

# 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.  $X \not\sim N(\mu, \sigma^2)$ .

V závislosti na rozsahu  $n$  náhodného výběru používáme k testování normality jeden ze tří nejpoužívanějších testů normality. Je-li rozsah náhodného výběru  $n \leq 30$ , použijeme na testování hypotézy  $H_0$  Shapiro-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  $H_0$  Lillieforsův test, který je implementován ve funkci `lillie.test()` v knihovně `nortest` nebo Anderson-Darlingův test, který je implementován ve funkci `ad.test()` taktéž v knihovně `nortest`. Přesnými algoritmy testů se zabývat nebudeme, vystačíme si pouze se znalostí zmíněných funkcí implementovaných v softwaru . Všechny tři funkce nám jako výstup poskytují  $p$ -hodnotu, na jejímž základě 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 třídicích 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 kvantilový graf, 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 odlehlými od referenční křivky na pravém nebo levém chvostu náhodného výběru.

#### 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  $H_0$  Shapiro-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

Shapiro-Wilkův test provedeme pomocí funkce `shapiro.test()`. Výstupem funkce je údaj o použité proměnné (`intorb.BB`), hodnota testovací statistiky Shapiro-Wilkova testu ( $W = 0.94268$ ) a výsledná  $p$ -hodnota ( $p$ -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 kvantilovým grafem.

Pro vykreslení histogramu nejprve stanovíme hranice třídících intervalů, do kterých rozdělíme naměřené hodnoty. Datový soubor rozdělíme na základě Sturgesova pravidla (viz sekce ??) do pěti ekvidistantních intervalů s šířkou 2mm 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 třídících intervalů (argument `at`) a měřítko osy  $y$  (argument `side = 2`) jehož hodnoty budou vykresleny horizontálně (argument `las = 1`).

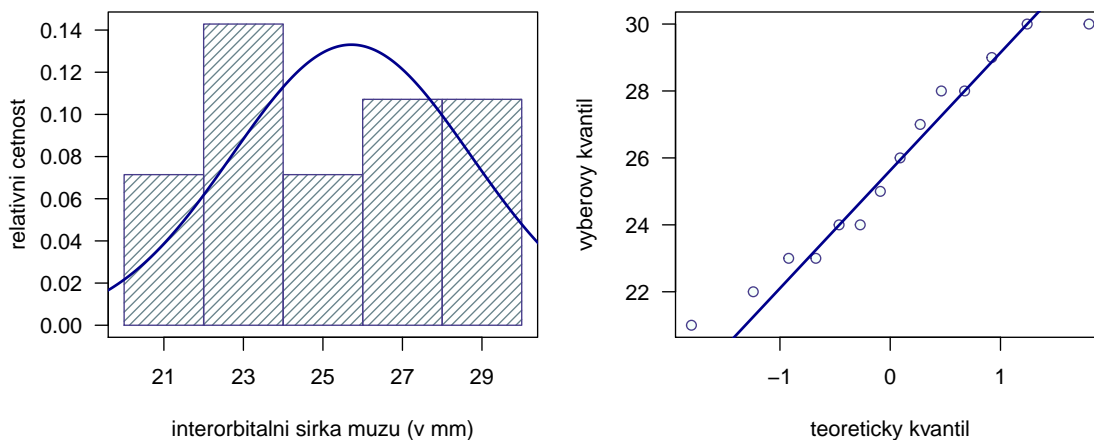
Nakonec do histogramu dokreslíme křivku hustoty normálního rozdělení  $N(\mu, \sigma^2)$ , kde parametry  $\mu$  a  $\sigma^2$  odhadneme jejich nestrannými odhady. Nejprve vygenerujeme dostatečně hustou posloupnost čísel pokrývající rozsah hodnot osy  $x$ , tj. posloupnost `xfit` tisíce čísel v rozsahu 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()`.

