Metody matematické statistiky (NMAI 061)

Zdeněk Hlávka

Univerzita Karlova v Praze
Matematicko-fyzikální fakulta
Katedra pravděpodobnosti a matematické statistiky
www.karlin.mff.cuni.cz/~hlavka

Plán přednášky

- Opakování: rozdělení náhodné veličiny.
- Normální rozdělení, centrální limitní věta.
- Odhady, testování hypotéz (t-test).
- Regresní analýza.
- Mnohorozměrné metody.

Doporučená literatura:

Jiří Anděl. *Matematická statistika*, SNTL/Alfa, Praha, 1985.

Jiří Anděl. *Statistické metody*, Matfyzpress, Praha, 1998.

Lenka Komárková, Arnošt Komárek, Vladislav Bína. *Základy analýzy dat a statistického úsudku, s příklady v R*, Skriptum VŠE FM, Jindřichův Hradec, 2006.

Karel Zvára. *Regrese*, Matfyzpress, Praha, 2008.

Doporučený software: R (`www.r-project.org`)

Týden 1

Opakování:

- prostor náhodných jevů,
- náhodná veličina,
- rozdělení,
- základní typy rozdělení (disk., spoj.),
- distribuční funkce,
- hustota,
- číselné charakteristiky (střední hodnota, rozptyl, momenty, kvantily),
- standardní normální rozdělení.

Pravděpodobnostní prostor

- Ω prostor elementárních jevů (všechny možné výsledky),
- ω elementární jevy,
- $\mathcal{A}$ σ-algebra (vhodný systém podmnožin Ω),
- P pravděpodobnostní míra

$(\Omega, \mathcal{A}, P)$ ... pravděpodobnostní prostor

$(\mathbb{R}, \mathcal{B})$ ... reálná čísla s borelovskou σ-algebrou

Náhodná veličina je měřitelné zobrazení $(\Omega, \mathcal{A}, P) \to (\mathbb{R}, \mathcal{B})$.

Příklady: hod mincí, hod kostkou, počasí, čas mezi událostmi, doba výpočtu, quincunx ...

Nezávislost

Náhodné jevy $A, B \subset \Omega$ nazýváme **nezávislé**, pokud

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

Jestliže jsou jevy A, B nezávislé a $P(A) > 0$, $P(B) > 0$, pak podmíněná pravděpodobnost

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

a podobně $P(B \mid A) = P(B)$.

Náhodné veličiny X a Y jsou **nezávislé** pokud jevy $\{X \le a\}$ a $\{Y \le b\}$ jsou nezávislé pro všechna a a b, tj. pokud

$$F(a,b) = P(\{X \le a\} \cap \{Y \le b\}) = P(\{X \le a\})P(\{Y \le b\}) = F_X(a)F_Y(b).$$

Distribuční funkce

Distribuční funkce náhodné veličiny X je

$$F(x) = P(X \le x), \qquad x \in \mathbb{R}$$

Distribuční funkce udává na prostoru $(\mathbb{R}, \mathcal{B})$ pravděpodobnostní míru, které se říká **rozdělení náhodné veličiny** X.

Z definice distribuční funkce je zřejmé, že $0 \le F(x) \le 1$ a pro $x_1 < x_2$ je $F(x_1) \le F(x_2)$ (distribuční funkce je neklesající). Lze odvodit i $\lim_{x \to -\infty} F(x) = 0$ a $\lim_{x \to \infty} F(x) = 1$.

Pomocí distribuční funkce můžeme snadno spočítat pravděpodobnost, že náhodná veličina padne do libovolného intervalu:

$$P(X \in (a, b\rangle) = P(X \le b) - P(X \le a) = F(b) - F(a).$$

Model náhodných dějů

Pravděpodobnostní rozdělení se často používají jako popis náhodných dějů, např.:

diskrétní (F je skokovitá funkce)
- alternativní (Bernoulli),
- binomické,
- Poissonovo,

spojitá (tj. existuje hustota f tak, že $F(x) = \int_0^x f(u)du$)
- exponenciální,
- Laplace, Cauchy, Weibull, Pareto, Erlang, ...
- normální (Gaussovo)

Příklad: Quincunx: binomické rozdělení umožňuje jednoduchý výpočet pravděpodobností a předpovědních intervalů (předpovídání).

R:

- pbinom
- dbinom
- rbinom
- qbinom

Později uvidíme, že binomické rozdělení lze pomocí centrální limitní věty dobře aproximovat normálním rozdělením.

Hustota

Rozdělení *spojité* náhodné veličiny se nejčastěji určuje **hustotou**.

Náhodná veličina X s distribuční funkcí $F(x)$ má hustotu $f(x)$, pokud

$$F(x) = \int_{-\infty}^{x} f(t)dt.$$

Hustota jednoznačně (a názorně) určuje rozdělení spojité náhodné veličiny.

Anglicky: density, probability density function, pdf.

Základní vlastnosti hustoty plynou z vlastností pravděpodobnosti. Pro $a < b$ máme:

$$\begin{aligned} P(X \in (a, b\rangle) &= P(X \in (-\infty, b\rangle) - P(X \in (-\infty, a\rangle) \\ &= P(X \leq b) - P(X \leq a) \\ &= F(b) - F(a) = \int_{-\infty}^{b} f(t)dt - \int_{-\infty}^{a} f(t)dt \\ &= \int_{a}^{b} f(t)dt. \end{aligned}$$

Hustota je zřejmě vždy nezáporná a $\int_{-\infty}^{+\infty} f(x)dx = 1$.

Příklad: Víme, že n.v. X má standardní normální rozdělení $N(0, 1)$, pokud má hustotu:

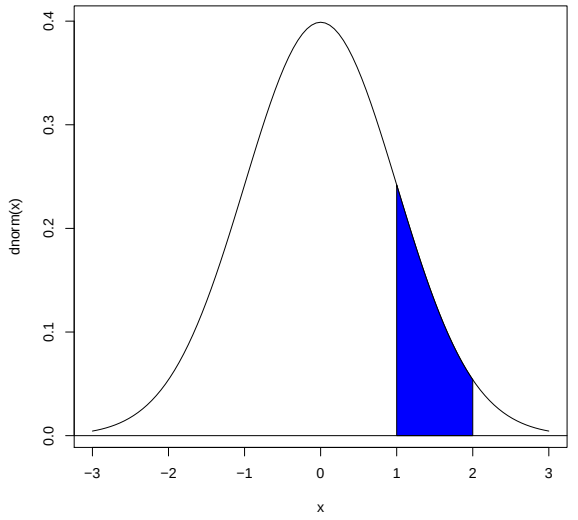
$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\{-x^2/2\}.$$

Distribuční funkce:

$$F(x) = \int_{-\infty}^{x} f(t)dt = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\{-t^2/2\}dt.$$

Jednoduše můžeme spočítat např.:

$$\begin{aligned} P(X \in (-1, 1)) &= \int_{-1}^{1} f(x)dx = F(1) - F(-1) = 2F(1) - 1, \\ P(X \in (0, 2)) &= \int_{0}^{2} f(x)dx = F(2) - F(0) = F(2) - 0.5. \end{aligned}$$



Charakteristiky rozdělení náhodné veličiny

Rozdělení náhodné veličiny je kompletně popsané distribuční funkcí (případně hustotou nebo pravděpodobnostní funkcí).

Zjednodušeně se rozdělení náhodných veličin popisuje pomocí vhodných měr polohy (například střední hodnota, medián, kvantily) a pomocí vhodných měr variability (například rozptyl, směrodatná odchylka, mezikvartilové rozpětí).

Důležité jsou zejména:

- momenty (střední hodnota, rozptyl a podobně),
- kvantily (například medián).

Mějme náhodnou veličinu X s distribuční funkcí $F(x)$, pak **střední hodnota** (expectation) n.v. X je:

$$EX = \int_{-\infty}^{+\infty} x dF(x) = \begin{cases} \int xf(x)dx & \text{pro spojité } X, \\ \sum_{i=1}^{+\infty} x_i P(X = x_i) & \text{pro diskrétní } X. \end{cases}$$

Střední hodnota transformované náhodné veličiny $g(X)$ je:

$$EX = \int_{-\infty}^{+\infty} g(x) dF(x) = \dots$$

Například **rozptyl** (variance) spojité n.v. X s hustotou $f(x)$ je:

$$Var(X) = E\{(X - EX)^2\} = \int_{-\infty}^{+\infty} (x - EX)^2 f(x) dx.$$

Obecně zavádíme k-tý **moment** náhodné veličiny X:

$$EX^k = \int_{-\infty}^{+\infty} x^k dF(x)$$

a k-tý **centrální moment** n.v. X:

$$\mu_k = E(X - EX)^k = \int_{-\infty}^{+\infty} (x - EX)^k dF(x).$$

Rozptyl je tedy druhý centrální moment ($Var(X) = \sigma^2 = \mu_2$).

Směrodatná odchylka (standard deviation) je $s.d. = \sqrt{Var(X)}$.

Šikmost (skewness) se definuje jako μ_3/σ^3 (míra „nesymetrie“).

Špičatost (kurtosis) se definuje jako μ_4/σ^4.

Příklad: Pro $X \sim N(0, 1)$, tj. standardní normální rozdělení máme:

$$EX = \int xf(x)dx = \int x \frac{1}{\sqrt{2\pi}} \exp\{-x^2/2\} dx = \dots$$

$$Var\, X = \int (x - EX)^2 \frac{1}{\sqrt{2\pi}} \exp\{-x^2/2\} dx = \dots$$

Střední hodnota je míra polohy.

Rozptyl (nebo směrodatná odchylka) je míra „rozptýlenosti“ (nebo měřítka).

Šikmost je míra nesymetrie rozdělení náhodné veličiny.

Pravidla pro počítání se středními hodnotami

Máme náhodné veličiny X a Y a konstanty a a b. Pak

- $Ea = a$,
- $EaX = \int axdF(x) = a\int xdF(x) = aEX$,
- $E(a + bX) = a + bEX$,
- $E(X + Y) = EX + EY$.

