Hipotēžu pārbaude par sadalījuma veidu

J. Valeinis[1]

[1]Latvijas Universitāte, Rīga

14.maijs, 2010

Hipotēžu pārbaude par sadalījumu: nostādne

Aplūkosim hipotēzi par populācijas sadalījuma veidu

$$H_0 : F(x) = F_0(x) \quad \text{pret} \quad H_1 : F(x) \neq F_0(x) \quad \forall x \in \mathbb{R},$$

- Hipotēžu iedalījums: vienkāršas (piemēram, $F_0 = N(0, 1)$) un saliktas (piemēram, $F_0 = N(\mu, \sigma^2)$)
- Testa statistikas: *Hī-kvadrāta*, *Kolmogorova-Smirnova*, Neimaņa, Lilliefora, Shapiro–Wilksa, Cramer-von-Mises, Andersona-Darlinga statistikas.

Hī-kvadrāta statistika

Izvēlamies $a_0 < a_1 < \ldots < a_{r-1}$, kur $a_i \in \mathbb{R}$. Parasti a_i izvēlas, sadalot F_0 vienādās daļās. Hī-kvadrāta statistika:

$$T = \sum_{i=1}^{r} \frac{(n_i - np_{i0})^2}{np_{i0}},$$

$p_{i0} = P(X \in [a_{i-1}, a_i))$, ja F_0 spēkā un n_i-novēroto datu skaits intervālā $[a_{i-1}, a_i)$.

- Ja H_0 spēkā, tad $T \to \chi^2_{r-1-k}$, citādi $T \to \infty$
- r parasti izvēlas tā, lai $np_{i0} \geq 5$, tas ir $r \leq n/5$. Problēma: r un a_i izvēle var ietekmēt lēmuma pieņemšanu!

Īrisu dati:histogramma

Īrisu dati apraksta īrisu kauslapu garumus ($n = 150$). Dati pētīti, lai noteiktu to ģeogrāfisko sadalījumu.

