## Osnova přednášky Vícerozměrné analogie t-testů

### I. Úlohy o jednom náhodném výběru z vícerozměrného rozložení

1. Charakteristiky p-rozměrného rozložení
2. Odhady charakteristik p-rozměrného rozložení
3. Základní poznatky o p-rozměrném normálním rozložení
4. Náhodný výběr z p-rozměrného normálního rozložení
5. Test hypotézy o vektoru středních hodnot
   Příklad na vícerozměrný jednovýběrový t-test
6. Test hypotézy o úplné nezávislosti sledovaných proměnných
   Příklad na test hypotézy o úplné nezávislosti sledovaných proměnných

### II. Úlohy o dvou nezávislých náhodných výběrech z vícerozměrného rozložení

1. Test hypotézy o rozdílu vektorů středních hodnot
2. Test hypotézy o shodě variančních matic
3. Příklad na Hotellingův $T^2$ test

# Vícerozměrné analogie t-testů

## I. Úlohy o jednom náhodném výběru z vícerozměrného rozložení

### 1. Charakteristiky p-rozměrného rozložení

Náhodný vektor $\mathbf{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}$ pochází z p-rozměrného rozložení s vektorem středních hodnot

$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix}$, varianční maticí

$$\operatorname{var} \mathbf{X} = \mathbf{\Sigma} = \begin{pmatrix} D(X_1) & C(X_1, X_2) & \dots & C(X_1, X_p) \\ \dots & \dots & \dots & \dots \\ C(X_p, X_1) & C(X_p, X_1) & \dots & D(X_p) \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1p} \\ \dots & \dots & \dots & \dots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_p^2 \end{pmatrix},$$

a korelační maticí

$$\operatorname{cor}\mathbf{X} = \boldsymbol{\rho} = \begin{pmatrix} 1 & R(X_1, X_2) & \dots & R(X_1, X_p) \\ \dots & \dots & \dots & \dots \\ R(X_p, X_1) & R(X_p, X_1) & \dots & 1 \end{pmatrix} = \begin{pmatrix} 1 & \rho_{12} & \dots & \rho_{1p} \\ \dots & \dots & \dots & \dots \\ \rho_{p1} & \rho_{p2} & \dots & 1 \end{pmatrix}.$$

(Matice var **X**, cor **X** jsou symetrické, cor **X** se dá vypočítat z var **X**: $\rho_{jk} = \dfrac{\sigma_{jk}}{\sigma_j \sigma_k}$.)

Počet charakteristik p-rozměrného rozložení tedy je:
p středních hodnot,
p rozptylů,
$\dfrac{p(p-1)}{2}$ kovariancí (kovariance je symetrická).

Celkem: $2p + \dfrac{p(p-1)}{2} = \dfrac{p^2 + 3p}{2}$.

Vidíme, že počet charakteristik roste kvadraticky s počtem složek náhodného vektoru. Např. pro p = 2 je jich $\dfrac{2^2 + 3\cdot 2}{2} = 5$, ale pro p = 10 už jich je $\dfrac{10^2 + 3\cdot 10}{2} = 65$.

## 2. Odhady charakteristik p-rozměrného rozložení

Vektor středních hodnot $\boldsymbol{\mu}$ a varianční matici $\boldsymbol{\Sigma}$ v praxi většinou neznáme, musíme je odhadnout na základě náhodného výběru. Pořídíme náhodný výběr $\mathbf{X}_1, \ldots, \mathbf{X}_n$ (kde $\mathbf{X}_i = \left(X_{i1}, \ldots, X_{ip}\right)^T, \ i = 1, \ldots, n$) z p-rozměrného rozložení s vektorem středních hodnot $\boldsymbol{\mu}$ a varianční maticí $\boldsymbol{\Sigma}$. Z těchto n náhodných vektorů utvoříme náhodnou matici

$$\begin{pmatrix} \mathbf{X}_1^T \\ \vdots \\ \mathbf{X}_n^T \end{pmatrix} = \begin{pmatrix} X_{11} & \ldots & X_{1p} \\ \ldots & \ldots & \ldots \\ X_{n1} & \ldots & X_{np} \end{pmatrix}.$$

K číselné realizaci této náhodné matice dospějeme tak, že na n objektech zjišťujeme hodnoty p proměnných. Např. náhodně vybereme n = 31 návštěvníků posilovny a zjišťujeme u nich hodnoty p = 4 proměnných: věk (v letech), hmotnost (v kg), doba cvičení (v min), maximální tep.

Znamená to, že i-tý objekt je charakterizován p-rozměrným vektorem pozorování $\mathbf{x}_i = \left(x_{i1}, \ldots, x_{ip}\right)^T, \ i = 1, \ldots, n$. Vektory pozorování uspořádáme do datové matice

$$\begin{pmatrix} x_{11} & \ldots & x_{1p} \\ \ldots & \ldots & \ldots \\ x_{n1} & \ldots & x_{np} \end{pmatrix},$$ kde řádky odpovídají jednotlivým objektům a sloupce proměnným.

