Statistika ve společenskovědním výzkumu 2

přednášející:	Martin Schindler
	KAP, tel. 48 535 2836, budova G
konzul. hodiny:	po dohodě
e-mail:	martin.schindler@tul.cz

naposledy upraveno: 20. února 2017

Požadavek na udělení zápočtu:

vypracování a obhájení semestrální práce.

Literatura

- Chráska, M.: Metody pedagogického výzkumu Praha:Grada, 2007.
- HENDL, J.: Přehled statistických metod zpracování dat. Portál: Praha, 2012 (4.vyd.)
- ZVÁRA K., ŠŤĚPÁN J.: Pravděpodobnost a matematická statistika. Praha: Matfyzpress, 2002.

Literatura online

- http://moodle.vsb.cz/vyuka/course/info.php?id=3
- http://www.studopory.vsb.cz/
- http://mathonline.fme.vutbr.cz/
- http://home.zcu.cz/ friesl/hpsb/tit.html
- tato prezentace: http://147.230.193.199/ ms/

Matematická statistika

Předpokládáme, že napozorovaná data $X_1, X_2, \ldots, X_n$ jsou náhodným vzorkem z populace a řídí se nějakým modelem s neznámými parametry

- Snaha: odhadnout tyto neznámé parametry
- Nejčastěji předpokládáme model tzv. normálního (Gaussova) nebo binomického rozdělení pravděpodobnosti.

Teorie pravděpodobnosti

- se zabývá tzv. **náhodnými pokusy**, tj. pokusy, u nichž výsledek není předem jednoznačně určen

- množinu všech možných výsledků náhodného pokusu označujeme Ω
- prvky Ω označujeme ω_i a nazýváme **elementární jevy**
- **náhodný jev** (ozn. A, B, atpd.) - tvrzení o výsledku náhodného pokusu, je to podmnožina Ω tvořena některými elem. jevy

Pravděpodobnost náhodného jevu A (ozn. $P(A)$): vyjadřuje míru očekávání, že nastane jev A.

- při velkém počtu opakování tohoto náhodného pokusu se relativní četnost jevu A blíží k $P(A)$.

Klasická pravděpodobnost

- množina všech výsledků náhodného pokusu Ω je složena z konečného počtu (n) elementárních jevů $\omega_1, \ldots, \omega_n$
- každý z těchto elementárních jevů je stejně pravděpodobný
- označme $m(A)$ počet elementárních jevů, které tvoří jev (jsou příznivé jevu) A

Potom

$$P(A) = \frac{m(A)}{n} = \frac{\text{počet příznivých elem. jevů}}{\text{počet všech elem. jevů}}$$

Příklad: hod kostkou

- jednou hodíme symetrickou šestistěnou kostkou s čísly $1, 2, \ldots, 6$
- jev A - padne šestka
- jev B - padne liché číslo
- každá z 6 možností, které mohou nastat, jsou stejně pravděpodobné
- určíme $m(A) = 1$ a $m(B) = 3$

Proto

$$P(A) = \frac{m(A)}{n} = \frac{1}{6}$$

a

$$P(B) = \frac{m(B)}{n} = \frac{3}{6} = \frac{1}{2}$$

Příklad (permutace)

Jaká je pravděpodobnost, že při náhodném seřazení písmen P, A, V, E, L vznikne slovo PAVEL?

- **faktoriál**: $n! = 1 \cdot 2 \cdot \ldots \cdot n$ je počet způsobů, jak uspořádat do řady n různých prvků - počet **permutací**
- počet všech možností seřazení je tedy $5! = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 = 120$, každá stejně pravděpodobná
- z nich příznivá je pouze jedna
- proto $P = \frac{1}{5!} = \frac{1}{120}$

Jak by to dopadlo s písmeny slova ANANAS?

- zde jde o **permutace s opakováním** (některé prvky se opakují), počet možností přeuspořádání je $\frac{6!}{2!\cdot 3!}$, z nich příznivá je pouze jedna
- proto $P = \frac{1}{\frac{6!}{2!\cdot 3!}} = \frac{2!\cdot 3!}{6!} = \frac{2\cdot 6}{720} = \frac{1}{60}$

Náhodná veličina

- použití jen náhodných jevů nestačí
- často je výsledkem náhodného pokusu číslo
- např. nás zajímá počet šestek při hodu deseti kostkami, nebo jak dlouho vydrží svítit žárovka

Náhodná veličina: číselné vyjádření výsledku náhodného pokusu

Rozdělení náhodné veličiny: udává jakých hodnot s jakou pravděpodobností veličina nabývá

- rozdělení lze jednoznačně určit např. pomocí distribuční funkce
- **Distribuční funkce** $F_X(x)$ náh. veličiny X určuje pro každé x pravděpodobnost, že je náh. veličina menší než číslo x:

$$F_X(x) = P(X < x) \qquad x \in R$$

kumulat. pravděpodobnost (představa: teoretický protějšek kumulativní relativní četnosti počítané v každém bodě R)

Typy rozdělení

Diskrétní rozdělení ($F_X(x)$ "schodovitá"): X je náhodná veličina s diskrétním rozdělením pravděpodobnosti, jestliže existuje seznam hodnot $x_1, x_2, \ldots$ a kladných pravděpodobností $P(X = x_1), P(X = x_2), \ldots$ splňujících $\sum_i P(X = x_i) = 1$.

Spojité rozdělení ($F_X(x)$ spojitá): existuje tzv. **hustota** $f_X(x)$, která udává "pravděpodobnost výsledku"
představa: teoretický protějšek hranice histogramu pro délku intervalů jdoucích k nule

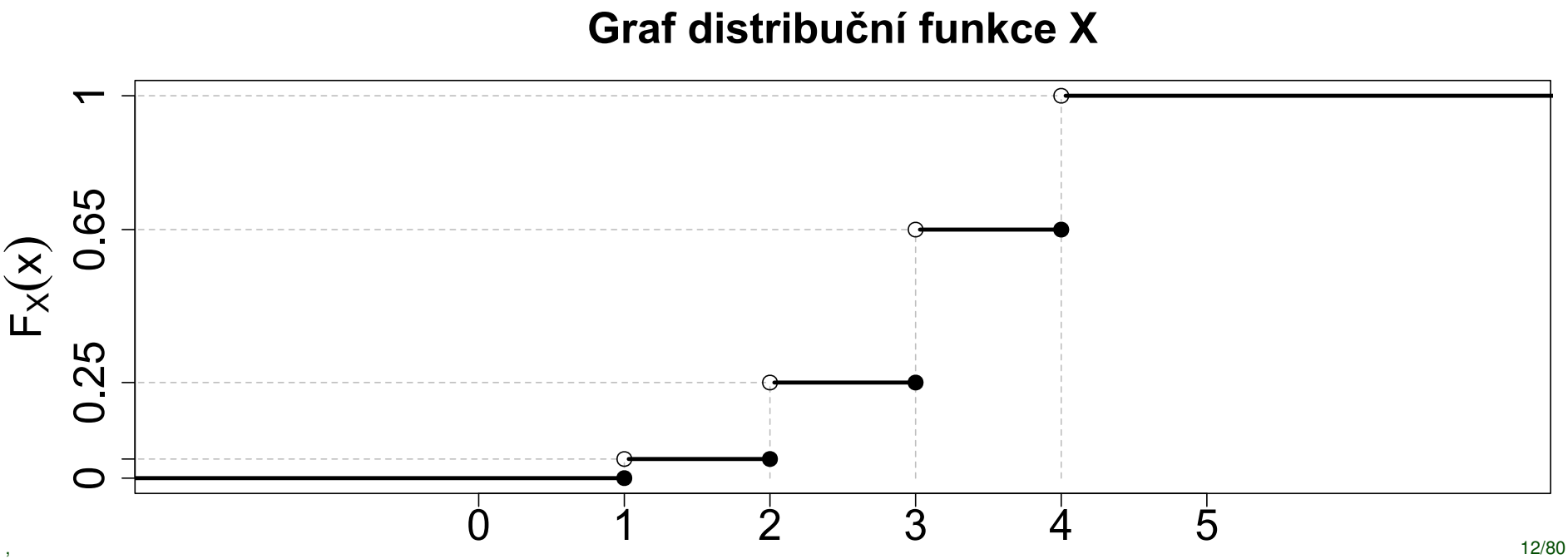
Příklad 1

(diskrétní rozdělení): Ze zkušenosti je známo, že rozdělení výsledku z předmětu MV2 u náhodně vybraného studenta (X) je následující:

x_i	1	2	3	4
$P(X = x_i)$	0,05	0,2	0,4	0,35

Určete $P(X < 3)$ a distribuční funkci náhodné veličiny X.

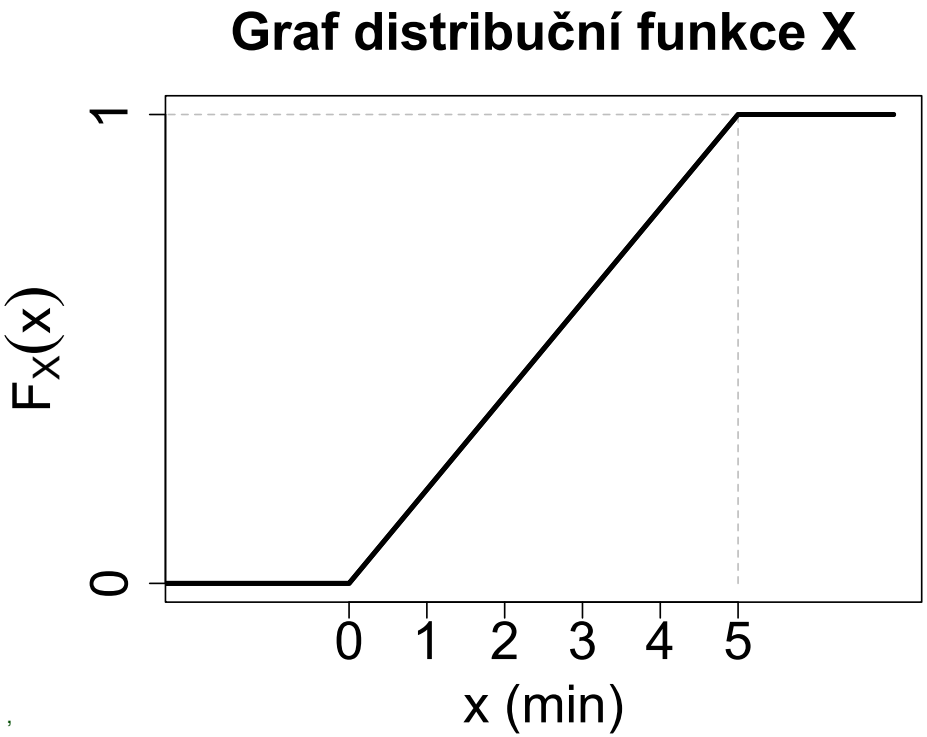
- $F_X(3) = P(X < 3) = P(X = 1) + P(X = 2) = 0{,}05 + 0{,}2 = 0{,}25$
- nutno určit $F_X(x) = P(X < x)$ pro každé $x \in R$

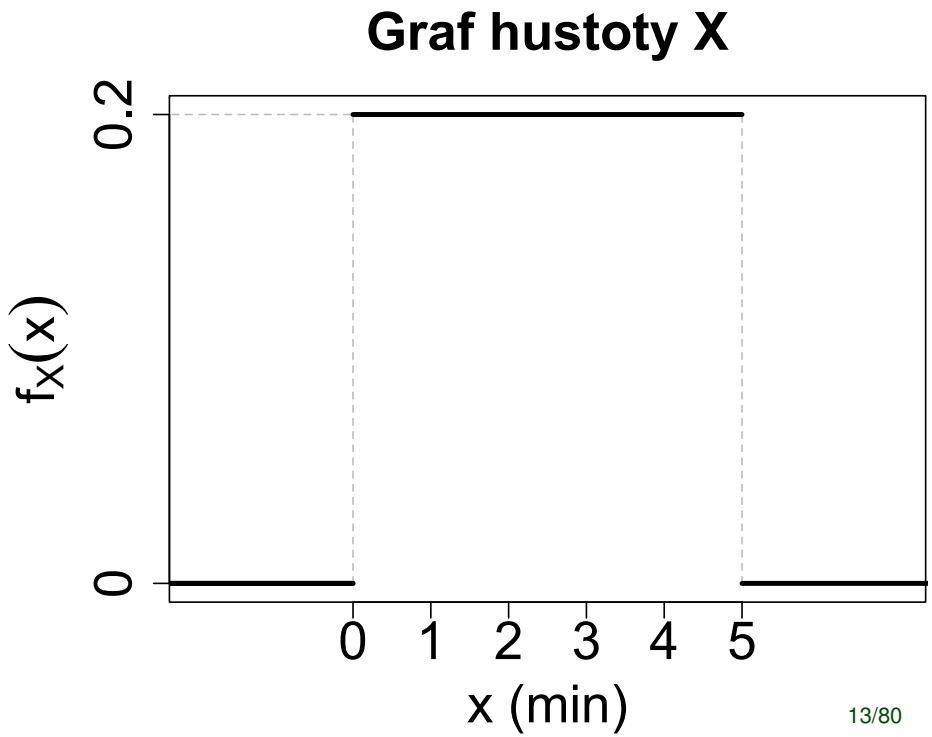


Příklad 2

(spojité rozdělení): Tramvaj jezdí v pravidelných pětiminutových intervalech. Předpokládejme, že čas našeho příchodu na zastávku je náhodný. Jaké je rozdělení náhodné veličiny X značící dobu čekání na tramvaj? ▸ k rovnoměrnému rozdělení

