Číselné charakteristiky náhodných veličin

Motivace

Doposud jsme poznali funkcionální charakteristiky náhodných veličin (např. distribuční funkce, pravděpodobnostní funkce, hustota pravděpodobnosti), které plně popisují pravděpodobnostní chování náhodné veličiny. Číselné charakteristiky vystihují pouze některé rysy tohoto chování, např. popisují polohu realizací náhodné veličiny na číselné ose či jejich proměnlivost (variabilitu). Jsou jednodušší než funkcionální charakteristiky, ale nesou jen částečnou informaci.
Podobně jako v popisné statistice volíme vhodnou číselnou charakteristiku podle toho, jakého typu je daná náhodná veličina - zda je ordinální nebo intervalová či poměrová. Číselné charakteristiky znaků mají své teoretické protějšky v číselných charakteristikách náhodných veličin.

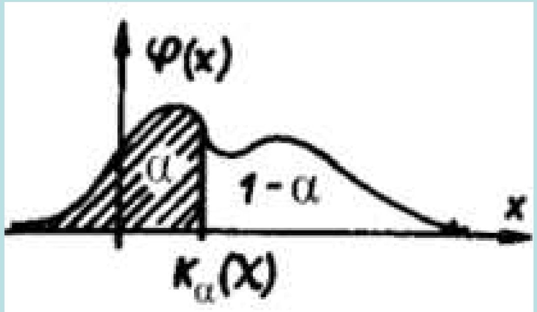
Číselné charakteristiky spojité náhodné veličiny aspoň ordinálního typu

Charakteristika polohy : α-kvantil
Nechť X je spojitá náhodná veličina aspoň ordinálního typu s distribuční funkcí $\Phi(x)$ a hustotou pravděpodobnosti $\varphi(x)$.
Nechť $\alpha \in (0, 1)$.
Číslo $K_\alpha(X)$, které splňuje podmínku

$$\alpha = \Phi(K_\alpha(X)) = \int_{-\infty}^{K_\alpha(X)} \varphi(x)\,dx,$$

se nazývá α-kvantil náhodné veličiny X.
$K_{0,50}(X)$ - medián,
$K_{0,25}(X)$ - dolní kvartil,
$K_{0,75}(X)$ - horní kvartil,
$K_{0,10}(X), \ldots, K_{0,90}(X)$ - 1. až 9. decil,
$K_{0,01}(X), \ldots, K_{0,99}(X)$ - 1. až 99. percentil.
Kterýkoliv α-kvantil je charakteristikou polohy číselných realizací náhodné veličiny na číselné ose.
Charakteristika variability: kvartilová odchylka $q = K_{0,75}(X) - K_{0,25}(X)$.

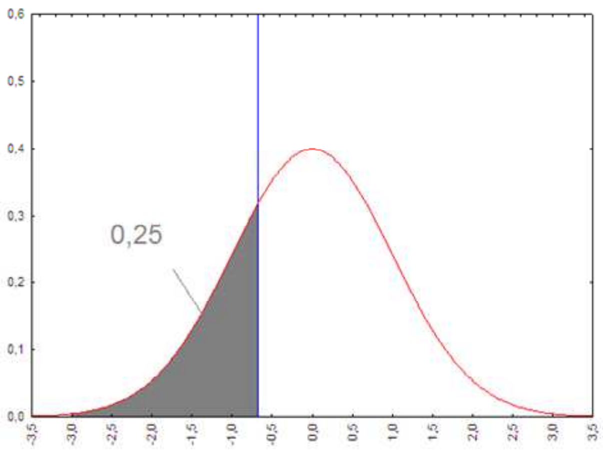
Ilustrace



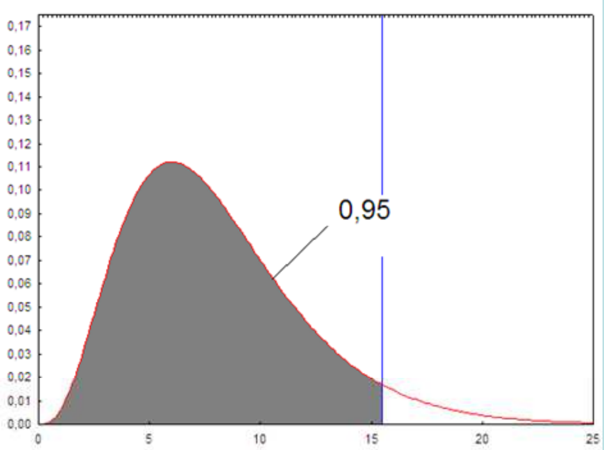
Označení pro kvantily speciálních rozložení

$X \sim N(0, 1) \Rightarrow K_\alpha(X) = u_\alpha$,
$X \sim \chi^2(n) \Rightarrow K_\alpha(X) = \chi^2_\alpha(n)$,
$X \sim t(n) \Rightarrow K_\alpha(X) = t_\alpha(n)$,
$X \sim F(n_1, n_2) \Rightarrow K_\alpha(X) = F_\alpha(n_1, n_2)$.
Tyto kvantily najdeme ve statistických tabulkách.
Používáme vztahy:
$u_\alpha = - u_{1-\alpha}$,
$t_\alpha(n) = - t_{1-\alpha}(n)$,
$F_\alpha(n_1, n_2) = \dfrac{1}{F_{1-\alpha}(n_2, n_1)}$.

