

UNIVERZITA OBRANY
Fakulta ekonomiky a managementu

Statistické zpracování dat v aplikaci STAT1

Výuková pomůcka pro předmět Statistika

Jiří Neubauer, Marek Sedláčik

3. 11. 2012

Obsah

Popis STAT1	1
Použití	2
Vložení dat	2
Popisné charakteristiky	3
Bodové rozdělení četností	4
Intervalové rozdělení četností	6
Bodové a intervalové odhady	8
Odhady parametrů normálního rozdělení	8
Odhady střední hodnoty pro výběry velkého rozsahu	10
Odhady parametru alternativního rozdělení	10
Testy statistických hypotéz	12
Jednovýběrové testy	12
Dvouvýběrové testy	12
Testy normality	13
Chí-kvadrát test nezávislosti v kontingenční tabulce	16
Statistické tabulky	17

Popis STAT1

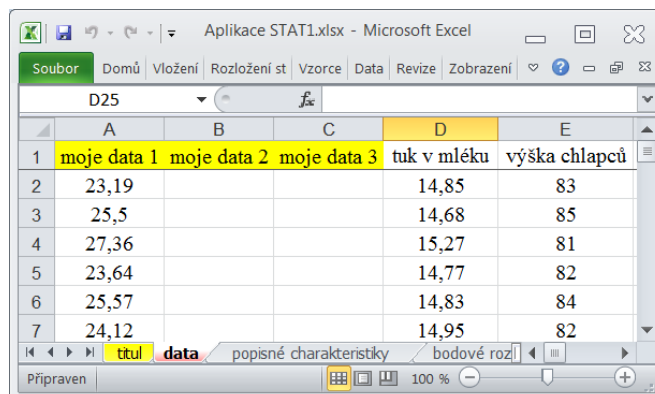
Aplikace STAT1 pracuje pod Microsoft Office Excel a je určena pro základní zpracování dat prostřednictvím exploratorní analýzy dat, metod jednorozměrné induktivní statistiky, dále jsou zde implementovány dvouvýběrové testy a chí-kvadrát test nezávislosti v kontingenční tabulce. Aplikace poskytuje řadu užitečných výstupů v podobě tabulek, grafů a statistických závěrů.

Uživatel může pomocí tohoto nástroje zpracovávat vlastní datové soubory, případně lze využít již vložených dat. Ovládání se provádí pomocí nabízených menu nebo pomocí parametrů, které jsou označeny červeně. Oporu lze najít rovněž v knize Neubauer, J., Sedlačík, M. a O. Kříž. *Základy statistiky: Aplikace v technických a ekonomických oborech*. Praha: Grada, 2012. ISBN 978-80-247-4273-1.

Použití

Vložení dat

Pro vložení vlastního datového souboru přejděte na list „data“. Do prvních třech sloupců označených „moje data 1“, „moje data 2“ a „moje data 3“ vložte data.



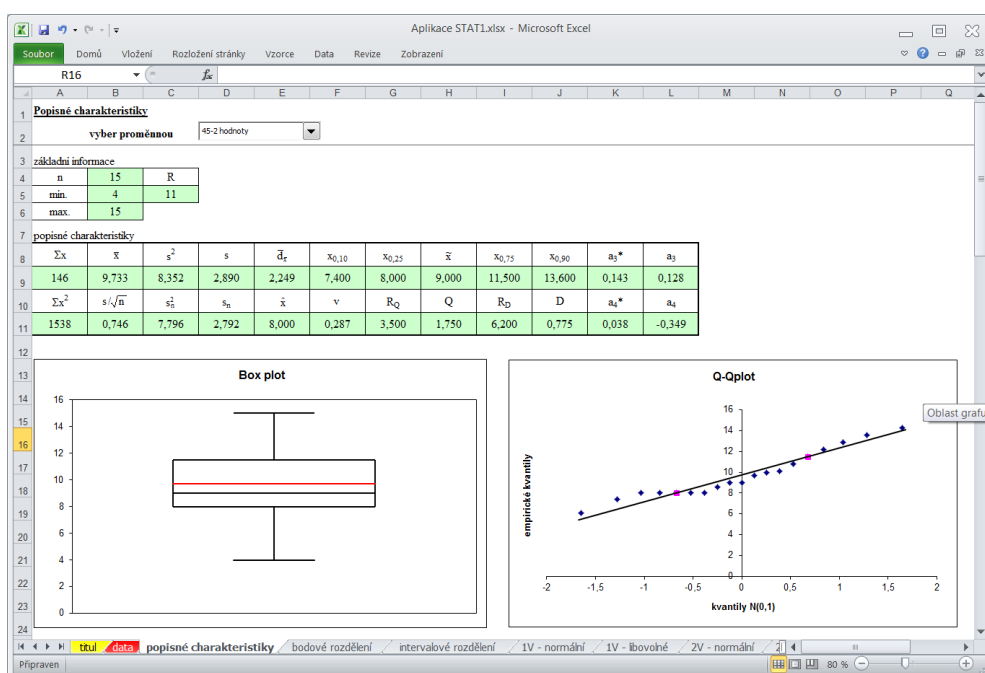
The screenshot shows a Microsoft Excel spreadsheet titled 'Aplikace STAT1.xlsx'. The active sheet is 'data'. The data is organized in a table with 5 columns and 7 rows. The first three columns are labeled 'moje data 1', 'moje data 2', and 'moje data 3'. The fourth column is labeled 'tuk v mléku' and the fifth column is labeled 'výška chlapců'. The data values are as follows:

