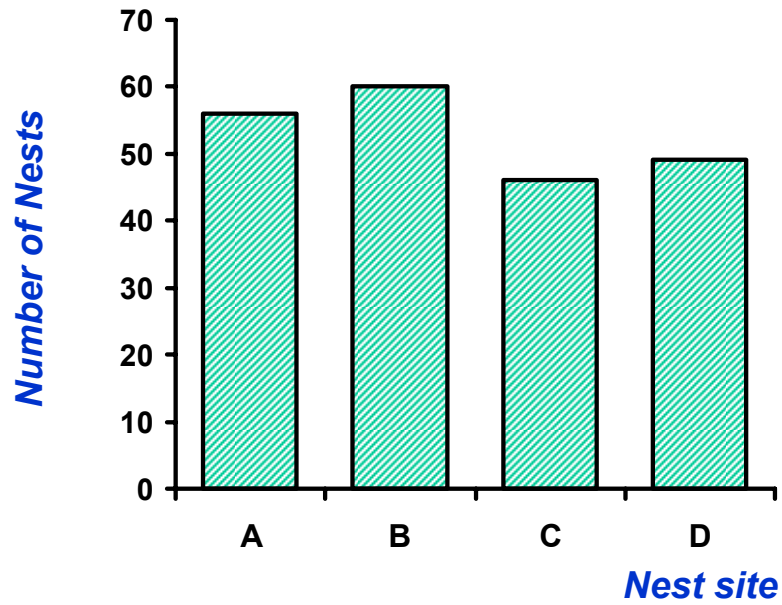
PŘÍKLADY I

A frequency table of nominal data

The location of sparrow nests.

Nest site		Number of nests observed
A.	Vines	**56**
B.	Building eaves	**60**
C.	Low tree branches	**46**
D.	Tree and building cavities	**49**

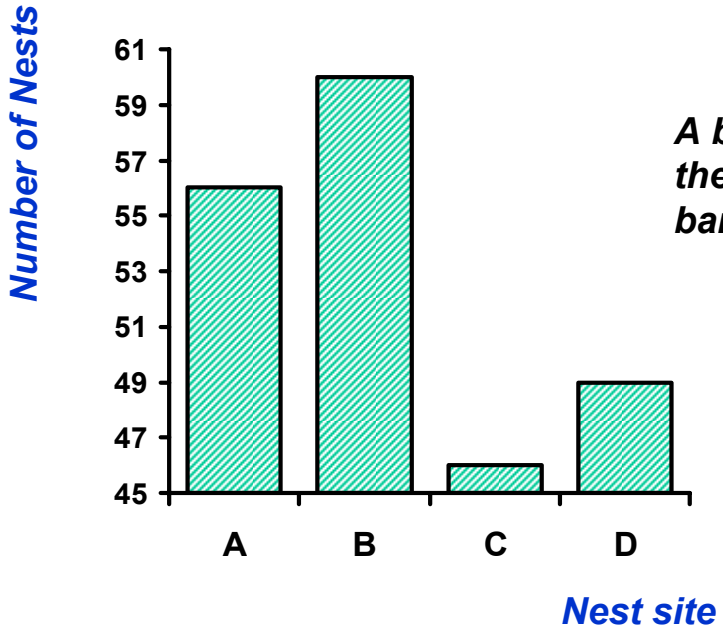


A bar graph of the sparrow nest data. An example of a <u>bar graph</u> for nominal data.

A frequency table of nominal data

The location of sparrow nests.

Nest site		Number of nests observed
A.	Vines	**56**
B.	Building eaves	**60**
C.	Low tree branches	**46**
D.	Tree and building cavities	**49**

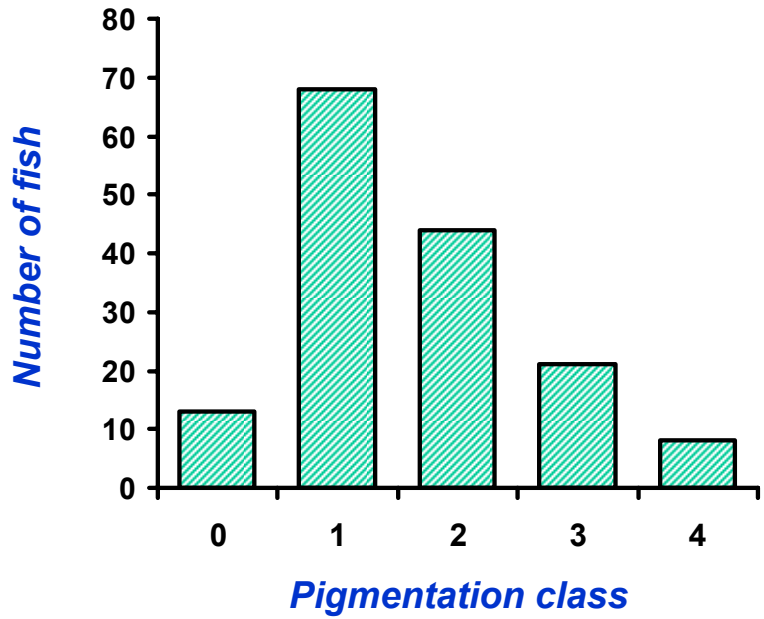


A bar graph of the sparrow nest data, drawn with the vertical axis starting at 45. Compare this with bar graph, where the axis starts at 0.

A frequency table of ordinal data

Numbers of sunfish, tabulated according to amount of black pigmentation

Pigmentation class	Amount of pigmentation	Number of fish
0	No black pigmentation	13
1	Faintly speckled	68
2	Moderately speckled	44
3	Heavily speckled	21
4	Solid black pigmentation	8

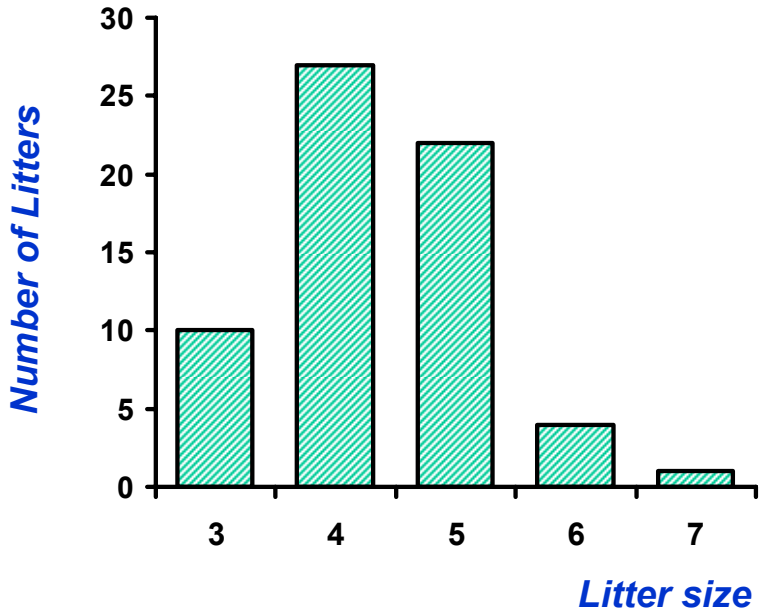


A bar graph of the sunfish pigmentation data. An example of a bar graph for ordinal data.

A frequency table of discrete data

Frequency of occurrence of various litter sizes in foxes

Litter size	Frequency
3	10
4	27
5	22
6	4
7	1



***A bar graph of the fox litter data. An example of a* bar graph for discrete, ratio scale, data.**

A frequency table of a discrete data

Number of aphids observed per clover plant

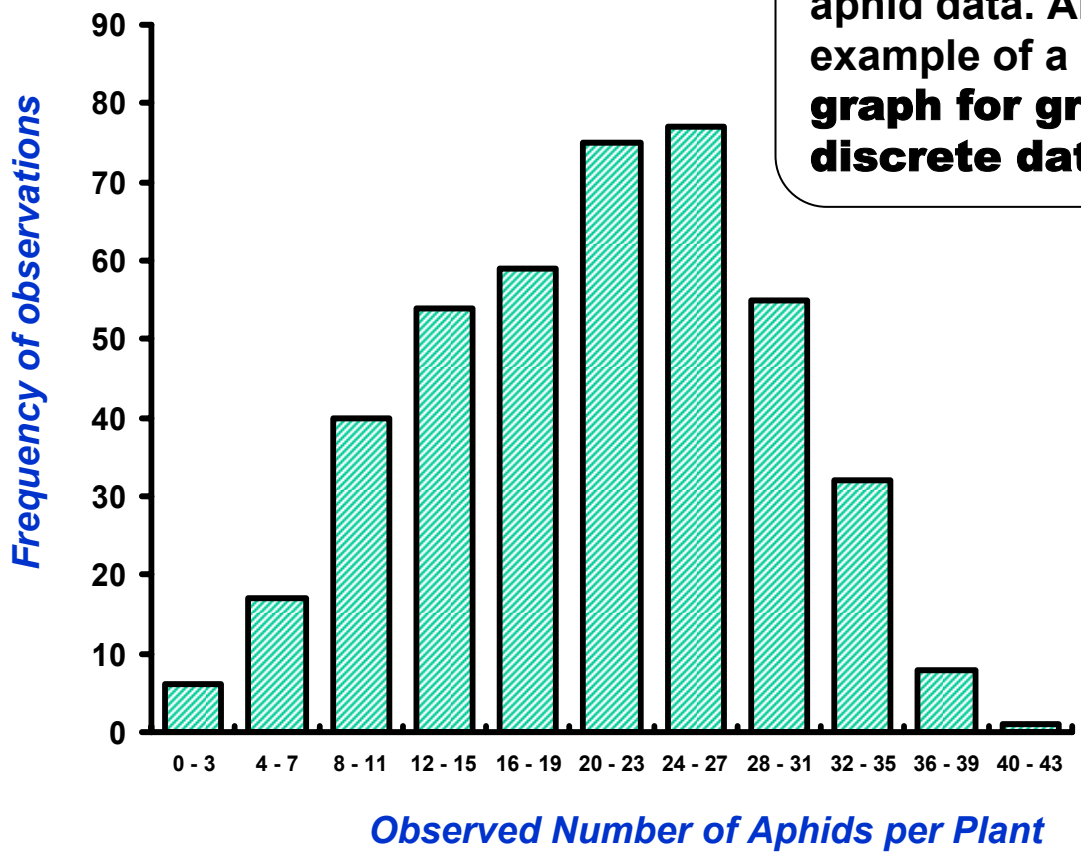
Number of aphids on a plant	Number of plants observed	Number of aphids on a plant	Number of plants observed
0	3	20	17
1	1	21	18
2	1	22	23
3	1	23	17
4	2	24	19
5	3	25	18
6	5	26	19
7	7	27	21
8	8	28	18
9	11	29	13
10	10	30	10
11	11	31	14
12	13	32	9
13	12	33	10
14	16	34	8
15	13	35	5
16	14	36	4
17	16	37	1
18	15	38	2
19	14	39	1
		40	0
		41	1

Total number of observations = 424

A frequency table of a discrete data

Number of aphids observed per clover plant

Number of aphids on a plant	Number of plants observed
0 - 3	6
4 - 7	17
8 - 11	40
12 - 15	54
16 - 19	59
20 - 23	75
24 - 27	77
28 - 31	55
32 - 35	32
36 - 39	8
40 - 43	1



A bar graph of the aphid data. An example of a bar **graph for grouped discrete data.**