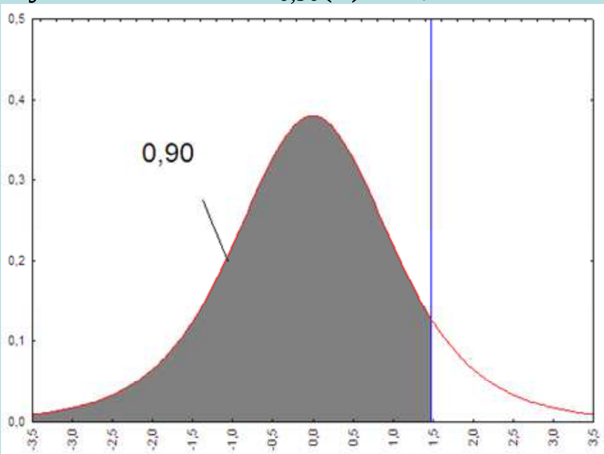
Význam kvantilu $u_{0,25} = -0,6745$



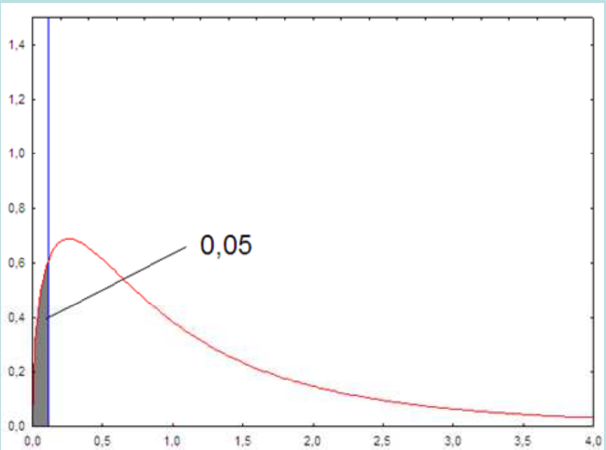
Význam kvantilu $\chi^2_{0,95}(8) = 15,5073$



Význam kvantilu $t_{0,90}(5) = 1,4759$



Význam kvantilu $F_{0,05}(3,7) = \dfrac{1}{F_{0,95}(7,3)} = \dfrac{1}{8,8867} = 0,1125$



Příklad: Nechť U ~ N(0, 1). Pomocí systému STATISTICA najděte 2. decil a první a poslední percentil.

První možnost: Použijeme Pravděpodobnostní kalkulátor. Do okénka průměr napíšeme 0, do okénka Sm. Odch. napíšeme 1, do okénka p napíšeme pro 2. decil 0,2, pro první percentil 0,01 a pro poslední percentil 0,99. V okénku X se objeví -0,841621 pro 2. decil, -2,326348 pro první percentil a 2,326348 pro poslední percentil.

Ilustrace pro poslední percentil:

Šedá plocha pod grafem hustoty má velikost 0,99 a hodnota distribuční funkce v bodě 2,326348 je 0,99 (značeno šrafovaně).

Druhá možnost: Otevřeme nový datový soubor o třech proměnné a jednom případu.
Do dlouhého jména první proměnné napíšeme =VNormal(0,2;0;1). Dostaneme -0,841621.
Do dlouhého jména druhé proměnné napíšeme =VNormal(0,01;0;1). Dostaneme -2,326348.
Do dlouhého jména třetí proměnné napíšeme =VNormal(0,99;0;1). Dostaneme 2,326348.

Příklad: Nechť X ~ N(-1, 4). Pomocí systému STATISTICA najděte horní kvartil.

První možnost: Spustíme Pravděpodobnostní kalkulátor, vybereme Rozdělení Normální. Do okénka průměr napíšeme -1, do okénka Sm. Odch. napíšeme 2, do okénka p napíšeme 0,75 a v okénku X se objeví 0,34898.
Druhá možnost: Otevřeme nový datový soubor o jedné proměnné a jednom případu.
Do dlouhého jména této proměnné napíšeme =VNormal(0,75;-1;2). Dostaneme 0,34898.

Příklad: Pomocí systému STATISTICA určete $\chi^2_{0,05}(5)$.

První možnost: Spustíme Pravděpodobnostní kalkulátor, vybereme Rozdělení Chi 2. Do okénka sv. napíšeme 5 a do okénka p napíšeme 0,05. V okénku Chi 2 se objeví 1,145476.
Druhá možnost: Otevřeme nový datový soubor o jedné proměnné a jednom případu.
Do dlouhého jména této proměnné napíšeme =VChi2(0,05;5). Dostaneme 1,145476.

Příklad: Pomocí systému STATISTICA určete $t_{0,975}(18)$ a $t_{0,01}(4)$.

První možnost: Spustíme Pravděpodobnostní kalkulátor, vybereme Rozdělení t (Studentovo). Do okénka sv. napíšeme 18 (resp. 4) a do okénka p napíšeme 0,975 (resp. 0,01). V okénku t se objeví 2,100922 (resp. -3,746947).

