1 Pravděpodobnost

1.1 Úvod do pravděpodobnosti

1.1.1 Náhodný jev, opačný a doplňkový

náhodný jev	=	tvrzení o výsledku náhodného pokusu
$A \cup B$	=	A nebo B
$A \cap B$	=	A a zároveň B
$\bar{A}$	=	jev opačný, doplňkový

1.1.2 Elementární jev

Definice:

Jev A se nazývá *elementární*, jestliže neexistují jevy A_1, A_2 takové, že $A_1 \neq A_2,\ A_1 \cup A_2 = A,\ A_1 \neq \varnothing$.

Př:

1. kostka
   padne sudé číslo - není elementární jev
   padne 2 - ano
2. životnost součástky bude méně než 1000 provoz. hodin - není elem.
   $\sim$ bude *přesně* 1000 prov. hodin - ano

Množinu všech el. jevů příslušných určitému náhodnému pokusu označíme Ω.

Věta: (de Morganova pravidla)

$$\overline{A_1 \cap A_2 \cap \ldots \cap A_n} = \overline{A_1} \cup \overline{A_2} \cup \ldots \cup \overline{A_n}$$
$$\overline{A_1 \cup A_2 \cup \ldots \cup A_n} = \overline{A_1} \cap \overline{A_2} \cap \ldots \cap \overline{A_n}$$

1.2 Neslučitelné, nemožné a jisté jevy

Definice:

Jevy A, B se nazývají *neslučitelné*, jestliže $A \cap B \neq \varnothing$; disjunktní jevy.

$\varnothing$ je jev *nemožný*.

Ω je jev *jistý*.

1.3 Statistická definice pravděpodobnosti

Definice: (Statistická definice pravděpodobnosti)

Buď A určitý jev příslušný určitému náhodnému pokusu. Provedeme n opakování pokusu a označíme n_A počet, kolikrát jev A nastal. Definujeme pravděpodobnost jevu A:

$$P(A) = \lim_{n \to \infty} \underbrace{\frac{n_A}{n}}$$

relativní četnost výskytu jevu A

Věta: (základní vlastnosti pravděpodobnosti)

1. vždy $0 \leq P(A) \leq 1$
2. $P(\varnothing) = 0, \quad P(\Omega) = 1$
3. $P(\bar{A}) = 1 - P(A)$
4. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
5. buďte $A_1, A_2, \ldots$ jevy, které jsou neslučitelné, pak $P(\bigcup_i A_i) = \sum_i P(A_i)$

Klasická definice pravděpodobnosti

Věta: (klasická „definice“ pravděpodobnosti)

Nechť Ω je konečná, skládá se z n elementárních jevů, a nechť všechny elem. jevy mají stejnou pravděpodobnost, pak pro libovolný jev A$\subset \Omega$ platí:

$$P(A) = \frac{m}{n}, \text{ kde } m \text{ je počet elem. jevů, ze kterých se jev } A \text{ skládá}$$

Dk. klasické „definice“ pravděpodobnosti:

$\Omega \ldots n$ případů
elem. jevy označíme $E_1, E_2, \ldots, E_n$; $P(E_1) = P(E_2) = \ldots = P(E_n) = x$
ukážeme, že $x = \frac{1}{n}$

$$\begin{aligned} E_1 \cup E_2 \cup \ldots \cup E_n &= \Omega \\ P(E_1 \cup E_2 \cup \ldots \cup E_n) &= 1 \\ P(E_1) + \ldots + P(E_n) &= 1 \qquad A = E_1 \cup \ldots \cup E_m \\ x + x + x + \ldots + x &= 1 \qquad P(A) = P(E_1) + \ldots + P(E_m) = \frac{1}{n} + \cdots + \frac{1}{n} = \\ x &= \frac{1}{n} \qquad = m \cdot \frac{1}{n} = \frac{m}{n} \end{aligned}$$

1.4 Podmíněná pravděpodobnost

Uvažujeme jevy $A, B, P(B) \neq 0$. Opakujme pokus n-kráte a všímejme si pouze těch případů, ve kterých nastal jev B. Jejich počet označme n_B, $n_B \neq 0$. Mezi pokusy, ve kterých nastal jev B si všímejme relativního zastoupení výskytu jevu A, které je vyjádřeno poměrem $\frac{n_{A\cap B}}{n_B} = \frac{\frac{n_{A\cap B}}{n}}{\frac{n_B}{n}}$.
Pro $n \to \infty$ tento podíl konverguje k číslu $\frac{P(A\cap B)}{P(B)}$.

Definice:

Podmíněná pravděpodobnost $P(A \mid B) = \frac{P(A\cap B)}{P(B)}$ je pravděpodobnost jevu A podmíněná jevem B.

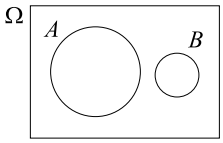
Poznámka: $P(A \cap B) = P(A \mid B) \cdot P(B)$
$P(A \cap B) = P(B \mid A) \cdot P(A)$

1.5 Nezávislé jevy

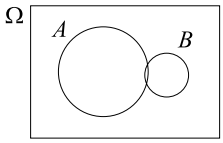
Definice:

Jevy A, B se nazývají *nezávislé*, jestliže $P(A \cap B) = P(A) \cdot P(B)$.

Poznámka: Jestliže $P(A) \neq 0, P(B) \neq 0$, pak A, B jsou nezávislé
$\iff P(A \mid B) = P(A)$
$[P(B \mid A) = P(B)]$



neslučitelné



nezávislé

Definice:

Jevy A, B, C se nazývají nezávislé, jestliže jsou nezávislé dvojice $A, B \quad A, C \quad B, C$ a dále $P(A \cap B \cap C) = P(A) \cdot P(B) \cdot P(C)$.

Věta:

Jestliže $A_1, A_2, \ldots, A_n$ jsou nezávislé, pak také $\bar{A}_1, \bar{A}_2, \ldots, \bar{A}_n$ jsou nezávislé jevy.

Př:
Linka se skládá z n dílů a funguje pouze tehdy, fungují-li všechny díly. Jsou známy pravděpodobnosti $p_i (i = 1, \ldots, n)$, že i-tý díl se porouchá během směny. Jaká je pravděpodobnost poruchy celé linky? Předpokládejme přitom, že poruchy jednotlivých dílů jsou navzájem nezívislé jevy.

řešení:

A_i – i-tý díl se porouchá

$P(A_i) = p_i$

A – jev, že linka se porouchá (během směny)

platí: $A = A_1 \cup A_2 \cup \ldots \cup A_n$
$\bar{A} = \bar{A}_1 \cup \bar{A}_2 \cup \ldots \cup \bar{A}_n$

nezávislost $\Rightarrow P(\bar{A}) = P(\bar{A}_1) \cdot P(\bar{A}_2) \cdot \ldots \cdot P(\bar{A}_n)$
$P(\bar{A}) = (1 - p_1) \cdot (1 - p_2) \cdot \ldots \cdot (1 - p_n)$
$P(A) = 1 - P(\bar{A}) = 1 - (1 - p_1) \cdot (1 - p_2) \cdot \ldots \cdot (1 - p_n)$

např. pro $n = 2 \;\; P(A) = 1 - (1 - p_1) \cdot (1 - p_2) = p_1 + p_2 - p_1 p_2$

Věta:

Pro libovolné (závislé i nezávislé) jevy $A_1, A_2, \ldots, A_n$ platí:

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

za předpokladu, že podmíněné pravděpodobnosti jsou definovány, nutno $P(A_1 \cap \ldots \cap A_{n-1}) \neq 0$.

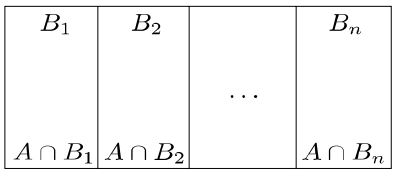
1.6 Věta o úplné pravděpodobnosti

Věta: (o úplné pravděpodobnosti)

Nechť $B_1, \ldots, B_n$ jsou jevy s kladnými pravděpodobnostmi, které jsou neslučitelné a jejich sjednocení je Ω (tzv. disjunktní rozklad Ω). Pak pro libovolný jev $A \subset \Omega$ platí:

$$P(A) = \sum_{i=1}^{n} P(A \mid B_i) \cdot P(B_i)$$

Dk.:



$\sum_{i=1}^{n} P(A \mid B_i) \cdot P(B_i)$

1.7 Bayesova věta

Věta: (Bayesova)

Nechť $B_1, \ldots, B_n$ tvoří disjunktní rozklad Ω, A buď jev s kladnou pravděpodobností. Pak pro $\forall j = 1, \ldots, n$ platí:

$$P(B_j \mid A) = \frac{P(A \mid B_j) \cdot P(B_j)}{P(A)} = \frac{P(A \mid B_j) \cdot P(B_j)}{\sum\limits_{i} P(A \mid B_i) \cdot P(B_i)}$$

Dk.: $P(B_j \mid A) = \frac{P(B_j \cap A)}{P(A)} = \frac{P(A|B_j) \cdot P(B_j)}{P(A)}$

1.8 Statistické soubory

Statistickým souborem rozsahu n rozumíme neuspořádanou n–tici čísel; nezáleží na pořadí, mohou se opakovat.

např.: $4, 8, 5, 8, 12$

Mějme $x_1, x_2, \ldots, x_n$ a definujme:

$\bar{x} = \frac{1}{n} \sum\limits_{i=1}^{n} x_i$ (aritmetický průměr)

$\sigma_n^2 = \frac{1}{n} \sum\limits_{i=1}^{n} (x_i - \bar{x})^2$ (rozptyl – průměr kvadratických odchylek od průměru)

$\sigma_n = \sqrt{\sigma_n^2}$ (směrodatná odchylka)

($\frac{1}{n-1}$ – výběrový(-á))

$s^2 = \frac{1}{n-1} \sum\limits_{i=1}^{n-1} (x_i - \bar{x})^2$ (výběrový rozptyl); zpravidla se užívá toho označení s

$s = \sqrt{s^2}$ (výběrová směrodatná odchylka)

$\sigma_n = \sqrt{\frac{n-1}{n}} \cdot s$

$s = \sqrt{\frac{n}{n-1}} \cdot \sigma_n$

Věta: (výpočetní tvar rozptylu)

$$\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\bar{x})^2 = \overline{x^2} - \bar{x}^2$$

Dk.: $\frac{1}{n}\sum(x_i - \bar{x})^2 = \frac{1}{n}\sum(x_i^2 - 2\bar{x}x_i + \bar{x}^2) = \frac{1}{n}\sum x_i^2 - \frac{1}{n}2\bar{x}\sum x_i + \frac{1}{n}\sum \bar{x}^2$

Variační rozpětí = největší hodnota − nejmenší hodnota

Šikmost statitického souboru $\Rightarrow \alpha = \frac{\frac{1}{n}\sum(x_i-\bar{x})^3}{s^3}$

3.mocnina $\rightarrow$ invariantní vůči změně měřítka

1.8.1 Použití vážených průměrů

Předpokládejme, že ve statistickém souboru rozsahu n se vskytla

hodnota		
	x_1	n_1 – kráte
	x_2	n_2 – kráte
	⋮	⋮
	x_k	n_k – kráte
		$n = n_1 + n_2 + \ldots + n_k$

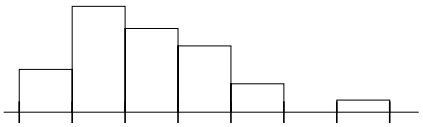
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k}(x_i \cdot n_i) = \sum_{i=1}^{k} x_i \frac{n_i}{n} = \sum_{i=1}^{k} x_i v_i, \text{ kde } v_i = \frac{n_i}{n} \text{ se nazývají vahami}$$

$\bar{x}$ je váženým průměrem hodnot $x_1, x_2, \ldots, x_k$

$\sum v_i = 1$

1.8.2 Histogram četností

Pro statistický soubor velkého rozsahu (alespoň několik desítek) lze rozložení souboru graficky přehledně vyjádřit pomocí tzv. histogramu četností (viz skripta Reif).



$\rightarrow$ intervaly = třídy
kraje *ne*celočíselné

Jestliže třídy četnosti vydělíme rozsahem souboru n, obdržíme relativní třídní četnosti.

Obsahy sloupců by měly být rovny relativním třídním četnostem.

Uvažujeme $n \rightarrow \infty$, zjemňujeme třídní dělení, v limitě obdržíme jistou křivku (hustota pravděpodobnosti).

1.9 Náhodné veličiny

Definice:

Náhodná veličina (NV) je funkce, která každému elementárnímu jevu určitého náhodného pokusu přiřazuje číslo.

— nejčastější označení X, Y; někdy ξ

Definice:

Jestliže NV může nabývat jen hodnot z konečné nebo spočetné množiny, nazývá se veličinou diskrétního typu.
Jestliže NV může nabývat všech hodnot v nějakém intervalu, nazývá se veličinou spojitého typu.

1.10 Veličina diskrétního typu

1.10.1 Pravděpodobnostní funkce

Definice:

Je-li X NV diskrétního typu, pak definujeme funkci $\mathcal{P}(x) = P(X = x),\ x \in (-\infty; +\infty)$, říkáme jí *pravděpodobnostní funkce*.

1.10.2 Střední hodnota

Definice:

Buď X NV diskrétního typu, která může nabývat s nenulovou pravděpodobností pouze hodnot $x_1, x_2, \ldots$ a je konečná nebo spočetná množina. Číslo $\sum_i x_i \cdot \mathcal{P}(x_i)$ se nazývá *střední hodnotou* a značí se $E(X)$.

$$E(X) = x_1 \cdot P(X = x_1) + x_2 \cdot P(X = x_2) + \ldots$$

...zobecnění váženého průměru

Je-li $h(x)$ nějaká reálná funkce, pak $E(h(X)) \overset{\text{def.}}{=\!=} \sum_i h(x_i) \cdot \mathcal{P}(x_i)$.

1.10.3 Obecné a centrální momenty

Definice:

Číslo $\mu'_k \overset{\text{def.}}{=\!=} \sum x_i^k \cdot \mathcal{P}(x_i) = E(X^k)$ se nazývá *k-tý obecný moment*.

Číslo $\mu_k \overset{\text{def.}}{=\!=} \sum (x_i - E(X))^k \cdot \mathcal{P}(x_i) = E[X - E(X)]^k$ se nazývá *k-tý centrální moment*.

