MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Nikola Hlaváčová

Testy dobré shody s exponenciálním rozdělením

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: RNDr. Šárka Hudecová, Ph.D.
Studijní program: Matematika
Studijní obor: Obecná matematika

Praha 2023

Prohlašuji, že jsem tuto bakalářskou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů. Tato práce nebyla využita k získání jiného nebo stejného titulu.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V .... dne ....

Podpis autora

Tímto bych ráda poděkovala vedoucí mé bakalářské práce RNDr. Šárce Hudecové, Ph.D. za trpělivost, vstřícnost a cenné rady, které mi při zpracování této práce poskytla.

Název práce: Testy dobré shody s exponenciálním rozdělením

Autor: Nikola Hlaváčová

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: RNDr. Šárka Hudecová, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Práce se zabývá testy dobré shody pro exponenciální rozdělení. V první části jsou zavedeny základní pojmy včetně jejich vlastností. Poté se věnujeme testům dobré shody založeným na empirické distribuční funkci. Tyto testy dělíme podle toho, zda je parametr rozdělení $\lambda > 0$ známý nebo neznámý. Je-li parametr neznámý, použijeme metodu tzv. parametrického bootstrapu, kdy neznámý parametr λ odhadneme pomocí metody maximální věrohodnosti. Následuje část, kde jsou představeny testy využívající některou z vlastností exponenciálního rozdělení např. Giniho index a střední zbytkovou životnost. V této části jsou odvozeny tvary jejich testových statistik a rozdělení za platnosti nulové hypotézy. Na závěr je uvedena simulační studie, která porovnává jednotlivé testy z hlediska hladiny a síly pro různá nastavení.

Klíčová slova: test dobré shody, exponenciální rozdělení, Kolmogorovův-Smirnovův test, Cramérův von Misesův test, Andersonův Darlingův test, Giniho index, střední zbytková životnost

Title: Goodness-of-fit tests for exponential distributions

Author: Nikola Hlaváčová

Department: Department of Probability and Mathematical Statistics

Supervisor: RNDr. Šárka Hudecová, Ph.D., Department of probability and mathematical Statistics

Abstract: The thesis deals with goodness-of-fit tests for the exponential distribution. In the first part, basic concepts are introduced, including their properties. Afterwards, goodness-of-fit tests based on the empirical distribution function are presented. These tests are divided depending on whether the parameter of the distribution $\lambda > 0$ is known or unknown. If the parameter is unknown, the method of parametric bootstrap is used, where the unknown parameter λ is estimated using the maximum likelihood method. The next section presents tests utilizing some properties of the exponential distribution, such as the Gini index and Mean Residual Life. In this part, the forms of their test statistics and distributions under the assumption of the null hypothesis are derived. Finally, a simulation study is presented that compares individual tests in terms of their level and power for different settings.

Keywords: goodness-of-fit test, exponential distribution, Kolmogorov-Smirnov test, Cramér-von Mises test, Anderson Darling test, Gini coefficient, Mean Residual Life

Obsah

Úvod

Exponenciální rozdělení je spojité rozdělení náhodné veličiny, které se běžně vyskytuje v teorii pravděpodobnosti a matematické statistice. Jako náhodnou veličinu zde typicky uvažujeme dobu, než nastane nějaký pozorovaný jev. Dá se tedy říci, že exponenciální rozdělení modeluje dobu čekání mezi dvěmi událostmi anebo dobu čekání na nastání nějakého jevu. Často se toto rozdělení označuje jako rozdělení „bez paměti“, neboť nám čas od začátku pozorování nezvyšuje ani nesnižuje riziko výskytu události v čase. Exponenciální rozdělení má své využití například při výpočtech času radioaktivního rozpadu, při výpočtech doby životnosti nějakého předmětu, v pojistné matematice. Hraje taky velmi důležitou roli v Markovových řetězcích se spojitým časem. Jedná se o dobu čekání na „poruchu“.

Testy dobré shody jsou statistické testy, které nám určují, zda při použití určitého modelu naše pozorované hodnoty odpovídají hodnotám, které očekáváme. Pomocí testů dobré shody můžeme například určit, jestli je námi vybraná skupina lidí skutečně reprezentativní pro celou populaci. Mezi nejznámější testy dobré shody patří Kolmogorovův-Smirnovův test, Chí-kvadrát test nebo Shapiro-Wilkův test. Cílem této práce je popsat testy dobré shody pro již zmíněné exponenciální rozdělení.

V první kapitole si zavedeme definici exponenciální rozdělení. Ukážeme jaké jsou jeho vlastnosti, mezi které patří například „bezpaměťovost“, rozdělení minima nebo rozdělení součtu. Zároveň s tím zavedeme empirickou distribuční funkci včetně jejích vlastností.

Poté se budeme věnovat testům založeným právě na již zmíněné empirické distribuční funkci. Mezi tyto testy patří Kolmogorovův-Smirnovův test, Cramérův von Misesův test a Andersonův Darlingův test. Všechny tyto tři testy budeme ještě dělit podle toho, zda parametr $\lambda > 0$ je známý a nebo neznámý. Pro neznámý parametr $\lambda > 0$ se setkáme s metodou tzv. parametrického bootstrapu.

Kapitola 3 bude zaměřena na testy dobré shody, které využívají některou z vlastností exponenciálního rozdělení. Jedná se o test založený na Giniho indexu a testy založené na střední zbytkové životnosti (Mean Residual Life).

Ve čtvrté kapitole uděláme simulační studii, kde využijeme testy a metody zmíněné v teoretické části. Tyto testy mezi sebou porovnáme. Pro jednotlivé testy se podíváme na jejich hladiny za platnosti nulové hypotézy pro různé rozsahy výběru a různé hodnoty parametru $\lambda > 0$. Zároveň se zaměříme i na jejich síly pro různé alternativy.

Vlastním přínosem této práce je rozepsání a podrobnější odvození některých tvrzení a vztahů týkajících se prezentovaných testových statistik a jejich vlastností. Doplněny jsou také ilustrační obrázky a příklady. V simulační studii jsou porovnány jednotlivé testy z hlediska hladiny a síly pro různá nastavení. Pro některé testy bylo třeba si příslušné kódy v programu R naimplementovat.

1. Základní pojmy

Nejprve si uvedeme definici exponenciálního rozdělení a empirické distribuční funkce s příslušnými charakteristikami. Tyto zavedené pojmy budeme používat v dalších částech práce.

1.1 Exponenciální rozdělení a jeho vlastnosti

Základní poznatky o exponenciálním rozdělení pochází z knihy Dupač a Hušková (2005).

Definice 1 (Exponenciální rozdělení)**.** *Nechť $\lambda > 0$. Nechť náhodná veličina X má hustotu*

$$f_X(x) = \begin{cases} \lambda e^{-x\lambda} & x > 0, \\ 0 & \textit{jinak.} \end{cases}$$

Pak říkáme, že X má exponenciální rozdělení. *Značíme $X \sim \mathsf{Exp}(\lambda)$.*

Věta 1. *Nechť $\lambda > 0$ a nechť náhodná veličina X má exponenciální rozdělení s parametrem λ. Distribuční funkce má tvar*

$$F_X(x) = \begin{cases} 1 - e^{-x\lambda} & x > 0, \\ 0 & \textit{jinak.} \end{cases}$$

Pro střední hodnotu dostáváme, že $\mathsf{E}\ X = \lambda^{-1}$ a rozptyl má tvar $\operatorname{var} X = \lambda^{-2}$.

Důkaz. K důkazu této věty použijeme definici 1. Distribuční funkci spočteme pomocí vztahu

$$F_X(x) = \int_{-\infty}^{x} f_x(y)dy.$$

Je-li $x \leq 0$, dostáváme

$$F_X(x) = \int_{-\infty}^{x} f_x(y)dy = F_X(x) = \int_{-\infty}^{x} 0dy = 0.$$

Je-li $x > 0$, dostáváme

$$\begin{aligned} \int_{-\infty}^{x} f_x(y)dy = \int_{-\infty}^{x} \lambda e^{-y\lambda} dy = \int_{-\infty}^{0} \lambda e^{-y\lambda} dy + \int_{0}^{x} \lambda e^{-y\lambda} dy \\ = 0 + \int_{0}^{x} \lambda e^{-y\lambda} dy = 1 - e^{-x\lambda}. \end{aligned}$$

K výpočtu střední hodnoty použijeme gama funkci a vztah

$$\mathsf{E}\ X = \int_{-\infty}^{\infty} x f_X(x)dx.$$

Dostáváme

$$\mathsf{E}\ X = \int_{-\infty}^{\infty} x f_X(x)dx = \int_{0}^{\infty} x\lambda e^{-x\lambda} dx = \int_{0}^{\infty} \frac{1}{\lambda} u e^{-u} du = \frac{1}{\lambda}\Gamma(2) = \frac{1}{\lambda},$$

kde ve 3. rovnosti jsme zavedli substituci $u = \lambda x$.

Zbývá spočítat rozptyl. Ten spočteme pomocí vzorce $\operatorname{var} X = \mathsf{E}\ X^2 - (\mathsf{E}\ X)^2$. Víme, že $(\mathsf{E}\ X)^2 = \frac{1}{\lambda^2}$ z výpočtu střední hodnoty $\mathsf{E}\ X$. Nyní spočteme hodnotu $\mathsf{E}\ X^2$. Dostáváme

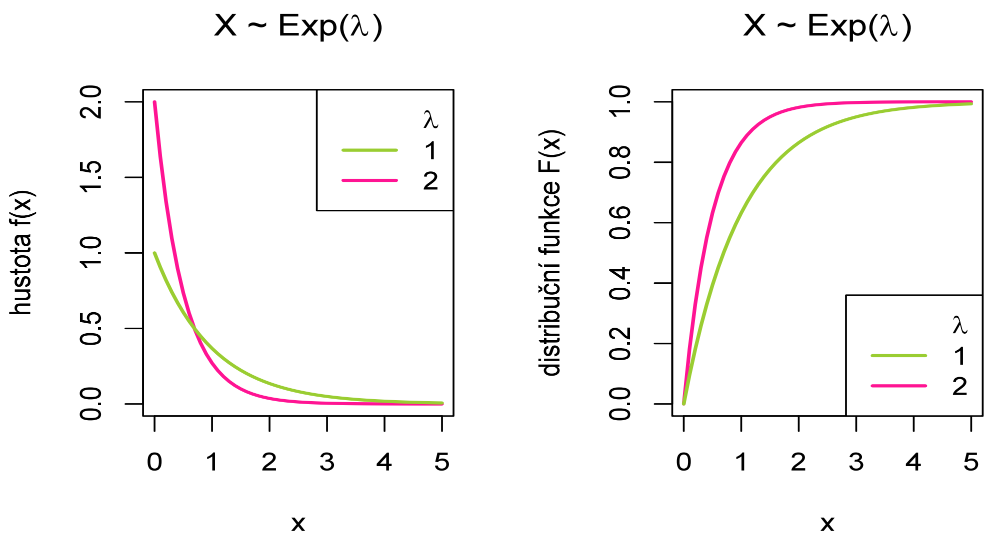
$$\mathsf{E}\ X^2 = \int_{-\infty}^{\infty} x^2 f_X(x)dx = \int_0^{\infty} x^2 \lambda e^{-x\lambda} dx = \frac{1}{\lambda^2}\int_0^{\infty} u^2 e^{-u} du = \frac{1}{\lambda^2}\Gamma(3) = \frac{2}{\lambda^2},$$

kde jsme opět zavedli substituci $u = \lambda x$. Celkem tedy dostáváme, že

$$\operatorname{var} X = \mathsf{E}\ X^2 - (\mathsf{E}\ X)^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

□

Na obrázku 1.1 můžeme vidět porovnání hustot a distribučních funkcí exponenciálního rozdělení s parametry 1 a 2.



Obrázek 1.1: Hustoty a distribuční funkce pro exponenciální rozdělení $\mathsf{Exp}(1)$ (zelená) a $\mathsf{Exp}(2)$(růžová).

Tvrzení 2. *Nechť X je náhodná veličina s exponenciálním rozdělením, potom platí*

$$P(X \geq s + t | X \geq t) = P(X \geq s)\ \forall s,t > 0.$$

Neboli se dá říci, že exponenciální rozdělení je rozdělení bez paměti.

Důkaz. Z věty 1 víme, že

$$F_X(x) = \begin{cases} 1 - e^{-x\lambda} & x > 0, \\ 0 & \text{jinak.} \end{cases}$$

Pravděpodobnost, že $X \geq x$ je rovna $P(X \geq x) = e^{-x\lambda}$, kde $x > 0$. Z definice podmíněné pravděpodobnosti dostáváme

$$\begin{aligned} P(X \geq s+t | X \geq t) &= \frac{P(X \geq s+t, X \geq t)}{P(X \geq t)} = \frac{P(X \geq s+t)}{P(X \geq t)} \\ &= \frac{e^{-(s+t)\lambda}}{e^{-t\lambda}} = e^{-s\lambda} = P(X \geq s), \end{aligned}$$

což jsme chtěli dokázat.