Pravidla pro počítání s rozptylem

- $Var(a) = 0$,
- $Var(aX) = a^2\, Var(X)$,
- $Var(b + X) = Var(X)$,
- $Var(X + Y) = Var(X) + 2E(X - EX)(Y - EY) + Var(y)$.

(Pravidla lze snadno ověřit pomocí definice střední hodnoty.)

Víme, že $P(X \in (a, b)) = F(b) - F(a)$. Pokud bychom chtěli najít interval (a, b) takový, že $P(X \in (a, b)) = 0.95$, můžeme zvolit a a b například tak, aby $F(b) = 0.975$ a $F(a) = 0.025$.

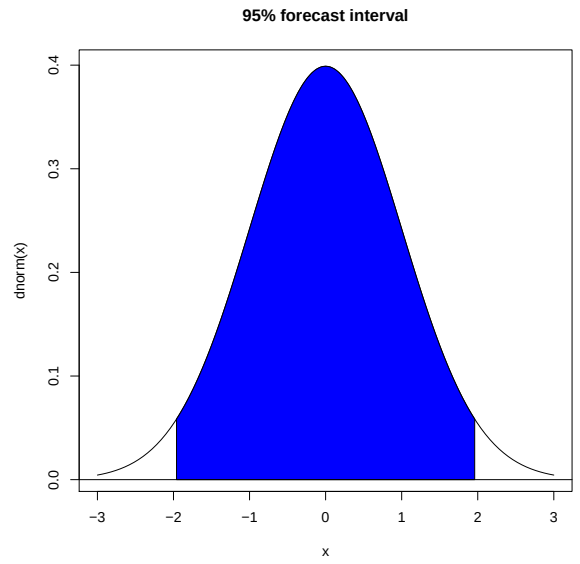
Při „předpovídání“ tedy potřebujeme vědět, ve kterých bodech nabývá distribuční funkce jistých hodnot.

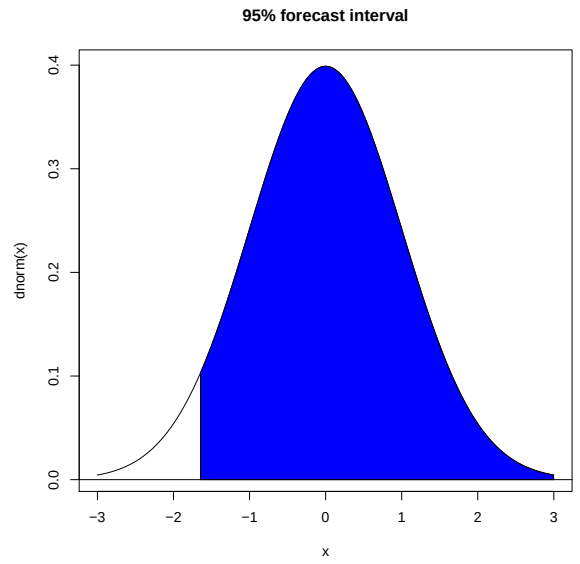
Pro náhodnou veličinu X s d.f. $F(x)$ a pro $\alpha \in (0, 1)$, je x_α tzv. α-**kvantil** rozdělení n.v. X, pokud

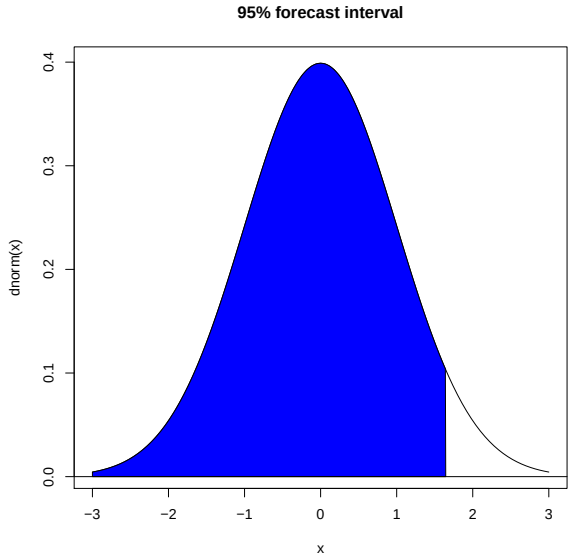
$$F(x_\alpha) = P(X \leq x_\alpha) = \alpha.$$

Příklad: Má-li X rozdělení $N(0, 1)$, pak $P(X \leq -1.645) \doteq 0.05$. Hodnota -1.645 je tedy 0.05-kvantil rozdělení $N(0, 1)$.

Příklad: Má-li X rozdělení $N(0, 1)$, pak $P(X \leq 1.96) \doteq 0.975$. Hodnota 1.96 je tedy 0.975-kvantil rozdělení $N(0, 1)$.







Důležitost normálního rozdělení plyne zejména z tzv. centrální limitní věty:

Věta: Nechť $\{X_1, X_2, \dots\}$ je posloupnost i.i.d. (nezávislých a stejně rozdělených) náhodných veličin s $EX_i = \mu$ a $Var(X_i) = \sigma^2 < +\infty$.

Pak pro $n \to \infty$ náhodná veličina $\sqrt{n}(\overline{X}_n - \mu)/\sigma$ konverguje (v distribuci) k normálnímu rozdělení $N(0, 1)$:

$$\sqrt{n}\left\{\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) - \mu\right\} \xrightarrow{D} N(0,\ \sigma^2).$$

Konvergence v distribuci k F = konvergence distribučních funkcí (v bodech spojitosti F)

Nezávislost náhodných veličin X_1 a X_2 = hodnoty náhodné veličiny X_1 neovlivňují rozdělení (distribuční funkci) X_2 = sdružená distribuční funkce je součin jednotlivých (marginálních) distribučních funkcí

Týden 2

Téma:

- kvantily,
- normální rozdělení,
- centrální limitní věta a její použití,
- náhodný výběr,
- bodový odhad,
- nestrannost,
- konzistence,
- příklad: konstrukce konfidenčního intervalu pro střední hodnotu.

Kvantil

α kvantil náhodné veličiny X je číslo x_α, které splňuje:

$$P(X \le x_\alpha) = \alpha.$$

Některé kvantily nemusí být definovány jednoznačně a pro diskrétní náhodné veličiny nemusí některé kvantily existovat.

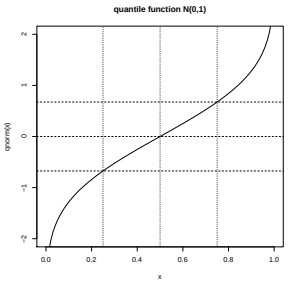
Obecněji (a jednoznačně) lze α-kvantil definovat např.

$$x_\alpha = \inf\{x : P(X \le x_\alpha) \ge \alpha\}.$$

Kvantilová funkce

obecně: F distribuční funkce $\longrightarrow F^{-1}$ je kvantilová funkce (pokud existuje)

Např. medián (50% kvantil), horní a dolní kvartil (25% a 75% kvantil), decily, percentily, atd.



α kvantil rozdělení N(0,1) budeme označovat u_α (v R: qnorm()).

Příklad: Nechť náhodná veličina X má standardní normální rozdělení. Z pravidel pro počítání středních hodnot víme, že $E(\mu + \sigma X) = \mu$ a $Var(\mu + \sigma X) = \sigma^2$. Jaké **rozdělení** ale má náhodná veličina $\mu + \sigma X = Y$?

Předpokládejme, že $\sigma > 0$. Podle **věty o hustotě transformované náhodné veličiny** (Anděl MS, Věta 3/46): „ *Nechť X má spojitou distribuční funkci $F(x)$. Předpokládejme, že $F'(x) = f(x)$ existuje všude s výjimkou nanejvýš konečně mnoha bodů. Budiž $t(x)$ ryze monotónní funkce, která má všude derivaci. Označme τ inverzní funkci k t. Pak náhodná veličina $Y = t(X)$ má hustotu $g(y) = f\{\tau(y)\}|\tau'(y)|$.*“ máme $X = (Y - \mu)/\sigma = \tau(Y)$ a tedy

$$g(y) = \frac{1}{\sqrt{2\pi}} \exp\{(y-\mu)^2/(2\sigma^2)\}/\sigma.$$

Normální rozdělení $N(\mu, \sigma^2)$
Řekneme, že náhodná veličina Y má rozdělení $N(\mu, \sigma^2)$, pokud má hustotu:

$$\phi(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-(y-\mu)^2/(2\sigma^2)\}.$$

Význam parametrů (protože $Y = \mu + \sigma X$, kde $X \sim N(0,1)$):

- $\mu = EY$, tj. střední hodnota,
- $\sigma^2 = Var(Y)$, tj. rozptyl.

Pro α-kvantil y_α n.v. $Y \sim N(\mu, \sigma^2)$ platí, že:

$$\alpha = P(Y \le y_\alpha) = P(\mu + \sigma X \le y_\alpha) = P(X \le (y_\alpha - \mu)/\sigma)$$

a proto $(y_\alpha - \mu)/\sigma = u_\alpha$ a tedy $y_\alpha = \mu + \sigma u_\alpha$ (proto jsou v tabulkách uvedeny pouze kvantily N(0,1)).

Příklad: Tvar normálního rozdělení: `curve(dnorm(x))`

Příklad: Pravidlo σ, 2σ, 3σ, ...

$P(Y \in (\mu - \sigma, \mu + \sigma)) = \cdots =$ `pnorm(1)-pnorm(-1) = 1-2*pnorm(-1)`

Důležitost normálního rozdělení plyne zejména z tzv. centrální limitní věty:

Věta: Nechť $\{X_1, X_2, \dots\}$ je posloupnost i.i.d. (nezávislých a stejně rozdělených) náhodných veličin s $EX_i = \mu$ a $Var(X_i) = \sigma^2 < +\infty$.

