PSA

1. Náhodný pokus a náhodný jev. Operace s náhodnými jevy
2. Pravděpodobnost a její vlastnosti. Klasická definice ppsti. Geometrická definice ppsti.
3. Axiomatická definice ppsti.
4. Nezávislost jevů. Podmíněná pravděpodobnost.
5. Věta o úplné pravděpodobnosti. Bayesova věta.
6. Náhodná veličina obecně, a speciálně DISKRÉTNÍHO a SPOJITÉHO TYPU.
7. Funkce hustoty a její vlastnosti. Pravděpodobnostní funkce a její vlastnosti.
8. Distribuční funkce a její vlastnosti. Kvantilová funkce a kvantily.
9. Charakteristiky polohy: střední hodnota, medián. Obecné a centrální momenty.
10. Charakteristiky variability: rozptyl, kvantilové rozpětí, variační koeficient. Šikmost a špičatost.
11. Binomické rozdělení. Hypergeometrické rozdělení.
12. Poissonovo rozdělení. Vztah Poissonova a exponenciálního rozdělení.
13. Rovnoměrné rozdělení. Exponenciální rozdělení.
14. Normální rozdělení. Rozdělení odvozená od normálního.
15. Vztahy mezi základními typy rozdělení.
16. Nezávislost náhodných veličin. Vícerozměrná rozdělení. Kovariance a korelace.
17. Vícerozměrné normální rozdělení.
18. Centrální limitní věta.
19. Transformace náhodných veličin.
20. Rozdělení minim a maxim náhodných veličin. Rozdělení součtu náhodných veličin.
21. Popisná statistika. Základní statistické grafy.
22. Náhodný výběr. Výběrové statistiky a jejich statistiky.
23. Bodové a intervalové odhady. Vlastnosti odhadů
24. Bodové odhady. Konstrukce odhadu metodou momentů a metodou maximální věrohodnosti.
25. Formulace statistických hypotéz. Chyba I. a II. druhu u statistických testů.
26. Statistické testování. Kritický obor. P-hodnota statistických testů.
27. Jednovýběrové parametrické testy.
28. Dvouvýběrové parametrické testy.
29. χ^2 test dobré shody.
30. Testování nezávislosti.
31. Formulace lineárního regresního modelu. Maticový zápis modelu.
32. Odhad parametrů lineárního regresního modelu. Gauss Markova věta.
33. Koeficient determinace.
34. Testování koeficientů lineárního regresního modelu.

① **NÁHODNÝ POKUS** • libovolný proces, jehož výsledek **není** jednoznačně **určen podmínkami**, za nichž je prováděn. Lze popsat množinu možných výsledků pokusu Ω

• proces, kde může nastat 2 nebo více výsledků (za stejných TP)

• hod kostkou, dotaz respondenta

NÁHODNÝ JEV = jakékoliv ověřitelné tvrzení o výsledku náhodného pokusu.

• „Nastal jev A." → A je pravdivé tvrzení o výsledku náhodného pokusu.

• $A \subset \Omega$

Všechny možné, dále nerozložitelné výsledky náhodného pokusu nazýváme **elementární jevy** - značíme ω

Obvykle volíme elementární jevy tak, aby každý jev A bylo možno napsat „pomocí" elementárních jevů.

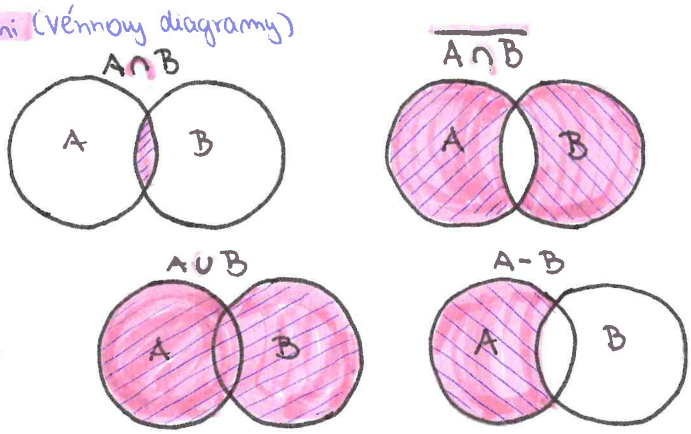
$A = \{\omega_1\} \cup \{\omega_2\} \cup \ldots$

Elementární jevy jsou disjunktní, tj. pro libovolné el. jevy platí: $\{\omega_1\} \cap \{\omega_2\} = \emptyset$.

Systém el. jevů je úplný, tj. $\Omega = \bigcup_{i=1}^{\infty} \{\omega_i\}$

OPERACE S JEVY - s jevy pracujeme jako s množinami (Vennovy diagramy)

- Ω - jev jistý
- $\emptyset$ - jev nemožný
- $\bar{A}$ (A^c, A') doplněk jevu $\Omega - A$ ⟶ neplatí A
- $A \cap B$ průnik jevů ⟶ A a zároveň B
- $A \cup B$ sjednocení jevů ⟶ A nebo B (nebo oba)
- $A \subset B$ A je podjevem B ⟶ když A, pak vždy B
- $A \cap B = \emptyset$ neslučitelné (disjunktní) jevy
- $A_1 \cup A_2 \cup \ldots \cup A_m = \Omega$ systém úplných jevů → všechny situace, které mohou nastat



VLASTNOSTI JEVŮ

1) KOMUTATIVITA $A \cup B \equiv B \cup A$ $\quad A \cap B \equiv B \cap A$

2) ASOCIATIVITA $(A \cup B) \cup C \equiv A \cup (B \cup C)$ $\quad (A \cap B) \cap C \equiv A \cap (B \cap C)$

3) DISTRIBUTIVITA $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$ $\quad (A \cup B) \cap C = (A \cap C) \cup (B \cap C)$

4) KOMPLEMENT $A \cap \bar{A} = \emptyset$ $\quad A \cup \bar{A} = \Omega$

5) de MORGANOVA PRAVIDLA $\overline{A \cup B} = \bar{A} \cap \bar{B}$ $\quad \overline{A \cap B} = \bar{A} \cup \bar{B}$

② **Pravděpodobnost jevu** • každému jevu A přiřazuji reálné číslo $P(A)$

• lze ji chápat jako předpověď poměrných četností výsledků při mnohonásobném opakování daného pokusu

• lze ji interpretovat jako kvantitativní ohodnocení stupně jistoty

Vlastnosti pravděpodobnosti Je-li $A, B \in \mathcal{A}$, pak platí:

1) $0 \leq P(A) \leq 1$ $\quad P(\emptyset) = 0$ $\quad P(\Omega) = 1$

2) ppst doplňku: $P(\bar{A}) = 1 - P(A)$

3) ppst sjednocení: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$  - průnik je 2x → musím 1x odečíst

4) pro $A \subset B$: $P(A) \leq P(B)$

Limitní vlastnosti ppsti: a) Jsou-li $A_1 \subset A_2 \subset \ldots \subset A_n \subset \ldots$ pak $P\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{n \to \infty} P(A_n)$

b) Jsou-li $A_1 \supset A_2 \supset \ldots \supset A_n \supset \ldots$ pak $P\left(\bigcap_{n=1}^{\infty} A_n\right) = \lim_{n \to \infty} P(A_n)$

• **KLASICKÁ DEFINICE PRAVDĚPODOBNOSTI**

Nechť Ω obsahuje konečný počet stejně možných elementárních jevů, pak ppst A:

$$P(A) = \frac{\text{počet příznivých výsledků}}{\text{počet všech možných výsledků}}$$

GEOMETRICKÁ DEFINICE PRAVDĚPODOBNOSTI

Nechť prostor Ω lze vyjádřit jako plochu v $\mathbb{R}^2$, pak ppst A:

$$P(A) = \frac{\text{obsah plochy } A}{\text{obsah plochy } \Omega}$$

Relativní četnost - **statistická definice ppsti**

Uvažujeme sérii n nezávislých náhodných pokusů, pak relativní četnost jevu A je

$$\frac{\text{počet pokusů, kdy výsledek je} = A}{\text{počet všech pokusů} = n}$$

③ **KOLMOGOROVOVA AXIOMATICKÁ DEFINICE PPSTI**

Trojici $(\Omega, \mathcal{A}, P)$ nazýváme pravděpodobnostní prostor, pokud platí

i) $\mathcal{A}$ je σ-algebra podmnožin množiny Ω, tj. $\mathcal{A}$ je systém podmnožin, pro které platí

- je-li $A \in \mathcal{A}$, pak také $\bar{A} \in \mathcal{A}$
- jsou-li A_i, $i = 1, 2, \ldots \in \mathcal{A}$, pak i $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$

ii) P je funkce, pro kterou platí

- $P(\Omega) = 1$
- $P(A) \geq 0$
- pro $\{A_i, i = 1, 2, \ldots\}$ takové, že $A_i \cap A_j = \emptyset$ pro $i \neq j$ platí $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$

Klasická definice ppsti je speciálním případem axiomatické definice, kdy $\Omega = \{\omega_1, \ldots, \omega_n\}$, $\mathcal{A}$ jsou všechny podmnožiny množiny Ω a $P(\{\omega_1\}) = \ldots = P(\{\omega_n\}) = 1/n$

(4) Dva nezávislé jevy, párová nezávislost

Dva jevy A a B jsou nezávislé, jestliže $P(A \cap B) = P(A) \cdot P(B)$

Pozor! rozlišovat nezávislé jevy a disjunktní jevy $\longrightarrow A \cap B = \emptyset \rightarrow P(A \cap B) = P(\emptyset) = 0$

Vlastnosti nezávislých jevů: pro dva nezávislé jevy A a B platí

i) $P(A|B) = P(A) \qquad P(B|A) = P(B)$

ii) $P(A \cup B) = P(A) + P(B) - P(A) \cdot P(B)$

iii) jevy $\bar{A}$ a $\bar{B}$ jsou nezávislé, jevy A a $\bar{B}$ jsou nezávislé, jevy $\bar{A}$ a B jsou nezávislé

Nezávislost více jevů, sdružená nezávislost

Náhodné jevy $A_1, A_2, \dots$ jsou (sdruženě) nezávislé, jestliže PRO každou k-tici náhodných jevů $A_{i_1}, \dots A_{i_k}$ platí:

$$P\left(\bigcap_{j=1}^{k} A_{i_j}\right) = \prod_{j=1}^{k} P(A_{i_j})$$

Z párové nezávislosti NEPLYNE sdružená nezávislost.

