Ústav matematiky a statistiky
Přírodovědecká fakulta
Masarykova univerzita

Aplikovaná statistika pro antropology I

Zadání zápočtového domácího úkolu
podzimní semestr 2019

Skupina A

Veronika Bendová

1. prosince 2019

Pokyny k řešení domácího úkolu

Domácí úkol sestává z šesti příkladů. Za vyřešení příkadů lze získat $6 + 6 + 6 + 9 + 13 + 8 = 48$ bodů + 8 bodů za celkovou úpravu a přehlednost úkolu, úpravu kódu, komentáře k postupům, apod. Celkem lze tedy získat 56 bodů.

Aby byl úkol uznán za splněný, je potřeba získat alespoň **42 bodů (75 %)**. Pokud student potřebných 42 bodů nezíská, bude mu úkol navrácen k opravě a dořešení příkladů na potřebný počet bodů. Pokud student ani po přepracování úkolu potřený počet bodů nezíská, nebude mu udělen zápočet. (Další, v pořadí druhé, přepracování úkolu nebude umožněno.)

Kompletní řešení domácího úkolu vložte, prosím, do odevzdávárny k předmětu MAS10c (cvičení z AS) nejpozději do 10. 12. 2019 23:59.

Kompletním řešením domácího úkolu je míněno dodání **zcela funkčního** R-Skriptu s názvem AS-2019-skupina-X-prijmeni-jmeno.R. Namísto X vložte verzi zadaného domácího úkolu (A nebo B). Zaslaný R-Skript bude obsahovat veškeré potřebné komentáře, popisy postupů, závěry testování a interpretace výsledků ve formátu R-kových komentářů. Před odesláním R-skriptu do odevzdávárny vyčistěte workspace (V RStudiu: Session → Clear Workspace) a všechny příkazy finálně projděte ještě jednou, abyste měli jistotu, že vše funguje, jak má. **Příklady, jejichž RSkript bude vyhazovat chybové hlášky, nebudou kontrolovány a automaticky budou vráceny k přepracování.**

Při vytváření řešení domácího úkolu se, prosím, striktně držte následujících pravidel:

- Na domácí úkol si vyhraďte dost času, pracujte na něm průběžně. Řešení úkolu není možné kvalitně zpracovat během jednoho či dvou dnů.
- Domácí úkol je vaší **samostatnou prací** a nahrazuje písemný test. Nepoužívejte kód, ani jeho části (týká se i částí obsahujících komentáře a interpretace výsledků) z řešení vašich spolužáků. Budou-li se kódy dvou řešení v libovolné části řešení shodovat, budou oba hodnoceny známkou **N**. Taktéž, bude-li se v kódu vyskytovat pasáž, která prokazatelně nezapadá konceptu kódu, bude úkol též hodnocen známkou N. Nárok na **zápočet** v takových případech **zaniká**.
- Striktně dodržte název odevzdávaného RSkriptu.
- Názvy datových souborů zanechte v původním znění, nepřejmenovávejte je.
- U jednotlivých úkolů, kde máte zjistit konkrétní výsledky, napište vaše výsledky stručně do komentářů za #. V celém Rskriptu (i v popiscích grafů) se vyvarujte diakritiky. Kódy s diakritikou budou automaticky **navráceny k přepracování**.
- Interpretace výsledků jsou nedílnou součástí příkladu a jsou hodnoceny celkem vysokým počtem bodů. **Absence interpretací výsledků tedy výrazně snižuje celkový počet bodů** z jinak správně vypracovaného příkladu.
- Při programování dodržujte jistou **přehlednost kódu**. Před a za symbolem <- uveďte vždy mezeru, taktéž jednotlivé argumenty funkcí oddělujte mezerami. Příklad správně a přehledně naprogramovaného kódu je k náhledu níže. Správné naprogramování kódu je v rámci úkolu bodově hodnoceno.

```
x  <- 1:15
px <- dbinom(x, size = 15, p = 0.5)

plot(x, px, type = 'h', lty = 2, lwd = 1,
     main = 'Pravdepodobnostni funkce binomickeho rozdeleni',
     cex.main = 0.9)
points(x, px, pch = 21, col = 'red', bg = 'salmon')

legend('topright', fill = c('salmon'), legend = c('binom'), bty = 'n')
```

