Ověření předpokladů o datech

Určení minimální velikosti výběru

Rozsah výběru n ovlivňuje přesnost odhadů parametrů polohy

Rozptyly odhadů jsou funkcí n^{-1}.

Metoda volby šíře intervalu přesnosti d:

(1) Z n_1 předběžných hodnot určí odhad rozptylu $s_0^2(x)$.

(2) Zvolí se číslo d tak, aby s pravděpodobností $(1 - \alpha)$ platilo

$$\mu - d \leq \bar{x} \leq \mu + d.$$

(3) Minimální velikost výběru se vyčíslí dle

$$n_{min} = \left[\frac{t_{1-\alpha/2}(n_1 - 1)}{d} \right]^2 s_0^2(x)$$

kde $t_{1-\alpha/2}(n_1 - 1)$ je kvantil Studentova rozdělení s $(n_1 - 1)$ stupni volnosti.

Metoda volby relativní směrodatné odchylky $\delta(s)$:

Relativní chyba směrodatné odchylky $\delta(s)$ má předepsanou hodnotu

$$n_{min} = \frac{g_2(x) - 1}{4\ \delta^2(s)} + 1$$

kde $g_2(x)$ je špičatost. Při $\delta(s) = 0.1$ (tj. 10%) je $n_{min} \approx 50$.

Minimální rozsahy výběru zajišťující relativní chybu směrodatné odchylky $\delta(s) = 10\ \%$

Rozdělení		Špičatost g_2	Minimální rozsah n_{min}
(A)	rovnoměrné	1.8	21
(B)	normální	3.0	51
(C)	exponenciální	6.0	126
(D)	Laplaceovo	9.0	176
(E)	lognormální	15.0	351

Ověření nezávislosti prvků výběru

Závislost měření je způsobena:

a) nestabilitou měřicího zařízení,

b) nekonstantností podmínek měření,

c) zanedbáním faktorů: objem vzorků, teplota, nečistota, ...

d) nesprávným, nenáhodným výběrem vzorků k měření,

e) časová závislost mezi prvky výběru.

TESTY NEZÁVISLOSTI

testují H_0:

Všechny prvky výběru jsou NAVZÁJEM nezávislé, ve výběru není autokorelace.

Používají se:

- testy autokorelace určitého řádu, např. pro autokorelaci I. řádu von Neumannův test
- testy významnosti autokorelačních koeficientů

Test významnosti autokorelačního koeficientu ρ_a:

Hypotéza: nulová H_0: $\rho_a = 0$, a alternativní H_A: $\rho_a \neq 0$.

Testační kritérium:

$$t_n = \frac{T_1 \sqrt{n+1}}{\sqrt{1 - T_1}} \qquad \text{kde} \qquad T_1 = \left(1 - \frac{T}{2}\right) \sqrt{\frac{n^2 - 1}{n^2 - 4}}$$

a T je von Neumannův poměr

$$T = \frac{\sum_{i=1}^{n-1} (x_{i+1} - x_i)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Testování: je-li

$$| t_n | > t_{1-\alpha/2}(n+1)$$

je nutno hypotézu o nezávislosti prvků výběru na hladině významnosti α zamítnout.

ZÁVISLOST A AUTOKORELACE

Obecná definice závislosti:

$$x_i = kF(x_1, x_2, \ldots, x_{i-1}) + e_i$$

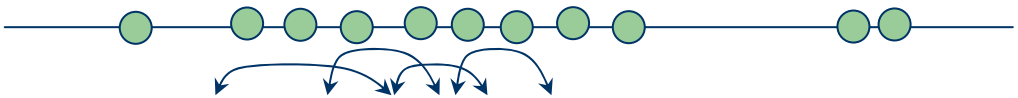
pokud platí k = 0, jedná se o data nezávislá