Druhá možnost: Otevřeme nový datový soubor o jedné proměnné a jednom případu.
Do Dlouhého jména této proměnné napíšeme =VStudent(0,975;18) (resp. VStudent(0,01;4)). Dostaneme 2,100922 (resp. -3,746947).

Příklad: Pomocí systému STATISTICA určete $F_{0,975}(3, 12)$ a $F_{0,05}(18, 20)$.

První možnost: Spustíme Pravděpodobnostní kalkulátor, vybereme Rozdělení F (Fisherovo). Do okénka sv1 napíšeme 3 (resp. 18), do okénka sv2 napíšeme 12 (resp. 20) a do okénka p napíšeme 0,975 (resp. 0,05).
V okénku F se objeví 4,474185 (resp. 0,456486).

Druhá možnost: Otevřeme nový datový soubor o jedné proměnné a dvou případech
Do Dlouhého jména první proměnné napíšeme =VF(0,975;3;12), do Dlouhého jména druhé proměnné napíšeme =VF(0,05;18;20).Dostaneme 4,474185 (resp. 0,456486).

Číselné charakteristiky diskrétních a spojitých náhodných veličin aspoň intervalového typu

Charakteristika polohy: střední hodnota E(X) – číslo, které charakterizuje polohu realizací náhodné veličiny na číselné ose s přihlédnutím k jejich pravděpodobnostem.

Diskrétní případ: náhodná veličina X má pravděpodobnostní funkci $\pi(x)$.

Střední hodnota $E(X) = \sum_{x=-\infty}^{\infty} x\pi(x)$, pokud je suma vpravo konečná.

Fyzikální význam: střední hodnota je těžiště soustavy hmotných bodů, jejichž celková hmotnost je 1 a bod o souřadnici x má hmotnost $\pi(x)$.

Spojitý případ: náhodná veličina X má hustotu pravděpodobnosti $\varphi(x)$.

Střední hodnota $E(X) = \int_{-\infty}^{\infty} x\varphi(x)dx$, pokud je integrál vpravo konečný.

Fyzikální význam: střední hodnota je těžiště hmotné přímky, jejíž celková hmotnost je 1 a hmota je na přímce rozprostřena podle předpisu φ (x).

Centrovaná náhodná veličina: Y = X - E(X).
(Pro náhodnou veličinu Y platí: E(Y) = 0.)

Střední hodnota transformované náhodné veličiny Y = g(X)

$$E(Y) = \left\langle \begin{array}{l} \sum\limits_{x=-\infty}^{\infty} g(x)\pi(x) \\ \int\limits_{-\infty}^{\infty} g(x)\varphi(x)dx \end{array} \right|$$

Střední hodnota transformované náhodné veličiny Y = g(X_1, X_2)

$$E(Y) = \left\langle \begin{array}{l} \sum\limits_{x_1=-\infty}^{\infty} \sum\limits_{x_2=-\infty}^{\infty} g(x_1, x_2)\pi(x_1, x_2) \\ \int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} g(x_1, x_2)\varphi(x_1, x_2)dx_1 dx_2 \end{array} \right|$$

Charakteristika variability: rozptyl D(X) - číslo, které charakterizuje proměnlivost realizací náhodné veličiny kolem její střední hodnoty s přihlédnutím k jejich pravděpodobnostem.

Definiční vzorec: $D(X) = E\left([X - E(X)]^2\right)$ (rozptyl je střední hodnota kvadrátu centrované náhodné veličiny).

Výpočetní vzorec: $D(X) = E\left(X^2\right) - [E(X)]^2$ (rozptyl je střední hodnota kvadrátu mínus kvadrát středních hodnot).

$$D(X) = \left\langle \begin{array}{l} \sum\limits_{x=-\infty}^{\infty} x^2 \pi(x) - \left[\sum\limits_{x=-\infty}^{\infty} x\pi(x)\right]^2 \\ \int\limits_{-\infty}^{\infty} x^2 \varphi(x)dx - \left[\int\limits_{-\infty}^{\infty} x\varphi(x)dx\right]^2 \end{array} \right|$$

Směrodatná odchylka $\sqrt{D(X)}$ - vyjadřuje průměrnou variabilitu realizací náhodné veličiny X kolem její střední hodnoty.

Standardizovaná náhodná veličina: $Z = \frac{X - E(X)}{\sqrt{D(X)}}$

(Pro náhodnou veličinu Z platí: E(Z) = 0, D(Z) = 1.)

Příklad na výpočet střední hodnoty a rozptylu diskrétní náhodné veličiny: Střelec střílí do terče až do prvního zásahu. Má v zásobě 4 náboje. Pravděpodobnost zásahu je při každém výstřelu 0,6. Náhodná veličina X udává počet nespotřebovaných nábojů. Vypočtěte střední hodnotu a rozptyl náhodné veličiny X.

