Statistika a spolehlivost v lékařství (A6M33SSL)

Osnova statistických cvičení s řešenými příklady

URČENO VÝHRADNĚ PRO STUDENTY V LS 2014/2015
verze 2015-03-31-1

T.Sieger

Obsah

Stručný přehled značení

X	náhodná veličina (zpravidla velké písmeno latinkou)
x	realizace náhodné veličiny (zpravidla malé písmeno latinkou)
$\chi_d f^2$	χ^2 rozdělení s df stupni volnosti
Φ	distribuční funkce normovaného normálního rozdělení
Φ^{-1}	kvantilová funkce normovaného normálního rozdělení
$Bi(n,p)$	binomické rozdělení popisující počet úspěchů v sérii nezávislých n pokusů s elementární pravděpodobností úspěchu p
$DXvarX$	rozptyl náhodné veličiny X
EX	střední hodnota náhodné veličiny X
F_X	distribuční funkce náhodné veličiny X
f_X	hustota pravděpodobnosti náhodné veličiny X
F_X^{-1}	kvantilová funkce náhodné veličiny X
$N(0,1)$	normované normální rozdělení
$N(\mu,\sigma^2)$	normální rozdělení se střední hodnotou μ a rozptylem σ^2
t_{df}	Studentovo t-rozdělení s df stupni volnosti

1 Opakování: pravděpodobnost

Vysvětlete termíny:

1.1 náhodný jev (elementární jev, prostor elementárních jevů)

Příklad: Házení kostkou.

1.2 pravděpodobnost

(funkce definovaná na podmnožinách prostoru elementárních jevů)

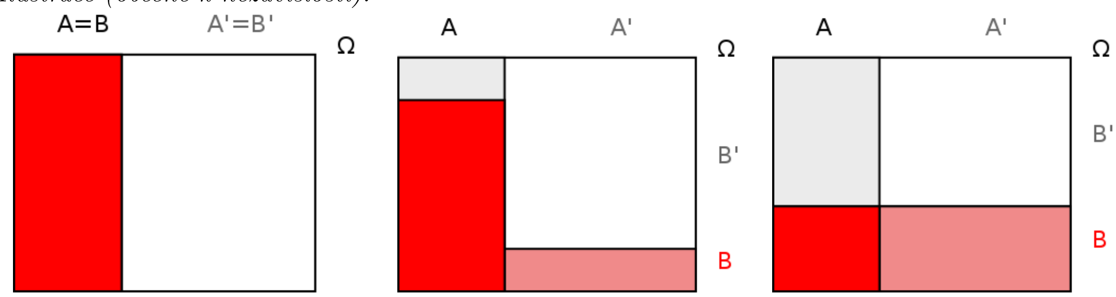
1.3 ⋆pravděpodobnostní prostor: (Ω, A, p)

1.4 nezávislost náhodných jevů

Příklad: Házení kostkou: jev "padne liché číslo" vs. jev "padne číslo větší než 3". Jsou tyto jevy nezávislé?

Nechat studenty nejprve hádat, zda jevy A a B jsou nezávislé. Pokud to nikdo jasně nevysvětlí, nechat je pokusit se definovat nezávislost. V případě neznalosti vyvracet očividně nepravdivé definice (vedoucí k nesmyslným závěrům) a snažit se, aby se se sami dobrali k tomu, co nezávislost znamená: že znalost nastálého jevu A neovlivňuje pravděpodobnost jevu B.

Ilustrace (obecně k nezávislosti):



identické jevy ("100%" závislé) ("částečně") závislé nezávislé

Lze doufat, že dojdou buď k tomu, že pravděpodobnost, že nastame jev A i B současně, je dána v geometrickém pojetí pravděpodobnosti jako určitá poměrná část plochy (prostoru elementárních jevů) odpovídající jevu A - poměrná v takevém poměru, jaký odpovídá pravděpodobnosti jevu B v rámci celého prostoru. To přímo znamená, že $p(A \cap B) = p(A)p(B)$. Intepretovat to lze tak, že rozhodovat o tom, zda nastane jev A resp. B je možno tak, že hodíme dvěma mincemi s pravděpodobnostmi, že padne panna $p(A)$ resp. $p(B)$. Mincemi hodíme nezávisle na sobě. Pravděpodobnost, že nastane A i B je dána prostě vynásobením pravděpodobností $p(A)$ a $p(B)$.

Alternativně je možné, že dojdou k tomu, že nezávislost znamená, že pravděpodobnost, že nastane jev B není ovlivňena tím, zda zároveň nastane jev A, tedy že $p(B) = p(B|A) = p(B|\bar{A})$. Z tohoto přes definici podmíněné pravděpodobnosti dostaneme "klasickou" definici nezávislosti.

$$p(B) = p(B|A) = \frac{p(B \cap A)}{p(A)} \qquad \text{(za podmínky } p(B) > 0) \tag{1}$$

a po vynásobení $p(A)$ dostáváme:

$$p(B)p(A) = p(B \cap A) \tag{2}$$

1.5 podmíněná pravděpodobnost

Ukázat přes geometrickou představu 2 množin A a B s neprázdným průnikem.

$$p(B) = p(B|A) = \frac{p(B \cap A)}{p(A)} \qquad \text{(za podmínky } p(B) > 0\text{)} \tag{3}$$

Příklad: Tenista má první podání úspěšné s pravděpodobností 0,6; druhé s pravděpodobností 0,8. S jakou pravděpodobností se dopustí dvojchyby?

Řešení: jev N_1 - chyba v prvním podání, jev N_2 - chyba ve druhém podání. $p(N_1) = 0,4$, $p(N_2|N_1) = 0,2$ (pozor, toto není $p(N_2)$!). $P(N_1 \cap N_2) = p(N_2|N_1)p(N_1) = 0,4 \cdot 0,2 = 0,08$.

1.6 úplná pravděpodobnost

aneb celková pravděpodobnost jevu A se dá "spočítat po kouskách a posčítat"

$$P(A) = \sum_i P(A \cap B_i) = \sum_i P(A|B_i)P(B_i)$$

Příklad: Spočtěte průměrné zastoupení studentek ve škole, která má 2 posluchárny a 2 šatny a v každé místnosti je jiné zastoupení dívek. (Předpokládáme, že nelze průměrné zastoupení studentek nelze zjistit přímo, např. anketou.)

Řešení: Jev D - náhodně vybraný(á) student(ka) je dívka. Jev M_i - student(ka) se nachází v místnosti i.

$$p(D) = \sum_{i=1}^{4} p(D \cap M_i) = \sum_{i=1}^{4} p(D|M_i)p(M_i) \tag{4}$$

1.7 Bayesův vzorec

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

aneb výpočet posteriorních podmíněných pravděpodobností z apriorních podmíněných pravděpodobností.

Označuje-li např. $P(B|A)$ podmíněnou pravděpodobnost určitého výsledku klinického testu (jev B) u zdravého člověka (podmínka A), pak Bayesův vzorec dává na základě výsledku testu (a apriorních pravděpodobností) podmíněnou pravděpodobnost $P(A|B)$ toho, že testovaná osoba je zdravá či nemocná.

Příklad: V obléhaném městě nepřítel kontaminoval jedem 10% vodovodů. Obyvatelé v postižených oblastech jsou v 90% otrávení (krom těch, kteří se vody zatím nenapili). Ve zbylých částech města je otrávených pouze 10% obyvatel (patrně z jiných zdrojů, než z vody). Jednomu obyvateli se z města podaří tajnou chodbou uniknout, avšak záhy umírá – byl otráven. Dá se odhadnout, zda tajná chodba vede do otrávené, nebo neotrávené části města? Jak se tato pravděpodobnost změní, vyjde-li z chodby další otrávený?

Řešení:

Jev, že chodba vede do jedem otrávené části města, označíme jako J a jev, že člověk přicházející chodbou je otrávený a zemře, jako M.

Za apriorní (předem očekávanou) pravděpodobnost toho, že chodba vede do otrávené části města, musíme (bez znalosti dodatečných informací) vzít poměr plochy otrávených částí města vůči celému městu, tj. 10%, tedy $p(J) = 0,1$. Dále víme, že $p(M|J) = 0,9$, $p(M|\bar{J}) = 0,1$. Poté, co chodbou přijde otrávený člověk, se pravděpodobnost, že chodba vede do otrávené části města, změní na $p(J|M)$:

$$p(J|M) \overset{Bayes}{=} \frac{p(M|J)p(J)}{p(M)} = \frac{p(M|J)p(J)}{\underbrace{p(M|J)p(J) + p(M|\bar{J})p(\bar{J})}_{\text{úplná pravděpodobnost}}} = \frac{0,9 \cdot 0,1}{0,9 \cdot 0,1 + 0,1 \cdot 0,9} = \frac{1}{2} \tag{5}$$

Pravděpodobnopst se zvýšila z apriorní pravděpodobnosti $0,1$ na posteriorní pravděpodobnosti $0,5$. Nyní však nemáme vůbec jasno, zda chodba vede do otrávené či neotrávené části města.

