Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE

Nadežda Langová

Empirická distribuční funkce

Katedra pravděpodobnosti a matematické statistiky
Vedoucí bakalářské práce: Mgr. Barbora Madurkayová

Studijní program: Matematika
Studijní obor: Obecná matematika

2009

Ďakujem vedúcej mojej bakalárskej práce, Mgr. Barbore Madurkayovej, za všetky cenné rady, pripomienky a predovšetkým za čas venovaný konzultáciám. Ďalej moja vďaka patrí Veronike Stankovianskej za pomoc s anglickým prekladom a Adriánovi Andrášovi za postrehy a nápady, ktoré mi pomohli pri upravovaní bakalárskej práce do finálnej podoby.

Prehlasujem, že som svoju bakalársku prácu napísala samostatne a výhradne s použitím citovaných prameňov. Súhlasím so zapožičiavaním práce a jej zverejňovaním.

V Prahe dňa 25. 5. 2009 Nadežda Langová

Obsah

Názov práce: Empirická distribuční funkce
Autor: Nadežda Langová
Katedra: Katedra pravděpodobnosti a matematické statistiky
Vedúci bakalárskej práce: Mgr. Barbora Madurkayová
e-mail vedúceho: madurka@karlin.mff.cuni.cz

Abstrakt: Táto práca sa zaoberá empirickou distribučnou funkciou, pričom prehľadne popisuje jej základné a asymptotické vlastnosti. Cieľom je čitateľa oboznámiť s jej využitím v matematickej štatistike, pozornosť je venovaná štúdiu jednovýberových testov dobrej zhody a dvojvýberových neparametrických testov, ktoré sú založené na empirickej distribučnej funkcii. Konkrétne ide o Kolmogorovov-Smirnovov a Cramerov-von Misesov test a ich najznámejšie modifikácie. Text je priebežne doplnený príkladmi a grafmi.

Kľúčové slová: empirická distribučná funkcia, testy dobrej zhody, Kolmogorovov-Smirnovov test, Lillieforsov test, Cramerov-von Misesov test, Andersonov-Darlingov test

Title: Empirical distribution function
Author: Nadežda Langová
Department: Department of Probability and Mathematical Statistics
Supervisor: Mgr. Barbora Madurkayová
Supervisor's e-mail address: madurka@karlin.mff.cuni.cz

Abstract: The thesis is concerned with the empirical distribution function and concise characterisation of its fundamental and asymptotic properties. The work's primary aim is to provide the reader with the applications of the empirical distribution function in the domain of mathematical statistics; putting emphasis on the study of one-sample goodness-of-fit and two-sample nonparametric tests. More precisely, the paper deals with Kolmogorov-Smirnov and Cramer-von Mises tests respectively, including some of their well-known modifications. The text is consecutively supplemented with examples and graphs.

Keywords: empirical distribution function, goodness-of-fit tests, Kolmogorov-Smirnov test, Lilliefors test, Cramer-von Mises test, Anderson-Darling test

Úvod

Klasické postupy štatistického uvažovania často predpokladajú presne špecifikované rozdelenie dát, napríklad exponenciálne rozdelenie s parametrom $\lambda = 5$ alebo normované normálne rozdelenie, o ktorom vieme, že má nulovú strednú hodnotu a jednotkový rozptyl. Tieto predpoklady však nie sú vždy splnené, hodnota parametra nie je známa, stredná hodnota a rozptyl nie sú bližšie určené. Poprípade vieme, že distribučná funkcia F nejakého rozdelenia je spojitá, ale nič viac. Všetky spomenuté problémy nás vedú k pojmu empirická distribučná funkcia, ktorá sa stala predmetom tejto práce.

V prvej kapitole zavedieme pojem empirická distribučná funkcia a popíšeme jej asymptotické vlastnosti. Zameriame sa na jej vzťah k pôvodnej distribučnej funkcii F náhodného výberu a správanie pri veľkom rozsahu. Zistíme, že je nestranným a konzistentným odhadom funkcie F.

Druhá kapitola je venovaná najznámejším jednovýberovým a dvojvýberovým testom založeným na empirickej distribučnej funkcii. Sú uvedené predpoklady ich prevedenia, nulové hypotézy, testovacie štatistiky a kritérium pre zamietnutie nulovej hypotézy. Vybrané testy sú doplnené príkladmi, na ktorých vyskúšame použitie jednotlivých testov a riešenia doplníme grafmi a tabuľkami.

V tretej kapitole nasimulujeme náhodné výbery z dvoch odlišných rozdelení s rôznymi rozsahmi a pomocou štatistického softwaru a jednovýberových testov z druhej kapitoly otestujeme hypotézu, že výbery pochádzajú z normálneho rozdelenia. Použité príkazy podrobne popíšeme, výsledky a kritické hodnoty usporiadame do prehľadných tabuliek a doplníme grafmi pre jednotlivé prípady. Takto vizuálne posúdime empirickú a teoretickú distribučnú funkciu rozdelenia, z ktorého náhodné dáta pochádzajú za nulovej hypotézy, a distribučnú funkciu rozdelenia, z ktorého boli naozaj nasimulované.

Riešenia príkladov, grafy aj simulácie v práci sú prevedené pomocou štatistického programu R 2.8.1, ktorý je voľne dostupný na webovej stránke http://cran.r-project.org/index.html.

Kapitola 1

Empirická distribučná funkcia a jej asymptotické vlastnosti

1.1 Úvod

V tejto kapitole sa zoznámime s empirickou distribučnou funkciou a zameriame sa na jej vlastnosti, o ktoré sa neskôr budeme opierať.

Nech $(\Omega, \mathcal{A}, P)$ je pravdepodobnostný priestor, kde Ω je neprázdna množina elementárnych javov, $\mathcal{A}$ je σ-algebra podmnožín Ω a P je pravdepodobnostná miera. Označme $(\mathcal{X}, \mathcal{B})$ výberový priestor, kde $\mathcal{X}$ je borelovská podmnožina $\mathbb{R}^n$ a $\mathcal{B}$ je systém borelovských podmnožín $\mathcal{X}$. Merateľné zobrazenie $\mathbf{X} : (\Omega, \mathcal{A}) \to (\mathcal{X}, \mathcal{B})$ sa nazýva n-rozmerný náhodný vektor.

1.2 Empirická distribučná funkcia

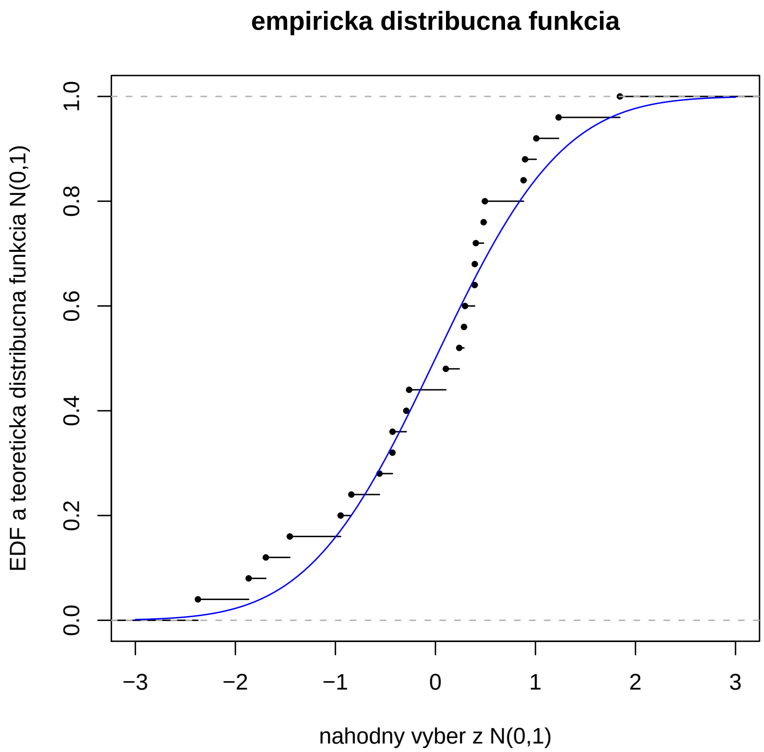
Nech $X_1, \ldots, X_n$ je náhodný výber z rozdelenia s distribučnou funkciou F, kde $n \in \mathbb{N}$. Nech je dané $x \in \mathbb{R}$ a nech

$$I\left(X \leq x\right) = \begin{cases} 0 & \text{pre } X > x, \\ 1 & \text{inak} \end{cases} \tag{1.1}$$

je tzv. indikátorová funkcia. Funkciu

$$\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x) \tag{1.2}$$

nazývame **empirická distribučná funkcia**.



Obr. 1.1 Ilustratívny obrázok grafu distribučnej funkcie normálneho rozdelenia $N(0,1)$ (vykreslená súvislou čiarou) a príslušnej empirickej distribučnej funkcie náhodného výberu z $N(0,1)$ s rozsahom 25 pozorovaní, ktorý bol nasimulovaný v programe R

Už z definície je zrejmé, že $\widehat{F}_n(x)$ je schodovitá sprava spojitá funkcia so skokmi v pozorovaných hodnotách X_i. Ak usporiadame hodnoty náhodných veličín $X_1, \ldots, X_n$ podľa veľkosti do postupnosti $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$, potom môžeme položiť

$$\widehat{F}_n(x) = \frac{k}{n} \qquad \text{pre} \quad x \in \left(x_{(k)}, x_{(k+1)}\right), k = 0, 1, \ldots, n \tag{1.3}$$

a $x_{(0)} = -\infty$, $x_{(n+1)} = \infty$, pričom $x_1, \ldots, x_n$ je realizácia náhodného výberu $X_1, \ldots, X_n$. Hodnota $n\widehat{F}_n(x)$ je počet elementov výberu $X_1, \ldots, X_n$, ktoré sú menšie ako x.

1.3 Vlastnosti empirickej distribučnej funkcie

Ukážeme (formálne aj graficky), že s rastúcim n sa funkcia $\widehat{F}_n(x)$ blíži k skutočnej distribučnej funkcii $F(x)$. Zvoľme pevné $x \in \mathbb{R}$ a pre naše potreby zapíšeme $\widehat{F}_n(x)$ v tvare

$$\widehat{F}_n(x) = \bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i, \tag{1.4}$$

kde $Y_i = I(X_i \leq x), i = 1, \ldots, n$ je definované vzťahom (1.1). Teda Y_i nadobúda len hodnoty 0 alebo 1 pre $i = 1, \ldots, n$ a má alternatívne rozdelenie $Alt(p)$ s parametrom $p = F(x)$.

