

# Úvod do testování hypotéz

- Datový soubor = Náhodný výběr → stanovíme předpoklady → ověřujeme, zda platí;
- předpoklady
  - o charakteristikách:  $\mu$ ,  $\sigma^2$ ,  $\sigma$ ,  $p$ , ...
  - o rozdělení: normální, Poissonovo, binomické, ...
  - o nezávislosti dvou znaků, ...

## Postup testování hypotéz:

1. literární rešerše, formulace problému ... přesná, jednoznačná
2. stanovení nulové hypotézy  $H_0$ 
  - hypotéza o níž test rozhodne, zda se zamítne, nebo ne
    - 1 náhodný výběr a publikovaná hodnota  $c$ ;  $H_0 : \mu = c$
    - 2 náhodné výběry se středními hodnotami  $\mu_1$  a  $\mu_2$ ;  $H_0 : \mu_1 = \mu_2$
3. stanovení alternativní hypotézy  $H_1$ 
  - alt. hypotézu přijímáme, pokud  $H_0$  zamítáme
    - $H_{11} : \mu_1 \neq c$  (oboustranná alt.);
    - $H_{12} : \mu_1 < c$  (levostranná alt.);
    - $H_{13} : \mu_1 > c$  (pravostranná alt.).
4. volba hladiny významnosti  $\alpha$ 
  - pst(riziko), že  $H_0$  zamítneme, když platí - snažíme se tuto hodnotu snížit na minimum
5. provedení měření; sběr dat
6. testování  $H_0$  (tři různé způsoby):
  - Kritický obor
  - Interval spolehlivosti
  - p-hodnota
7. rozhodnutí o zamítnutí/nezamítnutí  $H_0$
8. interpretace výsledků

# Přístupy k testování nulové hypotézy $H_0$

## Testování pomocí kritického oboru

- Testujeme hypotézu  $H_0 : \theta = c$  oproti  $H_1 : \theta \neq c$ , případně  $H_{12} : \theta < c$ , či  $H_{13} : \theta > c$
- vybereme vhodnou testovací statistiku  $T_0$
- vypočítáme hodnotu testovací statistiky  $t_0$
- stanovíme kritický obor  $W$ :
  - oboustranná alt.:  $W = (T_{min}; K_{\alpha/2}) \cup (K_{1-\alpha/2}; T_{max})$
  - levostranná alt.:  $W = (T_{min}; K_{\alpha})$
  - pravostranná alt.:  $W = (K_{1-\alpha}; T_{max})$
- Pokud  $t_0 \in W$ ,  $H_0$  zamítáme na hladině významnosti  $\alpha$ .

## Testování pomocí IS:

- Testujeme hypotézu  $H_0 : \theta = c$  oproti  $H_1 : \theta \neq c$ , případně  $H_{12} : \theta < c$ , či  $H_{13} : \theta > c$
- Sestrojíme  $100(1 - \alpha)\%$  IS:
  - oboustranná alt.  $H_{11} \rightarrow$  oboustranný IS
  - levostranná alt.  $H_{12} \rightarrow$  pravostranný IS
  - pravostranná alt.  $H_{13} \rightarrow$  levostranný IS
- pokud  $c \in IS$ ,  $H_0$  nezamítáme na hladině významnosti  $\alpha$ .

## Testování pomocí p-hodnoty


- Testujeme hypotézu  $H_0 : \theta = c$  oproti  $H_1 : \theta \neq c$ , případně  $H_{12} : \theta < c$ , či  $H_{13} : \theta > c$
- p-hodnota:
  - pro oboustrannou alt.  $H_{11}: p = 2 \min\{P(T_0 \leq t_0); P(T_0 > t_0)\}$
  - pro levostrannou alt.  $H_{12}: p = P(T_0 \leq t_0)$
  - pro pravostrannou alt.  $H_{13}: p = P(T_0 > t_0) = 1 - P(T_0 \leq t_0)$
- Je-li  $p \leq \alpha$ ,  $H_0$  zamítáme na hladině významnosti  $\alpha$ .

## 6 Testy normality

### 6.1 Testy jednorozměrné normality

Jak jsme si uvedli výše, jednorozměrná normalita náhodného výběru je stěžejním předpokladem umožňujícím testování nulové hypotézy o parametru střední hodnoty  $\mu$ , o parametru rozptylu  $\sigma^2$ , o rozdílu středních hodnot  $\mu_1 - \mu_2$  i o podílu rozptylů  $\sigma_1^2/\sigma_2^2$ ) pomocí parametrických testů. Dříve než použijeme libovolný ze zmíněných testů, musíme ověřit, zda námi naměřená data objektivně pochází z normálního rozdělení.

Nechť  $X_1, \dots, X_N$  je náhodný výběr z nějakého rozdělení  $L(\theta)$ , kde  $\theta$  je obecně vektor parametrů rozdělení  $L$ . Na hladině významnosti  $\alpha = 0.05$  testujeme hypotézu  $H_0$ : *Náhodný výběr pochází z normálního rozdělení.*, tj.  $X \sim N(\mu, \sigma^2)$ , tj.  $L(\theta) = N(\mu, \sigma^2)$ , kde  $\theta = (\mu, \sigma^2)^T$ , oproti alternativní hypotéze  $H_1$ : *Náhodný výběr nepochází z normálního rozdělení.*, tj.  $X \not\sim N(\mu, \sigma^2)$ .

Volbu vhodného testu normality určujeme primárně podle rozsahu náhodného výběru. Je-li rozsah náhodného výběru  $n \leq 30$ , použijeme na testování hypotézy  $H_0$  Shapiroův-Wilkův test. K tomu využijeme funkci `shapiro.test()`, která je součástí základní knihovny `stats`. Je-li rozsah náhodného výběru  $n > 30$ , použijeme k testování hypotézy Lillieforsův test implementovaný ve funkci `lillie.test()` `lillie.test`, Anderson-Darlingův test implementovaný ve funkci `ad.test()` nebo Cramerův-von Misesův test implementovaný ve funkci `svm.test()`. Všechny zmíněné funkce `lillie.test()`, `ad.test()`, `cvm.test()` a `sf.test()` pochází z balíčku `nortest`. Přesnými algoritmy testů se zabývat nebudeme, vystačíme si pouze se znalostí zmíněných funkcí implementovaných v softwaru . Všechny funkce zmíněné v tomto odstavci mají podobný výstup sestávající vždy z názvu použitého testu normality, názvu proměnné obsahující data, jejichž normalitu testujeme, hodnotu testovací statistiky testu a  $p$ -hodnotu. Právě na základě  $p$ -hodnoty rozhodneme o zamítnutí nebo nezamítnutí nulové hypotézy o normalitě náhodného výběru.

Pro lepší představu o skutečném rozdělení náhodného výběru doprovázíme častokrát výsledky testování normality vhodnými grafy. Od takových grafů očekáváme, že, kromě zobrazení skutečného tvaru dat, nám ukáží, zda má náhodný výběr charakter normálního rozdělení, či nikoliv. Jedním z grafů je histogram se stanoveným počtem třídících intervalů podle Sturgerova pravidla (viz kapitola ??), superporovaný křivkou hustoty normálního rozdělení  $N(\mu, \sigma^2)$ . V případě, že datový soubor pochází z normálního rozdělení, kopíruje tvar histogramu křivku hustoty normálního rozdělení. Čím více je tvar histogramu odlišný od tvaru křivky hustoty, tím více je pravděpodobné, že data z normálního rozdělení nepochází.

Druhým grafem je QQ-diagram, který zachycuje na ose  $x$  hodnoty kvantilů předpokládaného rozdělení náhodného výběru (v našem případě tedy normálního rozdělení) a na ose  $y$  hodnoty výběrových kvantilů vypočítaných na základě náhodného výběru. V případě, že náhodný výběr pochází z normálního rozdělení, je charakter teoretických kvantilů normálního rozdělení podobný charakteru výběrových kvantilů. Postavením takových kvantilů v jednom grafu proti sobě vede potom na tvar přímky. Naopak, pokud data nepochází z normálního rozdělení je charakter výběrových kvantilů odlišný od charakteru teoretických kvantilů a postavením takových kvantilů v jednom grafu proti sobě vede na křivku, která nemá tvar přímky. Nejčastějším tvarem ukazujícím na porušení normality dat bývá esovitý tvar křivky s body výrazně odlehlymi od referenční křivky na pravém nebo levém chvostu náhodného výběru.

*Poznámka:* Důvod, proč si zde prezentujeme tři testy normality, které je možné použít v případě, že rozsah náhodného výběru  $n \geq 30$  je, abychom čtenáři poskytli širší spektrum možných testů normality a umožnili mu vybrat si z testů takový, který mu přijde nejsympatičtější. I z toho důvodu u každého z následujících příkladů provádíme testování nulové hypotézy o normalitě naměřených dat pomocí Lillieforsova testu, Andersonova-Darlingova testu i Cramérova-von Misesova testu. Ve vlastních analýzách je ovšem situace jiná. Zde zpravidla volíme na otestování hypotézy o normalitě pouze jeden test, který používáme univerzálně pokaždé, když chceme hypotézu o normalitě naměřených dat testovat. V žádném případě neprovádíme ve vlastních analýzách všechny testy normality najednou a v případě rozporuplných závěrů testů si nevybíráme ten, jehož výsledek se nám aktuálně nejvíce hodí. Vždy mějme na paměti, že aby vlastní analýza nepůsobila účelově, nebo dokonce nebyla demagogií, musíme zvolené testy volit univerzálně a bez ohledu na to, jak příhodný či nepříhodný výsledek nám v konkrétní situaci předloží.

Více testů normality smíme provádět pouze v hraničních případech, kdy univerzálně zvolený test hypotézu o normalitě hraničně zamítá nebo nezamítá. Pak je vhodné provést i další testy, které nám mohou pomoci nahlédnout na normalitu náhodného výběru z jiného pohledu a lépe tak zvážit skutečný charakter rozdělení naměřených hodnot. Pokud se na základě dalších testů rozhodneme pro jiný závěr, než jaký nám poskytuje test univerzálně používaný na otestování hypotézy o normalitě v celé datové analýze, je vždy třeba svůj závěr podložit relevantními výsledky a, je-li to možné, taktéž grafickou vizualizací. Vždy mějme na paměti, že závěry o nulové hypotéze se snažíme vždy provádět co nejobektivněji tak, aby odrážely skutečnost, nikoliv naše přání.

*Poznámka:* Každá z výše uvedených testů normality má své přednosti i nedostatky vycházející z algoritmu, který se za testem skrývá. Lillieforsův test bývá například nejvíce benevolentní k zamítnutí hypotézy o normalitě náhodného výběru. To znamená, že hypotézu o normalitě zamítá zpravidla nejpomaleji, tedy při jeho využití je větší šance, že hypotézu o normalitě náhodného výběru nezamítne. Anderson-Darlingův test, je jako jediný z uvedených testů citlivý na odlehlá pozorování. Dokáže tedy nejlépe odhalit porušení normality náhodného výběru z důvodu odlehlých či extrémních pozorování.

Který test normality budeme favorizovat pro  $n \geq 30$  a proč? Lillieforsův, Anderson-Darlingův test, Shapiro-Wilkův test nebo Cramérův-von Misesův test? Jak je to se Shapiro-Wilkovým testem pro  $n > 30$ ? Někde se píše, že je nejlepším testem i pro  $n \geq 30$ . Zahreneme jej tedy také mezi testy pro  $n \geq 30$ ? Test, který budeme favorizovat, potom použijí na ověření normality u všech parametrických i neparametrických testů.

### Příklad 6.1. Test o jednorozměrné normalitě dat

Mějme datový soubor 19-more-samples-correlations-skull.txt a proměnnou intorb.B popisující interorbitální šířku mužů bantuské populace v mm (viz sekce ??). Na hladině významnosti  $\alpha = 0.05$  testujte hypotézu, že náhodný výběr naměřených hodnot interorbitální šířky pochází z normálního rozdělení.

#### Řešení příkladu 6.1

Nejprve načteme datový soubor a pomocí operátoru `[]` vybereme z tabulky pouze řádky týkající se mužů bantuské populace (`pop == 'ban'`) a sloupec obsahující údaje o interorbitální šířce, tj. `intorb.B`. Následně z vektoru naměřených hodnot odstraníme neznámé hodnoty pomocí příkazu `na.omit()` a zjistíme, jaký je rozsah tohoto náhodného výběru příkazem `length()`. Protože náhodný výběr obsahuje celkem 14 naměřených hodnot interorbitální šířky mužů

```
1 data <- read.delim('00-Data//19-more-samples-correlations-skull.txt')
2 intorb.BB <- data[data$pop == 'ban', 'intorb.B']
3 intorb.BB <- as.numeric(na.omit(intorb.BB))
4 n <- length(intorb.BB) # 14
```

bantuské populace, což je méně než 30, použijeme na testování hypotézy  $H_0$  Shapirův-Wilkův test. Proces testování si uvedeme v posloupnosti pěti kroků.

#### 1. Stanovení hypotéz

- **slovní formulace** nulové a alternativní hypotézy  
 $H_0$  : Data pochází z normálního rozdělení.  
 $H_1$  : Data nepochází z normálního rozdělení.
- **matematická formulace** nulové a alternativní hypotézy  
 $H_0$  :  $X \sim N(\mu, \sigma^2)$   
 $H_1$  :  $X \not\sim N(\mu, \sigma^2)$

#### 2. Volba hladiny významnosti

Podle zadání volíme hladinu významnosti  $\alpha = 0.05$ .

#### 3. Testování $p$ -hodnotou

Shapirův-Wilkův test provedeme pomocí funkce `shapiro.test()`. Výstupem funkce je údaj o použité proměnné (`intorb.BB`), hodnota testovací statistiky Shapirova-Wilkova testu ( $W = 0.94268$ ) a výsledná  $p$ -hodnota ( $p\text{-value} = 0.4537$ ).

```
5 shapiro.test(intorb.BB)
```

```
Shapiro-Wilk normality test
data:  intorb.BB
W = 0.94268, p-value = 0.4537
```

6  
7  
8  
9  
10

#### 4. Závěr testování

Protože  $p$ -hodnota = 0.4537105 je větší než  $\alpha = 0.05$ ,  $H_0$  nezamítáme na hladině významnosti  $\alpha = 0.05$ .

#### 5. Grafická vizualizace výsledků testování

Závěr testování nyní podpoříme histogramem superponovaným křivkou normálního rozdělení a QQ-diagramem.

Pro vykreslení histogramu nejprve stanovíme hranice třídících intervalů, do kterých rozdělíme naměřené hodnoty. Datový soubor rozdělíme na základě Sturgesova pravidla (viz sekce ??) do pěti ekvidistatních intervalů s šířkou 2 mm prostřednictvím hranic 20, 22, ..., 30. Histogram vykreslíme příkazem `hist()`, a to s výše uvedenými hranicemi třídících intervalů (argument `breaks`) a bez měřítek osy  $x$  a  $y$  (`axes = F`). Na osu  $y$  vyneseme hodnoty relativních četností (`prob = T`). Relativní škála na ose  $y$  nám zajistí správné měřítko k následnému vykreslení křivky hustoty normálního rozdělení. Příkazem `box()` dokreslíme okolo grafu černý rámeček. Příkazem

`axis()` doplníme měřítko osy  $x$  (argument `side = 1`) uvádějící hodnoty středů třídicích intervalů (argument `at`) a měřítko osy  $y$  (argument `side = 2`) jehož hodnoty budou vykresleny horizontálně (argument `las = 1`).

Nakonec do histogramu dokreslíme křivku hustoty normálního rozdělení  $N(\mu, \sigma^2)$ , kde parametry  $\mu$  a  $\sigma^2$  odhadneme jejich nestrannými odhady. Nejprve vygenerujeme dostatečně hustou posloupnost čísel pokrývající rozsah hodnot osy  $x$ , tj. posloupnost `xfit` tisíce čísel v rozsahu 15–35 mm. Dále stanovíme odhady parametrů  $\mu$  a  $\sigma^2$ . Parametr  $\mu$  odhadneme výběrovým průměrem pomocí funkce `mean()` a parametr  $\sigma^2$  odhadneme výběrovým rozptylem pomocí funkce `var()`. Pro každou hodnotu proměnné `xfit` nyní spočítáme hodnotu hustoty normálního rozdělení  $N(\mu, \sigma^2)$  pomocí příkazu `dnorm()`. Prvním argumentem `x` zadáváme posloupnost, nad kterou budeme hodnoty hustoty počítat (tj. `x = xfit`). Druhý argument (`mean`) specifikuje střední hodnotu  $\mu$  normálního rozdělení, tj. v našem případě odhad `m`. Konečně, třetí argument (`sd`) specifikuje směrodatnou odchylku  $\sigma$  normálního rozdělení, tj. v našem případě odmocnina z výběrového rozptylu `sqrt(s2)`. Křivku hustoty normálního rozdělení vykreslíme příkazem `lines()`.

QQ-diagram vykreslíme pomocí funkce `qqnorm()`. Prvním argumentem funkce je vektor naměřených hodnot, tj. `intorb.BB`. Pomocí argumentů `pch`, `bg` a `col` specifikuje kulatý typ vykreslovaných bodů se světle modrou výplní a tmavým odstínem obrysové barvy. Dále zakážeme vypsání nadpisu grafu (`main = ""`) a pomocí argumentů `xlab` a `ylob` změňme popisky os  $x$  a  $y$  z anglických na české ekvivalenty. Konečně argumentem `las` změňme směr popisků měřítka osy  $y$  na horizontální. Nakonec dokreslíme do grafu referenční čáru pomocí příkazu `qqline()`. Prvním argumentem funkce je opět vektor naměřených hodnot, na základě kterého bude spočítán sklon a posunutí referenční čáry. Tloušťku čáry zvýšíme o jedna argumentem `lwd = 2`.

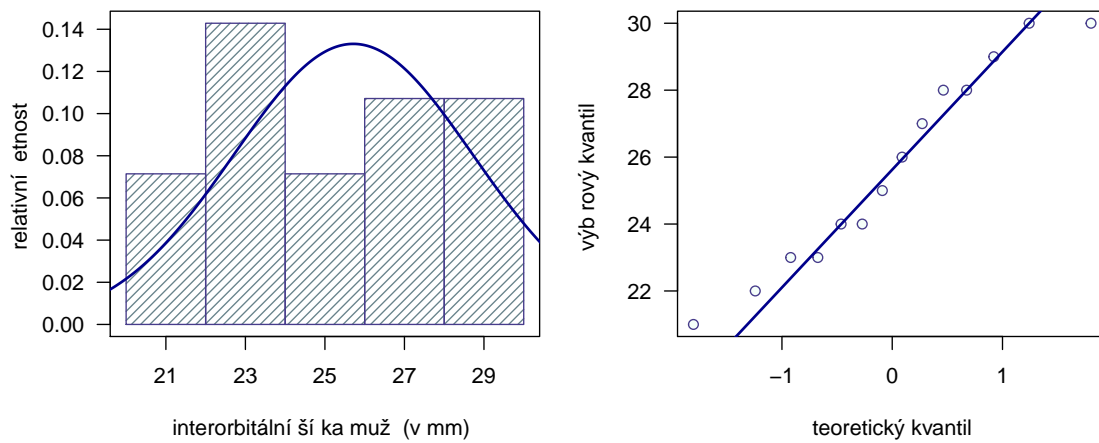
```
11 par(mar = c(4, 4, 1, 1))
12 hist(intorb.BB, breaks = seq(20, 30, by = 2), prob = T,
13       col = 'lightblue4', border = 'slateblue4', density = 20,
14       main = '', axes = F, ylab = 'relativní četnost',
15       xlab = 'interorbitální šířka mužů (v mm)')
16 box(bty = 'o')
17 axis(side = 1, at = seq(21, 29, by = 2))
18 axis(side = 2, las = 1)
19
20 xfit <- seq(15, 35, length = 512)
21 m <- mean(intorb.BB)
22 s2 <- var(intorb.BB)
23 yfit <- dnorm(x = xfit, mean = m, sd = sqrt(s2))
24 lines(xfit, yfit, col = 'darkblue', lwd = 2)
25
26 qqnorm(intorb.BB, pch = 21, bg = 'mintcream', col = 'slateblue4', main = '',
27        xlab = 'teoretický kvantil', ylab = 'výběrový kvantil', las = 1)
28 qqline(intorb.BB, col = 'darkblue', lwd = 2)
```

Ačkoli histogram na obrázku 1 nekopíruje přesný tvar křivky hustoty normálního rozdělení, nemůžeme říci, že by data nebyla normálně rozdělená. Při takto malém množství pozorování mohou data pocházet z normálního rozdělení, ačkoli tvar histogramu této skutečnosti úplně neodpovídá. Naopak z QQ-diagramu je normalita krásně viditelná. S výjimkou dvou krajních bodů se všechny body drží referenční přímky, což podporuje výsledek testování a normalitu náhodného výběru.

## 6. Interpretace výsledků

Datový soubor naměřených hodnot interorbitální šířky mužů bantuské populace pochází z normálního rozdělení.





Obrázek 1: Histogram a QQ-diagram interorbitální šířky u skeletů mužského pohlaví bantuské populace (v mm)

## Příklad 6.2. Test o jednorozměrné normalitě dat

Mějme datový soubor 15-anova-means-skull.txt a proměnnou upface.H popisující výšku horní části tváře mužů německé populace v mm (viz sekce ??). Na hladině významnosti  $\alpha = 0.05$  testujte hypotézu, že náhodný výběr naměřených hodnot pochází z normálního rozdělení.

### Řešení příkladu 6.2

Nejprve načteme datový soubor, příkazem na.omit() a pomocí operátoru [] vybereme z tabulky pouze řádky týkající se mužů německé populace (pop == 'nem') a sloupec obsahující údaje o výšce horní části tváře, tj. upface.H. Následně z vektoru naměřených hodnot výšek horní části tváře odstraníme chybějící údaje a zjistíme rozsah tohoto náhodného výběru. Protože náhodný výběr obsahuje celkem 19 naměřených hodnot výšky horní části tváře mužů

```
29 data <- read.delim('00-Data//15-anova-means-skull.txt')
30 upface.HN <- data[data$pop == 'nem', 'upface.H']
31 upface.HN <- as.numeric(na.omit(upface.HN))
32 n <- length(upface.HN) # 19
```

německé populace, což je méně než 30, použijeme na testování hypotézy o normalitě Shapirův-Wilkův test.

#### 1. Stanovení hypotéz

- **slovní formulace** nulové a alternativní hypotézy  
 $H_0$  : Data pochází z normálního rozdělení.  
 $H_1$  : Data nepochází z normálního rozdělení.
- **matematická formulace** nulové a alternativní hypotézy  
 $H_0$  :  $X \sim N(\mu, \sigma^2)$   
 $H_1$  :  $X \not\sim N(\mu, \sigma^2)$

#### 2. Volba hladiny významnosti

Podle zadání volíme hladinu významnosti  $\alpha = 0.05$ .

#### 3. Testování $p$ -hodnotou

Shapirův-Wilkův test provedeme pomocí funkce shapiro.test(). Výstupem funkce je údaj o použité proměnné (upface.HN), hodnota testovací statistiky Shapirova-Wilkova testu ( $W = 0.8964$ ) a  $p$ -hodnota ( $p$ -value = 0.0419).

```
33 shapiro.test(upface.HN)
```

```
Shapiro-Wilk normality test

data:  upface.HN
W = 0.8964, p-value = 0.0419
```

34  
35  
36  
37  
38

#### 4. Závěr testování

Protože  $p$ -hodnota = 0.0419 je menší než  $\alpha = 0.05$ ,  $H_0$  zamítáme na hladině významnosti  $\alpha = 0.05$ .

#### 5. Grafická vizualizace výsledků testování

Závěr testování nyní podpoříme histogramem superponovaným křivkou normálního rozdělení a QQ-diagramem. Analogicky jako v příkladu 6.1 vykreslíme histogram příkazem hist(), kde na základě Sturgesova pravidla rozdělíme naměřené hodnoty do pěti ekvidistantních intervalů s šířkou 3 mm prostřednictvím hranic 61, 66, ..., 76. Do histogramu dále pomocí příkazu lines() dokreslíme křivku hustoty normálního rozdělení  $N(\mu, \sigma^2)$ , dopočítanou funkcí dnorm(). Parametry  $\mu$  a  $\sigma^2$  odhadneme příkazy mean() a sd(). Nakonec vykreslíme QQ-diagram (prostřednictvím funkce qqnorm()) s referenční přímkou (prostřednictvím funkce qqline()).