- stačí určit distribuční funkci $F_X(x)$ nebo hustotu rozdělení $f_X(x)$ pro každé $x \in R$
- zřejmě pro $x \in (0, 5)$ platí $F_X(x) = P(X < x) = \frac{x}{5}$, a $f_X(x) = \frac{1}{5}$





Střední hodnota

Střední hodnota (očekávaná hodnota) náhodné veličiny X - hodnota, kolem které se kumulují hodnoty náhodné veličiny X

- pro diskrétní rozdělení: vážený průměr možných hodnot, váhami jsou pravděpodobnosti hodnot

$$EX = \sum_i x_i \cdot P(X = x_i) = x_1 \cdot P(X = x_1) + x_2 \cdot P(X = x_2) + \dots$$

u ▸ Př. 1: $EX = 1 \cdot 0{,}05 + 2 \cdot 0{,}2 + 3 \cdot 0{,}4 + 4 \cdot 0{,}35 = 3{,}05$ (střední, očekávaná známka)

- pro spojité rozdělení: integrál všech možných hodnot x, váhovou funkcí je hustota

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

u ▸ Př. 2: $EX = \int_{-\infty}^{0} x \cdot 0\, dx + \int_0^5 x \cdot \frac{1}{5}\, dx + \int_5^{\infty} x \cdot 0\, dx = \frac{5}{2}$ (střední, očekávaná doba čekání)

Rozptyl

Rozptyl náh. vel. X: $var\,X = E(X - EX)^2$ - udává variabilitu rozdělení náhodné veličiny X kolem její střední hodnoty, je to střední hodnota čtverců odchylek možných hodnot od střední hodnoty

- pro diskrétní rozdělení:

$$var\,X = E(X - EX)^2 = \sum_i (x_i - EX)^2 \cdot P(X = x_i) =$$

$$= (x_1 - EX)^2 \cdot P(X = x_1) + (x_2 - EX)^2 \cdot P(X = x_2) + \dots$$

u ▸ Př. 1:
$var\,X = 2{,}05^2 \cdot 0{,}05 + 1{,}05^2 \cdot 0{,}2 + 0{,}05^2 \cdot 0{,}4 + 0{,}95^2 \cdot 0{,}35 = 0{,}7475$

- pro spojité rozdělení:

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

u ▸ Př. 2:
$var\,X = \int_{-\infty}^{0} (x - \frac{5}{2})^2 \cdot 0\,dx + \int_0^5 (x - \frac{5}{2})^2 \cdot \frac{1}{5}\,dx + \int_5^{\infty} (x - \frac{5}{2})^2 \cdot 0\,dx \doteq 2{,}083$

$\sqrt{var\,X}$ se nazývá **směrodatná odchylka** náh. vel. X

Příklad

(binomické rozdělení): V testu je 5 otázek, na každou je správná právě jedna z odpovědí a), b), c), d). Jaká je pravděpodobnost, že odpovíme právě na 3 otázky správně, pokud tipujeme náhodně?

- ozn. počet správných odp. jako X
- na každou odpovíme správně s pravděpodobností $p = 1/4$
- odpovědi na jednotlivé otázky jsou nezávislé
- tj. pravděp., že ve třech daných (např. prvních třech) otázkách odpovíme správně a v ostatních nesprávně (ozn. 11100), je $p^3 \cdot (1-p)^2$
- mohli jsme se ale trefit i v jiných třech otázkách: počet způsobů, jak vybrat tři otázky z pěti, na které můžeme odpovědet správně je $\binom{5}{3} = 10$

$10\times$ {11100, 11010, 10110, 01110, 11001, 10101, 01101, 10011, 01011, 00111}

Tedy pravděp., že odpovíme právě na 3 otázky správně

$$P(X=3) = \binom{5}{3} \cdot p^3 \cdot (1-p)^2 = 10 \cdot (1/4)^3 \cdot (3/4)^2 = 0{,}088$$

Binomické rozdělení

Opakujeme nezávisle stejný náhodný pokus n-krát. Zajímá nás X četnost nějakého náhodného jevu v těchto n pokusech, jestliže je pravděpodobnost tohoto jevu ve všech pokusech stejná, rovna p. X může nabývat pouze hodnot $0, 1, \ldots, n$ a má rozdělení dané pravděpodobnostmi

$$P(X = i) = \binom{n}{i} \cdot p^i \cdot (1-p)^{n-i}, \quad i = 0, 1, \ldots, n; \qquad \text{kde } 0 < p < 1$$

- říkáme, že X má **binomické rozdělení** s parametry n a p
- zkráceně píšeme $X \sim Bi(n, p)$
- střední hodnota $EX = \sum_{i=0}^{n} i \cdot \binom{n}{i} \cdot p^i \cdot (1-p)^{n-i} = n \cdot p$
- rozptyl

$$var\, X = n \cdot p \cdot (1-p)$$

u Př.: $X \sim Bi(5, 1/4)$ $EX = \frac{5}{4}$ $var\, X = 5 \cdot \frac{1}{4} \cdot \frac{3}{4} = \frac{15}{16}$

Normální (Gaussovo) rozdělení

Nechť X je náhodná veličina se spojitým rozdělením s hustotou

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right), \quad \text{pro } x \in R.$$

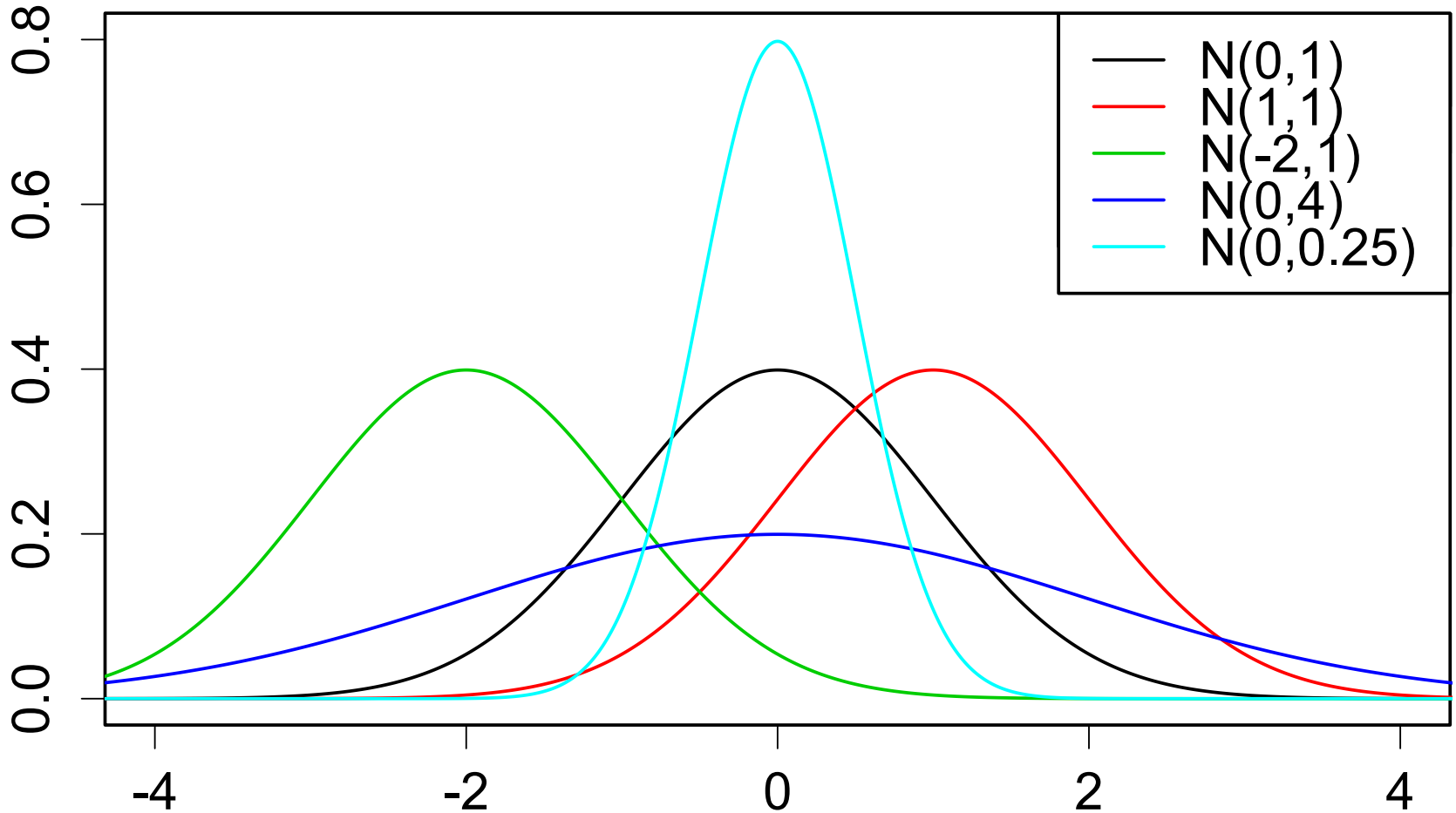
kde $\mu = EX$ a $\sigma^2 = var\ X$ jsou parametry rozdělení.

- říkáme, že X má **normální rozdělení** se stř. hod. μ a rozptylem σ^2
- zkráceně píšeme $X \sim N(\mu, \sigma^2)$
- pro distribuční funkcí $F_X(x)$ neexistuje explicitní vyjádření
- pro $N(0,1)$ jsou hodnoty přesně tabelovány
- nejdůležitější spojité rozdělení

▶ Vznik: součtem mnoha nepatrných příspěvků

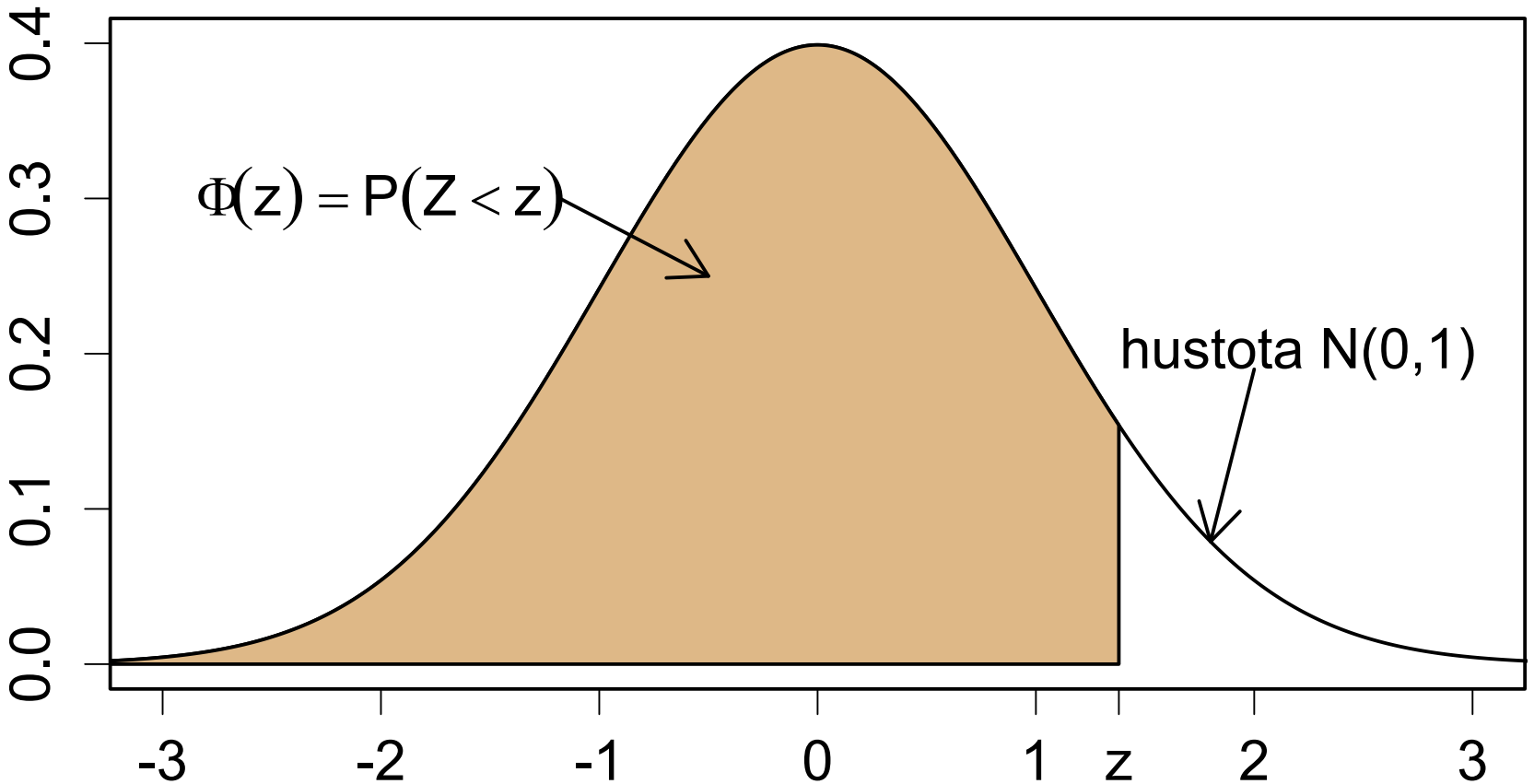
Grafy hustot normálního rozdělení $N(\mu, \sigma^2)$

