Část III. – Náhodný vektor

Náhodný vektor

- **Náhodným vektorem** budeme rozumět sloupcový vektor složený z náhodných veličin $X = (X_1, X_2, \ldots, X_n)^T$.
- Pro jednoduchost se omezíme pouze na dvousložkový náhodný vektor.
- Např. u osobních automobilů můžeme sledovat dvě proměnné – výkon motoru a maximální rychlost.

Náhodný vektor

Průzkum spokojenosti cestujících s MHD	
Hodnocení ceny jízdného	Hodnocení kvality cestování
1	1
1	2
1	3
2	2
2	3
2	3
3	2
3	2
3	3
3	3
3	3
3	3
3	3
3	4
3	4
4	3
4	3
4	4
4	4
4	4

Legenda	
1	Velmi nespokojen
2	Nespokojen
3	Spokojen
4	Velmi spokojen

Ukázka diskrétního náhodného vektoru – při průzkumu spokojenosti cestujících se systémem MHD nás zajímá více údajů, např. viz tabulka.

Náhodný vektor

- **Sdružená (simultánní) distribuční funkce** náhodných veličin X a Y je definována vztahem:

  $F(x,y) = P(X < x; Y < y)$.

- Vlastnosti sdružené distribuční funkce:

  1. $0 \le F(x,y) \le 1.$
  2. $F(x,-\infty) = F(-\infty, y) = F(-\infty,-\infty) = 0; F(\infty,\infty) = 1.$

Náhodný vektor

3. Funkce *F*(*x*,*y*) je neklesající v každé proměnné.
4. Funkce *F*(*x*,*y*) je zleva spojitá v každé proměnné.
5. $P\left(a_1 \leq X < b_1; a_2 \leq Y < b_2\right) = F\left(b_1; b_2\right) - F\left(b_1; a_2\right) - F\left(a_1; b_2\right) + F\left(a_1; a_2\right).$

Náhodný vektor

- Podle oboru hodnot, kterých můžou proměnné X a Y nabývat, rozeznáváme:
  - Náhodný vektor s diskrétním rozdělením.

$$F(x,y)=\sum_{x_i<x}\sum_{y_j<y}\underbrace{P\left(X=x_i;Y=y_j\right)}_{\text{Sdružená pravděpodobnostní funkce}}$$

  - Náhodný vektor se spojitým rozdělením.

$$F(x,y)=\int_{-\infty}^{x}\int_{-\infty}^{y}\underbrace{f(r,s)dr\,ds}_{\text{Sdružená hustota pravděpodobnosti}}$$

Náhodný vektor

- Pro sdruženou pravděpodobnostní funkci diskrétního náhodného vektoru dále platí:

1. $\sum_i \sum_j P(X = x_i, Y = y_j) = 1.$

2. $0 \le P(X = x_i, Y = y_j) \le 1.$

Náhodný vektor

- Pro sdruženou hustotu pravděpodobnosti spojitého náhodného vektoru dále platí:

1. $f(x, y) \geq 1.$

2. $\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} f(x, y)dxdy = 1.$

3. $P(a_1 \leq X < b_1; a_2 \leq Y < b_2) = \int\limits_{a_1}^{b_1}\int\limits_{a_2}^{b_2} f(x, y)dxdy.$

Náhodný vektor

- **Marginální distribuční funkcí** rozumíme distribuční funkci složky X nebo Y.

$$F_X(x) = P(X < x) = \lim_{y \to \infty} F(x, y)$$

$$F_Y(y) = P(Y < y) = \lim_{x \to \infty} F(x, y)$$

- Pro diskrétní náhodný vektor potom platí:

$$P_X(x) = \sum_{y_j} P(X = x; Y = y_j),$$

$$P_Y(Y) = \sum_{x_i} P(X = x_i; Y = y).$$

Marginální pravděpodobnostní funkce

Náhodný vektor

- Pro spojitý náhodný vektor potom platí:

$$\left.\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f(x,y)\,dy, \\ f_Y(y) &= \int_{-\infty}^{\infty} f(x,y)\,dx. \end{aligned}\right\}$$