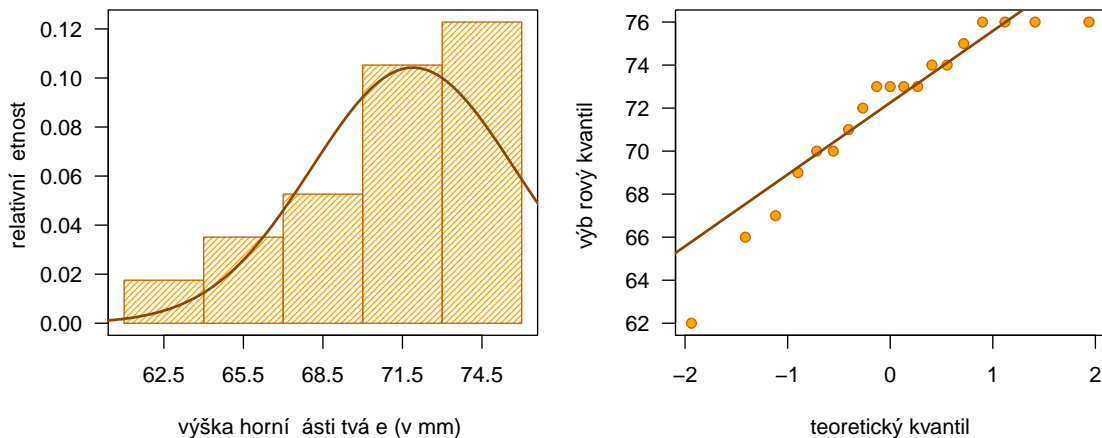
Ačkoli by se mohlo zdát, že histogram na obrázku 2 kopíruje tvar křivky hustoty normálního rozdělení celkem věrně, charakter normálního rozdělení dat není zachován. Na první pohled vidíme, že data jsou oproti



```

39 par(mar = c(4, 4, 1, 1))
40 hist(upface.HN, breaks = seq(61, 76, by = 3), prob = T,
41      col = 'orange', border = 'darkorange3', density = 30,
42      main = '', axes = F, ylab = 'relativní četnost',
43      xlab = 'výška horní části tváře (v mm)')
44 box(bty = 'o')
45 axis(side = 1, at = seq(62.5, 74.5, by = 3))
46 axis(side = 2, las = 1)
47
48 xfit <- seq(55, 80, length = 512)
49 m <- mean(upface.HN)
50 s2 <- var(upface.HN)
51 yfit <- dnorm(x = xfit, mean = m, sd = sqrt(s2))
52 lines(xfit, yfit, col = 'darkorange4', lwd = 2)
53
54 qqnorm(upface.HN, pch = 21, bg = 'orange', col = 'darkorange3', main = '',
55        xlab = 'teoretický kvantil', ylab = 'výběrový kvantil', las = 1)
56 qqline(upface.HN, col = 'darkorange4', lwd = 2)

```



Obrázek 2: Histogram a QQ-diagram výšky horní části tváře u skeletů mužského pohlaví německé populace (v mm)

klasickému normálnímu rozdělení vyšikmená doprava. To by nemusel být až tak závažný přestupek proti normalitě. Ovšem skutečnost, že vyšší hodnoty výšky horní části tváře mají s rostoucími hodnotami vyšší zastoupení, které se na pravém konci nesnižuje, je větším přestupkem proti normalitě. V QQ-diagramu potom vidíme nesymetrické odchýlení bodů od referenční křivky v levé dolní části grafu. Posunutí referenční přímky do levého horního rohu je důkazem výraznější a nesymetrické odlehlosti bodů umístěných navlevo, než tomu bylo u QQ-diagramu v příkladu 6.1. Závěr tohoto příkladu tedy je, že náhodný výběr nepochází z normálního rozdělení a jako takový nemůže být použit jako základ k parametrickým testům (viz kapitoly ?? a ??).

## 6. Interpretace výsledků

Datový soubor výšek horní části tváře mužů německé populace nepochází z normálního rozdělení.



### Příklad 6.3. Test o jednorozměrné normalitě dat

Mějme datový soubor 11-two-samples-means-skull.txt a proměnnou skull.H popisující výšku lebky v mm (viz sekce ??). Na hladině významnosti  $\alpha = 0.05$  testujte hypotézu, že náhodný výběr naměřených výšek lebky mužů pochází z normálního rozdělení.

#### Řešení příkladu 6.3

Nejprve načteme datový soubor a pomocí operátoru [] vybereme z tabulky pouze řádky týkající se mužů (sex == 'm') a sloupec obsahující výšku lebky, tj. skull.H. Příkazem na.omit() odstraníme z vektoru skull.HM chybějící údaje a příkazem length() zjistíme délku tohoto vektoru. Protože náhodný výběr obsahuje údaje o 215 výškách lebek

```
57 data <- read.delim('00-Data//11-two-samples-means-skull.txt')
58 skull.HM <- data[data$sex == 'm' , 'skull.H']
59 skull.HM <- as.numeric(na.omit(skull.HM))
60 n <- length(skull.HM) # 215
```

mužů, což je více než 30, použijeme na testování hypotézy o normalitě Lillieforsův test.

#### 1. Stanovení hypotéz

- **slovní formulace** nulové a alternativní hypotézy  
 $H_0$  : Data pochází z normálního rozdělení.  
 $H_1$  : Data nepochází z normálního rozdělení.
- **matematická formulace** nulové a alternativní hypotézy  
 $H_0$  :  $X \sim N(\mu, \sigma^2)$   
 $H_1$  :  $X \not\sim N(\mu, \sigma^2)$

#### 2. Volba hladiny významnosti

Podle zadání volíme hladinu významnosti  $\alpha = 0.05$ .

#### 3. Testování $p$ -hodnotou

Lillieforsův test provedeme pomocí funkce lillie.test(), která je implementovaná v knihovně nortest. Při použití funkce lillie.test() tedy musíme buď nejdříve načíst knihovnu nortest příkazem library(), nebo se při volání funkce lillie.test() na knihovnu nortest přímo odkázat. Zde použijeme druhý přístup přes odkaz pomocí názvu knihovny a operátoru ::. Výstupem funkce je údaj o použité proměnné (skull.HM), hodnota testovací statistiky Lillieforsova testu ( $D = 0.054341$ ) a  $p$ -hodnota ( $p\text{-value} = 0.1263$ ).

```
61 nortest::lillie.test(skull.HM)
```

```
Lilliefors (Kolmogorov-Smirnov) normality test
data: skull.HM
D = 0.054341, p-value = 0.1263
```

62  
63  
64  
65  
66

#### 4. Závěr testování

Protože  $p$ -hodnota = 0.1263 je větší než  $\alpha = 0.05$ ,  $H_0$  nezamítáme na hladině významnosti  $\alpha = 0.05$ .

#### 5. Grafická vizualizace výsledků testování

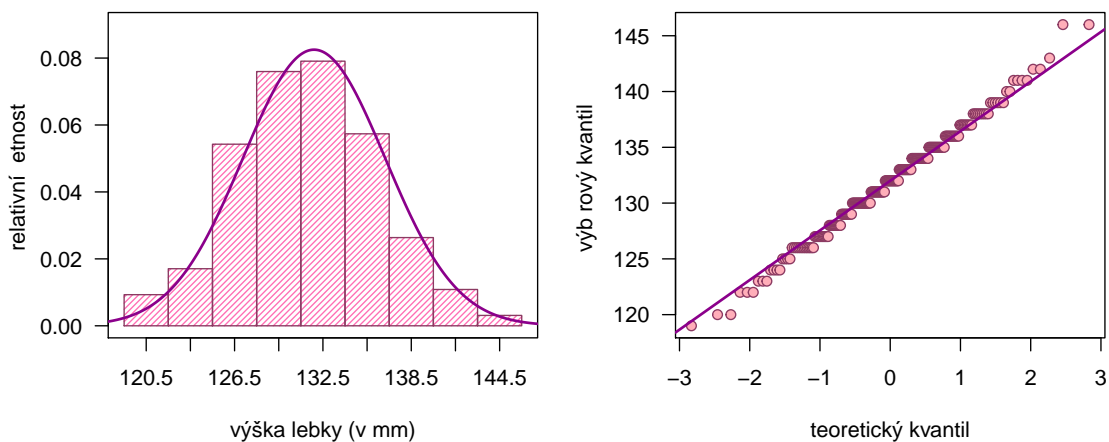
Normalitu náhodného výběru si vizualizujeme histogramem superponovaným křivkou normálního rozdělení a QQ-diagramem. Pro potřeby histogramu rozdělíme naměřené hodnoty na základě Sturgesova pravidla do devíti ekvidistantních intervalů s šířkou 3 mm prostřednictvím hranic 119, 122, ..., 146. Do histogramu dále dokreslíme křivku hustoty normálního rozdělení  $N(\mu, \sigma^2)$ , s parametry  $\mu$  a  $\sigma^2$  odhadnutými pomocí funkcí mean() a sd().

Rozdělení datového souboru ukázkově odpovídá normálnímu rozdělení. Na histogramu vykresleném na obrázku 3 je krásně viditelné, jak naměřená data kopírují tvar křivky hustoty normálního rozdělení. Taktéž QQ-diagram, kde se až na dvě, tři výjimky, drží všechny body v blízkém okolí referenční přímky, ukazuje na skoro až učebnicovou normalitu náhodného výběru.

```

67 par(mar = c(4, 4, 1, 1))
68 hist(skull.HM, breaks = seq(119, 146, by = 3), prob = T,
69     col = 'hotpink1', border = 'hotpink4', density = 30,
70     main = '', axes = F, ylab = 'relativní četnost',
71     xlab = 'výška lebky (v mm)', ylim = c(0, 0.09))
72 box(bty = 'o')
73 axis(side = 1, at = seq(120.5, 144.5, by = 3))
74 axis(side = 2, las = 1)
75
76 xfit <- seq(115, 150, length = 512)
77 yfit <- dnorm(x = xfit, mean = mean(skull.HM), sd = sd(skull.HM))
78 lines(xfit, yfit, col = 'darkmagenta', lwd = 2)
79
80 qqnorm(skull.HM, pch = 21, bg = 'lightpink1', col = 'hotpink4', main = '',
81     xlab = 'teoretický kvantil', ylab = 'výběrový kvantil', las = 1)
82 qqline(skull.HM, col = 'darkmagenta', lwd = 2)

```



Obrázek 3: Histogram a QQ-diagram výšky lebky u skeletů mužského pohlaví (v mm)

## 6. Interpretace výsledků

Datový soubor výšek lebky mužů pochází z normálního rozdělení.



#### Příklad 6.4. Test o jednorozměrné normalitě dat

Mějme datový soubor 10-two-samples-means-birth.txt a proměnnou birth.W popisující porodní hmotnost novorozenců-chlapců v g (viz sekce ??). Na hladině významnosti  $\alpha = 0.05$  testujte hypotézu, že náhodný výběr naměřených hmotností pochází z normálního rozdělení.

#### Řešení příkladu 6.4

Nejprve načteme datový soubor a pomocí operátoru [] vybereme z tabulky sloupec obsahující porodní hmotnost novorozenců, tj. birth.W. Dále příkazem na.omit() odstraníme z vektoru birth.W neznámé hodnoty a příkazem length() stanovíme délku tohoto vektoru. Protože náhodný výběr obsahuje celkem 573 naměřených hmotností novorozenců,

```
83 data <- read.delim('00-Data//10-two-samples-means-birth.txt')
84 birth.W <- data[, 'birth.W']
85 birth.W <- as.numeric(na.omit(birth.W))
86 n <- length(birth.W) # 573
```

což je více než 30, použijeme na testování hypotézy o normalitě Lillieforsův test.

#### 1. Stanovení hypotéz

- **slovní formulace** nulové a alternativní hypotézy  
 $H_0$  : Data pochází z normálního rozdělení.  
 $H_1$  : Data nepochází z normálního rozdělení.
- **matematická formulace** nulové a alternativní hypotézy  
 $H_0$  :  $X \sim N(\mu, \sigma^2)$   
 $H_1$  :  $X \not\sim N(\mu, \sigma^2)$

#### 2. Volba hladiny významnosti

Podle zadání volíme hladinu významnosti  $\alpha = 0.05$ .

#### 3. Testování $p$ -hodnotou

Lillieforsův test provedeme pomocí funkce lillie.test() z knihovny nortest. Výstupem funkce je údaj o použité proměnné (birth.W), hodnota testovací statistiky Lillieforsova testu ( $D = 0.083276$ ) a  $p$ -hodnota ( $p\text{-value} = 3.707 \times 10^{-10}$ ).

```
87 nortest::lillie.test(birth.W)
```

```
      Lilliefors (Kolmogorov-Smirnov) normality test
data:  birth.W
D = 0.083276, p-value = 3.707e-10
```

88  
89  
90  
91  
92

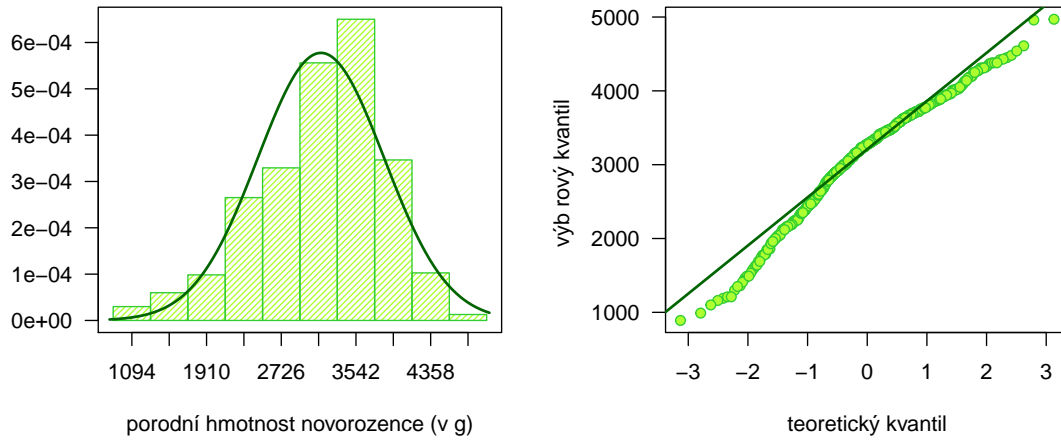
#### 4. Závěr testování

Protože  $p$ -hodnota  $= 3.707 \times 10^{-10}$  je menší než  $\alpha = 0.05$ ,  $H_0$  zamítáme na hladině významnosti  $\alpha = 0.05$ .

#### 5. Grafická vizualizace výsledků testování

Výsledek testování nyní podložíme ukázkou histogramu superponovaného křivkou normálního rozdělení a QQ-diagramu. Na základě Sturgesova pravidla rozdělíme naměřené hodnoty do deseti ekvidistantních intervalů s šířkou 408 g prostřednictvím hranic 890, 1298, ..., 4970. Tyto hranice použijeme při vykreslení histogramu, do kterého dále pomocí příkazu lines() dokreslíme křivku hustoty normálního rozdělení  $N(\mu, \sigma^2)$  s parametry  $\mu$  a  $\sigma^2$  odhadnutými jejich nestrannými odhady (viz kapitola ??). Nakonec vykreslíme výše zmíněný QQ-diagram spolu s referenční přímkou.

Z obou grafů je patrný nenormální charakter náhodného výběru. Na první pohled by se sice mohlo zdát, že histogram dostatečně kopíruje tvar křivky hustoty normálního rozdělení, ovšem při bližším pohledu vidíme, že histogram je oproti křivně normálního rozdělení posunutý doprava. Navíc interval se středem 3542 g obsahuje



Obrázek 4: Histogram a QQ-diagram porodní hmotnosti novorozenců (v g)

výrazně více hodnot, než by bylo vhodné. Při malém rozsahu náhodného výběru by to nebyl problém, ale při tak rozsáhlém výběru, jaký máme k dispozici jsou zmíněné nedostatky pro předpoklad normality fatální. QQ-diagram, ze kterého je odchylka bodů od referenční přímky na první pohled patrná, nás v odlišnosti od normálního rozdělení náhodného výběru jen utvrzuje. V tomto případě se tedy skutečně kloníme k závěru, že náhodný výběr nepochází z normálního rozdělení a jako takový není využitelný k parametrickému testování.

## 6. Interpretace výsledků

Datový soubor porodních hmotností novorozenců chlapců nepochází z normálního rozdělení.



## 6.2 Testy dvourozměrné normality

Dvourozměrná normalita náhodného výběru je stěžejním předpokladem umožňujícím testování nulové hypotézy o korelačním koeficientu  $\rho$  a rozdílu dvou korelačních koeficientů  $\rho_1 - \rho_2$  pomocí parametrických testů (viz kapitola ??). Dříve, než takový parametrický test použijeme, musíme ověřit, zda námi naměřená dvojice náhodných veličin pochází z dvourozměrného náhodného rozdělení.

Nechť  $(X_1, Y_1)^T, \dots, (X_n, Y_n)^T$  je náhodný výběr z rozdělení  $L(\theta)$ , kde  $\theta$  je obecně vektor parametrů rozdělení  $L$ . Na hladině významnosti  $\alpha = 0.05$  testujeme hypotézu  $H_0$ : *Náhodný výběr pochází z dvourozměrného normálního rozdělení.*, tj.  $X \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , tj.  $L(\theta) = N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , kde  $\theta = (\boldsymbol{\mu}, \boldsymbol{\Sigma}^2)^T$ , oproti alternativní hypotéze  $H_1$ : *Náhodný výběr nepochází z dvourozměrného normálního rozdělení.*, tj.  $X \not\sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ .

K testování hypotézy o dvourozměrné normalitě náhodného výběru použijeme primárně Mardiova test. Tento test nejprve stanoví hodnotu koeficientu šikmosti a koeficientu špičatosti a následně určí, zda jsou tyto hodnoty statisticky významné od nuly. V případě, že hodnota alespoň jednoho z koeficientů je statisticky významně odlišná od nuly, je nulová hypotéza o dvourozměrné normalitě náhodného výběru zamítnuta. Hypotézu o dvourozměrné normalitě může případně otestovat pomocí Henzova-Zirklerova testu nebo Roystonova testu vícerozměrné normality. Všechny tyto testy jsou k dispozici ve funkci `mvn()` implementované v knihovně `MVN`. Volbu Mardiova testu zvolíme nastavením argumentu `mvnTest = 'mardia'` (viz příklad 6.5), volbu Henzova-Zirklerova testu nastavením argumentu `mvnTest = 'hz'` a volbu Roystonova testu specifikujeme argumentem `mvnTest = 'royston'` ve funkci `mvn()`.

Dvourozměrné rozdělení náhodného výběru vizualizujeme 3D grafem a tečkovým diagramem superponovaným 95% elipsou spolehlivosti. Vizualně můžeme zhodnotit dvourozměrnou normalitu náhodného výběru právě pomocí tečkového digramu. Pokud alespoň 95 % hodnot náhodného výběru spadá do elipsy spolehlivosti, předpokládáme, že data pochází z dvourozměrného normálního rozdělení. Analogicky můžeme tečkový diagram superponovat 90% nebo 99% elipsou spolehlivosti a sledovat, zda do elipsy spolehlivosti spadá alespoň 90 % resp. alespoň 99 % naměřených hodnot.

*Poznámka:* Důvod, proč si zde prezentujeme tři testy dvourozměrné normality, je, abychom čtenáři poskytli širší spektrum možných testů dvourozměrné normality a umožnili mu vybrat si z testů takový, který mu přijde nejvhodnější. I z toho důvodu u každého z následujících příkladů provádíme testování nulové hypotézy o dvourozměrné normalitě naměřených dat pomocí Mardiova testu, Henzova-Zirklerova testu i Roystonova testu. Ve vlastních analýzách je ovšem situace jiná. Zde zpravidla volíme na otestování hypotézy o dvourozměrné normalitě pouze jeden test, který používáme univerzálně pokaždé, když chceme hypotézu o dvourozměrné normalitě testovat. V žádném případě neprovádíme ve vlastních analýzách všechny tři testy dvourozměrné normality najednou a v případě rozporupných závěrů testů si nevybíráme ten, jehož výsledek se nám aktuálně nejvíce hodí. Vždy mějme na paměti, že aby vlastní analýza nepůsobila účelově, nebo dokonce nebyla demagogií, musíme zvolené testy volit univerzálně a bez ohledu na to, jak příhodný či nepříhodný výsledek nám v konkrétní situaci předloží.

Více testů dvourozměrné normality smíme provádět pouze v hraničních případech, kdy univerzálně zvolený test hypotézu o dvourozměrné normalitě hraničně zamítá nebo nezamítá. Pak je vhodné provést i další testy, které nám mohou pomoci nahlédnout na dvourozměrnou normalitu náhodného výběru z jiného pohledu a lépe tak zvážit skutečný charakter dvourozměrného rozdělení naměřených hodnot. Pokud se na základě dalších testů rozhodneme pro jiný závěr, než jaký nám poskytuje test univerzálně používaný na otestování hypotézy o dvourozměrné normalitě v celé datové analýze, je vždy třeba svůj závěr podložit relevantními výsledky a, je-li to možné, taktéž grafickou vizualizací. Vždy mějme na paměti, že závěry o nulové hypotéze se snažíme vždy provádět co nejobektivněji tak, aby odrážely skutečnost, nikoliv naše přání.

*Poznámka:* Pokud si mezi prezentovanými testy nedokážete vybrat favoritní test a nevádí vám větší množství výsledků, doporučujeme vám zůstat u volby Mardiova testu (defaultní nastavení funkce `mvn()`). Tento test totiž poskytuje nejen globální pohled na dvourozměrnou normalitu náhodného výběru, ale také hodnotí dvourozměrnou normalitu marginálně z hlediska šikmosti a špičatosti. Tím nám může pomoci rychle zodpovědět otázku, zda v případě zamítnutí nulové hypotézy dvourozměrné normalitě náhodného výběru, byla tato hypotéza zamítnuta kvůli nevhodnému zešikmení nebo kvůli nevhodnému zešpičatění. Tato informace nám dále může posloužit při rozhodování o dalším postupu analýzy. Například je-li pro nás předpoklad dvourozměrné normality stěžejní, může nás porušení dvourozměrné normality z hlediska šikmosti navést na prozkoumání odlehlých pozorování a jejich případné odstranění. Naopak porušení dvourozměrné normality z hlediska špičatosti nás může motivovat k transformování dat do logaritmické, nebo odmocninové škály, nebo ke standardizaci naměřených hodnot.

### Příklad 6.5. Test o dvourozměrné normalitě dat

Mějme datový soubor 19-more-samples-correlations-skull.txt a proměnnou nose.H popisující výškou nosu v mm a proměnnou intorb.B popisující interorbitální šířku v mm mužů čínské populace (viz sekce ??). Na hladině významnosti  $\alpha = 0.05$  testujte hypotézu, že náhodný výběr naměřených hodnot pochází z dvourozměrného normálního rozdělení.

#### Řešení příkladu 6.5

Nejprve načteme datový soubor, pomocí operátoru [] vybereme z tabulky řádky týkající se čínské populace, tj. pop == 'cin' a sloupce obsahující naměřené hodnoty výšky nosu (nose.H) a hodnoty interorbitální šířky (intorb.B). Vše vložíme do proměnné udaje. Příkazem na.omit() odstraníme z proměnné udaje neznámé hodnoty a příkazem dim() zjistíme počet subjektů, jejichž údaje máme k dispozici.

```
93 data <- read.delim('00-Data//19-more-samples-correlations-skull.txt')
94 udaje <- data[data$pop == 'cin' , c('nose.H', 'intorb.B')]
95 udaje <- na.omit(udaje)
96 dim(udaje) # 19x2
97 range(udaje$nose.H) # 48-59
98 range(udaje$intorb.B) # 19-27
```

Náhodný výběr obsahuje údaje o dvou proměnných, tj. výšce nosu a interorbitální šířce, u 19 mužů čínské populace. Hodnoty výšky nosu se pohybují v rozmezí 48–59 mm, hodnoty interorbitální šířky se pohybují v rozmezí 19–27 mm. K otestování hypotézy o dvourozměrné normalitě použijeme Mardiaův test, Henzův-Zirklerův test a Roystonův test. Testování provedeme v posloupnosti pěti kroků.

#### 1. Stanovení hypotéz

- **slovní formulace** nulové a alternativní hypotézy  
 $H_0$  : Data pochází z dvourozměrného normálního rozdělení.  
 $H_1$  : Data nepochází z dvourozměrného normálního rozdělení.
- **matematická formulace** nulové a alternativní hypotézy  
 $H_0$  :  $(X, Y)^T \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$   
 $H_1$  :  $(X, Y)^T \not\sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

#### 2. Volba hladiny významnosti

Podle zadání volíme hladinu významnosti  $\alpha = 0.05$ .

#### 3. Testování $p$ -hodnotou

Mardiaův test provedeme pomocí funkce mvn() z knihovny MVN. Výstupem funkce jsou tři tabulky. První tabulka \$multivariateNormality poskytuje údaje o testu dvourozměrné normality, druhá tabulka \$univariateNormality poskytuje výsledky testu jednorozměrné normality provedeného zvlášť pro každou proměnnou a třetí tabulka \$Descriptives poskytuje výsledky popisné statistiky. Pro nás bude zajímavá první tabulka \$multivariateNormality.

```
99 MVN::mvn(udaje , mvnTest = 'mardia')$multivariateNormality
```

```
sROC 0.1-2 loaded
```

100

|   | Test            | Statistic          | p value           | Result |
|---|-----------------|--------------------|-------------------|--------|
| 1 | Mardia Skewness | 3.65691216526235   | 0.454423782959563 | YES    |
| 2 | Mardia Kurtosis | -0.335092668388565 | 0.737555183255058 | YES    |
| 3 | MVN             | <NA>               | <NA>              | YES    |

101

102

103

104

Tabulka \$multivariateNormality se skládá ze tří řádků a čtyř sloupců. V prvním řádku je uvedena hodnota testovací statistiky,  $p$ -hodnota a závěr testování nevýznamnosti koeficientu šikmosti (Mardia Skewness), v

druhém řádku je uvedena hodnota testovací statistiky,  $p$ -hodnota a závěr testování nevýznamnosti koeficientu špičatosti (Mardia Kurtosis). Ve třetím řádku je potom uveden celkový závěr testování hypotézy o dvourozměrné normalitě náhodného výběru.

Protože  $p$ -hodnota o nevýznamnosti koeficientu šikmosti, tj. 0.4544, je větší než 0.05, hypotézu o nevýznamnosti koeficientu šikmosti nezamítáme. Koeficient šikmosti tedy není statisticky významný a dvourozměrná data nejsou kladně ani záporně vyšikmená. Jelikož  $p$ -hodnota testu o nevýznamnosti koeficientu špičatosti, tj. 0.7376, je větší než  $\alpha = 0.05$ , hypotézu o nevýznamnosti koeficientu špičatosti též nezamítáme. Koeficient špičatosti tedy není statisticky významný a datový soubor nevykazuje abnormální špičatost nebo plochost. Protože náhodný výběr nevykazuje statisticky významné známky zešikmení ani zešpičatění, nezamítáme hypotézu o dvourozměrné normalitě náhodného výběru.

Dále si otestujeme hypotézu o dvourozměrné normalitě také pomocí Henzova-Zirklerova testu, a to pomocí funkce `mvn()` s nastavením argumentu `mvnTest = 'hz'`. Výstupem funkce jsou opět tabulky `$multivariateNormality`, `$univariateNormality` a `$Descriptives`, z nichž poslední dvě obsahují úplně stejné hodnoty jako v případě Mardiova testu. Opět se zaměříme pouze na tabulku `$multivariateNormality`.

```
105 MVN::mvn(udaje , mvnTest = 'hz')$multivariateNormality
```

|   | Test          | HZ      | p value   | MVN |
|---|---------------|---------|-----------|-----|
| 1 | Henze-Zirkler | 0.32743 | 0.6342473 | YES |

106  
107

Tabulka `$multivariateNormality` se nyní skládá pouze z jednoho řádku obsahujícího hodnotu testovací statistiky Henzova-Zirklerova testu,  $p$ -hodnotu a rozhodnutí o dvourozměrné normalitě náhodného výběru. Protože  $p$ -hodnota = 0.6342 je větší než 0.05, hypotézu  $H_0$  nezamítáme na hladině významnosti  $\alpha = 0.05$ .