Pak pro $n \to \infty$ náhodná veličina $\sqrt{n}(\overline{X}_n - \mu)/\sigma$ konverguje (v distribuci) k normálnímu rozdělení $N(0,1)$:

$$\sqrt{n}\left\{\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) - \mu\right\} \xrightarrow{D} N(0,\ \sigma^2).$$

Konvergence v distribuci k F = konvergence distribučních funkcí (v bodech spojitosti F)

Nezávislost náhodných veličin X_1 a X_2 = hodnoty náhodné veličiny X_1 neovlivňují rozdělení (distribuční funkci) X_2 = sdružená distribuční funkce je součin jednotlivých (marginálních) distribučních funkcí

S nezávislými veličinami X_1 a X_2 se dobře počítá:

$$E(X_1X_2) = E(X_1)E(X_2)$$

$$Cov(X_1, X_2) = 0$$

$$Var(X_1 + X_2) = Var\, X_1 + 2\, Cov(X_1, X_2) + Var\, X_2 = Var\, X_1 + Var\, X_2$$

$$Var(X_1 - X_2) = Var\, X_1 - 2\, Cov(X_1, X_2) + Var\, X_2 = Var\, X_1 + Var\, X_2$$

Veličiny $X_1, \dots, X_n$ jsou navzájem nezávislé, pokud je nezávislá každá jejich podmnožina (je to něco jiného než „po dvou“ nezávislé):

$$E(X_1X_2 \dots X_n) = E(X_1)E(X_2) \dots E(X_n)$$

$$Var\left(\sum X_i\right) = \sum Var(X_i)$$

CLV dobře aproximuje chování průměru nebo součtu nezávislých náhodných veličin (je to prakticky totéž, protože se to liší jenom známou konstantou).

Příklad: Quincunx: konečná poloha kuličky. Kvantily normálního rozdělení můžeme použít pro pravděpodobnostní předpověď.

Odhad

Cílem je odhadnout potřebné parametry (například střední hodnotu) na základě získaných pozorování.

Definice: Řekneme, že náhodné veličiny $X_1, \dots, X_n$ tvoří náhodný výběr, pokud X_i jsou navzájem nezávislé a mají stejné rozdělení.

Definice: Nechť $X_1, \dots, X_n$ je náhodný výběr z rozdělení s distribuční funkcí $F_\theta(x)$, kde θ je odhadovaný parametr. *Odhadem* nazveme libovolnou funkci $T(X_1, \dots, X_n)$ (důležité je, že funkce $T(.)$ nezávisí na neznámém parametru θ).

Odhady se snažíme zvolit tak, aby měly „dobré vlastnosti“.

Teoretické vlastnosti odhadů

Žádoucí vlastnosti odhadu:

- konzistence
- nestrannost

Míry kvality odhadu:

- vychýlení (bias)
- rozptyl
- MSE

Rozdělení průměru

Příklad: Mějme náhodný výběr o rozsahu n z normálního rozdělení. Jaké je rozdělení výběrového průměru?

Jednoduše lze spočítat střední hodnotu i rozptyl výběrového průměru.

Později si ukážeme, že lineární transformace „zachovává normalitu“ a výběrový průměr má tedy **normální rozdělení** $N(\mu, \sigma^2/n)$.

Příklad: Mějme náhodný výběr o rozsahu n z nějakého jiného rozdělení. Jaké je rozdělení výběrového průměru?

Pokud má zadané rozdělení střední hodnotu (EX), pak je střední hodnota výběrového průměru rovná této střední hodnotě.

Rozdělení výběrového průměru závisí na rozdělení náhodné veličiny (X). Pro rostoucí počet pozorování se rozdělení výběrového průměru rychle blíží rozdělení normálnímu podle CLV.

Navíc existují výpočetně náročné metody (bootstrap, subsampling), které nám umožní aproximovat rozdělení výběrového průměru i bez předpokladu znalosti rozdělení X.

Jednoduchý konfidenční interval

Příklad: konfidenční interval pro střední hodnotu, pokud známe rozptyl.

Quincunx:

1. Spočítáme teoretický rozptyl měřené náhodné veličiny.
2. Díky CLV získáme přibližné normální rozdělení průměru.
3. Pomocí kvantilů standardního normálního rozdělení a jednoduchých algebraických úprav odvodíme konfidenční inteval (náhodný interval, který s danou pravděpodobností překryje neznámou střední hodnotu).

Týden 3

Téma:

- rozdělení odvozená od normálního,
- výběrové charakteristiky.
- vlastnosti průměru a výběrového rozptylu,
- výběr z normálního rozdělení a jeho vlastnosti.
- konfidenční intervaly (jednostranné i oboustranné) pro parametry normálního rozdělení.

χ^2-rozdělení

Nechť jsou $X_1, X_2, \ldots$ a $Y_1, Y_2 \ldots$ nezávislé posloupnosti iid $N(0,1)$ n.v.

χ^2-rozdělení o n stupních volnosti je rozdělení náhodné veličiny

$$Z_n = \sum_{i=1}^{n} X_i^2$$

Značíme $Z_n \sim \chi^2_n$.

Víme, že $EZ_n = n$ a $Var\, Z_n = 2n$ (viz např. Wikipedia)

Kvantily budeme značit $\chi^2_{n;1-\alpha}$

t-rozdělení a F-rozdělení

t-rozdělení o n stupních volnosti: rozdělení podílu standardního normálního (tj. $X \sim N(0,1)$) a odmocniny nezávislého χ^2 rozdělení s n stupni volnosti: $T = X/\sqrt{Z_n/n} \sim t_n$;

$(1-\alpha)$ kvantil budeme značit $t_{n;1-\alpha}$.

Důležité je nezaměňovat kvantily a kritické hodnoty!!

F-rozdělení o n a m stupních volnosti: rozdělení podílu $(W_n/n)/(Z_m/m) \sim F_{n,m}$ dvou nezávislých náhodných veličin s χ^2-rozdělením ($W_n \sim \chi^2_n$, $Z_m \sim \chi^2_m$);

$(1-\alpha)$ kvantil budeme značit $F_{n,m;1-\alpha}$

Častý úkol ve statistice je „něco zjistit“ o parametrech nějakého rozdělení.

Takové „zjišťování“ bývá ve statistice založeno na opakovaných pozorováních nějaké náhodné veličiny. V nejjednodušší situaci můžeme předpokládat, že tato opakovaná pozorování jsou získána pomocí *náhodného výběru*.

Definice: Řekneme, že náhodné veličiny $X_1, \ldots, X_n$ tvoří náhodný výběr, pokud X_i jsou navzájem nezávislé a mají stejné rozdělení.

Naměřeným hodnotám $x_1, \ldots, x_n$ budeme říkat *realizace náhodného výběru*.

Příklad: quincunx, hod mincí (situace před a po)

Mějme náhodný výběr $X_1, \ldots, X_n$. Základní výběrové charakteristiky jsou:

- míry polohy (průměr, medián)
- míry měřítka (výběrový rozptyl, směrodatná odchylka, rozpětí, mezikvartilové rozpětí)
- výběrové kvantily (medián, kvartily, minimum, maximum)
- výběrové momenty a centrální momenty (průměr, rozptyl, šikmost, špičatost)

Pomocí metod teorie pravděpodobnosti je možné odvodit teoretické vlastnosti (pravděpodobnostní rozdělení) výběrových charakteristik, na které se díváme jako na náhodné veličiny.

To někdy bývá matoucí, proto je potřeba důsledně rozlišovat náhodné veličiny (např. průměr $\overline{X}_n$) a jejich realizace (naměřený a vypočtený průměr $\overline{x}_n$).

Příklad:

1/ Střední hodnota výběrového průměru a výběrového rozptylu.

2/ Rozptyl výběrového průměru.

3/ Rozdělení výběrového průměru (bez předpokladu normality): Nechť $X_1, \ldots, X_n$ je náhodný výběr splňující předpoklady centrální limitní věty. Pak rozdělení náhodné veličiny $\sqrt{n}(\overline{X}_n - \mu)$ konverguje v distribuci k $N(0, \sigma^2)$ a rozdělení výběrového průměru lze aproximovat pomocí rozdělení $N(\mu, \sigma^2/n)$.

Rozdělení výběrového průměru a rozptylu za předpokladu normality.

Věta: Nechť $X_1, \ldots, X_n$ je náhodný výběr z $N(\mu, \sigma^2)$. Pak platí:

- $\overline{X}_n \sim N(\mu, \sigma^2/n)$,
- $(n-1)S^2/\sigma^2$ má rozdělení χ^2_{n-1}, je-li $n > 1$ a $\sigma^2 > 0$,
- je-li $n > 1$, jsou veličiny $\overline{X}_n$ a S^2 nezávislé.

Důkaz: viz *Anděl: Matematická statistika, věta 18, strana 82, SNTL, 1985.*

Příklad: Nezávislost $X + Y$ a $X - Y$ za předpokladu normality.

Příklad: Z věty lze např. jednoduše spočítat rozptyl výběrového rozptylu za předpokladu normality.

Intervalový odhad

Mějme náhodný výběr $X_1, \ldots, X_n$ z rozdělení s distribuční funkcí $F_\theta(x)$, $\theta \in \Theta$.

Intervalový odhad je založen na dvou odhadech $T_1(X_1, \ldots, X_n)$ a $T_2(X_1, \ldots, X_n)$ (funkce, které neobsahují θ) a které splňují

$$P\{T_1(X_1, \ldots, X_n) < \theta < T_2(X_1, \ldots, X_n)\} = 1 - \alpha,$$

kde $1 - \alpha$ je spolehlivost (nejčastěji 0.95).

Horní odhad (jednostranný): $P\{\theta < T_h(X_1, \ldots, X_n)\} = 1 - \alpha$.

Dolní odhad: $P\{T_d(X_1, \ldots, X_n) < \theta\} = 1 - \alpha$.

