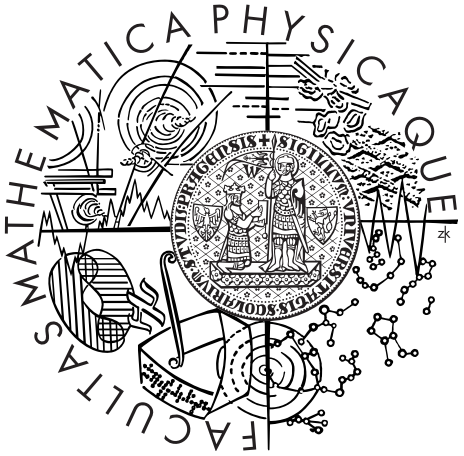
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE



Zdeněk Veselý

Permutační testy statistických hypotéz

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: prof. RNDr. Jana Jurečková DrSc.
Studijní program: Matematika
Studijní obor: MPMSE

Praha 2015

Děkuji paní profesorce Jurečkové za trpělivost.
Děkuji rodině za podporu i za netrpělivost.
Děkuji svému zaměstnavateli za flexibilitu.

Prohlašuji, že jsem tuto diplomovou práci vypracoval 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: Permutační testy statistických hypotéz

Autor: Zdeněk Veselý

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: prof. RNDr. Jana Jurečková DrSc., Katedra pravděpodobnosti a matematické statistiky

Abstrakt:

Tato práce seznamuje čtenáře s konceptem permutačních testů. Permutační test je zde demonstrován jako odpověď na zadání testu, kdy je nežádoucí dělat přísnější předpoklady na rozdělení dat. Pro taková zadání je často permutační test jediným zcela korektním testem. V práci je popsána obecná konstrukce permutačních testů i způsob, jak nalézt nejsilnější takový test oproti specifické alternativě. V druhé části pro vybrané problémy srovnává sílu permutačních testů s testy parametrickými i pořadovými. Toto je provedeno simulacemi. Výsledné síly parametrických a permutačních testů se velmi neliší, což jen potvrzuje užitečnost permutačních testů pro praxi.

Klíčová slova: Permutační testy, Exaktní testy, Testování hypotéz, Síla testu

Title: Permutation Tests of Statistical Hypotheses

Author: Zdeněk Veselý

Department: Department of Probability and Mathematical Statistics

Supervisor: prof. RNDr. Jana Jurečková DrSc., Department of Probability and Mathematical Statistics

Abstract:

This thesis presents permutation tests concept. Permutation test is demonstrated as response to testing problems where it is inconvenient to make any deeper presumptions on data probability distribution. For some of these problems it is even the only exact solution. The construction of permutation test is described in the thesis as well as approach to search of the most powerful tests to specific alternatives. In the second part of the thesis there are comparisons of powers of parametric, permutation and rank test using simulations. The result is that power of parametric and permutation test are very similar most of the times and that confirms that permutation tests are useful tool for praxis.

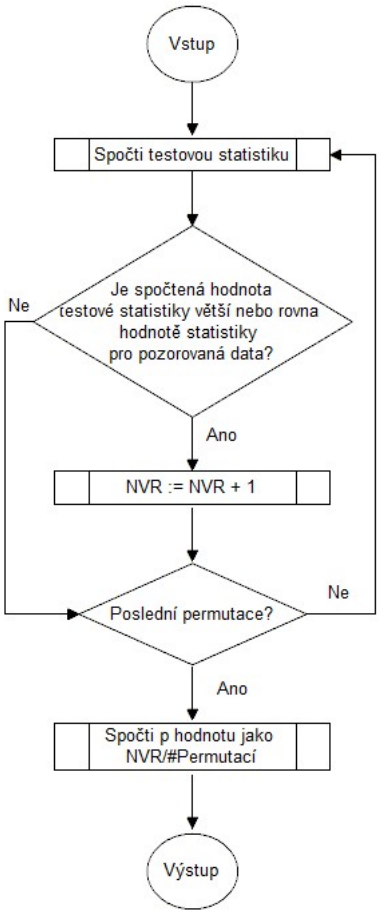
Keywords: Permutation tests, Exact tests, Hypothesis testing, Power of tests

Obsah

Úvod

Tato práce si klade za cíl shrnout základní poznatky o permutačních testech statistických hypotéz. Permutační testy poskytují statistikovi možnost udržet si jistotu, že používá správný nástroj i v případě, že nemůže předpokládat konkrétní rozdělení pozorovaných dat. Tato situace je v praxi velmi běžná, statistik se často musí rozhodovat, zda jsou jeho předpoklady pro parametrický test ještě odůvodnitelné.

Tato práce má ukázat, pro jaké problémy se dá využít permutačních testů. Bude zde také předvedena konstrukce testu (její jednoduchý náznak je v obrázku (1)).



Obrázek 1: Obecný postup vyhodnocení permutačního testu.

Vyhodnocení permutačního testu může být v praxi náročnější na výpočetní prostředky, než je tomu u parametrických testů, neboť je potřeba získat hodnotu testové statistiky pro mnoho permutací dat. Dnes už ale máme velmi silné počítače a tak s výpočetními možnostmi dnešní techniky roste i obliba tohoto druhu testů. V práci bude i představen způsob, jak provést test v případě, že není možné

vyhodnotit všechny permutace dat.

Navíc v případě velkých dat se může statistik uchýlit k využití parametrického testu, který sice nemusí být zcela korektní, ale asymptoticky se bude jeho chyba blížit požadované hodnotě. V případě malého počtu pozorování pak využije raději permutační test, který mu dá jistotu korektnosti.

Druhá část této práce bude zaměřena na porovnání sil parametrických, pořadových a permutačních testů. Silofunkce pořadových a permutačních testů se počítá velmi složitě a tak zde budou všechny silofunkce simulovány. Lze předpokládat, že parametrický test - tam, kde je korektní jej použít - bude dosahovat lepších výsledků. Otázkou ale je, jak velké účinnosti se statistik vzdává, když v daném případě odstoupí od parametrického testu k permutačnímu.

1. Úvod do permutačních testů

1.1 Problém testování hypotéz

Nejprve nastíníme problém testování hypotéz. Mějme $\mathbb{X} = (X_1, \ldots, X_n)$ vektor pozorování nabývající hodnot z prostoru $\mathfrak{X}$ s rozdělením $P_{\boldsymbol{\theta}}$, kde $\boldsymbol{\theta}$ leží v prostoru Θ. Na základě napozorovaných dat chceme rozhodnout zda platí nulová hypotéza

$$H_0 : \boldsymbol{\theta} \in \Theta_0$$

či alternativní hypotéza

$$H_1 : \boldsymbol{\theta} \in \Theta_1 = \Theta \setminus \Theta_0.$$

Rozhodování spočívá v konstrukci rozhodovací funkce

$$\Phi : \mathfrak{X} \to \{0; 1\}.$$

Nabývá-li Φ hodnotu 0, rozhodneme se pro nulovou hypotézu H_0, při hodnotě 1 se naopak rozhodneme pro alternativní hypotézu H_1 (neboli v používané terminologii zamítneme nulovou hypotézu).

Naše rozhodnutí může být správné či nesprávné, můžeme se dopustit dvou druhů chyb:

	Platí H_0	Platí H_1
$\Phi(\mathbf{x}) = 0$	Správné rozhodnutí	Chyba 2. druhu
$\Phi(\mathbf{x}) = 1$	Chyba 1. druhu	Správné rozhodnutí

Konstrukcí rozhodovací funkce Φ bychom rádi minimalizovali pravděpodobnosti obou chyb, to obvykle ale není možné. Zmenšováním pravděpodobnosti chyby jednoho druhu se typicky zvětšuje pravděpodobnost chyby opačného druhu. Testování obvykle nastavujeme tak, aby nulová hypotéza byla obecná a „neškodná“, zatímco alternativní hypotéza byla zajímavá, tu totiž chceme dokázat. Zvolíme si tedy hladinu testu (chybu 1. druhu) $\alpha \in (0; 1)$ a rozhodovací funkci budeme konstruovat tak, aby platilo

$$P_{\boldsymbol{\theta}}\left(\Phi(\mathbb{X}) = 1\right) \leq \alpha, \ \forall \boldsymbol{\theta} \in \Theta_0. \tag{1.1}$$

Nerovnost (1.1) budeme zapisovat ve zjednodušeném tvaru (1.2), přestože formálně výraz na levé straně nerovnosti (1.2) není dobře definován.

$$P\left(\Phi(\mathbb{X}) = 1 | \text{Platí } H_0\right) \leq \alpha. \tag{1.2}$$

Nerovnost (1.1) (resp. (1.2)) zajišťuje, aby pravděpodobnost chyby 1. druhu nepřekročila hladinu α. Nejčastější volby α jsou 0,05 nebo 0,01. Na čím menší hladině pracujeme, tím větší váhu má zamítnutí nulové hypotézy (ale zároveň je těžší ji zamítnout).

Důležitou vlastností testu je jeho nestrannost. Nestrannost zaručuje, že v případě platnosti alternativy H_1 budeme zamítat nejhůře se stejnou pravděpodobností jako v případě platnosti nulové hypotézy H_0.

Definice 1.1.1 (Nestrannost testu)**.** *Test na hladině α je nestranný právě tehdy, pokud platí* (1.2) *(podmínka na regulérnost testu) a zároveň*

$$P\left(\Phi(\mathbb{X}) = 1 | \textit{Platí } H_1\right) \geq \alpha.$$

Klasický přístup k testování hypotéz vede k volbě testové statistiky T, tedy kritéria, jehož hodnota má rozhodnout o přijetí či zamítnutí hypotézy. Testová statistika je funkcí pozorování, jejíž (přesné nebo přibližné) rozdělení za platnosti nulové hypotézy známe. Můžeme tedy nalézt množinu K takovou, že

$$P(T(\mathbb{X}) \in K | \text{Platí } H_0) \leq \alpha \tag{1.3}$$

Rozhodovací funkce pak má tvar

$$\Phi(\mathbf{x}) = \begin{cases} 0, & T(\mathbf{x}) \notin K \\ 1, & T(\mathbf{x}) \in K \end{cases} \tag{1.4}$$

Vhodnou volbou testové statistiky T a množiny K splňujících (1.3) se pak můžeme pokoušet minimalizovat chybu 2. druhu globálně, popřípadě lokálně (tj. jen pro určitá $\boldsymbol{\theta} \in \Theta_1$).

Ne vždy je jednoduché nalézt vhodnou testovou statistiku T takovou, abychom znali její rozdělení. Obvykle takovéto statistiky máme pro relativně malé rodiny rozdělení $\{P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta_0\}$. Ukážeme si to na následujícím příkladě,

Příklad 1.1.1 (Dvouvýběrový t-test)**.** *Nechť $X_1, \ldots, X_{n_1}$ je iid výběr z rozdělení $N(\mu_1, \sigma^2)$ a $Y_1, \ldots, Y_{n_2}$ iid výběr z rozdělení $N(\mu_2, \sigma^2)$, kde $\mu_1, \mu_2 \in \mathbb{R}, \sigma^2 > 0$ jsou neznámé konstanty. Výběry jsou navzájem nezávislé. Chceme testovat hypotézu*

$$H_0 : \mu_1 = \mu_2$$

proti jednostranné alternativě

$$H_1 : \mu_1 > \mu_2$$

V doposud uvedené terminologii test vypadá následovně:
$\mathbb{X} = (X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2})$, $\Theta = \{\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma^2), \mu_1, \mu_2 \in \mathbb{R}, \sigma^2 > 0\}$, $P_{\boldsymbol{\theta}}$ *je distribuční funkcí* $\mathbf{N}_n\left((\mu_1, \ldots, \mu_1, \mu_2, \ldots, \mu_2)^\top, \sigma^2 I_n\right)$, *kde $n = n_1 + n_2$. Jednotlivé hypotézy pak mají tvar*

$$H_0 : \boldsymbol{\theta} \in \Theta_0 = \left\{(\mu_1, \mu_2, \sigma^2), \mu_1, \mu_2 \in \mathbb{R}, \sigma^2 > 0, \mu_1 = \mu_2\right\}$$

$$H_1 : \boldsymbol{\theta} \in \Theta_1 = \left\{(\mu_1, \mu_2, \sigma^2), \mu_1, \mu_2 \in \mathbb{R}, \sigma^2 > 0, \mu_1 > \mu_2\right\}$$

Zvolíme testovou statistiku

$$T\left(\mathbb{X}\right) = \frac{\overline{X} - \overline{Y}}{\sqrt{(n_1 - 1)S_x^2 + (n_2 - 1)S_y^2}} \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}}, \tag{1.5}$$

kde

$$\overline{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i,$$
$$\overline{Y} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i,$$
$$S_x^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_i - \overline{X}),$$
$$S_y^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(Y_i - \overline{Y}).$$

Podle [Aff11] má pak statistika T za platnosti nulové hypotézy H_0 t-rozdělení o n_1+n_2-2 stupních volnosti. Nyní můžeme najít množinu K (zmíněnou v (1.3)*):*

$$K = (z, z > t_{n_1+n_2-2,1-\alpha}), \tag{1.6}$$

kde $t_{n_1+n_2-2,1-\alpha}$ je $1-\alpha$ kvantil t-rozdělení o n_1+n_2-2 stupních volnosti, tedy pro náhodnou veličinu Z s t-rozdělením o m stupních volnosti platí pro kvantil $P(Z \leq t_{m,p}) = p$.

Podle (1.4)*,* (1.5) *a* (1.6) *vidíme, že rozhodovací funkce pro tento příklad má tvar:*

$$\Phi(\mathbf{x}) = \begin{cases} 0, & T(\mathbf{x}) \leq t_{n_1+n_2-2,1-\alpha} \\ 1, & T(\mathbf{x}) > t_{n_1+n_2-2,1-\alpha} \end{cases}$$

Ukážeme, že tento test má opravdu hladinu α:

$$\begin{aligned} P\left(\Phi(\mathbb{X}) = 1 | \textit{Platí } H_0\right) =& P(T(\mathbb{X}) \in K | \textit{Platí } H_0) \\ =& P(T(\mathbb{X}) > t_{n_1+n_2-2,1-\alpha} | \textit{Platí } H_0) \\ =& \alpha \end{aligned}$$

V příkladě (1.1.1) jsme uvažovali poměrně úzkou rodinu rozdělení $\{P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\}$. Předpokládali jsme normální rozdělení, navíc u obou populací nezávislost a shodný rozptyl.

Uvažujme nyní širší rodiny rozdělení $\{P_\theta, \theta \in \Theta\}$, kde

$$\Theta^0 = \left\{ \theta = f : f(\mathbf{x}) = \prod_{i=1}^{n_1} f_1(x_i) \prod_{i=n_1+1}^{n_1+n_2} f_2(x_i), \right. \\ \left. \forall x \in \mathbb{R}\ f_k(x) \geq 0, \int f_k(x)\,\mathrm{d}x = 1,\ k = 1, 2 \right\} \tag{1.7}$$

a P_θ je rozdělení s hustotou θ. Jde tedy o rodinu absolutně spojitých rozdělení vektoru $\mathbb{X} = (X_1, \ldots, X_{n_1}, X_{n_1+1}, \ldots, X_{n_2})$, jehož prvky jsou nezávislé a $X_1, \ldots, X_{n_1}$ je náhodný výběr z jednoho rozdělení, zatímco $X_{n_1+1}, \ldots, X_{n_2}$ je náhodný výběr z obecně jiného rozdělení.

$$\Theta^1 = \Bigg\{ \theta = f : f(\mathbf{x}) = \prod_{i=1}^{n_1} f_1(x_i) \prod_{i=n_1+1}^{n_1+n_2} f_2(x_i),\ \forall x \in \mathbb{R}\ f_k(x) \geq 0,$$
$$\int f_k(x)\,\mathrm{d}x = 1,\ k = 1, 2,\ f_1(x) = f_2(x + d),\ d \in \mathbb{R} \Bigg\}$$

Tato rodina je zúžení rodiny definované v (1.7), kde rozdělení, ze kterých jsou generovány výběry, se liší jen posunutím o konstantu d.

Rodina nulové hypotézy bude ve tvaru:

$$\Theta_0 = \left\{ \theta = f : f(\mathbf{x}) = \prod_{i=1}^{n_1+n_2} f(x_i),\ f(x) \geq 0\ \forall x \in \mathbb{R},\ \int f(x)\,\mathrm{d}x = 1 \right\} \tag{1.8}$$

Chceme-li nyní testovat hypotézu

$$H_0 :\ f \in \Theta_0 \tag{1.9}$$

proti alternativě

$$H_1 :\ f \in \Theta_1^k = \Theta^k \setminus \Theta_0,$$

pro $k = 0, 1$, nelze již jednoduše použít t-test. Testová statistika definovaná v (1.5) nemá totiž za těchto obecnějších předpokladů t-rozdělení. Zúžíme-li rodinu požadavkem na konečnost několika momentů, pak t-test funguje asymptoticky (tj. skutečná hladina testu se sice nerovná požadovanému α, ale limitně se blíží k α s rostoucím $\min(n_1, n_2)$).

Ale to přináší další problémy. Předpokládáme další skutečnosti o rozdělení dat, o kterých v praxi nemusíme vědět, zda platí. Navíc je test pouze asymptoticky korektní. V praxi se proto musíme ještě zamýšlet nad tím, jestli je náš vzorek dostatečně velký na to, aby již byla hladina dostatečně blízko α. Nemluvě o tom, že pro různá rozdělení fungují často jiné testy lépe (ve smyslu síly - tj. chyby 2. druhu) než t-test.

1.2 Motivace pro permutační testy

Označme Π_n množinu všech permutací čísel $1, \ldots, n$. Pro $\mathbf{X} = (X_1, \ldots, X_n)$ uvažujme statistiku

$$\mathbf{X}^{()} = \left(X^{(1)}, X^{(2)}, \ldots, X^{(n)}\right),$$

kde

$$X^{(1)} \leq X^{(2)} \leq \ldots \leq X^{(n)}$$

a existuje $\pi \in \Pi_n$ takové, že $\mathbf{X} = \pi\left(X^{(1)}, X^{(2)}, \ldots, X^{(n)}\right)$. Tuto statistiku budeme nazývat vektorem pořádkových statistik, jedná se vlastně jen o vektor pozorování seřazený od nejnižší hodnoty po nejvyšší.

Věta 1.2.1. *Nechť* $\mathbf{X}$ *má rozdělení s hustotou* $f(\mathbf{x})$*. Pak má* $\mathbf{X}^{()}$ *rozdělení s hustotou*

$$p\left(x^{(1)}, x^{(2)}, \ldots, x^{(n)}\right) = \sum_{\pi \in \Pi_n} f\left(\pi\left(x^{(1)}, x^{(2)}, \ldots, x^{(n)}\right)\right)$$

Důkaz. Mějme

$$R_i = \sum_{j=1}^{n} I\left(X_i \leq X_j\right)$$

pořadí náhodné veličiny X_i ve vektoru $\mathbf{X}$. Nechť $\mathbf{R} = (R_1, \ldots, R_n)$ a $\mathcal{R}$ je množina všech možných $\mathbf{R}$.