úplný systém disjunktních jevů:
Systém jevů $A_1, A_2 \dots$ nazveme úplný systém disjunktních jevů pokud platí

$$\bigcup_{i \in I} A_i = \Omega \quad a \quad A_i \cap A_j = \emptyset \quad i \neq j$$

Podmíněná pravděpodobnost

Podmíněná pravděpodobnost jevu A za podmínky jevu B (kde $P(B) > 0$) je

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

pozn. $P(A|B) \neq P(B|A)$
$P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$

(5) Věta O ÚPLNÉ PRAVDĚPODOBNOSTI

Je-li $A_1, A_2 \dots$ takový úplný systém disjunktních jevů, že platí $P(A_i) > 0$, $i = 1, 2, \dots$ pak:

$$P(B) = P(B|A_1) \cdot P(A_1) + P(B|A_2) \cdot P(A_2) + P(B|A_3) \cdot P(A_3) + \dots = \sum_i P(B|A_i) \cdot P(A_i)$$

BAYESŮV VZOREC (O INVERZNÍ PRAVDĚPODOBNOSTI)

Je-li $A_1, A_2, \dots$ takový úplný systém disjunktních jevů, že platí $P(A_i) > 0$ $i = 1, 2, \dots$ a pokud $P(B) > 0$ pak

$$P(A_j|B) = \frac{P(B|A_j) \cdot P(A_j)}{\sum_i P(B|A_i) \cdot P(A_i)} = \frac{P(B|A_j) \cdot P(A_j)}{P(B)} \quad \forall j = 1, 2, \dots$$

DŮKAZ: $P(A_1|B)P(B) = P(B|A_1)P(A_1)$
$P(A_1 \cap B) = P(B \cap A_1)$

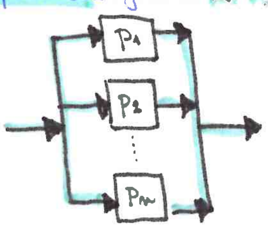
Spolehlivost PARALELNĚ a SÉRIOVĚ řazených nezávislých prvků

- Nechť $p_1, p_2, \dots, p_n$ jsou ppsti poruch prvků $P_1, P_2, \dots, P_n$. Předpokládáme, že poruchy jednotlivých prvků jsou na sobě nezávislé.

PARALELNÍ - značíme $P_1 \,p\, P_2 \,p\, P_3 \,p \dots p\, P_n$

Pravděpodobnost poruchy systému: $P = p_1 \cdot p_2 \cdot \dots \cdot p_n = \prod_i^n p_i$

Spolehlivost systému: $R = 1 - P = 1 - \prod_i p_i$



- systém funkční, pokud alespoň 1 prvek funkční
- pravděpodobnost poruchy systému je ppst toho, že dojde k poruše všech prvků

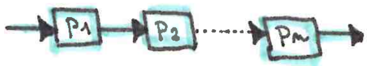
SÉRIOVÉ - značíme $P_1 \,s\, P_2 \,s \dots s\, P_n$

Pravděpodobnost poruchy systému:

$$P = 1 - (1-p_1) \cdot (1-p_2) \cdot \dots \cdot (1-p_n) = 1 - \prod_{i=1}^{n} (1-p_i) = 1 - R$$

Spolehlivost systému:

$$R = 1 - P = (1-p_1) \cdot (1-p_2) \cdot \dots \cdot (1-p_n) = \prod_{i=1}^{n} (1-p_i)$$



- systém funkční, pokud VŠECHNY prvky funkční
- ppst poruchy systému je ppst toho, že dojde k poruše ALESPOŇ 1 prvku
- ppst fungování systému je ppst toho, že fungují VŠECHNY prvky

⑥ NÁHODNÁ VELIČINA

Mějme pravděpodobnostní prostor $(\Omega, \mathcal{A}, P)$ pak reálnou měřitelnou funkci $X: \Omega \to \mathbb{R}$ nazveme NÁHODNÁ VELIČINA (měřitelná funkce $x \in \mathbb{R} \Rightarrow [X \leq x] \in \mathcal{A}$)

- výsledkům náhodného pokusu (náhodným jevům) přiřazujeme reálné číslo
- náhodnou veličinu značíme velkými latinskými písmeny z konce abecedy X, Y, Z / nebo X_1, X_2, nebo vybranými písmeny řecké abecedy ξ (ksi), η (eta), ζ (dzeta)
- realizace NV (je známa až po provedení náhodného pokusu) značíme malými písmeny x, y, z / $x_1, x_2, \dots$
- rozdělení NV = popis charakterizující pravděpodobnostní chování NV
- zápis $P(X = x)$ čteme „NV X nabývá hodnoty x"

DISKRÉTNÍ NV

Pokud existují vesměs různá čísla $x_1, x_2 \dots$ taková, že $\sum_{i=1}^{\infty} P(X = x_i) = 1$, pak náhodnou veličinu nazýváme DISKRÉTNÍ.

Funkce $P(x) \begin{cases} P(X = x_i) & \text{pro } x = x_i,\ i = 1, 2, \dots \\ 0 & \text{pro } x \neq x_i \end{cases}$ se nazývá PRAVDĚPODOBNOSTNÍ

- počítání pravděpodobností respektuje diskrétnost (uvažujeme $x_{(1)} < x_{(2)} < x_{(3)} < \dots$)
  - $P(X \leq x_i) = F(x_i) = P(X = x_1) + P(X = x_2) + \dots + P(X = x_i)$
  - $P(X < x_i) = F(x_{i-1}) = P(X = x_1) + P(X = x_2) + \dots + P(X = x_{i-1})$
  - $P(X \geq x_i) = 1 - P(X < x_i) = 1 - F(x_{i-1})$
  - $P(X > x_i) = 1 - P(X \leq x_i) = 1 - F(x_i)$

SPOJITÁ NV

Řekneme, že NV X je SPOJITÁ, pokud existuje funkce f tak, že $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\,dt$

Funkci f nazýváme FUNKCÍ HUSTOTY náhodné veličiny X.

- počítání pravděpodobností:

$$P(X = a) = 0$$

$$P(X \leq a) = P(X < a) = F(a) = \int_{-\infty}^{a} f(t)\,dt$$

$$P(X \geq a) = P(X > a) = 1 - F(a) = \int_{a}^{\infty} f(t)\,dt$$

$$P(a < X < b) = P(a \leq X \leq b) = F(b) - F(a) = \int_{a}^{b} f(t)\,dt$$

pravděpodobnost lze interpretovat jako obsah plochy pod funkcí hustoty

7 FUNKCE HUSTOTY

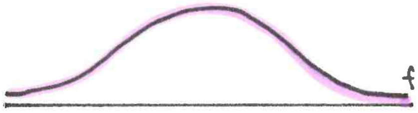
$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\,dt$$

vlastnosti distribuční funkce a funkce hustoty pro SPOJITOU NV

i) $f(x) = F'(x)$ pro skoro všechna x

ii) $f(x) \geq 0$ skoro všude a $\int_{-\infty}^{\infty} f(t)\,dt = 1$

iii) každá nezáporná (měřitelná) funkce $f(x)$ taková, že $\int_{-\infty}^{\infty} f(t)\,dt = 1$ je hustotou nějaké NV.



PRAVDĚPODOBNOSTNÍ FUNKCE

funkce $p(x) \begin{cases} P(X = x_i) & x = x_i,\ i = 1,2,\dots \\ 0 & x \neq x_i \end{cases}$

8 DISTRIBUČNÍ FUNKCE

$\forall x \in \mathbb{R}$ je distribuční fce NV X definována $F(x) = P(X \leq x)$

Vlastnosti:

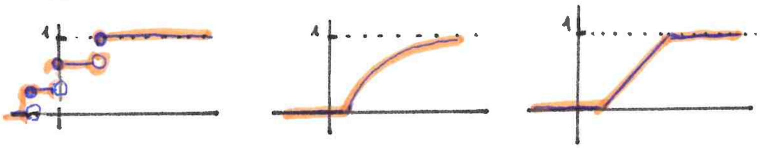
i) $0 \leq F(x) \leq 1$

ii) $\lim_{x \to -\infty} F(x) = 0$ a $\lim_{x \to +\infty} F(x) = 1$

iii) $F(x)$ je NEKLESAJÍCÍ na $\mathbb{R}$

iv) $F(x)$ je zprava spojitá

v) $F(x)$ má nejvýše spočetně bodů nespojitosti



Vlastnosti distribuční a pravděpodobnostní funkce pro DISKRÉTNÍ NV

i) distribuční fce je PO ČÁSTECH KONSTANTNÍ

ii) $F(x) = \sum_{x_i \leq x} P(X = x_i) = \sum_{x_i \leq x} P(x_i)$

iii) $P(X = x_i) = F(x_i) - \lim_{x \to x_i^-} F(x)$

KVANTILOVÁ FUNKCE A KVANTIL

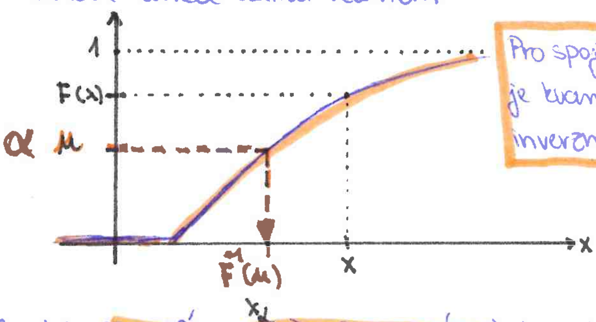
Pro náhodnou veličinu X s distribuční funkcí $F(x)$ je kvantilová funkce dána vztahem

$$F^{-1}(u) = \inf\{x \in \mathbb{R} : F(x) \geq u\}, \quad 0 < u < 1$$

Pro $0 < \alpha < 1$ se hodnota $F^{-1}(\alpha)$ nazývá α-kvantil.