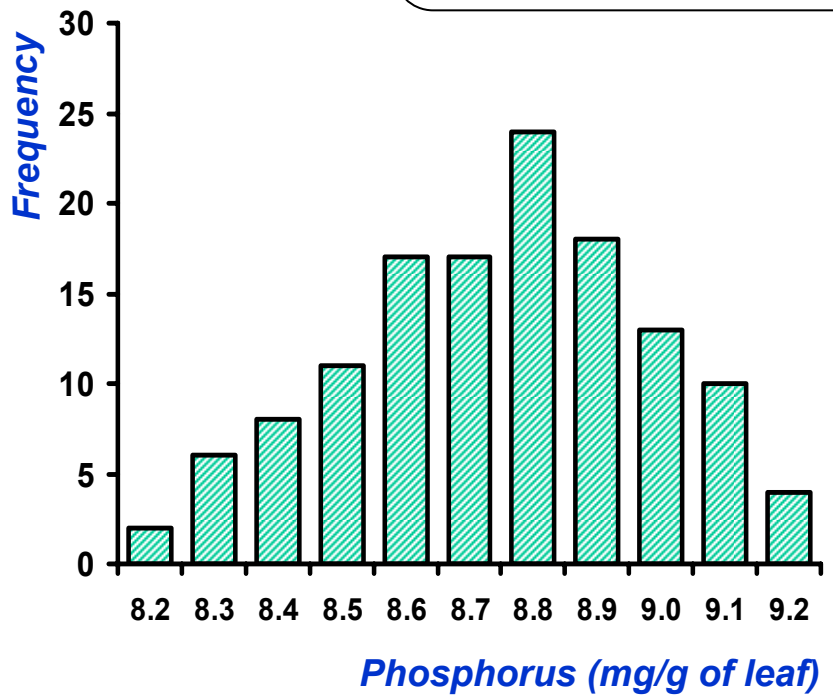
A frequency table of continuous data

Determinations of the amount of phosphorus in leaves

Phosphorus (mg/g of leaf)	Frequency (i.e.,number of determinations)	Cumulative frequency	
		Starting with low values	Starting with high values
8,15 - 8,25	2	2	130
8,25 – 8,35	6	8	128
8,35 – 8,45	8	16	122
8,45 – 8,55	11	27	114
8,55 – 8,65	17	44	103
8,65 – 8,75	17	61	86
8,75 – 8,85	24	85	69
8,85 – 8,95	18	103	45
8,95 – 9,05	13	116	27
9,05 – 9,15	10	126	14
9,15 – 9,25	4	130	4

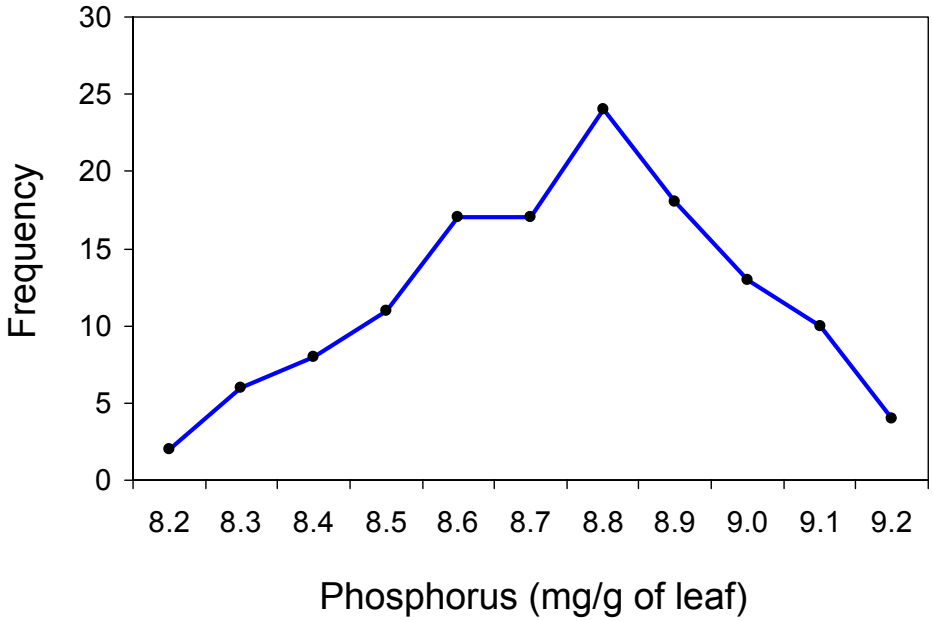
Total frequency = 130

A histogram of the leaf phosphorus data. An example of a histogram for continuous data (based on equal interval width)..



A frequency table of continuous data

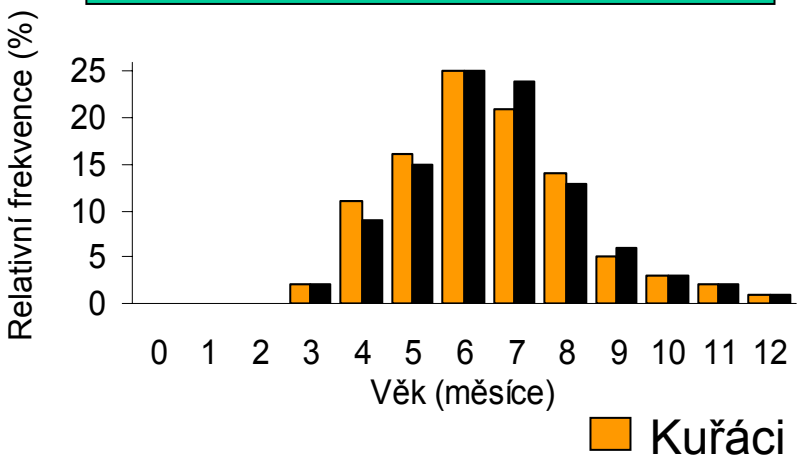
Determinations of the amount of phosphorus in leaves



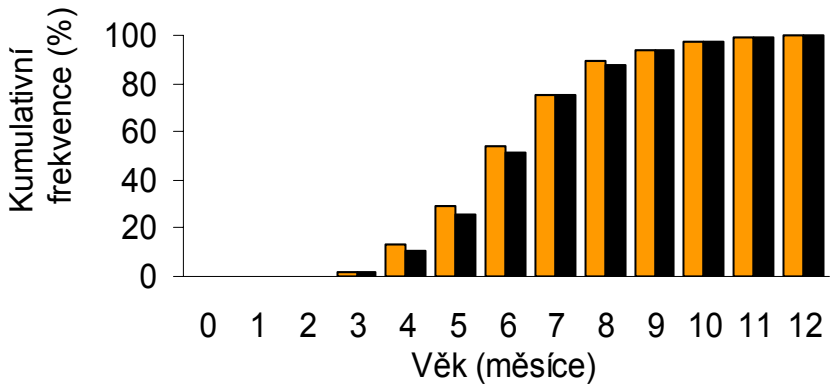
***A frequency polygon** for the leaf phosphorus data*

Grafický popis rozložení - příklad

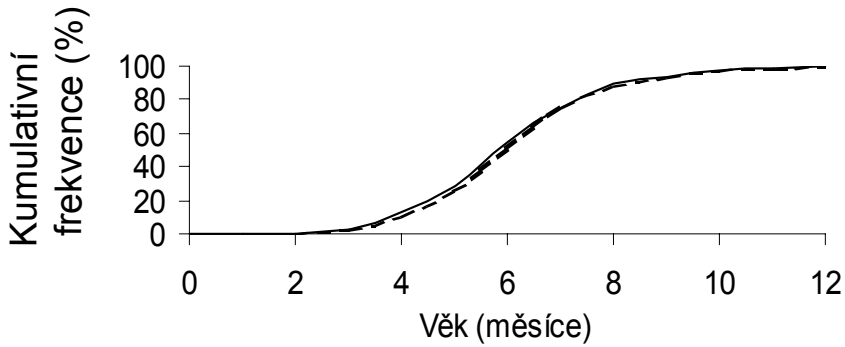
Histogram: relativní frekvence



Histogram: kumulativní relativní frekvence

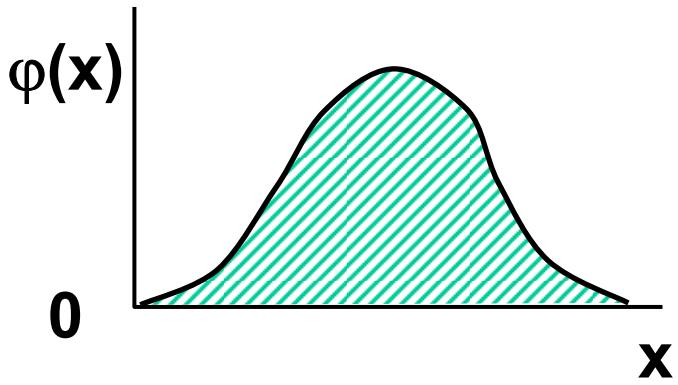


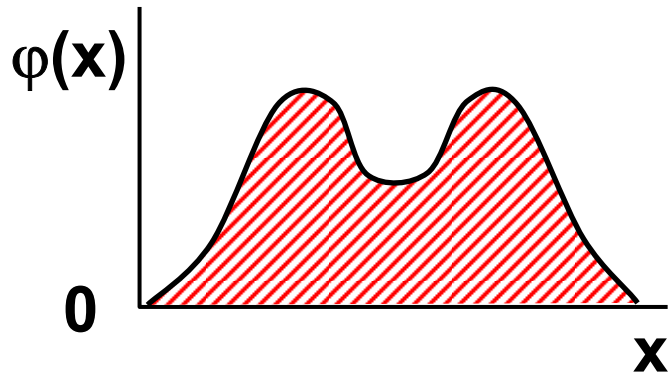
Křivka relativní kumulativní frekvence

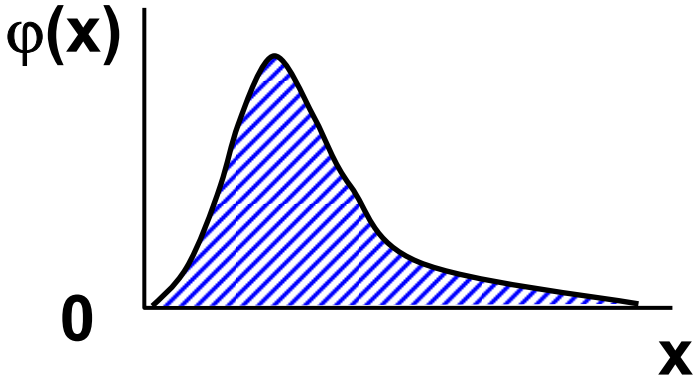


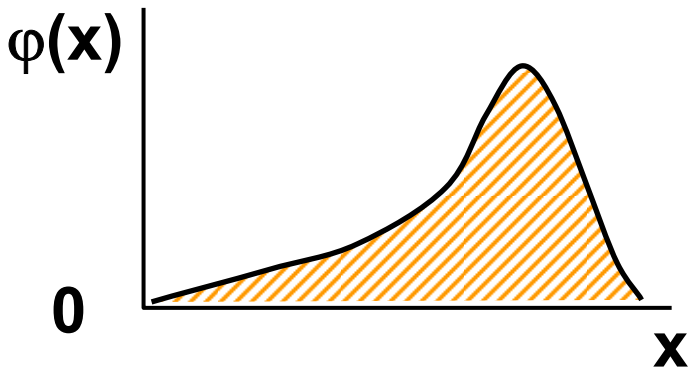
Věk prvního růstu zubů u dětí kuřáků (——) a nekuřáků (-------)
(Rantakalio and Mäkinen, 1984)

Příklad: spojitá čísla mohou mít různá rozložení

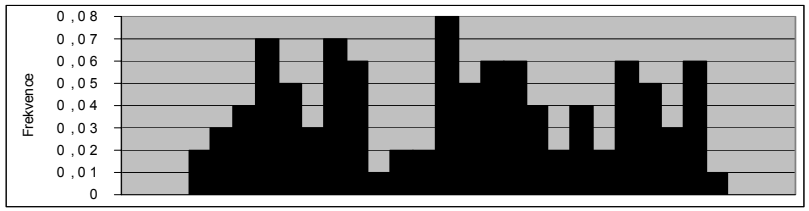




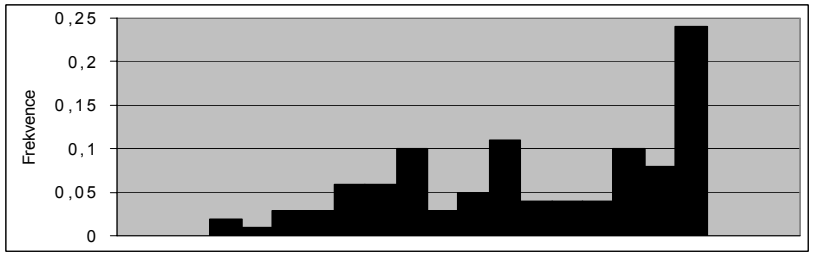




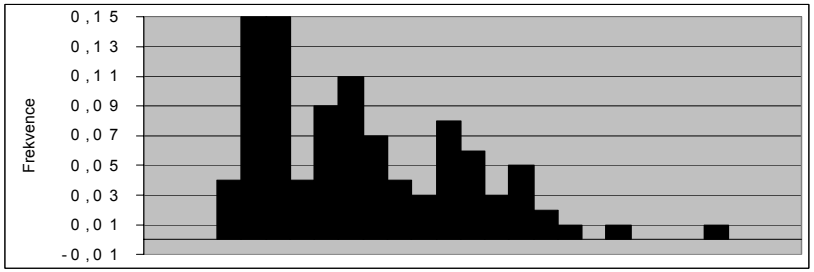
Histogram - tvar rozložení a relevantní ukazatel středu



