

Statistická inference II


*Zadání domácího úkolu
jarní semestr 2021*

Stanislav Katina
Veronika Bendová

katina@math.muni.cz

5. května 2021

Při vypracování domácího úkolu se řiďte striktně pokyky k domácímu úkolu uvedenými v samostatném .pdf souboru.

*DÚ je nejprve po formální stránce hodnocen cvičícím. Toto hodnocení zahrnuje: (1) přítomnost požadovaných souborů se správnými názvy, (2) kompletnost zpracování, (3) dostatečný opis vašich úvah, zvoleného postupu a interpretace výsledků, ať už tabulkových nebo grafických, (4) funkčnost, přehlednost a úpravu -kódu podle pravidel uváděných na kurzu Statistická inference I, případně v prezentaci *Standards of programming in R: R style guide*.*

DÚ je potřeba odevzdat do odevzdávacího 7 dní před termínem zkoušky, na který se přihlásíte.

Příklad 1. Test o pravděpodobnosti p ; praktický příklad

Mějme datový soubor `25-one-sample-probability-dermatoglyphs.txt` obsahující údaje o frekvenci výskytu dermatoglyfických vzorů *vír* (*whorl*), *smyčka* (*loop*) a *oblouček* (*arch*) na deseti prstech 470 jedinců (235 mužů a 235 žen) populace Bagathů z Araku Valley. Současně máme k dispozici hodnotu pravděpodobnosti výskytu dermatoglyfického vzoru *oblouček* u jedinců z populace Valmikis ($p_m = 0.0410$, $p_f = 0.0440$).

Na hladině významnosti $\alpha = 0.10$ otestujte, zda existuje rozdíl mezi frekvencemi výskytu dermatoglyfického vzoru *oblouček* u žen bagathské populace z Araku Valley a žen z populace Valmikis. Před testováním ověřte splnění Haldovy podmínky dobré aproximace. Testování proved'te pomocí (a) kritického oboru; (b) intervalu spolehlivosti; (c) p -hodnoty při použití testovacích statistik (1) Z_W , (2) U_S , (3) U_{LR} . Vykreslete graf zobrazující hranice a oblast 90 % empirického věrohodnostního intervalu spolehlivosti pro test o parametru p . Dále vykreslete sloupcový diagram relativních četností výskytu vzoru *oblouček* v populaci bagathských žen z Araku Valley. U každého typu testování (a)–(c) pro každý test (1)–(3) uveďte zdůvodněný závěr o H_0 a nakonec uveďte antropologickou interpretaci výsledku testování.

Poznámka: Při testování H_0 ověřujeme podmínku dobré aproximace ve tvaru $Np(1-p) > 9$, kde za p dosazujeme hodnotu p_0 z nulové hypotézy H_0 .

Poznámka: Sloupcový diagram relativních četností můžeme vykreslit například pomocí funkce `rel.barplot()` z `R`-Skriptu `Sbirka-AS-I-2018-funkce.R`. Jediným povinným vstupním argumentem funkce je vektor $(x, N-x)^T$, ostatní argumenty jsou volitelné. `R`-skript je vždy třeba automaticky načítat příkazem `source()`.

Požadovaná forma výstupu příkladu:

1. Test o pravděpodobnosti p :

- stanovení nulové a alternativní hypotézy (i) slovním zápisem; (ii) matematickým zápisem,
- tabulka výsledků (zaokrouhlená na čtyři desetinná místa) testování založeného na testovacích statistikách Z_W , U_S a U_{LR} pomocí (a) kritického oboru, (b) intervalu spolehlivosti, (c) p -hodnoty (viz tabulka 1),

Tabulka 1: Výsledek testu o pravděpodobnosti p při použití testovacích statistik Z_W , U_S a U_{LR}

Statistika	\hat{p}	statistika	W_{hh}	W_{dh}	IS_{dh}	IS_{hh}	p -hodnota
Z_W							
U_S							
U_{LR}							

- zdůvodněné rozhodnutí o nulové hypotéze (u všech tří způsobů testování (a)–(c) u všech tří testů (1)–(3); analogicky, jako bylo uváděno ve studijních materiálech ke cvičením) + interpretace výsledku testování.

