Cvičení 6.: Parametrické úlohy o jednom výběru a dvou nezávislých výběrech z alternativního rozložení

Úkol 1.: 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:

$X_1, \dots, X_{100}$ 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 $Y_{100} = X_1 + \dots + X_{100}$, $Y_{100} \sim Bi(100; 0{,}6)$ (viz skripta Teorie pravděpodobnosti a matematická statistika, sbírka příkladů, příklad 8.10.), $E(Y_{100}) = n\vartheta = 100.0{,}6 = 60$, $D(Y_{100}) = n\vartheta(1-\vartheta) = 100.0{,}6.0{,}4 =$ Označme $\Phi_{100}(y)$ distribuční funkci náhodné veličiny Y_{100}, $\Phi_{100}(y) = \sum_{t \le y} \binom{100}{t} 0{,}6^t\, 0{,}4^{100-t}$.

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 49.
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 Prom
1	0,016

Aproximativní výpočet

	1 Prom
1	0,012

Úkol 2.: 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, ..., 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

$$\left(\frac{m}{n}-\sqrt{\frac{\frac{m}{n}\left(1-\frac{m}{n}\right)}{n}}\,u_{1-\alpha};\ \infty\right)=\left(0{,}06-\sqrt{\frac{0{,}06(1-0{,}06)}{1000}}\,u_{0,95};\ \infty\right)$$ (viz skripta Základní statistické metody, důsledek 6.3.2.2.)

Postup ve STATISTICE:
1. možnost: 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.
2. možnost: Statistiky – Analýza síly testu – Odhad intervalu – Jeden podíl, Z, Chí-kvadrát test – OK – Pozorovaný podíl p: 0,06, Velik. Vzorku (N): 1000, Spolehlivost: 0,9 – Vypočítat. Dostaneme 0,0476.
S pravděpodobností přibližně 0,95 tedy ϑ > 0,047647. Protože tento interval zahrnuje i hodnoty nižší než 0,05, nelze vyloučit, že strana získá méně než 5% hlasů.

Úkol 3: 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

$$T_0=\frac{M-\vartheta}{\sqrt{\frac{\vartheta(1-\vartheta)}{n}}}$$, 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:

$$\frac{m-\vartheta}{\sqrt{\frac{\vartheta(1-\vartheta)}{n}}}=\frac{\frac{38}{150}-0{,}3}{\sqrt{\frac{0{,}3(1-0{,}3)}{150}}}=-1{,}247.$$ Kritický obor: $W=(-\infty, -u_{1-\alpha}]=(-\infty; -1{,}645]$.

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.

Postup ve STATISTICE:
Asymptotický způsob: Vytvoříme datový soubor o dvou proměnných (nazveme je t0 a kvantil) a jednom případu. Vypočteme realizaci testového kritéria tak, že do Dlouhého jména proměnné t0 napíšeme
=(38/150-0,3)/sqrt(0,3*0,7/150)
Do Dlouhého jména proměnné kvantil napíšeme
=VNormal(0,95;0;1)
Tím získáme kvantil $u_{0,95}$.

	1 t0	2 kvant
1	-1,2472	1,644

Jelikož realizace testového kritéria $t_0 = -1{,}24721913$ nepatří do kritického oboru $W = \dots 448\rangle$, H_0 nezamítáme na asymptotické hladině významnosti 0,05.

Přibližný způsob: Do nového datového souboru o jedné proměnné X a 150 případech uložíme 38 jedniček (indikují zájem o danou zemi) a 112 nul (indikují nezájem o danou zemi). Statistika – Základní statistiky a tabulky – t-test, samost. vzorek – OK – Proměnné X – OK, Test všech průměrů vůči 0,3 – Výpočet.

Proměi	Test průměrů vůči referenční konstantě (hodnot							
	Prům	Sm.od	N	Sm.chy	Referer konstai	t	SV	p
X	0,253	0,436	15	0,035	0,300	-1,30	14	0,192

Hodnota testové statistiky je při tomto přibližném způsobu -1,30976. Odpovídající p-hodnota je 0,1923, ovšem to je p-hodnota pro oboustranný test. Tuto p-hodnotu tedy musíme dělit dvěma a dostaneme 0,0961. Na asymptotické hladině významnosti 0,05 nelze zamítnout hypotézu, že zájem o danou zemi se nezměnil.

