Statistika

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

naposledy upraveno: 27. února 2020

Požadavek na udělení zápočtu (prezenční studium): v průběhu semestru budou znalosti prověřovány testy (2) z probírané látky. Termín každého testu bude dopředu oznámen cvičícím. Pro udělení zápočtu je nutné získat alespoň poloviční počet bodů z každého testu.

Požadavek na udělení zápočtu (kombinované studium): vypracování zápočtových prací, viz.
`http://147.230.193.199/~ms/stt.html`

Požadavky ke zkoušce (písemná i ústní část): znalost řešení úloh, vyložených pojmů a jejich vlastností v rozsahu daném přehledem přednášek.

Literatura

- KADEŘÁBEK J. *Statistika*. TUL, 2006.
- Hendl J. *Přehled statistických metod*. Praha: Portál, 2012.
- Anděl J. *Statistické metody*. Matfyzpress: Praha, 2007.
- CALDA E., DUPAČ V. *Matematika pro gymnázia : kombinatorika, pravděpodobnost a statistika*. Praha : Prometheus, 2004.
- Hebák P., Hustopecký J., Malá I. *Vícerozměrné statistické metody (2)*. Informatorium, Praha, 2005.
- Dalgaard P. *Introductory Statistics with R.*. 2008.
- ZVÁRA K., ŠŤĚPÁN J.: Pravděpodobnost a matematická statistika. Praha: Matfyzpress, 2002.

Literatura online

- http://147.230.193.199/~ms/stt.html
- http://www.studopory.vsb.cz
- http://mathonline.fme.vutbr.cz

Statistika

- **statistika** je jedním z oborů zabývajících se shromažďováním, zpracováním a analyzováním dat vznikajících při studiu tzv. **hromadných jevů**, což jsou jevy vyskytující se teprve u velkého souboru případů, ne jen u případů jednotlivých.
- **statistický soubor** je množina **statistických jednotek** (obyvatelé, obce, firmy,...), na nichž měříme (zjišťujeme) hodnoty **statistických znaků**(věk, počet obyvatel, obrat,...)
- zjištěnou hodnotu znaku vyjadřujeme ve vhodně zvoleném **měřítku** (stupnici).
- na jedné jednotce můžeme měřit několik znaků - to umožní vyšetřovat závislost (existuje souvislost mezi výškou a hmotností osob ve studované populaci?).

Ke studovanému datovému souboru lze přistoupit dvěma způsoby:

1. **Popisná statistika** - ze zjištěných dat chceme činit závěry pouze pro studovaný datový soubor (prošetřili jsme celou populaci, kterou chceme popsat)
2. **Matematická (inferenční) statistika** - Studovaný soubor chápeme jako **výběrový soubor** – množina prvků vybraných náhodně a nezávisle ze **základního souboru**, který je rozsáhlý (z důvodů časových, finančních, organizačních aj. nelze prozkoumat celý). Z hodnot proměnných zjištěných ve výběrovém souboru chceme činit závěry o základním souboru (v druhé půli semestru).

Typy měřítek

- **nula-jedničkové** (muž/žena, kuřák/nekuřák)
- **nominální** (rodinný stav, barva očí) - disjunktní kategorie, které nelze uspořádat
- **ordinální** (nejvyšší dosažené vzdělání, míra spokojenosti) - nominální měřítko s uspořádanými kategoriemi
- **intervalové** (teplota v Celsiové stupnici, rok narození) - možné hodnoty jsou číselně označeny, vzdálenost mezi sousedními hodnotami je konstatní
- **poměrové** (hmotnost, výška, počet obyvatel) - hodnoty jsou udávány v násobcích dohodnuté jednotky, nula znamená neexistenci měřené vlastnosti.
    - **Kvalitativní**: nula-jedničkové, nominální, ordinální
    - **Kvantitativní (spojité)**: intervalové, poměrové

Příklad - jednorozměrný

- jednorozměrná data (zajímá nás pouze jeden znak)

- zkoumáme IQ 62 žáků 8. tříd v jisté škole
- jak stručně popsat (zhodnotit), co mají data společného, nebo do jaké míry jsou odlišné?
- z naměřených hodnot zkoumaného znaku spočítáme charakteristiky (míry) některých jeho hromadných vlastností (charakteristiky polohy, variability, tvaru rozdělení, u vícerozměrných dat to budou i charakteristiky závislosti)
- charakteristiky (statistiky) jedním číslem vyjádří danou vlastnost

Příklad - naměřená data

naměřená data označme $x_1, x_2 \ldots, x_n$, nyní tedy $n = 62$.

107	141	105	111	112	96	103	140	136	92
92	72	123	140	112	127	120	106	117	92
107	108	117	141	109	109	106	113	112	119
138	109	80	111	86	111	120	96	103	112
104	103	125	101	132	113	108	106	97	121
134	84	108	84	129	116	107	112	128	133
96	94								

uspořádaný soubor označme $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$

72	80	84	84	86	92	92	92	94	96
96	96	97	101	103	103	103	104	105	106
106	106	107	107	107	108	108	108	109	109
109	111	111	111	112	112	112	112	112	113
113	116	117	117	119	120	120	121	123	125
127	128	129	132	133	134	136	138	140	140
141	141								

Třídní rozdělení četností

- Pokud se hodnoty často opakují, tak vytvoříme tzv. **četnostní tabulku**.
- Pokud jde o spojitou veličinu s velkým n (počtem naměřených hodnot), lze pro přehlednost obor hodnot dat rozdělit do M intervalů ohraničených body $a = a_0 < a_1 < a_2 < \ldots < a_{M-1} < a_M = b$.
- všechna pozorování z daného intervalu lze nahradit zástupnou hodnotou (zpravidla středem intervalu) x_i^*, $i = 1, \ldots, k$.
- nechť n_i označuje počet hodnot, které přísluší intervalu $\langle a_{i-1},\ a_i)$, $i = 1, \ldots, M$ – tzv. **třídní (absolutní) četnost** (jednotlivé intervaly se nazývají **třídy**).
- **kumulativní četnost** N_i udává počet hodnot v dané (i-té) třídě a třídách předcházejících
- čísla n_i/n označují **relativní četnost**.

Příklad - třídní rozdělení četností

Interval	x_i^*	absol. n_i	n_i/n	kumul. N_i	N_i/n
< 80	75	1	0.016	1	0.016
$\langle 80, 90)$	85	4	0.065	5	0.081
$\langle 90, 100)$	95	8	0.129	13	0.210
$\langle 100, 110)$	105	18	0.290	31	0.500
$\langle 110, 120)$	115	14	0.226	45	0.726
$\langle 120, 130)$	125	8	0.129	53	0.855
$\langle 130, 140)$	135	5	0.081	58	0.935
≥ 140	145	4	0.065	62	1.000

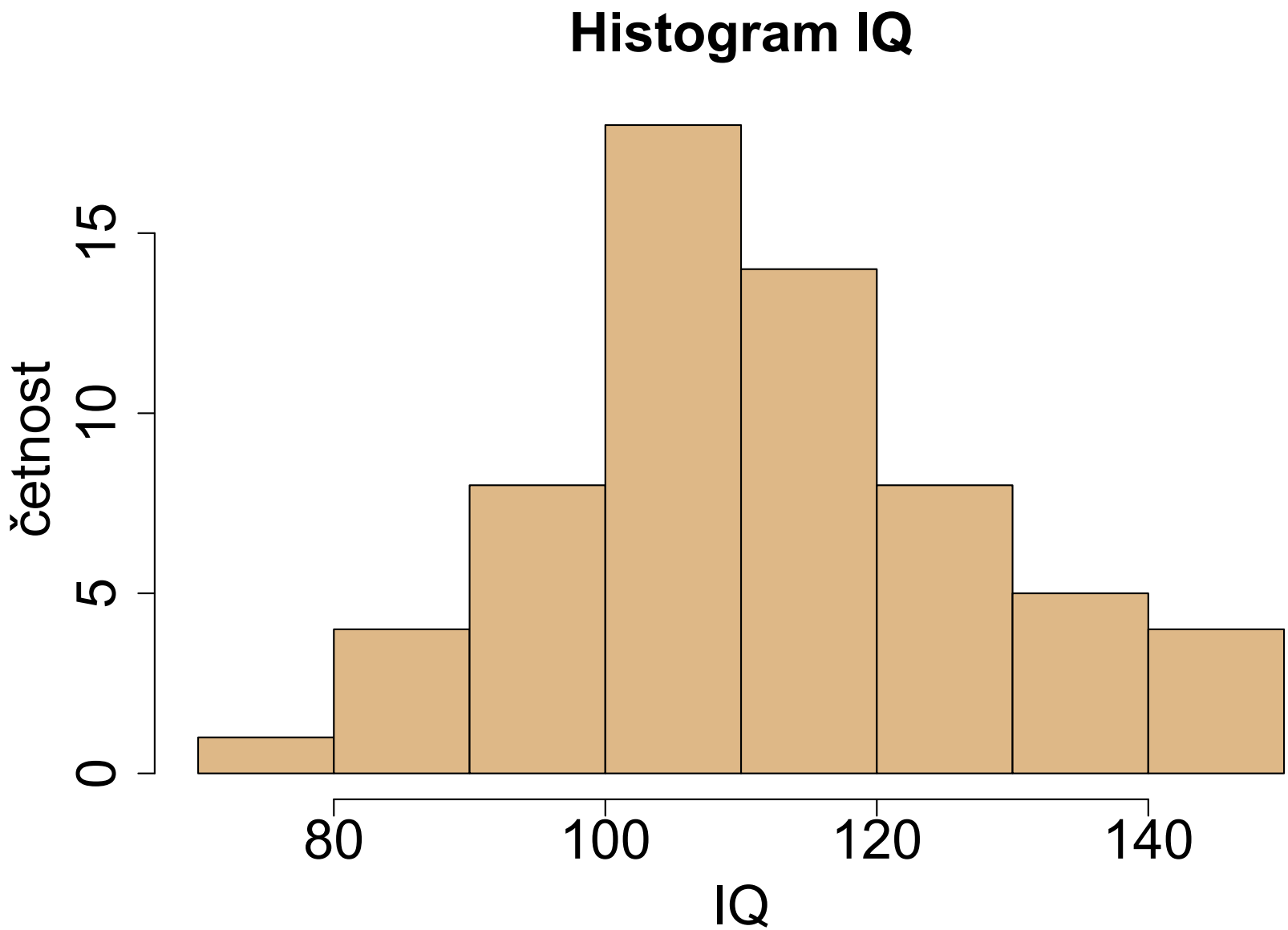
Histogram

- grafické znázornění třídních četností
- každému intervalu je přiřazen obdélníček tak, aby jeho plocha byla úměrná četnosti daného intervalu
- nejčastěji mají intervaly stejnou šířku (často vhodně zaokrouhlenou), pak výška obdélníků odpovídá četnostem.
- problém: volba počtu intervalů M
  lze použít např. tzv. Sturgesovo pravidlo:

$$M \approx 1 + 3.3 \log_{10}(n) \doteq 1 + \log_2(n)$$

- u našeho příkladu: $1 + \log_2(62) = 6.95$

Příklad - histogram



Charakteristiky polohy

- umožní charakterizovat úroveň číselné veličiny jedním číslem - ohodnocení, jak malých či velkých hodnot měření nabývají.
- pro charakteristiku polohy m souboru dat x by mělo platit, že se přirozeně mění se změnou měřítka, tj. že pro libovolné konstanty a, b:

$$m(a \cdot x + b) = a \cdot m(x) + b$$

- přičteme-li ke všem hodnotám konstantu b, tak se výsledná charakteristika zvětší o b
- vynásobíme-li každou hodnotu konstantou a, pak se výsledná charakteristika zvětší a-krát

Aritmetický průměr

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \ldots + x_n)$$

- u našeho příkladu: $\overline{x} = \frac{1}{62}(107 + 141 + \ldots + 94) = 111.0645$
- citlivý na hrubé chyby, odlehlá pozorování. Jen pro kvantitativní měřítka.
- z tabulky četností lze spočítat jako tzv. vážený průměr

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{M} n_i x_i^* = \frac{\sum_{i=1}^{M} n_i x_i^*}{\sum_{i=1}^{M} n_i} = \frac{1 \cdot 75 + 4 \cdot 85 + \ldots + 4 \cdot 145}{62} = 111.7742$$

- u nula-jedničkové veličiny: $\frac{\text{počet jedniček}}{\text{počet nul i jedniček}}$ = relativní četnost (procento) jedniček (pozorování s danou vlastností).
- u našeho příkladu $y_i = 0$ (i-tý žák je chlapec) , $y_i = 1$ (i-tý žák je dívka): $\overline{y} = \frac{32}{62} = 0.516$

Modus

- $\hat{x}$ - nejčastější hodnota
- má smysl určovat i pro nominální a ordinální měřítko
- není vždy jednoznačně určen
- u našeho příkladu:

72	80	84	84	86	92	92	92	94	96
96	96	97	101	103	103	103	104	105	106
106	106	107	107	107	108	108	108	109	109
109	111	111	111	112	112	112	112	112	113
113	116	117	117	119	120	120	121	123	125
127	128	129	132	133	134	136	138	140	140
141	141								

$$\hat{x} = 112$$

Medián

- $\tilde{x}$ - číslo, které dělí uspořádaný soubor na dvě stejně velké části. V uspořádaném výběru je uprostřed.

$$\tilde{x} = x_{(\frac{n+1}{2})} \qquad \text{pro } n \text{ liché}$$

$$\tilde{x} = \frac{1}{2}\left(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right) \qquad \text{pro } n \text{ sudé}$$

- robustní - není ovlivněn i velkými změnami několika hodnot. Lze často už i pro ordinální měřítko. U našeho příkladu:

72	80	84	84	86	92	92	92	94	96
96	96	97	101	103	103	103	104	105	106
106	106	107	107	107	108	108	108	109	109
109	111	111	111	112	112	112	112	112	113
113	116	117	117	119	120	120	121	123	125
127	128	129	132	133	134	136	138	140	140
141	141								

$$\tilde{x} = \frac{1}{2}\left(x_{(31)} + x_{(32)}\right) = 110$$

Kvantily: percentily, decily, kvartily

α-kvantil x_α ($\alpha \in (0,1)$) - dělí uspořádaný soubor na dvě části tak, že právě α-podíl těch nejmenších hodnot je menších než x_α

- $x_\alpha = x_{(\lceil \alpha n \rceil)}$,
  kde $\lceil a \rceil$ označuje a, pokud je to celé číslo, jinak nejbližší vyšší celé číslo.
- speciální případy kvantilů:

  percentily: $\alpha = 0.01, 0.02, \ldots, 0.99$
  decily: $\alpha = 0.1, 0.2, \ldots, 0.9$
  kvartily: $\alpha = 0.25, 0.5, 0.75$

  1. (dolní) kvartil značíme $Q_1 = x_{0.25}$
  3. (horní) kvartil značíme $Q_3 = x_{0.75}$
- medián je vlastně 50%-ní kvantil, 50-tý percentil, 5-tý decil a 2-hý kvartil

Příklad - kvantily

72	80	84	84	86	92	92	92	94	96
96	96	97	101	103	103	103	104	105	106
106	106	107	107	107	108	108	108	109	109
109	111	111	111	112	112	112	112	112	113
113	116	117	117	119	120	120	121	123	125
127	128	129	132	133	134	136	138	140	140
141	141								

- 1. kvartil $Q_1 = x_{0.25} = x_{(\lceil 0.25 \cdot 62 \rceil)} = x_{(\lceil 15.5 \rceil)} = x_{(16)} = 103$
- 3. kvartil $Q_3 = x_{0.75} = x_{(\lceil 0.75 \cdot 62 \rceil)} = x_{(\lceil 46.5 \rceil)} = x_{(47)} = 120$
- 1. decil (10%-ní kvantil)