α-kvantil značíme x_α:

- 0.5 kvantil – medián
- 0.25 kvantil – dolní kvartil
- 0.75 kvantil – horní kvartil



Pro spojitou NV je kvantilová funkce inverzní k $F(x)$

- kvantilová funkce je neklesající a pokud je distribuční funkce SPOJITÁ a OSTŘE ROSTOUCÍ, pak kvantilová funkce je totožná s obyčejnou inverzní funkcí k funkci F.
- Pro kvantily platí: $F(x_\alpha) = \alpha$ (speciálně pro medián $F(x_{0.5}) = 0.5$)

9) CHARAKTERISTIKY POLOHY:

• **STŘEDNÍ HODNOTA**

- pro diskrétní NV: $\mathbb{E}[X] = \sum_{i=1}^{\infty} x_i \cdot P(X=x_i)$

- pro spojitou NV: $\mathbb{E}[X] = \int_{-\infty}^{\infty} x \cdot f(x)\,dx$

pokud součet řady, resp. integrál existují

vlastnosti - pro náhodné NV X a Y, pro které existují střední hodnoty a pro libovolné $c \in \mathbb{R}$ platí

i) $\mathbb{E}c = c$

ii) $\mathbb{E}(cX) = c\mathbb{E}X$

iii) $\mathbb{E}(X+Y) = \mathbb{E}X + \mathbb{E}Y$

iv) $P(X \leq Y) = 1 \longrightarrow \mathbb{E}X \leq \mathbb{E}Y$

• **MEDIÁN** • α-kvantil, $F(x_{0,5}) = 0,5$

$x_{0,5} = F^{(-1)}(0,5) = \inf\{x \in \mathbb{R} : F(x) \geq 0,5\}$

MOMENTY NÁHODNÉ VELIČINY - pro náhodnou veličinu X a přirozené číslo n definujeme

n-tý obecný moment $\mu'_n = \mathbb{E}(X^n)$

n-tý centrální moment $\mu_n = \mathbb{E}[(X - \mathbb{E}X)^n]$

10) CHARAKTERISTIKY VARIABILITY:

• **ROZPTYL** - rozptyl NV X je definován vztahem $\mathbb{D}[X] = \mathbb{E}[(X - \mathbb{E}X)^2]$

směrodatná odchylka σ_X je odmocnina z rozptylu

vlastnosti rozptylu - pro NV X, pro kterou existuje rozptyl a pro libovolné $a \in \mathbb{R}$ platí:

i) $\mathbb{D}X \geq 0$ a navíc $\mathbb{D}X = 0 \Longleftrightarrow \exists c \in \mathbb{R} : P(X=c) = 1$

↳ důsledek $\mathbb{D}c = 0$

ii) $\boxed{\mathbb{D}X = \mathbb{E}(X^2) - (\mathbb{E}X)^2}$

iii) $\mathbb{D}(cX) = c^2\,\mathbb{D}(X)$

$\boxed{\mathbb{D}(X+Y) = \mathbb{D}X + \mathbb{D}Y \text{ - platí pro nezávislé NV}}$

- výpočtový tvar pro DISKRÉTNÍ: $\mathbb{D}X = \sum_1^{\infty} x_i^2 \cdot P(X=x_i) - (\mathbb{E}X)^2$

pro SPOJITOU: $\mathbb{D}X = \int_{-\infty}^{\infty} x^2 \cdot f(x)\,dx - (\mathbb{E}X)^2$

• **KVANTILOVÉ ROZPĚTÍ** $ICQ = x_{0,75} - x_{0,25}$

$= F^{-1}(0,75) - F^{-1}(0,25)$

• **VARIAČNÍ KOEFICIENT** $CV(X) = \dfrac{\mathrm{std}(X)}{\mathbb{E}X}$

ČEBYŠEVOVA NEROVNOST Nechť $\mathbb{D}X < \infty$ a nechť $\varepsilon > 0$, pak platí: $P(|X - \mathbb{E}X| \geq \varepsilon) \leq \dfrac{\mathbb{D}X}{\varepsilon^2}$

• **KOEFICIENT ŠIKMOSTI** $\gamma_3 = \dfrac{\mu_3}{(\sqrt{\mathbb{D}X})^3}$ ← 3. centrální moment

ŠPIČATOSTI $\gamma_4 = \dfrac{\mu_4}{(\mathbb{D}X)^2} - 3$ ← 4. centrální moment

(11) BINOMICKÉ ROZDĚLENÍ: $X \sim Bi(n,p)$

X – popisuje # úspěchů při n opakování

n – počet opakování

p – ppst úspěchu v jednom (každém) opakování

parametry $p \in (0,1)$ $n \in \mathbb{N}$, nosič $X \in \{1,2,0,3,4,5,\dots,n\}$

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

střední hodnota: $EX = n p$, rozptyl: $DX = n p (1-p)$

poznámky:
- $A_i \sim A(p)$, $i = 1,2,\dots,n$ nezávislé veličiny $\Longrightarrow X = \sum_1^n A_i \sim Bi(n,p)$
- $Bi(n, p = 1/2)$ - symetrické, $Bi(n, p \neq 1/2)$ - nesymetrické
- ppst. úspěchu v 1 pokusu je p, NV $X \sim Bi(n,p)$ charakterizuje # úspěšných pokusů při n nezávislých opakováních
- podíl výrobků s danou vlastností je v základním souboru p, NV $X \sim Bi(n,p)$ charakterizuje # výrobků s danou vlastností ve výběru rozsahu n, pokud prvky **VRACÍME ZPĚT**

➡ ALTERNATIVNÍ (BERNOULLIHO): $X \sim A(p)$, $p \in (0,1) \to X \in \{0,1\}$

ppstní funkce: $P(X=j) = p^j \cdot (1-p)^{1-j}$, $EX = p$, $DX = p(1-p)$

poznámky: • rozdělení alternativního jevu, platí X nepiaoi

— speciální případ BINOMICKÉHO, kde $n = 1$

HYPERGEOMETRICKÉ ROZDĚLENÍ $X \sim HG(N, K, n)$

X – počet prvků se sledovanou vlastností ve výběru

N – rozsah souboru

K – počet prvků v základním souboru se sledovanou vlastností

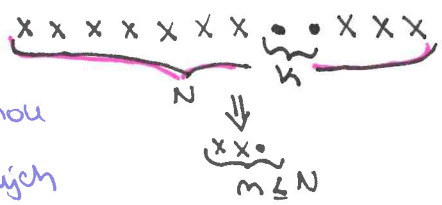
n – rozsah výběru

parametry: $0 < K < N$; $0 < n \leq N$; $n, K, N \in \mathbb{N}$, $X \in \{0,1,2,\dots,\min(K,n)\}$

ppstní funkce: $P(X=k) = \dfrac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$, kde $k = 0,1,2,\dots,\min(n,K)$

VÝBĚR BEZ OPAKOVÁNÍ

$EX = n \cdot \frac{K}{N}$, $DX = n \frac{K}{N}\left(1 - \frac{K}{N}\right)\frac{N-n}{N-1}$



poznámky: • pokud značíme $p = K/N$ procento prvků se sledovanou vlastností v souboru a dále uvažujeme, že vybíraných prvků je „tak moc", že nezáleží na tom, zda po výběru prvek vracíme či ne ➡ $HG(N,K,n) \approx Bi(n, p = \frac{K}{N})$ stačí $n \geq 30$, $p \leq 0{,}1$

(12) POISSONOVO ROZDĚLENÍ

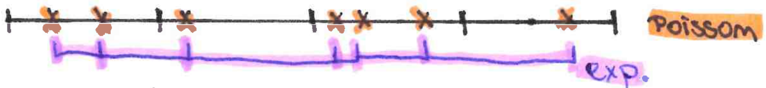
$X \sim Po(\lambda)$, $\lambda > 0$, $X \in \{0, 1, 2, \dots\}$ - nekonečné spočetné

ppstní funkce: $P(X = k) = e^{-\lambda} \cdot \frac{\lambda^k}{k!}$, kde $k = 0, 1, 2, \dots$

$EX = \lambda = DX$

poznámky:
- uvažujme náhodně se vyskytující událost, přesněji ppst výskytu události během časového intervalu je PŘÍMO ÚMĚRNÁ DÉLCE časového intervalu, průměrný počet výskytu událostí za konstantní jednotku času λ, pak NV $X \sim Po(\lambda)$ charakterizuje počet výskytu událostí za konstantní časovou jednotku
- rozdělení počtu úspěchů při velkém počtu nezávislých experimentů s malou pravděpodobností úspěchu
- pro $n \to \infty$ a $p \to 0$ je $Bi(n, p) \approx Po(\lambda = n \cdot p)$ - stačí $n \geq 30$, $p \leq 0{,}1$

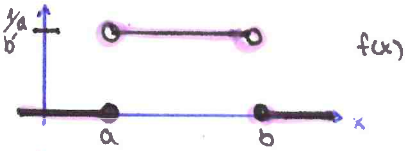
vztah Poissonova a exponenciálního rozdělení:
jestliže se počet výskytů událostí během nějakého časového intervalu řídí Poiss. rozdělením s parametrem λ, pak doba mezi výskytem 2 událostí se řídí EXP. ROZDĚLENÍM s $\delta = \frac{1}{\lambda}$



(13) ROVNOMĚRNÉ ROZDĚLENÍ

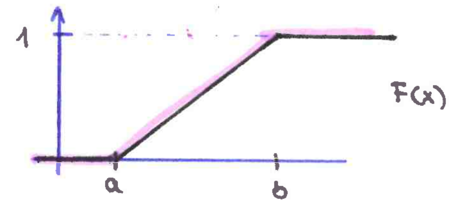
$X \sim Ro(a, b)$; $a, b \in \mathbb{R}$ $a < b$, $X \in (a, b)$

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



- střední hodnota: $EX = \frac{a+b}{2}$
- rozptyl: $DX = \frac{(a-b)^2}{12}$

distribuční funkce:
$$F(x) = \begin{cases} 0 & x \leq a \\ 1 & x \geq b \\ \frac{x-a}{b-a} & x \in (a, b) \end{cases}$$



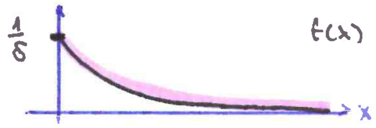
Nechť $X \sim Ro(0, 1)$ a F spojitá distribuční fce. Pak NV $Y = F^{-1}(X)$ má rozdělení s distr. fcí F.

poznámky: NV X může být libovolné reálné hodnoty z intervalu (a, b) a její výskyt na celém intervalu (a, b) je stejně možný

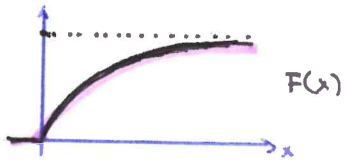
EXPONENCIÁLNÍ ROZDĚLENÍ

$X \sim Exp(\delta)$, $\delta > 0$, $X \in \langle 0, +\infty)$

funkce hustoty
$$f(x) = \begin{cases} \frac{1}{\delta} \cdot e^{-x/\delta} & x \geq 0 \\ 0 & x < 0 \end{cases}$$



distribuční funkce
$$F(x) = \begin{cases} 0 & x < 0 \\ 1 - e^{-x/\delta} & x \geq 0 \end{cases}$$

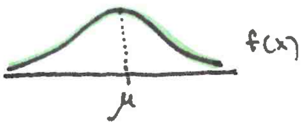


$EX = \delta$, $DX = \delta^2$

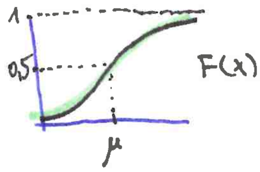
poznámky: popisuje například dobu čekání na určitou náhodnou událost nebo dobu životnosti součástek, které nepodléhají opotřebení. Parametr δ charakterizuje průměrnou dobu mezi výskytem dvou událostí.

