

# Statistika I / Statistika A

## Obsah

---

- Explorační analýza dat
- Teorie pravděpodobnosti
- Náhodná veličina
  - Číselné charakteristiky náhodné veličiny
- Náhodný vektor
  - Charakteristiky náhodného vektoru
- Diskrétní rozdělení pravděpodobnosti
  - Hypergeometrická náhodná veličina
  - Binomická náhodná veličina
  - Geometrická náhodná veličina
  - Negativně binomická náhodná veličina
  - Poissonovo rozdělení pravděpodobnosti
- Spojité rozdělení pravděpodobnosti
  - Rovnoměrné rozložení
  - Exponenciální rozdělení
  - Erlangovo rozdělení
  - Weibullovo rozdělení
  - Normální rozdělení
    - Normované normální rozdělení
  - $\chi^2$  rozdělení
  - Studentovo rozdělení
  - Fisherovo-Snedecorovo rozdělení
- Limitní věty
  - Centrální limitní věta
- Náhodné výběry a jejich zpracování
  - Teorie odhadu
    - Intervalový odhad
  - Testování hypotéz
  - ANOVA Analýza rozptylu
- Regresní analýza
  - Obecný lineární model

- Lineární regrese s jednou vysvětlující proměnnou

# Explorační analýza dat

Data představují výsledky **datově generačního procesu** – z množiny měřených objektů (domain) vybíráme proměnné měřených veličin. Množina měřených hodnot musí být vyčerpávající a vzájemně vylučující. Data vybírám z několika charakteristických typů:

- **Kvalitativní proměnná** – nabývá z předem daných hodnot, dělíme na **nominální** (má smysl hodnotu dané kategorie pojmenovat, k popisu slouží **četnost** proměnné) a **ordinální** (má smysl pořadí hodnoty dané kategorie). Dále se dělí na **alternativní** (vlastnosti, atributy, nabývají jedné ze dvou hodnot) nebo **množné**.
- **Kvantitativní proměnná** – nabývá hodnoty z množiny  $\mathbb{R}$ . Mohou být **diskrétní** (nabývají diskrétních hodnot v **konečném** počtu nebo **spočteném** počtu) nebo **spojité**.

Nominální kvalitativní proměnná nabývají absolutní četnosti  $n_i$ , přičemž platí  $\sum n_i = n$ . Relativní četnost  $p_i = \frac{n_i}{n}$ , přičemž  $\sum p_i = 1$ . Definujeme **modus** jako název varianty proměnné vykazující nejvyšší četnost. **Histogram** je klasickým grafem, v němž na jednu osu vynášíme varianty a na druhou jejich četnost. **Výsečový graf** prezentuje relativní četnosti jednotlivých variant pomocí plochami kruhových výsečí.

Ordinální kvalitativní proměnná využívá pro popis stejné charakteristiky jako pro popis nominální proměnné. **Kumulativní četnost**  $m_i$  definujeme jako počet hodnot proměnné, které nabývají varianty nižší nebo rovné dané variantě. Pokud  $x_1 < \dots < x_n$ , platí  $m_i = \sum_{j=1}^i n_j$ . **Kumulativní relativní četnost**  $F_i = \frac{m_i}{n}$ . **Polygon kumulativních četností** je spojnicovým grafem, v němž se na vodorovnou osu vynáší jednotlivé varianty v pořadí od „nejmenší“ do „největší“ a na svislou osu nanášíme kumulativní četnosti. **Paretův graf** je často užívaným grafem spojením histogramu a polygonu kumulativních četností, v němž na vodorovnou osu vynášíme v pořadí od „největšího významu“ po „nejmenší význam“.

Kvantitativní proměnné využívá stejné charakteristiky jako pro popis ordinální proměnné. Definujeme *míry polohy* určující typické rozložení hodnot proměnné a *míry variability* určující variabilitu hodnot kolem své typické polohy. **Aritmetický průměr** je mírou polohy  $\bar{x} = \frac{\sum x_i}{n}$ . **Modus pro diskrétní proměnnou** jako hodnotu nejčastější varianty proměnné. **Modus pro spojitě proměnné** považujeme za modus hodnotu, kolem níž je největší koncentrace hodnot proměnné. Pro určení hodnoty využijeme **shorth**, což je nejkratší interval, v němž leží alespoň 50 % hodnot proměnné. Modus  $\hat{x}$  definujeme jako střed shorthu.

**Kvantily**  $x_p$  jsou statistiky, které charakterizují polohu jednotlivých hodnot v rámci proměnné. Rozdělují datový soubor na dvě části - 100p% a zbytek

- **Dolní kvartil**  $x_{0,25}$  rozděluje datový soubor tak, že 25 % hodnot je menších než tento kvartil a zbytek, tj. 75 % větších nebo rovných.
- **Medián**  $x_{0,5}$  rozděluje datový soubor tak, že 50 % hodnot je menších než medián a zbytek, tj. 50 % větších nebo rovných.

- **Horní kvartil**  $x_{0.75}$  rozděluje datový soubor tak, že 75 % hodnot je menších než tento kvartil a zbytek, tj. 25 % větších nebo rovných.
- **Decily** rozdělují výběrový soubor na 10 stejně četných částí.
- **Percentily** dělí výběrový soubor na 100 stejně četných částí.

Lze říci, že hodnota  $p$  udává kumulativní relativní četnost kvantilu  $x_p$ . Kvantil a kumulativní relativní četnost jsou tedy inverzní hodnoty.

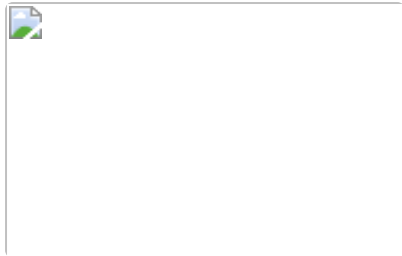
**Empirická distribuční funkce  $F(x)$  pro kvantitativní proměnnou** – označme si  $p(x_i)$  relativní četnost hodnoty  $x_i$ . Poté platí, že  $F(x) = \sum p(x_i)$ . **Interkvalitové rozpětí IQR** je mírou variability souboru a je definována jako vzdálenost mezi horním a dolním kvantilem  $IQR = x_{0.75} - x_{0.25}$ . **Median Absolute Deviation from median** (MAD) jakožto charakteristikou rozptýlenosti.

1. Výběrový soubor uspořádáme podle velikosti.
2. Určíme medián souboru  $x_{0.5}$ .
3. Pro každou hodnotu souboru určíme absolutní hodnotu její odchylky od mediánu  $|y_i - x_{0.5}|$ .
4. Absolutní odchylky od mediánu uspořádáme podle velikosti.
5. Určíme medián absolutních odchylek od mediánu, tj. MAD.

Mezi charakteristiky rozptýlenosti patří dále **výběrový rozptyl**  $s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$  a **směrodatná odchylka**  $s = \sqrt{s^2}$ . Odlehlou hodnotou **outlier** nazýváme hodnotu, která svou charakteristikou nepatří do datového souboru. Existují tři detekce outlier hodnot:

1. Za odlehlé pozorování lze považovat takovou hodnotu, jejíž absolutní hodnota **z-souřadnice** je větší než 3:  $z_i = \frac{x_i - \bar{x}}{s}$ . Z-souřadnici můžeme interpretovat jako počet směrodatných odchylek, o kolik se hodnota liší od průměru.
2. Za odlehlé pozorování lze považovat takovou hodnotu, jejíž absolutní hodnota **mediánové souřadnice** je větší než 3:  $m_i = \frac{x_i - x_{0.5}}{1.483 \cdot MAD}$ . Mediánová metoda je vhodnější než z-souřadnice díky menší závislosti na okrajových hodnotách.