1.10.4 Rozptyl

Hodnota $E([X - E(X)]^2)$ se nazývá *rozptyl* a značí se $D(X)$ (= ang. disperse) nebo $var(X)$ (variance) nebo $\sigma^2(X)$.

Odmocnina je směrodatná odchylka a značí se $\sigma(X) \overset{\text{def.}}{=\!=} \sqrt{D(X)}$.

Poznámka: $E(X + Y) = E(X) + E(Y)$, pokud střední hodnoty existují

$E(c \cdot X) = c \cdot E(X)$

$D(c \cdot X) = c^2 \cdot D(X)$, zde násobíme pouze jednu veličinu, NENÍ to součet veličin!

1.10.5 Nezávislé veličiny

Definice:

Veličiny X, Y se nazývají *nezávislé*, jestliže pro všechny dvojice reálných čísel x, y jsou jevy $X < x$, $Y < y$ nezávislé.

Jsou-li X, Y nezávislé, pak $D(X+Y) = D(X) + D(Y)$.
Obecně to nemusí platit.

Věta: (výpočetní tvar rozptylu)

$D(X) = E(X^2) - E^2(X)$

1.10.6 Příklady veličin diskrétního typu

I. *Alternativní rozdělení* pravděpodobnosti s parametrem $p \in (0;1)$

Může nabývat hodnot 0 nebo 1.

Značí se $X \sim A(p)$.

$$\mathcal{P}(1) = p \qquad E(X) = p$$
$$\mathcal{P}(0) = 1-p \qquad D(X) = p(1-p)$$

Př:
Při kontrole kvality — zmetek $\times$ dobrý $\equiv 1 \times 0$

II. *Hypergeometrické rozdělení* s parametry M, N, n

Nechť $M \leq N$, $n \leq N$, pak X může nabývat nezáporných celočíselných hodnot, $X \leq M$, $X \leq n$.

Značí se $X \sim H(M,N,n)$.

$$P(X=i) = \frac{\binom{M}{i} \cdot \binom{N-M}{n-i}}{\binom{N}{n}} \qquad \begin{matrix} \text{pro } i \geq 0 \\ i \leq M \\ i \leq n \end{matrix}$$

Př:
Uvažujme celkový soubor o N prvcích, mezi nimiž M prvků má určitou vlastnost v. Ze souboru se náhodně vybere n prvků. Potom počet X prvků, které mezi nimi mají uvažovanou vlastnost v má rozdělení $X \sim H(M,N,n)$. Hovoří se o tzv. náhodném výběru bez vracení zpět.

III. *Binomické rozdělení* s parametry n, p

X může nabývat pouze hodnot $0, 1, \ldots, n$.

Značí se $X \sim Bi(n,p)$.

Poznámka: $\sum \binom{n}{i} p^i (1-p)^{n-i} = [p + (1-p)]^n = 1^n = 1$

$$P(X=i) = \binom{n}{i} p^i (1-p)^{n-i} \text{pro } i = 0,1,\ldots,n \qquad E(X) = n \cdot p$$
$$D(X) = np(1-p)$$

Př:
Uvažujme jev, jehož pravděpodobnost je p. Provedeme n nezávislých pokusů. Počet X, kolikrát sledovaný jev nastal má rozdělení $X \sim Bi(n,p)$. $\rightarrow$ pro výběr s vracením zpět

IV. *Poissonovo rozdělení* s parametrem $\lambda > 0$

Může nabývat hodnot $0, 1, 2, \ldots$

Značí se $X \sim Po(\lambda)$.

$$P(X = i) = e^{-\lambda} \cdot \frac{\lambda^i}{i!} \qquad E(X) = \lambda$$
$$D(X) = \lambda$$

Př:
Počet určitých událostí za časovou jednotku (např. pořet poruch zařízení) má často Poissonovo rozdělení.

Věta:

$Po(\lambda)$ je limitním případem $Bi(n,p)$ pro $n \to \infty, p \to 0, n \cdot p \to \lambda$.

Důsledek: Pro velké n $(n \geq 30)$ a malé p $(p \leq 0,1)$ se používá aproximace $Bi(n,p) \approx Po(\lambda)$, kde $\lambda = n \cdot p$.

1.10.7 Distribuční funkce

Definice:

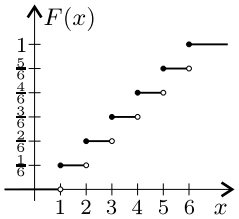
Buď X NV spojitého nebo diskrétního typu. Pak funkci $F(x)$ definovanou pro $x \in (-\infty; +\infty)$ předpisem

$$F(x) = P(X \leq x)$$

!

nazýváme *distribuční funkcí* NVy X.

Př:
X = počet bodů při hodu kostkou



Distribuční funkce diskrétní veličiny je nespojitá, distribuční funkce spojité NV je spojitá.

Věta:

Pro distribuční funkci vždy platí:

1. $0 \leq F(x) \leq 1$
2. F je neklesající
3. $\lim_{x \to +\infty} F(x) = 1$
4. $\lim_{x \to -\infty} F(x) = 0$

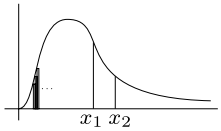
1.11 Veličina spojitého typu

Sledujeme náhodný pokus a mějme např. omezený interval. Ten rozdělíme a tím dostaneme relativní třídní četnosti, histogram. Interval dále zjemňujeme a sledujeme ho. Sloupce histogramu se ztenčují a v limitě vznikne křivka.

1.11.1 Hustota pravděpodobnosti

Hustota pravděpodobnosti $f(x)$ je limitním případem histogramu relativních třídních četností.

Př:



diskétní veličina nemá hustotu pravděpodobnosti

zjemňováním dostaneme nekonečně dlouhé a nekonečně tenké čáry $\rightarrow \int$ nemá smysl

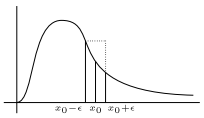
Pro $\forall x_1 < x_2$ platí:

$$\int_{x_1}^{x_2} f(x)dx = P(x_1 < X < x_2)$$

Věta:

Buď X NV spojitého typu. Pak platí:

1. $\forall x_0 \in (-\infty; +\infty)$ je pravděpodobnost $P(X = x_0) = 0$
2. $\forall x_1 < x_2 : P(x_1 < X < x_2) = P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f(x)dx = F(x_2) - F(x_1)$
3. $\int_{-\infty}^{+\infty} f(x)dx = 1$ — limitní případ histogramu
4. $F(x) = \int_{-\infty}^{x} f(t)dt$
5. $F'(x) = f(x)$ v některých bodech ne, pro skoro všechna x (ve všech bodech, ve kterých derivace existuje)



$P(X = x_0) \leq P(x_0 - \epsilon < X < x_0 + \epsilon)$

Otázka, že se trefím do nějakého konkrétního čísla (bodu) je nesmyslná

1.11.2 Střední hodnota

Definice:

Buď X NV spojitého typu. Pak

a) Číslo $\int_{-\infty}^{+\infty} xf(x)dx$ se nazývá *střední hodnotou* veličiny X a značí se $E(X)$ (= expectation).

b) Je-li $h(x)$ nějaká reálná funkce, pak $E(h(x)) = \int_{-\infty}^{+\infty} h(x)f(x)dx$.

c) $D(X) = var(X) = E([X - E(X)]^2)$ je tzv. *rozptyl.*

$$\sigma(X) = \sqrt{D(X)}$$

1.11.3 Obecné a centrální momenty

Definice:

Číslo $\mu'_k \overset{\text{def.}}{=\!=} \int x^k \cdot f(x)dx = E(X^k)$ se nazývá *k-tý obecný moment.*

Číslo $\mu_k \overset{\text{def.}}{=\!=} \int (x - E(X))^k \cdot f(x)dx = E[X - E(X)]^k$ se nazývá *k-tý centrální moment.*

1.11.4 Rozptyl

Věta:

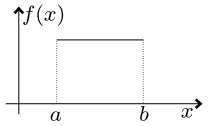
$D(X) = E(X^2) - E^2(X)$

1.11.5 Příklady veličin spojitého typu

I. *Rovnoměrné rozdělení* pravděpodobnosti na intervalu $(a; b)$

$X \sim R(a, b)$

$$\text{hustota } f(x) = \frac{1}{b-a} \qquad \text{pro } x \in (a; b)$$
$$f(x) = 0 \qquad \text{jinde}$$
$$E(X) = \frac{a+b}{2} \qquad D(X) = \frac{(b-a)^2}{12}$$



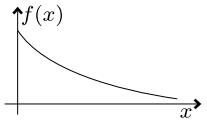
plný nebo prázdný puntík je jedno

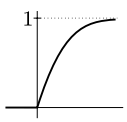
II. *Exponenciální rozdělení* pravděpodobnosti s parametrem $\delta > 0$

$X \sim Exp(\delta)$

$$f(x) = \frac{1}{\delta} e^{-\frac{x}{\delta}} \qquad \text{pro } x > 0$$
$$f(x) = 0 \qquad \text{pro } x < 0$$
$$F(x) = 1 - e^{-\frac{x}{\delta}} \qquad \text{pro } x > 0$$
$$F(x) = 0 \qquad \text{pro } x \leq 0$$
$$F(x) \text{ je spojitá} \qquad \downarrow$$

kam dám rovnítko je jedno





místo δ se někdy píše $\frac{1}{\lambda}$, tj. $\frac{1}{\delta} = \lambda$

$$E(X) = \delta = \frac{1}{\lambda}$$
$$D(X) = \delta^2 = \frac{1}{\lambda^2}$$

(λ . . . intenzita poruch)

Zaokrouhlovací chyba se řídí rovnoměrným rozdělením (čekání na tramvaj).

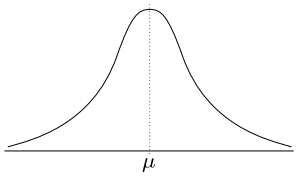
Použití: tento model (rozdělení NV) se používá pro životnost výrobků, které se neopotřebovávají a pro dobu bezporuchového chodu složitých zařízení; příchody prvků do systémů hromadné obsluhy

III. *Normální (Gaussovo) rozdělení* pravděpodobnosti s parametry $\mu, \sigma^2 \quad (\sigma^2 > 0)$

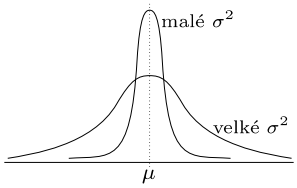
$X \sim N(\mu, \sigma^2)$

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

$$E(X) = \mu \qquad D(X) = \sigma^2$$



Použití: je-li některá veličina součtem velkého množství nezávislých vlivů, pak má přibližně normální rozdělení pravděpodobnosti

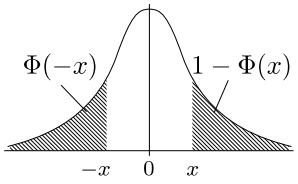


$N(0;1)$ se nazývá *normální normované* (standardizované) rozdělení

distribuční funkce veličiny $N(0;1)$ se značí Φ — bývá tabelována

$$\Phi(x) = \int_{-\infty}^{x} \varphi(t)dt \qquad \text{kde } \varphi(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}$$

$$\Phi(-x) = 1 - \Phi(x) \qquad \text{pro } \forall x$$



Věta:

Nechť $X \sim N(\mu, \sigma^2)$. Pak $\frac{X-\mu}{\sigma} \sim N(0;1)$.

Dk.:

$$P\left(\frac{X-\mu}{\sigma} \le x\right) \stackrel{?}{=} \Phi(x)$$

$$P\left(\frac{X-\mu}{\sigma} \le x\right) = P(X \le \mu + \sigma x) = \int_{-\infty}^{\mu+\sigma x} f(t)dt =$$

$$= \int_{-\infty}^{\mu+\sigma x} \frac{1}{\sqrt{2\pi} \cdot \sigma} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt =$$

$$= \left| \begin{array}{l} \frac{t-\mu}{\sigma} = u \\ \frac{1}{\sigma}dt = du \end{array} \right| = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi} \cdot \sigma} e^{-\frac{1}{2}u^2} du = \Phi(x)$$

1.11.6 Distribuční funkce obecného normálního rozdělení

Věta:

Nechť $X \sim N(\mu, \sigma^2)$. Pak X má distribuční funkci $F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$.

Dk.: $F(x) = P(X \leq x) = P\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) =$ podle předchozí věty $= \Phi(\frac{x-\mu}{\sigma})$.

Poznámka: Je-li $X \sim N(\mu, \sigma^2)$, pak

1. $P(\mu - \sigma < X < \mu + \sigma) \doteq \frac{2}{3}$
2. $P(\mu - 2\sigma < X < \mu + 2\sigma) \doteq 0,95$ (pravidlo 2σ — interval spolehlivosti)
3. $P(\mu - 3\sigma < X < \mu + 3\sigma) \doteq 0,997$

1.11.7 Čebyševova věta

Věta: (Čebyševova)

Buď $X_1, X_2 \ldots$ posloupnost NV, které jsou nezávislé, $E(X_i) = \mu$ pro $i = 1, 2, \ldots$ a existuje K, že $D(X_i) \leq K$ pro $\forall\ i$.

Pak pro libovolné $\epsilon > 0$ platí:

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \epsilon\right) = 0, \text{ tj.}$$

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \leq \epsilon\right) = 1.$$

Lemma:
Pro libovolnou NV X a $\epsilon > 0$ platí:

$$\lim_{n\to\infty} P\left(|X - E(X)| > \epsilon\right) \leq \frac{D(X)}{\epsilon^2}$$

Dk.: nechť X je spojitého typu
označme $Y = X - E(X)$; je $D(Y) \stackrel{?}{=} D(X)$
nechť $f(y)$ je hustota Y

$$P(|Y| > \epsilon) = \int_{-\infty}^{-\epsilon} f(y)dy + \int_{\epsilon}^{\infty} f(y)dy \leq$$

$$\leq \int_{-\infty}^{-\epsilon} \frac{y^2}{\epsilon^2} f(y)dy + \int_{\epsilon}^{\infty} \frac{y^2}{\epsilon^2} f(y)dy \leq \int_{-\infty}^{\infty} \frac{y^2}{\epsilon^2} f(y)dy =$$

$$= \frac{1}{\epsilon^2} \cdot E(Y^2) = \frac{1}{\epsilon^2} \cdot D(Y) \text{ neboť } E(Y) = 0$$

$$D(Y) = D(X)$$

Dk. Čebyševovy věty:

Označme $Y_n = \frac{1}{n}\sum X_i - \mu$.

