Základy pravděpodobnosti a statistiky

Ondřej Pangrác

6. června 2023

1 Diskrétní náhodné veličiny

Střední hodnota (reálné) náhodné veličiny X je

$$EX = \sum_{\omega \in \Omega} X(\omega) P(\{\omega\})$$

Ekvivalentně lze (pro konečné, resp. spočetné pravděpodobnostní prostory) psát

$$EX = \sum_{x \in Im(X)} x \cdot P(X = x)$$

a pokud X nabývá pouze nezáporných celočíselných hodnot, pak

$$EX = \sum_{n=0}^{\infty} P(X > n)$$

Linearita střední hodnoty

$$\begin{aligned} E(\alpha X) &= \alpha EX \\ E(X + Y) &= EX + EY \end{aligned}$$

Pokud jsou navíc X a Y nezávislé, pak také

$$E(XY) = EX \cdot EY$$

Rozptyl náhodné veličiny X

$$Var(X) = E((X - EX)^2)$$

lze také vyjádřit jako

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

potom $\sigma_X = \sqrt{Var(X)}$ se nazývá směrodatná odchylka.

Analogií linearity střední hodnoty je u rozptylu vztah

$$Var(\alpha X + \beta) = \alpha^2 Var(X)$$

Kovariance dvojice náhodných veličin X, Y

$$cov(X, Y) = E((X - EX)(Y - EY)) = E(XY) - EX \cdot EY$$

(je rovna 0, pokud jsou X, Y nezávislé).

Korelace

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

Pokud jsou X a Y nezávislé, je $\rho = 0$, obecně platí $-1 \leq \rho \leq 1$.

1.1 Příklady diskrétních rozdělení

Bernouliho (alternativní) rozdělení

$$Alt(p):$$
$$P(X=1)=p$$
$$P(X=0)=1-p$$
$$EX=p$$
$$Var(X)=p(1-p)$$

Binomické rozdělení

Počet úspěchů při n nezávislých opakováních téhož experimentu.

$$Binom(n,p):$$
$$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k} \text{ pro } k=0,1,\ldots,n$$
$$EX=np$$
$$Var(X)=np(1-p)$$

Hypergeometrické rozdělení

V osudí a bílých a b černých míčků, bez vracení vylosuji n. Kolik z nich je bílých?

$$Hyper(a,b,n):$$
$$P(X=k)=\frac{\binom{a}{k}\binom{n}{n-k}}{\binom{a+b}{n}} \text{ pro } k=max(0,n-b),...min(n,a)$$
$$EX=n\cdot\frac{a}{a+b}$$
$$Var(X)=n\cdot\frac{a}{a+b}\cdot\frac{b}{a+b}\cdot\frac{a+b-n}{a+b-1}$$

Geometrické rozdělení

Kolikátým opakováním dosáhnu poprvé úspěchu?

$$Geom(p):$$
$$P(X=k)=(1-p)^{1-k}p \text{ pro } k=1,2,\ldots$$
$$EX=\sum_{n=0}^{\infty}(1-p)^n=\frac{1}{1-(1-p)}=\frac{1}{p}$$

Poissonovo rozdělení

$Pois(\lambda):$
$P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}$ pro $k = 0, 1, \ldots$
$EX = \lambda$
$Var(X) = \lambda$

$Pois(\lambda)$ je limitou $Binom(n, \frac{\lambda}{n})$, $n \to \infty$

2 Spojité náhodné veličiny

Distribuční funkce náhodné veličiny X je $F_X : \mathbb{R} \to [0, 1]$, pro kterou platí $F_X(x) = P[X \leq x]$.

Vlastnosti distribuční funkce:

- F_X je neklesající
- $\lim_{x\to-\infty} F_X(x) = 0$
- $\lim_{x\to+\infty} F_X(x) = 1$

Nezáporná funkce f_X se nazývá hustoutou náhodné veličiny X, pokud je $F_X(t) = \int_{-\infty}^{t} f_X(x)\,dx$. Pak lze vyjádřit $P[a \leq X \leq b] = \int_a^b f(x)\,dx$.

Střední hodnota (spoj. náh. veličiny X)

$$E(X) = \int_{-\infty}^{\infty} x \cdot f_X(x)\,dx$$

(pokud tento integrál má smysl)

Rozptyl

$$var(X) = E((X - EX)^2) = E(X^2) - (EX)^2$$

kde $E(X^2) = \int_{-\infty}^{\infty} x^2 \cdot f_X(x)\,dx$.

Uniformní rozdělení na intervalu $[a, b]$

$f_X(x) = \frac{1}{b-a}$ pro $x \in [a, b]$, jinak je $f_X(x) = 0$

distribuční funkce je

$$\begin{aligned}
F_X(x) &= \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x \leq b \\ 1 & b < x \end{cases} \\
E(X) &= \ldots = \frac{b+a}{2} \\
E(X^2) &= \ldots = \frac{b^2 + ba + a^2}{3} \\
E(X) &= \ldots = \frac{(b-a)^2}{12}
\end{aligned}$$

Exponenciální rozdělení s parametrem $\lambda > 0$

$$\begin{aligned}
F_X(x) &= \begin{cases} 0 & x < 0 \\ 1 - e^{-\lambda x} & 0 \leq x \end{cases} \\
f_X(x) &= \begin{cases} 0 & x < 0 \\ \lambda \cdot e^{-\lambda x} & 0 \leq x \end{cases} \\
E(X) &= \ldots = \frac{1}{\lambda} \\
E(X^2) &= \ldots = \frac{2}{\lambda^2} \\
var(X) &= \ldots = \frac{1}{\lambda^2}
\end{aligned}$$

Normální rozdělení $N(\mu, \sigma)$

nejprve definujeme $N(0,1)$:

hustota $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, distribuční funkce $\Phi(x)$ nemá vzorec a lze spočítat numericky

obecné parametry

$$X \sim N(\mu, \sigma) \leftrightarrow Z = \frac{X - \mu}{\sigma} \sim N(0,1)$$

tedy $N(\mu, \sigma)$ má hustotu $\frac{1}{\sigma}\varphi(\frac{x-\mu}{\sigma})$

zjevně je μ střední hodnota a σ^2 rozptyl

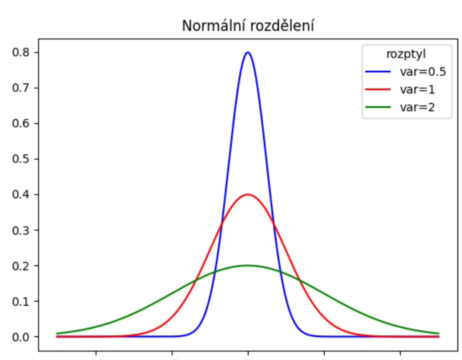
odolnost vůči součtu:

$X_i \sim N(\mu_i, \sigma_i), i = 1, \ldots, k$, pak $X = X_1 + \ldots + X_k \sim N(\mu, \sigma)$, kde $\mu = \mu_1 + \ldots + \mu_k$ a $\sigma^2 = \sigma_1^2 + \ldots + \sigma_k^2$

pravidlo 3σ:

pro $X \sim N(\mu, \sigma)$

$$\begin{aligned}
P\left[|X - \mu| < \sigma\right] &\doteq 0.68 \\
P\left[|X - \mu| < 2\sigma\right] &\doteq 0.95 \\
P\left[|X - \mu| < 3\sigma\right] &\doteq 0.997
\end{aligned}$$