Řešení:

X nabývá hodnot 0, 1, 2, 3 a její pravděpodobnostní funkce je

$\pi(0) = P(X=0) = 0{,}4^4 + 0{,}4^3 * 0{,}6 = 0{,}064$,

$\pi(1) = P(X=1) = 0{,}4^2 * 0{,}6 = 0{,}096$,

$\pi(2) = P(X=2) = 0{,}4 * 0{,}6 = 0{,}24$,

$\pi(3) = P(X=3) = 0{,}6$,

$\pi(x) = 0$ jinak

$E(X) = 0 * 0{,}064 + 1 * 0{,}096 + 2 * 0{,}24 + 3 * 0{,}6 = 2{,}376$

$D(X) = 0^2 * 0{,}064 + 1^2 * 0{,}096 + 2^2 * 0{,}24 + 3^2 * 0{,}6 - 2{,}376^2 = 0{,}8106$

Postup ve STATISTICE:

Otevřeme nový datový soubor o dvou proměnných X a cetnost a čtyřech případech. Do proměnné X napíšeme 0, 1, 2, 3, do proměnné cetnost napíšeme 64, 96, 240, 600.
Statistiky – Základní statistiky/tabulky – Popisné statistiky – OK – zavedeme proměnnou vah cetnost – OK - Proměnné X – OK – Detailní výsledky - zaškrtneme Průměr, Rozptyl – Výpočet.

	Popisné statistiky (Tabulka10)		
Proměnná	N platných	Průměr	Rozptyl
X	1000	2,376000	0,811435

Rozptyl však musíme upravit, musíme ho přenásobit číslem 999/1000. Do výstupní tabulky tedy přidáme za proměnnou Rozptyl novou proměnnou a do jejího Dlouhého jména napíšeme
=v3*999/1000

	Popisné statistiky (Tabulka10)			
Proměnná	N platných	Průměr	Rozptyl	NProm
X	1000	2,376000	0,811435	0,810624

Střední hodnoty a rozptyly vybraných diskrétních a spojitých rozložení

$X \sim Dg(\mu) \Rightarrow E(X) = \mu, D(X) = 0$

$X \sim A(\vartheta) \Rightarrow E(X) = \vartheta, D(X) = \vartheta(1-\vartheta)$

$X \sim Bi(n, \vartheta) \Rightarrow E(X) = n\vartheta, D(X) = n\vartheta(1-\vartheta)$

$X \sim Ge(\vartheta) \Rightarrow E(X) = \frac{1-\vartheta}{\vartheta}, \; D(X) = \frac{1-\vartheta}{\vartheta^2}$

$X \sim Hg(N, M, n) \Rightarrow E(X) = \frac{M}{N}n, \; D(X) = \frac{Mn}{N}\left(1-\frac{M}{N}\right)\frac{N-n}{N-1}$

$X \sim Po(\lambda) \Rightarrow E(X) = \lambda, \; D(X) = \lambda$

$X \sim Rd(G) \Rightarrow E(X) = \frac{n+1}{2}, \; D(X) = \frac{n^2-1}{12}$

$X \sim Rs(a, b) \Rightarrow E(X) = \frac{a+b}{2}, D(X) = \frac{(b-a)^2}{12}$

$X \sim Ex(\lambda) \Rightarrow E(X) = \frac{1}{\lambda}, \; D(X) = \frac{1}{\lambda^2}$

$X \sim N(\mu, \sigma^2) \Rightarrow E(X) = \mu, D(X) = \sigma^2$

$X \sim \chi^2(n) \Rightarrow E(X) = n, \; D(X) = 2n$

$X \sim t(n) \Rightarrow E(X) = 0$ pro $n \geq 2$ (pro $n = 1$ střední hodnota neexistuje), $D(X) = \frac{n}{n-2}$ pro $n \geq 3$ (pro $n = 1,2$ rozptyl neexistuje).

$X \sim F(n_1, n_2) \Rightarrow E(X) = \frac{n_2}{n_2 - 2}$ pro $n_2 \geq 3$ (pro $n_2 = 1,2$ střední hodnota neexistuje), $D(X) = \frac{2n_2^{\,2}(n_1 + n_2 - 2)}{n_1(n_2-2)^2(n_2-4)}$ pro $n_2 \geq 5$ (pro $n_2 = 1,2,3,4$ rozptyl neexistuje).

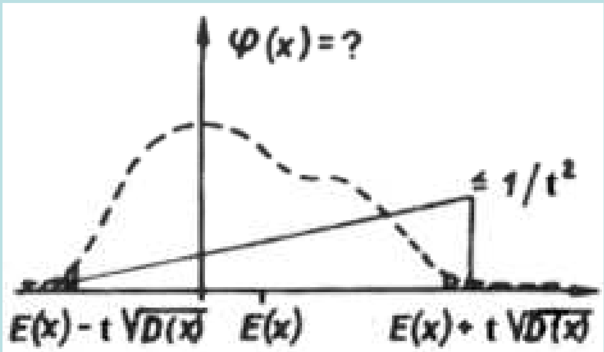
Čebyševova nerovnost:

Jestliže náhodná veličina X má střední hodnotu E(X) a rozptyl D(X), pak

$$\forall t > 0 : P\left(\left|X - E(X)\right| > t\sqrt{D(X)}\right) \leq \frac{1}{t^2}.$$

(Význam: pokud neznáme rozložení náhodné veličiny, ale známe její střední hodnotu a rozptyl, pak můžeme odhadnout pravděpodobnost, s jakou se od své střední hodnoty odchýlí o více než t-násobek své směrodatné odchylky.)

Ilustrace:



Příklad: Nechť $E(X) = \mu$, $D(X) = \sigma^2$.

