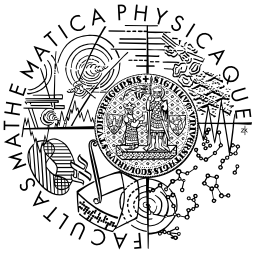


MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

**BAKALÁŘSKÁ PRÁCE**

Šárka Horská

# Rozdělení vzdálenosti mezi body

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: doc. RNDr. Zdeněk Hlávka, Ph.D.
Studijní program: Matematika
Studijní obor: Obecná matematika

Praha 2020

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

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

V ..... dne .....  .....

Podpis autora

Ráda bych poděkovala vedoucímu této bakalářské práce, doc. RNDr. Zdeňkovi Hlávkovi, Ph.D., za jeho ochotu, rady a v neposlední řadě za jeho čas.

Název práce: Rozdělení vzdálenosti mezi body

Autor: Šárka Horská

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: doc. RNDr. Zdeněk Hlávka, Ph.D., katedra pravděpodobnosti a matematické statistiky

Abstrakt: Tato bakalářská práce se zabývá rozdělením vzdálenosti mezi body z multinomického rozdělení a jeho vlastnostmi. Motivací ke studiu tohoto rozdělení je jeho použití při testování dat o velkém počtu kategorií a malém počtu pozorování. Pro takto řídká data není vhodné používat $\chi^2$-testy, ale můžeme použít například testy založené na vzdálenostech mezi body. Mezi takové patří test s testovou statistikou Biswas a Ghosh z roku 2014, které se budeme v práci věnovat.

Klíčová slova: Euklidovská vzdálenost; multinomické rozdělení; test dobré shody

Title: Distribution of interpoint distances

Author: Šárka Horská

Department: Department of Probability and Mathematical Statistics

Supervisor: doc. RNDr. Zdeněk Hlávka, Ph.D., department of Probability and Mathematical Statistics

Abstract: This thesis investigates basic properties of the interpoint distances between random vectors drawn from multinomial distribution. We also describe a possible application to testing sparse observations, i.e., a setup with small number of observations and large number of categories, where the classical $\chi^2$-test cannot be recommended. As an alternative, utilizing the multinomial interpoint distances, we will present the test statistic proposed by Biswas and Ghosh (2014).

Keywords: Euclidean distance; multinomial distribution; goodness-of-fit test

# Obsah

# Úvod

Chceme-li provést dvouvýběrový test pro kategoriální data, většinou použijeme některý z $\chi^2$-testů. Podle odborné literatury, např. článku Biswas a Ghosh (2014) nebo Modarres (2018), $\chi^2$-testy nejsou vhodné pro řídká data, kdy máme u nezanedbatelného množství kategorií nulový výskyt. Na testování řídkých dat o velkém počtu kategorií jsou vhodnější testy založené na vzdálenostech mezi body. V této práci se tedy budeme zabývat rozdělením vzdálenosti mezi body a testy, které s ním souvisí.

Uvažujme dva nezávislé náhodné výběry $\boldsymbol{X} = \{\boldsymbol{X}_i\}$ z rozdělení $F_X$ a $\boldsymbol{Y} = \{\boldsymbol{Y}_j\}$ z rozdělení $F_Y$ s rozsahy $N_X$ a $N_Y$. Nechť $\|\cdot\|$ značí Eukleidovskou normu. Pak definujme $D_{F_X}$ jako rozdělení náhodného vektoru $(\|\boldsymbol{X}_1 - \boldsymbol{X}_2\|^2, \|\boldsymbol{X}_1 - \boldsymbol{Y}_1\|^2)'$ a $D_{F_Y}$ jako rozdělení náhodného vektoru $(\|\boldsymbol{X}_1 - \boldsymbol{Y}_1\|^2, \|\boldsymbol{Y}_1 - \boldsymbol{Y}_2\|^2)'$. Střední hodnoty vektorů z $D_{F_X}$ a $D_{F_Y}$ budeme značit $\boldsymbol{\mu}_{D_{F_X}}$ a $\boldsymbol{\mu}_{D_{F_Y}}$.

Hypotéza $H_0 : F_X = F_Y$ proti alternativě $H_1 : F_X \neq F_Y$ je ekvivalentní s hypotézou $H_0^* : \mu_{D_{F_X}} = \mu_{D_{F_Y}}$ proti alternativě $H_1^* : \mu_{D_{F_X}} \neq \mu_{D_{F_Y}}$ (důkaz viz. Biswas a Ghosh (2014, str. 169)). Ke konstrukci testové statistiky si zadefinujme:

$$\hat{\boldsymbol{\mu}}_{D_{F_X}} = \left( \frac{\sum_{i=1}^{N_X} \sum_{j=i+1}^{N_X} \|\boldsymbol{X}_i - \boldsymbol{X}_j\|^2}{\binom{N_X}{2}}, \frac{\sum_{i=1}^{N_X} \sum_{j=1}^{N_Y} \|\boldsymbol{X}_i - \boldsymbol{Y}_j\|^2}{N_X N_Y} \right),$$

$$\hat{\boldsymbol{\mu}}_{D_{F_Y}} = \left( \frac{\sum_{i=1}^{N_X} \sum_{j=1}^{N_Y} \|\boldsymbol{X}_i - \boldsymbol{Y}_j\|^2}{N_X N_Y}, \frac{\sum_{i=1}^{N_Y} \sum_{j=i+1}^{N_Y} \|\boldsymbol{Y}_i - \boldsymbol{Y}_j\|^2}{\binom{N_Y}{2}} \right).$$

Tyto vektory jsou nestranné a konzistentní odhady $\boldsymbol{\mu}_{D_{F_X}}$ a $\boldsymbol{\mu}_{D_{F_Y}}$. Nyní můžeme definovat testovou statistiku (Biswas a Ghosh (2014)):

$$BG = \|\hat{\boldsymbol{\mu}}_{D_{F_X}} - \hat{\boldsymbol{\mu}}_{D_{F_Y}}\|.$$

Testování řídkých dat je tedy jedním z důvodů, proč bychom se měli o rozdělení vzdálenosti bodů zajímat. Ačkoliv se kritická hodnota testové statistiky $BG$ běžně počítá permutační metodou, neměli bychom být lhostejní k vlastnostem testových statistik, které používáme. Testová statistika $BG$ je však značně neprůhledná a její střední hodnota a rozptyl velmi špatně spočitatelné. Proto se zaměříme alespoň na náhodné vektory $\hat{\boldsymbol{\mu}}_{D_{F_X}}$ a $\hat{\boldsymbol{\mu}}_{D_{F_Y}}$, které nás přiblíží k poznání testové statistiky $BG$.

Celá bakalářská práce je rozdělena do čtyř kapitol. V první kapitole se seznámíme s multinomickým rozdělením a jeho vlastnostmi, protože v celé práci budeme uvažovat náhodné výběry právě z tohoto rozdělení. Stěžejní kapitolou celé práce je kapitola druhá, ve které se budeme věnovat rozdělení vzdálenosti mezi body. Ve třetí kapitole shrneme dosavadní poznatky a najdeme rozptyly a střední hodnoty složek náhodných vektorů $\hat{\boldsymbol{\mu}}_{D_{F_X}}$ a $\hat{\boldsymbol{\mu}}_{D_{F_Y}}$ za platnosti hypotézy $H_0$. Provedeme simulaci a vše vykreslíme do grafů. V poslední čtvrté kapitole budeme testovat hypotézu $H_0$ za pomoci testové statistiky $BG$.

# 1. Multinomické rozdělení

Pracujeme-li s náhodným výběrem z multinomického rozdělení, měli bychom se nejprve s tímto rozdělením řádně seznámit. Kromě základních vlastností multinomického rozdělení spočítáme také první čtyři momenty složek z náhodného vektoru, které využijeme později.

**Definice 1** (Omelka, 2020, str. 132)**.** *Nechť $n, k \in \mathbb{N}$, $n \geq 1$ a $K \geq 2$. Dále nechť $\boldsymbol{p} = (p_1, \ldots, p_K)^T$ je vektor konstant splňující $\sum_{k=1}^{K} p_k = 1$ a $p_k > 0$ pro $k \in \{1, \ldots, K\}$. Řekneme, že náhodný vektor $\boldsymbol{X} = (X_1, \ldots, X_K)^T$ má multinomické rozdělení $Mult_K(n, \boldsymbol{p})$, právě když jeho hustota vzhledem k součinové čítací míře na $\mathbb{Z}^K$ je*

$$P[X_1 = x_1, X_2 = x_2, \ldots, X_K = x_K] = \begin{cases} \frac{n!}{x_1! \cdots x_K!} \cdot p_1^{x_1} \cdots p_K^{x_K} & \sum_{k=1}^{K} x_k = n \\ & x_k \in N_0 \ \forall k, \\ 0 & jinak. \end{cases}$$

*Značíme $\boldsymbol{X} \sim Mult_K(n, \boldsymbol{p})$.*

Multinomické rozdělení si můžeme představit následovně. Máme $K$ přihrádek a $n$ nezávislých pokusů. Každý pokus musí nutně skončit právě v jedné z $K$ přihrádek. Pravděpodobnost, že daný pokus skončí v $k$-té přihrádce je $p_k$. Náhodná veličina $X_k$ tedy značí počet pokusů, které skončily v $k$-té přihrádce.

Je snadné nahlédnout, že binomické rozdělení je speciálním případem multinomického pro $K = 2$.

## 1.1 Vlastnosti multinomického rozdělení

**Věta 1** (Vlastnosti multinomického rozdělení)**.** *Nechť $\boldsymbol{X} \sim Mult_K(n, \boldsymbol{p})$. Potom má náhodný vektor $\boldsymbol{X}$ následující vlastnosti:*

1. *jednotlivé složky náhodného vektoru $\boldsymbol{X}$ jsou vzájemně závislé,*
2. *$X_k \sim Bi(n, p_k)$ pro $k \in \{1, \ldots, K\}$,*
3. *$E\boldsymbol{X} = n\boldsymbol{p}$,*
4. *$Var\boldsymbol{X} = n[diag(\boldsymbol{p}) - \boldsymbol{p}^{\otimes 2}]$*

$$= \begin{pmatrix} np_1(1-p_1) & -np_1p_2 & \cdots & -np_1p_K \\ -np_2p_1 & np_2(1-p_2) & \cdots & -np_1p_K \\ \vdots & \vdots & \ddots & \vdots \\ -np_Kp_1 & -np_Kp_2 & \cdots & np_K(1-p_K) \end{pmatrix}.$$

*Důkaz.* Závislost složek vektoru plyne z definice multinomického rozdělení, konkrétně z rovnosti $\sum_{k=1}^{K} X_{ik} = n$.