- symetrické kolem střední hodnoty



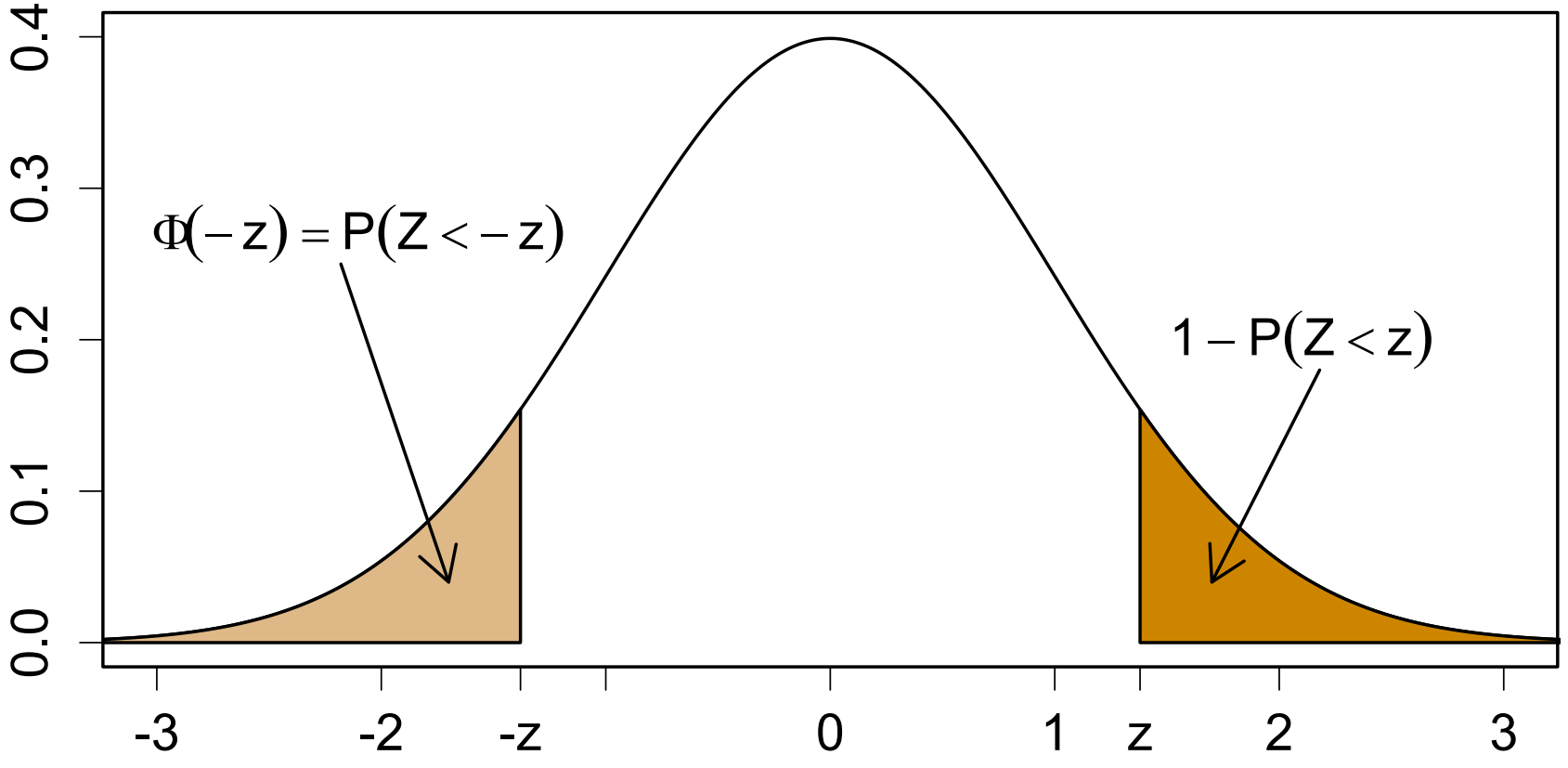
Normované normální rozdělení $Z \sim N(0,1)$

- distrib. funkce $N(0,1)$ značíme $\Phi(z) = P(Z < z)$
- např. $\Phi(1{,}38) = P(Z < 1{,}38) \overset{\text{z tabulek}}{=\!=} 0{,}916$



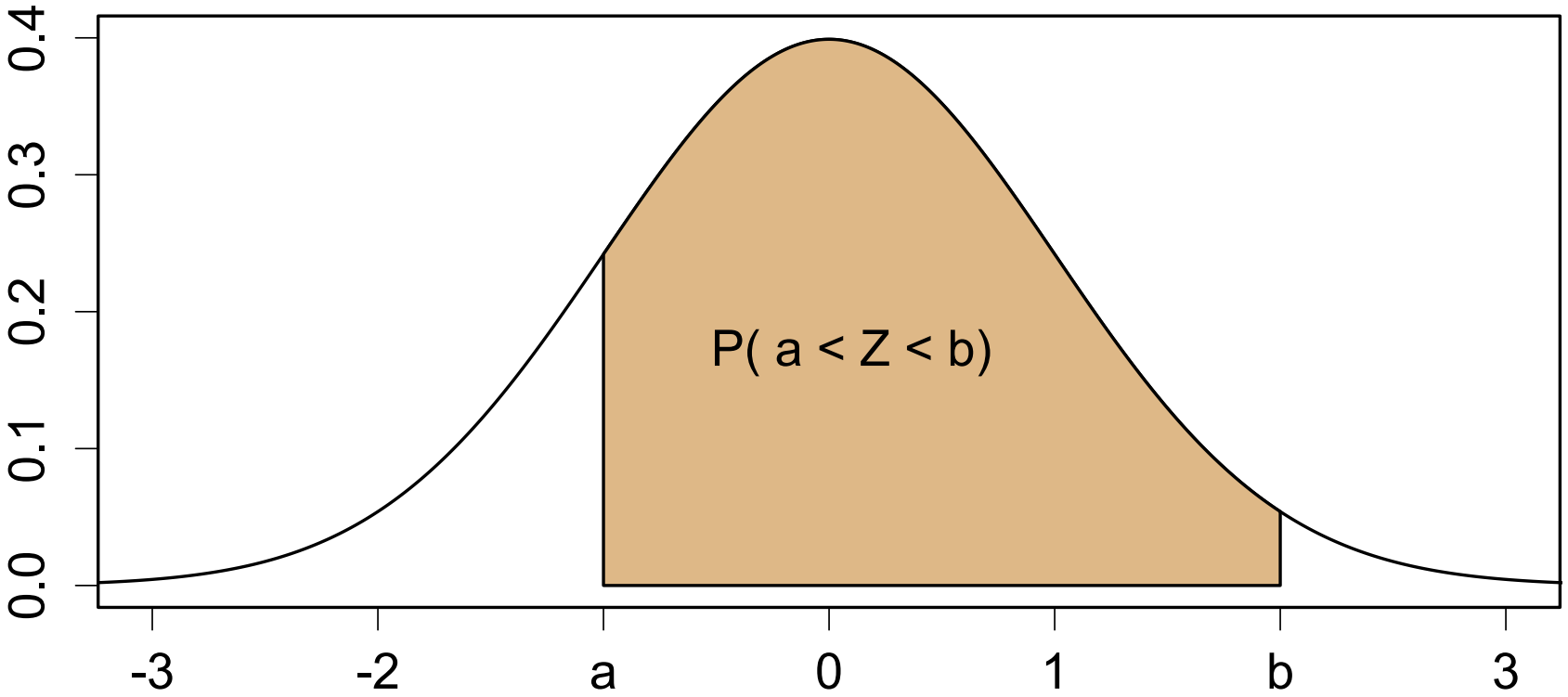
Normované normální rozdělení $Z \sim N(0,1)$

- ze symetrie $N(0,1)$ plyne: $\Phi(-z) = 1 - \Phi(z)$
- např. $P(Z < -1{,}38) = \Phi(-1{,}38) = 1 - \Phi(1{,}38) \overset{\text{z tab.}}{=\!=} 1 - 0{,}916 = 0{,}084$



Normované normální rozdělení $Z \sim N(0,1)$

- $P(a < Z < b) = P(Z < b) - P(Z < a) = \Phi(b) - \Phi(a)$
- např. $P(-1 < Z < 2) = \Phi(2) - \Phi(-1) \overset{\text{z tab.}}{=\!=} 0{,}977 - 0{,}158 = 0{,}819$



Obecné normální rozdělení $Z \sim N(\mu, \sigma^2)$

- pro $X \sim N(\mu, \sigma^2)$ platí, že

$$Z \overset{\text{ozn.}}{=\!=} \frac{X-\mu}{\sigma} \sim N(0,1)$$

- $P(X < x) = P(\frac{X-\mu}{\sigma} < \frac{x-\mu}{\sigma}) = \Phi\left(\frac{x-\mu}{\sigma}\right)$
- proto

$$P(a < X < b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

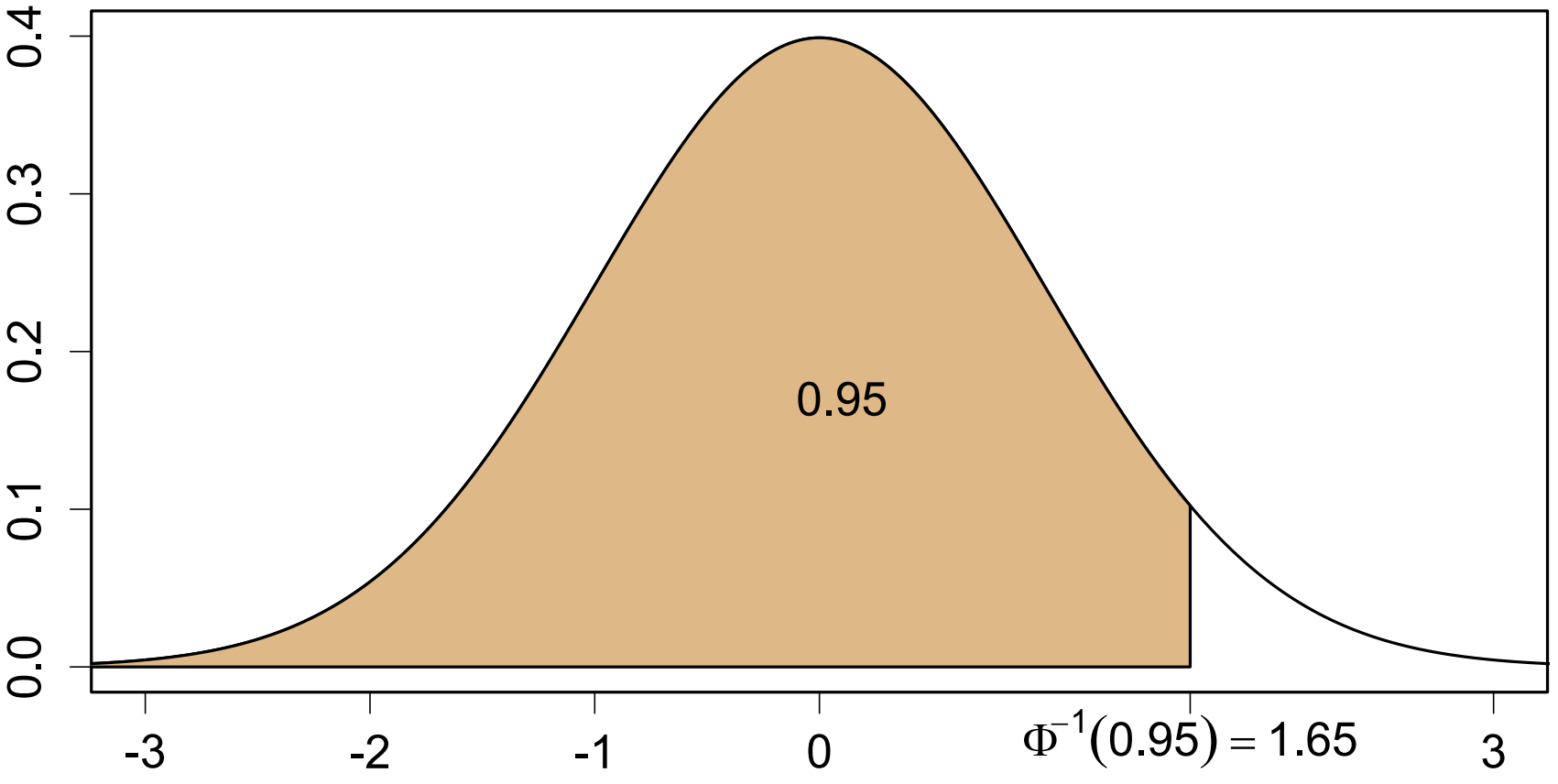
Př.: Výška chlapců v šesté třídě $X \sim N(\mu = 143, \sigma^2 = 49)$:
určeme $P(130 < X < 150) = \Phi\left(\frac{150-143}{7}\right) - \Phi\left(\frac{130-143}{7}\right) \doteq 0{,}81$
tedy mezi chlapci v šesté třídě je přibližně 81% vysokých 130 až 150 cm.

