Elektronický učební materiál

Počítačová podpora cvičení z předmětu M7521 Pravděpodobnost a statistika 1 s využitím programového systému STATISTICA

Autoři: Marie Budíková, Štěpán Mikoláš
Pracoviště: Katedra aplikované matematiky PřF MU v Brně

Základní informace o programovém systému STATISTICA 6

Systém má modulární stavbu. V multilicenci pro Masarykovu univerzitu jsou k dispozici moduly: Basic Statistics/Tables, Multiple Regression, ANOVA, Nonparametrics, Distribution Fitting, Advanced Linear / Nonlinear Models, Multivariate Explorartory Techniques, Industrial Statistics & Six Sigma.

Velké množství informací o systému STATISTICA lze najít na webové stránce společnosti StatSoft, která je jejím distributorem v České republice (www.statsoft.cz). Z této stránky vede rovněž odkaz na elektronickou učebnici statistiky.

STATISTICA 6 má několik typů oken:

- **spreadsheet** (datové okno, má příponu sta, jeho obsah však lze exportovat i v jiných formátech). Do datového okna lze načítat datové soubory nejrůznějších typů (např. z tabulkových procesorů, databázové soubory, ASCII soubory).
- **workbook** (má příponu stw). Do workbooku ukládají výstupy, tj. tabulky a grafy. Skládá se ze dvou oken, v levém okně je znázorněna stromová struktura výstupů, v pravém jsou samotné výstupy. V levém okně se lze pohybovat myší nebo kurzorem, mazat, přesouvat, editovat apod. Výstupy mohou sloužit jako vstupy pro další analýzy a grafy.
- **report** (má příponu str, lze ho uložit i ve formátu rtf, txt či htm). Pokud požadujeme, aby se výstupy ukládaly nejen do workbooku, ale i do reportu, postupujeme takto: Tools – Options – Output Manager – zaškrtneme Also send to Report Window – OK. Report se podobně jako workbook skládá ze dvou oken. Do reportu můžeme vkládat vlastní text, vysvětlující komentáře, poznámky apod. Tabulky a grafy lze v reportu i workbooku dále upravovat.
- **okno grafů** (přípona stg, lze ho uložit i jako bmp, jpg, png a wmf). Získá se tak, že ve workbooku klikneme pravým tlačítkem na graf a vybereme Clone Graph.
- **programovací okno** (přípona svb). Slouží pro zápis programů v jazyku STATISTICA Visual Basic.

  Mezi jednotlivými typy oken se přepínáme pomocí položky Window v hlavním menu.

Téma 1: Bodové zpracování četností

Vzorový příklad: U 20 studentů 1. ročníku byly zjišťovány známky z matematiky, angličtiny a údaje o pohlaví (viz skripta Popisná statistika, příklad 2.4). Příslušný datový soubor se jmenuje **znamky.sta**. Proveďte bodové zpracování četností.

Postup ve STATISTICE:

1. Do programu STATISTICA načtěte datový soubor **znamky.sta**.

2. Znaky nazvěte X, Y, Z, vytvořte jim návěští (X - známka z matematiky, Y - známka z angličtiny, Z - pohlaví studenta) a popište, co znamenají jednotlivé varianty (u znaků X a Y: 1 - výborně, 2 - velmi dobře, 3 - dobře, 4 - neprospěl, u znaku Z: 0 - žena, 1 - muž). Soubor uložte.
   Návod: Kurzor nastavíme na Var1 – 2x klikneme myší – Name X – Long Name známka z matematiky, Text label – výborně, Numeric – 1, velmi dobře, Numeric - 2, dobře, Numeric – 3, neprospěl, Numeric – 4, OK. U proměnné Y lze text label okopírovat z proměnné X – v Text Labels Editor zvolíme Copy from variable X.
   (Přepínání mezi číselnými hodnotami a jejich textovým popisem se děje pomocí tlačítka s obrázkem štítku.)

3. U znaků X a Y vypočtěte absolutní četnosti, relativní četnosti a relativní kumulativní četnosti.
   Návod: Statistics - Basic Statistics/Tables – Frequency tables – OK – Variables X, Y, OK – Summary.
   (Obě tabulky se uloží do workbooku a listovat v nich můžeme pomocí stromové struktury v levém okně.)

Řešení:

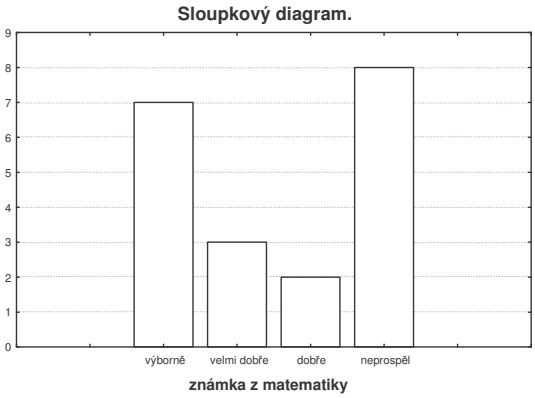
	Frequency table: X: známka z matematiky (zr			
Category	Count	Cumulative Count	Percent	Cumulative Percent
výborně:	7	7	35,00000	35,0000
velmi dobře:	3	10	15,00000	50,0000
dobře	2	12	10,00000	60,0000
neprospěl:	8	20	40,00000	100,0000
Missing	0	20	0,00000	100,0000

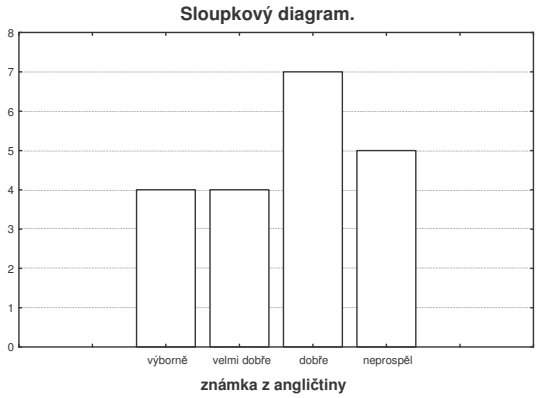
	Frequency table: Y: známka z angličtiny (znai			
Category	Count	Cumulative Count	Percent	Cumulative Percent
výborně:	4	4	20,00000	20,0000
velmi dobře:	4	8	20,00000	40,0000
dobře	7	15	35,00000	75,0000
neprospěl:	5	20	25,00000	100,0000
Missing	0	20	0,00000	100,0000

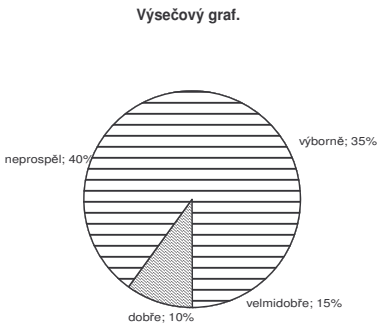
4. Vytvořte sloupkový diagram absolutních četností znaků X a Y.
   Návod: Graphs – Histograms – Variables X, Y – OK- vypneme Normal fit – Advanced – zaškrtneme Breaks between Columns, OK.
   Vytvořte výsečový diagram absolutních četností znaků X a Y.
   Návod: Graphs – 2D Graphs – Pie Charts – Variables X, Y – OK – Advanced – Pie legend Text and Percent (nebo Text and Value) – OK.

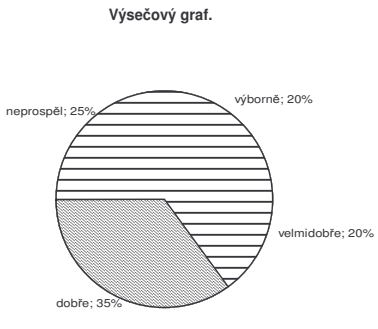
Vytvořte polygon absolutních četností znaků X a Y.
Návod: ve workbooku vstoupíme do tabulky rozložení četností proměnné X. Pomocí Edit – Delete - Cases vymažeme řádek označený Missing. Nastavíme se kurzorem na Count a kliknutím pravého tlačítka vstupíme do menu Line Plot: Entire Columns. Vykreslí se polygon četností.

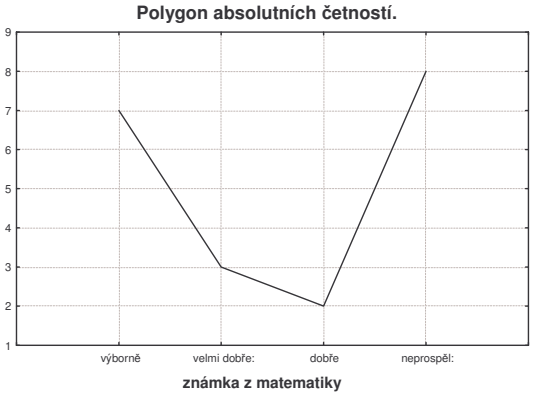
Řešení:

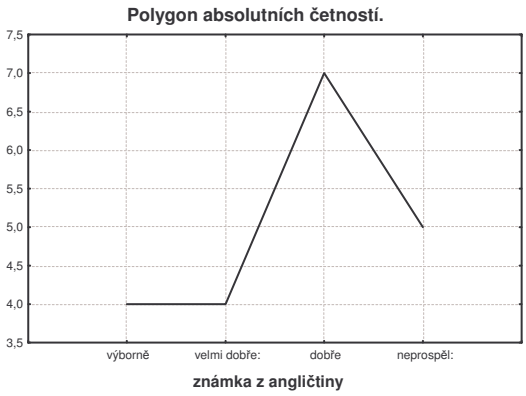








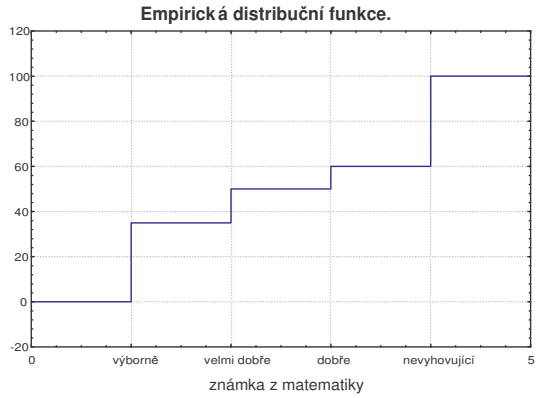


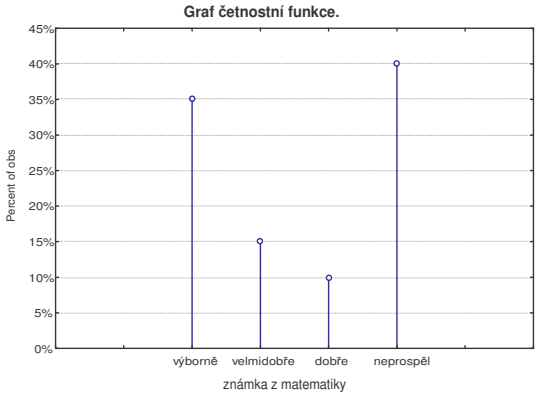


5. Vytvořte graf empirické distribuční funkce znaku X.
Návod: Při tvorbě histogramu zadáme v Advanced volbu Showing Type Cumulative, Y axis % - 2x klikneme myší na pozadí grafu – otevře se okno All Options – vybereme Plot: Bars – Type Rectangles.
Vytvořte graf četnostní funkce znaku X.

Návod: Při tvorbě histogramu zadáme v Advanced Y axis % - 2x klikneme myší na pozadí grafu – vybereme Plot General – zaškrtneme Markers – vybereme Plot: Bars – Type Lines.

Řešení:





6. Z datového souboru vyberte pouze ženy (pouze muže) a úkol 3 proveďte pro ženy (pro muže).
   Návod: Statistics - Basic Statistics/Tables – Frequency tables – OK – Variables X, Y, OK – Select Cases – zaškrtneme Selection Conditions – Include cases – zaškrtneme Specific, selected by Z = 0, OK.

Řešení:

Variační řady známek z matematiky a angličtiny pro ženy.

	Frequency table: X: známka z matematiky			
Category	Count	Cumulative Count	Percent	Cumulative Percent
výborně:	5	5	50,00000	50,0000
velmi dobře:	2	7	20,00000	70,0000
dobře	1	8	10,00000	80,0000
neprospěl:	2	10	20,00000	100,0000
Missing	0	10	0,00000	100,0000

	Frequency table: Y: známka z angličtiny			
Category	Count	Cumulative Count	Percent	Cumulative Percent
výborně:	4	4	40,00000	40,0000
velmi dobře:	2	6	20,00000	60,0000
dobře	1	7	10,00000	70,0000
neprospěl:	3	10	30,00000	100,0000
Missing	0	10	0,00000	100,0000

Variační řady známek z matematiky a z angličtiny pro muže.