Náhodnou veličinu $\boldsymbol{X}$ lze dle skript Omelka (2020, Věta 8.1) rozepsat do tvaru $\sum_{i=1}^{n} \boldsymbol{Y}_i$, kde $\boldsymbol{Y}_i \sim Mult_K(1, \boldsymbol{p})$. Pak platí rovnost $X_k = \sum_{i=1}^{n} Y_{ik}$, kde

$Y_{ik} \sim Alt(p_k)$, a odtud plyne $X_k \sim Bi(n, p_k)$. Potom je střední hodnota $EX_k = np_k$ a tedy $E\boldsymbol{X} = n\boldsymbol{p}$. Variační matice má na diagonále prvky $var(X_k)$ a mimo ní prvky $cov(X_k,X_l)$. Jelikož $X_k \sim Bi(n, p_k)$, známe $var(X_k) = np_k(1 - p_k)$. Pro výpočet kovariance si rozepišme

$$\begin{aligned} Cov(X_k,X_l) &= Cov(\sum_{i=1}^{n} Y_{ik}, \sum_{j=1}^{n} Y_{jl}) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} Cov(Y_{ik}, Y_{jl}) \\ &= \sum_{i=1}^{n} Cov(Y_{ik}, Y_{il}) \\ &= \sum_{i=1}^{n} E[Y_{ik}Y_{il}] - EY_{ik}EY_{il} \\ &= -np_kp_l. \end{aligned}$$

Ve výpočtu využíváme, že kovariance nezávislých náhodných veličin je nulová. Také $E[Y_{ik}Y_{il}] = 0$, protože $i$-tý pokus nemůže skončit v $k$-té a v $l$-té kategorii zároveň.

□

Zaveďme si značení pro rozptyl a kovarianci složek náhodného vektoru $\boldsymbol{X} \sim Mult_K(n, \boldsymbol{p})$. Pro $k,l \in \{1, \ldots, K\}$ budeme značit:

$$\begin{aligned} Var(X_k) &= \sigma_k, \\ Cov(X_k, X_l) &= \sigma_{kl}. \end{aligned}$$

**Lemma 2.** *Nechť $\boldsymbol{X} = (X_1, \ldots, X_K)^T$ je náhodný vektor z rozdělení $Mult_K(n, \boldsymbol{p})$ a $k \in \{1, \ldots, K\}$. Potom jsou první čtyři momenty náhodné veličiny $X_k$ tvaru*

$$\begin{aligned} &EX_k = np_k, \\ &EX_k^2 = np_k + n(n-1)p_k^2, \\ &EX_k^3 = np_k + 3n(n-1)p_k^2 + n(n-1)(n-2)p_k^3, \\ &EX_k^4 = np_k + 7n(n-1)p_k^2 + 6n(n-1)(n-2)p_k^3 + n(n-1)(n-2)(n-3)p_k^4. \end{aligned}$$

*Důkaz.* Momenty spočítáme pomocí momentové vytvořující funkce $M_{X_k}(t)$, pro kterou platí

$$E[X_k^a] = \frac{d^a M_{X_k}}{dt^a}(0).$$

Náhodná veličina $X_k \sim Bi(n, p_k)$ má momentovou vytvořující funkci

$$M_{X_k}(t) \;=\; (1 - \; p_k + \; p_k \; e^t)^n.$$

Hledané derivace snadno nalezneme pomocí Taylorova rozvoje v bodě nula, protože tvar Taylorova rozvoje funkce f v bodě a je:

$$f(a) + f'(a)(x-a) + \frac{f''(a)}{2}(x-a)^2 + \frac{f^{(3)}(a)}{3!}(x-a)^3 + o(x^4).$$

Taylorův rozvoj funkce $M_{X_k}(t)$ v bodě nula, ze kterého jsou ihned vidět momenty náhodné veličiny $X_k$:

$$1 + np_k t + \frac{1}{2!} np_k t^2[(n-1)p_k + 1] + \frac{1}{3!} np_k t^3[(n-2)(n-1)p_k^2 + 3(n-1)p_k + 1]$$
$$+ \frac{1}{4!} np_k t^4[(n-3)(n-2)(n-1)p_k^3 + 6(n-2)(n-1)p_k^2 + 7(n-1)p_k + 1] + o(t^5)$$

□

# 2. Vzdálenosti bodů

V této kapitole si zadefinujeme vzdálenost dvou náhodných vektorů z multinomického rozdělení, odvodíme její pravděpodobnostní rozdělení a základní vlastnosti. Budu postupovat podle článku Modarres (2018) a uvádět důkazy, kterým v článku nebyla věnována pozornost. Používám značení již zavedené v kapitole 1.

## 2.1 Rozdělení vzdálenosti bodů

Předpokládejme náhodný výběr vektorů $\{\boldsymbol{X}_i\}_{i=1}^N$ z multinomického rozdělení $Mult_K(n, \boldsymbol{p})$. Čtvercovou vzdáleností mezi $\boldsymbol{X}_i$ a $\boldsymbol{X}_j$ rozumíme

$$d^2_{(x)ij} = \sum_{k=1}^{K} T^2_{(x)ij}(k),$$

kde $T_{(x)ij}(k) = X_{ik} - X_{jk}$ pro $i,j \in \{1, \dots, N\}$.

Funkce $d^2_{(x)ij}$ nabývá přirozených hodnot a je nulová právě tehdy, když $\boldsymbol{X}_i = \boldsymbol{X}_j$. Maximální možná hodnota $d^2_{(x)ij}$ je $2n^2$. To lze nahlédnout následovně:

$$d^2_{(x)ij} = \sum_{k=1}^{K}(X_{ik} - X_{jk})^2 \leq \sum_{k=1}^{K}(X^2_{ik} + X^2_{jk}) \leq (\sum_{k=1}^{K} X_{ik})^2 + (\sum_{k=1}^{K} X_{jk})^2 = 2n^2$$

Taková situace nastane, pokud existují různá $k_1, k_2 \in \{1, \dots, K\}$ taková, že $X_{ik_1} = X_{jk_2} = n$. Z definice multinomického rozdělení vyplývá, že všechny ostatní složky obou náhodných vektorů jsou nulové. Pro takové vektory $\boldsymbol{X}_i$ a $\boldsymbol{X}_j$ nabývá $d^2_{(x)ij}$ hodnoty $2n^2$.

Vzdálenost bodů $\boldsymbol{X}_i$ a $\boldsymbol{X}_j$ ale není libovolná hodnota z množiny $\{0, \dots, 2n^2\}$, je součtem čtverců hodnot $T_{(x)ij}(k)$. Zadefinujme si proto množinu

$$\boldsymbol{I} = \{t = t_1^2 + \cdots + t_K^2 : t_i \in \{0, \dots, n\} \text{ pro } i \in \{1, \dots, K\}\}.$$

Nyní si postupně odvodíme rozdělení pro náhodné veličiny $T_{(x)ij}(k)$, $T^2_{(x)ij}(k)$ a $d^2_{(x)ij}$.

**Lemma 3.** *Rozdělení náhodné veličiny $T_{(x)ij}(k)$ je dáno vztahem*

$$P[T_{(x)ij}(k) = t_k] = \sum_{m=0}^{n} \binom{n}{t_k + m} p_k^{t_k+m}(1-p_k)^{n-(t_k+m)} \binom{n}{m} p_k^m (1-p_k)^{n-m}.$$

*Důkaz.* Rozepišme si

$$\begin{aligned}
P[T_{(x)ij}(k) = t_k] &= P[X_{ik} - X_{jk} = (t_k + m) - m)] \\
&= \sum_{m=0}^{n} P[X_{it} = t_k + m, X_{jt} = m] \\
&\overset{iid}{=} \sum_{m=0}^{n} P[X_{it} = t_k + m] P[X_{jt} = m] \\
&= \sum_{m=0}^{n} \binom{n}{t_k + m} p_k^{t_k+m}(1-p_k)^{n-(t_k+m)} \binom{n}{m} p_k^m (1-p_k)^{n-m}.
\end{aligned}$$

□

**Lemma 4.** *Rozdělení náhodné veličiny* $T^2_{(x)ij}(k)$ *je dáno vztahem*

$$P[T^2_{(x)ij}(k) = t_k^2] = (1+\delta_k) \sum_{m=0}^{n} \binom{n}{t_k+m} p_k^{t_k+m} (1-p_k)^{n-(t_k+m)} \binom{n}{m} p_k^m (1-p_k)^{n-m},$$

*kde* $\delta_k = 0$ *pro* $t_k = 0$, *jinak* $\delta_k = 1$.

*Důkaz.* Rozepišme si $T^2_{(x)ij}(k)$ pro dva případy

$$P[T^2_{(x)ij}(k) = t_k^2] = \begin{cases} P[T_{(x)ij} = 0] \text{ pro } t_k = 0, \\ P[T_{(x)ij} = t_k] + P[T_{(x)ij} = -t_k] = 2 \cdot P[T_{(x)ij} = t_k] \text{ jinak.} \end{cases}$$

Jelikož $P[T^2_{(x)ij}(k) = t_k^2] = (1 + \delta_k) \cdot P[T_{(x)ij}(k) = t_k]$, z lemmatu 3 pak plyne tvrzení.

□

**Věta 5** (O rozdělení vzdálenosti bodů)**.** *Nechť* $\{\boldsymbol{X}_i\}$ *je náhodný výběr z multinomického rozdělení* $Mult_K(n, \boldsymbol{p})$ *o rozsahu* $N_X$. *Označme množinu*

$$\boldsymbol{I} = \{t = t_1^2 + \ldots + t_K^2 : t_i \in \{0, \ldots, n\} \text{ pro } i \in \{1, \ldots, K\}\}.$$

*Potom rozdělení* $P[d^2_{(x)ij} = t]$ *náhodné veličiny* $d^2_{(x)ij}$ *je dáno vztahem*

$$\sum_{t \in \boldsymbol{I}} \prod_{k=1}^{K} (1+\delta_k) \sum_{m=0}^{n} \binom{n}{t_k+m} p_k^{t_k+m} (1-p_k)^{n-(t_k+m)} \binom{n}{m} p_k^m (1-p_k)^{n-m},$$

*kde* $\delta_k = 0$ *pro* $t_k = 0$, *jinak* $\delta_k = 1$ .

