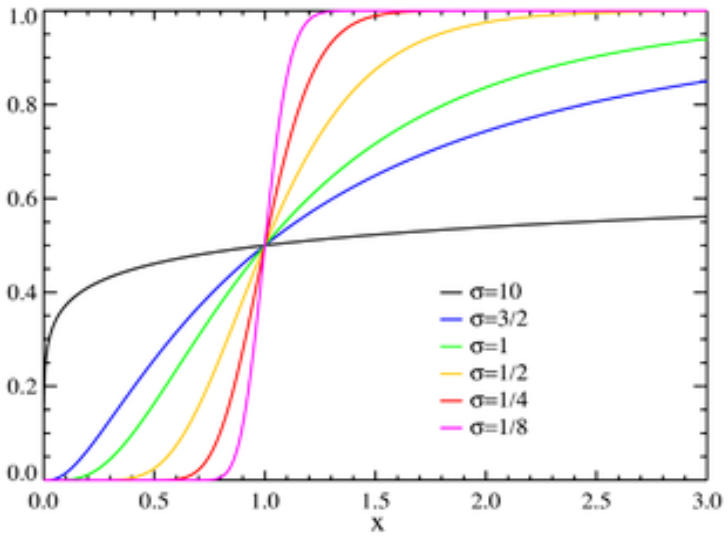


Logaritmicko normální rozložení

Logaritmicko normální rozložení: Náhodná veličina X ~ LN(μ, σ) vzniká v situacích, kdy kladná konstanta logaritmu μ je násobena velkým množstvím nezávislých náhodných veličin, kolísajících mírně kolem jedníčky. Variabilita jejich logaritmů je charakterizována parametrem σ. Logaritmicko normální rozdělení má hustotu

$$\varphi(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}}\, e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}, \quad x > 0$$



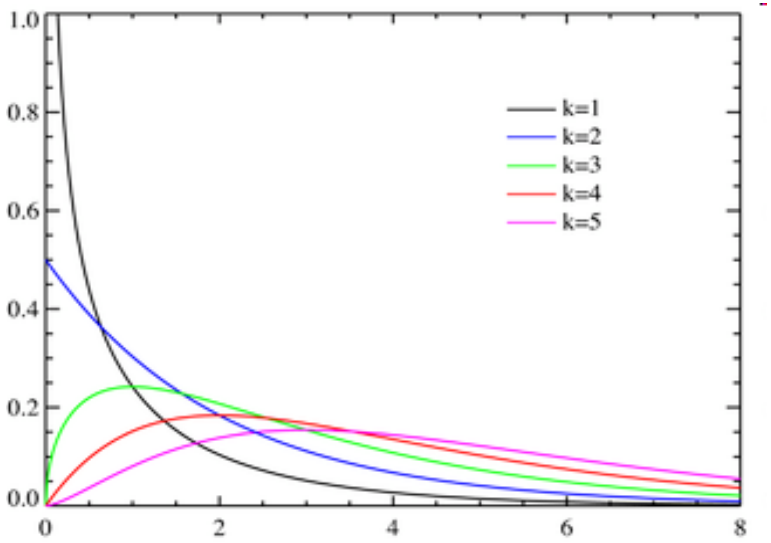


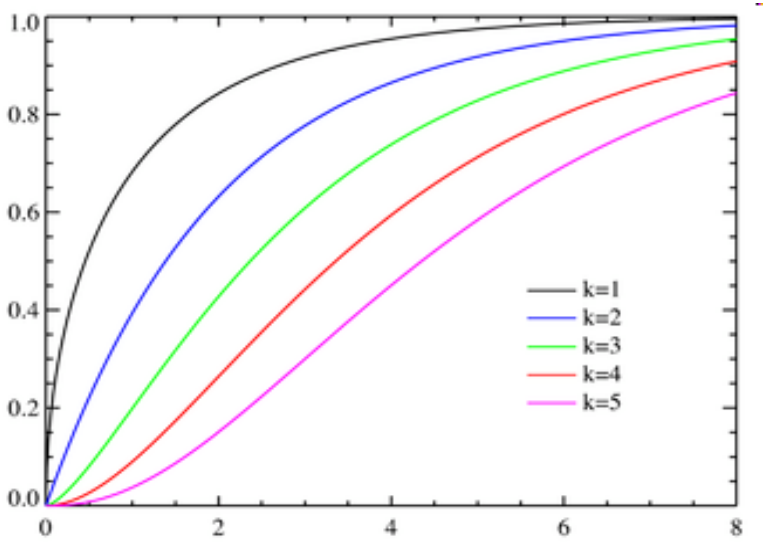
Pearsonovo χ^2 rozložení

Pearsonovo rozložení chí - kvadrát s n stupni volnosti: Nechť $X_1, ..., X_k$ jsou stochasticky nezávislé náhodné veličiny, $X_i \sim N(0, 1)$, $i = 1, ..., k$.

Pak náhodná veličina $X = X_1^2 + ... + X_k^2 \sim \chi^2(k)$.

$$\varphi(x,k) = \begin{cases} \dfrac{1}{2^{k/2} \cdot \Gamma(k/2)} \cdot x^{k/2-1} \cdot e^{-x/2} & x > 0 \\ 0 & jinak \end{cases} \qquad \Gamma(s) = \int_0^\infty e^{-t} \cdot t^{s-1}\, dt$$



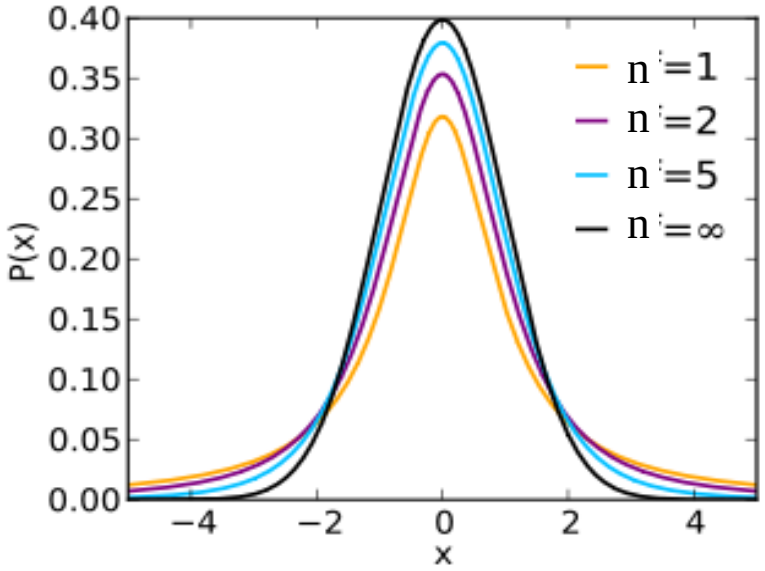


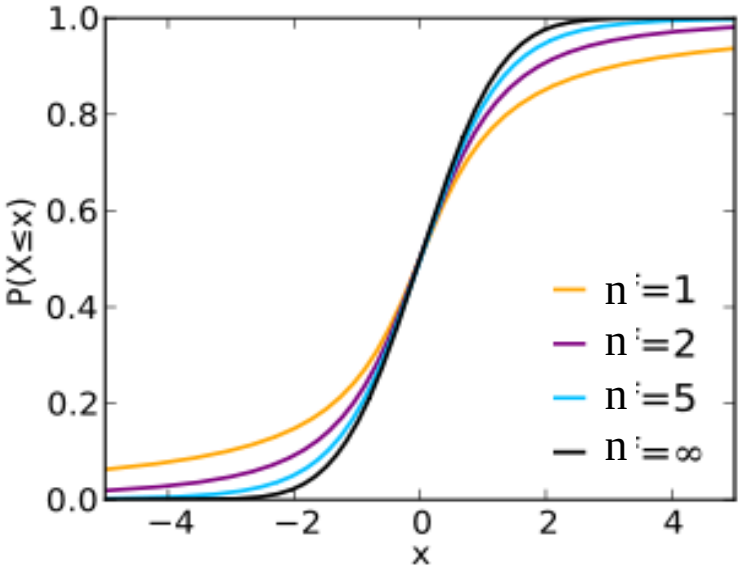
Studentovo rozložení

Studentovo rozložení s n stupni volnosti: Nechť X_1, X_2 jsou stochasticky nezávislé náhodné veličiny, $X_1 \sim N(0, 1)$, $X_2 \sim \chi^2(n)$.

Pak náhodná veličina $X = \frac{X_1}{\sqrt{\frac{X_2}{n}}} \sim t(n)$.

$$\varphi(x,n) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\cdot\Gamma(n/2)}\cdot\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}} \qquad x \in (-\infty,\infty)$$



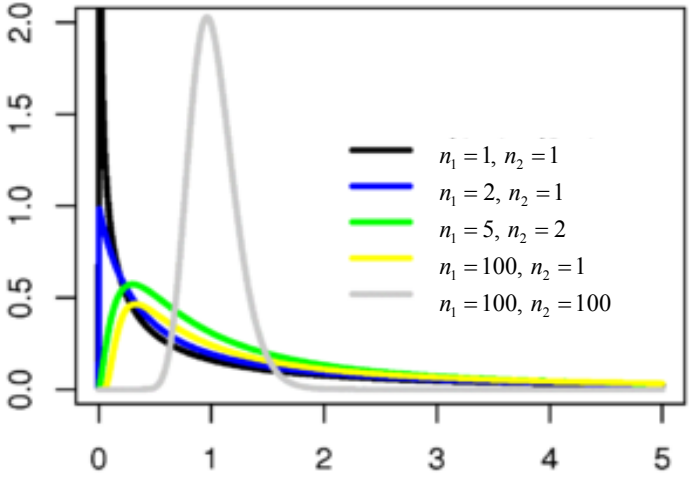


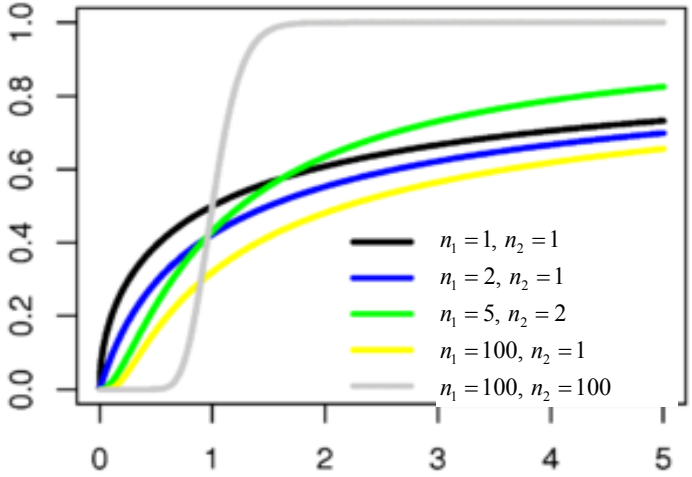
Fisher-Snedecorovo rozložení

Fisherovo-Snedecorovo rozložení s n_1 a n_2 stupni volnosti: Nechť X_1, X_2 jsou stochasticky nezávislé náhodné veličiny, $X_i \sim \chi^2(n_i)$, i = 1, 2.

Pak náhodná veličina $X = \frac{X_1 / n_1}{X_2 / n_2} \sim F(n_1, n_2)$.

$$\varphi(x, n_1, n_2) = \frac{\Gamma\left(\frac{n_1 + n_2}{2}\right) n_1^{n_1/2} n_2^{n_2/2}}{\Gamma(n_1/2)\Gamma(n_2/2)} \cdot \left(\frac{x^{(n_1-2)/2}}{(n_2 + n_1 x)^{(n_1+n_2)/2}} \right) \quad \text{pro } x > 0$$





Cauchyho rozložení

Cauchyho rozložení pravděpodobnosti s parametry x_0 a γ, pro $-\infty < x_0 < \infty$ a $\gamma > 0$, je definováno hustotou pravděpodobnosti ve tvaru

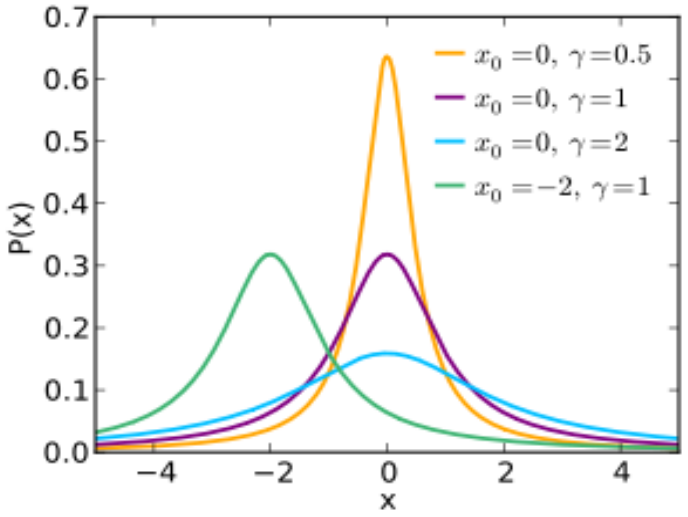
$$\varphi(x; x_0, \gamma) = \frac{1}{\pi\gamma\left[1 + \left(\frac{x-x_0}{\gamma}\right)^2\right]}.$$

$$= \frac{1}{\pi}\left[\frac{\gamma}{(x-x_0)^2 + \gamma^2}\right]$$

kde x_0 je parametr, určující umístění největší hodnoty rozdělení.
Zvláštní případ, kdy $x_0 = 0$ a $\gamma = 1$ se nazývá **standardní Cauchyho rozdělení** s hustotou pravděpodobnosti vyjádřenou vztahem

$$\varphi(x;0,1) = \frac{1}{\pi(1+x^2)}.$$

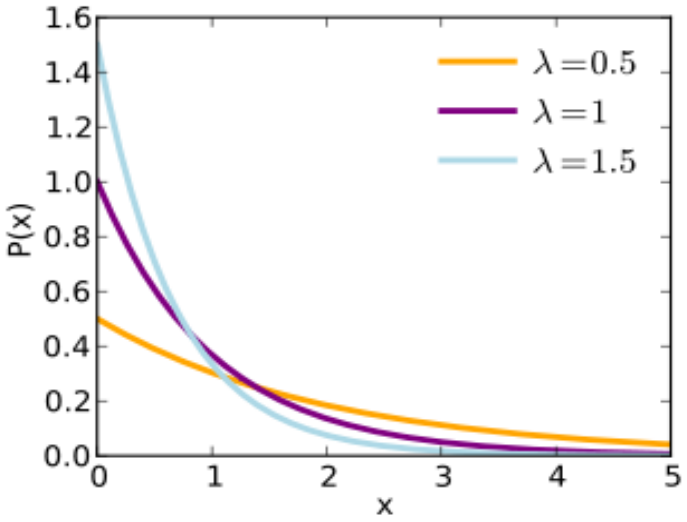
Standardní Cauchyho rozdělení je speciální případ Studentova rozdělení (pro n = 1).

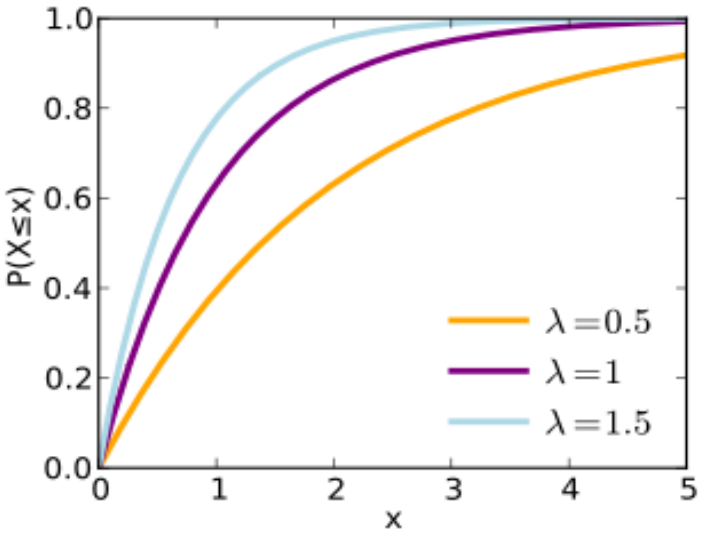


Exponenciální rozložení

Exponenciální rozložení: Náhodná veličina X udává dobu čekání na příchod nějaké události, která se může dostavit každým okamžikem se stejnou šancí bez ohledu na dosud pročekanou dobu. (Jde o tzv. čekání bez paměti.) Přitom $\frac{1}{\lambda}$ vyjadřuje střední dobu čekání. Píšeme X ~ Ex(λ)

$$\varphi(x) = \begin{cases} \lambda e^{-\lambda x} \text{ pro } x > 0 \\ 0 \text{ jinak} \end{cases}, \quad \Phi(x) = \begin{cases} 1 - e^{-\lambda x} \text{ pro } x > 0 \\ 0 \text{ jinak} \end{cases}$$





Příklad

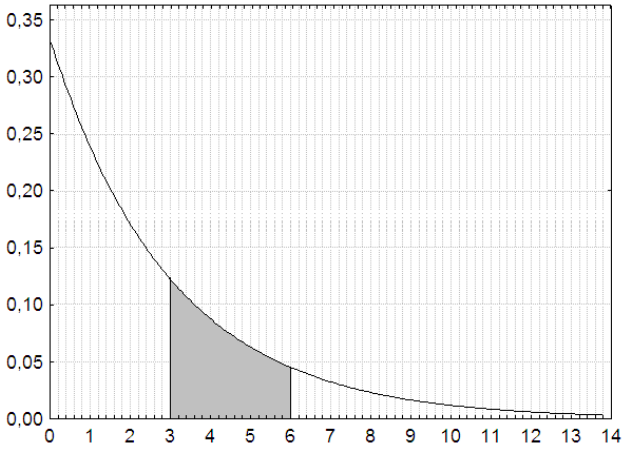
Příklad na exponenciální rozložení: Doba (v minutách) potřebná k obsloužení zákazníka v prodejně potravin je náhodná veličina, která se řídí rozložením $\mathrm{Ex}\left(\frac{1}{3}\right)$. Jaká je pravděpodobnost, že doba potřebná k obsloužení náhodně vybraného zákazníka v této prodejně bude v rozmezí od 3 do 6 minut?

Řešení:

X – doba potřebná k obsloužení náhodně vybraného zákazníka, $X \sim \mathrm{Ex}\left(\frac{1}{3}\right)$,

$$P(3 \le X \le 6) = \int_3^6 \frac{1}{3} e^{-\frac{x}{3}} dx = \frac{1}{3}(-3)\left[e^{-\frac{x}{3}}\right]_3^6 = -e^{-2} + e^{-1} = 0{,}233.$$

S pravděpodobností 0,233 bude zákazník obsloužen v době od 3 do 6 minut.

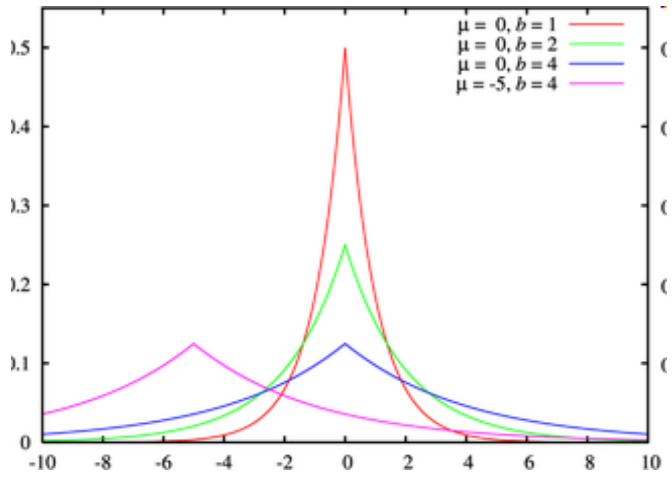


Laplaceovo rozložení

Laplaceovo rozložení: Náhodná veličina, která vznikne rozdílem dvou NV z exponenciálního rozložení, se řídí tímto rozložením. Využití ve fyzice, ekonomii – Brownův pohyb. Hustota je dána vzorcem

$$\varphi(x;\mu,b) = \frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right)$$

$$= \frac{1}{2b}\begin{cases}\exp\left(-\frac{\mu-x}{b}\right) & x < \mu \\ \exp\left(-\frac{x-\mu}{b}\right) & x \geq \mu\end{cases}$$



Platí např.:

$$X \sim Laplace(0,b) \Rightarrow |X| \sim Ex\left(\frac{1}{b}\right)$$

$$X_1 \sim Ex(\lambda_1),\ X_2 \sim Ex(\lambda_2) \Rightarrow \lambda_1 X_1 - \lambda_2 X_2 \sim Laplace(0,1)$$

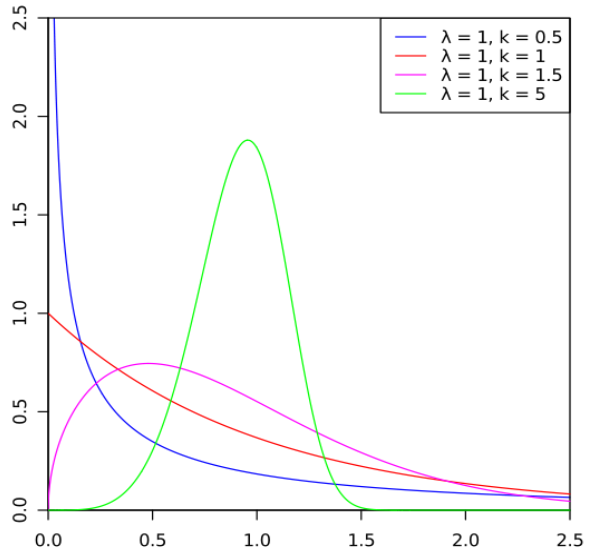
Weibullovo rozložení

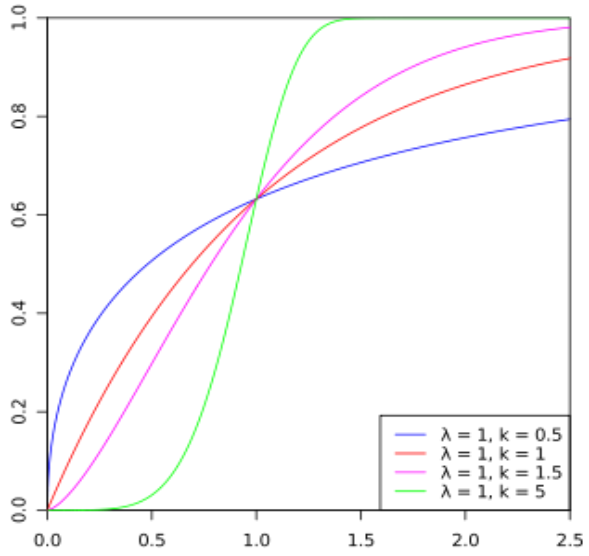
Weibullovo rozdělení: Náhodná veličina X ~ Wb(δ, ε) vyjadřuje dobu čekání na nějakou událost, která se každým okamžikem může dostavit se šancí úměrnou mocninné funkci pročekané doby. Přitom čísla δ > 0 a ε > 0 se nazývají parametry měřítka a formy.

$$\varphi(x;\delta,\varepsilon)=\begin{cases}\varepsilon\cdot\delta\cdot \mathrm{x}^{\varepsilon\text{-}1}e^{-\delta\cdot x^{\varepsilon}} & \text{pro } x>0\\ 0 & \text{pro } x\le 0\end{cases}$$

Jiná forma zápisu:

$$\varphi(x;\lambda,k)=\begin{cases}\frac{k}{\lambda}\left(\frac{\mathrm{x}}{\lambda}\right)^{k-1}e^{-\left(\frac{x}{\lambda}\right)^{k}} & \text{pro } x>0\\ 0 & \text{pro } x\le 0\end{cases}$$





9. Stochasticky nezávislé NV, generování realizací NV.

Motivace: Při provedení pokusu se může stát, že se realizace jedné náhodné veličiny Y dají jednoznačně určit ze známé realizace druhé náhodné veličiny X, tedy je mezi nimi funkční vztah Y = g(X). Takové náhodné veličiny se nazývají deterministicky závislé.
Jejich protipólem jsou náhodné veličiny stochasticky nezávislé: informace o realizaci jedné z nich nijak nemění šance, s nimiž při témž pokusu očekáváme realizaci druhé.
Např. náhodný pokus spočívá v hodu dvěma kostkami. Náhodná veličina X udává počet ok, která padla na 1. kostce a náhodná veličina Y udává počet ok, která padla na druhé kostce. Náhodné veličiny X, Y jsou stochasticky nezávislé.
Stochastickou nezávislost náhodných veličin zavádíme na základě analogie s četnostní nezávislostí znaků v daném výběrovém souboru, která se používá v popisné statistice. Musí platit multiplikativní vztah:

$\forall(x,y)\in R^2 : p(x,y)=p_1(x)p_2(y)$ pro bodové rozložení četností,

$\forall(x,y)\in R^2 : f(x,y)=f_1(x)f_2(y)$ pro intervalové rozložení četností.

V počtu pravděpodobnosti nahradíme četnostní funkci pravděpodobnostní funkcí resp. hustotu četnosti nahradíme hustotou pravděpodobnosti. Místo dvou náhodných veličin X, Y můžeme uvažovat n náhodných veličin:
Náhodné veličiny $X_1, \ldots, X_n$ jsou stochasticky nezávislé, když platí:

$\forall(x_1,\ldots,x_n)\in R^n : \pi(x_1,\ldots,x_n)=\pi_1(x_1)\cdot\ldots\cdot\pi_n(x_n)$ v diskrétním případě,

$\forall(x_1,\ldots,x_n)\in R^n : \varphi(x_1,\ldots,x_n)=\varphi_1(x_1)\cdot\ldots\cdot\varphi_n(x_n)$ ve spojitém případě,

$\forall(x_1,\ldots,x_n)\in R^n : \Phi(x_1,\ldots,x_n)=\Phi_1(x_1)\cdot\ldots\cdot\Phi_n(x_n)$ v obecném případě.

Stochasticky nezávislé náhodné veličiny

Definice:

a) Obecný případ: Řekneme, že náhodné veličiny $X_1, \ldots, X_n$ s marginálními distribučními funkcemi $\Phi_1(x_1), \ldots, \Phi_n(x_n)$ a simultánní distribuční funkcí $\Phi(x_1, \ldots, x_n)$ jsou **stochasticky nezávislé**, právě když

$\forall (x_1, \ldots, x_n) \in R^n : \Phi(x_1, \ldots, x_n) = \Phi_1(x_1) \ldots \Phi_n(x_n)$.

b) Diskrétní případ: Řekneme, že diskrétní náhodné veličiny $X_1, \ldots, X_n$ s marginálními pravděpodobnostními funkcemi $\pi_1(x_1), \ldots, \pi_n(x_n)$ a simultánní pravděpodobnostní funkcí $\pi(x_1, \ldots, x_n)$ jsou **stochasticky nezávislé**, právě když

$\forall (x_1, \ldots, x_n) \in R^n : \pi(x_1, \ldots, x_n) = \pi_1(x_1) \ldots \pi_n(x_n)$.

c) Spojitý případ: Řekneme, že spojité náhodné veličiny $X_1, \ldots, X_n$ s marginálními hustotami $\varphi_1(x_1), \ldots, \varphi_n(x_n)$ a simultánní hustotou $\varphi(x_1, \ldots, x_n)$ jsou **stochasticky nezávislé**, právě když

$\forall (x_1, \ldots, x_n) \in R^n : \varphi(x_1, \ldots, x_n) = \varphi_1(x_1) \ldots \varphi_n(x_n)$ s případnou výjimkou na množině bodů neovlivňujících integraci.

Definice:

Řekneme, že posloupnost $\{X_n\}_{n=1}^{\infty}$ je **posloupnost stochasticky nezávislých náhodných veličin**, právě když pro všechna přirozená n jsou stochasticky nezávislé náhodné veličiny $X_1, \ldots, X_n$.

Příklad

Příklad: Diskrétní náhodný vektor (X_1, X_2) má simultánní pravděpodobnostní funkci $\pi(x_1, x_2)$ danou hodnotami: $\pi(0,0) = \pi(0,2) = \pi(1,1) = \pi(2,0) = \pi(2,2) = 0$, $\pi(0,1) = \pi(1,0) = \pi(1,2) = \pi(2,1) = 0{,}25$. Jsou náhodné veličiny X_1, X_2 stochasticky nezávislé?

Řešení:

Sestavíme kontingenční tabulku, v níž budou hodnoty simultánní pravděpodobnostní funkce a obou marginálních pravděpodobnostních funkcí.

x_1	x_2			$\pi_1(x_1)$
	0	1	2	
0	0	0,25	0	0,25
1	0,25	0	0,25	0,5
2	0	0,25	0	0,25
$\pi_2(x_2)$	0,25	0,5	0,25	1

Ověříme splnění multiplikativního vztahu $\forall (x_1, x_2) \in R^2$: $\pi(x_1, x_2) = \pi_1(x_1)\, \pi_2(x_2)$. Již pro $x_1 = 0$, $x_2 = 0$ vztah splněn není, protože $\pi(0,0) = 0$, avšak $\pi_1(0) = 0{,}25$ a $\pi_2(0) = 0{,}25$. Veličiny X_1, X_2 tedy nejsou stochasticky nezávislé.

Příklad

Příklad:
Nechť spojitý vektor (X_1, X_2) má simultánní hustotu pravděpodobnosti

$\varphi(x_1, x_2) = \begin{cases} 24\, x_1^{\,2} x_2 (1 - x_1) \text{ pro } 0 \le x_1 < 1, 0 \le x_2 < 1 \\ 0 \text{ jinak} \end{cases}$. Dokažte, že náhodné veličiny X_1, X_2 jsou stochasticky nezávislé.

Řešení:
Vypočítáme obě marginální hustoty a ověříme platnost multiplikativního vztahu
$\forall (x_1, ..., x_n) \in R^n$: $\varphi(x_1, ..., x_n) = \varphi_1(x_1) \dots \varphi_n(x_n)$ s případnou výjimkou na množině bodů neovlivňujících integraci.

$$\varphi_1(x_1) = \int_0^1 24x_1^{\,2} x_2 (1 - x_1)\, dx_2 = 24x_1^{\,2} (1-x_1) \left[\frac{x_2^{\,2}}{2} \right]_0^1 = 12x_1^{\,2} (1-x_1) \text{ pro } 0 \le x_1 < 1,$$

$\varphi_1(x_1) = 0$ jinak.

$$\varphi_2(x_2) = \int_0^1 24x_1^{\,2} x_2 (1 - x_1)\, dx_1 = 24x_2 \left[\frac{x_1^{\,3}}{3} - \frac{x_1^{\,4}}{4} \right]_0^1 = 2x_2 \text{ pro } 0 \le x_2 < 1,$$

$\varphi_2(x_2) = 0$ jinak.
Vidíme, že multiplikativní vztah je splněn, tudíž veličiny X_1, X_2 jsou stochasticky nezávislé.

Stochasticky nezávislé náhodné vektory

Definice:

Nechť $(\Omega, \mathcal{A}, P)$ je pravděpodobnostní prostor, $\mathbf{X}_1 = (X_{11}, \ldots, X_{p_1 1}), \ldots, \mathbf{X}_n = (X_{1n}, \ldots, X_{p_n n})$ náhodné vektory definované na $(\Omega, \mathcal{A}, P)$. Řekneme, že tyto náhodné vektory jsou **stochasticky nezávislé**, právě když každá složka náhodného vektoru $\mathbf{X}_i$ je stochasticky nezávislá se všemi složkami náhodného vektoru $\mathbf{X}_k$ pro $\forall i \neq k$.

Věta:

Nechť $X_1, \ldots, X_n$ jsou stochasticky nezávislé náhodné veličiny, $g_1, \ldots, g_n$ borelovské funkce. Pak transformované náhodné veličiny $Y_1 = g_1(X_1), \ldots, Y_n = g_n(X_n)$ jsou opět stochasticky nezávislé náhodné veličiny.

(Tvrzení lze zobecnit i pro transformované náhodné vektory.)

Příklad

Příklad:
Nechť $X_1, \ldots, X_n$ jsou stochasticky nezávislé náhodné veličiny s distribučními funkcemi $\Phi_1(x_1), \ldots, \Phi_n(x_n)$. Zavedeme transformované náhodné veličiny $Y = \max\{X_1, \ldots, X_n\}$, $Z = \min\{X_1, \ldots, X_n\}$. Odvoďte jejich distribuční funkce $\Phi_{max}(y), \Phi_{min}(z)$.

Řešení:

$$\Phi_{max}(y) = P(Y \le y) = P(\max\{X_1, \ldots, X_n\} \le y) = P(X_1 \le y \wedge \ldots \wedge X_n \le y) = P(X_1 \le y) \cdot \ldots \cdot P(X_n \le y) = \Phi_1(y) \cdot \ldots \cdot \Phi_n(y)$$

$$\Phi_{min}(z) = P(Z \le z) = P(\min\{X_1, \ldots, X_n\} \le z) = P(X_1 \le z \vee \ldots \vee X_n \le z) = 1 - P(X_1 > z \wedge \ldots \wedge X_n > z) = 1 - P(X_1 > z) \cdot \ldots \cdot P(X_n > z) =$$
$$= 1 - [1 - P(X_1 \le z)] \cdot \ldots \cdot [1 - P(X_n \le z)] = 1 - [1 - \Phi_1(z)] \cdot \ldots \cdot [1 - \Phi_n(z)]$$

Příklad

Příklad:
Na automatické lince jsou láhve plněny mlékem. Je známo, že množství mléka v láhvích kolísá od 0,98 l do 1,02 l. V tomto intervalu považujeme každé množství mléka za stejně možné. Za 1 s se naplní 3 láhve. Jaká je pravděpodobnost, že
a) nejméně naplněná láhev obsahuje aspoň 1 l mléka,
b) v nejvíce naplněné láhvi není víc než 1,01 l mléka?

Řešení:
Náhodná veličina X_i udává množství mléka v i-té láhvi, i = 1, 2, 3. Je to spojitá náhodná veličina, její hustota pravděpodobnosti je konstantní na intervalu (0,98; 1 ,02). Z podmínky normovanosti S2 dostaneme, že hustota

$$\varphi(x)=\begin{cases}\frac{1}{40} & \text{pro } x\in(980,\,1020)\\ 0 & \text{jinak}\end{cases}$$

. Distribuční funkce:

$$\Phi(x)=\begin{cases}0 & \text{pro } x\in(-\infty,980\rangle\\ \int_{980}^{x}\frac{1}{40}\,dt=\frac{1}{40}[x]_{980}^{x}=\frac{x-980}{40} & \text{pro } x\in(980,1020)\\ 1 & \text{pro } x\in\langle 1020,\infty)\end{cases}$$

ad a) $P(Z\geq 1000)=1-P(Z<1000)=1-\Phi_{min}(1000)=[1-\Phi(1000)]^3=\left[1-\frac{1000-980}{40}\right]^3=\left(\frac{1}{2}\right)^3=\frac{1}{8}=0{,}125$

ad b) $P(Y\leq 1010)=\Phi_{max}(1010)=[\Phi(1010)]^3=\left(\frac{3}{4}\right)^3=\frac{27}{64}=0{,}42$

Rozložení transformovaných NV.

Motivace: Máme náhodnou veličinu X s distribuční funkcí $\Phi(x)$ (resp. pravděpodobnostní funkcí $\pi(x)$ v diskrétním případě resp. hustotou $\varphi(x)$ ve spojitém případě) a borelovskou funkci g: R → R. Zavedeme transformovanou náhodnou veličinu Y = g(X) a hledáme její distribuční funkci $\Phi_*(y)$ (resp. pravděpodobnostní funkcí $\pi_*(y)$ v diskrétním případě resp. hustotu $\varphi_*(y)$ ve spojitém případě).

Věta: Nechť X je diskrétní náhodná veličina s pravděpodobnostní funkcí $\pi(x)$ a g je borelovská ryze monotónní funkce, tedy v oblasti $C \subseteq R$ existuje inverzní funkce $g^{-1} = \tau$. Pak pravděpodobnostní funkce $\pi_*(y)$ transformované náhodné veličiny Y = g(X) má tvar:

$$\pi_*(y) = \begin{cases} \pi(\tau(y)) \text{pro} y \in C \\ 0 \text{jinak} \end{cases}.$$

Důkaz: $\pi_*(y) = P(Y = y) = P(g(X) = y) = P(X = g^{-1}(y)) = P(X = \tau(y)) = \pi(\tau(y))$ pro $y \in C$, $\pi_*(y) = 0$ jinak.