A na závěr pár doporučení a komentářů k zadání nebo k řešení úkolu:

- Zadání příkladů mohou obsahovat nadbytečné informace, které nejsou k řešení úkolu potřeba. Stejně tak datové soubory 30-goldman-alaska.csv a 30-goldman-poundbury.csv obsahují větší množství údajů, než jaké k vyřešení daného příkladu potřebujeme. Vždy je tedy třeba z datového souboru správně vybrat pouze údaje, které jsou potřebné k řešení příkladu.
- Názvy proměnných volte vždy tak, aby vystihovaly svůj obsah (rozhodně se vyvarujte zdrobnělin, názvů jako aa, nejake.cislo, bhg, cosi, apod.).
- V některých příkladech jsou uvedeny tipy na funkce, jejichž použití vám pomůže s řešením vybraných částí úkolu. Pokud jsme funkce nebrali na cvičeních, je třeba si jejich syntaxi nastudovat formou samostudia.
- Při práci s datovými soubory je třeba odstranit chybějící pozorování. Nikdy však neodstraňujeme automaticky všechna chybějící pozorování z celého datového souboru, přicházeli bychom tím o cenná data. NA hodnoty odstraňujeme vždy až po vyselektování proměnných nezbytných k provedení analýzy.
- Je-li součástí příkladu stanovení hypotéz H_0 a H_1, je tím vždy myšlen matematický zápis, nikoli slovní zápis. Pouze matematický zápis je tedy bodově hodnocen. Výjimku tvoří testy normality, kde H_0 a H_1 zadáváme výhradně slovně.
- Při vypracování grafů se řiďte vzhledem grafů uvedených v zadání úkolu. Čím vyšší bude shoda výsledného grafu s grafem v zadání (kromě barev, které mohou být voleny libovolně, ale rozumně), tím více bodů za graf získáte.
- Při vypracování příkladů na testování hypotéz je potřeba jednotlivé testy provést manuálním výpočtem v Rku, nikoli použitím funkcí jako jsou var.test(), t.test(), apod. Tyto funkce lze použít maximálně jako kontrolu vašich výsledků.

Přeji vám hodně zdaru při řešení příkladů :).

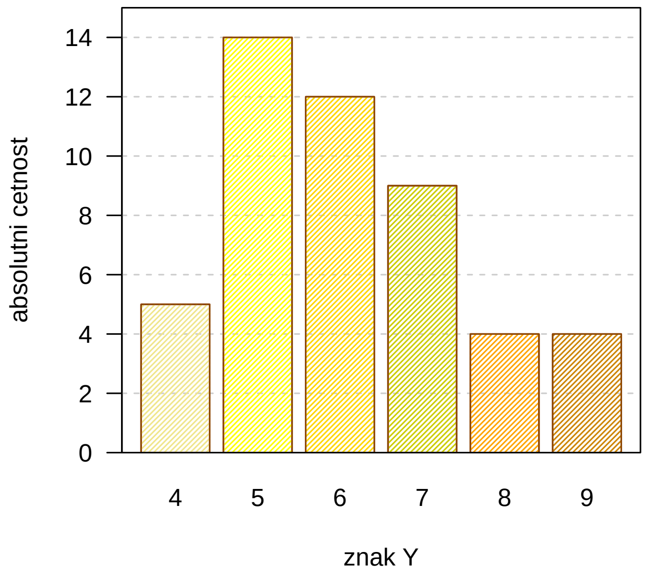
Příklad 1 (6 b). Znak Y nabývá variant 4, 5, 6, 7, 8, 9 s četnostmi 5, 14, 12, 9, 4, 4.

