

Základy biostatistiky

MS710P09

Monika Pecková

Ústav aplikací matematiky a výpočetní techniky
Přírodovědecká fakulta University Karlovy
léto 2021

(naposledy upraveno 23. března 2021)

Přednáška

- **Konzultace:** Út 13:15-14:00 nebo dle dohody, Albertov 6, místnost 208.
- **Přednáškové slajdy** v SIS. Počítejte s tím, že se slajdy budoucích přednášek mohou měnit, konečná verze bude až na přednášce.
- **Učebnice:** Karel Zvára: Základy statistiky v prostředí R (Biomedicínská statistika IV), Karolinum, Praha 2013.
Karel Zvára: Biostatistika. Karolinum 1998,...,2008.
- **Zkouška:** kombinovaná - samostatné řešení úloh a ústní zkouška (probíhá v počítačové laboratoři, úlohy řešíte s pomocí PC). Další informace později.
- Bez zápočtu nelze jít ke zkoušce.

Cvičení

- Cvičení v počítačové laboratoři B5.
- Na cvičení potřebujete *login* a *heslo*.
- Požadavky k zápočtu zahrnují aktivní účast na cvičení (maximálně dvě absence). Další požadavky (domácí úkoly apod.) upřesní cvičící.
- Cvičení používají volně šířitelný statistický program R (<https://cran.r-project.org/>)
- Dle preferencí cvičících možno používat balík RCommander (v rámci R) nebo prostředí RStudio (<https://www.rstudio.com/>).

Co je statistika?

Existuje mnoho, obvykle nepříliš lichotivých, charakteristik statistiky:

- Statistika nuda je ...
- Statistika je přesná analýza nepřesných čísel ...
- Lež, nestoudná lež, statistika ...

Tak co je opravdu statistika?

Statistika je věda o získávání a zpracování informace obsažené v empirických pozorováních skutečného světa

Kde se používá statistika?

Zkoumáme složitý systém,

- jehož funkci nelze jednoduše pochopit nebo popsat
- jenž se za stejných nebo podobných podmínek může projevovat odlišným způsobem
- lidské tělo, ekosystém, vědecký experiment, lidská společnost, ekonomika státu ...

Statistický přístup k řešení problémů

- Získáme pozorování - data (experiment, šetření).
- Pozorování považujeme za *náhodný výběr* ze všech možných.
- Stanovíme pravděpodobnostní model pro tato pozorování.
- V rámci modelu přesně zformulujeme problém, který chceme řešit.
- Data a model použijeme k vyřešení problému.

Druhy statistických úloh

- *Odhady parametrů*: Výpočet číselných charakteristik sledovaného systému
- *Testování hypotéz*: Ověřování pravdivosti výroků o chování systému
- *Predikce*: Předpovědi chování systému ve specifických podmínkách

Cíl přednášky:

- Porozumět základním principům statistických metod
- Zvládnout řešení některých jednoduchých problémů

Obsah přednášky

1 Popisná statistika

- vyjádření důležité vlastnosti dat pomocí několika čísel (míry polohy, variability, charakteristiky tvaru)
- vyjádření vlastností pomocí grafu (histogram, boxplot)
- popis závislostí pomocí grafu nebo čísel

2 Teoretická část

- náhodný jev, pravděpodobnost, podmíněná pravděpodobnost
- náhodná veličina, distribuční funkce, střední hodnota, rozptyl
- rozdelení náhodné veličiny, marginální a sdružené rozdelení

3 Statistická indukce

- náhodný výběr, odhad parametrů, intervaly spolehlivosti
- testování hypotéz (princip, hypotézy, hladina, p -hodnota, síla)
- t -testy o stř. hodnotě jednoho a dvou výběrů, neparametrické testy, korelační test
- analýza rozptylu (jednoduché a dvojné třídění)
- lineární regrese (model, testy v regresi, predikce)
- kontingenční tabulky (test dobré shody, test nezávislosti, čtyřpolní tabulky), poměr šancí, logistická regrese

Přednáška 1 (18.2.2020) - obsah

- Měřítka znaků, veličiny
- Diskrétní veličiny (četnosti, grafická zobrazení)
- Spojité veličiny (míry polohy, grafická zobrazení)
- Vlastnosti měr polohy

Jak vypadají data?

Na statistických jednotkách (osoba, pokusné zvíře, krevní vzorek, pokusné pole apod.) měříme znaky.

Příklad (Pacienti s CF)

<i>id</i>	<i>r.nar.</i>	<i>pohl.</i>	<i>mutace1</i>	<i>mutace2</i>	<i>plic.funkce</i>	<i>klin.stav</i>
37	2008	Ž	<i>F508del</i>	<i>F508del</i>	56.74	1
38	1989	Ž	<i>F508del</i>	<i>R553X</i>	88.61	1
39	1990	M	<i>N1303K</i>	<i>N1303K</i>	59.86	3
40	1985	Ž	<i>F508del</i>	<i>F508del</i>	65.91	2
41	2011	M	<i>F508del</i>	<i>F508del</i>	40.09	3
42	1982	M	<i>F508del</i>	<i>F508del</i>	55.23	4
43	2004	M	<i>F508del</i>	<i>G542X</i>	91.14	1

Měřítka znaků

- **Nominální** - pouze několik hodnot (úrovní). Př.: krevní skupina, pohlaví, genetická mutace. Pokud jsou pouze dvě úrovně, mluvíme o nula-jedničkovém měřítku.
- **Ordinální** - jako nominální, ale existuje přirozené uspořádání úrovní. Př.: vzdělání (VŠ/SŠ/ZŠ), kategorie 1-4 podle celkového stavu pacienta. Vzdálenosti mezi úrovněmi nemusí být stejné.
- **Intervalové** - pravidelně rozmištěné hodnoty. Má smysl se ptát o kolik se dvě hodnoty liší, 0 zvolena dohodou. Př.: rok narození, teplota.
- **Poměrové** - jako intervalové, ale 0 není zvolena libovolně. Má smysl se ptát, kolikrát je jedna hodnota vyšší. Př.: věk, výška, hmotnost, plicní funkce.

Měřítka znaků (zjednodušení)

- *Kvalitativní* - nula-jedničkové, nominální, ordinální. Udávají se četnosti úrovní.
- *Kvantitativní* - intervalové a poměrové. Udávají se číselné hodnoty.
- Měřítka znaků nejsou absolutní. V jiném kontextu mohou být jiná. Př.: barva - v biologii nominální, ve fyzice ordinální nebo poměrová.

Veličina

Veličina - číselně vyjádřený výsledek zjišťování.

- **Spojitá** - číselné hodnoty znaků v intervalovém nebo poměrovém měřítku. V praxi budou tyto veličiny měřené s určitou přesností, takže možných hodnot nebude nespočetně mnoho. Přesto s nimi budeme pracovat jako by byly skutečně spojité (zidealizovaný model).
- **Diskrétní** - hodnoty znaků v nula-jedničkovém, nominálním nebo ordinálním měřítku.
- Pro veličinu spočítanou z naměřených hodnot (funkci naměřených hodnot) se používá název **statistika**. Statistikou je třeba průměrná hodnota nebo minimum z naměřených hodnot.

Četnosti, relativní četnosti

- U kvalitativní veličiny pozorování y_1, y_2, \dots, y_n nabývají pouze úrovní A_1, A_2, \dots, A_k . Vzdálenosti mezi úrovněmi nemusí být stejné.
- Veličinu lze shrnout pomocí *četností* n_1, n_2, \dots, n_k , což jsou počty pozorování v kategoriích a *relativních četností* f_1, f_2, \dots, f_k , kde $f_i = \frac{n_i}{n}$.
- Pro četnosti a relativní četnosti samozřejmě platí:
 $n_1 + n_2 + \dots + n_k = n$ a $f_1 + f_2 + \dots + f_k = 1$.

Četnosti, relativní četnosti

Příklad (Krevní skupiny)

Při předoperačních vyšetřeních za poslední měsíc byly zjištěny krevní skupiny 80 pacientů. Takto vypadají data:

AB+	A+	0+	B+	0+	A+	0+	0-	A+	A+
0+	AB+	B+	A-	B+	A+	0+	B+	0+	A-
0+	0+	A+	0+	A+	B-	A+	A+	B+	0+
A+	A+	B+	0-	A+	A-	A-	A+	A-	A+
B+	0+	B+	A+	B+	0+	B-	A-	0+	A+
0+	A+	B-	A+	A+	A+	A+	B+	0+	A+
0+	0-	A-	0+	A+	0+	A+	B+	A+	0+
0+	AB+	AB+	0+	A+	A+	0+	B+	A+	0+

Přehlednější je tabulka četností a relativních četností nebo grafické znázornění pomocí barplotu nebo výsečového diagramu.

Četnosti, relativní četnosti

Příklad (Krevní skupiny)

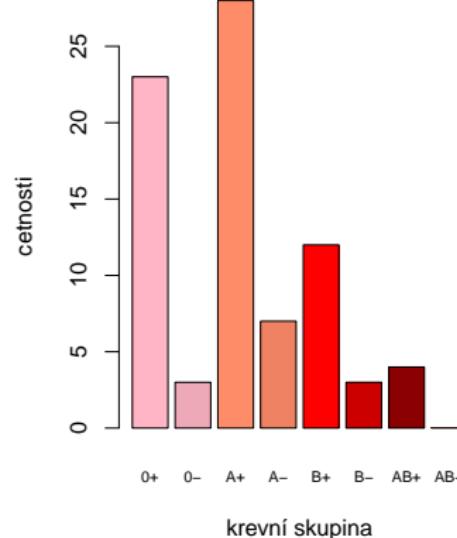
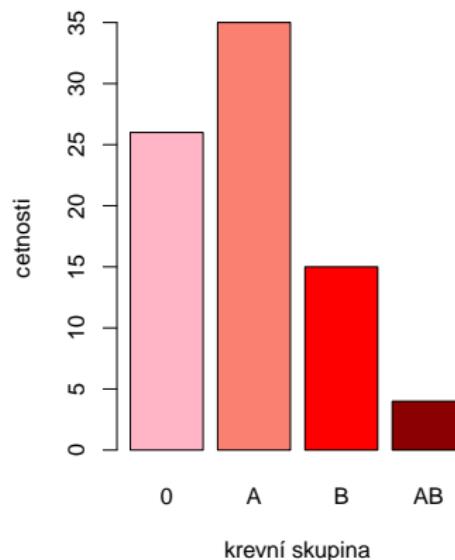
<i>skup.</i>	<i>čet.</i>	<i>rel.čet.</i>	<i>skup.</i>	<i>čet.</i>	<i>rel.čet.</i>
<i>A+</i>	28	0.3500	<i>B+</i>	12	0.1500
<i>A-</i>	7	0.0875	<i>B-</i>	3	0.0375
<i>O+</i>	23	0.2875	<i>AB+</i>	4	0.0500
<i>O-</i>	3	0.0375	<i>AB-</i>	0	0.0000

Nebo bez ohledu na Rh faktor:

<i>kr.skup</i>	<i>čet.</i>	<i>rel.čet.</i>
<i>A</i>	35	0.4375
<i>O</i>	26	0.3250
<i>B</i>	15	0.1875
<i>AB</i>	4	0.0500

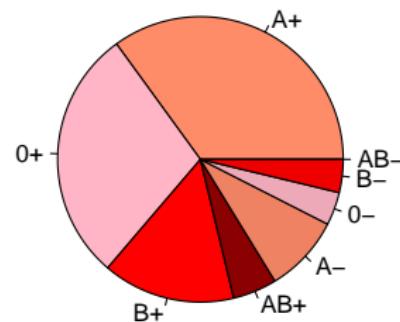
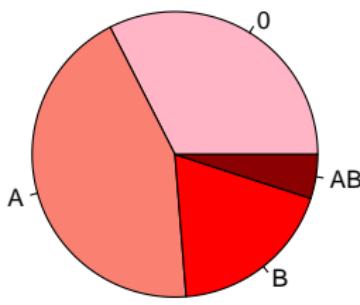
Grafické zobrazení kvalitativní veličiny

Obrázek: Sloupcový diagram (barplot) - Krevní skupiny



Grafické zobrazení kvalitativní veličiny

Obrázek: Výsečový diagram (piechart) - Krevní skupiny



Naměřená data

Pohled na nijak neupravené naměřené hodnoty y_1, y_2, \dots, y_n může být nepřehledný.

Příklad (Desetileté dívky - 100 výšek)

135.9	145.0	149.9	145.7	140.5	137.3	150.1	136.5	147.2	138.7
147.9	146.8	132.8	146.0	152.2	134.6	140.9	140.7	143.9	144.9
143.1	129.0	140.6	130.6	140.2	141.6	133.8	139.0	142.0	159.2
147.0	140.6	147.6	142.7	143.2	140.3	136.0	138.4	146.2	140.5
135.7	134.9	136.2	148.3	129.2	130.0	138.3	141.5	136.3	140.7
142.8	148.5	138.0	146.1	146.5	136.6	145.0	135.3	146.2	145.1
147.9	146.1	132.5	143.8	151.0	139.4	141.8	140.2	136.0	146.9
140.1	141.6	134.2	138.9	145.0	143.6	142.9	135.3	132.2	148.9
141.7	130.3	140.0	146.1	137.2	140.3	143.8	137.8	144.7	154.6
150.9	142.4	143.7	144.6	144.7	138.7	140.6	138.1	148.7	142.4

Variační řada

Lepší přehled získáme, když hodnoty uspořádáme.

Variační řada: $y_{(1)} \leq y_{(2)} \leq y_{(3)} \leq \dots \leq y_{(n)}$

Příklad (Desetileté dívky)

129.0	129.2	130.0	130.3	130.6	132.2	132.5	132.8	133.8	134.2
134.6	134.9	135.3	135.3	135.7	135.9	136.0	136.0	136.2	136.3
136.5	136.6	137.2	137.3	137.8	138.0	138.1	138.3	138.4	138.7
138.7	138.9	139.0	139.4	140.0	140.1	140.2	140.2	140.3	140.3
140.5	140.5	140.6	140.6	140.6	140.7	140.7	140.9	141.5	141.6
141.6	141.7	141.8	142.0	142.4	142.4	142.7	142.8	142.9	143.1
143.2	143.6	143.7	143.8	143.8	143.9	144.6	144.7	144.7	144.9
145.0	145.0	145.0	145.1	145.7	146.0	146.1	146.1	146.1	146.2
146.2	146.5	146.8	146.9	147.0	147.2	147.6	147.9	147.9	148.3
148.5	148.7	148.9	149.9	150.1	150.9	151.0	152.2	154.6	159.2

Pořadí

Pořadí: Jako kolikátá je hodnota ve variační řadě?

Pokud jsou v datech shodné hodnoty, obvykle se používá průměrné pořadí.

Příklad (Desetileté dívky - 10 výšek)

pozorování	pořadí	pozorování	pořadí
135.9	1	137.3	3
145.0	5.5	150.1	10
149.9	9	136.5	2
145.7	7	147.2	8
140.5	4	145.0	5.5

Medián, maximum a minimum

- **maximum** $y_{(n)}$ - nejvyšší hodnota
- **minimum** $y_{(1)}$ - nejnižší hodnota
- **medián** \tilde{y} - hodnota, která je ve variační řadě uprostřed
 $\tilde{y} = y_{(\frac{n+1}{2})}$ pro n lichá
 $\tilde{y} = \frac{1}{2}(y_{(\frac{n}{2})} + y_{(\frac{n}{2}+1)})$ pro n sudá

Pozor! Rozlišujeme y_i a $y_{(i)}$. y_i je pozorování zapsané jako i -té v našem datovém souboru, zatímco $y_{(i)}$ je i -té ve variační řadě.

Příklad (Desetileté dívky)

Minimum, maximum i medián okamžitě vidíme z variační řady:

- **minimum** $y_{(1)} = 129.0$
- **maximum** $y_{(100)} = 159.2$
- **medián** $\tilde{y} = \frac{1}{2}(y_{(50)} + y_{(51)}) = 141.6$

Percentily

- *Percentily* vydělují určité procento nejmenších hodnot.
- y_p je percentil, který vyděluje $100p\%$ nejmenších hodnot (p je číslo z $\langle 0, 1 \rangle$)

Speciální případy percentilů:

- *medián* $p = \frac{1}{2}$
- *kvartily* dolní kvartil Q_1 vyděluje čtvrtinu nejmenších pozorování $p = \frac{1}{4}$, horní kvartil Q_3 vyděluje čtvrtinu nejvyšších pozorování $p = \frac{3}{4}$
- *minimum a maximum* $p = 0$ a $p = 1$
- *decily* vydělují desetinu nejmenších (největších) pozorování $p = \frac{1}{10}$ ($p = \frac{9}{10}$)

Výpočet percentilu

Výpočet percentilu y_p může být komplikovaný. Číslo $100p$ nemusí být celé, v datech mohou být shodná pozorování. Existuje více způsobů, jak spočítat percentil.

Výpočet percentilu, který používá R:

- Najdeme k celé tak, že $\frac{k-1}{n-1} \leq p < \frac{k}{n-1}$.
- To znamená, že $k = [1 + (n - 1)p]$, kde $[a]$ označuje celou část čísla a .
- Percentil bude ležet mezi $y_{(k)}$ a $y_{(k+1)}$.
- Spočítáme ho jako lineární interpolaci

$$y_p = (1 - q)y_{(k)} + qy_{(k+1)}, \text{ kde}$$

$$q = 1 + (n - 1)p - k = \{1 + (n - 1)p\}, \text{ kde } \{a\} \text{ označuje zlomkovou část čísla } a.$$

Funguje výpočet pro medián?

Příklad (Desetileté dívky)

Medián ze 100 pozorování bude průměr 50. a 51. pozorování ve variační řadě. Zkusme použít obecný výpočet percentilu a ověřme, že funguje.

- Počítáme medián, $p = 0.5$.
- Hledáme k tak, aby $\frac{k-1}{99} \leq 0.5 < \frac{k}{99}$, tj. $k - 1 \leq 49.5 < k$, tj. $k = 50$.
- \tilde{y} leží mezi $y_{(50)}$ a $y_{(51)}$.
- $q = 1 + 49.5 - 50 = 0.5$
- $\tilde{y} = 0.5y_{(50)} + 0.5y_{(51)}$, což je skutečně průměr z 50. a 51. prvku variační řady.

Ve variační řadě desetiletých dívek je na 50. i 51. místě 141.6, $\tilde{y} = 141.6$.

Průměr

- *Aritmetický průměr:*

$$\bar{y} = \frac{1}{n}(y_1 + y_2 + \dots + y_n) = \frac{1}{n} \sum_{i=1}^n y_i$$

- *Vážený průměr \bar{y}_w s nezápornými vahami w_1, w_2, \dots, w_n :*

$$\bar{y}_w = \frac{\sum_{i=1}^n w_i y_i}{\sum_{i=1}^n w_i}$$

- Obyčejný aritmetický průměr je speciálním případem váženého s vahami $w_i = 1, i = 1, \dots, n$
- Vážený průměr používáme, když je měření znaku již vztaženo k nějakým menším, nestejně velkým jednotkám. Např. pokud jsou měření ve skutečnosti průměry ve skupinách, rodinách...

Vážený průměr

Příklad (Výšky dívek podle praktického lékaře)

Zajímá nás průměrná výška desetiletých dívek v určitém městě.

Nemůžeme získat jednotlivá měření, ale 3 praktičtí lékaři, kteří ve městě působí, jsou ochotni poskytnout průměrnou výšku desetiletých dívek, které je navštěvují:

lékař	prům. výška	počet dívek
1	142.3	37
2	140.6	148
3	140.1	115

Prostý aritmetický průměr výšek je $\bar{y} = (142.3 + 140.6 + 140.1)/3 = 141$.

To ovšem není průměr dívek ve městě, protože každý z lékařů počítal průměr z jiného počtu dívek. Abychom dostali skutečný průměr ze všech dívek, musíme vážit počty dívek u jednotlivých lékařů:

$$\bar{y}_w = (142.3 \cdot 37 + 140.6 \cdot 148 + 140.1 \cdot 115)/300 = 140.618.$$



Modus

- *Modus* ÿ je nejčastěji vyskytující se hodnota.
- U spojitých veličin záleží na zaokrouhlení, nemusí být smysluplný.

Příklad (Desetileté dívky)

Většina hodnot se vyskytuje pouze jednou, několik hodnot dvakrát, tři hodnoty (140.6, 145.0 a 146.1) se vyskytují třikrát. Vícekrát se žádná hodnota nevyskytuje. Máme tedy 3 modusy, ale nejedná se patrně o nijak zajímavé hodnoty.

Modus

Příklad (Věk otců - data Kojení)

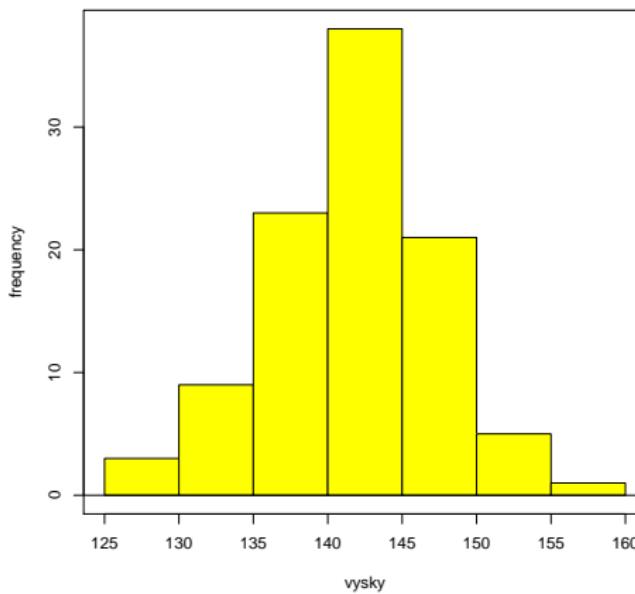
Tabulka: Četnosti (99 hodnot)

věk	čet	věk	čet	věk	čet	věk	čet
19	1	25	4	31	4	37	4
20	3	26	5	32	6	38	1
21	3	27	8	33	4	39	2
22	2	28	14	34	2	42	1
23	5	29	11	35	3	43	1
24	5	30	8	36	2		

Zde dává modus větší smysl, $\hat{y} = 28$. Velmi podobný průměru $\bar{y} = 28.89$ i medánu $\tilde{y} = 28$.

Histogram

Obrázek: Histogram výšek desetiletých dívek



Histogram

Jak histogram vznikl?

Příklad (Desetileté dívky)

- Podívám se, na jakém intervalu se vyskytují hodnoty:
 $\langle 129.0, 159.2 \rangle$
- Zvolím dělení, které tento interval pokrývá:
 $(125, 130) \cup (130, 135) \cup (135, 140) \cup \dots \cup (155, 160)$
- Zjistím četnosti v intervalech a vynesu do grafu

interval	četnost	interval	četnost
$(125, 130)$	3	$(145, 150)$	21
$(130, 135)$	9	$(150, 155)$	5
$(135, 140)$	23	$(155, 160)$	1
$(140, 145)$	38		

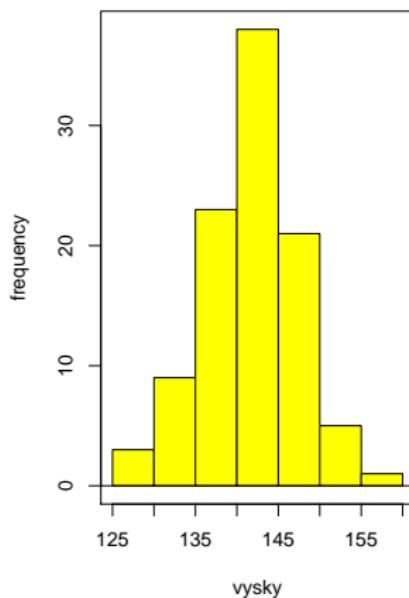
Histogram

- Dělení mohu volit dle svých preferencí. Různá dělení mohou poskytnout jiný obrázek stejných dat.
- Existují algoritmizované postupy hledání dělení pro histogram, např. Sturgesovo pravidlo (používá R).
- Místo četnosti mohu použít hustotu - měřítko na ose y se přepočítá tak, aby plocha celého histogramu byla 1.
- Použití hustoty umožňuje vytvořit histogram s nestejně dlouhými intervaly dělení. Je možné například prodloužit intervaly v oblastech, kde je řídký výskyt dat.

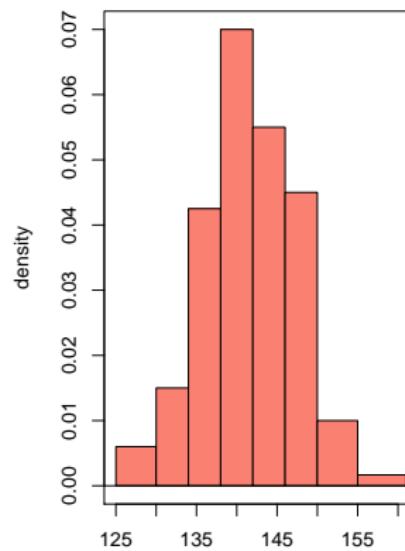
Histogram

Obrázek: Histogram výšek dívek (četnosti a hustota)

Cetnosti, Sturg. pravidlo

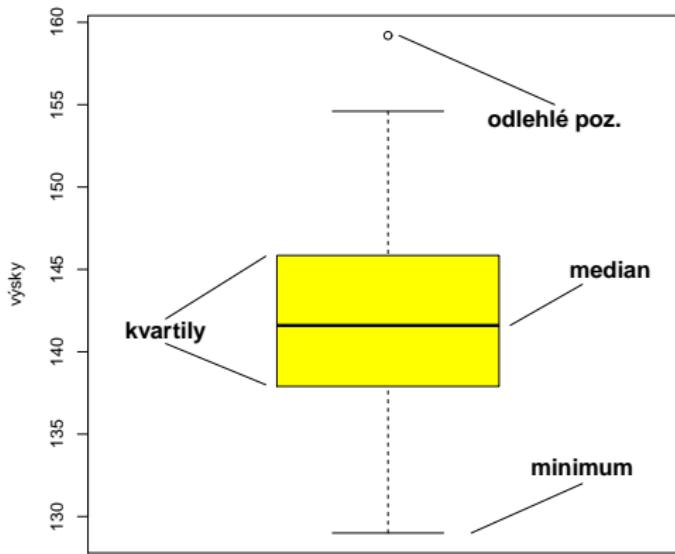


Hustota, zvolené intervaly



Krabicový diagram

Obrázek: Krabicový diagram - výšky dívek



Krubicový diagram

- Silná čára - medián
- Hranice krabice - dolní a horní quartil
- Konce tykadel - minimum a maximum, pokud žádná pozorování nejsou označena jako odlehlá. Pokud jsou odlehlá pozorování, tykadlo končí u posledního neodlehlého, odlehlá jsou označena jako samostatné body.
- Obvykle se označují jako odlehlá ta, která jsou vzdálena od quartilu (hranice krabice) o více než $\frac{3}{2}(Q_3 - Q_1)$ (jedenapůlnásobek rozdílu quartilů).

Empirická distribuční funkce

- *Empirická distribuční funkce* \hat{F}_n v bodě t je rovna podílu pozorování, která byla menší nebo rovna t .

$$\hat{F}_n(t) = \frac{\sharp(y_i \leq t)}{n},$$

kde \sharp značí počet.

- $\hat{F}_n(t)$ je po částech konstantní, skáče v každém bodě, kde se vyskytlo aspoň jedno pozorování. Začíná v 0, v bodě nejvyššího pozorování se dosahuje hodnoty 1.

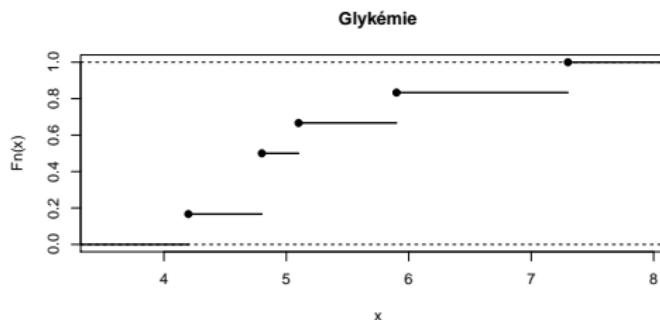
Příklad (Glykémie)

U šesti pacientů byla změřena ranní glykémie v mmol/l. Byly zjištěny tyto hodnoty: 5.1, 4.2, 4.8, 7.3, 4.8, 5.9

Empirická distribuční funkce bude nulová pro všechny hodnoty menší než 4.2 (minimum). V tomto bodě skočí o 1/6. Dále bude konstantní (rovna 1/6) až k bodu 4.8 (druhé nejmenší pozorování). V bodě 4.8 skočí o 2/6=1/3 (hodnotu měli 2 pacienti), a podobně pokračuje dále.

Empirická distribuční funkce

Obrázek: Empirická distribuční funkce

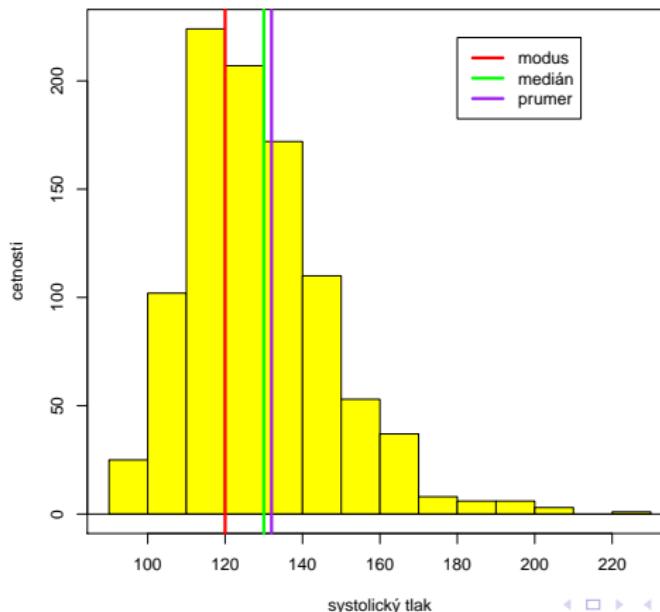


Porovnání měr polohy

- Průměr, percentily a modus jsou *míry polohy*, vypovídají něco o poloze dat.
- Medián, průměr a modus se pokouší nalézt “střed dat.” Každá z těchto měr to ovšem dělá jiným způsobem a mohou se výrazně lišit.
- U veličin se symetrickým histogramem bývají podobné.
- U nesymetrických veličin se často liší. Pokud je histogram sešikmený (odlehlejší pozorování na jedné straně), průměr \bar{y} je obvykle od mediánu \tilde{y} posunut směrem k odlehlym pozorováním. Modus \hat{y} (pokud dává smysl), bývá posunut od mediánu na opačnou stranu než průměr.

Porovnání měr polohy

Obrázek: Míry polohy - systolický tlak - data Stulong



Vlastnosti měr polohy

$\mu(y)$ je míra polohy spočítaná z y_1, y_2, \dots, y_n .

- ① *Posunutí o konstantu:* Posuneme všechna pozorování o konstantu a . Nová pozorování z_1, z_2, \dots, z_n splňují $z_i = y_i + a$ pro $i = 1, 2, \dots, n$. Rozumná míra polohy by se měla také posunout, tedy $\mu(z) = \mu(y) + a$.
- ② *Změna měřítka:* Všechna pozorování vynásobíme kladným číslem $b > 0$. Nová pozorování z_1, z_2, \dots, z_n splňují $z_i = by_i$ pro $i = 1, 2, \dots, n$. Rozumná míra polohy by měla příslušným způsobem změnit měřítko, tedy $\mu(z) = b\mu(y)$.

Je vidět, že průměr, percentily i modus tyto vlastnosti měr polohy splňují.

Přednáška 2 (23.2.2021) - obsah

- Míry variability
- Z-skóry
- Šikmost a špičatost
- Zkoumání závislosti dvou veličin
- Náhodné jevy
- Klasická definice pravděpodobnosti
- Počítání s pravděpodobnostmi

Variabilita (rozptylenost)

Jsou všechny naměřené hodnoty blízko mediánu? Jak jsou rozptylené? To by měly vyjadřovat míry variability. Z naměřených hodnot y_1, y_2, \dots, y_n spočítáme míru variability $\sigma(y)$

- ❶ Míra variability by neměla záviset na *posunutí*:
 $\sigma(y + a) = \sigma(a).$
- ❷ Při *změně měřítka* by se měla změnit rozumně:
 $\sigma(by) = b\sigma(y)$ pro $b > 0.$

Výběrový rozptyl

Výběrový rozptyl s_y^2 průměruje druhé mocniny odchylek pozorování od průměru.

$$\begin{aligned}s_y^2 &= \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \\&= \frac{1}{n-1} [(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \dots + (y_n - \bar{y})^2] \\&= \frac{1}{n-1} [\sum_{i=1}^n y_i^2 - n\bar{y}^2]\end{aligned}$$

Posunutí:

$$s_{(y+a)}^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i + a - \bar{y} - a)^2 = s_y^2$$

rozptyl se nemění.

Změna měřítka:

$$s_{(by)}^2 = \frac{1}{n-1} \sum_{i=1}^n (by_i - b\bar{y})^2 = b^2 s_y^2,$$

rozptyl se mění s druhou mocninou b .

Rozptyl a směrodatná odchylka

- Rozptyl lze vyjádřit také pomocí druhých mocnin vzájemných rozdílů:

$$s_y^2 = \frac{1}{2n(n-1)} \sum_{j=1}^n \sum_{i=1}^n (y_i - y_j)^2$$

- Odmocnina z rozptylu je *směrodatná odchylka* (standardní odchylka),

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

- Směrodatná odchylka se při změně měřítka mění s první mocninou parametru b , jak bychom to od míry variability očekávali.

Další míry variability

- *Rozpětí:*

$$R = \max(y) - \min(y) = y_{(n)} - y_{(1)}$$

- *Kvartilové (mezikvartilové) rozpětí:*

$$R_Q = Q_3 - Q_1$$

- Obě rozpětí splňují požadavky na míry variability.

- *Variační koeficient* je podíl směrodatné odchylky a průměru.

$$V = \frac{s_y}{\bar{y}}$$

- Variační koeficient nesplňuje požadavky na míry variability. Nemění se při změně měřítka, ale při posunutí se mění. Slouží k porovnání variability při různých polohách dat. Dává smysl pouze pro nezáporné veličiny.

Míry variability

Příklad (Glykémie)

Naměřené hodnoty: 5.1, 4.2, 4.8, 7.3, 4.8, 5.9, $\bar{y} = 5.35$

Rozpětí:

$$R = 7.3 - 4.2 = 3.1$$

Kvartilové rozpětí:

$$R_Q = 5.7 - 4.8 = 0.9$$

Rozptyl:

$$s_y^2 = \frac{1}{5}[(5.1 - 5.35)^2 + \dots + (5.9 - 5.35)^2] = 1.219$$

Směrodatná odchylka:

$$s_y = \sqrt{s_y^2} = 1.104$$

z-skóry

- Pozorování y_1, y_2, \dots, y_n posuneme o průměr \bar{y} a vydělíme směrodatnou odchylkou s_y .
- Nově vzniklá pozorování $z_i = \frac{y_i - \bar{y}}{s_y}$ nazveme **z-skóry**.
- Vzhledem k vlastnostem průměru a rozptylu snadno nahlédneme, že $\bar{z} = 0$ a $s_z^2 = s_z = 1$.
- z-skóry nezávisí na posunutí ani změně měřítka, jsou bezrozměrné.
- Lze je použít ke zkoumání vlastností nezávislých na poloze a variabilitě, třeba tvaru rozdělení veličiny.

Šikmost

- *Šikmost* je průměr třetích mocnin z-skórů:

$$g_1 = \frac{1}{n} \sum_{i=1}^n \left(\frac{y_i - \bar{y}}{s_y} \right)^3$$

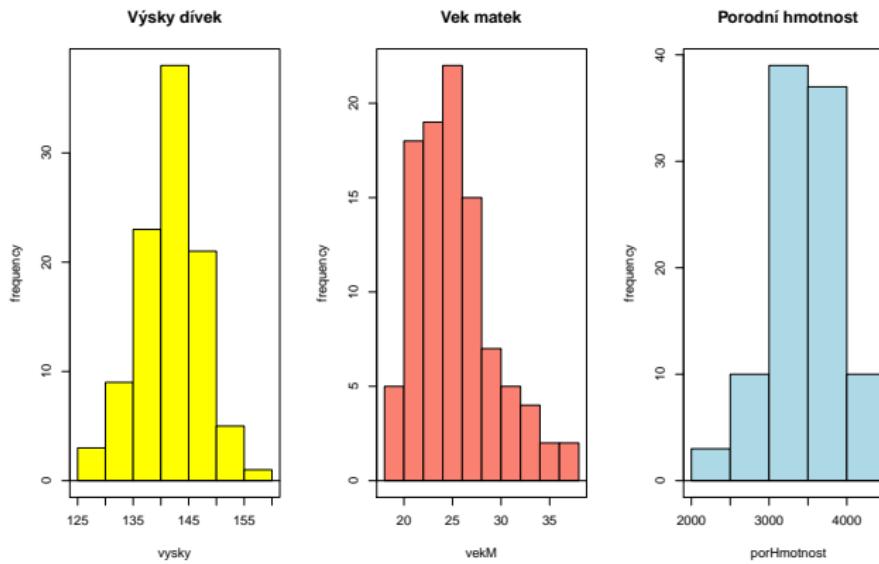
- Veličina se symetrickým histogramem má obvykle g_1 blízko 0.
- Je-li histogram protažený doprava, $g_1 > 0$.
- Je-li histogram protažený doleva, $g_1 < 0$.

Příklad (Různé šikmosti)

Výšky desetiletých dívek jsou dosti symetrické, mají šikmost $g_1 = 0.06$. Věk matek při porodu dítěte je sešikmen vpravo, $g_1 = 0.74$. Porodní hmotnosti mají sešikmení vlevo, $g_1 = -0.32$.

Šikmost

Obrázek: Histogramy veličin s různými šikmostmi



Špičatost

- *Špičatost* je průměr čtvrtých mocnin z-skórů zmenšený o trojku:

$$g_2 = \frac{1}{n} \sum_{i=1}^n \left(\frac{y_i - \bar{y}}{s_y} \right)^4 - 3$$

- Špičatost může být použita k posouzení, z jakého rozdělení data pocházejí. Normální rozdělení má šikmost i špičatost 0.
- Existují i jiné definice šikmosti a špičatosti, mohou se lišit konstantou použitou při průměrování. U špičatosti se někdy neodečítá 3.

Jak zkoumat závislosti

Vhodné grafy i metody prokazování závislostí se budou lišit podle měřítek zkoumaných znaků:

- *kvantitativní-kvantitativní*
rozptylový diagram (scatterplot)
korelace, regrese
- *kvantitativní-kvalitativní*
krabicové diagramy podle úrovní kvalitativní veličiny
t-testy, ANOVA
- *kvalitativní-kvalitativní*
kontingenční tabulka, paralelní barploty
 χ^2 -test, Fisherův test

Závislost dvou kvantitativních veličin

Na každé jednotce sledujeme dva kvantitativní znaky, pozorujeme tedy dvojice $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. Určitou představu o závislosti znaků získáme, když vyneseme pozorování (x_i, y_i) do grafu (rozptylový diagram).

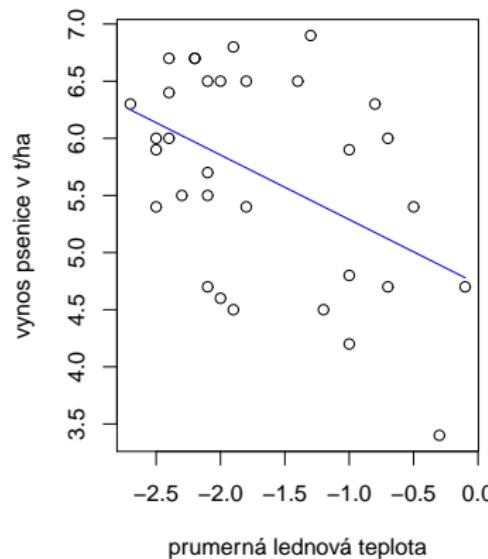
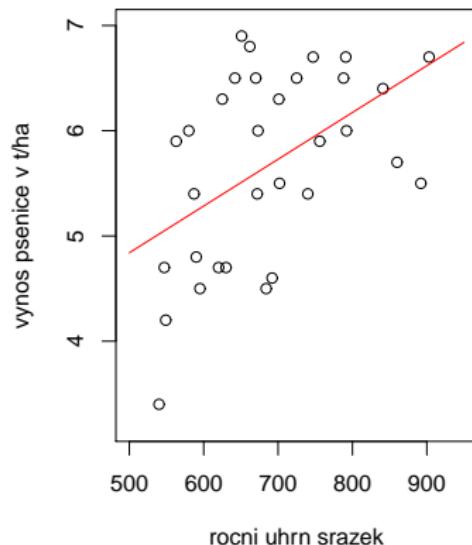
Příklad (Výnosy pšenice)

- *Zkoumáme souvislost průměrných výnosů pšenice (t/ha) ve vybraných oblastech ČR s počasím.*
- *U některých veličin (roční úhrn srážek) pozorujeme rostoucí trend. Srážky souvisí s výnosem v pozitivním smyslu (čím více srážek, tím větší úroda).*
- *U jiných pozorujeme klesající trend, tj. negativní souvislost - čím vyšší průměrná lednová teplota, tím nižší úroda.*
- *U veličin, které s úrodou nesouvisí budeme pozorovat zcela náhodný shluk bodů.*



Rozptylový diagram

Obrázek: Výnos pšenice (2017) v závislosti na srážkách a lednové teplotě



Výběrový korelační koeficient

- *Výběrová kovariance*

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

- Pokud veličiny souvisí v pozitivním smyslu, budeme očekávat, že kovariance je kladná, pokud souvisí v negativním smyslu, kovariance bude patrně záporná.
 - Veličiny, které spolu nesouvisí budou mít kovarianci blízkou 0.
 - *Korelační koeficient (Pearsonův)* je kovariancí z-skóru:
- $$r_{xy} = \frac{1}{n-1} \sum_{i=1}^n \frac{(x_i - \bar{x})}{s_x} \frac{(y_i - \bar{y})}{s_y} = \frac{s_{xy}}{s_x s_y}$$
- Vždy platí $-1 \leq r_{xy} \leq 1$.
 - Pokud body (x_i, y_i) leží na přímce, pak $r_{xy} = \pm 1$.
 - Pokud spolu veličiny nesouvisí, r_{xy} bude blízko 0.

Příklad (Výnosy pšenice)

Korelační koeficient mezi výnosy a srážkami je $r = 0.50$, mezi výnosy a lednovou teplotou je $r = -0.46$.



Závislost kvantitativní a kvalitativní veličiny

Problém lze formulovat jako porovnání několika souborů dat. Každou úroveň kvalitativní veličiny můžeme považovat za jeden soubor a porovnat míry polohy nebo krabicové diagramy.

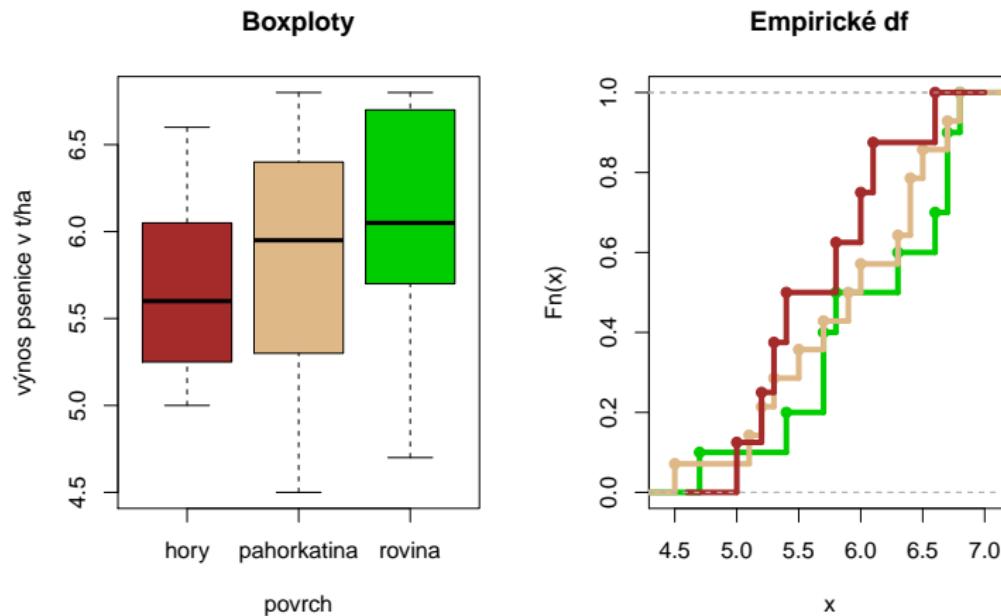
Příklad (Výnosy pšenice)

Zajímá nás, zda výnosy pšenice (t/ha) souvisí s reliéfem krajiny. Rozdělili jsme možné reliéfy na: hory, pahorkatiny a roviny. V tabulce jsou míry polohy výnosů pšenice podle reliéfů:

reliéf	průměr	min	Q_1	med	Q_3	max	n
hory	5.68	5.00	5.28	5.60	6.03	6.60	8
pahor.	5.88	4.50	5.35	5.95	6.40	6.80	14
rovina	6.04	4.70	5.70	6.05	6.68	6.80	10

Grafické znázornění

Obrázek: Výnos pšenice (2019) v závislosti na profilu



Kontingenční tabulka

Data lze shrnout do *kontingenční tabulky* podle kombinací úrovní obou veličin

Příklad (Výnosy pšenice)

Souvisí výnosy pšenice se suchem v oblasti? Nemáme přesné výnosy, ale pouze kategorie (vysoké, střední, nízké). Zasažení oblasti suchem kategorizujeme: výrazné sucho, mírné sucho, oblast nezasažena.

Kontingenční tabulka vypadá takto:

výnosy/sucho	žádné	mírné	výrazné	součet
nízký	2	6	8	16
střední	9	18	6	33
vysoký	11	6	2	19
<i>součet</i>	22	30	16	68

Kontingenční tabulka

Do kontingenční tabulky můžeme doplnit procenta po řádcích nebo po sloupcích. Pokud jsou kategorické veličiny nezávislé, mělo by percentuální zastoupení kategorií jedné veličiny ve všech kategoriích druhé veličiny být podobné.

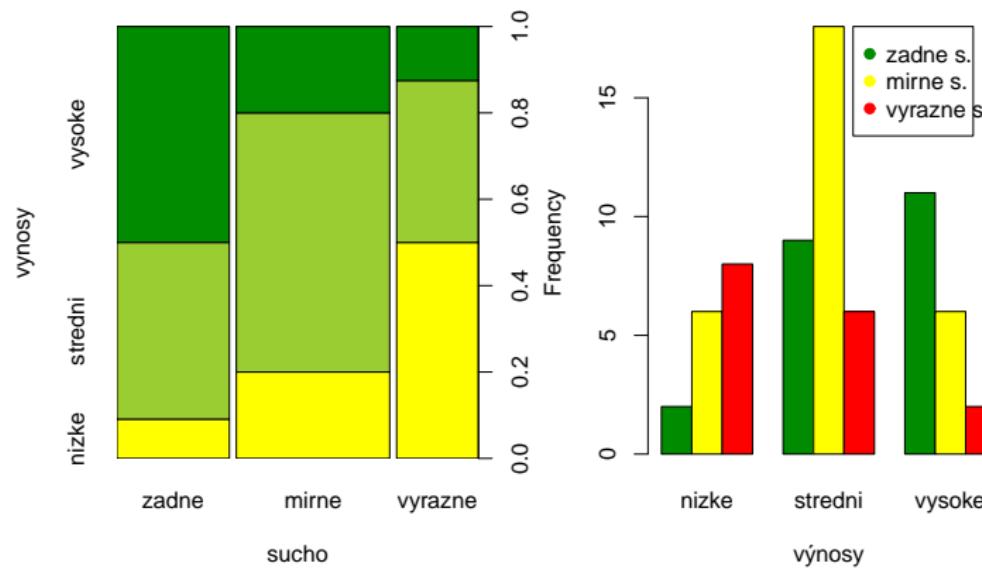
Příklad (Výnosy pšenice)

Doplníme-li procenta po sloupcích (percentuální zastoupení úrovní výnosů v kategoriích podle sucha), vidíme značné rozdíly. To svědčí pro souvislost výnosů se suchem.

výn./sucho	žádné	mírné	výrazné	součet
nízký	2 (9%)	6 (20%)	8 (50%)	16 (23.5%)
střední	9 (41%)	18 (60%)	6 (37.5%)	33 (48.5%)
vysoký	11 (50%)	6 (20%)	2 (12.5%)	19 (28%)
součet	22 (100%)	30 (100%)	16 (100%)	68 (100%)

Grafické znázornění

Obrázek: Výnos pšenice v závislosti na suchu (barploty)



Náhodný pokus

- **Náhodný pokus** je pokus konaný za přesně definovaných podmínek. Není předem známo, jak pokus dopadne.
- Př.1: **Hod kostkou** - může padnout 1, 2, 3, 4, 5 nebo 6.
- Př.2: **Výběr otázky u zkoušky** - student si tahá náhodně jednu z 20 otázek.
- Př.3: **Průzkum veřejného mínění** - náhodně vybraní občané budou kontaktováni a budou zjištovány jejich politické názory.

Náhodný jev

- Tvrzení o náhodném pokusu je *náhodný jev*.
- Př.1: **Hod kostkou** Náhodné jevy: Padne liché číslo. Padne číslo větší než 3.
- Př.2: **Výběr otázky** Náhodný jev: Student si vybere otázku, kterou umí.
- Nejmenší možné jevy, které nemohou nastat současně, ale vždy musí nastat jeden z nich, se nazývají *elementární jevy*.
- Př.1: **Hod kostkou** Elementární jevy: Padne 1. Padne 2....
- Př.2: **Výběr otázky** Elementární jevy: Student si vytáhne otázku č. 1. Student si vytáhne č. 2,...

Množinová představa o náhodných jevech

- Souhrn všech elementárních jevů nazveme Ω .
- Náhodné jevy jsou podmnožiny Ω

Příklad (Házení dvěma kostkami)

Hážeme dvěma kostkami, červenou a modrou. Elementární jevy jsou uspořádané dvojice (1,1), (1,2), ..., (6,6), kde na prvním místě je, co padlo na červené a na druhém, co padlo na modré. Množinu těchto dvojic nazveme Ω . Celkem existuje 36 elementárních jevů.

