Statistika pro informatiku

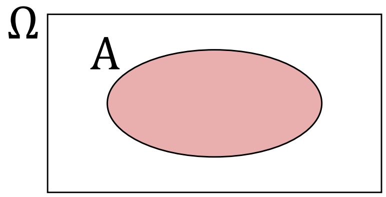
Souhrn látky

březen 2015

1 Základy statistiky a pravděpodobnosti

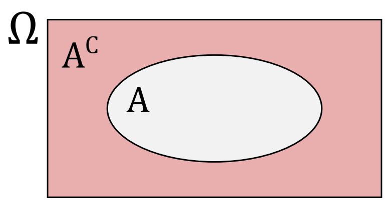
1.1 Pravděpodobnost jevu a jeho doplňku

$$\mathbb{P}(A) = \frac{size(A)}{size(\Omega)}$$



Obrázek 1: Vennův diagram základní pravděpodobnosti jevu

$$\mathbb{P}\left(A^C\right) = 1 - \mathbb{P}(A)$$

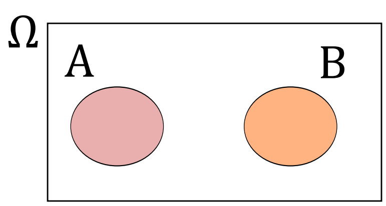


Obrázek 2: Vennův diagram doplňku jevu

1.2 Sjednocení jevů

Pro disjunktní jevy platí

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B).$$

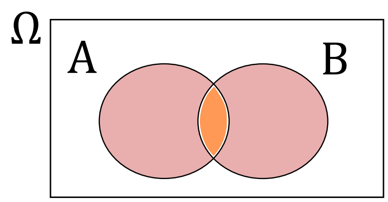


Obrázek 3: Dva disjunktní jevy

Jinak obecně platí (pro nedisjunktní jevy):

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

Oblast průniku by byla započítána dvakrát, proto je potřeba ji odečíst.



Obrázek 4: Sjednocení nedisjunktních jevů

1.3 Průnik jevů

$$\begin{aligned}
\mathbb{P}(A \cap B) &= \mathbb{P}(A|B)\,\mathbb{P}(B) \\
\mathbb{P}(A \cap B) &= \mathbb{P}(B|A)\,\mathbb{P}(A) \\
\mathbb{P}(A \cap B \cap C \ldots) &= \mathbb{P}(A)\,\mathbb{P}(B|A)\,\mathbb{P}(C|A \cap B) \ldots \\
\mathbb{P}(A \cap B \cap C \ldots) &= \mathbb{P}(A|B \cap C)\,\mathbb{P}(B|C)\,\mathbb{P}(C)
\end{aligned}$$

Obecně zapsáno:

$$\mathbb{P}(\text{intersection}) = \mathbb{P}(\text{event}|\text{condition}) * \mathbb{P}(\text{condition})$$

1.4 Nezávislost jevů

U nezávislých jevů platí

$$\begin{aligned}
\mathbb{P}(A|B) &= \mathbb{P}(A) \\
\mathbb{P}(B|A) &= \mathbb{P}(B),
\end{aligned}$$

a proto tedy:

$$\boxed{\mathbb{P}(A \cap B) = \mathbb{P}(A) * \mathbb{P}(B)}.$$

Pokud jsou dva jevy X a Y **spojité a nezávislé**, pak

$$\mathbb{P}(X = Y) = 0.$$

Pokud jsou dva jevy X a Y **stejně rozdělené a nezávislé**, pak

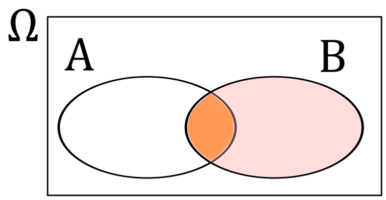
$$\mathbb{P}(X < Y) = \mathbb{P}(Y < X).$$

1.5 Podmíněná pravděpodobnost

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}, \mathbb{P}(B) \neq 0$$

„Pravděpodobnost jevu A za podmínky, že jsme v B a že jev B nastal.“

$$\mathbb{P}(A|B) = \mathbb{P}(B|A) * \mathbb{P}(A)$$



Obrázek 5: Podmíněná pravděpodobnost

1.6 Pravděpodobnostní míra

Pravděpodobnostní míra Q:

$$Q(A) = P(A|C)$$

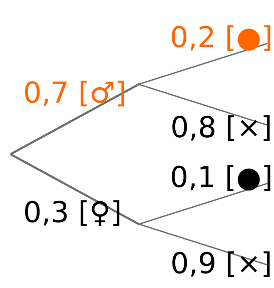
Platí

$$\begin{aligned} 0 &\leq Q(A) \leq 1 \\ Q(A) &= 1 \\ Q\left(\bigcup_{i=1}^{\infty} A_i\right) &= \sum_{i=1}^{\infty} Q(A_i), \text{ pokud jsou } A_i \text{ disjunktní jevy} \end{aligned}$$