Nakonec otestujeme hypotézu o dvourozměrné normalitě pomocí Roystonova testu, a to opět pomocí funkce `mvn()` s nastavením argumentu `mvnTest = 'royston'`. Výstupem funkce jsou opět tabulky `$multivariateNormality`, `$univariateNormality` a `$Descriptives`, z nichž poslední dvě tabulky obsahují úplně stejné hodnoty jako při použití Mardiova testu a Henzova-Zirklerova testu. Opět nás zajímá pouze tabulka `$multivariateNormality`.

```
108 MVN::mvn(udaje , mvnTest = 'royston')$multivariateNormality
```

|   | Test    | H        | p value   | MVN |
|---|---------|----------|-----------|-----|
| 1 | Royston | 2.035765 | 0.3637774 | YES |

109  
110

Tabulka `$multivariateNormality` obsahuje pouze jeden řádek s hodnotou testovací statistiky Roystonova testu,  $p$ -hodnotu a rozhodnutí o dvourozměrné normalitě náhodného výběru. Protože  $p$ -hodnota = 0.3638 je větší než 0.05, hypotézu  $H_0$  nezamítáme na hladině významnosti  $\alpha = 0.05$ .

#### 4. Grafická vizualizace výsledků testování

Rozdělení náhodného výběru vizualizujeme 3D grafem jádrového odhadu hustoty získaného na základě naměřených hodnot. Jádrový odhad vypočítáme pomocí funkce `kde2d()` z knihovny `MASS`. Vstupy funkce jsou nejprve naměřené hodnoty, nad kterými chceme jádrový odhad spočítat, tj. proměnné `nose.H` a `intorb.B`. Dále specifikujeme počet bodů (uzlů), ve kterých chceme jádrový odhad hustoty spočítat, nastavením argumentu `n = 50` (bodů). Posledním argumentem `lims` specifikujeme, na jaké ploše se má jádrový odhad vypočítat. Abchom získali komplexní pohled na data, necháme jádrový odhad spočítat ve směru proměnné `nose.H` v rozsahu 45–63 mm a ve směru proměnné `intorb.B` v rozsahu 16–30 mm. Výstupem funkce `kde2d()` jsou nové souřadnice  $x$ ,  $y$  a  $z$  jádrového odhadu dvourozměrné hustoty normálního rozdělení.

3D graf nyní vykreslíme pomocí funkce `persp()`. První tři argumenty jsou  $x$ -ová,  $y$ -ová a  $z$ -ová souřadnice hustoty, tj. proměnné `souradnice$x`, `souradnice$y` a `souradnice$z`. Dále argumenty `xlab`, `ylab` a `zlab` změním popisky os. Argumentem `theta` natočíme graf v horizontálním směru o  $20^\circ$  proti směru hodinových ručiček a argumentem `phi` natočíme graf ve vertikálním směru o  $30^\circ$  směrem k nám. Nakonec argumentem `col` nastavíme barevnou škálu grafu, a to tak, aby s rostoucí výškou grafu docházelo ke změně barev z palety `terrain.colors()` ve škále 12 odstínů. Rozdělení grafu podle výšky hustoty na 12 oblastí provedeme pomocí funkce `cut()`. Každá výšková oblast dostane potom přiřazený jeden odstín barvy z palety `terrain.colors`.



```

111 souradnice <- MASS::kde2d(udaje$nose.H, udaje$intorb.B, n = 50, lims = c(45, 63, 16, 30))
112 n <- dim(souradnice$z)[1]
113 vyska <- souradnice$z[-1, -1] + souradnice$z[-1, -n] + souradnice$z[-n, -1] + souradnice$z[-n, -n]
114 vyska <- cut(vyska, 12)
115
116 persp(souradnice$x, souradnice$y, souradnice$z, xlab = 'výška nosu (v mm)',
117       ylab = 'interorbitální šířka (v mm)', zlab = "relativní četnost",
118       theta = -20, phi = 30, col = terrain.colors(12)[vyska])

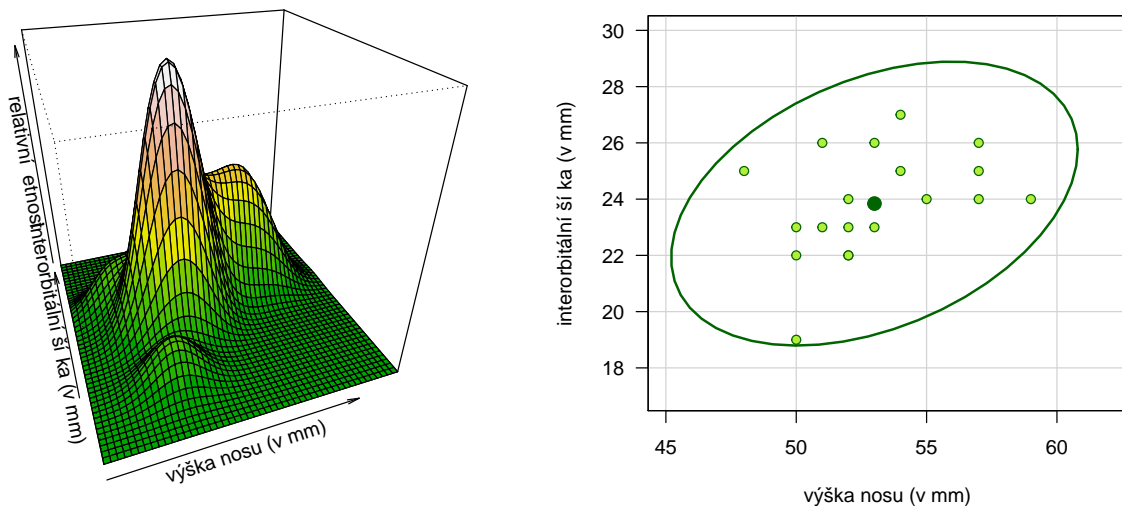
```

Dále vykreslíme tečkový graf superponovaný 95% elipsou spolehlivosti. K tomu použijeme funkci `dataEllipse()` z knihovny `car`. Prvními dvěma argumenty funkce `dataEllipse()` jsou proměnné, nad nimiž chceme elipsu sestrojít, tj. `nose.H` a `intorb.B`. Argumentem `level = 0.95` nastavíme hodnotu koeficientu spolehlivosti ( $1 - \alpha = 0.95$ ). Specifikací argumentů `xlim` a `ylim` nastavíme rozsahy os  $x$  a  $y$ , argumenty `xlab` a `ylab` změníme jejich popisky. Nastavením argumentu `main = ""` zakážeme vypsání nadpisu grafu. Konečně pomocí argumentů `pch`, `bg` a `col` nastavíme vykreslení kulatých bodů se světle zeleným vnitřkem a tmavě zeleným obrysem. Poznamenejme, že argument `col` ovlivňuje kromě barvy obrysu bodů také barvu elipsy a jejího středu. Nakonec argumentem `lwd` zvolíme silnější šířku obrysu elipsy a argumentem `las` nastavíme horizontální směr hodnot měřítka osy  $y$ .

```

119 car::dataEllipse(udaje$nose.H, udaje$intorb.B, level = 0.95, xlim = c(45, 62),
120                ylim = c(17, 30), xlab = 'výška nosu (v mm)',
121                ylab = 'interorbitální šířka (v mm)',
122                main = '', pch = 21, col = 'darkgreen', bg = 'olivedrab2', lwd = 2, las = 1)

```



Obrázek 5: 3D graf a tečkový diagram s 95% elipsou spolehlivosti pro výšku nosu a interorbitální šířku mužů čínské populace (v mm)

Z obou grafů je patrná dvourozměrná normalita náhodného výběru. Na 3D grafu můžeme vidět hlavní vrcholek doprovázený nižším vrcholkem. Tento jev by mohl ukazovat na směs dvou normálních rozdělení. Při takto malém počtu hodnot však nemůžeme s jistotou říci, že jde skutečně o směs. Malý vrcholek vpředu grafu značí odlehlé pozorování viditelné také na tečkovém diagramu. Přesuneme-li pozornost na tečkový diagram, vidíme, že všechny body se realizují uvnitř 95% elipsy spolehlivosti. To je souladu se závěrem testování a můžeme tedy říci, že grafy výsledek testování podporují.

## 5. Interpretace výsledků

Datový soubor výšky nosu a interorbitální šířky mužů čínské populace pochází z dvourozměrného normálního

rozdělení, a můžeme jej tedy použít jako základ k parametrickému testu o korelačním koeficientu  $\rho$  (kapitola ??).



### Příklad 6.6. Test o dvourozměrné normalitě dat

Mějme datový soubor 16-anova-head.txt a proměnnou head.L popisující délku hlavy v mm a proměnnou bizyg.W popisující šířku tváře v mm (viz sekce ??). Na hladině významnosti  $\alpha = 0.10$  testujte hypotézu, že náhodný výběr délek hlavy a šířek tváře žen pochází z dvourozměrného normálního rozdělení.

#### Řešení příkladu 6.6

Nejprve načteme datový soubor a pomocí operátoru `[]` vybereme z tabulky pouze řádky týkající se žen (`sex == 'f'`) a sloupce obsahující naměřené hodnoty délky hlavy (`head.L`) a šířky tváře (`bizyg.W`). Vše vložíme do proměnné `udaje`. Příkazem `na.omit()` odstraníme z proměnné `udaje` neznámé hodnoty a příkazem `dim()` zjistíme počet subjektů, jejichž údaje máme k dispozici.

```
123 data <- read.delim('00-Data//16-anova-head.txt')
124 udaje <- data[data$sex == 'f' , c('head.L', 'bizyg.W')]
125 udaje <- na.omit(udaje)
126 dim(udaje) # 100x2
127 range(udaje$head.L) # 170x205
128 range(udaje$bizyg.W) # 120-151
```

Náhodný výběr obsahuje údaje o dvou proměnných, tj. délce hlavy a šířce tváře, u 100 žen. Naměřené hodnoty délky hlavy se pohybují v rozmezí 170-205 mm, naměřené hodnoty šířky tváře se pohybují v rozmezí 120–151 mm. K otestování hypotézy o dvourozměrné normalitě použijeme Mardiov, Henziov-Zirklerův i Roystonův test.

#### 1. Stanovení hypotéz

- **slovní formulace** nulové a alternativní hypotézy  
 $H_0$  : Data pochází z dvourozměrného normálního rozdělení.  
 $H_1$  : Data nepochází z dvourozměrného normálního rozdělení.
- **matematická formulace** nulové a alternativní hypotézy  
 $H_0$  :  $(X, Y)^T \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$   
 $H_1$  :  $(X, Y)^T \not\sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

#### 2. Volba hladiny významnosti

V souladu se zadáním volíme hladinu významnosti  $\alpha = 0.10$ .

#### 3. Testování $p$ -hodnotou

Mardiov test provedeme pomocí funkce `mvn()` z knihovny MVN specifikací argumentu `mvnTest = 'mardia'`. Z výsledných tabulek nás zajímá pouze tabulka `$multivariateNormality` obsahující výsledky testu o nevýznamnosti koeficientu šikmosti, testu o nevýznamnosti koeficientu špičatosti, a v jejich důsledku také testu o dvourozměrné normalitě náhodného výběru.

```
129 MVN::mvn(udaje , mvnTest = 'mardia')$multivariateNormality
```

|   | Test            | Statistic         | p value           | Result |     |
|---|-----------------|-------------------|-------------------|--------|-----|
| 1 | Mardia Skewness | 9.8656660128771   | 0.042752347786637 | NO     | 130 |
| 2 | Mardia Kurtosis | 0.604444695534365 | 0.545548027690992 | YES    | 131 |
| 3 | MVN             | <NA>              | <NA>              | NO     | 132 |
|   |                 |                   |                   |        | 133 |

Protože  $p$ -hodnota testu o nevýznamnosti koeficientu šikmosti, tj. 0.04275, je menší než 0.10, hypotézu o nevýznamnosti koeficientu šikmosti zamítáme na hladině významnosti  $\alpha = 0.10$ . Koeficient šikmosti ukazuje na statisticky významné zešikmení dat. Jelikož  $p$ -hodnota testu o nevýznamnosti koeficientu špičatosti, tj. 0.5455, je větší než  $\alpha = 0.10$  hypotézu o nevýznamnosti koeficientu špičatosti nezamítáme na hladině významnosti  $\alpha = 0.10$ . Koeficient špičatosti ukazuje na statisticky nevýznamnou špičatost rozdělení náhodného výběru. Protože náhodný výběr vykazuje statisticky významné zešikmení, zamítáme hypotézu o dvourozměrné normalitě náhodného výběru.

*Poznámka:* V tabulce `$multivariateNormality` je ve sloupci `Result` uvedeno rozhodnutí, zda je na hladině významnosti  $\alpha = 0.05$  splněn předpoklad o nevýznamnosti koeficientu šikmosti, resp. koeficientu špičatosti a v

důsledku toho také předpoklad o normalitě dvourozměrného normálního rozdělení náhodného výběru. Z výše uvedené tabulky tedy vyplývá, že na hladině významnosti  $\alpha = 0.05$  není splněn předpoklad o nevýznamnosti s koeficientu šikmosti ani předpoklad o dvourozměrném normálním rozdělení. Předpoklad o nevýznamnosti s koeficientu špicatosti na hladině významnosti  $\alpha = 0.05$  splněn je. Ve funkci `mvn()` není bohužel možné změnit hladinu významnosti  $\alpha$  na 0.10 nebo 0.01. Pokud tedy testujeme hypotézu o dvourozměrné normalitě na jiné hladině významnosti než  $\alpha = 0.05$ , musíme hodnoty ve sloupci `Result` ignorovat a závěr testování stanovit na základě porovnání  $p$ -hodnoty s požadovanou hladinou významnosti, jako jsme to učinili výše.

Nyní testujeme nulovou hypotézu o dvourozměrné normalitě náhodného výběru pomocí Henzova-Zirklerova testu, a to použitím funkce `mvn()` s nastavením argumentu `mvnTest = 'hz'`.

```
134 MVN::mvn(udaje, mvnTest = 'hz')$multivariateNormality
```

|   | Test          | HZ        | p value    | MVN |
|---|---------------|-----------|------------|-----|
| 1 | Henze-Zirkler | 0.8516617 | 0.09943109 | YES |

135  
136

Protože  $p$ -hodnota = 0.09943 je menší než 0.10, hypotézu o dvourozměrné normalitě náhodného výběru zamítáme na hladině významnosti  $\alpha = 0.10$ .

Nakonec provedeme Roystonův test dvourozměrné normality pomocí funkce `mvn()` s nastavením argumentu `mvnTest = 'royston'`.

```
137 MVN::mvn(udaje, mvnTest = 'royston')$multivariateNormality
```

|   | Test    | H        | p value    | MVN |
|---|---------|----------|------------|-----|
| 1 | Royston | 6.417614 | 0.04040737 | NO  |

138  
139

Protože  $p$ -hodnota = 0.0404074 je menší než 0.10, hypotézu o dvourozměrné normalitě náhodného výběru zamítáme na hladině významnosti  $\alpha = 0.10$ .

#### 4. Grafická vizualizace výsledků testování

Rozdělení náhodného výběru vizualizujeme 3D grafem jádrového odhadu hustoty získaného na základě naměřených hodnot a tečkovým diagramem superponovaným elipsou spolehlivosti.

Jádrový odhad vypočítáme pomocí funkce `kde2d()` z knihovny `MASS`. Pro získání komplexnějšího pohledu na data spočítáme jádrový odhad ve směru proměnné `head.L` v rozsahu 166–211 mm a ve směru proměnné `bizyg.W` v rozsahu 115–154 mm. 3D graf vykreslíme pomocí funkce `persp()`, kde jako první tři argumenty zadáme  $x$ -ové,  $y$ -ové a  $z$ -ové souřadnice bodů jádrového odhadu hustoty. Argumentem `theta` natočíme graf v horizontálním směru o 20° po směru hodinových ručiček a argumentem `phi` natočíme graf ve vertikálním směru o 40° směrem k nám. Dále nastavíme barevnou škálu grafu tak, aby s rostoucí výškou grafu docházelo ke změně barev z palety `heat.colors()` ve škále 12 odstínů od červené až po žlutou.

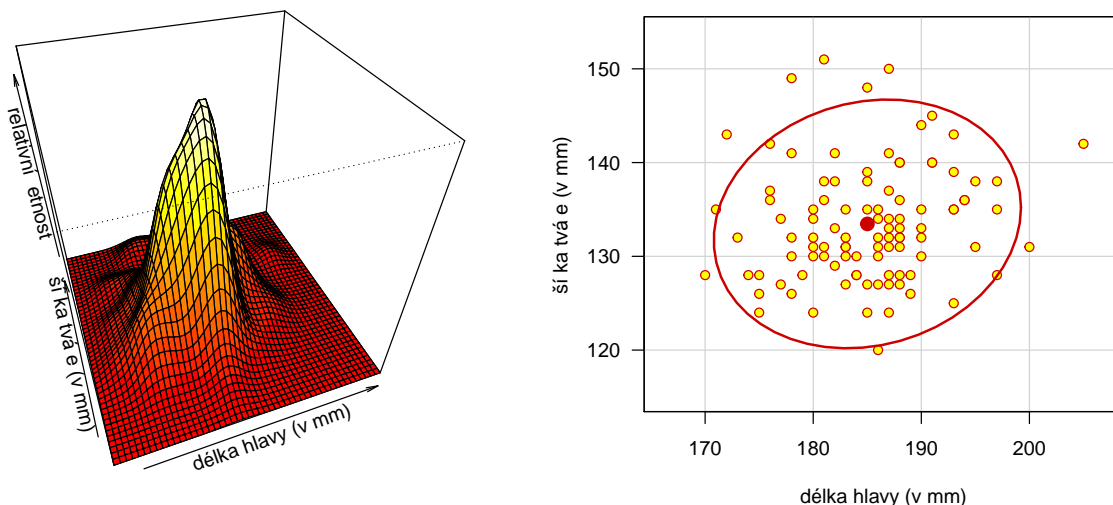
```
140 souradnice <- MASS::kde2d(udaje$head.L, udaje$bizyg.W, n = 50, lims = c(166, 211, 115, 154))
141 n <- dim(souradnice$z)[1]
142 vyska <- souradnice$z[-1, -1] + souradnice$z[-1, -n] + souradnice$z[-n, -1] + souradnice$z[-n, -n]
143 vyska <- cut(vyska, 12)
144
145 persp(souradnice$x, souradnice$y, souradnice$z, xlab = 'délka hlavy (v mm)',
146       ylab = 'šířka tváře (v mm)', zlab = "relativní četnost",
147       theta = -20, phi = 40, col = heat.colors(12)[vyska])
```

Dále vykreslíme tečkový diagram superponovaný 90% elipsou spolehlivosti pomocí funkce `dataEllipse()` z knihovny `car`. Prvními dvěma argumenty jsou proměnné, nad nimiž chceme elipsu sestrojít. Argumentem `level` dále stanovíme hodnotu koeficientu spolehlivosti  $(1 - \alpha) = 0.90$ . Pomocí argumentů `pch`, `bg` a `col` nastavíme vykreslení kulatých bodů se žlutým vnitřkem a tmavě červeným obrysem, tmavě červené kontury elipsy i jejího středu.

```

148 car::dataEllipse(udaje$head.L, udaje$bizyg.W, level = 0.90,
149                 xlim = c(166, 207), ylim = c(115, 154), main = '',
150                 xlab = 'délka hlavy (v mm)', ylab = 'šířka tváře (v mm)',
151                 pch = 21, col = 'red3', bg = 'yellow', lwd = 2, las = 1)

```



Obrázek 6: 3D graf a tečkový diagram s 90% elipsou spolehlivosti pro délku hlavy a šířku tváře žen (v mm)

Z obrázku 6 není porušení předpokladu normality příliš patrné. 3D graf vizualizuje data jako jeden pospolitý, i když poněkud špičatější vrchol s několika mírnými hrbolky značícími odlehlá pozorování. V tečkovém diagramu je pro splnění předpokladu dvourozměrné normality potřeba aby 90% elipsa spolehlivosti pokrývala alespoň 90 bodů (90 % bodů) a nejvýše 10 bodů smí ležet mimo elipsu. Mimo elipsu spolehlivosti leží právě 10 bodů, což je sice na hraně ale v pořádku. V tomto případě tedy grafická vizualizace není v přímém souladu s výsledky testování. I přesto se přikloníme k závěrům testů dvourozměrné normality.

## 5. Interpretace výsledků

Datový soubor obsahující údaje o délce hlavy a šířce tváře žen nepochází z dvourozměrného normálního rozdělení, a jako takový nemůže být použit jako základ k parametrickému testu o korelačním koeficientu  $\rho$ .



### Příklad 6.7. Test o dvourozměrné normalitě dat

Mějme datový soubor 01-one-sample-mean-skull-mf.txt a proměnnou skull.L popisující délku lebky v mm a proměnnou skull.B popisující šířku lebky v mm (viz sekce ??). Na hladině významnosti  $\alpha = 0.05$  testujte hypotézu, že náhodný výběr délek lebky a šířek lebky mužů starověké egyptské populace pochází z dvourozměrného normálního rozdělení.

#### Řešení příkladu 6.7

Nejprve načteme datový soubor a pomocí operátoru [] vybereme z tabulky pouze řádky týkající se mužů (sex == 'm') a sloupce obsahující naměřené hodnoty délky lebky (skull.L) a šířky lebky (skull.B). Vše vložíme do proměnné udaje. Příkazem na.omit() odstraníme z proměnné udaje neznámé hodnoty a příkazem dim() zjistíme počet subjektů, jejichž údaje máme k dispozici.

```
152 data <- read.delim('00-Data//01-one-sample-mean-skull-mf.txt')
153 udaje <- data[data$sex == 'm' , c('skull.L', 'skull.B')]
154 udaje <- na.omit(udaje)
155 dim(udaje) # 216x2
156 range(udaje$skull.L) # 164-199
157 range(udaje$skull.B) # 124-149
```

Náhodný výběr obsahuje údaje o dvou proměnných, délce lebky a šířce lebky, u 216 mužů. Naměřené hodnoty délky lebky se pohybují v rozmezí 164-199 mm, naměřené hodnoty šířky lebky se pohybují v rozmezí 124-149 mm. K otestování hypotézy o dvourozměrné normalitě použijeme Mardiov, Henzův-Zirklerův a Roystonův test.

#### 1. Stanovení hypotéz

- **slovní formulace** nulové a alternativní hypotézy  
 $H_0$  : Data pochází z dvourozměrného normálního rozdělení.  
 $H_1$  : Data nepochází z dvourozměrného normálního rozdělení.
- **matematická formulace** nulové a alternativní hypotézy  
 $H_0$  :  $(X, Y)^T \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$   
 $H_1$  :  $(X, Y)^T \not\sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

#### 2. Volba hladiny významnosti

Ze zadání volíme hladinu významnosti  $\alpha = 0.05$ .

#### 3. Testování $p$ -hodnotou

Mardiov test provedeme pomocí funkce mvn() se specifikací argumentu mvnTest = 'mardia'.

```
158 MVN::mvn(udaje , mvnTest = 'mardia')$multivariateNormality
```

|   | Test            | Statistic          | p value            | Result |     |
|---|-----------------|--------------------|--------------------|--------|-----|
| 1 | Mardia Skewness | 11.0266139489775   | 0.0262665328995546 | NO     | 159 |
| 2 | Mardia Kurtosis | -0.160547155018785 | 0.872450079003065  | YES    | 160 |
| 3 | MVN             | <NA>               | <NA>               | NO     | 161 |

Protože  $p$ -hodnota o nevýznamnosti koeficientu šikmosti, tj. 0.02627, je menší než 0.05, hypotézu o nevýznamnosti koeficientu šikmosti zamítáme na hladině významnosti  $\alpha = 0.05$ . Data jsou statisticky významně zešikmená. Jelikož  $p$ -hodnota testu o nevýznamnosti koeficientu špičatosti, tj. 0.8725, je větší než  $\alpha = 0.05$  hypotézu o nevýznamnosti koeficientu špičatosti nezamítáme na hladině významnosti  $\alpha = 0.05$ . Data nevykazují statisticky významnou špičatost. Protože náhodný výběr vykazuje statisticky významné zešikmení, zamítáme hypotézu o jeho dvourozměrné normalitě.

Henzův-Zirklerův test provedeme pomocí funkce mvn() se specifikací argumentu mvnTest = 'hz'.

```
163 MVN::mvn(udaje , mvnTest = 'hz')$multivariateNormality
```

|   | Test          | HZ        | p value  | MVN |
|---|---------------|-----------|----------|-----|
| 1 | Henze-Zirkler | 0.8213024 | 0.186025 | YES |

164  
165

Protože  $p$ -hodnota = 0.18602 je větší než 0.05, hypotézu o dvourozměrné normalitě náhodného výběru nezamítáme na hladině významnosti  $\alpha = 0.05$ .

Nakonec provedeme Roystonův test pomocí funkce `mvn()` se specifikací argumentu `mvnTest = 'royston'`.

```
166 MVN::mvn(udaje, mvnTest = 'royston')$multivariateNormality
```

|   | Test    | H        | p value   | MVN |
|---|---------|----------|-----------|-----|
| 1 | Royston | 2.320783 | 0.3134003 | YES |

167  
168

Protože  $p$ -hodnota = 0.3134 je větší než 0.05, hypotézu o dvourozměrné normalitě náhodného výběru nezamítáme na hladině významnosti  $\alpha = 0.05$ .

Henzův-Zirklerův test a Roystonův test  $H_0$  nezamítají, Mardiov test naopak  $H_0$  zamítá. Před stanovením závěru o rozdělení náhodného výběru se podíváme na grafickou vizualizaci dat.