2. Grafická vizualizace:

- graf zobrazující 90% věrohodnostní empirický interval spolehlivosti pro test o pravděpodobnosti p . V grafu bude vykreslena křivka testovací statistiky pro rostoucí posloupnost hodnot pravděpodobnosti p , barevně vyznačené hodnoty testovacích statistik, jejichž argumenty vyhovují intervalům spolehlivosti, referenční vodorovná přímka v hodnotě kvantilu $\chi_1^2(1-\alpha)$ a referenční svislá přímka v hodnotě p_0 spolu s bodem se souřadnicemi $[p_0, u_{LR}(p_0)]$. Barva svislé přímky a bodu bude (a) zelená, pokud $p_0 \in IS$; (b) červená, pokud $p_0 \notin IS$. Rozsah osy x volte rozumně, např. v rozmezí 0.02 až 0.08.
- sloupcový diagram relativních četností. Sloupec s výškou 1 bude rozdělen na dvě plochy, první plocha bude vizualizovat pravděpodobnost výskytu vzoru *oblouček* u žen bagathské populace z Araku Valley, druhá oblast bude vizualizovat pravděpodobnost výskytu jiného vzoru (*vír* + *smyčka*). Uprostřed každé oblasti bude umístěn popisek s hodnotou absolutní četnosti a relativní četnosti (v procentuální škále). Obě oblasti budou barevně odlišeny. V grafu bude dále vykreslena vodorovná červená čára ve výšce pravděpodobnosti výskytu vzoru *oblouček* u žen z populace Valmikis. Napravo od červené čáry bude umístěn červený popisek s hodnotou této pravděpodobnosti (v procentuální škále). Součástí diagramu bude legenda.
- komentář ke grafům. U grafu 95 % empirického věrohodnostního IS se zaměřte například na popis polohy p_0 vzhledem k IS a souvislost se závěrem o H_0 pomocí intervalu spolehlivosti. U sloupcového diagramu relativních četností například porovnejte vztah pravděpodobnosti výskytu vzoru *oblouček* u bagathské populace z Araku Valley a u populace Valmikis. Uveďte toto porovnání do souvislosti se závěrem testování o H_0 a s finální interpretací výsledku testování.

Příklad 2. Silofunkce pro test o korelačním koeficientu ρ

Předpokládejme, že $(X, Y)^T \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, kde $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$ je vektor středních hodnot a $\boldsymbol{\Sigma}$ je kovarianční matice. Na hladině významnosti $\alpha = 0.10$ testujeme všechny tři typy hypotéz

- a) $H_{01} : \rho = \rho_0$ oproti $H_{11} : \rho \neq \rho_0$ (oboustranná);
- b) $H_{02} : \rho \leq \rho_0$ oproti $H_{12} : \rho > \rho_0$ (pravostranná);
- c) $H_{03} : \rho \geq \rho_0$ oproti $H_{13} : \rho < \rho_0$ (levostranná).

1. Odvoďte tvary exaktních silofunkcí pro všechny tři typy hypotéz (a)–(c) při použití testovací statistiky pro rozdíl $\xi - \xi_0$, t.j. tvary $\beta_{11}^*(\xi)$, $\beta_{12}^*(\xi)$ a $\beta_{13}^*(\xi)$. Dále nakreslete silofunkce pro všechny tři typy hypotéz (a)–(c), kde $\rho_0 = -0.23$ a $\alpha = 0.10$. Do jednoho obrázku zakreslete vždy tvary silofunkcí pro $n = 7$, $n = 15$, $n = 25$, $n = 50$ a $n = 100$. Hodnoty ρ volte $\{-1, -0.999, -0.9998, \dots, 0.998, 0.999, 1\}$.
2. Pro situaci (a) odvoďte tvar přibližné (aproximované) silofunkce při použití testovací statistiky pro rozdíl $\xi - \xi_0$, t.j. tvar $\tilde{\beta}_{11}^*(\xi)$. Do jednoho grafu vykreslete křivky obou silofunkcí (exaktní a aproximované) a porovnejte jejich tvary. Výsledek slovně okomentujte. Hodnotu n zvolte 50, $\rho_0 = -0.23$. Rozsah osy x volte rozumně, pro globální pohled např. $\langle -1; 1 \rangle$, pro lokální pohled např. $\langle -0.45; 0.00 \rangle$. Rozsah osy y pro lokální pohled volte vhodně vzhledem k rozsahu osy x .