	A	B	C	D	E
1	moje data 1	moje data 2	moje data 3	tuk v mléku	výška chlapců
2	23,19			14,85	83
3	25,5			14,68	85
4	27,36			15,27	81
5	23,64			14,77	82
6	25,57			14,83	84
7	24,12			14,95	82

Data obsažená v knize Základy statistiky jsou uvedena v daném listu v pořadí, v jakém se v knize objevují.

Popisné charakteristiky

List „popisné charakteristiky“ nabízí výpočet vybraných číselných charakteristik datového souboru. Z nabízeného menu vyberte datový soubor, který máte v úmyslu zpracovávat. (Název datového souboru odpovídá názvu uvedenému v prvním řádku v listu „data“). Číselné charakteristiky v tabulkové podobě se spočítají automaticky.



Kromě těchto charakteristik lze na listu nalézt dva grafy: krabicový diagram (boxplot) a Q-Q plot. Krabicový diagram zachycuje minimální a maximální hodnotu datového souboru, dolní kvartil, medián, horní kvartil a aritmetický průměr (červená linka). Q-Q plot porovnává teoretické kvantily normovaného rozdělení $N(0,1)$ s empirickými kvantily určených z dat. Dále jsou spočteny testy normality založené na koeficientech šikmosti a špičatosti – viz **testy hypotéz**.