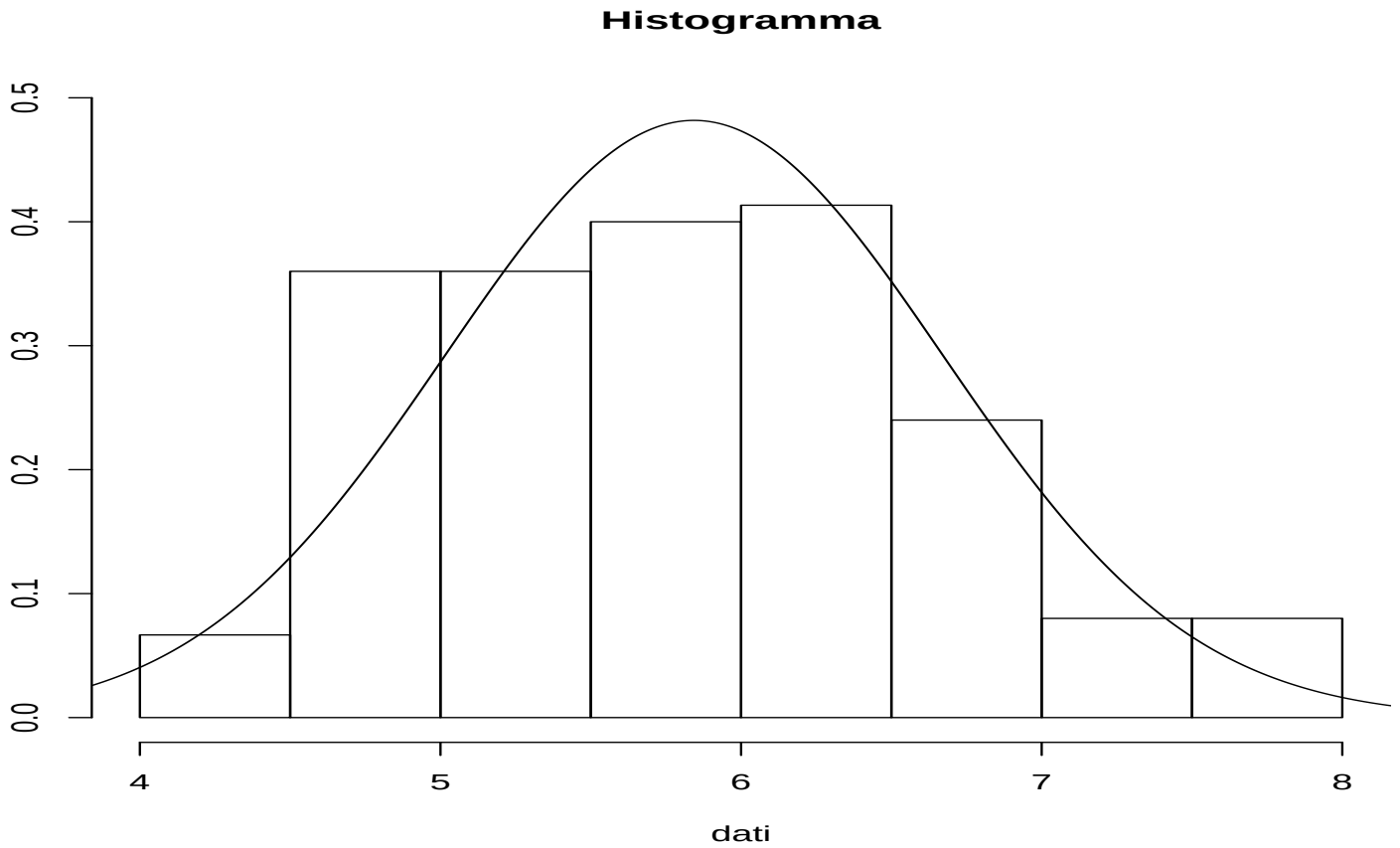


Figure: Histogramma ar pievienoto $N(\mu, \sigma^2)$ teorētisko blīvuma funkciju ar parametriem $\bar{X} = 5.84$, $S^2 = 0.68$

Īrisu dati:kvantiļu-kvantiļu (Q-Q) grafiks

Grafiku $\{F_1^{-1}(x), F_2^{-1}(x)\}$, kur $x \in (0,1)$ sauc par kvantiļu-kvantiļu grafiku divām sadalījuma funkcijām F_1 un F_2

