**Axiomy** $(\Omega, \mathcal{F} \subseteq \mathcal{P}(\Omega), P: \mathcal{F} \to [0,1])$

1) $\emptyset, \Omega \in \mathcal{F}$
2) $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$
3) $A_1, A_2, \ldots \in \mathcal{F} \Rightarrow \bigcup A_i \in \mathcal{F}$
4) $P(\Omega) = 1$
5) $P(\dot{\bigcup} A_i) = \sum_i P(A_i)$

**Podmíněná psť**

$P(A \cap B) = P(B) \cdot P(A|B)$

$P(A_1 \cap A_2 \cap A_3) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3 | A_1 \cap A_2)$

$P(A) = \sum_i P(B_i) \cdot P(A|B_i)$ , $B =$ rozklad $\Omega$

$P(B_j|A) = \frac{P(B_j \cap A)}{P(A)} = \frac{P(B_j) \cdot P(A|B_j)}{P(A)}$

$A \perp B \equiv P(A \cap B) = P(A) \cdot P(B)$

$A \perp_C B \equiv P(A \cap B | C) = P(A|C) \cdot P(B|C)$

**Diskrétní** ROZDĚLENÍ **Spojitá**

Diskrétní:

- $X \sim Ber(p)$
  $p_X(1) = p$ , $p_X(0) = 1-p$
  $\mathbb{E} = p$ $\quad Var = p(1-p)$
- $X \sim Geom(p)$
  $p_X(k) = (1-p)^{k-1} p$
  $\mathbb{E} = \frac{1}{p}$ $\quad Var = \frac{1-p}{p^2}$
- $X \sim Bin(n, p)$
  $p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$
  $\mathbb{E} = n \cdot p$ $\quad Var = np(1-p)$
- $X \sim Pois(\lambda) \approx Bin(n, \frac{\lambda}{n})$
  $p_X(k) = \frac{\lambda^k}{k!} e^{-\lambda}$ , $\lambda > 0$
  $\mathbb{E} = Var = \lambda$

Spojitá:

- $X \sim U(a, b)$
  $f_X(x) = \frac{1}{b-a}$ , $x \in [a, b]$
  $\mathbb{E} = \frac{a+b}{2}$ $\quad Var = \frac{(b-a)^2}{12}$
- $X \sim Exp(\lambda)$ ... $\lambda > 0$
  $F_X(x) = 1 - e^{-\lambda x}$ , $x > 0$
  $\mathbb{E} = \frac{1}{\lambda}$ $\quad Var = \frac{1}{\lambda^2}$
- $X \sim N(0, 1)$
  $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \rightsquigarrow \Phi$
- $X \sim N(\mu, \sigma^2) \to X = \mu + \sigma Z$
  $f_X(x) = \frac{1}{\sigma} \varphi\left(\frac{x-\mu}{\sigma}\right)$
  $F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$

$X_i \sim N(\mu_i, \sigma_i^2) \Rightarrow \sum X_i \sim N(\sum \mu_i, \sum \sigma_i^2)$

**Diskrétní n.v.** $X: \Omega \to \mathbb{R}$

1) $Im(X)$ je spočetná
2) $\forall x \in \mathbb{R}: \{X = x\} \in \mathcal{F}$

$\Rightarrow p_X(x): x \mapsto P[X = x]$

- **vlastnosti $\mathbb{E}$** – linearita
  $\mathbb{E}X = \sum_{n=0}^{\infty} P[X > n]$ , $Im(X) \subseteq \mathbb{N}$
  $X \perp Y \Rightarrow \mathbb{E}XY = \mathbb{E}X \cdot \mathbb{E}Y$
- **vlastnosti** $Var X := \mathbb{E}[(X - \mathbb{E}X)^2]$
  $var X = \mathbb{E}[X^2] - (\mathbb{E}X)^2$
  $= \mathbb{E}[X(X-1)] + \mathbb{E}X - (\mathbb{E}X)^2$

$Var(aX + b) = a^2 Var X$

$X \perp Y \Rightarrow Var(X+Y) = Var X + Var Y$

$CV_X := \frac{\sigma_X}{\mathbb{E}X}$ ... var. coef

$Var(X+Y) = Var X + Var Y + 2 cov(X, Y)$

$\to cov(X, Y) := \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)]$
$= \mathbb{E}XY - \mathbb{E}X \cdot \mathbb{E}Y$

$\to cov(X, aY + bZ + c) = a\, cov(X, Y) + b\, cov(X, Z)$

$\to \rho(X, Y) := \frac{cov(X, Y)}{\sigma_X \sigma_Y}$ ... korelace

**Obecné n.v.** $X: \Omega \to \mathbb{R}$

$\forall x \in \mathbb{R}: \{X \leq x\} \in \mathcal{F}$

$F_X(x): x \mapsto P[X \leq x]$

$\to F_X$ je: neklesající, zprava spoj.

$\to F_X(\infty) = 1$ , $F_X(-\infty) = 0$

**Diskrétní** VELIČINY **Spojité**

Diskrétní:

$\mathbb{E}X = \sum_{x \in Im X} x \cdot p_X(x)$

$\mathbb{E}X = \sum_{\omega \in \Omega} X(\omega) \cdot P(\{\omega\})$ $\quad \to |\Omega| \leq |\mathbb{N}|$

$Y := g(X) \Rightarrow p_Y(y) = \sum_{x: g(x) = y} p_X(x)$

$\mathbb{E}[g(X)] = \sum_x g(x)\, p_X(x)$

$\mathbb{E}[X|B] = \sum_x x \cdot P[X = x|B]$

$\mathbb{E}X = \sum_i P(B_i) \cdot \mathbb{E}[X|B_i]$

$p_{X,Y}(x, y) = P[X = x, Y = y]$

$p_X(x) = \sum_y p_{X,Y}(x, y)$

$P[g(X,Y) = z] = \sum_{x,y: g(x,y) = z} p_{X,Y}(x, y)$

$X \perp Y \equiv p_{X,Y} = p_X \cdot p_Y$

$P[X+Y = z] = \sum_x p_X(x)\, p_Y(z - x)$

$\mathbb{E}[g(X,Y)] = \sum_{x,y} g(x,y)\, p_{X,Y}(x,y)$

Spojité:

$\mathbb{E}X = \int_{-\infty}^{\infty} x f_X(x)\, dx$

$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx$

$Var(X) = \mathbb{E}[(X-\mu)^2] = \int_{-\infty}^{\infty} x^2 f_X(x)\, dx - \mu^2$

$F_{X,Y}(x, y) = P[X \leq x, Y \leq y]$

$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy$

$P[(X,Y) \in S] = \int_S f_{X,Y}(x, y)\, dx\, dy$

$f_{X,Y}(x, y) = \frac{\partial F_{X,Y}}{\partial x \partial y}(x, y)$

$\mathbb{E}[g(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)\, dx\, dy$

$X \perp Y \equiv F_{X,Y} = F_X \cdot F_Y \Leftrightarrow f_{X,Y} = f_X \cdot f_Y$

$Z = X + Y \Rightarrow f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x)\, dx$

$X$ je spojitá $\equiv F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt$

$f_X = F_X'$ , $f_X \geq 0$, $\int_{-\infty}^{\infty} f_X = 1$

**Kvantilová fce**

$Q_X(p) = \min \{x \in \mathbb{R} \mid p \leq F_X(x)\} \to$ medián $:= Q_X(\frac{1}{2})$

**Nerovnosti:**

Markov: $X \geq 0 \Rightarrow P[X \geq a] \leq \frac{\mathbb{E}X}{a} \Leftrightarrow P[X \geq k\mathbb{E}X] \leq \frac{1}{k}$

Čebyšev: $\lambda > 0 \Rightarrow P[|X - \mu| \geq \lambda \cdot \sigma] \leq \frac{1}{\lambda^2}$ → umocnit a před $\mathbb{E}$

$X \sim Bin(m, p), Y \sim Bin(n, p) \Rightarrow X + Y \sim Bin(n + m, p)$

## Zákony velkých čísel

silný: $P[\lim \overline{X}_n = \mu] = 1$

slabý: $P[|\overline{X}_n - \mu| > \varepsilon] = 0$

↳ $\overline{X}_n \xrightarrow{P} \mu$ Čebyšev

$\mathbb{E}\overline{X}_n = \frac{1}{n}(\sum \mathbb{E}X_i) = \frac{n}{n}\mu = \mu$

$\mathrm{Var}\,\overline{X}_n = \frac{1}{n^2}\sum \mathrm{Var}\,X_i = \frac{n}{n^2}\sigma^2 = \frac{\sigma^2}{n}$

$\Rightarrow \sigma(\overline{X}_n) = \sigma/\sqrt{n}$ → nez. $X_i$

## Centrální limitní věta

$X_1, \dots X_n$ stejně r. n. n. v.

$Y_n := \frac{\overline{X}_n - \mathbb{E}\overline{X}_n}{\sigma(\overline{X}_n)} = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \approx N(0,1)$

## Intervalové odhady

↳ mám $\hat{\theta}$ odhad $\theta$

↳ statistiky $D \leq H$ } dolní a horní odhad

$[D, H]$ je $(1-\alpha)$-CI $\equiv P[D \leq \theta \leq H] = 1-\alpha$

Věta: Nestranný $\hat{\theta}$ pro $\theta$, $\hat{\theta} \sim N(\mu, \sigma^2)$

$\Rightarrow \hat{\theta} \pm z_{\alpha/2}\,\sigma(\hat{\theta})$ je $(1-\alpha)$-CI

↳ $z_{\alpha/2} := \Phi^{-1}(1-\frac{\alpha}{2})$

Věta: $X_1, \dots X_n \sim N(\mu, \sigma^2)$, neznám $\sigma$, chci $\mu$

$\Rightarrow \overline{X}_n \pm t_{\alpha/2}\frac{\hat{S}_n}{\sqrt{n}}$ je $(1-\alpha)$-CI

↳ $t_{\alpha/2} := \Psi^{-1}_{n-1}(1-\frac{\alpha}{2})$

## Bodové odhady

... odhaduju parametr $\theta$

$\mathrm{bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$ ... $\hat{\theta}$ = statistika co to odhaduje

↳ nestranný $\equiv$ bias $= 0$

↳ asympt. n. $\equiv \lim_{n\to\infty} \mathbb{E}\hat{\theta}_n = \theta$

↳ konzistentní $\equiv \hat{\theta}_n \xrightarrow{P} \theta$ → ZVČ: $\overline{X}_n$ je konz. $\mu$

$MSE(\hat{\theta}) := \mathbb{E}[(\hat{\theta}-\theta)^2] = \mathrm{Var}\,\hat{\theta} + \mathrm{bias}(\hat{\theta})^2$

→ $\overline{X}_n$ ... nestr. o. pro $\mu$

→ $\hat{S}^2_n := \frac{1}{n-1}\sum_i (X_i - \overline{X}_n)^2$ ... nestr. o. pro $\sigma^2$

- **Metoda momentů**

$m_r := \mathbb{E}[X^r]$

$\hat{m}_r := \frac{1}{n}\sum_i X_i^r$

} položím rovnost $\Rightarrow$ soustava rovnic

↳ $m_1 = \mathbb{E}X$, $m_2 = \sigma^2 + \mu^2$

- **Maximum Likelihood**

→ max $P[X_1 = x_1 \,\&\, \dots \,\&\, X_n = x_n] \stackrel{\text{nez.}}{=} \prod_i P[X_i = x_i; \theta]$ ← svět štěstí

$\Rightarrow L(\theta, \overline{x}) = \prod_i P[X_i = x_i] = \prod_i f_{X_i}(x_i)$, $\ell(\theta, \overline{x}) = \log L$

## Testování hypotéz

$H_0 \times H_1$

- vybrat statistický model: $X \sim N(\mu, \sigma^2)$
- statistický test $T = h(X_1, \dots X_n)$
- kritický obor: $W \subseteq \mathbb{R}$

} $t \in W \Rightarrow$ zamítnu $H_0$

$\Rightarrow$ chyba 1. druhu $\alpha = P[T \in W | H_0]$

$\Rightarrow$ chyba 2. druhu $\beta = P[T \notin W | H_1]$

↳ malé $\alpha \Rightarrow$ slabý test – nezamítne hodně chyb

↳ $1-\beta$ = síla testu

$\Rightarrow$ p-value = min $\alpha$, pro kterou bych $H_0$ zamítl

## Jednovýběrový test

Jednovýběrový test ... $X_1, \dots X_n \sim N(\mu, \sigma^2) \rightarrow H_0: \mu = \mathfrak{J}$

Dvouvýběrový test ... $X_1, \dots X_n \sim N(\mu_X, \sigma^2)$, $Y_1, \dots Y_m \sim N(\mu_Y, \sigma^2) \rightarrow H_0: \mu_X = \mu_Y$

Párový test ... $X_1, \dots X_n \sim N(\mu_X, \sigma^2)$, $Y_1, \dots Y_n \sim N(\mu_Y, \sigma^2)$ → 1výb: $[\mu_X - \mu_Y] = 0$

↳ chci statistiku $\overline{X}_n \Rightarrow$ normalizace

Z-test $Z = \frac{\overline{X}_n - \mathbb{E}\overline{X}_n}{\sigma/\sqrt{n}}$ ($\mathbb{E}\overline{X}_n = \mu$) $\sim N(0,1) \Rightarrow W = \{x \in \mathbb{R} \mid |x| > z_{\alpha/2}\}$

T-test $T = \frac{\overline{X}_n - \mathbb{E}\overline{X}_n}{\hat{S}_n/\sqrt{n}} \sim T(n-1) \Rightarrow W = \{x \in \mathbb{R} \mid |x| > t_{\alpha/2}\}$ ← $\Psi^{-1}_{n-1}(1-\alpha/2)$

↳ neznám $\sigma \Rightarrow$ odhad

## Testy dobré shody

→ kategorická data

Observed: $O_1, \dots O_k$

Expected $E_1, \dots E_k$

G-test: $G = 2 \cdot \sum_i O_i \log \frac{O_i}{E_i}$

$\chi^2$-test: $\chi^2 = \sum_i \frac{(E_i - O_i)^2}{E_i}$

Kritický obor: $W = \{x \in \mathbb{R} \mid x > \gamma\}$

↳ chceme $P[T > \gamma; H_0] = \alpha$

# Pravděpodobnost

**Def:** Pravděpodobnostní prostor je $(\Omega, \mathcal{F}, P)$, kde

- $\Omega$ ... množina el. jevů (klidně nespočetná)
- $\mathcal{F} \subseteq \mathcal{P}(\Omega)$ ... množina jevů (jev = nastane jeden z těch el. jevů)

  ↳ $F = \mathcal{P}(\Omega)$ funguje dobře jen pro spočetné $\Omega$

  i) $\emptyset, \Omega \in \mathcal{F}$

  ii) $A \in \mathcal{F} \Rightarrow A^c := \Omega \setminus A \in \mathcal{F}$

  iii) $A_1, A_2, \ldots \in \mathcal{F} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$

- $P : \mathcal{F} \to [0,1]$ ... pravděpodobnost

  i) $P(\Omega) = 1$

  ii) $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$ pro $A_i \dot\cup A_j$ → po dvou disj. jevy

$\underline{P(\emptyset) = 0}$ ∵ $P\left(\bigcup_{i=1}^{\infty} \emptyset\right) = \sum_{i=1}^{\infty} P(\emptyset)$ ... jinak: něco $\in [0,1] = \infty$

$P(A_1 \cup A_2) = P(A_1) + P(A_2)$ pro $A_1 \dot\cup A_2$ $(A_3, A_4, \ldots = \emptyset)$

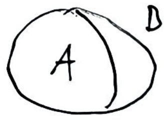
$A_1, A_2 \in \mathcal{F} \Rightarrow A_1 \cup A_2 \in \mathcal{F}$ ∵ $A_3, A_4, \ldots = \emptyset$

$\Rightarrow A_1 \cap A_2 \in \mathcal{F}$ ∵ $A_1 \cap A_2 = (A_1^c \cup A_2^c)^c$


**Tvrzení:** Nechť $(\Omega, \mathcal{F}, P)$ je p.p. a $A, B \in \mathcal{F}$. Potom

1) $P(A) + P(A^c) = 1$

2) $A \subseteq B \Rightarrow P(A) \leq P(B)$ & $P(B \setminus A) = P(B) - P(A)$.