Jev A: Na obou kostkách padlo sudé číslo.

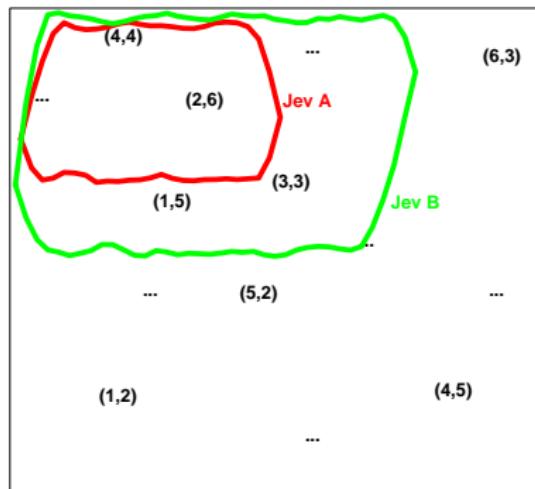
Jev A je podmnožina Ω , která obsahuje dvojice složené ze sudých čísel.

Jev B: Součet hodů na červené a modré kostce je sudý.

Jev B obsahuje dvojice složené ze dvou sudých nebo dvou lichých čísel. $A \subset B$, jev A je podjevem B. Pokud platí tvrzení A, určitě platí i tvrzení B.

Množinová představa o náhodných jevech

Obrázek: Hod dvěma kostkami



Pravděpodobnost

- *Pravděpodobnost* náhodného jevu je číslo z intervalu $\langle 0, 1 \rangle$, které vyjadřuje naše očekávání, jak často k jevu dochází.
- Pokud očekáváme, že všechny elementární jevy nastávají stejně často, můžeme použít *klasickou definici pravděpodobnosti*:
 - mpočet všech elementárních jevů
 - m_Apočet elementárních jevů, které tvoří jev A (počet elementárních jevů příznivých jevu A)
 - Pravděpodobnost jevu A je

$$P(A) = \frac{m_A}{m}$$

Pravděpodobnost

Příklad (Hod dvěma kostkami)

- Předpokládejme, že obě kostky jsou správné, tj., na obou padají všechna čísla se stejnou pravděpodobností $\frac{1}{6}$. Dále předpokládejme, že to, co padne na červené kostce nijak nesouvisí s tím, co padne na modré. Pak můžeme předpokládat, že všechny uspořádané dvojice - elementární jevy - mají stejnou pravděpodobnost.
- Jev A: Součet čísel na kostkách je menší než 5.
- Počet elementárních jevů $m = 36$.
- Jaké elementární jevy splňují, že součet je menší než 5? Snadno zjistíme, že pouze tyto: (1,1), (1,2), (1,3), (2,1), (2,2), (3,1). Takže $m_A = 6$.
- $P(A) = \frac{6}{36} = \frac{1}{6}$

Náhodné jevy

- Celé Ω je také náhodný jev a musí nastat vždy. Jedná se tedy o *jev jistý* a platí $P(\Omega) = 1$.
- Naopak \emptyset nenastane nikdy, je to *jev nemožný* a platí $P(\emptyset) = 0$.
- Jev \bar{A} je *jev opačný* k jevu A , pokud nastane právě tehdy, když nenastane A . Znamená to, že A a \bar{A} nemohou nastat současně a dohromady tvoří celé Ω . Množinově je \bar{A} doplňkem A . Platí $P(A) + P(\bar{A}) = 1$.
- Jevy A a B jsou *neslučitelné*, pokud nemohou nastat současně, znamená to $A \cap B = \emptyset$. Opačné jevy jsou samozřejmě neslučitelné.
- Pro neslučitelné jevy A a B platí $P(A \cup B) = P(A) + P(B)$.
- Jev A je *podjevem* B , pokud z A plyne B . Množinově platí $A \subseteq B$ a $P(A) \leq P(B)$.

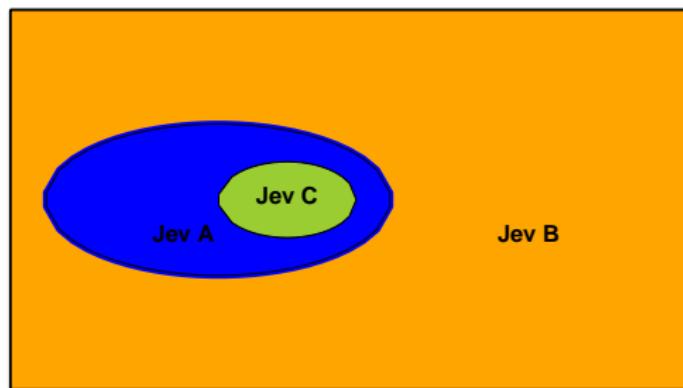
Náhodné jevy

Příklad (Hod dvěma kostkami)

- *Jev jistý: Součet hodů je menší než 15.*
- *Jev nemožný: Součet hodů je 1.*
- *Jevy opačné: Součet je lichý. Součet je sudý.*
- *Jevy neslučitelné, ale ne opačné: Součet je sudý. Rozdíl mezi hody je 1.*
- *Podjevy: Jev A: Na obou kostkách padne stejné číslo. Jev B: Součet je sudý. $A \subseteq B$.*

Vénovy diagramy

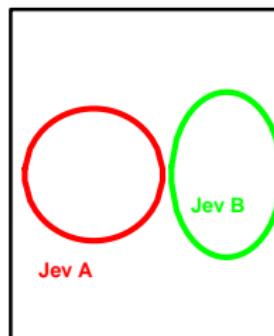
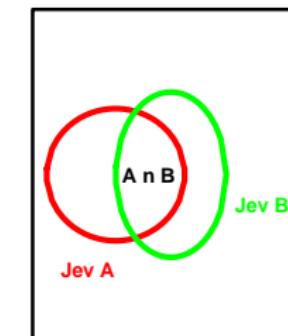
Náhodné jevy lze zobrazovat jako množiny pomocí Vénových diagramů. Celý prostor elementárních jevů Ω zobrazíme jako obdélník. Náhodné jevy jsou jeho podmnožiny: $B = \bar{A}$, $C \subset A$.

 Ω

Pravděpodobnost sjednocení jevů

Pro neslučitelné jevy platí $P(A \cup B) = P(A) + P(B)$. Pokud ovšem jevy nejsou neslučitelné, mají neprázdný průnik, kde platí oba současně. Obecně tedy platí:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

 Ω  Ω

Pravděpodobnost sjednocení jevů

Příklad (Hod dvěma kostkami)

- Zvítězíme ve hře, když součet hodů bude aspoň 10, nebo když na obou kostkách padne stejné číslo. Jakou má tento jev (jev A) pravděpodobnost?

- Jev B: Součet hodů je aspoň 10.

$$B = \{(6, 6), (6, 5), (5, 6), (5, 5), (4, 6), (6, 4)\}$$

$$P(B) = \frac{m_B}{m} = \frac{6}{36} = \frac{1}{6}$$

- Jev C: Na obou kostkách padlo stejné číslo.

$$C = \{(6, 6), (5, 5), (4, 4), (3, 3), (2, 2), (1, 1)\}$$

$$P(C) = \frac{m_C}{m} = \frac{6}{36} = \frac{1}{6}$$

- $B \cap C = \{(6, 6), (5, 5)\}$, $P(B \cap C) = \frac{m_{B \cap C}}{m} = \frac{2}{36} = \frac{1}{18}$

- $P(A) = P(B \cup C) = P(B) + P(C) - P(B \cap C) = \frac{1}{6} + \frac{1}{6} - \frac{1}{18} = \frac{5}{18}$

Přednáška 3 (2.3.2021) - obsah

- Opakování kombinatoriky, počítání s pravděpodobnostmi
- Podmíněná pravděpodobnost a nezávislost jevů
- Bayesův vzorec, vzorec pro úplnou pravděpodobnost
- Vlastnosti diagnostických testů
- Náhodná veličina, distribuční funkce
- Spojité rozdělení, hustota

Permutace

- Kolika způsoby mohu seřadit 15 studentů do řady?
- Počet *permutací* z n prvků je $n! = 1 \cdot 2 \cdot 3 \cdot \dots \cdot (n - 1) \cdot n$
- Faktoriály rostou velmi rychle.

$$3! = 1 \cdot 2 \cdot 3 = 6$$

$$5! = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 = 120$$

$$10! = 3628800$$

$$15! = 1.307674 \cdot 10^{12}$$

Kombinace

- Kolika způsoby mohu vybrat z deseti dětí dvojici? Nezáleží na pořadí ve vybrané dvojici.
- Počet k-prvkových *kombinací* z n prvků:

$${n \choose k} = \frac{n!}{(n-k)!k!}$$

(kombinační číslo, "n nad k")

- Proč?
- Počet dvojic z 10 dětí tedy je

$${10 \choose 2} = \frac{10!}{8!2!} = \frac{9 \cdot 10}{2} = 45$$

Počítání s pravděpodobnostmi

Příklad (Skautské vánoce)

Ve skautské družince je 8 kluků. Na vánoční schůzku skauti přinesou každý 2 zabalené dárky. Dárky se náhodně rozdají a každý dostane zase 2 dárky. Petr šel nakupovat dárky s Honzou a oba koupili jako oba dárky baterky. Jaká je pravděpodobnost, že si Petr přinese domů 2 baterky, když předpokládáme, že nikdo jiný než Honza a Petr baterky nenaděloval?

Počet dvojic dárků, které může Petr dostat (elementární jevy) označme m .

$$m = \binom{16}{2} = \frac{16!}{14!2!} = \frac{16 \cdot 15}{2} = 120$$

Jev A: Petr dostane 2 baterky. $m_A = \binom{4}{2} = \frac{4!}{2!2!} = \frac{4 \cdot 3}{2} = 6$,

$$P(A) = \frac{m_A}{m} = \frac{6}{120} = 0.05$$

Počítání s pravděpodobnostmi

Příklad (Skautské vánoce-pokračování)

Jaká je pravděpodobnost, že Petr dostane aspoň jednu baterku?

(Jev B)

Jev B obsahuje buď dvojice se dvěma baterkami, nebo s jednou baterkou. Jaká je pravděpodobnost dvou baterek již víme. Jaká je pravděpodobnost, že Petr dostane právě jednu baterku (Jev C)? V tom případě dostane jednu ze 4 baterek a jeden ze 12 ostatních dárků.

$$m_C = \binom{4}{1} \cdot \binom{12}{1} = 4 \cdot 12 = 48, P(C) = \frac{48}{120} = 0.4$$

Protože jevy A a C nemohou nastat současně, pravděpodobnost jevu B bude součtem pravděpodobností jevů A a C, tedy

$$P(B) = P(A) + P(C) = 0.05 + 0.4 = 0.45.$$

Pravděpodobnost, že si vytáhne aspoň jednu baterku bude tedy 0.45.

Lze spočítat $P(B)$ jednodušeji?



Počítání s pravděpodobnostmi

- Předpokládejme, že náhodný pokus spočívá ve vybrání k -tice z n prvků a že každá k -tice je stejně pravděpodobná.
- Počet možných k -tic je $m = \binom{n}{k}$, každá z nich je vybrána s pravděpodobností $1/m$.
- Skautský příklad předpokládal, že vybírané prvky jsou dvou druhů: baterky a nebaterky. Řekněme tedy, že máme a prvků 1. druhu a $n - a$ prvků 2. druhu a zajímá nás pravděpodobnost $P(A_j)$, že ve vybrané k -tici je právě j prvků prvního druhu.
- Pokud jsme vybrali právě j prvků 1. druhu, znamená to, že jsme vybrali také právě $k - j$ prvků 2. druhu. Možností, jak to udělat je $\binom{a}{j} \cdot \binom{n-a}{k-j}$ a bude tedy platit:

$$P(A_j) = \frac{\binom{a}{j} \cdot \binom{n-a}{k-j}}{\binom{n}{k}}$$

pro j , která splňují $\max(0, k + a - n) \leq j \leq \min(a, k)$

Počítání s pravděpodobnostmi

- Popsané schéma vybírání je jedním z klasických pravděpodobnostních modelů a uvidíme, že souvisí s hypergeometrickým rozdělením.
- Situací, kde se podobný model vyskytuje je mnoho (výběr otázky, kterou student umí; výběr karty při hře apod.)

Příklad (Počet ryb v rybníku)

Chceme odhadnout n - neznámý počet ryb v rybníku. Vylovíme určitý počet ryb (a), označíme je a pustíme zpět do rybníka. Po nějaké době vylovíme k ryb. Je-li počet označených mezi vylovenými j , můžeme předpokládat, že k/j (poměr vylovených ku označeným mezi vylovenými) je stejný jako n/a (poměr všech ryb ku všem označeným v rybníku) a odhadnout $\hat{n} = \frac{k \cdot a}{j}$.

Podmíněná pravděpodobnost

Podmíněná pravděpodobnost jevu A , víme-li, že nastal jev B , je

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

(předpokládáme, že $P(B) > 0$)

Příklad (Hod dvěma kostkami)

Vraťme se k jevům B (součet je aspoň 10) a C (na obou kostkách padlo stejné číslo) a spočítejme pravděpodobnost B , víme-li, že nastal C .

$$P(B|C) = \frac{P(B \cap C)}{P(C)} = \frac{\frac{1}{18}}{\frac{1}{6}} = \frac{6}{18} = \frac{1}{3}$$

Vidíme, že podmíněná pravděpodobnost je o dost vyšší než nepodmíněná ($P(B) = \frac{1}{6}$). Víme-li tedy, že na obou kostkách padlo stejné číslo, pravděpodobnost vysokého součtu je vyšší, než nevíme-li nic.

Podmíněná pravděpodobnost

Příklad (Hod dvěma kostkami)

Zkusme spočítat podmíněnou pravděpodobnost toho, že na obou kostkách padlo stejné číslo (jev C) za předpokladu, že na červené kostce padlo 4 (jev D). Je jasné, že $P(D) = \frac{1}{6}$ a $C \cap D$ obsahuje jedinou dvojici $(4, 4)$, takže $P(C \cap D) = \frac{1}{36}$.

$$P(C|D) = \frac{P(C \cap D)}{P(D)} = \frac{\frac{1}{36}}{\frac{1}{6}} = \frac{6}{36} = \frac{1}{6}$$

V tomto případě je podmíněná pravděpodobnost stejná jako nepodmíněná. Víme-li, že na červené kostce padla 4, nijak to neovlivní naše očekávání, zda nastane jev C . Jevy C a D jsou nezávislé.

Nezávislost jevů

- Jevy A a B jsou *nezávislé* právě tehdy, když

$$P(A) = P(A|B).$$

- Z definice podmíněné pravděpodobnosti vyplývá, že jevy jsou nezávislé právě tehdy, když

$$P(A) = \frac{P(A \cap B)}{P(B)},$$

tj. právě tehdy, když

$$P(A)P(B) = P(A \cap B).$$

- Dva jevy jsou tedy nezávislé, pokud pravděpodobnost jednoho z nich nijak neovlivní to, zda nastal ten druhý.

Nezávislost jevů

Příklad (Hod dvěma kostkami)

Jsou jevy, že na červené kostce padlo liché číslo (jev E) a že součet hodů je menší než 4 (jev F) nezávislé?

- $P(E) = \frac{1}{2}$, to je jasné.
- Jev $F = \{(1, 1), (1, 2), (2, 1)\}$, tedy $P(F) = \frac{3}{36} = \frac{1}{12}$.
- $(E \cap F) = \{(1, 1), (1, 2)\}$, takže
 $P(E \cap F) = \frac{2}{36} = \frac{1}{18} \neq P(E) \cdot P(F) = \frac{1}{2} \cdot \frac{1}{12} = \frac{1}{24}$.
- Jevy E a F jsou závislé.
- Závislost jevů E a F vypadá pochopitelně. Lichá čísla jsou ta menší z možných na kostce, takže patrně nižší součet bude pravděpodobnější, víme-li, že na jedné z kostek padlo něco lichého.

Vzorec pro celkovou pravděpodobnost

- Chceme spočítat $P(A)$ pravděpodobnost jevu A , ale známe pouze podmíněné pravděpodobnosti $P(A|H_i)$ jevu A pokud nastanou jevy H_i , $i = 1, 2, \dots, k$, pro nějaké jevy H_i , které jsou navzájem neslučitelné a jejich sjednocení je jev jistý ($H_1 \cup H_2 \cup \dots \cup H_k = \Omega$ a $H_i \cap H_j = \emptyset$, pro všechna i, j).
- Použijeme $P(A) = P(A \cap H_1) + \dots + P(A \cap H_k) = \sum_{i=1}^k P(A \cap H_i)$
- Z definice podmíněné pravděpodobnosti plyne, že $P(A \cap H_i) = P(A|H_i)P(H_i)$, takže platí

$$P(A) = \sum_{i=1}^k P(A|H_i)P(H_i)$$

- Tomuto vztahu se říká *vzorec pro celkovou pravděpodobnost*. Pokud je Ω sjednocením neslučitelných jevů H_1, \dots, H_k , pak $P(A)$ lze vyjádřit jako vážený průměr podmíněných pravděpodobností $P(A|H_i)$, kde jako váhy slouží pravděpodobnosti $P(H_i)$.

Bayesův vzorec

- Předpokládejme, že známe $P(A|H_i)$ pro H_i neslučitelné, jejichž sjednocení je celé Ω . Zajímají nás však $P(H_i|A)$, které neznáme.
- Použijeme definici podmíněné pravděpodobnosti a vzorec pro celkovou pravděpodobnost a odvodíme *Bayesův vzorec*:

$$P(H_i|A) = \frac{P(H_i \cap A)}{P(A)} = \frac{P(A|H_i)P(H_i)}{\sum_{i=1}^k P(A|H_i)P(H_i)}$$

- Bayesův vzorec umožňuje za dané situace spočítat "obrácené" podmíněné pravděpodobnosti.
- Jevy H_i mohou být nějaké hypotézy, z nichž nastává právě jedna. Pravděpodobnosti $P(H_i)$ nazveme *apriorními* (platné pokud nevíme, zda nastal jev A).
- Podmíněné pravděpodobnosti $P(H_i|A)$ pak nazveme *aposteriorními*, jsou to pravděpodobnosti hypotéz spočítané poté, co jsme zjistili, že nastal jev A.

Bayesův vzorec

Příklad (Podíl hereditární rakoviny prsu v populaci)

V ČR je přibližně jedna žena z 800 postižena mutací genu BRCA (přítomnost mutace označme M), která výrazně zvyšuje pravděpodobnost rakoviny prsu (BC). Pravděpodobnost, že žena s touto mutací dostane v průběhu života rakovinu prsu je $P(BC|M) = 0.87$. Pravděpodobnost rakoviny prsu u ženy bez mutace je $P(BC|\bar{M}) = 0.07$.

Jaká je pravděpodobnost, že žena má mutaci genu BRCA, pokud onemocněla rakovinou prsu? Nebo jinak: Jaký je podíl hereditárních případů mezi všemi případy?

$$\begin{aligned} P(M|BC) &= \frac{P(BC|M)P(M)}{P(BC|M)P(M) + P(BC|\bar{M})P(\bar{M})} = \\ &= \frac{0.87 \cdot 0.00125}{0.87 \cdot 0.00125 + 0.07 \cdot 0.99875} = 0.0153 \end{aligned}$$

Pravděpodobnost, že nemocná žena má mutaci genu BRCA je pouze 0.0153, takže podíl hereditárních případů je pouze asi 1.5%.



Senzitivita a specificita

- V medicíně se používají diagnostické testy na přítomnost nějaké choroby (kultivace výtěru - bakteriální zánět, mamograf - rakovina prsu, apod.).
- Testy obvykle nefungují stoprocentně. Občas vydají pozitivní výsledek, přestože osoba hledanou chorobu nemá - test je **nesprávně pozitivní**. Někdy naopak neodhalí nemoc, kterou testovaná osoba má - jsou **nesprávně negativní**.
- Použitelnost testu v praxi je závislá na tom, jak často dochází k těmto chybám. A také na tom, jaké mají tyto chyby důsledky.
- Označme T jev, že osoba má pozitivní test a N jev, že osoba má nemoc, kterou se snažíme odhalit. Vlastnosti testu shrnují následující míry:
 - **Senzitivita** - $P(T|N)$ pravděpodobnost, že osoba testuje pozitivně, pokud je opravdu nemocná.
 - **Specificita** - $P(\bar{T}|\bar{N})$ pravděpodobnost, že osoba testuje negativně, pokud je zdravá.

Senzitivita a specificita

- Ideální diagnostický test by měl senzitivitu i specificitu 1. Potom by odhalil všechny nemocné a žádné zdravé by neoznačil jako nemocné.
- Bohužel, obvykle tomu tak není. Kultivace bakterií bývá málo senzitivní (často bakterie není zachycena). Pokročilé zobrazovací metody (NMR, tomograf) pro detekci zhoubných nádorů bývají zase málo specifické (odhalí mnoho útvarů, které zhoubnými nádory nejsou).
- Proto je důležitá také “obrácená” podmíněná pravděpodobnost $P(N|T)$, tj. pravděpodobnost, že osoba je skutečně nemocná, pokud měla pozitivní výsledek testu.
- Použitím Bayesova vzorce dostaváme:

$$P(N|T) = \frac{P(T|N)P(N)}{P(T|N)P(N) + P(T|\bar{N})P(\bar{N})}$$

- K výpočtu potřebujeme specificitu, senzitivitu a také $P(N)$ - *prevalenci* choroby, tj. pravděpodobnost, že osoba ze sledované populace má testovanou nemoc.

Senzitivita a specificita

Příklad (Pozitivní výsledek mamografie)

Ženám starším 45 let se doporučuje každé 2 roky absolvovat preventivní mamografii. Zajímalo nás, jaká je pravděpodobnost, že žena, která má pozitivní nález na mamografii má opravdu rakovinu prsu.

Senzitivita a specificita preventivní mamografie se liší podle vyšetřované populace, přesného postupu, použitého přístroje apod. Pro náš výpočet použijeme údaje z dánské studie z let 1996-2009, která uvádí v populaci žen 50-70 let senzitivitu $P(T|N) = 0.905$, specificitu $P(\bar{T}|\bar{N}) = 0.966$ a prevalenci rakoviny prsu ve sledované populaci $P(N) = 0.0073$.

$$P(N|T) = \frac{0.905 \cdot 0.0073}{0.905 \cdot 0.0073 + 0.034 \cdot 0.9927} = 0.164$$

Pravděpodobnost, že žena s pozitivním nálezem na preventivní mamografii má opravdu rakovinu prsu je pouze 0.164.

Náhodná veličina

- **Náhodnou veličinou** rozumíme číselně vyjádřený výsledek náhodného pokusu. Značíme velkými písmeny (často X nebo Y).
- **Příklad 1:** X...počet ok na kostce. X může nabývat hodnot 1,2,3,4,5,6.
- **Příklad 2:** X...počet líců při hodu třemi mincemi. X může nabývat hodnot 0,1,2,3.
- **Příklad 3:** X...výška náhodně vybraného studenta. X může nabývat hodnoty v určitém intervalu.
- Pokud náhodná veličina může nabývat pouze konečně nebo spočetně mnoha hodnot, mluvíme o **diskrétní** náhodné veličině.
- Pokud může nabývat všech hodnot na intervalu nebo celém R, jedná se o **spojitou** náhodnou veličinu.

Rozdělení náhodné veličiny

- *Rozdělením náhodné veličiny* rozumíme pravděpodobnostní model, podle kterého se náhodná veličina chová.
- U diskrétních veličin je tento model určen, pokud víme, jakých hodnot s jakými pravděpodobnostmi může náhodná veličina nabývat.
- U spojitéch veličin bude znalost rozdělení znamenat, že dokážeme určit s jakou pravděpodobností bude náhodná veličina nabývat hodnot z jakéhokoliv intervalu v R.
- Budeme-li tedy mluvit o diskrétních nebo spojitéch rozděleních, míníme tím abstraktní modely, podle kterých se chovají náhodné veličiny. Později uvidíme, že některá rozdělení jsou častá a opakují se v různých situacích.

Distribuční funkce

- Rozdělení náhodné veličiny X se popisuje *distribuční funkcí* $F_X(t)$

$$F_X(t) = P(X \leq t), \quad t \in R.$$

- Distribuční funkce je neklesající, protože interval, jehož pravděpodobnost počítá, se zvětšuje s rostoucím t .
- Distribuční funkce je zprava spojitá - plyne z neostré nerovnosti.
- Platí $\lim_{t \rightarrow \infty} F_X(t) = 1$ a $\lim_{t \rightarrow -\infty} F_X(t) = 0$.

Distribuční funkce

Příklad (Sekretářky)

- Ve firmě pracují dvě sekretářky: Aneta a Bára. Chodí do práce pozdě s pravděpodobnostmi 0.1 a 0.2, nezávisle na sobě. Nechť X je náhodná veličina, označující počet přítomných sekretárek na začátku pracovní doby.
- X může nabývat hodnot 0, 1 nebo 2. Jedná se o diskrétní veličinu. Abychom mohli zkonstruovat distribuční funkci $F_X(t)$, potřebujeme pravděpodobnosti $P(X = k)$, pro $k = 0, 1, 2$.
- Označme A jev, že Aneta přišla pozdě a B jev, že Bára přišla pozdě. Ve výpočtech využijeme nezávislost těchto jevů, samozřejmě platí i nezávislost s jevy opačnými.
- $P(X = 0) = P(A \cap B) = P(A) \cdot P(B) = 0.1 \cdot 0.2 = 0.02$

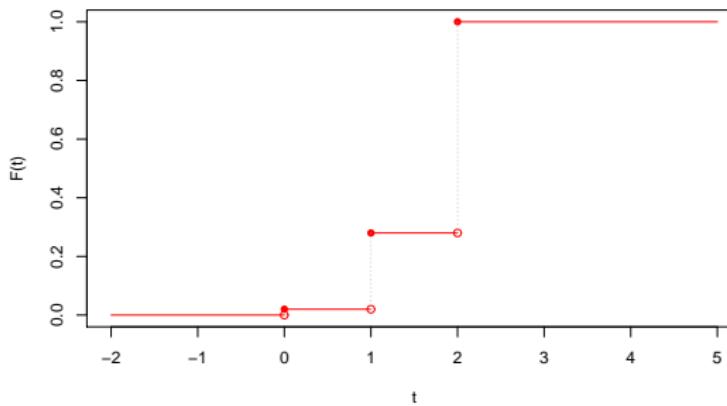
Distribuční funkce

Příklad (Sekretářky-pokračování)

- $P(X = 1) = P((A \cap \bar{B}) \cup (\bar{A} \cap B)) = P(A \cap \bar{B}) + P(\bar{A} \cap B) = P(A) \cdot P(\bar{B}) + P(\bar{A}) \cdot P(B) = 0.1 \cdot 0.8 + 0.9 \cdot 0.2 = 0.08 + 0.18 = 0.26$
- $P(X = 2) = P(\bar{A} \cap \bar{B}) = P(\bar{A}) \cdot P(\bar{B}) = 0.9 \cdot 0.8 = 0.72$
- $P(X = 0) + P(X = 1) + P(X = 2) = 0.02 + 0.26 + 0.72 = 1$,
což zjevně musí platit, protože některé z hodnot 0, 1 a 2
veličina X nabýt musí.
- Pravděpodobnosti použijeme k nákresu distribuční funkce $F_X(t)$.

Distribuční funkce

Obrázek: Distribuční funkce - Počet sekretárek



Distribuční funkce je po částech konstantní, skoky v bodech, kterých X nabývá s kladnou pravděpodobností, skok má velikost této pravděpodobnosti.

Hustota

- Rozdělení spojité náhodné veličiny nelze popsat pravděpodobnostmi možných hodnot. U spojité veličiny je možných hodnot nespočetně mnoho.
- Ve skutečnosti i spojité veličiny měříme s určitou přesností (výška v cm apod.), takže možných hodnot není nespočetně. Obvykle však s takovými veličinami pracujeme ve zidealizovaném modelu, kdy předpokládáme, že jsou skutečně spojité.
- K popisu spojitého rozdělení náhodné veličiny X používáme *hustotu* $f_X(u)$, což je taková funkce, že platí

$$F_X(t) = \int_{-\infty}^t f_X(u)du,$$

kde $F_X(t)$ je distribuční funkce náhodné veličiny X .

Hustota a distribuční funkce

- Distribuční funkce $F_X(t) = P(X \leq t)$ je integrál z hustoty na intervalu $(-\infty, t]$, tj. jedná se o plochu pod hustotou na tomto intervalu.
- Plocha pod celou hustotou je

$$\int_{-\infty}^{\infty} f_X(u)du = \lim_{t \rightarrow \infty} F_X(t) = 1.$$

- Plocha pod hustotou na intervalu (t, ∞) je tedy

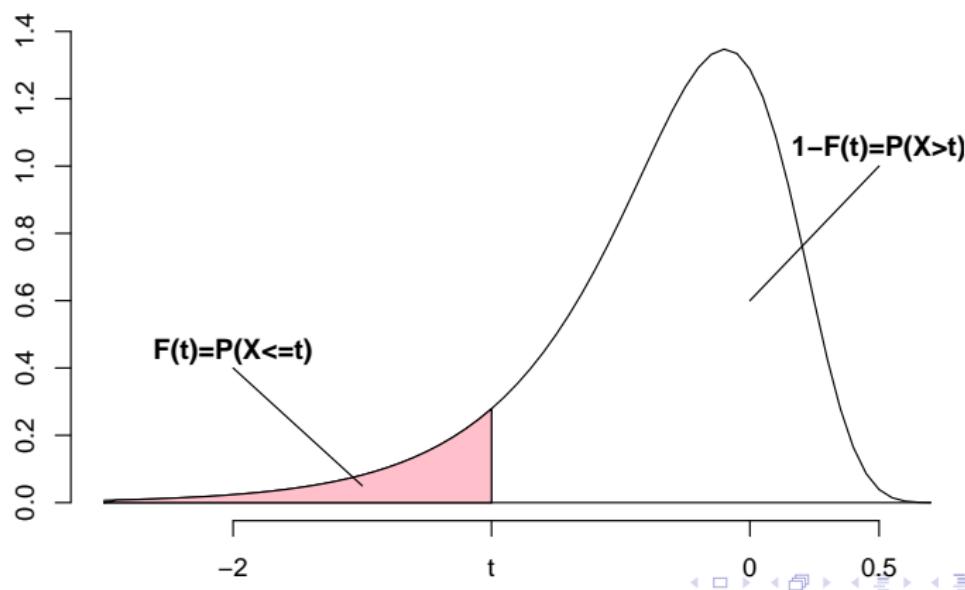
$$1 - F_X(t) = P(X > t).$$

- Hustota je vždy nezáporná, což plyne z faktu, že distribuční funkce je neklesající.
- Pokud je hustota na nějakém intervalu nulová, $F_X(t)$ je na tomto intervalu konstantní.
- Obráceně platí, že hustota je derivací distribuční funkce:

$$f_X(t) = \frac{d}{dt}F_X(t).$$

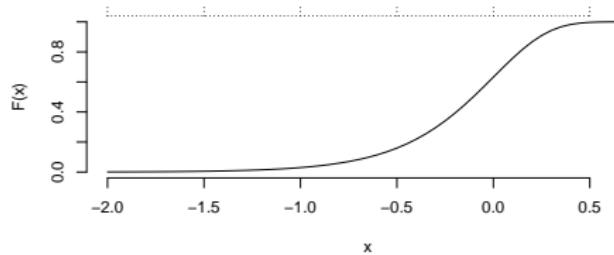
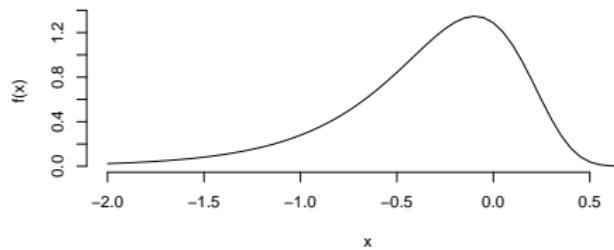
Hustota

Obrázek: Hustota spojitého rozdělení



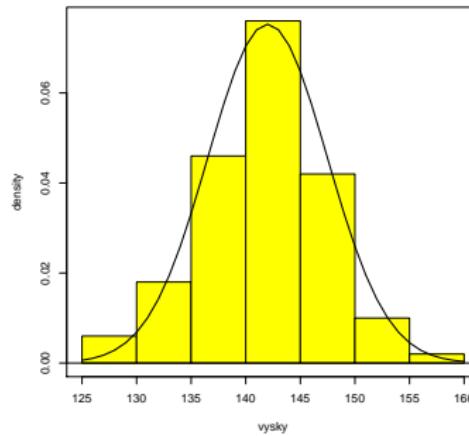
Hustota s distribuční funkce

Obrázek: Hustota a distribuční funkce spojitého rozdělení



Hustota a histogram

Obrázek: Histogram a odhad hustoty - Výšky 10letých dívek



Přesdtavu o hustotě rozdělení získáme z histogramu. Obálka idealizovaného histogramu (velké množství pozorování, velmi úzké intervaly) se bude podobat hustotě rozdělení.

Odhad distribuční funkce

- Odhadem distribuční funkce z pozorování x_1, x_2, \dots, x_n je *empirická distribuční funkce* \hat{F}_n , kterou jsme již používali k popisu dat:

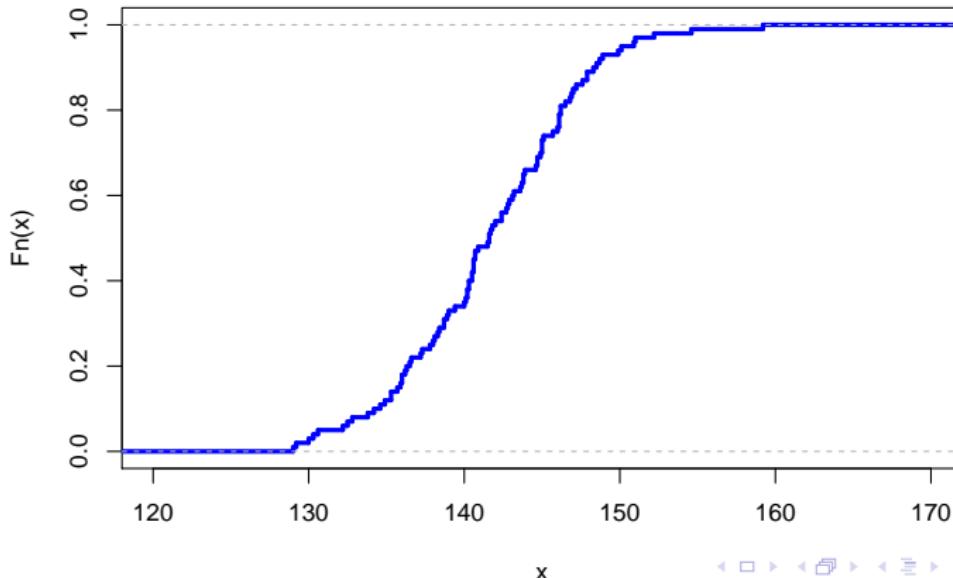
$$\hat{F}_n(t) = \frac{\sharp(x_i \leq t)}{n},$$

kde \sharp značí počet.

- Empirická distribuční funkce odhaduje v každém bodě skutečnou distribuční funkci $F(t) = P(X \leq t)$ pomocí relativní četnosti (podílu pozorování menších nebo rovných t).

Empirická distribuční funkce

Obrázek: Empirická distribuční funkce - výšky dívek

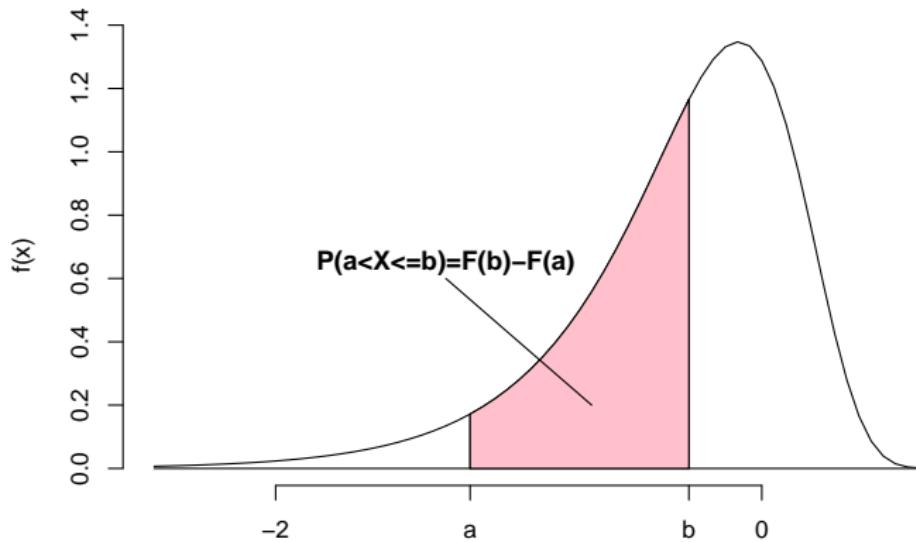


Pravděpodobnost intervalu

- U spojitého rozdělení s distribuční funkcí $F_X(t)$ a hustotou $f_X(t)$ spočítejme pravděpodobnost, že X padne do intervalu (a, b) .
- $P(X \in (a, b)) = P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F_X(b) - F_X(a)$
- Graficky si můžeme tuto pravděpodobnost představit jako plochu pod hustotou nad intervalom (a, b) .
- U spojitého rozdělení pro každý jednotlivý bod a platí $P(X = a) = 0$, nezáleží tedy na tom, zda v distribuční funkci používáme ostrou či neostrou nerovnost.

Pravděpodobnost intervalu

Obrázek: Pravděpodobnost intervalu - graficky



Přednáška 4 (9.3.2021) - obsah

- Populační kvantily
- Střední hodnota
- Populační rozptyl
- Sdružené a marginální rozdělení
- Populační kovariance, nezávislost náhodných veličin, populační korelační koeficient
- Střední hodnota a rozptyl součtu náhodných veličin
- Diskrétní rozdělení (alternativní, binomické, Poissonovo, hypergeometrické)

Populační medián

- Známe-li distribuční funkci náhodné veličiny, můžeme najít *kvantily* - teoretické (nebo také populační) protějšky percentilů.
- Pro populační medián $\tilde{\mu}$ by mělo platit:

$$F_X(\tilde{\mu}) = P(X \leq \tilde{\mu}) = 0.5$$

- U spojitých rozdělení takové číslo vždy existuje, distribuční funkce je spojitá a neklesající, takže bod, kde $F_X(t)$ překračuje 0.5 bude vyhovovat.
- Může ovšem existovat více bodů, které rovnici vyhovují. Pokud distribuční funkce právě neroste, vyhovuje požadavku celý interval bodů.
- U diskrétních rozdělení takový bod vůbec nemusí existovat, distribuční funkce může jednu polovinu přeskočit.

Populační medián a kvantily

- Abychom se vyhnuli popsaným problémům, definujeme *populační medián* $\tilde{\mu}$ jako nejmenší takové číslo, pro které platí:

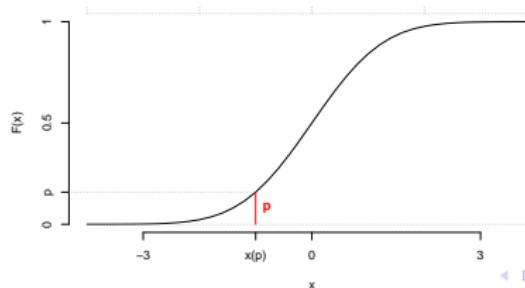
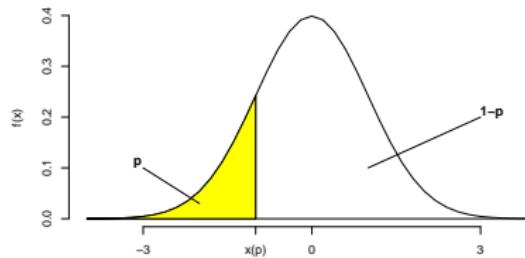
$$F_X(\tilde{\mu}) = P(X \leq \tilde{\mu}) \geq 0.5$$

- Podobnou definici můžeme použít pro obecný *populační kvantil* $x(p)$, $p \in (0, 1)$. $x(p)$ je nejmenší takové číslo, které splňuje

$$F_X(x(p)) = P(X \leq x(p)) \geq p$$

Kvantily

Obrázek: Kvantil - grafická představa



Střední hodnota

- *Střední hodnota* náhodné veličiny X je teoretický (populační) protějšek výběrového průměru.
- Budeme-li mnohokrát opakovaně provádět náhodný pokus, jehož výsledkem je X a pak spočítáme průměr, bude tento průměr blízko střední hodnoty.
- Střední hodnota náhodné veličiny X se označuje EX (z anglického expectation), případně μ_X .
- EX je teoretická hodnota, nenáhodná. V praxi může být neznámá.
- EX lze spočítat, pokud známe teoretické rozdělení náhodné veličiny.

Střední hodnota

- Pro diskrétní náhodnou veličinu, která může nabývat hodnot x_1, x_2, \dots je střední hodnota

$$\begin{aligned} EX &= x_1 \cdot P(X = x_1) + x_2 \cdot P(X = x_2) + \dots = \\ &= \sum_i x_i P(X = x_i) \end{aligned}$$

- Jedná se o vážený průměr, kdy možné hodnoty x_i vážíme jejich pravděpodobnostmi. Součet vah skutečně splňuje $\sum_i P(X = x_i) = 1$.
- Pro spojitou veličinu platí:

$$EX = \int_{-\infty}^{\infty} u f_X(u) du,$$

funkci vah zde přebírá hustota.

Střední hodnota

Příklad (Sekretářky)

- Firma má dvě sekretářky, které chodí pozdě do práce nezávisle na sobě, s pravděpodobnostmi 0.1 a 0.2.
- Minule jsme zjistili, že pokud X je počet sekretárek ve firmě na počátku pracovní doby, platí $P(X = 0) = 0.02$,
 $P(X = 1) = 0.26$ a $P(X = 2) = 0.72$
- Spočítejme střední hodnotu X :
$$EX = 0 \cdot P(X = 0) + 1 \cdot P(X = 1) + 2 \cdot P(X = 2) =$$
$$1 \cdot 0.26 + 2 \cdot 0.72 = 1.7$$
- Vidíme, že EX se nemusí rovnat možné hodnotě náhodné veličiny X . Zajisté nikdy neuvidíme ve firmě 1.7 sekretářky. Budeme-li však sledovat situaci opakováně mnohokrát a spočítáme-li průměr, měli bychom dostat číslo blízké 1.7.

Střední hodnota funkce náhodné veličiny

- Zřejmě platí $E(a + bX) = a + bEX$. Střední hodnota se chová rozumně při posunutí a změně měřítka.
- Obecněji, pokud náhodná veličina Y je funkcí X , tj. $Y = g(X)$, bude platit:

$$EY = Eg(X) = \sum_i g(x_i)P(X = x_i), \text{ pro } X \text{ diskrétní}$$

$$EY = Eg(X) = \int_{-\infty}^{\infty} g(u)f_X(u)du, \text{ pro } X \text{ spojítou.}$$

Populační rozptyl

- *Populační rozptyl* je teoretickou obdobou výběrového rozptylu. Definujeme ho jako střední hodnotou druhých mocnin odchylek od střední hodnoty:

$$\sigma_X^2 = E(X - \mu_X)^2$$

- Populační rozptyl značíme buď σ_X^2 nebo $\text{var}X$ (z anglického variance)
- Pro diskrétní náhodnou veličinu je rozptyl váženým průměrem druhých mocnin odchylek od střední hodnoty:

$$\sigma_X^2 = \text{var}X = \sum_i (x_i - \mu_X)^2 \cdot P(X = x_i)$$

- Pro spojitou náhodnou veličinu s hustotou $f(u)$ spočítáme rozptyl jako integrál:

$$\sigma_X^2 = \text{var}X = \int_{-\infty}^{\infty} (u - \mu_X)^2 f(u) du$$

Populační rozptyl

- *Populační směrodatná odchylka* je odmocninou z rozptylu:

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

- Zřejmě platí

$$\text{var}(a + bX) = \text{var}(bX) = b^2 \cdot \text{var}X$$

- Posunutí tedy nemění rozptyl a směrodatnou odchylku.
- Při změně měřítka se rozptyl mění s druhou mocninou změny měřítka.
- Pro směrodatnou odchylku platí:

$$\sigma_{a+bX} = |b| \cdot \sigma_X$$

Populační rozptyl

Příklad (Sekretářky)

Spočítejme rozptyl počtu sekretárek na počátku pracovní doby (náhodná veličina X). Použijeme již spočítanou střední hodnotu $EX = 1.7$:

$$\begin{aligned}\sigma_X^2 &= \sum_{i=0}^2 (i - \mu_X)^2 P(X = i) = \\ &= (0 - 1.7)^2 \cdot 0.02 + (1 - 1.7)^2 \cdot 0.26 + (2 - 1.7)^2 \cdot 0.72 = 0.25\end{aligned}$$

Sdružené a marginální rozdělení

- Zatím jsme zkoumali, jak se chovají jednotlivé náhodné veličiny. Často nás ovšem zajímají souvislosti mezi náhodnými veličinami. K jejich popisu potřebujeme vědět, jak se chovají dvě nebo více náhodných veličin současně.
- Model, který popisuje chování veličin X_1, X_2, \dots, X_k současně nazveme *sdruženým rozdělením* náhodného vektoru (X_1, X_2, \dots, X_k) .
- Jednotlivá rozdělení veličin nazveme *marginálními rozděleními*.
- Pokud známe sdružené rozdělení vektoru (X_1, X_2, \dots, X_k) , marginální rozdělení jednotlivých veličin jsou jím již určena.
- Obráceně to ale neplatí. Známe-li pouze marginální rozdělení jednotlivých veličin, chybí nám znalost souvislostí mezi veličinami a sdružené rozdělení určit nemůžeme.

Sdružené a marginální rozdělení

- Pro diskrétní náhodné veličiny (X, Y) je sdružené rozdělení popsáno pravděpodobnostmi $P(X = x_i \cap Y = y_j)$ pro všechna x_i , kterých X nabývá s kladnou pravděpodobností a všechna y_j , kterých Y nabývá s kladnou pravděpodobností.
- Pro spojité náhodné veličiny (X, Y) je sdružené rozdělení popsáno sdruženou hustotou $f_{X,Y}(x, y)$.
- Marginální rozdělení ze sdruženého lze spočítat. Pro diskrétní veličiny stačí sečít přes všechny možné hodnoty druhé veličiny:

$$P(X = x_i) = \sum_j P(X = x_i \cap Y = y_j)$$

$$P(Y = y_j) = \sum_i P(X = x_i \cap Y = y_j)$$

U spojitých veličin stačí integrovat přes jednu z proměnných:

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy, \quad f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx$$

Podmíněné rozdělení

- Známe-li sdružené rozdělení veličin X a Y , můžeme spočítat také *podmíněné rozdělení* jedné z veličin při dané hodnotě veličiny druhé.
- Pro diskrétní rozdělení platí:

$$P(X = x_i | Y = y_j) = \frac{P(X = x_i \cap Y = y_j)}{P(Y = y_j)}$$

- Povšimněme si, že tento vztah umožňuje spočítat sdružené rozdělení, pokud známe marginální rozdělení veličiny Y a podmíněné rozdělení veličiny X při všech možných hodnotách veličiny Y .

Sdružené rozdělení veličin

Příklad (Sekretářky-úplně jinak)

- Označme nyní v příkladu se sekretářkami X náhodnou veličinu, která nabývá pouze hodnot 0 (pokud sekretářka Aneta přijde pozdě) a 1 (pokud Aneta bude v práci včas). Podobně, Y bude nula-jedničková veličina monitorující příchody Báry.
- Platí, že $P(X = 1) = 0.9$, $P(X = 0) = 0.1$, $P(Y = 1) = 0.8$ a $P(Y = 0) = 0.2$, tím je dáno marginální rozdělení X a Y .
- Zrušme nyní předpoklad, že sekretářky chodí pozdě nezávisle na sobě a řekněme, že o tom nic nevíme. To znamená, že neznáme sdružené pravděpodobnosti $P(X = 0 \cap Y = 0)$ apod.

Sdružené rozdělení veličin

Příklad (Sekretářky-úplně jinak-pokračování)

- *Může se stát, že jezdí do práce stejným vlakem, a pokud chybí Aneta, určitě chybí i Bára, tedy $P(X = 0 \cap Y = 0) = 0.1$*
- *Nebo chodí pozdě zcela plánovaně a vždy se předem dohodnou, aby aspoň jedna z nich v práci byla, tedy $P(X = 0 \cap Y = 0) = 0$*
- *Nebo chodí pozdě zcela nezávisle na sobě, jak jsme dříve předpokládali a*
$$P(X = 0 \cap Y = 0) = P(X = 0) \cdot P(Y = 0) = 0.02$$
- *Ve všech těchto případech (a ještě spoustě jiných) mohou být marginální rozdělení X a Y stejná. Sdružené rozdělení se ale liší.*

Populační kovariance

- Pokud známe sdružené rozdělení X a Y , můžeme spočítat *populační kovariaci* σ_{XY} (také $\text{cov}(X, Y)$):

$$\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)]$$

- σ_{XY} je teoretickým protějškem výběrové kovariance s_{XY} , kterou jsme používali ke zkoumání závislosti dvou výběrů.
- Pro diskrétně rozdělené náhodné veličiny X a Y bude platit
$$\sigma_{XY} = \sum_i \sum_j (x_i - \mu_X)(y_j - \mu_Y) \cdot P(X = x_i \cap Y = y_j),$$
kde sčítáme přes všechny kombinace hodnot x_i a y_j , jakých X a Y nabývají s kladnou pravděpodobností.
- Je vidět, že $\sigma_{XX} = \sigma_X^2$.

Nezávislost náhodných veličin

- Pokud víme, že se veličiny X a Y chovají zcela nezávisle na sobě, pak nám marginální rozdělení stačí a sdružené snadno dopočítáme.
- **Sekretářky:** Pokud chodí pozdě nezávisle na sobě, pak $P(X = 0 \cap Y = 0) = P(X = 0) \cdot P(Y = 0) = 0.02$ a podobně dopočítáme ostatní sdružené pravděpodobnosti.
- **Hod dvěma kostkami:** Ve všech našich výpočtech jsme předpokládali, že to, co padlo na modré nesouvisí s tím, co padlo na červené. Tím pádem (pokud jsou kostky pravidelné) všechny uspořádané dvojice jsou stejně pravděpodobné a mohli jsme používat klasickou definici pravděpodobnosti.