- Vypočítejte 5% kvantil $y_{0.05}$, dolní kvartil $y_{0.25}$, medián $y_{0.5}$, horní kvartil $y_{0.75}$ a 95 % kvantil $y_{0.95}$ znaku Y. Hodnoty vložte do přehledné tabulky a řádně je interpretujte.
- Vykreslete sloupcový diagram absolutních četností znaku Y.

Požadovaná forma výstupu příkladu:

1. Tabulka s hodnotami požadovaných pěti kvantilů $y_{0.05}$, $y_{0.25}$, $y_{0.50}$, $y_{0.75}$, $y_{0.95}$. $(0.5 + 5 \times 0.3 + 0.5 = \mathbf{2.5\ b})$
2. Samostatná interpretace každého kvantilu. $(5 \times 0.3 = \mathbf{1.5\ b})$
3. Sloupcový diagram absolutních četností. $(\mathbf{2\ b})$

```
  5% kvantil dolni kvartil median horni kvartil 95% kvantil
1          4             5      6             7           9
```



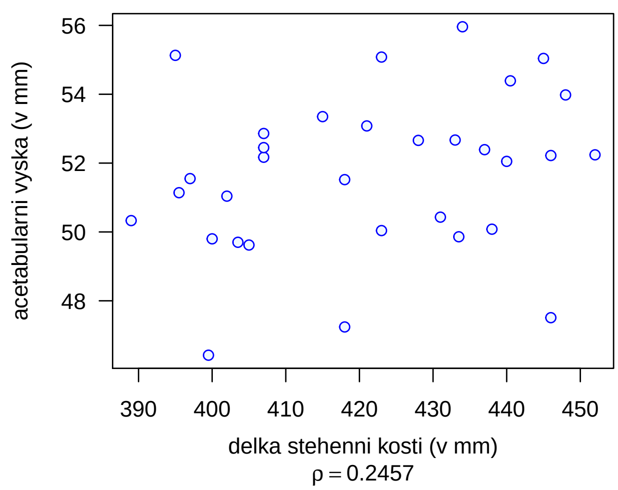
Příklad 2 (6 b). Máme k dispozici datový soubor 30-goldman-alaska.csv obsahující antropometrické údaje o délce kosti stehenní v mm (znak X spojitého typu (proměnná RFML)) a acetabulární výšce v mm (znak Y spojitého typu (proměnná RAcH)) z pravé strany u skeletů jedinců z aljašské populace (muži a ženy z kmenů Tigara a Ipituaq). Ze zadaných údajů byly dopočítány následující charakteristiky pro skelety **mužského** pohlaví: aritmetické průměry: $m_X = 421.1719\,\mathrm{mm}$, $m_Y = 51.6875\,\mathrm{mm}$; rozptyly: $s_X^2 = 18.6844^2\,\mathrm{mm}^2$, $s_Y^2 = 2.3072^2\,\mathrm{mm}^2$; kovariance: $s_{XY} = 10.5917$.

- Stanovte hodnotu odhadu korelačního koeficientu ρ a řádně ji interpretujte.
- Načtěte datový soubor 30-goldman-alaska.csv a vykreslete tečkový diagram zobrazující vztah délky kosti stehenní a acetabulární výšky pro skelety mužského pohlaví.

Požadovaná forma výstupu příkladu:

1. Název korelačního koeficientu, který jste vypočítali, a zdůvodnění, proč jste jej použili a proč je vhodnou statistikou použitelnou na stanovení míry závislosti mezi znaky X a Y. **(2 b)**
2. Výpočet korelačního koeficientu s výsledkem zaokrouhleným na čtyři desetinná místa. **(1.5 b)**
3. Kompletní interpretace vypočítaného koeficientu. **(1.5 b)**
4. Tečkový diagram. Součástí diagramu bude popisek (umístěný pod popiskem osy x) obsahující hodnotu vypočítaného korelačního koeficientu. Ten získáme pomocí příkazu mtext(bquote(paste(rho == .(rho))), side = ..., line = ...). **(1 b)**

```
[1] 0.2457
```



