Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE

Lukáš Vašek

Vztah empirické a teoretické distribuční funkce (jednorozměrný případ)

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: RNDr. Vlasta Kaňková, CSc.,
ÚTIA AV ČR

Studijní program: Matematika
Studijní obor: Obecná matematika

2006

Úvodem své bakalářské práce bych rád poděkoval
RNDr. Vlastě Kaňkové, CSc. za odborné vedení, cenné rady a připomínky.

Prohlašuji, že jsem svou bakalářskou práci napsal samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce a jejím zveřejňováním.

V Praze dne 20. 5. 2006 Lukáš Vašek

Obsah

Název práce: Vztah empirické a teoretické distribuční funkce (jednorozměrný případ)
Autor: Lukáš Vašek
Katedra: Katedra pravděpodobnosti a matematické statistiky
Vedoucí bakalářské práce: RNDr. Vlasta Kaňková, CSc., ÚTIA AV ČR
e-mail vedoucího: kankova@utia.cas.cz

Abstrakt: V předložené práci pojednáme o vztahu mezi empirickou a teoretickou distribuční funkcí. V první kapitole nejprve uvedeme základní vlastnosti těchto funkcí. Pomocí silného zákona velkých čísel ukážeme, že pro dostatečně velký rozsah výběru se bude empirická distribuční funkce „blížit" skoro jistě k teoretické distribuční funkci. Dokážeme však více. Podle Glivenkovy věty bude empirická distribuční funkce konvergovat stejnoměrně k teoretické distribuční funkci s pravděpodobností 1. O rychlosti této konvergence pojednáme ve druhé kapitole ve větách Kolmogorova a Smirnova. Na základě těchto vět jsou založeny testy dobré shody.

Klíčová slova: Empirická distribuční funkce, Glivenko, Kolmogorov.

Title: An Analysis of the Relationship between Empirical and Theoretical Distribution Functions (one-dimensional case)
Author: Lukáš Vašek
Department: Department of Probability and Mathematical Statistics
Supervisor: RNDr. Vlasta Kaňková, CSc., ÚTIA AV ČR
Supervisor's e-mail address: kankova@utia.cas.cz

Abstract: In the present work we analyze the relationship between empirical and theoretical distribution functions. In the first chapter we introduce the basic attributes of these functions. Due to the strong law of large numbers we show that the empirical distribution function of large random sample approximates the theoretical distribution function. We prove more as well. In accordance with the Glivenko theorem, the supremum of the absolute value of the difference between empirical and theoretical distribution functions converges to 0 with probability 1. In the second chapter we deal with the speed of this convergence according to the theorems of Kolmogorov and Smirnov. On these theorems are based tests on the distribution function.

Keywords: Empirical Distribution Function, Glivenko, Kolmogorov.

Kapitola 1

Teoretická a empirická distribuční funkce

1.1 Úvod

V této práci se budeme zabývat vztahem mezi empirickou a teoretickou distribuční funkcí. Ke každému náhodnému výběru, který pochází z rozdělení s distribuční funkcí $F(x)$, můžeme určit empirickou distribuční funkci. Ukážeme, že na základě silného zákona velkých čísel bude s rostoucím rozsahem výběru konvergovat empirická distribuční funkce k teoretické s pravděpodobností 1. Díky Glivenkově větě se dozvíme, že tato konvergence je dokonce stejnoměrná na celé reálné ose. O rychlosti konvergence však pojednáme až ve druhé kapitole ve větách, které dokázali ruští matematikové Smirnov a Kolmogorov. Na základě těchto vět budeme moci ověřovat, zda daný náhodný výběr pochází z rozdělení se spojitou distribuční funkcí F. Dále budeme moci testovat, zda dva náhodné výběry pocházejí z téhož rozdělení. Pokud neznáme teoretickou distribuční funkci, budeme moci na základě empirické distribuční funkce vymezit pás, ve kterém se bude nacházet neznámá distribuční funkce s předem požadovanou pravděpodobností.

1.2 Základní pojmy a označení

Nechť Ω je neprázdná množina a $\mathcal{A}$ je nějaká σ-algebra podmnožin mno-žiny Ω. Potom dvojici $(\Omega, \mathcal{A})$ nazveme měřitelný prostor. Prvky množiny Ω značíme ω a nazýváme elementární jevy. Prvky σ-algebry $\mathcal{A}$ nazýváme jevy (viz [6], strana 47).

Je-li $(\Omega, \mathcal{A})$ měřitelný prostor, definujeme pravděpodobnost P jako míru na $\mathcal{A}$ s vlastnostmi

$$\begin{aligned} &P(A) \geq 0, && A \in \mathcal{A}; \\ &P(\Omega) = 1, P(\emptyset) = 0; \\ &\textstyle\sum_{n=1}^{\infty} P(A_n) = P(\cup_{n=1}^{\infty} A_n), \end{aligned}$$

pro $\{A_n\}$ posloupnost po dvou disjunktních jevů ([7], strana 55).

Trojice $(\Omega, \mathcal{A}, P)$ se nazývá pravděpodobnostní prostor. Nechť $\mathbf{R}^n$ značí n-rozměrný eukleidovký prostor. σ-algebru generovanou systémem všech otevřených podmnožin v $\mathbf{R}^n$ značíme $\mathcal{B}^n$ a její prvky nazýváme borelovské množiny.

Měřitelné zobrazení X: $(\Omega, \mathcal{A}) \to (\mathbf{R}^1, \mathcal{B}^1)$ nazýváme reálná náhodná veličina, měřitelné zobrazení $\mathbf{X}$: $(\Omega, \mathcal{A}) \to (\mathbf{R}^n, \mathcal{B}^n)$ nazýváme n-rozměrný reálný náhodný vektor $(n = 1, 2, \ldots)$ ([7], strana 42).