Cauchyho rozdělení

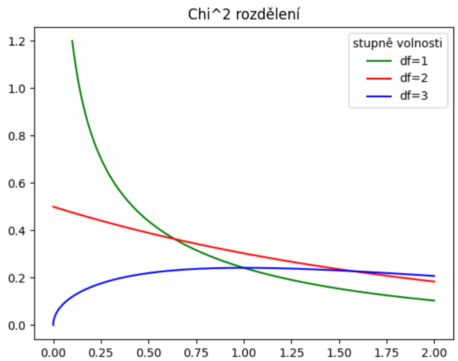
hustota $f(x) = \frac{1}{\pi(1+x^2)}$ připomíná (trochu rozpláclé) normální rozdělení, je to sudá funkce

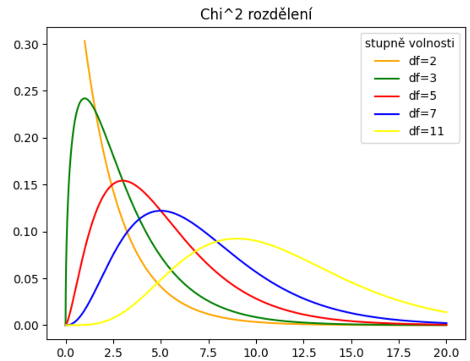
distribuční funkce je $F(x) = \frac{1}{\pi} \arctan x + \frac{1}{2}$

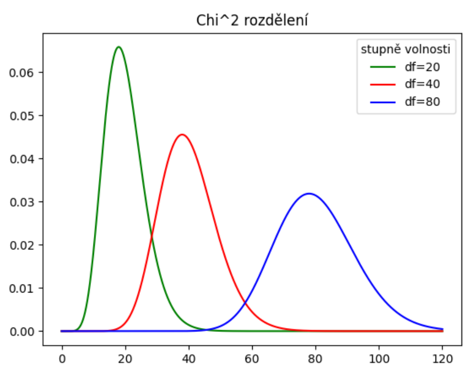
nemá ale střední hodnotu, neboť $\int_{-\infty}^{\infty} x \cdot f(x)\ dx$ není definován (výraz typu $\infty - \infty$) i přesto, že pro libovolné kladné t je $\int_{-t}^{t} x \cdot f(x)\ dx = 0$

χ^2 rozdělení

rozdělení součtu druhých mocnin nezávislých normálně rozdělených náhodných veličin, přesněji pro $X_1, \ldots, X_k \sim N(0,1)$ má $Q = \sum_{i=1}^{k} X_i^2$ rozdělení χ^2 s k stupni volnosti (degrees of freedom, df), $E(Q) = k, var(Q) = 2k$







3 Nerovnosti a limitní věty

Markovova nerovnost

X nezáporná reálná náhodná veličina, $a > 0$

$$P[X \geq a] \leq \frac{E(X)}{a}$$

(alternativní verze $P\left[X \geq t \cdot E(X)\right] \leq \frac{1}{t}$, pro $t \geq 1$)

Čebyševova nerovnost

$$P\left[|X - E(X)| \geq a \cdot \sqrt{var(X)}\right] \leq \frac{1}{a^2}$$

kde $a > 0$

Černovova nerovnost

X_i nabývá hodnot ± 1, obě s pravděpodobností $\frac{1}{2}$, $i = 1, \ldots, n$
$X = \sum_{i=1}^{n} X_i$, $t > 0$
pak $E(X) = 0$ a $\sigma^2 = var(X) = n$

$$P\left[X \leq -t\right] = P\left[X \geq t\right] \leq e^{\frac{-t^2}{2\sigma^2}}$$

Slabý zákon velkých čísel

$X_1, \ldots, X_n$ stejně rozdělené nezávislé náhodné veličiny se střední hodnotou μ a rozptylem σ^2. Ať $S_n = \frac{X_1 + \ldots + X_n}{n}$ (výběrový průměr), pak pro $\forall \varepsilon > 0$ je

$$\lim_{n \to \infty} P\left[|S_n - \mu| > \varepsilon\right] = 0$$

(použitím Čebyševovy nerovnosti dostaneme odhad $P\left[|S_n - \mu| > \varepsilon\right] \leq \frac{\sigma^2}{n\varepsilon^2}$)

Silný zákon velkých čísel

$X_1, \ldots, X_n$ stejně rozdělené nezávislé náhodné veličiny se střední hodnotou μ a rozptylem σ^2. Ať $S_n = \frac{X_1 + \ldots + X_n}{n}$, pak platí

$$\lim_{n \to \infty} S_n = \mu \text{ skoro jistě (tj. s pravděpodobností jedna)}$$

Centrální limitní věta

$X_1, \ldots, X_n$ stejně rozdělené nezávislé náhodné veličiny se střední hodnotou μ a rozptylem σ^2. Ať $Y_n = \frac{(X_1 + \ldots + X_n) - n \cdot \mu}{\sqrt{n}\sigma}$. Pak Y_n konverguje k $N(0, 1)$ v distribuci, tj. distribuční funkce $F_n = F_{Y_n}$ má limitu $\lim_{n \to \infty} F_n(x) = \Phi(x)$ pro $\forall x \in \mathbb{R}$.

4 Cvičení

Příklad 1:
Koupíme se po jednom losu ve dvou různých loteriích. V první je pravděpodobnost výhry 20%, ve druhé 10%. Jaká je pravděpodonost, že vyhraji v alespoň jedné z nich?

Příklad 2:
V osudí jsou míčky dvou barev, bílé a černé, od každé barvy stejný počet. Vylosuji jeden míček, podívám se na něj a vrátím ho do osudí. Poté vylosuji druhý míček. Jaká je pravděpodobnost, že oba mají stejnou barvu? Řešte pro tři barvy, respktive pro dvě barvy, ale bez vracení prvního míčku zpět do osudí.

Příklad 3:
Špatně seřízený automat vyrábí 30% zmetků. Jaká je pravděpodobnost, že při kontrole 3 výrobků zjistím nejvýše 1 zmetek?

Příklad 4:
V závislosti na $n \in \mathbb{N}$ určete pravděpodonost, že mezi n lidmi se najdou dva, kteří mají narozeniny ve stejný den? (narozeniny = číslo z množiny $\{1, 2, \ldots, 365\}$ volené náhodně a uniformě)

Příklad 5:
Balíček obsahuje n karet označených různými čísly. Balíček promícháme a poté otáčíme karty postupně od první až do k-té. Jaká je pravděpodobnost, že k-tá otočená karta je největší ze všech, jestliže ja největší z dosud otočených?

Příklad 6:
V učité skupině lidí je 55% žen a 45% mužů. Pro danou chorobu je pravděpodobnost onemocnění u ženy 1% a u muže 5%.

a) Jaká je pravděpodobnost, že náhodně vybraná osoba trpí touto chorobou?

b) Jaká je pravděpodobnost, že náhodně vybraná osoba trpící touto chorobou je muž?

Příklad 7:
Hodím dvěmi kostkami a jako jev A označím skutečnost, že součet hodů je sudý, jako jev B označím, že součet je dělitelný 3. Určete podmíněné pravděpodobnosti $P[A|B]$ a $P[B|A]$.

Příklad 8:
Test nemoci je pozitivní s pravděpodobností 99%, je-li testovaná osoba skutečně nemocná. Testuji 30 osob, všichni z nich jsou nemocní. Jaká je pravděpodobnost, že test bude ve všech případech pozitivní?

Příklad 9:
Mějme n osudí, v k-tém osudí je $k-1$ bílých a $n-k$ černých míčků (pro $k = 1, 2, \dots, n$). Nejprve vyberu náhodně (uniformě) jedno z osudí a z něj vylosuji dva míčky (bez vracení). Jaká je pravděpodobnost, že

a) první míček je černý,

b) druhý míček je černý,

c) druhý míček je černý, jestliže první je také černý.

Příklad 10:
Dokažte následující tvrzení.

a) Je-li jev A nezávislý sám na sobě, je jeho pravděpodobnost rovna 0 nebo 1.

b) Je-li $P[A]$ rovna 0 nebo 1, je A nezávislý na libovolném jevu B.