Symetrické rozložení,
medián je blízko průměru



Asymetrické rozložení,
kde průměr je
menší než medián

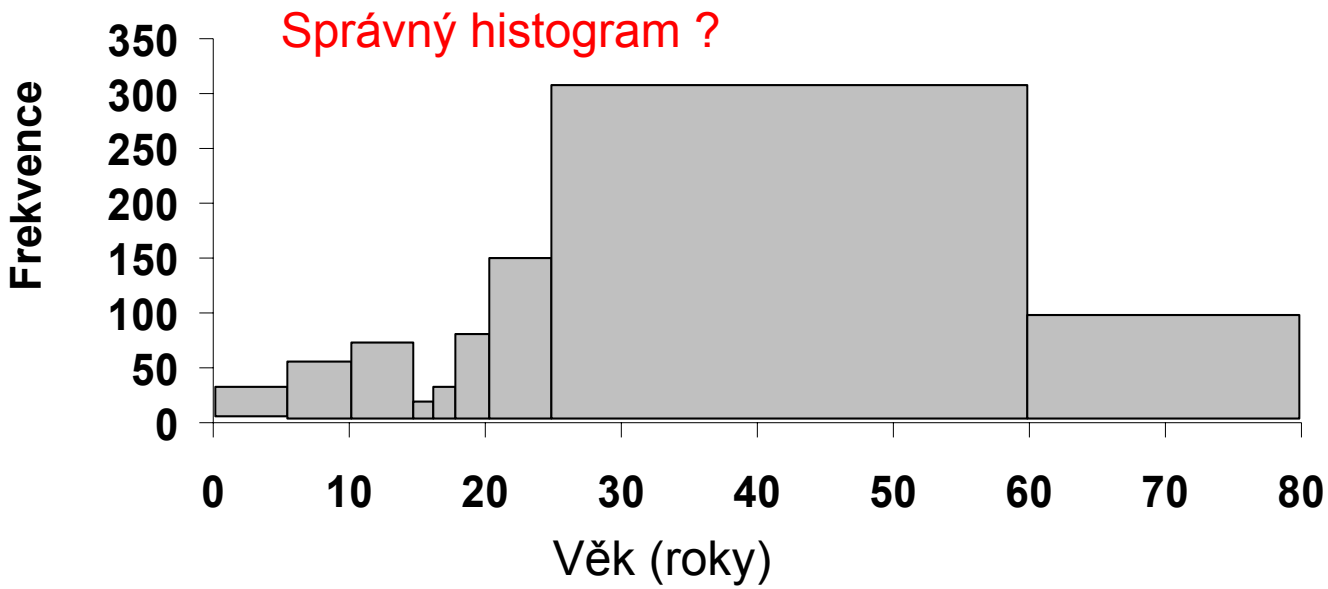


Asymetrické rozložení,
kde průměr je
větší než medián

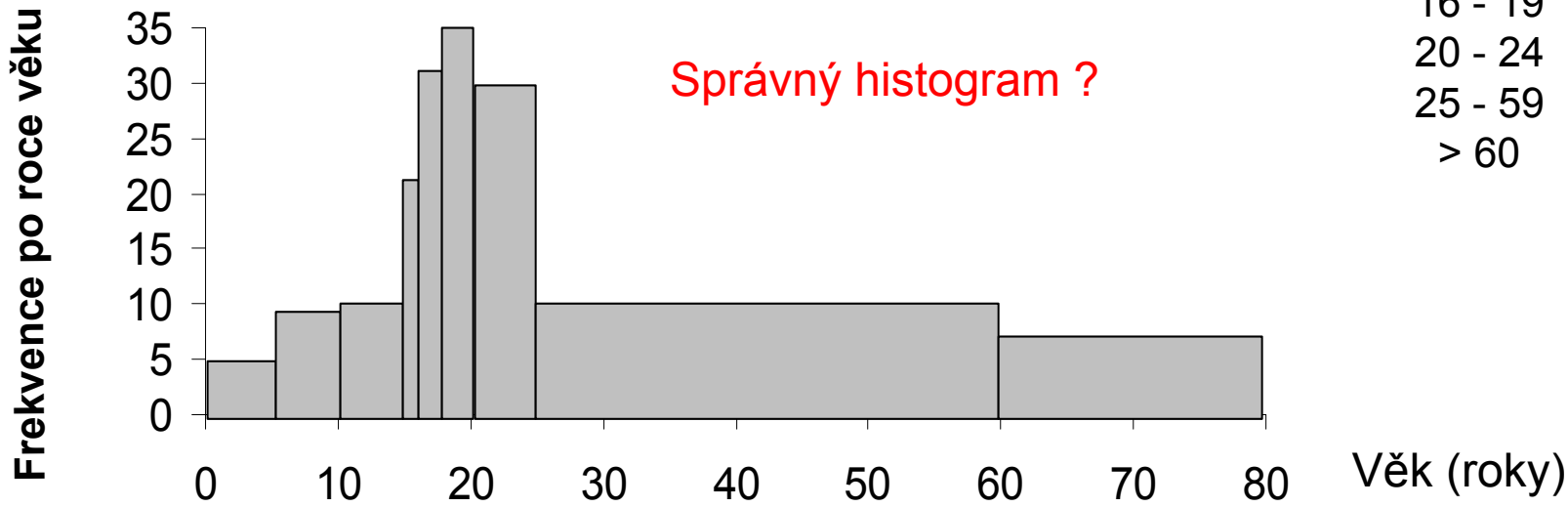
Reálný význam mediánu a průměru jako ukazatelů středu rozložení bude záviset na charakteru sledovaného znaku (např. znečištění vody v určité oblasti dusičnany; respirace půdy po ovlivnění kontaminantem; koncentrace látky v krvi pokusných zvířat).

Při posuzování rozložení sledovaného znaku v cílové populaci je nutné uvážit jak velký výběr (n), na základě kterého byly zobrazené histogramy spojeny.

Příklad: věk účastníků vážných dopravních nehod



Věk	f
0 - 4	28
5 - 9	46
10 - 15	58
16 - 19	20
20 - 24	114
25 - 59	316
> 60	103



PŘÍKLADY II

Sumární statistiky středu

Modus

Medián

Aritmetický průměr $$\overline{X} = \frac{\sum_i x_i}{n}$$

Geometrický průměr $$\overline{X}_G = \sqrt[n]{x_1 . x_2 . x_3 \ldots x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

Harmonický průměr $$\overline{X}_H = \frac{1}{\frac{1}{n}\sum_i \frac{1}{x_i}} = \frac{n}{\sum \frac{1}{x_i}}$$

Výpočet mediánu z primárních dat

A. Lichý počet (n)

Vzorek:

5; 1; 8; 3; 4

Medián - pořadí:

(n + 1) / 2 = 3. číslo
= 4

B. Sudý počet (n)

Vzorek:

1; 3; 4; 5; 7; 8

Medián - pořadí:

(n / 2) ; [(n + 2) / 2]
= (4 + 5)/2 = 4.5

Průměr a medián u frekvenčně tříděných dat

I. Dostupná původní data

x: Měsíční výdaje rodiny na bydlení
f: frekvence

xi	3,3	3,4	3,5	3,6	3,7	3,8	3,9	4,0	4,1	4,2	4,3	4,4	4,5
fi	1	0	1	2	1	3	3	4	3	2	2	2	1

Průměr: $\overline{x} = \frac{\sum x_i f_i}{\sum f_i} = 3,976$

Medián: **13-té číslo = 4,0**

Při současném odhadu mediánu a průměru jako ukazatelů středu symetrických rozložení je medián méně přesný než průměr.

Examples

Example 3.1

A sample from a population of butterfly wing lengths.

Xi(cm)	Xi(cm)
3.3	4.0
3.5	4.0
3.6	4.0
3.6	4.1
3.7	4.1
3.8	4.1
3.8	4.2
3.8	4.2
3.9	4.3
3.9	4.3
3.9	4.4
4.0	4.5

$$\sum X_i = 95.0 \ cm$$

$$n = 24$$

$$\overline{X} = \frac{\sum X_i}{n} = \frac{95.0 \ cm}{24} = 3.96 \ cm$$

Example 3.2

The data from Example 3.1 recorded as a frequency table.

X_i(cm)	f_i	$f_i X_i$(cm)
3.3	1	3.3
3.4	0	0
3.5	1	3.5
3.6	2	7.2
3.7	1	3.7
3.8	3	11.4
3.9	3	11.7
4.0	4	16.0
4.1	3	12.3
4.2	2	8.4
4.3	2	8.6
4.4	1	4.4
4.5	1	4.5
$\sum f_i = 24$		$\sum f_i X_i = 95.0 \ cm$

$$\sum f_i = n = 24$$

$$\overline{X} = \frac{\sum f_i X_i}{n} = \frac{95.0 \ cm}{24} = 3.96 \ cm$$

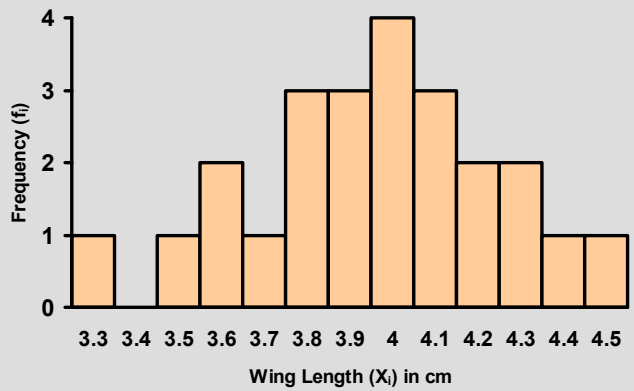
$$median = 3.95 \ cm + \left(\frac{1}{4}\right)(0.1 \ cm)$$

$$= 3.95 \ cm + 0.025 \ cm$$

$$= 3.975 \ cm$$

Figure 3.1

A histogram of the data in Example 3.2. The mean (3.96 cm) is the center of gravity of the histogram, and the median (3.975 cm) divides the histogram into two equal areas.



Examples

Example 3.3

Life expectancy of two hypothetical species of birds in captivity.

Species A X_i*(mo)*
34
36
37
39
40
41
42
43
79

$$n = 9$$

$$median = X_5 = 40 \ mo$$

$$\overline{X} = 43.4 \ mo$$

Species B X_i*(mo)*
34
36
37
39
40
41
42
43
44
45

$$n = 10$$

$$median = \frac{X_5 + X_6}{2}$$

$$= \frac{40 \ mo + \ 41 \ mo}{2}$$

$$= 40.5 \ mo$$

$$\overline{X} = 40.1 \ mo$$

PŘÍKLADY III

Rozložení náhodné veličiny, charakteristiky dat, Testy hypotéz, odhady

Příklad 1.

Nakreslete schematicky graf Gausovy křivky pro standardizované normální rozložení a pomocí symbolu A vyjádřete následující pravděpodobnosti:

Pravděpodobnost, že hodnota sled. veličiny	Symbol
leží mezi 0 a Z	A
leží mezi –Z a Z	2A
leží mimo interval –Z,+Z	1-2A
je menší než Z (Z je kladné)	2A+(1-2A)/2=1/2+A
je menší než Z (Z je záporné)	(1-2A)/2
je větší než Z (Z je kladné)	(1-2A)/2
je větší než Z (Z je záporné)	2A+(1-2A)/2=1/2+A

Příklady – rozložení, odhady

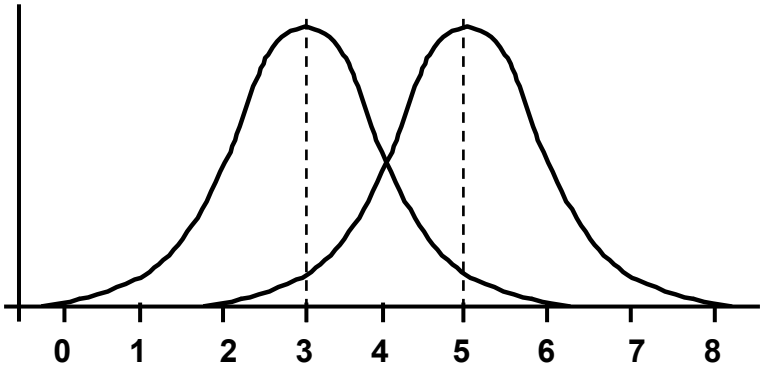
Příklad 2.