*Důkaz.* Jednotlivé kroky důkazu obsahuje článek Modarres (2018, str. 346) a využívá se v něm lemma 4.

□

## 2.2 Střední hodnota, rozptyl a kovariance

Nyní spočítáme střední hodnotu a rozptyl náhodné veličiny $d^2_{(x)ij}$. Budeme postupovat opět podle článku Modarres (2018). V případě kovariance naopak ukáži, proč si myslím, že je kovariance v článku nesprávně.

**Věta 6** (O střední hodnotě)**.** *Nechť $\boldsymbol{X}_i, \boldsymbol{X}_j$ jsou nezávislé náhodné vektory z rozdělení $Mult_K(n, \boldsymbol{p})$. Potom je střední hodnota vzdálenosti bodů $d^2_{(x)ij}$ dána vztahem:*

$$E[d^2_{(x)ij}] = 2\sum_{k=1}^{K} \sigma_k^2.$$

*Důkaz.* Výraz si rozepíšeme

$$E[d^2_{(x)ij}] = E\sum_{k=1}^{K}(X_{ik} - X_{jk})^2 = \sum_{k=1}^{K} EX_{ik}^2 - 2EX_{ik}X_{jk} + EX_{jk}^2 =$$

$$\overset{iid}{=} 2\sum_{k=1}^{K} EX_{ik}^2 - EX_{ik}X_{jk} \overset{\perp\!\!\!\perp}{=} 2\sum_{k=1}^{K} VarX_{ik} + [EX_{ik}]^2 - EX_{ik}EX_{jk} \overset{iid}{=} 2\sum_{k=1}^{K} \sigma_k^2.$$

V posledních dvou krocích jsem využila vzorce pro výpočet rozptylu $VarX_{ik} = EX_{ik}^2 - [EX_{ik}]^2$ a nezávislosti náhodných veličin $X_{ik}$ a $X_{jk}$, které mají stejné rozdělení.

□

Následující tři lemmata potřebuji k výpočtu rozptylu $d^2_{(x)ij}$.

**Lemma 7.** *Nechť $\boldsymbol{X}_i, \boldsymbol{X}_j$ jsou nezávislé náhodné vektory z rozdělení $Mult_K(n, \boldsymbol{p})$ a $k, l \in \{1, \ldots, K\}$. Potom platí následující rovnost:*

$$Cov\left[X_{ik}X_{jk}, X_{il}X_{jl}\right] = \sigma_{kl}\, np_kp_l\,(2n-1).$$

*Důkaz.* Výraz si v několika krocích rozepíšeme a použijeme větu 1.

$$\begin{aligned}
Cov\left[X_{ik}X_{jk}, X_{il}X_{jl}\right] &= E[X_{ik}X_{jk} \cdot X_{il}X_{jl}] - E[X_{ik}X_{jk}] \cdot E[X_{il}X_{jl}] \\
&\overset{\perp\!\!\!\perp}{=} E[X_{ik}X_{il}] \cdot E[X_{jk}X_{jl}] - EX_{ik} \cdot EX_{il} \cdot EX_{jk} \cdot EX_{jl} \\
&\overset{iid}{=} (E[X_{ik}X_{il}])^2 - (EX_{ik} \cdot EX_{il})^2 \\
&= (Cov\left[X_{ik}, X_{il}\right] + EX_{ik} \cdot EX_{il})^2 - (EX_{ik} \cdot EX_{il})^2 \\
&= (Cov\left[X_{ik}, X_{il}\right])^2 + 2\,Cov\left[X_{ik}, X_{il}\right] \cdot EX_{ik} \cdot EX_{il} \\
&= \sigma_{kl}(-np_kp_l + 2n^2p_kp_l) \\
&= \sigma_{kl}\, np_kp_l\,(2n-1)
\end{aligned}$$

Čtvrtá rovnost plyne ze vztahu $Cov[X_{ik},X_{il}] = E[X_{ik} \cdot X_{il}] - E[X_{ik}] \cdot E[X_{il}]$.

□

**Lemma 8.** *Nechť $\boldsymbol{X}_i, \boldsymbol{X}_j$ jsou nezávislé náhodné vektory z rozdělení $Mult_K(n, \boldsymbol{p})$ a $k, l \in \{1, \ldots, K\}$. Potom platí:*

$$Cov\,[X_{ik}^2, X_{il}^2] = \sigma_{kl} \cdot [1 + (2n-2)\,(p_k + p_l) + (4n^2 - 10n + 6)\,p_k p_l].$$

*Důkaz.* Jelikož jsou náhodné veličiny $X_{ik}$ a $X_{il}$ závislé, kovariance nejspíš nebude nulová. Výraz si tedy rozepíšeme a spočteme kažkou složku zvlášť.

$$Cov\,[X_{ik}^2, X_{il}^2] = E[X_{ik}^2 \cdot X_{il}^2] - EX_{ik}^2 \cdot EX_{il}^2.$$

$$\begin{aligned} EX_{ik}^2 \cdot EX_{il}^2 &\overset{2}{=} [n\,p_k + n(n-1)\,p_k^2] \cdot [n\,p_l + n(n-1)\,p_l^2] \\ &= np_kp_l[n + n(n-1)p_k + n(n-1)p_l + n(n-1)^2 p_k p_l] \\ &= -\,\sigma_{kl} \cdot [n + (n^2 - n)(p_k + p_l) + (n^3 - 2n^2 + n)p_k p_l] \end{aligned}$$

Pro výpočet $E[X_{ik}^2 \cdot X_{il}^2]$ definujme náhodnou veličinu $Y_{km} \sim Alt(p_k)$ pro $k \in \{1, \cdots, K\}, m \in \{1, \cdots, n\}$. Náhodná veličina nabývá hodnoty 1 právě tehdy, když $m$-tý pokus skončí v $k$-té kategorii. Náhodné veličiny $X_{ik}$ a $X_{il}$ si tedy rozepíšeme jako součet

$$X_{ik} = \sum_{m=1}^{n} Y_{km}\ , \qquad\qquad X_{il} = \sum_{j=1}^{n} Y_{lj}.$$

Pokud $m \neq j$, pak jsou náhodné veličiny $Y_{km}$ a $Y_{lj}$ nezávislé.

$$\begin{aligned} E[X_{ik}^2 \cdot X_{il}^2] &= E[(\sum_{m=1}^{n} Y_{km})^2 \cdot (\sum_{j=1}^{n} Y_{lj})^2] \\ &= E[(\sum_{m=1}^{n} Y_{km}^2 + \sum_{m=1}^{n}\sum_{\substack{t=1\\t\neq m}}^{n} Y_{km}Y_{kt}) \cdot (\sum_{j=1}^{n} Y_{lj}^2 + \sum_{j=1}^{n}\sum_{\substack{h=1\\h\neq j}}^{n} Y_{lj}Y_{lh})] \\ &= E[\sum_{m=1}^{n} Y_{km}^2 \cdot \sum_{j=1}^{n} Y_{lj}^2] + E[\sum_{m=1}^{n} Y_{km}^2 \cdot \sum_{j=1}^{n}\sum_{\substack{h=1\\h\neq j}}^{n} Y_{lj}Y_{lh}] \\ &\quad + E[\sum_{m=1}^{n}\sum_{\substack{t=1\\t\neq m}}^{n} Y_{km}Y_{kt} \cdot \sum_{j=1}^{n} Y_{lj}^2] + E[\sum_{m=1}^{n}\sum_{\substack{t=1\\t\neq m}}^{n} Y_{km}Y_{kt} \cdot \sum_{j=1}^{n}\sum_{\substack{h=1\\h\neq j}}^{n} Y_{lj}Y_{lh}] \\ &= \sum_{m=1}^{n}\sum_{j=1}^{n} E\,[\,Y_{km}^2 \cdot Y_{lj}^2\,] + \sum_{m=1}^{n}\sum_{j=1}^{n}\sum_{\substack{h=1\\h\neq j}}^{n} E\,[\,Y_{km}^2 \cdot Y_{lj}Y_{lh}\,] \\ &\quad + \sum_{m=1}^{n}\sum_{\substack{t=1\\t\neq m}}^{n}\sum_{j=1}^{n} E\,[\,Y_{km}Y_{kt} \cdot Y_{lj}^2\,] + \sum_{m=1}^{n}\sum_{\substack{t=1\\t\neq m}}^{n}\sum_{j=1}^{n}\sum_{\substack{h=1\\h\neq j}}^{n} E\,[\,Y_{km}Y_{kt} \cdot Y_{lj}Y_{lh}\,] \end{aligned}$$

Všechny střední hodnoty součinů závislých náhodných veličin jsou nulové, protože nemůže $m$-tý pokus skončit v $k$-té i $l$-té kategorii zároveň. Tyto střední hodnoty ze součtů tedy můžeme vyřadit. Jelikož nám pak zůstanou pouze střední hodnoty

součinů nezávislých náhodných veličin, můžeme je převést na součin středních hodnot.

$$E[X_{ik}^2 \cdot X_{il}^2] = \sum_{m=1}^{n} \sum_{\substack{j=1 \\ j\neq m}}^{n} EY_{km}^2 \cdot EY_{lj}^2 + \sum_{m=1}^{n} \sum_{\substack{j=1 \\ j\neq m}}^{n} \sum_{\substack{h=1 \\ h\neq j \\ h\neq m}}^{n} EY_{km}^2 \cdot EY_{lj} EY_{lh}$$
$$+ \sum_{m=1}^{n} \sum_{\substack{t=1 \\ t\neq m}}^{n} \sum_{\substack{j=1 \\ j\neq m \\ j\neq t}}^{n} EY_{km} EY_{kt} \cdot EY_{lj}^2 + \sum_{m=1}^{n} \sum_{\substack{t=1 \\ t\neq m}}^{n} \sum_{\substack{j=1 \\ j\neq m \\ j\neq t}}^{n} \sum_{\substack{h=1 \\ h\neq j \\ h\neq m \\ h\neq t}}^{n} EY_{km} EY_{kt} \cdot EY_{lj} EY_{lh}$$

Jelikož $Y_{km} \sim Alt(p_k)$, víme, že všechny momenty této náhodné veličiny jsou rovny $p_k$.

