Matematika IV – 11. přednáška
Normální rozdělení, limitní vlastnosti, zákony velkých čísel

Michal Bulant

Masarykova univerzita
Fakulta informatiky

4. 5. 2011

Obsah přednášky

Doporučené zdroje

- Martin Panák, Jan Slovák, **Drsná matematika**, *e*-text.
- Karel Zvára, Josef Štěpán, **Pravděpodobnost a matematická statistika**, Matfyzpress, 4. vydání, 2006, 230 stran, ISBN 80-867-3271-1.
- Marie Budíková, Štěpán Mikoláš, Pavel Osecký, **Popisná statistika**, Masarykova univerzita, 3. vydání, 2002, 48 stran, ISBN 80-210-1831-3.
- Marie Budíková, Štěpán Mikoláš, Pavel Osecký, **Teorie pravděpodobnosti a matematická statistika (sbírka příkladů)**, Masarykova univerzita, 3. vydání, 2004, 117 stran, ISBN 80-210-3313-4.

Plán přednášky

Charakteristiky náhodných veličin – připomenutí

- střední hodnota $E(X)$,

Charakteristiky náhodných veličin – připomenutí

- střední hodnota $E(X)$,
- rozptyl $D(X) = E([X - E(X)]^2)$, směrodatná odchylka $\sqrt{D(X)}$

Charakteristiky náhodných veličin – připomenutí

- střední hodnota $E(X)$,
- rozptyl $D(X) = E([X - E(X)]^2)$, směrodatná odchylka $\sqrt{D(X)}$
- kovariance $C(X, Y) = E([X - E(X)][Y - E(Y)])$, korelační koeficient $R(X, Y) = C(X, Y)/(\sqrt{D(X)}\sqrt{D(Y)})$, Cauchyova nerovnost $|R(X, Y)| \leq 1$,

Charakteristiky náhodných veličin – připomenutí

- střední hodnota $E(X)$,
- rozptyl $D(X) = E([X - E(X)]^2)$, směrodatná odchylka $\sqrt{D(X)}$
- kovariance $C(X, Y) = E([X - E(X)][Y - E(Y)])$, korelační koeficient $R(X, Y) = C(X, Y)/(\sqrt{D(X)}\sqrt{D(Y)})$, Cauchyova nerovnost $|R(X, Y)| \leq 1$,
- kvantily,

Charakteristiky náhodných veličin – připomenutí

- střední hodnota $E(X)$,
- rozptyl $D(X) = E([X - E(X)]^2)$, směrodatná odchylka $\sqrt{D(X)}$
- kovariance $C(X, Y) = E([X - E(X)][Y - E(Y)])$, korelační koeficient $R(X, Y) = C(X, Y)/(\sqrt{D(X)}\sqrt{D(Y)})$, Cauchyova nerovnost $|R(X, Y)| \leq 1$,
- kvantily,
- další momenty (obecné, centrální) - momentová vytvořující funkce $M_X(t) = E(e^{tX})$

Charakteristiky náhodných veličin – připomenutí

- střední hodnota $E(X)$,
- rozptyl $D(X) = E([X - E(X)]^2)$, směrodatná odchylka $\sqrt{D(X)}$
- kovariance $C(X, Y) = E([X - E(X)][Y - E(Y)])$, korelační koeficient $R(X, Y) = C(X, Y)/(\sqrt{D(X)}\sqrt{D(Y)})$, Cauchyova nerovnost $|R(X, Y)| \leq 1$,
- kvantily,
- další momenty (obecné, centrální) - momentová vytvořující funkce $M_X(t) = E(e^{tX})$

Věta

Charakteristiky náhodných veličin – připomenutí

- střední hodnota $E(X)$,
- rozptyl $D(X) = E([X - E(X)]^2)$, směrodatná odchylka $\sqrt{D(X)}$
- kovariance $C(X, Y) = E([X - E(X)][Y - E(Y)])$, korelační koeficient $R(X, Y) = C(X, Y)/(\sqrt{D(X)}\sqrt{D(Y)})$, Cauchyova nerovnost $|R(X, Y)| \leq 1$,
- kvantily,
- další momenty (obecné, centrální) - momentová vytvořující funkce $M_X(t) = E(e^{tX})$

Věta

- *Pro nezávislé náhodné veličiny platí* $M_{X+Y}(t) = M_X(t)M_Y(t)$.

Charakteristiky náhodných veličin – připomenutí

- střední hodnota $E(X)$,
- rozptyl $D(X) = E([X - E(X)]^2)$, směrodatná odchylka $\sqrt{D(X)}$
- kovariance $C(X, Y) = E([X - E(X)][Y - E(Y)])$, korelační koeficient $R(X, Y) = C(X, Y)/(\sqrt{D(X)}\sqrt{D(Y)})$, Cauchyova nerovnost $|R(X, Y)| \leq 1$,
- kvantily,
- další momenty (obecné, centrální) - momentová vytvořující funkce $M_X(t) = E(e^{tX})$

Věta

- *Pro nezávislé náhodné veličiny platí* $M_{X+Y}(t) = M_X(t)M_Y(t)$.
- *r-tý obecný moment* μ'_r *náhodné veličiny* X *je koeficient u* $\frac{t^r}{r!}$ *v rozvoji* M_X *do exponenciální mocninné řady (tedy např.* $EX = \mu'_1, DX = \mu'_2 - (\mu'_1)^2$*).*

Charakteristiky náhodných veličin – připomenutí

- střední hodnota $E(X)$,
- rozptyl $D(X) = E([X - E(X)]^2)$, směrodatná odchylka $\sqrt{D(X)}$
- kovariance $C(X, Y) = E([X - E(X)][Y - E(Y)])$, korelační koeficient $R(X, Y) = C(X, Y)/(\sqrt{D(X)}\sqrt{D(Y)})$, Cauchyova nerovnost $|R(X, Y)| \leq 1$,
- kvantily,
- další momenty (obecné, centrální) - momentová vytvořující funkce $M_X(t) = E(e^{tX})$

Věta

- *Pro nezávislé náhodné veličiny platí* $M_{X+Y}(t) = M_X(t)M_Y(t)$.
- *r-tý obecný moment* μ'_r *náhodné veličiny* X *je koeficient u* $\frac{t^r}{r!}$ *v rozvoji* M_X *do exponenciální mocninné řady (tedy např.* $EX = \mu'_1, DX = \mu'_2 - (\mu'_1)^2$*).*
- *Je-li* $Y = a + bX$*, pak* $M_Y(t) = e^{at} M_X(bt)$.

Momenty normálního rozdělení

Přímý výpočet střední hodnoty a rozptylu normovaného normálního rozdělení není triviální. S využitím momentové vytvořující funkce je ale poměrně jednoduchý.

Nechť $Z \sim N(0,1)$. Pak

$$M_Z(t) = \int_{-\infty}^{\infty} e^{tz} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \mathrm{d}z =$$
$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2 - 2tz + t^2 - t^2}{2}\right) \mathrm{d}z =$$
$$= \exp\left(\frac{t^2}{2}\right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z-t)^2}{2}\right) \mathrm{d}z = \exp\left(\frac{t^2}{2}\right).$$

Poslední integrál je roven 1 díky tomu, že na místě integrované funkce je funkce s vlastnostmi hustoty.

Střední hodnota a rozptyl normálního rozdělení

S využitím předchozího výpočtu $M_Z(t) = \exp\left(\frac{t^2}{2}\right)$ snadno spočítáme, že

$$M'_Z(t) = t\exp\left(\frac{t^2}{2}\right),$$

$$M''_Z(t) = t^2\exp\left(\frac{t^2}{2}\right) + \exp\left(\frac{t^2}{2}\right).$$

Dosazením $t = 0$ pak dostaneme

$$E(Z) = 0, D(Z) = 1.$$

Střední hodnota a rozptyl normálního rozdělení

S využitím předchozího výpočtu $M_Z(t) = \exp\left(\frac{t^2}{2}\right)$ snadno spočítáme, že

$$M'_Z(t) = t\exp\left(\frac{t^2}{2}\right),$$
$$M''_Z(t) = t^2\exp\left(\frac{t^2}{2}\right) + \exp\left(\frac{t^2}{2}\right).$$

Dosazením $t = 0$ pak dostaneme

$$E(Z) = 0, D(Z) = 1.$$

Pro transformovanou náhodnou veličinu $Y = \mu + \sigma Z \sim N(\mu, \sigma^2)$ pak snadno odvodíme z vlastností střední hodnoty, resp. rozptylu, že $E(Y) = \mu, D(Y) = \sigma^2$ (což zpětně zdůvodňuje zápis $N(\mu, \sigma^2)$). Momentová vytvořující funkce pro Y má tvar

$$M_Y(t) = \exp\left(\mu t + \sigma^2\frac{t^2}{2}\right).$$

Příklad

Určete rozdělení součtu nezávislých náhodných veličin
$X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$.