Pokud z chodby vystoupí další člověk a také zemře, změní se pravděpodobnost, že chodba vede do otrávené části města, dále na:

$$p(J|M_2, M_1) = \frac{p(M_2|J, M_1)p(J|M_1)}{p(M_2|J, M_1)p(J|M_1) + p(M_2|\bar{J}, M_1)p(\bar{J}|M_1)}. \tag{6}$$

Za předpokladu, že pravděpodobnost výskytu otráveného člověka v otrávené oblasti (resp. mimo otrávenou oblast) se po příchodu prvního otráveného nezmění (tj. pokud je ve městě hodně obyvatel a odchodem jednoho otráveného se pravděpodobnosti znatelně nezmění), platí $p(M_2|J, M_1) = p(M_2|J)$ a $p(M_2|\bar{J}, M_1) = p(M_2|\bar{J})$ a tedy

$$p(J|M_2, M_1) = \frac{p(M_2|J)p(J|M_1)}{p(M_2|J)p(J|M_1) + p(M|\bar{J})p(\bar{J}|M_1)} = \frac{0,9 \cdot 0,5}{0,9 \cdot 0,5 + 0,1 \cdot 0,5} = 0,9 \tag{7}$$

Pro srovnání: v případě, že by chodbou přišli (a vzápětí zemřeli) dva lidé nezávisle na sobě, můžeme položit

$$p(M_{1+2}|J) = p(M|J) \cdot p(M|J) = p(M|J)^2 \tag{8}$$

a

$$p(M_{1+2}|\bar{J}) = p(M|\bar{J}) \cdot p(M|\bar{J}) = p(M|\bar{J})^2 \tag{9}$$

a proto

$$\begin{aligned}
p(J|M_{1+2}) &= \frac{p(M_{1+2}|J)p(J)}{p(M_{1+2}|J)p(J) + p(M_{1+2}|\bar{J})p(\bar{J})} && (10)\\
&= \frac{p(M|J)^2 p(J)}{p(M|J)^2 p(J) + p(M|\bar{J})^2 p(\bar{J})} && (11)\\
&= \frac{0,9^2 \cdot 0,1}{0,9^2 \cdot 0,1 + 0,1^2 \cdot 0,9} && (12)\\
&= \frac{0,9^2}{0,9^2 + 0,1 \cdot 0,9} && (13)\\
&= \frac{0,9}{0,9 + 0,1} && (14)\\
&= 0,9 && (15)
\end{aligned}$$

1.8 náhodná veličina

(funkce definovaná na prostoru elementárních jevů)

★ měřitelná funkce $X : \Omega \to \mathbb{R}$ splňující $X^{-1}(I) = \{\omega \in \Omega : X(\omega)\} \in I$ pro každý interval I, tj. obrazem intervalů je množina ze σ-algebry A (systému podmnožin prostoru elementárních jevů Ω)

1.9 realizace náhodné veličiny

ilustrace: 3 světy (reálný, zjednodušený reálný (modelový), "statistický")

Příklad: Hmotnost studentů určitého kurzu – v modelovém a "statistickém" světě graficky vyznačte jednu konkrétní realizaci náhodné veličiny "hmotnost" a samotnou náhodnou veličinu "hmotnost".

Nechat studenty vyznačit na reálné ose jedno pozorování - hmotnost a pak to samé chtít pro náhodnou veličinu, aby byli konfrontováni s tím, že vlastně nějakou hodnotu nakreslit nelze, protože náhodná veličina je náhodná. Chová se náhodně (paralela s kvantovým světem a štěrbinovým experimentem). Jediné, co o ní víme, je jak se chová "typicky", "průměrně". Dojít k tomu, že někde je větší pravděpodobnost výskytu, jinde menší.

1.10 rozdělení pravděpodobnosti náhodné veličiny

Příklad: Definujte rozdělení pravděpodobnosti náhodné veličiny.
U spojitého rozdělení $p(X = x) = 0$, *proto používáme* $p(X \leq x)$.

1.11 distribuční funkce $F_X(x)$ náhodné veličiny X

$F_X(x) \stackrel{def.}{=} (p(X \leq x), x \in \mathbb{R}$

1.12 hustota pravděpodobnosti f_X náhodné veličiny X

V případě jakých rozdělení o ní hovoříme?
hustotu pravděpodobnosti definujeme u spojitých rozdělení.
U spojitého rozdělení $F_X(x) = \int_{-\infty}^{x} f_X(t)dt$.

Příklad: Určuje $f_X(x) = sin(x)$ pro $0 \leq x \leq \pi$ hustotu pravděpodobnosti nějaké náhodné veličiny?

Řešení: Ne, $\int_0^\pi sin(t)dt = [-cos(t)]_0^\pi = 1 - (-1) = 2$, tedy hustota se neintegruje do 1 a nejedná se o hustotu.

Příklad: Určuje $f_X(x) = 2 - x$ pro $0 \leq x \leq 2 + \sqrt{2}$ hustotu pravděpodobnosti nějaké náhodné veličiny?

Řešení: Ne, $\int_0^{2+\sqrt{2}} f_X(x)dx = 1$, ale $f_X(2 + \sqrt{2}) < 0$.

Vlastnosti distribuční funkce:

- $lim_{x \to -\infty} = 0$ *protože "*$p(X \leq -\infty) = 0$*"*
- $lim_{x \to -\infty} = 1$ *protože "*$p(X \leq \infty) = 1$*"*

1.13 kvantilová funkce F_X^{-1}

Pro spojité ryze rostoucí inverzní funkce k F_X.

Příklad: Je dáno $p(X \leq x) = x$ pro $0 \leq X \leq 1$. Odvoďte a načtrtněte distribuční funkci F_x, hustotu f_x a kvantilovou funkci F_X^{-1}.

Řešení: Distribuční funkce $F_X(x)$ je právě $p(X \leq x) = x$. Hustota $f_X(x)$ je derivací $F_X(x)$, tedy $f_X(x) = 1$. Kvantilová funkce je inverzí k distribuční funkci, tedy $F_X^{-1}(u) = u$ (má smysl samozřejmě jen pro $0 \leq u \leq 1$).

Příklad: Je dáno $f_x(x) = 3x$ pro $0 \leq X \leq 1$. Odvoďte a načtrtněte distribuční funkci F_x, $p(X \leq x)$ a kvantilovou funkci F_X^{-1}.

Řešení: Nelze řešit: f_x není hustota, neintegruje se do 1.

Příklad: př. Je dáno $f_x(x) = 2x$ pro $0 \leq X \leq 1$. Odvoďte a načtrtněte distribuční funkci F_x, $p(X \leq x)$ a kvantilovou funkci F_X^{-1}.

Řešení: Distribuční funkce je integrál hustoty, tedy $F_X(x) = x^2$. $p(X \leq x) = F_X(x)$ z definice. Kvantilová funkce F_X^{-1} je inverzí k F_X, tedy $F_X^{-1}(u) = \sqrt{u}$ (má smysl samozřejmě jen pro $0 \leq u \leq 1$).

Příklad: (Použití tabulek kvantilů a kritických hodnot:) Hmotnost vyráběné pilulky lze popsat normálním rozdělením se střední hodnotou $120mg$ a rozptylem $1mg^2$. Výstupní kontrola testuje, zda tomu tak skutečně je. Rozumně velký náhodný vzorek pilulek byl zvážen a setříděn podle narůstající hmotnosti.

1. V jakém rozmezí lze čekat hmotnost 10% nejlehčích pilulek?

2. V jakém rozmezí asi bude hmotnost 10% nejtěžších pilulek?
3. V jakém rozmezí lze čekat hmotnost 1% nejlehčích pilulek?
4. V jakém rozmezí lze čekat hmotnost $0,1\%$ nejlehčích pilulek?
5. V jakém rozmezí asi bude hmotnost 1% nejtěžších pilulek?
6. V jakém rozmezí asi bude hmotnost $0,1\%$ nejtěžších pilulek?
7. Jaká je pravděpodobnost, že nalezneme pilulku o hmotnosti 120mg?
8. Jaká je pravděpodobnost, že nalezneme pilulku těžší než 120mg?
9. Jaká je pravděpodobnost, že nalezneme pilulku těžší než 123mg?
10. Jaká je pravděpodobnost, že nalezneme pilulku o hmotnosti nižší než 117,5mg?

Řešení: *Uvažujme symetrii* $\Phi(x)$.

1. rozmezí hmotnosti 10% nejlehčích pilulek: $(max(0, 120 + \Phi^{-1}(0\%)); 120 + \Phi^{-1}(10\%)) = (0; 120 + (-\Phi^{-1}(90\%))) \doteq (0; 120 + (-1,282)) = (0; 118,718)$
2. rozmezí hmotnosti 10% nejtěžších pilulek: $(120 + \Phi^{-1}(90\%); 120 + \Phi^{-1}(100\%)) \doteq (1,282; \infty)$
3. rozmezí hmotnosti 1% nejlehčích pilulek: $(max(0, \Phi^{-1}(0\%)); 120 + \Phi^{-1}(1\%)) = (0; 120 + (-\Phi^{-1}(99\%))) \doteq (0; 120 + (-2,326)) = (0; 117,674)$
4. rozmezí hmotnosti 1% nejtěžších pilulek: $(120 + \Phi^{-1}(99\%); 120 + \Phi^{-1}(100\%)) \doteq (2,326; \infty)$
5. rozmezí hmotnosti $0,1\%$ nejlehčích pilulek: $(max(0, 120 + \Phi^{-1}(0\%)); 120 + \Phi^{-1}(0,1\%)) = (0; 120 + (-\Phi^{-1}(99,9\%))) \doteq (0; 120 + (-3,009)) = (0; 116,991)$
6. rozmezí hmotnosti $0,1\%$ nejtěžších pilulek: $(120+\Phi^{-1}(99,9\%); 120+\Phi^{-1}(100\%)) \doteq (3,009; \infty)$
7. Pravděpodobnost nálezu pilulky o hmotnosti 120mg: Pokud provádíme vážení naprosto přesně, pak tato pravděpodobnost je 0, protože u spojité náhodné veličiny nabývající nenulové pravděpodobnosti na nedegenerovaném intervalu platí, že pravděpodobnost jedné konkrétní realizace je nulová. *Srovnej s* hustotou pravděpodobnosti *- ta nulová není, ale aby se dal interpretovat jako* pravděpodobnost*, musí se zintegrovat na alespoň malém intervalu v okolí bodu, který nás zajímá.*

   Pokud vážíme pouze s určitou upřesností, například s přesností na desetiny gramu, bude naměření hmotnosti $120mg$ vlastně znamenat, že hmotnost pilulky je v rozsahu $< 119,95; 120,4\bar{9} >=< 120 - 0,1/2; 120 + 0,1/2 >$. Definujeme-li náhodnou veličinu $Y = X - 120mg$ (takže $EY = 0mg$), bude pravděpodobnost p nálezu pilulky o skutečné hmotnosti X a naměřené hmotnosti $120mg$

$$
\begin{aligned}
p &= P(119,95 \leq X < 120,05) \\
&= P(120 - 0,05 \leq X < 120 + 0,05) \\
&= P(-0,05 \leq Y < 0,05) \\
&= P(Y < 0,05) - p(Y < -0,05)
\end{aligned}
$$

   a protože $p(Y = y) = 0$, bude

$$
\begin{aligned}
p &= P(Y \leq 0,05) - P(Y \leq -0,05) \\
&= \Phi(0,05) - \Phi(-0,05)
\end{aligned}
$$

   Nyní využijeme symetrie normovaného normálního rozdělení kolem 0, tedy toho, že

$$\phi(x) = \phi(-x)$$

a

$$\Phi(x) = 1 - \Phi(-x),$$

a tak

$$\begin{aligned} p &= \Phi(0,05) - (1 - \Phi(0,05)) \\ &= 2\Phi(0,05) - 1. \end{aligned}$$