Definujme čísla se specifickým významem: **k-tý obecný výběrový moment** definujeme jako  $m'_k = \frac{1}{n} \sum (x_i)^k$ ,  $m'_0 = 1$ ,  $m'_1 = \bar{x}$  a **k-tý centrální výběrový moment** definujeme jako  $m_k = \frac{1}{n} \sum (x_i - \bar{x})^k$ ,  $m_0 = 1$ ,  $m_1 = 0$ ,  $m_2 = s_0^2 = \frac{1}{n} \sum (x_i - \bar{x})^2$ . **Výběrová šikmost** vyjadřuje asymetrii rozložení hodnot kolem jejího průměru  $\alpha = \frac{m_3}{s_0^3} = \frac{1}{n \cdot s^3} \cdot \sum (x_i - \bar{x})^3$ . Interpretujme: pokud  $\alpha = 0$ , tak jsou hodnoty proměnné kolem jejího průměru rozloženy symetricky; pokud  $\alpha > 0$ , tak u proměnné převažují hodnoty menší než průměr a pokud  $\alpha < 0$ , tak u proměnné převažují hodnoty větší než průměr. **Výběrová špičatost**  $\beta = \frac{m_4}{s_0^4} - 3 = \frac{1}{n} \cdot \frac{\sum (x_i - \bar{x})^4}{s^4} - 3$  vyjadřuje podobnost rozdělení k normálnímu rozdělení. Interpretujme: pokud  $\beta = 0$ , tak špičatost odpovídá normálnímu rozdělení; pokud  $\beta > 0$ , tak je proměnná rozdělena špičatě a pokud  $\beta < 0$ , tak je proměnná rozdělena plošně.



Kvalitativní proměnné vizualizujeme pomocí **box-and-whiskers** grafu, který reprezentuje minimum, dolní kvartil, medián, horní kvartil a maximum. Často se využívá s **histogramem četnosti** dělící datový soubor na třídy stejné délky a různé četnosti. **Číslicový histogram (stem and leaf plot)** dělí datový soubor na třídy stejné délky, v rámci každé třídy na lodyze máme listy určující jednotlivé položky v dané třídě.

## Teorie pravděpodobnosti

**Pokus** je konečný děj, který probíhá při určitém souboru fyzikálních podmínek. **Náhodný pokus** je takový pokus, jehož výsledek je náhodný při konstantních podmínkách. **Hromadný pokus** je pokus, který můžeme libovolněkrát opakovat při konstantních podmínkách. Výsledky pokusů musí být neslučitelné (k dvěma různým výsledkům nemůže dojít současně) a vyčerpávající (k nějakému výsledku dojít musí) – množinu všech výsledků nazýváme  $\Omega \neq \emptyset$  **základní prostor**. Jednoprvkové podmnožiny  $\omega \subset \Omega$  nazýváme **elementární jev**. Libovolné podmnožiny  $A \subset \Omega$  nazýváme **jevy**. **Jev nemožný**  $\emptyset$  nemůže nastat za žádných okolností. **Jev jistý**  $\Omega$  nastane při každé realizaci náhodného pokusu.

**Jevové pole**  $\mathcal{S}$  je systém podmnožin, pro který platí  $A \in \mathcal{S} \Rightarrow \overline{A} \in \mathcal{S}$  (systém je uzavřený vůči svým doplňkům) a  $(A_1, \dots), A_i \in \mathcal{S} \Rightarrow \bigcup A_i \in \mathcal{S}$ . Elementy jevového pole nazýváme **náhodnými jevy**. Uspořádaná trojice  $(\Omega, \mathcal{S}, P)$  tvoří **pravděpodobnostní prostor** náhodného pokusu, kde **pravděpodobnostní funkce**  $P: \mathcal{S} \rightarrow \mathbb{R}$  splňuje  $A \in \mathcal{S} : P(A) \geq 0, P(\Omega) = 1$  a  $(A_1, \dots), A_i \in \mathcal{S}, A_i \cap A_j = \emptyset : P(\bigcup A_i) = \sum P(A_i)$  (tzv. sigmaaditivita).

- $A, B \in \mathcal{S}, A \subset B \Rightarrow P(A) \leq P(B)$
- $P(\overline{A}) = 1 - P(A)$
- $P(A - B) = P(A) - P(A \cap B)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

**Podmíněná pravděpodobnost** značí vztah  $P(A|B) = \frac{P(A \cap B)}{P(B)}, P(B) \neq 0$ . Jevy jsou **nezávislé**, pokud  $P(A|B) = P(A)$  nebo  $P(B) = 0$ . Pro nezávislé jevy platí  $P(A \cap B) = P(A) \cdot P(B)$ . Jevy  $A_1, \dots, A_n$  jsou **stochasticky nezávislé** právě tehdy, když  $P(\bigcap A_i) = \prod P(A_i)$ .

Pro úplnou skupinu disjunktních jevů  $B_1, \dots, B_n, B_i \cup B_j = \emptyset$  vyslovme **Total Probability Theorem**:  $A \in \mathcal{S}, P(A) = \sum P(A|B_i) \cdot P(B_i) = \sum P(A \cap B_i)$  a **Bayes Theorem**:  $P(B_k|A) = \frac{P(A|B_k) \cdot P(B_k)}{\sum P(A|B_i) \cdot P(B_i)}$ .

# Náhodná veličina

Mějme pravděpodobnostní prostor  $(\Omega, S, P)$ . **Náhodná veličina**  $X$  je reálná funkce prvků  $\omega \in \Omega$  ze základního prostoru taková, že pro každé reálné  $x \in \mathbb{R}$  je množina  $\{\omega \in \Omega | X(\omega) < x\} \in S$ , tj. náhodným jevem. Náhodná veličina je zobrazením  $X : \Omega \rightarrow \mathbb{R}$  takové, že pro každé  $x \in \mathbb{R}$  platí  $X((-\infty, x)) = \{\omega \in \Omega | X(\omega) < x\} \in S$ . Množina  $\{x = X(\omega), \omega \in \Omega\}$  se nazývá **základní soubor**.

Nechť  $X$  je náhodná veličina. Reálnou funkci  $F(t)$  definovanou pro všechna reálná  $t \in \mathbb{R}$  vztahem

$$F(t) = P\{X \in (-\infty, t)\} = P(X < t)$$

Nazveme **distribuční funkci** náhodné veličiny  $X$ . Jedná se tedy o funkci, která každému reálnému číslu přiřazuje pravděpodobnost, že náhodná veličina nabude hodnoty menší než toto reálné číslo.