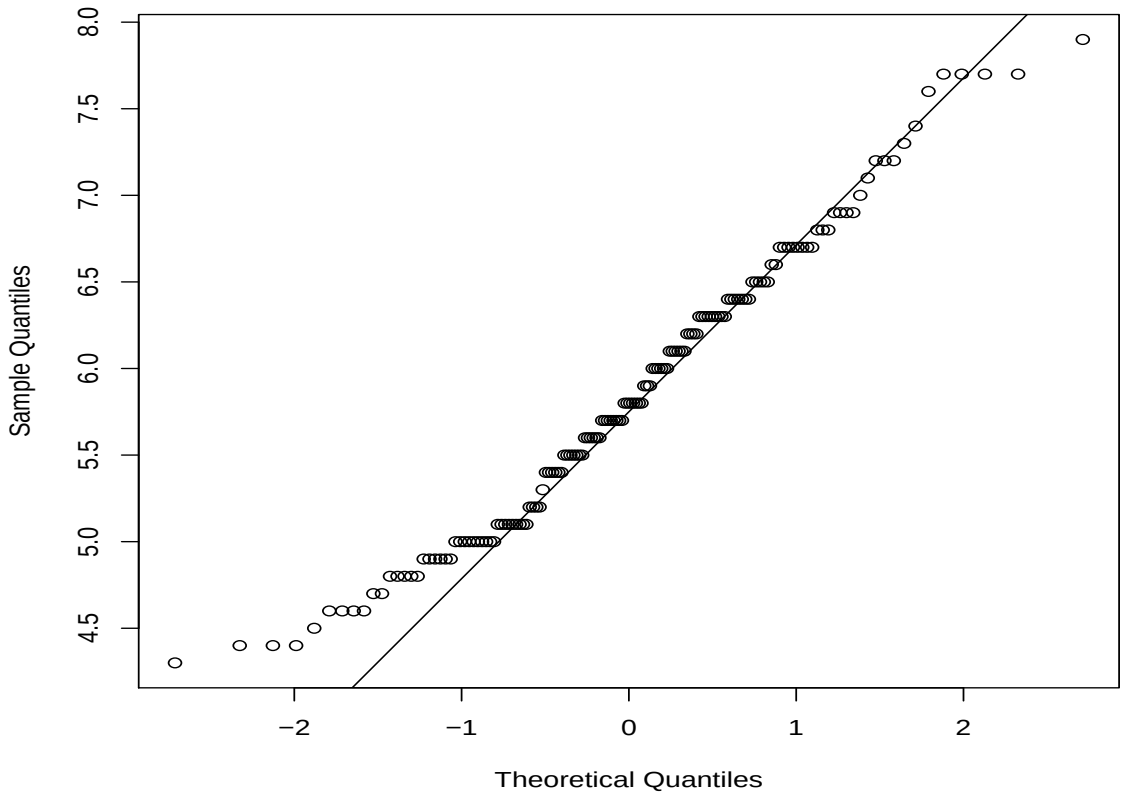


Figure: Q-Q grafiks Īrisu datiem ar pievienotu taisni, kas savieno pirmo un trešo kvartili

Īrisu dati: Empīriskā sadalījuma funkcija

Empīriskā sadalījuma funkcija $F_n(x) = n^{-1} \sum_{i=1}^{n} 1_{\{X_i \leq x\}}$ ir "labs" novērtējums populācijas sadalījumam $F(x)$!