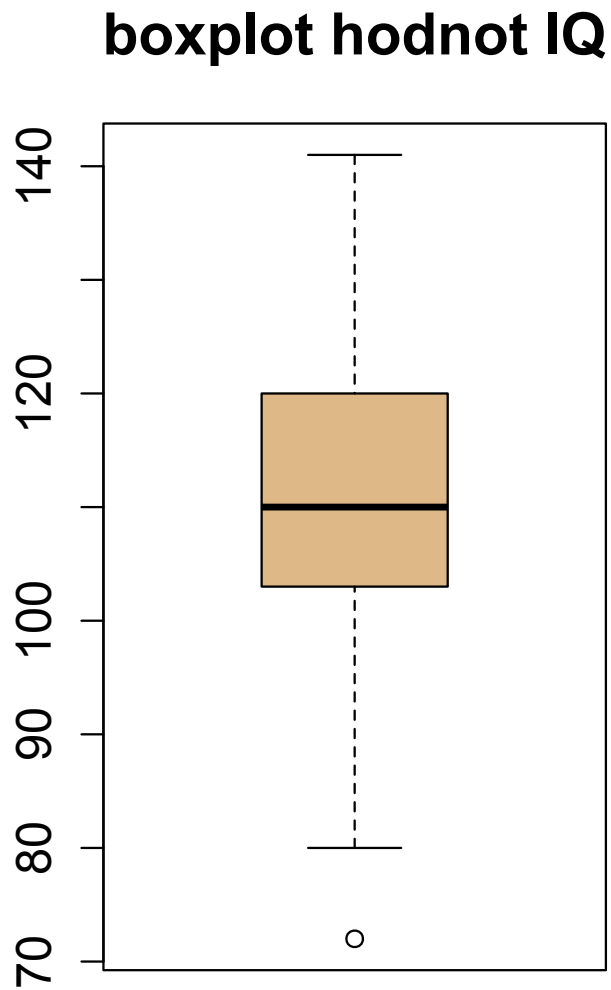
$$x_{0.1} = x_{(\lceil 0.1 \cdot 62 \rceil)} = x_{(\lceil 6.2 \rceil)} = x_{(7)} = 92$$

- 9. decil (90%-ní kvantil)

$$x_{0.9} = x_{(\lceil 0.9 \cdot 62 \rceil)} = x_{(\lceil 55.8 \rceil)} = x_{(56)} = 134$$

Boxplot

- česky **krabičkový diagram** - zobrazuje kvartily, medián, minimum, maximum a případně odlehlá pozorování (jsou od bližšího kvartilu dále než $1.5 \cdot (Q_3 - Q_1)$)
- u našeho příkladu: $Q_1 = 103, \tilde{x} = 110$, $Q_3 = 120$, 72 jako odlehlé pozorování



Charakteristiky variability

- měří rozptýlení, proměnlivost, nestejnost, variabilitu souboru dat.
- pro charakteristiku variability s souboru dat x by mělo platit, že pro libovolnou konstantu b a pro libovolnou kladnou konstantu $a > 0$:

$$s(a \cdot x + b) = a \cdot s(x)$$

- přičteme-li ke všem hodnotám konstantu b, tak se výsledná charakteristika nezmění
- vynásobíme-li každou hodnotu konstantou a, pak se výsledná charakteristika zvětší a-krát

Rozptyl (variance)

(populační) **rozptyl** $s_x^2 = var(x)$ - střední kvadratická odchylka od průměru

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right) = \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2\right) - \overline{x}^2$$

- u našeho příkladu:

$$s_x^2 = \frac{1}{62}\left[(107 - 111.0645)^2 + \ldots + (94 - 111.0645)^2\right] = 246.4797$$

- z naší tabulky četností:

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{M} n_i(x_i^* - \overline{x})^2 = \frac{1}{n}\left(\sum_{i=1}^{M} n_i x_i^{*2}\right) - \overline{x}^2$$
$$= (1 \cdot 75^2 + \ldots + 4 \cdot 145^2) - 111.7742^2 = 257.3361$$

- pro rozptyl platí $s_{a\cdot x+b}^2 = a^2 s_x^2$

Směrodatná odchylka, variační koeficient

(nevýběrová) **směrodatná odchylka**: odmocnina z rozptylu

$$s_x = \sqrt{s_x^2}$$

- stejný fyzikální rozměr jako původní data

variační koeficient:

$$v = \frac{s_x}{\overline{x}}$$

- definován pouze pro kladné hodnoty $x_1, \ldots, x_n > 0$
- nezávisí na volbě měřítka, lze použít na porovnání různých souborů

u našich dat: $s_x = \sqrt{246.4797} = 15.70$

$$v = \frac{15.70}{111.0645} = 0.1414$$

rozpětí: rozdíl maxima a minima souboru

$$R = x_{(n)} - x_{(1)}$$

mezikvartilové rozpětí: rozdíl třetího a prvního kvartilu

$$R_M = Q_3 - Q_1 = x_{0.75} - x_{0.25}$$

střední odchylka: průměr absolutních odchylek od mediánu

(nebo průměru)

$$d = \frac{1}{n} \sum_{i=1}^{n} |x_i - \tilde{x}|$$

u našich dat: $R = 141 - 72 = 69$ $R_M = 120 - 103 = 17$

$$d = \tfrac{1}{62}(|107 - 110| + \ldots + |94 - 110|) = 12.03$$

Míra variability pro kategoriální znak

Entropie

$$h = -\sum_{i=1}^{r} \frac{n_i}{n} \log\left(\frac{n_i}{n}\right)$$

Charakteristiky tvaru

- měří tvar rozdělení hodnot v souboru dat.
- pro charakteristiku tvaru γ souboru dat x by mělo platit, že pro libovolnou konstantu b a pro libovolnou kladnou konstantu $a > 0$:

$$\gamma(a \cdot x + b) = \gamma(x)$$

- vynásobíme-li každou hodnotu konstantou a nebo přičteme-li ke všem hodnotám konstantu b, tak se výsledná charakteristika nezmění.
- proto je počítáme ze standardizovaných hodnot

$$\frac{x_i - \overline{x}}{s_x}.$$

koeficient šikmosti: průměr třetích mocnin standardizovaných hodnot

$$g_1 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s_x}\right)^3$$

- měří zešikmení rozdělení (symetrické ≈ 0, pravý chvost > 0, levý chvost < 0)

koeficient špičatosti: průměr čtvrtých mocnin standardizovaných hodnot

$$g_2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s_x}\right)^4 - 3$$

- měří "špičatost" rozdělení (koncentrace kolem středu a na chvostech > 0, "ploché" rozdělení < 0)

Lze použít pro porovnání (ověření) s normálním rozdělením, pro které $g_1 \doteq g_2 \doteq 0$.

u našich dat: $g_1 = 0.0159$ $g_2 = -0.241$

Příklad - vícerozměrný

- vícerozměrná data (zajímá nás více znaků)

- zjištěno IQ, pohlaví, průměrná známka v pololetí v 7. a 8. třídě 62 žáků
- jak zhodnotit vztah (závislost) mezi jednotlivými znaky?
- vypočtením vhodné statistiky (čísla) nebo grafickým zobrazením

Příklad - naměřená vícerozměrná data

Dívka	1	0	0	1	0	1	0	0	1	1
Zn7	1	1	3.15	1.62	2.69	1.92	2.38	1	1.4	1.46
Zn8	1	1	3	1.73	2.09	2.09	2.55	1	1.9	1.45
IQ	107	141	105	111	112	96	103	140	136	92

Dívka	1	0	0	0	1	0	1	1	1	0
Zn7	1.85	3.15	1.15	1	1.69	1.6	1.62	1.38	1.7	3.23
Zn8	1.45	3.18	1.18	1	1.91	1.72	1.63	1.36	1.9	3.36
IQ	92	72	123	140	112	127	120	106	117	92

Dívka	0	0	1	1	1	1	0	1	0	1
Zn7	2.07	1.84	1.2	1.31	1.4	1.53	1.84	1	1.3	1.4
Zn8	2.45	1.9	1.36	1.45	1.73	1.6	1.54	1	1.45	1.82
IQ	107	108	117	141	109	109	106	113	112	119

Dívka	0	0	1	1	0	1	0	1	0	0
Zn7	1	2.92	2.23	1.69	2.61	1.07	1.46	2.15	1.69	1.38
Zn8	1	2.82	2.45	1.54	2.54	1	1.36	1.9	1.82	1.18
IQ	138	109	80	111	86	111	120	96	103	112

vícerozměrná data - pokračování

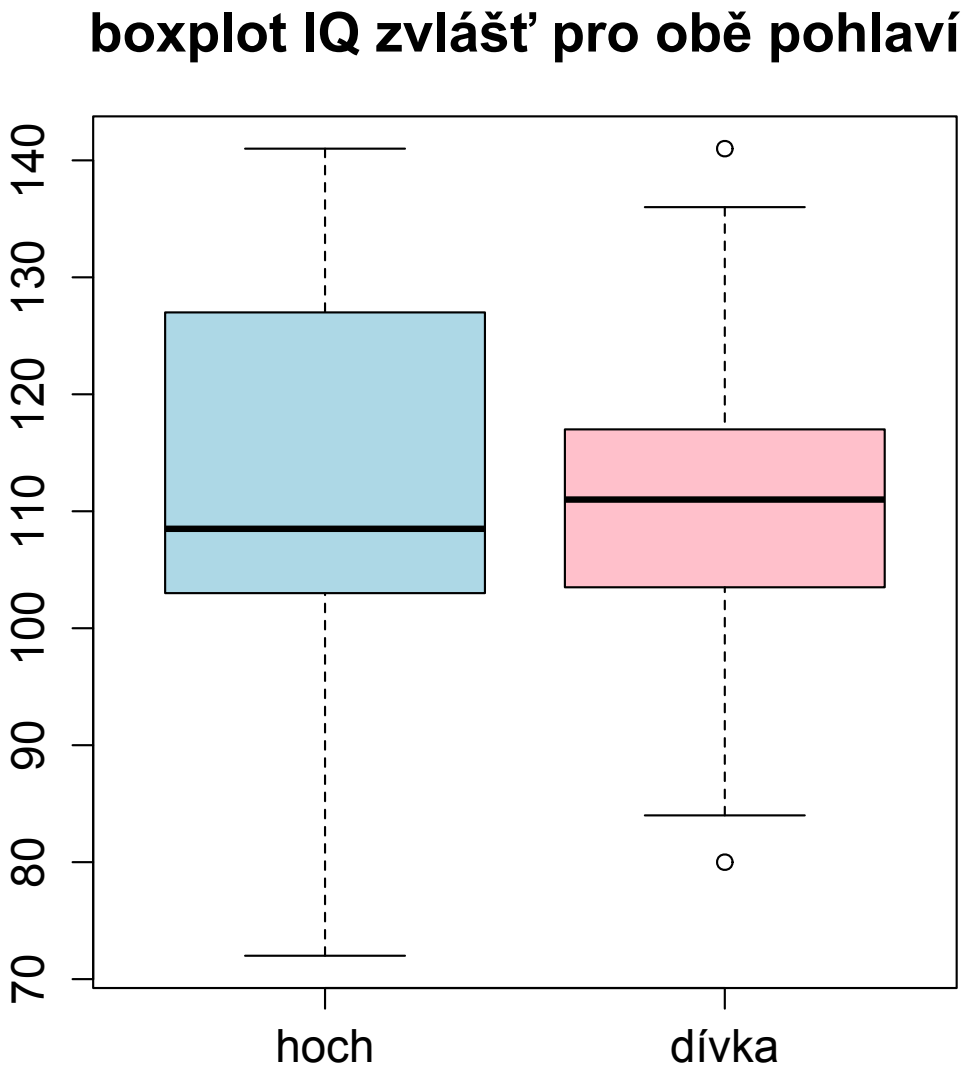
Dívka	1	1	1	0	0	1	0	1	1	0
Zn7	1.46	1.6	1.07	1.3	2.08	2	1.69	1.4	2.23	1.6
Zn8	1.54	1.63	1	1.27	1.54	2.09	1.91	1.45	2	1.81
IQ	104	103	125	101	132	113	108	106	97	121

Dívka	1	0	1	1	0	1	0	1	1	0
Zn7	1.07	3.13	1.84	1.8	1	1.92	2.2	1.53	1.3	1
Zn8	1.27	3.27	1.82	1.63	1	1.9	2.25	1.54	1.45	1.18
IQ	134	84	108	84	129	116	107	112	128	133

Dívka	0	0
Zn7	2.85	2.61
Zn8	2.91	2.81
IQ	96	94

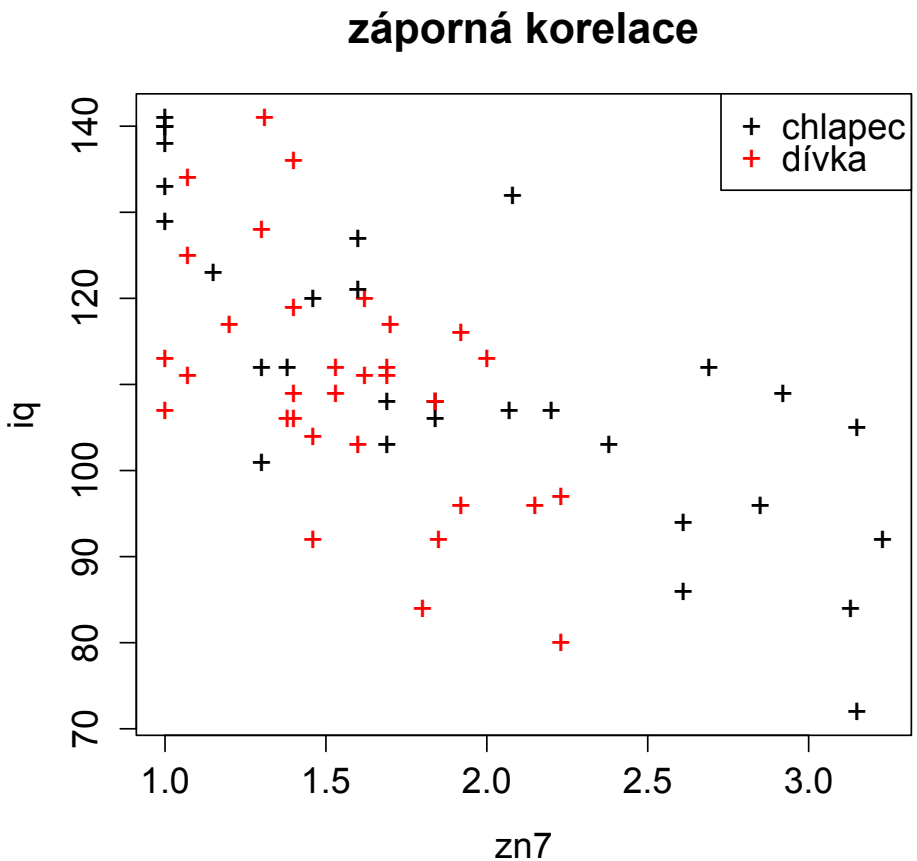
Grafické znázornění závislosti

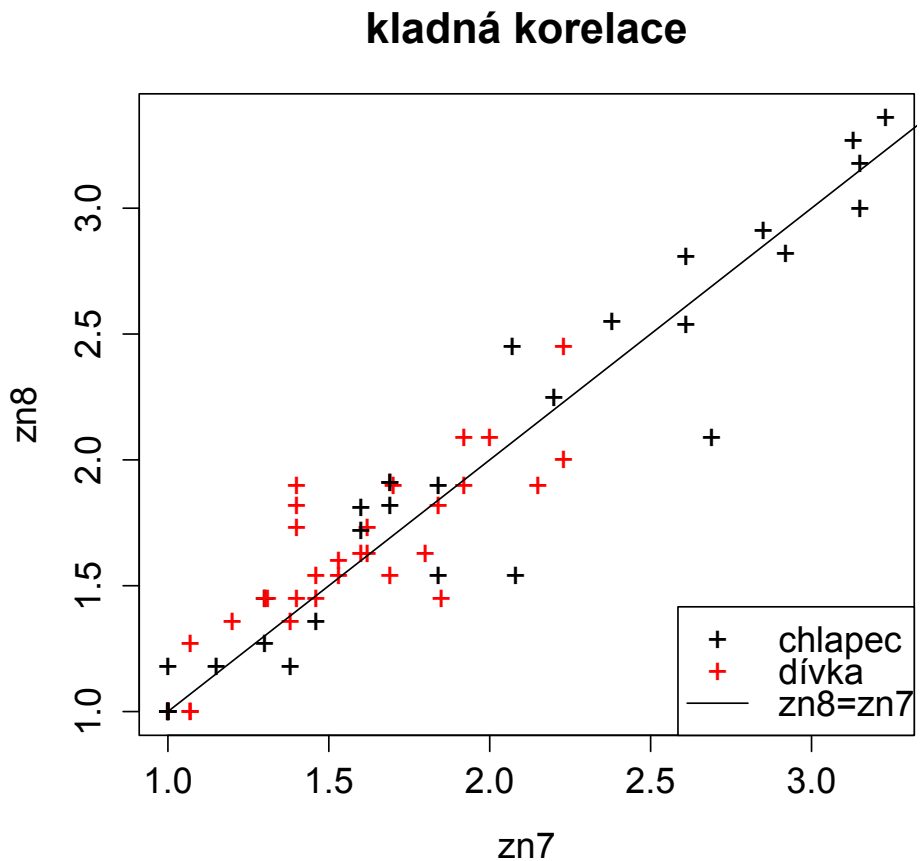
- Záleží na typu měřítka
- pro závislost kvantit. znaku na kvalitativním lze nakreslit boxplot/histogram pro každou kategorii kvalit. znaku
- zobrazení závislosti IQ na pohlaví
- $\overline{x}_{hoch} = 112.0$
  $\overline{x}_{divka} = 110.2$