1. Distribuční funkce je nezáporné číslo menší nebo rovno jedné.

$$0 \leq F(x) \leq 1$$

2. Distribuční funkce je neklesající.

$$\forall x_1, x_2 \in \mathbb{R} : x_1 < x_2 \Rightarrow F(x_1) \leq F(x_2)$$

3. Distribuční funkce je zleva spojitá.

$$\lim_{x \rightarrow \infty} F(x) = 1, \lim_{x \rightarrow -\infty} F(x) = 0$$

$$5. \forall a, b \in \mathbb{R}, a < b : P(a \leq X < b) = F(b) - F(a)$$

$$6. P(X = x_0) = \lim_{x \rightarrow x_0^+} F(x) - F(x_0)$$

Pro **diskrétní náhodnou veličinu** platí, že existuje konečná nebo spočetná množina reálných čísel  $M = \{x_1, \dots, x_n, \dots\}$  takových, že  $P(X = x_i) > 0, i = 1, \dots, n, \dots$  a  $\sum P(X = x_i) = 1$ . Funkce  $P(x_i) = P(X = x_i)$  se nazývá **pravděpodobnostní funkci** náhodné veličiny  $X$ . Distribuční funkce je schodovitá a platí pro ni  $F(x) = \sum_{x_i < x} P(X = x_i)$ .

Pro **spojitou náhodnou veličinu** platí, že distribuční funkce má tvar  $F(x) = \int_{-\infty}^x f(t)dt$ , kde  $f(x)$  je nezáporná funkce zvaná **hustota pravděpodobnosti**, pro kterou platí, že  $\int_{-\infty}^{\infty} f(x)dx = 1$ . Ve všech bodech, kde existuje derivace distribuční funkce platí  $f(x) = \frac{dF(x)}{dx}$ . Platí, že

$$1. P(X < a) = F(a) = \int_{-\infty}^a f(x)dx$$

$$2. P(X \geq a) = 1 - F(a) = \int_a^{\infty} f(x)dx$$

$$3. P(a \leq X < b) = F(b) - F(a) = \int_a^b f(x)dx$$

$$4. P(X = x) = 0$$

Dvě náhodné veličiny  $X, Y$  jsou **nezávislé**, pokud pro náhodný vektor (viz Náhodný vektor)  $A = (X, Y)$  platí  $F(x, y) = F_x(x) \cdot F_y(y)$ .

# Číselné charakteristiky náhodné veličiny

Obecný moment  $r$ -tého řádu  $\mu'_r = EX^r = \int_{-\infty}^{\infty} x^r f(x) dx$ .

Centrální moment  $r$ -tého řádu  $\mu_r = E(X - EX)^r = \int_{-\infty}^{\infty} (x - EX)^r f(x) dx$ .

Střední hodnota  $EX = \mu = \int_{-\infty}^{\infty} x f(x) dx$

- $E(aX + b) = aEX + b$
- $E(X_1 + X_2) = EX_1 + EX_2$
- $E(X_1 X_2) = E(X_1) E(X_2)$  pro nezávislé náhodné veličiny
- $Y = g(X) \Rightarrow EY = E(g(X)) = \int_{-\infty}^{\infty} g(x) f(x) dx$

Rozptyl  $DX = \mu_2 = E(X - EX)^2 = EX^2 - (EX)^2 = \int_{-\infty}^{\infty} x^2 f(x) dx - \left(\int_{-\infty}^{\infty} x f(x) dx\right)^2$ .

- $D(aX + b) = a^2 DX$
- $D(X_1 + X_2) = DX_1 + DX_2$  pro nezávislé náhodné veličiny

Směrodatná odchylka  $\sigma_x = \sqrt{DX}$

Šikmost (skewness) - mírou symetrie daného rozdělení  $a_3 = \frac{\mu_3}{\sigma_x^3}$  ( $a_3 = 0$  symetrický soubor,  $a_3 < 0$  negativně zešikmený soubor,  $a_3 > 0$  pozitivně zešikmený soubor).

Špičatost (kurtosis) - míra plochosti/špičatosti  $a_4 = \frac{\mu_4}{\sigma_x^4}$  ( $a_4 < 3$  plošší,  $a_4 > 3$  špičatější).

Kvantily jsou definovány jako v Explorační analýza dat  $F(x_p) = p$ .

Modus  $\hat{x}$  je pro diskrétní NV hodnota  $P(X = \hat{x}) \geq P(X = x_i)$ , pro spojitou NV  $f(\hat{x}) \geq f(x)$ .

## Náhodný vektor

**Náhodným vektorem** rozumíme sloupcový vektor složený z náhodných veličin  $\mathbf{X} = (X_1, X_2, \dots, X_n)$ .

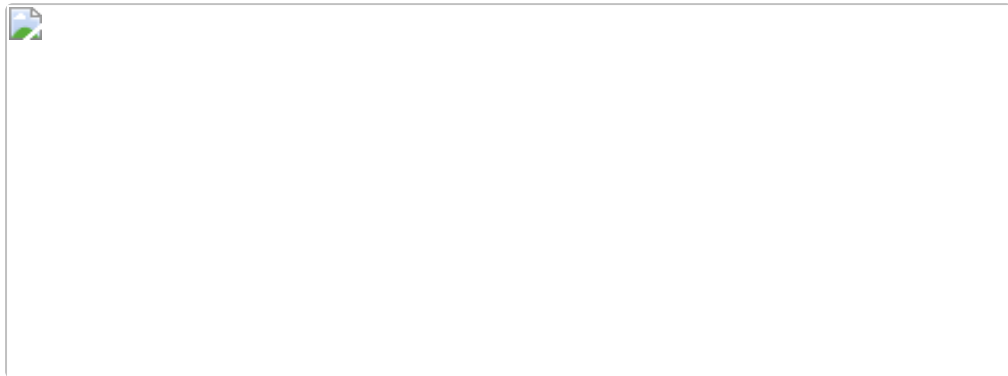
**Sdružená distribuční funkce** náhodných veličin  $F(x_1, \dots, x_n) = P(X_1 < x_1, \dots, X_n < x_n)$ .

- $\lim_{\mathbf{x} \rightarrow -\infty} F(\mathbf{x}) = 0$
- $\lim_{\mathbf{x} \rightarrow \infty} F(\mathbf{x}) = 1$
- Funkce je neklesající a zleva spojitá v každé proměnné
- $P(a_1 \leq X_1 < b_1, a_2 \leq Y < b_2) = F(b_1, b_2) - F(b_1, a_2) - F(a_1, b_2) + F(a_1, a_2)$

V případě **náhodného vektoru s diskrétním rozdělením** definujeme sdruženou distribuční funkci jako  $F(x_1, \dots, x_n) = \sum_{x_{1i} < x_1, \dots, x_{ni} < x_n} P(X_1 = x_{1i}, \dots, X_n = x_{ni})$ , kde  $P(x_{1i}, \dots, x_{ni})$  je **sdružená**



**pravděpodobnostní funkce.** V případě **náhodného vektoru se spojitým rozdělením** platí běžné definice rozložené do více dimenzí.



Chceme-li určit distribuční funkci veličiny  $X$  z dvousložkového vektoru, mluvíme o **marginální distribuční funkci**  $F_x(x) = P(X < x) = \lim_{y \rightarrow \infty} F(x, y)$ ,  $F_y(Y) = P(Y < y) = \lim_{x \rightarrow \infty} F(x, y)$ .

Chceme-li určit hustotu pravděpodobnosti veličiny  $X$  z dvousložkového vektoru, mluvíme o **marginální hustotě pravděpodobnosti**  $f_x(x) = \int_{y \rightarrow \infty} F(x, y)$ ,  $f_y(y) = \int_{x \rightarrow \infty} F(x, y)$ .

Složky  $X, Y$  dvousložkového náhodného vektoru jsou navzájem nezávislé právě tehdy, jsou-li nezávislé náhodné veličiny  $X, Y$ . Platí tedy, že  $F(x, y) = F_x(x)F_y(y)$ . Z těchto údajů můžeme vytvořit korelační tabulku.