$$\begin{aligned} E[X_{ik}^2 \cdot X_{il}^2] &= n(n-1)p_k p_l + n(n-1)(n-2)p_k p_l^2 + n(n-1)(n-2)p_k^2 p_l \\ &\quad + n(n-1)(n-2)(n-3)p_k^2 p_l^2 \\ &= np_k p_l \cdot [n-1+(n-1)(n-2)p_l + (n-1)(n-2)p_k \\ &\qquad + (n-1)(n-2)(n-3)p_k p_l] \\ &= -\sigma_{kl} \cdot [n-1+(n^2-3n+2)(p_k+p_l) + (n^3-6n^2+11n-6)\, p_k p_l] \end{aligned}$$

Stačí už jen dosadit

$$\begin{aligned} Cov\,[X_{ik}^2, X_{il}^2] &= E[X_{ik}^2 \cdot X_{il}^2] - EX_{ik}^2 \cdot EX_{il}^2 \\ &= -\sigma_{kl} \cdot [-1 + (-2n+2)\,(p_k+p_l) + (-4n^2+10n-6)\, p_k p_l] \\ &= \sigma_{kl} \cdot [1 + (2n-2)\,(p_k+p_l) + (4n^2-10n+6)\, p_k p_l] \end{aligned}$$

□

**Lemma 9.** *Ať $\boldsymbol{X}_i, \boldsymbol{X}_j$ jsou nezávislé náhodné vektory z rozdělení $Mult_K(n, \boldsymbol{p})$ a $k, l \in \{1, \ldots, K\}$. Potom platí:*

$$Cov\,[X_{ik}^2, X_{il} X_{jl}] = \sigma_{kl} \cdot [np_l + (2n^2 - 2n) p_k p_l].$$

*Důkaz.* Jelikož je náhodná veličina $X_{jl}$ nezávislá s náhodnými veličinami $X_{ik}$ a $X_{il}$, můžeme ji z kovariance vyjmout následujícím způsobem.

$$\begin{aligned} Cov\,[X_{ik}^2, X_{il} X_{jl}] &= E\,[\,X_{ik}^2 \cdot X_{il} X_{jl}] - EX_{ik}^2 \cdot E\,[\,X_{il} X_{jl}] \\ &\overset{\perp\!\!\!\perp}{=} E\,[\,X_{ik}^2 \cdot X_{il}] \cdot EX_{jl} - EX_{ik}^2 \cdot EX_{il} \cdot EX_{jl} \\ &= Cov\,[X_{ik}^2, X_{il}] \cdot EX_{jl} \end{aligned}$$

Pro výpočet kovariance si náhodné veličiny rozepíšeme jako v důkazu lemmatu 8.

$$X_{ik} = \sum_{m=1}^{n} Y_{km}\,, \qquad\qquad X_{il} = \sum_{j=1}^{n} Y_{lj},$$

kde $Y_{km} \sim Alt(p_k)$ pro $k \in \{1, \cdots, K\}, m \in \{1, \cdots, n\}$. Tedy náhodná veličina $Y_{km}$ nabývá hodnoty 1 právě tehdy, když $m$-tý pokus skončí v $k$-té kategorii.

Náhodné veličiny $Y_{km}$ a $Y_{lj}$ jsou závislé, pokud $m = j$.

$$\begin{aligned}
Cov\,[X_{ik}^2, X_{il}] &= Cov\,[(\sum_{m=1}^{n} Y_{km})^2\,,\, \sum_{j=1}^{n} Y_{lj}] \\
&= Cov\,[\sum_{m=1}^{n} Y_{km}^2 + \sum_{m=1}^{n}\sum_{\substack{t=1\\t\neq m}}^{n} Y_{km}Y_{kt}\,,\, \sum_{j=1}^{n} Y_{lj}] \\
&= Cov\,[\sum_{m=1}^{n} Y_{km}^2\,,\, \sum_{j=1}^{n} Y_{lj}] + Cov\,[\sum_{m=1}^{n}\sum_{\substack{t=1\\t\neq m}}^{n} Y_{km}Y_{kt}\,,\, \sum_{j=1}^{n} Y_{lj}] \\
&= \sum_{m=1}^{n}\sum_{j=1}^{n} Cov\,[Y_{km}^2\,,\, Y_{lj}] + \sum_{m=1}^{n}\sum_{\substack{t=1\\t\neq m}}^{n}\sum_{j=1}^{n} Cov\,[Y_{km}Y_{kt}\,,\, Y_{lj}]
\end{aligned}$$

Víme, že kovariance nezávislých náhodných veličin je rovna nule.

$$\begin{aligned}
Cov\,[X_{ik}^2, X_{il}] &= \sum_{m=1}^{n} Cov\,[Y_{km}^2\,,\, Y_{lm}] + 2\sum_{m=1}^{n}\sum_{\substack{t=1\\t\neq m}}^{n} Cov\,[Y_{km}Y_{kt}\,,\, Y_{lm}] \\
&= \sum_{m=1}^{n} E\,[Y_{km}^2 \cdot Y_{lm}] - EY_{km}^2 \cdot EY_{lm} \\
&\quad + 2\sum_{m=1}^{n}\sum_{\substack{t=1\\t\neq m}}^{n} E\,[Y_{km}Y_{kt} \cdot Y_{lm}] - E\,[Y_{km}Y_{kt}] \cdot EY_{lm} \\
&= -\sum_{m=1}^{n} p_k p_l - 2\sum_{m=1}^{n}\sum_{\substack{t=1\\t\neq m}}^{n} p_k^2 p_l \\
&= -n \cdot p_k p_l - 2n(n-1) \cdot p_k^2 p_l
\end{aligned}$$

Zbývá už jen dosadit.

$$\begin{aligned}
Cov\,[X_{ik}^2, X_{il}X_{jl}] &= Cov\,[X_{ik}^2, X_{il}] \cdot EX_{jl} \\
&= [-n \cdot p_k p_l - 2n(n-1) \cdot p_k^2 p_l] \cdot np_l \\
&= -np_k p_l \cdot [np_l + 2n(n-1)p_k p_l] \\
&= \sigma_{kl} \cdot [np_l + (2n^2 - 2n)p_k p_l]
\end{aligned}$$

□

**Věta 10** (O rozptylu)**.** *Nechť $\boldsymbol{X}_i, \boldsymbol{X}_j$ jsou nezávislé náhodné vektory z rozdělení $Mult_K(n, \boldsymbol{p})$. Potom je rozptyl $d_{(x)ij}^2$ dán vztahem:*

$$Var\; d_{(x)ij}^2 = \sum_{k=1}^{K} Var\, T_{(x)ij}^2(k) + 4\sum_{k=1}^{K}\sum_{l<k} \sigma_{kl} \cdot \phi_{kl},$$

$$\begin{aligned}
&kde\; Var\, T_{(x)ij}^2(k) = 2n\,[p_k + (4n-7)p_k^2 + 4(-2n+3)p_k^3 + 2(2n-3)p_k^4], \\
&\quad \phi_{kl} = 1 - 2(p_k + p_l) + (-4n+6)p_k p_l.
\end{aligned}$$

*Důkaz.* Nejprve si výraz rozepíšeme a rozdělíme na několik dílčích výpočtů:

$$\begin{aligned} Var\, d^2_{(x)ij} &= Var \sum_{k=1}^{K} T^2_{(x)ij}(k) \\ &= \sum_{k=1}^{K} Var\, T^2_{(x)ij}(k) + 2\sum_{k=1}^{K}\sum_{l<k} Cov\,[T^2_{(x)ij}(k), T^2_{(x)ij}(l)]. \end{aligned}$$

Začněme výpočtem rozptylu: $Var\, T^2_{(x)ij}(k) = E\, T^4_{(x)ij}(k) - (E\, T^2_{(x)ij}(k))^2$.

$$\begin{aligned} E\, T^2_{(x)ij}(k) &= E\,[X_{ik} - X_{jk}]^2 = E\, X^2_{ik} - 2E\, X_{ik}X_{jk} + E\, X^2_{jk} \\ &\overset{iid}{=} 2(EX^2_{ik} - (EX_{ik})^2) = 2\, Var\, X_{ik} = 2\, np_k(1-p_k) \\ E\, T^4_{(x)ij}(k) &= E\,(X_{ik} - X_{jk})^4 \\ &= EX^4_{ik} - 4\cdot E[X^3_{ik}X_{jk}] + 6\cdot E[X^2_{ik}X^2_{jk}] - 4\cdot E[X_{ik}X^3_{jk}] + EX^4_{jk} \\ &\overset{iid}{=} 2\cdot EX^4_{ik} - 8\cdot EX^3_{ik}EX_{jk} + 6\cdot (EX^2_{ik})^2 \end{aligned}$$

Hledané momenty dosadíme z lemmatu 2.

$$\begin{aligned} &E\, T^4_{(x)ij}(k) \\ &\quad = 2[np_k + 7n(n-1)p_k^2 + 6n(n-1)(n-2)p_k^3 + n(n-1)(n-2)(n-3)p_k^4] \\ &\qquad - 8[np_k + 3n(n-1)p_k^2 + n(n-1)(n-2)p_k^3][np_k] + 6[np_k + n(n-1)p_k^2] \\ &\quad = 2np_k + (12n^2 - 14n)p_k^2 + (-24n^2 + 24n)p_k^3 + (12n^2 - 12n)p_k^4 \end{aligned}$$

$$\begin{aligned} Var\, T^2_{(x)ij}(k) &= E\, T^4_{(x)ij}(k) - (E\, T^2_{(x)ij}(k))^2 \\ &= 2np_k + (12n^2 - 14n)p_k^2 + (-24n^2 + 24n)p_k^3 + (12n^2 - 12n)p_k^4 \\ &\quad - (2\, np_k(1-p_k))^2 \\ &= 2np_k + (8n^2 - 14n)p_k^2 + (-16n^2 + 24n)p_k^3 + (8n^2 - 12n)p_k^4 \\ &= 2n\,[p_k + (4n-7)p_k^2 + 4(-2n+3)p_k^3 + 2(2n-3)p_k^4] \end{aligned}$$

Nyní se zaměříme na výpočet $Cov\,[T^2_{(x)ij}(k), T^2_{(x)ij}(l)]$. Kovarianci součtu si můžeme rozdělit na součet kovariancí.

$$\begin{aligned} &Cov\,[T^2_{(x)ij}(k), T^2_{(x)ij}(l)] = Cov\,[(X_{ik} - X_{jk})^2, (X_{il} - X_{jl})^2] \\ &\quad = Cov[X^2_{ik},X^2_{il}] + Cov[X^2_{ik}, -2X_{il}X_{jl}] + Cov[X^2_{ik},X^2_{jl}] \\ &\qquad + Cov[-2X_{ik}X_{jk},X^2_{il}] + Cov[-2X_{ik}X_{jk}, -2X_{il}X_{jl}] + Cov[-2X_{ik}X_{jk},X^2_{jl}] \\ &\qquad + Cov[X^2_{jk},X^2_{il}] + Cov[X^2_{jk}, -2X_{il}X_{jl}] + Cov[X^2_{jk},X^2_{jl}] \end{aligned}$$

Rozmyslíme si, které členy jsou nulové a které se opakují. Uvědomme si, že můžeme beztrestně prohodit všechny indexy $i$ za $j$ a zároveň všechny indexy $j$ za $i$ a získáme výraz se stejným rozdělením jako měl výraz původní. To protože máme stejně rozdělené náhodné veličiny $\boldsymbol{X}_i$ a $\boldsymbol{X}_j$.