14 NORMÁLNÍ ROZDĚLENÍ (GAUSSOVO) $X \sim N(\mu, \sigma^2)$. $\mu, \sigma^2 \in \mathbb{R}$, $\sigma^2 > 0$; $x \in \mathbb{R}$

funkce hustoty: $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{1}{2} \cdot \frac{(x-\mu)^2}{\sigma^2}}$ $x \in \mathbb{R}$



Distribuční fce - nelze vyjádřit pomocí elementárních funkcí



$\mathbb{E}X = \mu$, $\mathbb{D}X = \sigma^2$

poznámky: NORMOVANÉ NORMÁLNÍ ROZDĚLENÍ $N(\mu=0, \sigma^2=1)$,
značíme obvykle U a příslušnou funkci hustoty značíme $\varphi(u)$, distribuční funkci $\Phi(u)$

- funkce hustoty symetrická kolem μ, tedy pro $x \in \mathbb{R}$: $\boxed{F(x-\mu) = 1 - F(x+\mu)}$
- symetrie funkce hustoty a distribuční funkce pro NORMOVANÉ NR

$\boxed{\varphi(-u) = \varphi(u)}$ $\boxed{\Phi(u) = 1 - \Phi(-u)}$

věta: Nechť $X \sim N(\mu, \sigma^2)$, pak $U = \frac{X-\mu}{\sigma} \sim N(0,1)$ a platí:

i) pro hodnoty distribuční funkce $F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$

ii) pro kvantily $x_\alpha = u_\alpha \cdot \sigma + \mu$

pravidlo 3-sigma - pro NV $X \sim N(\mu, \sigma^2)$ a číslo $k \in \mathbb{N}$ platí:

$$P(\mu - k\cdot\sigma < X < \mu + k\sigma) = F(\mu + k\sigma) - F(\mu - k\sigma) = \Phi(k) - \Phi(-k) = 2\cdot\Phi(k) - 1$$

Speciálně: $P(\mu - \sigma < X < \mu + \sigma) \doteq \frac{2}{3}$

$P(\mu - 2\sigma < X < \mu + 2\sigma) \doteq 0{,}95$

$P(\mu - 3\sigma < X < \mu + 3\sigma) \doteq 0{,}997$

15 Vztahy mezi základními typy rozdělení

$$\begin{tikzcd} Hg \arrow[r] & Bi \arrow[r] \arrow[dr] & Po \arrow[d, bend left] \\ & & N \end{tikzcd}$$

$A(p) = Bi(n=1, p)$

$HG(N, K, n) \approx Bi\left(p = \frac{N}{K}, n\right)$

$Bi(n, p) \approx Po(\lambda = n\cdot p)$

$N(\mu = n\cdot p, \sigma^2 = n\cdot p\cdot(1-p))$

$Po(\lambda) \approx N(\lambda, \lambda)$

(16) NEZÁVISLOST NÁHODNÝCH VELIČIN

→ NV $X_1, X_2, \dots, X_n$ jsou nezávislé, když $\forall x_1, x_2, \dots, x_n$ jsou nezávislé jevy $[X_1 \le x_1], [X_2 \le x_2], \dots [X_n \le x_n]$

- NV X a Y jsou nezávislé právě tehdy, když $\forall x, y \in \mathbb{R}^2$ platí

$$F(x,y) = F_X(x) \cdot F_Y(y)$$ – lze rozšířit pro n nezávislých veličin

$$P(X \le x_i \wedge Y \le y_j) = P(X \le x_i) \cdot P(Y \le y_j)$$

- DISKRÉTNÍ NV X a Y jsou nezávislé právě tehdy, když

$$P(X = x_i, Y = y_j) = P(X = x_i) \cdot P(Y = y_j) \qquad \forall i, j = 1, 2, \dots$$

- SPOJITÉ NV X a Y jsou nezávislé právě tehdy, když

$$f(x,y) = f(x) \cdot f(y) \qquad \forall (x,y) \in \mathbb{R}^2$$

VÍCEROZMĚRNÁ NÁHODNÁ VELIČINA – NÁHODNÝ VEKTOR

- SDRUŽENÁ DISTRIBUČNÍ FUNKCE – pro NV $X_1, X_2, \dots X_n$ se funkce $F(x_1, x_2, \dots, x_n) = P(X_1 \le x_1, X_2 \le x_2, \dots, X_n \le x_n)$ nazývá sdružená distribuční funkce
  - distribuční funkce jednotlivých veličin se nazývají **marginální distribuční funkce**
- SDRUŽENÁ FUNKCE HUSTOTY
  - pro SPOJITÉ NV se funkce $f(x_1, x_2, \dots, x_n) = \dfrac{\partial^n F(x_1, x_2, \dots, x_n)}{\partial x_1 \partial x_2 \partial x_3 \dots \partial x_n}$ nazývá sdružená funkce hustoty
  - dvourozměrná DISKRÉTNÍ NV – tabulka

$X \setminus Y$	y_1	y_2		y_m
x_1	$P(x_1, y_1)$	$P(x_1, y_2)$	....	$P(x_1, y_m)$
x_2	$P(x_2, y_1)$	$P(x_2, y_2)$	....	$P(x_2, y_m)$
$\vdots$				
x_m	$P(x_m, y_1)$	$P(x_m, y_2)$	...	$P(x_m, y_n)$

- vlastnosti sdružené distribuční funkce
  - $0 \le F(x_1, \dots, x_n) \le 1$
  - neklesající, zprava spojitá $\forall x_i$
  - $\lim\limits_{x_i \to -\infty} F(x_1, \dots, x_n) = 0 \quad \forall i = 1, 2, \dots, n \qquad \lim\limits_{x \to \infty} F(x) = 1$ kde $x \to \infty$ znamená $x_i \to \infty \quad \forall i = 1, 2, \dots, n$
- vlastnosti marginálních funkcí – uvažujeme dvourozměrnou NV (X,Y), pak pro marginální distribuční funkci platí:

$$F_X(x) = \lim_{y \to \infty} F(x,y) \qquad F_Y(y) = \lim_{x \to \infty} F(x,y)$$

  - pro DISKRÉTNÍ NV s pravděpodobnostní funkcí $P(X = x_i, Y = y_j)$, kde $i = 1, 2, \dots m \quad j = 1, 2, \dots n$

$$P_X(X = x_i) = \sum_j^n P(X = x_i, Y = y_j) \qquad P_Y(Y = y_j) = \sum_i^m P(X = x_i, Y = y_j)$$

  - pro SPOJITOU NV s funkcí hustoty $f(x,y)$

$$f_X(x) = \int_{\mathbb{R}} f(x,y)\,dy \qquad f_Y(y) = \int_{\mathbb{R}} f(x,y)\,dx$$

CHARAKTERISTIKY VÍCEROZMĚRNÉ NV

- střední hodnota náhodného vektoru $\mathbf{X}$ je vektor středních hodnot marginálních rozdělení $\mathbb{E}\mathbf{X} = (\mathbb{E}X_1, \mathbb{E}X_2, \dots)$
- rozptyl náhodného vektoru $\mathbf{X}$ je vektor $\mathbb{D}\mathbf{X} = (\mathbb{D}X_1, \mathbb{D}X_2, \dots)$

pro nezávislé NV X a Y platí: $\mathbb{E}(X \cdot Y) = \mathbb{E}X \cdot \mathbb{E}Y$

$$\mathbb{D}(X \pm Y) = \mathbb{D}X + \mathbb{D}Y$$

Budeme předpokládat, že NV mají konečné druhé momenty → tedy i konečné první momenty:
ze Schwarzovy nerovnosti pro střední hodnoty plyne, že $|\mathbb{E}(X\cdot Y)| \leq \sqrt{\mathbb{E}X^2 \cdot \mathbb{E}Y^2}$ a zvolím-li $Y=1$, dostaneme $|\mathbb{E}X| \leq \sqrt{\mathbb{E}X^2} < \infty$.

Dále předpokládáme, že veličiny mají nenulový rozptyl.

KOVARIANCE - pro NV X a Y: $\mathrm{cov}(X,Y) = \mathbb{E}[(X-\mathbb{E}X)\cdot(Y-\mathbb{E}Y)]$ součin mezi odchylkami marginálních veličin od jejich středních hodnot

i) $\mathrm{cov}(X,X) = \mathbb{D}X$
ii) $\mathrm{cov}(X,Y) = \mathrm{cov}(Y,X)$
iii) $\boxed{\mathrm{cov}(X,Y) = \mathbb{E}(X\cdot Y) - \mathbb{E}X\cdot\mathbb{E}Y}$
iv) $|\mathrm{cov}(X,Y)| \leq \sqrt{\mathbb{D}X\cdot\mathbb{D}Y}$
v) $\mathrm{cov}(a+b\cdot X, c+d\cdot Y) = b\cdot d\cdot\mathrm{cov}(X,Y)$ $a,b,c,d\in\mathbb{R}$
vi) $\mathbb{D}(X\pm Y) = \mathbb{D}X + \mathbb{D}Y \pm 2\cdot\mathrm{cov}(X,Y)$

- pro nezávislé NV $\Longrightarrow$ $\mathrm{cov}(X,Y)=0$
- pro vícerozměrnou veličinu $\mathbf{X} = X_1, X_2, \dots, X_n$ definujeme kovarianční matici

$$\mathrm{var}\,\mathbf{X} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m1} & \sigma_{m2} & \cdots & \sigma_{mm} \end{bmatrix}$$

$\sigma_{ij} = \mathrm{cov}(X_i, X_j)$ $i=j:$ $\sigma_{ij} = \mathbb{D}X$

KORELACE - pro NV X a Y ($\mathbb{D}X>0$, $\mathbb{D}Y>0$): $\mathrm{cor}(X,Y) = \dfrac{\mathrm{cov}(X,Y)}{\sqrt{\mathbb{D}X}\cdot\sqrt{\mathbb{D}Y}}$

- normovaná kovariance
- X_1, X_2 nezávislé $\Longrightarrow$ $\mathrm{cor}(X_1,X_2)=0$ - nekorelované veličiny

i) $\mathrm{cor}(X,X)=1$
ii) $\mathrm{cor}(X,Y)=\mathrm{cor}(Y,X)$
iii) $-1 \leq \mathrm{cor}(X,Y) \leq 1$
iv) $\mathrm{cor}(a+b\cdot X, c+d\cdot Y) = \mathrm{sgn}(b\cdot d)\cdot\mathrm{cor}(X,Y)$ $a,b,c,d\in\mathbb{R}$
v) $\mathrm{cor}(X,Y) = \pm 1 \iff Y = a \pm b\cdot X$, $b>0$

- korelační matice vektoru $\mathbf{X}$ je

$$\mathrm{cor}\,\mathbf{X} = \begin{bmatrix} 1 & \varrho_{12} & \cdots & \varrho_{1m} \\ \varrho_{21} & 1 & \cdots & \varrho_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \varrho_{m1} & \varrho_{m2} & \cdots & 1 \end{bmatrix}$$

17 VÍCEROZMĚRNÉ NORMÁLNÍ ROZDĚLENÍ

$\mathbf{X} = (X, Y) \sim N_2(\mu, \Sigma)$, kde $\mu = (\mu_x, \mu_y)$ je vektor středních hodnot

a $\Sigma = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}$ je kovarianční matice je popsáno sdruženou funkcí hustoty