3) $P(A \cup B) = P(A) - P(B) + P(A \cap B)$


4) $P\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} P(A_i)$ ... subaditivita

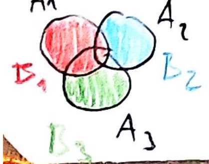
**Dk:**

1) $\Omega = A \dot\cup A^c \Rightarrow P(\Omega) = P(A \dot\cup A^c) = P(A) + P(A^c)$

2) $B = A \dot\cup (B \setminus A) \Rightarrow P(B) = P(A) + P(B \setminus A)$ & $P(B \setminus A) \geq 0$ ∵ to je pst.

3) $A \cup B = A \dot\cup (B \setminus (A \cap B)) \Rightarrow P(A \cup B) = P(A) + P(B \setminus (A \cap B)) = P(A) + P(B) - P(A \cap B)$

4) $B_i \subseteq A_i$: $B_1 := A_1$, $B_i := A_i \setminus \bigcup_{j<i} A_j \Rightarrow B_1, B_2, \ldots$ po dvou disj.




→ ale $\bigcup_{i=1}^{\infty} B_i = \bigcup_{i=1}^{\infty} A_i$ → přesně přes princip i. a e.

$\Rightarrow P(\bigcup A) = P(\bigcup B) = \sum_i P(B_i) \overset{B_i \subseteq A_i}{\leq} \sum_i P(A_i)$

Příklady p.p.:

- klasický: $\Omega$ konečná & uniformní $P$ $\Rightarrow P(A) = \frac{|A|}{|\Omega|}$
- diskrétní: $\Omega$ spočetná

$$p : \Omega \to [0,1], \quad \sum_{\omega \in \Omega} p(\omega) = 1, \quad P(A) = \sum_{\omega \in A} p(\omega)$$

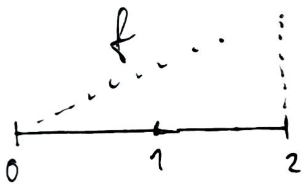
- geometrický: $\Omega \subseteq \mathbb{R}^d$ ...  $\subseteq \mathbb{R}^2$

$\mathcal{F} = \{A \in \mathcal{P}(\Omega) \mid \mathrm{Vol}(A) \text{ je definovaný}\}$ ... různé cihly a tak

$$P(A) = \frac{\mathrm{Vol}(A)}{\mathrm{Vol}(\Omega)}$$

- spojitý: $\Omega \subseteq \mathbb{R}^d$ ... $[0,1] \subseteq \mathbb{R}$

$$f : \Omega \to [0,\infty), \quad \int_\Omega f = 1, \quad P(A) = \int_A f$$




$f(x) = \frac{1}{2}x$ $\quad \int_0^2 f(x)\,dx = 1$ ✓

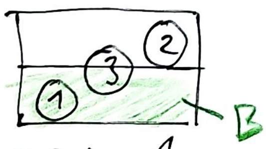
$\Omega = [0,2]$

$$\Rightarrow P([1,2]) = \int_1^2 f(x)\,dx = \frac{x^2}{4}\Big|_1^2 = 0.75$$

- Podmíněná pravděpodobnost

Def: Pst jevu A za podmínky, že nastal jev B je

$$P(A \mid B) := \frac{P(A \cap B)}{P(B)}, \quad P(B) \neq 0.$$

Ukázka: 


$P(1 \mid B) = \frac{P(1)}{P(B)} = \frac{\frac{1}{10}}{\frac{1}{2}} = \frac{1}{5}$ $\quad P(3 \mid B) = \frac{P(3 \cap B)}{P(B)} = \frac{1}{10}.$

$P(2 \mid B) = \frac{P(\emptyset)}{P(B)} = 0$

$P(B) = \frac{1}{2}$

$P(1) = P(2) = P(3) = \frac{1}{10}$

Tvrzení: Nechť je $(\Omega, \mathcal{F}, P)$ p.p., $B \in \mathcal{F}$ a $P(B) > 0$. Pro $A \in \mathcal{F}$ definujeme $Q(A) := P(A|B)$. Tvrdíme, že $(\Omega, \mathcal{F}, Q)$ je také p.p.

Dk: Funkce $Q$ musí být pravděpodobnost, tedy musí splňovat

i) $Q(\Omega) = 1$ ... $Q(\Omega) = P(\Omega|B) = \frac{P(\Omega \cap B)}{P(B)} = 1$ ✓

ii) $Q(\dot{\bigcup}_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} Q(A_i)$, $A_i \,\dot\cup\, A_j$ — de-Morgan — disjunkt

$$\to Q(\bigcup A_i) = P(\bigcup A_i | B) = \frac{P(\bigcup_i A_i \cap B)}{P(B)} = \frac{1}{P(B)} \cdot P\Big(\bigcup_i (A_i \cap B)\Big) =$$

$$= \frac{1}{P(B)} \sum_i P(A_i \cap B) = \sum_i P(A_i|B) = \sum_i Q(A_i). \quad \blacksquare$$

Příklad: $A_i \sim$ i-tá karta v balíčku 32 karet je ♡.

$$P(A_1 \cap A_2) = P(A_1) \cdot P(A_2|A_1) = \frac{8}{32} \cdot \frac{7}{31}$$

↳ kdybych tohle chtěl z definice, tak jsem si nepomohl
⇒ vlastně to počítám v novém p.p.

Věta (o zřetězeném podmiňování): Nechť $A_1, \dots A_n \in \mathcal{F}$, $P(A_1 \cap \dots \cap A_{n-1}) > 0$.

$$P(A_1 \cap \dots \cap A_n) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1 \cap A_2) \cdot \ldots \cdot P(A_n | A_1 \cap \dots \cap A_{n-1}).$$

Dk: indukcí.

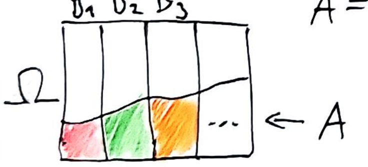
Def: Rozklad množiny $\Omega$ je systém disj. množin $\mathcal{B} = \{B_1, B_2, \dots\}$ spočetně mnoho t.ž.:

$\forall i: B_i \subseteq \Omega$ & $\bigcup \mathcal{B} = \Omega$

Věta (o rozboru případů): Nechť $\mathcal{B}$ je rozklad $\Omega$ a $A \in \mathcal{F}$. Potom platí

$$P(A) = \sum_i P(B_i) \cdot P(A|B_i),$$ kde pro $P(B_i) = 0$ dodefinujeme $P(A|B_i) := 0$.

Dk:




$A = (A \cap B_1) \cup (A \cap B_2) \cup (A \cap B_3) \cup \dots$ disj.

$$\Rightarrow P(A) = \sum_i P(A \cap B_i) = \sum_i P(B_i) \cdot P(A|B_i). \quad \blacksquare$$

Věta (Bayesova): Nechť $\mathcal{B}$ je rozklad $\Omega$ a $A \in \mathcal{F}$. Platí

$$P(B_j|A) = \frac{P(B_j \cap A)}{P(A)} = \frac{P(B_j) \cdot P(A|B_j)}{\sum_i P(B_i) \cdot P(A|B_i)}.$$

$\to$ speciálně $P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)}$.

**Příklad:** Gambler's ruin.

→ férová hra: vsadí \$1 a je 50% šance vyhrát \$1 zpět.

$p(a) :=$ P[odejde s \$N, když začal s \$a]  $0 \le a \le N$

$A \sim p(a)$
$B \sim$ výhra v 1. kole

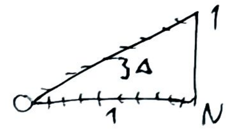
$$P(A) = P(B)\cdot P(A|B) + P(B^c)\cdot P(A|B^c)$$ prohra v 1. kole

$$\Rightarrow p(a) = \frac{1}{2}\cdot p(a+1) + \frac{1}{2}\cdot p(a-1)$$

$p(0) = 0$
$p(N) = 1$

$$2p(a) = p(a+1) + p(a-1)$$

$$\Delta = p(a+1) - p(a) = p(a) - p(a-1)$$




$$\Rightarrow p(a) = \frac{a}{N}$$

**Příklad:**

$a = n+1$ - bitové číslo  101 . . . . . 0|1
$b = n$ - bitové číslo  011 . . . . . 1|

$P[|a|_1 > |b|_1] = ?$  $\Rightarrow X :=$ pst, že ve 2n-bit číslech je stejný #1

→ poslední bit a je buď  / pst, že v prvním z nich je více

- 0 $\Rightarrow P[|a|_1 > |b|_1 \mid 0] = \frac{1-X}{2}$
- 1 $\Rightarrow P[|a|_1 > |b|_1 \mid 1] = X + \frac{1-X}{2}$ → stejně nebo v a víc

$$\Rightarrow P[|a|_1 > |b|_1] = \frac{1}{2}\cdot\frac{1-X}{2} + \frac{1}{2}\left(X + \frac{1-X}{2}\right) = \frac{1}{2}\cdot(X + 1 - X) = \frac{1}{2}$$

• Nezávislost jevů

**Def:** Jevy $A, B \in \mathcal{F}$ jsou nezávislé, $A \perp B \equiv P(A \cap B) = P(A)\cdot P(B)$

☼ $A \perp B \iff \frac{P(A\cap B)}{P(B)} = P(A) \iff P(A|B) = P(A)$

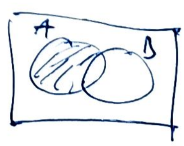
**Def:** Pokud $P(A|B) > P(A)$ ... $A, B$ jsou pozitivně korelované
$P(A|B) < P(A)$ ... $A, B$ jsou negativně korelované

**Def:** Nechť $I$ je indexová množina. Jevy $\{A_i \mid i \in I\}$ jsou nezávislé $\equiv$

$$\forall J \subseteq I : P\left(\bigcap_{j\in J} A_j\right) = \prod_{j \in J} P(A_j).$$

**Def:** Jevy $A, B$ jsou nezávislé za podmínky $C$, $A \perp_C B \equiv$

$$P(A \cap B \mid C) = P(A|C)\cdot P(B|C)$$




**Věta:** $A \perp B \Rightarrow A \perp B^c$ & $A^c \perp B^c$

**Dk:** $P(A \cap B^c) = P(A \setminus (A\cap B)) = P(A) - P(A)\cdot P(B) = P(A)\cdot(1 - P(B)) = P(A)\cdot P(B^c)$

$A^c \perp B^c$ ∵ $A \perp B^c \Rightarrow A^c \perp B^c$.

• Diskrétní náhodné veličiny

Def: Nechť $(\Omega, \mathcal{F}, P)$ je p.p. Funkce $X: \Omega \to \mathbb{R}$ je diskrétní n.v. $\equiv$

1, $Im(X)$ je nejvýše spočetný ... obor hodnot

2, $\forall x \in \mathbb{R}: \{X = x\} := \{\omega \in \Omega \,|\, X(\omega) = x\} \in \mathcal{F}$ $\quad X^{-1}(x)$

↳ to chceme, abychom se mohli ptát na $P[X=x]$

Poznámka: Náhodná veličina X určuje p.p. $(\Omega', \mathcal{F}', p)$, kde

$\Omega' = Im(X)$, $\mathcal{F}' = \mathcal{P}(\Omega')$, $p(x) = P[X=x]$

→ tento prostor, resp. fci p, nazýváme rozdělení / distribuce X.