#### 4. Grafická vizualizace výsledků testování

Rozdělení náhodného výběru vizualizujeme 3D grafem jádrového odhadu hustoty a tečkovým diagramem. Jádrový odhad vypočítáme pomocí funkce `kde2d()` na ploše o rozsahu 160–203 mm ve směru proměnné `skull.L` a rozsahu 120–153 mm ve směru proměnné `skull.B`. 3D graf vykreslíme pomocí funkce `persp()`, kde argumentem `col` nastavíme barevnou škálu grafu tak, aby s rostoucí výškou grafu docházelo ke změně barev v devíti odstínech 'GnBu' z palety `brewer.pal` z knihovny `RColorBrewer`, a to vzestupně od modré po bílou.

```
169 souradnice <- MASS::kde2d(udaje$skull.L, udaje$skull.B, n = 50,
170                          lims = c(160, 203, 120, 153))
171 n <- dim(souradnice$z)[1]
172 vyska <- souradnice$z[-1, -1] + souradnice$z[-1, -n] +
173        souradnice$z[-n, -1] + souradnice$z[-n, -n]
174 vyska <- cut(vyska, 9)
175 persp(souradnice$x, souradnice$y, souradnice$z, xlab = 'délka lebky (v mm)',
176       ylab = 'šířka lebky (v mm)', zlab = "relativní četnost",
177       theta = 0, phi = 40, col = rev(RColorBrewer::brewer.pal(9, 'GnBu'))[vyska])
```

Dále vykreslíme tečkový diagram superponovaný 95% elipsou spolehlivosti.

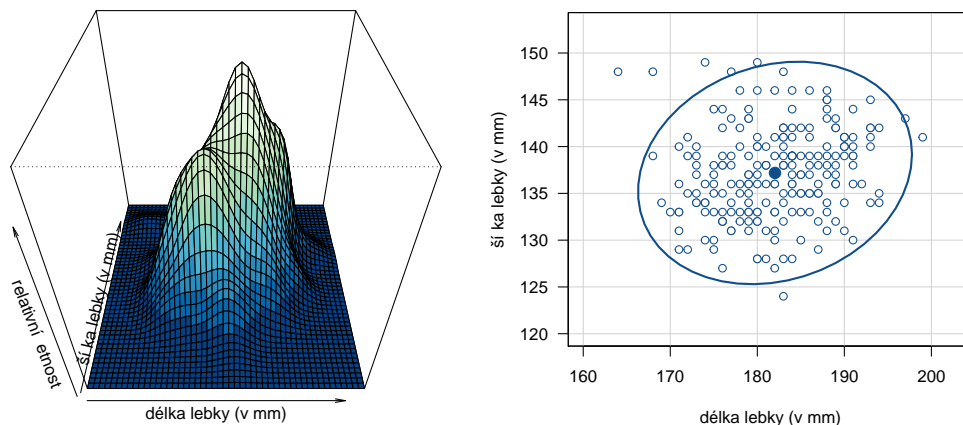
```
178 car::dataEllipse(udaje$skull.L, udaje$skull.B, level = 0.95, xlim = c(160, 203),
179                 ylim = c(120, 153), main = '', xlab = 'délka lebky (v mm)',
180                 ylab = 'šířka lebky (v mm)', pch = 21, col = 'dodgerblue4',
181                 bg = 'white', lwd = 2, las = 1)
```

3D graf nám ukazuje pospolitě normální rozdělení. V tečkovém grafu by alespoň 95% hodnot, tj. 205 bodů, mělo ležet uvnitř elipsy spolehlivosti a nejvýše 11 bodů smí ležet mimo oblast elipsy. V našem případě leží mimo elipsu spolehlivosti pouze 8 hodnot. Po shlédnutí 3D grafu a tečkového diagramu se kloníme k závěru Henzova-Zirklerova a Roystonova testu.

#### 5. Interpretace výsledků

Náhodný výběr délek lebky a šířek lebky mužů starověké egyptské populace pochází z dvourozměrného normálního rozdělení.





Obrázek 7: 3D graf a tečkový diagram s 95% elipsou spolehlivosti pro délku lebky a šířku lebky mužů (v mm)

*Poznámka:* Balíček RColorBrewer disponuje širokou nabídkou barevných palet, jako např. 'YlOrRd' poskytující odstíny od světle žluté po sytě červenou, 'YlGn' pokrývající odstíny od světle žluté po sytě zelenou, 'PuRd' poskytující odstíny od bílé po purpurovou, nebo 'Blues' pokrývající odstíny od bílé po tmavě modrou. Přehled všech barevných škál, které balíček RColorBrewer poskytuje, lze zobrazit příkazem `display.brewer.all(n = NULL, type = 'all')`.



### Příklad 6.8. Test o dvourozměrné normalitě dat

Mějme datový soubor 19-more-samples-correlations-skull.txt a proměnnou nose.B popisující šířku nosu v mm a proměnnou intorb.B popisující interorbitální šířku v mm (viz sekce ??). Na hladině významnosti  $\alpha = 0.05$  testujte hypotézu, že náhodný výběr šířky nosu a interorbitální šířky mužů peruánské populace pochází z dvourozměrného normálního rozdělení.

#### Řešení příkladu 6.8

Nejprve načteme datový soubor a vybereme z něj pouze řádky týkající se mužů peruánské populace (`pop == 'per'`) a sloupce obsahující naměřené hodnoty šířky nosu (`nose.B`) a interorbitální šířky (`intorb.B`). Vše vložíme do proměnné udaje. Z proměnné udaje odstraníme neznámé hodnoty a zjistíme počet subjektů, jejichž údaje máme k dispozici.

```
182 data <- read.delim('00-Data//19-more-samples-correlations-skull.txt')
183 udaje <- data[data$pop == 'per', c('nose.B', 'intorb.B')]
184 udaje <- na.omit(udaje)
185 dim(udaje) # 46x2
186 range(udaje$nose.B) # 19-26
187 range(udaje$intorb.B) # 19-28
```

Náhodný výběr obsahuje údaje o šířce nosu a interorbitální šířce u 46 mužů peruánské populace. Naměřené hodnoty šířky nosu se pohybují v rozmezí 19–26 mm, naměřené hodnoty interorbitální šířky se pohybují v rozmezí 19–28 mm. K otestování hypotézy o dvourozměrné normalitě použijeme Mardiov, Henzův-Zirklerův i Roystonův test.

#### 1. Stanovení hypotéz

- **slovní formulace** nulové a alternativní hypotézy  
 $H_0$  : Data pochází z dvourozměrného normálního rozdělení.  
 $H_1$  : Data nepochází z dvourozměrného normálního rozdělení.
- **matematická formulace** nulové a alternativní hypotézy  
 $H_0$  :  $(X, Y)^T \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$   
 $H_1$  :  $(X, Y)^T \not\sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

#### 2. Volba hladiny významnosti

Podle zadání volíme hladinu významnosti  $\alpha = 0.05$ .

#### 3. Testování $p$ -hodnotou

Nejprve provedeme Mardiov test.

```
188 MVN::mvn(udaje, mvnTest = 'mardia')$multivariateNormality
```

|   | Test            | Statistic           | p value           | Result |
|---|-----------------|---------------------|-------------------|--------|
| 1 | Mardia Skewness | 4.27819772855481    | 0.369663150730262 | YES    |
| 2 | Mardia Kurtosis | -0.0684871107744411 | 0.945397880096616 | YES    |
| 3 | MVN             | <NA>                | <NA>              | YES    |

189  
190  
191  
192

Protože  $p$ -hodnota o nevýznamnosti koeficientu šikmosti, tj. 0.3697, je větší než 0.05, hypotézu o nevýznamnosti koeficientu šikmosti nezamítáme na hladině významnosti  $\alpha = 0.05$ . Jelikož  $p$ -hodnota testu o nevýznamnosti koeficientu špičatosti, tj. 0.9454, je větší než 0.05, hypotézu o nevýznamnosti koeficientu špičatosti nezamítáme na hladině významnosti  $\alpha = 0.05$ . Protože náhodný výběr nevykazuje statisticky významné zešikmení ani zešpičatění, nezamítáme hypotézu o dvourozměrné normalitě náhodného výběru.

Nyní provedeme Henzův-Zirklerův test dvourozměrné normality.

```
193 MVN::mvn(udaje, mvnTest = 'hz')$multivariateNormality
```

|   | Test          | HZ        | p value | MVN |
|---|---------------|-----------|---------|-----|
| 1 | Henze-Zirkler | 0.5159026 | 0.39524 | YES |

194  
195

Protože  $p$ -hodnota = 0.3952 je větší než 0.05, hypotézu o dvourozměrné normalitě náhodného výběru nezamítáme na hladině významnosti  $\alpha = 0.05$ .

Nakonec otestujeme hypotézu o dvourozměrné normalitě pomocí Roystonova testu.

```
196 MVN::mvn(udaje, mvnTest = 'royston')$multivariateNormality
```

|   | Test    | H        | p value    | MVN |
|---|---------|----------|------------|-----|
| 1 | Royston | 8.396178 | 0.01505224 | NO  |

197  
198

Protože  $p$ -hodnota = 0.015052 je menší než 0.05, hypotézu o dvourozměrné normalitě náhodného výběru zamítáme na hladině významnosti  $\alpha = 0.05$ .

Mardiův test a Henzův-Zirklerův test  $H_0$  nezamítají, Roystonův test naopak  $H_0$  zamítá. Před stanovením závěru o nulové hypotéze se podíváme na grafickou vizualizaci dat.

#### 4. Grafická vizualizace výsledků testování

Rozdělení náhodného výběru vizualizujeme 3D grafem jádrového odhadu hustoty získaného na základě naměřených hodnot a tečkovým diagramem s 95% elipsou spolehlivosti. 3D graf vykreslíme pomocí funkce `persp()`, kde argumentem `col` nastavíme barevnou škálu grafu tak, aby s rostoucí výškou grafu docházelo ke změně barev v devíti odstínech 'YlOrBr' z palety `brewer.pal` z knihovny `RColorBrewer`, a to vzestupně od hnědé po bílou.

```
199 souradnice <- MASS::kde2d(udaje$nose.B, udaje$intorb.B, n = 50,
200                          lims = c(17, 28, 17, 30))
201 n <- dim(souradnice$z)[1]
202 vyska <- souradnice$z[-1, -1] + souradnice$z[-1, -n] +
203          souradnice$z[-n, -1] + souradnice$z[-n, -n]
204 vyska <- cut(vyska, 9)
205 par(mar = c(4, 4, 1, 1))
206 persp(souradnice$x, souradnice$y, souradnice$z, xlab = 'šířka nosu (v mm)',
207        ylab = 'interorbitální šířka (v mm)', zlab = "relativní četnost",
208        col = rev(RColorBrewer::brewer.pal(9, 'YlOrBr'))[vyska],
209        theta = 20, phi = 40)
210
211 car::dataEllipse(udaje$nose.B, udaje$intorb.B, level = 0.95,
212                  xlim = c(18, 28), ylim = c(17, 29), xlab = 'šířka nosu (v mm)',
213                  ylab = 'interorbitální šířka (v mm)', main = '', pch = 21,
214                  col = 'sienna4', bg = 'navajowhite1', lwd = 2, las = 1)
```

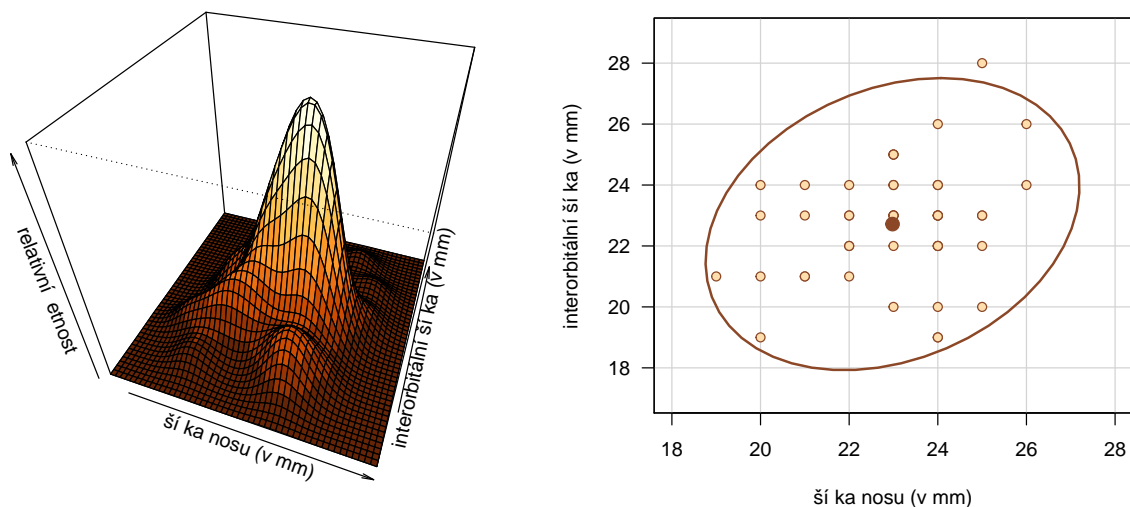
3D graf zachycuje charakter dvourozměrného normálního rozdělení náhodného výběru s několika odlehlými hodnotami, které však ještě spadají do 95% elipsy spolehlivosti (viz tečkový diagram). V tečkovém diagramu by alespoň 95 % hodnot, tj. 44 bodů, mělo ležet uvnitř elipsy spolehlivosti a nejvýše dva body se smí realizovat mimo oblast elipsy. Z diagramu vidíme, že mimo elipsu spolehlivosti leží pouze jeden bod. Na základě pohledu na 3D graf a tečkový diagram se kloníme k výsledkům Mardiaova a Henzova-Zirklerova testu o dvourozměrné normalitě náhodného výběru.

#### 5. Interpretace výsledků

Náhodný výběr popisující šířku nosu a interorbitální šířku u mužů peruánské populace pochází z dvourozměrného normálního rozdělení.



*Poznámka:* V příkladech 6.6, 6.7 a 6.8 jsme se setkali se situací, kdy se testy dvourozměrné normality rozcházejí v názoru na zamítnutí či nezamítnutí nulové hypotézy. Při analýze reálných dat nejde o zcela neobvyklý jev. Každý ze



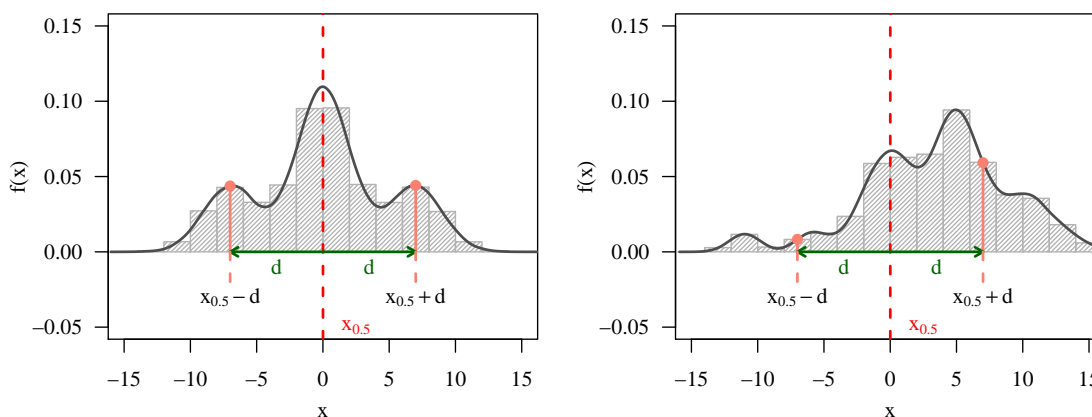
Obrázek 8: 3D graf a tečkový diagram s 95% elipsou spolehlivosti pro šířku nosu a interorbitální šířku (v mm) mužů peruánské populace

zmíněných testů je založen na jiném algoritmu, v rámci kterého vyhodnocuje potenciální dvourozměrnou normalitu náhodného výběru na základě svých vlastních kritérií (nevýznamnost koeficientu šikmosti a koeficientu špičatosti, porovnání teoretické hustoty a jádrového odhadu hustoty, apod). Proto při posuzování dvouzorněrné normality dat je vhodné provést více testů, nezanedbat ani grafickou vizualizaci dat a závěr o nulové hypotéze stanovit po důkladném zvážení všech aspektů a vyhodnocení získaných výsledků.

### 6.3 Testy symetrie

Nepochází-li datový soubor z normálního rozdělení, musíme hypotézu o parametru  $\mu$  otestovat pomocí jednovýběrového neparametrického testu (viz kapitola ??). Nejčastěji používaným testem je v takovém Wilcoxonův jednovýběrový test (viz sekce ?? pro exaktní variantu testu a sekce ?? pro asymptotickou variantu testu). Předpokladem Wilcoxonova jednovýběrového testu je pouhá symetrie náhodného výběru okolo mediánu. V případě, že ani tento předpoklad není splněn, musíme na otestování hypotézy o parametru  $\mu$  použít znaménkový test (viz sekce ?? pro exaktní variantu testu a sekce ?? pro asymptotickou variantu testu).

Nechť  $X_1, \dots, X_N$  je náhodný výběr z nějakého rozdělení  $L(\theta)$ , kde  $\theta$  je obecně vektor parametrů rozdělení  $L$ . Na hladině významnosti  $\alpha = 0.05$  testujeme hypotézu  $H_0$ : *Náhodný výběr pochází z rozdělení symetrického okolo mediánu*. oproti alternativní hypotéze  $H_1$ : *Náhodný výběr nepochází z rozdělení symetrického okolo mediánu*. Matematický zápis nulové a alternativní hypotézy definujeme pomocí hustoty  $f(x)$ . Pokud jsou data symetrická okolo mediánu, znamená to, že hodnota hustoty v libovolném bodě vzdáleném od mediánu  $x_{0.50}$  o vzdálenost  $d$  směrem doleva (značíme  $x_{0.50} - d$ ) je stejná jako hodnota hustoty v bodě vzdáleném od mediánu  $x_{0.50}$  o vzdálenost  $d$  směrem doprava (značíme  $x_{0.50} + d$ ), tj.  $f(x_{0.50} - d) = f(x_{0.50} + d)$ . Naopak pokud data nejsou symetrická okolo mediánu, potom se hodnoty hustoty v bodech  $x_{0.50} - d$  a  $x_{0.50} + d$  nerovnájí, tj.  $f(x_{0.50} - d) \neq f(x_{0.50} + d)$ . Graficky je ukázka hustoty v bodech  $x_{0.50} - d$  a  $x_{0.50} + d$  při symetrickém i nesymetrickém rozdělení znázorněna na obrázku 9.



Obrázek 9: Ukázka hodnot hustoty v bodech  $x_{0.50} - d$  a  $x_{0.50} + d$  u (a) symetrického rozdělení (vlevo) náhodného výběru; (b) nesymetrického rozdělení (vpravo) náhodného výběru

K testování hypotézy o symetrii můžeme použít Miaoův test (Miao et al., 2006), Cabiliův-Masarův test (Cabilio a Masaro, 1996) nebo Miraův test (Mira, 1999) symetrie okolo neznámého mediánu. Všechny tři testy jsou k dispozici ve funkci `symmetry.test()`. Prvním, a jediným povinným vstupním argumentem funkce je vektor naměřených hodnot. Volbu Miaoův testu zvolíme nastavením argumentu `option = 'MGG'` (viz příklad 6.5). Poznamenejme, že tento test je defaultním testem, který se provede, pokud parametr `option` nespecifikujeme. Volbu Cabiliova-Masarova testu nastavením argumentu `option = 'CM'` a volbu Miraův testu specifikujeme argumentem `option = 'M'` ve funkci `symmetry.test()`. Dalším argumentem, který můžeme specifikovat je argument `side`. Nastavením hodnoty argumentu `side == 'both'` (defaultní nastavení) testujeme nulovou hypotézu o symetrii oproti alternativní hypotéze  $H_1$ : *Náhodný výběr nepochází ze symetrického rozdělení*. Nastavením hodnoty argumentu `side == 'right'` testujeme nulovou hypotézu o symetrii oproti alternativní hypotéze  $H_1$ : *Náhodný výběr pochází z rozdělení vyšikmeného doprava*. Nastavením hodnoty argumentu `side == 'left'` potom testujeme nulovou hypotézu o symetrii oproti alternativní hypotéze  $H_1$ : *Náhodný výběr pochází z rozdělení vyšikmeného doleva*. Posledním argumentem, který je nutné specifikovat, je argument `boot`. Nastavením tohoto argumentu na hodnotu `F`, tj. `boot = F` zakážeme (defaultní) výpočet  $p$ -hodnoty na základě bootstrapové metody. Bootstrapová metoda je poměrně novou statistickou metodou, jejímž principem se zde detailněji nebudeme zabývat. Poznamenejme akorát, že základní princip bootstrapové metody je založen na náhodném převzorkování náhodného výběru. Pointa je, že při každém spuštění metody dochází k novému převzorkování náhodného výběru, které se může lišit od převzorkování vytvořeného při předchozím spuštění. Proto i distribuční funkce, kterou odhadujeme nikoliv na základě původního náhodného výběru, ale na základě převzorkovaného náhodného výběru, je při každém spuštění mírně odlišná, stejně jako výsledná  $p$ -hodnota

vypočítaná právě na základě takto odhadnuté distribuční funkce. Pro naše potřeby však bootstrapová metoda není potřebná, proto ji nastavením argumentu `boot = F` vypneme a k výpočtu  $p$ -hodnoty zvolíme standardní exaktní výpočet.

Symetrické rozdělení náhodného výběru vizualizujeme histogramem superponovaným křivkou jádrového odhadu funkce hustoty doplněným o svislou referenční přímkou v hodnotě mediánu. Pokud bude křivka jádrového odhadu symetrická okolo mediánu, můžeme předpokládat symetrické rozdělení naměřených hodnot. Pro doplnění představy o symetrii náhodného výběru můžeme také vykreslit krabicový diagram. Data považujeme za symetrická, je-li diagram symetrický okolo mediánu. Oba grafy nám tak pomohou udělat si náhled na symetrii náhodného výběru a případně odhalit příčinu jejího porušení.

*Poznámka:* Závěrem poznamenejme, že důvodem, proč si zde prezentujeme tři testy symetrie, je poskytnout čtenáři širší spektrum možných testů symetrie a umožnit mu tak vybrat si z předložených testů ten nejsympatičtější. I z toho důvodu u každého z následujících příkladů provádíme testování nulové hypotézy o symetrii naměřených dat okolo mediánu pomocí Miaové testu, Cabliova-Masarova testu i Miraové testu. Při provádění vlastních analýz ovšem nastává odlišná situace. Zde zpravidla volíme na otestování hypotézy o symetrii pouze jeden univerzální test, který používáme pokaždé, když chceme hypotézu o symetrii testovat. V žádném případě neprovádíme všechny tři testy symetrie najednou a v případě rozporuplných závěrů testů (jako je tomu například v příkladu 6.10) si nevybíráme ten, jehož výsledek se nám aktuálně nejvíce hodí. Vždy mějme na paměti, že ve vlastní analýze volíme použité testy univerzálně, bez ohledu na to, jak příhodný či nepříhodný výsledek nám v konkrétní situaci předloží.

Více testů symetrie smíme provádět pouze v hraničních případech, kdy univerzálně zvolený test hypotézu o symetrii hraničně zamítá nebo nezamítá. Pak je vhodné provést i další testy, které nám mohou pomoci nahlédnout na symetrii náhodného výběru z jiného pohledu a lépe tak zvážit skutečný charakter symetrie naměřených hodnot. Pokud se na základě testů rozhodneme pro jiný závěr, než jaký nám poskytuje test univerzálně používaný na otestování hypotézy o symetrii v celé datové analýze, je vždy třeba svůj závěr podložit relevantními výsledky a, je-li to možné, taktéž grafickou vizualizací. Vždy mějme na paměti, že závěry o nulové hypotéze se snažíme vždy provádět co nejobektivněji tak, aby odrážely skutečnost, nikoliv naše přání.

*Poznámka:* Pokud si mezi prezentovanými testy nedokážete vybrat favoritní test, doporučujeme vám zůstat u volby Miaové testu (defaultní nastavení funkce `symmetry.test()`). Jde o nejnovější test, který disponuje nejhodnějšími vlastnostmi. Je však třeba mít na paměti, že je velmi citlivý na extrémní pozorování. **Můj tip, pokud testy detailněji znáte a doporučili byste mi jiný test jako favoritní, můžeme doporučení změnit.**

### Příklad 6.9. Test symetrie

Máme datový soubor 21-goldman-measures.csv obsahující naměřené hodnoty anteroposteriorního průměru diafýzy pažní kosti z pravé strany (humer.ADR) a z levé strany (humer.ADL) u skeletů mužského a ženského pohlaví z pohřebiště z oblasti Staubing. Na hladině významnosti  $\alpha = 0.05$  testujte hypotézu, že náhodný výběr anteroposteriorního průměru diafýzy pažní kosti z pravé strany u skeletů mužského pohlaví pochází ze symetrického rozdělení.

### Řešení příkladu 6.9

Nejprve načteme datový soubor příkazem `read.delim()`, a následně pomocí operátoru `[]` vybereme z tabulky řádky týkající se skeletů mužského pohlaví (`sex == 'm'`) z pohřebiště z oblasti Staubing (`pop == 'Staubing'`) a sloupec `humer.ADR` obsahující naměřené hodnoty anteroposteriorního průměru diafýzy pažní kosti. Tento výběr vložíme do proměnné `humer.ADRM`. Příkazem `na.omit()` odstraníme z proměnné `humer.ADRM` neznámé hodnoty a příkazem `length()` zjistíme rozsah náhodného výběru a příkazem `range()` rozpětí naměřených hodnot. Nakonec, jelikož budeme testovat hypotézu o symetrickém rozdělení naměřených hodnot okolo mediánu, si vypočítáme hodnotu mediánu pomocí funkce `median()` s volbou argumentu `type = 5`.

```
215 data <- read.delim('00-Data//21-goldman-measures.csv', sep = ';', dec = '.')
216 humer.ADRM <- data[data$sex == 'm' & data$pop == 'Staubing', 'humer.ADR']
217 humer.ADRM <- na.omit(humer.ADRM)
218 n <- length(humer.ADRM) # 12
219 range(humer.ADRM) # 20.32-24.99
220 x0.50 <- median(humer.ADRM, type = 5) # 22.59
```

Datový soubor obsahuje údaje o anteroposteriorním průměru diafýzy pažní kosti z pravé strany u 12 skeletů mužského pohlaví. Naměřené hodnoty se pohybují v rozmezí 20.32–24.99 mm. Medián naměřených hodnot  $x_{0.50} = 22.59$  mm. K otestování hypotézy o symetrickém rozdělení náhodného výběru použijeme postupně Miaové test, Cabiliův-Masarův test i Miraové test. Testování provedeme v posloupnosti pěti kroků.