Hodnotu $\Phi(0,05)$ v tabulkách nemáme, a tak odhadneme kýženou pravděpodobnost konzervativně raději větší[1], než skutečnou pomocí $\Phi(0,06)$:

$$\begin{aligned} p &\approx 2\Phi(0,06) - 1 \\ &\approx 2 \cdot 0,5239 - 1 = 0,0476. \end{aligned}$$

Při nepřesném měření je tedy pravděpodobnost nálezu určité konkrétní realizace náhodné veličiny nenulová, v našem případě měření s přesností na $0,1mg$ je pravděpodobnost nálezu pilulky o hmotnosti $120mg$ necelých 5%. *Zamyslete se, jak by se tato pravděpodobnost změnila, kdybychom měřili s větší přesností? A jak by se změnila, kdyby hmotnost pilulek neměla jednotkový rozptyl, ale větší (např. $10mg^2$)?*

1.14 charakteristiky rozdělení (střední hodnota, rozptyl, směrodatná odchylka)

Střední hodnota popisuje typickou[2] hodnotu náhodné veličiny. Počítá se jako vážený průměr všech možných relizací náhodné veličiny. Pro spojitou náhodnou veličinu X to je:

$$EX = \int_{-\infty}^{\infty} x f_X(x) dx \tag{16}$$

za předpokladu, že

$$\int_{-\infty}^{\infty} |x| f_X(x) dx < \infty \tag{17}$$

Pro diskrétní náhodnou veličinu X neprůměrujeme přes všechna reálná čísla, ale jen přes jednotlivé diskrétní realizace:

$$EX = \sum_i x_i p(X = x_i) \tag{18}$$

Pozn.: pomocí distribuční funkce můžeme definici sjednotit na:

$$EX = \int_{-\infty}^{\infty} x dF_X(x) dx \tag{19}$$

Vlastnosti střední hodnoty:

$$E(a + bX) = a + bEX \tag{20}$$
$$E(X + Y) = EX + EY \tag{21}$$
$$E(XY) = EXEY, \qquad \text{pokud jsou } X, Y \text{ nezávislé} \tag{22}$$
$$\tag{23}$$

(a, b nenáhodné konstanty, X, Y nezávislé náhodné veličiny)

Příklad: Uvedené vztahy odvoďte.

Pro charakterizaci kolísání náhodné veličiny kolem střední hodnoty se používá rozptyl $varX$ definovaný jako:

[1] poznamenejme, že skutečná pravděpodobnost je přibližně rovna $0,0399$

[2] anglicky *expected*, odtud značení EX

$$var(X) = E(X - EX)^2 \tag{24}$$

Pozn.: Proč se používá čtverec a ne přímo $E(X - EX)$? Protože $E(X - EX) = E(X) - E(EX)) = EX - EX = 0$)

Lze jednoduše ukázat, že $E(X - a)^2$ je minimalizován při volbě $a = EX$. V tomto smyslu je tedy střední hodnota opravdu "středem" náhodné veličiny.

Vlastnosti rozptylu:

$$var(a + bX) = b^2 varX \tag{25}$$

$$varX = (EX^2) - (EX)^2 \tag{26}$$

$$\tag{27}$$

(a, b nenáhodné konstanty, X náhodná veličina)

Příklad: Dokažte uvedená tvrzení.

1.15 charakteristiky realizace náhodné veličiny

- (výběrový) průměr ($\bar{x}$, $\bar{x}_n$)
- (výběrový) medián ($\tilde{x}$, $\tilde{x}_n$)
- (výběrový) modus
- (výběrový) rozptyl (S^2)
- (výběrová) směrodatná odchylka
- (výběrové) kvantily
- (výběrové) inter-kvartilové rozpětí (IQR)

Ukázat na příkladech.

Příklad: př. hmotnost studentů – v modelovém světě graficky vyznačte výběr (několik realizací náhodné veličiny "hmotnost") a sumarizujte

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{28}$$

je nevychýlený odhad. Ukážeme to pomocí tvrzení:

$$\begin{aligned}
\sum_{i=1}^{n} (x_i - \mu)^2 &= \sum_{i=1}^{n} (x_i - \bar{x} + \bar{x} - \mu)^2 \\
&= \sum_{i=1}^{n} [(x_i - \bar{x})^2 + 2(x_i - \bar{x})(\bar{x} - \mu) + (\bar{x} - \mu)^2] \\
&= \sum_{i=1}^{n} (x_i - \bar{x})^2 + 2 \sum_{i=1}^{n} (x_i - \bar{x})(\bar{x} - \mu) + \sum_{i=1}^{n} (\bar{x} - \mu)^2 \\
&= \sum_{i=1}^{n} (x_i - \bar{x})^2 + 2(\bar{x} - \mu) \sum_{i=1}^{n} \underbrace{(x_i - \bar{x})}_{=0} + n(\bar{x} - \mu)^2 \\
&= \sum_{i=1}^{n} (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2
\end{aligned}$$

a tedy

$$\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}(x_i-\mu)^2-n(\bar{x}-\mu)^2$$

z čehož

$$\begin{aligned}
ES^2 &= E\bigg(\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\bigg) && (29)\\
&= \frac{1}{n-1}E\bigg(\sum_{i=1}^{n}(x_i-\mu)^2-n(\bar{x}-\mu)^2\bigg) && (30)\\
&= \frac{1}{n-1}\bigg(\sum_{i=1}^{n}E(x_i-\mu)^2-nE(\bar{x}-\mu)^2\bigg) && (31)\\
&= \frac{1}{n-1}\big(n\sigma^2-n\cdot var(\bar{x})\big) && (32)\\
&= \frac{1}{n-1}\big(n\sigma^2-n\frac{\sigma^2}{n}) && (33)\\
&= \frac{1}{n-1}\big((n-1)\sigma^2\big) && (34)\\
&= \sigma^2 && (35)
\end{aligned}$$

Někdy se používá vychýlený odhad:

$$S^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 \qquad (36)$$

1.16 centrální limitní věta a význam normálního rozdělení.

Ukázat na simulacích.

1.17 Čebyševova nerovnost

$$p(|X-EX|\geq\epsilon)\leq\frac{varX}{\epsilon^2} \qquad (37)$$

Důkaz pro diskrétní rozdělení:

$$\begin{aligned}
varX &= E(X-EX)^2 && (38)\\
&= \sum_{i}(X_i-EX)^2p_i && (39)\\
&\geq \sum_{i:|X_i-EX|\geq\epsilon}(X_i-EX)^2p_i && (40)\\
&\geq \sum_{i:|X_i-EX|\geq\epsilon}\epsilon^2p_i && (41)\\
&= \epsilon^2\sum_{i:|X_i-EX|\geq\epsilon}p_i && (42)\\
&= \epsilon^2p(|X_i-EX|\geq\epsilon) && (43)\\
& && (44)
\end{aligned}$$

a tedy

$$p(|X-EX|\geq\epsilon)\leq\frac{varX}{\epsilon^2} \qquad (45)$$

Příklad: Demonstrovat sílu Čebyševovy nerovnosti ve srovnání s tím, co o sobě říká normální rozdělení. $X \sim N(0,1)$.

1. Jaká je $p(X > 0)$?
2. Jaká je $p(X > 1)$?
3. Jaká je $p(X > 2)$?
4. Jaká je $p(X > 3)$?

Řešení:

1. $p(X > 0)$: Z Čebyševovy nerovnosti dostáváme

$$p(|X - EX| \geq 0) \leq \frac{varX}{0^2} = \frac{1}{0} \tag{46}$$

a z distribuční funkce

$$p(|X - EX| \geq 0) = 2p(X - EX > 0) = 2(1 - \Phi(0)) = 2(1 - 0,5) = 1. \tag{47}$$

(Tedy Čebyševova nerovnost vlastně nepřináší žádnou novou informaci: říká, že pravděpodobnost je konečná, což jsme věděli už dříve (věděli jsme dokonce více - že pravděpodobnost je nanejvýš 1).)

2. $p(X > 1)$:

$$p(|X - EX| \geq 1) \leq \frac{varX}{1^2} = \frac{1}{1} = 1 \tag{48}$$

a z distribuční funkce

$$p(|X - EX| \geq 1) = 2p(X - EX > 1) = 2(1 - \Phi(1)) \doteq 2(1 - 0,84134) \doteq 0,31731. \tag{49}$$

3. $p(X > 2)$:

$$p(|X - EX| \geq 2) \leq \frac{var(X)}{2^2} = \frac{1}{4} = 0,25 \tag{50}$$

a z distribuční funkce

$$p(|X - EX| \geq 2) = 2p(X - EX > 2) = 2(1 - \Phi(2)) \doteq 2(1 - 0,97725) = 0,04550. \tag{51}$$

4. $p(X > 3)$:

$$p(|X - EX| \geq 3) \leq \frac{var(X)}{3^2} = \frac{1}{9} = 0,111\ldots \tag{52}$$

a z distribuční funkce

$$p(|X - EX| \geq 3) = 2p(X - EX > 3) = 2(1 - \Phi(3))) = 1 - 0,99865 = 0,00270. \tag{53}$$

Na co tedy Čebyševova nerovnost je, když distribuční funkce nám dodá mnohem více informací?

2 Odhady parametrů.

2.1 odhad, druhy odhadů, vychýlení a rozptyl odhadu, konzistentní odhad

Příklad: Střelba na terč:

- vychýlená s velkým rozptylem,
- vychýlená s malým rozptylem,
- nevychýlená s velkým rozptylem,
- nevychýlená s malým rozptylem.

2.2 rozklad střední kvadratické chyby odhadu na systematickou chybu (vychýlení) a rozptyl

θ - skutečná (neznámá) hodnota parametru, T_i - odhady parametrů, ET - střední hodnota odhadů

$$\begin{aligned}
E(T-\theta)^2 &= E(T-ET+ET-\theta)^2 && (54)\\
&= E[(T-ET)^2 + 2(T-ET)(ET-\theta) + (ET-\theta)^2] && (55)\\
&= E(T-ET)^2 + 2E(T-ET)\underbrace{(ET-\theta)}_{konstanta} + E\underbrace{(ET-\theta)^2}_{konstanta} && (56)\\
&= E(T-ET)^2 + 2(ET-\theta)\underbrace{E(T-ET)}_{=0} + (ET-\theta)^2 && (57)\\
&= var(T) + (ET-\theta)^2 && (58)\\
& && (59)
\end{aligned}$$

tedy střední kvadratická chyba je součtem rozptylu odhadu a čtverce vychýlení.

