Téma 6: Parametrické úlohy o jednom náhodném výběru

Úkol 1.: Vlastnosti výběrového průměru z normálního rozložení

Předpokládejme, že velký ročník na vysoké škole má výsledky ze statistiky normálně rozloženy kolem střední hodnoty 72 bodů se směrodatnou odchylkou 9 bodů. Najděte pravděpodobnost, že průměr výsledků náhodného výběru 10 studentů bude větší než 80 bodů.

Návod:

$X_1, ..., X_{10}$ je náhodný výběr z N(72, 81). Počítáme P(M > 80), přičemž výběrový průměr M má normální rozložení se střední hodnotou E(M) = μ = 72 a rozptylem D(M) $= \frac{\sigma^2}{n} = \frac{81}{10} =$ 8,1 (viz skripta Základní statistické metody, věta 6.1.1.1., bod 2).

Tedy $P(M > 80) = 1 - P(M \le 80) = 1 - \Phi(80)$, kde $\Phi(80)$ je hodnota distribuční funkce rozložení N(72; 8,1) v bodě 80.

Vytvoříme datový soubor o jedné proměnné a o jednom případu. Do Dlouhého jména této proměnné napíšeme =1 – INormal(80;72;sqrt(8,1)). Zjistíme, že $1 - \Phi(80) = 0{,}00247005$.

Funkce INormal(x;μ;σ) počítá hodnotu distribuční funkce rozložení $N(\mu,\sigma^2)$ v bodě x.

	1 Prom1
1	0,00247

Úkol 2.: Vlastnosti výběrového průměru z alternativního rozložení

Mezi americkými voliči 60% osob volí republikány a 40% demokraty. Jaká je pravděpodobnost, že v náhodném výběru 100 amerických voličů budou voliči republikánů v menšině? Výpočet proveďte jak přesně, tak pomocí aproximace normálním rozložením.

Návod:

$X_1, ..., X_{100}$ je náhodný výběr z A(0,6), $X_i = 1$, když i-tá osoba volí republikány, $X_i = 0$ jinak, i = 1, ..., 100. Zavedeme statistiku $Y_{100} = X_1 + ... + X_{100}$, $Y_{100} \sim Bi(100; 0{,}6)$ (viz skripta Teorie pravděpodobnosti a matematická statistika, sbírka příkladů, příklad 8.10.),

$E(Y_{100}) = n\vartheta = 100.0{,}6 = 60$, $D(Y_{100}) = n\vartheta(1-\vartheta) = 100.0{,}6.0{,}4 = 24$ Označme $\Phi_{100}(y)$ distribuční funkci náhodné veličiny Y_{100}, $\Phi_{100}(y) = \sum_{t=0}^{y} \binom{100}{t} 0{,}6^t 0{,}4^{100-t}$.

Přesný výpočet: $P(Y_{100} < 50) = P(Y_{100} \le 49) = \Phi_{100}(49) = 0{,}016761686$.

Vytvoříme datový soubor o jedné proměnné a o jednom případu. Do Dlouhého jména této proměnné napíšeme =IBinom(49;0,6;100). Funkce IBinom(x;p;n) počítá hodnotu distribuční funkce rozložení Bi(n,p) v bodě x.

Přibližný výpočet: užijeme důsledek Moivreovy - Laplaceovy integrální věty (viz skripta Základní statistické metody, věta 6.3.1.1.). Nejdříve ověříme splnění podmínky dobré aproximace $n\vartheta(1-\vartheta) = 100.0{,}6.0{,}4 = 24 > 9$. Podmínka je splněna.

$P(Y_{100} < 50) = P(Y_{100} \le 49) \approx \Phi(49)$, kde $\Phi(49)$ je hodnota distribuční funkce rozložení N(60; 24) v bodě 80.

Vytvoříme datový soubor o jedné proměnné a o jednom případu. Do Dlouhého jména této proměnné napíšeme =INormal(49;60;sqrt(24)).

Zjistíme, že $\Phi(49) = 0{,}012372$.

Přesný výpočet

	1 Prom1
1	0,016762

Aproximativní výpočet

	1 Prom1
1	0,012372

Úkol 3.: Intervaly spolehlivosti pro parametry μ, σ^2 **normálního rozložení**

Z populace stejně starých selat téhož plemene bylo vylosováno šest selat a po dobu půl roku jim byla podávána táž výkrmná dieta. Byly zaznamenávány průměrné denní přírůstky hmotnosti v Dg. Z dřívějších pokusů je známo , že v populaci mívají takové přírůstky normální rozložení, avšak střední hodnota i rozptyl se měnívají. Přírůstky v Dg: 62, 54, 55, 60, 53, 58.

a) Najděte 95% empirický levostranný interval spolehlivosti pro neznámou střední hodnotu μ při neznámé směrodatné odchylce σ.
b) Najděte 95% empirický interval spolehlivosti pro směrodatnou odchylku σ.