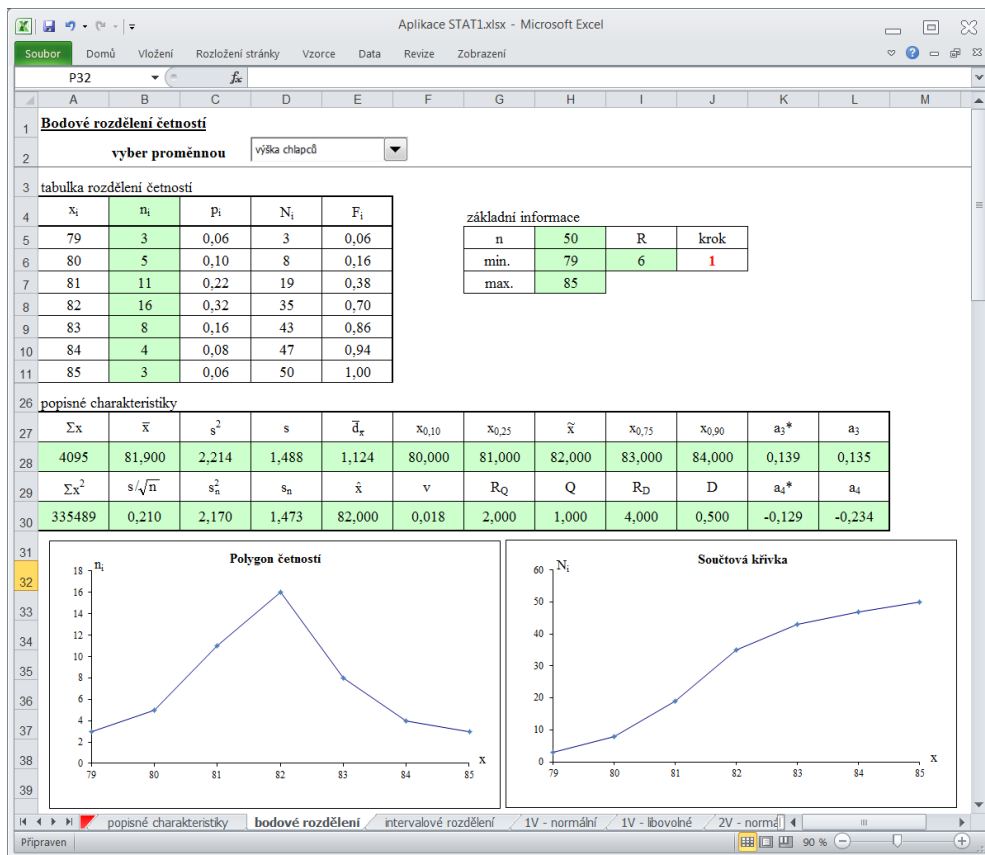
Bodové rozdělení četností

Pro vytvoření tabulky bodového rozdělení četností a grafů popisující toto rozdělení přejděte na list „bodové rozdělení“. Z nabízeného menu vyberte datový soubor, který máte v úmyslu zpracovávat. (Název datového souboru odpovídá názvu uvedenému v prvním řádku v listu „data“).

x_i	n_i	p_i	F_i
79	3	0,06	0,16
80	5	0,10	0,16
81	11	0,22	0,38
82	16	0,32	0,70
83	8	0,16	0,86

základní informace			
n	50	R	krok
min.	79	6	1
max.	85		

Tabulka rozdělení četností se vytvoří automaticky s krokem uvedeným v políčku „krok“, nastavenou hodnotu „1“ lze měnit v závislosti na vlastnostech datového souboru. Spolu s tabulkou se vypočítají základní popisné charakteristiky a vytvoří se dva grafy: polygon četností a součtová křivka. Řádky s nulovými četnostmi je možné skrýt (tyto hodnoty se nebudou objevovat v grafech).



Dále jsou spočteny testy normality založené na koeficientech šikmosti a špičatosti – viz **testy hypotéz**.

Intervalové rozdělení četností

Pro vytvoření tabulky intervalového rozdělení četností a grafů popisující toto rozdělení přejděte na list „intervalové rozdělení“. Z nabízeného menu vyberte datový soubor, který máte v úmyslu zpracovávat. (Název datového souboru odpovídá názvu uvedenému v prvním řádku v listu „data“). Ke správnému vytvoření tabulky rozdělení četností je potřeba zadat následující tři parametry (označené červeně): k ... plánovaný počet tříd (řádků) v tabulce, h ... šířka třídy (intervalu), a ... počáteční hodnota, od které se začne tabulka vytvářet. Jako pomůcka pro určení optimálního počtu tříd jsou zde uvedena dvě pravidla. Konkrétní volba potom závisí na zpracovateli.

The screenshot shows the 'intervalové rozdělení' worksheet in Microsoft Excel. It features a menu bar, a dropdown for selecting the variable 'prach ve vzduchu', and a table of distribution with columns for class boundaries, midpoints, frequencies, relative frequencies, cumulative frequencies, and cumulative relative frequencies. To the right, there are summary statistics and parameters for class determination.

x_d	x_h	x_i	n_i	p_i	N_i	F_i
1	1,1	1,05	7	0,12	7	0,12
6	1,1	1,2	8	0,13	15	0,25
7	1,2	1,3	11	0,18	26	0,43
8	1,3	1,4	14	0,23	40	0,67
9	1,4	1,5	9	0,15	49	0,82
10	1,5	1,6	9	0,15	58	0,97
11	1,6	1,7	2	0,03	60	1,00
12	1,7	1,8	0	0,00	60	1,00
13	1,8	1,9	0	0,00	60	1,00
14	1,9	2	0	0,00	60	1,00