1.7 Bayessova věta

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

$$\mathbb{P}(♂ \cap \bullet) = \mathbb{P}(\bullet|♂) * \mathbb{P}(♂) = \mathbb{P}(♂) * \mathbb{P}(\bullet|♂) = 0,7 * 0,2 = \underline{\underline{0,14}}$$



Obrázek 6: Bayessova věta pomocí stromu

1.8 Shrnutí

Jev	**Sjednocení** ($\cup$)	**Průnik** ($\cap$)
Disjunktní jevy	$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$	$\mathbb{P}(A \cap B) = \emptyset$
Nedisjunktní jevy	$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$	$\mathbb{P}(A \cap B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cup B)$
Závislé jevy	$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$	$\mathbb{P}(A \cap B) = \mathbb{P}(A\|B) * \mathbb{P}(B)$
Nezávislé jevy	$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$	$\mathbb{P}(A \cap B) = \mathbb{P}(A) * \mathbb{P}(B)$

Tabulka 1: Shrnutí operací nad různými jevy

2 Vlastnosti

2.1 Střední hodnota $\mathbb{E}X$

Pro diskrétní veličiny

$$\mathbb{E}X = \sum_i p_i x_i = \sum_i x_i * \mathbb{P}(X = x_i)$$

Pro spojité veličiny

$$\mathbb{E}X = \int_{-\infty}^{+\infty} x * f_x(x)\,\mathrm{d}x$$

(P a f jsou funkce hustoty.)

Pro libovolné náhodné veličiny platí:

$$\begin{aligned}
\mathbb{E}(aX + Y) &= a\mathbb{E}(X) + \mathbb{E}(Y) \text{ (linearita)} \\
\mathbb{E}(X \pm Y) &= \mathbb{E}(X) \pm \mathbb{E}(Y) \\
\mathbb{E}(X + Y) &= \mathbb{E}(max\{X, Y\}) + \mathbb{E}(min\{X, Y\}) \\
\mathbb{E}X^2 &= \sum_i p_i x_i^2 \text{ (pro diskrétní jevy)} \\
\mathbb{E}(\max\{X, Y\}) &= \mathbb{E}(X) + E(Y) - \mathbb{E}(\min\{X, Y\}) \\
\mathbb{E}(XY) &= \mathbb{E}X * \mathbb{E}Y \text{ (platí jen pro nezávislé jevy)}
\end{aligned}$$

2.2 Rozptyl

Pro diskrétní náhodnou veličinu jej můžeme definovat vztahem:

$$\sigma^2 = \sum_{i=1}^{n} [x_i - E(X)]^2 p_i = \sum_{i=1}^{n} x_i^2 p_i - [E(X)]^2$$

Pro spojitou náhodnou veličinu definujeme rozptyl vztahem:

$$\sigma^2 = \int_{-\infty}^{+\infty} [x_i - E(X)]^2 f(x)\,\mathrm{d}x = \int_{-\infty}^{+\infty} x_i^2 f(x)\,\mathrm{d}x - [E(X)]^2$$

Dále platí:

$$\begin{aligned} var\left(X\right) &= E\left[\left(X-E\left(X\right)\right)^{2}\right]=E\left(X^{2}\right)-\left(E\left(X\right)\right)^{2} \\ var\left(X\right) &= cov\left(X,\,X\right) \\ var\left(aX\right) &= a^{2}var\left(X\right) \\ var\left(X+a\right) &= var\left(X\right) \\ var\left(X\pm Y\right) &= var\left(X\right)+var\left(Y\right)\pm2cov\left(X,\,Y\right) \end{aligned}$$

2.3 Distribuční funkce

- Funkce je zprava spojitá.