- Náhodné veličiny $X_{ik}^2$ a $X_{jl}^2$ jsou vzájemně nezávislé, proto

$$Cov\left[X_{ik}^2,X_{jl}^2\right] = Cov\left[X_{jk}^2,X_{il}^2\right] = 0.$$

- Rovné si jsou následující kovariance:

$$\begin{aligned} Cov[X_{ik}^2,X_{il}^2] &= Cov[X_{jk}^2,X_{jl}^2], \\ Cov[X_{ik}^2,-2X_{il}X_{jl}] &= Cov[X_{jk}^2,-2X_{il}X_{jl}], \\ Cov[-2X_{ik}X_{jk},X_{il}^2] &= Cov[-2X_{ik}X_{jk},X_{jl}^2]. \end{aligned}$$

Výpočet se nám značně zkrátil na

$$\begin{aligned} Cov\left[T_{(x)ij}^2(k),T_{(x)ij}^2(l)\right] &= 2\,Cov\left[X_{ik}^2,X_{il}^2\right] + 2\,Cov\left[X_{ik}^2,-2X_{il}X_{jl}\right] \\ &\quad + 2\,Cov\left[-2X_{ik}X_{jk},X_{il}^2\right] + Cov\left[-2X_{ik}X_{jk},-2X_{il}X_{jl}\right] \\ &= 2\,Cov\left[X_{ik}^2,X_{il}^2\right] - 4\,Cov\left[X_{ik}^2,X_{il}X_{jl}\right] \\ &\quad - 4\,Cov\left[X_{ik}X_{jk},X_{il}^2\right] + 4\,Cov\left[X_{ik}X_{jk},X_{il}X_{jl}\right]. \end{aligned}$$

Použijeme lemma 9 a připravíme si následující součet:

$$\begin{aligned} &Cov\left[X_{ik}^2,X_{il}X_{jl}\right] + Cov\left[X_{ik}X_{jk},X_{il}^2\right] = \\ &= \sigma_{kl}\cdot[np_l + (2n^2-2n)p_kp_l + np_k + (2n^2-2n)p_kp_l] \\ &= \sigma_{kl}\cdot[n(p_k+p_l) + (4n^2-4n)p_kp_l]. \end{aligned}$$

Dosadíme výsledky lemmat 7, 8 a 9:

$$\begin{aligned} Cov\left[T_{(x)ij}^2(k),T_{(x)ij}^2(l)\right] &= 2\,\sigma_{kl}\cdot\Big[[1+(2n-2)\,(p_k+p_l) + (4n^2-10n+6)\,p_kp_l] \\ &\quad - 2[n(p_k+p_l) + (4n^2-4n)p_kp_l] \\ &\quad + 2[np_kp_l\,(2n-1)]\Big] \\ &= 2\,\sigma_{kl}\cdot[1-2(p_k+p_l) + (-4n+6)p_kp_l]. \end{aligned}$$

□

**Lemma 11.** *Nechť $\boldsymbol{p} = (\frac{1}{K},\ldots,\frac{1}{K})$ je vektor konstant o $K$-složkách a $\{\boldsymbol{X}_i\}$ je náhodný výběr z multinomického rozdělení $Mult_K(n,\boldsymbol{p})$ o rozsahu $N_X$. Nechť indexy $i,j \in \{1,\ldots,N_X\}$ a platí $i \neq j$. Potom platí:*

$$\begin{aligned} Ed_{(x)ij}^2 &= 2n\,\left(1-\frac{1}{K}\right), \\ Var\,d_{(x)ij}^2 &= 4n(2n-1)\,\left(\frac{1}{K}-\frac{1}{K^2}\right). \end{aligned}$$

*Důkaz.* Výsledky lemmatu plynou přímo z vět 6 a 10.

□

Pro výpočty v kapitole 3 budeme kromě střední hodnoty a rozptylu náhodné veličiny $d^2_{(x)ij}$ potřebovat také její kovarianci. V článku Modarres (2018, str. 349) je uvedená tato kovariance:

$$Cov[d^2_{(x)ij},d^2_{(x)ih}] = 2\sum_{k=1}^{K}\sum_{l=1}^{K}\sigma_{kl}\cdot(1-2p_k)(1-2p_l).$$

Uvažujme speciální případ s pravděpodobnostním vektorem $\boldsymbol{p} = (\frac{1}{K},\ldots,\frac{1}{K})$. Potom můžeme kovarianci s pomocí věty 1 zjednodušit:

$$Cov[d^2_{(x)ij},d^2_{(x)ih}] = -2n(1-2/K)^2.$$

Potom by mělo pro kažkou dvojici parametrů ($K$,$n$) platit, že absolutní hodnota korelace $d^2_{(x)ij}$ a $d^2_{(x)ih}$ je menší rovná jedné, tedy má platit nerovnost:

$$\left|\frac{-2n(1-\frac{2}{K})^2}{4n(2n-1)\left(\frac{1}{K}-\frac{1}{K^2}\right)}\right| \leq 1.$$

Například pro $K = 10$ a $n = 1$ požadovaná nerovnost neplatí. Výpočtem rozptylu i kovariance na reálných datech generovaných z příslušných rozdělení jsem zjistila, že chybný je vzorec pro kovarianci. Pro účely v kapitole 3 teoretickou hodnotu kovariance alespoň odhadneme hodnotou kovariance napočítané z velkého množství dat generovaných z rozdělení $Mult_K(n,\boldsymbol{p})$:

```
#napocitani covariance
n=7
pr.x=rep(1/k, k)
s=100000
d.1=numeric(s)
d.2=numeric(s)
sim.c=numeric(10)
for (q in 1:10) {
        for (t in 1:s) {
                x1=rmultinom(1,n,pr.x)
                x2=rmultinom(1,n,pr.x)
                x3=rmultinom(1,n,pr.x)
                d.1[t]=sum((x1-x2)^2)
                d.2[t]=sum((x1-x3)^2)
        }
        sim.c[q]=cov(d.1,d.2)
}
c.d2x=mean(sim.c)
```

# 3. Simulace

Nyní se podíváme, jak vypadá realizace náhodného vektoru $\hat{\boldsymbol{\mu}}_{D_{F_X}}$, který jsme v úvodu definovali pro testovou statistiku $BG$ jako vektor:

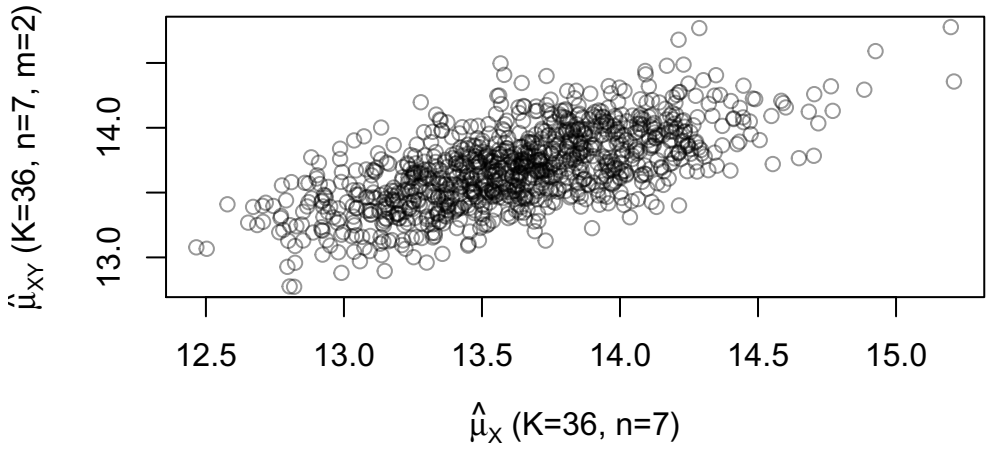
$$\hat{\boldsymbol{\mu}}_{D_{F_X}} = \left( \frac{\sum_{i=1}^{N_X} \sum_{j=i+1}^{N_X} \|\boldsymbol{X}_i - \boldsymbol{X}_j\|^2}{\binom{N_X}{2}}, \frac{\sum_{i=1}^{N_X} \sum_{j=1}^{N_Y} \|\boldsymbol{X}_i - \boldsymbol{Y}_j\|^2}{N_X N_Y} \right)'.$$

Složky náhodného vektoru $\hat{\boldsymbol{\mu}}_{D_{F_X}}$ si označme $\hat{\mu}_X$ a $\hat{\mu}_{XY}$. Nyní provedeme simulaci. Zopakujeme 1000-krát proces, kdy generujeme hodnoty náhodných veličin z rozdělení $Mult_K(n,\boldsymbol{p}_x)$ a $Mult_K(n,\boldsymbol{p}_y)$, vždy s rozsahem 50 pozorování, a z nich napočítáme $\hat{\mu}_X$ a $\hat{\mu}_{XY}$. Vznikne nám tak graf s osami $\hat{\mu}_X$ a $\hat{\mu}_{XY}$. Graf obsahuje 1000 hodnot, které se však často překrývají, proto jsou jednotlivé hodnoty odstínované podle intenzity jejich výskytu.