Category	Frequency table: X: známka z matematiky			
	Count	Cumulative Count	Percent	Cumulative Percent
výborně:	2	2	20,00000	20,0000
velmi dobře:	1	3	10,00000	30,0000
dobře	1	4	10,00000	40,0000
neprospěl:	6	10	60,00000	100,0000
Missing	0	10	0,00000	100,0000

Category	Frequency table: Y: známka z angličtiny			
	Count	Cumulative Count	Percent	Cumulative Percent
velmi dobře:	2	2	20,00000	20,0000
dobře	6	8	60,00000	80,0000
neprospěl:	2	10	20,00000	100,0000
Missing	0	10	0,00000	100,0000

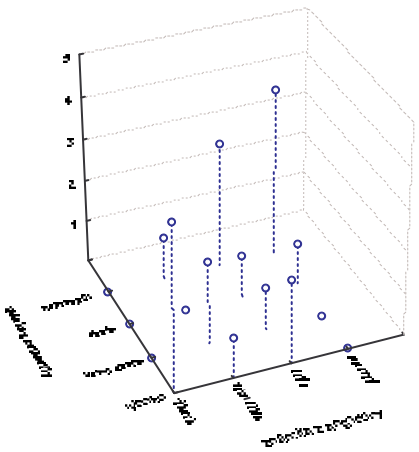
7. Nadále pracujte s celým datovým souborem. Vytvořte kontingenční tabulku absolutních četností znaků X a Y a graf simultánní četností funkce.
   Návod: Statistics - Basic Statistics/Tables – Tables and banners – OK – Select cases – All – OK – Specify tables - List 1 X, List 2 Y, OK, Summary.
   Vytvoření grafu simultánní četnostní funkce: Návrat do Crosstabulation Tables Result – 3D histograms – vybereme Axis Scaling – Mode Manual – Minimum 0 (a totéž provedeme pro Axis Y) – dále vybereme Graph Layout – Type – Spikes – OK. Graf lze natáčet pomocí Point of View.
   Vytvořte kontingenční tabulku sloupcově a řádkově podmíněných relativních četností znaků X a Y.
   Návod: Návrat do Crosstabulation Tables Result – Options - zaškrtneme ve sloupci Compute tables volbu Percentages of column counts (resp. Percentages of row counts).

Řešení:

Kontingenční tabulka absolutních četností známek z matematiky a z angličtiny.

Summary Frequency Table (znamky) Marked cells have counts > 10 (Marginal summaries are not marked)					
X	Y výborně	Y velmi dobře	Y dobře	Y neprospěl	Row Totals
výborně	4	1	2	0	7
velmi dobře	0	2	1	0	3
dobře	0	0	1	1	2
neprospěl	0	1	3	4	8
All Grps	4	4	7	5	20

Simultánní četnostní funkce.



Kontingenční tabulka sloupcově a řádkově podmíněných relativních četností.

	Summary Frequency Table (znamky) Marked cells have counts > 10 (Marginal summaries are not marked)					
	X	Y výborně	Y velmi dobře	Y dobře	Y neprospěl	Row Totals
Count	výborně	4	1	2	0	7
Column Percent		100,00%	25,00%	28,57%	0,00%	
Row Percent		57,14%	14,29%	28,57%	0,00%	
Count	velmi dobře	0	2	1	0	3
Column Percent		0,00%	50,00%	14,29%	0,00%	
Row Percent		0,00%	66,67%	33,33%	0,00%	
Count	dobře	0	0	1	1	2
Column Percent		0,00%	0,00%	14,29%	20,00%	
Row Percent		0,00%	0,00%	50,00%	50,00%	
Count	neprospěl	0	1	3	4	8
Column Percent		0,00%	25,00%	42,86%	80,00%	
Row Percent		0,00%	12,50%	37,50%	50,00%	

Téma 2: Intervalové zpracování četností

Vzorový příklad: U 60 vzorků oceli byly zjišťovány hodnoty meze plasticity a meze pevnosti v $kpcm^{-2}$ (viz skripta Popisná statistika, př. 2.5). Datový soubor se jmenuje **ocel.sta**. Proveďte intervalové zpracování četností.

Postup ve STATISTICE:

1. Načtěte soubor **ocel.sta**. Proměnným X a Y vytvořte návěští „mez plasticity" a „mez pevnosti".

2. Pro X a Y použijeme intervalové zpracování četností. Podle Sturgesova pravidla je optimální počet třídicích intervalů 7. Musíme zjistit minimum a maximum, abychom vhodně stanovili třídicí intervaly.
   Návod: Statistics – Basic Statistics/Tables – Descriptive statistics - Variables X,Y – zaškrtneme Minimum&maximum – Summary. (Pro X je minimum 33 a maximum 160, tedy vhodná volba třídicích intervalů je (30,50>, (50,70>, ..., (150,170>, pro Y je minimum 52 a maximum 189, tedy třídicí intervaly zvolíme (50,70>, (70,90>, ... , (170,190>)

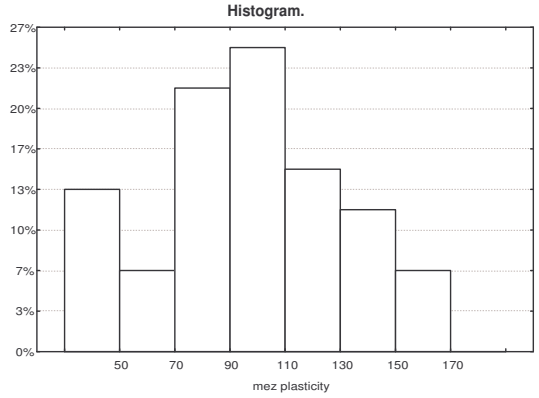
Řešení:

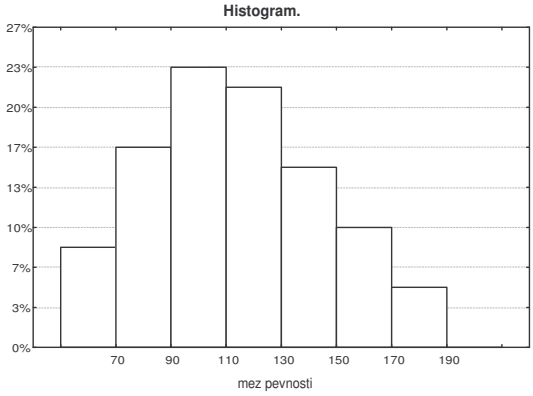
Variable	Descriptive Statistics (c	
	Minimum	Maximum
X	33,00000	160,0000
Y	52,00000	189,0000

U znaku X volíme dolní mez prvního třídicího intervalu 30, horní mez posledního třídicího intervalu 170. U znaku Y volíme dolní mez prvního třídicího intervalu 50, horní mez posledního třídicího intervalu 190.
Celkem tedy třídicí intervaly znak X budou:
(30,50>, (50,70>, (70,90>, (90,110>, (110,130>, (130,150>, (150,170>
a pro znak Y:
(50,70>, (70,90>, (90,110>, (110,130>, (130,150>, (150,170>, (170,190>.

3. Vytvořte histogram pro X a pro Y.
   Návod: Graphs – Histograms – Variables X – vypneme Normal fit – Advanced – zaškrtneme Boundaries – Specify Boundaries – 50 70 90 110 130 150 170 OK – Y Axis %. Po vykreslení histogramu lze 2 x klepnout na pozadí grafu a ve volbě All Options měnit různé vlastnosti grafu.

Řešení:





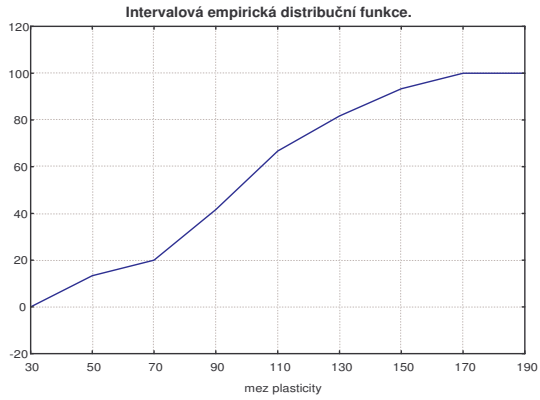
4. Proveďte zakódování hodnot proměnných X a Y do příslušných třídicích intervalů.
   Návod: Insert – Add Variables – 2 – After Y – OK – přejmenujeme je na RX a RY. Nastavíme se kurzorem na RX – Data – Recode - vyplníme podmínky pro všech 7 kategorií. (Pozor – podmínky se musí psát ve tvaru X > 30 and X <= 50 atd.). Pak klepneme na OK. Analogicky pro Y.

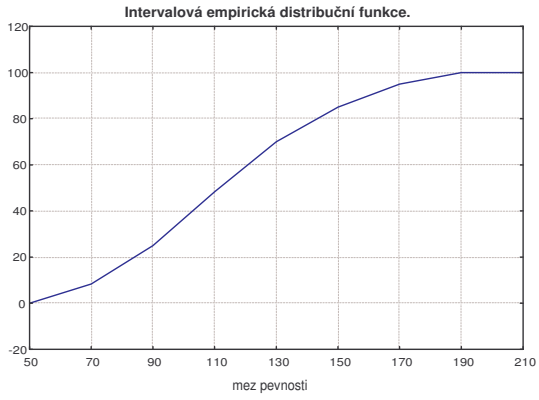
5. Vytvořte graf intervalové empirické empirické distribuční funkce pro X.
   Návod: Vytvoříme Frequency table pro RX. Před 1. případ vložíme řádek, kde do Category napíšeme 0 a do Cumulative Percent také 0. Nastavíme se kurzorem na Cumulative Percent – Graphs – Graphs of Block Data – Custom Graph from Block by Column – Line Plots (Variables) – OK.

Řešení:

	Tabulka četností:RX (ocel)			
Kategorie	Četnost	Kumulativní četnost	Rel.četnost	Kumulativní rel.četnost
0				0,0000
1	8	8	13,33333	13,3333
2	4	12	6,66667	20,0000
3	13	25	21,66667	41,6667
4	15	40	25,00000	66,6667
5	9	49	15,00000	81,6667
6	7	56	11,66667	93,3333
7	4	60	6,66667	100,0000
ChD	0	60	0,00000	100,0000

	Tabulka četností:RY (ocel)			
Kategorie	Četnost	Kumulativní četnost	Rel.četnost	Kumulativní rel.četnost
0				0,0000
1	5	5	8,33333	8,3333
2	10	15	16,66667	25,0000
3	14	29	23,33333	48,3333
4	13	42	21,66667	70,0000
5	9	51	15,00000	85,0000
6	6	57	10,00000	95,0000
7	3	60	5,00000	100,0000
ChD	0	60	0,00000	100,0000





6. Sestavte kontingenční tabulky absolutních četností (relativních četností, sloupcově a řádkově podmíněných relativních četností) dvourozměrných třídicích intervalů pro (X,Y).
Návod: Viz úkol č. 7 v tématu 1, kde budeme pracovat s proměnnými RX a RY.

Řešení:

Kontingenční tabulky absolutních a relativních četností.

	Summary Frequency Table (ocel) Table: RX(7) x RY(7)								
	RX	RY (50,70>	RY (70,90>	RY (90,110>	RY 110,130	RY 130,150	RY 150,170	RY 70,190	Row Totals
Count	(30,50>	5	3	0	0	0	0	0	8
Total Percent		8,33%	5,00%	0,00%	0,00%	0,00%	0,00%	0,00%	13,33%
Count	(50,70>	0	3	1	0	0	0	0	4
Total Percent		0,00%	5,00%	1,67%	0,00%	0,00%	0,00%	0,00%	6,67%
Count	(70,90>	0	4	7	1	1	0	0	13
Total Percent		0,00%	6,67%	11,67%	1,67%	1,67%	0,00%	0,00%	21,67%
Count	(90,110>	0	0	6	8	1	0	0	15
Total Percent		0,00%	0,00%	10,00%	13,33%	1,67%	0,00%	0,00%	25,00%
Count	110,130>	0	0	0	4	5	0	0	9
Total Percent		0,00%	0,00%	0,00%	6,67%	8,33%	0,00%	0,00%	15,00%
Count	(130,150	0	0	0	0	2	5	0	7
Total Percent		0,00%	0,00%	0,00%	0,00%	3,33%	8,33%	0,00%	11,67%
Count	(150,170	0	0	0	0	0	1	3	4
Total Percent		0,00%	0,00%	0,00%	0,00%	0,00%	1,67%	5,00%	6,67%
Count	All Grps	5	10	14	13	9	6	3	60
Total Percent		8,33%	16,67%	23,33%	21,67%	15,00%	10,00%	5,00%	

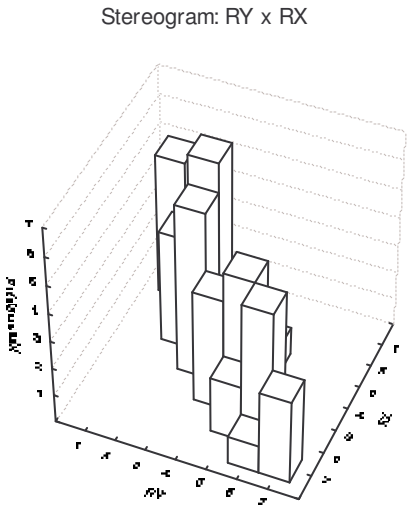
Kontingenční tabulky řádkově a sloupcově podmíněných relativních četností.

