M6120 – 5. cvičení : **M6120cv05** (*Testování jednoho a dvou náhodných výběrů - parmetrické i neparametrické testy*)

A. Inference pro výběry z normálního rozdělení.

Induktivní statistické metody jsou založeny na vztazích mezi následujícími teoretickými rozděleními:

Normální a odvozená rozdělení

Nechť $k, n \in \mathbb{N}, \nu, \nu_1, \nu_2, \ldots, \nu_k \in \mathbb{N}$, $b_0, b_1, \ldots, b_n \in \mathbb{R}$, $\exists i \in \{1, \ldots, n\} : b_i \neq 0$

Normální rozdělení $X \sim N(\mu, \sigma^2) \sim f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ $x \in \mathbb{R}$; $EX = \mu, DX = \sigma^2$

$$\perp\!\!\!\perp \{X_1, \ldots, X_n\} \wedge X_i \sim N(\mu_i, \sigma_i^2) \quad \Rightarrow \quad b_0 + \sum_{i=1}^{n} b_i X_i \sim N\left(b_0 + \sum_{i=1}^{n} b_i \mu_i, \sum_{i=1}^{n} b_i^2 \sigma_i^2\right)$$

$$X \sim N(\mu, \sigma^2) \quad \Rightarrow \quad U = \frac{X-\mu}{\sigma} \sim N(0,1)$$

χ^2 rozdělení

$$\perp\!\!\!\perp \{U_1, \ldots, U_\nu\} \simeq N(0,1) \quad \Rightarrow \quad K = U_1^2 + \cdots + U_\nu^2 \sim \chi^2(\nu)$$

$$\perp\!\!\!\perp \{X_1, \ldots, X_n\} \left\{K_1 \sim \chi^2(\nu_1), \ldots, K_k \sim \chi^2(\nu_k)\right\} \quad \Rightarrow \quad K = K_1 + \cdots + K_k \sim \chi^2(\nu_1 + \cdots + \nu_k)$$

Studentovo t-rozdělení

$$U \sim N(0,1) \perp K \sim \chi^2(\nu) \quad \Rightarrow \quad T = \frac{U}{\sqrt{\frac{K}{\nu}}} \sim t(\nu)$$

Fisherovo-Snedecorovo F-rozdělení

$$K_1 \sim \chi^2(\nu_1) \perp K_2 \sim \chi^2(\nu_2) \quad \Rightarrow \quad F = \frac{K_1/\nu_1}{K_2/\nu_2} \sim F(\nu_1, \nu_2)$$

Výběrová rozdělení

Usuzování o parametrech (popř. parametrických funkcích) se děje na základě **náhodných výběrů**, ze kterých jsou vytvářeny **statistiky**. Mezi základní statistiky patří **výběrový průměr** a **výběrový rozptyl**, které jsou nestrannými odhady střední hodnoty a rozptylu.

Náhodný výběr $\perp\!\!\!\perp \{X_1, \ldots, X_n\} \simeq \mathcal{L}\left(\mu(\boldsymbol{\theta}), \sigma^2(\boldsymbol{\theta})\right)$

Výběrové statistiky

Výběrový průměr: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \Rightarrow \quad E\bar{X} = \mu(\boldsymbol{\theta}), \quad D\bar{X} = \frac{\sigma^2(\boldsymbol{\theta})}{n}$

Výběrový rozptyl: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \quad \Rightarrow \quad ES^2 = \sigma^2(\boldsymbol{\theta})$

Mezi náhodnými výběry hrají velmi důležitou roli náhodné výběry z normálního rozdělení. Pokud provádíme úsudky o poloze a variabilitě, můžeme tak činit na základě jednoho či více výběrů.

VLASTNOSTI VÝBĚROVÝCH STATISTIK V PŘÍPADĚ 1 VÝBĚRU Z NORMÁLNÍHO ROZDĚLENÍ

$$\perp\!\!\!\perp \{X_1, \ldots, X_n\} \simeq \mathcal{L}\left(\mu(\boldsymbol{\theta}), \sigma^2(\boldsymbol{\theta})\right) = N(\mu, \sigma^2)\,,$$

tj. $\mu(\boldsymbol{\theta})=\mu$, $\sigma^2(\boldsymbol{\theta})=\sigma^2$, kde $\boldsymbol{\theta}=(\theta_1,\theta_2)=(\mu,\sigma^2)$, přičemž $\mu \in \mathbb{R}$, $\sigma^2 > 0$

$$\Rightarrow \quad \begin{array}{ll} (i) & \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \Rightarrow \quad U_{\bar{X}} = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1) \\ (ii) & \bar{X} \text{ a } S^2 \text{ jsou stochasticky nezávislé, tj. } \bar{X} \perp S^2 \\ (iii) & K = \frac{n-1}{\sigma^2} S^2 = \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma}\right)^2 \sim \chi^2(n-1) \\ (iv) & T = \frac{U_{\bar{X}}}{\sqrt{K/(n-1)}} = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1) \end{array}$$

VLASTNOSTI VÝBĚROVÝCH STATISTIK V PŘÍPADĚ 1 VÝBĚRU Z DVOUROZMĚRNÉHO NORMÁLNÍHO ROZDĚLENÍ

$$\perp\!\!\!\perp \{X_1, \ldots, X_n\} \left\{\mathbf{X}_1 = \begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \ldots, \mathbf{X}_n = \begin{pmatrix} X_n \\ Y_n \end{pmatrix}\right\} \simeq N_2\left(\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right)$$

tj. $\mu(\boldsymbol{\theta}) = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$, $\sigma^2(\boldsymbol{\theta}) = \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$ a $\boldsymbol{\theta}(\theta_1, \ldots, \theta_5) = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$,
přičemž $\mu_{,}\mu_2 \in \mathbb{R}$, $\sigma_1^2 > 0$, $\sigma_2^2 > 0$, $\rho \in (0,1)$.

Tento tzv. „**párový**“ náhodný výběr převedeme na „rozdílový“:

$$\perp\!\!\!\perp \{X_1, \ldots, X_n\} \, (Z_1 = X_1 - Y_1, \ldots, Z_n = X_n - Y_n) \simeq N(\mu_Z = \mu_1 - \mu_2, \sigma_Z^2 = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)$$

$$\Rightarrow \quad \begin{array}{ll} (i) & \bar{Z} \sim N\left(\mu_Z = \mu_1 - \mu_2, \frac{\sigma_Z^2}{n}\right) \quad \Rightarrow \quad U_{\bar{Z}} = \frac{\bar{Z}-\mu_Z}{\sigma_Z/\sqrt{n}} = \frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sigma_Z/\sqrt{n}} \sim N(0,1) \\ (ii) & \bar{Z} = \bar{X} - \bar{Y} \text{ a } S_Z^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Z_i - \bar{Z})^2 \text{ jsou stochasticky nezávislé, tj. } \bar{Z} \perp S_Z^2 \\ (iii) & K_Z = \frac{n-1}{\sigma_Z^2} S_Z^2 \sim \chi^2(n-1) \\ (iv) & T_Z = \frac{U_{\bar{Z}}}{\sqrt{K_Z/(n-1)}} = \frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{S_Z/\sqrt{n}} \sim t(n-1) \end{array}$$

VLASTNOSTI VÝBĚROVÝCH STATISTIK V PŘÍPADĚ 2 NEZÁVISLÝCH VÝBĚRŮ Z NORMÁLNÍHO ROZDĚLENÍ

$$\perp \begin{cases} \perp\!\!\!\perp \{X_1,\dots,X_{n_1}\} \simeq N(\mu_1,\sigma_1^2) & \bar{X},\ S_X^2 \\ \perp\!\!\!\perp \{Y_1,\dots,Y_{n_2}\} \simeq N(\mu_2,\sigma_2^2) & \bar{Y},\ S_Y^2 \end{cases}$$

(i) $\quad \bar{X}-\bar{Y} \sim N\left(\mu_1-\mu_2, \frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right) \Rightarrow U_{\bar{X}-\bar{Y}} = \frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim N(0,1)$

(ii) $\quad \perp \begin{cases} K_X = \frac{n_1-1}{\sigma_1^2}S_X^2 & \sim & \chi^2(n_1-1) \\ K_Y = \frac{n_2-1}{\sigma_2^2}S_Y^2 & \sim & \chi^2(n_2-1) \end{cases} \Rightarrow \quad F = \frac{K_X/(n_1-1)}{K_Y/(n_2-1)} \sim F(n_1-1, n_2-1)$