Nezávislost náhodných veličin

- Co přesně myslíme tím, když řekneme, že se veličiny X a Y chovají nezávisle?
- *Veličiny X a Y jsou nezávislé*, pokud všechny jevy, které jsou tvrzeními o X jsou nezávislé na všech jevech, které jsou tvrzeními o Y . (Co je to nezávislost jevů, to již víme.)
- Pro diskrétní náhodné veličiny platí, že jsou nezávislé právě tehdy, když platí:

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

pro všechny kombinace hodnot x_i a y_j , jakých X a Y nabývají s nenulovou pravděpodobností.

Populační kovariance a nezávislost

- Pokud jsou X a Y nezávislé náhodné veličiny, pak platí:

$$EXY = EX \cdot EY$$

- Tvrzení snadno ukážeme pro diskrétní veličiny:

$$\begin{aligned} EXY &= \sum_i \sum_j x_i y_j \cdot P(X = x_i \cap Y = y_j) = \\ &= \sum_i \sum_j x_i y_j \cdot P(X = x_i) \cdot P(Y = y_j) = \\ &= \sum_i x_i P(X = x_i) \cdot \sum_j y_j P(Y = y_j) = EX \cdot EY \end{aligned}$$

- Pro kovarianci nezávislých veličin pak platí

$$\begin{aligned} \sigma_{XY} &= E(X - \mu_X)(Y - \mu_Y) = \\ &= E(XY - X\mu_Y - \mu_X Y + \mu_X \mu_Y) = \\ &= EXY - EX\mu_Y - \mu_X EY + \mu_X \mu_Y = 0 \end{aligned}$$

- Jsou-li X a Y nezávislé, pak $\sigma_{XY} = 0$.

- **Pozor!** Opačná implikace neplatí.

Populační korelační koeficient

- Odečtením střední hodnoty a vydělením směrodatnou odchylkou dostaneme *normovanou* náhodnou veličinu.

$$Z = \frac{X - \mu_X}{\sigma_X}.$$

- Z je populační obdoba z-skóru a platí $\mu_Z = 0$, $\sigma_Z = 1$.
- Populační korelační koeficient* ρ_{XY} definujeme jako kovarianci normovaných veličin:

$$\rho_{XY} = cov\left(\frac{X - \mu_X}{\sigma_X}, \frac{Y - \mu_Y}{\sigma_Y}\right) = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y}$$

- Vždy platí $-1 \leq \rho_{XY} \leq 1$.
- Pokud jsou X a Y nezávislé, platí $\sigma_{XY} = 0$, tedy i $\rho_{XY} = 0$.
- Opačná implikace neplatí, existují i závislé náhodné veličiny, pro které $\rho_{XY} = 0$.

Populační šikmost a špičatost

- Normovanou náhodnou veličinu lze použít k výpočtu populačních charakteristik, které nezávisí na střední hodnotě a rozptylu, např. šiknosti a špičatosti.
- Populační šikmost* je

$$\gamma_1 = E \left(\frac{X - \mu_X}{\sigma_X} \right)^3 = \frac{E(X - \mu_X)^3}{\sigma_X^3}.$$

- Populační špičatost* je

$$\gamma_2 = E \left(\frac{X - \mu_X}{\sigma_X} \right)^4 - 3 = \frac{E(X - \mu_X)^4}{\sigma_X^4} - 3.$$

Vlastnosti střední hodnoty a rozptylu

- Mějme dvě náhodné veličiny X a Y , jejichž střední hodnoty a rozptyly známe. Zkusme spočítat střední hodnotu a rozptyl jejich součtu.
- Z vlastností váženého průměru a integrálu lze snadno ukázat:

$$E(X + Y) = EX + EY.$$

- Rozptyl $X + Y$ splňuje:

$$\begin{aligned} \text{var}(X + Y) &= E(X + Y - \mu_X - \mu_Y)^2 = \\ &= E[(X - \mu_X) + (Y - \mu_Y)]^2 = \\ &= E(X - \mu_X)^2 + E(Y - \mu_Y)^2 + 2 \cdot E[(X - \mu_X)(Y - \mu_Y)] = \\ &= \sigma_X^2 + \sigma_Y^2 + 2 \cdot \sigma_{XY} \end{aligned}$$

- Vidíme, že pouze pokud $\sigma_{XY} = 0$, bude platit

$$\text{var}(X + Y) = \sigma_X^2 + \sigma_Y^2.$$

Rozptyl součtu nezávislých veličin

- Pokud jsou X a Y nezávislé, pak je kovariance nulová a platí:

$$\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2 + 2\sigma_{XY} = \sigma_X^2 + \sigma_Y^2,$$

- Na rozdíl od středních hodnot, kde tato vlastnost platí vždy, u rozptylu potřebujeme nezávislost X a Y .
- Rozdíl nezávislých náhodných veličin má stejný rozptyl jako jejich součet:

$$\begin{aligned}\sigma_{X-Y}^2 &= E(X - Y - (\mu_X - \mu_Y))^2 = \\&= E[(X - \mu_X) - (Y - \mu_Y)]^2 = \\&= E(X - \mu_X)^2 + E(Y - \mu_Y)^2 - 2E[(X - \mu_X)(Y - \mu_Y)] = \\&= \sigma_X^2 + \sigma_Y^2\end{aligned}$$

Sdružené a marginální rozdělení

Příklad (Diabetes)

V jisté populaci seniorů je pravděpodobnost vysokého krevního tlaku (VKT) 0.5 a pravděpodobnost diabetu (dia) 0.3.

Pravděpodobnost, že senior má zároveň VKT i dia je 0.25.

Označme X indikátor VKT ($X = 1$, pokud má senior VKT a $X = 0$, pokud nemá VKT). Podobně Y nechť je indikátor diabetu.

Marginální rozdělení X a Y jsou zřejmá:

$$P(X = 1) = P(X = 0) = 0.5,$$

$$P(Y = 1) = 0.3, \quad P(Y = 0) = 0.7.$$

Střední hodnoty a rozptyly X a Y:

$$\mu_X = 1 \cdot P(X = 1) + 0 \cdot P(X = 0) = 0.5$$

$$\mu_Y = 0.3$$

$$\sigma_X^2 = (1 - 0.5)^2 \cdot 0.5 + (0 - 0.5)^2 \cdot 0.5 = 0.25$$

$$\sigma_Y^2 = 0.21$$

Sdružené a marginální rozdělení

Příklad (Diabetes - pokračování)

Sdružené rozdělení získáme z informace, že

$P(X = 1 \cap Y = 1) = 0.25$, doplněním následující tabulky:

		$Y (\text{dia})$		marginální
		1	0	
$X (\text{VKT})$	1	0.25	?	0.5
	0	?	?	0.5
marginální		0.3	0.7	1

		$Y (\text{dia})$		marginální
		1	0	
$X (\text{VKT})$	1	0.25	0.25	0.5
	0	0.05	0.45	0.5
marginální		0.3	0.7	1

Sdružené a marginální rozdělení

Příklad (Diabetes - pokračování)

- Jsou X a Y nezávislé? Pokud by byly, muselo by například platit $P(X = 1 \cap Y = 1) = P(X = 1) \cdot P(Y = 1)$
To ale neplatí, $0.25 \neq 0.5 \cdot 0.3 = 0.15$, X a Y jsou závislé.
Situace, kdy pacient trpí oběma chorobami je častější, než pokud by byly veličiny nezávislé.
- Spočítejme střední hodnotu $X + Y$ (počet chorob).
 $\mu_{X+Y} = \mu_X + \mu_Y = 0.5 + 0.3 = 0.8$
- Kovariance X a Y je

$$\begin{aligned}\sigma_{XY} &= \sum_i \sum_j (x_i - \mu_X)(y_j - \mu_Y)P(X = x_i \cap Y = y_j) = \\ &= (1 - 0.5)(1 - 0.3) \cdot 0.25 + (1 - 0.5)(0 - 0.3) \cdot 0.25 + \\ &+ (0 - 0.5)(1 - 0.3) \cdot 0.05 + (0 - 0.5)(0 - 0.3) \cdot 0.45 = 0.1\end{aligned}$$
- Rozptyl $X + Y$
 $\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2 + 2 \cdot \sigma_{XY} = 0.25 + 0.21 + 2 \cdot 0.1 = 0.66$

Alternativní rozdělení

- Náhodná veličina X má *alternativní rozdělení* (nebo také Bernoulliovo nebo nula-jedničkové), pokud nabývá pouze hodnot 1 a 0 a pravděpodobnostmi $P(X = 1) = \pi$, $P(X = 0) = 1 - \pi$, pro nějaké $\pi \in (0, 1)$.
- Alternativní rozdělení má jediný parametr π . Fakt, že X má alternativní rozdělení s parametrem π můžeme zapsat třeba takto $X \sim Alt(\pi)$.
- Střední hodnota a rozptyl:
$$\mu_X = 1 \cdot P(X = 1) + 0 \cdot P(X = 0) = P(X = 1) = \pi$$
$$\sigma_X^2 = (1 - \pi)^2 \cdot P(X = 1) + (0 - \pi)^2 \cdot P(X = 0) = \\ = (1 - \pi)^2 \pi + \pi^2 (1 - \pi) = \pi(1 - \pi)$$
- Příklady: Házení mincí rub/líc. Sekretářka přítomna/nepřítomna. Pacient má/nemá diabetes.

Binomické rozdělení

- *Binomické rozdělení* vznikne, pokud opakujeme nula-jedničkový (alternativní) pokus n-krát, každý pokus nezávisle na ostatních, a pak sečteme počet jedniček (úspěšných pokusů).
- Binomické rozdělení má 2 parametry: počet pokusů n a pravděpodobnost úspěchu π , rozdělení se značí $Bi(n, \pi)$
- **Příklad:** Hodím deseti mincemi. Počet líců, které padly bude náhodná veličina X . X může nabýt hodnot $0, 1, 2, \dots, 10$. Pokud jsou mince správné, padne na každé minci líc s pravděpodobností 0.5. $X \sim Bi(10, 0.5)$.
- Alternativní rozdělení je speciálním případem binomického: $Alt(\pi) = Bi(1, \pi)$.
- Binomické rozdělení je součtem nezávislých alternativních. Pokud $X_i \sim Alt(\pi)$, nezávislé pro $i = 1, 2, \dots, n$, pak $X = X_1 + X_2 + \dots + X_n \sim Bi(n, \pi)$.

Binomické rozdělení

- Střední hodnotu $X \sim Bi(n, \pi)$ spočítáme jako součet středních hodnot alternativních veličin:
$$EX = E(X_1 + X_2 + \dots + X_n) = EX_1 + EX_2 + \dots + EX_n = n\pi$$
- Pro výpočet rozptylu můžeme také použít součet rozptylů X_i , protože víme, že X_1, X_2, \dots, X_n jsou nezávislé.
$$\sigma_X^2 = \sigma_{X_1}^2 + \sigma_{X_2}^2 + \dots + \sigma_{X_n}^2 = n\pi(1 - \pi)$$
- Pomocí kombinatoriky lze odvodit, že

$$P(X = k) = \binom{n}{k} \pi^k (1 - \pi)^{n-k},$$

pro $k = 0, 1, \dots, n$

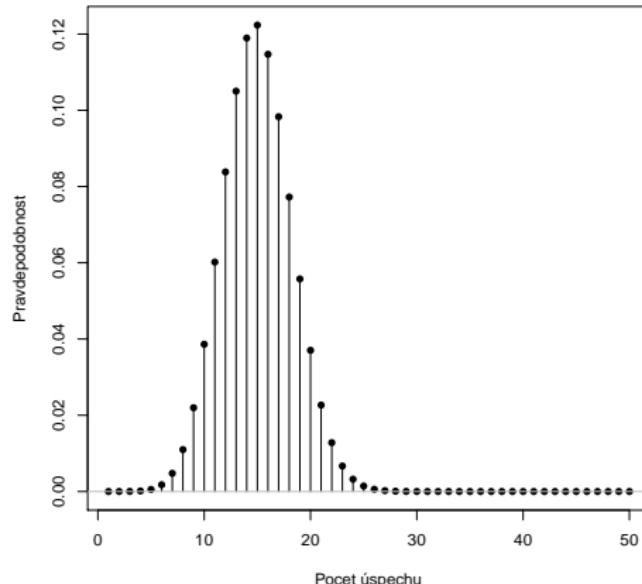
Binomické rozdělení

Příklad (Diabetes)

- Ve sledované populaci je pravděpodobnost diabetu 0.3. Náhodně vyberu 50 lidí z této populace. Počet lidí s cukrovkou ve svém výběru označím X .
- $X \sim Bi(50, 0.3)$. Střední hodnota je $EX = 50 \cdot 0.3 = 15$.
- Očekávali bychom, že ve výběru bude okolo 15 diabetiků. Pravděpodobnost, že jich bude přesně 15 ale není vysoká, $P(X = 15) = 0.122$.
- Vyneseme-li pravděpodobnosti $P(X = k)$ do grafu, vidíme, že pravděpodobnosti velmi nízkého nebo vysokého počtu diabetiků ve výběru jsou velmi malé. Např.
 $1 - F_X(25) = P(X > 25) \doteq 0.0009$.

Binomické rozdělení

Obrázek: Pravděpodobnosti $\text{Bi}(50, 0.3)$ -počet diabetiků



Poissonovo rozdělení

- Poissonovo rozdělení se používá k popisu výskytu řídkých jevů (počet jevů za časový interval, v jednotkovém objemu apod.)
- Rozdělení předpokládá, že počet výskytů v jednom časovém intervalu (objemu aj.) nesouvisí s výskyty v jiných intervalech.
- Rozdělení má jediný parametr λ , označujeme ho $Po(\lambda)$. Veličina $X \sim Po(\lambda)$ nabývá celých nezáporných hodnot s pravděpodobnostmi:

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \dots$$

- Platí $\mu_X = \lambda$ a $\sigma_X^2 = \lambda$.
- Binomické rozdělení s vysokým n a malým π lze approximovat $Po(n\pi)$.

Poissonovo a binomické rozdělení

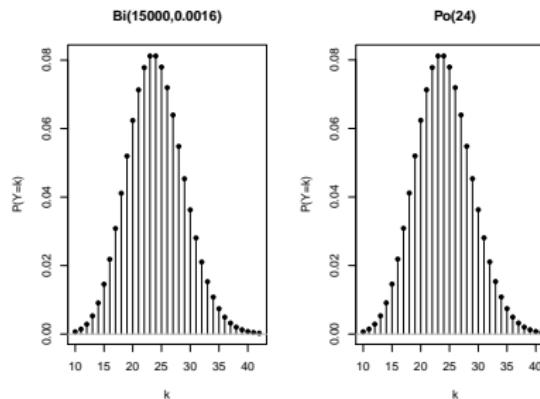
Příklad (Počet dětí narozených s rozštěpem patra)

- Pravděpodobnost, že se dítě narodí s rozštěpem patra je 0.0016. Specializované pracoviště, které se věnuje plastické chirurgii zajímají každoroční počty dětí, které v Praze narodí s touto vadou.
- Pokud známe celkový počet dětí n , které se v Praze narodí, můžeme použít binomické rozdělení $Bi(n, 0.0016)$. Každé z narozených dětí si můžeme představit jako jeden alternativní pokus, kdy s pravděpodobností 0.0016 bude mít dítě rozštěp patra. V Praze se narodí kolem 15 000 dětí ročně.
- Vzhledem k tomu, že n je velmi vysoké a π malé, můžeme approximovat Poissonovým rozdělením s $\lambda = 15000 \cdot 0.0016 = 24$.
- Lékaře zajímá, kolik maximálně (s pravděpodobností 0.95) mohou v Praze každý rok očekávat dětí s rozštěpem. Pro $Bi(15000, 0.0016)$ je to 32 ($P(Y \leq 32) = 0.9534$), pokud approximujeme pomocí $Po(24)$, je to také 32 ($P(Y \leq 32) = 0.9533$).



Poissonovo a binomické rozdělení

Obrázek: Porovnání binomického a Poissonova rozdělení



Příklad (Počet dětí narozených s rozštěpem patra)

V roce 2018 se nakonec v Praze narodilo více dětí (15540). Aproximace pomocí $Po(24)$ však docela vyhovuje, pro $Bi(15540, 0.0016)$ je $P(Y \leq 32) = 0.9325$.

Hypergeometrické rozdělení

- Vybíráme-li náhodně k subjektů z celkového počtu n , kde a subjektů je označených a $n - a$ neoznačených, pak počet označených mezi taženými má hypergeometrické rozdělení $Hyp(n, a, k)$.
- Pokud $Y \sim Hyp(n, a, k)$, pak

$$P(Y = m) = \frac{\binom{a}{m} \binom{n-a}{k-m}}{\binom{n}{k}},$$

pro m celá, splňující

$$\max(0, k + a - n) \leq m \leq \min(k, a)$$

- Jedná o již známé rozdělení vylovených označených ryb z rybníka.

Přednáška 5 (16.3.2021) - obsah

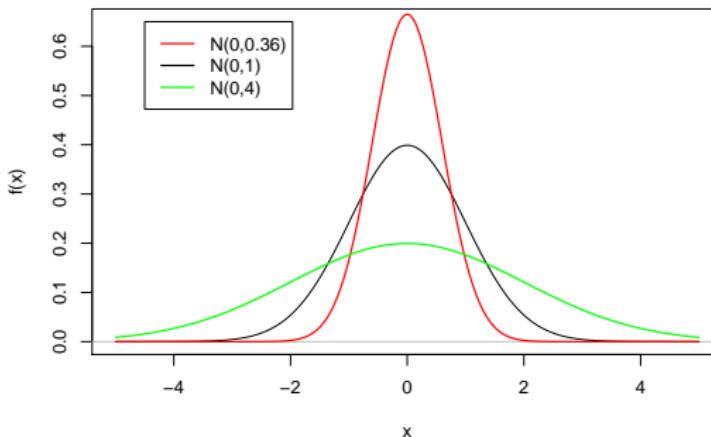
- Normální rozdělení
- Některá další spojitá rozdělení (log-normální, t -rozdělení, χ^2 -rozdělení, F -rozdělení)
- Náhodný výběr
- Průměr z náhodného výběru
- Centrální limitní věta
- Interval spolehlivosti pro střední hodnotu normálního rozdělení

Normální rozdělení

- *Normální rozdělení* je spojité rozdělení se dvěma parametry: μ (střední hodnota) a σ^2 (rozptyl), značíme $X \sim N(\mu, \sigma^2)$.
- Hustota normálního rozdělení $N(\mu, \sigma^2)$ je
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \text{ pro } x \in R.$$
- Hustota $f(x)$ je symetrická kolem střední hodnoty, kladná na celém R , ale mimo určitý interval je velmi blízká 0.
- *Normované normální rozdělení* je normální rozdělení s nulovou střední hodnotou a jednotkovým rozptylem $N(0, 1)$.

Hustota normálního rozdělení

Obrázek: Hustota normálního rozdělení - různé rozptyly



Změníme-li střední hodnotu, tvar hustoty se nezmění, hustota se pouze posune. Změníme-li rozptyl, tvar se změní.

Vlastnosti normálního rozdělení

- Hustota $N(\mu, \sigma^2)$ dosahuje maxima v bodě μ a maximální hodnota je $f(\mu) = \frac{1}{\sqrt{2\pi}\sigma} \doteq \frac{0.4}{\sigma}$.
- Distribuční funkce $\Phi(x)$ normovaného normálního rozdělení je tabelována.
- Distribuční funkci jakéhokoliv jiného normálního rozdělení lze odvodit z $\Phi(x)$. Platí totiž, že pokud $X \sim N(\mu, \sigma^2)$, potom

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

- Tabelovány jsou také kvantily $z(p)$ normovaného normálního rozdělení, které je možné použít k výpočtu kvantilů jakéhokoliv jiného normálního rozdělení.

Distribuční funkce normálního rozdělení

Příklad (Výšky desetiletých dívek)

Řekněme, že víme, že výšky desetiletých dívek mají normální rozdělení $N(141, 36)$. Jaká je pravděpodobnost, že náhodně vybraná dívka nebude větší než 135 cm?

$X \sim N(141, 36)$. Hledáme $P(X \leq 135) = F_X(135)$.

Použijeme hodnoty $\Phi(x)$, distribuční funkce $N(0, 1)$, která je tabelována:

$$P(X \leq 135) = P\left(\frac{X-141}{6} \leq \frac{135-141}{6}\right) = P(Y \leq -1) = \Phi(-1) \doteq 0.16, \text{ (použili jsme, že } Y \sim N(0, 1)\text{)}.$$

Statistický software často umí počítat hodnoty distribučních funkcí jakéhokoliv normálního rozdělení sám, obvykle nemusíme ani přepočítávat na normované normální.

Kvantily normálního rozdělení

Příklad (Výšky desetiletých dívek)

Lékaři chtějí identifikovat děti s nenormálním vzhledem kvůli hormonálním poruchám. Chtějí 3 % nejvyšších a 3 % nejnižších dětí v populaci posílat na vyšetření. Jak vysoká nebo jak malá musí být desetiletá dívka, aby byla odeslaná na vyšetření?

Hledáme $x(0.03)$ a $x(0.97)$, kvantily rozdělení $N(141, 36)$. Kvantily normovaného normálního rozdělení $z(p)$ jsou tabelovány, čehož využijeme:

$$\begin{aligned} 0.03 &= P(X \leq x(0.03)) = P\left(\frac{X-141}{6} \leq \frac{x(0.03)-141}{6}\right) = \\ &= P(Y \leq \frac{x(0.03)-141}{6}). \end{aligned}$$

Protože $Y \sim N(0, 1)$, musí platit: $\frac{x(0.03)-141}{6} = z(0.03)$, což je tabelovaný kvantil $N(0, 1)$, $z(0.03) = -1.8808$

Potom $x(0.03) = 6 \cdot (-1.8808) + 141 \doteq 129.72$.

Kvantily normálního rozdělení

Příklad (Výšky desetiletých dívek- pokračování)

Podobně spočítáme i druhý kvantil: $x(0.97) \doteq 152.28$.

Pokud měříme s přesností na cm, asi bychom poslali na další vyšetření dívky vyšší než 152 cm a dívky nižší než 130 cm.

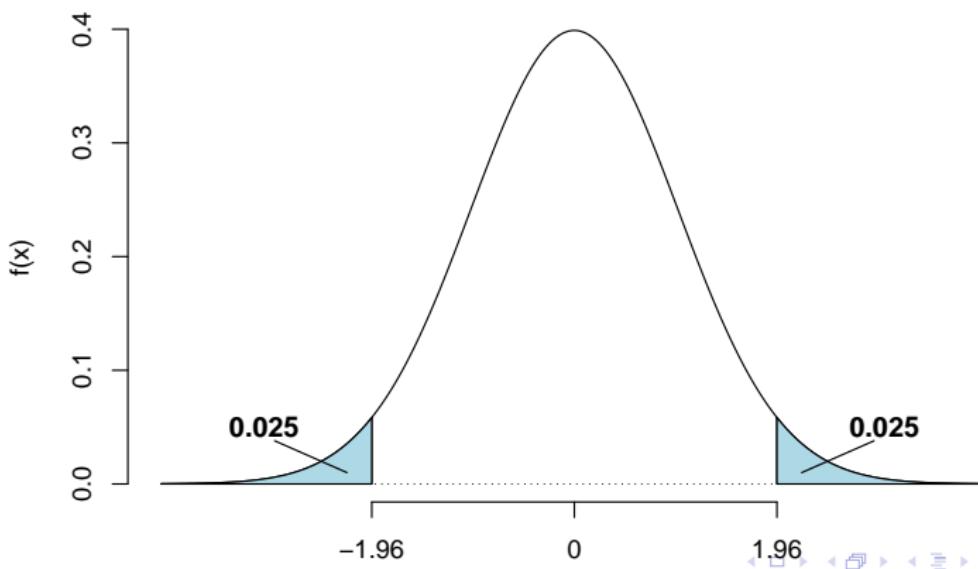
Podobně jako pro hodnoty distribuční funkce, hledání kvantilů různých normálních rozdělení bývá s dnešním statistickým softwarem snazší než s tabulkami.

Kvantily normálního rozdělení

- Jak záhy uvidíme, některé kvantily normovaného normálního rozdělení jsou ve statistice velmi používané. Nezaškodí jich několik uvést.
 - $z(0.95) = 1.645$.
Pro $Y \sim N(0, 1)$ tedy platí $P(Y \leq 1.645) = 0.95$.
Díky symetrii Y kolem 0 bude platit také
 $P(Y \geq -1.645) = 0.95$, tedy $P(|Y| \leq 1.645) = 0.9$
Pro $X \sim N(\mu, \sigma^2)$ pak bude platit
 $P(|X - \mu| \leq 1.645 \cdot \sigma) = 0.9$.
 - $z(0.975) = 1.960$. Tento kvantil je možná nejpoužívanější.
Platí, že $P(|X - \mu| \leq 1.96 \cdot \sigma) = 0.95$
 - Další důležité kvantily: $z(0.99) = 2.326$ a $z(0.995) = 2.576$

Kvantily normálního rozdělení

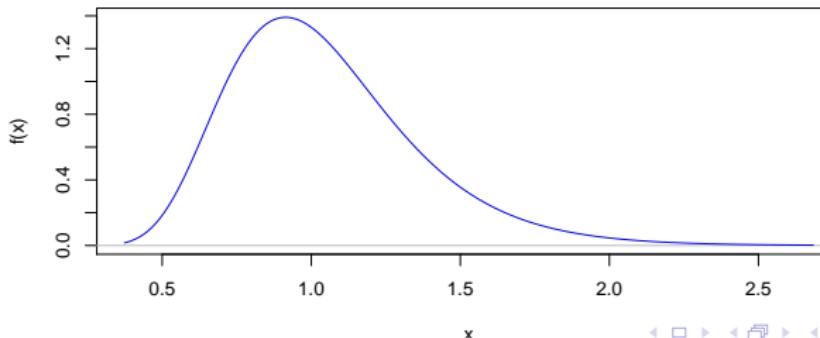
Obrázek: Kvantily $z(0.025)$ a $z(0.975)$ rozdělení $N(0, 1)$



Logaritmicko-normální rozdělení

- X má *logaritmicko-normální rozdělení*, pokud $\log X \sim N(\mu, \sigma^2)$, pro nějaké parametry μ a σ .
- Log-normální rozdělení nabývá pouze kladných hodnot (jinak nelze logaritmovat).
- Nemá symetrickou hustotu, je sešikmené.

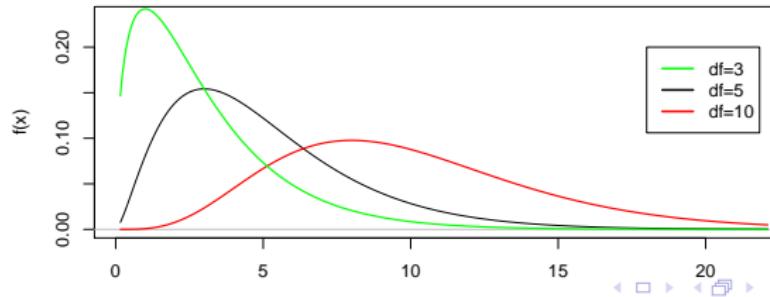
Obrázek: Hustota log-normálního rozdělení ($\log N(0,0.09)$)



χ^2 -rozdělení

- **χ^2 -rozdělení** (chí kvadrát) dostaneme jako rozdělení součtu druhých mocnin nezávislých náhodných veličin s normovaným normálním rozdělením.
- Pokud $X_i \sim N(0, 1)$, X_i nezávislé pro $i = 1, 2, \dots, k$, pak $X = X_1^2 + X_2^2 + \dots + X_k^2 \sim \chi_k^2$ (chí kvadrát s k stupni volnosti)
- χ^2 -rozdělení nabývá pouze kladných hodnot.

Obrázek: Hustoty rozdělení χ^2



Studentovo t -rozdělení

- Máme-li $X \sim N(0, 1)$ a $W \sim \chi^2_k$; X a W nezávislé, pak veličina T :

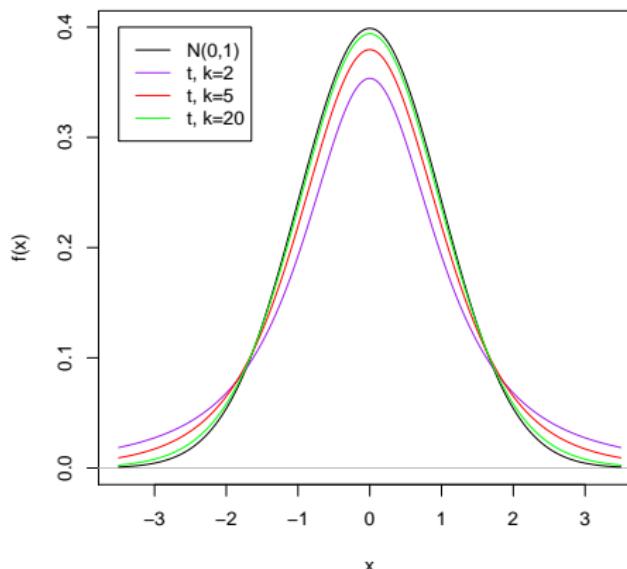
$$T = \frac{X}{\sqrt{W/k}} \sim t_k,$$

(T má *Studentovo t -rozdělení* s k stupni volnosti).

- Hustota t -rozdělení je symetrická kolem 0.
- Pro rostoucí počet stupňů volnosti se hustota blíží hustotě $N(0, 1)$.
- Používá se, pokud chceme normovat normálně rozdělenou veličinu směrodatnou odchylkou. Pokud směrodatnou odchylku neznáme, dělíme veličinu jejím odhadem a pak můžeme dostat veličinu s t -rozdělením.

Studentovo t -rozdělení

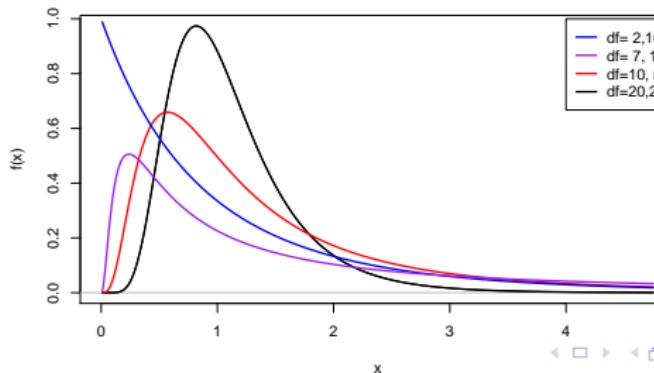
Obrázek: Srovnání hustot t -rozdělení a $N(0, 1)$



Fisherovo-Snedecorovo F -rozdělení

- **F -rozdělení** vznikne jako podíl dvou nezávislých náhodných veličin s χ^2 -rozdělením.
- $V \sim \chi^2_k$, $W \sim \chi^2_m$, V a W nezávislé, pak:
$$F = \frac{V/k}{W/m} \sim F_{k,m}$$
- Rozdělení se používá při porovnávání dvou odhadů rozptylů.

Obrázek: Hustoty F -rozdělení



Náhodný výběr

- Základní soubor - populace (všichni lidé na světě, všichni občané ČR, všichni diabetici, všechny myši světa apod.) O této populaci chceme něco zjistit, rozhodnout. Obvykle ji ale nemáme celou k dispozici.
- Máme pouze *náhodný výběr* - podsoubor vybraný z celé populace zcela náhodným postupem.
- Skutečné rozdělení veličiny, o kterou se zajímáme, neznáme.
- Pomocí náhodného výběru ale můžeme usoudit něco o rozdělení veličiny v celé populaci, odhadnout parametry rozdělení a rozhodnout hypotézy o celé populaci.
- To, co je v tomto postupu náhodné, je vybírání náhodného výběru.

Odhady parametrů pomocí náhodného výběru

- Výběr můžeme použít k odhadům parametrů rozdělení náhodné veličiny, ze které vybíráme, např. střední hodnoty nebo rozptylu.
- Pokud se střední hodnota odhadu rovná odhadovanému parametru, říkáme, že jde o odhad *nevychýlený*, nebo také *nestranný*.
- Nevychýlený odhad skutečně odhaduje hledaný parametr. Naopak, vychýlené odhady odhadují něco jiného, než hledaný parametr.
- Dobrý odhad by měl být nevychýlený a jeho rozptyl by se měl zmenšovat s rostoucím počtem jednotek ve výběru.

Náhodný výběr

Příklad (Výšky desetiletých dívek)

- Celá populace, o které chceme něco usoudit, jsou všechny desetileté dívky v ČR. Zajímá nás, jaké rozdělení mají jejich výšky, jakou mají střední hodnotu a jaký rozptyl.
- Desetiletých dívek v ČR bude kolem 50 000, tyto údaje k dispozici nemáme. Máme pouze výšky 100 dívek: X_1, X_2, \dots, X_{100} v náhodném výběru.
- Usoudíme, že výšky bývají v populaci normálně rozděleny a že by náš výběr mohl pocházet z rozdělení $N(\mu, \sigma^2)$, jehož parametry neznáme.
- Náhodný výběr použijeme k odhadům parametrů: pomocí $\bar{X} = 141.6$ odhadneme μ_X a pomocí $s_X^2 = 32.8$ odhadneme σ_X^2 .
- Budeme-li mít jiný výběr, dostaneme jiný průměr \bar{X} a jiný výběrový rozptyl s_X^2 . Naopak, skutečné, populační parametry μ_X a σ_X^2 náhodné nejsou, nezávisí na tom, jaký výběr jsme vybrali.

Vlastnosti náhodného výběru

- Výběr musí být *reprezentativní*, tj. reprezentovat celou populaci. Není rozumné vybírat pouze basketbalistky, chci-li usoudit něco o výškách všech dívek v ČR.
- Podobně není rozumné vybírat dívky z jediné školy, nebo pouze z Liberce. Výběr by měl odpovídat složení celé populace (bydliště, etnický původ apod.)
- Pokud výběr není reprezentativní, odhady jsou vychýlené.
- Důležitý je *rozsah výběru*, tj. počet jednotek ve výběru.
- Rozsah výběru ovlivní variabilitu odhadu. Máme-li malé výběry, odhady se mohou dost lišit.
- Rozsah také ovlivní naši schopnost rozhodovat o hypotézách. Nemusíme být schopni zamítнуть neplatnou hypotézu, protože v malém výběru je těžké rozhodnout, zda to, co pozorujeme není pouze náhoda.
- Malý rozsah výběru ale nezpůsobí vychýlenost odhadů.

Náhodný výběr

Příklad (Věk matek)

Máme velký soubor, 10 916 věků matek. Předpokládejme, že je to celá populace. Populační průměr je $\mu = 25.40$.

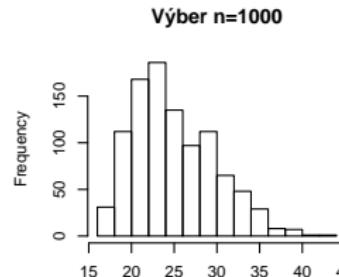
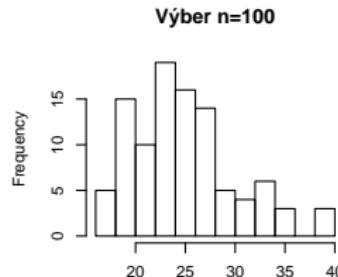
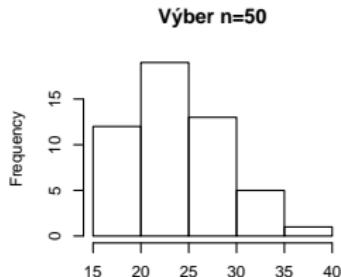
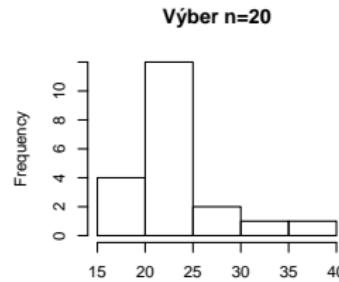
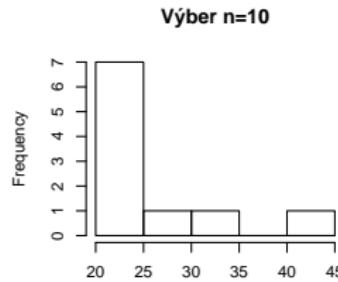
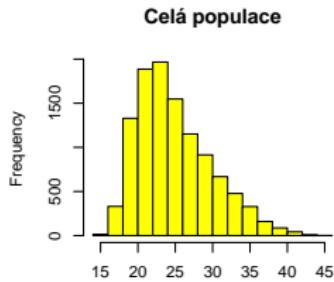
Vybírejme nyní náhodné výběry o různých rozsazích, vždy pětkrát. Tabulka obsahuje jejich průměry - odhady populačního průměru:

rozsah	průměry				
5	26.40	26.00	32.20	24.40	23.80
10	26.50	24.50	25.50	24.70	27.30
20	26.90	26.15	25.85	24.95	25.05
50	26.00	24.84	25.60	25.78	24.44
100	25.02	24.97	25.59	25.71	25.28
1000	25.41	25.40	25.42	25.38	25.47

Při malých výběrech se odhady μ dosti liší, čím větší je rozsah výběru, tím přesněji jsme schopni parametr odhadnout.

Náhodný výběr

Obrázek: Histogramy celé populace a výběrů



Průměr z náhodného výběru

- Nechť X_1, X_2, \dots, X_n je náhodný výběr, tj. veličiny X_1, X_2, \dots, X_n jsou stejně rozdělené a nezávislé. Předpokládejme, že mají střední hodnotu μ_X a rozptyl σ_X^2 .
- Spočítejme *střední hodnotu výběrového průměru*:
$$E\bar{X} = E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} E \sum_{i=1}^n X_i = \frac{1}{n} \cdot n \cdot \mu_X = \mu_X$$
(Použili jsme vlastnosti stř. hodnoty.)
- Střední hodnota průměru se rovná střední hodnotě rozdělení, ze kterého vybíráme výběr. \bar{X} je nestranným (nevychýleným) odhadem μ_X .

Průměr z náhodného výběru

- Spočítejme *rozptyl výběrového průměru \bar{X}* :

$$\text{var} \bar{X} = \text{var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \text{var} \sum_{i=1}^n X_i = \frac{1}{n^2} \cdot n \cdot \sigma_X^2 = \frac{\sigma_X^2}{n}$$

(Použili jsme vlastnosti rozptylu a nezávislost veličin X_1, X_2, \dots, X_n)

- *Směrodatná odchylka průměru* se rovná

$$\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}$$

- Směrodatná odchylka průměru se označuje také jako *střední chyba průměru*, (S.E.M. - standard error of mean).
- Obecně rozumíme *střední chybou odhadu* směrodatnou odchylku tohoto odhadu.

Průměr z náhodného výběru

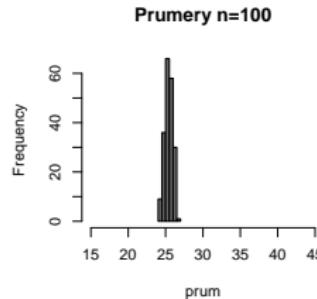
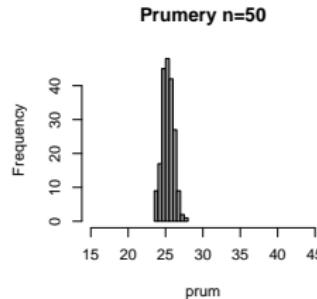
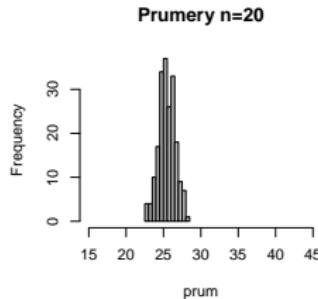
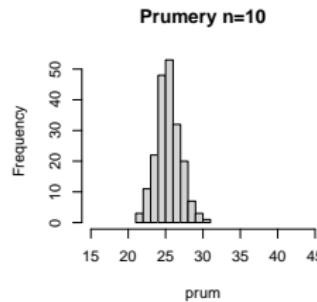
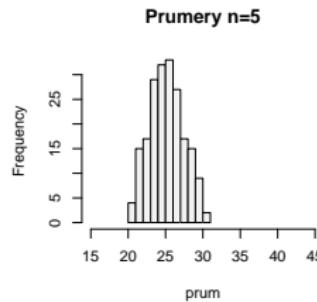
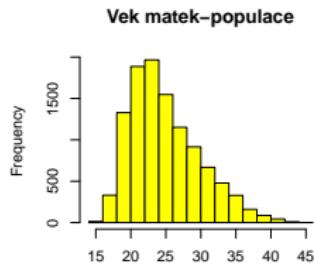
Příklad (Věk matek)

Znovu budeme generovat výběry o různých velikostech z populace matek. Tentokrát nás zajímá chování \bar{X} , takže pro každý rozsah vygenerujeme výběr 100krát a po každé spočítáme \bar{X} . Ze 100 hodnot \bar{X} pak spočítáme průměr a výběrovou směrodatnou odchylku. Populační stř. hodnota je $\mu = 25.40$, populační směrodatná odchylka je $\sigma_X = 4.94$. Výsledky jsou v tabulce:

<i>rozsah</i>	<i>prům \bar{X}</i>	$s_{\bar{X}}$	σ_X / \sqrt{n}
1	25.97	5.19	4.94
10	25.73	1.72	1.56
50	25.28	0.70	0.70
100	25.42	0.50	0.49
1000	25.40	0.16	0.16

Chování průměrů náhodných výběrů

Obrázek: Histogramy průměrů výběrů z populace matek (200 výb.)



Centrální limitní věta

- Nechť X_1, X_2, \dots, X_n je náhodný výběr z rozdělení se střední hodnotou μ_X a rozptylem σ_X^2 .
- Již víme, že výběrový průměr $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ má střední hodnotu μ_X a rozptyl $\frac{\sigma_X^2}{n}$.
- *Centrální limitní věta* (CLV) říká, že rozdělení znormovaného průměru

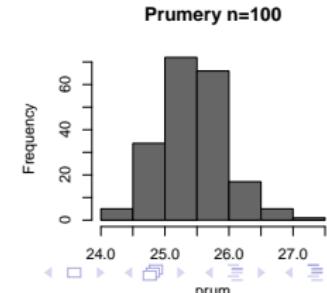
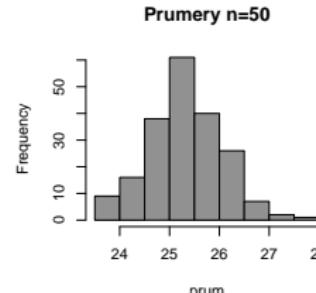
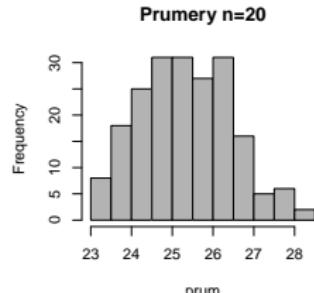
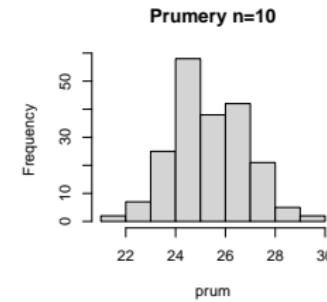
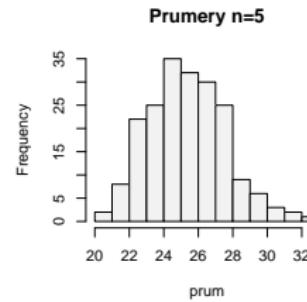
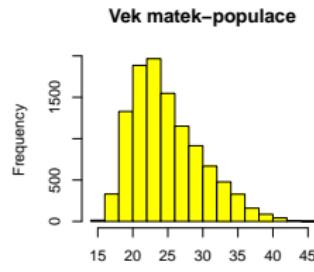
$$Z_n = \frac{\bar{X}_n - \mu_X}{\sigma} \cdot \sqrt{n}$$

se s rostoucím n blíží normovanému normálnímu $N(0, 1)$.

- Pro velká n tedy přibližně platí $\bar{X}_n \sim N(\mu_X, \sigma_X^2/n)$.
- Tento výsledek je velmi obecný, platí i pro diskrétní a zcela nenormální rozdělení.

Centrální limitní věta

Obrázek: Histogramy průměrů výběrů z populace matek (200 výb.) (různá měřítka)



Aproximace binomického rozdělení normálním

- Znalost CLV nám umožňuje approximovat binomické rozdělení pomocí normálního.
- Předpokládejme, že $Y \sim Bi(n, \pi)$ a n je dostatečně velké. Y je součtem n nezávislých alternativních veličin $Alt(\pi)$, takže $\frac{Y}{n}$ je průměrem n nezávislých veličin.
- Podle CLV platí pro dost vysoké n přibližně

$$\frac{Y}{n} \sim N\left(\pi, \frac{\pi(1-\pi)}{n}\right),$$

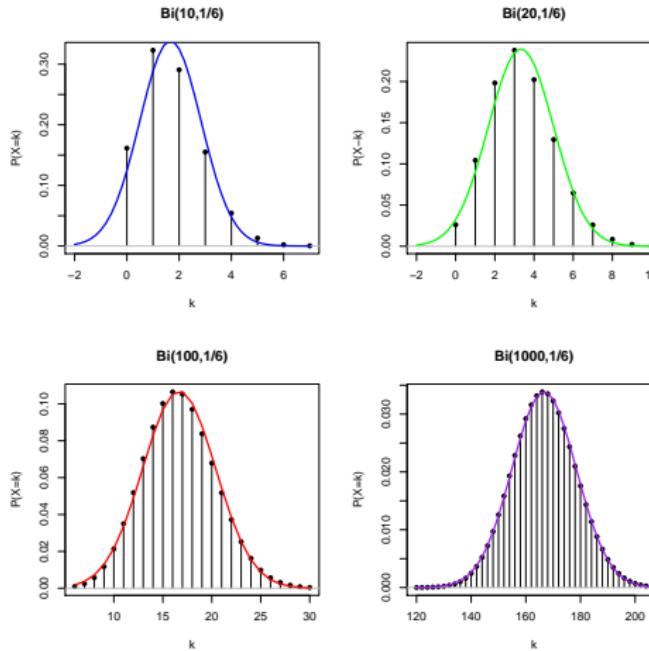
neboli

$$Y \sim N(n\pi, n\pi(1-\pi))$$

- Tato approximace se běžně používá, zjednodušuje se tím pracný přesný výpočet hodnot distribuční funkce a kvantilů binomického rozdělení při vyšších hodnotách n .

Aproximace binomického rozdělení normálním

Obrázek: Aproximace $\text{Bi}(n, 1/6)$ - pravd. šestky při hodu kostkou



Nestrannost výběrového rozptylu

- Upravme výběrový rozptyl S_n^2 následujícím způsobem:

$$\begin{aligned} S_n^2 &= \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = \\ &= \frac{1}{n-1} \sum_{i=1}^n (X_i - \mu_X + \mu_X - \bar{X})^2 = \\ &= \frac{1}{n-1} \sum_{i=1}^n (X_i - \mu_X)^2 - \frac{2}{n-1} \sum_{i=1}^n (X_i - \mu_X)(\bar{X} - \mu_X) + \\ &\quad + \frac{1}{n-1} \sum_{i=1}^n (\bar{X} - \mu_X)^2 = \\ &= \frac{1}{n-1} \sum_{i=1}^n (X_i - \mu_X)^2 - \frac{n}{n-1}(\bar{X} - \mu_X)^2 \end{aligned}$$

- Nyní použijeme vlastnosti střední hodnoty a spočítáme:

$$\begin{aligned} E S_n^2 &= \frac{1}{n-1} \sum_{i=1}^n E(X_i - \mu_X)^2 - \frac{n}{n-1} E(\bar{X} - \mu_X)^2 = \\ &= \frac{1}{n-1} n \cdot \sigma_X^2 - \frac{n}{n-1} \cdot \frac{\sigma_X^2}{n} = \frac{n \cdot \sigma_X^2}{n-1} - \frac{\sigma_X^2}{n-1} = \sigma_X^2 \end{aligned}$$

- Výběrový rozptyl je tedy nestranným odhadem populačního rozptylu (proto používáme při výpočtu výběrového rozptylu $n-1$).

Přesnost odhadu

- Mějme výběr X_1, X_2, \dots, X_n z normálního rozdělení $N(\mu, \sigma^2)$. Představme si, že σ^2 známe, ale μ neznáme a chceme jej odhadnout.
- Přirozeným odhadem μ bude \bar{X} . Již víme, že je to odhad nestranný, tj. $E\bar{X} = \mu$, takže skutečně odhaduje střední hodnotu. Nevíme však nic o přesnosti tohoto odhadu.
- Umíme si představit, že pokud je rozsah výběru $n = 5$, může být \bar{X} docela daleko od μ , zatímco pokud $n = 1000$, bude \bar{X} patrně velmi blízko μ .
- Jak blízko nebo jak daleko asi může být odhad od odhadovaného parametru nám pomohou zjistit intervaly spolehlivosti.

Int. spolehlivosti pro μ při známém σ

Pokud výběr X_1, X_2, \dots, X_n pochází z normálního rozdělení $N(\mu, \sigma^2)$, pak i průměr \bar{X} bude normální a bude platit $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$, tedy $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$.

Použijeme-li $z(\alpha)$ - kvantily $N(0, 1)$, vidíme, že

$$P\left(z\left(\frac{\alpha}{2}\right) < Z < z\left(1 - \frac{\alpha}{2}\right)\right) = 1 - \alpha$$

a ze symetrie $N(0, 1)$ kolem 0 dostaneme

$$\begin{aligned} 1 - \alpha &= P(|Z| < z\left(1 - \frac{\alpha}{2}\right)) = P\left(\frac{|\bar{X} - \mu|}{\sigma/\sqrt{n}} < z\left(1 - \frac{\alpha}{2}\right)\right) = \\ &= P\left(|\bar{X} - \mu| < \frac{\sigma}{\sqrt{n}}z\left(1 - \frac{\alpha}{2}\right)\right) \end{aligned}$$

To je ekvivalentní vztahu:

$$P\left(\bar{X} - \frac{\sigma}{\sqrt{n}}z\left(1 - \frac{\alpha}{2}\right) < \mu < \bar{X} + \frac{\sigma}{\sqrt{n}}z\left(1 - \frac{\alpha}{2}\right)\right) = 1 - \alpha$$

Int. spolehlivosti pro μ při známém σ

- Zjistili jsme, že interval

$$\left(\bar{X} - \frac{\sigma}{\sqrt{n}} z(1 - \frac{\alpha}{2}), \bar{X} + \frac{\sigma}{\sqrt{n}} z(1 - \frac{\alpha}{2}) \right)$$

pokrývá skutečnou hodnotu μ s pravděpodobností $1 - \alpha$.

