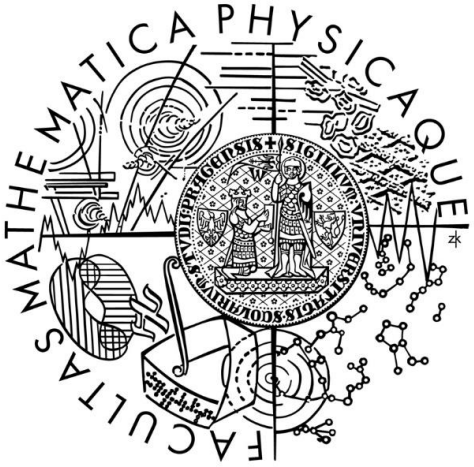
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Kateřina Malíková

Shapirův-Wilkův test normality

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce:	doc. RNDr. Arnošt Komárek, Ph.D.
Studijní program:	Matematika
Studijní obor:	Finanční matematika

Praha 2014

Děkuji panu doc. RNDr. Arnoštu Komárkovi, Ph.D., vedoucímu mé bakalářské práce, za odborné vedení, cenné rady a připomínky, vstřícnost, trpělivost a věnovaný čas, neboť tím významnou měrou přispěl ke zpracování této práce. Dále děkuji své rodině za poskytnuté zázemí a motivaci a všem, kteří mě při psaní práce jakýmkoliv způsobem podpořili.

Prohlašuji, že jsem tuto bakalářskou práci vypracovala samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

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 v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ..... Podpis autora

Název práce: Shapirův-Wilkův test normality

Autor: Kateřina Malíková

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: doc. RNDr. Arnošt Komárek, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: V práci představíme Shapirův-Wilkův test ověřující normalitu zkoumaného statistického výběru. Nejprve uvedeme základní informace o normálním rozdělení. Dále popíšeme samotný test a odvodíme tvar testové statistiky W a některé její analytické vlastnosti včetně dvou momentů a maximální a minimální přípustné hodnoty, kterých může nabývat. Seznámíme se též s některými aproximacemi různých koeficientů používanými k výpočtům a vyhodnocování testu a také s odhadem rozdělení testové statistiky Shapirova-Wilkova testu. Na závěr ukážeme příklad testování i způsob implementace v počítačových programech.

Klíčová slova: Shapirův-Wilkův test, test normality, normální rozdělení

Title: Shapiro-Wilk test of normality

Author: Kateřina Malíková

Department: Department of Probability and Mathematical Statistics

Supervisor: doc. RNDr. Arnošt Komárek, Ph.D., Department of Probability and Mathematical Statistics

Abstract: In this work we introduce Shapiro-Wilk normality test testing examined statistical sampling. At first, we state the basic information about the normal distribution. Further, we describe the test and we derive the shape of the test statistic W and some of its analytical properties, including two moments and the maximum and minimum allowable values that it may take. We also find some approximations of various coefficients used for calculations and evaluation of the test and also the estimate of the distribution of the test statistics of Shapiro-Wilk test. In conclusion, we show an example of testing and method of implementation in computer programs.

Keywords: Shapiro-Wilk test, normality test, normal distribution

Obsah

Použité značení

- $N(\mu,\sigma^2)$ normální rozdělení s parametry μ a σ^2
- $N(0,1)$ normované normální rozdělení
- $D(\boldsymbol{Y})$ rozdělení náhodného vektoru $\boldsymbol{Y}$
- $D(\boldsymbol{Y}) \sim N(\mu,\sigma^2)$ rozdělení $\boldsymbol{Y}$ je normální
- $Y_{(1)},\ldots,Y_{(n)}$ uspořádaný náhodný výběr o rozsahu n
- $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ výběrový průměr
- $\varphi(x)$ hustota normovaného normálního rozdělení v bodě x
- $\Phi(x)$ distribuční funkce normovaného normálního rozdělení v bodě x

Úvod

Statistika je v současnosti jedním z velmi rozšířených vědních oborů uplatňujícím se v nejrůznějších odvětvích. Používá spoustu metod k testovaní hypotéz o nashromážděných datech a k odhadům jednotlivých parametrů a také k tvorbě modelů využívaných jako podklady pro dobré rozhodování.

Mnoho statistických metod a testů je vázáno na nějaké předpoklady o datech, např. o jejich konkrétním rozdělení. Velká část procedur má sice varianty i pro výběry z neznámých rozdělení, můžeme-li se však spolehnout na některé konkrétní předpoklady, dosáhneme přesnějších a uspokojivějších výsledků.

Nejdříve je však třeba požadované předpoklady o datech ověřit. Existuje řada různých metod (grafy, výpočty, testy), z nichž nejspolehlivější bývá některá jejich kombinace.

K nejdůležitějším a nejpožadovanějším předpokladům pro různé statistické procedury patří normalita statistického výběru. To, že data pocházejí z normálního rozdělení, lze opět ověřit mnoha způsoby. Jedním z nich je Shapirův-Wilkův test, kterým se budeme v této práci podrobněji zabývat.

Shapirův-Wilkův test je založen na analýze rozptylu, neboť pro data z normálního (obecně symetrického) rozdělení platí některé užitečné vlastnosti, kterých test využívá.

V práci si nejprve připomeneme základní informace o normálním rozdělení, které využijeme pro pochopení dalšího textu. V další kapitole se seznámíme s Shapirovým-Wilkovým testem a jeho testovou statistikou W, kterou si postupně odvodíme s využitím vlastností normálního rozdělení. Dále popíšeme a dokážeme některé analytické vlastnosti testové statistiky, včetně rozsahu možných hodnot a dalších později využitých informací.

Důležitou součástí práce jsou aproximace váhových koeficientů a dalších konstant tvořících vyjádření testové statistiky, jejichž hodnoty se liší pro různé délky zkoumaných statistických výběrů, a také aproximace rozdělení celé testové statistiky. Na závěr si ještě ukážeme příklad výpočtu testové statistiky a rozhodnutí o zamítnutí či nezamítnutí nulové hypotézy a také způsob implementace testu v některých počítačových programech.

Kapitola 1

Normální rozdělení

Normální, neboli Gaussovo, rozdělení je jedním ze spojitých rozdělení pravděpodobnosti náhodných veličin.

Definice 1. *Hustota normálního rozdělení je dána vztahem*

$$\varphi(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right], \quad x \in \mathbb{R}.$$

Tato funkce zakreslená do grafu má tvar Gaussovy křivky. Funkce je symetrická kolem svého jediného lokálního maxima, kterým je střední hodnota.

Normální rozdělení je jednoznačně dáno svými dvěma parametry, kterými jsou střední hodnota μ ($\mu \in \mathbb{R}$) a rozptyl σ^2 ($\sigma^2 > 0$). Jsou-li tyto parametry rovny hodnotám $\mu = 0$ a $\sigma^2 = 1$, pak se jedná o tzv. normované normální rozdělení.

Distribuční funkci

$$\Phi(x) = \int_{-\infty}^{x} \varphi(t)dt$$

nelze vyjádřit pomocí primitivních funkcí. Její hodnoty pro normované normální rozdělení jsou tabelovány. Součty nezávislých stejně rozdělených náhodných veličin z některých rozdělení za určitých podmínek konvergují v distribuci k normálnímu rozdělení (viz Centrální limitní věta (Dupač a Hušková, 2009, Věta 4.9)).

Kapitola 2

Testová statistika W

2.1 Test

Předmětem testování testu pojmenovaného po S. S. Shapirovi a M. B. Wilkovi je rozdělení náhodného výběru ze spojitého rozdělení. Nechť $Y_1, \ldots, Y_n$ označuje nezávislé stejně rozdělené náhodné veličiny se střední hodnotou $\mu \in \mathbb{R}$ a rozptylem $0 < \sigma^2 < \infty$. Chceme zjistit, zda pozorovaný náhodný výběr $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ pochází z normálního rozdělení. Zformulujeme nulovou hypotézu a alternativu.

Nulová hypotéza.

$$H_0 : \ D(\boldsymbol{Y}) \sim N(\mu, \sigma^2).$$

Pozorovaný náhodný výběr pochází z normálního rozdělení s libovolnými parametry μ a σ^2.

Alternativní hypotéza.

$$H_1 : \ D(\boldsymbol{Y}) \not\sim N(\mu, \sigma^2).$$

Pozorovaný náhodný výběr nepochází z normálního rozdělení s libovolnými parametry μ a σ^2.