Pravidlo, které každé hodnotě nebo množině hodnot z každého intervalu přiřazuje pravděpodobnost, že náhodná veličina nabude této hodnoty nebo hodnoty z tohoto intervalu, nazýváme rozdělením náhodné veličiny ([4], strana 57).

Řekneme, že náhodné veličiny $X_1, X_2, \ldots, X_n$ jsou nezávislé ([7], strana 142), jestliže pro každé $(x_1, x_2, \ldots, x_n) \in \mathbf{R}^n$ platí

$$P(X_1 < x_1, X_2 < x_2, \ldots, X_n < x_n) = \prod_{i=1}^{n} P(X_i < x_i).$$

n-rozměrný náhodný vektor $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, kde náhodné veličiny $X_1, \ldots, X_n$ jsou vzájemně nezávislé a všechny mají stejné rozdělení, nazýváme náhodným výběrem rozsahu n z tohoto rozdělení ([4], strana 149).

Reálná náhodná veličina X má alternativní rozdělení, jestliže nabývá hodnot 0 a 1 s pravděpodobností $1-p$ a p. Značíme $X \sim Alt(p)$, kde p je parametr rozdělení, $0 < p < 1$.

Binomické rozdělení má náhodná veličina X, která nabývá hodnot $0 \leq k \leq n$ s pravděpodobnostmi $\binom{n}{k} p^k (1-p)^{n-k}$. Značíme $X \sim Bi(n, p)$, kde $0 < p < 1$ a n je přirozené.

1.3 Teoretická distribuční funkce

Je-li X náhodná veličina, pak její distribuční funkcí nazveme funkci $F(x)$, která je pro všechna reálná x definována vztahem ([4], kapitola 5.2)

$$F(x) = P(X < x).$$

Zde $P(X < x)$ značí pravděpodobnost jevu, že X nabývá hodnoty menší než x. Je-li X náhodná veličina, která nabývá hodnot x_n s pravděpodobnostmi $p_n = P(X = x_n)$, kde $n = 1, 2, \ldots$, platí

$$F(x) = \sum_{x_n < x} p_n.$$

Sčítáme přes takové indexy n, pro něž je $x_n < x$. Distribuční funkce, pro kterou platí předchozí nerovnost, odpovídá diskrétnímu rozdělení.

Pokud existuje funkce $f(x)$ taková, že pro každé reálné x platí

$$F(x) = \int_{-\infty}^{x} f(t)dt,$$

pak říkáme, že distribuční funkce $F(x)$ odpovídá spojitému rozdělení. Funkci $f(x)$ nazýváme hustota náhodné veličiny X. Vlastnost, že $F(x)$ je distribuční funkce náhodné veličiny X budeme značit $F_X(x)$. Distribuční funkce je vždy neklesající, zleva spojitá, omezená a platí pro ni

$$\lim_{x \to -\infty} F(x) = 0, \qquad \lim_{x \to +\infty} F(x) = 1.$$

Poznamenejme, že v některé literatuře ([2], strana 188) je distribuční funkce definována tak, že je zprava spojitá. V této práci se však touto definicí zabývat nebudeme.

1.4 Empirická distribuční funkce

Nechť $X_1, \ldots, X_n$ je náhodný výběr rozsahu n s distribuční funkcí $F(x)$. Uspořádejme hodnoty $X_1, \ldots, X_n$ do neklesající posloupnosti

$$X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}.$$

Empirickou distribuční funkci $F_n(x)$ tohoto výběru ([5], strana 99) definujme takto:

$$F_n(x) = \begin{cases} 0 & \text{pro všechna } x \leq X_{(1)}; \\ \frac{k}{n} & \text{pro všechna } X_{(k)} < x \leq X_{(k+1)}, \qquad k = 1, 2, \ldots, n-1; \\ 1 & \text{pro všechna } x > X_{(n)}. \end{cases}$$

Empirická distribuční funkce je stejně jako teoretická distribuční funkce neklesající, zleva spojitá, omezená a platí

$$\lim_{x \to -\infty} F_n(x) = 0, \qquad \lim_{x \to +\infty} F_n(x) = 1.$$

$F_n(x)$ je po částech konstantní. Pokud jsou všechny hodnoty $X_1, X_2, \ldots, X_n$ od sebe různé, pak v každé z nich má $F_n(x)$ skok o velikosti $\frac{1}{n}$. Pokud se však hodnota X_i v souboru $X_1, \ldots, X_n$ (pro $i = 1, \ldots, n$) vyskytuje právě k-krát, pak $F_n(x)$ má v bodě $x = X_i$ skok o velikosti $\frac{k}{n}$. Poznamenejme, že empirická distribuční funkce závisí na rozsahu výběru a také na náhodě.

1.5 Vlastnosti empirické distribuční funkce

Mějme $X_1, \ldots, X_n$ náhodný výběr z rozdělení s distribuční funkcí F. Nechť x je dané reálné číslo. Definujme náhodnou veličinu $\xi_i(x)$ takto

$$\xi_i(x) = \begin{cases} 1 & \text{je-li } X_i < x; \\ 0 & \text{je-li } X_i \geq x, \qquad i = 1, 2, \ldots, n. \end{cases}$$

Potom empirická distribuční funkce náhodného výběru $X_1, \ldots, X_n$ bude

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \xi_i(x).$$

Pro konkrétní realizaci náhodného výběru je funkce $F_n(x)$ totožná s empirickou distribuční funkcí, kterou jsme zavedli v odstavci 1.4.