Příklad 3 (6 b). Máme k dispozici naměřené údaje o délce kyčelní kosti (v mm) z levé strany u mužských skeletů tří japonských populací (9 skeletů z populace Tsugumo Shell Mound, 7 skeletů z populace Yoshigo Shell Mound a 3 skelety z populace Yasaki Shell Mound). Ze zadaných údajů byly dopočítány následující charakteristiky: (a) Tsugumo SM: aritmetický průměr: $m_1 = 149.22\,\mathrm{mm}$; směrodatná odchylka: $s_1 = 5.67\,\mathrm{mm}$; (b) Yashigo SM: $m_2 = 151.00\,\mathrm{mm}$; $s_2 = 4.55\,\mathrm{mm}$; (c) Yasaki SM: $m_3 = 154.00\,\mathrm{mm}$; $s_3 = 2.00\,\mathrm{mm}$.

- Stanovte hodnotu váženého průměru výběrových rozptylů řádně ji interpretujte.
- Stanovte hodnotu variačního koeficientu $v = \frac{s}{m}$, kde s je výběrová směrodatná odchylka a m je výběrový průměr, pro délku levé kyčelní kosti mužských skeletů z populace Tsugumo Shell Mound. Na základě hodnoty koeficientu variace v zhodnoťte, jak velký je rozptyl vzhledem k aritmetickému průměru. Co nám hodnota koeficientu variace v říká o náhodném výběru?

Požadovaná forma výstupu příkladu:

1. Výpočet váženého průměru výběrových rozptylů s výsledkem zaokrouhleným na čtyři desetinná místa. (**2.5 b**)
2. Odpověď celou větou. (**0.5 b**)
3. Výpočet variačního koeficientu s výsledkem zaokrouhleným na čtyři desetinná místa. (**1 b**)
4. Odpovědi na dvě otázky. ($2 \times 1 =$ **2 b**)

```
[1] 24.3379
```
13

```
[1] 0.038
```
14

Příklad 4 (9 b). Předpokládejme, že diametrální rozměr hlavice pažní kosti u mužů je normálně rozdělený okolo střední hodnoty 45 mm s rozptylem 3.45^2 mm^2.

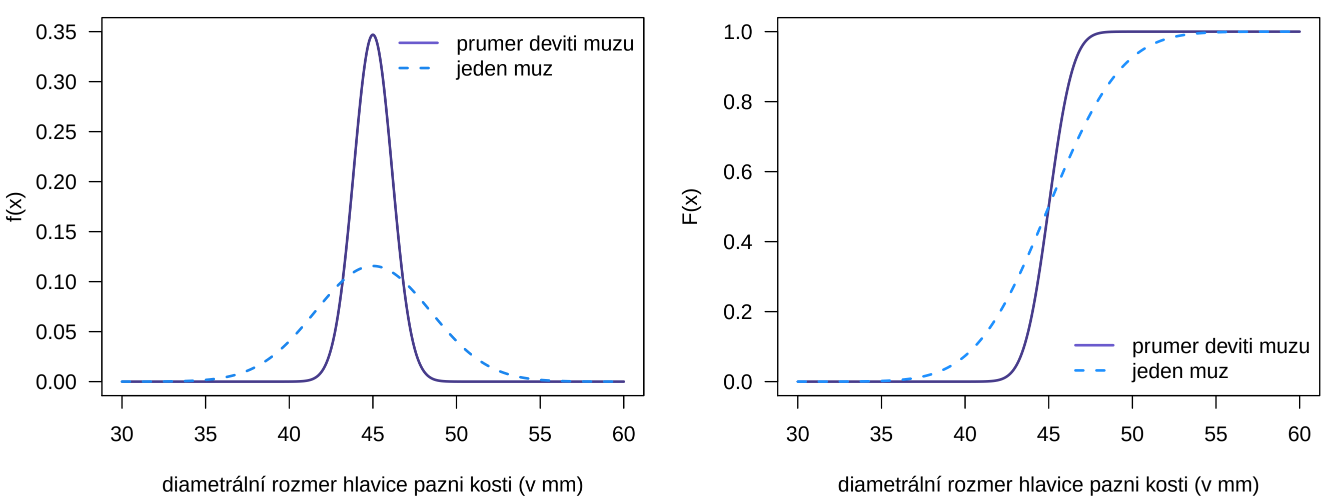
- (1) Jaká je pravděpodobnost, že **diametrální rozměr** hlavice pažní kosti náhodně vybraného muže bude alespoň 46.6 mm?
- (2) Jaká je pravděpodobnost, že **průměr diametrálního rozměru** hlavice pažní kosti devíti náhodně vybraných mužů bude alespoň 46.6 mm?
- Vykreslete graf hustoty normálního rozdělení průměru diametrálního rozměru hlavice pažní kosti devíti mužů. Na osu x naneste posloupnost 1000 hodnot od 30 mm do 60 mm a na osu y hodnoty hustoty normálního rozdělení průměru diametrálního rozměru devíti mužů ($\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$). Do grafu dokreslete také křivku hustoty normálního rozdělení pro diametrální rozměr jednoho muže ($n = 1$).
- Vykreslete graf distribuční funkce normálního rozdělení průměru diametrálního rozměru hlavice pažní kosti devíti mužů. Na osu x naneste posloupnost 1000 hodnot od 30 mm do 60 mm a na osu y hodnoty distribuční funkce normálního rozdělení průměru diametrálního rozměru devíti mužů ($\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$). Do grafu dokreslete také křivku distribuční funkce normálního rozdělení pro diametrální rozměr jednoho muže ($n = 1$).