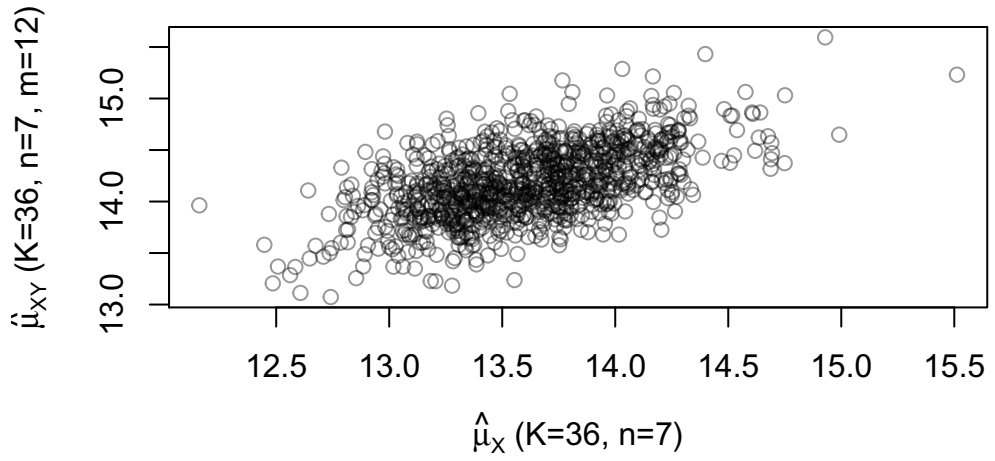
Nyní se zaměřme na to, z jakých rozdělení budeme hodnoty náhodných veličin generovat pro naši simulaci. Pro jednoduchost položme pravděpodobnostní vektor $\boldsymbol{p}_x$ takový, že $p_{(x)k} = 1/K$ pro $k \in \{1, \ldots, K\}$. Pro pevné $m \in \{1, \ldots, K-1\}$ pak uvažujme pravděpodobnostní vektor $\boldsymbol{p}_y$ takový, že

$$p_{(y)k} = \begin{cases} \frac{1}{K-m} & \text{pokud } k \in \{1, \ldots, K-m\}, \\ 0 & \text{pokud } k \in \{K-m+1, \ldots, K\}. \end{cases}$$

Pro srovnání vytvořme grafy dva. Pro oba položme $K = 36$ a $n = 7$. Jediný parametr, ve kterém se budou grafy lišit, je pravděpodobnostní vektor $\boldsymbol{p}_y$. V prvním případě chceme napočítat hodnoty $\hat{\boldsymbol{\mu}}_{D_{F_X}}$ z dat, která pocházejí z podobných rozdělení, proto definujme $\boldsymbol{p}_y$ s $m = 2$. V druhém případě položme $\boldsymbol{p}_y$ s $m = 12$. Získali jsme tak grafy 3.1 a 3.2:




Obrázek 3.1: Graf značí hodnoty $\hat{\boldsymbol{\mu}}_{D_{F_X}}$.




Obrázek 3.2: Graf značí hodnoty $\hat{\boldsymbol{\mu}}_{D_{F_X}}$.

Uvažujme hypotézu, že rozdělení $Mult_K(n,\boldsymbol{p}_x)$ a $Mult_K(n,\boldsymbol{p}_y)$, ze kterých pocházejí náhodné výběry $\{\boldsymbol{X}_i\}$ a $\{\boldsymbol{Y}_j\}$, se rovnají. Potom za hypotézy mají náhodné veličiny $d^2_{(x)ij}$ a $d^2_{(xy)ij}$ stejné střední hodnoty, rozptyly i kovariance. Jelikož $\hat{\mu}_X$ a $\hat{\mu}_{XY}$ jsou průměry náhodných veličin $d^2_{(x)ij}$ a $d^2_{(xy)ij}$, pro jejich střední hodnoty za hypotézy platí:

$$E\hat{\mu}_X = E\hat{\mu}_{XY} = Ed^2_{(x)ij}.$$

Rozepišme si rozptyl náhodné veličiny $\hat{\mu}_X$ za platnosti hypotézy:

$$\begin{aligned} Var\,\hat{\mu}_X &= \binom{N_X}{2}^{-2} \sum_{i=1}^{N_X}\sum_{j=i+1}^{N_X}\sum_{k=1}^{N_X}\sum_{h=k+1}^{N_X} Cov\left[\|\boldsymbol{X}_i-\boldsymbol{X}_j\|^2, \|\boldsymbol{X}_k-\boldsymbol{X}_h\|^2\right] \\ &= \binom{N_X}{2}^{-2} \sum_{i=1}^{N_X}\sum_{j=i+1}^{N_X}\sum_{k=1}^{N_X}\sum_{h=k+1}^{N_X} Cov\left[d^2_{(x)ij}, d^2_{(x)kh}\right] \end{aligned}$$

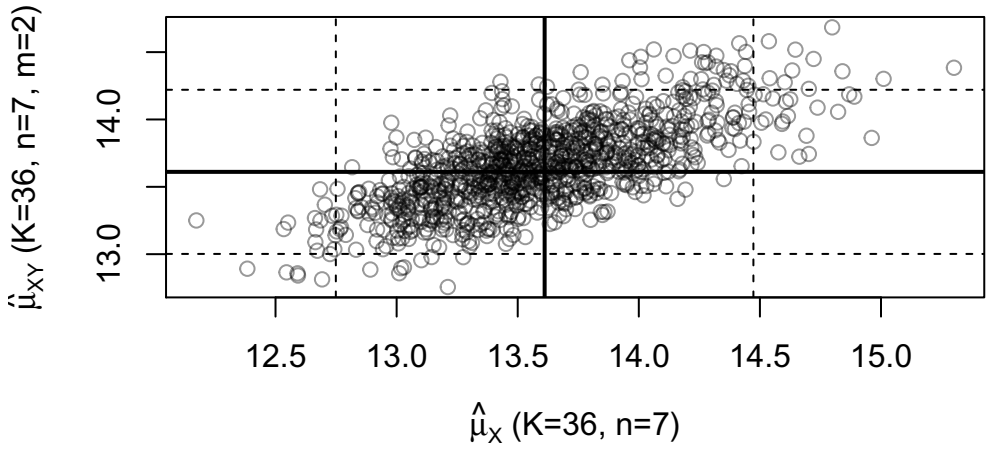
Nyní je zapotřebí spočítat, kolik suma přes indexy $i$,$j$,$k$,$h$ obsahuje nenulových kovariancí. Výraz $Cov\left[d^2_{(x)ij}, d^2_{(x)kh}\right]$ je rozptylem, pokud $i = k$ a $j = h$. Případ, kdy platí $i = h$ a $j = k$, nemůže nastat, protože vede ke sporu $h = i < j = k$ s nerovností $k < h$. Máme tedy celkem $\binom{N_X}{2}$ rozptylů $d^2_{(x)ij}$. Počet nenulových kovariancí je roven počtu případů, kdy mezi indexy nastává právě jedna rovnost. Navíc musí být splněny podmínky $i < j$ a $k < h$. Výpočet je tedy celkem nepřehledný a snadno bychom mohli něco opomenout nebo napočítat vícekrát. Proto nenulové kovariance napočítáme raději v Rstudiu:

```
n.x=50              #rozsah nahodneho vyberu
v=0                 #v je pocitadlo, znaci pocet var a cov
for (i in 1:n.x) {for (j in 1:n.x) {if(j>i) {
        for (k in 1:n.x) {for (h in 1:n.x) {if(h>k){
                if((i==k)|(i==h)|(j==k)|(j==h)){v=v+1} }}} }}}
```

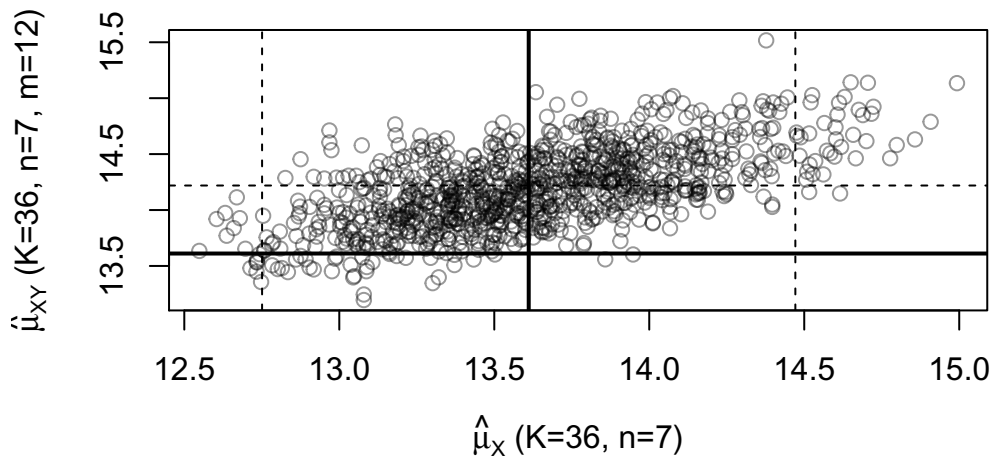
Dále se zaměřme na rozptyl $\hat{\mu}_{XY}$ za platnosti hypotézy:

$$\begin{aligned} Var\,\hat{\mu}_{XY} &= (N_X N_Y)^{-2} \sum_{i=1}^{N_X} \sum_{j=1}^{N_Y} \sum_{k=1}^{N_X} \sum_{h=1}^{N_Y} Cov\left[\|\boldsymbol{X}_i - \boldsymbol{Y}_j\|^2, \|\boldsymbol{X}_k - \boldsymbol{Y}_h\|^2\right] \\ &= (N_X N_Y)^{-2} \sum_{i=1}^{N_X} \sum_{j=1}^{N_Y} \sum_{k=1}^{N_X} \sum_{h=1}^{N_Y} Cov\left[d^2_{(xy)ij}, d^2_{(xy)kh}\right] \\ &= (N_X N_Y)^{-1} \left[Var\, d^2_{(x)12} + (N_X + N_Y - 2) \cdot Cov(d^2_{(x)12}, d^2_{(x)13})\right] \end{aligned}$$

Jelikož z kapitoly 2 známe $Var\, d^2_{(x)12}$ a $Cov(d^2_{(x)12}, d^2_{(x)13})$, umíme tedy za platnosti hypotézy spočítat i rozptyly náhodných veličin $\hat{\mu}_X$ a $\hat{\mu}_{XY}$. Zanesme je do našich grafů. Nechť v grafu silná čára značí střední hodnotu a přerušované čáry označují vzdálenost od střední hodnoty $\pm 2\sqrt{Var\,\hat{\mu}_X}$ nebo $\pm 2\sqrt{Var\,\hat{\mu}_{XY}}$. Získali jsme tak grafy 3.3 a 3.4. Je zřejmé, že střední hodnota a rozptyl kolmé k ose $\hat{\mu}_X$ budou data aproximovat hezky. Pokud střední hodnota a rozptyl kolmé k ose $\hat{\mu}_{XY}$ také dobře aproximují data, pak můžeme usoudit, že hypotéza by nemusela být zamítnuta. Opačný případ by měl svědčit v neprospěch hypotézy.