Příklad 11:
Dokažte, že pro dvě nezávislé náhodné veličiny X, Y platí $E(XY) = EX \cdot EY$.

Příklad 12:
V osudí je 200 losů, z nichž 10 vyhrává. Jaká je pravděpodobnost, že alespoň jeden z 10 zakoupených losů bude výherní?

Příklad 13:
Máme tři šestistěné kostky. První má na všech stěnách číslo 1, druhá na třech stěnách 1 a na třech 2 a třetí má na dvou stěnách 1, na dvou 2 a na dvou 3. Hodíme všemi třemi a spočteme součet hozených čísel. Určete střední hodnotu výsledku.

Příklad 14:
Hodím dvěmi šestistěnými kostkami, černou a bílou. Jaká je pravděpodonost, že na černé padlo větší číslo, než na bílé? Jaká je střední hodnota součtu na obou kostkách a střední hodnota součinu?

Příklad 15:
Lék na dané onemocnění je účiný v 90% případů. Jaká je pravděpodobnost, že pomůže alespoň 8 nemocným z 10, kterým byl podán?

Příklad 16:
V bedně je 50 výrobků. Odběratel vybere náhodně 10 z nich a zkontorluje je. Dodávku přijme, pokud žádný z testovaných není vadný. Jaká je pravděpodobnost přijetí dodávky, je-li ve skutečnosti v bedně 5 vadných výrobků?

Příklad 17:
Z 10 čísel v osudí napíši na papír 5, poté se vylosuje také 5 čísel. Určete střední hodnotu počtu uhodntých čísel.

5 Statistika

Statistika je nejen vědní obor, ale také má toto slovo technický význam, znamenající libovolnou funkci spočtenou z náhodného výběru.

Náhodný výběr – uniformě náhodně bez vracení / s vracením

Náh. vektor $X = (X_1, \ldots, X_n)$ je náhodný výběr z pravděpodobnostního rozdělení P, pokud (i) $X_1, \ldots, X_n$ jsou nezávislé náhodné veličiny a (ii) všechny mají stejné rozdělení pravděpodobnosti P.

- bodové odhady – chceme určit hodnotu parametru (např. střední hodnota nějakého rozdělení apod.)
- intervalové odhady – interval, v kterém se parametr nachází s určitou pravděpodobností
- testování hypotéz – podporují nebo vyvracejí naměřená data naši hypotézu?
- regrese – závislost mezi dvěmi naměřenými veličinami

parametrické vs. neparametrické modely – hledáme model vysvětlující naměřená data

- parametrické – hledáme vhodný model z určité množiny (např. předpokládáme, že data jsou normálně rozdělená a hledáme parametry μ, σ pro $N(\mu, \sigma)$)
- neparametrické – výběr ze všech možných distribučních funkcí

Geoge Box: „Všechny modely jsou špatné, ale některé jsou užitečné.“

6 Bodové odhady

$X_1, \ldots, X_n$ náhodný výběr, ϑ parametr rozdělení, θ_n odhad tohoto parametru z náhodného výběru
střední kvadratická chyba $MSE_\vartheta(\theta_n) = E((\theta_n - \vartheta)^2)$

Výběrový průměr a rozptyl

- výběrový průměr $\overline{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$
- výběrový rozptyl $S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \overline{X})^2$
- výběrový rozptyl $\overline{S}_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \overline{X})^2$

6.1 Metoda momentů

r-tý moment $m_r = E(X^r)$, jeho odhad $\overline{m}_r = \frac{1}{n}\sum_{i=1}^{n} X_i^r$
pokud má hledaná statistika k parametrů, obvykle vyjdeme ze soustavy

$$m_r = \overline{m}_r \qquad r = 1, \ldots, k$$

Příklad.
$X_1, \ldots, X_n$ náhodný výběr z $N(\mu, \sigma)$, chceme odhadnout parametry μ, σ

$$\begin{aligned} m_1 &= \mu \\ m_2 &= EX^2 = var(X) + (EX)^2 = \sigma^2 + \mu^2 \end{aligned}$$

takže

$$\begin{aligned} \mu &= m_1 = \overline{m}_1 = \frac{X_1 + \ldots + X_n}{n} \\ \sigma^2 + \mu^2 &= m_2 = \overline{m}_2 = \frac{X_1^2 + \ldots + X_n^2}{n} \end{aligned}$$

z toho

$$\begin{aligned} \overline{\mu} &= \frac{X_1 + \ldots + X_n}{n} = \overline{X}_n \\ \overline{\sigma} &= \sqrt{m_2 - m_1^2} = \sqrt{\frac{X_1^2 + \ldots + X_n^2}{n} - \overline{X}_n} \end{aligned}$$

7 Intervalové odhady

Dvojice $t_1 \leq t_2$ určuje intervalový odhad statistiky ϑ o spolehlivosti $1 - \alpha$ ($1 - \alpha$ confidence interval), pokud $P[t_1 \leq \vartheta \leq t_2] \geq 1 - \alpha$. Mimo dostatečné spolehlivosti chceme mít interval $[t_1, t_2]$ co nejkratší. Tyto požadavky jsou ale svým způsobem protichůdné.

Nejprve se zaměříme na vlastnosti normálního rozdělení:

$$\begin{aligned} X_i \sim N(\mu_i, \sigma_i), i = 1, \ldots, n &\Rightarrow \sum_{i=1}^{n} \alpha_i X_i \sim N(\mu, \sigma), \mu = \sum_{1}^{n} \alpha_i \mu_i, \sigma^2 = \sum_{1}^{n} \alpha_i^2 \sigma_i^2 \\ X \sim N(\mu, \sigma) &\Rightarrow U = \frac{X - \mu}{\sigma} \sim N(0, 1) \\ X_1, \ldots, x_n \sim N(0, 1) &\Rightarrow K = \sum_{i=1}^{n} X_i^2 \sim \chi^2(n) \\ K_i \sim \chi^2(n_i), i = 1, \ldots, k &\Rightarrow K = \sum_{i=1}^{n} K_i \sim \chi^2(n_1 + \ldots, n_k) \end{aligned}$$

$$U \sim N(0,1), K \sim \chi^2(n) \quad \Rightarrow \quad T = \frac{U}{\sqrt{K/n}} \sim t(n)$$
$$K_i \sim \chi^2(n_i), i = 1,2 \quad \Rightarrow \quad F = \frac{K_1/n_1}{K_2/n_2} \sim F(n_1, n_2)$$

kde $t(n)$ je Studentovo t-rozdělení a $F(n_1, n_2)$ Fischerovo rozdělení (známá rozdělení s hodnotami v tabulkách a softwareových knihovnách).

Poznamenejme na okraj, že pro F-rozdělení bývají v tabulkách uváděny pouze hodnoty $\alpha \geq 0.5$. Z definice je patrné, že $F^* = \frac{K_2/n_2}{K_1/n_1} = \frac{1}{F} \sim\sim F(n_2, n_1)$ a tedy $F_{1-\alpha}(n_1, n_2) = 1/F_\alpha(n_2, n_1)$.