Požadovaná forma výstupu příkladu:

1. Tři odvození silofunkcí pro všechny tři typy hypotéz (a)–(c). Každé odvození bude začínat řádkem $\beta_{1i}^*(\xi) = \Pr(H_0 \text{ zamítáme} | H_0 \text{ neplatí})$ a končit příslušným vzorcem exaktní síly uvedeným v řešení příkladu 7.4 (viz cvičení 7).
2. Vlastnoručně naprogramovaná funkce `sila.exakt()`, která pro zadané hodnoty parametrů ρ_0 , ρ , n a α a zvolený typ alternativní hypotézy vrátí hodnotu exaktní silofunkce.
3. Tři grafy silofunkcí ((a) pro oboustrannou; (b) pro pravostrannou; (c) pro levostrannou alternativu). V každém grafu bude pět barevně odlišených křivek silofunkcí ($n = 7, 15, 25, 50, 100$), vodorovná referenční přímka hodnotě α a legenda.
4. Podrobné komentáře popisující tvary silofunkcí a porovnávající silofunkce pro různá n . Závěrem zmiňte, (a) kde se v grafech projevilo zvolení parametru $\rho_0 = -0.23$ a jak by se grafy změnily při zvolení hodnoty parametru $\rho_0 = 0.23$; (b) kde se v grafech projevilo zvolení hladiny významnosti $\alpha = 0.10$ a jak by se grafy změnily při zvolení hladiny významnosti $\alpha = 0.05$.
5. Odvození aproximované silofunkce. Odvození bude začínat vzorcem exaktní silofunkce uvedené v příkladu 7.4 (viz cvičení 7) a končit tvarem aproximované silofunkce uvedené v příkladu 7.4 (viz cvičení 7).
6. Vlastnoručně naprogramovaná funkce `sila.aprox()`, která pro zadané hodnoty parametrů ρ_0 , ρ , n a α vrátí hodnotu aproximované silofunkce $\tilde{\beta}_{11}^*(\xi)$.
7. Dva grafy. První graf bude zobrazovat globální pohled na porovnání aproximované a exaktní silofunkce, druhý graf bude zobrazovat lokální pohled na porovnání aproximované a exaktní silofunkce. Oba grafy budou obsahovat dvě barevně odlišené křivky (exaktní a aproximované síly), vodorovnou referenční přímku v hodnotě α a legendu.
8. Podrobné komentáře popisující rozdíly mezi exaktní a aproximovanou silofunkcí. Závěrem uveďte, kdy můžeme použít aproximovanou silofunkci namísto exaktní silofunkce, aniž bychom se dopustili výrazné chyby a naopak, kdy bychom se použitím aproximované síly namísto exaktní síly výrazné chyby dopustili.

Příklad 3. Minimální rozsah náhodného výběru

Předpokládejme, že $X_1 \sim N(\mu_1, \sigma_1^2)$ a $X_2 \sim N(\mu_2, \sigma_2^2)$. Testujeme všechny tři typy hypotéz:

1. $H_{01} : \mu_1 - \mu_2 = \mu_0$ oproti $H_{11} : \mu_1 - \mu_2 \neq \mu_0$ (oboustranná),
2. $H_{02} : \mu_1 - \mu_2 \leq \mu_0$ oproti $H_{12} : \mu_1 - \mu_2 > \mu_0$ (pravostranná),
3. $H_{03} : \mu_1 - \mu_2 \geq \mu_0$ oproti $H_{13} : \mu_1 - \mu_2 < \mu_0$ (levostranná).

Pro hypotézy (a), (b), (c) odvoďte vztah pro minimální rozsah n_1 náhodného výběru X_1 a pro minimální rozsah n_2 náhodného výběru X_2 při použití testovací statistiky Z_W pro rozdíl $\mu_1 - \mu_2$. Naprogramujte funkci `min.n()`, která pro libovolnou z hypotéz (a), (b), (c) vypočítá minimální potřebné rozsahy n_1 a n_2 . Vstupní argumenty funkce budou předpokládaná hodnota výběrového rozdílu $\mu_1 - \mu_2$ (`diff`), hodnota μ_0 z H_0 (`mu0`), σ_1 (`sigma1`), σ_2 (`sigma2`), k (`k`), α (`alpha`) a $\beta^*(\mu_1 - \mu_2)$ (`sila`). Parametr k je předpokládaný poměr velikostí rozsahů $n_2 : n_1$.