Náhodné veličiny $\xi_i(x)$ mají stejné rozdělení. Platí $nF_n(x) = \sum_{i=1}^{n} \xi_i(x)$. Pro libovolné i označme $p_n = P(\xi_i(x) = 1) = F_{\xi_i}(x)$. Potom zřejmě $\xi_i(x) \sim Alt(p_n)$.

Dále ukažme, že $nF_n(x) \sim Bi(n, p_n)$. To plyne z toho, že součin $n.F_n(x)$ je definován jako součet náhodných veličin s alternativním rozdělením s parametrem p_n. Platí

$$P\Big(\sum_{i=1}^{n} \xi_i(x) = k\Big) = \binom{n}{k} p_n^k (1 - p_n)^{n-k}, \qquad 0 \leq k \leq n.$$

Ukážeme, že pro dostatečně velký rozsah výběru bude empirická distribuční funkce konvergovat „velmi rychle" k teoretické distribuční funkci. Podle [3], strana 208, budeme pro dostatečně velké n prakticky přesvědčeni, že platí nerovnost

$$\sup_{-\infty < x < +\infty} |F_n(x) - F(x)| < \varepsilon,$$

pro každé $\varepsilon > 0$. Přesné tvrzení dokážeme v odstavci 1.7.

1.6 Kolmogorovův silný zákon velkých čísel

Silný zákon velkých čísel uvedeme ve tvaru pro nezávislé stejně rozdělené náhodné veličiny.

Věta 1.1 (Kolmogorovova) *Nechť $X_1, X_2, \ldots$ je posloupnost nezávislých stejně rozdělených náhodných veličin s konečnou střední hodnotou $EX_i = \mu$. Potom platí*

$$P\Big(\lim_{n \to +\infty} \frac{X_1 + \cdots + X_n}{n} = \mu\Big) = 1.$$

Důkaz této věty nalezneme v [6], věta 3, strana 336.

V následující větě ukážeme, že se s rostoucím n funkce $F_n(x)$ „blíží" k teoretické distribuční funkci $F(x)$ s pravděpodobností 1.

Věta 1.2 *Pro každé reálné x platí*

$$P\Big(\lim_{n\to+\infty} F_n(x) = F(x)\Big) = 1,$$

jinak řečeno $F_n(x) \to F(x)$ skoro jistě pro $n \to \infty$.

Důkaz. ([1], věta 8.9, strana 105) Využijeme definice empirické distribuční funkce z odstavce 1.5. Pro každé pevné x jsou veličiny $\xi_i(x)$ nezávislé stejně rozdělené. Platí

$$\mathrm{P}\big(\xi_i(x) = 1\big) = F(x) \qquad a \qquad \mathrm{E}\xi_i(x) = F(x).$$

Z Kolmogorovovy věty 1.1 platí

$$P\Big(\lim_{n\to+\infty} \frac{\xi_1(x) + \cdots + \xi_n(x)}{n} = \mathrm{E}\xi_i(x)\Big) = 1,$$

odkud již plyne dokazované tvrzení.

1.7 Glivenkova věta

V následující větě se dozvíme, že z dostatečně velkého náhodného výběru můžeme získat libovolně podrobnou informaci o distribuční funkci, neboť empirická distribuční funkce konverguje s pravděpodobností 1 stejnoměrně na celé reálné ose k distribuční funkci statistického souboru, vzrůstá-li rozsah výběru do nekonečna.

Věta 1.3 (Glivenkova) *Nechť $X_1, X_2, \ldots, X_n$ jsou nezávislé náhodné veličiny se stejným rozdělením, jehož distribuční funkce je $F(x)$. Nechť dále $F_n(x)$ značí empirickou distribuční funkci náhodného výběru $X_1, \ldots, X_n$. Označme*

$$D_n = \sup_{-\infty<x<+\infty} |F_n(x) - F(x)|.$$

Potom platí

$$P(\lim_{n\to\infty} D_n = 0) = 1.$$

Důkaz. ([2], strana 269; [6], strana 339) Uvažujme $F(x)$ distribuční funkci, která není degenerovaná. Tedy existují alespoň dva body x, pro něž pro každé $h > 0$ platí $F(x) < F(x+h)$. (viz [6], strana 161) Podle odstavců 1.3 a 1.4 víme, že $F(x)$ a $F_n(x)$ jsou neklesající, zleva spojité, omezené funkce nabývající hodnot mezi 0 a 1. Označme $F(x+0)$, $F_n(x+0)$ limity funkcí v bodě x zprava.

Zvolme libovolné M přirozené a nechť $k = 0, 1, \ldots, M$. Nyní rozdělme interval [0,1] na M stejných částí délky $\frac{1}{M}$. Protože $F(x)$ nemusí být v libovolném bodě spojitá, volme $x_{M,k}$ nejmenší x, které vyhovuje nerovnosti

$$F(x) \leq \frac{k}{M} \leq F(x+0).$$

Zřejmě platí $-\infty = x_{M,0} < x_{M,1} \leq x_{M,2} \leq \cdots \leq x_{M,M} \leq +\infty$, přičemž pro dostečně velké M je alespoň jedna z nerovností mezi $x_{M,1}$ a $x_{M,M}$ ostrá vzhledem k tomu, že $F(x)$ má alespoň dva body růstu.

Vyšetříme tedy rozdíl $|F_n(x) - F(x)|$ a také limitu tohoto rozdílu zprava pouze v bodech $x_{M,k}$.

Volme

$$D_n^{(1)} = \max_{1 \leq k \leq M} |F_n(x_{M,k}) - F(x_{M,k})|, D_n^{(2)} = \max_{1 \leq k \leq M} |F_n(x_{M,k}+0) - F(x_{M,k}+0)|$$

a označme $D_{n,MAX} = max(D_n^{(1)}, D_n^{(2)})$.