A. Zakreslete schematicky následující dvojice rozložení:

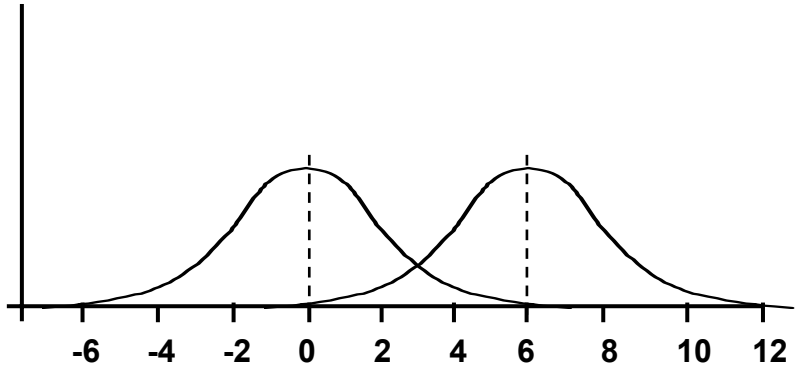
a) N(μ = 5,σ = 1) a N(μ = 3, σ = 1)

b) N(μ = 0, σ = 2) a N(μ = 6, σ = 2)

a)



b)



Příklad 2.

B. Najděte následující kvantily.

a) 95 % kvantil Studentova rozložení pro výběr o n = 20 $t_{0,95}^{(20-1)} = 1,7291$

b) 95 % kvantil Studentova rozložení pro výběr o n = 120 $t_{0,95}^{(120-1)} = 1,6578$

Příklady – rozložení, odhady

C. „Z skóre“

Hodnota kostní dřeně je u pacientů s určitým typem onkologického onemocnění hodnocena podle tzv. „Z skóre“, vycházejícího z přepočtu na standardizované normální rozložení

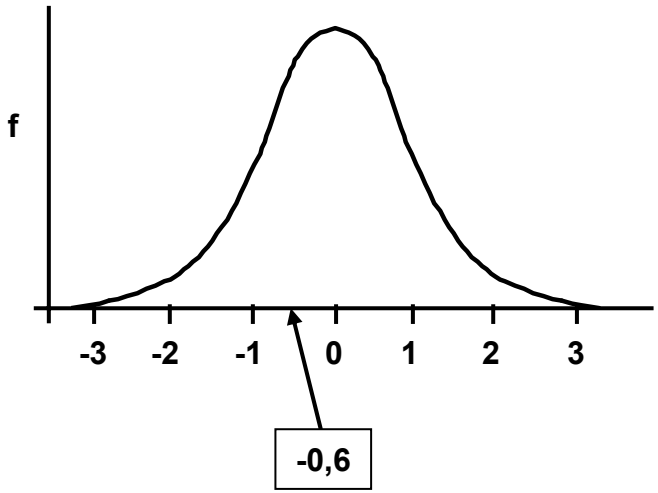
a) Vysvětlete jakou formou takové hodnoty vznikají, jaký mají smysl a jak probíhá hodnocení konkrétního pacienta

b) Jakou pravděpodobnostní pozici má v dané populaci jedinec s hodnotou Z skóre – 0.6

c) Je porovnávání jedinců z různých populací pomocí Z skóre závislé na variabilitě (rozptylu) původních dat ?

Hodnoty Z-skóre vycházejí z přepočtu na standardizované normální rozdělení. Pro jejich získání se od každé odečte střední hodnota souboru a podělí směrodatnou odchylkou souboru. $z = \frac{x - \mu}{\sigma}$

Tyto hodnoty mají potom střední hodnotu nulovou s jednotkovým rozptylem. Z grafu rozložení je potom možné odečítat jednotlivé hodnoty Z-skóre. Z-skóre je závislé na variabilitě původních dat.



Příklady – rozložení, odhady

Příklad 3.

a) Jak velká část hodnot náhodné veličiny X, která má normální rozložení, leží mezi –1,76s a +1.76s?

1.76 je hodnota kvantilu normálního rozložení u_p pro p=0,96, tedy v intervalu –1,76s a +1.76s leží 96% hodnot náhodné veličiny X

b) Koncentrace toxické chemikálie v tkáních ryb z jezera, které je kontaminováno továrnou produkující celulózu, byla shledána přibližně normální s průměrem 67.56 ng/kg tkáně a směrodatnou odchylkou 2.57 ng/kg. Rozložení této sledované veličiny bylo odhadováno na základě mnohonásobné analýzy vzorků ryb (každý *o* 30 rybách); výsledkem analýzy každého vzorku je průměrná koncentrace látky na 1 kg tkáně. Jak velký podíl vzorků má koncentrace nižší než 62 ng/kg?

m = 67.56; s = 2.57

$$P\left(X < \frac{62-67.56}{2.57}\right) = P(X < -2.16) = P(X > 2.16) = 1 - F(2.16) = 0.015$$

tedy vzorků s koncentrací nižší než 62ng/kg je 1.5%.

Najděte takovou koncentraci chemikálie, kterou může v jezeře překročit 5 % populace ryb.

hledáme hodnotu, pro kterou bude platit, že 95% vzorků má nižší koncentraci než tato hodnota, tedy:

$$0.05 = P\left(X > \frac{\mu-67.56}{2.57}\right) = 1 - P\left(X < \frac{\mu-67.56}{2.57}\right) = 1 - F\left(\frac{\mu-67.56}{2.57}\right)$$

$$0.95 = F\left(\frac{\mu-67.56}{2.57}\right) = F(1.65) \quad \Rightarrow \quad \mu = 1.65*2.57 + 67.56 = 71.08$$

5% populace ryb překročí hodnotu chemikálie 71.08ng/kg

Příklady – rozložení, odhady

Příklad 3.

c) Předpokládejme, že podle mezinárodních norem nesmí koncentrace vysoce toxických látek v mléčných výrobcích překročit hranici 30 pg/kg tuku (jde o vymyšlené hodnoty). Výrobce, který hodlá začít zpracovávat mléko od nového dodavatele zjistil, že je schopen produkovat výrobky s průměrnou koncentrací 28 pg/kg, ale se směrodatnou odchylkou 1.6 pg/kg.
Jaký podíl jeho nových výrobků by pravděpodobně nesplnil podmínky pro uvedení na trh?

$$P\left(X > \frac{30-28}{1.6}\right) = 1 - P\left(X < \frac{30-28}{1.6}\right) = 1 - F(1.25) = 1 - 0.8943 = 0.1057$$

10.6% nových výrobků nesplní podmínky pro uvedení na trh

Zavedením přísné kontroly dodávaného mléka by bylo možné snížit rozptyl hodnot při zachování průměrné koncentrace sledovaných látek v mléce na 28 pg/kg.
Jaká by musela být směrodatná odchylka, aby pouze 2 % nové produkce překračovalo povolený limit?

$$0.02 = P\left(X > \frac{30-28}{\sigma}\right) = 1 - P\left(X < \frac{30-28}{\sigma}\right) = 1 - F\left(\frac{30-28}{\sigma}\right)$$

$$0.98 = F\left(\frac{30-28}{\sigma}\right) = F(2.06) \quad \Rightarrow \quad \sigma = 2 / 2.06 = 0.97$$

aby produkce překračovala povolený limit pouze o 2% musí být sm. odchylka jen 0.97 pg/kg

Příklady – rozložení, odhady

Příklad 4.

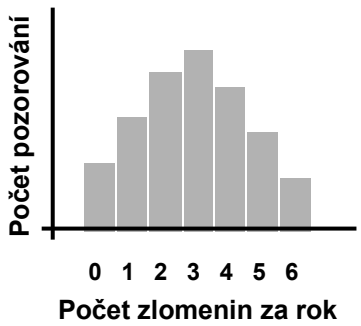
a) U následujícího souboru dat (koncentrace zinku v půdě na deseti sousedících kontaminovaných lokalitách) navrhněte vhodné charakteristiky polohy a rozptylu a vypočítejte je.

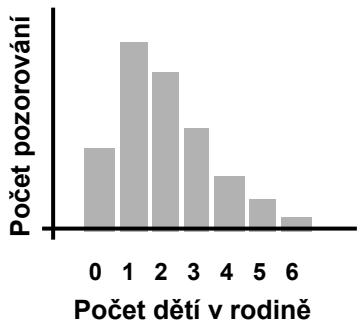
40.60, 40.29, 37.51, 38.90, 38.13, 38.15, 34.81, 37.00, 39.95, 40.43

jako charakteristiku polohy použijeme průměr: $$X = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{10}385{,}77 = 38{,}58$$

jako charakteristiku rozptylu použijeme směrodatnou odchylku: $$S_{\bar{X}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2} = \sqrt{\frac{1}{10}30{,}65} = 1{,}75$$

b) Jaké charakteristiky souboru dat lze přibližně zjistit z histogramu četností? Popište co nejpřesněji soubory dat, které jsou zobrazeny na následujících histogramech:





Z histogramu četností se dá přibližně zjistit modus, minimální a maximální hodnota, kvantily.

Příklady – rozložení, odhady

Příklad 5.

a) Při stanovení průměrného obsahu dusičnanů v říční vodě iontově selektivní elektrodou má měření směrodatnou odchylku σ = 1.5 mg/l.
Kolik vzorků vody musí badatel odebrat (n = ?), pokud požaduje odhad průměrné hodnoty se směrodatnou odchylkou 0.2 mg/l?

$$\sigma = 1.5 mg/l \qquad s_{\bar{x}} = 0.2 mg/l \qquad s_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \quad \Rightarrow \quad n = \left(\frac{\sigma}{s_{\bar{x}}}\right)^2 \doteq 57$$

Badatel musí odebrat 57 vzorků, pokud požaduje odhad průměrné hodnoty se směrodatnou odchylkou 0.2mgl.

b) Odběr jednoho vzorku půdy na běžné stanovení minerálních forem dusíku má cenu 120 Kč.
Na průzkum poměrně rozsáhlé lokality máte k dispozici 12 000 Kč.

1. Máte dostatečné finanční prostředky k odhadu průměrné koncentrace minerálního dusíku na lokalitě tak přesnému, že 95% interval spolehlivosti má šířku 4 jednotky (jednotky o rozměru koncentrace, ve kterém je výsledek vyjádřen); předpokládejte rozptyl σ = 12.0 jednotek.
2. Jak se změní situace, použijeme-li 90 % interval spolehlivosti?

$P(L1 < \mu < L2) = 1 - \alpha$ pro 95% interval spolehlivosti je α=0.05. Pro L1 a L2 platí $L1, L2 = \bar{x} \pm u_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$

$$4 = L2 - L1 = \bar{x} + u_{0,975}\frac{\sigma}{\sqrt{n}} - \left(\bar{x} - u_{0,975}\frac{\sigma}{\sqrt{n}}\right) = 2u_{0,975}\frac{\sigma}{\sqrt{n}} \quad \Rightarrow \quad n = \left(\frac{1}{2}1{,}96 \cdot 12\right)^2 = 139$$

Pokud chceme odhadnout průměrnou koncentraci na lokalitě tak, aby 95% interval spolehlivosti měl šířku 4 jednotky, potřebujeme k tomu 139 vzorků. **Finanční prostředky** vystačí pouze na 100 vzorků, tedy **jsou nedostatečné.**

Budeme-li uvažovat jen **90% interval spolehlivosti**, $u_{0,95}$=1,645. Počet vzorků získáme stejným výpočtem (n=98).
V tomto případě budou **finanční prostředky dostatečné**.