Konstrukce intervalového odhadu

Máme odhad $T(X_1, \ldots, X_n)$ a ze znalosti rozdělení nějaké jeho funkce $h(T, \theta)$ najdeme c_1 a c_2 tak, aby

$$P(c_1 < h(T, \theta) < c_2) = 1 - \alpha$$

a jednoduchými algebraickými úpravami získáme $\tilde{c}_1$ a $\tilde{c}_2$ tak, aby

$$P(\tilde{c}_1 < \theta < \tilde{c}_2) = 1 - \alpha.$$

Příklad: střední hodnota a rozptyl normálního rozdělení.

Konfidenční interval pro střední hodnotu normálního rozdělení

Použitím t-rozdělení získáme oboustranný konfidenční interval:

$$(\overline{X}_n \pm t_{n-1;1-\alpha/2} S_n/\sqrt{n})$$

Jednostranné intervaly:

$$(-\infty, \overline{X}_n + t_{n-1;1-\alpha} S_n/\sqrt{n})$$

$$(\overline{X}_n - t_{n-1;1-\alpha} S_n/\sqrt{n}, \infty)$$

Bez předpokladu normality (pro rozdělení s konečným rozptylem), lze použít přibližný interval založený na CLV:

$$(\overline{X}_n \pm u_{1-\alpha/2} S_n/\sqrt{n}).$$

Podobné intervaly (přesné nebo asymptotické - se σ a kvantily normálního rozdělení) vychází i pokud je rozptyl známý (taková situace je ale v praxi spíše neobvyklá).

Příklad: Interval spolehlivosti pro rozptyl: horní odhad, oboustranný interval (problém volby vhodného kvantilu).

Oboustranný konfidenční interval pro rozptyl normálního rozdělení:

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

Horní odhad (obdobně): $(n-1)S_n^2/\chi^2_{n-1;\alpha}$.

Týden 4

Téma:

- odhadování,
- momentová metoda a delta metoda,
- metoda maximální věrohodnosti (diskrétní i spojitá rozdělení),
- vlastnosti maximálně věrohodného odhadu.

Situace: máme náhodný výběr z rozdělení s distribuční funkcí $F_\theta(x)$ a chceme odhadnout (vektorový) parametr $\theta = (\theta_1, \ldots, \theta_p)$.

Nejobvyklejší metody konstrukce odhadu:

momentová metoda (odhad se konstruuje pomocí srovnání teoretických a výběrových momentů),

metoda maximální věrohodnosti.

Momentová metoda

Princip metody: za odhad se zvolí taková hodnota parametru θ, která vede ke shodě prvních p teoretických a výběrových momentů (buď centrálních nebo necentrálních). Teoretické momenty $m_j(\theta)$ závisí na θ, výběrové momenty $\sum_{i=1}^n X_i^j/n$ jsou funkce náhodného výběru.

Odhad $\hat{\theta}$ získáme řešením soustavy rovnic $m_i(\hat{\theta}) = \sum X_i^j/n$.

Příklad: Odhad parametru exponenciálního rozdělení s hustotou $\lambda e^{-\lambda x}$ pro $x > 0$.

Asymptotické rozdělení $\hat{\theta}$ lze často spočítat použitím CLV a delta-metody.

Delta metoda v jednorozměrném případě

Pokud posloupnost náhodných veličin X_n splňuje

$$\sqrt{n}[X_n - \theta] \xrightarrow{D} N(0, \sigma^2),$$

pak

$$\sqrt{n}[g(X_n) - g(\theta)] \xrightarrow{D} \mathcal{N}(0, \sigma^2[g'(\theta)]^2)$$

pro každou funkci $g(.)$, která má derivaci $g'(\theta) \neq 0$.

Příklad: Asymptotické rozdělení odhadu $\hat{\lambda}$ parametru λ exponenciálního rozdělení.

Delta metoda ve vícerozměrném případě

Pro informaci (mnohorozměrné normální rozdělení bude později):

Pomocí mnohorozměrné centrální limitní věty můžeme získat:

$$\sqrt{n}\left(\hat{\xi} - \xi\right) \xrightarrow{D} N\left(0, \Sigma\right),$$

kde Σ je pozitivně definitní varianční matice.

Pak nám delta metoda dává asymptotické rozdělení věktoru $\hat{\theta} = h(\hat{\xi})$:

$$\sqrt{n}\left(h(\hat{\xi}) - h(\xi)\right) \xrightarrow{D} N\left(0, \nabla h(\beta)^T \cdot \Sigma \cdot \nabla h(\beta)\right).$$

Metoda maximální věrohodnosti

Princip metody: odhad je „nejpravděpodobnější“ hodnota parametru, tj. hodnota parametru, která maximalizuje sdruženou hustotu (nebo pravděpodobnost) pozorovaného náhodného výběru.

Věrohodnostní funkce:

$$L(\theta; X_1, \ldots, X_n) = \prod f_\theta(X_i)$$

(pro diskrétní n.v. použijeme pravděpodobnostní funkcí místo hustoty)

Maximálně věrohodný odhad $\hat{\theta}$ takový, že

$$L(\hat{\theta}; X_1, \ldots, X_n) \geq L(\theta; X_1, \ldots, X_n), \quad \forall \theta \in \Theta.$$

Při hledání maximálně věrohodného odhadu se většinou lépe pracuje s logaritmickou věrohodnostní funkcí.

Logaritmická věrohodnostní funkce:

$$l(\theta; X_1, \ldots, X_n) = \log L(\theta; X_1, \ldots, X_n) = \sum \log f_\theta(X_i)$$

Obvyklý postup: spočítáme derivaci logaritmické věrohodnostní funkce a položíme ji rovnou nule. Vyřešením této soustavy rovnic získáme maximálně věrohodné odhady jednotlivých parametrů.

Příklad: Odhad střední hodnoty exponenciálního rozdělení.

Příklad: Odhad střední hodnoty a rozptylu normálního rozdělení.

Příklad: Odhad střední hodnoty Poissonova rozdělení (diskrétní rozdělení).

Asymptotické rozdělení ML odhadu

Věta: Za jistých předpokladů má maximálně věrohodný odhad $\hat{\theta}$ parametru θ asymptotické rozdělení:

$$n^{1/2}(\hat{\theta} - \theta) \xrightarrow{D} N(0, 1/J(\theta)),$$

kde

$$J(\theta) = -E\left\{\frac{\partial^2 \log f_\theta(X)}{\partial \theta^2}\right\}$$

je Fisherova míra informace o parametru θ, která je obsažena v náhodné veličině X s hustotou $f_\theta(x)$.

Důkaz a všechny předpoklady: viz například Anděl (1985, věta XV.6.10, str. 268)

Pozn.: ($J_n(\theta) = nJ(\theta)$ je Fisherova míra informace o parametru θ, která je obsažena v náhodném výběru $X_1, \ldots, X_n$ z rozdělení s hustotou $f_\theta(x)$).

Asymptotické rozdělení ML odhadu

Poznámky k maximálně věrohodným odhadům:

Pokud věrohodnostní matici maximalizujeme numericky, tak můžeme numericky získat i odhad Fisherovy informace (střední hodnotu odhadneme průměrem a vyjde nám druhá derivace věrohodnostní funkce).

Pokud odhadujeme vektorový parametr, pak má maximálně věrohodný odhad asymptoticky mnohorozměrné normální rozdělení (s nulovou střední hodnotou) a varianční matice je inverze tzv. Fisherovy informační matice.

V praxi se jako odhad varianční matice často používá tzv. sandwichový odhad $J^{-1}(\theta)V(\theta)J^{-1}(\theta)$, který má výhodnější vlastnosti ($V(\theta) = Var\left\{\frac{\partial \log f_\theta(X)}{\partial \theta}\right\}$).

Týden 5

Téma:

- princip testování hypotéz,
- nulová a alternativní hypotéza,
- chyba prvního a druhého druhu,
- hladina testu,
- síla testu (silofunkce),
- jednovýběrový t-test,
- p-hodnota.

Testování hypotéz

hypotéza = výrok o parametru (nebo typu) rozdělení

nulová hypotéza ... H_0

alternativní hypotéza ... H_1

Máme $X_1, \ldots, X_n$ náhodný výběr z rozdělení, jehož distribuční funkce závisí na $\theta \in \Theta$

Chceme otestovat H_0 versus H_1:
$H_0 : \theta = \theta_0$
$H_1 : \theta \in \{\Theta \setminus \theta_0\}$

Testovací kritérium (statistika) $T(X_1, \ldots, X_n)$. Nulovou hypotézu (H_0) *zamítneme* ve prospěch alternativní hypotézy (H_1), pokud testová statistika padne do předem určeného *kritického oboru* K (tj. $T \in K$).

Rozhodování přináší možnost chyby:
chyba 1.druhu = H_0 zamítneme i když platí = $P(T \in K|H_0)$
chyba 2.druhu = H_0 nezamítneme když neplatí= $P(T \notin K|H_1)$

Hladina testu (významnosti) = P(chyba 1.druhu) = α (obvykle $0.05, 0.01, \ldots$)

Síla testu = $P(T \in K|H_1)$

Obvykle požadujeme, aby chyba 1. druhu nebyla moc velká ($P(T \in K|H_0) \leq \alpha$) a přitom byla síla testu co největší.

P-hodnota = hraniční hladina testu, na které ještě zamítáne nulovou hypotézu (tj. nulovou hypotézu zamítáme, pokud je p-hodnota $\leq \alpha$)

Obvyklý (rozumný) postup při odvozování testu:

1. zvolíme rozumnou testovou statistiku T (obvykle: „malá za H_0, velká za H_1“),
2. určíme kritický obor K tak, aby $P(T \in K|H_0) = \alpha$.

Postup při testování: zamítneme H_0, pokud $T \in K$.

Zamítnutí nulové hypotézy znamená: „prokázali jsme, že platí H_1“.

Nezamítnutí nulové hypotézy znamená: buď H_0 platí nebo nemáme dost pozorování, abychom mohli H_0 vyvrátit (v praxi se vyplatí naplánovat experiment tak, abychom „zajímavý rozdíl“ prokázali s dostatečně velkou pravděpodobností (sílou)).