Kvantilový graf 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 obrysově barvy. Dále zakážeme vypsání nadpisu grafu (`main = ""`) a pomocí argumentů `xlab` a `ylab` změníme popisky os  $x$  a  $y$  z anglických na české ekvivalenty. Konečně argumentem `las` změníme 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 = 'relativni cetnost',
15     xlab = 'interorbitalni sirka muzu (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 = 'teoreticky kvantil', ylab = 'vyberovy kvantil', las = 1)
28 qqline(intorb.BB, col = 'darkblue', lwd = 2)

```



Obrázek 1: Histogram a kvantilový graf interorbitalní šířky u skeletů mužského pohlaví bantuské populace (v mm)

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 kvantilového grafu 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 interorbitalní šířky mužů bantuské populace pochází z normálního rozdělení.



## 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 Shapiro-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

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

```
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

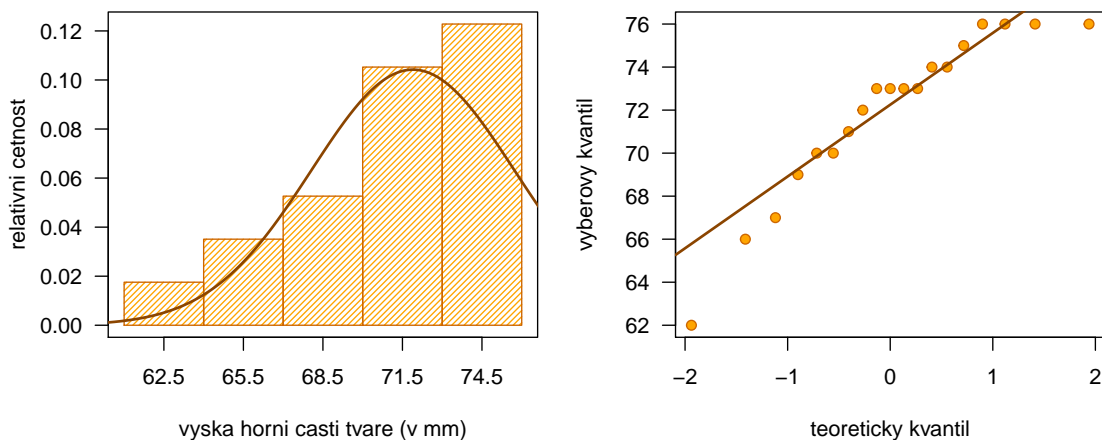
Závěr testování nyní podpoříme histogramem superponovaným křivkou normálního rozdělení a kvantilovým grafem. Analogicky jako v příkladu 6.1 vykreslíme histogram příkazem hist(), kde na základě Sturgesova pravidla rozdělíme naměřené hodnoty do pěti ekvidistantních intervalů s šířkou 3 mm prostřednictvím 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 kvantilový graf (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 = 'relativni cetnost',
43     xlab = 'vyska horni casti tvare (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 = 'teoreticky kvantil', ylab = 'vyberovy kvantil', las = 1)
56 qqline(upface.HN, col = 'darkorange4', lwd = 2)

```



Obrázek 2: Histogram a kvantilový graf 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 kvantilovém grafu 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 kvantilového grafu 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 kvantilovým grafem. 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().

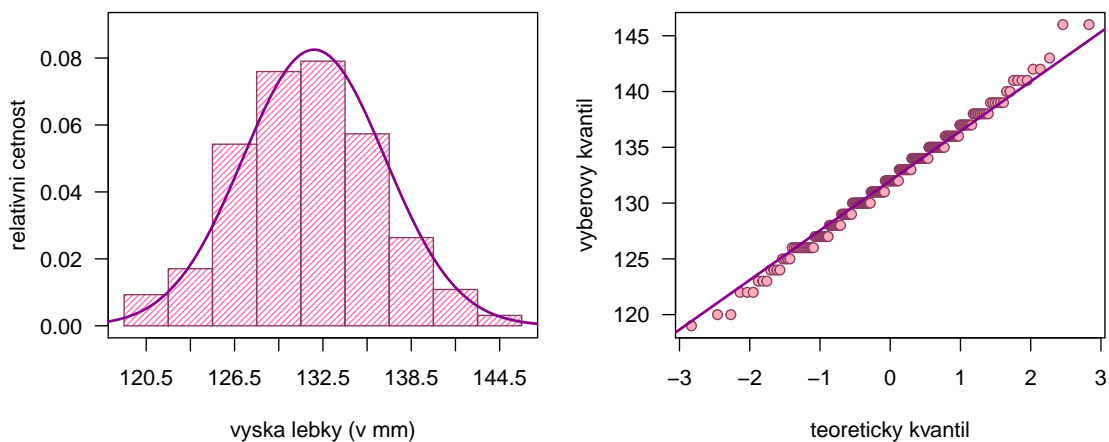
Rozdělení datového souboru ukázkově odpovídá normálnímu rozdělení. Na histogramu vykresleném na obrázku 3 je krásně viditelné, jak naměřená data kopírují tvar křivky hustoty normálního rozdělení. Taktéž kvantilový graf, kde se až na dvě, tři výjimky, drží všechny body v blízkém okolí referenční přímky, 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 = 'relativni cetnost',
71     xlab = 'vyska 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 = 'teoreticky kvantil', ylab = 'vyberovy kvantil', las = 1)
82 qqline(skull.HM, col = 'darkmagenta', lwd = 2)