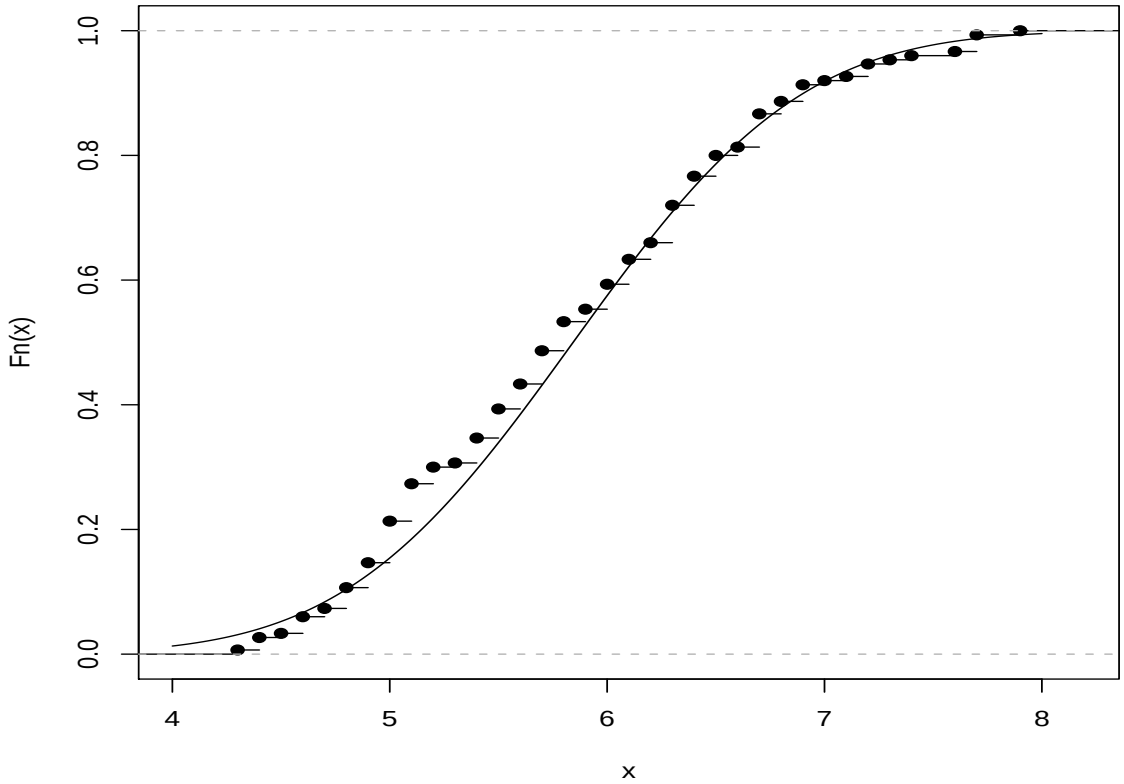


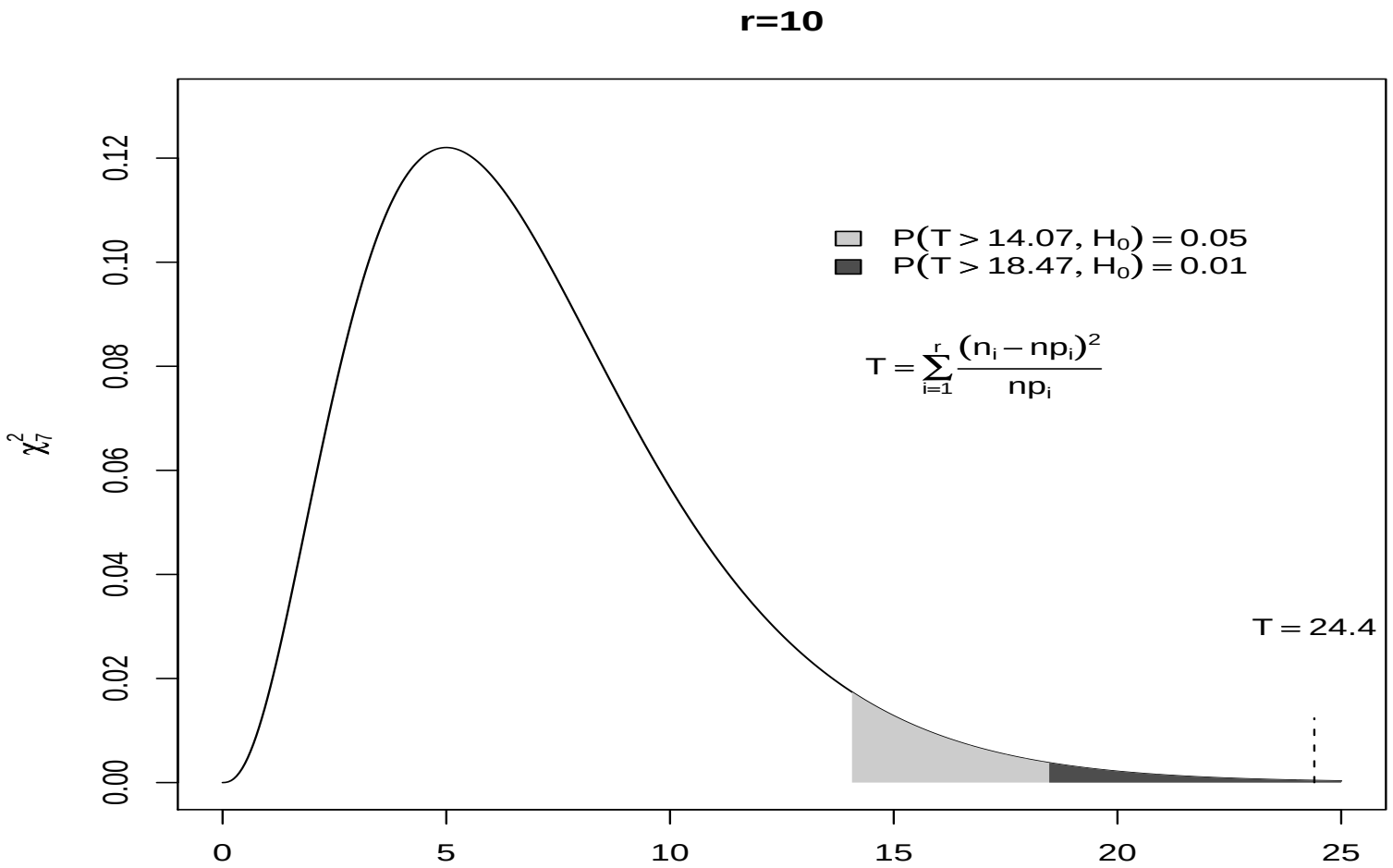
Figure: Empīriskā sadalījma funkcija $F_n(x)$ ar pievienoto teorētisko $N(\mu, \sigma^2)$ sadalījuma funkciju ar novērtētiem parametriem