Obrázek 3.3: Graf značí hodnoty $\hat{\boldsymbol{\mu}}_{D_{F_X}}$ a jsou v něm zakresleny střední hodnoty a rozptyly náhodných veličin $\hat{\mu}_X$ a $\hat{\mu}_{XY}$.




Obrázek 3.4: Graf značí hodnoty $\hat{\boldsymbol{\mu}}_{D_{F_X}}$ a jsou v něm zakresleny střední hodnoty a rozptyly náhodných veličin $\hat{\mu}_X$ a $\hat{\mu}_{XY}$.

O platnosti hypotézy ale rozhodneme permutačním testem(viz. A.1). Výsledky testu nejsou překvapivé. V prvním případě hypotézu na základě permutačního testu nezamítáme a ve druhém případě hypotézu zamítáme. Je ale zapotřebí zmínit, že ne vždy je z grafu jednoznačně poznat, jestli by měla být hypotéza zamítnuta.

# 4. Testování pomocí testové statistiky BG

Díky projektu EXPRO (Staré mýty, nová fakta: české země v centru hudebního dění 15. století) jsem se dostala k tabulce o výskytu písní v konkrétních renesančních zpěvnících z 15. a 16. století. Tuto tabulku jsem přepracovala do tabulky A.1, která se zaměřuje jen na 15. století a popisuje, kolikrát byla daná píseň opsána do všech zpěvníků dohromady v 1. nebo 2. polovině 15. století.

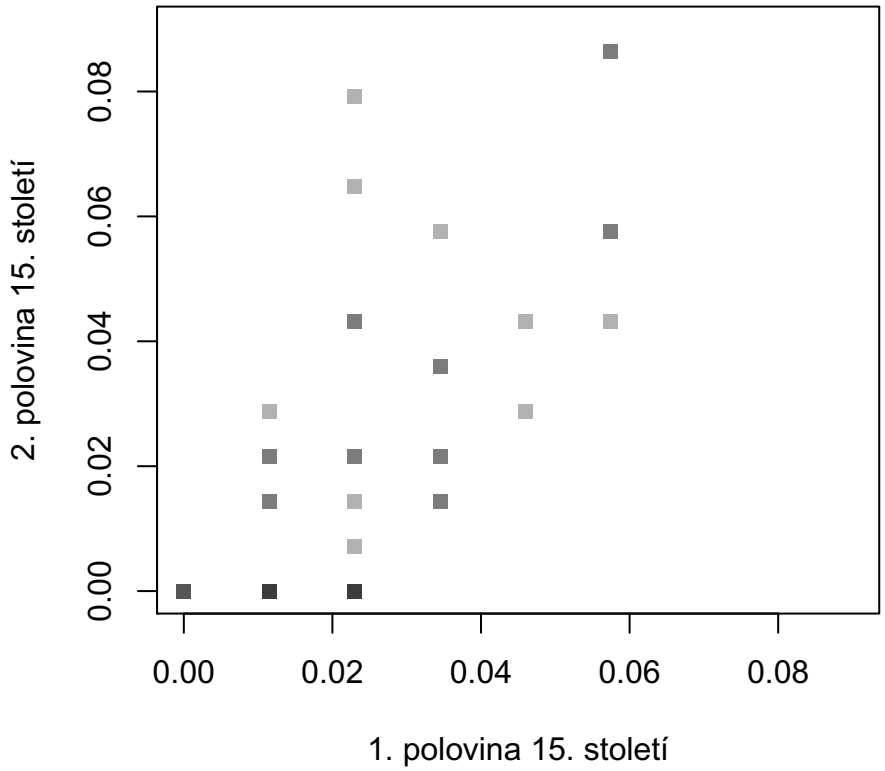
Chceme testovat, jestli byly písničky v 1. polovině 15. století zařazovány do zpěvníků se stejnou pravděpodobností jako ve 2. polovině 15. století. Položme toto tvrzení jako hypotézu. Nejprve ji otestujeme pomocí $\chi^2$ -testu. Abychom ho mohli aplikovat, nahradíme nulové hodnoty malými kladnými hodnotami (viz. článek Modarres (2018, str. 342)). Nahraďme tedy nuly například hodnotou 0.001. Testujme na hladině $\alpha = 0.05$. Pak je $p$-hodnota rovná 0.6593, tedy hypotézu podle $\chi^2$-testu nemůžeme zamítnout.

Nyní bychom chtěli hypotézu ověřit pomocí $BG$ testové statistiky. Musíme ale nejdříve upravit naše data. Vektor, který udává počty opsání jednotlivých písní v dané polovině 15. století si rozepíšeme na takové jednotkové vektory, že jejich součtem získáme vektor původní. Pokud máme realizaci náhodného vektoru $(2,5,2,\ldots,2,0,0,0)^T$, který má 38 složek a jejich součet je 87, pak jsou nové vektory tvaru:

$$\boldsymbol{X}_1 = e_1,\ \boldsymbol{X}_2 = e_1,\ \boldsymbol{X}_3 = e_2,\ \ldots,\ \boldsymbol{X}_7 = e_2,\ \ldots,\ \boldsymbol{X}_{86} = e_{35}, \boldsymbol{X}_{87} = e_{35}.$$

Složky každého vektoru se sečtou na jedno společné $n \in \mathbb{N}$, a to konkrétně na $n = 1$. Tedy $\boldsymbol{X}_i$ jsou realizacemi náhodných vektorů z multinomického rozdělení $Mult_{38}(1, \boldsymbol{p}_x)$ s rozsahem 87. Stejně tak vytvoříme 139 vektorů $\boldsymbol{Y}_i$, které budou realizacemi náhodných vektorů z rozdělení $Mult_{38}(1, \boldsymbol{p}_y)$. Nyní můžeme formulovat naší hypotézu jako $H_0 : \boldsymbol{p}_x = \boldsymbol{p}_y$. Testujeme tedy tvrzení, že obě rozdělení se rovnají. Tuto hypotézu testuje $BG$ statistika, kterou napočítáme pro $\{\boldsymbol{X}_i\}_{i=1}^{87}$ a $\{\boldsymbol{Y}_i\}_{i=1}^{139}$ (viz. A.2). Za pomoci permutační metody jsem získala $p$-hodnotu rovnou 0.012. Jelikož $0.012 < 0.05$ hypotézu o rovnosti rozdělení zamítáme.

Každý test rozhodl o zamítnutí hypotézy jinak a $p$-hodnoty jsou hodně rozdílné. Znázorněme si tedy relativní četnosti písní do grafu. Pak bod grafu o souřadnicích $(i,j)$ je písní, která byla v 1. polovině 15. století $(i \cdot 87)$-krát opsána a $(j \cdot 139)$-krát opsána ve 2. polovině 15. století. Jelikož se body často překrývají, je jejich četnost znázorněna stupni šedi. Pokud Vám z dat připadá, že písničky byly v 1. polovině 15. století zařazovány do zpěvníků se stejnou pravděpodobností jako ve 2. polovině 15. století, pak jistě očekáváte graf, který se podobá grafu souměrnému podle funkce $y = x$. Graf 4.1 se takto souměrnému grafu příliš nepodobá.




Obrázek 4.1: Graf značí počty opisů jednotlivých písní dělené celkovým počtem opisů za danou polovinu 15. století.

Potvrdilo se nám tedy, co již známe z úvodu. Tedy, že $\chi^2$-test není vhodný pro řídká data a v takovém případě je lepší použít test založený na $BG$ statistice.

# Závěr

V práci jsme se důkladně seznámili s pojmem vzdálenosti mezi body a dokázali jsme jeho základní vlastnosti. Tyto důkazy, které v článku Modarres (2018) chyběly nebo byly velmi zkratkovité, tvoří stěžejní část celé práce.

Dále jsme si ukázali, jak ze vzdálenosti bodů sestavit testovou statistiku $BG$ definovanou v úvodu. Testová statistika $BG$ se napočítává z náhodných vektorů $\hat{\boldsymbol{\mu}}_{D_{F_X}}$ a $\hat{\boldsymbol{\mu}}_{D_{F_Y}}$. V kapitole 3 jsme se věnovali náhodnému vektoru $\hat{\boldsymbol{\mu}}_{D_{F_X}}$. Nejprve jsme generovali data z rozdělení $Mult_K(n,\boldsymbol{p}_x)$ a $Mult_K(n,\boldsymbol{p}_y)$. Z těchto dat jsme napočítali hodnoty $\hat{\boldsymbol{\mu}}_{D_{F_X}}$, které jsme zakreslili do grafu. Dále jsme si ukázali, jak napočítat střední hodnotu a rozptyl jednotlivých složek vektoru $\hat{\boldsymbol{\mu}}_{D_{F_X}}$ za hypotézy rovnosti multinomických rozdělení. Tyto hodnoty jsme také zakreslili do grafu. Podle toho, jak dobře vykreslené teoretické hodnoty aproximují data, si lze udělat obrázek o hypotéze. Nakonec jsme rozhodnutí o zamítnutí hypotézy vyvodili z výsledku permutačního testu, který počítal $BG$ testovou statistiku.

V poslední kapitole 4 jsme se zabývali testováním řídkých dat. Tentokrát jsme testovali data skutečná, ne simulovaná. Testovali jsme hypotézu, že písničky byly v 1. polovině 15. století zařazovány do zpěvníků se stejnou pravděpodobností jako ve 2. polovině 15. století. Nejprve jsme použili $\chi^2$-test. Potom jsme naše data upravili do takového tvaru, abychom mohli napočítat testovou statistiku $BG$ a permutační metodou určit $p$-hodnotu. Každý test rozhodl o zamítnutí hypotézy jinak. Proto jsme si znázornili relativní četnosti písní do grafu, ze kterého je patrné, že pravděpodobnost zařazení písniček se v obou polovinách 15. století značně lišila. Potvrdili jsme si tak, že test založený na vzdálenostech mezi body je pro řídká data vhodnější než $\chi^2$-test.