Pro $0 \leq k \leq M-1$ za předpokladu $x_{M,k} < x_{M,k+1}$ zřejmě platí

$$F(x_{M,k+1}) - F(x_{M,k}+0) \leq \frac{1}{M}.$$

Nyní využijeme vlastnosti, že $F_n(x)$ je neklesající. Jestliže
$x_{M,k} < x \leq x_{M,k+1}$, potom platí

$$F_n(x) \leq F_n(x_{M,k+1}) \leq F(x_{M,k+1}) + D_{n,MAX} \leq F(x) + \frac{1}{M} + D_{n,MAX}$$

a

$$F_n(x) \geq F_n(x_{M,k}+0) \geq F(x_{M,k}+0) - D_{n,MAX} \geq F(x) - \frac{1}{M} - D_{n,MAX}.$$

Předchozí úvahy nás vedou k nerovnosti $D_n \leq D_{n,MAX} + \frac{1}{M}$. Poznamenejme, že náhodné veličiny D_n a $D_{n,MAX}$ závisejí na ω, tedy předchozí nerovnosti platí skoro jistě.

Podle věty 1.2 je pro každé pevné x

$$P\Big(\lim_{n\to\infty} F_n(x) = F(x)\Big) = 1 \quad a \quad P\Big(\lim_{n\to\infty} F_n(x+0) = F(x+0)\Big) = 1.$$

Vidíme tedy, že $D_{n,MAX}$ konverguje k 0 skoro jistě, pro $n \to +\infty$, proto

$$P(\limsup_{n\to\infty} D_n > \frac{1}{M}) = 0$$

pro každé přirozené M, odkud již plyne

$$P(\lim_{n\to\infty} D_n = 0) = 1.$$

K úplnosti důkazu dodejme, že je-li distribuční funkce $F(x)$ degenerovaná, potom platí $F_n(x) = F(x)$ pro každé $n \geq 1$ a každé reálné x.

Glivenkovu větu můžeme vyslovit (podle [6], strana 340) také takto: K daným kladným číslům ε a δ existuje N_0 tak, že platí

$$P(\sup_{n\geq N_0} D_n \leq \varepsilon) > 1 - \delta.$$

Tímto vzorcem je zdůrazněn praktický význam Glivenkovy věty v matematické statistice. Ovšem na druhé straně nám nepodává žádnou informaci o závislosti čísla N_0 na ε a δ. Tento nedostatek odstraníme větami Kolmogorova (věta 2.2) a Smirnova (věta 2.1).

1.8 Konvergence v distribuci

Distribuční funkce je podle odstavce 1.3 jednoznačně určena pravděpodobnostní mírou P. Slabou konvergenci pravděpodobnostních měr tedy budeme definovat přímo pro distribuční funkce.

Pouze v tomto odstavci budeme distribuční funkci reálné náhodné veličiny X_k značit $F_k(x)$ $(k = 1, 2 \ldots)$. Dále nechť $F(x)$ je distribuční funkce reálné náhodné veličiny X. Řekneme, že F_k konvergují slabě k F ([7], strana 200), je-li splněna

$$\lim_{k \to +\infty} \int_{\mathbf{R}} f dF_k = \int_{\mathbf{R}} f dF,$$

pro každou f spojitou omezenou funkci na $\mathbf{R}$.

Jak je dokázáno v [7], strana 219, lze slabou konvergenci distribučních funkcí vyjádřit také následujícím způsobem. Řekneme, že distribuční funkce $F_k(x)$ konverguje slabě k $F(x)$, je-li splněno

$$\lim_{k \to \infty} F_k(x) = F(x)$$

pro každý bod $x \in \mathbf{R}$, v němž je funkce $F(x)$ spojitá. Značíme

$$F_k(x) \longrightarrow^w F(x).$$

Nechť X_k a X jsou náhodné veličiny, jejichž distribuční funkce jsou $F_k(x)$ a $F(x)$. Pokud $F_k(x) \longrightarrow^w F(x)$, potom X_k konvergují k X v distribuci. Tuto vlastnost budeme značit $X_k \longrightarrow^{\mathcal{D}} X$ ([2], strana 329).

Vztah konvergence v distribuci náhodných veličin X_k a X a slabé konvergence k nim příslušných distribučních funkcí $F_k(x)$ a $F(x)$ je ekvivalentní, jak ukazuje následující věta.

Věta 1.4 *Nechť X_k jsou reálné náhodné veličiny s distribuční funkcí $F_k(x)$ a nechť X je reálná náhodná veličina s distribuční funkcí $F(x)$, $(k = 1, 2, \ldots)$. Potom platí*

$$X_k \longrightarrow^{\mathcal{D}} X \qquad \Leftrightarrow \qquad F_k(x) \longrightarrow^w F(x).$$

Důkaz nalezneme v [7], strana 219.

Kapitola 2

Jednovýběrové a dvouvýběrové testy

2.1 Limitní věty pro jeden výběr

Nejprve uvedeme dvě důležité věty, které vyšetřují maximální odchylku empirické a teoretické distribuční funkce.

Zavedeme předpoklad (A): Nechť $X_1, X_2, \ldots, X_n$ jsou nezávislé stejně rozdělené náhodné veličiny se spojitou distribuční funkcí $F(x)$ a nechť $F_n(x)$ je empirická distribuční funkce odpovídající náhodnému výběru $X_1, \ldots, X_n$. Podle [6], strana 423, platí tyto věty.