□

Tvrzení 3. *Mějme dány dvě nezávislé stejně rozdělené náhodná veličiny X, Y s exponenciální rozdělení $X \sim \mathsf{Exp}(\lambda_X)$ a $Y \sim \mathsf{Exp}(\lambda_Y)$. Nechť $Z = \min(X, Y)$ je náhodná veličina. Potom $Z \sim \mathsf{Exp}(\lambda_X + \lambda_Y)$.*

Důkaz. K důkazu tohoto tvrzení budeme potřebovat tvar distribuční funkce pro exponenciální rozdělení zavedený ve větě 1. Budeme chtít najít tvar distribuční funkce pro náhodnou veličinu Z. Nechť $z > 0$. Potom

$$\begin{aligned} F_Z(z) &= P(Z \leq z) = P(\min(X, Y) \leq z) = 1 - P(X > z, Y > z) \\ &= 1 - P(X > z)P(Y > z) = 1 - e^{-z\lambda_X} e^{-z\lambda_Y} = 1 - e^{-z(\lambda_X + \lambda_Y)}, \end{aligned}$$

kde ve 4. rovnosti jsme využili nezávislost náhodných veličin X a Y. Dostáváme tedy, že náhodná veličina Z má také exponenciální rozdělení s parametrem $\lambda_X + \lambda_Y$. Pokud $z \leq 0$, tak $F_Z(z) = 0$.

□

Tvrzení 4. *Mějme dány dvě nezávislé stejně rozdělené náhodné veličiny X, Y s exponenciálním rozdělením $X \sim \mathsf{Exp}(\lambda)$ a $Y \sim \mathsf{Exp}(\lambda)$. Nechť $U = X + Y$ je náhodná veličina. Potom $U \sim \Gamma(2, \lambda)$.*

Důkaz. Nejprve spočteme distribuční funkci $F_U(x)$ náhodné veličiny U. Pokud $u \leq 0$, tak $F_U(u) = 0$. Nechť $u > 0$. Potom

$$\begin{aligned} F_U(u) &= P(U \leq u) = P(X + Y \leq u) \\ &= \int_0^\infty P(X + Y \leq u | X = x) f_X(x) dx \\ &= \int_0^\infty P(X + Y \leq u | X = x) \lambda e^{-\lambda x} dx \\ &= \int_0^\infty P(Y \leq u - x | X = x) \lambda e^{-\lambda x} dx \\ &= \int_0^\infty (1 - e^{-\lambda(u-x)}) \lambda e^{-\lambda x} dx \\ &= (1 - e^{-\lambda u}) - \lambda u e^{-\lambda u}. \end{aligned}$$

Zderivováním distribuční funkce $F_U(u)$ podle parametru u, dostaneme hustotu $f_U(u) = \lambda^2 u e^{-\lambda u}$. Odtud vidíme, že $U \sim \Gamma(2, \lambda)$.

□

1.2 Empirická distribuční funkce

V této kapitole si zavedeme definici empirické distribuční funkce s příslušnými charakteristikami, které využijeme při popisování testů dobré shody s exponenciálním rozdělením.

Definice 2 (Empirická distribuční funkce)**.** *Mějme náhodný výběr* $X_1, ..., X_n$*. Funkce*

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(X_i \leq x),\ x \in \mathbb{R},$$

se nazývá empirická distribuční funkce*.*

Věta 5. *Nechť* $x \in \mathbb{R}$ *a* $\widehat{F}_n(x)$ *je empirická distribuční funkce náhodného výběru* $X_1, ..., X_n$ *s distribuční funkcí* $F_X(x)$*. Potom platí:*

$$\begin{aligned}
&(i) \quad \mathsf{E}\,\widehat{F}_n(x) = F_X(x),\\
&(ii) \quad \widehat{F}_n(x) \xrightarrow[n\to\infty]{P} F_X(x),\\
&(iii) \quad \sup_{x\in\mathbb{R}} |\widehat{F}_n(x) - F_X(x)| \xrightarrow[n\to\infty]{P} 0.
\end{aligned}$$

Důkaz. Pro body (i) a (ii) je dobré si uvědomit, že pro pevné x se dá hodnota $\widehat{F}_n(x)$ chápat jako relativní četnost jevu $[X_i \leq x]$ pro daných n pozorování. Empirickou distribuční funkci tedy můžeme přepsat do tvaru

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(X_i \leq x) = \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}_n.$$

Náhodné veličiny Y_i jsou nezávislé a stejně rozdělené s alternativním rozdělením $\mathsf{Alt}(F_X(x))$. Odtud a z vlastností relativní četnosti dostáváme, že $\mathsf{E}\,\widehat{F}_n(x) = F_X(x)$.

K důkazu bodu (ii) využijeme Čebyševovu nerovnost, kterou můžeme najít v knize Dupač a Hušková (2005). Nechť $\epsilon > 0$, platí

$$P(|\widehat{F}_n(x) - F_X(x)| > \epsilon) \leq \frac{F_X(x)(1 - F_X(x))}{n\epsilon^2} \xrightarrow[n\to\infty]{} 0,$$

neboli $\widehat{F}_n(x) \xrightarrow[n\to\infty]{P} F_X(x)$. Tato část důkazu se dá ukázat i pomocí silného zákona velkých čísel pro veličiny Y_i.

Bod (iii) se též nazývá Glivenkova-Cantelliho věta a důkaz lze nalézt například v knize Borokov (2019).

□

2. Testy založené na empirické distribuční funkci

Cílem následujících dvou kapitol je představit testy dobré shody s exponenciálním rozdělním. Máme dán nějaký náhodný výběr a zajímá nás, zda pochází z exponenciálního rozdělení. V úvodu jsme si ukázali, že exponenciální rozdělení je důležité, neboť má několik specifických vlastností a objevuje se i v náhodných procesech jako například doba do další události. Jedná se rovněž o základní model modelování doby do poruchy.

Mějmě dán náhodný výběr $X_1, X_2, ..., X_n$ nezávislých stejně rozdělených náhodných veličin se spojitou distribuční funkcí $F_X(x)$. Při testování budeme rozlišovat dvě možnosti: Buď můžeme testovat nulovou hypotézu H_0 proti alternativě H_1, kde máme předem daný paramater $\lambda_0 > 0$:

$$H_0 : F_X(x) = 1 - e^{-x\lambda_0} \ \forall x > 0, \tag{2.1}$$

$$H_1 : \exists x > 0 : F_X(x) \neq 1 - e^{-x\lambda_0},$$

nebo testujeme tzv. složenou nulovou hypotézu $\widetilde{H_0}$ proti složené alternativě $\widetilde{H_1}$, kde $\lambda > 0$ je neznámý (předem nespecifikovaný) parametr:

$$\widetilde{H_0} : \exists \lambda > 0 : F_X(x) = 1 - e^{-x\lambda} \ \forall x > 0, \tag{2.2}$$

$$\widetilde{H_1} : \forall \lambda > 0 \ \exists x > 0 : F_X(x) \neq 1 - e^{-x\lambda}.$$

V této kapitole ukážeme aplikaci obecných testů dobré shody a v další kapitole představíme testy, které využívají některou z vlastností exponenciálního rozdělení.

2.1 Testy dobré shody se známým parametrem

Nejprve se zaměříme na testy dobré shody se známým parametrem λ_0, kde $\lambda_0 > 0$. Jako F_0 zde budeme uvažovat distribuční funkci exponenciálního rozdělení se známým parametrem λ_0, neboli $F_0 = 1 - e^{-x\lambda_0}$.

2.1.1 Kolmogorovův-Smirnovův test

Kolmogorovův-Smirnovův test (KS) je test založený na testování shody empirické distribuční funkce s předem zvolenou distribuční funkcí, v našem případě s distribuční funkcí exponenciálního rozdělení. Tvar distribuční funkce pro exponenciální rozdělení jsme si zavedli ve větě 1. KS test je odvozen za předpokladu spojitosti rozdělení distribuční funkce $F_X(x)$. Testová statistika, kterou KS test využívá, je tvaru

$$K_n = \sup_{x \in \mathbb{R}} |\widehat{F}_n(x) - F_0(x)|,$$

kde $\widehat{F}_n(x)$ značí empirickou distribuční funkci z definice 2. Se supremem se omezíme pouze na kladná x. V opačném případě je distribuční funkce exponenciálního rozdělení rovna 0 a tedy i celý rozdíl $|\widehat{F}_n(x) - F_0(x)| = 0$, neboť $\widehat{F}_n(x) = 0$. Kdyby $\widehat{F}_n(x) \neq 0$, znamenalo by to, že jsme napozorovali nějaké záporné hodnoty a to by automaticky znamenalo, že se nemůže jednat o exponenciální rozdělení. Dostáváme tedy testovou statistiku ve tvaru

$$K_n = \sup_{x>0} |\widehat{F}_n(x) - F_0(x)|. \tag{2.3}$$

Máme, že $K_n = \max(K_n^+, K_n^-)$, kde

$$K_n^+ = \sup_{x>0} (\widehat{F}_n(x) - F_0(x)) \text{ a } K_n^- = \sup_{x>0} (F_0(x) - \widehat{F}_n(x)).$$

Testujeme hypotézu H_0 proti alternativě H_1 uvedené v rovnici (2.1). Zajímá nás tedy, jak moc se empirická distribuční funkce liší od předem dané spojité distribuční funkce exponenciálního rozdělení. K tomuto výpočtu potřebujeme určit hodnotu testové statistiky K_n. Tato hodnota se nepočítá pomocí výše uvedeného vztahu, ale pomocí následující věty.

Věta 6. *Je-li F_0 spojitá, tak platí*

$$K_n^+ = \max_{1\le i\le n} \left(\frac{i}{n} - F_0(X_{(i)})\right),\ K_n^- = \max_{1\le i\le n} \left(F_0(X_{(i)}) - \frac{i-1}{n}\right).$$

Důkaz. Důkaz lze najít v knize Gibbons a Chakraborti (2003, Kapitola 4).

□

Nulovou hypotézu H_0 zamítáme, pokud se empirická distribuční funkce příliš liší od F_0, neboli pokud je hodnota testové statistiky příliš velká, jak můžeme vidět na obrázku 2.1. Naopak pokud data skutečně pochází z exponenciálního rozdělení s parametrem λ_0, je hodnota testové statistiky malá. Ilustraci tohoto případu můžeme vidět na obrázku 2.2.

Spočítat přesné rozdělení testové statistiky za platnosti nulové hypotézy býva často obtížné, proto se častěji používá asymptotické rozdělení testové statistiky za platnosti nulové hypotézy.

Věta 7. *Nechť $X_1, X_2, ..., X_n$ je náhodný výběr ze spojitého rozdělení s distribuční funkcí F_X. Potom*

$$\sqrt{n} \sup_{x\in\mathbb{R}} |\widehat{F}_n(x) - F_X(x)| \xrightarrow[n\to\infty]{d} Z,$$

kde náhodná veličina Z má distribuční funkci danou předpisem

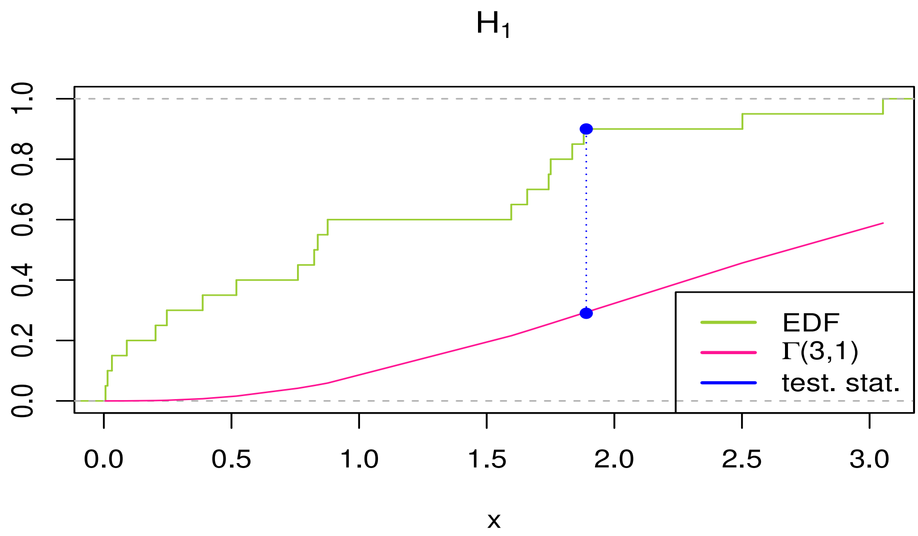
$$G(y) = 1 - 2\sum_{k=1}^{\infty} (-1)^{k+1} e^{-2k^2y^2},\ y > 0,$$

v opačném případě je $G(y) = 0$.

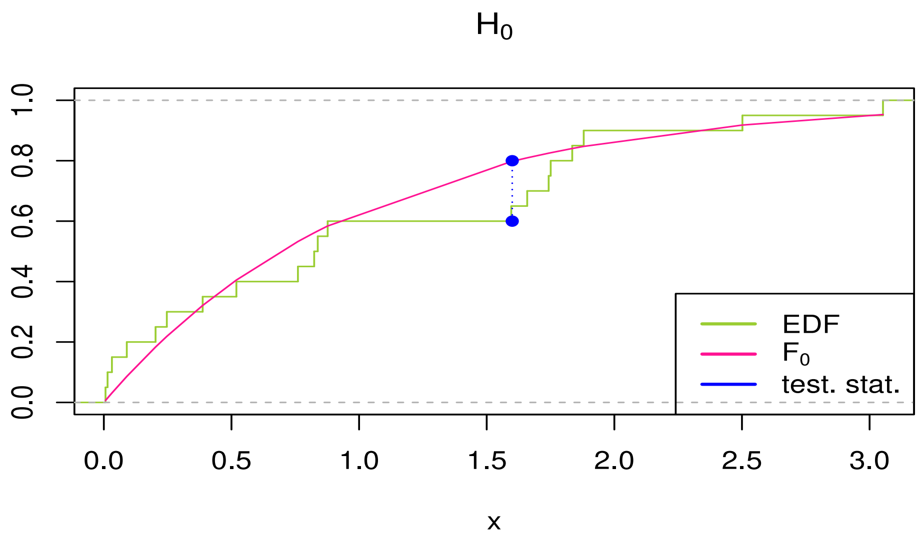
Poznámka. Tato věta se někdy také nazývá Smirnovova věta a její důkaz lze najít v knize Hájek a Šidák (1967).