Grafické znázornění závislosti - 2

Rozptylový diagram: závislost dvou kvantitativních znaků





Charakteristiky závislosti

dva znaky na každé jednotce, tj. máme $(x_1, y_1), \ldots, (x_n, y_n)$

kovariance: měří směr závislosti, ovlivněna změnou měřítka

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i\right) - \overline{x}\cdot\overline{y},$$

- Platí $s_{xx} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = s_x^2, \quad s_{yy} = s_y^2$

(Pearsonův) korelační koeficient: normovaná kovariance, měří směr i velikost závislosti

$$r_{x,y} = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}} = \frac{s_{xy}}{s_x s_y} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s_x}\right)\cdot\left(\frac{y_i - \overline{y}}{s_y}\right)$$

- u našich dat pro znaky IQ a zn7:

$$r_{IQ,zn7} = \frac{-6.2876}{15.6997 \cdot 0.6106} = -0.6559$$

Korelační koeficient

- měří směr a míru lineární závislosti
- nabývá jen hodnot z intervalu $\langle -1, 1 \rangle$
- $r_{x,y} \approx 0$ (znaky x a y vzájemně nezávislé)
- $r_{x,y}$ blízko 1 (kladná závislost: s rostoucím x znak y v průměru roste)
- $r_{x,y}$ blízko -1 (záporná závislost: s rostoucím x znak y v průměru klesá)

U našich dat lze spočítat pro každou dvojici znaků dívka, iq, zn7, zn8: tzv. **korelační matice**

	dívka	iq	zn7	zn8
dívka	1.0000	-0.0597	-0.3054	-0.2661
iq	-0.0597	1.0000	-0.6559	-0.6236
zn7	-0.3054	-0.6559	1.0000	0.9481
zn8	-0.2661	-0.6236	0.9481	1.0000

Regresní přímka - metoda nejmenších čtverců

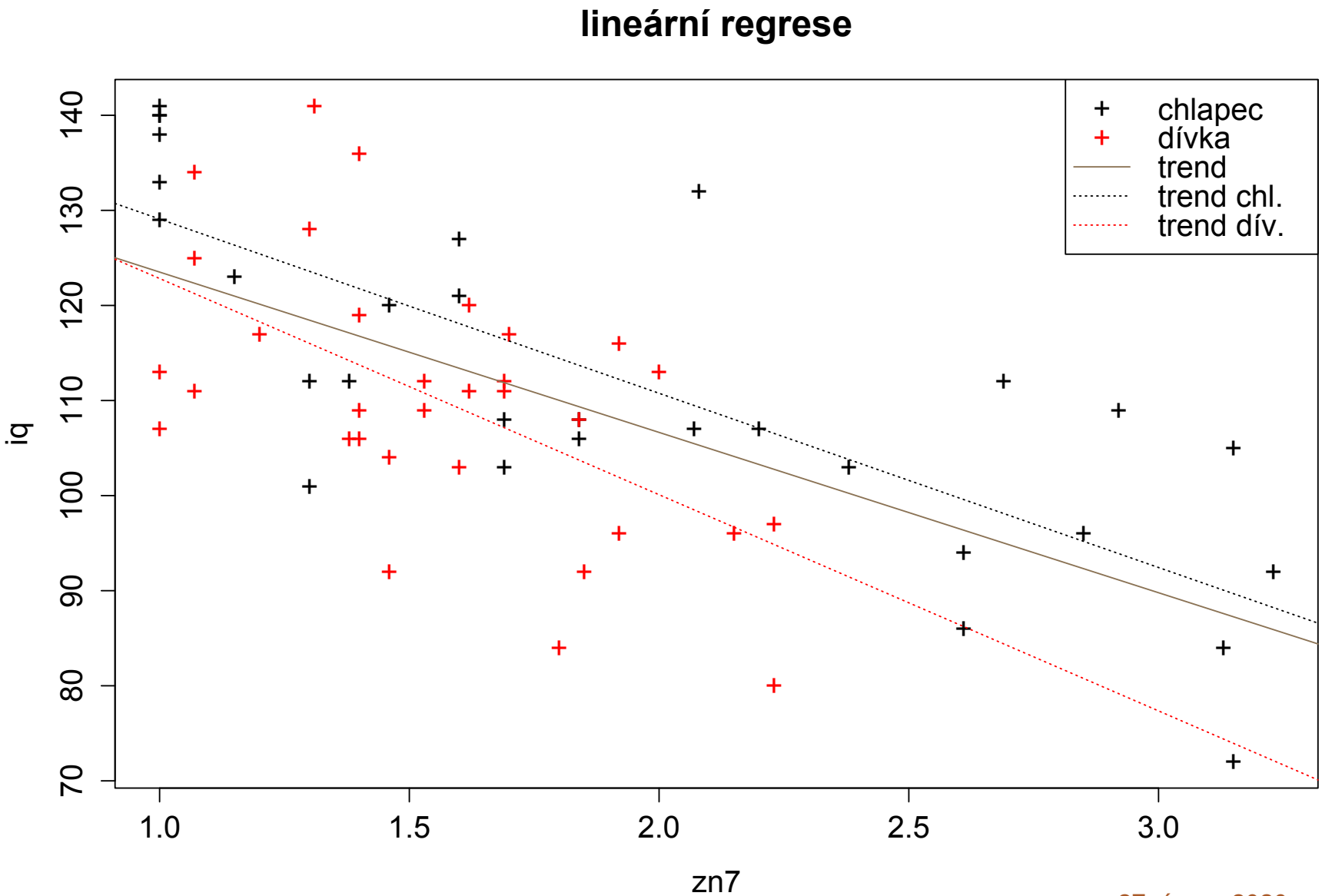
- Máme sadu dvojic $(x_i, y_i), i = 1, \ldots, n$. Chceme z daných hodnot znaku x odhadnout hodnoty znaku y. Předpokládáme lineární závislost y na x, tj. že přibližně platí

$$y \doteq a + b \cdot x$$

- Parametry a a b regresní přímky se odhadnou **metodou nejmenších čtverců**, tj. hledáme hodnoty, pro které je výraz $\sum_{i=1}^{n}(y_i - (a + b \cdot x_i))^2$ minimální. Řešením jsou:

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i \cdot y_i) - n \cdot \overline{x} \cdot \overline{y}}{\sum_{i=1}^{n} x_i^2 - n \cdot \overline{x}^2} = \frac{S_{xy}}{S_x^2} \qquad \hat{a} = \overline{y} - \hat{b} \cdot \overline{x}$$

Regresní přímka: znázornění lineární závislosti dvou kvantitativních znaků



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

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

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 (reálná funkce def. na Ω)

Rozdělení náhodné veličiny: udává jakých hodnot s jakou pravděpodobností veličina nabývá (množinová funkce: každé podmožině R přiřadí pravděpodobnost)

- 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

Vlastnosti distribuční funkce $F_X(x)$:

- neklesající, zleva spojitá
- $\lim_{x\to-\infty} F_X(x) = 0, \qquad \lim_{x\to\infty} F_X(x) = 1$

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): náhodná veličina X má spojité rozdělení, jestliže existuje tzv. **hustota** $f_X(x)$, pro kterou platí

$$F_X(x) = P(X < x) = \int_{-\infty}^{x} f_X(t)\,dt$$

- $f_X(x) = F'_X(x)$ pro každý x bod spojitosti $f_X(x)$
- $f_X(x) \geq 0\ \forall x, \quad P(a < X < b) = \int_a^b f_X(x)\,dx, \quad \int_{-\infty}^{\infty} f_X(x)\,dx = 1$
- $P(X = a) = 0$ pro každé $a \in R$ (představa: teoretický protějšek hranice histogramu pro délku intervalů jdoucích k nule)

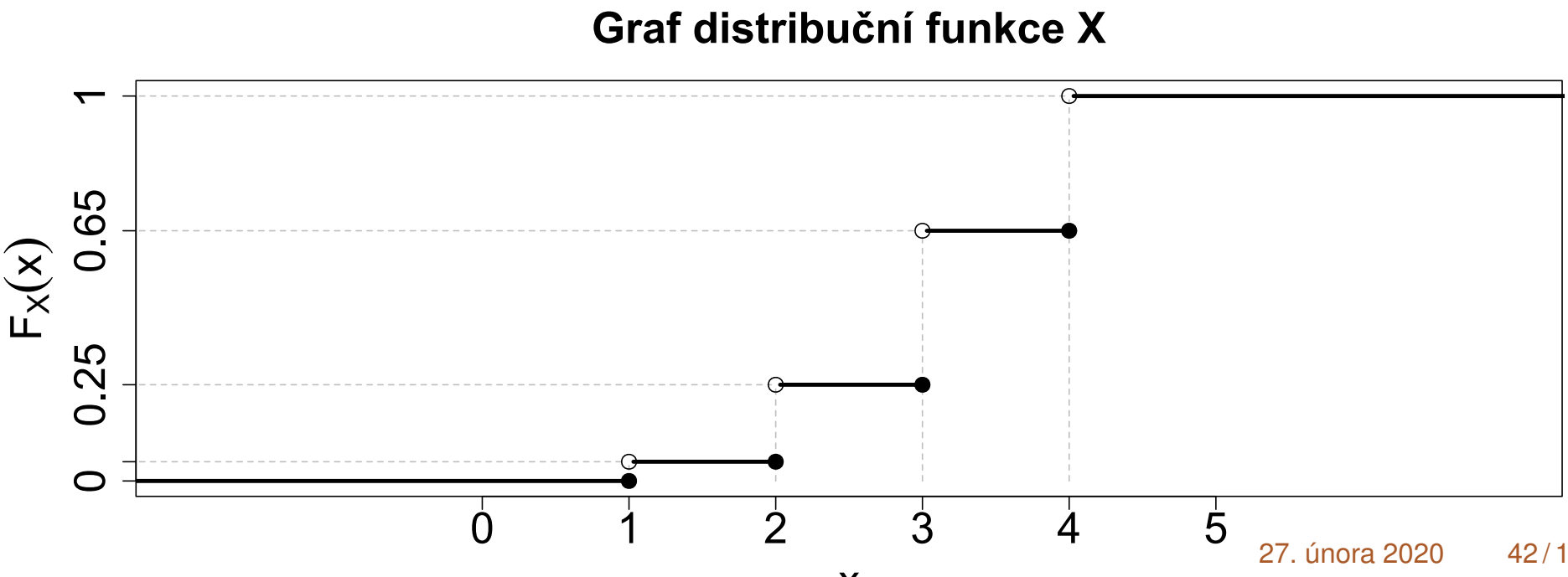
Příklad 1

(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.

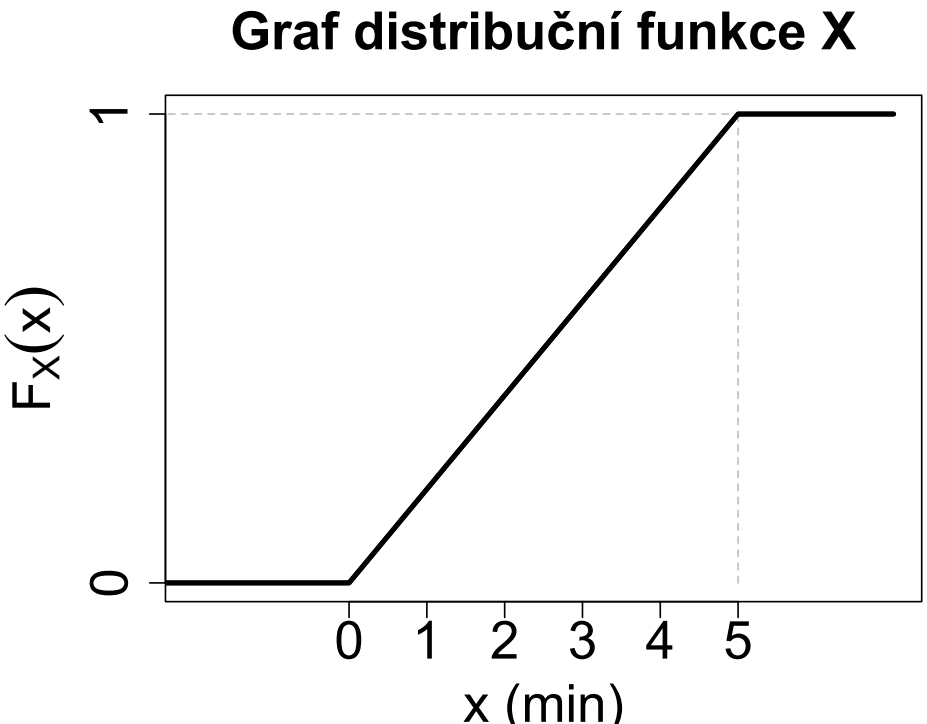
- $F_X(3) = P(X < 3) = P(X = 1) + P(X = 2) = 0{,}05 + 0{,}2 = 0{,}25$
- nutno určit $F_X(x) = P(X < x)$ pro každé $x \in R$

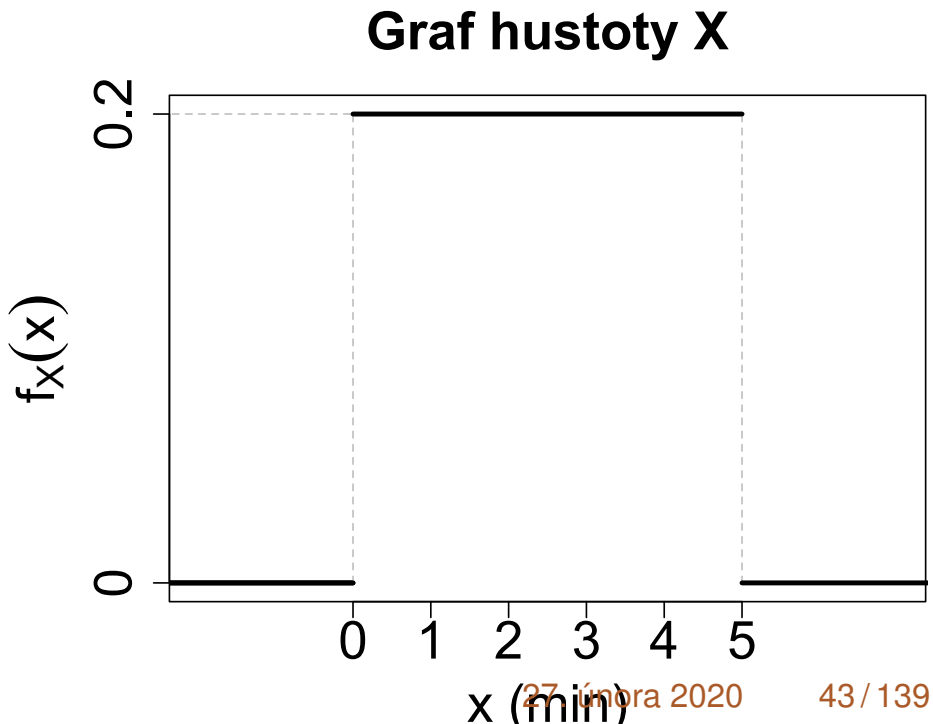


Příklad 2

(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 $F_X(x) = P(X < x) = \frac{x}{5}$, tedy $f_X(x) = \frac{1}{5}$