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \cdot e^{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - 2\rho\frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right]}$$

- Pro dvourozměrné normální rozdělení platí
  - marginální rozdělení jsou normální
  - X a Y jsou nezávislé právě tehdy, když jsou nekorelované

    $\mathrm{cov}(X,Y) = 0 \Leftrightarrow$ X a Y jsou nezávislé

18 Konvergence v pravděpodobnosti

- Posloupnost NV $\{X_n\}_{n=1}^{\infty}$ konvergují v ppsti k NV $\mathbf{X}$ pro $n \to \infty$ právě tehdy, když

$\forall \varepsilon > 0: \lim_{n\to\infty} P(|X_n - X| > \varepsilon) = 0$ – značíme $X_n \xrightarrow{P} X$

- **Konvergence skoro jistě** – Posloupnost NV $\{X_n\}_1^{\infty}$ konvergují skoro jistě k NV $\mathbf{X}$ pro $n \to \infty$ právě tehdy, když

  $P(\lim_{n\to\infty} |X_n - X| = 0) = 1$ – značíme $X_n \xrightarrow{sj} X$

- **Konvergence v distribuci** – Posloup. NV $\{X_n\}_1^{\infty}$ konvergují v distribuci k NV $\mathbf{X}$ pro $n \to \infty$ právě tehdy, když

  $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ v každém bodě, v němž je $F_X(x)$ spojitá – značíme $X_n \xrightarrow{D} X$

VZTAHY KONVERGENCÍ:
i) $X_n \xrightarrow{sj} X \Rightarrow X_n \xrightarrow{P} X$

ii) $X_n \xrightarrow{P} X \Rightarrow X_n \xrightarrow{D} X$

iii) pokud X konstanta μ, pak $X_n \xrightarrow{D} \mu \Rightarrow X_n \xrightarrow{P} \mu$

Zákony velkých čísel se zabývají rozdělením průměrů z NV. Označíme $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ průměr z prvních n veličin.

- **ČEBYŠEVŮV SLABÝ ZÁKON VELKÝCH ČÍSEL** – Nechť $X_1, X_2, X_3 \ldots$ je posloupnost nezávislých NV se střední hodnotou $\mathbb{E}X_i = \mu$ a omezeným rozptylem $\mathbb{D}X_i \le C$, pak platí $\bar{X}_n \xrightarrow{P} \mu$

- **CHINČINŮV SLABÝ ZÁKON VELKÝCH ČÍSEL** – Nechť $X_1, X_2, X_3 \ldots$ je posloupnost nezávislých stejně rozdělených NV se střední hodnotou $\mathbb{E}X_i = \mu$, pak platí $\bar{X}_n \xrightarrow{P} \mu$

- **KOLMOGOROVOVŮV SILNÝ ZÁKON VELKÝCH ČÍSEL** – Nechť $X_1, X_2, X_3 \ldots$ je posloup. nezávislých stejně rozdělených NV s $\mathbb{E}X_i = \mu$, pak platí

  $\bar{X}_n \xrightarrow{sj} \mu$

- **CENTRÁLNÍ LIMITNÍ VĚTA** – Nechť $\{X_n\}_1^{\infty}$ jsou vzájemně nezávislé NV, pro které platí $\mathbb{E}X_i = \mu$, $\mathbb{D}X_i = \sigma^2 > 0$ a $\mathbb{E}|X_i|^3 < \infty$. Pak pro distribuční funkci NV

  $S_n = \frac{1}{\sqrt{n}} \cdot \sum_1^n \left(\frac{X_i - \mu}{\sigma}\right)$ platí $\lim_{n\to\infty} P(S_n \le x) = \Phi(x)$ – větu lze interpretovat tak, že pro velká n přibližně platí
  - $\sum_1^n X_i \overset{a}{\sim} N(n\mu, n\sigma^2)$
  - $\bar{X}_n \overset{a}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$

  příklad aproximace normálním rozdělením: $Bi(n,p) \approx N(np, np(1-p))$ pro $np(1-p) \ge 9$

  – Nechť $\{X_n\}_1^{\infty}$ jsou vzájemně nezávislé, stejně rozdělené NV s $\mathbb{E}X_i = \mu$ a $\mathbb{D}X_i = \sigma^2$. Pak

$$\frac{1}{\sqrt{n}}\sum_1^n (X_i - \mu) = \sqrt{n}(\bar{X}_n - \mu) \xrightarrow{D} N(0, \sigma^2)$$

- **CENTR. LIM. VĚTA PRO NÁHODNÉ VEKTORY** – $\{\mathbf{X}_n\}_1^{\infty}$ jsou vzájemně nezávislé, stejně rozdělené k-dimensionální N. vektory s $\mathbb{E}\mathbf{X}_i = \mu$ a konečnou kovarianční maticí Σ. Pak platí $\frac{1}{\sqrt{n}}\sum_1^n (\mathbf{X}_i - \mu) = \sqrt{n}(\bar{\mathbf{X}}_n - \mu) \xrightarrow{D} N_k(0, \Sigma)$

Δ METODA
- 1D případ – $\{T_n\}_1^{\infty}$ splňuje $\sqrt{n}(T_n - \mu) \xrightarrow{D} N(0, \sigma^2)$ pro hodnoty μ, σ^2. Pak $\sqrt{n}(g(T_n) - g(\mu)) \xrightarrow{D} N(0, \sigma^2 \cdot (g'(\mu))^2)$
- k-dimensionální – $\{\mathbf{T}_n\}_1^{\infty}$ splňuje $\sqrt{n}(\mathbf{T}_n - \mu) \xrightarrow{D} N_k(\mathbf{0}, \Sigma)$ pro vektor konstant μ a matici Σ.

  Nechť g je spojitě diferencovatelná fce $\mathbb{R}^k \to \mathbb{R}^p$, označíme $D(x) = \frac{\partial g(x)}{\partial x}$. Pak platí

$$\sqrt{n}(g(\mathbf{T}_n) - g(\mu)) \xrightarrow{D} N_p(\mathbf{0}, D(\mu)\Sigma D(\mu)^T)$$

19 TRANSFORMACE NÁHODNÝCH VELIČIN

- pokud X je NV a $Y=h(X)$ pak platí: $F_Y(y)=P(Y\le y)=P(h(X)\le y)$
- pokud $h(x)$ rostoucí: $F_Y(y)=P(Y\le y)=P(h(X)\le y)=P(X\le h^{-1}(y))=F_X(h^{-1}(y))$

  a pro funkci hustoty platí $f_Y(y)=f_X(h^{-1}(y))\left|\frac{dh^{-1}(y)}{dy}\right|$
- pro vícerozměrné NV platí $f_Y(y_1,y_2,\dots,y_m)=f_X(h_1^{-1}(y_1,\dots,y_m),\dots,h_m^{-1}(y_1,\dots,y_m))\,|J(y_1,\dots,y_m)|$

  kde
$$J(y_1,\dots,y_m)=\begin{pmatrix}\frac{dh_1^{-1}}{dy_1} & \frac{dh_1^{-1}}{dy_2} & \cdots & \frac{dh_1^{-1}}{dy_m}\\ \vdots & & \ddots & \vdots\\ \frac{dh_m^{-1}}{dy_1} & \cdots & \cdots & \frac{dh_m^{-1}}{dy_m}\end{pmatrix}$$

- Nechť X je NV s pravděpodobnostní funkcí p, resp. funkcí hustoty f a nechť $h(X)$ je TRANSFORMACE NV X. Pak střední hodnota transformované NV:
  - DISKRÉTNÍ: $\mathbb{E}[h(X)]=\sum_i h(x_i)\,p(X=x_i)$
  - SPOJITÉ: $\mathbb{E}(h(X))=\int h(x)f(x)\,dx$

SPECIÁLNÍ PŘÍPADY:
- NV $Y=a+bX \rightarrow G(y)=\begin{cases}F\left(\frac{y-a}{b}\right) & b>0\\ 1-F\left(\frac{y-a}{b}\right) & b<0\end{cases}$
- $Y=|X| \rightarrow G(y)=\begin{cases}0 & y<0\\ F(y)-F_-(-y) & y\ge 0\end{cases}$ $\qquad F_-(x)=\lim_{y\to x_-}F(y)$
- $Y=X^2 \rightarrow G(y)=\begin{cases}0 & y<0\\ F(\sqrt{y})-F_-(-\sqrt{y}) & y\ge 0\end{cases}$

20

- Nechť X a Y jsou NV s funkcemi hustoty f_X, f_Y, pak funkce hustoty pro součet $W=X+Y$ má tvar
$$g(w)=\int_{\mathbb{R}} f_X(x)\,f_Y(w-x)\,dx$$

Rozdělení součtu nezávislých NV:
- $X_i,\ i=1,2,\dots,m \sim$ Bernoulliho $A(p)$ .... $\sum_1^m X_i \sim$ binomické $Bi(m,p)$
- $X_1\sim Bi(m_1,p),\ X_2\sim Bi(m_2,p)$ ... $X_1+X_2\sim Bi(m_1+m_2,p)$
- $X_1\sim Po(\lambda_1),\ X_2\sim Po(\lambda_2)$ ... $X_1+X_2\sim Po(\lambda_1+\lambda_2)$
- $X_1\sim N(\mu_1,\sigma_1^2),\ X_2\sim N(\mu_2,\sigma_2^2)$ ... $X_1+X_2\sim N(\mu_1+\mu_2,\sigma_1^2+\sigma_2^2)$
- $X_1, X_2 \sim$ rovnoměrné ... $X_1+X_2\sim$ trojúhelníkové
- $X_1\sim f_1,\ X_2\sim f_2$ .... $X_1+X_2\sim f(y)=\int f_1(x)\,f_2(y-x)\,dx$

Rozdělení odvozená od normálního rozdělení
- $Z_i\sim N(0,1)$ pro $i=1,2,\dots,k$ jsou NV - nezávislé, pak NV $X^2=Z_1^2+Z_2^2+\dots+Z_k^2$ má χ^2 rozdělení se stupněm volnosti k
$$X^2=Z_1^2+Z_2^2+\dots+Z_k^2\sim\chi^2(\nu=k)$$
- $Z\sim N(0,1)$ a $X^2\sim\chi^2(\nu=k)$ jsou nezávislé NV, pak NV $T=\frac{Z}{\sqrt{X^2/k}}$ má STUDENTOVO t-rozdělení se stupněm volnosti k $\quad T=\frac{Z}{\sqrt{X^2/k}}\sim t(\nu=k)$
- $X_k^2\sim\chi^2(\nu=k)$ a $X_m^2\sim\chi^2(\nu=m)$ jsou nezávislé NV, pak $F=\frac{X_k^2/k}{X_m^2/m}\sim F(\nu_1=k,\nu_2=m)$ $\quad F$ má FISHEROVO-SNEDECOROVO ROZDĚLENÍ

Nechť $X_i\sim F_i(x)$ jsou nezávislé NV s distribučními funkcemi F_i, pak pro NV $Y=\max(X_1,X_2,\dots,X_m)$ platí
$$Y\sim F_Y(y)=F_1(y)\cdot F_2(y)\cdot\ldots\cdot F_m(y)$$
a pro NV $Y=\min(X_1,X_2,\dots,X_m)$ platí
$$Y\sim F_Y(y)=1-(1-F_1(y))\cdot(1-F_2(y))\cdots(1-F_m(y))$$

(21) DESKRIPTIVNÍ (POPISNÁ) STATISTIKA

Předpokládáme, že máme k dispozici data - značíme $x_1, x_2, \dots, x_n$, v případě, že data budou uspořádána podle VELIKOSTI, zdůrazníme to značením $x_{(1)}, x_{(2)}, \dots$ a platí $x_{(1)} \leq x_{(2)} \leq \dots x_{(n)}$.
Charakteristiky datového souboru, nazývané též POPISNÉ MÍRY DAT jsou čísla spočtená z dat.