základní informace

n	60	R
min.	1,01	0,62
max.	1,63	

stanovení počtu tříd

$1 + 3,32 \log n$	6,903	k	7
$5 \log n$	8,891	h	0,1

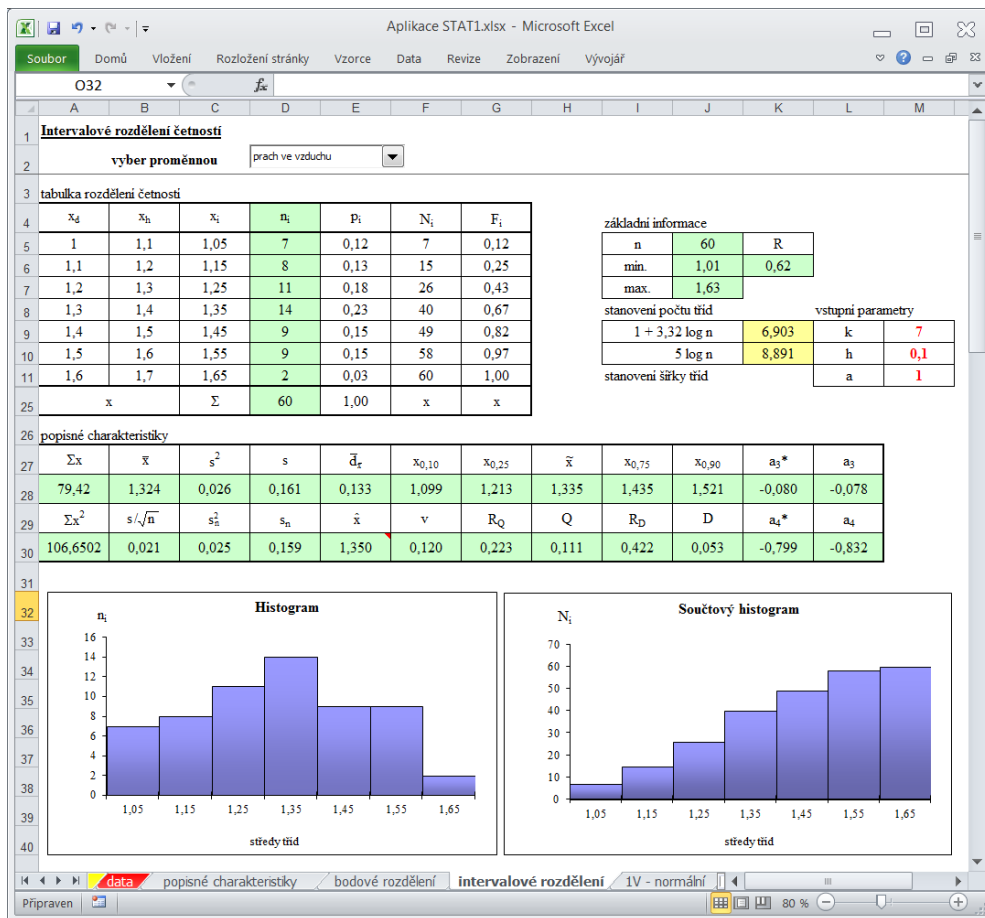
stanovení šířky tříd

h	0,089
---	-------

vstupní parametry

a	1
---	---

Spolu s tabulkou se vypočítají základní popisné charakteristiky a vytvoří se dva grafy: histogram a součtový histogram. Řádky s nulovými četnostmi je možné skrýt (tyto hodnoty se nebudou objevovat v grafech).



Dále jsou spočteny testy normality založené na koeficientech šikmosti a špičatosti – viz **testy hypotéz**.

Bodové a intervalové odhady

Odhady parametrů normálního rozdělení

Bodové a intervalové odhady střední hodnoty a rozptylu (příp. směrodatné odchyly) získáme přepnutím na list „1V – normální“. Poté, co vybereme analyzovaný datový soubor a zadáme riziko odhadu α (implicitně nastaveno na hodnotu 0,05), bodové odhady a intervaly spolehlivosti pro střední hodnotu, rozptyl i směrodatnou odchytku (oboustranný, dolní i horní) se vypočítají.

Výběr z normálního rozdělení

vyber proměnnou: benzín

n	\bar{x}	s	s^2
20	6,180	0,399	0,160

α
0,05

1. Bodové odhady parametrů

$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{\sigma}$	estSE
6,180	0,160	0,399	0,089

2. Velikost výběru

připustná chyba Δ	p-st	min.n
0,2	95%	18

3. Intervalové odhady pro střední hodnotu

p-st	oboustranný	Δ	dolní	horní	Δ	
95%	5,993	6,367	0,187	6,026	6,334	0,154

kvantily $t_p(v)$

v	$t_{1-\alpha}(v)$	$t_{1-\alpha/2}(v)$
19	1,729	2,093

5. Intervalové odhady pro rozptyl

p-st	oboustranný	dolní	horní	
95%	0,092	0,340	0,101	0,300

kvantily $\chi_p^2(v)$

v	$\chi_{1-\alpha/2}^2(v)$	$\chi_{\alpha}^2(v)$	$\chi_{1-\alpha}^2(v)$	$\chi_{1-\alpha/2}^2(v)$
19	8,91	10,12	30,14	32,85

6. Intervalové odhady pro směrodatnou odchytku

p-st	oboustranný	dolní	horní	
95%	0,304	0,583	0,317	0,547

Odhady parametrů lze také získat přímým zadáním číselných charakteristik (rozsahu, aritmetického průměru a výběrové směrodatné odchyly).