Věta 2.1 (Smirnova) *Za předpokladu (A) platí*

$$\lim_{n\to+\infty} P\Big(\sqrt{n} \sup_{-\infty<x<+\infty} (F_n(x) - F(x)) < y\Big) = \begin{cases} 1 - e^{-2y^2} & \text{pro } y > 0, \\ 0 & \text{jinak.} \end{cases}$$

Věta 2.2 (Kolmogorova) *Za předpokladu (A) platí*

$$\lim_{n\to+\infty} P\Big(\sqrt{n} \sup_{-\infty<x<+\infty} |F_n(x) - F(x)| < y\Big) = \begin{cases} K(y) & \text{pro } y > 0, \\ 0 & \text{jinak,} \end{cases}$$

kde

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

$K(y)$ je tedy distribuční funkce veličiny $\sqrt{n}\sup_x |F_n(x) - F(x)|$ pro $n \to \infty$. Všimněme si, že věta 2.1 vyšetřuje supremum rozdílu $F_n(x)$ a $F(x)$, zatím co věta 2.2 se týká suprema absolutní hodnoty tohoto rozdílu. Dále uveďme, že limitní rozdělení v obou větách nezávisí na $F(x)$, pouze se předpokládá spojitost distribuční funkce $F(x)$.

2.2 Jednovýběrový Kolmogorovův-Smirnovův test

V tomto odstavci ukážeme, že empirické distribuční funkce náhodného výběru lze užít k ověření, zda náhodný výběr je z rozdělení s distribuční funkcí $F(x)$. Pojednáme o Kolmogorově-Smirnově jednovýběrovém testu. Budeme testovat v modelu s libovolnou spojitou distribuční funkcí $F(x)$.

Mějme $X_1, \ldots, X_n$ náhodný výběr z rozdělení se spojitou distribuční funkcí. Naším cílem bude testovat hypotézu H_0, že tato distribuční funkce je $F(x)$, proti alternativě H_1, že tomu tak není. Je-li $F_n(x)$ empirická distribuční funkce odpovídající výběru $X_1, \ldots, X_n$, položíme

$$D_n = \sup_{-\infty<x<+\infty} |F_n(x) - F(x)|.$$

Podle vět 1.2 a 1.3 svědčí velké hodnoty D_n proti hypotéze H_0. Pro náhodnou veličinu D_n stanovíme kritické hodnoty $D_n(\alpha)$ na hladině α, pro které platí vztah ([5], strana 101)

$$P\big(D_n < D_n(\alpha)\big) = 1 - \alpha.$$

Pro $n = 1, \ldots, 100$ nalezneme kritické hodnoty $D_n(\alpha)$ v [5] v tabulce 37. Při větších hodnotách n využijeme limitního vztahu ve větě 2.2. Zřejmě platí

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

Nyní podle věty 2.2 za předpokladu $y > 0$ platí

$$\lim_{n\to+\infty} P\big(\sqrt{n}D_n < y\big) = \lim_{n\to+\infty} P\Big(D_n < \frac{y}{\sqrt{n}}\Big) = 1 - 2\sum_{k=1}^{\infty}(-1)^{k+1}e^{-2k^2y^2}.$$

Položíme-li $D_n(\alpha) = y/\sqrt{n}$, dostáváme

$$\lim_{n\to+\infty} P\big(D_n < D_n(\alpha)\big) = 1 - 2\sum_{k=1}^{\infty}(-1)^{k+1}e^{-2k^2nD_n^2(\alpha)}.$$

Pro velké hodnoty n se kritické hodnoty $D_n(\alpha)$ aproximují pomocí prvního členu řady v předchozím výrazu. Řešením rovnice $1-\alpha = 1-2e^{-2nD_n^2(\alpha)}$ dostáváme ([1], strana 164)

$$D_n(\alpha) \doteq \sqrt{\frac{1}{2n}\ln\frac{2}{\alpha}}.$$

Hypotézu H_0 zamítáme, je-li $D_n \geq D_n(\alpha)$.

Nyní se zaměříme na jednostranný test. Jestliže očekáváme, že rozdíl $F_n(x) - F(x)$ bude převážně nezáporný, označíme

$$D_n^+ = \sup_{-\infty<x<+\infty}(F_n(x) - F(x)).$$

Budeme testovat hypotézu na hladině významnosti α, že náhodný výběr je z rozdělení se spojitou distribuční funkcí $F(x)$. Test je založen proti alternativě, že náhodný výběr pochází z rozdělení s distribuční funkcí, která nabývá ve všech bodech x hodnoty větší než $F(x)$. K náhodné veličině D_n^+ stanovíme kritické hodnoty $D_n^+(\alpha)$, pro které platí vztah ([5], strana 101)

$$D_n^+(\alpha) = D_n(2\alpha).$$

Zamítáme testovanou hypotézu, je-li $D_n^+ \geq D_n^+(\alpha)$.

Očekáváme-li, že rozdíl $F_n(x) - F(x)$ bude převážně nekladný, můžeme testovat hypotézu, že náhodný výběr pochází z rozdělení se spojitou distribuční funkcí $F(x)$, proti alternativě, že náhodný výběr pochází z rozdělení, jehož distribuční funkce je ve všech bodech menší než $F(x)$. Definujeme náhodnou veličinu

$$D_n^- = \sup_{-\infty<x<+\infty}(F(x) - F_n(x)).$$

Postup testu bude stejný, dosadíme-li v předchozím D_n^- namísto D_n^+ a $D_n^-(\alpha) = D_n^+(\alpha)$, kde $D_n^-(\alpha)$ je kritická hodnota náhodné veličiny D_n^-.

2.3 Limitní věty pro dva výběry

Máme-li dva náhodné výběry rozsahu n a m, můžeme podle následujících vět, které dokázal Smirnov, porovnáním empirických distribučních funkcí obou výběrů rozhodnout, zda pocházejí ze stejného statistického souboru.