Pro náhodný vektor je definována **podmíněná pravděpodobnostní funkce**  $f(x|y) = \frac{f(x,y)}{f_y(y)}$ .

## Charakteristiky náhodného vektoru

Smíšený obecný moment řádu  $k, n$ :  $\mu'_{kn} = E(X^k Y^n)$ .

Smíšený centrální moment řádu  $k, n$ :  $\mu_{kn} = E[(X - EX)^k (Y - EY)^n]$

Kovariance je nejjednodušším ukazatelem souvislosti dvou náhodných veličin  $Cov(X, Y) = \mu_{11} = E[(X - EX)(Y - EY)]$ . Kladná hodnota kovariance znamená, že se zvětšením hodnoty  $X$  se pravděpodobně zvýší i hodnota  $Y$ , oproti tomu záporná hodnota kovariance znamená, že se zvětšením hodnoty  $X$  se pravděpodobně sníží hodnota  $Y$ . Často definujeme kovarianční matici

$$\begin{bmatrix} DX & Cov(X, Y) \\ Cov(X, Y) & DY \end{bmatrix}.$$

Jednoduchý korelační koeficient je mírou lineární závislosti dvou náhodných veličin definovaný jako  $\rho_{x,y} = \frac{Cov(X,Y)}{\sqrt{DX \cdot DY}}$ . Mohou být nekorelované, pozitivně korelované a negativně korelované.

## Diskrétní rozdělení pravděpodobnosti

Definujeme **Bernoulliho pokusy** posloupnost nezávislých pokusů majících pouze 2 možné výsledky a pravděpodobnost výskytu události  $p$  je konstantní v každém pokuse.



**Poissonův proces** popisuje výskyt náhodných událostí na nějakém pevném časovém intervalu – speciální případ bodového procesu. Každý proces musí dodržet následující předpoklady – rychlost výskytu událostí je konstantní v průběhu celého intervalu a jednotlivé události musí být nezávislé.

## Hypergeometrická náhodná veličina

---

Předpokládejme, že v souboru  $N$  prvků  $M$  prvků s danou vlastností a zbylých  $(N - M)$  prvků tuto vlastnost nemá. Postupně vybereme ze souboru  $n$  prvků, z nichž žádný nevracíme zpět. Definujme náhodnou veličinu  $X$  jako počet se sledovanou vlastností ve výběru  $n$  prvků, pak tato veličina má hypergeometrické rozdělení s parametry  $N, M, n$ , což značíme  $X \rightarrow H(N; M; n)$ .

$$P(X = k) = \frac{\binom{M}{k} \binom{N-M}{n-k}}{\binom{N}{n}}$$

Hypergeometrické rozdělení využijeme při statistické kontrole jakosti, když zkoumáme jakost malého počtu výrobků nebo když kontrola má ráz destrukční zkoušky.

## Binomická náhodná veličina

---

Binomická náhodná veličina  $X$  je definována jako počet výskytu události v  $n$  Bernoulliho pokusech. Pro rozložení veličiny  $X \rightarrow Bi(n, p)$  musíme znát počet pokusů  $n$  a pravděpodobnost výskytu události  $p$ .

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

Je-li výběrový poměr  $\frac{n}{N}$  v hypergeometrickém rozdělení menší než 0,05, lze hypergeometrické rozdělení nahradit binomickým  $H(N; M; n) \rightarrow Bi(n; \frac{M}{N})$ .

Variantou binomické veličiny pro  $n = 1$  je **alternativní náhodná veličina**. Pokud  $X \rightarrow A(p)$ , poté  $P(X = 1) = p, P(X = 0) = 1 - p$ .

## Geometrická náhodná veličina

---

Geometrická náhodná veličina  $X$  je definovaná jako počet Bernoulliho pokusů do prvního výskytu události, **včetně něj**. Značíme  $X \rightarrow G(p)$ , kde  $p$  je pravděpodobnost výskytu události.

$$P(X = n) = p(1 - p)^{n-1}$$

# Negativně binomická náhodná veličina

Negativně binomická náhodná veličina  $X$  je definována jako počet Bernoulliho pokusů; do  $k$ -tého výskytu události, včetně  $k$ -tého výskytu. Geometrická náhodná veličina je speciálním případem negativně binomické náhodné veličiny pro  $k = 1$ . Značíme  $X \rightarrow NB(k, p)$ , kde  $k$  je požadovaný počet výskytů události a  $p$  je pravděpodobnost výskytu události.

$$P(X = n) = \binom{n-1}{k-1} p^k (1-p)^{n-k}$$

## Poissonovo rozdělení pravděpodobnosti

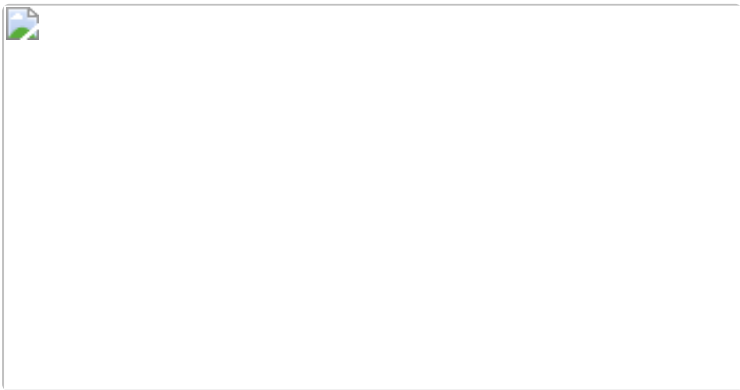
Definujme si náhodný pokus jako Poissonův proces probíhající v čase  $t$  s rychlostí výskytu  $\lambda$ . Pokud veličina  $X$  značí počet výskytu události v časovém intervalu  $t$  poté  $X \rightarrow Po(\lambda t)$ .

$$P(X = k) = \frac{(\lambda t)^k e^{-\lambda t}}{k!}$$

Je-li počet pokusů  $n \rightarrow \infty$  a pravděpodobnost výskytu události  $p \rightarrow 0$ , poté můžeme binomické rozdělení aproximovat Poissonovým rozdělením  $Bi(n, p) \sim Po(\lambda)$ ,  $\lambda = np$ . Dobrou aproximaci splňují podmínky  $n > 30$  a  $p < 0.3$ .

## Spojité rozdělení pravděpodobnosti

Pro nezápornou náhodnou veličinu  $X$  se spojitým rozdělením definujeme pro  $F(t) \neq 1$  **intenzitu poruch**  $\lambda(t) = \frac{f(t)}{1-F(t)}$ . Představuje-li náhodná veličina  $X$  dobu do poruchy nějakého zařízení, pak intenzita poruch vyjadřuje, že pokud do času  $t$  nedošlo k žádné poruše, tak pravděpodobnost, že k ní dojde v následujícím okamžiku malé délky  $dt$ , je přibližně  $\lambda(t) \cdot dt$ .



Křivka na obrázku se nazývá **vanová křivka** a obvykle se dělí na tři úseky.

1. V prvním úseku křivka poruch klesá. Odpovídající časový interval se nazývá **období časných poruch**. Příčinou zvětšené intenzity poruch v tomto období jsou poruchy v důsledku výrobních vad, nesprávné

montáže, chyb při návrhu nebo při výrobě.