#### 1. Stanovení hypotéz

- **slovní formulace** nulové a alternativní hypotézy  
 $H_0$  : Data pochází z rozdělení symetrického okolo mediánu.  
 $H_1$  : Data nepochází z rozdělení symetrického okolo mediánu.
- **matematická formulace** nulové a alternativní hypotézy  
 $H_0$  :  $f(x_{0.50} - x) = f(x + x_{0.50})$   
 $H_1$  :  $f(x_{0.50} - x) \neq f(x + x_{0.50})$

#### 2. Volba hladiny významnosti

V souladu se zadáním volíme hladinu významnosti  $\alpha = 0.05$ .

#### 3. Testování $p$ -hodnotou

Miaové test provedeme pomocí funkce `symmetry.test()` z knihovny `lawstat`. Vstupní parametry bude vektor naměřených hodnot `humer.ADRM`, argument `boot` nastavený na hodnotu `F` (`FALSE`) a argument `option = 'MGG'` značící volbu Miaové testu. Výstupem funkce je údaj o názvu proměnné použité k testování hypotézy o symetrickém rozdělení naměřených hodnot okolo mediánu (`humer.ADRM`), hodnota testovací statistiky Miaové testu `Test statistic = -0.10873`,  $p$ -hodnota `p-value = 0.9134` a znění alternativní hypotézy `alternative hypothesis: the distribution is asymmetric`.

```
221 lawstat::symmetry.test(humer.ADRM, boot = F, option = 'MGG')
```

```
Symmetry test by Miao, Gel, and Gastwirth (2006)
data: humer.ADRM
Test statistic = -0.10873, p-value = 0.9134
alternative hypothesis: the distribution is asymmetric.
```

222  
223  
224  
225  
226  
227

### Závěr testování:

Protože  $p$ -hodnota = 0.9134 je větší než  $\alpha = 0.05$ ,  $H_0$  nezamítáme na hladině významnosti  $\alpha = 0.05$ .

Dále otestujeme hypotézu o symetrickém rozdělení naměřených hodnot okolo mediánu také pomocí Cabiliova-Masarova testu, a to pomocí funkce `symmetry.test()` s nastavením argumentu `option = 'CM'`. Výstupem funkce je název použité proměnné `humer.ADRM`, hodnota testovací statistiky Cabiliova-Masarova testu `Test statistic = -0.10459`,  $p$ -hodnota `p-value = 0.9167` a znění alternativní hypotézy.

```
228 lawstat::symmetry.test(humer.ADRM, boot = F, option = 'CM')
```

```
Symmetry test by Cabilio and Masaro (1996)
data: humer.ADRM
Test statistic = -0.10459, p-value = 0.9167
alternative hypothesis: the distribution is asymmetric.
```

229  
230  
231  
232  
233  
234

### Závěr testování:

Protože  $p$ -hodnota = 0.9167 je větší než  $\alpha = 0.05$ ,  $H_0$  nezamítáme na hladině významnosti  $\alpha = 0.05$ .

Nakonec otestujeme hypotézu o symetrickém rozdělení naměřených hodnot okolo mediánu pomocí Miraové testu. Opět použijeme funkci `symmetry.test()`, tentokrát s nastavením argumentu `option = 'M'`. Jako u předchozích dvou testů je výstupem funkce opět název použité proměnné `humer.ADRM`, dále hodnota testovací statistiky Miraové testu `Test statistic = -0.11668`,  $p$ -hodnota `p-value = 0.9071` a znění alternativní hypotézy `alternative hypothesis: the distribution is asymmetric`.

```
235 lawstat::symmetry.test(humer.ADRM, boot = F, option = 'M')
```

```
Symmetry test by Mira (1999)
data: humer.ADRM
Test statistic = -0.11668, p-value = 0.9071
alternative hypothesis: the distribution is asymmetric.
```

236  
237  
238  
239  
240  
241

### Závěr testování:

Protože  $p$ -hodnota = 0.9071 je větší než  $\alpha = 0.05$ ,  $H_0$  nezamítáme na hladině významnosti  $\alpha = 0.05$ .

## 4. Grafická vizualizace výsledků testování

Rozdělení náhodného výběru vizualizujeme pomocí histogramu superponovaného křivkou jádrového odhadu hustoty a pomocí krabicového diagramu (viz obrázek 10). Histogram vykreslíme příkazem `hist()` analogicky, jako například v příkladu 6.1 nebo v kapitole ????. Podle Sturgesova pravidla rozdělíme datový soubor do pěti třídicích intervalů se šířkou 1 mm. Hranice intervalů budou 20, 21, ..., 25. Jádrový odhad hustoty vypočítáme příkazem `density()`. Funkce `density()` má pouze jeden povinný parametr, a sice vektor naměřených hodnot, na základě kterých jádrový odhad hustoty konstruujeme. Zde jde tedy o vektor naměřených anteroposteriorních průměrů diafýzy pažní kosti z pravé strany `humer.ADRM`. Křivku hustoty vykreslíme příkazem `lines()`. Do grafu dále pomocí příkazu `abline()` doplníme vodorovnou referenční čáru v hodnotě mediánu  $x_{0.50} = 22.59$ . Krabicový diagram potom vykreslíme příkazem `boxplot()`.

Z obou grafů na obrázku 10 je patrná symetrie naměřených hodnot okolo mediánu  $x_{0.05} = 22.59$ . Z grafu histogramu vidíme, že histogram i jádrový odhad hustoty vykazují výraznou symetrii okolo vodorovné referenční příčky v hodnotě mediánu. Krabicový diagram je potom ukázkově symetrický okolo vodorovné červené přerušované referenční příčky zastupující medián. To je v naprostém souladu se závěrem testování.

## 5. Interpretace výsledků

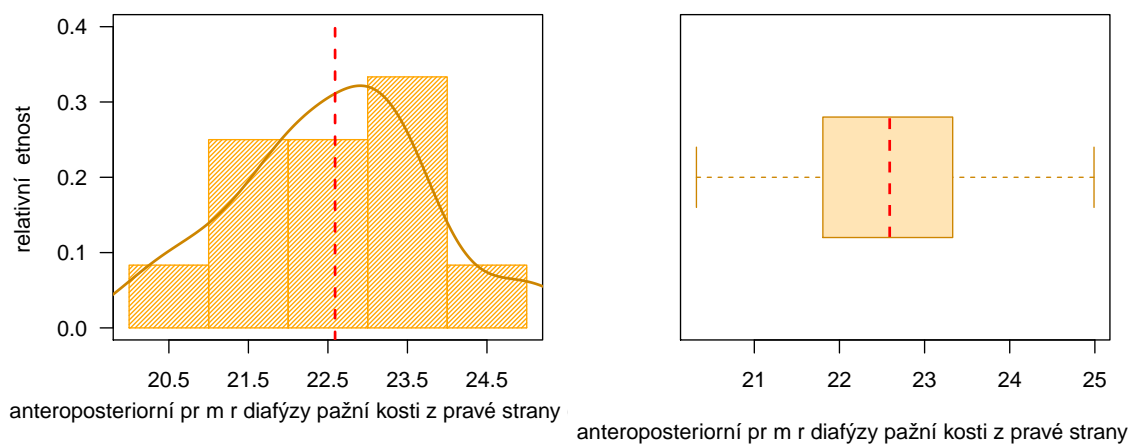
Datový soubor anteroposteriorních průměrů diafýzy pažní kosti z pravé strany u skeletů mužského pohlaví z pohřebiště v oblasti Staubing pochází z rozdělení symetrického okolo mediánu.



```

242 par(mar = c(4, 4, 1, 1))
243 hist(humer.ADRM, breaks = seq(20, 25, by = 1), prob = T, col = 'orange',
244       density = 40, main = '', ylab = 'relativní četnost', axes = F,
245       ylim = c(0, 0.40), xlab = '')
246 box(bty = 'o')
247 axis(1, seq(20.5, 24.5, by = 1))
248 axis(2, las = 1)
249 lines(density(humer.ADRM), col = 'orange3', lwd = 2)
250 abline(v = x0.50, col = 'red', lwd = 2, lty = 2)
251 mtext('anteroposteriorní průměr diafýzy pažní kosti z pravé strany (v mm)',
252       side = 1, line = 2.2)
253
254 boxplot(humer.ADRM, col = 'moccasin', horizontal = T, border = 'orange3',
255         medcol = 'red', medlty = 2, medlwd = 2, type = 5,
256         xlab = 'anteroposteriorní průměr diafýzy pažní kosti z pravé strany (v mm)')

```



Obrázek 10: Histogram a krabicový diagram anteroposteriorních průměrů diafýzy pažní kosti z pravé strany u skeletů mužského pohlaví



### Příklad 6.10. Test symetrie

Máme datový soubor 21-goldman-measures.csv obsahující údaje o epikondylární šířce pažní kosti na pravé straně (humer.EBR) a na levé straně (humer.EBL) u skeletů mužského a ženského pohlaví z aljašské populace z kmene Tigara (viz sekce ??). Na hladině významnosti  $\alpha = 0.05$  testujte hypotézu, že náhodný výběr epikondylárních šířek pažní kosti na pravé straně u skeletů ženského pohlaví pochází ze symetrického rozdělení.

### Řešení příkladu 6.10

Nejprve načteme datový soubor, a následně pomocí operátoru `[]` vybereme z tabulky naměřené hodnoty epikondylární šířce pažní kosti na pravé straně (humer.EBR) u skeletů ženského pohlaví (`sex == 'f'`) z aljašské populace z kmene Tigara (`pop == 'Tigara'`). Tento výběr vložíme do proměnné `humer.EBRF`. Z této proměnné odstraníme NA pozorování (`na.omit()`), zjistíme rozsah náhodného výběru (`length()`) a rozpětí naměřených hodnot (`range()`). Nakonec vypočítáme hodnotu mediánu (`median()`).

```
257 data <- read.delim('00-Data//21-goldman-measures.csv', sep = ';', dec = '.')
258 humer.EBRF <- data[data$sex == 'f' & data$pop == 'Tigara', 'humer.EBR']
259 humer.EBRF <- na.omit(humer.EBRF)
260 n <- length(humer.EBRF) # 25
261 range(humer.EBRF) # 44-58
262 x0.50 <- median(humer.EBRF, type = 5) # 54
```

Datový soubor obsahuje naměřené hodnoty epikondylární šířky pažní kosti na pravé straně 25 skeletů ženského pohlaví. Naměřené hodnoty se pohybují v rozmezí 44–58 mm s mediánem  $x_{0.50} = 54$  mm. K otestování hypotézy o symetrickém rozdělení náhodného výběru použijeme postupně Mioové test, Cabiliův-Masarův test i Miraové test.

#### 1. Stanovení hypotéz

- **slovní formulace** nulové a alternativní hypotézy  
 $H_0$  : Data pochází z rozdělení symetrického okolo mediánu.  
 $H_1$  : Data nepochází z rozdělení symetrického okolo mediánu.
- **matematická formulace** nulové a alternativní hypotézy  
 $H_0$  :  $f(x_{0.50} - x) = f(x + x_{0.50})$   
 $H_1$  :  $f(x_{0.50} - x) \neq f(x + x_{0.50})$

#### 2. Volba hladiny významnosti

Hladina významnosti  $\alpha = 0.05$ .

#### 3. Testování $p$ -hodnotou

Mioové test provedeme pomocí funkce `symmetry.test()` z knihovny `lawstat`. Vstupním parametry bude vektor naměřených hodnot `humer.EBRF`, argument `boot` nastavený na hodnotu `F` (potlačení výpočtu  $p$ -hodnoty bootstrapovou metodou) a argument `option = 'MGG'` značící volbu Mioové testu. Z výstupů popsaných v předchozím příkladu nás nejvíce zajímá hodnota tetovací statistiky Mioové testu `Test statistic = -3.3427` a  $p$ -hodnota `p-value = 0.0008298`.

```
263 lawstat::symmetry.test(humer.EBRF, boot = F, option = 'MGG')
```

```
                Symmetry test by Miao, Gel, and Gastwirth (2006)
data: humer.EBRF
Test statistic = -3.3427, p-value = 0.0008298
alternative hypothesis: the distribution is asymmetric.
```

#### Závěr testování:

Protože  $p$ -hodnota = 0.0008298 je menší než  $\alpha = 0.05$ ,  $H_0$  zamítáme na hladině významnosti  $\alpha = 0.05$ .

Dále otestujeme hypotézu o symetrickém rozdělení naměřených hodnot okolo mediánu pomocí Cabiliova-Masarova testu. Opět použijeme funkci `symmetry.test()` tentokrát s nastavením argumentu `option = 'CM'`. Hodnota testovací statistiky Cabiliova-Masarova testu `Test statistic = -2.9952`,  $p$ -hodnota `p-value = 0.002743`.

```
270 lawstat::symmetry.test(humer.EBRF, boot = F, option = 'CM')
```

```
Symmetry test by Cabilio and Masaro (1996)
data: humer.EBRF
Test statistic = -2.9952, p-value = 0.002743
alternative hypothesis: the distribution is asymmetric.
```

271  
272  
273  
274  
275  
276

#### Závěr testování:

Protože  $p$ -hodnota = 0.002743 je menší než  $\alpha = 0.05$ ,  $H_0$  zamítáme na hladině významnosti  $\alpha = 0.05$ .

Nakonec otestujeme hypotézu o symetrickém rozdělení pomocí Miraové testu prostřednictvím funkce `symmetry.test()` s volbou argumentu `option = 'M'`. Hodnota testovací statistiky Miraové testu `Test statistic = -1.9394`,  $p$ -hodnota `p-value = 0.05245` a znění alternativní hypotézy `alternative hypothesis: the distribution is asymmetric`.

```
277 lawstat::symmetry.test(humer.EBRF, boot = F, option = 'M')
```

```
Symmetry test by Mira (1999)
data: humer.EBRF
Test statistic = -1.9394, p-value = 0.05245
alternative hypothesis: the distribution is asymmetric.
```

278  
279  
280  
281  
282  
283

#### Závěr testování:

Protože  $p$ -hodnota = 0.05245 je větší než  $\alpha = 0.05$ ,  $H_0$  nezamítáme na hladině významnosti  $\alpha = 0.05$ .

Z uvedených závěrů testování vidíme, že u testů symetrie došlo k rozporným závěrům týkajícím se nulové hypotézy o symetrii naměřených hodnot okolo mediánu. Zatímco Miraové test a Cabiliova-Masarův test hypotézu o symetrii zamítli, Miraové test hypotézu o symetrii nezamítl. Zde je důležité si uvědomit, že taková situace může reálně nastat. Každý z testů je totiž založen na jiném principu, tedy i kritérium rozhodující o nulové hypotéze se liší a v hraničních případech může dojít k rozporu mezi závěry jednotlivých testů. V takovém případě může být velmi nápomocná právě grafické vizualizace naměřených hodnot, která nám umožní udělat si obrázek o naměřených datech.

#### 4. Grafická vizualizace výsledků testování

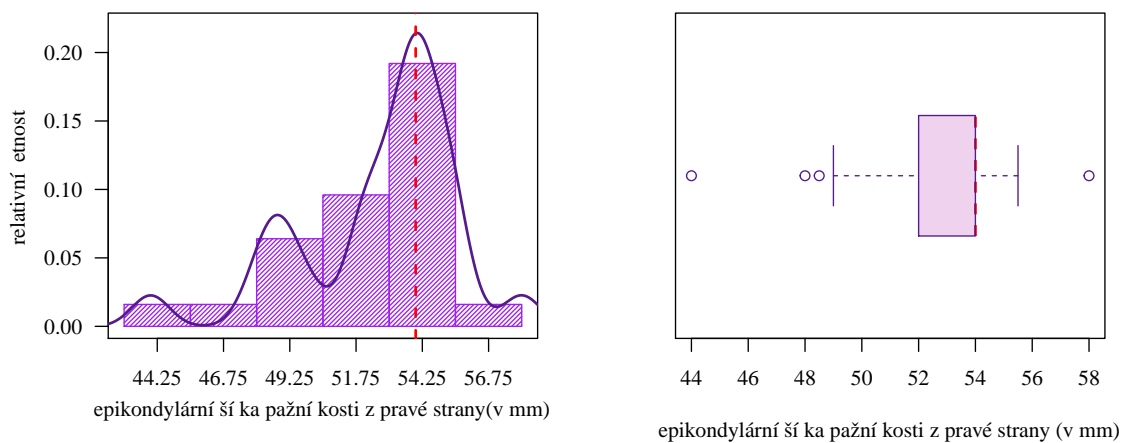
Rozdělení náhodného výběru vizualizujeme pomocí histogramu superponovaného křivkou jádrového odhadu hustoty a pomocí krabicového diagramu (viz obrázek 11). Histogram vykreslíme příkazem `hist()`. V souladu se Sturgesovým pravidlem rozdělíme naměřená data do šesti ekvidistantních třídicích intervalů o šířce 2.5 mm. Hranice třídicích intervalů zvolíme 43.0, 45.5, . . . , 58.0. Jádrový odhad hustoty vypočítáme příkazem `density()`. Vektorem naměřených hodnot, na základě kterých jádrový odhad hustoty konstruujeme, zde bude vektor epikondylárních šířek pažní kosti na pravé straně `humer.EBRF`. Do grafu dále doplníme svislou referenční čáru (`abline()`) v hodnotě mediánu  $x_{0.50} = 54$ . Krabicový diagram zkonstruujeme příkazem `boxplot()`.

Z obou grafů na obrázku 11 je patrná silná asymetrie naměřených hodnot okolo mediánu  $x_{0.05} = 54$ . Z grafu histogramu vidíme, silné vyšikmení dat směrem doprava. Z krabicového diagramu je potom patrné posunutí mediánu směrem vpravo. Hodnota mediánu splývá s hodnotou horního kvartilu, zatímco nalevo od mediánu se tvoří dlouhý chvost odlehlých pozorování a několika extrémních měření. Na základě grafické vizualizace se tedy kloníme k závěrům Miraové testu a Cabiliova-Masarova testu, které hypotézu o symetrii náhodného výběru okolo mediánu zamítají. Polehčující okolností k závěru Miraové testu budiž fakt, že hypotéza o symetrii nebyla zamítnuta pouze hraničně ( $p$ -hodnota Miraové testu  $p = 0.05245$  je velmi blízká hladině významnosti  $\alpha = 0.05$ ).

```

284 par(mar = c(4, 4, 1, 1), family = 'Times')
285 hist(humer.EBRF, breaks = seq(43, 58, by = 2.5), prob = T, col = 'purple',
286       density = 40, main = '', ylab = 'relativní četnost', axes = F,
287       ylim = c(0, 0.22), xlab = '')
288 box(bty = 'o')
289 axis(1, seq(44.25, 56.75, by = 2.5))
290 axis(2, las = 1)
291 lines(density(humer.EBRF), col = 'purple4', lwd = 2)
292 abline(v = x0.50, col = 'red', lwd = 2, lty = 2)
293 mtext('epikondylární šířka pažní kosti z pravé strany(v mm)', side = 1, line = 2.2)
294
295 boxplot(humer.EBRF, col = 'thistle2', horizontal = T, border = 'purple4',
296         medcol = 'red', medlty = 2, medlwd = 2,
297         xlab = 'epikondylární šířka pažní kosti z pravé strany (v mm)')

```



Obrázek 11: Histogram a krabicový diagram epikondylárních šířek pažní kosti z pravé strany u skeletů ženského pohlaví

## 5. Interpretace výsledků

Datový soubor epikondylárních šířek pažní kosti z pravé strany u skeletů ženského pohlaví aljašské populace z kmene Tigara nepochází z rozdělení symetrického okolo mediánu.



### Příklad 6.11. Test symetrie

Mějme datový soubor 21-goldman-measures.csv obsahující údaje o délce kyčelní kosti z pravé strany (iblade.LR) a z levé strany (iblade.LL) u skeletů mužského a ženského pohlaví z japonské populace Yoshigo Shell Mound. Na hladině významnosti  $\alpha = 0.05$  testujte hypotézu, že náhodný výběr délek kyčelní kosti z levé strany u skeletů ženského pohlaví pochází ze symetrického rozdělení.

#### Řešení příkladu 6.11

Nejprve načteme datový soubor, a následně pomocí operátoru `[]` vybereme z tabulky údaje o délce kyčelní kosti z levé strany (iblade.LL) u skeletů ženského pohlaví (`sex == 'f'`) z japonské populace Yoshigo Shell Mound (`pop == 'Yoshigo Shell Mound'`). Následně odstraníme chybějící pozorování a zjistíme rozsah náhodného výběru, rozpětí naměřených hodnot a hodnotu mediánu.

```
298 data <- read.delim('00-Data//21-goldman-measures.csv', sep = ';', dec = '.')
299 iblade.LLF <- data[data$sex == 'f' & data$pop == 'Yoshigo Shell Mound', 'iblade.LL']
300 iblade.LLF <- na.omit(iblade.LLF)
301 n <- length(iblade.LLF) # 8
302 range(iblade.LLF) # 136-152
303 x0.50 <- median(iblade.LLF, type = 5) # 147.5
```

Datový soubor obsahuje 8 údajů o délce kyčelní kosti z levé strany u skeletů ženského pohlaví z populace Yoshigo Shell Mound. Naměřené hodnoty se pohybují v rozmezí 136-152 mm okolo mediánu  $x_{0.50} = 147.5$  mm. Hypotézu o symetrickém rozdělení náhodného výběru okolo mediánu otestujeme v pěti krocích pomocí Mioové testu, Cabiliova-Masarova testu a Miraové testu.

#### 1. Stanovení hypotéz

- **slovní formulace** nulové a alternativní hypotézy  
 $H_0$  : Data pochází z rozdělení symetrického okolo mediánu.  
 $H_1$  : Data nepochází z rozdělení symetrického okolo mediánu.
- **matematická formulace** nulové a alternativní hypotézy  
 $H_0$  :  $f(x_{0.50} - x) = f(x + x_{0.50})$   
 $H_1$  :  $f(x_{0.50} - x) \neq f(x + x_{0.50})$

#### 2. Volba hladiny významnosti

Hladina významnosti  $\alpha = 0.05$ .

#### 3. Testování $p$ -hodnotou

K provedení Mioové testu použijeme funkci `symmetry.test()` z knihovny `lawstat`. Vstupními parametry bude vektor naměřených hodnot `iblade.LLF`, argument `boot = F` (potlačení výpočtu  $p$ -hodnoty bootstrapovou metodou) a argument `option = 'MGG'` (volba Mioové testu). Z výstupu funkce vidíme, že hodnota testovací statistiky Mioové testu `Test statistic = -0.70284` a  $p$ -hodnota testu `p-value = 0.4822`.

```
304 lawstat::symmetry.test(iblade.LLF, boot = F, option = 'MGG')
```

```

Symmetry test by Miao, Gel, and Gastwirth (2006)
data:  iblade.LLF
Test statistic = -0.70284, p-value = 0.4822
alternative hypothesis: the distribution is asymmetric.
```

#### Závěr testování:

Protože  $p$ -hodnota = 0.4822 je větší než  $\alpha = 0.05$ ,  $H_0$  nezamítáme na hladině významnosti  $\alpha = 0.05$ .

Dále otestujeme hypotézu o symetrickém rozdělení pomocí Cabiliova-Masarova testu (funkce `symmetry.test()` s nastavením argumentu `option = 'CM'`). Z níže uvedeného výstupu vidíme, že hodnota testovací statistiky Cabiliova-Masarova testu `Test statistic = -0.64476` a výsledná  $p$ -hodnota testu `p-value = 0.5191`.

```
311 lawstat::symmetry.test(iblade.LLF, boot = F, option = 'CM')
```

```

Symmetry test by Cabilio and Masaro (1996)
data:  iblade.LLF
Test statistic = -0.64476, p-value = 0.5191
alternative hypothesis: the distribution is asymmetric.

```

312  
313  
314  
315  
316  
317

#### Závěr testování:

Protože  $p$ -hodnota = 0.5191 je větší než  $\alpha = 0.05$ ,  $H_0$  nezamítáme na hladině významnosti  $\alpha = 0.05$ .

Nakonec otestujeme nulovou hypotézu Miraové testu (funkce `symmetry.test()` s volbou argumentu `option = 'M'`). Hodnota testovací statistiky Miraové testu `Test statistic = -0.49728`,  $p$ -hodnota `p-value = 0.6190` a znění alternativní hypotézy `alternative hypothesis: the distribution is asymmetric`.

```
318 lawstat::symmetry.test(iblade.LLF, boot = F, option = 'M')
```

```

Symmetry test by Mira (1999)
data:  iblade.LLF
Test statistic = -0.49728, p-value = 0.619
alternative hypothesis: the distribution is asymmetric.

```

319  
320  
321  
322  
323  
324

#### Závěr testování:

Protože  $p$ -hodnota = 0.61905 je větší než  $\alpha = 0.05$ ,  $H_0$  nezamítáme na hladině významnosti  $\alpha = 0.05$ .