S použitím vzťahu (1.4) teraz už ľahko dokážeme, že $n\widehat{F}_n \sim Bi(n, F(x))$, pretože pre pevne zvolené $x \in \mathbb{R}$ je $n\widehat{F}_n(x) = \sum_{i=1}^{n} Y_i$ súčtom nezávislých rovnako rozdelených náhodných veličín $Y_i \sim Alt(F(x))$, kde $i = 1, \ldots, n$.

Pre strednú hodnotu a rozptyl empirickej distribučnej funkcie platí:

$$E\widehat{F}_n(x) = E\bar{Y}_n = EY_1 = F(x), \tag{1.5}$$

$$var\widehat{F}_n(x) = var\bar{Y}_n = \frac{1}{n}varY_1 = \frac{F(x)(1 - F(x))}{n}. \tag{1.6}$$

Zo vzťahu (1.5) plynie, že $\widehat{F}_n(x)$ je nestranným odhadom funkcie $F(x)$ a zo vzťahu (1.6) vidíme, že $var\widehat{F}_n(x) \longrightarrow 0$ pre $n \to \infty$. Spojenie tohto poznatku s Čebyševovou nerovnosťou nasledujúcim spôsobom, kde pre každý $\epsilon > 0$ platí

$$P\left(|\widehat{F}_n(x) - F(x)| > \epsilon\right) \leq \frac{F(x)(1 - F(x))}{n\epsilon^2} \to 0 \text{ pre } n \to \infty,$$

nám ukazuje, že $\widehat{F}_n(x)$ je aj konzistentným odhadom $F(x)$, dokonca pre každé $x \in \mathbb{R}$ platí

$$\widehat{F}_n(x) \to F(x) \text{ s.v. pre } n \to \infty. \tag{1.7}$$

Vzťah (1.7) plynie z aplikácie silného zákona veľkých čísel pre nezávislé rovnako rozdelené náhodné veličiny.

Nasledujúca veta predstavuje ešte silnejšie tvrdenie, pri ktorom sa nemusíme obmedzovať na pevné x.

Veta 1.1 (*Glivenko-Cantelli*).

$$\sup_{x\in\mathbb{R}} |\widehat{F}_n(x) - F(x)| \xrightarrow{s.v.} 0 \quad \textit{pre} \quad n \to \infty.$$

Dôkaz: Viď Borovkov [3], str. 5–6. □

Ak je $F(x)$ spojitá, potom platí

$$P\left(\limsup_{n\to\infty} \sqrt{\frac{2n}{\ln\ln n}} \cdot \sup_x \left|\widehat{F}_n(x) - F(x)\right| = 1\right) = 1.$$

Tento vzťah (viď Borovkov [3], str. 12) nám približuje rýchlosť konvergencie z vety 1.1. Z Moivre-Laplaceovej vety (viď Appendix, Veta A.3) odvodíme asymptotické normálne rozdelenie pre pevné x a $n \to \infty$:

$$\sqrt{n}\left(\widehat{F}_n(x) - F(x)\right) = \sqrt{n}\left(\bar{Y}_n - EY_i\right) \xrightarrow{d} N\left(0, F(x)\left(1 - F(x)\right)\right).$$

Interval spoľahlivosti: Nech $X_1, \ldots, X_n$ je náhodný výber z rozdelenia s distribučnou funkciou F. Zostrojíme približný interval spoľahlivosti pre $F(x)$ metódou, ktorú popísali napr. Dupač, Hušková [4], str. 106.

Nech x je ľubovoľne zvolené reálne číslo. Z tvrdenia odvodeného z centrálnej limitnej vety

$$\sqrt{n}\left(\widehat{F}_n(x) - F(x)\right) \xrightarrow{d} N\left(0, F(x)\left(1 - F(x)\right)\right)$$

plynie

$$\sqrt{n}\left(\frac{\widehat{F}_n(x) - F(x)}{\sqrt{F(x)(1-F(x))}}\right) \xrightarrow{d} N(0,1).$$

Nasledujúci krok umožňuje Cramérova-Sluckého veta (viď Appendix, Veta A.4)

$$\sqrt{n}\left(\frac{\widehat{F}_n(x) - F(x)}{\sqrt{\widehat{F}_n(x)\left(1-\widehat{F}_n(x)\right)}}\right) \xrightarrow{d} N(0,1).$$

Nech u_β je kvantil normovaného normálneho rozdelenia, kde v našom prípade $\beta = 1-\alpha/2$ a $0<\alpha<1$. Potom

$$P\left[-u_{1-\frac{\alpha}{2}} < \sqrt{n}\left(\frac{\widehat{F}_n(x) - F(x)}{\sqrt{\widehat{F}_n(x)\left(1-\widehat{F}_n(x)\right)}}\right) < u_{1-\frac{\alpha}{2}}\right] \longrightarrow 1-\alpha \quad \text{pre } n \to \infty,$$

z čoho plynie

$$P\left[\widehat{F}_n(x) - u_{1-\frac{\alpha}{2}}\frac{\sqrt{\widehat{F}_n(x)\left(1-\widehat{F}_n(x)\right)}}{\sqrt{n}} < F(x) < \widehat{F}_n(x) + u_{1-\frac{\alpha}{2}}\frac{\sqrt{\widehat{F}_n(x)\left(1-\widehat{F}_n(x)\right)}}{\sqrt{n}}\right]$$

$\longrightarrow 1-\alpha \quad$ pre $n \to \infty$.

Dostaneme intervalový odhad $F(x)$ o asymptotickej spoľahlivosti $1-\alpha$

$$\left(\widehat{F}_n(x) - u_{1-\frac{\alpha}{2}}\frac{\sqrt{\widehat{F}_n(x)\left(1-\widehat{F}_n(x)\right)}}{\sqrt{n}},\ \widehat{F}_n(x) + u_{1-\frac{\alpha}{2}}\frac{\sqrt{\widehat{F}_n(x)\left(1-\widehat{F}_n(x)\right)}}{\sqrt{n}}\right).$$

Empirické odhady parametrov: Nech $X_1, \ldots, X_n$ je náhodný výber z rozdelenia s neznámou distribučnou funkciou $F(x)$ a $\theta = t(F)$ parameter, ktorý odhadneme pomocou funkcie $\widehat{F}_n(x)$. Namiesto $t(F)$ spočítame $\hat{\theta}_n = t(\widehat{F}_n)$. Odhad $\hat{\theta}_n$ sa nazýva empirický odhad parametra θ (t označuje funkcionál).

Najskôr

$$\int_{-\infty}^{+\infty} h(x)\, d\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} h(X_i) \quad \text{pre pevné } X_i.$$

Označme $X := X_1$, kde X_1 je podľa predpokladu náhodná veličina s rozdelením s distribučnou funkciou F. Teraz už ľahko odvodíme empirické odhady k-tého momentu $a_k = EX^k$ a k-tého centrálneho momentu $b_k = E(X-EX)^k$ ako

$$\hat{a}_k := \hat{a}_k(X) = \int_{-\infty}^{+\infty} x^k\, d\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} X_i^k,$$

$$\hat{b}_k := \hat{b}_k(X) = \int_{-\infty}^{+\infty} (x-\hat{a}_1)^k\, d\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} (X_i-\hat{a}_1)^k.$$

Pre momenty $\hat{a}_1$ a $\hat{b}_2$ sa často v literatúre používajú špeciálne symboly

$$\bar{X} = \hat{a}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \hat{S}^2 = \hat{b}_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2.$$

Ďalšou charakteristikou náhodného výberu je β-kvantil príslušného rozdelenia (napr. pri odvodení približného intervalu spoľahlivosti pre $F(x)$ bol spomenutý kvantil normovaného normálneho rozdelenia), definovaný predpisom

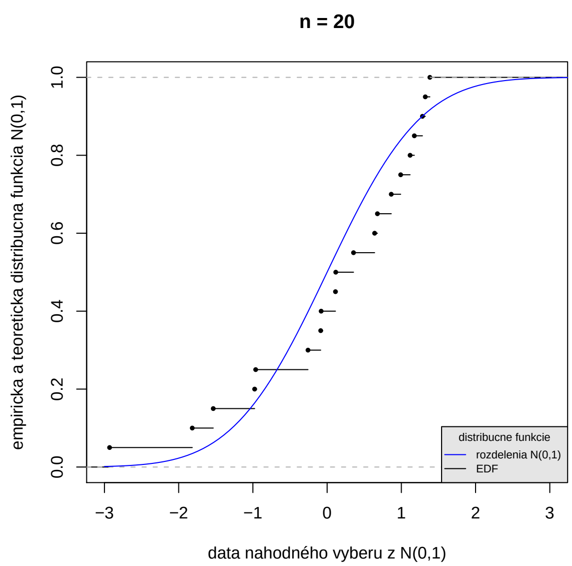
$$x_\beta = F^{-1}(\beta) = \inf\{x : F(x) \geq \beta\}, \quad 0 < \beta < 1.$$

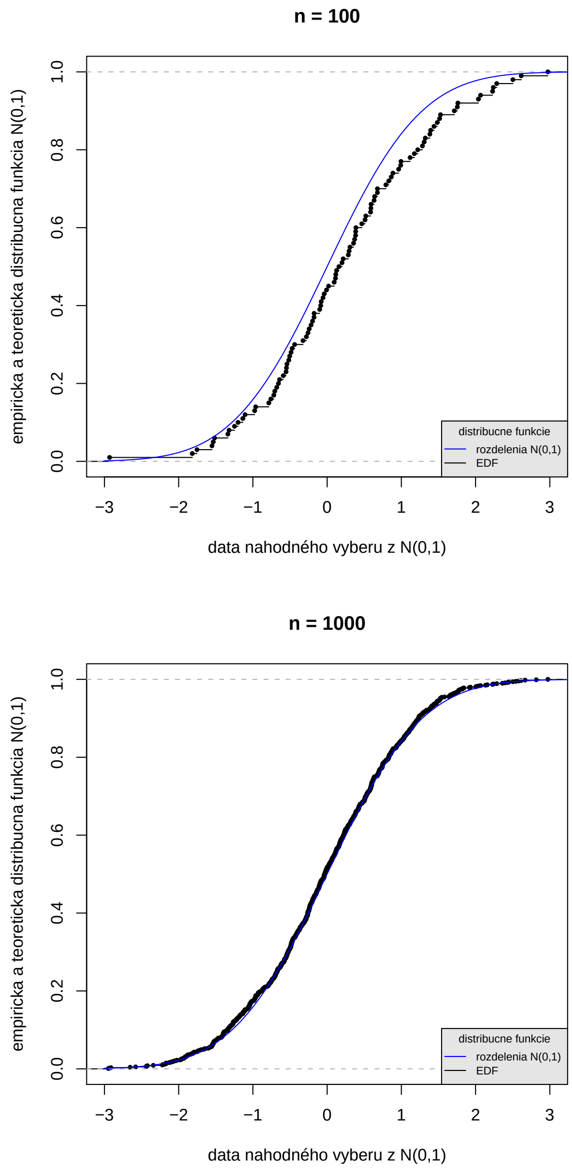
Hodnotu

$$\hat{x}_\beta = \widehat{F}_n^{-1}(x) = \inf\left\{x : \widehat{F}_n(x) \geq \beta\right\}, \quad 0 < \beta < 1,$$

vyjadrenú tiež pomocou empirickej distribučnej funkcie, nazveme výberový kvantil. Viac príkladov na výberové charakteristiky uviedol Borovkov [3], str.7.