Příklad

Určete rozdělení součtu nezávislých náhodných veličin $X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$.

Řešení

Z vlastností momentové vytvořující funkce dostáváme

$$M_{X+Y}(t) = \exp\big(\mu_X t + \sigma_X^2 \frac{t^2}{2}\big) \exp\big(\mu_Y t + \sigma_Y^2 \frac{t^2}{2}\big) = \\ = \exp\big((\mu_X + \mu_Y)t + (\sigma_X^2 + \sigma_Y^2)\frac{t^2}{2}\big).$$

Proto $X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$.

Γ (gamma) rozdělení

Γ rozdělení se často používá u modelů čekání (např. v pojistné matematice je čas dožití často modelován pomocí gamma rozdělení).

Příklad

Určete konstantu c tak, aby funkce $cx^{a-1}e^{-bx}$ pro $x > 0$ a nulová jinde ($a, b > 0$ jsou parametry) byla hustotou náhodné veličiny.

Γ (gamma) rozdělení

Γ rozdělení se často používá u modelů čekání (např. v pojistné matematice je čas dožití často modelován pomocí gamma rozdělení).

Příklad

Určete konstantu c tak, aby funkce $cx^{a-1}e^{-bx}$ pro $x > 0$ a nulová jinde ($a, b > 0$ jsou parametry) byla hustotou náhodné veličiny.

Řešení

Hustota musí splňovat

$$1 = \int_0^\infty cx^{a-1}e^{-bx}\,\mathrm{d}x = \int_0^\infty c\left(\frac{t}{b}\right)^{a-1} e^{-t}\frac{1}{b}\,\mathrm{d}t =$$
$$= \frac{c}{b^a}\int_0^\infty t^{a-1}e^{-t}\,\mathrm{d}t = \frac{c}{b^a}\Gamma(a),$$

proto $c = \frac{b^a}{\Gamma(a)}$.

Poznámka

Funkce Γ je zobecnění faktoriálu ($\Gamma(n) = (n-1)!$ pro $n \in \mathbb{N}$), definované předpisem $\Gamma(a) = \int_0^\infty x^{a-1} e^{-x}\,\mathrm{d}x$. Často počítáme hodnoty této funkce s využitím vlastností $\Gamma(1/2) = \sqrt{\pi}$, $\Gamma(a+1) = a \cdot \Gamma(a)$.

Poznámka

Funkce Γ je zobecnění faktoriálu ($\Gamma(n) = (n-1)!$ pro $n \in \mathbb{N}$), definované předpisem $\Gamma(a) = \int_0^\infty x^{a-1}e^{-x}\,\mathrm{d}x$. Často počítáme hodnoty této funkce s využitím vlastností
$\Gamma(1/2) = \sqrt{\pi}$, $\Gamma(a+1) = a \cdot \Gamma(a)$.

Definice

Rozdělení náhodné veličiny s hustotou

$$f(x) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}$$

spočítanou v předchozím příkladu nazýváme **gamma rozdělení** s parametry a, b a značíme $\Gamma(a, b)$.

Poznámka

Funkce Γ je zobecnění faktoriálu ($\Gamma(n) = (n-1)!$ pro $n \in \mathbb{N}$), definované předpisem $\Gamma(a) = \int_0^\infty x^{a-1} e^{-x}\, dx$. Často počítáme hodnoty této funkce s využitím vlastností $\Gamma(1/2) = \sqrt{\pi}$, $\Gamma(a+1) = a \cdot \Gamma(a)$.

Definice

Rozdělení náhodné veličiny s hustotou

$$f(x) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}$$

spočítanou v předchozím příkladu nazýváme **gamma rozdělení** s parametry a, b a značíme $\Gamma(a, b)$. Momentová vytvořující funkce je pak $M(t) = (b/b - t)^a$, střední hodnota $E(X) = a/b$ a rozptyl $D(X) = a/b^2$.

Příklad (rozdělení χ^2 podruhé)

Nechť Z má normované normální rozdělení. Určete hustotu transformované náhodné veličiny $X = Z^2$.

Příklad (rozdělení χ^2 podruhé)

Nechť Z má normované normální rozdělení. Určete hustotu transformované náhodné veličiny $X = Z^2$.

Řešení

Již dříve jsme vypočetli přímým výpočtem přes distribuční funkci, že hustota

$$f_X(x) = \frac{1}{\sqrt{2\pi}} x^{-\frac{1}{2}} \mathrm{e}^{-\frac{x}{2}}$$

a řekli jsme si, že jde o (Pearsonovo) χ^2 rozdělení s jedním stupněm volnosti, které značíme $X \sim \chi^2(1)$.

Příklad (rozdělení χ^2 podruhé)

Nechť Z má normované normální rozdělení. Určete hustotu transformované náhodné veličiny $X = Z^2$.

Řešení

Již dříve jsme vypočetli přímým výpočtem přes distribuční funkci, že hustota

$$f_X(x) = \frac{1}{\sqrt{2\pi}} x^{-\frac{1}{2}} e^{-\frac{x}{2}}$$

a řekli jsme si, že jde o (Pearsonovo) χ^2 rozdělení s jedním stupněm volnosti, které značíme $X \sim \chi^2(1)$. Nyní vidíme, že jde o speciální případ Γ-rozdělení, totiž $\Gamma(1/2, 1/2)$.

Obecně pro součet Y čtverců n nezávislých náhodných veličin s rozdělením $N(0,1)$ obdobně odvodíme, že má rozdělení $\Gamma(n/2, 1/2)$ a říkáme, že Y má rozdělení $\chi^2(n)$ (*chí kvadrát s n stupni volnosti*). Toto rozdělení se ve statistice používá velmi často.

Další důležitá rozdělení

F-rozdělení

Jsou-li X, Y nezávislé náhodné veličiny s rozděleními $X \sim \chi^2(k)$, $Y \sim \chi^2(m)$, pak má transformovaná náhodná veličina

$$U = \frac{X/k}{Y/m}$$

takzvané Fisher-Snedecorovo F-rozdělení $F(k, m)$ s k a m stupni volnosti.

Další důležitá rozdělení

F-rozdělení

Jsou-li X, Y nezávislé náhodné veličiny s rozděleními $X \sim \chi^2(k)$, $Y \sim \chi^2(m)$, pak má transformovaná náhodná veličina

$$U = \frac{X/k}{Y/m}$$

takzvané Fisher-Snedecorovo F-rozdělení $F(k, m)$ s k a m stupni volnosti.

Studentovo **t-rozdělení**

Jsou-li $Z \sim N(0, 1)$ a $X \sim \chi^2(n)$ nezávislé náhodné veličiny, pak má veličina

$$T = \frac{Z}{\sqrt{X/n}}$$

tzv. Studentovo t-rozdělení $t(n)$ s n stupni volnosti.

Přehled rozdělení odvozených od normálního

$Z_1, \ldots, Z_k \sim N(0,1)$ . . . . . **nezávislá** normovaná normální

$X_k^2 = \sum_{i=1}^k Z_i^2 \sim \chi^2(k)$ . . . . chí-kvadrát o k stupních volnosti

$F_{k,m} = \frac{X_k^2/k}{X_m^2/m} \sim F(k,m)$ . . F-rozdělení s k a m stupni volnosti

$T_k = \frac{Z}{\sqrt{X_k^2/k}} \sim t(k)$ . . . . . . t-rozdělení s k stupni volnosti

Přehled rozdělení odvozených od normálního

$Z_1, \ldots, Z_k \sim N(0,1)$ . . . . . **nezávislá** normovaná normální

$X_k^2 = \sum_{i=1}^{k} Z_i^2 \sim \chi^2(k)$ . . . . chí-kvadrát o k stupních volnosti

$F_{k,m} = \frac{X_k^2/k}{X_m^2/m} \sim F(k,m)$ . . F-rozdělení s k a m stupni volnosti

$T_k = \frac{Z}{\sqrt{X_k^2/k}} \sim t(k)$ . . . . . . t-rozdělení s k stupni volnosti

Odtud zejména $Z^2 \sim \chi^2(1)$ a $T_k^2 \sim F(1,k)$.