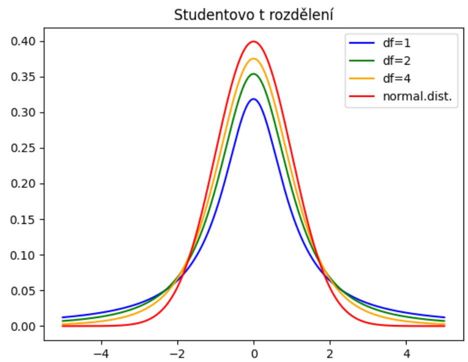
Z výše uvedených vztahů lze odvodit rozdělení výběrových statistik normálního rozdělení. Jsou-li $X_1, \ldots, X_n \sim N(\mu, \sigma)$, $\overline{X} = \overline{X}_n = \frac{1}{n}\sum_1^n X_i$ výběrový průměr a $\overline{S}^2 = \overline{S}_n^2 = \frac{1}{n-1}\sum_1^n (X_i - \overline{X})^2$ výběrový rozptyl, pak

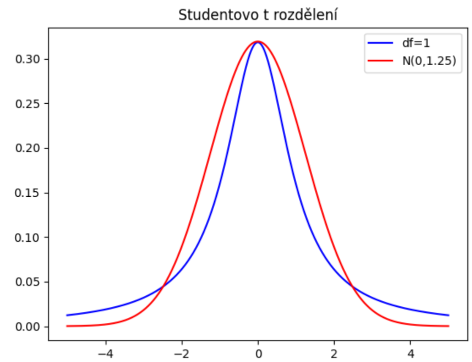
výběrový průměr	$\overline{X} \sim N(\mu, \sigma/\sqrt{n})$
statistika	$U = \dfrac{\overline{X} - \mu}{\sigma}\sqrt{n} \sim N(0,1)$
statistika	$K = \dfrac{n-1}{\sigma^2}\overline{S}^2 \sim \chi^2(n-1)$
statistika	$T = \dfrac{\overline{X} - \mu}{\overline{S}}\sqrt{n} \sim t(n-1)$

Statistiky U, K, T se nazývají pivotové statistiky, přičemž U je pivotovou stat. pro μ při známém σ, K je pivotovou stat. pro σ^2 a T je pivotovou stat. pro μ při neznámém σ. Příslušné intervalové odhady jsou následující.

Věta 7.1 *Mějme $X_1, \ldots, X_n$ náhodný výběr z $N(\mu, \sigma)$, přičemž σ je známá a μ chceme odhadnout. Pro $\alpha \in (0,1)$ mějme $z_{\alpha/2}$ takové, že $\Phi(z_{\alpha/2}) = 1 - \frac{\alpha}{2}$. Pak interval $C = [\overline{X}_n - z_{\alpha/2} \cdot \sigma/\sqrt{n}, \overline{X}_n + z_{\alpha/2} \cdot \sigma/\sqrt{n}]$ je intervalovým odhadem μ o spolehlivosti $1 - \alpha$, tedy $P[\mu \in C] = 1 - \alpha$.*

Chceme-li odhadovat střední hodnotu normálně rozdělené náhodné veličiny s neznámým roptylem, použijeme statistiku T a distribuční funkci Studentova rozdělení s $n-1$ stupni volnosti, která se označuje Ψ_{n-1} (a je v tabulkách a softwareových knihovnách). Pokud bychom v tomto vzorci statistiky T nahradili $\overline{S}_n$ směrodatnou odchylkou σ, pak $\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$ má rozdělení $N(0,1)$. Jelikož $\overline{S}_n$ je asymptotickou aproximací σ, bude Ψ_n konvergovat (s rostoucím n) k distribuci $N(0,1)$.





Věta 7.2 *Mějme $X_1, \dots, X_n$ náhodný výběr z $N(\mu, \sigma)$, přičemž σ je neznámá a μ chceme odhadnout. Pro $\alpha \in (0,1)$ mějme $t_{\alpha/2}$ takové, že $\Psi_{n-1}(t_{\alpha/2}) = 1 - \frac{\alpha}{2}$ a $\delta = t_{\alpha/2} \cdot \frac{\overline{S}_n}{\sqrt{n}}$. Pak $P\left[\mu \in [\overline{X}_n - \delta, \overline{X}_n + \delta]\right] = 1 - \alpha$.*

Pro intervalové odhady rozptylu (resp. směrodatné odchylky) se využijeme pivotovou statistiku K.

Věta 7.3 *Nechť $X_1, \dots, X_n$ je z normální distribuce a $\overline{S} = \overline{S}_n^2$ výběrový rozptyl. Ať a, b jsou taková, že $\chi^2(n-1)(a) = \alpha/2$ a $\chi^2(n-1)(b) = 1 - \alpha/2$. Pak intervalový odhad rozptylu o spolehlivosti $1 - \alpha$ je $[\frac{(n-1)\overline{S}^2}{a}, \frac{(n-1)\overline{S}^2}{b}]$ a směrodatné odchylky $[\overline{S} \cdot \sqrt{\frac{n-1}{a}}, \overline{S} \cdot \sqrt{\frac{n-1}{b}}]$.*

Pokud nemáme garantované normální rozdělení, lze k odhadu využít Centrální limitní věty.

Věta 7.4 *Mějme $X_1, \dots, X_n$ náhodný výběr z nějakého rozdělení, přičemž jeho směrodatná odchylka σ je známá a střední hodnotu μ chceme odhadnout. Pro $\alpha \in (0,1)$ mějme $z_{\alpha/2}$ takové, že $\Phi(z_{\alpha/2}) = 1 - \frac{\alpha}{2}$. Označme $\overline{X}_n = \frac{X_1 + \ldots + X_n}{n}$ a $C = [\overline{X}_n - z_{\alpha/2} \cdot \sigma/\sqrt{n}, \overline{X}_n + z_{\alpha/2} \cdot \sigma/\sqrt{n}]$.*

Pak $\lim_{n \to \infty} P[\mu \in C] = 1 - \alpha$.

Na rozdíl od předchozí věty zde pravděpodobnost nabývá požadované hodnoty jen v limitě.

8 Testování hypotéz

Nulová hypotéza H_0 (null hypothesis) – výchozí předpoklad, že se naměřená data shodují s modelem

Alternativní hypotéza H_1 – naplatí H_0

výsledek testu pak může mít dva možné výsledky

- přijetí nulové hypotézy – data jsou v souladu s naším předpokladem
- zamítnutí nulové hypotézy – data odporují, model je chybný

data jsou výsledkem náhodného procesu, takže mohou nastat chyby

chyba 1. druhu (Type 1 error) – zamítneme H_0, přestože platí

chyba 2. druhu (Type 2 error) – přijmeme H_0, i když neplatí

pravděpodobnost chyby 1. druhu se nazývá hladina významnosti testu (significance level) – základní parametr testu, který si obvykle můžeme nastavit (klasicky se volí na úrovni 5%, R. Fisher 1925)

postup testu pak vypadá (může vypadat):

1. máme testovanou statistiku $T = f(X_1, \ldots, X_n)$
2. nastavíme hladinu významnosti α
3. urrčíme kritický obor (rejection region) W
4. „naměříme“ data $X_1, \ldots, X_n$
5. pokud $f(X_1, \ldots, X_n) \in W$, zamítneme H_0

pravděpodobnost chyby 1. druhu je (volena) $P\left[f(X) \in W|H_0\right] = \alpha$, pak pravděpodobnost chyby 2. druhu $P\left[f(X) \notin W|H_1\right] = \beta$ závisí na vlastnostech statistického testu a určuje tzv. sílu testu $1 - \beta$

Základní testy (zde pro střední hodnotu) postupují stejně jako intervalové odhady.

8.1 Z test

pro střední hodnotu

Jednovýběrový test „Je střední hodnota rovna μ?“

Modelem je $N(\mu, \sigma)$ (známe oba parametry) a otázkou je, zda data $X_1, \ldots, X_n$ odpovídají modelu se danou střední hodnotou μ. Pokud by tomu tak bylo, pak $Z = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$ by měla distibuci $N(0, 1)$. Na základě $\alpha \in (0, 1)$ určíme $z_{\alpha/2}$ (viz. výše u intervalových odhodů) a $W = \{x : |x| > z_{\alpha/2}\}$. H_0 přijmeme, pokud $Z \notin W$.