Marginální hustoty pravděpodobnosti

Náhodný vektor

- Nezávislost složek náhodného vektoru:
  - Složky X a Y náhodného vektoru jsou navzájem nezávislé, jsou-li nezávislé náhodné proměnné X a Y, tedy:

    $$F(x, y) = F_X(x) \cdot F_Y(y).$$

  - Náhodný vektor s diskrétním rozdělením:

    $$P(X = x_i, Y = y_j) = P_X(X = x_i) \cdot P_Y(Y = y_j).$$

  - Náhodný vektor se spojitým rozdělením:

    $$f(x, y) = f_X(x) \cdot f_Y(y).$$

Náhodný vektor

- Sdružená pravděpodobnostní funkce diskrétního dvousložkové náhodného vektoru se často zobrazuje v podobě **korelační tabulky**.

Náhodný vektor

X / Y	y_1	y_2	$\cdots\cdots$	y_n	$\sum_{j=1}^{n} P(X = x_i; Y = y_j)$
x_1	$P(X = x_1; Y = y_1)$	$P(X = x_1; Y = y_2)$		$P(X = x_1; Y = y_n)$	$P_X(x_1)$
x_2	$P(X = x_2; Y = y_1)$	$P(X = x_2; Y = y_2)$		$P(X = x_2; Y = y_n)$	$P_X(x_2)$
$\vdots$					$\vdots$
x_m	$P(X = x_m; Y = y_1)$	$P(X = x_m; Y = y_2)$		$P(X = x_m; Y = y_n)$	$P_X(x_m)$
$\sum_{i=1}^{m} P(X = x_i; Y = y_j)$	$P_Y(y_1)$	$P_Y(y_2)$	$\cdots\cdots$	$P_Y(y_n)$	$\sum_{i=1}^{m} P_X(x_i) = \sum_{j=1}^{n} P_Y(y_j) = 1$

Náhodný vektor

- Podmíněné rozdělení pravděpodobnosti:
  - Pro náhodný vektor s diskrétním rozdělením je definována podmíněná pravděpodobnostní funkce ve tvaru:

$$P\left(X = x \mid Y = y\right) = \frac{P\left(X = x, Y = y\right)}{P_Y\left(Y = y\right)} \text{ pro } P_Y\left(Y = y\right) \neq 0,$$

$$P\left(Y = y \mid X = x\right) = \frac{P\left(X = x, Y = y\right)}{P_X\left(X = x\right)} \text{ pro } P_X\left(X = x\right) \neq 0.$$

Náhodný vektor

- Podmíněné rozdělení pravděpodobnosti:
  - Pro náhodný vektor se spojitým rozdělením je definována podmíněná hustota pravděpodobnosti:

$$f(x|y) = \frac{f(x,y)}{f_Y(y)} \text{ pro } f_Y(y) \neq 0,$$

$$f(y|x) = \frac{f(x,y)}{f_X(x)} \text{ pro } f_X(x) \neq 0.$$

Náhodný vektor

- Číselné charakteristiky náhodného vektoru:
  - **Smíšený počáteční (obecný) moment řádu *k,n***:
    $\mu_{k,n} = E\left(X^k \cdot Y^n\right).$
  - Pro diskrétní náhodný vektor potom platí:
    $$\mu_{k,n} = \sum_{i=1}^{m} \sum_{j=1}^{n} x_i^k \cdot y_j^n \cdot P\left(X = x_i; Y = y_j\right).$$
  - Pro spojitý náhodný vektor platí:
    $$\mu_{k,n} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^k \cdot y^n \cdot f\left(x, y\right) dx\, dy.$$

Náhodný vektor

– **Smíšený centrální moment řádu *k,n***:

$$\nu_{k,n} = E\left[(X - EX)^k \cdot (Y - EY)^n\right].$$

– Pro diskrétní náhodný vektor potom platí:

$$\nu_{k,n} = \sum_{i=1}^{m} \sum_{j=1}^{n} (x_i - EX)^k \cdot (y_j - EY)^n \cdot P(X = x_i, Y = y_j).$$

– Pro spojitý náhodný vektor platí:

$$\nu_{k,n} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - EX)^k \cdot (y - EY)^n \cdot f(x, y)\, dx\, dy.$$

Náhodný vektor

- Marginální číselné charakteristiky složek náhodného vektoru:
  - Střední hodnota proměnné X:

$$EX = \mu_{1,0} = \sum_{i=1}^{m} x_i \cdot P_X(x_i) -$$ pro diskrétní proměnnou.

$$EX = \mu_{1,0} = \int_{-\infty}^{\infty} x \cdot f_X(x)dx -$$ pro spojitou proměnnou.