Výpočet pravděpodobnosti 1

u ▸ Př. 1 (diskrétní rozdělení): Určete pravděpodobnost, že studentova známka

- bude lepší než 4 ale ne lepší než 2:

$$P(2 \leq X < 4) \overset{\text{z distr. f.}}{=\!=} P(X < 4) - P(X < 2) = F_X(4) - F_X(2) = 0{,}65 - 0{,}05 = 0{,}6$$
$$\overset{\text{z tab. pravd.}}{=\!=} P(X = 3) + P(X = 2) = 0{,}4 + 0{,}2 = 0{,}6$$

- nebude lepší než 3:

$$P(X \geq 3) \overset{\text{z distr. f.}}{=\!=} 1 - P(X < 3) = 1 - F_X(3) = 1 - 0{,}25 = 0{,}75$$
$$\overset{\text{z tab. pravd.}}{=\!=} P(X = 3) + P(X = 4) = 0{,}4 + 0{,}35 = 0{,}75$$

- bude rovna 4:

$$P(X = 4) \overset{\text{z tab. pravd.}}{=\!=} 0{,}35$$ je to výška skoku distr. funkce v bodě 4

Výpočet pravděpodobnosti 2

u ▸ Př. 2 (spojité rozdělení): Určete pravděpodobnost, že budeme čekat

- méně než 4 ale více než 2 minuty:

$$P(2 < X < 4) \overset{P(X=2)=0}{=\!=} P(X < 4) - P(X < 2) \overset{\text{z distr. f.}}{=\!=} F_X(4) - F_X(2) = \frac{4}{5} - \frac{2}{5} = \frac{2}{5}$$

$$\overset{\text{z hustoty}}{=\!=} \int_2^4 f_X(x)\,dx = \int_2^4 \frac{1}{5}\,dx = \frac{2}{5}$$

- více než 4 minuty:

$$P(X > 4) \overset{\text{z distr. f.}}{=\!=} 1 - P(X < 4) = 1 - F_X(4) = 1 - \frac{4}{5} = \frac{1}{5}$$

$$\overset{\text{z hustoty}}{=\!=} \int_4^\infty f_X(x)\,dx = \int_4^5 \frac{1}{5}\,dx + \int_5^\infty 0\,dx = \frac{1}{5}$$

- přesně 4 minuty:

$$P(X = 4) = \int_4^4 \frac{1}{5}\,dx = 0$$ výška skoku distr. funkce v bodě 4 je rovna 0

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)

Střední hodnota funkce $Y = g(X)$ náhodné veličiny X - hodnota, kolem které se kumulují hodnoty náhodné veličiny $g(X)$

- pro diskrétní rozdělení: vážený průměr funkčních hodnot

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

- pro spojité rozdělení: integrál všech možných hodnot $g(x)$, váhovou funkcí je hustota

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

u ▸ Př. 1: předpokládejme, že nás nezajímá střední (očekávaná) známka, ale střední (očekávaná) výše školného, které, dejme tomu, je vázáno na známku funkčním předpisem $g(x) = 1000 \cdot x^2$ Kč
$Eg(X) = 1000 \cdot 1^2 \cdot 0{,}05 + 1000 \cdot 2^2 \cdot 0{,}2 + 1000 \cdot 3^2 \cdot 0{,}4 + 1000 \cdot 4^2 \cdot 0{,}35 =$ 10 050 Kč

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

Nezávislost náhodných veličin

Podobně jako u náh. jevů lze hovořit o nezávislosti náhodných veličin, a to tehdy, když výsledky jedné veličiny neovlivní pravděpodobnost výsledku druhé.

Říkáme, že náh. veličiny X a Y jsou **nezávislé**, jestliže pro každé $x, y \in R$ platí

$$P(X < x, Y < y) = P(X < x) \cdot P(Y < y)$$

speciálně pro diskrétní rozdělení lze nahradit podmínkou, že pro všechna i, j platí

$$P(X = x_i, Y = y_j) = P(X = x_i) \cdot P(Y = y_j)$$

Vlastnosti střední hodnoty a rozptylu

Nechť $a, b \in R$ a X je libovolná náhodná veličina, potom

1) $E(a + b \cdot X) = a + b \cdot EX$
2) $var\,(a + b \cdot X) = b^2 \cdot var\,X$
3) $var\,X \geq 0$
4) $var\,X = EX^2 - (EX)^2$
5) $E(X + Y) = EX + EY$
6) pro nezávislé X, Y: $var\,(X + Y) = var\,X + var\,Y$

Důkaz: 1), 2), 4) a 5) plyne z linearity sumace resp. integrálu:
ad 1) např. pro spojité rozdělení:

$$E(a + b \cdot X) = \int_{-\infty}^{\infty} (a + b \cdot x) \cdot f_X(x)\,dx \overset{\text{lin. int.}}{=\!=\!=}$$

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

ad 2):

$$var(a + b \cdot X) = E[a + b \cdot X - E(a + b \cdot X)]^2 \overset{1)}{=} E[a + b \cdot X - (a + b \cdot EX)]^2 =$$

$$= E[b \cdot (X - EX)]^2 = b^2 \cdot var\,X$$

ad 3): plyne z faktu, že $var\,X$ je integrál (suma) nezáporné funkce (hodnot)
ad 4): podobně jako 1) a 2) (dom. cvičení), ad 5) a 6): bez důkazu

Kvantily rozdělení

Nechť náhodná veličina X má distribuční funkci F_X. Potom funkce F_X^{-1} daná vztahem

$$F_X^{-1}(\alpha) = \inf\{x \in R : F_X(x) \geq \alpha\} \qquad 0 < \alpha < 1,$$

se nazývá **kvantilová funkce**

Infimum množiny A, inf *A: je maximum z těch prvků, které jsou menší nebo rovny všem prvkům v A.*

- Hodnotám funkce $F_X^{-1}(\alpha)$ říká α**-kvantil** (nebo $100 \cdot \alpha$ %-ní kvantil)
- V případě spojitého rozdělení je to přímo inverzní funkce a platí, že

  $$P(X < F_X^{-1}(\alpha)) = \alpha$$

  α-kvantil je tedy taková hodnota, pod kterou je veličina s pravděpodobností α
- speciálně $F_X^{-1}(0{,}5)$ se nazývá **medián** rozdělení.

u ▸ Př. 1: $F_X^{-1}(0{,}5) = \inf\{x : F_X(x) \geq 0{,}5\} \overset{\text{z grafu } F_X}{=\!=\!=} 3$

u ▸ Př. 2: $F_X^{-1}(0{,}5) = \inf\{x : F_X(x) \geq 0{,}5\} \overset{\text{z inv. funkce k } F_X}{=\!=\!=\!=\!=} 5 \cdot 0{,}5 = 2{,}5$

je 50 %-ní pravděpodobnost, že budu čekat méně než 2,5 minuty

Alternativní rozdělení

Příklad: na otázku je správná právě jedna z odpovědí a), b), c), d). Jaká je pravděpodobnost, že odpovíme správně, pokud tipujeme náhodně?

Položme $X = 1$ (nebo 0), jestli odpovíme správně (resp. nesprávně)

$$P(X = 1) = 1/4, \quad P(X = 0) = 3/4$$

X má tzv. alternativní rozdělení s parametrem $p = 1/4$

Obecně: jde o diskrétní rozdělení, také se nazývá nula-jedničkové

X má **alternativní rozdělení** s param. p, ozn. $X \sim Alt(p)$, pokud

$$P(X = 1) = p, \qquad P(X = 0) = 1 - p, \qquad 0 < p < 1$$

- střední hodnota $EX = 1 \cdot P(X = 1) + 0 \cdot P(X = 0) = p$
- rozptyl

$$var\, X = EX^2 - (EX)^2 = 1^2 \cdot P(X = 1) + 0^2 \cdot P(X = 0) - p^2 =$$
$$= p - p^2 = p \cdot (1 - p)$$

u Př.: $EX = \frac{1}{4}$ $\qquad var\, X = \frac{1}{4} \cdot \frac{3}{4} = \frac{3}{16}$

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 $X \sim Bi(n, p)$
- dá se chápat jako součet n nezávislých náh. vel. $\sim Alt(p)$
- střední hodnota $EX = \sum_{i=0}^{n} i \cdot \binom{n}{i} \cdot p^i \cdot (1-p)^{n-i} = n \cdot p$
- rozptyl

$$var\, X = EX^2 - (EX)^2 = \sum_{i=0}^{n} i^2 \cdot \binom{n}{i} \cdot p^i \cdot (1-p)^{n-i} - (n \cdot p)^2 = n \cdot p \cdot (1-p)$$

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

Příklad

(geometrické rozdělení): Na každou z otázek je správná právě jedna z odpovědí a), b), c), d). Postupně na otázky odpovídáme a to tak, že tipujeme náhodně, dokud na nějakou správně neodpovíme. Jaká je pravděpodobnost, že poprvé odpovíme správně až na třetí otázku? (tzn. před první úspěšně zodpovězenou otázkou přesně dvě otázky nezodpovíme správně)

- ozn. X počet špatných odpovědí před první správnou
- na každou odpovíme správně s pravděpodobností $p = 1/4$
- odpovědi na jednotlivé otázky jsou nezávislé
- musíme se netrefit v prvních dvou otázkách a ve třetí odpovědet správně

Tedy hledaná pravděpodobnost je

$$(1-p)^2 \cdot p = (3/4)^2 \cdot (1/4) \doteq 0{,}14$$

Geometrické rozdělení

Opakujeme nezávisle stejný náhodný pokus. Sledujeme počet pokusů X (v nichž nenastane daný jev) před prvním pokusem, ve kterém daný jev nastane. Platí přitom, že pravděpodobnost tohoto jevu ve všech pokusech stejná, rovna p. Zajímá nás tedy počet "neúspěšných" pokusů X před prvním "úspěchem". X může nabývat pouze hodnot $0, 1, \ldots$ a má rozdělení dané pravděpodobnostmi

$$P(X = i) = (1 - p)^i \cdot p, \quad i = 0, 1, \ldots$$

kde $0 < p < 1$

- říkáme, že X má **geometrické rozdělení** s parametrem p
- zkráceně píšeme $X \sim Ge(p)$
- střední hodnota $EX = \sum_{i=0}^{\infty} i \cdot (1 - p)^i \cdot p = \frac{1-p}{p}$
- rozptyl

$$var\, X = EX^2 - (EX)^2 = \sum_{i=0}^{\infty} i^2 \cdot (1 - p)^i \cdot p - \left(\frac{1 - p}{p}\right)^2 = \frac{1 - p}{p^2}$$

u Př.: $X \sim Ge(1/4)$ $EX = 3$ $var\, X = \frac{3}{4}/(\frac{1}{4})^2 = 12$

Příklad

(hypergeometrické rozdělení): V hrnci je 30 sladkých knedlíků, z nichž 10 je jahodových a 20 švestkových. Z hrnce náhodně vybereme 6 knedlíků. Jaká je pravděp., že méně než dva z nich budou jahodové?

- ozn. X počet jahod. knedlíků mezi vybranými
- vybíráme "bez vracení", tj. jednotlivé "tahy" nejsou nezávislé
- hledáme $P(X < 2) = P(X = 0) + P(X = 1)$
- $P(X = 0)$ nebo $P(X = 1)$ je možné spočítat z klasické def. pravd.

$$P(X = 0) = \frac{\binom{10}{0} \cdot \binom{20}{6}}{\binom{30}{6}} \doteq 0{,}065 \quad \text{resp. } P(X = 1) = \frac{\binom{10}{1} \cdot \binom{20}{5}}{\binom{30}{6}} \doteq 0{,}261$$

hledaná pravděpodobnost je tedy

$$P(X < 2) \doteq 0{,}065 + 0{,}261 = 0{,}326$$

Hypergeometrické rozdělení

Toto rozdělení je možné popsat následující situací. Uvažujme množinu, která obsahuje N objektů, z nichž M má jistou vlastnost. Vybereme náhodně z této množiny n objektů. Potom X označuje počet vybraných objektů mající uvažovanou vlastnost. X může nabývat pouze celočíselných hodnot s pravděpodobnostmi

$$P(X = i) = \frac{\binom{M}{i} \cdot \binom{N-M}{n-i}}{\binom{N}{n}}, \quad \text{pro} \quad \max(0, M + n - N) \leq i \leq \min(M, n)$$

- říkáme, že X má **hypergeometrické rozdělení** s parametry N, M a n
- zkráceně píšeme $X \sim Hg(N, M, n)$
- střední hodnota $EX = \sum_i i \cdot \frac{\binom{M}{i} \cdot \binom{N-M}{n-i}}{\binom{N}{n}} = \frac{n \cdot M}{N}$
- rozptyl $var\, X = \frac{n \cdot M \cdot (N-M)}{N^2} \cdot \left(1 - \frac{n-1}{N-1}\right)$

u Př.:

$X \sim Hg(N = 30, M = 10, n = 6)$ $EX = \frac{6 \cdot 10}{30} = 2$ $var\, X \doteq 1{,}103$

Poissonovo rozdělení

Nechť X je náhodná veličina nabývající pouze hodnot $i = 0, 1, 2, \ldots$ a to s pravděpodobnostmi

$$P(X = i) = \frac{\lambda^i}{i!} e^{-\lambda}, \qquad i = 0, 1, 2, \ldots$$

kde $\lambda > 0$ je dané číslo.

- říkáme, že X má **Poissonovo rozdělení** s parametrem λ
- zkráceně píšeme $X \sim Po(\lambda)$
- střední hodnota a rozptyl $\quad EX = var\, X = \lambda$

Nechť $Y_n \sim Bi(n, p)$, kde n je velké a p malé tak, že $n \cdot p = \lambda$.
Pak platí $\lim_{n\to\infty} P(Y_n = i) = P(X = i)$.
Tj. pro n velké a p malé lze rozdělení $Bi(n, p)$ nahradit rozdělením $Po(n \cdot p)$
Např. pro $\quad Y \sim Bi(20, 0{,}1) \quad$ a $\quad X \sim Po(20 \cdot 0{,}1) = Po(2)$
je $\quad P(Y = 3) \doteq 0.19 \quad$ a $\quad P(X = 3) \doteq 0.18$

▶ Nejčastěji se používá pro popis pravděpodobnosti počtu událostí v nějakém časovém intervalu, pokud události nastávají v náhodných okamžicích a nezávisle s intenzitou λ (počet telefonních hovorů, dopravních nehod, příchodů zákazníků do obchodu apod.)

Příklad

(Poissonovo rozdělení): Během pracovního dne do call centra přijde v průměru 30 hovorů za hodinu. Jaká je pravděpodobnost, že během jedné minuty přijde více než jeden hovor?