# Seznam použité literatury


Biswas, M. a Ghosh, A. K. (2014). A nonparametric two-sample test applicable to high dimensional data. *Journal of Multivariate Analysis*, **123**, 160 – 171. ISSN 0047-259X. doi: https://doi.org/10.1016/j.jmva.2013.09.004. URL http://www.sciencedirect.com/science/article/pii/S0047259X13001966.

Modarres, R. (2018). Multinomial interpoint distances. *Statist. Papers*, **59** (1), 341–360. ISSN 0932-5026. doi: 10.1007/s00362-016-0766-7. URL https://doi.org/10.1007/s00362-016-0766-7.

Omelka, M. (2020). NMSA331 Matematická statistika 1 Poznámky k přednášce. [online]. URL http://msekce.karlin.mff.cuni.cz/~omelka/Soubory/nmsa331/ms1.pdf.

# A. Přílohy

## A.1 Vykreslení dat a testování hypotézy z kapitoly 3

```
n.sim=1000
n.x=50                    #rozsah vyberu X
n.y=n.x                   #rozsah vyberu Y
x.m=numeric(n.sim)
y.m=numeric(n.sim)
xy.m=numeric(n.sim)
k=36                      #pocet kategorii
n=7
i=12
pr.x=rep(1/k, k)          #pstni vektory
pr.y=c(rep((1/(k-i)),(k-i)),rep(0,i))

bg.f=numeric(n.sim)
bg.s=numeric(n.sim)
x.m.perm=numeric(n.sim)
y.m.perm=numeric(n.sim)
xy.m.perm=numeric(n.sim)


for (t in 1:n.sim) {
        x=rmultinom(n.x,n,pr.x) #generuji n.v X
        y=rmultinom(n.y,n,pr.y) #generuji n.v Y
        s.x=0                   #pomocne promenne pro castecne soucty
        s.y=0
        s.xy=0
        for(i in 1:n.x) {
                for(j in 1:n.y) {
                        if(j>i) {
                                s.x=s.x + sum((x[,i]-x[,j])^2)
                                s.y=s.y + sum((y[,i]-y[,j])^2)
                        }
                        s.xy=s.xy + sum((x[,i]-y[,j])^2)
                }
        }
        x.m[t]=s.x/choose(n.x,2)
        y.m[t]=s.y/choose(n.x,2)
        xy.m[t]=s.xy/(n.x*n.y)
        bg.f[t]=sqrt((x.m[t]-xy.m[t])^2+(xy.m[t]-y.m[t])^2)


        z=cbind(x,y)
        m=ncol(z)
        perm=sample(1:m,m,replace = FALSE)
        x.perm=z[,perm[1:n.x]]
```

```
        y.perm=z[,perm[(n.x+1):m]]

        s.x=0
        s.y=0
        s.xy=0
        for(i in 1:n.x) {
                for(j in 1:n.y) {
                        if(j>i) {
                          s.x=s.x + sum((x.perm[,i]-x.perm[,j])^2)
                          s.y=s.y + sum((y.perm[,i]-y.perm[,j])^2)
                        }
                        s.xy=s.xy + sum((x.perm[,i]-y.perm[,j])^2)
                }
        }
        x.m.perm[t]=s.x/choose(n.x,2)
        y.m.perm[t]=s.y/choose(n.x,2)
        xy.m.perm[t]=s.xy/(n.x*n.y)
        bg.s[t]=sqrt((x.m.perm[t]-xy.m.perm[t])^2
                     +(xy.m.perm[t]-y.m.perm[t])^2)
}

# warp speed monte carlo (jen jedna permutace v kazde simulaci)

val=mean(bg.s>=bg.f) #empricka sila testu

#v je pocet kovarianci a variaci
v=0
for (i in 1:n.x) {for (j in 1:n.x) {if(j>i) {
        for (l in 1:n.x) {for (h in 1:n.x) {if(h>l){
                if((i==l)|(i==h)|(j==l)|(j==h)){v=v+1}
        }}}
}}}

#napocitani covariance ze simulace dat
s=100000
d.1=numeric(s)
d.2=numeric(s)
sim.c=numeric(10)
for (q in 1:10) {
        for (t in 1:s) {
                x1=rmultinom(1,n,pr.x)
                x2=rmultinom(1,n,pr.x)
                x3=rmultinom(1,n,pr.x)
                d.1[t]=sum((x1-x2)^2)
                d.2[t]=sum((x1-x3)^2)
        }
        sim.c[q]=cov(d.1,d.2)
}
c.d2x=mean(sim.c)

#vypocet vlastnosti slozek statistiky
```

```
e.d2x=2*n*(1-1/k)
v.d2x=4*n*(2*n-1)*(1/k)*(1-(1/k))

v.mux=((choose(n.x,2))^(-2))*((choose(n.x,2)*v.d2x)
                              + ((v-choose(n.x,2)) *c.d2x))
v.x=2*sqrt(v.mux)

v.muxy=n.x^(-2)*(v.d2x+(2*(n.x-1)*c.d2x))
v.xy=2*sqrt(v.muxy)

#graf s teoretickymi hodnotami
plot(x.m,xy.m,pch = 1, col = rgb(0, 0, 0, 0.4),
     xlab = expression(paste(hat(mu)[X],' (K=36, n=7)')),
     ylab = expression(paste(hat(mu)[XY],' (K=36, n=7, m=12)')))
abline(h=e.d2x,v=e.d2x,lwd=2)
abline(h=(e.d2x + v.xy),lty=2)
abline(v=(e.d2x + v.x),lty=2)
abline(h=(e.d2x - v.xy),lty=2)
abline(v=(e.d2x - v.x),lty=2)
```

## A.2 Testování hypotézy z kapitoly 4 permutační metodou

```
x=diag(38)[,unlist(lapply(1:38,function(i)rep(i,sumy[i,1])))]
y=diag(38)[,unlist(lapply(1:38,function(i)rep(i,sumy[i,2])))]
n.x=ncol(x)
n.y=ncol(y)
s=500
bg.p=numeric(s)

#vypocet testove statistiky z dat v tabulce
s.x=0
for(i in 1:n.x) {
        for(j in 1:n.x) {
                if(j>i) {s.x=s.x + sum((x[,i]-x[,j])^2)}
        }
}
x.m=s.x/choose(n.x,2)

s.y=0
for(i in 1:n.y) {
        for(j in 1:n.y) {
                if(j>i) {s.y=s.y + sum((y[,i]-y[,j])^2)}
        }
}
y.m=s.y/choose(n.y,2)

s.xy=0
for(i in 1:n.x) {
```

```
        for(j in 1:n.y) {
                s.xy=s.xy + sum((x[,i]-y[,j])^2)
        }
}
xy.m=s.xy/(n.x*n.y)

bg=sqrt((x.m-xy.m)^2+(xy.m-y.m)^2) #testova statistika

#permutacni test
z=cbind(x,y)
n=ncol(z)

for (t in 1:s) {
        perm=sample(1:n,n,replace = FALSE)
        x.p=z[,perm[1:n.x]]
        y.p=z[,perm[(n.x+1):n]]

        s.x=0
        for(i in 1:n.x) {
                for(j in 1:n.x) {
                        if(j>i) {s.x=s.x + sum((x.p[,i]-x.p[,j])^2)}
                }
        }
        x.m=s.x/choose(n.x,2)

        s.y=0
        for(i in 1:n.y) {
                for(j in 1:n.y) {
                        if(j>i) {s.y=s.y + sum((y.p[,i]-y.p[,j])^2)}
                }
        }
        y.m=s.y/choose(n.y,2)

        s.xy=0
        for(i in 1:n.x) {
                for(j in 1:n.y) {
                        s.xy=s.xy + sum((x.p[,i]-y.p[,j])^2)
                }
        }
        xy.m=s.xy/(n.x*n.y)

        bg.p[t]=sqrt((x.m-xy.m)^2+(xy.m-y.m)^2) #testove statistiky
}

p=mean(bg.p>bg) #p hodnota permutacni metodou
```

## A.3 Tabulky

Tabulka A.1: Tabulka uvádí, kolikrát byly písně přepsány do zpěvníků v 1. a 2. polovině 15. století.

| **Píseň** | **1400 - 1449** | **1450-1499** |
|---|---|---|
| Felici peccatrici | 2 | 11 |
| Ihesus Christus nostra salus | 5 | 12 |
| Cedit hiems eminus | 2 | 6 |
| Dies est leticie | 5 | 12 |
| Ad honorem et decorem | 3 | 8 |
| Veni dulcis consolator | 5 | 8 |
| Ave yerarchia | 2 | 9 |
| Salve regina glorie | 5 | 8 |
| Ave Maris stella lucens | 5 | 6 |
| Imperatrix gloriosa | 2 | 6 |
| Ave trinitatis cubile | 3 | 3 |
| Nunc festum celebremus | 4 | 6 |
| Candens ebur castitatis | 3 | 5 |
| Ex legis observancia | 2 | 3 |
| Mittitur archangelus | 1 | 3 |
| Pueri nativitatem | 4 | 4 |
| Constat ethereis | 1 | 2 |
| E morte pater divinus | 2 | 3 |
| Omnes attendite | 2 | 2 |
| Stupefactus inferni dux | 3 | 2 |
| Ave rosa in yericho | 1 | 4 |
| Prima declinacio | 1 | 2 |
| Stalať se jest | 3 | 5 |
| Gaude regina glorie | 2 | 1 |
| Puer nobis nascitur | 3 | 3 |
| Ursula speciosa | 2 | 0 |
| Iam prestolantes gloriam | 1 | 0 |
| Sol de stella | 1 | 3 |
| Cum gaudio concurrite | 3 | 2 |
| Resurrexit dominus | 1 | 0 |
| Laus domino resonet | 2 | 0 |
| Quidam triplo metro | 2 | 0 |
| Dyvo flagrans numine | 1 | 0 |
| Sampsonis honestissima | 1 | 0 |
| Jezu krysste styedry knyeze | 2 | 0 |
| Imperatrix gloriosa2 | 0 | 0 |
| Ad honorem et decorem2 | 0 | 0 |
| Ihesus Christus nostra salus2 | 0 | 0 |