- Tomuto intervalu se říká *($1 - \alpha$)100% interval spolehlivosti pro střední hodnotu μ .*
- Náhodnou složkou v těchto úvahách je výběr X_1, X_2, \dots, X_n , náhodný je tedy \bar{X} a následně hranice intervalu spolehlivosti. Interval spolehlivosti je tedy náhodný. Naopak, odhadovaná hodnota μ náhodná není.
- *($1 - \alpha$)100% interval spolehlivosti můžeme definovat jako náhodný interval, který pokrývá odhadovaný parametr s pravděpodobností $1 - \alpha$.*

Interval spolehlivosti

- Známe-li interval spolehlivosti, získáme informaci, jak daleko od \bar{X} nejvýše může (s rozumnou pravděpodobností) být μ .
- Běžně se užívají 95% intervaly spolehlivosti ($\alpha = 0.05$).
- Při odvození intervalu spolehlivosti jsme nepoužívali normalitu původních veličin X_1, X_2, \dots, X_n , ale pouze normalitu průměru \bar{X} . Podle CLV víme, že při dostatečném rozsahu výběru bude průměr normální i při nenormálně rozděleném výběru. Máme-li tedy dost velký výběr, uvedený interval spolehlivosti bude fungovat i pro nenormální rozdělení.
- Pro interval spolehlivosti se někdy používá zkratka CI (z angl. confidence interval).

Šířka intervalu spolehlivosti

- Šířka intervalu spolehlivosti závisí na směrodatné odchylce σ , na α (skrze kvantil $z(1 - \frac{\alpha}{2})$) a na počtu pozorování n .
- Větší rozptyl σ^2 rozdělení výběru bude znamenat širší interval spolehlivosti.
- Zvolíme-li menší α , dostaneme širší interval spolehlivosti, protože se snažíme najít interval, který bude pokrývat μ s vyšší pravděpodobností.
- Větší rozsah výběru n bude zužovat interval spolehlivosti.

Intervaly spolehlivosti

Příklad (Výšky desetiletých dívek)

Víme, že rozdělení výšek desetiletých dívek je $N(141, 36)$.

Představme si hypotetickou situaci, že neznáme střední hodnotu.

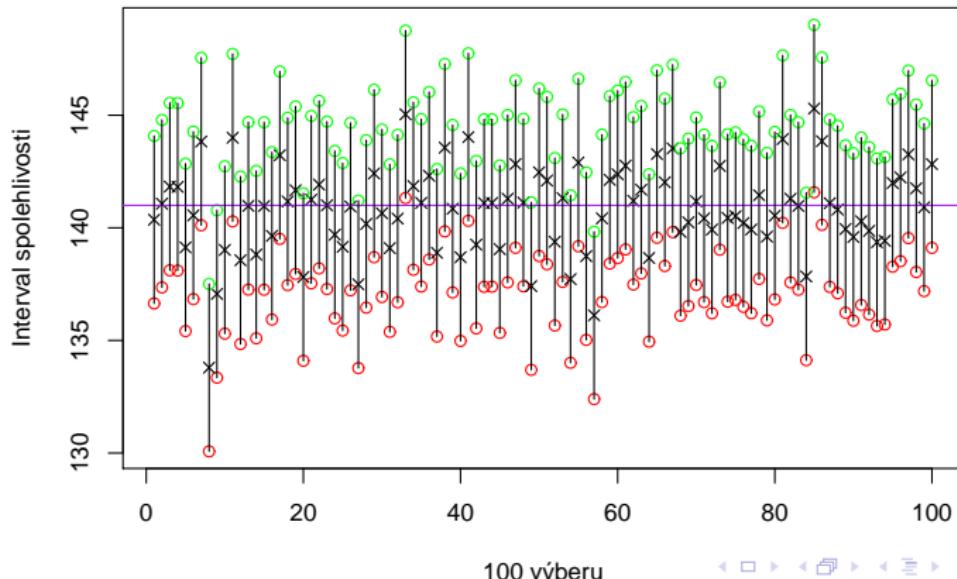
Abychom ji odhadli, vybereme náhodný výběr o rozsahu 10, spočítáme průměr a 95% interval spolehlivosti. Ten by měl pokrýt skutečnou hodnotu μ s pravděpodobností 0.95. Pokud tedy postup opakujeme 100krát, přibližně 95 intervalů pokryje μ .

Vzhledem k tomu, že ve skutečnosti μ známe, můžeme tento fakt zkontolovat.

V následujícím obrázku si všimněte, že všech 100 intervalů spolehlivosti má stejnou šířku, liší se pouze posunutím, díky různým hodnotám \bar{X} .

Intervaly spolehlivosti

Obrázek: Intervaly spolehlivosti, n=10



Šířka intervalu spolehlivosti

- Šířku intervalu spolehlivosti obvykle můžeme ovlivnit především rozsahem výběru.
- Šířka intervalu spolehlivosti

$$\delta = 2 \cdot \frac{\sigma}{\sqrt{n}} z\left(1 - \frac{\alpha}{2}\right)$$

- Pokud má mít interval spolehlivosti s požadovanou šířku δ , musíme zvolit

$$\sqrt{n} = \frac{2\sigma}{\delta} z\left(1 - \frac{\alpha}{2}\right),$$

$$n = \left(\frac{2\sigma}{\delta} z\left(1 - \frac{\alpha}{2}\right)\right)^2$$

Šířka intervalu spolehlivosti

Příklad (Výšky desetiletých dívek)

Pro $n = 10$ jsme dostali šířku 95% intervalu spolehlivosti:

$$\delta_{10} = \frac{2.6}{\sqrt{10}} \cdot 1.96 \doteq 7.44 \text{ (cm)}$$

To je velmi široký interval a rozhodně bychom chtěli střední hodnotu výšky odhadnout přesněji. Řekněme, že bychom chtěli, aby šířka intervalu spolehlivosti byla menší než 2 cm, tj., aby \bar{X} s 95% spolehlivostí padl do vzdálenosti menší než 1 cm od odhadované hodnoty μ . Potom:

$$n \geq \left(\frac{2\sigma}{\delta} z(1 - \frac{\alpha}{2}) \right)^2 = \left(\frac{2.6}{2} \cdot 1.96 \right)^2 \doteq 138.3.$$

Potřebovali bychom tedy výběr o rozsahu nejméně 139 dívek.

Int. spolehlivosti pro μ při neznámém σ

- Dosavadní úvahy byly poněkud nerealistické, protože se nedá předpokládat, že bychom znali rozptyl σ^2 zkoumané veličiny. Ve skutečnosti ho obvykle neznáme a musíme ho odhadnout pomocí výběrového rozptylu.
- Nahradíme-li ve znormovaném průměru

$$Z = \frac{\bar{X} - \mu}{\sigma} \sqrt{n}$$

směrodatnou odchylku σ jejím odhadem

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

dostáváme veličinu

$$T = \frac{\bar{X} - \mu}{S} \sqrt{n}.$$

- Veličina T nemá normální rozdělení, protože S je náhodná veličina, nikoliv konstanta. Dá se ale odvodit, že T má Studentovo t -rozdělení s $n - 1$ stupni volnosti.

Přednáška 6 (23.3.2021) - obsah

- Interval spolehlivosti pro μ při neznámém σ
- Interval spolehlivosti pro pravděpodobnost
- Testování hypotéz - hypotézy, chyby, hladina a síla testu
- Test o pravděpodobnosti v binomickém rozdělení (testová procedura, p -hodnota, síla, normální approximace)
- Test o střední hodnotě normálního rozdělení - jednovýběrový t -test (testová procedura, souvislost s CI, p -hodnota)

Int. spolehlivosti pro μ při neznámém σ

- Použijeme-li tedy místo kvantilů normálního rozdělení kvantily rozdělení t_{n-1} , stejným postupem dostaneme interval spolehlivosti.
- $(1 - \alpha)100\%$ *int. spolehlivosti pro μ při neznámém σ* je

$$\left(\bar{X} - \frac{S}{\sqrt{n}} t_{n-1}(1 - \frac{\alpha}{2}), \bar{X} + \frac{S}{\sqrt{n}} t_{n-1}(1 - \frac{\alpha}{2}) \right)$$
- Hustota t -rozdělení klesá pomaleji než hustota normálního rozdělení, takže $t_{n-1}(1 - \frac{\alpha}{2}) > z(1 - \frac{\alpha}{2})$, a interval spolehlivosti při neznámém σ je o něco širší, než by byl při známém rozptylu. Takto jsme penalizováni za to, že jsme neznámou hodnotu nahradili jejím odhadem.
- Naštěstí při vysokém počtu pozorování bude rozdíl mezi kvantilem t -rozdělení a normálního rozdělení velmi malý. t -rozdělení se k normálnímu blíží pro $n \rightarrow \infty$.
- Délku intervalu spolehlivosti určuje S/\sqrt{n} , což je odhad střední chyby průměru.

Int. spolehlivosti pro μ

Příklad (Glykémie)

Lékaře zajímá střední hodnota lačné glykémie u léčených pacientů s diabetem typu 2. U výběru 15 diabetických pacientů byla změřena glykémie nalačno (v mmol/l). Předpokládáme, že glykémie mají rozdělení $N(\mu, \sigma^2)$, ani jeden z parametrů neznáme, ale zajímá nás pouze μ . Toto jsou zjištěné hodnoty:

4.3	6.9	5.6	7.6	7.3	7.9	7.0	4.7	5.4
7.4	6.5	4.3	5.9	6.8	6.0			

$$n = 15, \bar{X} = 6.24, S = 1.18, t_{14}(0.975) = 2.14$$

95% interval spolehlivosti pro μ je

$$(6.24 - \frac{1.18}{\sqrt{15}} \cdot 2.14, 6.24 + \frac{1.18}{\sqrt{15}} \cdot 2.14) = (5.58, 6.90)$$

Význam intervalu spolehlivosti

- Všimněme si, že v příkladu pouze 5 z 15 zjištěných glykémií leží uvnitř intervalu spolehlivosti pro μ .
- S rostoucím n se interval spolehlivosti zužuje a většina pozorování bude ležet vně intervalu spolehlivosti pro stř. hodnotu.
- Interval spolehlivosti neříká nic o tom, kde leží pozorování!
- Interval spolehlivosti říká něco o poloze odhadovaného parametru.
- Je dobré se dívat na interval spolehlivosti jako na *intervalový odhad parametru*. Pokud chceme rozlišovat, můžeme odhad parametru jedním číslem (např. \bar{X}) nazývat *bodovým odhadem parametru*.

Interval spolehlivosti pro pravděpodobnost

Příklad (Diabetici)

Nevíme, jaký podíl diabetiků je ve sledované populaci. Vybereme náhodný výběr o rozsahu n . Počet diabetiků ve výběru označme Y .

Rozumný odhad pravděpodobnosti diabetu (π) ve sledované populaci bude relativní četnost $\frac{Y}{n} = \hat{\pi}$.

Podobně, jako když jsme odhadovali střední hodnotu normálního rozdělení, zatím nevíme nic o přesnosti tohoto odhadu.

Zkonstruujme tedy interval spolehlivosti pro π .

Centrální limitní věta pro relativní četnost

- Mějme náhodný výběr n subjektů, subjekty jsou nezávislé. Chceme odhadnout pravděpodobnost π nějakého znaku. Označme Y absolutní četnost znaku ve výběru.
- Pak $Y = X_1 + X_2 + \dots + X_n$, kde $X_i = 1$, pokud i -tý subjekt má zkoumaný znak, $X_i = 0$, pokud i -tý subjekt zkoumaný znak nemá.
- Potom $X_i \sim Alt(\pi)$ a $Y \sim Bi(n, \pi)$.
- Přirozeným odhadem π je relativní četnost $\hat{\pi} = \frac{Y}{n} = \frac{\sum_{i=1}^n X_i}{n}$, která je průměrem nezávislých, stejně rozdělených náhodných veličin X_i , $EX_i = \pi$, $varX_i = \pi(1 - \pi)$.
- Pak ovšem můžeme na $\hat{\pi}$ aplikovat centrální limitní větu a zjišťujeme, že pro velké n bude přibližně platit

$$\hat{\pi} \sim N\left(\pi, \frac{\pi(1-\pi)}{n}\right).$$

Interval spolehlivosti pro pravděpodobnost

- Zjistili jsme, že podle centrální limitní věty (CLV) pro velká n přibližně platí:

$$\hat{\pi} \sim N\left(\pi, \frac{\pi(1-\pi)}{n}\right).$$

- Pokud v rozptylu nahradíme neznámé π jeho odhadem $\hat{\pi}$, zjišťujeme, že přibližně platí:

$$P\left(\frac{|\hat{\pi}-\pi|}{\sqrt{\hat{\pi}(1-\hat{\pi})}}\sqrt{n} < z(1 - \frac{\alpha}{2})\right) = 1 - \alpha$$

- Známým postupem pak dostaneme *přibližný interval spolehlivosti pro π* :

$$\left(\hat{\pi} - \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}z(1 - \frac{\alpha}{2}), \hat{\pi} + \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}z(1 - \frac{\alpha}{2})\right).$$

- Skutečná pravděpodobnost π bude pokryta tímto intervalom s pravděpodobností $1 - \alpha$, pro n dost velké (použili jsme CLV).

Interval spolehlivosti pro pravděpodobnost

Příklad (Diabetici)

Vybrali jsme 50 lidí ze zkoumané populace, 9 z nich byli diabetici.

Bodový odhad pravděpodobnosti diabetu je:

$$\hat{\pi} = \frac{Y}{n} = \frac{9}{50} = 0.18$$

95% interval spolehlivosti je:

$$(0.18 - 1.96\sqrt{\frac{0.18 \cdot 0.82}{50}}, 0.18 + 1.96\sqrt{\frac{0.18 \cdot 0.82}{50}}) = (0.074, 0.286)$$

Testování hypotéz

Příklad (Kostky)

Chci zjistit, zda padá šestka na červené a modré kostce skutečně s pravděpodobností 1/6.

Hodím každou z nich 100krát, na červené kostce padla šestka 16krát, na modré pouze 9krát. Již umím spočítat 95% interval spolehlivosti pro pravděpodobnost:

*Červená kostka: $n = 100$, $Y_c = 16$, $\hat{\pi}_c = 0.16$, 95% interval spolehlivosti:
 $(0.16 - 1.96\sqrt{\frac{0.16 \cdot 0.84}{100}}, 0.16 + 1.96\sqrt{\frac{0.16 \cdot 0.84}{100}}) = (0.088, 0.232)$*

*Modrá kostka: $n = 100$, $Y_m = 9$, $\hat{\pi}_m = 0.09$, 95% interval spolehlivosti:
 $(0.09 - 1.96\sqrt{\frac{0.09 \cdot 0.91}{100}}, 0.1 + 1.96\sqrt{\frac{0.09 \cdot 0.91}{100}}) = (0.034, 0.146)$*

U modré kostky $\frac{1}{6}$ není v intervalu spolehlivosti, u červené ano. Co by to mohlo znamenat?

Testování

Příklad (Glykémie)

- Zkoumali jsme lačné glykémie u skupiny pacientů léčených pro diabetes typu 2 a zjistili jsme, že průměrná lačná glykémie ve výběru byla 6.24 mmol/l , 95% int. spolehlivosti (5.58,6.90).
- Chtěli bychom ji porovnat s lačnou glykémií pacientů, kteří kromě toho, že jsou léčeni na diabetes typu 2 mají také celiakii a dodržují bezlepkovou dietu.
- Vybrali jsme 18 takových pacientů a zjistili jsme, že průměrná lačná glykémie byla 5.27 mmol/l , 95% int. spolehlivosti (4.84,5.70).
- Zdá se, že pacienti s celiakií měli v průměru nižší lačnou glykémii. Můžeme rozhodnout, že rozdělení glykémie v těchto skupinách se opravdu liší?
- Intervaly spolehlivosti pro stř. hodnoty ve dvou skupinách se trochu překrývají. Znamená to, že by střední hodnoty mohly být stejné?

Rozhodování o populaci

- Chěli bychom rozhodnout něco o celé populaci, k dispozici máme ale pouze náhodný výběr.
- Nemůžeme na základě výběru něco s jistotou tvrdit o celé populaci, ale můžeme použít rozhodovací proceduru, ve které budeme kontrolovat pravděpodobnost, že uděláme chybu.
- Taková procedura nám umožní rozhodnout, když připustíme, že se můžeme s nějakou známou, malou pravděpodobností mylit.
- Standardizovanou rozhodovací proceduru nazveme *statistickým testem*.

Volba hypotéz

- Budeme rozhodovat o tom, zda platí tvrzení, které nazveme *nulovou hypotézou H_0* , nebo jiné tvrzení, které nazveme *alternativní hypotézou H_1* .
- Nulová hypotéza: neděje se nic zajímavého (kostka je správná ($\pi = \frac{1}{6}$), obě skupiny mají stejnou stř. hodnotu lačné glykémie, dvě veličiny jsou nezávislé apod.)
- Obvykle se snažíme nulovou hypotézu vyvrátit.
- Alternativní hypotéza je obvykle to, co chceme prokázat, opak nulové hypotézy (dvě populace se liší, veličiny jsou závislé, na kostce padá šestka méně často, než by měla).

Možná rozhodnutí

- Rozhodovací procedura skončí buď tím, že *zamítneme H_0* , nebo tím, že *nezamítneme H_0* .
- Zamítneme-li H_0 , znamená to, že naše data svědčí proti H_0 . Shromáždili jsme dostatek důkazů proti H_0 (u soudu: obviněný je odsouzen).
- Nezamítneme-li H_0 , znamená to, že nemáme dost důkazů proti H_0 (u soudu: obviněný není odsouzen).
- Nezamítneme-li H_0 , neznamená to ještě, že H_0 platí. Je možné, že jen nemáme důkazy (málo dat apod.)
- Zamítneme-li H_0 , je situace jasnější, znamená to, že jsme shromáždili dost důkazů proti H_0 , H_0 tedy nejspíš neplatí.
- Rozhodovací procedura se nechová k hypotézám symetricky!
- Pokud zamítneme H_0 , mluvíme o *statisticky významném* nebo také *signifikantním* výsledku.

Chyby v rozhodovací proceduře

- Můžeme se dopustit *chyby 1. druhu*, pokud zamítneme H_0 , přestože H_0 platí,
- nebo *chyby 2. druhu*, pokud nezamítneme H_0 , přestože platí H_1 .
- Obvykle chceme držet nízkou hlavně pravděpodobnost chyby 1. druhu. Nechceme mylně prokazovat, že H_0 neplatí.
- **Př.- Klinické zkoušky:** Porovnáváme novou léčbu s používanou standardní. H_0 : Nová léčba není lepší. H_1 : Nová léčba je lepší. Nechceme se dopuštět chyby 1. druhu často, jinak bude zbytečně měněna standardní léčba, bez zlepšení výsledků.

Hladina a síla testu

- Abychom kontrolovali pravděpodobnost chyby 1. druhu, stanovíme *hladinu testu α* (obvykle se volí $\alpha = 0.05$), což je maximální přípustná pravděpodobnost chyby 1. druhu.
- Hladina testu zpravidla definuje testovací proceduru a chybu 2. druhu již kontrolovat nemůžeme. Může se stát, že pravděpodobnost chyby 2. druhu je vysoká. Při plánování experimentu/šetření je dobré o chybách 2. druhu přemýšlet.
- *Síla testu $1 - \beta$* je pravděpodobnost, že zamítneme H_0 , pokud platí H_1 . Bude záviset na konkrétní alternativě. Zkoumání síly testu pro různé alternativy nám může dát představu, jak často se budeme dopouštět chyby 2. druhu.

Schéma testu

	Skutečnost	
Rozhodnutí	H_0 platí	H_0 neplatí
zam. H_0	dopouštím se chyby 1. druhu $P \leq \alpha$ (hladina)	správné rozhodnutí $P = 1 - \beta$ (síla)
nezam. H_0	správné rozhodnutí $P \geq 1 - \alpha$	dopouštím se chyby 2. druhu $P = \beta$

Síla testu

Příklad (Kostky)

Chceme testovat, zda pravděpodobnost šestky je $\frac{1}{6}$. Zvolíme nulovou hypotézu $H_0 : \pi = \frac{1}{6}$ a alternativní $H_1 : \pi \neq \frac{1}{6}$. Provedeme 100 hodů a chceme rozhodnout.

Řekněme, že je skutečná pravděpodobnost šestky $\pi = \frac{1}{2}$. Pak si umíme představit, že šestka bude padat výrazně víckrát než by měla a celkem snadno zamítneme H_0 . Budeme mít velkou sílu.

Řekněme, že skutečná pravděpodobnost šestky je $\pi = \frac{1}{5}$. Šestka bude padat trochu více, než by měla, ale nebude tak snadné rozhodnout, zda to není jen náhoda. Možná výsledek nebude přesvědčivý a H_0 nezamítneme. V tomto případě máme sílu nižší. (Jak bychom si mohli pomoci, máme-li podezření na nízkou sílu testu?)

Síla tedy závisí na skutečné hodnotě testovaného parametru.

Výsledek testu

- Je třeba mít neustále na zřeteli, že z náhodného výběru nemůžeme rozhodnout hypotézu o celé populaci s absolutní jistotou.
- Za platnosti hypotézy H_0 testová procedura kontroluje pravděpodobnost, že uděláme chybu (pravd. je menší než α).
- Za platnosti H_1 pravděpodobnost, že uděláme chybu (β), pod kontrolou nemáme.
- Sílu testu $(1 - \beta)$ lze ovlivnit rozsahem výběru, je třeba provést úvahy o síle testu již při plánování experimentu/šetření.
- Pokud testujeme hodnotu nějakého parametru (jako π u kostek), je dobré po provedení testu spočítat interval spolehlivosti pro parametr. Tím získáme určitou představu o tom, jak přesně odhadujeme a testujeme.

Testování hypotéz

Testování hypotéz - postup při rozhodování:

- ① Zvolit nulovou hypotézu H_0 a alternativní H_1 .
- ② Zvolit hladinu testu α .
- ③ Zvolit test, tj. pravidlo, podle kterého rozhodneme (obvykle testovou statistiku a kritický obor).
- ④ Spočítat testovou statistiku a rozhodnout.
- ⑤ Interpretovat výsledek (co to znamená, že jsme zamítli/nezamítli H_0). Velmi důležité, bez tohoto bodu je naše námaha k ničemu.

Testová statistika a kritický obor

- Test je obvykle založen na nějaké *testové statistice (T)*, což je náhodná veličina spočítaná z náhodného výběru. Její rozdělení závisí na tom, zda platí, či neplatí H_0 .
- *Kritický obor (K)* je množina, do které když padne T , tak zamítneme H_0 .
- Statistiku T volíme tak, aby její rozdělení za nulové hypotézy bylo známé.
- Potom je K obvykle určen požadavkem na hladinu testu a lze ho najít pomocí známého rozdělení T za H_0 .

Test o pravděpodobnosti v binomickém rozdělení

- Chceme testovat nulovou hypotézu, že pravděpodobnost π nějakého jevu je rovna nějakému konkrétnímu číslu π_0 . Volíme tedy $H_0: \pi = \pi_0$.
- Máme podezření, že π je menší, testujeme tedy proti *jednostranné alternativě* $H_1: \pi < \pi_0$.
- Budeme rozhodovat na základě pokusu nebo šetření, které jsme provedli. Řekněme, že pozorujeme n situací, ve kterých sledovaný jev nastane nebo nenastane a Y bude náhodná veličina označující počet situací, kdy sledovaný jev nastal.
- Proti H_0 ve prospěch H_1 bude svědčit nízká hodnota Y .
- Založíme tedy testovou statistiku na Y a budeme zamítat H_0 pokud Y bude nízké, kritický obor bude nějaký interval $K = \langle 0, k \rangle$ a budeme zamítat H_0 , pokud $Y \in K$.

Test o pravděpodobnosti v binomickém rozdělení

- Má-li test dodržet hladinu α , musí platit, že pravděpodobnost chyby 1. druhu nepřesáhne α . Musí tedy platit:
$$P(\text{zam. } H_0 | H_0) = P(Y \epsilon K | H_0) = P(Y \leq k | H_0) \leq \alpha$$
- Pokud platí H_0 , pak $Y \sim Bi(n, \pi_0)$ a $P(Y \leq k | H_0) = F_0(k)$, kde F_0 je distribuční funkce rozdělení $Bi(n, \pi_0)$.
- Největší takové celé k , aby $F_0(k) \leq \alpha$ tedy získáme ze znalosti binomického rozdělení (výpočtem nebo z tabulek) a takto určíme kritický obor.

P-hodnota

- Povšimněme si, že nemusíme nutně kritický obor znát, přesto dokážeme učinit rozhodnutí.
- Řekněme, že jsme napozorovali $Y = y$. Pokud $F_0(y) \leq \alpha$, musí nutně platit $y \leq k$ (k bylo největší celé číslo, pro které nerovnost platí), takže nastala situace, kdy $Y \in K$ a H_0 zamítneme.
- Pokud naopak $F_0(y) > \alpha$, pak musí platit $y > k$ (distribuční funkce je neklesající), tj. H_0 nezamítneme.
- V tomto případě nám tedy stačí spočítat $F_0(y)$ a víme, zda zamítnout H_0 nebo ne.
- Tomuto postupu se říká zamítání pomocí p-hodnoty. Hodnota distribuční funkce $F_0(y)$ je v tomto případě p-hodnotou.

P-hodnota

- **P-hodnota** může pro různé testy vypadat různě. Její význam lze formulovat třeba takto: P-hodnota je pravděpodobnost, že za H_0 bude mít testová statistika takovou hodnotu, jakou jsme dostali, nebo hodnotu ještě více svědčící proti nulové hypotéze ve prospěch alternativy.
- P-hodnota je nejmenší hladina α , na které bychom na základě napozorovaných dat zamítli H_0 .
- Pokud je p-hodnota menší nebo rovna hladině testu α , zamítáme H_0 . Znamená to, že za H_0 je hodnota testové statistiky, kterou jsme dostali, nepravděpodobná.
- Pokud je p-hodnota větší než α , nezamítáme H_0 , znamená to, že se za H_0 takováto hodnota testové statistiky může vyskytnout s pravděpodobností vyšší než je hladina.
- P-hodnota bude záležet na testové statistice, ale také na kritickém oboru. Může se lišit, pokud změníme alternativu, například z jednostranného na oboustranný test.

Test o pravděpodobnosti v binomickém rozdělení

Příklad (Kostky)

Máme podezření, že šestka na modré kostce padá méně často, než by měla. Nechť π je pravděpodobnost, že padne šestka. Zkusíme testovat $H_0: \pi = \frac{1}{6}$ proti alternativě $H_1: \pi < \frac{1}{6}$.

S kostkou jsme provedli náhodný pokus, hodili jsme 100krát. Y je počet šestek ze 100 pokusů (absolutní četnost). Kritický obor bude interval $K = \langle 0, k \rangle$. Pokud $Y \in K$ ($Y \leq k$), zamítneme H_0 .

Nyní najdeme k tak, aby pravděpodobnost chyby 1. druhu nepřekročila hladinu $\alpha = 0.05$.

$$P(\text{ch.1.druhu}) = P(\text{zam.} H_0 | H_0) = P(Y \leq k | H_0) = F_0(k) \leq 0.05$$

F_0 je distribuční funkce Y za nulové hypotézy, je to tedy distribuční funkce $Bi(100, \frac{1}{6})$. Použijeme tedy největší číslo k takové, že $F_0(k) \leq 0.05$.

Test o pravděpodobnosti v binomickém rozdělení

Příklad (Kostky - pokračování)

Binomické rozdělení je diskrétní, možná nenajdeme k tak, aby $F_0(k) = 0.05$ přesně. Z tabulek zjistíme, že $P(Y \leq 10|H_0) = 0.043$ a $P(Y \leq 11|H_0) = 0.078$. Zvolíme tedy $k = 10$, kritický obor bude $K = \langle 0, 10 \rangle$.

Pro modrou kostku $Y = 9$, tedy $Y \in K$. To znamená, že zamítáme H_0 ve prospěch alternativy H_1 . Modrá kostka tedy patrně není správná. Na hladině $\alpha = 0.05$ jsme prokázali, že pravděpodobnost, se kterou padá šestka, je nižší než $\frac{1}{6}$.

Mohli jsme rozhodnout pomocí p-hodnoty. P-hodnota v tomto případě je $F_0(9) = 0.021 \leq 0.05$, zamítáme H_0 .

U červené kostky je $Y = 16$, nulovou hypotézu bychom nezamítli (p-hodnota 0.494). U červené kostky tedy náš experiment nikterak nesvědčí proti H_0 .

Síla testu

Příklad (Kostky)

Počítejme sílu testu s kostkami:

Síla testu je $P(\text{zam. } H_0 | H_1) = P(Y \leq 10 | H_1)$. Abychom mohli sílu spočítat, musíme znát rozdělení Y za H_1 . Za H_1 bude $Y \sim Bi(100, \pi)$, a bude záležet na konkrétní hodnotě π .

Zvolme $\pi = 0.05$, to by znamenalo, že šestka padá přibližně v jednom z 20 případů. $P(Y \leq 10 | \pi = 0.05) = 0.99$. V tomto případě je tedy síla vysoká, pravděpodobnost chyby 2. druhu pouze 0.01.

Pro $\pi = 0.1$ dostaneme $P(Y \leq 10 | \pi = 0.1) = 0.58$, takže pokud šestka padá přibližně v jednom z deseti případů, často její chybnost neodhalíme, ve 42 % experimentů bychom nic neprokázali.

Pro $\pi = 0.15$ bude síla $P(Y \leq 10 | \pi = 0.15) = 0.10$, v 90 % případů náš experiment chybnost kostky neodhalí.

Síla testu

Příklad (Kostky - pokračování)

Zvětšení rozsahu výběru vede ke zvětšení síly. Zkusme zvýšit počet hodů kostkou na 200. Pak za nulové hypotézy bude $Y \sim Bi(200, \frac{1}{6})$.

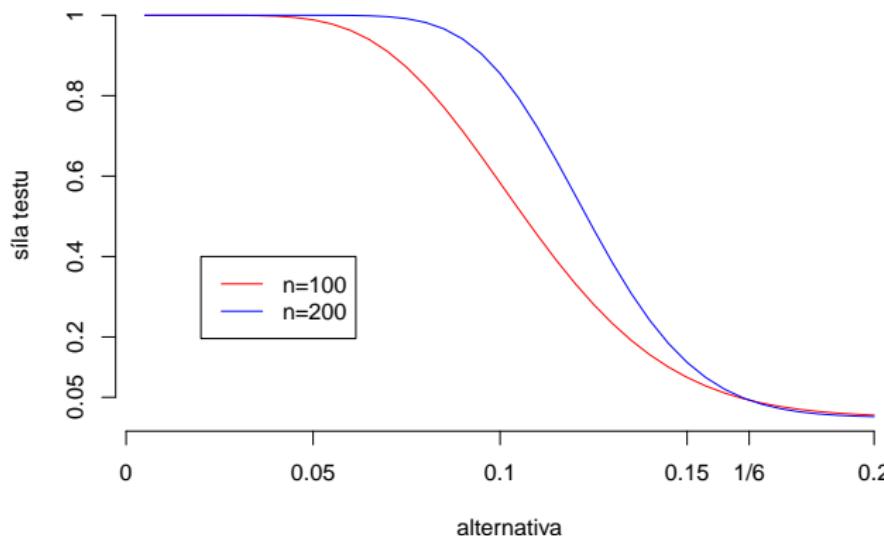
Kritický obor testu se změní, abychom dodrželi hladinu $\alpha = 0.05$, bude $k = 24$, $K = \langle 0, 24 \rangle$.

Pro $\pi = 0.1$ dostaneme $P(Y \leq 24 | \pi = 0.1) = 0.86$, takže v tomto případě se naše schopnost detegovat chybnou kostku výrazně zvýšila.

Pro $\pi = 0.15$ dostaneme $P(Y \leq 24 | \pi = 0.15) = 0.14$, zde je síla pořád malá. Abychom pro tuto alternativu dosáhli síly aspoň 0.5, potřebovali bychom více než 1300 hodů kostkou.

Síla testu

Obrázek: Síla testu jako funkce π - příklad kostky



Oboustranná alternativa

- Zatím jsme testovali nulovou hypotézu typu $H_0: \pi = \pi_0$, proti **jednostranné alternativě** $H_1: \pi < \pi_0$.
- Jednostrannou alternativu volíme, pokud nás opačná situace nezajímá nebo nemůže nastat (u kostek nás zajímalo pouze, není-li pravděpodobnost šestky nižší).
- Pokud nás zajímá zcela obecně, zda $\pi = \pi_0$ nebo ne, budeme testovat proti **oboustranné alternativě** $H_1: \pi \neq \pi_0$.
- V takovém případě se změní kritický obor, protože budeme zamítat jak pro příliš malé četnosti, tak pro příliš velké. Kritický obor bude $\langle 0, k_1 \rangle \cup \langle k_2, n \rangle$.
- k_1 a k_2 získáme z požadavku na pravděpodobnost chyby 1. druhu: $P(\text{zam.} H_0 | H_0) = P(Y \leq k_1 | H_0) + P(Y \geq k_2 | H_0) \leq \alpha$
- Obvykle se rozdělí α na poloviny a kritický obor se najde tak, aby k_1 bylo největší celé splňující $P(Y \leq k_1 | H_0) \leq \frac{\alpha}{2}$ a k_2 nejmenší celé takové, že $P(Y \geq k_2 | H_0) \leq \frac{\alpha}{2}$.

Oboustranná alternativa

Příklad (Kostky)

Testujme tedy oboustranně, $H_0: \pi = \frac{1}{6}$ proti $H_1: \pi \neq \frac{1}{6}$.

Zamítneme H_0 pro Y bud' příliš velké, nebo příliš malé. Kritický obor bude $K = \langle 0, k_1 \rangle \cup \langle k_2, 100 \rangle$.

Najdeme tedy největší k_1 takové, aby

$P(Y \leq k_1 | H_0) \leq \alpha/2 = 0.025$ a nejmenší k_2 takové, aby

$P(Y \geq k_2 | H_0) \leq 0.025$. Zjistíme, že $k_1 = 9$ a $k_2 = 25$. Kritický obor je tedy $K = \langle 0, 9 \rangle \cup \langle 25, 100 \rangle$.

Kdybychom testovali proti oboustranné alternativě, stále bychom pro modrou kostku H_0 zamítali (šestka padla 9krát).

Normální approximace

- Uvažujeme testování pravděpodobnosti π nějakého jevu.
 $H_0: \pi = \pi_0$ proti oboustranné alternativě $H_1: \pi \neq \pi_0$.
- Testová statistika Y (četnost jevu) má binomické rozdělení, za H_0 platí $Y \sim Bi(n, \pi_0)$. Toto rozdělení můžeme podle CLV approximovat normálním, protože $\frac{Y}{n}$ je průměrem nezávislých náhodných veličin s rozdělením $Alt(\pi_0)$.
- Pro velká n bude tedy za H_0 přibližně platit

$$Y \sim N(n\pi_0, n\pi_0(1 - \pi_0)),$$

a pro znormovanou četnost

$$Z = \frac{Y - n\pi_0}{\sqrt{n\pi_0(1 - \pi_0)}} \sim N(0, 1)$$

- Můžeme použít Z jako testovou statistiku a jako hranice kritického oboru kvantily normálního rozdělení.
- Budeme tedy zamítat pro $|Z| \geq z(1 - \frac{\alpha}{2})$. Pro $\alpha = 0.05$ to znamená zamítnout, pokud $|Z| \geq 1.96$.

Normální approximace

Příklad (Kostky)

Pro modrou kostku:

$$Z = \frac{9 - 100 \cdot \frac{1}{6}}{\sqrt{100 \cdot \frac{1}{6} \left(1 - \frac{1}{6}\right)}} = -2.06$$

Z je v absolutní hodnotě vyšší, než 1.96, takže bychom zamítli H_0 stejně jako při použití přesného testu.

Pro červenou kostku:

$$Z = \frac{16 - 100 \cdot \frac{1}{6}}{\sqrt{100 \cdot \frac{1}{6} \left(1 - \frac{1}{6}\right)}} = -0.18$$

V tomto případě bychom, samozřejmě H_0 nezamítli.

Test o střední hodnotě normálního rozdělení

- Nechť X_1, X_2, \dots, X_n jsou nezávislé náhodné veličiny s rozdělením $N(\mu, \sigma^2)$ (náhodný výběr).
- Neznáme ani μ ani σ^2 , ale σ^2 nás nezajímá.
- Chceme testovat, zda je střední hodnota μ rovna nějaké konkrétní hodnotě μ_0 .
- Nulová hypotéza: $H_0: \mu = \mu_0$.
Alternativní hypotéza: $H_1: \mu \neq \mu_0$ (oboustranná alternativa).
- Pokud máme důvod, můžeme testovat proti některé z jednostranných alternativ $H_1: \mu < \mu_0$ nebo $H_1: \mu > \mu_0$.
- Důvod k volbě jednostranné alternativy musí vycházet z logiky problému, který studujeme (opačná alternativa nemůže nastat, nebo nás nezajímá), nikoliv z pohledu na data!

Test o střední hodnotě normálního rozdělení

Příklad (Glykémie)

Lékaře zajímá, zda jeho současní pacienti mají stejné rozdělení lačné glykémie jaké se uvádí v literatuře u léčených pacientů s diabetem typu 2. Zjistil, že se rozdělení obvykle považuje za normální se střední hodnotou 6.0.

Chceme tedy testovat, zda jeho pacienti mají střední hodnotu glykémie 6.0. Zvolíme $H_0: \mu = 6.0$ proti $H_1: \mu \neq 6.0$.

Jednovýběrový t-test

- Základem testové statistiky bude $\bar{X} - \mu_0$. Pokud bude $|\bar{X} - \mu_0|$ příliš vysoké, bude to svědčit proti H_0 .
- Za H_0 bude mít $\bar{X} - \mu_0$ normální rozdělení s nulovou stř. hodnotou a neznámým rozptylem.
- Ke znormování veličiny použijeme $\frac{s_x^2}{n}$, což je odhad rozptylu \bar{X} a dostaváme testovou statistiku

$$T = \frac{\bar{X} - \mu_0}{s_x} \sqrt{n}$$

- Lze ukázat, že za H_0 platí $T \sim t_{n-1}$, testová statistika má Studentovo t-rozdělení o $n - 1$ stupních volnosti. Můžeme tedy použít kvantily t-rozdělení jako hranice kritického oboru.
- Popsaný test nazveme *jednovýběrovým t-testem*.

Jednovýběrový t-test

- Jednovýběrový t -test tedy používá testovou statistiku

$$T = \frac{\bar{X} - \mu_0}{S_X} \sqrt{n}$$

- Při oboustranné alternativě bude zamítat pro $|T| \geq k$, kde k získáme z požadavku na dodržení stanovené hladiny α :

$$\begin{aligned} P(\text{chyby 1.druhu}) &= P(|T| \geq k | H_0) = \\ &= P(T \leq -k | H_0) + P(T \geq k | H_0) = \alpha \end{aligned}$$

- Vzhledem k rozdělení statistiky T za H_0 bude tato rovnost splněna, zvolíme-li $k = t_{n-1}(1 - \frac{\alpha}{2})$. Podobně postupujeme v jednostranných případech a dostáváme kritické obory:

alternativa	test zamítne, pokud
$\mu \neq \mu_0$	$ T \geq t_{n-1}(1 - \frac{\alpha}{2})$
$\mu > \mu_0$	$T \geq t_{n-1}(1 - \alpha)$
$\mu < \mu_0$	$T \leq -t_{n-1}(1 - \alpha)$

Jednovýběrový t-test a interval spolehlivosti

- V situaci X_1, X_2, \dots, X_n , kde μ i σ^2 jsou neznámé jsme odvodili $(1 - \alpha)100\%$ int. spolehlivosti pro μ :

$$\left(\bar{X} - \frac{s_x}{\sqrt{n}} \cdot t_{n-1}(1 - \frac{\alpha}{2}), \bar{X} + \frac{s_x}{\sqrt{n}} \cdot t_{n-1}(1 - \frac{\alpha}{2}) \right).$$

- Pro všechny body m z int. spolehlivosti platí:

$$\bar{X} - \frac{s_x}{\sqrt{n}} \cdot t_{n-1}(1 - \frac{\alpha}{2}) < m < \bar{X} + \frac{s_x}{\sqrt{n}} \cdot t_{n-1}(1 - \frac{\alpha}{2})$$

- Úpravou této nerovnosti dostáváme, že pro všechny body m z int. spolehlivosti platí:

$$|T| = \frac{|\bar{X} - m|}{s_x} \sqrt{n} < t_{n-1}(1 - \frac{\alpha}{2})$$

- Pro všechny body m z int. spol. tedy platí, že pokud bychom testovali $H_0: \mu = m$ proti oboustranné alternativě na hladině α , jednovýběrový t -test by nezamítl. Obráceně samozřejmě platí, že pro všechny body mimo int. spol. by test zamítl.

Jednovýběrový t-test a interval spolehlivosti

- Mezi intervalem spolehlivosti pro μ a jednovýběrovým t -testem tedy existuje jednoznačný vztah, jsou založeny na stejné statistice T .
- Výsledek jednovýběrového t -testu je vidět na odpovídajícím int. spolehlivosti, stačí se podívat, zda testovaná hodnota μ_0 je, či není pokryta int. spolehlivosti.
- Na $(1 - \alpha)100\%$ int. spolehlivosti pro μ se můžeme dívat jako na množinu všech takových bodů m , pro které by jednovýběrový t -test na hladině α nezamítl $H_0: \mu = m$.

Jednostranné testy a jednostranné intervaly spolehlivosti

- Při použití jednostranného t-testu je přirozeně množina všech bodů m , pro které by test neměl zamítat $H_0: \mu = m$ z jedné strany neomezená.
- Pokud testujeme proti alternativě $H_1: \mu > m$, bude test zamítat, pokud $T \geq t_{n-1}(1 - \alpha)$, tj. pokud $\bar{X} - \frac{S_x}{\sqrt{n}} t_{n-1}(1 - \alpha) \geq m$.

Test tedy nezamítne H_0 právě tehdy, když

$$m \in (\bar{X} - \frac{S_x}{\sqrt{n}} \cdot t_{n-1}(1 - \alpha), \infty).$$

- Tento jednostranný interval spolehlivosti pokrývá skutečnou střední hodnotu μ s pravděpodobností $1 - \alpha$. Na rozdíl od oboustranného CI si “vybírá” možnost nepokrytí μ pouze na jedné straně.
- Podobně, testujeme-li proti alternativě $H_1: \mu < m$, odpovídá testu obrácený jednostranný interval spolehlivosti:

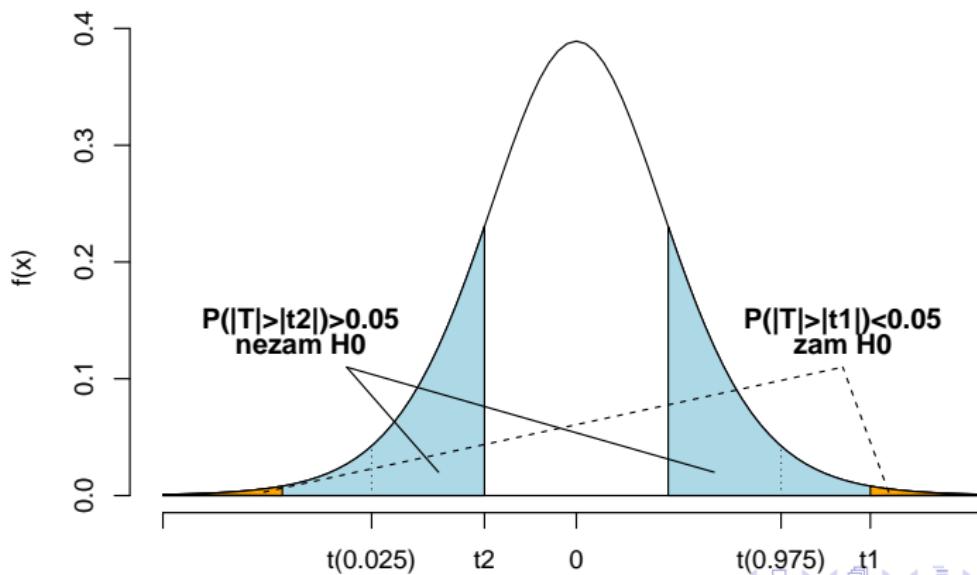
$$(-\infty, \bar{X} - \frac{S_x}{\sqrt{n}} \cdot t_{n-1}(1 - \alpha)).$$

Rozhodování pomocí p -hodnoty

- Obvykle rozhodujeme o výsledku testu pomocí p -hodnoty, což je pravděpodobnost, že za nulové hypotézy bude hodnota testové statistiky právě taková, jakou jsme dostali, nebo ještě více svědčící proti nulové hypotéze ve prospěch alternativní.
- Řekněme, že jsme v testu dostali hodnotu testové statistiky $T = t$. Potom hodnoty svědčící stejně nebo ještě více proti H_0 (při oboustranné alternativě) jsou hodnoty v absolutní hodnotě větší nebo rovné absolutní hodnotě t .
- P -hodnota je tedy $P(|T| \geq |t|)$ spočítaná za platnosti H_0 .
- Pokud je p -hodnota menší nebo rovna α , pak zamítáme H_0 , znamená to, že $|t|$ musí být větší než $t_{n-1}(1 - \frac{\alpha}{2})$. T tedy padla do kritického oboru.
- Pokud je p -hodnota větší než α , pak nezamítáme H_0 , protože T není v kritickém oboru.

P-hodnota

Obrázek: Rozhodování pomocí p-hodnoty



P-hodnota

- Velmi nízká p -hodnota nemusí znamenat, že je μ velmi daleko od μ_0 .
- P-hodnota bude velmi nízká, pokud máme přesvědčivé důkazy, že H_0 neplatí. To může být proto, že μ je daleko od μ_0 , ale také proto, že máme velký výběr a pak je třeba i malý rozdíl průkazný.
- Představu o tom, jak je skutečná μ daleko od μ_0 získáme z intervalu spolehlivosti pro μ , nikoliv z p -hodnoty.

Jednovýběrový t -test

Příklad (Glykémie)

Testujme na hladině $\alpha = 0.05$ hypotézu $H_0: \mu = 6.0$, (střední hodnota lačné glykémie je 6.0 mmol/l), proti oboustranné alternativě $H_1: \mu \neq 6.0$.

Testová statistika $T=0.7845$, p-hodnota 0.4458 , 95% interval spolehlivosti $(5.58, 6.90)$, $\bar{X} = 6.24$.

P-hodnota je větší než $0.05 \rightarrow$ na hladině 0.05 nezamítáme H_0 , je možné, že glykémie mají stř. hodnotu 6.0 . Tomu odpovídá CI, který obsahuje testovanou hodnotu $\mu_0 = 6.0$.

Kdyby lékař chtěl jenom testovat, zda jeho pacienti nemají horší (vyšší) stř. hodnotu glykémie než je popsáno v literatuře, mohl by použít jednostranný t-test. Testová statistika zůstává stejná, p-hodnota se ale mění: $p = 0.2229$, CI: $(5.70, \infty)$. Ani v tomto případě by H_0 nezamítl.



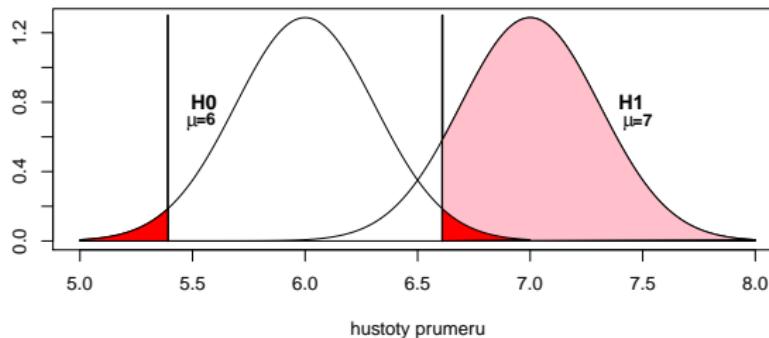
Přednáška 7 (31.3.2020) - obsah

- Síla jednovýběrového t -testu
- Předpoklady jednovýběrového t -testu, ověřování normality
- Dvouvýběrový t -test (testová procedura, souvislost s intervalem spolehlivosti, předpoklady, Welchův test)
- Párový t -test (testová procedura, předpoklady)
- Neparametrické testy (Wilcoxonův dvouvýběrový)

Síla jednovýběrového t -testu

- Síla testu $1 - \beta$ je pravděpodobnost, že zamítneme H_0 , pokud platí H_1 . Alternativní hypotéza H_1 obsahuje nekonečně mnoho možností, jak může vypadat skutečná μ (cokoliv jiného/většího/menšího než μ_0) a síla bude záležet na konkrétní hodnotě μ .
- Pokud skutečná hodnota μ bude daleko od testované hodnoty μ_0 , budeme zamítat nulovou hypotézu častěji (síla bude vyšší), než pokud μ bude blízko μ_0 .

Síla jednovýběrového t -testu



Příklad (Glykémie)

Kdybychom předpokládali známou směr. odchylku glykémií ($\sigma = 1.2$), můžeme nakreslit hustotu průměru z 15 pozorování za $H_0: \mu = 6$ a za alternativy $H_1: \mu = 7$. Svislé čáry ukazují oblasti zamítnutí. Za H_0 zamítáme s pravděpodobností 0.05 (červené oblasti), za uvedené H_1 s pravděpodobností 0.896 (růžová oblast).

Síla jednovýběrového t -testu

- Síla závisí na hladině testu, skutečném rozptylu a rozsahu výběru. Z těchto parametrů je obvykle možné ovlivnit pouze rozsah výběru.
- Pokud σ^2 je skutečný rozptyl a pro rozsah výběru n platí:

$$n \geq \left(\frac{z(1 - \alpha/2) + z(1 - \beta)}{\mu_1 - \mu_0} \right)^2 \cdot \sigma^2,$$

kde z jsou kvantily $N(0, 1)$, pak test na hladině α dosáhne síly aspoň $1 - \beta$ při skutečné stř. hodnotě μ_1 .

- Tento výpočet lze použít při plánování experimentu. Je ovšem třeba znát aspoň přibližně rozptyl.

Síla jednovýběrového t -testu

Příklad (Glykémie)

Kdybychom předpokládali známou směr. odchylku glykémií ($\sigma = 1.2$) a rozhodli se, že chceme sílu 0.8, pokud skutečná $\mu = 6.5$ ($H_0: \mu = 6$), potřebovali bychom aspoň

$$n \geq \left(\frac{1.96 + 0.84}{0.5} \right)^2 \cdot 1.44 = 45.16,$$

tedy aspoň 46 pozorování. Je vidět, že na detekci takového rozdílu je rozsah výběru 15 zcela nedostačující.