$\Rightarrow$

(iii) $\quad \sigma_1^2=\sigma_2^2=\sigma^2 \Rightarrow K = K_X+K_Y = \frac{1}{\sigma^2}\underbrace{\left[(n_1-1)S_X^2+(n_2-1)S_Y^2\right]}_{=(n_1+n_2-2)S_{XY}^2} \sim \chi^2(n_1+n_2-2)$

(iv) $\quad \sigma_1^2=\sigma_2^2=\sigma^2 \Rightarrow T = \frac{U_{\bar{X}-\bar{Y}}}{\sqrt{K/(n_1+n_2-2)}} = \frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{S_{XY}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t(n_1+n_2-2)$

Kvantily některých důležitých rozdělení

Je–li F distribuční funkcí a $\alpha \in (0,1)$, pak $F^{-1}(\alpha) = Q(\alpha) = \inf\{x \in \mathbb{R} : F(x) \geq \alpha\}$ je **kvantilová funkce** a číslo $x_\alpha = Q(\alpha)$ je α-**kvantilem** rozdělení s distribuční funkcí $F(x)$.

u_α	`qnorm`	standardizované normální rozdělení
$\chi^2_\alpha(\nu)$	`qchisq`	χ^2 rozdělení o ν stupních volnosti
$t_\alpha(\nu)$	`qt`	Studentovo rozdělení o ν stupních volnosti
$F_\alpha(\nu_1,\nu_2)$	`qf`	Fisherovo-Snedecorova rozdělení o ν_1 a ν_2 stupních volnosti

Je-li distribuční funkce F absolutně spojitá a ryze monotónní a je-li příslušná hustota f **sudá funkce**, pak pro $x \in \mathbb{R}$ platí $F(x) = 1 - F(-x)$ a odtud pro $\alpha \in (0,1)$ $x_\alpha = -x_{1-\alpha}$, což speciálně platí pro **normální** a **Studentovo rozdělení**.

Intervalové odhady pro normální výběry

Intervalové odhady pro σ^2 z 1 náhodného výběru z normálního rozdělení		
oboustranný	levostranný	pravostranný
$\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}, \frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}$	$\frac{(n-1)S^2}{\chi^2_{1-\alpha}(n-1)}$	$\frac{(n-1)S^2}{\chi^2_{\alpha}(n-1)}$
Intervalové odhady pro σ_1^2/σ_2^2 ze 2 náhodných výběrů z normálního rozdělení		
oboustranný	levostranný	pravostranný
$\frac{S_X^2}{S_Y^2}\frac{1}{F_{1-\frac{\alpha}{2}}(n_1-1,n_2-1)}, \frac{S_X^2}{S_Y^2}\frac{1}{F_{\frac{\alpha}{2}}(n_1-1,n_2-1)}$	$\frac{S_X^2}{S_Y^2}\frac{1}{F_{1-\alpha}(n_1-1,n_2-1)}$	$\frac{S_X^2}{S_Y^2}\frac{1}{F_{\alpha}(n_1-1,n_2-1)}$

Intervalové odhady pro μ z 1 náhodného výběru z normálního rozdělení			
při	typ	odhad	$c(\alpha)$
známém σ^2	oboustranný	$\bar{X}-c(\alpha), \bar{X}+c(\alpha)$	$u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$
	levostranný	$\bar{X}-c(\alpha)$	$u_{1-\alpha}\frac{\sigma}{\sqrt{n}}$
	pravostranný	$\bar{X}+c(\alpha)$	$u_{1-\alpha}\frac{\sigma}{\sqrt{n}}$
neznámém σ^2	oboustranný	$\bar{X}-c(\alpha), \bar{X}+c(\alpha)$	$t_{1-\alpha/2}(n-1)\frac{S}{\sqrt{n}}$,
	levostranný	$\bar{X}-c(\alpha)$	$t_{1-\alpha}(n-1)\frac{S}{\sqrt{n}}$
	pravostranný	$\bar{X}+c(\alpha)$	$t_{1-\alpha}(n-1)\frac{S}{\sqrt{n}}$
Intervalové odhady pro $\mu_1-\mu_2$ ze 2 náhodných výběrů z normálního rozdělení			
při	typ	odhad	$c(\alpha)$
známém σ_1^2, σ_2^2	oboustranný	$\bar{X}-\bar{Y}-c(\alpha), \bar{X}-\bar{Y}+c(\alpha)$	$u_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$
	levostranný	$\bar{X}-\bar{Y}-c(\alpha)$	$u_{1-\alpha}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$
	pravostranný	$\bar{X}-\bar{Y}+c(\alpha)$	$u_{1-\alpha}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$
neznámém σ_1^2, σ_2^2	oboustranný	$\bar{X}-\bar{Y}-c(\alpha), \bar{X}-\bar{Y}+c(\alpha)$	$t_{1-\frac{\alpha}{2}}(n_1+n_2-2)\ S_{XY}\sqrt{\frac{n_1+n_2}{n_1 n_2}}$
	levostranný	$\bar{X}-\bar{Y}-c(\alpha)$	$t_{1-\alpha}(n_1+n_2-2)S_{XY}\sqrt{\frac{n_1+n_2}{n_1 n_2}}$
	pravostranný	$\bar{X}-\bar{Y}+c(\alpha)$	$t_{1-\alpha}(n_1+n_2-2)S_{XY}\sqrt{\frac{n_1+n_2}{n_1 n_2}}$

Testování hypotéz

Mějme náhodný výběr $\mathbf{X}=(X_1,\ldots,X_n)'$ rozsahu n z rozdělení, které závisí na parametru $\boldsymbol{\theta}=(\theta_1,\ldots,\theta_m)'\in\boldsymbol{\Theta}$ a parametrickou funkci $\gamma(\boldsymbol{\theta})$.

(a) Hypotéza $\boxed{H_0}:\gamma(\boldsymbol{\theta})=\gamma(\boldsymbol{\theta}_0)$ proti *oboustranné* alternativě $\boxed{H_1}:\gamma(\boldsymbol{\theta})\neq\gamma(\boldsymbol{\theta}_0)$:

Mějme **intervalový odhad** $(D_n(\mathbf{X}),H_n(\mathbf{X}))$ parametrické funkce $\gamma(\boldsymbol{\theta})$ o spolehlivosti $\boxed{1-\alpha}$. Pokud platí nulová hypotéza, pak $1-\alpha=P_{\boldsymbol{\theta}}\left(D_n(\mathbf{X})\leq\gamma(\boldsymbol{\theta}_0)\leq H_n(\mathbf{X})\right)$, takže **kritický obor** tohoto testu má tvar $W_\alpha=\{\mathbf{X}\in\mathbb{R}^n:\gamma(\boldsymbol{\theta}_0)\notin(D_n(\mathbf{X}),H_n(\mathbf{X}))\}$.

Zjistíme-li v konkrétní situaci, že $\gamma(\boldsymbol{\theta}_0)\notin(d_n(\mathbf{x}),h_n(\mathbf{x}))$ tj. realizace $\mathbf{x}\in W_\alpha$, potom

- buď nastal jev, který má pravděpodobnost α (volí se blízká nule),
- nebo neplatí nulová hypotéza.

Protože při obvyklé volbě $\alpha=0.05$ nebo $\alpha=0.01$ je tento jev „prakticky nemožný“, proto nulovou hypotézu $\boxed{H_0}$ **zamítáme ve prospěch alternativy** $\boxed{H_1}$.

V opačném případě, tj. pokud $\gamma(\boldsymbol{\theta}_0)\in(d_n(\mathbf{x}),h_n(\mathbf{x}))$ tj. realizace $\mathbf{x}\notin W_\alpha$, nulovou hypotézu $\boxed{H_0}$ **nezamítáme**.

(b) Hypotéza $\boxed{H_0}: \gamma(\boldsymbol{\theta}) = \gamma(\boldsymbol{\theta}_0)$ proti *levostranné* alternativě $\boxed{H_1}: \gamma(\boldsymbol{\theta}) > \gamma(\boldsymbol{\theta}_0)$: V tomto případě využijeme **dolní odhad** $D_n(\mathbf{X})$ parametrické funkce $\gamma(\boldsymbol{\theta})$ o spolehlivosti $\boxed{1-\alpha}$. Pokud platí nulová hypotéza, pak $1-\alpha = P_{\boldsymbol{\theta}}\left(D_n(\mathbf{X}) \leq \gamma(\boldsymbol{\theta}_0)\right)$, takže **kritický obor** tohoto testu má tvar: $W_\alpha = \{\mathbf{X} \in \mathbb{R}^n : D_n(\mathbf{X}) > \gamma(\boldsymbol{\theta}_0)\}$.