Pak pro $B \in \mathcal{B}_n$ platí

$$\begin{aligned} P\left(\mathbf{X}^{()} \in B\right) &= \sum_{\mathbf{r} \in \mathcal{R}} P\left(\mathbf{X}^{()} \in B, \mathbf{R} = \mathbf{r}\right) \\ &= \sum_{r \in \mathcal{R}} \int_{\substack{\mathbf{x}^{()} \in B \\ \mathbf{R} = \mathbf{r}}} \cdots \int f\left(x_1, \ldots, x_n\right)\, dx_1 \ldots\, dx_n \\ &= \sum_{r \in \mathcal{R}} \int_{B} \cdots \int f\left(x_{r_1}, \ldots, x_{r_n}\right)\, dx_{r_1} \ldots\, dx_{r_n} \\ &= \int_{B} \cdots \int p\left(x_{r_1}, \ldots, x_{r_n}\right)\, dx_{r_1} \ldots\, dx_{r_n} \end{aligned}$$

□

Uvědomíme si, že pokud je výše uvedené rozdělení symetrické podle jednotlivých souřadnic, tj.

$$f(\mathbf{x}) = f\left(\pi\left(\mathbf{x}\right)\right), \forall \mathbf{x}, \forall \pi \in \Pi_n, \tag{1.10}$$

pak pro hustotu rozdělení $\mathbf{X}^{()}$ platí

$$p\left(\mathbf{x}^{()}\right) = n! f\left(\mathbf{x}\right)$$

Zároveň je zřejmé, že pro hustotu rozdělení vektoru $\left(\mathbf{X}, \mathbf{X}^{()}\right)$ platí:

$$f_{\mathbf{X}, \mathbf{X}^{()}}\left(\mathbf{x}, \mathbf{x}^{()}\right) = \begin{cases} 0 & \nexists \pi \in \Pi_n : \pi\left(\mathbf{x}\right) = \mathbf{x}^{()} \\ f\left(\mathbf{x}\right) & \exists \pi \in \Pi_n : \pi\left(\mathbf{x}\right) = \mathbf{x}^{()} \end{cases}$$

Naším cílem je ukázat na tomto místě, že statistika $\mathbf{X}^{()}$ má velmi šikovné vlastnosti. Konkrétně budeme prověřovat postačitelnost a úplnost. Nejprve si však tyto pojmy definujme.

Definice 1.2.1 (Postačující statistika)**.** *Statistika* T *je postačující pro parametr* $\boldsymbol{\theta}$ *právě tehdy, když podmíněné rozdělení vektoru* $\mathbf{X} = (X_1, \ldots, X_n)$ *při daném* T *nezávisí na* $\boldsymbol{\theta}$*.*

Definice 1.2.2 (Úplná statistika)**.** *Mějme náhodnou veličinu X s rozdělením $P_{\boldsymbol{\theta}}$, $\boldsymbol{\theta} \in \Theta$. Mějme statistiku $T(X)$. Řekneme, že statistika T je úplná, právě tehdy, když platí pro každou měřitelnou g*

$$\{E[g(T(X))] = 0,\ \forall \boldsymbol{\theta} \in \Theta\} \Rightarrow \{g(T(X)) = 0 \text{ skoro jistě},\ \forall \boldsymbol{\theta} \in \Theta\}.$$

Věta 1.2.2. *Pro rodinu spojitých rozdělení náhodného vektoru $\mathbf{X}$ symetrických podle souřadnic (viz.* (1.10)*) je $\mathbf{X}^{()}$ postačující statistikou.*

Důkaz.

$$\begin{aligned} f_{\mathbf{X}}\left(\mathbf{x}|\mathbf{X}^{()} = \mathbf{x}^{()}\right) =& \frac{f_{\mathbf{X},\mathbf{X}^{()}}\left(\mathbf{x}, \mathbf{x}^{()}\right)}{p\left(\mathbf{x}^{()}\right)} \\ =& \frac{f_{\mathbf{X},\mathbf{X}^{()}}\left(\mathbf{x}, \mathbf{x}^{()}\right)}{n! f(\mathbf{x})} \\ =& \begin{cases} 0 & \nexists \pi \in \Pi_n : \pi(\mathbf{x}) = \mathbf{x}^{()} \\ \frac{1}{n!} & \exists \pi \in \Pi_n : \pi(\mathbf{x}) = \mathbf{x}^{()} \end{cases} \end{aligned}$$

Platí tedy, že $f_{\mathbf{X}}\left(\mathbf{x}|\mathbf{X}^{()} = \mathbf{x}^{()}\right)$ nezávisí na f a tedy $\mathbf{X}^{()}$ je postačující statistikou. □

Věta 1.2.3. *Mějme rodinu rozdělení s hustotou ve tvaru*

$$f(\mathbf{x}) = c(\boldsymbol{\theta})\, exp\left[\sum_{i=1}^{k} \theta_i T_i(\mathbf{x})\right],$$

kde $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k) \in \Theta$. Nechť zároveň prostor parametrů Θ obsahuje neprázdný k-rozměrný obdélník. Pak je statistika $T(\mathbf{x}) = (T_1(\mathbf{x}), \ldots, T_k(\mathbf{x}))$ úplnou statistikou.

Důkaz. Bez újmy na obecnosti lze uvažovat, že prostor parametrů Θ obsahuje množinu

$$I = \{(\theta_1, \ldots, \theta_k), -a \leq \theta_i \leq a, \forall i = 1, \ldots, k\}.$$

Mějme funkci $g(t)$, pro kterou platí

$$Eg(T) = 0,\ \forall \boldsymbol{\theta} \in I$$

Rozložme funkci g na kladnou a zápornou část g^+ a g^- tak, že platí

$$g(t) = g^+(t) - g^-(t).$$

Pak pro všechny $\boldsymbol{\theta} \in I$ platí

$$\int e^{\sum \theta_i t_i} g^+(t)\, \mathrm{d}\nu(t) = \int e^{\sum \theta_i t_i} g^-(t)\, \mathrm{d}\nu(t) \tag{1.11}$$

a tedy i pro $\boldsymbol{\theta} = \mathbf{0}$

$$\int g^+(t)\, \mathrm{d}\nu(t) = \int g^-(t)\, \mathrm{d}\nu(t).$$

Definujme si pravděpodobnostní míry

$$\mathrm{d}P^{+}(t) = cg^{+}(t)\,\mathrm{d}\nu(t)$$
$$\mathrm{d}P^{-}(t) = cg^{-}(t)\,\mathrm{d}\nu(t),$$

kde c je normující konstanta. Rovnost (1.11) lze pak přepsat

$$\int e^{\sum \theta_i t_i}\mathrm{d}P^{+}(t) = \int e^{\sum \theta_i t_i}\mathrm{d}P^{-}(t) \tag{1.12}$$

Budeme nyní považovat $\boldsymbol{\theta}$ za vektor komplexních čísel. Z věty 9 v kapitole 2 v knize [Leh86] pak plyne, že rovnost (1.12) platí i pro I_c

$$I_c = \{(\theta_1, \ldots, \theta_k), \theta_j = \epsilon_j + i\mu_j, -a \leq \epsilon_j \leq a, \mu_j \in \mathbb{R}\}.$$

Specificky pak tedy platí rovnost

$$\int e^{i\sum \mu_j t_j}\mathrm{d}P^{+}(t) = \int e^{i\sum \mu_j t_j}\mathrm{d}P^{-}(t).$$

Tyto integrály jsou zároveň charakteristickými funkcemi rozdělení P^{+}, resp. P^{-}. Z jejich rovnosti plyne i rovnost samotných rozdělení. Proto tedy i $g^{+} = g^{-}$ a na začátku zvolená funkce g musí být rovna nule, což jsme chtěli dokázat. □

Věta 1.2.4. *Mějme $X_1, \ldots, X_n$ náhodný výběr ze spojitého rozdělení $F \in \mathcal{F}$, kde $\mathcal{F}$ je rodina všech absolutně spojitých rozdělení. Pak statistika $\mathbf{X}^{()}$ je úplnou postačující statistikou pro $\mathcal{F}$.*

Uvedeme zde hlavní kroky důkazu věty (1.2.4). Podrobný důkaz lze nalézt v [Leh86].

Důkaz. Z věty (1.2.2) plyne, že $\mathbf{X}^{()}$ je postačující statistika i pro rodinu $\mathcal{F}$.

Úplnost lze dokázat ve více krocích. Nejdříve odvodíme, že statistika

$$Q(\mathbf{X}) = \mathbf{X}^{()} = \left(\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(n)}\right)$$

je ekvivalentní statistice

$$T(\mathbf{X}) = \left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2, \ldots, \sum_{i=1}^{n} X_i^n\right)$$

ve smyslu, že obě generují stejný obraz:

$$Q^{-1}(Q(\mathbf{x})) = \{\ \pi(\mathbf{x}), \pi \in \Pi_n\}$$
$$T^{-1}(T(\mathbf{x})) = \{\ \pi(\mathbf{x}), \pi \in \Pi_n\}$$

Odtud plyne, že statistika Q je úplná právě tehdy, když je statistika T úplná.

Dalším krokem je důkaz, že T je úplná pro rodinu

$$\mathcal{F}_n = \left\{f : \exists\theta_1, \ldots, \theta_n, f(x) = C(\theta_1, \ldots, \theta_n)\, e^{-x^{2n}+\theta_1 x+\ldots+\theta_n x^n}\right\}.$$

Náhodný výběr z tohoto rozdělení má hustotu

$$f(x_1, \ldots, x_n) = C^n(\theta_1, \ldots, \theta_n)\, e^{-\sum x_j^{2n}+\theta_1\sum x_j+\theta_2\sum x_j^2+\ldots+\theta_n\sum x_j^n}$$

Z věty (1.2.3) plyne, že statistika T je úplná pro tento systém hustot.

Rodina rozdělení $\mathcal{F}_n$ už je natolik hustá, že vlastnost úplnosti se dá rozšířit na rodinu všech spojitých hustot $\mathcal{F}$. Tím pak je dokázáno, že statistika $\mathbf{X}^{()}$ je úplná pro rodinu $\mathcal{F}$. □

V kapitole (1.1) jsme naznačili, v čem spočívá omezení parametrických testů. Nyní si tedy stanovme náš požadavek na test. Budeme chtít nestrannost pro všechna spojitá rozdělení splňující nulovou hypotézu. Navážeme na příklad (1.1.1). Budeme chtít tento příklad rozšířit na rodinu všech spojitých rozdělení (1.8):

$$\Theta_0 = \left\{ \theta = f : f(\mathbf{x}) = \prod_{i=1}^{n_1+n_2} f(x_i),\ f(x) \geq 0\, \forall x \in \mathbb{R},\ \int f(x)\,\mathrm{d}x = 1 \right\}$$

Pro test $\Phi(\mathbf{x})$ musí tedy platit:

$$\int \Phi(\mathbf{x})\, f(\mathbf{x})\,\mathrm{d}\mathbf{x} = \alpha, \tag{1.13}$$

pro všechny hustoty $f \in \Theta_0$.

Věta 1.2.5. *Rovnost* (1.13) *je splněna pro všechny* $f \in \Theta_0$*, právě tehdy, když platí*

$$\frac{1}{n!} \sum_{\pi \in \Pi_n} \Phi(\pi(\mathbf{x})) = \alpha.$$

Důkaz. Z věty (1.2.4) víme, že statistika $\mathbf{X}^{()}$ je úplná a postačující pro rodinu rozdělení Θ_0. Proto rovnost (1.13) je plněna právě tehdy, když platí:

$$E\left[\Phi(\mathbf{X}) \,|\mathbf{X}^{()}\right] = \alpha$$

pro všechny hustoty. Jelikož

$$f_{\mathbf{X}}\left(\mathbf{x}|\mathbf{X}^{()} = \mathbf{x}^{()}\right) = \begin{cases} 0 & \nexists \pi \in \Pi_n : \pi(\mathbf{x}) = \mathbf{x}^{()} \\ \frac{1}{n!} & \exists \pi \in \Pi_n : \pi(\mathbf{x}) = \mathbf{x}^{()}, \end{cases}$$

pak

$$E\left[\Phi(\mathbf{X}) \,|\mathbf{X}^{()}\right] = \frac{1}{n!} \sum_{\pi \in \Pi_n} \Phi(\pi(\mathbf{x})).$$

A tedy

$$\frac{1}{n!} \sum_{\pi \in \Pi_n} \Phi(\pi(\mathbf{x})) = \alpha.$$

□

Věta (1.2.5) říká, jaký tvar musí mít test, aby byl nestranný pro tak velkou rodinu rozdělení. V kapitole (1.3) si zavedeme jiné značení, ale stále se budeme zabývat testy v tomto tvaru, tedy tzv. permutačními testy.

Nyní si zpřísníme náš požadavek na test. Budeme chtít nejsilnější nestranný test pro obecnou nulovou hypotézu a danou alternativní hypotézu. Řekněme, že za platnosti H_1 má $\mathbf{X}$ rozdělení s hustotou $g(\mathbf{x})$. Samozřejmě $g(\mathbf{x}) \notin \Theta_0$.

Síla testu je pak

$$\int \Phi(\mathbf{x})\, g(\mathbf{x})\, \mathrm{d}\mathbf{x} = \int E\left[\mathbf{X}|\mathbf{X}^{()} = \mathbf{x}^{()}\right] \mathrm{d}P_{\mathbf{X}^{()}}\left(\mathbf{x}^{()}\right)$$

Jelikož

$$E\left[\mathbf{X}|\mathbf{X}^{()} = \mathbf{x}^{()}\right] = \frac{\sum_{\pi\in\Pi_n} \Phi(\pi(\mathbf{x}))\, g(\pi(\mathbf{x}))}{\sum_{\pi\in\Pi_n} g(\pi(\mathbf{x}))}$$

tak se úloha hledání nejsilnějšího nestranného testu redukuje na hledání rozhodovací funkce $\Phi(\mathbf{x})$, která splňuje

$$\frac{1}{n!}\sum_{\pi\in\Pi_n} \Phi(\pi(\mathbf{x})) = \alpha$$

a maximalizuje

$$\frac{\sum_{\pi\in\Pi_n} \Phi(\pi(\mathbf{x}))\, g(\pi(\mathbf{x}))}{\sum_{\pi\in\Pi_n} g(\pi(\mathbf{x}))}$$

pro všechna $\mathbf{x}^{()}$.

Z Neyman–Pearsonova lemmatu (viz [Leh86]) víme, že maximální síly dosáhneme takovou rozhodovací funkcí, která bude H_0 zamítat v případě vysoké hodnoty výrazu

$$\frac{n! g(\mathbf{x})}{\sum_{\pi\in\Pi_n} g(\pi(\mathbf{x}))}$$

Rozhodovací funkci lze tedy napsat ve tvaru

$$\Phi(\mathbf{x}) = \begin{cases} 1 & g(\mathbf{x}) > C\left(\mathbf{x}^{()}\right) \\ \gamma & g(\mathbf{x}) = C\left(\mathbf{x}^{()}\right) \\ 0 & g(\mathbf{x}) < C\left(\mathbf{x}^{()}\right), \end{cases}$$

kde $\gamma \in (0;1)$.

Prakticky to tedy vypadá tak, že se vyhodnotí hodnota funkce g pro všechny permutace pozorování $\mathbf{x}$. Zamítat nulovou hypotézu H_0 budeme tehdy, pokud hodnota $g(\mathbf{x})$ bude mezi k nejvyššími nebo s pravděpodobností γ tehdy, pokud bude $k+1$ nejvyšší. Pro k a γ platí:

$$\begin{aligned} k &\in \mathbf{N} \\ \gamma &\in (0;1) \\ k + \gamma &= \alpha \cdot n! \end{aligned} \tag{1.14}$$

1.3 Konstrukce permutačního testu

Definice 1.3.1 (Orbita)**.** *Mějme $\mathbf{x} \in \mathfrak{X}$ a rodinu rozdělení $P_{\boldsymbol{\theta}}$, $\boldsymbol{\theta} \in \Theta$.*

Orbita příslušící $\mathbf{x}$ jsou takové prvky $x^* \in \mathfrak{X}$, které splňují

$$\frac{f_{P_{\boldsymbol{\theta}}}(\mathbf{x}^*)}{f_{P_{\boldsymbol{\theta}}}(\mathbf{x})} = 1,\ \forall \boldsymbol{\theta} \in \Theta.$$

Značení: $\mathfrak{X}_{|\mathbf{x}} = \left\{\mathbf{x}^*,\ \frac{f_{P_{\boldsymbol{\theta}}}(\mathbf{x}^*)}{f_{P_{\boldsymbol{\theta}}}(\mathbf{x})} = 1,\ \forall \boldsymbol{\theta} \in \Theta\right\}$

Na orbitu můžeme nahlížet jako na jistou třídu ekvivalence, neboť prvky nacházející se ve stejné orbitě podávají o rozdělení stejnou informaci. Uvažujeme-li rodinu rozdělení Θ_0^0 resp. Θ_0^1 (viz. (1.9), $\mathbf{x}$ vzniká realizací iid veličin), pak

$$\mathfrak{X}_{|\mathbf{x}} = \{\mathbf{x}^*, \mathbf{x}^* = \pi(\mathbf{x}), \pi \in \Pi_n\}. \tag{1.15}$$

Jelikož jsou pozorování iid, permutací se poměr věrohodností nezmění. Zároveň rodina rozdělení je natolik široká, že pro každé $\mathbf{x}^{**} \neq \pi(\mathbf{x})$, $\forall \pi \in \Pi_n$ najdeme $\boldsymbol{\theta} \in \Theta_0^k$ takovou, aby $\frac{f_{P_{\boldsymbol{\theta}}}(\mathbf{x}^{**})}{f_{P_{\boldsymbol{\theta}}}(\mathbf{x})} \neq 1$.

Nyní budeme touto orbitou podmiňovat. Za platnosti H_0 (1.9) z Bayesovy věty a definice orbity (1.3.1) plyne:

$$P_{\boldsymbol{\theta}}\left\{\mathbf{x}^* = \mathbf{x}' | \mathbf{x}^* \in \mathfrak{X}_{|\mathbf{x}}\right\} = \frac{\#\left[\mathbf{x}^* = \mathbf{x}', \mathbf{x}^* \in \mathfrak{X}_{|\mathbf{x}}\right]}{\#\left[\mathbf{x}^* \in \mathfrak{X}_{|\mathbf{x}}\right]}.$$

Symbolem $\#[A]$ označujeme počet prvků s vlastností A.

Jak vidíme, tento výraz nezávisí na $\boldsymbol{\theta} \in \Theta_0^k$. Počet prvků v orbitě $\#\left[\mathbf{x}^* \in \mathfrak{X}_{|\mathbf{x}}\right]$ je konečný. Pokud $\mathbf{x}$ neobsahuje shody (tj. $x_i \neq x_j \Leftrightarrow i \neq j$), je tento počet pro náš případ $n!$. Pravděpodobnost shody je pro tento příklad nulová (veličiny iid ze spojitého rozdělení).

Máme-li statistiku $T : \mathfrak{X} \to \mathbb{R}$, pak pro ni obdobně platí:

$$P_{\boldsymbol{\theta}}\left\{T(\mathbf{x}^*) = t | \mathbf{x}^* \in \mathfrak{X}_{|\mathbf{x}}\right\} = \frac{\#\left[\mathbf{x}^* : T(\mathbf{x}^*) = t, \mathbf{x}^* \in \mathfrak{X}_{|\mathbf{x}}\right]}{\#\left[\mathbf{x}^* \in \mathfrak{X}_{|\mathbf{x}}\right]}, \tag{1.16}$$

opět nezávisle na $\boldsymbol{\theta} \in \Theta_0^k$.

Klasické testy jsou postaveny na znalosti (alespoň přibližného) rozdělení testové statistiky T za platnosti nulové hypotézy. Oproti tomuto přístupu stojí permutační testy, které jsou založeny na znalosti rozdělení podmíněné testové statistiky $T|\mathfrak{X}_{|\mathbf{x}}$. Toto rozdělení je diskrétní, jelikož je zde jen konečně mnoho bodů, které mají kladnou pravděpodobnost.