Nyní již můžeme určit kritický obor. Nulovou hypotézu zamítáme pro příliš velké hodnoty testové statistiky, tedy pro

$$H_0 \text{ zamítáme} \Leftrightarrow \sqrt{n}K_n \ge k_{1-\alpha},$$



Obrázek 2.1: Vyznačení testové statistiky KS testu (modrá) za platnosti alternativy H_1, kdy máme dána data $X_1, X_2, ..., X_{20}$ generovaná z $\Gamma(3,1)$ (růžová) a emprickou distribuční funkcí $\widehat{F}_{20}(x)$ (zelená). Testujeme, zda $\lambda_0 = 1$.



Obrázek 2.2: Vyznačení testové statistiky KS testu (modrá) za platnosti nulové hypotézy H_0, kdy máme dána data $X_1, X_2, ..., X_{20}$ generovaná z $\mathsf{Exp}(1)$ (růžová) a empirickou distribuční funkci $\widehat{F}_{20}(x)$ (zelená). Testujeme, zda $\lambda_0 = 1$.

kde $k_{1-\alpha}$ označuje (1-α)-kvantil rozdělení s distribuční funkcí G. Tento test dodržuje hladinu α asymptoticky.

V praxi si zvolíme konečnou hodnotu K, např. $K = 10\,000$, čímž si aproximujeme nekonečnou sumu v předpisu distribuční funkce G a pak se dá již numericky řešit rovnice

$$1 - 2\sum_{k=1}^{10\,000}(-1)^{k+1}e^{-2k^2y^2} = 1 - \alpha.$$

S rostoucí hodnotou α se kritická hodnota $k_{1-\alpha}$ snižuje (viz tabulka 2.1).

α	0,01	0,05	0,1
$k_{1-\alpha}$	1,6276	1,3581	1,2239

Tabulka 2.1: Kritické hodnoty $k_{1-\alpha}$ pro různé hodnoty α.

Pro náš test dostáváme asymptotickou p-hodnotu danou předpisem

$$p = \inf(\alpha \in (0{,}1) : \sqrt{n}k_n \geq k_{1-\alpha}) = 1 - G(\sqrt{n}k_n),$$

kde k_n je pozorovaná hodnota testové statistiky K_n.

Tvrzení 8. *Mějme dán náhodný výběr $X_i \sim \mathsf{Exp}(\lambda_0)$. Potom rozdělení testové statistiky KS testu za platnosti nulové hypotézy H_0 nezávisí na parametru $\lambda_0 > 0$.*

Důkaz. Předpokládejme, že platí nulová hypotéza H_0 (2.2). Z toho, že $X_i \sim \mathsf{Exp}(\lambda_0)$ víme, že $F_0(x) = 1 - e^{-\lambda_0 x}$ pro $x > 0$. Označme $Y_i = \frac{X_i}{\lambda_0}$. Dále nechť $\widehat{F}'_n(y) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(Y_i \leq y)$, $y \in \mathbb{R}$, označuje empirickou distribuční funkci veličin Y_i a $F'_0(y) = 1 - e^{y}$ pro $y > 0$ označuje distribuční funkci $\mathsf{Exp}(1)$. Potom testovou statistiku 2.3 můžeme přepsat do tvaru

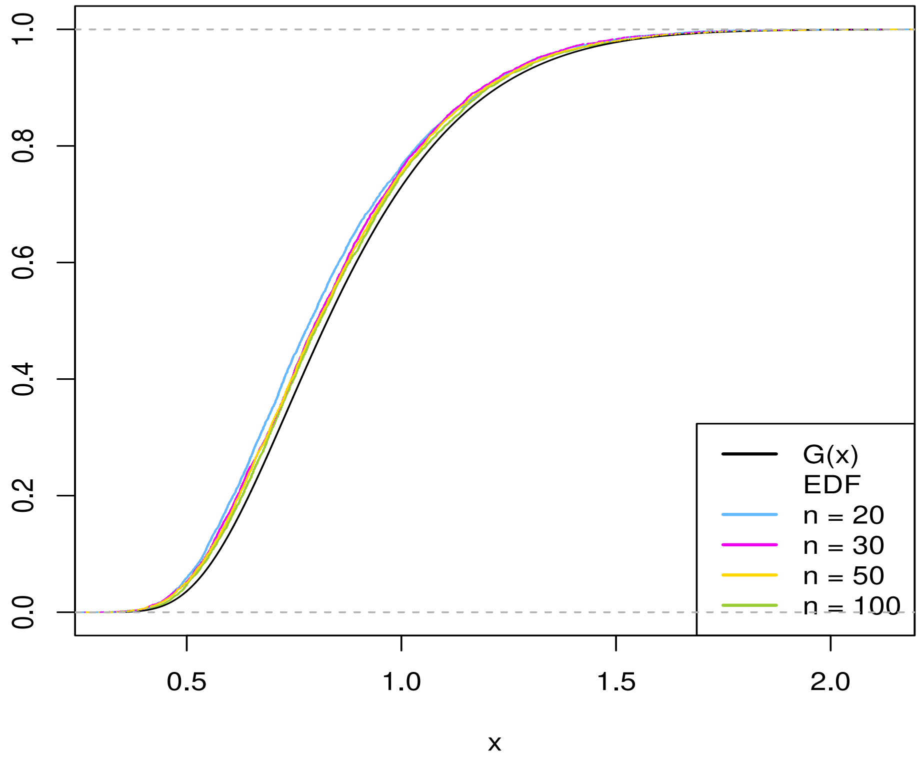
$$K_n = \sup_{\frac{x}{\lambda_0}>0}\left|\widehat{F}_n\left(\frac{x}{\lambda_0}\right) - F_0\left(\frac{x}{\lambda_0}\right)\right| = \sup_{y>0}|\widehat{F}'_n(y) - F'_0(y)|.$$

Tedy rozdělení testové statistiky KS testu za platnosti nulové hypotézy H_0 skutečně nezávisí na parametru $\lambda_0 > 0$.

$\square$

Pokud nás zajímá, jak dobře funkce G z Věty 7 aproximuje empirickou distribuční funkci testových statistik $\sqrt{n}K_n$ pro různé rozsahy výběru například $n = 20, 30, 50, 100$, si nagenerujeme $N = 5\,000$ výběrů o stejném rozsahu n. Z těchto výběrů spočteme příslušné testové statistiky $\sqrt{n}K_n$ a sestrojíme jejich empirickou distribuční funkci a aproximační funkci G. Z obrázku 2.3 je vidět, že čím bude hodnota rozsahu výběru n větší, tím bude empirická distribuční funkce více podobná funkci G a bude se tedy jednat o lepší aproximaci. Pro $n = 100$ už je to velmi rozumné.

Chceme-li určit přesné kritické hodnoty pro dané n, můžeme je dostat pomocí simulace. Pro zadané λ_0 a n nasimulujeme $N = 10\,000$ výběrů a pak kritickou hodnotu určíme jako kvantil příslušných hodnot testových statistik $\sqrt{n}K_n$. Z tabulky 2.2 je vidět, že se kritické hodnoty blíží k hodnotám v tabulce 2.1. Můžeme si také všimnout, že nejprve se začnou podobat kritické hodnoty u $\alpha = 0{,}1$, nejpozději pak kritické hodnoty u $\alpha = 0{,}01$.



Obrázek 2.3: Graf empir. distribuční funkce z hodnot testové statistiky $\sqrt{n}K_n$ o rozsahu výběru $n = 20, 30, 50, 100$ s parametrem $\lambda_0 = 1$ a graf funkce G, která ji aproximuje.

$\lambda_0 = 1$				
$k_{1-\alpha}$	$n = 20$	$n = 30$	$n = 50$	$n = 100$
$\alpha = 0{,}01$	1,5774	1,5249	1,5787	1,6272
$\alpha = 0{,}05$	1,3172	1,3169	1,3372	1,3346
$\alpha = 0{,}10$	1,1844	1,1819	1,2001	1,2024

Tabulka 2.2: Nasimulované kritické hodnoty $k_{1-\alpha}$ pro různé hodnoty α, n a parametr $\lambda_0 = 1$.

2.1.2 Testy založené na integrální vzdálenosti

Nechť ω_n^2 je tvaru:

$$\omega_n^2 = \int_{-\infty}^{\infty} (\widehat{F}_n(x) - F_0(x))^2 w(F_0(x)) dF_0(x), \tag{2.4}$$

kde ω je náhodná veličina a $w(y)$ označuje nezápornou váhovou funkci pro $0 \leq y \leq 1$.

Pokud $w(y) = 1$, jedná se o Cramérův von Misesův test (CvM) (D'Agostino a Stephens (1986, Kapitola 4)). CvM test je založený na podobném principu jako KS test. Opět testujeme shodu empirické distribuční funkce s předem zvolenou distribuční funkcí, v našem případě s distribuční funkcí exponenciálního rozdělení, tj. H_0 proti H_1. Pro $w(y) = 1$ dostáváme, že

$$\omega_n^2 = \int_0^\infty (\widehat{F}_n(x) - (1 - e^{\lambda_0 x}))^2 \lambda_0 e^{-x\lambda_0} dx.$$

Dá se ukázat, že testová statistika, kterou CvM test využívá, se dá upravit do následujícího tvaru:

$$T_n = n\omega_n^2 = \frac{1}{12n} + \sum_{i=1}^n \left(F_0(X_{(i)}) - \frac{2i-1}{2n} \right)^2.$$

Tento tvar lze nalézt v knize D'Agostino a Stephens (1986, Kapitola 4).

Pokud váhová funkce z rovnosti (2.4) splňuje $w(y) = [(1-y)y]^{-1}$, jedná se o Andersonův-Darlingův (AD) test (D'Agostino a Stephens (1986, Kapitola 4)). Testová statistika, kterou AD test využívá, je tvaru:

$$\begin{aligned}
A_n &= n \int_{-\infty}^\infty \frac{(\widehat{F}_n(x) - F_0(x))^2}{(F_0(x)(1 - F_0(x))} dF_0(x) \\
&= n \int_0^\infty \frac{(\widehat{F}_n(x) - (1 - e^{\lambda_0 x}))^2}{((1 - e^{\lambda_0 x})(e^{\lambda_0 x}))} \lambda_0 e^{-x\lambda_0} dx \\
&= -n - \sum_{i=1}^n \frac{2i-1}{n} (\log(F_0(X_{(i)})) + \log(1 - F_0(X_{(n+1-i)}))).
\end{aligned}$$

Tento tvar lze nalézt opět v knize D'Agostino a Stephens (1986, Kapitola 4).

Tvrzení 9. *Mějme dán náhodný výběr $X_i \sim \mathsf{Exp}(\lambda_0)$. Potom rozdělení testové statistiky* (2.4) *za platnosti nulové hypotézy H_0 nezávisí na parametru $\lambda_0 > 0$.*

Důkaz. Důkaz je analogický důkazu ve Tvrzení 8.

□

Podobně jako ve větě 7 se dá ukázat, že testové statistiky CvM testu a AD testu konvergují v distribuci ke konečnému rozdělení G(jiné než u KS testu), které je stejné pro všechna spojitá rozdělení F_X (viz Lehmann a Romano (2005, Kapitola 14)). Příslušné kritické hodnoty jsou tabelovány. Můžeme ale také použít podobný postup, jaký je uvedený v části o KS testu. Pro konečné n jsme schopni si sami kritické hodnoty nasimulovat (viz tabulka 2.3 a tabulka 2.4). Kritické hodnoty opět, jako u KS testu, nezávisí na parametru λ_0 (tvrzení 9). Tyto kritické hodnoty můžeme najít také například v tabulkách v článku Stephens (1974). Při porovnání je vidět, že se nasimulované hodnoty liší jen nepatrně od hodnot tabelovaných.

2.2 Testy dobré shody s neznámým parametrem

Nyní se zaměříme na testy dobré shody s neznámým parametrem $\lambda > 0$. Tento parametr λ si odhadneme pomocí metody maximální věrohodnosti. Pro exponenciální rozdělení bychom stejný odhad dostali i použitím momentové metody.

$\lambda_0 = 1$					tab
	$n = 20$	$n = 30$	$n = 50$	$n = 100$	$n \to \infty$
$\alpha = 0{,}01$	0,719	0,685	0,707	0,729	0,743
$\alpha = 0{,}05$	0,455	0,440	0,468	0,462	0,461
$\alpha = 0{,}10$	0,344	0,342	0,350	0,351	0,347

Tabulka 2.3: Nasimulované kritické hodnoty CvM testu pro různé hodnoty α, n a parametr $\lambda_0 = 1$ a tabelované hodnoty tab pro $n \to \infty$.

$\lambda_0 = 1$					tab
	$n = 20$	$n = 30$	$n = 50$	$n = 100$	$n \to \infty$
$\alpha = 0{,}01$	3,893	3,569	3,719	3,740	3,857
$\alpha = 0{,}05$	2,507	2,436	2,499	2,451	2,492
$\alpha = 0{,}10$	1,930	1,913	1,953	1,903	1,933

Tabulka 2.4: Nasimulované kritické hodnoty AD testu pro různé hodnoty α, n a parametr $\lambda_0 = 1$ a tabelované hodnoty tab pro $n \to \infty$.