Požadovaná forma výstupu příkladu:

1. Výpočet pravděpodobnosti + odpověď celou větou na otázku (1). (1 + 0.5 = **1.5 b**)
2. Výpočet pravděpodobnosti + odpověď celou větou na otázku (2). (1.5 + 0.5 = **2 b**)
3. Graf s dvěma křivkami funkcí hustot + legenda. ($2 \times 0.5 + 0.5 =$ **1.5 b**)
4. Graf s dvěma křivkami distribučních funkcí + legenda. ($2 \times 0.5 + 0.5 =$ **1.5 b**)
5. Podrobný popis obou grafů + popis propojení grafů s výsledky pravděpodobností (1) a (2). Jaký je vztah mezi křivkou hustoty pro průměr diametrálního rozměru hlavice pažní kosti devíti mužů a křivkou hustoty pro diametrální rozměr jednoho muže? Jakým způsobem souvisí tvary křivek hustot, resp. distribučních funkcí s vypočítanými pravděpodobnostmi? (**2.5 b**)

```
[1] 0.3214069
```
15

```
[1] 0.08206658
```
16



Příklad 5 (13 b). Máme k dispozici datový soubor 30-goldman-poundbury.csv obsahující antropometrické údaje o délce holenní kosti v mm (RTML a LTML) a délce stehenní kosti v mm (RFML a LFML) z pravé a levé strany u skeletů mužského a ženského pohlaví z římského pohřebiště v Poundbury. Na hladině významnosti $\alpha = 0.05$ testujte, zda mezi délkou holenní kosti z pravé a levé strany u žen existuje statisticky významný rozdíl.

Tip: Datový soubor obsahuje neznámé (tzv. NA) hodnoty. Po vyselektování sledovaných proměnných je potřeba řádky s NA hodnotami odstranit.

Požadovaná forma výstupu příkladu:

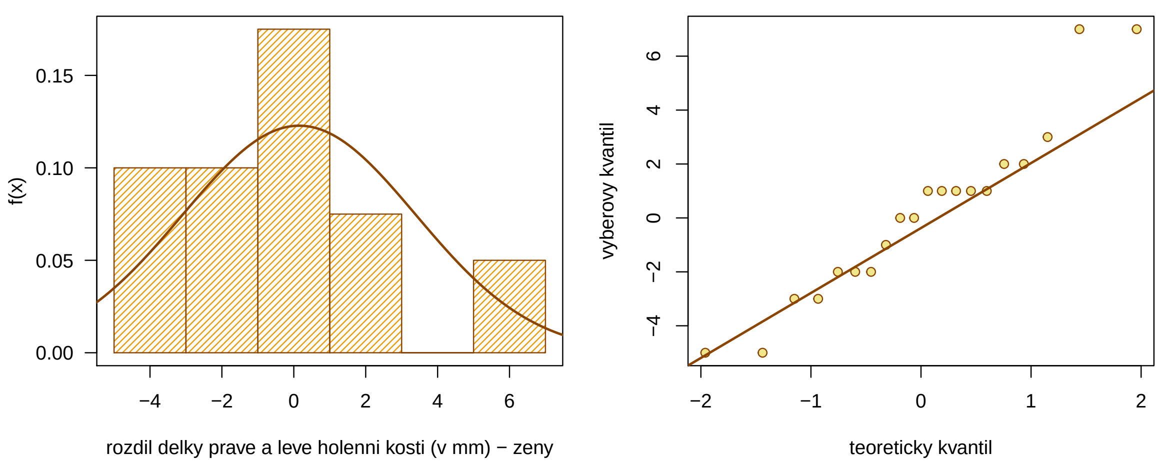
1. **Testování normality:** Správně zvolený test normality se zdůvodněním volby testu + H_0, H_1 + zdůvodněné rozhodnutí o zamítnutí/nezamítnutí H_0 + interpretace výsledku testování + grafická vizualizace normality dat (histogram + Q-Q graf).

$$((1 + 0.5 + 0.5 + 0.25 + (1.25 + 0.5)) = \mathbf{4\,b})$$

```
[1] 0.1948109
```
17

Poznámka: Histogram bude vykreslen se správným počtem třídicích intervalů (viz Sturgesovo pravidlo) a se zaznamenanými hodnotami středů třídicích intervalů. Dále bude superponován křivkou hustoty normálního rozdělení $N(\mu, \sigma^2)$, kde odhad parametrů μ a σ^2 získáte z dat.

Tip: Aby se vám křivka vykreslila správně, musíte v příkazu hist() zadat argument prob=T. Tento argument převede měřítko y-ové osy z absolutní škály (na ose y jsou defaultně nastaveny absolutní četnosti) na relativní škálu (na ose y budou relativní četnosti).



2. **Test hypotézy ze zadání**: Volba vhodného testu na základě charakteru dat a výsledku testu normality se zdůvodněním volby testu + H_0, H_1 + kompletní test (a) kritickým oborem; (b) intervalem spolehlivosti; (c) p-hodnotou se zdůvodněným rozhodnutím o zamítnutí/nezamítnutí H_0 (u všech tří typů testování) + interpretace výsledku testování.

$$(1 + 2 + 3 + 1 = \mathbf{7\,b})$$

```
[1] "Testovaci␣statistika:"
```
18

```
[1] 0.206477
```
19

```
[1] "Kriticky␣obor:"
```
20

```
[1] -2.093024
```
21

```
[1] 2.093024
```
22

```
[1] "Interval␣spolehlivosti:"
```
23

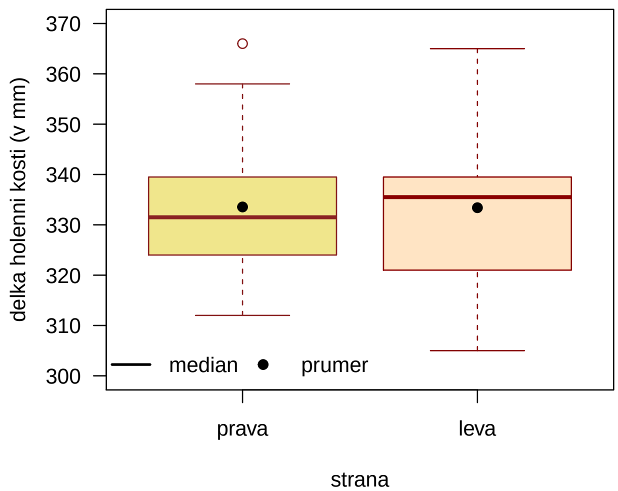
```
[1] 1.670526
```

```
[1] -1.370526
```

```
[1] "p-hodnota:"
```

```
[1] 0.8386148
```