Jednovýběrový t-test (one-sample t-test)

$X_1, \ldots, X_n$ náhodný výběr z $N(\mu, \sigma^2)$, rozptyl σ^2 neznáme.

$H_0 : \mu = \mu_0$
$H_1 : \mu \neq \mu_0$ (oboustranná alternativa)

Víme, že za platnosti H_0 má $T = \sqrt{n}(\overline{X}_n - \mu_0)/S$ rozdělení t_{n-1}.

$P(|T| \geq t_{n-1;1-\alpha/2}|H_0) = \alpha$ a kritický obor je tedy $(-\infty, t_{n-1;\alpha/2}) \cap (t_{n-1;1-\alpha/2}, \infty)$.

p-hodnota $= P(|T_{n-1}| > t)$, kde $T_{n-1} \sim t_{n-1}$ a $t = \sqrt{n}(\overline{x}_n - \mu_0)/s$ je pozorovaná hodnota testové statistiky.

Příklad

Příklad: Opakované vážení vzorku:

15.23 15.21 15.19 15.16 15.26 15.22 15.23 15.26 15.23 15.29

Chceme otestovat, jestli skutečná hmotnost vzorku je $\mu_0 = 15.2$.

testová statistika $T = \ldots$

kritická hodnota (určuje kritický obor)

p-hodnota

Jednostranný jednovýběrový t-test (one-sided one-sample t-test)

$X_1, \ldots, X_n$ náhodný výběr z $N(\mu, \sigma^2)$, rozptyl σ^2 neznáme.

$H_0 : \mu = \mu_0$ $H_1 : \mu > \mu_0$ (jednostranná alternativa)

H_0 zamítáme, pokud $T \geq t_{n-1;1-\alpha}$

Porovnání s dvouvýběrovým testem: na jedné straně má jednostranný test větší sílu, ale na druhé straně (pokud vyjde $\overline{X}_n < \mu_0$) nulovou hypotézu vůbec zamítnout nemůžeme.

Párový t-test (paired t-test)

Dvojice (páry) pozorování na každém objektu (X_i, Y_i), $i = 1, \ldots, n$.

Platí: $Var(X_i - Y_i) = Var(X_i) + Var(Y_i) - 2\,Cov(X_i, Y_i)$ (čím jsou pozorování v páru závislejší, tím menší je rozptyl jejich rozdílu).

$H_0 : EX_i = EY_i + \Delta$ je totéž jako $H_0 : E(X_i - Y_i) = \Delta$.

Můžeme tedy použít jednovýběrový t-test na $Z_i = X_i - Y_i$.

Wilcoxonův test (signed rank test)

$X_1, \ldots, X_n$ náhodný výběr

$H_0 : \text{med}(X) = \mu_0$
$H_1 : \text{med}(X) \neq \mu_0$

Testová statistika je založena pouze na pořadích a není tedy citlivá na odlehlá pozorování.

1. $Z_i = X_i - \mu_0$,
2. Z_i se seřadí podle absolutních hodnot (R_i - pořadí i-tého pozorování),
3. $S^+ = \sum_{i:Z_i>0} R_i$, $S^- = \sum_{i:Z_i<0} R_i$

Za platnosti H_0 by S^+ a S^- měly být podobné. Rozdělení S^+ za předpokladu symetrie kolem μ_0 lze snadno vypočítat (i když je výpočetně náročné).

Hypotézy o rozptylu

$X_1, \ldots, X_n$ náhodný výběr z $N(\mu, \sigma^2)$.

$H_0 : \sigma^2 = \sigma_0^2$
$H_1 : \sigma > \sigma_0^2$ (jednostranná alternativa)

Nulovou hypotézu zamítneme, pokud bude výběrový rozptyl S^2 moc velký ($S^2 > c$).

Za platnosti H_0 víme, že

$$\frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}$$

Kritickou hodnotu c teď snadno spočítáme tak, aby $P(S^2 > c|H_0) = \alpha$.

P-hodnota = ?

Příklad: Odvození testu proti oboustranné alternativě $H_1 : \sigma > \sigma_0^2$.

Hypotéza shody rozptylů

$X_1, \ldots, X_n$ a $Y_1, \ldots, Y_m$ dva nezávislé náhodné výběry z $N(\mu_1, \sigma_1^2)$ a $N(\mu_2, \sigma_2^2)$.

$H_0 : \sigma_1^2 = \sigma_2^2$
$H_1 : \sigma_1 \neq \sigma_2^2$

Nulovou hypotézu zamítneme, pokud S_1^2/S_2^2 bude daleko od 1 (rozdělení podílu za H_0 umíme snadno spočítat).

Kritickou hodnotu spočítáme ze známého rozdělení S_1^2/S_2^2 za nulové hypotézy tak, aby $P(\text{zamítneme } H_0|H_0 \text{ platí}) = \alpha$.

Odvození kritické hodnoty:

$$\frac{(n-1)S_1^2}{\sigma_1^2} \sim \chi^2_{n-1}, \quad \frac{(m-1)S_2^2}{\sigma_2^2} \sim \chi^2_{m-1},$$

kde S_1 a S_2 jsou nezávislé (spočítané z nezávislých náhodných výběrů)

$$\frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} \sim F_{n-1,m-1}$$

Za H_0 máme $\sigma_1^2 = \sigma^2$ a tedy $S_1^2/S_2^2 \sim F_{n-1,m-1}$

Kritické hodnoty nyní snadno určíme tak, aby

$$\alpha = P\left(\frac{S_1^2}{S_2^2} < c_1 \text{ nebo } \frac{S_1^2}{S_2^2} > c_2 | H_0\right)$$

$c_1 = F_{n-1,m-1;\alpha/2}$ a $c_1 = F_{n-1,m-1;1-\alpha/2}$

Poznámky:

Kritické hodnoty a kvantily jsou uvedeny v tabulkách, ale v každých tabulkách se může používat jiné značení (POZOR!)

Statistické programy prakticky vždy uvádí p-hodnotu (pak kritickou hodnotu nepotřebujeme).

Statistická významnost není totéž jako praktická důležitost (s dostatečně velkým počtem pozorování lze statisticky prokázat i naprosto nezajímavý rozdíl).

V praxi se doporučuje experimenty plánovat tak, aby rozsah výběru byl rozumný (s ohledem na sílu testu) a tak, aby vyhodnocení dat (primární analýza) bylo co nejjednodušší.

Pro diskrétní proměnné se často používá test nezávislosti v kontingenční tabulce.

V mírně nestandardních situacích lze často použít test poměrem věrohodností (likelihood ratio test).

Týden 6

Téma:

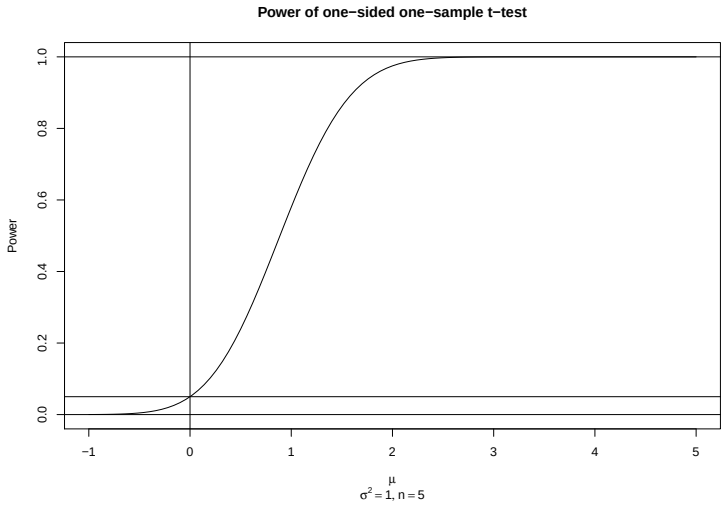
- síla testu (silofunkce),
- párový a dvouvýběrový t-test,
- ověřování předpokladů:
  - shoda rozptylů,
  - normalita,
  - nezávislost.
- princip použití pořadových testů (podrobněji na cvičení).

Síla jednovýběrového t-testu

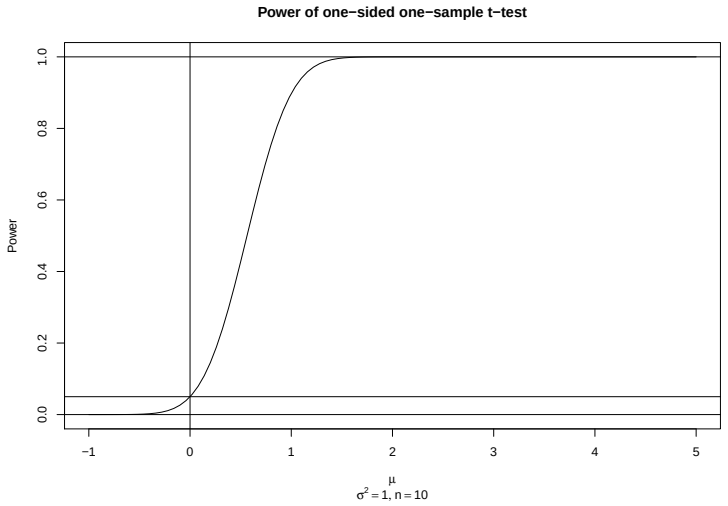
Síla testu je pravděpodobnost, že testová statistika překročí kritickou hodnotu (tj. pravděpodobnost zamítnutí nulové hypotézy).

Za předpokladu normality lze sílu jednovýběrového t-testu vypočítat jako funkci skutečné střední hodnoty μ, rozptylu σ^2 a počtu pozorování n.