Na záver tejto kapitoly si ešte môžeme pre lepšiu predstavu graficky naznačiť, ako sa pre čoraz väčšie n funkcia $\widehat{F}_n(x)$ blíži k pôvodnej distribučnej funkcii $F(x)$ (viď obr. 1.2).





Obr. 1.2 Grafy distribučnej funkcie normálneho rozdelenia $N(0,1)$ (vykreslená súvislou modrou farbou) a príslušnej empirickej distribučnej funkcie s naznačeným rozsahom náhodného výberu, ktorý bol nasimulovaný v programe R

Kapitola 2

Testy založené na empirickej distribučnej funkcii

2.1 Úvod do neparametrických testov

V predchádzajúcej kapitole sme si vymenovali vlastnosti empirickej distribučnej funkcie (EDF), ktorá má ešte jednu výnimočnú vlastnosť a tou je vhodný tvar z hľadiska aplikácií a využitia v štatistike.

V tejto kapitole sa budeme zaoberať testami, ktoré spočívajú v posudzovaní zhody EDF náhodného výberu s nejakou teoretickou distribučnou funkciou. Patria do skupiny tzv. **testov dobrej zhody**, ktoré overujú hypotézu, že daný náhodný výber pochádza z nejakého teoretického rozdelenia. Okrem toho často testujeme hypotézy, ktoré tvrdia, že dva nezávislé náhodné výbery pochádzajú z toho istého rozdelenia. Voľba vhodného testu potom úzko súvisí s príslušnou alternatívou, ktorá závisí na tom, čo môžeme a chceme o týchto rozdeleniach predpokladať.

Klasické postupy štatistického usudzovania často predpokladajú normálne rozdelenie dát, prípadne nejaké iné, ale taký predpoklad nie je skoro nikdy splnený. Preto boli vyvinuté neparametrické metódy usudzovania, tzv. **neparametrické testy**, ktorých výhoda spočíva v tom, že na tvar rozdelenia dát nemusíme klásť špeciálne požiadavky.

Teória testovnia hypotéz je dopodrobna rozoberaná vo viacerých publikáciách. Základné definície a pojmy, ktoré budeme v tejto časti používať, uvádza napr. Anděl [1], kap. 8, Dupač, Hušková [4], kap. 6, viaceré si vysvetlíme či už v priamo texte alebo v appendixe na konci práce. Text je doplnený príkladmi a riešenia sú prevedené pomocou softwaru R 2.8.1.

2.2 Kolmogorovov-Smirnovov test

Medzi najznámejšie testy, ktoré si v tejto práci rozoberieme, nesporne patrí Kolmogorovov-Smirnovov (ďalej K-S) test, ktorý sa stal východiskom pre vznik ďalších testov založených na podobnom princípe.

2.2.1 Jednovýberový Kolmogorovov-Smirnovov test

Pozrime sa najskôr na jeden náhodný výber $X_1, \ldots, X_n$ s neznámou distribučnou funkciou $F(x)$, ktorú odhadneme pomocou EDF $\widehat{F}_n(x)$. Chceme otestovať hypotézu $H_0 : F(x) = F_0(x)$ proti alternatíve $H_1 : F(x) \neq F_0(x)$, kde $F_0(x)$ je distribučná funkcia nejakého spojitého rozdelenia. Testovacia štatistika jednovýberového K-S testu má tvar

$$K_n = \sup_{-\infty<x<\infty} |\widehat{F}_n(x) - F_0(x)|, \tag{2.1}$$

a pri jednostrannej alternatíve

$$\begin{aligned} K_n^- &= \sup_{-\infty<x<\infty} (\widehat{F}_n(x) - F_0(x)), \\ K_n^+ &= \sup_{-\infty<x<\infty} (F_0(x) - \widehat{F}_n(x)). \end{aligned}$$

Pri testovaní štatistických hypotéz sa snažíme hľadať postupy, ktoré nám pomôžu mimimalizovať hodnoty pravdepodobností chýb 1. a 2. druhu, čo však súčasne pre obe hodnoty nie je možné dosiahnuť. Zníženie jednej môže mať za následok zvýšenie druhej a naopak. Aby sa štatistiky K-S testu dali používať v praxi, bolo treba docieliť odpovedajúcu hladinu významnosti, tzn. obmedziť chybu 1. druhu nejakým dopredu daným malým číslom α, $\alpha \in (0, 1)$.

Kolmogorovov tvar štatistiky K-S testu umožnil tabelizovanie rozdelenia K_n, Smirnov získal tvar pre K_n^+ a K_n^-.

Veta 2.1. *Pre* $\forall x > 0$

1. *Kolmogorov (1933)*

$$\lim_{n\to\infty} P(K_n \leq x) = \sum_{j=-\infty}^{\infty} (-1)^j e^{-2j^2x^2}, \tag{2.2}$$

2. Smirnov (1941)

$$\lim_{n\to\infty} P(K_n^+ > x) = e^{-2x^2}. \tag{2.3}$$

Ako uvádza Serfling [9], Kolmogorovov pôvodný dôkaz (2.2) bol založený na vyjadrení funkcie $\widehat{F}_n$ pomocou podmieneného Poissonovho procesu. Smirnov použil presné vyjadrenie $P(K_n^+ > x)$, aby ukázal (2.3).

K-S test spočíva v porovnaní najväčšej vzdialenosti EDF $\widehat{F}_n(x)$ a teoretickej distribučnej funkcie $F_0(x)$ s príslušnou kritickou hodnotou $K_n(\alpha)$. Pre menšie rozsahy n je uvedená v tabuľkách (napr. Sheskin [10], str. 1167) alebo ju môžeme aproximovať číslom

$$K_n(\alpha) = \sqrt{\frac{1}{2n}\ln\frac{2}{\alpha}} \qquad \text{pre} \quad n \geq 30.$$

Veľké hodnoty K_n také, že $K_n \geq K_n(\alpha)$, svedčia v neprospech H_0 a túto hypotézu môžeme na základe K-S testu zamietnuť na príslušnej hladine α.

Príklad 1. Niekoľkými meraniami sa zisťovala dĺžka špeciálnej automobilovej súčiastky. Výsledky (viď [11]) v cm boli nasledovné:

5,83	5,80	5,85	5,88	5,84	5,83	5,98	5,78
5,82	5,81	5,86	5,82	5,90			

Úlohou je na hladine $\alpha = 0,05$ overiť hypotézu o normalite rozdelenia náhodného výberu nameraných hodnôt, ak predpísaná dĺžka súčiastky je $5,8$ cm a povolená odchýlka od normy je $0,01$ cm.

Riešenie:
Príklady budeme riešiť pomocou programu R 2.8.1., kde je väčšina štatistických testov priamo naprogramovaná. Celý nasledujúci postup podrobnejšie vysvetlíme, aby sme si ukázali, aké uplatnenie majú vzťahy z prvej časti práce.

Všetkých hodnôt je 13, teda rozsah výberu je $n = 13$. Najskôr namerané hodnoty usporiadame podľa veľkosti od najmenšej po najväčšiu. To využijeme a zo vzťahu (1.3) spočítame EDF náhodného výberu $\widehat{F}_{13}(x)$. Ďalším krokom je zistiť, aké hodnoty nadobúda teoretická distribučná funkcia $F_0(x)$ normálneho rozdelenia so strednou hodnotou $5,8$ a rozptylom $0,01$. Potom už ľahko spočítame samotné rozdiely medzi $\widehat{F}_{13}(x)$ a $F_0(x)$, stačí sa obmedziť na body skokov. Podľa vzťahu (2.1) práve suprémum týchto rozdielov určí hodnotu štatistiky K_{13}.

V programe R 2.8.1. zapíšeme namerané dáta do vektoru x a zoradíme ich príkazom `sort(x)`. Spočítame hodnoty empirickej a teoretickej distribučnej funkcie príkazmi `ecdf(y)` a `pnorm(y,EX,sqrt(var))`, kde `y=sort(x)` symbolizuje vektor zoradených dát podľa veľkosti, `EX` $= 5,8$ a `var` $= 0,01$. Výsledky zapíšeme do tabuľky 2.1.

i	$x_{(i)}$	$F_0(x_{(i)})$	$\widehat{F}_n(x_{(i)})$	$\widehat{F}_n(x_{(i+1)})$	Δ_i^1	Δ_i^2
1	5,78	0,4207	0	1/13	0,3438	0,4207
2	5,80	0,5000	1/13	2/13	0,3461	**0,4231**
3	5,81	0,5398	2/13	3/13	0,3091	0,3860
4	5,82	0,5792	3/13	5/13	0,1946	0,3485
5	5,82	0,5792	5/13	5/13	0,1946	0,1946
6	5,83	0,6179	5/13	7/13	0,0794	0,2333
7	5,83	0,6179	7/13	7/13	0,0794	0,0794
8	5,84	0,6554	7/13	8/13	0,0400	0,1170
9	5,85	0,6915	8/13	9/13	0,0008	0,0761
10	5,86	0,7257	9/13	10/13	0,0435	0,0334
11	5,88	0,7881	10/13	11/13	0,0580	0,0189
12	5,90	0,8413	11/13	12/13	0,0817	0,0048
13	5,98	0,9641	12/13	1	0,0359	0,0410

Tab. 2.1 Hodnoty empirickej a teoretickej distribučnej funkcie

Pre vysvetlenie[1] $\Delta_i^1 = |\widehat{F}_n(x_{(i+1)}) - F_0(x_{(i)})|$ a $\Delta_i^2 = |\widehat{F}_n(x_{(i)}) - F_0(x_{(i)})|$. Vyznačená hodnota v poslednom stĺpci odpovedá testovej štatistike K_{13}, teda najväčšiemu rozdielu medzi $\widehat{F}_{13}(x)$ a $F_0(x)$. Z tabuľky 2.1 sa dá vyčítať, že toto suprémum nastalo v bode $x_{(2)}$.