Pomocí funkce `min.n()` vypočítejte minimální rozsahy obou náhodných výběrů pro hypotézy (a)–(c) při $\alpha = 0.01$, $\beta^*(\mu_1 - \mu_2) = 0.80$, $k = \frac{n_2}{n_1} = 1.25$. Parametr $\sigma_1^2 = 5.5^2$, $\sigma_2^2 = 2.5^2$, $\mu_0 = 0$ a předpokládaný rozdíl $\mu_1 - \mu_2 \in \{-7.0, -6.0, \dots, -2.0, -1.0, -0.5, 0.5, 1.0, 2.0, \dots, 6.0, 7.0\}$. Minimální rozsahy obou náhodných výběrů zakreslete jako bodový graf, kde na osu x vynesete rozdíl $\mu_1 - \mu_2 - \mu_0$, na osu y vypočítané minimální rozsahy n_1 a n_2 .

Určete, (1) jaké rozsahy náhodných výběrů X_1 a X_2 potřebujeme na otestování nulové hypotézy $H_0 : \mu_1 - \mu_2 \geq \mu_0$, kde $\mu_0 = 1$, chceme-li, aby síla testu $\beta^*(\mu_1 - \mu_2) = 0.85$ při stanovené hladině významnosti $\alpha = 0.10$, předpokládáme-li, že výběrový rozdíl $\mu_1 - \mu_2$ bude 0.7, poměr $k = \frac{n_2}{n_1} = 1$, $\sigma_1^2 = 3^2$, $\sigma_2^2 = 2^2$; (2) jaké rozsahy náhodných výběrů X_1 a X_2 potřebujeme na otestování nulové hypotézy $H_0 : \mu_1 - \mu_2 \leq \mu_0$, kde $\mu_0 = 2$, chceme-li, aby síla testu $\beta^*(\mu_1 - \mu_2) = 0.90$ při stanovené hladině významnosti $\alpha = 0.05$, předpokládáme-li, že výběrový rozdíl $\mu_1 - \mu_2$ bude nabývat hodnoty 4.0, poměr $k = \frac{n_2}{n_1} = 0.33$, $\sigma_1^2 = 4^2$ a $\sigma_2^2 = 7^2$; (3) jaké rozsahy náhodných výběrů X_1 a X_2 potřebujeme na otestování nulové hypotézy $H_0 : \mu_1 - \mu_2 = \mu_0$, kde $\mu_0 = 3$, chceme-li, aby síla testu $\beta^*(\mu_1 - \mu_2) = 0.95$ při stanovené hladině významnosti $\alpha = 0.01$, předpokládáme-li, že výběrový rozdíl $\mu_1 - \mu_2$ bude nabývat hodnoty 1.4, poměr $k = \frac{n_2}{n_1} = 0.75$, $\sigma_1^2 = 1^2$ a $\sigma_2^2 = 3^2$. Výsledné rozsahy pro přehlednost vložte spolu se zadanými parametry do přehledné tabulky.

Požadovaná forma výstupu příkladu:

1. Trojice grafů vizualizujících rozsahy náhodných výběrů n_1 a n_2 pro situace (a)–(c). Grafy budou bodového typu, na ose x budou hodnoty $\mu_1 - \mu_2 - \mu_0 \in \{-7.0, -6.0, \dots, -1.0, -0.5, 0.5, 1.0, \dots, 6.0, 7.0\}$, na ose y hodnoty minimálních rozsahů n_1 a n_2 pro každý rozdíl $\mu_1 - \mu_2 - \mu_0$. Body vizualizující rozsahy náhodného výběru X_1 a X_2 budou od sebe barevně odlišeny. Nad oběma body v dostatečné výšce budou uvedeny popisky s přesnými hodnotami rozsahů náhodných výběrů n_1 a n_2 opět barevně odlišené (barva popisků s hodnotami rozsahů n_1 , resp. n_2 bude ve stejném odstínu jako barva bodů pro n_1 , resp. pro n_2). Součástí grafu bude legenda.
2. Souhrnná tabulka výsledků minimálních rozsahů n_1 a n_2 náhodných výběrů X_1 a X_2 vypočítaných pro situace (1)–(3), viz tabulka 2.