Īrisu dati: p-vērtības

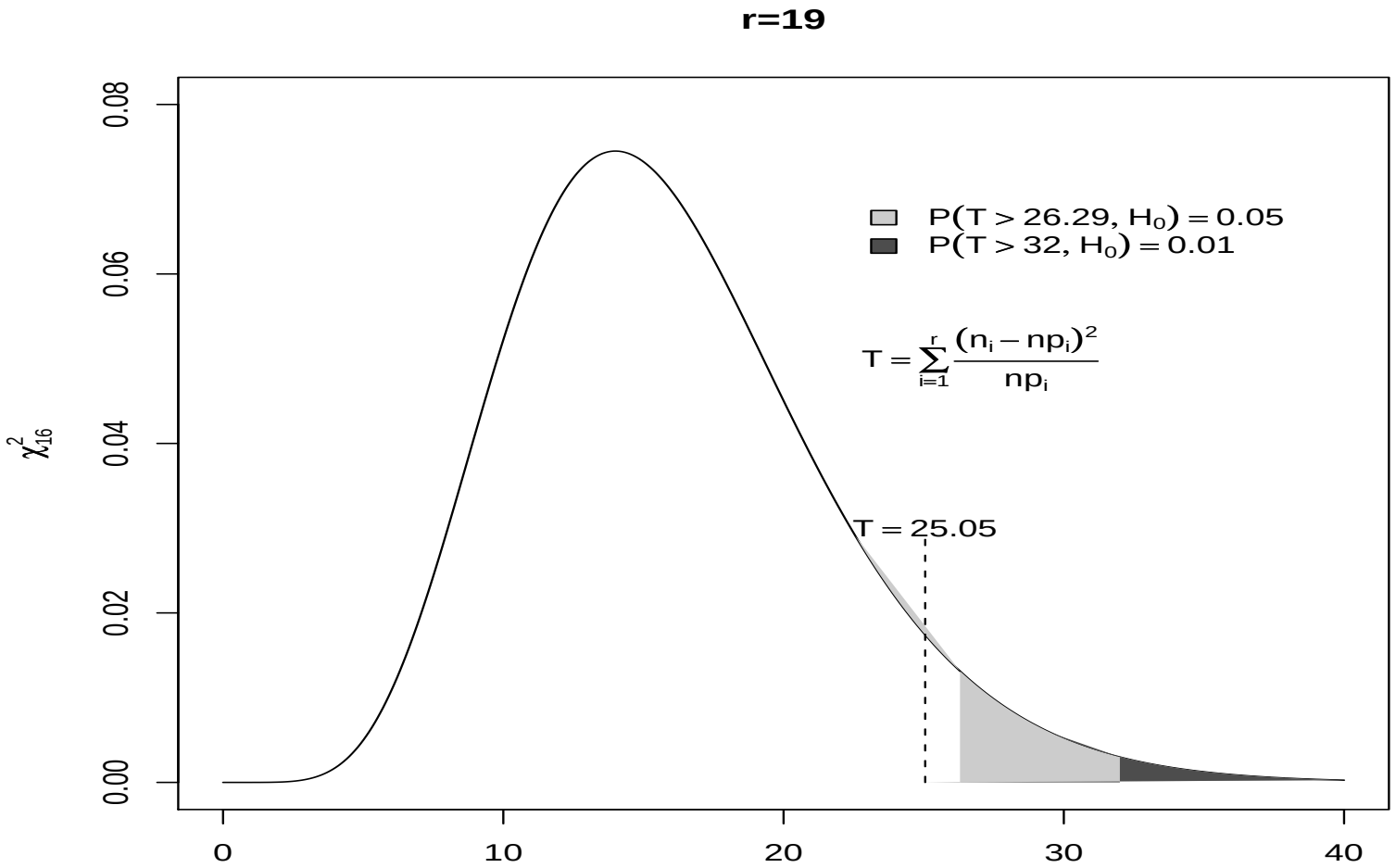
- Kritiskais apgabals formā $[c_\alpha, \infty]$, kur c_α - kritiskā vērtība
- $\alpha = 0.01; 0.05$
- p-vērtības dažādai intevālu r izvēlei:

r	p-vērtība
30	0.0001
25	0.0010
19	**0.0700**
17	0.0030
15	**0.1350**
14	0.0100

Hī-kvadrāta statistika: $r = 10$



Hī-kvadrāta statistika: $r = 19$



Kolmogorova-Smirnova statistika

Teorēma (Kolmogorova-Smirnova statistika)