Definice 1.3.2. *Statistiky T_1 a $T_2 : \mathfrak{X} \to \mathbb{R}$ jsou permutačně ekvivalentní, pokud pro každé $\mathbf{x} \in \mathfrak{X}$ a $\mathbf{x}^* \in \mathfrak{X}_{|\mathbf{x}}$ platí $\{T_1(\mathbf{x}) \leq T_1(\mathbf{x}^*)\} \Leftrightarrow \{T_2(\mathbf{x}) \leq T_2(\mathbf{x}^*)\}$. Budeme značit $T_1 \approx T_2$.*

Možná lepší název by mohl být *ekvivalentní vůči orbitám*, protože orbity nejsou vždy jen všechny permutace. Ale když se celá třída testů jmenuje permutační testy, tak budeme pro vztah statistik používat pojem permutačně ekvivalentní.

Uvědomme si, že relace $\approx$ je reflexivní, symetrická a tranzitivní. Určitě také platí, že je-li $T_1 = \varphi(T_2)$, kde φ je ryze rostoucí, pak $T_1 \approx T_2$.

Definici (1.3.2) bychom mohli rozšířit o statistiky, pro které platí

$$\{T_1(\mathbf{x}) \leq T_1(\mathbf{x}^*)\} \Leftrightarrow \{T_2(\mathbf{x}) \geq T_2(\mathbf{x}^*)\},$$

pak bychom mohli uvažovat také φ ryze klesající.

Rozhodovací funkce permutačního testu má tedy tvar

$$\Phi(\mathbf{x}) = \begin{cases} 0, & T(\mathbf{x}) \leq T_\alpha(\mathbf{x}) \\ 1, & T(\mathbf{x}) > T_\alpha(\mathbf{x}) \end{cases},$$

kde $T_\alpha(\mathbf{x})$ je α-kvantil rozdělení $T|\mathfrak{X}_{|\mathbf{x}}$, tedy

$$\begin{aligned} T_\alpha(\mathbf{x}) &= \inf\left\{t : P\left[T(\mathbf{x}^*) \geq t | \mathbf{x}^* \in \mathfrak{X}_{|\mathbf{x}}\right] \leq \alpha\right\} \\ &= \inf\left\{t : \frac{\#\left\{T(\mathbf{x}^*) \geq t,\ \mathbf{x}^* \in \mathfrak{X}_{|\mathbf{x}}\right\}}{\#\left\{\mathbf{x}^* \in \mathfrak{X}_{|\mathbf{x}}\right\}} \leq \alpha\right\}. \end{aligned}$$

$\mathfrak{X}_{|\mathbf{x}}$ zde přísluší rodině rozdělení při nulové hypotéze.

Ukážeme si, že tento test opravdu má hladinu α. Nejprve si uvědomíme, že za platnosti nulové hypotézy platí pro podmíněnou střední hodnotu

$$\begin{aligned} E\left[\Phi(\mathbf{X}^*)|\mathbf{X}^* \in \mathfrak{X}_{|\mathbf{x}}\right] &= P\left[\Phi(\mathbf{X}^*) = 1 | \mathbf{X}^* \in \mathfrak{X}_{|\mathbf{x}}\right] \\ &= P\left[T(\mathbf{X}^*) > T_\alpha(\mathbf{X}) | \mathbf{X}^* \in \mathfrak{X}_{|\mathbf{x}}\right] \leq \alpha. \end{aligned}$$

Pro nepodmíněnou pak

$$E\left[\Phi(\mathbf{X})\right] = E\left[E\left[\Phi(\mathbf{X}^*)|\mathbf{X}^* \in \mathfrak{X}_{|\mathbf{x}}\right]\right] \leq \alpha.$$

Dořešme nyní příklad (1.1.1) pro obecnější zadání. Uvažujme nejprve jednostranně

$$\begin{aligned} H_0 &: \theta \in \Theta_0 \\ H_1 &: \theta \in \Theta_1, \end{aligned}$$

kde Θ_0 je jako v (1.8), zatímco pro Θ_1 platí:

$$\begin{aligned} \Theta_1 = \Bigg\{ \theta = f : f(\mathbf{x}) &= \prod_{i=1}^{n_1} f_1(x_i) \prod_{i=n_1+1}^{n_1+n_2} f_1(x_i - \delta), \\ \delta > 0, f_1(x) &= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \mu \in \mathbf{R}, \sigma > 0 \Bigg\} \end{aligned}$$

Za nulové hypotézy je tedy $\mathbf{X} = (X_1, \ldots, X_{n_1}, X_{n_1+1}, \ldots, X_{n_1+n_2})$ iid výběr z libovolného spojitého rozdělení. Za alternativní rozdělení je výběr $X_1, \ldots, X_{n_1}$ iid z $N(\mu, \sigma^2)$ a výběr $X_{n_1+1}, \ldots, X_{n_1+n_2}$ iid z posunutého $N(\mu - \delta, \sigma^2)$. Navzájem jsou nezávislé.

Hustota rozdělení vektoru $\mathbf{X}$ za alternativní hypotézy má tvar

$$\begin{aligned}
f(\mathbf{x}) =& \frac{1}{\sigma^n (2\pi)^{n/2}} exp\left\{ -\frac{1}{2\sigma} \sum_{i=1}^{n_1} (x_i - \mu)^2 - \frac{1}{2\sigma} \sum_{i=n_1+1}^{n} (x_i - \mu + \delta)^2 \right\} \\
=& \frac{1}{\sigma^n (2\pi)^{n/2}} exp\left\{ -\frac{1}{2\sigma} \sum_{i=1}^{n_1} (x_i - \mu)^2 \right. \\
& \left. -\frac{1}{2\sigma} \sum_{i=n_1+1}^{n} \left((x_i - \mu)^2 + 2(x_i - \mu)\delta + \delta^2 \right) \right\} \\
=& \frac{1}{\sigma^n (2\pi)^{n/2}} exp\left\{ -\frac{1}{2\sigma} \sum_{i=1}^{n} (x_i - \mu)^2 - \frac{1}{2\sigma} \sum_{i=n_1+1}^{n} \left(2(x_i - \mu)\delta + \delta^2 \right) \right\} \\
=& \frac{1}{\sigma^n (2\pi)^{n/2}} exp\left\{ -\frac{1}{2\sigma} \left(\sum_{i=1}^{n} (x_i - \mu)^2 + \sum_{i=n_1+1}^{n} \left(-2\mu\delta + \delta^2 \right) \right) \right. \\
& \left. -\frac{1}{\sigma} \sum_{i=n_1+1}^{n} x_i \delta \right\} \\
=& \frac{1}{\sigma^n (2\pi)^{n/2}} exp\left\{ -\frac{1}{2\sigma} \left(\sum_{i=1}^{n} (x_i - \mu)^2 + \sum_{i=n_1+1}^{n} \left(-2\mu\delta + \delta^2 \right) \right) \right\} \\
& \cdot exp\left\{ -\frac{1}{\sigma} \delta \sum_{i=n_1+1}^{n} x_i \right\}.
\end{aligned}$$

Až na poslední exponenciálu je celý tento výraz nezávislý na permutaci $\mathbf{x}$, lze tedy říci, že $f(\mathbf{x})$ je permutačně ekvivalentní $exp\left\{ -\frac{1}{\sigma}\delta \sum_{i=n_1+1}^{n} x_i \right\}$. Jelikož je e^x rostoucí funkce a σ i δ kladné parametry, tak je $f(\mathbf{x})$ permutačně ekvivalentní $-\sum_{i=n_1+1}^{n} x_i$ a tím i $\sum_{i=1}^{n_1} x_i$, neboť

$$\sum_{i=1}^{n_1} x_i = \sum_{i=1}^{n} x_i - \sum_{i=n_1+1}^{n} x_i$$

a výraz $\sum_{i=1}^{n} x_i$ je nezávislý na permutaci pozorování. Uvažujme tedy statistiku

$$T(\mathbf{x}) = \sum_{i=1}^{n_1} x_i$$

Nalezneme k, γ odpovídající (1.14). Rozhodovací funkce bude:

$$\Phi(\mathbf{x}) = \begin{cases} 1 & \sum_{\pi \in \Pi_n} I\left(T(\mathbf{x}) \leq T(\pi(\mathbf{x}))\right) < k \\ \gamma & \sum_{\pi \in \Pi_n} I\left(T(\mathbf{x}) \leq T(\pi(\mathbf{x}))\right) = k \\ 0 & \sum_{\pi \in \Pi_n} I\left(T(\mathbf{x}) \leq T(\pi(\mathbf{x}))\right) > k. \end{cases}$$

Zamítat nulovou hypotézu budeme tehdy, když hodnota statistiky T bude vysoká.

Pro oboustranný test (tedy variantu za $H_1 : \delta \neq 0$) bychom došli ke statistice

$$T(\mathbf{x}) = \left| \sum_{i=1}^{n_1} x_i \right|.$$

1.4 Metoda Monte Carlo pro rozdělení testové statistiky

V (1.16) jsme odvodili podmíněné rozdělení testové statistiky T. Abychom získali toto rozdělení v konkrétním případě, musíme vyčíslit T ve všech bodech z $\mathfrak{X}_{|\mathbf{x}}$. Těch je ale často mnoho (např.: $n!$, 2^n, kde n je počet napozorovaných náhodných veličin). Pokud máme málo pozorování, lze tedy vyčíslit rozdělení T přesně a my získáváme přesný test. Je to výhoda oproti parametrickým testům, kde známe jen asymptotické rozdělení statistiky T a nemáme dostatek pozorování, aby nám ta asymptotika zafungovala.

Máme-li tolik pozorování, že není možné kvůli omezenosti výpočetních prostředků vyčíslit hodnoty T ve všech bodech orbity, pak se spokojíme s přibližným rozdělením. V některých případech známe asymptotické rozdělení statistiky. Nejčastěji však použijeme aproximaci Monte Carlo, kde budeme dělat výběr z $\mathfrak{X}_{|\mathbf{x}}$. Lze použít různých typů výběru, například náhodný výběr bez vracení, který vede na bootstrap.

My budeme používat pouze náhodný výběr s vracením. Máme tedy jedno konkrétní pozorování $\mathbf{x}$ náhodného vektoru $\mathbb{X}$ a zvolenou testovou statistiku T. Zvolíme si B velikost náhodného výběru, náhodně vybereme $\{\mathbf{x}_1^*, \ldots, \mathbf{x}_B^*\}$ z $\mathfrak{X}_{|\mathbf{x}}$. Vyčíslíme

$$T_i^* = T(\mathbf{x}_i^*),\ i = 1, \ldots, B.$$

Pak

$$\hat{F}_B(z) = \sum_{i=1}^{B} \mathbf{1}\left(T_i^* \leq z\right)/B$$

je nestranný a konzistentní odhad

$$F_T\left(z|\mathbb{X}\right) = P\left(T(\mathbf{x}^*) \leq z|\mathbf{x}^* \in \mathfrak{X}_{|\mathbf{x}}\right),\ \forall z \in \mathbb{R}.$$

Označme $T_0 = T(\mathbf{x})$, tedy hodnotu testové statistiky pro naše skutečné pozorování. Pak bude platit, že

$$\hat{\lambda} = \sum_{i=1}^{B} \mathbf{1}\left(T_i^* > T_0\right)/B$$

je nestranný a konzistentní odhad p-hodnoty

$$\lambda = P\left(T(\mathbf{x}^*) > T_0|\mathbf{x}^* \in \mathfrak{X}_{|\mathbf{x}}\right).$$

S dnešními výpočetními prostředky již obvykle nebývá problém provést výpočet pro vysokou hodnotu B, přesto se ale podívejme na přesnost takového odhadu. $\{\mathbf{1}\left(T_i^* > T_0\right), i\}$ jsou iid náhodné veličiny s alternativním rozdělením s parametrem λ. Platí tedy

$$\begin{aligned} E\hat{\lambda} =& \lambda \\ var\left(\hat{\lambda}\right) =& \frac{1}{B}\lambda\left(1-\lambda\right) \leq \frac{1}{4B}, \end{aligned}$$

jelikož

$$max_{\lambda\in\langle 0,1\rangle}\left(\lambda\left(1-\lambda\right)\right)=\frac{1}{4}.$$

Užitím centrální limitní věty aproximujeme rozdělení $\hat{\lambda}$ normálním rozdělením $N\left(\lambda,\frac{1}{B}\lambda\left(1-\lambda\right)\right)$. Lze tedy uvažovat 95% interval spolehlivosti odhadu jako

$$\left(\hat{\lambda}-\frac{1}{B}\hat{\lambda}\left(1-\hat{\lambda}\right)u_{0.95},\hat{\lambda}+\frac{1}{B}\hat{\lambda}\left(1-\hat{\lambda}\right)u_{0.95}\right),$$

kde $u_{0.95}\doteq 1.645$ je 95% kvantil normálního normovaného rozdělení. Šířka tohoto intervalu je tedy

$$2\frac{1}{B}\hat{\lambda}\left(1-\hat{\lambda}\right)u_{0.95}\leq\frac{1}{2B}u_{0.95}\doteq 0.822/B$$

Pokud se budeme pohybovat kolem p-hodnoty $\lambda=0.05$, pak se šířka ještě zúží na

$$2\frac{1}{B}0.05\left(1-0.05\right)u_{0.95}=\frac{0.095}{B}u_{0.95}\doteq 0.039/B$$

Pokud bychom interval spolehlivosti konstruovali na hladině 99%, pak je šířka $0.055/B$. Pro účely simulování sil testů budeme zde nejčastěji využívat $B=1000$, při této hodnotě je již nepřesnost Monte Carlo odhadu zanedbatelná.

Rozdělení náhodné veličiny $T(\mathbf{X}^*)|\mathbf{X}^*\in\mathfrak{X}_{|\mathbf{x}}$ je diskrétní, nabývá jen konečně mnoha různých hodnot. V případě menších výběrů může být těchto hodnot pouze velmi málo, takže námi zvolená hladina testu α nemusí být dosažena. Mějme tedy množinu

$$H=\left\{t_1<t_2<\ldots<t_m,\forall x\in\mathfrak{X}_{|\mathbf{x}}\exists i\in\{1,\ldots,m\}:T(x)=t_i\right\}$$

a množinu

$$P_H=\left\{0=f_0<f_1<\ldots<f_m=1,f_0=0,f_i=P(T(\mathbf{X}^*)\leq t_i|\mathbf{X}^*\in\mathfrak{X}_{|\mathbf{x}})\right\}$$

Hodnoty množiny P_H odhadneme metodou Monte Carlo. Je-li $k:t_k=T_0=T(\mathbf{X})$, pak lze pro randomizovaný test brát p-hodnotu $\lambda_R=u,u\sim U(f_{k-1},f_k)$, kde $U(a,b)$ je rovnoměrné rozdělení na intervalu $\langle a,b\rangle$. Takto lze zkonstruovat test pro jakoukoliv hladinu $\alpha\in\langle 0,1\rangle$.

2. Jednovýběrové testy

2.1 Test symetrie

Nyní se podíváme na permutační test symetrie rozdělení. Mějme $\mathbb{X} = (X_1, \ldots, X_n)$ iid výběr z rozdělení, jehož symetrii kolem nuly chceme testovat. Formulujme tedy problém formálně: $\mathbb{X} \sim F_{\boldsymbol{\theta}}$,

$$\Theta_0 = \left\{ \boldsymbol{\theta} = f_n : f_n(\mathbf{x}) = \prod_i f(x_i),\ f(x) = f(-x), \forall x \in \mathfrak{X} \right\}. \tag{2.1}$$

f budeme obvykle značit hustoty, nebudeme proto již psát podmínky

$$f(x) \geq 0, \forall x \in \mathfrak{X},\ \int f(x)\,\mathrm{d}x = 1.$$

Celý prostor parametrů pak může mít tvar

$$\Theta^1 = \left\{ \boldsymbol{\theta} = f_n : f_n(\mathbf{x}) = \prod_i f(x_i),\ \exists d \in \mathbb{R} :\ f(d+x) = f(d-x), \forall x \in \mathfrak{X} \right\}$$

nebo

$$\Theta^2 = \left\{ \boldsymbol{\theta} = f_n : f_n(\mathbf{x}) = \prod_i f(x_i),\ f \text{ je hustota} \right\}.$$

Θ^1 je rodina všech absolutně spojitých symetrických rozdělení, nulová hypotéza by se pak dala zapsat jako

$$H_0 : \boldsymbol{\theta} \in \Theta^1,\ d = 0$$
$$H_1 : \boldsymbol{\theta} \in \Theta^1,\ d \neq 0,$$

testujeme tedy, zda střed symetrie je v nule. Pokud bychom chtěli testovat, zda je střed symetrie v předem daném d_0, pak bychom transformovali data na

$$\mathbb{X}_{d_0} = (X_1 - d_0, \ldots, X_n - d_0)$$

a na tyto data pak vystavěli hypotézu jako výše.

Θ^2 je obecná rodina všech spojitých rozdělení.

Orbita příslušící rodině nulové hypotézy je

$$\mathfrak{X}_{|\mathbf{x}} = \{ \mathbf{x}^* :\ \mathbf{x}^* = \pi(S_1 x_1, \ldots, S_n x_n), \pi \in \Pi_n, S_i \in [-1; 1] \}. \tag{2.2}$$

S_i zde hraje roli znaménka, jistě za platnosti H_0 platí $f(x_i) = f(-x_i) = f(S_i x_i),\ S_i \in [-1; 1],\ \forall i$. Pak tedy

$$\mathbf{x}^* \in \mathfrak{X}_{|\mathbf{x}} :\ \frac{f_n(\mathbf{x}^*)}{f_n(\mathbf{x})} = \frac{\prod_i f(S_{u_i} x_{u_i})}{\prod_i f(x_i)} = \frac{\prod_i f(x_i)}{\prod_i f(x_i)} = 1,$$

kde $(u_1, \ldots, u_n) = \pi(1, \ldots, n)$.

Rodina Θ_0 definovaná v (2.1) je opět natolik bohatá, že pokud $\mathbf{x}^{**}$ není ve tvaru popsaném v (2.2), pak existuje $f_n \in \Theta^0$ taková, že $\frac{f_n(\mathbf{x}^{**})}{f_n(\mathbf{x})} \neq 1$.

Uvědomme si, že i za platnosti alternativní hypotézy je stále rozdělení invariantní vůči permutacím dat. To plyne z toho, že předpokládáme iid výběr.

Za testovou statistiku můžeme volit například

$$\begin{aligned} T_1(\mathbf{x}) &= med(x_1, \ldots, x_n) \\ T_2(\mathbf{x}) &= \sum_i sign(x_i) \\ T_3(\mathbf{x}) &= \sum_i x_i. \end{aligned}$$

Test založený na T_2 je vlastně znaménkový test (viz [Aff11]). Testová statistika T_3 je permutačně ekvivalentní statistice

$$T_3'(\mathbf{x}) = \overline{\mathbf{x}}$$

a také statistice

$$T_3''(\mathbf{x}) = \frac{\sum_i x_i}{\sqrt{\sum_i x_i^2}}.$$

Za nulové hypotézy a dalších předpokladů (konečný rozptyl $var\,(X_i)$, nízký poměr $\sum X_i^4 / \left(\sum X_i^2\right)^2$) má $T_3''\,(\mathbf{X}^*)\,|\mathbf{X}^* \in \mathfrak{X}_{|\mathbf{x}}$ asymptoticky rozdělení $N(0,1)$ (viz [Pes01]).