Přehled rozdělení odvozených od normálního

$Z_1, \dots, Z_k \sim N(0,1)$ . . . . . **nezávislá** normovaná normální

$X_k^2 = \sum_{i=1}^k Z_i^2 \sim \chi^2(k)$ . . . . chí-kvadrát o k stupních volnosti

$F_{k,m} = \frac{X_k^2/k}{X_m^2/m} \sim F(k,m)$ . . F-rozdělení s k a m stupni volnosti

$T_k = \frac{Z}{\sqrt{X_k^2/k}} \sim t(k)$ . . . . . . t-rozdělení s k stupni volnosti

Odtud zejména $Z^2 \sim \chi^2(1)$ a $T_k^2 \sim F(1,k)$.

rozdělení	střední hodnota	rozptyl
$N(\mu,\sigma^2)$	μ	σ^2
$\chi^2(k)$	k	$2k$
$t(k)$	0	$k/(k-2)$
$F(k,m)$	$m/(m-2)$	$2m^2(k+m-2)/k(m-2)^2(m-4)$

Plán přednášky

Motivace

S jedním případem limitní věty jsme se již setkali –
de Moivre-Laplaceova věta říká, že binomické rozdělení $\mathrm{Bi}(n, p)$ lze za určitých podmínek aproximovat normovaným normálním rozdělením. Obvykle se k aproximaci přistupuje při splnění podmínek $np(1-p) > 9$ a $\frac{1}{n+1} < p < \frac{n}{n+1}$.

Motivace

S jedním případem limitní věty jsme se již setkali – de Moivre-Laplaceova věta říká, že binomické rozdělení $\mathrm{Bi}(n, p)$ lze za určitých podmínek aproximovat normovaným normálním rozdělením. Obvykle se k aproximaci přistupuje při splnění podmínek $np(1-p) > 9$ a $\frac{1}{n+1} < p < \frac{n}{n+1}$.
V této kapitole zformulujeme zobecnění této věty a rovněž další tvrzení umožňující odhadovat chování náhodných veličin při velkém počtu nezávislých opakování náhodného pokusu.

Čebyševova nerovnost

Věta

Pro libovolné $\epsilon > 0$ *platí*

$$P(|X - EX| \geq \epsilon) \leq \frac{DX}{\epsilon^2}.$$

Čebyševova nerovnost

Věta

Pro libovolné $\epsilon > 0$ *platí*

$$P(|X - EX| \geq \epsilon) \leq \frac{DX}{\epsilon^2}.$$

Důkaz.

Budeme odhadovat rozptyl DX ve spojitém případě (diskrétní analogicky):

$$DX = \int_{-\infty}^{\infty} (X - EX)^2 f(x)\,\mathrm{d}x \geq \int_{|x-EX|\geq\epsilon} (X - EX)^2 f(x)\,\mathrm{d}x \geq$$

$$\geq \int_{|x-EX|\geq\epsilon} \epsilon^2 f(x)\,\mathrm{d}x = \epsilon^2 P(|X - EX| \geq \epsilon).$$

$\square$

Pomocí Čebyševovy nerovnosti můžeme odhadovat pravděpodobnost, s jakou se náhodná veličina s neznámým rozdělením odchýlí od své střední hodnoty o více než k-násobek směrodatné odchylky (zřejmě je totiž $P(|X - E(X)| \geq k\sigma) \leq \frac{1}{k^2}$).

Pomocí Čebyševovy nerovnosti můžeme odhadovat pravděpodobnost, s jakou se náhodná veličina s neznámým rozdělením odchýlí od své střední hodnoty o více než k-násobek směrodatné odchylky (zřejmě je totiž $P(|X - E(X)| \geq k\sigma) \leq \frac{1}{k^2}$).

Příklad

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

1. Odhadněte $P(|X - \mu| \geq 3\sigma)$.

Pomocí Čebyševovy nerovnosti můžeme odhadovat pravděpodobnost, s jakou se náhodná veličina s neznámým rozdělením odchýlí od své střední hodnoty o více než k-násobek směrodatné odchylky (zřejmě je totiž $P(|X - E(X)| \geq k\sigma) \leq \frac{1}{k^2}$).

Příklad

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

1. Odhadněte $P(|X - \mu| \geq 3\sigma)$.
2. Vypočtěte $P(|X - \mu| \geq 3\sigma)$, jestliže navíc víte, že $X \sim N(0, 1)$.

Pomocí Čebyševovy nerovnosti můžeme odhadovat pravděpodobnost, s jakou se náhodná veličina s neznámým rozdělením odchýlí od své střední hodnoty o více než k-násobek směrodatné odchylky (zřejmě je totiž $P(|X - E(X)| \geq k\sigma) \leq \frac{1}{k^2}$).

Příklad

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

1. Odhadněte $P(|X - \mu| \geq 3\sigma)$.
2. Vypočtěte $P(|X - \mu| \geq 3\sigma)$, jestliže navíc víte, že $X \sim N(0, 1)$.

Pomocí Čebyševovy nerovnosti můžeme odhadovat pravděpodobnost, s jakou se náhodná veličina s neznámým rozdělením odchýlí od své střední hodnoty o více než k-násobek směrodatné odchylky (zřejmě je totiž $P(|X - E(X)| \geq k\sigma) \leq \frac{1}{k^2}$).

Příklad

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

1. Odhadněte $P(|X - \mu| \geq 3\sigma)$.
2. Vypočtěte $P(|X - \mu| \geq 3\sigma)$, jestliže navíc víte, že $X \sim N(0, 1)$.

Řešení

1. $1/9$,
2. 0,0027.

Zákon velkých čísel

Věta (Čebyševova – slabý zákon velkých čísel)

Nechť jsou $X_1, X_2, \ldots$ po dvou nezávislé náhodné veličiny, které mají všechny stejnou střední hodnotu μ a rozptyl shora ohraničený stejnou hodnotou σ^2. Pak pro libovolné $\epsilon > 0$ platí

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

Říkáme, že posloupnost aritmetických průměrů konverguje podle pravděpodobnosti ke střední hodnotě μ.

Zákon velkých čísel

Věta (Čebyševova – slabý zákon velkých čísel)

Nechť jsou $X_1, X_2, \dots$ po dvou nezávislé náhodné veličiny, které mají všechny stejnou střední hodnotu μ a rozptyl shora ohraničený stejnou hodnotou σ^2. Pak pro libovolné $\epsilon > 0$ platí

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

Říkáme, že posloupnost aritmetických průměrů konverguje podle pravděpodobnosti ke střední hodnotě μ.

Speciálním případem této věty je Bernoulliova věta, která říká, že je-li $Y_n \sim \mathrm{Bi}(n, p)$, pak posloupnost relativních četností Y_n/n konverguje podle pravděpodobnosti k p.

Věta (Bernoulliova)

Pro náhodnou veličinu s binomickým rozdělením $Y_n \sim \mathrm{Bi}(n, p)$ *a pro libovolné* $\epsilon > 0$ *platí*

$$P\left(\left|\frac{Y_n}{n} - p\right| > \epsilon\right) \leq \frac{p(1-p)}{n\epsilon^2}.$$

Věta (Bernoulliova)

Pro náhodnou veličinu s binomickým rozdělením $Y_n \sim \mathrm{Bi}(n, p)$ *a pro libovolné* $\epsilon > 0$ *platí*

$$P\left(\left|\frac{Y_n}{n} - p\right| > \epsilon\right) \leq \frac{p(1-p)}{n\epsilon^2}.$$

Důkaz.

Plyne snadno z Čebyševovy nerovnosti, neboť
$E(Y_n/n) = np/n = p$ a
$D(Y_n/n) = np(1-p)/n^2 = p(1-p)/n.$ □

Příklad

Při zkoušce bylo zjištěno, že mezi 600 kontrolovanými studenty je 5 studentů, kteří neumí ani malou násobilku. Odhadněte pravděpodobnost, že relativní četnost takových studentů se od jejich pravděpodobnosti výskytu liší o více než 0,01? (Můžete předpokládat, že pravděpodobnost výskytu studenta bez znalosti násobilky je menší než 0,02).

Centrální limitní věta

Centrální limitní věta dá odpověď na otázku, proč je normální rozdělení nejdůležitějším rozdělením. Ukazuje totiž, že rozdělení součtu dostatečně velkého počtu nezávislých a stejně rozdělených náhodných veličin lze aproximovat normálním rozdělením.

Věta

Nechť je $Y_1, Y_2, \ldots$ posloupnost **nezávislých stejně rozdělených** *náhodných veličin se střední hodnotou μ a rozptylem σ^2. Pak pro* **normované** *náhodné veličiny*

$$S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{Y_i - \mu}{\sigma}$$

platí

$$\lim_{n\to\infty} P(S_n < x) = \Phi(x),$$

kde Φ je distribuční funkce rozdělení $N(0,1)$.

Příklad

Mezi matematiky v ČR je jich 10% s příjmem přesahujícím celostátní průměr. Kolik matematiků je třeba pozvat na konferenci, aby s pravděpodobností aspoň 0,95 mezi nimi bylo 8 až 12 procent s nadprůměrným příjmem?

Příklad

Mezi matematiky v ČR je jich 10% s příjmem přesahujícím celostátní průměr. Kolik matematiků je třeba pozvat na konferenci, aby s pravděpodobností aspoň 0,95 mezi nimi bylo 8 až 12 procent s nadprůměrným příjmem?