Př.: Jaké výšky dosahuje jen 5% chlapců v šesté třídě?

. . . 95%-ní kvantil rozdělení $N(\mu = 143, \sigma^2 = 49)$

Kvantily normovaného normálního rozdělení 1

- kvantilovou funkci náh. vel. $Z \sim N(0,1)$ značíme $\Phi^{-1}(\alpha)$
- platí $P(Z < \Phi^{-1}(\alpha)) = \Phi(\Phi^{-1}(\alpha)) = \alpha$
- lze najít v tabulkách $\Phi(x)$ inverzním postupem
- často používané: $\Phi^{-1}(0{,}95) = 1{,}65$ a $\Phi^{-1}(0{,}975) = 1{,}96$

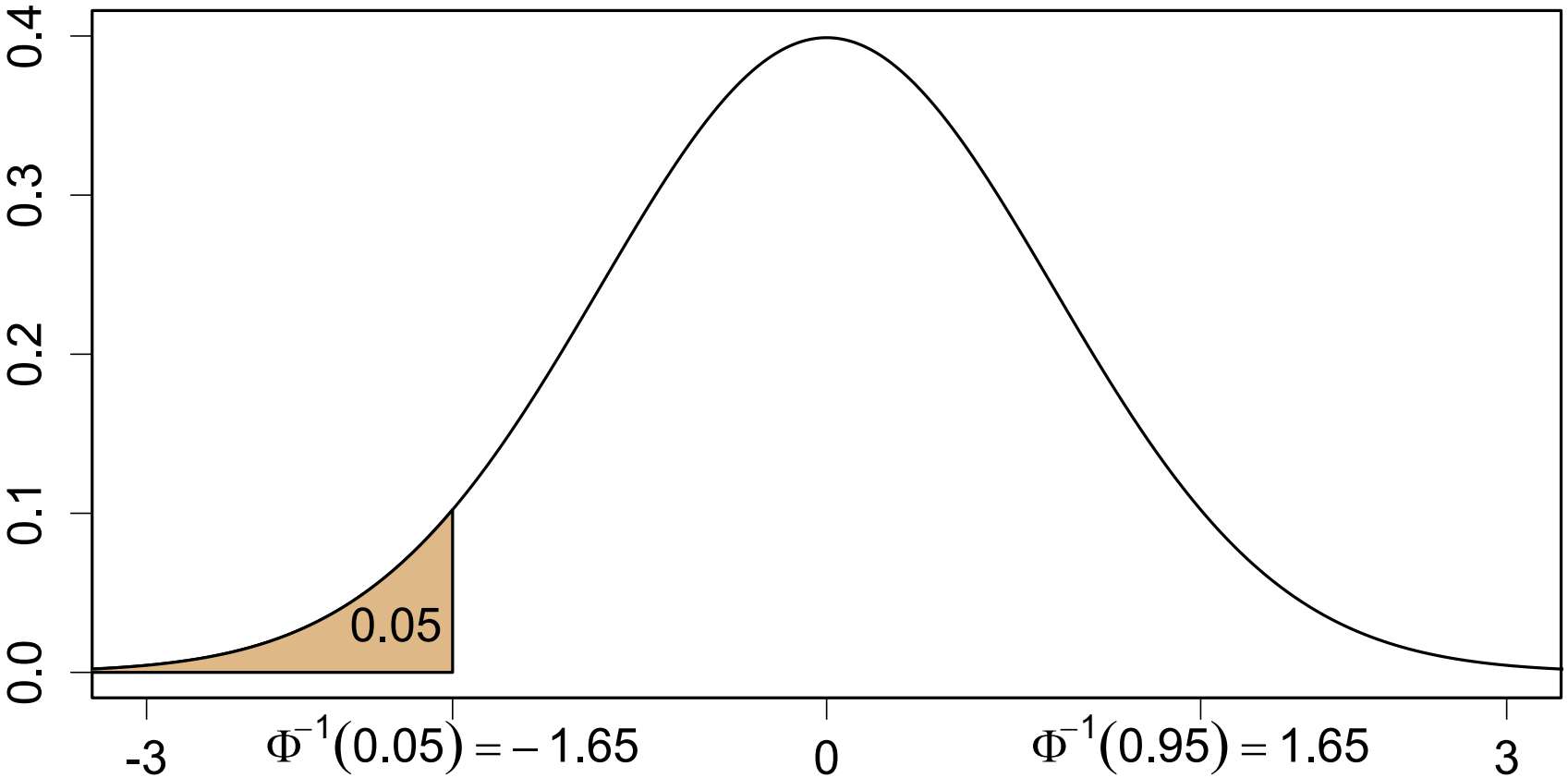


Kvantily normovaného normálního rozdělení 2

- v tabulkách často jen kvantily pro $\alpha \geq 0{,}5$
- pro $\alpha < 0{,}5$ lze využít vztahu (plyne ze symetrie rozdělení):

$$\Phi^{-1}(\alpha) = -\Phi^{-1}(1-\alpha)$$

- např: 5%-ní kvantil $N(0,1)$ je $\Phi^{-1}(0{,}05) = -\Phi^{-1}(0{,}95) = -1{,}65$



Kvantily obecného normálního rozdělení

- pro $X \sim N(\mu, \sigma^2)$ platí, že $Z \overset{\text{ozn.}}{=\!=} \frac{X-\mu}{\sigma} \sim N(0,1)$
- α-kvantil náh. vel X je taková hodnota h, pro kterou platí

$$P(X < h) = \alpha \qquad \Phi\left(\frac{h-\mu}{\sigma}\right) = \alpha$$

$$P\left(\frac{X-\mu}{\sigma} < \frac{h-\mu}{\sigma}\right) = \alpha \qquad \frac{h-\mu}{\sigma} = \Phi^{-1}(\alpha)$$

$$P\left(Z < \frac{h-\mu}{\sigma}\right) = \alpha \qquad h = \sigma \cdot \Phi^{-1}(\alpha) + \mu$$

Př.: Určeme 95%-ní kvantil rozdělení $N(\mu = 143, \sigma^2 = 49)$
je roven $\sigma \cdot \Phi^{-1}(0{,}95) + \mu = 7 \cdot 1{,}65 + 143 = 154{,}5$
tedy jen 5% chlapců v šesté třídě měří více než 154,5 cm.

Náhodný výběr

Náhodný výběr je n-tice $X_1, X_2, \ldots, X_n$ náhodných veličin, které jsou nezávislé a mají stejné rozdělení.

- Př. 1: Výška chlapců šestých tříd, velká populace, náhodně vybereme n chlapců u nichž změříme výšku X_i
- Př. 2: Měření pevnosti tkaniny, změříme pevnost na n náhodně vybraných vzorcích

- počet veličin **n** označujeme pojmem **rozsah výběru**
- parametry rozdělení (stř. hodnotu μ, rozptyl σ^2, atd.) náh. veličin X_i často neznáme
- z náhodného výběru lze tyto neznáme parametry rozdělení odhadnout
- **výběrový průměr** $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ je (bodovým) odhadem střední hodnoty (výšky, pevnosti)
- **výběrový rozptyl** $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ je (bodovým) odhadem rozptylu rozdělení
- $\overline{X}$ a S^2 jsou také náhodné veličiny

Vlastnosti výběrového průměru

Nechť $X_1, X_2, \ldots, X_n$ je náhodný výběr z rozdělení se střední hodnotou μ a rozptylem σ^2. Potom

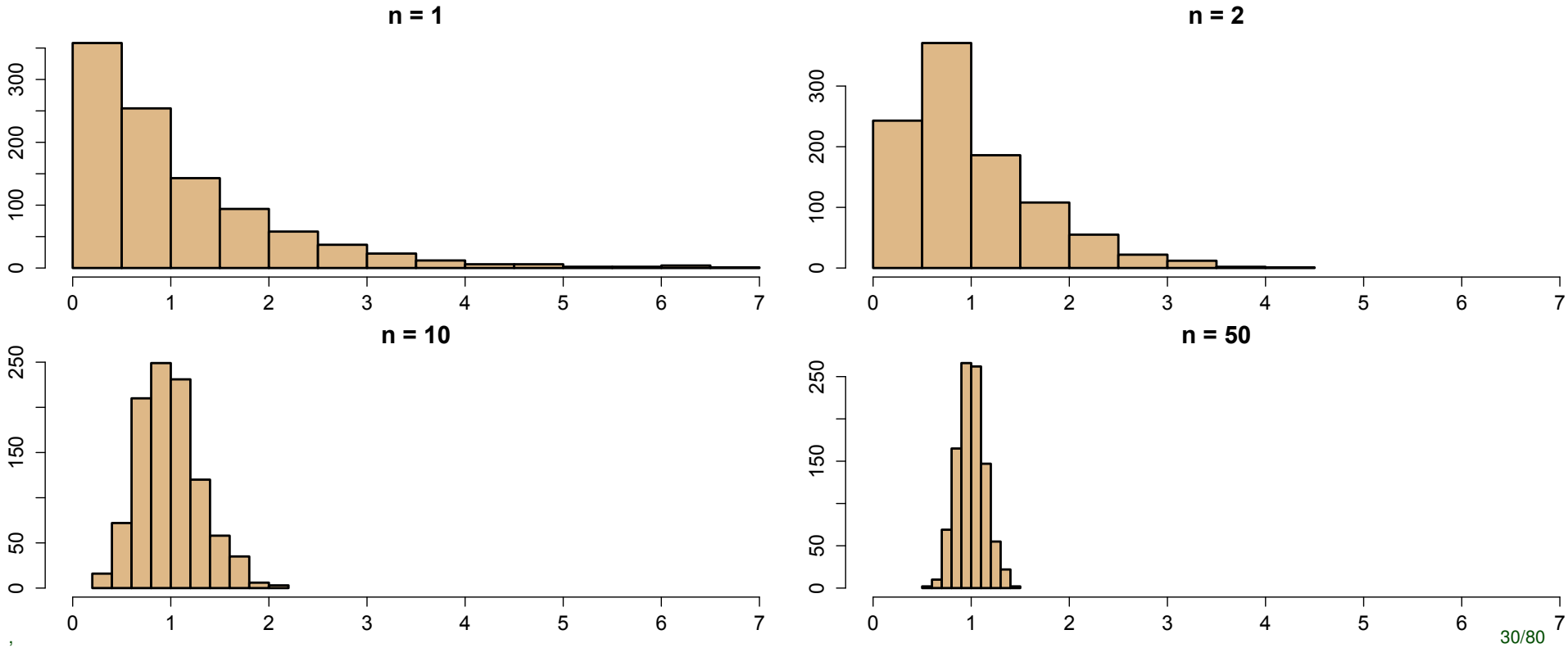
1) $E\overline{X} = \mu$ ($\overline{X}$ je nestranný odhad μ)
2) $var\,(\overline{X}) = \frac{\sigma^2}{n}$

podobně lze dokázat nestrannost výběr. rozptylu, tj. $ES^2 = \sigma^2$

Histogramy průměrů

Př.: Zajímá nás životnost vyráběných zářivek, náhodně vybereme n zářivek, otestujeme je a spočítáme jejich průměrnou životnost. Takových průměrů spočítáme 1 000 a nakreslíme jejich histogram. (Data vygenerována z exponenciálního rozdělení se střední hodnotou 1)

- z obrázku patrno, že s rostoucím n klesá variabilita průměru a zlepšuje se normalita (centrální limitní věta)



Př.: Česká obchodní inspekce chce zkontrolovat výrobce coly, zda nešidí zákazníky. Chce proto odhadnout střední množství coly v dvoulitrové lahvi a zkontrolovat tak, zda je plnící automat správně nastaven. Náhodně bylo za tímto účelem vybráno 100 lahví a byl zjištěn jejich průměrný obsah $\overline{X} = 1{,}982$ litrů. O daném plnícím automatu je navíc známo, že směrodatná odchylka množství plněného do dvoulitrových lahví je $\sigma = 0{,}05$ litrů (tedy rozptyl $\sigma^2 = 0{,}0025$ litrů^2) a množství nápoje v jedné lahvi se dá považovat za normálně rozdělenou náhodnou veličinu $N(\mu, \sigma^2 = 0{,}0025)$. Potvrzují data domněnku, že je automat špatně nastaven a výrobce tak šidí spotřebitele?