Uvědomme si, že $\mathfrak{X}_{|\mathbf{x}}$ zde obsahuje až $2^n n!$ prvků. Jelikož je ale rozdělení i za platnosti alternativní hypotézy invariantní vůči permutacím, zároveň i výše uvedené testové statistiky jsou takto invariantní, pak se můžeme omezit pouze na prvky tvaru

$$\mathbf{x}^* = (S_1 x_1, \ldots, S_n x_n).$$

Těch je „pouze“ 2^n.

Ukážeme si nestrannost testu

$$\begin{aligned} H_0 &: \boldsymbol{\theta} \in \Theta^1,\ d = 0 \\ H_1 &: \boldsymbol{\theta} \in \Theta^1,\ d > 0 \end{aligned}$$

Máme $\mathbb{X}_0 = \{X_1, \ldots, X_n\}$ náhodný výběr z rozdělení symetrického kolem nuly a dále $\mathbb{X}_d = \{X_1 + d, \ldots, X_n + d\}$. Test založíme na statistice $T(\mathbf{x}) = \sum x_i$. Označme

$$\begin{aligned} T_0 &= T(\mathbb{X}_0), \\ T_d &= T(\mathbb{X}_d) = \sum (X_i + d) = T_0 + nd, \\ T_0^* &= T(\mathbb{X}_0^*) = \sum S_i X_i, \\ T_d^* &= T(\mathbb{X}_d^*) = \sum \left(S_i \left(X_i + d\right)\right) = T_0^* + d \sum S_i, \end{aligned}$$

kde $\mathbb{X}_0^* \in \mathfrak{X}_{|\mathbb{X}_0}$, $\mathbb{X}_d^* \in \mathfrak{X}_{|\mathbb{X}_d}$.

Nyní si definujme nejnižší hladinu, na které ještě zamítáme nulovou hypotézu (tzv. p-hodnotu) jako

$$\lambda(\mathbf{x}) = \inf\{\alpha : T(\mathbf{x}) > T_\alpha\}.$$

λ budeme podle potřeby brát jako funkci $\mathbf{x}$ popřípadě $T_0\,(= T(\mathbf{x}))$. Nestrannost testu je ekvivalentní vlastnosti $\lambda(\mathbb{X}_d) \leq \lambda(\mathbb{X}_0)$ a její platnost nyní ověříme:

$$\begin{aligned}
\lambda(\mathbb{X}_d) &= P\left(T_d^* > T_d | \mathbb{X}_d^* \in \mathfrak{X}_{|\mathbb{X}_d}\right) \\
&= P\left(T_0^* + d\sum S_i > T_0 + nd | \mathbb{X}_0^* \in \mathfrak{X}_{|\mathbb{X}_0}\right) \\
&= P\left(T_0^* + d\sum (S_i - 1) > T_0 | \mathbb{X}_0^* \in \mathfrak{X}_{|\mathbb{X}_0}\right) \\
&\leq P\left(T_0^* > T_0 | \mathbb{X}_0^* \in \mathfrak{X}_{|\mathbb{X}_0}\right) = \lambda(\mathbb{X}_0). \qquad (2.3)
\end{aligned}$$

Nerovnost (2.3) plyne z toho, že $S_i \in [-1; 1]$, tedy $(S_i - 1) \leq 0$.

Pro ilustraci uveďme srovnání permutačního testu postaveném na testové statistice T_3 s t-testem a Wilcoxonovým testem. Uvažujme $\{z_i, i \in 1, \ldots, n\}$ jsou iid z dvojitého exponenciálního rozdělení (hustota ve tvaru $f(x) = \frac{1}{2}e^{-|x|}$).

Mějme

$$x_i = \delta + z_i$$

Budeme pro $n = 5, 10, 20, 50$ a $\delta = 0, 0.1, 0.25, 0.5, 0.75$ porovnávat sílu jednotlivých testů. Výsledek (viz grafy (2.1), (2.2), (2.3) a (2.4)) ukazuje různé chování různých testů. Wilcoxonův test funguje zde lépe při větším počtu pozorování, ale při velmi malém jej nemá smysl uvažovat.[1] T-test se zdá být příliš konzervativní, při $n = 5, \delta = 0$, tedy za platnosti nulové hypotézy, je jeho síla pouze 0.0319, tedy hodně pod hladinou testu $\alpha = 0.05$.

Permutační test dopadl nejstabilněji, i při malém počtu pozorování je dodržena hladina testu, při větším počtu pozorování je jeho síla srovnatelná s t-testem.

2.2 Test nezávislosti

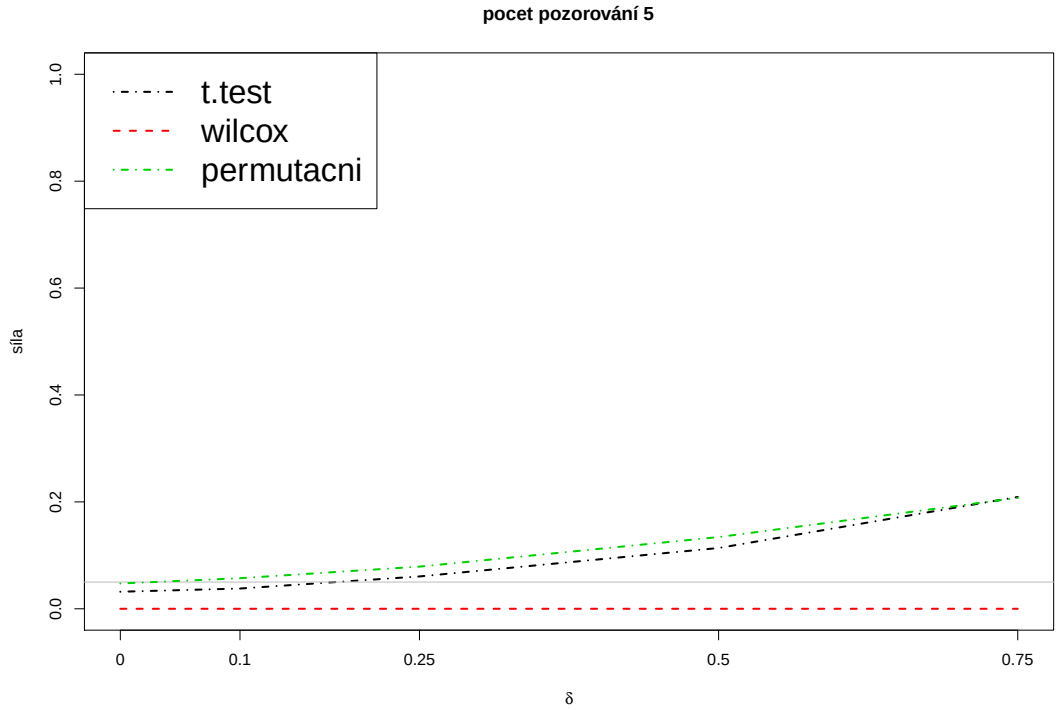
Nyní budeme chtít otestovat nezávislost dvou náhodných veličin. Mějme $\mathbb{X} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ iid náhodné vektory. Zajímá nás nyní, zda jsou náhodné veličiny X_i, Y_i nezávislé.

Formálně máme tedy $\mathbb{X} \sim F_{\boldsymbol{\theta}}$,

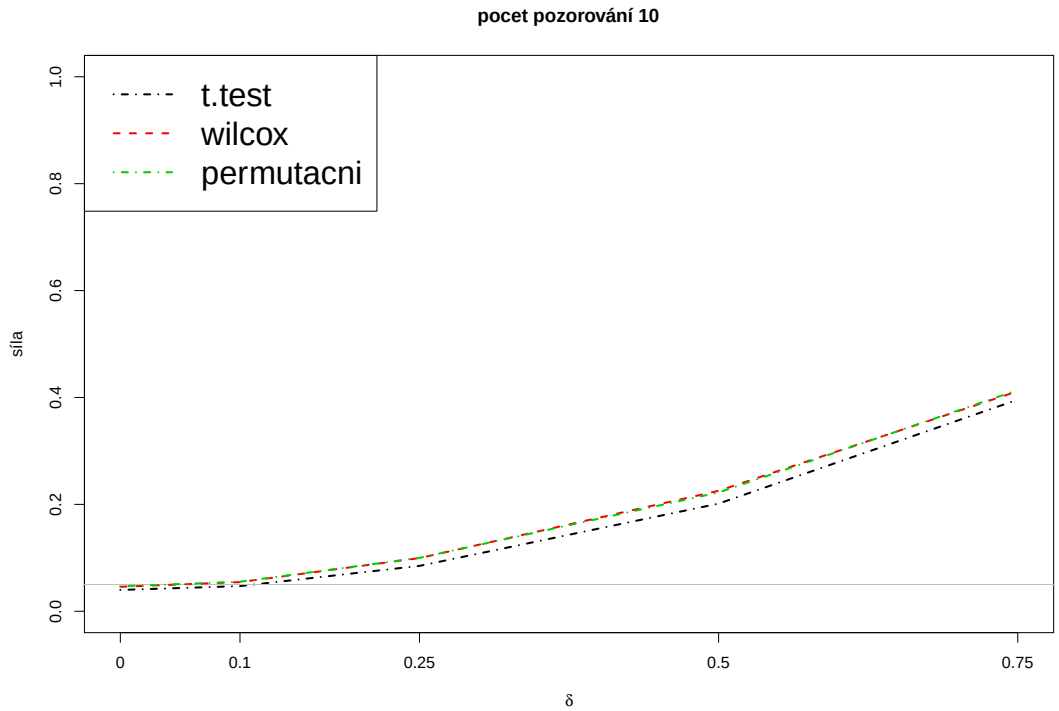
$$\Theta_0 = \left\{\boldsymbol{\theta} = f_n : \; f_n(\mathbf{x}) = \prod_i f^1(x_i) f^2(y_i)\right\}$$

$$\Theta = \left\{\boldsymbol{\theta} = f_n : \; f_n(\mathbf{x}) = \prod_i f(x_i, y_i)\right\}$$

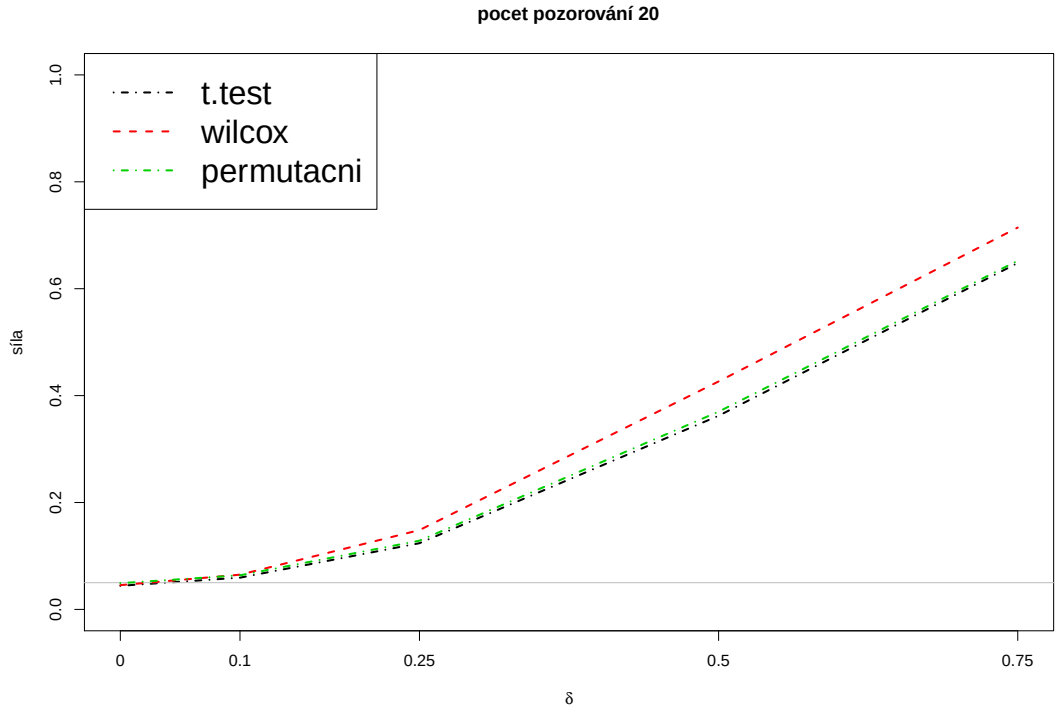
[1] Primárně nám v těchto grafech jde o porovnávání parametrické a permutační varianty testu. Pořadové testy zde uvádíme pouze pro ilustraci. Implementace pořadových testů je převzata z knihoven softwaru R a není předmětem této práce.



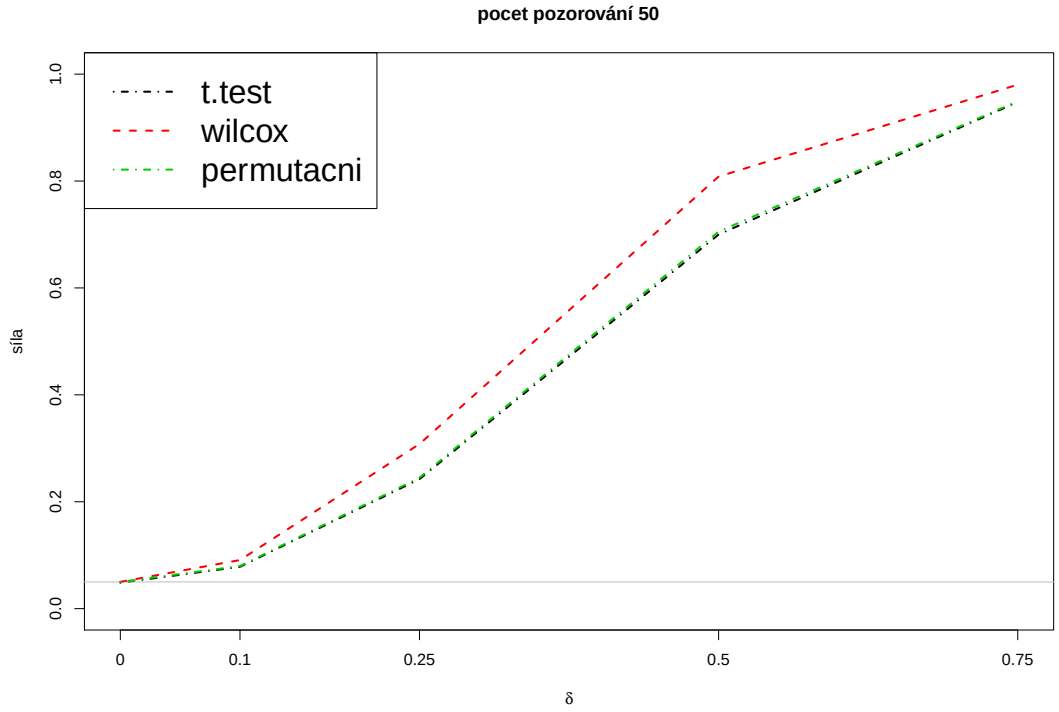
Obrázek 2.1: Síly testu symetrie při pěti pozorováních a hladině testu 0.05: Síla testu byla odhadnuta pro parametry $\delta = 0, 0.1, 0.25, 0.5, 0.75$ při počtu 10000 simulací pro každou hodnotu δ. Wilcoxonův test pro tak malý počet pozorování není schopen zamítnout nulovou hypotézu.



Obrázek 2.2: Síly testu symetrie při deseti pozorováních a hladině testu 0.05: Síla testu byla odhadnuta pro parametry $\delta = 0, 0.1, 0.25, 0.5, 0.75$ při počtu 10000 simulací pro každou hodnotu δ.



Obrázek 2.3: Síly testu symetrie při dvaceti pozorováních a hladině testu 0.05: Síla testu byla odhadnuta pro parametry $\delta = 0, 0.1, 0.25, 0.5, 0.75$ při počtu 10000 simulací pro každou hodnotu δ.



Obrázek 2.4: Síly testu symetrie při padesáti pozorováních a hladině testu 0.05: Síla testu byla odhadnuta pro parametry $\delta = 0, 0.1, 0.25, 0.5, 0.75$ při počtu 10000 simulací pro každou hodnotu δ.

Test bude ve tvaru:

$$H_0 : \boldsymbol{\theta} \in \Theta_0$$
$$H_1 : \boldsymbol{\theta} \in \Theta \setminus \Theta_0$$

Orbita zde vypadá následovně

$$\mathfrak{X}_{|\mathbf{x}} = \{\mathbf{x}^* : \mathbf{x}^* = \{(x_1^*, y_1^*), \ldots, (x_n^*, y_n^*)\}, (x_1^*, \ldots, x_n^*) = \pi_1 (x_1, \ldots, x_n), \\ (y_1^*, \ldots, y_n^*) = \pi_2 (y_1, \ldots, y_n), \ \pi_1, \pi_2 \in \Pi_n\} \tag{2.4}$$

Za H_0 platí

$$\mathbf{x}^* \in \mathfrak{X}_{|\mathbf{x}} : \ \frac{f_n(\mathbf{x}^*)}{f_n(\mathbf{x})} = \frac{\prod_i f^1(x_{u_i}) f^2(y_{v_i})}{\prod_i f^1(x_i) f^2(y_i)} = \frac{\prod_i f^1(x_i) \prod_i f^2(y_i)}{\prod_i f^1(x_i) \prod_i f^2(y_i)} = 1,$$

kde $(u_1, \ldots, u_n) = \pi_1(1, \ldots, n)$, $(v_1, \ldots, v_n) = \pi_2(1, \ldots, n)$.

Pro $x^{**} \notin \mathfrak{X}_{|\mathbf{x}}$ definované v (2.4) opět existuje hustota z Θ_0 taková, že poměr věrohodností se nerovná jedné.

Za testovou statistiku lze brát například $T(\mathbf{x}) = \sum_i x_i y_i$. Všechny body obrazu statistiky T mají tvar

$$T(\mathbf{x}^*) = \sum_i x_i^* y_i^* = \sum_i x_{u_i} y_{v_i} = \sum_i x_i y_{w_i}, \ (w_1, \ldots, w_n) = \pi(1, \ldots, n).$$

Lze tedy opět brát jen $n!$ prvků tvaru $\{(x_1, y_{w_1}), \ldots, (x_n, y_{w_n})\}$ namísto všech prvků orbity, kterých je $2\,n!$.