Aplikace STAT1.xlsx - Microsoft Excel

Soubor Domů Vložení Rozložení stránky Vzorce Data Revize Zobrazení Vývojář

N25

Výběr z normálního rozdělení

vyber proměnnou benzín

Zadejte hodnoty charakteristik:

n	\bar{x}	s	s^2	α	Pomůcka:
15	42,460	0,863	0,745	0,05	rozpyl 0,745 odchylka 0,863134

1. Bodové odhady parametrů

$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{\sigma}$	estSE
42,460	0,745	0,863	0,223

2. Velikost výběru

připustná chyba Δ	p-st	min.n
0,2	95%	86

3. Intervalové odhady pro střední hodnotu

p-st	oboustranný	Δ	dolní	horní	Δ
95%	41,982	42,938	0,478	42,067	42,853

kvantily $t_p(v)$

v	$t_{1-\alpha}(v)$	$t_{1-\alpha/2}(v)$
14	1,761	2,145

1V - normální

Odhady střední hodnoty pro výběry velkého rozsahu

Bodové a intervalové odhady střední hodnoty získáme přepnutím na list „1V – libovolné“. Ovládání je obdobné jako u odhadů parametrů normálního rozdělení.

Odhady parametru alternativního rozdělení

Bodové a intervalové odhady parametru alternativního rozdělení získáme přepnutím na list „1V a 2V – podíl“. Zde je nutné zadat vstupní n a m , kde podíl m/n je bodovým odhadem parametru π alternativního rozdělení.

Aplikace STAT1.xlsx - Microsoft Excel

Soubor Domů Vložení Rozložení stránky Vzorce Data Revize Zobrazení Vývojář

G31

Odhady a testy pro podíl jednotek v populaci - výběry z $A(\pi)$

Jeden výběr z $A(\pi)$ - výpočty z charakteristik

Zadejte rozsah a hodnotu četnosti:

n	m	p	$np(1-p)$	α
250	20	0,080	18,400	0,05

1. Bodový odhad podílu

$\hat{\pi}$	estSE(p)
0,080	0,017

2. Velikost výběru

připustná chyba Δ	p-st	min.n
0,03	95%	315

3. Intervalové odhady pro podíl

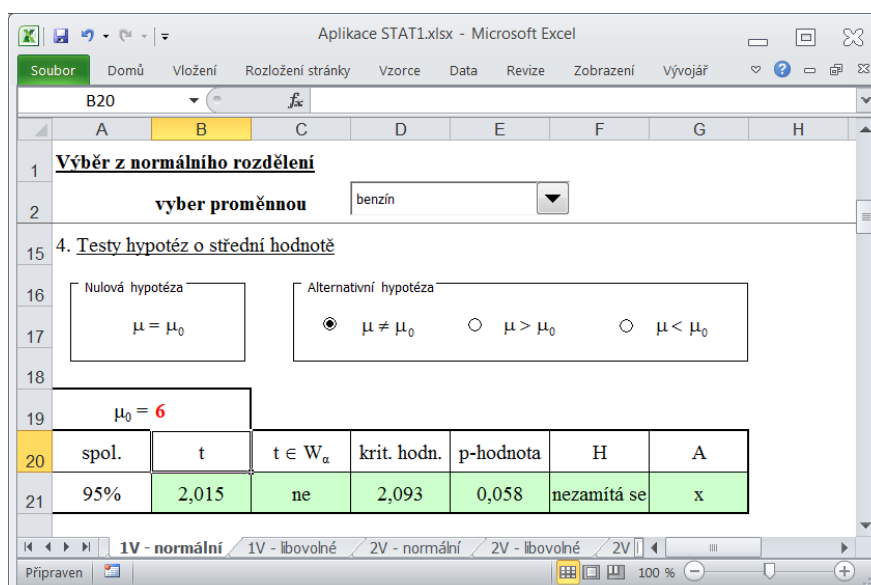
p-st	oboustranný	Δ	dolní	horní	Δ
95%	0,046	0,114	0,034	0,052	0,108