Distribuční funkce pro diskrétní veličiny

$$F = \mathbb{P}(X\leq x_{i})=\sum_{x_{i}\leq x}p_{x}\left(x_{i}\right)$$

Distribuční funkce pro spojité veličiny

$$F=\mathbb{P}(X\leq x_{i})=\int\limits_{-\infty}^{x}f_{x}\left(u\right)\mathrm{d}u\;\forall x\in\mathbb{R}$$

(X je náhodná veličina, x_i je číslo)

2.4 Hustota

Funkce hustoty pro diskrétní veličiny

$$p\left(X\right)=\mathbb{P}\left(X=x\right)$$

Funkce hustoty pro spojité veličiny

$$f\left(x\right)=F_{x}'(x)$$

2.5 Kovariance

Definice
Střední hodnota součinu odchylek obou náhodných veličin X a Y od jejich středních hodnot.[Novovičová-1999]

$$cov\left(X,\,Y\right)=E\left[\left(X-E\left(X\right)\right)*\left(Y-E\left(Y\right)\right)\right]=E\left(XY\right)-E\left(X\right)E\left(Y\right)$$

Platí, že pokud

$$cov\left(X,\,Y\right)=0$$

pak

$$E\left(XY\right)=E\left(X\right)*E\left(Y\right)$$

a X a Y jsou nezávislé.

Dále platí, že

$$\begin{aligned} cov\left(X.Y\right) &= cov\left(Y,\,X\right) \\ cov\left(X,\,X\right) &= var\left(X\right) \end{aligned}$$

2.6 Korelační koeficient

$$\rho(X, Y) = \frac{cov(X, Y)}{\sigma_x * \sigma_y} = \frac{cov(X, Y)}{\sqrt{var(X)} * \sqrt{var(Y)}}$$

3 Rozdělení pravděpodobnosti

Distribuční funkce (F)	$X \leq k$
Hustota	$X = k$
Funkce přežití	$X > k$

Tabulka 2: Funkce a nerovnosti

3.1 Diskrétní (nespojité) rozdělení

Diskrétní veličiny mohou nabývat pouze spočetného počtu hodnot (i nekonečného).

Rozdělení	**Funkce hustoty**	**Distribuční funkce** (F)	$\mathbb{E}X$	$varX$
Bernoulliho, $X \sim Be(p)$	$\mathbb{P}(0) = 1 - p$, $\mathbb{P}(1) = p$	$\times$	p	$p(1-p)$
Binomické, $X \sim Bi(n,\ p)$	$\binom{n}{k} p^k (1-p)^{n-k}$	$I_{1-p}(n-k,\ 1+k)$	$\mathbb{E}X = n * p$	$varX = np(1-p)$
Geometrické, $X \sim geom(p)$	$(1-p)^{k-1} * p$	$\mathbb{P}(T \leq n) = 1 - (1-p)^n$ $\mathbb{P}(T > n) = (1-p)^n$	$\mathbb{E}X = \frac{1}{p}$	$varX = \frac{1-p}{p^2}$
Poissonovo, $X \sim Pois(\lambda)$	$\frac{\lambda^k}{k!} e^{-\lambda}$	$Q(\lfloor k+1 \rfloor,\ \lambda)$	λ	λ

Obrázek 7: Diskrétní rozdělení

3.2 Spojité rozdělení

Spojité náhodné veličiny nabývají na rozdíl od diskrétních veličin nějakého intervalu.

Rozdělení	Funkce hustoty	Distribuční funkce (F)	$\mathbb{E}X$	$varX$
Rovnoměrné, $X \sim Unif\,(a,b)$	$\frac{1}{b-a};\ x \in [a,b]$	$F(X)\begin{cases} 0 & \text{pro } x \leq a \\ \frac{x-a}{b-a} & \text{pro } a < x < b \\ 1 & \text{pro } x \geq b \end{cases}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponenciální, $X \sim Exp\,(n,\ p)$	$\lambda e^{-\lambda x};\ x \in [0,+\infty)$	$\mathbb{P}(X \leq x) = 1 - e^{-\lambda x}$ $\mathbb{P}(X > x) = e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Normální (Gaussovo), $X \sim Geom\,(p)$	$\frac{1}{\sigma\sqrt{2\pi}} * e^{\frac{-(x-\mu)^2}{2\sigma^2}}$	$\times$	μ	σ^2

Obrázek 8: Spojité rozdělení

4 Entropie

Entropie diskrétní veličiny

$$\boxed{H_b(X) = -\sum_{\text{all } i} p_i \log_b p_i}$$

Entropie spojité veličiny

$$\boxed{H_b(X) = -\int_{-\infty}^{+\infty} f(x) \log_b f(x)\,\mathrm{d}x}$$

(b je základ abecedy pro kódová slova, nejčastěji používáme binární abecedu, tedy $b = 2$)