#### 4. Grafická vizualizace výsledků testování

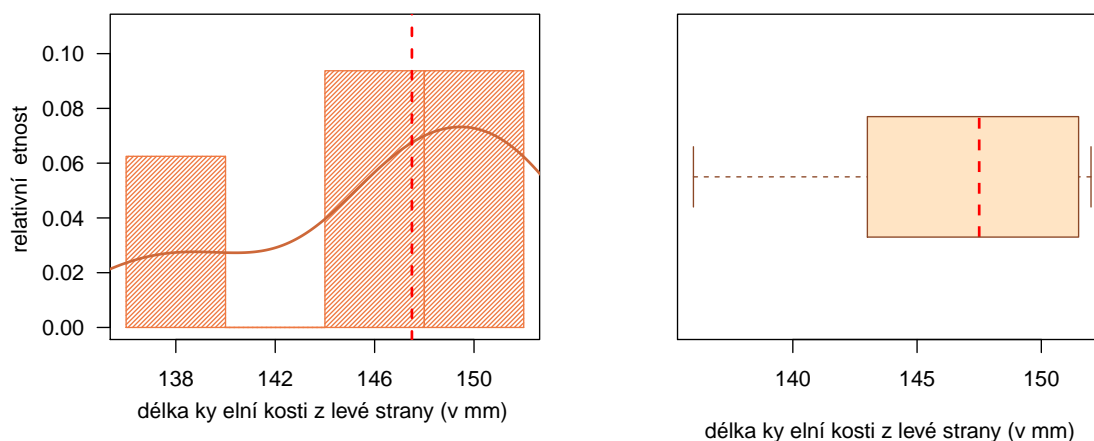
Stejně jako v předchozích příkladech vizualizujeme rozdělení náhodného výběru pomocí histogramu superponovaného křivkou jádrového odhadu hustoty a pomocí krabicového diagramu (viz obrázek 12). Pro účely histogramu rozdělíme naměřené délky kyčelních kostí z levé strany do čtyř ekvidistantních třídicích intervalů o šířce 4 mm prostřednictvím stanovených hranic 136, 140, ..., 152 mm. Jádrový odhad hustoty vypočítáme příkazem `density()` s vektorem délek kyčelních kostí z levé strany `iblade.LLF` jako vstupním argumentem. Nakonec do histogramu doplníme svislou referenční čáru v hodnotě mediánu  $x_{0.50} = 147.5$  (`abline()`). Krabicový diagram vykreslíme příkazem `boxplot()`.

```

325 par(mar = c(4, 4, 1, 1))
326 hist(iblade.LLF, breaks = seq(136, 152, by = 4), prob = T, col = 'sienna2',
327     density = 40, main = '', ylab = 'relativní četnost', axes = F,
328     ylim = c(0, 0.11), xlab = '')
329 box(bty = 'o')
330 axis(1, seq(138, 150, by = 4))
331 axis(2, las = 1)
332 lines(density(iblade.LLF), col = 'sienna3', lwd = 2)
333 abline(v = x0.50, col = 'red', lwd = 2, lty = 2)
334 mtext('délka kyčelní kosti z levé strany (v mm)', side = 1, line = 2.2)
335
336 boxplot(iblade.LLF, col = 'bisque', horizontal = T, border = 'sienna4',
337     medcol = 'red', medlty = 2, medlwd = 2,
338     xlab = 'délka kyčelní kosti z levé strany (v mm)')

```

Na základě grafické vizualizace uvedené na obrázku 12 bychom pravděpodobně symetrii náhodného výběru okolo mediánu neusuzovali. Zobrazený histogram i křivka jádrového odhadu jsou na první pohled vyšikmené



Obrázek 12: Histogram a krabicový diagram délek kyčelní kostí z levé strany u skeletů ženského pohlaví

směrem doprava s prodlouženým levým chvostem. Krabicový diagram taktéž ukazuje na prodloužený levý konec dat s odlehlými pozorováními vyskytující se nalevo od hodnoty mediánu  $x_{0,50} = 147.5$ .

Nabízí se tedy otázka, jak je možné, že závěry všech tří testů poměrně suverénně hovoří o symetrickém rozdělení náhodného výběru, když grafická vizualizace ukazuje spíše na nesymetrické rozdělení. Odpověď se skrývá v malém rozsahu náhodného výběru. Připomeňme si, že celkem máme k dispozici pouze osm měření délek kyčelních kostí s levé strany. Je důležité si uvědomit, že při takto malém počtu pozorování může i jedna hodnota přidaná k náhodnému výběru zcela změnit stávající charakter dat. Navíc je třeba mít stále na paměti základní myšlenku testování, která říká, že nezamítnutí nulové hypotézy o symetrii pouze svědčí o tom, že nemáme dostatek indicií, abychom hypotézu o symetrii zamítli. Můžeme to tedy brát tak, že libovolný z uvedených testů symetrie (stejně jako kterýkoli jiný statistický test, který si v této sbírce představíme) je v případě malého množství pozorování velmi benevolentní a nulovou hypotézu zamítá opravdu pouze při velmi hrubém porušení předpokládaného charakteru dat. Vzhledem k nízkému počtu pozorování není předpoklad symetrie porušen tak fatálně, aby došlo k zamítnutí hypotézy o symetrickém rozdělení náhodného výběru okolo hodnoty mediánu.

## 5. Interpretace výsledků

Datový soubor délek kyčelní kosti z levé strany u skeletů ženského pohlaví z japonské populace Yoshigo Shell Mound pochází z rozdělení symetrického okolo mediánu.

*Poznámka:* Situace v příkladu 12 se zdá být velmi podobná situaci z příkladu 11. Stěžejní rozdíl je však právě v počtu pozorování. Zatímco v příkladu 12 jsme měli k dispozici pouze osm pozorování, v příkladu 11 jsme jich měli k dispozici 25, což už je znatelně větší rozsah náhodného výběru. Pokud by rozsah náhodného výběru v příkladu 12 byl rovněž 25, můžeme při stávajícím vzhledu histogramu a krabicového diagramu předpokládat, že by testy hypotézu o symetrickém charakteru dat zamítly.

*Poznámka:* Příklad 12 pro nás nese poučení o tom, že zejména v případě malých rozsahů náhodných výběrů může být grafická vizualizace velmi zavádějící a závěry vytvořené na jejím základě mohou být velmi subjektivní. Z toho důvodu vždy považujeme grafickou vizualizaci za pouhý doplněk testování, který nám pomáhá vytvořit si konkrétnější představu o charakteru datového souboru, případně nám pomoci při rozhodování ve sporných situacích. Grafickou vizualizaci však nikdy nebereme za hlavní kritérium rozhodující o nulové hypotéze. Tím může být vždy a jedině statistický test.



### Příklad 6.12. Test symetrie

Mějme datový soubor 21-goldman-measures.csv obsahující údaje o délce vřetenní kosti z pravé strany (radius.LR) a z levé strany (radius.LL) mužů a žen pohřbených v oblasti Indian Knoll v Kentucky. Na hladině významnosti  $\alpha = 0.05$  testujte hypotézu, že náhodný výběr délek vřetenní kosti z pravé strany u mužů z oblasti Indian Knoll v Kentucky pochází ze symetrického rozdělení.

**Řešení příkladu 6.12** Nejprve načteme datový soubor, a následně vybereme z tabulky naměřené hodnoty délky vřetenní kosti z pravé strany (radius.LR) u mužů (sex == 'm') pohřbených v oblasti Indian Knoll v Kentucky (pop == 'Indian Knoll'). Následně odstraníme chybějící pozorování, zjistíme rozsah náhodného výběru, rozpětí naměřených hodnot a hodnotu mediánu.

```
339 data <- read.delim('00-Data//21-goldman-measures.csv', sep = ';', dec = '.')
340 radius.LRM <- data[data$sex == 'm' & data$pop == 'Indian Knoll', 'radius.LR']
341 radius.LRM <- na.omit(radius.LRM)
342 n <- length(radius.LRM) # 25
343 range(radius.LRM) # 231.0-270.5
344 x0.50 <- median(radius.LRM, type = 5) # 244
```

Datový soubor obsahuje údaje o délce vřetenní kosti z pravé strany u 25 mužů pohřbených v oblasti Indian Knoll v Kentucky. Naměřené hodnoty se pohybují okolo mediánu  $x_{0.50} = 244$  mm v celkovém rozmezí 231.0-270.5 mm. Hypotézu o symetrickém rozdělení náhodného výběru okolo mediánu otestujeme opět pomocí Mioové testu, Cabiliova-Masarova testu a Miraové testu.

#### 1. Stanovení hypotéz

- **slovní formulace** nulové a alternativní hypotézy  
 $H_0$  : Data pochází z rozdělení symetrického okolo mediánu.  
 $H_1$  : Data nepochází z rozdělení symetrického okolo mediánu.
- **matematická formulace** nulové a alternativní hypotézy  
 $H_0$  :  $f(x_{0.50} - x) = f(x + x_{0.50})$   
 $H_1$  :  $f(x_{0.50} - x) \neq f(x + x_{0.50})$

#### 2. Volba hladiny významnosti

Hladina významnosti  $\alpha = 0.05$ .

#### 3. Testování $p$ -hodnotou

K provedení Mioové testu použijeme funkci `symmetry.test()` z knihovny `lawstat`. Vstupními parametry bude vektor naměřených hodnot `radius.LRM`, argument `boot = F` a argument `option = 'MGG'`. Z výstupu funkce vidíme, že hodnota testovací statistiky Mioové testu `Test statistic = 1.8791` a  $p$ -hodnota testu `p-value = 0.06023`.

```
345 lawstat::symmetry.test(radius.LRM, boot = F, option = 'MGG')
```

```
Symmetry test by Miao, Gel, and Gastwirth (2006)
data: radius.LRM
Test statistic = 1.8791, p-value = 0.06023
alternative hypothesis: the distribution is asymmetric.
```

#### Závěr testování:

Protože  $p$ -hodnota = 0.06023 je větší než  $\alpha = 0.05$ ,  $H_0$  nezamítáme na hladině významnosti  $\alpha = 0.05$ .

Dále otestujeme hypotézu o symetrickém rozdělení pomocí Cabiliova-Masarova testu (funkce `symmetry.test()` s nastavením argumentu `option = 'CM'`). Hodnota testovací statistiky Cabiliova-Masarova testu `Test statistic = 2.0304` a výsledná  $p$ -hodnota testu `p-value = 0.04231`.

```
352 lawstat::symmetry.test(radius.LRM, boot = F, option = 'CM')
```

```
Symmetry test by Cabilio and Masaro (1996)
data: radius.LRM
Test statistic = 2.0304, p-value = 0.04231
alternative hypothesis: the distribution is asymmetric.
```

353  
354  
355  
356  
357  
358

#### Závěr testování:

Protože  $p$ -hodnota = 0.04231 je menší než  $\alpha = 0.05$ ,  $H_0$  zamítáme na hladině významnosti  $\alpha = 0.05$ .

Nakonec otestujeme nulovou hypotézu Miraové testu (funkce `symmetry.test()` s volbou argumentu `option = 'M'`). Hodnota testovací statistiky Miraové testu `Test statistic = 1.5274` s  $p$ -hodnotou `p-value = 0.1267`.

```
359 lawstat::symmetry.test(radius.LRM, boot = F, option = 'M')
```

```
Symmetry test by Mira (1999)
data: radius.LRM
Test statistic = 1.5274, p-value = 0.1267
alternative hypothesis: the distribution is asymmetric.
```

360  
361  
362  
363  
364  
365

#### Závěr testování:

Protože  $p$ -hodnota = 0.1267 je větší než  $\alpha = 0.05$ ,  $H_0$  nezamítáme na hladině významnosti  $\alpha = 0.05$ .

Porovnáme-li výsledek Miaové testu, Cabiliova-Masarova testu a Miraové testu, vidíme, že opět dochází k rozporným závěrům týkajících se nulové hypotézy o symetrii náhodného výběru. Zatímco Miaové test a Miraové test hypotézu o symetrii nezamítá, Cabiliova-Masarův test ji zamítá. Zamítnutí hypotézy je však opět hraniční, k čemuž můžeme při finálním rozhodnutí o nulové hypotéze přihlídnout. Než učiníme závěr o symetrii náhodného výběru, provedeme grafickou vizualizaci naměřených hodnot.

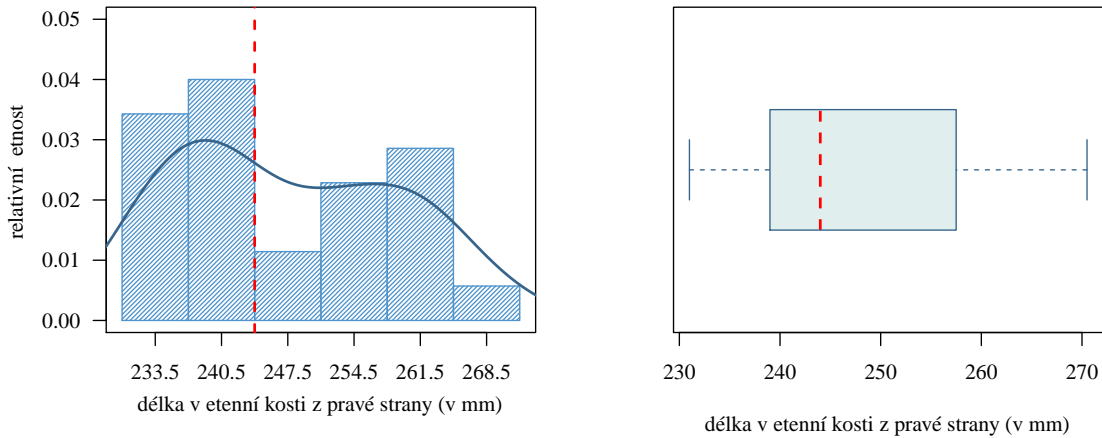
#### 4. Grafická vizualizace výsledků testování

Nejprve vykreslíme histogram superponovaný křivkou jádrového odhadu hustoty a následně krabicový diagram (viz obrázek 13). Pro účely histogramu rozdělíme naměřené hodnoty délky vřetenní kosti z pravé strany do šesti ekvidistantních třídících intervalů s hranicemi 230, 237, ..., 272 mm. Šířka každého intervalu je 7 mm. Do histogramu doplníme jádrový odhad hustoty a svislou referenční čáru v hodnotě mediánu  $x_{0.50} = 244$ . Nakonec vykreslíme zmiňovaný krabicový diagram.

```
366 par(mar = c(4, 4, 1, 1), family = 'Times')
367 hist(radius.LRM, breaks = seq(230, 272, by = 7), prob = T, col = 'steelblue3',
368     density = 40, main = '', ylab = 'relativní četnost', axes = F, xlab = '',
369     ylim = c(0, 0.05))
370 box(bty = 'o')
371 axis(1, seq(233.5, 268.5, by = 7))
372 axis(2, las = 1)
373 lines(density(radius.LRM), col = 'steelblue4', lwd = 2)
374 abline(v = median(radius.LRM, type = 5), col = 'red', lwd = 2, lty = 2)
375 mtext('délka vřetenní kosti z pravé strany (v mm)', side = 1, line = 2.2)
376
377 boxplot(radius.LRM, col = 'azure2', horizontal = T, border = 'steelblue4',
378     medcol = 'red', medlty = 2, medlwd = 2,
379     xlab = 'délka vřetenní kosti z pravé strany (v mm)')
```

Grafická vizualizace uvedená na obrázku 13 taktéž nehovoří příliš jednoznačně. Z grafů rozhodně není patrná zřejmá symetrie, ale na typický případ asymetrických dat okolo mediánu to také nevypadá. Jde tedy skutečně o hraniční případ a jakékoli rozhodnutí na základě grafické vizualizace by mohlo být naprosto subjektivní.





Obrázek 13: Histogram a krabicový diagram délek vřetenní kostí z pravé strany u skeletů mužského pohlaví

I s takto spornými případy se však v praxi můžeme běžně setkat a je dobré být na ně připravený. V tomto případě podpoříme závěr Miaoové testu a Miraové testu a hypotézu o symetrii nezamítáme. Důvodem našeho rozhodnutí je fakt, že dva ze tří provedených testů hypotézu o symetrii náhodného výběru nezamítly a třetí test hypotézu zamítl hraničně. Grafická vizualizace potom nevykazuje závažné porušení symetrického charakteru naměřených dat.

##### 5. Interpretace výsledků

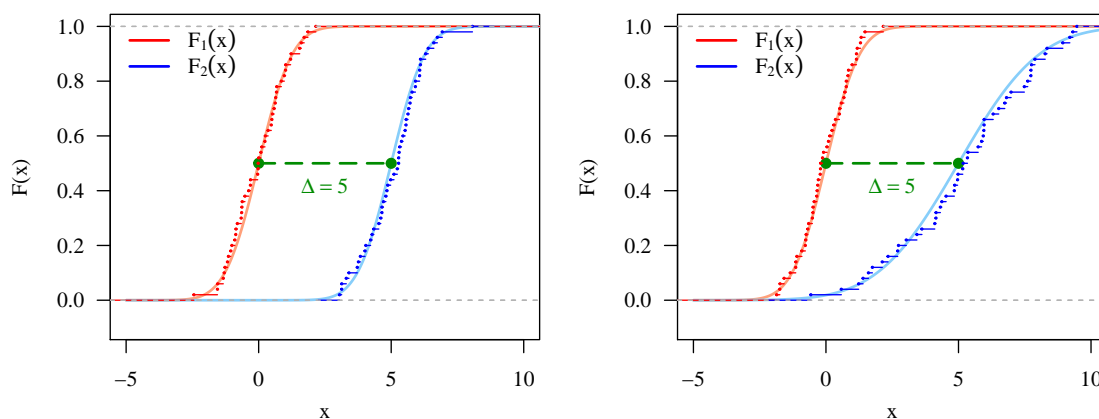
Datový soubor délek vřetenních kostí z pravé strany u mužů pohřbených v oblasti Indian Knoll pochází z rozdělení symetrického okolo mediánu.




## 6.4 Test o shodě distribučních funkcí dvou náhodných výběrů

Předpokládejme nyní, že máme dva náhodné výběry a chceme otestovat hypotézu a shodě středních hodnot těchto dvou náhodných výběrů. Pochází-li oba náhodné výběry z normálního rozdělení, použijeme na test hypotézy o shodě středních hodnot těchto výběrů parametrický dvouvýběrový test, a to konkrétně buď klasický dvouvýběrový  $t$ -test (viz sekce ??) nebo dvouvýběrový  $t$ -test s Welchovou aproximací stupňů volnosti (viz sekce ??). Naopak, nepochází-li alespoň jeden z náhodných výběrů z normálního rozdělení, nemůžeme dvouvýběrové parametrické testy z kapitoly ?? použít, a musíme použít některý z neparametrických testů uvedených v kapitole ?. Konkrétně volíme buď Wilcoxonův dvouvýběrový test (viz sekce ?? pro exaktní variantu testu a sekce ?? pro asymptotickou variantu testu), nebo Mannův-Whitneyův dvouvýběrový test (viz sekce ??). Předpokladem obou uvedených neparametrických testů je pouhá shoda distribučních funkcí obou náhodných výběrů, přičemž je povoleno, aby se distribuční funkce navzájem lišily pouze posunutím. Naštěstí pro nás bývá tento předpoklad ve valné většině případů splněn. Málokdy v reálných situacích dochází k tomu, že se tvary distribučních funkcí porovnávaných náhodných výběrů statisticky významně liší. V případě, že by však předpoklad shody distribučních funkcí dvou porovnávaných náhodných výběrů přeci jen nebyl splněn, museli bychom hledat metodu vhodnou na porovnání takovýchto výběrů v řadách pokročilých statistických metod.

Nechť  $X_{11}, \dots, X_{1n_1}$ ,  $n_1 \geq 2$  je náhodný výběr ze spojitého rozdělení s distribuční funkcí  $F_1(x)$  a  $X_{21}, \dots, X_{2n_2}$ ,  $n_2 \geq 2$  je na něm nezávislý náhodný výběr ze spojitého rozdělení s distribuční funkcí  $F_2(x)$ . Na hladině významnosti  $\alpha = 0.05$  testujeme hypotézu  $H_0$ : *Distribuční funkce  $F_1(x)$  a  $F_2(x)$  se liší pouze posunutím o vzdálenost  $\Delta$* . oproti  $H_1$ : *Distribuční funkce  $F_1(x)$  a  $F_2(x)$  se neliší pouze posunutím o vzdálenost  $\Delta$* . Matematické ekvivalenty zápisu nulové a alternativní hypotézy definujeme právě pomocí distribučních funkcí, tj.  $H_0 : F_1(x) = F_2(x + \Delta), \forall x \in \mathbb{R}$ , kde  $\Delta \in \mathbb{R}$ , oproti  $H_1 : F_1(x) \neq F_2(x + \Delta)$ , pro nějaké  $x \in \mathbb{R}$ , přičemž  $\Delta \in \mathbb{R}$ . Graficky je ukázka dvou distribučních funkcí lišících se pouze posunutím o vzdálenost  $\Delta$  a dvou distribučních funkcí lišících se i jinak než pouhým posunutím o vzdálenost  $\Delta$  vyobrazena na obrázku 14.



Obrázek 14: Ukázka dvojice distribučních funkcí  $F_1(x)$  a  $F_2(x)$  (a) lišících se pouze posunutím o vzdálenost  $\Delta = 5$ ; (b) lišících se více než pouhým posunutím o vzdálenost  $\Delta = 5$

Proces testování nulové hypotézy o shodě dvou distribučních funkcí s výjimkou posunutí o vzdálenost  $\Delta$  bude sestávat ze dvou kroků. V prvním kroku odstraníme vliv posunutí mezi oběma náhodnými výběry, a tedy také mezi jim příslušnými distribučními funkcemi, metodou centrování (viz sekce ???). Tuto metodu aplikujeme zvlášť na každý náhodný výběr. Připomeňme, že centrování náhodného výběru docílíme tak, že z hodnot náhodného výběru vypočítáme aritmetický průměr, který následně odečteme od každé naměřené hodnoty náhodného výběru. V druhém kroku potom distribuční funkce obou centrovaných náhodných výběrů ( $F_{1c}(x)$  a  $F_{2c}(x)$ ) podrobíme Kolmogorovu-Smirnovovu testu (K-S testu) o shodě dvou distribučních funkcí. Na hladině významnosti  $\alpha = 0.05$  testujeme nulovou hypotézu  $H_0$ : *Distribuční funkce obou centrovaných náhodných výběrů jsou shodné.*, tj.  $H_0 : F_{1c}(x) = F_{2c}(x), \forall x \in \mathbb{R}$  oproti  $H_1$ : *Distribuční funkce obou centrovaných náhodných výběrů nejsou shodné.*, tj.  $H_1 : F_{1c}(x) \neq F_{2c}(x)$  pro nějaké  $x \in \mathbb{R}$ . Algoritmus Kolmogorova-Smirnovova testu zde detailněji rozebírat nebudeme. Test provedeme pomocí funkce `ks.test()` implementované v základní  knihovně `stats`. Funkce `ks.test()` má dva povinné vstupní argumenty. Prvním povinným argumentem je vektor naměřených hodnot prvního náhodného výběru `x`, druhým povinným argumentem je vektor naměřených hodnot druhého náhodného výběru `y`. Nepovinným argumentem funkce je potom

argument `alternative` umožňující volbu alternativní hypotézy (oboustranná alternativa (`alternative = 'two.sided'`; defaultní nastavení) odpovídající alternativní hypotéze  $H_{11} : F_{1c}(x) \neq F_{2c}(x)$ , pravostranná alternativa (`alternative = 'greater'`) odpovídající alternativní hypotéze  $H_{12} : F_{1c}(x) \geq F_{2c}(x)$  nebo levostranná alternativa (`alternative = 'less'`) odpovídající alternativní hypotéze  $H_{13} : F_{1c}(x) \leq F_{2c}(x)$ ). Posledním nepovinným argumentem je argument `exact`, kterým určujeme, zda v případě jednostranných alternativních hypotéz bude  $p$ -hodnota Kolmogorova-Smirnovova testu vypočítána exaktně (`exact = T`), nebo asymptoticky (`exact = F`). Vzhledem k tomu, že pro naše účely budeme využívat pouze oboustrannou alternativní hypotézu, není třeba ve funkci `ks.test()` argumenty `alternative` a `exact` jakkoli specifikovat.

Graficky shodu distribučních funkcí dvou náhodných výběrů s výjimkou posunutí o vzdálenost  $\Delta$  vizualizujeme pomocí dvojice histogramů zakreslených do jednoho grafu, přičemž každý histogram bude superponován vlastní křivkou jádrového odhadu funkce hustoty. Z kapitoly ??? víme že, distribuční funkce  $F(x)$  a hustota  $f(x)$  jsou navzájem velmi úzce spjaté. Pokud tedy histogramy, potažmo křivky jádrových odhadů hustot mají podobný tvar (s výjimkou posunutí), budou i distribuční funkce obou náhodných výběrů podobné (opět s výjimkou posunutí). Naopak výrazná odlišnost v histogramech, potažmo v křivkách jádrových odhadů hustot ukazuje též na odlišnost distribučních funkcí obou náhodných výběrů. Druhým grafem vizualizujícím míru shody distribučních funkcí dvou náhodných výběrů s výjimkou posunutí bude potom přímo graf zobrazující dvojici výběrových distribučních funkcí. Z tohoto grafu bude potom na první pohled patrná podobnost či odlišnost distribučních funkcí obou náhodných výběrů.

### Příklad 6.13. Test o shodě distribučních funkcí dvou náhodných výběrů

Mějme datový soubor 21-goldman-measures.csv obsahující údaje o délce stehenní kosti z pravé strany (femur.LR) a z levé strany (femur.LL) u skeletů mužského a ženského pohlaví z pohřebiště Pueblo Bonito a Puye z oblasti Nového Mexika. Na hladině významnosti  $\alpha = 0.05$  testujte hypotézu, o shodě distribučních funkcí délky stehenní kosti z levé strany u skeletů ženského pohlaví z pohřebiště Pueblo Bonito a Puye.