Je $E(Y_n) = 0,\ D(Y_n) \leq \left(\frac{1}{n}\right)^2 \cdot n \cdot K = \frac{K}{n} \to 0$

$P\left(|Y_n| > \epsilon\right) \leq \frac{D(Y_n)}{\epsilon^2} \leq \frac{1}{\epsilon^2} \cdot \frac{K}{n} \to 0$ pro $n \to \infty$.

1.11.8 Momentová funkce

Definice:

Buď X NV. Funkce $M_x(t) = E(e^{tX})$ se nazývá *momentovou funkcí*.
Funkce $\psi_x(t) = E(e^{itX})$ se nazývá *charakteristickou funkcí* veličiny X.

1.11.9 Necentrální moment

Věta:

a) $\left.\frac{d^k M_x(t)}{dt^k}\right|_{t=0} = \mu'_k$ — k–tý necentrální moment

b) $\left.\frac{d^k \psi_x(t)}{dt^k}\right|_{t=0} = i^k \mu'_k$

Př: Určete charakteristickou funkci rozdělení $N(0;1)$.

$$\psi(t) = E\left(e^{itX}\right) = \int_{-\infty}^{\infty} e^{itx} \cdot f(x)dx = \int_{-\infty}^{\infty} (\cos tx + i \sin tx)\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx =$$

$$= \int_{-\infty}^{\infty} \cos tx \cdot e^{-\frac{1}{2}x^2} dt \qquad (\textstyle\int \text{ podle parametru})$$

$$\frac{d\psi}{dt} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (-\sin tx) x e^{-\frac{1}{2}x^2} dx = \frac{1}{\sqrt{2\pi}} \left[\sin tx \cdot e^{-\frac{1}{2}x^2}\right]_{-\infty}^{+\infty} -$$

$$- \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t \cdot \cos tx \cdot e^{-\frac{1}{2}x^2} dx = -t \cdot \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \cos tx \cdot e^{-\frac{1}{2}x^2} dx = -t \cdot \psi(t)$$

$\frac{d\psi}{dt} = -t \cdot \psi(t)$ je separovatelná dif. rovnice, $\psi(0) = 1$

$\frac{d\psi}{dt} = -t dt \quad \ldots \quad \ln|\psi| = -\frac{1}{2}t^2 \quad \ldots \quad \psi = e^{-\frac{1}{2}t^2}$

Pozor! charakteristická funkce pro $X \sim N(\mu, \sigma^2)$ je $\psi(t) = e^{i\mu - \frac{1}{2}t^2\sigma^2}$

Poznámka: Buďte $X_1, \ldots, X_n$ nezávislé NV, $E(X_i) = \mu$ a $D(X_i) = \sigma_0^2$.

Pak $E(\overline{X}) = \frac{1}{n}(\mu + \ldots + \mu) = \mu$

$D(\overline{X}) = D(\frac{1}{n}\sum X_i) = \left(\frac{1}{n}\right)^2 \cdot \sum D(X_i) = \left(\frac{1}{n}\right)^2 \cdot n \cdot \sigma_0^2 = \frac{\sigma_0^2}{n}$

Označme $X_\Sigma = \sum_{i=1}^{n} X_i$.

Pak $E(X_\Sigma) = n \cdot \mu$

$D(X_\Sigma) = n \cdot \sigma_0^2$

Věta: (centrální limitní (verze Lyndeberg-Léni))

Buď X_n posloupnost nezávislých NV mající stejné rozdělení pravděpodobnosti se střední hodnotou μ a rozptylem σ_0^2 (každé jednotlivé NV). Pak pro posloupnost $Y_n = \sum_{i=1}^{n} X_i$ platí:

$\forall x \in \mathbf{R} \quad \lim_{n\to\infty} P\left(\frac{Y_n - n\mu}{\sqrt{n\sigma_0^2}} \le x\right) = \Phi(x)$ — speciální druh konvergence funkcí, tj.

$\frac{Y_n - n\mu}{\sqrt{n\sigma_0^2}}$ tzv. konverguje k distribuci

Důsledek: Pro velké n je $\sum_{i=1}^{n} X_i \approx N(n\mu, n\sigma_0^2)$, $\overline{X} \approx N\left(\mu, \frac{\sigma_0^2}{n}\right)$.

Dk. (náznak):

$U_j = \frac{X_j - \mu}{\sigma_0} \quad j = 1, 2, \ldots$

$E(U_j) = 0, \quad \sigma(U_j) = 1$

Buď $\psi(t)$ char. funkce U_j a $\psi(t) = E\left(e^{itU_j}\right)$ derivováním slejzá U_j a i

$\psi'(0) = 0$

$\psi''(0) = 1 \cdot i^2 = -1$

$\psi(t) \approx 1 - \frac{1}{2}t^2$ pro $t \to 0$ (Taylorův rozvoj)

$$Z_n = \frac{\sum_{j=1}^{n} U_j}{\sqrt{n}}$$

ψ_n buď char. funkce Z_n

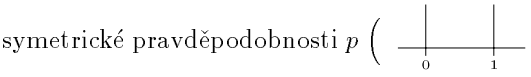
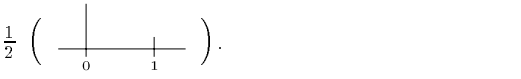
$$\psi_n(t) = E\left(e^{itZ_n}\right) = E\left(e^{\frac{1}{\sqrt{n}}it\sum U_j}\right) = E\left(\prod_{j=1}^{n} e^{\frac{1}{\sqrt{n}}itU_j}\right) =$$

$$\overset{\left(\text{věta že } \frac{E(XY) =}{E(X)E(Y)}\right)}{=\!=\!=} \prod_{j=1}^{n} E\left(e^{\frac{1}{\sqrt{n}}itU_j}\right) = \left[\psi(\frac{1}{\sqrt{n}}\right]^n \approx \left[1 - \frac{1}{2}\left(\frac{t}{\sqrt{n}}\right)^2\right]^n =$$

Aproximace některých NV pomocí normálního rozdělení:

Je-li n velké ($n \ge 30$) a $np(1-p) \le 9$:

$Bi(n,p) \approx N\big(np, np(1-p)\big)$

Konvergence při symetrické pravděpodobnosti p () je rychlejší než konvergence při $p \ll \frac{1}{2}$ ().

Je-li $\lambda \ge 9$, pak $Po(\lambda) \approx N(\lambda, \lambda)$.

Některá další rozdělení NV:

Definice:

NV X má tzv. *logaritmicko–normální rozdělení* pravděpodobnosti s parametry μ, σ^2, jestliže X nabývá jen kladných hodnot a

$$\ln X \sim N(\mu, \sigma^2).$$

Budeme psát $X \sim LN(\mu, \sigma^2)$.

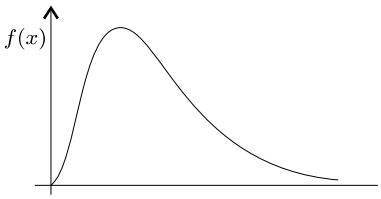
Odvození vzorců pro F, f:

pro $x \leq 0$ je $F(x) = 0$

pro $x > 0$:

$$F(x) = P(X \leq x) = P(\ln X \leq \ln x) = \Phi\left(\frac{\ln x - \mu}{\sigma}\right)$$

$$f(x) = F'(x) = \begin{cases} 0 & \text{pro } x < 0 \\ \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln x - \mu}{\sigma}\right)^2} \cdot \frac{1}{\sigma} \cdot \frac{1}{x} & \text{pro } x > 0 \end{cases}$$



$$E(X) = e^{\mu + \frac{\sigma^2}{2}}$$

Použití: Veličina, která je součinem velkého počtu nezávislých veličin srovnatelné velikosti (žádná není dominantní) má přibližně toto rozdělení.

Gama funkce:

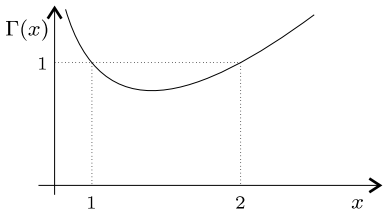
$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt, \quad x > 0$$

$$\Gamma(x+1) = \int_0^\infty t^x e^{-t} dt \overset{\text{per partes}}{=\!=} \left[-t^y e^{-t}\right]_0^\infty + \int_0^\infty x t^{x-1} e^{-t} dt = x\Gamma(x)$$

$\Gamma(1) = 1$

$\Gamma(2) = 1$

$\Gamma(3) = 2 \ldots \Gamma(n+1) = n!$



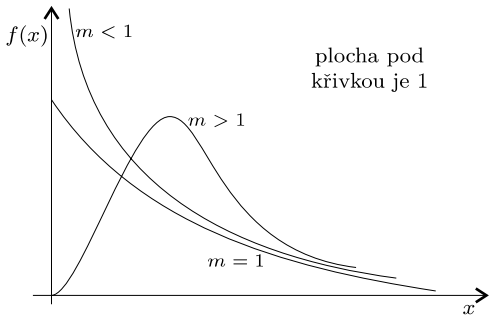
$$\left[\text{např.:} \quad \int_0^{\frac{\pi}{2}} \sin^{r-1} x \cos^{s-1} x dx = \frac{1}{2} \cdot \frac{\Gamma\left(\frac{r}{2}\right) \cdot \Gamma\left(\frac{S}{2}\right)}{\Gamma\left(\frac{r+s}{2}\right)} \quad \ldots \text{ to fakt platí} \right]$$

POZNÁMKA: $\int_0^\infty x^{m-1} \cdot e^{-x} dx = \Gamma(m)$

$$\int_0^\infty \frac{1}{\Gamma(m)} x^{m-1} \cdot e^{-x} dx = 1$$

Definice:

NV X má tzv. *Gama rozdělení* pravděpodobnosti s parametry $m > 0$, $\delta > 0$, jestliže má hustotu pravděpodobnosti $f(x) = 0$ pro $x \leq 0$, $f(x) = \frac{1}{\Gamma(m)} \cdot \frac{1}{\delta^m} \cdot x^{m-1} \cdot e^{-\frac{x}{\delta}}$ pro $x > 0$.



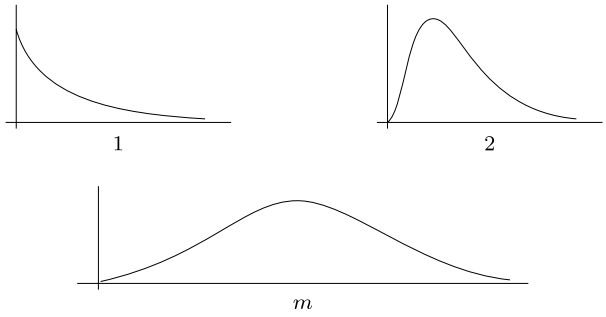
Pro $m = 1$ dostáváme exponenciální rozdělení.

Použití: k teorii spolehlivosti; doba životnosti součástek, kde není opotřebování

Mají-li $X_1, \ldots, X_m$ rozdělení $Exp(\delta)$ a jsou nezávislé, pak $\sum X_i$ má Gama rozdělení s parametry m, δ.

$E(X) = m \cdot \delta$

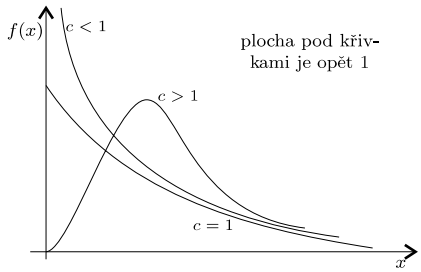
$D(X) = m \cdot \delta^2$



Definice:

NV X má *Weibullovo rozdělení* pravděpodobnosti s parametry $c > 0$, $\delta > 0$, jestliže má

$$F(x) = 0 \qquad \text{pro } x \leq 0$$
$$F(x) = 1 - e^{-\left(\frac{x}{\delta}\right)^c} \qquad \text{pro } x > 0$$
$$f(x) = F'(x) = \frac{c}{x} \cdot \left(\frac{x}{\delta}\right)^c \cdot e^{-\left(\frac{x}{\delta}\right)^c} \text{ pro } x > 0.$$



Pro $c = 1$ dostáváme exponenciální rozdělení.

– lze dokázat, že za určitých podmínek má minimum z n náhodných veličin pro N přibližně Weibullovo rozdělení (pružnost lan a mostů — praskne v minimálním článku?)

– odvodíme střední hodnotu:

$$E(X) = \int_{-\infty}^{\infty} xf(x)\mathrm{d}x = \int_{0}^{\infty} c \cdot \frac{1}{\delta^c} \cdot x^c \cdot e^{-\left(\frac{x}{\delta}\right)^c} \mathrm{d}x = \left|\text{subst. } \frac{x}{\delta} = t, \frac{\mathrm{d}x}{\delta} = \mathrm{d}t\right| =$$

$$= c \cdot \delta \int_{0}^{\infty} t^c e^{-t^c} \mathrm{d}t = \left|\text{subst. } t^c = u, et^{c-1}\mathrm{d}t = \mathrm{d}u\right| = \delta \int_{0}^{\infty} u^{\frac{1}{c}} e^{-u} \mathrm{d}u =$$

$$= \delta \int_{0}^{\infty} u^{\left(\frac{1}{c}+1\right)} e^{-u} \mathrm{d}u = \delta\Gamma\left(\frac{1}{c}+1\right)$$

$$E(X^2) = \int_{0}^{\infty} x^2 f(x)\mathrm{d}x = \ldots = \delta^2\Gamma\left(\frac{2}{c}+1\right)$$

$$D(X) = E(X^2) - E^2(X) = \delta^2\left[\Gamma\left(\frac{2}{c}+1\right) - \Gamma\left(\frac{1}{c}+1\right)\right]$$

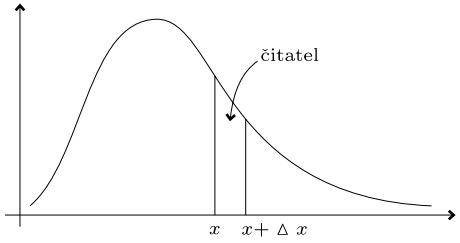
1.11.10 Intenzita poruch

poznámka: Buď $\Delta\, x$ krátký časový interval a spočteme podmíněnou pravděpodobnostporuchy v intervalu $(x; x+\Delta\, x)$ za předpokladu, že v intervalu $(0; x)$ dosud porucha nenastala.