Tabulka 2: Minimální rozsahy n_1 a n_2 náhodných výběrů X_1 a X_2 při předem zvolených hodnotách $\mu_1 - \mu_2$, μ_0 , σ_1^2 , σ_2^2 , k , α a β^*

	$\mu_1 - \mu_2$	μ_0	σ_1^2	σ_2^2	k	α	$\beta^*(\mu_1 - \mu_2)$	$\beta(\mu_1 - \mu_2)$	n_1	n_2
$H_0 : \mu_1 - \mu_2 \leq \mu_0$										
$H_0 : \mu_1 - \mu_2 = \mu_0$										
$H_0 : \mu_1 - \mu_2 \geq \mu_0$										

3. Komentáře popisující grafy minimálních rozsahů n_1 a n_2 . V komentářích se zaměřte například na porovnání velikostí rozsahů pro oboustrannou alternativu a jednostranné alternativy. Dále popište změnu velikostí rozsahů n_1 a n_2 s $\mu_1 - \mu_2 \rightarrow \mu_0$. Překvapil vás prudký vzrůst rozsahů n_1 a n_2 v hodnotě $\mu_1 - \mu_2 - \mu_0 = -0.5$ resp. v hodnotě 0.5? Vysvětlíte, proč s $\mu_1 - \mu_2 \rightarrow \mu_0$ dochází k navyšování velikostí rozsahů a jak by situace pokračovala kdybychom předpokládali hodnotu výběrového rozdílu $\mu_1 - \mu_2$ ještě bližší nule. Jaký důsledek má tento trend v praxi a proč je důležité počítat minimální rozsahy náhodných výběrů ideálně ještě ve fázi plánování experimentu?
4. Tři odpovědi (celou větou) na otázky (1)–(3).

Příklad 4. Rozdělení testovacích statistik Z_W , U_S a U_{LR} testu o parametru λ Poissonova rozdělení

Na hladině významnosti $\alpha = 0.05$ testujeme $H_0: \lambda = \lambda_0 = 2.5$ oproti $H_1: \lambda \neq \lambda_0$. Na základě simulační studie proveďte, zda za předpokladu, že

- $X \sim \text{Poiss}(\lambda)$, kde $\lambda = 2.5$;
- $X \sim p\text{Poiss}(\lambda) + (1-p)\text{Poiss}(\lambda_2)$, kde $\lambda = 2.5$, $\lambda_2 = 1.3$, $p = 0.9$;
- $X \sim \text{NegBin}(k, p)$, kde $k = 2.5$ a $p = 0.5$;
- $X \sim \text{Bin}(N, p)$, kde $N = 25$ a $p = 0.10$;

platí, že (1) $Z_W = \frac{\bar{X} - \lambda_0}{\sqrt{\bar{X}/N}} \sim N(0, 1)$; (2) $U_S = \frac{(\bar{X} - \lambda_0)^2}{\lambda_0/N} \sim \chi_1^2$; (3) $U_{LR} = 2N \left(\bar{x} \ln \frac{\bar{x}}{\lambda_0} - \bar{x} + \lambda_0 \right) \sim \chi_1^2$. Pro každou simulaci náhodného výběru X_m o rozsahu $n = 100$ vypočítejte (a) $Z_{W, \text{poz}, m}$; (b) $U_{S, \text{poz}, m}$; (c) $U_{LR, \text{poz}, m}$, $m = 1, 2, \dots, M$, kde $M = 1000$ a vykreslete histogramy těchto M testovacích statistik v relativní škále. Každý histogram superponujte teoretickou křivkou hustoty příslušného rozdělení. Nakonec spočítejte empirickou pravděpodobnost pokrytí (spolehlivost) $1 - \hat{\alpha}$ každého testu, a to jako počet testovacích statistik, které nenáleží do kritického oboru \mathcal{W} , dělený počtem simulací M . U každého testu (1)–(3) následně rozhodněte, zda je pro danou situaci (a)–(d) konzervativní, liberální, nebo ani jedno.