Řešení

$Y_n \sim \mathrm{Bi}(n; 0{,}1)$, $E(Y_n) = 0{,}1 \cdot n$, $D(Y_n) = 0{,}1 \cdot 0{,}9 \cdot n$. Pak

$$0{,}95 \leq P(0{,}08n \leq Y_n \leq 0{,}12n) =$$

$$= P\left(\frac{0{,}08 - 0{,}1}{\sqrt{0{,}09n}} n \leq \frac{Y_n - 0{,}1n}{\sqrt{0{,}09n}} \leq \frac{0{,}12 - 0{,}1}{\sqrt{0{,}09n}} n\right) =$$

$$= P\left(\frac{-\sqrt{n}}{15} \leq \frac{Y_n - 0{,}1n}{\sqrt{0{,}09n}} \leq \frac{\sqrt{n}}{15}\right) \approx \Phi\left(\frac{\sqrt{n}}{15}\right) - \Phi\left(-\frac{\sqrt{n}}{15}\right).$$

Je tedy $\Phi\left(\frac{\sqrt{n}}{15}\right) \geq 0{,}975$, což je ekvivalentní $\sqrt{n}/15 \geq 1{,}96$, tj. $n \geq 865$.

Řešení (Pomocí Bernoulliovy nerovnosti)

Nyní využijme Bernoulliovu nerovnost – ta dává

$$P\left(\left|\frac{Y_n}{n} - 0{,}1\right| \leq 0{,}02\right) \geq 1 - \frac{0{,}1 \cdot 0{,}9}{n \cdot 0{,}02^2},$$

což má být alespoň 0,95. Odtud

$$n \geq \frac{0{,}09}{0{,}05 \cdot 0{,}02^2} = 4500.$$

Řešení (Pomocí Bernoulliovy nerovnosti)

Nyní využijme Bernoulliovu nerovnost – ta dává

$$P\left(\left|\frac{Y_n}{n} - 0{,}1\right| \leq 0{,}02\right) \geq 1 - \frac{0{,}1 \cdot 0{,}9}{n \cdot 0{,}02^2},$$

což má být alespoň 0,95. Odtud

$$n \geq \frac{0{,}09}{0{,}05 \cdot 0{,}02^2} = 4500.$$

Vidíme, že odhad prostřednictvím Bernoulliovy nerovnosti je podstatně slabší než odhad s využitím centrální limitní věty (resp. de Moivre-Laplaceovy věty).

Plán přednášky

Statistika zkoumá jevy na rozsáhlých **souborech** případů a zkoumá **statistické znaky** jednotlivých statistických **jednotek**. Obvykle nelze testovat všechny jednotky **základního souboru**, proto se omezujeme na prozkoumání některého **výběrového souboru** rozsahu n.

Statistika zkoumá jevy na rozsáhlých **souborech** případů a zkoumá **statistické znaky** jednotlivých statistických **jednotek**. Obvykle nelze testovat všechny jednotky **základního souboru**, proto se omezujeme na prozkoumání některého **výběrového souboru** rozsahu n.

Předpokládejme, že jsme na n statistických jednotkách naměřili **soubor hodnot**

$$x_1, \dots, x_n$$

daného znaku. Znaky obvykle dělíme na *kvalitativní* (nominální, ordinální) a *kvantitativní* (intervalové, poměrové).

Počtu prvků souboru říkáme **rozsah**.

Základní pojmy popisné statistiky

- absolutní (relativní) četnosti, četnostní tabulka

Základní pojmy popisné statistiky

- absolutní (relativní) četnosti, četnostní tabulka
- histogram

Základní pojmy popisné statistiky

- absolutní (relativní) četnosti, četnostní tabulka
- histogram
- (výběrový) průměr, geometrický, harmonický průměr

Základní pojmy popisné statistiky

- absolutní (relativní) četnosti, četnostní tabulka
- histogram
- (výběrový) průměr, geometrický, harmonický průměr
- medián, p-tý kvantil, percentil, kvartil

Základní pojmy popisné statistiky

- absolutní (relativní) četnosti, četnostní tabulka
- histogram
- (výběrový) průměr, geometrický, harmonický průměr
- medián, *p*-tý kvantil, percentil, kvartil
- modus

Základní pojmy popisné statistiky

- absolutní (relativní) četnosti, četnostní tabulka
- histogram
- (výběrový) průměr, geometrický, harmonický průměr
- medián, p-tý kvantil, percentil, kvartil
- modus
- rozptyl s_x^2, resp. $n/(n-1)s_x^2$

Základní pojmy popisné statistiky

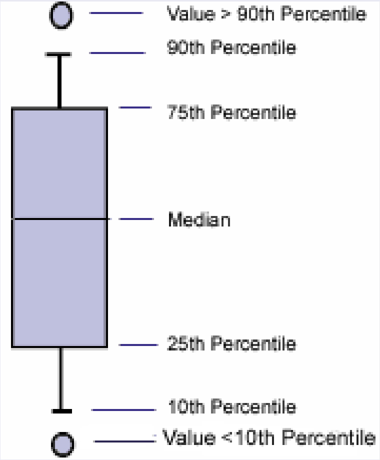
- absolutní (relativní) četnosti, četnostní tabulka
- histogram
- (výběrový) průměr, geometrický, harmonický průměr
- medián, p-tý kvantil, percentil, kvartil
- modus
- rozptyl s_x^2, resp. $n/(n-1)s_x^2$
- rozpětí, kvartilové rozpětí, průměrná odchylka (od mediánu)

Základní pojmy popisné statistiky

- absolutní (relativní) četnosti, četnostní tabulka
- histogram
- (výběrový) průměr, geometrický, harmonický průměr
- medián, p-tý kvantil, percentil, kvartil
- modus
- rozptyl s_x^2, resp. $n/(n-1)s_x^2$
- rozpětí, kvartilové rozpětí, průměrná odchylka (od mediánu)
- koeficient šikmosti, špičatosti

Diagramy

Krabicový diagram, box plot



Plán přednášky

Náhodný vektor – připomenutí

Je-li $(\Omega, \mathcal{A}, P)$ pravděpodobnostní prostor a $X_1, \dots, X_n$ na něm definované náhodné veličiny s distribučními funkcemi $F_1, \dots, F_n$, pak **náhodným vektorem** je n-tice $X = (X_1, \dots, X_n)$ s distribuční funkcí definovanou vztahem

$$F_X(x_1, \dots, x_n) = P(X_1 \leq x_1, \dots, X_n \leq x_n).$$

V tomto kontextu nazýváme F *simultánní distribuční funkcí* náhodného vektoru X a F_i marginální distribuční funkcí náhodné veličiny X_i.

Podobně jako v případě diskrétní náhodné veličiny označuje $p(x_1, \dots, x_n)$ pravděpodobnostní funkci **diskrétního náhodného vektoru** X, je-li

$$F_X(x_1, \dots, x_n) = \sum_{t_1 \le x_1} \cdots \sum_{t_n \le x_n} p(t_1, \dots, t_n).$$

Podobně jako v případě diskrétní náhodné veličiny označuje $p(x_1, \ldots, x_n)$ pravděpodobnostní funkci **diskrétního náhodného vektoru** X, je-li

$$F_X(x_1, \ldots, x_n) = \sum_{t_1 \leq x_1} \cdots \sum_{t_n \leq x_n} p(t_1, \ldots, t_n).$$

Funkci f_X nazveme **hustotou** normálního vektoru X, pokud pro libovolnou n-tici $(x_1, \ldots, x_n)$ platí

$$F_X(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_X(t_1, \ldots, t_n)\, \mathrm{d}t_1 \ldots \mathrm{d}t_n.$$

Uvážíme-li diskrétní náhodný vektor [1] (X, Y), pak je vztah mezi sdruženým rozdělením vektoru (X, Y) a marginálním rozdělením promenné X určen rovností $P(X = x_i) = \sum_{j=1}^{\infty} P(X = x_i, Y = y_j)$, kde $y_1, \ldots$ tvoří úplný systém jevů.

[1] Obvykle zapisujeme ve statistice vektory do sloupců, proto bychom spíše měli psát $(X, Y)^T$.

Uvážíme-li diskrétní náhodný vektor [1] (X, Y), pak je vztah mezi sdruženým rozdělením vektoru (X, Y) a marginálním rozdělením promenné X určen rovností $P(X = x_i) = \sum_{j=1}^{\infty} P(X = x_i, Y = y_j)$, kde $y_1, \ldots$ tvoří úplný systém jevů. Vztah pro spojitě rozdělený náhodný vektor je analogický.

[1] Obvykle zapisujeme ve statistice vektory do sloupců, proto bychom spíše měli psát $(X, Y)^T$.