	Summary Frequency Table (ocel) Table: RX(7) x RY(7)								
	RX	RY 1	RY 2	RY 3	RY 4	RY 5	RY 6	RY 7	Row Totals
Count	1	5	3	0	0	0	0	0	8
Row Percent		62,50%	37,50%	0,00%	0,00%	0,00%	0,00%	0,00%	
Count	2	0	3	1	0	0	0	0	4
Row Percent		0,00%	75,00%	25,00%	0,00%	0,00%	0,00%	0,00%	
Count	3	0	4	7	1	1	0	0	13
Row Percent		0,00%	30,77%	53,85%	7,69%	7,69%	0,00%	0,00%	
Count	4	0	0	6	8	1	0	0	15
Row Percent		0,00%	0,00%	40,00%	53,33%	6,67%	0,00%	0,00%	
Count	5	0	0	0	4	5	0	0	9
Row Percent		0,00%	0,00%	0,00%	44,44%	55,56%	0,00%	0,00%	
Count	6	0	0	0	0	2	5	0	7
Row Percent		0,00%	0,00%	0,00%	0,00%	28,57%	71,43%	0,00%	
Count	7	0	0	0	0	0	1	3	4
Row Percent		0,00%	0,00%	0,00%	0,00%	0,00%	25,00%	75,00%	
Count	All Grps	5	10	14	13	9	6	3	60

	Summary Frequency Table (ocel) Table: RX(7) x RY(7)								
	RX	RY 1	RY 2	RY 3	RY 4	RY 5	RY 6	RY 7	Row Totals
Count	1	5	3	0	0	0	0	0	8
Column Percent		100,00%	30,00%	0,00%	0,00%	0,00%	0,00%	0,00%	
Count	2	0	3	1	0	0	0	0	4
Column Percent		0,00%	30,00%	7,14%	0,00%	0,00%	0,00%	0,00%	
Count	3	0	4	7	1	1	0	0	13
Column Percent		0,00%	40,00%	50,00%	7,69%	11,11%	0,00%	0,00%	
Count	4	0	0	6	8	1	0	0	15
Column Percent		0,00%	0,00%	42,86%	61,54%	11,11%	0,00%	0,00%	
Count	5	0	0	0	4	5	0	0	9
Column Percent		0,00%	0,00%	0,00%	30,77%	55,56%	0,00%	0,00%	
Count	6	0	0	0	0	2	5	0	7
Column Percent		0,00%	0,00%	0,00%	0,00%	22,22%	83,33%	0,00%	
Count	7	0	0	0	0	0	1	3	4
Column Percent		0,00%	0,00%	0,00%	0,00%	0,00%	16,67%	100,00%	
Count	All Grps	5	10	14	13	9	6	3	60

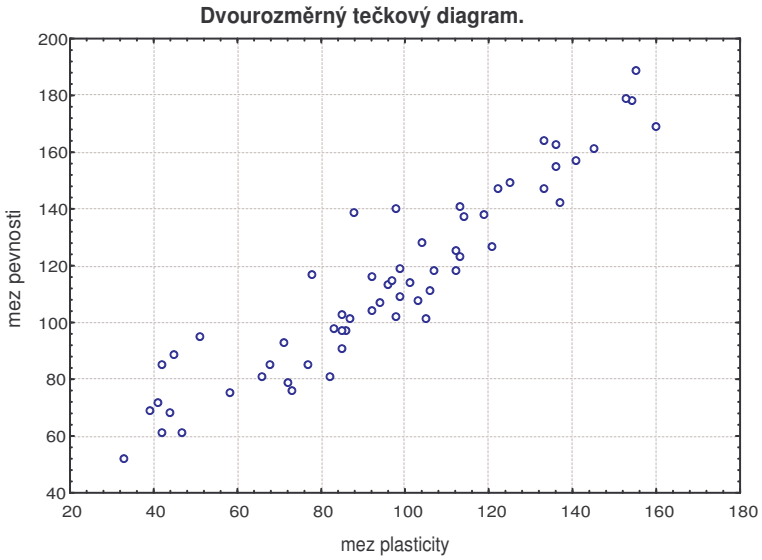
7. Vytvořte stereogram pro (RX,RY).
   Návod: V tabulce Crosstabulation Tables Result zaškrtneme 3D histograms. Ve volbě Axis Scaling (pro RX i pro RY) změníme Mode na Manual – Minimum 0.

Řešení:



8. Nakreslete dvourozměrný tečkový diagram pro (X,Y).
   Návod: Graphs – Scatterplots – Variables X,Y – OK vypneme Linear fit – OK.

Řešení:



Téma 3: Výpočet číselných charakteristik jednorozměrného datového souboru

Vzorový příklad: Pro následující datové soubory vypočtěte číselné charakteristiky.

Postup ve STATISTICE:

1. Načtěte soubor **znamky.sta**. Pro známky z matematiky a angličtiny vypočtěte medián, dolní a horní kvartil a kvartilovou odchylku. Výsledky porovnejte s údaji ve skriptech Popisná statistika (viz str. 28).
   Návod: Stastistics – Basic Statistics/Tables – Descriptive Statistics – OK - Variables X, Y, OK – zaškrtneme Median, Lower & upper quartiles, Quartile range – Summary.

Řešení:

Variable	Descriptive Statistics (znamky)			
	Median	Lower Quartile	Upper Quartile	Quartile Range
X	2,500000	1,000000	4,000000	3,000000
Y	3,000000	2,000000	3,500000	1,500000

2. Načtěte soubor **ocel.sta**. Pro mez plasticity a mez pevnosti vypočtěte aritmetické průměry, směrodatné odchylky a rozptyly. Výsledky porovnejte s údaji ve skriptech Popisná statistika (viz str. 30).
   Návod: Stastistics – Basic Statistics/Tables – Descriptive Statistics – OK - Variables X, Y, OK – zaškrtneme Mean, Standard Deviation, Variance – Summary.
   Vysvětlení: Rozptyl a směrodatná odchylka vyjdou ve STATISTICE jinak než ve skriptech, protože STATISTICA ve vzorci pro výpočet rozptylu nepoužívá 1/n, ale 1/(n-1).

Řešení:

Variable	Descriptive Statistics (ocel)		
	Mean	Variance	Std.Dev.
X	95,8833	1070,240	32,71453
Y	114,4000	1075,125	32,78911

3. Je třeba si uvědomit, že průměr a rozptyl nepopisují rozložení četností jednoznačně. Existují datové soubory, které mají shodný průměr i rozptyl, ale přesto se jejich rozložení četností velmi liší. Tuto skutečnost dobře ilustruje následující příklad: Tři skupiny studentů o počtech 149, 69 a 11 odpovídaly při testu na 10 otázek. Znak X je počet správně zodpovězených otázek. Známe absolutní četnosti znaku X ve všech třech skupinách.

č. sk.	X										
	0	1	2	3	4	5	6	7	8	9	10
1	2	5	15	20	25	15	25	20	15	5	2
2	4	3	2	1	0	49	0	1	2	3	4
3	1	0	0	0	0	9	0	0	0	0	1

Vypočtěte průměr (mean), rozptyl (variance), šikmost (skewness) a špičatost (kurtosis) počtu správně zodpovězených otázek ve všech třech skupinách. Nakreslete sloupkové diagramy absolutních četností.

Návod: Při zadávání dat do STATISTIKY utvořte čtyři proměnné a 11 případů. V 1. sloupci budou varianty znaku X (tj. 0 až 10), v dalších sloupcích pak absolutní četnosti. Proměnné pojmenujeme X, SK1, SK2, SK3. V tabulce Descriptive Statistics zadáme Variable X a klepneme na tlačítko W, abychom program upozornili, že budeme pracovat s daty zadanými pomocí absolutních četností. Zadáme Weight variable SK1, zaškrtneme Status On, OK – zaškrtneme Mean, Variance, Skewness, Kurtosis – Summary. Dále pro znak X nakreslíme sloupkový diagram – viz úkol č. 4 v tématu „Bodové rozložení četností". Tytéž úkoly provedeme s Weight variable SK2 a SK3.

Řešení:

1. skupina (X weightet by SK1)

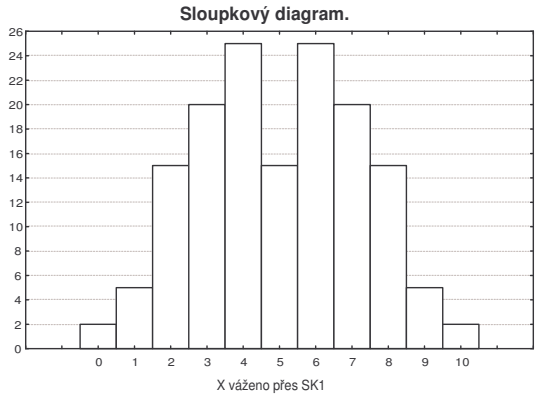
	Descriptive Statistics			
Variable	Mean	Variance	Skewness	Kurtosis
X	5,000000	5,000000	-0,000000	-0,759500

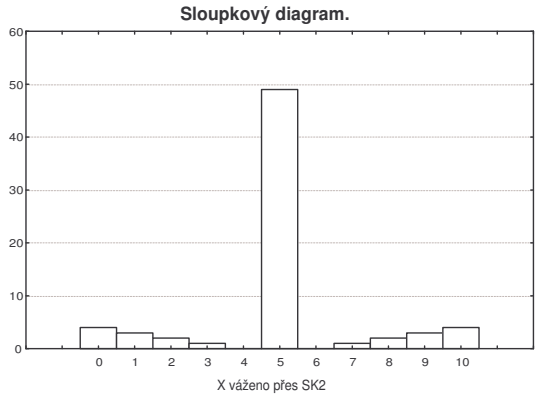
2. skupina (X weightet by SK2)

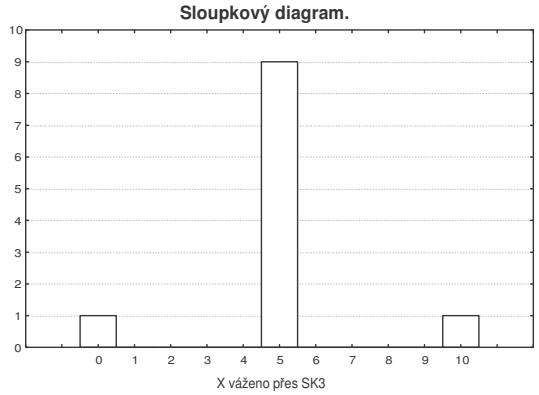
	Descriptive Statistics			
Variable	Mean	Variance	Skewness	Kurtosis
X	5,000000	5,000000	-0,000000	1,291133

3. skupina (X weightet by SK3)

	Descriptive Statistics (cischar)			
Variable	Mean	Variance	Skewness	Kurtosis
X	5,000000	5,000000	-0,000000	5,000000







Interpretace: Všechny tři skupiny mají týž průměr, rozptyl a šikmost, liší se pouze ve špičatosti. Rozložení četností počtu správně zodpovězených otázek je ovšem velmi rozdílné.

Téma 4: Korelace a regrese

Vzorový příklad: Pro následující datové soubory proveďte korelační, resp. regresní analýzu.

Postup ve STATISTICE:

1. Načtěte soubor **znamky.sta**. Vypočtěte Spearmanův korelační koeficient známek z matematiky a angličtiny pro všechny studenty, pak zvlášť pro muže a zvlášť pro ženy. Získané výsledky interpretujte.
   (Spearmanův korelační koeficient měří těsnost lineární závislosti dvou ordinálních proměnných x, y a počítá se podle vzorce:

   $$r_S = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (R_i - Q_i)^2 ,$$

   kde R_i je pořadí x_i- tj. počet těch hodnot $x_1, ..., x_n$, které jsou $\leq x_i$ a Q_i je pořadí y_i.)