Pokud bychom chtěli sílu 0.9 na stejnou alternativu, pak

$$n \geq \left(\frac{1.96 + 1.28}{0.5} \right)^2 \cdot 1.44 = 60.47.$$

Pokud by nám stačila síla 0.8 na alternativu $\mu = 7$, pak nám 15 pozorování stačí, $n \geq (1.96 + 0.84)^2 \cdot 1.44 = 11.29$.

Předpoklady jednovýběrového t -testu

Jednovýběrový t -test bude fungovat tak jak má, pokud data opravdu pocházejí z modelu, který předpokládáme. Pokud tomu tak není, není zaručeno, že test dodržuje předepsanou hladinu.
Jaký model tedy předpokládáme?

- ① X_1, X_2, \dots, X_n jsou nezávislé.
- ② X_1, X_2, \dots, X_n mají všechny $N(\mu, \sigma^2)$

Z těchto předpokladů je důležitější nezávislost.

Předpoklad nezávislosti

- Nezávislost se z dat ověřit nedá, musíme usoudit na nezávislost z toho, jak byla data sebrána a z logiky věci.
- Pozor na skupinové závislosti. Př.- provádí experiment na krysách a mám krysy z několika vrhů. Pak data nebudou nezávislá, krysy z jednoho vrchu se budou možná chovat podobně, měření na krysách ze stejného vrchu mohou být závislá.
- Pozor na časové závislosti. Př.- měřím každý den koncentrace CO_2 na určitém místě. Pozorování, která jsou časově blízko, budou patrně podobná, budu mít v datech závislosti.
- Závislá data se musí analyzovat jinak, testy k tomu určenými.

Předpoklad normality

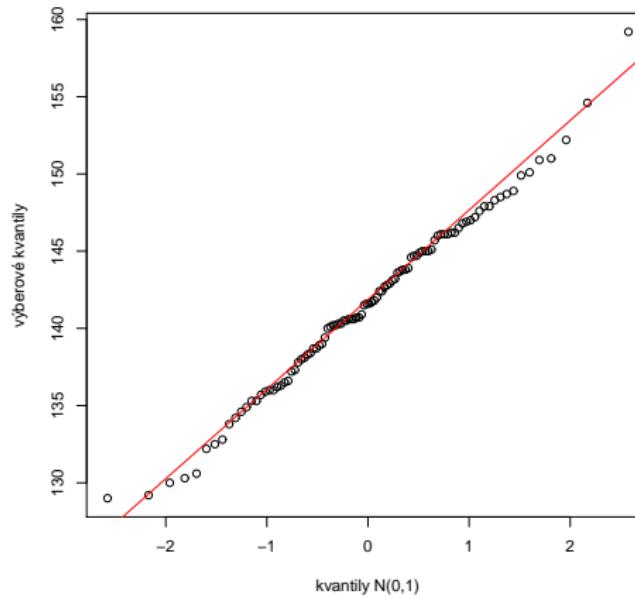
- U jednovýběrového t -testu i intervalu spolehlivosti potřebujeme normalitu k tomu, aby T statistika měla opravdu t -rozdělení.
- Stačí k tomu ovšem, aby normální byl průměr \bar{X} . Pokud máme dost pozorování, CLV nám zajistí normalitu průměru i když původní data normální nejsou.
- Máme-li tedy dost pozorování, nemusíme být s požadavkem normality dat příliš striktní.
- Nejnebezpečnější jsou situace, kdy máme málo dat, nebo situace, kdy máme velmi odlehlá pozorování.
- Normalitu můžeme ověřit na datech.

Ověřování normality

- Normalitu lze ověřit pomocí *normálního diagramu*.
- Jedná se o graf, kde vynášíme variační řadu pozorování (výběrové kvantily) proti kvantilům $N(0,1)$ odpovídajícím počtu pozorování (teoretické kvantily).
- Pokud data pocházejí z normálního rozdělení, body by měly ležet přibližně na přímce

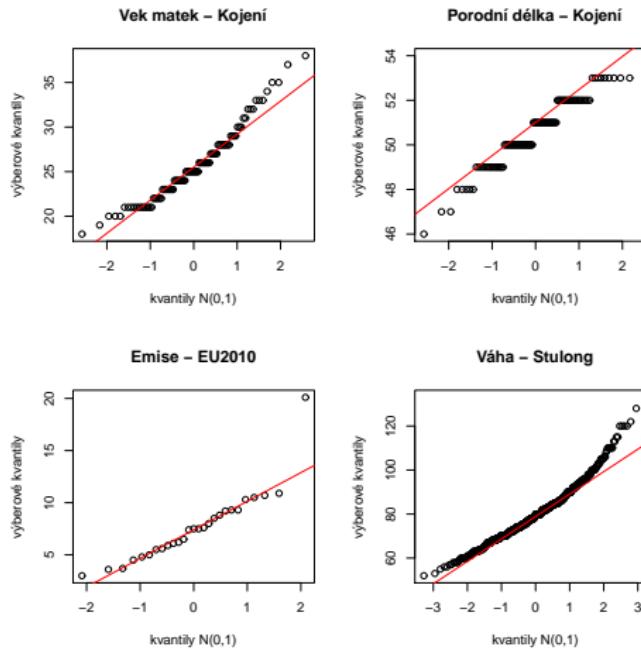
Normální diagram

Obrázek: Normální diagram - výšky dívek



Normální diagram

Obrázek: Normální diagramy - různá porušení normality



Test normality

- Existují testy normality, např, *Shapirův-Wilkův test*.
- U těchto testů je H_0 : Výběr má normální rozdělení, H_1 : Výběr nemá normální rozdělení.
- **Pozor!** Máme-li dost pozorování, test bude zamítat i na malé odchytky od normality, přitom nám malé odchytky nevadí, máme CLV.
- Většinou stačí podívat se na normální diagram. Pokud uděláte test normality, interpretujte výsledek opatrně, vezměte v úvahu kolik máte pozorování a **vždy** si nakreslete také normální diagram, abyste viděli, co se děje.

Dvouvýběrový problém

Příklad (Glykémie - 2 výběry)

Máme údaje o lačných glykémiích pacientů s diabetem 2 a údaje o lačných glykémiích pacientů, kteří kromě diabetu typu 2 mají také celiakii a jsou na bezlepkové dietě. Zajímá nás, zda pacienti obou skupin mají stejné rozdělení ranní glykémie.

Pouze diabetes

4.3	6.9	5.6	7.6	7.3	7.9	7.0	4.7	5.4
7.4	6.5	4.3	5.9	6.8	6.0			

Diabetes+celiakie

6.0	5.0	5.1	6.4	6.1	5.8	4.0	3.9	5.3
4.9	4.1	6.5	5.7	4.5	6.3	5.6	5.6	4.1

Budeme předpokládat, že v obou skupinách má ranní glykémie normální rozdělení se shodným rozptylem. Zajímá nás, zda také střední hodnota je v obou skupinách stejná. To zatím neumíme testovat.

Dvouvýběrový problém

- Předpokládejme, že máme dva výběry: X_1, X_2, \dots, X_n a Y_1, Y_2, \dots, Y_m z rozdělení $X \sim N(\mu_X, \sigma^2)$ a $Y \sim N(\mu_Y, \sigma^2)$.
Předpokládejme, že všechny veličiny X_i a Y_j jsou nezávislé (tedy nejen veličiny uvnitř výběrů, ale i veličiny z různých výběrů).
- Chceme testovat $H_0: \mu_X = \mu_Y$, proti alternativě $H_1: \mu_X \neq \mu_Y$ (případně proti jednostranným alternativám $H_1: \mu_X > \mu_Y$ nebo $H_1: \mu_X < \mu_Y$)
- O rozptylu σ^2 předpokládáme, že je stejný pro oba výběry, ale jeho hodnotu neznáme a nezajímáme se o ni.
- \bar{X} a \bar{Y} jsou nestranné odhady μ_X a μ_Y . Zdá se tedy rozumné založit testovou statistiku na rozdílu $\bar{X} - \bar{Y}$, který bude mít za H_0 normální rozdělení s nulovou střední hodnotou.
Potřebujeme ještě odhad rozptylu.

Testová statistika dvouvýběrového t -testu

- Rozptyl veličiny $\bar{X} - \bar{Y}$ odhadneme pomocí váženého průměru odhadů rozptylu v obou výběrech:

$$\begin{aligned} S^2 &= \frac{n-1}{n+m-2} S_X^2 + \frac{m-1}{n+m-2} S_Y^2 = \\ &= \frac{1}{n+m-2} \left(\sum_{i=1}^n (\bar{X}_i - \bar{X})^2 + \sum_{j=1}^m (\bar{Y}_j - \bar{Y})^2 \right) \end{aligned}$$

- Dá se ukázat, že $S \sqrt{\frac{n+m}{nm}}$ je střední chyba rozdílu $\bar{X} - \bar{Y}$.
- Dvouvýběrový t-test* založíme na statistice

$$T = \frac{\bar{X} - \bar{Y}}{S} \sqrt{\frac{nm}{n+m}},$$

která má za nulové hypotézy rozdělení t_{n+m-2} .

- Kritický obor testu určíme pomocí kvantilů t_{n+m-2} :

alternativa	test zamítne, pokud
$\mu_X \neq \mu_Y$	$ T \geq t_{n+m-2}(1 - \frac{\alpha}{2})$
$\mu_X > \mu_Y$	$T \geq t_{n+m-2}(1 - \alpha)$
$\mu_X < \mu_Y$	$T \leq -t_{n+m-2}(1 - \alpha)$

Dvouvýběrový t-test a interval spolehlivosti

- Označme $\delta = \mu_X - \mu_Y$, skutečný rozdíl středních hodnot výběrů. Nestranný bodový odhad δ je $\bar{X} - \bar{Y}$.
- Použijeme rozdělení testové statistiky T ke konstrukci intervalu spolehlivosti a dostaneme $(1 - \alpha)100\%$ interval spolehlivosti pro δ :
$$(\bar{X} - \bar{Y} - S \sqrt{\frac{n+m}{nm}} \cdot t_{n+m-2}(1 - \frac{\alpha}{2}), \bar{X} - \bar{Y} + S \sqrt{\frac{n+m}{nm}} \cdot t_{n+m-2}(1 - \frac{\alpha}{2}))$$
- Tento interval pokrývá skutečnou hodnotu δ s pravděpodobností $1 - \alpha$.
- Interval spolehlivosti je založen na testové statistice dvouvýběrového t -testu, existuje vzájemně jednoznačný vztah mezi testem a int. spolehlivosti: Dvouvýběrový t -test zamítá H_0 právě tehdy, když 0 není pokryta intervalm spolehlivosti.

Dvouvýběrový t-test

Příklad (Glykémie - 2 výběry)

Testujeme, zda stř. hodnoty lačných glykémií v obou skupinách jsou stejné: $H_0: \mu_X = \mu_Y$ proti oboustranné alternativě, že se střední hodnoty liší: $H_0: \mu_X \neq \mu_Y$.

Ve skupině diabetiků je průměrná glykémie $\bar{X} = 6.24 \text{ mmol/l}$, mezi diabetiky-celiaky $\bar{Y} = 5.27 \text{ mmol/l}$. Testová statistika dvouvýběrového t-testu: $T=2.7047$, p -hodnota 0.0110 .
95% interval spolehlivosti pro δ : $(0.24, 1.70)$

P -hodnota je nižší než $0.05 \rightarrow$ zamítáme H_0 . Na hladině 0.05 jsme prokázali, že stř. hodnoty nejsou stejné. Diabetici s celiakií mají stř. hodnotu glykémie nižší (plyne z CI), odhadem asi o 0.97 (rozdíl průměrů).

Předpoklady dvouvýběrového t-testu

Dvouvýběrový *t*-test bude fungovat tak jak má, pokud budou data skutečně pocházet ze situace, kterou předpokládáme:

- ❶ Všechny veličiny $X_1, X_2, \dots, X_n, Y_1, \dots, Y_m$ jsou mezi sebou nezávislé.
- ❷ Každý výběr má normální rozdělení.

Pozor! To neznamená, že oba výběry spojené dohromady jsou normální. Pokud jsou oba výběry normální a H_0 neplatí (stř. hodnoty jsou různé), spojený výběr normální nebude.

- ❸ Oba výběry mají shodné rozptyly.

Ověřování předpokladů dvouvýběrového t -testu

- Nezávislost je třeba posoudit z logiky experimentu a sběru dat.
- Normalitu lze ověřit pomocí normálních diagramů pro každý výběr zvlášť, případně pomocí testů normality. Dvouvýběrový test není příliš citlivý na odchylky od normality, při dostatečných počtech pozorování se průměry normalizují díky CLV a nehrozí problémy.
- Shodné rozptyly jsou problematičtějším předpokladem. Pokud nejsou splněny, mohou test zásadně poškodit, T statistika pak nemá očekávané rozdělení a p -hodnota neodpovídá skutečnosti.
- Neshodné rozptyly nepředstavují zásadní problém, pokud máme vyvážený experiment/šetření, tj. pokud mají obě skupiny stejný nebo aspoň přibližně stejný, dostatečně velký rozsah. Potom bude test přibližně dodržovat hladinu i při nestejných rozptylech.

Shodnost rozptylů u dvou výběrů

- Shodnost rozptylů lze posoudit graficky, např. pomocí boxplotů každého z výběrů.
- Klasický *F-test na shodnost rozptylů* testuje $H_0: \sigma_X^2 = \sigma_Y^2$ proti $H_1: \sigma_X^2 \neq \sigma_Y^2$ pomocí testové statistiky založené na poměru odhadů rozptylů:

$$F = \frac{S_X^2}{S_Y^2}$$

- Za H_0 bude F blízko 1, zamítáme pro F vzdálené od 1. Za nulové hypotézy $F \sim F_{n-1, m-1}$. Zamítáme H_0 pokud $\frac{S_X^2}{S_Y^2} \geq F_{n-1, m-1}(1 - \frac{\alpha}{2})$ nebo $\frac{S_Y^2}{S_X^2} \geq F_{m-1, n-1}(1 - \frac{\alpha}{2})$.
- Nevýhodou F -testu je jeho velká citlivost na normalitu dat (je mnohem citlivější než t -testy).
- Existují i jiné testy na shodnost rozptylů, např. *Bartlettův test*, který je však také citlivý na odchylky dat od normality. Více se doporučuje *Levenův test*, což je dvouvýběrový t -test provedený na odchylky od mediánu (či průměru).

Dvouvýběrový t-test při neshodných rozptylech výběrů

- Elegantním způsobem jak obejít přepoklad shodných rozptylů je použití *Welchova t-testu*.
- Welchův test nepředpokládá shodné rozptyly, rozptyl $\bar{X} - \bar{Y}$ odhaduje jako součet odhadů rozptylů průměrů. Testová statistika je potom:

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}}$$

- Za H_0 má T také t-rozdělení, počet stupňů volnosti však závisí na rozptylech a obecně není celočíselný.
- Welchův test se doporučuje, pokud nevíme, zda jsou rozptyly stejné, což je ale skoro vždy. Vzhledem k tomu, že za použití Welchova testu nejsme nijak penalizováni, doporučují používat ho jako první volbu, tím odpadne problém ověřování shodnosti rozptylů.
- R používá Welchův test. Pokud chcete klasický t-test, musíte změnit parametr příkazu.

Předpoklad shodných rozptylů

Příklad (Glykémie)

Odhadneme-li rozptyl glykémií pro pacienty, kteří mají pouze diabetes, dostáváme $S_X^2 = 1.404$, odhad pro diabetiky s celiakií je $S_Y^2 = 0.754$. Zdá se tedy, že předpoklad shodnosti rozptylů nejspíš splněn není.

P-hodnota F-testu je 0.2233, Levenova testu 0.2160. Žádný z testů nezmítí shodnost rozptylů, ale vzhledem k tomu, že odhady rozptylů se výrazně liší, je možné, že jsme nezmítli pouze kvůli malým výběrům. Zkusíme použít Welchův test.

Testová statistika Welchova dvouvýběrového t-testu: $T=2.6293$, p-hodnota 0.0144. 95% interval spolehlivosti pro δ : (0.21, 1.73) Závěr je stejný, zamítáme H_0 , střední hodnoty nejsou stejné.

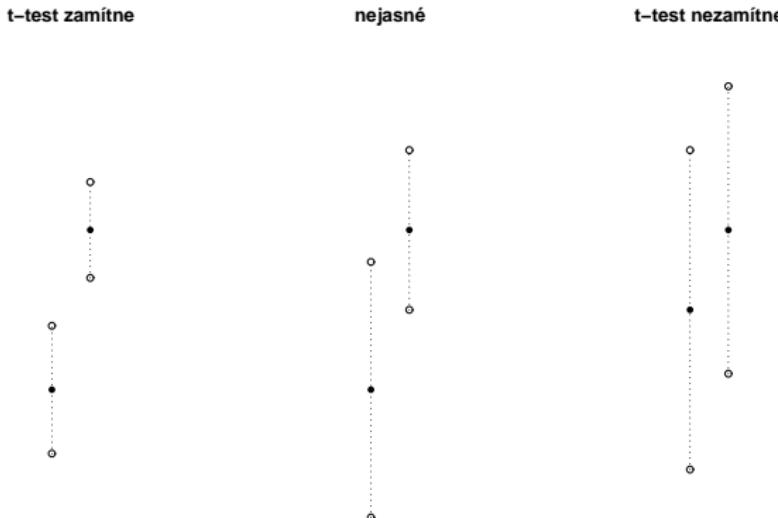
Vidíme, že T je trochu nižší, p-hodnota trochu vyšší a CI širší, než když jsme použili klasický t-test.

Dvouvýběrový test a intervaly spolehlivosti pro μ_X a μ_Y

- Lze z intervalů spolehlivosti pro střední hodnoty v každém výběru usoudit na výsledek dvouvýběrového t-testu?
- Pokud se $(1 - \alpha)\%$ CI pro stř. hodnoty obou výběrů vůbec nepřekrývají, pak dvouvýběrový t-test zamítne $H_0: \mu_X = \mu_Y$ na hladině α .
- Pokud se $(1 - \alpha)\%$ CI pro střední hodnoty překrývají, výsledek dvouvýběrového t-testu není jasný.
- Pokud se ovšem $(1 - \alpha)\%$ CI pro stř. hodnoty překrývají takovým způsobem, že každý z nich pokryje průměr druhého výběru, pak dvouvýběrový t-test na hladině α H_0 nezamítne.
- Všechna tato pravidla platí pro klasický dvouvýběrový t-test, nikoliv pro Welchův.

Dvouvýběrový test a intervaly spolehlivosti pro μ_X a μ_Y

Obrázek: Rozhodnutí podle intervalů spolehlivosti ve výběrech



Párová situace

Příklad (Glykémie - párová situace)

Tentokrát nás zajímá, zda se snížila střední hodnota lačné glykémie poté, co byl pacientům s diabetem nasazen nový lék. Každému pacientu byla znova změřena lačná glykémie po 2 měsících, kdy užívali nový typ léku na podporu tvorby inzulinu:

pacient	1	2	3	4	5	6	7	8
původně	4.3	6.9	5.6	7.6	7.3	7.9	7.0	4.7
po 2 měs.	4.5	6.9	4.4	7.7	6.4	7.5	7.5	4.1
pacient	9	10	11	12	13	14	15	
původně	5.4	7.4	6.5	4.3	5.9	6.8	6.0	
po 2 měs.	5.0	7.6	6.0	3.9	4.9	6.1	5.0	

Lze použít dvouvýběrový t-test? **Ne!** Nejsou splněny předpoklady, měření na jednom pacientovi nejsou nezávislá.

Párový t -test

- Máme dva výběry X_1, X_2, \dots, X_n a Y_1, Y_2, \dots, Y_n o stejném rozsahu, kde X_i a Y_i pro stejné i jsou potenciálně závislé (pozorování na stejném subjektu apod.).
- Můžeme se na data dívat jako na výběr páru $(\begin{smallmatrix} X_1 \\ Y_1 \end{smallmatrix}), (\begin{smallmatrix} X_2 \\ Y_2 \end{smallmatrix}), \dots, (\begin{smallmatrix} X_n \\ Y_n \end{smallmatrix})$, které jsou mezi sebou nezávislé. Uvnitř páru však závislost být může.
- V tomto případě **nelze** použít dvouvýběrový t -test pro porovnání středních hodnot X a Y , protože je porušen předpoklad nezávislosti.
- Veličiny $Z_i = X_i - Y_i$, $i = 1, 2, \dots, n$ budou ovšem nezávislé.
- Na veličinách Z_i pak můžeme testovat pomocí jednovýběrového t -testu hypotézu $H_0: \mu_Z = 0$, která je ekvivalentní testu $H_0: \mu_X = \mu_Y$.
- Tento test nazveme **párovým t -testem** a nejdá se o nic jiného než o jednovýběrový t -test provedený na rozdíly.

Párový t-test

Příklad (Glykémie)

Použijeme párový t-test na testování $H_0: \mu_X = \mu_Y$, kde μ_X je střední hodnota původních glykémií a μ_Y je stř. hodnota glykémií po 2 měsících na novém léku.

Průměr rozdílů $Z_i = X_i - Y_i$ je $\bar{Z} = 0.41$, testová statistika $T = 3.0681$, p -hodnota 0.0083 , $95\% \text{ CI}$ pro μ_Z je $(0.12, 0.69)$.

P-hodnota je menší než $0.05 \rightarrow$ zamítáme H_0 na pětiprocentní hladině. Nula není pokryta intervalem spolehlivosti, což odpovídá výsledku testu. Lék patrně snížil stř. hodnotu lačné glykémie.

Odhadnutý rozdíl mezi stř. hodnotami je asi 0.41 mmol/l . Podle CI to vypadá, že skutečný rozdíl mezi středními hodnotami je aspoň 0.12 a nejvýše 0.69 .

Předpoklady párového *t*-testu

Předpoklady párového *t*-testu jsou shodné s předpoklady jednovýběrového *t*-testu, aplikovanými na rozdíly $Z_i = X_i - Y_i$.

- ❶ Z_1, Z_2, \dots, Z_n jsou nezávislé náhodné veličiny, což je splněno, pokud páry $(\frac{X_1}{Y_1}), (\frac{X_2}{Y_2}), \dots, (\frac{X_n}{Y_n})$ jsou mezi sebou nezávislé.
- ❷ Všechny Z_i mají rozdělení $N(\mu_Z, \sigma_Z^2)$. Tento předpoklad bude splněn, pokud X a Y jsou normální, ale není to nutné. Je řada situací, kdy Z je normální, aniž by X a Y byly normální, takže pokud ověřujete normalitu, dívejte se na rozdíly Z .
Podobně jako u jiných *t*-testů, test není příliš citlivý na odchylky od normality. Při dostatečném počtu pozorování je možné se spolehnout na CLV.

Neparametrické testy

- Pokud máme výběry z rozdělení, které se výrazně odchyluje od normálního a nemáme dost pozorování na to, abychom se mohli spolehnout na CLV, můžeme použít pro porovnání dvou výběrů neparametrické testy.
- *Neparametrické testy* nepředpokládají žádné specifické rozdělení výběrů. Musíme se však smířit s tím, že jejich možnosti jsou omezené.
- Nulová hypotéza vypadá jinak než u t -testů, netestujeme tvrzení o středních hodnotách a nemůžeme tudíž dostat intervaly spolehlivosti pro μ nebo jejich rozdíl.
- Proto je praktičtější použít t -test, pokud je to možné, neparametrické testy by měly být až druhou volbou.
- Neparametrické testy často využívají pořadí pozorování ve výběrech, v těchto případech mluvíme také o *pořadových testech*.

Wilcoxonův dvouvýběrový test

- Nechť X_1, X_2, \dots, X_n a Y_1, Y_2, \dots, Y_m jsou dva nezávislé výběry ze spojitého rozdělení. Chceme porovnat jejich rozdělení, ale víme, že nejsou normální a nemůžeme použít dvouvýběrový t -test.
- Budeme testovat obecnou hypotézu H_0 : Rozdělení X a Y je stejné, (tj. není rozdíl v mích plohy, speciálně v mediánech) proti alternativě, že stejně není.
- Najdeme pořadí pozorování R_1, \dots, R_{n+m} ve spojeném výběru $X_1, \dots, X_n, Y_1, \dots, Y_m$.
- Testová statistika *Wilcoxonova dvouvýběrového testu* je

$$W_X = \sum_{i=1}^n R_i$$

tj., je to součet pořadí prvního výběru.

Wilcoxonův dvouvýběrový test

- Pokud platí H_0 , bude pořadí náhodnou permutací čísel $1, 2, \dots, n + m$. Pokud se rozdelení v mediánu liší, např. X má vyšší medián, budou pořadí prvního výběru patrně vyšší a naopak. Příliš vysoké nebo příliš nízké hodnoty W_X budou tedy svědčit proti H_0 .
- Kritický obor nebo p -hodnotu získáme z přesného rozdělení W_X za H_0 nebo pomocí normální approximace.
- *Mannův-Whitneyův test* používá ve stejné situaci jako testovou statistiku počet takových dvojic (X_i, Y_j) , kde $X_i > Y_j$.
- Dá se ukázat, že test založený na Mannově-Whitneyově statistice je ekvivalentní Wilcoxonovu testu.

Použití Wilcoxonova dvouvýběrového testu

- Wilcoxonův dvouvýběrový test je vhodný zejména v situacích, kdy máme výrazně odlehlá pozorování a předpokládáme podobný tvar rozdělení v obou skupinách.
- Je možné použít Wilc. test, pokud data mají ve skutečnosti spíše ordinální než intervalové měřítko, tj. pokud vzdálenosti mezi měřeními nejsou jasně definovány a známe pouze pořadí.
- Wilcoxonův test bude fungovat velmi dobře, pokud X a Y mají stejné rozdělení, pouze posunuté v mediánu.
- Pokud test zamítne nulovou hypotézu, nemusí být jednoduché kvantifikovat, jak se výběry liší. Je možné uvést rozdíl mediánů, pokud věříme, že rozdíly výběrů spočívají hlavně v posunutí.
- Pokud jsou ve výběrech shodná pozorování, výpočet testu může být komplikovanější.

Wilcoxonův dvouvýběrový test

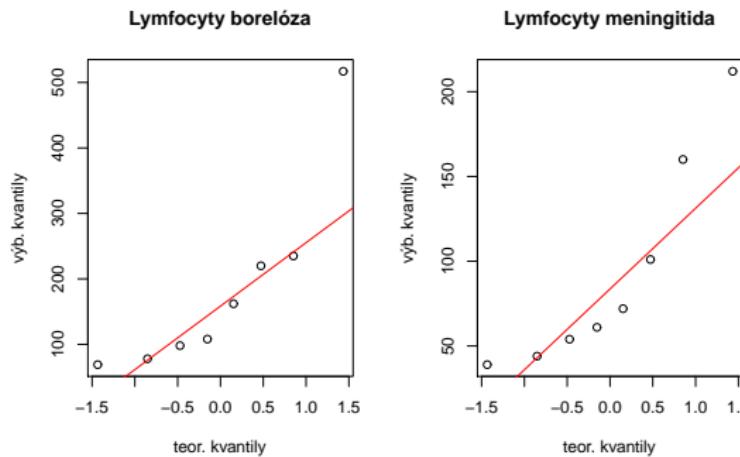
Příklad (Infekce CNS)

Lékaři zkoumali, zda se liší imunitní reakce při různých infekcích mozku, napozorovali počty lymfocytů v 1 mm^3 mozkomíšního moku u pacientů s boreliózou a meningitidou:

borelioza	220	517	69	235	162	108	98	78
meningitida	54	39	44	72	212	160	61	101

Je vidět, že počty lymfocytů vykazují v obou skupinách rozdělení se značně odlehlymi hodnotami.

Wilcoxonův dvouvýběrový test



Normální diagramy jasně ukazují nenormalitu dat, výběry jsou malé, usoudíme, že bude vhodné použít Wilcoxonův dvouvýběrový test.

Wilcoxonův dvouvýběrový test

Příklad (Infekce CNS)

Pořadí boreliáz mezi všemi infekcemi podle lymfocytů: 14, 16, 5, 15, 12, 10, 8, 7. Součet těchto pořadí je $W_X = 87$. Použijeme-li přesné rozdělení statistiky W_X , dostaneme p-hodnotu 0.04988.

Na hladině 0.05 tedy zamítáme hypotézu H_0 , že obě infekce vykazují stejné rozdělení lymfocytů. Rozdílem mezi mediány ($\tilde{X} = 135$ pro boreliózu a $\tilde{X} = 66.5$ pro meningitidu) můžeme popsat rozdíl mezi skupinami.

Přednáška 8 (7.4.2020) - obsah

- Neparametrické testy (znaménkový, Wilcoxonův párový, Kolmogorovův-Smirnovův)
- Analýza rozptylu - jednoduché třídění (formulace problému, model, rozklad součtu čtverců, tabulka analýzy rozptylu, předpoklady)
- Kruskalův-Wallisův test

Porovnání rozdělení dvou výběrů

- Velmi obecné porovnání rozdělení dvou výběrů X_1, \dots, X_n a Y_1, \dots, Y_m poskytuje *Kolmogorovův-Smirnovův test*, který využívá odhadů distribučních funkcí v obou výběrech.
- Testovou statistikou je maximální rozdíl empirických distribučních funkcí (největší svislá vzdálenost):

$$D = \max |\hat{F}_X(t) - \hat{F}_Y(t)|$$

- Zatímco dvouvýběrový Wilcoxonův test je citlivý zejména na posunutí v poloze výběrů, Kolmogorovův-Smirnovův je citlivý na jakékoli rozdíly mezi rozděleními výběrů.
- Pokud se rozdělení výběrů liší posunutím, pak je ovšem lepší použít Wilcoxonův test, který bude mít větší sílu.
Kolmogorovův-Smirnovův test je tedy vhodný pouze když chceme detegovat jakékoli neshody mezi rozděleními.

Kolmogorovův-Smirnovův test

Příklad (Věk matek podle délky kojení)

V datech Kojení nás zajímá, zda se liší rozdělení věku matek, které kojily aspoň 24 týdnů od rozdělení věku matek, které kojily kratší dobu. Máme podezření, že se nebude jednat o posunutí.

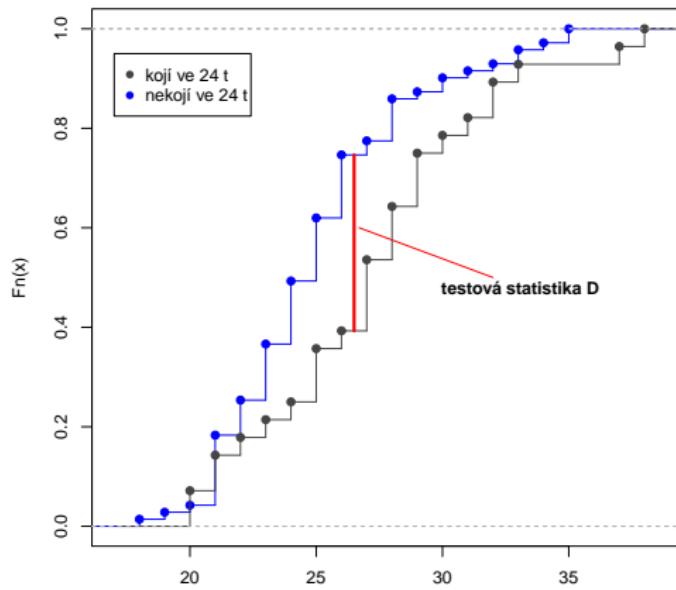
50 ze 71 matek, které již ve 24 týdnech nekojily (tj. 70.4%) bylo věkem mezi 21-26 lety. Žádnou takovou koncentraci mezi matkami, které ještě ve 24 týdnech kojily (28 matek) nepozorujeme.

Použijeme-li obecný Kolmogorovův-Smirnovův test, zjistíme, že maximální svislá vzdálenost mezi empirickými distribučními funkcemi je $D = 0.3536$, p -hodnota testu 0.0132. Test tedy zamítá shodnost rozdělení, neposkytuje však žádnou další informaci.

Obrázek empirických dist. fcí potvrzuje, že se nejedná o posunutí, svislá vzdálenost je na začátku a na konci sledovaného intervalu minimální.

Kolmogorovův-Smirnovův test

Obrázek: Empirické distribuční funkce kojící/nekojící ve 24 t.



Znaménkový test

- Mějme nyní párovou situaci: $(\frac{X_1}{Y_1}, \frac{X_2}{Y_2}, \dots, \frac{X_n}{Y_n})$, výběr nezávislých párů. Předpokládejme, že rozdíly $X_i - Y_i$ mají spojité, ale výrazně nenormální rozdělení a pozorování není dost na to, abychom se odvolali na CLV.
- Pokud X a Y mají stejné rozdělení, pak rozdíly Z_i mají rozdělení symetrické kolem 0, populační medián rozdělení Z je 0. Tuto hypotézu lze testovat pomocí *znaménkového testu*
- Znaménkový test testuje hypotézu H_0 :
 $P(X_i \geq Y_i) = P(Y_i \geq X_i) = \frac{1}{2}$. Testová statistika W je počet dvojic, kde $X_i > Y_i$. Pokud nejsou žádné dvojice, kde $X_i = Y_i$, za H_0 platí $W \sim Bi(n, \frac{1}{2})$. Pokud někde nastává rovnost, tyto dvojice vynecháme a snížíme počet pokusů v binomickém rozdělení.

Znaménkový test

- P-hodnotu nebo kritický obor znaménkového testu určíme pomocí binomického rozdělení W za nulové hypotézy..
- Je možné approximovat binomické normálním. Přibližně platí:

$$Z = \frac{W - \frac{n_1}{2}}{\sqrt{\frac{n_1}{4}}} \sim N(0, 1),$$

kde n_1 je počet párů, kde $X_i \neq Y_i$.

- Znaménkový test používá pouze pořadí uvnitř párů, je tedy použitelný i v situaci, kdy výběry nejsou k dispozici a známe pouze pořadí uvnitř párů.
- Znaménkový test je možné použít i v jednovýběrové situaci, kdy u rozdělení výběru X_1, X_2, \dots, X_n chceme testovat, že populační medián je roven x_0 , tj.

$$H_0: P(X_i \leq x_0) = P(X_i \geq x_0) = \frac{1}{2}.$$

Wilcoxonův párový test

- V párové situaci s neznámým rozdělením výběrů lze použít také *Wilcoxonův párový test*. Testuje
 $H_0: Z_i = X_i - Y_i$ má symetrické rozdělení kolem 0.
- Test používá pořadí R_1, R_2, \dots, R_n výběru $|Z_1|, |Z_2|, \dots, |Z_n|$. Testová statistika je

$$W = \sum_{i:X_i > Y_i} R_i$$

- Pokud bychom místo pořadí absolutních hodnot přidávali jedničku za každu dvojici, kde $X_i > Y_i$, dostaneme znaménkový test. Wilcoxonův párový se tedy liší od znaménkového tím, že bere v úvahu také pořadí velikosti rozdílů, nejen jejich znaménko.
- P-hodnota testu se získá bud' z přesného rozdělení statistiky W nebo pomocí normální approximace.
- Wilcoxonův párový test lze použít i v jednovýběrové situaci.

Párové neparametrické testy

Příklad (Leukocyty)

Lékaři zkoumali, zda v průběhu infekční mononukleózy klesají počty leukocytů v krvi. U 15 pacientů měli měření při diagnóze nemoci a po dvou týdnech nemoci. V tabulce jsou počty bílých krvinek v $10^9/l$.

<i>začátek</i>	4.7	8.9	7.3	3.8	4.3	9.1	8.7	6.0
<i>po 14 dnech</i>	4.1	8.7	4.2	4.3	6.6	7.9	8.4	6.2
<i>začátek</i>	6.9	5.4	8.8	7.2	4.5	6.5	8.0	
<i>po 14 dnech</i>	6.0	5.2	8.7	6.8	3.7	6.4	3.1	

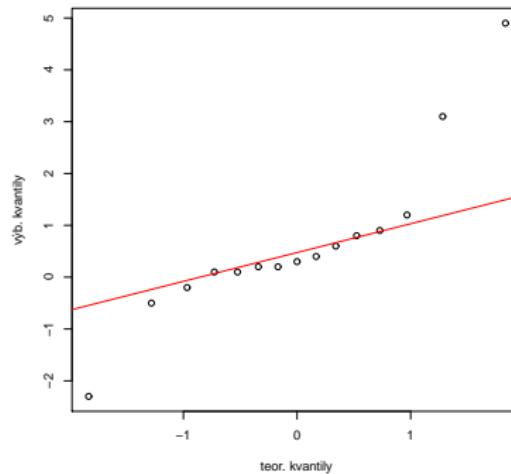
Rozdíly v počtu bílých krvinek u pacientů (uspořádané):

<i>rozdíl 1.-2.</i>	-2.3	-0.5	-0.2	0.1	0.1	0.2	0.2	0.3
	0.4	0.6	0.8	0.9	1.2	3.1	4.9	

Ve variační řadě vidíme odlehlá pozorování, rozdíly patrně nebudou normální.

Párové neparametrické testy

Obrázek: Normální diagram rozdílů počtu leukocytů



Normální diagram ukazuje rozdělení s těžkými chvosty, vzhledem k malému počtu pozorování použijeme raději neparametrické testy.

Párové neparametrické testy

Příklad (Leukocyty)

Použijeme nejprve znaménkový test. Testujeme H_0 : medián Z_i je 0 proti jednostranné alternativě H_1 : medián Z_i je větší než 0, tj. medián počtu leukocytů se snížil.

U 12 pacientů se počty leukocytů snížily, u 3 se zvýšily. $W = 12$, za H_0 platí $W \sim Bi(15, 0.5)$. P-hodnota je $P(W \geq 12 | H_0) = 0.0176$, takže bychom zamítli H_0 ve prospěch alternativy, že medián je větší než 0, tj. počet leukocytů na počátku nemoci má vyšší medián.

Normální approximace vydá testovou statistiku $Z = 2.3238$, p-hodnota pro jednostranný test je $P(Z \geq 2.3238 | H_0) = 0.0102$, samozřejmě také zamítáme H_0 .

Kdybychom použili Wilcoxonův párový test, dostaneme $W=95$, p-hodnota jednostranného testu 0.0249, takže také zamítáme H_0 .

Určitý odhad rozdílu v míře polohy obou rozdělení nám dá rozdíl mediánů 0.7 ($10^9/l$), CI ale nezískáme.

Přehled *t*-testů

Tabulka: Testy o poloze jednoho nebo dvou výběrů

situace	normální výběry	nenormální výb, n nízké
1 výběr	jednovýběrový <i>t</i> -test	znaménkový nebo Wilcoxonův párový
2 výběry nezávislé	dvouvýběrový <i>t</i> -test	Wilcoxonův dvouvýběrový (Kolmogorovův-Smirnovův)
2 výběry párové závislosti	párový <i>t</i> -test	znaménkový nebo Wilcoxonův párový

Porovnání středních hodnot více výběrů

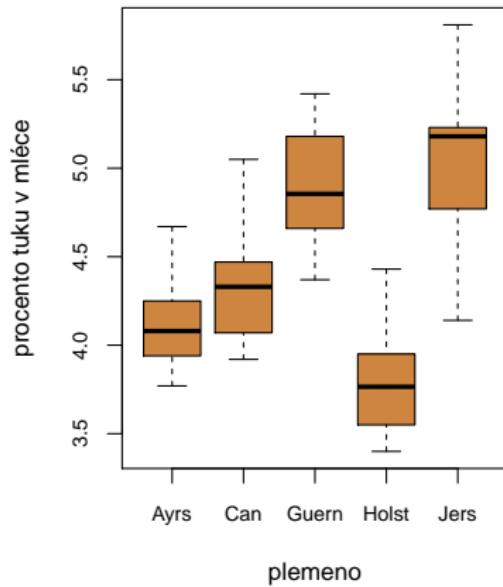
Příklad (Tuk v mléce podle plemena krávy)

- Byla sebrána data o obsahu tuku v mléce padesáti krav pěti různých plemen: skot Ayrshirský, Kanadský, Guernseyský, Holštýnský a Jerseyský. Chtěli bychom testovat, zda jsou střední hodnoty obsahu tuku u všech plemen stejné nebo zda se liší.
- Mohlo by se zdát, že by šlo použít dvouvýběrový *t-test* na každou dvojici výběrů. Uvědomme si však, že při takovém postupu bychom prováděli $\binom{5}{2} = 10$ testů. Každý z testů má hladinu α , takže pravděpodobnost chyby 1. druhu by mohla být při takovém testování dosti vysoká, v krajním případě až 10α .
- Potřebujeme test, který dodrží předepsanou hladinu α , test který bude testovat shodnost stř. hodnot k výběrům najednou ($k > 2$).
- Problém se dá formulovat také jako problém závislosti kvantitativní veličiny a veličiny kvalitativní (závislost obsahu tuku v mléce na plemenu krávy).

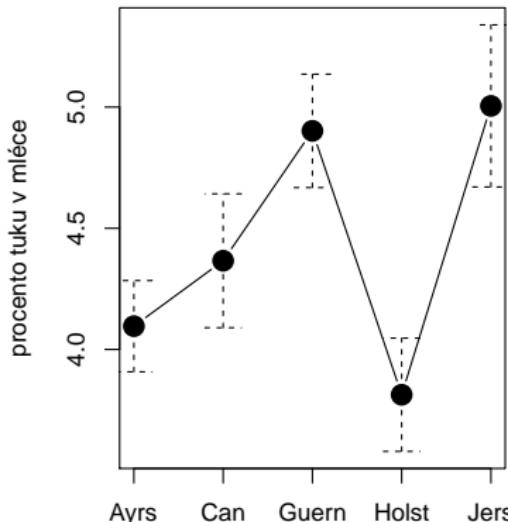
Porovnání středních hodnot více výběrů

Obrázek: Porovnání procenta tuku v mléce pěti plemen skotu

Krabicové diagramy



Prumery a CI (95%)



Model pro jednoduché třídění

- Obecně máme k nezávislých náhodných výběrů:
 $Y_{11}, Y_{12}, \dots, Y_{1n_1}$ z rozdělení $N(\mu_1, \sigma^2)$
 $Y_{21}, Y_{22}, \dots, Y_{2n_2}$ z rozdělení $N(\mu_2, \sigma^2)$
...
 $Y_{k1}, Y_{k2}, \dots, Y_{kn_k}$ z rozdělení $N(\mu_k, \sigma^2)$.
- Předpokládáme, že každý z výběrů pochází z normálního rozdělení a že mají stejný rozptyl σ^2 . Důležitý je předpoklad nezávislosti všech veličin $Y_{11}, Y_{12}, \dots, Y_{kn_k}$.
- Chceme testovat $H_0: \mu_1 = \mu_2 = \dots = \mu_k$ proti $H_1:$ neplatí H_0 , tj. aspoň pro nějaké dva výběry střední hodnota není stejná.
- Potřebujeme tedy metodu, která je zobecněním dvouvýběrového t -testu na více výběrů.

Model pro jednoduché třídění

- Uvedenou úlohu můžeme vyjádřit také následujícím způsobem: Pro všechna $i = 1, \dots, k$, $j = 1, \dots, n_i$ splňují Y_{ij} model:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij},$$

kde ε_{ij} jsou nezávislé náhodné veličiny s rozdelením $N(0, \sigma^2)$.

- Tato parametrizace odpovídá představě, že

$$\mu_i = \mu + \alpha_i,$$

tj., že střední hodnoty výběrů jsou součtem nějaké celkové úrovně výběrů (μ) a i -tého efektu (α_i). α_i je odchylka i -tého výběru od celkové úrovně, vliv i -té úrovně kvalitativního znaku na stř. hodnotu znaku kvantitativního.

- Bez dalších podmínek by modelu odpovídalo nekonečně mnoho parametrů, proto zavádíme reparametizační podmítku:
 $\sum_{i=1}^k \alpha_i = 0$. Pak budou α_i skutečně odchylky od celkové úrovně.
- Chceme testovat $H_0: \alpha_1 = \alpha_2 = \dots = \alpha_k = 0$.

Rozklad součtu čtverců

- Označme průměry jednotlivých výběrů a celkový průměr:

$$\bar{Y}_{i\cdot} = \frac{\sum_{j=1}^{n_i} Y_{ij}}{n_i} \quad \text{a} \quad \bar{Y}_{..} = \frac{\sum_{i=1}^k \sum_{j=1}^{n_i} Y_{ij}}{n}$$

- Celkovou variabilitu v datech (totální součet čtverců S_T) lze vyjádřit jako součet dvou součtů čtverců ($S_A + S_e$):

$$S_T = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2 = \sum_{i=1}^k n_i (\bar{Y}_{i\cdot} - \bar{Y}_{..})^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\cdot})^2$$

- Součet čtverců S_A bude vysoký, pokud se průměry ve výběrech hodně liší od celkového průměru. *Reziduální součet čtverců S_e* bude vysoký, pokud je velká variabilita uvnitř výběrů.
- Každému ze součtů čtverců odpovídá určitý počet stupňů volnosti (označujeme f): $f_T = n - 1$, $f_A = k - 1$, $f_e = n - k$.

Testová statistika

- Test bude založen na porovnání variability mezi výběry a uvnitř výběrů. Protože tato metoda využívá rozklad rozptylu dat, nazývá se *analýza rozptylu*, nebo *ANOVA (analysis of variance)*.
- Testová statistika je:

$$F_A = \frac{S_A/(k-1)}{S_e/(n-k)}$$

- Za H_0 bude $F_A \sim F_{k-1, n-k}$. Proti H_0 budou svědčit vysoké hodnoty F_A .
- Kritický obor tedy určí kvantil F-rozdělení, H_0 zamítneme, pokud

$$F_A \geq F_{k-1, n-k}(1-\alpha).$$

- Tento nejjednodušší případ analýzy rozptylu nazýváme *jednoduchým tříděním* - třídíme pozorování podle jediného faktoru s k úrovněmi odpovídajícími výběrům.

Tabulka analýzy rozptylu

- Všechny výpočty lze shrnout do tabulky:

variabilita	součet čtverců S	počet st. volnosti f	prům. čtverec S/f	testová stat. F	p-hodnota
výběry reziduální	S_A S_e	$f_A = k - 1$ $f_e = n - k$	S_A/f_A S_e/f_e	F_A	p
celková	S_T	$n - 1$			

- Reziduální součet čtverců lze použít také k odhadu společného rozptylu výběrů. Odhadem rozptylu je $\hat{\sigma}^2 = S^2 = S_e/f_e$.

Tabulka analýzy rozptylu

Příklad (Tuk v mléce podle plemena krávy)

Výsledky pro test, zda stř. hodnota obsahu tuku v mléce souvisí s plemenem krávy jsou v tabulce analýzy rozptylu:

<i>variabilita</i>	<i>součet čtverců</i> <i>S</i>	<i>počet st. volnosti</i> <i>f</i>	<i>prům. čtverec</i> <i>S/f</i>	<i>testová stat.</i> <i>F</i>	<i>p-hodnota</i>
<i>plemena reziduální</i>	10.495	4	2.624	20.16	$1.43 \cdot 10^{-9}$
	5.857	45	0.130		
<i>celková</i>	16.352	49			

P-hodnota je velmi nízká, zamítáme nulovou hypotézu, na hladině 0.05 prokazujeme, že se plemena liší co do stř. hodnoty obsahu tuku v mléce.

Bonferroniho metoda

- F -test zodpoví základní otázku, zda se liší alespoň některé stř. hodnoty výběrů. Nedá nám však odpověď na to, které dvojice výběrů to jsou. Tuto otázku zodpoví *mnohonásobné porovnávání*, které provádíme, pokud F -test zamítne H_0 .
- Provádění opakovaných t -testů nevede ke správnému řešení, nebyla by dodržena hladina testu, prováděli bychom $r = k(k - 1)/2$ testů.
- Metod mnohonásobného porovnávání existuje víc. Můžeme použít *Bonferroniho metodu*, která zamítá shodnost stř. hodnoty i -tého a j -tého výběru, pokud

$$|\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot}| \geq t_{n-k}(1 - \alpha/r) \sqrt{S^2 \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}$$

- Bonferroniho metoda (na rozdíl od r dvouvýběrových t -testů) změní kvantil t -rozdělení podle počtu porovnání a pro všechny testy používá stejný odhad rozptylu S^2 spočítaný ze všech výběrů.

Tukeyova metoda

- Při vysokém k může být kvantil používaný Bonferronihho metodou až příliš vysoký. Na druhou stranu, pokud nechceme porovnávat úplně všechny skupiny mezi sebou (třeba jen každou s nich s jedinou kontrolní), můžeme r snížit podle skutečného počtu porovnání.
- Mnohonásobné porovnávání je možné provést také *Tukeyovou metodou*. Ta zamítá shodnost stř. hodnoty i -tého a j -tého výběru, pokud

$$|\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot}| \geq q_{k,n-k}(1-\alpha) \sqrt{\frac{S^2}{2} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)},$$

kde $q_{k,n-k}(1-\alpha)$ je kvantil studentizovaného rozpětí (hodnoty jsou tabelovány).

Mnohonásobná porovnávání

Příklad (Tuk v mléce podle plemena krávy)

Celkový průměr procent tuku v mléce je $\bar{Y}_{..} = 4.437$. V tabulce jsou průměry ve výběrech jednotlivých plemen:

plemeno \bar{Y}_i	Ayrshire	Canadian	Guernsey	Holstein	Jersey
4.096	4.366	4.902	3.813	5.005	

Provedli jsme 10 porovnání pomocí Tukeyovy metody. V následující tabulce jsou odhady rozdílů stř. hodnot a p-hodnoty testů:

plemena	odh rozd.	p-hod	plemena	odh rozd.	p-hod
A-C	-0.270	0.4602	C-H	0.553	0.0110*
A-G	-0.806	0.0001*	C-J	-0.639	0.0024*
A-H	0.283	0.4125	G-H	1.089	0.0001*
A-J	-0.909	0.0001*	G-J	-0.103	0.9679
C-G	-0.536	0.0146*	H-J	-1.192	0.0001*

Jersey a Guernsey mají vyšší stř. hodnotu tuku než ostatní plemená.

Rozdíl mezi Guernsey a Jersey jsme neprokázali.

U tří plemen s nižším obsahem tuku jsme prokázali pouze rozdíl mezi Canadian a Holstein, další rozdíly průkazné nebyly.

Předpoklady jednoduchého třídění

Předpoklady jednoduchého třídění vyplývají z předpokládaného modelu:

- ① Nezávislost všech veličin: $Y_{11}, Y_{12}, \dots, Y_{kn_k}$ (i z různých výběrů).
- ② Normalita všech výběrů.
- ③ Stejný rozptyl σ^2 u všech k výběrů.