(stochastická) Nezávislost náhodných veličin

Dříve uvedenou definici nezávislosti náhodných veličin $X_1, \dots, X_n$ pomocí vztahu

$$P(X_1 = x_1, \dots, X_n = x_n) = P(X_1 = x_1) \cdots P(X_n = x_n)$$

pro libovolné $x_1, \dots, x_n$, tak můžeme nyní přepsat pomocí vztahu mezi sdruženou distribuční funkcí náhodného vektoru $X = (X_1, \dots, X_n)$ a marginálních distribučních funkcí náhodných veličin $X_1, \dots, X_n$:

$$F_X(x_1, \dots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n).$$

(stochastická) Nezávislost náhodných veličin

Dříve uvedenou definici nezávislosti náhodných veličin $X_1, \dots, X_n$ pomocí vztahu

$$P(X_1 = x_1, \dots, X_n = x_n) = P(X_1 = x_1) \cdots P(X_n = x_n)$$

pro libovolné $x_1, \dots, x_n$, tak můžeme nyní přepsat pomocí vztahu mezi sdruženou distribuční funkcí náhodného vektoru $X = (X_1, \dots, X_n)$ a marginálních distribučních funkcí náhodných veličin $X_1, \dots, X_n$:

$$F_X(x_1, \dots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n).$$

Příklad

Házíme dvěma běžnými kostkami, jako náhodnou veličinu X označme součet bodů na obou kostkách, jako náhodnou veličinu Y absolutní hodnotu rozdílu. Určete sdružené rozdělení náhodného vektoru (X, Y), obě marginální rozdělení a odvoďte, jsou-li X a Y nezávislé.

Číselné charakteristiky náhodných vektorů

$E(X) = (E(X_1), \dots, E(X_n))$ se nazývá vektor středních hodnot,

Číselné charakteristiky náhodných vektorů

$E(X) = (E(X_1), \ldots, E(X_n))$ se nazývá vektor středních hodnot,

$$\mathrm{var}(X) = \begin{pmatrix} D(X_1) & C(X_1, X_2) & \cdots & C(X_1, X_n) \\ \multicolumn{4}{c}{\dotfill} \\ C(X_n, X_1) & C(X_n, X_2) & \cdots & D(X_n) \end{pmatrix}$$

varianční (rozptylová) matice a

Číselné charakteristiky náhodných vektorů

$E(X) = (E(X_1), \dots, E(X_n))$ se nazývá vektor středních hodnot,

$$\operatorname{var}(X) = \begin{pmatrix} D(X_1) & C(X_1, X_2) & \cdots & C(X_1, X_n) \\ \dots & \dots & \dots & \dots \\ C(X_n, X_1) & C(X_n, X_2) & \cdots & D(X_n) \end{pmatrix}$$

varianční (rozptylová) matice a

$$\operatorname{cor} X = \begin{pmatrix} 1 & R(X_1, X_2) & \cdots & R(X_1, X_n) \\ \dots & \dots & \dots & \dots \\ R(X_n, X_1) & R(X_n, X_2) & \cdots & 1 \end{pmatrix}$$

je korelační matice.

Číselné charakteristiky náhodných vektorů

$E(X) = (E(X_1), \dots, E(X_n))$ se nazývá vektor středních hodnot,

$$\mathrm{var}(X) = \begin{pmatrix} D(X_1) & C(X_1, X_2) & \cdots & C(X_1, X_n) \\ \multicolumn{4}{c}{\dotfill} \\ C(X_n, X_1) & C(X_n, X_2) & \cdots & D(X_n) \end{pmatrix}$$

varianční (rozptylová) matice a

$$\mathrm{cor}\, X = \begin{pmatrix} 1 & R(X_1, X_2) & \cdots & R(X_1, X_n) \\ \multicolumn{4}{c}{\dotfill} \\ R(X_n, X_1) & R(X_n, X_2) & \cdots & 1 \end{pmatrix}$$

je korelační matice.
Snadno je po rozepsání po jednotlivých složkách vidět, že $\mathrm{var}(X) = E((X - E(X)) \cdot (X - E(X))^T)$

Ukážeme na příkladech, že pravděpodobnostní struktura náhodného vektoru (X, Y) není určena pouze marginálními rozděleními veličin X a Y. Podstatný je rovněž pravděpodobnostní vztah mezi X a Y, který je částečně popsán např. prostřednictvím korelačního koeficientu.

Ukážeme na příkladech, že pravděpodobnostní struktura náhodného vektoru (X, Y) není určena pouze marginálními rozděleními veličin X a Y. Podstatný je rovněž pravděpodobnostní vztah mezi X a Y, který je částečně popsán např. prostřednictvím korelačního koeficientu.

Příklad

Jsou-li X a Y náhodné veličiny, nabývající hodnot 0 a 1, pak

$$P(X = 1, Y = 1) - P(X = 1)P(Y = 1) = E(XY) - E(X)E(Y) =$$
$$= \operatorname{cov}(X, Y).$$

Ukážeme na příkladech, že pravděpodobnostní struktura náhodného vektoru (X, Y) není určena pouze marginálními rozděleními veličin X a Y. Podstatný je rovněž pravděpodobnostní vztah mezi X a Y, který je částečně popsán např. prostřednictvím korelačního koeficientu.

Příklad

Jsou-li X a Y náhodné veličiny, nabývající hodnot 0 a 1, pak

$$P(X = 1, Y = 1) - P(X = 1)P(Y = 1) = E(XY) - E(X)E(Y) =$$
$$= \operatorname{cov}(X, Y).$$

Odtud je snadno vidět, že pokud jsou X a Y nekorelované, jsou i nezávislé (což obecně neplatí).

Uveďme ještě příklad, ilustrující, že nekorelovanost nemusí implikovat nezávislost:

Uveďme ještě příklad, ilustrující, že nekorelovanost nemusí implikovat nezávislost:

Příklad

Buďte A a X nezávislé náhodné veličiny, splňující $X \sim N(0,1)$ a $P(A=1) = P(A=-1) = 1/2$. Položíme-li $Y = AX$, pak

$$P(Y < y) = \frac{1}{2}P(X < y) + \frac{1}{2}P(-X < y) = \Phi(y),$$

proto má rovněž Y rozdělení $N(0,1)$.

Uveďme ještě příklad, ilustrující, že nekorelovanost nemusí implikovat nezávislost:

Příklad

Buďte A a X nezávislé náhodné veličiny, splňující $X \sim N(0,1)$ a $P(A=1) = P(A=-1) = 1/2$. Položíme-li $Y = AX$, pak

$$P(Y < y) = \frac{1}{2}P(X < y) + \frac{1}{2}P(-X < y) = \Phi(y),$$

proto má rovněž Y rozdělení $N(0,1)$.

Dále $\operatorname{cov}(X,Y) = E(XY) - E(X)E(Y) = E(AX^2) = E(A)E(X^2) = 0 \cdot 1 = 0$, přitom $P(X = Y) = P(X = -Y) = 1/2$ a X, Y zřejmě nejsou nezávislé.

Příklad

Nechť (X, Y) je náhodný vektor, který má rovnoměrné rozdělení na jednotkovém kruhu $K = \{(x, y) : x^2 + y^2 \leq 1\}$. Zřejmě je hustota tohoto rozdělení rovna $1/\pi$ pro $(x, y) \in K$ a 0 jinde a je rovněž vidět, že X, Y **nejsou nezávislé.**

Příklad

Nechť (X, Y) je náhodný vektor, který má rovnoměrné rozdělení na jednotkovém kruhu $K = \{(x, y) : x^2 + y^2 \leq 1\}$. Zřejmě je hustota tohoto rozdělení rovna $1/\pi$ pro $(x, y) \in K$ a 0 jinde a je rovněž vidět, že X, Y **nejsou nezávislé.** Označme $R = R(X, Y)$ a $\Phi = \Phi(X, Y)$ polární souřadnice náhodného vektoru (X, Y) a určíme rozdělení vektoru (R, Φ).

Příklad

Nechť (X, Y) je náhodný vektor, který má rovnoměrné rozdělení na jednotkovém kruhu $K = \{(x, y) : x^2 + y^2 \leq 1\}$. Zřejmě je hustota tohoto rozdělení rovna $1/\pi$ pro $(x, y) \in K$ a 0 jinde a je rovněž vidět, že X, Y **nejsou nezávislé.** Označme $R = R(X, Y)$ a $\Phi = \Phi(X, Y)$ polární souřadnice náhodného vektoru (X, Y) a určíme rozdělení vektoru (R, Φ).
Pro $0 < r_1 \leq 1$ a $0 < \varphi_1 \leq 2\pi$ je

$$\begin{aligned} P(R < r_1, \Phi \leq \varphi_1) &= \frac{1}{\pi}\pi r_1^2 \frac{\varphi_1}{2\pi} = \\ &= \int_0^{r_1} \int_0^{\varphi_1} \frac{1}{2\pi} 2r \, \mathrm{d}\varphi \, \mathrm{d}r. \end{aligned}$$

Hustota je tedy rovna $f(r, \varphi) = \frac{r}{\pi}$ pro $0 < r \leq 1$, $0 < \varphi \leq 2\pi$ a rovna 0 všude jinde.