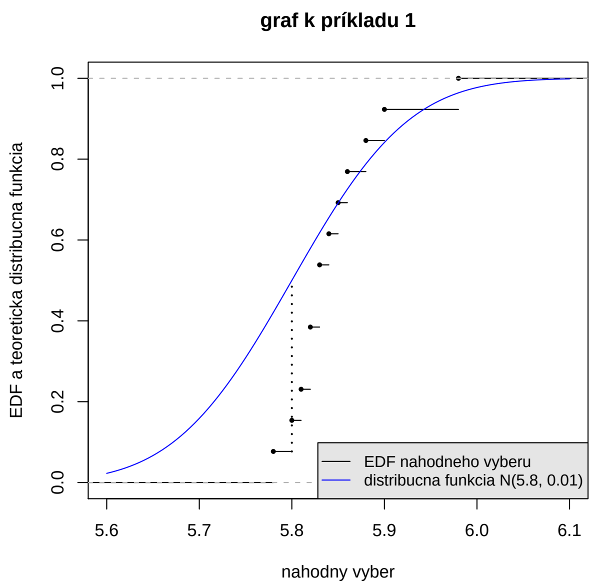
Samotný K-S test vykonáme príkazom `ks.test(y,pnorm,EX,sqrt(var))`, čo znamená, že program R prevedie jednovýberový K-S test s obojstrannou alternatívou a otestuje, či dáta uložené vo vektore `y` pochádzajú z normálneho rozdelenia s distribučnou funkciou označenou ako `pnorm`, so strednou hodnotou `EX` a rozptylom `var`. Výsledkom budú nasledujúce výstupné parametre, kde `D` udáva hodnotu testovacej štatistiky K_{13} a parameter `p-value` označuje p-hodnotu:

[1] V tabuľke 2.1 namiesto výrazu $|\widehat{F}_n(x_i) - F_0(x_i)|$ by sme mohli rovnako napísať $|\widehat{F}_n(x_i - \epsilon) - F_0(x_i)|$ a $|\widehat{F}_n(x_{i+1}) - F_0(x_i)|$ by sme mohli vymeniť za $|\widehat{F}_n(x_i + \epsilon) - F_0(x_i)|$, kde v oboch prípadoch $\epsilon \to 0$. V prvom prípade však môžeme ϵ vynechať kvôli spojitosti $\widehat{F}_n(x)$ sprava. Túto vlastnosť sme uviedli v 1. kapitole.

One-sample Kolmogorov-Smirnov test
data: y
D = 0.4231, p-value = 0.01905
alternative hypothesis: two-sided
Warning message: In ks.test(y, pnorm, EX, sqrt(var)) :
cannot compute correct p-values with ties

Na konci výstupu vypísal program R varovné hlásenie, že pomocou funkcie `ks.test` nie je možné vypočítať správnu p-hodnotu kvôli zhodným pozorovaniam. Teraz však tento problém ignorujeme, lebo o zamietnutí H_0 rozhodneme na základe porovnania kritickej hodnoty a hodnoty K_{13}.

Záver: Kritickú hodnotu jednovýberového K-S testu $K_{13}(0,05) = 0,3610$ sme vyhľadali v tabuľke A21, Sheskin [10]. Keďže $K_{13} = 0,4231$, potom $K_{13} > K_{13}(0,05)$. Na základe tohto výsledku zamietame nulovú hypotézu na hladine $\alpha = 0,05$.



Obr. 2.1 Graf $\widehat{F}_{13}(x)$ a $F_0(x)$

Do grafu na obrázku (2.1), ktorý vizuálne posudzuje zhodu EDF a teoretickej distribučnej funkcie z príkladu 1 je vertikálnou prerušovanou čiarou vyznačené $\sup_{-\infty<x<\infty} |\widehat{F}_n(x) - F_0(x)|$.

2.2.2 Modifikácia – Lillieforsov test

Občas sa stane, že nepoznáme základné parametre rozdelenia, z ktorého pochádza pozorovaný náhodný výber, tzn. strednú hodnotu a rozptyl (napr. u normálneho rozdelenia) a musíme tieto parametre odhadnúť. Tým sa zmení rozdelenie štatistiky K_n, za inak rovnakých podmienok dôjde k zmenšeniu sily testu a vtedy je vhodné použiť namiesto K-S testu jednu z jeho modifikácií pre distribučnú funkciu s odhadovanými parametrami.

Lillieforsov test, pomenovaný po Hubertovi Lillieforsovi, má rovnaký tvar testovacej štatistiky ako jednovýberový K-S test. Používa sa však na testovanie nulovej hypotézy, ktorá hovorí, že skúmané dáta pochádzajú z normálneho rozdelenia, pričom stredná hodnota a rozptyl nie sú bližšie špecifikované. Tieto parametre pred prevedením testu odhadneme nasledujúcim spôsobom: strednú hodnotu pomocou výberového priemeru

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{2.4}$$

a rozptyl pomocou výberového rozptylu

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2. \tag{2.5}$$

Z tohto dôvodu využíva Lillieforsov test odlišné kritické hodnoty ako jednovýberový K-S test. Je možné ich vyhľadať v tabuľkách, napr. v knihe Sheskin [10]. Ak testovaciu štatistiku Lillieforsovho testu označíme symbolom K_n^*, môžeme ju alternatívne vyjadriť ako

$$K_n^* = \max\left\{K_n^{+*}, K_n^{-*}\right\},$$

kde

$$K_n^{+*} = \max_{i=1,\dots,n}\left\{\frac{i}{n} - p_{(i)}\right\}, \quad K_n^{-*} = \max_{i=1,\dots,n}\left\{p_{(i)} - \frac{i-1}{n}\right\},$$

pričom $p_{(i)} = \Phi\left(\left(X_{(i)} - \bar{X}\right)/\sigma\right)$, kde Φ je distribučná funkcia normálneho rozdelenia so strednou hodnotou $\bar{X}$ a smerodajnou odchýlkou σ.

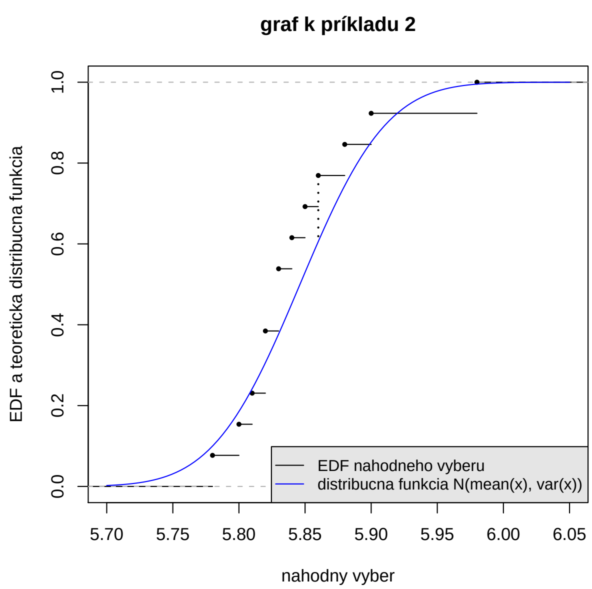
Príklad 2. Pozmeníme zadanie predchádzajúceho príkladu nasledovne: Niekoľkými meraniami sa zisťovala dĺžka špeciálnej automobilovej súčiastky.

Výsledky v cm viď príklad 1. Úlohou je na hladine $\alpha = 0,05$ overiť hypotézu, že dáta pochádzajú z normálneho rozdelenia.

Riešenie:
Normálne rozdelenie, z ktorého za hypotézy H_0 pochádza náhodný výber nameraných hodnôt, nemá presne určené hodnoty parametrov. Na testovanie nulovej hypotézy použijeme Lillieforsov test, ktorý prevedieme pomocou programu R 2.8.1. Uvažujeme hladinu $\alpha = 0,05$. Pri niektorých výpočtoch bude možné využiť aj vzťahy z predchádzajúcich kapitol.

Strednú hodnotu a rozptyl odhadneme výberovými parametrami. V tomto prípade, keď je rozsah malý, nie je ťažké spočítať výberový priemer zo vzťahu (2.4) a výberový rozptyl z (2.5). V programe R použijeme v prvom prípade príkaz `mean(x)` a v druhom `var(x)`.

Analogicky, ako v predchádzajúcom príklade, by sme spočítali vzdialenosti medzi EDF $\widehat{F}_n(x)$ a teoretickou distribučnou funkciou (označenou napr. ako $F_L(x)$) normálneho rozdelenia, tentokrát s odhadnutými parametrami. Skôr než prejdeme na prevedenie samotného testu, si pre lepšiu predstavu môžeme graficky v programe R vytvoriť vizuálne porovnanie $\widehat{F}_n(x)$ a $F_L(x)$ aj s najväčšou vzdialenosťou v jednom z 13 skokov. Na obr. 2.1 sme ju vyznačili prerušovanou vertikálnou čiarou.



Obr. 2.1 Graf $\widehat{F}_{13}(x)$ a $F_L(x)$

Na prevedenie Lillieforsovho testu v programe R potrebujeme špeciálnu knižnicu `nortest`[2], ktorá obsahuje viacero jednovýberových testov na testovanie hypotézy, že náhodný výber pochádza z normálneho rozdelenia. Zadaním príkazu `lillie.test(x)`, kde `x` je vektor, do ktorého sme uložili 13 nameraných hodnôt, dostaneme výsledok jednovýberového testu s nasledujúcim výstupom.

```
Lilliefors (Kolmogorov-Smirnov) normality test
data: x
D = 0.1634, p-value = 0.4478
```

`D` označuje hodnotu testovacej štatistiky K_n^* a `p-value` je p-hodnota. V knihe Sheskin [10] v tabuľke A22 kritických hodnôt Lillieforsoveho testu nájdeme hodnotu $K_{13}^*(0,05) = 0,234$.