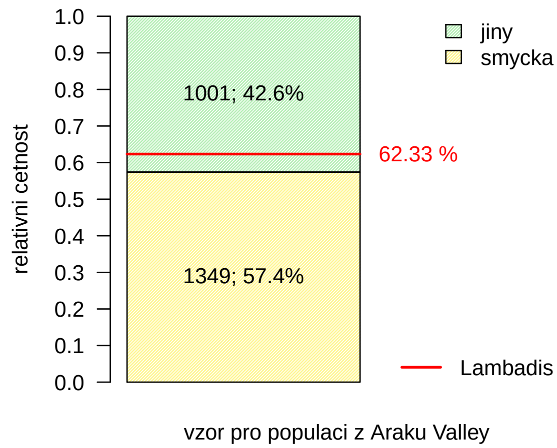
3. Krabicový diagram porovnávající délku holenní kosti u žen z pravé a levé strany. (**2 b**)



Příklad 6 (**8 b**). Mějme datový soubor 25-one-sample-probability-dermatoglyphs.txt obsahující údaje o frekvenci výskytu dermatoglyfického vzoru *smyčka* na 10 prstech 470 jedinců z populace Bagathů z Araku Valley (celkem 4700 otisků prstů). Současně máme k dispozici hodnotu pravděpodobnosti výskytu dermatoglyfického vzoru *smyčka* na prstech mužů a žen z populace Lambadis ($p_m = 0.5618$, $p_f = 0.6233$). Na hladině významnosti $\alpha = 0.01$ zjistěte, zda je u žen bagathské populace z Araku Valley menší frekvence výskytu dermatoglyfického vzoru *smyčka* než u žen z populace Lambadis.

Požadovaná forma výstupu příkladu:

1. Sloupcový graf relativních četností výskytu dermatoglyfického vzoru *smyčka* u žen bagathské populace z Araku Valley. Do grafu doplňte také referenční čáru pro pravděpodobnost výskytu vzoru *smyčka* u žen z populace Lambadis (příkaz segments()) s popiskem obsahujícím hodnotu pravděpodobnosti v procentuální škále (příkaz text()). (0.5 + 0.5 = **1 b**)



2. **Ověření Haldovy podmínky dobré aproximace:** Výpočet + závěr (podmínka je/není splněná). ($1 + 0.5 = \mathbf{1.5\ b}$)

3. **Test o pravděpodobnosti:** H_0, H_1 + kompletní test (a) kritickým oborem; (b) intervalem spolehlivosti; (c) p-hodnotou se zdůvodněným rozhodnutím o zamítnutí/nezamítnutí H_0 (u všech tří typů testování) + interpretace výsledku testování. ($2 \times 1 + 3 + 0.5 = \mathbf{5.5\ b}$)

```
[1] "Testovaci statistika:"
```

```
[1] -4.927872
```

```
[1] "Kriticky obor:"
```

```
[1] -2.326348
```

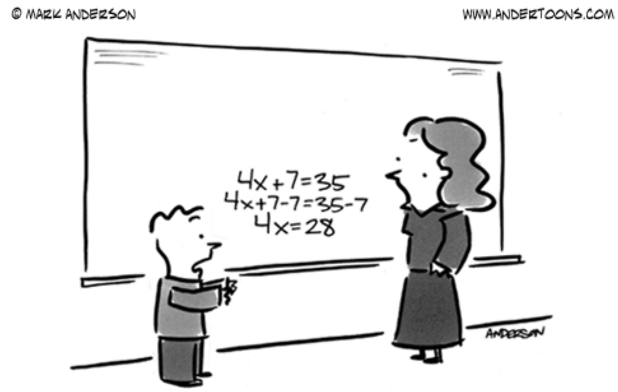
```
[1] "Interval spolehlivosti:"
```

```
[1] 0.5977725
```

```
[1] "p-hodnota:"
```

```
[1] 4.156494e-07
```

School Cartoon #6446



"You *knew* X was 7 the *whole time* and you *never said anything*?!"