Příklad: $X \sim \pi(x)$, $Y = a + bX$, $\pi_*(y) = ?$

Řešení:

a) $b \neq 0$: $\pi_*(y) = P(Y = y) = P(a + bX = y) = P\left(X = \frac{y-a}{b}\right) = \pi\left(\frac{y-a}{b}\right)$

b) $b = 0$: $Y = a \Rightarrow Y \sim Dg(a)$

Rozložení transformované spojité náhodné veličiny

Věta: Nechť X je spojitá náhodná veličina s hustotou φ(x) a g je borelovská ryze monotónní funkce se spojitou a nenulovou derivací v R, tedy v oblasti C ⊆ R existuje inverzní funkce $g^{-1} = \tau$ se spojitou a nenulovou derivací. Pak hustota $\varphi_*(y)$ transformované náhodné veličiny Y = g(X) má tvar:

$$\varphi_*(y) = \begin{cases} \varphi(\tau(y))|\tau'(y)| \text{pro} y \in C \\ 0 \text{ jinak} \end{cases}.$$

Důkaz:

$$\Phi_*(y) = P(Y \le y) = P(g(X) \le y) = \begin{cases} P(X \le \tau(y)) = \Phi(\tau(y)) \text{ prog rostoucí} \\ P(X \ge \tau(y)) = 1 - \Phi(\tau(y)) \text{ prog klesající} \end{cases}$$

$$\varphi_*(y) = \frac{d\Phi_*(y)}{dy} = \left.\begin{cases} \varphi(\tau(y))\tau'(y) \text{ prog rostoucí} \\ -\varphi(\tau(y))\tau'(y) \text{ prog klesající} \end{cases}\right\} = \varphi(\tau(y))|\tau'(y)| \text{ pro } y \in C,\ \varphi_*(y) = 0 \text{ jinak}$$

Příklad: $X \sim Rs\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$, Y = tg X, $\varphi_*(y)$ = ?

Řešení:

$$\varphi(x) = \begin{cases} \frac{1}{\pi} \text{ pro } x \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right), \\ 0 \text{ jinak} \end{cases}$$

$$\Phi_*(y) = P(Y \le y) = P(\text{tg}(X) \le y) = P(X \le \text{arctg}(y)) = \Phi(\text{arctg}(y))$$

$$\varphi_*(y) = \frac{d\Phi_*(y)}{dy} = \varphi(\text{arctg}(y))\frac{1}{1+y^2} = \frac{1}{\pi(1+y^2)}$$

Říkáme, že Y má Cauchyovo rozložení, píšeme Y ~ t(1).

Nemonotónní transformace

Věta: Není-li transformační funkce g ryze monotónní, pak mezi X a Y neexistuje vzájemně jednoznačný vztah. Distribuční funkce transformované náhodné veličiny Y se vypočte podle vzorce: $\Phi_*(y) = P(X \in \Delta_1) + P(X \in \Delta_2) + \ldots$, kde $\Delta_1, \Delta_2, \ldots$ jsou ty intervaly, pro které $Y \le y$.

Příklad: $X \sim N(0,1)$, $Y = X^2$, $\varphi_*(y) = ?$

Řešení:

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}},$$

$$\Phi_*(y) = P(Y \le y) = P(X^2 \le y) = P(-\sqrt{y} \le X \le \sqrt{y}) = \Phi(\sqrt{y}) - \Phi(-\sqrt{y}) = \Phi(\sqrt{y}) - [1 - \Phi(\sqrt{y})] =$$

$$= 2\Phi(\sqrt{y}) - 1$$

$$\varphi_*(y) = \frac{d\Phi_*(y)}{dy} = 2\varphi(\sqrt{y})\frac{1}{2\sqrt{y}} = \frac{1}{\sqrt{2\pi}} e^{-\frac{y}{2}} \frac{1}{\sqrt{y}} = \frac{1}{\sqrt{2\pi y}} e^{-\frac{y}{2}}$$ pro $y > 0$, $\varphi_*(y) = 0$ jinak

Y má χ^2 rozložení s jedním stupněm volnosti, píšeme $Y \sim \chi^2(1)$.

$X \sim \chi^2(k)$:

$$\varphi(x,k) = \begin{cases} \frac{1}{2^{k/2} \cdot \Gamma(k/2)} \cdot x^{k/2-1} \cdot e^{-x/2} & x > 0 \\ 0 & jinak \end{cases} \qquad \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$$

Rozložení transformovaného náhodného vektoru

Věta (transformace náhodného vektoru $\mathbf{X} = (X_1, \ldots, X_n)$ na skalární náhodnou veličinu $Y = g(X_1, \ldots, X_n)$)

a) Diskrétní případ: $\mathbf{X} = (X_1, \ldots, X_n) \sim \pi(x_1, \ldots, x_n)$, $g: R^n \to R$ je borelovská funkce $\Rightarrow$

$$Y = g(x_1, \ldots, x_n) \sim \pi_*(y_1, \ldots, y_n) = \sum_{(x_1,\ldots x_n)\in S(y)} \ldots \sum \pi(x_1,\ldots,x_n), \text{ kde}$$

$$S(y) = \{(x_1,\ldots,x_n) \in R^n; g(x_1,\ldots,x_n) = y\}$$

b) Spojitý případ: $\mathbf{X} = (X_1, \ldots, X_n) \sim \varphi(x_1, \ldots, x_n)$, $g: R^n \to R$ je borelovská funkce $\Rightarrow$

$$Y = g(x_1, \ldots, x_n) \sim \varphi_*(y_1, \ldots, y_n) = \frac{d}{dy}\int_{S(y)} \ldots \int \varphi(x_1,\ldots,x_n)dx_1 \ldots dx_n \text{ , kde}$$

$$S(y) = \{(x_1,\ldots,x_n) \in R^n; g(x_1,\ldots,x_n) \le y\}$$

Věta o konvoluci

Věta (věta o konvoluci)

a) Diskrétní případ: X_1, X_2 jsou stochasticky nezávislé náhodné veličiny, $X_i \sim \pi_i(x_i)$, $i = 1, 2 \Rightarrow$

$$Y = X_1 + X_2 \sim \pi_*(y) = \sum_{x_1=-\infty}^{\infty} \pi_1(x_1)\pi_2(y - x_1) = \sum_{x_2=-\infty}^{\infty} \pi_1(y - x_2)\pi_2(x_2)$$

$\pi_*(y)$ se nazývá konvoluce funkcí $\pi_1(x_1)$, $\pi_2(x_2)$.

b) Spojitý případ: X_1, X_2 jsou stochasticky nezávislé náhodné veličiny, $X_i \sim \varphi_i(x_i)$, $i = 1, 2 \Rightarrow$

$$Y = X_1 + X_2 \sim \varphi_*(y) = \int_{-\infty}^{\infty} \varphi_1(x_1)\varphi_2(y - x_1)dx_1 = \int_{-\infty}^{\infty} \varphi_1(y - x_2)\varphi_2(x_2)dx_2$$

$\varphi_*(y)$ se nazývá konvoluce funkcí $\varphi_1(x_1)$, $\varphi_2(x_2)$.

Příklad

Příklad: X_1, X_2 jsou stochasticky nezávislé náhodné veličiny, $X_i \sim Po(\lambda_i)$, $i = 1, 2$, $Y = X_1 + X_2$, $\pi_*(y) = ?$

Řešení:

$$\pi_i(x_i) = \begin{cases} \dfrac{\lambda_i^{x_i}}{x_i!} e^{-\lambda_i} & \text{pro } x_i = 0, 1, \ldots \\ 0 & \text{jinak} \end{cases}$$

$$\pi_*(y) = \sum_{x_1=-\infty}^{\infty} \pi_1(x_1)\pi_2(y - x_1) = |x_1 \geq 0, y - x_1 \geq 0 \Rightarrow 0 \leq x_1 \leq y| = \sum_{x_1=0}^{y} \pi_1(x_1)\pi_2(y - x_1) =$$

$$= \sum_{x_1=0}^{y} \frac{\lambda_1^{x_1}}{x_1!} e^{-\lambda_1} \frac{\lambda_2^{y-x_1}}{(y - x_1)!} e^{-\lambda_2} = e^{-(\lambda_1+\lambda_2)} \frac{1}{y!} \sum_{x_1=0}^{y} \binom{y}{x_1} \lambda_1^{x_1} \lambda_2^{y-x_1} = \frac{(\lambda_1 + \lambda_2)^y}{y!} e^{-(\lambda_1+\lambda_2)} \text{ pro } y = 0, 1, \ldots$$

$\pi_*(y) = 0$ jinak.

Vidíme, že $Y \sim Po(\lambda_1 + \lambda_2)$.

Zobecnění: $X_1, \ldots, X_n$ jsou stochasticky nezávislé náhodné veličiny, $X_i \sim Po(\lambda_i)$, $i = 1, 2, \ldots, n$ $\Rightarrow$
$Y = X_1 + \ldots + X_n \sim Po(\lambda_1 + \ldots + \lambda_n)$.

Lineární transformace náhodného vektoru

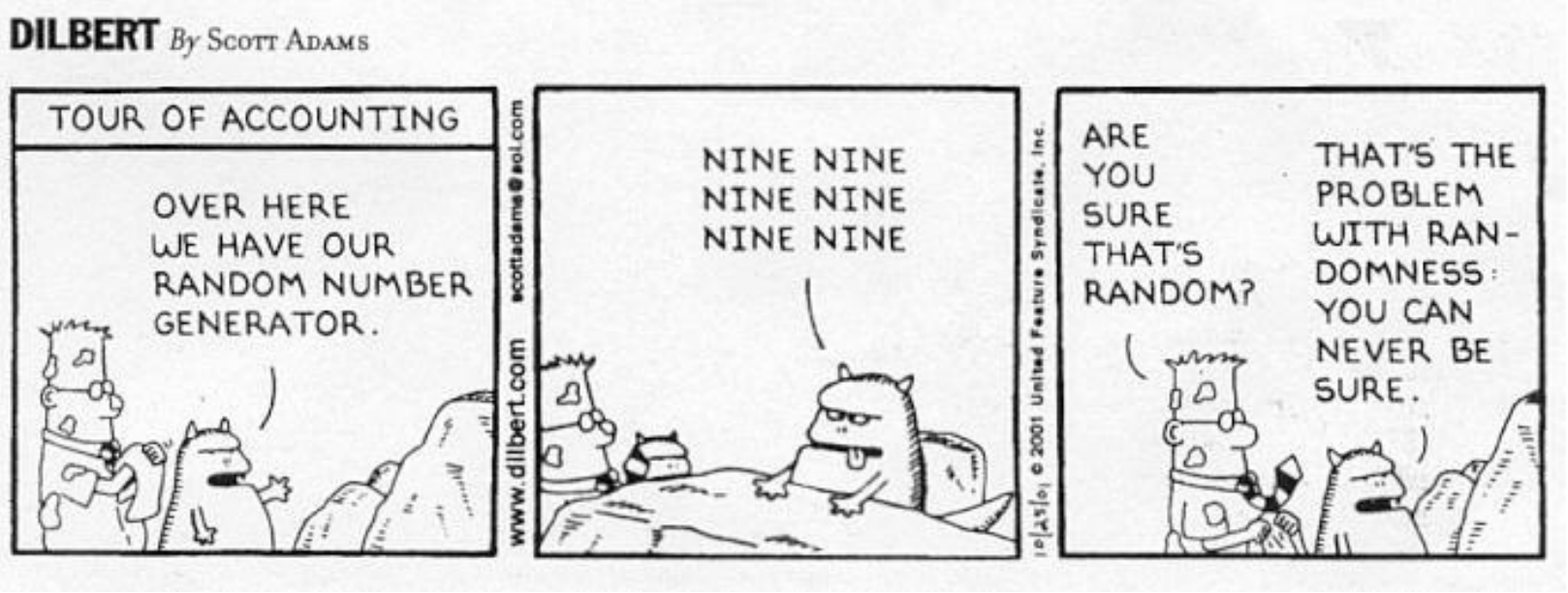
Věta (lineární transformace n - rozměrného náhodného vektoru)
Nechť $\mathbf{X} = (X_1, \ldots, X_n)'$ je náhodný vektor, $\mathbf{a} = (a_1, \ldots, a_n)'$ je reálný vektor a $\mathbf{B} = (b_{ij})_{i,j=1,\ldots,n}$ je reálná čtvercová pozitivně definitní matice (tj. $\forall \mathbf{x} \in \mathbf{R}^n$ je kvadratická funkce $\mathbf{x}'\mathbf{B}\mathbf{x} > 0$). Pak pro rozložení pravděpodobností transformovaného náhodného vektoru $\mathbf{Y} = \mathbf{a} + \mathbf{B}\mathbf{X}$ platí:

a) Diskrétní případ: $\pi_*(\mathbf{y}) = \pi(\mathbf{B}^{-1}(\mathbf{y} - \mathbf{a}))$

b) Spojitý případ: $$\varphi_*(y) = \det(B)^{-1}\varphi\left(B^{-1}(y-a)\right)$$

Věta: Nech ť náhodný vektor $\mathbf{X} = (X_1, \ldots, X_n)'$ má n -rozměrné normální rozložení $N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Položme $\mathbf{Y} = \mathbf{a} + \mathbf{B}\mathbf{X}$. Pak $\mathbf{Y} \sim N_n(\mathbf{a} + \mathbf{B}\boldsymbol{\mu}, \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}')$.

Generování náhodných čísel



Random Number Functions

- SAS can generate random observations from discrete and continuous distributions.

Binomial (n,p)	ranbin(seed,n,p)
Multinomické (p1,...pk)	rantbl(seed, p1,...,pk)
Exponential (1)	ranexp(seed)
Standard Normal (μ=0; sigmaSquared=1)	rannor(seed)
Poisson (mean > 0)	ranpoi(seed, mean)
Uniform (interval (0,1))	ranuni(seed)
Cauchy (0,1)	rancau(seed)
Gamma(a)	rangam(seed,a)

Seeds

- A **SEED -** is a number used by the random number generator to start the algorithm
- They can be any **POSITIVE NUMBER** or **Zero**
  - **o seed** = a different series of numbers each time you run the program.
  - **Any positive seed** = a repeatable series of numbers each time you run the program.

Funkce RAND

Syntax

RAND (*dist*, *parm-1*,...,*parm-k*)

Required Arguments

dist

is a character constant, variable, or expression that identifies the distribution. Valid distributions are as follows:

Distribution	Argument
Bernoulli	BERNOULLI
Beta	BETA
Binomial	BINOMIAL
Cauchy	CAUCHY
Chi-Square	CHISQUARE
Erlang	ERLANG
Exponential	EXPONENTIAL
F	F
Gamma	GAMMA
Geometric	GEOMETRIC
Hypergeometric	HYPERGEOMETRIC
Lognormal	LOGNORMAL
Negative Binomial	NEGBINOMIAL
Normal	NORMAL|GAUSSIAN
Poisson	POISSON
T	T
Tabled	TABLE
Triangular	TRIANGLE
Uniform	UNIFORM
Weibull	WEIBULL

Note: Except for T and F, you can minimally identify any distribution by its first four characters.

parm-1,...,parm-k

are *shape*, *location*, or *scale* parameters appropriate for the specific distribution.

Více viz:
http://support.sas.com/documentation/cdl/en/lrdict/64316/HTML/default/viewer.htm#a001466748.htm

Funkce RAND

Details

Generating Random Numbers
The RAND function generates random numbers from various continuous and discrete distributions. Wherever possible, the simplest form of the distribution is used.
The RAND function uses the Mersenne-Twister random number generator (RNG) that was developed by Matsumoto and Nishimura (1998). The random number generator has a very long period ($2^{19937} - 1$) and very good statistical properties. The period is a Mersenne prime, which contributes to the naming of the RNG. The algorithm is a twisted generalized feedback shift register (TGFSR) that explains the latter part of the name. The TGFSR gives the RNG a very high order of equidistribution (623-dimensional with 32-bit accuracy), which means that there is a very small correlation between successive vectors of 623 pseudo-random numbers.
The RAND function is started with a single seed. However, the state of the process cannot be captured by a single seed. You cannot stop and restart the generator from its stopping point.

Reproducing a Random Number Stream
If you want to create reproducible streams of random numbers, then use the CALL STREAMINIT routine to specify a seed value for random number generation. Use the CALL STREAMINIT routine once per DATA step before any invocation of the RAND function. If you omit the call to the CALL STREAMINIT routine (or if you specify a non-positive seed value in the CALL STREAMINIT routine), then RAND uses a call to the system clock to seed itself.

Duplicate Values in the Mersenne-Twister RNG Algorithm
The Mersenne-Twister RNG algorithm has an extremely long period, but this does not imply that large random samples are devoid of duplicate values. The RAND function returns at most 2^{32} distinct values. In a random uniform sample of size 10^5, the chance of drawing at least one duplicate is greater than 50%. The expected number of duplicates in a random uniform sample of size M is approximately $M^2/2^{33}$ when M is much less than 2^{32}. For example, you should expect about 115 duplicates in a random uniform sample of size $M=10^6$. These results are consequences of the famous “birthday matching problem” in probability theory.

Funkce RAND

Bernoulli Distribution

x = **RAND**('BERNOULLI',*p*)

Arguments

x

is an observation from the distribution with the following probability density function:

$$f(x) = \begin{cases} 1 & p = 0, x = 0 \\ p^x (1-p)^{1-x} & 0 < p < 1, x = 0, 1 \\ 1 & p = 1, x = 1 \end{cases}$$

Range: *x* = 0, 1

p

is a numeric probability of success.

Range: $0 \le p \le 1$

Beta Distribution

x = **RAND**('BETA',*a*,*b*)

Arguments

x

is an observation from the distribution with the following probability density function:

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1} (1-x)^{b-1}$$

Range: $0 < x < 1$

a

is a numeric shape parameter.

Range: $a > 0$

b

is a numeric shape parameter.

Range: $b > 0$

Funkce RAND

Binomial Distribution

x = **RAND**('BINOMIAL',*p*,*n*)

Arguments

x

is an integer observation from the distribution with the following probability density function:

$$f(x) = \begin{cases} 1 & p = 0,\ x = 0 \\ \binom{n}{x} p^x (1-p)^{n-x} & 0 < p < 1,\ x = 0, \ldots, n \\ 1 & p = 1,\ x = n \end{cases}$$

Range: $x = 0, 1, \ldots, n$

p

is a numeric probability of success.

Range: $0 \le p \le 1$

n

is an integer parameter that counts the number of independent Bernoulli trials.

Range: $n = 1, 2, \ldots$

Cauchy Distribution

x = **RAND**('CAUCHY')

Arguments

x

is an observation from the distribution with the following probability density function:

$$f(x) = \frac{1}{\pi\left(1+x^2\right)}$$

Range: $-\infty < x < \infty$

Funkce RAND

Chi-Square Distribution

x = **RAND**('CHISQUARE',*df*)

Arguments

x

is an observation from the distribution with the following probability density function:

$$f(x) = \frac{2^{-df/2}}{\Gamma\left(\frac{df}{2}\right)} x^{df/2-1} e^{-x/2}$$

Range: $x > 0$

df

is a numeric degrees of freedom parameter.

Range: $df > 0$

Erlang Distribution

x = **RAND**('ERLANG',*a*)

Arguments

x

is an observation from the distribution with the following probability density function:

$$f(x) = \frac{1}{\Gamma(a)} x^{a-1} e^{-x}$$

Range: $x > 0$

a

is an integer numeric shape parameter.

Range: $a = 1, 2, \ldots$

Funkce RAND

Exponential Distribution

x = **RAND**('EXPONENTIAL')

Arguments

x

is an observation from the distribution with the following probability density function:

$$f(x) = e^{-x}$$

Range: $x > 0$

F Distribution

x = **RAND**('F',*n*, *d*)

Arguments

x

is an observation from the distribution with the following probability density function:

$$f(x) = \frac{\Gamma\left(\frac{n+d}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{d}{2}\right)} \frac{n^{n/2} d^{d/2} x^{n/2-1}}{(d+nx)^{(n+d)/2}}$$

Range: $x > 0$

n

is a numeric numerator degrees of freedom parameter.

Range: $n > 0$

d

is a numeric denominator degrees of freedom parameter.

Range: $d > 0$

Funkce RAND

Geometric Distribution

x = **RAND**('GEOMETRIC',*p*)

Arguments

x

is an integer count that denotes the number of trials that are needed to obtain one success. *X* is an integer observation from the distribution with the following probability density function:

$$f(x) = \begin{cases} (1-p)^{x-1}p & 0 < p < 1,\ x = 1, 2, \ldots \\ 1 & p = 1,\ x = 1 \end{cases}$$

Range: *x* = 1, 2, ...

p

is a numeric probability of success.

Range: $0 < p \leq 1$

Gamma Distribution

x = **RAND**('GAMMA',*a*)

Arguments

x

is an observation from the distribution with the following probability density function:

$$f(x) = \frac{1}{\Gamma(a)} x^{a-1} e^{-x}$$

Range: *x* > 0

a

is a numeric shape parameter.

Range: *a* > 0

Funkce RAND

Hypergeometric Distribution

x = **RAND**('HYPER',*N*,*R*,*n*)

Arguments

x

is an integer observation from the distribution with the following probability density function:

$$f(x) = \frac{\binom{R}{x}\binom{N-R}{n-x}}{\binom{N}{n}}$$

Range: $x = \max(0, (n - (N - R))), \ldots, \min(n, R)$

N

is an integer population size parameter.

Range: $N = 1, 2, \ldots$

R

is an integer number of items in the category of interest.

Range: $R = 0, 1, \ldots, N$

n

is an integer sample size parameter.

Range: $n = 1, 2, \ldots, N$

Lognormal Distribution

x = **RAND**('LOGNORMAL')

Arguments

x

is an observation from the distribution with the following probability density function:

$$f(x) = \frac{e^{-\ln^2(x)/2}}{x\sqrt{2\pi}}$$

Range: $x > 0$

The hypergeometric distribution is a mathematical formalization of an experiment in which you draw *n* balls from an urn that contains *N* balls, *R* of which are red. The hypergeometric distribution is the distribution of the number of red balls in the sample of *n*.

Funkce RAND

Negative Binomial Distribution

x = **RAND**('NEGBINOMIAL',*p*,*k*)

Arguments

x

is an integer observation from the distribution with the following probability density function:

$$f(x) = \begin{cases} \binom{x+k-1}{k-1}(1-p)^x p^k & 0 < p < 1,\ x = 0, 1, \ldots \\ 1 & p = 1,\ x = 0 \end{cases}$$

Range: *x* = 0, 1, ...

k

is an integer parameter that is the number of successes. However, non-integer *k* values are allowed as well.

Range: *k* = 1, 2, ...

p

is a numeric probability of success.

Range: $0 < p \leq 1$

The negative binomial distribution is the distribution of the number of failures before *k* successes occur in sequential independent trials, all with the same probability of success, *p*.

Funkce RAND

Normal Distribution

x = **RAND**('NORMAL',<,*θ*,*λ*>)

Arguments

x

is an observation from the normal distribution with a mean of θ and a standard deviation of λ that has the following probability density function:

$$f(x) = \frac{1}{\lambda\sqrt{2\pi}} exp\left(-\frac{(x-\theta)^2}{2\lambda^2}\right)$$

Range: $-\infty < x < \infty$

θ

is the mean parameter.

Default: 0

λ

is the standard deviation parameter.

Default: 1

Range: λ > 0

Poisson Distribution

x = **RAND**('POISSON',*m*)

Arguments

x

is an integer observation from the distribution with the following probability density function:

$$f(x) = \frac{m^x e^{-m}}{x!}$$

Range: *x* = 0, 1, ...

m

is a numeric mean parameter.

Range: *m* > 0

Funkce RAND

Tabled Distribution

x = **RAND**('TABLE',*p1*,*p2*, ...)

Arguments

x

is an integer observation from one of the following distributions:

If $\sum_{i=1}^{n} p_i < 1$, then *x* is an observation from this probability density function:

$$f(i) = p_i, \quad i = 1, 2, \ldots, n$$

and

$$f(n+1) = 1 - \sum_{i=1}^{n} p_i$$

If for some index $\sum_{i=1}^{n} p_i \geq 1$, then *x* is an observation from this probability density function:

$$f(i) = p_i, \quad i = 1, 2, \ldots, n-1$$

and

$$f(n) = 1 - \sum_{i=1}^{n-1} p_i$$

p1, p2, ...

are numeric probability values.

Range: $0 \leq p1, p2, \ldots \leq 1$

Restriction: The maximum number of probability parameters depends on your operating environment, but the maximum number of parameters is at least 32,767.

The tabled distribution takes on the values 1, 2, ..., *n* with specified probabilities.

Note: By using the FORMAT statement, you can map the set {1, 2, ..., *n*} to any set of *n* or fewer elements.

T Distribution

x = **RAND**('T',*df*)

Arguments

x

is an observation from the distribution with the following probability density function:

$$f(x) = \frac{\Gamma\left(\frac{df+1}{2}\right)}{\sqrt{df\,\pi}\,\Gamma\left(\frac{df}{2}\right)}\left(1 + \frac{x^2}{df}\right)^{-\frac{df+1}{2}}$$

Range: $-\infty < x < \infty$

df

is a numeric degrees of freedom parameter.

Range: $df > 0$

Funkce RAND

Triangular Distribution

x = **RAND**('TRIANGLE',*h*)

Arguments

x

is an observation from the distribution with the following probability density function:

$$f(x) = \begin{cases} \frac{2x}{h} & 0 \le x \le h \\ \frac{2(1-x)}{1-h} & h < x \le 1 \end{cases}$$

In this equation, $0 \le h \le 1$.

Range: $0 \le x \le 1$

Note: The distribution can be easily shifted and scaled.

h

is the horizontal location of the peak of the triangle.

Range: $0 \le h \le 1$

Funkce RAND

Uniform Distribution

x = **RAND**('UNIFORM')

Arguments

x

is an observation from the distribution with the following probability density function:

$$f(x) = 1$$

Range: $0 < x < 1$

The uniform random number generator that the RAND function uses is the Mersenne-Twister (Matsumoto and Nishimura 1998). This generator has a period of $2^{19937} - 1$ and 623-dimensional equidistribution up to 32-bit accuracy. This algorithm underlies the generators for the other available distributions in the RAND function.

Weibull Distribution

x = **RAND**('WEIBULL',*a*,*b*)

Arguments

x

is an observation from the distribution with the following probability density function:

$$f(x) = \frac{a}{b^a} x^{a-1} e^{-\left(\frac{x}{b}\right)^a}$$

Range: $x \geq 0$

a

is a numeric shape parameter.

Range: $a > 0$

b

is a numeric scale parameter.

Range: $b > 0$

Generování v Excelu

Suppose you want to model a discrete uniform distribution of demand where the values of 8 through 12 all have the same probability of occurring (uniform, equally likely).
The spreadsheet has a function, =RAND(), that returns a random number between 0 and 1. However, this will result in a continuous uniform distribution.

To create a discrete uniform distribution, use the INT() function. For example:

Values for RAND()	(=INT(8+5*RAND())
0 <=RAND() < 0.2	8
0.2 <=RAND() < 0.4	9
0.4 <=RAND() < 0.6	10
0.6 <=RAND() < 0.8	11
0.8 <=RAND() < 1.0	12

In general, if you want a discrete, uniform distribution of integer values between *x* and *y*, use the formula:

$$\mathbf{INT}(x + (y - x + 1)*\mathbf{RAND}())$$

Generování v Excelu

Generating from the Normal Distribution. The normal distribution plays an important role in many simulation and analytic models. Normality is often assumed.

Consider drawing a random demand from a normal distribution with a mean (m) of 1000 and a standard deviation (s) of 100.

If Z is a unit normal random variable (normally distributed with a mean of 0 and a standard deviation of 1) then m + Zs is a normal random variable with mean m and standard deviation s.

So, we can draw from a unit normal distribution. Excel has a built-in function that can do this:

= NORMINV(RAND() , 1000, 100)

Excel will automatically return a normally distributed random number with mean 1000 and std. dev. 100.

Generování v Excelu (2007)

If you want to generate random numbers in Excel between, say 1 and 10, use the RANDBETWEEN function. This function allows you to specify the range of numbers it is to pick from.

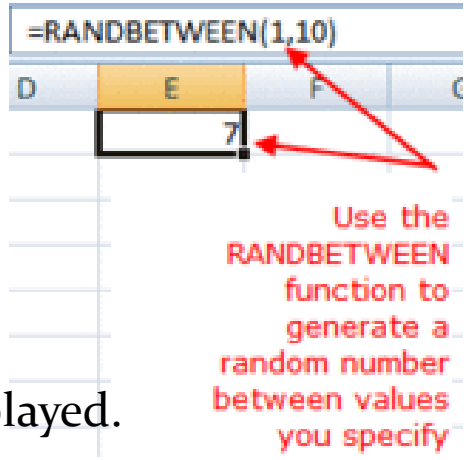
The syntax for the RAND function is:

= **RANDBETWEEN (Bottom , Top)**

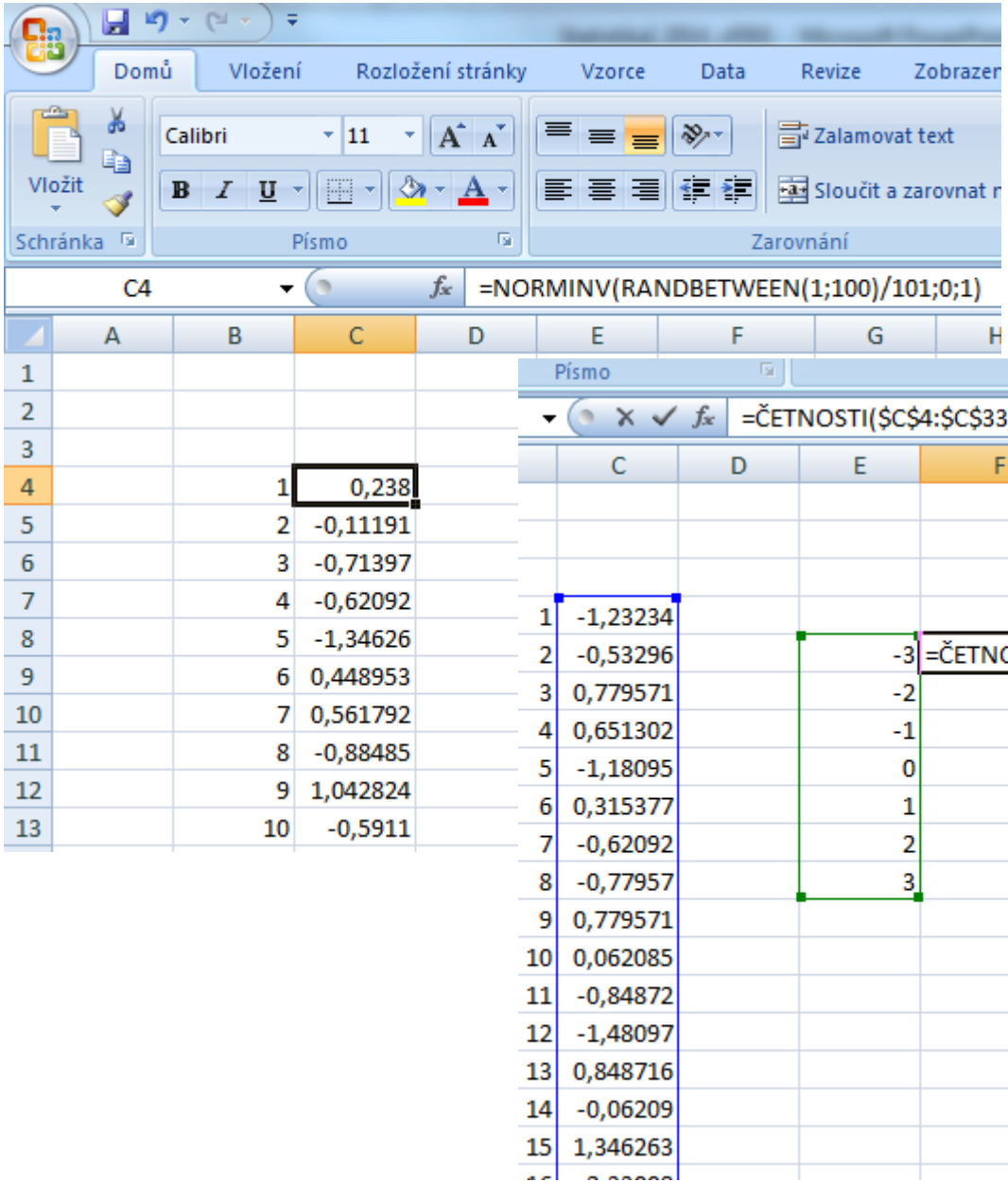
Bottom - the lowest number the function is to use.

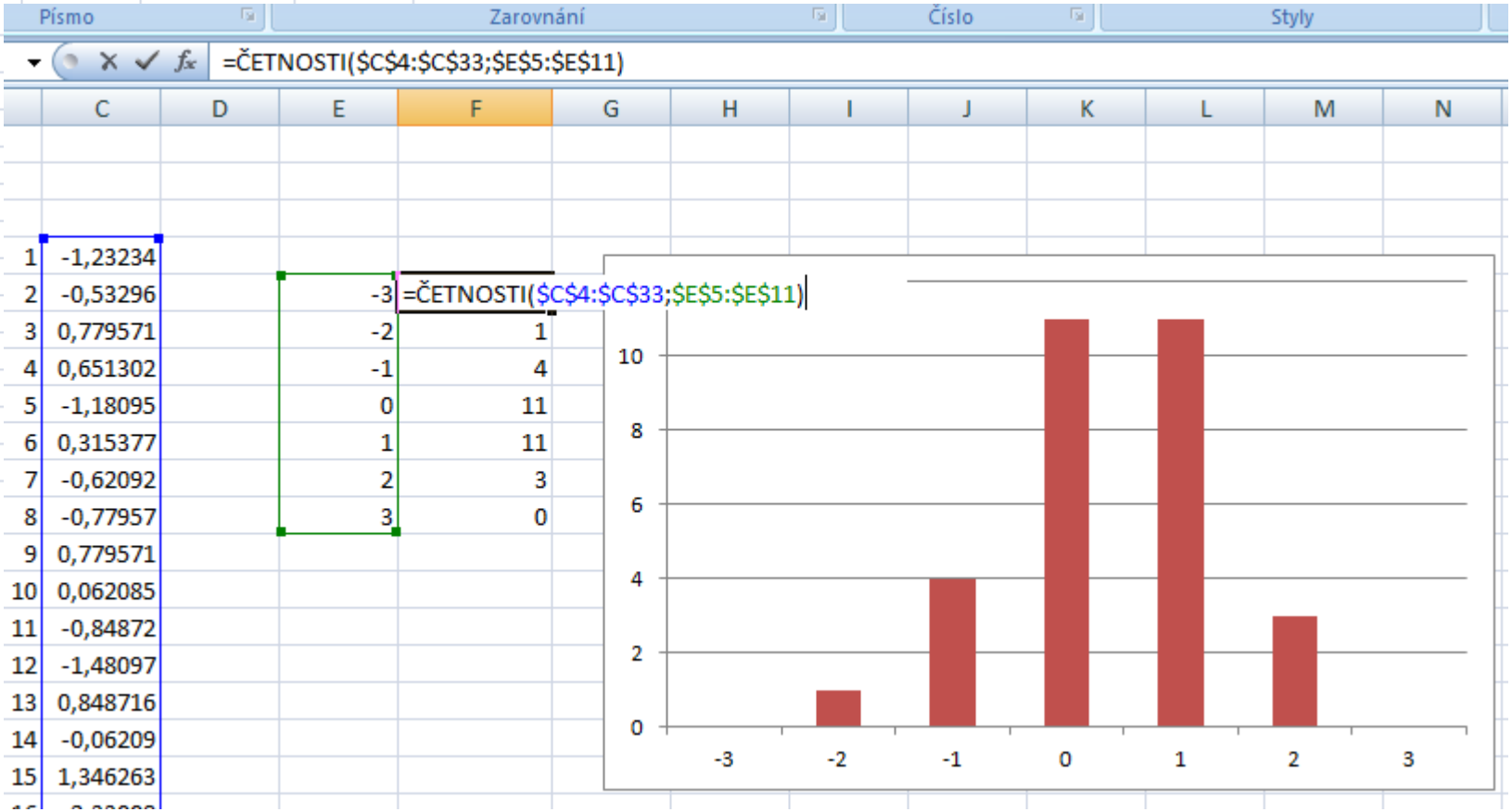
Top - the highest number the function is to use.

Example Using Excel's RANDBETWEEN Function:



1. Click on cell E1 in the spreadsheet - the location where the results will be displayed.
2. Click on the *Formulas* tab of the ribbon menu.
3. Choose **Math & Trig** from the ribbon to open the function drop down list.
4. Click on *RANDBETWEEN* in the list to bring up the function's dialog box.
5. Click on the "Bottom" line in the dialog box.
6. Type the number 1 (one) on this line.
7. Click on the "Top" line in the dialog box.
8. Type the number 10 (ten) on this line.
9. Click OK.
10. A random number between 1 and 10 should appear in cell E1.
11. To generate another random number, press the **F9** key on the keyboard.
12. When you click on cell E1 the complete function = **RANDBETWEEN (1 , 10)** appears in the formula bar above the worksheet.