Příklady – rozložení, odhady

Příklad 5.

c/ Limit EPA pro vypouštění suspendovaných pevných odpadů do řek je maximálně 60 mg na litr denně, s maximálním měsíčním průměrem 30 mg na litr denně. Předpokládejte, že chcete testovat náhodně vybrané vzorky vody z jedné řeky s cílem odhadnout průměrnou denní dávku pevných kontaminantů, které pocházejí z těžebních závodů na břehu řeky.
Pokud chcete získat 95 % interval pro průměr s šířkou 2 mg, jak velký počet vzorků vody musíte zpracovat ?
Předchozí zkoušky prokázaly, že výsledky analýzy vodních vzorků jsou přibližně normálně rozloženy se směrodatnou odchylkou 5 mg.

obdobně jako v předchozím příkladu platí

$$2 = L2 - L1 = \bar{x} + u_{0,975}\frac{\sigma}{\sqrt{n}} - \left(\bar{x} - u_{0,975}\frac{\sigma}{\sqrt{n}}\right) = 2u_{0,975}\frac{\sigma}{\sqrt{n}} \quad \Rightarrow \quad n = (1{,}96 \cdot 5)^2 = 96$$

Tedy pro získání 95% intervalu spolehlivosti potřebujeme získat 96 vzorků.

d/ Podle Food and Drug Administration (FDA) obsahuje průměrný šálek kávy (7 g kávy) 115 mg kofeinu, a tato hodnota kolísá od 60 do 170 mg (rozsah výsledků provedených analýz). Máte za úkol tyto testy zopakovat tak, aby přesnost vašich závěrů byla v rozsahu 5 mg s 95% pravděpodobností.
Kolik šálků kávy musíte přibližně analyzovat k dosažení takových výsledků?

min=60
max=170
m=115mg

Z rozsahu minimálních a maximálních hodnot vypočítáme směrodatnou odchylku. Platí, že ±3s pokrývají 99,9% všech hodnot normálního rozložení. Tedy 170-60=6s → **s=18,3**

rozsah 95% intervalu spolehlivosti je 5mg, pro dosažení obdobných výsledků bude zapotřebí **206 šálků kávy**.

$$5 = L2 - L1 = \bar{x} + u_{0,975}\frac{\sigma}{\sqrt{n}} - \left(\bar{x} - u_{0,975}\frac{\sigma}{\sqrt{n}}\right) = 2u_{0,975}\frac{\sigma}{\sqrt{n}} \quad \Rightarrow \quad n = \left(\frac{2}{5}1{,}96 \cdot 18{,}3\right)^2 \doteq 206$$

Příklady – rozložení, odhady

Příklad 6.

Jsou naměřena následující čísla (opakovaná měření délky jednoho objektu v cm): 15; 13; 12; 11

a) Vypočítejte aritmetický průměr, směrodatnou odchylku a standardní chybu.
b) Vyjádřete správně přesnost odhadu průměru a vysvětlete použitý způsob vyjádření.
c) Jaký význam v tomto případě má interval spolehlivosti pro odhad průměru?
d) Změnil by se odhad ukazatelů variability při měření na 1 desetinné místo? (např. 15.3; 12.7; 12.2; 10.8)
e) Změnil by se odhad ukazatelů variability při zvětšení počtu měření?

a)

$$X = \frac{1}{n}\sum_{i=1}^{4} X_i = 12{,}75$$

$$S^2 = \frac{1}{n}\sum_{i=1}^{4} (X_i - X)^2 = 2{,}19 \quad \Rightarrow \quad S = 1{,}48$$

$$SE = \frac{S}{\sqrt{n}} = 0{,}74$$

d)

$$X = \frac{1}{n}\sum_{i=1}^{4} X_i = 12{,}75$$

$$S^2 = \frac{1}{n}\sum_{i=1}^{4} (X_i - X)^2 = 2{,}65 \quad \Rightarrow \quad S = 1{,}63$$

$$SE = \frac{S}{\sqrt{n}} = 0{,}82$$

Variabilita při měření na 1 desetinné místo vzroste.

c) interval spolehlivosti pro odhad průměru nám říká, pokud budeme znovu provádět vzorkování na souboru, ze kterého byl interval spolehlivosti spočítán, průměrná hodnota nového souboru se bude s 95% pravděpodobností vyskytovat v daném intervalu spolehlivosti

e) Při zvětšení počtu měření vy variabilita klesla.

Příklady – rozložení, odhady

Příklad 7.

Měření vzorku 25ti malých semenáčků ve školce (zadáno jako odhad pro celou výsadbu přibližně 600 jedinců) vedlo k následujícím výsledkům:
Průměr: 62.8 cm; SD: 11,8 cm
Vypočítejte 95% interval spolehlivosti pro odhad průměru.

$$L1 = \bar{x} - u_{0,975}\frac{\sigma}{\sqrt{n}} = 62{,}8 - 1{,}96\frac{11{,}8}{\sqrt{25}} = 58{,}17$$

$$L2 = \bar{x} + u_{0,975}\frac{\sigma}{\sqrt{n}} = 62{,}8 + 1{,}96\frac{11{,}8}{\sqrt{25}} = 67{,}43$$

95% interval spolehlivosti: (58,17; 67,43)

Příklad 8.

Bylo provedeno vzorkování na dvou polních lokalitách s cílem posoudit aktivitu extracelulární ureázy v půdě. Na každé lokalitě bylo odebráno 10 vzorků s následujícími výsledky:
a) Průměr 15,1 U/g (d.w.) / SD 3,1 U/g (d.w.)
b) Průměr 241 U/g (d.w.) / SD 25,8 U/g (d.w.)
Která z obou lokalit je v daném znaku variabilnější ?
Má smysl porovnávat intervaly spolehlivosti pro odhad průměru mezi lokalitami A a B?

Pro posouzení variability určíme koeficient variance: $c = s/\bar{x}$

$$c_a = s/\bar{x} = 3{,}1/15{,}1 = 0{,}205$$

$$c_b = s/\bar{x} = 25{,}8/241 = 0{,}107$$

Větší variabilitu má vzorek a. Spíše než intervaly spolehlivosti samotné by bylo lepší porovnávat šířku těchto intervalů.

Příklady – rozložení, odhady

Příklad 9.

Když je medián 15. číslo ve vzestupně seřazeném souboru, jak velký je celkový vzorek (n = ?) Spočítejte medián pro následující vzorky

a) Vzorek I: 5; 1; 8; 3; 4

b) Vzorek II:1; 3; 4; 5; 7; 8

Pokud je medián 15.číslo, celkový vzorek obsahuje 2n-1=29 čísel.

a) medián vzorku I. je 4, protože při lichém počtu prvků je medián (n+1)/2 prvek

b) medián vzorku II. je 4,5, protože při sudém počtu prvků je medián průměrem n/2 prvku a n/2+1 prvku.

Příklad 10.

Je naměřena následující sada hodnot (hmotnosti rostlin v g) a je třeba vyjádřit výsledek měření v sumarizované podobě jako odhad průměru a odpovídající interval spolehlivosti (95%). Posuďte možnost správného vyjádření a toto proveďte.

	X: originální data (g)	Ln (X)
	10,2	2,32
	15,3	2,73
	14,1	2,65
	11,2	2,42
	18,2	2,90
	11,2	2,42
	22,5	2,97
	23,5	3,02
	27,5	3,11
Průměr	**17,1**	**2,78**
Medián	**15,3**	**2,72**
SD	**6,2**	**0,36**
SE	**2,1**	**0,12**

Výpočet mediánu z frekvenčních dat

a) Určete medián tohoto souboru dat: 1,3,4,5,7,8 [4,5]

b) Určete medián tohoto souboru dat: 5,1,8,3,4 [4]

c) Tento příklad je ukázkou výpočtu mediánu u velkého souboru dat. V následující tabulce je uveden rozbor rozložení souboru dat od 179 krav, kde sledovanou veličinou byl počet dní od narození telete do znovuobnovení menstruačního cyklu. Uvedená data jsou velmi zjednodušena a jsou zde uvedena pouze pro ilustraci:

Class limits (days)	0,5-20,5	20,5-40,5	40,5-60,5	60,5-80,5	80,5-100,5	100,5-120,5	120,5-140,5	140,5-160,5	160,5-180,5	180,5-200,5	200,5-220,5
Frequency	8	33	50	32	15	20	11	6	2	1	1
Cumulative frequency	8	41	91	123	138	158	169	175	177	178	179

Frekvence zastoupení dosahuje nejvyšší hodnoty u třídy od 40,5 – 60,5 dnů. Druhý (menší) frekvenční pík lze pozorovat u intervalu od 100,5 do 120,5 dní. Existence dvou maxim (bimodální data) je důkazem nenormality tohoto konkrétního souboru.

Výpočet mediánu z frekvenčních dat

Jelikož n =179, pak je medián devadesátá hodnota od počátku souboru, a dále je zřejmé, že bude velmi blízko horní hranici třídy 40,5 – 60,5 dní. Za předpokladu, že 50 hodnot této třídy je v ní rovnoměrně rozmístěno lze použít následující vzorec:

$$M = X_L + \frac{gl}{f}, \; kde$$

X_L = hodnota X (sledované veličiny) na spodní hranici třídy obsahující medián: zde 40,5 dní

g = pořadová hodnota mediánu minus kumulativní frekvence do horní hranice předchozí třídy, tj. 90 - 41= 49

l = třídní interval: 20 dní

f = frekvence ve třídě obsahující medián

- Dosadíme-li do uvedeného vzorce, získáme odhad mediánu jako 60 dní. Průměr tohoto datového souboru je 69,9, což je významně odlišná hodnota, a potvrzuje znovu nenormální charakter dat.
- U velkých vzorků z normálních populací je výběrový odhad mediánu normálně rozložen kolem populační hodnoty se směrodatnou odchylkou $1{,}253\ \sigma / \sqrt{n}$. U normálního rozložení, kde medián i průměr představují odhad stejné hodnoty, je medián méně přesný než průměr. Proto hlavní význam mediánu spočívá u nesymetrických distribucí.
- Existuje velmi jednoduchá metoda pro výpočet intervalu spolehlivosti pro odhad mediánu a jako horní a spodní hranice slouží pořadová čísla vypočítaná podle následujícího vztahu:

$$\frac{(n+1)}{2} \pm \frac{z\sqrt{n}}{2}, \; kde$$

n představuje velikost datového souboru, **z** je kvantil standardizovaného normálního rozložení pro příslušnou pravděpodobnost. U našeho příkladu je n = 179 a pro 95% interval spolehlivosti je z přibližně rovno 2. Horní a spodní limit pro odhad mediánu tedy je $90 \pm \sqrt{179} = 77$ a 103. 95% interval spolehlivosti je tedy tvořen počty dní, které mají pořadí 77 a 103:

77: Počet dní = 40,5+(36)(20)/50 = 55 dní

103: Počet dní = 60,5+(12)(20)/32 = 68 dní

Medián cílové populace byl tedy odhadnut 95% intervalem spolehlivosti jako hodnota ležící mezi 55 a 68 dny. Interpretujte tento výsledek.

Průměr a medián u frekvenčně tříděných dat: příklad

II. Symetrická rozložení

✓ **X: třídně uspořádaná koncentrace látky zjišťovaná v n = 27 jedincích**

Třída	f_i
1,85 - 1,95	2
1,95 - 2,05	1
2,05 - 2,15	2
2,15 - 2,25	3
2,25 - 2,35	5
2,35 - 2,45	6
2,45 - 2,55	4
2,55 - 2,65	3
2,65 - 2,75	1

Medián (M) ~ 14. číslo

$$M = X_L + \frac{g \,.\, l}{f} = 2{,}35 + \frac{1 \,.\, 0{,}1}{6} = 2{,}367$$

Průměr = 2,33
Modus = 2,4

X_L … hodnota x na spodní hranici třídy obsahující medián
g … požadovaná hodnota mediánu - kumulativní frekvence do horní hranice předchozí třídy
l … třídní interval
f … frekvence ve třídě obsahující medián

Příklady – rozložení a testy pro dva výběry

Příklady

Příklad 1:

Hodnotili by jste následující sumární statistiky jako smysluplné (tedy jako interpretovatelné a správně spočítané ?) Je-li to možné, pojmenujte typ rozložení pro každou takto specifikovanou proměnnou.

ZNAK X1
= počet dnů v roce s deštivým počasím
- hodnoceno pro 20 relativně hodně vzdálených lokalit (n = 20)

Průměr: 189,6
Medián: 142
SD: 85,3

log-normální rozložení

ZNAK X2

= hmotnost myší pod vlivem určitého typu diety
- hodnoceno pro 20 jedinců (n = 20)

Průměr: 100
MIN / MAX: 20 / 180
SE: 15,9

normální rozložení

ZNAK X3
= nosnost slepic za určité období
- hodnoceno pro 20 jedinců (n = 20)

Geometrický průměr: 42,3
Medián: 38
MIN / MAX : 15 / 114

log-normální rozložení

Příklady

Příklad 2:

Čtete vědeckou literaturu a v ní naleznete následující údaj o výšce rostliny:

n = 20
Geometrický průměr: 42,3
MIN / MAX : 10 / 114

Dovedete přibližně určit v jakých hranicích se pohybuje spolehlivý odhad průměru (uvažujte pro výpočet 95 % spolehlivost) ?