(c) Hypotéza $\boxed{H_0}: \gamma(\boldsymbol{\theta}) = \gamma(\boldsymbol{\theta}_0)$ proti *pravostranné* alternativě $\boxed{H_1}: \gamma(\boldsymbol{\theta}) < \gamma(\boldsymbol{\theta}_0)$ V tomto případě využijeme **horní odhad** $H_n(\mathbf{X})$ parametrické funkce $\gamma(\boldsymbol{\theta})$ o spolehlivosti $\boxed{1-\alpha}$. Pokud platí nulová hypotéza, pak $1-\alpha = P_{\boldsymbol{\theta}}\left(\gamma(\boldsymbol{\theta}_0) \leq H_n(\mathbf{X})\right)$, takže **kritický obor** tohoto testu má tvar: $W_\alpha = \{\mathbf{X} \in \mathbb{R}^n : H_n(\mathbf{X}) < \gamma(\boldsymbol{\theta}_0)\}$.

H_0	H_1	Hypotézu H_0 zamítáme, pomocí	
		intervalu spolehlivosti	kritické oblasti, tj. pokud $\mathbf{x} \in W_\alpha$, kde $W_\alpha =$
$\gamma(\boldsymbol{\theta}) = \gamma(\boldsymbol{\theta}_0)$	$\gamma(\boldsymbol{\theta}) \neq \gamma(\boldsymbol{\theta}_0)$	$\gamma(\boldsymbol{\theta}_0) \notin (d_n(\mathbf{x}), h_n(\mathbf{x}))$	$\{\mathbf{X} \in \mathbb{R}^n : \gamma(\boldsymbol{\theta}_0) \notin (D_n(\mathbf{X}), H_n(\mathbf{X}))\}$
$\gamma(\boldsymbol{\theta}) = \gamma(\boldsymbol{\theta}_0)$	$\gamma(\boldsymbol{\theta}) > \gamma(\boldsymbol{\theta}_0)$	$\gamma(\boldsymbol{\theta}_0) < d_n(\mathbf{x})$	$\{\mathbf{X} \in \mathbb{R}^n : D_n(\mathbf{X}) > \gamma(\boldsymbol{\theta}_0)\}$
$\gamma(\boldsymbol{\theta}) = \gamma(\boldsymbol{\theta}_0)$	$\gamma(\boldsymbol{\theta}) < \gamma(\boldsymbol{\theta}_0)$	$\gamma(\boldsymbol{\theta}_0) > h_n(\mathbf{x})$	$\{\mathbf{X} \in \mathbb{R}^n : H_n(\mathbf{X}) < \gamma(\boldsymbol{\theta}_0)\}$

P–HODNOTA

Nebo taky *p–value* je pravděpodobnost, že získáme stejné nebo vyšší kritérium než vypočítané za předpokladu, že platí H_0.

Pro testování nulové hypotézy

- musíme mít vždy odvozenu vhodnou statistiku $T_n = T(X_1, \ldots, X_n)$
- musíme znát její rozdělení (popř. alespoň její asymptotické rozdělení) za platnosti nulové hypotézy.

Předpokládejme, že pro konkrétní realizaci náhodného výběru $\mathbf{x} = (x_1, \ldots, x_n)$ tato statistika nabyla hodnoty $t_n = T_n(\mathbf{x})$, pak p–hodnota je rovna

$$p\text{-}value = P(T_n(\mathbf{X}) \geq t_n | H_0)$$

(a) V případě spojitého rozdělení

$$p\text{-}value = P(T_n(\mathbf{X}) \geq t_n | H_0) = 1 - F_{H_0}(t_n)$$

(b) V případě diskrétního rozdělení

$$p\text{-}value = P(T_n(\mathbf{X}) \geq t_n | H_0) = 1 - F_{H_0}(t_n) + \lim_{y \to t_n^-} F_{H_0}(y)$$

B. Jeden náhodný výběr

Příklad 1: Posouzeni výsledků Amthaureova IQ testu

Vrátíme se k příkladu z předchozího cvičení, který se týkal IQ testu pro 98 studentů. Postupně načteme popisný a datový soubor.

```
> fileTxt <- paste(data.library, "s204.inf", sep = "")
> con <- file(fileTxt)
> (popis <- readLines(con))
```

```
[1] "S204: Posouzeni vysledku Amthaureova testu u 98 studentu"
[2] "[1] skore Amthaureova testu"
[3] "V ramci prijimaciho rizeni absolvuji uchazeci o studium na vysoke skole "
[4] "Amthaueruv test struktury inteligence. Vysledky tohoto testu se vyjadruji"
[5] "prostrednictvim  tzv. hrubeho skore. Ze studentu prijatych ke studiu"
[6] "behem 4 let byl proveden nahodny vyber 98 studentu."
[7] ""
```

```
> close(con)
> fileDat <- paste(data.library, "s204.txt", sep = "")
> skore <- scan(fileDat)
> str(skore)
```

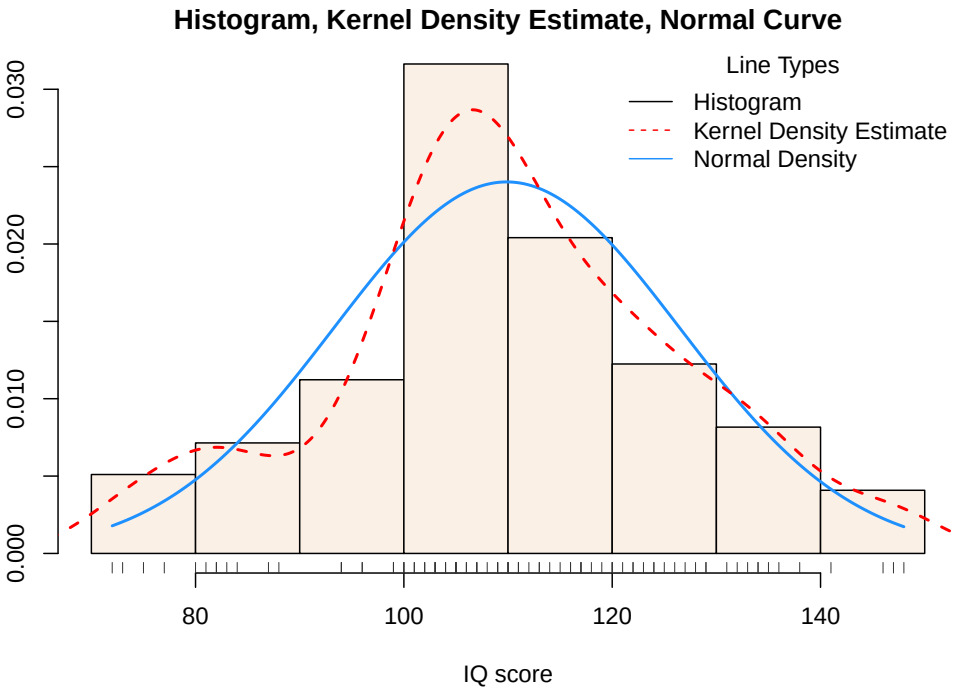
```
 num [1:98] 77 105 110 88 128 104 94 104 129 96 ...
```

Chceme testovat hypotézu, že průměrné skóre je 110 bodů. Musíme si stanovit nulovou a alternativní hypotézu:

$$H_0: \ \mu = 110 \quad vs \quad H_1: \ \mu \neq 110$$

Chceme-li použít t–test, měli bychom ověřit normalitu. Nejprve to rychle provedeme graficky pomocí histogramu, jádrových odhadů a také pomocí kvantilového Q–Q grafu.

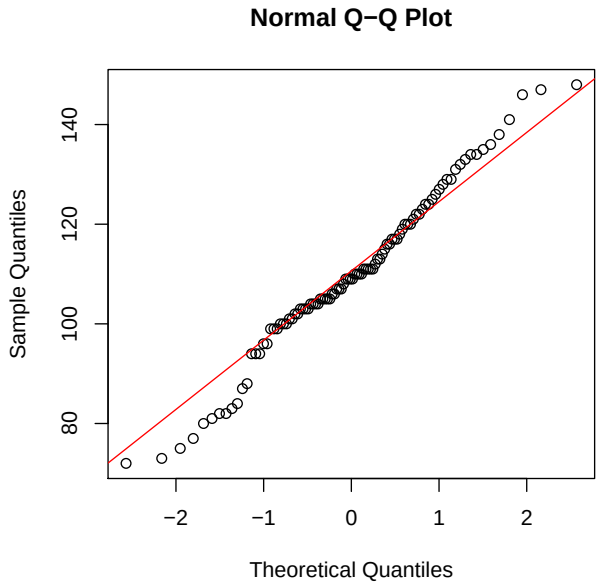
```
> options(width = 70)
> x <- skore
> par(mar = c(4, 2, 1, 0) + 0.75)
> h <- hist(x, probability = TRUE, breaks = "FD", col = "linen",
    xlab = "IQ score", main = "Histogram, Kernel Density Estimate, Normal Curve")
> xfit <- seq(min(x), max(x), length = 512)
> yfit <- dnorm(xfit, mean = mean(x), sd = sd(x))
> lines(xfit, yfit, col = "dodgerblue", lwd = 2)
> lines(density(x, n = 512), lwd = 2, col = "red", lty = 2)
> rug(x, side = 1, ticksize = 0.02, col = "grey20")
> legend("topright", legend = c("Histogram", "Kernel Density Estimate",
    "Normal Density"), col = c("black", "red", "dodgerblue"),
    lty = c(1, 2, 1), bty = "n", title = "Line Types")
```



Obrázek 1: Histogram, jádrový odhad hustoty, normální hustota pro hodnoty Amthauerova IQ testu.