1V a 2V - podíl

Testy statistických hypotéz

Jednovýběrové testy

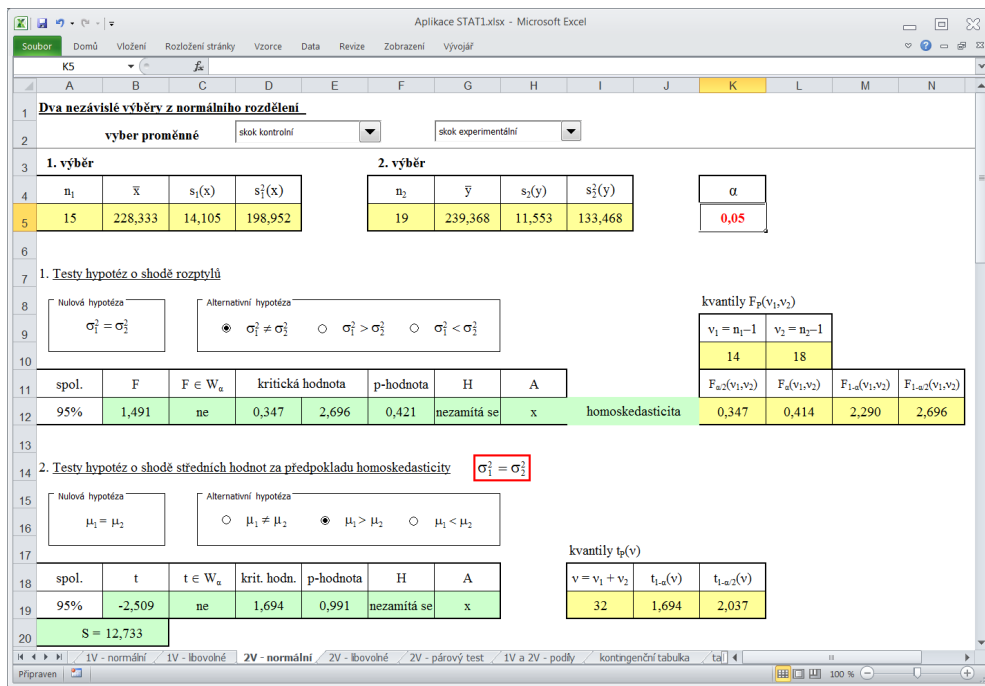
Aplikace STAT1 obsahuje tyto jednovýběrové testy hypotéz: test střední hodnoty a rozptylu normálního rozdělení (list „1V – normální“), test střední hodnoty pro velké výběry (list „1V – libovolné“) a test parametru alternativního rozdělení pro velké výběry (list 1V a 2V – podíly). Testování se ve všech případech provádí podobně, zaměříme se na jeden konkrétní – test střední hodnoty normálního rozdělení. Přejdeme na list „1V – normální“ a vybereme datový soubor. Zvolíme hladinu významnosti α (implicitně nastaveno na hodnotu 0,05), zadáme nulovou hypotézu H a vybereme jednu ze tří nabízených alternativních hypotéz A.



Jako výstup obdržíme hodnotu testového kritéria, kritickou hodnotu, p-hodnotu a slovní odpověď (H se nezamítá, nebo H se zamítá A se přijímá). Testy je možné také počítat zadáním číselných charakteristik (v dolní části listu).