Předpoklady jsou tedy velmi podobné předpokladům douveyběrového t -testu, pouze převedené do situace více výběrů.

Ověření předpokladů

- Předpoklad nezávislosti nelze ověřit z dat, nezávislost můžeme pouze předpokládat z logiky experimentu nebo způsobu sběru dat.
- Normalitu lze ověřit pomocí normálních diagramů pro každý výběr zvlášť.
- Je také možné spočítat *rezidua*:

$$\hat{\varepsilon}_{ij} = Y_{ij} - \bar{Y}_i.$$

Rezidua jsou odhady "chyb" ε_{ij} , které by měly mít všechny stejné normální rozdělení. Můžeme tedy použít rezidua ze všech výběrů a nakreslit jeden normální diagram.

- Analýza rozptylu není příliš citlivá na mírné odchylky od normality. Obezřetní musíme být hlavně v situaci s malými rozsahy výběrů a odlehlymi pozorováními.

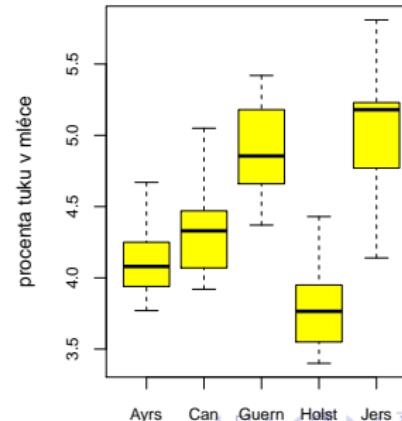
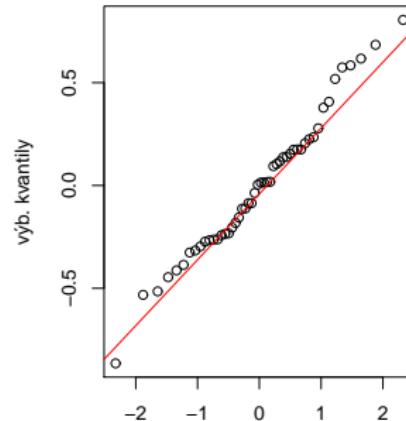
Ověření předpokladů

- Analýza rozptylu je citlivá na neshodné rozptyly ve výběrech. Pokud tento předpoklad není splněn, skutečná hladina testu může být jiná, test funguje špatně.
- Shodnost rozptylů lze ověřit z dat. Je možné nakreslit krabicové diagramy všech výběrů a porovnat je. Častým problémem může být zvětšování rozptylu při vyšších středních hodnotách, což se na krabicových diagramech projeví.
- Lze také testovat shodnost rozptylů, např. pomocí Levenova testu (jednoduché třídění provedené na absolutní hodnoty odchylek od průměrů v každém výběru) nebo pomocí Bartlettova testu (ten je však dosti citlivý na porušení normality).

Ověření předpokladů

Příklad (Tuk v mléce podle plemena krávy)

- ① Normalitu ověříme pomocí normálního diagramu reziduí.
- ② Pro ověření shodných rozptylů použijeme boxploty. Levenův test vydal p -hodnotu 0.7259, výsledek je v souladu se shodnými rozptyly.



Co když nejsou splněny předpoklady?

- Pokud jsou v datech závislosti, je třeba použít jinou metodu. Např. při skupinových závislostech je možné použít analýzu rozptylu s náhodnými bloky.
- Mírné odchyly od normality obvykle nevadí. Při zásadnějších problémech s normalitou lze použít monotónní transformaci, která data znormalizuje (logaritmus). Tento postup často také vyřeší problém s vyššími rozptyly ve výběrech s vyššími průměry.
- Jiná možnost je použít neparametrický test pro porovnání několika výběrů (Kruskalův-Wallisův test).
- V případě, že nejsou splněny shodné rozptyly, ale normalita výběru vcelku odpovídá, je možné použít Welchovu úpravu jednoduchého třídění, která shodné rozptyly nepožaduje. Není však příliš vhodná při malých rozsazích výběrů.
- Pokud jsou výběry dostatečně velké a data jsou vyvážená (všechny výběry mají přibližně stejný rozsah), ani neshodné rozptyly obvykle příliš nevadí.

Vícevýběrový problém

- Pokud máme výrazný problém s normalitou výběrů, můžeme použít neparametrický *Kruskalův-Wallisův test*. Jedná se o zobecnění Wilcoxonova dvouvýběrového testu na více výběrů.
- Jsme v situaci, kdy máme k nezávislých náhodných výběrů:
 $Y_{11}, Y_{12}, \dots, Y_{1n_1}$
 $Y_{21}, Y_{22}, \dots, Y_{2n_2}$
...
 $Y_{k1}, Y_{k2}, \dots, Y_{kn_k}.$
- Předpokládáme, že každý z výběrů pochází nějakého spojitého rozdělení. Označme celkový počet pozorování
 $n = n_1 + \dots + n_k.$
- Chceme testovat H_0 : Všechny výběry pocházejí ze stejného rozdělení, proti alternativě, že tomu tak není.

Kruskalův-Wallisův test

- Testová statistika je založena na pořadích. Označme R_{ij} pořadí pozorování Y_{ij} ve spojeném výběru. Dále označme

$$T_i = \sum_{j=1}^{n_i} R_{ij}$$

součet pořadí v i -té výběru.

- Testová statistika *Kruskalova-Wallisova* testu je

$$Q = \frac{12}{n(n+1)} \sum_{i=1}^k \frac{T_i^2}{n_i} - 3(n+1)$$

- Za H_0 platí přibližně: $Q \sim \chi_{k-1}^2$. Proti nulové hypotéze svědčí vysoké hodnoty Q , zamítáme H_0 pro $Q \geq \chi_{k-1}^2(1-\alpha)$.

Kruskalův-Wallisův test

Příklad (Kvalita života dětí s diabetem)

Lékaři zkoumali, jak léčba ovlivňuje kvalitu života dětí s diabetem. Z dotazníků byly spočítány skóry kvality života na stupnici 0-100.

Řada dětí s diabetem měla také autoimunitní onemocnění štítné žlázy (TD -thyroid disease). Lékaře zajímalo, zda toto onemocnění, případně jeho léčba, má vliv na kvalitu života. Rozdělili děti na 3 skupiny: pouze diabetes (10 dětí), diabetes+TD bez léčby (8 dětí), diabetes+TD léčené (8 dětí) a chtěli testovat, zda se rozdělení skóřů ve skupinách liší.

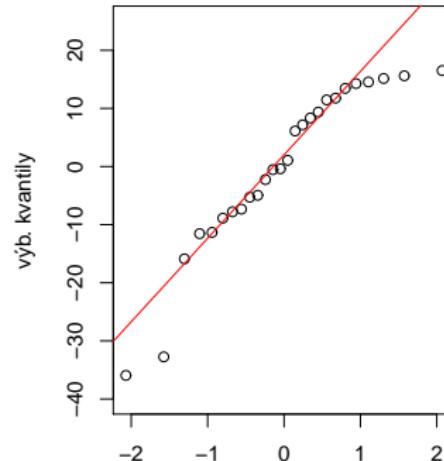
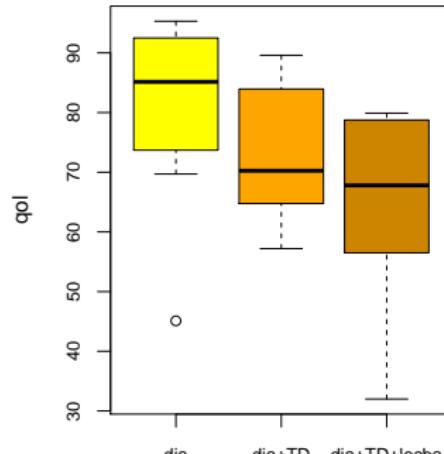
V tabulce jsou míry polohy skóřů ve 3 skupinách dětí:

skupina	min	Q1	med	prům	Q3	max
dia	45.10	75.40	85.15	81.03	91.98	95.30
dia+TD	57.20	65.03	70.25	73.09	81.58	89.60
dia+TD+léčba	32.00	58.15	67.80	64.75	78.48	79.90

Kruskalův-Wallisův test

Příklad (Kvalita života dětí s diabetem)

Podle boxplotů a normálního diagramu reziduí se zdá, že rozdělení výběrů nejsou symetrická, jsou výrazně sešikmená vlevo.



Kruskalův-Wallisův test

Příklad (Kvalita života dětí s diabetem)

Vzhledem k porušení normality a nízkému počtu pozorování jsme se rozhodli provést Kruskalův-Wallisův test.

Testová statistika vyšla $Q = 6.19$, p-hodnota 0.0453 . Na hladině 0.05 tedy zamítáme H_0 , že výběry pochází ze stejného rozdělení. Podle boxplotů, mediánů a průměrů se zdá, že skupina dětí, které měly pouze diabetes má nejvyšší skóry a skupina léčená na TD nejnižší. Test ovšem nedává odpověď na to, které dvojice se opravdu se významně liší.

Přednáška 9 (14.4.2020) - obsah

- Analýza rozptylu - dvojné třídění (model s bez interakcí, s interakcemi, tabulka analýzy rozptylu, předpoklady)
- Analýza rozptylu - dvojné třídění s náhodnými efekty
- Friedmanův test
- Korelační test (Pearsonův, Spearmanův)

Situace pro dvojně třídění

Příklad (Obsah tuku a produkce mléka)

Obsah tuku v mléce krávy patrně souvisí i s jinými faktory než je plemeno krávy. Máme podezření, že u krav, které jsou po prvním teleti (prvotelky) se obsah tuku liší od krav, které se již otelily vícekrát. Chceme sestavit model, ve kterém by bylo možné testovat závislost obsahu tuku na obou faktorech (plemeno i první tele).

V datech máme vždy 10 pozorování od každého plemene, z toho je vždy 5 od prvotekl a 5 od krav, které se telily vícekrát. Máme tedy zcela vyvážená data, v každé kombinaci úrovní faktorů máme stejně pozorování.

Kromě obsahu tuku máme od krav také jejich průměrnou denní produkci mléka. Podobně jako u obsahu tuku nás zajímá, zda i produkce mléka souvisí s plemenem a s tím, zda je kráva prvotelka.

Model dvojnáho třídění bez interakcí

- Máme tedy $k \times r$ výběrů spojité náhodné veličiny Y , každý z nich je vybrán pro jednu kombinaci úrovní faktorů A a B (A má k úrovní, B má r úrovní).
- Zajímá nás, zda se stř. hodnota Y liší pro různé úrovně A a B .
- Pro i -tou úroveň faktoru A a j -tou úroveň faktoru B má výběr rozsah n_{ij} a veličiny v tomto výběru označíme Y_{ijt} pro $t = 1, \dots, n_{ij}$.
- Pokud faktory A a B mají na veličinu Y aditivní vliv, můžeme model zapsat:

$$Y_{ijt} = \mu + \alpha_i + \beta_j + \varepsilon_{ijt},$$

kde μ je celková střední hodnota, α_i efekty úrovní faktoru A , β_j efekty úrovní faktoru B a ε_{ijt} jsou nezávislé náhodné veličiny s rozdělením $N(0, \sigma^2)$.

- Tento model nazveme modelem *dvojnáho třídění bez interakcí*.

Model dvojn ého třídění bez interakcí

- Model má takto píliš mnoho parametrů, zavedeme proto ještě podmínky:

$$\sum_{i=1}^k \alpha_i = 0 \quad a \quad \sum_{j=1}^r \beta_j = 0.$$

- Model bez interakcí odpovídá situaci, kdy pro každé i platí, že efekt i -té úrovně A je vždy stejný, bez ohledu na úroveň B a podobně efekt j -té úrovně B je stejný pro všechny úrovně A .
- V modelu chceme testovat hypotézy:

$$H_A: \alpha_1 = \alpha_2 = \dots = \alpha_k = 0, \text{ (faktor } A \text{ nemá vliv na stř. hod. } Y\text{)}$$

$$H_B: \beta_1 = \beta_2 = \dots = \beta_r = 0, \text{ (faktor } B \text{ nemá vliv na stř. hod. } Y\text{)}$$
- Obdobně jako u jednoduchého třídění mžeme sestavit součty čtverců S_T , S_A , S_B a S_e , které kvantifikují celkovou variabilitu, variabilitu vysvětlenou faktorem A , variabilitu vysvětlenou faktorem B a reziduální variabilitu.

Dvojné třídění bez interakcí

- Rozklad součtů čtverců $S_T = S_A + S_B + S_e$ obecně neplatí. Pokud máme ale pro všechny kombinace úrovní faktorů stejný počet pozorování (n_{ij} stejná pro všechna i a j), pak tento vztah platit bude.
- Pro testování použijeme opět F -testy založené na podílech součtů čtverců. Používáme tabulku:

variabilita	součet čtverců S	počet st. volnosti f	prům. čtverec S/f	testová stat. F	p-hod
faktor A	S_A	$f_A = k - 1$	S_A/f_A	F_A	p
faktor B	S_B	$f_B = r - 1$	S_B/f_B	F_B	p
reziduální	S_e	$f_e = n - k - r + 1$	S_e/f_e		
celková	S_T	$n - 1$			

Dvojné třídění s interakcemi

- Často nemůžeme předpokládat, že vlivy faktorů jsou aditivní. Faktor A může mít různý vliv na Y při různých úrovních faktoru B .
- Pak musíme do modelu přidat další parametry γ_{ij} , které vyjádří vliv kombinace faktorů na Y . Model *dvojněho třídění s interakcemi* pak je:

$$Y_{ijt} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijt}$$

- Přidáme další reparametizační podmínky: $\sum_{i=1}^k \gamma_{ij} = 0$ a $\sum_{j=1}^r \gamma_{ij} = 0$.
- V takovém modelu nejprve testujeme hypotézu $H_{AB}: \gamma_{11} = \gamma_{12} = \dots = \gamma_{kr} = 0$, která znamená, že interakce v modelu nejsou potřeba a vliv faktorů A a B je aditivní.

Dvojné třídění s interakcemi

Obdobně jako u předchozích modelů lze spočítat součty čtverců a sestavit tabulku analýzy rozptylu:

variabilita	součet čtverců S	počet st. volnosti f	prům. čtverec S/f	test stat. F	p-hod
faktor A	S_A	$f_A = k - 1$	S_A/f_A	F_A	p
faktor B	S_B	$f_B = r - 1$	S_B/f_B	F_B	p
interakce	S_{AB}	$f_{AB} = (k - 1)(r - 1)$	S_{AB}/f_{AB}	F_{AB}	p
reziduální	S_e	$f_e = n - kr$	S_e/f_e		
celková	S_T	$n - 1$			

- Používání dvojněho třídění vyžaduje jistou obezřetnost, zvláště v případě, kdy data nejsou vyvážená (n_{ij} nejsou stejná pro všechny kombinace i a j). Existuje více možností, jak počítat součty čtverců a jejich interpretace může být různá.

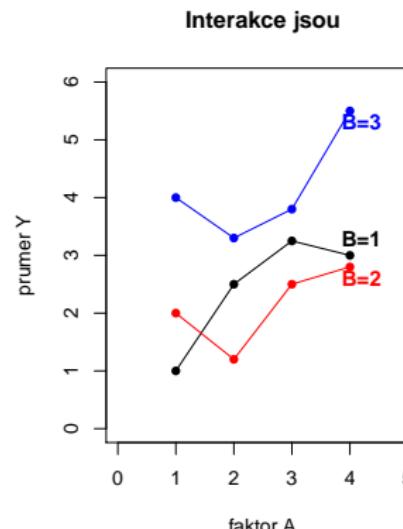
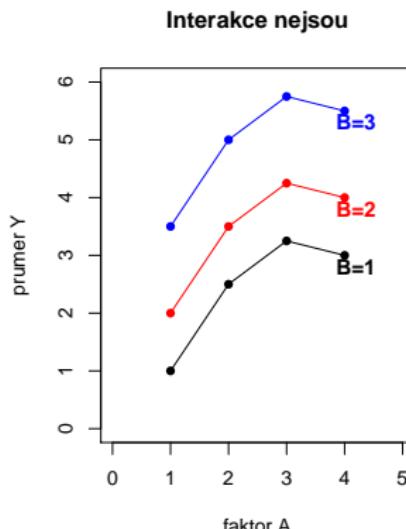
Jak postupovat v situaci dvojn ho t id n ?

- 1 Pokud nem me n jak  z asadn  d vod pro aditivitu vlivu obou faktor , m li bychom nejprve vyzkou et model s interakcemi. V modelu otestujeme, zda jsou interakce nulov  (hypot za H_{AB})
- 2 Pokud H_{AB} nezm tneme, data jsou v souladu s aditivitou a m žeme p ej t k modelu bez interakc  a testovat H_A a H_B .
- 3 Pokud H_{AB} zam tneme, model zjednodu it nem žeme. Pokud se to hod , m žeme p ej t k modelu jednoduch ho t id n  s kombinovan m faktorem ($k \times r$  rov ) a testovat mnohon sobn m porovn v n m, kter  kombinace faktor  A a B se od sebe li .

Grafy průměrů

Vhodný nástroj ke zkoumání situace dvojnho třídění je diagram průměrů Y podle kombinací úrovní obou faktorů A a B .

Obrázek: Diagramy průměrů podle úrovní faktorů



Předpoklady dvojn ého třídění

Předpoklady dvojn ého třídění jsou obdobné jako u jednoduchého třídění

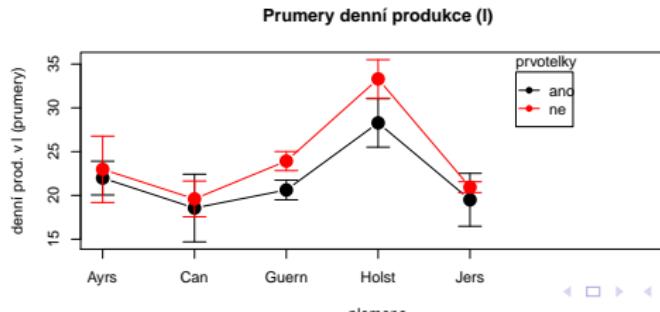
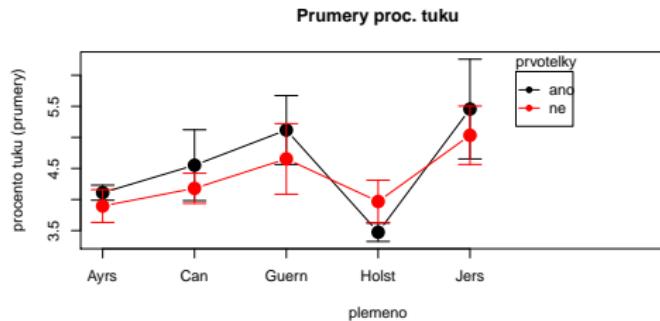
- ① Nezávislost všech veličin $Y_{111}, Y_{112}, \dots, Y_{krn_{kr}}$.
- ② Normalita všech veličin $Y_{111}, Y_{112}, \dots, Y_{krn_{kr}}$.
- ③ Shodné rozptyly všech veličin.

Podobně také m ůžeme ov ěrit normalitu a shodné rozptyly:

- Náhodné veličiny ε_{ijt} musí být normální $N(0, \sigma^2)$, což lze ov ěrit na odhadech těchto veličin - reziduích $\hat{\varepsilon}_{ijt} = Y_{ijt} - \bar{Y}_{ij}$. (\bar{Y}_{ij} je průměr všech pozorování s i -tou úrovní A a j -tou úrovní B .)
- Normalitu lze ov ěrit normálním diagramem reziduí.
- Shodné rozptyly m ůžeme ov ěrit pomocí boxplotů pro všechny kombinace faktorů nebo vynesením absolutních hodnot reziduí proti průměrům ve všech kombinacích faktorů.
- Je možné také testovat normalitu reziduí (Shapir ův-Wilk ův test) a shodné rozptyly pro všechny kombinace úrovní faktorů A a B (např. Leven ův test).

Tuk v mléce a denní produkce mléka

Obrázek: Diagramy průměrů a 95% CI ve skupinách



Denní produkce mléka

Příklad (Denní produkce mléka)

Zkoumali jsme vliv plemene krávy a toho, zda je kráva prvotelka, na denní produkci mléka. Z grafu průměrů se zdá, že prvotelky měly spíše nižší denní produkci mléka, plemena se také lišila, interakce spíše přítomny nebudou. Když spočítáme model s interakcemi, skutečně dostáváme p-hodnotu na přítomnost interakcí 0.1228.

Přejdeme tedy k modelu bez interakcí. Oba faktory jsou v modelu významné (plemeno $p=1.3 \cdot 10^{-15}$, prvotelka $p=0.0002$).

Prokazujeme tedy (na hladině 0.05), že prvotelky mají nižší stř. hodnotu denní produkce mléka (odhad efektu prvotelky -2.36 l) a že plemena mají různé stř. hodnoty denní produkce mléka.

Tuk v mléce

Příklad (Tuk v mléce)

Zkoumali jsme také vliv plemene krávy a toho, zda je kráva prvotelka na procento tuku v mléce. V tomto případě se zdá, že by zde interakce být mohla, u Holštýnských krav mají prvotelky nižší obsah tuku, zatímco u ostatních plemen vždy vyšší. Interakce mají p-hodnotu 0.0342. To znamená, že jsme prokázali přítomnost interakce a model zjednodušit nemůžeme.

Je možné přejít k jednoduchému třídění s kombinovaným faktorem a pak provést mnohonásobná porovnání a zjistit, které kombinace úrovní faktorů se od sebe liší. Kombinace má ale 10 úrovní, takže porovnání je 45 a výsledek je dosti nepřehledný. V souladu s grafem průměrů se zdá, že Holštýnský skot se vymyká. Zatímco Holštýnské krávy, které se telily vícekrát se od většiny ostatních plemen neliší, Holštýnské prvotelky mají signifikantně nižší obsah tuku v mléce než všechny ostatní (s výjimkou Ayrshirských).

Skupinové závislosti v jednoduchém třídění

- V situaci jednoduchého třídění, kdy zkoumáme vliv jednoho faktoru na náhodnou veličinu Y , můžeme mít v datech skupinové závislosti.
- *Př.*: Měříme zda nějaké laboratorní hodnoty souvisí s typem ošetření. Pokus provádíme na laboratorních krysách, ale vždy máme několik krys ze stejného vrchu, které rozdělíme (náhodně) na různá ošetření. V takovém případě nemůžeme tato pozorování považovat za nezávislá, očekáváme, že se příbuzná zvířata budou chovat podobně.
- *Př.*: Zkoumáme, zda se liší kvalita života diabetických dětí podle toho zda trpí dalšími onemocněními. Ve výběru máme děti, které se léčí v různých diabetologických ordinacích a očekáváme, že údaje od dětí ze stejné ordinace mohou být z různých důvodů závislé.
- Nezajímá nás vliv konkrétních skupinek (vrhu zvířat, ordinací apod) na Y , pouze chceme vzít v úvahu, že tam vliv může být.
- Považujeme tedy efekty skupinek - **náhodných bloků** za realizaci náhodné veličiny.
- Úrovně faktoru A , jejichž vliv na Y zkoumáme, nazveme *ošetřením*.

Model pro dvojné třídění s náhodnými efekty

- Zkoumáme vliv ošetření na Y , ale máme v modelu náhodné bloky.
Model pro jednoduché třídění bude vypadat podobně jako model pro dvojné třídění bez interakcí:

$$Y_{ij} = \mu + \alpha_i + B_j + \varepsilon_{ij},$$

kde ε_{ij} jsou nezávislé náhodné veličiny s rozdelením $N(0, \sigma^2)$, α_i jsou efekty úrovní faktoru A a B_j jsou efekty závislých bloků. B_j nyní nepovažujeme za pevné, ale za náhodné (proto označujeme velkým písmenem).

- Podobně jako u jednoduchého třídění zavedeme podmínu $\sum_{i=1}^k \alpha_i = 0$ a budeme předpokládat, že B_1, \dots, B_r je výběr z $N(0, \sigma_B^2)$. Nulová stř. hodnota zde odpovídá druhé reparametrizační podmínce dvojného třídění.
- Testujeme $H_A: \alpha_1 = \alpha_2 = \dots = \alpha_k = 0$ (faktor A nemá vliv na Y).
- Test na vliv bloků odpovídá testu $H_B: \sigma_B^2 = 0$ (bloky se neliší, nemají vliv na Y).

Testy v modelu s náhodnými efekty

- Testy H_A a H_B provádíme pomocí F statistik stejně jako v situaci dvojněho třídění s pevnými efekty.
- Tak jako u dvojněho třídění s pevnými efekty je výhodné, když jsou data vyvážená, tj. když jsou bloky stejně velké a stejně rozložené mezi ošetření. Pokud tedy provádíme experiment a můžeme počty ovlivnit, je výhodné takto postupovat.
- Rozdíl mezi modelem s náhodnými efekty a obyčejným dvojněm tříděním je hlavně v interpretaci. V případě závislých bloků nás nezajímá, jak konkrétní bloky ovlivňují Y , zajímá nás maximálně, zda ji ovlivňují nebo nikoliv.
- Zatímco u dvojněho třídění pevně volíme úrovně obou faktorů, u modelu s náhodnými bloky úrovně bloků neovlivňujeme, pokud experiment/šetření opakujeme, mohou se nastavit zcela jinak.

Model s náhodnými efekty

- Pokud ve dvojném třídění máme pro každou kombinaci úrovní faktorů právě jedno pozorování, mluvíme o dvojném třídění bez opakování. V případě, že efekty jednoho z faktorů považujeme za náhodné, je náš model zobecněním párového t -testu na více než 2 výběry.
- Předpoklady dvojněho třídění s náhodnými efekty jsou obdobné jako u dvojněho třídění s pevnými efekty. Přibývá pouze předpoklad na normální rozdělení náhodných efektů.

Model s náhodnými efekty

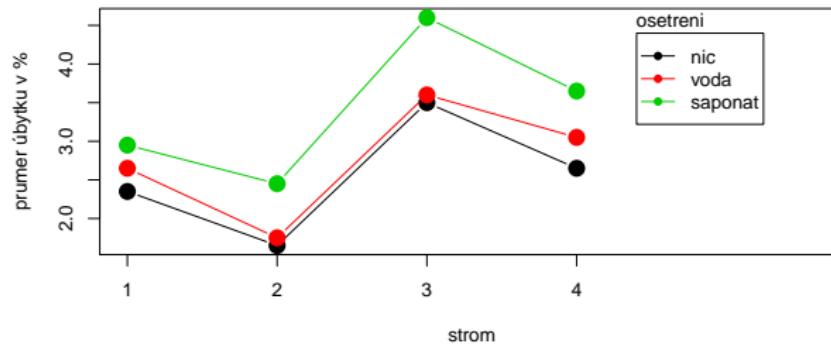
Příklad (Jablka)

Žáci ve škole prováděli pokus s jablkami. Účelem bylo ukázat, že na slupce jablek je ochranná vrstva, která brání vysychání. Na začátku pokusu třetinu jablek nechali bez ošetření, třetinu pečlivě umyli čistou vodou a třetinu pečlivě omyli saponátem. Potom jablka zvážili a nechali dva týdny ležet za oknem. Po dvou týdnech je znovu zvážili a zaznamenali, kolik procent váhy jablka ztratila.

K pokusu použili 24 jablek ze 4 stromů (z každého stromu vždy 6 jablek). Děti nevěděly nic o odrůdách jablek, ale uvědomovaly si, že se jablka z různých stromů mohou svými vlastnostmi lišit, proto z každého stromu vždy po 2 jablkách dali na každé ošetření. Získaly následující data (průměry úbytků ze 2 jablek v %).

	strom 1	strom 2	strom 3	strom 4
nic	2.35	1.65	3.50	2.65
voda	2.65	1.75	3.60	3.05
saponát	2.95	2.45	4.60	3.65

Model s náhodnými efekty



Příklad (Jablka)

Kdybychom nevzali v úvahu různé stromy a provedli jednoduché třídění na tři různá ošetření jablek, zjistili bychom, že rozdíl mezi ošetřeními není průkazný (p -hodnota 0.1010). Když ale použijeme model s náhodnými efekty pro stromy, zjištujeme, že rozdíl mezi ošetřeními je významný ($p = 0.000025$). Prokážeme také vliv stromů ($p = 7.2 \cdot 10^{-9}$).



Neparametrický test v situaci náhodných bloků

- Chceme zkoumat, zda se rozdělení veličiny Y liší podle toho, které z několika ošetření bylo použito
- Předpokládáme situaci náhodných bloků (k ošetření a r bloků, každý blok o k členech náhodně rozdělen mezi ošetření) a víme, že rozdělení Y není normální. Pak můžeme použít pro porovnání ošetření neparametrický *Friedmanův test*.
- Friedmanův test je zobecněním znaménkového testu.
- Princip testu spočívá v tom, že určíme pořadí pozorování v každém bloku (každé odpovídá jinému ošetření). Pokud se ošetření neliší, budou to náhodné permutace o k prvcích. Pokud se ošetření liší, budou pořadí vykazovat podobný vzorec ve všech blocích.
- Testová statistika používá součty pořadí jednotlivých ošetření ve všech blocích.

Friedmanův test

- Chceme testovat hypotézu H_0 : Rozdělení Y nezávisí na ošetření.
- Označme R_{ij} pořadí i -tého ošetření v j -tém bloku ($i=1,\dots,k$, $j=1,\dots,r$).
- Testová statistika Friedmanova testu je

$$Q = \frac{12}{k(k+1)r} \sum_{i=1}^k \left(\sum_{j=1}^r R_{ij} \right)^2 - 3r(k+1)$$

- Za H_0 platí $Q \sim \chi_{k-1}^2$. Proti H_0 svědčí vysoké hodnoty Q , budeme tedy zamítat H_0 , pokud $Q \geq \chi_{k-1}^2(1-\alpha)$.
- Friedmanův test nepoužívá skutečné hodnoty Y , pouze jejich pořadí uvnitř bloků, je tedy možné jej použít i v případě, kdy veličina Y má ordinální měřítko.

Friedmanův test

Příklad (Degustace vína)

6 degustátorů hodnotí víno stejné odrůdy od 4 různých vinařů. Hodnotí tak, že seřadí 4 vína podle chuti od nejlepšího k nejhoršímu. Chceme testovat, zda je rozdíl v chuti mezi víny od testovaných vinařů.

Ošetření budou vinaři ($k = 4$), náhodné bloky budou degustátoři ($r = 6$). Výsledkem pokusu nejsou číselné hodnoty, ale pouze pořadí ošetření uvnitř bloků - ideální situace pro Friedmanův test:

vinaři/degustátoři	1	2	3	4	5	6	\sum
A	4	3	1	4	3	1	16
B	1	2	3	1	1	2	10
C	2	1	2	2	2	3	12
D	3	4	4	3	4	4	22

$$Q = \frac{12}{4 \cdot 5 \cdot 6} (16^2 + 10^2 + 12^2 + 22^2) - 3 \cdot 6 \cdot 5 = 8.4, \quad p = 0.0384.$$

Prokázali jsme tedy na hladině 0.05, že mezi vinaři je rozdíl v chuti vína.

Přehled testů na rozdíly v poloze výběrů

Tabulka: Testy o poloze výběrů

situace	normální výběry	nenormální výběry
1 výběr	jednovýběrový <i>t</i> -test	znaménkový nebo Wilcoxonův párový
2 výběry nezávislé	dvouvýběrový <i>t</i> -test	Wilcoxonův dvouvýběrový (Kolmogorovův-Smirnovův)
2 výběry párové závislosti	párový <i>t</i> -test	znaménkový nebo Wilcoxonův párový
>2 výběry nezávislé	ANOVA jedn. třídění	Kruskalův-Wallisův
>2 výběry náh. bloky	ANOVA s náh. efekty	Friedmanův

Testování závislosti

- Způsob testování závislosti dvou veličin závisí na měřítku veličin.
- Závislost kvantitativní a kvalitativní veličiny již testovat umíme. Jedná se o porovnání rozdělení několika výběrů podle úrovní kvalitativní veličiny (faktoru), což se dá udělat pomocí jednoduchého třídění (v případě 0-1 faktoru dvouvýběrovým t -testem).
- Nyní se budeme zabývat testováním a popisem závislosti dvou kvantitativních veličin.

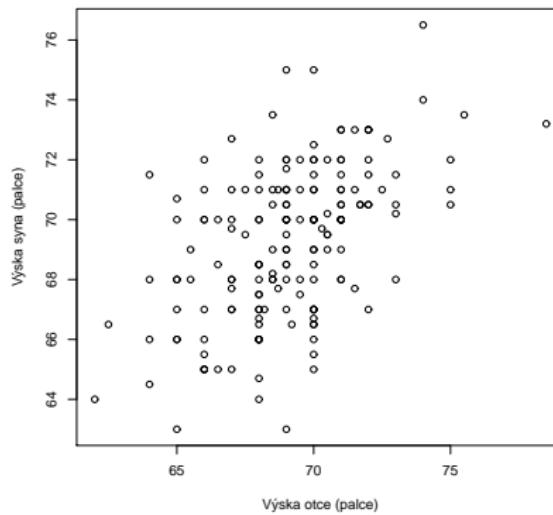
Korelační koeficient

- Chceme zkoumat závislost svou spojitého veličin X a Y . Máme k dispozici měření těchto veličin na výběru n nezávislých subjektů, tj. máme dvojice $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$.
- Již známe *výběrový korelační koeficient (Pearsonův)*:

$$r_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \cdot \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2}}$$
- Víme, že $-1 \leq r_{XY} \leq 1$. Proti nezávislosti svědčí vyšší hodnoty $|r_{XY}|$. Hodnoty $r_{XY} > 0$, svědčí o pozitivní závislosti (veličiny jdou spolu), $r_{XY} < 0$, svědčí o negativní závislosti (veličiny jdou proti sobě).
- Vyšší hodnota $|r_{XY}|$ nemusí znamenat, že jsou veličiny opravdu závislé. Co když je to náhoda? Potřebujeme standardizovaný test pro testování závislosti.

Závislost dvou veličin

Obrázek: Závislost výšky syna a otce - Data Galton



Korelační koeficient $r_{XY} = 0.505$. Co to znamená?

Korelační test

- Označme ρ_{XY} *populační korelační koeficient* (teoretická korelace). Pokud jsou X a Y nezávislé, pak $\rho_{XY} = 0$.
- Chceme testovat $H_0: \rho_{XY} = 0$ proti alternativě $H_1: \rho_{XY} \neq 0$ (případně proti jednostranným alternativám).
- Za předpokladu, že (X, Y) má dvourozměrné normální rozdělení, platí za H_0 :

$$T = \frac{r_{XY}}{\sqrt{1-r_{XY}^2}} \sqrt{n-2} \sim t_{n-2}$$

- Můžeme tedy založit *korelační test* na statistice T a zamítat H_0 , pokud $|T| \geq t_{n-2}(1 - \frac{\alpha}{2})$.

Korelační test

Příklad (Otcové a synové - Galton)

V případě otců a synů vychází $T=7.6506$, p-hodnota $1.39 \cdot 10^{-12}$, tedy extrémně nízká. Samozřejmě zamítáme nezávislost. V datovém souboru je 173 dvojic otec/syn, zjevně dost na jasné prokázání souvislosti.

Spočítáme-li 95% interval spolehlivosti pro ρ_{XY} , vychází (0.38, 0.61). Populační korelační koeficient bude pokryt tímto intervalem s pravděpodobností 0.95.

Interpretace korelačního testu

- Velikost korelačního koeficientu souvisí s tím, jak moc pozorování leží na přímce. Pokud by ležela přesně na přímce, pak $r_{XY} = 1$ nebo $r_{XY} = -1$. Čím více budou pozorování rozrážena kolem přímky, tím nižší bude $|r_{XY}|$. Korelační koeficient tedy měří sílu závislosti ve smyslu linearity.
- P -hodnota korelačního testu měří průkaznost závislosti veličin. Když dostaneme velmi nízkou p -hodnotu (jako v datech Galton), nemusí to znamenat vysokou hodnotu $|r_{XY}|$. Znamená to jen, že máme velmi silné důkazy, že souvislost mezi veličinami skutečně existuje. Máme-li dost pozorování, podaří se nám prokázat i relativně slabší souvislosti.

Předpoklady korelačního testu

- Testová statistika T bude mít za H_0 předpokládané rozdělení t_{n-2} , pokud sdružené rozdělení (X, Y) bude dvourozměrné normální.
- K tomu, aby sdružené rozdělení bylo normální, nestačí normalita X a Y , musely by být normální i jejich lineární kombinace.
- Ověření sdružené normality je nesnadné, naštěstí korelační test není příliš citlivý na odchylky od normality, zvláště při dostatečném počtu pozorování.
- Problematická by mohla být výrazně odlehlá pozorování nebo vykousnuté části scatterplotu (vodorovný řez dvourozměrnou normální hustotou by měla být kružnice nebo elipsa).
- Korelační test nebude fungovat moc dobře, pokud závislost není lineární.
- V případě jasné nenormality nebo monotónní ale nelineární závislosti bude lépe než Pearsonův korelační koeficient fungovat pořadový (Spearmanův) korelační koeficient.

Spearmanův korelační test

- *Spearmanův korelační test* používá Spearmanův pořadový korelační koeficient r_{XY}^S .
- Tento koeficient dostaneme, pokud místo výběrů X_1, X_2, \dots, X_n a Y_1, Y_2, \dots, Y_n spočítáme běžný korelační koeficient (Pearsonův) z jejich pořadí v každém výběru (R_1, R_2, \dots, R_n a Q_1, Q_2, \dots, Q_n):

$$r_{XY}^S = r_{RQ}$$

- Pokud v datech nejsou shodná pozorování, lze ukázat:

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

- P -hodnotu Spearanova testu nezávislosti můžeme najít pomocí normální approximace a zamítat pro

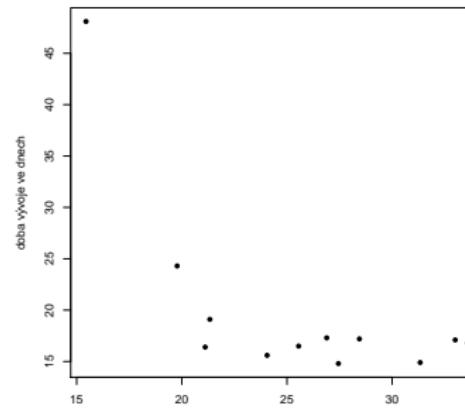
$$|r_{XY}^S| \sqrt{n-1} \geq z(1 - \frac{\alpha}{2})$$

Korelační test

Příklad (Závislost vývoje kříска bramborového na teplotě)

13 vajíček kříска (drobný hmyz škodící rostlinám) bylo ponecháno za různých teplot a bylo zaznamenáno, jak dlouho trval vývoj hmyzu od vajíčka až po dospělce (ve dnech).

Obrázek: Závislost doby vývoje kříска na teplotě



Korelační test

Příklad (Závislost vývoje kříска bramborového na teplotě)

Pearsonův kor. koeficient je $r_{XY} = -0.67$, p-hodnota kor. testu 0.0170. Zamítli bychom nezávislost na pětiprocentní hladině a usoudili bychom na silnou negativní závislost, čím vyšší teplota, tím kratší doba vývoje. Ze scatterplotu je ovšem zřejmé, že výsledek bude nejspíš ovlivněn jedním výrazně odlehlym pozorováním. Nebýt tohoto pozorování, závislost není tak zřejmá.

Spearmanův kor. koeficient je: $r_{XY}^S = -0.45$, p-hodnota 0.1472. Na základě Spearmanova testu bychom tedy nezamítli nezávislost.

Pokud bychom vynechali odlehlé pozorování, ani Pearsonův kor. test nezamítne nezávislost, $r_{XY-1} = -0.52$ a p-hodnota 0.0992.

Podle obrázku se zdá, že případná závislost nejspíš nebude lineární, proto Pearsonův kor. koeficient příliš vhodný není. Možná se doba vývoje mění pouze pro teploty nižší než 25 stupňů C, pak už zůstává stejná. Rozhodně bychom potřebovali více pozorování při nižších teplotách.

Přednáška 10 (21.4.2020) - obsah

- Jednoduchá lineární regrese (model, metoda nejmenších čtverců, testování nezávislosti v regresi, koeficient determinace, předpoklady, transformace)
- Mnohonásobná regrese (model, testování přínosu regresorů, analýza rozptylu, předpoklady)

Regres - motivace

- Nechceme jen prokázat, že závislost existuje, chceme ji popsat. Když se X zvýší o 1, jak se změní Y ?
- Jedná-li se skutečně o lineární závislost, můžeme se pokusit najít přímku, která ji popisuje.
- U korelace nezáleží na pořadí korelovaných veličin, metoda se k nim chová stejně, u regrese tomu tak nebude.
- V regresi zvolíme jednu veličinu , které budeme říkat *vysvětlovaná proměnná* (závisle proměnná, odezva, angl. response) a tu se pokusíme vysvětlit pomocí *vysvětlující proměnné* (nezávisle proměnná, regresor, prediktor)
- Procedura se nebude chovat k oběma veličinám symetricky.

Regres - motivace

Příklad (Otcové s synové - Galton)

Chceme predikovat výšku syna, pokud známe výšku otce. Zvolíme tedy jako vysvětlovanou proměnnou Y výšku syna a jako vysvětlující proměnnou X výšku otce.

Chtěli bychom daty proložit přímku. Víme, že přesně to možné nebude, data v přímce neleží. Je jasné, že se nám nikdy nepodaří přesně předpovědět výšku syna na základě výšky otce. Výška syna bude záviset na dalších okolnostech, o kterých nic nevíme (výška matky, další genetické okolnosti, výživa, fyzická aktivita apod.). Představme si, že všechny tyto vlivy, které neznáme, způsobí určitou odchylku výšky syna od lineární závislosti. Tyto odchylky můžeme považovat za realizace náhodné veličiny. Pokud stanovíme pro odchylky (chyby) rozumné přepoklady, podaří se nám přímku daty proložit.

Regresní model

- Každému x_i odpovídá nikoliv jedna hodnota y_i , ale náhodná veličina Y_i .
- Předpokládáme, že existují β_0 a β_1 tak, že $EY_i = \beta_0 + \beta_1 x_i$. Střední hodnota Y je tedy lineárně závislá na x . Hodnoty Y_i ovšem přesně na přímce ležet nebudou.
- Budeme předpokládat *regresní model*:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

kde $\varepsilon_i \sim N(0, \sigma^2)$, ε_i nezávislé pro všechna $i = 1, \dots, n$.

- Jediná náhodná část Y_i je náhodná chyba ε_i . Ta má nulovou střední hodnotu a stejný rozptyl pro všechna $i = 1, \dots, n$.
- *Regresní koeficienty* β_0 a β_1 neznáme, musíme je odhadnout.
- Regresní koeficient β_1 vyjadřuje změnu stř. hodnoty Y při jednotkové změně X .
- Regresní koeficient β_0 je střední hodnotou Y , pokud $X = 0$ (samozřejmě pouze v případě, že X může nabýt hodnoty 0).

Odhad regresních koeficientů

- Odhad regresních koeficientů se provádí *metodou nejmenších čtverců*, která hledá β_0 a β_1 tak, aby součet čtverců

$$\sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i)^2$$

byl minimalizován.

- Hodnoty b_0 a b_1 , které součet čtverců minimalizují, použijeme jako odhad regresních koeficientů β_0 a β_1 . Přímku $y = b_0 + b_1 x$ nazveme *regresní přímkou* a použijeme ji jako odhad závislosti Y na X .
- Minimální součet čtverců

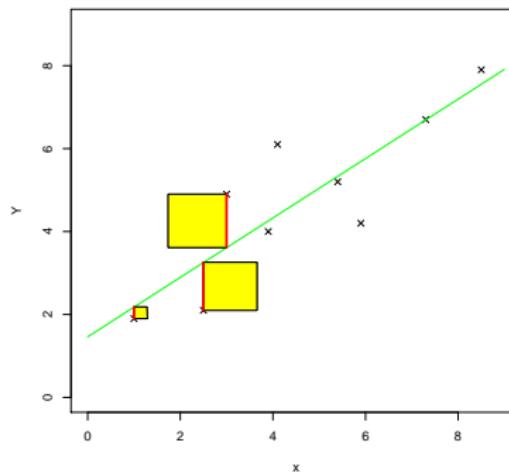
$$S_e = \sum_{i=1}^n (Y_i - b_0 - b_1 x_i)^2$$

nazveme *reziduálním součtem čtverců*.

- Všimněme si souvislosti s ANOVou. Tam jsme nazývali res. součtem čtverců variabilitu uvnitř výběrů, tedy podobně jako zde, variabilitu, kterou nedokážeme vysvětlit pomocí vysvětlující proměnné.

Metoda nejmenších čtverců

Obrázek: Metoda nejmenších čtverců



Metoda nejmenších čtverců hledá přímku, která minimalizuje součet čtverců svislých vzdáleností od pozorování.

Vyrovnанé hodnoty, rezidua

- Svislé vzdálenosti pozorování od přímky

$$\hat{\varepsilon}_i = Y_i - b_0 - b_1 x_i, \text{ pro } i = 1, \dots, n$$

se nazývají *rezidua*. Rezidua odhadují chyby ε_i a budeme je používat ke zkoumání vlastností ε_i .

- Body na regresní přímce $\hat{Y}_i = b_0 + b_1 x_i$ nazveme *odhadnutými* nebo *vyrovnanými* hodnotami.
- Rezidua jsou tedy $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$.
- Reziduální součet čtverců lze použít k odhadu neznámého rozptylu chyb σ^2 . Odhadneme jej pomocí $S^2 = \frac{S_e}{n-2}$.

Testování nezávislosti v regresním modelu

- Nezávislost mezi vysvětlovanou a vysvětlující proměnnou můžeme testovat přímo v regresním modelu. Nezávislost v regresním modelu je ekvivalentní faktu $\beta_1 = 0$. Můžeme tedy testovat $H_0: \beta_1 = 0$ proti alternativě $H_1: \beta_1 \neq 0$.
- Jako testovou statistiku použijeme b_1 (odhad β_1), znormovanou odhadem směrodatné odchylky:

$$T = \frac{b_1}{S.E.(b_1)} = \frac{b_1}{S} \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}$$

- Za H_0 má veličina T rozdělení t_{n-2} . H_0 zamítáme pro $|T| \geq t_{n-2}(1 - \frac{\alpha}{2})$.
- Test nezávislosti v jednoduché regresi (jeden regresor v modelu), je ekvivalentní korelačnímu testu (Pearsonovu).
- Podobně je možné testovat hypotézu $H_0: \beta_0 = 0$, což by znamenalo, že regresní přímka prochází počátkem. To ale obvykle není příliš zajímavé.

Koeficient determinace

- *Koeficient determinace* R^2 vyjadřuje, jaký podíl variability veličiny Y se dá vysvětlit pomocí regresního modelu (lineární závislosti na x).
- $R^2 = 1 - \frac{S_e}{S_T} = 1 - \frac{S_e}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$,
kde S_T je totální součet čtverců - celková variabilita ve výběru Y_1, Y_2, \dots, Y_n .
- R^2 je tedy podíl variability, kterou se podařilo vysvětlit regresním modelem, vždy platí $0 \leq R^2 \leq 1$.
- Je-li R^2 vysoké, znamená to, že x dobře vysvětuje Y , body na scatterplotu leží blízko přímky. Takový model je vhodný k predikci. Je-li R^2 nízké, mnoho variability v datech je nevysvětleno, model nebude predikovat dobře.
- V jednoduché regresi (jeden regresor v modelu) je $R^2 = r_{XY}^2$, tedy koeficient determinace se rovná druhé mocnině Pearsonova korelačního koeficientu.

Analýza rozptylu

- Podobně jako v analýze rozptylu lze totální součet čtverců rozložit:

$$S_T = \sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = S_R + S_e$$

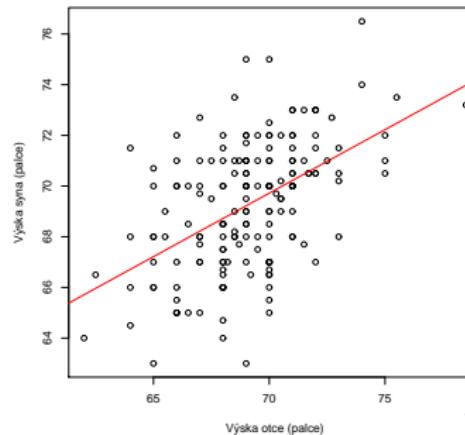
- Součet čtverců S_R odpovídá variabilitě vysvětlené regresním modelem, S_e je reziduální součet čtverců, který obsahuje variabilitu, která zůstává nevysvětlena.
- Každému ze součtů čtverců odpovídá příslušný počet stupňů volnosti ($f_T = n - 1$, $f_R = 1$, $f_e = n - 2$).
- Podobným přístupem jako v analýze rozptylu můžeme porovnat S_R a S_e a na základě toho usoudit, zda regrese vysvětluje dostatečné množství variability v porovnání se zbytkovou variabilitou, tj. zda Y vůbec souvisí s X . Testová statistika $F = \frac{S_R/f_r}{S_e/f_e}$ má rozdělení $F_{1,n-2}$, pokud platí H_0 : nezávislost Y a X .
- U jednoduché regrese (jeden regresor) je test ekvivalentní popsanému t -testu testujícímu $H_0: \beta_1 = 0$.

Regresce

Příklad (Výšky otců a synů - Galton)

*Pomocí metody nejmenších čtverců jsme odhadli regresní přímku:
 $y = 34.6523 + 0.5010x$. Zakreslena ve scatterplotu:*

Obrázek: Regrese výška syna/ výška otce -Galton



Regresy

Příklad (Výšky otců a synů - Galton -pokračování)

Test nezávislosti ve regresi $H_0: \beta_1 = 0$ vydal testovou statistiku $T = 7.651$, p -hodnota $1.39 \cdot 10^{-12}$. P -hodnota je stejná jako u korelačního testu i F testu na celý model ($F=58.3$, $df=1, 171$). Nezávislost samozřejmě zamítáme.

Parametr β_1 jsme odhadli jako 0.5010. To znamená, že je-li otec o 1 palec vyšší, synovu výšku odhadneme o 0.5 palců vyšší. Pomocí rozdělení statistiky T lze získat interval spolehlivosti pro β_1 : (0.37, 0.63). Skutečná směrnice závislosti tedy bude patrně pokryta tímto intervalom.

Reziduální součet čtverců je $S_e = 815.41$, odhad σ^2 : $S^2 = 4.77$.