Dobu do poruchy označíme „X“ (NV)

$$P(X < x+\Delta\, x | X > x) = \frac{P(x < X < x+\Delta\, x)}{P(X > x)} = \frac{F(x+\Delta\, x) - F(x)}{1 - F(x)} \approx$$

$$\approx \frac{f(x)\ \Delta\, x}{1 - F(x)} = \frac{f(x)}{1 - F(x)} \cdot \Delta\, x \text{ — přímá úměrnost délky intervalu}$$



Definice:

Uvažujme spojitou NV s hustotou pravděpodobnosti$f(x)$ a distribuční funkcí $F(x)$. Podíl $\lambda(x) = \dfrac{f(x)}{1-F(x)}$ se nazývá *intenzitou poruch*.

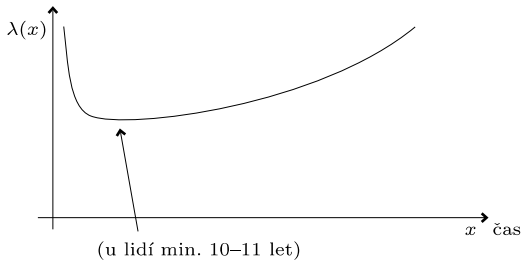
Př:

Nechť $X \sim W(c, \delta)$. Pak $\lambda(x) = \frac{c}{\delta}\left(\frac{x}{\delta}\right)^{c-1} = \frac{c}{\delta^c}x^{c-1}$, $x > 0$.

Pro $\mathbf{c} = 1$ (exp. rozd.) je $\lambda = \frac{1}{\delta}$, tj. *konstantní* intenzita poruch (vhodný model tam, kde se výrobky neopotřebovávají).

Pro $\mathbf{c} > 1$ je in.p. rostoucí, pro $\mathbf{c} < 1$ klesající (Temelín — nejdříve hodně poruch a pak čím dál tím méně).

V praxi mívá často in.p. složitější průběh a obvykle má tvar:



1.11.11 Funkce beta (Eulerův integrál 1. druhu)

$$B(p,q) = \int_0^1 x^{p-1}(1-x)^{q-1}\mathrm{d}x, \quad p > 0, q > 0$$

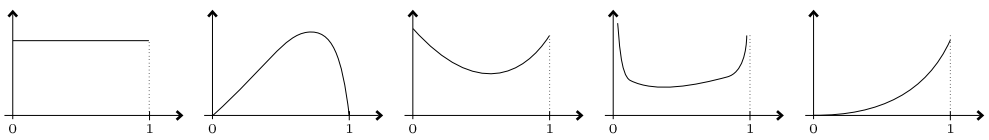
$$B(p,q) = \frac{\Gamma(p)\cdot\Gamma(q)}{\Gamma(p+q)}$$

- *Beta rozdělení pravděpodobnosti* s parametry $p, q > 0$

  $X \sim Be(p,q)$

  $f(x) = \text{konst} \cdot x^{p-1}(1-x)^{q-1}$ pro $x \in (0;1)$

  $f(x) = 0$ pro $x \notin (0;1)$, kde konst $= \dfrac{1}{B(p,q)}$



Používá se jako model pro různé veličiny nabývající hodnot z intervalu $< 0; 1 >$.

$$E(X) = \frac{p}{p+q}$$

$$D(X) = \frac{p\cdot q}{(p+q)^2(p+q+1)}$$

1.11.12 Kvantily spojitých veličin

Definice:

Buď X NV spojitého typu, $p \in (0;1)$. Pak $100p\%$ *kvantil* je číslo x_p takové, že

$$P(X \leq x_p) = p.$$

Tedy $F(x_p) = p$.

Poznámka: $50p\%$ kvantil se nazývá *medián* a značí se $\tilde{x}$

$x_{0,25}$ $\sim$ *dolní kvartil*

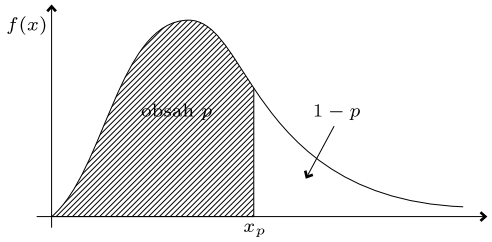
$x_{0,75}$ $\sim$ *horní kvartil*

Př:

Pro $\tilde{x} \sim Exp(\delta)$ spočtěte x_p.

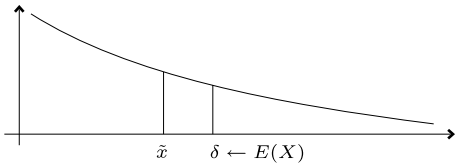
$\Rightarrow\ F(x_p) = p \quad \rightarrow \quad 1 - e^{-\frac{x_p}{\delta}} = p \quad \rightarrow \quad x_p = -\delta \cdot \ln(1-p)$

např. $\tilde{x} = -\delta \cdot \ln \frac{1}{2} = \delta \cot \ln 2 \quad \rightarrow$ medián je nižší než střední hodnota ($< \delta$; rozdíl mezi průměrným platem a tím, kdy polovina lidí má plat pod nějakou hranicí a druhá polovina nad ní)



Poznámka: Kvantily rozdělení $N(0;1>$ jsou tabelovány pro $p \geq 0,5$ a značí se u_p.

Platí $u_p = 1 - u_{1-p}$ pro každé $p \in (0;1)$. Např. $u_{0,05} = -u_{0,95} \doteq 1,645$.



Věta:

Pro $X \sim N(\mu, \sigma^2)$ je $x_p = \mu + \sigma \cdot u_p$.

Důkaz:

$$F(x_p) = p \quad \leftarrow \text{ chci} \qquad \Phi\left(\frac{x_p - \mu}{\sigma}\right) = p$$

$$\frac{x_p - \mu}{\sigma} = u_p \quad \rightarrow \quad x_p = \mu + \sigma \cdot u_p$$

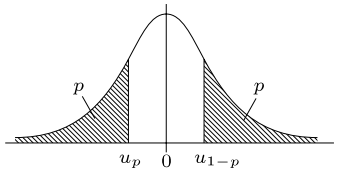
Př:

Za předpokladu, že pro rozměr určitých výrobků platí:

$$X \approx N(\mu, \sigma^2), \text{ kde } \mu = 15 \text{ a } \sigma = 0,2.$$

Nalezněte c takové, aby přibližně 95% výrobků v budoucnu vyráběných mělo rozměr větší než c.

$$\Rightarrow \quad c \approx x_{0,05} = \mu + \sigma \cdot u_{0,95} = \mu - 1,645\sigma \doteq 14,67$$



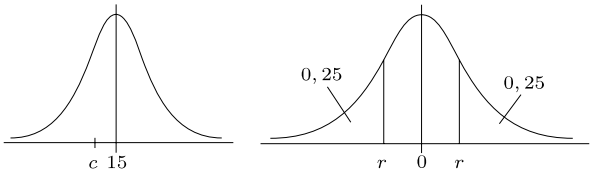
Poznámka: V kartografii se používá termín „pravděpodobné chyba“.

Číslo r se nazývá p.ch. pro NV X, jestliže $P(|X| < r) = \frac{1}{2}$.

$$|X| < r \Longleftrightarrow X \in (-r; r)$$
$$p(x) \in (-r; r) = \frac{1}{2}$$
$$P(X > r) = \frac{1}{4}$$
$$P(X \leq r) = 0,75$$

Pro $X \sim N(0; \sigma^2)$ nalezněme r:

$$r = x_{0,75} = \mu + \sigma \cdot u_{0,75} = \sigma \cdot 0,674$$



2 Statistika

2.1 Teorie odhadu

Předpokládejme, že NV X má nějaké rozdělení pravděpodobnosti, které známe až na hodnotu jednoho nebo několika parametrů. Je změřeno n hodnot a chceme pomocí nich odhadnout neznámé parametry.

Definice:

NVy $X_1, \ldots, X_n$ se nazývají *nezávislé*, jestliže pro všechna reálná čísla $c_1, \ldots, c_n$ jsou jevy $X_1 \leq c_1, \ldots, X_n \leq c_n$ nezávislé.

Definice:

Buď X NV a $X_1, \ldots, X_n$ posloupnost nezávislých NV takových, že všechny mají stejné rozdělení pravděpodobnostijako NV X. Pak posloupnost $X_1, \ldots, X_n$ se nazývá *náhodným výběrem v rozsahu n z rozdělení NV X*.

Definice:

Buďte $X_1, \ldots, X_n$ NV a g nějaká funkce n proměnných. Pak NV $g(X_1, \ldots, X_n)$ se nazývá *statistika*.

Př.: $\overbrace{\frac{1}{n}\sum_{i=1}^{n} X_i}^{\text{aritm. } \varnothing}$ je statistika

2.1.1 Bodové odhady

Buď $X_1, \ldots, X_n$ náhodný výběr NV X a θ nechť je nějaký papametr (např. μ, rozptyl). Statistika $g(X_1, \ldots, X_n)$ se nazývá *konzistentním odhadem parametru* θ, jestliže pro $\forall\ \epsilon > 0$ je $\lim_{n\to\infty} P(|g(X_1, \ldots, X_n) - \theta| < \epsilon) = 1$ (statistika konverguje k θ když $n \to \infty$).

Př:

$D(X) < \infty$, pak $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, kde $X_1, \ldots, X_n$ je náhodný výběr, je konzistentním odhadem parametru $\theta = E(X)$. Důkaz viz Čebyševova věta.

Definice:

Buď $X_1, \ldots, X_n$ náhodný výběr z rozdělení NV X a θ buď nějaký parametr. Statistika $g(X_1, \ldots, X_n)$ se nazývá *nestranným (nevychýleným) odhadem parametru* θ, jestliže pro všechny přípustné hodnoty θ platí:

$$E(g) = \theta$$

Definice:

Statistika g se nazývá *asymptoticky nestranným odhadem* θ, jestliže $\lim_{n\to\infty} E(g) = \theta$.

Př:

Je-li $X_1, \ldots, X_n$ náhodný výběr s rozdělením NV X, pak statistika $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$ je nestranným odhadem parametru $\theta = E(X)$.

neboť:

$$E(\overline{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} \cdot n \cdot E(X) = E(X)$$

parametr μ se odhaduje aritmetickým průměrem

$\sim$ δ $\sim$

$\sim$ λ $\sim$

Tvrzení: Buď $X_1, \ldots, X_n$ náhodný výběr z rozdělení NV X, která má rozptyl $D(X)$ $[\sigma^2]$. Pak platí:

$$E\left(\sum_{i=1}^{n}[X_i - \overline{X}]^2\right) = (n-1) \cdot \sigma^2$$

Dk. označme $\mu = E(X)$

Platí:

$$\begin{aligned}\sum_{i=1}^{n}[X_i - \overline{X}]^2 &= \sum_{i=1}^{n}\Big[(X_i - \mu) + (\mu - \overline{X})\Big]^2 = \\ &= \sum_{i=1}^{n}(X_i - \mu)^2 + n(\mu - \overline{X})^2 + 2(\mu - \overline{X})\sum_{i=1}^{n}(X_i - \mu) = \\ &= \sum_{i=1}^{n}(X_i - \mu)^2 + n(\mu - \overline{X})^2 + 2n(\mu - \overline{X})(\overline{X} - \mu) = \\ &= \sum_{i=1}^{n}(X_i - \mu)^2 - n(\mu - \overline{X})^2\end{aligned}$$

$$E\left(\sum_{i=1}^{n}[X_i - \overline{X}]^2\right) = n \cdot \sigma^2 = n \cdot D(\overline{X}) = n \cdot \sigma^2 - n \cdot \frac{\sigma^2}{n} = (n-1) \cdot \sigma^2$$

Důsledek: Statistika $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ je nestranným odhadem rozptylu veličiny X.

Statistika $M_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ je vychýleným ($\neg$ nestranným) odhadem hodnoty $D(X)$, který je však asymptoticky nestranný (podhodnocuje), tj.

$$\lim_{n\to\infty} E(M_2) = D(X)$$

neboť

$$E(M_2) = E\left(\frac{n-1}{n} \cdot S^2\right) = \frac{n-1}{n} \cdot E(S^2) = \frac{n-1}{n} \cdot D(X) = \xrightarrow{\text{konverguje}} 1 \cdot D(X)$$

2.1.2 Bodový odhad parametrů metodou maximální věrohodnosti

angl. MLE = *maximum likehood estimators* (= odhadovač) [estimate = odhad, číslo]

Uvažujme NV X s $f(x,\theta)$ resp. pravděpodobnostní funkcí $\mathcal{P}(x,\theta)$, kde θ je neznámý parametr nebo vektor neznámých parametrů. Nechť $x_1, x_2, \ldots, x_n$ je realizace náhodného výběru s rozdělením NV X. Funkce

$$L(\theta) = \prod_{i=1}^{n} f(x_i,\theta), \quad \text{resp.} \quad L(\theta) = \prod_{i=1}^{n} \mathcal{P}(x_i,\theta)$$

se nazývá *věrohodnostní funkcí*.

Je-li $\hat{\theta}$ takové, že $L(\hat{\theta}) \geqq L(\theta)$ pro všechna přípustná θ, pak $\hat{\theta}$ se nazývá odhadem θ metodou max. věrohodnosti. V praxi je obvykle výhodnější pracovat s $\ln L(\theta)$.

Příklady: Pro $X \sim N(\mu,\sigma^2)$ jsou MLE odhady $\hat{\mu} = \hat{x}$, $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

Pro $X \sim Exp(\delta)$ je MLE odhad $\hat{\delta} = \hat{x}$.

Pro $X \sim Po(\lambda)$ je $\sim \hat{\lambda} = \hat{x}$.

Pro $X \sim LN(\mu,\sigma^2)$ je $\sim \hat{\mu} = \overline{\ln x} = \frac{1}{n}\sum_{i=1}^{n} \ln x_i$, $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\ln x_i - \overline{\ln x}\right)^2$.

Pro Weibulla vychází netriviální řešení.