2.3 metoda momentů

Příklad: Při vykopávkách bylo odhaleno několik kostí dinosaura. Archeologové mají představu o proporcích dinosaura a chtějí na základě nálezu restaurovat jeho celkovou velikost.

Řešení: Vzít jednu nebo více kostí (funkci realizací náhodných veličin) a dát je do souvislosti s teoretickým modelem dinosaura (teoretickými protějšky odpovídajících náhodných veličin).

Je lepší vzít málo, nebo více kostí? (Více - proto se výběr charakterizuje momenty, které stavějí nad hodnotami všech pozorování, a ne třeba nad kvantily (které zohledňují uspořádání, ale ne přímo hodnoty všech jednotlivých realizací).)

Příklad: Odhad parametrů $N(\mu, \sigma^2)$.

2.4 metoda maximální věrohodnosti

Věrohodnostní funkce (jako funkce dat resp. parametrů).

Příklad: Kuřáci a nedopalek I. Potkáme dva známé, o nichž víme, že jeden z nich je kuřák a druhý nekuřák. Mezi nimi leží nedopalek. Co je pravděpodobnější - že nedopalek odhodil kuřák, nebo že se na zemi ocitl vinou nekuřáka?

Věrohodnostní funkce $L(x|p) = L(nedopalek|osoba\ (ne)kouří)$ jako funkce dat za daných parametrů.

Příklad: Kuřáci a nedopalek II. Potkáme dva známé, o nichž nevíme, zda kouří, nebo nekouří. Pod jedním z nich leží nedopalek. Co si pomyslíme o tom, který asi kouří - ten, pod který leží nedopalek, nebo ten druhý?

Věrohodnostní funkce jako funkce neznámých parametrů za daných dat: $L(p|x) = L(osoba\ (ne)kouří|nedopalek)$. Funkci vyhodnotíme dvakrát pro různou hodnotu parametru a zjistíme, která varianta je věrohodnější.

2.5 příklady na obě metody

Příklad: Z jediné realizace x_1 náhodné veličiny X o níž víte, že pochází z normálního rozdělení $N(\mu, 1)$, odhadněte parametr μ momentovou metodou a metodou maximální věrohodnosti.

1. Budou odhady nestranné?
2. Jaké budou mít rozptyly?
3. Budou konzistentní?

Kolik a jakých momentů použijete?

Řešení:

1. Momentovou metodou: Stačí jediný moment: $EX = \mu$ odhadneme pomocí x_1: $\hat{\mu} = m_X = \frac{1}{1}\sum_{i=1}^{1} x_i = x_1$. $\hat{\mu} = x_1$ je nestranný odhad, protože $Ex_1 = \mu$.
   Metodou maximální věrohodnosti: věrohodnostní funkce je
   $$L(x_1|\mu, \sigma^2) = f_X(x_1|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x_1-\mu}{2\sigma^2}}$$
   potom logaritmická věrohodnostní funkce je
   $$l(x_1|\mu, \sigma^2) = \frac{-1}{2} ln(2\pi) - \frac{1}{2} ln(\sigma^2) - \frac{(x_1-\mu)^2}{2\sigma^2}$$
   a hodnotu parametru μ nalezneme, položíme-li parciální derivaci podle μ rovnu 0:
   $$\begin{aligned} \frac{\partial l(x_1|\mu, \sigma^2)}{\partial \mu} &= -\frac{2(x_1-\mu)(-1)}{2\sigma^2} = \frac{(x_1-\mu)}{\sigma^2} \\ \frac{(x_1-\hat{\mu})}{\sigma^2} &= 0 \\ \hat{\mu} &= x_1 \end{aligned}$$
2. Rozptyl odhadu: $var\hat{\mu} = varx_1 = \sigma^2$.
3. Budou odhady konzistentní? Nelze přímo určit, protože ke konzistenci bychom potřebovali vysledovat chování odhadů pro zvětšující se velikosti výběrů, ale my máme k dispozici pouze jediné pozorování.

Příklad: Ze dvou realizací x_1, x_2 náhodné veličiny X o níž víte, že pochází z normálního rozdělení $N(\mu, 1)$, odhadněte parametr μ momentovou metodou a metodou maximální věrohodnosti.

Řešení:

1. Momentovou metodou: Opět stačí jediný moment: $EX = \mu$ odhadneme pomocí prvního výběrového momentu: $\hat{\mu} = m_X = \frac{1}{2}\sum_{i=1}^{2} x_i = \frac{x_1+x_2}{2}$. $\hat{\mu}$ je nestranný odhad, protože $E\frac{x_1+x_2}{2} = \frac{Ex_1+Ex_2}{2} = \frac{\mu+\mu}{2} = \mu$.
   Metodou maximální věrohodnosti: věrohodnostní funkce je
   $$L(x_1, x_2|\mu, \sigma^2) = \prod_{i=1}^{2} f_X(x_i|\mu, \sigma^2) = \prod_{i=1}^{2} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x_i-\mu}{2\sigma^2}}$$
   potom logaritmická věrohodnostní funkce je
   $$\begin{aligned} l(x_1, x_2|\mu, \sigma^2) &= \sum_{i=1}^{2} \Big(-\frac{1}{2} ln(2\pi) - \frac{1}{2} ln(\sigma^2) - \frac{(x_i-\mu)^2}{2\sigma^2} \Big) \\ &= -\frac{n}{2} ln(2\pi) - \frac{n}{2} ln(\sigma^2) - \sum_{i=1}^{2} \frac{(x_i-\mu)^2}{2\sigma^2} \end{aligned}$$

a hodnotu parametru μ nalezneme, položíme-li parciální derivaci podle μ rovnu 0:

$$\frac{\partial l(x_1, x_2|\mu, \sigma^2)}{\partial \mu} = \sum_{i=1}^{2} \frac{2(x_1 - \mu)(-1)}{2\sigma^2} = \sum_{i=1}^{2} \frac{(x_1 - \mu)}{\sigma^2}$$

$$\sum_{i=1}^{2} \frac{(x_1 - \hat{\mu})}{\sigma^2} = 0$$

$$\hat{\mu} = \frac{\sum_{i=1}^{2} x_i}{2}$$

2. Rozptyl odhadu: $var\hat{\mu} = var\frac{x_1+x_2}{2} = \frac{varx_1+varx_2}{4} = \frac{\sigma^2+\sigma^2}{4} = \frac{\sigma^2}{2}$. Tedy tento odhad bude mít rozptyl poloviční ve srovnání s odhadem pořízeným u jediné realizace.

3. Budou odhady konzistentní? Nelze přímo určit, protože ke konzistenci bychom potřebovali vysledovat chování odhadů pro zvětšující se velikosti výběrů, ale my máme k dispozici pouze jediné pozorování.

TODO

Příklad: (Pokračování přechozího příkladu.) Co kdybyste pro odhad parametr použili pouze poslední realizaci x_2?

- Jak se změní odhady?
- Budou lepší, než minulé odhady?
- Budou nestranné?
- Jaké budou mít rozptyly?
- Budou konzistentní?

Příklad: (Pokračování přechozího příkladu.) Co kdybyste pro odhad parametrů použili větší z realizací, tj. $max(x_1, x_2)$? Jak se změní odhady?

Příklad: Z jediné realizace x_1 náhodné veličiny X o níž víte, že pochází z normálního rozdělení $N(\mu, \sigma^2)$, odhadněte parametry μ a σ^2 momentovou metodou a metodou maximální věrohodnosti. $\hat{\sigma^2} = 0$. *Souvislost s uvíznutím E-M algoritmu v lokálním minimu odpovídajícím singulárním variančním maticím.*

Příklad: Odhad parametrů $N(\mu, \sigma^2)$ z více realizací.

2.6 intervalové odhady

Příklad: Odhadněte co nejpřesněji glykémii (koncentraci glukózy v krvi $[mmol/l]$) u určité skupiny pacientů s těžkou cukrovkou. Víme, že u zdravých lidí jsou hodnoty glykemie typicky v rozmezí cca $3 - 6mmol/l$, ale u pacientů lze očekávat mnohem vyšší hodnoty. Měřit naneštěstí nemůžeme dostatečně přesně, chybu měření odhadujeme pomocí směrodatné odchylky na $s = 4mmol/l$. Navíc prozatím máme k dispozici pouze jediné měření: $x_1 = 11,3mmol/l$.

1. Vhodným způsobem graficky znázorněte rozložení pravděpodobnosti náhodné veličiny “glykemie” - načtrtněte hustotu.

2. Naznačte oblast A, pro kterou $P(X \in A) = 95\%$. Je oblast symetrická? Lze najít jinou takovou oblast? Interpretujte danou oblast. Ověřte, že jste oblast našli správně pomocí numerické simulace.

3. Zkonstruujte 95% interval spolehlivosti pro střední hodnotu glykemie. Interpretujte daný interval. Přesnost je malá – jak to zlepšit?

Řešení: Oblast A je definována jako interval $< q_d, q_h >$, kde $\Phi(q_h) - \Phi(q_l) = 0,95$.

Oblasti mohou být různé, typicky symetrické (pro oboustrannou alternativu, tj. když chyby na obou stranách jsou stejně důležité či očekávatelné).

Interpretace oblasti A: při opakovaných náhodných výběrech budou realizace náhodné veličiny glykémie ležet v oblasti A v 95% případů. V tomto pohledu je tedy pravděpodobnost, že daná náhodná veličina nabývá hodnoty y daného intervalu, rovna 95%.

Podobně lze konstruovat oblast, v níž se budou často nacházet výběrové průměry náhodných výběrů dané velikosti.

Inteval spolehlivosti pro střední hodnotu naproti tomu bude definován kolem výběrového průměru. $(\bar{X} + \frac{\sigma}{\sqrt{n}}\Phi^{-1}(\frac{\alpha}{2}), \bar{X} + \frac{\sigma}{\sqrt{n}}\Phi^{-1}(1 - \frac{\alpha}{2}))$

Příklad: (Pokračování.) Na další schůzce jste seznámeni s požadavkem naměřit glykémii s danou přesností – s přesností na 1 desetinné místo. ***Řešení:*** Tento požadavek nelze splnit. Pokud

by skutečná hodnota ležela blízko hranice, kde se láme zaokrouhlování, ...