Aditivita entropie

$$\begin{aligned} H(X,Y) &= H(X) + H(Y|X) \\ H(X,Y) &= H(X) + H(Y) \text{ (speciálně jen pro nezávislé náhodné veličiny)} \end{aligned}$$

4.1 Sdružená entropie

$$H(X,Y) = -\sum_{i,j} p_{i,j} \log p_{i,j}$$

(sdružená hustota $p_{i,j} = P(X = x_i,\ Y = Y_j)$)

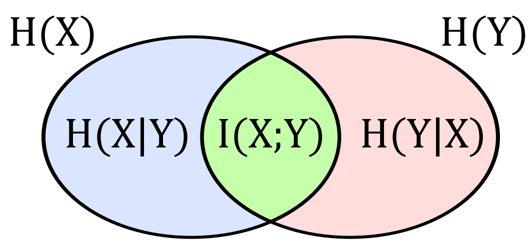
4.2 Podmíněná entropie

$$
\begin{aligned}
H(X|Y) &= -\sum_{i,j} p(x_i, y_i) \log p(y_j|x_i) \\
H(X|Y) &= H(X, Y) - H(Y) \\
\mathbb{P}(x_i|y_j) &= \frac{\mathbb{P}(x_i,\, y_j)}{\mathbb{P}(y_i)} \quad \begin{matrix}\text{sdružená hustota} \\ \text{marginální hustota}\end{matrix} \\
\mathbb{P}(X, Y) &= \mathbb{P}(X|Y) * \mathbb{P}(Y)
\end{aligned}
$$

4.3 Vzájemná informace

$$
\begin{aligned}
I(X; Y) &= H(X) - H(X|Y) \\
I(X; Y) &= H(Y) - H(Y|X) \\
I(X; Y) &= H(X) + H(Y) - H(X, Y) \\
I(X; Y) &= I(Y; X) \\
I(X; X) &= H(X)
\end{aligned}
$$

$$
I(X, Y) = \sum_{i,\, j} p_{i,j} \log \frac{p_{i,j}}{p_i * p_j} = \ldots = H(X) - H(X|Y)
$$



Obrázek 9: Vzájemná informace a entropie (tedy $H(X, Y)$)

4.4 Kódování

Střední délka kódového slova

$$
\begin{aligned}
L(C) &= \mathbb{E}\ell(X) = \sum_i \ell(x_i) * \mathbb{P}(X = x_i) \\
L(C) &\geq H_D(X)
\end{aligned}
$$

Kódování je optimální, pokud se střední délka kódového slova a entropie rovnají ($L(C) = H_D(X)$).

5 Náhodné procesy

Značení procesu

$$
X(t,\ \omega) = X_t = X(t)
$$

Střední hodnota

$$\begin{aligned}\eta_x(t) &= \mathbb{E}X(t) = \int x(t) * f_{X_t}(x)\,\mathrm{d}x \\ &\quad \mathbb{E}X(t) = \sum x_i(t)\,\mathbb{P}(X_t = x_i(t))\end{aligned}$$

Pokud je střední hodnota **nezávislá** na t, **je proces stacionární** v průměru. Tedy

$$\eta(t) = \eta_x\,\forall t$$

Střední hodnota integrálu

$$\mathbb{E}\int_a^b X(t)\,\mathrm{d}t = \int_a^b \mathbb{E}X(t)\,\mathrm{d}t$$

Autokorelační funkce

$$\begin{aligned}R_{xx}(t_1,\ t_2) &= \mathbb{E}\left[X(t_1) * \overline{X(t_2)}\right]\ \text{ v } \mathbb{C} \\ &= \mathbb{E}[X(t_1) * X(t_2)] = \sum_i x_i(t) * P(X(t_1) * X(t_2) = x_i(t))\ \text{ v } \mathbb{R}\end{aligned}$$

V ukázkovém příkladu se autokorelační funkce spočítala pomocí následujícího „triku“ ($0 \leq t_1 \leq t_2 \leq 2$)

$$\begin{aligned}R_x(t_1,\,t_2) &= \mathbb{E}[X(t_1) * X(t_2)] = 0 * \mathbb{P}(X(t_1) * X(t_2) = 0) + 1 * \mathbb{P}(X(t_1) * X(t_2) = 1) \\ &= \mathbb{P}(X(t_1) * X(t_2) = 1) = \mathbb{P}(X(t_1) = 1,\,X(t_2) = 1) = \\ &= \boxed{\mathbb{P}(t_1 \leq A,\,t_2 \leq A) = \mathbb{P}(A \geq t_1,\,t_2 \geq A) = P(A \geq t_2)}\end{aligned}$$