V našem případě máme datovou matici tvaru:

| | 1<br>věk | 2<br>hmotnost | 3<br>doba cvičení | 4<br>max. tep |
|---|---|---|---|---|
| 1 | 51 | 75,1 | 12,6 | 180 |
| 2 | 48 | 81,1 | 11,2 | 161 |
| 3 | 46 | 78 | 9,6 | 171 |
| 4 | 44 | 72,6 | 8,9 | 157 |
| 5 | 45 | 69 | 11,1 | 175 |
| 6 | 48 | 93,3 | 12,9 | 172 |
| 7 | 45 | 75,4 | 10,5 | 191 |
| 8 | 51 | 60,8 | 9,9 | 153 |
| 9 | 43 | 78 | 9,4 | 196 |
| 10 | 42 | 62,9 | 11,5 | 178 |
| 11 | 46 | 84,5 | 10,5 | 174 |
| 12 | 38 | 74,7 | 10,1 | 177 |
| 13 | 39 | 89,4 | 14 | 190 |
| 14 | 39 | 68,2 | 11,1 | 189 |
| 15 | 41 | 80,9 | 10,6 | 169 |
| 16 | 48 | 84,8 | 10,3 | 179 |
| 17 | 43 | 83,1 | 9 | 193 |
| 18 | 45 | 71,3 | 11 | 176 |
| 19 | 45 | 79,6 | 10 | 171 |
| 20 | 42 | 93,3 | 10,3 | 168 |
| 21 | 43 | 75,1 | 10,1 | 179 |
| 22 | 38 | 91,2 | 11,4 | 185 |
| 23 | 34 | 76,8 | 10,1 | 194 |
| 24 | 38 | 87,5 | 8,7 | 162 |
| 25 | 36 | 69,9 | 8,2 | 173 |
| 26 | 32 | 90,7 | 9,2 | 185 |
| 27 | 41 | 79,2 | 11,6 | 184 |
| 28 | 34 | 77,7 | 12 | 186 |
| 29 | 37 | 82,9 | 10,9 | 168 |
| 30 | 38 | 83,1 | 13,1 | 184 |
| 31 | 32 | 83,6 | 8,6 | 179 |

Zavedeme následující označení:

$M_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij}$ … výběrový průměr j-té proměnné, $j = 1, \ldots, p$

$\mathbf{M} = \begin{pmatrix} M_1 & \ldots & M_p \end{pmatrix}^T$ … vektor výběrových průměrů

(V našem případě: $\mathbf{m} = \begin{pmatrix} 41{,}7 & 79{,}2 & 10{,}6 & 177{,}4 \end{pmatrix}^T$, tedy průměrný věk je 41,7 roku, průměrná hmotnost je 79,2 kg, průměrná doba cvičení je 40,6 min a průměrný maximální tep je 177,4. )

$S_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_{ij} - M_j\right)^2$ … výběrový rozptyl j-té proměnné, $j = 1, \ldots, p$

$S_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_{ij} - M_j\right)\left(X_{ik} - M_k\right)$… výběrová kovariance j-té a k-té proměnné, $j, k = 1, \ldots, p$

$$\mathbf{S} = \begin{pmatrix} S_1^2 & S_{12} & \ldots & S_{1p} \\ \ldots & \ldots & \ldots & \ldots \\ S_{p1} & S_{p2} & \ldots & S_p^2 \end{pmatrix} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\mathbf{X}_i - \mathbf{M}\right)\left(\mathbf{X}_i - \mathbf{M}\right)^T$$ … výběrová varianční matice

(Matice $\mathbf{W} = \sum_{i=1}^{n}\left(\mathbf{X}_i - \mathbf{M}\right)\left(\mathbf{X}_i - \mathbf{M}\right)^T = (n-1)\mathbf{S}$ se nazývá Wishartova matice.)

(V našem případě: $\mathbf{s} = \begin{pmatrix} 27{,}16 & -10{,}16 & 1{,}34 & -21{,}20 \\ & 69{,}24 & 1{,}71 & 12{,}02 \\ & & 1{,}92 & 3{,}40 \\ & & & 118{,}31 \end{pmatrix}$.)

$R_{jk} = \dfrac{S_{jk}}{S_j S_k}$ … výběrový koeficient korelace j-té a k-té proměnné, $j, k = 1, \ldots, p$

$\mathbf{R} = \begin{pmatrix} 1 & R_{12} & \ldots & R_{1p} \\ \ldots & \ldots & \ldots & \ldots \\ R_{p1} & R_{p2} & \ldots & 1 \end{pmatrix}$ … výběrová korelační matice

(V našem případě: $\mathbf{r} = \begin{pmatrix} 1 & -0{,}23 & 0{,}19 & -0{,}37 \\ & 1 & 0{,}15 & 0{,}13 \\ & & 1 & 0{,}23 \\ & & & 1 \end{pmatrix}$, tedy věk záporně koreluje s hmotností a tepem, ale kladně s dobou cvičení, hmotnost kladně koreluje s dobou cvičení a tepem a doba cvičení kladně koreluje s tepem. )

Lze dokázat, že

- vektor výběrových průměrů $\mathbf{M}$ je nestranným odhadem vektoru středních hodnot $\boldsymbol{\mu}$, tj. $E(\mathbf{M}) = \boldsymbol{\mu}$;
- výběrová varianční matice $\mathbf{S}$ je nestranným odhadem varianční matice $\boldsymbol{\Sigma}$, tj. $E(\mathbf{S}) = \boldsymbol{\Sigma}$;
- výběrová korelační matice $\mathbf{R}$ je vychýleným odhadem korelační matice $\boldsymbol{\rho}$, tj. $E(\mathbf{R}) \approx \boldsymbol{\rho}$.