MÍRY POLOHY

- **Průměry** - vhodné pro kvantitativní typ dat
  - ARITMETICKÝ $\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$
  - GEOMETRICKÝ $\bar{x}_G = \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n}$ (procentuální změny)
  - HARMONICKÝ $\bar{x}_H = \left(\frac{1}{n}\sum_1^n \left(\frac{1}{x_i}\right)\right)^{-1}$ (průměrná rychlost)
  - smysl pro kladná čísla

  $\bar{x}_H \leq \bar{x}_G \leq \bar{x}$

- **Medián** = hodnota, když uspořádáme data podle velikosti < přesně polovina menší než medián / přesně polovina větší než medián

$$\tilde{x} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & n \text{ liché} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right) & n \text{ sudé} \end{cases}$$

- **Výběrové kvantily (percentily)** = α-procentní kvantil Q_α je taková hodnota, že u dat uspořádaných podle velikosti je α procent hodnot menší než Q_α a $(100-\alpha)$ procent hodnot je větší než Q_α kvantil.
  Q_{25} / Q_{75} = dolní/horní kvartil, Q_{50} = medián, Q_{10} / Q_{90} = dolní/horní decil
- **Modus** = nejčetnější hodnota (objevuje se v souboru nejčastěji)

MÍRY VARIABILITY

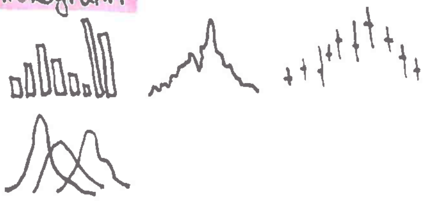
- **rozptyl** $s_p^2 = \frac{1}{n}\sum_1^n (x_i - \bar{x})^2$
- **výběrový rozptyl** $s^2 = \frac{1}{n-1}\sum_1^n (x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum_1^n x_i^2 - n\bar{x}^2\right)$
- **výběrová směrodatná odchylka** $s = \sqrt{s^2}$
- **výběrový variační koeficient** $v = \frac{s}{\bar{x}}$
- **rozpětí** $R = Q_{100} - Q_0 = x_{(n)} - x_{(1)} = \max(x) - \min(x)$
- **kvartilové / decilové rozpětí** $R_Q = Q_{75} - Q_{25}$ / $R_D = Q_{90} - Q_{10}$
- **průměrná absolutní odchylka od mediánu** $d_{\tilde{x}} = \frac{1}{n}\sum_1^n |x_i - \tilde{x}|$
- **statistická entropie** (Claude Shannon) - vhodná zejména pro kvalitativní data, využitelná v teorii informací $H = -\sum_1^m \frac{n_j}{n} \log_b\left(\frac{n_j}{n}\right)$ kde n_j jsou četnosti nenulové četnosti hodnot měřené veličiny. Pro $b=2$ v bitech, $b=e$ v natech, $b=10$ v banech.
  Entropii lze normalizovat vydělením délkou informace → metrika entropie = míra náhodnosti informace
  - funkce $S(t) = \frac{1}{n}\sum_1^n (x_i - t)^2$ nabývá minima $\iff t = \bar{x}$
    $T(t) = \frac{1}{n}\sum_1^n |x_i - t|$ nabývá minima $\iff t = \tilde{x}$
  - Entropie maximální $\iff \frac{n_1}{n} = \frac{n_2}{n} = \dots = \frac{n_m}{n}$

VIZUALIZAČNÍ NÁSTROJE

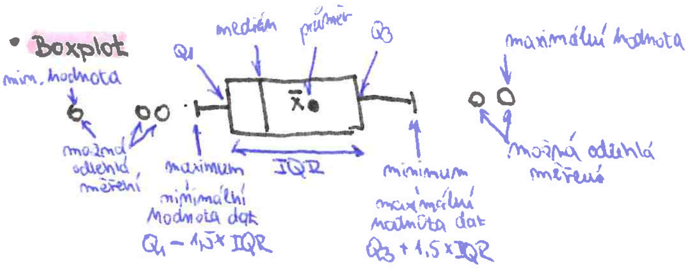
- Matplotlib - jazyk Python
- Plotly - Python, R
- Seaborn - Python, postavena na Matplotlib
- ggplot2 - R

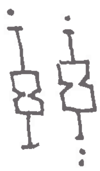
STATISTICKÉ GRAFY

- **Histogram**

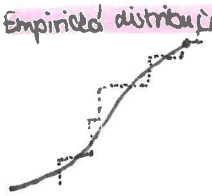


- **Boxplot**





- **Empirická distribuční fce**



- **Probability plot**

(22) NÁHODNÝ VÝBĚR

NÁHODNÝ VÝBĚR - vektor $X_1, X_2, \dots, X_n$ složený z n nezávislých stejně rozdělených NV nazýváme **NÁHODNÝ VÝBĚR ROZSAHU n**.

- NV X_i jsou popsány distribuční funkcí (případně hustoty / pravděpodobnostní) s parametry $\Theta = (\Theta_1, \dots, \Theta_k)$, tedy $X_i \sim F(\Theta_1, \Theta_2, \dots, \Theta_k)$, střední hodnota [illegible], rozptyl $\sigma^2(\Theta)$, kde Θ jsou parametry rozdělení
- realizace náhodného výběru $x_1, x_2, \dots, x_n$ tvoří datový soubor
- náhodný výběr a datový soubor lze popsat pomocí charakteristik = funkce výběru $X_1, X_2 \dots$ resp. funkce hodnot dat. soubor

STATISTIKA - libovolnou NV T_n, která vznikne jako funkce náhodného výběru $X_1, \dots, X_n$ nazýváme STATISTIKOU $T_n = T(X_1 \dots X$

VÝBĚROVÉ CHARAKTERISTIKY - Nechť $X_1, X_2, \dots, X_n$ je náhodný výběr rozsahu n z rozdělení s distribuční funkcí $F(x, \theta)$. Pak statistika

$\bar{X}_n = \frac{1}{n}\sum_1^n X_i$ výběrový průměr

$S_n^2 = \frac{1}{n-1}\sum_1^n (X_i - \bar{X}_n)^2$ výběrový rozptyl

$S_n = \sqrt{S_n^2}$ výběrová směrodatná odchylka

Střední hodnota a rozptyl výběrového průměru a rozptylu - Nechť $X_1, \dots, X_n$ je náhodný výběr rozsahu n z rozdělení se střední hodnotou $\mu = \mu(\theta)$ a rozptylem $\sigma^2 = \sigma^2(\theta)$. Pak

i) výběrový průměr: $\mathbb{E}(\bar{X}_n) = \mu$ $\quad \mathbb{D}(\bar{X}_n) = \sigma^2/n$

ii) výběrový rozptyl: $\mathbb{E}(S_n^2) = \sigma^2$ $\quad \mathbb{D}(S_n^2) = \frac{\mu_4}{n} - \frac{\sigma^4(n-3)}{n(n-1)}$, kde μ_4 je čtvrtý centrální moment $\mu_4 = \mathbb{E}[(X-\mu)^4]$

ROZDĚLENÍ VÝBĚROVÝCH CHARAKTERISTIK - pro náhodný výběr rozsahu n z normálního rozdělení $X \sim N(\mu, \sigma^2)$ platí:

$\bar{X}_n = \frac{1}{n}\sum_1^n X_i$ má rozdělení	$N(\mu, \sigma^2/n)$ - normální
$Z = \frac{\bar{X}_n - \mu}{\sigma}\sqrt{n}$	$N(0,1)$ - normální
$t = \frac{\bar{X}_n - \mu}{S_n}\sqrt{n}$	$t(n-1)$ - Studentovo t-rozdělení se stupněm volnosti $\nu = n-1$
$\frac{(n-1)S_n^2}{\sigma^2}$	$\chi^2(n-1)$ - χ^2 rozdělení se st. volnosti $\nu = n-1$

23 BODOVÝ ODHAD

BODOVÝ ODHAD - Uvažujeme NV X s fcí hustoty $f(x,\Theta)$ s neznámým parametrem Θ.
Pak bodovým odhadem Θ rozumíme jakoukoliv statistiku (fci) ${}_\Theta T_m = T(X_1,\dots,X_m)$, která nezávisí Θ.
Pak bod. odhad Θ založený na statistice ${}_\Theta T_m$ a dané realizaci výběru je hodnota statistiky T s dosazenou realizací $\mathbb{X}$
a značíme $\hat{\Theta} = {}_\Theta T_m(\mathbb{X})$

Vlastnosti i) statistika $T = {}_\Theta T_m$ poskytuje **nestranný odhad**, pokud platí $\mathbb{E}(T(X_1,\dots,X_m)) = \Theta$ $\forall$ možné Θ
Vychýlení odhadu $Bias[{}_\Theta T_m]$ je rozdíl mezi střední hodnotou odhadu a odhadovým parametrem
$Bias[{}_\Theta T_m] = \mathbb{E}[{}_\Theta T_m(X_1,X_2,\dots,X_m)] - \Theta$
Střední kvadratická chyba odhadu je $MSE({}_\Theta T_m) = \mathbb{E}({}_\Theta T_m - \Theta)^2 = \mathbb{D}({}_\Theta T_m) - (Bias({}_\Theta T_m))^2$
ii) statistika T poskytuje **konzistentní odhad**, pokud s rostoucím m „zpřesňujeme odhad", tj
$\lim_{m\to\infty} P(|{}_\Theta T_m - \Theta| < \varepsilon) = 1$ $\forall$ možné Θ

iii) statistika ${}_\Theta T_m$ poskytuje **lepší odhad** než statistika ${}_\Theta U_m$, pokud $\mathbb{D}({}_\Theta T_m) < \mathbb{D}({}_\Theta U_m)$