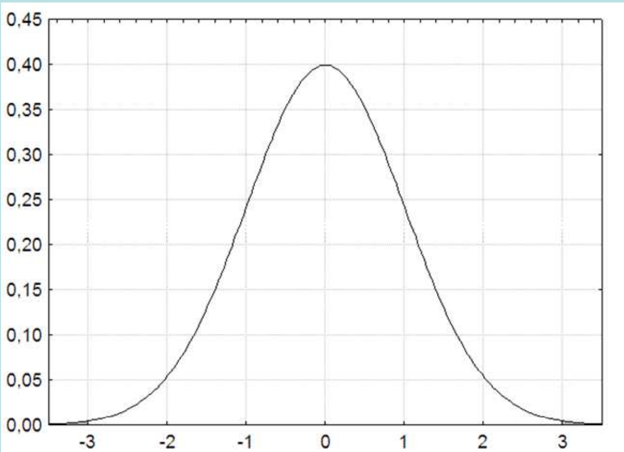
a) Odhadněte $P(|X - \mu| > 3\sigma)$.

b) Jestliže $X \sim N(\mu, \sigma^2)$, vypočtěte $P(|X - \mu| > 3\sigma)$.

Řešení:

ad a) $P(|X - \mu| > 3\sigma) \leq \frac{1}{3^2} = \frac{1}{9} = 0,\bar{1}$. (Tento výsledek je znám jako pravidlo 3σ a říká, že nejvýše 11,1% realizací náhodné veličiny leží vně intervalu $(\mu - 3\sigma, \mu + 3\sigma)$.)

ad b) $P(|X - \mu| > 3\sigma) = 1 - P(-3\sigma \leq X - \mu \leq 3\sigma) = 1 - P(-3 \leq \frac{X - \mu}{\sigma} \leq 3) = 1 - \Phi(3) + \Phi(-3) = 2[1 - \Phi(3)] =$
$= 2(1 - 0,99865) = 0,0027$. (Má-li náhodná veličina normální rozložení, pak pouze 0,27% realizací leží vně intervalu $(\mu - 3\sigma, \mu + 3\sigma)$.)



Charakteristika společné variability: kovariance $C(X_1, X_2)$ – číslo, které charakterizuje variabilitu realizací dvou náhodných veličin X_1, X_2 kolem jejich středních hodnot s přihlédnutím k pravděpodobnostem těchto realizací.

Definiční vzorec: $C(X_1, X_2) = E([X_1 - E(X_1)][X_2 - E(X_2)])$ (kovariance je střední hodnota součinu centrovaných náhodných veličin).

Výpočetní vzorec: $C(X_1, X_2) = E(X_1 X_2) - E(X_1)E(X_2)$ (kovariance je střední hodnota součinu mínus součin středních hodnot).

$$C(X_1, X_2) = \left\langle \begin{array}{l} \sum\limits_{x_1=-\infty}^{\infty} \sum\limits_{x_2=-\infty}^{\infty} x_1 x_2 \pi(x_1, x_2) - \sum\limits_{x_1=-\infty}^{\infty} x_1 \pi_1(x_1) \sum\limits_{x_2=-\infty}^{\infty} x_2 \pi_2(x_2) \\ \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} x_1 x_2 \varphi(x_1, x_2) dx_1 dx_2 - \int\limits_{-\infty}^{\infty} x_1 \varphi_1(x_1) dx_1 \int\limits_{-\infty}^{\infty} x_2 \varphi_2(x_2) dx_2 \end{array} \right|$$

Význam kovariance: Je-li kovariance kladná (záporná), pak to svědčí o existenci jistého stupně přímé (nepřímé) lineární závislosti mezi realizacemi náhodných veličin X_1, X_2. Je-li kovariance nulová, pak říkáme, že náhodné veličiny X_1, X_2 jsou nekorelované a znamená to, že mezi jejich realizacemi není žádný lineární vztah. Pozor – z nekorelovanosti nevyplývá stochastická nezávislost, zatímco ze stochastické nezávislosti plyne nekorelovanost.

Charakteristika těsnosti lineárního vztahu: koeficient korelace $R(X_1, X_2)$ - číslo, které charakterizuje těsnost lineární závislosti realizací náhodných veličin X_1, X_2. Čím bližší je 1, tím těsnější je přímá lineární závislost, čím bližší je -1, tím těsnější je nepřímá lineární závislost.

Definiční vzorec: $$R(X_1, X_2) = E\left(\frac{X_1 - E(X_1)}{\sqrt{D(X_1)}} \cdot \frac{X_2 - E(X_2)}{\sqrt{D(X_2)}}\right)$$ pro kladné směrodatné odchylky, jinak klademe $R(X_1, X_2) = 0$ (koeficient korelace je střední hodnota součinu standardizovaných náhodných veličin).

Výpočetní vzorec: $$R(X_1, X_2) = \frac{C(X_1, X_2)}{\sqrt{D(X_1)}\sqrt{D(X_2)}}$$ (koeficient korelace je podíl kovariance a součinu směrodatných odchylek).