Příklad: (Pokračování.) Na další schůzce je navrženo, abyste garantovali, že vzdálenost odhadu od skutečné hodnoty nebyla větší 0,1. Spočtěte počet potřebný počet pokusů. ***Řešení:*** Tento

požadavek také nelze splnit. Nelze dát garanci, jen omezit pravděpodobnost chyby. ...

3 Testování hypotéz

3.1 nulová a alternativní hypotéza

3.2 testová statistika a její rozdělení

3.3 kritická hodnota testu

3.4 P-hodnota testu

3.5 chyba 1. a 2. druhu, síla testu

3.6 ROC křivka

3.7 jednovýběrový t-test

Příklad: Hmotnost vyráběné pilulky lze popsat normálním rozdělením se střední hodnotou $120mg$ a rozptylem $36mg^2$. Výstupní kontrola testuje, zda tomu tak skutečně je. Náhodný vzorek 10 pilulek byl zvážen a byla spočtena jejich průměrná hmotnost $124mg$.

1. Odpovídá vzorek požadované kvalitě pilulek?
   (a) formulujte nulovou a alternativní hypotézu
   (b) navrhněte vhodnou testovou statistiku
   (c) načrtněte rozdělení testové statistiky
   (d) najděte kritickou hodnotu (kritický obor) testu
   (e) proveďte test a vyslovte závěr

2. Odpovídá vzorek požadované kvalitě i v případě, když nebudeme vědět, jaký rozptyl má hmotnost pilulek? (Tj. pouze víme, že hmotnost vyráběné pilulky lze popsat normálním rozdělením se střední hodnotou $120mg$ a s neznámým rozptylem.) Předpokládejme, že si navíc spočteme (nevychýlený) výběrový rozptyl v hodnotě $36mg^2$. *Co kdybychom použili vychýlený výběrový rozptyl?*

3. Jaká je pravděpodobnost chyby 1. druhu výše uvedených testů?

4. Jaká je pravděpodobnost chyby 2. druhu výše uvedených testů? *Potřebujeme k tomu znát něco navíc? (Navíc budeme předpokládat, že očekáváme problém s příliš těžkými pilulkami se střední hodnotou $125mg$.)*

5. Kolik pilulek bychom museli odebrat, aby pravděpdobnost odhalení nekvalitního vzorku byla max. 10%?
   - určete pro případ se známým rozptylem i bez něj

Řešení:

1. Odpovídá vzorek požadované kvalitě pilulek?
   (a) Uvažujeme-li model $X \sim N(\mu, \sigma^2)$, kde X je (náhodná veličina) hmotnost pilulky, μ je střední hodnota veličiny X (za optimistického předpokladu správně vyrobených pilulek je rovna $120mg$), a $\sigma^2 = 36mg^2$. Nulová hypotéza potom vyjadřuje naše optimistické očekávání, tedy:

   $$H_0 : \mu = 120mg. \tag{60}$$

   Druhou možností je alternativní hypotéza:

   $$H_A : \mu \neq 120mg \tag{61}$$

   (b) Testová statistika je nějaká vhodná funkce náhodného výběru (vzorku 10ti pilulek), která vhodně sumarizuje vše důležité z tohoto vzorku do jediného čísla. V našem případě bude roli testové statistiky hrát průměrná hmotnost 10ti pilulek, $\bar{X}_{10}$. Testovou statistiku budeme chápat jednak jako náhodnou veličinu $\bar{X}_{10}$ (tedy něco, co neznáme, co nemáme naměřeno, co je náhodné, ale přesto o tom můžeme ledacos říct - můžeme to zkoumat probabilisticky), jednak jako její realizaci $\bar{x}_{10}$ (průměr konkrétního vzorku 10ti vybraných pilulek).
   (c) Rozdělení testové statistiky za nulové hypotézy můžeme snadno odvodit ze znalosti rozdělení hmotnosti jednotlivých pilulek. Protože za nulové hypotézy víme, jaké rozdělení má náhodná veličina X (hmotnost jedné pilulky):

   $$X \sim N(\mu, \sigma^2), \tag{62}$$

   víme, že

   $$\bar{X}_{10} \sim N(\mu, \frac{\sigma^2}{10}), \tag{63}$$

   protože

   $$\begin{aligned} E\bar{X}_{10} &= E\frac{1}{10}\sum_{i=1}^{10} X_i & (64)\\ &= E\frac{1}{10}10X & (65)\\ &= EX & (66)\\ &= \mu & (67)\end{aligned}$$

   a

   $$var\bar{X}_{10} = \frac{varX}{10}, \tag{68}$$

   což pro ilustraci ukážeme pro průměr dvou stejně rozdělených nezávislých náhodných veličin X_1 a X_2:

$$\begin{aligned}
var\frac{X_1+X_2}{2} &= \frac{1}{4}var(X_1+X_2)\\
&= \frac{E(X_1+X_2-EX_1-EX_2)^2}{4}\\
&= \frac{E[(X_1-EX_1)+(X_2-EX_2)]^2}{4}\\
&= \frac{E[(X_1-EX_1)^2+2(X_1-EX_1)(X_2-EX_2)+(X_2-EX_2)^2]}{4}\\
&= \frac{E(X_1-EX_1)^2+2E(X_1-EX_1)(X_2-EX_2)+E(X_2-EX_2)^2}{4}\\
&= \frac{varX_1+2cov(X_1,X_2)+varX_2}{4}\\
&= \frac{\sigma^2+0+\sigma^2}{4} \text{ (protože pro } X_1, X_2 \text{ nezávislé } cov(X_1,X_2)=0)\\
&= \frac{\sigma^2}{2}
\end{aligned}$$

Rozdělení testové statistiky je tedy normální rozdělení se střední hodnotou $120mg$ a rozptylem $3,6mg^2$.

(d) Kritický obor je taková "oblast" (množina reálných čísel) testových statistik, při kterých budeme zamítat nulovou hypotézu. Budeme chtít, aby za platnosti nulové hypotézy testová statistika do kritického oboru padala co nejméně (s maximální definovou chybou 1. druhu α, označovanou jako hladinu testu a odpovídající falešné pozitivitě), ale za platnosti alternativní hypotézy do něj padala naopak co nejvíce (aby test měl velkou "šanci" neplatnou nulovou hypotézu zamítnout - mluvíme o síle testu a chybě 2. druhu β odpovídající falešné negativitě).

Kritický obor stanovíme na základě požadavku na maximální dovolenou chybu 1.druhu, tj. hladinu testu, kterou standardně volíme $\alpha = 0.05$.

Kritický obor bude část reálné osy mimo oblast, v níž lze očekávat $100(1-\alpha)\%$ testových statistik $\bar{X}_{10}$, tedy sjednocení

$$\begin{aligned}
(-\infty, F^{-1}_{N(\mu,\frac{\sigma^2}{10})}(\alpha/2)] &\quad\cup\quad [F^{-1}_{N(\mu,\frac{\sigma^2}{10})}(1-\alpha/2),\infty)\\
=(-\infty,\mu+\frac{\sigma}{\sqrt{10}}\Phi^{-1}\big(\alpha/2\big)] &\quad\cup\quad [\mu+\frac{\sigma}{\sqrt{10}}\Phi^{-1}\big(1-\alpha/2\big)],\infty)\\
=(-\infty,120+\frac{6}{\sqrt{10}}-1,96] &\quad\cup\quad [120+\frac{6}{\sqrt{10}}1.96,\infty)\\
=(-\infty,116,281] &\quad\cup\quad [123,719,\infty)
\end{aligned}$$

Můžeme však postupovat i tak, že nekonstruujeme kritický obor v měřítku testové statistiky (tedy tak, že škálujeme a posouváme kvantily notrmálního rozdělení), ale naopak tak, že realizaci testové statistiky normujeme tak, aby "seděla" s kvantily normovaného normálního rozdělení:

$$\begin{aligned}
t &= \frac{\bar{x}_{10}-120}{\frac{\sigma}{\sqrt{10}}}\\
&= \frac{\bar{x}_{10}-120}{\sigma}\sqrt{10}\\
&= \frac{124-120}{6}\sqrt{10}\\
&= 2,108
\end{aligned}$$

takže ji v našem případě oboustranného testu můžeme jednoduše porovnat s kritickou hodnotou a H_0 zamítnout, když

$$|t| > \Phi^{-1}\big(1-\alpha/2\big)$$

(e) Na základě kritického oboru:

$$\bar{X}_{10} = 124mg \in (-\infty, 116,281] \quad \cup \quad [123,719,\infty)$$

a (alternativně) i kritické hodnoty:

$$|t| > \Phi^{-1}\big(1-\alpha/2\big)2,108 \qquad\qquad > 1,96$$

zamítáme na hladině 5% nulovou hypotézu, že hmotnost pilulek je rovna jejich nominální hodnotě.

2. V případě, že nebudeme předem znát rozptyl hmostnosti jednotlivých pilulek, musíme rozptyl odhadnout z dat a zohlednit to, že jedná o (nepřesný) odhad, nikoli o přesné číslo. Rozdělení testové statistiky $\bar{X}_{10}$ nyní nebude normální, ale bude to t-rozdělení (s $10-1$ stupni volnosti).

Normovaná testová statistika bude

$$\begin{aligned} t &= \frac{\bar{x}_{10}-120}{s_x}\sqrt{10} \\ &= \frac{124-120}{6}\sqrt{10} \\ &= 2,108 \end{aligned}$$

a kritická hodnota bude

$$F_{t_9}^{-1}\big(1-\alpha/2\big) = 2,26,$$

takže

$$|t| = 2,108 < F_{t_9}^{-1}\big(1-\alpha/2\big) = 2,26$$

a nulovou hypotézu na hladině 5% nezamítáme.

Kdybychom místo nevychýlého výběrového rozptylu

$$s_X^2 = \frac{1}{9}\sum_{i=1}^{10}(x_i - \bar{x}_{10})^2 = 36$$

použili vychýlený výběrový rozptyl

$$s_X^2 = \frac{1}{10}\sum_{i=1}^{10}(x_i - \bar{x}_{10} = 36\frac{9}{10} = 32,4,$$

který je z principu menší, než nevychýlený, (a bylo by nesprávné jej použít, protože by uměle snižoval naši nejistotu o rozptylu), dostali bychom normovanou testovou statistiku

$$\begin{aligned} t &= \frac{\bar{x}_{10}-120}{s_x}\sqrt{10} \\ &= \frac{124-120}{\sqrt{32,4}}\sqrt{10} \\ &= 2,222 \end{aligned}$$

a nulovou hypotézu na hladině 5% bychom zamítli také.