V dalším textu odvodíme podobu testové statistiky, která je založena na různých odhadech rozptylu. Jeden odhad je obecně platný pro všechny nezávislé stejně rozdělené náhodné veličiny, druhý je spojený s předpokladem, že výběr pochází z normálního rozdělení. Připomeňme si nejprve, co je to výběrový rozptyl.

2.2 Výběrový rozptyl

Označme $\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$ výběrový průměr náhodného výběru $Y_1, \ldots, Y_n$ z rozdělení se střední hodnotou $\mu \in \mathbb{R}$ a rozptylem $0 < \sigma^2 < \infty$.

Definice 2. *Hodnotu*

$$(S^*)^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

nazýváme výběrový rozptyl.

Tvrzení 1. *Výběrový rozptyl je nestranným odhadem rozptylu σ^2 nezávislých stejně rozdělených náhodných veličin $Y_1, \ldots, Y_n$.*

Důkaz. Snadný výpočet, který ukáže $\mathsf{E}\,(S^*)^2 = \sigma^2$.

❑

Důsledek.

$$S^2 = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \tag{2.1}$$

je nestranným odhadem $(n-1)\sigma^2$.

2.3 Odhad parametrů normálního rozdělení

K rozhodnutí o zamítnutí či nezamítnutí nulové hypotézy nám pomůže i odhad parametrů náhodného výběru za předpokladu normálního rozdělení. V celé podkapitole předpokládejme, že náhodný výběr $Y_1,\ldots,Y_n$ pochází z normálního rozdělení se střední hodnotou $\mu \in \mathbb{R}$ a rozptylem $0 < \sigma^2 < \infty$.

Nechť $Y_{(1)},\ldots,Y_{(n)}$ značí uspořádaný náhodný výběr a $\boldsymbol{Y}_{(n)} = (Y_{(1)},\ldots,Y_{(n)})^\top$ vektor těchto pořádkových statistik délky n.

Tvrzení 2. *Pokud* $Y_1,\ldots,Y_n \overset{iid}{\sim} N(\mu,\sigma^2)$, *potom lze psát*

$$Y_{(i)} = \mu + \sigma X_{(i)}, \quad i = 1,2,\ldots,n,$$

kde $X_1,\ldots,X_n \overset{iid}{\sim} N(0,1)$, *a*

$$\mathsf{E}\, Y_{(i)} = \mu + \sigma\, \mathsf{E}\, X_{(i)}, \quad i = 1,2,\ldots,n,$$

$$\mathrm{var} Y_{(i)} = \sigma^2 \mathrm{var} X_{(i)}, \quad i = 1,2,\ldots,n.$$

Důkaz. První část tvrzení je zřejmá. Ke druhé využijeme vlastností střední hodnoty a rozptylu.

$$\mathsf{E}\, Y_{(i)} = \mathsf{E}\,[\mu + \sigma X_{(i)}] = \mu + \sigma\, \mathsf{E}\, X_{(i)},$$

kde rovnosti plynou z vlastností integrálu. Totéž využijeme u rozptylu

$$\mathrm{var}\, Y_{(i)} = \mathrm{var}[\mu + \sigma X_{(i)}] = \sigma^2\, \mathrm{var}\, X_{(i)}.$$

❑

Zavedeme značení $\boldsymbol{m} = (m_1,\ldots,m_n)^\top$ pro vektor středních hodnot pořádkových statistik $(X_{(1)},\ldots,X_{(n)})$ z normovaného normálního rozdělení a $\boldsymbol{V} = (v_{ij})_{i,j=1}^n$ pro jejich kovarianční matici. Potom získáme následující důsledek.

Důsledek.

$$\begin{aligned} \mathsf{E}\, Y_{(i)} &= \mu + \sigma m_i, \quad i = 1,2,\ldots,n, \\ \mathrm{var}\, Y_{(i)} &= \sigma^2 v_{ii}, \quad i = 1,2,\ldots,n, \\ \mathrm{cov}(Y_{(i)},Y_{(j)}) &= \sigma^2 v_{ij}, \quad i,j = 1,2,\ldots,n, \quad i \neq j. \end{aligned} \tag{2.2}$$

Věta 3. *Uvažujme lineární regresi* $\mathsf{E}\, Y_{(i)} = \mu + \sigma m_i$, *kde* $\boldsymbol{\beta} = (\mu,\sigma)^\top$ *jsou neznámé regresní parametry, a* $\boldsymbol{Y}_{(n)} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{V})$. *Potom odhad* $\boldsymbol{\beta}$ *metodou nejmenších čtverců je*

$$\hat{\boldsymbol{\beta}} = (\hat{\mu},\hat{\sigma})^\top = \left(\frac{\mathbf{1}^\top \boldsymbol{V}^{-1}\boldsymbol{Y}_{(n)}}{\mathbf{1}^\top \boldsymbol{V}^{-1}\mathbf{1}}, \frac{\boldsymbol{m}^\top \boldsymbol{V}^{-1}\boldsymbol{Y}_{(n)}}{\boldsymbol{m}^\top \boldsymbol{V}^{-1}\boldsymbol{m}} \right)^\top,$$

kde $\mathbf{1} = (1,\ldots,1)^\top$ *je jednotkový vektor délky* n.

Důkaz. Ve větě uvažujeme lineární regresi (2.2), tedy regresní model

$$\mathsf{E}\, \boldsymbol{Y}_{(n)} = \boldsymbol{X}\boldsymbol{\beta},$$

kde

$$\boldsymbol{X} = \begin{pmatrix} 1 & m_1 \\ 1 & m_2 \\ \vdots & \vdots \\ 1 & m_n \end{pmatrix}.$$

Dle Aitkenovy věty (viz Zvára, 2008, část 2.8) je za předpokladu $\mathrm{var}\, \boldsymbol{Y}_{(n)} = \sigma^2\boldsymbol{V}$ nejlepším nestranným lineárním odhadem $\hat{\boldsymbol{\beta}}$ zobecněnou metodou nejmenších čtverců

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{V}^{-1}\boldsymbol{Y}_{(n)}.$$

Postupným výpočtem

$$(\boldsymbol{X}^\top \boldsymbol{V}^{-1}\boldsymbol{X})^{-1} = \begin{pmatrix} \frac{\boldsymbol{m}^\top \boldsymbol{V}^{-1}\boldsymbol{m}}{\mathbf{1}^\top \boldsymbol{V}^{-1}\mathbf{1}\boldsymbol{m}^\top \boldsymbol{V}^{-1}\boldsymbol{m} - (\mathbf{1}^\top \boldsymbol{V}^{-1}\boldsymbol{m})^2} & -\frac{\boldsymbol{m}^\top \boldsymbol{V}^{-1}\mathbf{1}}{\mathbf{1}^\top \boldsymbol{V}^{-1}\mathbf{1}\boldsymbol{m}^\top \boldsymbol{V}^{-1}\boldsymbol{m} - (\mathbf{1}^\top \boldsymbol{V}^{-1}\boldsymbol{m})^2} \\ -\frac{\mathbf{1}^\top \boldsymbol{V}^{-1}\boldsymbol{m}}{\mathbf{1}^\top \boldsymbol{V}^{-1}\mathbf{1}\boldsymbol{m}^\top \boldsymbol{V}^{-1}\boldsymbol{m} - (\mathbf{1}^\top \boldsymbol{V}^{-1}\boldsymbol{m})^2} & \frac{\mathbf{1}^\top \boldsymbol{V}^{-1}\mathbf{1}}{\mathbf{1}^\top \boldsymbol{V}^{-1}\mathbf{1}\boldsymbol{m}^\top \boldsymbol{V}^{-1}\boldsymbol{m} - (\mathbf{1}^\top \boldsymbol{V}^{-1}\boldsymbol{m})^2} \end{pmatrix},$$

$$\boldsymbol{X}^\top \boldsymbol{V}^{-1}\boldsymbol{Y}_{(n)} = \begin{pmatrix} \mathbf{1}^\top \boldsymbol{V}^{-1}\boldsymbol{Y}_{(n)} \\ \boldsymbol{m}^\top \boldsymbol{V}^{-1}\boldsymbol{Y}_{(n)} \end{pmatrix}$$

získáme

$$\hat{\mu} = \frac{\boldsymbol{m}^\top \boldsymbol{V}^{-1}(\boldsymbol{m}\mathbf{1}^\top - \mathbf{1}\boldsymbol{m}^\top)\boldsymbol{V}^{-1}\boldsymbol{Y}_{(n)}}{\mathbf{1}^\top \boldsymbol{V}^{-1}\mathbf{1}\boldsymbol{m}^\top \boldsymbol{V}^{-1}\boldsymbol{m} - (\mathbf{1}^\top \boldsymbol{V}^{-1}\boldsymbol{m})^2},$$

$$\hat{\sigma} = \frac{\mathbf{1}^\top \boldsymbol{V}^{-1}(\mathbf{1}\boldsymbol{m}^\top - \boldsymbol{m}\mathbf{1}^\top)\boldsymbol{V}^{-1}\boldsymbol{Y}_{(n)}}{\mathbf{1}^\top \boldsymbol{V}^{-1}\mathbf{1}\boldsymbol{m}^\top \boldsymbol{V}^{-1}\boldsymbol{m} - (\mathbf{1}^\top \boldsymbol{V}^{-1}\boldsymbol{m})^2}.$$