Regresy

Příklad (Výšky otců a synů - Galton -pokračování)

Koefficient determinace je $R^2 = 0.255$, což odpovídá korelačnímu koeficientu $r_{XY} = 0.505$ ($0.505^2 = 0.255$). Výška otce tedy vysvětluje asi jen čtvrtinu variability výšky syna. Predikce patrně nebudou moc dobré, což jsme očekávali. Předpověď výšky syna jenom na základě výšky otce asi moc přesná nebude.

Přesto můžeme vyzkoušet, jak by predikce (odhad) fungovala. Pokud otec měří 72 palce (asi 182.9 cm), výšku syna by model předpověděl jako $34.6523 + 0.5010 \cdot 72 = 70.72$ palců (179.6 cm).

Předpoklady regresního modelu

Předpoklady regresního modelu jsme v podstatě vyslovili při jeho formulaci:

- ❶ Nezávislost veličin Y_1, Y_2, \dots, Y_n .
- ❷ Linearita závislosti EY a x .
- ❸ Normální rozdělení chyb ε_i .
- ❹ Shodné rozptyly chyb ε_i pro všechna $i = 1, \dots, n$.

Nezávislost

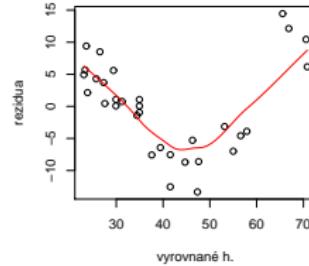
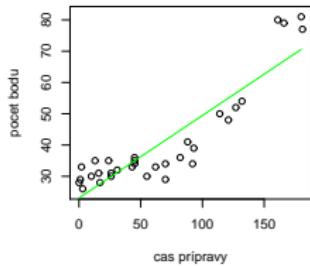
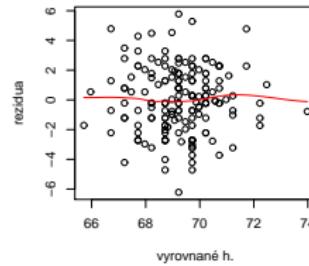
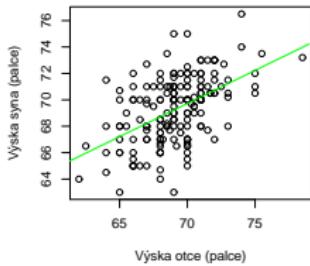
- Nezávislost veličin Y_1, Y_2, \dots, Y_n musíme přepokládat z logiky experimentu/šetření a způsobu sbírání dat.
- Pozor na časové závislosti (opakovaná měření v různém čase apod.) a skupinové závislosti (rodiny apod.).
- Pro závislá data nelze běžný regresní model použít. Existují složitější modely, které taková data zvládnou.
- Všimněte si, že nezávislost regresorů x_1, x_2, \dots, x_n nepožadujeme. Regresní model pracuje s regresory jako s pevnými hodnotami, nemusí to být náhodné veličiny, ale třeba pevně zvolené hodnoty při experimentu.

Předpoklad linearity závislosti

- Nelze prokládat rozumnou přímku závislostí, která není lineární.
- Linearitu závislosti lze ověřit pomocí reziduí $\hat{\varepsilon}_i$.
- Vhodné je vynést rezidua do grafu proti vyrovnaným hodnotám \hat{Y}_i . Při lineární závislosti budou náhodně rozptýlena. Pokud bude mít závislost jiný tvar, uvidíme určitý vzor.
- Pokud máme podezření, že závislost je jiná (kvadratická apod.), můžeme ztransformovat x .
- Rezidua proti vyrovnaným hodnotám jsme vynesli pro modely výška syn/otec (lineární závislost) a výsledek testu/čas přípravy (nelineární závislost).

Předpoklad linearity závislosti

Obrázek: Ověření linearity: syn/otec a výsledek/čas přípravy



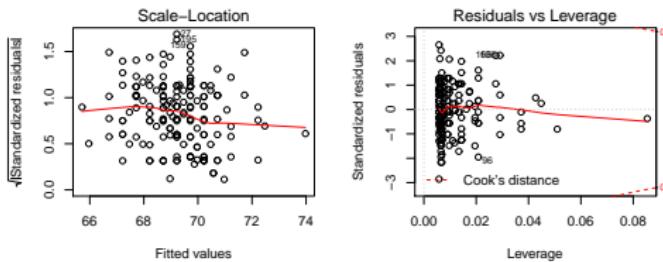
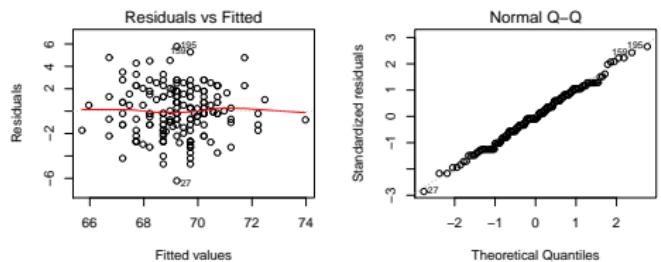
Předpoklad normality chyb a stability rozptylu

- Normalitu chyb lze ověřit normálním diagramem reziduí $\hat{\varepsilon}_i$, případně testem normality.
- Mírné odchylky od normality obvykle nebývají velkým problémem.
- Stabilita rozptylu (homoskedasticita) se neověřuje snadno, existují různé možnosti, jak může být tento předpoklad porušen.
- Častá je situace, kdy variabilita chyb roste s rostoucím \hat{Y} . Lze ověřit grafem $|\hat{\varepsilon}_i|$ vnesených proti \hat{Y}_i . Neměli bychom pozorovat rostoucí trend.
- Existují testy na stabilitu rozptylů (např. Breusch-Pagan).
- R vydá diagnostické grafy jediným příkazem.

Diagnostické grafy v R

Obrázek: Model: syn ~ otec (data Galton)

lm(syn ~ otec)



Co dělat, když jsou porušeny předpoklady modelu?

- Někdy pomůže transformace Y , může spravit normalitu chyb i zvyšující se rozptyly zároveň. Používá se hlavně logaritmus (lze použít pouze pro kladné veličiny). To znamená přechod k modelu

$$\ln Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Je třeba dát pozor při interpretaci vzhledem k původní vysvětlované veličině Y .

- Někdy je třeba použít složitější modely (zobecněné lineární modely apod.)

Transformace

Příklad (Úroveň polyaminů v plazmě u dětí)

Data obsahují úrovně polyaminů v plazmě u 25 dětí ve věku od 0 do 4 let. Zajímá nás souvislost věku s úrovní polyaminů v plazmatu

Sestavili jsme regresní model (Y je úroveň polyaminů, x věk), regresní přímka $y = 13.76 - 2.29x$, p-hodnota testu nezávislosti: $3.44 \cdot 10^{-8}$, koeficient determinace: $R^2 = 0.74$.

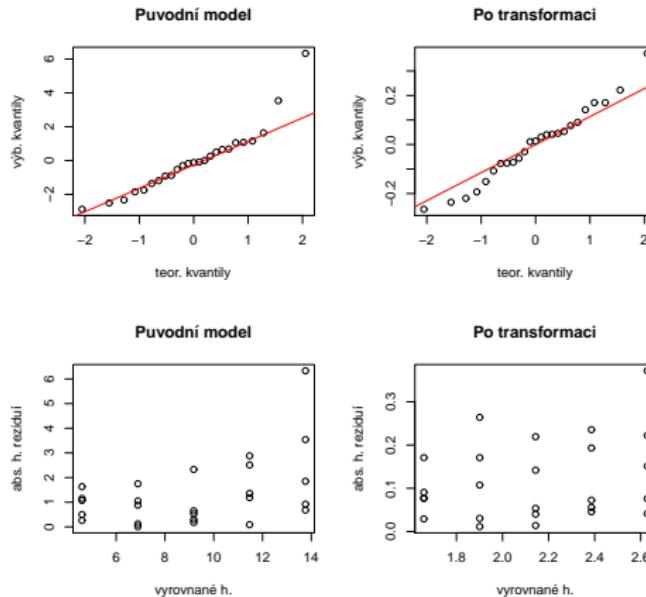
Usoudili jsme, že věk souvisí s úrovní polyaminů, zdá se, že dítě, které je o rok starší bude mít úroveň polyaminů v plazmě odhadem o 2.29 nižší. Věkem lze vysvětlit asi 74% variability polyaminů v plazmě.

Při ověřování předpokladů vyšlo najevo, že normalita chyb moc neplatí, chyby mají těžký chvost doprava, p-hodnota Shapirova-Wilkova testu 0.0148. Také rozptyly jsou podle grafu reziduí nestabilní.

Rozhodli jsme se zlogaritmovat úrovně polyaminů a zkusit, zda model funguje lépe.

Transformace

Obrázek: Normální diagramy reziduí a abs. hodnoty reziduí



Transformace

Příklad (Úroveň polyaminů v plazmě u dětí)

Regresní model se zlogaritmovanou úrovní polyaminů odhadl regresní přímku: $\ln(y) = 2.63 - 0.24x$, p -hodnota testu nezávislosti: $1.2 \cdot 10^{-10}$, koeficient determinace: $R^2 = 0.84$.

Normalita chyb se výrazně zlepšila, stále trochu těžší pravý chvost, ale p -hodnota Shapirova-Wilkova testu 0.8331, rozhodně přijatelnější. Také rozptyly reziduů nevykazují rostoucí trend s vyrovnanými hodnotami.

Model po transformaci lépe vyhovuje předpokladům regrese. Chceme-li jej použít k odhadování úrovně polyaminů, musíme zpět odlogaritmovat. Dítě ve věku x let bude mít odhadem úroveň polyaminů:

$$\hat{y} = e^{2.63 - 0.24x}$$

Mnohonásobná regrese

- Často máme k dispozici několik potenciálních vysvětlujících proměnných (regresorů) x_1, x_2, \dots, x_k a chceme najít model, který bude co nejlépe vysvětlovat veličinu Y .
- Model pak vypadá takto:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + \varepsilon_i,$$

kde $\varepsilon_i \sim N(0, \sigma^2)$, pro $i = 1, \dots, n$, jsou nezávislé náhodné veličiny.

- Regresní koeficienty β_j vyjadřují změnu střední hodnoty Y při jednotkové změně j -tého regresoru, pokud ostatní regresory zůstávají stejné.
- Regresní funkce v mnohorozměrném případě již není přímka. Při dvou regresorech se jedná o rovinu v prostoru, při více regresorech je to těžko představitelná vícerozměrná plocha.
- Odhad regresních koeficientů označíme b_0, b_1, \dots, b_k .

Testy v mnohonásobné regresi

- Odhadu regresních parametrů můžeme použít pro testování přínosu jednotlivých regresorů. Pro j -tý regresor ($j = 1, \dots, k$) testujeme $H_0: \beta_j = 0$, tj. hypotézu, že j -tý regresor, za přítomnosti ostatních regresorů již nic dalšího (o veličině Y) nevysvětluje.
- Testová statistika je

$$T_j = \frac{b_j}{S.E.(b_j)},$$

nulovou hypotézu zamítneme na hladině α , pokud
 $|T_j| \geq t_{n-k-1}(1 - \alpha/2)$.

- Testovat $H_0: \beta_0 = 0$ lze samozřejmě také, ale obvykle to není nijak zajímavé.

Odhadnuté hodnoty, rezidua, součty čtverců

- Podobně jako v jednoduché regresi, odhadnuté hodnoty označíme:

$$\hat{Y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \dots + b_k x_{ki}.$$

- Rezidua pak budou

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 x_{1i} + b_2 x_{2i} + \dots + b_k x_{ki}).$$

- Stejně bude fungovat rozklad totálního součtu čtverců na variabilitu vysvětlenou regresí (S_R) a variabilitu zůstávající nevysvětlenou (reziduální) (S_e):

$$S_T = \sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = S_R + S_e$$

Analýza rozptylu v regresi

- Podobně jako v analýze rozptylu lze použít součty čtverců k testování hypotézy, zda všechny regresory dohromady vůbec něco vysvětluje o Y , tj., zda má vůbec uvažovaný regresní model smysl.
- Nulová hypotéza je tedy $H_0: Y$ nesouvisí s žádným z regresorů.
- Testová statistika je

$$F = \frac{S_R/k}{S_e/(n - k - 1)},$$

- nulovou hypotézu zamítneme, pokud $F \geq F_{k,n-k-1}(1 - \alpha)$.
- Se součty čtverců souvisí také koeficient determinace R^2 , který vyjadřuje podíl celkové variability Y , který se podařilo vysvětlit pomocí závislostí Y na regresorech x_1, x_2, \dots, x_k .

$$R^2 = 1 - \frac{S_e}{S_T} = \frac{S_R}{S_T}$$

Analýza rozptylu v regresi

Můžeme sestavit tabulku analýzy rozptylu:

variabilita	součet čtverců S	počet st. volnosti f	prům. čtverec S/f	testová stat. F	p-hod
regrese	S_R	$f_R = k$	S_R/f_R	F_R	p
reziduální	S_e	$f_e = n - k - 1$	S_e/f_e		
celková	S_T	$n - 1$			

- V mnohonásobné regresi F -test vypovídá o celém modelu (všech regresorech). Liší se tedy od t -testů o přínosech jednotlivých regresorů. F -test může být signifikantní i když některé z regresorů vůbec s Y nesouvisí.
- Také R^2 se týká celého modelu, zkoumá jaký podíl variability veličiny Y vysvětlují všechny regresory v modelu dohromady.

Mnohonásobná regrese

Příklad (Výšky otce a syna - data Galton)

Zkusme přidat do modelu také výšku matky. Tabulka odhadů regresních parametrů:

	odhad	S.E.(b)	T	p-hod.
b_0	20.5199	5.8149	3.529	0.000537 ***
b_1 (otec)	0.4737	0.0636	7.445	$4.6 \cdot 10^{-12}$ ***
b_2 (matka)	0.2504	0.0680	3.683	0.000309 ***

Výška otce i výška matky jsou v modelu důležité, obě souvisí s výškou syna ($p = 4.65 \cdot 10^{-12}$ pro výšku otce a 0.0003 pro výšku matky).

Koefficient determinace $R^2 = 0.3101$. Zahrnutí výšky matky do modelu zlepšilo schopnost predikovat, výška obou rodičů vysvětluje asi 31% variability výšky syna.

F-test celého modelu vydal statistiku $F = 38.2$ ($df=2, 170$), $p = 2 \cdot 10^{-14}$, potvrzuje tedy, že výška syna souvisí s výškou matky a otce.

Předpoklady regresního modelu

Předpoklady mnohonásobné regrese jsou obdobné předpokladům jednoduché regrese.

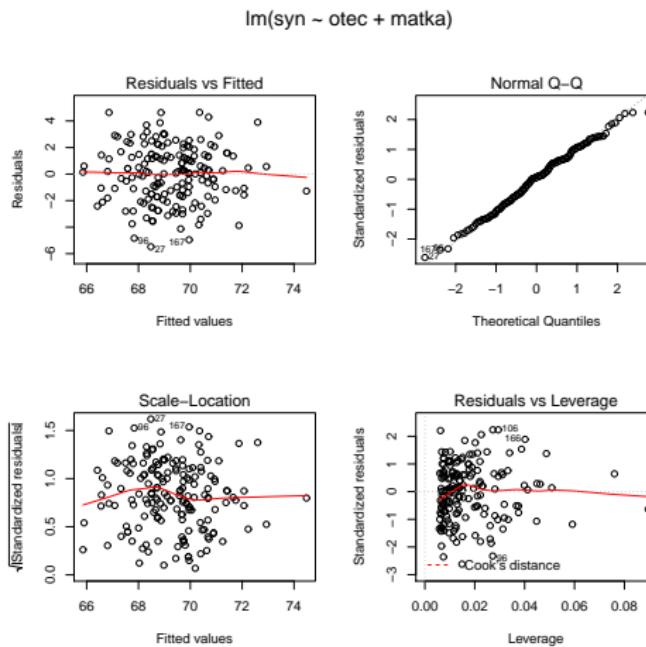
- ① Nezávislost všech veličin Y_1, \dots, Y_n
- ② Linearita závislosti Y na každém z regresorů za přítomnosti ostatních regresorů.
- ③ Normalita veličin Y_1, \dots, Y_n , což odpovídá normalitě chyb ε_i .
- ④ Shodné rozptyly všech veličin Y_1, \dots, Y_n , což, spolu s předchozím předpokladem odpovídá shodnému rozdělení chyb $\varepsilon_i \sim N(0, \sigma^2)$ pro nějaké σ^2 .

Ověřování předpokladů regresního modelu

- Předpoklady normality a shodných rozptylů se ověřují na reziduích podobně jako v situaci jednoduché regrese (normální diagram reziduí, graf abs. hodnot reziduí proti odhadnutým hodnotám).
- V R lze použít panel diagnostických grafů modelu.
- Nezávislost Y_1, \dots, Y_n musíme předpokládat podle logiky experimentu nebo sběru dat.
- Regresní diagnostika se může výrazně změnit, pokud zahrneme do modelu další důležitý regresor, nebo naopak nějaký regresor vynecháme.
- Problémy s normalitou a nestabilními rozptyly někdy vyřeší vhodná transformace Y (nejčastěji logaritmus).

Diagnostické grafy v R

Obrázek: Model: syn ~ otec + matka (data Galton)



Přednáška 11 (28.4.2020) - obsah

- Mnohonásobná regrese (kvadratická regrese, lineární modely výběr modelu)
- Multinomické rozdělení, test dobré shody

Kvadratická regrese

- Někdy je jasné, že vysvětlovaná veličina Y nejspíš závisí na vysvětlující veličině X , ale závislost nevypadá lineárně.
- Pokud máme podezření na jinou funkční závislost, můžeme sestavit model, který takovou závislost bude vyjadřovat.
- Zkusme sestavit model kvadratické regrese, tedy takový, kde střední hodnoty Y budou kvadratickou funkcí x .
- V modelu mnohonásobné regrese jako regresory použijeme x a x^2 . Model *kvadratické regrese* tedy vypadá:

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i,$$

kde $\varepsilon_i \sim N(0, \sigma^2)$, pro $i = 1, \dots, n$, jsou nezávislé náhodné veličiny.

- V takovém modelu můžeme testovat $H_0: \beta_2 = 0$, tj. nejedná se o kvadratickou závislost (kvadratický člen je 0).
- Hypotézy $H_0: \beta_1 = 0$ a $H_0: \beta_0 = 0$ lze sice formálně testovat také, ale zpravidla nemají žádnou rozumnou interpretaci.

Kvadratická regrese

Příklad (Závislost doby vývoje kříска na teplotě)

Kdysi jsme počítali korelací doby vývoje kříска bramborového na teplotě a usoudili jsme, že pořebujeme více dat pro nižší teploty. Získali jsme několik dalších měření a pokusili jsme se popsat závislost pomocí jednoduchého lineárního modelu. Dostali následující výsledky:

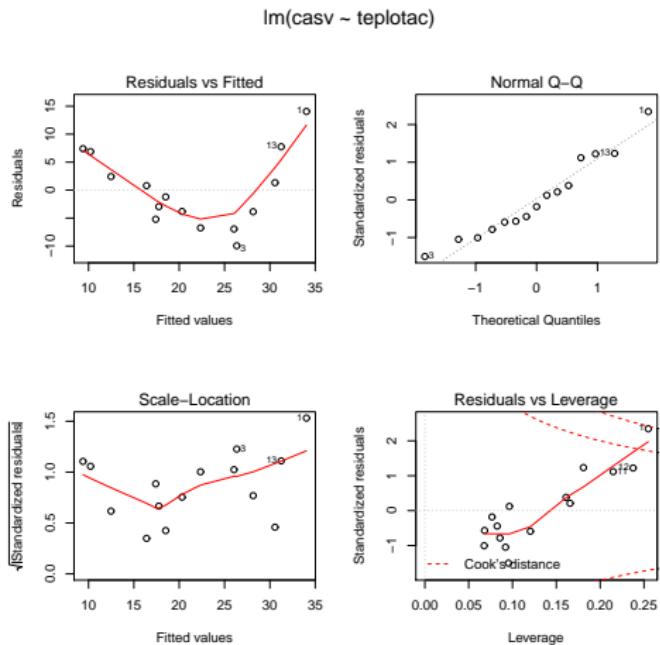
	odhad	S.E.(b)	T	p-hod.
b_0	54.98	8.22	6.688	0.00002 ***
b_1 (teplota)	-1.36	0.32	-4.184	0.00107 **

Koefficient determinace byl $R^2 = 0.574$, F-test celého modelu $F=17.51$ ($df=1, 13$), $p = 0.00107$.

Diagnostické grafy ale nevypadají dobře. Zdá se, že závislost není lineární. Model nevypadá rozumně.

Kvadratická regrese

Obrázek: Diagnostika: Doba vývoje kříска/teplota



Kvadratická regrese

Příklad (Závislost doby vývoje kříска na teplotě)

Zkusili jsme sestavit model kvadratické regrese. V tabulce jsou odhadovány parametry:

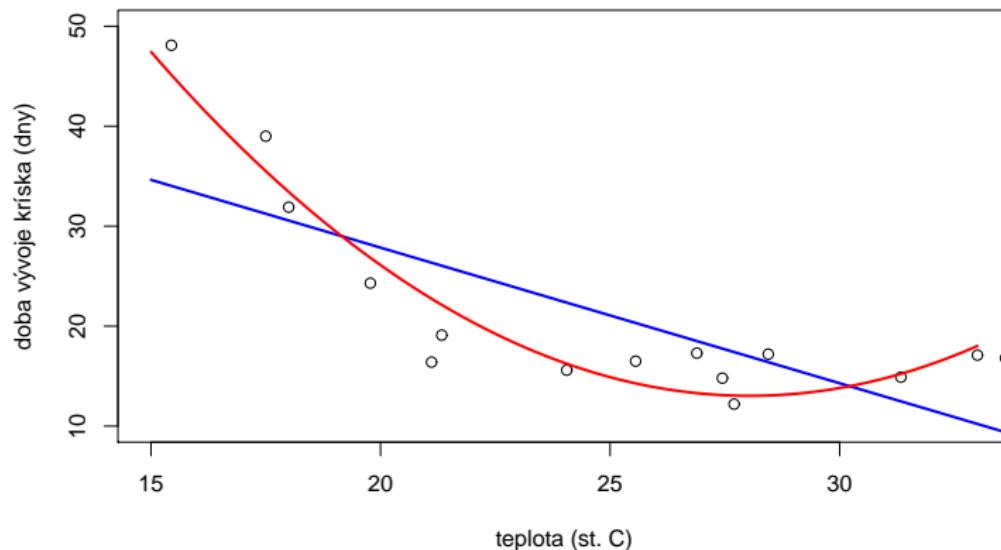
	odhad	S.E.(b)	T	p-hod.
b_0	172.20	17.60	9.786	$4.5 \cdot 10^{-7}$ ***
$b_1 (t)$	-11.35	1.47	-7.713	$5.4 \cdot 10^{-6}$ ***
$b_2 (t^2)$	0.20	0.03	6.828	0.00002 ***

Koefficient determinace $R^2 = 0.912$, F-test celého modelu
 $F=62.795$ ($df=2, 12$), $p = 4.4 \cdot 10^{-7}$

*Model sedí lépe, i když stále není ideální. Diagnostika se zlepšila.
 Závislost je kvadratickým modelem popsána lépe, alespoň na určitém intervalu.*

Kvadratická regrese

Obrázek: Regresní přímka a kvadratická křivka - vývoj kříска



Polynomiální regrese

- Podobně jako kvadratickou regresi lze sestavit regresní model používající jinou funkční závislost mezi Y a X .
- Nejčastěji se používají polynomy, které mohou popisovat i značně komplikované závislosti mezi Y a X .
- Opatrnost je ale na místě, v případě polynomiální regrese jsou regresory různými mocninami téže veličiny, takže jsou mezi sebou korelované, což může způsobovat různé problémy.

Souvislost regrese s analýzou rozptylu

- Jak jsme viděli, v regresním modelu můžeme použít stejnou techniku jako v modelech analýzy rozptylu a rozkládat celkovou variabilitu (S_T) na část vysvětlenou modelem (v regresi spojitými regresory a v ANOVě diskrétními faktory) a část, kterou jsme nevysvětlili - variabilitu reziduální (v regresi druhé mocniny odchylek od regresní přímky, v ANOVě druhé mocniny odchylek od průměrů ve všech výběrech).
- Jak regresní modely, tak modely analýzy rozptylu patří do rodiny *lineárních modelů* a liší se pouze tím, jaké měřítko mají vysvětlující proměnné. U spojitéh veličin obvykle mluvíme o regresi, u kategoriálních veličin o analýze rozptylu.

Jednoduché třídění jako lineární model

- Pokud A je faktor s k úrovněmi, můžeme zavést indikátory X_i , což jsou nula-jedničkové veličiny $X_i = 1$, pokud má subjekt i -tou úroveň A , jinak $X_i = 0$.
- Model jednoduchého třídění pak můžeme parametrizovat třeba takto:

$$Y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \dots + \beta_{k-1} x_{(k-1)j} + \varepsilon_j,$$

kde ε_j jsou náhodné chyby a x_{ij} indikátory i -té úrovně faktoru A .

- Pokud je tedy Y_j v i -té kategorii faktoru A , pak všechny indikátory kromě i -tého budou rovny 0 a zjistíme, že stř. hodnota je

$$EY_j = \beta_0 + \beta_i.$$

- Všimněte si, že indikátor k -té úrovně již v modelu není, pokud jsou všechny ostatní indikátory 0, je jasné, že subjekt má k -tou úrovneň A . Pokud je tedy Y_j v k -té úrovni A , platí v této parametrizaci $EY_j = \beta_0$ a $\beta_1, \dots, \beta_{k-1}$ jsou odchylky stř. hodnot ostatních úrovní A od k -té.

Lineární modely

- Vidíme tedy, že model jednoduchého třídění lze parametrisovat tak, aby měl tvar regresního modelu.
- Obecný lineární model může používat spojité i faktoriální regresory.
- Máme-li v modelu nula-jedničkový regresor, znamená to posunutí odhadu o konstantu, tedy aditivní vliv příslušného faktoru bez ohledu na hodnoty ostatních regresorů.
- Nemůžeme-li aditivitu faktoru předpokládat, můžeme použít interakce s ostatními regresory podobně jako v analýze rozptylu.

Lineární modely

Příklad (Hmotnost dětí - data Kojení)

Použijeme data Kojení a pokusíme se najít model, který by vysvětloval hmotnost miminek v půl roce pomocí jejich délky v půl roce a pohlaví. Mohli bychom sestavit například model:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i,$$

kde Y je hmotnost miminka v g, x_1 je délka miminka v cm a x_2 je indikátor pohlaví: $x_2 = 1$ pro hochy, $x_2 = 0$ pro dívky.

Uvědomme si, že podle modelu stř. hodnota hmotnosti pro dívky je $EY_i = \beta_0 + \beta_1 x_{1i}$ a pro chlapce $EY_i = \beta_0 + \beta_1 x_{1i} + \beta_2$. V obou případech bude mít regresní přímka závislosti hmotnosti na délce stejnou směrnici (β_1), přímka bude pouze pro chlapce posunutá o β_2 .

Lineární modely

Příklad (Hmotnost dětí - data Kojení)

Kdybychom chtěli dovolit pro chlapce i jinou směrnici závislosti hmotnosti na délce, musíme přidat ještě jeden člen do modelu (interakci pohlaví a délky):

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i}x_{2i} + \varepsilon_i,$$

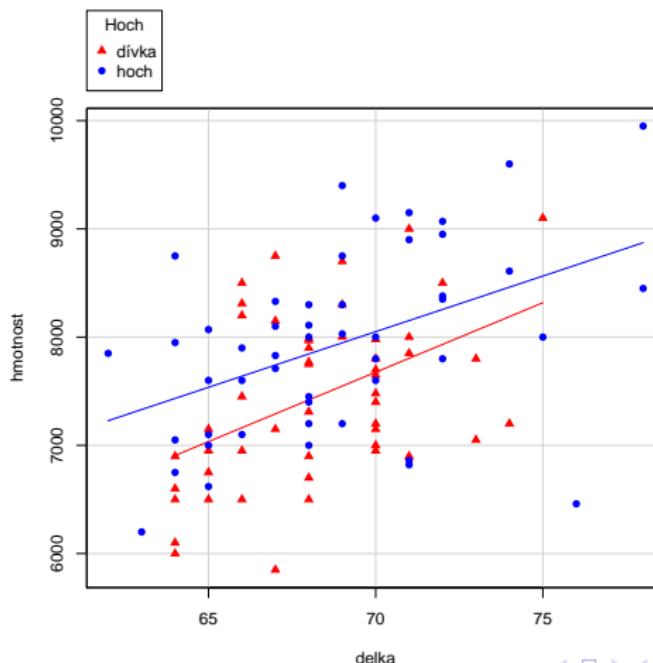
kde Y je hmotnost miminka v g, x_1 je délka miminka v cm a x_2 je indikátor pohlaví: $x_2 = 1$ pro hochy, $x_2 = 0$ pro dívky.

V tomto modelu stř. hodnota hmotnosti pro dívky je $EY_i = \beta_0 + \beta_1 x_{1i}$ a pro chlapce $EY_i = \beta_0 + \beta_2 + (\beta_1 + \beta_3)x_{1i}$. Tento model tedy dovoluje pro chlapce zcela odlišnou regresní přímku závislosti hmotnosti na délce než pro dívky.

Zkusíme tedy nejprve spočítat tento model s interakcemi. Regresní přímky pro chlapce a dívky zvlášť jsou zakreslené ve scatterplotu.

Lineární modely

Obrázek: Regrese hmotnost/délka (hoši a dívky zvlášt')



Lineární modely

Příklad (Hmotnost dětí - data Kojení)

Podle scatterplotu se směrnice regresních přímek závislosti hmotnost/délka pro chlapce a dívky se liší málo. Odhadu modelu s interakcemi jsou v tabulce

	odhad	S.E.(b)	T	p-hod.
b_0	-1304.71	2608.19	-0.500	0.6181
b_1 (délka)	128.29	38.17	3.361	0.0011 **
b_2 (pohl=hoch)	2158.82	3254.87	0.663	0.5088
b_3 (délka*hoch)	-25.49	47.49	-0.537	0.5927

Interakce nejsou v modelu důležité a můžeme je vyněchat (odpovídá podobným směrnicím přímek). Ostatní p-hodnoty zatím nehodnotíme, mohou se změnit po vynechání interakcí. Model bez interakcí:

	odhad	S.E.(b)	T	p-hod.
b_0	-180.26	1548.29	-0.116	0.9076
b_1 (délka)	111.83	22.62	4.943	$3.3 \cdot 10^{-6}$ ***
b_2 (pohl=hoch)	413.34	147.55	2.801	0.0062 **



Lineární modely

Příklad (Hmotnost dětí - data Kojení)

Vidíme, že odhad parametrů i testy o přínosech parametrů se výrazně změnily po vynechání interakcí. Oba regresory jsou teď v modelu významné. Zdá se tedy, že hmotnost souvisí s délkou i s pohlavím, ale interakce v modelu potřeba nejsou. Neprokázali jsme, že by regresní přímka závislosti hmotnosti na délce měla jinou směrnici pro dívky než pro chlapce.

Odhadujeme, že chlapci, při stejné délce, budou mít stř. hodnotu hmotnosti asi o 413 g vyšší než dívky. Pokud se délka dítěte zvětší o 1 cm, odhadujeme zvýšení stř. hodnoty váhy o 112 g.

Koefficient determinace pro model s interakcemi byl $R_1^2 = 0.268$, po vynechání interakcí se snížil velmi málo $R_2^2 = 0.266$. Interakce tedy opravdu téměř nic navíc nevysvětlují.

Závislosti mezi regresory

- Jsou-li regresory v modelu korelované, může být obtížné rozpoznat skutečnou souvislost vysvětlované proměnné a regresorů.
- V příkladu s hmotností a výškou miminek jsme viděli, jak výrazně se změnily odhady koeficientů, když jsme vynechali interakci. Stalo se to proto, že interakce je korelovaná s oběma regresory a část efektu regresorů byla v jejich interakci.
- Obecně, pokud přidáme do modelu regresor, který koreluje s jiným regresorem, může to zcela zásadně ovlivnit odhad regresního koeficientu i test o přínosnosti regresoru v modelu, dokonce může dojít k tomu, že závislost vysvětlované proměnné na původním regresoru změní znaménko.
- U dvojněho třídění korelacím mezi regresory odpovídají situace s nevyváženými daty (některé kombinace úrovní faktorů jsou častější).
- Pokud provádíme experiment a můžeme hodnoty regresorů ovlivnit, je dobré nastavit experiment tak, aby závislost mezi regresory byla co nejmenší.

Závislost mezi regresory

Příklad (Hmotnost dětí - data Kojení)

Řekněme, že bychom tentokrát chtěli odhadovat hmotnost dětí v půl roce pomocí jejich údajů při porodu (porodní váha a porodní délka).

Kdybychom sestavili jednoduchý regresní model pouze s porodní hmotností, zjistíme, že veličiny jsou závislé, $R^2 = 0.184$, odhady parametrů:

	odhad	S.E.(b)	T	p-hod.
b_0	4839.3388	614.37	7.877	$4.9 \cdot 10^{-12} ***$
b_1 (porHmotnost)	0.8215	0.1757	4.676	$9.5 \cdot 10^{-6} ***$

Kdybychom sestavili jednoduchý model pouze s porodní délkou, zjistíme také signifikantní souvislost, $R^2 = 0.167$, odhady parametrů:

	odhad	S.E.(b)	T	p-hod.
b_0	3564.42	2555.90	-1.395	0.1660
b_1 (porDelka)	222.43	50.49	4.405	0.00003 ***

Závislost mezi regresory

Příklad (Hmotnost dětí - data Kojení)

Model s porodní hmotností i porodní délkou ovšem vypadá takto:

	odhad	S.E.(b)	T	p-hod.
b_0	720.317	3387.581	0.213	0.8321
b_1 (porDelka)	100.577	81.353	1.236	0.2194
b_2 (porHmotnost)	0.542	0.286	1.895	0.0611

Mohlo by zdát, že ani porodní délka ani porodní hmotnost s hmotností v půl roce nesouvisí. My ovšem víme, že obě veličiny s ní souvisí. Navíc F-test celého modelu vydal p-hodnotu 0.000027 a $R^2 = 0.197$.

Zdánlivá nepříenosnost regresorů se objevila kvůli jejich silné korelacii ($r = 0.79$). Když zahrneme do modelu jeden z nich, druhý už se jeví nepříenosný (t-testy testují přínos regresoru za přítomnosti ostatních regresorů v modelu). Vidíme že koeficient determinace se zvýšil velmi málo, když jsme zahrnuli druhou veličinu. Obě veličiny tedy v modelu nepotřebujeme, je-li tam jedna, druhá již mnoho nepřidává. Vynecháme spíš porodní délku, protože porodní hmotnost vysvětluje více variability hmotnosti v půl roce než porodní délka.

Výběr modelu

- Často řešíme situaci, kdy máme k dispozici několik (třeba i větší množství) potenciálních regresorů a chceme vybrat model, který bude "ideální". Měl by obsahovat pouze regresory, které souvisí s vysvětlovanou proměnnou, měl by dobře predikovat a ještě být snadno interpretovatelný.
- Z úvah v předchozích částech asi tušíme, že výběr modelu nemusí být snadná záležitost.
- Můžeme použít sestupný výběr. Nejprve vytvoříme model se všemi regresory a potom postupně vynecháváme regresory od těch, které jsou nejméně přínosné pro model.
- Nebo naopak můžeme postupovat vzestupně, od regresoru nejlépe vysvětlujícího odpověď'.
- Zcela automatizované postupy však někdy vedou k modelům, které jsou nevhodné nebo špatně interpretovatelné.

Rady pro výběr modelu

- Začněte úvahou, proč model vlastně sestavujete. Chcete popsat na čem všem vysvětlovaná veličina závisí? Chcete predikovat? Zajímá vás vliv jednoho regresoru za přítomnosti všech ostatních? Často zodpovězení této otázky zodpoví i otázku jak model vybírat.
- Někdy máte požadavky vyplývající ze zadání, např., že určité regresory v modelu musí být (i kdyby vám připadaly nepřínosné).
- Bud'te opatrní, pokud jsou některé regresory silně korelované. Někdy může být rozumnější vkládat do modelu jen jeden z nich.
- S interakcemi, případně vyššími mocninami některých regresorů šetřete, používejte je hlavně když máte na něco takového podezření z logiky řešeného problému.
- Pokud je v modelu významná interakce, vždy ponechte v modelu i příslušné interagující regresory, i kdyby náhodou nevypadaly v modelu přínosně (ve skutečnosti jsou).

Rady pro výběr modelu

- Pokud je regresorem nějaký faktor, reprezentovaný v modelu několika indikátory (více než 2 úrovně), chovějte se k těmto indikátorům v modelu jako k nedělitelné skupině (buď je v modelu ponechte všechny a nebo je všechny vynechte).
- Pokud chcete testovat, zda celá skupina regresorů je přínosná v modelu (např. skupina indikátorů odpovídající úrovním jednoho faktoru), můžete použít F -test porovnávající model s podmodelem bez příslušné skupiny regresorů. F -test je založen na porovnání variability, kterou vysvětluje testovaná skupina regresorů (za přítomnosti ostatních regresorů) s variabilitou reziduální.
- Není špatné zkontrolovat diagnostické grafy, i ty se budou měnit, pokud vynecháte z modelu něco důležitého.
- Nesnažte se vymyslet co nejkomplikovanější model. Přemýšlejte, zda lze vůbec model rozumně interpretovat. Často je jednodušší model vhodnější.

Kde jsme?

- Věnujeme se zkoumání závislostí mezi veličinami.
- Nejprve jsme se naučili zkoumat a testovat závislosti veličiny kvantitativní a kvalitativní (dvouvýběrový t -test nebo ANOVA).
- Pak jsme se naučili testovat závislosti dvou kvantitativních veličin (korelace, jednoduchá regrese) a pak jsme udělali malou odbočku k modelování (mnohonásobná regrese).
- Zbývají nám ještě testy závislostí mezi dvěma kvalitativními veličinami.
- Začněme nejprve rozdelením jedné kvalitativní veličiny.

Multinomické rozdělení

Příklad (Souvisí citlivost na infekci koronaviru s krevní skupinou?)

Pokud krevní skupiny nesouvisí s pravděpodobností, že se člověk nakazí koronavirem, měli bychom mezi nakaženými pozorovat stejné rozložení krevních skupin jako v populaci, ze které nakažení pocházejí. U skupiny 1775 nakažených v provincii Wuhan byl proveden test krevní skupiny, výsledky jsou v tabulce spolu a rozložením krevních skupin ve Wuhanu.

krev. skupina	A	B	AB	0	všichni
% ve Wuhanu	32.16	24.90	9.10	33.84	100.00
nakažení	670	469	178	458	1775

Multinomické rozdělení

- Počty nakažených v kategoriích podle krevní skupiny jsou realizací náhodného vektoru (Y_1, Y_2, \dots, Y_k) , Y_i je počet nakažených s i -tou krevní skupinou.
- Situace je podobná jako u binomického rozdělení, každý pokus ovšem může dopadnout k způsoby.
- Rozdělení (Y_1, Y_2, \dots, Y_k) nazveme *multinomickým rozdělením* s parametry n (počet pokusů) a $\pi_1, \pi_2, \dots, \pi_k$ (pravděpodobnosti možností). Musí platit $\pi_1 + \pi_2 + \dots + \pi_k = 1$, jedna z možností nastat musí.
- Z kombinatoriky plyne, že pro n_1, \dots, n_k takové, že $n_1 + n_2 + \dots + n_k = n$ platí:

$$P(Y_1 = n_1, Y_2 = n_2, \dots, Y_k = n_k) = \frac{n!}{n_1! n_2! \dots n_k!} \pi_1^{n_1} \pi_2^{n_2} \dots \pi_k^{n_k}$$

Multinomické rozdělení

- Binomické rozdělení je speciálním případem multinomického pro $k = 2$.
- Každá ze složek multinomického rozdělení má binomické: $Y_j \sim Bi(n, \pi_j)$
- Pro střední hodnoty a rozptyly tedy platí: $EY_j = n\pi_j$ a $\text{var} Y_j = n\pi_j(1 - \pi_j)$, pro $j = 1, 2, \dots, k$
- Veličiny Y_i a Y_j jsou závislé ($Y_1 + Y_2 + \dots + Y_k = n$).
 $\text{cov}(Y_i, Y_j) = -n\pi_i\pi_j$ (pro $i \neq j$).

Test dobré shody

- (Y_1, Y_2, \dots, Y_k) má multinomické rozdělení a chceme testovat, zda pravděpodobnosti kategorií se rovnají očekávaným, tj. testujeme $H_0: \pi_1 = \pi_1^0, \pi_2 = \pi_2^0, \dots, \pi_k = \pi_k^0$ proti alternativě $H_1: \text{Neplatí } H_0$.
- Pokud platí H_0 , pak očekávaná četnost v j -té kategorii bude $n\pi_j^0$. Bylo by tedy vhodné založit testovou statistiku na rozdílech $Y_j - n\pi_j^0$. Vysoké absolutní hodnoty těchto rozdílů budou svědčit proti H_0 .
- Testová statistika

$$X^2 = \sum_{j=1}^k \frac{(Y_j - n\pi_j^0)^2}{n\pi_j^0}$$

má za H_0 přibližně rozdělení χ^2_{k-1} , můžeme použít kvantily χ^2_{k-1} k nalezení kritického oboru nebo p -hodnoty.

- Test založený na X^2 nazveme *testem dobré shody*. Test bude zamítat na hladině α , pokud $X^2 \geq \chi^2_{k-1}(1 - \alpha)$.

Předpoklad testu dobré shody

- Rozdělení testové statistiky X^2 je pouze přibližné, nebude fungovat dobře při malých počtech pozorování.
- Důležité jsou očekávané četnosti za nulové hypotézy $n\pi_j^0$, které musí být dostatečně vysoké. Za dostatečné se považuje, pokud platí $n\pi_j^0 > 5$, pro všechna $j = 1, 2, \dots, k$.
- Minimální počet pozorování nutný pro fungování testu tedy závisí na tom, jak pravděpodobná má být nejméně častá kategorie.

Test dobré shody

Příklad (Koronavirus a krevní skupiny)

Počet nakažených ve studii: $n = 1775$.

Testujeme $H_0: \pi_1 = 0.3216, \pi_2 = 0.2490, \pi_3 = 0.0910,$
 $\pi_4 = 0.3384$.

Testová statistika:

$$\chi^2 = \frac{(670 - 1775 \cdot 0.3216)^2}{1775 \cdot 0.3216} + \frac{(469 - 1775 \cdot 0.2490)^2}{1775 \cdot 0.2490} + \frac{(178 - 1775 \cdot 0.0910)^2}{1775 \cdot 0.0910} + \frac{(458 - 1775 \cdot 0.3384)^2}{1775 \cdot 0.3384} = 54.44$$

P -hodnota je $9.0 \cdot 10^{-12}$, tedy přesvědčivě zamítáme H_0 . Nakažení ve studii nemají stejné rozdělení krevních skupin jako celá populace Wuhanu.

Test dobré shody

Příklad (Koronavirus a krevní skupiny)

V tabulce jsou skutečné počty nakažených podle krevních skupin a očekávané četnosti za H_0 :

Krev. skup.	A	B	AB	0
počet	670	469	178	458
oček.č.	570.8	442.0	161.5	600.7

Všechny očekávané četnosti > 5, předpoklad testu splněn.

Nakažených se skupinou 0 bylo výrazně méně než bychom očekávali. Naopak výrazně vyšší byl počet nakažených se skupinou A, trochu více bylo i nakažených se skupinou B a AB.

Možná je tedy skupina 0 odolnější proti koronaviru a nejméně odolná skupina A. Nicméně, počkala bych na potvrzení jinou studií. Šlo skutečně o náhodný výběr nakažených z Wuhanské populace?

Test dobré shody

Příklad (Kostky)

Kdysi jsme házeli 100krát červenou a modrou kostkou a testovali, zda pravděpodobnost, že padne šestka je skutečně $\frac{1}{6}$. U červené jsme H_0 nezamítli (šestka padla 16krát), u modré jsme zamítli (šestka padla pouze 9krát).

Co kdybychom chtěli testovat, že kostka je správná, tj., že všechna čísla padají s pravděpodobností $\frac{1}{6}$? Bylo by rozumné udělat 6 testů na binomickou pravděpodobnost, podobně jako pro šestku? Proč?

Testujeme správnost kostky testem dobré shody.

$$H_0: \pi_1 = \pi_2 = \dots = \pi_6 = \frac{1}{6}.$$

Test dobré shody

Příklad (Kostky - pokračování)

Na modré kostce padla čísla 1-6 s následujícími četnostmi :
(26, 18, 14, 12, 21, 9), na červené (12, 10, 23, 18, 21, 16).

$$X_M^2 = \frac{1}{16.67} [(25 - 16.67)^2 + (18 - 16.67)^2 + \dots + (9 - 16.67)^2] \\ = 11.72$$

P-hodnota je 0.0388, zamítáme tedy, že modrá kostka je správná.

X_C^2 = 7.64, p-hodnota 0.1772. Nezamítáme, že červená kostka je správná.

Všechny očekávané četnosti byly $100 \cdot \frac{1}{6} \doteq 16.67$, žádné problémy s předpoklady nenastaly.

Přednáška 12 (5.5.2020) - obsah

- χ^2 test nezávislosti v kontingenční tabulce
- Test homogenity, test symetrie
- Čtyřpolní tabulka, Fisherův test
- Podíl šancí a relativní riziko
- Epidemiologické studie

Závislost dvou kvalitativních veličin

Příklad (Svižníci)

Dospělí svižníci druhu *Cicindela fulgida* mají zpočátku jasně červenou barvu a ve stáří tmavnou a zbarvují se do hněda. Líhnou se z kukel v období duben-září a jako dospělci přezimují do dalšího roku. Aby se zjistilo, zda líhnutí dospělých brouků probíhá rovnoměrně celém časovém intervalu, byli sbíráni brouci do pastí v určitých obdobích a kategorizováni podle barvy. Zajímá nás, zda období sběru souvisí s barvou svižníků. Následující tabulka obsahuje výsledky sběru:

	<i>jasně červená</i>	<i>tmavočervená</i>	<i>hnědá</i>	<i>všichni</i>
<i>duben-květen</i>	28	7	5	40
<i>červen-červenec</i>	28	26	14	68
<i>srpen-září</i>	68	22	38	128
<i>všichni</i>	124	55	57	236

Kontingenční tabulka

Máme-li 2 kvalitativní (kategorické) veličiny A a B , můžeme data shrnout do *kontingenční tabulky*. Nechť A má k kategorií A_1, A_2, \dots, A_k a B má r kategorií B_1, \dots, B_r . Kontingenční tabulka vypadá takto:

	B_1	B_2	...	B_r	všechny
A_1	n_{11}	n_{12}	...	n_{1r}	$n_{1.}$
A_2	n_{21}	n_{22}	...	n_{2r}	$n_{2.}$
...
A_k	n_{k1}	n_{k2}	...	n_{kr}	$n_{k.}$
všechny	$n_{.1}$	$n_{.2}$...	$n_{.r}$	$n_{..} = n$

Četnosti kombinací faktorů $A_i \cap B_j$ označíme n_{ij} . Poslední sloupec (řádek) obsahuje marginální četnosti, tj. četnosti v celých kategoriích A_i (B_j).

χ^2 test nezávislosti

- Četnosti $(n_{11}, n_{12}, \dots, n_{kr})$ mají multinomické rozdělení s parametry n a π_{ij} , $i = 1, \dots, k$ a $j = 1, \dots, r$, což jsou pravděpodobnosti $A_i \cap B_j$.
- Chceme testovat H_0 : A a B jsou nezávislé, proti alternativě H_1 : A a B jsou závislé.
- Použijeme opět statistiku X^2 podobně jako v testu dobré shody, jenom změníme očekávané četnosti tak, aby odpovídaly H_0 . Prozatím označíme o_{ij} očekávané četnosti za H_0 , potom

$$X^2 = \sum_{i=1}^k \sum_{j=1}^r \frac{(n_{ij} - o_{ij})^2}{o_{ij}}.$$

- Za H_0 má X^2 přibližně rozdělení $\chi^2_{(k-1)(r-1)}$.

Očekávané četnosti za hypotézy nezávislosti

- Zbývá nám rozmyslet si, co jsou o_{ij} , očekávané četnosti za hypotézy nezávislosti.
- Pokud jsou A a B nezávislé, pak musí platit

$$P(A_i \cap B_j) = P(A_i)P(B_j)$$

pro všechna $i = 1, \dots, k$ a $j = 1, \dots, r$.

Označíme-li marginální pravděpodobnosti $\pi_{i\cdot}$ a $\pi_{\cdot j}$, lze totéž vyjádřit rovnicí: $\pi_{ij} = \pi_{i\cdot} \cdot \pi_{\cdot j}$.

- Žádné z těchto pravděpodobností neznáme, ale můžeme je odhadnout: $\hat{\pi}_{ij} = \frac{n_{ij}}{n}$, $\hat{\pi}_{i\cdot} = \frac{n_{i\cdot}}{n}$, $\hat{\pi}_{\cdot j} = \frac{n_{\cdot j}}{n}$.
- Za H_0 budeme tedy přibližně očekávat $\frac{n_{ij}}{n} \doteq \frac{n_{i\cdot}}{n} \cdot \frac{n_{\cdot j}}{n}$, což je ekvivalentní $n_{ij} \doteq \frac{n_{i\cdot} \cdot n_{\cdot j}}{n}$.
- Jako očekávanou četnost za H_0 tedy použijeme $\hat{o}_{ij} = \frac{n_{i\cdot} \cdot n_{\cdot j}}{n}$.

χ^2 test nezávislosti

- χ^2 *test nezávislosti* v kontingenční tabulce bude tedy založen na testové statistice:

$$\chi^2 = \sum_{i=1}^k \sum_{j=1}^r \frac{(o_{ij} - \frac{n_i \cdot n_j}{n})^2}{\frac{n_i \cdot n_j}{n}}$$

- Proti H_0 svědčí vysoké hodnoty χ^2 . Test bude zamítat pro $\chi^2 \geq \chi^2_{(k-1)(r-1)}(1 - \alpha)$.
- χ^2 rozdělení testové statistiky za H_0 platí pouze přibližně. Za dostačující počet pozorování, aby test fungoval, se považuje takové n , že $o_{ij} > 5$ pro všechna i a j , tj., všchny očekávané četnosti jsou větší než 5.

