Ústav matematiky a statistiky
Přírodovědecká fakulta
Masarykova univerzita

Aplikovaná statistika pro antropology I

Zadání zápočtového domácího úkolu
podzimní semestr 2019

Skupina B

Veronika Bendová

28. listopadu 2019

Pokyny k řešení domácího úkolu

Domácí úkol sestává z pěti příkladů. Za vyřešení příkadů lze získat $6+6+6+9+21 = 48$ 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 (rozhoně 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 :).

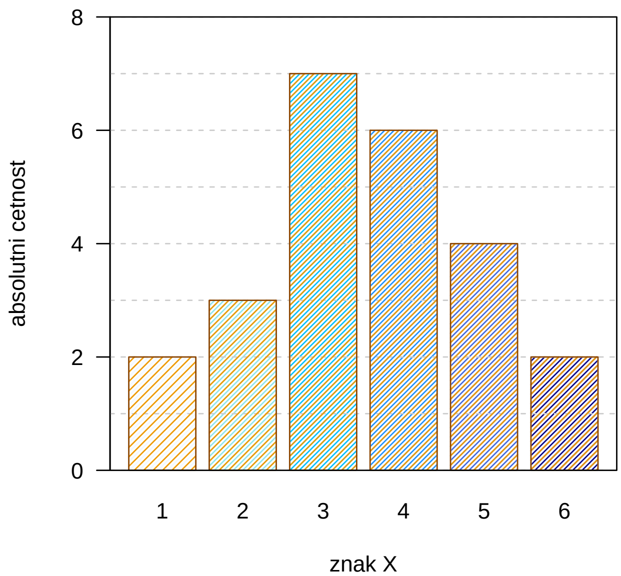
Příklad 1. (**6 b**) Znak X nabývá hodnot 4, 3, 1, 3, 3, 6, 4, 5, 5, 5, 2, 6, 2, 3, 4, 4, 2, 3, 3, 3, 4, 5, 4, 1.

- Vypočítejte druhý decil $x_{0.20}$, dolní kvartil $x_{0.25}$, medián $x_{0.5}$, horní kvartil $x_{0.75}$ a osmý decil $x_{0.80}$ znaku X. Hodnoty vložte do přehledné tabulky a řádně je interpretujte.
- Vykreslete sloupcový diagram absolutních četností znaku X.

Požadovaná forma výstupu příkladu:

1. Tabulka s hodnotami požadovaných pěti kvantilů $x_{0.20}$, $x_{0.25}$, $x_{0.50}$, $x_{0.75}$, $x_{0.80}$. $(0.5 + 5 \times 0.3 + 0.5 = \mathbf{2.5\,b})$
2. Samostatná interpretace každého kvantilu. $(5 \times 0.3 = \mathbf{1.5\,b})$
3. Sloupcový diagram absolutních četností. $(\mathbf{2\,b})$

```
  2. decil dolni kvartil median horni kvartil 8. decil
1        2             3    3.5           4.5        5
```



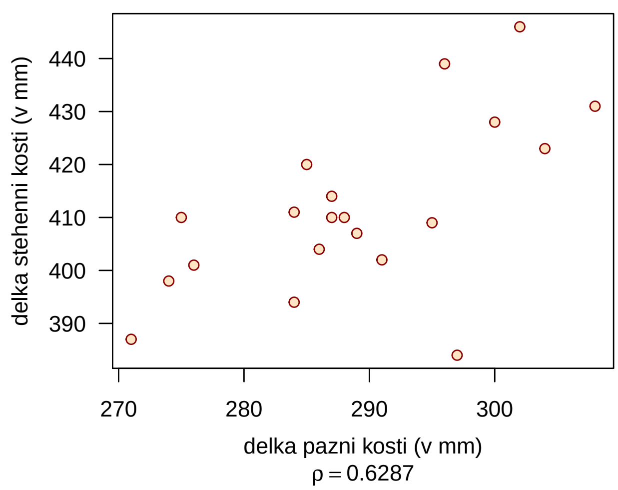
Příklad 2 (6 b). Máme k dispozici datový soubor 30-goldman-poundbury.csv obsahující antropometrické údaje o délce kosti pažní v mm (znak X spojitého typu (proměnná LHML)) a délce kosti stehenní v mm (znak Y spojitého typu (proměnná LFML)) z levé strany u skeletů z římského pohřebiště v Poundbury. Ze zadaných údajů byly dopočítány následující charakteristiky pro skelety ženského pohlaví: aritmetické průměry: $m_X = 288.9500$ mm, $m_Y = 411.4000$ mm; směrodatné odchylky: $s_X = 10.3287$ mm, $s_Y = 16.1323$ mm; kovariance: $s_{XY} = 104.7579$.