- ozn. X počet příchozích hovorů za 1 min.
- X značí počet událostí za časový interval, tedy $X \sim Po(\lambda)$
- λ neznáme, ale víme, že $EX = \lambda$
- střední počet hovorů za 1 minutu $EX = \lambda$ můžeme odhadnout číslem $\frac{30}{60} = 0{,}5$
- položíme $\lambda = 0{,}5$ a spočítáme

$$P(X > 1) = 1 - [P(X = 0) + P(X = 1)] =$$

$$= 1 - \left[\frac{0{,}5^0}{0!}e^{-0{,}5} + \frac{0{,}5^1}{1!}e^{-0{,}5}\right] \doteq 1 - 0{,}606 - 0{,}303 \doteq 0{,}09$$

Rovnoměrné rozdělení

V ▸ příkladu šlo o rovnoměrné rozdělení na intervalu $(0, 5)$

Nechť X je náhodná veličina se spojitým rozdělením s hustotou

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{pro } a < x < b \\ 0 & \text{pro } x \leq a \text{ nebo } x \geq b. \end{cases}$$

a distribuční funkcí

$$F_X(x) = \begin{cases} 0 & x \leq a \\ \frac{x-a}{b-a} & a < x < b \\ 1 & x \geq b. \end{cases}$$

- říkáme, že X má **rovnoměrné rozdělení** na intervalu (a, b)
- zkráceně píšeme $X \sim R(a, b)$
- střední hodnota a rozptyl (odvození - dom. cvičení)

$$EX = \frac{(a+b)}{2}, \quad var(X) = \frac{(b-a)^2}{12}$$

Příklad: náh. veličina značící chybu při zaokrouhlování

Exponenciální rozdělení

Nechť X je náhodná veličina se spojitým rozdělením s hustotou

$$f_X(x) = \begin{cases} \lambda \cdot e^{-\lambda \cdot x} & x \geq 0 \\ 0 & \textit{jinak}, \end{cases}$$

a distribuční funkcí

$$F_X(x) = \int_{-\infty}^{x} f(t)\,dt = \begin{cases} 1 - e^{-\lambda \cdot x} & x \geq 0 \\ 0 & x < 0. \end{cases}$$

kde $\lambda > 0$ je dané číslo

- říkáme, že X má **exponenciální rozdělení** s parametrem λ
- zkráceně píšeme $X \sim Exp(\lambda)$
- střední hodnota $EX = \int_{-\infty}^{\infty} x \cdot f_X(x)\,dx = \int_0^{\infty} x \cdot \lambda \cdot e^{-\lambda \cdot x}\,dx \overset{\text{p. p.}}{=\!=} \frac{1}{\lambda}$
- rozptyl $var\,X = EX^2 - (EX)^2 = \int_0^{\infty} x^2 \cdot \lambda \cdot e^{-\lambda \cdot x}\,dx - \left(\frac{1}{\lambda}\right)^2 \overset{2\times \text{ p. p.}}{=\!=} \frac{1}{\lambda^2}$

▶ lze chápat jako limitní (spojitý) případ geometrického rozdělení. Používá se pro popis doby čekání na událost nebo doby mezi událostmi, jestliže události nastávají v náhodných okamžicích a nezávisle (doba do příchodu tel. hovoru, příchodu zákazníka, doba do poruchy apod.)

Příklad

(exponenciální rozdělení): Životnost výrobku je v průměru 14 let a dá se modelovat exponenciálním rozdělením. Určete

a) pravděp., že se pokazí v prvním roce po skončení dvouleté záruky

b) jakou maximální záruční dobu může prodejce stanovit tak, aby se během ní nepokazilo více jak 20% výrobků

- ozn. X jako životnost výrobku, $X \sim Exp(\lambda)$
- λ neznáme, ale víme, že $EX = 1/\lambda$
- střední životnost $EX = 1/\lambda$ můžeme odhadnout číslem 14
- položíme tedy $\lambda = \frac{1}{14}$ a spočítáme

a)
$$P(X \in (2,3)) = \int_2^3 f_X(x)\,dx = \int_2^3 \frac{1}{14} \cdot e^{-\frac{x}{14}}\,dx = e^{-\frac{2}{14}} - e^{-\frac{3}{14}}$$

nebo $= P(X < 3) - P(X < 2) = F_X(3) - F_X(2) = 1 - e^{-\frac{3}{14}} - \left(1 - e^{-\frac{2}{14}}\right) \doteq 0{,}06$

b) hledáme zár. dobu z tak, aby $P(X < z) = 0{,}2$

- tedy $z = F_X^{-1}(0{,}2)$ (20%-ní kvantil rozdělení $Exp(\lambda)$)
- $F_X^{-1}(u)$ se určí inverzní k $F_X(x)$: $F_X^{-1}(u) = -\frac{1}{\lambda} \cdot \ln(1-u)$

hledaná záruční doba je $z = -14 \cdot \ln(0{,}8) \doteq 3{,}12 \doteq 3$ roky a 1,5 měsíce

Normální (Gaussovo) rozdělení

Nechť X je náhodná veličina se spojitým rozdělením s hustotou

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

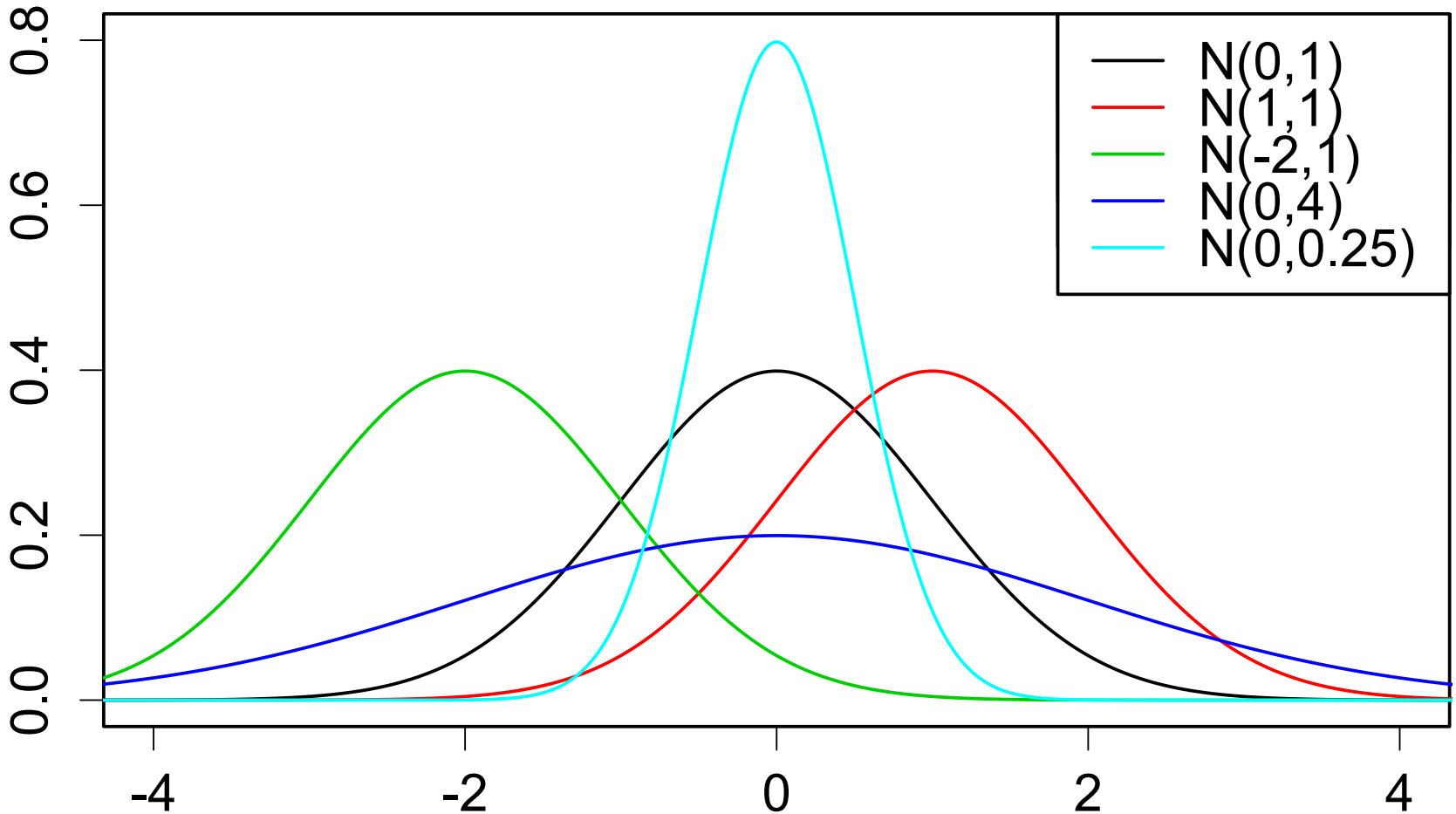
kde $\mu = EX$ a $\sigma^2 = var\, X$ jsou parametry rozdělení.

- 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 $F_X(x) = \int_{-\infty}^{x} f(t)\, dt$ neexistuje explicitní vyjádření
- pro $N(0,1)$ jsou hodnoty přesně tabelovány
- nejdůležitější spojité rozdělení

▶ Vznik: součtem mnoha nepatrných příspěvků

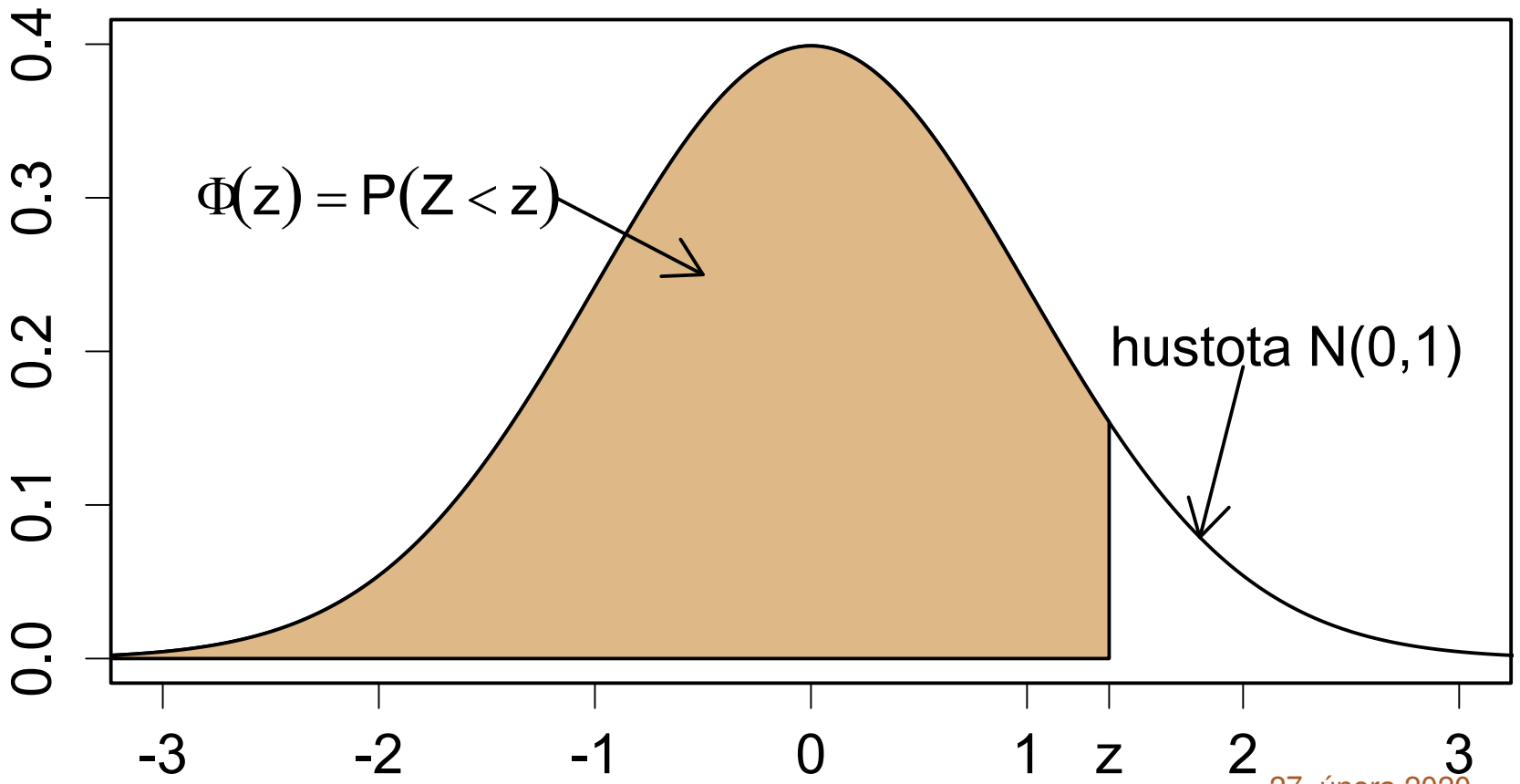
Grafy hustot normálního rozdělení $N(\mu, \sigma^2)$

- symetrické kolem střední hodnoty



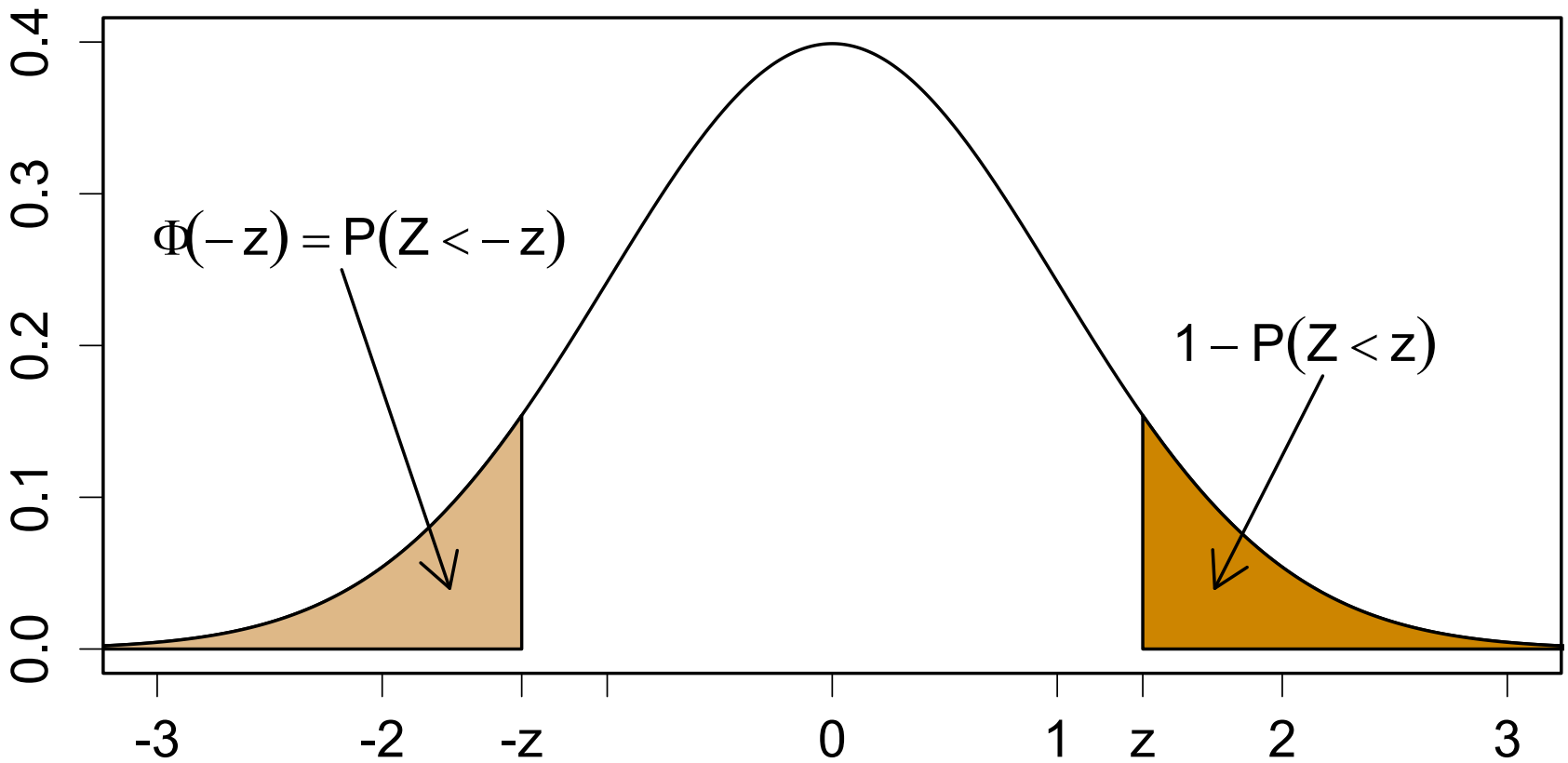
Normované normální rozdělení $Z \sim N(0,1)$