POZNÁMKA: $\frac{\mathrm{d}}{\mathrm{d}\theta}\ln L(\theta)$, resp. $\frac{\partial}{\partial\theta_i}\ln L(\theta) = 0$ pro $i = 1,\ldots,k$, je-li $\theta = (\theta_1,\ldots,\theta_k)$, se nazývají věrohodnostní rovnice.

Za poměrně velmi obecných předpokladů mají věrohodnostní rovnice právě jedno řešení $(\hat{\theta})$ a $\hat{\theta}$ je asymptoticky nestranný odhad θ a jde o konzistentní odhad θ, řád konvergence je $\frac{\text{konst.}}{\sqrt{n}}$ (analyticky $\frac{|a_n - a|}{\frac{c}{\sqrt{n}}} \to 1$), tj. pro velká n $|\hat{\theta} - \theta| \approx \frac{\text{konst.}}{\sqrt{n}}$.

Za určitých předpokladů platí, že MLE odhady mají mezi všemi asymptoticky nestrannými odhady nejnižší rozptyl.

2.1.3 Bodové odhady parametrů metodou momentů

Př:

Chtějme odhadnout parametry c, δ Weibullova rozdělení. Máme realizaci $x_1,\ldots,x_n$ náhodného výběru. Víme, že

$$E(X) = \Gamma\left(1+\tfrac{1}{c}\right)\delta$$
$$D(X) = \left[\Gamma\left(1+\tfrac{2}{c}\right) - \Gamma^2\left(1+\tfrac{1}{c}\right)\right]\delta^2$$

Z rovnic $E(X) = \bar{x}$

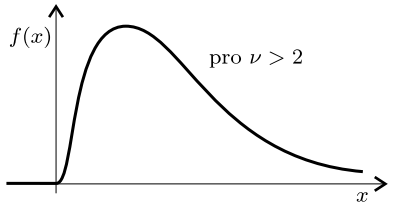
$D(X) = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ spočteme odhady parametrů c, δ.

$$\frac{D(X)}{E^2(X)} = \frac{\Gamma\left(1+\frac{2}{c}\right)}{\Gamma^2\left(1+\frac{1}{c}\right)} - 1 = \frac{s^2}{\bar{x}^2}$$

Příklady dalších rozdělení:

I. Rozdělení $\chi^2(\nu)$, kde ν je celočíselný parametr, tzv. počet stupňů volnosti. Toto rozdělení je speciálním případem Γ–rozdělení pro $m = \frac{\nu}{2}$, $\delta = 2$.

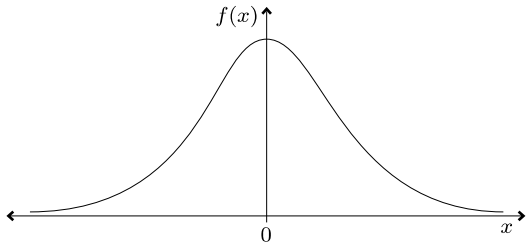
$$E(X) = \nu$$
$$D(X) = 2\nu$$



Pro velké ν je $\chi^2(\nu) \approx N(\nu, 2\nu)$.

II. Rozdělení $t(\nu)$ (*Studentovo*) s parametrem ν, který je stejný jako u rozd. χ^2

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \cdot \frac{1}{\sqrt{\pi \cdot \nu}} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \quad x \in (-\infty; +\infty)$$



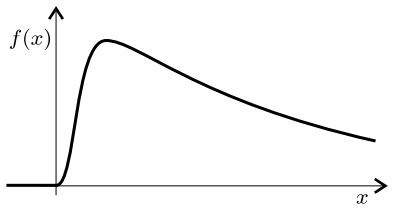
Pro velké ν je $t(\nu) \approx N(0; 1)$.

Pro $\nu = 1$ je $f(x) = \frac{1}{\sqrt{\pi}} \cdot \frac{1}{\sqrt{\pi}}(1 + x^2)^{-1}$, $x \in \mathbf{R}$ — Cauchyho rozdělení.

Pro Cauchyho rozdělení je $E(X) = \int\limits_{-\infty}^{\infty} x f(x)\mathrm{d}x = \frac{1}{2\pi} \int\limits_{-\infty}^{\infty} \frac{2x}{1+x^2}\mathrm{d}x = \frac{1}{2\pi}\left[\ln(1 + x^2)\right]_{-\infty}^{\infty}$ neexistuje.

III. Rozdělení $F(\nu_1, \nu_2)$ (*F–rozdělení, Fisherovo, Fisherovo–Snedeckorovo*), $\nu_1, \nu_2, \ldots$

$f(x)$ neuvedeme, jelikož je příliš složitá



$$F(\nu_1, \nu_2) \neq F(\nu_2, \nu_1)$$

Věta: (1)

Nechť $X_1, \ldots, X_n \sim N(0;1)$ a jsou nezávislé. Pak

$$X_1^2 + \ldots + X_n^2 \sim \chi^2(\nu), \text{ kde } \nu = n.$$

Věta: (2)

Nechť $X \sim N(0;1)$, $Y \sim \chi^2(\nu)$ a X, Y nechť jsou nezávislé. Pak

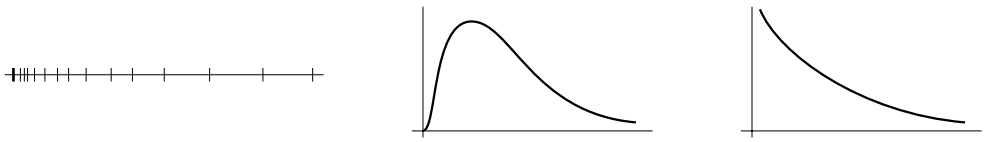
$$\frac{X}{\sqrt{\frac{Y}{\nu}}} \sim t(\nu). \qquad \text{— intervalové odhady } E(X)$$

Věta: (3)

Nechť X, Y jsou nezávislé, $X \sim \chi^2(\nu_1)$, $Y \sim \chi^2(\nu_2)$. Pak

$$\frac{\frac{X}{\nu_1}}{\frac{Y}{\nu_2}} \sim F(\nu_1, \nu_2). \qquad \text{— např. testování 2 } D(X), \text{ podělím, pak je rozdíl v šířce zvonu}$$

– pro malá n se těžko rozlišují např. exponenciální a LN



pro počet hodnot kolem 10 je chybovost značná (asi 40 %), vyhrává neprávem LN, protože má 2 parametry a tím se lépe přizpůsobí

→ zavedeme jakousi pokutu $\quad L_1 \overset{?}{\sim} L_2$ – pokuta(n)

$E(1)$

$LN(0, \sigma^2)$ – pokuta závisí pouze na σ^2 a při vhodném zvolení σ^2 je ta pokuta až 100 % úspěšná, ale to neznám, vlastně nevím jak tu pokutu volit – vhodně odhadnout σ

— Exp a LN jsou celkem jednoduchá; pro Weib. už je to značně náročné

2.1.4 Intervalové odhady parametrů

Definice:

Buď $0 < p < 1$. Pak $100p\%$ intervalem spolehlivosti pro parametr θ je takový interval (α, β), pro který platí $P(\alpha < \theta < \beta) = p$.

Intervaly spolehlivosti pro střední hodnotu

Nechť X je NV se střední hodnotou μ a rozptylem σ^2. Nechť σ^2 je známé. Mějme náhodný výběr $X_1, \ldots, X_n$.

Pro velké n je

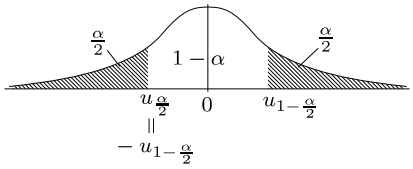
$$\overline{X} \approx N\big(\mu, \frac{\sigma^2}{n}\big)$$

$$\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \approx N(0,1)$$

$$\frac{\mu-\overline{X}}{\frac{\sigma}{\sqrt{n}}} \approx N(0,1) \iff N(0,1) \text{ je symetrické kolem osy } y$$

S pravděpodobností přibližně $1-\alpha$ je

$$u_{\frac{\alpha}{2}} < \frac{\mu-\overline{X}}{\sigma} \cdot \sqrt{n} < u_{1-\frac{\alpha}{2}}$$



$$\overline{X} - u_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + u_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \qquad (*)$$

$\rightarrow$ pro velká n jde o přibližný (použití aproximace) $100(1-\alpha)\%$ oboustranný (žádné $-\infty$ ani $+\infty$) interval spolehlivosti pro parametr μ — musím znát σ

Případ s neznámým σ^2:

Jde-li o výběr z $N(\mu,\sigma^2)$, pak $\frac{\overline{X}-\mu}{s} \cdot \sqrt{n} \approx t(n-1)$. Oboustranný $100(1-\alpha)\%$ interval spolehlivosti pro μ je

$$\overline{X} - t_{1-\frac{\alpha}{2}}(n-1) \cdot \frac{s}{\sqrt{n}} < \mu < \overline{X} + t_{1-\frac{\alpha}{2}}(n-1) \cdot \frac{s}{\sqrt{n}} \qquad (**)$$

– pro n velké to konverguje k $(*)$

Pokud rozdělení není normální, platí výše uvedený interval ($(**)$ pro $100(1-\alpha)\%$) přibližně pro velká n.

Dále se odhaduje σ^2.

Přibližné intervaly spolehlivosti pro parametr p alternativního rozdělení lze pro velká n počítat následovně:

označme m počet jedniček ve výběru $x_1, \ldots, x_n$ z $A(p)$

$\hat{p} = \frac{m}{n}$ je bodový odhad p

nechť $n\hat{p}(1-\hat{p}) \geq 9$

Pak m je realizací NV s rozdělením $Bi(n,p) \approx N\big(np, np(1-p)\big)$

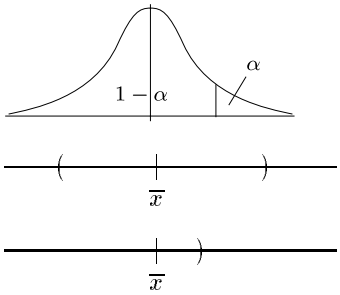
$$\frac{np-m}{\sqrt{np(1-p)}} \approx N(0;1)$$

Platí, že

$$\frac{np - m}{\sqrt{np(1-p)}} \approx N(0;1).$$

– odtud se snadno získají přibližné intervaly spolehlivosti

Jmenovatel je méně náchylný ke změně p než čitatel; jinak by to byla nepěkná středoškolská matematika.



... pouze horní mez – kolik maximálně zmetků → jednostranné intervaly spolehlivosti

– zde mám hodnotu zprava přesnější, ale oželil jsem druhou mez

2.2 Testování hypotéz

Ověřujme určitou hypotézu týkající se nějaké NV X na základě náhodného výběru o rozsahu n. Můžeme se při tom dopustit chyby dvojího druhu:

	Hypotézu *přijmeme*	Hypotézu *nepřijmeme*
Hypotéza *platí*	O.K.	chyba ***1. druhu***
Hypotéza *neplatí*	chyba ***2. druhu***	O.K.

Definice:

Pravděpodobnost chyby 1. druhu se nazývá *hladinou významnosti testu* a značí se α.

Říkáme, že test provádíme na hladině významnosti α.

Poznámka: Nejčastěji se používá $\alpha = 0{,}05$, popřípadě $\alpha = 0{,}01$.

Formální postup při testování hypotézy:

1) zformulujeme hypotézu ve tvaru, ve kterém se předpokládá platnost určitého modelu
2) zvolíme vhodnou statistiku (funkci naměřených hodnot)
3) zvolíme vhodnou hladinu významnosti testu α ($0{,}05 \times 0{,}01$) a kritický obor pro uvažovanou statistiku takový, aby pravděpodobnost, že uvažovaná statistika za platnosti hypotézy padla do kritického bodu s pravděpodobností α (nebo $\leq \alpha$, jen o málo)
4) na základě naměřených hodnot zjistíme hodnotu statistiky

   leží-li v kritickém oboru, hypotézu zamítneme (na hladině významnosti α)

   neleží-li v kritickém oboru, hypotézu nezamítáme (neprokázal jsem ji)

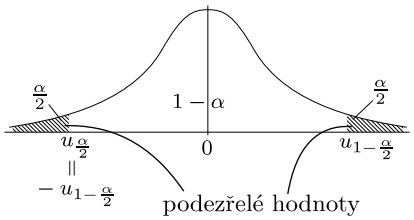
2.2.1 Testy o střední hodnotě

Buď $x_1, \ldots, x_n$ realizace náhodného výběru z $N(\mu, \sigma^2)$, kde σ^2 je známé. Chtějme testovat hypotézu

$$H_0\colon \mu = \mu_0$$

Využijeme toho, že

$$\frac{\overline{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0;1), \text{ tj. } u = \frac{\mu-\overline{X}}{\sigma}\cdot\sqrt{n} \sim N(0;1).$$



Kritický obor pro statistiku $u = (\mu_0 - \overline{X})\cdot\frac{1}{\sigma}\sqrt{n}$ je obor

$$|u| > u_{1-\frac{\alpha}{2}}.$$

Dosadím hodnoty a mohu rozhodnout.

Pokud není σ známé, používá se statistika

$$t = \frac{\mu_0-\overline{X}}{s}\cdot\sqrt{n}$$
$$\text{krit. obor: } |t| > t_{1-\frac{\alpha}{2}}(n-1)$$

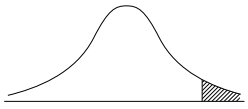
– hovoří se o tzv. jednovýběrovém t–testu, přesněji jde o oboustranný t–test

Je možno konstruovat i jednostranné testy.

2.2.2 p–hodnota testu (p–value)

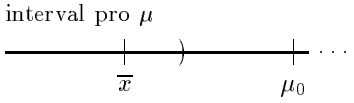
Při použití SW se zpravidla nemusí předem zadávat hladina významnosti testu, program vyčíslí tzv. p–hodnotu. Hypotézu lze zamítnout na hladině významnosti $\alpha \iff p$–hodnota $\leq \alpha$ (rozhoduje se, zda chceme jednostranný či oboustranný).

Poznámka: Týká-li se hypotéza hodnoty nějakého parametru θ, lze ji testovat ekvivalentně pomocí intervalů spolehlivosti. Hypotézu $\theta = \theta_0$ zamítneme na hladině významnosti α, jestliže θ_0 nepatří do $100(1-\alpha)\%$ intervalu spolehlivosti pro parametr θ.