2. Ve druhém úseku dochází k běžnému využívání zaběhnutého výrobku, k poruchám dochází většinou z vnějších příčin, nedochází k opotřebení, které by změnilo funkční vlastnosti výrobku. Např. exponenciální rozdělení.
3. Ve třetím úseku procesy stárnutí a opotřebení mění funkční vlastnosti výrobku, projevují se nastřádané otřesy, trhliny a intenzita poruch vzrůstá. Např. Erlangovo rozdělení.

## Rovnoměrné rozložení

Rozložení s hustotou pravděpodobností je konstantní na intervalu  $\langle a, b \rangle$ . Náhodnou veličinu s tímto rozdělením značíme  $X \rightarrow R(a, b)$

$$f(x) = \begin{cases} \frac{1}{b-a} & x \in \langle a, b \rangle \\ 0 & \text{jinde} \end{cases}$$

## Exponenciální rozdělení

Mějme Poissonův proces, tj. v určitém časovém intervalu se s konstantní rychlostí výskytu  $\lambda$  objevují události, které jsou na sobě nezávislé. Poté exponenciální rozdělení značí dobu do výskytu první události. Náhodnou veličinu  $X$  s exponenciálním rozdělením značíme  $X \rightarrow E(\lambda)$ , kde  $\lambda$  je parametrem Poissonova procesu.

$$f(t) = \lambda e^{-\lambda t}$$

Exponenciální rozdělení bývá někdy nazýváno **rozdělení bez paměti**  $P(X > (t_1 + t_2) | X > t_1) = P(X > t_2)$ . Toto rozdělení dobře popisuje dobu života zařízení, u kterých dochází k poruše ze zcela náhodných příčin.

Exponenciální rozdělení je využito v teorii hromadné obsluhy nebo v teorii spolehlivosti.

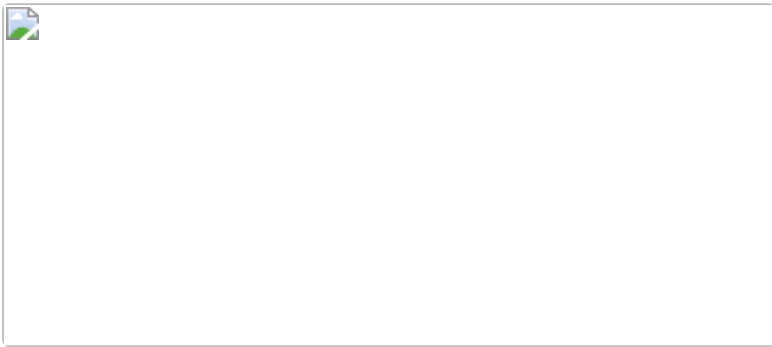
## Erlangovo rozdělení

Určitým zobecněním exponenciální náhodné veličiny je veličina s Erlangovým rozdělením, která popisuje dobu do výskytu  $k$ -té události v Poissonově procesu. Erlangovo rozdělení je speciálním typem tzv. Gamma rozdělení pro  $k$  z množiny celých čísel. Značíme  $X_k \rightarrow Erlang(k, \lambda)$ , kde  $k$  je počet událostí (parametr tvaru) a  $\lambda$  je rychlost výskytu těchto událostí.

$$f(t) = \lambda e^{-\lambda t} \cdot \frac{(\lambda t)^{k-1}}{(k-1)!}$$

Intenzita poruch je v případě Erlangova rozdělení rostoucí funkce a proto je toto rozdělení vhodné pro modelování procesů stárnutí.

# Weibullovo rozdělení



Weibullovo rozdělení je velmi flexibilní a proto se jím popisují veličiny jako doba do poruchy. Používá se při popisu komponent v období raných poruch nebo v období stárnutí. Weibullovo rozdělení má dva parametry  $\Theta$  – parametr měřítka, scale, závisí na materiálu, namáhání a podmínkách užívání – a  $\beta$  – parametr tvaru, shape, na jeho hodnotě závisí tvar intenzity poruch a tím i vhodnost použití pro určité období doby života. Veličinu značíme  $X \rightarrow W(\Theta, \beta)$ .

$$F(t) = 1 - e^{-\left(\frac{t}{\Theta}\right)^{\beta}}$$

Pro intenzitu poruch platí  $\lambda(t) = konst. \cdot t^{\beta-1}$ , tudíž tvar intenzity poruch závisí na volbě parametru  $\beta$ .

$0 < \beta < 1$	období dětských nemocí	klesající funkce
$\beta = 1$	období stabilního života	exponenciální rozdělení
$1 < \beta < 2$	období stárnutí	konvexní, rostoucí funkce
$\beta = 2$	období stárnutí	lineárně rostoucí funkce
$\beta > 2$	období stárnutí	konkávní, rostoucí funkce

# Normální rozdělení

Lze říci, že normální rozdělení je vhodným pravděpodobnostním modelem tehdy, působí-li na kolísání náhodné veličiny velký počet nepatrných a vzájemně nezávislých vlivů. Za určitých podmínek lze pomocí něj aproximovat řadu jiných spojitých i nespojitých rozdělení. Normální rozdělení má dva parametry:  $\mu$  – střední hodnotu charakterizující polohu a  $\sigma^2$  – rozptyl. Náhodnou veličinu s normálním rozdělením značíme  $X \rightarrow N(\mu; \sigma^2)$ .

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)^2}$$

## Normované normální rozdělení

Normální rozdělení se středním hodnotou rovnou nule a jednotkovým rozptylem. To, že má náhodná veličina  $Z \rightarrow N(0, 1)$ .

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{x^2}{2}}$$

Nechť  $X \rightarrow N(\mu, \sigma^2)$ , poté definujme  $Z = \frac{X-\mu}{\sigma}$  se stejným, ale normovaným rozdělením. Mezi distribuční funkci normální a normované normální náhodné veličiny platí vztah  $F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$ .

Pravidlo  $6\sigma$  je jedním ze základních principů, na nichž stojí kontrola kvality a jakosti. Máme-li data pocházející z normálního rozdělení o parametrech  $\mu, \sigma^2$ , pak téměř všechna (99,8 %) leží v intervalu  $\mu \pm 3\sigma$ .

## $\chi^2$ rozdělení

Nechť  $Z_1, \dots, Z_n$  jsou nezávislé náhodné veličiny,  $Z_i \rightarrow N(0, 1)$ . Poté náhodná veličina  $\chi_n^2 = \sum Z_i^2 \rightarrow \chi^2(n)$  má chí-kvadrát rozdělení o  $n$  stupních volnosti.  $E(\chi_n^2) = n, D(\chi_n^2) = 2n$ .

## Studentovo rozdělení

Náhodná veličina  $t_n = \frac{Z}{\sqrt{\frac{\chi_n^2}{n}}}$  má Studentovo rozdělení o  $n$  stupních volnosti.  $E(\chi_n^2) = 0, D(\chi_n^2) = \frac{n}{n-2}$ .

Tvarem Studentova rozdělení je také symetrická zvonovitá křivka, stejně jako o normálního rozdělení. Pro velká  $n$  je rozdělení blízké k  $N(0, 1)$ .

## Fisherovo-Snedecorovo rozdělení

Náhodná veličina  $F_{n,m} = \frac{\frac{\chi_m^2}{m}}{\frac{\chi_n^2}{n}}$  má Fisherovo-Snedecorovo rozdělení o  $m$  a  $n$  stupních volnosti.  $E(F_{m,n}) = \frac{m}{m-2}$ . Pro velká  $m$  se střední hodnota blíží k 1.