- $\overline{X} = 1{,}982$ se dá považovat za bodový odhad středního množství v lahvi μ. Při každém náhodném výběru lahví vyjde jiný odhad (průměr). Co teď?
- Nelze najít např. nějaký interval (...intervalový odhad), o kterém bychom dokázali říct, že pokrývá neznámé střední množství μ s velkou pravděpodobností?
- Jak ověřit domněnku (...testování hypotéz), že výrobce špatným nastavením automatu šidí zákazníky?

Matematická statistika

Předpokládejme, že $X_1, X_2, \ldots, X_n$ je náhodný výběr z nějakého rozdělení většinou s neznámými parametry
Většinou předpokládáme, že náh. výběr pochází z pevně daného rozdělení (nejčastěji normálního) a snažíme se odhadnout neznámé parametry tohoto rozdělení nebo ověřit (testovat) hypotézy o těchto parametrech (u norm. rozd. půjde o střední hodnotou μ a rozptyl σ^2)

- **bodový odhad** neznámého parametru je jedna hodnota, kterou spočítáme z hodnot realizovaného náhodného výběru, např. $\overline{X}$ je bodovým odhadem μ
- **intervalový odhad** neznámého parametru (také **interval spolehlivosti**) je interval (jehož hranice také závisí na náhodném výběru), který pokrývá hodnotu neznámého parametru s předepsanou pravděpodobností
- v **testování hypotéz** se snažíme rozhodnout mezi dvěma odporujícími si tvrzeními (hypotézami) o daném parametru rozdělení, např. zda je automat na plnění lahví správně nastaven ($\mu = 2$ litry) nebo není ($\mu \neq 2$ litry)

Interval spol. pro μ, když σ^2 známe, u $N(\mu, \sigma^2)$

Pro náhodný výběr $X_1, X_2, \ldots, X_n$ z rozdělení $N(\mu, \sigma^2)$ platí

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$ z CLV platí pro n velké i pro nenormální data

proto

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

a tedy platí, že

$$P\left(-\Phi^{-1}(1-\alpha/2) < \frac{\overline{X} - \mu}{\sigma} \cdot \sqrt{n} < \Phi^{-1}(1-\alpha/2)\right) = 1 - \alpha$$

$100(1-\alpha)\%$-ní interval spolehlivosti pro μ a známé σ^2 je tedy

$$\left(\overline{X} - \Phi^{-1}(1-\alpha/2) \cdot \frac{\sigma}{\sqrt{n}};\ \overline{X} + \Phi^{-1}(1-\alpha/2) \cdot \frac{\sigma}{\sqrt{n}}\right)$$

tento interval (je náhodný) pokrývá neznámou střední hodnotu μ s pravděpodobností $1-\alpha$

- jen zhruba $100(1-\alpha)\%$ takových intervalů obsahuje neznámé μ

zpět k Př.: Náhodně vybráno 100 lahví coly a byl zjištěn jejich průměrný obsah $\overline{X} = 1{,}982$ litrů. Naměřené hodnoty považujeme za realizaci náhodného výběru z rozdělení $N(\mu, \sigma^2 = 0{,}0025)$. Spočítejme 95%-ní interval spolehlivosti pro střední množství coly v jedné lahvi μ.

- $100(1-\alpha)$%-ní int. spol. je $\left(\overline{X} - \Phi^{-1}(1-\alpha/2) \cdot \frac{\sigma}{\sqrt{n}};\ \overline{X} + \Phi^{-1}(1-\alpha/2) \cdot \frac{\sigma}{\sqrt{n}}\right)$
- pro 95%-ní int. spol. položíme $\alpha = 0{,}05$ a najdeme tedy $\Phi^{-1}(1-0{,}05/2) = \Phi^{-1}(0{,}975) = 1{,}96$
- dosadíme za $\overline{X} = 1{,}982$, $\sigma = 0{,}05$ a $n = 100$:

$$\left(1{,}982 - 1{,}96 \cdot \frac{0{,}05}{\sqrt{100}};\ 1{,}982 + 1{,}96 \cdot \frac{0{,}05}{\sqrt{100}}\right) \doteq$$
$$\doteq (1{,}982 - 0{,}010;\ 1{,}982 + 0{,}010) =$$
$$= (1{,}972;\ 1{,}992)$$

S pravděpodobností 95% tento interval obsahuje neznámou střední hodnotu μ, ale neobsahuje hodnotu 2. Lze tedy s velkou jistotou tvrdit, že automat není správně nastaven.

Př.: Z populace jedenáctiletých chlapců bylo náhodně vybráno 16 a byla zjištěna jejich hmotnost (v kilogramech):

33,1	36,7	34,5	30,5	35,9	36,5	40,5	37,9
38,2	39,5	28,9	36,3	35,5	35,8	45,8	43,4

Měření budeme považovat za realizaci náh. výběru z rozdělení $N(\mu, \sigma^2)$. Chceme 95%-ní interval spolehlivosti pro střední hmotnost jedenáctiletých chlapců.

Problém: nelze použít předchozí postup, protože neznáme směrodatnou odchylku měření σ.

Interval spol. pro μ, když σ^2 neznáme, u $N(\mu, \sigma^2)$

neznámé σ nahradíme odhadem, tzv. **výběrovou směrodatnou odchylkou**

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

$100(1-\alpha)\%$-ní interval spolehlivosti pro μ a neznámé σ^2 pro výběr z normálního rozdělení je

$$\left(\overline{X} - t_{n-1}(1-\alpha/2)\cdot\frac{S}{\sqrt{n}};\ \overline{X} + t_{n-1}(1-\alpha/2)\cdot\frac{S}{\sqrt{n}}\right)$$

- nahrazení kvantilu $\Phi^{-1}(1-\alpha/2)$ kvantilem $t_{n-1}(1-\alpha/2)$ (je větší $\rightarrow$ širší interval) je daní za to, že neznámou hodnotu σ nahrazujeme jejím odhadem S.
- $t_n(\alpha)$ označuje α-kvantil tzv. (Studentova) t-rozdělení o n stupních volnosti; najdeme ho v tabulkách
- interpretace je stejná jako u předchozího intervalu

zpět k Př.: Z 16 naměřených hodnot chceme spočítat 95%-ní interval spolehlivosti pro střední hmotnost.

- spočítáme $\overline{X} = 36{,}8125$, $S = 4{,}2711$ a položíme $n = 16$
- pro 95%-ní int. spol. položíme $\alpha = 0{,}05$ a najdeme $t_{15}(1 - 0{,}05/2) = t_{15}(0{,}975) \doteq 2{,}13$

Tedy s 95%-ní pravděpodobností je střední hmotnost pokryta intervalem:

$$\left(\overline{X} - t_{n-1}(1-\alpha/2)\cdot\frac{S}{\sqrt{n}};\ \overline{X} + t_{n-1}(1-\alpha/2)\cdot\frac{S}{\sqrt{n}}\right) \doteq$$

$$\doteq \left(36{,}8125 - 2{,}13\cdot\frac{4{,}2711}{\sqrt{16}};\ 36{,}8125 + 2{,}13\cdot\frac{4{,}2711}{\sqrt{16}}\right) \doteq$$

$$\doteq (36{,}8125 - 2{,}274;\ 36{,}8125 + 2{,}274) \doteq$$

$$\doteq (34{,}54;\ 39{,}09)$$

- pro 99%-ní int. spol. je $\alpha = 0{,}01$ a $t_{15}(1 - 0{,}01/2) = t_{15}(0{,}995) = 2{,}95$

tedy 99%-ní interval spolehlivosti pro μ je $(33{,}66;\ 39{,}96)$

,

párová data

Někdy máme k dispozici dvě sady dat (měření) a snažíme se je porovnat (jejich střední hodnoty). Označme napozorované veličiny $(X_1, Y_1), \ldots, (X_n, Y_n)$ a předpokládejme, že veličiny X a Y se stejným indexem nelze považovat za nezávislé (často proto, že jsou měřena na jednom objektu), ale veličiny s různými indexy za nezávislé považovat již lze (měření spolu nesouvisející, např. proto, že jsou provedena na různých objektech).

Př.: Náhodně vybráno 8 lidí, kteří byli podrobeni dietě. Byla zaznamenána jejich hmotnost (v kg) před dietou a po ní.

Osoba	1	2	3	4	5	6	7	8
Před	81	85	92	82	86	88	79	85
Po	84	68	73	79	71	80	71	72

Chtěli bychom zjistit, zda má dieta vliv na hmotnost.

Interval spol. pro $\mu = \mu_X - \mu_Y$, párová data

Předpokládejme, že máme dvourozměrný náhodný výběr (X_1, Y_1), $\ldots, (X_n, Y_n)$ takový, že X a Y tvoří páry, které nelze považovat za nezávislé. Označme $\mu_X = EX_i$ a $\mu_Y = EY_i$.

Dále položme $Z_1 = X_1 - Y_1, \ldots, Z_n = X_n - Y_n$ a předpokládejme, že veličiny Z se dají považovat za náhodný výběr z rozdělení $N(\mu, \sigma^2)$, kde $\mu = \mu_X - \mu_Y$.
Zajímá nás intervalový odhad rozdílu $\mu = \mu_X - \mu_Y$ (podobně jako pro jeden výběr s neznámým rozptylem).
$100(1-\alpha)\%$-ní interval spolehlivosti pro $\mu = \mu_X - \mu_Y$:

$$\left(\overline{Z} - t_{n-1}(1-\alpha/2) \cdot \frac{S_Z}{\sqrt{n}};\ \overline{Z} + t_{n-1}(1-\alpha/2) \cdot \frac{S_Z}{\sqrt{n}}\right)$$

tedy 95%-ní interval spolehlivosti pro $\mu = \mu_X - \mu_Y$:

$$= \left(10 - 2{,}365 \cdot \frac{7{,}4642}{\sqrt{8}}, 10 + 2{,}365 \cdot \frac{7{,}4642}{\sqrt{8}}\right) = (3{,}76;\ 16{,}24)$$

tj. střední úbytek hmotnosti je mezi 3,76 a 16,24 kg.

,

Dva nezávislé výběry

Někdy máme k dispozici dvě sady dat (měření), které se snažíme porovnat (jejich střední hodnoty), přičemž veličiny nejsou párově závislé a nemusí jich být stejný počet. Označme napozorované veličiny $X_1, \ldots, X_n$ a $Y_1 \ldots, Y_m$ a budeme je považovat za dva nezávislé náhodné výběry (všechny veličiny jsou mezi sebou nezávislé).

Př.: Ve třídě byly zjištěny následující výšky žáků (v cm):

Chlapci	130	140	136	141	139	133	149	151
Dívky	135	141	143	132	146	146	151	141
Chlapci	139	136	138	142	127	139	147	
Dívky	141	131	142	141				

Ověřme, zda jsou chlapci a dívky v průměru stejně vysocí. Volte $\alpha = 0{,}05$.

Interval spol. pro $\mu = \mu_X - \mu_Y$, dva nezávislé výběry

Předpokládejme, že máme náhodný výběr $X_1, \ldots, X_n \sim N(\mu_X, \sigma^2)$ a náhodný výběr $Y_1, \ldots, Y_m \sim N(\mu_Y, \sigma^2)$ a tyto dva výběry jsou nezávislé se stejným rozptylem.
Položíme

$$S^{*2} = \frac{1}{n+m-2} \cdot \left((n-1) \cdot S_X^2 + (m-1) \cdot S_Y^2 \right),$$

kde $S_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$ a $S_Y^2 = \frac{1}{m-1} \sum_{i=1}^{m} (Y_i - \overline{Y})^2$.

$100(1-\alpha)\%$-ní interval spolehlivosti pro $\mu_X - \mu_Y$:

$$\left(\overline{X} - \overline{Y} - t_{n+m-2}(1-\alpha/2) \cdot S^* \cdot \sqrt{\frac{n+m}{n \cdot m}};\ \overline{X} - \overline{Y} + t_{n+m-2}(1-\alpha/2) \cdot S^* \cdot \sqrt{\frac{n+m}{n \cdot m}} \right)$$

zpět k ▸ Př.: na hladině $\alpha = 0{,}05$ testujte, zda jsou chlapci a dívky v průměru stejně vysocí.

Chlapci	130	140	136	141	139	133	149	151
Dívky	135	141	143	132	146	146	151	141
Chlapci	139	136	138	142	127	139	147	
Dívky	141	131	142	141				

Jsou chlapci a dívky v průměru stejně vysocí? Spočteme
$\overline{X} = 139{,}133; \quad \overline{Y} = 140{,}833; \quad S_X^2 = 42{,}981; \quad S_Y^2 = 33{,}788;$

$$S^* = \sqrt{\frac{1}{n+m-2} \cdot ((n-1) \cdot S_X^2 + (m-1) \cdot S_Y^2)} = \sqrt{\frac{1}{25}(14 \cdot 42{,}981 + 11 \cdot 33{,}788)} = 6{,}240$$

tedy 95%-ní interval spolehlivosti pro $\mu = \mu_X - \mu_Y$:

$$\left(139{,}133 - 140{,}833 \mp t_{25}(0{,}975) \cdot 6{,}240 \cdot \sqrt{\frac{15+12}{15 \cdot 12}}\right) = (-14{,}55;\ 11{,}15)$$

tj. střední rozdíl střední výšky chlapců a dívek je mezi $-14{,}55$ a $11{,}15$ cm.
Závěr: je možné, že chlapci a dívky jsou v průměru stejně vysocí.