3. Hladina chyby 1. druhu je z principu rovna hladině významnosti α a je tedy 5%.

4. Hladina chyby 2. druhu nelze zjistit pouze na základě znalosti rozdělení testové statistiky při nulové hypotéze. Potřebujeme totiž zjistit pravděpodobnost, že za platnosti alternativní hypotézy uděláme chybu: nezamítneme nulovou hypotézu, přestože bychom ji zamítnout měli.

   Když budeme předpokládat alternativu příliš těžkých pilulek (se střední hodnotou $125mg$), můžeme pravděpodobnost chyby 2. druhu β spočítat:

   TODO

5. Nyní zjistíme, jak velký výběr pilulek by výstupní kontrola musela provést, aby se pravděpodobnost chyby 2. druhu β snížila a byla nanejvýš 10%. Chybu 2. druhu uděláme, pokud za platnosti alternativní hypotézy H_A nezamítneme H_0. Tedy když

$$H_d \leq \bar{X}_n \leq H_h,$$

kde

$$H_d(n) = F^{-1}_{N(120,36/n)}(\alpha/2) = 120 + \Phi^{-1}(\alpha/2)\frac{6}{\sqrt{n}}$$

a

$$H_h(n) = F^{-1}_{N(120,36/n)}(1-\alpha/2) = 120 + \Phi^{-1}(1-\alpha/2)\frac{6}{\sqrt{n}}$$

Potom chceme, aby pravděpodobnost chyby 2. druhu byla

$$\beta = P(H_d(n) \leq \bar{X}_n \leq H_h(n)|H_A) \leq 0,1$$

přitom

$$P(H_d(n) \leq \bar{X}_n \leq H_h(n)|H_A) < P(\bar{X}_n \leq H_h(n)|H_A)$$

ale rozdíl není velký, např. pro $n = 10$ je

$$\begin{aligned}
P(H_d(n) > \bar{X}_n)|H_A) &= P(\bar{X}_n < H_d(n)|H_A) \\
&\approx P(\bar{X}_n < 120 - 1,96 \cdot 6/\sqrt{10}|H_A) \\
&= F_{N(125,36/10)}(120 - 1,96 \cdot 6/\sqrt{10}) \\
&= \Phi\big(\frac{120 - 1,96 \cdot 6/\sqrt{10}}{6/\sqrt{10}} - 125\big) \\
&= \Phi\big(-1,96 - \frac{5\sqrt{10}}{6}\big) \\
&\approx \Phi\big(-4,595\big) \\
&\approx 0,00000216
\end{aligned}$$

$$\begin{aligned}
0,1 \geq P(H_d(n) \leq \bar{X}_n \leq H_h(n)|H_A) &\approx P(\bar{X}_n \leq H_h(n)|H_A) \\
&= F_{N(125,36/n)}(H_h(n)) \\
&= \Phi\Big(\frac{H_h(n) - 125}{6/\sqrt{n}}\Big) \\
&= \Phi\Big(\frac{120 + \Phi^{-1}_{N(0,1)}(1-\alpha/2)\frac{6}{\sqrt{n}} - 125}{6/\sqrt{n}}\Big) \\
&= \Phi\big(\Phi^{-1}_{N(0,1)}(1-\alpha/2) - \frac{5\sqrt{n}}{6}\big)
\end{aligned}$$

Nyní budeme obě strany rovnice chápat jako argument funkce Φ^{-1}:

$$\begin{aligned}
\Phi^{-1}(0,1) &\geq \Phi^{-1}_{N(0,1)}(1-\alpha/2) - \frac{5\sqrt{n}}{6} \\
-1,282 &\geq 1,960 - \frac{5\sqrt{n}}{6} \\
\frac{5\sqrt{n}}{6} &\geq 1,960 + 1,282 \\
n &\geq \left[\frac{6}{5}(1,960 + 1,282)\right]^2 \\
n &\geq 15.1
\end{aligned}$$

TODO: zkouška

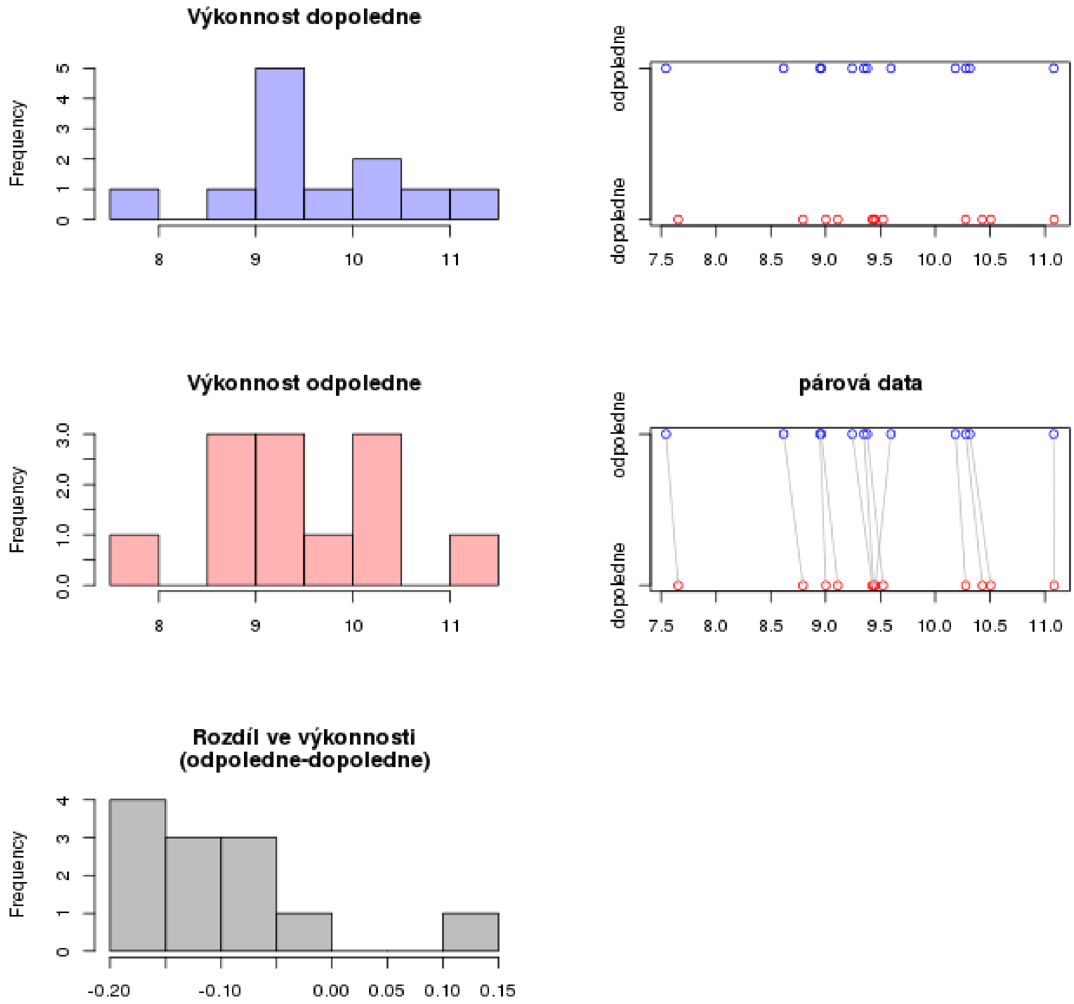
3.8 párový a dvouvýběrový t-test

Příklad: Vedení továrny zjišťuje, zda pracovní výkonnost po obědě klesá. U pracovníků sleduje výkonnost dopoledne a odpoledne. Byly naměřeny následující hodnoty výkonnosti: dopoledne: 8,79; 10,28; 11,08; 7,65; 10,43; 10,51; 9,43; 9,45; 9,44; 9,11; 9,52; 9,00 ((výběrový) průměr $\bar{x}_1 \approx 9,557$, (výběrová) směrodatná odchylka $0,919$); odpoledne: 8,62; 10,18; 11,08; 7,54; 100,28; 10,31; 9,24; 9,59; 9,35; 8,96; 9,38; 8,95 ((výběrový) průměr $\bar{x}_2 \approx 9,457$, (výběrový) rozptyl $0,866$). Směrodatná odchylka rozdílů ve výkonnosti mezi odpolednem a dopolednem je $0,095$. Na hladině 5% otestujte hypotézu, že výkonnost odpoledne klesá.

1. Zamyslete se nad tím, co vlastně chcete zkoumat a jaká data máte k dispozici. Případně nejasnosti konzultujte se zadavatelem (cvičícím). *Představují data párová nebo nepárová pozorování, tj. byly výkonnosti odpoledne naměřeny na* stejných *pracovnících jako dopoledne, nebo ne? Co by bylo správnější? Co bychom mohli v jednotlivých případech testovat a jak? Jaký výsledek byste v jednotlivých případech očekávali a proč?*
2. Zkontrolujte data.
3. Formulujte nulovou a alternativní hypotézu.
4. Proveďte test a vyslovte závěr.

Řešení:

1. Pokud by data představovala nepárová pozorování, měli bychom úlohu znesnadněnou variabilitou výkonnosti mezi jednotlivými pracovníky, takže test by na daných datech nejspíš žádný rozdíl nezaznamenal, protože rozdíl v průměrech je ve srovnání se směrodatnou odchylkou relativně malý. Pokud by data představovala párová pozorování, variabilita ve výkonnosti mezi jednotlivými pracovníky by nehrála žádnou roli, protože by se eliminovala při výpočtu rozdílu ve výkonnosti mezi dopolednem a odpolednem.
2. Kontrola dat: jedna z odpoledních výkonností $(100,28)$ je odlehlá, patrně překlep - po konzultaci se zadavatelem opravíme na $10,28$.



3. Nulová a alternativní hypotéza:

   H_0 : výkonnost dopoledne = výkonnost odpoledneH_A : výkonnost dopoledne > výkonnost odpoledne

4. Proveďte test a vyslovte závěr.

$$(111)p^1 0(1-p)^1$$

   Pozor, u dopoledních výkonností je uvedena výběrová směrodatná odchylka, u odpoledních výkonností naopak výběrový rozptyl!