```



Obrázek 3: Histogram a kvantilový graf 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

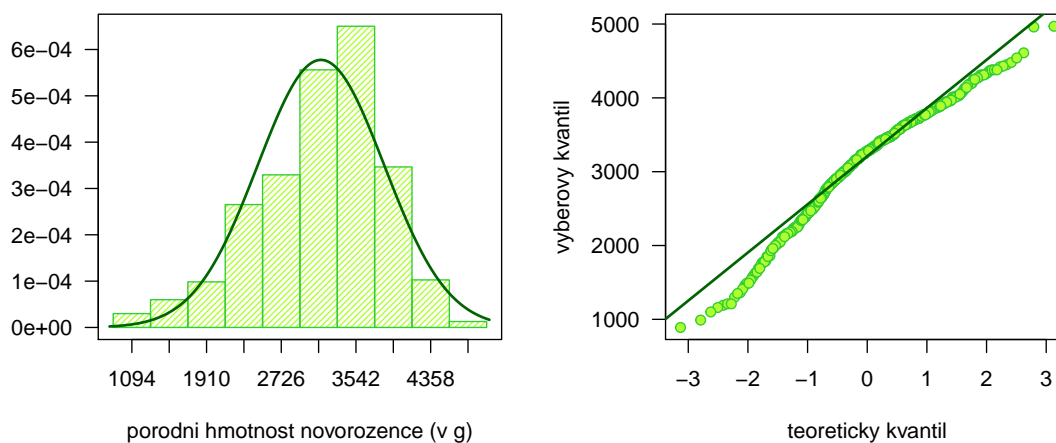
#### 4. Závěr testování

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 kvantilového grafu. 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 kvantilový graf 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 kvantilový graf 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. Kvantilový graf, 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 novorozenech 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 Mardiaův 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 Henze-Zirklerova testu nebo Roystonova testu vícerozměrné normality. Všechny tyto testy jsou k dispozici ve funkci `mvn()` implementované v knihovně `MVN`. Volbu Mardiaova testu zvolíme nastavením argumentu `mvnTest = 'mardia'` (viz příklad 6.5), volbu Henze-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.