Def: Pravděpodobnostní fce (probability mass function = pmf) d.n.v. X

je $p_X : \mathbb{R} \to [0,1]$, $x \mapsto P[X=x]$

$$\sum_{x \in ImX} p_X(x) = 1 \quad \because \quad \Omega = \bigcup_{x \in ImX} \{X = x\}$$

P P

$\Omega$ $\boxed{X=1 \,|\, X=2 \,|\, X=3 \,|\, \dots}$

→ rozklad $\Omega$

Bernoulliho rozdělení

$X \sim Ber(p)$ ... $P[X=1] = p_X(1) = p$

$P[X=0] = p_X(0) = 1-p$

Indikátor jevu $A \in \mathcal{F}$

$$I_A(\omega) = \begin{cases} 1 \dots \omega \in A & \dots A \text{ nastal} \\ 0 \dots \omega \notin A & \dots A \text{ nenastal} \end{cases}$$

$I_A \sim Ber(P(A))$

Geometrické rozdělení

X = na kolikátý pokus se poprvé něco povedlo (pokusy jsou nezávislé s pstí p)

$X \sim Geom(p)$

$p_X(1) = p$

$p_X(2) = (1-p)p$

$p_X(3) = (1-p)^2 p$

$$\Rightarrow p_X(k) = \begin{cases} (1-p)^{k-1} p & , \text{ pro } k \in \mathbb{N}^+ \\ 0 & , \text{ jinak} \end{cases}$$

→ ještě potřebujeme $\sum_{x \in ImX} p_X(x) = 1$

$$\sum_{k=1}^{\infty} p_X(k) = \sum_{k=1}^{\infty} (1-p)^{k-1} \cdot p = p \cdot \sum_{k=0}^{\infty} (1-p)^k = p \cdot \frac{1}{1-(1-p)} = 1 \quad \checkmark$$

## Binomické rozdělení

$X = \#$ úspěchů při $n$ nezávislých pokusech s pstí $p$

$X \sim Bin(n,p)$, $\quad p_X(k) = \binom{n}{k} \cdot p^k (1-p)^{n-k} \quad \Rightarrow \sum_{k=0}^{\infty} p_X(k) = (p + (1-p))^n = 1$

Součet nez. bin. roz. je bin. rozdělení

$X \sim Bin(m,p)$, $\quad Y \sim Bin(n,p)$, $\quad X \perp Y \Rightarrow Z = X+Y \sim Bin(n+m, p)$

$$p_Z(k) = P[X+Y=k] = \sum_{i=0}^{k} P[X=i \;\&\; Y=k-i] = \sum_{i=0}^{k} p_X(i) \cdot p_Y(k-i)$$

$$= \sum_{i=0}^{k} \binom{m}{i} p^i (1-p)^{m-i} \binom{n}{k-i} \cdot p^{k-i} (1-p)^{n-k+i} = p^k (1-p)^{m+n-k} \underbrace{\sum_{i=0}^{k} \binom{m}{i}\binom{n}{k-i}}_{\binom{m+n}{k}}$$

## Poissonovo rozdělení

→ aproximace bin. r. pro velké $n$ a malou $p$

$X \sim Pois(\lambda) \quad \dots \quad p_X(k) = \frac{\lambda^k}{k!} e^{-\lambda}$, $\quad \lambda > 0$

$$\hookrightarrow \sum_{k=0}^{\infty} p_X(k) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} \cdot e^{\lambda} = 1 \quad \checkmark$$

$Pois(\lambda) \approx Bin(n, \frac{\lambda}{n})$

↳ nechť $X \sim Pois(\lambda)$ a $X_n \sim Bin(n, \frac{\lambda}{n}) \Rightarrow$ hledíme $\lim_{n\to\infty} p_{X_n}(k) = p_X(k)$

$$p_{X_n}(k) = \binom{n}{k} \cdot \left(\frac{\lambda}{n}\right)^k \cdot \left(1 - \frac{\lambda}{n}\right)^{n-k} = \frac{n^{\underline{k}}}{k!} \cdot \frac{\lambda^k}{n^k} \cdot \left(1 - \frac{\lambda}{n}\right)^n \cdot \cancel{\left(1 - \frac{\lambda}{n}\right)^{-k}} \to 1 \qquad n \to \infty, \; k \in \mathbb{N}_0$$

$$= \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{n}\right)^n \cdot \cancel{\frac{n^{\underline{k}}}{n^k}} \to 1 \; = \frac{\lambda^k}{k!} \cdot e^{-\lambda} = p_X(k) \quad \dots \quad \left(1 + \frac{x}{n}\right)^n \to e^x$$

## • Střední hodnota

**Def:** Střední hodnota d.n.v. $X$ je $\mathbb{E}X := \sum_{x \in \mathrm{Im}X} x \cdot p_X(x)$.

↳ pokud řada nekonverguje absolutně, tak není definovaná

👁 Pokud je $X$ d.n.v. na diskrétním p.p. (tedy $\Omega$ je nejvýše spočetná), tak

$$\mathbb{E}X = \sum_{\omega \in \Omega} X(\omega) \cdot P(\{\omega\})$$

Dk: $$\sum_{\omega \in \Omega} X(\omega) P(\{\omega\}) = \sum_{x \in \mathrm{Im}X} \sum_{\underbrace{\omega : X(\omega) = x}_{= X^{-1}(x)}} X(\omega) P(\{\omega\}) = \sum_{x \in \mathrm{Im}X} x \sum_{\omega \in X^{-1}(x)} P(\{\omega\}) = \sum_{x \in \mathrm{Im}X} x \cdot P[X = x]$$ ∎

Příklady:

① $X \sim \mathrm{Ber}(p) \Rightarrow \mathbb{E}X = 1 \cdot p + 0 \cdot (1-p) = p$

② $X \sim \mathrm{Bin}(n,p) \Rightarrow \mathbb{E}X = \sum_{k=0}^{n} k \cdot \binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=1}^{n} \underbrace{k \cdot \binom{n}{k}}_{n\binom{n-1}{k-1}} p^k (1-p)^{n-k} =$

$\mathbb{E}X = p \cdot n$

$$= n \sum_{k=1}^{n} \binom{n-1}{k-1} p^k (1-p)^{n-k} = n \sum_{k=0}^{n-1} \binom{n-1}{k} p^{k+1} (1-p)^{n-1-k}$$

$$= n \cdot p \sum_{k=0}^{n-1} \binom{n-1}{k} p^k (1-p)^{(n-1)-k} = n \cdot p \cdot (p + 1 - p)^{n-1} = np$$

Lemma o džbánu:

③ $X \sim \mathrm{Geom}(p) \Rightarrow \mathbb{E}X = \sum_{k=1}^{\infty} k \cdot p_X(k) = \sum_{k=1}^{\infty} k \cdot (1-p)^{k-1} p = p \cdot \sum_{k=1}^{\infty} k (1-p)^{k-1}$

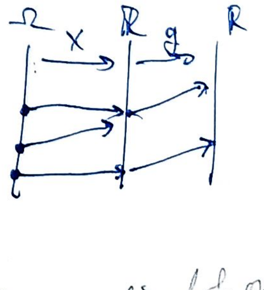
$\sum_{k=0}^{\infty} q^k = \frac{1}{1-q} \Rightarrow \frac{d}{dx} \sum = \sum_{k=1}^{\infty} k q^{k-1} = \frac{1}{(1-q)^2}$ $\Big|$ $= p \cdot \frac{1}{(1-(1-p))^2} = \frac{1}{p}$

## • Pravidlo naivního statistika – PNS

👁 Pokud je $X$ d.n.v. a $g: \mathbb{R} \to \mathbb{R}$, potom je $Y := g(X)$ také d.n.v.

Dk: Chceme $Y$ d.n.v., tedy

1) $Y : \Omega \to \mathbb{R}$ ✓
2) $\mathrm{Im}(Y)$ je nejvýše spočetný $\because$ $|\mathrm{Im}(X \circ g)| \leq |\mathrm{Im}(X)|$




3) pro $\forall y \in \mathbb{R}$: $Y^{-1}(y) = \{Y = y\} \in \mathcal{F}$

$Y^{-1}(y) = \bigcup_{\substack{x \in \mathrm{Im}X \\ g(x) = y}} X^{-1}(x)$ & $X^{-1}(x) \in \mathcal{F} \Rightarrow Y^{-1}(y) \in \mathcal{F}$

↳ spočetné sjednocení ↳ jeden z axiomů pravděpodobnosti ∎

**Věta (PNS):** Pokud $X$ je d.n.v. a $g: \mathbb{R} \to \mathbb{R}$, potom
$$\mathbb{E}[g(X)] = \sum_{x \in \mathrm{Im}X} g(x)\, p_X(x)$$

y je určeno jednoznačně

$$\sum_x \sum_{y: g(x)=y} g(x)\, p_X(x) = \sum_x{}' g(x)\, p_X(x)$$

**Dk:** Označme $Y := g(X)$. Z definice
$$\mathbb{E}[Y] = \sum_{y \in \mathrm{Im}Y} y \cdot P[Y=y] = \sum_y y \cdot \sum_{x: g(x)=y} P[X=x] = \sum_y \sum_{x: g(x)=y} g(x)\, p_X(x)$$

disj.

$\to$ jako v minulém důkazu: $\{Y=y\} = Y^{-1}(y) = \bigcup_{\substack{x \in \mathrm{Im}X \\ g(x)=y}} X^{-1}(x) = \dot{\bigcup_{\substack{x \in \mathrm{Im}X \\ g(x)=y}}} \{X=x\}$ ∎

**Věta (vlastnosti $\mathbb{E}$):** $X, Y$ jsou d.n.v.

① $P[X \geq 0] = 1$ & $\mathbb{E}X = 0 \Rightarrow P[X=0] = 1$

② $\mathbb{E}X \geq a \Rightarrow P[X \geq a] > 0$

③ $\mathbb{E}[aX+b] = a\mathbb{E}X + b$

④ $\mathbb{E}[X+Y] = \mathbb{E}X + \mathbb{E}Y$ ... zatím pouze pro diskrétní p.p.

**Dk:** ① $\mathbb{E}X = \sum_x x \cdot p_X(x) = 0$, všechny členy $\geq 0 \Rightarrow$ všechny členy $= 0$

② $\mathbb{E}X = \sum_x x\, p_X(x) \geq a$. Kdyby $P[X \geq a] = 0$, tak $\forall x \geq a: P(X=x) = 0$

$\Rightarrow$ tedy $\sum_x x\, p_X(x) < \sum_x a\, p_X(x) = a \sum_x p_X(x) = a \Rightarrow a \leq \sum < a$ ↯

③ $\mathbb{E}[aX+b] = \mathbb{E}[g(X)] = \sum_x (ax+b)\, p_X(x) = a \sum_x x\, p_X(x) + b \sum_x p_X(x) = a\mathbb{E}X + b$.

④ $\mathbb{E}[X+Y] = \sum_{\omega \in \Omega} (X(\omega) + Y(\omega)) \cdot P(\{\omega\}) = \sum_\omega X(\omega) P(\omega) + \sum_\omega Y(\omega) P(\omega) = \mathbb{E}X + \mathbb{E}Y$ ∎

• **Linearita střední hodnoty:** $\mathbb{E}\left[\sum_i a_i X_i\right] = \sum_i a_i \mathbb{E}X_i$

$\to$ aplikace: $X \sim \mathrm{Bin}(n,p) \longrightarrow \mathbb{E}X = ?$ (víme $n \cdot p$)

$X = \#$ úspěchů v $n$ pokusech $\Rightarrow X_i =$ Indikátor, za $i$-tý pokus úspěl

$\hookrightarrow X_i \sim \mathrm{Ber}(p) \Rightarrow \mathbb{E}X_i = p$

$\mathbb{E}X = \sum_{i=1}^n \mathbb{E}X_i = n \cdot p$

**Def (podmíněná $\mathbb{E}$):** Nechť je $X$ d.n.v., $B$ jev, $P(B) > 0$. Definujeme
$$\mathbb{E}[X|B] := \sum_{x \in \mathrm{Im}X} x \cdot P[X=x|B]$$
$\to$ opět pokud to absolutně konverguje

**Věta (o celkové $\mathbb{E}$):** Nechť je $B_1, B_2, \ldots$ rozklad $\Omega$ a $X$ je d.n.v. Potom
$$\mathbb{E}X = \sum_i P(B_i) \cdot \mathbb{E}[X|B_i]$$

**Dk:** $\sum_x x P[X=x] = \sum_x x \cdot \sum_i P(B_i) \cdot P[X=x|B_i] = \sum_i P(B_i) \cdot \underbrace{\sum_x x \cdot P[X=x|B_i]}_{\mathbb{E}[X|B_i]}$ ∎

$\hookrightarrow$ věta o celkové pst.

Příklady:

① $X =$ celkové skóre při házení kostkou, opakuji, pokud padla 6

$$\mathbb{E}X = P(\text{⚅})\cdot\mathbb{E}[X \mid \text{⚅}] + (1-P(\text{⚅}))\cdot\mathbb{E}[X \mid \text{⚅}^c]$$

$$= \frac{1}{6}\cdot(6+\mathbb{E}X) + \frac{5}{6}\cdot 3 \qquad \hookrightarrow \frac{1}{5}(1+2+3+4+5) = 3$$

$$= \frac{1}{6}\mathbb{E}X + \frac{7}{2} \implies \mathbb{E}X = \frac{42}{10} = \underline{\underline{4.2}}$$

② $n$-bit číslo, $P(1) = p$

$$\begin{matrix} 1 & 2 & 3 & 4 & & \\ 1 & 0 & 1 & 1 & 0 & 0 \end{matrix}$$

$X = \#$ podřetězců $101 \implies \mathbb{E}X = ?$

$A_i :=$ na $i$-té pozici začíná $101$. $\longrightarrow P[A_i] = p(1-p)p = p^2(1-p)$

$$\Downarrow \mathbb{E} I_{A_i} = p^2(1-p)$$

$$\longrightarrow X = \sum_{i=1}^{n-2} I_{A_i} \implies \mathbb{E}X = \sum_{i=1}^{n-2} \mathbb{E} I_{A_i} = \underline{\underline{(n-2)\cdot p^2(1-p)}}$$

③ Nechť $A$ je jev a $I_A$ jeho indikátor. Pokud $A = A_1 \cup A_2 \cup \ldots \cup A_n$, tak

$$1 - I_A = \prod_{i=1}^{n} (1 - I_{A_i}) \quad \ldots \text{ celkem obvious}$$

$$\mathbb{E}[1 - I_A] = 1 - P(A) = \mathbb{E}\left[\prod_{i=1}^{n} (1 - I_{A_i})\right]$$

$\hookrightarrow$ roznásobit + lin $\mathbb{E} \implies$ princip in a exkluze

Větička: Pokud je $X$ d.n.v. a $\mathrm{Im}(X) \subseteq \mathbb{N}$, potom

$$\mathbb{E}X = \sum_{n=0}^{\infty} P[X > n] \longrightarrow s(t) := P[X > t] \ \ldots \text{ survival function}$$

Dk: $\mathbb{E}X = 1\cdot p_X(1) + 2p_X(2) + 3p_X(3) + \ldots$

$$\begin{aligned} &= p_1 + p_2 + p_3 + p_4 + \ldots &&= P(X>0) \\ &\quad + p_2 + p_3 + p_4 + \ldots &&= P(X>1) \\ &\quad\quad + p_3 + p_4 + \ldots &&= P(X>2) \\ &\quad\quad\quad + p_4 + \ldots &&= P(X>3) \end{aligned}$$

Příklad: $X \sim \mathrm{Geom}(p)$

$$\mathbb{E}X = \sum_{n=0}^{\infty} P(X > n) = \sum_{n=0}^{\infty} (1-p)^n = \frac{1}{1-(1-p)} = \frac{1}{p}$$

**Def:** Rozptyl d.n.v. $X$ je $var(X) := \mathbb{E}[(X-\mathbb{E}X)^2]$

Směrodatná odchylka $\sigma_X := \sqrt{var X}$

Variační koeficient $CV_X := \frac{\sigma_X}{\mathbb{E}X}$ ... pouze pro $\mathbb{E}X > 0$

**Věta:** $var(X) = \mathbb{E}[X^2] - (\mathbb{E}X)^2 = \mathbb{E}[X(X-1)] + \mathbb{E}X - (\mathbb{E}X)^2$

**Dk:** Označme $\mu := \mathbb{E}X$

$$var X = \mathbb{E}[(X-\mu)^2] = \mathbb{E}[X^2 - 2X\mu + \mu^2] = \mathbb{E}[X^2] - 2\mu\mathbb{E}X + \mu^2 = \mathbb{E}[X^2] - (\mathbb{E}X)^2$$

$$\Rightarrow \mathbb{E}[X(X-1)] = \mathbb{E}[X^2] - \mu \Rightarrow \mathbb{E}[X(X-1)] + \mu - (\mathbb{E}X)^2 = \mathbb{E}[X^2] - (\mathbb{E}X)^2 = var X$$