- Nechť statistika ${}_\Theta T_m$ je nestranný nebo asymptoticky nestranný odhad parametru Θ a platí $\lim_{m\to\infty} \mathbb{D}({}_\Theta T_m) = 0$.
Pak ${}_\Theta T_m$ je konzistentním odhadem Θ

INTERVALOVÉ ODHADY - Platí-li pro statistiky $D = D(X_1,\dots,X_m)$ a $H = H(X_1,\dots,X_m)$ vztah $P(D \le \Theta \le H) = 1-\alpha$
říkáme, že $(D;H)$ je **intervalovým odhadem parametru** Θ s koeficientem spolehlivosti $1-\alpha$.
Intervaly tvaru $(D;H)$ se nazývají dvoustranné, tvaru $(-\infty,H)\,/\,(D,+\infty)$ jednostranné. Přesnost je dána $|H-D|$

- pro parametr μ normálního rozdělení (známý rozptyl σ^2) $\bar{X} \pm u_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{m}}$
(neznámý σ^2) $\bar{X} \pm t_{1-\frac{\alpha}{2}}(\nu = m-1) \cdot \frac{S}{\sqrt{m}}$
- pro parametr σ^2 normálního rozdělení $\left(\frac{(m-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}(\nu=m-1)} ; \frac{(m-1)S^2}{\chi^2_{\frac{\alpha}{2}}(\nu=m-1)} \right)$
- pro parametr p alternativního rozdělení $\hat{p} \pm u_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{m}}$

24 METODY KONSTRUKCE BODOVÝCH ODHADŮ

- **MOMENTOVÁ METODA** - porovnává momenty NV a výběrové momenty
- **METODA MAXIMÁLNÍ VĚROHODNOSTI** - založena na maximalizaci věrohodnostní funkce L, kde tato funkce pro náhodný výběr má tvar

$$L(\Theta) = \prod_{i=1}^{m} f(x_i,\Theta) = f(x_1,\Theta)\cdot f(x_2,\Theta)\cdot\ldots\cdot f(x_m;\Theta)$$

25 FORMULACE STATISTICKÝCH HYPOTÉZ

- na základě realizace náh. výběru $X_1,X_2,\dots,X_n$ rozsahu m ověřujeme určitou hypotézu týkající se NV X.
- statistická hypotéza je tvrzení
  - o parametrech (parametrické testy) pozorované NV pocházející ze základního souboru
  - o tvaru rozdělení znaku v základním souboru (nepar. testy) na základě pozorované NV

- testování statist. hypotéz je jednoduchý rozhodovací postup, při němž se na základě výsledku získaných náhodným výběrem vysloví buď pro testovanou (nulovou) hypotézu nebo pro alternativní hypotézu

CHYBY I. a II. DRUHU u STATISTICKÝCH TESTŮ

úsudek o H_0 \ skutečnost	H_0 je pravdivá	H_0 je nepravdivá
nezamítá se H_0	správné rozhodnutí $P = 1-\alpha$	chyba II. druhu $P = \beta$ ← dopočítává se
zamítá se H_0	chyba I. druhu $P = \alpha$ pevné	správné rozhodnutí $P = \underbrace{1-\beta}_{\text{síla testu}}$

α - ppst. chyby I. druhu
β - ppst. chyby II. druhu

26 STATISTICKÉ TESTOVÁNÍ

- Předpokládáme, že $X_1, X_2, \dots, X_n$ je náhodný výběr z rozdělení s distribuční funkcí F_θ, $\theta \in \Theta$. Dále předpokládáme, že o parametru θ ∃ dvě konkurující si hypotézy H_0 je $\theta \in \Theta_0$ a H_1 je $\theta \in \Theta - \Theta_0$. Testem rozumíme rozhodovací postup, který nám na základě realizace náh. výběru umožní zamítnout nebo nezamítnout platnost hypotézy H_0.

Rozhodovací algoritmus testů

- **testová statistika** = stat. vypočtená z náh. výběru $T = T(X_1, X_2, \dots, X_n)$
- statistické rozhodování probíhá na základě **kritického oboru** W, kdy kritický obor volíme tak, aby při platnosti H_1 hodnota testové statistiky padla do ↑ co nejčastěji a při platnosti H_0 zřídka

  pokud $T \in W$ ⟶ zamítám nulovou hypotézu a pokud $T \notin W$ ⟶ nezamítám H_0
- velikost W volíme tak, abychom platnou hypotézu H_0 zamítali nejvýše s ppstí α - jedná se o maximální povolenou chybu I. druhu a nazýváme ji **hladina testu**

  volíme obvykle $\alpha = 5\%$ (případně 1% nebo 10%)
- výpočetní SW poskytují info o nejmenší hladině, při které bychom hypotézu H_0 zamítli, tato nejmenší hladina se nazývá **p-hodnota testu**
  - udává mezní hodnotu hladiny významnosti, při které zamítáme H_0
  - H_0 zamítáme na hladině α ⟺ hodnota $p < \alpha$
  - nízké hodnoty p ⟶ zamítám nulovou hypotézu

    vysoké p-hodnoty ⟶ nezamítám nulovou hypotézu

27 JEDNOVÝBĚROVÉ PARAMETRICKÉ TESTY

- Pokud $X_1, \dots, X_n$ je náh. výběr z normálního rozdělení se známým rozptylem σ^2

- **Test o střední hodnotě při známém rozptylu ... Z-test**

nulová hypotéza H_0	alternativní H_1	testová statistika $T(X_1, X_2, \dots, X_n)$	kritický obor $T \in W$
	$\mu \neq \mu_0$		$\lvert U \rvert > u_{1-\alpha/2}$
$\mu = \mu_0$	$\mu > \mu_0$	$U = \frac{\bar{x} - \mu_0}{\sigma} \sqrt{n}$	$U > u_{1-\alpha}$
	$\mu < \mu_0$		$U < -u_{1-\alpha}$

- **Test o stř. hodnotě při neznámém rozptylu ... T-test**

H_0	H_1	$T(X_1, \dots, X_n)$	$T \in W$
	$\mu \neq \mu_0$		$\lvert T \rvert > t_{1-\alpha/2}\ (\nu = n-1)$
$\mu = \mu_0$	$\mu > \mu_0$	$T = \frac{\bar{x} - \mu_0}{S} \sqrt{n}$	$T > t_{1-\alpha}\ (\nu = n-1)$
	$\mu < \mu_0$		$T < -t_{1-\alpha}\ (\nu = n-1)$

- **Test o rozptylu** - neznámá střední hodnota

H_0	H_1	$T(X_1, \dots, X_n)$	$T \in W$
	$\sigma^2 \neq \sigma_0^2$		$T \in (\chi^2_{\alpha/2}(\nu = n-1); \chi^2_{1-\alpha/2}(\nu = n-1))$
$\sigma^2 = \sigma_0^2$	$\sigma^2 > \sigma_0^2$	$T = \frac{(n-1)S^2}{\sigma_0^2}$	$T > \chi^2_{1-\alpha}(\nu = n-1)$
	$\sigma^2 < \sigma_0^2$		$T < \chi^2_{\alpha}(\nu = n-1)$

(28) DVOUVÝBĚROVÉ PARAMETRICKÉ TESTY

• **Test o shodě středních hodnot - T-test** - $X_1,\dots,X_{n_1}$ je náh. výběr z normálního rozdělení $N(\mu_1,\sigma^2)$ a $Y_1,\dots,Y_{n_2}$ je na něm nezávislý náh. výběr z $N(\mu_2,\sigma^2)$

nulová hypotéza H_0	alternativní H_1	testová statistika T	kritický obor $T\in W$
	$\mu_1\neq\mu_2$		$\lvert T\rvert > t_{1-\alpha/2}\,(\nu=\min(n_1,n_2)-1)$
$\mu_1=\mu_2$	$\mu_1>\mu_2$	$T=\dfrac{\bar{x}-\bar{y}}{\sqrt{S_1^2/n_1+S_2^2/n_2}}$	$T>t_{1-\alpha}\,(\nu=\min(n_1,n_2)-1)$
	$\mu_1<\mu_2$		$T<-t_{1-\alpha}\,(\nu=\min(n_1,n_2)-1)$

• **Test o shodě rozptylů F-test** - -||- ↑

H_0	H_1	T	$T\in W$
$\sigma_1^2=\sigma_2^2$	$\sigma_1^2\neq\sigma_2^2$		$F\notin(F_{\alpha/2}\ ;\ F_{1-\alpha/2}\ (\nu_1=n_1-1,\nu_2=n_2-1)$
	$\sigma_1^2>\sigma_2^2$	$F=S_1^2\,/\,S_2^2$	$F>F_{1-\alpha}\,(\nu_1=n_1-1,\nu_2=n_2-1)$
	$\sigma_1^2<\sigma_2^2$		$F<F_\alpha\,(\nu_1=n_1-1,\nu_2=n_2-1)$

• **Test o koeficientu korelace (lineární závislosti)** - Pro NV X a Y lze lineární vazbu vyjádřit pomocí kovariance

$$\mathrm{cov}(X,Y)=\mathbb{E}[(X-\mathbb{E}X)\cdot(Y-\mathbb{E}Y)]=\mathbb{E}(X\cdot Y)-\mathbb{E}(X)\cdot\mathbb{E}(Y)$$ nebo korelace

$$\mathrm{cor}(X,Y)=\varrho=\frac{\mathrm{cov}(X,Y)}{\sqrt{\mathbb{D}(X)\cdot\mathbb{D}(Y)}}$$

• **Pearsonův koeficient korelace** - Pokud $(X_1,Y_1),(X_2,Y_2),\dots,(X_n,Y_n)$ je náhodný výběr z dvourozměrného rozdělení, pak Pearsonův koeficient korelace je definován

$$r=\frac{\sum_1^n(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_1^n(X_i-\bar{X})^2}\cdot\sqrt{\sum_1^n(Y_i-\bar{Y})^2}}$$

$\bar{X},\bar{Y}$ - aritmetické průměry prvních a druhých měření

-ekvivalentně lze vyjádřit pomocí součinů z-skórů

$$r=\frac{1}{n-1}\sum_1^n\left(\frac{X_i-\bar{X}}{S_X}\right)\left(\frac{Y_i-\bar{Y}}{S_Y}\right),\quad S_X=\sqrt{\frac{1}{n-1}\sum_1^n(X_i-\bar{X})^2}$$ směrodatná odchylka

• v případě, kdy 1 ze znaků je binární (např. sledujeme-li závislost binární položky Y a celkového počtu bodů X, tzv index RIT) se kor. koef. redukuje na **bodově-biseriální korelační koef.**

$$r_{bis}=\frac{(\bar{X}_1-\bar{X}_0)(n_0 n_1)}{S\cdot n\cdot(n-1)}$$

$\bar{X}_1$ - průměr celkového počtu bodů studentů, kteří mají hodnotu sledované položky rovnou 1