Mějme náhodný výběr $X_1, X_2, ..., X_n$ z exponenciálního rozdělení s neznámým parametrem λ. Z Definice 1 víme, že

$$f_X(x) = \begin{cases} \lambda e^{-x\lambda} & x > 0, \\ 0 & \text{jinak.} \end{cases},$$

takže

$$L(\lambda, \mathbf{x}) = \prod_{i=1}^{n} \lambda e^{-x_i\lambda} \mathbf{1}(x_i > 0) = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i} \mathbf{1}(\min x_i > 0),$$

kde $\mathbf{x} \in \mathbb{R}^n$ je dané a

$$l(\lambda,\mathbf{x}) = n \log(\lambda) - \lambda \sum_{i=1}^{n} x_i + c,$$

kde c je konstanta nezávisející na parametru λ. Zderivujeme funkci $l(\lambda,\mathbf{x})$ podle parametru λ a řešíme rovnici

$$\frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0$$

dostáváme

$$\hat{\lambda}_n = \frac{1}{\overline{X}_n}. \tag{2.5}$$

Nalezli jsme tedy maximálně věrohodný odhad neznámého parametru λ. Je vidět, že námi nalezený odhad je maximum. To plyne například z toho, že pokud uděláme druhou derivaci funkce $l(\lambda,\mathbf{x})$ podle parametru λ, tak je záporná pro všechna $\lambda > 0$. Funkce $l(\lambda,\mathbf{x})$ je tedy konkávní pro všechna $\lambda > 0$ a námi nalezený maximálně věrohodný odhad $\hat{\lambda}_n = \frac{1}{\overline{X}_n}$ je jejím maximem.

Všechny testové statistiky zmíněné v části 2.1 se dají použít i pro testy dobré shody s neznámým parametrem λ. Místo distribuční funkce F_0 uvažujeme distribuční funkci $F_{\hat{\lambda}_n}(x) = 1 - e^{-\hat{\lambda}_n x}$ pro $x > 0$ s odhadnutým parametrem $\hat{\lambda}_n$. Opět nás zajímá, jak se od sebe liší empirická distribuční funkce a distribuční funkce exponenciálního rozdělení s parametrem $\hat{\lambda}_n$.

Testová statistika KS testu má tvar

$$K_n^\star = \sup_{x>0} |\widehat{F}_n(x) - F_{\hat{\lambda}_n}(x)|. \tag{2.6}$$

Testová statistika CvM testu má tvar

$$T_n^\star = \frac{1}{12n} + \sum_{i=1}^{n} \left(F_{\hat{\lambda}_n}(X_{(i)}) - \frac{2i-1}{2n} \right)^2.$$

Testová statistika AD testu má tvar

$$A_n^\star = -n - \sum_{i=1}^{n} \frac{2i-1}{n} (\log(F_{\hat{\lambda}_n}(X_{(i)})) + \log(1 - F_{\hat{\lambda}_n}(X_{(n+1-i)})))$$

Podíváme se blíže na KS test. Pro CvM a AD test budou následující postupy fungovat opět stejně, jako tomu bylo v části 2.1.

Testová statistika (2.6) nemá asymptoticky stejné rozdělení, které je uvedené ve Větě 7, jako testová statistika KS testu se známým parametrem λ_0. Pokud by se Věta 7 použila, byl by test konzervativní, neboli skutečná hodnota α by byla mnohem menší než bychom po testu chtěli, tím pádem by nám test nezamítal, i když by měl. Například mějme data $X_1, X_2, ..., X_{100}$ z exponenciálního rozdělení s parametrem $\lambda = 1$ a testujme složenou nulovou hypotézu $\widetilde{H}_0 : \exists\lambda > 0\ \forall x > 0 : F_X(x) = 1 - e^{-x\lambda}$ na hladině $\alpha = 0{,}05$. Tuto hypotézu by KS test neměl zamítnout. Pokud bychom použili větu 7, tak příslušná p-hodnota je rovna 0,005, neboli p-hodnota $< \alpha$ a tedy zamítáme $\widetilde{H}_0$.

Příslušnou p-hodnotu testu spočítáme pomocí metody tzv. parametrického bootstrapu (Omelka (2022, Kapitola 8) nebo Stute a kol. (1993)). Hlavní myšlenka je taková, že máme náhodný výběr $X_1, X_2, ..., X_n$ a z něj odhadneme parametr $\hat{\lambda}_n$. Chceme vědět, jaké je rozdělení testové statistiky (2.6). Postupně si generujeme B bootstrapových výběrů $X_1^{(b)\star}, X_2^{(b)\star}, ..., X_n^{(b)\star}$ pro $b = 1, ..., B$ o stejném rozsahu n, které mají exponenciální rozdělení s parametrem $\hat{\lambda}_n$. Pro každý bootstrapový výběr si spočteme parametr $\hat{\lambda}_n^{(b)\star}$ (analogicky jako pro náhodný výběr $X_1, X_2, ..., X_n$) a pomocí tohoto $\hat{\lambda}_n^{(b)\star}$ spočteme testovou statistiku $K_n^{(b)\star}$. Zamítáme pokud naše testová statistika bude větší než výběrový kvantil na hladině $1 - \alpha$ spočtený z $K_n^{(b)\star}$, kde $b = 1, ..., B$. P-hodnotu potom spočteme jako relativní četnost případů, kdy je hodnota testové statistiky $K_n^{(b)\star}$ větší než hodnota testové statistiky K_n, neboli p-hodnotu můžeme vyjádřit jako

$$\frac{\#b : [K_n^{(b)\star} > K_n]}{B}.$$

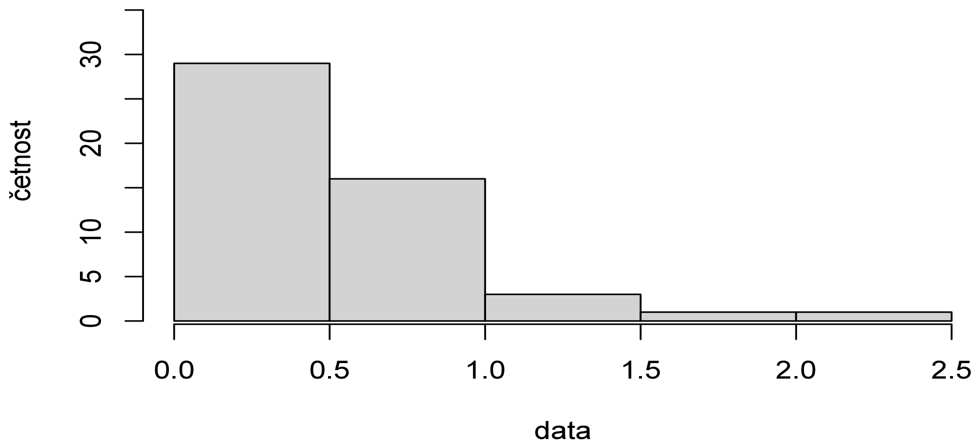
Hodnotu testové statistiky K_n spočteme z náhodného výběru $X_1, X_2, ..., X_n$ a odhadnutého parametru $\hat{\lambda}_n$.

Příklad. Mějme nagenerovaná data $X_1, X_2, ..., X_{50}$ z $\mathsf{Exp}(2)$ (obrázek 2.4).

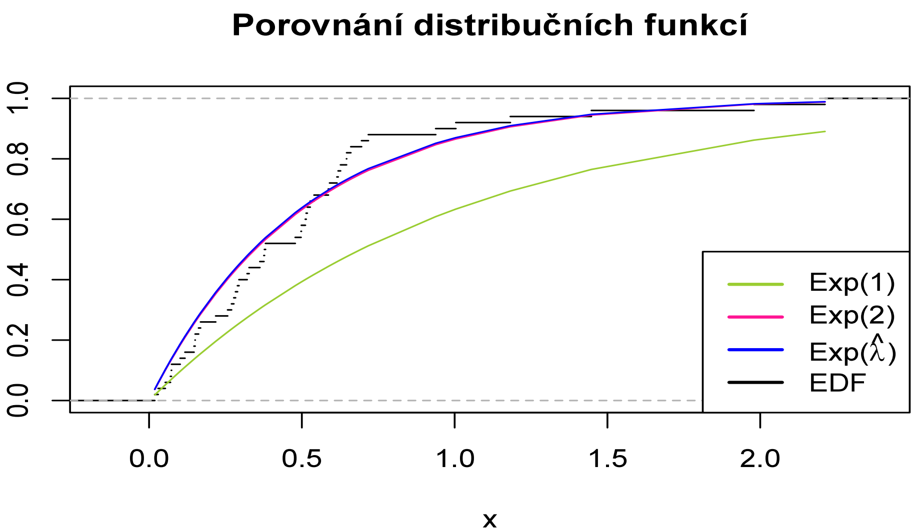
Nejprve testujeme nulovou hypotézu $H_0 : F_X(x) = 1 - e^{-x}$ pro $x > 0$, tzn. data pocházejí z $\mathsf{Exp}(1)$. Testujeme na hladině $\alpha = 0{,}05$. Tuto hypotézu by měly všechny tři zmíněné testy zamítnout. P-hodnota pro KS test je rovna $1 - G(\sqrt{n}K_n)$, kde $n = 50$ a G je funkce z věty 7. Pro KS test, CvM test a AD test vychází p-hodnota $< 0{,}001$. Ve všech třech případech je p-hodnota menší než $\alpha = 0{,}05$, a tedy zamítáme nulovou hypotézu H_0, data skutečně nepocházejí z $\mathsf{Exp}(1)$.

Nyní se podívejme na test složené nulové hypotézy $\widetilde{H}_0 : \exists\lambda > 0\ \forall x > 0 : F_X(x) = 1 - e^{-x\lambda}$. Tuto hypotézu by všechny tři zmíněné testy neměly zamítnout. Pro výpočet p-hodnoty se použije metoda parametrického bootstrapu zmíněná výše, neboť parametr λ je neznámý. Uvažujme $B = 1\,000$ bootstrapových výběrů. Pro KS test vychází p-hodnota 0,156. Pro CvM test vychází p-hodnota 0,090. Pro AD test vychází p-hodnota 0,103. Ve všech třech případech je p-hodnota větší než $\alpha = 0{,}05$, a tedy nemůžeme zamítnout složenou nulovou hypotézu $\widetilde{H}_0$, že data pocházejí z $\mathsf{Exp}(\lambda)$.

Na Obrázku 2.5 můžeme vidět rozdíly mezi jednotlivými distribučními funkcemi zmíněnými v tomto příkladě. Je vidět, že data skutečně nepocházejí z $\mathsf{Exp}(1)$. Naopak nemůžeme zamítnout složenou nulovou hypotézu $\widetilde{H}_0$, jak je vidět na obrázku funkce $\mathsf{Exp}(\hat{\lambda}_n)$ a $\mathsf{Exp}(2)$ jsou totožné.



Obrázek 2.4: Histogram četností dat $X_1, X_2, ..., X_{50}$ generovaných z $\mathsf{Exp}(2)$.



Obrázek 2.5: Porovnání distribučních funkcí $\mathsf{Exp}(1), \mathsf{Exp}(2), \mathsf{Exp}(\hat{\lambda}_n)$ a empirické distribuční funkce $\widehat{F}_{50}(x)$ dat generovaných z $\mathsf{Exp}(2)$. Distribuční funkce $\mathsf{Exp}(2)$ a $\mathsf{Exp}(\hat{\lambda}_n)$ se překrývají.

3. Testy využívající vlastnosti exponenciálního rozdělení

V této části práce se zaměříme na testy dobré shody, které využívají některou z vlastností exponenciálního rozdělení. Mějme opět dán náhodný výběr $X_1, X_2, ..., X_n$ nezávislých stejně rozdělených náhodných veličin s distribuční funkcí exponenciálního rozdělení $F_X(x)$ s parametrem $\lambda > 0$. Předpokládejme, že parametr $\lambda > 0$ je neznámý a nechť $\hat{\lambda}_n = 1/\overline{X_n}$ je jeho maximálně věrohodný odhad spočtený dle (2.5). Dále nechť pro $1 \leq j \leq n$

$$Y_j = X_j \hat{\lambda}_n = X_j / \overline{X_n}.$$

3.1 Test založený na Giniho indexu

Následující test je založený na Giniho indexu.