Domů
Vložení
Rozložení stránky
Vzorce
Data
Revize
Zobrazen
Calibri
11
Zalamovat text
Sloučit a zarovnat
Vložit
Schránka
Písmo
Zarovnání
C4
=NORMINV(RANDBETWEEN(1;100)/101;0;1)
A B C D E F G H
1
2
3
4 1 0,238
5 2 -0,11191
6 3 -0,71397
7 4 -0,62092
8 5 -1,34626
9 6 0,448953
10 7 0,561792
11 8 -0,88485
12 9 1,042824
13 10 -0,5911


Písmo
Zarovnání
Číslo
Styly
=ČETNOSTI($C$4:$C$33;$E$5:$E$11)
C D E F G H I J K L M N
1 -1,23234
2 -0,53296
3 0,779571
4 0,651302
5 -1,18095
6 0,315377
7 -0,62092
8 -0,77957
9 0,779571
10 0,062085
11 -0,84872
12 -1,48097
13 0,848716
14 -0,06209
15 1,346263
-3
-2
-1
0
1
2
3
=ČETNOSTI($C$4:$C$33;$E$5:$E$11)
1
4
11
11
3
0
10
8
6
4
2
0
-3
-2
-1
0
1
2
3

Inverse-Transform Method for Generating Non-$U(0,1)$ Random Numbers

- Let $F(x)$ be distribution function of X
- Define inverse function of F by

$$F^{-1}(y) = \inf\{x : F(x) \geq y\}, 0 \leq y \leq 1.$$

- Generate X by $X = F^{-1}(U)$

- Example: exponential distribution

$$F(x) = 1 - e^{-\lambda x}$$

$$X = F^{-1}(U) = -\frac{1}{\lambda}\log(1-U)$$

10. Číselné charakteristiky NV

Číselné charakteristiky náhodných veličin

Motivace: Doposud jsme pracovali s funkcionálními charakteristikami náhodných veličin (např. distribuční funkce, pravděpodobnostní funkce, hustota pravděpodobnosti), které plně popisují pravděpodobnostní chování náhodné veličiny. Číselné charakteristiky vystihují pouze některé rysy tohoto chování, např. popisují polohu realizací náhodné veličiny na číselné ose či jejich proměnlivost (variabilitu). Jsou jednodušší než funkcionální charakteristiky, ale nesou jen částečnou informaci. Podobně jako v popisné statistice volíme vhodnou číselnou charakteristiku podle toho, jakého typu je daná náhodná veličina - zda je ordinální nebo intervalová či poměrová. Číselné charakteristiky znaků mají své teoretické protějšky v číselných charakteristikách náhodných veličin.

Definice:
Nechť X je náhodná veličina aspoň ordinálního charakteru a $\alpha \in (0,1)$. Číslo $K_\alpha(X)$ se nazývá α-kvantil náhodné veličiny X, jestliže splňuje nerovnosti:

$$P(X \le K_\alpha(X)) \ge \alpha \wedge P(X \ge K_\alpha(X)) \ge 1-\alpha$$

Jiné možné označení kvantilu: x_α

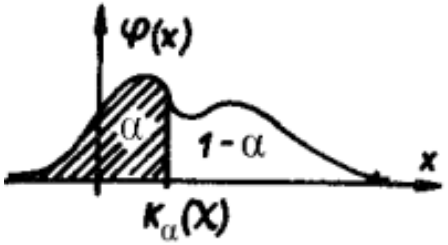
Kvantil $K_{0,50}(X)$ se nazývá medián,
$K_{0,25}(X)$ dolní kvartil,
$K_{0,75}(X)$ horní kvartil,
kvantily $K_{0,10}(X)$, ..., $K_{0,90}(X)$ jsou decily,
$K_{0,01}(X)$, ..., $K_{0,99}(X)$ jsou percentily.
Kterýkoliv α-kvantil je charakteristikou polohy číselných realizací náhodné veličiny na číselné ose.
Jako charakteristika variability slouží kvartilová odchylka $q = K_{0,75}(X) - K_{0,25}(X)$.

Kvantil spojité NV

Důsledek: (pro spojitou náhodnou veličinu)

Je-li X spojitá náhodná veličina, pak $K_\alpha(X)$ je takové číslo, pro které platí: $\alpha = \Phi(K_\alpha(X)) = \int_{-\infty}^{K_\alpha(X)} \varphi(x)dx$.

Ilustrace:



Příklad

Příklad: Nechť X ~ Ex(1). Určete medián a kvartilovou odchylku.

Řešení: $\varphi(x)=\begin{cases} e^{-x} & \text{prox}>0 \\ 0 & \text{jinak} \end{cases}$, $\Phi(x)=\begin{cases} 1-e^{-x} & \text{prox}>0 \\ 0 & \text{jinak} \end{cases}$

$$\alpha=\Phi(K_\alpha(X))=1-e^{-K_\alpha(X)} \Rightarrow K_\alpha(X)=-\ln(1-\alpha)$$

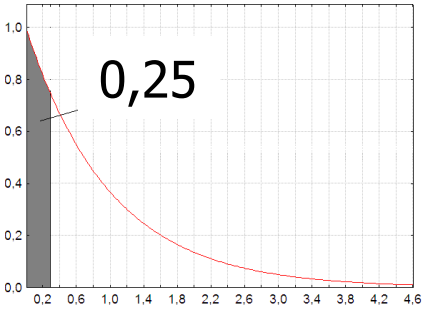
$$K_{0,50}(X)=-\ln(1-0,5)=-\ln\frac{1}{2}=\ln 2=0,693$$

$$K_{0,25}(X)=-\ln(1-0,25)=-\ln\frac{3}{4}=\ln 4-\ln 3=0,288$$

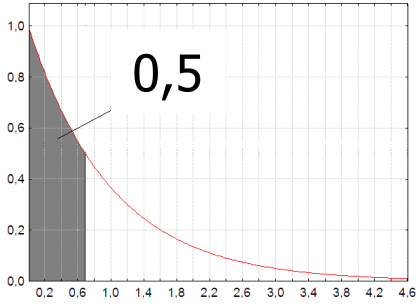
$$K_{0,75}(X)=-\ln(1-0,75)=-\ln\frac{1}{4}=\ln 4=1,386$$

$$q=K_{0,75}(X)-K_{0,25}(X)=1,386-0,288=1,098$$

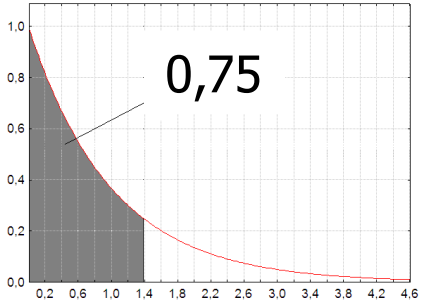
Dolní kvartil



Medián



Horní kvartil



Kvantily vybraných rozložení NV

Označení:

$X \sim N(0, 1) \Rightarrow K_\alpha(X) = u_\alpha$,
$X \sim \chi^2(n) \Rightarrow K_\alpha(X) = \chi^2_\alpha(n)$,
$X \sim t(n) \Rightarrow K_\alpha(X) = t_\alpha(n)$,
$X \sim F(n_1, n_2) \Rightarrow K_\alpha(X) = F_\alpha(n_1, n_2)$.
Tyto kvantily najdeme ve statistických tabulkách. Při jejich hledání používáme vztahy:

$u_\alpha = - u_{1-\alpha}$,
$t_\alpha(n) = - t_{1-\alpha}(n)$,
$F_\alpha(n_1, n_2) = \frac{1}{F_{1-\alpha}(n_2, n_1)}$.

Kvantily lze také vypočítat pomocí statistického software.

Příklad:

a) Nechť $U \sim N(0, 1)$. Najděte medián a horní a dolní kvartil.
b) Určete $\chi^2_{0,025}(25)$.
c) Určete $t_{0,99}(30)$ a $t_{0,05}(14)$.
d) Určete $F_{0,975}(5, 20)$ a $F_{0,05}(2, 10)$.

Řešení:

ad a) $u_{0,50} = 0$, $u_{0,25} = -0,67449$, $u_{0,75} = 0,67449$
ad b) $\chi^2_{0,025}(25) = 13,12$
ad c) $t_{0,99}(30) = 2,4573$, $t_{0,05}(24) = -1,7613$
ad d) $F_{0,975}(5, 20) = 3,2891$, $F_{0,05}(2, 10) = 0,05156$

Kvantily transformované NV

Věta:
Nechť X je spojitá náhodná veličina s distribuční funkcí $\Phi(x)$, $\alpha \in (0,1)$ a $g: R \to R$ ryze monotónní borelovská funkce. Pak pro α-kvantil transformované náhodné veličiny Y = g(X) platí:

a) Je-li g všude rostoucí funkce, pak $K_\alpha(Y) = g(K_\alpha(X))$.

b) Je-li g všude klesající funkce, pak $K_\alpha(Y) = g(K_{1-\alpha}(X))$.

Důkaz:

ad a) $\alpha = \Phi(K_\alpha(X)) = P(X \le K_\alpha(X)) = P(g(X) \le g(K_\alpha(X))) = P(Y \le g(K_\alpha(X))) = \Phi_*(g(K_\alpha(X))) \Rightarrow g(K_\alpha(X)) = K_\alpha(Y)$

ad b) $1-\alpha = \Phi(K_{1-\alpha}(X)) = P(X \le K_{1-\alpha}(X)) = P(g(X) \ge g(K_{1-\alpha}(X))) = 1 - P(Y \le g(K_{1-\alpha}(X))) = 1 - \Phi_*(g(K_{1-\alpha}(X))) \Rightarrow g(K_{1-\alpha}(X)) = K_\alpha(Y)$

Příklad:
Nechť U ~ N(0, 1). Najděte 9. decil transformované náhodné veličiny Y = 3 + 2U.

Řešení:
Funkce y = 3 + 2u je všude rostoucí funkce, tedy $K_{0,90}(Y) = 3 + 2\,u_{0,90} = 3 + 2 \times 1{,}28155 = 5{,}5631$.

Střední hodnota NV

Definice:

Nechť $(\Omega, \mathcal{A}, P)$ je pravděpodobnostní prostor, X náhodná veličina aspoň intervalového typu definovaná na měřitelném prostoru $(\Omega, \mathcal{A})$.

a) Je-li X diskrétní náhodná veličina s pravděpodobnostní funkcí $\pi(x)$, pak její střední hodnota (vzhledem k P) je číslo $E(X) = \sum_{x=-\infty}^{\infty} x\pi(x)$, pokud suma vpravo je konečná nebo absolutně konverguje. Jinak řekneme, že střední hodnota neexistuje.

b) Je-li X spojitá náhodná veličina s hustotou pravděpodobnosti $\varphi(x)$, pak její střední hodnota (vzhledem k P) je číslo $E(X) = \int_{-\infty}^{\infty} x\varphi(x)dx$, pokud integrál vpravo je konečný nebo absolutně konverguje. Jinak řekneme, že střední hodnota neexistuje.

(Střední hodnota je číslo, které charakterizuje polohu realizací náhodné veličiny na číselné ose s přihlédnutím k jejich pravděpodobnostem. V diskrétním případě představuje střední hodnota těžiště soustavy hmotných bodů, jejichž hmotnost je popsána pravděpodobnostní funkcí $\pi(x)$ a ve spojitém případě je střední hodnota těžištěm hmotné přímky, na níž je rozprostření hmoty popsáno hustotou pravděpodobnosti $\varphi(x)$. Střední hodnota je teoretickým protějškem váženého aritmetického průměru.)

Příklad

Příklad:
Náhodná veličina X udává počet ok při hodu kostkou. Vypočtěte její střední hodnotu.
Řešení :

$$\pi(x)=\begin{cases}\frac{1}{6} \text{ pro } x=1,\ldots,6\\ 0 \text{ jinak}\end{cases},\quad E(X)=\sum_{x=1}^{6} x\pi(x)=\frac{1}{6}(1+2+3+4+5+6)=\frac{7}{2}=3{,}5.$$

Příklad:

Rozložení náhodné veličiny X je dáno hustotou φ(x) = 2x+2 na (−1, 0) a nulovou jinde. Vypočtěte její střední hodnotu.

Řešení

$$E(X)=\int_R x\varphi(x)\,dx=\int_{-1}^{0} x(2x+2)\,dx=\left[2\frac{x^3}{3}+x^2\right]_{-1}^{0}=\frac{2}{3}-1=-\frac{1}{3}$$

Střední hodnota transformované NV

Věta:

a) Diskrétní případ: Nechť X je diskrétní náhodná veličina s pravděpodobnostní funkcí $\pi(x)$ (resp. $(X_1,\ldots,X_n)$ je diskrétní náhodný vektor s pravděpodobnostní funkcí $\pi(x_1,\ldots,x_n)$). Nechť $g : R \mapsto R$ je borelovská funkce, $Y = g(X)$ je transformovaná náhodná veličina (resp. $g : R^n \mapsto R$ je borelovská funkce, $Y = g(X_1,\ldots,X_n)$ je transformovaná náhodná veličina). Pak $E(Y) = \sum_{x=-\infty}^{\infty} g(x)\pi(x)$, pokud součet vpravo je konečný nebo absolutně konvergentní (resp. $E(Y) = \sum_{x_1=-\infty}^{\infty} \ldots \sum_{x_n=-\infty}^{\infty} g(x_1,\ldots,x_n)\pi(x_1,\ldots,x_n)$, pokud součet vpravo je konečný nebo absolutně konvergentní).

a) Spojitý případ: Nechť X je spojitá náhodná veličina s hustotou $\varphi(x)$ (resp. $(X_1,\ldots,X_n)$ je spojitý náhodný vektor s hustotou $\varphi(x_1,\ldots,x_n)$). Nechť $g : R \mapsto R$ je borelovská funkce, $Y = g(X)$ je transformovaná náhodná veličina (resp. $g : R^n \mapsto R$ je borelovská funkce, $Y = g(X_1,\ldots,X_n)$ je transformovaná náhodná veličina). Pak $E(Y) = \int_{-\infty}^{\infty} g(x)\varphi(x)dx$, pokud integrál vpravo je konečný nebo absolutně konvergentní (resp. $E(Y) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} g(x_1,\ldots,x_n)\varphi(x_1,\ldots,x_n)dx_1\ldots dx_n$, pokud integrál vpravo je konečný nebo absolutně konvergentní).

Příklad

Příklad:
Nechť $X \sim Ex(\lambda)$, $Y = e^{-\gamma X}$, kde $\gamma > 0$ je konstanta. Vypočtěte $E(Y)$.

Řešení:
$$\varphi(x) = \begin{cases} \lambda e^{-\lambda x} \text{ pro } x > 0 \\ 0 \text{ jinak} \end{cases}, \quad E(Y) = \int_0^{\infty} e^{-\gamma x} \lambda e^{-\lambda x} dx = \frac{\lambda}{\lambda + \gamma}.$$

Rozptyl NV

Definice:

Nechť (Ω , $\mathcal{A}$, P) je pravděpodobnostní prostor, X náhodná veličina aspoň intervalového typu definovaná na měřitelném prostoru (Ω , $\mathcal{A}$), která má střední hodnotu E(X). Rozptylem náhodné veličiny X rozumíme číslo $D(X) = E([X-E(X)]^2)$, pokud střední hodnota vpravo existuje. Číslo $\sqrt{D(X)}$ se nazývá směrodatná odchylka.

(Rozptyl je číslo, které charakterizuje proměnlivost realizací náhodné veličiny kolem její střední hodnoty s přihlédnutím k jejich pravděpodobnostem. Je teoretickým protějškem váženého rozptylu. Je vhodnější počítat rozptyl podle vzorce $D(X) = E(X^2) - [E(X)]^2$, jak bude ukázáno později.)

Důsledek:

V diskrétním případě je rozptyl dán vzorcem $D(X) = \sum_{x=-\infty}^{\infty} [x - E(X)]^2 \pi(x) = \sum_{x=-\infty}^{\infty} x^2 \pi(x) - [E(X)]^2$

a ve spojitém případě vzorcem: $D(X) = \int_{-\infty}^{\infty} [x - E(X)]^2 \varphi(x)\, dx = \int_{-\infty}^{\infty} x^2 \varphi(x)\, dx - [E(X)]^2$

(pokud suma či integrál vpravo absolutně konvergují).

Centrovaná a standardizovaná NV

Definice:

Transformovaná náhodná veličina $X - E(X)$ se nazývá centrovaná náhodná veličina .

Transformovaná náhodná veličina $\frac{X - E(X)}{\sqrt{D(X)}}$ se nazývá standardizovaná náhodná veličina .

Příklad: Náhodná veličina X udává počet ok při hodu kostkou. Vypočtěte její rozptyl.

Řešení : $\pi(x) = \begin{cases} \frac{1}{6} & \text{pro } x = 1, \ldots, 6 \\ 0 & \text{jinak} \end{cases}$, $E(X) = 3{,}5$ (viz př. 12.10.), $D(X) = \sum_{x=1}^{6} x^2 \frac{1}{6} - 3{,}5^2 = \ldots = \frac{35}{12} = 2{,}92.$

Kovariance a korelace NV

Definice: Kovariancí náhodných veličin X_1, X_2, které mají střední hodnoty $E(X_1)$, $E(X_2)$, rozumíme číslo $C(X_1, X_2) = E([X_1 - E(X_1)][X_2 - E(X_2)])$ (pokud střední hodnoty vpravo existují).

Kovariance je číslo, které charakterizuje proměnlivost realizací náhodných veličin X_1, X_2 kolem jejich středních hodnot s přihlédnutím k jejich pravděpodobnostem. Je-li kovariance kladná (záporná), pak to svědčí o existenci jistého stupně přímé (nepřímé) lineární závislosti mezi realizacemi náhodných veličin X_1, X_2. Je-li kovariance nulová, pak říkáme, že náhodné veličiny X_1, X_2 jsou nekorelované a znamená to, že mezi jejich realizacemi není žádný lineární vztah. Pozor – z nekorelovanosti nevyplývá stochastická nezávislost, zatímco ze stochastické nezávislosti plyne nekorelovanost. Kovariance je teoretickým protějškem vážené kovariance. Pro výpočet je vhodné použít vzorec

$$C(X_1, X_2) = E(X_1 X_2) - E(X_1)E(X_2).$$

Koeficientem korelace náhodných veličin X_1, X_2 rozumíme číslo

$$R(X_1, X_2) = \begin{cases} E\left(\dfrac{X_1 - E(X_1)}{\sqrt{D(X_1)}} \cdot \dfrac{X_2 - E(X_2)}{\sqrt{D(X_2)}}\right) & \text{pro } \sqrt{D(X_1)}\sqrt{D(X_2)} > 0 \\ 0 & \text{jinak} \end{cases}$$

, pokud střední hodnoty vpravo existují.

Koeficient korelace je číslo, které charakterizuje těsnost lineární závislosti realizací náhodných veličin X_1, X_2. Čím bližší je 1, tím těsnější je přímá lineární závislost, čím bližší je -1, tím těsnější je nepřímá lineární závislost. Je vhodnější počítat koeficient korelace podle vzorce $R(X_1, X_2) = \dfrac{C(X_1, X_2)}{\sqrt{D(X_1)}\sqrt{D(X_2)}}$

Kovariance NV

Důsledek:

V diskrétním případě je kovariance dána vzorcem

$$C(X_1, X_2) = \sum_{x_1=-\infty}^{\infty} \sum_{x_2=-\infty}^{\infty} [x_1 - E(X_1)] \cdot [x_2 - E(X_2)] \pi(x_1, x_2) = \sum_{x_1=-\infty}^{\infty} \sum_{x_2=-\infty}^{\infty} x_1 x_2 \pi(x_1, x_2) - E(X_1)\, E(X_2)$$

a ve spojitém případě vzorcem

$$C(X_1, X_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [x_1 - E(X_1)] \cdot [x_2 - E(X_1)] \varphi(x_1, x_2) dx_1 dx_2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1 x_2 \varphi(x_1, x_2) dx_1 dx_2 - E(X_1)\, E(X_2)$$

Příklad

Příklad ♫:
Náhodná veličina X udává příjem manžela (v tisících dolarů) a náhodná veličina Y příjem manželky (v tisících dolarů. Je známa simultánní pravděpodobnostní funkce $\pi(x,y)$ diskrétního náhodného vektoru (X,Y): $\pi(10,10) = 0{,}2$, $\pi(10,20) = 0{,}04$, $\pi(10,30) = 0{,}01$, $\pi(10,40) = 0$, $\pi(20,10) = 0{,}1$, $\pi(20,20) = 0{,}36$, $\pi(20,30) = 0{,}09$, $\pi(20,40) = 0$, $\pi(30,10) = 0$, $\pi(30,20) = 0{,}05$, $\pi(30,30) = 0{,}1$, $\pi(30,40) = 0$, $\pi(40,10) = 0$, $\pi(40,20) = 0$, $\pi(40,30) = 0$, $\pi(40,40) = 0{,}05$, $\pi(x,y) = 0$ jinak. Vypočtěte koeficient korelace příjmů manžela a manželky.
Řešení: Náhodná veličina X i náhodná veličina Y nabývají hodnot 10, 20, 30, 40. Sestavíme kontingenční tabulku:

X	Y				
	10	20	30	40	$\pi_1(x)$.
10	0,20	0,04	0,01	0,00	0,25
20	0,10	0,36	0,09	0,00	0,55
30	0,00	0,05	0,10	0,00	0,15
40	0,00	0,00	0,00	0,05	0,05
$\pi_2(y)$	0,30	0,45	0,20	0,05	1,00

Spočteme
$E(X) = 10.0{,}25+20.0{,}55+30.0{,}15+40.0{,}05 = 20$, $E(Y) = 10.0{,}30+20.0{,}45+30.0{,}20+40.0{,}05 = 20$,
$D(X) = 10^2.0{,}25+20^2.0{,}55+30^2.0{,}15+40^2.0{,}05 - 20^2 = 60$, $D(Y) = 10^2.0{,}30+20^2.0{,}45+30^2.0{,}20+40^2.0{,}05 - 20^2 = 70$,
$C(X,Y) = 10.10.0{,}20 + 10.20.0{,}04 + \ldots 40.40.0{,}05 - 20.20 = 49$,
$R(X,Y) = 49/\sqrt{60}\sqrt{70} = 0{,}76$.

Střední hodnota a rozptyl vybraných typů rozložení NV

Poznámka: Uvedeme střední hodnoty a rozptyly vybraných typů diskrétních a spojitých rozložení:

a) $X \sim Dg(\mu) \Rightarrow E(X) = \mu,\ D(X) = 0$

b) $X \sim A(\vartheta) \Rightarrow E(X) = \vartheta,\ D(X) = \vartheta\,(1-\vartheta)$

c) $X \sim Bi(n, \vartheta) \Rightarrow E(X) = n\vartheta,\ D(X) = n\vartheta\,(1-\vartheta)$

d) $X \sim Ge(\vartheta) \Rightarrow E(X) = \frac{1-\vartheta}{\vartheta},\ D(X) = \frac{1-\vartheta}{\vartheta^2}$

e) $X \sim Hg(N,M,n) \Rightarrow E(X) = \frac{M}{N}n,\ D(X) = \frac{Mn}{N}\left(1-\frac{M}{N}\right)\frac{N-n}{N-1}$

f) $X \sim Rd(G) \Rightarrow E(X) = \frac{n-1}{2},\ D(X) = \frac{n^2-1}{12}$

g) $X \sim Po(\lambda) \Rightarrow E(X) = \lambda,\ D(X) = \lambda$

h) $X \sim Rs(a, b) \Rightarrow E(X) = \frac{a+b}{2},\ D(X) = \frac{(b-a)^2}{12}$

i) $X \sim Ex(\lambda) \Rightarrow E(X) = \frac{1}{\lambda},\ D(X) = \frac{1}{\lambda^2}$

j) $X \sim N(\mu, \sigma^2) \Rightarrow E(X) = \mu,\ D(X) = \sigma^2$

k) $X \sim \chi^2(n) \Rightarrow E(X) = n,\ D(X) = 2n$

l) $X \sim t(n) \Rightarrow E(X) = 0$ pro $n \geq 2$, pro $n = 1$ $E(X)$ neexistuje, $D(X) = \frac{n}{n-2}$ pro $n \geq 3$, pro $n = 1, 2$ $D(X)$ neexistuje

m) $X \sim F(n_1, n_2) \Rightarrow E(X) = \frac{n_2}{n_2-2}$ pro $n_2 \geq 3$, pro $n_2 = 1, 2$ $E(X)$ neexistuje, $D(X) = \frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)^2(n_2-4)}$ pro $n_2 \geq 5$, pro $n_2 = 1, 2, 3, 4$ $D(X)$ neexistuje.

Příklad

Příklad:
V sadě 15 výrobků je 5 zmetků. Náhodně vybereme 4 výrobky. Určete střední hodnotu a rozptyl náhodné veličiny X, která udává počet zmetků, jestliže výběr provádíme
a) bez vracení,
b) s vracením.

Řešení:
ad a) X ~ Hg(N, M, n), N = 15, M = 5, n = 4

$$E(X) = \frac{M}{N} n = \frac{5}{15} 4 = \frac{4}{3} = 1,\bar{3}, \quad D(X) = \frac{Mn}{N}\left(1 - \frac{M}{N}\right)\frac{N-n}{N-1} = \frac{4}{3}\left(1 - \frac{5}{15}\right)\frac{11}{14} = \frac{44}{63} = 0,6984$$

ad b) X ~ Bi(n, ϑ), n = 4, $\vartheta = \frac{5}{15} = \frac{1}{3}$

$$E(X) = n\vartheta = 4\frac{1}{3} = 1,\bar{3}, \quad D(X) = n\vartheta(1-\vartheta) = \frac{4}{3}\left(1 - \frac{1}{3}\right) = \frac{8}{9} = 0,\bar{8}$$

Příklad

Najděte medián rozložení určeného hustotou $\varphi(x) = 1 - x/2,\ 0 < x < 2$.

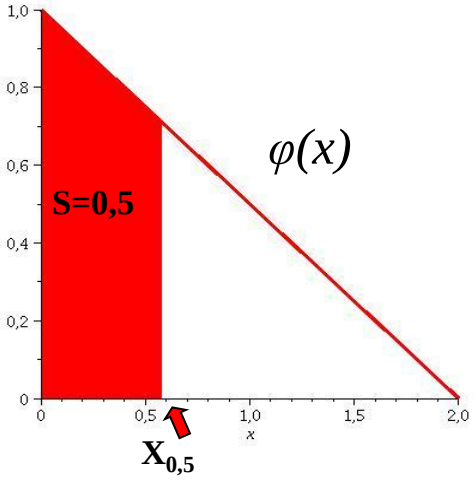
Řešení: Distribuční funkce: $F(x) = 0$ pro $x \le 0$, $F(x) = 1$ pro $x \ge 2$ a

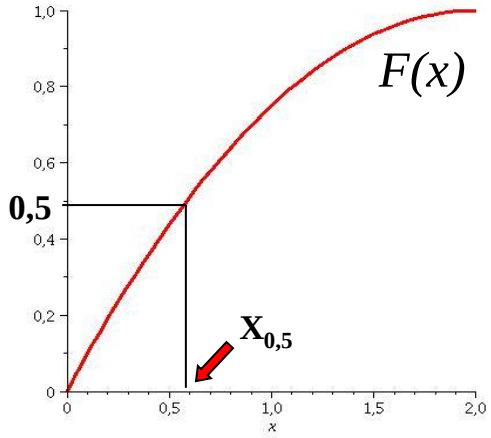
$$F(x) = \int_0^x 1 - \frac{t}{2}\,dt = x - \frac{x^2}{4} \qquad \text{pro} \quad x \in (0,2)$$

Medián $x_{0,5}$ je řešením rovnice $F(x) = 0{,}5$, tedy

$$x^2 - 4x + 2 = 0 \quad \Rightarrow x_{1,2} = \frac{4 \pm \sqrt{16-8}}{2} = 2 \pm \sqrt{2} = \begin{cases} 3{,}4142 \\ 0{,}5857 \end{cases}$$

Protože $3{,}4142 > 2$, je hledaným řešením $x_{0,5} = 0{,}5857$.



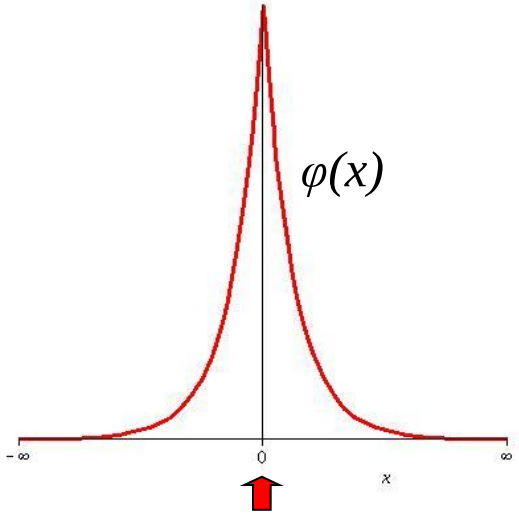


Příklad

Náhodná veličina má hustotu $\varphi(x) = a \cdot e^{-|x|} \quad x \in (-\infty, \infty)$ Určete a, střední hodnotu a rozptyl.

Řešení: $a = 1 \Big/ \left(2 \int_0^{\infty} e^{-x}\, dx \right) = \frac{1}{2}$

$$E(X) = \int_{-\infty}^{\infty} x\varphi(x)\, dx = \int_{-\infty}^{\infty} x \frac{1}{2} e^{-|x|}\, dx = 0 \qquad D(X) = E(X^2) = \int_{-\infty}^{\infty} x^2 \frac{1}{2} e^{-|x|}\, dx = 2$$



E(X)

Příklad

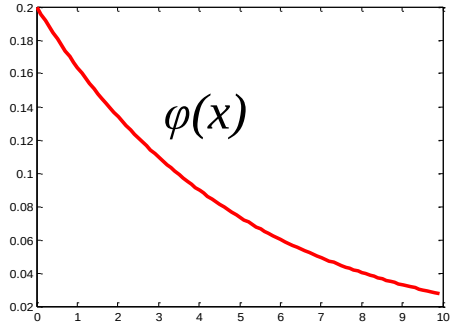
Nechť životnost (v letech) výrobků se řídí exponenciálním rozložením s distribuční funkcí $F(x) = 1-e^{-x/5}$, $x > 0$. Tj. střední doba životnosti je 5 let. Tvar distribuční funkce znamená, že k poruše výrobku dojde s velkou pravděpodobností velmi brzy po jeho prodeji. Jakou záruční dobu stanoví výrobce, nemá-li počet reklamovaných výrobků překročit 10%?

Řešení: Náhodná veličina X udává životnost výrobku. Hledáme takové x, aby platilo

$P(X \le x) = 0{,}1$ Tedy hledáme 10% kvantil.

$$\Rightarrow \quad 0{,}1 = 1 - e^{-x/5}$$

$$\Rightarrow \quad x = -5\ln(0{,}9) = -5\cdot(-0{,}10536) = 0{,}5268$$



Pro splnění požadované podmínky je třeba stanovit záruční dobu na cca ½ roku.

Příklad

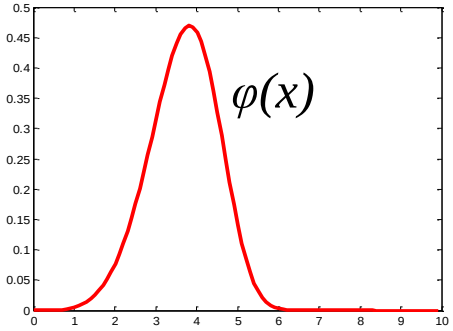
Nechť životnost (v letech) výrobků se řídí Weibullovým rozložením s distribuční funkcí $F(x) = 1\text{-}e^{-(x/4)^{\wedge}5}$, $x > 0$. Tj. střední doba životnosti je cca 3.67 let. Tvar distribuční funkce znamená, že k poruše výrobku pravděpodobně nedojde hned po jeho prodeji, ale až po nějaké době. Jakou záruční dobu stanoví výrobce, nemá-li počet reklamovaných výrobků překročit 10%?

Řešení: Náhodná veličina X udává životnost výrobku. Hledáme takové x, aby platilo

$P(X \le x) = 0{,}1$ Tedy hledáme 10% kvantil.

$$\Rightarrow \quad 0{,}1 = 1 - e^{-(x/4)^5}$$

$$\Rightarrow \quad x = 4 \cdot \sqrt[5]{-\ln(0{,}9)} = 2{,}55$$



Pro splnění požadované podmínky je třeba stanovit záruční dobu na cca 2,5 roku.

Momenty, šikmost a špičatost NV

Definice:

Nechť X, X_1, X_2 jsou náhodné veličiny, $k, k_1, k_2 \in R, r, s \in N$.

a) Číslo $E([X-k]^r)$ se nazývá **r-tý moment** náhodné veličiny X kolem konstanty k. Je-li $k = 0$, jde o **r-tý počáteční moment**, je-li $k = E(X)$, jedná se o **r-tý centrální moment**.

b) Číslo $E([X_1-k_1]^r[X_2-k_2]^s)$ se nazývá **$r \times s$-tý moment** náhodných veličin X_1, X_2 kolem konstant k_1, k_2. Je-li $k_1 = k_2 = 0$, jde o **$r \times s$-tý počáteční moment**, je-li $k_1 = E(X_1), k_2 = E(X_2)$, jedná se o **$r \times s$-tý centrální moment**.

Číslo $A_3(X) = \frac{E\left([X-E(X)]^3\right)}{\left[\sqrt{D(X)}\right]^3}$ se nazývá šikmost náhodné veličiny X.

Číslo $A_4(X) = \frac{E\left([X-E(X)]^4\right)}{\left[\sqrt{D(X)}\right]^4} - 3$ se nazývá špičatost náhodné veličiny X.

Je-li $A_3(X) = 0$, jde o symetrické rozložení. Je-li $A_3(X) > 0$, jde o kladně sešikmené rozložení a je-li $A_3(X) < 0$, jde o záporně sešikmené rozložení.

Je-li $A_4(X) = 0$, jde o rozložení s normální špičatostí. Je-li $A_4(X) > 0$, jde o špičaté rozložení a je-li $A_4(X) < 0$, jde o ploché rozložení.

Vektor středních hodnot, variační a korelační matice náhodného vektoru

Definice:

Nechť $\mathbf{X} = (X_1, \ldots, X_n)'$ je náhodný vektor. Reálný vektor $E(\mathbf{X}) = (E(X_1), \ldots, E(X_n))'$ se nazývá **vektor středních hodnot**. Reálná čtvercová symetrická matice

$$var(\mathbf{X}) = \begin{pmatrix} D(X_1) & C(X_1, X_2) & \ldots & C(X_1, X_n) \\ \ldots & \ldots & \ldots & \ldots \\ C(X_n, X_1) & C(X_n, X_2) & \ldots & D(X_n) \end{pmatrix}$$

se nazývá **varianční matice** náhodného vektoru $\mathbf{X}$ a reálná čtvercová symetrická matice

$$cor(\mathbf{X}) = \begin{pmatrix} 1 & R(X_1, X_2) & \ldots & R(X_1, X_n) \\ \ldots & \ldots & \ldots & \ldots \\ R(X_n, X_1) & R(X_n, X_2) & \ldots & 1 \end{pmatrix}$$

se nazývá **korelační matice** náhodného vektoru $\mathbf{X}$.

Příklad

Příklad:
Pro náhodný vektor (X, Y) z příkladu ♫ (str. 341). najděte vektor středních hodnot, varianční a korelační matici.

Řešení:
Bylo spočteno, že
E(X) = 20, E(Y) = 20,
D(X) = 60, D(Y) = 70,
C(X,Y) = 49,
R(X,Y) = 0,76.

Řešením jsou tedy:

$$\mathrm{E}(\mathbf{X}) = \begin{pmatrix} 20 \\ 20 \end{pmatrix}, \mathrm{var}(\mathbf{X}) = \begin{pmatrix} 60 & 49 \\ 49 & 70 \end{pmatrix}, \mathrm{cor}(\mathbf{X}) = \begin{pmatrix} 1 & 0{,}76 \\ 0{,}76 & 1 \end{pmatrix}$$

Vlastnosti číselných charakteristik NV

Věta: Nechť a, a_1, a_2, b, b_1, b_2 jsou reálná čísla, X, X_1, ..., X_n, Y_1, ..., Y_m jsou náhodné veličiny definované na témž pravděpodobnostním prostoru. V následujících vzorcích vždy z existence číselných charakteristik na pravé straně vyplývá existence výrazu na levé straně.