Př.: U stroje na výrobu součástek by měla být podle normy jeho chybovost (tj. pravděpodobnost, že vyrobí zmetek) nejvýše 10%. Při kontrole náhodného vzorku 400 součástek bylo mezi nimi zjištěno 42 zmetků. Jak určit 95%-ní a 99%-ní interval spolehlivosti pro chybovost stroje.

- označme jako p neznámou chybovost stroje
- náh. vybráno $n = 400$ součástek, každá s pravděp. p zmetek
- tedy celkový počet zmetků mezi vybranými $Y \sim Bi(n = 400, p)$
- náh. výběrem zjištěn počet zmetků ve výběru (absolutní četnost) $y = 42$ (realizací Y zjištěna hodnota y)

▶ bodovým odhadem p je relativní četnost $\hat{p} = \frac{y}{n} = \frac{42}{400} = 0{,}105$

▶ jak bychom mohli odhadnout p intervalem?

- z Centrální limitní věty: pro $Y \sim Bi(n, p)$ má $Y \dot{\sim} N(n \cdot p, n \cdot p \cdot (1 - p))$ pro dostatečně velké n
- tedy $\frac{Y}{n} \dot{\sim} N(p, \frac{p \cdot (1-p)}{n})$

Interval spol. pro parametr p binomického rozdělení

Máme-li náh. veličinu Y z rozdělení $Bi(n, p)$, pak $\frac{Y}{n} \dot{\sim} N(p, \frac{p\cdot(1-p)}{n})$ a protože rozptyl (kvůli neznámému p) tohoto rozdělení neznáme, nahradíme p v rozptylu odhadem $\hat{p}$. Tedy $\frac{Y}{n} \dot{\sim} N(p, \frac{\hat{p}\cdot(1-\hat{p})}{n})$ a platí

$$P\left(-\Phi^{-1}(1-\alpha/2) < \frac{\frac{Y}{n} - p}{\sqrt{\hat{p}\cdot(1-\hat{p})}} \cdot \sqrt{n} < \Phi^{-1}(1-\alpha/2)\right) = 1-\alpha$$

za $\frac{Y}{n}$ pak dosadíme napozorovanou relativní četnost $\frac{y}{n} = \hat{p}$ a dostaneme:
100$(1-\alpha)$%-ní int. spol. pro parametr p binomického rozdělení je

$$\left(\hat{p} - \Phi^{-1}(1-\alpha/2)\cdot\sqrt{\frac{\hat{p}\cdot(1-\hat{p})}{n}};\ \hat{p} + \Phi^{-1}(1-\alpha/2)\cdot\sqrt{\frac{\hat{p}\cdot(1-\hat{p})}{n}}\right)$$

- interpretace je podobná jako u předchozích intervalů

zpět k Př.: Ze 400 náh. vybraných součástek bylo 42 zmetků. Chceme spočítat 95%-ní a 99%-ní interval spolehlivosti pro chybovost stroje.

- bodovým odhadem chybovosti p je podíl vadných ve výběru $\hat{p} = \frac{y}{n} = \frac{42}{400} = 0{,}105$
- pro 95%-ní (resp. 99%-ní) int. spol. položíme $\alpha = 0{,}05$ (resp. $\alpha = 0{,}01$)
- a najdeme $\Phi^{-1}(1 - 0{,}05/2) = \Phi^{-1}(0{,}975) = 1{,}96$ a $\Phi^{-1}(1 - 0{,}01/2) = \Phi^{-1}(0{,}995) = 2{,}58$

Tedy 95%-ní int. spol. pro chybovost p stroje je:

$$\left(\hat{p} - \Phi^{-1}(1-\alpha/2) \cdot \sqrt{\frac{\hat{p}\cdot(1-\hat{p})}{n}};\ \hat{p} - \Phi^{-1}(1-\alpha/2) \cdot \sqrt{\frac{\hat{p}\cdot(1-\hat{p})}{n}}\right) \doteq$$

$$\doteq \left(0{,}105 - 1{,}96 \cdot \sqrt{\frac{0{,}105\cdot(1-0{,}105)}{400}};\ 0{,}105 + 1{,}96 \cdot \sqrt{\frac{0{,}105\cdot(1-0{,}105)}{400}}\right)$$

$$\doteq (0{,}075;\ 0{,}135) = (7{,}5\%;\ 13{,}5\%)$$

resp. 99%-ní int. spol. by vyšel $(0{,}065;\ 0{,}145) = (6{,}5\%;\ 14{,}5\%)$

,

Porovnání dvou populačních pravděpodobností

Někdy Chceme porovnat, zda výskyt nějakého jevu je stejně pravděpodobný ve dvou různých populacích.

Př.: Do průzkumu veřejného mínění bylo zapojeno 800 náhodně vybraných osob. Odpovídali na otázku, zda by se měla zvýšit daň z tabáku. Z 605 nekuřáků odpovědělo 351 ano. Ze 195 kuřáků odpovědělo ano 41. Je to dostatečný důkaz, abychom na hladině významnosti $\alpha = 0{,}05$ mohli tvrdit, že se v této otázce populace kuřáků a populace nekuřáků významně liší?

Počet nekuřáků s kladnou odpovědí má $Bi(n = 605, p_1)$. Počet kuřáků s kladnou odpovědí má $Bi(n = 195, p_2)$. Chceme ověřit, zda $p_1 = p_2$.

Porovnání dvou populačních pravděpodobností

Předpokládejme, že máme napozorované nezávislé náhodné veličiny $Y_1 \sim Bi(n_1, p_1)$ a $Y_2 \sim Bi(n_2, p_2)$.
Položíme $\hat{p}_1 = Y_1/n_1$, $\hat{p}_2 = Y_2/n_2$ a $\hat{p} = (Y_1 + Y_2)/(n_1 + n_2)$

$100(1-\alpha)\%$-ní interval spolehlivosti pro $p_1 - p_2$:

$$\left(\hat{p}_1 - \hat{p}_2 \mp \Phi^{-1}(1-\alpha/2) \cdot \sqrt{\frac{\hat{p}_1 \cdot (1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2 \cdot (1-\hat{p}_2)}{n_2}} \right)$$

zpět k Př.: Z 605 nekuřáků odpovědělo 351 ano. Ze 195 kuřáků odpovědělo ano 41. Lze tvrdit, že se názor kuřáků a nekuřáků významně liší?

95%-ní interval spolehlivosti pro $p_1 - p_2$:

$$\left(0{,}58 - 0{,}21 \mp 1{,}96 \cdot \sqrt{\frac{0{,}58 \cdot (1 - 0{,}58)}{605} + \frac{0{,}21 \cdot (1 - 0{,}21)}{195}}\right) = (0{,}30;\ 0{,}44)$$

Tj. v populaci nekuřáků je o 30 až 44% více příznivců zvýšení daně. Jejich názory se tedy významně liší.

Vlastnosti intervalů spolehlivosti

- šířka intervalu roste s vyšší požadovanou spolehlivostí (viz. poslední příklad)
- šířka intervalu klesá s vyšším n (počtem pozorování)
  ▶ např. u intervalu pro μ u $N(\mu, \sigma^2)$ nebo pro p u $Bi(n, p)$ je šířka nepřímo úměrná odmocnině z n; a tedy k získání dvakrát užšího (přesnějšího) intervalu spolehlivosti je třeba 4-krát více pozorování
- v některých případech lze z požadavku na šířku intervalu odhadnout potřebný počet pozorování n.

Jak ověřovat hypotézy?

- jak rozhodnout, zda platí tvrzení o neznámém parametru rozdělení?
- spočítali jsme intervalový odhad pro střední množství μ coly v lahvi: (1,972; 1,992)
- lze (a s jakou jistotou) tvrdit, že je automat špatně nastaven?
- požadavek: chtěli bychom např., aby pravděpodobnost “křivého obvinění” byla malá
- proto: zavádíme standardizované postupy pro takové rozhodování

Testování hypotéz

$X_1, X_2, \ldots, X_n$ je náh. výb. z rozdělení s nezn. parametrem(y).
Máme dvě odporující si hypotézy o parametru(ech) daného rozdělení:

- tzv. **nulovou hypotézu H_0**: parametr se rovná určité hodnotě, parametry se rovnají,...
- tzv. **alternativní hypotézu H_1**: opak nulové hypotézy, často to, co se snažíme prokázat

Podle typu H_0 a H_1 se zvolí rozhodovací kritérium (test, testové kritérium), které závisí na (vypočtu ho z) realizovaném náhodném výběru (napozorovaných datech).
Možná rozhodnutí:

- zamítáme H_0, pokud data (a tedy i test) svědčí proti ní
- nezamítáme H_0, pokud data (a tedy i test) neposkytují dostatek "důkazů" proti H_0

Postup a možné chyby při rozhodování

- **chyba 1. druhu**: H_0 platí a my ji zamítneme
- **chyba 2. druhu**: H_0 neplatí a my ji nezamítneme

hladina testu: označujeme ji α (tu volíme, nejčastěji $= 0{,}05$), je nejvyšší přípustná pravděpodobnost chyby 1. druhu

rozhodnutí\skutečnost	H_0 platí	H_0 neplatí
nezamítáme H_0	správně	chyba 2. druhu
zamítáme H_0	chyba 1. druhu $\leq \alpha$	správně

Postup: Podle toho, co chceme zjistit, zformulujeme H_0 a H_1 a zvolíme α. Pak zvolíme vhodné rozhodovací kritérium: tj. z testů, jejichž hladina je menší než α vybereme obvykle ten s minimální pravděpodobností chyby 2. druhu

zpět k ▸ Př.: Náhodně vybráno 100 lahví coly a byl zjištěn jejich průměrný obsah $\overline{X} = 1{,}982$ litrů. Naměřené hodnoty považujeme za realizaci náhodného výběru z rozdělení $N(\mu, \sigma^2 = 0{,}0025)$. Dá se tvrdit, že je automat špatně nastaven?

Chtěli bychom provést na hladině $\alpha = 0{,}05$ test hypotézy

- $H_0 : \mu = 2$ litry (automat je správně nastaven)

proti alternativě

- $H_1 : \mu \neq 2$ litry (automat není správně nastaven)

Jak zvolit testové kritérium?

Z-test: jednovýběrový test střední hodnoty (σ^2 známe)

$X_1, X_2, \ldots, X_n$ je náh. výb. z rozdělení $N(\mu, \sigma^2)$, kde σ^2 známe. Z již odvozeného plyne, že

$$P\left(\frac{|\overline{X} - \mu|}{\sigma} \cdot \sqrt{n} \geq \Phi^{-1}(1 - \alpha/2)\right) = \alpha$$

Tedy pro test hypotézy $H_0 : \mu = \mu_0$ proti alternativě $H_1 : \mu \neq \mu_0$ lze použít testovou statistiku

$$Z = \frac{\overline{X} - \mu_0}{\sigma} \cdot \sqrt{n}$$

a na hladině α zamítáme hypotézu H_0 (přikloníme se k H_1), pokud $|Z| \geq \Phi^{-1}(1 - \alpha/2)$

- pokud $|Z| < \Phi^{-1}(1 - \alpha/2)$, tak H_0 nezamítáme. Závěr: H_0 může platit
- Pozn.: Pro dostatečně velká n platí díky Centrální limitní větě i pro jiná rozdělení než normální

zpět k ▸ Př.: Náhodně vybráno 100 lahví coly, $\overline{X} = 1{,}982$ litrů. Předp, že data pocházejí z rozdělení $N(\mu, \sigma^2 = 0{,}0025)$. Dá se tvrdit, že je automat špatně nastaven?

Chtěli bychom provést na hladině $\alpha = 0{,}05$ test hypotézy

- $H_0 : \mu = 2$ litry (automat je správně nastaven)

proti alternativě

- $H_1 : \mu \neq 2$ litry (automat není správně nastaven)

Testové kritérium (testová statistika) je

$$Z = \frac{\overline{X} - \mu_0}{\sigma} \cdot \sqrt{n} = \frac{1{,}982 - 2}{0{,}05} \cdot \sqrt{100} = -3{,}6$$

Tedy

$$|Z| = 3{,}6 \geq \Phi^{-1}(1 - \alpha/2) = \Phi^{-1}(0{,}975) = 1{,}96$$

a proto na hladině 0,05 zamítáme H_0 a přikláníme se k H_1

Závěr: automat není správně nastaven

Lze usoudit i z toho, že: $2 \notin (1{,}972;\ 1{,}992)$ (95%-ní int. spol. pro μ)

zpět k ▸ Př.: Byla změřena hmotnost 16 jedenáctiletých chlapců. Měření považujeme za realizaci náh. výběru z rozdělení $N(\mu, \sigma^2)$. Lze tvrdit, že se jejich hmotnost změnila oproti době před 25 lety, kdy byla střední hmotnost jedenáctiletých 34 kg? Volme hladinu testu $\alpha = 0{,}01$

Chtěli bychom tedy provést na hladině $\alpha = 0{,}01$ test hypotézy

- $H_0 : \mu = 34$ kg (hmotnost je rovna hmotnosti před 25 lety)

proti alternativě

- $H_1 : \mu \neq 34$ kg (hmotnost není rovna hmotnosti před 25 lety)

Problém: nelze použít předchozí postup, protože neznáme směrodatnou odchylku měření σ.