$\bar{X}_0$ - -||- .. rovnou 0

S^2 - STD odchylka celkových bodových zisků, n_1 - #1, n_0 - #0

• Pokud $(X_1,Y_1),\dots,(X_n,Y_n)$ je náh. výběr z dvourozměrného normálního rozdělení, kde kor. koeficient $\varrho=0$

H_0	H_1	T	$T\in W$
	$\varrho\neq0$		$\lvert R\rvert>t_{1-\alpha/2}\,(\nu=n-2)$
$\varrho=0$	$\varrho>0$	$R=\dfrac{r}{\sqrt{1-r^2}}\cdot\sqrt{n-2}$	$R>t_{1-\alpha}\,(\nu=n-2)$
	$\varrho<0$		$R<-t_{1-\alpha}\,(\nu=n-2)$

(29) χ^2 TEST DOBRÉ SHODY

Ho ppsti výsledků jsou $p_1, p_2, \dots, p_k$ (data mají rozdělení dané těmito ppstmi). Porovnáváme očekávané četnosti $n_i^0 > 5$ určené pomocí ppsti p_i a naměřené četnosti n_i. Postup je:

- rozdělíme obor hodnot NV X na k nepřekrývajících se tříd
- zjistíme, kolik hodnot realizovaného náh. výběru se nachází v jednotlivých třídách. Počty prvků v jednotlivých třídách označíme n_i
- pokud je $m > 0$ parametrů rozdělení neznámých → odhadneme je (k dispozici máme tedy po tomto kroku ppsti p_i dané tímto rozdělením)
- pro každou třídu spočteme očekávaný # hodnot v této třídě, ozn. n_i^0. Platí $n_i^0 = n p_i$
- v případě, že je v některé třídě # očekávaných hodnot < 5 → tuto třídu sdružíme s jinou
- testovací statistika má tvar $\chi^2 = \sum_1^k \frac{(n_i - n_i^0)^2}{n_i^0}$
- test. statistika má za platnosti Ho asymptoticky χ^2-rozdělení se stupněm volnosti $\nu = k-1-m$
- Hypotézu, že se NV řídí předpokládaným modelem zamítáme na hladině významnosti α, je-li $\chi^2 > \chi^2_{1-\alpha}(\nu)$ kde $\chi^2_{1-\alpha}(\nu)$ je kvantil χ^2 rozdělení s volností $\nu = k-1-m$

(30) TESTOVÁNÍ NEZÁVISLOSTI

– Testujeme nezávislost 2 NV X a Y, pro které máme k dispozici údaje o četnostech n_{ij} obsažené v kontingenční tabulce

	y_1	y_2	...	y_j	součty
x_1	n_{11}	n_{12}	...	n_{1J}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	...	n_{2J}	$n_{2\bullet}$
...	...	...	...	...	...
x_I	n_{I1}	n_{I2}	...	n_{IJ}	$n_{I\bullet}$
součty	$n_{\bullet 1}$	$n_{\bullet 2}$	...	$n_{\bullet J}$	n

- Při platnosti nezávislosti lze očekávané četnosti vypočítat podle vztahu (požadujeme $n_i^0 > 5$) $n_{ij}^0 = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n}$
- testovací statistika má tvar: $\chi^2 = \sum^I \sum^J \frac{(n_{ij} - n_{ij}^0)^2}{n_{ij}^0}$
- test. stat. má za platnosti nulové hypotézy (nezávislost obou veličin) asymptoticky χ^2-rozdělení se stupněm volnosti $\nu = (I-1)(J-1)$
- hypotézu, že se NV nezávislé, zamítáme na hladině významnosti α, je-li $\chi^2 > \chi^2_{1-\alpha}(\nu)$ kde $\chi^2_{1-\alpha}(\nu)$ je kvantil χ^2 rozdělení s volností $\nu = (I-1)(J-1)$

(31) LINEÁRNÍ REGRESNÍ MODEL

– zjišťujeme, zda hodnoty Y (vysvětlovaná proměnná) závisí na hodnotě x (vysvětlující prom.) Uvažujeme model $Y = f(\beta_0, \beta_1, \dots, \beta_k; x) + \varepsilon$, kde $\varepsilon \sim N(0, \sigma^2)$ reprezentuje nesystematické chyby.

Model odhadujeme na základě sady měření $(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)$

Příklady lineárních modelů

$Y = \beta_0 + \beta_1 x + \varepsilon$ – lineární model

$Y = \beta_0 + \beta_1 \ln(x) + \varepsilon$ – logaritmický m.

$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$ – polynomiální stupně 2

$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon$ – modely s více vysvětlujícími proměnnými

Příklady nelineárních modelů

$Y = \beta_0 \cdot e^{\beta_1 x} + \varepsilon$ – exponenciální

$Y = \beta_0 \cdot x^{\beta_1} + \varepsilon$ – mocninný

MATICOVÝ ZÁPIS LIN. REG. MODELU

vektor náhodné složky

$$\underbrace{\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_m \end{bmatrix}}_{} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{m1} & \cdots & x_{mk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

matice typu $n \times (k+1)$ hodnot vysvětlujících proměnných – regres. matice, # sloupců = p = k+1

β = vektor hledaných k+1 parametrů

- $Y = [Y_1, Y_2, \dots, Y_n]^T$ je vektor vysvětlované proměnné

- **Předpoklady regresního modelu a normálního reg. modelu**

(P1) $E(\varepsilon_i) = 0 \quad \forall i = 1, 2, \dots, n$

(P2) $D(\varepsilon_i) = \sigma^2$ a $cov(\varepsilon_i, \varepsilon_j) = 0$ pro $i \neq j$

(*P1+P2) $\varepsilon \sim N(0, \sigma^2 I_n)$

(P3) ∄ přesná lineární vazba mezi sloupci matice, matice X má plnou hodnost

(P4) X je nestochastická

(32) ODHAD PARAMETRŮ LIN. REG. MODELU

$\beta_0, \beta_1, \dots, \beta_k$ lze odhadnout např. metodou nejmenších čtverců, která minimalizuje výraz

$$S(\beta_0, \dots, \beta_k) = \sum_i^n (y_i - f(\beta_0, \beta_1, \dots, \beta_k; x_i))^2$$

- **Gaussova Markovova věta** - uvažujeme LIN. REG. M. v maticovém tvaru $\mathbb{Y} = \mathbb{X}\beta + \varepsilon$ s předpoklady (P1)-(P4)
  - i) odhad param. metodou nejmenších čtverců má tvar $\mathbb{b} = (\mathbb{X}^T \cdot \mathbb{X})^{-1} \mathbb{X}^T \mathbb{Y}$
  - ii) $\mathbb{b}$ je nestranným odhadem β (tj. $\mathbb{E}(\mathbb{b}) = \beta$) a $\mathbb{D}(\mathbb{b}) = \sigma^2 (\mathbb{X}^T \cdot \mathbb{X})^{-1}$
  - iii) $S_R^2 = \frac{RSE}{n-k-1}$ je nestranným odhadem σ^2 (tj. $\mathbb{E}(S_R^2) = \sigma^2$), kde $RSE = (\mathbb{Y} - \mathbb{X}\mathbb{b})^T (\mathbb{Y} - \mathbb{X}\mathbb{b})$ - reziduální součet čtverců odchylek

- **Gaussova Markovova věta s předpokladem normality** - uvažujeme LIN. REG. NORMÁLNÍ MODEL v maticovém tvaru $\mathbb{Y} = \mathbb{X}\beta + \varepsilon$ s předpoklady $(^*P_1 + P_2) - (P4)$. Pak
  - i) $\mathbb{b} \sim N(\beta; \sigma^2 (\mathbb{X}^T \cdot \mathbb{X})^{-1})$
  - ii) $\frac{RSE}{\sigma^2} = \frac{S_R^2 (n-k-1)}{\sigma^2} \sim \chi^2$ se stupni volnosti $\nu = n-k-1$
  - iii) Pro $j = 0, 1, \dots, k$ platí $\frac{b_j - \beta_j}{\sqrt{S_R^2 \cdot diag_j (X^T X)^{-1}}} \sim t(\nu = n-k-1)$ kde $diag_j (X^T X)^{-1}$ je prvek na pozici jj matice $(\mathbb{X}^T \mathbb{X})^{-1}$

- Pro lin. reg. model značíme
  - $TSS = \sum_i (y_i - \bar{y})^2$ - celkový součet čtverců $\rightarrow S_T^2 = \frac{TSS}{n-1}$ celkový rozptyl
  - $ESS = \sum_i (\hat{y}_i - \bar{y})^2$ - vysvětlený součet čtverců $\rightarrow S_E^2 = \frac{ESS}{k}$ regresní rozptyl
  - $RSE = \sum_i (y_i - \hat{y}_i)^2$ - reziduální součet čtverců odchylek $\rightarrow S_R^2 = \frac{RSE}{n-k-1}$ residuální rozptyl

Platí: $TSS = ESS + RSE$

(33) KOEFICIENT DETERMINACE $R^2 = 1 - \frac{RSE}{TSS}$, UPRAVENÝ KOEF. DET. $R_a^2 = 1 - \frac{RSE/(n-k-1)}{TSS/(n-1)}$

Pro koef. det. platí
- i) $R^2 = \frac{ESS}{TSS} = \frac{S_E^2 (k)}{S_T^2 (n-1)} = 1 - \frac{S_R^2 (n-k-1)}{S_T^2 (n-1)}$
- ii) $0 \leq R^2 \leq 1$
- iii) $k=1$ (přímková regrese) platí $\sqrt{R^2} = |r|$,

pro $\uparrow$ platí: $R_a^2 = 1 - S_R^2 / S_T^2$

kde r je výběrový korelační koeficient

$$r = \frac{S_{xy}}{S_x \cdot S_y} = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i^n (x_i - \bar{x})^2} \cdot \sqrt{\sum_i^n (y_i - \bar{y})^2}}$$

(34)

- Pro **test významnosti regrese**, kde $H_0: \beta_1 = \beta_2 = \dots = \beta_k$ ver. H_1: alespoň 1 koef. je nenulový se využije testová statistika

$$F = \frac{S_E^2}{S_R^2} = \frac{ESS/k}{RSE/(n-k-1)} = \frac{R^2}{1-R^2} \cdot \frac{n-k-1}{k}$$

kde $F|_{H_0} \sim F(\nu_1 = k, \nu_2 = n-k-1)$ Fisherovo rozdělení

- Pro **test významnosti pro jednotlivé koeficienty**, kde $H_0: \beta_j = 0$ ver. $H_1: \beta_j \neq 0$ se využije testová statistika

$$t = \frac{b_j}{se(b_j)}$$

směrodatná odchylka odhadu koeficientu $se(b_j) = \sqrt{S_R^2 \cdot diag_j (X^T X)^{-1}}$

a platí $t|_{H_0} \sim t(\nu = n-k-1)$