Využili jsme rovnost $\mathbf{1}^\top \boldsymbol{V}^{-1}\boldsymbol{m} = \boldsymbol{m}^\top \boldsymbol{V}^{-1}\mathbf{1}$ která plyne ze symetrie kovarianční matice $\boldsymbol{V}$. Pro normální (obecně symetrická) rozdělení platí $\mathbf{1}^\top \boldsymbol{V}^{-1}\boldsymbol{m} = 0$, odhady tedy můžeme ještě upravit na

$$\hat{\mu} = \frac{\boldsymbol{m}^\top \boldsymbol{V}^{-1}\boldsymbol{m}\mathbf{1}^\top \boldsymbol{V}^{-1}\boldsymbol{Y}_{(n)}}{\mathbf{1}^\top \boldsymbol{V}^{-1}\mathbf{1}\boldsymbol{m}^\top \boldsymbol{V}^{-1}\boldsymbol{m}} = \frac{\mathbf{1}^\top \boldsymbol{V}^{-1}\boldsymbol{Y}_{(n)}}{\mathbf{1}^\top \boldsymbol{V}^{-1}\mathbf{1}},$$

$$\hat{\sigma} = \frac{\mathbf{1}^\top \boldsymbol{V}^{-1}\mathbf{1}\boldsymbol{m}^\top \boldsymbol{V}^{-1}\boldsymbol{Y}_{(n)}}{\mathbf{1}^\top \boldsymbol{V}^{-1}\mathbf{1}\boldsymbol{m}^\top \boldsymbol{V}^{-1}\boldsymbol{m}} = \frac{\boldsymbol{m}^\top \boldsymbol{V}^{-1}\boldsymbol{Y}_{(n)}}{\boldsymbol{m}^\top \boldsymbol{V}^{-1}\boldsymbol{m}}.$$

❑

Jako důsledek věty 3 a Aitkenovy věty (viz Zvára, 2008, část 2.8) vyplývá, že odhad $\hat{\sigma}$ (2.3) je za předpokladu, že náhodný výběr $Y_1, \ldots, Y_n$ pochází z normálního rozdělení, nejlepším lineárním nestranným odhadem směrodatné odchylky σ.

Dále použijeme značení

$$R^2 = \boldsymbol{m}^\top \boldsymbol{V}^{-1} \boldsymbol{m},$$
$$C^2 = \boldsymbol{m}^\top \boldsymbol{V}^{-1} \boldsymbol{V}^{-1} \boldsymbol{m}$$

a získáme následující důsledek.

Důsledek. Označme

$$b = \frac{R^2 \hat{\sigma}}{C^2} = \frac{\boldsymbol{m}^\top \boldsymbol{V}^{-1} \boldsymbol{Y}_{(n)}}{\boldsymbol{m}^\top \boldsymbol{V}^{-1} \boldsymbol{V}^{-1} \boldsymbol{m}}. \tag{2.3}$$

Je-li náhodný výběr $Y_1, \ldots, Y_n$ z normálního rozdělení, pak b je až na konstantu nejlepším lineárním nestranným odhadem směrodatné odchylky σ.

2.4 Testová statistika

V části 2.2 a 2.3 jsme odvodili informace potřebné k sestavení testové statistiky pro test normality z části 2.1. Položme

$$W = \frac{b^2}{S^2} = \frac{R^4 \hat{\sigma}^2}{C^2 S^2} = \frac{\left(a^\top \boldsymbol{Y}_{(n)}\right)^2}{S^2} = \frac{\left(\sum_{i=1}^n a_i Y_{(i)}\right)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2},$$

kde platí vztah

$$\boldsymbol{a}^\top = (a_1, \ldots, a_n) = \frac{\boldsymbol{m}^\top \boldsymbol{V}^{-1}}{(\boldsymbol{m}^\top \boldsymbol{V}^{-1} \boldsymbol{V}^{-1} \boldsymbol{m})^{1/2}} \tag{2.4}$$

a b je odhad směrodatné odchylky určený rovnicí (2.3) a S^2 je odhad rozptylu definovaný vzorcem (2.1).

Pochází-li náhodný výběr $Y_1, \ldots, Y_n$ z normálního rozdělení, pak je testová statistika W podílem dvou odhadů téhož parametru, a to rozptylu σ^2.

Odhad S^2 je až na konstantu nestranným odhadem rozptylu pro výběry z libovolného spojitého rozdělení. Pro odhad b^2 téhož parametru jsme potřebovali předpoklad normality pozorovaných dat. Proto jsou-li data normálně rozdělená, bude hodnota testové statistiky s velkou pravděpodobností blízko 1. Naopak hodnota výrazně odlišná od 1 povede k zamítnutí nulové hypotézy.

Rozsáhlé studie testové statistiky založené na testování mnoha náhodných výběrů z různých rozdělení navíc ukázaly, že pozitivní korelace mezi čitatelem a jmenovatelem testové statistiky je u výběrů z normálního rozdělení větší než pro výběry s jiným rozdělením. Rozptyl hodnot testové statistiky W je tedy za platnosti nulové hypotézy menší něž za platnosti alternativy.

2.5 Vlastnosti koeficientů a_i

Nyní odvodíme některé vlastnosti koeficientů a_i z testové statistiky W. Těchto znalostí využijeme v dalších kapitolách, zvláště k důkazu některých tvrzení, ke zjednodušení výpočtu a k pochopení různých aproximací.

Tvrzení 4. *Koeficienty a_i z testové statistiky W lze vyjádřit jako*

$$a_i = \frac{\sum_{j=1}^{n} m_j v_{ij}}{C},$$

kde $C = (\boldsymbol{m}^\top \boldsymbol{V}^{-1}\boldsymbol{V}^{-1}\boldsymbol{m})^{\frac{1}{2}}$.

Z vyjádření vidíme, že k určení jednotlivých koeficientů a_i potřebujeme znát vektor středních hodnot $\boldsymbol{m}$ a kovarianční matici $\boldsymbol{V}$. Prvky kovarianční matice jsou vypočítané a tabelované pro výběry do délky $n = 20$. Pro rozsáhlejší výběry je lepší místo složitých výpočtů použít aproximace, jejichž vyčíslení je mnohem jednodušší. Navíc odhady jsou velmi přesné. Těmito aproximacemi se budeme zabývat v kapitole 4.

Tvrzení 5.

$$\boldsymbol{a}^\top \boldsymbol{a} = \sum_{i=1}^{n} a_i^2 = 1.$$

Důkaz.

$$\boldsymbol{a}^\top = (a_1, \ldots, a_n) = \frac{\boldsymbol{m}^\top \boldsymbol{V}^{-1}}{(\boldsymbol{m}^\top \boldsymbol{V}^{-1}\boldsymbol{V}^{-1}\boldsymbol{m})^{1/2}}.$$

$$\boldsymbol{a}^\top \boldsymbol{a} = \frac{\boldsymbol{m}^\top \boldsymbol{V}^{-1}}{(\boldsymbol{m}^\top \boldsymbol{V}^{-1}\boldsymbol{V}^{-1}\boldsymbol{m})^{1/2}} \frac{(\boldsymbol{m}^\top \boldsymbol{V}^{-1})^\top}{(\boldsymbol{m}^\top \boldsymbol{V}^{-1}\boldsymbol{V}^{-1}\boldsymbol{m})^{1/2}} = \frac{(\boldsymbol{m}^\top \boldsymbol{V}^{-1}\boldsymbol{V}^{-1}\boldsymbol{m})}{(\boldsymbol{m}^\top \boldsymbol{V}^{-1}\boldsymbol{V}^{-1}\boldsymbol{m})} = 1.$$

❑

Tvrzení 6. *Platí vztah*

$$a_i = -a_{n-i+1}, \quad i = 1, \ldots, n.$$

Důkaz. Náhodné veličiny $-X_1, \ldots, -X_n$ mají zřejmě stejné rozdělení jako veličiny $X_1, \ldots, X_n$. K nim příslušné pořádkové statistiky však jsou uspořádány v opačném pořadí. Tedy $X_{(i)}$ má stejné rozdělení jako $-X_{(n-i+1)}$. Odtud

$$m_i = -m_{n-i+1}, \quad i = 1, \ldots, n.$$

Ze stejného důvodu při záměně indexů řádků a sloupců kovarianční matice $\boldsymbol{V}$ a tedy i její inverze zůstanou tyto matice nezměněny. Vyčíslujeme-li hodnoty a_i a $-a_{n-i+1}$, sčítáme pro oba výběry tytéž hodnoty, které podělíme stejnou konstantou. Rovnost tedy platí.