Definice 3 (Giniho index)**.** *Nechť X_1, X_2 jsou nezávislé stejně rozdělené náhodné veličiny z rozdělení P_X s konečnou nenulovou střední hodnotou* $\mathsf{E}\ X$. Giniho index *rozdělení P_X definujeme předpisem*

$$G_X = \frac{1}{2} \frac{\mathsf{E}\ |X_1 - X_2|}{\mathsf{E}\ X_1}.$$

Poznámka. Ekvivalentní zápisy tvaru Giniho indexu můžeme najít v článku Dorfman (1979), odkud pochází i následující zápis vhodnější pro výpočet. Nechť náhodná veličina X pochází ze spojitého rozdělení s distribuční funkcí $F_X(x)$ a konečnou nenulovou střední hodnotou $\mathsf{E}\ X$, potom

$$G_X = 1 - \frac{1}{\mathsf{E}\ X} \int_0^\infty (1 - F_X(x))^2 dx. \tag{3.1}$$

Spočteme si hodnotu Giniho indexu exponenciálního rozdělení s parametrem $\lambda > 0$. Víme, že $F_X(x) = 1 - e^{-\lambda x}$ pro $x > 0$ a $\mathsf{E}\ X = \frac{1}{\lambda}$. Dosazením do předpisu (3.1) dostáváme

$$G_X = 1 - \lambda \int_0^\infty [1 - (1 - e^{-\lambda x})]^2 dx = 1 - \lambda \int_0^\infty e^{-2\lambda x} dx = 1 - \lambda \frac{1}{2\lambda} = 1 - \frac{1}{2} = \frac{1}{2}.$$

Definice 4 (Výběrový Giniho index)**.** *Nechť $X_1, X_2, ..., X_n$ je náhodný výběr ze spojitého rozdělení.* Výběrový Giniho index *definujeme předpisem*

$$G_n = \frac{1}{\binom{n}{2}} \frac{1}{2} \sum\sum_{i<j} \frac{|X_i - X_j|}{\overline{X_n}}.$$

Poznámka. Formálně správně bychom měli v definici 4 dodefinovat G_n pro případ, kdy $\overline{X_n} = 0$. Případ $\overline{X_n} = 0$ nastává s pravděpodobností 0, neboť data pocházejí ze spojitého rozdělení. Proto toto dodefinování neprovádíme.

Výběrový Giniho index zapsaný pomocí náhodných veličin Y_j je tvaru

$$G_n = \frac{1}{2n(n-1)} \sum_{j,k=1}^{n} |Y_j - Y_k|,$$

neboť $\frac{1}{\binom{n}{2}}\frac{1}{2} = \frac{1}{n(n-1)}$ a $\sum\sum_{i<j} |X_i - X_j|$ označuje sčítání přes všechny různé dvojice, zatímco $\sum_{j,k=1}^{n} |Y_j - Y_k|$ označuje sčítání přes všechny dvojice, těch je dvakrát více, proto dostáváme ještě $\frac{1}{2}$.

Lemma 10. *Mějme dán náhodný výběr $X_1, X_2, ..., X_n$ z rozdělení s distribuční funkcí F_X s konečnou nenulovou střední hodnotou $\mathsf{E}\ X_1$ a s konečným rozptylem. Potom pro statistiku*

$$T_n = \frac{1}{\binom{n}{2}} \sum\sum_{i<j} |X_i - X_j| \tag{3.2}$$

platí, že

$$\mathsf{E}\ T_n = \mathsf{E}\ |X_1 - X_2|$$

a

$$\operatorname{var} T_n = \frac{4}{n(n-1)} \left[\operatorname{var} X_1 + (n-2)\,\mathsf{E}\,(|X_1 - X_2||X_1 - X_3|) - \frac{2n-3}{2}(\mathsf{E}\ T_n)^2 \right].$$

Pokud náhodný výběr $X_1, X_2, ..., X_n$ pochází z exponenciálního rozdělení s parametrem $\lambda > 0$, potom

$$\operatorname{var} T_n = \frac{2(2n-1)}{\lambda^2 3n(n-1)}$$

a

$$\mathsf{E}\ T_n = \frac{1}{\lambda}.$$

Důkaz. Chceme ukázat, že $\mathsf{E}\ T_n = \mathsf{E}\ |X_1 - X_2|$. Víme, že průměr odhaduje střední hodnoty. O veličinách T_n můžeme říci, že jsou zobecněný průměr, uvažujeme totiž průměr přes všechny dvojice, neboť $\binom{n}{2}$ označuje počet všech různých dvojic a $\sum\sum_{i<j} |X_i - X_j|$ označuje sčítání přes všechny různé dvojice. Pokud počítáme střední hodnotu, můžeme s ní jít dovnitř a dostáváme

$$\mathsf{E}\ T_n = \frac{1}{\binom{n}{2}} \binom{n}{2} \mathsf{E}\ |X_1 - X_2| = \mathsf{E}\ |X_1 - X_2|.$$

Jedná se vlastně o průměr všech dvojic.

Pokud data pochází z exponenciálního rozdělení s parametrem $\lambda > 0$, tak víme, že $G_X = \frac{1}{2}$. Neboli

$$\frac{1}{2} = G_X = \frac{1}{2}\frac{\mathsf{E}\ |X_1 - X_2|}{\mathsf{E}\ X_1} = \frac{\lambda}{2}\,\mathsf{E}\ |X_1 - X_2| = \frac{\lambda}{2}\,\mathsf{E}\ T_n.$$

Odsud již snadno vidíme, že $\mathsf{E}\ T_n = \frac{1}{\lambda}$.

Důkaz pro obecný rozptyl a rozptyl, kdy data pochází z exponenciálního rozdělení s parametrem $\lambda > 0$, lze najít v článku Zenga a kol. (2004).

$\square$

Věta 11. *Mějme dán náhodný výběr $X_1, X_2, ..., X_n$ z rozdělení s distribuční funkcí F_X s konečnou nenulovou střední hodnotou $\mathsf{E}\ X_1$. Potom pro Giniho index G_X a výběrový Giniho index G_n platí*

$$G_n \xrightarrow[n\to\infty]{P} G_X = \frac{1}{2}\frac{\mathsf{E}\ |X_1 - X_2|}{\mathsf{E}\ X_1}. \tag{3.3}$$

Důkaz. Je zřejmé, že $\frac{1}{2} \xrightarrow[n\to\infty]{} \frac{1}{2}$. Ze silného zákona velkých čísel platí, že $\overline{X_n} \xrightarrow[n\to\infty]{s.j.} \mathsf{E}\ X_1$. Uvažujme funkci $g(t) = \frac{1}{t}$. Potom podle věty o spojité transformaci, s využitím funkce g, platí

$$\frac{1}{\overline{X_n}} \xrightarrow[n\to\infty]{s.j.} \frac{1}{\mathsf{E}\ X_1}.$$

Označme

$$T_n = \frac{1}{\binom{n}{2}} \sum\sum_{i<j} |X_i - X_j|.$$

Podle lemmatu 10 platí, že $\mathsf{E}\ T_n = \mathsf{E}\ |X_1 - X_2|$ a pro obecý rozptyl T_n platí $\operatorname{var} T_n = \frac{4}{n(n-1)}\left[\operatorname{var} X_1 + (n-2)\,\mathsf{E}\,(|X_1 - X_2||X_1 - X_3|) - \frac{2n-3}{2}(\mathsf{E}\ T_n)^2\right]$. Máme, že T_n je nestranným odhadem a $\operatorname{var} T_n \xrightarrow[n\to\infty]{} 0$. Odsud dostáváme $T_n \xrightarrow[n\to\infty]{P} \mathsf{E}\ |X_1 - X_2|$. Protože konvergence skoro jistě implikuje konvergenci v pravděpodobnosti

$$\left(\frac{1}{\overline{X_n}}, T_n\right)^T \xrightarrow[n\to\infty]{P} \left(\frac{1}{\mathsf{E}\ X_1}, \mathsf{E}\ |X_1 - X_2|\right)^T.$$

Opět podle věty o spojité transformaci s funkcí $h(s,t) = st$ dostaneme

$$\frac{T_n}{\overline{X_n}} \xrightarrow[n\to\infty]{P} \frac{\mathsf{E}\ |X_1 - X_2|}{\mathsf{E}\ X_1}.$$

Celkem tedy platí

$$G_n = \frac{1}{\binom{n}{2}} \sum\sum_{i<j} \frac{1}{2}\frac{|X_i - X_j|}{\overline{X_n}} \xrightarrow[n\to\infty]{P} \frac{1}{2}\frac{\mathsf{E}\ |X_1 - X_2|}{\mathsf{E}\ X_1} = G_X,$$

což jsme chtěli dokázat.

□

Důsledek. Podle věty 11 platí, že pokud náhodný výběr $X_1, X_2, ..., X_n$ pochází z exponenciáního rozdělení s parametrem $\lambda > 0$, potom

$$G_n \xrightarrow[n\to\infty]{P} \frac{1}{2}.$$

Tvrzení 12. *Nechť $X_1, X_2, ..., X_n$ je náhodný výběr z rozdělení s distribuční funkcí F_X, nenulovou střední hodnotou a $\mathsf{E}\ X^2 < \infty$. Potom pro výběrový Giniho index G_n platí*

$$\frac{G_n - \mathsf{E}\ G_n}{\sqrt{\operatorname{var} G_n}} \xrightarrow[n\to\infty]{d} N(0,1).$$

Důkaz. Důkaz tohoto obecného tvrzení lze najít v článku Hoeffding (1948).

□

My budeme chtít toto tvrzení dokázat pro náhodný výběr z exponenciálního rozdělení s parametrem $\lambda > 0$. Odvodit rozptyl a střední hodnotu G_n není snadné, a proto použijeme přepis tohoto testu pomocí tzv. normalizovaných rozestupů. Seřadíme-li si hodnoty $X_1, X_2, ...X_n$ od nejmenší po největší, dostaneme uspořádaný výběr. Symbolem $X_{(j)}$ rozumíme j-tou nejměnší hodnotu mezi pozorovanými $X_1, X_2, ..., X_n$. Rozestupy mezi hodnotami $X_{(j)}$ jsou definovány vztahem $d_j = X_{(j)} - X_{(j-1)}$, kde $j = 1, 2, ..., n$ a $X_{(0)} = 0$. Normalizovaný rozestup získáme z d_j přenásobením číslem $(n + 1 - j)$, neboli dostáváme

$$D_j = (n + 1 - j)d_j = (n + 1 - j)(X_{(j)} - X_{(j-1)}).$$

Dále spočteme hodnoty

$$U_k = \sum_{j=1}^{k} D_j / \sum_{j=1}^{n} X_j, \text{ kde } 1 \leq k \leq n - 1.$$

Tvrzení 13. *Nechť platí nulová hypotéza H_0. Potom veličiny U_k pro $k = 1, ..., n-1$ tvoří uspořádaný náhodný výběr z rovnoměrného rozdělení na intervalu* (0,1).

Důkaz. Důkaz tohoto tvrzení lze najít v knize D'Agostino a Stephens (1986, Kapitola 10).

□

Definujme si náhodnou veličinu S_n jako výraz

$$S_n = \sum_{j=1}^{n-1} U_j = 2n - \frac{2}{n} \sum_{j=1}^{n} jY_{(j)}.$$

Odvození tohoto vztahu lze najít v knize D'Agostino a Stephens (1986, Kapitola 10).

Předpokládejme, že platí nulová hypotéza H_0, potom U_j tvoří uspořádaný náhodný výběr. Nám ovšem nevadí, že U_j tvoří uspořádaný náhodný výběr, neboť sčítání uspořádaného náhodného výběru je to samé jako sčítání neuspořádaného náhodného výběru. Pokud si toto uvědomíme, tak můžeme použít centrální limitní větu pro výběr o rozsahu $n - 1$ z rovnoměrného rozdělení na intervalu (0,1), která za platnosti nulové hypotézy dává

$$\frac{S_n - (n-1)\frac{1}{2}}{\sqrt{(n-1)\frac{1}{12}}} \xrightarrow[n \to \infty]{d} N(0,1),$$

protože pro rovnoměrné rozdělení na intervalu (0,1) máme, že střední hodnota je rovna 1/2 a rozptyl je roven 1/12.

Následující tvrzení ukáže spojitost náhodné veličiny S_n s výběrovým Giniho indexem G_n.

Tvrzení 14. *Platí*

$$\frac{S_n}{n-1} = 1 - G_n. \tag{3.4}$$

Důkaz. Z definice výběrového Giniho indexu můžeme psát

$$G_n = \frac{1}{2n(n-1)} \sum_{j=1}^{n} \sum_{k=1}^{n} |Y_j - Y_k|.$$

Pokud budeme dále tento výraz upravovat, dostáváme následující

$$\begin{aligned}
G_n &= \frac{1}{2n(n-1)} \sum_{j=1}^{n} \left[\sum_{k:Y_k<Y_j} (Y_j - Y_k) + \sum_{k:Y_k>Y_j} (Y_k - Y_j) \right] \\
&= \frac{1}{2n(n-1)} \sum_{j=1}^{n} \left[Y_j \left(\sum_{k:Y_k<Y_j} 1 \right) - \sum_{k:Y_k<Y_j} Y_k + \sum_{k:Y_k>Y_j} Y_k - Y_j \left(\sum_{k:Y_k>Y_j} 1 \right) \right] \\
&= \frac{1}{2n(n-1)} \sum_{j=1}^{n} \left[Y_j(R_j - 1) - \sum_{k:Y_k<Y_j} Y_k + \sum_{k:Y_k>Y_j} Y_k - Y_j(n - R_j) \right],
\end{aligned}$$

kde R_j označuje pořadí náhodné veličiny Y_j, neboli platí $Y_j = Y_{(R_j)}$. Dále si všimněme, že platí

$$n = \frac{\sum_{k=1}^{n} X_k}{\overline{X}_n} = \sum_{k=1}^{n} Y_k = Y_j + \sum_{k:Y_k<Y_j} Y_k + \sum_{k:Y_k>Y_j} Y_k$$

a

$$\sum_{j=1}^{n} Y_j R_j = \sum_{j=1}^{n} Y_{(j)} j.$$

S využitím těchto pozorování můžeme dále G_n upravovat a dostaneme

$$\begin{aligned}
G_n &= \frac{1}{2n(n-1)} \sum_{j=1}^{n} \left[Y_j(R_j - 1) - \sum_{k:Y_k<Y_j} Y_k + n - Y_j - \sum_{k:Y_k<Y_j} Y_k - Y_j(n - R_j) \right] \\
&= \frac{1}{2n(n-1)} \sum_{j=1}^{n} \left[Y_j R_j - Y_j - 2 \sum_{k:Y_k<Y_j} Y_k + n - Y_j - Y_j n + Y_j R_j \right] \\
&= \frac{1}{2n(n-1)} \left[\sum_{j=1}^{n} 2 Y_j R_j - 2 \sum_{j=1}^{n} Y_j - 2 \sum_{j=1}^{n} \sum_{k:Y_k<Y_j} Y_k + \sum_{j=1}^{n} n - \sum_{j=1}^{n} Y_j n \right] \\
&= \frac{1}{2n(n-1)} \left[2 \sum_{j=1}^{n} j Y_{(j)} - 2n - 2 \sum_{j=1}^{n} \sum_{k:Y_k<Y_j} Y_k + n^2 - n^2 \right] \\
&= \frac{1}{n(n-1)} \left[\sum_{j=1}^{n} j Y_{(j)} - n - \sum_{j=1}^{n} \sum_{k:Y_k<Y_j} Y_k \right]
\end{aligned}$$