   Návod: Po načtení souboru zvolíme Statistics – Nonparametrics – Correlations – OK – Variables First variable list X, Second variable list Y – OK – Spearman R. Počítáme-li r_S pro muže, vybereme v tabulce tabulce Nonparametric Correlation tlačítko Select Cases – Specific, select by Z=1.

Řešení:

Pro všechny

Pair of Variables	Spearman Rank Order Correlations (zna MD pairwise deleted Marked correlations are significant at p <			
	Valid N	Spearman R	t(N-2)	p-level
X & Y	20	0,688442	4,027090	0,000791

Pro muže (if Z=1)

Pair of Variables	Spearman Rank Order Correlations (zna MD pairwise deleted Marked correlations are significant at p <			
	Valid N	Spearman R	t(N-2)	p-level
X & Y	10	0,373544	1,138990	0,287662

Pro ženy (if Z=0)

Pair of Variables	Spearman Rank Order Correlations (zna MD pairwise deleted Marked correlations are significant at p <			
	Valid N	Spearman R	t(N-2)	p-level
X & Y	10	0,860314	4,773446	0,001402

2. Vysvětlení významu Pearsonova korelačního koeficientu: Načtěte soubor **korkoef.sta**, který obsahuje proměnné X,Y1,Y2,Y3,Y4, X4. Vypočtěte Pearsonovy korelační koeficienty dvojic proměnných (X,Y1), (X,Y2), (X,Y3), (X4,Y4) a pro každou z uvedených dvojic proměnných

nakreslete dvourozměrný tečkový diagram. Pro které dvojice proměnných se hodí Pearsonův korelační koeficient jako vhodná míra těsnosti lineární závislosti?
Návod: Statistics – Basis Statistics/Tables – Correlation matrices – OK – One variable list X, Y1 – OK – Summary: Correlation matrix – Návrat do Product-Moment and Partial Correlations – Advanced/plot – 2D Scatterplots – OK – First X, Second Y1 – OK. Analogicky pro ostatní dvojice proměnných.

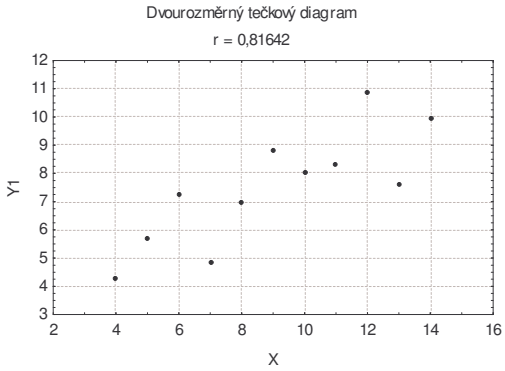
Řešení:

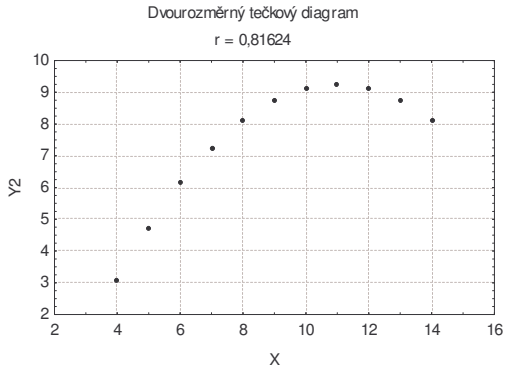
Variable	Correlations (korkoe	
	X	Y1
X	1,000000	0,816421
Y1	0,816421	1,000000

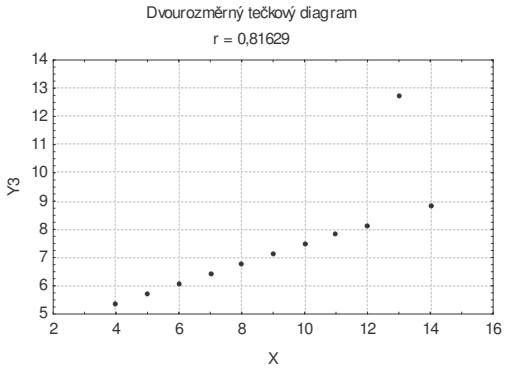
Variable	Correlations (korkoe	
	X	Y2
X	1,000000	0,816237
Y2	0,816237	1,000000

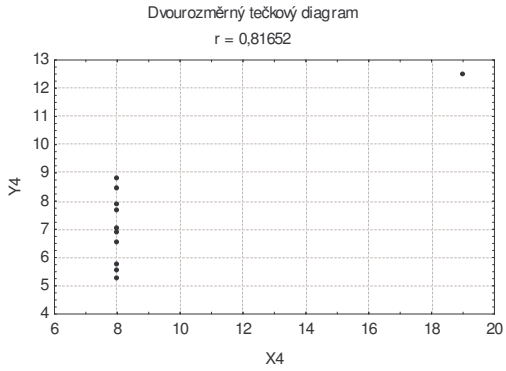
Variable	Correlations (korkoe	
	X	Y3
X	1,000000	0,816287
Y3	0,816287	1,000000

Variable	Correlations (korkoe	
	X4	Y4
X4	1,000000	0,816521
Y4	0,816521	1,000000









3. Načtěte do STATISTIKY soubor **ocel.sta**. Vypočtěte kovarianci a Pearsonův koeficient korelace meze plasticity a meze pevnosti. Porovnejte s výsledky ve skriptech Popisná statistika (str. 30).
Návod: Po načtení souboru zvolíme Statistics - Multiple Regression - Variables Independent X, Dependent Y – OK – OK – Residuals/assumption-prediction – Descriptive statistics – Covariances. Pro získání korelačního koeficientu zvolíme Correlation místo Covariances. Vysvětlení: Kovariance vyjde ve STATISTICE jinak než ve skriptech, protože ve STATISTICE se ve vzorci pro výpočet kovariance nepoužívá 1/n, ale 1/(n-1).

Řešení:

Variable	Correlations (ocel)	
	X	Y
X	1,000000	0,934548
Y	0,934548	1,000000

Variable	Covariances (ocel)	
	X	Y
X	1070,240	1002,471
Y	1002,471	1075,125

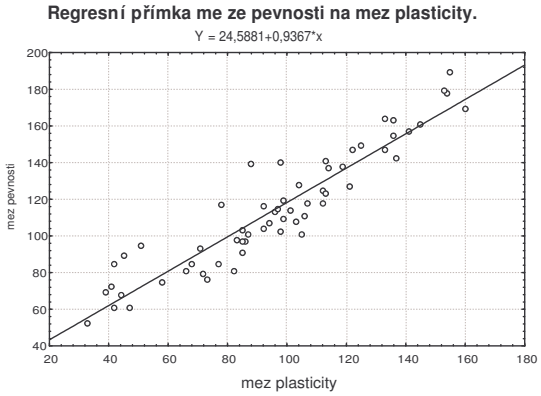
4. Určete koeficienty regresní přímky meze pevnosti na mez plasticity a stanovte index determinace. Určete regresní odhad meze pevnosti, je-li mez plasticity 110. Nakreslete regresní přímku do dvourozměrného tečkového diagramu.
   Návod: V tabulce Multiple Regression zvolíme Variables Independent X, Dependent Y – OK – Summary:Regression results. Ve výstupní tabulce najdeme koeficient b_0 ve sloupci B na řádku označeném Intercept, koeficient b_1 ve sloupci B na řádku označeném X, index determinace pod označením R2.
   Pro výpočet predikované hodnoty zvolíme Residuals/assumption/prediction Predict dependent variable X:110 - OK. Ve výstupní tabulce je hledaná hodnota označena jako Predictd.
   Nakreslení regresní přímky: Návrat do Multiple Regression – Residuals / assumption / prediction – Perform residuals analysis – Scatterplots – Bivariate correlation – X, Y – OK.
   Jiný způsob: Do dvourozměrného tečkového diagramu nakreslíme regresní přímku tak, že v tabulce 2D Scatterplots zvolíme Fit Linear, OK.

Řešení:

Statistic	Summary
	Value
Multiple R	0,9345
Multiple R2	0,8734
Adjusted R2	0,8712
F(1,58)	400,0641
p	0,0000
Std.Err. of Estimate	11,7677

Variable	Predicting Values for (ocel) variable:Y		
	B-Weight	Value	B-Weight * Value
X	0,936679	110,0000	103,0346
Intercept			24,5881
Predicted			127,6228
-95,0%CL			124,3063
+95,0%CL			130,9392

N=60	Regression Summary for Dependent Variable:Y (ocel) R= ,93454811 R2= ,87338017 Adjusted R2= ,87119707 F(1,58)=400,06 p<0,0000 Std.Error of estimate: 11,768					
	Beta	Std.Err. of Beta	B	Std.Err. of B	t(58)	p-level
Intercept			24,58814	4,740272	5,18707	0,000003
X	0,934548	0,046724	0,93668	0,046830	20,00160	0,000000



5. U sedmi náhodně vybraných strojů v určitém podniku se zjišťovalo stáří stroje v letech (proměnná x) a týdenní náklady v Kč na údržbu stroje (proměnná y). Data:
(1,35), (1,52), (3,81), (3,105), (5,100), (6,125), (7, 120)
Data znázorněte graficky. Vyzkoušejte následující čtyři modely:
$y = \beta_0 + \beta_1 x$, $y = \beta_0 + \beta_1 \sqrt{x}$, $y = \beta_0 + \beta_1 \log_{10} x$, $y = \beta_0 + \beta_1 1/x$. Vyberte ten model, který poskytuje nejvyšší index determinace. Určete regresní odhad týdenních nákladů pro stroj starý čtyři roky.
Návod: Datový soubor s proměnnými X a Y doplňte o proměnné SQRTX, LOGX a INVX. Hodnoty proměnné SQRTX získáte tak, že do Long Name napíšete =sqrt(x). (Analogicky pro ostatní proměnné.) Regresní analýzu provedete tak, že roli nezávisle proměnné bude hrát proměnná X, pak SQRTX, LOGX a nakonec INVX.

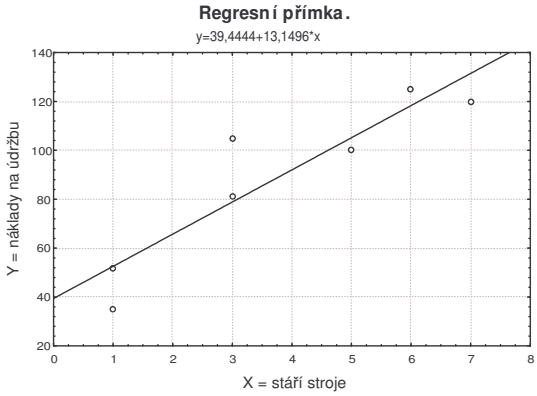
Řešení:

Model s proměnnou X

Statistic	Summary Value
Multiple R	0,91004
Multiple R2	0,82817
Adjusted R2	0,79381
F(1,5)	24,09909
p	0,00444
Std.Err. of Estimate	15,48711

Variable	Predicting Values for (stroje) variable: Y		
	B-Weight	Value	B-Weight * Value
X	13,14957	4,000000	52,5983
Intercept			39,4444
Predicted			92,0427
-95,0%CL			76,8676
+95,0%CL			107,2179

N=7	Regression Summary for Dependent Variable: Y (stroje) R= ,91004028 R2= ,82817331 Adjusted R2= ,79380797 F(1,5)=24,099 p<,00444 Std.Error of estimate: 15,487					
	Beta	Std.Err. of Beta	B	Std.Err. of B	t(5)	p-level
Intercept			39,44444	11,54341	3,417054	0,018898
X	0,910040	0,185379	13,14957	2,67862	4,909082	0,004439

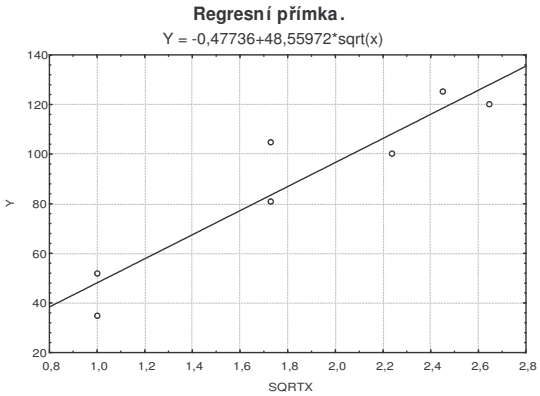


Model s odmocninou

Statistic	Summary Value
Multiple R	0,93924
Multiple R2	0,88217
Adjusted R2	0,85860
F(1,5)	37,43261
p	0,00169
Std.Err. of Estimate	12,82508

Variable	Predicting Values for (stroje) variable: Y		
	B-Weight	Value	B-Weight * Value
SQRTX	48,55972	2,000000	97,1194
Intercept			-0,4774
Predicted			96,6421
-95,0%CL			83,6962
+95,0%CL			109,5880