## Limitní věty

Definujme **konvergenci podle pravděpodobnosti ke konstantě**: je dána posloupnost náhodných veličin  $\{X_n\} = (X_1, \dots, X_n)$  a reálné číslo  $a$ , poté pokud pro  $\varepsilon > 0$  platí  $\lim_{n \rightarrow \infty} P(|X_n - a| < \varepsilon) = 1$ , pak říkáme posloupnost  $\{X_n\}$  konverguje k  $a$  podle pravděpodobnosti. Značíme  $X_n \xrightarrow{p} \mu$ .

Definujeme **konvergenci v distribuci**: je dána posloupnost náhodných veličin  $\{X_n\}$  a náhodná veličina  $X$  s distribuční funkcí  $F(x)$ . Jestliže  $\lim_{n \rightarrow \infty} F_n(x) = F(x)$ , pak říkáme, že posloupnost náhodných veličin  $\{X_n\}$  konverguje k náhodné veličině  $X$  v distribuci a  $F(x)$  nazýváme **asymptotickou distribuční funkcí**. Poté můžeme náhodnou veličinu  $X_n$  aproximovat asymptotickou distribuční funkcí.

Je-li  $X$  libovolná náhodná veličina se střední hodnotou  $EX$  a konečným rozptylem  $DX = \sigma^2$ , pak **Čebyševova nerovnost** odhaduje pravděpodobnost odchylky náhodné veličiny  $X$  od její střední hodnoty.

$$\forall \epsilon > 0 : P(|X - EX| \geq \epsilon) \leq \frac{DX}{\epsilon^2}$$

Čebyševova nerovnost pro případ, kdy chceme odhadnout pravděpodobnost, že náhodná veličina  $X$  je od své střední hodnoty vzdálená o více než  $k$ -násobek směrodatná odchylky  $\sigma$

$$\forall \sigma, k > 0 : P(|X - EX| \geq k\sigma) \leq \frac{1}{k^2}$$

**Zákon velkých čísel** označuje tvrzení o konvergenci průměru v posloupnosti náhodných veličin:  $X_1, \dots, X_n$  jsou nezávislé náhodné veličiny, jejichž střední hodnoty jsou rovny  $\mu$ . Jestliže  $\overline{X_n}$  definujeme jako  $\overline{X_n} = \frac{1}{n} \sum_{j=1}^n X_j$ , pak posloupnost  $(\overline{X_n} \xrightarrow{p} \mu)$ . Posloupnost nemusí mít stejné rozdělení, a zároveň nemáme žádné požadavky na jejich rozptyl.

Důsledkem zákona velkých čísel je **Bernoulliho věta**, která tvrdí, že relativní četnost sledovaného jevu stochasticky konverguje (konverguje podle pravděpodobnosti) k jeho pravděpodobnosti. Necht'  $X_1, X_2, \dots$  jsou nezávislé náhodné veličiny s alternativním rozdělením s parametrem  $p$ , jestliže  $\overline{X_n}$  definujeme jako  $\overline{X_n} = \frac{1}{n} \sum X_j$ , pak  $(\overline{X_n} \xrightarrow{p} p)$ .

## Centrální limitní věta

O náhodných veličinách, jež konvergují v distribuci k normálnímu rozdělení, říkáme, že mají **asymptoticky normální rozdělení**.

**Lindenberg-Lévy** – jestliže  $X_1, \dots, X_n$  jsou nezávislé náhodné veličiny se stejnými středními hodnotami  $\mu$  a se stejnými rozptyly  $\sigma^2$ , pak platí

$$Y_n = \frac{\sum X_i - n\mu}{\sigma\sqrt{n}} \Rightarrow \lim_{n \rightarrow \infty} P(Y_n < u) = \Phi(u)$$

Čili  $Y_n$  má asymptoticky normální rozdělení  $N(0, 1)$ . Proto platí, že

1.  $X = \sum X_i \Rightarrow EX = n\mu, DX = n\sigma^2$ ,  
rozdělení náhodné veličiny  $X$  lze aproximovat rozdělením  $N(n\mu; n\sigma^2)$
2.  $\overline{X} = \frac{\sum X_i}{n} = \frac{X}{n} \Rightarrow E\overline{X} = \mu, D\overline{X} = \frac{\sigma^2}{n}$   
rozdělení náhodné veličiny  $\overline{X}$  lze aproximovat rozdělením  $N(\mu, \frac{\sigma^2}{n})$

**Moivre-Laplace** – necht'  $X \rightarrow Bi(n; p)$ ,  $EX = np$ ;  $DX = np(1 - p)$ , potom pro velká  $n$  platí, že:

$$U = \frac{X - np}{\sqrt{np(1-p)}} \rightarrow N(0, 1)$$

Aproximace binomického rozdělení normálním se zlepšuje s rostoucím rozptylem. Poměrně dobré výsledky dává tato aproximace v případě, že  $np(1-p) > 9$  nebo  $\min\{np; n(1-p)\} > 5$ .

**Aproximace rozdělení výběrové relativní četnosti normálním rozdělením** – máme-li  $n$  Bernoulliho pokusů, při kterých nastane  $k$  výskytů nějaké události, můžeme určit výběrovou relativní četnost  $p = \frac{k}{n} = \frac{\sum X_i}{n}$ , kde  $X_i \rightarrow A(\pi)$ . Na základě Lindenberg-Lévy můžeme tento součet aproximovat normálním rozdělením  $\sum X_i = N(n\pi, n\pi(1-\pi))$ , jejich průměr  $p \rightarrow N\left(\pi, \frac{\pi(1-\pi)}{n}\right)$  a  $\frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \rightarrow N(0, 1)$ . Pro přesnější výpočty se provádí **oprava na spojitost**.

$$P\left(k_1 < \sum X_i < k_2\right) = F(k_2) - F(k_1) \approx \Phi\left(\frac{k_2 + 0.5 - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k_1 - 0.5 - np}{\sqrt{np(1-p)}}\right)$$

**Aproximace Poissonova rozdělení normálním rozdělením** – pokud interval  $(0, t)$  je dostatečně velký, lze Poissonovo rozdělení aproximovat  $X \rightarrow Po(\lambda t) \rightarrow N(\lambda t, \lambda t)$  pro dostatečně velké  $t$ . Dále lze průměrný počet výskytů událostí za časovou jednotku aproximovat normálním rozdělením  $Y = \frac{X}{t} \rightarrow N\left(\lambda, \frac{\lambda}{t}\right)$ .

$$P(k_1 < X < k_2) = F(k_2) - F(k_1) \approx \Phi\left(\frac{k_2 + 0.5 - \lambda t}{\sqrt{\lambda t}}\right) - \Phi\left(\frac{k_1 - 0.5 - \lambda t}{\sqrt{\lambda t}}\right)$$

## Náhodné výběry a jejich zpracování

**Náhodný výběr** je speciální náhodný vektor, jehož složky jsou nezávislé náhodné veličiny se stejným rozdělením pravděpodobnosti. Opakujeme-li  $n$ -krát nezávisle pokus, jehož výsledkem je náhodná veličina  $X$  s distribuční funkcí  $F(x)$ , sledujeme náhodný výběr  $X = (X_1, \dots, X_n)$ ,  $X_i \sim F(x)$  z rozdělení  $F(x)$ , kde  $n$  značí rozsah výběru. Obvykle rozdělujeme náhodné výběry na **malé** pro  $n \leq 30$  a **velké** pro  $n > 30$ . Náhodný výběr má simultánní distribuční funkci  $F(x) = \prod F(x_i)$  a simultánní hustotu pravděpodobností  $f(x) = \prod f(x_i)$ .