Lineární závislost prvků jednoho souboru - **AUTOKORELACE**

$$x_i = \rho_k \, x_{i-k} + e_i$$

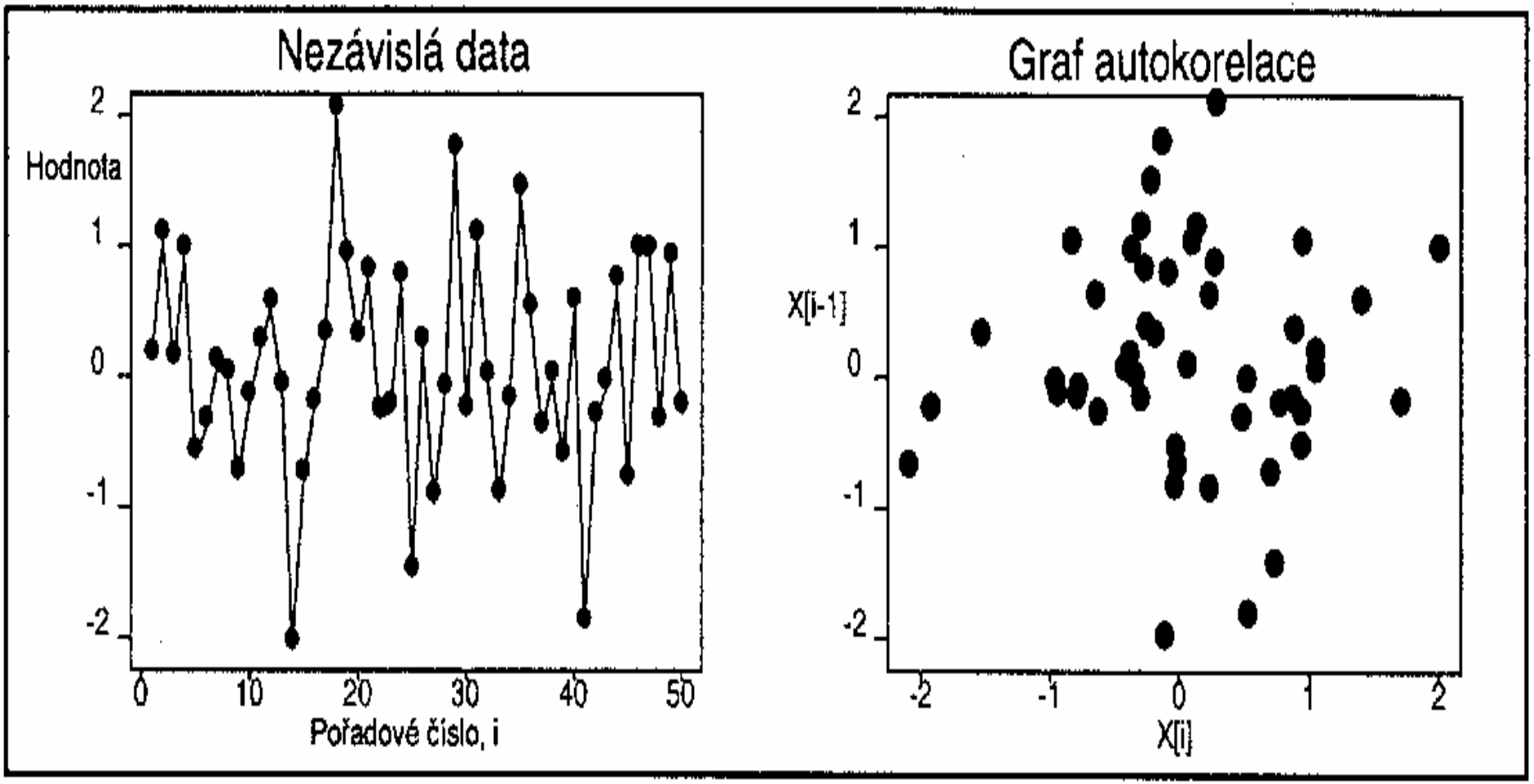
ρ_k autokorelační koeficient k-tého řádu