N=7	Regression Summary for Dependent Variable: Y (stroje) R= ,93923698 R2= ,88216611 Adjusted R2= ,85859933 F(1,5)=37,433 p<,00169 Std.Error of estimate: 12,825					
	Beta	Std.Err. of Beta	B	Std.Err. of B	t(5)	p-level
Intercept			-0,47736	15,29638	-0,031207	0,976312
SQRTX	0,939237	0,153515	48,55972	7,93690	6,118220	0,001691

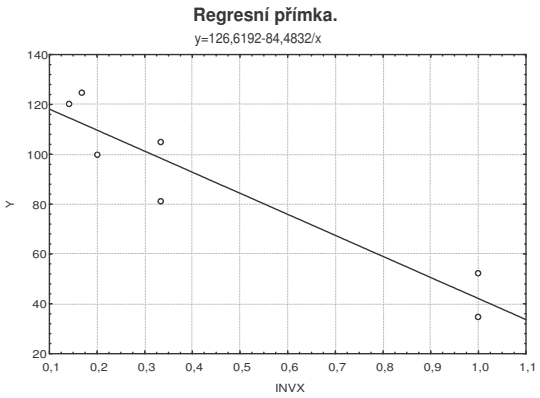


Model s převrácenou hodnotou

Statistic	Summary Value
Multiple R	0,94282
Multiple R2	0,88891
Adjusted R2	0,86670
F(1,5)	40,01016
p	0,00146
Std.Err. of Estimate	12,45245

Variable	Predicting Values for (stroje) variable: Y		
	B-Weight	Value	B-Weight * Value
INVX	-84,4832	0,250000	-21,1208
Intercept			126,6192
Predicted			105,4984
-95,0%CL			91,5231
+95,0%CL			119,4738

N=7	Regression Summary for Dependent Variable: Y (stroje) R= ,94282234 R2= ,88891396 Adjusted R2= ,86669676 F(1,5)=40,010 p<,00146 Std.Error of estimate: 12,452					
	Beta	Std.Err. of Beta	B	Std.Err. of B	t(5)	p-level
Intercept			126,6192	7,67327	16,50134	0,000015
INVX	-0,942822	0,149054	-84,4832	13,35627	-6,32536	0,001456

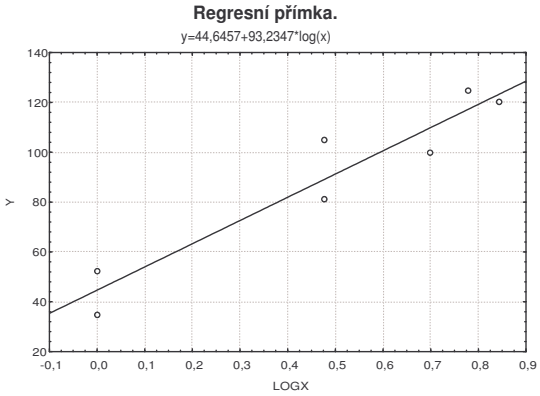


Model s logaritmem

Statistic	Summary Value
Multiple R	0,95349
Multiple R2	0,90915
Adjusted R2	0,89097
F(1,5)	50,03321
p	0,00087
Std.Err. of Estimate	11,26153

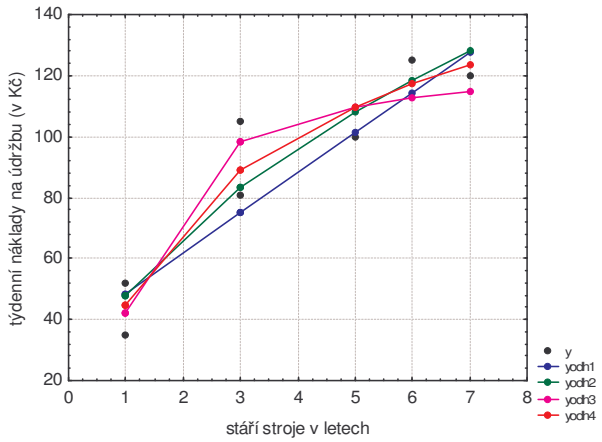
Variable	Predicting Values for (stroje) variable: Y		
	B-Weight	Value	B-Weight * Value
LOGX	93,23472	0,602060	56,1329
Intercept			44,6457
Predicted			100,7786
-95,0%CL			88,9325
+95,0%CL			112,6247

N=7	Regression Summary for Dependent Variable: Y (stroje) R= ,95349135 R2= ,90914576 Adjusted R2= ,89097491 F(1,5)=50,033 p<,00087 Std.Error of estimate: 11,262					
	Beta	Std.Err. of Beta	B	Std.Err. of B	t(5)	p-level
Intercept			44,64571	7,49541	5,956407	0,001907
LOGX	0,953491	0,134799	93,23472	13,18100	7,073415	0,000874



Nejvyšší hodnotu indexu determinace vykazuje model s logaritmem.

Výsledky všech čtyř modelů:



Téma 5: Výpočty pravděpodobností pomocí distribuční funkce binomického rozložení

STATISTICA poskytuje hodnoty distribučních funkcí mnoha rozložení. Omezíme se na binomické rozložení (funkce IBinom(x,p,n), kde x ... počet úspěchů, p ... pravděpodobnost úspěchu v jednom pokusu, n ... celkový počet pokusů).

Vzorový příklad na binomické rozložení: Pojišťovna zjistila, že 12% pojistných událostí je způsobeno vloupáním. Jaká je pravděpodobnost, že mezi 30 náhodně vybranými pojistnými událostmi bude způsobeno vloupáním

a) nejvýše 6,
b) aspoň 6,
c) právě 6,
d) od dvou do pěti?

Řešení:

Náhodná veličina X udává počet pojistných událostí způsobených vloupáním,
$X \sim Bi(30; 0{,}12)$.
ad a) $P(X \leq 6) = \Phi(6) = 0{,}9393$,
ad b) $P(X \geq 6) = 1 - P(X \leq 5) = 1 - \Phi(5) = 0{,}1431$,
ad c) $P(X=6) = \Phi(6) - \Phi(5) = 0{,}0825$,
ad d) $P(2 \leq X \leq 5) = \Phi(5) - \Phi(1) = 0{,}7469$.

Postup ve STATISTICE:
Otevřeme nový datový soubor se čtyřmi proměnnými a o jednom případu.

Řešení:

Do Long Name 1. proměnné napíšeme =IBinom(6;0,12;30).
Do Long Name 2. proměnné napíšeme =1-IBinom(5;0,12;30).
Do Long Name 3. proměnné napíšeme =IBinom(6;0,12;30)-IBinom(5;0,12;30).
Do Long Name 3. proměnné napíšeme =IBinom(5;0,12;30)-IBinom(1;0,12;30).

a) $P(X \leq 6) = \Phi(6) =$ IBinom(6;0,12;30) = 0,939393
b) $P(X \geq 6) = 1 - P(X \leq 5) = 1 - \Phi(5) =$ 1–IBinom(5;0,12;30) = 0,143077
c) $P(X=6) = P(X \leq 6) - P(X \leq 5) =$ IBinom(6;0,12;30)-IBinom(5;0,12;30) = 0,082470
d) $P(2 \leq X \leq 5) = P(1 < X \leq 5) = \Phi(5) - \Phi(1) =$ IBinom(5;0,12;30)-IBinom (1;0,12;30) = 0,746953

Příklady ze skript Teorie pravděpodobnosti a matematická statistika, kapitola 4:

Příklad 4.10.

n = 10, úspěch = narození chlapce, pravděpodobnost úspěchu $\vartheta = 0{,}5$
X udává počet narozených chlapců

Řešení:

a) $P(X=5) = P(X \leq 5) - P(X \leq 4) = \Phi(5) - \Phi(4) =$ IBinom(5;0,5;10)-IBinom(4;0,5;10) = 0,246094
b) $P(3 \leq X \leq 8) = P(2 < X \leq 8) = \Phi(8) - \Phi(2) =$ IBinom(8;0,5;10) - IBinom(2;0,5;10) = 0,934570

Příklad 4.11.

n = 7, úspěch = potkání dvou vlaků během 24 hodin, pravděpodobnost úspěchu $\vartheta = 0,2$
X udává počet potkání dvou vlaků během týdne

Řešení:

$P(X=3) = P(X\leq3) - P(X\leq2) = \Phi(3) - \Phi(2) = IBinom(3;0,2;7) - IBinom(2;0,2;7) = 0,11468$
$P(X\leq3) = IBinom(3;0,2;7) = 0,966656$
$P(X\geq3) = 1 - P(X<3) = 1 - P(X\leq2) = 1 - \Phi(2) = 1 - IBinom(2;0,2;7) = 0,148032$

Příklad 4.12.

Úspěch je výhra partie se stejně silným soupeřem, když remíza je vyloučena
Pravděpodobnost úspěchu $\vartheta = 0,5$, X udává počet úspěchů

a) n = 4
b) n = 8

Řešení:

ad a) $P(X=3) = P(X\leq3) - P(X\leq2) = \Phi(3)-\Phi(2) = IBinom(3;0,5;4) - IBinom(2;0,5;4) = 0,250000$
ad b) $P(X=5) = P(X\leq5) - P(X\leq4) = \Phi(5)-\Phi(4) = IBinom(5;0,5;8) - IBinom(4;0,5;8) = 0,218750$

Příklad 4.13.

n = 20, úspěch je padnutí tří líců při hodu třemi mincemi, $\vartheta = 1/8 = 0,125$,
X udává počet úspěchů

Řešení:

$P(X\geq1) = 1 - P(X<1) = 1 - P(X\leq0) = 1 - IBinom(0;0,125;20) = 0,930791$

Příklad 4.14.

n = 5, úspěch je padnutí tří jedniček při hodu třemi kostkami, $\vartheta = 1/6^3 = 1/216$,
X udává počet úspěchů

Řešení:

$P(X=2) = P(X\leq2) - P(X\leq1) = \Phi(2) - \Phi(1) = IBinom(2;1/216;5) - IBinom(1;1/216;5) = 0,000211$

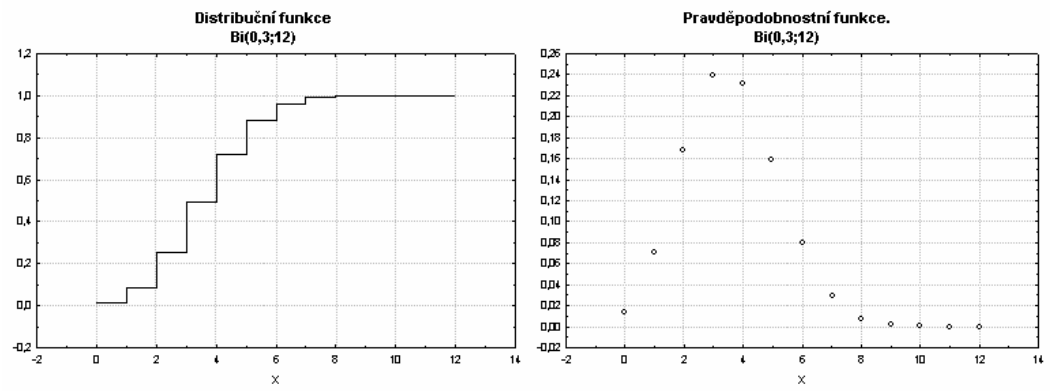
Téma 6: Kreslení grafů distribuční funkce a pravděpodobnostní funkce binomického rozložení

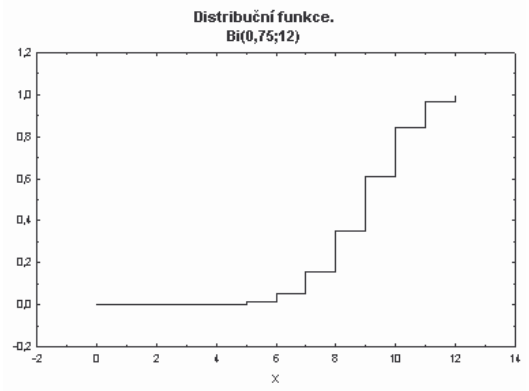
Vzorový příklad: Nakreslete graf distribuční funkce a pravděpodobnostní funkce náhodné veličiny X ~ Bi(12; 0,3)

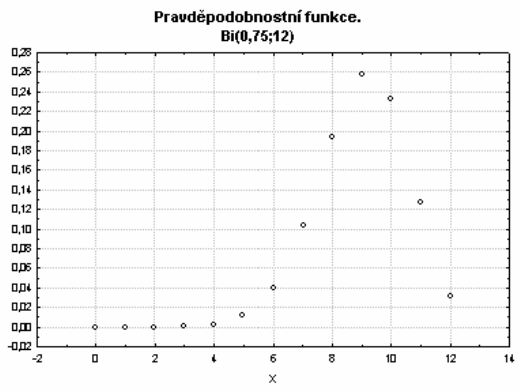
Postup ve STATISTICE: Vytvoříme nový datový soubor o 3 proměnných a 13 případech. První proměnnou nazveme X a uložíme do ní hodnoty 0, 1, ..., 12 (do Long Name napíšeme =v0-1). Druhou proměnnou nazveme DF a uložíme do ní hodnoty distribuční funkce (do Long Name napíšeme příkaz =IBinom(x;0,3;12)). Třetí proměnnou nazveme PF a uložíme do ní hodnoty pravděpodobnostní funkce (do Long Name napíšeme příkaz =Binom(x;0,3;12)).
Graf distribuční funkce: Graphs – Scatterplots – Variables X, DF – OK – vypneme Linear fit – OK – 2x klikneme na pozadí grafu – Plot:General – zaškrtneme Line – Line Type: Step – OK.
Graf pravděpodobnostní funkce: Graphs – Scatterplots – Variables X, PF – OK – vypneme Linear fit - OK.