Normalitu pomocí Q–Q grafu ověříme napsáním příkazů

```
> qqnorm(skore)
> qqline(skore, col = "red")
```



Obrázek 2: Ověření normality pro hodnoty Amthauerova IQ testu.

Předchozí dva grafy nesvědčí o tom, že by mezi předpokládaným normálním rozdělením a odhadnutým rozdělením byla ideální shoda.

Proto se budeme snažit provést testování normality na základě nějakých vhodných testů. V tom případě máme k dispozici více možností.

Jedna skupina testů je založena na empirických distribučních funkcích, jako zástupce můžeme uvést Kolmogorův–Smirnovův test, popř. Shapiro–Wilksův test.

Další testy jsou založeny na momentových charakteristikách, především na šikmosti či špičatosti. Příkladem může být d'Agostinův test.

V základním balíku R-base najdeme dva známé testy normality: Shapiro-Wilkův test a Kolmogorův-Smirnovův test. Podíváme se na ten první.

Shapiro–Wilkův test pro testování normality

Shapiro–Wilkův test je založen na statistice

$$W = \frac{\left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2},$$

kde $X_{(i)}$ jsou pořádkové statistiky a a_i jsou váhy, které jsou odvozeny ze středních hodnot a varianční matice pořádkových statistik prostého náhodného výběru z $N(0,1)$ rozsahu n. Tyto hodnoty bývají tabelovány.

Na testovou statistiku W lze pohlížet jako na korelaci mezi pozorovanými hodnotami a jejich normálními skóry.

Testová statistika dosahuje hodnoty 1 v případě, že data vykazují perfektní shodu s normálních rozdělením. Je–li W statisticky významně nižší než 1, zamítáme nulovou hypotézu o shodě s normálním rozdělením.

```
> x <- skore
> shapiro.test(x)


        Shapiro-Wilk normality test

data:  x
W = 0.9834, p-value = 0.2536
```

Z výsledků Shapiro–Wilkova testu je patrné, že nelze zamítnout hypotézu, že náhodný výběr pochází z normálního rozdělení, protože p–hodnota není menší než 0.05. Také hodnota statistiky W je velmi blízká k jedničce. Toto zjištění nás opravňuje k použití klasického t–testu.

t-test v prostředí R

Pomocí příkazu `t.test()` vhodnou volbou parametrů lze provádět

- testování střední hodnoty jednoho výběru (zadáním parametru `mu`);
- testování shody středních hodnot dvou výběrů (zadáním x i y);
- testování párových výběrů (zadáním `paired = TRUE`);
- testy mohou být oboustranné, levostranné či pravostranné `alternative = c("two.sided", "less", "greater")`;
- procedura se dokáže vyrovnat i s nestejnými rozptyly: `var.equal = FALSE` (implicitně nastaveno);
- implicitní hladina významnosti `conf.level = 0.95`.

Nyní nastavíme vhodně parametry a provedeme testování

$$H_0: \ \mu = 110 \quad vs \quad H_1: \ \mu \neq 110$$

```
> x <- skore
> mu0 <- 110
> t.test(x, mu = mu0)
```

```
        One Sample t-test

data:  x
t = -0.073, df = 97, p-value = 0.942
alternative hypothesis: true mean is not equal to 110
95 percent confidence interval:
 106.5465 113.2086
sample estimates:
mean of x
 109.8776
```

Na základě

- p-hodnoty (není menší než 0.05)
- či 95% intervalu spolehlivosti (obsahuje testovanou hodnotu 110)

můžeme konstatovat, že pomocí t–textu nelze zamítnout tvrzení, že průměrná hodnota IQ je 110 bodů.

Nebo to můžeme vyslovit i tak, že data nejsou v rozporu s tvrzením nulové hypotézy.

Uvažovanou hypotézu můžeme testovat i neparametricky. Opět máme k dispozici celou řadu testů. Jedním z nich je Wilcoxonův test nebo znaménkový test.

Jednovýběrový Wilcoxonův test

Předpokládejme, že máme náhodný výběr ze spojitého rozdělení s hustotou f, která je symetrická kolem bodu a. Platí tedy

$$f(a+x) = f(a-x).$$

Z toho plyne, že $\boxed{a}$ musí být rovno mediánu $\boxed{x_{med}}$. Jednovýběrový Wilcoxonův test je tedy určen k testování hypotézy

$$H_0: \; x_{med} = x_0 \quad vs \quad H_1: \; x_{med} \neq x_0$$

Nejprve předpokládejme, že žádná z veličin X_i není rovna x_0. Položme

$$Y_i = X_i - x_0.$$

Veličiny Y_i seřadíme do neklesající posloupnosti podle jejich absolutní hodnoty

$$|Y|_{(1)} \leq |Y|_{(2)} \leq \ldots \leq |Y|_{(n)}.$$

Označme $\boxed{R_i^+}$ pořadí veličiny $|Y_i|$. Zaveďme dále značení

$$S^+ = \sum_{Y_i \geq 0} R_i^+ \quad \text{a} \quad S^- = \sum_{Y_i < 0} R_i^+.$$

Zřejmě platí

$$S^+ + S^- = \frac{1}{2} n(n+1).$$

Je–li číslo $\boxed{min(S^+, S^-)}$ menší nebo rovno tabelované kritické hodnotě $\boxed{w_n(\alpha)}$, pak **zamítáme nulovou hypotézu**.

Dá se také ukázat, že statistika $\boxed{S^+}$ má asymptoticky normální rozdělení, takže testování nulové hypotézy lze rovněž založit na veličině

$$U = \frac{S^+ - ES^+}{\sqrt{DS^+}} \overset{A}{\sim} N(0,1),$$

kde

$$ES^+ = \frac{1}{4} n(n+1) \quad \text{a} \quad DS^+ = \frac{1}{24} n(n+1)(2n+1).$$

Jestliže

$$|U| \geq u_{\frac{\alpha}{2}} \quad \Rightarrow \quad \text{zamítáme nulovou hypotézu.}$$

Je třeba zdůraznit, že jedním z předpokladů jednovýběrového Wilcoxova testu je i symetrie hustoty kolem mediánu. K zamítnutí nulové hypotézy může tedy oprávněně dojít i tehdy, je–li medián roven x_0, ale hustota je výrazně nesymetrická.

Je–li některá z veličin X_i rovna x_0, zpravidla se toto pozorování vynechává.

Vrátíme se k našemu příkladu a provedeme pomocí Wilcoxonova testu testování nulové hypotézy, že medián IQ skórů je roven 110 bodům. Parametry u testu `wilcox.test()` jsou analogické jako u testu `t.test()`.

```
> x <- skore
> mu0 <- 110
> wilcox.test(x, mu = mu0)


        Wilcoxon signed rank test with continuity correction

data:  x
V = 2205, p-value = 0.9189
alternative hypothesis: true location is not equal to 110
```

Z výstupu je zřejmé, že stejně jako u t–testu, p–hodnota se blíží k jedničce, takže nás nic neopravňuje zamítnout nulovou hypotézu.

Dalším důležitým neparametrickým testem je velmi jednoduchý, tzv. znaménkový test.

Znaménkový test

Tento test se opět vztahuje k testování mediánu a předpokládá pouze spojitou hustotu. Opět utvoříme rozdíly

$$X_1 - x_0, \ldots, X_n - x_0.$$

Pokud je některý z rozdílů nulový, zpravidla se vynechává. Počet rozdílů s kladným znaménkem označíme $\boxed{Y}$. Platí–li nulová hypotéza, pak tato náhodná veličina má binomické rozdělení

$$Y \sim Bi(n, \pi), \quad \text{kde} \quad \pi = \frac{1}{2}.$$

Nulovou hypotézu zamítneme, bude–li Y blízko nule nebo blízko číslu n.