Označme (B) následující předpoklady: Nechť $X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, Y_m$ jsou nezávislé náhodné veličiny. Nechť X_i $(i = 1, \ldots, n)$ resp. Y_j $(j = 1, \ldots, m)$ mají spojité distribuční funkce $F(x)$ resp. $G(x)$. Buď $F_n(x)$ empirická distribuční funkce výběru $X_1, \ldots, X_n$ a $G_m(x)$ empirická distribuční funkce výběru $Y_1, \ldots, Y_m$. Pokud při $n \to +\infty$, $m \to +\infty$ zůstává podíl $\frac{n}{m} = \tau$, kde $0 < \tau < \infty$, potom platí podle [5], strana 100, následující věty.

Věta 2.3 *Je-li $F(x) \equiv G(x)$ a platí-li (B), potom*

$$\lim_{n,m\to+\infty} P\Big(\sqrt{\frac{nm}{n+m}} \sup_{-\infty<x<+\infty} (F_n(x)-G_m(x)) < y\Big) = \begin{cases} 1 - e^{-2y^2} & \text{pro } y > 0, \\ 0 & \text{jinak.} \end{cases}$$

Věta 2.4 *Je-li $F(x) \equiv G(x)$ a platí-li (B), potom*

$$\lim_{n,m\to+\infty} P\Big(\sqrt{\frac{nm}{n+m}} \sup_{-\infty<x<+\infty} |F_n(x)-G_m(x)| < y\Big) = \begin{cases} K(y) & \text{pro } y > 0, \\ 0 & \text{jinak,} \end{cases}$$

kde

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

Pro stejně velké rozsahy výběrů, kde $n = m$, můžeme předchozí věty za předpokladu $F(x) \equiv G(x)$ vyjádřit takto:

$$\lim_{n\to+\infty} P\Big(\sqrt{\frac{n}{2}} \sup_{-\infty<x<+\infty} (F_n(x) - G_n(x)) < y\Big) = \begin{cases} 1 - e^{-2y^2} & \text{pro } y > 0, \\ 0 & \text{jinak} \end{cases}$$

a

$$\lim_{n\to+\infty} P\Big(\sqrt{\frac{n}{2}} \sup_{-\infty<x<+\infty} |F_n(x) - G_n(x)| < y\Big) = \begin{cases} K(y) & \text{pro } y > 0, \\ 0 & \text{jinak.} \end{cases}$$

2.4 Dvouvýběrový Kolmogorovův-Smirnovův test

Nechť $X_1, \ldots, X_n$ je náhodný výběr z rozdělení se spojitou distribuční funkcí $F(x)$ a nechť $Y_1, \ldots, Y_m$ je na něm nezávislý náhodný výběr z rozdělení se spojitou distribuční funkcí $G(x)$. Nechť $F_n(x)$ a $G_m(x)$ značí empirické distribuční funkce obou výběrů. Budeme testovat hypotézu $H_0 : F = G$ proti alternativě $H_1 : F \neq G$.

Označme

$$D_{n,m} = \sup_{-\infty<x<+\infty} |F_n(x) - G_m(x)| \qquad a \qquad M = \sqrt{\frac{n.m}{n+m}}.$$

Lze očekávat, že velké hodnoty náhodné veličiny $D_{n,m}$ povedou k zamítnutí hypotézy H_0. Pro $D_{n,m}$ stanovíme kritickou hodnotu $D_{n,m}(\alpha)$ na hladině α, ze vztahu

$$P\big(D_{n,m} < D_{n,m}(\alpha)\big) = 1 - \alpha.$$

Pro malé rozsahy výběrů nalezneme kritické hodnoty $D_{n,m}(\alpha)$ v [1] v tabulkách T9 a T10 nebo v [5] v tabulce 38. Pro větší hodnoty n a m použijeme limitní větu 2.4. Pro $y > 0$ platí

$$\lim_{n,m\to\infty} P\big(\sqrt{M} D_{n,m} < y\big) = \lim_{n,m\to\infty} P\Big(D_{n,m} < \frac{y}{\sqrt{M}}\Big) = 1 - 2\sum_{k=1}^{\infty} (-1)^{k+1} e^{-2k^2y^2}.$$

Položíme-li $D_{n,m}(\alpha) = y/\sqrt{M}$, potom

$$\lim_{n,m\to+\infty} P\big(D_{n,m} < D_{n,m}(\alpha)\big) = 1 - 2\sum_{k=1}^{\infty} (-1)^{k+1} e^{-2k^2 n D^2_{n,m}(\alpha)}.$$

Obdobně jako v odstavci 2.2 aproximujeme předchozí rovnici pomocí prvního členu uvedené řady. Řešíme tedy $1 - \alpha = 1 - 2e^{-2MD^2_{n,m}(\alpha)}$, což nás vede k vyjádření ([1], strana 106)

$$D_{n,m}(\alpha) \doteq \sqrt{\frac{1}{2M} \ln \frac{2}{\alpha}}.$$

Hodnoty Kolmogorovy funkce $K(y)$ uvedené v odstavcích 2.1 a 2.3 lze nalézt v [5] v tabulce 39.

Hypotézu H_0 zamítáme, je-li $D_{n,m} \geq D_{n,m}(\alpha)$.