Příklad (pokr.)

Marginální hustoty $g(r)$ a $h(\varphi)$ veličin R a Φ se nyní snadno dopočtou:

$$g(r) = \int_{-\infty}^{\infty} f(r,\varphi)\,\mathrm{d}\varphi = \int_0^{2\pi} \frac{r}{\pi}\,\mathrm{d}\varphi = 2r$$

$$h(\varphi) = \int_{-\infty}^{\infty} f(r,\varphi)\,\mathrm{d}r = \int_0^{1} \frac{r}{\pi}\,\mathrm{d}r = \frac{1}{2\pi}.$$

Příklad (pokr.)

Marginální hustoty $g(r)$ a $h(\varphi)$ veličin R a Φ se nyní snadno dopočtou:

$$g(r) = \int_{-\infty}^{\infty} f(r, \varphi)\, \mathrm{d}\varphi = \int_0^{2\pi} \frac{r}{\pi}\, \mathrm{d}\varphi = 2r$$

$$h(\varphi) = \int_{-\infty}^{\infty} f(r, \varphi)\, \mathrm{d}r = \int_0^{1} \frac{r}{\pi}\, \mathrm{d}r = \frac{1}{2\pi}.$$

Veličina Φ má rovnoměrné rozdělení na $(0, 2\pi\rangle$, odkud $E(\Phi) = \pi$ a $D(\Phi) = \pi^2/3$, snadno rovněž odvodíme $E(R) = 2/3$, $D(R) = 1/18$.

Příklad (pokr.)

Marginální hustoty $g(r)$ a $h(\varphi)$ veličin R a Φ se nyní snadno dopočtou:

$$g(r) = \int_{-\infty}^{\infty} f(r,\varphi)\,\mathrm{d}\varphi = \int_0^{2\pi} \frac{r}{\pi}\,\mathrm{d}\varphi = 2r$$

$$h(\varphi) = \int_{-\infty}^{\infty} f(r,\varphi)\,\mathrm{d}r = \int_0^{1} \frac{r}{\pi}\,\mathrm{d}r = \frac{1}{2\pi}.$$

Veličina Φ má rovnoměrné rozdělení na $(0, 2\pi\rangle$, odkud $E(\Phi) = \pi$ a $D(\Phi) = \pi^2/3$, snadno rovněž odvodíme $E(R) = 2/3$, $D(R) = 1/18$.
Všimněme si ale zejména, že $f(r,\varphi) = g(r)h(\varphi)$, což znamená **nezávislost veličin R a Φ.**

Vlastnosti charakteristik náhodného vektoru

Věta

Pro náhodné vektory X, Y stejné dimenze, konstantní matici B a konstantní vektor a (odpovídajících dimenzí) platí

- $E(X+Y) = E(X) + E(Y)$,

Vlastnosti charakteristik náhodného vektoru

Věta

Pro náhodné vektory X, Y stejné dimenze, konstantní matici B a konstantní vektor a (odpovídajících dimenzí) platí

- $E(X+Y) = E(X) + E(Y)$,
- $E(a+BX) = a + B \cdot E(X)$,

Vlastnosti charakteristik náhodného vektoru

Věta

Pro náhodné vektory X, Y stejné dimenze, konstantní matici B a konstantní vektor a (odpovídajících dimenzí) platí

- $E(X+Y) = E(X) + E(Y)$,
- $E(a + BX) = a + B \cdot E(X)$,
- $\operatorname{var}(a + B \cdot X) = B \operatorname{var}(X) B^T$.

Vlastnosti charakteristik náhodného vektoru

Věta

Pro náhodné vektory X, Y stejné dimenze, konstantní matici B a konstantní vektor a (odpovídajících dimenzí) platí

- $E(X+Y) = E(X) + E(Y)$,
- $E(a+BX) = a + B \cdot E(X)$,
- $\operatorname{var}(a + B \cdot X) = B \operatorname{var}(X) B^T$.

Vlastnosti charakteristik náhodného vektoru

Věta

Pro náhodné vektory X, Y stejné dimenze, konstantní matici B a konstantní vektor a (odpovídajících dimenzí) platí

- $E(X+Y) = E(X) + E(Y)$,
- $E(a+BX) = a + B \cdot E(X)$,
- $\operatorname{var}(a + B \cdot X) = B \operatorname{var}(X) B^T$.

Důkaz.

Důkaz vyplývá z vlastností náhodných veličin a ze vztahu $\operatorname{var}(X) = E((X - E(X))(X - E(X))^T)$. $\square$

Plán přednášky

Definice

Náhodným výběrem rozsahu n rozumíme n-tici **nezávislých** a **stejně rozdělených** náhodných veličin $X_1, \ldots, X_n \sim F_X(x)$ (někdy také hovoříme o n nezávislých kopiích náhodné veličiny X).

Definice

Náhodným výběrem rozsahu n rozumíme n-tici **nezávislých** a **stejně rozdělených** náhodných veličin $X_1, \dots, X_n \sim F_X(x)$ (někdy také hovoříme o n nezávislých kopiích náhodné veličiny X).

Náhodným výběrem rozsahu n **z** p-**rozměrného rozdělení** rozumíme n-tici **nezávislých** a **stejně rozdělených** p-rozměrných náhodných vektorů.

Definice

Náhodným výběrem rozsahu n rozumíme n-tici **nezávislých** a **stejně rozdělených** náhodných veličin $X_1, \ldots, X_n \sim F_X(x)$ (někdy také hovoříme o n nezávislých kopiích náhodné veličiny X).
Náhodným výběrem rozsahu n **z** p**-rozměrného rozdělení** rozumíme n-tici **nezávislých** a **stejně rozdělených** p-rozměrných náhodných vektorů.

V matematické statistice často pracujeme s transformacemi náhodného výběru, takovým náhodným veličinám (příp. vektorům) říkáme **statistiky**. V následujícím zavedeme několik důležitých statistik a ukážeme jejich souvislost s číselnými charakteristikami náhodných veličin.

Základní statistiky

Definice

Nechť $X_1, \ldots, X_n$ je náhodný výběr. Statistiku

$$M = \frac{1}{n}\sum_{i=1}^{n} X_i$$

nazýváme **výběrový průměr**, statistiku

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M)^2$$

výběrový rozptyl a statistiku $S = \sqrt{S^2}$ **výběrová směrodatná odchylka**. Analogicky se definují i výběrová kovariance, příp. výběrový korelační koeficient pro dvourozměrný náhodný výběr.

Vlastnosti statistik

Protože jsou uvedené statistiky náhodnými veličinami, lze se přirozeně ptát po jejich číselných charakteristikách.

Věta

Nechť $X_1, \ldots, X_n$ je náhodný výběr rozsahu n z rozdělení se střední hodnotou μ a rozptylem σ^2 . Pak platí:

- $E(M) = \mu$,

Vlastnosti statistik

Protože jsou uvedené statistiky náhodnými veličinami, lze se přirozeně ptát po jejich číselných characteristikách.

Věta

Nechť $X_1, \ldots, X_n$ je náhodný výběr rozsahu n z rozdělení se střední hodnotou μ a rozptylem σ^2 . Pak platí:

- $E(M) = \mu$,
- $D(M) = \mathrm{var}(M) = \sigma^2/n$,

Vlastnosti statistik

Protože jsou uvedené statistiky náhodnými veličinami, lze se přirozeně ptát po jejich číselných characteristikách.

Věta

Nechť $X_1, \ldots, X_n$ je náhodný výběr rozsahu n z rozdělení se střední hodnotou μ a rozptylem σ^2 . Pak platí:

- $E(M) = \mu$,
- $D(M) = \mathrm{var}(M) = \sigma^2/n$,
- $E(S^2) = \sigma^2$.

Důkaz.

Ukážeme jen (nejsložitější) 3. tvrzení.
Snadno se odvodí, že platí

$$\sum (X_i - \mu)^2 = \sum (X_i - M)^2 + n(M - \mu)^2.$$

Důkaz.

Ukážeme jen (nejsložitější) 3. tvrzení.
Snadno se odvodí, že platí

$$\sum (X_i - \mu)^2 = \sum (X_i - M)^2 + n(M - \mu)^2.$$

Proto je

$$E(S^2) = \frac{1}{n-1} E\left(\sum (X_i - \mu)^2\right) - \frac{n}{n-1} E(M - \mu)^2 =$$
$$= \frac{1}{n-1} \sum D(X_i) - \frac{n}{n-1} D(M) =$$
$$= \frac{n}{n-1}\sigma^2 - \frac{1}{n-1}\sigma^2 = \sigma^2.$$

□

V předchozí větě jsme ukázali, že výběrový průměr M splňuje $E(M) = \mu$, jeho střední hodnota je tedy rovna odhadovanému parametru μ. V takovém případě říkáme, že statistika M je **nestranným odhadem** parametru μ.