- Obdobně bychom definovali vztahy pro proměnnou Y.

Náhodný vektor

– Rozptyl proměnné X:

$$DX = \nu_{2,0} = \sum_{i=1}^{m} \left(x_i - EX\right)^2 \cdot P_X\left(x_i\right)$$ – pro diskrétní

proměnnou.

$$DX = \nu_{2,0} = \int_{-\infty}^{\infty} \left(x - EX\right)^2 \cdot f_X\left(x\right) dx$$ – pro spojitou

proměnnou.

- Obdobně bychom definovali vztahy pro proměnnou Y.

Náhodný vektor

- Nejdůležitější číselnou charakteristikou náhodného vektoru je **kovariance** – smíšený centrální moment řádu 1,1:

$$Cov(X,Y)=\nu_{1,1}=E\left[(X-EX)\cdot(Y-EY)\right].$$

  – Pro diskrétní náhodný vektor platí:

$$Cov(X,Y)=\sum_{i=1}^{m}\sum_{j=1}^{n}(x_i-EX)\cdot(y_j-EY)\cdot P(X=x_i,Y=y_j).$$

  – Pro spojitý náhodný vektor platí:

$$Cov(X,Y)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-EX)\cdot(y-EY)\cdot f(x,y)\,dx\,dy.$$

Náhodný vektor

- Vlastnosti kovariance náhodného vektoru:
  1. $Cov(X,Y) = E(X \cdot Y) - EX \cdot EY$,
  2. $Cov(X,X) = DX$, $Cov(Y,Y) = DY$,
  3. $Cov(X,Y) = Cov(Y,X)$,
  4. Jsou-li složky X a Y náhodného vektoru nezávislé, potom platí: $Cov(X,Y) = 0$.

Náhodný vektor

- Kovariance představuje nejjednodušší ukazatel souvislosti dvou náhodných veličin.
- Kladná hodnota kovariance znamená, že s rostoucí proměnnou X roste i proměnná Y.
- Záporná hodnota kovariance znamená, že s rostoucím X klesá Y.
- **Kovarianční matice** – $\begin{vmatrix} DX & Cov(X,Y) \\ Cov(X,Y) & DY \end{vmatrix}$.

Náhodný vektor

- **Jednoduchý korelační koeficient** – vyjadřuje míru **lineární** závislosti (korelaci) mezi složkami X a Y náhodného vektoru, je definován:

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sqrt{DX \cdot DY}} \text{ pro } DX, DY \neq 0$$

Pro jednoduchý korelační koeficient platí:

1. $\rho_{X,Y} \in \langle -1;1 \rangle$.
2. $\rho_{X,Y} = \rho_{Y,X}$.

Náhodný vektor

- Pokud $\rho_{X,Y} = 0$, proměnné X a Y náhodného vektoru jsou nekorelované (lineárně nezávislé).
- Pokud $\rho_{X,Y} > 0$, proměnné X a Y náhodného vektoru jsou pozitivně korelované.
- Pokud $\rho_{X,Y} < 0$, proměnné X a Y náhodného vektoru jsou negativně korelované.

Náhodný vektor

- Je-li potřeba testovat vzájemnou nezávislost složek náhodného vektoru X a Y, kde X a Y jsou kategoriální proměnné, můžeme použít $\boldsymbol{\chi}^2$ **test nezávislosti v kombinační (kontingenční) tabulce**.
- Kategoriální proměnná je proměnná, kterou nemůžeme měřit, ale pouze zařadit do tříd (např. známka ve škole apod.). Nechť složka X nabývá *k* různých variant, složka Y *m* různých variant.
- Kombinační tabulka je v podstatě tabulka sdružených a marginálních četností.