Nejprve vytvoříme funkci `sign.test`, kterou později použijeme. Budeme se snažit přizpůsobit jména parametrů zvyklostem z předchozích testů. Uvnitř těla funkce použije `binom.test`, který je exaktním testem pro testování nulové hypotézy, že parametr π binomického rozdělení je roven nějaké konkrétní hodnotě π_0.

```
> sign.test <- function(x, y = NULL, mu = 0, alternative = c("two.sided",
    "less", "greater"), conf.level = 0.95) {
    if (is.null(y))
        d <- x - mu
    else d <- x - y
    binom.test(sum(d > 0), length(d), p = 0.5, alternative = alternative,
        conf.level = conf.level)
}
```

Znaménkový test je dobré používat především v případě, kdy rozdělení testovaných dat je výrazně zešikmené.

Použijeme právě vytvořený znaménkový test pro testování nulové hypotézy, že medián IQ skóre je roven 110 bodům.

```
> x <- skore
> mu0 <- 110
> sign.test(x, mu = mu0)
```

```
        Exact binomial test

data:  sum(d > 0) and length(d)
number of successes = 44, number of trials = 98, p-value =
0.3634
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.3483325 0.5527931
sample estimates:
probability of success
             0.4489796
```

Z výsledků znaménkového testu je patrné, že

- p-hodnota není menší než 0.05,
- 95% interval spolehlivosti obsahuje testovanou hodnotu $\pi_0 = \frac{1}{2}$,

takže i tímto testem nezamítáme nulovou hypotézu.

Závěrem bychom tedy mohli říci, že na základě všech parametrických i neparametrických testů lze konstatovat, že získané hodnoty IQ testu nejsou v rozporu s tvrzením, že jeho průměrná hodnota je rovna 110 bodům.

C. Párové náhodné výběry

Příklad 2: Hladina penicilinu v séru pacientů po 50 a 90 minutách aplikace

Nejprve načteme popisný a datový soubor.

```
> fileTxt <- paste(data.library, "b2089.inf", sep = "")
> con <- file(fileTxt)
> (popis <- readLines(con))
```

```
[1] "B208,209: Hladina penicilinu v seru pacientu po 50 a 90 minutach aplikace"
[2] "[1] hladina penicilinu (mg/l) po 50 minutach aplikace"
[3] "[2] hladina penicilinu (mg/l) po 90 minutach aplikace"
[4] "Pri studii biologicke dostupnosti leku byla stanovena hladina koncentrace"
[5] "penicilinu v seru zdravych dobrovolniku vysokotlakou kapalinovou "
[6] "chromatografii."
[7] ""
```

```
> close(con)
> fileDat <- paste(data.library, "b2089.txt", sep = "")
> HladinaPenicilinu <- read.table(fileDat)
> str(HladinaPenicilinu)
```

```
'data.frame':        30 obs. of  2 variables:
 $ V1: num  2.1 0.9 1.98 1.89 1.11 ...
 $ V2: num  0.732 0.732 0.712 0.753 0.654 0.72 0.701 0.762 0.77 0.704 ...
```

```
> x <- HladinaPenicilinu$V1
> y <- HladinaPenicilinu$V2
```

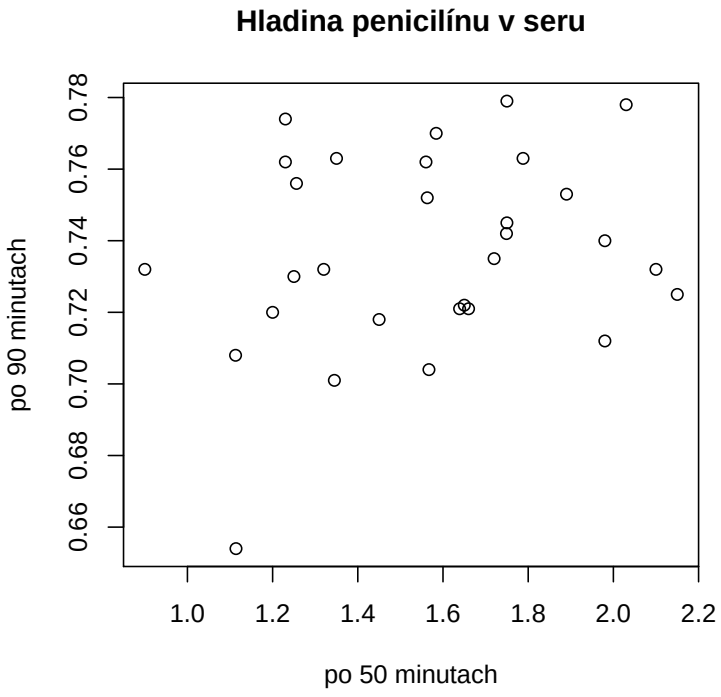
Protože z párových náhodných výběrů rozdílem $X_i - Y_i$ prakticky dostanem jediný náhodný výběr, tak k testování nulové hypotézy

$$H_0: \ \mu_1 = \mu_2 \quad vs \quad H_1: \ \mu_1 \neq \mu_2$$

použijeme tytéž testy, jako v případě jediného náhodného výběru.

Nejprve však data vykreslíme

```
> plot(x, y, main = "Hladina penicilínu v seru", xlab = "po 50 minutach",
     ylab = "po 90 minutach")
```



Obrázek 3: Hladiny penicilínu v séru po 50 a 90 minutách.

Již z grafu je patrné, že hypotézu nejspíše zamítneme, protože se hodnoty rozhodně nenacházejí kolem přímky $x = y$.

Pokud budeme chtít použít t–test, je potřeba také otestovat normalitu. Správně bychom měli otestovat, že jde o dvourozměrné normální rozdělení. Ale spokojíme se se dvěma testy, jednou pro hodnoty hladiny penicilínu po 50 minutách, podruhé pro hodnoty hladiny penicilínu po 90 minutách.

```
> shapiro.test(x)

        Shapiro-Wilk normality test

data:  x
W = 0.9726, p-value = 0.6128


> shapiro.test(y)

        Shapiro-Wilk normality test

data:  y
W = 0.9478, p-value = 0.1473
```

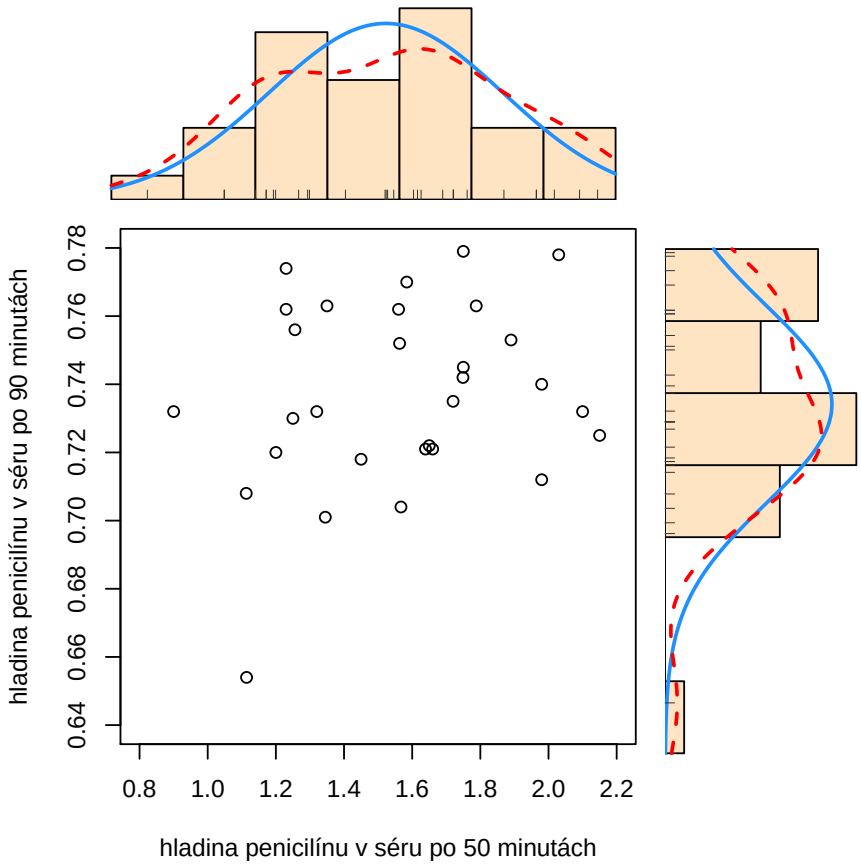
Protože p-hodnoty ani v jednom případě nejsou menší než 0.05 a také hodnoty statistiky W jsou docela blízké k 1, nulovou hypotézu o normalitě nemůžeme zamítnout.