■

👁 $var(X) \geq 0 \Rightarrow \mathbb{E}[X^2] \geq (\mathbb{E}X)^2$

**Důsledek:** nerovnost mezi aritmetickým a kvadratickým průměrem

$$\frac{x_1^2 + x_2^2 + \dots + x_n^2}{n} \geq \left(\frac{x_1 + x_2 + \dots + x_n}{n}\right)^2 \Rightarrow \frac{1}{n}\sum_i x_i \leq \sqrt{\frac{1}{n}\sum_i x_i^2}$$

→ dokonce platí i pro vážený průměr: $\sum_i x_i w_i \leq \sqrt{\sum_i x_i^2 w_i}$, $\sum_i w_i = 1$

👁 $var(X) = 0 \iff X$ je konstantní fce

**Příklady:**

① $X \sim Ber(p) \Rightarrow \mathbb{E}X = p$

↳ $Var(X) = \mathbb{E}[(X-p)^2] = p \cdot (1-p)^2 + (1-p) \cdot p^2 = \underline{p(1-p)}$

↳ $X^2 = X \Rightarrow Var(X) = \mathbb{E}[X^2] - (\mathbb{E}X)^2 = p - p^2$

docela hnus

② $X \sim Bin(n, p) \Rightarrow \mathbb{E}X = n \cdot p$

a) z definice: PNS: $g(x) = (x - np)^2 \Rightarrow var X = \sum_{k=0}^{n} (k-np)^2 \binom{n}{k} p^k (1-p)^{n-k}$

b) $var X = \mathbb{E}[X(X-1)] + np - (np)^2$

$$\mathbb{E}[X(X-1)] \overset{PNS}{=} \sum_{k=0}^{n} k(k-1)\binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=2}^{n} (k-1)\underbrace{k\binom{n}{k}}_{n\binom{n-1}{k-1}} p^k (1-p)^{n-k}$$

$$= (n-1)n \sum_{k=2}^{n} \binom{n-2}{k-2} p^k (1-p)^{n-k} = n(n-1)p^2 \sum_{k=0}^{n-2} \binom{n-2}{k} p^k (1-p)^{n-2-k}$$

$$= n(n-1)p^2 \cdot (p + (1-p))^{n-2} = n(n-1)p^2$$

$\Rightarrow var X = p^2 n(n-1) + pn - p^2 n^2 = \underline{n \cdot p \cdot (1-p)}$ → $n$-krát víc než $Ber(p)$

③ $X \sim Pois(\lambda)$

→ obdobně $\mathbb{E}X = Var X = \lambda$ ⇝ $Pois(\lambda) \approx Bin(n, \frac{\lambda}{n})$ ↗ $\mathbb{E}X = \lambda = n \cdot \frac{\lambda}{n}$ ↘ $Var X = \lambda \approx \lambda \cdot (1 - \cancel{\frac{\lambda}{n}})$

④ $X \sim Geom(p) \Rightarrow \mathbb{E}X = \frac{1}{p}$

$Var X = \mathbb{E}[X^2] - (\mathbb{E}X)^2 = \mathbb{E}[X^2] - \frac{1}{p^2}$

$\Rightarrow B_1 :=$ 1. pokus úspěch, $B_2 := B_1^c$ } $\mathbb{E}[X^2] = P[B_1] \cdot \mathbb{E}[X^2 | B_1] + P[B_2] \cdot \mathbb{E}[X^2 | B_2] =$

$= p \cdot 1^2 + (1-p) \cdot \mathbb{E}[(1+X)^2] =$

$= p + (1-p) \cdot (1 + 2 \cdot \frac{1}{p} + \mathbb{E}[X^2])$

$\Rightarrow \mathbb{E}[X^2](1-(1-p)) = p + 1 - p + \frac{2(1-p)}{p} \Rightarrow \mathbb{E}[X^2] = \frac{2-p}{p^2}$

$\Rightarrow Var(X) = \frac{2-p}{p^2} - \frac{1}{p^2} = \underline{\underline{\frac{1-p}{p^2}}}$

⑤ $Y = aX + b$

$\Rightarrow Var(Y) = a^2 Var(X)$ } imunní vůči posunům

$\sigma(Y) = |a| \sigma_X$

$CV_Y = CV_X$, pro $a \geq 0$ $\Rightarrow$ CV je imunní vůči škálování

• **Náhodné vektory** ↗ na stejném p.p.

**Def:** Pro d.n.v. $X, Y$ definujeme sdruženou pstní fci (joint pmf) jako

$$p_{X,Y}(x,y) := P[X = x \ \& \ Y = y]$$

Aby to byla pst, tak musí pro $\forall x, y \in \mathbb{R}$: $\{X = x \ \& \ Y = y\} \in \mathcal{F}$

**Ukázka:**

| X\Y | 0 | 1 | 2 |
|---|---|---|---|
| 0 | 1/4 | 1/6 | 1/12 |
| 1 | 1/6 | 1/4 | 1/12 |

$\rightarrow p_X(0) = P[X=0] = \frac{1}{4} + \frac{1}{6} + \frac{1}{12}$

$p_Y(2) = P[Y=2] = 1/12 + 1/12$

**Def:** Rozdělení $p_{X,Y}$ říkáme sdružené rozdělení. Rozdělení jednotlivých složek $p_X$ a $p_Y$ nazýváme marginální rozdělení

👁 $p_X(x) = \sum_{y \in Im Y} p_{X,Y}(x,y)$ $\qquad p_Y(y) = \sum_{x \in Im X} p_{X,Y}(x,y)$

↖ přes disj. sjednocení

👁 Z m. r. nelze odvodit sdružené r.

| | 0 | 1 |
|---|---|---|
| 0 | 1 | 0 |
| 1 | 0 | 1 |

a

| | 0 | 1 |
|---|---|---|
| 0 | ½ | ½ |
| 1 | ½ | ½ |

mají stejná marg. rozdělení

↳ jde to, pokud jsou $X, Y$ nezávislé

Věta: Pokud je $(X,Y)$ d.n. vektor a $g: \mathbb{R}^2 \to \mathbb{R}$, potom je $g(X,Y)$ d.n.v. a platí

$$p_Z(z) = P[g(X,Y) = z] = \sum_{x,y: g(x,y) = z} p_{X,Y}(x,y)$$

Dk: Zjevně je $\{g(X,Y) = z\}$ disj. sjednocením jevů $\{X = x \ \& \ Y = y\}$ pro dvojice, pro které $g(x,y) = z$. Navíc se stačí omezit na $x \in \mathrm{Im}\,X$ a $y \in \mathrm{Im}\,Y$, takže to sjednocení je spočetné $\Rightarrow$ lze použít axiom $P(\bigcup) = \sum P$. ∎

- Nezávislost d.n.v.

Def: $X, Y$ jsou nezávislé d.n.v. $X \perp Y \equiv \forall x, y \in \mathbb{R}: \{X = x\} \perp \{Y = y\}$

$X \perp Y \iff \forall x, y: p_{X,Y}(x,y) = p_X(x)\, p_Y(y)$

Def: $X_1, \dots X_n$ jsou nezávislé $\equiv$ jevy $\{X_1 = x_1\}, \dots \{X_n = x_n\}$ jsou nezávislé.

- Multinomické rozdělení

Házíme $n$-krát kostkou ... férovou $\Rightarrow p_1 = p_2 = \dots = p_6 = \frac{1}{6}$

$$\left.\begin{array}{l} X_1 := \#\text{ jedniček} \sim Bin(n, p_1) \\ \vdots \\ X_n := \#\text{ šestek} \sim Bin(n, p_6) \end{array}\right\} (X_1, \dots X_6) \sim Mult(n, p_1, \dots p_6)$$

$$\Rightarrow p_{X_1, \dots X_6}(k_1, \dots k_6) = \binom{n}{k_1, \dots k_6} \cdot p_1^{k_1} \cdot p_2^{k_2} \cdot \dots \cdot p_6^{k_6}$$

Def: Multinomický koeficient $\binom{n}{k_1, \dots k_m} := \frac{n!}{k_1! \cdots k_m!}$, $\sum_i k_i = n$

↳ intuice: # způsobů jak rozdělit $n$ prvků do $m$ množin o $k_1, \dots k_m$ prvcích

Věta (konvoluční vzorec): Nechť $X, Y$ jsou d.n.v. a $Z := X + Y$. Potom platí

$$p_Z(z) = P[X + Y = z] = \sum_{x \in \mathrm{Im}\,X} P[X = x \ \& \ Y = z - x]$$

pokud $X \perp Y$, potom $p_Z(z) = \sum_x p_X(x)\, p_Y(z - x)$.

Dk: Použijeme předchozí větu.

**Věta** (PNS pro vektory): Nechť $(X,Y)$ je d.n. vektor a $g:\mathbb{R}^2\to\mathbb{R}$. Potom platí

$$\mathbb{E}[g(X,Y)]=\sum_{x,y} g(x,y)\cdot p_{X,Y}(x,y)$$

**Dk**: Označme $Z:=g(X,Y)$ → podle věty z předchozí stránky

$$\mathbb{E}Z=\sum_z z\, p_Z(z)=\sum_z z\cdot\sum_{x,y:g(x,y)=z} p_{X,Y}(x,y)=\sum_z\sum_{x,y:g(x,y)=z} z\cdot p_{X,Y}(x,y)$$

$$=\sum_{x,y}\sum_{z:g(x,y)=z} z\cdot p_{X,Y}(x,y)=\sum_{x,y} g(x,y)\, p_{X,Y}(x,y).$$ ∎

→ dopsud jen diskrétní

**Věta**: Nechť jsou $X,Y$ d.n.v. v libovolném p.p., $a,b\in\mathbb{R}$. Potom

$$\mathbb{E}[aX+bY]=a\mathbb{E}X+b\mathbb{E}Y.$$

**Dk**: Použijeme PNS s $g(x,y):=ax+by$.

$$\mathbb{E}[aX+bY]=\sum_{x,y}(ax+by)\,p_{X,Y}(x,y)=a\sum_{x,y}x\,p_{X,Y}(x,y)+b\sum_{x,y}y\,p_{X,Y}(x,y)$$

$$\rightsquigarrow\sum_{x,y}x\,p_{X,Y}(x,y)=\sum_x x\cdot\sum_y p_{X,Y}(x,y)=\sum_x x\,p_X(x)=\mathbb{E}X$$ ∎

**Věta**: $X\perp Y\Rightarrow \mathbb{E}XY=\mathbb{E}X\cdot\mathbb{E}Y$.

**Dk**: PNS pro funkci $g(x,y)=xy$:

$$\mathbb{E}[XY]=\sum_{x,y}xy\cdot p_{X,Y}(x,y)=\sum_{x,y}x\cdot y\cdot p_X(x)\,p_Y(y)=\sum_x\sum_y x\,p_X(x)\,y\,p_Y(y)$$

$$=\sum_x x\,p_X(x)\cdot\sum_y y\,p_Y(y)=\mathbb{E}X\cdot\mathbb{E}Y$$

- **Kovariance**

**Def**: Pro n.v. $X,Y$ definujeme jejich kovarianci jako

$$\operatorname{cov}(X,Y):=\mathbb{E}[(X-\mathbb{E}X)(Y-\mathbb{E}Y)]$$

**Tvrzení**: $\operatorname{cov}(X,Y)=\mathbb{E}XY-\mathbb{E}X\cdot\mathbb{E}Y$ ... linearita střední hodnoty

👁 $X\perp Y\Rightarrow\operatorname{cov}(X,Y)=0$

**Tvrzení**: Vlastnosti kovariance:

1) $\operatorname{var}(X)=\operatorname{cov}(X,X)$

2) $\operatorname{cov}(X,aY+bZ+c)=a\cdot\operatorname{cov}(X,Y)+b\cdot\operatorname{cov}(X,Z)$ ... linearita

**Věta**: Pro d.n.v. $X, Y$ platí

$$\mathrm{var}(X+Y) = \mathrm{var}(X) + \mathrm{var}(Y) + 2\cdot \mathrm{cov}(X,Y)$$

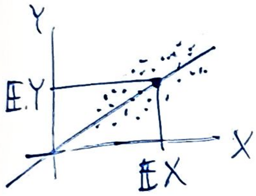
**Dk**: $\mathrm{var}(X+Y) = \mathbb{E}\left[(X+Y-\mathbb{E}(X+Y))^2\right] = \mathbb{E}\left[(X+Y-\mathbb{E}X-\mathbb{E}Y)^2\right] =$

$$= \mathbb{E}\left[X^2-2X\mathbb{E}X+(\mathbb{E}X)^2\right] + \mathbb{E}\left[Y^2-2Y\mathbb{E}Y+(\mathbb{E}Y)^2\right] + \mathbb{E}\left[2(XY-X\mathbb{E}Y-Y\mathbb{E}X+\mathbb{E}X\mathbb{E}Y)\right]$$

$$= \mathbb{E}\left[(X-\mathbb{E}X)^2\right] + \mathbb{E}\left[(Y-\mathbb{E}Y)^2\right] + 2\cdot\mathbb{E}\left[(X-\mathbb{E}X)(Y-\mathbb{E}Y)\right] \quad \blacksquare$$

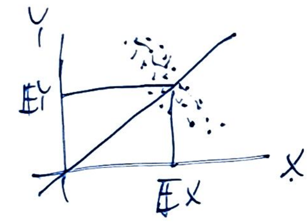
**Důsledek**: $X \perp Y \Rightarrow \mathrm{var}(X+Y) = \mathrm{var}(X) + \mathrm{var}(Y)$.

**Intuice**: $\mathrm{cov}(X,Y) = \mathbb{E}\left[(X-\mathbb{E}X)(Y-\mathbb{E}Y)\right]$ ... označíme $r(x,y) := (x-\mathbb{E}X)(y-\mathbb{E}Y)$




$x > \mathbb{E}X \Rightarrow$ nejspíš $y > \mathbb{E}Y \Rightarrow r > 0$
$x < \mathbb{E}X \Rightarrow$ nejspíš $y < \mathbb{E}Y \Rightarrow r > 0$ $\Big\}$ $\mathrm{cov}(X,Y) > 0$

$\Rightarrow$ velké $x \Rightarrow$ velké $y$ ... zvyšuje rozptyl součtu




$x > \mathbb{E}X \Rightarrow$ nejspíš $y < \mathbb{E}Y \Rightarrow r < 0$
$x < \mathbb{E}X \Rightarrow$ nejspíš $y > \mathbb{E}Y \Rightarrow r < 0$ $\Big\}$ $\mathrm{cov}(X,Y) < 0$

$\Rightarrow$ velké $x \Rightarrow$ malé $y$ ... snižuje rozptyl součtu

• **Korelace vs Kauzalita**

**Def**: **Korelace** n.v. $X, Y$ je $\rho(X,Y) := \dfrac{\mathrm{cov}(X,Y)}{\sigma_X \cdot \sigma_Y}$.

$-1 \leq \rho(X,Y) \leq 1$

**Dk**: Použijeme Cauchyho nerovnost, která říká, že $|\mathbb{E}XY| \leq \sqrt{\mathbb{E}[X^2]\cdot\mathbb{E}[Y^2]}$, kde dosadíme $X' := X-\mathbb{E}X$ a $Y' := Y-\mathbb{E}Y$, z čehož $|\mathrm{cov}(X,Y)| \leq \sigma_X \sigma_Y$ $\blacksquare$

**!** Korelace $\not\Rightarrow$ Kauzalitu ... zrádné je, že $\rho(X,Y) = \rho(Y,X)$

↳ navíc korelace je často způsobena nějakou společnou příčinou

## Spojité náhodné veličiny

**Def:** Náhodná veličina (obecná) na p.p. $(\Omega, \mathcal{F}, P)$ je funkce $X: \Omega \to \mathbb{R}$ t.ž. $\forall x \in \mathbb{R}: \{X \leq x\} \in \mathcal{F}$.