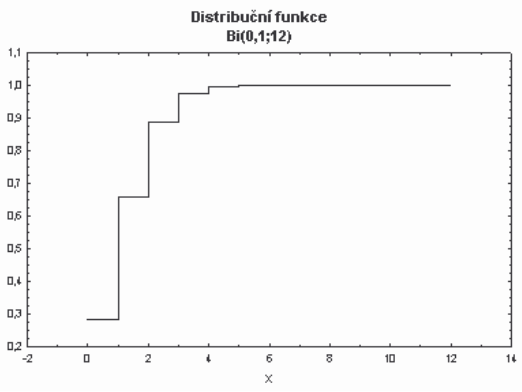
Podle tohoto návodu nakreslete grafy distribučních a pravděpodobnostních funkcí binomického rozložení pro různá n a p, např. n=5, p=0,5 (resp. 0,75) apod. Sledujte vliv parametrů na vzhled grafů.

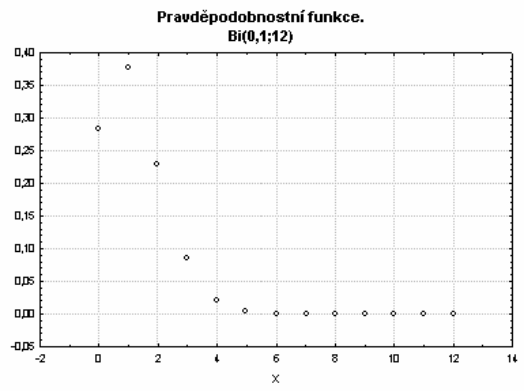
Řešení:

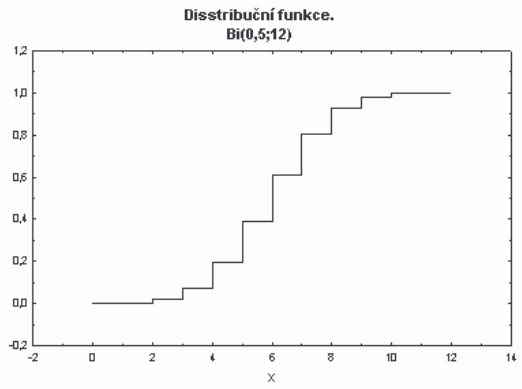


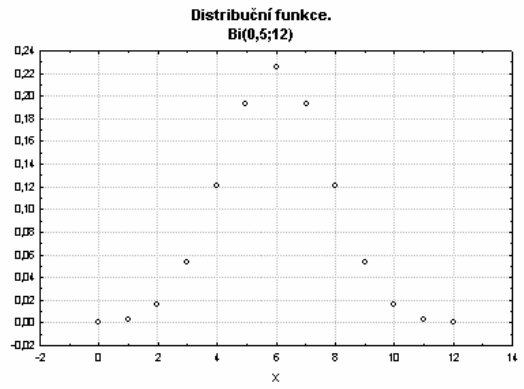


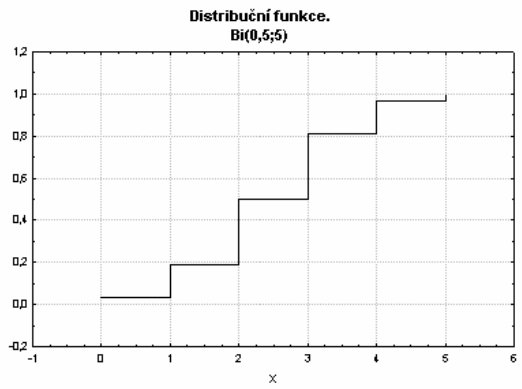


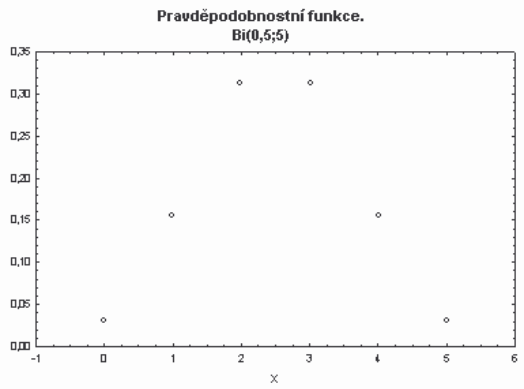

Distribuční funkce
Bi(0,1;12)
X

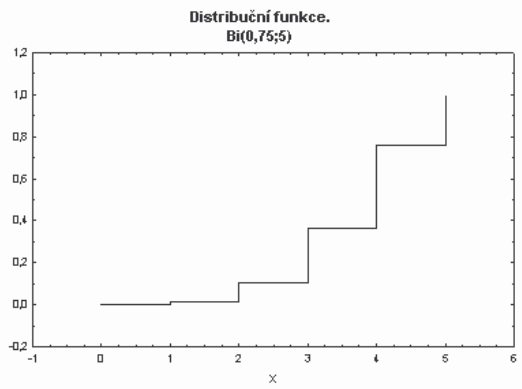

Pravděpodobnostní funkce.
Bi(0,1;12)
X

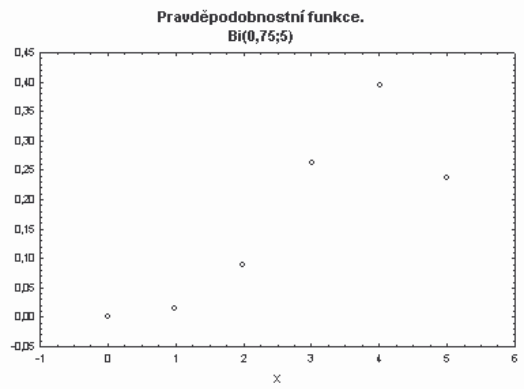

Disstribuční funkce.
Bi(0,5;12)
X

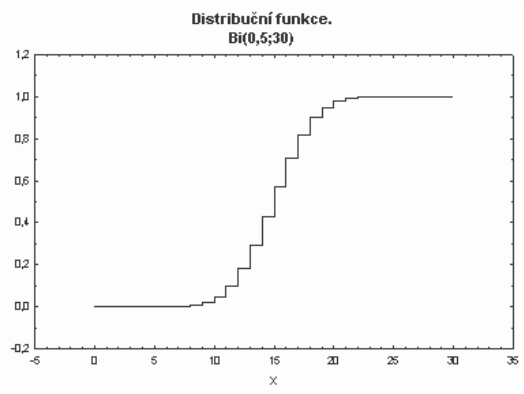

Distribuční funkce.
Bi(0,5;12)
X

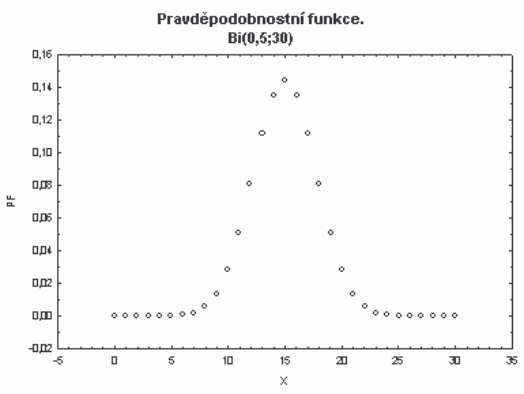

Distribuční funkce.
Bi(0,5;5)
X

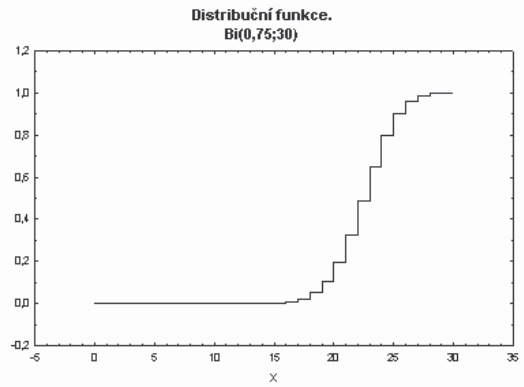

Pravděpodobnostní funkce.
Bi(0,5;5)
X

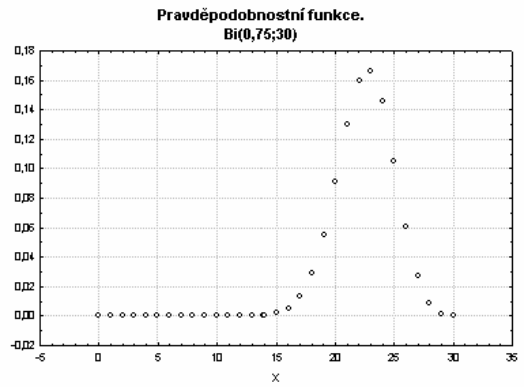

Distribuční funkce.
Bi(0,75;5)
X


Pravděpodobnostní funkce.
Bi(0,75;5)
X


Distribuční funkce.
Bi(0,5;30)
X


Pravděpodobnostní funkce.
Bi(0,5;30)
P F
X


Distribuční funkce.
Bi(0,75;30)
X


Pravděpodobnostní funkce.
Bi(0,75;30)
X

Téma 7: Výpočet střední hodnoty, rozptylu, kovariance a koeficientu korelace u diskrétních náhodných veličin.

Vzorový příklad 1. Postupně se zkouší spolehlivost čtyř přístrojů. Další se zkouší jen tehdy, když předchozí je spolehlivý. Každý z přístrojů vydrží zkoušku s pravděpodobností 0,8. Náhodná veličina X udává počet zkoušených přístrojů. Vypočtěte střední hodnotu a rozptyl náhodné veličiny X.

Řešení:

X nabývá hodnot 1, 2, 3, 4 a její pravděpodobnostní funkce je $\pi(1) = 0,2$, $\pi(2) = 0,8*0,2 = 0,16$, $\pi(3) = 0,8^2*0,2 = 0,128$, $\pi(4) = 0,8^3*0,2 + 0,8^4 = 0,512$, $\pi(0) = 0$ jinak
$E(X) = 1*0,2 + 2*0,16 + 3*0,128 + 4*0,512 = 2,952$
$D(X) = 1^2*0,2 + 2^2*0,16 + 3^2*0,128 + 4^2*0,512 - 2,952^2 = 1,4697$

Postup ve STATISTICE*:*
Otevřeme nový datový soubor o čtyřech případech a pěti proměnných, které nazveme x,pi(x),x*pi(x),xkvadrat,xkvadrat*pi(x). První proměnnou naplníme hodnotami náhodné veličiny X, druhou hodnotami její pravděpodobnostní funkce. Do třetí proměnné uložíme součin $x\pi(x)$ (do Long name napišeme =v1*v2), do čtvrté x^2 (do Long name napíšeme =v1^2), do páté součin $x^2\pi(x)$ (do Long name napíšeme v4*v2).

x	pi(x)	x*pi(x)	xkvadrat	xkvadrat*pi(x)
1	0,2	0,2	1	0,2
2	0,16	0,32	4	0,64
3	0,128	0,384	9	1,152
4	0,512	2,048	16	8,192

Výpočty E(X) a D(X) provedeme takto:

Statistics – Basic Statistics/Tables – Descriptive Statistics –Variables x*pi(x), xkvadrat*pi(x) – OK, zaškrtneme Sum - Summary

Proměnnou Sum ve workbooku transponujeme: Data – Transpose – File.
Proměnou x*pi(x) přejmenujeme na E(X) (vidíme, že E(X) = 2,952). Přidáme (ve workbooku) proměnnou D(X) a do jejího Long name napíšeme = v2-v1^2.
Vidíme, že D(X) = 1,4697.