V předchozí větě jsme ukázali, že výběrový průměr M splňuje $E(M) = \mu$, jeho střední hodnota je tedy rovna odhadovanému parametru μ. V takovém případě říkáme, že statistika M je **nestranným odhadem** parametru μ.
Podobně jsme viděli, že S^2 je nestranným odhadem parametru σ^2.

V předchozí větě jsme ukázali, že výběrový průměr M splňuje $E(M) = \mu$, jeho střední hodnota je tedy rovna odhadovanému parametru μ. V takovém případě říkáme, že statistika M je **nestranným odhadem** parametru μ.
Podobně jsme viděli, že S^2 je nestranným odhadem parametru σ^2.
Všimněme si rovněž, že „přirozeněji" definovaná statistika $\frac{1}{n}\sum(X_i - M)^2$ není nestranným odhadem σ^2, její střední hodnota je totiž $\frac{n-1}{n}\sigma^2$. Rozmyslete si, je-li S nestranným odhadem směrodatné odchylky σ.

Náhodný výběr z normálního rozdělení

Uvažme nyní speciální případ, kdy je $X_1, \dots, X_n$ náhodný výběr z normálního rozdělení $N(\mu, \sigma^2)$.

Věta

- *M a S^2 jsou nezávislé náhodné veličiny.*

Náhodný výběr z normálního rozdělení

Uvažme nyní speciální případ, kdy je $X_1, \dots, X_n$ náhodný výběr z normálního rozdělení $N(\mu, \sigma^2)$.

Věta

- *M a S^2 jsou nezávislé náhodné veličiny.*
- *$M \sim N(\mu, \sigma^2/n)$, a tedy $U = (M - \mu)/(\sigma/\sqrt{n}) \sim N(0, 1)$.*

Náhodný výběr z normálního rozdělení

Uvažme nyní speciální případ, kdy je $X_1, \dots, X_n$ náhodný výběr z normálního rozdělení $N(\mu, \sigma^2)$.

Věta

- *M a S^2 jsou nezávislé náhodné veličiny.*
- *$M \sim N(\mu, \sigma^2/n)$, a tedy $U = (M - \mu)/(\sigma/\sqrt{n}) \sim N(0, 1)$.*
- *$K = (n-1)S^2/\sigma^2 \sim \chi^2(n-1)$.*

Náhodný výběr z normálního rozdělení

Uvažme nyní speciální případ, kdy je $X_1, \dots, X_n$ náhodný výběr z normálního rozdělení $N(\mu, \sigma^2)$.

Věta

- *M a S^2 jsou nezávislé náhodné veličiny.*
- *$M \sim N(\mu, \sigma^2/n)$, a tedy $U = (M - \mu)/(\sigma/\sqrt{n}) \sim N(0, 1)$.*
- *$K = (n-1)S^2/\sigma^2 \sim \chi^2(n-1)$.*
- *$\sum(X_i - \mu)^2/\sigma^2 \sim \chi^2(n)$.*

Náhodný výběr z normálního rozdělení

Uvažme nyní speciální případ, kdy je $X_1, \dots, X_n$ náhodný výběr z normálního rozdělení $N(\mu, \sigma^2)$.

Věta

- *M a S^2 jsou nezávislé náhodné veličiny.*
- $M \sim N(\mu, \sigma^2/n)$, *a tedy* $U = (M - \mu)/(\sigma/\sqrt{n}) \sim N(0, 1)$.
- $K = (n-1)S^2/\sigma^2 \sim \chi^2(n-1)$.
- $\sum (X_i - \mu)^2/\sigma^2 \sim \chi^2(n)$.
- $T = (M - \mu)/(S/\sqrt{n}) \sim t(n-1)$.

Náhodný výběr z normálního rozdělení

Uvažme nyní speciální případ, kdy je $X_1, \ldots, X_n$ náhodný výběr z normálního rozdělení $N(\mu, \sigma^2)$.

Věta

- *M a S^2 jsou nezávislé náhodné veličiny.*
- $M \sim N(\mu, \sigma^2/n)$, *a tedy* $U = (M - \mu)/(\sigma/\sqrt{n}) \sim N(0, 1)$.
- $K = (n - 1)S^2/\sigma^2 \sim \chi^2(n - 1)$.
- $\sum (X_i - \mu)^2/\sigma^2 \sim \chi^2(n)$.
- $T = (M - \mu)/(S/\sqrt{n}) \sim t(n - 1)$.

Náhodný výběr z normálního rozdělení

Uvažme nyní speciální případ, kdy je $X_1, \dots, X_n$ náhodný výběr z normálního rozdělení $N(\mu, \sigma^2)$.

Věta

- *M a S^2 jsou nezávislé náhodné veličiny.*
- *$M \sim N(\mu, \sigma^2/n)$, a tedy $U = (M - \mu)/(\sigma/\sqrt{n}) \sim N(0, 1)$.*
- *$K = (n-1)S^2/\sigma^2 \sim \chi^2(n-1)$.*
- *$\sum (X_i - \mu)^2/\sigma^2 \sim \chi^2(n)$.*
- *$T = (M - \mu)/(S/\sqrt{n}) \sim t(n-1)$.*

Poznámka

K odhadu μ, známe-li σ^2, slouží U, v opačném případě T.

Náhodný výběr z normálního rozdělení

Uvažme nyní speciální případ, kdy je $X_1, \ldots, X_n$ náhodný výběr z normálního rozdělení $N(\mu, \sigma^2)$.

Věta

- *M a S^2 jsou nezávislé náhodné veličiny.*
- $M \sim N(\mu, \sigma^2/n)$, *a tedy* $U = (M - \mu)/(\sigma/\sqrt{n}) \sim N(0, 1)$.
- $K = (n-1)S^2/\sigma^2 \sim \chi^2(n-1)$.
- $\sum(X_i - \mu)^2/\sigma^2 \sim \chi^2(n)$.
- $T = (M - \mu)/(S/\sqrt{n}) \sim t(n-1)$.

Poznámka

K odhadu μ, známe-li σ^2, slouží U, v opačném případě T.
K odhadu σ^2, neznáme-li μ, slouží K, v opačném případě následující (bezejmenná?) statistika, která je vlastně statistikou K, v níž místo odhadu M použijeme přímo μ.

Příklad

V roce 1951 bylo rozsáhlým statistickým průzkumem zjištěno, že střední hodnota výšky desetiletých chlapců je 136,1 cm se směrodatnou odchylkou $\sigma = 6{,}4$ cm.

Příklad

V roce 1951 bylo rozsáhlým statistickým průzkumem zjištěno, že střední hodnota výšky desetiletých chlapců je 136,1 cm se směrodatnou odchylkou $\sigma = 6{,}4$ cm.
V roce 1961 byla zjištěna výška pouze u 15 náhodně vybraných chlapců:

130	140	136	141	139	133	149	151
139	136	138	142	127	139	147	

Otázkou je, zda se v porovnání s rokem 1951 změnila střední výška chlapců, pokud předpokládáme, že variabilita výšek se v různých generacích příliš nemění.

Řešení

Vzhledem k tomu, že základní soubor všech desetiletých chlapců je rozsáhlý, lze zmíněná data považovat za náhodný výběr.Zjistíme, že $M = 139{,}133$, $n = 15$ a s využitím statistiky U dostáváme, že s 95% pravděpodobností leží hodnota μ v intervalu

$$(M - 1,96\sigma/\sqrt{n}; M + 1,96\sigma/\sqrt{n}) = (135{,}9; 142{,}4).$$

Protože i střední hodnota výšek z roku 1951 leží v tomto intervalu, nemá vážný důvod tvrdit, že se střední výška změnila.

Řešení

Vzhledem k tomu, že základní soubor všech desetiletých chlapců je rozsáhlý, lze zmíněná data považovat za náhodný výběr.Zjistíme, že $M = 139{,}133$, $n = 15$ a s využitím statistiky U dostáváme, že s 95% pravděpodobností leží hodnota μ v intervalu

$$(M - 1,96\sigma/\sqrt{n}; M + 1,96\sigma/\sqrt{n}) = (135{,}9; 142{,}4).$$

Protože i střední hodnota výšek z roku 1951 leží v tomto intervalu, nemá vážný důvod tvrdit, že se střední výška změnila. Pokud bychom ovšem připustili vyšší možnost omylu a stanovili interval se spolehlivostí pouze 90%, pak bychom na této hladině hypotézu, že se střední výška změnila, přijali – interval je nyní (136,41;141,85).