Pieņemsim, ka $X_1, \ldots, X_n$ ir gadījuma izlase ar F. Tad

$$KS = \sup_{-\infty<x<\infty} \sqrt{n}|F_n(x) - F(x)| \to_d \sup_{0<t<1} |B(t)|,$$

kur $B(t)$ - Brauna tilts, $F_n(x) = n^{-1} \sum_{i=1}^{n} 1_{\{X_i \leq x\}}$ - empīriskā sadalījuma funkcija

- Teorēma ir spēkā, ja H_0 spēkā, tas ir pie patiesā sadalījuma $F(x)$ ar zināmiem parametriem (vienkārša hipotēze).
- Ja H_0 nav spēkā, tad $KS \to \infty$.
- Praksē parametri ir jānovērtē (salikta hipotēze) un robežsadalījums jāaproksimē ar simulāciju palīdzību.

Kolmogorova-Smirnova statistika: piezīmes

- Standarta Brauna tilta $B(t)$ sadalījums ir tāds pats kā procesam $W(t) - tW(1)$, kur $W(t)_{t\geq 0}$ apzīmē Brauna kustību (procesu, kuram ir $N(0,t)$ sadalījums, pieaugumi ir stacionāri un neatkarīgi).
- Brauna tilts ir sasiets galos (tas ir $B(0) = B(1) = 0$), pieaugumi ir atkarīgi, $B(t) \sim N(0, t(1-t))$
- Brauna kustība un Brauna tilts statistikā tiek izmantots plaši dēļ KS testa statistikas un Donskera teorēmas