Pokusíme se ještě graficky posoudit normalitu. Proto vytvoříme scatter plot obou proměnných a po boku dokreslíme jejich histogram, jádrový odhad hustoty a křivku normální hustoty založené na výběrových průměrech a výběrových rozptylech.

Všimněte si také, jakým způsobem bylo nutno se vyrovnat s faktem, že graf `barplot()` na ose x je v rozmezí od nuly do počtu subintervalů.

```
> x <- HladinaPenicilinu$V1
> y <- HladinaPenicilinu$V2
> xhist <- hist(x, breaks = "FD", plot = FALSE)
> yhist <- hist(y, breaks = "FD", plot = FALSE)
> nHx <- length(xhist$breaks)
> nHy <- length(yhist$breaks)
> xr <- c(xhist$breaks[1], xhist$breaks[nHx])
> yr <- c(yhist$breaks[1], yhist$breaks[nHy])
> nD <- 200
> xdens <- density(x, n = nD, from = xr[1], to = xr[2])
> ydens <- density(y, n = nD, from = yr[1], to = yr[2])
> nxfit <- dnorm(xdens$x, mean = mean(x), sd = sd(x))
> nyfit <- dnorm(ydens$x, mean = mean(y), sd = sd(y))
> topx <- max(c(xhist$density, nxfit, xdens$y))
> topy <- max(c(yhist$density, nyfit, ydens$y))
> trX <- (nHx - 1) * (xdens$x - xr[1])/diff(xr)
> trY <- (nHy - 1) * (ydens$x - yr[1])/diff(yr)
```

```
> trx <- (nHx - 1) * (x - xr[1])/diff(xr)
> try <- (nHy - 1) * (y - yr[1])/diff(yr)
> def.par <- par(no.readonly = TRUE)
> layout(matrix(c(2, 0, 1, 3), 2, 2, byrow = TRUE), width = c(3,
    1), height = c(1, 3), respect = TRUE)
> par(mar = c(4, 4, 1, 1))
> plot(x, y, xlim = xr, ylim = yr, xlab = "hladina penicilínu v séru po 50 minutách",
    ylab = "hladina penicilínu v séru po 90 minutách")
> par(mar = c(0, 3, 1, 1))
> barplot(xhist$density, axes = FALSE, ylim = c(0, topx),
    space = 0, col = "bisque")
> lines(trX, nxfit, col = "dodgerblue", lwd = 2)
> lines(trX, xdens$y, lwd = 2, col = "red", lty = 2)
> rug(trx, side = 1, col = "grey20", ticksize = 0.05)
> par(mar = c(3, 0, 1, 1))
> barplot(yhist$density, axes = FALSE, xlim = c(0, topy),
    space = 0, horiz = TRUE, col = "bisque")
> lines(nyfit, trY, col = "dodgerblue", lwd = 2)
> lines(ydens$y, trY, lwd = 2, col = "red", lty = 2)
> rug(try, side = 2, col = "grey20", ticksize = 0.05)
> par(def.par)
```



Obrázek 4: Scatter plot a marginální histogramy pro hladiny penicilínu v séru po 50 a 90 minutách.

Po grafickém posouzení normality postupně provedeme nejprve dva neparametrické testy a pak t–test.

```
> sign.test(x, y)


        Exact binomial test

data:  sum(d > 0) and length(d)
number of successes = 30, number of trials = 30, p-value =
1.863e-09
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.8842967 1.0000000
sample estimates:
probability of success
                     1


> wilcox.test(x, y, paired = TRUE)


        Wilcoxon signed rank test

data:  x and y
V = 465, p-value = 1.863e-09
alternative hypothesis: true location shift is not equal to 0


> t.test(x, y, paired = TRUE)


        Paired t-test

data:  x and y
t = 14.1341, df = 29, p-value = 1.541e-14
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.7059607 0.9448326
sample estimates:
mean of the differences
              0.8253967
```

Z výsledků je ihned vidět, že vždy zamítneme nulovou hypotézu o shodě středních hodnot, nebot'

`sign.test`
- p-hodnota $<$ 0.05
- hodnota $\pi_0 = \frac{1}{2}$ neleží uvnitř 95% intervalu spolehlivosti

`wilcox.test`
- p-hodnota $<$ 0.05

`t.test`
- p-hodnota $<$ 0.05
- nula neleží uvnitř 95% intervalu spolehlivosti.

D. Dva nezávislé náhodné výběry

Příklad 3: Přesnost balícího automatu před a po seřízení

Nejprve načteme popisný a datový soubor.

```
> fileTxt <- paste(data.library, "automat2.inf", sep = "")
> con <- file(fileTxt)
> (popis <- readLines(con))
```

```
[1] "Presnost baliciho automatu pred a po serizeni"
[2] "Vazenim se ziskali udaje o presnem mnozstvi potravinarskych vyrobku"
[3] "urciteho druhu, automaticky balenych u vyrobku nahodne vybranych"
[4] "pred a po serizeni baliciho automatu. Na 5% hladine vyznamnosti"
[5] "prokazte, ze"
[6] "a) pred serizenim automatu stredni hodnota prekracuje 250g a "
[7] "   smerodatna odchylka prekracuje 1 g"
[8] "b) kolisavost mnozstvi se serizenim automatu snizila"
[9] "c) stredni hodnota se serizenin automatu zmenila."
```

```
> close(con)
> fileDat <- paste(data.library, "automat2.txt", sep = "")
> x <- scan(fileDat, nlines = 1)
> y <- scan(fileDat, skip = 1, nlines = 1)
```

Celkovou charakteristiku dat získáme příkazem `summary()`

```
> summary(x)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  243.2   248.2   251.3   250.3   252.8   254.0
```

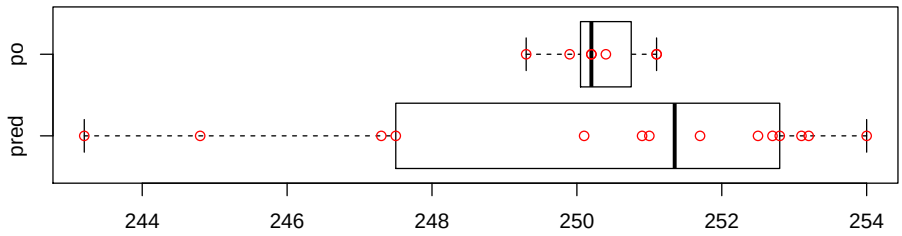
```
> summary(y)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  249.3   250.1   250.2   250.3   250.8   251.1
```

Všimněme si, že oba dva výběry mají (po zaokrouhlení) stejné výběrové průměry.

Lepší představu o datech však získáme graficky. Proto nejprve vykreslíme krabicové grafy. Abychom mohli oba dva boxploty zakreslit do jediného grafu, vytvoříme nové proměnné.

```
> xy <- c(x, y)
> id <- c(rep(1, length(x)), rep(2, length(y)))
> idf <- factor(id, labels = c("pred", "po"))
> boxplot(xy ~ idf, horizontal = TRUE)
> points(x, rep(1, length(x)), col = "red")
> points(y, rep(2, length(y)), col = "red")
```

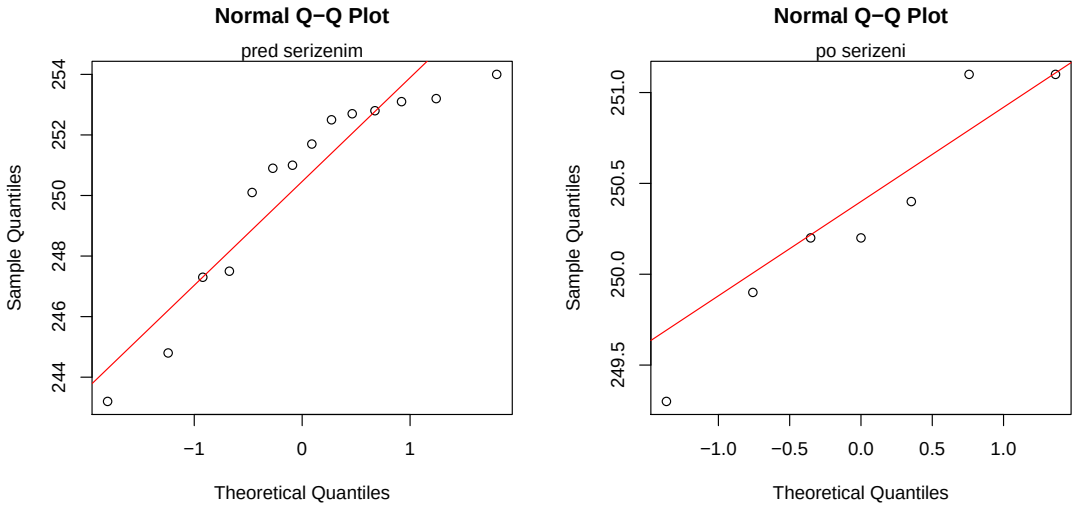


Obrázek 5: Krabicové grafy pro data měřící přesnost balícího automatu před a po seřízení.