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



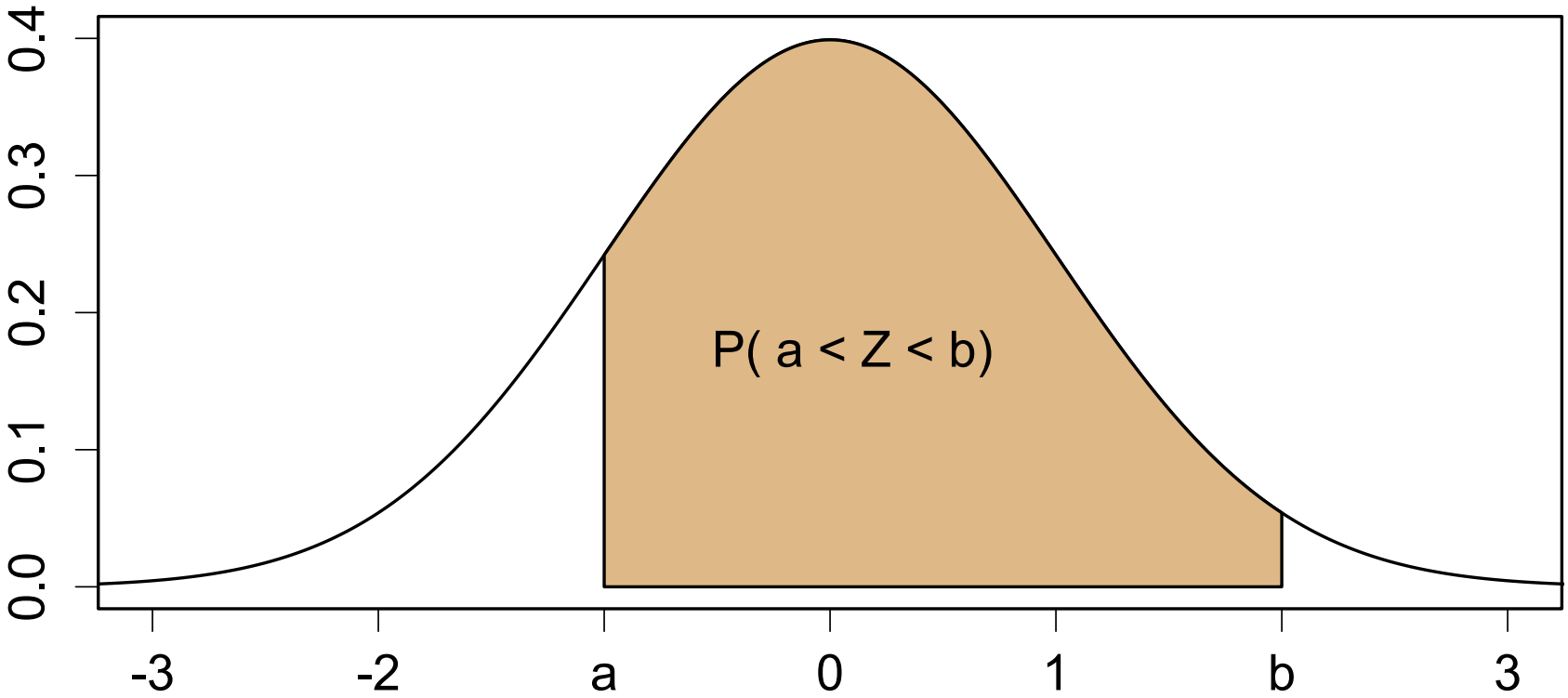
Normované normální rozdělení $Z \sim N(0,1)$

- 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

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

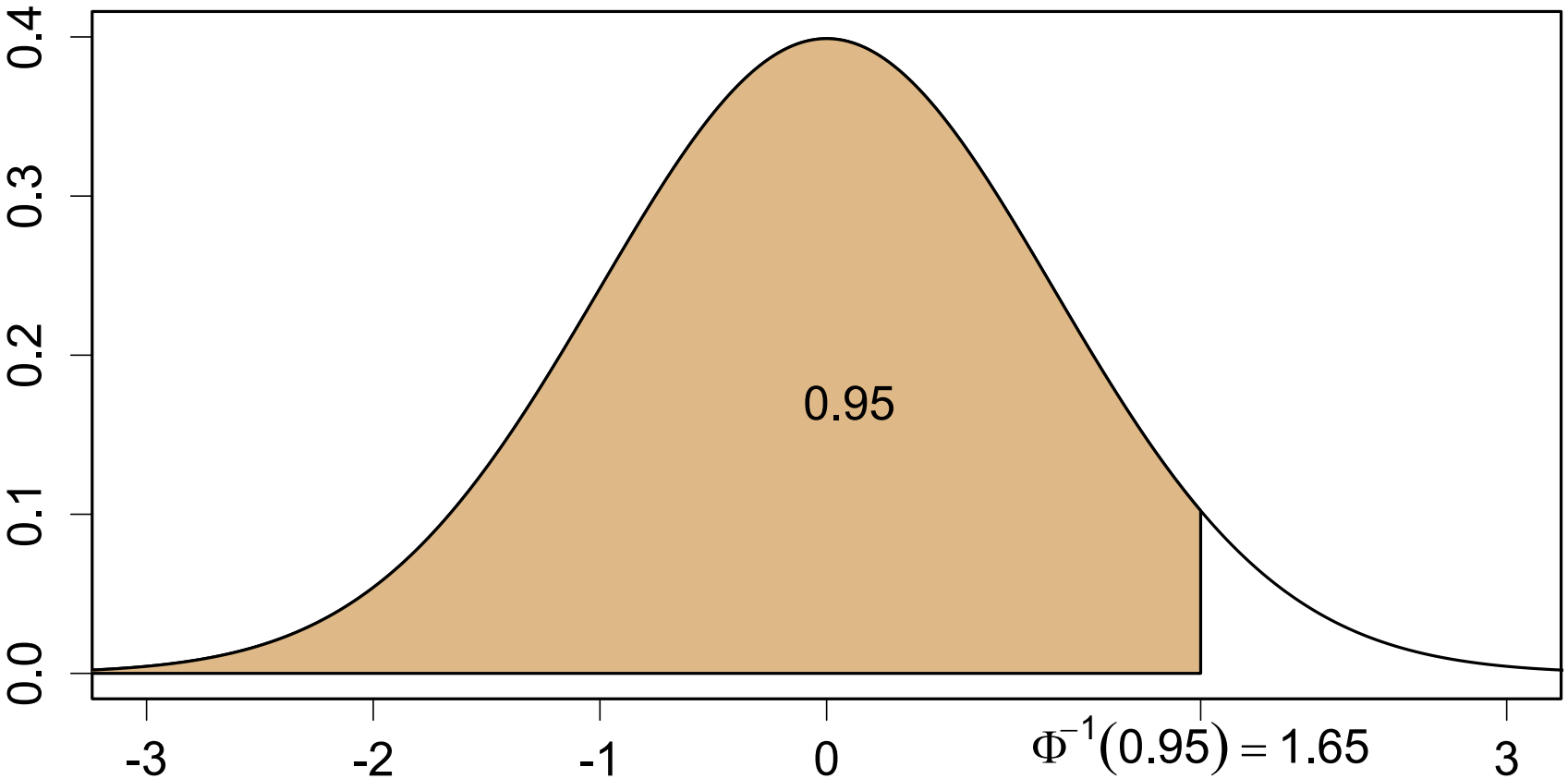
Př.: Výška chlapců v šesté třídě $X \sim N(\mu = 143, \sigma^2 = 49)$: určeme $P(130 < X < 150) = \Phi\left(\frac{150-143}{7}\right) - \Phi\left(\frac{130-143}{7}\right) \doteq 0{,}81$ tedy mezi chlapci v šesté třídě je přibližně 81% vysokých 130 až 150 cm.

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 $\Phi^{-1}(0{,}975) = 1{,}96$

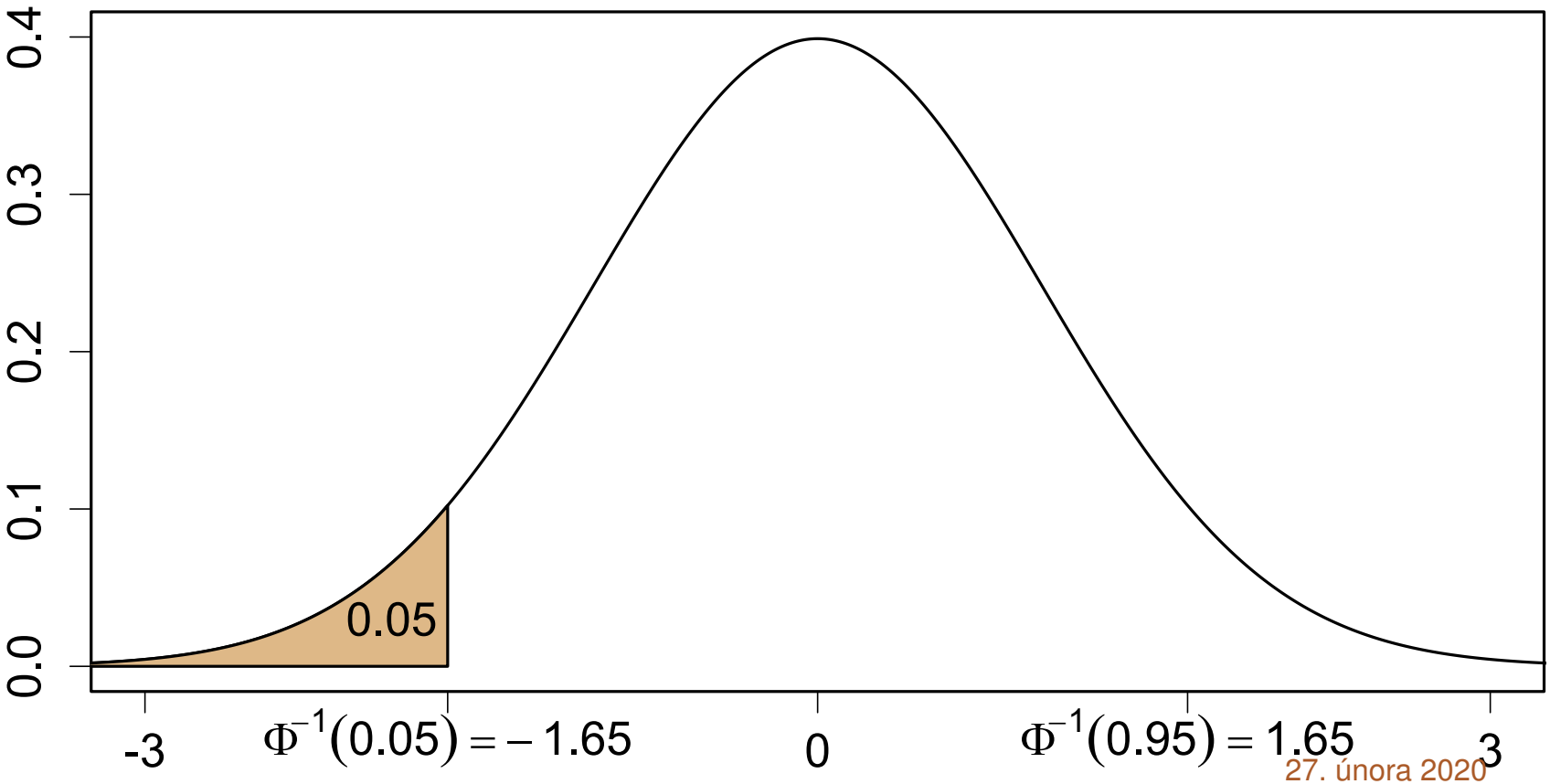


Kvantily normovaného normálního rozdělení 2

- 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) $var\,(\overline{X}) = \frac{\sigma^2}{n}$

Důkaz:

ad 1) Z ▸ vlastností stř. hodnoty (body 1) a 5)) plyne:

$$E\overline{X} = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} EX_i = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

ad 2) Z ▸ vlastností roptylu (body 2) a 6)) plyne:

$$var\,(\overline{X}) = var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} var\,X_i = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}$$

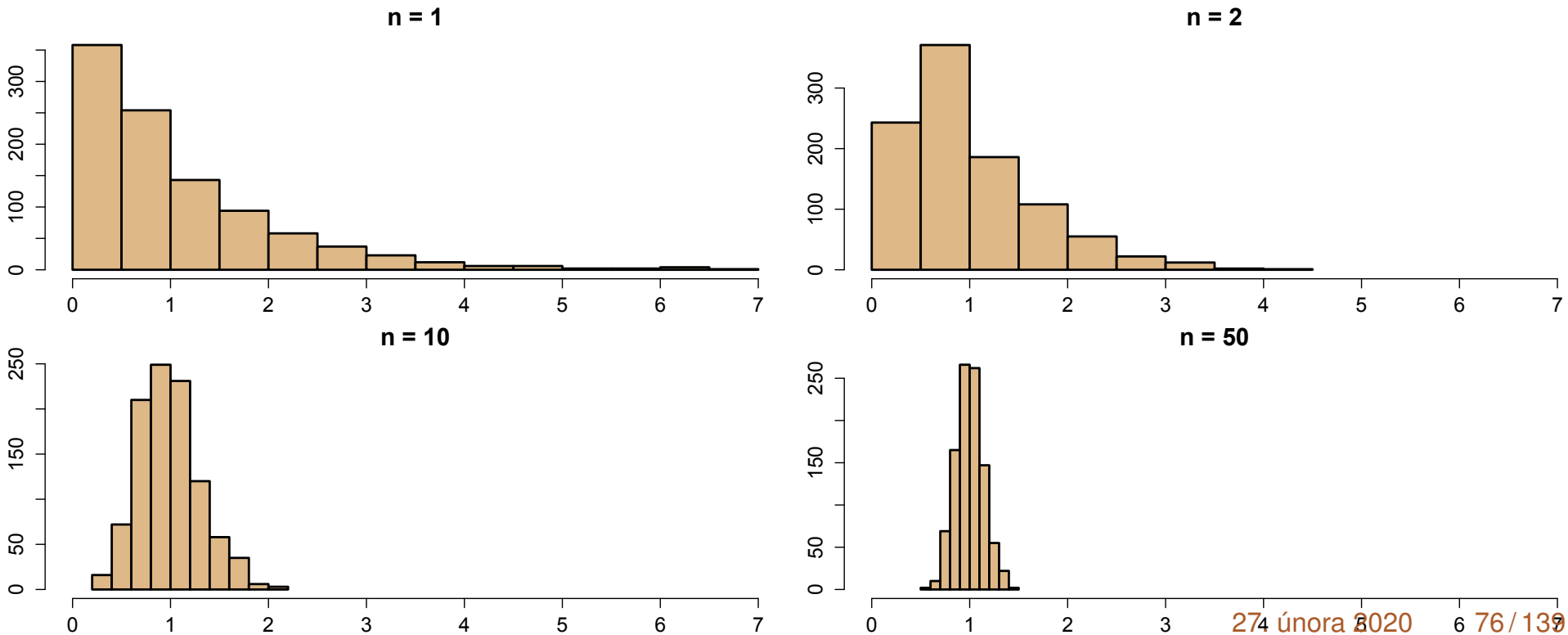
Poznámka:

- z důkazu plyne, že $E(\sum_{i=1}^{n} X_i) = n \cdot \mu$ a $var\,(\sum_{i=1}^{n} X_i) = n \cdot \sigma^2$
- 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 $Exp(\lambda = 1)$)

- z obrázku patrno, že s rostoucím n klesá variabilita průměru a zlepšuje se normalita (platí 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

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

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 $(33{,}66;\ 39{,}96)$

Jak spočítat interval spolehlivosti pro rozptyl (variabilitu hmotností) σ^2?