Záver: Keďže $K_{13}^* < K_{13}^*(0,05)$, tzn. že hodnota testovacej štatistiky neleží v kritickom obore, nezamietame nulovú hypotézu na uvažovanej hladine $\alpha = 0,05$.

2.2.3 Dvojvýberový Kolmogorovov-Smirnovov test

K-S test nedosahuje dobré výsledky len pri testovaní rozdelenia jedného náhodného výberu. Poznáme aj tzv. dvojvýberový K-S test, ktorý meria maximálne odchýlky dvoch empirických distribučných funkcií.

Budeme sa zaoberať zrovnávaním dvoch nezávislých náhodných výberov $X_1, \dots, X_n$ a $Y_1, \dots, Y_m$ so spojitými distribučnými funkciami $F(x)$ a $G(x)$. Testujeme hypotézu $H_0 : F(x) = G(x)$ proti alternatíve $H_1 : F(x) \neq G(x)$. Táto alternatíva hovorí, že F a G nie su rovnaké, ale nemôžeme prijať žiadny zvláštny predpoklad o spôsobe, akým sa odlišujú.

Nech $\widehat{F}_n(x)$ je EDF prvého výberu a $\widehat{G}_m(x)$ je EDF druhého výberu, kde $\widehat{F}_n(x)$ a $\widehat{G}_m(x)$ sú zavedené vzťahom (1.2). Dvojvýberový K-S test má štatistiku tvaru

$$K_{n,m} \quad = \sup_{-\infty<x<\infty} |\widehat{F}_n(x) - \widehat{G}_m(x)|,$$

[2] http://cran.r-project.org/web/packages/nortest/

a pri jednostrannej alternatíve

$$
\begin{aligned}
K^{+}_{n,m} &= \sup_{-\infty<x<\infty} (\widehat{F}_n(x) - \widehat{G}_m(x)), \\
K^{-}_{n,m} &= \sup_{-\infty<x<\infty} (\widehat{G}_m(x) - \widehat{F}_n(x)).
\end{aligned}
$$

Za platnosti H_0 teda podľa vety 1.1 $K_{n,m} \to 0$ s pravdepodobnosťou 1 pre $m \to \infty, n \to \infty$. Toto tvrdenie si upresníme neskôr vo vete 2.2..
Označíme kritickú hodnotu dvojvýberového K-S testu symbolom $K_{n,m}(\alpha)$. Hypotézu H_0 zamietame na hladine významnosti α, ak $K_{n,m} \geq K_{n,m}(\alpha)$.

Ako ukázala Jurečková [6], K-S test je tzv. poradový test. To znamená, že závisí len na poradiach.

Označme $(X_1, \ldots, X_n, Y_1, \ldots, Y_m) = (Z_1, \ldots, Z_N)$, kde $N = m+n$. Usporiadajme pozorovania $Z_1, \ldots, Z_N$ do postupnosti $Z_{(1)} < \cdots < Z_{(N)}$. Vektor $(Z_{(1)}, \ldots, Z_{(N)})$ je vektor poradí týchto pozorovaní. Nech $V_j = 0$ ak Z_j pochádza z prvého výberu a $V_j = 1$ ak Z_j pochádza z druhého výberu. Pretože $\widehat{F}_n(x)$ a $\widehat{G}_m(x)$ sú neklesajúce, schodovité a skoky nastávajú len v niektorých z bodov $Z_j, j = 1, \ldots, N$, stačí hľadať suprémum rozdielu $\widehat{F}_n(x) - \widehat{G}_m(x)$ v týchto bodoch. Teda platí

$$
\begin{aligned}
\widehat{F}_n(Z_{(j)}) &= \frac{1}{n}\left[(1-V_1) + (1-V_2) + \cdots + (1-V_j)\right], \\
\widehat{G}_m(Z_{(j)}) &= \frac{1}{m}\left(V_1 + \cdots + V_j\right), \qquad j = 1, \ldots, N.
\end{aligned}
$$

Z týchto dvoch rovností ľahko odvodíme rozdiel

$$
\widehat{F}_n(Z_{(j)}) - \widehat{G}_m(Z_{(j)}) = \frac{n+m}{mn}\left[j\frac{m}{m+n} - V_1 - \cdots - V_j\right], \quad j = 1, \ldots, N \tag{2.6}
$$

a spočítame

$$
K_{n,m} = \frac{m+n}{mn} \cdot \sup_{1\leq j\leq N} \left|j\frac{m}{m+n} - V_1 - \cdots - V_j\right|, \quad j = 1, \ldots, N. \tag{2.7}
$$

Keďže veličiny $V_1, \ldots, V_N$ závisia len na poradiach, plynie zo vzťahu (2.7), že aj $K_{n,m}$ je len funkciou poradí a že kritické hodnoty nezávisia na F, pretože rozdelenie vektora poradí za platnosti H_0 nezávisí na F. Podobne sa dá odvodiť vzťah (2.7) aj pre $K^+_{n,m}$ a $K^-_{n,m}$. Absolútne hodnoty stačí zameniť za obyčajné zátvorky. Tabuľky kritických hodnôt uviedol napr. Sheskin [10], str. 1169-1170.

V nasledujúcej vete si ukážeme, že rozdelenie testovacej štatistiky $K_{n,m}$ pri $n \to \infty, m \to \infty$ nezávisí na rozdelení pôvodného výberu.

Veta 2.2. *Za platnosti H_0*

$$\lim_{n\to\infty, m\to\infty} P\left(\sqrt{\frac{nm}{n+m}}K_{n,m} < x\right) = K(x), \quad x > 0, \tag{2.8}$$

kde $K(x) = 1 + 2\sum_{k=1}^{\infty}(-1)^k e^{-2k^2x^2}$.

Dôkaz: Viď Borovkov [3], str. 384, Veta 1. □

Funkciu $K(x)$ môžeme aproximovať pomocou počiatočných členov $1-2e^{-2x^2}$. Potom dostaneme

$$P\left(K_{n,m} < \frac{x}{\sqrt{\frac{mn}{m+n}}}\right) \doteq 1 - 2e^{-2x^2},$$

pričom výraz na pravej strane sa rovná $1-\alpha$ pre $x = x' = \sqrt{\frac{1}{2}\ln\frac{2}{\alpha}}$. Pre väčší rozsah n, m sa kritická hodnota $K_{n,m}$ aproximuje číslom

$$K'_{n,m}(\alpha) = \frac{x'}{\sqrt{\frac{mn}{m+n}}} = \sqrt{\frac{m+n}{2mn}\ln\frac{2}{\alpha}}.$$

H_0 zamietneme na príslušnej hladine α, keď $K_{n,m} \geq K'_{n,m}(\alpha)$.

V čom teda spočíva Kolmogorov-Smirnov test? Podľa Anděla [1] môže byť praktické prevedenie nasledovné: z výberov $X_1, \ldots, X_n$ a $Y_1, \ldots, Y_m$ sa spočítajú $\widehat{F}_n(x)$, $\widehat{G}_m(x)$ a veličina $K_{n,m}$. Pre malé m a n sa $K_{n,m}$ porovná s presnými kritickými hodnotami $K_{n,m}(\alpha)$.

Pre veľké m a n sa využije veta 2.2 a položí sa $x_0 = \sqrt{\frac{mn}{m+n}}K_{n,m}$. Vypočíta sa hodnota $K(x_0)$. Ak $K(x_0) \geq 1-\alpha$, potom zamietneme H_0 na asymptotickej hladine α.

2.3 Cramerov-von Misesov test

2.3.1 Jednovýberový Cramerov-von Misesov test

Vráťme sa k počiatočnému problému testovania hypotézy $H_0 : F = F_0$, kde náhodný výber n nezávislých pozorovaní X_i má rozdelenie s distribučnou funkciou $F(x)$ a EDF $\widehat{F}_n(x)$. Pri testovaní hypotézy H_0 proti alternatíve $H_1 : F \neq F_0$, kde F_0 je distribučná funkcia nejakého spojitého rozdelenia, Smirnov modifikoval štatistiku, ktorú vytvorili Cramer a von Mises, do tvaru

$$W_n^2 = n \int\limits_{-\infty}^{\infty} \psi(F_0(x)) \left[\widehat{F}_n(x) - F_0(x)\right]^2 dF_0(x), \tag{2.9}$$

kde $\psi(u), 0 \leq u \leq 1$ je nezáporná váhová funkcia. V literatúre (viď napr. Durbin [5]) je venovaná zvláštna pozornosť hlavne dvom variantom štatistiky W_n^2 so špeciálnym tvarom funkcie ψ. Pokiaľ $\psi(F_0(x)) = [F_0(x)\,(1 - F_0(x))]^{-1}$, potom ide o tzv. Andersonovu-Darlingovu štatistiku, ktorej sa budeme venovať neskôr. Keď $\psi = 1$, potom sa jedná o tzv. Cramerovu-von Misesovu (ďalej C-vM) štatistiku[3] a píšeme

$$W_n^2 = n \int\limits_{-\infty}^{\infty} \left[\widehat{F}_n(x) - F_0(x)\right]^2 dF_0(x).$$

V praxi sa často využíva nasledujúci alternatívny zápis

$$W_n^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left(p_{(i)} - \frac{2i-1}{2n}\right), \tag{2.10}$$

kde $p_{(i)} = F_0(x_{(i)})$ a $x_{(i)}$ je v poradí i-tá najvyššia hodnota náhodného výberu pre $i = 1, \ldots, n$.

Hypotézu H_0 zamietneme na hladine α, ak hodnota štatistiky W_n^2, ktorú dostaneme prevedením testu, je väčšia ako kritická hodnota jednovýberového C-vM testu $W_n^2(\alpha)$, ktorú nájdeme v tabuľkách, napr. Stephens [12].