**Poznámka**: V některých situacích pracujeme s lineární kombinací složek náhodného vektoru $\mathbf{X}$: $c_1X_1 + \ldots + c_pX_p = \mathbf{c}^T\mathbf{X}$. Pak střední hodnota náhodné veličiny $\mathbf{c}^T\mathbf{X}$ je $\mathbf{c}^T\boldsymbol{\mu}$ a rozptyl je $\mathbf{c}^T\boldsymbol{\Sigma}\mathbf{c}$. Nestranným odhadem střední hodnoty $\mathbf{c}^T\boldsymbol{\mu}$ je $\mathbf{c}^T\mathbf{M}$ a nestranným odhadem rozptylu $\mathbf{c}^T\boldsymbol{\Sigma}\mathbf{c}$ je $\mathbf{c}^T\mathbf{S}\mathbf{c}$.

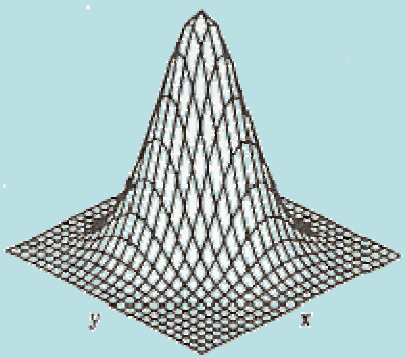
## 3. Základní poznatky o p-rozměrném normálním rozložení

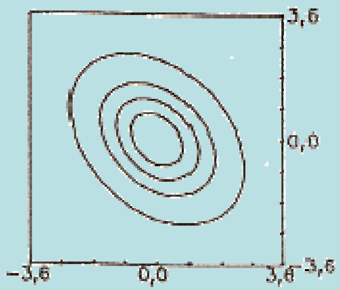
Náhodný vektor $\mathbf{X} = \left(X_1, \ldots, X_p\right)^T$ se řídí p-rozměrným normálním rozložením $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, kde parametr $\boldsymbol{\mu}$ je vektor středních hodnot a parametr $\boldsymbol{\Sigma}$ je varianční matice, když jeho hustota má tvar:

$$\varphi(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{p}{2}} \left|\boldsymbol{\Sigma}\right|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})}.$$

**Ilustrace pro dvourozměrné normální rozložení**

Graf hustoty a vrstevnice dvourozměrného normálního rozložení s parametry $\mu_1 = 0$, $\mu_2 = 0$, $\sigma_1^2 = 1$, $\sigma_2^2 = 1$, $\rho = -0,75$:



**Důležité vlastnosti p-rozměrného normálního rozložení:**

a) Všechna marginální (a podmíněná) rozložení jsou normální.
b) Lineární transformací $\mathbf{Y} = \mathbf{a} + \mathbf{B}\mathbf{X}$, kde $\mathbf{a}$ je p-rozměrný sloupcový reálný vektor a $\mathbf{B}$ je reálná čtvercová matice řádu p, se normalita neporuší: $\mathbf{Y} \sim \mathrm{N}_p\left(a + B\boldsymbol{\mu}, B\boldsymbol{\Sigma}B^T\right)$
c) Je-li varianční matice $\boldsymbol{\Sigma}$ diagonální, jsou náhodné veličiny $X_1, \ldots, X_p$ stochasticky nezávislé.
d) Sečteme-li n stochasticky nezávislých p-rozměrných náhodných vektorů, z nichž každý se řídí p-rozměrným normálním rozložením, pak výsledný součet má také p-rozměrné normální rozložení.

## 4. Náhodný výběr z p-rozměrného normálního rozložení

Nechť náhodný výběr $\mathbf{X_1}, \ldots, \mathbf{X}_n$ pochází z rozložení $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Označme $\mathbf{M}$ vektor výběrových průměrů a $\mathbf{S}$ výběrovou varianční matici. Pak platí:

a) Wishartova matice $\mathbf{W} = (n-1)\mathbf{S}$ má p-rozměrné Wishartovo rozložení s n-1 stupni volnosti a parametrem $\boldsymbol{\Sigma}$, píšeme $\mathbf{W} \sim W_p(n-1, \boldsymbol{\Sigma})$. (Wishartovo rozložení je zobecněním $\chi^2$- rozložení. Je-li p = 1 a $\boldsymbol{\Sigma} = (1)$, jde o rozložení $\chi^2(n-1)$.)

b) Statistika $T^2 = n(\mathbf{M}-\boldsymbol{\mu})^T \mathbf{S}^{-1}(\mathbf{M}-\boldsymbol{\mu})$ má Hotellingovo rozložení s p a n-1 stupni volnosti, píšeme $T^2 \sim T^2(p, n\text{-}1)$. (Hotellingovo rozložení je zobecněním Studentova rozložení.)