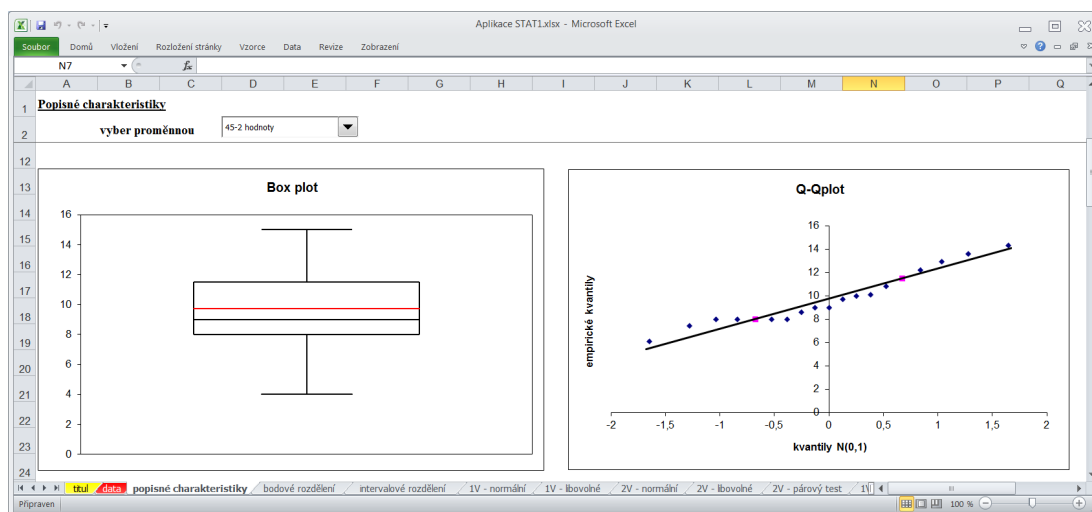
Dvouvýběrové testy

Aplikace STAT1 obsahuje tyto dvouvýběrové testy hypotéz: test shody dvou rozptylů nezávislých normálních rozdělení (list „2V – normální“), test shody dvou středních hodnot nezávislých normálních rozdělení (za předpokladu homoskedasticity a heteroskedasticity – list „2V – normální“), test shody dvou středních hodnot pro velké nezávislé výběry (list „2V – libovolné“), test shody dvou středních hodnot pro závislé výběry (párový test – list „2V – párový test“) a test shody dvou parametrů alternativního rozdělení pro velké nezávislé výběry (list 1V a 2V – podíly). Testování se provádí podobně jako u jednovýběrových testů, zde je třeba vybrat dva datové soubory.



Testy normality

Základní představu o tvaru rozdělení datového souboru můžeme získat konstrukcí histogramu, případně polygonu četností (viz intervalové a bodové rozdělení četností). V listu „popisné charakteristiky“ lze nalézt kromě krabicového diagramu i Q-Q plot porovnávající teoretické kvantily normovaného rozdělení $N(0,1)$ s empirickými kvantily určených z dat. Leží-li tyto body přibližně na přímce, můžeme usoudit, že zkoumaný náhodný výběr pochází z normálního rozdělení.



Listy „popisné charakteristiky“, „bodové rozdělení“ a „intervalové rozdělení“ obsahují v dolní části testy normality založené na výběrových koeficientech šik-

mosti a špičatosti.

Aplikace STAT1.xlsx - Microsoft Excel

Soubor Domů Vložení Rozložení stránky Vzorce Data Revize Zobrazení

N7

1 **Popisné charakteristiky**

2 **vyber proměnnou** 45-2 hodnoty

26 test o nulové šikmosti

a_3	$D(a_3)$	u_3	$u_{1-a/2}$	p-hodnota	
0,128	0,271	0,246	1,960	0,806	→ nulová šikmost se nezamítá

29 test o nulové špičatosti

a_4	$D(a_4)$	u_4	$u_{1-a/2}$	p-hodnota	
-0,349	0,609	0,033	1,960	0,973	→ nulová špičatost se nezamítá

32 kombinovaný test o šikmosti a špičatosti: C-test

u_3	u_4	C	$\chi^2_{1-\alpha}(2)$	p-hodnota	
0,246	0,033	0,062	5,991	0,970	→ normální rozdělení se nezamítá

36 modifikovaný test o nulové šikmosti

b	W^2	δ	a	z_3	$u_{1-a/2}$	p-hodnota	
3,524	1,247	3,010	2,846	0,260	1,960	0,795	→ nulová šikmost se nezamítá

39 modifikovaný test o nulové špičatosti

B	A	z_4	$u_{1-a/2}$	p-hodnota	
1,660	19,359	0,262	1,960	0,793	→ nulová špičatost se nezamítá

42 kombinovaný test o šikmosti a špičatosti: modifikovaný C-test

z_3	z_4	C'	$\chi^2_{1-\alpha}(2)$	p-hodnota	
0,260	0,262	0,136	5,991	0,934	→ normální rozdělení se nezamítá

ttul data popisné charakteristiky bodové rozdělení intervalové rozdělení 1V - normální

Připraven Přepočít

Výpočet těchto testů zadáním potřebných charakteristik (rozsah, koeficient šikmosti a špičatosti) lze provést v dolní části listu „popisné charakteristiky“.

Chí-kvadrát test nezávislosti v kontingenční tabulce

List „kontingenční tabulka“ je určen pro testování nezávislosti v kontingenční tabulce užitím tzv. chí-kvadrát testu nezávislosti dvou statistických znaků. Tento test patří mezi neparametrické metody, to znamená, že nevyžaduje znalost rozdělení zkoumaných statistických proměnných. Při chí-kvadrát testu nezávislosti testujeme nulovou hypotézu H_0 , že sledované znaky jsou nezávislé, proti alternativní hypotéze H_1 , která je naopak hypotézou o jejich závislosti. Uživatel doplní poze hodnoty absolutních četností do připravené kontingenční tabulky a zvolí hladinu významnosti testu α (implicitně nastaveno na hodnotu 0,05).