[3] Je známy aj tzv. L_1-variant Cramerovho-von Misesovho testu so štatistikou M tvaru $M = \sqrt{n} \int\limits_{-\infty}^{\infty} |\widehat{F}_n(x) - F_0(x)|\, dF_0(x)$. Viac v článku Schmid, Trede [8].

Veta 2.3. *Za platnosti H_0 platí*

$$\lim_{n\to\infty} P\left(W_n^2 < \lambda\right) = P\left(\sum_{i=1}^{\infty} \frac{S_i^2}{i^2\pi^2} < \lambda\right), \qquad \lambda > 0,$$

kde $S_1, S_2, \ldots$ sú nezávislé náhodné veličiny s normálnym normovaným rozdelením $N(0,1)$.

Dôkaz: Viď Serfling [9], str. 64. $\square$

2.3.2 Modifikácia – Andersonov-Darlingov test

Nech je hypotéza a alternatíva rovnaká ako v predchádzajúcej podkapitole. Ako sme už uviedli, keď váhovú funkciu $\psi(F_0(x))$ vo vzťahu (2.9) nahradíme výrazom $\left[F_0(x)\,(1-F_0(x))\right]^{-1}$, dostaneme tvar testovacej štatistiky jednovýberového Andersonovho-Darlingovho (ďalej A-D) testu

$$A^2 = n \int_{-\infty}^{\infty} \frac{\left[\widehat{F}_n(x) - F_0(x)\right]^2}{F_0(x)\,(1-F_0\,(x))}\, dF_0(x).$$

Nasledujúci alternatívny zápis, vypočítaný na základe postupu, ktorý podrobnejšie popísal Durbin [5], sa v praxi pre výpočetné účely používa častejšie. Teda

$$A^2 = -n - \sum_{i=1}^{n} (2i-1)\left(\ln z_i + \ln\left(1 - z_{n+1-i}\right)\right)/n, \tag{2.11}$$

kde $z_i = F_0(x_{(i)})$ a $x_{(i)}$ je i-tá hodnota pozorovania náhodného výberu s rozsahom n, $i = 1, \ldots, n$.

Hypotézu H_0 zamietneme na príslušnej hladine α, ak platí $A^2 \geq A^2(\alpha)$, pričom $A^2(\alpha)$ je kritická hodnota jednovýberového A-D testu. Tabuľka kritických hodnôt je uvedená v článku Stephens [12].

Príklad 3. Vezmime si zadanie z príkladu 2. Na hladine $\alpha = 0,05$ otestujeme hypotézu H_0, že namerané dĺžky súčiastky pochádzajú z normálneho rozdelenia. Použijeme jednovýberový C-vM a A-D test.

Riešenie:
V programe R 2.8.1. načítame knižnicu `nortest`, ktorá oba tieto testy obsahuje. Do vektora `x` uložíme namerané hodnoty. Príkazom `cvm.test(x)`

otestujeme nulovú hypotézu, že 13 hodnôt pochádza z normálneho rozdelenia s odhadnutými parametrami, ktoré pomocou príkazov `mean(x)` pre odhad strednej hodnoty a `var(x)` pre odhad rozptylu, vypočíta program R sám. Potom na tej istej hladine $\alpha = 0,05$ prevedieme A-D test príkazom `ad.test(x)`, ktorý tiež používa výberové parametre. Výsledky zapíšeme do spoločnej tabuľky 2.2.

TEST	test. štat.	krit. hodnota	p-hodnota	rozhodnutie o H_0
C-vM	$W_{13}^2 = 0,9040$	$W_{13}^2(0,05) = 0,1213$	$p = 0,1365$	H_0 nezamietame
A-D	$A^2 = 0,5719$	$A^2(0,05) = 0,6830$	$p = 0,1112$	H_0 nezamietame

Tab. 2.2 Výstupné parametre testov v programe R na hladine $\alpha = 0,05$

Výsledok C-vM testu, prevedeného v programe R, je zaznamenaný v druhom riadku tabuľky 2.2 spolu s kritickou hodnotou $W_{13}^2(0,05) = 0,1213$. Vypočítali sme ju na základe postupu, ktorý uviedol Stephens [12], ako

$$W_n^2(0,05) = \frac{W^2(0,05)}{(1 + 0,5/n)}, \tag{2.12}$$

kde n udáva rozsah výberu. V našom prípade dosadíme $n = 13$ a $W^2(0,05) = 0,126$. Na základe porovnania $0,9040 = W_{13}^2 \leq W_{13}^2(0,05) = 0,1213$ nezamietame nulovú hypotézu na uvažovanej hladine $\alpha = 0,05$.

Výstupné parametre a rozhodnutie o H_0 na základe A-D testu nájdeme v treťom riadku tabuľky 2.2. Keďže $0,5719 = A^2 \leq A^2(0,05) = 0,6830$, nezamietame H_0 na hladine $\alpha = 0,05$.

2.3.3 Dvojvýberový Cramerov-von Misesov test

Podľa Andersona [2] je v prípade dvojvýberového testu C-vM test považovaný za silnejší ako dvojvýberový K-S test. Nech $X_1, \ldots, X_n$ je náhodný výber s distribučnou funkciou $F(x)$ a EDF $\widehat{F}_n(x)$ a $Y_1, \ldots, Y_m$ je náhodný výber s distribučnou funkciou $G(x)$ a EDF $\widehat{G}_m(x)$. Uvažujeme hypotézu, podľa ktorej oba výbery majú to isté (nešpecifikované) spojité rozdelenie. Potom štatistiku takéhoto dvojvýberového Cramer-von Misesovho testu proti obojstrannej alternatíve $H_1 : F \neq G$ zapíšeme v tvare

$$W_{n,m} = \left[\frac{nm}{n+m}\right] \int_{-\infty}^{\infty} \left[\widehat{F}_n(x) - \widehat{G}_m(x)\right]^2 d\widehat{H}_{n+m}(x), \tag{2.13}$$

kde $\widehat{H}_{n+m}(x)$ je EDF obidvoch výberov dohromady, tzn. $(n+m)\widehat{H}_{n+m}(x) = n\widehat{F}_n(x)+m\widehat{G}_m(x)$ a dáva každému pozorovaniu $X_1,\ldots,X_n,Y_1,\ldots,Y_m$ váhu $1/(n+m)$. Integrál (2.13) sa dá prepísať na tvar

$$W_{n,m} = \left[\frac{nm}{(n+m)^2}\right]\left\{\sum_{i=1}^{n}\left[\widehat{F}_n(X_i)-\widehat{G}_m(X_i)\right]^2+\sum_{j=1}^{m}\left[\widehat{F}_n(Y_j)-\widehat{G}_m(Y_j)\right]^2\right\}.$$

S použitím vzťahu (2.6) testovaciu štatistiku prepíšeme nasledovne

$$W_{n,m} = \frac{1}{nm}\sum_{j=1}^{n+m}\left(j\frac{n}{n+m}-V_1-\cdots-V_j\right)^2.$$

Odtiaľto, rovnako ako v predchádzajúcom dvojvýberovom K-S teste, plynie, že štatistika Cramer-von Misesovho testu závisí len na poradiach. Tabuľky kritických hodnôt sú väčšinou uvádzané len pre malé výbery, (viď napr. Anderson [2]). Pri väčšom rozsahu m a n sa musia dopočítavať, najčastejšie metódou Monte Carlo alebo bootstrapom. Pre veľké m,n je možné odvodiť limitné kritické hodnoty.

Veta 2.4. *Za platnosti H_0 a pre ľubovoľné $\lambda > 0$ platí*

$$\lim_{n\to\infty,m\to\infty} P\left(W_{n,m}<\lambda\right) = P\left(\sum_{j=1}^{\infty}\frac{U_j^2}{j^2\pi^2}<\lambda\right),$$

kde $U_1,U_2,\ldots$ sú nezávislé náhodné veličiny s normálnym normovaným rozdelením $N(0,1)$.

Dôkaz: Dôkaz je možné nájsť v knihe Jurečková [6], str. 64., Veta 6. □

Poznámka: Ako uviedol Lehmann [7], Kolmogorovov-Smirnovov, Cramerov-von Misesov a Andersonov-Darlingov test sú konzistentné testy.

Kapitola 3

Simulácie testov o normalite rozdelenia náhodného výberu v programe R 2.8.1

3.1 Náhodný výber z normálneho rozdelenia $N(0,1)$

V programe R 2.8.1. vygenerujeme príkazom `rnorm(n,EX,sqrt(varX))` náhodný výber z normálneho rozdelenia so strednou honotou $EX = 0$ a rozptylom $varX = 1$. Generovanie prevedieme trikrát pre rôzne rozsahy n. Aby sme ten istý výber o n dátach mohli použiť viackrát, je vhodné si ho uložiť do vektoru napr. `x.norm` priradením `x.norm = rnorm(n,EX,sqrt(varX))`. Pri opätovnom generovaní by sme však dostali iné dáta (napríklad, ak by sme simuláciu spustili opäť od začiatku). Ak sa takej situácii potrebujeme vyhnúť, pomôže nám príkaz `set.seed(k)` s parametrom `k`, zadaný hneď na začiatku, po ktorom `rnorm(n,EX,sqrt(varX))` vygeneruje vždy tie isté dáta pre daný rozsah n aj pri opätovnom spustení programu, čo má viacero výhod.

Zvolíme `set.seed(1111)`. Na uvažovanej hladine $\alpha = 0,05$ postupne otestujeme hypotézu H_0, že náhodný výbery s rozsahmi 20, 100 a 1000 pochádza z normálneho rozdelenia proti alternatíve H_1, že výber pochádza z iného než normálneho rozdelenia. Príkazom `require(nortest)` sa otvorí knižnica `nortest` s ďalšími testami pre normalitu. Potom už môžeme použiť K-S test pre normálne rozdelenie s parametrami 0 a 1 a Lillieforsov test, C-vM test a A-D pre rozdelenie s odhadnutými parametrami `mean(x.norm)` namiesto strednej hodnoty a `var(x.norm)` namiesto rozptylu, ktoré program R vypočíta na základe nasimulovaných dát. Medzi vstupné dáta pre každý test patrí testovaný vektor `x.norm`. Po ich prevedení zapíšeme hodnoty testovacích štatistík, kritické hodnoty, p-hodnoty a rozhodnutie o zamietnutí či nezamietnutí H_0 do tabuľky 2.3. Kritické hodnoty K-S a Lillieforsovho testu uviedol Sheskin [10], pre C-vM (použijeme vzťah (2.12)) a A-D test viď Stephens [12].