**Poznámka:** Mezi Hotellingovým a Fisherovým – Snedecorovým rozložením platí vztah:

$X \sim T^2(\nu_1, \nu_2) \Rightarrow Y = \dfrac{\nu_2 - \nu_1 + 1}{\nu_1 \nu_2} X \sim F(\nu_1, \nu_2 - \nu_1 + 1)$. Statistiku $T^2$ tedy můžeme transformovat na statistiku s F-S rozložením:

$$\frac{n-p}{p(n-1)} T^2 = \frac{n(n-p)}{p(n-1)} (\mathbf{M}-\boldsymbol{\mu})^T \mathbf{S}^{-1}(\mathbf{M}-\boldsymbol{\mu}) \sim F(p, n-p).$$

## 5. Test hypotézy o vektoru středních hodnot

Tento test je p-rozměrnou analogií jednovýběrového t-testu. Pro připomenutí:

Náhodný výběr $X_1, \ldots, X_n$ pochází z rozložení $N(\mu, \sigma^2)$, kde parametry $\mu, \sigma^2$ neznáme. Na hladině významnosti $\alpha$ testujeme hypotézu $H_0 : \mu = c$ proti alternativě $H_1 : \mu \neq c$.

Testová statistika: $T_0 = \dfrac{M - c}{\frac{S}{\sqrt{n}}}$ se za platnosti $H_0$ řídí rozložením $t(n-1)$.

Kritický obor: $W = \left(-\infty, -t_{1-\alpha/2}(n-1)\right\rangle \cup \left\langle t_{1-\alpha/2}(n-1), \infty\right)$.

Jestliže $t_0 \in W$, $H_0$ zamítáme na hladině významnosti $\alpha$.

**Poznámka:** Vzhledem k tomu, že platí tvrzení: $X \sim t(n) \Rightarrow Y = X^2 \sim F(1, n)$, můžeme $H_0$ zamítnout na hladině významnosti $\alpha$, když $t_0^{\,2} \in \left\langle F_{1-\alpha}(1, n-1), \infty\right)$.

p-rozměrný případ:

Náhodný výběr $\mathbf{X}_1, \ldots, \mathbf{X}_n$ pochází z rozložení $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, kde parametry $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ neznáme. Na hladině významnosti $\alpha$ testujeme hypotézu $H_0: \boldsymbol{\mu} = \mathbf{c}$ proti alternativě $H_1: \boldsymbol{\mu} \neq \mathbf{c}$, kde $\mathbf{c} = (c_1, \ldots, c_p)^T$ je vektor reálných konstant. (Alternativa vlastně tvrdí, že aspoň jedna složka vektoru středních hodnot neodpovídá ověřovanému předpokladu.)

Testová statistika $T_0 = \dfrac{n(n-p)}{p(n-1)}(\mathbf{M}-\mathbf{c})^T \mathbf{S}^{-1}(\mathbf{M}-\mathbf{c})$ se za platnosti $H_0$ řídí rozložením $F(p, n-p)$.

Kritický obor: $W = \langle F_{1-\alpha}(p, n-p), \infty)$.

Jestliže $t_0 \in W$, $H_0$ zamítáme na hladině významnosti $\alpha$.

**Poznámka:** Test $H_0: \boldsymbol{\mu} = \mathbf{c}$ proti $H_1: \boldsymbol{\mu} \neq \mathbf{c}$ nelze nahradit p jednorozměrnými t-testy $H_{0j}: \mu_j = c_j$ proti $H_{1j}: \mu_j \neq c_j$, $j = 1, \ldots, p$, protože při tomto postupu by pravděpodobnost chyby 1. druhu byla větší než $\alpha$, dokonce až $1-(1-\alpha)^p$.

Pokud na dané hladině významnosti $\alpha$ zamítneme vícerozměrnou hypotézu $H_0 : \boldsymbol{\mu} = \mathbf{c}$ ve prospěch alternativy $H_1 : \boldsymbol{\mu} \neq \mathbf{c}$, zjistíme, vzhledem ke kterým složkám vektoru $\boldsymbol{\mu}$ byla nulová hypotéza zamítnuta.

K tomu lze použít p jednorozměrných t-testů $H_{0j} : \mu_j = c_j$ proti $H_{1j} : \mu_j \neq c_j$, $j = 1, \ldots, p$, u nichž hladinu významnosti $\alpha$ upravíme pomocí Bonferroniho korekce:

$H_{0j}$ zamítneme na hladině významnosti $\alpha$, když vypočtená p-hodnota bude $\leq \dfrac{\alpha}{p}$.

## Příklad na vícerozměrný jednovýběrový t-test

Výrobce určitého typu součástek uvádí, že nejdůležitější čtyři rozměry nabývají těchto hodnot: 9,50 mm, 6,35 mm, 5,98 mm a 4,40 mm. Náhodně bylo vybráno 15 součástek, byly u nich zjištěny hodnoty těchto rozměrů a zapsány do proměnných X1, X2, X3, X4. Údaje jsou uloženy v souboru soucastky.sta.

Za předpokladu, že data pocházejí ze čtyřrozměrného normálního rozložení s neznámým vektorem středních hodnot $\boldsymbol{\mu} = \begin{pmatrix} \mu_1 & \mu_2 & \mu_3 & \mu_4 \end{pmatrix}^T$ a neznámou varianční maticí

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 \end{pmatrix},$$

na hladině významnosti 0,05 testujte hypotézu, že tvrzení výrobce je pravdivé. V případě zamítnutí nulové hypotézy zjistěte, které rozměry přispěly k jejímu zamítnutí.