### 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, Henze-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 Henze-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 Henze-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 `$Descriptives`, z nichž poslední dvě tabulky obsahují úplně stejné hodnoty jako při použití Mardiova testu a Henze-Zirklerova testu. Opět nás zajímá pouze tabulka `$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 = 'vyska nosu (v mm)',
117       ylab = 'interorbitalni sirka (v mm)', zlab = "relativni cetnost",
118       theta = -20, phi = 30, col = terrain.colors(12)[vyska])

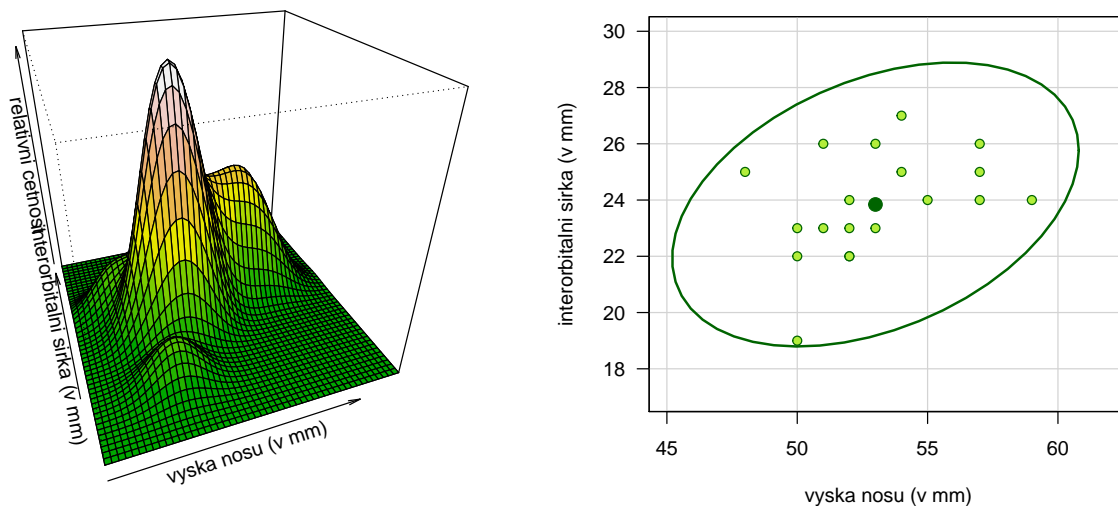
```

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 = 'vyska nosu (v mm)', ylab = 'interorbitalni sirka (v mm)',
121                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 interorbitalní 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 rozdělání. Při takto malém počtu hodnot však nemůžeme s jistotou říci, že jde skutečně o směs. Malý vrcholek vpředu grafu značí odlehle 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 interorbitalní 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.

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

Mardiaův 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.

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

	Test	Statistic	p value	Result	
1	Mardia Skewness	9.8656660128771	0.042752347786637	NO	129
2	Mardia Kurtosis	0.604444695534365	0.545548027690992	YES	130
3	MVN	<NA>	<NA>	NO	131

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 Henze-Zirklerova testu, a to použitím funkce `mvn()` s nastavením argumentu `mvnTest = 'hz'`.

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

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

134  
135

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'`.

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

	Test	H	p value	MVN
1	Royston	6.417614	0.04040737	NO

137  
138

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.

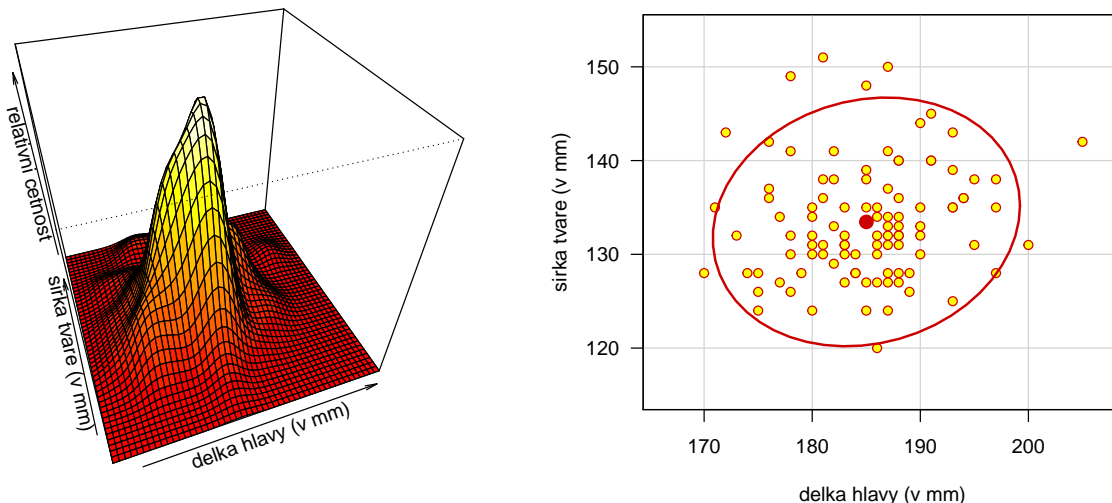
```
