Pravděpodobnost

Náhodný pokus
Náhodný jev (=jeden výsledek)
Množina všech jevů (všech možných výsledků) Ω
Jev jistý = celá množina Ω
Jev nemožný
Morganovo pravidlo

Pravděpodobnost $\boldsymbol{0 \leq p \leq 1}$
Podmíněná pravděpodobnost $\boldsymbol{P(Y|X)}$
Apriorní vs. aposteriorní pravděpodobnost
Bayesův vzorec

Pravděpodobnost

Náhodný pokus
Náhodný jev
Množina všech elementárních jevů Ω
Jev jistý = celá množina Ω
Jev nemožný
Morganovo pravidlo

Pravděpodobnost
Podmíněná pravděpodobnost
Apriorní vs. aposteriorní pravděpodobnost
Bayesův vzorec

Náhodná veličina

Když chceme popsat hodně velký (až nekonečný) základní soubor, pracujeme většinou jen s výběrem, výběrovým souborem.
Z výběrových dat potom spočítáme nějakou charakteristiku, která má reprezentovat vlastnost celého základního souboru.
Takový výběr můžeme mnohokrát opakovat. Zvolená charakteristika pak bude mít pokaždé trochu jinou hodnotu, protože na výsledku se podílí i náhoda – ve výběru subjektů do výběrového souboru.
Proto všechny hodnoty pozorované nebo měřené na náhodně vybraných subjektech (počet listů na rostlině, délka zobáku kosa) nazýváme **náhodná veličina** nebo **náhodná proměnná** [random variable] a konkrétní zjištěnou hodnotu realizace náhodné veličiny.

Někdy mluvíme také o výsledku náhodného procesu. Tím se myslí například měření rychlosti, kterou padá semeno trubicí (proces). Výsledkem je potom ta rychlost.

Náhodná veličina

Matematika vidí náhodnou veličinu jako funkci, která každému subjektu přiřadí hodnotu z množiny možných hodnot. Chování náhodné veličiny potom popisuje jako výčet přípustných hodnot a pravděpodobností, s nimiž mohou subjekty těchto hodnot nabývat.

Příklad: 1/ házení kostkou;

2/ holčičky (●) a kluci (□) v rodině se 3 dětmi:

(●,●,●); (●,●,□); (●,□,●); (□,●,●); (●,□,□); (□,□,●); (□,●,□); (□,□,□).

→ $\mathrm{P}(0\ \text{kluků}) = \frac{1}{8} = 0{,}125$ $\mathrm{P}(2\ \text{kluci}) = \frac{3}{8} = 0{,}375$

$\mathrm{P}(1\ \text{kluk}) = \frac{3}{8} = 0{,}375$ $\mathrm{P}(3\ \text{kluci}) = \frac{1}{8} = 0{,}125$

[čti: pravděpodobnost, že v rodině budou právě 3 kluci, se rovná…]

Toto se nazývá **rozdělení pravděpodobnosti** [probability distribution], často jenom rozdělení nebo distribuce.

Poznámka: V teorii přísně rozlišujeme náhodné veličiny s diskrétním rozdělením pravděpodobností (házení kostkou) a se spojitým rozdělením prstí (výška člověka). V praxi se často použije spojité rozdělení pro diskrétní data tam, kde dostávám alespoň desítky různých hodnot (např. počty krvinek).

Značení

Náhodnou veličinu označujeme většinou **X, Y, Z,** tj. velká písmena z konce abecedy.

Zápis: $P(X = x_i) = p_i, x_i \in \{x_1, x_2, \cdots, x_m\}$

[čti: pravděpodobnost, že náhodná veličina X nabyde hodnoty x_i je p_i. x_i leží v množině hodnot $x_1, \cdots, x_m$, kterých může nabývat veličina X.]

Diskrétní rozdělení pravděpodobností - tabulkou

- Typické pro data na nominální a ordinální stupnici, ale i čísla…
- Hodnoty x_i jsou od sebe jasně odděleny, je jich nejvýše spočetně
- Chování diskrétní náhodné veličiny mohu popsat „tabulkou“ (např. házení kostkou, počet kluků v rodině), ale také vzorečkem.

Vlastnosti:

každé $\boldsymbol{p_i \geq 0}$ (tedy neznáme zápornou pravděpodobnost)

součet $\boldsymbol{\sum_{i=1}^{m} p_i = 1}$ (tedy množina $\{x_1, x_2, \cdots, x_m\}$ popisuje všechny možnosti, hodnoty pro náhodnou veličinu X)

Proto máme v dotazníku políčko „jiné“, popíšeme tak všechny možnosti.