#### Řešení příkladu 6.13

Nejprve pomocí příkazu read.delim() načteme datový soubor 21-goldman-measures.csv. Z načtených dat vybereme naměřené hodnoty délky stehenní kosti z levé strany femur.LL skeletů ženského pohlaví (sex == 'f') z pohřebiště Pueblo Bonito (pop == 'Pueblo Bonito, Chaco Canyon') a vložíme je do proměnné femur.LFPB. Analogicky vytvoříme vektor femur.LFP obsahující naměřené hodnoty délky stehenní kosti femur.LL skeletů ženského pohlaví (sex == 'f') z pohřebiště Puye (pop == 'Puye'). Z obou vektorů odstraníme chybějící pozorování příkazem na.omit() a pomocí příkazů length(), range() a mean() zjistíme rozsahy obou náhodných výběrů a rozpětí a aritmetický průměr naměřených hodnot.

```
380 data <- read.delim('00-Data//21-goldman-measures.csv', sep = ';', dec = '.')
381 femur.LFPB <- na.omit(data[data$sex == 'f' & data$pop == 'Pueblo Bonito, Chaco Canyon',
382                          'femur.LL'])
383 femur.LFP <- na.omit(data[data$sex == 'f' & data$pop == 'Puye', 'femur.LL'])
384 n1 <- length(femur.LFPB) # 7
385 n2 <- length(femur.LFP) # 12
386 range(femur.LFPB) # 408-439
387 range(femur.LFP) # 368-410
388 m1 <- mean(femur.LFPB) # 427.3571
389 m2 <- mean(femur.LFP) # 391.25
```

Datový soubor obsahuje naměřené hodnoty délky stehenní kosti z levé strany u 7 skeletů ženského pohlaví z pohřebiště Pueblo Bonito a u 12 skeletů ženského pohlaví z pohřebiště Puye. Naměřené délky stehenních kostí u skeletů z oblasti Pueblo Bonito se realizují v okolí aritmetického průměru 427.3571 mm v rozmezí 408-439 mm. Naměřené délky stehenních kostí u skeletů z oblasti Puye se realizují v okolí aritmetického průměru 391.25 mm v rozmezí 368-410 mm.

Na základě výše uvedených údajů můžeme konstatovat, že máme k dispozici náhodný výběr  $X_{11}, X_{12}, \dots, X_{1n_1}$ ,  $n_1 = 7$ , popisující délku stehenní kosti z levé strany u skeletů ženského pohlaví z pohřebiště Pueblo Bonito. Náhodná veličina  $X_{11}$  popisuje délku stehenní kosti z levé strany u prvního skeletu ženského pohlaví,  $\dots$ , náhodná veličina  $X_{17}$  popisuje délku stehenní kosti z levé strany u sedmého skeletu ženského pohlaví. Předpokládejme nyní, že tento náhodný výběr pochází z nějakého nám neznámého rozdělení  $L(\theta_1)$  s distribuční funkcí  $F_1(x)$ . Dále máme k dispozici náhodný výběr  $X_{21}, X_{22}, \dots, X_{2n_2}$ ,  $n_2 = 12$ , popisující délku stehenní kosti z levé strany u skeletů ženského pohlaví z pohřebiště Puye. Náhodná veličina  $X_{21}$  popisuje délku stehenní kosti z levé strany u prvního skeletu ženského pohlaví,  $\dots$ , náhodná veličina  $X_{112}$  popisuje délku stehenní kosti z levé strany u dvanáctého skeletu ženského pohlaví. Předpokládejme, že tento náhodný výběr pochází z nějakého nám neznámého rozdělení  $L(\theta_2)$  s distribuční funkcí  $F_2(x)$ . Na hladině významnosti  $\alpha = 0.05$  chceme v souladu se zadáním testovat nulovou hypotézu o shodě distribučních funkcí  $F_1(x)$  a  $F_2(x)$ , přičemž je dovoleno, aby se distribuční funkce lišily pouze posunutím o vzdálenost  $\Delta$ . Testování provedeme v posloupnosti pěti kroků pomocí Kolmogorova-Smirnovova testu implementovaného v  $\mathbb{R}$  funkci ks.test().

#### 1. Stanovení hypotéz

- **slovní formulace** nulové a alternativní hypotézy

$H_0$  : Distribuční funkce délky stehenní kosti z levé strany u skeletů ženského pohlaví z pohřebiště Pueblo Bonito a Puye se liší pouze posunutím o vzdálenost  $\Delta$ .

$H_1$  : Distribuční funkce délky stehenní kosti z levé strany u skeletů ženského pohlaví z pohřebiště Pueblo Bonito a Puye se neliší pouze posunutím o vzdálenost  $\Delta$ .

- **matematická formulace** nulové a alternativní hypotézy

$H_0 : F_1(x) = F_2(x + \Delta), \forall x \in \mathbb{R}, \text{ kde } \Delta \in \mathbb{R}.$

$$H_1 : F_1(x) \neq F_2(x + \Delta), \text{ pro nějaké } x \in \mathbb{R}, \text{ přičemž } \Delta \in \mathbb{R}.$$

## 2. Volba hladiny významnosti

Hladina významnosti  $\alpha = 0.05$ .


## 3. Testování $p$ -hodnotou

Z úvodního rozboru příkladu víme, že náhodný výběr délek stehenních kostí skeletů z pohřebiště Pueblo Bonito se pohybuje okolo hodnoty 427.3571 mm, zatímco náhodný výběr skeletů z pohřebiště Puye se pohybuje okolo hodnoty 391.25 mm. Před otestováním nulové hypotézy musíme nejprve zajistit, aby rozdílné polohy obou náhodných výběrů negativně neovlivnily výsledek testu o shodě obou distribučních funkcí (s výjimkou posunutí). Vliv rozdílné polohy eliminujeme metodou centrování. Náhodný výběr délek stehenních kostí u skeletů z oblasti Pueblo Bonito  $X_{11}, \dots, X_{1n_1}$  centrujeme tak, že od každé hodnoty tohoto náhodného výběru odečteme hodnotu aritmetického průměru 427.3571 mm. Získáme tak náhodný výběr centrovaných náhodných veličin  $X_{11,c}, \dots, X_{1n_1,c}$ .

$$\begin{aligned} X_{11,c} &= X_{11} - m_1 = 408.0 - 427.3571 = -19.357143 \\ X_{12,c} &= X_{12} - m_1 = 439.0 - 427.3571 = 11.642857 \\ X_{13,c} &= X_{13} - m_1 = 421.0 - 427.3571 = -6.357143 \\ X_{14,c} &= X_{14} - m_1 = 436.5 - 427.3571 = 9.142857 \\ X_{15,c} &= X_{15} - m_1 = 429.0 - 427.3571 = 1.642857 \\ X_{16,c} &= X_{16} - m_1 = 435.0 - 427.3571 = 7.642857 \\ X_{17,c} &= X_{17} - m_1 = 423.0 - 427.3571 = -4.357143 \end{aligned}$$

Analogicky centrujeme náhodný výběr délek stehenních kostí u skeletů z oblasti Puye  $X_{21}, \dots, X_{2n_2}$  tak, že od každé hodnoty tohoto náhodného výběru odečteme hodnotu aritmetického průměru 391.25 mm. Získáme tak náhodný výběr centrovaných náhodných veličin  $X_{21,c}, \dots, X_{2n_2,c}$ .

$$\begin{aligned} X_{21,c} &= X_{21} - m_2 = 378.0 - 391.25 = -13.25 \\ X_{22,c} &= X_{22} - m_2 = 398.0 - 391.25 = 6.75 \\ &\vdots \\ X_{212,c} &= X_{212} - m_2 = 394.5 - 391.25 = 3.25 \end{aligned}$$

Procesem centrování jsme docílili toho, že se hodnoty obou centrovaných náhodných výběrů  $X_{11,c}, \dots, X_{1n_1,c}$  a  $X_{21,c}, \dots, X_{2n_2,c}$  realizují okolo nuly. Konkrétně aritmetický průměr centrovaných délek stehenních kostí z pohřebiště Pueblo Bonito je roven hodnotě  $2.436147e-14$  mm zatímco aritmetický průměr centrovaných délek stehenních kostí z pohřebiště Puye je roven přesně 0 mm. Nyní si ukažme, jak lze vytvořit vektory centrovaných délek stehenních kostí z pohřebiště Pueblo Bonito (femur.LFPB.c) a z pohřebiště Puye (femur.LFP.c) pomocí softwaru .

```
390 femur.LFPB.c <- femur.LFPB - m1
391 femur.LFP.c <- femur.LFP - m2
392 mean(femur.LFPB.c) # 2.436147e-14
393 mean(femur.LFP.c) # 0
```

V tomto okamžiku je vhodné si uvědomit, že nahrazením původních náhodných výběrů jejich centrovanými náhodnými výběry a následným aplikováním Kolmogorova-Smirnova testu právě na tyto centrované náhodné výběry dochází k modifikaci původní nulové hypotézy (a tedy také původní alternativní hypotézy) na následující znění:

- **slovní formulace** modifikované nulové a alternativní hypotézy

$H_0$  : Distribuční funkce centrované délky stehenní kosti z levé strany u skeletů ženského pohlaví z pohřebiště Pueblo Bonito a Puye jsou shodné.

$H_1$  : Distribuční funkce centrované délky stehenní kosti z levé strany u skeletů ženského pohlaví z pohřebiště Pueblo Bonito a Puye nejsou shodné.

- **matematická formulace** modifikované nulové a alternativní hypotézy

$H_0 : F_{1c}(x) = F_{2c}(x), \forall x \in \mathbb{R}, \text{ kde } \Delta \in \mathbb{R}.$

$H_1 : F_{1c}(x) \neq F_{2c}(x), \text{ pro nějaké } x \in \mathbb{R}, \text{ přičemž } \Delta \in \mathbb{R}.$

Nyní otestujeme modifikovanou nulovou hypotézu o shodě dvou distribučních funkcí centrovaných náhodných výběrů pomocí Kolmogorova-Smirnovova testu. Tím, že místo na původní náhodné výběry aplikujeme Kolmogorův-Smirnovův test právě na centrované náhodné výběry, kde je již vliv posunutí odfiltrován, snadno zjistíme, zda se distribuční funkce původních náhodných výběrů liší jinak než posunutím. Kolmogorův-Smirnovův test provedeme pomocí funkce `ks.test()`. Vstupními argumenty funkce budou pouze vektory centrovaných náhodných výběrů `femur.LFPB.c` a `femur.LFP.c`. Jak jsme uvedli výše, další argumenty funkce `ks.test`, jako jsou `alternative` a `exact` není nutné specifikovat. Výstupem funkce `ks.test()` je přesný název použitého testu (Two-sample Kolmogorov-Smirnov test), přesné názvy proměnných použitých pro testování (`femur.LFPB.c` a `femur.LFP.c`), hodnota testovací statistiky K-S testu ( $D = 0.25$ ),  $p$ -hodnota testu ( $p\text{-value} = 0.8869$ ) a zvolená alternativní hypotéza (`alternative hypothesis: two-sided`; defaultní nastavení použité v případě, že argument `alternative` není specifikován).

```
394 ks.test(femur.LFPB.c, femur.LFP.c)
```

```

Two-sample Kolmogorov-Smirnov test
data: femur.LFPB.c and femur.LFP.c
D = 0.25, p-value = 0.8869
alternative hypothesis: two-sided

```

395  
396  
397  
398  
399  
400

#### Závěr testování:

Protože  $p$ -hodnota = 0.8869 je větší než  $\alpha = 0.05$ ,  $H_0$  nezamítáme na hladině významnosti  $\alpha = 0.05$ .

#### 4. Grafická vizualizace výsledků testování

Tvary hustot obou náhodných výběrů (ze kterých potom vychází také tvar distribučních funkcí) zobrazíme pomocí dvojice histogramů vykreslené do jednoho obrázku. Každý histogram bude superponován křivkou jádrového odhadu hustoty. Pro tento účel roztřídíme náhodný výběr délek stehenních kostí z pohřebiště Pueblo Bonito do čtyř ekvidistantních třídicích intervalů o šířce 8 mm prostřednictvím hranic 408, 416, ..., 440 mm a náhodný výběr délek stehenních kostí z pohřebiště Puye do pěti ekvidistantních třídicích intervalů o šířce 9 mm prostřednictvím hranic 366, 375, ..., 410 mm. Protože histogramy nemají stejné středy třídicích intervalů, porušíme v tomto případě konvenční pravidlo stanovující, že měřítko osy  $x$  zobrazuje hodnoty středů třídicích intervalů. Aby byl graf přehledný zvolíme na výplň sloupců každého histogramu barvu s určitou mírou průhlednosti. K tomu využijeme funkci `rgb()` z knihovny `grDevices`. V syntaxi funkce `rgb()` budeme specifikovat hodnoty čtyř vstupních parametrů. Specifikováním argumentu `red` určujeme množství červené barvy, která se podílí na vzniku barvy finální. Hodnotu argumentu `red` volíme z intervalu  $\langle 0; 1 \rangle$ , přičemž čím je tato hodnota bližší nule, tím menší je podíl červené barvy na finální barvě. Analogicky definujeme podíl zelené barvy (argument `green`) a podíl modré barvy (argument `blue`) na finální barvě. Konečně, specifikací argumentu `alpha` definujeme míru průhlednosti výsledné barvy. Hodnotu argumentu `alpha` volíme opět z intervalu  $\langle 0; 1 \rangle$ , přičemž čím je hodnota argumentu `alpha` bližší nule, tím více bude finální barva průhledná. V kombinaci s argumentem `density = 80` v příkazu `hist()` doporučujeme pro dosažení optimální míry průhlednosti histogramů nastavit hodnotu argumentu `alpha = 0.2`.

Druhým vykresleným grafem bude graf výběrových distribučních funkcí  $F_1(x)$  a  $F_2(x)$ . Hodnoty výběrové distribuční funkce  $F_1(x)$  pro délky stehenních kostí z pohřebiště Pueblo Bonito vypočítáme pomocí funkce `ecdf()`

```

401 par(mar = c(4, 4, 1, 1), family = 'Times')
402 hist(femur.LFPB, breaks = seq(408, 440, by = 8), prob = T,
403       col = grDevices::rgb(1, 0, 0, alpha = 0.2), density = 80,
404       main = '', ylab = 'relativní četnost', axes = F, xlab = '',
405       ylim = c(0, 0.06), xlim = c(365, 442))
406 box(bty = 'o')
407 axis(1, seq(360, 450, by = 10))
408 axis(2, las = 1)
409 lines(density(femur.LFPB), col = 'red', lwd = 2)
410 mtext('délka stehenní kosti z levé strany (v mm)', side = 1, line = 2.2)
411
412 hist(femur.LFP, breaks = seq(366, 411, by = 9), prob = T, density = 80,
413       col = grDevices::rgb(0, 0, 1, alpha = 0.2), add = T)
414 lines(density(femur.LFP), col = 'blue', lwd = 2)
415
416 legend('topleft', fill = c(rgb(1, 0, 0, 0.4), rgb(0, 0, 1, 0.4)),
417       legend = c('Pueblo Bonito', 'Puye'), bty = 'n')

```

aplikované na vektor naměřených hodnot `femur.LFPB`. Příkazem `plot(ecdf(femur.LFPB))` rovnou vykreslíme graf zobrazující křivku výběrové distribuční funkce  $F_1(x)$ . Finální vzhled grafu doladíme specifikací známých argumentů, jako jsou například argumenty `col`, `pch`, `cex`, `lwd`, `lty`, `xlim`, `xlab`, `ylab`, `las`, apod. Poznamenejme, že graf rovnou připravujeme také pro vykreslení křivky výběrové distribuční funkce  $F_2(x)$  pro data z pohřebiště Puye. Proto je nutné v příkazu `plot()` zadat rozsah osy  $x$  natolik široký, aby se do něj křivka výběrové distribuční funkce  $F_2(x)$  vešla (například nastavení argumentu `xlim = c(365, 442)`). Hodnoty výběrové distribuční funkce  $F_2(x)$  pro délky stehenních kostí z pohřebiště Puye opět vypočítáme pomocí funkce `ecdf()` aplikované tentokrát na vektor naměřených hodnot `femur.LFP`. Příkazem `lines(ecdf(femur.LFP))` doplníme křivku výběrové distribuční funkce  $F_2(x)$  do stávajícího grafu. Finální vzhled křivky doladíme například specifikací argumentů `col`, `pch`, `lwd`, `lty`, apod.

```

418 plot(ecdf(femur.LFPB), col = 'red', main = '', ylab = 'F(x)',
419       xlab = '', las = 1, lwd = 1, cex = 0.4, xlim = c(365, 442))
420 lines(ecdf(femur.LFP), col = 'blue', lwd = 1, cex = 0.4)
421 mtext('délka stehenní kosti z levé strany', side = 1, line = 2.2)
422
423 legend('topleft', lwd = 2, col = c('red', 'blue'),
424       legend = c('Pueblo Bonito', 'Puye'), bty = 'n' )

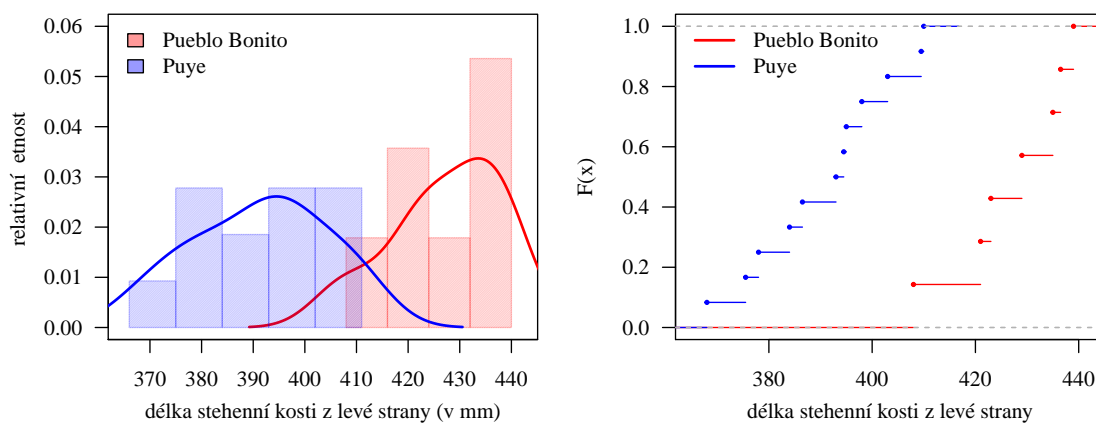
```

Grafická vizualizace uvedená na obrázku 15 podporuje výsledek testování. Histogramy i jádrové odhady hustoty obou výběrů jsou si tvarem velmi podobné. Taktéž křivky výběrových distribučních funkcí mají podobný tvar.

## 5. Interpretace výsledků

Distribuční funkce délky stehenní kosti z levé strany u skeletů ženského pohlaví z pohřebiště Pueblo Bonito a Puye se liší pouze posunutím.





Obrázek 15: Histogramy (vlevo) a distribuční funkce (vpravo) náhodných výběrů délek stehenních kostí z levé strany u skeletů ženského pohlaví (a) z pohřebiště Pueblo Bonito (červená barva); (b) z pohřebiště Puye (modrá barva)



### Příklad 6.14. Test o shodě distribučních funkcí dvou náhodných výběrů

Mějme datový soubor 05-one-sample-correlation-skull-mf.txt obsahující údaje o největší výšce mozkovny (skull.pH) a morfologické výšce tváře (face.H) u jedinců starověké egyptské populace (viz dekce ??). Na hladině významnosti  $\alpha = 0.05$  testujte hypotézu, o shodě distribučních funkcí morfologické výšky tváře u mužů a žen starověké egyptské populace.

#### Řešení příkladu 6.14

Nejprve načteme datový soubor 05-one-sample-correlation-skull-mf.txt. Z načtených dat vybereme naměřené hodnoty morfologické výšky tváře face.H mužů `sex == 'm'` a vložíme je do proměnné (face.HM). Analogicky vytvoříme vektor face.HF obsahující naměřené hodnoty morfologické výšky tváře face.H žen (`sex == 'f'`). Z obou vektorů odstraníme chybějící pozorování (`na.omit()`) a zjistíme rozsahy (`length()`), rozpětí naměřených hodnot (`range()`) a aritmetické průměry (`mean()`) obou náhodných výběrů.

```
425 data <- read.delim('00-Data//05-one-sample-correlation-skull-mf.txt')
426 face.HM <- na.omit(data[data$sex == 'm', 'face.H'])
427 face.HF <- na.omit(data[data$sex == 'f', 'face.H'])
428 n.M <- length(face.HM) # 166
429 n.F <- length(face.HF) # 80
430 range(face.HM) # 85-136
431 range(face.HF) # 95-119
432 m1 <- mean(face.HM) # 114.5361
433 m2 <- mean(face.HF) # 107.5625
```

Datový soubor obsahuje naměřené hodnoty morfologické výšky tváře u 166 mužů a 80 žen starověké egyptské populace. Naměřené morfologické výšky tváře mužů se pohybují v rozmezí 85–136 mm okolo hodnoty 114.5361 mm. Naměřené morfologické výšky tváře žen se pohybují v rozmezí 95–119 mm okolo hodnoty 107.5625 mm.

K dispozici máme tentokrát náhodný výběr  $X_{11}, X_{12}, \dots, X_{1n_1}$ ,  $n_1 = 166$ , popisující morfologickou výšku tváře u mužů. Náhodná veličina  $X_{11}$  popisuje morfologickou výšku tváře prvního muže,  $\dots$ , náhodná veličina  $X_{1166}$  popisuje morfologickou výšku tváře u 166. muže. Předpokládejme, že náhodný výběr morfologické výšky tváře mužů pochází z nějakého nám neznámého rozdělení  $L(\theta_1)$  s distribuční funkcí  $F_1(x)$ . Dále máme k dispozici náhodný výběr  $X_{21}, X_{22}, \dots, X_{2n_2}$ ,  $n_2 = 80$ , popisující morfologickou výšku tváře u žen. Náhodná veličina  $X_{21}$  popisuje morfologickou výšku tváře první ženy,  $\dots$ , náhodná veličina  $X_{180}$  popisuje morfologickou výšku tváře 80. ženy. Předpokládejme, že náhodný výběr morfologické výšky tváře žen pochází z nějakého nám neznámého rozdělení  $L(\theta_2)$  s distribuční funkcí  $F_2(x)$ . Na hladině významnosti  $\alpha = 0.05$  chceme ze zadání testovat nulovou hypotézu o shodě distribučních funkcí  $F_1(x)$  a  $F_2(x)$ , přičemž je dovoleno, aby se distribuční funkce lišily pouze posunutím o vzdálenost  $\Delta$ . Testování provedeme v posloupnosti pěti kroků pomocí Kolmogorova-Smirnovova testu.

#### 1. Stanovení hypotéz

- **slovní formulace** nulové a alternativní hypotézy

$H_0$  : Distribuční funkce morfologické výšky tváře mužů a žen starověké egyptské populace se liší pouze posunutím o vzdálenost  $\Delta$ .

$H_1$  : Distribuční funkce morfologické výšky tváře mužů a žen starověké egyptské populace se neliší pouze posunutím o vzdálenost  $\Delta$ .

- **matematická formulace** nulové a alternativní hypotézy

$H_0$  :  $F_1(x) = F_2(x + \Delta), \forall x \in \mathbb{R}$ , kde  $\Delta \in \mathbb{R}$ .

$H_1$  :  $F_1(x) \neq F_2(x + \Delta)$ , pro nějaké  $x \in \mathbb{R}$ , přičemž  $\Delta \in \mathbb{R}$ .

#### 2. Volba hladiny významnosti

Hladina významnosti  $\alpha = 0.05$ .

#### 3. Testování $p$ -hodnotou


Z úvodního rozboru příkladu víme, že náhodný výběr morfologické výšky tváře mužů se pohybuje okolo hodnoty 114.5361 mm, zatímco náhodný výběr morfologické výšky tváře žen se pohybuje okolo hodnoty

107.5625 mm. Před otestováním nulové hypotézy musíme zajistit, aby rozdílné polohy obou náhodných výběrů negativně neovlivnily výsledek testu o shodě obou distribučních funkcí (s výjimkou posunutí). Oba náhodné výběry tedy metodou centrování posuneme do nuly. Nejprve od každé naměřené hodnoty morfologické výšky tváře mužů  $X_{11}, \dots, X_{1n_1}$  odečteme hodnotu aritmetického průměru 114.5361 mm. Získáme tak náhodný výběr centrovaných náhodných veličin  $X_{11,c}, \dots, X_{1n_1,c}$ .

$$\begin{aligned} X_{11,c} &= X_{11} - m_1 = 123 - 114.5361 = 8.4639 \\ X_{12,c} &= X_{12} - m_1 = 112 - 114.5361 = -2.5361 \\ &\vdots \\ X_{1166,c} &= X_{1166} - m_1 = 107 - 114.5361 = -7.5361 \end{aligned}$$

Analogicky od každé naměřené hodnoty morfologické výšky tváře žen  $X_{21}, \dots, X_{2n_2}$  odečteme hodnotu aritmetického průměru 107.5625 mm, čímž získáme náhodný výběr centrovaných náhodných veličin  $X_{21,c}, \dots, X_{2n_2,c}$ .

$$\begin{aligned} X_{21,c} &= X_{21} - m_2 = 112 - 107.5625 = 4.4375 \\ X_{22,c} &= X_{22} - m_2 = 104 - 107.5625 = -3.5625 \\ &\vdots \\ X_{280,c} &= X_{280} - m_2 = 111 - 107.5625 = 3.4375 \end{aligned}$$

Jak jsme uvedli výše, procesem centrování jsme docílili toho, že se hodnoty obou centrovaných náhodných výběrů  $X_{11,c}, \dots, X_{1n_1,c}$  a  $X_{21,c}, \dots, X_{2n_2,c}$  realizují okolo nuly. Konkrétně aritmetický průměr centrovaných morfologických výšek tváře mužů je roven hodnotě  $-2.2267 \times 10^{-15}$  mm zatímco aritmetický průměr centrovaných morfologických výšek tváře žen je roven přesně 0 mm. Nyní vytvoříme vektory centrovaných morfologických výšek tváře mužů (`face.HM.c`) a žen (`face.HF.c`) pomocí softwaru .

```
434 face.HM.c <- face.HM - m1
435 face.HF.c <- face.HF - m2
436 mean(face.HM.c) # -2.226727e-15
437 mean(face.HF.c) # 0
```

Nahrazením původních náhodných výběrů jejich centrovanými náhodnými výběry a následným aplikováním Kolmogorova-Smirnovova testu právě na tyto centrované náhodné výběry dochází k modifikaci původní nulové a alternativní hypotézy na následující znění:

- **slovní formulace** modifikované nulové a alternativní hypotézy  
 $H_0$  : Distribuční funkce centrované morfologické výšky tváře mužů a žen starověké egyptské populace jsou shodné.  
 $H_1$  : Distribuční funkce centrované morfologické výšky tváře mužů a žen starověké egyptské populace nejsou shodné.
- **matematická formulace** modifikované nulové a alternativní hypotézy  
 $H_0$  :  $F_{1c}(x) = F_{2c}(x), \forall x \in \mathbb{R}$ , kde  $\Delta \in \mathbb{R}$ .  
 $H_1$  :  $F_{1c}(x) \neq F_{2c}(x)$ , pro nějaké  $x \in \mathbb{R}$ , přičemž  $\Delta \in \mathbb{R}$ .