χ^2 test nezávislosti

Příklad (Svižníci)

V tabulce uvádíme kromě četností také očekávané četnosti za hypotézy nezávislosti (v závorce), všechny jsou větší než 5.

	jasně červená	tmavočervená	hnědá	všichni
duben-květen	28 (21.0)	7 (9.3)	5 (9.7)	40
červen-červenec	28 (35.7)	26 (15.8)	14 (16.4)	68
srpen-září	68 (67.3)	22 (29.8)	38 (30.1)	128
všichni	124	55	57	236

Testová statistika $X^2 = 17.37$, $p = 0.0016$. Zamítáme H_0 na hladině 0.05. Barva souvisí s obdobím sběru, svižníci se nejspíš nelíhnou rovnoměrně.

Skutečné a očekávané četnosti dávají obrázek závislosti. Zdá se, že na jaře je více jasně červených svižníků, zpočátku léta naopak méně. Větší procento tmavočervených na počátku léta a hnědých na konci léta odpovídá větší intenzitě líhnutí na jaře, ale nejspíš také na konci léta, kdy je jasně červených opět dost.



Test homogeneity

- Mohli bychom otázku formulovat trochu jinak. Mohli bychom se ptát, zda je barevné rozložení svižníků stejné v každém časovém období. Pak bychom pevně stanovili, že chytíme třeba 50 svižníků na jaře, 50 na počátku léta a 50 v pozdním létě a budeme zkoumat rozložení barev.
- V tomto případě bychom testovali

$$H_0: \pi_{11} = \pi_{21} = \dots = \pi_{k1},$$

$$\pi_{12} = \pi_{22} = \dots = \pi_{k2},$$

.....

$$\pi_{1r} = \pi_{2r} = \dots = \pi_{kr},$$

tj. hypotézu, že všechna multinomická rozdělení v řádcích tabulky (nebo ve sloupcích) mají stejné pravděpodobnosti. Takovému testu se říká *test homogeneity*.

Test homogeneity

- Na rozdíl od testu nezávislosti, u testu homogeneity považujeme jedny z marginálních četností za pevné, máme tedy několik výběrů z multinomického rozdělení.
- Formálně se test homogeneity provádí stejně jako test nezávislosti, očekávané četnosti za H_0 jsou stejné o_{ij} jako u testu nezávislosti. Používáme stejnou testovou statistiku:

$$\chi^2 = \sum_{i=1}^k \sum_{j=1}^r \frac{(n_{ij} - o_{ij})^2}{o_{ij}},$$

která má za H_0 přibližně rozdělení $\chi^2_{(k-1)(r-1)}$.

- Volba mezi testem nezávislosti a homogeneity závisí na filosofii řešeného problému a způsobu sběru dat.

Test homogeneity

Příklad (Kostky)

Chěli bychom testovat, zda červená a modrá kostka z našeho pokusu mají stejné pravděpodobnosti všech čísel, bez ohledu na to zda jsou kostky správné.

V tomto případě předem určíme, kolikrát budeme kterou kostkou házet, budeme tedy provádět test homogeneity. Nulová hypotéza bude $H_0: \pi_1^M = \pi_1^C, \pi_2^M = \pi_2^C, \dots, \pi_6^M = \pi_6^C$. Již jsme každou kostkou hodili 100krát. Výsledky jsou v tabulce:

	1	2	3	4	5	6	vše
M	26	18	14	12	21	9	100
Č	12	10	23	18	21	16	100
obě	38	28	37	30	42	25	200

Test homogeneity

Příklad (Kostky - pokračování)

Testová statistika $X^2 = 12.806$, za H_0 by měla mít rozdělení χ^2 s $(k - 1)(r - 1) = 5 \cdot 1 = 5$ stupni volnosti. Dostáváme p-hodnotu 0.0253. Na hladině 0.05 tedy zamítáme, že kostky mají stejné multinomické rozdělení.

Očekávané četnosti budou pro každý počet ok stejné pro obě kostky, rovné průměru z četností příslušného počtu ok na obou kostkách (stejný počet hodů na obou kostkách). Všechny očekávané četnosti jsou větší než 5.

Test symetrie

Příklad (Porovnání PCR a kultivace u vzorků tkáně z ortopedie)

Bakterie ve vzorcích tkáně ortopedických pacientů byly hledány jak pomocí kultivace, tak pomocí PCR metody. Vzorky pro obě metody byly od pacienta odebrány ve stejný čas a ze stejného místa. U všech vzorků, které byly aspoň jednou metodou pozitivní, se infekce potvrdila i klinicky. Zajímá nás, jestli u ortopedických vzorků některá z metod byla úspěšnější v odhalování patogenů. V následující tabulce jsou výsledky obou metod na těchto vzorcích:

		Kultivace		
		pozitivní	negativní	
PCR	pozitivní	27	46	73
	negativní	24	54	78
		51	100	151

Test symetrie

- V tomto případě potřebujeme testovat testovat $H_0: \pi_{12} = \pi_{21}$ proti $H_1: \pi_{12} \neq \pi_{21}$, tj. test symetrie ve čtyřpolní tabulce.
- Za H_0 by mělo přibližně platit $n_{12} = n_{21}$. Testová statistika má tvar:

$$X^2 = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}},$$

za H_0 má rozdělení χ_1^2 . Tento test se nazývá *McNemarův*.

- Proti H_0 budou svědčit vysoké hodnoty X^2 . Budeme zamítat pro $X^2 \geq \chi_1^2(1 - \alpha)$.
- Všimněme si, že se jedná o obdobu párového testu pro nula-jedničkové veličiny. Vzorky odebrané ve stejnou dobu na stejném pacientovi jsou závislé páry. Výsledek je 0/1 (neg./poz. test).
- Test symetrie lze zobecnit i pro větší čtvercové tabulky, takový test se nazývá Bowkerův.

McNemarův test

Příklad (Porovnání PCR a kultivace u vzorků tkáně z ortopedie)

Testová satatistika McNemarova testu je $X^2 = \frac{(46-24)^2}{46+24} = 6.91$,
 p -hodnota je 0.0086.

Zamítáme tedy hypotézu, že obě metody jsou v odhalování patogenů stejně úspěšné. Na hladině 0.05 jsme prokázali, že PCR byla na ortopedických vzorcích úspěšnější.

- Podobně jako v případě spojitých veličin, je důležité odhalit závislé páry a použít příslušný test, jinak můžeme dospat k chybným závěrům.

McNemarův test

Příklad (Krční mandle a Hodgkinův lymfom (J.Rice 1995))

Od 70. let minulého století se dává do souvislosti odstranění krčních mandlí (tonsilektomie) s propuknutím Hodgkinova lymfomu. Jedna ze studií provedla následující šetření: U 85 pacientů s Hodgkinovým lymfomem, kteří měli zdravého sourozence stejného pohlaví bylo zjištěno, zda pacienti nebo jejich sourozenci prodělali tonsilektomii. Použitím sourozenců jako kontrol měly být eliminovány případné genetické dispozice. Ve studii byla uvedena následující tabulka:

	tonsilektomie	nic	všichni
Hodgkin lym.	41	44	85
sourozenci	33	52	85
všichni	74	96	170

Použili test homogeneity (Mají pacienti stejný podíl odstranění mandlí jako jejich zdraví sourozenci?), dostali $X^2 = 1.53$, p-hodnota=0.2159, takže neprokázali souvislost Hodgkinova lymfomu s odstraněním mandlí.



McNemarův test

Příklad (Krční mandle a Hodgkinův lymfom)

Uvedená analýza však není v pořádku, nebere v úvahu souvislost mezi sourozenci. Ve skutečnosti jde o výběr závislých párů, na kterých měříme nula-jedničkový znak (tonsilektomie/nic). Potřebujeme jinou tabulku:

		Sourozenec		všichni
Pacient	tonsilektomie	26	15	41
	nic	7	37	44
všichni		33	52	85

Z McNemarova testu, dostaneme $X^2 = 2.91$, $p=0.0880$. Sice stále nezmítáme hypotézu, že zdraví sourozenci měli stejně často odstraněné mandle jako nemocní, ale p -hodnota je výrazně nižší a problém rozhodně stojí za další zkoumání.

Souvislost tonsilektomie a Hodgkinova lym. se dále řeší, otázka je, zda je to souvislost příčinná nebo ne. Možná jsou to spíš záněty mandlí (vedoucí k jejich odstranění), které mohou mít příčinnou souvislost s lymfomem...



Čtyřpolní tabulka

Nejjednodušší kontingenční tabulka je 2x2 - *čtyřpolní*:

		faktor B		
		B_1	B_2	
faktor A	A_1	a	b	$a+b$
	A_2	c	d	$c+d$
		a+c	b+d	n

Testovou statistiku testu nezávislosti lze přepsat jako:

$$\chi^2 = \frac{n(ad-bc)^2}{(a+b)(c+b)(a+c)(b+d)},$$

za hypotézy nezávislosti má χ^2 přibližně rozdělení χ_1^2 .

Yatesova korekce

- Rozdělení X^2 platí pouze pro dostatečně vysoké očekávané četnosti. Pokud jsou nižší (obzvláště pokud jsou nižší než 5), doporučuje se použít tzv. *opravu na spojitost* (také Yatesova korekce).
- Testová statistika má pak tvar:

$$X^2 = \frac{n(|ad-bc|-n/2)^2}{(a+b)(c+b)(a+c)(b+d)},$$

- Korekce snižuje hodnotu X^2 , takže zvyšuje p -hodnotu. Při menších četnostech je tak spolehlivěji zajištěna příslušná hladina testu.

Fisherův faktoriálový test

- Místo χ^2 testu nezávislosti je možné při testování nezávislosti použít *Fisherův faktoriálový test* (také Fisherův exaktní test).
- Tento test fixuje marginální četnosti a hledá všechny možné tabulky, které ještě více odporují H_0 než tabulka, kterou jsme dostali. Tímto způsobem lze napočítat přímo p -hodnotu. Fisherův test nepoužívá žádnou testovou statistiku.
- Pro čtyřpolní tabulku je princip jasný, ale může být výpočetně náročný.
- Existují zobecnění pro větší tabulky.
- Fisherův test funguje i při malých očekávaných četnostech, je možné jej použít, pokud předpoklad χ^2 testu není splněn.

Testy ve čtyřpolní tabulce

Příklad (Zkoušky)

Chceme testovat, zda výsledek zkoušek ze statistiky souvisí s výsledkem z matematiky. Výběr 30 studentů vydal následující tabulku:

		Matematika		
		udělal	neudělal	
Statistika	udělal	14	5	19
	neudělal	3	8	11
		17	13	30

$X^2 = 6.11$, p -hodnota je 0.0134. Zdá se tedy, že jasně zamítáme H_0 , a prokazujeme na hladině 0.05, že výsledky zkoušek spolu souvisí.

Testy ve čtyřpolní tabulce

Příklad (Zkoušky - pokračování)

Spočítáme-li očekávané četnosti (uvedeny v závorce), vidíme, že jedna z nich je menší než 5. P-hodnota možná není správně.

		Matematika		
		udělal	neudělal	
Statistika	udělal	14 (10.77)	5 (8.23)	19
	neudělal	3 (6.23)	8 (4.77)	11
		17	13	30

Statistika s Yatesovou korekcí je $X^2 = 4.37$, p-hodnota 0.0366.

Fisherův test vydá p-hodnotu 0.0228. V obou případech jsme dostali vyšší p-hodnotu, nicméně stále zamítáme H_0 . Výsledky zkoušek souvisí. Případů, kdy student obě zkoušky udělá nebo obě neudělá, je více, než by jich bylo za nezávislosti.

Podíl šancí

- Kromě testování potřebujeme také popsat míru závislosti kvalitativních veličin. Zajímá nás souvislost dvou nula-jedničkových veličin A a B a můžeme si jejich dvě úrovně představit tak, že nějaký jev A (nebo jev B) nastal nebo nenastal (úrovně ano/ne). Nula jedničkové veličiny jsou tedy indikátory těchto jevů.
- Z náhodného výběru jsme spočítali čtyřpolní kontingenční tabulku:

		B		
		ano	ne	
A	ano	a	b	a+b
	ne	c	d	c+d
		a+c	b+d	n

- Podíl a/c je podíl počtu případů, kdy nastal jev A ku případům, kdy nenastal jev A , měřeno za situace, kdy nastal jev B . a/c je tedy odhad šance jevu A v situaci, kdy nastal jev B . Podobně podíl b/d je odhad šance A v situaci, kdy B nenastal.

Podíl šancí

- Pokud šance a/c a b/d budou podobné, potom to, zda nastal jev B nejspíš neovlivňuje šanci jevu A , čili jevy (a také náhodné veličiny, které jsou jejich indikátory) nejspíš nebudou závislé.
- Jako míru závislosti veličin A a B můžeme použít *podíl šancí*, angl. *odds ratio*:

$$\hat{\theta} = \frac{ad}{bc}.$$

- $\hat{\theta}$ je výběrová míra závislosti dvou nula-jedničkových veličin a odpovídá jí populační (teoretický) podíl šancí θ , který bude roven 1, pokud jsou veličiny A a B nezávislé.
- $\hat{\theta}$ můžeme považovat pro nula-jedničkové veličiny za jakousi obdobu výběrového korelačního koeficientu.

Interval spolehlivosti pro podíl šancí

- Pro představu, jak přesně jsme odhadli podíl šancí, je dobré zkonztruovat interval spolehlivosti pro θ .
- Využijeme, že střední chyba přirozeného logaritmu $\log(\hat{\theta})$ je přibližně:

$$S.E.\log(\hat{\theta}) \doteq \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}.$$

- Přibližný interval spolehlivosti pro $\log(\theta)$ je

$$(\log(\hat{\theta}) - S.E.\log(\hat{\theta})z(1-\alpha/2), \log(\hat{\theta}) + S.E.\log(\hat{\theta})z(1-\alpha/2))$$

a interval spolehlivosti pro θ dostaneme odlogaritmováním mezi.

- Interval spolehlivosti lze použít také k testování $H_0: \theta = 1$, tj. testování nezávislosti. Budeme zamítat H_0 , pokud CI nepokryje 1.

Podíl šancí

Příklad (Zkoušky)

Vrátíme se k příkladu se zkouškami ze statistiky a matematiky.

Podíl šancí je

$$\hat{\theta} = \frac{14 \cdot 8}{3 \cdot 5} = 7.47$$

Je tedy vysoký, což odpovídá tomu, že už jsme pomocí testů prokázali závislost.

95% interval spolehlivosti pro $\log(\theta)$ vychází: (0.3361, 3.6848), což po odlogaritmování vydá interval pro θ : (1.3995, 39.8372).

Vidíme, že interval nepokryvá 1, takže potvrzujeme i tímto způsobem, že výsledky zkoušky ze statistiky a matematiky jsou závislé. Interval je ovšem velmi široký, takže skutečné θ jsme odhadli nepřesně.

Relativní riziko

- Podíl šancí se často používá v epidemiologii, kde nás zajímá souvislost nějakého rizikového faktoru (expozice) E a nemoci N .
- Označme tedy E jev, že subjekt byl vystaven tomuto rizikovému faktoru a N jev, že dostal zkoumanou nemoc.
- V tomoto problému nás obvykle zajímají pravděpodobnosti $P(N|E)$ a $P(N|\bar{E})$, tj. pravděpodobnosti, že osoba onemocní, pokud byla/nebyla vystavena rizikovému faktoru. Tyto pravděpodobnosti označujeme jako *absolutní rizika*.
- Jako míra asociace mezi E a N se používá *relativní riziko*

$$RR = \frac{P(N|E)}{P(N|\bar{E})},$$

které měří kolikrát vyšší mají pravděpodobnost nemoci lidé vystaveni E .

Relativní riziko

- Výsledky našich pozorování shrneme do čtyřpolní tabulky:

		E		
		ano	ne	
N	ano	a	b	a+b
	ne	c	d	c+d
		a+c	b+d	n

- Pak RR můžeme odhadnout tak, že odhadneme pravděpodobnosti pomocí relativních četností a dostáváme:

$$\hat{RR} = \frac{a/(a+c)}{b/(b+d)}$$

- Pokud je nemoc ve sledované populaci vzácná, pak a bude velmi malé v porovnání s c a podobně b bude velmi malé v porovnání s d . V takovém případě bude $\hat{RR} \sim \hat{\theta}$, podíl šancí bude podobný relativnímu riziku. Toho se využívá ve studiích, které díky způsobu výběru subjektů neumožňují přímý odhad RR .

Randomizovaný experiment

- Plánujeme-li studii na zkoumání souvislosti rizikového faktoru E s nemocí N (nemocí N teď rozumíme obecně nějakou událost, nemusí to být nutně diagnóza nemoci, může to být třeba smrt nebo přítomnost komplikací apod.), můžeme zvolit různé přístupy.
- Ideální situace nastane, když můžeme provést *randomizovaný experiment*. Subjekty náhodně přiřadíme do skupiny, která je vystavena E a skupiny, která není vystavena E , a sledujeme, zda nastane N .
- Tento přístup, který je standardem při testování nových léků (klinické zkoušky), však v epidemiologii často není možný.
- Někdy nemáme možnost ovlivnit expozici subjektů (nemohu lidi náhodně přiřadit na kouření/nekouření nebo užívání/neužívání hormonální antikoncepcie).
- Často je nemoc N velmi vzácná a museli bychom provést experiment s mnoha tisíci účastníků, aby se nám podařilo získat užitečná data.
- U živých subjektů, musíme mít na zřeteli také etiku experimentu.

Randomizovaný experiment

- Pokud ovšem randomizovaný experiment provést mohu, jsem v dobré situaci, protože data z randomizovaného experimentu se nejlépe interpretují.
- Jsou-li subjekty ve studii rozumným výběrem ze sledované populace, mohu odhadnout absolutní rizika i relativní riziko.
- Díky randomizaci se exponovaná skupina patrně nebude lišit od neexponované v žádných charakteristikách subjektů. To je důležité, protože pokud by se skupiny lišily, mohli bychom pozorovat rozdílná rizika ve skupinách z jiných důvodů, než je sledovaná expozice E.

Kohortová studie

- Pokud nemohu ovlivňovat expozici, zvolím observační přístup.
- Vyberu skupinu subjektů, kteří jsou/byli vystaveni E a skupinu těch, kteří nejsou/nebyli vystaveni E a sleduji, zda u nich nastane N . Takovému přístupu se říká *kohortová studie* angl. *cohort study*.
- Nevýhoda oproti randomizovanému experimentu je v tom, že skupiny nevybírám náhodně a mohou se lišit v různých jiných charakteristikách. Potenciálním rozdílem mezi skupinami se pak musím věnovat při analýze a interpretaci výsledků.
- Kohortová studie ovšem umožňuje odhadnout relativní riziko. Vybírali jsme sice podle expozice, takže podíl exponovaných ve studii nemusí být stejný jako v populaci, ale pokud jsou exponovaní i neexponovaní rozumným výběrem z příslušných populací, odhad relativního rizika bude fungovat správně.
- Kohortové studie budou vhodné, když je expozice v populaci vzácná (vyberu počet exponovaných podle potřeby). Nebudou ale vhodné, pokud je vzácná nemoc (rozsah by mohl být příliš velký).

Studie případů a kontrol

- Pokud je nemoc vzácná, je vhodnější zvolit jiný postup. Nejprve identifikuj skupinu nemocných (případy) a ze stejné populace, z jaké jsem vybrala nemocné, vybírám skupinu zdravých (kontroly). Potom zjistím, kolik případů a kolik kontrol bylo vystaveno rizikovému faktoru E . Tyto studie se nazývají *studie případů a kontrol*, angl. *case-control study*.
- Nevýhodou takto vybrané studie je fakt, že podíl nemocných (případů) neodpovídá populaci. Odhadы absolutních a relativního rizika nemoci se nebudou vztahovat k populaci, ze které jsme vybírali, budou špatně.
- Podíl šancí θ však z tabulky odhadneme správně. Uvědomme si, že θ je stejný, pokud tabulku obrátíme (odhadujeme-li šance expozice a ne šance nemoci). A rizika a šance expozice pro nemocné a zdravé dokážeme ve studii případů a kontrol odhadovat správně. V těchto studiích tedy využijeme fakt, že nemoc je vzácná a approximujeme relativní riziko podílem šancí ($\hat{RR} \sim \hat{\theta}$).

Příklady studií

Příklad (Women's Health Initiative)

- Od roku 1993 probíhala v USA Women's Health Initiative. Studie měla několik větví, ale hlavní část se věnovala vlivu hormonální substituční terapie (předpisovaná ženám v přechodu a po přechodu) na srdeční choroby, rakovinu prsu a osteoporózu.
- Bylo randomizováno 27 347 žen (50-70 let) na hormonální terapii a placebo. Jednalo o randomizovaný experiment.
- Očekávalo se mírné zvýšení pravděpodobnosti rakoviny prsu a snížení pravděpodobnosti kardiovaskulárních chorob a osteoporózy.
- Pokus zastaven předčasně (2002) pro špatně vypadající výsledky. Odhad relativních rizik s 95% CI: rak. prsu 1.26 (1.00,1.59), isch. choroba srdeční 1.29 (1.02,1.63), mrtvice 1.41 (1.07,1.85), plic. embolie 2.13 (1.39,3.25), zlomenina krčku 0.66 (0.45,0.98).
- Oproti očekávání se riziko kardiovask. chorob zvýšilo. Pokus změnil pohled na předpisování hormonální substituční terapie.

Příklady studií

Příklad (Statiny a osteoporóza)

- *Statiny jsou léčiva, která se užívají ke snižování hladiny cholesterolu v krvi. Studie na zvířatech ukazovaly, že by mohly také redukovat řídnutí kostí u starších lidí.*
- *Kohortová studie z r. 2000 identifikovala kohortu pacientů, kteří brali statiny a další kohortu pacientů, kteří se léčili také na vysoký cholesterol, ale nebrali statiny a kohortu osob, kteří se neléčili s vysokým cholesterolom (všichni z populace jedné zdravotní pojišťovny). Byli sledováni na výskyt zlomenin nohy po dobu 2 let.*
- *U pacientů se statiny byl odhad abs. rizika zlomeniny během 2 let 0.0020, u pacientů s vys. cholesterolom a jinými léky 0.0018 a u zdravých 0.0035.*
- *Je tedy nejasné, zda statiny skutečně redukují řídnutí kostí, možná ano, ale jiné léky na vys. cholesterol by pak fungovaly podobně. Nebo je to vysoký cholesterol, který je protektivní.*



Příklady studií

Příklad (Antikoncepcie a žilní trombóza)

- Studie na závislost žilní trombózy s hormonální antikoncepcí identifikovala všechny případy žilní trombózy u žen 18-45 let za 2 roky u klientek určité zdravotní pojišťovny a kontroly ve stejné věkové skupině u stejné zdravotní pojišťovny. Pak dohledali, zda ženy užívaly v těchto 2 letech hormonální antikoncepci.
- Jedná se o studii případů a kontrol.

		Horm. antikoncepcie		
		ano	ne	
trombóza	ano	69	84	153
	ne	64	247	311
		133	331	464

- Nemůžeme odhadnout přímo relativní riziko, ale odhad podílu šancí je $\hat{\theta} = 3.17$, 95% CI (2.08, 4.83). Zvýšené riziko trombózy u žen na horm. antikoncepci jsme tedy prokázali.

Příklady studií

- Při zkoumání vztahu antikoncepce a žilní trombózy se studie případů a kontrol nabízí. Žilní trombózy jsou vzácné onemocnění u mladých žen, zatímco hormonální antikoncepce je běžná.
- U statinů a zlomenin to tak jasné není, bylo by patrně možné řešit stejný problém i studií případů a kontrol.
- Studie s tonsilektomií a Hodgkinovým lymfomem byla studie případů a kontrol. Kontroly byli zdraví sourozenci.
- Na rozdíl od randomizovaných studií, oba typy observačních studií jsou velmi citlivé na výběr subjektů. U studie případů a kontrol je to asi nejnáročnější, protože kontroly musí být vybrány náhodně ze stejné populace, ze které pocházejí případy a tuto populaci nemusí být snadné identifikovat. Ale i kohortové studie spoléhají na to, že jsou obě porovnávané kohorty co nejpodobnější, až na expozici.
- Randomizované studie těmito problémy netrpí, randomizace zajistí srovnatelnost skupin. Zato však jsou nejnákladnější a v mnoha případech zcela neproveditelné.

Přednáška 13 (12.5.2020) - obsah

- Modelování alternativních veličin - logistická regrese
- Souvislost a příčinnost, confounding
- Shrnutí učiva

Studie na rakovinu jícnu

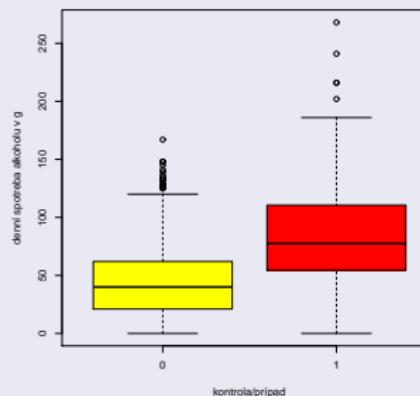
Příklad (Studie Ille-et-Villaine (Tuyns et al. 1977))

- *Bretaň je známá vysokými počty nemocných s rakovinou jícnu. V 70. letech byla provedena v bretaňském okrese Ille-et-Villaine studie případů a kontrol na zkoumání rizikových faktorů souvisejících s touto nemocí.*
- *Případy tvořilo 200 mužů, kteří byli diagnostikováni v místní nemocnici v období leden 1972-duben 1974. Jako kontroly bylo vybráno 775 dospělých mužů ze stejného okresu podle volebních seznamů.*
- *U mužů byla sebrána hlavně data týkající se konzumace alkoholu a kouření, což jsou potenciální rizikové faktory pro rakovinu jícnu.*
- *Začneme tím, že se budeme zajímat o souvislost rakoviny jícnu s průměrnou denní spotřebou alkoholu.*

Studie na rakovinu jícnu

Příklad (Studie Ille-et-Villaine)

Začneme tím, co už umíme. Mohli bychom se podívat, jak se liší rozdělení denní spotřeby alkoholu mezi případy a kontrolami:



Průměr mezi kontrolami je 44.43 g, mezi případy 85.08 g, medián mezi kontrolami 40.0 g, mezi případy 77.5 g. Zdá se, že se rozdělení spotřeby alkoholu značně liší.

Studie na rakovinu jícnu

Příklad (Studie Ille-et-Villaine)

Střední hodnoty spotřeby alkoholu samozřejmě nejsou stejné, Welchův *t-test* velmi jasně zamítá (*p-hod.* < 0.0001). Nás však spíš zajímá, jak souvisí spotřeba alkoholu s pravděpodobností, že muž dostane rakovinu jícnu. Na to nám naše analýza zatím odpověď nedává.
 Mohli bychom kategorizovat denní spotřebu alkoholu. Nabízí se kategorie: $\leq 40g$ a $> 40g$ ($40g$ odpovídá dvěma nápojům - "málo rizikové pití").
 Potom dostaneme čtyřpolní tabulku:

		spotř. alkoholu		
		> 40g	$\leq 40g$	
rakovina jícnu	ano	171	29	200
	ne	375	400	775
		546	429	975

Podíl šancí vychází $\hat{\theta} = \frac{171 \cdot 400}{29 \cdot 375} = 6.29$, 95%CI: (4.14, 9.55). Prokázali jsme tedy velmi silnou závislost rakoviny jícnu na tom, zda muž pije více než 40g alkoholu denně.

Studie na rakovinu jícnu

Příklad (Studie Ille-et-Villaine)

- Tímto způsobem jsme sice dostali odhad podílu šancí, ale pouze pro jednu konkrétní možnost kategorizace. Tím jsme velkou část informace ztratili. Bylo by dobré vědět, jak se mění šance onemocnění, bude-li se měnit spotřeba alkoholu.
- Kontroly ve studii neměly stejné věkové rozložení jako případy. Věkový průměr případů byl 60.04 roku, zatímco kontrol 50.21 roku. Očekáváme, že riziko onemocnění bude souviset s věkem. Rádi bychom nějak odstranili vliv rozdílu věku mezi skupinami.
- Ve druhé fázi bychom se rádi zabývali vlivem dalšího rizikového faktoru - kouření. Rádi bychom zkoumali vliv obou rizikových faktorů najednou. Je aditivní nebo spolu oba rizikové faktory interagují?
- Tyto otázky by se nám podařilo zodpovědět, kdybychom dokázali sestavit model, kde vysvětlovaná proměnná bude alternativní (případ/kontrola, tj. nemocný/zdravý).



Model logistické regrese

- Předpokládejme, že máme veličiny Y_1, Y_2, \dots, Y_n , každá z nich má alternativní rozdělení s pravděpodobností úspěchu $\pi_1, \pi_2, \dots, \pi_n$. Chtěli bychom sestavit model, který by vysvětloval $EY_i = \pi_i$ pomocí jednoho nebo více regresorů, podobně jako v lineární regresi vysvětlujeme stř. hodnotu normálně rozdělených veličin.
- Není rozumné snažit se modelovat přímo π_i pomocí lineární funkce regresorů. Víme, že π_i je číslo z intervalu $\langle 0, 1 \rangle$, zatímco lineární funkce je neomezená. Mohli bychom tedy dostat odhady, které jsou mimo interval $\langle 0, 1 \rangle$.
- Ukazuje se, že je vhodné modelovat logaritmus šancí. Model *logistické regrese* pro 1 regresor pak vypadá takto:

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 x_i,$$

kde \log je přirozený logaritmus.

Model logistické regrese

- Funkci

$$f(\pi) = \log\left(\frac{\pi}{1 - \pi}\right)$$

(logaritmus podílu šancí) se říká logit.

- Kdybychom chtěli vyjádřit pravděpodobnost úspěchu π z rovnice modelu jako funkci prediktoru x , zjistíme, že

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

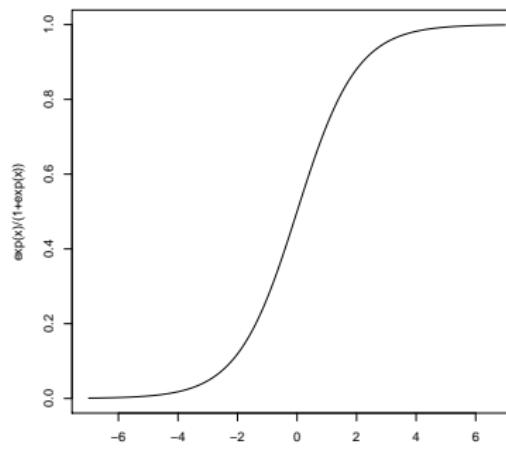
- Definiční obor této funkce je R (regresor může nabývat jakékoli reálné hodnoty), ale obor hodnot je interval $(0, 1)$. Pro libovolnou reálnou hodnotu regresoru tedy dostanu $\pi(x) \in (0, 1)$.

Model logistické regrese

Závislost $\pi(x)$ na regresoru pro $\beta_0 = 0$ a $\beta_1 = 1$ ukazuje graf.

Pokud vám připomíná distribuční funkci, pak se díváte správně. Je to distribuční funkce rozdělení, které nazýváme logistické.

Obrázek: π jako funkce regresoru



Význam regresních koeficientů

- Z modelu odvodíme význam regresního koeficientu β_1 . Šance úspěchu splňuje:

$$\frac{\pi(x)}{1 - \pi(x)} = e^{\beta_0 + \beta_1 x}$$

- Pokud se regresor x zvýší o 1, šance úspěchu je

$$\frac{\pi(x+1)}{1 - \pi(x+1)} = e^{\beta_0 + \beta_1(x+1)} = e^{\beta_1} \cdot e^{\beta_0 + \beta_1 x}$$

- e^{β_1} je tedy podíl šancí úspěchu při hodnotách regresoru $x+1$ a x . Je vlastností modelu, že bude tento podíl šancí je stejný pro všechna x .
- Pokud x může nabývat hodnoty 0, pak je e^{β_0} šance úspěchu pro $x = 0$.

Koeficienty regresního modelu

- Odhadování koeficientů logistické regrese se provádí numericky pomocí metody maximální věrohodnosti.
- Získané odhady regresních koeficientů b_0 a b_1 můžeme použít k testování hypotéz o regresních koeficientech.
- Hypotéza $H_0: \beta_1 = 0$ je hypotéza nezávislosti. Pokud platí H_0 , pak šance úspěchu nezávisí na x . Test je obdobou t-testu v normální regresi, testová statistika $Z = \frac{b_1}{S.E.(b_1)}$ má za H_0 přibližně rozdělení $N(0, 1)$.
- Hypotéza $H_0: \beta_0 = 0$ obvykle není zajímavá, znamenala by, že šance úspěchu pro $x = 0$ je 1.

Odbočka: Metoda maximální věrohodnosti

- Metoda maximální věrohodnosti je jedna z nejužívanějších odhadovacích metod ve statistice.
- Chceme odhadnout parametr θ nějakého rozdělení. θ může být střední hodnota, nebo rozptyl, nebo cokoliv, na čem hustota rozdělení nějak závisí.
- Hustotu zkoumaného rozdělení tedy můžeme vyjádřit jako $f(x; \theta)$.
- Máme náhodný výběr X_1, X_2, \dots, X_n z tohoto rozdělení.
- Veličiny ve výběru jsou nezávislé, takže sdružená hustota bude součin hustot:

$$f(x_1, \dots, x_n; \theta) = f(x_1; \theta)f(x_2; \theta)\dots f(x_n; \theta)$$

- Budeme hledat takové θ , které tuto sdruženou hustotu maximalizuje, přičemž hodnoty x_1, \dots, x_n budou pevné, rovné hodnotám získaným z výběru.

Odbočka: Metoda maximální věrohodnosti

- Na sdruženou hustotu se tedy díváme jako na funkci θ . Hodnoty x_1, \dots, x_n získané z výběru považujeme za pevné. Sdruženou hustotu $L(\theta; x_1, \dots, x_n)$ jako funkci parametru θ nazveme **věrohodnostní funkcí**, angl. *likelihood (function)*.
- Pokoušíme se tedy pro dané hodnoty x_1, \dots, x_n hledat θ tak, aby hodnota sdružené hustoty v x_1, \dots, x_n byla co nejvyšší.
- Pokud je rozdělení diskrétní, hustotě odpovídají pravděpodobnosti bodů x_1, \dots, x_n .
- Často se maximalizuje spíš $\log(L(\theta; x_1, \dots, x_n)) = \ell(\theta; x_1, \dots, x_n)$, protože součiny hustot převádí na součty jejich logaritmů. Logaritmus je monotónní funkce, takže θ , které maximalizuje $\log(L(\theta))$ bude maximalizovat i $L(\theta)$.

Odbočka: Metoda maximální věrohodnosti

Příklad (Odhad pravděpodobnosti jevu metodou max. věrohodnosti)

Chceme odhadnout pravděpodobnost π nějakého jevu.

Máme pouze 3 pozorování alternativního rozdělení $X \sim Alt(\pi)$ (1/0 jev nastal/nenastal): $x_1 = 1, x_2 = 0, x_3 = 1$

Pro alternativní rozdělení platí: $P(X = 1; \pi) = \pi$ a

$P(X = 0; \pi) = 1 - \pi$

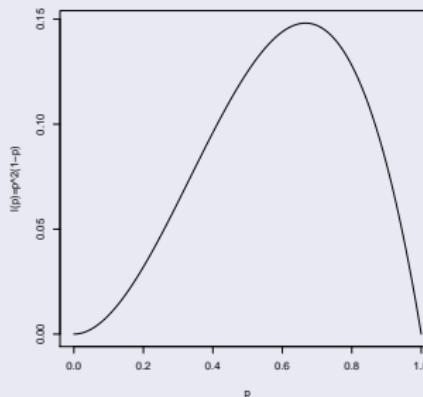
Věrohodnostní funkce tedy bude

$$L(\pi; x_1, x_2, x_3) = \pi \cdot (1 - \pi) \cdot \pi = \pi^2(1 - \pi)$$

Odbočka: Metoda maximální věrohodnosti

Příklad (Odhad pravděpodobnosti jevu metodou max. věrohodnosti)

Na intervalu $(0, 1)$ (přípustné hodnoty π) vypadá věrohodnostní funkce:



Maxima nabývá v bodě $\pi = 2/3$ (ověřte analyticky, stačí zderivovat).

Maximálně věrohodný odhad bude tedy $\hat{\pi}_{ML} = 2/3$. Vidíme, že se velmi rozumně rovná relativní četnosti.

Model logistické regrese

- Věrohodnostní funkce v případě odhadu parametrů logistické regrese je složitější, neodhadujeme přímo π , ale parametry β_0 a β_1 .
- V bodě (b_0, b_1) nabývá logaritmus věrohodnostní funkce svého maxima, $\ell(b_0, b_1; y_1, \dots, y_n)$ je tedy hodnota tohoto maxima.
- Pro hodnocení modelu se používá *deviance*, což je dvojnásobek rozdílu logaritmu věrohodnosti maximálního možného modelu (stejně parametrů jako pozorování) a zkoumaného modelu.

$$D = 2(\ell(y_1, \dots, y_n; y_1, \dots, y_n) - \ell(b_0, b_1; y_1, \dots, y_n))$$

- Na devianci se můžeme dívat jako na obdobu reziduálního součtu čtverců u lineárních modelů.
- Rozdíl deviancí nulového modelu (pouze intercept) a zkoumaného modelu má χ^2_1 , pokud regresor nesouvisí s vysvětlovanou proměnnou. Lze ho tedy použít k testování celého modelu podobně jako se používá F -test v lineárních modelech.

Model logistické regrese

- Logistickou regresi lze samozřejmě použít i v případě k regresorů.
Model má pak tvar:

$$\log \left(\frac{\pi_i}{1 - \pi_i} \right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki},$$

Rozdíl deviance modelu od modelu nulového pak bude mít χ^2_k rozdělení, pokud žádný z regresorů nesouvisí s pravděpodobností úspěchu.

- Logistická regrese patří do rodiny zobecněných lineárních modelů, které modelují veličiny s jiným než normálním rozdělením.
- Posouzení, zda model funguje na datech rozumně, je obtížnější než u normálních lineárních modelů. Rezidua se nedají jednoduše použít k diagnostice různých potíží s modelem, u logistické regrese nemají jasnou interpretaci ani nepocházejí z jednoho rozdělení.

Studie na rakovinu jícnu

Příklad (Studie Ille-et-Villaine)

Zkusme použít logistickou regresi ke zkoumání závislosti rakoviny jícnu na denní spotřebě alkoholu. Odhadnuté parametry:

	odhad	S.E.(b)	Z	p-hod.
b_0	-2.9689	0.1808	-18.42	$< 2 \cdot 10^{-16}$ ***
b_1 (alk v g)	0.0260	0.0023	11.22	$< 2 \cdot 10^{-16}$ ***

Prokázali jsme tedy souvislost denní spotřeby alkoholu s pravděpodobností rakoviny jícnu. Pokud se zvýší spotřeba alkoholu o 1 g, odhadujeme, že šance rakoviny je $e^{0.0260} = 1.0264$ krát vyšší.

Zajímavější by asi bylo spočítat, kolikrát se zvýší šance při zvýšení spotřeby alkoholu o 20g denně (jeden alk. nápoj). Odhadujeme, že v takovém případě je šance rakoviny $e^{20 \cdot 0.0260} = 1.6825$ krát vyšší.

Studie na rakovinu jícnu

Příklad (Studie Ille-et-Villaine)

Vzhledem k tomu, že očekáváme vliv věku na pravděpodobnost rakoviny, zkusíme vložit do modelu také věk.

	odhad	S.E.(b)	Z	p-hod.
b_0	-6.7885	0.5329	-12.74	$< 2 \cdot 10^{-16}$ ***
b_1 (alk v g)	0.0276	0.0025	11.15	$< 2 \cdot 10^{-16}$ ***
b_2 (věk)	0.0666	0.0079	8.49	$< 2 \cdot 10^{-16}$ ***

Vidíme, že věk souvisí s pravděpodobností rakoviny jícnu také. Muž starší o 1 rok (se stejnou spotřebou alkoholu) má šanci rakoviny $e^{0.0666} = 1.07$ krát vyšší.

Spotřeba alkoholu je i za přítomnosti věku významným prediktorem. Koefficient se nepatrně zvýšil. Muž stejného věku, který konzumuje o jeden nápoj denně více (20g), má šanci rakoviny $e^{0.0276 \cdot 20} = 1.74$ krát vyšší.



Studie na rakovinu jícnu

Příklad (Studie Ille-et-Villaine)

Chceme zjistit, jaký je vliv kouření na rakovinu jícnu. Tabulka ukazuje počty kuřáků a nekuřáků ve skupinách:

		kouření		
		ano	ne	
rakovina jícnu	ano	191	9	200
	ne	520	255	775
		711	264	975

Odhad podílu šancí pro kuřáky vs. nekuřáky je

$\hat{\theta} = \frac{191 \cdot 255}{9 \cdot 520} = 10.41$, tedy skutečně vysoké číslo. Naše tabulka však nebude v úvahu věk a spotřebu alkoholu. Uvidíme, zda tento vysoký vliv kouření potvrdíme i v modelu.

Studie na rakovinu jícnu

Příklad (Studie Ille-et-Villaine)

Model s kouřením:

	odhad	S.E.(b)	Z	p-hod.
b_0	-8.2265	0.6514	-12.63	$< 2 \cdot 10^{-16}$ ***
b_1 (alk v g)	0.0260	0.0025	10.22	$< 2 \cdot 10^{-16}$ ***
b_2 (věk)	0.0646	0.0082	7.90	$2.8 \cdot 10^{-15}$ ***
b_3 (kuřák)	1.9277	0.3663	5.26	$1.4 \cdot 10^{-7}$ ***

I za přítomnosti věku a spotřeby alkoholu bylo kouření významným prediktorem. Kuřák bude mít šanci, že onemocní rakovinou jícnu odhadem $e^{1.9277} = 6.8$ krát vyšší než stejně starý nekuřák se stejnou spotřebou alkoholu.

Zkusili jsme ještě vložit do modelu interakci alkoholu s kouřením. (Má alkohol jiný vliv na rakovinu jícnu u kuřáků než u nekuřáků?) Interakce ovšem významná nebyla ($p=0.1096$). Zdá se tedy, že vliv alkoholu se neliší pro kuřáky a nekuřáky.



Logistická regrese

- Na příkladu jsme viděli, že logistická regrese umožňuje zkoumat vliv spojitéch rizikových faktorů na šanci nemoci.
- Model přirozeně odhaduje podíly šancí, což je zvláště u studií případů a kontrol, které neumožňují odhad relativního rizika, velmi vhodné.
- Můžeme zkoumat také více rizikových faktorů najednou a případné interakce mezi nimi.
- Logistická regrese má samozřejmě mnohem širší využití než jenom v epidemiologii. Je vhodná, pokud chceme popsat nebo testovat souvislosti nějakých veličin s výskytem určitého jevu.

Souvislost a příčinnost

- Statistické testy nám umožňují prokázat (na určité hladině) souvislost dvou veličin. To, že je nějaká souvislost průkazná ovšem nevypovídá nic o příčinnosti (kauzalitě) mezi souvisejícími veličinami.
- Prokázat kauzální spojení mezi dvěma veličinami je mnohem složitější, než jenom prokázat souvislost.
- Př.: Chci zkoumat, zda pravidelné návštěvy zubního lékaře působí jako prevence zubních kazů. Udělám náhodný výběr z nějaké populace a zjistím pro každého člověka, kolikrát byl za posledních 5 let u zubaře a kolik má zubních kazů. "Překvapivě" zjistím, že počet zubních kazů koreluje s počtem návštěv zubaře pozitivně. Čím více lidé chodí k zubaři, tím více mají zubních kazů. Zde je chyba v úvaze odhalitelná snadno. Zaměnili jsme příčinu a důsledek. Tato asociace je příčinná, ale patrně spíš opačným směrem. V tomto případě by asi pomohlo věnovat více pozornosti časové posloupnosti.

Souvislost a příčinnost

- Př. Zkoumáme různé veličiny, které by mohly souviseť s tím, že člověk onemocní rakovinou plic. Zjistíme velmi silnou asociaci s nošením zapalovače v kapse. V tomto případě časová posloupnost sedí. Přesto se náš rozum vzpírá tomu, že by zákaz nošení zapalovačů po kapsách mohl ochránit lidi před rakovinou.
- V tomto případě jsme narazili na problém, kterému se anglicky říká *confounding*, česky (nepříliš povedeně) *matení*. Jde o to, že existuje nějaká další veličina (*confounder*, pěkný český výraz neexistuje, někdy se používá *zavádějící faktor*), která je kauzálně spojena s oběma veličinami, jejichž asociaci zkoumám.
- V příkladu s rakovinou plic jde samozřejmě o kouření. Kouření způsobuje rakovinu plic a také způsobuje, že člověk má v kapse zapalovač. Proto pozorujeme asociaci mezi zapalovačem a rakovinou plic a kauzálního v ní není vůbec nic.
- Bohužel, v praxi je někdy těžké odhalit takové zavádějící veličiny a některé asociace mohou být chybně označeny za kauzální.

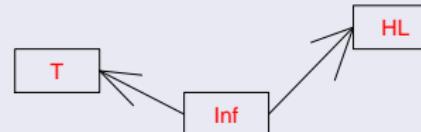
Souvislost a příčinnost

Příklad (Tonsilektomie a Hodgkinův lymfom)

Případ tonsilektomie a Hodgkinova lymfomu může být také ovlivněn zavádějícím faktorem. Některé studie prokázaly souvislost mezi odstraněním mandlí a Hodgkinovým lymfomem. Časová posloupnost je v pořádku, mohlo by existovat kauzální spojení:



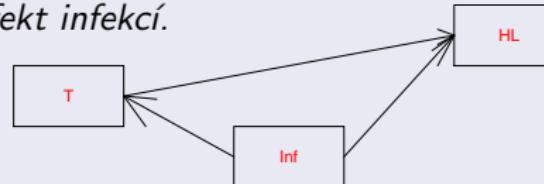
Je ale možné, že ve skutečnosti jsou příčinou Hodgkinova lymfomu opakované infekce dýchacích cest, které také vedou vedou k tonsilektomii. Mezi tonsilektomií a Hodgkinovým lymfomem možná příčinnost není žádná.



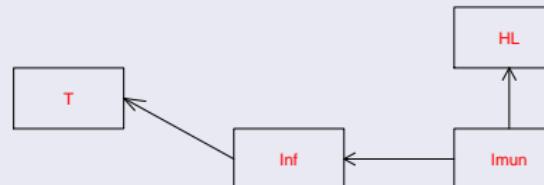
Souvislost a příčinnost

Příklad (Tonsilektomie a Hodgkinův lymfom)

Vzhledem k tomu, že mandle jsou imunitní orgán, umíme si představit, že přeče jen odstranění mandlí samo o sobě může mít vliv na vývoj Hodgkinova lymfomu. Možná tedy část pozorovaného efektu skutečně způsobí tonsilektomie, ale efekt, který pozorujeme je vyšší, protože pozorujeme ještě efekt infekcí.



Nebo je všechno ještě složitější a hlavní příčinou Hodgkinova lymfomu je imunitní selhání, které také způsobí opakované infekce a následně tonsilektomii, tedy "o úroveň vyšší" zavádějící faktor.



Souvislost a příčinnost

- Z předchozích příkladů vidíme, že prokázat příčinnou souvislost může být obtížné.
- Nejspolehlivější nástroj na prokazování kauzality je randomizovaný pokus, kde můžeme manipulovat jednou proměnnou a náhodně přiřazovat subjekty na různé úrovňě této proměnné a pak pozorovat, jaký vliv to má na druhou proměnnou.
- Spoustu asociací však takto zkoumat nelze. Nemůžeme lidem nařídit aby kouřili nebo aby si nechali odstranit mandle. Potom si musíme vystačit s pozorovacími studiemi, s jejichž interpretací musíme být opatrní.
- Mohou pomoci modely, ve kterých jako regresory použijeme potenciální zavádějící faktory a zjistíme, jestli zkoumaná příčina stále ovlivňuje odpověď i za přítomnosti těchto regresorů.

Co svědčí ve prospěch kauzální souvislosti v pozorovací studii

- Pokud má být příčinná souvislost mezi dvěma veličinami, tak samozřejmě příčina musí předcházet důsledek. Musíme tedy dávat pozor na časovou posloupnost.
- Vysvětlení mechanismu. Pokud nemáme nejmenší tušení, jak by mohla jedna veličina druhou ovlivňovat, je to podezřelé.
- Souvisí vyšší hodnoty příčiny s vyšší pravděpodobností důsledku? Je dobré ověřit, pokud to dává smysl. Asi nelze zkoumat, zda více vytržených mandlí souvisí s vyšší pravděpodobností Hodgkinova lymfomu.
- Absence potenciálního zavádějícího faktoru, i když to někdy bývá těžké zjistit.
- Dostatečně silná asociace. Slabší souvislosti mohou být dílem zavádějících faktorů.

Přehled probraných metod

Tabulka: Testy o poloze výběrů

situace	normální výběry	nenormální výb, n nízké
1 výběr	jednovýběrový t -test	znaménkový nebo Wilcoxonův párový
2 výběry nezávislé	dvouvýběrový t -test	Wilcoxonův dvouvýběrový (Kolmogorovův-Smirnovův)
2 výběry párové závislosti	párový t -test	znaménkový nebo Wilcoxonův párový
> 2 výběry nezávislé	jednoduché třídění	Kruskalův- Wallisův test
> 2 výběry skup. závislosti	náhodné bloky	Friedmanův test

Přehled probraných metod

Tabulka: Zkoumání závislostí

veličiny	metoda
spojitá/spojitá	korelační test (Pearson/Spearman) lineární regrese
spojitá/kategorická	analýza rozptylu
spojitá/nula-jedničková	douvýb. t-test
kategorická/kategorická	χ^2 test nezávislosti Fisherův test
nula-jedničková/nula-jedničková	χ^2 test nezávislosti Fisherův test podíl šancí
nula-jedničková/spoj. nebo kat.	logistická regrese

Doporučení pro život

- Přemýšlejte o statistické analýze již při plánování experimentu (sběru dat). Uvažujte, jaká data budete potřebovat. Uvažujte o síle testu na alternativu, kterou chcete prokázat. To vám dá představu o potřebném rozsahu výběru.
- Před volbou statistické metody si rozmyslete, na jakou otázku hledáte odpověď'.
- Pokud to jde, kreslete grafy, obrázky, diagramy, teprve pak se pouštějte do testování a modelování.
- Pokud je to možné, s výsledkem testu vždy uved'te interval spolehlivosti pro testovaný parametr.
- A když si nejste jisti, zeptejte se statistika.

Konec přednášky