Spojité rozdělení pravděpodobností - funkcí

Příklad: hmotnost novorozenců, váha biomasy apod.

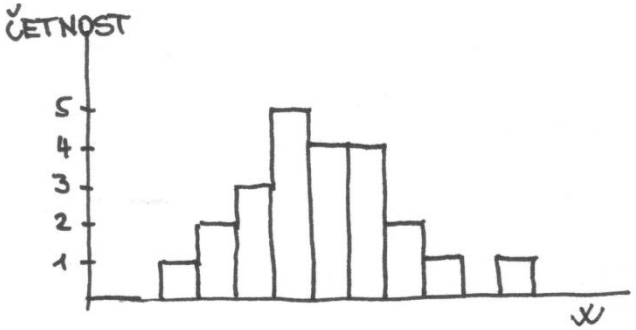
Toto rozdělení „prstí“ nemohu popsat tabulkou, protože veličina X může nabývat nekonečně mnoha hodnot (milimetr a půl, milimetr a $^3/_4$).

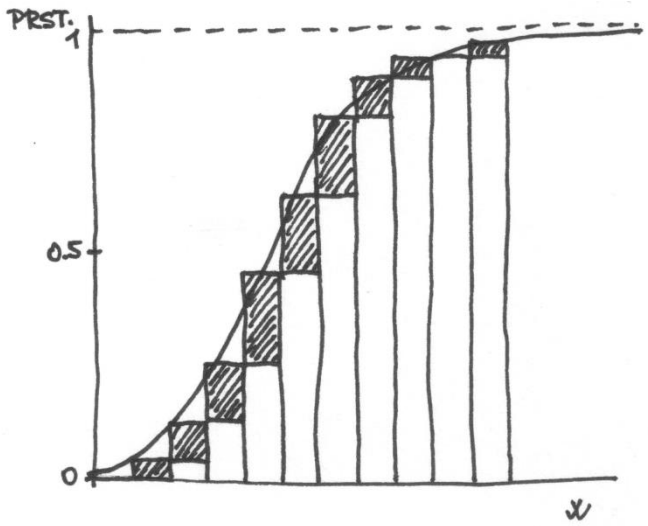
Neptáme se na jednu hodnotu $P(X = 5)$, ale spíše na $P(X \leq x_i)$.

Spojité rozdělení popisujeme funkcí.

Známe už histogram četností.

Z něho sestrojíme histogram kumulativních (relativních) četností:

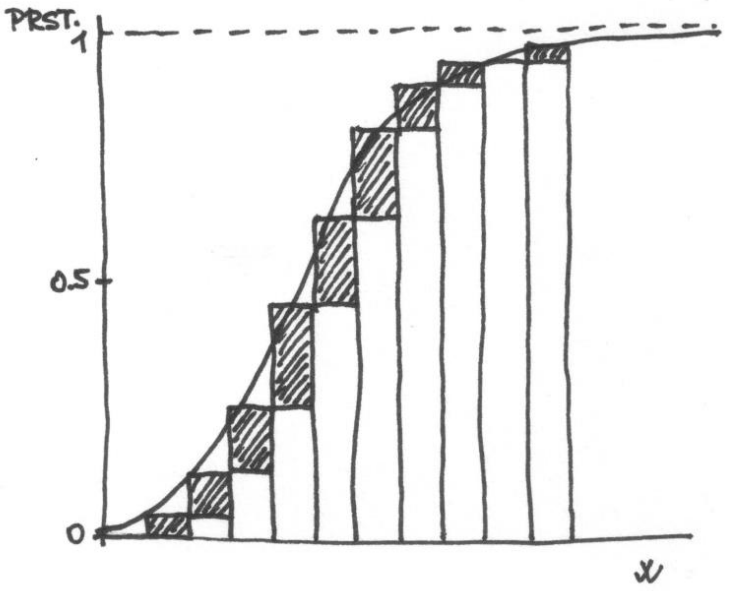




Distribuční funkce náhodné veličiny X

[(cumulative) distribution function]

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



… někteří autoři píší $F(x) = P(X < x)$. Je to také dobře, protože pro spojitou distribuční funkci platí $\boldsymbol{P(X = x) = 0}$, tedy prst, že veličina nabyde hodnoty právě $\boldsymbol{x}$, je nulová.

Pozor, pro diskrétní náhodnou veličinu musím definovat, kam patří mezní hodnoty, proto je tam rozdíl mezi $P(X \leq x)$ a $P(X < x)$.