**Řešení:**

Na hladině významnosti 0,05 testujeme hypotézu $H_0$: $\begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{pmatrix} = \begin{pmatrix} 9{,}50 \\ 6{,}35 \\ 5{,}98 \\ 4{,}40 \end{pmatrix}$

proti alternativě $H_1$: $\begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{pmatrix} \neq \begin{pmatrix} 9{,}50 \\ 6{,}35 \\ 5{,}98 \\ 4{,}40 \end{pmatrix}$.

Hodnotu testové statistiky $T_0 = \frac{n(n-p)}{p(n-1)}(\mathbf{M}-\mathbf{c})^T \mathbf{S}^{-1}(\mathbf{M}-\mathbf{c})$ a odpovídající p-hodnotu vypočteme pomocí systému STATISTICA.

Statistiky – Základní statistiky a tabulky – t-test, samost. vzorek – OK – Proměnné X1, X2, X3, X4 – OK – záložka Možnosti – zvolíme Test průměrů vůči různým volitelným konstantám Specif. X1: 9,5, X2: 6,35, X3: 5,98, X4: 4,4 – OK – zaškrtneme Vícerozměrný test (Hotellingovo $T^2$) – Výpočet. Dostaneme výstupní tabulku:

| | Test průměrů vůči referenční konstantě (hodnotě) (soucastky.sta) T2(celé případy ChD)=19,2432 F(4,11)=3,7799 p<,03597 | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Proměnná | Průměr | Sm.odch. | N | Sm.chyba | Referenční konstanta | t | SV | p |
| X1 | 9,491833 | 0,010695 | 15 | 0,002761 | 9,500000 | -2,95748 | 14 | 0,010391 |
| X2 | 6,357433 | 0,011481 | 15 | 0,002964 | 6,350000 | 2,50752 | 14 | 0,025099 |
| X3 | 5,981467 | 0,011129 | 15 | 0,002873 | 5,980000 | 0,51043 | 14 | 0,617706 |
| X4 | 4,400327 | 0,007024 | 15 | 0,001814 | 4,400000 | 0,18011 | 14 | 0,859646 |

Testová statistika vícerozměrného jednovýběrového t-testu se realizuje hodnotou 3,7799, odpovídající p-hodnota je 0,03597, tedy s rizikem omylu nejvýše 5 % považujeme za prokázané, že rozměry součástky neodpovídají deklarovaným hodnotám.
Protože jsme zamítli nulovou hypotézu, v dalším kroku zjistíme, které rozměry přispěly k jejímu zamítnutí. Budeme tedy simultánně testovat hypotézy $H_{01}: \mu_1 = 9{,}5$, $H_{02}: \mu_2 = 6{,}35$, $H_{03}: \mu_3 = 5{,}98$, $H_{04}: \mu_4 = 4{,}4$ proti $H_{11}: \mu_1 \neq 9{,}5$, $H_{12}: \mu_2 \neq 6{,}35$, $H_{13}: \mu_3 \neq 5{,}98$, $H_{14}: \mu_4 \neq 4{,}4$. $H_{0j}$ zamítneme na hladině významnosti $\alpha = 0{,}05$, když vypočtená p-hodnota bude menší nebo rovna $\frac{\alpha}{\text{počet testů}} = \frac{0{,}05}{4} = 0{,}0125$. Vidíme, že vícerozměrná hypotéza byla zamítnuta kvůli X1.

## 6. Test hypotézy o úplné nezávislosti sledovaných proměnných

Řada statistických úloh vede na zkoumání závislosti mezi p sledovanými proměnnými. Nejdříve by se mělo zjistit, zda se nejedná o systém nezávislých proměnných. V takovém případě by bylo zbytečné pokračovat v analýze závislostí.
Na hladině významnosti 0,05 testujeme $H_0 : \mathrm{cor}\mathbf{X} = \mathbf{I}$ proti $H_0 : \mathrm{cor}\mathbf{X} \neq \mathbf{I}$ ($\mathbf{I}$ je jednotková matice řádu p).

Testová statistika $T_0 = -n \ln|\mathbf{R}|$ se za platnosti $H_0$ asymptoticky řídí rozložením $\chi^2\left(\frac{p(p-1)}{2}\right)$.

Kritický obor: $W = \left\langle \chi^2_{1-\alpha}\left(\frac{p(p-1)}{2}\right), \infty \right)$

Jestliže $t_0 \in W$, $H_0$ zamítáme na hladině významnosti $\alpha$.

**Poznámka:** Aproximaci $\chi^2$-rozložením můžeme zpřesnit, když testovou statistiku $T_0$ vynásobíme konstantou $1 - \frac{2p+11}{6n}$.

## 6. Test hypotézy o úplné nezávislosti sledovaných proměnných

Řada statistických úloh vede na zkoumání závislosti mezi p sledovanými proměnnými. Nejdříve by se mělo zjistit, zda se nejedná o systém nezávislých proměnných. V takovém případě by bylo zbytečné pokračovat v analýze závislostí.
Na hladině významnosti 0,05 testujeme $H_0 : \mathrm{cor}\mathbf{X} = \mathbf{I}$ proti $H_0 : \mathrm{cor}\mathbf{X} \neq \mathbf{I}$ ($\mathbf{I}$ je jednotková matice řádu p).

Testová statistika $T_0 = -n \ln|\mathbf{R}|$ se za platnosti $H_0$ asymptoticky řídí rozložením $\chi^2\left(\frac{p(p-1)}{2}\right)$.