Podle krabicových grafů také výběrové mediány se liší. Výrazně je rozdílná i variabilita dat.

Normalitu nejdříve posoudíme graficky, následně pomocí testů.

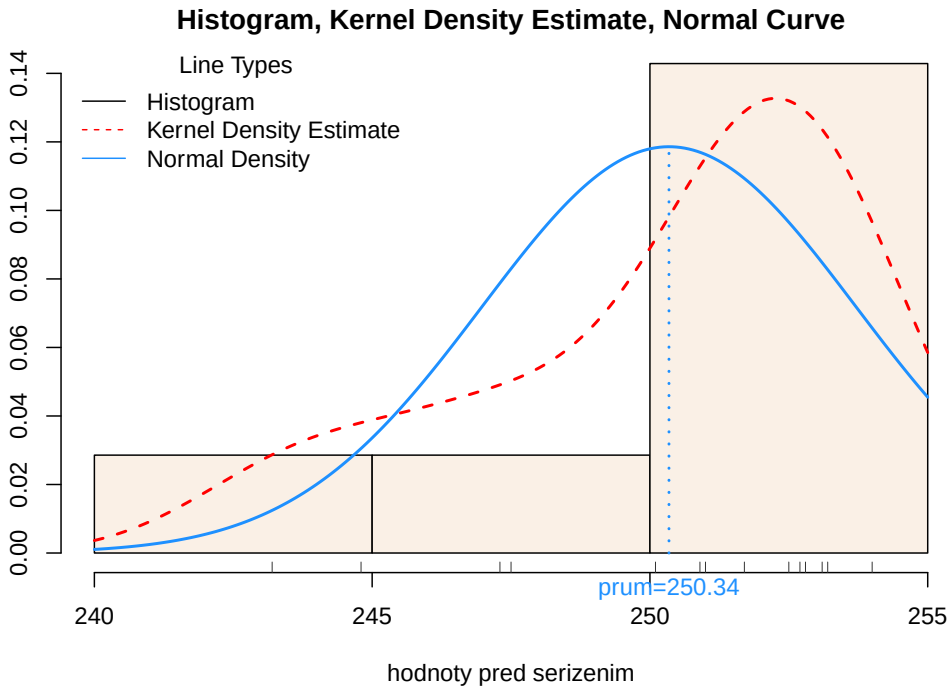
```
> par(mfrow = c(1, 2))
> qqnorm(x)
> qqline(x, col = "red")
> mtext("pred serizenim")
> qqnorm(y)
> qqline(y, col = "red")
> mtext("po serizeni")
```



Obrázek 6: Ověření normality pro data měřící přesnost balícího automatu před a po seřízení.

```
> options(width = 70)
> par(mar = c(4, 2, 1, 0) + 0.75)
> h <- hist(x, probability = TRUE, breaks = "FD", col = "linen",
      xlab = "hodnoty pred serizenim", main = "Histogram, Kernel Density Estimate, Normal Curve")
```

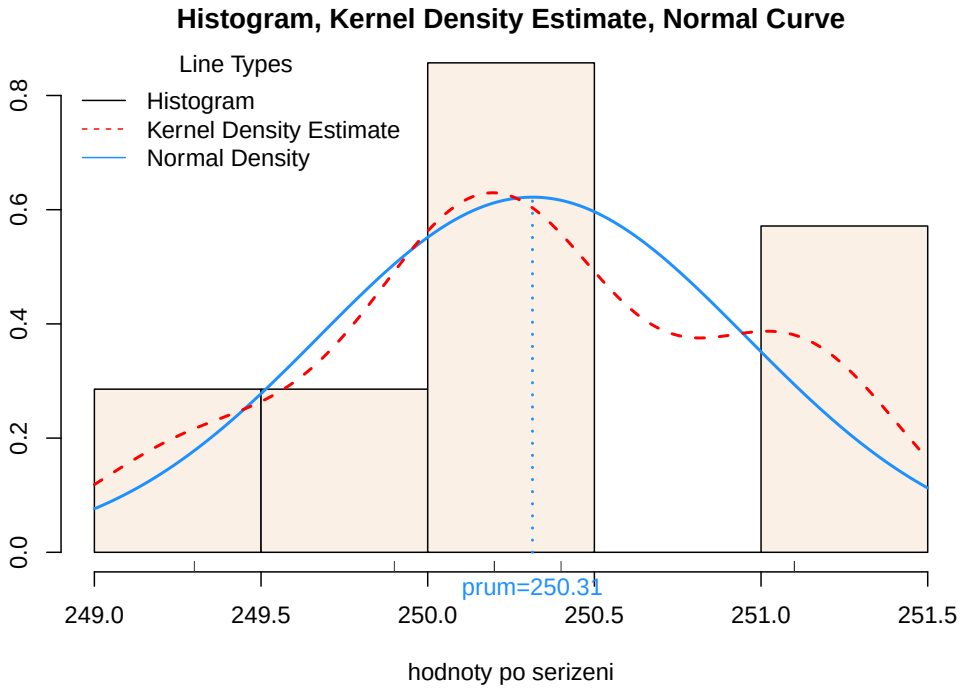
```
> nH <- length(h$breaks)
> nD <- 200
> mx <- mean(x)
> sx <- sd(x)
> mxn <- dnorm(mx, mean = mx, sd = sx)
> xfit <- seq(h$breaks[1], h$breaks[nH], length = nD)
> yfit <- dnorm(xfit, mean = mx, sd = sx)
> lines(xfit, yfit, col = "dodgerblue", lwd = 2)
> lines(density(x, n = nD, from = xfit[1], to = xfit[nD]),
    lwd = 2, col = "red", lty = 2)
> rug(x, side = 1, ticksize = 0.02, col = "grey20")
> legend("topleft", legend = c("Histogram", "Kernel Density Estimate",
    "Normal Density"), col = c("black", "red", "dodgerblue"),
    lty = c(1, 2, 1), bty = "n", title = "Line Types")
> mtext(paste("prum=", round(mx, 2), sep = ""), at = mx,
    side = 1, line = 0, col = "dodgerblue")
> lines(rep(mx, 2), c(0, mxn), lty = 3, col = "dodgerblue",
    lwd = 2)
```



Obrázek 7: Histogram, jádrový odhad hustoty, normální hustota pro data měřící přesnost balícího automatu před seřízením.

```
> options(width = 70)
> par(mar = c(4, 2, 1, 0) + 0.75)
> h <- hist(y, probability = TRUE, breaks = "FD", col = "linen",
    xlab = "hodnoty po serizeni", main = "Histogram, Kernel Density Estimate, Normal Curve")
> nH <- length(h$breaks)
> nD <- 200
> mx <- mean(y)
> sx <- sd(y)
> mxn <- dnorm(mx, mean = mx, sd = sx)
> xfit <- seq(h$breaks[1], h$breaks[nH], length = nD)
```