autokorelace I. řádu
sousední hodnoty

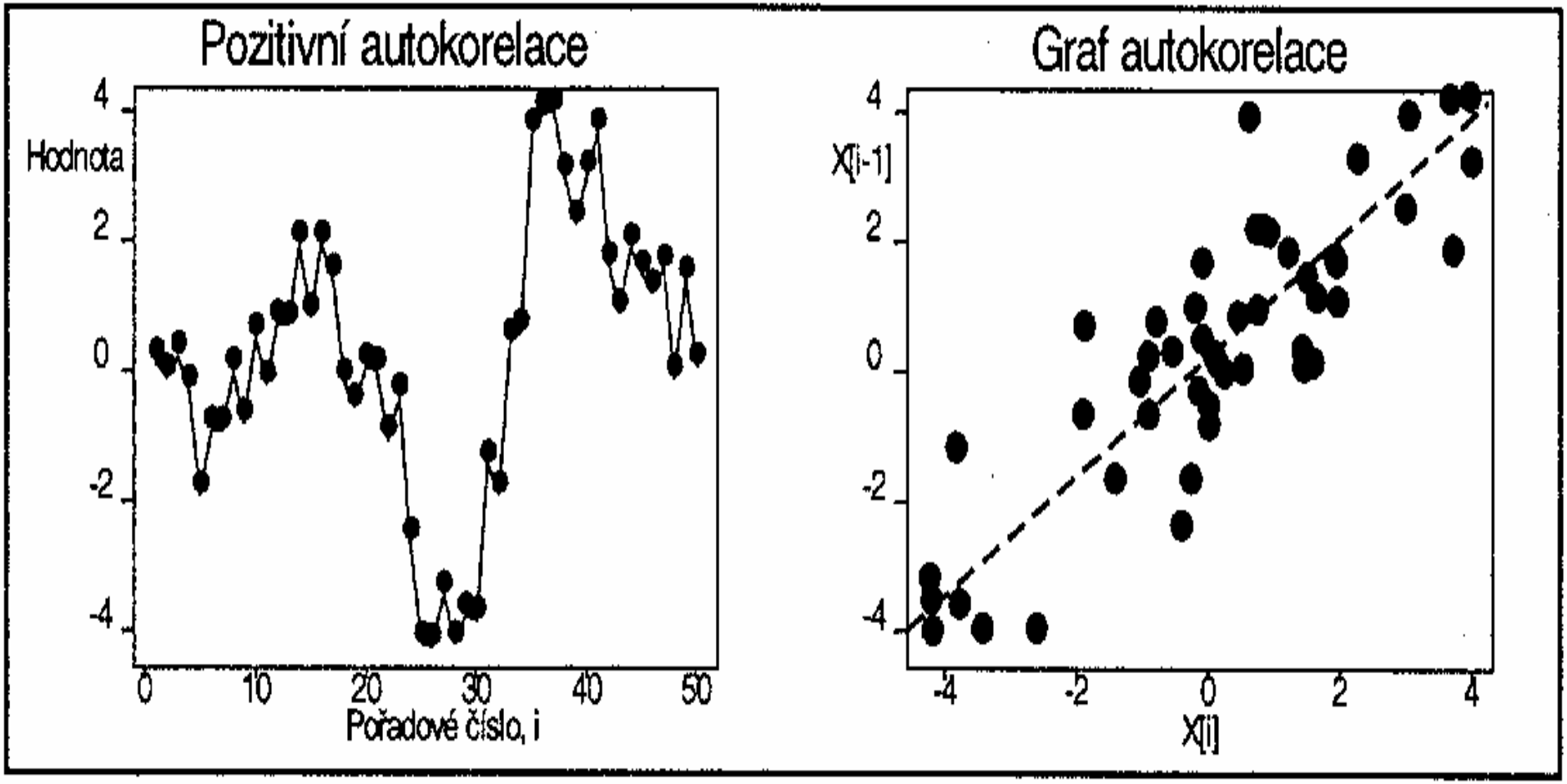


autokorelace II. řádu
hodnoty „přes jednu“

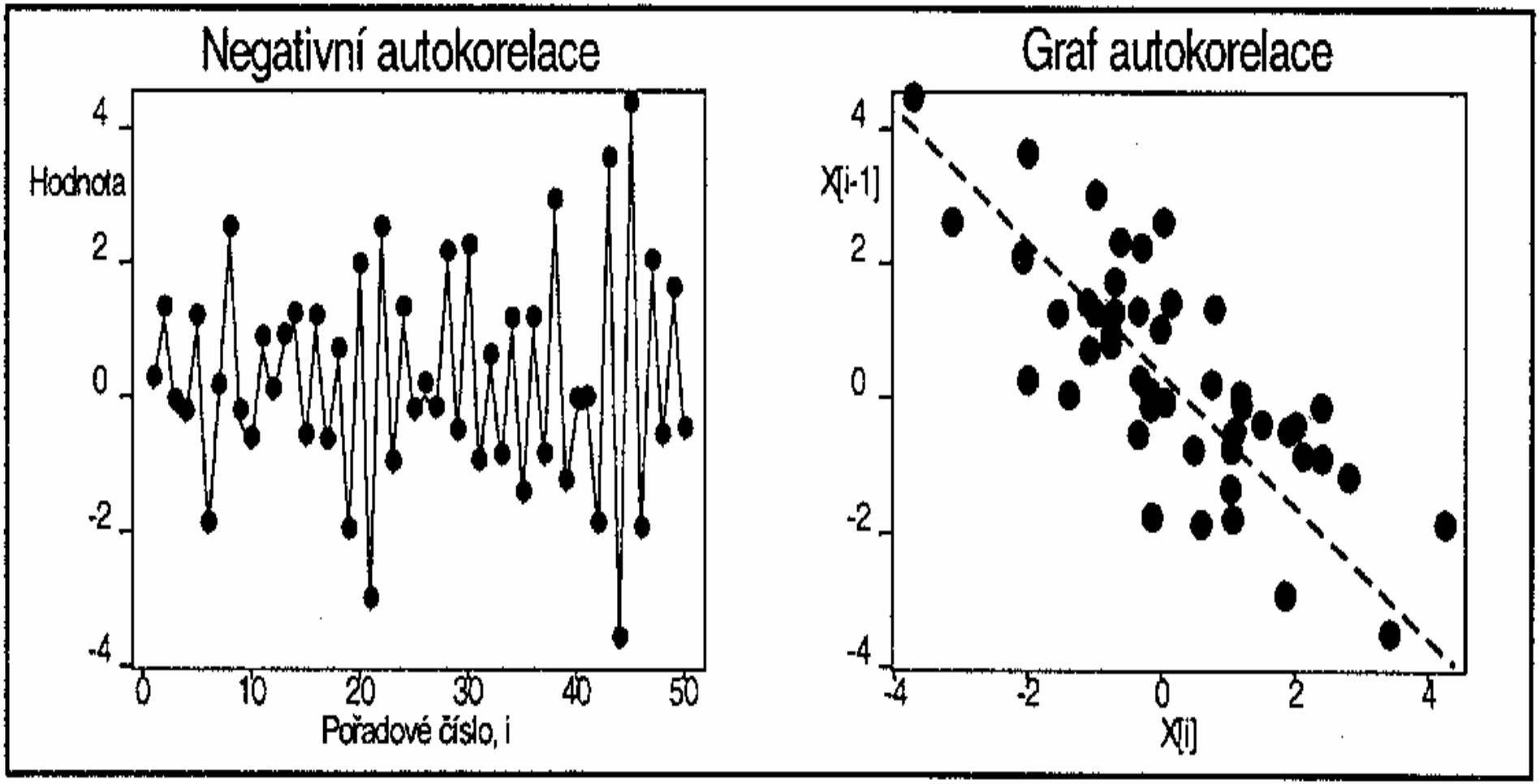
PŘÍKLADY AUTOKORELACE



PŘÍKLADY AUTOKORELACE



PŘÍKLADY AUTOKORELACE



Ověření normality výběru

Existují dva základní typy testů normality:

1. Je-li typ odchylek od normality předem znám, používají se tzv. ***směrové testy***.

2. Není-li typ odchylek od normality předem znám, používají se tzv. ***omnibus-testy***.

1. pravidlo: ***Testy jsou obecně vždy méně citlivé na odchylky od normality než diagnostické grafy.***

2. pravidlo: ***Když není normalita rozdělení prokázána, je nutné hlouběji analyzovat data.***

1. Test kombinace výběrové šikmosti a špičatosti.

Hypotéza: nulová H_0: normalita rozdělení výběru, *vs.* H_A: ...

Testovací kritérium: je definováno

$$C_1 = \frac{\hat{g}_1^2}{D(\hat{g}_1)} + \frac{[\hat{g}_2 - E(\hat{g}_2)]^2}{D(\hat{g}_2)}$$

kde výběrová šikmost a její rozptyl $\hat{g}_1$, $D(\hat{g}_1)$, resp. výběrová špičatost a její střední hodnota resp. rozptyl $\hat{g}_2$, $E(\hat{g}_2)$, $D(\hat{g}_2)$.

Testování: při $C_1 > \chi^2_{1-\alpha}(2)$, je nutno hypotézu o normalitě rozdělení výběru zamítnout.