Kritický obor: $W = \left\langle \chi^2_{1-\alpha}\left(\frac{p(p-1)}{2}\right), \infty\right)$

Jestliže $t_0 \in W$, $H_0$ zamítáme na hladině významnosti $\alpha$.

**Poznámka:** Aproximaci $\chi^2$-rozložením můžeme zpřesnit, když testovou statistiku $T_0$ vynásobíme konstantou $1 - \frac{2p+11}{6n}$.

**Příklad:** Na základě dat z příkladu o rozměrech součástek testujte hypotézu, že mezi sledovanými čtyřmi rozměry není žádná závislost.

**Řešení:**

Logaritmus determinantu výběrové korelační matice získáme v systému STATISTICA takto: Statistiky – Vícerozměrné průzkumné techniky – Hlavní komponenty & klasifikační analýza – Proměnné X1, X2, X3, X4 – OK – OK – Popis. statistiky – Korelační matice Inverzní.

| | Inverzní korelační matice (soucastky.sta)<br>Aktivní proměnné<br>Log(Determinant) korelační matice: -,10371221 | | | |
|---|---|---|---|---|
| Proměnná | X1 | X2 | X3 | X4 |
| X1 | 1,051183 | -0,116527 | -0,221930 | 0,034663 |
| X2 | -0,116527 | 1,065407 | 0,229398 | 0,101462 |
| X3 | -0,221930 | 0,229398 | 1,089346 | -0,038291 |
| X4 | 0,034663 | 0,101462 | -0,038291 | 1,014320 |

V záhlaví výstupní tabulky je číslo $\ln|\mathbf{R}|$ = -0,10371221.

K dalším výpočtům použijeme STATISTIKU jako inteligentní kalkulačku.
Otevřeme nový datový soubor o 3 proměnných a 1 případu. Do Dlouhého jména 1. proměnné napíšeme =-0,103712221 (tj. $\ln|\mathbf{R}|$), do Dlouhého jména druhé proměnné napíšeme =-15*v1 (tj. $T_0 = -n \ln|\mathbf{R}|$) a Dlouhého jména třetí proměnné napíšeme =VCHi2(0,95;6) (tj. kvantil $\chi^2_{0,95}(6)$).

| | 1<br>Prom1 | 2<br>Prom2 | 3<br>Prom3 |
|---|---|---|---|
| 1 | -0,1037122 | 1,55568315 | 12,5915872 |

Protože testová statistika 1,5557 nepatří do kritického oboru $\langle 12,5916; \infty)$, hypotézu o úplné nezávislosti čtyř rozměrů součástek nezamítáme na hladině významnosti 0,05.

# II. Úlohy o dvou nezávislých náhodných výběrech z vícerozměrného rozložení

## 1. Test hypotézy o rozdílu vektorů středních hodnot

Tento test je p-rozměrnou analogií dvouvýběrového t-testu. Pro připomenutí:
Náhodný výběr $X_{11},\ldots,X_{1n_1}$ pochází z rozložení $N(\mu_1,\sigma^2)$, na něm nezávislý náhodný výběr $X_{21},\ldots,X_{2n_2}$ pochází z rozložení $N(\mu_2,\sigma^2)$, přičemž parametry $\mu_1,\mu_2,\sigma^2$ neznáme. Označíme $M_1,M_2$ výběrové průměry, $S_1^2,S_2^2$ výběrové rozptyly, $S_*^2=\dfrac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$ vážený průměr výběrových rozptylů. Na hladině významnosti $\alpha$ testujeme hypotézu $H_0:\mu_1=\mu_2$ proti alternativě $H_1:\mu_1\neq\mu_2$.

Testová statistika: $T_0=\dfrac{M_1-M_2}{S_*\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}$ se za platnosti $H_0$ řídí rozložením $t(n_1+n_2-2)$.

Kritický obor: $W=\left(-\infty,-t_{1-\alpha/2}(n_1+n_2-2)\right\rangle\cup\left\langle t_{1-\alpha/2}(n_1+n_2-2),\infty\right)$.

Jestliže $t_0\in W$, $H_0$ zamítáme na hladině významnosti $\alpha$.

**Upozornění:** Předpoklad, že rozptyly obou rozložení jsou shodné (tj. test nulové hypotézy $H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1$ proti alternativě $H_1: \frac{\sigma_1^2}{\sigma_2^2} \neq 1$) ověřujeme F-testem.

Testová statistika $T_0: \frac{S_1^2}{S_2^2}$ se v případě platnosti $H_0$ řídí rozložením $F(n_1 - 1, n_2 - 1)$.

Kritický obor: $W = \langle 0, F_{\alpha/2}(n_1 - 1, n_2 - 1) \rangle \cup \langle F_{1-\alpha/2}(n_1 - 1, n_2 - 1), \infty)$.

Jestliže $t_0 \in W$, $H_0$ zamítáme na hladině významnosti $\alpha$.

p- rozměrný případ (Hotellingův $T^2$ test)

Máme náhodný výběr $\mathbf{X}_{11},\ldots,\mathbf{X}_{1n_1}$ (kde $\mathbf{X}_{1i} = \left(X_{1i1},\ldots,X_{1ip}\right)^T$, $i = 1,\ldots,n_1$) z $N_p(\boldsymbol{\mu}_1,\boldsymbol{\Sigma})$ a dále na něm nezávislý náhodný výběr $\mathbf{X}_{21},\ldots,\mathbf{X}_{2n_2}$ (kde $\mathbf{X}_{2i} = \left(X_{2i1},\ldots,X_{2ip}\right)^T$, $i = 1,\ldots,n_2$) z $N_p(\boldsymbol{\mu}_2,\boldsymbol{\Sigma})$, přičemž parametry $\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\boldsymbol{\Sigma}$ neznáme.