Dvouvýběrový test „Mají dvě distribuce stejnou střední hodnotu?“

$X_1, \ldots, X_n$ náhodný výběr z $N(\mu_1, \sigma)$, $Y_1 \ldots, Y_m$ náhodný výběr z $N(\mu_2, \sigma)$, přičemž rozptyl je v obou ditribucích stejný a známý.

H_0: $\mu_1 = \mu_2$ (H_1: $\mu_1 \neq \mu_2$)

pro $S = \overline{X}_n - \overline{Y}_n$ je, pokud platí H_0, střední hdonota rovna 0 a rozptyl

$$var(S) = var(\overline{X}_n) + var(\overline{Y}_m) = \frac{\sigma^2}{n} + \frac{\sigma^2}{m}$$

(z nezávislosti výběrů a tedy nez. veličin $\overline{X}_n$ a $\overline{Y}_m$)

Dále již postupujeme stejně jako u jednovýběrového testu, protože (při platnosti H_0) má $Z = \frac{S}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}} = \frac{\overline{X}_n-\overline{Y}_m}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}}$ distibuci $N(0,1)$. Další postup (určení $z_{\alpha/2}$ a W) je totožný.

Párový test

Náhodný výběr se skládá z dvojic $(X_1, Y_1), \ldots, (X_n, Y_n)$, přičemž uvnitř dvojic nemusí být veličiny nezávislé. Ověřujeme hypotézu, že rozdíl hodnot ve dvojidích je nulový, nebo obecněji roven nějaké konstantě. Tento případ převedeme na jednovýběrový test střední hodnoty náhodné veličiny $X_i - Y_i$

8.2 Studentův T test

odstraňuje předpoklad známého rozptylu

Jednovýběrový test „Je střední hodnota rovna μ?“

Nulová hypotéza předpokládá, že náhodný výběr $X_1, \ldots, X_n$ pochází z distribuce $N(\mu, \sigma)$, kde předpokládáme znalost střední hodnoty μ, ale neznáma σ.

Postup testu je totožný s jednovýběrovým Z testem, jen místo známého rozptylu σ^2 použijeme jeho odhad $\overline{S}_n^2 = \frac{\sum_{i=1}^n (X_n-\overline{X}_n)^2}{n-1}$.

Platí-li hypotéza H_0, pak má $T = \frac{\overline{X}_n-\mu}{\overline{S}_n/\sqrt{n}}$ Studentovo rozdělení s $n-1$ stupni volnosti. Pro určení kritického oboru W pak místo distribuce Φ použijeme Ψ_{n-1}.

Dvouvýběrový test pro shodné rozptyly

Předpokládáme, že náhodný výběr $X_1, \ldots, X_n$ je z distribuce $N(\mu_1, \sigma)$ a $Y_1 \ldots, Y_m$ z distribuce $N(\mu_2, \sigma)$, obě rozdělení mají stejný rozptyl. Ten můžeme odhadnout

$$\overline{S}^2 = \frac{\sum_{i=1}^n (X_i - \overline{X}_n)^2 + \sum_{i=1}^m (Y_i - \overline{Y}_n)^2}{n+m-2} = \frac{(n-1)\overline{S}_X^2 + (m-1)\overline{S}_Y^2}{n+m-2}$$

a $T = \frac{\overline{X}_n-\overline{Y}_m}{\overline{S}\sqrt{\frac{1}{n}+\frac{1}{m}}}$ má (pokud platí H_0: $\mu_1 = \mu_2$) distribuci Studentova rozdělení s n+m-2 stupni volnosti.

Dvouvýběrový test pro neshodné rozptyly

Jako odhad směrodatné odchylky použijeme $SE(\overline{X}_n - \overline{Y}_m) = \sqrt{\frac{\overline{S}_X^2}{n} + \frac{\overline{S}_Y^2}{m}}$ a $T = \frac{\overline{X}_n-\overline{Y}_m}{SE(\overline{X}_n-\overline{Y}_m)}$ má přibližně Studento rozdělení, počet stupňů volnosti závisí na $n, m, \overline{S}_X, \overline{S}_Y$ a je nejvýše $n+m-2$.

8.3 Test dobré shody

Na rozdíl od předchozích testů určených pro numerická data, je smyslem tohoto testu interpretovat „kategoriální data“. Umožňuje nám ověřit, zda náhodná veličina má dané rozdělení u kterého přepokládáme, že generuje několik (konečně mnoho ne nutně numerických) hodnot. Příkladem může být například test, zda hrací kostka háže každou z hodnot se stejnou pravděpodobností.

Model je takový, že experiment má k možných výsledků s pravděpodobnostmi $p_1, \ldots, p_k$ ($\sum_1^k p_i = 1$). Opakuji n-krát pokus a označím X_i počet dosažených výsledků i-tého typu. Pak tento model má multinomické rozdělení, tedy

$$P\left[X_i = x_i, i = 1, \ldots, k\right] = \binom{n}{x_1, \ldots, x_n} p_1^{x_1} \ldots p_k^{x_k}$$

pro $(x_1, \ldots, x_k)$ splňující $\sum_1^k x_i = n$. Tento vztah lze využít pro tzv. přesný test shody pouze pro malé hodnoty n a k a není příliš praktický.

V obecném případě použijeme χ^2 test. Poznamejme jen, že existují i jiné typy testů dobré shody, toto je základní příklad. Budeme porovnávat očekávané četnosti v jednotlivých kategoriích np_i se skutečně pozorovanými četnostmi X_i a spočteme statistiku

$$\chi^2 = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i}$$

Pokud platí hypotéza H_0, tedy testovaná náhodná veličina má očekávané rozdělení, pak má χ^2 tzv. „chí kvadrát“ rozdělení s $k-1$ stupni volnosti (toto rozdělení je tabulizováno). Pro danou hladinu významnosti α určíme kritický obor W a H_0 zamítneme, pokud naše hodnota $\chi^2 \in W$.

Uvádí se, že test lze použít, pokud všechny očekávané četnosti np_i jsou alespoň 5. Případně se někdy v literatuře objevuje předpoklad, že alespoň 80% z očekávaných četností je větších než 5 a všechny musí být větší než 1.

8.4 Test normality (Shapiro–Wilk 1965)

$X_1, \ldots, X_n$ náhodný výběr, hypotéza H_0 je, že pochází z normálního rozdělení $N(\mu, \sigma)$

testovaná statistika (nemá žádné jméno) je

$$W = \frac{\left(\sum_{i=1}^n a_i X_{(i)}\right)^2}{\sum_{i=1}^n (X_i - \overline{X})^2}$$

kde $\overline{X}$ je výběrový průměr, $X_{(i)}$ i-té nejmenší číslo z $X_1, \ldots, X_n$ (i-th order statistic)

$(a_1, \ldots, a_n) = \frac{m^T V^{-1}}{c}$, kde $m = (m_1, \ldots, m_n)^T$ vektor středních hodnot order statistics $N(0,1)$, V je kovarianční matice m a $c = ||V^{-1}m|| = \sqrt{m^t V^{-1} V^{-1} m}$

Testovaná statistika W nabýcá hodnot od 0 do 1, přičemž 1 znamená úplnou shodu. Prahové hodnoty pro danou hladinu významnosti α se pro statistiku W počítají pomocí Monte Carlo simulací, vektory koeficientů lze také určit pomocí simulace. Poznamenejme, že normální rozdělení je spojité a příliš hrubá diskretizace může tento test ovlivnit.

8.5 ANOVA (analýza rozptylu)

Data mohou být rozdělena do několika skupin podle (jedné nebo více) „popisných proměnných“ a zkoumaná číselná náhodná proměnná na nich může záviset. Popisné proměnné mohou být číselné, ale často jsou kategoriální. Jako nulovou hypotézu budeme uvažovat, že rozdělení do skupin nemá vliv na hodnoty náhodné proměnné, tedy „ve všech skupinách má zkoumaná náh. proměnná stejnou střední hodnotu“.