Vlastnosti distribuční funkce

- $0 \leq F(x) \leq 1$
- Je neklesající
- $\lim_{x \to -\infty} F(x) = 0; \quad \lim_{x \to +\infty} F(x) = 1$
- $P(x_1 < X \leq x_2) = F(x_2) - F(x_1)$

Pro X s diskrétním rozdělením píšeme:

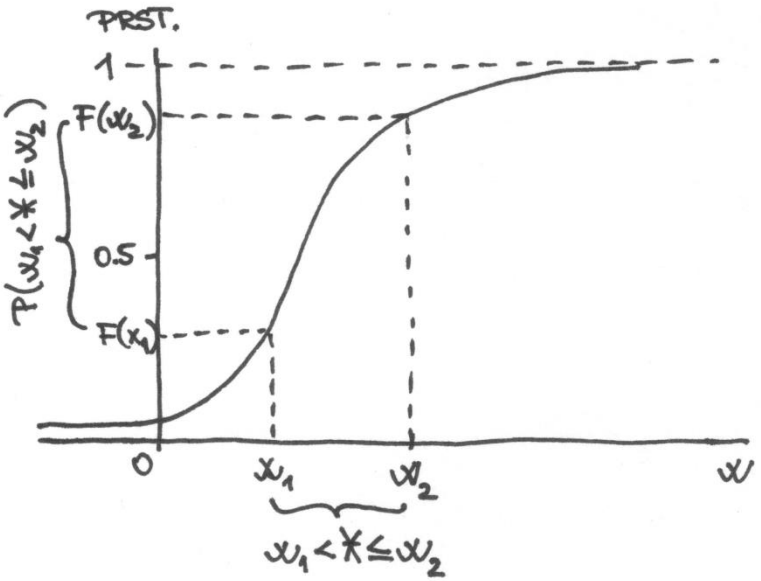
$$F(x) = \sum_{t \leq x} p(t) = \sum_{x_i \leq x} p_i = \sum_{x_i \leq x} P(X = x_i)$$

funkce $p(t)$ se nazývá pravděpodobnostní funkce [probability mass function]

Pro X se spojitým rozdělením lze zapsat:

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

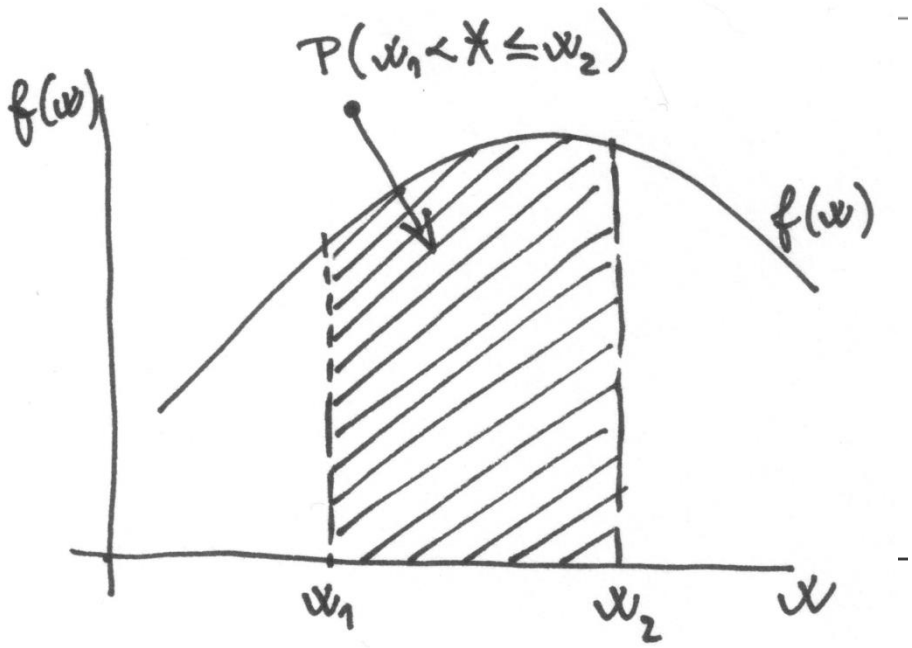
kde funkce *f(x)* je derivací distribuční funkce F(x): $F'(x) = f(x)$



Funkci ***f(x)*** nazýváme **hustotou pravděpodobností** náhodné veličiny X [probability density function].

Platí: $\boldsymbol{P(x_1 < X \leq x_2) = F(x_2) - F(x_1) = \int_{x_1}^{x_2} f(t)dt}$

a $\boldsymbol{\int_{-\infty}^{\infty} f(x)dx = 1}$ … hodnota X leží jistě mezi $-\infty$ a $+\infty$.