   Nejprve provedem (nesprávný) nepárový test:

$$
\begin{aligned}
t &= \frac{\bar{x}_2 - \bar{x}_1}{\sqrt{\frac{11s_1^2+11s_2^2}{11+11}(\frac{1}{11}+\frac{1}{11})}} \\
&= \frac{\bar{x}_2 - \bar{x}_1}{\sqrt{\frac{s_1^2+s_2^2}{2}\frac{2}{11}}} \\
&= \frac{\bar{x}_2 - \bar{x}_1}{\sqrt{\frac{s_1^2+s_2^2}{11}}} \\
&= \frac{9,457 - 9,557}{\sqrt{\frac{0,919^2+0,866}{11}}} \\
&\approx -0,254
\end{aligned}
$$

kvantil t-rozdělení:

$$q_{t_{22}}^{-1}(\alpha) = -1.72$$

Protože $t > q_{t_{22}}(\alpha)$, H_0 nezamítáme.

V případě, že data budeme správně chápat jako párová pozorování, dostaneme:

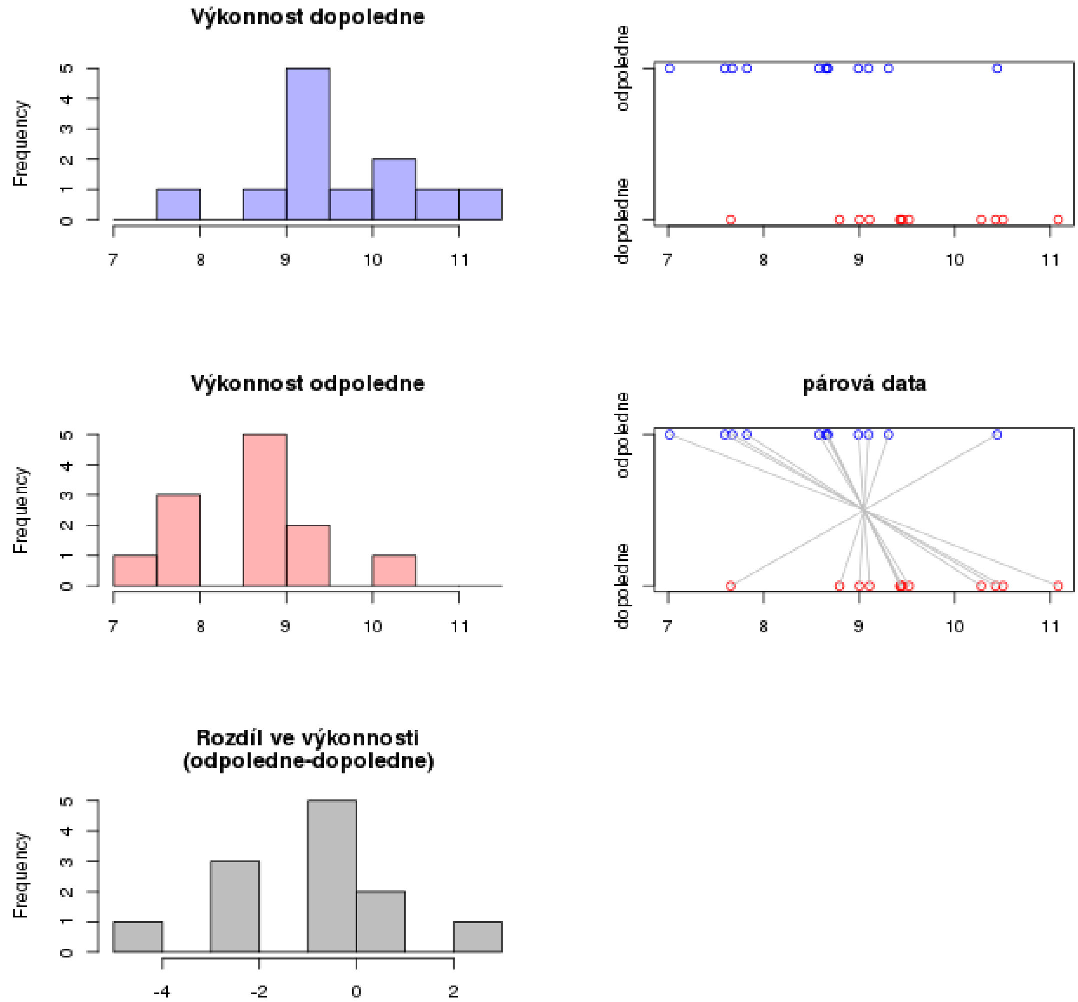
$$
\begin{aligned}
t &= \frac{\bar{x}_2 - \bar{x}_1}{s_d}\sqrt{11} \\
t &= \frac{9,457 - 9,557}{0,095}\sqrt{11} \\
&\approx -3,491
\end{aligned}
$$

kvantil t-rozdělení je přitom $q_{t_11}(\alpha) = -1.8$

Protože $t < q_{t_{11}}(\alpha)$, H_0 na hladině 5% zamítáme.

Příklad: (Pokračování předchozího příkladu:) Po čase byly naměřeny následující hodnoty výkonnosti: dopoledne: 8,79; 10,28; 11,08; 7,65; 10,43; 10,51; 9,43; 9,45; 9,44; 9,11; 9,52; 9,00 (průměr $\bar{x}_1 = 9,557$, směrodatná odchylka 0,919); odpoledne: 9,31; 7,82; 7,02; 10,45; 7,67; 7,59; 8,67; 8,65; 8,66; 8,99; 8,58; 9,10 (průměr $\bar{x}_2 = 8,543$, rozptyl 0,845). Na hladině 5% otestujte hypotézu, že výkonnost odpoledne klesá.

1. Opět se zamyslete, jaká data máte k dispozici a co chcete zkoumat.
2. Formulujte nulovou a alternativní hypotézu.
3. Proveďte test a vyslovte závěr.



Řešení:

3.9 χ^2 test dobré shody

Příklad: Ze sta hodů mincí padla panna 44-krát. Lze z tohoto výsledku usuzovat na to, zda je mince falešná? Znáte-li více možností, jak na tuto otázku odpovědět, použijte všechny takové přístupy.

Řešení:
$X_i \sim Bi(n,p) = Bi(100, 0,5)$
Za platnosti $H_0 : p = 0,5$ je pravděpdobnost, že padne 44 nebo méně panen,

$$p = \sum_{i=0}^{44} \binom{100}{i} p^i (1-p)^{100-i} \approx 0,136$$

tedy p-hodnota testu v binomickém rozdělení je vyšší než 5% a hypotézu nemůžeme zamítnout.
Aproximací binomického rozdělení normálním:

$$X_i \sim N(p, \frac{p(1-p)}{n})$$

a můžeme zkonstruovat interval spolehlivosti pro střední hodnotu normálního rozdělení:

$P = 0,113$
Pomocí χ^2 testu:

$$\chi^2 = \sum_{i=1}^{2} \frac{(n_i - np_i)^2}{np_i} = \sum_{i=1}^{2} \frac{(n_i - 100 \cdot 0,5)^2}{100 \cdot 0,5} = \frac{(44-50)^2 + (56-50)^2}{50} = 1,44$$

Porovnáním s kvantilem $q_{\chi^2_1}(0,95) = 3,84$ vychází, že nemůžeme zamítnout hypotézu, že na minci padají stejně pravděpdobně panny a orli.
$P = 0,230$

Příklad: Krevní skupiny
Ve vybraných nemocnicích byly u ambulantně ošetřených pacientů v určitém dni sledovány podíly jednotlivých krevních skupin. Dostali jsme následující tabulku:

místo	A	B	0	AB
Karlovy Vary	33	6	56	5
Ostrov nad Ohří	9	1	16	1
Olomouc	54	14	52	5

Otestujte, zda jsou podíly jednotlivých krevních skupin v jednotlivých nemocnicích stejné.

Řešení:
Vidíme, že v Ostrově nad Ohří bylo vyšetřeno relativně málo pacientů a některé krevní skupiny zde byly zastoupeny jen jediným pacientem. Protože χ^2 test dobré shody je testem asymptotickým a vyžaduje, aby v každé buňce tabulky bylo alespoň 5 pozorování, musíme přistoupit ke slučování některých řádků a nebo sloupců. Nejpřitozenějším se jeví sloučit data z nemocnice v Ostrově nad Ohří s daty z Karlových Varů. Dostáváme tedy novou tabulku:

místo	A	B	0	AB
Karlovy Vary a Ostrov nad Ohří	42	7	72	6
Olomouc	54	14	52	5

Nyní jsou předpoklady testu splněny.
Nulovou hypotézou bude, že zastoupení krevních skupin se mezi nemocnicemi neliší, tedy že rozdělení pacientů s jednotlivými krevními skupinami je v obou nemocnicích stejné (homogenní). To je ekvivalentní tomu, že rozdělení krevních skupin nezávisí na tom, v jaké nemocnici jej zjišťujeme, jinými slovy, jev naměřená krevní skupina je nezávislá na tom, v jaké nemocnici byla měřena. Z nezávislosti přímo plyne, že pravděpodobnost $p_{i,j}$ naměření j-té krevní skupiny v i-té nemocnici je dáno jako součin pravděpodobnosti p_j^J, že měříme j-tou krevní skupinu, a pravděpodobnosti p_i^I, že měříme v i-té nemocnici:

$$H_0 : p_{i,j} = p_i^I p_j^J$$
$$H_A : p_{i,j} \neq p_i^I p_j^J$$

Odhadneme jednotlivé pravděpodobnosti:

$$p_i^I = \frac{\sum_{j=1}^4 n_{ij}}{n}$$
$$p_j^J = \frac{\sum_{i=1}^2 n_{ij}}{n}$$

a spočteme testovou statistiku

$$T = \sum_{i=1}^{2} \sum_{j=1}^{4} \frac{(n_{ij} - np_{i,j})^2}{np_{i,j}} = 7,1346$$

Kritickou hodnotu určíme jako kvantil χ^2 rozdělení na $q_{\chi^2_{(4-1)(2-1)}}(1 - \alpha/2) = 7,81.$

Protože hodnota testové statiky T nepřesahuje kritickou hodnotu, H_0 nezamítáme. To ovšem neznamená, že bychom H_0 prokázali! Pouze jsme v datech nenalezli rozpor s nulovou hypotézou. (Jedna z možností je, že opravdu nulová hypotéza platí. Druhou možností by bylo, že test nemá dostatečnou sílu nulovou hypotézu vyvrátit. Protože mezi těmito možnostmi nedokážeme rozlišit, nelze nezamítnutí nulové hypotézy interpretovat jako její platnost!)

3.10 test korelačního koeficientu

Příklad: Deprese a kouření - korelace.