Dále ukážeme jednostranný test hypotézy. Pokud nás zajímají zejména nezáporné hodnoty rozdílu mezi $F_n(x)$ a $G_m(x)$, potom můžeme testovat hypotézu, že dva náhodné výběry pocházejí ze stejného rozdělení se spojitou distribuční funkcí. Test je založen proti alternativě, že první výběr pochází z rozdělení, jehož distribuční funkce bude nabývat ve všech bodech větší hodnoty než distribuční funkce, která určuje rozdělení, z něhož pochází druhý náhodný výběr. Označme

$$D^+_{n,m} = \sup_{-\infty<x<+\infty} (F_n(x) - G_m(x)).$$

K náhodné veličině $D^+_{n,m}$ nalezneme kritickou hodnotu $D^+_{n,m}(\alpha)$ na hladině významnosti α. Platí vztah

$$D^+_{n,m}(\alpha) = D_{n,m}(2\alpha).$$

Hypotézu zamítneme, je-li $D^+_{n,m} \geq D^+_{n,m}(\alpha)$.

V případě, že očekáváme spíše nekladné hodnoty rozdílu mezi $F_n(x)$ a $G_m(x)$, potom označme

$$D^-_{n,m} = \sup_{-\infty<x<+\infty} (G_m(x) - F_n(x)).$$

Můžeme testovat stejnou hypotézu, tentokrát však proti alternativě, že první výběr pochází z rozdělení, jehož distribuční funkce bude nabývat ve všech bodech menší hodnoty než distribuční funkce, která určuje rozdělení, z něhož pochází druhý náhodný výběr. Postup testu bude stejný, dosadíme-li v předchozím $D^-_{n,m}$ namísto $D^+_{n,m}$ a $D^-_{n,m}(\alpha) = D^+_{n,m}(\alpha)$, kde $D^-_{n,m}(\alpha)$ je kritická hodnota náhodné veličiny $D^-_{n,m}$.

2.5 Pásy spolehlivosti

Nechť $X_1, \ldots, X_n$ je náhodný výběr rozsahu n z rozdělení s distribuční funkcí $F(x)$. Empirická distribuční funkce tohoto výběru je $F_n(x)$. Označme

$$D_n = \sup_{-\infty<x<+\infty} |F(x) - F_n(x)|.$$

Pro $D_n(\alpha)$ nechť platí pro zvolenou hladinu α vztah

$$P\big(D_n < D_n(\alpha)\big) = 1 - \alpha.$$

Položíme $D_n(\alpha) = y/\sqrt{n}$. Nyní můžeme k distribuční funkci $F(x)$ vymezit pás, jehož hranice je vymezena křivkami

$$F(x) + \frac{y}{\sqrt{n}} \qquad a \qquad F(x) - \frac{y}{\sqrt{n}}.$$

Potom

$$P\Big(\sup_{-\infty<x<+\infty} |F(x) - F_n(x)| < \frac{y}{\sqrt{n}}\Big)$$

je zřejmě pravděpodobnost, že empirická distribuční funkce $F_n(x)$ bude celá ležet uvnitř tohoto pásu.

Na druhou stranu k funkci $F_n(x)$ dostaneme pás s hraničními křivkami

$$F_n(x) + D_n(\alpha) \qquad a \qquad F_n(x) - D_n(\alpha),$$

který s pravděpodobností $1 - \alpha$ pokryje distribuční funkci $F(x)$.

Někdy je užitečné určit, minimální rozsah výběru n tak, aby takový pás spolehlivosti nepřekročil šířku $2D$, kde D je předem zvolené kladné číslo. Ze vztahu $D = D_n(\alpha) = y/\sqrt{n}$ určíme $n = y^2/D^2$ ([5], strana 106). Hodnotu y určíme ze vztahu

$$P\Big(D_n < \frac{y}{\sqrt{n}}\Big) = 1 - \alpha,$$

kde $1 - \alpha$ je požadovaná pravděpodobnost, že pás pokryje distribuční funkci $F(x)$. Použitím limitní věty 2.2 dostaneme pro n přibližný vztah

$$P\big(D_n < D\big) = 1 - \alpha \doteq K(\sqrt{n}D),$$

kde $K(y)$ je Kolmogorova funkce definovaná ve větě 2.2. Příslušné hodnoty Kolmogorovy funkce jsou tabelovány v [5] v tabulce 39.

2.6 Rozdělení pro konečné rozsahy výběrů

Mějme $X_1, \ldots, X_n$ a $Y_1, \ldots, Y_n$ dva nezávislé náhodné výběry stejného rozsahu z téhož rozdělení. Nechť $F_n(x)$ a $G_n(x)$ jsou příslušné empirické distribuční funkce obou výběrů. Označíme

$$D_n = \sup_{-\infty<x<+\infty} |F_n(x) - G_n(x)|, \quad D_n^* = \sup_{-\infty<x<+\infty} (F_n(x) - G_n(x)).$$

Pro konečné rozsahy výběrů je užitečné znát přesné rozdělení veličin D_n i D_n^*. Touto problematikou se zabývali Gněděnko a Koroljuk (viz [6], strana 426).

Položíme-li $c = y\sqrt{2n}$, potom podle [5], strana 100, platí následující vztahy.

$$\Phi_n^*(y) = P\Big(\sqrt{\frac{n}{2}} D_n^* < y\Big) =$$

$$= \begin{cases} 0 & \text{pro} \quad \mathrm{y} \le \frac{1}{\sqrt{2\mathrm{n}}}, \\ 1 - \frac{\binom{2n}{n-c}}{\binom{2n}{n}} & \text{pro} \quad \frac{1}{\sqrt{2\mathrm{n}}} < \mathrm{y} \le \sqrt{\frac{\mathrm{n}}{2}}, \\ 1 & \text{jinak} \end{cases}$$

a

$$\Phi_n(y) = P\Big(\sqrt{\frac{n}{2}} D_n < y\Big) =$$

$$= \begin{cases} 0 & \text{pro} \quad \mathrm{y} \le \frac{1}{\sqrt{2\mathrm{n}}}, \\ \frac{1}{\binom{2n}{n}} \sum_{k=-[n/c]}^{+[n/c]} (-1)^k \binom{2n}{n-kc} & \text{pro} \quad \frac{1}{\sqrt{2\mathrm{n}}} < \mathrm{y} \le \sqrt{\frac{\mathrm{n}}{2}}, \\ 1 & \text{jinak.} \end{cases}$$

Na základě těchto vztahů lze dokázat Smirnovovy věty 2.3 a 2.4 limitním přechodem. Důkaz je uveden v [6], strana 426.