Zavedeme označení:

$n = n_1 + n_2$ … celkový rozsah obou výběrů

$M_{hj} = \frac{1}{n_h}\sum_{i=1}^{n_h} X_{hij}$ … výběrový průměr j-té proměnné v h-tém výběru, $h = 1,2$, $j = 1,\ldots,p$

$\mathbf{M}_h = \left(M_{h1} \quad \ldots \quad M_{hp}\right)^T$ … vektor výběrových průměrů v h-tém výběru, $h = 1,2$

$\mathbf{S}_h = \frac{1}{n_h - 1}\sum_{i=1}^{n_h}\left(\mathbf{X}_{hi} - \mathbf{M}_h\right)\left(\mathbf{X}_{hi} - \mathbf{M}_h\right)^T$ … výběrová varianční matice v h-tém výběru, $h = 1,2$

$\mathbf{S} = \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{n - 2}$ … společná výběrová varianční matice

Na hladině významnosti $\alpha$ testujeme hypotézu $H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ proti alternativě $H_1: \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$.

Statistika $\frac{n_1 n_2}{n}(\mathbf{M}_1 - \mathbf{M}_2)^T \mathbf{S}^{-1}(\mathbf{M}_1 - \mathbf{M}_2)$ se řídí Hotellingovým rozložením $T^2(p, n-2)$, když $H_0$ platí.

Vzhledem ke vztahu mezi Hotellingovým a F-S rozložením vynásobíme tuto statistiku konstantou $\frac{n-p-1}{p(n-2)}$ a získáme testovou statistiku:

$T_0 = \frac{n-p-1}{p(n-2)} \cdot \frac{n_1 n_2}{n}(\mathbf{M}_1 - \mathbf{M}_2)^T \mathbf{S}^{-1}(\mathbf{M}_1 - \mathbf{M}_2)$, která se za platnosti $H_0$ řídí rozložením $F(p, n-p-1)$.

Kritický obor: $W = \langle F_{1-\alpha}(p, n-p-1), \infty)$.

Jestliže $t_0 \in W$, $H_0$ zamítáme na hladině významnosti $\alpha$.

## 2. Test hypotézy o shodě varianční matic

Předpoklad o shodě varianční matic můžeme ověřit pomocí Boxova M-testu.
Na hladině významnosti $\alpha$ testujeme hypotézu $H_0 : \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ proti alternativě $H_1 : \mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$.

Testová statistika má tvar: $T_0 = \frac{1}{C_p}\left[(n-2)\ln|\mathbf{S}| - (n_1 - 1)\ln|\mathbf{S}_1| - (n_2 - 1)\ln|\mathbf{S}_2|\right]$, kde

$C_p = 1 + \frac{2p^2 + 3p - 1}{6(p+1)}\left(\frac{1}{n_1 - 1} + \frac{1}{n_2 - 1} - \frac{1}{n-2}\right)$ je konstanta zlepšující aproximaci.

V případě platnosti $H_0$ se statistika $T_0$ asymptoticky řídí rozložením $\chi^2\left(\frac{p(p+1)}{2}\right)$. Pokud

$t_0 \in \left\langle \chi^2_{1-\alpha}\left(\frac{p(p+1)}{2}\right), \infty\right)$, hypotézu o shodě varianční matic zamítneme na asymptotické

hladině významnosti $\alpha$. Aproximace je vyhovující, když rozsahy výběrů jsou aspoň 20 a počet proměnných je nejvýše 5.

V případě, že rozsahy výběrů jsou shodné, nemusíme Boxův test provádět.

Simultánní t-testy:

Pokud na dané hladině významnosti $\alpha$ zamítneme hypotézu $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ ve prospěch alternativy $H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$, zjistíme, které proměnné jsou příčinou jejího zamítnutí.
V této situaci provedeme p simultánních testů $H_{0j} : \mu_{1j} = \mu_{2j}$ proti $H_{1j} : \mu_{1j} \neq \mu_{2j}$, $j = 1, \ldots, p$ pomocí testové statistiky $T_{0j} = \frac{n-p-1}{p(n-2)} \cdot \frac{n_1 n_2}{n} \cdot \frac{(M_{1j} - M_{2j})^2}{S_{*j}^2}$, která se za platnosti $H_{0j}$ řídí rozložením $F(p, n-p-1)$.

Kritický obor: $W = \langle F_{1-\alpha}(p, n-p-1), \infty)$.
Jestliže $t_{0j} \in W$, $H_{0j}$ zamítáme na hladině významnosti $\alpha$.

**Příklad na Hotellingův $T^2$ test**

23 náhodně vybraných mužů a 22 náhodně vybraných žen mělo posoudit podobné výrobky od tří firem – označme je A, B, C – na škále 0 bodů (naprosto nevyhovující) až 10 bodů (zcela vyhovující). Výsledky jsou uloženy v souboru hodnoceni_vyrobku.sta.
Za předpokladu, že data tvoří realizace dvou nezávislých náhodných výběrů ze dvou třírozměrných normálních rozložení se stejnými variančními maticemi, Hotellingovým $T^2$ testem ověřte na hladině významnosti 0,05 hypotézu, že hodnocení mužů a žen se neliší. Pokud dojde k zamítnutí nulové hypotézy, zjistěte, které firmy se v hodnocení mužů a žen liší.