❑

Důsledek. Platí $\sum_{i=1}^{n} a_i = 0$.

Poznámka. Koefient a_n (a tedy $|a_1|$) má vždy nejvyšší hodnotu a dále se hodnota a_{n-i+1} (a tedy $|a_i|$) s rostoucím i postupně snižuje. Pro náhodné výběry liché délky $n = 2k + 1$ je vždy prostřední koeficient

$$a_{k+1} = 0.$$

Kapitola 3

Některé vlastnosti testové statistiky W

V této kapitole odvodíme některé analytické vlastnosti testové statistiky W, které byly poprvé odvozeny v práci Shapira a Wilka (Shapiro a Wilk, 1965). Nejdříve se seznámíme s vlastností, díky které platí další důležitá tvrzení, a jejími důsledky.

Věta 7. *Testová statistika W je invariantní vůči změně polohy a měřítka.*

Důkaz. Tvrzení jinými slovy říká, že hodnota testové statistiky pro náhodný vektor $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ a pro náhodný vektor $\boldsymbol{Z} = (Z_1, \ldots, Z_n)^\top$ závislý na $\boldsymbol{Y}$ způsobem $\boldsymbol{Z} = a\boldsymbol{Y} + b$, kde konstanta a určuje měřítko a konstanta b polohu, je stejná. Tedy $W(\boldsymbol{Y}) = W(a\boldsymbol{Y} + b)$.

Označme W_Z testovou statistiku pro vektor náhodných veličin $\boldsymbol{Z}$ a W_Y testovou statistiku pro vektor náhodných veličin $\boldsymbol{Y}$. Podobným způsobem rozlišme i b_Z^2, b_Y^2, S_Z^2, S_Y^2, čitatele a jmenovatele těchto statistik.

Jmenovatele upravíme snadno, s pomocí

$$\bar{Z} = \frac{1}{n}\sum_{i=1}^{n}(cY_i + d) = d + \frac{c}{n}\sum_{i=1}^{n} Y_i = d + c\bar{Y}$$

získáme

$$S_Z^2 = \sum_{i=1}^{n}(Z_i - \bar{Z})^2 = \sum_{i=1}^{n}(cY_i + d - c\bar{Y} - d)^2 = d^2\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = d^2 S_Y^2.$$

Pro úpravu čitatele využijeme vlastnost koeficientů z tvrzení 6, $-a_i = a_{n-i+1}$, pak

$$\begin{aligned} b_Z &= \sum_{i=1}^{n} a_{n-i+1}(Z_{n-i+1} - Z_i) = \sum_{i=1}^{n} a_{n-i+1}(cY_{n-i+1} + d - cY_i - d) = \\ &= \sum_{i=1}^{n} a_{n-i+1} d(Y_{n-i+1} - Y_i) = d b_Y, \end{aligned}$$

kde první rovnost platí pro n sudé i liché, neboť pro $n = 2k + 1$ je vždy hodnota prostředního koeficientu $a_{k+1} = 0$. Dosazením

$$W_Z = \frac{b_Z^2}{S_Z^2} = \frac{d^2 b_Y^2}{d^2 S_Y^2} = \frac{b_Y^2}{S_Y^2} = W_Y.$$

❑

Důsledek. Pro výběry z normálního rozdělení má testová statistika W rozdělení, které závisí pouze na velikosti výběru n.

Důsledek. Pro výběry z normálního rozdělení je testová statistika W statisticky nezávislá na S^2 a $\bar{Y}$.

Důsledek.

$$\mathsf{E}\,(W^r) = \frac{\mathsf{E}\,(b^{2r})}{\mathsf{E}\,(S^{2r})}. \tag{3.1}$$

U testové statistiky W dokážeme odvodit její maximální a dokonce i minimální možnou hodnotu.

Věta 8. *Maximální hodnota, které testová statistika W může nabývat, je 1.*

Důkaz. Podle věty 7 je testová statistika W invariantní vůči změně polohy. Bez újmy na obecnosti tedy předpokládajme, že $\bar{Y} = 0$. Potom

$$W = \left[\sum_{i=1}^{n} a_i Y_{(i)}\right]^2 / \sum_{i=1}^{n} Y_i^2.$$

Z Cauchy-Schwarzovy nerovnosti máme

$$\left(\sum_i a_i Y_{(i)}\right)^2 \leq \sum_i a_i^2 \sum_i Y_{(i)}^2 = \sum_i Y_{(i)}^2.$$

Rovnost v posledním vztahu plyne z $\sum_i a_i^2 = a^\top a = 1$ z tvrzení 5. Hodnota testové statistiky je tedy omezena 1. Této maximální hodnoty dosáhne v případě, že $y_i = \eta a_i$ pro libovolné η. Opět s využitím zmiňované vlastnosti máme

$$\left(\eta \sum_i a_i^2\right)^2 = \eta^2 \sum_i a_i^2.$$

❑

Věta 9. *Minimální hodnota, které testová statistika W může nabývat, je rovna $na_1^2/(n-1)$.*

Důkaz. Věta byla původně intuitivním dohadem ověřeným numerickými studiemi. Níže uvedený důkaz přidal C. L. Mallows (Shapiro a Wilk, 1965, část 2.3, Lemma 3).

Dle věty 7 je testová statistika W invariantní vůči změně polohy a měřítka. Předpokládejme tedy, že $\sum Y_i = 0$ a $\sum a_i Y_{(i)} = 1$. Tím získáme $W = 1/\sum Y_i^2$. Úloha se nám tedy mění na nalezení maximální hodnoty $\sum_{i=1}^{n} Y_i^2$ za uvedených

podmínek. Je to konvexní funkce na konvexní množině. Maximum se nachází v jednom z vrcholů, kterými jsou

$$\left(\frac{n-1}{na_1}, \frac{-1}{na_1}, \ldots, \frac{-1}{na_1}\right)^\top,$$

$$\left(\frac{n-2}{n(a_1+a_2)}, \frac{n-2}{n(a_1+a_2)}, \frac{-2}{n(a_1+a_2)}, \ldots, \frac{-2}{n(a_1+a_2)}\right)^\top,$$

$$\vdots$$

$$\left(\frac{1}{n(a_1+\cdots+a_{n-1})}, \frac{1}{n(a_1+\cdots+a_{n-1})}, \ldots, \frac{-(n-1)}{n(a_1+\cdots+a_{n-1})}\right)^\top.$$

Numerickým výpočtem můžeme zjistit, že hledaná maximální hodnota se nachází v prvním uvedeném bodě. Součtem získáme

$$\sum_{i=1}^{n} Y_i^2 = \frac{(n-1)^2}{n^2 a_1^2} + \sum_{i=1}^{n-1}\left(\frac{-1}{na_1}\right)^2 = \frac{(n-1)^2+(n-1)}{n^2 a_1^2} = \frac{(n-1)}{na_1^2}.$$

Minimální hodnota testové statistiky je

$$W = 1/\sum Y_i^2 = na_1^2/(n-1)$$

a tvrzení je dokázáno.

❑

Přesné rozdělení testové statistiky W nelze přímo odvodit. Známe alespoň některé její momenty. Aproximací rozdělení se budeme zabývat v kapitole 4.

Věta 10. *Momenty $\mathsf{E}(W^{\frac{1}{2}})$ a $\mathsf{E}(W)$ testové statistiky W jsou rovny*

$$\mathsf{E}\, W^{\frac{1}{2}} = \frac{R^2\Gamma[\frac{1}{2}(n-1)]}{C\Gamma(\frac{1}{2}n)\sqrt{2}},$$

$$\mathsf{E}\, W = \frac{R^2(R^2+1)}{C^2(n-1)},$$

kde $R^2 = \boldsymbol{m}^\top \boldsymbol{V}^{-1}\boldsymbol{m}$ a $C^2 = \boldsymbol{m}^\top \boldsymbol{V}^{-1}\boldsymbol{V}^{-1}\boldsymbol{m}$.