Pokud si rozmyslíme, kolikrát započteme $Y_{(1)}, Y_{(2)}, ...$, dostáváme rovnost

$$\sum_{j=1}^{n} \sum_{k:Y_k<Y_j} Y_k = \sum_{j=1}^{n} Y_j (n - R_j).$$

Upravujme dále G_n

$$
\begin{aligned}
&= \frac{1}{n(n-1)}\left[\sum_{j=1}^{n} jY_{(j)} - n - \sum_{j=1}^{n} Y_j(n-R_j)\right] \\
&= \frac{1}{n(n-1)}\left[\sum_{j=1}^{n} jY_{(j)} - n - \sum_{j=1}^{n} Y_j n - \sum_{j=1}^{n} Y_j R_j\right] \\
&= \frac{1}{n(n-1)}\left[\sum_{j=1}^{n} jY_{(j)} - n - n^2 - \sum_{j=1}^{n} Y_{(j)} j\right] \\
&= \frac{1}{n(n-1)}\left[2\sum_{j=1}^{n} jY_{(j)} - n - n^2\right].
\end{aligned}
$$

Pro $1 - G_n$ platí

$$
\begin{aligned}
1 - G_n &= \frac{1}{n(n-1)}\left[n^2 - n - 2\sum_{j=1}^{n} jY_{(j)} + n + n^2\right] \\
&= \frac{1}{n(n-1)}\left[2n^2 - 2\sum_{j=1}^{n} jY_{(j)}\right].
\end{aligned}
$$

Pro $S_n/(n-1)$ platí

$$
\begin{aligned}
\frac{S_n}{n-1} &= \frac{1}{n-1}\left[2n - \frac{2}{n}\sum_{j=1}^{n} jY_{(j)}\right] \\
&= \frac{1}{n(n-1)}\left[2n^2 - 2\sum_{j=1}^{n} jY_{(j)}\right].
\end{aligned}
$$

Dokázali jsme vztah

$$\frac{S_n}{n-1} = 1 - G_n.$$

□

Nyní již můžeme dokázat následující tvrzení.

Tvrzení 15. *Nechť $X_1, X_2, ..., X_n$ je náhodný výběr z exponenciálního rozdělení s parametrem $\lambda > 0$. Potom pro výběrový Giniho index G_n platí*

$$\frac{G_n - 1/2}{\sqrt{\frac{1}{12(n-1)}}} \xrightarrow[n\to\infty]{d} N(0{,}1).$$

Důkaz. Z použití centrální limitní věty na testovou statistiku S_n víme, že platí

$$\frac{S_n - (n-1)\frac{1}{2}}{\sqrt{(n-1)\frac{1}{12}}} \xrightarrow[n\to\infty]{d} N(0{,}1).$$

Vytkneme-li z čitatele i jmenovatele výraz $(n-1)$ dostáváme

$$\frac{\frac{S_n}{n-1}-\frac{1}{2}}{\sqrt{\frac{1}{12(n-1)}}} \xrightarrow[n\to\infty]{d} N(0,1).$$

Ve tvrzení 14 máme uvedený vztah (3.4) mezi S_n a G_n. S využitím tohoto vztahu lze nyní již snadno odvodit rozdělení pro G_n. Dostáváme

$$\frac{1-G_n-1/2}{\sqrt{\frac{1}{12(n-1)}}} = \frac{1/2-G_n}{\sqrt{\frac{1}{12(n-1)}}} \xrightarrow[n\to\infty]{d} N(0,1).$$

Protože normované normální rozdělení je symetrické kolem počátku, můžeme psát

$$\frac{G_n-1/2}{\sqrt{\frac{1}{12(n-1)}}} \xrightarrow[n\to\infty]{d} N(0,1).$$

□

Pro náhodný výběr $X_1, X_2, ..., X_n$ z exponenciálního rozdělení s parametrem $\lambda > 0$ označme

$$F_n = \frac{G_n-1/2}{\sqrt{\frac{1}{12(n-1)}}} = \sqrt{12(n-1)}(G_n-1/2).$$

Testujeme nulovou hypotézu H_0 (2.1) proti alternativě H_1. Spočteme si hodnotu výběrového Giniho indexu G_n. Za platnosti nulové hypotézy H_0 by tato hodnota měla konvergovat v pravděpodobnosti k hodnotě Giniho indexu $G_X = \frac{1}{2}$. Pokud tomu tak je, tak víme, že pro výběrový Giniho index G_n platí, že F_n konverguje v distribuci k normovanému normálnímu rozdělení. Nyní již můžeme určit kritický obor. Nulovou hypotézu zamítáme pro příliš malé a příliš velké hodnoty testové statistiky, tedy

$$H_0 \text{ zamítáme} \Leftrightarrow |F_n| \geq u_{1-\alpha/2},$$

kde $u_{1-\alpha/2}$ označuje $(1-\alpha/2)$-kvantil normovaného normálního rozdělení. Tento test dodržuje hladinu α asymptoticky. Pro náš test dostáváme asymptotickou p-hodnotu danou předpisem

$$p = \inf(\alpha \in (0,1) : f_n \leq u_{\alpha/2} \text{ nebo } f_n \geq u_{1-\alpha/2}) = 2\min(1-\Phi(f_n-), \Phi(f_n)),$$

kde f_n označuje pozorovanou hodnotu testové statistiky F_n.

Test pro exponenciální rozdělení založený na S_n a G_n není obecně konzistentní, jak je zmíněno v knize D'Agostino a Stephens (1986, Kapitola 10). Test totiž testuje, zda Giniho index je roven 1/2. Pokud bychom měli jiné rozdělení, které není exponenciální a má také Giniho index roven 1/2, tak proti tomuto rozdělení nebude tento test fungovat.

3.2 Testy založené na střední zbytkové životnosti

Mezi další testy využívající některou z vlastností exponenciálního rozdělení patří testy založené na střední zbytkové životnosti (Mean Residual Life). Pojem střední zbytková životnost si můžeme lépe vysvětlit např. na životnosti výrobku. Střední zbytková životnost je zbývající očekávaná doba životnosti výrobku za předpokladu, že výrobek byl funkční do nějakého časového okamžiku t. Zbytkovou životnost lze také chápat jako dobu dalšího trvání a je definována vztahem

$$m(t) = \mathsf{E}\ (X - t|X > t) \text{ pro } t > 0.$$

Tvrzení 16. *Nechť X je nezáporná náhodná veličina s konečnou nenulovou střední hodnotou $\mathsf{E}\ X$. Potom rozdělení F_X náhodné veličiny X je exponenciální právě tehdy, když*

$$\mathsf{E}\ (X - t|X > t) = \mathsf{E}\ X\ \forall t > 0. \tag{3.5}$$

Důkaz. Předpokládejme, že nezáporná náhodná veličina X pochází z exponenciálního rozdělení s parametrem λ. Z linearity střední hodnoty můžeme psát $\mathsf{E}\ (X - t|X > t) = \mathsf{E}\ (X|X > t) - t$. Dále dostáváme

$$\begin{aligned} \mathsf{E}\ (X|X > t) - t &= \frac{\mathsf{E}\ (\mathbf{1}(X > t)X)}{1 - P(X \le t)} - t = \frac{\int_t^\infty x\lambda e^{-\lambda x}dx}{1 - (1 - e^{-\lambda t})} - t = \frac{e^{-\lambda t}(\frac{1}{\lambda} + t)}{e^{-\lambda t}} - t \\ &= \frac{1}{\lambda} = \mathsf{E}\ X. \end{aligned}$$

Naopak předpokládejme, že $m(t) = \mathsf{E}\ (X - t|X > t) = \mathsf{E}\ X$. Chceme ukázat, že náhodný výběr $X_1, X_2, ..., X_n$ pochází z exponenciálního rozdělení s konečnou nenulovou střední hodnotou $\mathsf{E}\ X$. S využitím vlastnosti

$$1 - F(t) = \frac{m(0)}{m(t)} e^{-\int_0^t \frac{1}{m(u)}du},$$

uvedené v článku Guess a Proschan (1988), a definice $m(t)$ můžeme psát

$$\begin{aligned} 1 - F(t) &= \frac{\mathsf{E}\ X}{\mathsf{E}\ X} e^{-\int_0^t \frac{1}{\mathsf{E}\ X}du} \\ &= e^{-\frac{1}{\mathsf{E}\ X}t}. \end{aligned}$$

Odtud pak snadno plyne, že $F(t) = 1 - e^{-\frac{1}{\mathsf{E}\ X}t}$, neboli distribuční funkce je ve tvaru exponenciálního rozdělení s parametrem $\lambda = 1/\,\mathsf{E}\ X$.

□

Tvrzení 17. *Podmínka* (3.5) *uvedená ve tvrzení 16 se dá zaměnit za podmínku*

$$\mathsf{E}\,(\min(X, t)) = \mathsf{E}\ X \cdot F(t)\ \forall t > 0. \tag{3.6}$$

Důkaz. Nechť $t > 0$. Nejprve si rozepišme

$$\begin{aligned} \mathsf{E}\,(\min(X,t)) &= \mathsf{E}\,(\min(X,t)|X \le t)P(X \le t) + \mathsf{E}\,(\min(X,t)|X > t)P(X > t) \\ &= \mathsf{E}\,(X|X \le t)P(X \le t) + \mathsf{E}\,(t|X > t)P(X > t) \\ &= \mathsf{E}\,(X|X \le t)P(X \le t) + tP(X > t) \\ &= \mathsf{E}\,(X|X \le t)F(t) + t(1 - F(t)) \\ &= \frac{\mathsf{E}\ (\mathbf{1}(X \le t)X)}{F(t)} F(t) + t(1 - F(t)) \\ &= \mathsf{E}\ (\mathbf{1}(X \le t)X) + t(1 - F(t)) \\ &= \mathsf{E}\ ((1 - \mathbf{1}(X > t))X) + t(1 - F(t)) \\ &= \mathsf{E}\ X - \mathsf{E}\ (\mathbf{1}(X > t)X) + t(1 - F(t)). \end{aligned}$$

Z důkazu tvrzení 16 máme, že

$$\mathsf{E}\,(\mathbf{1}(X > t)X) = \mathsf{E}\,(X|X > t)(1 - F(t)).$$

S využitím této vlastnosti můžeme psát

$$\begin{aligned}
\mathsf{E}\,(\min(X,t)) &= \mathsf{E}\,X - \mathsf{E}\,(X|X > t)(1 - F(t)) + t(1 - F(t)) \\
&= \mathsf{E}\,X - (1 - F(t))(\mathsf{E}\,(X|X > t) - t) \\
&= \mathsf{E}\,X - (1 - F(t))(\mathsf{E}\,(X - t|X > t).
\end{aligned}$$

Chceme ukázat, že platí (3.5) právě tehdy, když platí (3.6). Předpokládejme nejprve, že platí (3.5), tedy $\mathsf{E}\,(X - t|X > t) = \mathsf{E}\,X$. Potom

$$\begin{aligned}
\mathsf{E}\,(\min(X,t)) &= \mathsf{E}\,X - (1 - F(t))(\mathsf{E}\,(X - t|X > t) \\
&= \mathsf{E}\,X - (1 - F(t))\,\mathsf{E}\,X \\
&= \mathsf{E}\,X \cdot F(t).
\end{aligned}$$

Nyní předpokládejme, že platí (3.6), tedy $\mathsf{E}\,(\min(X,t)) = \mathsf{E}\,X \cdot F(t)$. Potom

$$\begin{aligned}
\mathsf{E}\,(X - t|X > t) &= \mathsf{E}\,(X|X > t) - t \\
&= \frac{\mathsf{E}\,(\mathbf{1}(X > t)X)}{1 - F(t)} - t \\
\mathsf{E}\,(X - t|X > t)(1 - F(t)) &= \mathsf{E}\,(\mathbf{1}(X > t)X) - t(1 - F(t)) \\
&= -(\mathsf{E}\,X - \mathsf{E}\,(\mathbf{1}(X > t)X) + t(1 - F(t)) - \mathsf{E}\,X) \\
&= -(\mathsf{E}\,(\min(X,t)) - \mathsf{E}\,X) \\
&= \mathsf{E}\,X - \mathsf{E}\,X \cdot F(t) \\
&= \mathsf{E}\,X(1 - F(t)).
\end{aligned}$$

Tedy dostáváme $\mathsf{E}\,(X - t|X > t) = \mathsf{E}\,X$. Tím je ekvivalence dokázána.