Předpoklady použití ANOVA testu:

1. náhodný výběr měření, pokud možno rovnoměrně přes všechny skupiny (četnost výskytu hodnot popisné proměnné je uniformní),

2. v každé skupině je zkoumaná náhodná proměnná normálně rozdělená

3. a to se steným rozptylem.

Body 2 a 3 nemusí platit striktně, ANOVA je vůčí nim částečně robustní. Před spuštěním samotného testu je možné provést test předpokladů (v různém software bývá implementován, ale často se neprovádí automaticky).

Značení a výpočty

k počet skupin

n_i velikost jednotlivých skupin, tj. počet vzorků v i-té skupině, $i = 1, \dots, k$

n celkový počet vzorků, $n = \sum_{i=1}^{k} n_i$

x_{ij} j-tý vzorek z i-té skupiny

$\overline{x}$ výběrová střední hodnota (celková), $\overline{x} = \sum_{i,j} x_{ij}$

$\overline{x_i}$ výběrová střední hodnota ve skupině i, $\overline{x_i} = \frac{1}{n} \sum_{j=1}^{n_i} x_{ij}$

s_i^2 výběrový rozptyl ve skupině i, $\overline{s_i} = \frac{1}{n-1} \sum_{j=1}^{n_i} (x_{ij} - \overline{x_i})^2$

SST sum of squres total, $SST = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \overline{x})^2$

SSG sum of sq. between groups, $SSG = \sum_{i=1}^{k} n_i (\overline{x_i} - \overline{x})^2$

SSE sum of sq. inside groups, $SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \overline{x_i})^2 = \sum_{i=1}^{k} (n-1) s_i^2$

stupně volnosti (df=degree of freedom) $df(SSG) = k - 1$, $df(SSE) = n - k$, $df(SST) = n - 1$

MS mean of squares, $MS = SS/df$, tedy

$$\begin{aligned} MSG &= SSG/(k-1) \\ MSE &= SSE/(n-k) = \frac{(n_1-1)s_1^2 + \ldots + (n_k-1)s_k^2}{(n_1-1) + \ldots + (n_k-1)} \end{aligned}$$

F-statistika $F = MSG/MSE$, má tabulizovaný průběh pro dané stupně volnosti $k-1$ a $n-k$

P-value jednostranný odhad pravděpodobnosti pro F-statistiku, tj. pravděpodobnost, že F je větší rovna dané hodnotě, pokud je P-value menší než α, zamítáme nulovou hypotézu

Tabulka s výsledky ANOVA testu vypadá následovně.

faktor	SS	df	MS	F	P-value
Group/model	SSG	k-1	MSG	MSG/MSE	p
Error/residual	SSE	n-k	MSE		
Total	SST	n-1			

Při zamítnutí nulové hypotézy o rovnosti středních hodnot v jednotlivých skupinách lze následně provést testy porovnávající dvojice skupin mezi sebou. Nejběžnější je Tukeyho test, který umožňuje porovnat všechny dvojice skupin naráz.

Vícefaktorová ANOVA

Pokud máme data rozdělena do skupin pomocí dvou (případně i více) popisných proměnných lze posuzovat míru vlivu těchto faktorů. To lze pro každý faktor zvlášť, ale může se stát, že vliv jednotlivých faktorů se jednoduše nesčítá, ale kombinuje složitějším způsobem. Potom provádíme ANOVA test se skupinami určeným kartézským součinem oborů hodnout obou faktorů. Zde je třeba dát pozor na podobnou velikost skupin, pokud se faktory ovlivňují navzájem, nemusí být předpoklad splněn.

V tabulce (příklad níže) uvažujeme, že faktor A má k skupin a faktor B má ℓ skupin, SSE se počítá společně podle rozdělení do $k\ell$ skupin. Obvykle faktory řadíme podle jejich vlivu na střední hodnoty ve skupinách, tento vliv můžeme odhadnout nebo použít pro každý z faktorů jednofaktorový ANOVA test.

faktor	SS	df	MS	F	P-value
A	SSA	k-1	MSA	MSA/MSE	p_A
B	SSB	l-1	MSB	MSB/MSE	p_B
A×B	SSAB	(k-1)(l-1)	MSAB	MSAB/MSE	p_{AB}
Error	SSE	n-k	MSE		
Total	SST	n-1			

Pro ilustraci uvažme analýzu platu na pohlaví a dosaženám vzdělání. Faktor pohlaví má dvě možnosti, pro vzdělání máme tři hodnoty (bez maturity, s maturitou, vysokoškolské). Konkrétní data uvádět nebudeme, ale základní charakteristiky jsou: v každé ze šesti skupin je 5 hodnot, střední hodnota platu je 19.7, mezi ženami 18.9 a mezi muži 20.5, při rozdělení podle vzdělání jsou střední hodnoty 14.2 (bez maturity), 17.1 (s maturitou) a 27.8 (VŠ). Nyní se podíváme na tabulky ANOVA testu.

Provedeme-li test pro faktor pohlaví, výsledky mohou vypadat takto.

faktor	SS	df	MS	F	P-value
pohl.	17.6	1	17.6	0.437	0.515
Error	1130.7	28	40.4		

Pro faktor vzdělání dostaneme následující výsledky.

faktor	SS	df	MS	F	P-value
vzd.	1026	2	513	113	< 0.001
Error	122	27	4.52		

Pokud provedeme dvoufaktorový test, bude výsledená tabulka takováto.

faktor	SS	df	MS	F	P-value
vzd.	1026.20	2	513.10	121.68	< 0.001
pohl.	17.63	1	17.63	4.78	0.052
oba	3.27	2	1.63	0.39	0.683
Error	101.20	24	4.22		

Nejprve volíme hladinu významnosti a to tradičně $\alpha = 0.05$. Při interpretaci výsledku postupujeme odspodu a hledáme nejjednodušší, ale data vysvětlující model. Protože P-value na řádku s kombinací obou faktorů je větší než α, lze přijmout nulovou hypotézu, respektive zamítnout alternativní hypotézu, že „kombinace obou faktorů je nutná k popisu modelu“. Na dalším řádku pro faktor pohlaví je P-value opět větší než α, i když v tomto případě jen o málo. Tím pádem můžeme též zamítnout vliv pohlaví na střední hodnotu platu a zbývá nám pouze faktor vzdělání s P-value menší než α, kde zamítneme nulovou hypotézu (tedy vzdělání má podstatný vliv na výši platu). Kdyby na řádku s faktorem pohlaví byla též hodnota menší než α použili bysme model s aditivním vlivem obou faktorů (jak vzdělání tak pohlaví by mělo určitý vliv na plat a tyt vlivy by se ve výsledku sčítali).

8.6 Levenův test homogenity rozptylů

Při některých testech je součástí předpokladů rovnost rozptylů pro dvě nebo více skupin, do kterých náhodný výběr rozdělíme (např. ANOVA). Předpokládáme, že v každé skupině jsou data normálně rozdělená s nějakou střední hodnotou a hypotéza H_0 je, že všechny skupiny mají stejný rozptyl.

Jde v podstatě o jednoduchý ANOVA test pro náhodnou veličinu Z_{ij} definovanou následovně: máme náhodný výběr X_{ij}, označující j-tý prvek ze skupiny i z toho $Z_{ij} = |X_{ij} - \overline{X}_i|$, kde $\overline{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}$ (výběrový průměr i-té skupiny).

(Pozn.: Někdy se místo $\overline{X}_i$ používá $\tilde{X}_i$ medián i-té skupiny. Tato varianta se nazývá Brown–Forsythův test a funguje lépe např. pro χ^2 rozdělení.)