KS un Brauna tilts

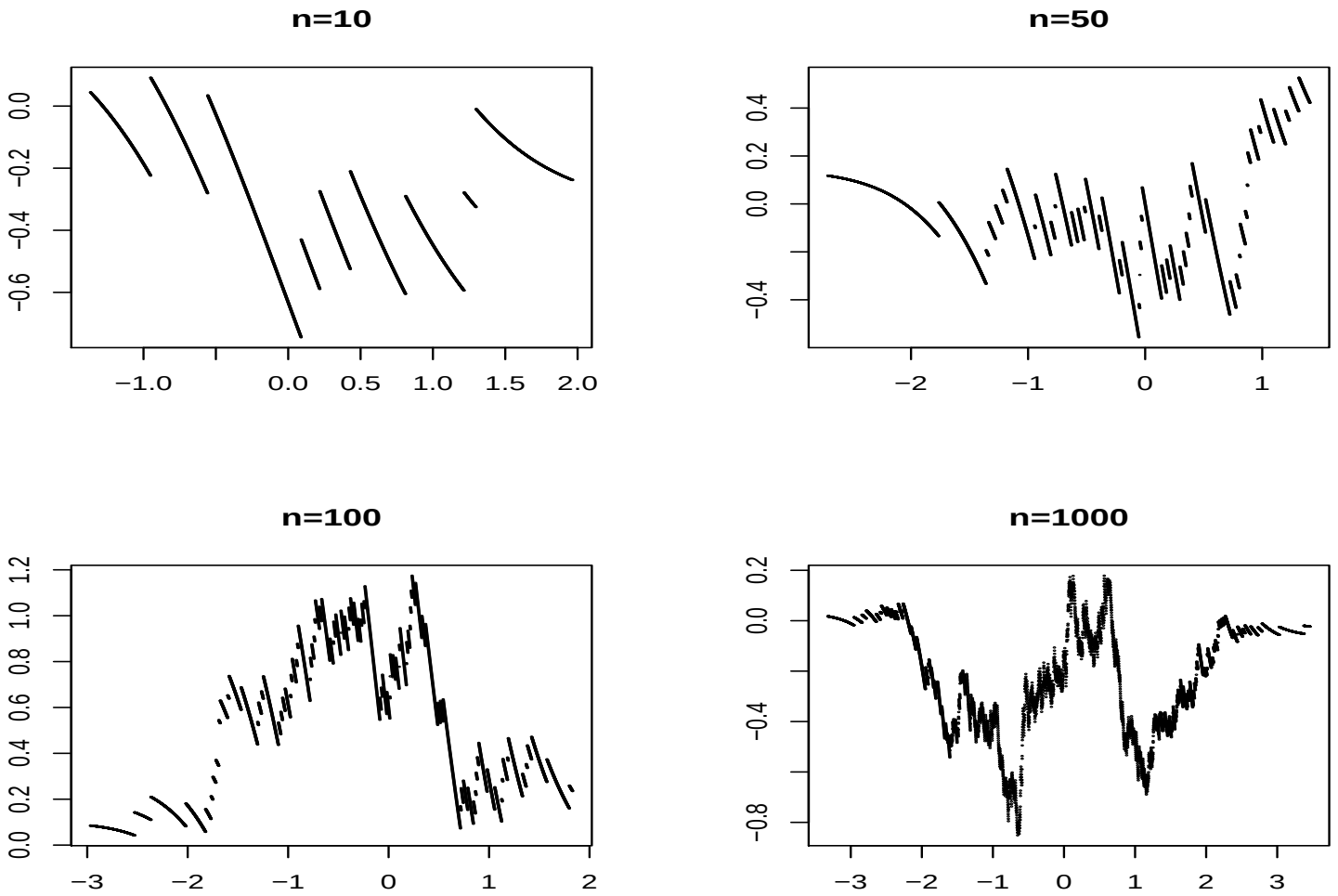


Figure: Statistikas $\sqrt{n}(F_n(x) - F(x))$ uzvedība

Donskera teorēma

Teorēma (Donskera teorēma: invariances pricips jeb funkcionālā centrālā robežteorēma)

Ja $X_1, X_2, ..., X_n$ ir neatkarīgi, vienādi sadalīti gadījuma lielumi ar $E(X_i) = 0$. Definēsim $S_n = X_1 + X_2 + \ldots + X_n$ un gadījuma definēsim gadījuma elementu

$$Y_n(t, w) = \frac{1}{\sigma\sqrt{n}} S_{[nt]}(w) + (nt - [nt]) \frac{1}{\sigma\sqrt{n}} X_{[nt]+1}(w).$$

Tad $Y_n \to_d W$, kur W apzīmē Brauna kustības jeb Vīnera procesa mēru.