Pro reálné velikosti výběrů je vhodnější testační kritérium C_2 dle

$$C_2 = Z^2(\hat{g}_1) + Z^2(\hat{g}_2)$$

kde $Z(\hat{g}_1)$, $Z(\hat{g}_2)$ jsou normální aproximace $\hat{g}_1$ a $\hat{g}_2$ dle postupu:

Vyčíslení $Z(\hat{g}_1)$:

1. Vypočtou se pomocné veličiny

$$Y = \hat{g}_1 \sqrt{\frac{(n + 1)(n + 3)}{6(n - 2)}}$$

$$G = \frac{3(n^2 + 27n - 70)(n + 1)(n + 3)}{(n - 2)(n + 5)(n + 7)(n + 9)}$$

$$W = -1 + \sqrt{2G - 1}$$

$$A = \sqrt{\frac{2}{W^2 - 1}}$$

2. Vypočítá se normální aproximace

$$Z(\hat{g}_1) = \frac{1}{\sqrt{\ln W}} \ln\left(\frac{Y}{A} + \sqrt{\left(\frac{Y}{A}\right)^2 + 1}\right)$$

Pokud data pocházejí z normálního rozdělení, má $Z(\hat{g}_1)$ normované normální rozdělení.

Vyčíslení $Z(\hat{g}_2)$:

1. Vypočte se střední hodnota $E(\hat{g}_2)$ a rozptyl $D(\hat{g}_2)$.
2. Určí se standardizovaná veličina

$$S = \frac{\hat{g}_2 - E(\hat{g}_2)}{\sqrt{D(\hat{g}_2)}}$$

3. Vypočte se šikmost veličiny S podle

$$g_1(S) = \frac{6\,(n^2 - 5\,n + 2)}{(n + 7)\,(n + 9)} \sqrt{\frac{6\,(n + 3)\,(n + 5)}{n\,(n - 2)\,(n - 3)}}$$

a dále pomocná veličina

$$A = 6 + \frac{8}{g_1(S)} \left[\frac{2}{g_1(S)} + \sqrt{1 + \frac{4}{g_1^2(S)}} \right]$$

4. Vypočítá se normální aproximace

$$Z(\hat{g}_2) = \frac{\left(1 - \frac{2}{9\,A}\right) - \sqrt[3]{\frac{1 - \frac{2}{A}}{1 + S\sqrt{\frac{2}{A-4}}}}}{\sqrt{\frac{2}{9\,A}}}$$

Pokud data pocházejí z normálního rozdělení, má $Z(\hat{g}_2)$ normované normální rozdělení.

2. *Andersonův-Darlingův test*

Hypotéza: nulová H_0: $F_E(x) = F_T(x)$ ***vs.*** alternativní H_A: $F_E(x) \neq F_T(x)$, kde $F_T(x)$ je teoretické a $F_E(x)$ je rozdělení výběru.

Testovací kritérium: je definováno

$$AD = -\frac{\sum_{i=1}^{n} (2\,i - 1)\,(\ln Z_i + \ln(1 - Z_{n-i+1}))}{n} - n$$

kde Z_i jsou

$$Z_i = F_T(x_{(i)})$$

Pro testaci normality je nulová hypotéza H_0: $F_E(x) = N(\bar{x}, s^2)$.

Veličiny $Z_i = \Phi[(x_{(i)} - \bar{x})/s]$ jsou hodnoty distribuční funkce normálního rozdělení.

Testování: $AD > D_{1-\alpha}$, nulová hypotéza se zamítá.

(Pro hodnotu kvantilu $D_{0.95}$ platí pro velký rozsah n vztah

$$D_{0.95} = 1.0348 \left(1 - \frac{1.013}{n} - \frac{0.93}{n^2}\right)$$

3. Shapirův-Wilkův test

Testovací kritérium: pro malé výběry $3 \leq n \leq 50$ má tvar

$$W = \frac{\left[\sum_{i=1}^{n} a_i \, x_{(i)}\right]^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Koeficienty a_i jsou tabelovány např. v ČSN 01 0225.

Testování: pokud je W menší než kvantil $W^*_{1-\alpha}$, zamítá se hypotéza o normalitě výběru.

Ověření homogenity výběru

Modifikace vnitřních hradeb B_D^* a B_H^*

$$B_D^* = \tilde{x}_{0.25} - K\,(\tilde{x}_{0.75} - \tilde{x}_{0.25})$$

$$B_H^* = \tilde{x}_{0.75} + K\,(\tilde{x}_{0.75} - \tilde{x}_{0.25})$$

Parametr K: volí se tak, aby pravděpodobnost P(n, K), že z výběru velikosti n pocházejícího z normálního rozdělení nebude žádný prvek mimo vnitřní hradby $[B_D^*, B_H^*]$, byla dostatečně vysoká, např. 0.95.