Příklad 3:

Chemický experiment (n = 5)
Výsledky jednotlivých opakování:
X1 = 5,3; X2 = 5,6; X3 = 5,9; X4 = 8,2; X5 = 5,0

Do jaké míry by mohlo být oprávněné vyloučit hodnotu X4 = 8,2 ?

Příklady

Příklad 4:

Toxikologická laboratoř musela přejít na nový způsob chovu morčat, které používala na průzkum vlivu organických kontaminantů na tělesnou hmotnost organismu v době intenzivního růstu. Dvacet těchto nových morčat ze specializované laboratoře je živeno touto speciální kontaminovanou dietou. jejich průměrný přírůstek na hmotnosti je během dvou měsíců 28g. V předchozích experimentech s bývalou populací o relativně velkém rozsahu (n > 500) byl průměrný přírůstek morčat za těchto podmínek 29,8g a rozptyl $s^2 = 25$. Testujte hypotézu, zda je nová populace srovnatelná s předchozí.

Test se bude provádět využitím jednovýběrového t-testu s nulovou hypotézou: $\bar{x} = \mu$

$$t = \frac{\bar{x} - \mu}{s}\sqrt{n} = \frac{28 - 29{,}8}{\sqrt{25}}\sqrt{20} = -1{,}609$$

$$t_{0,975}^{(19)} = 2{,}093$$

protože $|t| < t_{0,975}^{(19)}$ nulovou hypotézu nezamítáme.

Nová populace je srovnatelná s předchozí.

Příklady

Příklad 5:

Máte za úkol testovat, zda nově vyvinuté antibiotikum proniká do mléka, je-li podáváno kravám po dobu dvou týdnů. Stanovte cíl experimentu, typ sledované veličiny a uspořádání experimentu. Diskutujte pravděpodobnosti a význam možných chyb. Dále diskutujte předpokládané rozložení sledované veličiny a navrhněte způsob testování. Za normálních podmínek se antibiotikum v mléce vůbec nevyskytuje.

Formulujte hypotézu a systém testování pro následující situace:

a) již stopový průnik antibiotika mléko znehodnotí

b) antibiotikum znehodnotí mléko až od koncentrace C_k

Cílem experimentu bude ověřit hypotézu, že antibiotikum do mléka neproniká. Experiment můžeme uspořádat jako párový test, tedy vyšetřit skupinu krav před podáváním antibiotika a po podávání antibiotika. Tím zajistíme, že výskyt antibiotika v mléce po jeho podávání nebude ovlivněn jeho přítomností před podáváním.

hypotéza a) množství antibiotika v mléce po jeho podávání je nulové

hypotéza b) množství antibiotika v mléce po jeho podávání není větší než koncentrace c_k

Příklad 6:

Na Iontově selektivní elektrodě je napsáno, že průměrný obsah dusičnanů ve vzorku naměří se směrodatnou odchylkou 1,5 mg/ml. Jak velký počet opakovaných měření musíte udělat, je-li stanovení průměrné koncentrace požadováno s přesností danou standardní chybou 0,2 mg/ml ?

$$\sigma = 1.5 mg/l \qquad se = 0.2 mg/l$$

$$se = \frac{\sigma}{\sqrt{n}} \quad \Rightarrow \quad n = \left(\frac{\sigma}{se}\right)^2 \doteq 57$$

musíme udělat 57 měření, pokud požaduje odhad průměrné hodnoty se směrodatnou odchylkou 0.2mgl.

Příklady

Příklad 7:

Nepříliš čistá protilátka není jako směs proteinů přesně definovaná a její složení je kolísavé, což zvyšuje variabilitu opakovaných stanovení při reakci s antigenem. Jelikož není k dispozici lepší zdroj, je nutné před každým pokusem danou šarži testovat na standard – tedy na antigenní vzorek o přesně známé koncentraci. Je-li rozptyl stanovení pod hranicí σ_0^2, lze látku použít k nejdůležitějším stanovením a naopak, překročí-li hodnotu σ_m^2, nelze daný preparát použít vůbec.

Navrhněte standardní způsob testování takových experimentů i pro následující konkrétní situaci:

Pravdivá koncentrace testovaného antigenu je 115,2 µg/ml ve standardním vzorku.

n = 10
σ_0^2 = 4,5
σ_m^2 = 14,5
s^2 = 8,87

Zjištěný odhad průměru standardního vzorku: 110,8 µg/ml

Nejprve je vhodné vyloučit možnost, že rozptyl naměřené koncentrace překročí hodnotu **σ_m^2.** Jinak by musel být preparát úplně vyřazen. Toto provedeme F-testem s jednostrannou hypotézou **$s^2 < \sigma_m^2$** Pokud hypotézu nevyvrátíme, můžeme testovat hypotézu, že rozptyl naměřené koncentrace je nižší než **σ_0^2** Pokud hypotézu nezamítneme, je preparát možné použít úplně, pokud hypotézu zamítneme, můžeme preparát použít jen omezeně.

Test I

1. Pro vektor čísel: 1,2,5,8,7,4,3,6,11,12,9:

- neexistuje distribuční funkce, protože není uspořádán vzestupně
- neexistuje distribuční funkce, protože tato má smysl pouze pro výběry větší než n =100 vzorků
- existuje distribuční funkce a lze ji bez jakýchkoliv problémů sestrojit
- existuje distribuční funkce, ale pokud jde o výběr z populace, pak má pouze orientační význam
- jsou definovány kvantily, a tudíž i medián
- neexistuje rozložení
- nelze smysluplně předpokládat normální rozložení

2. Máte za úkol uspořádat pokus, který má porovnat pH vody mezi vodními nádržemi ošetřenými fosforem a neošetřenými fosforem.

a) Navrhněte uspořádání pokusu tak, aby bylo párové.

b) Navrhněte uspořádání pokusu tak, aby bylo nezávislé.

c) Můžete studovat n = 100 nádrží: navrhněte možnosti hodnocení takového pokusu.

d) Zdůvodněte výhody a nevýhody nezávislého a párového uspořádání.
Jak se interpretačně liší ? Který design více vypovídá o realitě v přírodě ?

Test II

3. Máte za úkol uspořádat experiment porovnávající koncentraci protilátek proti tetanu po očkování novou vakcínou, která je ve stadiu klinického testování.

a) Naplánujte pokus pro párové uspořádání.

b) Naplánujte pokus pro nezávislé uspořádání.

c) Představte si, že vyberete z populace n = 100 sourozenců ochotných k testování vakcíny. Mezi získanými vektory dat (koncentrace protilátek) je zjištěn koeficient korelace r = 0,123, což je nevýznamná hodnota. Jaké má toto zjištění důsledky pro uspořádání pokusu ?

4. Máte dva vektory dat (n = 150 párů) z pokusu, kde byl sledován rozklad toxických látek v odpadní vodě bakteriemi. Jako kontrola slouží voda neošetřená bakteriemi.

a) Autor pokusu tvrdí, že má pokus uspořádán párově.
Jakým způsobem ho tedy musel provést ?
Jak by jste ověřili významnost tohoto párování ?

b) Předpokládejme párové uspořádání takového pokusu, tj., výstupem jsou dva vektory čísel (úbytek toxických látek v průběhu pokusu v ng), každý vektor o délce n =150.
Nechali jste vypočítat koeficienty korelace mezi oběma vektory dat a byly vám předloženy následující výsledky:

Pearsonova korelace: r = 0,956
Spearmanova korelace r_s= 0,196

Co z těchto výsledků vyplývá pro hodnocení pokusu a jaké další hodnocení zvolíte ?

Srovnání parametrů dvou výběrů
Experimenty pro dva pokusné zásahy

Příklad 1:

a) Máte k dispozici počty jedinců kůrovce, kteří byly polapeni do dvou pastí umístěných v zamořené oblasti. Vaším úkolem je srovnat rozptyl obou proměnných: $H_0 : \sigma_1^2 = \sigma_2^2$

Počty jedinců											
Past 1	41	34	33	36	40	25	31	37	34	30	38
Past 2	52	57	62	55	64	57	65	55	-	-	-

K otestování shody rozptylu použijeme tzv. F– test pro poměr rozptylů (Variance ratio test):

$$n_1 = 11;\ v_1 = 10$$

$$n_2 = 8;\ v_2 = 7$$

$$s_1^2 = 21{,}87;\ \ s_2^2 = 15{,}36$$

$$F = \frac{\text{Max}\left(s_1^2 \cdot s_2^2\right)}{\text{Min}\left(s_1^2 \cdot s_2^2\right)} = \frac{21{,}87}{15{,}36} = 1{,}42$$

$$F(0{,}05)[10\ \ ;7] = 4{,}76$$

$$p > 0{,}5$$

Nezamítáme nulovou hypotézu shody rozptylů.

Je tedy možné vypočítat společný rozptyl jako vážený průměr rozptylů obou proměnných:

$$s_p^2 = 19{,}19$$

Tabulka

Pozn: Vzhledem k tomu, že naše H_0 byla oboustranná, je třeba k testování použít tabulky (F-rozdělení):

f_2 = d. f. for Smaller Mean Square	f_1 = d.f. for Larger Mean Square									
	2	4	6	8	10	12	15	20	30	nekon.
2	39,00	39,25	39,33	39,37	39,40	39,42	39,43	39,45	39,46	39,50
3	16,04	15,10	14,74	14,54	14,42	14,34	14,25	14,17	14,08	13,90
4	10,65	9,60	9,20	8,98	8,84	8,75	8,66	8,56	8,46	8,26
5	8,43	7,39	6,98	6,76	6,62	6,52	6,43	6,33	6,23	6,02
6	7,26	6,23	5,82	5,60	5,46	5,37	5,27	5,17	5,07	4,85
7	6,54	5,52	5,12	4,90	4,76	4,67	4,57	4,47	4,36	4,14
8	6,06	5,05	4,65	4,43	4,30	4,20	4,10	4,00	3,89	3,67
9	5,71	4,72	4,32	4,10	3,96	3,87	3,77	3,67	3,56	3,33
10	5,46	4,47	4,07	3,85	3,72	3,62	3,52	3,42	3,31	3,08
12	5,10	4,12	3,73	3,51	3,37	3,28	3,18	3,07	2,96	2,72
15	4,76	3,80	3,41	3,20	3,06	2,96	2,86	2,76	2,64	2,40
20	4,46	3,51	3,13	2,91	2,77	2,68	2,57	2,46	2,35	2,09
30	4,18	3,25	2,87	2,65	2,51	2,41	2,31	2,20	2,07	1,79
nekon.	3,69	2,79	2,41	2,19	2,05	1,94	1,83	1,71	1,57	1,00

Pro vypočítaný poměr obou rozptylů (1,42) lze vypočítat interval spolehlivosti. Interpretujte výsledek tohoto výpočtu vyjádřený jako:

$$P\left(0{,}298 \le \frac{\sigma_1^2}{\sigma_2^2} \le 5{,}61\right) = 0{,}95$$

Srovnání parametrů dvou výběrů
Experimenty pro dva pokusné zásahy

Příklad 2:

Pomocí F–testu uvedeného v úloze 1, lze rovněž testovat rovnost dvou koeficientů variance:

$$F = \frac{(s^2_{\log})_1}{(s^2_{\log})_2};$$

Je třeba ověřit, zda má koncentrace Zn nalezená v kontaminovaných půdách stejný rozptyl jako obsah mikrobiální biomasy naměřený na stejných lokalitách (srovnání často nutné pro správnou volbu metody současné analýzy obou proměnných). Nulovou hypotézu budeme testovat srovnáním koeficientů variance podle výše uvedeného vztahu:

Obsah Zn (mg/kg)	Log (Zn)	Obsah biomasy (mg C/kg)	Log (biomasa)
72,5	1,86034	183,0	2,26245
71,7	1,85552	172,3	2,23629
60,8	1,78390	180,1	2,25551
63,2	1,80072	190,2	2,27921
71,4	1,85370	191,4	2,28194
73,1	1,86392	169,9	2,22943
77,9	1,89154	166,4	2,22115
75,7	1,89910	177,6	2,24944
72,0	1,85733	-	-
69,0	1,84	-	-

Srovnání parametrů dvou výběrů
Experimenty pro dva pokusné zásahy

$\nu_1 = 9$

$\overline{X}_1 = 70{,}73\text{kg}$

$SS_1 = 246{,}1610\text{kg}$

$s_1^2 = 27{,}3512\text{kg}^2$

$s_1 = 5{,}23\text{kg}$

$V_1 = 0{,}0739$

$(SS_{\log})_1 = 0{,}00987026\,32$

$(s_{\log}^2)_1 = 0{,}00109669\,59$

$\nu_2 = 7$

$\overline{X}_2 = 178{,}82\text{cm}$

$SS_2 = 590{,}1350\text{cm}^2$

$s_2^2 = 84{,}3050\text{cm}^2$

$s_2 = 9{,}18\text{cm}$

$V_2 = 0{,}0513$

$(SS_{\log})_2 = 0{,}0034727534$

$(s_{\log}^2)_2 = 0{,}004961076$

$$F = \frac{0{,}00109669\,59}{0{,}0004961076} = 2{,}21$$

$F_{0{,}05(2)} = 4{,}82$

$0{,}20 < p < 0{,}50$

Nezamítáme H_0.