Variable	Descriptiv Sum
x*pi(x)	2,95200
xkvadrat*pi(x)	10,18400

Vzorový příklad 2. 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, π(30,30) = 0,1, π(30,40) = 0, π(40,10) = 0, π(40,20) = 0, π(40,30) = 0, π(40,40) = 0,05, π(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. Stanovíme hodnoty marginálních pravděpodobnostních funkcí: $\pi_1(10) = 0,25$, $\pi_1(20)=0,55$, $\pi_1(30) = 0,15$, $\pi_1(40) = 0,05$, $\pi_1(x) = 0$ jinak, $\pi_2(10) = 0,3$, $\pi_2(20) = 0,45$, $\pi_2(30) = 0,2$, $\pi_2(10) = 0,05$, $\pi_2(y) = 0$ jinak. Spočteme E(X) = 20, E(Y) = 20, D(X) = 60, D(Y) = 70. Dosazením do vzorce pro výpočet kovariance zjistíme, že C(X,Y) = 49, tedy koeficient korelace $R(X,Y) = 49/\sqrt{60}\sqrt{70} = 0,76$.

Postup ve STATISTICE:
Budeme potřebovat dva nové soubory. První pro výpočet středních hodnot a rozptylů, druhý pro výpočet kovariance a koeficientu korelace.
První soubor bude mít 4 případy a 10 proměnných.
Zde jsou pro výpočet středních hodnot a rozptylů použity dva soubory vzhledem k přílišné délce tabulky pro obě náhodné veličiny.

x	pi(x)	x*pi(x)	xkvadrat	xkvadrat*pi(x)
10	0,25	2,5	100	25
20	0,55	11	400	220
30	0,15	4,5	900	135
40	0,05	2	1600	80

Variable	Descriptiv Sum
x*pi(x)	20,0000
xkvadrat*pi(x)	460,0000

y	pi(y)	y*pi(y)	ykvadrat	ykvadrat*pi(y)
10	0,3	3	100	30
20	0,45	9	400	180
30	0,2	6	900	180
40	0,05	2	1600	80

Variable	Descriptiv Sum
y*pi(y)	20,0000
ykvadrat*pi(y)	470,0000

Nyní vytvoříme nový datový soubor o 16 případech a 4 proměnných, které nazveme x,y,pi(x,y), a x*y*pi(x,y). Do první proměnné napíšeme 10, 10, 10, 10, 20, 20, 20, 20, 30, 30, 30, 30, 40, 40, 40, 40 a do druhé proměnné 10, 20, 30, 40, 10, 20, 30, 40, 10, 20, 30, 40, 10, 20, 30, 40.
Do třetí proměnné zapíšeme hodnoty simultánní pravděpodobnostní funkce π(x,y) a do čtvrté proměnné uložíme součin xyπ(x,y) (do Long name napíšeme =v1*v2*v3).

x	y	pi(x,y)	x*y*pi(x,y)
10	10	0,2	20
10	20	0,04	8
10	30	0,01	3
10	40	0	0
20	10	0,1	20
20	20	0,36	144
20	30	0,09	54
20	40	0	0
30	10	0	0
30	20	0,05	30
30	30	0,1	90
30	40	0	0
40	10	0	0
40	20	0	0
40	30	0	0
40	40	0,05	80

Statistics – Basic Statistics/Tables – Variables x*y*pi(x,y) – OK , zaškrtneme Sum – Summary.

Variable	Descriptiv Sum
x*y*pi(x,y)	449,0000

Proměnnou Sum ve workbooku přejmenujeme na E(X,Y) a přidáme k ní 6 nových proměnných E(X), E(Y), D(X), D(Y), C(X,Y), R(X,Y). Do proměnných E(X), E(Y), D(X), D(Y) napíšeme vypočtené střední hodnoty a rozptyly. Do Long name proměnné C(X,Y) napíšeme=v1-vv2*v3 a do Long name proměnné R(X,Y) napíšeme =v6/sqrt(v4*v5).

	E(X,Y)	E(X)	E(Y)	D(X)	D(Y)	C(X,Y)	R(X,Y)
x*y*pi(x,y)	449	20	20	60	70	49	0,756086

Vzorový příklad 3. Náhodná veličina X udává počet ok při hodu kostkou. Vypočtěte její střední hodnotu a rozptyl.

Řešení:

Náhodná veličina X nabývá hodnot 1, 2, 3, 4, 5, 6. Její pravděpodobnostní funkce je $\pi(1) = 1/6$, $\pi(2) = 1/6$, $\pi(3) = 1/6$, $\pi(4) = 1/6$, $\pi(5) = 1/6$, $\pi(6) = 1/6$, $\pi(x) = 0$ jinak
$E(X) = (1/6)(1+2+3+4+5+6) = 21/6 = 3,5$
$E(X^2) = (1/6)(1+4+9+16+25+36) = 91/6$
$D(X) = E(X^2) – [E(X)]^2 = 91/6 – 49/4 = 35/12$

Postup ve STATISTICE:
Otevřeme nový datový soubor o čtyřech případech a pěti proměnných, které nazveme x, pi(x), x*pi(x), xkvadrat, xkvadrat*pi(x). První proměnnou naplníme hodnotami náhodné veličiny X, druhou hodnotami její pravděpodobnostní funkce (do Long name napíšeme =1/6). Do třetí proměnné uložíme součin $x\pi(x)$ (do Long name napišeme =v1*v2), do čtvrté x^2 (do

Long name napíšeme =v1^2), do páté součin x^2 π(x) (do Long name napíšeme v4*v2).

x	pi(x)	x*pi(x)	xkvadrat	xkvadrat*pi(x)
1	0,166666667	0,166666667	1	0,166666667
2	0,166666667	0,333333333	4	0,666666667
3	0,166666667	0,5	9	1,5
4	0,166666667	0,666666667	16	2,66666667
5	0,166666667	0,833333333	25	4,16666667
6	0,166666667	1	36	6

Výpočty E(X) a D(X) provedeme takto:

Statistics – Basic Statistics/Tables – Descriptive Statistics –Variables x*pi(x), xkvadrat*pi(x) – OK, zaškrtneme Sum – Summary

Proměnnou Sum ve workbooku transponujeme: Data – Transpose – File.
Proměnou x*pi(x) přejmenujeme na E(X) (vidíme, že E(X) = 2,952). Přidáme (ve workbooku) proměnnou D(X) a do jejího Long name napíšeme = v2-v1^2.
Vidíme, že D(X) = 1,4697.

Variable	Descriptiv Sum
x*pi(x)	3,50000
xkvadrat*pi(x)	15,16667

Vzorový příklad 4. Diskrétní náhodný vektor (X_1,X_2) má simultánní pravděpodobnostní funkci s hodnotami π(0,-1) = c, π(0,0) = π(0,1) = π(1,-1) = π(2,-1) = 0, π(1,0) = π(1,1) = = π(2,1) = 2c, π(2,0) = 3c, π(x,y) = 0 jinak. Určete konstantu c a vypočtěte $R(X_1,X_2)$.

Řešení:

Náhodná veličina X_1 nabývá hodnot 0, 1, 2, náhodná veličina X_2 nabývá hodnot –1, 0, 1. Součet hodnot simultánní pravděpodobnostní funkce musí být roven jedné a odtud 10c = 1, tedy c = 0,1.
Stanovíme hodnoty marginálních pravděpodobnostních funkcí: $\pi_1(0) = 0,1$, $\pi_1(1) = 0,4$, $\pi_1(2) = 0,5$, $\pi_1(x) = 0$ jinak, $\pi_2(-1) = 0,1$, $\pi_2(0) = 0,5$, $\pi_2(1) = 0,4$, $\pi_2(y) = 0$ jinak.
Spočteme E(X) = 20, E(Y) = 20, D(X) = 60, D(Y) = 70. Dosazením do vzorce pro výpočet kovariance zjistíme, že C(X,Y) = 49, tedy koeficient korelace $R(X,Y) = 49/\sqrt{60}\sqrt{70} = 0,76$.

Postup ve STATISTICE:
Budeme potřebovat tři nové soubory. První dva pro výpočet středních hodnot a rozptylů, třetí pro výpočet kovariance a koeficientu korelace.
První dva soubory bude mít po 3 případy a 5 proměnných.

x1	pi(x1)	x1*pi(x1)	x1kvadrat	x1kvadrat*pi(x1)
0	0,1	0	0	0
1	0,4	0,4	1	0,4
2	0,5	1	4	2

Variable	Descriptiv Sum
x1*pi(x1)	1,400000
x1kvadrat*pi(x1)	2,400000

x2	pi(x2)	x2*pi(x2)	x2kvadrat	x2kvadrat*pi(x2)
-1	0,1	-0,1	1	0,1
0	0,5	0	0	0
1	0,4	0,4	1	0,4

Variable	Descriptiv Sum
x2*pi(x2)	0,300000
x2kvadrat*pi(x2)	0,500000

Nyní vytvoříme nový datový soubor o 9 případech a 4 proměnných, které nazveme x1, x2, pi(x1,x2), x1*x2*pi(x1,x2). Do první proměnné napíšeme 0, 0, 0, 1, 1, 1, 2, 2, 2 a do druhé proměnné –1, 0, 1, -1, 0, 1, -1, 0, 1. Do třetí proměnné zapíšeme hodnoty simultánní pravděpodobnostní funkce $\pi(x_1,x_2)$ a do čtvrté proměnné uložíme součin $x_1x_2\pi(x_1,x_2)$ (do Long name napíšeme =v1*v2*v3)).

x1	x2	pi(x1,x2)	x1*x2*pi(x1,x2)
0	-1	0,1	0
0	0	0	0
0	1	0	0
1	-1	0	0
1	0	0,2	0
1	1	0,2	0,2
2	-1	0	0
2	0	0,3	0
2	1	0,2	0,4

Statistics – Basic Statistics/Tables – Variables x1*x2*pi(x1,x2) – OK , zaškrtneme Sum – Summary.

Variable	Descriptiv Sum
x1*x2*pi(x1,x2)	0,600000

Proměnnou Sum ve workbooku přejmenujeme na E(X1*X2) a přidáme k ní 6 nových proměnných E(X1), E(X2), D(X1), D(X2), C(X1,X2), R(X1,X2). Do proměnných E(X1), E(X2), D(X1), D(X2) napíšeme vypočtené střední hodnoty a rozptyly. Do Long name proměnné C(X,Y) napíšeme =v1-v2*v3 a do Long name proměnné R(X,Y) napíšeme =v6/sqrt(v4*v5).

	E(X1*X2)	E(X1)	E(X2)	D(X1)	D(X2)	C(X1,X2)	R(X1,X2)
x1*x2*pi(x1,x2)	0,600000	1,4	0,3	0,44	0,41	0,18	0,42379

Téma 8: Ilustrace empirického zákona velkých čísel

Empirický zákon velkých čísel: Se vzrůstajícím počtem pokusů se relativní četnost úspěchu ustaluje kolem pravděpodobnosti úspěchu.

Modelová situace: Provádíme n nezávislých hodů mincí. Padnutí líce považujeme za úspěch. Tento pokus budeme simulovat pomocí programu STATISTICA a budeme sledovat závislost relativní četnosti úspěchu na počtu pokusů. (Počet pokusů volíme 2, 5, 10, 20, 50, 100, 200, 500,1000, 2000.)

Postup: Vygenerujeme n náhodných čísel mezi 0 a 1. Nabude-li náhodné číslo hodnotu z intervalu <0,5; 1>, pokus považujeme za úspěšný - tzn., že padl líc. Zjistíme relativní četnost úspěchu. Postup opakujeme pro různá n a nakonec znázorníme graficky závislost relativní četnosti úspěchu na počtu pokusů.

Návod: File – New – Number of variables 2, Number of cases 2000 – OK. 1. proměnnou přejmenujeme na NC, do Long Name napíšeme =Rnd(1), OK. (Funkce Rnd(1) vygeneruje náhodné číslo mezi 0 a 1.) 2. proměnnou přejmenujeme na POCET. Data – Recode - Category 1: Include If NC >=0,5, Category 2: Include If NC < 0,5, New Value 2, value 0, OK. (Proměnná POCET indikuje, zda nastal úspěch nebo neúspěch.) Vypočítáme průměr proměnné POCET (tj. relativní četnost úspěchu). Poznamenáme si počet pokusů n a relativní četnost úspěchu p. Nyní vymažeme posledních 1000 případů. Edit – Delete – Cases - From Case 1001 To Case 2000, OK. Znovu naplníme proměnné NC a POCET a spočteme průměr proměnné POCET. Postup opakujeme, až nám zbudou jen dva případy. Pak vytvoříme nový datový soubor o dvou proměnných n a p a 10 případech, kam zapíšeme hodnoty n a p. Nakonec nakreslíme dvourozměrný tečkový diagram závislosti p na n.