KS un Brauna tilts

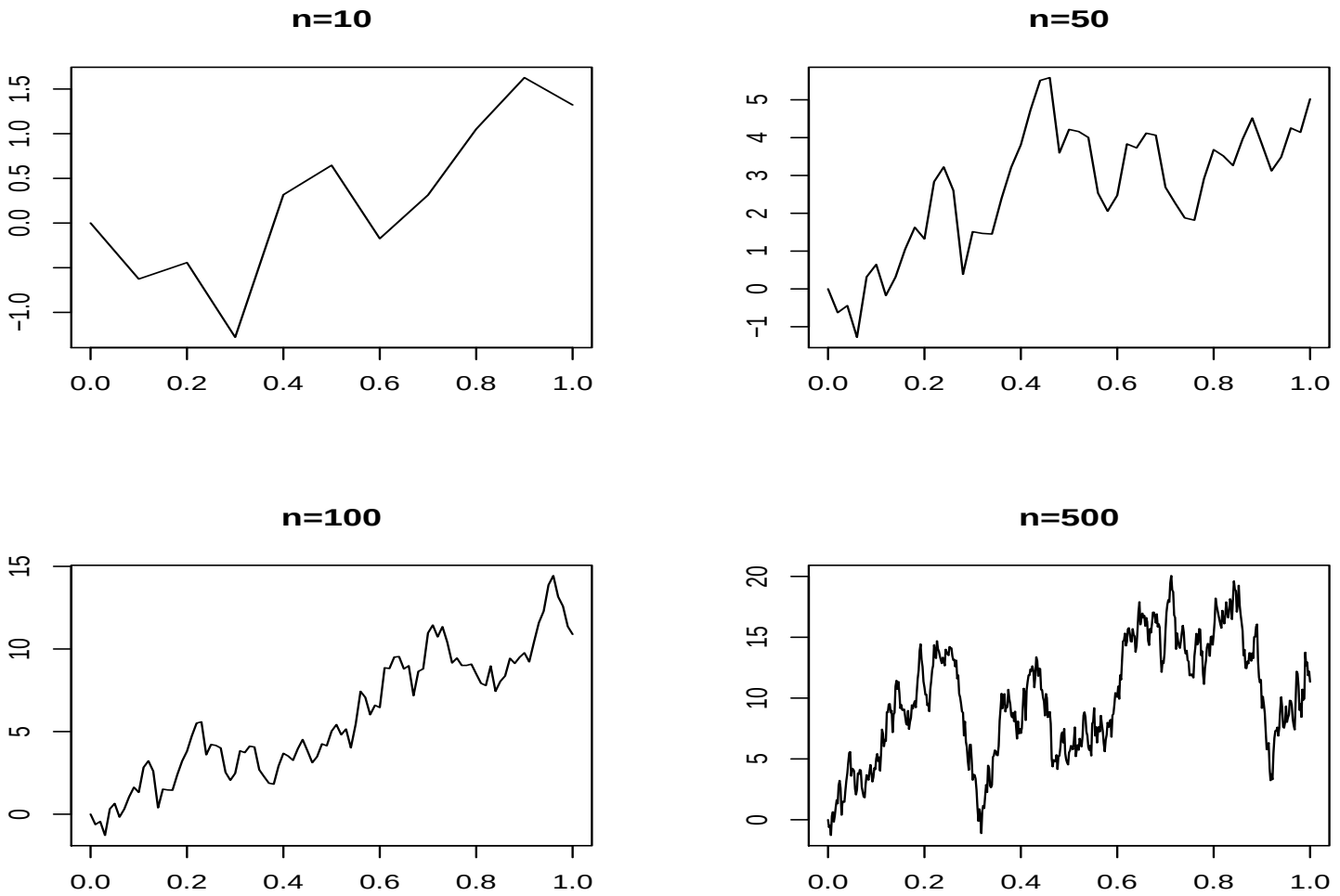


Figure: Donskera teorēma

Īrisu dati: p-vērtības un secinājumi

Testa statistikas	p-vērtības
Anderson-Darling	0.023
Lilliefors (Kolmogorov-Smirnov)	0.006
Shapiro-Francia	0.026
Cramer-von-Mises	0.047

Secinājumi:

- Īrisu datiem noraidam normalitāti!
- Ja p-vērtība nav pārāk liela, lietot dažādas testa statistikas. Uzmanīties ar Hī-kvadrāta statistiku!

Cits tests: Neimaņa tests neatkarīgiem un atkarīgiem datiem

$X_1, \ldots, X_n$ - atkarīgi novērojumi, $\phi_1, \phi_2, \ldots$ - ortonormāla polinomu sistēma. Neimaņa statistika

$$N_k = (12\sigma^2)^{-1} R_k = (12\sigma^2)^{-1} \sum_{j=1}^{k} \left\{ n^{-\frac{1}{2}} \sum_{i=1}^{n} \phi_j(X_i) \right\}^2$$

$$\sigma^2 = \sum_{t=-\infty}^{+\infty} \mathrm{Cov}(X_0, X_t).$$

- Munk, Stockis, Valeinis, Gieze (2009), Neyman smooth goodness-of-fit tests for the marginal distribution of dependent data, *Annals of the Institute of Statistical Mathematics*, pieņemta publicēšanai
- Mağistra darbi: Sintija Cīrule (2009), Inese Edule (2009)

Nākošās tēmas

- Neparametriskā statistika: neparametrisks blīvuma funkcijas novērtējums un neparametriskā regresija (prognozēšana laikrindām)
- Brauna kustības pārbaude un Frakcionālā Brauna kustība (ilglaicīgās atmiņas procesi laikrindu analīzē)
- Laikrindu analīze: prognozes veikšana ar ARIMA modeļiem un spektrālā analīze (spektru salīdzināšana)