Náhodný vektor

X / Y	y_1	y_2	. . .	y_m	$\sum$
x_1	n_{11}	n_{12}		n_{1m}	$n_{1\bullet}$
x_2	n_{21}	n_{22}		n_{2m}	$n_{2\bullet}$
$\vdots$					
x_k	n_{k1}	n_{k2}		n_{km}	$n_{k\bullet}$
$\sum$	$n_{\bullet 1}$	$n_{\bullet 2}$		$n_{\bullet m}$	n

Náhodný vektor

- Test je založen na porovnání empirických (pozorovaných) četností s četnostmi teoretickými.
- V případě nezávislosti platí:

$$n_{ij}^{*} = \left(\frac{n_{i\bullet}}{n} \cdot \frac{n_{\bullet j}}{n} \right) \cdot n = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n},$$

kde n_{ij}^{*} je příslušná teoretická četnost.

Náhodný vektor

1. Volba nulové a alternativní hypotézy:

   H_0 – Náhodné veličiny (znaky) X a Y v kombinační tabulce jsou nezávislé.

   H_1 – Náhodné veličiny (složky náhodného vektoru) X a Y v kombinační tabulce jsou závislé.

Náhodný vektor

2. Volba testového kritéria a jeho nulového rozdělení (za předpokladu platnosti nulové hypotézy):

$$K = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{\left(n_{ij} - n_{ij}^{*}\right)^2}{n_{ij}^{*}} \rightarrow \chi^2_{(k-1)\cdot(m-1)} .$$

- S rostoucí hodnotou testové statistiky K roste rozpor naměřených dat s nulovou hypotézou.