Úkol 4.: Asymptotický interval spolehlivosti pro parametrickou funkci $\vartheta_1 - \vartheta_2$

Při výstupní kontrole bylo náhodně vybráno 150 výrobků vyrobených na ranní směně a rovněž 150 výrobků vyrobených na odpolední směně. U ranní směny bylo zjištěno 16 zmetků a u odpolední 12 zmetků. Sestrojte 95% asymptotického interval spolehlivosti pro rozdíl pravděpodobností vyrobení zmetku v obou směnách.

Návod: Zavedeme náhodnou veličinu X_{1i}, která bude nabývat hodnoty 1, když i-tý výrobek z ranní směny je zmetek, 0 jinak, i = 1, …, 150. Náhodné veličiny $X_{1,1}, \dots, X_{1,150}$ tvoří náhodný výběr z rozložení $A(\vartheta_1)$. Dále zavedeme náhodnou veličinu X_{2i}, která bude nabývat hodnoty 1, když i-tý výrobek z odpolední směny je zmetek, 0 jinak, i = 1, …, 150.. Náhodné veličiny $X_{2,1}, \dots, X_{2,150}$ tvoří náhodný výběr z rozložení $A(\vartheta_2)$.

$n_1 = 150$, $n_2 = 150$, $m_1 = 16/150 = 0{,}1067$, $m_2 = 12/150 = 0{,}08$.

Ověření podmínek $n_1\vartheta_1(1-\vartheta_1) > 9$ a $n_2\vartheta_2(1-\vartheta_2) > 9$: Parametry ϑ_1 a ϑ_2 neznáme, nahradíme je odhady m_1 a m_2: 16.(1-16/150) = 14,29 > 9, 12.(1-12/150) = 11,04 > 9.

Meze 100(1-α)% asymptotického empirického intervalu spolehlivosti pro parametrickou funkci $\vartheta_1 - \vartheta_2$ jsou:

$$d = m_1 - m_2 - \sqrt{\frac{m_1(1-m_1)}{n_1} + \frac{m_2(1-m_2)}{n_2}}\,u_{1-\alpha/2} =$$

$$= \frac{16}{150} - \frac{12}{150} - \sqrt{\frac{\frac{16}{150}(1-\frac{16}{150})}{150} + \frac{\frac{12}{150}(1-\frac{12}{150})}{150}}\,1{,}96 = -0{,}039$$

$$h = m_1 - m_2 + \sqrt{\frac{m_1(1-m_1)}{n_1} + \frac{m_2(1-m_2)}{n_2}}\,u_{1-\alpha/2} =$$

$$= \frac{16}{150} - \frac{12}{150} + \sqrt{\frac{\frac{16}{150}(1-\frac{16}{150})}{150} + \frac{\frac{12}{150}(1-\frac{12}{150})}{150}}\,1{,}96 = 0{,}092$$

Zjistili jsme tedy, že s pravděpodobností přibližně 0,95: $-0{,}039 < \vartheta_1 - \vartheta_2 < 0{,}092$.

Postup ve STATISTICE:
Otevřeme nový datový soubor se dvěma proměnnými d a h a o jednom případu. Do Dlouhého jména proměnné d napíšeme:
=16/150-12/150-sqrt((16/150)*(134/150)/150+(12/150)*(138/150)/150)*VNormal(0,975;0;1)
Do Dlouhého jména proměnné h napíšeme:
=16/150-
12/150+sqrt((16/150)*(134/150)/150+(12/150)*(138/150)/150)*VNormal(0,975;0;1)
Dostaneme tabulku