139 souradnice <- MASS::kde2d(udaje$head.L, udaje$bizyg.W, n = 50, lims = c(166, 211, 115, 154))
140 n <- dim(souradnice$z)[1]
141 vyska <- souradnice$z[-1, -1] + souradnice$z[-1, -n] + souradnice$z[-n, -1] + souradnice$z[-n, -n]
142 vyska <- cut(vyska, 12)
143
144 persp(souradnice$x, souradnice$y, souradnice$z, xlab = 'delka hlavy (v mm)',
145       ylab = 'sirka tvare (v mm)', zlab = "relativni cetnost",
146       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.

```

147 car::dataEllipse(udaje$head.L, udaje$bizyg.W, level = 0.90,
148                   xlim = c(166, 207), ylim = c(115, 154), main = '',
149                   xlab = 'delka hlavy (v mm)', ylab = 'sirka tvare (v mm)',
150                   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.

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

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

	Test	Statistic	p value	Result	
1	Mardia Skewness	11.0266139489775	0.0262665328995546	NO	158
2	Mardia Kurtosis	-0.160547155018785	0.872450079003065	YES	159
3	MVN	<NA>	<NA>	NO	160
					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.

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

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

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

163  
164

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'`.

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

	Test	H	p value	MVN
1	Royston	2.320783	0.3134003	YES

166  
167

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

Henze-Zirklerův test a Roystonův test  $H_0$  nezamítají, Mardiuův 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.

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

Dále vykreslíme tečkový diagram superponovaný 95% elipsou spolehlivosti.

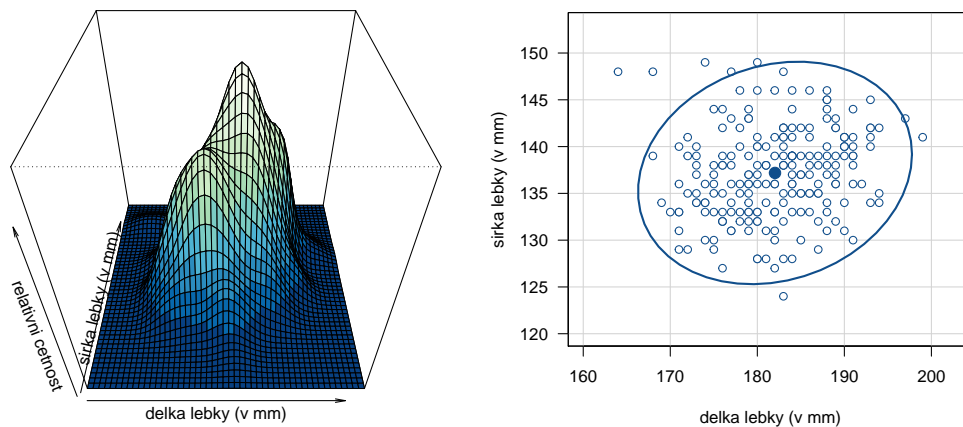
```
177 car::dataEllipse(udaje$skull.L, udaje$skull.B, level = 0.95, xlim = c(160, 203),
178                  ylim = c(120, 153), main = '', xlab = 'delka lebky (v mm)',
179                  ylab = 'sirka lebky (v mm)', pch = 21, col = 'dodgerblue4',
180                  bg = 'white', lwd = 2, las =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