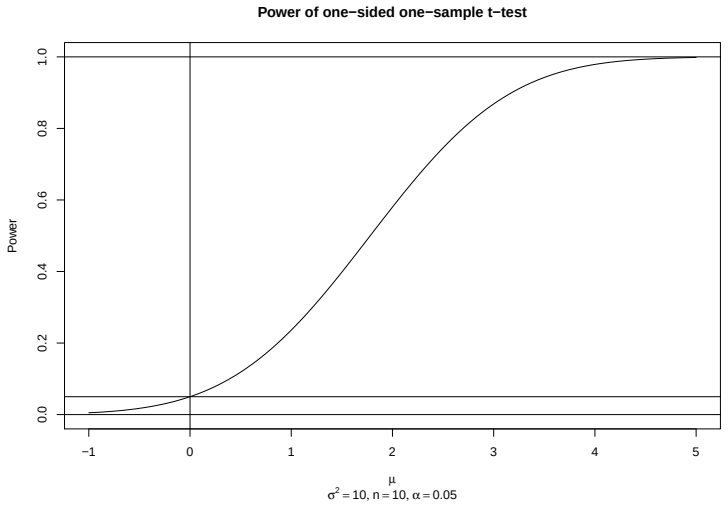
Síla jednostranného jednovýběrového t-testu



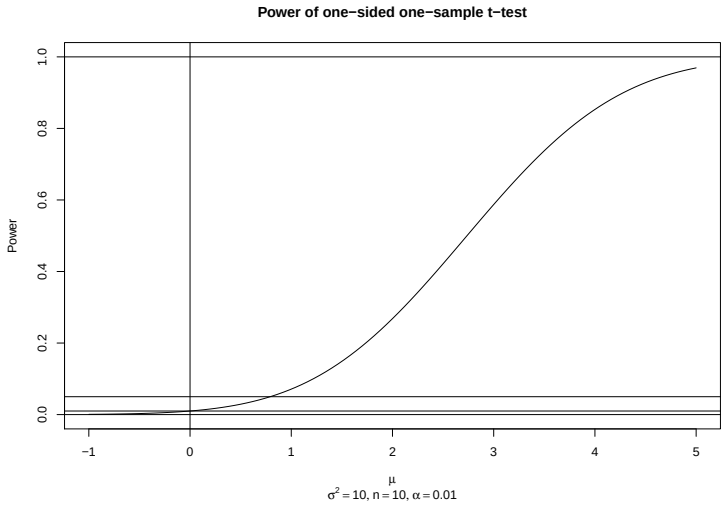
Síla jednostranného jednovýběrového t-testu



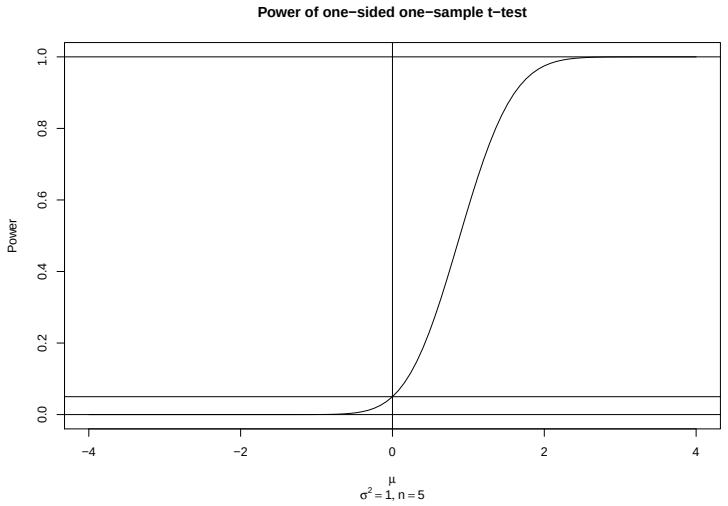
Síla jednostranného jednovýběrového t-testu



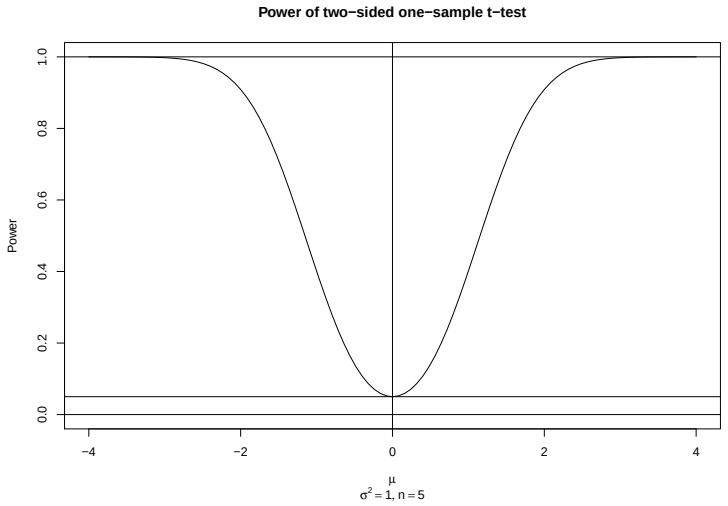
Síla jednostranného jednovýběrového t-testu



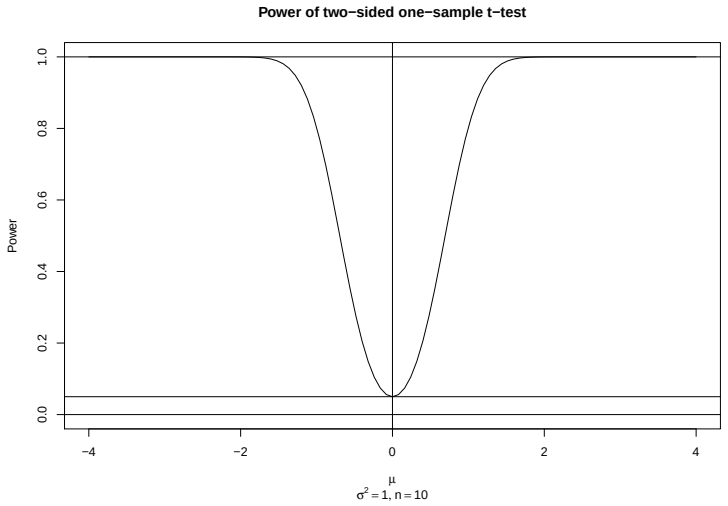
Síla jednostranného jednovýběrového t-testu



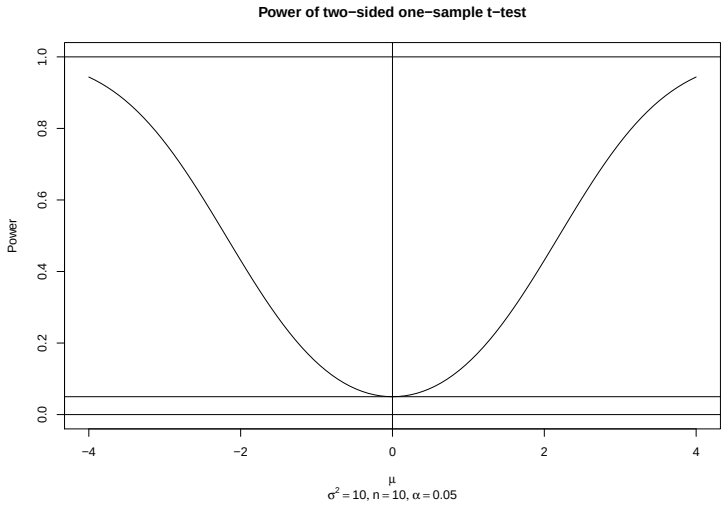
Síla oboustranného jednovýběrového t-testu



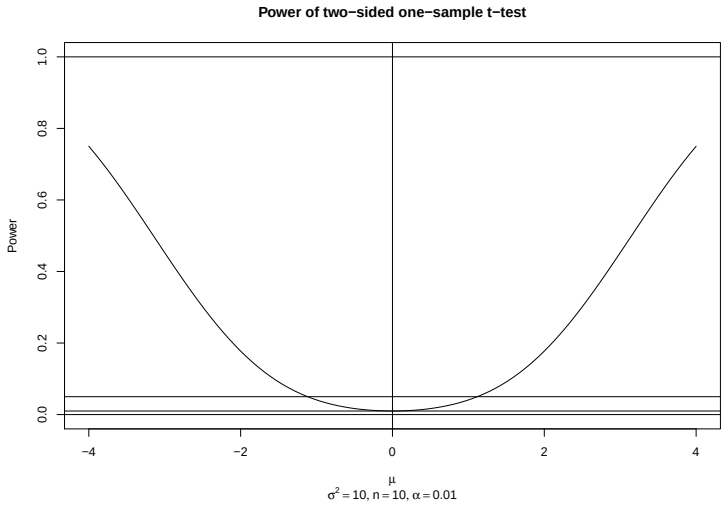
Síla oboustranného jednovýběrového t-testu



Síla oboustranného jednovýběrového t-testu



Síla oboustranného jednovýběrového t-testu



Předpoklady pro jednovýběrový a párový t-test

Předpoklady pro jednovýběrový t-test:

- normalita,
- nezávislost pozorování.

Díky CLV nemívá porušení předpokladu normality závažný vliv na vlastnosti jednovýběrového t-testu (ten v praxi selhává pouze v přítomnosti velkých odlehlých pozorování).

V praxi lze vlastnosti t-testu (sílu) zlepšit použitím vhodné transformace dat (Box-Cox, logaritmus, odmocnina), pak se ale testuje hypotéza o střední hodnotě transformovaných dat (to obvykle vůbec nevadí např. u párového testu).

Předpoklady pro použití párového testu jsou splněné, pokud rozdíly $(Y_i - X_i)$ splňují předpoklady pro použití jednovýběrového testu.

Předpoklady pro jednovýběrový a párový pořadový test

Wilcoxonův test lze použít jako náhradu za t-test pro nenormální data (je to vhodné, pokud hrozí přítomnost velkých odlehlých pozorování).

Předpoklady pro jednovýběrový Wilcoxonův test:

- symetrie kolem testované hodnoty,
- nezávislost pozorování.

Při porušení předpokladu symetrie nemusí být zcela jasné, jestli nulovou hypotézu nezamítáme spíše kvůli nesymetrii dat.