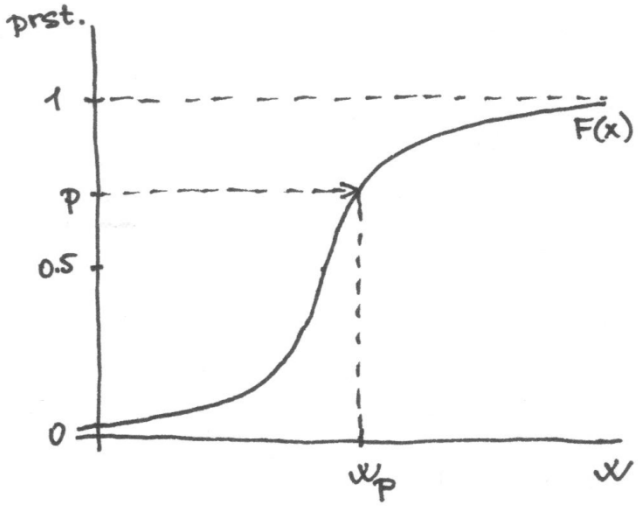
! <u>Pozor na měřítka</u>:
histogram četností ukazuje počty hodnot v intervalu,
histogram <u>relativních</u> četností ukazuje pravděpodobnost, že vyberu hodnotu z daného intervalu.
Funkci hustoty pravděpodobností ***f(x)*** můžeme chápat jako idealizovaný histogram relativních četností pro nekonečně velký základní soubor.

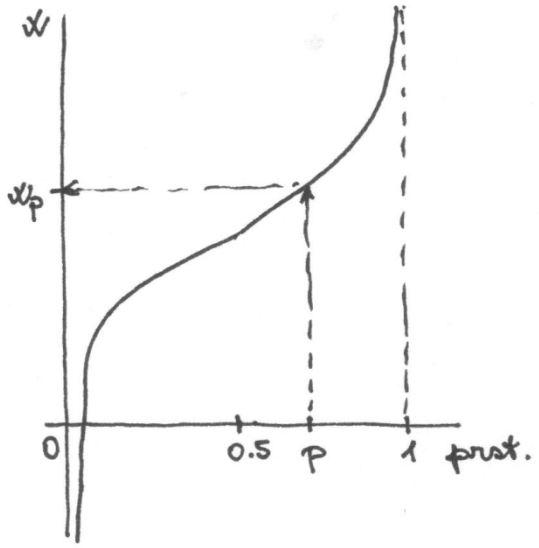
Kvantilová funkce [quantile function]

Je to inverzní funkce k distribuční funkci ***F(x)***

Značíme ji $\boldsymbol{\Phi(p)}$ *[čti: fí]* a funguje takto:

$$\boldsymbol{\Phi(F(x)) = x}$$
$$\boldsymbol{F(x) = p}$$
$$\boldsymbol{\Phi(p) = x}$$





V souvislosti
s kvantilovou funkcí
říkáme hodnotám ***x*** **kvantily**.

p-kvantil je taková hodnota $\boldsymbol{x_p}$, pro kterou $\boldsymbol{F(x_p) = \mathbf{P}(\mathbf{X} \leq x_p) = p}$

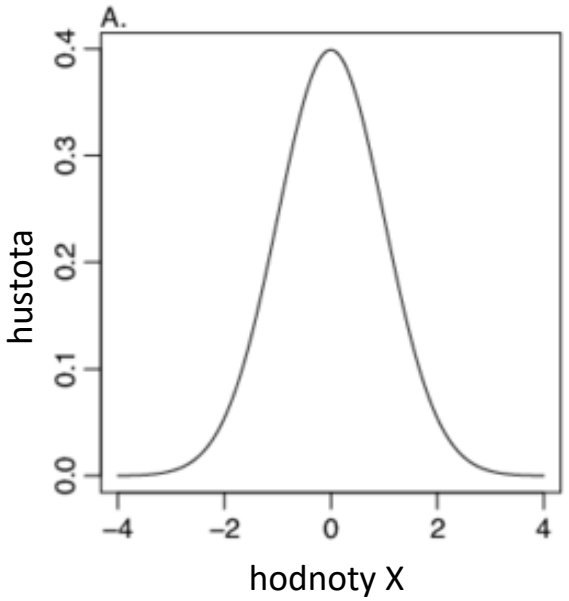
$\boldsymbol{95\%\ kvantil = x_{0,95}: P(X \leq x_{0,95}) = 0,95}$