Vytvořte animaci demonstrující případnou konvergenci rozdělení testovacích statistik Z_W , U_S a U_{LR} k příslušnému rozdělení při rostoucí hodnotě rozsahu n . Hodnoty n volte $n = 10, 25, 30, 50, 100, 150, 250, 500$.

Požadovaná forma výstupu příkladu:

- Vlastnoručně naprogramovaná funkce `LambdaTest()`, která pro zvolený typ rozdělení (`distr = 'poiss', 'mix.pois', 'neg.binom', 'binom'`) vykreslí příslušnou trojici histogramů testovacích statistik Z_W , U_S a U_{LR} . Přesnější popis histogramů viz níže. Povinný vstup funkce bude `lambda = ...`; volitelné vstupy budou `lambda0 = ...`, `p = 0.9`, `mix.lambda = 1.3`, `neg.k = 2.5`, `neg.p = ...`, `bin.N = 25`, `bin.p = ...`, `M = ...`, `n = ...`, `alpha = ...`, `distr = ...`, `plot = T`.
- Čtyři animace zobrazující chování testovacích statistik Z_W , U_S a U_{LR} při $n \rightarrow \infty$. V jednom kroku animace se zobrazí vedle sebe histogram Z_W , histogram U_S a histogram U_{LR} spočítané z jedné sady M náhodných výběrů o rozsazích n . Každý histogram bude superponován křivkou příslušného rozdělení. Pod každým histogramem budou (na ose x) čtyři popisky. První popisek ponese název použité testovací statistiky. Druhý popisek ponese údaj o rozsahu náhodného výběru n (hodnota se bude měnit podle zvoleného rozsahu n) a o hodnotě parametru λ_0 (hodnota se bude měnit podle parametru λ_0). Třetí popisek ponese název zvoleného rozdělení s konkrétními hodnotami parametrů. Například pro situaci (a) se vypíše popisek $X \sim \text{Poiss}(2.5)$. Popisek se bude měnit automaticky v závislosti na zvolené hodnotě argumentu `distr`, tj. podle zvoleného typu rozdělení (a)–(d). Dále se budou automaticky měnit hodnoty parametrů v popisku, dojde-li ke změně hodnot vstupních argumentů funkce `LambdaTest()`. Čtvrtý popisek bude ve tvaru $\text{Pr}(\text{pokrytí}) = \dots$, kde za (...) bude dosazena hodnota empirické pravděpodobnosti pokrytí.
- Podrobné komentáře hodnotící, zda pro vybrané rozdělení (a)–(d) dochází k uspokojivé shodě skutečného rozdělení testovací statistiky (1)–(3) s příslušným předpokládaným rozdělením. Hodnocení uveďte konkrétně ke každé z 12 kombinací: rozdělení (a)–(d) \times testovací statistika (1)–(3). Dále uveďte, jak se kvalita shody skutečného a předpokládaného rozdělení mění s $n \rightarrow \infty$ (vždy uveďte, zda se kvalita shody zlepšuje, zhoršuje, či nemění).
- Zaměřte se na situaci (d) a zavzpomínejte na poznatky ze Statistické inference I. Čím si vysvětlujete vysokou kvalitu shody skutečného a teoretického rozdělení testovacích statistik Z_W , U_S a U_{LR} v situaci, kdy testujeme H_0 o parametru λ Poissonova rozdělení, ačkoli náhodná veličina X pochází z binomického rozdělení? Jde o náhodu, nebo to lze logicky vysvětlit? Uveďte svůj názor :).
- Slovní vyjádření, zda pro $n = 100$ je pro vybrané rozdělení (a)–(d) test (1)–(3) konzervativní, liberální, nebo ani jedno. Slovní vyjádření uveďte konkrétně ke každé z 12 kombinací: rozdělení (a)–(d) \times testovací statistika (1)–(3). Pro každou z 12 kombinací také uveďte, jak se míra konzervativnosti, liberálnosti (případně ani jednoho) mění při $n \rightarrow \infty$ (test se stává více konzervativní, více liberální, je stále stejný, mění se z konzervativního na liberální, apod.). Vždy uveďte, jak si případnou změnu vysvětlujete, nebo naopak, jak si vysvětlujete, že ke změně s $n \rightarrow \infty$ nedochází?