Návod:

Vytvoříme datový soubor o 4 proměnných a 6 případech. První proměnnou nazveme hmotnost, druhou dm1, třetí dm2 a čtvrtou hm2. Do proměnné hmotnost zapíšeme zjištěné údaje. Pomocí Popisných statistik zjistíme realizace výběrového průměru a výběrové směrodatné odchylky.

	Popisné statistiky	
Proměnná	Průměr	Sm. odch.
hmotnost	57,00000	3,577709

ad a) Dolní mez 100(1-α)% empirického levostranného intervalu spolehlivosti pro μ při neznámém σ^2 je $m-\frac{s}{\sqrt{n}}t_{1-\alpha}(n-1)$ (viz skripta Základní statistické metody, věta 6.1.2.1., bod 2 b), tedy v našem případě $57-\frac{3{,}577709}{\sqrt{6}}t_{0,95}(5)=54{,}06$

Do Dlouhého jména proměnné dm1 zapíšeme výraz = 57 – 3,577709* VStudent(0,95;5)/sqrt(6) Funkce VStudent(x;df) počítá x-kvantil rozložení t(df). Dostaneme výsledek 54,05682, tedy $\mu > 54{,}06$ Dg s pravděpodobností aspoň 0,95.

ad b) Meze 100(1-α)% empirického oboustranný intervalu spolehlivosti pro σ při neznámém μ jsou $\left(\frac{\sqrt{n-1}\,s}{\sqrt{\chi^2_{1-\alpha/2}(n-1)}};\frac{\sqrt{n-1}\,s}{\sqrt{\chi^2_{\alpha/2}(n-1)}}\right)$ (viz skripta Základní statistické metody, věta 6.1.2.1., bod 3 a – nutno odmocnit).

Do Dlouhého jména proměnné dm2 zapíšeme výraz =3,577709*sqrt(5)/sqrt(VChi2(0,975;5)). Vyjde 2,233235. Podobně do Dlouhého jména proměnné hm2 zapíšeme výraz =3,577709*sqrt(5)/sqrt(VChi2(0,025;5)) Vyjde 8,774739 . Funkce VChi2(x;nu) počítá x-kvantil rozložení χ^2(nu).

Dostaneme výsledek: 2,23 g < σ < 8,77 g s pravděpodobností aspoň 0,95.

	1 dm1	2 dm2	3 hm2
1	54,05683	2,233235	8,774739

Úkol 4.: Interval spolehlivosti pro rozdíl parametrů $\mu_1 - \mu_2$ dvourozměrného normálního rozložení

Bylo vylosováno 6 vrhů selat a z nich vždy dva sourozenci. Jeden z nich vždy dostal náhodně dietu č. 1 a druhý dietu č. 2. Přírůstky v Dg jsou následující: (62,52), (54,56), (55,49), (60,50), (53,51), (58,50). Za předpokladu, že uvedené dvojice tvoří náhodný výběr z dvourozměrného normálního rozložení s vektorem středních hodnot (μ_1, μ_2), sestrojte 95% interval spolehlivosti pro rozdíl středních hodnot.

Návod:

Vytvoříme datový soubor o třech proměnných a šesti případech. Do proměnných v1 a v2 zapíšeme naměřené přírůstky, do proměnné v3 uložíme rozdíly v1 - v2.

Ve STATISTICE je implementován výpočet oboustranného intervalu spolehlivosti pro μ, když σ^2 neznáme (viz skripta Základní statistické metody, věta 6.1.2.1., bod 2 a). Pomocí Popisných statistik zjistíme meze 95% intervalu spolehlivosti pro střední hodnotu proměnné v3 tak, že zaškrtneme Meze spoleh. prům.

	Popisné statistiky	
Proměnná	Int. spolehl. -95,000%	Int. spolehl. +95,000%
Prom3	0,626461	10,70687

Dostaneme výsledek: $0{,}63 \text{ Dg} < \mu < 10{,}71 \text{ Dg}$ s pravděpodobností aspoň 0,95.

Úkol 5.: Asymptotický interval spolehlivosti pro parametr ϑ alternativního rozložení

Může politická strana, pro niž se v předvolebním průzkumu vyslovilo 60 z 1000 dotázaných osob, očekávat se spolehlivostí aspoň 0,95, že by v této době ve volbách překročila 5% hranici pro vstup do parlamentu?