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

```
187 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

188  
189  
190  
191

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 Henze-Zirklerův test dvourozměrné normality.

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

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

193  
194

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.

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

	Test	H	p value	MVN
1	Royston	8.396178	0.01505224	NO

196  
197

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 Henze-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.

```
198 souradnice <- MASS::kde2d(udaje$nose.B, udaje$intorb.B, n = 50,
199                          lims = c(17, 28, 17, 30))
200 n <- dim(souradnice$z)[1]
201 vyska <- souradnice$z[-1, -1] + souradnice$z[-1, -n] +
202         souradnice$z[-n, -1] + souradnice$z[-n, -n]
203 vyska <- cut(vyska, 9)
204 par(mar = c(4, 4, 1, 1))
205 persp(souradnice$x, souradnice$y, souradnice$z, xlab = 'sirka nosu (v mm)',
206       ylab = 'interorbitalni sirka (v mm)', zlab = "relativni cetnost",
207       col = rev(RColorBrewer::brewer.pal(9, 'YlOrBr'))[vyska],
208       theta = 20, phi = 40)
209
210 car::dataEllipse(udaje$nose.B, udaje$intorb.B, level = 0.95,
211                 xlim = c(18, 28), ylim = c(17, 29), xlab = 'sirka nosu (v mm)',
212                 ylab = 'interorbitalni sirka (v mm)', main = '', pch = 21,
213                 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 Mardiiova a Henze-Zirklerova testu o dvourozměrné normalitě náhodného výběru.

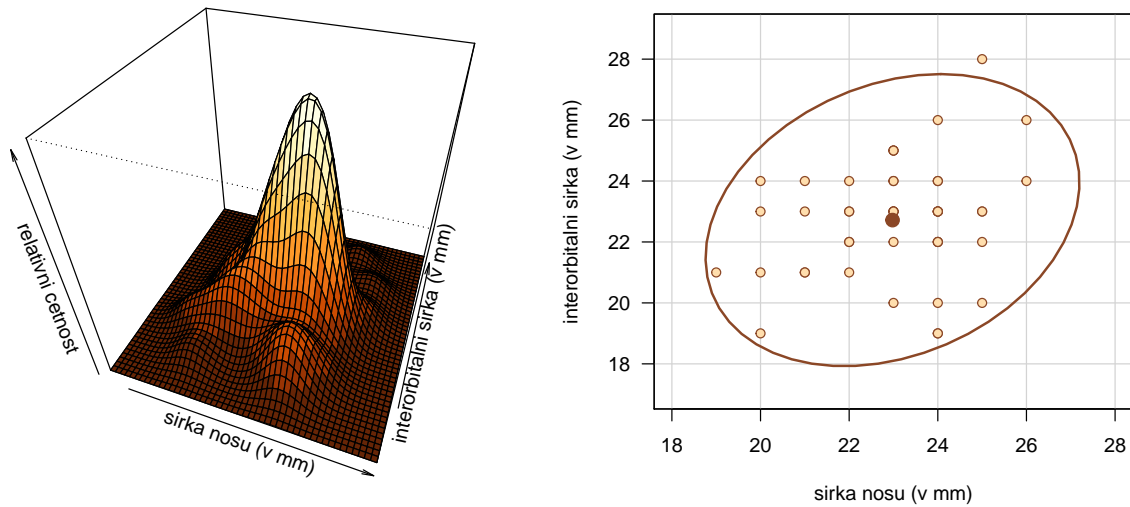
#### 5. Interpretace výsledků

Náhodný výběr popisující šířku nosu a interorbitalní 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.