Příklad – two sample test (párový x nepárový)

Pokus na zvířatech – srovnání dvou variant (n = 7 jedinců)

kontrola před ošetřením: $\overline{X}_1 = 8{,}74;\ s_1^2 = 4{,}026;\ s_{x_1}^2\ 0{,}575$

kontrola po ošetření: $\overline{X}_2 = 7{,}73;\ s_2^2 = 2{,}904;\ s_{x_2}^2\ 0{,}415$

$r = 0{,}981$

$Cov = 3{,}352$

$\overline{D} = 1{,}01;\ s_{D^2} = 0{,}225$

$s_{\overline{D}} = 0{,}179$

$t = \overline{D}/s_{\overline{D}} = 5{,}639$

p < 0.01

$$s_p^2 = \frac{6 \cdot s_1^2 + 6 \cdot s_2^2}{12} = 3{,}464$$

$$s_{\overline{X}_1 - \overline{X}_2} = \sqrt{2} \cdot s_p / \sqrt{n} = 0{,}995$$

$$t = \frac{1{,}01}{0{,}995} = 1{,}016$$

p < 0.328

Příklad

Průměrný denní příjem výživy odhadovaný 10 dní před lékařským zásahem a 10 dní po lékařském zásahu.

Pacient	před lékařským zásahem	po lékařském zásahu	diference
1	5260	3910	1350
2	5470	4220	1250
3	5640	3885	1755
4	6180	5160	1020
5	6390	5645	745
6	6515	4680	1835
7	6805	5265	1540
8	7515	5975	1540
9	7515	6790	725
10	8230	6900	1330
11	8770	7335	1435
Průměr	**6753,6**	**5433,2**	**1320,5**
Medián	**6515**	**5265**	**1350**
SD	**1142,1**	**1216,8**	**366,7**

Odhadněte 95% interval spolehlivosti pro rozdíl mezi průměry.
Ověřte zda je rozdíl statisticky významný (testujte nulovou hypotézu).
Pearsonova korelace: r = 0,9536

Je možné použít neparametrickou alternativu pro tyto testy ?

Test provedeme využitím párového t-testu nebo jednovýběrovým t-testem s nulovou hypotézou, že průměrná hodnota diferencí se neliší od nuly

Příklady I

Příklad 1.

Při sledování určitého fyziologického parametru souvisejícího s činností srdce, nesmí rozptyl hodnot přesáhnout stanovený limit, aby možné odchylky od normálu nezanikly v šumu. Tato limitní hodnota je **σ_0 = 4.5**. Po zakoupení nového přístroje testovala klinika měření na **n = 30** pacientech s výsledkem **s = 4.0**. Je možné dále pokračovat ve vyšetřování na novém přístroji, nebo je nutné tento test provádět na přesnějším stroji?
Proveďte komplexní rozbor situace, včetně závěrů o dalším postupu měření. (Jde v podstatě o test shody výběrového odhadu rozptylu a rozptylu cílové populace.)

Testujeme nulovou hypotézu H_0: $s^2 \le \sigma^2$ na 5% hladině významnosti, jako testovou statistiku použijeme $\chi^2 = \frac{(n-1).s^2}{\sigma^2}$

$\chi^2 = \frac{(n-1).s^2}{\sigma^2} = \frac{29 \cdot 4^2}{4{,}5^2} = 22{,}91$ Porovnáme-li hodnotu testové statistiky s kvantilem $\chi^2_{1-\alpha}{}^{(n-1)} = \chi^2_{0{,}95}{}^{(29)} = 42{,}557$

platí, že $\chi^2 < \chi^2_{0{,}95}{}^{(29)}$ tedy nulovou hypotézu nezamítneme. Je možné dále pokračovat ve vyšetřování na novém přístroji.

Příklad 2.

Aby bylo podávané antibiotikum účinné proti bakteriím v ledvinách, musí jeho koncentrace v krvi dosáhnout alespoň hodnoty **18 jednotek/ ml**. Z dřívějších rozsáhlých výzkumů víme, že stanovení obsahu antibiotika v krvi vykazuje směrodatnou odchylku **σ = 3.3**. Při testování nové varianty antibiotika na myších byla u **n = 9** myší nalezena **průměrná koncentrace látky v krvi 10.2 jednotky.**
Při α = 0.05 testujte, zda je tato hodnota dostatečná pro účinnost antibiotika v ledvinách.

Testujeme nulovou hypotézu H_0: $\bar{x} \ge \mu$ na 5% hladině významnosti, jako testovou statistiku použijeme $t = \frac{\bar{x}-\mu}{s}\sqrt{n}$

$t = \frac{\bar{x}-\mu}{s}\sqrt{n} = \frac{10{,}2-18}{3{,}3}\sqrt{9} = -7{,}09$ Porovnáme-li hodnotu testové statistiky s kvantilem $t_\alpha{}^{(n-1)} = t_{0{,}05}{}^{(8)} = -1{,}86$

platí, že $t < t_{0{,}05}{}^{(8)}$ tedy nulovou hypotézu zamítneme. Tato hodnota není dostatečná pro účinnost antibiotika v ledvinách.

Příklady II

Příklad 3.

a) Deset myší bylo testováno na přítomnost jater poškozující toxické látky, která se může vyskytnout v jednom druhu masových konzerv. K testování bylo odebráno 100 konzerv a po deseti dnech uhynuly 2 myši, tzn. ve dvou konzervách byla látka prokázána.
Jaký je interval spolehlivosti výskytu této látky v celém souboru konzerv?

n=100
r=2

$$\hat{p} = 0{,}02 \qquad s_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}} = \sqrt{\frac{0{,}02 \cdot 0{,}98}{99}} = 0{,}014$$

interval spolehlivosti (1-α)100% π : $\hat{p} \pm Z_{1-\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n-1}} = 0{,}02 \pm Z_{1-\alpha/2} \cdot 0{,}014$

b) Bylo zkoumáno 115 žen starších 36 let, zda měly potíže s chrupem během těhotenství. Kladně odpovědělo 46 žen.
Jaké jsou vaše závěry o celé populaci žen tohoto věku při 99% spolehlivosti? (Vypočítejte interval spolehlivosti pro p)

n=115
r=46

$$\hat{p} = 0{,}4 \qquad s_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}} = \sqrt{\frac{0{,}4 \cdot 0{,}6}{114}} = 0{,}046$$

interval spolehlivosti 99% π : $\hat{p} \pm Z_{1-\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n-1}} = 0{,}4 \pm Z_{0{,}995} \cdot 0{,}046 = 0{,}4 \pm 2{,}58 \cdot 0{,}046 = 0{,}4 \pm 0{,}12$

28%-52% žen tohoto věku má potíže s chrupem během těhotenství při 99% spolehlivosti.

Příklady III

Příklad 3.

c) Pravděpodobnost narození chlapce je asi 1/2. Máte zhodnotit výsledky průzkumu populace, která žije v silně poškozeném životním prostředí. Průzkum se týká 1000 náhodně vybraných rodin a zjištěný podíl narozených chlapců je 0.41.
Jaké jsou vaše závěry o této populaci?
Jak se váš odhad zpřesní, když použijete vzorek n = 10 000 rodin při zachování odhadu p = 0.41?

Použijeme jednovýběrový binomický test s nulovou hypotézou H_0: p=π, hladina významnosti α=0,05

testová statistika $Z = \frac{n \cdot \hat{p} - n \cdot \pi}{\sqrt{n \cdot \hat{p}(1-\hat{p})}} = \frac{1000 \cdot 0{,}41 - 1000 \cdot 0{,}5}{\sqrt{1000 \cdot 0{,}41 \cdot 0{,}59}} = -5{,}79$ a příslušný kvantil $Z_{1-\frac{\alpha}{2}} = Z_{0{,}975} = 1{,}96$

protože $|Z| > Z_{0{,}975}$ nulovou hypotézu zamítáme. Chlapci se ve zkoumavé populaci nerodí s pravděpodobností 0,5.

interval spolehlivosti π: $\hat{p} \pm Z_{1-\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n-1}} = 0{,}4 \pm Z_{0{,}975} \cdot 0{,}046 = 0{,}41 \pm 1{,}96 \cdot 0{,}016 = 0{,}41 \pm 0{,}03$

pokud použijeme n=10 000, bude int. spolehlivosti užší π: $\hat{p} \pm Z_{1-\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n-1}} = 0{,}41 \pm 1{,}96 \cdot 0{,}005 = 0{,}41 \pm 0{,}01$

d) Jaká je pravděpodobnost, že rodina se třemi dětmi bude mít 2 (3) chlapce?
Podrobně analyzujte problém a použijte obecného definičního vztahu pro binomické rozložení.

n = 3
r = 2
p=0,5 (stejná pravděpodobnost narození chlapce jako narození dívky)

$$P(r) = \binom{n}{r} \cdot p^r \cdot (1-p)^{(n-r)} = \frac{n!}{r!(n-r)!} \cdot p^r \cdot q^{(n-r)}$$

$$P(2) = \binom{3}{2} \cdot 0{,}5^2 \cdot 0{,}5^{(1)} = \frac{3!}{2!(1)!} \cdot 0{,}5^2 \cdot 0{,}5^{(1)} = 0{,}375$$

pravděpodobnost narození 2 chlapců v rodině se třemi dětmi je 0,375

r = 3 platí $P(3) = \binom{3}{3} \cdot 0{,}5^3 \cdot 0{,}5^0 = 1 \cdot 0{,}5^3 \cdot 0{,}5^0 = 0{,}125$

pravděpodobnost narození 3 chlapců v rodině se třemi dětmi je 0,125

Příklady IV

Příklad 3.

e) Předpokládá se, že lidé trpící určitou krevní chorobou mají abnormální jeden z chromozómů. S cílem odhadnout podíl takto postižených chromozómů bylo studováno 5 buněk od každého ze 120 pacientů a byl zjišťován počet buněk s postiženým chromozómem (tento počet = sledovaný jev = r). Výsledky jsou uvedeny v následující tabulce. Odhadněte podíl postižených chromozómů u populace nemocných lidí.

r(četnost jevu)	0	1	2	3	4	5	celkem
f(poč. pacientů)	6	31	42	29	10	2	120

Pro odhad p se používá vztah $\hat{p} = \dfrac{\sum_{i=1}^{k} f_i X_i \Big/ \sum_{i=1}^{k} f_i}{n}$

X_i	fi	$X_i f_i$
0	6	0
1	31	31
2	42	84
3	29	87
4	10	40
5	2	10

$$\sum_{i=1}^{k} f_i X_i = 252$$

$$\sum_{i=1}^{k} f_i = 120$$

$$n = 5$$

$$\hat{p} = \frac{252/120}{5} = 0{,}42$$

pravděpodobnost výskytu postiženého chromozómu

Příklady – různé

Příklad 1.

Two-tailed t test for significant diference between a mean and a hypothesized population mean of μ_0=22 year.