TEST	test. štat.	krit. hodnota	p-hodnota	rozhodnutie o H_0
n = 20				
K-S	$0,1888$	$0,2940$	$0,4218$	H_0 nezamietame
Lill	$0,1477$	$0,2340$	$0,3029$	H_0 nezamietame
C-vM	$0,1007$	$0,1229$	$0,1016$	H_0 nezamietame
A-D	$0,6415$	$0,7040$	$0,0806$	H_0 nezamietame
n = 100				
K-S	$0,1004$	$0,1360$	$0,2655$	H_0 nezamietame
Lill	$0,0473$	$0,0860$	$0,8422$	H_0 nezamietame
C-vM	$0,0311$	$0,1254$	$0,8306$	H_0 nezamietame
A-D	$0,2051$	$0,7540$	$0,8690$	H_0 nezamietame
n = 1000				
K-S	$0,0198$	$0,0430$	$0,8281$	H_0 nezamietame
Lill	$0,0152$	$0,0280$	$0,8310$	H_0 nezamietame
C-vM	$0,0408$	$0,1259$	$0,6664$	H_0 nezamietame
A-D	$0,3094$	$0,7870$	$0,5567$	H_0 nezamietame

Tab. 2.3 Výsledky testov na uvažovanej hladine $\alpha = 0,05$ o nulovej hypotéze, že náhodný výber s rozsahom 20, 100 a 1000 pochádza z normálneho rozdelenia

Z tabuľky 2.3 vidíme, že testy nezamietli nulovú hypotézu o normálnom rozdelení dát. Tieto výsledky sme čakali, lebo dáta boli vygenerované ako náhodný výber z $N(0,1)$. Svedčí o tom aj obrázok (1.2), na ktorom je znázornená distribučná funkcia normovaného normálneho rozdelenia, od ktorej sa empirická distribučná funkcia náhodného výberu so zväčšujúcim sa rozsahom dát odkláňa čoraz menej.

3.2 Náhodný výber z Laplaceovho rozdelenia $L(1,1)$

Pre program R je Laplaceovo rozdelenie obsiahnuté v knižnici `VGAM`[1]. Za n postupne dosadíme hodnoty 20, 100 a 1000. Použijeme `set.seed(1111)` a vygenerujeme náhodný výber o rozsahu n pomocou jednoduchého príkazu `rlaplace(n,location,scale)`, kde *location* a *scale* sú parametre rozdelenia a obe majú hodnotu 1. Hodnoty si uložíme do vektoru `x.lap`. Samostatne pre každé n otestujeme hypotézu H_0, že náhodný výber pochádza z normálneho rozdelenia a použijeme Lillieforsov, C-vM a A-D test pre jeden výber s odhadnutými parametrami a vstupnými dátami `x.lap`. Uvažujeme hladinu významnosti $\alpha = 0,05$. Aby sme získali prehľad o výsledných dátach, zapíšeme ich do tabuľky:

TEST	test. štat.	krit. hodnota	p-hodnota	rozhodnutie o H_0
n = 20				
Lill	$0,1742$	$0,1900$	$0,11510$	H_0 nezamietame
C-vM	$0,1159$	$0,1229$	$0,06227$	H_0 nezamietame
A-D	$0,8005$	$0,7040$	$0,03142$	H_0 zamietame
n = 100				
Lill	$0,0861$	$0,0886$	$0,06470$	H_0 zamietame
C-vM	$0,7575$	$0,1254$	$0,01194$	H_0 zamietame
A-D	$4,529$	$0,7540$	$0,01861$	H_0 zamietame
n = 1000				
Lill	$0,0682$	$0,0280$	$7,349.10^{-12}$	H_0 zamietame
C-vM	$1,6844$	$0,1259$	$1,772.10^{-09}$	H_0 zamietame
A-D	$9,1383$	$0,7870$	$2,200.10^{-16}$	H_0 zamietame

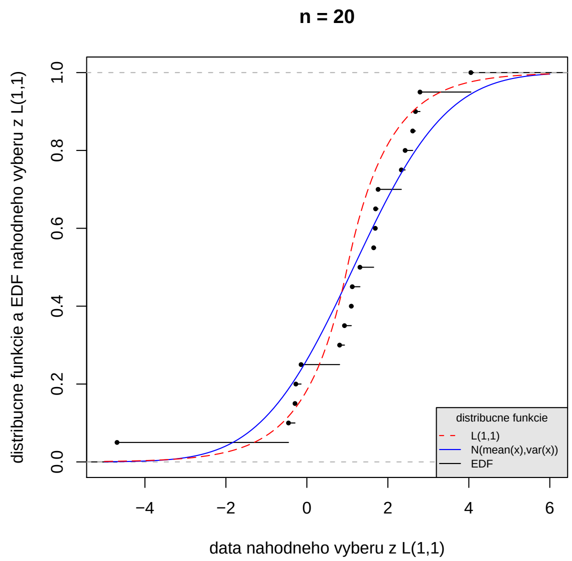
Tab. 2.4 Výsledky testov na uvažovanej hladine $\alpha = 0,05$ o nulovej hypotéze, že náhodný výber s rozsahom 20, 100 a 1000 pochádza z normálneho rozdelenia

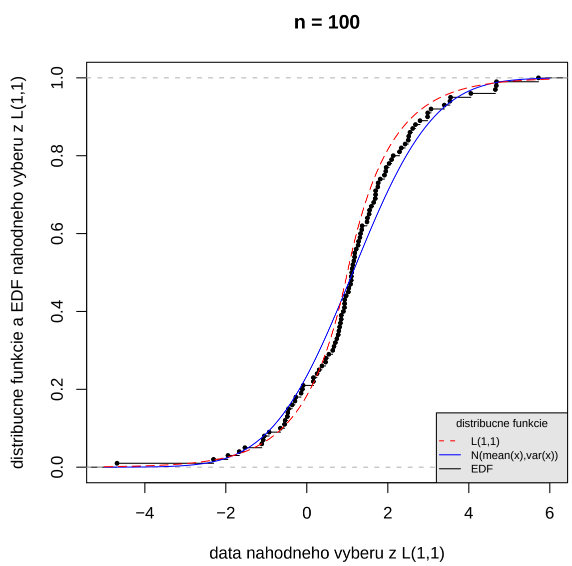
Kritické hodnoty sme získali rovnako ako v podkapitole 3.1. Pre väčšie rozsahy, teda $n = 100$ a $n = 1000$ všetky testy zamietli nulovú hypotézu na uvažovanej hladine $\alpha = 0,05$, čo znamená, že potvrdili alternatívu, že náhodný výber je z iného než normálneho rozdelenia. V našej simulácii šlo o Laplacovo rozdelenie $L(1,1)$. Pri simulácii s 20 pozorovaniami je však

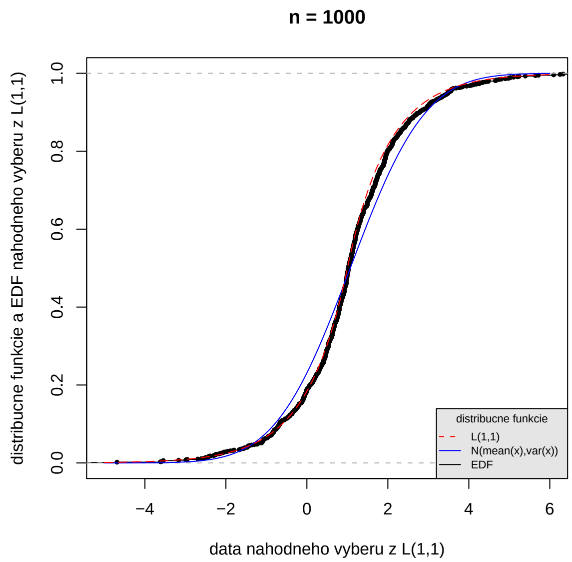
[1] http://cran.r-project.org/web/packages/VGAM/

výsledok iný. Mohli by sme povedať, že sa prvé dva testy, Lillieforsov a C-vM, dopustili chyby 2. druhu, tzn. nezamietli hypotézu H_0 napriek tomu, že platí H_1. Tretí test, tj. A-D test, nulovú hypotézu zamietol aj pri $n = 20$.

Pre grafické posúdenie empirickej a teoretickej distribučnej funkcie sú uvedené nasledujúce grafy.







Obr. 1.2 Grafy distribučnej funkcie normálneho rozdelenia s výberovými parametrami (vykreslená modrou súvislou čiarou), distribučnej funkcie Laplaceovho rozdelenia $L(1,1)$ (vykreslená červenou prerušovanou čiarou) a príslušnej empirickej distribučnej funkcie s naznačeným rozsahom nasimulovaného náhodného výberu

Záver

Ústredným motívom celej práce bola empirická distribučná funkcia. Dôkladne sme popísali jej základné a asymptotické vlastnosti, zamerali sme sa na jej vzťah ku skutočnej distribučnej funkcii náhodného výberu z nejakého rozdelenia. Ukázali sme, že pri pevne zvolenej reálnej hodnote x je empirická distribučná funkcia nestranným a konzistentným odhadom tejto funkcie. Navyše, pri veľkom rozsahu, je jej dobrou aproximáciou. V príkladoch sme odvodili empirické odhady rôznych momentov a intervalový odhad neznámej distribučnej funkcie.

Zvláštnu pozornosť sme venovali testom dobrej zhody, menovite šlo o Kolmogorovov-Smirnovov (K-S), Lillieforsov, Cramerov-von Misesovov (C-vM) a Andersonov-Darlingov test, ktoré testujú hypotézu, že náhodný výber pochádza z nejakého rozdelenia so spojitou distribučnou funkciou. Zaoberali sme sa aj dvojvýberovým K-S a C-vM testom, ktoré skúmajú hypotézu, že dva nezávislé náhodné výbery pochádzajú z toho istého rozdelenia. Táto časť práce sa zameriavala na popisnú stránku testov, zámerom nebolo hodnotenie, ktorý test je lepšie použiť v akej situácii, alebo ktorý z nich má väčšie vypovedajúce schopnosti ako ostatné. Hlavným cieľom bakalárskej práce bolo čitateľa oboznámiť s využitím empirickej distribučnej funkcie v matematickej štatistike pomocou menovaných testov, ktoré sú založené práve na empirickej distribučnej funkcii a majú nemalé využitie pri testovaní spomínaných hypotéz.