Test tedy vypadá následovně:

$$\begin{aligned} k &= \text{počet skupin} \\ n_i &= \text{velikost } i\text{–té skupiny} \\ N &= \sum_{i=1}^{k} n_i \\ \overline{Z}_i &= \frac{1}{n_i}\sum_{j=1}^{n_i} Z_{ij} \\ \overline{Z} &= \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} Z_{ij} \\ W &= \frac{\sum_{i=1}^{k} n_i(\overline{Z}_i - \overline{Z})^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Z_{ij} - \overline{Z}_i)^2} \cdot \frac{N-k}{k-1} \end{aligned}$$

Statistika W má (přibližně) F-distirbudi s $k-1$ a $N-k$ stupni volnosti.

9 Příklady

Příklad 1:
Ve skupině deseti studentu byla změřena jejich výška (v centimetrech) a hmotnost (v kilogramech). Spočtěte korelaci mezi těmito dvěma náhodnými veličinami, pokud výsledky měření byly následující:

výška (cm)	151	154	165	146	158	172	142	169	176	138
váha (kg)	43	48	56	42	55	54	44	49	55	38

Řešení: Pro výpočet korelace $\rho(X,Y) = \frac{cov(X,Y)}{\sqrt{var(X)var(Y)}}$ použijeme odhady střední hodnoty $\overline{X} = \frac{1}{n}\sum_{i=1} nX_i$, rozptylu $\overline{S}^2_X = \frac{1}{n-1}\sum_{i=1} n(X_i - \overline{X})^2$ (analogicky pro druhou náh. proměnnou Y) a kovariance $\frac{1}{n-1}\sum_{i=1} n(X_i - \overline{X})(Y_i - \overline{Y})$. Jak odhady rozptylů tak kovariance mají shodný multiplikativní faktor $\frac{1}{n-1}$, který lze po dosazeni vykrátit a můžeme přímo počítat $\rho(X,Y) = \frac{\sum(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum(X_i-\overline{X})^2\sum(Y_i-\overline{Y})^2}}$.

Potřebné výpočty lze provést automaticky pomocí vhodného statistického software, my si je zde shrneme do tabulky.

	X_i	Y_i	$X_i - \overline{X}$	$Y_i - \overline{Y}$	$var(X)$	$var(Y)$	$cov(X,Y)$
	151	43	−6.1	−5.4	37.21	29.16	32.94
	154	48	−3.1	−0.4	9.61	0.16	1.24
	165	56	7.9	7.6	62.41	57.76	60.04
	146	42	−11.1	−6.4	123.21	40.96	71.04
	158	55	0.9	6.6	0.81	43.56	5.94
	172	54	14.9	5.6	222.01	31.36	83.44
	142	44	−15.1	−4.4	228.01	19.36	66.44
	169	49	11.9	0.6	141.61	0.36	7.14
	176	55	18.9	6.6	357.21	43.56	124.74
	138	38	−19.1	−10.4	364.81	108.16	198.64
součet	1571	484			1546.9	374.4	651.6
výběr. odhad	157.1	48.4			171.9	41.6	72.4

Z hodnot pak spočteme korelaci $\rho = \frac{651.6}{\sqrt{1546.9 \cdot 374.4}} = 0.856$. To znamená poměrně výraznou pozitivní korelaci mezi náhodnými veličinami ($\rho > 0$ pozitivní korelace znamenajjící, že „vyšší hodnota X vede k vyšší hodnotě Y“, $\rho < 0$ negativní korelace, „vyšší hodnota X vede k nižší hodnotě Y“).

Příklad 2:
Letecká společnost odhadovala množství pasažérů na určitém letu. Z dat získaných v jednom měsíci za 20 letů spočetla, že průměrný počet pasažérů byl 112 a výběrový rozptyl 25. Určete 95% oboustraný interval spolehlivosti pro střední hodnotu počtu pasažérů.

Řešení: Protože máme k dispozici výběrové odhady střední hodnoty a rozptylu, použijeme distribuci Studentova rozdělení s $n-1$ stupni volnosti. Určíme $\delta = t_{\alpha/2} \cdot \frac{\overline{S}}{\sqrt{n}}$ a interval spolehlivosti $C = [\overline{X} - \delta, \overline{X} + \delta]$. Konkrétně zde použijeme hodnoty $n = 20$, $\alpha = 0.05$, $\overline{X} = 112$ a $\overline{S} = \sqrt{25} = 5$, určíme $t_{\alpha/2} = 2.093$ ze Studentova rozdělení s 19 stupni volnosti, dosadíme a spočteme $\delta = 2.34$. Z toho $C = [109.66, 114.34]$.

Příklad 3:
Při měření el. odporu kabelu byly získány následující hodnoty (v ohmech): 0.139, 0.144, 0.139, 0.140, 0.136, 0.143, 0.141, 0.136. Předpokládejme, že naměřené hodnoty jsou realizací náhodné veličiny s normálním rozdělením s neznámou střední hodnotou a rozptylem. Odhadněte tyto parametry a určete 95% interval spohlivosti pro střední hodnotu.

Řešení: Budeme postupovat obdobně jako v předchozím příkladu, jen nejprve musíme spočítat výběrový průměr a rozptyl, tedy $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = 0,13975$, $\overline{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - X)^2 = 8.5 \cdot 10^{-6}$.

Pro $n = 8$ a $\alpha = 0.05$ určíme $t_{\alpha/2} = 2.365$ ze Studentova rozdělení se 7 stupni volnosti, dosadíme a spočteme $\delta = 0.00244$ a $C = [0.13731, 0, 14219]$.

Příklad 4:
Při řezání metrových prken na pile byly naměřeny tyto hodnoty (v mm): 992, 1010, 1005, 996, 998, 1000, 1003, 999, 1000, 997. Odhadněte směrodatnou odchylku a určete 90% inteval spolehlivosti pro tento odhad.

Řešení: Pro odhad rozptylu normálně rozdělené veličiny (považujme, že naše měření je z normální distribuce) se použije χ^2 rozdělení (s $n-1$ stupni volnosti). Pokud je $\overline{S}^2$ výběrový rozptyl, pak $(1-\alpha)$ interval spolehlivosti určíme následovně. Ať a, b jsou taková, že $\chi^2(a, n-1) = \alpha/2$ a $\chi^2(b, n-1) = 1-\alpha/2$. Pak intervalový odhad rozptylu je $[\frac{(n-1)\overline{S}^2}{a}, \frac{(n-1)\overline{S}^2}{b}]$ a směrodatné odchylky $[\overline{S} \cdot \sqrt{\frac{n-1}{a}}, \overline{S} \cdot \sqrt{\frac{n-1}{b}}]$.

S hodnotami v zadání máme $n = 10$, $\alpha = 0.1$ a určíme $a = 16,919$, $b = 3,325$ (z tabulek pro χ^2 rozdělení s 9 stupni volnosti). Pak 90% interval spolehlivosti pro rozptyl je $[13.476, 68, 571]$ a pro směrodatnou odchylku $[3.671, 8.281]$.