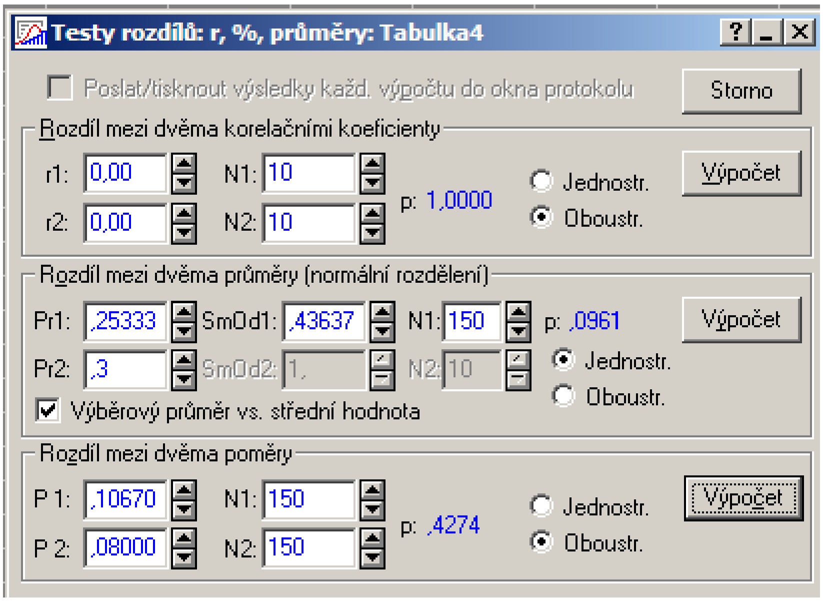
	1 d	2 h
1	-0,0:	0,092

S pravděpodobností přibližně 0,95 se rozdíl pravděpodobností vyrobení zmetku na ranní a odpolední směně nachází v intervalu (-0,039; 0,092).

Úkol 5.: Testování hypotézy o parametrické funkci ϱ

Pro údaje z úkolu 4 testujte na asymptotické hladině významnosti 0,05 hypotézu, že pravděpodobnost vyrobení zmetků v obou směnách je táž.

Postup ve STATISTICE:
Statistiky – Základní statistiky a tabulky – Testy rozdílů: r, %, průměry – OK – vybereme Rozdíl mezi dvěma poměry – do políčka P 1 napíšeme 0,1067, do políčka N1 napíšeme 150, do políčka P 2 napíšeme 0,08, do políčka N2 napíšeme 150 –Výpočet. Dostaneme p-hodnotu 0,4274, tedy nezamítáme nulovou hypotézu na hladině významnosti 0,05.



Úkol k samostatnému řešení: Přírůstky cen akcií na burze (v %) u 10 náhodně vybraných společností dosáhly těchto hodnot: 10, 16, 5, 10, 12, 8, 4, 6, 5, 4. Sestrojte 95% asymptotický empirický interval spolehlivosti pro pravděpodobnost, že přírůstek ceny akcie překročí 8,5%.
Výsledek: 0,096 < ϱ < 0,704 s pravděpodobností aspoň 0,95.

Znamená to, že pravděpodobnost, že přírůstek ceny akcie překročí 8,5%, je aspoň 9,6% a nanejvýš 70,4% (při spolehlivosti 95%.)

Úkol k samostatnému řešení: Z 28 studentek oboru národní hospodářství mělo z matematiky trojku 17 studentek, zatímco z 20 studentek oboru informatika mělo z matematiky trojku jen 6 studentek. Na asymptotické hladině významnosti 0,05 testujte hypotézu, že pravděpodobnost získání trojky z matematiky je obě skupiny studenek stejná.
Výsledek: Testová statistika se realizuje hodnotou t_0 = 2,100009, kritický obor je $W = (-\infty; -1,96\rangle \cup \langle 1,96; \infty)$. Protože $t_0 \in W$, zamítáme nulovou hypotézu na asymptotické hladině významnosti 0,05.
Použijeme-li v systému STATISTICA aplikaci Testy rozdílů, dostaneme p-hodnotu 0,0358, tedy na asymptotické hladině významnosti 0,05 zamítáme nulovou hypotézu.