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

Předpokládejme, že $X_1, X_2, \ldots, X_n$ je náh. výběr z rozdělení $N(\mu, \sigma^2)$. Lze dokázat, že platí

$$P\left(\chi^2_{n-1}(\alpha/2) < \frac{(n-1)\cdot S^2}{\sigma^2} < \chi^2_{n-1}(1-\alpha/2))\right) = 1-\alpha$$

- kde $\chi^2_n(\alpha)$ označuje α-kvantil tzv. χ^2-rozdělení [čti: chí-kvadrát] o n stupních volnosti; najdeme ho v tabulkách

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

$$\left(\frac{(n-1)\cdot S^2}{\chi^2_{n-1}(1-\alpha/2)}; \frac{(n-1)\cdot S^2}{\chi^2_{n-1}(\alpha/2)}\right)$$

- interpretace je podobná jako u předchozích intervalů

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

- spočítali jsme $\overline{X} = 36{,}8125$, $S^2 = 4{,}2711^2$ a máme $n = 16$
- pro 95%-ní int. spol. položíme $\alpha = 0{,}05$
- a najdeme $\chi^2_{15}(1 - 0{,}05/2) = \chi^2_{15}(0{,}975) = 27{,}49$ a $\chi^2_{15}(0{,}05/2) = \chi^2_{15}(0{,}025) = 6{,}26$

Tedy s 95%-ní pravděpodobností je rozptyl hmotností jedenáctiletých chlapců pokryt intervalem:

$$\left(\frac{(n-1)\cdot S^2}{\chi^2_{n-1}(1-\alpha/2)}; \frac{(n-1)\cdot S^2}{\chi^2_{n-1}(\alpha/2)} \right) \doteq$$

$$\doteq \left(\frac{15 \cdot 4{,}2711^2}{27{,}49}; \frac{15 \cdot 4{,}2711^2}{6{,}26} \right) \doteq$$

$$\doteq (9{,}95;\ 43{,}71)$$

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 CLV (Moivreovy-Laplaceovy ▸ 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 $(0{,}065;\ 0{,}145) = (6{,}5\%;\ 14{,}5\%)$

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

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 $|T| = 2{,}634 \geq t_{n-1}(1 - \alpha/2) = t_{15}(0{,}975) = 2{,}131$

Párový t-test

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.
Jak zvolit testové kritérium?

Párový t-test

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, intervalový odhad

Někdy nás zajímá intervalový odhad rozdílu $\mu = \mu_X - \mu_Y$ (podobně jako pro jeden výběr s neznámým rozptylem).
$100(1-\alpha)\%$-ní interval spolehlivosti pro $\mu = \mu_X - \mu_Y$:

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

Tento interval lze použít i 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:
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

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				

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

Jak nyní zvolit testové kritérium?

Dvouvýběrový t-test

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)

Dvouvýběrový t-test, intervalový odhad

Někdy nás zajímá intervalový odhad rozdílu $\mu_X - \mu_Y$.
$100(1-\alpha)\%$-ní interval spolehlivosti pro $\mu_X - \mu_Y$:

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

Tento interval lze použít i 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:
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 $\overline{X} = 139{,}133; \quad \overline{Y} = 140{,}833; \quad S_X^2 = 42{,}981; \quad S_Y^2 = 33{,}788;$

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

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 ▸ Moivreovy-Laplaceovy 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 $|U| \geq \Phi^{-1}(1 - \alpha/2)$

Znaménkový test - exaktní (přesný)

- Používá se, pokud je n malé

Vycházíme z toho, že za platnosti H_0 má počet rozdílů s kladným znaménkem $Y \sim Bi(n, p = 1/2)$ a tedy očekáváme, že zjištěná hodnota Y bude blízko své střední hodnoty $n/2$.
Přikloníme se tedy k $H_1 : \tilde{x} \neq x_0$, pokud bude Y moc malé ($\leq k_1$) nebo moc velké ($\geq k_2$).
Zvolíme hladinu testu α.
Pak k_1 se volí jako největší číslo, pro které ještě platí, že

- $P(Y \leq k_1) \leq \alpha/2$

a k_2 se volí jako nejmenší číslo, pro které ještě platí, že

- $P(Y \geq k_2) \leq \alpha/2$

Zamítáme H_0 na hladině α, pokud $Y \leq k_1$ nebo $Y \geq k_2$.
Pozn.: Skutečná pravd. chyby prvního druhu je často menší než α

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.

Exaktní test:

Máme $Y \sim Bi(n = 46, p = 1/2)$, $\alpha/2 = 0{,}025$ a určíme k_1 a k_2

k	14	**15**	16	...	30	**31**	32
$P(Y = k)$	0,003	0,007	0,014	...	0,014	0,007	0,003
$P(Y \leq k)$	0,006	0,013	0,027	...	0,987	0,994	0,998
$P(Y \geq k)$	0,998	0,994	0,987	...	0,027	0,013	0,006

Protože $k_1 = 15 < Y = 19 < k_2 = 31$, nezamítáme H_0 na hladině 0,05

Pozn.: skutečná hladina testu (pravd. chyba 1. druhu) je jen $2 \cdot 0{,}013 = 0{,}026$.

Asymptotický test: Spočteme

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

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

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.

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

Z ▸ Moivreovy-Laplaceovy 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

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

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

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

Test o parametru p binom. rozd. - exaktní (přesný)

- Používá se, pokud je n malé

Za platnosti H_0 má počet událostí $Y \sim Bi(n, p)$ a tedy očekáváme, že zjištěná hodnota Y bude blízko své střední hodnoty $n \cdot p$.
Zamítneme tedy $H_0 : p = p_0$ ve prospěch alternativy $H_1 : p \neq p_0$, pokud bude Y moc malé ($\leq k_1$) nebo moc velké ($\geq k_2$).
Zvolíme hladinu testu α.
Pak k_1 se volí jako největší číslo, pro které ještě platí, že

- $P(Y \leq k_1) \leq \alpha/2$

a k_2 se volí jako nejmenší číslo, pro které ještě platí, že

- $P(Y \geq k_2) \leq \alpha/2$

Zamítáme H_0 na hladině α, pokud $Y \leq k_1$ nebo $Y \geq k_2$.

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 $H_1 : p \neq 1/6$.

Exaktní test:
Máme $Y \sim Bi(n = 600, p = 1/6)$, $\alpha/2 = 0{,}025$ a určí se $k_1 = 81$ a $k_2 = 119$. Tedy H_0 bychom zamítali.

Zde se ale můžeme spolehnout jen na *Asymptotický test:*
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 $|U| = 4{,}053 \geq \Phi^{-1}(0{,}975) = 1{,}960$

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

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

zpět k ▸ Př.: Byla změřena hmotnost 16 jedenáctiletých chlapců. Lze tvrdit, že se jejich hmotnost změnila oproti době před 25 lety, kdy byla střední hmotnost jedenáctiletých 34 kg? Volme hladinu testu $\alpha = 0{,}01$

Chceme testovat $H_0 : \tilde{x} = 34$ kg proti $H_1 : \tilde{x} \neq 34$ kg. Přitom hodnot větších než 34 je $Y = 13$

Exaktní: $Y \sim Bi(n = 16, p = 1/2)$, $\alpha/2 = 0{,}005$ a určíme k_1 a k_2

k	**2**	3	4	...	12	13	**14**
$P(Y = k)$	0,002	0,009	0,028	...	0,028	0,009	0,002
$P(Y \leq k)$	0,002	0,011 0,011	0,038	...	0,989	0,998	1,000
$P(Y \geq k)$	1,000	0,998	0,989	...	0,038	0,011 0,011	0,002

Protože $k_1 = 2 < Y = 13 < k_2 = 14$, nezamítáme H_0 na hladině 0,01.
Asymptotický (pro $n = 16$ málo věrohodný): Spočteme

$$U = \frac{2Y - n}{\sqrt{n}} = \frac{2 \cdot 13 - 16}{\sqrt{16}} = 2{,}5002{,}500$$

ani nyní H_0 nezamítáme, protože $|U| = 2{,}500 \not\geq \Phi^{-1}(0{,}995) = 2{,}576$

Pozn.: Na hladině 0,05 ($k_1 = 3 < Y = 13 \geq k_2 = 13$ a $\Phi^{-1}(0{,}975) = 1{,}961{,}96$) by oba testy zamítaly H_0.

zpět k ▸ Př.: $n = 8$ lidí bylo podrobeno dietě. Má dieta vliv na hmotnost? Volíme $\alpha = 0{,}05$

Chceme testovat hyp. o mediánu shozených kilogramů $H_0 : \tilde{z} = 0$ kg (nemá vliv) proti $H_1 : \tilde{z} \neq 0$ kg (má vliv). Stačí vědět, že zhublo $Y = 7$ lidí.

Exaktní: $Y \sim Bi(n = 8, p = 1/2)$, $\alpha/2 = 0{,}025$ a určíme k_1 a k_2

k	**0**	1	2	3	4	5	6	7	**8**
$P(Y = k)$	0,004	0,031	0,109	0,219	0,273	0,219	0,109	0,031	0,004
$P(Y \leq k)$	0,004	0,035	0,145	0,363	0,637	0,855	0,965	0,996	1,000
$P(Y \geq k)$	1,000	0,996	0,965	0,855	0,637	0,363	0,145	0,035	0,004

Protože $k_1 = 0 < Y = 7 < k_2 = 8$, nezamítáme H_0 na hladině 0,05 (tedy ani na hladině 0,01).
Asymptotický (pro $n = 8$ velmi málo věrohodný): Spočteme

$$U = \frac{2Y - n}{\sqrt{n}} = \frac{2 \cdot 7 - 8}{\sqrt{8}} = 2{,}121$$

nyní H_0 na hl. 0,05 zamítáme, protože $|U| = 2{,}121 \geq \Phi^{-1}(0{,}975) = 1{,}96$ ale na hl. 0,01 bychom už nezamítali ($\Phi^{-1}(0{,}995) = 2{,}576$)

Jaký test vybrat?

Je lepší jednovýběrový (párový) t-test nebo znaménkový test?
Záleží na situaci.

- při normalitě je vhodnější t-test (díky menší chybě 2. druhu)
- někdy nejsou k dospozici přesná měření, ale jen počet kladných znamének (hodnot větších než hypotetický medián), např. počet nadmírových piv (ne o kolik se lišily od míry) nebo počet pacientů, kteří zhubli po dietě (ne o kolik přesně se změnila jejich hmotnost). Pak nezbývá, než použít znaménkový test.
- pokud data nepocházejí z normálního rozdělení, ale známe přesné hodnoty lze použít tzv. jednovýběrový Wilcoxonův test

jednovýběrový Wilcoxonův test:

- je založen na pořadí hodnot, nepožaduje se normalita
- spolu se znaménkovým testem je zástupcem tzv. *neparametrických testů* (testy, které nepředpokládají, že data pocházejí z nějakého daného rozdělení s parametry, které je nutné odhadovat)
- je lepší než znaménkový test (má menší chybu 2. druhu)

Jednovýběrový Wilcoxonův test - asymptotický

Máme veličiny $X_1, \ldots, X_n$ ze spojitého rozdělení se symetrickou hustotou s mediánem $\tilde{x}$. Testujeme hypotézu $H_0 : \tilde{x} = x_0$ proti $H_1 : \tilde{x} \neq x_0$, kde x_0 je dané číslo.

1. Vyloučíme z dalšího zpracování pozorování, pro něž $X_i = x_0$ a příslušně snížíme rozsah n.
2. Určíme (průměrná) pořadí R_i^+ veličin $|X_i - x_0|$.
3. Test je založen na součtu pořadí R_i^+ těch veličin $|X_i - x_0|$, pro které je $X_i - x_0 > 0$, tj.

$$S = \sum_{i:x_i > x_0} R_i^+$$

Vypočteme statistiku, která má za H_0 asymptoticky norm. normální rozdělení:

$$U = \frac{S - \frac{n\cdot(n+1)}{4}}{\sqrt{\frac{n\cdot(n+1)\cdot(2n+1)}{24}}}$$

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

Použijme Wilcoxonův test na data z ▸ příkladu, kde bylo 8 lidí podrobeno dietě. Má dieta vliv na hmotnost? Testujeme hyp., že medián shozených kilogramů je nula, volíme $\alpha = 0{,}05$

Osoba	1	2	3	4	5	6	7	8
$X_i =$ Před	81	85	92	82	86	88	79	85
$Y_i =$ Po	84	68	73	79	71	80	71	72
$Z_i =$ Rozdíl	-3	17	19	3	15	8	8	13
$\lvert Z_i - 0 \rvert$	3	17	19	3	15	8	8	13
R_i^+	1,5	7	8	1,5	6	3,5	3,5	5

Spočteme $S = \sum_{i:z_i>0} R_i^+ = 7 + 8 + 1{,}5 + 6 + 3{,}5 + 3{,}5 + 5 = 34{,}5.$

$$U = \frac{S - \frac{n\cdot(n+1)}{4}}{\sqrt{\frac{n\cdot(n+1)\cdot(2n+1)}{24}}} = \frac{34{,}5 - 18}{\sqrt{51}} = 2{,}310$$

H_0 na hl. 0,05 zamítáme, protože $|U| = 2{,}310 \geq \Phi^{-1}(0{,}975) = 1{,}96$ ale na hl. 0,01 bychom už nezamítali ($\Phi^{-1}(0{,}995) = 2{,}576$)

- Wilcoxonův test je nejvhodnější volbou, pokud nelze předpokládat normalitu

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

- Jak ale postupovat, pokud nelze předpokládat normalitu?

Spearmanův korelační koeficient

Máme dvourozměrný výběr $(X_1, Y_1), \ldots, (X_n, Y_n)$. Označme $(R_1, \ldots, R_n)$ pořadí veličin $(X_1, \ldots, X_n)$ a $(Q_1, \ldots, Q_n)$ pořadí veličin $(Y_1, \ldots, Y_n)$.

Spearmanův korelační koeficient r_S se spočítá jako Pearsovův kor. koef. počítaný z dvojic $(R_1, Q_1), \ldots, (R_n, Q_n)$. Pokud se ani v jednom souboru nevyskytují shodná pozorování, lze jej zjednodušit na

$$r_S = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (R_i - Q_i)^2$$

- nepožaduje se normalita dat, neparametrická alternativa k r
- také měří míru závislosti, vhodný ale i pro zachycení nelineární monotonní závislosti

Označím-li

$$U = \sqrt{n-1} \cdot r_S,$$

pak hypotézu nezávislosti veličin X a Y na hl. α zamítám, pokud $|U| \geq \Phi^{-1}(1 - \alpha/2)$.

- Podobná míra závislosti: Kendallovo τ - je složitější na výpočet, ale o něco lepší vlastnosti

Na hladině $\alpha = 0{,}05$ testujme hypotézu nezávislosti mezi výsledky z jazykového a matematického testy z ▸ příkladu, pokud nelze předpokládat normalitu.

X_i =Jaz. test	50	23	28	34	14	54	46	52	53
R_i	6	2	3	4	1	9	5	7	8
Y_i =Mat. test	38	28	14	26	18	40	23	30	27
Q_i	8	6	1	4	2	9	3	7	5

Spočteme

$$r_S = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (R_i - Q_i)^2 = 1 - \frac{6}{9(9^2-1)} \left[(6-8)^2 + \ldots + (8-5)^2 \right] = 0{,}683$$

Protože

$$|U| = |\sqrt{n-1} \cdot r_S| = |\sqrt{8} \cdot 0{,}683| = 1{,}933 \not\geq \Phi^{-1}(1-\alpha/2) = 1{,}960$$

hypotézu nezávislosti na hl. 0,05 nezamítáme. Nelze tedy tvrdit, že existuje významná 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:

pohlaví	**strach z matematiky** 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.