Jednovýběrový t-test: test stř. hodnoty (σ^2 neznáme)

$X_1, X_2, \ldots, X_n$ je náh. výb. z rozdělení $N(\mu, \sigma^2)$, kde σ^2 neznáme. Platí, že $\frac{\overline{X}-\mu}{S} \cdot \sqrt{n} \sim t_{n-1}$, z čehož podobně jako u Z-testu plyne:

$$P\left(\frac{|\overline{X} - \mu|}{S} \cdot \sqrt{n} \geq t_{n-1}(1 - \alpha/2)\right) = \alpha$$

Tedy pro test hypotézy $H_0 : \mu = \mu_0$ proti alternativě $H_1 : \mu \neq \mu_0$ lze použít testovou statistiku

$$T = \frac{\overline{X} - \mu_0}{S} \cdot \sqrt{n}$$

a na hladině α zamítáme hypotézu H_0 (přikloníme se k H_1), pokud $|T| \geq t_{n-1}(1 - \alpha/2)$

- pokud $|T| < t_{n-1}(1 - \alpha/2)$, tak H_0 nezamítáme. Závěr: H_0 může platit

zpět k Př.: Byla změřena hmotnost 16 jedenáctiletých chlapců. Měření pocházejí z rozdělení $N(\mu, \sigma^2)$. Lze tvrdit, že se jejich hmotnost změnila oproti době před 25 lety, kdy byla střední hmotnost jedenáctiletých 34 kg?

Chtěli bychom tedy provést na hladině $\alpha = 0{,}01$ test hypotézy

- $H_0 : \mu = 34$ kg (hmotnost je rovna hmotnosti před 25 lety)

proti alternativě

- $H_1 : \mu \neq 34$ kg (hmotnost není rovna hmotnosti před 25 lety)

Testové kritérium (testová statistika) je

$$T = \frac{\overline{X} - \mu_0}{S} \cdot \sqrt{n} = \frac{36{,}8125 - 34}{4{,}2711} \cdot \sqrt{16} = 2{,}634$$

Tedy

$$|T| = 2{,}634 < t_{n-1}(1 - \alpha/2) = t_{15}(0{,}995) = 2{,}947$$

a proto na hladině 0,01 nezamítáme H_0

Závěr: Nevylučujeme, že je hmotnost rovna hmotnosti před 25 lety

- Pozn.: na hladině $\alpha = 0{,}05$ bychom H_0 zamítli (přiklonili se k H_1), protože $|T| = 2{,}634 \geq t_{n-1}(1 - \alpha/2) = t_{15}(0{,}975) = 2{,}131$ (ekviv. $34 \notin (34{,}54;\ 39{,}09)$)

Párový t-test

zpět k ▸ Př.: Předpokládejme, že máme dvourozměrný náhodný výběr $(X_1, Y_1), \ldots, (X_n, Y_n)$ takový, že X a Y tvoří páry, které nelze považovat za nezávislé. Označme $\mu_X = EX_i$ a $\mu_Y = EY_i$.

Dále položme $Z_1 = X_1 - Y_1, \ldots, Z_n = X_n - Y_n$ a předpokládejme, že veličiny Z se dají považovat za náhodný výběr z rozdělení $N(\mu, \sigma^2)$, kde $\mu = \mu_X - \mu_Y$.
Tedy test hypotézy, že obě sady měření pocházejí z rozdělení o stejné střední hodnotě $H_0 : \mu_X - \mu_Y = 0$ je totéž jako test hypotézy $H_0 : \mu = 0$. Test hypotézy $H_0 : \mu = 0$ proti alternativě $H_1 : \mu \neq 0$ je úlohou jednovýběrového t-testu.

Tedy spočítáme $\overline{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i$ a $S_Z^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \overline{Z})^2$ a pokud

$$|T| = \frac{|\overline{Z} - 0|}{S_Z} \cdot \sqrt{n} \geq t_{n-1}(1 - \alpha/2)$$

tak na hladině α zamítáme hypotézu H_0 (přikloníme se k $H_1 : \mu_X \neq \mu_Y$)

,

Párový t-test

Pro test hypotézy, že obě sady měření pocházejí z rozdělení o stejné střední hodnotě $H_0 : \mu = 0$ proti alternativě $H_1 : \mu \neq 0$ na hladině α, lze použít i interval spolehlivosti:
Pokud $0 \notin \left(\overline{Z} - t_{n-1}(1-\alpha/2) \cdot \frac{S_Z}{\sqrt{n}};\ \overline{Z} + t_{n-1}(1-\alpha/2) \cdot \frac{S_Z}{\sqrt{n}}\right)$ tak na hladině α zamítáme hypotézu H_0 (přikloníme se k $H_1 : \mu_X \neq \mu_Y$)

zpět k Př.: 8 lidí podrobeno dietě. Má dieta vliv na hmotnost?

Osoba	1	2	3	4	5	6	7	8
X=Před	81	85	92	82	86	88	79	85
Y=Po	84	68	73	79	71	80	71	72
Z=Rozdíl	-3	17	19	3	15	8	8	13

Provedeme na hladině $\alpha = 0{,}05$ test hypotézy

- $H_0 : \mu = \mu_X - \mu_Y = 0$ kg (dieta nemá vliv na hmotnost)
- proti $H_1 : \mu = \mu_X - \mu_Y \neq 0$ kg (dieta má vliv na hmotnost)

Spočteme $\overline{Z} = 10$ a $S_Z = \sqrt{S_Z^2} = \sqrt{55{,}71429} = 7{,}4642$

Testová statistika je

$$T = \frac{\overline{Z} - 0}{S_Z} \cdot \sqrt{n} = \frac{10 - 0}{7{,}4642} \cdot \sqrt{8} = 3{,}789$$

Tedy

$$|T| = 3{,}789 \geq t_{n-1}(1 - \alpha/2) = t_7(0{,}975) = 2{,}365$$

a proto na hladině 0,05 zamítáme H_0.

Závěr: dieta má vliv na hmotnost.

- Pozn.: i pro $\alpha = 0{,}01$ bychom H_0 zamítali ($t_7(0{,}995) = 3{,}499$)

Dvouvýběrový t-test

zpět k ▸ Př.:Předpokládejme, že máme náhodný výběr $X_1, \ldots, X_n \sim N(\mu_X, \sigma^2)$ a náhodný výběr $Y_1, \ldots, Y_m \sim N(\mu_Y, \sigma^2)$ a tyto dva výběry jsou nezávislé se stejným rozptylem.
Položíme

$$S^{*2} = \frac{1}{n+m-2} \cdot \left((n-1) \cdot S_X^2 + (m-1) \cdot S_Y^2 \right),$$

kde $S_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$ a $S_Y^2 = \frac{1}{m-1} \sum_{i=1}^{m} (Y_i - \overline{Y})^2$.

Pro test hypotézy, že obě sady měření pocházejí z rozdělení o stejné střední hodnotě $H_0 : \mu_X - \mu_Y = 0$ proti alternativě $H_1 : \mu_X - \mu_Y \neq 0$ lze použít statistiku:

$$T = \frac{\overline{X} - \overline{Y} - 0}{S^*} \cdot \sqrt{\frac{n \cdot m}{n+m}}$$

a pokud $|T| \geq t_{n+m-2}(1 - \alpha/2)$, tak na hladině α zamítáme hypotézu H_0 (přikloníme se k $H_1 : \mu_X \neq \mu_Y$ střední hodnoty nejsou stejné)

,

Pro test hypotézy, že obě sady měření pocházejí z rozdělení o stejné střední hodnotě $H_0 : \mu_X - \mu_Y = 0$ proti alternativě $H_1 : \mu_X - \mu_Y \neq 0$ na hladině α, lze použít i interval spolehlivosti:

$$\left(\overline{X} - \overline{Y} - t_{n+m-2}(1-\alpha/2) \cdot S^* \cdot \sqrt{\frac{n+m}{n \cdot m}};\ \overline{X} - \overline{Y} + t_{n+m-2}(1-\alpha/2) \cdot S^* \cdot \sqrt{\frac{n+m}{n \cdot m}} \right)$$

Pokud 0 neleží v tomto $100(1-\alpha)\%$-ním intervalu spolehlivosti pro $\mu_X - \mu_Y$, tak na hladině α zamítáme hypotézu H_0 (přikloníme se k $H_1 : \mu_X \neq \mu_Y$)

zpět k ▸ Př.: na hladině $\alpha = 0{,}05$ testujte, zda jsou chlapci a dívky v průměru stejně vysocí.

Chlapci	130	140	136	141	139	133	149	151
Dívky	135	141	143	132	146	146	151	141
Chlapci	139	136	138	142	127	139	147	
Dívky	141	131	142	141				

- test $H_0 : \mu_X - \mu_Y = 0$ cm (jsou stejně vysocí)
- proti $H_1 : \mu_X - \mu_Y \neq 0$ cm (nejsou stejně vysocí)

Spočteme $\overline{X} = 139{,}133$; $\overline{Y} = 140{,}833$; $S_X^2 = 42{,}981$; $S_Y^2 = 33{,}788$;

$$S^* = \sqrt{\frac{1}{n+m-2} \cdot ((n-1) \cdot S_X^2 + (m-1) \cdot S_Y^2)} = \sqrt{\frac{1}{25}(14 \cdot 42{,}981 + 11 \cdot 33{,}788)} = 6{,}240$$

Testová statistika je

$$T = \frac{\overline{X} - \overline{Y} - 0}{S^*} \cdot \sqrt{\frac{n \cdot m}{n+m}} = \frac{139{,}133 - 140{,}833 - 0}{6{,}240} \cdot \sqrt{\frac{15 \cdot 12}{15+12}} = -0{,}703$$

Tedy $|T| = 0{,}703 < t_{n+m-2}(1-\alpha/2) = t_{25}(0{,}975) = 2{,}060$ a proto na hladině 0,05 nezamítáme H_0.

Závěr: je možné, že chlapci a dívky jsou v průměru stejně vysocí.

- Na každé nižší hladině (i $\alpha = 0{,}01$) bychom H_0 tím spíše nezamítli

,

Znaménkový test

Někdy máme k dispozici jen informaci, kolikrát při velkém počtu nezávislých opakování zkoumaná veličina překročila (+) nebo byla menší (-) než nějaká daná hodnota. A chceme testovat hypotézu, že obojí nastává se stejnou pravděpodobností, tj. že medián (50%-ní kvantil) rozdělení je roven té dané hodnotě.

Př.: Ze 46 piv, které se u vašeho stolu večer vypily, bylo 27 podmírových a 19 nadmírových. Lze tvrdit, že výčepní systematicky šidí (ať už zákazníky nebo majitele hospody)?
Chceme ověřit, zda medián množství piva ve sklenici může být půl litru. Známe přitom jen počet piv pod a nad touto mírou. Jak zvolit testové kritérium?

Znaménkový test - asymptotický (pro velké n)

Máme veličiny $X_1, \ldots, X_n$ ze spojitého rozdělení s mediánem $\tilde{x}$. Tedy platí

$$P(X_i < \tilde{x}) = P(X_i > \tilde{x}) = \frac{1}{2} \qquad i = 1, \ldots, n$$

Chceme testovat hypotézu $H_0 : \tilde{x} = x_0$ proti $H_1 : \tilde{x} \neq x_0$, kde x_0 je dané číslo.

Utvoří se rozdíly $X_1 - x_0, \ldots, X_n - x_0$ a ty nulové se vynechají (a příslušně se zmenší n).

Za platnosti H_0 má počet rozdílů s kladným znaménkem $Y \sim Bi(n, p = 1/2)$ a tedy podle Centrální limitní věty pro velké n platí: Y má přibližně normální rozdělení $N(n/2, n/4)$

Za platnosti H_0 tedy

$$U = \frac{Y - n/2}{\sqrt{n/4}} = \frac{2Y - n}{\sqrt{n}} \dot{\sim} N(0, 1)$$

$H_0 : \tilde{x} = x_0$ na hladině α zamítneme, pokud $|U| \geq \Phi^{-1}(1 - \alpha/2)$

zpět k ▸ Př.: Ze 46 piv bylo 27 podmírových a 19 nadmírových. Lze tvrdit, že výčepní nedodržuje míru (ať už jedním nebo druhým směrem)?

Na hladině $\alpha = 0{,}05$ testovat $H_0 : \tilde{x} = 500$ ml proti $H_1 : \tilde{x} \neq 500$ ml.