	Frequency table: POCET (Ezvc)			
Category	Count	Cumulative Count	Percent	Cumulative Percent
0	1005	1005	50,25000	50,2500
1	995	2000	49,75000	100,0000
Missing	0	2000	0,00000	100,0000

	Frequency table: POCET (Ezvc)			
Category	Count	Cumulative Count	Percent	Cumulative Percent
0	510	510	51,00000	51,0000
1	490	1000	49,00000	100,0000
Missing	0	1000	0,00000	100,0000

	Frequency table: POCET (Ezvc)			
Category	Count	Cumulative Count	Percent	Cumulative Percent
0	256	256	51,20000	51,2000
1	244	500	48,80000	100,0000
Missing	0	500	0,00000	100,0000

Category	Frequency table: POCET (Ezvc)			
	Count	Cumulative Count	Percent	Cumulative Percent
0	100	100	50,00000	50,0000
1	100	200	50,00000	100,0000
Missing	0	200	0,00000	100,0000

Category	Frequency table: POCET (Ezvc)			
	Count	Cumulative Count	Percent	Cumulative Percent
0	42	42	42,00000	42,0000
1	58	100	58,00000	100,0000
Missing	0	100	0,00000	100,0000

Category	Frequency table: POCET (Ezvc)			
	Count	Cumulative Count	Percent	Cumulative Percent
0	23	23	46,00000	46,0000
1	27	50	54,00000	100,0000
Missing	0	50	0,00000	100,0000

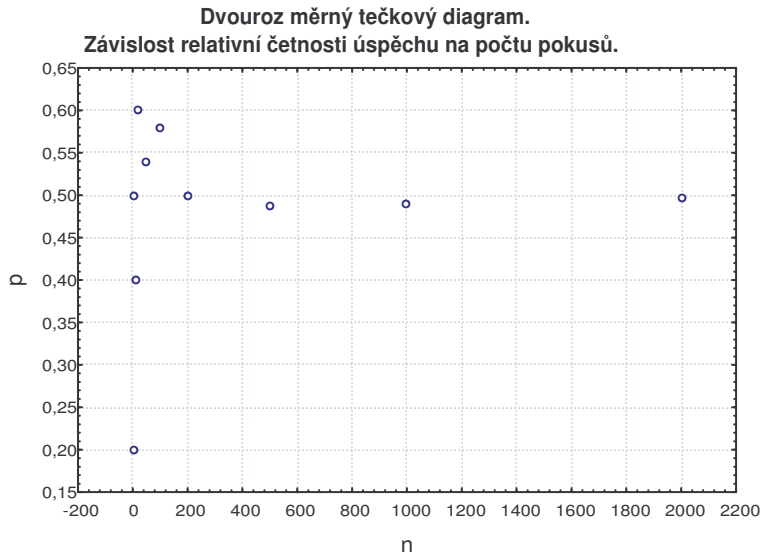
Category	Frequency table: POCET (Ezvc)			
	Count	Cumulative Count	Percent	Cumulative Percent
0	8	8	40,00000	40,0000
1	12	20	60,00000	100,0000
Missing	0	20	0,00000	100,0000

Category	Frequency table: POCET (Ezvc)			
	Count	Cumulative Count	Percent	Cumulative Percent
0	6	6	60,00000	60,0000
1	4	10	40,00000	100,0000
Missing	0	10	0,00000	100,0000

Category	Frequency table: POCET (Ezvc)			
	Count	Cumulative Count	Percent	Cumulative Percent
0	1	1	50,00000	50,0000
1	1	2	50,00000	100,0000
Missing	0	2	0,00000	100,0000

Category	Frequency table: POCET (Ezvc)			
	Count	Cumulative Count	Percent	Cumulative Percent
0	4	4	80,00000	80,0000
1	1	5	20,00000	100,0000
Missing	0	5	0,00000	100,0000

n	2000	1000	500	200	100	50	20	10	5	2
p	0,4975	0,4900	0,4880	0,5000	0,5800	0,5400	0,6000	0,4000	0,2000	0,5000


Dvouroz měrný tečkový diagram.
Závislost relativní četnosti úspěchu na počtu pokusů.
p
n
0,65
0,60
0,55
0,50
0,45
0,40
0,35
0,30
0,25
0,20
0,15
-200
0
200
400
600
800
1000
1200
1400
1600
1800
2000
2200

Téma 9: Centrální limitní věta

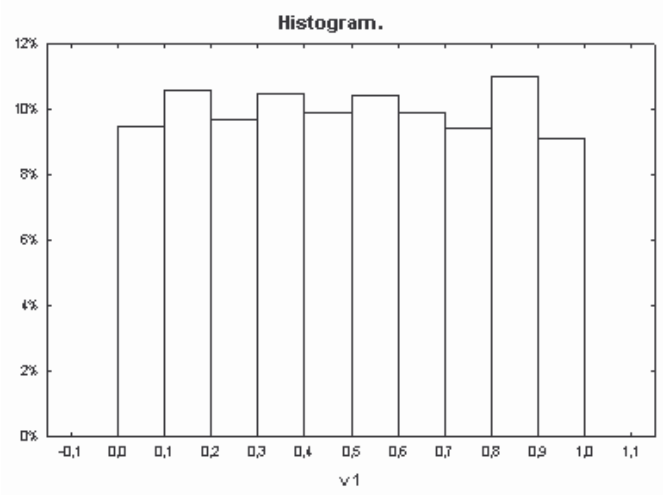
Ilustrace centrální limitní věty

Vygenerujeme 12 x 1000 realizací náhodných veličin X_1, ..., X_{12}, $X_i \sim Rs(0,1)$, i=1, ..., 12. Podle centrální limitní věty má náhodná veličina $X = X_1 + ... + X_{12} - 6$ přibližně rozložení N(0,1).
Návod: Vytvoříme nový datový soubor o 13 proměnných a 1000 případech. Otevřeme programovací okno STATISTICA VISUAL BASIC (File – New – Macro (SVB) Program – Name clv – OK) a do okna napíšeme příkazy:

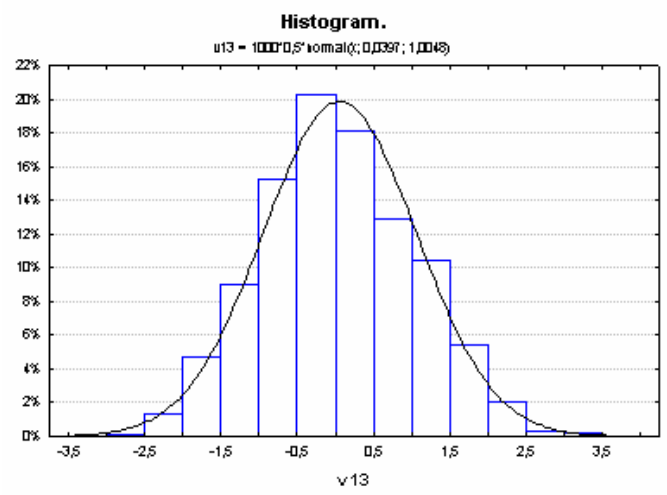
```
Dim s As Spreadsheet
Set s = ActiveSpreadsheet
For i = 1 To 12
  s.Variable(i).FillRandomValues
  ' do proměnných v1 až v12 se uloží náhodná čísla
  ' z intervalu(0,1)
Next i
s.VariableLongName(13) = "=Sum(v1:v12)-6"
' do proměnné v13 se uloží součet proměnných v1 až v12
' zmenšený o 6
s.Recalculate
```

Znázorníme histogramy proměnných v1 a v13 a porovnáme jejich vzhled s tvarem hustot rozložení Rs(0,1), N(0,1).
Dále spočteme průměry a rozptyly proměnných v1 a v13 a porovnáme je s teoretickou střední hodnotou a rozptylem náhodné veličiny s rozložením Rs(0,1) (E(X)=0,5, D(X)=1/12=0,833) a náhodné veličiny s rozložením N(0,1) (E(X)=0, D(X)=1).

Řešení:



Jedná se o 1000 náhodných čísel vygenerovaných z intervalu (0,1).
Jejich aritmetický průměr je m = 0,497491 a rozptyl $s^2 = 0,082374$.
Střední hodnota Rs(0,1) je E(X) = 0,5 a rozptyl D(X) = 1/12 = 0,08333.



Variable	Descriptive Statistics	
	Mean	Variance
v1	0,497471	0,082374
v13	0,039656	1,009721

Jedná se o náhodnou veličinu v13 = v1+v2+...+v12 – 6, která podle centrální limitní věty má rozložení N(0,1). (přesněji řečeno, posloupnost standardizovaných součtů konverguje v distribuci ke standardizované normální náhodné veličině.)
Aritmetický průměr v13 vyšel m = 0,039656, rozptyl s^2 = 1,009721.
Střední hodnota X ~N(0,1) je E(X) = 0, rozptyl D(X) = 1.

Aplikace Moivreovy - Laplaceovy integrální věty

Pomocí STATISTIKY spočteme př. 11.2. ze skript Teorie pravděpodobnosti a matematická statistika:
Y_{100} – počet úspěchů v posloupnosti n = 100 opakovaných nezávislých pokusů, pravděpodobnost úspěchu ϑ = 0,3, $Y_{100} \sim$ Bi(100, 0,3), $E(Y_{100}) = n\vartheta = 30$, $D(Y_{100}) = n\vartheta(1-\vartheta) = 21$.

Aproximativní výpočet:

$$P(20 \le Y_{100} \le 40) = P\left(\frac{19-30}{\sqrt{21}} \le \frac{Y_{100}-30}{\sqrt{21}} \le \frac{40-30}{\sqrt{21}}\right) \approx \Phi\left(\frac{10}{\sqrt{21}}\right) - \Phi\left(-\frac{11}{\sqrt{21}}\right) = 0{,}9773,$$

kde Φ(x) je distribuční funkce rozložení N(0,1).

Postup ve STATISTICE:
File – New – Number of variables 2, Number of cases 1 – OK.
Nastavíme se kurzorem na 1. sloupec.
Long Name =INormal(10/sqrt(21);0;1)- INormal(-11/sqrt(21);0;1) OK. (Funkce INormal(x;mu;sigma) poskytuje hodnotu distribuční funkce v bodě x normálního rozložení se střední hodnotou mu a směrodatnou odchylkou sigma.)

Přesný výpočet:

$$P(20 \le Y_{100} \le 40) = P(19 < Y_{100} \le 40) = \Phi(40) - \Phi(19) = 0{,}978614,$$

kde Φ(x) je distribuční funkce rozložení Bi(100, 0,3).

Postup ve STATISTICE:
Nastavíme se kurzorem na 2. sloupec.
Long Name =IBinom(40;0,3;100)- IBinom(19;0.3;100). (Funkce IBinom(x;p;n) poskytuje hodnotu distribuční funkce v bodě x binomického rozložení s parametry p a n.)

Podle tohoto návodu vyřešte příklady 11.3., 11.5., 11.6.

Př. 11.3.
n = 400, ϑ = 0,2, úspěch je nutnost opravy v záruční době
$n\vartheta$ = 80, $n\vartheta(1-\vartheta)$ = 64
aproximativní výpočet: $P(Y_{400} > 96) \approx$ 1 - INormal(16/8;0;1) = 0,022750
přesný výpočet: $P(Y_{400} > 96) =$ 1 – IBinom(96;0,2;400) = 0,024640

Př. 11.5.
n = 10000, ϑ = 0,515, úspěch je narození chlapce
$n\vartheta$ = 5150, $n\vartheta(1-\vartheta)$ = 2497,75
Úkol (a)
aproximativní výpočet: $P(Y_{10000} \leq 5000) \approx$ INormal(-150/sqrt(2497,75);0;1) = 0,001344
přesný výpočet: $P(Y_{10000} \leq 5000)$ = IBinom(5000;0,515;10000) = 0,001347
Úkol (b)
aproximativní výpočet: $P(4999 < Y_{10000} \leq 5300) \approx$ INormal(150/sqrt(2497,75);0;1) - INormal(-151/sqrt(2497,75);0;1) = 0,997399
přesný výpočet: $P(4999 < Y_{10000} \leq 5300) =$ IBinom(5300;0;1) – IBinom(4999;0;1) = 0,997400

Př. 11.6.
n = 1000, ϑ = 0,05, úspěch je zhotovení vadného výrobku
$n\vartheta$ = 50, $n\vartheta(1-\vartheta)$ = 47,5
aproximativní výpočet: $P(Y_{1000} \leq 70) \approx$ INormal(20/sqrt(47,5);0;1) = 0,998145
přesný výpočet: $P(Y_{1000} \leq 70)$ = IBinom(70;0,05;1000) = 0,997670