Důkaz. Z důsledku věty 7 víme

$$\mathsf{E}\, W^{\frac{1}{2}} = \frac{\mathsf{E}\, b}{\mathsf{E}\, S},$$

$$\mathsf{E}\, W = \frac{\mathsf{E}\, b^2}{\mathsf{E}\, S^2}.$$

Oba čitatele upravíme

$$\mathsf{E}\, b = \mathsf{E}\left(\frac{R^2\hat{\sigma}}{C}\right) = \frac{R^2}{C}\,\mathsf{E}\,\hat{\sigma} = \frac{R^2}{C}\sigma,$$

$$\mathsf{E}\ b^2 = \mathsf{E}\left(\frac{R^4\hat{\sigma}^2}{C^2}\right) = \frac{R^4}{C^2}\,\mathsf{E}\ \hat{\sigma}^2 = \frac{R^4}{C^2}\left[\mathrm{var}(\hat{\sigma}) + (\mathsf{E}\ \hat{\sigma})^2\right] = \sigma^2 R^2(R^2+1)/C^2.$$

V poslední rovnosti využíváme vztah $\mathrm{var}(\hat{\sigma}) = \sigma^2/\boldsymbol{m}^\top \boldsymbol{V}^{-1}\boldsymbol{m} = \sigma^2/R^2$. Jmenovatele prvních dvou vztahů získáme ze znalosti rozdělení

$$Z = \frac{(n-1)(S^*)^2}{\sigma^2} = \frac{S^2}{\sigma^2} \sim \chi^2_{n-1},$$

kde S^* je výběrový rozptyl. Odtud

$$\mathsf{E}\ \frac{S}{\sigma} = \mathsf{E}\ \sqrt{Z} = \int_0^\infty \sqrt{Z} f_{\chi^2_{n-1}}(Z)dz = \int_0^\infty \sqrt{Z}\frac{1}{2^{(n-1)/2}\Gamma(\frac{n-1}{2})} Z^{(n-3)/2}e^{Z/2}dz =$$
$$= \frac{\sqrt{2}\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})}\int_0^\infty \frac{1}{2^{n/2}\Gamma(\frac{n}{2})} Z^{(n-2)/2}e^{Z/2}dz = \frac{\sqrt{2}\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})}\int_0^\infty f_{\chi^2_n}dz = \frac{\sqrt{2}\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})},$$

kde $f_{\chi^2_r}(Z) = \frac{1}{2^{r/2}\Gamma(r/2)} Z^{(r-2)/2}e^{Z/2}$ je hustota χ^2_r, a tedy

$$\mathsf{E}\ S = \sigma\sqrt{2}\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})}.$$

Z důsledku tvrzení 1 máme též

$$\mathsf{E}\ S^2 = (n-1)\sigma^2.$$

Z podílů jednotlivých výsledků získáme dokazované tvrzení.

❑

Kapitola 4

Aproximace

Jak již bylo zmíněno v části 2.5, k určení koeficientů a_i a některých dalších hodnot pořebujeme znát prvky kovarianční matice $\boldsymbol{V}$, které jsou tabelované pro výběry do délky $n = 20$. Pro rozsáhlejší výběry se používají aproximace. Některé z nich uvedeme v této kapitole. Dále také popíšeme, jakým způsobem bylo aproximováno rozdělení testové statistiky W. Začneme s odhady, které ve své práci uvedli Shapiro a Wilk (Shapiro a Wilk, 1965).

4.1 Shapirova-Wilkova aproximace

Označme

$$\boldsymbol{a}^* = (a_1^*, \ldots, a_n^*)^\top = \boldsymbol{m}^\top \boldsymbol{V}^{-1}$$

čitatele ze vzorce (2.4), který bude předmětem několika aproximací. Odsud také přímo vidíme, že jmenovatel z téhož vzorce $C^2 = \boldsymbol{m}^\top \boldsymbol{V}^{-1}\boldsymbol{V}^{-1}\boldsymbol{m} = (\boldsymbol{a}^*)^\top \boldsymbol{a}^*$.

Aproximace (Shapiro-Wilk, aproximace a_i^*)**.**

$$\hat{a}_i^* = 2m_i, \quad i = 2,3,\ldots,n.$$

Koeficienty a_i^* jsou závislé na vektoru středních hodnot $\boldsymbol{m}$ a kovarianční matici $\boldsymbol{V}$. Shapiro a Wilk použili k aproximaci pouze první z těchto nezbytných parametrů pro výpočet, střední hodnotu, a to pro všechna i, pro která platí $2 \leq i \leq n-1$.

Porovnání. Z porovnání přesných hodnot a_i^* a odhadnutých hodnot $\hat{a}_i^2$, kde $i \neq 1$, pro výběry délky $n \leq 20$ vyplývá, že chyby aproximací jsou jen velmi malé (pro $n = 20$ viz tabulka 4.1) a s rostoucím n jsou odhady přesnější. Proto věříme, že užití těchto aproximací je vhodné i pro výběry délky $n > 20$.

Hodnoty $\hat{a}_i^*$ je třeba pro odvození testové statistiky W nejprve vydělit hodnotou $C = (\boldsymbol{m}^\top \boldsymbol{V}^{-1}\boldsymbol{V}^{-1}\boldsymbol{m})^{1/2}$, která však pro dlouhé výběry opět závisí na neznámé kovarianční matici V. Můžeme využít vlastnosti $C^2 = (\boldsymbol{a}^*)^\top \boldsymbol{a}^*$, nebo použít některou jinou aproximaci této konstanty.

Z důvodu možných odlehlých pozorování je také vhodné váhy náhodných veličin s nejnižšími a nejvyššími hodnotami aproximovat zvlášť.

n	i	$\hat{a}_i^*$	a_i^*	${\tilde{a}_i}^*$	n	i	$\hat{a}_i^*$	${\tilde{a}_i}^*$
20	1	−4,223	−4,2013	−4,215	30	1	−4,655	−4,671
	2	−2,815	−2,8494	−2,764		2	−3,231	−3,170
	3	−2,262	−2,2765	−2,237		3	−2,730	−2,768
	4	−1,842	−1,8502	−1,820		4	−2,357	−2,369
	5	−1,491	−1,4960	−1,476		5	−2,052	−2,013
	6	−1,181	−1,1841	−1,169		6	−1,789	−1,760
	7	−0,897	−0,8990	−0,887		7	−1,553	−1,528
	8	−0,630	−0,6314	−0,622		8	−1,338	−1,334
	9	−0,374	−0,3784	−0,370		9	−1,137	−1,132
	10	−0,124	−0,1243	−0,123		10	−0,947	−0,941
						11	−0,765	−0,759
						12	−0,589	−0,582
						13	−0,418	−0,413
						14	−0,249	−0,249
						15	−0,083	−0,082

Tabulka 4.1: Porovnání aproximací a^*. *Převzato z článku Shapiro a Wilk (1965).*

Aproximace (Shapiro-Wilk, aproximace a_1)**.**

$$\hat{a}_1^2 = \hat{a}_n^2 = \begin{cases} \dfrac{\Gamma\left(\frac{1}{2}n\right)}{\sqrt{2}\Gamma\left[\frac{1}{2}(n+1)\right]}, & n \leq 20, \\ \dfrac{\Gamma\left[\frac{1}{2}(n+1)\right]}{\sqrt{2}\Gamma\left(\frac{1}{2}n+1\right)}, & n > 20. \end{cases}$$

Rovnost mezi $\hat{a}_1^*$ a $\hat{a}_n^*$ vyplývá z tvrzení 6. Hodnoty koeficientů závisí na délce výběru n. Pozměněná aproximace pro výběry délky větší než 20 byla zapříčiněna srovnáním s jinými aproximacemi (např. Plackettova aproximace, viz následující část 4.2), v této podobě se jeví jako přesnější.

Porovnání. Pro $n \leq 20$ porovnáme přesné a odhadnuté hodnoty a_1^2 a $\hat{a}_1^2$. Chyby jsou jen velmi malé (viz tabulka 4.2). Vidíme ale, že hodnoty odhadů rostou rychleji než hodnoty skutečné. Pro velká n by se mohla aproximace od přesných hodnot odchýlit. Bylo tedy vhodné odhady pro $n > 20$ upravit dle porovnání s jinými aproximacemi.