V 15 vybraných okresech ČR bylo sledováno procento obyvatel, kteří kouří, a zároveň index depresivity v daném okresu. Dostali jsme následující data:

procento kouření	index výskytu deprese
0,76	1,27
0,57	1,06
0,93	1,73
0,64	1,21
0,70	0,81
0,48	1,29
0,85	1,42
0,42	0,63
0,03	0,78
0,26	0,57
0,33	0,82
0,13	1,12
0,50	0,92
0,80	1,04
0,34	0,56

Průměrné procento kouření $\bar{k} = 0,516$, průměrný index depresivity $1,015$. Korelační koeficient $r = 0,656$. Na hladině 5% otestujte, zda jsou míra kouření a výskyt depresí navzájem korelované. Co se dá z výsledku usuzovat? ***Řešení:***

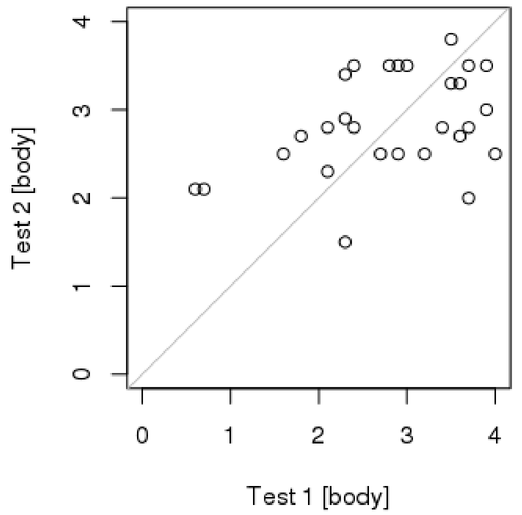
Testová statistika t je

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0,656\sqrt{15-2}}{\sqrt{1-0,656^2}} \approx 3,13$$

a kritická hodnota je $q_{t_13}(0,975) = 2,16$. Protože $|t| > q_{t_13}(0,975)$, zamítáme nulovou hypotézu o nulovosti korelačního koeficientu.

Příklad: Srovnání výsledků testu 1 a 2 zadaných na cvičení SSL v LS 2014/15

výsledek testu 1	výsledek testu 2	rozdíl test 2 - test 1
3.7	2.8	-0.9
3.5	3.8	0.3
3.0	3.5	0.5
2.7	2.5	-0.2
3.3	NA	NA
4.0	2.5	-1.5
NA	1.8	NA
3.2	2.5	-0.7
2.4	3.5	1.1
0.7	2.1	1.4
2.1	2.3	0.2
3.4	2.8	-0.6
2.9	2.5	-0.4
0.6	2.1	1.5
NA	2.5	NA
3.6	3.3	-0.3
NA	NA	NA
3.7	2.0	-1.7
1.8	2.7	0.9
3.9	3.5	-0.4
2.3	1.5	-0.8
NA	NA	NA
3.9	3.0	-0.9
3.3	NA	NA
NA	2.7	NA
2.3	3.4	1.1
1.6	2.5	0.9
3.6	2.7	-0.9
3.7	3.5	-0.2
2.4	2.8	0.4
2.8	3.5	0.7
2.9	3.5	0.6
2.7	NA	NA
2.3	2.9	0.6
NA	2.0	NA
2.1	2.8	0.7
3.5	3.3	-0.2



Na hladině 5% otestujte, zda jsou výsledky testu 1 srovnatelné s výsledky testu 2, a zda je mezi výsledky nějaká závislost. Co se dá z výsledku usuzovat?

Nápověda:

- průměr z testu 1 $\bar{t_1} \approx 2,835$,
- výběrová směrodatná odchylka výsledků testu 1 $s_1 \approx 0,882$,
- průměr z testu 2 $\bar{t_1} = 2,775$,
- výběrová směrodatná odchylka výsledků testu 2 $s_2 \approx 0,585$,
- průměrný rozdíl (test 2 - test 1) $\bar{d} \approx 0,0429$, výběrová směrodatná odchylka rozdílů $s_d \approx 0,861$,
- korelační koeficient $r = 0,411$.

Řešení:

Jedná se o párová pozorování, tedy musíme použít párový test. Můžeme váhat mezi parametrickým t-testem a nějakým neparametrickým testem, např. Wilcoxonovým testem. Proti normalitě mluví to, že data se nacházejí se v omezeném intervalu $< 0; 4 >$ bodů, a nelze asi určit typickou hodnotu, které budou nabývat. Naproti tomu víme, že výsledek je dán jako součet bodů dosažených v jednotlivých testových otázkách, takže pokud by byly jednotlivé otázky nezávislé, lze ze znalosti centrální limitní věty očekávat, že součet se bude blížit k normalitě (i když pro malý počet otázek jen pomalu).

Zvolíme nejprve párový t-test.

Budeme předpokládat, že $d\ N(\mu, \sigma^2)$. Máme k dispozici 28 úplných pozorování.

Za nulové hypotézy budeme očekávat, že výsledky v obou testech budou srovnatelné, tedy že rozdíly budou nulové:

$$H_0 : d = 0$$
$$H_A : d \neq 0$$

Testová statistika

$$t = \frac{\bar{d} - 0}{s_d}\sqrt{28}$$
$$t = \frac{0,0429}{0,861}\sqrt{28}$$
$$\approx 0,264$$

Kritická hodnota představovaná $97,5\%$ kvantilem t-rozdělení je $q_{t_{28-1}}(1-\alpha/2) \approx 2,05$

Protože $|t| \leq q_{t_{27}}(0,975)$, H_0 na hladině 5% nezamítáme. P-hodnota je $2(1 - F_{t_{27}}(0,264)) \approx 0,794$.

Pokud bychom použili párový Wilcoxonův test a předpokládali bychom, že rozdíly pocházejí z nějakého symetrického rozdělení s mediánem m, tedy $d\ F(m)$, testovali bychom stejnou nulovou hypotézu:

$$H_0 : d = 0$$
$$H_A : d \neq 0$$

Ani tento test by nulovou hypotézu nezamítl. P-hodnota by v tomto případě vyšla velmi podobná: $p \approx 0,776$.

Z výsledků usuzujeme, že mezi výsledky testů nebyl významný rozdíl, což může např. znamenat, že náročnost obou testů byla srovnatelná.

Dále budeme testovat, zda výsledek v prvním testu souvisel s výsledkem ve druhém testu. Protože výsledky jsou kvantitativní (číselné) a ne kvalitativní (např. prospěl/neprospěl), můžeme k testu závislosti použít test výběrového korelačního koeficientu.

Testová statistika t je

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0,411\sqrt{28-2}}{\sqrt{1-0,411^2}} \approx 2,09$$

a kritická hodnota je $q_{t_26}(0,975) = 2,06$. Protože $|t| > q_{t_26}(0,975)$, zamítáme nulovou hypotézu o nulovosti korelačního koeficientu.

3.11 lineární regrese

Příklad: Srovnání výsledků testu 1 a 2 zadaných na cvičení SSL v LS 2014/15 - lineární regrese (Pokračování předchozího příkladu.)

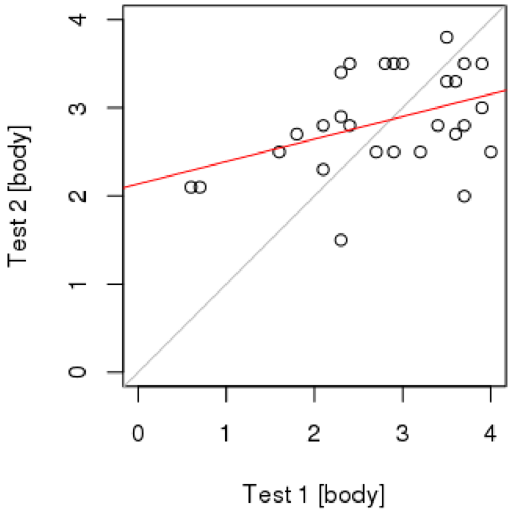
Modelujte výsledek testu 2 pomocí výsledku testu 1.

- formulujte model a vysvětlete jej
- odhadněte parametry modelu (jsou významné?)
- interpretujte parametry modelu
- kolik procent variability v datech je schopen model vysvětlit?
- kdybychom uvažovali obrácený model (výsledek testu 1 vysvětlovaný pomocí výsledku testu 2), jaká by byla hodnota koeficientu příslušného k výsledku testu 2?

Co se dá z výsledků usuzovat?

Nápověda:

- $\widehat{\theta_0} \approx 2,135$, $\widehat{\sigma^2_{\widehat{\theta_1}}} \approx 0,327^2$, $t_{df} \approx 6,523$
- $\widehat{\theta_1} \approx 0,255$, $\widehat{\sigma^2_{\widehat{\theta_1}}} \approx 0,111^2$, $t_{df} \approx 2,296$
- koeficient determinace $r^2_{1,2} = \frac{\widehat{\sigma^2_{t_2}}}{\sigma^2_{t_2}} \approx 0,169$.
- korelační koeficient $r = 0,411$.



Řešení:

- model: $T2_i = \theta_0 + \theta_1 T1_i + \epsilon_i$
- parametry modelu: $\widehat{\theta_0} \approx 2,135$, $\widehat{\sigma^2_{\widehat{\theta_1}}} \approx 0,327^2$, $t_{df=26} \approx 6,523$, $p < 0,00001$. $\widehat{\theta_1} \approx 0,255$, $\widehat{\sigma^2_{\widehat{\theta_1}}} \approx 0,111^2$, $t_{df} \approx 2,296$, $p \approx 0,03$.
- interpretace θ_0: v druhém testu průměrně bodovali i studenti, kteří v prvním testu nezískali žádný bod interpretace θ_1: výsledek prvního testu významně předpovídá výsledek druhého testu, i když významnost není veliká
- koeficient determinace $r^2_{1,2} = \frac{\widehat{\sigma^2_{\widehat{t_2}}}}{\sigma^2_{t_2}} \approx 0,169$, tedy výsledek prvního testu vysvětuje pouze cca 17% variability ve výsledcích druhého testu. Druhý test se patrně týkal jiné látky, než test první.
- kdybychom uvažovali obrácený model (výsledek testu 1 vysvětlovaný pomocí výsledku testu 2), byl by odhad koeficientu příslušného k výsledku testu 2 daný jako $\frac{r^2}{\widehat{\theta_1}} = \frac{0{,}411^2}{0{,}255} \approx 0,664$