**Def:** Pro n.v. $X$ definujeme distribuční funkci (cumulative dist. f. CDF) jako $F_X: \mathbb{R} \to [0,1]$, $x \to P[X \leq x]$

**Věta:** Nechť $X$ je n.v. Pak

① $F_X$ je neklesající

② $\lim_{x \to \infty} F_X(x) = 1$

③ $\lim_{x \to -\infty} F_X(x) = 0$

④ $F_X$ je zprava spojitá

**Dk**

① $a < b \Rightarrow \{X \leq a\} \subseteq \{X \leq b\} \Rightarrow P[X \leq a] \leq P[X \leq b]$

② $A_n := \{X \leq n\}$ ... $A_1 \subseteq A_2 \subseteq \dots$

$\hookrightarrow \lim_{n \to \infty} P(A_n) = P\left(\bigcup_{n \in \mathbb{N}} A_n\right) = P(\Omega) = 1$

③ $B_n := \{X \leq -n\}$ ... $B_1 \supseteq B_2 \supseteq \dots$

$\hookrightarrow \lim_{n \to \infty} P(B_n) = P\left(\bigcap_{n \in \mathbb{N}} B_n\right) = P(\emptyset) = 0$

④ Nechť $x \in \mathbb{R}$. Chceme $\lim_{\varepsilon \to 0^+} F_X(x+\varepsilon) = F_X(x)$

$C_n := \{X \leq x + \frac{1}{n}\} \Rightarrow \lim_{\varepsilon \to 0^+} F_X(x+\varepsilon) = \lim_{n \to \infty} P(C_n) = P(\bigcap C_n) = F_X(x)$ ∎

**Def:** N.v. $X$ je spojitá $\equiv \exists f_X: \mathbb{R} \to \mathbb{R}$ t.ž.

$$F_X(x) = \int_{-\infty}^{x} f_X(s)\, ds \quad \dots \quad f_X \text{ je hustota } X$$

pdf

$\hookrightarrow$ probability density function

👁 $\frac{d}{dx} F_X(x) = f_X(x) \Rightarrow F_X$ je spojitá.

👁 $f_X \geq 0$ ... $\because F_X$ je neklesající

👁 $\int_{-\infty}^{\infty} f(s)\, ds = 1$ $\because \lim_{n \to \infty} F_X(n) = 1$

Intuice:

$$F_X(x) \approx \frac{\#\,\text{čísel} \le x}{\#\,\text{pokusů}} \quad \dots \text{ empiric CDF = ECDF}$$

$\Rightarrow F_X$ je něco jako limita ECDF

$\rightarrow f_X$ je něco jako limita histogramu těch pokusů

Věta: Pro s.n.v. $X$ platí

① $\forall x \in \mathbb{R}: P(X=x) = 0$

② $P(a \le X \le b) = \int_a^b f_X(t)\,dt$

Dk: ① $P(X=x) \le P(x-\frac{1}{n} < X \le x+\frac{1}{n}) = \int_{x-\frac{1}{n}}^{x+\frac{1}{n}} f(t)\,dt \xrightarrow[\text{Lagrange}]{n\to\infty} \frac{2}{n} f(x) \longrightarrow 0$

↳ to je vidět pro omezené $f$, pro neomezené detaily vynecháme

② $\int_a^b f_X(t)\,dt = F_X(b) - F_X(a) = P[a < X \le b] = P[a \le X \le b] - \underbrace{\cancel{P[X=a]}}_{\to 0}$ $\blacksquare$

Def: Střední hodnota s.n.v. $X$ je $\mathbb{E}X := \int_{-\infty}^{\infty} x\, f_X(x)\,dx$ (pokud to je dobře def.)

Věta (PNS): $\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x)\, f_X(x)\,dx$

Def: $\mathrm{Var}(X) := \mathbb{E}[(X-\mu)^2]$, $\mu := \mathbb{E}X$

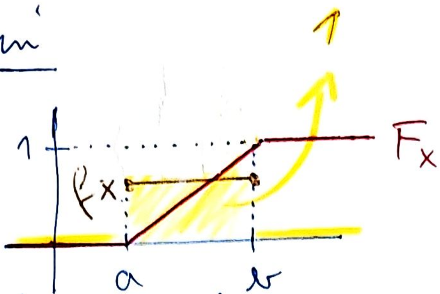
↳ výpočet: $\int_{-\infty}^{\infty} (x-\mu)^2 f_X(x)\,dx$

Tvrzení: $\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}X)^2$ ... díky linearitě, kterou zatím nemáme

# Vzorky spoj. rozdělení

## ① Uniformní

$X \sim \mathrm{Unif}(a,b)$




$$F_X(x) = \begin{cases} 0 & , x < a \\ \frac{x-a}{b-a} & , x \in [a,b] \\ 1 & , x > b \end{cases}$$

$$f_X(x) = \begin{cases} \frac{1}{b-a} & , x \in [a,b] \\ 0 & , \text{jinak} \end{cases}$$

$$\mathbb{E}X = \int_{-\infty}^{\infty} x f(x)\,dx = \int_{-\infty}^{a} + \int_{a}^{b} + \int_{b}^{\infty} = \int_a^b \frac{x}{b-a}\,dx = \frac{1}{2}\left[\frac{x^2}{b-a}\right]_a^b = \underline{\underline{\frac{a+b}{2}}}$$

$$\mathrm{Var}X = \mathbb{E}[X^2] - (\mathbb{E}X)^2 = \int_{-\infty}^{\infty} x^2 f(x)\,dx - \left(\frac{a+b}{2}\right)^2 = \dots = \underline{\underline{\frac{(b-a)^2}{12}}}$$

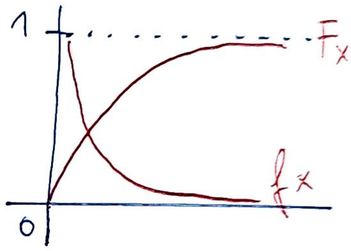
## ② Exponenciální ... obdoba geometrického

$X \sim \mathrm{Exp}(\lambda)$ ... $\lambda > 0$

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x} & , x > 0 \\ 0 & , x \le 0 \end{cases} \Rightarrow f_X(x) = F' = \lambda e^{-\lambda x}$$

$$\mathbb{E}X = \int_0^{\infty} x \lambda e^{-\lambda x}\,dx = \dots = \frac{1}{\lambda}$$

$$\mathrm{Var}X = \int_0^{\infty} x^2 \lambda e^{-\lambda x}\,dx - \frac{1}{\lambda^2} = \dots = \frac{1}{\lambda^2}$$



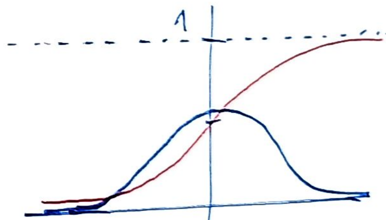

Nemá paměť: $P[X > s+t \mid X \geq t] = P[X > s]$

## ③ Standardní normální

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

distribuce: $\Phi(x) = \int_{-\infty}^{x} \varphi(s)\,ds$

$X \sim N(0,1)$ → $\mathbb{E}X = 0$, $\mathrm{Var}X = 1$




↳ k $\varphi(x)$ se dá dojít aproximací Bin pomocí Stirlinga

## ④ Obecné normální

$Z \sim N(0,1)$

$X \sim N(\mu, \sigma^2) \equiv X = \mu + \sigma \cdot Z$ ($\mu$ → shift = střed, $\sigma$ → roztažení)

$\mathbb{E}X = \mu$

$\mathrm{Var}X = \sigma^2$

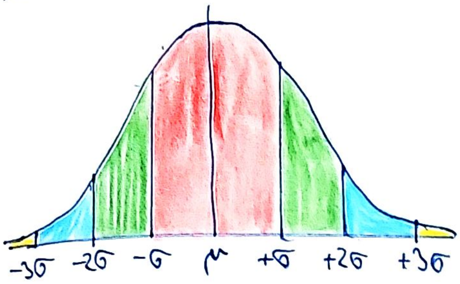
$$f_X(x) = \frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right)$$

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\sigma}\varphi\left(\frac{s-\mu}{\sigma}\right) ds = \int_{-\infty}^{\frac{x-\mu}{\sigma}} \varphi(u)\,du = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

**Věta (odolnost vůči součtu):** $X_1, \dots X_n$ jsou nez. n. v. $X_i \sim N(\mu_i, \sigma_i^2)$

$$\Rightarrow X_1 + \dots + X_n \sim N(\mu, \sigma^2), \quad \mu = \mu_1 + \dots + \mu_n, \quad \sigma^2 = \sigma_1^2 + \dots + \sigma_n^2$$

## Pravidlo $3\sigma$ 68–95–99.7 rule




$$P(\mu-\sigma < X < \mu+\sigma) = \Phi(1) - \Phi(-1) \approx 0.68$$

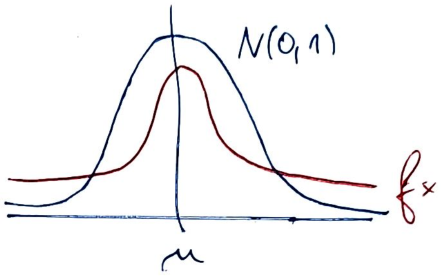
$$P(\mu-2\sigma < X < \mu+2\sigma) = \Phi(2) - \Phi(-2) \approx 0.95$$

$$P(\mu-3\sigma < X < \mu+3\sigma) = \Phi(3) - \Phi(-3) \approx 0.997$$

## (5) Cauchyho rozdělení

$$f_X(x) = \frac{1}{\pi(1+x^2)}$$

$$F_X(x) = \int_{-\infty}^{x} f_X = \frac{1}{\pi}\operatorname{arctg} x + \frac{1}{2}$$



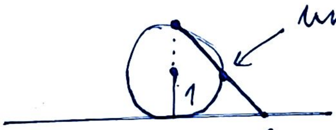

→ vypadá podobně jako $N(0,1)$

**!** střední hodnota neexistuje

$$\mathbb{E}X = \int_{-\infty}^{\infty} x f_X(x)\,dx = \int_{-\infty}^{0} x f_X + \int_{0}^{\infty} x f_X = \dots = \infty - \infty$$

$$\hookrightarrow \int_0^\infty x f_X = \frac{1}{\pi}\int_0^\infty \frac{x}{1+x^2}\,dx = \left| \begin{matrix} u = 1+x^2 \\ du = 2x\,dx \end{matrix} \right| = \frac{1}{2\pi}\int_1^\infty \frac{du}{u} = \frac{1}{2\pi} \ln|u| \Big|_1^\infty = \infty$$

→ k čemu je dobré?



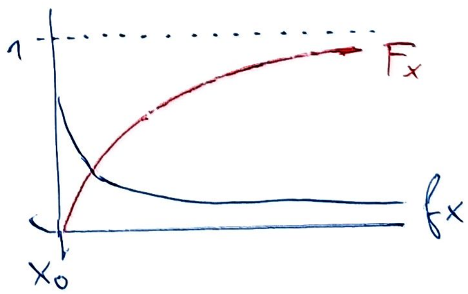

unif. náhodný bod na kružnici

Cauchyho rozdělení

## (6) Paretovo rozdělení

$$F_X(x) = \begin{cases} 1 - \left(\frac{x_0}{x}\right)^\alpha, & x \geq x_0 \\ 0, & x < x_0 \end{cases}$$

$\alpha, x_0 > 0$




$$f_X(x) = \begin{cases} \frac{\alpha}{x} \cdot \left(\frac{x_0}{x}\right)^\alpha, & x \geq x_0 \\ 0, & x < x_0 \end{cases}$$

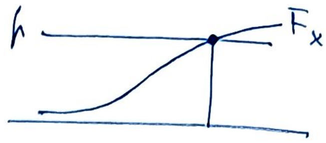
☼ $F_X$ roste k 1 mnohem pomaleji než Exp ⇒ long tail = dlouhý chvost

$$\mathbb{E}X = x_0 \frac{\alpha}{\alpha-1}$$

• Kvantilová fce

Def: Pro n.v. X definujeme kvantilovou fci $Q_X:[0,1]\to\mathbb{R}$ jako

$$Q_X(p) := \min\{x\in\mathbb{R} \mid p \le F_X(x)\}$$




Pro spojitou X je $Q_X(p)=F_X^{-1}(p)$

Pro $\forall x\in\mathbb{R}$: $Q_X(p)\le x \iff p\le F_X(x)=P[X\le x]$

Def: Medián je hodnota $m:=Q_X(\frac{1}{2})$

Pro spojitou n.v. to je nejmenší takové číslo, že

$$P[X\le m]\ge \frac{1}{2} \text{ a } P[X>m]\le\frac{1}{2}$$

→ pro d.n.v. může být takových hodnot víc

Def: První kvartil je hodnota $q:=Q_X(\frac{1}{4})$ ... 1. čtvrtina hodnot $\le q$

Desátý percentil je $Q_X(10/100)$ ... hodnota > než 10 % hodnot ...