5.1 Exponenciální závody

Spojité, bez paměti (memoryless), bez intenzit.

$$S \sim Exp(\lambda),\ T \sim Exp(\mu)$$

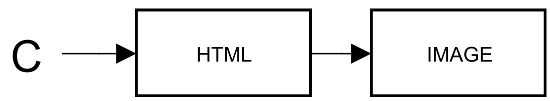
(Náhodné veličiny S a T jsou exponenciálně rozdělené.)

$$\mathbb{E}(S) = \frac{1}{\lambda},\ \mathbb{E}(T) = \frac{1}{\mu}$$

$$\begin{aligned}\mathbb{E}\max\{S,\,T\} &= \mathbb{E}(S) + \mathbb{E}(T) - \mathbb{E}\min\{S,\,T\} \\ \mathbb{E}\min\{S,\,T\} &= \frac{1}{\lambda+\mu} \\ \mathbb{P}(T < S) &= \frac{\mu}{\lambda+\mu} \\ \mathbb{P}(S < T) &= \frac{\lambda}{\lambda+\mu}\end{aligned}$$

5.1.1 Grafická reprezentace

Často pomocí následujícího diagramu (příklad):



Obrázek 10: Diagram exponenciálních závodů

6 Markovovy řetězce

Markovova podmínka

$$P\left(X_n = s_j | X_0 = s_0,\, X_1 = s_1,\, \ldots,\, X_{n-1} = s_i\right) = P\left(X_n = s_j | X_{n-1} = s_i\right) = P_{ij}$$

Matice přechodů

$$\mathbb{P}_{i,j} = p\left(i,\, j\right) = P\left(X_{n+1} = s_j | X_n = s_i\right) = p_{i,j}$$

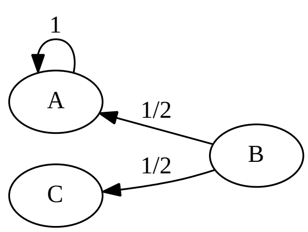
Řetězec zapisujeme

- diagramem,
- maticí přechodů.

6.1 Absorbující Markovův řetězec

V absorbujícím řetězci je pravděpodobnost, že bude proces absorbován 1. Tedy:

$$\lim_{n\to\infty} = Q^n = 0$$



Obrázek 11: Absorbující Markovův řetězec obsahuje rekurentní stavy

$$P = \left(\begin{array}{c|c} Q & R \\ \hline 0 & 1 \end{array}\right)$$

(P – matice přechodů, Q – tranzientní stavy, R – rekurentní /absorbující/ stavy, 1 – jednotková matice, 0 – matice nul)

Fundamentální matice
Předpokládaný počet průchodů v tranzientním (Q) stavu vypočítáme pomocí **fundamentální matice** N:

$$N = (1 - Q)^{-1}$$

$N_{i,j} = \mathbb{E}(\text{počet průchodů stavem j}|\text{začínáme ve stavu i})$, zároveň součet řádku je střední doba absorpce, pokud začínáme v určitém stavu.

Inverzní matice

$$A^{-1} = \frac{1}{\det A} * \operatorname{adj} A$$

Adjungovaná matice je **transponovaná**[a] matice subdeterminantů vynásobená mřížkou znamének:

$$\begin{pmatrix} + & - & + & \cdots \\ - & + & - & \cdots \\ + & - & + & \cdots \\ \vdots & \vdots & \vdots & \cdots \end{pmatrix}^{\mathbf{T}}$$

[a]Prohodíme sloupce a řádky

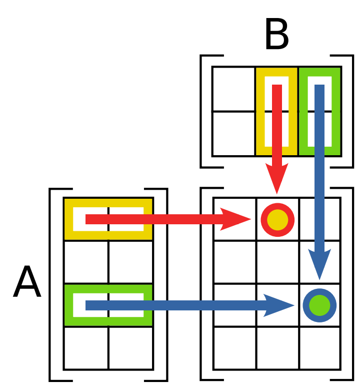
Pravděpodobnostní matice
Pravděpodobnost, že spadneme do rekurentního stavu stavu „j“, začínáme-li v „i“ je

$$B_{i,j} = \mathbb{P}(\text{pohlcení v } s_j | \text{start v } s_i)$$

Matici B vypočítáme jako

$$B = N * R$$

(Nezapomeňme, že matice B kopíruje označení sloupců a řádků z matice přechodů P)



Obrázek 12: Násobení matic, vynásobíme prvek z A s prvkem z B a výsledky sečteme

6.2 Stacionární distribuce

Po čase se některé Markovovy řetězce ustálí v nějakém stavu.