$$1. T_1(X) = \overline{X_n} = \frac{\sum_i X_i}{n}, E\overline{X_n} = EX_i$$

výběrový průměr pro odhad střední hodnoty

$$2. T_2(X) = \frac{1}{n-1} \sum_i (X_i - \overline{X_n})^2, E(T_2(X)) = DX_i$$

výběrový rozptyl,  $T_3(X) = \sqrt{T_2(X)} = S$  výběrová směrodatná odchylka

Předpokládejme, že  $X = (X_1, \dots, X_n)$ ,  $X_i \rightarrow N(\mu, \sigma^2)$ . Definujme výběrová rozdělení

$$1. \overline{X_n} \rightarrow N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$2. Z_n = \frac{\overline{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}} \rightarrow N(0, 1)$$



$$3. \frac{S_x^2}{\sigma_x^2}(n-1) \rightarrow \chi^2(n-1)$$

jeden stupeň volnosti ztrácíme na náhradu  $\mu$  za  $\overline{X}_n$

$$4. \frac{(\overline{X}_n - \mu)}{S} \sqrt{n} \rightarrow t_{n-1}$$

Předpokládejme navíc, že  $Y = (Y_1, \dots, Y_m), Y_i \rightarrow N(\mu', \sigma'^2)$ .

$$5. \frac{\overline{X} - \overline{Y} - (\mu - \mu')}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma'^2}{m}}} \rightarrow N(0, 1)$$

$$6. \frac{\frac{S_x^2}{\sigma_x^2}(n-1)}{\frac{S_y^2}{\sigma_y^2}(m-1)} = \frac{\frac{S_x^2}{\sigma_x^2}}{\frac{S_y^2}{\sigma_y^2}} \rightarrow F_{n-1, m-1}$$

Předpokládejme navíc, že  $\sigma_x^2 = \sigma_y^2$ .

$$7. \frac{\overline{X} - \overline{Y} - (\mu - \mu')}{\sqrt{S_x^2(n-1) + S_y^2(m-1)}} \cdot \sqrt{\frac{nm}{n+m}} \cdot \sqrt{n+m-2} \rightarrow t_{n+m-2}$$

## Teorie odhadu

Nechť máme náhodný výběr  $X = (X_1, \dots, X_n) \sim F(x, \Theta)$ . Cílem naší teorie odhadu je pro známé pravděpodobnostní rozdělení  $F$  najít parametr  $\Theta \in \Omega$ , kde  $\Omega$  značí parametrický prostor pomocí výběrové charakteristiky **odhadu**  $\widehat{\Theta} = T(X)$  pro nalezení **bodového odhadu**  $t(x)$ . Aby odhad byl přesný, požadujeme splnění tří vlastností odhadu – nestrannost, konzistence a efektivita.

**Nestrannost odhadu** – požadujeme, aby  $\forall \Theta \in \Omega : E\widehat{\Theta} = \Theta$ . Často požadujeme, aby byl odhad alespoň **asymptoticky nestranný**  $\forall \Theta \in \Omega : \lim_{n \rightarrow \infty} E\widehat{\Theta}_n = \Theta$ .

- Výběrový průměr  $\widehat{\Theta} = \overline{X}$  je nestranným odhadem střední hodnoty.

**Konzistence odhadu** – požadujeme, aby byl nestranný nebo asymptoticky nestranný a zároveň  $\lim_{n \rightarrow \infty} D\widehat{\Theta}_n = 0$ .

- Výběrový průměr  $\widehat{\Theta} = \overline{X}$  je konzistentním odhadem střední hodnoty, neboť  $D\widehat{\Theta} = \frac{\sigma^2}{n} \rightarrow 0$ .

**Efektivnost odhadu** – odhad je efektivní, značíme  $\widehat{\Theta}_0$ , právě tehdy když je nestranný a zároveň  $\forall \widehat{\Theta}_1, E\widehat{\Theta}_1 = \Theta : D\widehat{\Theta}_0 \leq D\widehat{\Theta}_1$ .

## Intervalový odhad

Často hledáme **intervalový odhad** – hledáme funkce  $T_D(X)$  a  $T_H(X)$  tak, aby  $P(T_D \leq \Theta \leq T_H) = 1 - \alpha$ , kde  $\alpha$  je nejčastěji 0.01. Těmito mezím říkáme **interval spolehlivosti pro  $\Theta$**  se spolehlivostí  $1 - \alpha$ . Konkrétní reprezentaci  $t_D(X)$  a  $t_H(X)$  nazýváme **intervalový odhad pro  $\Theta$**  se spolehlivostí  $1 - \alpha$ .

Rozdělme  $\alpha = \alpha_1 + \alpha_2; \alpha_1, \alpha_2 \geq 0$  tak, aby  $P(\Theta \leq T_H(X)) = 1 - \alpha_2$  a  $P(\Theta < T_D(X)) = \alpha_1$ . Nejčastěji volíme  $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$  nebo  $\alpha_1 = \alpha, \alpha_2 = 0$  pro jednostranný interval spolehlivosti.

1. Zvolme vhodnou  $T(X)$ , ze které jsme schopni odvodit  $T_D$  a  $T_H$ .

2. Algebraickou metodou najděme  $T_D$  a  $T_H$ .

## Testování hypotéz

Testování hypotéz je pojat jako rozhodovací proces, v němž proti sobě stojí dvě tvrzení. **Nulová hypotéza**  $H_0$  představuje rovnovážný stav, bývá vyjádřena rovností. Jde o tvrzení o populaci, které je bráno jako předpoklad při testování. Oproti ní stavíme tzv. **alternativní hypotézu**  $H_A$ . Ta představuje porušení rovnovážného stavu a zapisujeme je nerovností nebo nerovnicí. Alternativní hypotézu volíme v souladu s daty.

**Čistý test významnosti** zodpovídá otázku, zda získaný náhodný výběr  $X$  je či není extrémní s ohledem na testovanou hypotézu (zda zjištěné údaje podporují nulovou hypotézu).

1. Formulace nulové hypotézy  $H_0$  a alternativní hypotézy  $H_A$ .

2. Volba testové statistiky  $T(X)$  – funkce výběru, která vyjadřuje sílu platnosti nulové hypotézy ve srovnání s hypotézou alternativní. Je třeba znát nulové rozdělení  $F_0(x) = P(T(X) < x | H_0)$ .

3. Výpočet pozorované hodnoty testové statistiky  $x_{OBS}$ .

1. Je-li  $H_A$  ve tvaru „<“:  $p_{value} = F_0(x_{OBS})$ .

2. Je-li  $H_A$  ve tvaru „>“:  $p_{value} = 1 - F_0(x_{OBS})$ .

3. Je-li  $H_A$  ve tvaru „≠“ a nulové rozdělení je symetrické:  $p_{value} = 2 \min \{F_0(x_{OBS}); 1 - F_0(x_{OBS})\}$ .