Vlastnosti střední hodnoty

a) $E(a) = a$

b) $E(a + bX) = a + bE(X)$

c) $E(X - E(X)) = 0$

d) $E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i)$

e) Jsou-li náhodné veličiny $X_1, ..., X_n$ stochasticky nezávislé, pak $E\left(\prod_{i=1}^{n} X_i\right) = \prod_{i=1}^{n} E(X_i)$

Vlastnosti číselných charakteristik NV – kovariance

Vlastnosti kovariance

a) $C(a_1, X_2) = C(X_1, a_2) = C(a_1, a_2) = 0$

b) $C(a_1 + b_1X_1, a_2 + b_2X_2) = b_1b_2C(X_1, X_2)$

c) $C(X, X) = D(X)$

d) $C(X_1, X_2) = C(X_2, X_1)$

e) $C(X_1, X_2) = E(X_1X_2) - E(X_1)E(X_2)$

f) $C\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} C(X_i, Y_j)$

Vlastnosti číselných charakteristik NV – rozptyl

Vlastnosti rozptylu

a) $D(a) = 0$

b) $D(a + bX) = b^2 D(X)$

c) $D(X) = E(X^2) - [E(X)]^2$

d) $D\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} D(X_i) + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} C(X_i, X_j)$ (jsou -li náhodné veličiny $X_1, ..., X_n$ nekorelované, pak $D\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} D(X_i)$)

Vlastnosti číselných charakteristik NV – korelace

Vlastnosti koeficientu korelace

a) $R(a_1, X_2) = R(X_1, a_2) = R(a_1, a_2) = 0$

b) $R(a_1 + b_1X_1, a_2 + b_2X_2) = \mathrm{sgn}(b_1b_2)\, R(X_1, X_2)$

c) $R(X, X) = 1$ pro $D(X) \neq 0$, $R(X, X) = 0$ jinak

d) $R(X_1, X_2) = R(X_2, X_1)$

e) $R(X_1, X_2) = \begin{cases} \dfrac{C(X_1, X_2]}{\sqrt{D(X_1)}\sqrt{D(X_2)}} & \text{pro } \sqrt{D(X_1)}\sqrt{D(X_2)} > 0 \\ 0 \text{ jinak} \end{cases}$

Vlastnosti střední hodnoty - důkaz

Důkaz:

Pro vlastnosti střední hodnoty

ad a) $X \sim Dg(a)$, $\pi(x) = \begin{cases} 1 \text{ pro } x = a \\ 0 \text{ jinak} \end{cases}$, $E(X) = \sum_{x=-\infty}^{\infty} x\pi(x) = a\pi(a) = a \cdot 1 = a$

ad b) Diskrétní případ: $E(a + bX) = \sum_{x=-\infty}^{\infty} (a + bx)\pi(x) = \sum_{x=-\infty}^{\infty} a\pi(x) + \sum_{x=-\infty}^{\infty} bx\pi(x) = a\sum_{x=-\infty}^{\infty} \pi(x) + b\sum_{x=-\infty}^{\infty} x\pi(x) = a + bE(X)$

Spojitý případ: $$E(a+bX) = \int_{-\infty}^{\infty} (a+bx)\varphi(x)dx = a\int_{-\infty}^{\infty} \varphi(x)dx + b\int_{-\infty}^{\infty} x\varphi(x)dx = a + bE(X)$$

ad c) Plyne z (b), kde a = -E(X), b = 1.

ad d) Spojitý případ:

$$E\left(\sum_{i=1}^{n} X_i\right) = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} (x_1 + \ldots + x_n)\varphi(x_1,\ldots,x_n)dx_1\cdots dx_n =$$

$$= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} x_1\varphi(x_1,\ldots,x_n)dx_1\cdots dx_n + \cdots + \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} x_n\varphi(x_1,\ldots,x_n)dx_1\cdots dx_n =$$

$$= \int_{-\infty}^{\infty} x_1\left[\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \varphi(x_1,\ldots,x_n)dx_2\cdots dx_n\right]dx_1 + \cdots + \int_{-\infty}^{\infty} x_n\left[\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \varphi(x_1,\ldots,x_n)dx_1\cdots dx_{n-1}\right]dx_n =$$

$$= \int_{-\infty}^{\infty} x_1\varphi_1(x_1)dx_1 + \cdots + \int_{-\infty}^{\infty} x_n\varphi_n(x_n)dx_n = E(X_1) + \cdots + E(X_n) = \sum_{i=1}^{n} E(X_i)$$

Vlastnosti střední hodnoty - důkaz

ad d) Diskrétní případ: analogicky jako ve spojitém případě.

ad e) Spojitý případ:

$$E\left(\prod_{i=1}^{n} X_i\right) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (x_1 \cdot \ldots \cdot x_n)\varphi(x_1, \ldots, x_n)\, dx_1 \cdots dx_n =$$

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (x_1 \cdot \ldots \cdot x_n)\varphi(x_1) \cdot \ldots \cdot \varphi(x_n)\, dx_1 \cdots dx_n =$$

$$= \int_{-\infty}^{\infty} x_1 \varphi(x_1)\, dx_1 \cdot \ldots \cdot \int_{-\infty}^{\infty} x_n \varphi(x_n)\, dx_n =$$

$$= \prod_{i=1}^{n} E(X_i)$$

ad e) Diskrétní případ: analogicky jako ve spojitém případě.

Vlastnosti kovariance - důkaz

Pro vlastnosti kovariance:

ad a) $C(a_1, X_2) = E([a_1 - E(a_1)][X_2 - E(X_2)]) = E([a_1 - a_1][X_2 - E(X_2)]) = E(0) = 0$

ad b)

$$C(a_1 + b_1X_1, a_2 + b_2X_2) = E([a_1 + b_1X_1 - (a_1 + b_1E(X_1))][a_2 + b_2X_2 - (a_2 + b_2E(X_2))]) = b_1b_2E([X_1 - E(X_1)][X_2 - E(X_2)]) = b_1b_2C(X_1, X_2)$$

ad c) $C(X, X) = E([X - E(X)][X - E(X)]) = E([X - E(X)]^2) = D(X)$

ad d) $C(X_1, X_2) = E([X_1 - E(X_1)][X_2 - E(X_2)]) = E([X_2 - E(X_2)][X_1 - E(X_1)]) = C(X_2, X_1)$

ad e)

$$C(X_1, X_2) = E([X_1 - E(X_1)][X_2 - E(X_2)]) = E(X_1X_2 - X_2E(X_1) - X_1E(X_2) + E(X_1)E(X_2)) = E(X_1X_2) - E(X_2X_1) - E(X_1)E(X_2) + E(X_1)E(X_2) =$$
$$= E(X_1X_2) - E(X_1)E(X_2)$$

ad f)

$$C\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = E\left(\left[\sum_{i=1}^{n} X_i - E\left(\sum_{i=1}^{n} X_i\right)\right]\left[\sum_{j=1}^{m} Y_j - E\left(\sum_{j=1}^{m} Y_j\right)\right]\right) = E\left(\sum_{i=1}^{n}[X_i - E(X_i)]\sum_{j=1}^{m}[Y_j - E(Y_j)]\right) = \sum_{i=1}^{n}\sum_{j=1}^{m}[X_i - E(X_i)][Y_j - E(Y_j)] =$$
$$= \sum_{i=1}^{n}\sum_{j=1}^{m} C(X_i, Y_j)$$

Vlastnosti rozptylu - důkaz

Pro vlastnosti rozptylu:

ad a) $$D(a) = E\left(\left[a - E(a)\right]^2\right) = E\left(\left[a - a\right]^2\right) = E(0) = 0$$

ad b) $$D(a+bX) = E\left(\left[a+bX - E(a+bX)\right]^2\right) = E\left(\left[a+bX - a - bE(X)\right]^2\right) =$$
$$= E\left(b^2\left[X - E(X)\right]^2\right) = b^2 D(X)$$

ad c) $$D(X) = E\left(\left[X - E(X)\right]^2\right) = E\left(X^2 - 2XE(X) + \left[E(X)\right]^2\right) =$$
$$= E\left(X^2\right) - 2E(X)E(X) + \left[E(X)\right]^2 = E\left(X^2\right) - \left[E(X)\right]^2$$

ad d) $$D\left(\sum_{i=1}^{n} X_i\right) = C\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} C\left(X_i, X_j\right) =$$
$$= C(X_1, X_1) + C(X_1, X_2) + \ldots + C(X_1, X_n) + \ldots +$$
$$+ C(X_n, X_1) + C(X_n, X_2) + \ldots C(X_n, X_n) =$$
$$= \sum_{i=1}^{n} D(X_i) + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} C\left(X_i, X_j\right)$$

Vlastnosti korelace - důkaz

Pro vlastnosti koeficientu korelace:

ad a) Plyne přímo z definice, protože $D(a_1) = D(a_2) = 0$,

ad b)
$$R(a_1 + b_1X_1, a_2 + b_2X_2) = E\left(\frac{a_1 + b_1X_1 - E(a_1 + b_1X_1)}{\sqrt{D(a_1 + b_1X_1)}} \cdot \frac{a_2 + b_2X_2 - E(a_2 + b_2X_2)}{\sqrt{D(a_2 + b_2X_2)}} \right) =$$

$$= E\left(\frac{a_1 + b_1X_1 - a_1 - b_1E(X_1)}{\sqrt{b_1{}^2 D(X_1)}} \cdot \frac{a_2 + b_2X_1 - a_2 - b_2E(X_2)}{\sqrt{b_2{}^2 D(X_2)}} \right) =$$

$$= \frac{b_1}{|b_1|} \cdot \frac{b_2}{|b_2|} R(X_1, X_2) = \operatorname{sgn}(b_1 \cdot b_2) R(X_1, X_2)$$

Vlastnosti korelace - důkaz

ad c) Pro D(X) = 0 plyne přímo z definice, jinak platí

$$R(X,X) = E\left(\frac{X - E(X)}{\sqrt{D(X)}} \cdot \frac{X - E(X)}{\sqrt{D(X)}} \right) = \frac{1}{D(X)} E\left(\left[X - E(X) \right]^2 \right) = \frac{1}{D(X)} D(X) = 1$$

ad d) Zřejmé.

ad e)
$$R(X_1,X_2) = E\left(\frac{X_1 - E(X_1)}{\sqrt{D(X_1)}} \cdot \frac{X_2 - E(X_2)}{\sqrt{D(X_2)}} \right) =$$

$$= \frac{E\left(\left[X_1 - E(X_1) \right] \cdot \left[X_2 - E(X_2) \right] \right)}{\sqrt{D(X_1)} \cdot \sqrt{D(X_2)}} = \frac{C(X_1,X_2)}{\sqrt{D(X_1)} \cdot \sqrt{D(X_2)}}$$

Příklad

Příklad:
Vypočtěte střední hodnotu a rozptyl
a) centrované náhodné veličiny $Y = X - E(X)$,
b) standardizované náhodné veličiny $U = \frac{X - E(X)}{\sqrt{D(X)}}$.

Řešení:
ad a) $E(Y) = E(X - \mu) = E(X) - E(\mu) = \mu - \mu = 0$, $D(Y) = D(X - \mu) = D(X) = \sigma^2$,
ad b) $E(U) = E(\frac{X - \mu}{\sigma}) = \frac{1}{\sigma} E(X - \mu) = \frac{1}{\sigma} . 0 = 0$, $D(U) = D(\frac{X - \mu}{\sigma}) = \frac{1}{\sigma^2} D(X - \mu) = \frac{1}{\sigma^2} . \sigma^2 = 1$.

Příklad:
Náhodné veličiny X, Z jsou náhodné chyby, které vznikají na vstupním zařízení. Mají střední hodnoty $E(X) = -2$, $E(Y) = 4$ a rozptyly $D(X) = 4$, $D(Y) = 9$. Koeficient korelace těchto chyb je $R(X,Y) = -0,5$. Chyba na výstupu zařízení souvisí s chybami na vstupu funkční závislostí $Z = 3X^2 - 2XY + Y^2 - 3$. Najděte střední hodnotu chyby na výstupu.

Řešení: $E(Z) = E(3X^2 - 2XY + Y^2 - 3) = 3E(X^2) - 2E(XY) + E(Y^2) - E(3) = 3\{D(X) + [E(X)]^2\} - 2[C(X,Y) + E(X)E(Y)] + D(Y) + [E(Y)]^2 - 3 = 3[D(X) + [E(X)]^2] - 2[R(X,Y)\sqrt{D(X)}\sqrt{D(Y)} + E(X)E(Y)] + D(Y) + [E(Y)]^2 - 3 = 3(4 + 4) - 2[-0,5 \times 2 \times 3 + (-2) \times 4] + 9 + 16 - 3 = 24 + 22 + 25 - 3 = 68$

Příklad

Náhodná veličina X udává počet ok při hodu kostkou. NV Y = 2+ 3X. Vypočtěte:

a) E(X) a D(X),

b) E(Y) a D(Y),

c) C(X,Y),

d) R(X,Y).

Řešení:

a) $E(X) = \frac{1}{6}\sum_{x=1}^{6} x = 3{,}5 \qquad D(X) = \frac{1}{6}\sum_{x=1}^{6} x^2 - E(X)^2 = \frac{91}{6} - 3{,}5^2 = 2{,}9167$

b) $E(Y) = E(2+3X) = 2+3E(X) = 2+3\cdot 3{,}5 = 12{,}5$

$D(Y) = D(2+3X) = 3^2 D(X) = 9\cdot 2{,}9167 = 26{,}25$

c) $C(X,Y) = C(X,2+3X) = 3C(X,X) = 3D(X) = 3\cdot 2{,}9167 = 8{,}7501$

d) $R(X,Y) = R(X,2+3X) = \mathrm{sgn}(3)R(X,X) = 1\cdot 1 = 1$

Příklad

Náhodná veličina X udává součet počtu ok při hodu 2 kostkami. Vypočtěte E(X).

Řešení: X_i … počet ok při i-tém hodu, i = 1,…,6

$$E(X_i) = 3{,}5$$

$$E(X) = E\left(\sum_{i=1}^{2} X_i\right) = \sum_{i=1}^{2} E(X_i) = \sum_{i=1}^{2} 3{,}5 = 7$$

Nebo:

součet	počet možností	možnosti
2	1	11
3	2	12 21
4	3	22 13 31
5	4	23 32 41 14
6	5	33 24 42 51 15
7	6	34 43 25 52 16 61
8	5	44 35 53 26 62
9	4	54 45 36 63
10	3	55 64 46
11	2	56 65
12	1	66
Celkem	**36**	

$$E(X) = \sum_{x=2}^{12} x\pi(x) = \frac{1}{36}(2\cdot 1 + 3\cdot 2 + 4\cdot 3 + 5\cdot 4 + 6\cdot 5$$

$$+ 7\cdot 6 + 8\cdot 5 + 9\cdot 4 + 10\cdot 3 + 11\cdot 2 + 12\cdot 1) = \frac{252}{36} = 7$$

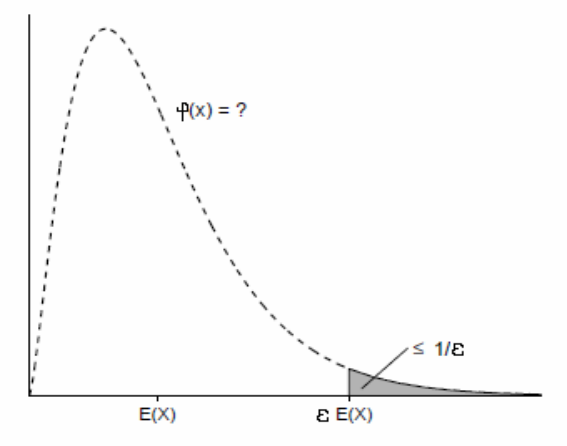
Markovova nerovnost

Věta (Markovova nerovnost):

Nechť pro náhodnou veličinu X se střední hodnotou E(X) platí P(X > 0) = 1. Pak platí Markovova nerovnost:

$$\forall \varepsilon > 0 : \boxed{P(X > \varepsilon E(X)) \leq \frac{1}{\varepsilon}}.$$

Ilustrace pro spojitý případ:



Důkaz: Pro spojitý případ:

$$E(X) = \int_0^\infty x\varphi(x)dx \geq \int_{\varepsilon E(X)}^\infty x\varphi(x)dx \geq \int_{\varepsilon E(X)}^\infty \varepsilon E(X)\varphi(x)dx = \varepsilon E(X)\int_{\varepsilon E(X)}^\infty \varphi(x)dx = \varepsilon E(X)P(X > \varepsilon E(X))$$

$$\Rightarrow P(X > \varepsilon E(X)) \leq \frac{1}{\varepsilon}$$

Příklad

Příklad:

Nechť $P(X > 0) = 1$ a $E(X) = \delta$, kde $\delta > 0$ je konstanta.

a) Odhadněte $P(X > 3\delta)$.

b) Nechť $X \sim Ex\left(\frac{1}{\delta}\right)$. Vypočtěte $P(X > 3\delta)$.

Řešení:

ad a) $P(X > 3\delta) \leq \frac{1}{3} = 0,\overline{3}$

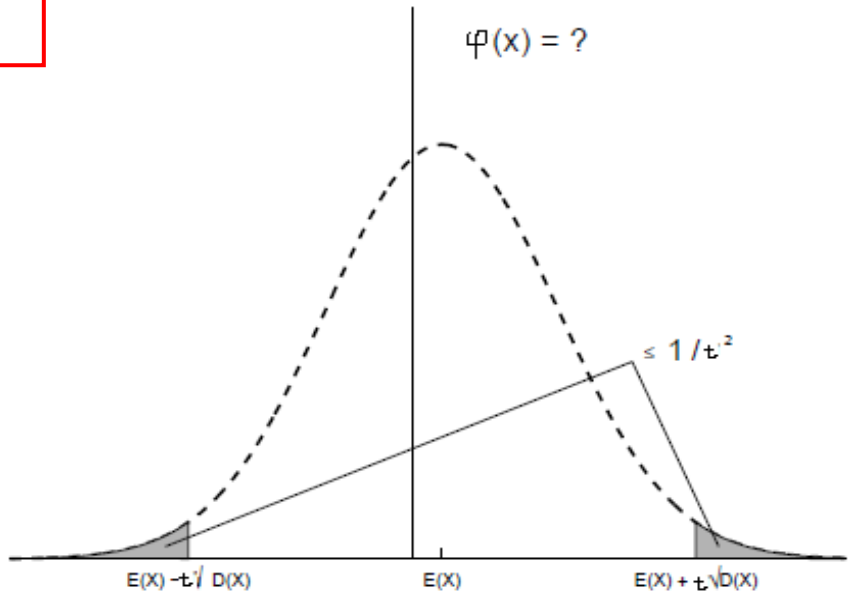
ad b) $X \sim Ex\left(\frac{1}{\delta}\right) \Rightarrow \varphi(x) = \begin{cases} \frac{1}{\delta} e^{-\frac{x}{\delta}} \text{ pro } x > 0 \\ 0 \text{ jinak} \end{cases}$, $E(X) = \delta, P(X > 3\delta) = \int_{3\delta}^{\infty} \frac{1}{\delta} e^{-\frac{x}{\delta}} dx = \left[-e^{-\frac{x}{\delta}} \right]_{3\delta}^{\infty} = e^{-3} = 0,04975$

Čebyševova nerovnost

Věta (Čebyševova nerovnost):
Nechť náhodná veličina X má střední hodnotu E(X) a rozptyl D(X). Pak platí Čebyševova nerovnost:

$$\forall t > 0 : P\left(\left|X - E(X)\right| > t\sqrt{D(X)}\right) \leq \frac{1}{t^2}.$$

Ilustrace pro spojitý případ:



Důkaz: Pro spojitý případ: Plyne z Markovovy nerovnosti, kde položíme $Y = [X - E(X)]^2$. Pak $P(Y > 0) = 1$ a pro $\forall \varepsilon > 0 : P(Y > \varepsilon E(Y)) \leq \frac{1}{\varepsilon}$, tj. pro $\forall \varepsilon > 0 : P\left([X - E(X)]^2 > \varepsilon E\left([X - E(X)]^2\right)\right) \leq \frac{1}{\varepsilon}$. Položme $\varepsilon = t^2$. Po odmocnění máme $\forall t > 0 : P\left(\left|X - E(X)\right| > t\sqrt{D(X)}\right) \leq \frac{1}{t^2}$.

Příklad

Příklad: Nechť $E(X) = \mu$, $D(X) = \sigma^2$.
a) Odhadněte $P(|X - \mu| > 3\sigma)$.
b) Jestliže $X \sim N(\mu, \sigma^2)$, vypočtěte $P(|X - \mu| > 3\sigma)$.

Řešení:

ad a) $P(|X - \mu| > 3\sigma) \leq \frac{1}{3^2} = \frac{1}{9} = 0,\bar{1}$.

(Tomuto výsledku se říká pravidlo 3σ a říká, že nejvýše 11,1% realizací náhodné veličiny leží vně intervalu $(\mu - 3\sigma, \mu + 3\sigma)$.)

ad b) $P(|X - \mu| > 3\sigma) = 1 - P(-3\sigma \leq X - \mu \leq 3\sigma) = 1 - P(-3 \leq \frac{X - \mu}{\sigma} \leq 3) = 1 - \Phi(3) + \Phi(-3) = 2[1 - \Phi(3)] = 2(1 - 0,99865) =$ 0,0027. (Má-li náhodná veličina normální rozložení, pak pouze 0,27% realizací leží vně intervalu $(\mu - 3\sigma, \mu + 3\sigma)$.)

Cauchy – Schwarzova – Buňakovského nerovnost

Věta (Cauchyova – Schwarzova – Buňakovského nerovnost):
Nechť $R(X_1, X_2)$ je koeficient korelace náhodných veličin X_1, X_2. Pak $|R(X_1, X_2)| \le 1$ a rovnost nastane tehdy a jen tehdy, když mezi veličinami X_1, X_2 existuje s pravděpodobností 1 úplná lineární závislost, tj. existují konstanty a, b tak, že $P(X_2 = a + bX_1) = 1$.

Důkaz: Zavedeme standardizované náhodné veličiny $U_i = \frac{X_i - E(X_i)}{\sqrt{D(X_i)}}$, i = 1, 2.

$$0 \le D(U_1 \pm U_2) = D(U_1) \pm 2C(U_1, U_2) + D(U_2) = 2[1 \pm R(X_1, X_2)] \Rightarrow |R(X_1, X_2)| \le 1.$$

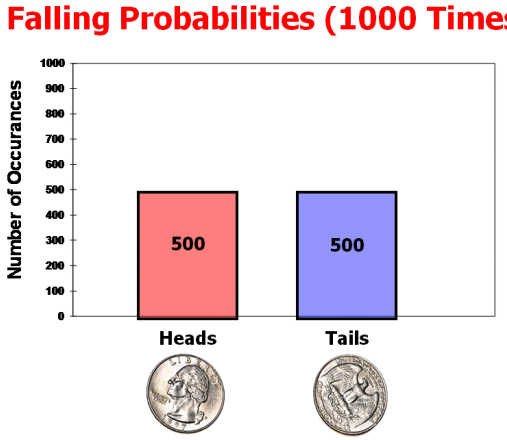
Předpokládejme nejprve, že $R(X_1, X_2) = 1$. V tomto případě počítáme $D(U_1 - U_2) = 2[1 - R(X_1, X_2)] = 0$. To je možné jen tak, že $P(U_1 = U_2) = 1$, tj. $P(U_1 - U_2 = 0) = 1$, tj. $1 = P\left(\frac{X_1 - E(X_1)}{\sqrt{D(X_1)}} = \frac{X_2 - E(X_2)}{\sqrt{D(X_2)}}\right) = P\left(X_2 = E(X_2) - \frac{\sqrt{D(X_2)}}{\sqrt{D(X_1)}}E(X_1) + \frac{\sqrt{D(X_2)}}{\sqrt{D(X_1)}}X_1\right)$, tudíž

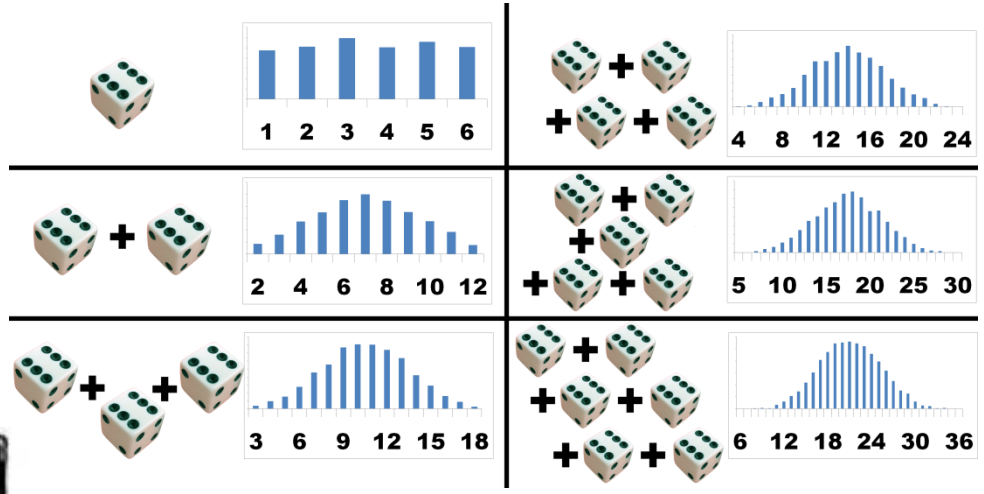
$$a = E(X_2) - \frac{\sqrt{D(X_2)}}{\sqrt{D(X_1)}}E(X_1), \; b = \frac{\sqrt{D(X_2)}}{\sqrt{D(X_1)}}.$$

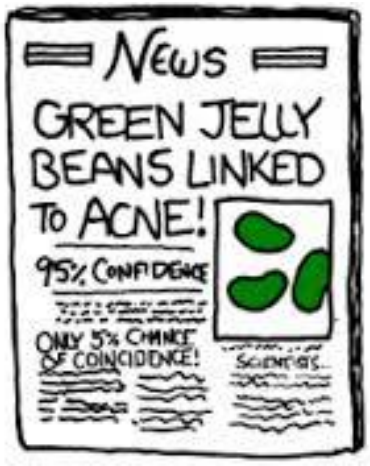
Předpokládáme-li, že $R(X_1, X_2) = -1$, pak počítáme $D(U_1 + U_2)$.

Nechť naopak $P(X_2 = a + bX_1) = 1$. Pak $R(X_1, X_2) = R(X_1, a + bX_1) = \mathrm{sgn}(b) R(X_1, X_1) = \mathrm{sgn}(b) = \begin{cases} 1 \text{ pro } b > 0 \\ -1 \text{ pro } b < 0 \end{cases}$.

11. Slabý zákon velkých čísel a centrální limitní věta, úvod do testování hypotéz







Slabý zákon velkých čísel a centrální limitní věta

S rostoucím počtem opakovaných nezávislých pokusů zjišťujeme, že empirické charakteristiky, které popisují výsledky těchto pokusů, se blíží teoretickým charakteristikám. Například relativní četnost úspěchu se blíží pravděpodobnosti úspěchu; průměr měření zatížených náhodnou chybou se blíží hledané neznámé střední hodnotě; empirická distribuční funkce se blíží distribuční funkci. Těmito skutečnostmi se zabývá Slabý zákon velkých čísel, specifikovaný např. Čebyševovou větou, nebo Bernoulliovou větou.

Podstatou centrální limitní věty je tvrzení, že náhodná veličina X, která vznikla jako součet velkého počtu vzájemně nezávislých náhodných veličin $X_1, X_2, \ldots, X_n$ má za velmi obecných podmínek přibližně normální rozdělení. Nejjednodušší specifikací centrální limitní věty je Moivre-Laplaceova věta. Zobecněním Moivre-Laplaceovy věty je věta Lindbergova-Lévyova. Nejobecněji centrální limitní větu formuloval Ljapunov, jeho větu však nebudeme uvádět. V současné době, kdy databáze mají ohromné množství položek, je aplikace CLV nesmírně užitečná.

Při uvedení zmíněných vět se neobejdeme bez pojmu konvergence posloupnosti náhodných veličin. V počtu pravděpodobnosti se nabízí řada způsobů, jak konvergenci posloupnosti náhodných veličin definovat, my si uvedeme následující tři.

Typy konvergence posloupnosti NV

Definice:

Říkáme, že náhodná posloupnost $(X_1, X_2, \ldots, X_n, \ldots)$ konverguje k náhodné veličině X

(i.) *jistě*, právě když všechny realizace náhodné posloupnosti $(X_1(\omega), X_2(\omega), \ldots, X_n(\omega), \ldots)$ konvergují k realizaci náhodné veličiny $X(\omega)$. Tedy platí:

$$\forall \omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)$$

[Jedná se o "obyčejnou" konvergenci číselné posloupnosti]

(ii.) *podle pravděpodobnosti*, právě když pro každé $\varepsilon > 0$ platí:

$$\lim_{n\to\infty} P(|X_n - X| \leq \varepsilon) = 1$$

[Při vzrůstajícím počtu pokusů jsou větší odchylky X_n od X krajně nepravděpodobné]

(iii.) *v distribuci*, právě když pro distribuční funkce $F_1(x_1) \sim X_1, \ldots, F_n(x_n) \sim X_n, \ldots$, popř. $F(x) \sim X$ platí:

$$\lim_{n\to\infty} F_n(x) = F(x) \text{ pro všechna } x, \text{ kde je funkce } F \text{ spojitá}$$

[Jedná se o nejslabší z uvedených typů konvergence, definuje se jen s užitím distribučních funkcí]

Typy konvergence posloupnosti NV

Poznámka:
Náhodná posloupnost $(X_1, X_2, \ldots, X_n, \ldots)$ může konvergovat i ke konstantě, což je v předchozí definici zahrnuto. Stačí uvažovat náhodnou veličinu X degenerovanou.

Věta:
Nechť $(X_1, X_2, \ldots, X_n, \ldots)$ je náhodná posloupnost.

1. Jestliže tato náhodná posloupnost konverguje k náh. vel. X jistě, pak k ní nutně konverguje i podle pravděpodobnosti. Konverguje-li k X podle pravděpodobnosti, pak k ní nutně konverguje i v distribuci. [Obrácené implikace obecně neplatí.]

2. K tomu, aby náhodná posloupnost $(X_1, X_2, \ldots, X_n, \ldots)$ konvergovala podle pravděpodobnosti k číslu μ, stačí splnění podmínek

$$\lim_{n\to\infty} E(X_n) = \mu \quad \wedge \quad \lim_{n\to\infty} D(X_n) = 0$$

Slabý zákon velkých čísel – Čebyševova věta

Věta: Čebyševova (slabý zákon velkých čísel)

Nechť náhodná posloupnost $(X_1, X_2, \ldots, X_n, \ldots)$ je posloupnost stochasticky nezávislých a stejně rozložených náhodných veličin se stejnou střední hodnotou μ a stejným rozptylem σ^2. Potom náhodná posloupnost aritmetických průměrů $(X_1, \frac{1}{2}\sum_{i=1}^{2} X_i, \ldots, \frac{1}{n}\sum_{i=1}^{n} X_i, \ldots)$ konverguje podle pravděpodobnosti ke střední hodnotě μ. Tedy pro každé $\varepsilon > 0$ platí:

$$P(|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu| < \varepsilon) \geq 1 - \frac{\sigma^2}{n\varepsilon^2}$$

neboli

$$\lim_{n\to\infty} P(|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu| < \varepsilon) = 1$$

[Při velkém počtu nezávislých pokusů můžeme téměř jistě očekávat, že aritmetický průměr jednotlivých pokusů se bude od střední hodnoty μ lišit krajně nepatrně. Proto při dostatečně velkém n lze střední hodnotu μ odhadnout průměrem výsledků jednotlivých pokusů.]

Bernoulliova věta

Věta: Bernoulliova (důsledek Čebyševovy věty)

Nechť náhodná veličina Y_n udává počet úspěchů v posloupnosti n nezávislých opakovaných pokusů, kdy úspěch nastává v každém pokusu s pravděpodobností ϑ, $0 < \vartheta < 1$. Pak posloupnost relativních četností $(Y_1, \frac{Y_2}{2}, \ldots \frac{Y_n}{n}, \ldots)$ konverguje podle pravděpodobnosti k pravděpodobnosti úspěchu ϑ. Tedy pro každé $\varepsilon > 0$ platí:

$$P(|\frac{Y_n}{n} - \vartheta| < \varepsilon) \geq 1 - \frac{\vartheta(1-\vartheta)}{n\varepsilon^2}$$

neboli

$$\lim_{n\to\infty} P(|\frac{Y_n}{n} - \vartheta| < \varepsilon) = 1$$

Příklad

Příklad:

Pravděpodobnost vyrobení zmetku je $\frac{12}{3000}$. Při výstupní kontrole bylo testováno 3000 výrobků. Jaká je pravděpodobnost, že relativní četnost výskytu zmetků se od pravděpodobnosti výskytu zmetku liší nejvýše o 0,01?

Řešení:

Označme Y_{3000} náhodnou veličinu udávající počet zmetků (úspěchů) v 3000 pokusech.
Potom $Y_{3000} \sim Bi(3000, \frac{12}{3000})$

Relativní četnost úspěchů by se s rostoucím n měla blížit k pravděpodobnosti úspěchu. My chceme určit pravděpodobnost, že pro $n = 3000$ se relativní četnost úspěchů od pravděpodobnosti úspěchu neodchýlí o více, než o 0,01. Tedy v Bernoulliove větě budeme za ε volit 0,01.

Pro každé $\varepsilon > 0$ platí: $P(|\frac{Y_n}{n} - \vartheta| < \varepsilon) \geq 1 - \frac{\vartheta(1-\vartheta)}{n\varepsilon^2}$.
Tedy $P(|\frac{Y_{3000}}{3000} - \frac{12}{3000}| < 0,01) \geq 1 - \frac{\frac{12}{3000}(1-\frac{12}{3000})}{3000 \cdot 0,01^2} \doteq 0,872$

Pokud bychom chtěli využít přímo Čebyševovu větu, pak bychom za X_i volili náhodnou veličinu s alternativním rozložením, kde jednička symbolizuje vyrobení zmetku (úspěch) a nula vyrobení kvalitního výrobku.
Tedy $X_i \sim A(\frac{12}{3000})$, $i = 1, \ldots, 3000$ $E(X_i) = \frac{12}{3000}$ $D(X_i) = \frac{12}{3000}(1 - \frac{12}{3000})$; $X_1, \ldots, X_{3000}$ jsou stoch. nezávislé. Dále stačí za ε volit 0,01 a dosadit do Čebyševovy věty. (Uvědomte si, že binomická náhodná veličina vzniká jako součet nezávislých, stejně rozložených alternativních náhodných veličin.)

Centrální limitní věta

Věta: Lindbergova-Lévyova (centrální limitní věta)
Nechť náhodná posloupnost $(X_1, \ldots, X_n, \ldots)$ je posloupnost stochasticky nezávislých a stejně rozložených náhodných veličin se stejnou střední hodnotou μ a stejným rozptylem σ^2. Uvažme součet $X = \sum_{i=1}^{n} X_i$ a odvoďme střední hodnotu a rozptyl nové náhodné veličiny X.

$$E(X) = E(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} E(X_i) = \sum_{i=1}^{n} \mu = n\mu$$

$$D(X) = D(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} D(X_i) = \sum_{i=1}^{n} \sigma^2 = n\sigma^2$$

Nyní uvažme standardizovaný součet $U_n = \frac{X - E(X)}{\sqrt{D(X)}} = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}}$ [n můžeme libovolně zvětšovat]
Potom náhodná posloupnost standardizovaných součtů $(U_1, U_2, \ldots, U_n, \ldots)$ konverguje v distribuci k náhodné veličině $U \sim N(0,1)$. Tedy

$$\forall u \in \mathrm{R}: \quad \lim_{n\to\infty} P(U_n \leq u) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt$$

Zkráceně píšeme $U_n \approx N(0,1)$ a říkáme, že U_n se asymptoticky řídí normálním standardizovaným rozložením. [Všimněte si, že $U_n = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i - \mu}{\frac{\sigma}{\sqrt{n}}}$. Centrální limitní věta tedy tvrdí, že s rostoucím n se distribuční funkce průměrů náhodných veličin $X_1, \ldots, X_n$ blíží distribuční funkci normálního rozložení se střední hodnotou μ a rozptylem $\frac{\sigma^2}{n}$. Toto nastává bez ohledu na původní rozložení náhodné veličiny X.]