Věk smrti (v letech) 25 koní určitého druhu:

17,2, 18,0, 18,7, 19,8, 20,3, 20,9, 21,0, 21,7, 22,3, 22,6, 23,1, 23,4, 23,8, 24,2, 24,6, 25,8, 26,0, 26,3, 27,2, 27,6, 28,1, 28,6, 29,3, 30,1, 35,1.

$$H_0 : \mu = 22$$

$$H_A : \mu \neq 22$$

$$\alpha = 0{,}05$$

$$n = 25$$

$$\overline{X} = 24{,}23$$

$$s^2 = 18{,}0388$$

$$s_{\bar{x}} = \sqrt{\frac{18{,}0388^{\,2}}{25}} = 0{,}85$$

$$t = \frac{\overline{X} - \mu}{s_{\bar{x}}} = \frac{24{,}23 - 22}{0{,}85} = 2{,}624$$

$$v = n - 1 = 25 - 1 = 24$$

$$t_{0{,}05(2),24} = 2{,}064$$

Protože $|t| > t_{0,05(2),24}$, zamítáme H_0 a usuzujeme, že soubor 25 životních délek koňů pochází z populace jejíž průměr, μ, není 22 let.

$$0{,}01 < P(|t| \geq 2{,}624) < 0{,}02$$

Příklad 2.

A two-tailed t test for significant diference between a sample mean and a hypothesized population mean of zero.

Hmotnostní změny 12 potkanů po pobytu v režimu nuceného cvičení. Každá změna hmotnosti (v gramech) je definována jako hmotnost po cvičení mínus hmotnost před cvičením.

1,7
0,7
-0,4
-1,8
0,2
0,9
-1,2
-0,9
-1,8
-1,4
-1,8
-2,0

$$H_0 : \mu = 0$$

$$H_A : \mu \neq 0$$

$$\alpha = 0{,}05$$

$$n = 12$$

$$\overline{X} = -0{,}65g$$

$$s^2 = 1{,}5682g^2$$

$$t = \frac{\overline{X} - \mu}{s_{\bar{x}}} = \frac{-0{,}65g}{0{,}36g} = -1{,}81$$

$$s_{\bar{x}} = \sqrt{\frac{1{,}5682g^2}{12}} = 0{,}36g$$

$$v = n - 1 = 11$$

$$t_{0,05\,(2),11} = 2{,}201$$

Protože $|t| < t_{0.05\,(2),11}$ **, nezamítáme H_0.**

$0{,}05 < P < 0.10$

Příklad 3.

A one-tailed t test for the hypotheses H_0: $\mu \leq 45\,\text{sec}$ and $H_A: \mu > 45\,\text{sec}$

Doby rozpustnosti (v sekundách) drogy v žaludeční šťávě::

42,7, 43,4, 44,6, 45,1, 45,6, 45,9, 46,8, 47,6.

$H_0: \mu \leq 45 sek$

$H_A = \mu > 45 sek$

$\alpha = 0{,}05$

$n = 8$

$\overline{X} = 45{,}21 sek$

$SS = 18{,}8288 sek^2$

$s^2 = 2{,}6898 sek^2$

$s_{\bar{x}} = 0{,}58 sek$

$$t = \frac{45{,}21 sek - 45 sek}{0{,}58 sek} = 0{,}36$$

$v = 7$

$t_{0{,}05(1),7} = 1{,}895$

Když $t \geq t_{0,05\,(1),\,7}$, zamítáme H_0.

Závěr: nezamítáme H_0. $P\,(t \geq 0{,}36\,) > 0{,}50$

Příklad 4.

A two-tailed variance ratio test for the hypothese H0: $\sigma_1^2 = \sigma_2^2$ and $H_A : \sigma_1^2 \neq \sigma_2^2$. The data are the numbers of moths caught dutiny the night by 11 traps of one style and 8 traps of a second style.

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_A : \sigma_1^2 \neq \sigma_2^2$$

$$\alpha = 0{,}05$$

Trap type 1	Trap type 2
41	52
34	57
33	62
36	55
40	64
25	57
31	56
37	55
34	
30	
38	

$$n_1 = 11 \qquad n_2 = 8$$

$$v_1 = 10 \qquad v_2 = 7$$

$$SS_1 = 218{,}73 moths^2 \qquad SS_2 = 107{,}50 moths^2$$

$$s_1^2 = 21{,}87 moths^2 \qquad s_2^2 = 15{,}36 moths^2$$

$$F = \frac{s_1^2}{s_2^2} = \frac{21{,}87}{15{,}36} = 1{,}42$$

$$F_{0{,}05(2),10,7} = 4{,}76$$

$$P(F \geq 1{,}42) > 0{,}50$$

$$s_p^2 = \frac{218{,}73 moths^2 + 107{,}50 moths^2}{10 + 7} = 19{,}19 moths^2$$

The hypotheses may be submitted to the variance ratio test, for which one calculates

$$F = \frac{s_1^2}{s_2^2}$$ **or** $$F = \frac{s_2^2}{s_1^2}$$ **, whichever is larger.**

Srovnání parametrů dvou výběrů
Experimenty pro dva pokusné zásahy

Příklad 1.

a) Máte k dispozici počty jedinců kůrovce, které byli polapeny do dvou pastí umístněných v zamořené oblasti. Vaším úkolem je srovnat rozptyl obou proměnných

$$H_0 : \sigma_1^2 = \sigma_2^2$$

	Počty jedinců										
Past 1	41	34	33	36	40	25	31	37	34	30	38
Past 2	52	57	62	55	64	57	56	55	-	-	-

K otestování shody rozptylů použijeme tzv. F-test pro poměr rozptylů (Variance ratio test)

$$n_1 = 11, v_1 = 10$$
$$n_2 = 8, v_2 = 7$$
$$s_1^2 = 21{,}87; s_2^2 = 15{,}36$$

$$F = \frac{Max\ (s_1^2 . s_2^2)}{Min\ (s_1^2 . s_2^2)} = \frac{21{,}87}{15{,}36} = 1{,}42$$

$$F(0{,}05)[10;7] = 4{,}76$$

$$P > 0{,}5$$

Nezamítáme nulovou hypotézu shody rozptylu.

Je tedy možné vypočítat společný rozptyl jako vážený průměr rozptylů obou proměnných:

$$s_p^2 = 19{,}19$$

Srovnání parametrů dvou výběrů
Experimenty pro dva pokusné zásahy

***Pozn.*: Vzhledem k tomu, že naše H_0 byla oboustranná, je třeba k testování použít tabulky:**

5 LEVEL (TWO-TAILED) OF THE DISTRIBUTION OF F

f2 = d.f. for Smaller Mean Square	f_1 = d.f. for Larger Mean Square									
	2	4	6	8	10	12	15	20	30	∞
2	39,00	39,25	39,33	39,37	39,40	39,42	39,43	39,45	39,46	39,50
3	16,04	15,10	14,74	14,54	14,42	14,34	14,25	14,17	14,08	13,90
4	10,65	9,60	9,20	8,98	8,84	8,75	8,66	8,56	8,46	8,26
5	8,43	7,39	6,98	6,76	6,62	6,52	6,43	6,33	6,23	6,02
6	7,26	6,23	5,82	5,60	5,46	5,37	5,27	5,17	5,07	4,85
7	6,54	5,52	5,12	4,90	4,76	4,67	4,57	4,47	4,36	4,14
8	6,06	5,05	4,65	4,43	4,30	4,20	4,10	4,00	3,89	3,67
9	5,71	4,72	4,32	4,10	3,96	3,87	3,77	3,67	3,56	3,33
10	5,46	4,47	4,07	3,85	3,72	3,62	3,52	3,42	3,31	3,08
12	5,10	4,12	3,73	3,51	3,37	3,28	3,18	3,07	2,96	2,72
15	4,76	3,80	3,41	3,20	3,06	2,96	2,86	2,76	2,64	2,40
20	4,46	3,51	3,13	2,91	2,77	2,68	2,57	2,46	2,35	2,09
30	4,18	3,25	2,87	2,65	2,51	2,41	2,31	2,20	2,07	1,79
∞	3,96	2,79	2,41	2,19	2,05	1,94	1,83	1,71	1,57	1,00

16

Pro vypočítaný poměr obou rozptylů (1,42) lze vypočítat interval spolehlivosti. Interpretujte výsledek tohoto výpočtu vyjádřený jako:

$$P(0{,}298 \leq \frac{\sigma_1^2}{\sigma_2^2} \leq 5{,}61) = 0{,}95$$

Srovnání parametrů dvou výběrů
Experimenty pro dva pokusné zásahy

Příklad 2.

Pomocí F-testu uvedeného v Příkladu 1 této kapitoly lze rovněž testovat rovnost dvou koeficientů variance:

$$F = \frac{\left(s^2_{\log}\right)_1}{\left(s^2_{\log}\right)_2} =; \; jmenovatel \; < čitatel$$

Je třeba ověřit, zda má koncentrace Zn nalezená v kontaminovaných půdách stejný rozptyl jako obsah mikrobiální biomasy naměřený na stejných lokalitách (srovnání často nutné pro správnou volbu metody současné analýzy obou proměnných). Nulovou hypotézu budeme testovat srovnáním koeficientů variance podle výše uvedeného vztahu:

Koeficient variance je zde označen jako **V**:

$v_1 = 9$

$\overline{X}_1 = 70{,}73 kg$

$SS_1 = 246{,}1610 kg^2$

$s_1^2 = 27{,}3512 kg^2$

$s_1 = 5{,}23 kg$

$V_1 = 0{,}0739$

$(SS_{\log})_1 = 0{,}0098702632$

$(s^2_{\log})_1 = 0{,}0010966959$

$F = \frac{0{,}0010966959}{0{,}0004961076} = 2{,}21$

$F_{0{,}05\,(2),9,7} = 4{,}82$

$0{,}20 < P < 0.50$

Nezamítáme H_0.

$v_2 = 7$

$\overline{X}_2 = 178{,}82 cm$

$SS_2 = 590{,}1350 cm^2$

$s_2^2 = 84{,}3050 cm^2$

$s_2 = 9{,}18 cm$

$V_2 = 0{,}0513$

$(SS_{\log})_2 = 0{,}0034727534$

$(s^2_{\log})_2 = 0{,}0004961076$

Obsah Zn (mg/kg)	Log (Zn)	Obsah biomasy (mg C/kg)	Log (biomasa)
72,5	1,86034	183,0	2,26245
71,7	1,85552	172,3	2,23629
60,8	1,78390	180,1	2,25551
63,2	1,80072	190,2	2,27921
71,4	1,85370	191,4	2,28194
73,1	1,86392	169,9	2,22943
77,9	1,89154	166,4	2,22115
75,7	1,89910	177,6	2,24944
72,0	1,85733	-	-
69,0	1,84	-	-

Příklad 5.

A one-tailed variance ratio test for the hypothesis that duck clutch size is less variable in captive than in wild birds.

$$H_0 : \sigma_1^2 \geq \sigma_2^2$$

$$H_A : \sigma_1^2 < \sigma_2^2$$

$$\alpha = 0{,}05$$

Clutch Size of Ducks	
Captive	**Wild**
10	9
11	8
12	11
11	12
10	10
11	13
11	11
-	10
-	12

$$n_1 = 7 \qquad n_2 = 9$$

$$v_1 = 6 \qquad v_2 = 8$$

$$SS_1 = 2{,}86 eggs^2 \qquad SS_2 = 20{,}00 eggs^2$$

$$s_1^2 = 0{,}48 eggs^2 \qquad s_2^2 = 2{,}50 eggs^2$$

$$F = \frac{2{,}50}{0{,}48} = 5{,}21$$

$$F_{0{,}05(1),8,6} = 4{,}15$$

Therefore, reject H0

$$0{,}025 < P(F \geq 5{,}21) < 0{,}05$$

Confidence interval for variance ratio

A 1-α confidence interval for the variance ratio, σ_1^2 / σ_2^2, is defined by its lower confidence limit,

$$L_1 = \left(\frac{s_1^2}{s_2^2} \right) \left(\frac{1}{F_{\alpha(2),v_1,v_2}} \right)$$

and its upper confidence limit,

$$L_2 = \left(\frac{s_1^2}{s_2^2} \right) F_{\alpha(2),v_2,v_1}$$

In Example 9.1, $s_1^2 / s_2^2 = 1{,}42$, $F_{0{,}05(2),10,7} = 4{,}76$, and $F_{0{,}05(2),7,10} = 3{,}95$. Therefore, we would calculate L_1=0,298 and L_2=5,61, and we could state

$$P\left(0{,}298 \leq \frac{\sigma_1^2}{\sigma_2^2} \leq 5{,}61 \right) = 0{,}95$$