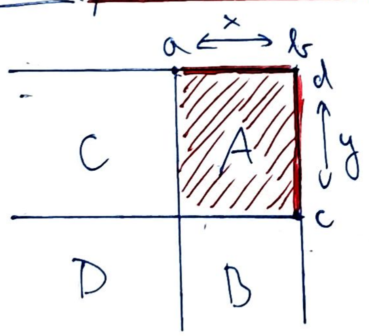
• Náhodné vektory

→ joint cdf

Def: Pro n.v. X, Y na p.p. $(\Omega,\mathcal{F},P)$ definujeme sdruženou distribuční fci

$$F_{X,Y}:\mathbb{R}^2\to[0,1], \quad F(x,y)=P[X\le x \;\&\; Y\le y]$$

Věta (pst obdélníku):




$$P[(X,Y)\in A]=P[X\in(a,b] \;\&\; Y\in(c,d]]$$

$$=F(b,d)-F(a,d)-F(b,c)+F(a,c)$$

$F:=F_{X,Y}$

Přes hustotu:

$$P[(X,Y)\in S]=\int_S f_{X,Y}(x,y)\,dx\,dy$$

$$f_{X,Y}(x,y)=\frac{\partial^2 F_{X,Y}}{\partial x\,\partial y}(x,y)$$

Marginální hustota

$$f_X(x)=\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy$$

$$f_Y(y)=\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx$$

**Věta (PNS):** Pro střední hodnotu funkce dvou n.v. platí

$$\mathbb{E}[g(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)\, f_{X,Y}(x,y)\, dx\, dy$$

**Důsledek (linearita $\mathbb{E}$)** $\quad \mathbb{E}[aX + bY + c] = a\mathbb{E}X + b\mathbb{E}Y + c$

- **Nezávislost**

**Def:** Náhodné vel. $X, Y$ jsou nezávislé, $X \perp Y \equiv$

$$\forall x, y \in \mathbb{R} : \{X \le x\} \perp \{Y \le y\}$$

$$X \perp Y \iff F_{X,Y}(x,y) = F_X(x) \cdot F_Y(y).$$

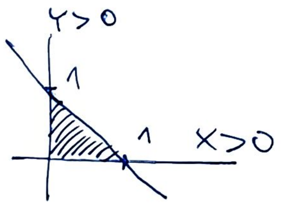
**Věta:** $X \perp Y \iff f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y).$

**Věta (konvoluce):** Nechť $X, Y$ jsou nezávislé spojité n.v. Pak $Z := X + Y$ je také s.n.v. a platí

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x)\, f_Y(z-x)\, dx.$$

**Příklady:**

① $f_{X,Y}(x,y) = \begin{cases} e^{-x-y}, & x, y > 0 \\ 0, & \text{jinak} \end{cases} \quad \to P(X+Y \le 1) = ?$




$$P(\underbrace{X+Y \le 1}_{A}) = \int_A f_{X,Y}(x,y)\, dx\, dy = \int_0^1 \int_0^{1-x} e^{-x-y}\, dy\, dx = \dots = \frac{e-2}{e}$$

② $X, Y \sim N(0,1) \Rightarrow Z := X + Y \sim ?$

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x)\, f_Y(z-x)\, dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-x)^2}{2}} dx = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-x^2 - \frac{z^2}{2} + xz} dx$$

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-(x - \frac{z}{2})^2} e^{-\frac{z^2}{4}} dx = \frac{e^{-z^2/4}}{2\pi} \int_{-\infty}^{\infty} e^{-(x-\frac{z}{2})^2} dx =$$

$$= \frac{e^{-z^2/4}}{2\pi} \int_{-\infty}^{\infty} e^{-u^2} du = \frac{1}{2\sqrt{\pi}} e^{-\frac{z^2}{4}} \Rightarrow \sigma = 2$$

$$\Rightarrow Z \sim N(0, 2)$$

$$X \sim N(0, \sigma^2) \dots f_X = \frac{1}{\sigma} \varphi\left(\frac{x}{\sigma}\right) = \frac{1}{\sigma} \cdot \frac{1}{2\pi} \cdot e^{-\frac{1}{2} \cdot \left(\frac{x}{\sigma}\right)^2}$$

- Vícerozměrné std. norm. rozdělení

$Z_1, \dots Z_n \to Z_i \sim N(0,1) \Rightarrow f_{Z_i}(t) = \varphi(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$

↳ nezávislé $\Rightarrow f_{X,Y} = f_X \cdot f_Y$

↳ $Z := (Z_1, \dots Z_n)$

$$f_Z(t_1, \dots t_n) = \prod_{i=1}^{n} f_Z(t_i) = (\sqrt{2\pi})^{-n} e^{-\frac{t_1^2 + \dots + t_n^2}{2}} = \underline{\underline{(\sqrt{2\pi})^{-n} \cdot e^{-\frac{\|\vec{t}\|}{2}}}}$$

$\Rightarrow f_Z$ záleží pouze na $\|\vec{t}\|$, tedy vzdálenosti od počátku

↳ je to radiální symetrická funkce

→ pokud $n=2$ (2 dimenze), tak $f_Z$ je Bell-curve

$\Rightarrow$ díky této symetrii dává $W := \frac{Z}{\|Z\|}$

unif. náhodný bod na jednotkové sféře v $\mathbb{R}^n$

- Podmiňování

Def: $X$ je n.v. na $(\Omega, \mathcal{F}, P)$, $B \in \mathcal{F}$ a $P(B) > 0$. Definujeme

$$F_{X|B}(x) := P[X \leq x \mid B]$$

→ k tomu přísluší $f_{X|B}$.

Věta (o rozkladu hustoty): $X$ je s.n.v. a $B_1, B_2, \dots$ je rozklad $\Omega$. Platí

$$F_X(x) = \sum_i P(B_i) \cdot F_{X|B_i}(x)$$

$$f_X(x) = \sum_i P(B_i) \cdot f_{X|B_i}(x)$$

Dk: $F_X$ ... věta o úplné psti pro $A = \{X \leq x\}$

$f_X$ ... zderivuju $F_X$

Def: $\mathbb{E}[X|B] := \int_{-\infty}^{\infty} x f_{X|B}(x)\, dx$

Věta: Pokud je $B_1, B_2, \dots$ rozklad $\Omega$, tak

$$\mathbb{E}X = \sum_i P(B_i)\, \mathbb{E}[X|B_i]$$

• **Nerovnosti**

Příklad: Může být 99 % lidí starších než průměr?

↳ Ano ... 1 člověk ... 1 rok
99 lidí ... 20 let } průměr < 20

Může být 51 % lidí starších než 2× průměr? → NE!

Věta (Markovova ner.) Nechť n.v. $X$ splňuje $X \geq 0$, $a \in \mathbb{R}$. Pak

$$P[X \geq a] \leq \frac{\mathbb{E}X}{a}$$

Důsledek: $P[X \geq b \cdot \mathbb{E}X] \leq \frac{1}{b}$ ... volbou $a := b\mathbb{E}X$

Dk: $\mathbb{E}X = P[X \geq a] \cdot \mathbb{E}[X \mid X \geq a] + P(X < a) \cdot \mathbb{E}[X \mid X < a]$

$\geq P[X \geq a] \cdot \mathbb{E}[X \mid X \geq a] \geq P[X \geq a] \cdot a$ ∎

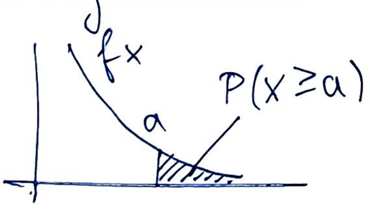
Věta (Čebyševova ner.) Nechť $X$ má $\mu := \mathbb{E}X$ a $\sigma^2 := \mathrm{Var}(X)$, $\lambda > 0$. Pak

$$P[|X - \mu| \geq \lambda \cdot \sigma] \leq \frac{1}{\lambda^2}$$

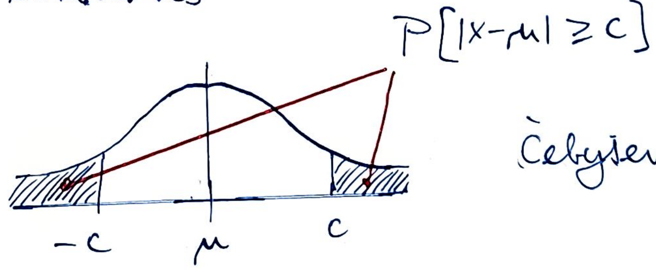
Dk: $|X-\mu| \geq \underbrace{\lambda \cdot \sigma}_{\geq 0} \iff (X-\mu)^2 \geq \lambda^2 \sigma^2 \implies Y := (X-\mu)^2 \implies \mathbb{E}Y = \sigma^2$

$\implies P[\ldots] = P[Y \geq \lambda^2 \sigma^2] = P[Y \geq \lambda^2 \mathbb{E}Y] \leq \frac{1}{\lambda^2}$ ... Markov. ∎

Odhady chvostů = tail estimates




Markov




Čebyšev

Věta (Chernoffova ner.): Nechť $X_i = \begin{cases} +1, & \text{pst} = 50\% \\ -1, & \text{pst} = 50\% \end{cases}$. $X := \sum_{i=1}^{n} X_i$, $\lambda > 0$

$$P[X \leq -\lambda \cdot \sigma_X] = P[X \geq \lambda \cdot \sigma_X] \leq e^{-\frac{\lambda^2}{2}}, \quad \sigma_X^2 = n \implies \sigma = \sqrt{n}$$

↳ Čebyšev je kvadratický odhad, tohle je exponenciální

• Zákony velkých čísel

Def: Výběrový průměr (sample mean) n.v. $X_1, \dots X_n$ je $\overline{X}_n := \frac{1}{n}\sum_i X_i$.

Věta (silný ZVČ): Nechť $X_1, X_2, \dots$ jsou stejně rozdělené nezávislé n.v. se $\mathbb{E}X_i = \mu$ a $\operatorname{Var} X_i = \sigma^2$. Pak platí

$$P\left[\lim_{n\to\infty} \overline{X}_n = \mu\right] = 1 \quad \dots \lim_{n\to\infty} \overline{X}_n = \mu \text{ skoro jistě}$$

Aplikace: Monte-Carlo integrování

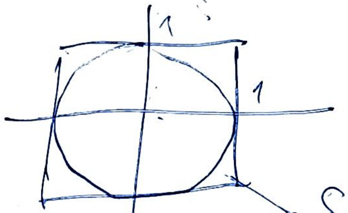
↳ chci spočítat $I := \int_{\bar{x}\in S} g(\bar{x})\,d\bar{x}$, $S \subseteq \mathbb{R}^d$

↳ objem S … $V := \operatorname{vol}(S) = \int_S d\bar{x}$

⇒ $\bar{X}_1, \dots \bar{X}_n$ jsou samply $\in S$ ⇒ $I \approx Q_n = V \cdot \frac{1}{n}\sum_i f(\bar{x}_i)$

Př: Odhad obsahu kruhu

$$g(x,y) := \begin{cases} 1, & x^2+y^2 \le 1 \\ 0, & \text{jinak} \end{cases}$$




S = čtverec

⇒ $V = 4$, $\frac{1}{n}\sum_i f(\bar{x}_i) = \frac{\#\text{bodů v kruhu}}{\#\text{bodů celkem}}$

Věta (slabý ZVČ): Nechť $X_1, X_2, \dots$ jsou stejně rozdělené nezávislé n.v. se $\mathbb{E}X_i = \mu$ a $\operatorname{Var} X_i = \sigma^2$. Pak pro $\forall \varepsilon > 0$ platí

$$\lim_{n\to\infty} P\left[|\overline{X}_n - \mu| > \varepsilon\right] = 0, \quad \text{píšeme } \overline{X}_n \xrightarrow{P} \mu.$$

⇒ Říkáme, že posloupnost $\overline{X}_n$ konverguje k $\mu$ v pravděpodobnosti.

Dk:

$X_1, \dots X_n$ stejně r. ⇒ $\mathbb{E}\overline{X}_n = \frac{1}{n}(\mathbb{E}X_1 + \dots + \mathbb{E}X_n) = \mu$

↳ pro Var: $\operatorname{var}(\overline{X}_n) = \frac{1}{n^2}\operatorname{Var}\left(\sum_i X_i\right) \overset{\text{nez.}}{=} \frac{1}{n^2}(\operatorname{var} X_1 + \dots + \operatorname{var} X_n) = \frac{\sigma^2}{n}$

⇒ chceme použít Čebyševovu n. $P[|X-\mu| > t\sigma] \le \frac{1}{t^2}$

↳ $\sigma_{\overline{X}_n} = \frac{\sigma}{\sqrt{n}}$ ⇒ $t := \frac{\varepsilon\sqrt{n}}{\sigma}$

$$\Rightarrow P\left[|\overline{X}_n - \mu| > t\frac{\sigma}{\sqrt{n}}\right] = P\left[|\overline{X}_n - \mu| > \varepsilon\right] \le \frac{\sigma^2}{n\varepsilon^2} \longrightarrow 0. \quad ∎$$

• Centrální limitní věta

Věta: Nechť $X_1, X_2, \dots$ jsou stejně rozdělené nezávislé n.v. se $\mathbb{E}X_i = \mu$ a $\operatorname{Var} X_i = \sigma^2$. Označme

$$Y_n := \frac{\overline{X}_n - \mu}{\underbrace{\sigma/\sqrt{n}}_{=\sigma(\overline{X}_n)}} = \frac{X_1 + \dots + X_n - n\mu}{\sqrt{n}\,\sigma}, \quad F_n := \text{distr. fce } Y_n$$

Potom $\forall x \in \mathbb{R}: \lim_{n\to\infty} F_n(x) = \Phi(x)$, píšeme $Y_n \xrightarrow{d} N(0,1)$ a říkáme, že posloupnost $Y_n$ konverguje k $N(0,1)$ v distribuci.

Intuice:

① $\mathbb{E}Y_n = n \cdot \frac{\mathbb{E}\overline{X}_n - \mu}{\sqrt{n}\sigma} = n \cdot \frac{0}{\dots} = 0$

$\operatorname{Var} Y_n = \left(\frac{\sqrt{n}}{\sigma}\right)^2 \cdot \operatorname{Var}\overline{X}_n = \frac{n}{\sigma^2} \cdot \frac{\sigma^2}{n} = 1$

} $Y_n$ „má šanci" být $N(0,1)$

② pokud $X_i \sim N(\mu, \sigma^2)$, tak (odolnost vůči součtu) bude $Y_n$ také normální.

↳ díky ① bude $Y_n \sim N(0,1)$

$\Rightarrow$ pokud CLV platí pro nějaké rozdělení, tak jedině $N(0,1)$.

Příklad: výtah má max. nosnost 1000 kg.

↳ nastoupí 13 osob, $\mathbb{E}X_n = 75 = \mu$, $\sigma_X = 10$
s hmotnostmi $X_1, \dots X_{13}$

$\Rightarrow P(X_1 + \dots + X_{13} \geq 1000) = ?$ $\Rightarrow S_{13} := X_1 + \dots + X_{13}$

↳ $Y_{13} := \frac{S_{13} - 13 \cdot 75}{10\sqrt{13}}$ ... $S_{13} > 1000 \iff Y_n > \frac{1000 - 13 \cdot 75}{10\sqrt{13}} \approx 0.69$

CLV: $Y_{13} \approx N(0,1) \Rightarrow P(S_{13} > 1000) \approx P(Y_n > 0.7) = 1 - P(Y_n \leq 0.7) = 1 - \Phi(0.7)$

# STATISTIKA

1) Explorační analýza ... přehledné znázornění nějakých dat

2) Konfirmační analýza ... matematická disciplína

↳ snažíme se z dat zjistit nějaké skryté skutečnosti a odhalit nepravdy způsobené náhodným kolísáním dat

→ základ: jak vybírat z dané populace objekty, co budeme měřit?

a) uniformně náhodně bez vracení

b) uniformně náhodně s vracením → lépe se analyzuje

Def: Posloupnost nez. n.v. $X_1, \dots X_n$ se stejnou distribuční funkcí $F$ je náhodný výběr s rozsahem $n$. Píšeme $X_1, \dots X_n \sim F$.