Moivre – Laplaceova věta

Věta: Moivre-Laplaceova (důsledek Lindbergovy-Lévyovy věty)

Nechť $Y_n \sim Bi(n, \vartheta)$, $n = 1, 2, \ldots$. Potom $E(Y_n) = n\vartheta \quad D(Y_n) = n\vartheta(1-\vartheta)$
a $U_n = \frac{Y_n - n\vartheta}{\sqrt{n\vartheta(1-\vartheta)}} \approx N(0,1)$

[Moivre-Laplaceova věta říká, že při dostatečně velkém počtu nezávislých pokusů konverguje v distribuci binomické rozdělení k normálnímu.]

Poznámka:

Na základě Moivre-Laplaceova věty se používá přibližný vzorec, který nahrazuje pracný výpočet distribuční funkce binomického rozložení jednoduchým hledáním v tabulkách distribuční funkce normálního standardizovaného rozložení. Porovnejte

Přesný výpočet:

$P(Y_n \leq y) = \sum_{t=0}^{y} \binom{n}{t} \vartheta^t (1-\vartheta)^{n-t}$ … náročná sumace

Aproximace normálním rozložením:

$$P(Y_n \leq y) = P\left(\frac{Y_n - n\vartheta}{\sqrt{n\vartheta(1-\vartheta)}} \leq \frac{y - n\vartheta}{\sqrt{n\vartheta(1-\vartheta)}}\right) \approx \Phi\left(\frac{y - n\vartheta}{\sqrt{n\vartheta(1-\vartheta)}}\right) \sim N(0,1),$$

kde $\Phi(u)$ je tabelovaná distribuční funkce standardizovaného normálního rozložení.

Aproximaci je vhodné použít pokud jsou splněny následující podmínky:
$n\vartheta(1-\vartheta) > 9 \quad \wedge \quad \frac{1}{n+1} < \vartheta < \frac{n}{n+1}$.

Příklad

V určité skupině zaměstnanců je 10% s příjmem, který překračuje celostátní průměr. Kolik zaměstnanců z této skupiny je třeba vybrat, aby s pravděpodobností aspoň 0,95 bylo mezi nimi 8% až 12% zaměstnanců s nadprůměrným příjmem?

Řešení:

X – počet zaměstnanců s nadprůměrným příjmem, X ~ Bi(n, 0,1), E(X) = 0,1n, D(X) = 0,09n,

$$0{,}95 \le P\left(0{,}08 \le \frac{X}{n} \le 0{,}12\right) = P\left(0{,}08n \le X \le 0{,}12n\right) = P\left(\frac{0{,}08n - 0{,}1n}{\sqrt{0{,}09n}} \le \frac{X - 0{,}1n}{\sqrt{0{,}09n}} \le \frac{0{,}12n - 0{,}1n}{\sqrt{0{,}09n}}\right) =$$

$$P\left(\frac{-\sqrt{n}}{15} \le \frac{X - 0{,}1n}{\sqrt{0{,}09n}} \le \frac{\sqrt{n}}{15}\right) \approx \Phi\left(\frac{\sqrt{n}}{15}\right) - \Phi\left(-\frac{\sqrt{n}}{15}\right) = 2\Phi\left(\frac{\sqrt{n}}{15}\right) - 1 \Rightarrow \Phi\left(\frac{\sqrt{n}}{15}\right) \ge 0{,}975,$$

$$\text{tedy } \frac{\sqrt{n}}{15} \ge u_{0{,}975} = 1{,}96 \Rightarrow \sqrt{n} \ge 29{,}4 \Rightarrow n \ge 865.$$

Příklad

100-krát nezávisle na sobě házíme kostkou. Jaká je pravděpodobnost, že šestka padne aspoň 20-krát?

Řešení:

Označme Y_{100} náhodnou veličinu, udávající počet padnutých šestek ve 100 hodech, $Y_{100} \sim Bi(100, \frac{1}{6})$.

Nejdříve ověříme podmínky pro použití aproximace normálním rozložením:

$n\vartheta(1-\vartheta) = 100 \cdot \frac{1}{6}(1-\frac{1}{6}) = \frac{500}{36} > 9 \quad \wedge \quad \frac{1}{101} < \frac{1}{6} < \frac{100}{101}$, tedy obě podmínky jsou splněny.

Hledanou pravděpodobnost odhadneme pomocí Moivre-Laplaceovy věty.

$$P(Y_{100} \geq 20) = 1 - P(Y_{100} < 20) = P(Y_{100} \leq 19) = 1 - P\left(\frac{Y_{100} - 100/6}{\sqrt{100 \cdot 1/6 \cdot 5/6}} \leq \frac{19 - 100/6}{\sqrt{100 \cdot 1/6 \cdot 5/6}}\right) =$$

$$1 - P(U_n \leq 0,626) \approx 1 - \Phi(0,626) = 1 - 0,73565 = 0,2635.$$

(Přesný výpočet pomocí softwaru by vyšel 0,2198.)

Aproximace binomického rozložení normálním rozložením nemusí být vždy nejvhodnější. Pro extrémně malé pravděpodobnosti úspěchu ϑ užíváme přibližný vzorec, který vychází z Poissonovy věty.

Poissonova věta

Věta: Poissonova

Nechť $Y_1, Y_2, \ldots$ je posloupnost stochasticky nezávislých náhodných veličin, $Y_n \sim Bi(n, \vartheta_n)$, $n = 1, 2, \ldots$ a platí $\lim_{n\to\infty} n\vartheta_n = \lambda$. Pak posloupnost $Y_1, Y_2, \ldots Y_n, \ldots$ konverguje v distribuci k náhodné veličině $Y \sim Po(\lambda)$, tedy $Y_n \approx Po(\lambda)$.

[Náhodná veličina Y má Poissonovo rozložení s parametrem λ, náhodnou veličinu Y_n s binomickým rozložením lze aproximovat Poissonovým rozložením.]

Příklad

Poznámka:

Na základě Poissonovy věty se používá přibližný vzorec, který nahrazuje pracný výpočet distribuční (resp. pravděpodobnostní) funkce binomického rozložení jednoduchým hledáním v tabulkách distribuční (resp. pravděpodobnostní) funkce Poissonova rozložení.

- $P(Y_n \leq y) = \sum_{t=0}^{y} \binom{n}{t} \vartheta^t (1-\vartheta)^{n-t} \approx F_{n\vartheta}(y) \sim Po(n\vartheta)$, kde $F_{n\vartheta}(y)$ je distribuční funkce Poissonova rozložení s parametrem $\lambda = n\vartheta$
- $P(Y_n = y) = \binom{n}{y} \vartheta^y (1-\vartheta)^{n-y} \approx \frac{(n\vartheta)^y}{y!} e^{-n\vartheta}$ (Srovnej s pravděpodobnostní funkcí Poissonova rozložení, která je tabelovaná.)

Aproximaci je vhodné použít, pokud jsou splněny následující podmínky:
$n \geq 30 \wedge \vartheta \leq 0,1$.

Příklad

Během zkoušky spolehlivosti se přístroj porouchá s pravděpodobností 0,05. Jaká je pravděpodobnost, že při zkoušení 100 přístrojů se jich porouchá právě 5?

Řešení:

Označme Y_{100} náhodnou veličinu, udávající počet porouchaných přístrojů ve 100 zkouškách, $Y_{100} \sim Bi(100; 0,05)$.

Nejdříve ověříme podmínky pro použití aproximace Poissonovým rozložením:

$100 \geq 30 \quad \wedge \quad 0.05 \leq 0,1.$

Určení hledané pravděpodobnosti aproximací Poissonovým rozložením:

$P(Y_{100} = 5) \approx \frac{(100 \cdot 0,5)^5}{5!} e^{-100 \cdot 0,05}$, což nemusíme počítat, jelikož jde o pravděpodobnostní funkci Poissonova rozložení v bodě 5 s parametrem $\lambda = 100 \cdot 0,05$, která je v tabulkách. Tedy $p_5(5) = 0,17547$

Určení hledané pravděpodobnosti přesným výpočtem:

$P(Y_{100} = 5) = \binom{100}{5} 0,05^5 (1 - 0,05)^{95} = \ldots = 0,18$

Testování hypotéz

Motivace: Častým úkolem statistika je na základě dat ověřit předpoklady o parametrech nebo typu rozložení, z něhož pochází náhodný výběr. Takovému předpokladu se říká nulová hypotéza. Nulová hypotéza vyjadřuje nějaký teoretický předpoklad, často skeptického rázu a uživatel ji musí stanovit předem, bez přihlédnutí k datovému souboru. Proti nulové hypotéze stavíme alternativní hypotézu, která říká, co platí, když neplatí nulová hypotéza. Alternativní hypotéza je formulována tak, aby mohla platit jenom jedna z těchto dvou hypotéz. Pravdivost alternativní hypotézy by znamenala objevení nějakých nových skutečností nebo zásadnější změnu v dosavadních představách.

Např. výzkumník by chtěl na základě dat prověřit tezi (nový objev), že pasivní kouření škodí zdraví. Jako nulovou hypotézu tedy položí tvrzení, že pasivní kouření neškodí zdraví a proti nulové hypotéze postaví alternativní, že pasivní kouření škodí zdraví.

Testováním hypotéz se myslí rozhodovací postup, který je založen na daném náhodném výběru a s jehož pomocí rozhodneme o zamítnutí či nezamítnutí nulové hypotézy.

Definice nulové a alternativní hypotézy

Nechť $X_1, ..., X_n$ je náhodný výběr z rozložení $L(\vartheta)$, kde parametr $\vartheta \in \Xi$ neznáme. Nechť $h(\vartheta)$ je parametrická funkce a c daná reálná konstanta.

a) Oboustranná alternativa: Tvrzení H_0: $h(\vartheta) = c$ se nazývá jednoduchá nulová hypotéza. Proti nulové hypotéze postavíme složenou oboustrannou alternativní hypotézu H_1: $h(\vartheta) \neq c$.

b) Levostranná alternativa: Tvrzení H_0: $h(\vartheta) \geq c$ se nazývá složená pravostranná nulová hypotéza. Proti jednoduché nebo složené pravostranné nulové hypotéze postavíme složenou levostrannou alternativní hypotézu H_1: $h(\vartheta) < c$.

c) Pravostranná alternativa: Tvrzení H_0: $h(\vartheta) \leq c$ se nazývá složená levostranná nulová hypotéza. Proti jednoduché nebo složené levostranné nulové hypotéze postavíme složenou pravostrannou alternativní hypotézu H_1: $h(\vartheta) > c$.

Testování nulové a alternativní hypotézy

Testováním H_0 proti H_1 rozumíme rozhodovací postup založený na náhodném výběru $X_1, ..., X_n$, s jehož pomocí zamítneme či nezamítneme platnost nulové hypotézy.

(Volba alternativní hypotézy není libovolná, ale vyplývá z konkrétní situace. Např. při současné technologii je pravděpodobnost vyrobení zmetku $\vartheta = 0{,}01$.

a) Po rekonstrukci výrobní linky byla obnovena výroba, přičemž technologie zůstala stejná. Chceme ověřit, zda se změnila kvalita výrobků. Testujeme H_0: $\vartheta = 0{,}01$ proti H_1: $\vartheta \neq 0{,}01$.

b) Byly provedeny změny v technologii výroby s cílem zvýšit kvalitu. V tomto případě tedy testujeme H_0: $\vartheta = 0{,}01$ proti H_1: $\vartheta < 0{,}01$.

c) Byly provedeny změny v technologii výroby s cílem snížit náklady. V této situaci testujeme H_0: $\vartheta = 0{,}01$ proti H_1: $\vartheta > 0{,}01$.)

Definice chyby 1. a 2. druhu

Při testování H_0 proti H_1 se můžeme dopustit jedné ze dvou chyb: chyba 1. druhu spočívá v tom, že H_0 zamítneme, ač ve skutečnosti platí a chyba 2. druhu spočívá v tom, že H_0 nezamítneme, ač ve skutečnosti neplatí. Situaci přehledně znázorňuje tabulka:

skutečnost	rozhodnutí	
	H_0 nezamítáme	H_0 zamítáme
H_0 platí	správné rozhodnutí	**chyba 1. druhu**
H_0 neplatí	**chyba 2. druhu**	správné rozhodnutí

Pravděpodobnost chyby 1. druhu se značí **α** a nazývá se **hladina významnosti testu** (většinou bývá $\alpha = 0{,}05$, méně často 0,1 či 0,01). Pravděpodobnost chyby 2. druhu se značí **β**. Číslo **1–β** se nazývá **síla testu** a vyjadřuje pravděpodobnost, že bude H_0 zamítnuta za předpokladu, že neplatí. Obvykle se snažíme, aby síla testu byla aspoň 0,8. Obě hodnoty, α i 1–β, závisí na velikosti efektu, který se snažíme detekovat. Čím drobnější efekt, tím musí být větší rozsah náhodného výběru.

Poznámka: Testování nulové hypotézy proti alternativní hypotéze třemi způsoby.

Testování nulové hypotézy proti alternativní hypotéze lze provést pomocí kritického oboru, pomocí intervalu spolehlivosti nebo pomocí p-hodnoty.

Definice testového kritéria, oboru nezamítnutí, kritického oboru a kritických hodnot

Statistika $T_0 = T_0(X_1, ..., X_n)$ se nazývá testovým kritériem. Množina všech hodnot, jichž může testové kritérium nabýt, se rozpadá na obor nezamítnutí nulové hypotézy (značí se V) a obor zamítnutí nulové hypotézy (značí se W a nazývá se též kritický obor). Tyto dva obory jsou odděleny kritickými hodnotami (pro danou hladinu významnosti α je lze najít ve statistických tabulkách).

Rozhodnutí o nulové hypotéze pomocí realizace testového kritéria v oboru nezamítnutí či v kritickém oboru

Jestliže číselná realizace t_o testového kritéria T_o padne do kritického oboru W, pak nulovou hypotézu zamítáme na hladině významnosti α a znamená to skutečné vyvrácení testované hypotézy. Jestliže t_o padne do oboru nezamítnutí V, pak jde o pouhé mlčení, které platnost nulové hypotézy jenom připouští.

Stanovení kritického oboru v případě oboustranné alternativy, levostranné alternativy, pravostranné alternativy

Kritický obor v případě oboustranné alternativy má tvar

$W = (t_{\min}, K_{\alpha/2}(T)\rangle \cup \langle K_{1-\alpha/2}(T), t_{\max})$, kde $K_{\alpha/2}(T)$ a $K_{1-\alpha/2}(T)$ jsou kvantily rozložení, jímž se řídí testové kritérium T_0, je-li nulová hypotéza pravdivá.

Kritický obor v případě levostranné alternativy má tvar:

$W = (t_{\min}, K_{\alpha}(T)\rangle$.

Kritický obor v případě pravostranné alternativy má tvar:

$W = \langle K_{1-\alpha}(T), t_{\max})$.

Doporučený postup při testování nulové hypotézy proti alternativní hypotéze pomocí kritického oboru

- Stanovíme nulovou hypotézu a alternativní hypotézu. Přitom je vhodné zvolit jako alternativní hypotézu ten předpoklad, jehož přijetí znamená závažné opatření a mělo by k němu dojít jen s malým rizikem omylu.
- Zvolíme hladinu významnosti α. Zpravidla volíme $\alpha = 0{,}05$, méně často 0,1 nebo 0,01.
- Najdeme vhodné testové kritérium a na základě zjištěných dat vypočítáme jeho realizaci.
- Jestliže realizace testového kritéria padla do kritického oboru, nulovou hypotézu zamítáme na hladině významnosti α a přijímáme alternativní hypotézu. V opačném případě nulovou hypotézu nezamítáme na hladině významnosti α.
- Na základě rozhodnutí, které jsme učinili o nulové hypotéze, učiníme nějaké konkrétní opatření, např. seřídíme obráběcí stroj.
- (Při testování hypotéz musíme mít k dispozici odpovídající nástroje, nejlépe vhodný statistický software. Nemáme-li ho k dispozici, musíme znát příslušné vzorce. Dále potřebujeme statistické tabulky a kalkulačku.)

Testování nulové hypotézy proti alternativní hypotéze pomocí 100(1-α)% empirického intervalu spolehlivosti pro parametrickou funkci h(ϑ)

Sestrojíme 100(1-α)% empirický interval spolehlivosti pro parametrickou funkci h(ϑ). Pokryje-li tento interval hodnotu c, pak H_0 nezamítáme na hladině významnosti α, v opačném případě H_0 zamítáme na hladině významnosti α.

Pro test H_0 proti oboustranné alternativě sestrojíme oboustranný interval spolehlivosti.

Pro test H_0 proti levostranné alternativě sestrojíme pravostranný interval spolehlivosti.

Pro test H_0 proti pravostranné alternativě sestrojíme levostranný interval spolehlivosti.

Testování nulové hypotézy proti alternativní hypotéze pomocí p-hodnoty

p-hodnota udává nejnižší možnou hladinu významnosti pro zamítnutí nulové hypotézy. Je to riziko, že bude zamítnuta H_0 za předpokladu, že platí (riziko planého poplachu). Jestliže p-hodnota $\leq \alpha$, pak H_0 zamítáme na hladině významnosti α, je-li p-hodnota $> \alpha$, pak H_0 nezamítáme na hladině významnosti α.

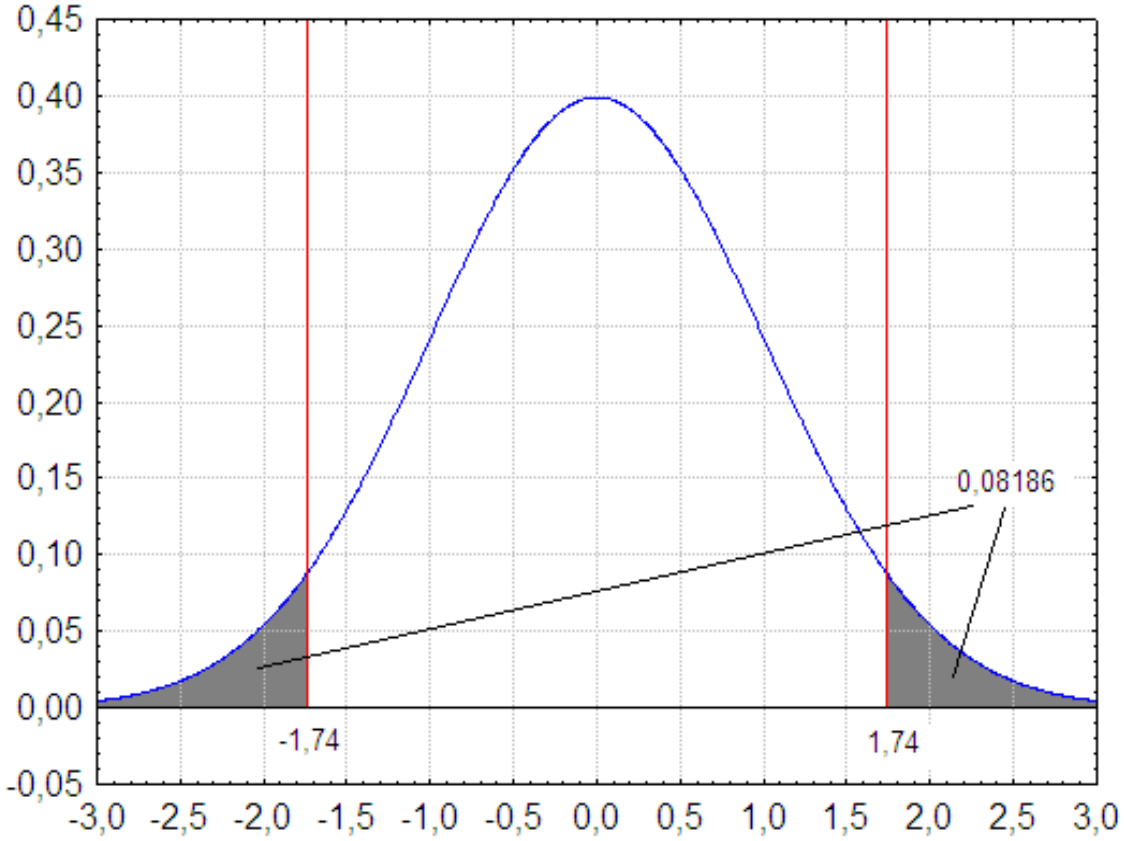
Způsob výpočtu p-hodnoty:

- Pro oboustrannou alternativu $p = 2 \min\{P(T_0 \leq t_0), P(T_0 \geq t_0)\}$.
- Pro levostrannou alternativu $p = P(T_0 \leq t_0)$.
- Pro pravostrannou alternativu $p = P(T_0 \geq t_0)$.

(p-hodnota vyjadřuje pravděpodobnost, s jakou číselné realizace $x_1, ..., x_n$ náhodného výběru $X_1, ..., X_n$ podporují H_0, je-li pravdivá. Statistické programové systémy poskytují ve svých výstupech p-hodnotu. Její výpočet vyžaduje znalost distribuční funkce rozložení, kterým se řídí testové kritérium T_0, je-li H_0 pravdivá. Vzhledem k tomu, že v běžných statistických tabulkách jsou uvedeny pouze hodnoty distribuční funkce standardizovaného normálního rozložení, bez použití speciálního software jsme schopni vypočítat p-hodnotu pouze pro test hypotézy o střední hodnotě normálního rozložení při známém rozptylu.)

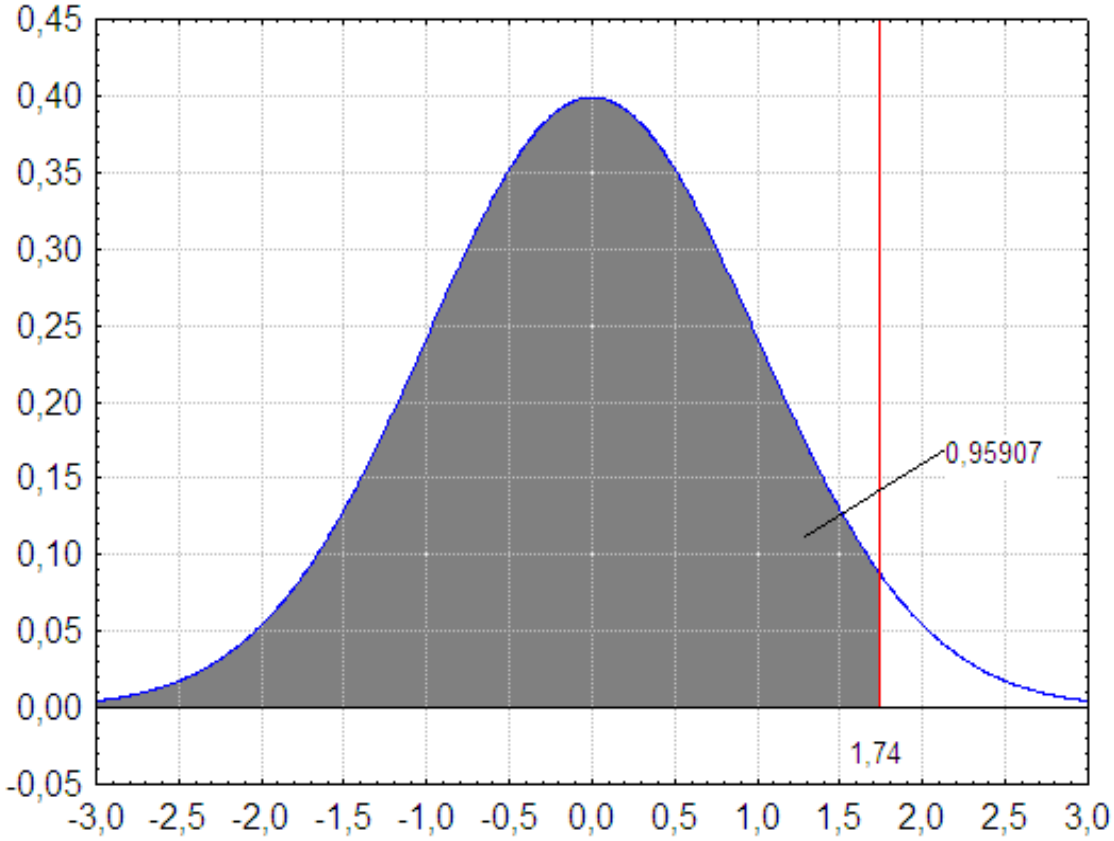
Ilustrace významu p-hodnoty (1)

Oboustranný test:



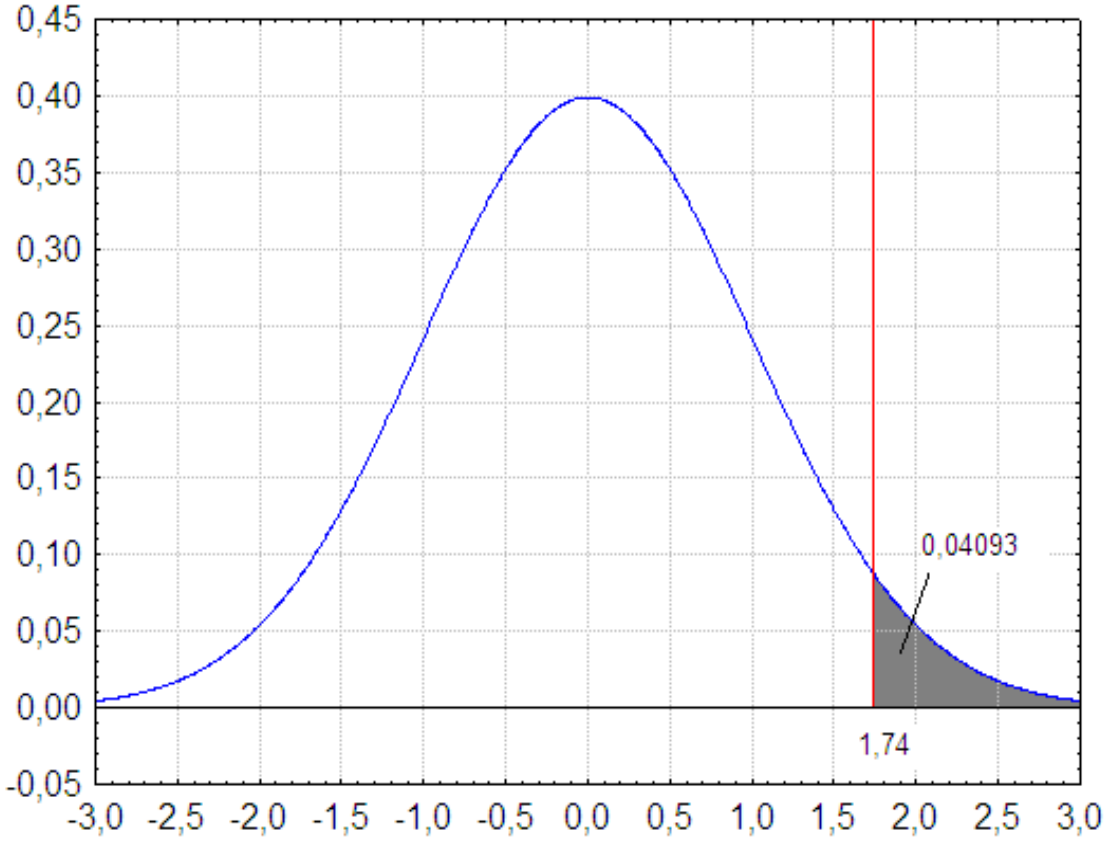
Ilustrace významu p-hodnoty (2)

Levostranný test:



Ilustrace významu p-hodnoty (3)

Pravostranný test:



Příklad (1)

Nechť $X_1, ..., X_{400}$ je náhodný výběr z $N(\mu,0{,}01)$. Je známo, že výběrový průměr se realizoval hodnotou 0,01. Na hladině významnosti 0,05 testujte hypotézu H_0: $\mu = 0$ proti pravostranné alternativě H_1: $\mu > 0$

a) pomocí intervalu spolehlivosti
b) pomocí kritického oboru
c) pomocí p-hodnoty.

Řešení:

ad a) Při testování nulové hypotézy proti pravostranné alternativě používáme levostranný interval spolehlivosti.

$$d = m - \frac{\sigma}{\sqrt{n}} u_{1-\alpha} = 0{,}01 - \frac{0{,}1}{\sqrt{400}} u_{0{,}95} = 0{,}01 - \frac{0{,}1}{20} 1{,}64485 = 0{,}0018.$$

Protože číslo c = 0 neleží v intervalu $(0{,}0018; \infty)$, H_0 zamítáme na hladině významnosti 0,05.

Příklad (2)

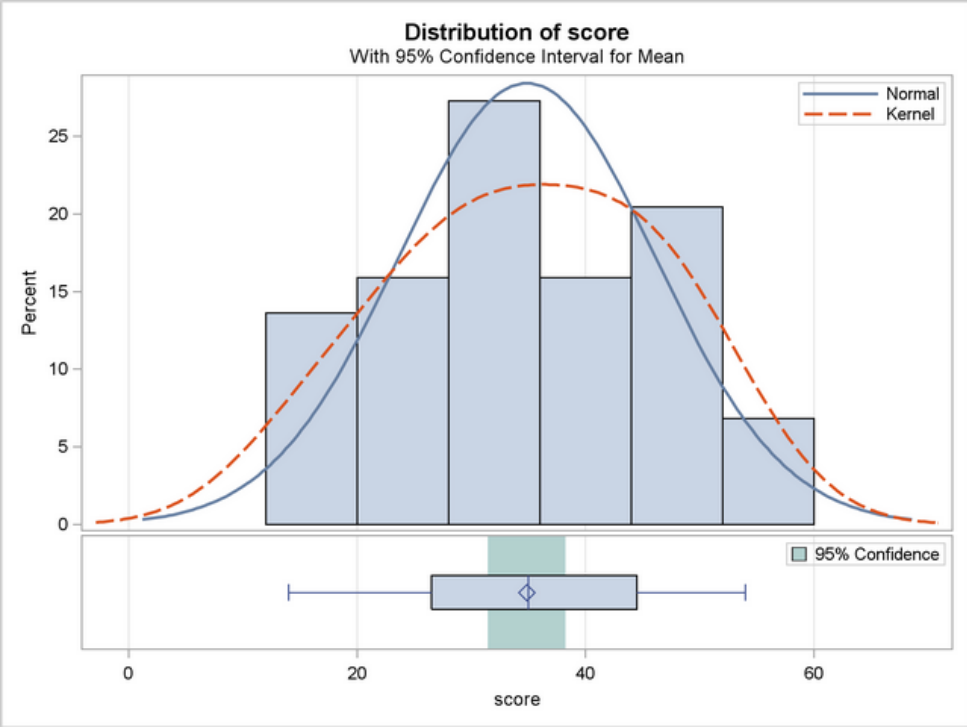
ad b) Vypočteme realizaci testové statistiky: $t_0 = \frac{m-c}{\frac{\sigma}{\sqrt{n}}} = \frac{0{,}01-0}{\frac{0{,}1}{\sqrt{400}}} = \frac{0{,}01 \cdot 20}{0{,}1} = 2$

Stanovíme kritický obor: $W = \langle u_{1-\alpha}, \infty) = \langle u_{0{,}95}, \infty) = \langle 1{,}64485, \infty)$

Protože testová statistika se realizuje v kritickém oboru, H_0 zamítáme na hladině významnosti 0,05.

ad c) Při testování nulové hypotézy proti pravostranné alternativě se p-hodnota počítá podle vzorce: $p = P(T_0 \geq t_0)$. V našem případě: $p = P(T_0 \geq 2) = 1 - \Phi(2) = 1 - 0{,}97725 = 0{,}02275$. Protože p-hodnota je menší než hladina významnosti 0,05, H_0 zamítáme na hladině významnosti 0,05.

12. Testování hypotéz v MS Excel a SAS



Inference about the Slope (pro regresní přímku): t Test

- t test for a population slope
  - Is there a linear relationship between x and y?
- Null and alternative hypotheses
  - H_0: $\beta_1 = 0$ (no linear relationship)
  - H_1: $\beta_1 \neq 0$ (linear relationship does exist)
- Test statistic
  - $$t = \frac{b_1 - \beta_1}{s_{b_1}}$$
  - $$d.f. = n - 2$$

where:

b_1 = Sample regression slope coefficient

β_1 = Hypothesized slope

s_{b1} = Estimator of the standard error of the slope

Inference about the Slope: t Test

(continued)

House Price in $1000s (y)	Square Feet (x)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

Estimated Regression Equation:

$$\widehat{\text{houseprice}} = 98.25 + 0.1098(\text{sq.ft.})$$

The slope of this model is 0.1098

Does square footage of the house affect its sales price?

Inferences about the Slope: t Test Example

Test Statistic: **t = 3.329**

$H_o: \beta_1 = o$

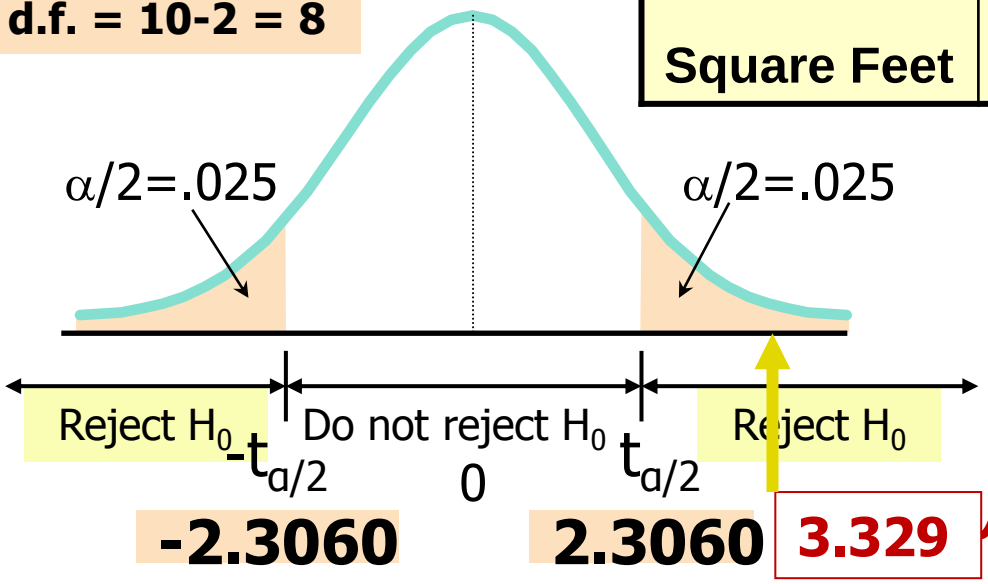
$H_A: \beta_1 \neq o$

From Excel output:

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	98.24833	58.03348	1.69296	0.12892
Square Feet	0.10977	0.03297	3.32938	0.01039

b_1 S_{b_1} t

d.f. = 10-2 = 8



Decision: Reject H_0 **0.01<0.05**

Conclusion:

There is sufficient evidence that square footage affects house price

Regression Analysis for Description

Confidence Interval Estimate of the Slope:

$$b_1 \pm t_{\alpha/2} s_{b_1}$$

d.f. = n - 2

Excel Printout for House Prices:

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	**98.24833**	**58.03348**	**1.69296**	**0.12892**	**-35.57720**	**232.07386**
Square Feet	**0.10977**	**0.03297**	**3.32938**	**0.01039**	**0.03374**	**0.18580**

At 95% level of confidence, the confidence interval for the slope is (0.0337, 0.1858)

Regression Analysis for Description

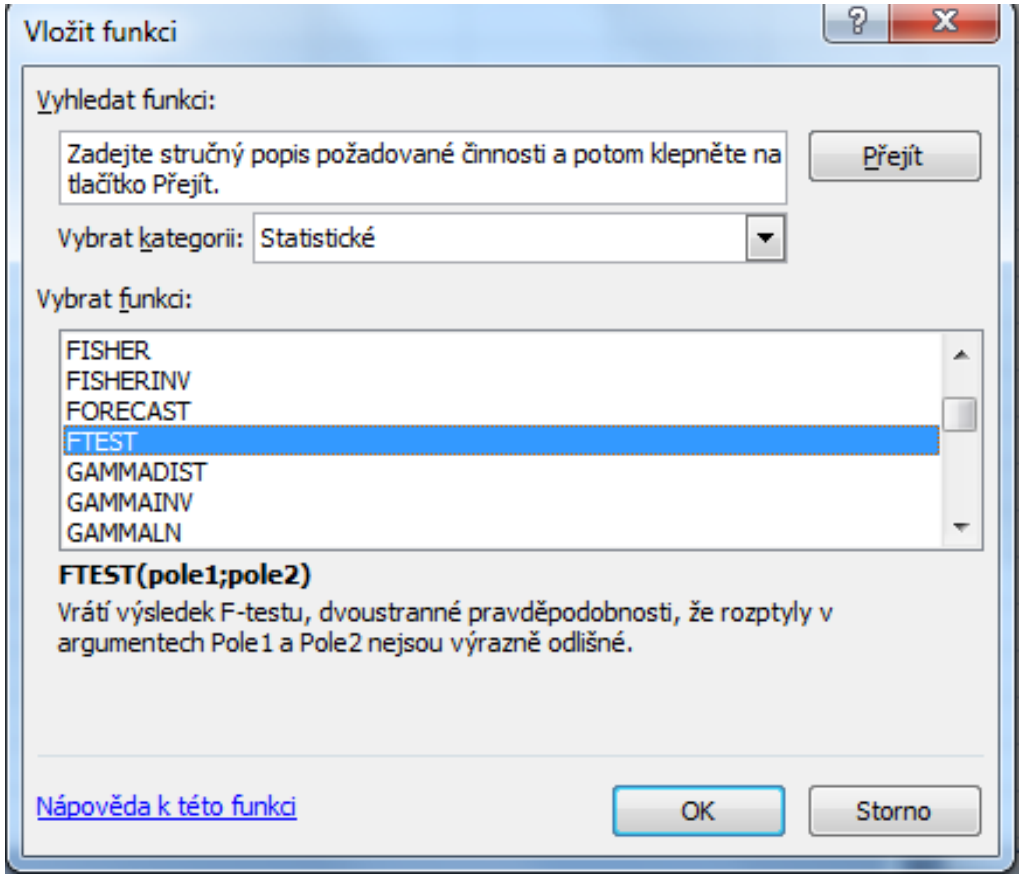
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	**98.24833**	**58.03348**	**1.69296**	**0.12892**	**-35.57720**	**232.07386**
Square Feet	**0.10977**	**0.03297**	**3.32938**	**0.01039**	**0.03374**	**0.18580**

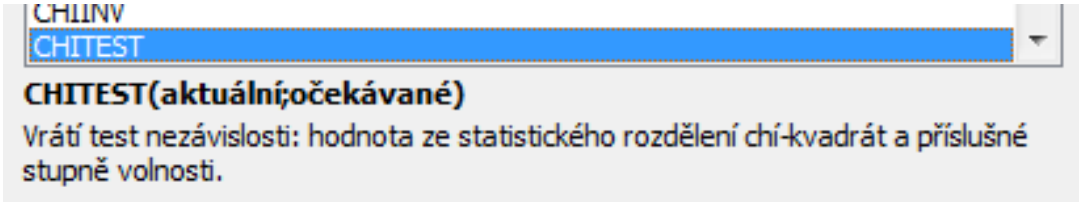
Since the units of the house price variable is $1000s, we are 95% confident that the average impact on sales price is between $33.70 and $185.80 per square foot of house size

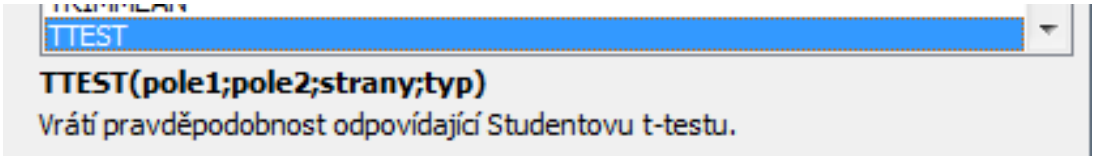
This 95% confidence interval does not include 0.