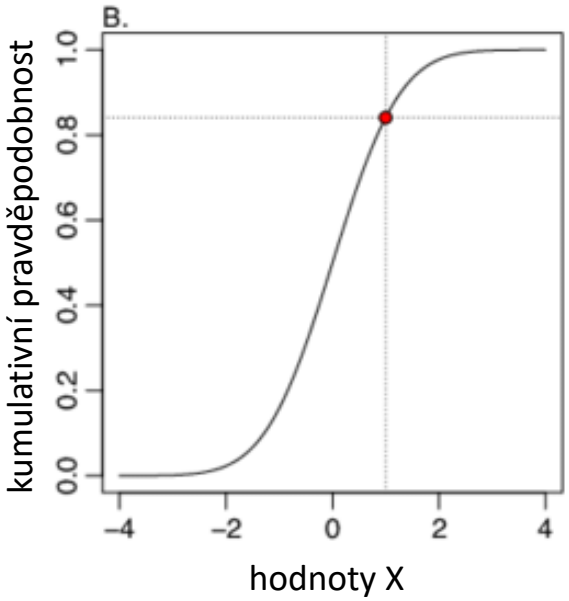
a také $\boldsymbol{P(X > x_{0,95}) = 0,05}$

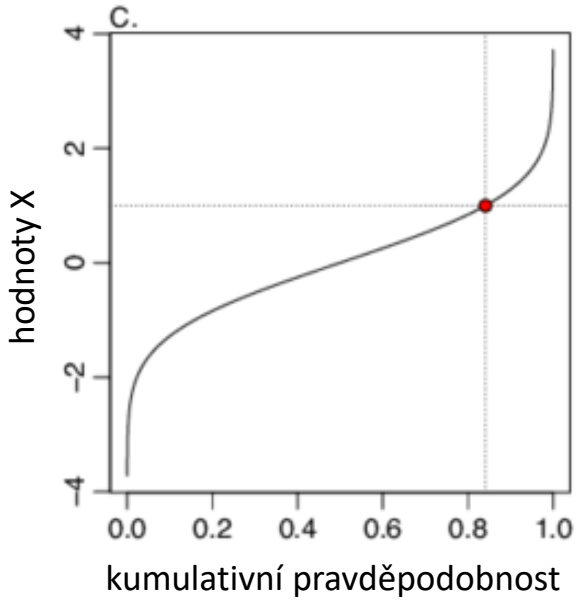
Tuto vlastnost využíváme při testování hypotéz.

Trojice funkcí: **Hustota** – **distribuční funkce** – **kvantilová funkce**

Příklad: normální rozdělení







rovnoměrné rozdělení

Charakteristiky náhodné veličiny

Střední hodnota [expected value, mean value]

$$\mu_X = EX = \sum_{j=1}^{\infty} x_j^* P(X = x_j^*)$$

veličina **X** s diskrétním rozdělením
… x_j^* je typická hodnota

$$\mu_X = EX = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

veličina **X** se spojitým rozdělením
f(x) je hustota

- Je to „vážený průměr“ všech možných hodnot veličiny **X**, vážíme pomocí pravděpodobnosti, že hodnota nastane.
- Jiná interpretace: je to těžiště možných hodnot veličiny **X**.
- Odvodíme to cestou z výběrového průměru na populační průměr:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i \xrightarrow{\text{celá populace}} \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{N}\sum_{j=1}^{m} n_j \cdot x_j^* = \sum_{j=1}^{m} \frac{n_j}{N} \cdot x_j^*$$

relativní četnost $x_j^* \approx$ pravděpodobnost

Příklad s rovnoměrným rozdělením

(Populační) rozptyl [variance, dispersion]

různé typy zápisu

$$\sigma^2_X = varX = D(X) = E(X - EX)^2 = E(X - \mu_X)^2$$

- Je to průměrný čtverec vzdálenosti náhodné veličiny ***X*** od její střední hodnoty ***EX.***
- populační ≈ typicky neznámá hodnota, výpočet zahrnuje všechny možné hodnoty a jejich pravděpodobnosti
- Rozptyl má jiné měřítko než původní náhodná veličina, je to mocnina.

(Populační) směrodatná odchylka [standard deviation]

$$\sigma_X = \sqrt{\sigma^2_X} = SD(X) = \sqrt{D(X)}$$

- Odmocnina z populačního rozptylu
- Má stejné měřítko jako původní náhodná veličina

Míry diverzity

- Variabilita veličiny na nominální stupnici (např. barva květů)

Měli jsme entropii výběrového souboru:

$$H = -\sum_{j=1}^{m} \frac{n_j}{n} \cdot ln \frac{n_j}{n}$$

! Kategorie jsou očíslovány, ale výpočet charakteristik není ovlivněn pořadím kategorií.