Aproximace (Shapiro-Wilk, aproximace C^2 a R^2)**.**

$$C^2 = -2{,}722 + 4{,}083n,$$

$$R^2 = -2{,}411 + 1{,}981n.$$

C^2 a stejně tak i R^2 se jako funkce proměnné n chovají přibližně lineárně. Použijeme tedy metodu nejmenších čtverců na známé hodnoty těchto funkcí a zíkáme aproximace použitelné i pro velká n.

Porovnání. K ověření vhodnosti aproximace pro větší hodnoty n využijeme srovnání extrapolace aproximující rovnice s hodnotami získanými pomocí rovnice

n	přesně	aproximace	n	přesně	aproximace
7	0,388	0,392	14	0,276	0,272
8	0,366	0,365	15	0,265	0,261
9	0,347	0,343	16	0,256	0,254
10	0,329	0,324	17	0,247	0,245
11	0,314	0,308	18	0,239	0,237
12	0,300	0,295	19	0,231	0,231
13	0,287	0,283	20	0,224	0,226

Tabulka 4.2: Porovnání a_1^2 a aproximace $\hat{a}_1^2$. *Převzato z článku Shapiro a Wilk (1965).*

$C^2 = (\boldsymbol{a}^*)^\top \boldsymbol{a}^*$, kde použijeme

$$(\hat{a}_1^*)^2 = \frac{\hat{a}_1^2}{1 - 2\hat{a}_1^2} \sum_{i=2}^{n-1} (\hat{a}_i^*)^2.$$

Např. pro $n = 30$ dostaneme hodnoty 119,77 a 120,47, které se liší jen nepatrně, což podporuje důvěru v obě aproximace.

4.2 Plackettova aproximace

Další možnou aproximací je aproximace R. L. Placketta, který navrhl obecnější odhady pro různá rozdělení. Zde uvedeme pouze odhady ve tvaru vhodném pro náš případ s normálním rozdělením.

Aproximace (Plackett, aproximace a_i^*)**.**

$$\tilde{a}_j^* = n m_j [\Phi(m_{j+1}) - \Phi(m_{j-1})], \quad j = 2,3,\ldots,n-1,$$

$$\tilde{a}_j^* = n \frac{m_j \varphi(m_j)^2}{\Phi(m_j)} + m_j^2 \varphi(m_j) - \varphi(m_j) + m_j [\Phi(m_{j+1}) - \Phi(m_j)], \quad j = 1,$$

kde $\Phi(m_j)$ *je distribuční funkce normovaného normálního rozdělení v bodě* m_j *a* $\varphi(m_j)$ *je hustota normovaného normálního rozdělení v bodě* m_j *a* $\tilde{a}_1^* = -\tilde{a}_n^*$.

Porovnání. Pokud srovnáme výsledky Plackettovy aproximace $\tilde{a}_i^*$, výše zmíněné aproximace $\hat{a}_i^*$ a přesné hodnoty pro $n \leq 20$, zjistíme, že oba odhady jsou velmi dobré. Např. pro $n = 20$ (viz tabulka 4.1) vidíme, že rozdíly v obou aproximacích jsou zanedbatelné. A srovnáme-li dále výsledky obou aproximací pro větší rozsah výběru n, např. pro $n = 30$, výsledky jsou opět velmi podobné. To jen dokazuje, že těmto aproximacím můžeme směle důvěřovat.

Aproximace (Plackett, aproximace R^2)**.**

$$\tilde{R}^2 = 2 \frac{m_1^2 \varphi(m_1)^2}{\Phi(m_1)} + m_1^3 \varphi(m_1) + m_1 \varphi(m_1) - 2\Phi(m_1) + 1.$$

Porovnání. Placketovy odhady hodnot R^2 jsou opět velmi dobré. Např. odhad pro $n = 20$ je 36,09, aproximace zmíněná výše dává pro tento případ 37,21, skutečná hodnota je 37,26. Existující rozdíly v aproximacích nejsou příliš velké, tedy výsledky se spíše podporují a lze je použít.

4.3 Aproximace rozdělení testové statistiky W

Zatím jsme se nezmínili o rozdělení testové statistiky. Neznáme ho ovšem přesně. Chceme-li aproximovat rozdělení testové statistiky W, vycházíme ze znalosti pouze prvního a druhého momentu pořádkové statistiky, z čehož odvodíme pro testovou statistiku W jen momenty $\mathsf{E}\, W^{\frac{1}{2}}$ a $\mathsf{E}\, W$. To nám znemožňuje použít některé metody pro aproximaci a rozšíření znalostí např. o další momenty testové statistiky. Jako nejvhodnější pro odvození rozdělení testové statistiky W se jeví empirické odhady. Způsob, jakým byla aproximace provedena, je popsán níže.

Pro momenty testové statistiky zaveďme značení

$$\mu_{\frac{1}{2}} = \mathsf{E}\,(W^{\frac{1}{2}}), \qquad \mu_1 = \mathsf{E}\,(W),$$

$$\mu_2 = \mathsf{E}\,(W - \mathsf{E}\, W)^2,$$

$$\mu_3 = \frac{\mathsf{E}\,(W - \mathsf{E}\, W)^3}{(\operatorname{var} W)^{3/2}}, \qquad \mu_4 = \frac{\mathsf{E}\,(W - \mathsf{E}\, W)^4}{(\operatorname{var} W)^2}.$$

Aproximace (empirické odhady, aproximace testové statistiky W)**.** *Pro řešení tohoto problému bylo použito velké množství náhodných výběrů z normálního rozdělení. Přesněji pro* $3 \leq n \leq 20$ *byl počet výběrů roven* $m = 5000$*, pro* $21 \leq n \leq 50$ *byla hodnota* $m = [100000/n]$.

Pro $3 \leq n \leq 50$ *byly vypočteny opakované hodnoty testové statistiky* W *a dále byly určeny empirické kvantily pro každé* n*. Takto získané body distribučních funkcí testových statistik* W *byly proloženy hladkými křivkami, čímž je odhadnuto rozdělení* W *pro všechna zmiňovaná* n*. Počet potřebných náhodných výběrů pro odvození rozdělení testové statistiky byl určen dle předchozích pozorování a zkušeností odborníků.*

Porovnání. Kontrolní porovnání momentů $\hat{\mu}_i$ odhadnutých z aproximovaného rozdělení testové statistiky s odpovídajícími známými teoretickými momenty μ_i potvrdilo vhodnost a správnost použití empirického odhadování, neboť výsledky jsou velmi přesné. Stejnou metodou lze rozšířit znalosti o rozdělení W i o další momenty (viz tabulka 4.3).

n	$\mu_{\frac{1}{2}}$ $\cdot 10^{-4}$	$\hat{\mu}_{\frac{1}{2}}$ $\cdot 10^{-4}$	$\mu^*_{\frac{1}{2}}$ $\cdot 10^{-4}$	μ_1 $\cdot 10^{-4}$	$\hat{\mu}_1$ $\cdot 10^{-4}$	$\tilde{\mu}_1$ $\cdot 10^{-4}$	μ^*_1 $\cdot 10^{-4}$	$\hat{\mu}_2$ $\cdot 10^{-6}$	$\sqrt{\tilde{\mu}_2}$ $\cdot 10^{-4}$	$\sqrt{\mu^*_2}$ $\cdot 10^{-4}$	$\hat{\mu}_3$	$\tilde{\mu}_3$	μ^*_3	$\hat{\mu}_4$	$\tilde{\mu}_4$	μ^*_4
3	9549	9547	9557	9135	9130	—	9150	5698	—	754	−0,593	—	−0,592	2,375	—	2,064
4	9486	9489	9482	9012	9019	—	9006	5166	—	717	−0,894	—	−0,913	3,723	—	3,542
5	9494	9491	9484	9026	9021	—	9007	4491	—	668	−0,818	—	−1,075	7,813	—	4,223
6	9521	9525	9516	9072	9082	—	9065	3390	—	591	−1,179	—	−1,161	5,430	—	4,832
7	9547	9545	9551	9123	9120	9113	9130	2995	542	535	−1,323	−1,23	−1,226	6,410	5,44	5,055
8	9574	9575	9575	9174	9175	9174	9175	2470	499	496	−1,384	−1,27	−1,235	7,109	5,44	5,127
9	9600	9596	9601	9221	9215	9227	9225	2293	466	463	−1,599	−1,48	−1,403	8,448	6,49	6,162
10	9622	9620	9626	9264	9260	9277	9271	1972	432	426	−1,666	−1,47	−1,408	9,281	6,45	6,365
15	9706	9705	9705	9424	9422	9423	9421	1023	323	320	−1,888	−1,52	−1,631	16,738	6,47	8,241
20	9757	9760	9760	9523	9527	9523	9527	651	253	251	−2,276	−1,55	−1,670	32,591	7,76	8,867
30	—	9811	9824	—	9626	9631	9651	344	183	179	−2,729	−1,46	−1,537	71,771	6,73	6,957
40	—	9839	9856	—	9682	9684	9715	229	147	141	−3,172	−1,34	−1,488	136,479	6,17	6,688
50	—	9855	9881	—	9714	9718	9765	154	126	117	−3,324	−1,20	−1,537	212,428	6,05	7,238
70	—	—	9909	—	—	9755	9818	—	99	91	—	−1,02	−1,809	—	5,64	9,104
100	—	—	9933	—	—	9786	9866	—	78	64	—	−0,73	−1,584	—	3,86	7,411
200	—	—	9963	—	—	9821	9925	—	54	34	—	−0,44	−1,569	—	3,60	7,663
300	—	—	9974	—	—	9835	9949	—	44	23	—	−0,25	−1,440	—	3,18	6,528
500	—	—	9984	—	—	9850	9967	—	33	14	—	−0,10	−1,386	—	3,00	6,249
750	—	—	9989	—	—	9860	9978	—	26	9	—	−0,02	−1,509	—	3,06	7,169
1000	—	—	9991	—	—	9866	9983	—	22	7	—	0,05	−1,301	—	3,16	5,549
1500	—	—	9994	—	—	9876	9988	—	18	5	—	0,18	−1,316	—	3,09	6,094
2000	—	—	9995	—	—	9884	9991	—	15	4	—	0,24	−1,301	—	3,16	5,736