Návod:

Zavedeme náhodné veličiny $X_1, \dots, X_{1000}$, přičemž $X_i = 1$, když i-tá osoba se vysloví pro danou politickou stranu a $X_i = 0$ jinak, $i = 1, \dots, 1000$. Tyto náhodné veličiny tvoří náhodný výběr z rozložení $A(\vartheta)$. V tomto případě $n = 1000$, $m = 60/1000 = 0{,}06$, $\alpha = 0{,}05$, $u_{1-\alpha} = u_{0,95} = 1{,}645$.

Ověření podmínky $n\vartheta(1-\vartheta) > 9$: parametr ϑ neznáme, musíme ho nahradit výběrovým průměrem. Pak $1000.0{,}06.0{,}94 = 56{,}4 > 9$.

95% levostranný interval spolehlivosti pro ϑ je

$$\left(m - \sqrt{\frac{m(1-m)}{n}}\, u_{1-\alpha} \;;\; \infty \right) = \left(0{,}06 - \sqrt{\frac{0{,}06(1-0{,}06)}{1000}}\, u_{0,95} \;;\; \infty \right)$$

(viz skripta Základní statistické metody, důsledek 6.3.2.2.)

Vytvoříme datový soubor o jedné proměnné a o jednom případu. Do Dlouhého jména této proměnné napíšeme =0,06-sqrt(0,06*0,94/1000)*VNormal(0,95;0;1). Vyjde 0,047647.

S pravděpodobností přibližně 0,95 tedy $\vartheta > 0{,}03$. Protože tento interval zahrnuje i hodnoty nižší než 0,05, nelze vyloučit, že strana získá méně než 5% hlasů.

Úkol 6.: Testování hypotézy o parametru μ normálního rozložení
Systematická chyba měřicího přístroje se eliminuje nastavením přístroje a měřením etalonu, jehož správná hodnota je $\mu = 10{,}00$. Nezávislými měřeními za stejných podmínek byly získány hodnoty: 10,24 10,12 9,91 10,19 9,78 10,14 9,86 10,17 10,05, které považujeme za realizace náhodného výběru rozsahu 9 z rozložení $N(\mu, \sigma^2)$. Je možné při riziku 0,05 vysvětlit odchylky od hodnoty 10,00 působením náhodných vlivů?

Návod:
Na hladině významnosti 0,05 testujeme hypotézu H_0: $\mu = 10$ proti oboustranné alternativě H_1: $\mu \neq 10$. Jde o úlohu na jednovýběrový t-test (viz skripta Základní statistické metody, návod 6.1.4.2.). Ten je ve STATISTICE implementován. Vytvoříme datový soubor o jedné proměnné a devíti případech, kam zapíšeme naměřené hodnoty. V Základních statistikách/tabulkách vybereme t-test, samostatný vzorek. Do Referenčních hodnot zapíšeme 10. Ve výstupu se podíváme na hodnotu testového kritéria a na p-hodnotu. Pokud p-hodnota bude menší nebo rovna 0,05, zamítneme hypotézu H_0: $\mu = 10$ ve prospěch oboustranné alternativní hypotézy H_1: $\mu \neq 10$ na hladině významnosti 0,05. V opačném případě H_0 nezamítáme. V našem případě je

	Test průměrů vůči referenční konstantě (hodnotě)							
Proměnná	Průměr	Sm.odch.	N	Sm.chyba	Referenční konstanta	t	SV	p
Prom1	10,05111	0,162669	9	0,054223	10,00000	0,942611	8	0,373470

Protože p-hodnota 0,373470 > 0,05 nulovou hypotézu nezamítáme na hladině významnosti 0,05. S rizikem omylu nejvýše 5% lze tedy odchylky od hodnoty 10 vysvětlit působením náhodných vlivů.
Všimněme si ještě hodnoty testového kriteria: $t_0 = 0{,}942611$. Kritický obor

$$W = \left(-\infty, -t_{1-\alpha/2}(n-1)\right\rangle \cup \left\langle t_{1-\alpha/2}(n-1), \infty\right) = \left(-\infty, -t_{0,975}(8)\right\rangle \cup \left\langle t_{0,975}(8), \infty\right) =$$
$$= \left(-\infty, -2{,}306\right\rangle \cup \left\langle 2{,}306, \infty\right)$$

Protože $t_0 \notin W$, nezamítáme na hladině významnosti 0,05 hypotézu H_0 .

Úkol 7.: Testování hypotézy o rozdíl parametrů μ_1 - μ_2 dvourozměrného normálního rozložení
Pro data z úkolu 4. testujte na hladině významnosti 0,05 hypotézu, že obě výkrmné diety mají stejný vliv.