$\frac{\mu_0-\overline{X}}{\cdots}$

$\overline{X} \ll \mu_0$ jsou podezřelé



Testy hypotéz o rozptylu σ^2 rozdělení $N(\mu,\sigma^2)$ se provádějí pomocí tzv. F–testu.

2.2.3 χ^2–test dobré shody

Předpokládejme hypotézu, že nějaká NV X se řídí určitým modelem, který známe až na hodnoty několika parametrů, počet neznámých parametrů označíme m $(m \geq 0)$.

Mějme náhodný výběr $x_1, \ldots, x_n$.

Obor hodnot NV X rozdělme na několik intervalů, jejich počet označme k. Označme $n_i (i = 1, \ldots, k)$ počet naměřených hodnot v i–tém intervalu (je $n_1 + n_2 + \ldots + n_k = n$).

Pro každý interval spočteme tzv. očekávané počty $o_i (i = 1, \ldots, k)$, $o_i = n \cdot p_i$, kde p_i je pravděpodobnost, že NV X padne do i–tého intervalu.

Pro velká n má součet

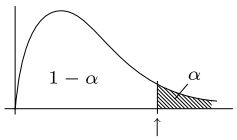
$$S = \sum_{i=1}^{k} \frac{(o_i - n_i)^2}{o_i}$$

přibližně rozdělení $\chi^2(\nu)$, kde $\nu = k - 1 - m$.

Za dostatečně velké n se považuje takové, že $o_i \geq 5$ pro $\forall\ i = 1, \ldots, k$. Za kritické hodnoty se považují velké hodnoty S.

Hypotézu o platnosti daného modelu zamítáme na hladině významnosti α, jestliže

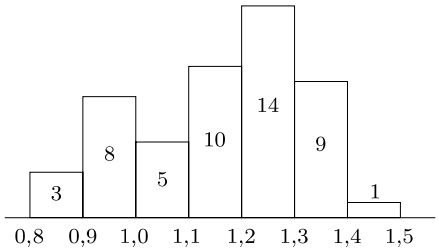
$$S > \chi^2_{1-\alpha}(\nu).$$



Poznámka: Jestliže při výpočtu hodnot o_i zjistíme, že $o_i < 5$ pro některou třídu, lze sdružit některé třídy a sečíst příslušná o_i.

Př.:

Bylo naměřeno 50 hodnot, $\overline{x} = 1{,}16$ a $s = 0{,}158$. Máme dále k dispozici následující histogram četností:



Testujte hypotézu, že NV má normální rozdělení.

Řešení: parametry μ, σ odhadneme pomocí $\overline{x}, s$, tj. $m = 2$.

Třídy jsou:

	n_i	p_i	o_i	
$(-\infty; 0{,}8)$	0	0,012	0,6	$n_i = 11$
$(0{,}8; 0{,}9)$	3	0,038	1,9	$o_i = 8$
$(0{,}9; 1{,}0)$	8	0,11	5,5	
$(1{,}0; 1{,}1)$	5	0,19	9,5	
$(1{,}1; 1{,}2)$	10	0,25	14,5	
$(1{,}2; 1{,}3)$	14	0,21	10,5	
$(1{,}3; 1{,}4)$	9	0,13	6,5	$n_i = 10$
$(1{,}4; 1{,}5)$	1	0,045	2,2	$o_i = 9{,}5$
$(1{,}5; +\infty)$	0	0,016	0,8	

$$P(X < b) = \Phi\left(\tfrac{b-\overline{x}}{s}\right)$$
$$P(a < X < b) = \Phi\left(\tfrac{b-\overline{x}}{s}\right) - \Phi\left(\tfrac{a-\overline{x}}{s}\right)$$
$$P(X > b) = 1 - \Phi\left(\tfrac{a-\overline{x}}{s}\right)$$

$$S = \frac{(11-8)^2}{8} + \frac{(5-9{,}5)^2}{9{,}5} + \ldots + \frac{(10-9{,}5)^2}{9{,}5} \doteq 4{,}995$$

$$\nu = 5 - 1 - 2 = 2$$

Volme $\alpha = 0{,}05$.

$\chi^2_{0,95}(2) = 5{,}99$

$S < \chi^2_{0,95}(2) \Rightarrow$ hypotézu o normálním rozdělení veličiny nezamítáme

2.2.4 Test nezávislosti dvou veličin pomocí kontingenční tabulky

Uvažujme diskrétní nebo spojité NV X, Y. Chceme testovat jejich nezávislost. Jde-li o spojité NV, rozdělíme jejich obor do tříd. Mějme n dvojic (X, Y), každá dvojice nechť padne do jednoho pole dvourozměrného rozdělení četností. Příslušná tabulka četností se nazývá *kontingenční tabulka*.

Př.:

známky z MA & FYA		
1,3	3,3	3,1
3,1	3,3	3,2
2,2	3,2	3,2
3,1	3,3	2,3

	1	2	3
1			|
2		|	|
3	|||	|||	||||

Mějme např. (pro $n = 50$) kontingenční tabulku:

1	0	5	6
1	2	5	8
10	13	13	36
12	15	23	50

Četnost v i–tém řádku $(i = 1, \ldots, I)$ a j–tém sloupci $(j = 1, \ldots, J)$ označme n_{ij}.

Označme (tradiční značení):

$$n_{i\bullet} = \sum_{j=1}^{J} n_{ij}$$

$$n_{\bullet j} = \sum_{i=1}^{I} n_{ij}$$

$$n = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}$$

Za předpokladu nezávislosti NV X, Y jsou očekávané počty

$$o_{ij} = \frac{n_{i\bullet}}{n} \cdot \frac{n_{\bullet j}}{n} \cdot n = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n}.$$

Označme

$$S = \sum_{i,j} \frac{(o_{ij} - n_{ij})^2}{o_{ij}}.$$

Pro velká n má součet S za předpokladu nezávislosti X, Y přibližně rozdělení $\chi^2(\nu)$, kde

$$\nu = (I-1)(J-1)$$

Hypotézu o nezávislosti NV X, Y zamítneme na hladině významnosti α, jestliže

$$S > \chi^2_{1-\alpha}(\nu).$$

Za dostatečně velké n se považuje takové, že $o_{ij} \geq 5$ pro $\forall\ i, j$. Není-li tato podmínka splněna, je možno sdružit některé řádky, resp. sloupce.

$$o_{11} = \frac{6 \cdot 12}{50} = \frac{72}{50} = 1{,}44 \qquad o_{21} = \frac{96}{50} = 1{,}92 \qquad o_{31} = 8{,}64$$
$$o_{12} = \frac{6 \cdot 15}{50} = \frac{90}{50} = 1{,}8 \qquad o_{22} = \frac{120}{50} = 2{,}4$$
$$o_{13} = \frac{6 \cdot 23}{50} = \frac{138}{50} = 2{,}76 \qquad o_{23} = \frac{184}{50} = 3{,}68$$

o_{ij}

↓	1,40	1,80	2,76	6
↑	1,92	2,40	3,68	8
	8,64	10,80	16,56	36
	12	15	23	

$\longrightarrow$

	1	2	3
M 1 ∨ 2	3,86	4,2	6,44
M 3	8,64	10,8	16,56

$\longrightarrow$

$\longrightarrow$ o_{ij}

	F 1 ∨ 2	F 3
M 1 ∨ 2	7,56	6,44
M 3	19,44	16,56

n_{ij}

4	10
23	13

$$S = \frac{(7{,}56-4)^2}{7{,}56} + \frac{(6{,}44-10)^2}{6{,}44} + \frac{(19{,}44-23)^2}{19{,}44} + \frac{(16{,}56-13)^2}{16{,}56} = \ldots S \in (5;6)$$

$\alpha = 0{,}05$

$\chi^2_{0,95}(1) = 3{,}84 \quad \Rightarrow S > \chi^2_{0,95}(1) \Rightarrow$

hypotézu o nezávislosti známek z MA a FYA zamítáme na hladině významnosti 0,05

2.2.5 Síla testu, silofunkce

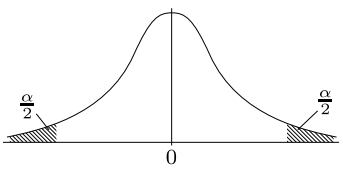
Pravděpodobnost chyby 2. druhu se značí β a závisí na n a na tom, jak je vzdálena skutečnost od hypotézy.

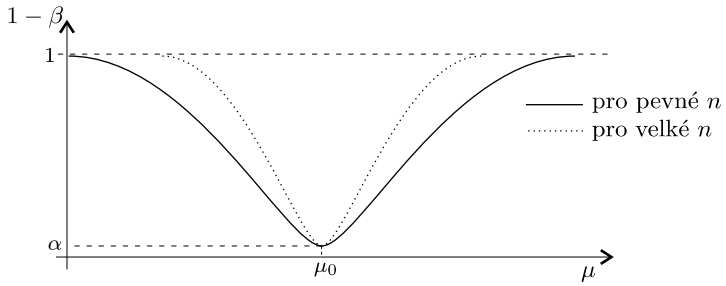
$1 - \beta$ = pravděpodobnost zamítnutí hypotézy, která neplatí

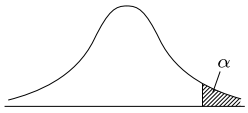
$1 - \beta$ se nazývá sílou testu, a protože závisí na skutečné hodnotě parametru nazývá se tato závislost *silofunkcí*.

Uvažujme např. test $\mu = \mu_0$ o parametru $N(\mu, \sigma^2)$.

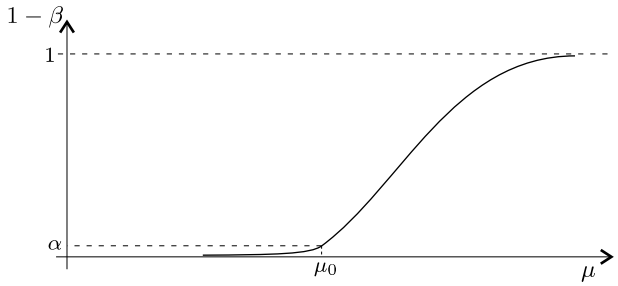
Uvažujme oboustrannou alternativu, tj. oboustranný kritický obor statistiky $t = \frac{\overline{X} - \mu_0}{s}\sqrt{n}$.





Volme jednostranný test 

.



2.2.6 Testy shody dvou středních hodnot

a) Případ dvou nazávislých výběrů.

Jsou-li $x_1, \ldots, x_n$ a $y_1, \ldots, y_n$ dva nezávislé výběry ze dvou rozdělení a chceme testovat hypotézu shody teoretických středních hodnot $\mu_x = \mu_y$, pak pro velká n, m je možno použít statistiku

$$t = \frac{|\overline{x} - \overline{y}|}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{n}}}$$

s kritickým oborem

$$|t| > u_{1-\frac{\alpha}{2}}$$

pro veliké n a m, a pak ji zamítám.

Zdůvodnění:

$$\overline{x} \approx N\big(\mu_x, \frac{\sigma_x^2}{n}\big) \approx N\big(\mu_x, \frac{s_x^2}{n}\big)$$

$$\overline{y} \approx N\big(\mu_y, \frac{\sigma_y^2}{m}\big) \approx N\big(\mu_y, \frac{s_y^2}{m}\big)$$

$$\overline{x} - \overline{y} \approx N\big(\mu_x - \mu_y, \frac{s_x^2}{n} + \frac{s_y^2}{m}\big) \quad \text{[příčítám } -\text{násobek, proto se rozptyly sčítají]}$$

Nejsou-li n, m dostatečně velká, je vhodnější použít kritické obory

$$t_{1-\frac{\alpha}{2}}(\nu), \text{ kde } \nu = \min\{n, m\} - 1.$$

b) Případ dvou obecně závislých výběrů (tzv. *párový test*).

Nechť $x_1, \ldots, x_n$ a $y_1, \ldots, y_n$ jsou takové, že $(x_i, y_i), i = 1, \ldots, n$, jsou náhodným výběrem dvourozměrné veličiny (X, Y). Pak pro test hypotézy $\mu_x = \mu_y$ lze použít následující postup:

Spočteme $z_i = x_i - y_i (i = 1, \ldots, n)$ a pomocí statistiky

$$t = \frac{\overline{z}}{s_z} \cdot \sqrt{n}$$

testujeme hypotézu, že jde o náhodný výběr z rozdělení s nulovou střední hodnotou.

Poznámka: Test shody dvou rozptylů se provádí pomocí F–testu používajícího statistiku $F = \frac{s_x^2}{s_y^2} \cdot \text{konst.}$

2.3 Náhodný vektor (vícerozměrná NV)

$\vec{X} = (X_1, X_2, \ldots, X_k)$

Definice:

Distribuční funkce $\overrightarrow{\text{NV}}$ (náh. vektor) $\vec{X}$ je definována

$$F(x_1, \ldots, x_k) = P(X_1 \leq x_1 \wedge X_2 \leq x_2 \wedge \ldots \wedge X_k \leq x_k).$$

Definice:

Nechť $\vec{X}$ je $\overrightarrow{\text{NV}}$, který může nabývat pouze hodnot $\vec{c_1}, \vec{c_2}, \ldots$ Pak pravděpodobnostní funkce je definována následovně:

$$\mathcal{P}(\vec{x}) = P(\vec{X} = \vec{x}).$$

Poznámka: $\mathcal{P}(\vec{x}) = 0$ pro $\vec{x} \notin \{\vec{c_1}, \vec{c_2}, \ldots\}$
$\mathcal{P}(\vec{c_1}) + \mathcal{P}(\vec{c_2}) + \ldots = 1$

Definice:

Pro $\overrightarrow{\text{NV}}$ $\vec{X}$ spojitého typu se definuje hustota pravděpodobnosti $f(x_1, \ldots, x_k)$. Jde o funkci, pro kterou platí

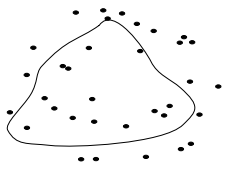
$$F(x_1, \ldots, x_k) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \ldots \int_{-\infty}^{x_k} f(t_1, t_2, \ldots, t_k)\, \mathrm{d}t_1 \mathrm{d}t_2 \ldots \mathrm{d}t_k.$$

Věta:

Buď $\vec{X}$ $\overrightarrow{\text{NV}}$ a $A \subset \mathbf{R}^k$.

a) Je-li $\vec{X}$ diskrétního typu, pak

$$P(\vec{X} \in A) = \sum_{\vec{x} \in A} \mathcal{P}(\vec{x})$$



b) Je-li $\vec{X}$ spojitého typu, pak

$$P(\vec{X} \in A) = \int_A f(\vec{x}) \, \mathrm{d}\vec{x} = \int_{\mathbf{R}^2} \cdots \int$$

Poznámka: Výše uvedená distribuční funkce $F(\vec{x})$, pravděpodobnostní funkce $\mathcal{P}(\vec{x})$ a hustota $f(\vec{x})$ se často uvádějí s přívlastkem *sdružená*...