- Stanovte hodnotu odhadu korelačního koeficientu ρ a řádně ji interpretujte.
- Načtěte datový soubor 30-goldman-poundbury.csv a vykreslete tečkový diagram zobrazující vztah délky pažní kosti a stehenní kosti pro skelety ženského 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] 0.6287
```



Příklad 3 (6 b). Máme k dispozici naměřené údaje o acetabulární výšce (v mm) z pravé strany u mužských skeletů ze tří pohřebišť na území Nového Mexika (19 skeletů s pohřebiště Hawikuh, 4 skelety z pohřebiště Pueblo Bonito a 7 skeletů z pohřebiště Puye). Ze zadaných údajů byly dopočítány následující charakteristiky: (a) Hawikuh: aritmetický průměr: $m_1 = 47.98\,\text{mm}$; rozptyl: $s_1^2 = 2.15^2\,\text{mm}^2$; (b) Pueblo Bonito: $m_2 = 51.08\,\text{mm}$; $s_2^2 = 1.83^2\,\text{mm}^2$; (c) Puye: $m_3 = 46.20\,\text{mm}$; $s_3^2 = 2.73^2\,\text{mm}^2$.

- 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 acetabulární výšku z pravé strany mužských skeletů z pohřebiště Puye. 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 × 1 = **2 b**)

```
[1] 5.11
```
13

```
[1] 0.0591
```
14

Příklad 4 (9 b). Předpokládejme, že délka holenní kosti u žen je normálně rozdělená okolo střední hodnoty 333 mm se směrodatnou odchylkou 22 mm.

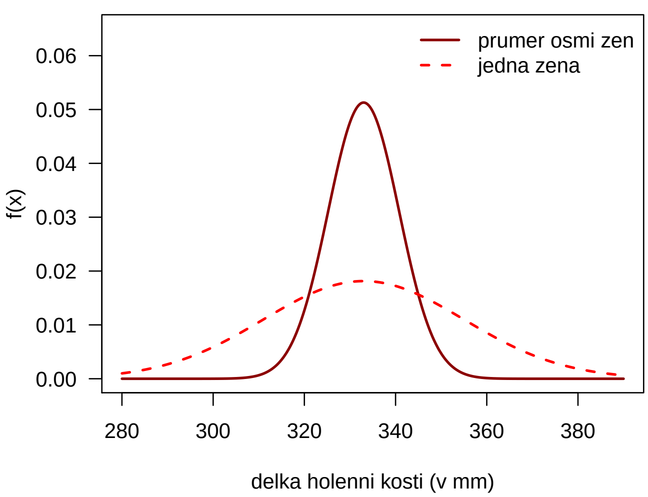
- (1) Jaká je pravděpodobnost, že **délka holenní kosti** náhodně vybrané ženy bude nejvýše 340 mm?
- (2) Jaká je pravděpodobnost, že **průměrná délka holenní kosti** osmi náhodně vybraných žen bude nejvýše 340 mm?
- Vykreslete graf hustoty normálního rozdělení průměrné délky holenní kosti u osmi žen. Na osu x naneste posloupnost 1000 hodnot od 280 mm do 390 mm a na osu y hodnoty hustoty normálního rozdělení průměrné délky holenní kosti u osmi žen ($\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$). Do grafu dokreslete také křivku hustoty normálního rozdělení pro délku holenní kosti pro jednu ženu ($n = 1$).
- Vykreslete graf distribuční funkce normálního rozdělení průměrné délky holenní kosti u osmi žen. Na osu x naneste posloupnost 1000 hodnot od 280 mm do 390 mm a na osu y hodnoty distribuční funkce normálního rozdělení průměrné délky holenní kosti u osmi žen ($\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$). Do grafu dokreslete také křivku distribuční funkce normálního rozdělení pro délku holenní kosti jedné ženy ($n = 1$).

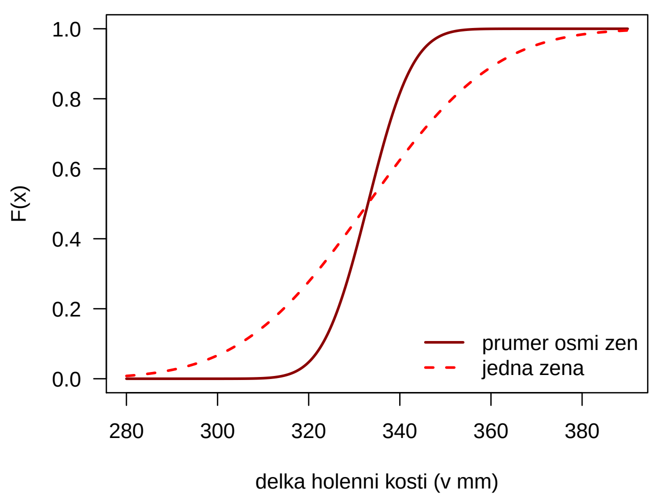
Požadovaná forma výstupu příkladu:

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 hustoty + 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 průměrnou délku holenní kosti osmi žen a křivkou hustoty pro délku holenní kosti jedné ženy? Jakým způsobem souvisí tvary křivek hustot, resp. distribučních funkcí s vypočítanými pravděpodobnostmi? (**2.5 b**)