Tabulka 4.3: Některé momenty μ_i testové statistiky W a jejich Shapiro-Wilkovy $\hat{\mu}_i$, Roystonovy $\tilde{\mu}_i$ a mé vlastní μ^*_i aproximace.

4.4 Aproximace pro velká n

Pro každou délku výběru n je třeba nalézt příslušné koeficienty a_i, rozdělení testové statistiky W a empirické kvantily. Díky předchozí aproximaci testové statistiky známe její rozdělení a kvantily pro $n \leq 50$.

J. P. Royston představil nový způsob získání odhadů potřebných k vyčíslení testové statistiky založený na množství aproximací (Royston, 1982a). To umožnilo rozšíření aproximace rozdělení W i pro $7 \leq n \leq 2000$. Byl zde popsán i výpočet hladiny významnosti pro $n < 7$.

Royston využil Shapirovy a Wilkovy aproximace koeficientů

$$\hat{a_i}^* = 2m_i, \quad i = 2,3,\ldots,n-1,$$

$$\hat{a_1}^2 = \hat{a_n}^2 = \begin{cases} g(n-1), & n \leq 20, \\ g(n), & n > 20, \end{cases}$$

kde

$$g(n) = \frac{\Gamma\left[\frac{1}{2}(n+1)\right]}{\sqrt{2}\Gamma\left(\frac{1}{2}n+1\right)}.$$

Funkci $g(n)$ však dále aproximoval a zjednodušil.

Aproximace.

$$g(n) = \left[\frac{6n+7}{6n+13}\right]\left(\frac{\exp(1)}{n+2}\left[\frac{n+1}{n+2}\right]^{n-2}\right)^{1/2}.$$

Aproximace byly použity pro všechna n, a to i pro rozsahy výběrů, u nichž známe hodnoty koeficientů a_i přesně. Tyto přesné hodnoty jsou zohledňovány pouze pro velmi krátké výběry s $n < 7$.

Další aproximací, která byla při sestavování nového testování použita, byl odhad středních hodnot m_i. Počítání přesných hodnot zahrnuje numerické integrace a součty řad. Aproximace pro celý rozsah $2 \leq n \leq 2000$ je jednodušší a sjednocuje způsob počítání pro kratší i rozsáhlejší výběry.

S použitím předchozích odhadů byla vypočítána testová statistika W.

Aproximace. *Pro určení rozdělení testové statistiky byla vygenerována pseudonáhodná čísla a byly použity výše uvedené aproximace. Pro každé uvažované n tak bylo získáno 6000 hodnot W. Rozsahy výběrů, kterých se simulace týkaly byly: všechna n v rozsahu 7-30, n dělitelná pěti pro interval 30-100 a dále hodnoty 125, 150, 200, 300, 400, 500, 600, 750, 1000, 1250, 1500, 2000. Přitom vždy byla vygenerována jiná sada pseudonáhodných čísel, aby se předešlo možným závislostem mezi výsledky.*

Porovnání. Roystonovy empirické první a druhé momenty $\tilde{\mu}_i$ vyšly ze srovnání s přesnými μ_i a s Shapirovými a Wilkovými hodnotami $\hat{\mu}_i$ velmi dobře. Odhady šikmosti a špičatosti se však velmi výrazně lišily (viz tabulka 4.3). Neuvažujeme-li nejkratší výběry, šikmost a špičatost se dle nové metody odhadu mírně zmenšují až do $n = 750$. Podle Shapira a Wilka se však stále zvětšují, u špičatosti dokonce velmi prudce. Empirické kumulativní rozdělení potvrzuje, že šikmost W se s rostoucím n nejprve skutečně zmenšuje, což vede k doměnce, že odhady Shapira a Wilka nebyly správné; mohly být ovlivněny např. zaokrouhlovacími chybami. Z výsledků Roystonovy aproximace vidíme, že testová statistika W by mohla mít normální rozdělení okolo $n = 750$.

4.5 Hodnoty testové statistiky pro malá n

J. P. Royston využil pro výpočet koeficientů a_i spoustu aproximací. Jejich vliv na výsledky pro rozsáhlé výběry je zanedbatelný. Ovšem pro malá n je jejich použití už méně vhodné, což vyplynulo z porovnání odhadů prvního momentu s jeho skutečnými hodnotami.

Proto pro $3 < n < 7$ využijeme některé hodnoty Shapira a Wilka, kteří vypočítali

$$u_n = \log\left[\frac{(W - \varepsilon_n)}{(1 - W)}\right], \tag{4.1}$$

kde

$$\varepsilon_n = a_1^2 n/(n-1),$$

což je nejmenší možná hodnota W (viz věta 9) pro rozsah výběru n. Odvodili také hustotu W pro $n = 3$,

$$f(W) = \frac{3}{\pi}(1 - W)^{-\frac{1}{2}} W^{-\frac{1}{2}}, \quad \frac{3}{4} \leq W \leq 1,$$

a tedy distribuční funkci

$$p = \frac{6}{\pi}\left[\sin^{-1}\sqrt{W} - \sin^{-1}\sqrt{\frac{3}{4}}\right] = \frac{6}{\pi}\left[\sin^{-1}\sqrt{\frac{\exp(u) + \frac{3}{4}}{1 + \exp(u)}} - \sin^{-1}\sqrt{\frac{3}{4}}\right],$$

kde jsme v poslední rovnosti využili vztah (4.1) pro $n = 3$. Pomocí vhodných funkcí s tabulovanými koeficienty od Shapira a Wilka lze u_n převést na u_3. Tím dokážeme též vypočítat p-hodnoty pro všechna $n = 3, 4, 5, 6$ postupem

$$W \to u_n \to u_3 \to p.$$

4.6 Vlastní aproximace rozdělení W

Jako součást práce jsem se sama pokusila aproximovat rozdělení testové statistiky W, resp. hodnoty jejích momentů, a porovnala je s výsledky Shapirových-Wilkových a Roystonových odhadů (viz Shapiro a Wilk, 1965; Royston, 1982a).

K aproximaci jsem použila matematický software R. Zde jsem vygenerovala vždy 5000 pseudonáhodných výběrů pro různé rozsahy n v hodnotách 3, 4,. . . , 10, 15, 20, 30, 40, 50, 70, 100, 200, 300, 500, 750, 1000, 1500, 2000. K výpočtu testové statistiky jsem využila zabudovanou funkci, která je v tomto programu založena na algorimtu J. P. Roystona (viz Royston, 1995). Získala jsem tak hodnoty testové statistiky W a z nich odvodila její momenty.

Tyto aproximace μ_i^* se pro $n \leq 50$ ve většině případů příliš neliší od aproximací $\hat{\mu}_i$ Shapira a Wilka nebo od Roystonových odhadů $\tilde{\mu}_i$ (viz tabulka 4.3). V některých případech se odhady dokonce přibližně rovnají, výjimečně jsou i blíže skutečným hodnotám μ_i. Pokud se odhady Shapira a Wilka a Roystona rozcházejí (např. u špičatosti), mé aproximace sledují spíše Roystona.