↳ naměřená data $x_1, \dots x_n$ = realizace, $x_i = X(\omega)$.

↳ na základě realizace se snažíme určit $F$ = model

① Neparametrický model ... nic o $F$ nepředpokládáme

↳ dost dobře nejde

② Parametrický model ... $F$ je z nějaké množiny $\{F_\theta \mid \theta \in \Theta\}$ → možné světy

$Exp(\lambda)$ ... $\theta = \lambda$ ; $\Theta = \mathbb{R}^+$

$U(a,b)$ ... $\theta = (a,b)$ ; $\Theta = \mathbb{R}^2$

$U(0,b)$ ... $\theta = b$ ; $\Theta = \mathbb{R}$

$N(\mu, \sigma^2)$ ... $\theta = (\mu, \sigma)$ ; $\Theta = \mathbb{R} \times \mathbb{R}^+$

Def: Statistika je libovolná fce náhodného výběru $T = T(X_1, \dots X_n)$.

↳ tedy to je n.v., ale svoji náhodnost čerpá pouze z $X_1, \dots X_n$

⇒ výběrový průměr, medián, maximum, ...

Příklad: $X_1, \dots X_n$ doby běhu programu ⇒ $p = P[X > 100\text{ ms}] = ?$

1) $\hat{p} = \frac{\# i : x_i > 100\text{ ms}}{n}$ ... $x_1, \dots x_n$ = naměřená data

2) $X \sim N(\mu, \sigma^2) \Rightarrow P(X > 100) = 1 - P(X \leq 100) = 1 - \Phi\left(\frac{100 - \mu}{\sigma}\right)$

↳ neznám $\mu, \sigma$ ⇒ odhad

$\hat{\mu} = \overline{X_n}$ $\quad \hat{\sigma}^2 = \hat{S}_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - X_n)^2$ ... výběrový rozptyl

zdůvodnění později

• **Bodové odhady**

$\hat{\theta}$ ... odhad parametru $\theta$ $\rightarrow$ estimator

↳ náhodná v. závislá na naměřených datech $\Rightarrow$ statistika

Def: Bias (vychýlení) odhadu $\theta$ je $\mathbb{E}[\hat{\theta} - \theta] = \mathbb{E}[\hat{\theta}] - \theta$

Def: Pro náhodný výběr $X_1, \dots X_n \sim F_\theta$ je bodový odhad $\hat{\theta}_n$

1) nestranný $\equiv \mathbb{E}\hat{\theta}_n = \theta$ ... bias $= 0$

2) asymptoticky nestranný $\equiv \lim_{n\to\infty} \mathbb{E}\hat{\theta}_n = \theta$ ... bias $\longrightarrow 0$

3) konzistentní $\equiv \hat{\theta}_n \xrightarrow{P} \theta$, neboli

$\forall \varepsilon > 0 : P[|\hat{\theta}_n - \theta| > \varepsilon] \to 0.$ $\Rightarrow$ malý rozptyl

odhad může být pořád biased:

↳ unbiased

ZVČ říká, že $\overline{X}_n$ je konzistentní odhad $\mathbb{E}X$

Def: Střední kvadratická odchylka $MSE(\hat{\theta}) := \mathbb{E}[(\hat{\theta} - \theta)^2]$.

Věta: $MSE(\hat{\theta}) = \text{Var}\,\hat{\theta} + \text{bias}(\hat{\theta})^2$

Dk: $\text{Var}(\hat{\theta}) = \text{Var}(\hat{\theta} - \theta) = \mathbb{E}[(\hat{\theta} - \theta)^2] - \mathbb{E}[(\hat{\theta} - \theta)]^2 = MSE(\hat{\theta}) - \text{bias}(\hat{\theta})^2$

Věta: Mějme náhodný výběr $X_1, \dots X_n$. Platí

1. $\overline{X}_n := \frac{1}{n}\sum_i X_i$ ... nestranný & konzistentní odhad $\mu$

   ↳ x.mean() $\rightarrow$ numpy

2. $S_n^2 := \frac{1}{n}\sum_i (X_i - \overline{X}_n)^2$ ... asym. nestr. & konz. odhad $\sigma^2$

   ↳ x.var(ddof=0)

3. $\hat{S}_n^2 := \frac{1}{n-1}\sum_i (X_i - \overline{X}_n)^2$ ... nestranný & konz. odhad $\sigma^2$

   výběrový rozptyl

   ↳ x.var(ddof=1)

Dk: (1) $\mathbb{E}\overline{X}_n = \mu$ z linearity, konzistence ze ZVČ

(2), (3) } konzistenci vynecháme, nestrannost $\hat{S}_n^2$ je na další straně

$\rightarrow$ proč se tam sakra dělí $n-1$? ↗

Př: Kolik je $\mathbb{E}\left[\sum_i (X_i - \overline{X}_n)^2\right]$ ?

$$\mathbb{E}[\dots] = \mathbb{E}\sum_i (X_i^2 - 2X_i\overline{X}_n + \overline{X}_n^2) = \mathbb{E}\Big[\sum_i X_i^2 - 2\overline{X}_n \underbrace{\sum_i X_i}_{n\cdot\overline{X}_n} + n\cdot\overline{X}_n^2\Big]$$

→ využiju, že $X_1, \dots X_n \sim F_\theta$ ... stejná distr.

$$= n\cdot \mathbb{E}X_i^2 - n\,\mathbb{E}\overline{X}_n^2$$

→ využiju $\operatorname{Var} X = \mathbb{E}[X^2] - (\mathbb{E}X)^2 \Rightarrow \mathbb{E}X^2 = \operatorname{Var} X + (\mathbb{E}X)^2 = \sigma^2 + \mu^2$

$$= n(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right) = (n-1)\,\sigma^2$$

nezávislost

$$\operatorname{Var} X_n = \frac{1}{n^2}\operatorname{Var}(X_1 + \dots + X_n) = \frac{1}{n^2} n\cdot\sigma^2 = \frac{\sigma^2}{n}$$

$\Rightarrow \mathbb{E}\hat{S}_n = \sigma^2 \Rightarrow \hat{S}_n$ je nestr.

- Metoda momentů

$r$-tý moment $X$ ... $m_r := \mathbb{E}[X^r]$

$r$-tý výběr. m. $X$ ... $\hat{m}_r := \frac{1}{n}\sum_i X_i^r$

$$\underset{\text{pro } r=1:}{} \quad \mathbb{E}X = m_1, \quad \overline{X}_n = \hat{m}_1$$

⇒ pro odhad $\theta$ vyřešíme rovnici $m_1(\theta) = \hat{m}_1(\theta)$

↳ případně soustavu rovnic $m_r(\theta) = \hat{m}_r(\theta)$ pro $r = 1, 2, \dots$

Př: ① $X_i \sim \mathrm{Ber}(p)$ → $\theta = p$

$$\left.\begin{aligned} m_1 &= \mathbb{E}X_i = p \\ \hat{m}_1 &= \overline{X}_n \end{aligned}\right\} \hat{p} = \overline{X}_n$$

... díky ZVČ to je nestr. konz. odhad

② $X_1, \dots X_n \sim N(\mu, \sigma^2)$ → dva parametry ⇒ 2 rovnice

$$m_1 = \mu$$
$$m_2 = \mathbb{E}X^2 = \operatorname{var}(X) + (\mathbb{E}X)^2 = \sigma^2 + \mu^2$$

$$\Rightarrow \left.\begin{aligned} \hat{m}_1 &= \overline{X}_n = \mu \\ \hat{m}_2 &= \frac{1}{n}\sum_i X_i^2 = \mu^2 + \sigma^2 = \overline{X}_n^2 + \sigma^2 \end{aligned}\right\} \begin{aligned} \hat{\mu} &= \overline{X}_n \\ \hat{\sigma}^2 &= \hat{m}_2 - \hat{\mu}^2 \end{aligned}$$

- Metoda max. věrohodnosti – max. likelihood

→ náhodný výběr $X_1, \dots X_n \sim F_\theta$ a realizace $x_1, \dots x_n$

↳ předpokládáme $\sim F_\theta$ ⇒ to dává distribuci

① diskrétní n.v.: Chceme max. $P[X_1 = x_1 \ \& \ X_2 = x_2 \ \& \dots \& \ X_n = x_n]$

↳ $X_1, \dots X_n$ nezávislé ⇒ $P[\forall i : X_i = x_i] = \prod P[X_i = x_i]$

$$\Rightarrow L(\theta, \vec{x}) := \prod_i P[X_i = x_i ; \theta]$$

↳ ve smyslu, kde $X_i \sim F_\theta$

② spojitá n.v.: $L(\theta, \vec{x}) := \prod_i f_{X_i;\theta}(x_i)$

⇒ hledám max $L(\theta, \bar{x})$, → derivuju podle $\theta$

respektive $\ell(\theta, \vec{x}) := \log L(\theta, \bar{x}) = \sum_i \log(f_{X_i}(x_i))$

Příklady

① $X_1, \dots X_n \sim Ber(p) \Rightarrow \theta = p$

$k := \#1$

$$P(\bar{X} = \bar{x}) = L(p, \bar{x}) = P[X_1 = x_1, \dots X_n = x_n] = p^k (1-p)^{n-k}$$

$$\Rightarrow \ell(p, \bar{x}) = k \cdot \ln(p) + (n-k)\ln(1-p)$$

$$\Rightarrow \frac{\partial \ell}{\partial p}(p, \bar{x}) = \frac{k}{p} + \frac{k-n}{1-p} = 0 \Rightarrow \dots \Rightarrow \hat{p} = \frac{k}{n}$$

↓ v max

② $X_1, \dots X_n \sim U(0, \theta) \Rightarrow f(x) = \begin{cases} \frac{1}{\theta}, & x \in [0, \theta] \\ 0, & \text{jinak} \end{cases}$

$L(\theta, \bar{x}) = \prod_i f(x_i)$ ... aby to nebyla 0, tak $\theta \geq \max_i x_i$

$= \prod_i \frac{1}{\theta} = \frac{1}{\theta^n}$ ... max ⇒ $\theta$ co nejmenší ⇒ $\hat{\theta} := \max_i X_i$

③ $X_1, \dots X_n \sim Geom(p) \Rightarrow P[X_i = k] = (1-p)^{k-1} \cdot p$

$$L(p, \bar{x}) = \prod_i P[X_i = x_i] = p^n (1-p)^{(\sum x_i) - n}$$

$$\Rightarrow \ell(p, \bar{x}) = n \log(p) + (\sum x_i - n) \log(1-p)$$

→ přeškáluju: $\log(p) + (\bar{X}_n - 1)\log(1-p)$

→ derivace: $\frac{1}{p} + \frac{1 - \bar{X}_n}{1-p} = 0 \Rightarrow \hat{p} = \frac{1}{\bar{X}_n}$

• <u>Intervalové odhady</u>

→ přes bodový odhad odhadnu $\theta$ jako $\hat{\theta}$ ... ale typicky $\hat{\theta} \neq \theta$

⇒ chci si být dost jistý, že $\hat{\theta}$ není moc mimo

<u>Def</u>: Nechť $D \leq H$ jsou statistiky. Říkáme, že určují $[D, H]$ <u>intervalový odhad</u> o <u>spolehlivosti</u> $(1-\alpha) \equiv P[D \leq \theta \leq H] = 1-\alpha$.

↳ umožňují chybu $\alpha \in (0,1)$

↳ anglicky $(1-\alpha)$-confidence interval ⇒ <u>$(1-\alpha)$-CI</u>

↳ někdy dává smysl uvažovat jednostranné odhady $(D = -\infty \vee H = \infty)$

<u>Značení</u>: Často budeme chtít CI typu $X \pm \delta := [X-\delta, X+\delta]$.

<u>Věta</u>: Máme <u>nestranný</u> bodový odhad $\hat{\theta}$ pro parametr $\theta$.

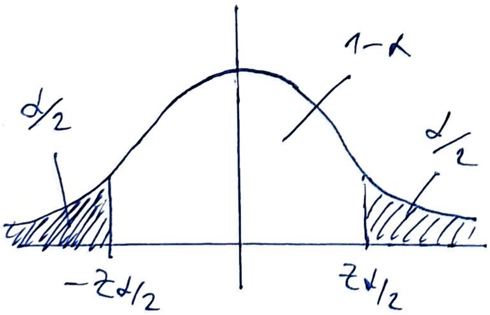
Pokud je $\hat{\theta} \sim N(\mu, \sigma^2)$, tak

$$\hat{\theta} \pm z_{\alpha/2}\, \sigma(\hat{\theta}) \text{ je } (1-\alpha)\text{-CI}, \quad z_{\alpha/2} := \Phi\left(1-\frac{\alpha}{2}\right).$$

Pokud není normálně rozdělené, tak řekneme, že skoro je (CLV)

<u>Dk</u>: Nejprve provedeme standardizaci, $E\hat{\theta} = \theta$ díky nestrannosti

$$Z = \text{stand}(\hat{\theta}) = \frac{\hat{\theta} - E\hat{\theta}}{\sigma(\hat{\theta})} = \frac{\hat{\theta} - \theta}{\sigma(\hat{\theta})} \sim N(0,1)$$




$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1-\alpha$$

$$P\left(-z_{\alpha/2} \leq \frac{\hat{\theta} - \theta}{\sigma(\hat{\theta})} \leq z_{\alpha/2}\right) = 1-\alpha$$

$$P(\underbrace{\hat{\theta} - z_{\alpha/2}\sigma}_{D} \leq \theta \leq \underbrace{\hat{\theta} + z_{\alpha/2}\sigma}_{H}) = 1-\alpha$$

∎

Příklady:

① $X_1, \dots X_n \sim N(\mu, \sigma^2) \quad \rightarrow \theta = \mu$

↳ máme nestranný odhad $\hat{\theta} = \overline{X}_n \sim N(?, ?)$

↳ $\mathbb{E}X_n = \mu$, $Var(X_n) = \frac{\sigma^2}{n} \Rightarrow \hat{\theta} \sim N(\mu, \frac{\sigma^2}{n})$

$\Rightarrow \overline{X}_n \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$ je $(1-\alpha)$-CI.

② $X_1, \dots X_n$ libovolné rozdělení

$\left.\begin{array}{l}\sigma \text{ známe} \\ n \text{ velké}\end{array}\right\}$ hledáme $\theta := \mathbb{E}X \Rightarrow$ opět odhad $\overline{X}_n$ pro $\hat{\theta}$

→ podle CLV je $\overline{X}_n$ přibližně normálně rozdělený ... $\frac{\overline{X}_n - \mu}{\sigma(\overline{X}_n)} \approx N(0,1)$

$\Rightarrow \overline{X}_n \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ je přibližně $(1-\alpha)$-CI

↳ $\lim_{n \to \infty} P(D \leq \theta \leq H) = 1-\alpha \quad \dots \quad \infty \doteq 30$

později

→ tady je pro větší rezervu lepší použít tzv. Student-t rozdělení

Pozn: Najdeme CI tvaru $X \pm \Delta$.