Pro ilustraci uvažujme dvě náhodné veličiny X a Y, pro které platí:

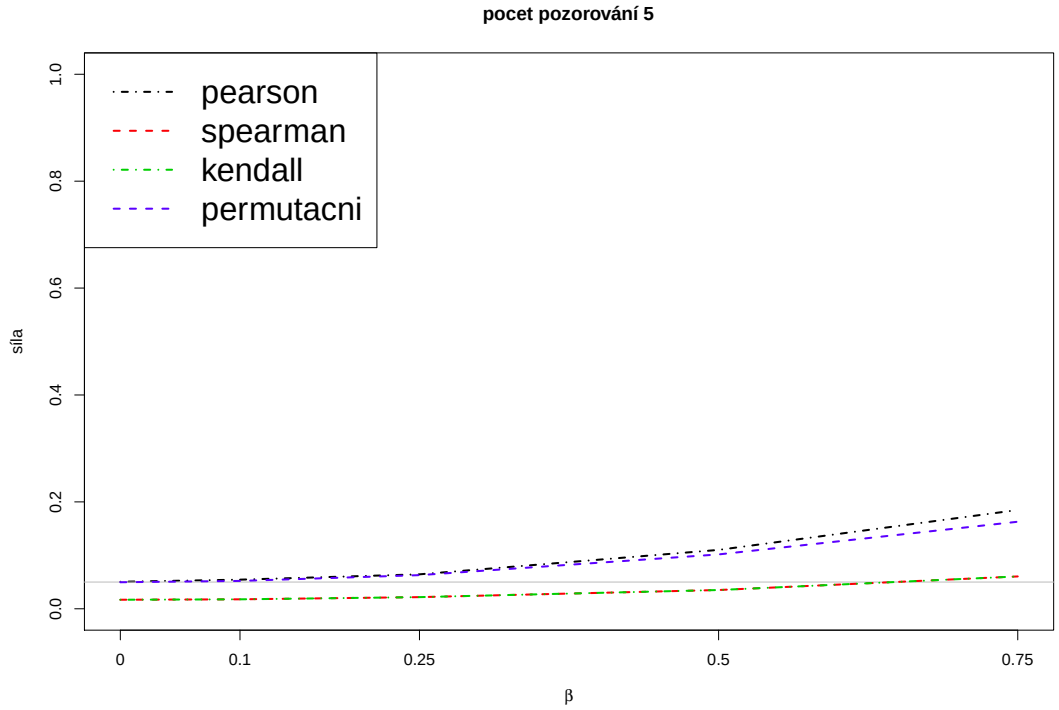
$$X = \epsilon_1$$
$$Y = \beta X + \epsilon_2,$$

kde ϵ_1, ϵ_2 jsou iid $N(0, 1)$. Chceme testovat alternativní hypotézu $H_1 : \beta \neq 0$ oproti nulové hypotéze $H_0 : \beta = 0$. V tomto případě můžeme využít korelační testy jako jsou Pearsonův test, Spearmanův test či Kendallův test. Pearsonův test předpokládá normální rozdělení, naopak další uvedené testy jsou založeny na pořadových statistikách a normalitu rozdělení pro své korektní fungování předpokládat nepotřebují. Zároveň však tím, že využívají pořadové statistiky, ztrácí část informace o pozorováních a tedy dosahují nižších sil testu než Pearsonův test.

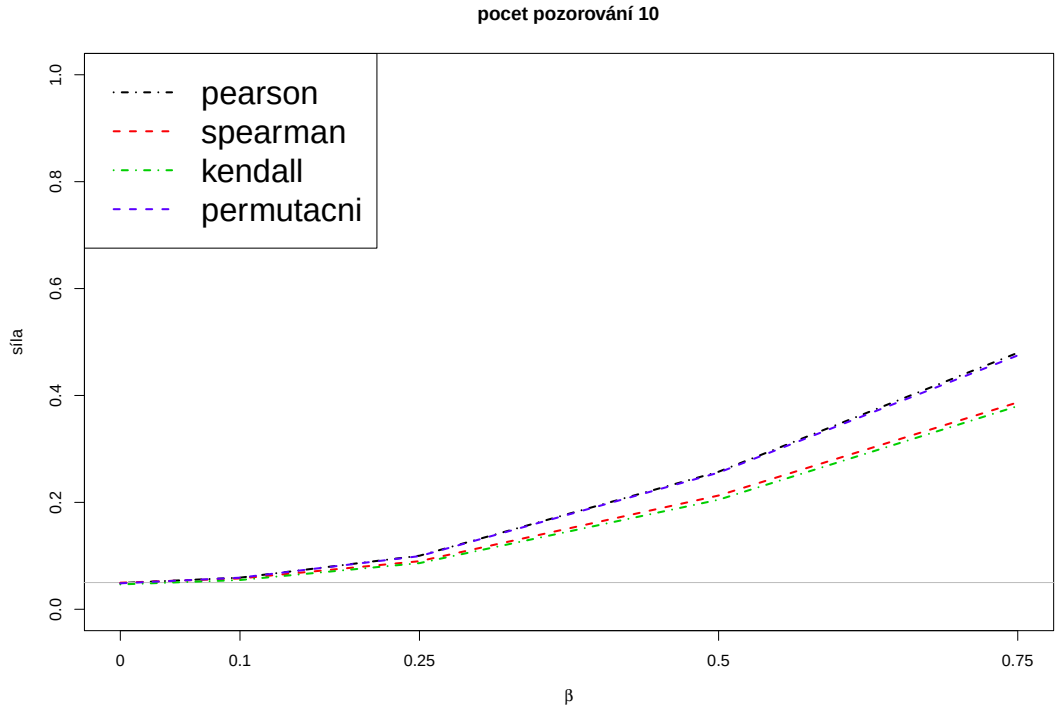
Permutační test založený na výše uvedené statistice T normalitu také předpokládat nepotřebuje, přesto využívá skutečných hodnot pozorování. Jeho síla pak dosahuje hodnot srovnatelných s Pearsonovým testem - viz grafy (2.5), (2.6), (2.7), (2.8). Permutační test popsaný výše je vlastně permutační obdobou Pearsonova testu. Pearsonův test vychází z výběrového korelačního koeficientu

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}},$$

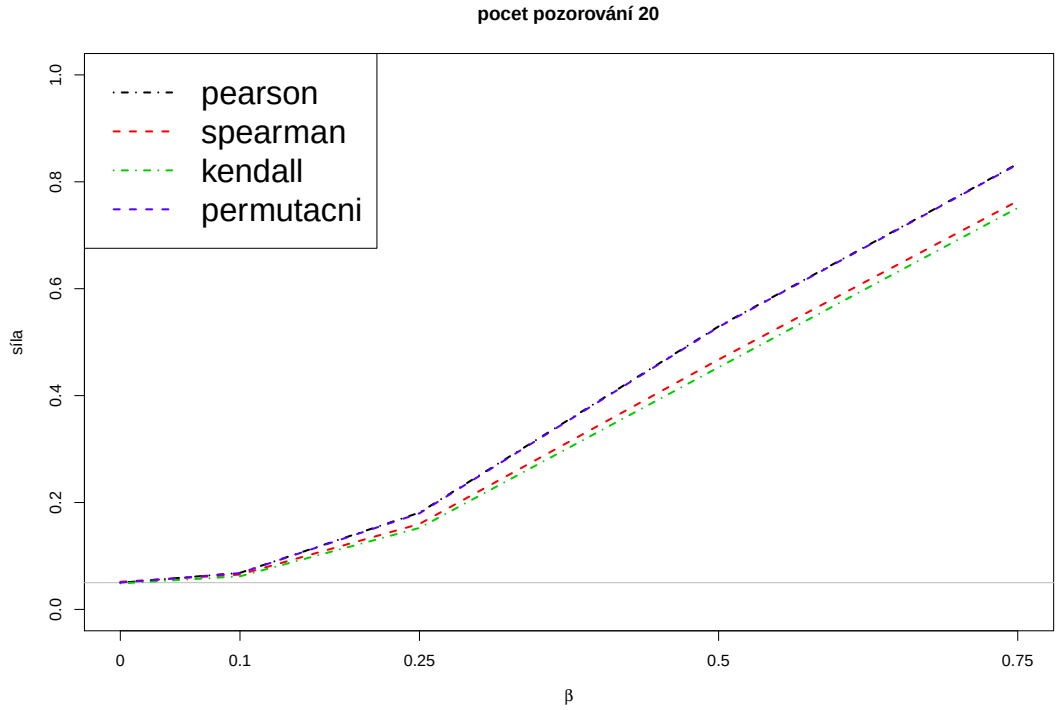
který je permutačně ekvivalentní statistice $T(\mathbf{X}) = \sum_i X_i Y_i$. Použitím této permutační obdoby získáme jistotu, že náš test bude korektní i bez předpokladu normality. Zaplatíme za to pouze velmi malým snížením síly testu.



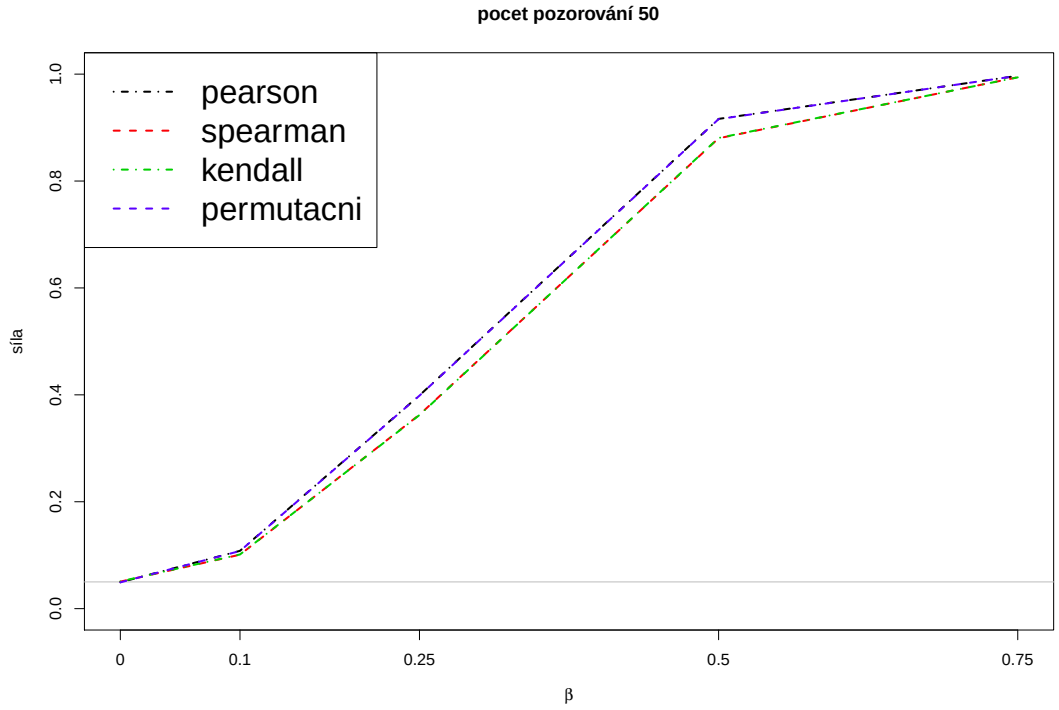
Obrázek 2.5: Síly testu korelace při pěti pozorováních a hladině testu 0.05: Hodnoty sil Spearmanova a Kendallova testu pod hodnotou 0.05 jsou dány velký počtem remíz. Síla testu byla odhadnuta pro parametry $\beta = 0, 0.1, 0.25, 0.5, 0.75$ při počtu 20000 simulací pro každou hodnotu β.



Obrázek 2.6: Síly testu korelace při deseti pozorováních a hladině testu 0.05: Síla testu byla odhadnuta pro parametry $\beta = 0, 0.1, 0.25, 0.5, 0.75$ při počtu 20000 simulací pro každou hodnotu β.



Obrázek 2.7: Síly testu korelace při dvaceti pozorováních a hladině testu 0.05: Síla testu byla odhadnuta pro parametry $\beta = 0, 0.1, 0.25, 0.5, 0.75$ při počtu 20000 simulací pro každou hodnotu β.



Obrázek 2.8: Síly testu korelace při padesáti pozorováních a hladině testu 0.05: Síla testu byla odhadnuta pro parametry $\beta = 0, 0.1, 0.25, 0.5, 0.75$ při počtu 20000 simulací pro každou hodnotu β.

2.3 Bod změny

Mějme $\mathbb{X} = (X_1, \ldots, X_n)$ nezávislé. Data zde mohou představovat měření nějaké vlastnosti výrobků sjíždějících z výrobní linky. V ideálním případě mají $X_1, \ldots, X_n$ stejné rozdělení. My budeme testovat, zda se mezi I-tým a $I+1$-ním výrobkem nezměnilo rozdělení měřené vlastnosti, tedy zda nedošlo například k nějaké poruše výrobní linky. Pokud bychom znali I, pak se bavíme o klasickém dvouvýběrovém testu. My nyní I neznáme, vyšetřujeme, zda vůbec bod zlomu I existuje.

Nulová a alternativní hypotéza mají tedy tvar:

$$\begin{aligned} &H_0 : X_1, \ldots, X_n \sim F \\ &H_1 : \exists I \in \{1, 2, \ldots n-1\} : X_1, \ldots, X_I \sim F_1, \\ &\qquad X_{I+1}, \ldots, X_n \sim F_2,\ F_1 \neq F_2 \end{aligned}$$

Formálně pak $\mathbb{X} \sim F_{\boldsymbol{\theta}}$

$$\Theta_0 = \left\{ \boldsymbol{\theta} = f_n :\ f_n(\mathbf{x}) = \prod_i f(x_i) \right\}$$

$$\Theta_1 = \left\{ \boldsymbol{\theta} = f_n :\ \exists I \in \{1, 2, \ldots n-1\} : \right.$$

$$\left. f_n(\mathbf{x}) = \prod_{i=1}^{I} f^1(x_i) \prod_{j=I+1}^{n} f^2(x_j),\ f^1 \neq f^2 \right\}$$

Pokud budeme předpokládat, že v bodě zlomu I nastala pouze změna střední hodnoty (resp. obecně polohy, existenci střední hodnoty nemusíme předpokládat), můžeme alternativu zúžit:

$$\Theta_1^1 = \left\{ \boldsymbol{\theta} \in \Theta_1, \exists \delta \neq 0 :\ f_1(x) = f_2(x - \delta)\ \forall x,\ \int x f(x)\,\mathrm{d}x \in \mathbb{R} \right\}$$

V tomto tvaru můžeme dělat i jednostranné testy.

Rozdělení za nulové hypotézy odpovídá případu (1.15), taktéž tedy orbita.

Pro alternativu Θ_1^1 lze užít testovou statistiku ve tvaru

$$T(\mathbf{x}) = \max_{1 \leq i < n} \left[\left(\frac{i}{n} W_n - W_i \right)^2 \cdot \{ i(n-i) \}^{-1} \right], \tag{2.5}$$

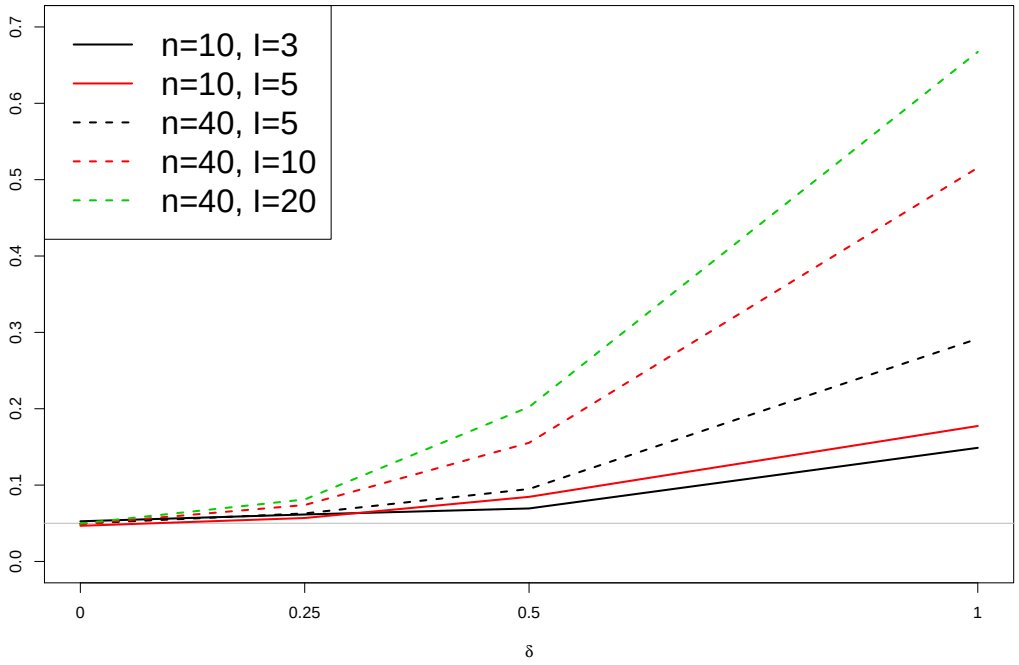
kde $W_i = x_1 + \ldots + x_i$.

Parametrické testy se zde konstruují velmi obtížně. Například testová statistika (2.5) je vlastně maximem $n-1$ závislých náhodných veličin a rozdělení takové veličiny se nepočítá pěkně. Navíc je rozdělení maxima citlivé na rozdělení dat a tedy na dodržení předpokladů. Test se dá založit i na metodách maximální věrohodnosti (viz [FJK10]). Neparametrické přístupy k tomuto problému lze nalézt například v [Pet79].

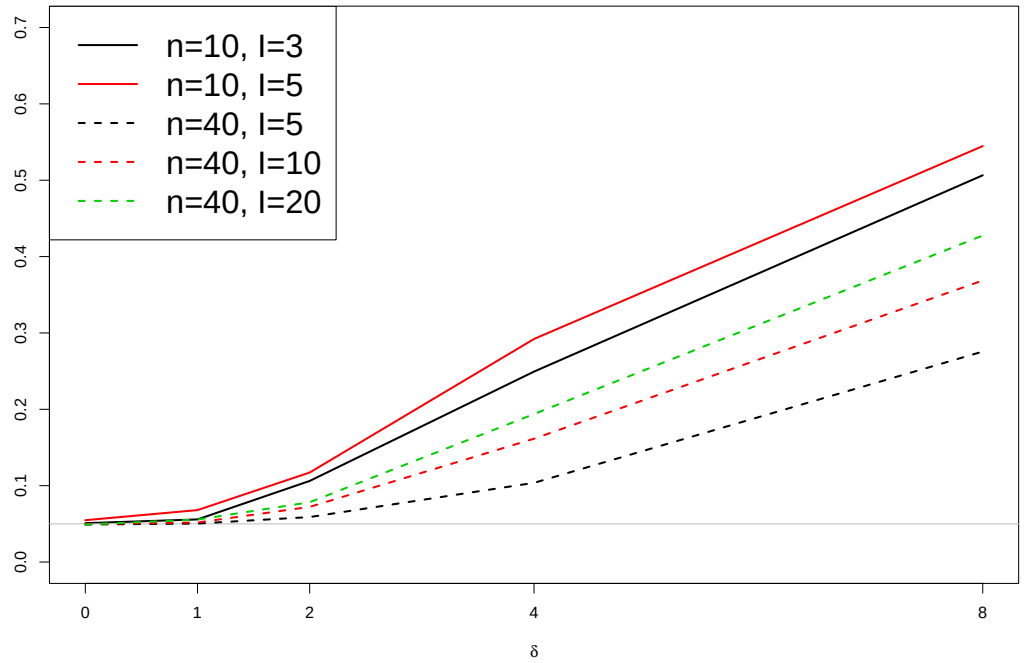
Pro ilustraci síly proveďme simulace případu:

$$\begin{aligned} X_1, \ldots, X_I &\sim N(0, 1) \\ X_{I+1}, \ldots, X_n &\sim N(\delta, 1), \end{aligned}$$

pro $\delta = 0, 0.25, 0.5, 1$ a pro několik dvojic parametrů n, I. Výsledek je v obrázku (2.9). Vidíme, že test dodržuje hladinu testu. Pro porovnání je v obrázku (2.10) stejná simulace pro Cauchyho rozdělení. Parametr δ v tomto případě odpovídá mediánu rozdělení a síla byla odhadnuta pro hodnoty $\delta = 0, 1, 2, 4, 8$. Cauchyho rozdělení má mnohem těžší ocasy než normální rozdělení, je tedy potřeba většího posunu k dosažení stejné síly jako při normálním rozdělení.



Obrázek 2.9: Síly testu existence bodu změny při normálním rozdělení a hladině testu 0.05: Síla testu byla odhadnuta pro parametry $\delta = 0, 0.25, 0.5, 1$ při počtu 10000 simulací pro každou hodnotu δ.