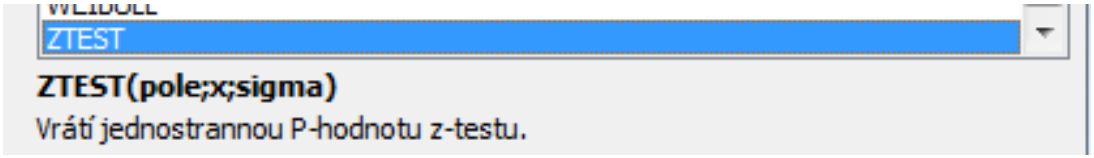
Conclusion: There is a significant relationship between house price and square feet at the .05 level of significance

Testy v MS Excel



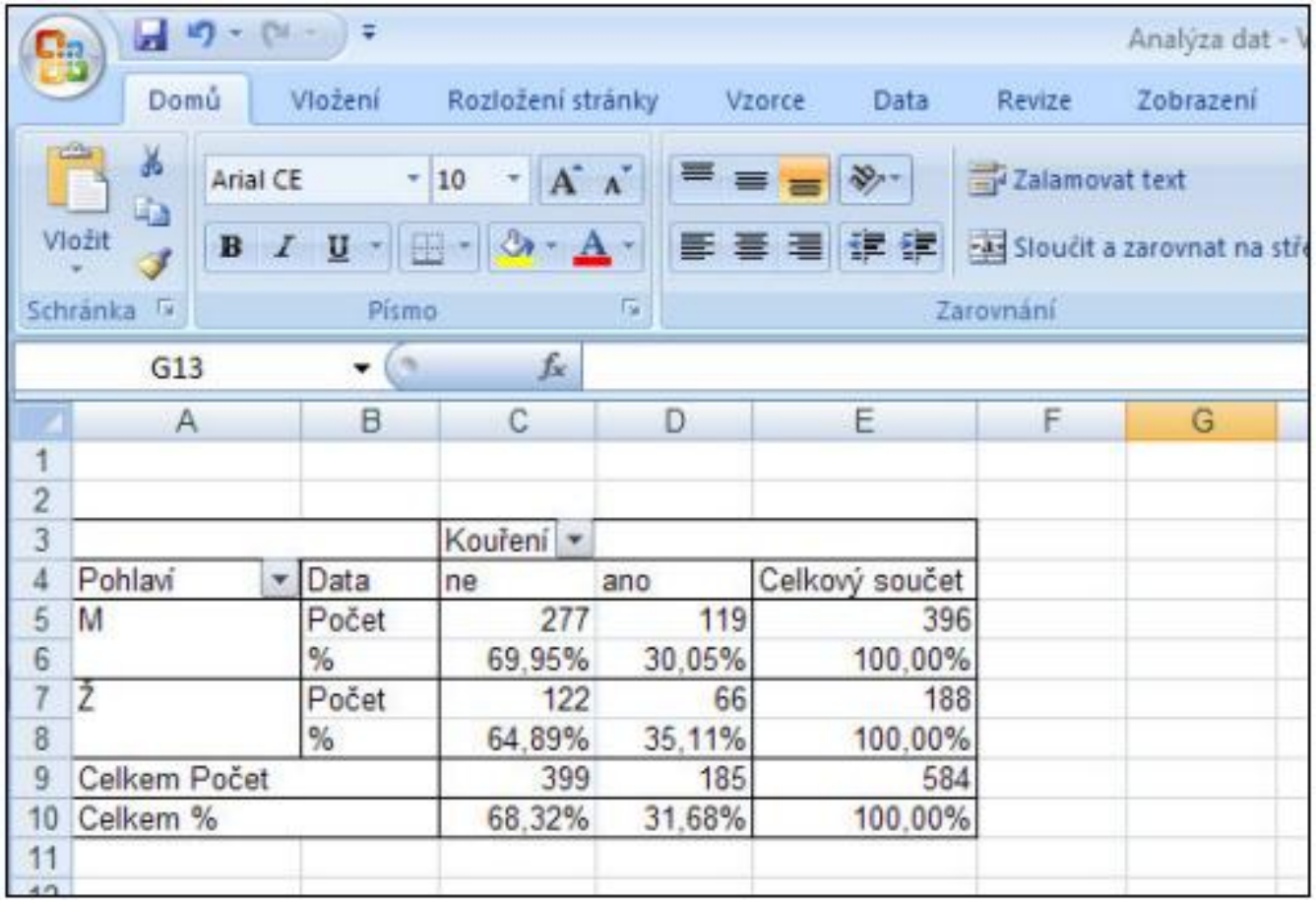






Testy v MS Excel

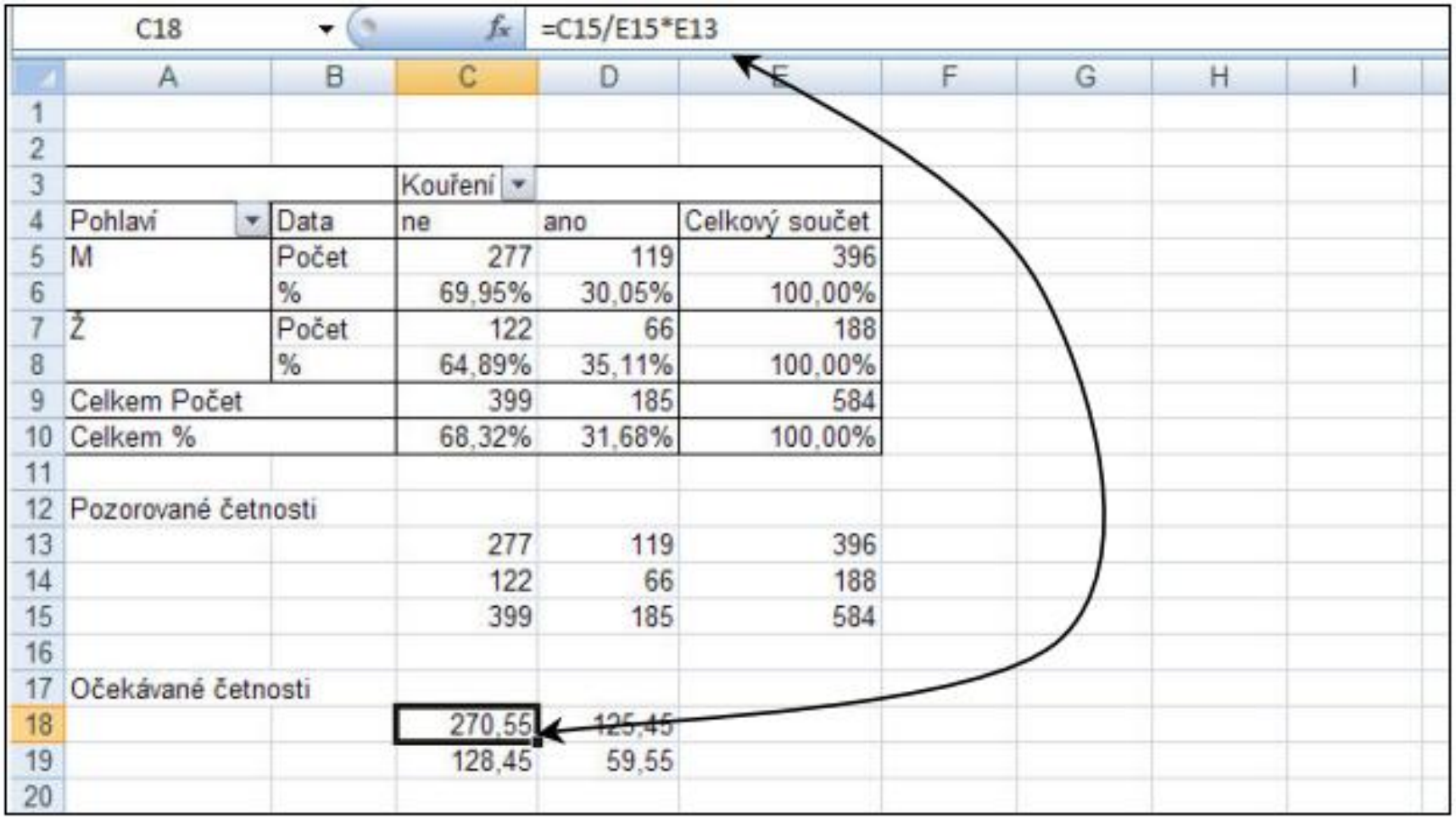
Nulová hypotéza: Podíl kuřáků je stejný u mužů i žen.
Alternativní hypotéza: Podíl kuřáků u mužů a u žen se liší.



	A	B	C	D	E	F	G
1							
2							
3			Kouření				
4	Pohlavi	Data	ne	ano	Celkový součet		
5	M	Počet	277	119	396		
6		%	69,95%	30,05%	100,00%		
7	Ž	Počet	122	66	188		
8		%	64,89%	35,11%	100,00%		
9	Celkem Počet		399	185	584		
10	Celkem %		68,32%	31,68%	100,00%		
11							

Zdroj: http://ulb.upol.cz/praktikum/statistika3.pdf

Testy v MS Excel



C18 =C15/E15*E13

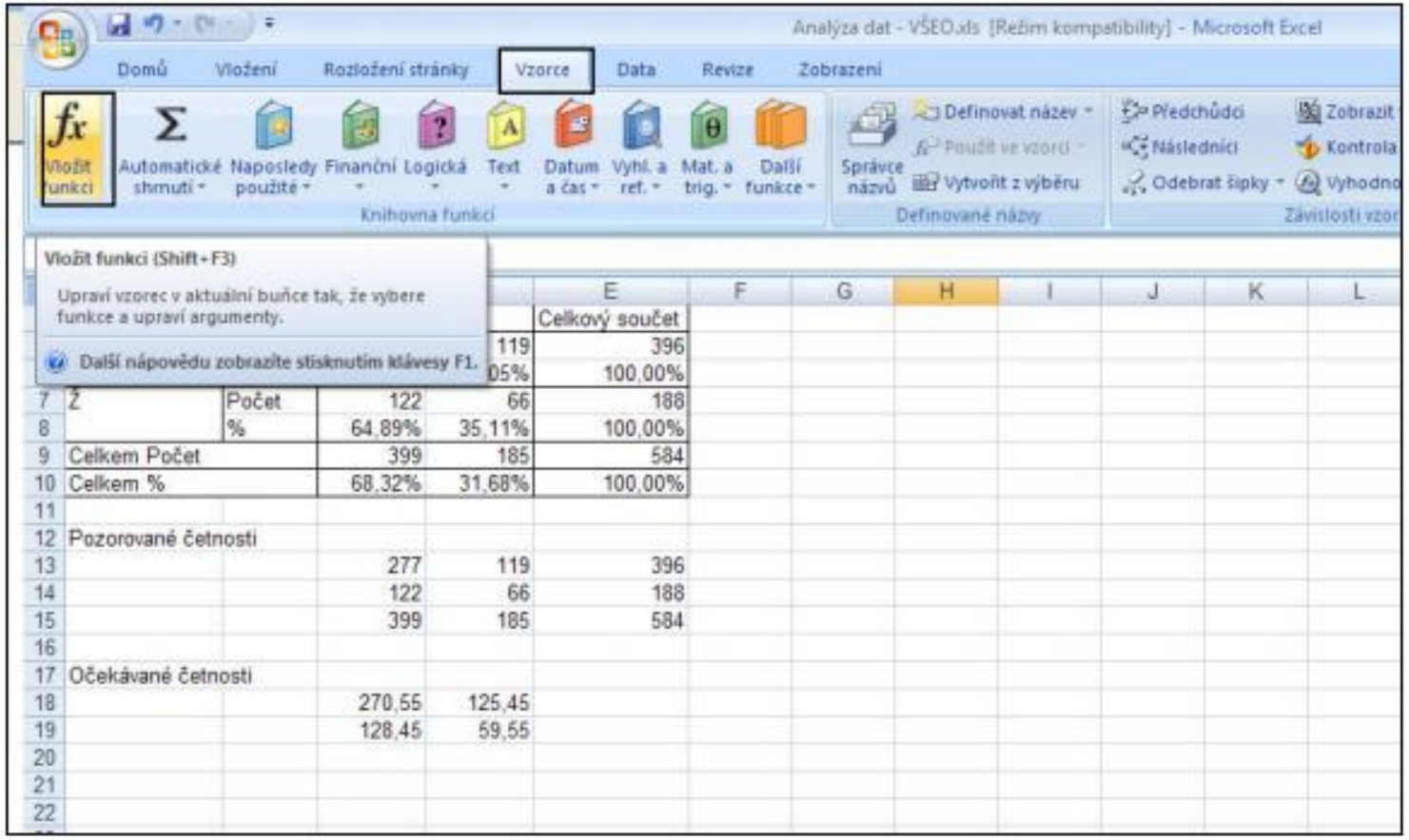
	A	B	C	D	E	F	G	H	I
1									
2									
3			Kouření						
4	Pohlaví	Data	ne	ano	Celkový součet				
5	M	Počet	277	119	396				
6		%	69,95%	30,05%	100,00%				
7	Ž	Počet	122	66	188				
8		%	64,89%	35,11%	100,00%				
9	Celkem Počet		399	185	584				
10	Celkem %		68,32%	31,68%	100,00%				
11									
12	Pozorované četnosti								
13			277	119	396				
14			122	66	188				
15			399	185	584				
16									
17	Očekávané četnosti								
18			270,55	125,45					
19			128,45	59,55					
20									

Testy v MS Excel

K výpočtu dosažené hladiny statistické významnosti, neboli signifikance (tzv. *p*-hodnoty), použijeme funkci **CHITEST**.

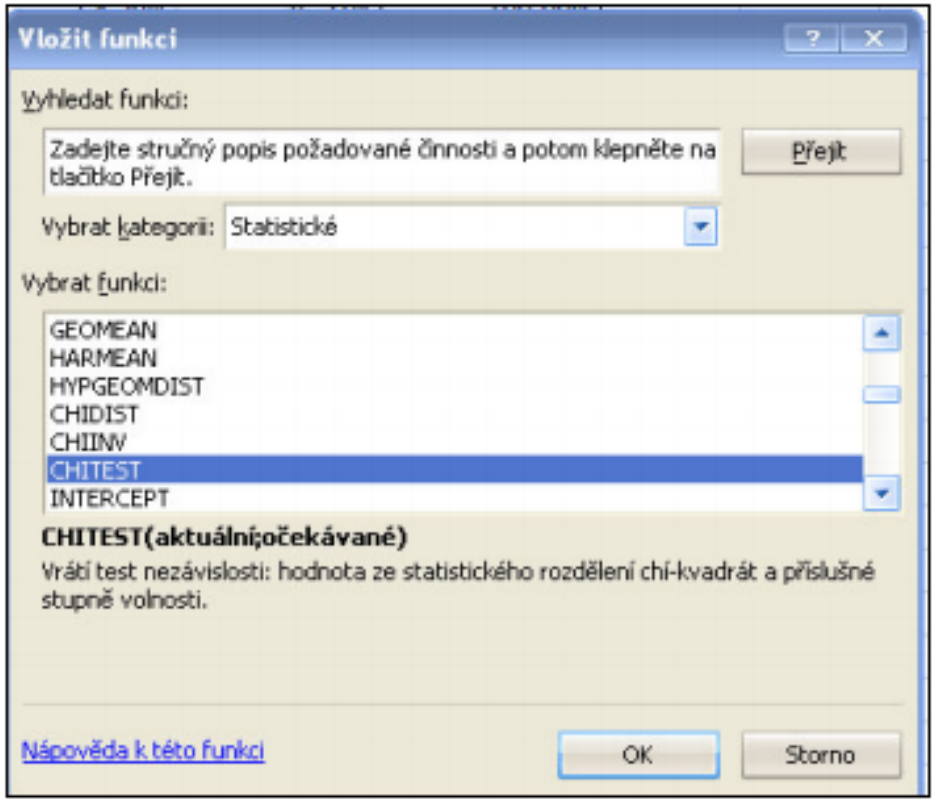
Klikněte do buňky, kam chcete umístit hodnotu signifikance (např. do buňky E21).

Z řádkového menu zvolte **Vzorce** a klikněte na ikonu **Vložit funkci**.



					Celkový součet
				119	396
				05%	100,00%
7	Ž	Počet	122	66	188
8		%	64,89%	35,11%	100,00%
9	Celkem Počet		399	185	584
10	Celkem %		68,32%	31,68%	100,00%
11					
12	Pozorované četnosti				
13			277	119	396
14			122	66	188
15			399	185	584
16					
17	Očekávané četnosti				
18			270,55	125,45	
19			128,45	59,55	
20					
21					
22					

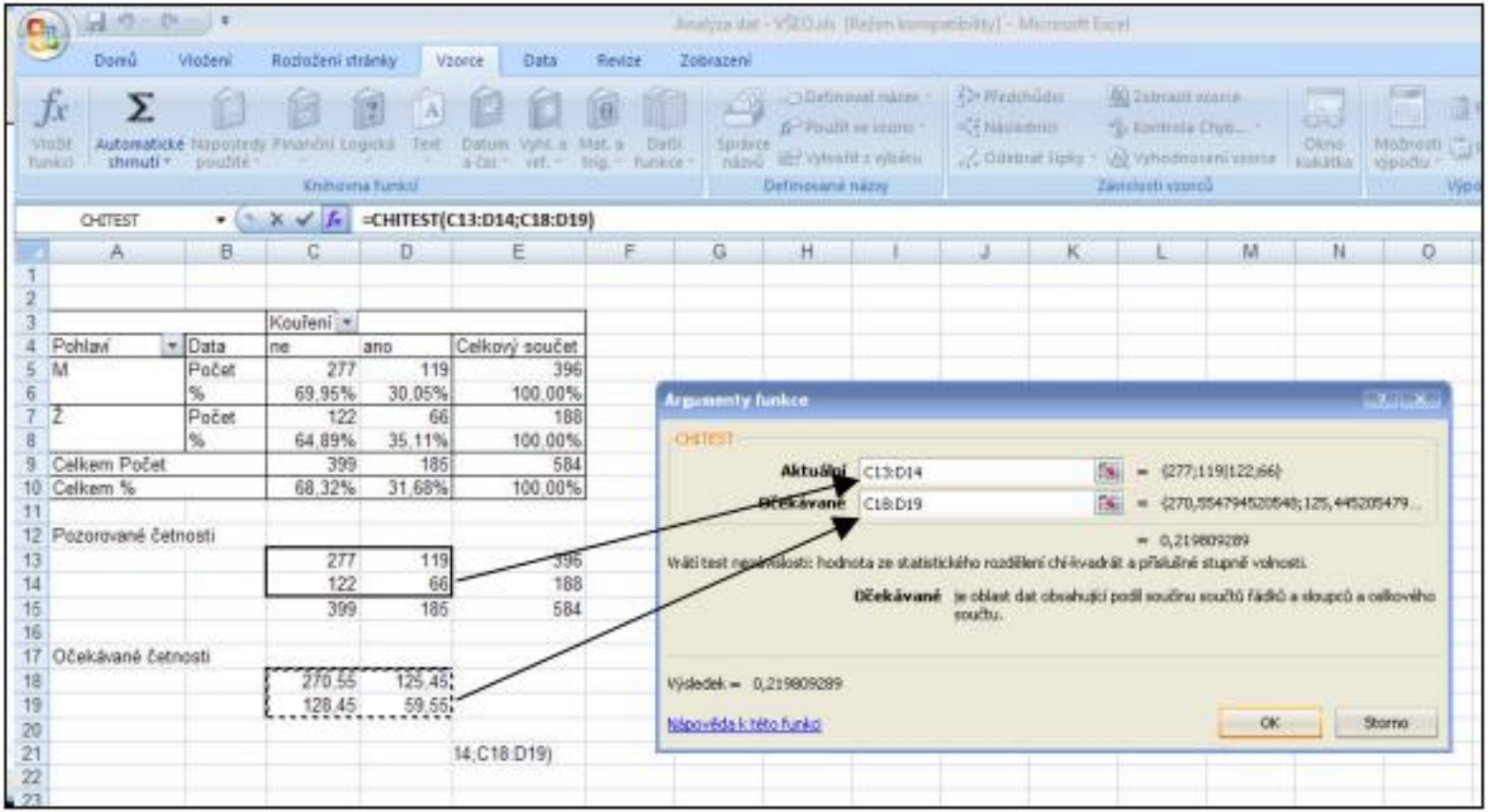
Testy v MS Excel



Otevřete dialogové okno **Argumenty funkce**. Do pole **Aktuální** zadejte adresu oblasti buněk s pozorovanými četnostmi C13:D14 (pouze čtyři hodnoty!).

Testy v MS Excel

Do pole **Očekávané** zadejte adresu oblasti buněk s vypočítanými očekávanými četnostmi C18:D19 (také čtyři hodnoty).



Testy v MS Excel

Klikněte na **OK.**

Tabulky s výslednou hodnotou signifikance:

	A	B	C	D	E	F
1						
2						
3			Kouření			
4	Pohlaví	Data	ne	ano	Celkový součet	
5	M	Počet	277	119	396	
6		%	69,95%	30,05%	100,00%	
7	Ž	Počet	122	66	188	
8		%	64,89%	35,11%	100,00%	
9	Celkem Počet		399	185	584	
10	Celkem %		68,32%	31,68%	100,00%	
11						
12	Pozorované četnosti					
13			277	119	396	
14			122	66	188	
15			399	185	584	
16						
17	Očekávané četnosti					
18			270,55	125,45		
19			128,45	59,55		
20						
21	**Signifikance chí-kvadrát testu:**				**0,220**	
22						

Před vypočítanou hodnotu (např. do buňky A21) napište text „Signifikance chí-kvadrát testu:“ Hodnotu signifikance zaokrouhlete na 3 desetinná místa.

Funkce chí-kvadrát test v Excelu nezobrazuje hodnotu testového kritéria χ^2, zobrazí pouze *p*-hodnotu.

Výsledek, tedy dosaženou hladinu statistické významnosti, porovnáme s hodnotou 0,05.
Je-li dosažená hladina statistické významnosti menší než 0,05, nulovou hypotézu
zamítáme, v opačném případě nulovou hypotézu zamítnout nemůžeme. V tomto příkladu
p = 0,220, nulovou hypotézu tedy zamítnout nemůžeme.
Závěr testování zní: Podíl kuřáků je stejný v populaci mužů i žen.

Introduction

Suppose you want to answer the following questions:

- Does a new headache medicine provide the typical time to relief of 100 minutes, or is it different?
- Does a weekend training session have an effect on performance on an exam?
- Does a new headache medicine differ in time to relief from a standard headache treatment?

Zdroj: http://www.biostat.umn.edu/~susant/Lab6415/Lab3.ppt

T-tests can be used to answer all of these questions.

There are three main types of t-tests:

1. One-sample
2. Matched Pairs
3. Two-sample

One-Sample T-test

A one-sample t-test is used to compare a sample to an average or general population. You may know the average height of men in the U.S., and you could test whether a sample of professional basketball players differ significantly in height from the general U.S. population. A significant difference would indicate that basketball players belong to a different distribution of heights than the general U.S. population.

Matched Pairs T-test

A matched pairs t-test usually involves the same subjects being measured on some factor at two points in time. For example, subjects could be tested on short-term memory, receive a brief tutorial on memory aids, then have their short-term memory re-tested. A significant difference in score (after-before) would indicate that the tutorial had an effect.

Two-Sample T-test

A two-sample t-test compares two groups on some factor. For example, one group could receive an experimental treatment and the second group could receive a standard of care treatment or placebo.

Notice that in a two-sample t-test, two distinct groups are being compared, as opposed to the one-sample, where one group is compared to a general average, or a matched-pairs, where only one group is being measured twice.

One-sample T-test in SAS

We want to test whether a new headache medicine provides a relief time equal to or different from the standard of 100 minutes.

Ho: $\mu=100$

Ha: $\mu\neq100$

We have 10 observations of time to relief. Before we can test our hypothesis, however, we have to test the data for normality.

Type the following code in SAS:

```
DATA relieftime;
      INPUT relief;
DATALINES;
90
93
93
99
98
100
103
104
99
102
;
PROC UNIVARIATE DATA = relieftime normal plot;
      VAR relief;
      histogram relief / midpoints = 80 to 120 by 5 normal;
RUN;
```

Tests for Normality

- The histogram shows most observations falling at the peak of the normal curve.
- The box-plot shows that the mean falls on the median (*--+--*), indicating no skewed data.
- The formal tests of normality in the output are non-significant, indicating these data come from a normal distribution.
- We can assume the data are normally distributed and proceed with the one-sample t-test.

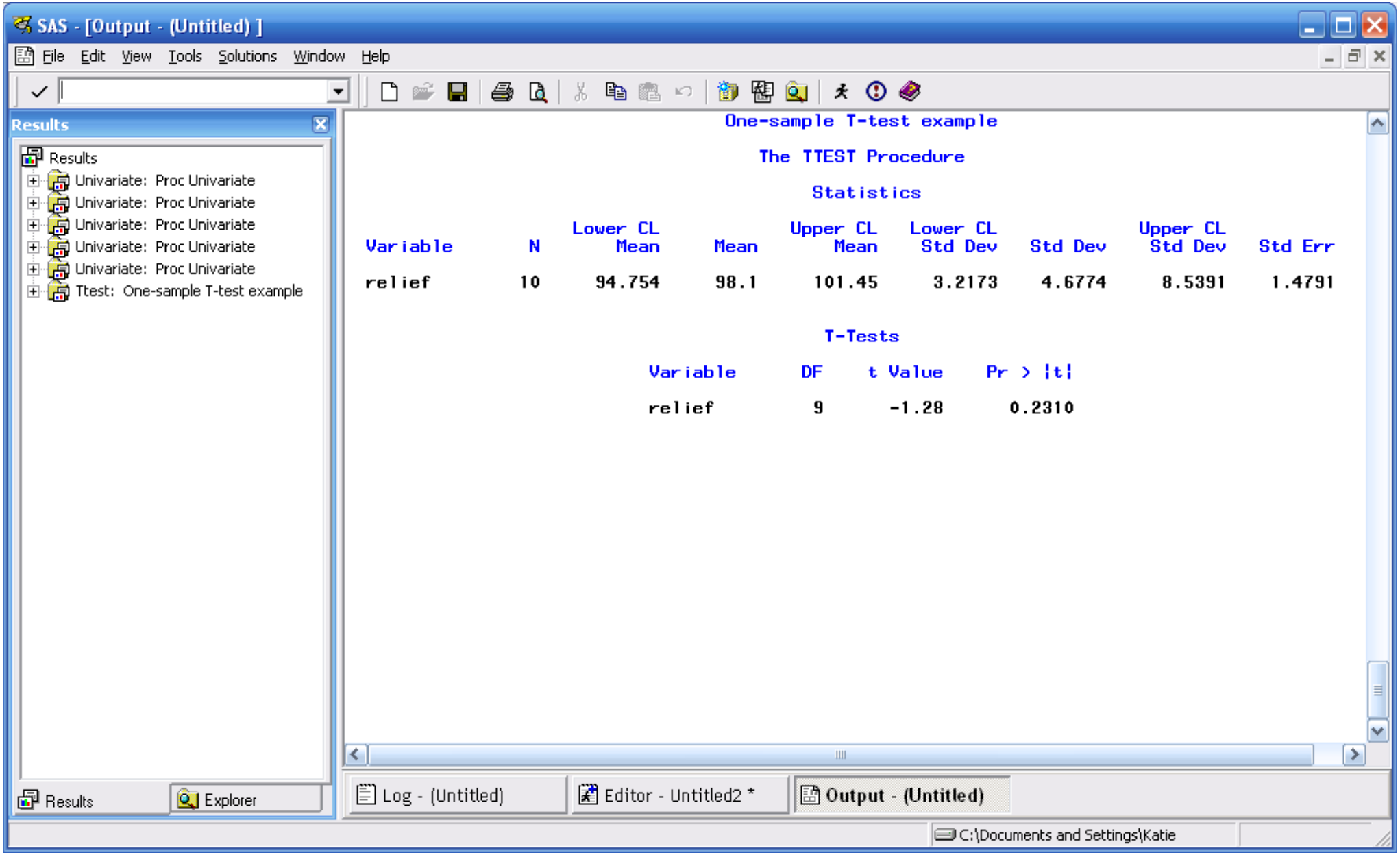
SAS Code for a One-Sample T-test

```
PROC TTEST DATA = relieftime ho=100;
   TITLE 'One-sample T-test example' ;
   VAR relief;
RUN;
```

The code is telling SAS to run a t-test procedure on the variable relief, and the mean value of relief should be compared to a null value of 100.

After running this program, check your log for errors, then look at the output.

SAS Output for One-sample T-test



Interpreting Output

From the SAS output, you can see that the mean relief time of the 10 subjects is 98.1 minutes. The calculated t* value = -1.28, and this test statistic has a p-value of 0.23 (this value is found under the label "Pr > |t|" which stands for the probability of getting a value greater than the absolute value of t*). This is a two-sided test. If this were a one-sided test, you would simply divide the p-value by 2.

Conclusion

If alpha = 0.05, then our p-value is greater than alpha. Therefore, we fail to reject the null hypothesis. The new headache medicine does not provide a different time to relief from 100 minutes.

Matched Pairs T-test in SAS

We want to determine whether a weekend study session improves students' test scores. Six students are given a math test before the session, then they are re-tested after the weekend training. This is a matched pairs t-test, because the same subjects are being measured before and after some intervention.

Ho: $\mu_{before} = \mu_{after}$

Ha: $\mu_{before} \neq \mu_{after}$

Again, before we can analyze the data, we have to determine whether we can assume the data come from a normal distribution.

Type the following code into SAS and run the program

```
DATA study;
      INPUT before after;
DATALINES;
90 95
87 92
100 104
80 89
95 101
90 105
;
PROC UNIVARIATE DATA = study normal plot;
      VAR before after;
      histogram before after / normal;
RUN;
```

Tests for Normality

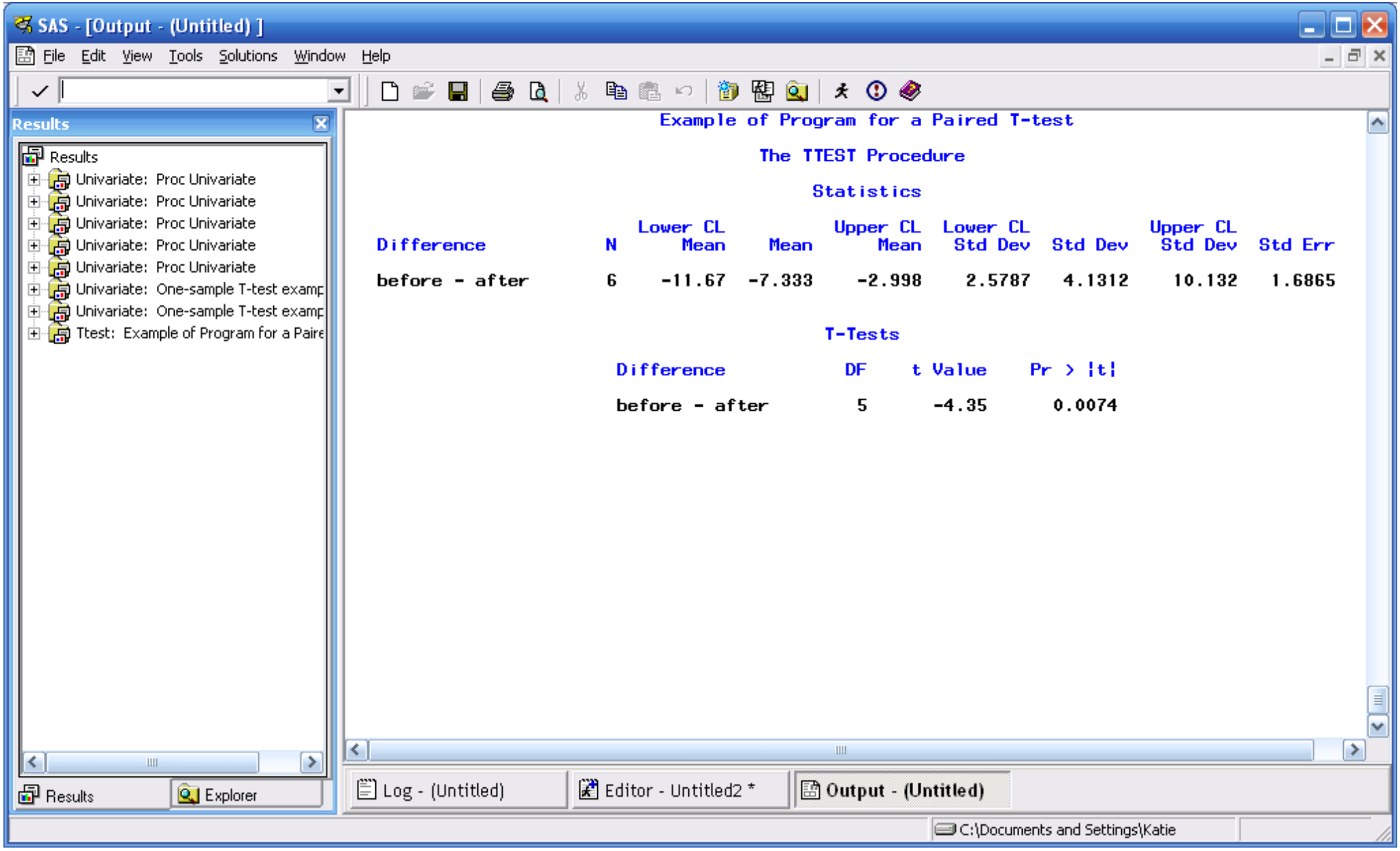
- There are so few data points that the histograms are difficult to interpret.
- The box-plots for before and after both show the mean very close to the median, suggesting the data are not skewed.
- The tests of normality for before and after have p-values > alpha, indicating we do not reject the assumption of normality.
- We can proceed with the matched pairs t-test.