Asymptotický test: Spočteme

$$U = \frac{2Y - n}{\sqrt{n}} = \frac{2 \cdot 19 - 46}{\sqrt{46}} = -1{,}180$$

H_0 nezamítáme, protože $|U| = 1{,}180 \not\geq \Phi^{-1}(0{,}975) = 1{,}960$

Znaménkový test - možné použití

- test o mediánu u náh. výběru $X_1, \ldots, X_n$ ze spojitého rozdělení.
- lze použít i namísto jednovýběrového (resp. párového) t-testu
- výhoda: nevyžaduje se normální rozdělení výběru
- nevýhoda: u normálně rozděleného výběru je o něco vyšší chyba 2. druhu v porovnání s t-testem
- Pokud jsme si jisti normalitou dat, je tedy nejlepší použít t-test

Zkuste použít znaménkový test na příklady, na které byly použity jednovýběrový nebo párový t-test

Test o parametru p binomického rozdělení

Někdy máme k dispozici jen informaci, kolikrát při velkém počtu nezávislých opakování nastal nějaký jev a zajímá nás pravděpodobnost (chceme testovat hypotézu o pravděpodobnosti), že daný jev nastane.

Př.: Při 600 hodech kostkou padla šestka 137-krát. Testujte hypotézu, že šestka padá na této kostce s pravděpodobností $1/6$

Počet šestek má $Bi(n = 600, p)$. Chceme ověřit, zda $p = 1/6$. Jak zvolit testové kritérium?

Test o parametru p binom. rozd. (asymptotický)

Předpokládejme, že máme napozorovanou realizaci náhodné veličiny $Y \sim Bi(n, p)$, tj. např. počet událostí v n stejných nezávislých pokusech. $\hat{p} = Y/n$

Chceme testovat hypotézu o pravděpodobnosti p, že událost nastane $H_0 : p = p_0$ proti alternativě $H_1 : p \neq p_0$

Z Centrální limitní věty pro velké n platí: Y má přibližně normální rozdělení

$$N(n \cdot p, n \cdot p \cdot (1 - p))$$

Za platnosti H_0 tedy

$$U = \frac{Y - n \cdot p_0}{\sqrt{n \cdot p_0 \cdot (1 - p_0)}} = \frac{\hat{p} - p_0}{\sqrt{p_0 \cdot (1 - p_0)/n}} \dot{\sim} N(0, 1)$$

$H_0 : p = p_0$ na hladině α zamítneme, pokud $|U| \geq \Phi^{-1}(1 - \alpha/2)$

- Pozn.: Znaménkový test je speciálním případem pro $p_0 = 1/2$

zpět k ▸ Př.: Při 600 hodech kostkou padla 137-krát šestka. Ověřme, zda šestka padá na této kostce s pravděpodobností $1/6$.
Na hladině $\alpha = 0{,}05$ testovat $H_0 : p = 1/6$ proti $H_1 : p \neq 1/6$.

(Asymptotický) test toho, že parametr p binom. rozd. je roven $1/6$*:*
Spočteme

$$U = \frac{137 - 600 \cdot 1/6}{\sqrt{600 \cdot 1/6 \cdot 5/6}} = \frac{137 - 100}{\sqrt{83{,}33}} = 4{,}053$$

a H_0 zamítáme, protože $|U| = 4{,}053 \geq \Phi^{-1}(0{,}975) = 1{,}960$

Porovnání dvou populačních pravděpodobností

viz. ▸ příklad:

Předpokládejme, že máme napozorované nezávislé náhodné veličiny $Y_1 \sim Bi(n_1, p_1)$ a $Y_2 \sim Bi(n_2, p_2)$.
Položíme $\hat{p}_1 = Y_1/n_1$, $\hat{p}_2 = Y_2/n_2$ a $\hat{p} = (Y_1 + Y_2)/(n_1 + n_2)$
Chceme testovat hypotézu o pravděpodobnostech $H_0 : p_1 = p_2$ proti alternativě $H_1 : p_1 \neq p_2$
Pro velké n_1 a n_2 lze opět využít Centrální limitní větu.
Za platnosti H_0 je

$$U = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p} \cdot (1 - \hat{p}) \cdot (1/n_1 + 1/n_2)}} \dot{\sim} N(0, 1)$$

$H_0 : p_1 = p_2$ na hladině α zamítneme, pokud $|U| \geq \Phi^{-1}(1 - \alpha/2)$

zpět k ▸ Př.: Z 605 nekuřáků odpovědělo 351 ano. Ze 195 kuřáků odpovědělo ano 41. Lze tvrdit, že se názor kuřáků a nekuřáků významně liší?

Na hladině $\alpha = 0{,}05$ testovat $H_0 : p_1 = p_2$ proti $H_1 : p_1 \neq p_2$.

Spočteme

$$U = \frac{351/605 - 41/195}{\sqrt{\frac{392}{800} \cdot (1 - \frac{392}{800}) \cdot (\frac{1}{605} + \frac{1}{195})}} = \frac{0{,}58 - 0{,}21}{0{,}0412} = 8{,}99$$

a H_0 na hladině $\alpha = 0{,}05$ zamítáme, protože
$|U| = 8{,}99 \geq \Phi^{-1}(0{,}975) = 1{,}96$. Jejich názory se významně liší.

95%-ní interval spolehlivosti pro $p_1 - p_2$ vyšel $(0{,}30;\ 0{,}44)$.

Testy nezávislosti

Někdy máme k dispozici sadu dvojrozměrných veličin (opakovaná měření dvou znaků) a snažíme se zjistit, zda existuje závislost (korelace) mezi těmi dvěma znaky. Označme napozorované veličiny $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Př.: Ze studentů statistiky bylo náhodně vybráno 9 a byl jim dán matematický a jazykový test s následujícími výsledky:

Číslo studenta	1	2	3	4	5	6	7	8	9
Jazykový test	50	23	28	34	14	54	46	52	53
Matematický test	38	28	14	26	18	40	23	30	27.

Chtěli bychom zjistit, zda u studentů existuje závislost mezi výsledky jazykového a matematického testu.

Pozn.: je to jiná úloha, než rozhodnout, zda jsou u studentů výsledky jazykového a matematického testu na stejné úrovni (v tom případě by bylo na místě použít např. párový t-test příp. neparametrickou alternativu)

Jak zde zvolit testové kritérium?

(Pearsonův) korelační koeficient

Předpokládejme, že máme dvourozměrný náhodný výběr $(X_1, Y_1), \ldots, (X_n, Y_n)$, tj. veličiny s různými indexy jsou nezávislé. Označme S_X^2 a S_Y^2 výběrové rozptyly X a Y a dále **výběrovou kovarianci** mezi X a Y jako

$$S_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}\right) \cdot \left(Y_i - \overline{Y}\right) = \frac{1}{n-1} \left[\sum_{i=1}^{n} (X_i \cdot Y_i) - n \cdot \overline{X} \cdot \overline{Y}\right]$$

(Pearsonův) výběrový korelační koeficient:

$$r_{XY} = r = \frac{S_{XY}}{\sqrt{S_x^2 \cdot S_y^2}} = \frac{\sum_{i=1}^{n} (X_i \cdot Y_i) - n \cdot \overline{X} \cdot \overline{Y}}{\sqrt{\left(\sum_{i=1}^{n} X_i^2 - n \cdot \overline{X}^2\right)\left(\sum_{i=1}^{n} Y_i^2 - n \cdot \overline{Y}^2\right)}}$$

Za předpokladu normality spočítáme

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

a hypotézu nezávislosti veličin X a Y na hladině α zamítáme, jestliže $|T| \geq t_{n-2}(1 - \alpha/2)$

Na hladině $\alpha = 0{,}05$ testujme hypotézu nezávislosti mezi výsledky z jazykového a matematického testy z ▸ příkladu, kde bylo bylo vybráno a podrobeno oběma testům 9 studentů.

Jazykový test	50	23	28	34	14	54	46	52	53
Matematický test	38	28	14	26	18	40	23	30	27

Spočteme $S_X^2 = 223{,}25$ a $S_Y^2 = 70{,}86$ a
$S_{XY} = \frac{1}{8}(50 \cdot 38 + \ldots + 53 \cdot 27 - 9 \cdot 39{,}33 \cdot 27{,}11) = 85{,}46$
korelační koeficient je tedy $r = \frac{S_{XY}}{\sqrt{S_x^2 \cdot S_y^2}} = \frac{85{,}46}{14{,}94 \cdot 8{,}42} = 0{,}679$

Spočítáme

$$T = \frac{r}{\sqrt{1 - r^2}} \cdot \sqrt{n-2} = \frac{0{,}679}{\sqrt{1 - 0{,}679^2}} \cdot \sqrt{7} = 2{,}450$$

a protože $|T| = 2{,}450 \geq t_{n-2}(0{,}975) = 2{,}365$, tak hypotézu nezávislosti na hl. 0,05 zamítáme. Lze tedy tvrdit, že existuje závislost mezi výsledkem jazykového a matematického testu

,

Test nezávislosti v kontingenční tabulce

Někdy máme k dispozici data v kontingenční tabulce, např. proto, že měříme současně dva znaky v nominálním měřítku na n nezávislých objektech. Cílem je opět zjistit, zda existuje závislost mezi těmito dvěma znaky.

Př.: Za účelem zjištění, zda existuje vztah mezi pohlavím a úrovní strachu z matematiky bylo náhodně vybráno 100 středoškolských studentů, kteří byli podrobeni psychologickému testu, kterým byla zjištěna úrovně strachu (nízká, střední, vysoká), který v nich vyvolává matematika. Výsledky byly následující:

	strach z matematiky			
pohlaví	nízký	střední	vysoký	součet
muž	10	26	20	56
žena	4	10	30	44
součet	14	36	50	100

Ize použít χ^2-test dobré shody: porovnává napozorované četnosti s očekávanými za nezávislosti znaků

pohlaví	strach z mat			součet
	níz	stř	vys	
muž	10	26	20	56
žena	4	10	30	44
součet	14	36	50	100

pohlaví	strach z mat			součet
	níz	stř	vys	
muž	18%	46%	36%	100%
žena	9%	23%	68%	100%
celkem	14%	36%	50%	100%

- existuje vztah mezi pohlavím a strachem z matematiky?
- pokud jsou tyto dva znaky nezávislé, rozdělení procent pro obě pohlaví by mělo být podobné
- odhad pravděp., že pohl. studenta je ženské
  $P(\text{pohl.} = \text{ž}) = 44/100$
- odhad pravd., že strach studenta je vys. $P(\text{strach} = \text{v}) = 50/100$
- tedy odhad pravděp. (za nezávislosti), že studentem je žena s vysokým strachem
  $P(\text{pohl.} = \text{ž} \cap \text{strach} = \text{v}) = (44/100) \cdot (50/100) = 0{,}22$
- tedy mezi 100 studenty bych takových studentů očekával
  $100 \cdot (44/100) \cdot (50/100) = 22$
- podobně: očekávané četnosti pro 5 zbývajících buněk.

χ^2 test nezávislosti v kontingenční tabulce

- označme n_{ij} četnost v i-tém řádku a j-tém sloupci tabulky (celkem I řádků a J sloupců)
- označme n_{i+} (resp. n_{+j}) součet četností v i-tém řádku (resp. j-tém sloupci)
- očekávaná četnost v i-tém řádku a j-tém sloupci za hypotézy nezávislosti je

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

Testová statistika je mírou shody mezi n_{ij} a o_{ij}:

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - o_{ij})^2}{o_{ij}}$$

Pokud $\chi^2 \geq \chi^2_{(I-1)\cdot(J-1)}(1-\alpha)$, tak zamítáme hypotézu nezávislosti dvou znaků na hladině α.

▶ pro věrohodnost testu se požaduje, aby všechny očekávané četnosti byly větší než 5

Na hladině $\alpha = 0{,}05$ testujme hypotézu nezávislosti mezi pohlavím a strachem před matematikou z ▸ příkladu.

Napozorované (resp. očekávané) četnosti jsou:

pohlaví	**strach z matematiky** nízký	střední	vysoký	součet
muž	10 (7,84)	26 (20,16)	20 (28)	56
žena	4 (6,16)	10 (15,84)	30 (22)	44
součet	14	36	50	100

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(n_{ij}-o_{ij})^2}{o_{ij}} = \frac{(10-7{,}84)^2}{7{,}84} + \frac{(26-20{,}16)^2}{20{,}16} +$$

$$+ \frac{(20-28)^2}{28} + \frac{(4-6{,}16)^2}{6{,}16} + \frac{(10-15{,}84)^2}{15{,}84} + \frac{(30-22)^2}{22} = 10{,}39$$

Zjistíme dále, že $\chi^2 = 10{,}39 \geq \chi^2_{(I-1)\cdot(J-1)}(1-\alpha) = \chi^2_2(0{,}95) = 5{,}99$

Proto zamítáme hypotézu nezávislosti na hl. 5%. Existuje vztah mezi pohlavím a strachem z matematiky.

- Dá se říct, že strach z matem. je ovlivněn pohlavím.