Obrázek 2.10: Síly testu existence bodu změny při Cauchyho rozdělení a hladině testu 0.05: Síla testu byla odhadnuta pro parametry $\delta = 0, 1, 2, 4, 8$ při počtu 10000 simulací pro každou hodnotu δ.

3. Vícevýběrové testy

3.1 ANOVA

3.1.1 Jednoduché třídění

Formulujme si nejdříve problém analýzy rozptylu jednoduchého třídění. Máme $C \geq 3$ skupin, v i-té skupině n_i iid pozorování, tedy

$$\mathbb{X}_i = (X_{i1}, \ldots, X_{in_i}).$$

Všechna pozorování jsou navzájem nezávislá. Naším cílem je otestovat hypotézu, že všechny pozorování pocházejí ze stejného rozdělení (tj.

$$\mathbb{X} = (X_{11}, \ldots, X_{1n_1}, X_{21}, \ldots, X_{cn_C})$$

jsou iid) proti alternativě, že se rozdělení liší podle skupin.

Označme si nyní $X_{ij} \sim F_i$, $i = 1, \ldots, C$, $j = 1, \ldots, n_i$. Nulová hypotéza má pak tvar

$$H_0 : F_1 = F_2 = \ldots = F_C.$$

Alternativa závisí na tom, v jaké se pohybujeme rodině rozdělení. Nejčastěji předpokládáme normální rozdělení u všech pozorování, které se mezi skupinami mohou lišit pouze střední hodnotou. Alternativní hypotéza tedy říká, že existuje alespoň jedna dvojice skupin, která se liší středními hodnotami. V tomto případě pro testování užijeme F-test (viz [Aff11]).

Přejdeme k obecnější situaci. Máme $\mathbb{X} \sim F_{\boldsymbol{\theta}}$. Nulové hypotéze odpovídá prostor parametrů

$$\Theta_0 = \left\{ \boldsymbol{\theta} = F_n,\ F_n(\mathbf{x}) = \prod_{i=1}^{n} F(x_i) \right\},$$

kde $n = n_1 + n_2 + \ldots + n_C$. Alternativou může být, že se skupiny liší pouze posunem, tedy $X_{ij} = \delta_i + Z_{ij}$, kde $Z_{ij}, \forall i, j$ jsou iid náhodné veličiny a existuje alespoň jedna dvojice skupin k, l taková, že $\delta_k \neq \delta_l$. Formálně zapíšeme

$$\begin{aligned}
\Theta^1 = \Bigg\{ & \boldsymbol{\theta} = F_n,\ F_n(\mathbf{x}) = \prod_{i=1}^{n_1} F^1(x_{1i}) \cdot \prod_{i=1}^{n_2} F^2(x_{2i}) \ldots \prod_{i=1}^{n_C} F^C(x_{Ci}), \\
& \mathbf{x} = (x_{11}, \ldots, x_{1n_1}, x_{21}, \ldots, x_{Cn_C}) \\
& \exists \delta_1, \ldots, \delta_C \in \mathbb{R} : F^i = F(x - \delta_i), i = 1, \ldots, C \Bigg\} \\
\Theta_1^1 = & \Theta^1 \setminus \Theta_0.
\end{aligned} \tag{3.1}$$

Obecnějším případem může být alternativa, kde namísto posunu předpokládáme dominanci v distribuci, tedy že alespoň pro jednu dvojici k, l skupin platí $X_{k1} \overset{d}{>} X_{l1}$. Formálně

$$\Theta^2 = \Bigg\{ \boldsymbol{\theta} = F_n,\ F_n(\mathbf{x}) = \prod_{i=1}^{n_1} F^1(x_{1i}) \cdot \prod_{i=1}^{n_2} F^2(x_{2i}) \dots \prod_{i=1}^{n_C} F^C(x_{Ci}),$$
$$\mathbf{x} = (x_{11}, \dots, x_{1n_1}, x_{21}, \dots, x_{Cn_C}),\ \forall 1 \le k, l \le C :$$
$$\left\{F^k(x) \ge F^l(x), \forall x \in \mathbb{R}\right\} \text{nebo} \left\{F^k(x) \le F^l(x), \forall x \in \mathbb{R}\right\} \Bigg\}$$
$$\Theta_1^2 = \Theta^2 \setminus \Theta_0.$$

Nulová hypotéza a tedy i orbita je opět stejná jako v (1.15), tj.

$$\mathfrak{X}_{|\mathbf{x}} = \left\{\mathbf{x}^*, \mathbf{x}^* = \pi\left(\mathbf{x}\right), \pi \in \Pi_n\right\}.$$

Vhodnou statistikou je

$$S(\mathbf{x}) = \sum_{j=1}^{C} \left(\overline{Y}_j - \overline{Y}_\cdot\right)^2 n_j,$$

kde $\mathbf{x} = (x_{11}, \dots, x_{1n_1}, x_{21}, \dots, x_{Cn_C})$, $\overline{Y_j} = \sum_{i=1}^{n_j} x_{ji}/n_j$, je průměr v j-té skupině a $\overline{Y}_\cdot = \sum_{i=1}^{n} x_i/n$ je celkový průměr. Pro naše účely si můžeme statistiku S zjednodušit

$$\begin{aligned}
S(\mathbf{x}) &= \sum_{j=1}^{C} \left(\overline{Y}_j - \overline{Y}_\cdot\right)^2 n_j \\
&= \sum_{j=1}^{C} \left(\overline{Y}_j^2 - 2\overline{Y}_j\overline{Y}_\cdot + \overline{Y}_\cdot^2\right) n_j \\
&= \sum_{j=1}^{C} n_j\overline{Y}_j^2 - 2\overline{Y}_\cdot \sum_{j=1}^{C} n_j\overline{Y}_j + \overline{Y}_\cdot^2 \sum_{j=1}^{C} n_j \\
&= \sum_{j=1}^{C} n_j\overline{Y}_j^2 - 2n\overline{Y}_\cdot^2 + n\overline{Y}_\cdot^2 \\
&= \sum_{j=1}^{C} n_j\overline{Y}_j^2 - n\overline{Y}_\cdot^2.
\end{aligned}$$

Avšak $\overline{Y}_\cdot$ nezávisí na permutaci $\mathbf{x}$ (je tedy stejná pro všechny $\mathbf{x}^* \in \mathfrak{X}_{|\mathbf{x}}$), proto statistika S je permutačně ekvivalentní se statistikou

$$T(\mathbf{x}) = \sum_{j=1}^{C} n_j\overline{Y}_j^2. \tag{3.2}$$

Vidíme, že obě tyto statistiky a rozdělení dat i za platnosti alternativní hypotézy jsou invariantní vůči permutacím v rámci skupin. Není tedy nezbytné vyčíslovat statistiku ve všech $n!$ bodech orbity $\mathfrak{X}_{|\mathbf{x}}$, ale pouze v přerovnáních $\mathbb{X}$ do C skupin o $\{n_1, \dots, n_C\}$ prvcích, těch je $n!/\left(n_1! \cdot n_2! \dots \cdot n_C!\right)$.

Neparametrickou variantou testu je Kruskal-Wallisův test. Tento test se zakládá na statistice

$$Q = \frac{12}{n\,(n+1)} \sum_{i=1}^{C} \frac{T_i^2}{n_i} - 3\,(n+1)\,,$$

kde

$$T_j = \sum_{k=1}^{n_j} R_{jk}$$
$$R_{ij} = \sum_{r=1}^{C} \sum_{k=1}^{n_r} I\,(X_{ij} \leq X_{rk})\,.$$

Testová statistika má podle [HŠ67] za platnosti H_0 asymptoticky (s $\min(n_i)$ jdoucím do nekonečna) χ^2_{C-1} rozdělení.

Porovnejme si síly permutačního testu založeném na statistice T (viz (3.2)) s parametrickým F-testem a neparametrickým Kruskal - Wallis testem. Pro simulaci volíme alternativu ve tvaru Θ_1^1 (viz (3.1)), kde F je distribuční funkce $N(0,1)$ rozdělení,

$$C = 3,\ n_1 = n_2 = n_3 \in \{5, 10, 20, 50\}\,,$$
$$\delta_1 = \delta_2 = 0, \delta_3 \in \{0, 0.1, 0.25, 0.5, 0.75\}\,.$$

Výsledné grafy se nachází na obrázcích (3.1), (3.2), (3.3) a (3.4).Ačkoliv síla všech tří testů se v tomto případě od sebe liší jen minimálně, lze pozorovat, že permutační test nám dává většinou lepší výsledky než neparametrický Kruskal - Wallis a téměř totožné jako parametrický F-test. Jistota správnosti použití testu v případě, kdy by předpoklad normality dat nemusel být splněn, jistě převáží malou ztrátu síly při použití permutační obdoby parametrického testu.

3.1.2 Dvojité třídění

Problém dvojitého třídění, resp. opakovaných pozorování spočívá v tom, že máme n subjektů a na každém subjektu uděláme K měření v K různých podmínkách. Připouštíme, že se subjekty mezi sebou mohou lišit, avšak předmětem zkoumání je pouze vliv podmínek na odezvu.

Máme tedy pozorování

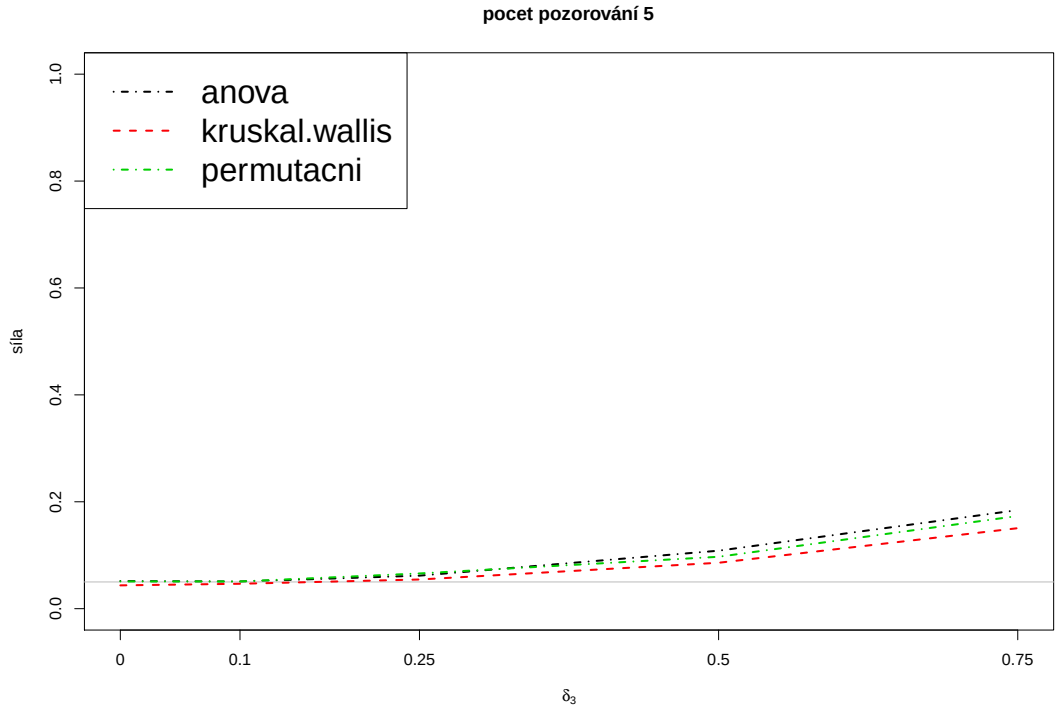
$$\begin{aligned}\mathbb{X} &= \{X_{ij}, i = 1, \ldots, n, j = 1, \ldots, K\} \\ &= \{\mathbb{X}_i = (X_{i1}, \ldots, X_{iK})\,, i = 1, \ldots, n\}\,,\end{aligned}$$

kde $\mathbb{X}_i$ jsou měření na jednom subjektu. Předpokládáme model

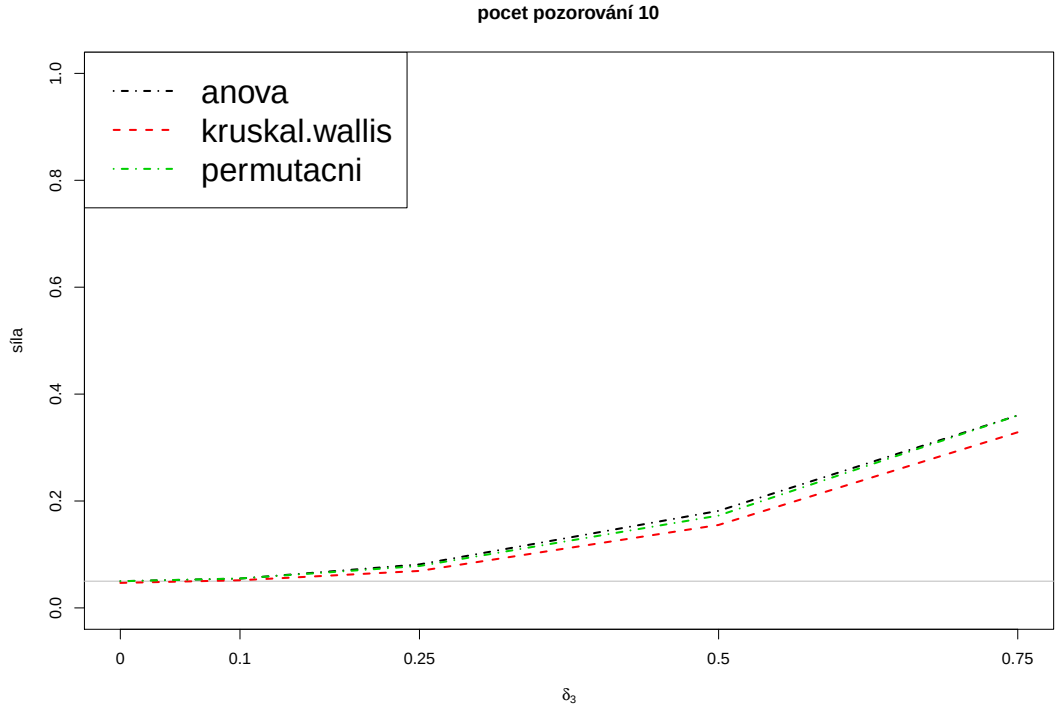
$$X_{ij} = \mu_i + \delta_j + \sigma_i Z_{ij},$$

kde $\{Z_{ij}, i, j\}$ jsou iid. Hypotézy pak jsou

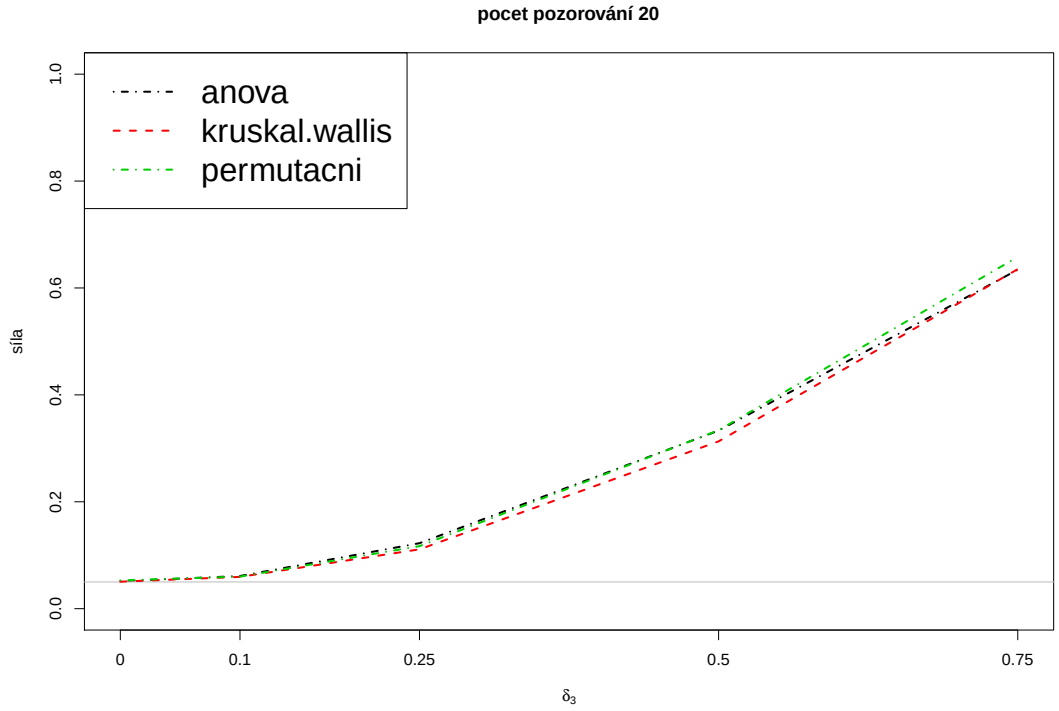
$$\begin{aligned}H_0 &: \delta_j = \delta_k, \forall j, k = 1, \ldots, K \\ &\quad \left(\Leftrightarrow \left\{X_{i1} \overset{d}{=} X_{i2} \overset{d}{=} \ldots \overset{d}{=} X_{iK}, \forall i = 1, \ldots, n\right\}\right) \\ H_1 &: \text{neplatí } H_0\end{aligned}$$



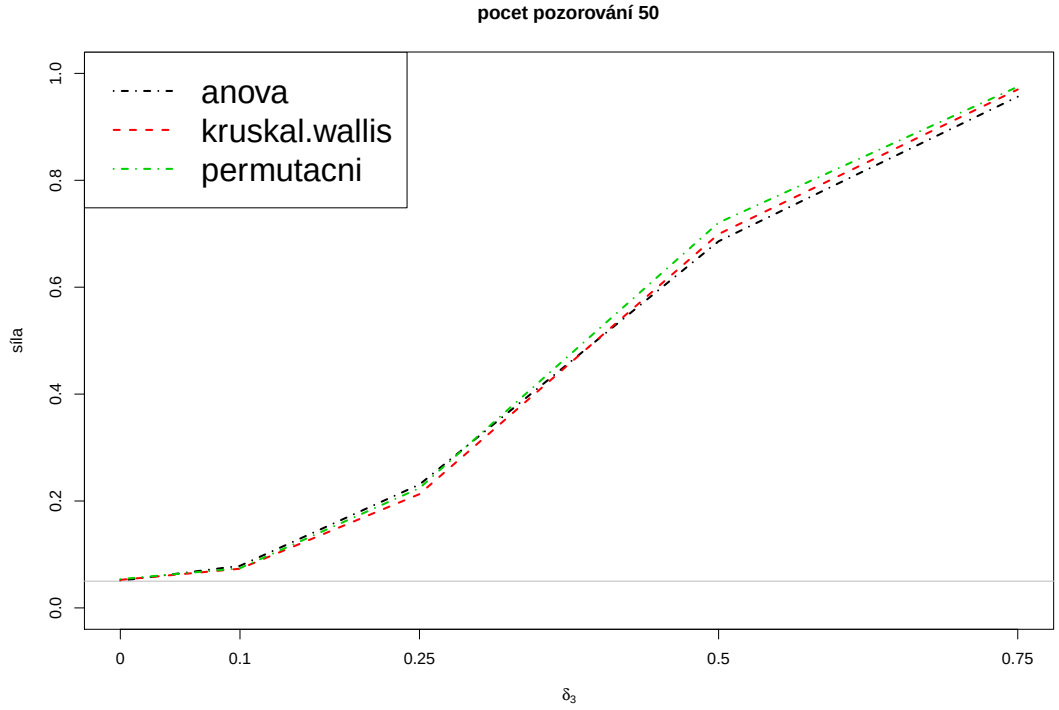
Obrázek 3.1: Síly analýzy rozptylu při třech skupinách o pěti pozorováních a hladině testu 0.05. Bylo provedeno 10000 simulací pro každou hodnotu δ_3.