Řešení

Vzhledem k tomu, že základní soubor všech desetiletých chlapců je rozsáhlý, lze zmíněná data považovat za náhodný výběr.Zjistíme, že $M = 139{,}133$, $n = 15$ a s využitím statistiky U dostáváme, že s 95% pravděpodobností leží hodnota μ v intervalu

$$(M - 1,96\sigma/\sqrt{n}; M + 1,96\sigma/\sqrt{n}) = (135{,}9; 142{,}4).$$

Protože i střední hodnota výšek z roku 1951 leží v tomto intervalu, nemá vážný důvod tvrdit, že se střední výška změnila. Pokud bychom ovšem připustili vyšší možnost omylu a stanovili interval se spolehlivostí pouze 90%, pak bychom na této hladině hypotézu, že se střední výška změnila, přijali – interval je nyní (136,41;141,85). Podobně, pokud nás zajímá pouze **dolní odhad** střední hodnoty výšek chlapců (a vůbec tedy nepřipouštíme možnost, že by se střední výška snížila), pak s 95% pravděpodobností je střední výška větší než 136,41, a tedy nyní opět přijímáme hypotézu, že se střední výška zvýšila.

Dva nezávislé výběry z normálního rozdělení

Věta

Nechť je $X_{11}, \dots, X_{m1}$ *náhodný výběr rozsahu* m *z rozdělení* $N(\mu, \sigma_1^2)$ *a* $X_{12}, \dots, X_{n2}$ *je na něm nezávislý náhodný výběr rozsahu* n *z rozdělení* $N(\mu, \sigma_2^2)$*, přičemž* $m, n \geq 2$*. Označme* M_1, M_2 *jejich výběrové průměry a* S_1^2, S_2^2 *výběrové rozptyly. Dále nechť je*

$$S_*^2 = \frac{(m-1)S_1^2 + (n-1)S_2^2}{m+n-2}$$

vážený průměr výběrových rozptylů. *Pak platí:*

Dva nezávislé výběry z normálního rozdělení

Věta

Nechť je $X_{11}, \dots, X_{m1}$ *náhodný výběr rozsahu* m *z rozdělení* $N(\mu, \sigma_1^2)$ *a* $X_{12}, \dots, X_{n2}$ *je na něm nezávislý náhodný výběr rozsahu* n *z rozdělení* $N(\mu, \sigma_2^2)$*, přičemž* $m, n \geq 2$*. Označme* M_1, M_2 *jejich výběrové průměry a* S_1^2, S_2^2 *výběrové rozptyly. Dále nechť je*

$$S_*^2 = \frac{(m-1)S_1^2 + (n-1)S_2^2}{m+n-2}$$

vážený průměr výběrových rozptylů. *Pak platí:*

- $M_1 - M_2$ *a* S_*^2 *jsou stochasticky nezávislé,*

Dva nezávislé výběry z normálního rozdělení

Věta

Nechť je $X_{11}, \ldots, X_{m1}$ náhodný výběr rozsahu m z rozdělení $N(\mu, \sigma_1^2)$ a $X_{12}, \ldots, X_{n2}$ je na něm nezávislý náhodný výběr rozsahu n z rozdělení $N(\mu, \sigma_2^2)$, přičemž $m, n \geq 2$. Označme M_1, M_2 jejich výběrové průměry a S_1^2, S_2^2 výběrové rozptyly. Dále nechť je

$$S_*^2 = \frac{(m-1)S_1^2 + (n-1)S_2^2}{m+n-2}$$

vážený průměr výběrových rozptylů. *Pak platí:*

- $M_1 - M_2$ *a* S_*^2 *jsou stochasticky nezávislé,*
- $M_1 - M_2 \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n})$,

Dva nezávislé výběry z normálního rozdělení

Věta

Nechť je $X_{11}, \dots, X_{m1}$ náhodný výběr rozsahu m z rozdělení $N(\mu, \sigma_1^2)$ a $X_{12}, \dots, X_{n2}$ je na něm nezávislý náhodný výběr rozsahu n z rozdělení $N(\mu, \sigma_2^2)$, přičemž $m, n \geq 2$. Označme M_1, M_2 jejich výběrové průměry a S_1^2, S_2^2 výběrové rozptyly. Dále nechť je

$$S_*^2 = \frac{(m-1)S_1^2 + (n-1)S_2^2}{m+n-2}$$

vážený průměr výběrových rozptylů. *Pak platí:*

- $M_1 - M_2$ *a* S_*^2 *jsou stochasticky nezávislé,*
- $M_1 - M_2 \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n})$,
- *je-li* $\sigma_1^2 = \sigma_2^2 = \sigma^2$, *pak*
  $K = (m+n-2)S_*^2/\sigma^2 \sim \chi^2(m+n-2)$,

Dva nezávislé výběry z normálního rozdělení

Věta

Nechť je $X_{11}, \dots, X_{m1}$ *náhodný výběr rozsahu* m *z rozdělení* $N(\mu, \sigma_1^2)$ *a* $X_{12}, \dots, X_{n2}$ *je na něm nezávislý náhodný výběr rozsahu* n *z rozdělení* $N(\mu, \sigma_2^2)$*, přičemž* $m, n \geq 2$*. Označme* M_1, M_2 *jejich výběrové průměry a* S_1^2, S_2^2 *výběrové rozptyly. Dále nechť je*

$$S_*^2 = \frac{(m-1)S_1^2 + (n-1)S_2^2}{m+n-2}$$

vážený průměr výběrových rozptylů. *Pak platí:*

- $M_1 - M_2$ *a* S_*^2 *jsou stochasticky nezávislé,*
- $M_1 - M_2 \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n})$,
- *je-li* $\sigma_1^2 = \sigma_2^2 = \sigma^2$*, pak*
  $K = (m+n-2)S_*^2/\sigma^2 \sim \chi^2(m+n-2)$,
- $F = \frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F(m-1, n-1).$

Užití statistik dvou nezávislých výběrů

- Statistika U, vzniklá normováním $M_1 - M_2$, se používá pro odhad rozdílu $\mu_1 - \mu_2$, známe-li rozptyly σ_1^2, σ_2^2.

Užití statistik dvou nezávislých výběrů

- Statistika U, vzniklá normováním $M_1 - M_2$, se používá pro odhad rozdílu $\mu_1 - \mu_2$, známe-li rozptyly σ_1^2, σ_2^2.
- Je-li $\sigma_1^2 = \sigma_2^2 = \sigma^2$, pak statistika T (vzniklá z U nahrazením teoretického společného rozptylu σ^2 váženým průměrem výběrových rozptylů S_*^2) slouží pro odhad rozdílu $\mu_1 - \mu_2$, neznáme-li rozptyl σ^2.

Užití statistik dvou nezávislých výběrů

- Statistika U, vzniklá normováním $M_1 - M_2$, se používá pro odhad rozdílu $\mu_1 - \mu_2$, známe-li rozptyly σ_1^2, σ_2^2.
- Je-li $\sigma_1^2 = \sigma_2^2 = \sigma^2$, pak statistika T (vzniklá z U nahrazením teoretického společného rozptylu σ^2 váženým průměrem výběrových rozptylů S_*^2) slouží pro odhad rozdílu $\mu_1 - \mu_2$, neznáme-li rozptyl σ^2.
- Statistika $K = (m + n - 2)S_*^2/\sigma^2$ slouží k odhadu společného rozptylu σ^2.

Užití statistik dvou nezávislých výběrů

- Statistika U, vzniklá normováním $M_1 - M_2$, se používá pro odhad rozdílu $\mu_1 - \mu_2$, známe-li rozptyly σ_1^2, σ_2^2.
- Je-li $\sigma_1^2 = \sigma_2^2 = \sigma^2$, pak statistika T (vzniklá z U nahrazením teoretického společného rozptylu σ^2 váženým průměrem výběrových rozptylů S_*^2) slouží pro odhad rozdílu $\mu_1 - \mu_2$, neznáme-li rozptyl σ^2.
- Statistika $K = (m+n-2)S_*^2/\sigma^2$ slouží k odhadu společného rozptylu σ^2.
- Statistika $F = \frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2}$ slouží k odhadu podílu rozptylů σ_1^2/σ_2^2.

Příklad

Mějme dva nezávislé náhodné výběry; první rozsahu 10 z rozdělení $N(2; 1{,}5)$ a druhý rozsahu 5 z rozdělení $N(3, 4)$. Určete pravděpodobnost, že výběrový průměr prvního výběru bude menší než výběrový průměr druhého výběru.

Příklad

Mějme dva nezávislé náhodné výběry; první rozsahu 10 z rozdělení $N(2; 1{,}5)$ a druhý rozsahu 5 z rozdělení $N(3, 4)$. Určete pravděpodobnost, že výběrový průměr prvního výběru bude menší než výběrový průměr druhého výběru.

Řešení

$$P(M_1 < M_2) = P(M_1 - M_2 < 0) =$$

$$= P\left(\frac{(M_1 - M_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} < \frac{0 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}}\right) =$$

$$P\left(U < \frac{-2+3}{\sqrt{\frac{1{,}5}{10} + \frac{4}{5}}}\right) = P(U < 1{,}05) =$$

$$= \Phi(1{,}05) = 0{,}853.$$