Shannonova entropie je populační obdoba, místo relativních četností má přímo pravděpodobnosti

$$H = -\sum_{j=1}^{m} p_j \cdot ln(p_j)$$

Používá se jako míra diverzity, míra bohatosti společenstva.

Simpsonův index (Giniho index)

$$1 - \sum_{j=1}^{m} {p_j}^2 = \sum_{j=1}^{m} p_j \cdot (1 - p_j)$$

Pravá strana rovnice popisuje prst mezidruhového setkání při neomezeně velkém společenstvu.

Několik náhodných veličin, náhodný vektor

Na **jednom subjektu** měřím více znaků => náhodný vektor ***(X, Y, Z, …)***

Sdružené rozdělení [joint distrib.] prstí náhodného vektoru ***(X, Y)***:

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y) \qquad \text{… "a zároveň„}$$

Marginální rozdělení [marginal distrib.] nazýváme rozdělení prstí samotných veličin ***X*** a ***Y***

$$F(x) = P(X \leq x), \qquad F(y) = P(Y \leq y)$$

Chování dvojice náh. vel. ***(X, Y)*** nemůžeme popsat (v obecném případě) chováním jednotlivých náh. vel. ***X*** a ***Y***. Nic bychom nevěděli o jejich vztahu.

Náh. vel. ***X, Y*** s diskrétním rozdělením:

Sdružené rozdělení zapíšu jako trojici: typické hodnoty + prst. výskytu

$$x_i^*, y_j^*, P\left(X = x_i^*, Y = y_j^*\right) \qquad \text{… pro každé } i \text{ a } j$$

Náh. vel. ***X, Y*** se spojitým rozdělením:

Sdružené rozdělení: $F_{X,Y}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(x, y)\, dy\, dx$

Když chceme z $F_{X,Y}$ získat prst $P(X \leq x)$, musíme se zbavit omezení na náh vel Y, tj. hodnoty Y omezit co možná nejméně:

$$\lim_{y\to\infty} F_{X,Y}(x,y) = \lim_{y\to\infty} P(X \leq x, Y \leq y) = P(X \leq x) = F_X(x)$$

Podobně pro Y:

$$\lim_{x\to\infty} F_{X,Y}(x,y) = \lim_{x\to\infty} P(X \leq x, Y \leq y) = P(Y \leq y) = F_Y(y)$$

Ještě vzorečky:

Náh vel X, Y s diskrétním rozdělením:

Marginální rozdělení: $\sum_j P(X = x_i^*, Y = y_j^*) = P(X = x_i^*)$

$$\sum_i P(X = x_i^*, Y = y_j^*) = P(Y = y_j^*)$$

Náh vel X, Y se spojitým rozdělením:

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

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

Nezávislost náhodných veličin

Máme-li popsat chování náhodného vektoru, musíme vědět, jak spolu složky náhodného vektoru interagují.

U diskrétních veličin je teoreticky možné získat relativní četnosti pro všechny dvojice $(X = x_i^*, \mathrm{Y} = y_j^*)$. Ale spojitou sdruženou hustotu $\boldsymbol{f_{X,Y}}$ většinou získat neumíme.

Situace se velmi zjednoduší, když ukážeme, že naměřená hodnota ***Y*** je nezávislá na tom, jakou hodnotu jsme naměřili pro veličinu ***X*** (přesto, že jsou naměřené na stejném subjektu).

Potom je chování veličiny ***X*** **nezávislé** na chování veličiny ***Y*** a platí:

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

Přesněji pro diskrétní veličiny:

$$\boldsymbol{P\left(X = x_i^*, \mathrm{Y} = y_j^*\right) = P\left(X = x_i^*\right) \cdot P\left(\mathrm{Y} = y_j^*\right)} \ \ldots \text{ pro každé } \boldsymbol{x_i^*} \text{ a } \boldsymbol{y_j^*}$$

a pro spojité:

$$\boldsymbol{f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y)} \quad \ldots \text{pro každé } \boldsymbol{x} \text{ a } \boldsymbol{y}$$

Kovariance náh. vel. X a Y [covariance]

popisuje vzájemnou závislost či nezávislost X a Y

$$\sigma_{XY} = cov(X,Y) = E(X-EX)(Y-EY) = E(X-\mu_X)(Y-\mu_Y)$$

Všimněte si, že $cov(X,X) = E(X-EX)(X-EX) = E(X-EX)^2 = var\ X$

Jsou-li X, Y nezávislé, potom platí, že $\sigma_{XY} = cov(X,Y) = 0$.

! Neplatí naopak:

i když je $cov(X,Y) = 0$, přesto X a Y mohou být závislé!