Příklad: Chování t-testu a Wilcoxonova testu v přítomnosti jediného hodně velkého odlehlého pozorování.

Plánování experimentu

V praxi se doporučuje experiment naplánovat tak, aby:

- bylo možné výsledky vyhodnotit jednoduše (je vhodné zajistit např. nezávislost jednotlivých měření),
- test měl rozumnou sílu proti rozumným alternativám (rozumná většinou znamená 80% pravděpodobnost zamítnutí nulové hypotézy při vhodně zvolené prakticky zajímavé alternativě).

Podmínkou ovšem je, abychom přesně věděli, co vlastně chceme zkoumat (testovat)!

Dvouvýběrový t-test

Máme dva **nezávislé** náhodné výběry $X_1, \ldots, X_n$ a $Y_1, \ldots, Y_m$ z $N(\mu_x, \sigma^2)$ a $N(\mu_y, \sigma^2)$.

$H_0 : \mu_x = \mu_y + \Delta$
$H_1 : \mu_x \neq \mu_y + \Delta$

Přirozená testová statistika je založená na rozdílu průměrů vyděleném odhadem směrodatné odchylky (rozdílu průměrů):

$$T = \frac{\bar{X}_n - \bar{Y}_m - \Delta}{S_{\text{pooled}}\sqrt{\frac{1}{n} + \frac{1}{m}}},$$

kde $S^2_{\text{pooled}} = \{(n-1)S^2_X + (m-1)S^2_Y\}/(n+m-2)$.

Za platnosti H_0 má testová statistika rozdělení t_{n+m-2}. Nulovou hypotézu tedy zamítneme, pokud $|T| > t_{n+m-2;1-\alpha/2}$.

Dvouvýběrový t-test: jednostranné varianty

Levostranná alternativa:

$H_0 : \mu_x = \mu_y + \Delta$
$H_L : \mu_x < \mu_y + \Delta$

Za platnosti H_0 má testová statistika rozdělení t_{n+m-2}. Nulovou hypotézu zamítáme, pokud $T < -t_{n+m-2;1-\alpha}$.

Pravostranná alternativa:

$H_0 : \mu_x = \mu_y + \Delta$
$H_L : \mu_x > \mu_y + \Delta$

Za platnosti H_0 má testová statistika rozdělení t_{n+m-2}. Nulovou hypotézu zamítáme, pokud $T > t_{n+m-2;1-\alpha}$.

Dvouvýběrový t-test: jednostranné varianty

POZOR: jednostrannou alternativu si musíme vybrat předem.

Pokud si jednostrannou alternativu zvolíme až podle naměřených hodnot, tak bude mít jednostranný test ve skutečnosti dvakrát vyšší pravděpodobnost chyby prvního druhu.

Dvouvýběrový vs. párový t-test

POZOR: použití párového nebo dvouvýběrového testu závisí na způsobu sběru dat. Pokud mají oba výběry stejný rozsah (typická situace ve zkouškové písemce i v praxi), tak žádný program sám správný test nevybere!

Chybné použití dvouvýběrového t-testu (místo párového t-testu) snižuje sílu.

Při chybném použití párového t-testu (místo dvouvýběrového t-testu) můžou vycházet naprosté nesmysly - záleží pak na uspořádání dat v obou výběrech, tj. na tom, jaké hodnoty se od sebe budou odečítat.

Příklad: Při zjišťování vlivu kouření na nervovou soustavu se u dvanácti osob měřil počet záchvěvů ruky před vykouřením a po vykouření cigarety. Výsledky měření jsou v následující tabulce:

před	44	54	37	62	40	44	49	53	23	69	51	28
po	50	63	52	83	48	43	55	47	25	71	58	37

Zvolte vhodný test a rozhodněte, jestli je mezi průměrným počtem záchvěvů před a po vykouření cigarety významný rozdíl na hladině významnosti $\alpha = 0.01$.

Příklad: Jak známo, nedoporučuje se rychle za sebou střídat požívání horkého jídla a studeného nápoje, protože jsou přitom zuby vystavovány teplotním šokům, které mohou snižovat odolnost zubní skloviny. Byl proveden experiment, ve kterém osm vytržených neplombovaných zubů bylo opakovaně vystavováno teplotním šokům tak, že byly střídavě ponořovány do vařící a ledové vody. Osm jiných zubů bylo naopak pomalu vařeno. Nakonec byly všechny zuby drceny v lisu a přitom byla změřena síla, při které každý zub prasknul:

pomalu uvařené	27.4	26.2	26.2	29.4	30.1	28.2	27.0	28.4
po teplotním šoku	25.9	26.4	27.0	27.8	26.3	27.6	27.0	25.6

Rozhodněte, jestli teplotní šoky opravdu snižují pevnost zubu a spočítejte 95% konfidenční interval pro rozdíl středních hodnot síly potřebné k rozdrcení zubu.

Předpoklady

Předpoklady pro dvouvýběrový t-test:

- shoda rozptylů (test shody rozptylů),
- normalita (test normality),
- nezávislost.

Předpoklad shody rozptylů

Welchův test (default v R):

$$T = \frac{\overline{X}_n - \overline{Y}_m}{S_{\text{m.diff}}},$$

kde

$$S^2_{\text{m.diff}} = \frac{S^2_X}{n} + \frac{S^2_Y}{m}.$$

Za platnosti H_0 (i bez předpokladu shody rozptylů) má testová statistika přibližně t-rozdělení s počtem stupňů volnosti:

$$\text{W} = \frac{(S^2_X/n + S^2_Y/m)^2}{(S^2_X/n)^2/(n-1) + (S^2_Y/m)^2/(m-1)},$$

tj. H_0 zamítáme, pokud $|T| > t_{W,1-\alpha/2}$.

Předpoklad normality

V R je implementováno mnoho různých testů normality. V praxi se nejčastěji doporučuje test Shapiro-Wilkův.

POZOR: V případě dvouvýběrového testu se test normality samozřejmě používá zvlášť na každý výběr (při platnosti alternativy není sloučený výběr normální, ani když oba výběry normální jsou).

Porušení normality lze často řešit vhodnou transformací (která „opraví“ sešikmení dat): Box-Coxy mocninné transformace, logaritmus. Obvykle příliš nezáleží na tom, jestli testujeme shodnost středních hodnot pro původní a transformovaná data.

V přítomnosti odlehlých pozorování můžeme použít dvouvýběrový pořadový (Wilcoxonův) test.

Předpoklad nezávislosti

Nezávislost po sobě jdoucích pozorování se testuje Durbin-Watsonovým testem, ale v praxi můžou být data závislá i „jinak“.

- časové řady,
- longitudinální data, tj. opakovaná měření na jednotlivých subjektech.

Další způsob porušení předpokladů může být například cenzorování nebo závislost na dalších veličinách nebo chybějící pozorování nebo spousta dalších problémů...

Dvouvýběrový Wilcoxonův test (rank sum test, Mann-Whitney)

$X_1, \ldots, X_n$ a $Y_1, \ldots, Y_m$ nezávislé náhodné výběry s posunutými distribučními funkcemi $F_X(x)$ a $G_Y(x) = F_X(x + \delta)$.

$H_0 : \delta = \delta_0$
$H_1 : \delta \neq \delta_0$

1. $Z_1, \ldots, Z_{n+m}$ je spojený výběr $X_i - \delta_0$ a Y_i,
2. seřadíme Z_i podle velikosti,
3. S_X je součet pořadí odpovídající výběru X a S_Y je součet pořadí odpovídající výběru Y.

Za platnosti H_0 lze pro dané n a m vypočítat rozdělení S_X a S_Y.

Testová statistika S_X (nebo $W_{n,m} = S_X - n(n+1)/2$) je založena pouze na pořadích a není tedy citlivá na odlehlá pozorování.

Poznámky:

T-testy lze zobecnit i pro vícerozměrná data (Hotellingovo T^2, F-test).

Pro jiné situace lze často jednoduše odvodit *test poměrem věrohodnosti (likelihood ratio test)*: za jistých předpokladů má za platnosti H_0 testová statistika $-2\log(L_0/L_1)$ rozdělení $\chi^2_{r_1-r_0}$...

V praxi se často používají testy nezávislosti v kontingenční tabulce (budeme mít na konci semestru).

Testování hypotéz se hodně používá i v lineární regresi (t-testy významnosti regresních koeficientů, testy podmodelů) - to budeme mít asi za měsíc.

Vzhledem k tomu, že pravděpodobnost chyby prvního druhu se většinou volí α =5%, tak při provedení většího množství testů nakonec vždy najdeme významnou závislost, která ve skutečnosti neexistuje (zde pak pomáhá např. Bonferroniho nebo Holmova korekce na mnohonásobné testování).

Mnohonásobné testování

Při testování většího počtu hypotéz dochází ke zvýšení celkové pravděpodobnosti chyby prvního druhu (FWER = family-wise error rate).

Proto se v praxi často musí získané p-hodnoty upravovat. Často se používá například Bonferroniho metoda (p-hodnoty se vynásobí počtem testů), která je vhodná hlavně pro menší počet porovnání.

Pro velké počty porovnání (např. v genetice) se místo FWER kontroluje FDR (false discovery rate).

Pro některé prakticky důležité situace (zejména k-výběrový problém) se používají speciální postupy: analýza rozptylu, Tukeyho a Scheffého metoda, atd.

Týden 7

Téma:

- náhodné vektory,
- pravidla pro počítání s vektory středních hodnot a s variančními maticemi,
- sdružené, marginální a podmíněné rozdělení,
- kovariance a korelace,
- grafické znázornění mnohorozměrných dat.

Mnohorozměrná data

$\mathcal{X}$ datová matice (n pozorování p-tice náhodných veličin, tzv. náhodného vektoru)

$$\mathcal{X} = \begin{pmatrix} x_{11} & \dots & x_{1p} \\ x_{21} & \dots & x_{2p} \\ \vdots & \dots & \vdots \\ x_{n1} & \dots & x_{np} \end{pmatrix}$$

Příklad: Bankovky, kosatce, ...