Příklad na výpočet koeficientu korelace diskrétního náhodného vektoru: Náhodná veličina X udává počet skleniček tvrdého alkoholu, které během jedné hodiny pobytu v restauraci vypije náhodně vybraný host a náhodná veličina Y udává počet vykouřených cigaret za 1 h u téhož hosta. Hodnoty simultánní pravděpodobnostní funkce $\pi(x, y)$ jsou dány v kontingenční tabulce:

X	Y			
	0	1	2	$\pi_1(x)$
1	0,02	0,08	0,05	0,15
2	0,02	0,28	0,25	0,55
3	0,01	0,13	0,16	0,30
$\pi_2(y)$	0,05	0,49	0,46	1

Vypočtěte koeficient korelace náhodných veličin X, Y.

Řešení:

E(X) = 1.0,15 + 2.0,55 + 3.0,30 = 2,15

E(Y) = 0.0,05 + 1.0,49 + 2.0,46 = 1,41

D(X) = 1^2.0,15 + 2^2.0,55 + 3^2.0,30 – $2{,}15^2$ = 0,4275

D(Y) = 0^2.0,05 + 1^2.0,49 + 2^2.0,46 – $1{,}41^2$ = 0,3419

C(X,Y) = 1.0.0,02 + 1.1.0,08 + … + 3.2.0,16 – 2,15.1,41 = 0,0585,

R(X,Y) = 0,0585/√0,4275√0,3419 = 0,153.

Vidíme, že lineární závislost mezi počtem vypitých skleniček tvrdého alkoholu a počtem vykouřených cigaret za 1 h pobytu v restauraci je jenom slabá.

Postup ve STATISTICE:

Vytvoříme nový datový soubor o třech proměnných X, Y, cetnost a 16 případech. Do proměnné X napíšeme 1, 1, 1, 2, 2, 2, 3, 3, 3, do proměnné Y 3x pod sebe 0, 1, 2 a do proměnné cetnost 2, 8, 5, 2, 28, 25, 1, 13, 16. Statistiky - Základní statistiky/tabulky – zavedeme proměnnou vah cetnost – OK - Korelační matice – OK – 1 seznam proměnných – X, Y – OK.

	Korelace (alkohol_cigarety.sta) Označ. korelace jsou významné na hlad. p < ,05000 N=100 (Celé případy vynechány u ChD)	
Proměnná	X	Y
X	1,00	0,15
Y	0,15	1,00

Vlastnosti střední hodnoty

a) $E(a) = a$

b) $E(a + bX) = a + bE(X)$

c) $E(X - E(X)) = 0$

d) $E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i)$

e) Jsou-li náhodné veličiny $X_1, ..., X_n$ stochasticky nezávislé, pak $E\left(\prod_{i=1}^{n} X_i\right) = \prod_{i=1}^{n} E(X_i)$

Vlastnosti kovariance

a) $C(a_1, X_2) = C(X_1, a_2) = C(a_1, a_2) = 0$

b) $C(a_1 + b_1X_1, a_2 + b_2X_2) = b_1b_2C(X_1, X_2)$

c) $C(X, X) = D(X)$

d) $C(X_1, X_2) = C(X_2, X_1)$

e) $C(X_1, X_2) = E(X_1X_2) - E(X_1)E(X_2)$

f) $C\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} C(X_i, Y_j)$

Vlastnosti rozptylu

a) $D(a) = 0$

b) $D(a + bX) = b^2 D(X)$

c) $D(X) = E(X^2) - [E(X)]^2$

d) $D\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} D(X_i) + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} C(X_i, X_j)$ (jsou-li náhodné veličiny $X_1, \ldots, X_n$ nekorelované, pak $D\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} D(X_i)$)

Vlastnosti koeficientu korelace

a) $R(a_1, X_2) = R(X_1, a_2) = R(a_1, a_2) = 0$

b) $R(a_1 + b_1X_1, a_2 + b_2X_2) = \mathrm{sgn}(b_1b_2)\, R(X_1, X_2)$

c) $R(X, X) = 1$ pro $D(X) \neq 0$, $R(X, X) = 0$ jinak

d) $R(X_1, X_2) = R(X_2, X_1)$

e) $R(X_1, X_2) = \begin{cases} \dfrac{C(X_1, X_2)}{\sqrt{D(X_1)}\sqrt{D(X_2)}} & \text{pro } \sqrt{D(X_1)}\sqrt{D(X_2)} > 0 \\ 0 & \text{jinak} \end{cases}$

f) $|R(X_1, X_2)| \leq 1$ a rovnost nastane tehdy a jen tehdy, když mezi veličinami X_1, X_2 existuje s pravděpodobností 1 úplná lineární závislost, tj. existují konstanty a_1, a_2 tak, že $P(X_2 = a_1 + a_2X_1) = 1$. (Uvedená nerovnost se nazývá Cauchyova – Schwarzova – Buňakovského nerovnost.)

Příklad na využití vlastností číselných charakteristik: Náhodné veličiny X, Y jsou náhodné chyby, které vznikají na vstupním zařízení. Mají střední hodnoty $E(X) = -2$, $E(Y) = 4$ a rozptyly $D(X) = 4$, $D(Y) = 9$. Koeficient korelace těchto chyb je $R(X,Y) = -0{,}5$. Chyba na výstupu zařízení souvisí s chybami na vstupu funkční závislostí $Z = 3X^2 – 2XY + Y^2 - 3$. Najděte střední hodnotu chyby na výstupu.