Pro nezávislé X, Y s diskrétním rozdělením počítáme takto:

$$cov(X,Y) = E(X-\mu_X)(Y-\mu_Y) =$$
$$= \sum_i \sum_j (x_i^* - \mu_X)(y_j^* - \mu_Y) P(X = x_i^*, Y = y_j^*) =$$
$$= \sum_i \sum_j (x_i^* - \mu_X)(y_j^* - \mu_Y) P(X = x_i^*) P(Y = y_j^*) =$$
$$= \sum_i (x_i^* - \mu_X) \cdot P(X = x_i^*) \cdot \sum_j (y_j^* - \mu_Y) \cdot P(Y = y_j^*) =$$
$$= \sum_i x_i^* \cdot P(X = x_i^*) - \mu_X \cdot \sum_i P(X = x_i^*) \cdot \sum_j y_j^* \cdot P(Y = y_j^*) - \mu_Y \cdot 1 =$$
$$= (EX - \mu_X)(EY - \mu_Y) = 0.$$

tady uplatním vlastnost nezávislosti

Kovariance

Příklad, jak poznám, že jsou veličiny závislé:

P(X=0) P(Y=0) se nerovná P(X=0, Y=0)

Vlastnosti střední hodnoty

- Využijeme při změně měřítka náh. vel. X (transformace), při normování (Z-skóry) i při součtu náhodných veličin
- Změna měřítka: $\alpha + \beta \cdot X$, tedy α~posunutí, β~násobení

Platí:

Střední hodnota: $\mu_{\alpha+\beta X} = E(\alpha + \beta X) = \alpha + \beta\, EX = \alpha + \beta\, \mu_X$

Rozptyl: $\sigma^2{}_{\alpha+\beta X} = var(\alpha + \beta X) = \beta^2\, varX = \beta^2\, \sigma^2{}_X$

Směrodatná odch.: $\sigma_{\alpha+\beta X} = sd(\alpha + \beta X) = |\beta|\, sd(X) = |\beta|\, \sigma_X$

Součet náh. veličin: $\mu_{X+Y} = E(X + Y) = EX + EY = \mu_X + \mu_Y$

$$\sigma^2{}_{X+Y} = var(X + Y) =$$

$$= var(X) + var(Y) + 2\, cov(X, Y)$$

$$= \sigma^2{}_X + \sigma^2{}_Y + 2\, \sigma_{XY}$$

Vlastnosti střední hodnoty

Alespoň jedno odvození na procvičení ☺

Normování

$$Z = \frac{X - \mu_X}{\sigma_X}$$

$$\mu_Z = E\left(\frac{X - \mu_X}{\sigma_X}\right) = \frac{1}{\sigma_X}(EX - \mu_X) = 0$$

$$\sigma^2{}_Z = E(Z - \mu_Z)^2 = EZ^2 = E\left(\frac{X - \mu_X}{\sigma_X}\right)^2$$

$$= \frac{1}{\sigma^2{}_X} E(X - \mu_X)^2 = \frac{varX}{\sigma^2{}_X} = 1$$

= změna měřítka náhodné veličiny tak, že nová veličina má nulovou střední hodnotu a jednotkový rozptyl, tedy $\mu = 0$ a $\sigma^2{}_Z = 1$.

- Z je fyzikálně bezrozměrná veličina
- Tvar pro výběrový soubor jsou z-skóry (bylo dříve)
- Normovaný tvar náh. veličiny se používá k výpočtu vlastností, které nezávisí na populačním průměru ani na populačním rozptylu.
- Příkladem jsou **koeficienty šikmosti a špičatosti** [skewness, kurtosis]

$$\gamma_1 = EZ^3 = \frac{E(X-\mu_X)^3}{\sigma^3{}_X} \qquad \gamma_2 = EZ^4 - 3 = \frac{E(X-\mu_X)^4}{\sigma^4{}_X} - 3$$

Tvar pro výběrový soubor jsme uvedli jako g_1 a g_2.

Korelační koeficient [coefficient of correlation]