Obrázek 3.2: Síly analýzy rozptylu při třech skupinách o deseti pozorováních a hladině testu 0.05. Bylo provedeno 10000 simulací pro každou hodnotu δ_3.



Obrázek 3.3: Síly analýzy rozptylu při třech skupinách o dvaceti pozorováních a hladině testu 0.05. Bylo provedeno 10000 simulací pro každou hodnotu δ_3.



Obrázek 3.4: Síly analýzy rozptylu při třech skupinách o padesáti pozorováních a hladině testu 0.05. Bylo provedeno 10000 simulací pro každou hodnotu δ_3.

V námi používaném značení vypadají hypotézy následovně: $\mathbb{X} \sim F_{\boldsymbol{\theta}}$

$$H_0 : \boldsymbol{\theta} \in \Theta_0$$
$$H_1 : \boldsymbol{\theta} \in \Theta_1 = \Theta \setminus \Theta_0,$$

kde

$$\Theta_0 = \left\{ \boldsymbol{\theta} = F_n,\ F_n(\mathbf{x}) = \prod_{i,j} F_i(x_{ij}),\ \mathbf{x} = \{x_{ij}, i = 1, \ldots, n, j = 1, \ldots, K\}, \right.$$
$$\left. \exists \mu_i, \sigma_i,\ i = 1, ..., n :\ F_i\left(\frac{x - \mu_i}{\sigma_i}\right) = F(x), \forall x \in \mathbb{R} \right\}$$

$$\Theta = \left\{ \boldsymbol{\theta} = F_n,\ F_n(\mathbf{x}) = \prod_{i,j} F_{i,j}(x_{ij}),\ \mathbf{x} = \{x_{ij}, i = 1, \ldots, n, j = 1, \ldots, K\}, \right.$$
$$\left. \exists \mu_i, \sigma_i, \delta_j :\ F_{ij}\left(\frac{x - \mu_i - \delta_j}{\sigma_i}\right) = F(x), \forall x \in \mathbb{R} \right\}.$$

Za nulové hypotézy jsou měření uvnitř subjektu záměnné (zde jsou iid, někdy stačí předpokládat jen záměnnost), mezi subjekty se ale zaměňovat nemohou. Orbita tedy má tvar

$$\mathfrak{X}_{|\mathbf{x}} = \left\{ \mathbf{x}^*,\ \mathbf{x}^* = \left\{ x^*_{ij}, i = 1, \ldots, n, j = 1, \ldots, k \right\} : \right.$$
$$\left. (x^*_{i1}, \ldots, x^*_{iK}) = \pi^i (x_{i1}, \ldots, x_{iK}), \pi^i \in \Pi_K, \forall i \right\}$$

a obsahuje $(K!)^n$ bodů.

Při volbě testové statistiky se můžeme opět inspirovat parametrickým řešením a vzít statistiku

$$T(\mathbf{x}) = \frac{\sum_{j=1}^{K} \left(\overline{X}_{\cdot j} - \overline{X}_{\cdot\cdot}\right)^2}{\sum_{i,j} \left(X_{ij} - \overline{X}_{i\cdot} - \overline{X}_{\cdot j} + \overline{X}_{\cdot\cdot}\right)^2},$$

kde

$$\overline{X}_{i\cdot} = \sum_{j=1}^{K} X_{ij}/K,$$
$$\overline{X}_{\cdot j} = \sum_{i=1}^{n} X_{ij}/n,$$
$$\overline{X}_{\cdot\cdot} = \sum_{i,j} X_{ij}/nK.$$

Testová statistika je zde až na konstantu podíl odhadu rozptylu mezi podmínkami a odhadu rozptylu náhodné veličiny Z_{ij}.

Pokud bychom zde chtěli použít pořadový test, pak bychom nejspíše sáhli po Friedmanově testu. Ten je založený na pořadích v rámci vektoru $\mathbb{X}_i$.

3.2 Testy homogenity

3.2.1 Test homogenity pro veličiny s diskrétními uspořádatelnými odezvami

Mějme náhodné veličiny, jejichž možné odezvy označíme $a_1, \ldots, a_K$. U těchto odezev předpokládáme jistou uspořádanost, tedy že má smysl psát $a_i < a_j$, $i < j$.

Může se tedy jednat například o anketové otázky typu: „Jak velkou bolest pociťujete po podání daného léku?“ s předem nastavenými možnostmi pro odpověď od „žádnou“ po „velmi velkou“. V tomto případě budeme chtít testovat, zda se liší účinnost dvou léků tlumících bolest.

Mějme tedy nezávislé veličiny

$$X_{ji},\ j = 1, 2,\ i = 1, \ldots, n_j,$$

pro které platí

$$P\left(X_{ji} = a_k\right) = p_{jk}, \forall i, j, k.$$

Pozorování jsou tedy ve dvou skupinách (každá přísluší jednomu léku), v rámci skupiny jde o iid náhodné veličiny. Označme

$$F_{jk} = F_j(a_k) = P\left(X_{j1} \leq a_k\right) = \sum_{l=1}^{k} p_{jl}.$$

Chceme testovat hypotézu

$$H_0 :\ F_{1k} = F_{2k},\ \forall k = 1, \ldots, K$$

proti alternativě

$$H_1 :\ F_{1k} \leq F_{2k},\ \forall k = 1, \ldots, K,\ \exists l :\ F_{1l} < F_{2l}.$$

Nulová hypotéza vlastně odpovídá $X_{11} \stackrel{d}{=} X_{21}$, alternativní $X_{11} \stackrel{d}{>} X_{21}$.

Označme si $\mathbb{X} = (X_{11}, \ldots, X_{1n_1}, X_{21}, \ldots, X_{2n_2})$ a $n = n_1 + n_2$. Máme $\mathbb{X} \sim F_{\boldsymbol{\theta}}$. Nulovou hypotézu vyjadřuje prostor parametrů

$$\Theta_0 = \left\{ \boldsymbol{\theta} = F_n : F_n(\mathbf{x}) = \prod_{i=1}^{n} F(x_i), \right\}$$

alternativní hypotézu pak

$$\Theta_1 = \Theta \setminus \Theta_0,$$

kde

$$\Theta = \left\{ \boldsymbol{\theta} = F_n : F_n(\mathbf{x}) = \prod_{i=1}^{n_1} F^1(x_i) \prod_{i=n_1+1}^{n} F^2(x_i), \right.$$
$$\left. F^1(x) \leq F^2(x), \forall x \in \{a_1, \ldots, a_K\} \right\}$$

.

Orbita je klasická, tedy $\mathfrak{X}_{|\mathbf{x}} = \{\mathbf{x}^*, \mathbf{x}^* = \pi\left(\mathbf{x}\right), \pi \in \Pi_n\}$.

Test založíme na skóre přiřazeném jednotlivým skupinám, l-té skupině přiřadíme skóre ω_l. Pro skóre musí platit $\omega_i < \omega_j,\ i < j$. Označme si nyní

$$\boldsymbol{\omega} = (\omega_1, \ldots, \omega_K)$$
$$f_{jk} = \sum_{i=1}^{n_j} \mathbf{1}\left(X_{ji} = a_k\right)$$
$$f_{.k} = f_{1k} + f_{2k}.$$

Testová statistika může následně vypadat takto

$$T_{\boldsymbol{\omega}}(\mathbf{x}) = \sum_{l=1}^{K} \omega_l \left(f_{1k} - f_{2k}\right) = \omega^1 - \omega^2,$$

kde ω^j je součet skóre pozorování z j-té skupiny. Uvědomme si, že $\omega^1 + \omega^2$ je konstantní pro všechny prvky $\mathfrak{X}_{|\mathbf{x}}$ a tedy statistika $T_{\boldsymbol{\omega}}$ je permutačně ekvivalentní se statistikou $T'_{\boldsymbol{\omega}} = \omega^1$. Nevýhodou testu je závislost na volbě skóre. Lze například použít triviální skóre $\omega_i = i$.

Můžeme se inspirovat dalšími parametrickými testy a použít statistiky bez nutnosti volby skórů jako například:

$$\begin{aligned} T_1(\mathbf{x}) =& \sum_{i=1}^{K-1} (\hat{F}_{2i} - \hat{F}_{1i}) \left(\hat{F}_{\cdot i} \left(1 - \hat{F}_{\cdot i}\right)\right)^{-\frac{1}{2}} \\ T_2(\mathbf{x}) =& \sup_i \left(\hat{F}_{2i} - \hat{F}_{1i}\right), \end{aligned}$$

kde

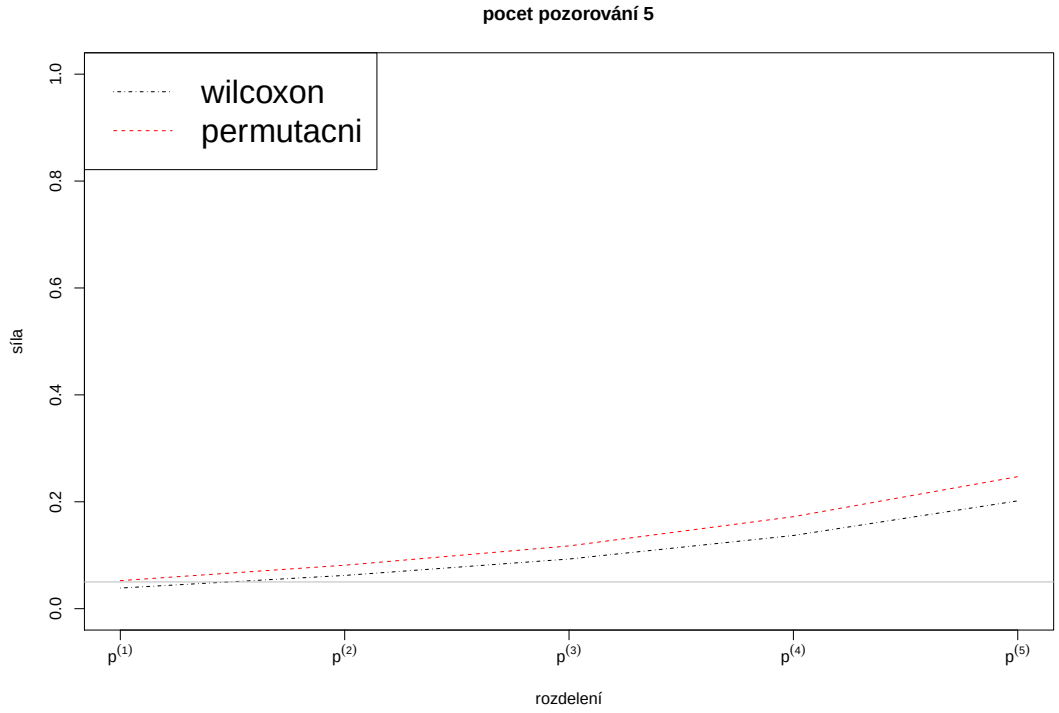
$$\begin{aligned} \hat{F}_{ji} =& N_{ji}/n_j, \\ \hat{F}_{\cdot i} =& N_{\cdot i}/n, \\ N_{ji} =& \sum_{k=1}^{i} f_{jk}, \\ N_{\cdot i} =& N_{1i} + N_{2i}. \end{aligned}$$

V praxi opět nemusíme vyčíslovat všech $n!$ permutací, nezáleží na pořadí pozorování v rámci populace ani při platnosti H_1. Stačí nám všechny kombinace, jakými lze rozdělit n prvkovou množinu pozorování na dvě části o n_1 a $n_2 = n - n_1$ prvcích. Těch je tedy $\binom{n}{n_1}$.

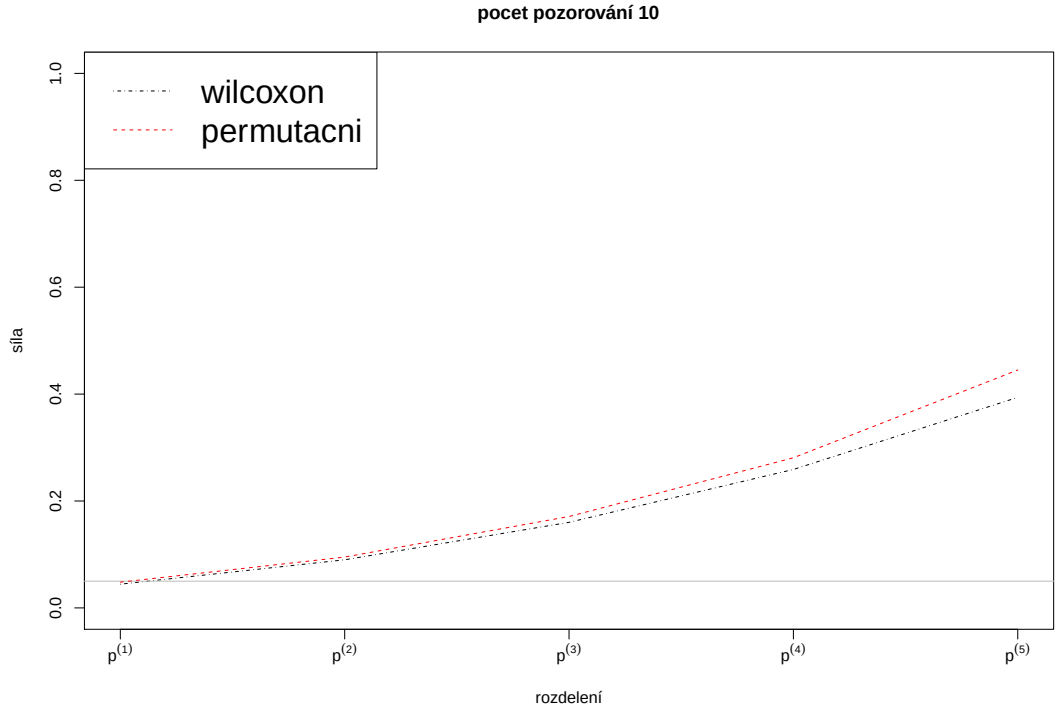
	p_{21}	p_{22}	p_{23}	p_{24}
$p^{(1)}$	0.250	0.250	0.250	0.250
$p^{(2)}$	0.300	0.275	0.225	0.200
$p^{(3)}$	0.350	0.300	0.200	0.150
$p^{(4)}$	0.400	0.325	0.175	0.100
$p^{(5)}$	0.450	0.350	0.150	0.050

Tabulka 3.1: Tabulka pravděpodobností pro simulaci síly testu homogenity

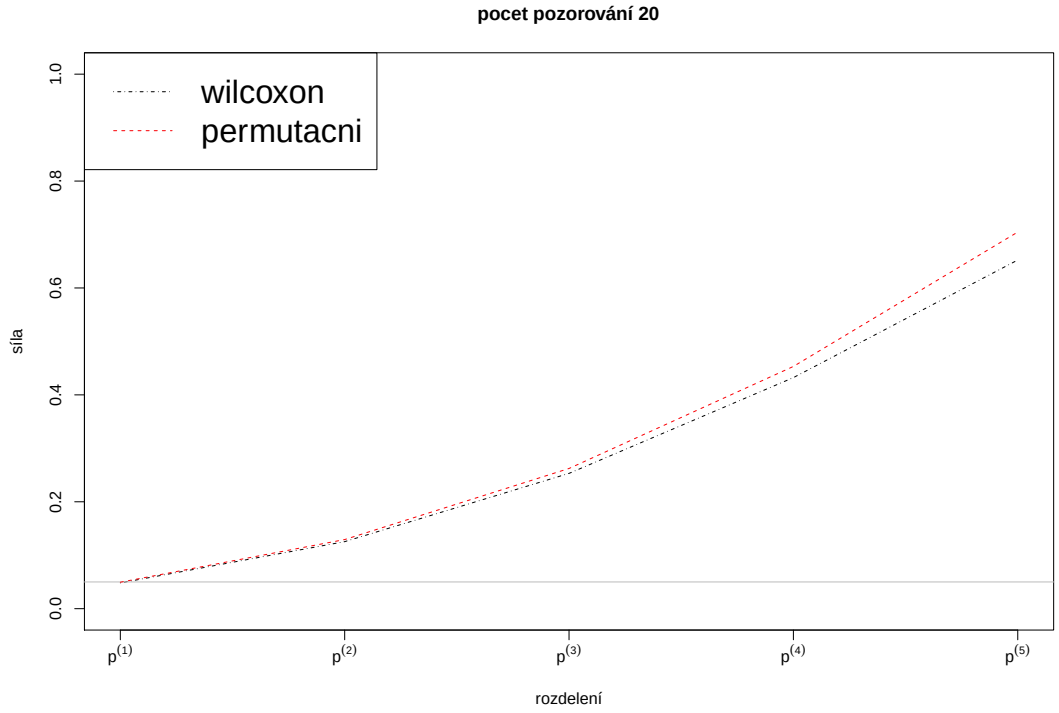
Porovnejme sílu testu založeném na statistice T_1 s dvouvýběrovým Wilcoxonovým testem. Budeme uvažovat dva výběry z diskrétního rozdělení nabývajícího hodnot (kategorií) 1,2,3,4. Platí $p_{1k} = 0.25, k = 1, \ldots, 4$. U rozdělení druhého náhodného výběru měníme pravděpodobnosti podle tabulky (3.1): od $p^{(1)}$ (odpovídá nulové hypotéze) po $p^{(5)}$ (je nejmenší v distribuci). Testujeme jednostrannou alternativu. Výsledky najdeme v grafech (3.5), (3.6), (3.7) a (3.8). Vidíme, že permutační test si zde vedl lépe než neparametrický test.



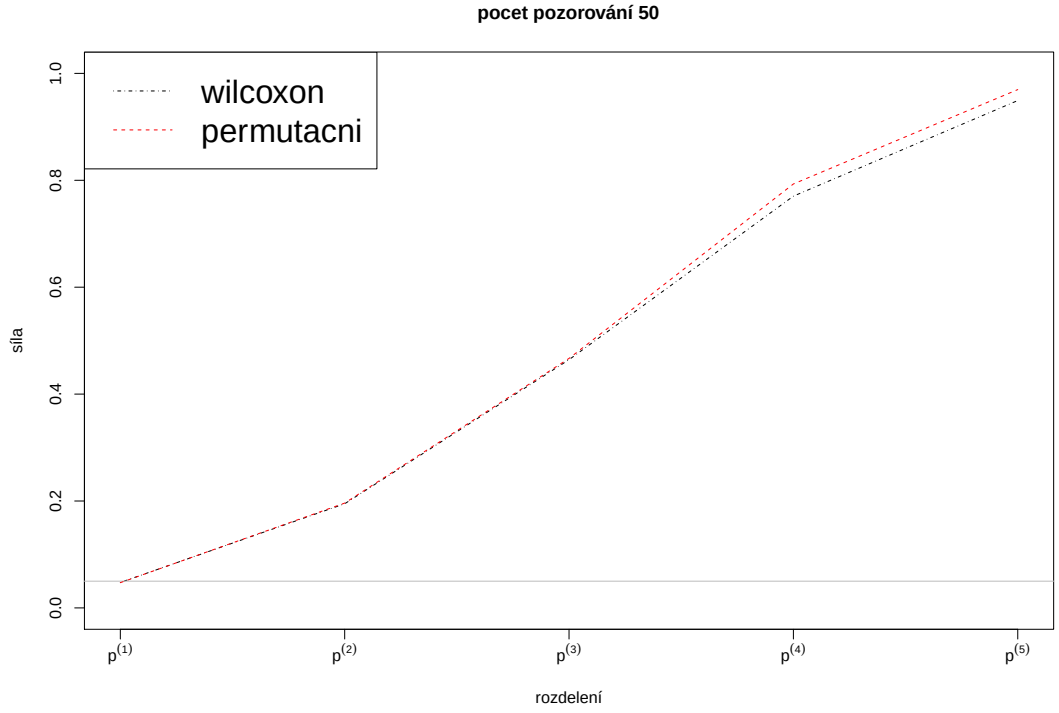
Obrázek 3.5: Síly testu homogenity při dvou skupinách o pěti pozorováních a hladině testu 0.05. Bylo provedeno 10000 simulací pro každou hodnotu $p^{(i)}$.