□

Na základě náhodného výběru $X_1, X_2, ..., X_n$ můžeme pro pevné $t > 0$ levou stranu (3.6) odhadnout výrazem

$$\frac{1}{n}\sum_{i=1}^{n} \min(X_i,t)$$

a pravou stranu rovnosti (3.6) odhadneme výrazem

$$\overline{X_n} \cdot \widehat{F}_n(t) = \overline{X_n} \cdot \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(X_i \le t).$$

Pro každé $t > 0$ můžeme tyto dva výrazy porovnat. Podobně jako v kapitole 2 se můžeme dívat na jejich supremální vzdálenost, anebo na integrální vzdálenost. Nejprve se podívejme na supremální vzdálenost

$$\begin{aligned}
&\sup_{t>0}\left|\frac{1}{n}\sum_{i=1}^{n} \min(X_i,t) - \overline{X_n}\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(X_i \le t)\right| \\
&= \overline{X_n}\sup_{t>0}\left|\frac{1}{n}\sum_{i=1}^{n} \frac{\min(X_i,t)}{\overline{X_n}} - \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(X_i \le t)\right| \\
&= \overline{X_n}\sup_{t>0}\left|\frac{1}{n}\sum_{i=1}^{n} \min\left(\frac{X_i}{\overline{X_n}},\frac{t}{\overline{X_n}}\right) - \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(X_i \le t)\right|.
\end{aligned}$$

Výběrový průměr $\overline{X_n}$ uvedený před supremem vynecháme. Zaveďme si substituci $y = t/\overline{X_n}$, potom můžeme dále psát

$$\begin{aligned}\overline{k_n} &= \sup_{y>0}\left|\frac{1}{n}\sum_{i=1}^{n}\min(Y_i,y) - \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(X_i \le y\overline{X_n})\right| \\ &= \sup_{y>0}\left|\frac{1}{n}\sum_{i=1}^{n}\min(Y_i,y) - \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(Y_i \le y)\right|.\end{aligned}$$

V článku Baringhaus a Henze (2000) jsou zmíněny dvě testové statistiky využívající vlastnost (3.6). První z nich je statistika Kolmogorovova-Smirnovova typu, která má tvar

$$\begin{aligned}\overline{K_n} &= \sqrt{n}\sup_{y>0}\left|\frac{1}{n}\sum_{j=1}^{n}\min(Y_j,y) - \frac{1}{n}\sum_{j=1}^{n}\mathbf{1}(Y_j \le y)\right| \\ &= \overline{k_n}\sqrt{n}.\end{aligned}$$

Dá se ukázat, že $\overline{K_n}$ lze vyjádřit jako

$$\overline{K_n} = \sqrt{n}\max(K_n^+, K_n^-),$$

kde

$$\begin{aligned}\overline{K_n}^+ &= \max_{j=0,1,\dots,n-1}\left[\frac{1}{n}(Y_{(1)} + \dots + Y_{(j)}) + Y_{(j+1)}\left(1-\frac{j}{n}\right) - \frac{j}{n}\right], \\ \overline{K_n}^- &= \max_{j=0,1,\dots,n-1}\left[\frac{j}{n} - \frac{1}{n}(Y_{(1)} + \dots + Y_{(j)}) - Y_{(j)}\left(1-\frac{j}{n}\right)\right].\end{aligned}$$

Podobně jako u Kolmogorovovy-Smirnovovy testové statistiky uvedené v části 2.1 se stačí dívat na výraz

$$Y_{(j+1)}\left(1-\frac{j}{n}\right),$$

který má skoky v pořádkových statistikách.

Druhá statistika, která uvedenou vlastnost využívá je založena na integrální vzdálenosti a jedná se o statistiku Cramérova von Misesova typu, která má tvar

$$\overline{T_n} = n\int_0^\infty \left(\frac{1}{n}\sum_{j=1}^{n}\min(Y_j,y) - \frac{1}{n}\sum_{j=1}^{n}\mathbf{1}(Y_j \le y)\right)^2 e^{-y}dy.$$

Tvrzení 18. *Platí*

$$\begin{aligned}\overline{T_n} &= n\int_0^\infty \left(\frac{1}{n}\sum_{j=1}^{n}\min(Y_j,y) - \frac{1}{n}\sum_{j=1}^{n}\mathbf{1}(Y_j \le y)\right)^2 e^{-y}dy \\ &= \frac{1}{n}\sum_{j,k=1}^{n}\left[2 - 3e^{-\min(Y_j,Y_k)} - 2\min(Y_j,Y_k)(e^{-Y_j} + e^{-Y_k}) + 2e^{-\max(Y_j,Y_k)}\right].\end{aligned}$$

Důkaz. Rozepišme si

$$
\begin{aligned}
\overline{T_n} &= n\int_0^\infty \left(\frac{1}{n}\sum_{j=1}^n \min{(Y_j,y)} - \frac{1}{n}\sum_{j=1}^n \mathbf{1}(Y_j \leq y)\right)^2 e^{-y}dy \\
&= n\int_0^\infty \left[\left(\frac{1}{n}\sum_{j=1}^n \min{(Y_j,y)}\right)^2 - 2\left(\frac{1}{n}\sum_{j=1}^n \min{(Y_j,y)}\right)\frac{1}{n}\sum_{j=1}^n \mathbf{1}(Y_j \leq y)\right. \\
&+ \left.\left(\frac{1}{n}\sum_{j=1}^n \mathbf{1}(Y_j \leq y)\right)^2\right] e^{-y}dy \\
&= n\int_0^\infty \left[\frac{1}{n^2}\sum_{j,k=1}^n \min{(Y_j,y)}\min{(Y_k,y)} - \frac{2}{n^2}\sum_{j,k=1}^n \min{(Y_j,y)}\mathbf{1}(Y_k \leq y)\right. \\
&+ \left.\frac{1}{n^2}\sum_{j,k=1}^n \mathbf{1}(Y_j \leq y)\mathbf{1}(Y_k \leq y)\right] e^{-y}dy \\
&= \frac{1}{n}\sum_{j,k=1}^n \int_0^\infty [\min{(Y_j,y)}\min{(Y_k,y)} - 2\min{(Y_j,y)}\mathbf{1}(Y_k \leq y) \qquad (3.7) \\
&+ \mathbf{1}(Y_j \leq y)\mathbf{1}(Y_k \leq y)]\, e^{-y}dy.
\end{aligned}
$$

Tento výraz roztrhneme na jednotlivé integrály a dostáváme

$$
\begin{aligned}
\int_0^\infty \mathbf{1}(Y_j \leq y)\mathbf{1}(Y_k \leq y)e^{-y}dy &= \int_0^\infty \mathbf{1}(\max{(Y_j,Y_k)} \leq y)e^{-y}dy \\
&= \int_{\max{(Y_j,Y_k)}}^\infty e^{-y}dy \\
&= e^{-\max{(Y_j,Y_k)}}.
\end{aligned}
$$

Dále upravme prostřední integrál

$$
\int_0^\infty 2\min{(Y_j,y)}\mathbf{1}(Y_k \leq y)e^{-y}dy = \int_{Y_k}^\infty 2\min{(Y_j,y)}e^{-y}dy,
$$

pro $Y_j < Y_k$ dostáváme

$$
\int_{Y_k}^\infty 2Y_j e^{-y}dy = 2Y_j e^{-Y_k}
$$

a pro $Y_k < Y_j$ dostáváme

$$
\begin{aligned}
\int_{Y_k}^{Y_j} 2ye^{-y}dy + \int_{Y_j}^\infty 2Y_j e^{-y}dy &= 2(-Y_j e^{-Y_j} - e^{-Y_j} + Y_k e^{-Y_k} + e^{-Y_k}) + 2Y_j e^{-Y_j} \\
&= -2e^{-Y_j} + 2Y_k e^{-Y_k} + 2e^{-Y_k}.
\end{aligned}
$$

Zbývá upravit poslední výraz

$$
\int_0^\infty \min{(Y_j,y)}\min{(Y_k,y)}e^{-y}dy,
$$

pro $Y_j < Y_k$ dostáváme

$$\int_0^{Y_j} y^2 e^{-y} dy + \int_{Y_j}^{Y_k} y Y_j e^{-y} dy + \int_{Y_k}^{\infty} Y_k Y_j e^{-y} dy = 2 - Y_j e^{-Y_j} - 2e^{-Y_j} - Y_j e^{-Y_k}$$

a pro $Y_k < Y_j$ dostáváme

$$\int_0^{Y_k} y^2 e^{-y} dy + \int_{Y_k}^{Y_j} y Y_k e^{-y} dy + \int_{Y_j}^{\infty} Y_k Y_j e^{-y} dy = 2 - Y_k e^{-Y_k} - 2e^{-Y_k} - Y_k e^{-Y_j}.$$

Dosazením do rovnosti (3.7) dostaneme požadovanou rovnost.

□

Tvrzení 19. *Nechť $X_1, X_2, ..., X_n$ je náhodný výběr ze spojitého rozdělení s distribuční funkcí F_X. Testová statistika $\overline{K_n}$ konverguje v distribuci ke stejnému rozdělení, jaké je uvedené ve větě 7, neboli*

$$\sqrt{n} \sup_{y>0} \left| \frac{1}{n} \sum_{j=1}^{n} \min(Y_j, y) - \frac{1}{n} \sum_{j=1}^{n} \mathbf{1}(Y_j \leq y) \right| \xrightarrow[n\to\infty]{d} Z,$$

kde náhodná veličina Z má distribuční funkci danou předpisem

$$G(y) = 1 - 2\sum_{k=1}^{\infty} (-1)^{k+1} e^{-2k^2y^2},\ y > 0,$$

v opačném případě je $G(y) = 0$. Testovou statistika $\overline{T_n}$ má asymptoticky stejné rozdělení jako testová statistika T_n uvedená v kapitole 2.

Důkaz. Důkaz tohoto tvrzení lze najít v článku Baringhaus a Henze (2000).

□

Nulovou hypotézu zamítáme pro příliš velké hodnoty testové statistiky, tedy

$$H_0 \text{ zamítáme} \Leftrightarrow \overline{K_n}\sqrt{n} \geq k_{1-\alpha},$$

kde $k_{1-\alpha}$ označuje (1-α)-kvantil rozdělení s distribuční funkcí G uvedená ve větě 7. Tento test dodržuje hladinu α asymptoticky. Pro testovou statistiku $\overline{T_n}$ uvažujme stejné kritické hodnoty jako pro testovou statistiku T_n uvedenou v kapitole 2, kde jsme si tyto hodnoty nasimulovali (viz tabulka (2.3)).

Pro náš test dostáváme asymptotickou p-hodnotu danou předpisem

$$p = \inf(\alpha \in (0,1) : \overline{k_n}\sqrt{n} \geq k_{1-\alpha}) = 1 - G(\overline{k_n}\sqrt{n}),$$

kde $\overline{k_n}$ je pozorovaná hodnota testové statistiky $\overline{K_n}$. Analogicky bychom dostali p-hodnotu pro testovou statistiku $\overline{T_n}$.

Stejně jako pro dříve zmíněný Kolmogorovův-Smirnovův test a Cramérův von Misesův test platí, že oba testy jsou konzistentní proti každé alternativě, což plyne z tvrzení 16.

4. Praktická část

V této kapitole se podíváme na porovnání testů zmíněných v předchozích kapitolách. Konkrétně se podíváme na Kolmogorovův-Smirnovův test s testovou statistikou K_n, Cramérův von Misesův test s testovou statistikou T_n z kapitoly 2 a na test založený na Giniho indexu s testovou statistikou F_n, na statistiku $\overline{K_n}$ Kolmogorovova-Smirnovova typu a statistiku $\overline{T_n}$ Cramérova von Misesova typu z kapitoly 3. Jako první se budeme zabývat hladinou těchto testů za platnosti nulové hypotézy

$$\widetilde{H}_0 : \exists \lambda > 0 : F_X(x) = 1 - e^{-x\lambda} \; \forall x > 0.$$

V druhé části se pak podíváme na síly testů za platnosti alternativy

$$\widetilde{H}_1 : \forall \lambda > 0 \; \exists x > 0 : F_X(x) \neq 1 - e^{-x\lambda}.$$

Parametr $\lambda > 0$ je v obou případech neznámý (předem nespecifikovaný). Simulace jsou provedeny v programu R (R Core Team, 2022). K dílčím výpočtům je použita i knihovna goftest (Faraway a kol., 2021).

4.1 Porovnání hladin testů za platnosti $\widetilde{H}_0$

Předpokládejeme, že naše data skutečně pochází z exponenciálního rozdělení s parametrem $\lambda > 0$, neboli platí nulová hypotéza. Parametr $\lambda > 0$ je neznámý. Stejně jako v části 2.2 si neznámý parametr λ odhadneme pomocí metody maximální věrohodnosti. Víme, že maximálně věrohodným odhadem parametru λ je $1/\overline{X_n}$. Příslušnou p-hodnotu opět spočteme pomocí metody parametrického bootstrapu, kdy z daného náhodného výběru $X_1, X_2, ..., X_n$ z $\mathsf{Exp}(\lambda)$ odhadneme parametr $\hat{\lambda}_n$. Postupně si generujeme B bootstrapových výběrů $X_1^{(b)\star}, X_2^{(b)\star}, ..., X_n^{(b)\star}$ pro $b = 1, ..., 1\,000$ o stejném rozsahu n, které mají exponenciální rozdělení s parametrem $\hat{\lambda}_n$. Pro každý bootstrapový výběr si spočteme parametr $\hat{\lambda}_n^{(b)\star}$ (analogicky jako pro náhodný výběr $X_1, X_2, ..., X_n$) a pomocí tohoto $\hat{\lambda}_n^{(b)\star}$ spočteme příslušnou testovou statistiku. Jakmile budeme znát testovou statistiku, určíme p-hodnotu. Celý tento postup opakujeme Nkrát. Dostaneme tedy N p-hodnot z nichž určíme empirickou hladinu testu. Tu spočteme jako relativní četnost případů, kdy je p-hodnota menší než α, neboli jako počet p-hodnot, pro které zamítáme nulovou hypotézu, vydělený počtem opakování N. Tento postup použijeme pro Kolmogorovův-Smirnovův test a Cramérův von Misesův test.