Řešení:

$E(Z) = E(3X^2 – 2XY + Y^2 – 3) = 3E(X^2) – 2E(XY) + E(Y^2) – E(3) =$

$= 3\{D(X) + [E(X)]^2\} – 2[C(X,Y) + E(X)E(Y)] + D(Y) + [E(Y)]^2 – 3 =$

$= 3[D(X) + [E(X)]^2] – 2[R(X,Y)\sqrt{D(X)}\sqrt{D(Y)} + E(X)E(Y)] + D(Y) + [E(Y)]^2 - 3 =$

$3(4 + 4) -2[-0{,}5\times 2\times 3 + (-2) \times 4] + 9 + 16 – 3 = 24 + 22 + 25 – 3 = 68$

Centrální limitní věta:

Jsou-li náhodné veličiny $X_1, \ldots, X_n$ stochasticky nezávislé a všechny mají stejné rozložení se střední hodnotou μ a rozptylem σ^2, pak pro velká n ($n \geq 30$) lze rozložení součtu $\sum_{i=1}^{n} X_i$ aproximovat normálním rozložením $N(n\mu, n\sigma^2)$. Zkráceně píšeme

$$\sum_{i=1}^{n} X_i \approx N\left(n\mu, n\sigma^2\right).$$

Pokud součet $\sum_{i=1}^{n} X_i$ standardizujeme, tj. vytvoříme náhodnou veličinu $U_n = \dfrac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}}$, pak rozložení této náhodné veličiny lze aproximovat standardizovaným normálním rozložením. Zkráceně píšeme $U_n \approx N(0,1)$

Normální rozložení je tedy rozložením limitním, k němuž se blíží všechna rozložení, proto hraje velmi důležitou roli v počtu pravděpodobnosti a matematické statistice.

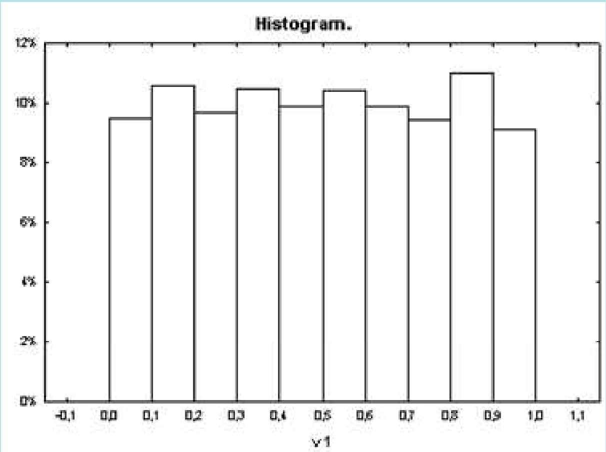
Ilustrace centrální limitní věty:

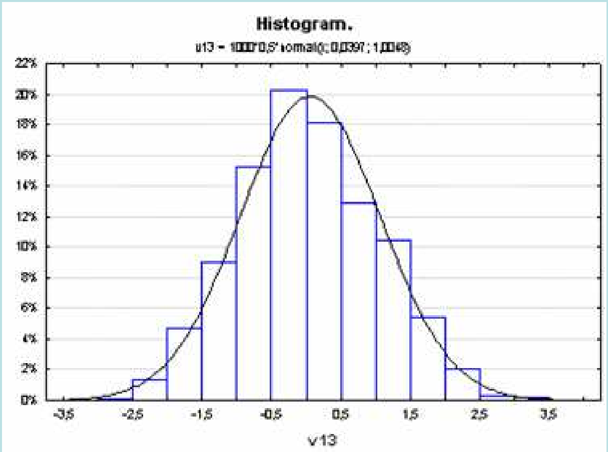
Uvažme n stochasticky nezávislých náhodných veličin $X_1, \dots, X_n$, přičemž každá z nich má rovnoměrné spojité rozložení na intervalu (0,1), tj. $X_i \sim Rs(0,1)$, $i = 1, \dots, n$. Protože $E(X_i) = \frac{1}{2}$ a $D(X_i) = \frac{1}{12}$, podle centrální limitní věty

náhodná veličina $U_n = \dfrac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} = \dfrac{\sum_{i=1}^{n} X_i - \frac{n}{2}}{\sqrt{\frac{n}{12}}} \approx N(0,1)$.

Položíme-li n = 12, pak $U_{12} = \sum_{i=1}^{12} X_i - 6 \approx N(0,1)$.

Při ilustraci působení centrální limitní věty na počítači postupujeme tak, že pro každou z 12 veličin $X_i \sim Rs(0,1)$, $i = 1, \dots, 12$ vygenerujeme dostatečně velký počet realizací , např. 1000 a uložíme je do proměnných v_1 až v_{12}. Do proměnné v_{13} uložíme součet proměnných v_1 až v_{12} zmenšený o 6. Histogram kterékoliv proměnné v_1 až v_{12} se svým tvarem bude blížit obdélníku, zatímco histogram proměnné v_{13} se svým tvarem bude blížit Gaussově křivce.





Důsledkem centrální limitní věty je **Moivreova – Laplaceova věta:**
Nechť $X_1, \ldots, X_n$ jsou stochasticky nezávislé náhodné veličiny, všechny se řídí alternativním rozložením A(ϑ). Pak jejich součet $Y_n = \sum_{i=1}^{n} X_i$ má binomické rozložení Bi(n, ϑ). Střední hodnota veličiny Y_n je $E(Y_n) = n\vartheta$, rozptyl je $D(Y_n) = n\vartheta(1-\vartheta)$.