2.7 Příklad

Uvedeme test hypotézy na hladině $\alpha = 0.05$, že výběr rozsahu $n = 16$ pochází z rozdělení se spojitou distribuční funkcí $F(x)$, proti všeobecné alternativě. Kladným i záporným odchylkám mezi $F_n(x)$ a $F(x)$ se přikládá stejná důležitost. $F(x)$ je definována následovně.

$$F(x) = \begin{cases} 0 & \text{pro } x \leq 0; \\ x & \text{pro } 0 < x \leq 1; \\ 1 & \text{pro } x > 1. \end{cases}$$

Výběrové hodnoty jsou (převzato z [5], strana 103)

$0,692;\quad 0,259;\quad 0,463;\quad 0,986;\quad 0,084;\quad 0,466;\quad 0,884;\quad 0,219;$

$0,992;\quad 0,108;\quad 0,680;\quad 0,792;\quad 0,468;\quad 0,328;\quad 0,565;\quad 0,554.$

Vytvoříme přehlednou tabulku.

i	x_i	$F_n(x_i)$	$F(x_i)$	$F_n(x_i) - F(x_i)$	$F_n(x_i + 0) - F(x_i)$
1	0,084	0,0000	0,084	-0,0840	-0,0215
2	0,108	0,0625	0,108	-0,0455	0,0170
3	0,219	0,1250	0,219	-0,0940	-0,0315
4	0,259	0,1875	0,259	-0,0715	-0,0090
5	0,328	0,2500	0,328	-0,0780	-0,0155
6	0,463	0,3125	0,463	**-0,1505**	-0,0880
7	0,466	0,3750	0,466	-0,0910	-0,0285
8	0,468	0,4375	0,468	-0,0305	0,0320
9	0,554	0,5000	0,554	-0,0540	0,0085
10	0,565	0,5625	0,565	-0,0025	0,0600
11	0,680	0,6250	0,680	-0,0550	0,0075
12	0,692	0,6875	0,692	0,0045	0,0580
13	0,792	0,7500	0,792	-0,0420	0,0205
14	0,884	0,8125	0,884	-0,0715	-0,0090
15	0,986	0,8750	0,986	-0,1110	-0,0485
16	0,992	0,9375	0,992	-0,0545	0,0080

Vidíme, že $D_{16} = 0,1505$. Kritická hodnota na hladině $\alpha = 0,05$ je podle tabulky 37 v [5] $D_{16}(0,05) = 0,32733$. Platí $D_{16} < D_{16}(0,05)$, podle odstavce 2.2 tedy není důvod k zamítnutí hypotézy, že uvedený náhodný výběr pochází z rozdělení s distribuční funkcí $F(x)$.

Uvažujme nyní stejný příklad, tentokrát však předpokládejme, že rozdíl $F_n(x) - F(x)$ bude převážně nekladný. Testujme hypotézu, že náhodný výběr pochází z rozdělení, jehož distribuční funkce je $F(x)$, proti alternativě, že náhodný výběr pochází z rozdělení, jehož distribuční funkce je v každém bodě menší než funkce $F(x)$. Podle odstavce 2.2 definujme

$$D_n^- = \sup_{-\infty<x<+\infty} \Big(F(x) - F_n(x)\Big).$$

Z tabulky dostaneme $D_{16}^- = 0,1505$. Ze vztahu $D_n^-(\alpha) = D_n(2\alpha)$ dostaneme

$$P\Big(D_{16}^- < D_{16}^-(0,05)\Big) = P\Big(D_{16}^- < D_{16}(0,10)\Big) = 0,95.$$

Podle tabulky 37 v [5] je kritická hodnota $D_{16}(0,10) = 0,29472 = D_{16}^-(0,05)$. Protože $D_{16}^- < D_{16}^-(0,05)$, dostáváme, že test nevede k zamítnutí hypotézy.

2.8 Závěr

Ukázali jsme, že empirická distribuční funkce náhodného výběru je pro dostatečně velký rozsah výběru „podobná" teoretické distribuční funkci. Máme-li náhodný výběr z rozdělení, jehož distribuční funkce nám není známa, je možné nahradit tuto neznámou distribuční funkci empirickou distribuční funkcí náhodného výběru. V praxi se ukazuje, že empirická distribuční funkce poskytuje nejlepší možný odhad teoretické distribuční funkce.

Literatura

[1] Anděl J.: *Statistické metody*, Matfyzpress, Praha, 2003.

[2] Billingsley P.: *Probability and Measure*, Wiley, New York, 1995.

[3] Glivenko V. I.: *Theorie pravděpodobnosti*, Přírodovědecké nakladatelství, Praha, 1950.

[4] Hátle J., Likeš J.: *Základy počtu pravděpodobnosti a matematické statistiky*, SNTL/Alfa, Praha, 1972.

[5] Janko J.: *Statistické tabulky*, NČSAV, Praha, 1958.

[6] Rényi A.: *Teorie pravděpodobnosti*, Academia, Praha, 1972.

[7] Štěpán J.: *Teorie pravděpodobnosti*, Academia, Praha, 1987.