Applikace STAT1.xlsx - Microsoft Excel

Soubor Domů Vložení Rozložení stránky Vzorce Data Revize Zobrazení Vývojář

P8 0,05

1 χ^2 -test nezávislosti v kontingenční tabulce

2

3 Empirické četnosti

	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	$n_{i\cdot}$
4 x_1	13	17	10	11							51
5 x_2	36	14	9	2							61
6 x_3											0
7 x_4											0
8 x_5											0
9 x_6											0
10 x_7											0
11 x_8											0
12 x_9											0
13 x_{10}											0
14 $n_{\cdot j}$	49	31	19	13	0	0	0	0	0	0	112

15

16

Nulová hypotéza: Znaky X a Y jsou nezávislé
Alternativní hypotéza: Znaky X a Y jsou závislé

spolehlivost	χ^2	$\chi^2 \in W_\alpha$	st. volnosti	krit. hodnota	p-hodnota	H	A
95%	16,609	ano	3	7,815	0,001	zamítá se	se přijme

hladina významnosti 0,05

Závislost v tabulce je statisticky významná

1V - normální 1V - libovolné 2V - normální 2V - libovolné 2V - párový test 1V a 2V - podly kontingenční tabulka tabulky

Připraven

Statistické tabulky

Poslední list „tabulky“ obsahuje hodnoty pravděpodobnostních a distribučních funkcí Poissonova, binomického a hypergeometrického rozdělení dále funkce hustoty pravděpodobnosti, distribuční funkce a kvantily rozdělení rovnoměrného, exponenciálního, normálního a log-normálního (u verze pro MS Office 2003 a 2007 - STAT.xls - chybí distribuční funkce hypergeometrického rozdělení a funkce hustoty pravděpodobnosti log-normálního rozdělení, která nejsou dispozici). Jsou zde uvedeny i kvantily Pearsonova, Studentova a Fisher-Snedecorova rozdělení.

The screenshot shows a Microsoft Excel spreadsheet titled "Applikace STAT.xls - Microsoft Excel". The spreadsheet is organized into sections for different types of probability distributions. Each section contains a table with parameters and calculated values. The values are color-coded: red for input parameters and green for calculated results.

Statistické tabulky

Diskrétní modely

Poissonovo rozdělení

λ	2	$p(x)$	$F(x)$
x	3	0,18045	0,85712

binomické rozdělení

n	5	$p(x)$	$F(x)$
π	0,51	0,31850	0,79975
x	3		

hypergeometrické rozdělení

N	100	$p(x)$	$F(x)$
M	4	0,17649	0,98837
n	5		
x	1		

Spojitě modely

rovnorné rozdělení

α	0	$f(x)$	$F(x)$
β <td>7</td> <td>0,14286</td> <td>0,71429</td>	7	0,14286	0,71429
x	5	kvantil x_p	
p	0,95	6,65000	

exponenciální rozdělení

α	0	$f(x)$	$F(x)$
δ <td>5</td> <td>0,13406</td> <td>0,32968</td>	5	0,13406	0,32968
x	2	kvantil x_p	
p	0,95	14,97866	

normální rozdělení

μ	50	$f(x)$	$F(x)$
σ^2 <td>25</td> <td>0,01579</td> <td>0,03593</td>	25	0,01579	0,03593
x	41	kvantil x_p	
p	0,3	47,37800	

logaritnicko-normální rozdělení

μ	2	$f(x)$	$F(x)$
σ^2 <td>0,49</td> <td>0,09702</td> <td>0,19032</td>	0,49	0,09702	0,19032
x	4	kvantil x_p	
p	0,3	5,11880	

kvantily Pearsonova rozdělení

v	12	kvantil x_p
p	0,95	21,02607

kvantily Studentova rozdělení

v	20	kvantil x_p
p	0,975	2,08596

kvantily Fisher-Snedecorova rozdělení

v_1	6	kvantil x_p
v_2 <td>16</td> <td>2,74131</td>	16	2,74131
p	0,95	

Použité zdroje

- Anděl, J. *Základy matematické statistiky*. 1. vyd. Praha: Matfyzpress, 2005. ISBN 80-86732-40-1.
- Budíková, M., M. Králová a B. MAROŠ. *Průvodce základními statistickými metodami*. 1. vyd. Praha: Grada, 2010. ISBN 978-80-247-3243-5.
- Chajdiak, J. *Štatistika v Exceli 2007*. 1. vyd. Bratislava: Statis, 2009. ISBN 978-80-85659-49-8.
- Neubauer, J., Sedlačík, M. a O. Kříž *Základy statistiky: Aplikace v technických a ekonomických oborech*. Praha: Grada, 2012. ISBN 978-80-247-4273-1.
- Schels, I. *Excel 2007 – vzorce a funkce*. 1. vyd. Praha: Grada, 2008. ISBN 978-80-247-2074-6.