Podle centrální limitní věty se standardizovaná náhodná veličina $\frac{Y_n - n\vartheta}{\sqrt{n\vartheta(1-\vartheta)}}$ asymptoticky řídí standardizovaným normálním rozložením N(0,1).

Aproximace se považuje za vyhovující, když jsou splněny podmínky: $\frac{1}{n+1} < \vartheta < \frac{n}{n+1}$ a $n\vartheta(1-\vartheta) > 9$.

Na základě Moivreovy – Laplaceovy věty se používá aproximativní vzorec, který složitý výpočet distribuční funkce binomického rozložení nahrazuje jednoduchým hledáním v tabulkách hodnot distribuční funkce standardizovaného normálního rozložení.

Máme náhodnou veličinu $Y_n \sim$ Bi(n, ϑ). Pak

pravděpodobnostní funkce $P(Y_n = y) = \binom{n}{y}\vartheta^y(1-\vartheta)^{n-y}$ pro y = 0, 1, …, n,

distribuční funkce $P(Y \le y) = \sum_{t=0}^{y} P(Y = t) = \sum_{t=0}^{y}\binom{n}{t}\vartheta^t(1-\vartheta)^{n-t}$ - složitý výpočet

Aproximativní vzorec: $P(Y_n \le y) \approx \Phi\left(\frac{y - n\vartheta}{\sqrt{n\vartheta(1-\vartheta)}}\right)$.

Příklad na aplikaci Moivreovy – Laplaceovy věty: 100 x nezávisle na sobě hodíme hrací kostkou. Jaká je pravděpodobnost, že šestka padne aspoň 20 x?

Řešení:

Y_{100} – počet šestek ve 100 hodech, $Y_{100} \sim Bi(100, \frac{1}{6})$.

Ověření podmínek dobré aproximace: $\frac{1}{n+1} < \vartheta < \frac{n}{n+1}$ a $n\vartheta(1-\vartheta) > 9$

$\frac{1}{101} < \frac{1}{6} < \frac{100}{101}$ a $100 \cdot \frac{1}{6} \cdot \left(1 - \frac{1}{6}\right) = 13{,}\overline{8} > 9$.

Aproximativní výpočet:

$$P(Y_{100} \geq 20) = 1 - P(Y_{100} \leq 19) = 1 - P\left(\frac{Y_{100} - \frac{100}{6}}{\sqrt{100 \cdot \frac{1}{6} \cdot \frac{5}{6}}} \leq \frac{19 - \frac{100}{6}}{\sqrt{100 \cdot \frac{1}{6} \cdot \frac{5}{6}}}\right) = 1 - P(U_{100} \leq 0{,}626) \approx$$

$$\approx 1 - \Phi(0{,}626) = 1 - 0{,}73565 = 0{,}26435$$

Přesný výpočet:

$$P(Y_{100} \geq 20) = 1 - P(Y_{100} \leq 19) = 1 - \sum_{t=0}^{19} \binom{100}{t} \left(\frac{1}{6}\right)^{t} \left(\frac{5}{6}\right)^{100-t} = 0{,}219751$$

Příklad na aplikaci Moivreovy – Laplaceovy věty: Osobě prohlašující, že má proutkařské schopnosti, předložíme 100 dvojic zakrytých nádob. V každé dvojici je jedna nádoba prázdná a druhá naplněná vodou. Výsledky proutkaře srovnáme s výsledky hypotetické osoby, která pracuje zcela náhodně. Nechť náhodná veličina Y_{100} udává počet úspěšně identifikovaných dvojic nádob. Jaká je pravděpodobnost, že Y_{100} překročí přirozené číslo y, y = 0, 1, …, 100?

Řešení: Je zřejmé, že $Y_{100} \sim \mathrm{Bi}\left(100, \frac{1}{2}\right)$.

$$P(Y_{100} > y) = 1 - P(Y_{100} \le y) = 1 - P\left(\frac{Y_{100} - 100 \cdot 0{,}5}{\sqrt{100 \cdot 0{,}5 \cdot (1 - 0{,}5)}} \le \frac{y - 100 \cdot 0{,}5}{\sqrt{100 \cdot 0{,}5 \cdot (1 - 0{,}5)}}\right) = 1 - P\left(\frac{Y_{100} - 50}{5} \le \frac{y - 50}{5}\right) \approx 1 - \Phi\left(\frac{y - 50}{5}\right)$$

y = 50: $P(Y_{100} > 50) \approx 1 - \Phi(0) = 1 - 0{,}5 = 0{,}5$

y = 55: $P(Y_{100} > 55) \approx 1 - \Phi(1) = 1 - 0{,}84134 = 0{,}15866$

y = 60: $P(Y_{100} > 60) \approx 1 - \Phi(2) = 1 - 0{,}97725 = 0{,}02275$

y = 65: $P(Y_{100} > 65) \approx 1 - \Phi(3) = 1 - 0{,}99865 = 0{,}00135$

Ilustrace: závislost $P(Y_{100} > y)$ na y