Z tohto dôvodu bola práca doplnená nielen príkladmi, ale aj simuláciami jednovýberových testov s riešeniami a grafmi prevedenými pomocou programu R 2.8.1, vďaka ktorým si čitateľ lepšie predstaví správanie empirickej distribučnej funkcie voči teoretickej distribučnej funkcii rozdelenia, z ktorého pochádzal nasimulovaný náhodný výber.

Appendix A

Základné definície a tvrdenia

A.1 Limitná časť

V tejto časti si vysvetlíme základné definície a pojmy z teórie pravdepodobnosti. Nech $(\Omega, \mathcal{A}, P)$ je pravdepodobnostný priestor, na ktorom je definovaná postupnosť reálnych náhodných veličín $X_1, X_2, \ldots$ a reálna náhodná veličina X.

Definícia A.1. Povieme, že X_n konvergujú k X s pravdepodobnosťou 1 (skoro všade, silno), ak $P(\lim_{n\to\infty} X_n = X) = 1$. Píšeme $X_n \xrightarrow{s.v.} X$.

Definícia A.2. Povieme, že X_n konvergujú k X v pravdepodobnosti, ak $\lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1$ pre každý $\epsilon > 0$. Píšeme $X_n \xrightarrow{P} X$.

Definícia A.3. Nech X_n má distribučnú funkciu F_n a X má distribučnú funkciu F. Povieme, že X_n konvergujú k X v distribúcii (konvergujú slabo), ak $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ pre každé x také, že F_X je spojitá v x. Píšeme $X_n \xrightarrow{d} X$.

Vety A.1-4 sú uvedené v rovnakom alebo podobnom znení spolu s dôkazmi (prípadne s odkazmi na dôkazy) v knihe Anděl [1], dodatok B.

Veta A.1 (*Čebyševova nerovnosť*)**.** *Nech reálna náhodná veličina X má strednú hodnotu μ a konečný rozptyl σ^2. Potom pre každý $\epsilon > 0$ platí, že*

$$P(|X - \mu| > \epsilon) \leq \frac{\sigma^2}{\epsilon^2}.$$

Veta A.2 (*Silný zákon veľkých čísel pre rovnako rozdelené náhodné veličiny*)**.** *Nech $X_1, \ldots, X_n$ sú nezávislé rovnako rozdelené náhodné veličiny s konečnou strednou hodnotou μ. Ak $n \to \infty$, potom $\frac{1}{n} \sum_{i=1}^{n} X_n \to \mu$ s pravdepodobnosťou 1.*

Veta A.3 je jedným z vyjadrení centrálnej limitnej vety (označujeme CLV) pre veličiny s binomickým rozdelením. Hovorí, že pri veľkom počte nezávislých pokusov konverguje binomické rozdelenie k normálnemu v distribúcii.

Veta A.3 (*Moivreova-Laplaceova*)**.** *Nech pre každé $n \geq 1$ je $Y_n \sim Bi(n, p)$, kde $p \in (0, 1)$. Potom*

$$\lim_{n \to \infty} P\left(\frac{Y_n - np}{\sqrt{np\,(1-p)}} \leq x\right) = \phi(x) \qquad x \in \mathbb{R},$$

kde $\phi(x)$ je distribučná funkcia normovaného normálneho rozdelenia $N(0, 1)$.

Veta A.4 (*Cramérova-Sluckého*)**.** *Nech $X_1, X_2, \ldots$ je postupnosť náhodných veličín s distribučnými funkciami $F_1, F_2, \ldots$, nech F je distribučná funkcia a k konštanta. Nech F_n konvergujú slabo k F. Nech pre postupnosť náhodných veličín $Y_1, Y_2, \ldots$ platí, že $Y_n \xrightarrow{P} k$. Definujeme $R_n = X_n + Y_n$, $S_n = X_n Y_n$ a $T_n = X_n / Y_n$, kde F_n^R, F_n^S a F_n^T sú po rade distribučné funkcie veličín R_n, S_n a T_n. Potom $F_n^R(x)$ konvergujú slabo k $F(x - k)$. Ak $k > 0$, potom F_n^S konvergujú slabo k $F(x/k)$ a F_n^T konvergujú slabo k $F(kx)$.*

A.2 Teória odhadov

Definícia A.4. Nech $X_1, \ldots, X_n$ sú nezávislé rovnako rozdelené náhodné veličiny z rozdelenia s distribučnou funkciou F. Potom hovoríme, že ide o náhodný výber z tohto rozdelenia.

Chceme odhadnúť nejakú charakteristiku rozdelenia F_0, na základe náhodného výberu $X_1, \ldots, X_n$. Predpokladáme, že $F_0 \in \mathcal{F}$, kde $\mathcal{F}$ je množina uvažovaných rozdelení, ktoré nazývame model (rodina). Všeobecne $\mathcal{F} = \left\{F_\theta, \theta \in \Theta \subset \mathbb{R}^k\right\}$, kde θ je neznámy parameter.

Definícia A.5. Bodovým odhadom $\hat{\theta}_n$ parametra $\theta_0 = t(F_0)$ rozumieme ľubovoľnú merateľnú funkciu dát $\hat{\theta}_n = T_n(X_1, \ldots, X_n) \in \mathbb{R}^k$, pre ľubovoľné $k \in \mathbb{N}$.

Definícia A.6. Povieme, že $\hat{\theta}_n$ je nestranný odhad parametru $\theta_0 = t(F_0)$ v modeli $\mathcal{F}$, ak platí $E\hat{\theta}_n = \theta_0$, $\forall F_0 \in \mathcal{F}$.

Definícia A.7. Povieme, že $\hat{\theta}_n$ je konzistentný odhad parametru $\theta_0 = t(F_0)$ v modeli $\mathcal{F}$, ak platí $\hat{\theta}_n \xrightarrow{P} \theta_0$, $\forall F_0 \in \mathcal{F}$.

Definícia A.8. Interval $B \equiv (C_L(X_1, \ldots, X_n), C_U(X_1, \ldots, X_n)) \in \mathbb{R}$ sa nazýva intervalový odhad parametru $\theta_0 = t(F_0) \in \mathbb{R}$ s pravdepodobnosťou pokrytia $1 - \alpha \in (0, 1)$, práve keď $P_{F_0}(B \ni \theta_0) = 1 - \alpha$. Intervalu B sa tiež hovorí interval spoľahlivosti a $1 - \alpha$ sa nazýva spoľahlivosť.

A.3 Teória testovania hypotéz

Uvažujme nulovú hypotézu $H_0 : F \in \mathcal{F}$ a alternatívu $H_1 : G \in \mathcal{G}$, kde $\mathcal{F}$ a $\mathcal{G}$ sú dva disjunktné modely rozdelení. Nech $\mathbf{X} = (X_1, \ldots, X_n)$ je náhodný vektor.

Test hypotézy je daný testovacou štatistikou $S(\mathbf{X}) \in \mathbb{R}$ a kritickým oborom $C \in \mathbb{R}$. Pokiaľ $S(\mathbf{X}) \in C \Rightarrow$ zamietame H_0 v prospech H_1, pokiaľ $S(\mathbf{X}) \notin C \Rightarrow$ nezamietame H_0.

Pri testovaní sa môžeme dopustiť dvoch chýb: chyby 1. druhu, keď zamietneme H_0 aj napriek tomu, že H_0 platí a chyby 2. druhu, keď nezamietneme H_0 napriek tomu, že H_0 neplatí. Kvalitu testu potom posudzujeme podľa pravdepodobností, s ktorými tieto dve možné chyby nastanú.

Definícia A.9. Nech $\alpha \in (0, 1)$ je dopredu stanovené číslo. Potom číslo $\sup_{P \in \mathcal{F}} P(S(\mathbf{X}) \in C) = \alpha$ sa nazýva hladina významnosti testu (skrátene hladina testu). Pre približné testy (asymptotické) je ich hladina α približná (asymptotická) pre $n \to \infty$.

Definícia A.10. Nech $P \in \mathcal{G}$. Číslo $\beta(P) = P(S(\mathbf{X}) \in C)$ sa nazýva sila testu proti alternatíve.

Definícia A.11. Test nazveme konzistentným práve vtedy, keď $\beta(P) \geq \alpha$ pre $\forall P \in \mathcal{G}$.

Definícia A.12. P-hodnotou p nazývame najmenšiu hladinu testu, pri ktorej ešte nezamietneme hypotézu H_0.

Tzn. že ak je $p < \alpha$, potom zamietame nulovú hypotézu H_0 na hladine významnosti α.

Literatúra

[1] Anděl, J.: *Základy matematické statistiky*, Matfyzpress, Praha, 2005.

[2] Anderson, T. W.: *On the Distribution of the Two-Sample Cramér-von Mises Criterion*, The Annals of Mathematical Statistics **33** (1962), 1148–1159.

[3] Borovkov, A. A.: *Mathematical Statistics*, Gordon and Breach, Amsterdam, 1998.

[4] Dupač, V., Hušková, M.: *Pravděpodobnost a matematická statistika*, Karolinum, Praha, 2005.

[5] Durbin, J.: *Distribution theory for tests based on the sample distribution function*, Society for Industrial and Applied Mathematics, Philadelphia, Pa, 1973.

[6] Jurečková, J.: *Pořadové testy*, Státní pedagogické nakladatelství, Praha, 1981.

[7] Lehmann, E.L.: *Testing statistical hypotheses*, Wiley, New York, 1997.

[8] Schmid, F., Trede, M.: *An L_1-variant of the Cramer-von Mises test*, Statistics Probability Letters **26** (1996) 91–96.

[9] Serfling, R. J.: *Approximation Theorems of Mathematical Statistics*, Wiley, New York, 2002.

[10] Sheskin, D. J.: *Parametric and Nonparametric Statistical Procedures*, Chapman & Hall/CRC, Boca Raton, London, New York, Washington, D.C., 2004.

[11] Stavebná fakulta, Slovenská technická univerzita v Bratislave. Kapitola 12. Dostupné na: http://mpm.svf.stuba.sk/∼tunega06/OLD/ [Náhľad na stránku dňa 22.2.2009].

[12] Stephens, M. A.: *EDF Statistics for Goodness of Fit and Some Comparisons*, Journal of the American Statistical Association, **69** (1974), 730–737.