Nyní otestujeme modifikovanou nulovou hypotézu o shodě dvou distribučních funkcí centrovaných náhodných výběrů pomocí Kolmogorova-Smirnovova testu. Tím, že místo na původní náhodné výběry aplikujeme Kolmogorovův-Smirnovův test právě na centrované náhodné výběry, kde je již vliv posunutí odfiltrován, snadno

zjistíme, zda se distribuční funkce původních náhodných výběrů liší jinak než posunutím. Kolmogorov-Smirnovův test provedeme pomocí funkce `ks.test()`. Vstupními argumenty funkce budou vektory centrovaných náhodných výběrů `face.HM.c` a `face.HF.c`. Další argumenty není třeba specifikovat. Ze všech výstupů funkce `ks.test()` je pro nás nejrelevantnější hodnota testovací statistiky ( $D = 0.12063$ ) a  $p$ -hodnota ( $p\text{-value} = 0.4119$ ).

```
438 ks.test(face.HM.c, face.HF.c)
```

```

Two-sample Kolmogorov-Smirnov test
data: face.HM.c and face.HF.c
D = 0.12063, p-value = 0.4119
alternative hypothesis: two-sided

```

439  
440  
441  
442  
443  
444

#### Závěr testování:

Protože  $p$ -hodnota = 0.4119 je větší než  $\alpha = 0.05$ ,  $H_0$  nezamítáme na hladině významnosti  $\alpha = 0.05$ .

*Poznámka:* Výstup funkce `ks.test()` je tentokrát doprovázen varovnou hláškou `Warning message: In ks.test(face.HM.c, face.HF.c): p-value will be approximate in the presence of ties.` Hláška nás upozorňuje na to, naměřené hodnoty v alespoň jednom ze dvou náhodných výběrů obsahují shodná měření (`ties`) v jejichž důsledku byla  $p$ -hodnota testu vypočítána aproximativně, namísto exaktně. Aproximativní výpočet byl v tomto případě umožněn, protože rozsahy obou náhodných výběrů jsou dostatečně vysoké. Použití aproximativní metody na výpočet  $p$ -hodnoty nám však nijak nevádí, proto varovnou hlášku sice bereme na vědomí, ale praktický důsledek z ní pro nás nevyplývá.

#### 4. Grafická vizualizace výsledků testování

Tvary hustot obou náhodných výběrů zobrazíme pomocí dvojice histogramů superponovaných příslušnými křivkami jádrových odhadů hustoty a pomocí grafu dvojice distribučních funkcí  $F_1(x)$  a  $F_2(x)$  (viz obrázek 16). Pro účel vykreslení histogramů roztrídíme náhodný výběr morfologických výšek tváře mužů do osmi ekvidistantních třídících intervalů o šířce 7 mm prostřednictvím hranic 82, 89, ..., 138 mm a náhodný výběr morfologických výšek tváře žen do sedmi ekvidistantních třídících intervalů o šířce 3 mm prostřednictvím hranic 93, 96, ..., 121 mm. Oba histogramy opět vybarvíme průhlednými barvami vytvořenými prostřednictvím funkce `rgb()` z knihovny `grDevices`. Uvnitř funkce `rgb()` opět specifikujeme argumenty `red`, `green`, `blue` a `alpha`.

```

445 par(mar = c(4, 4, 1, 1), family = 'Times')
446 hist(face.HM, breaks = seq(82, 138, by = 7), prob = T,
447       col = grDevices::rgb(1, 0, 0, alpha = 0.2), density = 80,
448       main = '', ylab = 'relativní četnost', axes = F, xlab = '',
449       ylim = c(0, 0.08), xlim = c(80, 140))
450 box(bty = 'o')
451 axis(1, seq(80, 140, by = 10))
452 axis(2, las = 1)
453 lines(density(face.HM), col = 'red', lwd = 2)
454 mtext('morfologická výška tváře (v mm)', side = 1, line = 2.2)
455
456 hist(face.HF, breaks = seq(93, 121, by = 3), prob = T, density = 80,
457       col = grDevices::rgb(0, 0, 1, alpha = 0.2), add = T)
458 lines(density(face.HF), col = 'blue', lwd = 2)
459
460 legend('topleft', fill = c(rgb(0, 0, 1, 0.4), rgb(1, 0, 0, 0.4)),
461       legend = c('muži', 'ženy'), bty = 'n')

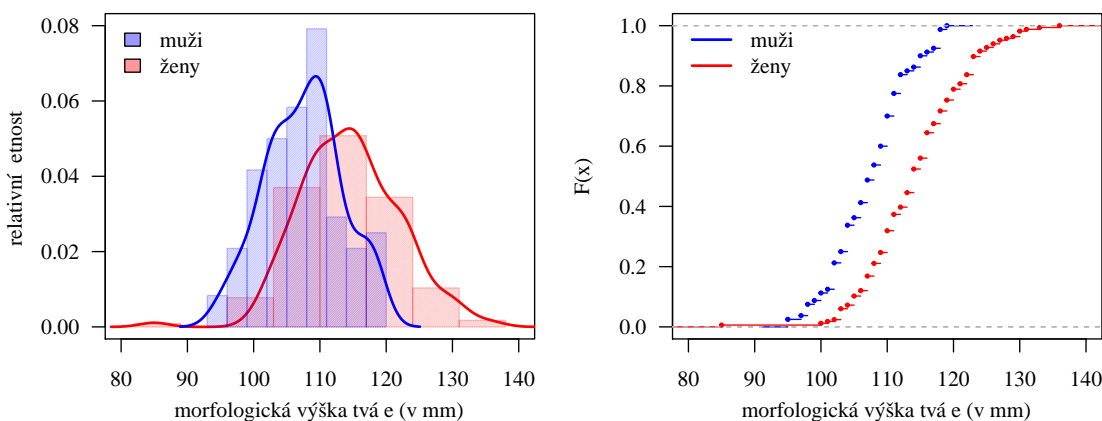
```

Druhým vykresleným grafem bude graf výběrových distribučních funkcí  $F_1(x)$  a  $F_2(x)$ . Hodnoty výběrové distribuční funkce  $F_1(x)$  pro morfologické výšky tváře mužů vypočítáme pomocí funkce `ecdf()` aplikované na vektor naměřených hodnot `face.HM`. Příkazem `plot(ecdf(face.HM))` rovnou vykreslíme graf zobrazující křivku výběrové distribuční funkce  $F_1(x)$ . Nezapomeneme opět rozšířit rozsah osy  $x$  tak, aby se do grafu vešla také křivka výběrové distribuční funkce  $F_2(x)$  pro morfologickou výšku tváře žen (například nastavení argumentu

$xlim = c(80, 140)$ ). Hodnoty výběrové distribuční funkce  $F_2(x)$  vypočítáme opět pomocí funkce `ecdf()` aplikované tentokrát na vektor naměřených hodnot `face.HM`. Příkazem `lines()` potom doplníme křivku výběrové distribuční funkce  $F_2(x)$  do stávajícího grafu.

```

462 plot(ecdf(face.HM), col = 'red', main = '', ylab = 'F(x)',
463       xlab = '', las = 1, lwd = 1, cex = 0.4, xlim = c(80, 140))
464 lines(ecdf(face.HF), col = 'blue', lwd = 1, cex = 0.4)
465 mtext('morfologická výška tváře (v mm)', side = 1, line = 2.2)
466
467 legend('topleft', lwd = 2, col = c('blue', 'red'),
468       legend = c('muži', 'ženy'), bty = 'n')
```



Obrázek 16: Histogramy (vlevo) a distribuční funkce (vpravo) náhodných výběrů morfologických výšek tváře (a) mužů (modrá barva); (b) žen (červená barva)

Grafická vizualizace uvedená na obrázku 16 podporuje výsledek testování. Histogramy i jádrové odhady hustoty obou výběrů jsou si tvarem velmi podobné. Taktéž křivky výběrových distribučních funkcí mají podobný tvar.

*Poznámka:* V grafu zobrazujícím výběrové distribuční funkce by nás mohlo zarazit, že hodnoty výběrové distribuční funkce morfologické výšky tváře mužů jsou vypočítány a vykresleny pouze v 35 hodnotách, zatímco původní vektor naměřených morfologických výšek tváře mužů je tvořen 166 hodnotami. Je to proto, že soubor 166 naměřených šířek dolní čelisti obsahuje pouze 35 unikátních hodnot (zbylých 131 měření nabývají hodnot stejných jako některá z 35 unikátních hodnot; jde o tzv. duplicitní hodnoty, které jsou ve funkci `ks.test` nazývané právě jako `ties`). Hodnoty výběrové distribuční funkce  $F_1(x)$  jsou z definice počítány pouze v 35 unikátních hodnotách (viz kapitola ???, sekce ???). Analogicky jsou potom hodnoty výběrové distribuční funkce morfologické výšky tváře žen vypočítány a vykresleny pouze ve 24 hodnotách, ačkoli původní vektor naměřených šířek dolní čelisti je tvořen 80 hodnotami. Soubor 80 naměřených šířek dolní čelisti obsahuje totiž pouze 24 unikátních hodnot, zatímco zbylých 56 hodnot je duplicitních. Závěrem poznamenejme, že v předchozím příkladě 6.13 tento problém nenastal, protože v žádném z obou náhodných výběrů nedošlo k výskytu jediné duplicitní hodnoty.

## 5. Interpretace výsledků

Distribuční funkce morfologické výšky tváře mužů a žen starověké egyptské populace se liší pouze posunutím.



### Příklad 6.15. Test o shodě distribučních funkcí dvou náhodných výběrů

Mějme datový soubor 16-anova-head.txt obsahující údaje o šířce dolní čelisti (bizyg.W) a sexuální orientaci (sexor) u mladých dospělých jedinců (mužů a žen), převážně studentů vysokých škol z Brna a Ostravy (viz sekce ??). Na hladině významnosti  $\alpha = 0.05$  testujte hypotézu o shodě distribučních funkcí šířky dolní čelisti u žen s výlučně heterosexuální orientací a u žen s jinou sexuální orientací.

#### Řešení příkladu 6.15

Nejprve načteme datový soubor 16-anova-head.txt. Z datového souboru vybereme naměřené hodnoty šířky dolní čelisti (bigo.W) žen (sex == 'f') s výlučně heterosexuální orientací (sexor == 'op') a vložíme je do proměnné bigo.WFO. Analogicky vytvoříme vektor bigo.WFS obsahující naměřené hodnoty šířky dolní čelisti (bigo.W) žen (sex == 'f') s jinou sexuální orientací (sexor == 'sa'). Z obou vektorů odstraníme chybějící pozorování a zjistíme rozsahy náhodných výběrů, rozpětí naměřených hodnot a aritmetické průměry.

```
469 data <- read.delim('00-Data//16-anova-head.txt')
470 bigo.WFO <- na.omit(data[data$sex == 'f' & data$sexor == 'op', 'bigo.W'])
471 bigo.WFS <- na.omit(data[data$sex == 'f' & data$sexor == 'sa', 'bigo.W'])
472 n1 <- length(bigo.WFO) # 62
473 n2 <- length(bigo.WFS) # 23
474 range(bigo.WFO) # 90-111
475 range(bigo.WFS) # 91-111
476 m1 <- mean(bigo.WFO) # 99.5
477 m2 <- mean(bigo.WFS) # 103.087
```

Datový soubor obsahuje údaje o šířce dolní čelisti u 62 žen s heterosexuální orientací a 23 žen s jinou sexuální orientací. Naměřené údaje o šířce dolní čelisti žen s heterosexuální orientací se pohybují v rozmezí 90-111 mm okolo hodnoty 99.5 mm. Naměřené údaje o šířce dolní čelisti žen s jinou sexuální orientací se pohybují v rozmezí 91-111 mm okolo hodnoty 103.087 mm.

Náhodný výběr  $X_{11}, X_{12}, \dots, X_{1n_1}$ ,  $n_1 = 62$ , popisuje šířku dolní čelisti žen s heterosexuální orientací. Náhodná veličina  $X_{11}$  popisuje šířku dolní čelisti první ženy,  $\dots$ , náhodná veličina  $X_{162}$  popisuje šířku dolní čelisti 62. ženy. Předpokládejme, že tento náhodný výběr pochází z nějakého nám neznámého rozdělení  $L(\theta_1)$  s distribuční funkcí  $F_1(x)$ . Náhodný výběr  $X_{21}, X_{22}, \dots, X_{2n_2}$ ,  $n_2 = 23$  popisuje šířku dolní čelisti žen s jinou sexuální orientací. Náhodná veličina  $X_{21}$  popisuje šířku dolní čelisti první ženy,  $\dots$ , náhodná veličina  $X_{123}$  popisuje šířku dolní čelisti 23. ženy. Předpokládejme, že tento náhodný výběr pochází z nějakého nám neznámého rozdělení  $L(\theta_2)$  s distribuční funkcí  $F_2(x)$ . Na hladině významnosti  $\alpha = 0.05$  chceme testovat nulovou hypotézu o shodě distribučních funkcí  $F_1(x)$  a  $F_2(x)$ , přičemž je dovoleno, aby se distribuční funkce lišily pouze posunutím o vzdálenost  $\Delta$ . Testování provedeme pomocí Kolmogorova-Smirnovova testu.

#### 1. Stanovení hypotéz

- **slovní formulace** nulové a alternativní hypotézy

$H_0$  : Distribuční funkce šířky dolní čelisti žen s výlučně heterosexuální orientací a s jinou sexuální orientací se liší pouze posunutím o vzdálenost  $\Delta$ .

$H_1$  : Distribuční funkce šířky dolní čelisti žen s výlučně heterosexuální orientací a s jinou sexuální orientací se neliší pouze posunutím o vzdálenost  $\Delta$ .

- **matematická formulace** nulové a alternativní hypotézy

$H_0$  :  $F_1(x) = F_2(x + \Delta), \forall x \in \mathbb{R}$ , kde  $\Delta \in \mathbb{R}$ .

$H_1$  :  $F_1(x) \neq F_2(x + \Delta)$ , pro nějaké  $x \in \mathbb{R}$ , přičemž  $\Delta \in \mathbb{R}$ .

#### 2. Volba hladiny významnosti

Hladina významnosti  $\alpha = 0.05$ .

#### 3. Testování $p$ -hodnotou

Z úvodního rozboru příkladu víme, že náhodný výběr šířky dolní čelisti žen s heterosexuální orientací se pohybuje okolo hodnoty 99.5 mm, zatímco náhodný výběr šířky dolní čelisti žen s jinou sexuální orientací se pohybuje okolo hodnoty 103.087 mm. Před otestováním nulové hypotézy musíme zajistit, aby rozdílné polohy

obou náhodných výběrů negativně neovlivnily výsledek testu o shodě obou distribučních funkcí (s výjimkou posunutí). Oba náhodné výběry tedy metodou centrování posuneme do nuly.

```
478 bigo.WFO.c <- bigo.WFO - m1
479 bigo.WFS.c <- bigo.WFS - m2
480 mean(bigo.WFO.c) # 0
481 mean(bigo.WFS.c) # 4.942755e-15
```

Z nově vypočítaných aritmetických průměrů vidíme, že hodnoty obou centrovaných náhodných výběrů  $X_{11,c}, \dots, X_{1n_1,c}$  a  $X_{21,c}, \dots, X_{2n_2,c}$  se skutečně realizují okolo nuly. Konkrétně aritmetický průměr centrovaných šířek dolní čelisti žen s heterosexuální orientací je roven přesně 0 mm zatímco aritmetický průměr šířek dolní čelisti žen s jinou sexuální orientací je roven hodnotě  $4.9428 \times 10^{-15}$  mm. Nahrazením původních náhodných výběrů centrovanými náhodnými výběry a následným aplikováním Kolmogorova-Smirnovova testu právě na tyto centrované náhodné výběry dochází k modifikaci původní nulové a alternativní hypotézy na následující znění:

- **slovní formulace** modifikované nulové a alternativní hypotézy  
 $H_0$  : Distribuční funkce šířky dolní čelisti žen s výlučně heterosexuální orientací a s jinou sexuální orientací jsou shodné.  
 $H_1$  : Distribuční funkce šířky dolní čelisti žen s výlučně heterosexuální orientací a jinou sexuální orientací nejsou shodné.
- **matematická formulace** modifikované nulové a alternativní hypotézy  
 $H_0$  :  $F_{1c}(x) = F_{2c}(x), \forall x \in \mathbb{R}$ , kde  $\Delta \in \mathbb{R}$ .  
 $H_1$  :  $F_{1c}(x) \neq F_{2c}(x)$ , pro nějaké  $x \in \mathbb{R}$ , přičemž  $\Delta \in \mathbb{R}$ .

Nyní otestujeme modifikovanou nulovou hypotézu o shodě dvou distribučních funkcí centrovaných náhodných výběrů pomocí Kolmogorova-Smirnovova testu. Tím, že místo na původní náhodné výběry aplikujeme Kolmogorův-Smirnovův test právě na centrované náhodné výběry, kde je již vliv posunutí odfiltrován, snadno zjistíme, zda se distribuční funkce původních náhodných výběrů liší jinak než posunutím. Kolmogorův-Smirnovův test provedeme pomocí funkce `ks.test()`, kterou aplikujeme na vektory centrovaných náhodných veličin `bigo.WFO.c` a `bigo.WFS.c`. Z výstupu funkce `ks.test()` vidíme, že hodnota testovací statistiky K-S testu ( $D = 0.29453$ ) a  $p$ -hodnota ( $p$ -value = 0.1089).

```
482 ks.test(bigo.WFO.c, bigo.WFS.c)
```

```
Two-sample Kolmogorov-Smirnov test
data:  bigo.WFO.c and bigo.WFS.c
D = 0.29453, p-value = 0.1089
alternative hypothesis: two-sided
```

483  
484  
485  
486  
487  
488

#### Závěr testování:

Protože  $p$ -hodnota = 0.1089 je větší než  $\alpha = 0.05$ ,  $H_0$  nezamítáme na hladině významnosti  $\alpha = 0.05$ .

*Poznámka:* Výstup funkce `ks.test()` je tentokrát doprovázen varovnou hláškou `Warning message: In ks.test(bigo.WFO.c, bigo.WFS.c) : cannot compute exact p-value with ties`. Hláška nás upozorňuje na to, že naměřené hodnoty v alespoň jednom ze dvou náhodných výběrů obsahují shodná měření (`ties`), v jejichž důsledku nebyla  $p$ -hodnota testu vypočítána exaktně. To nám však pro naše potřeby nijak nevádí, proto varovnou hlášku sice bereme na vědomí, ale praktický důsledek z ní pro nás nevyplývá.

#### 4. Grafická vizualizace výsledků testování

Tvary hustot obou náhodných výběrů zobrazíme pomocí dvojice histogramů superponovaných příslušnými křivkami jádrových odhadů hustoty a pomocí grafu dvojice distribučních funkcí (viz obrázek 17). Pro účel vykreslení histogramů roztrídíme náhodný výběr šířek dolní čelisti žen s heterosexuální orientací do sedmi ekvidistantních třídících intervalů o šířce 3 mm prostřednictvím hranic 90, 93, ..., 111 mm a náhodný výběr

```

489 par(mar = c(4, 4, 1, 1), family = 'Times')
490 hist(bigo.WF0, breaks = seq(90, 111, by = 3), prob = T,
491       col = grDevices::rgb(1, 0, 0, alpha = 0.2), density = 80,
492       main = '', ylab = 'relativní četnost', axes = F, xlab = '',
493       ylim = c(0, 0.14), xlim = c(75, 125))
494 box(bty = 'o')
495 axis(1, seq(80, 120, by = 10))
496 axis(2, las = 1)
497 lines(density(bigo.WF0), col = 'red', lwd = 2)
498 mtext('šířka dolní čelisti (v mm)', side = 1, line = 2.2)
499
500 hist(bigo.WFS, breaks = seq(91, 111, by = 4), prob = T, density = 80,
501       col = grDevices::rgb(0, 0, 1, alpha = 0.2), add = T)
502 lines(density(bigo.WFS), col = 'blue', lwd = 2)
503
504 legend('topleft', fill = c(rgb(0, 0, 1, 0.4), rgb(1, 0, 0, 0.4)),
505       legend = c('ženy s heterosexuální o.', 'ženy s jinou sexuální o.'), bty = 'n')

```

šířek dolní čelisti žen s jinou sexuální orientací do pěti ekvidistantních třídících intervalů o šířce 4 mm prostřednictvím hranic 91, 95, ..., 111 mm.

Druhým vykresleným grafem bude graf výběrových distribučních funkcí  $F_1(x)$  a  $F_2(x)$ . Hodnoty výběrové distribuční funkce  $F_1(x)$  pro šířku dolní čelisti žen s heterosexuální orientací vypočítáme pomocí funkce `ecdf()` aplikované na vektor naměřených hodnot `bigo.WF0`. Příkazem `plot()` potom vykreslíme graf zobrazující křivku této výběrové distribuční funkce. Nezapomeneme opět rozšířit rozsah osy  $x$  tak, aby se do grafu vešla také křivka výběrové distribuční funkce  $F_2(x)$  pro šířku dolní čelisti žen s jinou sexuální orientací (například nastavením argumentu `xlim = c(80, 120)`). Hodnoty výběrové distribuční funkce  $F_2(x)$  vypočítáme rovněž pomocí funkce `ecdf()` aplikované tentokrát na vektor naměřených hodnot `bigo.WFS`.

```

506 plot(ecdf(bigo.WF0), col = 'red', main = '', ylab = 'F(x)',
507       xlab = '', las = 1, lwd = 1, cex = 0.4, xlim = c(80, 120))
508 lines(ecdf(bigo.WFS), col = 'blue', lwd = 1, cex = 0.4)
509 mtext('šířka dolní čelisti (v mm)', side = 1, line = 2.2)
510
511 legend('topleft', lwd = 2, col = c('blue', 'red'),
512       legend = c('ženy s heterosexuální o.', 'ženy s jinou sexuální o.'), bty = 'n')

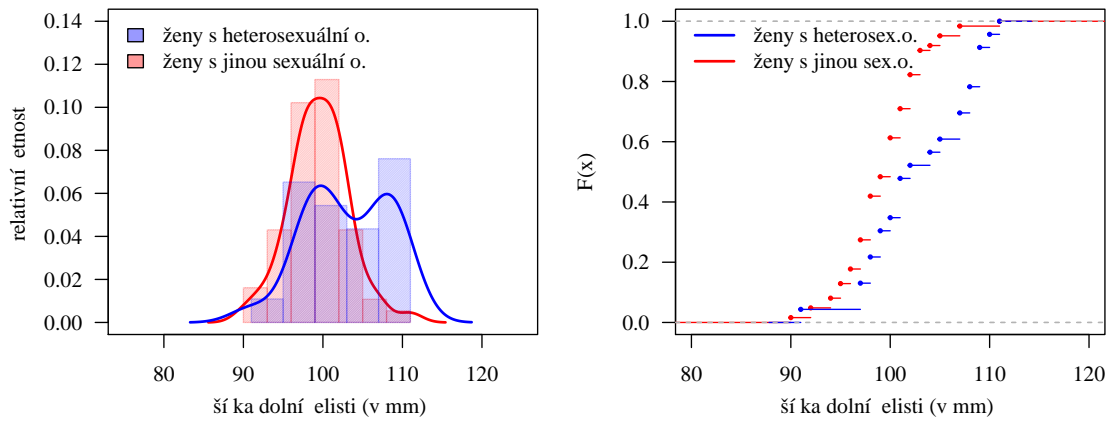
```

Z grafu porovnávajícího histogramy a jádrové odhady hustot dvou náhodných výběrů, který je vykreslen na obrázku 17, bychom usuzovali spíše na rozdílnost tvarů rozdělení obou náhodných výběrů. Histogramy i křivky jádrových odhadů hustot si tvarem příliš podobné nejsou. Rozdílnost v distribučních funkcích však není natolik patrná jako v grafu s histogramy, nicméně je značně patrnější než v předchozích dvou příkladech. Zcela jistě v tomto příkladě nejde o ukázkový případ shody tvaru dvou distribučních funkcí, čemuž odpovídá i výsledná  $p$ -hodnota K-S testu blízká hladině významnosti  $\alpha = 0.10$ . Grafická vizualizace tak podporuje závěr testování, že ačkoli shoda distribučních funkcí obou náhodných výběrů není tak suverénní jako v předchozích dvou příkladech, oba náhodné výběry ještě pochází z rozdělení se shodnými distribučními funkcemi.

*Poznámka:* Analogicky jako tomu bylo v příkladu 6.14, i zde oba náhodné výběry obsahují duplicitní hodnoty (ties). Konkrétně náhodný výběr šířky dolní čelisti žen s heterosexuální orientací obsahuje 16 unikátních hodnot a 46 duplicitních hodnot. Proto je křivka výběrové distribuční funkce  $F_1(x)$  vypočítána a vykreslena pouze v 16 bodech. Náhodný výběr šířky dolní čelisti žen s jinou sexuální orientací obsahuje potom 14 unikátních hodnot a 9 duplicitních hodnot. Proto je křivka výběrové distribuční funkce  $F_2(x)$  vypočítána a vykreslena pouze ve 14 bodech.

## 5. Interpretace výsledků

Distribuční funkce šířky dolní čelisti žen s výlučně heterosexuální orientací a jinou sexuální orientací se liší pouze posunutím.



Obrázek 17: Histogramy (vlevo) a distribuční funkce (vpravo) náhodných výběrů šířek dolní čelisti žen (a) s heterosexuální orientací (modrá barva); (b) s jinou sexuální orientací