```
[1] 0.6248265
```
15

```
[1] 0.8159277
```
16





Příklad 5 (21 b). Máme k dispozici datový soubor 30-goldman-alaska.csv obsahující antropometrické údaje o acetabulární výšce z pravé strany (proměnná RAcH) a z levé strany (proměnná LAcH) u skeletů jedinců z aljašské populace (muži a ženy z kmenů Tigara a Ipituaq).

Na hladině významnosti $\alpha = 0.05$ ověřte, zda je acetabulární výška z pravé strany u žen z kmene Tigara větší než u žen z kmene Ipituaq.

Tip: Datový soubor obsahuje neznámé (tzv. NA) hodnoty. Před řešením příkladu je vhodné tyto hodnoty ze sledovaných proměnných odstranit.

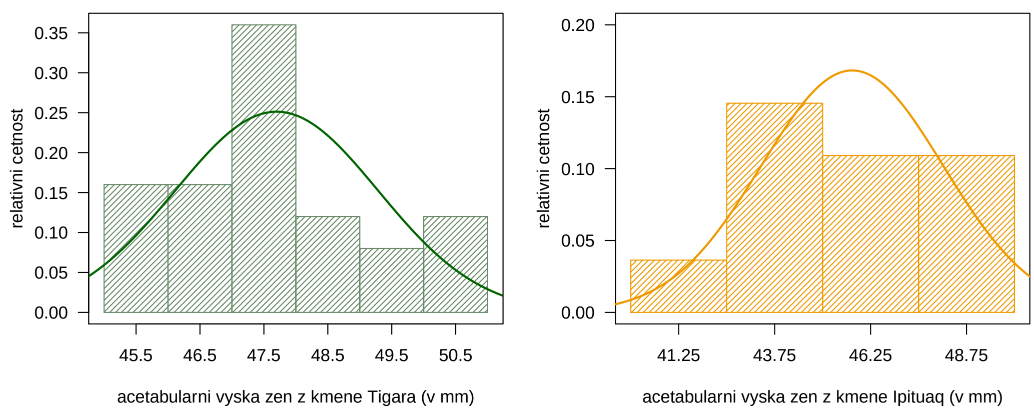
Požadovaná forma výstupu příkladu:

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 (zvlášt pro populaci z kmene Tigara a zvlášť pro populaci z kmene Ipituaq).

$$((0.5 + 0.5 + 0.5 + 0.25 + 1.25 + 0.5) \times 2 = \mathbf{7\,b})$$

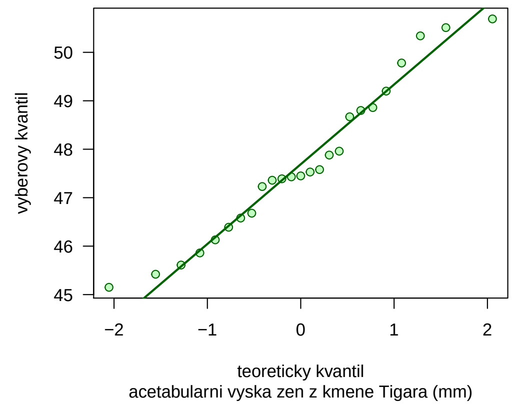
Poznámka: Histogramy budou vykresleny se správným počtem třídicích intervalů (viz Sturgesovo pravidlo) a se zaznamenanými hodnotami středů třídicích intervalů. Histogram pro acetabulární výšku pro ženy z kmene Tigara bude superponován křivkou normálního rozdělení $N(\mu_1, \sigma_1^2)$, kde odhad parametrů μ_1 a σ_1^2 získáte z dat. Histogram pro acetabulární výšku pro ženy z kmene Ipituaq bude superponován křivkou normálního rozdělení $N(\mu_2, \sigma_2^2)$, kde odhad parametrů μ_2 a σ_2^2 získáte z dat.

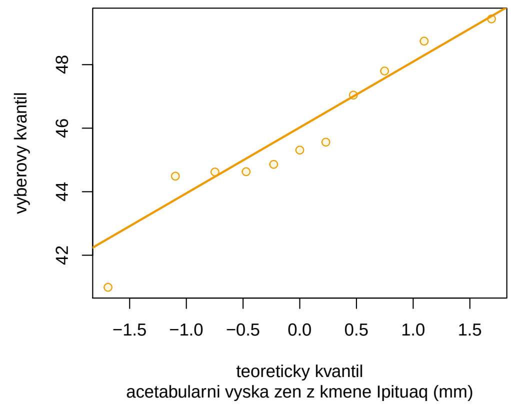
Tip: Aby se vám křivky vykreslily správně, musíte v příkazu hist() zadat argument prob=T. Tento argument převede měřítko y-ové osy z absolutní škály (na ose y jsou defaultně nastaveny absolutní četnosti) na relativní škálu (na ose y budou relativní četnosti).