↳ jaké $n$ nám takový odhad zajistí pro dané $\Delta$?

$$\Delta = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \Rightarrow n \geq \left(\frac{z_{\alpha/2}\,\sigma}{\Delta}\right)^2 = \frac{const}{\Delta^2}$$

$\Rightarrow$ pokud chci $\Delta$ zmenšit 10x, a zachovat spolehlivost $\alpha$, tak potřebuji 100x víc měření ($n$)

③ Dvouvýběrový test

$\left.\begin{array}{l} X_1, \dots X_n \sim N(\mu_X, \sigma_X^2) \\ Y_1, \dots Y_m \sim N(\mu_Y, \sigma_Y^2)\end{array}\right\}$ známe $\sigma_X, \sigma_Y$, chceme odhad pro $\theta := \mu_X - \mu_Y \Rightarrow \hat{\theta} = \overline{X}_n - \overline{Y}_m$

→ máme nestranný odhad pro $\hat{\theta}$, protože $X_i, Y_i$ je norm, tak $\hat{\theta}$ taky normální

→ protože $X_i, Y_j$ jsou všechny nezávislé, tak

$$Var\,\hat{\theta} = Var(\overline{X}_n - \overline{Y}_m) = Var(\overline{X}_n) - Var(\overline{Y}_m) = \frac{\sigma_X^2}{n} - \frac{\sigma_Y^2}{m}$$

$$\Rightarrow \sigma(\hat{\theta}) = \sqrt{\frac{\sigma_X^2}{n} - \frac{\sigma_Y^2}{m}}$$

$\Rightarrow \overline{X}_n - \overline{Y}_m \pm z_{\alpha/2}\, \sigma(\hat{\theta})$ je $(1-\alpha)$-CI.

- Intervalové odhady, když neznám rozptyl

(1) Plug-in estimate

$X_1, \dots X_n \sim Ber(p) \Rightarrow \theta = p$, ale neznám $\sigma$:

$\Rightarrow \sigma(\hat{p}) = ?$ ... $Var\, X_i = p \cdot (1-p)$ $\because X_i \sim Ber(p)$

↳ $\hat{p} = \overline{X}_n \Rightarrow Var\, \overline{X}_n = \frac{1}{n} \cdot p(1-p)$

$\Rightarrow \sigma(\hat{p}) = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ ... protože neznám $p$

$\Rightarrow \overline{X}_n \pm z_{\alpha/2} \sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}$ je $(1-\alpha)$-CI

(2) Studentův test

$X_1, \dots X_n \sim N(\theta, \sigma^2)$ ... $\sigma$ neznám

↳ nestranný odhad $\hat{\theta} = \overline{X}_n$, $\sigma(\hat{\theta}) = \frac{\sigma}{\sqrt{n}}$, ale $\sigma$ neznám

$\Rightarrow$ odhad $\hat{\sigma} = \hat{S}_n = \sqrt{\frac{1}{n-1} \sum_i (X_i - \overline{X}_n)^2} \Rightarrow \sigma(\hat{\theta}) \approx \frac{\hat{S}_n}{\sqrt{n}}$

! v důkazu věty dělám normalizaci $\hat{\theta}$ na stand. normální r.

$\Rightarrow T_n := \frac{\hat{\theta} - E\hat{\theta}}{\sigma(\hat{\theta})} = \frac{\overline{X}_n - \mu}{\hat{S}_n / \sqrt{n}}$ → tohle není konst, ale n. veličina

$\Rightarrow T_n$ není $N(0,1)$, ale něco podobného

Def: Students t-distribution s $n-1$ stupni volnosti je přesně toto rozdělení.

→ distribuční funkci značíme $\Psi_{n-1}(x)$.

↳ python: scipy.stats.t.cdf(x, n-1)

$\Rightarrow$ použijeme odhad

→ analogie pro $\Phi^{-1}$

$\hat{\theta} \pm t_{\alpha/2} \cdot \frac{\hat{S}_n}{\sqrt{n}}$, kde $t_{\alpha/2} := \Psi^{-1}_{n-1}(1-\alpha/2)$

↳ analogie pro $z_{\alpha/2}$

! $t_{\alpha/2} > z_{\alpha/2} \Rightarrow$ interval je větší

↳ neznáme totiž $\sigma$

Věta: Nechť $X_1, \dots X_n \sim N(\mu, \sigma^2)$. Neznáme $\mu$ ani $\sigma$, chceme určit $\mu$. Opět máme $\alpha \in (0,1)$. Nechť $\Psi_{n-1}(t_{\alpha/2}) = 1 - \alpha/2$. Pak

$$\overline{X}_n \pm t_{\alpha/2} \frac{\hat{S}_n}{\sqrt{n}} \quad \text{je } (1-\alpha)\text{-CI pro } \mu.$$

• Testování hypotéz

1) otázka ... nalévají správnou míru piva?

$H_0$ ... nulová hypotéza (konzervativní model) ... $\mathbb{E}X = 500$ ml

$H_1$ ... alternativní hypotéza (významná odchylka) ... $\mathbb{E}X < 470$ ml

2) design experimentu – rozhodnout před začátkem měření

- vybereme vhodný statistický model ... $X \sim N(500, \sigma^2)$
- chyba 1. druhu: Zamítneme $H_0$, i když platí ... „Trapas"

$\Rightarrow$ zvolíme hladinu významnosti $\alpha$ ... typicky $\alpha = 0.05$

↳ test konstruujeme tak, aby $P[\text{ch. 1. d.}] = \alpha$

- chyba 2. druhu: Nezamítneme $H_0$, ale ona neplatí ... „Promarněná příležitost"

$\Rightarrow$ zvolíme parametr $\beta$ $\Rightarrow$ pak $P[\text{ch. 2. d.}] = \beta$

↳ hodnota $1-\beta$ se nazývá síla testu

- určíme statistický test $T = h(X_1, \dots X_n)$ ... $T = \overline{X}_n$
- určíme kritický obor (rejection region) – množinu $W$ podle $\alpha, \beta$

3) naměříme data $x_1, \dots x_n$ → realizace $X_1, \dots X_n$

4) vyhodnocení: pokud $t = h(x_1, \dots x_n) \in W \Rightarrow$ zamítneme $H_0$

$\Rightarrow \alpha = P[T \in W ; H_0]$ ... ; znamená: ve světě, kde platí $H_0$

$\Rightarrow \beta = P[T \notin W ; H_1]$ ... ve světě, kde platí $H_1$

5) výpočet dalších závěrů ↗ pokud $p = 0.09$ a $\alpha = 0.05 \Rightarrow$ nezamítnu
$\alpha = 0.11 \Rightarrow$ zamítnu

p-value := min. $\alpha$, pro kterou bychom $H_0$ zamítli
↳ pro menší $\alpha$ už ne

# Příklady:

## ① Jednovýběrový test

Nalili mi správnou míru piva?

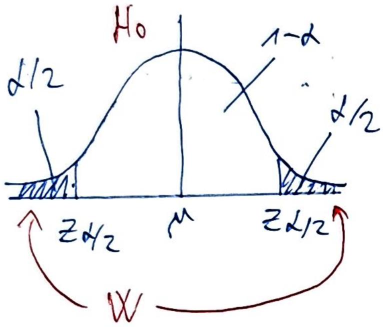
$X_1, \dots X_n \sim N(\theta, \sigma^2)$ , řekněme, že $\sigma$ známe ze zkušenosti

$H_0: \theta = 500$ ml → $\mu := 500$ ml

$H_1: \theta \neq 500$ ml → pro jednoduchost

→ chci statistiku něco jako $\overline{X}_n$, ale musím pak udělat $W$

$$\Rightarrow Z = \frac{\overline{X}_n - \mathbb{E}\overline{X}_n}{\sigma(\overline{X}_n)} = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \to \text{podle CLV: } Z \sim N(0,1)$$




→ chci $P[Z \in W ; H_0] = \alpha \Rightarrow P[Z \notin W ; H_0] = 1-\alpha$

↳ $(1-\alpha)$–CI použiju

$\Rightarrow W = \{x \in \mathbb{R} \mid |x| > z_{\alpha/2}\}$

→ $\beta$ závisí na tom, jak blízko je $\theta$ k 500 ml.

## ② Dvouvýběrový test

Nalili mi stejně jako kamarádovi?

$$\left.\begin{aligned} X_1, \dots X_n &\sim N(\theta_X, \sigma^2) \\ Y_1, \dots Y_m &\sim N(\theta_Y, \sigma^2) \end{aligned}\right\} \begin{aligned} H_0&: \theta_X = \theta_Y \\ H_1&: \theta_X \neq \theta_Y \end{aligned}$$ , předpokládáme, že známe $\sigma$

→ zase chci statistiku něco jako $\overline{X}_n - \overline{Y}_m =: S$

↳ pokud platí $H_0$, tak $\mathbb{E}S = 0$

↳ $\operatorname{Var} S = \operatorname{Var} \overline{X}_n - \operatorname{Var} \overline{Y}_m = \frac{\sigma^2}{n} + \frac{\sigma^2}{m}$

$$\Rightarrow Z := \frac{S - \mathbb{E}S}{\sigma(S)} = \frac{\overline{X}_n - \overline{Y}_m}{\sigma\sqrt{1/n + 1/m}} \to \text{podle CLV: } Z \sim N(0,1)$$

• **Z-test**: jako v ①: $W = \{x \in \mathbb{R} \mid |x| < z_{\alpha/2}\}$

## ③ Co kdybychom $\sigma$ neznali?

↳ v ① a ② jsme vyrobili Z-test. Teď ①, ale $\sigma$ neznáme

$\Rightarrow$ **T-test**: $\sigma \approx \hat{S}_n = \frac{1}{n-1}\sum_i (X_i - \overline{X}_n)^2 \Rightarrow \sigma(\overline{X}_n) \approx \frac{\hat{S}_n}{\sqrt{n}}$

$$\Rightarrow T := \frac{\overline{X}_n - \mu}{\hat{S}_n/\sqrt{n}} \quad \text{a} \quad W := \{x \in \mathbb{R} \mid |x| > t_{\alpha/2}\}, \quad t_{\alpha/2} = \Psi_{n-1}^{-1}(1-\alpha/2)$$

## • Hypotézy pro kategorická data – testy dobré shody

→ doposud numerická data: kolik nalili, # dětí, doba běhu programu

⇒ házím kostkou ... je spravedlivá? → 6 kategorií

Observed: $O_1, \dots O_k$.
Expected: $E_1, \dots E_k$

↳

| 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| 92 | 120 | 88 | 98 | 95 | 107 |
| 100 | 100 | 100 | 100 | 100 | 100 |

→ $H_0$ = kostka je spravedlivá

→ dá se odvodit nějaký rozumný vzoreček pro statistický test

G-test: $G = 2 \cdot \sum_i O_i \cdot \log \frac{O_i}{E_i}$ $\qquad \forall i\ E_i = O_i \Rightarrow G = \chi^2 = \emptyset$

$\chi^2$-test $\chi^2 = \sum_i \frac{(E_i - O_i)^2}{E_i}$ $\Big\}$ Taylor. p.

↳ měří vzdálenost od ideálního stavu $\qquad \chi^2 \doteq 6.86$

→ jak zvolit kritický obor pro dané $\alpha$?

⇒ chceme $P[T > \gamma \,;\, H_0] = \alpha$ ... pro $T = G$ nebo $T = \chi^2$

↳ jak určit $\gamma$? ... co už je moc velké?

(1) exaktní test – pro malé $n$ = # samplů $\qquad$ pro náš příklad $\gamma \doteq 11.1 \Rightarrow H_0$

→ projdu všech $(\#\text{kategorií})^n$ možných výsledků

↳ respektive $\binom{n+k-1}{n}$ ... musí se to sečíst na $n$

→ pro $\forall$ výsledek vyhodnotím $T$ a určím $\gamma$, aby $T > \gamma$ pouze v $100\alpha$ procentech případů

(2) velké $n$ ⇒ udělám (1) s nějakou množinou náhodných vzorků

(3) velké $n$ & $T = \chi^2$ → aproximace pomocí stv. $\chi^2$ rozdělení s $k-1$ stupni volnosti → poslední kategorii dopočítám

$\chi^2_{k-1}$ je rozdělení n.v. $Q = Z_1^2 + \dots + Z_{k-1}^2$, $Z_i \sim N(0,1)$

↳ dá se ukázat, že $\chi^2 \approx Q$. (pokud platí $H_0$)

⇒ potom $\gamma = F_Q^{-1}(1-\alpha)$ ⇒ $P[Q > \gamma] \doteq P[\chi^2 > \gamma] \doteq \alpha$.

## Druhy testů

- **1-výběrový test** $X_1, \dots X_n \sim N(\mu, \sigma^2) \rightarrow H_0: \mu = 5$

stats.ttest_1samp(x, popmean=5)

- **2-výběrový test**
$$\left.\begin{array}{l} X_1, \dots X_n \sim N(\mu_X, \sigma^2) \\ Y_1, \dots Y_m \sim N(\mu_Y, \sigma^2) \end{array}\right\} H_0: \mu_X = \mu_Y$$

stats.ttest_ind(x, y, equal_var=True)

↳ pokud $\sigma_X \neq \sigma_Y \Rightarrow$ equal_var=False

- **párový test**
$$\left.\begin{array}{l} X_1, \dots X_n \sim N(\mu_X, \sigma^2) \\ Y_1, \dots Y_n \sim N(\mu_Y, \sigma^2) \end{array}\right\} H_0: \mu_X = \mu_Y$$

→ $X_i, Y_i$ spolu nějak souvisí … běh algoritmu před/po optimali…

⇒ mohl bych udělat 2-výběr. t-test, ale jde to líp

⇒ udělám $D_i := X_i - Y_i$ a dělám

1-výběrový t-test $D_1, \dots D_n$ s $H_0: \overbrace{\mu_X - \mu_Y}^{\mu_D} = 0$

stats.ttest_rel(x, y)

- **Lineární regrese**

data: $(x_i, y_i)$ pro $i = 1, \dots n$

↳ cíl: $y = ax + b$ ... $x$ = nezávislá proměnná - predictor

$y$ = závislá proměnná - response

↳ chybu regrese měříme pomocí MSE

$$\sum_{i=1}^{n} (y_i - (ax_i + b))^2$$ ⇒ chceme to minimalizovat

**řešení:**

spíš výběrová cov a var

$$b = \overline{x_n} / \overline{y_n}$$

$$a = \frac{\sum_i (x_i - \overline{x_n})(y_i - \overline{y_n})}{\sum_i (x_i - \overline{x_n})^2} = \frac{\mathrm{cov}(\overline{x_n}, \overline{y_n})}{\mathrm{var}(\overline{x_n})}$$