Pro rozsáhlejší výběry s $n > 50$ se mé odhady někdy výrazně liší, s rostoucím n stále více. U střední hodnoty rostou rychleji, u rozptylu naopak rychleji klesají. Mé odhady pro špičatost stále zůstavají daleko v záporných hodnotách, zatímco Roystonovy aproximace rostou i do kladných čísel. U špičatosti mé hodnoty kolísají naopak nad hodnotami Roystonovými.

Kapitola 5

Praktická ilustrace

5.1 Příklad

Na tomto místě ukážeme postup testování. K dosažení výsledku testu potřebujeme tabulku koeficientů a_i Shapirova-Wilkova testu a též tabulku kvantilů pro danou délku testovaného výběru n.

Příklad. Ordinaci lékaře navštívilo během jednoho dne 15 žen. Odpovídá rozdělení jejich výšky normálnímu rozdělení? Pozorované hodnoty jsou po řadě

$$x_1 = 157, \quad x_2 = 153, \quad x_3 = 163, \quad x_4 = 169, \quad x_5 = 164,$$
$$x_6 = 175, \quad x_7 = 178, \quad x_8 = 171, \quad x_9 = 162, \quad x_{10} = 159,$$
$$x_{11} = 157, \quad x_{12} = 167, \quad x_{13} = 150, \quad x_{14} = 164, \quad x_{15} = 168.$$

Použijeme Shapirův-Wilkův test.

1. Uspořádáme pozorování a získáme

$$y_1 = 150, \quad y_2 = 153, \quad y_3 = 157, \quad y_4 = 157, \quad y_5 = 159,$$
$$y_6 = 175, \quad y_7 = 178, \quad y_8 = 171, \quad y_9 = 162, \quad y_{10} = 159,$$
$$y_{11} = 157, \quad y_{12} = 167, \quad y_{13} = 150, \quad y_{14} = 164, \quad y_{15} = 168.$$

2. $$S^2 = \sum_{i=1}^{15}(y_i - \bar{y})^2 = \sum_{i=1}^{15} y_i^2 - \frac{1}{15}\left(\sum_{i=1}^{15} y_i\right)^2 = 403317 - 402456{,}6 = 860{,}4.$$

3. Pro $n = 15$ v tabulkách najdeme koeficienty

$$a_{15} = 0{,}5150, \quad a_{14} = 0{,}3306, \quad a_{13} = 0{,}2495, \quad a_{12} = 0{,}1878,$$
$$a_{11} = 0{,}1353, \quad a_{10} = 0{,}0880, \quad a_9 = 0{,}0433, \quad a_8 = 0{,}0000.$$

   Z tvrzení 6 víme, že $a_i = -a_{n-i+1}$.

$$b = \sum_{i=1}^{15}(a_i y_i) = \sum_{i=1}^{7} a_{n-i+1}(y_{n-i+1} - y_i) =$$
$$= 0{,}5150(178 - 150) + 0{,}3306(175 - 153) + 0{,}2495(171 - 157) +$$
$$+0{,}1878(169 - 157) + 0{,}1353(168 - 159) + 0{,}0880(167 - 162) +$$
$$+0{,}0433(164 - 163) = 29{,}1408.$$

4.

$$W = \frac{b^2}{S^2} = \frac{29{,}1408^2}{860{,}4} = 0{,}9869668.$$

5. Pro $n = 15$ nalezneme kvantily p_α pro různé hladiny α

$$p_{0,01} = 0{,}835, \quad p_{0,02} = 0{,}855, \quad p_{0,05} = 0{,}881,$$
$$p_{0,1} = 0{,}901, \quad p_{0,5} = 0{,}950, \quad p_{0,9} = 0{,}975,$$
$$p_{0,95} = 0{,}980, \quad p_{0,98} = 0{,}984, \quad p_{0,99} = 0{,}987.$$

Porovnáme-li hodnotu testové statistiky s hodnotami nalezenými v tabulce, vidíme, že výsledek testu se nachází výrazně nad hladinou 0,05. Hypotézu o normalitě dat tedy nezamítáme.

5.2 Implementace v počítačových programech

V části 4.4 jsme představili aproximace J. P. Roystona publikované v roce 1982 (Royston, 1982a), které usnadnily výpočet a vzhledem k jejich kvalitě neovlivnily výsledky testování. J. P. Royston ve stejném roce sepsal také kód programu, který popsané aproximace využívá (Royston, 1982b). Tento program byl schopný testovat data o rozsahu $3 \leq n \leq 2000$.

V roce 1995 byl program upraven o některé podmínky a rozšířen pro výběry délky $3 \leq n \leq 5000$ (Royston, 1995). Dnes je základem pro spoustu Shapirových-Wilkových testů implementovaných v nejrůznějších počítačových matematických a statistických programech.

Například v dokumentaci programu R (R Core Team, 2014) můžeme nalézt: „Použitý algoritmus je překlad pro C kódu popsaného v článku Royston (1995). Výpočet p-hodnoty je přesný pro $n = 3$, jinak jsou použity aproximace, zvlášť pro $4 \leq n \leq 11$ a $n \geq 12$.“

Z dokumentace programu Matlab (www.mathworks.com) lze též vyčíst, že implementace Shapirova-Wilkova testu je založena na tomtéž článku. Jiným příkladem je program Statistica (www.statsoft.com), který využívá pouze první verzi Roystonova programu z roku 1982, a tedy je schopný, narozdíl od předchozích dvou zmiňovaných programů, testovat pouze data o rozsahu menším než 2000.

Závěr

V práci jsme se věnovali Shapirovu-Wilkovu statistickému testu, který ověřuje normalitu dat. Test je vhodný k ověření předpokladu o datech, jímž právě normalita často bývá. Odvodili jsme tvar testové statistiky W a některé její důležité vlastnosti, jakými jsou například její minimální a maximální hodnota.

K výpočtu testové statistiky jsou pro každý rozsah náhodného výběru potřeba jiné hodnoty činitelů v jejím čitateli. Shapiro a Wilk k nalezení hodnoty W pro výběry do rozsahu 50 využili aproximace těchto koeficientů (viz Shapiro a Wilk, 1965). Dále také určili empirické kvantily.

Jejich aproximace později upravil pro výběry do rozsahu 2000 a později i dále rozšířil J. P. Royston (viz Royston, 1982a). Výpočet testové statistiky též naprogramoval (viz Royston, 1982b, 1995). Věnoval se i rozdělení testové statistiky a určení empirických kvantilů. V současnosti se k výpočtu testové statistiky nejčastěji využívají odhady J. P. Roystona, jak jsme ukázali na příkladech některých softwarových programů.

Testovou statistiku W lze upravit pro verzi s nekompletními výběry W^*, čímž se Shapiro a Wilk také zabývali. Některé další testy i pro jiná rozdělení jsou též založena na podobném principu jako výše popisovaný test. J. P. Royston dále navrhl využít testovou statistiku k přeměně dat, vyžadujeme-li alespoň přibližnou normalitu.

Proběhlo několik studií hodnotící testy ověřující předpoklad normality. Ukázalo se, že Shapirův-Wilkův test je silný a velmi spolehlivý, dokonce i pro velmi krátké výběry ($n \leq 20$). Je citlivý zvláště k nesymetrickým rozdělením, rozdělením s těžkými chvosty a též k odlehlým pozorováním.

Samotní autoři testové statistiky však předpokládají, že „test, jako nástroj k hodnocení normality dat, bude využíván jako doplněk k vykreslování pravděpodobnosti normality, ne jako náhrada za něj“ (Shapiro a Wilk, 1965, část 6.1).

Literatura

Dupač, V. a Hušková, M. (2009). *Pravděpodobnost a matematická statistika*. 4.dotisk 1.vydání. Karolinum, Praha. ISBN 978-80-246-0009-3.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org/`.

Royston, J. P. (1982a). An Extension of Shapiro and Wilk's W Test for Normality to Large Samples. *Applied Statistics*, **31**(2), 115–124.

Royston, J. P. (1982b). Algorithm AS 181: The W Test for Normality. *Applied Statistics*, **31**(2), 176–180.

Royston, J. P. (1995). Remark AS R94: A Remark on Algorithm AS 181: The W Test for Normality,. *Applied Statistics*, **44**(4), 547–551.

Shapiro, S. S. a Wilk, M. B. (1965). An Analysis of Variance Test for Normality (Complete Samples). *Biometrica*, **52**(3/4), 591–611.

Zvára, K. (2008). *Regrese*. 1.vydání. Matfyzpress, Praha. ISBN 978-80-7378-041-8.

Seznam tabulek