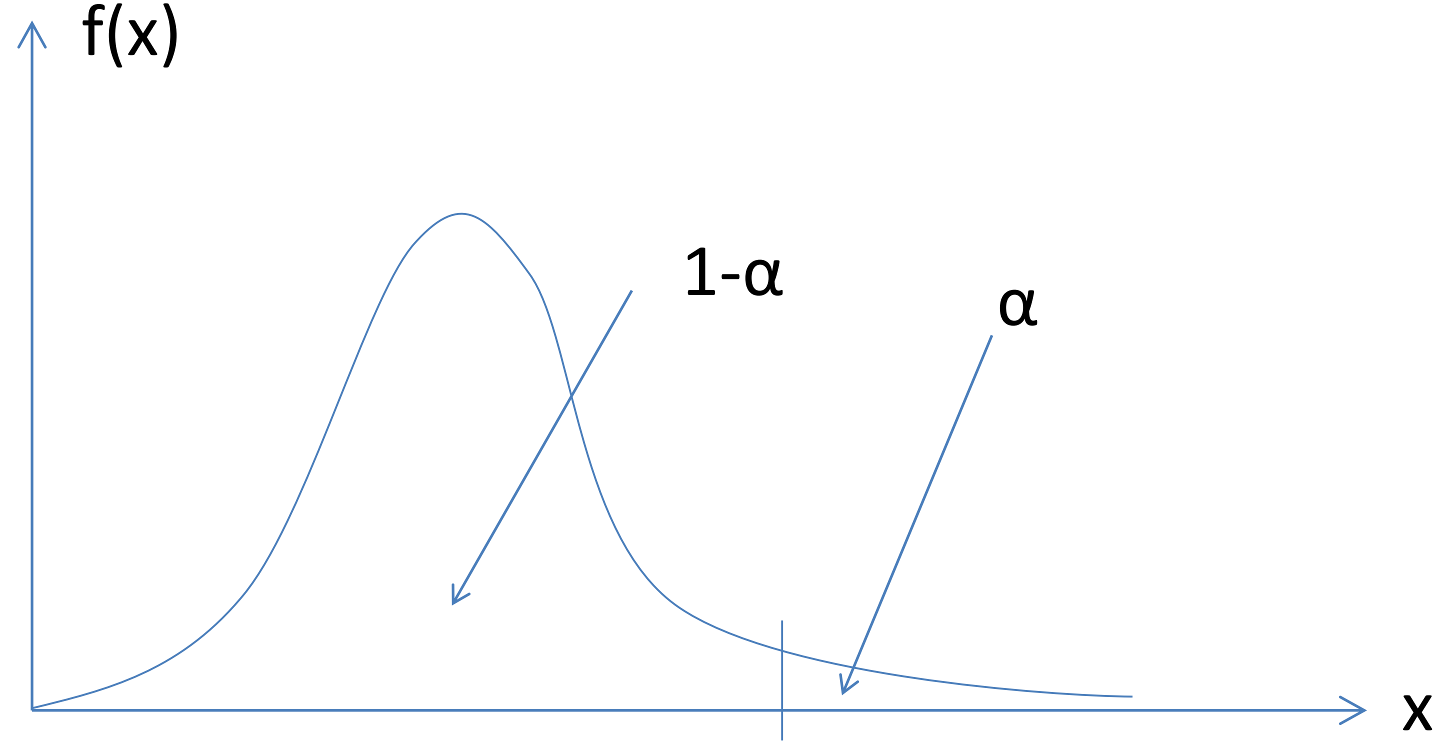
Náhodný vektor

- Předpoklad testu:
  - Žádná teoretická četnost nesmí být menší než 2
  - Alespoň 80% teoretických četností musí být větší než 5.

3. Sestavení oboru přijetí a kritického oboru:

$$x_{krit} = \chi^2_{(1-\alpha);(k-1)\cdot(m-1)} = CHIINV\left[\alpha;(k-1)\cdot(m-1)\right].$$

Náhodný vektor



x_{krit}

Obor přijetí - Oblast nezamítnutí H_0

Kritický obor - Oblast zamítnutí H_0

Náhodný vektor

4. Výpočet pozorované hodnoty testové statistiky x_{obs} (za předpokladu platnosti nulové hypotézy).
5. Formulace závěru testu:
   - $x_{obs} < x_{krit}$ – nezamítáme nulovou hypotézu o nezávislosti náhodných proměnných.
   - $x_{obs} \geq x_{krit}$ – zamítáme nulovou hypotézu ve prospěch alternativní hypotézy.

Náhodný vektor

- V případě, že není splněn předpoklad výše uvedeného testu, lze použít tzv. Yatesovu korekci, která snižuje hodnotu testového kritéria, je tedy obtížnější zamítnout nulovou hypotézu. Testová statistika je v tomto případě ve tvaru:

$$K_{Yates} = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{\left(n_{ij} - n_{ij}^{*} - 0,5\right)^2}{n_{ij}^{*}} \rightarrow \chi^2_{(k-1)\cdot(m-1)}$$

Náhodný vektor

- Uvedený test pouze slouží k rozhodnutí o tom, zda jsou znaky X a Y závislé či nezávislé, ale již nestanovuje, o jak silnou závislost se jedná.
- Pro stanovení síly závislosti zavádíme různé koeficienty, jež jsou obdobou korelačního koeficientu.

Náhodný vektor

- Zavádíme koeficient kontingence definovaný:

$$CC = \sqrt{\frac{K}{K+n}},$$

kde *K* je hodnota testového kritéria a *n* rozsah výběru. Pro čtvercovou kombinační tabulku (*k*=*m*) nabývá hodnot $\langle 0;1 \rangle$.

Náhodný vektor

- Pro obdélníkovou kombinační tabulku je ovšem maximální hodnota koeficientu kontingence rovna:
  $$CC_{\max} = \sqrt{\frac{\min(k,m)-1}{\min(k,m)}},$$
  proto zavádíme korigovaný koeficient kontingence ve tvaru:
  $$CC_{cor} = \frac{CC}{CC_{\max}},$$
  který je už v intervalu $\langle 0;1 \rangle$.

Náhodný vektor

- Jako další míra síly závislosti se používá Cramerův koeficient definovaný ve tvaru:

$$V=\sqrt{\frac{K}{n\cdot\left[\min\left(k,m\right)-1\right]}},$$

jež nabývá hodnot z intervalu $\langle 0;1\rangle$.