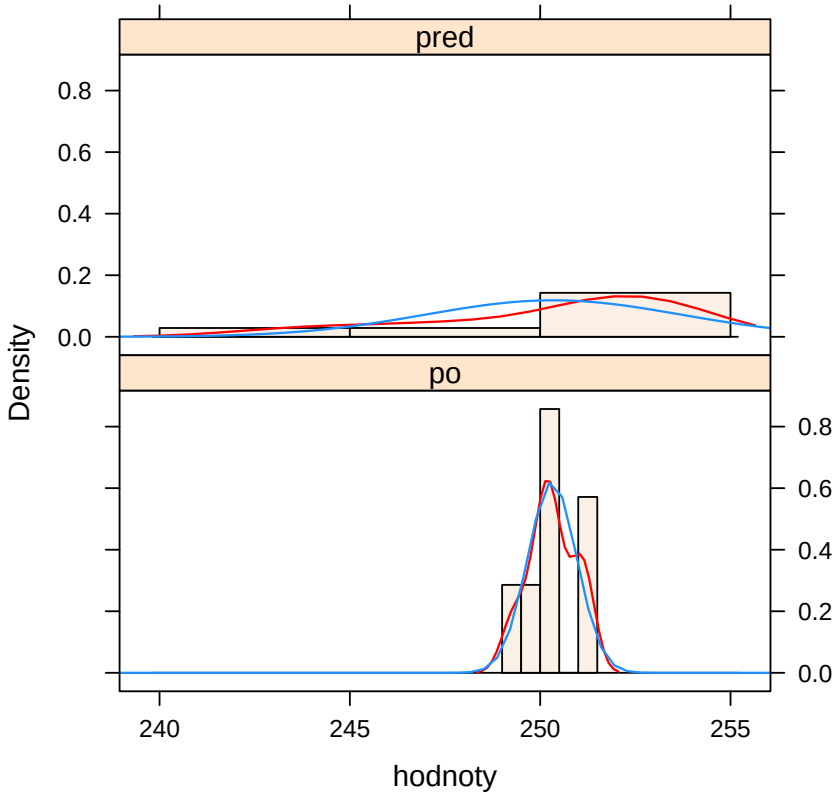
```
> yfit <- dnorm(xfit, mean = mx, sd = sx)
> lines(xfit, yfit, col = "dodgerblue", lwd = 2)
> lines(density(y, n = nD, from = xfit[1], to = xfit[nD]),
    lwd = 2, col = "red", lty = 2)
> rug(y, side = 1, ticksize = 0.02, col = "grey20")
> legend("topleft", legend = c("Histogram", "Kernel Density Estimate",
    "Normal Density"), col = c("black", "red", "dodgerblue"),
    lty = c(1, 2, 1), bty = "n", title = "Line Types")
> mtext(paste("prum=", round(mx, 2), sep = ""), at = mx,
    side = 1, line = 0, col = "dodgerblue")
> lines(rep(mx, 2), c(0, mxn), lty = 3, col = "dodgerblue",
    lwd = 2)
```



Obrázek 8: Histogram, jádrový odhad hustoty, normální hustota pro data měřící přesnost balícího automatu po seřízení.

Pomocí grafu `histogram()` z knihovny `lattice` ukážeme obě rozdělení ve stejném měřítku.

```
> library(lattice)
> data <- data.frame(hodnoty = xy, mereni = idf)
> print(histogram(~hodnoty | mereni, data, as.table = TRUE,
    layout = c(1, 2), breaks = "FD", type = "density",
    panel = function(x, ...) {
        panel.histogram(x, col = "linen", ...)
        panel.densityplot(x, col = "red", lty = 1, lwd = 1.25,
            ...)
        panel.mathdensity(dmath = dnorm, col = "dodgerblue",
            lty = 1, lwd = 1.25, args = list(mean = mean(x),
                sd = sd(x)))
    }))
```



Obrázek 9: Grafické posouzení normality pro data měřící přesnost balícího automatu před a po seřízení.

Z přechozích grafů jasně vidíme, že jde o rozdílná rozdělení a především měření před seřízením mají rozdělení velmi asymetrické.

Normalitu nakonec otestujeme pomocí Shapiro–Wilkova testu.

```
> shapiro.test(x)

        Shapiro-Wilk normality test

data:  x
W = 0.8673, p-value = 0.03841

> shapiro.test(y)

        Shapiro-Wilk normality test

data:  y
W = 0.9279, p-value = 0.5329
```

Přesně podle očekávání, normalita byla zamítnuta u měřeních před seřízením. Proto k testování shody obou výběrů použijeme pouze neparametrické testy. (Ale je třeba říci, že těžko dělat úsudek na základě tak malých výběrů.)

Wilcoxonův dvouvýběrový test

Princip testu spočívá v tom, že z náhodných výběrů $X_1, \ldots, X_{n_1}$ a $Y_1, \ldots, Y_{n_2}$ vytvoříme sdružený výběr, který vzestupně uspořádáme. Takto seřazeným hodnotám přiřadíme pořadí. Pokud se neliší jejich rozdělení, pak budou mít i shodné průměrné pořadí.

Symbolem T_1 a T_2 označme součet pořadí prvního a druhého výběru. Zřejmě platí

$$T_1 + T_2 = \frac{1}{2}(n_1 + n_2)(n_1 + n_2 + 1)$$

K testování se obvykle používá statistika

$$U_1 = n_1 n_2 + \frac{1}{2} n_1 (n_1 + 1) - T_1$$

Testu založeném na U_1 se říká Mannův–Whitneův test.

Položme dále

$$U_2 = n_1 n_2 + \frac{1}{2} n_2 (n_2 + 1) - T_2, \qquad \text{přitom platí} \quad U_1 + U_2 = n_1 n_2.$$

Pokud $\boxed{min(U_1, U_2)}$ je menší nebo rovno tabelované kritické hodnotě, zamítá se nulová hypotéza

$$H_0 : \; F_X = F_Y \quad vs \quad H_1 : \; F_X \neq F_Y.$$

Dá se také ukázat, že statistika $\boxed{U_1}$ má asymptoticky normální rozdělení, takže testování nulové hypotézy lze rovněž založit na veličině

$$U_{MW} = \frac{U_1 - EU_1}{\sqrt{DU_1}} \overset{A}{\sim} N(0,1),$$

kde

$$EU_1 = \frac{1}{2} n_1 n_2 \quad \text{a} \quad DU_1 = \frac{1}{12} n_1 n_2 (n_1 + n_2 + 1).$$

Jestliže

$$|U| \geq u_{\frac{\alpha}{2}} \quad \Rightarrow \quad \text{zamítáme nulovou hypotézu.}$$

I když je Wilcoxonův test formulován jako test proti obecné alternativě, je citlivý zejména na tzv. alternativu posunutí

$$H_1^* : \; F_X(x) = F_Y(x - \Delta), \quad \text{kde } \Delta \neq 0.$$

Pro případné jiné alternativy, např. když se F_X liší od F_Y spíše rozptylem nebo tvarem, se raději doporučuje dvouvýběrový Kolmogorův–Smirnovův test.

Nyní na naše data, která se týkají měření před a po seřízení stroje, aplikujme dvouvýběrový Wilcoxonův test.

```
> wilcox.test(x, y)
```

```
        Wilcoxon rank sum test with continuity correction

data:  x and y
W = 61, p-value = 0.3906
alternative hypothesis: true location shift is not equal to 0
```

Protože p–hodnota není menší než 0.05, pomocí dvouvýběrového Wilcoxonova testu se nám nepodařilo zamítnout hypotézu o shodě rozdělení.

Kolmogorův–Smirnovův test

Tento test je založen na výběrových (empirických) distribučních funkcích. Označme výběrovou distribuční funkci náhodného výběru $X_1, \ldots, X_{n_1}$ symbolem F_{n_1} a obdobně výběrovou distribuční funkci náhodného výběru $Y_1, \ldots, Y_{n_2}$ symbolem F_{n_2}.

Kolmogorův–Smirnovův test je založen na statistice

$$D_{n_1,n_2} = \sup_x |F_{n_1}(x) - F_{n_2}(x)|$$

Za platnosti nulové hypotézy $D_{n_1,n_2} \to 0$ skoro jistě při $n_1 \to \infty$, $n_2 \to \infty$. Kritické hodnoty tohoto testu bývají tabelovány.

Nyní na naše data, která se týkají měření před a po seřízení stroje, aplikujme dvouvýběrový Kolmogorův–Smirnovův test.

```
> ks.test(x, y)
```

```
        Two-sample Kolmogorov-Smirnov test

data:  x and y
D = 0.5, p-value = 0.1938
alternative hypothesis: two-sided
```

Protože p–hodnota není menší než 0.05, pomocí dvouvýběrového Kolmogorova–Smirnova testu se nám nepodařilo zamítnout hypotézu o shodě rozdělení.

Poznámka

Pokud bychom použili k testování shody polohy klasický dvouvýběrový t–test, bylo by dobré nejprve ověřit shodu rozptylů pomocí testu `var.test` (testuje $H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1$ pomocí F–testu), i když funkce `t.test` implicitně počítá s nestejnými rozptyly.

E. Úkol:

1. ODHAD MINUTY
   (a) Načtěte soubor informací `signal.inf` a dat `signal.txt`. Prohledněte si oba soubory.
   (b) Ověřte graficky i pomocí testu normalitu.
   (c) Testujte hypotézu, že polovina osob délku jedné minuty podhodnotí a polovina nadhodnotí.
2. DVA ZPŮSOBY HNOJENI
   (a) Načtěte soubor informací `hnojeni.inf` a dat `hnojeni.txt`. Prohledněte si oba soubory.
   (b) Ověřte graficky i pomocí testu normalitu.
   (c) Testujte hypotézu, že způsob hnojení nemá vliv na výnos pšenice.