Rovnovážná distribuce

- **Diskrétní**:$\pi : \pi * Q = \pi$
- **Spojité**: $\pi : \pi * Q = 0$ (Zadáno pomocí intenzit λ)

U obou platí

$$\pi_i + \pi_2 + \ldots + \pi_n = 1$$

Detailní rovnováha

$$\pi_i p_{i,j} = \pi_j p_{j,i} \, \forall i, j$$

7 Systémy hromadné obsluhy

Tři modely systému:

- $M/M/1$
- $M/M/m$
- $M/G/\infty$

Parametry systémů

$$\rho = \frac{\lambda}{\mu}$$

(**míra vytížení**; ρ musí být $\rho < 1$, aby systém pracoval)
(λ – **intenzita příchodů**)

$$\mu = \frac{1}{T_S}$$

(μ – **intenzita obsluhy**, T_S – **doba obsluhy**)

$$N = \rho * \frac{1}{1-\rho}$$

(N – průměrný celkový **počet požadavků** v systému)

Průměrný čas ve frontě/systému

$$\boxed{\rho = N_S}$$

$$\begin{aligned} N &= \lambda * T \\ T &= T_Q + T_S \\ N &= \underbrace{N_Q}_{\text{Fronta}} + \underbrace{N_S}_{\text{Obsluha}} \end{aligned}$$

(T – celkový průměrný čas v systému, T_Q – průměrný čas ve frontě/čekání, T_S – průměrný čas zpracování požadavku)

Little's law (střední počet požadavků v systému):

$$\begin{aligned} N &= \sum_{n=0}^{\infty} n * \pi_n \\ N_S &= \lambda * T_S \end{aligned}$$

8 Statistika

$$\overline{X} = \frac{\sum}{N}$$

8.1 Konfidenční intervaly

σ^2 známe:

$$\mu \in \left(\overline{X}_n - \mathcal{Z}_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}; \overline{X}_n + \mathcal{Z}_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right)$$

($\mathcal{Z}$ – tabulky normálního rozdělení)

σ^2 neznáme:

$$\mu \in \left(\overline{X}_n - \mathcal{T}_{\frac{\alpha}{2};n-1}\frac{\mathcal{S}_n}{\sqrt{n}}; \overline{X}_n + \mathcal{T}_{\frac{\alpha}{2};n-1}\frac{\mathcal{S}_n}{\sqrt{n}}\right)$$

($\mathcal{T}$ – tabulky Studentova $\mathcal{T}$ rozdělení)
($;n-1$ je stupeň volnosti studentova rozdělení)

8.2 Testování hypotéz

Testovací statistika, **neznáme** σ:

$$T = \frac{\overline{X}_n - \mu_0}{\frac{\mathcal{S}}{\sqrt{n}}}$$

Testovací statistika, **známe** σ:

$$Z = \frac{\overline{X}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

9 Ostatní

9.1 Řady

Výpočet řad

$$\begin{aligned} \sum_{n=0}^{\infty} r^n &= \frac{1}{1-r} \\ \sum_{n=0}^{\infty} n * r^n &= \frac{r}{(1-r)^2} \end{aligned}$$

Součet nekonečné geometrické řady:

$$S_n = \frac{a_0}{1-q} \text{ pro } |q| < 1$$

(a_0 – první prvek, q – kvocient)

9.2 Logaritmy

$$\begin{aligned} \log_a(x_1 * x_2) &= \log_a x_1 + \log_a x_2 \\ \log_a\left(\frac{x_1}{x_2}\right) &= \log_a x_1 - \log_a x_2 \\ \log_a x^r &= r * \log_a x \\ \log_a a &= 1 \\ \log_a 1 &= 0 \end{aligned}$$

Reference

[Novovičová-1999] NOVOVIČOVÁ, CSC. *Pravděpodobnost a matematická statistika.* Praha, 1999. Skripta. České vysoké učení technické v Praze. Dostupné [online] z http://euler.fd.cvut.cz/publikace/files/skripta3.pdf