SAS Code for Matched Pairs T-test

```
PROC TTEST DATA = study;
   TITLE "Example of Program for a Paired T-test" ;
   PAIRED before * after;
RUN;
```

The code tells SAS to do a paired t-test on the data set study, and it will compare the difference of the means between before and after.

SAS Output of a Matched Pairs T-test



Interpreting Output

The difference of the mean score (d-bar: before-after) is -7.33; on average the scores before the weekend were lower than the scores after the training session. (If in your paired statement you had typed "after*before" the average difference would be 7.33.)

Is this difference statistically significant? To answer that question, look at the p-value. The t* for the test is -4.35, and the p-value is 0.0074.

Conclusion

If alpha = 0.05, then the p-value < alpha, and we reject the null hypothesis. Therefore, we can conclude that average scores are different before and after the weekend session, and the training does improve test scores.

Two-Sample T-test in SAS

We want to determine whether a new headache medicine provides a different time to relief than a control medicine. Two groups of five subjects each are either given the treatment or control.

Ho: $\mu_1 = \mu_2$

Ha: $\mu_1 \neq \mu_2$

Before we can conduct the two-sample t-test, however, we must determine whether the data come from a normal distribution.

Type the following code into SAS and run the program:

```
DATA response;
   INPUT group $ time;
DATALINES;
c 80
c 93
c 83
c 89
c 98
t 100
t 103
t 104
t 99
t 102
;
PROC UNIVARIATE DATA = response normal plot;
   class group;
   var time;
   histogram time / midpoints = 80 to 120 by 5 normal;
RUN;
```

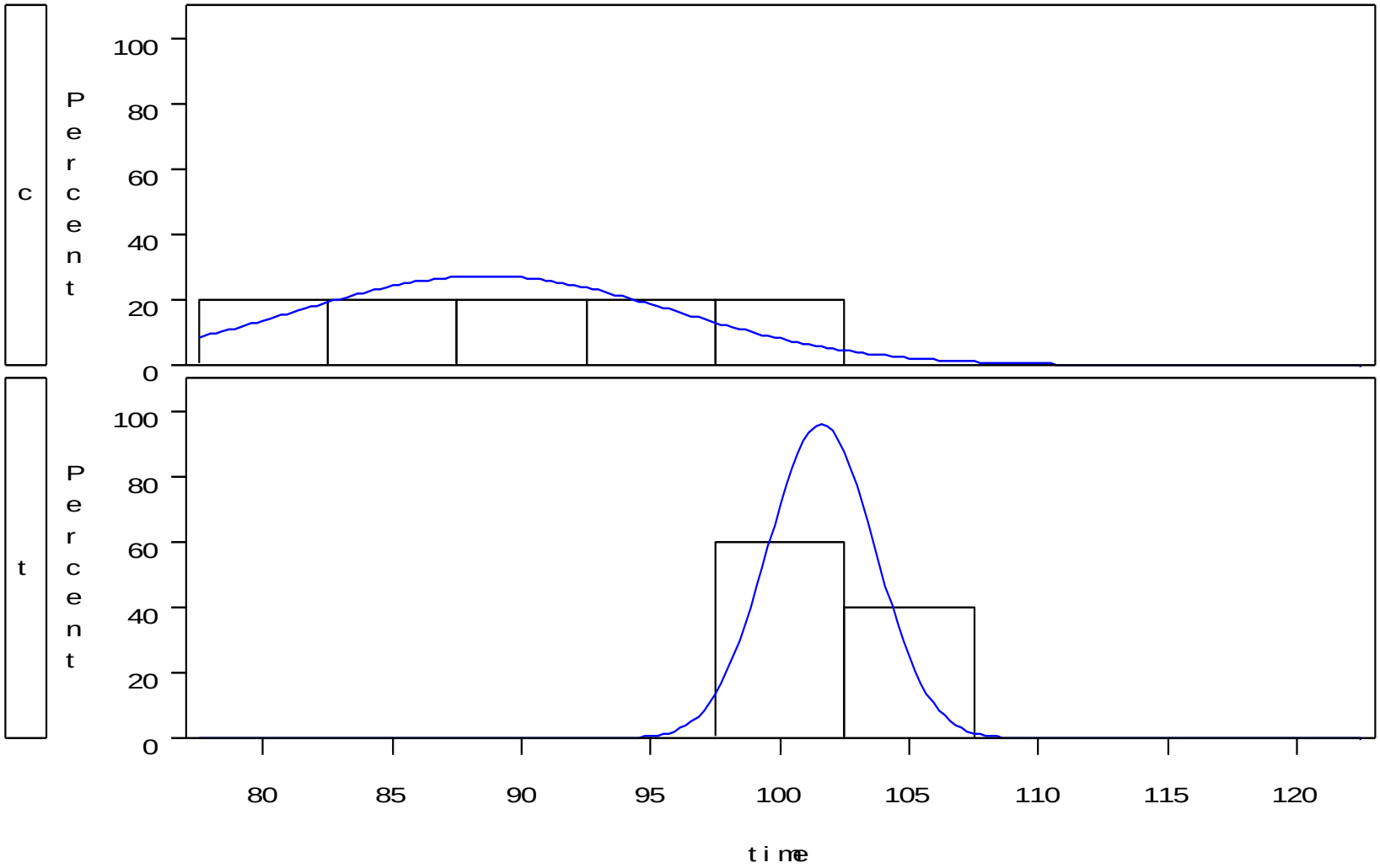
A few notes:

- Notice the variable “group” is followed by a “$” because it is a categorical variable
- The code has specified that the univariate procedure be performed on the variable *time*, but that it is done by the *class* “group.” This way you will have separate summary statistics, plots and histograms for the treatment and control groups.

Tests for Normality

- The tests for normality for both the treatment and control groups are non-significant (p-value > alpha), indicating we can assume they come from a normal distribution.
- Because each group only has 5 subjects, the histograms are difficult to interpret, but there is no indication of non-normality.
- Proceed with the two-sample t-test

Histograms for control and treatment groups

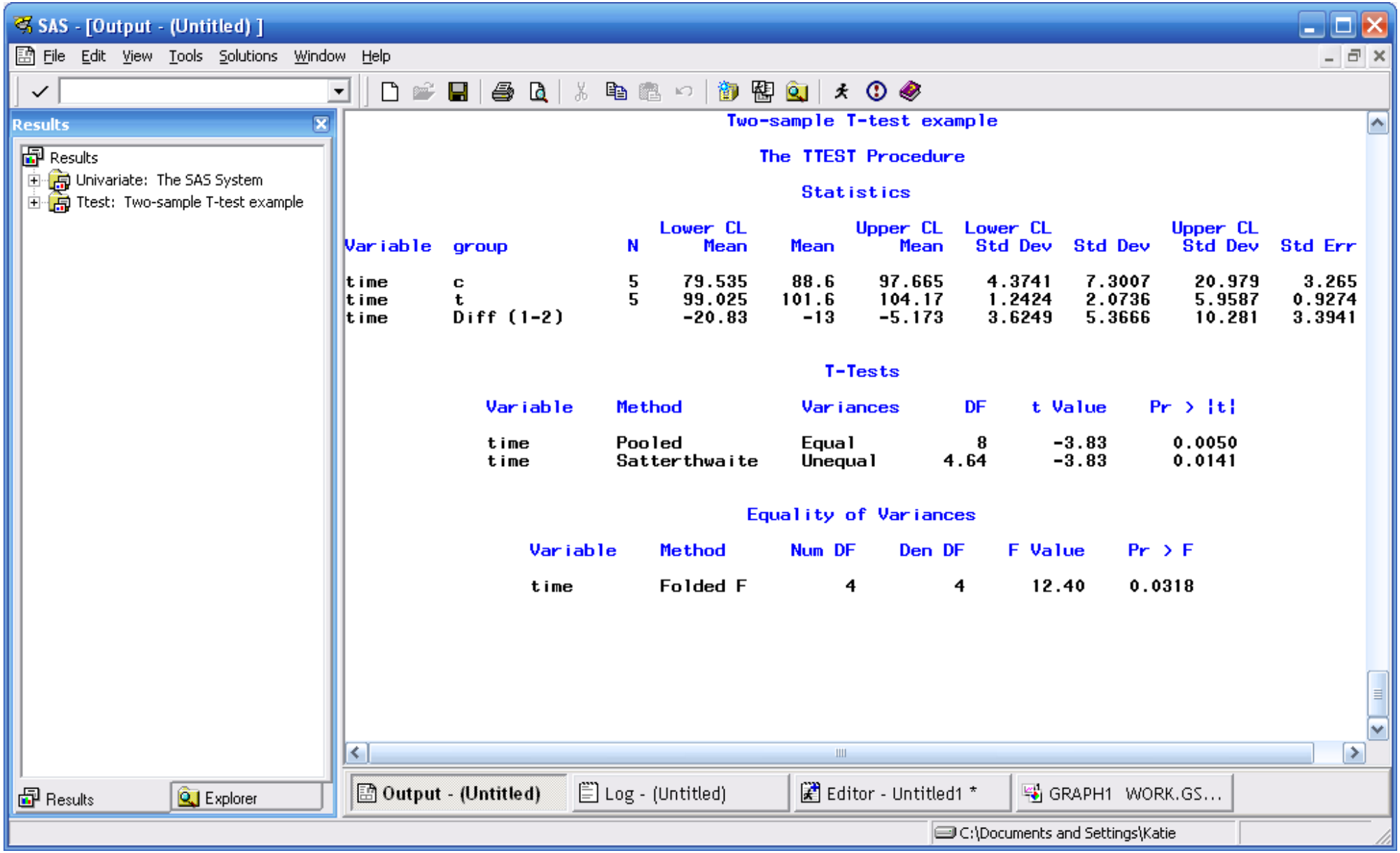


SAS Code for Two-Sample T-test

```
PROC TTEST DATA = response;
   TITLE 'Two-sample T-test example';
   class group;
   var time;
RUN;
```

- Notice for a two-sample t-test you must specify what distinguishes the two samples; in this case we compare the two samples defined by “group” (treatment and control), and we tell SAS to compare their mean “time” to relief.

SAS Output for a Two-Sample T-test



Interpreting the Output: Pooled vs. Unpooled Variance

Before you can interpret your test statistic and reach a conclusion, you must determine whether to use the **pooled** or **unpooled** variances test statistic. If we can assume the two samples have *equal variances*, then we use the *pooled* t*. If, on the other hand, we determine that the two samples have *unequal variances*, then we must use the *unpooled* t*.

SAS conducts a formal F-test to determine whether the two groups have equal variances:

Ho: $\sigma_1^2 = \sigma_2^2$ vs. Ha: $\sigma_1^2 \neq \sigma_2^2$

If the p-value > 0.05, we fail to reject the null and can conclude the variances of the two groups are equal; thus we use the pooled variances t*.

If the p-value < 0.05, we reject the null and conclude the variances of the two groups are unequal; thus we use the unpooled variances t*.

You find the F-test under the heading "Equality of Variances" in your SAS output. In our case, the p-value (Pr > F) is 0.03, which is less than 0.05; we cannot assume $\sigma_1^2 = \sigma_2^2$. We need to use the "t Value" from the "Unpooled" Method.

Conclusion

- The t* value for unpooled variances is -3.83, and the corresponding p-value = 0.0141, which is less than alpha (0.05). Therefore, we reject the null and conclude that the treatment group differs significantly from the control group in time to relief from headache.
- Notice from the SAS output that the treatment group took an average of about 20 minutes longer to feel relief than the control group ("Diff (1-2)"), implying the treatment is significantly worse than the control.

The TTEST Procedure

- General form of the TTEST procedure:

```
PROC TTEST DATA=SAS-data-set;
        CLASS variable;
        PAIRED variables;
        VAR variables;
RUN;
```

Chi-square test of independence

- What is the Chi-square test of independence?

  Ans. It tests whether the variable in the row and column are independent or related
- What is the null hypothesis?

  Ans. The variables in the row and column are independent: there is no relationship between row and column frequencies
- The command for SAS to test this is provided in the option of “proc freq”. Simply use **chisq**.
- To display the expected cell frequency for each cell use the option “expected.”

Zdroj: http://web.uri.edu/its/files/ppt/statistics.sas.ppt

Chi-square test of independence: exercise

There are 34 students in the classroom and there was a vote on whether they wanted to have a turtle in their classroom as a pet. The data file “vote.txt” contains the result of the vote (Yes=y, No=n), and gender of the students (male=m, female=f).

- Q1 Import the file “vote.txt” into SAS and name the variables “answers” and “gender.”

- Q2 Using the option “chisq,” test whether or not the answers to the vote and gender are associated with each other.

Answers

Q1

```
data vote;
infile 'c:/vote.txt';
input answers $ gender $;
run;
```

Q2

```
proc freq data=vote;
tables answers*gender /expected chisq;
run;
```

Results

Output - (Untitled)

The SAS System 13:14

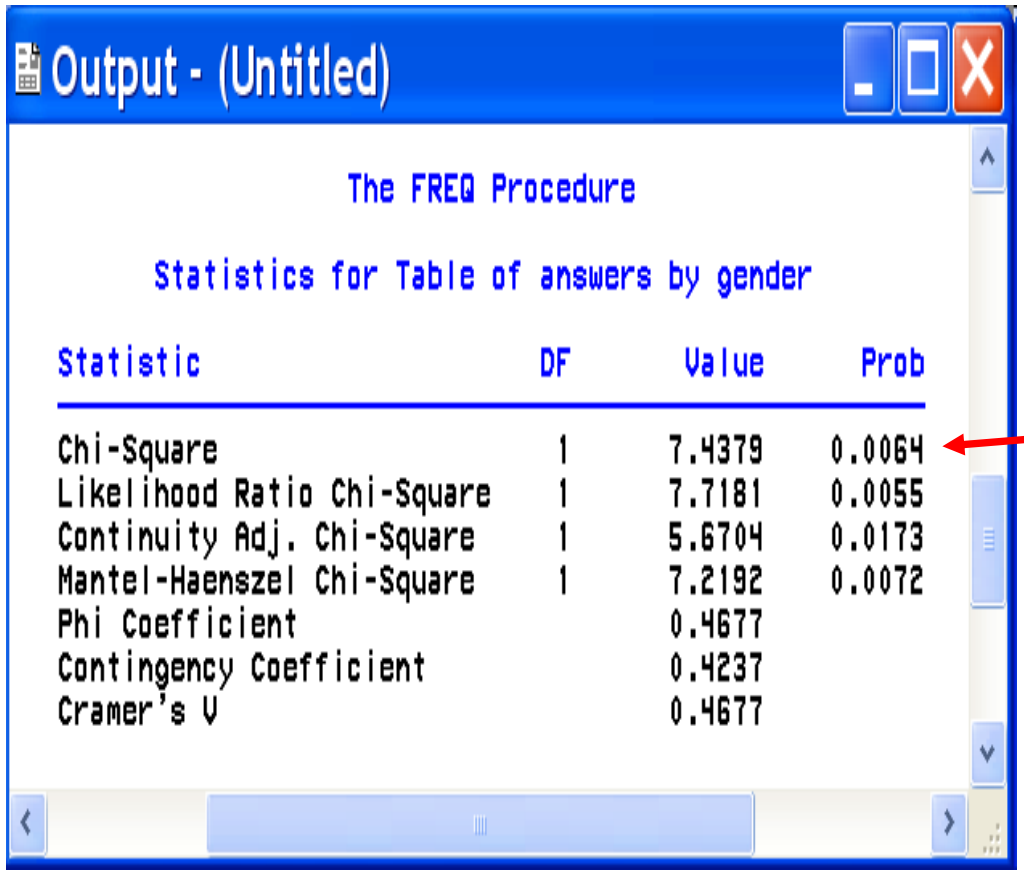
The FREQ Procedure

Table of answers by gender

answers Frequency Expected Percent Row Pct Col Pct	gender f	 m	 Total
n	11 7.0588 32.35 73.33 68.75	4 7.9412 11.76 26.67 22.22	15 44.12
y	5 8.9412 14.71 26.32 31.25	14 10.059 41.18 73.68 77.78	19 55.88
Total	16 47.06	18 52.94	34 100.00

$$\text{Expect Freq} = \frac{\text{Row total}(15) \times \text{Column total}(16)}{\text{Table total}(34)}$$

What does the result tell you?



- The null hypothesis that the two variables are independent is rejected at even 1% significance level.

This is lower than 0.01

- The two variables "answers" and "gender" are associated with each other (They are dependent).

13. Statistické tabulky

Následující tabulky obsahují hodnoty:

- Pravděpodobnostní funkce Binomického rozložení
- Pravděpodobnostní funkce Poissonova rozložení
- Distribuční funkce standardizivaného normálního rozložení
- Kvantilů standardizivaného normálního rozložení
- Kvantilů rozložení $\chi 2$ rozložení
- Kvantilů Studentova rozložení
- Kvantilů Fisherova-Snedecorova rozložení

Pravděpodobnostní funkce binomického rozložení Bi(n,p) 1. část

		p												
n	x	.01	.05	.10	.15	.20	.25	.30	1/3	.35	.40	.45	.49	.50
2	0	.9801	.9025	.8100	.7225	.6400	.5625	.4900	.4444	.4225	.3600	.3025	.2601	.2500
	1	.0198	.0950	.1800	.2550	.3200	.3750	.4200	.4444	.4550	.4800	.4950	.4998	.5000
	2	.0001	.0225	.0100	.0225	.0400	.0625	.0900	.1111	.1225	.1600	.2025	.2401	.2500
3	0	.9703	.8574	.7290	.6141	.5120	.4219	.3430	.2963	.2746	.2160	.1664	.1327	.1250
	1	.0294	.1354	.2430	.3251	.3840	.4219	.4410	.4444	.4436	.4320	.4084	.3823	.3750
	2	.0003	.0071	.0270	.0574	.0960	.1406	.1850	.2222	.2389	.2880	.3341	.3674	.3750
	3	.0000	.0001	.0010	.0034	.0080	.0156	.0270	.0370	.0429	.0640	.0911	.1176	.1250
4	0	.9606	.8145	.6561	.5220	.4096	.3164	.2401	.1975	.1785	.1296	.0915	.0677	.0625
	1	.0388	.1715	.2916	.3685	.4096	.4219	.4116	.3951	.3845	.3456	.2995	.2600	.2500
	2	.0006	.0135	.0486	.0975	.1536	.2109	.2646	.2963	.3105	.3456	.3675	.3747	.3750
	3	.0000	.0005	.0036	.0115	.0256	.0469	.0756	.0988	.1125	.1536	.2005	.2400	.2500
	4	.0000	.0000	.0001	.0005	.0016	.0039	.0081	.0123	.0150	.0256	.0410	.0576	.0625
5	0	.9510	.7738	.5905	.4437	.3277	.2373	.1681	.1317	.1160	.0778	.0503	.0345	.0312
	1	.0480	.2036	.3280	.3915	.4096	.3955	.3602	.3292	.3124	.2592	.2059	.1657	.1562
	2	.0010	.0214	.0729	.1382	.2048	.2637	.3087	.3292	.3364	.3456	.3369	.3185	.3125
	3	.0000	.0011	.0081	.0244	.0512	.0879	.1323	.1646	.1811	.2304	.2757	.3060	.3125
	4	.0000	.0000	.0004	.0022	.0064	.0146	.0284	.0412	.0488	.0768	.1128	.1470	.1562
	5	.0000	.0000	.0000	.0001	.0003	.0010	.0024	.0041	.0053	.0102	.0185	.0283	.0312

Pravděpodobnostní funkce binomického rozložení Bi(n,p) 2. část

		p												
6	0	.9415	.7351	.5314	.3771	.2621	.1780	.1176	.0878	.0754	.0467	.0277	.0176	.0156
	1	.0571	.2321	.3543	.3993	.3932	.3560	.3025	.2634	.2437	.1866	.1359	.1014	.0938
	2	.0014	.0305	.0984	.1762	.2458	.2966	.3241	.3292	.3280	.3110	.2780	.2437	.2344
	3	.0000	.0021	.0146	.0425	.0819	.1318	.1852	.2195	.2355	.2765	.3032	.3121	.3125
	4	.0000	.0001	.0012	.0055	.0154	.0330	.0595	.0823	.0951	.1382	.1861	.2249	.2344
	5	.0000	.0000	.0001	.0004	.0015	.0044	.0102	.0165	.0205	.0369	.0609	.0864	.0938
	6	.0000	.0000	.0000	.0000	.0001	.0002	.0007	.0014	.0018	.0041	.0083	.0139	.0156
7	0	.9321	.6983	.4783	.3206	.2097	.1335	.0824	.0585	.0490	.0280	.0152	.0090	.0078
	1	.0659	.2573	.3720	.3960	.3670	.3115	.2471	.2048	.1848	.1306	.0872	.0603	.0547
	2	.0020	.0406	.1240	.2097	.2753	.3115	.3171	.3073	.2985	.2613	.2140	.1740	.1641
	3	.0000	.0036	.0230	.0617	.1147	.1730	.2269	.2561	.2679	.2903	.2918	.2786	.2734
	4	.0000	.0002	.0026	.0109	.0287	.0577	.0972	.1280	.1442	.1935	.2388	.2676	.2734
	5	.0000	.0000	.0002	.0012	.0043	.0115	.0250	.0384	.0466	.0774	.1172	.1543	.1641
	6	.0000	.0000	.0000	.0001	.0004	.0013	.0036	.0064	.0084	.0172	.0320	.0494	.0547
	7	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0005	.0006	.0016	.0037	.0068	.0078
8	0	.9227	.6634	.4305	.2725	.1678	.1001	.0576	.0390	.0319	0168	.0084	.0046	.0039
	1	.0746	.2793	.3826	.3847	.3355	.2670	.1977	.1561	.1373	.0896	.0548	.0352	.0312
	2	.0026	.0515	.1488	.2376	.2936	.3115	.2965	.2731	.2587	.2090	.1569	.1183	.1094
	3	.0001	.0054	.0331	.0839	.1468	.2076	.2541	.2731	.2786	.2787	.2568	.2273	.2188
	4	.0000	.0004	.0046	.0185	.0459	.0865	.1361	.1707	.1875	.2322	.2627	.2730	.2734
	5	.0000	.0000	.0004	.0026	.0092	.0231	.0467	.0683	.0808	.1239	.1719	.2098	.2188
	6	.0000	.0000	.0000	.0002	.0011	.0038	.0100	.0171	.0217	.0413	.0703	.1008	.1094
	7	.0000	.0000	.0000	.0000	.0001	.0004	.0012	.0024	.0033	.0079	.0164	.0277	.0312
	8	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0002	.0007	.0017	.0033	.0039

Pravděpodobnostní funkce binomického rozložení Bi(n,p) 3. část

		p											
9	0	.9135	.6302	.3874	.2316	.1342	.0751	.0404	.0260	.0207	.0101	.0046	.0023	.0020
	1	.0830	.2985	.3874	.3679	.3020	.2253	.1556	.1171	.1004	.0605	.0339	.0202	.0176
	2	.0034	.0629	.1722	.2597	.3020	.3003	.2668	.2341	.2162	.1612	.1110	.0776	.0703
	3	.0001	.0077	.0446	.1069	.1762	.2336	.2668	.2731	.2716	.2508	.2119	.1739	.1641
	4	.0000	.0006	.0074	.0283	.0661	.1168	.1715	.2048	.2194	.2508	.2600	.2506	.2461
	5	.0000	.0000	.0008	.0050	.0165	.0389	.0735	.1024	.1181	.1672	.2128	.2408	.2461
	6	.0000	.0000	.0001	.0006	.0028	.0087	.0210	.0341	.0424	.0743	.1160	.1542	.1641
	7	.0000	.0000	.0000	.0000	.0003	.0012	.0039	.0073	.0098	.0212	.0407	.0635	.0703
	8	.0000	.0000	.0000	.0000	.0000	.0001	.0004	.0009	.0013	.0035	.0083	.0153	.0176
	9	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0003	.0008	.0016	.0020
10	0	.9044	.5987	.3487	.1969	.1074	.0563	.0282	.0173	.0135	.0060	.0025	.0012	.0010
	1	.0914	.3151	.3874	.3474	.2684	.1877	.1211	.0867	.0725	.0403	.0207	.0114	.0098
	2	.0042	.0746	.1937	.2759	.3020	.2816	.2335	.1951	.1757	.1209	.0763	.0495	.0439
	3	.0001	.0105	.0574	.1298	.2013	.2503	.2668	.2601	.2522	.2150	.1665	.1267	.1172
	4	.0000	.010	.0112	.0401	.0881	.1460	.2001	.2276	.2377	.2508	.2384	.2130	.2051
	5	.0000	.0001	.0015	.0085	.0264	.0584	.1029	.1366	.1536	.2007	.2340	.2456	.2461
	6	.0000	.0000	.0001	.0012	.0055	.0162	.0368	.0569	.0689	.1115	.1596	.1966	.2051
	7	.0000	.0000	.0000	.0001	.0008	.0031	.0090	.0163	.0212	.0425	.0746	.1080	.1172
	8	.0000	.0000	.0000	.0000	.0001	.0004	.0014	.0030	.0043	.0106	.0229	.0389	.0439
	9	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0005	.0016	.0042	.0083	.0098
	10	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0008	.0010

Pravděpodobnostní funkce Poissonova rozložení Po(λ) 1. část

	λ									
x	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0
0	0,9048	8187	7408	6703	6065	5488	4966	4493	4066	3679
1	0905	1637	2222	2681	3033	3293	3476	3595	3659	3679
2	0045	0164	0333	0536	0758	0988	1217	1438	1647	1839
3	0002	0011	0033	0072	0126	0198	0284	0383	0494	0613
4	0000	0001	0003	0007	0016	0030	0050	0077	0111	0153
5	0000	0000	0000	0001	0002	0004	0007	0012	0020	0031
6	0000	0000	0000	0000	0000	0000	0001	0002	0003	0005
7	0000	0000	0000	0000	0000	0000	0000	0000	0000	0001

x	1,1	1,2	1,3	1,4	1,5	1,6	1,7	1,8	1,9	2,0
0	0,3329	3012	2725	2466	2231	2019	1827	1653	1496	1353
1	3662	3614	3543	3452	3347	3230	3106	2975	2842	2707
2	2014	2169	2303	2417	2510	2584	2640	2678	2700	2707
3	0738	0867	0998	1128	1255	1378	1496	1607	1710	1804
4	0203	0260	0324	0395	0471	0551	0636	0723	0812	0902
5	0045	0062	0084	0111	0141	0176	0216	0260	0309	0361
6	0008	0012	0018	0026	0035	0047	0061	0078	0098	0120
7	0001	0002	0003	0005	0008	0011	0015	0020	0027	0034
8	0000	0000	0001	0001	0001	0002	0003	0005	0006	0009
9	0000	0000	0000	0000	0000	0000	0001	0001	0001	0002

Pravděpodobnostní funkce Poissonova rozložení Po(λ) 2. část

	λ									
x	3,0	4,0	5,0	6,0	7,0	8,0	9,0	10,0	11,0	12,0
0	0,0498	0183	0067	0025	0009	0003	0001	0000	0000	0000
1	1494	0733	0337	0149	0064	0027	0011	0005	0002	0001
2	2240	1465	0842	0446	0223	0107	0050	0023	0010	0004
3	2240	1954	1404	0892	0521	0286	0150	0076	0037	0018
4	1680	1954	1755	1339	0912	0573	0337	0189	0102	0053
5	1008	1563	1755	1606	1277	0916	0607	0378	0224	0127
6	0504	1042	1462	1606	1490	1221	0911	0631	0411	0255
7	0216	0595	1044	1377	1490	1396	1171	0901	0646	0437
8	0081	0298	0653	1033	1304	1396	1318	1126	0888	0655
9	0027	0132	0363	0688	1014	1241	1318	1251	1085	0874
10	0008	0053	0181	0413	0710	0993	1186	1251	1194	1048
11	0002	0019	0082	0225	0452	0722	0970	1137	1194	1144
12	0002	0006	0034	0113	0264	0481	0728	0948	1094	1144
13		0002	0013	0052	0142	0296	0504	0729	0926	1056
14		0001	0005	0022	0071	0169	0324	0521	0728	0905
15			0002	0009	0033	0090	0194	0347	0534	0724
16				0003	0014	0045	0109	0217	0367	0543
17				0001	0006	0021	0058	0128	0237	0383
18					0002	0009	0029	0071	0145	0256
19					0001	0004	0014	0037	0084	0161
20						0002	0006	0019	0046	0097
21						0001	0003	0009	0024	0055
22							0001	0004	0012	0030
23								0002	0006	0016
24								0001	0003	0008
25									0001	0004
26										0002

Distribuční funkce Φ(u) rozložení N(0,1)

	Φ(u)	u	Φ(u)	u	Φ(u)	u	Φ(u)
0,00	0,50000	0,40	0,65542	0,80	0,78814	1,20	0,88493
0,01	0,50399	0,41	0,65910	0,81	0,79103	1,21	0,88686
0,02	0,50798	0,42	0,66276	0,82	0,79389	1,22	0,88877
0,03	0,51197	0,43	0,66640	0,83	0,79673	1,23	0,89065
0,04	0,51595	0,44	0,67003	0,84	0,79955	1,24	0,89251
0,05	0,51994	0,45	0,67364	0,85	0,80234	1,25	0,79435
0,06	0,52392	0,46	0,67724	0,86	0,80511	1,26	0,89617
0,07	0,52790	0,47	0,68082	0,87	0,80785	1,27	0,89796
0,08	0,53188	0,48	0,68439	0,88	0,81057	1,28	0,89973
0,09	0,53586	0,49	0,68793	0,89	0,81327	1,29	0,90147
0,10	0,53983	0,50	0,69146	0,90	0,81594	1,30	0,90320
0,11	0,54380	0,51	0,69497	0,91	0,81859	1,31	0,90490
0,12	0,54776	0,52	0,69847	0,92	0,82121	1,32	0,90658
0,13	0,55172	0,53	0,70194	0,93	0,82381	1,33	0,90824
0,14	0,55567	0,54	0,70540	0,94	0,82639	1,34	0,90988
0,15	0,55962	0,55	0,70884	0,95	0,82894	1,35	0,91149
0,16	0,56356	0,56	0,71226	0,96	0,83147	1,36	0,91309
0,17	0,56749	0,57	0,71655	0,97	0,83398	1,37	0,91466
0,18	0,57142	0,58	0,71904	0,98	0,83646	1,38	0,91621
0,19	0,57535	0,59	0,72240	0,99	0,83891	1,39	0,91774
0,20	0,57926	0,60	0,72575	1,00	0,84134	1,40	0,91924
0,21	0,58317	0,61	0,72907	1,01	0,84375	1,41	0,92073
0,22	0,58706	0,62	0,73237	1,02	0,84614	1,42	0,92220
0,23	0,59095	0,63	0,73565	1,03	0,84850	1,43	0,92364
0,24	0,59483	0,64	0,73891	1,04	0,85083	1,44	0,92507
0,25	0,59871	0,65	0,74215	1,05	0,85314	1,45	0,92647
0,26	0,60257	0,66	0,74537	1,06	0,85543	1,46	0,92786
0,27	0,60642	0,67	0,74857	1,07	0,85769	1,47	0,92922
0,28	0,61026	0,68	0,75175	1,08	0,85993	1,48	0,93056
0,29	0,61409	0,69	0,75490	1,09	0,86214	1,49	0,93189
0,30	0,61791	0,70	0,75804	1,10	0,86433	1,50	0,93319
0,31	0,62172	0,71	0,76115	1,11	0,86650	1,51	0,93448
0,32	0,62552	0,72	0,76424	1,12	0,86864	1,52	0,93574
0,33	0,62930	0,73	0,76730	1,13	0,87076	1,53	0,93699
0,34	0,63307	0,74	0,77035	1,14	0,87286	1,54	0,93822
0,35	0,63683	0,75	0,77337	1,15	0,87493	1,55	0,93943
0,36	0,64058	0,76	0,77637	1,16	0,87698	1,56	0,94062
0,37	0,64431	0,77	0,77935	1,17	0,87900	1,57	0,94179
0,38	0,64803	0,78	0,78230	1,18	0,88100	1,58	0,94295
0,39	0,65173	0,79	0,78524	1,19	0,88298	1,59	0,94408

u	Φ(u)	u	Φ(u)	u	Φ(u)	u	Φ(u)
1,60	0,94520	2,00	0,97725	2,40	0,99180	3,10	0,99903
1,61	0,94630	2,01	0,97778	2,41	0,99202	3,12	0,99910
1,62	0,94738	2,02	0,97831	2,42	0,99224	3,14	0,99916
1,63	0,94845	2,03	0,97882	2,43	0,99245	3,16	0,99921
1,64	0,94950	2,04	0,97932	2,44	0,99266	3,18	0,99926
1,65	0,95053	2,05	0,97982	2,45	0,99286	3,20	0,99931
1,66	0,95154	2,06	0,98030	2,46	0,99305	3,22	0,99936
1,67	0,95254	2,07	0,98077	2,47	0,99324	3,24	0,99940
1,68	0,95352	2,08	0,98124	2,48	0,99343	3,26	0,99944
1,69	0,95449	2,09	0,98169	2,49	0,99361	3,28	0,99948
1,70	0,95543	2,10	0,98214	2,50	0,99379	3,30	0,99952
1,71	0,95637	2,11	0,98257	2,52	0,99413	3,32	0,99955
1,72	0,95728	2,12	0,98300	2,54	0,99446	3,34	0,99958
1,73	0,95818	2,13	0,98341	2,56	0,99477	3,36	0,99961
1,74	0,95907	2,14	0,98382	2,58	0,99506	3,38	0,99964
1,75	0,95994	2,15	0,98422	2,60	0,99534	3,40	0,99966
1,76	0,96080	2,16	0,98461	2,62	0,99560	3,42	0,99969
1,77	0,96164	2,17	0,98500	2,64	0,99585	3,44	0,99971
1,78	0,96246	2,18	0,98537	2,66	0,99609	3,46	0,99973
1,79	0,96327	2,19	0,98574	2,68	0,99632	3,48	0,99975
1,80	0,96407	2,20	0,98610	2,70	0,99653	3,50	0,99977
1,81	0,96485	2,21	0,98645	2,72	0,99674	3,55	0,99981
1,82	0,96562	2,22	0,98679	2,74	0,99683	3,60	0,99984
1,83	0,96638	2,23	0,98713	2,76	0,99711	3,65	0,99987
1,84	0,96712	2,24	0,98745	2,78	0,99728	3,70	0,99989
1,85	0,96784	2,25	0,98778	2,80	0,99744	3,72	0,99991
1,86	0,96856	2,26	0,98809	2,82	0,99760	3,80	0,99993
1,87	0,96926	2,27	0,98840	2,84	0,99774	3,85	0,99994
1,88	0,96995	2,28	0,98870	2,86	0,99788	3,90	0,99995
1,89	0,97062	2,29	0,98899	2,88	0,99801	3,95	0,99996
1,90	0,97128	2,30	0,98928	2,90	0,99813	4,00	0,99997
1,91	0,97193	2,31	0,98956	2,92	0,99825	4,05	0,99997
1,92	0,97257	2,32	0,98983	2,94	0,99836	4,10	0,99998
1,93	0,97320	2,33	0,99010	2,96	0,99846	4,15	0,99998
1,94	0,97381	2,34	0,99036	2,98	0,99856	4,20	0,99999
1,95	0,97441	2,35	0,99061	3,00	0,99865	4,25	0,99999
1,96	0,97500	2,36	0,99086	3,02	0,99874	4,30	0,99999
1,97	0,97558	2,37	0,99111	3,04	0,99882	4,35	0,99999
1,98	0,97615	2,38	0,99134	3,06	0,99889	4,40	0,99999
1,99	0,97670	2,39	0,99158	3,08	0,99897	4,45	1,00000