= kovariance normovaných tvarů dvou náhodných veličin X, Y

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

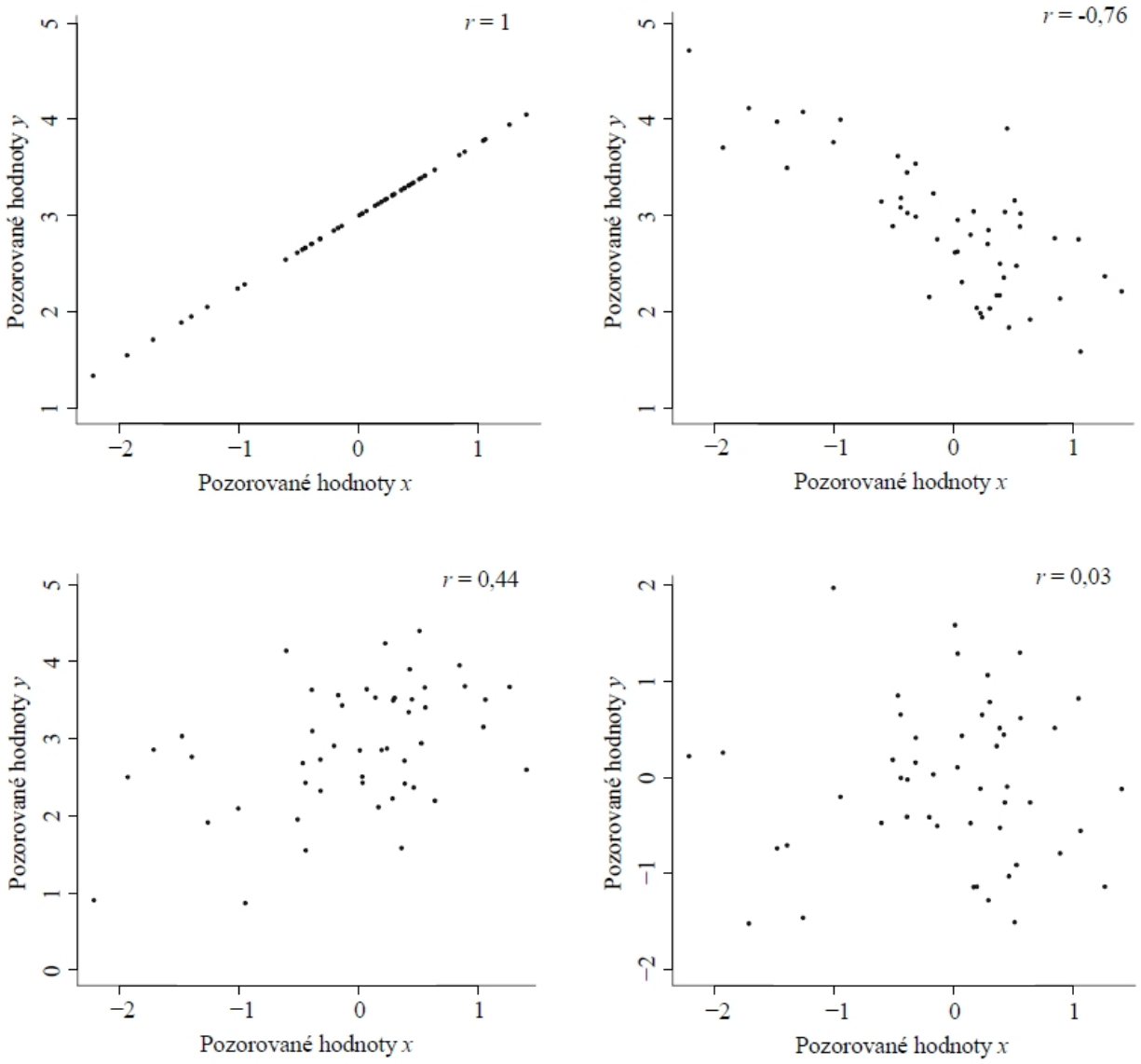
- platí $\rho_{XY} \in \langle -1, 1 \rangle$, tedy $-1 \leq \rho_{XY} \leq 1$

- měří sílu lineární závislosti X a Y
- můžeme tak srovnávat sílu závislosti mezi znaky o různých průměrech a rozptylech
- nezávislé X, Y mají $\sigma_{XY} = 0$, tedy také $\rho_{XY} = 0$. Naopak to neplatí!
- svých extrémních hodnot ± 1 nabývá tehdy, když $Y = \alpha \pm \beta X$
- bude v kapitole o regresi a závislosti (korelaci)

Odvození horní rovnosti:

$$\rho_{XY} = cov\left(\frac{X - \mu_X}{\sigma_X}, \frac{Y - \mu_Y}{\sigma_Y}\right) = E\left[\left(\frac{X - \mu_X}{\sigma_X} - 0\right) \cdot \left(\frac{Y - \mu_Y}{\sigma_Y} - 0\right)\right] =$$

$$= \frac{1}{\sigma_X \sigma_Y} E[(X - \mu_X)(Y - \mu_Y)] = \frac{cov(X, Y)}{sd(X) sd(Y)} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$



Obr. 10.2 Ukázky realizací náhodných veličin X a Y a vypočtené výběrové korelační koeficienty.

Pavlík & Jarkovský, Biostatistika, str. 120