Příklad 5:
Máme šestistěnou hrací kostku a chceme ověřit, zda je férová, tedy že každé z čísel padá se stejnou pravděpodobností. Provedeme 60 hodů a dosáhneme násedujících frekvencí výsledků

výsledek	1	2	3	4	5	6
četnost	12	7	7	8	18	8

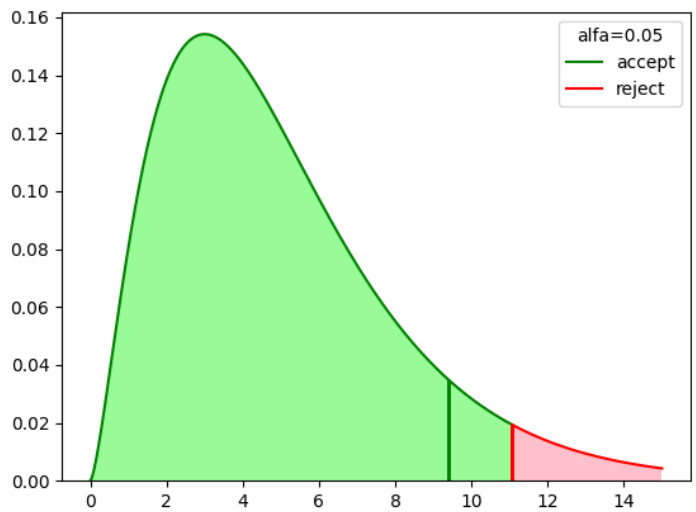
Fakt, že 18-krát padla pětka nám přijde podezřelý. Je možné zamítnou hypotézu férové kostky na hladině významnosti 95%?

Řešení: Při 60 hodech kostkou jsou učekávané četnosti pro všechny hodnoty rovny 10. Ty porovnáme s výsledky pokusu, kde jako X_i označíme skutečnou četnost výsledku i, pomocí testu dobré schody. Pro hladinu významnosti 95% máme $\alpha = 0.05$. Spočteme

$$\chi^2 = \sum_{i=1}^{6} \frac{(X_1 - 10)^2}{10} = 9.4$$

přičemž počet stupňů volnosti je 5. To odpovídá P-value $p = 0.094$. Protože $p > \alpha$, přijímáme nulovou hypotézu a kostku lze považovat za férovou.

Poznamenejme na závěr, že hraniční hodnota P-value pro hladinu významnosti 95% a 5 stupňů volnosti by byla $\chi^2 = 11.07$.



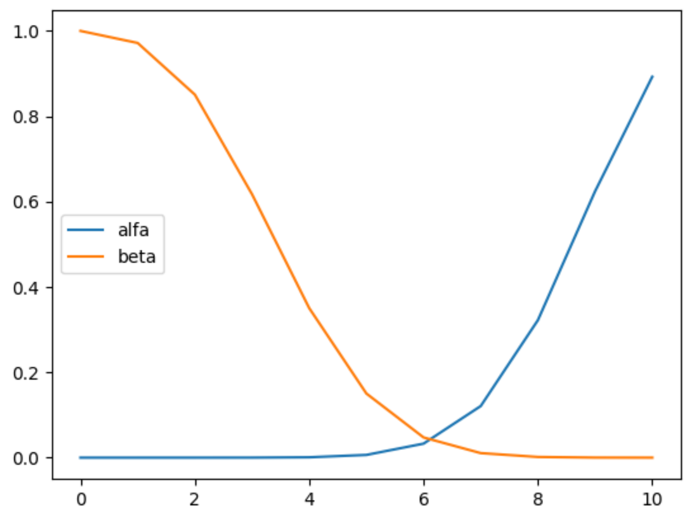
Příklad 6:
Máme dva pytlíky s kuličkami, jeden s 80 bílými a 20 černými kuličkami (pytlík A) a druhý s 30 bílými a 70 černými kuličkami (pytlík B). Máme jeden z nich a provedeme následující test: vylosujeme z něho postupně 10 kuliček (s vracením) a pokud jich je alespň k bílých, přijmeme hypotézu, že je to pytlík A, pokud jich je méně než k, tuto hypotézu odmítneme.

V závislosti na parametru k určete pravděpodobnost chyby prvního a druhého druhu a najděte k pro které je test vyvážený, tj. pravděpodobnost chyby prvního a druhého druhu je (přibližně) stejná.

Řešení: Označme Y náhodnou veličinu rovnou počtu bílých kuliček, které v n tazích ($n = 10$ v našem příkladu) vylosujeme, $Y \in \{0, 1, \ldots, n\}$. Náhodná veličina Y má binomické rozdělení s parametry n, p, kde $p_A = 0.8$ a $p_B = 0.3$.

$$P(Y = y) = \binom{n}{y} p^y (1-p)^{n-y}\ ,\ y = 0, 1, \ldots, n$$

$$\alpha_k = P(Y < k|H_0) = \sum_{i=0}^{k-1} \binom{n}{i} p_A^i (1-p_A)^{n-i}$$
$$\beta_k = P(Y \geq k|H_1) = \sum_{i=k}^{n} \binom{n}{i} p_B^i (1-p_B)^{n-i}$$



Příklad 7:
Máme připravené tři různé verze písemky, označené A, B, C, a chceme otestovat, zda jsou stejně obtížné. Ve skupině je celkem 33 studentů, které náhodně rozdělíme do tří skupin po 11 a každá skupina napíše jeden test. Výsledky testů jsou následující:

A	$1, 2, 3, 3, 4, 4, 4, 5, 5, 6, 7$
B	$2, 3, 4, 4, 5, 5, 5, 6, 6, 7, 8$
C	$3, 4, 5, 5, 6, 6, 6, 7, 7, 8, 9$

Na hladině významnosti 90% proveďte vhodný test na určení rovnosti středních hodnot výsledků všech tří verzí písemky.

Řešení: Nejprve spočteme odhady parametrů rozdělení pro jednotlivé testy. Výběrové průměry jsou

$$\overline{X}_A = 4 \qquad \overline{X}_B = 5 \qquad \overline{X}_C = 6$$

a výběrové rozptyly

$$\overline{S}_A^2 = \overline{S}_B^2 = \overline{S}_C^2 = 3$$

Můžeme tedy předpokládat rovnost rozptylů.

První možností je pro každou dvojici skupin provést dvouvýběrový t-test se shodnými rozptyly. Pro skupiny A–B bude vypadat výpočet následovně

$$t_{AB} = \frac{\overline{X}_A - \overline{X}_B}{\overline{S}\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} = \frac{4-5}{\sqrt{3} \cdot \sqrt{\frac{1}{11} + \frac{1}{11}}} \doteq 1.35$$

To při $df_{AB} = (n_A - 1) + (n_B - 1) = 20$ stupních volnosti odpovídá $p_{AB} = 0.191 > \alpha = 0.1$. Hypotézu o rovnosti středních hodnot (na hladině významnosti 90%) přijímáme.

Při porovnání skupin B–C dostaneme stejné výsledky $t_{BC} \doteq 1.35$, $p_{BC} = 0.191 > \alpha$. Při porovnání A–C je dvojnásobný čitatel zlomku a tak vyjde $T_{AC} \doteq 2.71$ s $p = 0.014 < \alpha$ a zde nulovou hypotéza zamítáme.

Druhou možností je porovnat vše najednou pomocí ANOVA testu. Potřebuje ještě průměr celého výběru $\overline{X} = 5$ a spočtem dále

$$\begin{aligned}
SSG &= \sum_{i=A,B,C} n_i \cdot (\overline{X}_i - \overline{X})^2 = 11 \cdot (-1)^2 + 11 \cdot 0^2 + 11 \cdot 1^2 = 22 \\
df &= k - 1 = 2 \\
MSG &= SSG/df = 11 \\
SSE &= \sum_{i=A,B,C} \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_i)^2 = \sum_{i=A,B,C} (n_i - 1)\overline{S}_i^2 = \\
&= 10 \cdot 3 + 10 \cdot 3 + 10 \cdot 3 = 90 \\
df &= n - k = 30 \\
MSE &= SSE/df = 3 \\
F &= \frac{MSG}{MSE} = \frac{11}{3} \doteq 3.67 \\
p &\doteq 0.038
\end{aligned}$$

Protože je hodnota $p < \alpha$, zamítáme hypotézu o rovnosti středních hodnot všech tří skupin. Pro porovnání dvojic skupin lze místo výše uvedeného t-testu použít některý post hoc test, které se ve statistických programech nabízejí (např. Tukey).