Odhad ceny domu

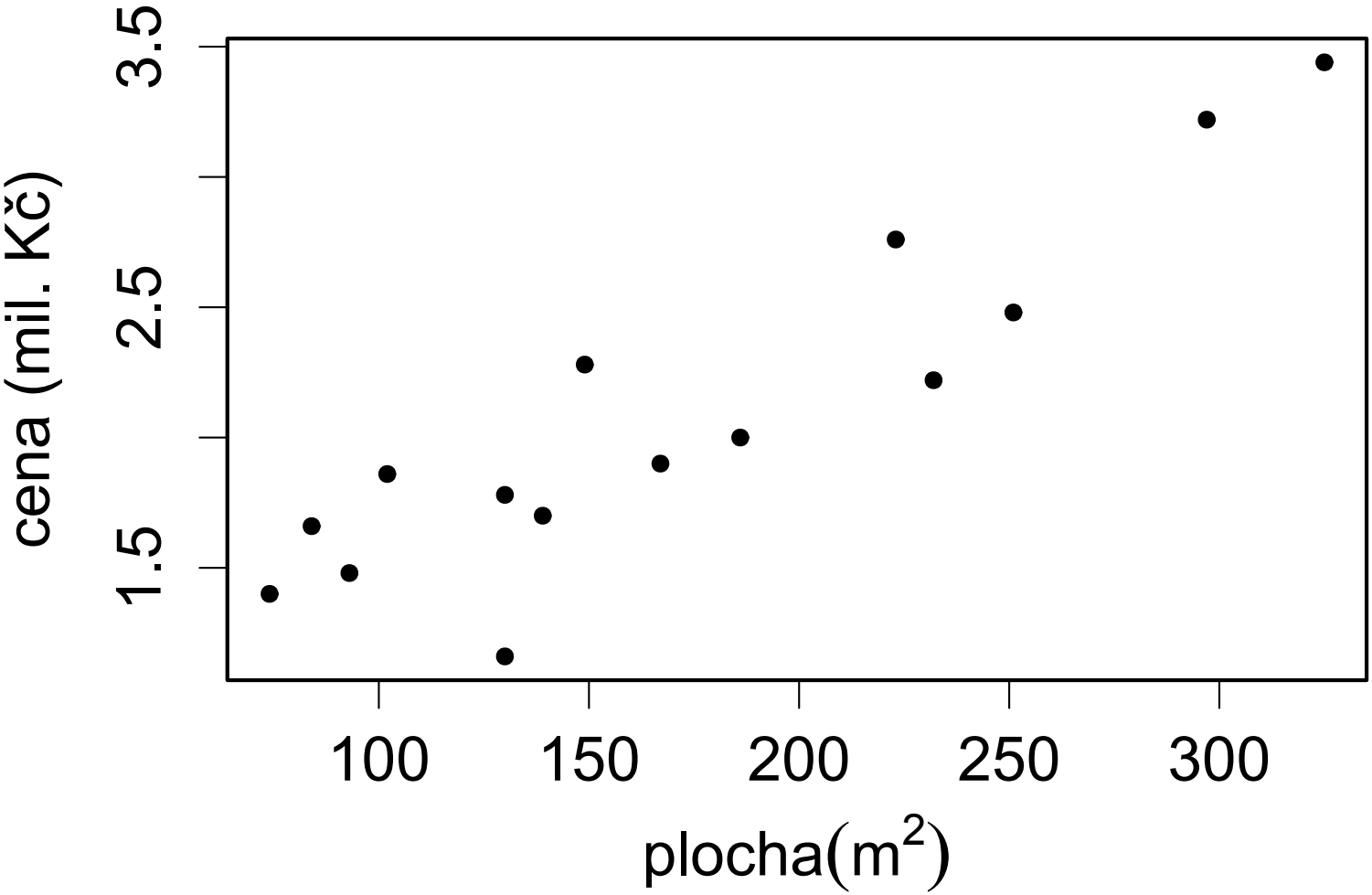
Chcete prodat dům o velikosti $200 m^2$. Jak odhadnout jeho prodejní cenu? K dispozici máte jen velikost a cenu několika jiných domů.

- cenu domů ovlivňuje spousta faktorů (okolí, velikost, stav objektu, atd.)
- pro jednoduchost použijme k odhadu ceny domu pouze jeho velikost
- Jak cenu odhadnout? Stačí provést odhad jen tak od oka, nebo existuje nějaký exaktnější postup?
- prodejní ceny domů (v mil. Kč) a jejich plochy (v m^2) byly:

Plocha (x_i)	Cena(Y_i)
74	1,40
84	1,66
93	1,48
102	1,86
130	1,78
130	1,16
139	1,70
149	2,28
167	1,90
186	2,00
223	2,76
232	2,22
251	2,48
297	3,22
325	3,44

Závislost ceny domu na velikosti

mnohem užitečnější je podívat se na obrázek:



- lze předpokládat, že se cena lineárně mění s plochou

Regresní přímka - metoda nejmenších čtverců

- Máme sadu dvojic $(x_i, Y_i), i = 1, \ldots, n$. Chceme z daných hodnot regresorů x_i (nezávislé proměnné) odhadnout hodnoty závislé proměnné Y_i (vysvětlované proměnné)
- předpoklad: každé ploše domu x_i odpovídá nějaká průměrná (střední) cena Y_i, která závisí na ploše x_i lineárně:

$$EY_i = a + b \cdot x_i, \qquad i = 1, \ldots, n$$

- Navíc předpokládejme, že Y_i jsou nezávislé $Y_i \sim N(a + b \cdot x_i, \sigma^2), \quad i = 1, \ldots, n$
- Parametry a a b regresní přímky se odhadnou **metodou nejmenších čtverců**, tj. hledáme hodnoty, pro které je výraz $\sum_{i=1}^{n}(Y_i - (a + b \cdot x_i))^2$ minimální. Řešením jsou:

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i \cdot Y_i) - n \cdot \overline{x} \cdot \overline{Y}}{\sum_{i=1}^{n} x_i^2 - n \cdot \overline{x}^2} = \frac{S_{xY}}{S_x^2} \qquad \hat{a} = \overline{Y} - \hat{b} \cdot \overline{x}$$

- **Reziduální součet čtverců** (nevysvětlená variabilita Y): $S_e = \sum_{i=1}^{n}(Y_i - (\hat{a} + \hat{b} \cdot x_i))^2$ min. hodnota součtu čtverců
- **Reziduální rozptyl:** $s^2 = S_e/(n-2)$
- rovnice přímky odhadující závislost: $y = \hat{a} + \hat{b} \cdot x$
- Je tato závislost významná? Testujeme $H_0 : b = 0$ proti $H_1 : b \neq 0$ pomocí statistiky

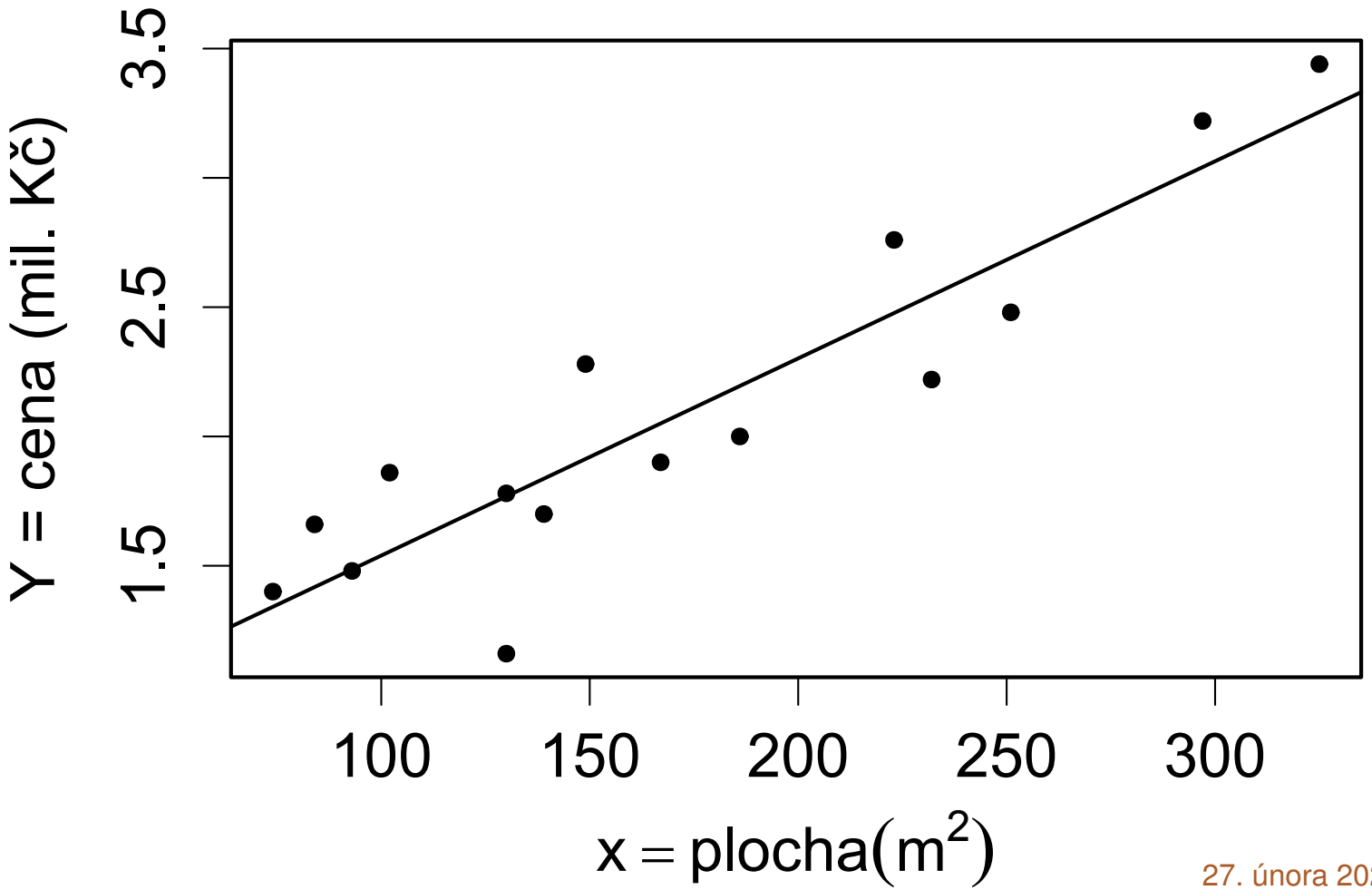
$$T = \frac{\hat{b}}{s} \cdot \sqrt{\sum_{i=1}^{n} x_i^2 - n \cdot \overline{x}^2}$$

hypotézu H_0 (že Y na x nezávisí) na hladině α zamítáme, pokud $|T| \geq t_{n-2}(1 - \alpha/2)$

- **Koeficient determinace:** jaká část celkové variability vysvětlované proměnné ($\sum_{i=1}^{n}(Y_i - \overline{Y})^2$) je závislostí vysvětlena:

$$R^2 = 1 - \frac{S_e}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2} (= r_{xY}^2)$$

zpět k ▸ příkladu. Chceme odhad lineární regresní přímky závislosti ceny domu na jeho ploše. Spočteme $\hat{b} = 0{,}0076(mil./m^2)$ a $\hat{a} = 0{,}777(mil.)$ rovnice přímky odhadující závislost: $y = 0{,}777 + 0{,}0076 \cdot x$
interpretace $\hat{b}$: s každým m^2 roste střední cena domu o 7 600 Kč
interpretace $\hat{a}$ (ne vždy smysluplné): cena domu o 0 m^2 je 777 tis. Kč?



- spočítáme reziduální součet čtverců: $S_e = 1{,}036$
- reziduální rozptyl: $s^2 = S_e/(n-2) = 0.0797$
- Je tato lineární závislost významná? Testujeme $H_0 : b = 0$ proti $H_1 : b \neq 0$ pomocí statistiky

$$T = \frac{\hat{b}}{s} \cdot \sqrt{\sum_{i=1}^{n} x_i^2 - n \cdot \overline{x}^2} = \frac{0{,}0076}{0{,}282} \cdot \sqrt{529780 - 15 \cdot 29629{,}88} = 7{,}9$$

a protože $|T| = 7{,}9 \geq t_{13}(0{,}975) = 2{,}16$, tak hypotézu $H_0 : b = 0$ (že cena na ploše nezávisí) na hladině 0,05 zamítáme.

- koeficient determinace:

$$R^2 = 1 - \frac{S_e}{\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2} = 1 - \frac{1{,}036}{5.997} = 0{,}8272$$

tedy 83% variability ceny je vysvětleno lineární závislostí na ploše.

- odhad střední ceny domu o $200 m^2$:
$\hat{Y} = 0{,}777 + 0{,}0076 \cdot 200 = 2{,}297$

Porovnání více nezávislých výběrů

Dvacet dva pacientů, kteří podstoupili operaci srdce, bylo náhodně rozděleno do tří skupin.
Skupina 1: Pacienti dostali 50 % oxidu dusného a 50 % kyslíkové směsi nepřetržitě po dobu 24 hodin;
Skupina 2: Pacienti dostali 50 % oxidu dusného a 50 % kyslíkové směsi pouze během operace;
Skupina 3: Pacienti nedostali žádný oxid dusný, ale dostali 35 – 50 % kyslíku po dobu 24 hodin.
Tabulka ukazuje koncentraci soli kyseliny listové v červených krvinkách ve všech třech skupinách po uplynutí 24 hodin ventilace. Je mezi ošetřeními rozdíl?

1.	276	280	275	291	347	354	380	330	
2.	206	210	226	249	255	273	285	295	309
3.	241	246	270	293	328				

Porovnání více nezávislých výběrů

Zobecnění dvouvýběrového t-testu na více výběrů.
Pozorujeme I nezávislých náh. výběrů, tj. nezávislé náhodné veličiny

$$Y_{ip} \sim N(\mu + \alpha_i, \sigma^2), \quad i = 1, \ldots, I, \quad p = 1, \ldots, n_i.$$

Chceme testovat $H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_I$ proti obecné alternativě (tj. že alespoň jedna rovnost neplatí).
Položíme $n = \sum_{i=1}^{I} n_i$,

$$y_{i.} = \frac{1}{n_i} \sum_{p=1}^{n_i} Y_{ip} \quad a \quad y_{..} = \frac{1}{n} \sum_{i=1}^{I} \sum_{p=1}^{n_i} Y_{ip}.$$

Dále položíme $f_A = I - 1$, $f_e = n - I$ a také

$$S_A = \sum_{i=1}^{I} n_i y_{i.}^2 - n y_{..}^2 \quad a \quad S_e = \sum_{i=1}^{I} \sum_{p=1}^{n_i} Y_{ip}^2 - \sum_{i=1}^{I} n_i y_{i.}^2$$

Porovnání více nezávislých výběrů

Vypočteme testovou statistiku

$$F_A = \frac{S_A/f_A}{S_e/f_e} = \frac{(\sum_{i=1}^{I} n_i y_{i.}^2 - n y_{..}^2)/(I-1)}{(\sum_{i=1}^{I}\sum_{p=1}^{n_i} Y_{ip}^2 - \sum_{i=1}^{I} n_i y_{i.}^2)/(n-I)}$$

Pokud $F_A > F_{I-1,n-I}(1-\alpha)$, tak zamítáme H_0 na hladině α.

Označíme $s^2 = S_e/f_e$

Pokud zamítneme H_0, tak lze zjistit, které výběry se od sebe liší, použitím metody mnohonásobného porovnávání, např. Tukeyho:

$$|y_{i.} - y_{j.}| \geq s \cdot q_{I,n-I}(1-\alpha)\sqrt{\frac{1}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)},$$

kde $q_{m,\nu}(\alpha)$ je α-kvantil tzv. studentizovaného rozpětí (např. v programu R: qtukey(α, m, ν))

zpět k ▸ příkladu. Spočteme

$$F_A = \frac{S_A/f_A}{S_e/f_e} = \frac{15663/2}{28073/19} = 5{,}301$$

Protože $5{,}301 > F_{I-1,n-I}(1-\alpha) = F_{2,19}(0{,}95) = 3{,}522$, tak zamítáme H_0 na hladině 0,05.

Které skupiny se od sebe liší?
Spočteme $y_{1.} = 316{,}62$, $y_{2.} = 256{,}44$, $y_{3.} = 275{,}60$, $s = 38{,}44$ a $q_{3,19}(0{,}95) = 3.593$.

$$|y_{1.} - y_{2.}| = 60{,}18 \geq 47{,}45,$$

$$|y_{1.} - y_{3.}| = 41{,}02 < 55{,}67,$$

$$|y_{2.} - y_{3.}| = 19{,}16 < 54{,}47.$$