Obrázek 3.6: Síly testu homogenity při dvou skupinách o deseti pozorováních a hladině testu 0.05. Bylo provedeno 10000 simulací pro každou hodnotu $p^{(i)}$.



Obrázek 3.7: Síly testu homogenity při dvou skupinách o dvaceti pozorováních a hladině testu 0.05. Bylo provedeno 10000 simulací pro každou hodnotu $p^{(i)}$.



Obrázek 3.8: Síly testu homogenity při dvou skupinách o padesáti pozorováních a hladině testu 0.05. Bylo provedeno 10000 simulací pro každou hodnotu $p^{(i)}$.

Můžeme také testovat širší alternativu

$$H_1 : \exists k : F_{1k} \neq F_{2k}. \tag{3.3}$$

Ta není jen oboustranným rozšířením předchozího případu, nemusí vždy dojít k dominanci v distribuci. Odpovídající prostor parametrů má tvar

$$\Theta = \left\{ \boldsymbol{\theta} = F_n : F_n(\mathbf{x}) = \prod_{i=1}^{n_1} F^1(x_i) \prod_{i=n_1+1}^{n} F^2(x_i), \right.$$
$$\left. \exists l \in \{1, \ldots, K\} : \; F^1(a_l) \neq F^2(a_l) \right\}.$$

Pro tento případ lze výše uvedené statistiky jednoduše upravit na

$$\begin{aligned} T'_w =& |T_w|, \\ T'_1 =& |T_1|, \\ T'_2 =& \sup_i \left| \hat{F}_{2i} - \hat{F}_{1i} \right|. \end{aligned}$$

3.2.2 Test homogenity pro veličiny s neuspořádatelnými odezvami

Obdobně jako v podkapitole (3.2.1) mějme odezvy náhodné veličiny $a_1, \ldots, a_K$, nyní ale předpokládejme, že mezi nimi neexistuje uspořádání. Můžeme si představit například anketovou otázku: „Která z těchto barev se Vám líbí nejvíce?“ s nabídnutými možnostmi. Opět uvažujme dvě populace (například muže a ženy). Chtějme testovat, zda se tyto populace liší.

Uvažujeme nezávislé náhodné veličiny

$$X_{ji}, j = 1, 2, i = 1, \ldots n_j,$$

nabývající hodnot $a_1, \ldots, a_K$, kde platí

$$P\left(X_{ji} = a_k\right) = p_{jk}, \forall i, j, k.$$

Hypotézy pak mají následující tvar

$$\begin{aligned} H_0 &: p_{1k} = p_{2k}, \forall k = 1, \ldots, K \\ H_1 &: \exists k \in \{1, \ldots, K\} : p_{1k} \neq p_{2k} \end{aligned}$$

Oproti případu v podkapitole (3.2.1) nemá smysl uvažovat dominanci v distribuci, alternativní hypotéza tedy říká, že se populace liší. Orbita má v tomto případě stejný tvar jako v podkapitole (3.2.1), $\mathfrak{X}_{|\mathbf{x}} = \{\mathbf{x}^*, \mathbf{x}^* = \pi\left(\mathbf{x}\right), \pi \in \Pi_n\}$.

Parametrický test obvykle zakládáme na χ^2 statistice:

$$T_p\left(\mathbf{x}\right) = \sum_{\substack{j=1,2 \\ k=1,\ldots,K}} \frac{n_j \left(\hat{p}_{jk} - \hat{p}_{.k}\right)^2}{\hat{p}_{.k}},$$

kde

$$\hat{p}_{jk} = \frac{f_{jk}}{n_j}$$
$$\hat{p}_{\cdot k} = \frac{f_{\cdot k}}{n},$$

$f_{jk}, f_{\cdot k}$ jako v podkapitole (3.2.1).

Alternativně se lze inspirovat Anderson-Darling testem

$$T_{ad}(\mathbf{x}) = \sum_{\substack{j=1,2 \\ k=1,\ldots,K}} \left(\hat{p}_{jk} - \hat{p}_{\cdot k}\right)^2 / \left[f_{\cdot k}\left(n - f_{\cdot k}\right)\left(n - n_j\right)/n_j\right],$$

Tyto testy je samozřejmě možné použít také pro testování hypotézy uvedené v (3.3).

3.2.3 Test homogenity pro spojité náhodné veličiny

Test homogenity pro spojité náhodné veličiny lze převést na situaci (3.2.1) rozdělením odezvy na kategorie. V této podkapitole se podíváme na další možné testy.

Mějme tedy dvě populace $X_{ji}, j = 1, 2, i = 1, \ldots n_j$, kde X_{ji} jsou nezávislé náhodné veličiny se spojitým rozdělením. Definujme si distribuční funkce F_1, F_2:

$$F_j(x) = P(X_{ji} \leq x), \forall j = 1, 2, i = 1, \ldots n_j$$

Hypotézy pak mají tvar:

$$H_0 : F_1(x) = F_2(x), \forall x \in \mathbb{R}$$
$$H_1 : \exists x \in \mathbb{R} : F_1(x) \neq F_2(x)$$

Orbita je stejná jako v předcházejících podkapitolách. Opět se inspirujme známými parametrickými testy. Statistika založená na Kolmogorov-Smirnov testu:

$$T_{cs} = \sup_{A \in \mathcal{A}} \left|\hat{P}_1(A) - \hat{P}_2(A)\right|,$$

kde $\mathcal{A}$ je odpovídající množina jevů a

$$\hat{P}_j(A) = \int_A \mathrm{d}\hat{F}_j(x)$$
$$\hat{F}_j(x) = \#\left(X_{ji} \leq x\right)/n_j, j = 1, 2$$

Alternativně test založíme na Anderson-Darling testové statistice:

$$T^2_{ad}(x) = \int_{-\infty}^{\infty} \left(\hat{F}_1(x) - \hat{F}_2(x)\right)^2 / \left(\hat{F}(x)\left[1 - \hat{F}(x)\right]\right) \mathrm{d}\hat{F}(x),$$

kde

$$\hat{F}(x) = \frac{n_1\hat{F}_1(x) + n_2\hat{F}_2(x)}{n}$$

Je zřemé, že odhad $\hat{F}$ je stejný pro všechny $\mathbf{x}^* \in \mathfrak{X}_{|\mathbf{x}}$.

Úpravou statistiky T^2_{ad} lze testovat i jednostrannou alternativu:

$$T_{ad}(x) = \int_{-\infty}^{\infty} \left(\hat{F}_1(x) - \hat{F}_2(x)\right) / \left(\hat{F}(x)\left[1 - \hat{F}(x)\right]\right)^{-1/2} \mathrm{d}\hat{F}(x),$$

4. Regresní modely

4.1 Absolutní chyby, čtverce chyb

V této kapitole se budeme zabývat lineárními regresními modely ve tvaru

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{4.1}$$

kde $\mathbf{Y}$ je náhodný vektor délky n (odezva), $\mathbf{X}$ je matice regresorů $n \times k$, $\boldsymbol{\beta}$ je vektor parametrů délky k, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\mathsf{T}$ je vektor chyb.

Nejprve ale budeme uvažovat pouze jeden regresor, tak můžeme zapisovat

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

Naměřené hodnoty budeme označovat malými písmeny x_i, y_i a pro chyby (rezidua) e_i.

$$y_i = \alpha + \beta x_i + e_i$$

V tomto typu úlohy při zadaných (naměřených) $\{(x_i, y_i), i = 1, \ldots, n\}$ chceme nalézt takové hodnoty parametrů α, β, které minimalizují v nějakém smyslu chyby $\{e_i, i = 1, \ldots, n\}$. Nejčastěji se setkáváme s úkolem minimalizovat čtverce chyb, tedy výraz

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left[y_i - (\alpha + \beta x_i)\right]^2 \tag{4.2}$$

Hodnoty parametrů minimalizující (4.2) budeme označovat $\hat{\alpha}, \hat{\beta}$. Tato metoda odhadu parametrů α, β je nazývána metoda nejmenších čtverců (MSE) a je oblíbená proto, že existuje jednoduché vyjádření $\hat{\alpha}, \hat{\beta}$:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$
$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x},$$

kde $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ a $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Historicky ale MSE předcházelo jiné měření chyby. Šlo o úlohu minimalizovat absolutní chyby (MAE):

$$\sum_{i=1}^{n} |e_i| = \sum_{i=1}^{n} |y_i - (\alpha + \beta x_i)| \tag{4.3}$$

Zde budeme označovat hodnoty parametrů minimalizující (4.3) $\tilde{\alpha}, \tilde{\beta}$. Na rozdíl od $\hat{\alpha}, \hat{\beta}$ ale neexistuje jednoduché vyjádření $\tilde{\alpha}, \tilde{\beta}$.

Tyto hodnoty lze nalézt iterativním procesem, proto je dlouhodobě využívanější MSE. Do nedávna nebylo úplně jednoduché počítat $\tilde{\alpha}, \tilde{\beta}$, ale s dnešní výpočetní silou to již není tak velký problém (závisí samozřejmě na velikosti dat).

Vlastnosti těchto dvou metod se v mnohém liší. Porovnání metod nalezneme například v [CR96]. MSE odhad je velmi citlivý na odlehlá pozorování, MAE je robustnější. MAE odhady mají lepší vlastnosti pro asymetrická rozdělení chyb i pro některá symetrická (např. Cauchyho rozdělení, dvojité exponenciální rozdělení), MSE pro normální, trojúhelníkové či rovnoměrné rozdělení. V případě, že chyby jsou *iid* z dvojitého exponenciálního rozdělení, platí $med\,(\tilde{\alpha}) = \alpha, med\left(\tilde{\beta}\right) = \beta$.

Zatímco obzvláště pro normální rozdělení chyb a MSE existují vhodné testy pro testování podmodelů (např. hypotéza $H_0 : \beta = 0$ oproti alternativě $H_1 : \beta \neq 0$), pro MAE je vhodných parametrických testů jen málo. Zde je tedy potenciál permutačních testů.

4.2 Testování podmodelů pro MAE odhady

4.2.1 Testová statistika

Budeme vycházet ze značení (4.1). Předpokládejme, že $\{\epsilon_i, i = 1, \ldots, n\}$ jsou *iid* náhodné veličiny z neznámého rozdělení, pro které platí $med\,(\epsilon_i) = 0$ (na rozdíl od MSE, kde bychom předpokládali $E\epsilon_i = 0$, zde ani nepředpokládáme existenci střední hodnoty).

Pro stanovení podmodelu si rozdělíme regresory

$$\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)\,,$$

kde $\mathbf{X}_1$ je matice $n \times k_1$ a $\mathbf{X}_2$ je matice $n \times k_2$. Platí $k_1 + k_2 = k$, $k_1 > 0, k_2 > 0$.

$$\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$$

kde $\boldsymbol{\beta}_1$ je vektor délky k_1 a $\boldsymbol{\beta}_2$ je vektor délky k_2.

Hypotézy pak postavíme následovně:

$$H_0 :\ \boldsymbol{\beta}_2 = \mathbf{0}$$
$$H_1 :\ \boldsymbol{\beta}_2 \neq \mathbf{0}$$

Nulová hypotéza tedy říká, že odezva závisí jen na prvních k_1 regresorech, na dalších k_2 regresorech nezávisí. Podmodel má tedy tvar:

$$\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}$$

Uvažujme tuto statistiku

$$T\,(\mathbf{X}, \mathbf{Y}) = \frac{SA_{H_0} - SA_{H_1}}{SA_{H_1}} \tag{4.4}$$

kde

$$SA_{H_0} = \min_{b_1,\ldots,b_{k_1} \in \mathbf{R}} \sum_{i=1}^{n} \left| y_i - \sum_{j=1}^{k_1} b_j x_{ij} \right|$$

$$SA_{H_1} = \min_{b_1,\ldots,b_{k_1+k_2} \in \mathbf{R}} \sum_{i=1}^{n} \left| y_i - \sum_{j=1}^{k_1+k_2} b_j x_{ij} \right|$$

Statistika T (4.4) není nepodobná F statistice požívané při testování podmodelu v případě MSE lineární regrese ([Aff11]). Vysoká hodnota statistiky říká, že je velký rozdíl mezi velikostí chyb původního modelu a podmodelu, a svědčí tedy proti nulové hypotéze H_0.

4.2.2 Test nulového podmodelu

Budeme zde předpokládat, že první sloupec matice $\mathbf{X}$ je vektor jedniček, dále $k_1 = 1$. Podmodel tedy lze zapsat ve tvaru:

$$\mathbf{Y} = \alpha + \boldsymbol{\epsilon} \tag{4.5}$$

Za platnosti H_0 jsou tedy jednotlivá pozorování odezvy zcela záměnná, protože nezávisí na žádném regresoru. To nás vede k orbitě

$$\mathfrak{X}_{|\mathbf{y}} = \{\mathbf{y}^*, \mathbf{y}^* = \pi(\mathbf{y}), \pi \in \Pi_n\}$$

Test tedy založíme na testové statistice T (4.4), kterou budeme vyčíslovat pro různé permutace pozorované odezvy $\mathbf{y}$. Uvědomme si, že SA_{H_0} je konstantní

$$\begin{aligned} SA_{H_0} &= \min_{b_1,\ldots,b_{k_1} \in \mathbf{R}} \sum_{i=1}^{n} \left| y_i - \sum_{j=1}^{k_1} b_j x_{ij} \right| \\ &= \min_{\alpha \in \mathbf{R}} \sum_{i=1}^{n} |y_i - \alpha|, \end{aligned}$$

Výraz SA_{H_0} se tedy nemění s permutací $\mathbf{y}$.

Zato výraz SA_{H_1} se bude měnit s každou permutací. Součástí jeho výpočtu je nalezení minimalizujících parametrů. To je ale iterativní proces, který může v případě větší hodnoty n nebo k mít vyšší nároky na výpočetní techniku, neboť jej budeme mnohokrát opakovat pro vyčíslení dostatečného množství $T(\mathbf{x}, \mathbf{y}^*)$.

4.2.3 Test nenulového podmodelu

Zde se budeme zabývat testováním podmodelu, který se nedá zapsat vzorcem (4.5). Zde již orbita není vůbec jednoduchá. Jeden přístup předpokládá více pozorování pro všechny kombinace nezávislé veličiny X (viz [BM82]).

Další uvedené přístupy nejsou přímo permutačními testy tak, jak jsme si je tady definovali. Testy se zakládají na permutaci reziduí podmodelu, na permutaci reziduí modelu nebo na permutaci odezvy $\mathbf{Y}$. Z uvedených metod má nejblíže k principům permutačních testů metoda permutace reziduí podmodelu, která v případě nulového podmodelu odpovídá metodě popsané v kapitole (4.2.2). Porovnání sil této metody s klasickou MSE metodou pomocí simulací provedl Cade a Richards [CR96]. Dochází zde k závěru, že tam, kde je užití klasického modelu opodstatněné (rozdělení chyb je *i.i.d.* s normálním rozdělením nebo s rozdělením, které k normálnímu nemá daleko), dochází užitím MAE metody k mírné ztrátě síly. Jinak je ale MAE metoda robustnější. Podrobnější analýzu efektu heteroskedasticity či odlehlých pozorování na sílu testů lze najít ve zmíněném článku [CR96].

Závěr

Tato práce se dá rozdělit na dva úseky - v první kapitole se zaměřuje na poznatky o pořádkových statistikách, které jsou základním stavebním kamenem permutačních testů. Práce ukazuje, že pro zadání problému testu nulové hypotézy, která obsahuje všechny spojité rozdělení, je permutační test jediným zcela korektním řešením. První kapitola dále demonstruje, jak konstruovat nejsilnější test proti dané alternativě.

Pro zcela korektní vyhodnocení testu je potřeba vyčíslit testovou funkci pro všechny permutace pozorovaných dat. V práci je tedy ukázána metoda Monte Carlo, pomocí které lze redukovat nároky na výpočetní techniku. Zároveň je zde ukázáno, jaké nepřesnosti se dopouštíme využitím této metody. Při počtu $B = 1000$ výběrů z množiny všech permutací dat je chyba již pro praxi zanedbatelná. Ve většině zde uvažovaných testů pro „rozumně“ velký vektor pozorování proběhne takový test na stolním počítači pod 1 sekundu. Pro jednotlivý test nepředstavuje tedy problém zvýšit přesnost i řádově, čas výpočtu se zvýší lineárně s počtem výběrů B.

Druhý úsek se zaměřuje na výkonnost permutačních testů. Porovnávají se zde síly parametrických, permutačních a pořadových testů pro různá zadání. Obvykle platí, že zavedený parametrický test bývá nejsilnější. Motivací pro použití permutačního nebo pořadového testu může být neznámé rozdělení dat. V takovém případě se v praxi buď učiní předpoklady na data, které ale nemusejí být nutně splněny, nebo se může použít nějaký neparametrický test. Simulacemi síly jednotlivých testů se ukazuje, že permutační varianty testů se blíží těm parametrickým. Použitím permutační varianty parametrického testu se jeví pro praxi jako velmi výhodné - dojde k získání jistoty korektnosti testu ve smyslu udržení chyby 1. druhu za cenu jen velmi malé ztráty síly.

Z toho všeho plyne, že permutační testy mají velký potenciál pro praxi. Jejich krása je v nenáročnosti na předpoklady. Lze skoro těžce vymyslet příklad, kdy by bylo použití takového testu neopodstatněné. Parametrické testy bývají v rukou výzkumníků velmi ohýbány pro testování hypotéz, pro které nejsou vůbec vhodné. Pro korektnost jejich výsledků by bylo přínosem, pokud by jejich testy byly nahrazeny permutačními variantami.

Seznam použité literatury

[Aff11] J. Anděl and Univerzita Karlova. Matematicko fyzikální fakulta. *Základy matematické statistiky*. Matfyzpress, 2011.

[BM82] B. M. Brown and J. S. Maritz. Distribution-free methods in regression. *Australian Journal of Statistics*, 24(3):318–331, 1982.

[CR96] Brian S Cade and Jon D Richards. Permutation tests for least absolute deviation regression. *Biometrics*, pages 886–902, 1996.

[FJK10] Stergios B Fotopoulos, Venkata K Jandhyala, and Elena Khapalova. Exact asymptotic distribution of change-point mle for change in the mean of gaussian sequences. *The Annals of Applied Statistics*, pages 1081–1104, 2010.

[HŠ67] Jaroslav Hájek and Zbyněk Šidák. *Theory of rank tests*. Academia, 1967.

[Leh86] E.L. Lehmann. *Testing Statistical Hypotheses*. Springer texts in statistics. Springer, 1986.

[Pes01] F. Pesarin. *Multivariate Permutation Tests: With Applications in Biostatistics*. Wiley, 2001.

[Pet79] AN Pettitt. A non-parametric approach to the change-point problem. *Applied statistics*, pages 126–135, 1979.