Jednorozměrné distribuční funkce atd. jednotlivých složek $X_i, i = 1, \ldots, k$, se uvádějí s přívlastkem *marginální*...

A indexují se — $F_i, \mathcal{P}_i, f_i$.

Věta:

$$F_1(x) = \lim_{y \to \infty} F(x, y), \qquad F_2(y) = \lim_{x \to \infty} F(x, y)$$
$$\mathcal{P}_1(x) = \sum_y \mathcal{P}(x, y), \qquad \mathcal{P}_2(y) = \sum_x \mathcal{P}(x, y)$$
$$f_1(x) = \int_{-\infty}^{\infty} f(x, y) \, \mathrm{d}y, \quad f_2(y) = \int_{-\infty}^{\infty} f(x, y) \, \mathrm{d}x$$

Definice:

Veličiny X, Y se nazývají nezávislé, jestliže pro všechny podmnožiny A, B reálných čísel platí

$$P(X \in A \wedge Y \in B) = P(X \in A) \cdot P(Y \in B).$$

Věta:

X, Y jsou nezávislé $\Leftrightarrow$ pro $\forall\, x, y$ platí $F(x, y) = F_1(x) \cdot F_2(y)$.

Věta:

Je-li (X, Y) diskrétního typu, pak X, Y jsou nezávislé $\Leftrightarrow \forall\, x, y$ platí $\mathcal{P}(x, y) = \mathcal{P}_1(x) \cdot \mathcal{P}_2(y)$.

Věta:

Je-li (X, Y) spojitého typu, pak X, Y jsou nezávislé $\Leftrightarrow \forall\, x, y$ platí $f(x, y) = f_1(x) \cdot f_2(y)$, ale pokud hustotu pravděpodobnostizměním v konečně mnoha bodech, tak se to také nezmění (pro skoro všechna x, y).

2.3.1 Charakteristika náhodných vektorů

Uvažujme vektor (X, Y). Vektor $\big(E(X), E(Y)\big)$ se nazývá vektorem středních hodnot.

Definice:

Nechť $h(x,y)$ je nějaká funkce a (X,Y) je diskrétního typu. Pak

$$E\big(h(X,Y)\big) \overset{\text{def.}}{=\!=} \sum_{(x,y)} h(x,y)\cdot \mathcal{P}(x,y).$$

Je-li (X,Y) spojitého typu, pak

$$E\big(h(X,Y)\big) \overset{\text{def.}}{=\!=} \iint\limits_{\mathbf{R}^2} h(x,y)\cdot f(x,y)\ \mathrm{d}x\mathrm{d}y.$$

Definice:

Číslo

$$E\big(\big[X-E(X)\big]\big[Y-E(Y)\big]\big)$$

se nazývá *kovariancí* veličin X,Y a značí se $cov(X,Y)$.

Je-li $\sigma(X)\neq 0$ a $\sigma(Y)\neq 0$, pak číslo

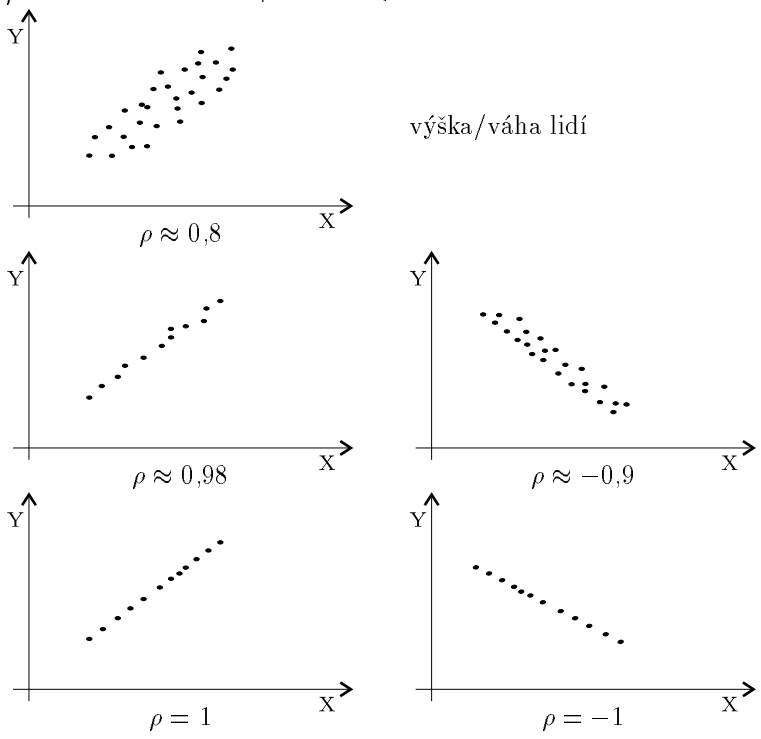
$$\rho=\rho(X,Y)=\frac{cov(X,Y)}{\sigma(X)\sigma(Y)}$$

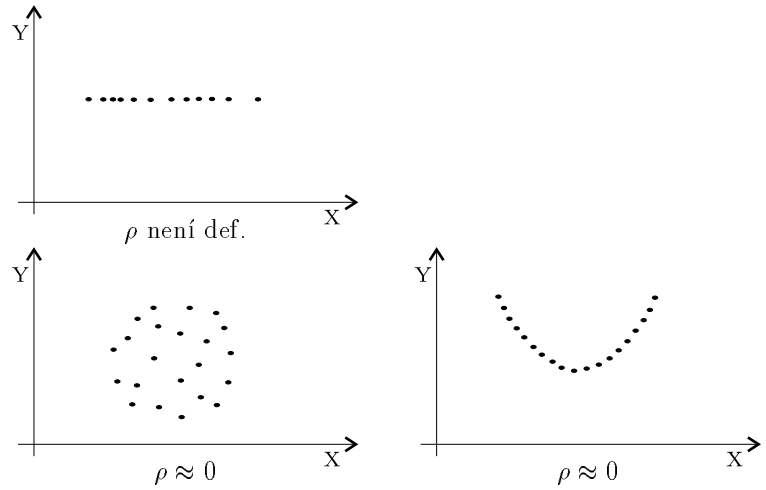
se nazývá *korelačním koeficientem*.

Je-li $\rho(X,Y)=0$, nazývají se veličiny nekorelovanými.

Poznámka:

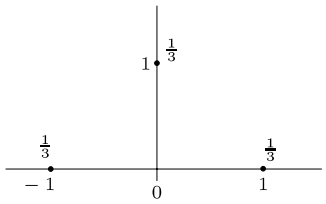
1. Vždy $-1\leq\rho(X,Y)\leq 1$.
2. X,Y nezávislé $\overset{\Rightarrow}{\not\Leftarrow}$ $cov(X,Y)=0$ (tj. $\rho=0$)
3. $\rho=1\iff Y=aX+b$ s $a>0$
   $\rho=-1\iff Y=aX+b$ s $a<0$





Př.:

Nechť (X, Y) je diskrétního typu, $\mathcal{P}(1;0) = \mathcal{P}(-1;0) = \mathcal{P}(0;1) = \frac{1}{3}, \mathcal{P}(x,y) = 0$ jinak.



$E(X) = -1 \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} = 0$
$E(Y) = 0 \cdot \frac{2}{3} + 1 \cdot \frac{1}{3} = \frac{1}{3}$
$\big[X - E(X)\big]\big[Y - E(Y)\big] = X\big(Y - \frac{1}{3}\big)$
$cov(X,Y) = (-1) \cdot (0 - \frac{1}{3}) \cdot \frac{1}{3} + 0 \cdot (1 - \frac{1}{3}) \cdot \frac{1}{3} + 1 \cdot (0 - \frac{1}{3}) \cdot \frac{1}{3} = 0$
$\rho(X,Y) = 0$
? $\mathcal{P}(0;1) = \mathcal{P}_1(0) \cdot \mathcal{P}_2(1)$
$\frac{1}{3} \stackrel{?}{=} \frac{1}{3} \cdot \frac{1}{3} \qquad \Longrightarrow$ závislé veličiny

$cov(X,X) = E\big(\big[X - E(X)\big] \cdot \big[X - E(X)\big]\big) = D(X)$

Věta: (výpočetní tvar kovariance)

$cov(X,Y) = E(XY) - E(X)E(Y)$ $\qquad$ (pokud $Y = X$, pak $cov(X,X) = D(X)$ a to je potom speciální případ)

Důsledek: X, Y nezávislé $\Rightarrow E(XY) = E(X)E(Y)$

Věta:

Je-li $(X_i, Y_i), i = 1, \ldots, n$, náhodný výběr pro dvourozměrný $\overrightarrow{\text{NV}}$ (X,Y), pak

$$\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})$$

je nestranný odhad výběrové kovariance $cov(X,Y)$.

Poznámka: Je-li $\sigma(X) \neq 0$ a $\sigma(Y) \neq 0$, pak pro $\rho(X,Y)$ se odhaduje pomocí

$$\frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2 \cdot \frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\overline{Y})^2}} \stackrel{\text{def.}}{=\!=} r(X,Y)$$

a to se nazývá *výběrový korelační koeficient*, ale to už není nestranným odhadem $\rho(X,Y)$, ale je to asymptoticky nestranným odhadem ρ.

Definice:

Buď $\vec{X} = (X_1, \ldots, X_k)$ $\overrightarrow{\text{NV}}$. Definujme matici

$$\sigma_{ij} = cov(X_i, Y_i)$$

pro $i,j = 1,\ldots,k$. Matice

$$\Sigma = \sigma_{ij}$$

se nazývá *kovarianční matice* (někdy varianční). Tato matice je symetrická. Na diagonále jsou rozptyly.

Matice

$$\rho_{ij} = \rho(X_i, Y_i)$$

pro $i,j = 1,\ldots,k$ se nazývá *korelační matice*. Je též symetrická a na diagonále má jedničky. Dají se definovat i výběrové kovarianční a korelační matice.

2.3.2 Vliv chyb při měření na přesnost výsledku

Chtějme zjistit hodnotu funkce k proměnných $y = f(x_1,\ldots,x_k)$, ve které hodnoty x_i jsou měřeny s náhodnými odchylkami ε_i, o kterých předpokládáme, že mají nulové střední hodnoty a jejich rozptyly označme $\sigma_i^2, i = 1,\ldots,k$. Předpokládejme, že σ_i^2 jsou malé. Chceme zjistit jakou E a D má výsledná chyba $\varepsilon_f = f(x_1+\varepsilon_1,\ldots,x_k+\varepsilon_k) - f(x_1,\ldots,x_k)$. Pomocí Taylorova rozvoje (totálního diferenciálu) dostaneme $\varepsilon_f \approx \sum_{i=1}^{k} \frac{\partial f}{\partial x_i}(\vec{x}) \cdot \varepsilon_i$. Pišme f_i místo $\frac{\partial f}{\partial x_i}(\vec{x})$, pak $\varepsilon_f \approx \sum_{i=1}^{k} f_i \cdot \varepsilon_i$. Potom

$$E(\varepsilon_f) = \sum_{i=1}^{k} f_i \cdot E(\varepsilon_i) = 0$$

$$D(\varepsilon_f) = E(\varepsilon_f^2) - E^2(\varepsilon_f) = E(\varepsilon_f^2) = E\left(\left[\sum_{i=1}^{k} f_i \cdot \varepsilon_i\right]^2\right) = E\left(\left[\sum_{i=1}^{k} f_i \cdot \varepsilon_i\right] \cdot \left[\sum_{j=1}^{k} f_j \cdot \varepsilon_j\right]\right) =$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{k} f_i \cdot f_j \cdot E(\varepsilon_i \varepsilon_j)$$

Předpokládejme, že chyby $\varepsilon_1,\ldots,\varepsilon_k$ jsou navzájem nezávislé. Pak

$$D(\varepsilon_f) = \underbrace{\sum_{i\neq j}\sum f_i f_j E(\varepsilon_i)E(\varepsilon_j)}_{\text{0, mimo diagonálu}} \neq \sum_i f_i^2 \cdot E(\varepsilon_i^2) =$$

$$= \sum_{i=1}^{k} f_i^2 \cdot \sigma_i^2 .$$

Př.:

Hodnota x_1 nechť je měřena s „chybou“ $\sigma_1 = 0{,}03 \cdot x_1$ a x_2 s „chybou“ $\sigma_2 = 0{,}02 \cdot x_2$.
Uvažujme $f(x_1, x_2) = \frac{x_1}{x_2}$, pak $f_1 = \frac{1}{x_2}, f_2 = -\frac{x_1}{x_2}$.
Potom $D(\varepsilon_f) = \frac{1}{x_2^2}(0{,}03 \cdot x_1)^2 + \frac{x_1^2}{x_2^4}(0{,}02 \cdot x_2)^2 = \frac{x_1^2}{x_2^2}(0{,}03^2 + 0{,}02^2)$
$\frac{\sqrt{D(X)}}{\frac{x_1}{x_2}} = \sqrt{0{,}03^2 + 0{,}02^2}$

2.3.3 Vícerozměrné normální rozdělení

Nechť $\vec{\mu} = (\mu_1, \ldots, \mu_k)^\top$ (transpozice ve statistice ale většinou ') a $\Sigma = \sigma_{ij}$ $(i, j = 1, \ldots, k)$ je symetrická pozitivně–definitní matice. Nechť $\vec{X} = (X_1, \ldots, X_k)^\top$ má k–rozměrné normální rozdělení s parametry $\vec{\mu}, \Sigma$, jestliže $\vec{X}$ má hustotu

$$f(x_1, \ldots, x_k) = \frac{1}{\sqrt{2\pi \cdot \det\Sigma}} \cdot e^{-\frac{1}{2}\overbrace{(\vec{x} - \vec{\mu})^\top \cdot \Sigma^{-1} \cdot (\vec{x} - \vec{\mu})}^{\text{pro } k=1 \text{ je to } (\frac{x-\mu}{\sigma})^2}}.$$

Věta:

Má-li vektor (X_1, X_2) normální rozdělení, pak X_1, X_2 jsou nezávislé $\iff cov(X_1, x_2) = 0$.