**Řešení:**

Na hladině významnosti 0,05 testujeme hypotézu

$H_0$: $\begin{pmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{13} \end{pmatrix} = \begin{pmatrix} \mu_{21} \\ \mu_{22} \\ \mu_{23} \end{pmatrix}$ proti alternativě $H_1$: $\begin{pmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{13} \end{pmatrix} \neq \begin{pmatrix} \mu_{21} \\ \mu_{22} \\ \mu_{23} \end{pmatrix}$.

Hodnotu testové statistiky $T_0 = \frac{n-p-1}{p(n-2)} \cdot \frac{n_1 n_2}{n} (\mathbf{M}_1 - \mathbf{M}_2)^T \mathbf{S}^{-1} (\mathbf{M}_1 - \mathbf{M}_2)$ a odpovídající p-hodnotu vypočteme pomocí systému STATISTICA.

Statistiky – Základní statistiky a tabulky – t-test, nezávislé, dle skupin – OK – Proměnné – Závisle proměnné X1, X2, X3, Grupovací proměnná ID – OK – na záložce Možnosti zaškrtneme Vícerozměrný test (Hotellingovo $T^2$) – Výpočet

| | t-testy; grupováno: ID: pohlaví respondenta (hodnoceni_vyrobku.sta)<br>Skup. 1: muž; Skup. 2: žena<br>Hotellingovo 15,5599 F(3,41)=4,9454 p<,00506 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Proměnná | Průměr muž | Průměr žena | t | sv | p | Poč.plat muž | Poč.plat. žena | Sm.odch. muž | Sm.odch. žena | F-poměr Rozptyly | p Rozptyly |
| X1 | 5,086957 | 4,545455 | 0,697666 | 43 | 0,489142 | 23 | 22 | 2,574579 | 2,631807 | 1,044950 | 0,917081 |
| X2 | 5,434783 | 3,818182 | 2,098562 | 43 | 0,041766 | 23 | 22 | 2,642762 | 2,519190 | 1,100510 | 0,829044 |
| X3 | 5,304348 | 3,045455 | 3,117687 | 43 | 0,003246 | 23 | 22 | 2,770540 | 2,011332 | 1,897411 | 0,147512 |

Testová statistika Hotellingova testu nabývá hodnoty 4,9454, odpovídající p-hodnota je menší než 0,00506, tedy na hladině významnosti 0,05 zamítáme hypotézu, že vektory středních hodnot proměnných X1, X2, X3 jsou v obou skupinách shodné. S rizikem omylu nejvýše 5 % jsme tedy prokázali, že mezi muži a ženami existuje rozdíl v hodnocení výrobků tří firem. (Vidíme, že hodnocení mužů je příznivější než hodnocení žen.)

Nyní pomocí simultánních testů zjistíme, které firmy jsou rozdílně hodnoceny muži a ženami.

Simultánní testy založené na statistice $T_{0j} = \frac{n-p-1}{p(n-2)} \cdot \frac{n_1 n_2}{n} \cdot \frac{(M_{1j} - M_{2j})^2}{S_{*j}^2}$ STATISTICA neposkytuje. (V našem případě n = 45, p = 3, $n_1$ = 23, $n_2$ = 22, tedy $\frac{n-p-1}{p(n-2)} \cdot \frac{n_1 n_2}{n} = \frac{20746}{5805}$ .)

S pomocí STATISTIKY však můžeme vypočítat vektory výběrových průměrů a směrodatných odchylek. V této tabulce ponecháme pouze proměnné obsahující průměry a směrodatné odchylky. Dále za poslední proměnnou vložíme dvě nové proměnné $T_{0j}$ a kvantil. Do Dlouhého jména proměnné $T_{0j}$ napíšeme:

=(20746/5805)*(v1-v2)^2/((22*v3^2+21*v4^2)/45)

Do Dlouhého jména proměnné kvantil napíšeme:

=VF(0,95;3;41)

| Proměnná | t-testy; grupováno: ID: pohlaví respondenta (hodnoceni_vyrobku.sta)<br>Skup. 1: muž; Skup. 2: žena<br>Hotellingovo 15,5599 F(3,41)=4,9454 p<,00506 | | | | | |
|---|---|---|---|---|---|---|
| | Průměr muž | Průměr žena | Sm.odch. muž | Sm.odch. žena | T0j =(20746/58 | kvantil =VF(0,95;3; |
| X1 | 5,086957 | 4,545455 | 2,574579 | 2,631807 | 0,161895 | 2,832747 |
| X2 | 5,434783 | 3,818182 | 2,642762 | 2,519190 | 1,464812 | 2,832747 |
| X3 | 5,304348 | 3,045455 | 2,770540 | 2,011332 | 3,232981 | 2,832747 |

Vidíme, že statistika T03 se realizuje v kritickém oboru $W = \langle 2{,}8327; \infty)$. S rizikem omylu nejvýše 5 % jsme tedy prokázali, že výrobky firmy C jsou odlišně hodnoceny muži a ženami.