Návod: Označme $\mu = \mu_1 - \mu_2$. Na hladině významnosti 0,05 testujeme hypotézu H_0: $\mu = 0$ proti oboustranné alternativě H_1: $\mu \neq 0$. Jde o úlohu na párový t-test (viz skripta Základní statistické metody, odstavec 6.2.2.). Ten je ve STATISTICE implementován.Vytvoříme datový soubor o dvou proměnných a šesti případech. Do proměnných v1 a v2 zapíšeme naměřené přírůstky. V Základních statistikách/tabulkách vybereme t-test, závislé vzorky. Zadáme názvy obou proměnných a ve výstupu se podíváme na hodnotu testového kritéria a na p-hodnotu.

Proměnná	t-test pro závislé vzorky Označ. rozdíly jsou významné na hlad. p < ,05000							
	Průměr	Sm.odch.	N	Rozdíl	Sm.odch. rozdílu	t	sv	p
Prom1	57,00000	3,577709						
Prom2	51,33333	2,503331	6	5,666667	4,802777	2,890087	5	0,034183

Protože p-hodnota 0,034183 < 0,05 zamítáme hypotézu H_0: $\mu = 0$ ve prospěch alternativní hypotézy H_1: $\mu \neq 0$ na hladině významnosti 0,05. Znamená to, že jsme s rizikem omylu nejvýše 5% prokázali rozdíl v účinnosti obou výkrmných diet.
Všimněme si ještě hodnoty testového kriteria: $t_0 = 2{,}890087$. Kritický obor

$$W = \left(-\infty, -t_{1-\alpha/2}(n-1)\right\rangle \cup \left\langle t_{1-\alpha/2}(n-1), \infty\right) = \left(-\infty, -t_{0,975}(5)\right\rangle \cup \left\langle t_{0,975}(5), \infty\right) =$$
$$= \left(-\infty, -2{,}5706\right\rangle \cup \left\langle 2{,}5706, \infty\right)$$

Protože $t_0 \in W$, zamítáme na hladině významnosti 0,05 hypotézu H_0.

Úkol 8: Testování hypotézy o parametru ϑ alternativního rozložení

Určitá cestovní kancelář organizuje zahraniční zájezdy podle individuálních přání zákazníků. Z několika minulých let ví, že 30% všech takto organizovaných zájezdů má za cíl zemi X. Po zhoršení politických podmínek v této zemi se cestovní kancelář obává, že se zájem o tuto zemi mezi zákazníky sníží. Ze 150 náhodně vybraných zákazníků v tomto roce má 38 za cíl právě zemi X. Potvrzují nejnovější data pokles zájmu o tuto zemi? Volte hladinu významnosti 0,05.

Návod:
Máme náhodný výběr $X_1, \dots, X_{150}$ z rozložení A(0,3). Testujeme H_0: $\vartheta = 0{,}3$ proti levostranné alternativě H_1: $\vartheta < 0{,}3$. V tomto případě je testovým kritériem statistika

$$T_0 = \frac{M - c}{\sqrt{\frac{c(1-c)}{n}}}$$

, která v případě platnosti nulové hypotézy má asymptoticky rozložení N(0,1) (viz skripta Základní statistické metody, věta 6.3.3.1.). Musíme ověřit splnění podmínky $n\vartheta(1-\vartheta) > 9$: 150.0,3.0,7 = 31,5 > 9. Vypočteme realizaci testového kritéria:

$$\frac{m - c}{\sqrt{\frac{c(1-c)}{n}}} = \frac{\frac{38}{150} - 0{,}3}{\sqrt{\frac{0{,}3(1-0{,}3)}{150}}} = -1{,}24722 .$$ Kritický obor: $W = \left(-\infty, -u_{1-\alpha}\right\rangle = \left(-\infty, -1{,}645\right\rangle$.

Protože testové kritérium nepatří do kritického oboru, H_0 nezamítáme na asymptotické hladině významnosti 0,05. S rizikem omylu nejvýše 5% tedy naše data neprokázala pokles zájmu zákazníků cestovní kanceláře o zemi X.
Vytvoříme datový soubor o dvou proměnných a jednom případu. Vypočteme realizaci testového kritéria tak, že do Dlouhého jména první proměnné zapíšeme odpovídající vzorec, tj. =(38/150-0,3)/sqrt(0,3*0,7/150). Do Dlouhého jména druhé proměnné napíšeme =VNormal(0,95;0;1)**,** čímž získáme kvantil $u_{0,95}$ a testové kritérium porovnáme s opačnou hodnotou tohoto kvantilu.

	1 Prom1	2 Prom2
1	-1,24722	1,644854

Protože testové kritérium není menší než opačná hodnota příslušného kvantilu, nulovou hypotézu nezamítáme na asymptotické hladině významnosti 0,05.