```
[1] 0.3715629
```
17

```
[1] 0.4749806
```
18





2. **Test o shodě rozptylů σ_1^2 a σ_2^2:** Hladinu významnosti zvolte $\alpha = 0.05$. Stanovené hypotézy 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.

$(2 \times 0.5 + 3 + 1 = \mathbf{5\,b})$

```
[1] "Testovaci␣statistika:"
```
```
[1] 0.448568
```
```
[1] "Kriticky␣obor:"
```
```
[1] 0.3788467
```
```
[1] 3.365369
```
```
[1] "Interval␣spolehlivosti:"
```
```
[1] 0.1332894
```
```
[1] 1.184036
```
```
[1] "p-hodnota:"
```
```
[1] 0.104749
```

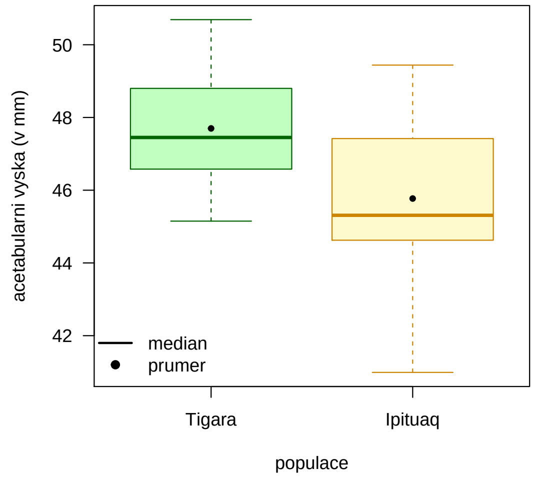
3. **Test hypotézy ze zadání:** Volba vhodného testu na základě výsledků testů normality a testu o shodě rozptylů se zdůvodněním volby testu + H_0, H_1 + kompletní test (a) kritickým oborem; (b) intervalem spolehlivosti; (c) p-hodnotou se zdůvodněným rozhodnutím o zamítnutí/nezamítnutí H_0 (u všech tří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

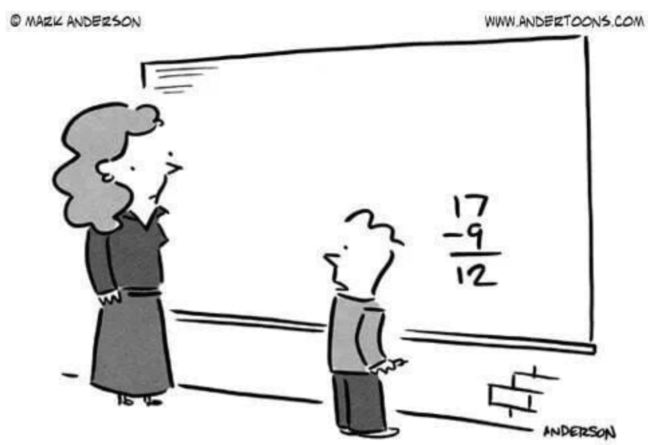
$(1 + 2 \times 1 + 3 + 1 = \mathbf{7\,b})$

```
[1] "Testovaci␣statistika:"
```
```
[1] 2.876229
```
```
[1] "Kriticky␣obor:"
```
```
[1] 1.690924
```
```
[1] "Interval␣spolehlivosti:"
```
```
[1] 0.7946559
```
```
[1] "p-hodnota:"
```

```
[1] 0.00345072
```
36

4. Krabicový diagram porovnávající acetabulátní výšku z pravé strany u žen z kmene Tigara a u žen z kmene Ipituaq. (**2 b**)





"I know it's wrong, I'm just waiting for the autocorrect."