Kvantily standardizovaného normálního rozložení u_P

P	u_P	P	u_P	P	u_P	P	u_P
0,50	0,000	0,75	0,674	0,950	1,645	0,975	1,960
0,51	0,025	0,76	0,706	0,951	1,655	0,976	1,970
0,52	0,050	0,77	0,739	0,952	1,665	0,977	1,995
0,53	0,075	0,78	0,772	0,953	1,675	0,978	2,014
0,54	0,100	0,79	0,806	0,954	1,685	0,979	2,034
0,55	0,126	0,80	0,842	0,955	1,695	0,980	2,054
0,56	0,151	0,81	0,878	0,956	1,706	0,981	2,075
0,57	0,176	0,82	0,915	0,957	1,717	0,982	2,097
0,58	0,202	0,83	0,954	0,958	1,728	0,983	2,120
0,59	0,228	0,84	0,994	0,959	1,739	0,984	2,144
0,60	0,253	0,85	1,036	0,960	1,751	0,985	2,170
0,61	0,279	0,86	1,080	0,961	1,762	0,986	2,197
0,62	0,305	0,87	1,126	0,962	1,774	0,987	2,226
0,63	0,332	0,88	1,175	0,963	1,787	0,988	2,257
0,64	0,358	0,89	1,227	0,964	1,799	0,989	2,290
0,65	0,385	0,90	1,282	0,965	1,812	0,990	2,326
0,66	0,412	0,905	1,311	0,966	1,825	0,991	2,366
0,67	0,440	0,910	1,341	0,967	1,838	0,992	2,409
0,68	0,468	0,915	1,372	0,968	1,852	0,993	2,457
0,69	0,496	0,920	1,405	0,969	1,866	0,994	2,512
0,70	0,524	0,925	1,440	0,970	1,881	0,995	2,576
0,71	0,553	0,930	1,476	0,971	1,896	0,996	2,652
0,72	0,583	0,935	1,514	0,972	1,911	0,997	2,748
0,73	0,613	0,940	1,555	0,973	1,927	0,998	2,878
0,74	0,643	0,945	1,598	0,974	1,943	0,999	3,090

Kvantily Pearsonova rozložení $\chi^2(v)$

stupně volnosti	pravděpodobnost				
	0,005	0,01	0,025	0,05	0,1
1	0,0000	0,0002	0,0010	0,0039	0,0158
2	0,0100	0,0201	0,0506	0,1026	0,2107
3	0,0717	0,1148	0,2158	0,3519	0,5844
4	0,2070	0,2971	0,4844	0,7107	1,0636
5	0,4117	0,5543	0,8312	1,1455	1,6103
6	0,6757	0,8721	1,2373	1,6354	2,2041
7	0,9893	1,2390	1,6899	2,1673	2,8331
8	1,3444	1,6465	2,1797	2,7326	3,4895
9	1,7349	2,0879	2,7004	3,3251	4,1682
10	2,1559	2,5582	3,2470	3,9403	4,8652
11	2,6032	3,0535	3,8157	4,5748	5,5778
12	3,0738	3,5706	4,4038	5,2260	6,3038
13	3,5650	4,1069	5,0088	5,8919	7,0415
14	4,0747	4,6604	5,6287	6,5706	7,7895
15	4,6009	5,2293	6,2621	7,2609	8,5468
16	5,1422	5,8122	6,9077	7,9616	9,3122
17	5,6972	6,4078	7,5642	8,6718	10,085
18	6,2648	7,0149	8,2307	9,3905	10,865
19	6,8440	7,6327	8,9065	10,117	11,651
20	7,4338	8,2604	9,5908	10,851	12,443
21	8,0337	8,8972	10,283	11,591	13,240
22	8,6427	9,5425	10,982	12,338	14,041
23	9,2604	10,196	11,689	13,091	14,848
24	9,8862	10,856	12,401	13,848	15,659
25	10,520	11,524	13,120	14,611	16,473
26	11,160	12,198	13,844	15,379	17,292
27	11,808	12,879	14,573	16,151	18,114
28	12,461	13,565	15,308	16,928	18,939
29	13,121	14,256	16,047	17,708	19,768
30	13,787	14,953	16,791	18,493	20,599
40	20,707	22,164	24,433	26,509	29,051
50	27,991	29,707	32,357	34,764	37,689
60	35,534	37,485	40,482	43,188	46,459
70	43,275	45,442	48,758	51,739	55,329
80	51,172	53,540	57,153	60,391	64,278
90	59,196	61,754	65,647	69,126	73,291
100	67,328	70,065	74,222	77,929	82,358
200	152,24	156,43	162,73	168,28	174,84
300	240,66	245,97	253,91	260,88	269,07
500	422,30	429,39	439,94	449,15	459,93

stupně volnosti	pravděpodobnost				
	0,90	0,95	0,975	0,99	0,995
1	2,706	3,841	5,024	6,635	7,879
2	4,605	5,991	7,378	9,210	10,597
3	6,251	7,814	9,348	11,345	12,838
4	7,779	9,488	11,143	13,277	14,860
5	9,236	11,070	12,833	15,086	16,750
6	10,645	12,592	14,449	16,812	18,548
7	12,017	14,067	16,013	18,475	20,278
8	13,362	15,507	17,535	20,090	21,955
9	14,684	16,919	19,023	21,666	23,589
10	15,987	18,307	20,483	23,209	25,188
11	17,275	19,675	21,920	24,725	26,757
12	18,549	21,026	23,337	26,217	28,300
13	19,812	22,362	24,736	27,688	29,819
14	21,064	23,685	26,119	29,141	31,319
15	22,307	27,996	27,488	30,578	32,801
16	23,542	26,296	28,845	32,000	34,267
17	24,769	27,587	30,191	33,409	35,718
18	25,989	28,869	31,526	34,805	37,156
19	27,204	30,144	32,852	36,191	38,582
20	28,412	31,410	34,170	37,566	39,997
21	29,615	32,671	35,479	38,932	41,401
22	30,813	33,924	36,781	40,289	42,796
23	32,007	35,172	38,076	41,638	44,181
24	33,196	36,415	39,364	42,980	45,599
25	34,382	37,652	40,646	44,314	46,928
26	35,563	38,885	41,923	45,642	48,290
27	36,741	40,113	43,195	46,963	49,645
28	37,916	41,337	44,461	48,278	50,993
29	39,087	42,557	45,722	49,588	52,336
30	40,256	43,773	46,979	50,892	53,672
40	51,805	55,758	59,342	63,691	66,766
50	63,167	67,505	71,420	76,154	79,490
60	74,397	79,082	83,298	88,379	91,952
70	85,527	90,531	95,023	100,43	104,21
80	96,578	101,88	106,63	112,33	116,32
90	107,57	113,15	118,14	124,12	128,30
100	118,50	124,34	129,56	135,81	140,17
200	226,02	233,99	241,06	249,45	255,26
300	331,79	341,40	349,87	359,91	366,84
500	540,93	553,13	563,85	576,49	585,21

Kvantily Studentova rozložení t(n)

stupně volnosti	pravděpodobnost				
	0,90	0,95	0,975	0,99	0,995
1	3,078	6,314	12,706	31,821	63,657
2	1,886	2,920	4,303	6,965	9,925
3	1,638	2,353	3,182	4,541	5,841
4	1,533	2,132	2,776	3,747	4,604
5	1,476	2,015	2,571	3,365	4,032
6	1,440	1,943	2,447	3,143	3,707
7	1,415	1,895	2,365	2,998	3,499
8	1,397	1,860	2,306	2,896	3,355
9	1,303	1,833	2,262	2,821	3,250
10	1,372	1,812	2,228	2,764	3,169
11	1,363	1,796	2,201	2,718	3,106
12	1,356	1,782	2,179	2,681	3,055
13	1,350	1,771	2,160	2,650	3,012
14	1,345	1,761	2,145	2,624	2,977
15	1,341	1,753	2,131	2,602	2,947
16	1,337	1,746	2,120	2,583	2,921
17	1,333	1,740	2,110	2,567	2,898
18	1,330	1,734	2,101	2,552	2,878
19	1,328	1,729	2,093	2,539	2,961
20	1,325	1,725	2,086	2,528	2,845
21	1,323	1,721	2,080	2,518	2,831
22	1,321	1,717	2,074	2,508	2,819
23	1,319	1,714	2,069	2,500	2,807
24	1,318	1,711	2,064	2,492	2,797
25	1,316	1,708	2,060	2,485	2,878
26	1,315	1,706	2,056	2,479	2,779
27	1,314	1,703	2,052	2,473	2,771
28	1,313	1,701	2,048	2,467	2,763
29	1,311	1,699	2,045	2,462	2,756
30	1,310	1,697	2,042	2,457	2,750
40	1,303	1,684	2,021	2,423	2,704
60	1,296	1,671	2,000	2,390	2,660
120	1,289	1,658	1,980	2,358	2,617
∞	1,282	1,645	1,960	2,326	2,576

Kvantily rozložení $F_{0,95}(\nu_1,\nu_2)$ - 1. část

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9
1	161,45	199,50	215,71	224,58	230,16	233,99	236,77	238,88	240,54
2	18,513	19,000	19,164	19,247	19,296	19,330	19,353	19,371	19,385
3	10,128	9,552	9,277	9,117	9,014	8,941	8,887	8,845	8,812
4	7,709	6,944	6,591	6,388	6,256	6,163	6,094	6,041	5,999
5	6,608	5,786	5,410	5,192	5,050	4,950	4,876	4,818	4,773
6	5,987	5,143	4,757	4,534	4,387	4,284	4,207	4,147	4,099
7	5,591	4,737	4,347	4,120	3,972	3,866	3,787	3,726	3,677
8	5,318	4,459	4,066	3,838	3,688	3,581	3,501	3,438	3,388
9	5,117	4,257	3,863	3,633	3,482	2,274	3,293	3,230	3,179
10	4,965	4,103	3,708	3,478	3,326	3,217	3,136	3,072	3,020
11	4,844	3,982	3,587	3,357	3,204	3,095	3,012	2,948	2,896
12	4,747	3,885	3,490	3,259	3,106	2,996	2,913	2,849	2,796
13	4,667	3,806	3,411	3,179	3,025	2,915	2,832	2,767	2,714
14	4,600	3,739	3,344	3,112	2,958	2,848	2,764	2,699	2,646
15	4,543	3,682	3,287	3,056	2,901	2,791	2,707	2,641	2,588
16	4,494	3,634	3,239	3,007	2,852	2,741	2,657	2,591	2,538
17	4,451	3,592	3,197	2,965	2,810	2,699	2,614	2,548	2,494
18	4,414	3,555	3,160	2,928	2,773	2,661	2,577	2,510	2,456
19	4,381	3,522	3,127	2,895	2,740	2,628	2,544	2,477	2,423
20	4,351	3,493	3,098	2,866	2,711	2,599	2,514	2,447	2,393
21	4,325	3,467	3,073	2,840	2,685	2,573	2,488	2,421	2,366
22	4,301	3,443	3,049	2,817	2,661	2,549	2,464	2,397	2,342
23	4,279	3,422	3,028	2,796	2,640	2,528	2,442	2,375	2,320
24	4,260	3,403	3,009	2,776	2,621	2,508	2,423	2,355	2,300
25	4,242	3,385	2,991	2,759	2,603	2,490	2,405	2,337	2,282
26	4,225	3,369	2,975	2,743	2,587	2,474	2,388	2,321	2,266
27	4,210	3,354	2,960	2,728	2,572	2,459	2,373	2,305	2,250
28	4,196	3,340	2,947	2,714	2,558	2,445	2,359	2,291	2,236
29	4,183	3,328	2,934	2,701	2,545	2,432	2,346	2,278	2,223
30	4,171	3,316	2,922	2,690	2,534	2,421	2,334	2,266	2,211
40	4,085	3,232	2,839	2,606	2,450	2,336	2,249	2,180	2,124
60	4,001	3,150	2,758	2,525	2,368	2,254	2,167	2,097	2,040
120	3,920	3,072	2,680	2,447	2,290	2,175	2,087	2,016	1,959
∞	3,842	2,996	2,605	2,372	2,214	2,099	2,010	1,938	1,880

Kvantily rozložení $F_{0,95}(\nu_1,\nu_2)$ - 2. část

ν_2 \ ν_1	10	12	15	20	24	30	40	60	120	∞
1	241´,88	243,91	245,95	248,01	249,05	250,09	251,14	252,20	253,25	254,32
2	19,396	19,413	19,429	19,446	19,454	19,462	19,471	19,479	19,487	19,496
3	8,786	8,745	8,703	8,660	8,639	8,617	8,594	8,572	8,549	8,527
4	5,964	5,912	5,858	5,803	5,774	5,746	5,717	5,688	5,658	5,628
5	4,735	4,678	4,619	4,558	4,527	4,496	4,464	4,431	4,398	4,365
6	4,060	4,000	3,938	3,874	3,842	3,808	3,774	3,740	3,705	3,669
7	3,637	3,575	3,511	3,445	3,411	3,376	3,340	3,304	3,267	3,230
8	3,347	3,284	3,218	3,150	3,115	3,079	3,043	3,005	2,967	2,928
9	3,137	3,073	3,006	2,937	2,901	2,864	2,826	2,787	2,748	2,707
10	2,978	2,913	2,845	2,774	2,737	2,700	2,661	2,621	2,580	2,538
11	2,854	2,788	2,719	2,646	2,609	2,571	2,531	2,490	2,448	2,405
12	2,753	2,687	2,617	2,544	2,506	2,466	2,426	2,384	2,341	2,296
13	2,671	2,604	2,533	2,459	2,420	2,380	2,339	2,297	2,252	2,206
14	2,602	2,534	2,463	2,388	2,349	2,308	2,266	2,223	2,178	2,131
15	2,544	2,475	2,404	2,328	2,288	2,247	2,204	2,160	2,114	2,066
16	2,494	2,425	2,352	2,276	2,235	2,194	2,151	2,106	2,059	2,010
17	2,450	2,381	2,308	2,230	2,190	2,148	2,104	2,058	2,011	1,960
18	2,412	2,342	2,269	2,191	2,150	2,107	2,063	2,017	1,968	1,917
19	2,378	2,308	2,234	2,156	2,114	2,071	2,026	1,980	1,930	1,878
20	2,348	2,278	2,203	2,124	2,083	2,039	1,994	1,946	1,896	1,843
21	2,321	2,250	2,176	2,096	2,054	2,010	1,965	1,917	1,866	1,812
22	2,297	2,226	2,151	2,071	2,028	1,984	1,938	1,890	1,838	1,783
23	2,275	2,204	2,128	2,048	2,005	1,961	1,914	1,865	1,813	1,757
24	2,255	2,183	2,108	2,027	1,984	1,939	1,892	1,842	1,790	1,733
25	2,237	2,165	2,089	2,008	1,964	1,919	1,872	1,822	1,768	1,711
26	2,220	2,148	2,072	1,990	1,946	1,901	1,853	1,803	1,749	1,691
27	2,204	2,132	2,056	1,974	1,930	1,884	1,836	1,785	1,731	1,672
28	2,190	2,118	2,041	1,959	1,915	1,869	1,820	1,769	1,714	1,654
29	2,177	2,105	2,028	1,945	1,901	1,854	1,806	1,754	1,698	1,638
30	2,165	2,092	2,015	1,932	1,887	1,841	1,792	1,740	1,684	1,622
40	2,077	2,004	1,925	1,839	1,793	1,744	1,693	1,637	1,577	1,509
60	1,993	1,917	1,836	1,748	1,700	1,649	1,594	1,534	1,467	1,389
120	1,911	1,834	1,751	1,659	1,608	1,554	1,495	1,429	1,352	1,254
∞	1,831	1,752	1,666	1,571	1,517	1,459	1,394	1,318	1,221	1,000

Kvantily rozložení $F_{0,975}(\nu_1,\nu_2)$ - 1. část

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9
1	647,79	799,50	864,16	899,58	921,85	937,11	948,22	956,66	963,28
2	38,506	39,000	39,165	39,248	39,298	39,331	39,355	39,373	39,387
3	17,443	16,044	15,439	15,101	14,885	14,735	14,624	14,540	14,473
4	12,218	10,649	9,979	9,605	9,365	9,197	9,074	8,980	8,905
5	10,007	8,434	7,764	7,388	7,146	6,978	6,853	6,757	6,681
6	8,813	7,260	6,599	6,227	5,988	5,820	5,696	5,600	5,523
7	8,073	6,542	5,890	5,523	5,285	5,119	4,995	4,899	4,823
8	7,571	6,060	5,416	5,053	4,817	4,652	4,529	4,433	4,357
9	7,209	5,715	5,078	4,718	4,484	4,320	4,197	4,102	4,026
10	6,937	5,456	4,826	4,468	4,236	4,072	3,950	3,855	3,779
11	6,724	5,256	4,630	4,275	4,044	3,881	3,759	3,664	3,588
12	6,554	5,096	4,474	4,121	3,891	3,728	3,607	3,512	3,436
13	6,414	4,965	4,347	3,996	3,767	3,604	3,483	3,388	3,312
14	6,298	4,857	4,242	3,892	3,663	3,501	3,380	3,285	3,209
15	6,200	4,765	4,153	3,804	3,576	3,415	3,293	3,199	3,123
16	6,115	4,687	4,077	3,729	3,502	3,341	3,219	3,125	3,049
17	6,042	4,619	4,011	3,665	3,438	3,277	3,156	3,061	2,985
18	5,978	4,560	3,954	3,608	3,382	3,221	3,100	3,005	2,929
19	5,922	4,508	3,903	3,559	3,333	3,172	3,051	2,956	2,880
20	5,872	4,461	3,859	3,515	3,289	3,128	3,007	2,913	2,837
21	5,827	4,420	3,819	3,475	3,250	3,090	2,969	2,874	2,798
22	5,786	4,383	3,783	3,440	3,215	3,055	2,934	2,839	2,763
23	5,750	4,349	3,751	3,408	3,184	3,023	2,902	2,808	2,731
24	5,717	4,319	3,721	3,379	3,155	2,995	2,874	2,779	2,703
25	5,686	4,291	3,694	3,353	3,129	2,969	2,848	2,753	2,677
26	5,659	4,266	3,670	3,329	3,105	2,945	2,824	2,729	2,653
27	5,633	4,242	3,647	3,307	3,083	2,923	2,802	2,707	2,631
28	5,610	4,221	3,626	3,286	3,063	2,903	2,782	2,687	2,611
29	5,588	4,201	3,607	3,267	3,044	2,884	2,763	2,669	2,592
30	5,568	4,182	3,589	3,250	3,027	2,867	2,746	2,651	2,557
40	5,424	4,051	3,463	3,126	2,904	2,744	2,624	2,529	2,452
60	5,286	3,925	3,343	3,008	2,786	2,627	2,507	2,412	2,334
120	5,152	3,805	3,227	2,894	2,674	2,515	2,395	2,299	2,222
∞	5,024	3,689	3,116	2,786	2,567	2,408	2,288	2,192	2,114

Kvantily rozložení $F_{0,975}(\nu_1,\nu_2)$ - 2. část

ν_2 \ ν_1	10	12	15	20	24	30	40	60	120	∞
1	968,93	976,71	984,87	993,10	997,25	1001,4	1005,6	1009,8	1014,0	1018,3
2	39,398	39,415	39,431	39,448	39,456	39,465	39,473	39,481	39,490	39,498
3	14,419	14,337	14,253	14,167	14,124	14,081	14,037	13,992	13,947	13,902
4	8,844	8,751	8,657	8,560	8,511	8,461	8,411	8,360	8,309	8,257
5	6,619	6,525	6,428	6,329	6,278	6,227	6,175	6,125	6,069	6,0115
6	5,461	5,366	5,269	5,168	5,117	5,065	5,013	4,959	4,905	4,849
7	4,761	4,666	4,568	4,467	4,415	4,362	4,309	4,256	4,199	4,142
8	4,295	4,200	4,101	4,000	3,947	3,894	3,840	3,784	3,728	3,670
9	3,964	3,868	3,769	3,667	3,614	3,560	3,506	3,449	3,392	3,333
10	3,717	3,621	3,522	3,419	3,365	3,311	3,255	3,198	3,140	3,080
11	3,526	3,430	3,330	3,226	3,173	3,118	3,061	3,004	2,944	2,883
12	3,374	3,277	3,177	3,073	3,019	2,963	2,906	2,848	2,787	2,725
13	3,250	3,153	3,053	2,948	2,893	2,837	2,7880	2,720	2,659	2,596
14	3,147	3,050	2,949	2,844	2,789	2,732	2,674	2,614	2,552	2,487
15	3,060	2,963	2,862	2,756	2,701	2,644	2,585	2,524	2,461	2,395
16	2,986	2,889	2,788	2,681	2,625	2,568	2,509	2,447	2,383	2,316
17	2,922	2,825	2,723	2,616	2,560	2,502	2,442	2,380	2,315	2,247
18	2,866	2,769	2,667	2,559	2,503	2,445	2,384	2,321	2,256	2,187
19	2,817	2,720	2,617	2,509	2,452	2,394	2,333	2,270	2,203	2,133
20	2,774	2,676	2,573	2,465	2408	2,349	2,287	2,223	2,156	2,085
21	2,735	2,637	2,534	2,425	2,368	2,308	2,247	2,182	2,114	2,042
22	2,700	2,602	2,498	2,389	2,332	2,272	2,210	2,145	2,076	2,003
23	2,668	2,570	2,467	2,357	2,299	2,239	2,176	2,111	2,042	1,968
24	2,640	2,541	2,437	2,327	2,269	2,209	2,146	2,080	2,010	1,935
25	2,614	2,515	2,411	2,301	2,242	2,182	2,118	2,052	1,981	1,906
26	2,590	2,491	2,387	2,276	2,217	2,157	2,093	2,026	1,955	1,878
27	2,568	2,469	2,364	2,253	2,195	2,133	2,069	2,002	1,930	1,853
28	2,547	2,448	2,344	2,232	2,174	2,112	2,048	1,980	1,907	1,829
29	2,529	2,430	2,325	2,213	2,154	2,092	2,028	1,959	1,886	1,807
30	2,511	2,412	2,307	2,195	2,136	2,074	2,009	1,940	1,866	1,787
40	2,388	2,288	2,182	2,068	2,007	1,943	1,875	1,803	1,724	1,637
60	2,270	2,169	2,061	1,945	1,882	1,815	1,744	1,667	1,581	1,482
120	2,157	2,055	1,945	1,825	1,760	1,690	1,614	1,530	1,433	1,310
∞	2,048	1,945	1,833	1,709	1,640	1,556	1,484	1,388	1,268	1,000

Kvantily rozložení $F_{0,99}(\nu_1,\nu_2)$ - 1. část

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9
1	4052,2	4999,5	5403,5	5624,6	5763,7	5859,0	5928,3	5981,6	6022,5
2	98,503	99,000	99,166	99,249	99,299	99,332	99,356	99,374	99,388
3	34,116	30,817	29,457	28,710	28,237	27,911	27,672	27,489	27,345
4	21,198	18,000	16,694	15,977	15,522	15,207	14,976	14,799	14,639
5	16,258	13,274	12,060	11,392	10,967	10,672	10,456	10,289	10,158
6	13,745	10,925	9,780	9,148	8,746	8,466	8,260	8,102	7,976
7	12,246	9,547	8,451	7,847	7,460	7,191	6,993	6,840	6,719
8	11,259	8,649	7,591	7,006	6,632	6,371	6,178	6,029	5,911
9	10,561	8,022	6,992	6,422	6,057	5,802	5,613	5,467	5,351
10	10,044	7,559	6,552	5,994	5,636	5,386	5,200	5,057	4,942
11	9,646	7,206	6,217	5,668	5,316	5,069	4,886	4,745	4,632
12	9,330	6,927	5,953	5,412	5,064	4,821	4,640	4,499	4,388
13	9,074	6,701	5,739	5,205	4,862	4,620	4,441	4,302	4,191
14	8,862	6,515	5,564	5,035	4,695	4,456	4,278	4,140	4,030
15	8,683	6,359	5,417	4,893	4,556	4,318	4,142	4,005	3,895
16	8,531	6,226	5,292	4,773	4,437	4,202	4,026	3,890	3,780
17	8,400	6,112	5,185	4,669	4,336	4,102	3,927	3,791	3,682
18	8,285	6,013	5,092	4,579	4,428	4,015	3,841	3,705	3,597
19	8,185	5,926	5,010	4,500	4,171	3,939	3,765	3,631	3,523
20	8,096	5,849	4,938	4,431	4,103	3,871	3,699	3,564	3,457
21	8,017	5,780	4,874	4,369	4,042	3,812	3,640	3,506	3,398
22	7,945	5,719	7,817	4,313	3,988	3,758	3,587	3,453	3,346
23	7,881	5,664	4,765	4,264	3,939	3,710	3,539	3,406	3,299
24	7,823	5,614	4,718	4,218	3,895	3,667	3,496	3,363	3,256
25	7,770	5,568	4,676	4,177	3,855	3,627	3,457	3,324	3,217
26	7,721	5,526	4,637	4,140	3,818	3,591	3,421	3,288	3,182
27	7,677	5,488	4,601	4,106	3,785	3,558	3,388	3,256	3,149
28	7,636	5,453	4,568	4,074	3,754	3,528	3,358	3,226	3,120
29	7,598	5,421	4,538	4,045	3,725	3,500	3,330	3,198	3,092
30	7,563	5,390	4,510	4,018	3,699	3,474	3,305	3,173	3,067
40	7,314	5,179	4,313	3,828	3,514	3,291	3,124	2,993	2,888
60	7,077	4,977	4,126	3,649	3,339	3,119	2,953	2,823	2,719
120	6,851	4,787	3,949	3,480	3,174	2,956	2,792	2,663	2,559
∞	6,635	4,605	3,782	3,319	3,017	2,802	2,639	2,511	2,407

Kvantily rozložení $F_{0,99}(\nu_1,\nu_2)$ - 2. část

ν_2 \ ν_1	10	12	15	20	24	30	40	60	120	∞
1	6055,8	6106,3	6157,3	6208,7	6234,6	6260,7	6286,8	6313,0	6339,4	6366,0
2	99,399	99,416	99,432	99,449	99,458	99,466	99,474	99,483	99,491	99,501
3	27,229	27,052	26,872	26,690	26,598	26,505	26,411	26,316	26,221	26,125
4	14,546	14,374	14,198	14,020	13,929	13,838	13,745	13,652	13,558	13,463
5	10,051	9,888	9,722	9,553	9,467	9,379	9,291	9,202	9,112	9,020
6	7,874	7,718	7,559	7,396	7,313	7,229	7,143	7,057	6,969	6,880
7	6,620	6,469	6,314	6,155	6,074	5,992	5,908	5,824	5,737	5,650
8	5,814	5,667	5,515	5,359	5,279	5,198	5,116	5,032	4,946	4,859
9	5,257	5,111	4,962	4,808	4,729	4,649	4,567	4,483	4,398	4,311
10	4,849	4,706	4,558	4,405	7,327	4,247	4,165	4,082	3,997	3,909
11	4,539	4,397	4,251	4,099	4,021	3,941	3,860	3,776	3,690	3,603
12	4,296	4,155	4,010	3,858	3,781	3,701	3,619	3,536	3,449	3,361
13	4,100	3,960	3,815	3,665	3,587	3,507	3,425	3,341	3,255	3,165
14	3,939	3,800	3,656	3,505	3,427	3,348	3,266	3,181	3,094	3,004
15	3,805	3,666	3,522	3,372	3,294	3,214	3,132	3,047	2,960	2,868
16	3,691	3,553	3,409	3,259	3,181	3,101	3,018	2,933	2,845	2,753
17	3,593	3,455	3,312	3,162	3,084	3,003	2,921	2,835	2,746	2,653
18	3,508	3,371	3,227	3,077	2,999	2,919	2,835	2,749	2,660	2,566
19	3,434	3,297	3,153	3,003	2,925	2,844	2,761	2,674	2,584	2,489
20	3,368	3,231	3,088	2,938	2,859	2,779	2,695	2,608	2,517	2,421
21	3,310	3,173	3,030	2,880	2,801	2,720	2,636	2,548	2,457	2,360
22	3,258	3,121	2,978	2,827	2,749	2,668	2,583	2,495	2,403	2,306
23	3,211	3,074	2,931	2,781	2,702	2,620	2,536	2,447	2,354	2,256
24	3,168	3,032	2,889	2,738	2,659	2,577	2,492	2,404	2,310	2,211
25	3,129	2,993	2,850	2,699	2,620	2,538	2,453	2,364	2,270	2,169
26	3,094	2,958	2,815	2,664	2,585	2,503	2,417	2,327	2,233	2,132
27	3,062	2,926	2,783	2,632	2,552	2,470	2,384	2,294	2,198	2,097
28	3,032	2,896	2,753	2,602	2,522	2,440	2,354	2,263	2,167	2,064
29	3,005	2,869	2,726	2,574	2,495	2,412	2,325	2,234	2,138	2,034
30	2,979	2,843	2,700	2,549	2,469	2,386	2,299	2,208	2,111	2,006
40	2,801	2,665	2,522	2,369	2,288	2,203	2,114	2,019	1,917	1,805
60	2,632	2,496	2,352	2,198	2,115	2,029	1,936	1,836	1,726	1,601
120	2,472	2,336	2,192	2,035	1,950	1,860	1,763	1,656	1,533	1,381
∞	2,321	2,185	2,039	1,878	1,791	1,696	1,592	1,473	1,325	1,000

Kvantily rozložení $F_{0,995}(\nu_1,\nu_2)$ - 1. část

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9
1	16211	20000	21615	22500	23056	23437	23715	23925	24091
2	198,50	199,00	199,17	199,25	199,30	199,33	199,36	199,37	199,39
3	55,552	49,799	47,467	46,196	45,392	44,838	44,434	44,126	43,882
4	31,333	26,284	24,259	23,155	22,456	21,975	21,622	21,352	21,139
5	22,785	18,314	16,530	15,556	14,940	14,513	14,200	13,961	13,772
6	18,635	14,544	12,917	12,028	11,464	11,073	10,786	10,566	10,391
7	16,236	12,404	10,882	10,050	9,522	9,155	8,885	8,678	8,514
8	14,688	11,042	9,597	8,805	8,3302	7,952	7,694	7,496	7,339
9	13,614	10,107	8,717	7,956	7,471	7,134	6,885	6,693	6,541
10	12,826	9,427	8,081	7,343	6,872	6,545	6,303	6,116	5,968
11	12,226	8,912	7,600	6,881	6,422	6,102	5,865	5,682	5,537
12	11,754	8,510	7,226	6,521	6,071	5,757	5,525	5,345	5,202
13	11,374	8,187	6,926	6,234	5,791	5,482	5,253	5,076	4,935
14	11,060	7,922	6,680	5,998	5,562	5,257	5,031	4,857	4,717
15	10,798	7,701	6,476	5,803	5,372	5,071	4,847	4,674	4,536
16	10,575	7,514	6,303	5,638	5,212	4,913	4,692	4,521	4,384
17	10,384	7,354	6,156	5,497	5,075	4,779	4,559	4,389	4,254
18	10,218	7,215	6,028	5,375	4,956	4,663	4,445	4,276	4,141
19	10,073	7,094	5,916	5,268	4,853	4,561	4,345	4,177	4,043
20	9,944	6,987	5,818	5,174	4,762	4,472	4,257	4,090	3,956
21	9,830	6,891	5,730	5,091	4,681	4,393	4,179	4,013	3,880
22	9,727	6,806	5,652	5,017	4,609	4,323	4,109	3,944	3,812
23	9,635	6,730	5,582	4,950	4,544	4,259	4,047	3,882	3,750
24	9,551	6,661	5,519	4,890	4,486	4,202	3,991	3,826	3,695
25	9,475	6,598	5,462	4,835	4,433	4,150	3,939	3,776	3,645
26	9,406	6,541	5,409	4,785	4,384	4,103	3,893	3,730	3,599
27	9,342	6,489	5,361	4,740	4,340	4,059	3,850	3,688	3,557
28	9,284	6,440	5,317	4,698	4,300	4,020	3,811	3,649	3,519
29	9,230	6,396	5,276	4,659	4,262	3,983	3,775	3,613	3,483
30	9,180	6,355	5,239	4,623	4,228	3,949	3,742	3,580	3,451
40	8,828	6,066	4,976	4,374	3,986	3,713	3,509	3,350	3,222
60	8,495	5,795	4,729	4,140	3,760	3,492	3,291	3,134	3,008
120	8,179	5,539	4,497	3,921	3,548	3,285	3,087	2,933	2,808
∞	7,879	5,298	4,279	3,715	3,350	3,091	2,897	2,744	2,621

Kvantily rozložení $F_{0,995}(\nu_1,\nu_2)$ - 2. část

ν_2 \ ν_1	10	12	15	20	24	30	40	60	120	∞
1	24,224	24426	24630	24836	24940	25044	25148	25253	25359	25465
2	199,40	199,42	199,43	199,45	199,46	199,47	199,47	199,48	199,49	199,51
3	43,686	43,387	43,085	42,778	42,622	42,466	42,308	42,149	41,989	41,829
4	20,967	20,705	20,438	20,167	20,030	19,892	19,752	19,611	19,468	19,325
5	13,618	13,384	13,146	12,903	12,780	12,656	12,530	12,402	12,274	12,144
6	10,250	10,034	9,814	9,589	9,474	9,358	9,241	9,122	9,002	8,879
7	8,380	8,176	7,968	7,754	7,645	7,735	7,423	7,309	7,193	7,076
8	7,211	7,015	6,814	6,608	6,503	6,396	6,288	6,177	6,065	5,951
9	6,417	6,227	6,033	5,832	5,729	5,625	5,519	5,410	5,300	5,188
10	5,847	5,661	5,471	5,274	5,173	5,071	4,966	4,859	4,750	4,639
11	5,418	5,236	5,049	4,855	4,756	4,654	4,551	4,445	4,337	4,226
12	5,086	4,906	4,721	4,530	4,432	4,331	4,228	4,123	4,015	3,904
13	4,820	4,643	4,460	4,270	4,173	4,073	3,970	3,866	3,758	3,647
14	4,603	4,428	4,247	4,059	3,961	3,862	3,760	3,655	3,547	3,436
15	4,424	4,250	4,070	3,883	3,786	3,687	3,585	3,480	3,372	3,260
16	4,272	4,099	3,921	3,734	3,638	3,538	3,437	3,332	3,224	3,112
17	4,142	3,971	3,793	3,607	3,511	3,412	3,311	3,206	3,097	2,984
18	4,031	3,860	3,683	3,498	3,402	3,303	3,201	3,096	2,987	2,873
19	3,933	3,763	3,587	3,402	3,306	3,208	3,106	3,000	2,891	2,776
20	3,847	3,678	3,502	3,318	3,222	3,123	3,022	2,916	2,806	2,690
21	3,771	3,602	3,427	3,243	3,147	3,049	2,947	2,841	2,730	2,614
22	3,703	3,535	3,360	3,176	3,081	2,982	2,880	2,774	2,663	2,546
23	3,642	3,475	3,300	3,117	3,021	2,922	2,820	2,713	2,602	2,484
24	3,587	3,420	3,246	3,062	2,967	2,868	2,765	2,659	2,546	2,428
25	3,537	3,370	3,196	3,013	2,918	2,819	2,716	2,609	2,496	2,377
26	3,492	3,325	3,152	2,969	2,873	2,774	2,671	2,563	2,450	2,330
27	3,450	3,284	3,110	2,928	2,832	2,733	2,630	2,522	2,408	2,287
28	3,412	3,246	3,073	2,890	2,794	2,695	2,592	2,483	2,369	2,247
29	3,377	3,211	3,038	2,855	2,759	2,660	2,557	2,448	2,333	2,210
30	3,344	3,179	3,006	2,823	2,727	2,628	2,524	2,415	2,300	2,176
40	3,117	2,953	2,781	2,598	2,502	2,402	2,296	2,184	3,064	1,932
60	2,904	2,742	2,571	2,387	2,290	2,187	2,079	1,962	1,834	1,688
120	2,705	2,544	2,373	2,188	2,089	1,984	1,871	1,747	1,606	1,431
∞	2,519	2,358	2,187	2,000	1,898	1,789	1,669	1,533	1,364	1,000

Literatura

- Budíková, Marie - Mikoláš, Štěpán - Osecký, Pavel. *Popisná statistika*. 3., doplněné vyd. Brno : Masarykova univerzita, 1998. 52 s. ISBN 80-210-1831-3.

- Budíková, Marie - Mikoláš, Štěpán - Osecký, Pavel. *Teorie pravděpodobnosti a matematická statistika. Sbírka příkladů*. 3. vyd. Brno : Masarykova univerzita, 2004. 127 s. ISBN 80-210-3313-4.

- Michal Friesl - výukové texty (např. Pravděpodobnost a statistika, Posbírané příklady z pravděpodobnosti a statistiky,…): http://home.zcu.cz/~friesl/Archiv/DldTeach.html

- Blanka Šedivá - Pravděpodobnost a statistika: http://home.zcu.cz/~sediva/pse/

- Michal Čihák - výukové texty: http://www.cihak.com/michal/

- Petr Otipka, Vladislav Šmajstrla - Pravděpodobnost a statistika: http://homen.vsb.cz/~oti73/cdpast1/

- Jana Novovičová - Pravděpodobnost a matematická statistika: http://euler.fd.cvut.cz/publikace /files/skripta3.pdf