Při volbě $P(n, K) = 0.95$ lze v rozmezí $8 \leq n \leq 100$ použít aproximace

$$K \approx 2.25 - \frac{3.6}{n}$$

Pro takto určený parametr K se všechny prvky výběru ležící mimo hradby $[B_D^*, B_H^*]$ považují za vybočující.

TESTY ODLEHLÝCH HODNOT

$\mathbf{H_0}$: Odchylka extrémní hodnoty je náhodná

GRUBBSŮV TEST (předpokládá normální rozdělení)

$$T_n = \frac{x_n - \bar{x}}{S} \qquad T_1 = \frac{\bar{x} - x_1}{S}$$

Hypotéza je přijata, když $T_1 < T_{1,\alpha}$, resp. $T_n < T_{n.\alpha}$.

DIXONŮV TEST (nepředpokládá normální rozdělení)

$$Q_n = \frac{x_n - x_{n-1}}{x_n - x_1} \qquad Q_1 = \frac{x_2 - x_1}{x_n - x_1}$$

Hypotéza je přijata, když $Q_1 < Q_{1,\alpha}$ resp. $Q_n < Q_{n,\alpha}$.

TEST ODLEHLÝCH HODNOT V EDA

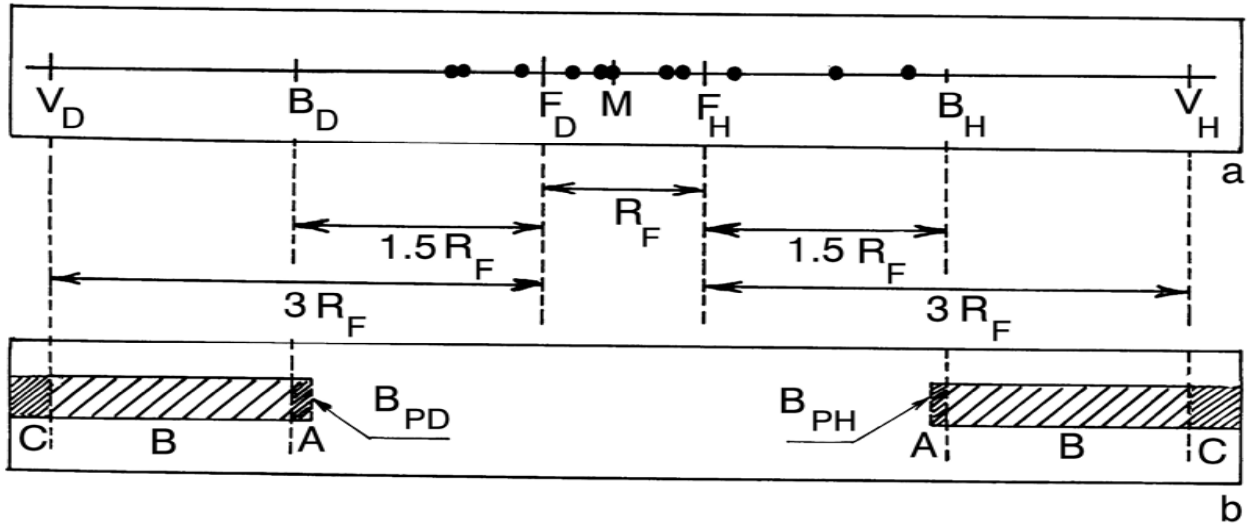
metoda **modifikovaných vnitřních hradeb**

$$B_D^* = F_D - K R_F \qquad B_H = F_H + K R_F$$

Při volbě $P(n, K) = 0.95$ lze v rozmezí $8 \leq n \leq 100$ použít aproximace

$$K \approx 2.25 - \frac{3.6}{n}$$

Pro takto určený parametr K se všechny prvky výběru ležící mimo hradby $[B_D^*, B_H^*]$ považují za vybočující.



Přilehlé hodnoty:

$$B_H = F_H + 1.5\ R_F$$ Horní přilehlá hodnota

$$B_D = F_D - 1.5\ R_F$$ Dolní přilehlá hodnota

Vnitřní hradby:

$$V_H = F_H + 3\ R_F$$ Horní vnitřní hradba

$$V_D = F_D - 3\ R_F$$ Dolní vnitřní hradba

"Podezřelá měření" (=vybočující hodnoty):

Body mimo interval $\{V_D, V_H\}$ vnitřních hradeb