2.3.4 Test významnosti výběrového koeficientu korelace r

Pro dvourozměrné normální rozdělení vektoru (X, Y) lze testovat nazávislost složek X, Y tak, že testujeme hypotézu $\rho = 0$.

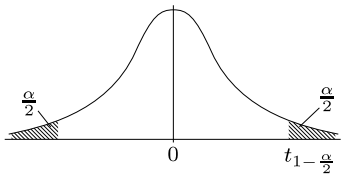
Buď r výběrový korelační koeficient spočtený pro $n > 2$. Pak statistika

$$t = \frac{r}{\sqrt{1 - r^2}}\sqrt{n - 2}$$

má t–rozdělení pravděpodobnosti s $\nu = n - 2$ (za předpokladu $\rho = 0$).

Hypotézu $\rho = 0$ zamítáme na hladině významnosti α, je-li

$$|t| > t_{1-\frac{\alpha}{2}}(n - 2)$$



Př.:

Ze šesti bodů a pro $r = 0{,}8$ nelze hypotézu $\rho = 0$ zamítnout na hladině významnosti α, ale pro $n = 7$ už hranici přelezeme a hypotézu zamítneme.

2.3.5 Konvoluce

Definice:

Nechť f, g jsou definovány na $(-\infty; \infty)$. Tzv. *konvoluce* funkcí f, g je funkce

$$h(t) = \int_{-\infty}^{\infty} f(t-x)g(x)\ \mathrm{d}x.$$

Při $t - x = u$ je

$$h(t) = \int_{\infty}^{-\infty} f(u)g(t-u)(-\mathrm{d}u) = \int_{-\infty}^{\infty} f(u)g(t-u)\ \mathrm{d}u.$$

Píšeme $h = f * g$.

Věta:

Nechť X, Y jsou NV spojitého typu s hustotami f, g.

Pak $X + Y$ má hustotu $f * g$.

$$\left[\begin{array}{l} H(t) = P(X+Y \le t) \\ \text{nezáv.} \Leftrightarrow \iint\limits_{x+y \le t} f(x)g(y)\ \mathrm{d}x\mathrm{d}y, \text{ pak } \underset{y}{x} \text{ běhá od } -\infty \text{ do } \underset{t-x}{t-y} \quad h(t) = H'(t) \end{array} \right]$$

2.3.6 Podmíněná pravděpodobnostní funkce a podmíněná hustota

Definice:

Nechť $\vec{X} = (X, Y)$ je diskrétního typu s pravděpodobnostní funkcí $\mathcal{P}(x, y)$. Definujme podmíněnou pravděpodobnostní funkci

$$\mathcal{P}(x|y) = P(X = x|Y = y) = \frac{P(X = x \wedge Y = y)}{P(Y = y)} =$$
$$= \frac{\mathcal{P}(x, y)}{\mathcal{P}_2(y)}$$
$$\mathcal{P}(y|x) = \frac{\mathcal{P}(x, y)}{\mathcal{P}_1(x)}$$

Pro spojitý $\overrightarrow{\mathrm{NV}}$ se definují podmíněné hustoty:

$$f(x|y) = \frac{f(x, y)}{f_2(y)}$$
$$f(y|x) = \frac{f(x, y)}{f_1(y)}$$

Poznámka: $f(x|y) = \frac{f(y|x)f_1(x)}{f_2(y)}$, všechny musí existovat.

Poznámka: Nechť $f(x|y)$ existuje a $f(y|x)$ také. Pak
X, Y jsou nezávislé $\Longleftrightarrow f(x|y) = f_1(x)$ pro $\forall\ y$
$X, Y \Longleftrightarrow f(y|x) = f_2(y)$ pro $\forall\ x$,
pro která to má smysl.

2.4 Regrese

Definice:

Podmíněná střední hodnota

$$E(X|y) \overset{\text{def.}}{=\!=} \int_{-\infty}^{\infty} x \cdot f(x|y)\ \mathrm{d}x$$

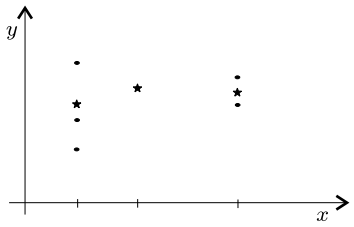
se nazývá *regresní hodnotou*.

Taktéž

$$E(Y|x) = \int_{-\infty}^{\infty} y \cdot f(y|x)\ \mathrm{d}x$$

$$\text{ev.} \sum_{y} y \cdot \mathcal{P}(y|x)$$

$$\vdots$$



a to budeme užívat častěji.

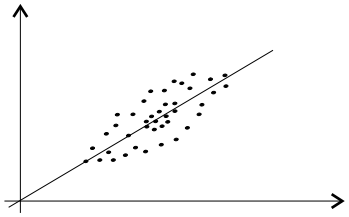
Funkce

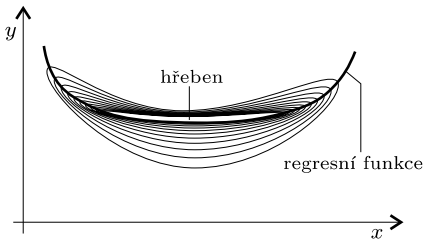
$$y \mapsto E(X|y), \text{ či } x \mapsto E(Y|x)$$

se nazývají *regresními funkcemi*.

Pro dvourozměrné normální rozdělení pravděpodobnosti vyjde

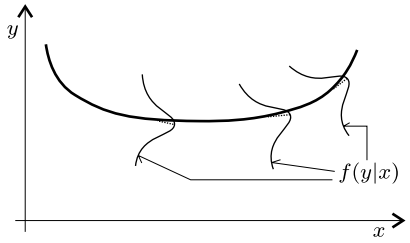
$$E(Y|x) = \mu_2 + \frac{\sigma_{12}}{\sigma_{11}}(x - \mu_1) \qquad \left(\begin{smallmatrix}\text{jednoduchý vzoreček po}\\ \text{několikastránkové integraci}\end{smallmatrix}\right)$$





– vrstevnice plochy

– dvourozměrný $\overrightarrow{\mathrm{NV}}$ s dvourozměrným rozdělením, ne s normálním



regrese = návrat zpět (zkoumání výšky otce a syna; když je otec nadprůměrný, syn je také, ale jejich průměr je blíž k E a opačně; návrat zpět ke střední hodnotě)

2.4.1 Jednoduchá regrese (přímka)

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (i = 1, \ldots, n)$; zde se na to dívám jako na číslo (malá y_i, x_i)

x_i ... vysvětlující proměnná

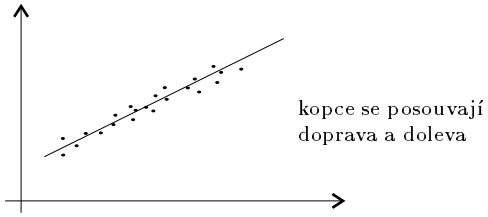
y_i ... vysvětlovaná proměnná

ε_i ... náhodné odchylky

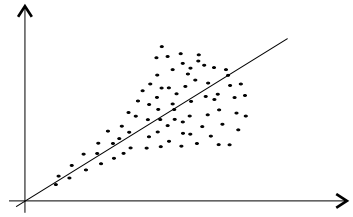
β_0, β_1 ... regresní koeficienty

Odchylky ε_i nechť splňují podmínky:

1. $E(\varepsilon_i) = 0$
2. $D(\varepsilon_i) = \sigma^2$ (konstantní, nezávisí na i) — *homoskedasticita*
3. $\varepsilon_1, \ldots, \varepsilon_n$ jsou nezávislé



splněna homoskedasticita
měření odpovídá homosked. modelu



měření neodpovídá homosked. modelu
– heteroskedastický model

Odhady koeficientů β_0, β_1 spočteme pomocí MNČ (metody nejmenších čtverců).

MNČ: b_0, b_1 – jde o hodnoty, které minimalizují součet

$\hat{y}_i = b_0 + b_1 x_i$ – očekávaná hodnota

y_i – naměřená hodnota

$$\sum_{i=1}^{n} \left[y_i - \hat{y}_i\right]^2 = \sum_{i=1}^{n} \left[y_i - (b_0 - b_1 x_i)\right]^2$$

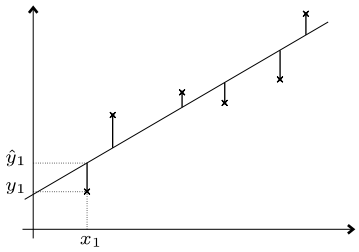
– v dnešní době se začínají užívat absolutní hodnoty, ale je to výpočetně náročné – rozsekání na intervaly; ale lépe to vzdoruje chybám při měření, kvadrát tu chybu ještě zveličí

$\rightarrow$ zderivování podle b_i

Koeficienty b_0, b_1 spočtené MNČ jsou nestrannými odhady β_0, β_1, jsou lineární vzhledem k $y_1, \ldots, y_n$ a mezi všemi nestrannými odhady lineárními vzhledem k $y_1, \ldots, y_n$ mají nejmenší rozptyl $\longrightarrow$ nejlepší nestranné lineární odhady.

Definujme tzv. vyrovnané (očekávané) hodnoty

$\hat{y}_i = b_0 + b_1 x_i$



– odchylky kolmo

Rozdíly $e_i = y_i - \hat{y}_i$ se nazývají *rezidua*.

Součet $\sum_{i=1}^{n} e_i^2$ se značí RSS (*residual sum of squares*, reziduální součet čtverců), SSE (*sum of square errors*), S_e.

Statistika $s^2 = \dfrac{\text{RSS}}{n-2}$ je nestranným odhadem parametru σ^2 a nazývá se *reziduální rozptyl*.

Hodnota $s = \sqrt{s^2}$ (odmocnina z reziduálního rozptylu) se v Excelu uvádí pod názvem chyba střední hodnoty.

2.4.2 Koeficient determinace

Značí se $R^2 = 1 - \dfrac{\text{RSS}}{\sum (y_i - \overline{y})^2}$.

Udává, jaký podíl rozptylu hodnot y_i se podařilo vysvětlit pomocí regresní závislosti.

$R = \sqrt{R^2}$ se nazývá *koeficient vícenásobné korelace*.

Vždy $0 \leq R^2 \leq 1$.

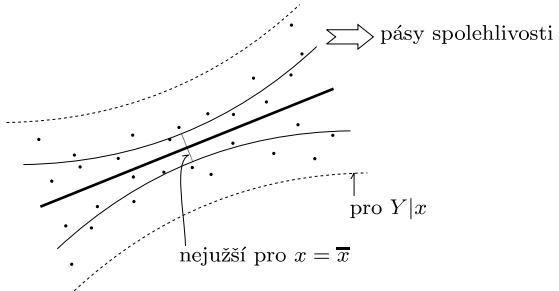
adjusted $R^2 = \underbrace{\text{upravené } R^2}$ (s $n-2$ je to normálně rozšířené, k je počet regresních koeficientů)

$$= 1 - \frac{\frac{\sum e_i^2}{n-2-(k+1)}}{\frac{\sum (y_i - \overline{y})^2}{n-2}}$$

Předpovědní intervaly

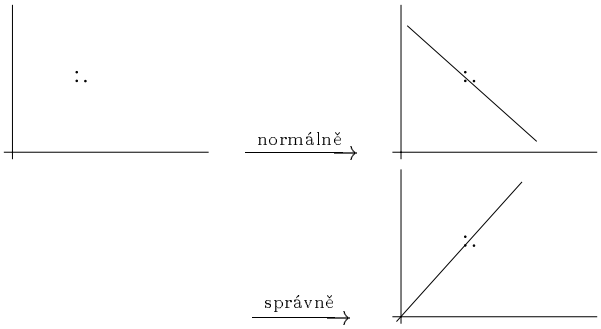
Konstruují se pro $E(Y|x)$, tj. pro regresní hodnotu nebo pro $Y|x$, tj. pro jednu náhodnou hodnotu (budoucí pozorování).

– interval užší pro $E(Y|x)$ než pro $Y|x$



Přímka procházející počátkem

$y_i = \beta_1 x_i + \varepsilon_i$



TO NE !

MNČ:

$$\sum [y_i - b_1 x_i]^2$$

$$b_1 = \frac{\sum x_i y_i}{\sum x_i^2}$$

Excel pak počítá ale R^2 pro přímku s úsekem a to mu vyjde záporně. Hodnotu b_1 spočítá správně.

2.4.3 Maticový zápis vícenásobné regrese

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ $(i = 1, \ldots, n)$ lze zapsat v maticovém tvaru:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Pak můžeme psát

$$\vec{y} = X \cdot \vec{\beta} + \vec{\varepsilon},$$

kde X se nazývá *regresní matice*.

Obecněji lze uvažovat více vysvětlujících proměnných.

$$y_i = x_{i0}\beta_0 + x_{i1}\beta_1 + \ldots + x_{ik}\beta_k + \varepsilon_i \quad (i = 1, \ldots, n), \text{ kde obvykle } x_{i0} = 1.$$

Předpokládejme obecný model $\vec{y} = X \cdot \vec{\beta} + \vec{\varepsilon}$, kde X je matice typu $n/k+1, n > k+1$ a má maximální možnou hodnost, tj. $k+1$, sloupce jsou lineárně nezávislé.

Př.:

$y_i = \beta_0 + \beta_1 u_i + \beta_2 v_i + \varepsilon_i$

y_i ... výška syna

u_i ... výška otce

v_i ... výška matky

$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$ — závislost spotřeby na rychlosti auta (ale asi bez β_0)

$$X = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}$$

Poznámka: Při zadávání regresní matice X se v programech zpravidla nezadává sloupec $(1, \ldots, 1)^\top$, který je přidán automaticky programem.

Anova slouží k testování hypotézy, že vysvětlovaná proměnná nezávisí na žádné vysvětlující a regrese je tvaru $y_i = \beta_0 + \varepsilon_i$.

Obsah