4.  $p_{value}$  určuje minimální hladinu významnosti, na níž bychom při daném výběrovém souboru mohli nulovou hypotézu zamítnout. Čím menší je  $p_{value}$ , tím silnější je výpověď náhodného výběru proti nulové hypotéze. Nejběžněji:

1. Je-li  $p_{value} < 0.01 = \alpha$ : zamítáme  $H_0$ .

2. Je-li  $0.01 < p_{value} < 0.05$ : nedokážeme rozhodnout.

3. Je-li  $p_{value} > 0.05$ : nezamítáme  $H_0$ .

Jelikož při rozhodování o nulové hypotéze vycházíme z výběrového souboru, který nemusí dostatečně přesně odpovídat vlastnostem základního souboru, můžeme se při rozhodování dopustit chyby.

### VÝSLEDEK TESTU

SKUTEČNOST	Nezamítáme $H_0$		Zamítáme $H_0$	
	Platí $H_0$	OK, pravděpodobnost $1 - \alpha$ zvaná <i>spolehlivost</i>	<b>Chyba 1. druhu,</b> pravděpodobnost $\alpha$ zvaná <i>hladina významnosti</i>	OK, pravděpodobnost $\gamma := 1 - \beta$ zvaná <i>síla testu</i>
	Platí $H_A$	<b>Chyba 2. druhu,</b> pravděpodobnost $\beta$		



Pravděpodobnost chyby 2. druhu závisí na přesné hodnotě alternativní hypotézy.

Dokážeme určit  $\beta$  pro případ, že alternativní hypotéza je přesně specifikována. **Operativní charakteristika** je závislost pravděpodobnosti chyby 2. druhu na přesné specifikaci alternativní hypotézy.

Při testování více než dvou hypotéz nelze použít testování „po dvojicích“, neboť  $P(\min p_1, \dots, p_k < p) \approx np \sim 1$ .

## ANOVA Analýza rozptylu

Předpokládejme  $k$  datových tříd z normálního rozdělení a mající stejný rozptyl – homeskedasticitu, každá z nich  $n_i$  hodnot ( $N = \sum_i n_i$ ). Testujeme hypotézu  $H_0 : \mu_1 = \dots = \mu_k$  proti alternativně  $H_A : \neg H_0$ . Hledáme takovou testovou statistiku  $F$ , která nejen umožní implementaci  $H_0$ , ale je i citlivá na platnost  $H_0$ .

Sestavme **totální variabilitu**  $SS_T = \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2$ , kde  $\bar{X}$  je výběrový průměr ze všech pozorovaných hodnot. Rozdělme  $SS_T = SS_W + SS_B = \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2 + \sum_{i=1}^k n_i (\bar{X}_i - \bar{X})^2$ , kde  $SS_W$  je vnitřní variabilita a  $SS_B$  je mezitřídní variabilita. Zavedme **vnitřní výběrový rozptyl** jako  $S_W^2 = \frac{SS_W}{N-k}$ , **mezitřídní výběrový rozptyl**  $S_B^2 = \frac{SS_B}{k-1}$  a **F-poměr** jako  $F = \frac{S_B^2}{S_W^2} \rightarrow F_{k-1, N-k}$ .

**Post Hoc** je proces, který provádíme v případě, že zamítneme  $H_0$ . Cílem je vytvořit takové rozdělení datových tříd, v rámci kterých platí  $H_0$ . Můžeme využít statistiku  $LSD_{i,j} = \frac{\bar{X}_i - \bar{X}_j - (\mu_i - \mu_j)}{\sqrt{\frac{\sigma^2}{n_i} + \frac{\sigma^2}{n_j}}} \cdot \frac{1}{\sqrt{\frac{S_W^2}{\sigma^2} \frac{(N-k)}{N-k}}} = \frac{\bar{X}_i - \bar{X}_j}{S_W \cdot \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}}$

pro porovnání.

V případě, že nejsou splněny požadavky ANOVA analýzy, můžeme využít **Kruskal-Wallisův test**, kde rozhodujeme o  $H_0 : x_{0.51} = \dots = x_{0.5k}$ .

## Regresní analýza

**Regrese** značí systematické změny jedné veličiny při změnách jiných veličin a popis těchto změn matematickými funkcemi. Snažíme se tedy napozorované hodnoty vyrovnat vhodnou matematickou funkcí.

**Vysvětlovaná (závisle) proměnná** – proměnná v regresním modelu, jejíž chování se snažíme vysvětlit, popsat matematickou křivkou. Jedná se o proměnnou na levé straně regresní funkce a většinou ji označujeme symbolem  $A$ . **Vysvětlující (nezávisle) proměnné** – proměnné v regresním modelu, jejichž chování vysvětluje chování závisle proměnné  $Y$ . Jedná se o proměnné na pravé straně regresní funkce a většinou je označujeme symboly  $X, Z, \dots$ .

# Obecný lineární model

---

$$Y = \mathbf{X}\beta + e$$

- $Y$  je náhodný vektor  $n$  hodnot vysvětlované proměnné
- $\mathbf{X}$  je matice zadaných hodnot vysvětlujících proměnných o rozměrech  $n \times k$
- $\beta$  je vektor  $p = k$  neznámých parametrů
- $e$  je vektor  $n$  hodnot náhodných chyb

Předpoklady obecného lineárního modelu

1.  $\forall i \leq n : Ee_i = 0$   
náhodná složka nepůsobí systematickým způsobem na hodnoty vysvětlované proměnné  $Y$
2.  $\forall i \leq n : De_i = \sigma^2$   
homoskedasticita náhodných složek, variabilita náhodné složky nezávisí na hodnotách vysvětlujících proměnných
3.  $\forall i, j \leq n; i \neq j : Cov(e_i, e_j) = 0$   
náhodné složky jsou nekorelované
4.  $\mathbf{X}$  je nestochastická matice.
5. Parametry  $\beta_j, j \leq k$  nabývají libovolných hodnot.

Pokud platí předpoklady 6 a 7, nazýváme model **regresní model**.

6.  $h(\mathbf{X}) = k \wedge n > k$   
mezi vysvětlujícími proměnnými nebyla funkční lineární závislost
7.  $\forall i \leq n : e_i \sim N$   
toto také implikuje normalitu proměnné  $Y$

Mezi několik regresních modelů patří:

- Obecná regresní přímka, nebo lineární regrese s jednou vysvětlující proměnnou

$$Y_i = \beta_0 + \beta_1 x_i + e_i, \mathbf{X} = \begin{pmatrix} 1 & x_i \\ \cdots & \cdots \\ 1 & x_n \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

- Kvadratická regrese

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_i$$

# Lineární regrese s jednou vysvětlující proměnnou

---

Mějme  $n > 2$  pozorování, tedy  $n$  dvojic  $(Y_i, x_i)$ , ze kterých sestavíme model  $Y_i = \beta_0 + \beta_1 x_i + e_i$ . Pro jeho určení využijeme metody nejmenších čtverců, tj.  $\min \sum_{i=1}^n e_i^2$ . Toto vede na soustavu normálních rovnic vedoucí k řešení  $b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$  a  $b_0 = \bar{Y} - b_1 \bar{x}$ . Odhadem hodnoty  $E(Y|x)$  je poté statistika  $\widehat{Y}(x) = b_0 + b_1 x$ . Jako vektor **reziduí** považujeme  $\widehat{e}_i = Y_i - \widehat{Y}_i$ .

Pro hledání intervalového odhadu pro  $E(Y|x)$  budeme vycházet ze statistiky  $\frac{\widehat{Y}(x) - \beta_0 - \beta_1 x}{S_{\widehat{Y}}} \sim t_{n-2}$ .