Pro zbylé tři testy použijeme asymptotické rozdělení uvedené v kapitole 3.

Naše data budou pocházet z exponenciálního rozdělení s parametrem $\lambda \in \{1{,}5\}$ o rozsahu výběru $n \in \{20{,}50{,}100\}$. Počet bootstrapových výběrů bude $B = 1\,000$ a počet opakování bude $N = 1\,000$. Testujeme na hladině $\alpha = 0{,}05$.

Z tabulky (4.1) je patrné, že nám hladina testu jednotlivých statistik za platnosti nulové hypotézy nezávisí na parametru $\lambda > 0$. Pro všechny testy je vidět, že s rostoucím rozsahem výběru se hodnoty hladin pro jednotlivé testy příliš neliší a hladina testů je velmi podobná a blízká nominální hodnotě 5%.

$\lambda = 1$					
	K_n	T_n	F_n	$\overline{K_n}$	$\overline{T_n}$
$n = 20$	0,05	0,06	0,06	0,06	0,06
$n = 50$	0,04	0,05	0,06	0,05	0,05
$n = 100$	0,05	0,05	0,06	0,05	0,05
$\lambda = 5$					
	K_n	T_n	F_n	$\overline{K_n}$	$\overline{T_n}$
$n = 20$	0,05	0,06	0,06	0,06	0,06
$n = 50$	0,04	0,05	0,06	0,05	0,05
$n = 100$	0,05	0,05	0,06	0,05	0,05

Tabulka 4.1: Porovnání hladin testů za platnosti nulové hypotézy pro data pocházející z exponenciálního rozdělení s parametrem $\lambda \in \{1,5\}$ o rozsahu výběru $n \in \{20,50,100\}$.

4.2 Porovnání sil testů za platnosti $\widetilde{H}_1$

Uvažujme analogický postup jako v předchozí části. Tentokrát ovšem naše data pochází z gamma rozdělení $\Gamma(a,1)$, kde $a \in \{1/2,3/4,3/2\}$, neboli platí alternativa $\widetilde{H}_1$.

$\Gamma(1/2,1)$					
	K_n	T_n	F_n	$\overline{K_n}$	$\overline{T_n}$
$n = 20$	0,45	0,53	0,59	0,37	0,55
$n = 50$	0,84	0,90	0,88	0,81	0,88
$n = 100$	0,99	1,00	0,99	0,98	0,99

Tabulka 4.2: Porovnání sil testů za platnosti alternativy pro data pocházející z $\Gamma(1/2,1)$ o rozsahu výběru $n \in \{20,50,100\}$.

$\Gamma(3/2,1)$					
	K_n	T_n	F_n	$\overline{K_n}$	$\overline{T_n}$
$n = 20$	0,16	0,19	0,18	0,19	0,17
$n = 50$	0,37	0,42	0,44	0,39	0,42
$n = 100$	0,68	0,77	0,77	0,68	0,76

Tabulka 4.3: Porovnání sil testů za platnosti alternativy pro data pocházející z $\Gamma(3/2,1)$ o rozsahu výběru $n \in \{20,50,100\}$.

Z tabulek (4.2) a (4.3) je patrné, že s rostoucím rozsahem výběru se síla testů zvětšuje. Pro parametr $a = 1/2$ je síla testu pro $n = 50$ přibližně rovna 0,9, zatímco pro parametr $a = 3/2$ je síla testu pro $n = 50$ přibližně rovna 0,4. Tedy pokud je skutečné rozdělení gamma s parametrem $a = 3/2$, tak potřebujeme výrazně více pozorování, abychom byli schopni odhalit porušení nulové hypotézy.

$\Gamma(3/4{,}1)$					
	K_n	T_n	F_n	$\overline{K_n}$	$\overline{T_n}$
$n = 20$	0,10	0,12	0,16	0,08	0,13
$n = 50$	0,23	0,26	0,31	0,19	0,28
$n = 100$	0,42	0,49	0,51	0,37	0,47

Tabulka 4.4: Porovnání sil testů za platnosti alternativy pro data pocházející z $\Gamma(3/4{,}1)$ o rozsahu výběru $n \in \{20{,}50{,}100\}$.

Z definice 5 víme, že hustota gamma rozdělení $\Gamma(a, p)$ je tvaru

$$f(x) = \frac{1}{p^a \Gamma(a)} x^{a-1} e^{-x/p},$$

kde $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$. My v této kapitole uvažujeme parametr $p = 1$. Pokud bychom zvolili i parametr $a = 1$, dostáváme hustotu exponenciálního rozdělení s parametrem $\lambda = 1$. Exponenciální rozdělení je speciálním případem gamma rozdělení. Odsud plyne, že pro parametr $a = 3/4$ se dostáváme blíže nulové hypotéze než pro parametr $a = 1/2$ a jak vidíme porovnáním tabulek (4.2) a (4.4), dostáváme nižší sílu.

Z tabulky (4.2) je vidět, že pro $n = 100$ je jedno který z uvedených testů zvolíme, neboť jejich síla je přibližně rovna jedné. Z tabulek (4.3) a (4.4) pro $n = 100$ vidíme, že nejsilnější je Cramérův von Misesův test, test založený na Giniho indexu a test s testovou statistikou Cramérova von Misesova typu založený na střední zbytkové životnosti. Předchozí věta platí i pokud uvažujeme $n \in \{20{,}50\}$.

Nyní předpokládejme, že naše data pochází z Weibullova rozdělení $W(a, \sigma = 1)$, kde $a \in \{1/2, 3/4, 3/2\}$, neboli platí alternativa $\widetilde{H}_1$. Z definice Weibullova rozdělení 6 vidíme, že exponenciální rozdělení je jeho speciálním případem pro parametr $a = 1$.

$W(1/2{,}1)$					
	K_n	T_n	F_n	$\overline{K_n}$	$\overline{T_n}$
$n = 20$	0,88	0,90	0,90	0,80	0,90
$n = 50$	1,00	1,00	1,00	1,00	1,00
$n = 100$	1,00	1,00	1,00	1,00	1,00

Tabulka 4.5: Porovnání sil testů za platnosti alternativy pro data pocházející z $W(1/2{,}1)$ o rozsahu výběru $n \in \{20{,}50{,}100\}$.

V tabulce (4.5), kdy parametr $a = 1/2$, vidíme, že pro rozsah výběru $n = 100$ je síla testu rovna jedné pro všechny námi vybrané testy. Podobně je tomu i v tabulce (4.6), kdy parametr $a = 3/2$. Naopak pokud parametr $a = 3/4$, tak síla testu není rovna jedné pro žádné $n \in \{20{,}50{,}100\}$, v tomto případě potřebujeme více pozorování, abychom byli schopni odhalit porušení nulové hypotézy. Je vidět, že s rostoucím rozsahem výběru se síla testů opět zvětšuje.

$W(3/2,1)$					
	K_n	T_n	F_n	$\overline{K_n}$	$\overline{T_n}$
$n = 20$	0,38	0,50	0,48	0,45	0,46
$n = 50$	0,80	0,90	0,93	0,88	0,92
$n = 100$	0,98	1,00	1,00	1,00	1,00

Tabulka 4.6: Porovnání sil testů za platnosti alternativy pro data pocházející z $W(3/2{,}1)$ o rozsahu výběru $n \in \{20{,}50{,}100\}$.

$W(3/4,1)$					
	K_n	T_n	F_n	$\overline{K_n}$	$\overline{T_n}$
$n = 20$	0,27	0,27	0,36	0,19	0,31
$n = 50$	0,54	0,65	0,70	0,52	0,66
$n = 100$	0,84	0,90	0,93	0,85	0,92

Tabulka 4.7: Porovnání sil testů za platnosti alternativy pro data pocházející z $W(3/4{,}1)$ o rozsahu výběru $n \in \{20{,}50{,}100\}$.

Opět, jako tomu bylo u gamma rozdělení, nejsilnější je Cramérův von Misesův test, test založený na Giniho indexu a test s testovou statistikou Cramérova von Misesova typu založený na střední zbytkové životnosti.

Závěr

Cílem této práce bylo popsat testy dobré shody pro exponenciální rozdělení s parametrem $\lambda > 0$.

Jako první jsme zavedli definici exponenciálního rozdělení a empirické distribuční funkce včetně jejich vlastností. Tyto vlastnosti jsme pak využili v dalších částech práce. Následně jsme přešli na testy dobré shody založené na empirické distribuční funkci se známým parametrem $\lambda > 0$. Nasimulovali jsme si kritické hodnoty pro jednotlivé testy a porovnali je s tabelovanými hodnotami. Všechny zmíněné testy pro známý parametr $\lambda > 0$ jsme zobecnili i pro neznámý parametr λ pomocí metody tzv. parametrického bootstrapu. Neznámý parametr λ jsme odhadli pomocí metody maximální věrohodnosti. Zároveň s tím jsme ukázali i příklad. V další části jsme přešli na testy využívající některou z vlastností exponenciálního rozdělení. Jednalo se o testy založené na Giniho indexu a střední zbytkové životnosti. Odvodili jsme tvary jejich testových statistik a rozdělení za platnosti nulové hypotézy.

Na závěr jsme zjištěné teoretické poznatky zkoumali v simulační studii, kdy jsme sledovali hladinu jednotlivých testů za platnosti nulové hypotézy a sílu testů za platnosti alternativy. Pro alternativu jsme uvažovali gamma rozdělení a Weibullovo rozdělení, které mají exponenciální rozdělení jako speciální případ. Při zkoumání hladin testů jsme zjistili, že všechny hladiny jsou blízké nominální hodnotě 5%. Naopak pokud jsme se podívali na síly námi vybraných testů, tak nejsilnější testy jsou Cramérův von Misesův test, test založený na Giniho indexu a test s testovou statistikou Cramérova von Misesova typu založený na střední zbytkové životnosti.

Seznam použité literatury

Baringhaus, L. a Henze, N. (2000). Tests of fit for exponentiality based on a characterization via the mean residual life function. *Statistical Papers*, **41**, 225–236.

Borokov, A. A. (2019). *Mathematical statistics*. Routledge, London.

Dorfman, R. (1979). A Formula for the Gini coefficient. *The Review of Economics and Statistics*, **61**(1), 146–149.

Dupač, V. a Hušková, M. (2005). *Pravděpodobnost a matematická statistika*. Karolinum, Praha. ISBN 978-80-246-0009-3.

D'Agostino, R. B. a Stephens, M. A. (1986). *Goodness-of-fit techniques*. Marcel Dekker, Inc., New York. ISBN 0-8247-8705-6.

Faraway, J., Marsaglia, G., Marsaglia, J. a Baddeley, A. (2021). *goftest: Classical Goodness-of-Fit Tests for Univariate Distributions*. URL `https://CRAN.R-project.org/package=goftest`. R package version 1.2-3.

Gibbons, J. D. a Chakraborti, S. (2003). *Nonparametric statistical inference*. Fourth Edition. Dekker, New York, NY. ISBN 0-8247-4052-1.

Guess, F. a Proschan, F. (1988). Mean residual life: Theory and applications. *Handbook of Statistics*, pages 215–224.

Hájek, J. a Šidák, Z. (1967). *Theory of rank tests*. First Edition. Academic Press, New York.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, **19**(3), 293–325.

Lehmann, E. L. a Romano, J. P. (2005). *Testing statistical hypotheses*. Second edition. New York.

Omelka, M. (2022). *Modern statistical methods*. `https://www2.karlin.mff.cuni.cz/~omelka/Soubory/nmst434/nmst434_course-notes.pdf`, Datum stažení: 08-04-2023.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

Stephens, M. A. (1974). EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, **69**(347), 730–737.

Stute, W., Manteiga, W. G. a Quindimil, M. P. (1993). Bootstrap based goodness-of-fit-tests. *Metrika*, **40**(1), 243–256.

Zenga, M., Polisicchio, M. a Greselin, F. (2004). The variance of Gini's mean difference and its estimators. *Statistica*, **64**(3), 455–475.

Seznam obrázků

Seznam tabulek

A. Přílohy

A.1 Gama rozdělení

Definice 5 (Gama rozdělení)**.** *Nechť* $a > 0, p > 0$. Gama rozdělení $\Gamma(a,p)$ *definujeme pomocí hustoty*

$$f(x) = \begin{cases} \frac{1}{p^a\Gamma(a)} x^{a-1} e^{-x/p} & x > 0, \\ 0 & \textit{jinak,} \end{cases}$$

kde $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$.

A.2 Weibullovo rozdělení

Definice 6 (Weibullovo rozdělení)**.** *Nechť* $a > 0, \sigma > 0$. Weibullovo rozdělení $W(a,\sigma)$ *definujeme pomocí hustoty*

$$f(x) = \begin{cases} \frac{a}{\sigma} \left(\frac{x}{\sigma}\right)^{a-1} e^{-(x/\sigma)^a} & x > 0, \\ 0 & \textit{jinak.} \end{cases}$$