Grafické znázornění: grafy v R, ggobi.

Náhodný vektor $X \in \mathbb{R}^p$

(Mnohorozměrná) sdružená distribuční funkce:
$F(x) = P(X \le x) = P(X_1 \le x_1, X_2 \le x_2, \dots, X_p \le x_p)$

$f(x)$ je sdružená hustota X, t.j., $F(x) = \int_{-\infty}^{x} f(u)du$

$\int_{-\infty}^{\infty} f(u)\,du = 1$, $P\{X \in (a,b)\} = \int_a^b f(x)dx$

Ve vícerozměrném prostoru potřebujeme navíc další pojmy:

$X = (X_1, X_2)^\top, \quad X_1 \in \mathbb{R}^k \quad X_2 \in \mathbb{R}^{p-k}$
marginální hustota X_1 je $f_{X_1}(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2)dx_2$

podmíněná hustota X_2 (za podmínky $X_1 = x_1$) je
$f_{X_2|X_1=x_1}(x_2) = f(x_1, x_2)/f_{X_1}(x_1)$

Příklad:

$$f(x_1,x_2)=\begin{cases}\frac{1}{2}x_1+\frac{3}{2}x_2 & 0\le x_1,x_2\le 1,\\ 0 & \text{jinak.}\end{cases}$$

$f(x_1,x_2)$ je skutečně pravděpodobnostní hustota, protože $f(x_1,x_2)\ge 0$ a

$$\int f(x_1,x_2)dx_1x_2=\frac{1}{2}\left[\frac{x_1^2}{2}\right]_0^1+\frac{3}{2}\left[\frac{x_2^2}{2}\right]_0^1=\frac{1}{4}+\frac{3}{4}=1.$$

Marginální hustoty jsou:

$$f_{X_1}(x_1) = \int f(x_1,x_2)dx_2=\int_0^1\left(\frac{1}{2}x_1+\frac{3}{2}x_2\right)dx_2=\frac{1}{2}x_1+\frac{3}{4};$$
$$f_{X_2}(x_2) = \int f(x_1,x_2)dx_1=\int_0^1\left(\frac{1}{2}x_1+\frac{3}{2}x_2\right)dx_1=\frac{3}{2}x_2+\frac{1}{4}.$$

Podmíněné hustoty:

$$f(x_2\mid x_1)=\frac{\frac{1}{2}x_1+\frac{3}{2}x_2}{\frac{1}{2}x_1+\frac{3}{4}}\quad\text{and}\quad f(x_1\mid x_2)=\frac{\frac{1}{2}x_1+\frac{3}{2}x_2}{\frac{3}{2}x_2+\frac{1}{4}}.$$

Nezávislost

Náhodné veličiny X_1, X_2 jsou nezávislé tehdy a jen tehdy pokud $f(x)=f(x_1,x_2)=f_{X_1}(x_1)f_{X_2}(x_2)$.
Totéž jinými slovy: všechna podmíněná rozdělení jsou stejná jako rozdělení marginální ($f(x_2\mid x_1)=f_{X_2}(x_2)$).

POZOR: Dva náhodné vektory mohou mít stejná marginální rozdělení a přitom různá sdružená rozdělení.

Příklad:

$$f(x_1,x_2)=1,\quad 0<x_1,x_2<1,$$
$$f(x_1,x_2)=1+\alpha(2x_1-1)(2x_2-1),\quad 0<x_1,\ x_2<1,\quad -1\le\alpha\le 1.$$
$$f_{X_1}(x_1)=1,\quad f_{X_2}(x_2)=1.$$
$$\int_0^1 1+\alpha(2x_1-1)(2x_2-1)dx_2=1+\alpha(2x_1-1)[x_2^2-x_2]_0^1=1.$$

Vektor středních hodnot

$EX\in\mathbb{R}^p$ je p-rozměrný vektor středních hodnot náhodného vektoru X

$$EX=\begin{pmatrix}EX_1\\ \vdots\\ EX_p\end{pmatrix}=\int xf(x)dx=\begin{pmatrix}\int x_1f(x)dx\\ \vdots\\ \int x_pf(x)dx\end{pmatrix}=\mu.$$

Poznámka: zřejmě $\int x_1f(x)dx=\cdots=\int x_1f_{X_1}(x_1)dx_1$.

Vlastnosti vektoru středních hodnot plynou z vlastností integrálu (nebo z vlastností střední hodnoty náhodné veličiny):

$$E(\mathcal{A}X+b) = \mathcal{A}E(X)+b$$
$$E(X+Y) = E(X)+E(Y)$$

Jsou-li náhodné vektory X a Y nezávislé, pak

$$E(XY^\top)=\int xy^\top f(x,y)dxdy$$
$$=\int xf(x)dx\int y^\top f(y)dy=EXEY^\top$$

Varianční matice (Σ)

$$\Sigma=Var(X)=E(X-\mu)(X-\mu)^\top$$

Budeme říkat, že náhodný vektor X má rozdělení s vektorem středních hodnot $EX=\mu$ a s varianční maticí $Var(X)=\Sigma$, t.j.,

$$X\sim(\mu,\Sigma)$$

(Ko)varianční matice lineární transformace

Kovarianční matice: $Cov(X, Y) = E(X - EX)(Y - EY)^\top$

Varianční (rozptylová) matice: $Cov(X, X) = Var(X)$

Vlastnosti:

$$Var(a^\top X) = a^\top \; Var(X) \; a = \sum_{i,j} a_i a_j \sigma_{X_i X_j}$$
$$Var(\mathcal{A}X + b) = \mathcal{A} \; Var(X) \; \mathcal{A}^\top$$
$$Cov(X + Y, Z) = Cov(X, Z) + Cov(Y, Z)$$
$$Var(X + Y) = Var(X) + Cov(X, Y) + Cov(Y, X) + Var(Y)$$
$$Cov(\mathcal{A}X, \mathcal{B}Y) = \mathcal{A} \; Cov(X, Y) \; \mathcal{B}^\top.$$

Prvky matice Σ jsou rozptyly a kovariance složek náhodného vektoru X: $\Sigma = (\sigma_{X_i X_j})$

(rozptyl $\sigma_{X_i X_j} = Cov(X_i, X_j)$, kovariance $\sigma_{X_i X_i} = Var(X_i)$)

Výpočetní vzorec:

$$\Sigma = E(XX^\top) - \mu\mu^\top$$

Varianční matice je pozitivně semidefinitní: $\Sigma \geq 0$
(rozptyl $a^\top \Sigma a$ libovolné lineární kombinace $a^\top X$ nemůže být záporný).

Transformace

Hustota transformovaného vektoru se (v případě potřeby) spočítá podobně jako hustota transformované náhodné veličiny.

$X \sim f_X$, zajímá nás hustota (prosté) trasformace $Y = t(X)$?

Pokud $\mathcal{J}$ je jakobián zpětné transformace $X = u(Y)$, t.j.,

$$\mathcal{J} = \left(\frac{\partial x_i}{\partial y_j} \right) = \left(\frac{\partial u_i(y)}{\partial y_j} \right),$$

pak hustota $Y = t(X)$ je:

$$f_Y(y) = \text{abs}(|\mathcal{J}|) f_X\{u(y)\}$$

Mnohorozměrné normální rozdělení

Hustota mnohorozměrného normálního rozdělení (za předpokladu plné hodnosti Σ) je:

$$f(x) = |2\pi\Sigma|^{-1/2} \exp\left\{ -\frac{1}{2}(x - \mu)^\top \Sigma^{-1} (x - \mu) \right\}.$$

$X \sim N_p(\mu, \Sigma)$

Vektor středních hodnot $EX = \mu$

Varianční matice X je $Var\{X\} = \Sigma > 0$

Příklad: Jaké je marginální rozdělení každé složky náhodného vektoru?

Příklad: Čemu odpovídá kvadratická forma $(x - \mu)^T \Sigma^{-1} (x - \mu)$ ve vzorci pro hustotu?

Hustota $N_p(\mu, \Sigma)$ je konstantní na elipsoidech

$$(x-\mu)^\top \Sigma^{-1}(x-\mu) = d^2$$

Pokud $X \sim N_p(\mu, \Sigma)$, pak náhodný vektor

$$Y = (X-\mu)^\top \Sigma^{-1}(X-\mu)$$

má rozdělení χ^2_p (protože tzv. Mahalanobisova transformace $Z = \Sigma^{-1/2}(X-\mu) \sim N_p(0, \mathcal{I}_p)$ a $Y = Z^T Z = \sum_{j=1}^p Z_j^2$).

(Pozn.: pokud varianční matici nahradíme odhadem, tak získáme tzv. Hotellingovo rozdělení a mnohorozměrnou verzi t-testu.)

Centrální limitní věta

Centrální limitní věta popisuje asymptotické rozdělení výběrového průměru.

$X_1, X_2, \ldots, X_n$, i.i.d. z rozdělení $X_i \sim (\mu, \Sigma)$

$$\sqrt{n}(\overline{x} - \mu) \xrightarrow{\mathcal{L}} N_p(0, \Sigma) \qquad \text{for} \quad n \longrightarrow \infty.$$

CLV lze použít ke konstrukci konfidenčních elipsoidů (nepraktické) nebo k testování.

Normální rozdělení hraje ve statistice centrální úlohu.

Mnohorozměrná delta metoda

Pokud $\sqrt{n}(t-\mu) \xrightarrow{\mathcal{L}} N_p(0, \Sigma)$ a $f = (f_1, \ldots, f_q)^\top : \mathbb{R}^p \to \mathbb{R}^q$ jsou reálné funkce diferencovatelné v $\mu \in \mathbb{R}^p$, pak $f(t)$ je asymptoticky normální se stř. hodnotou $f(\mu)$ a varianční maticí $\mathcal{D}^\top \Sigma \mathcal{D}$, t.j.,

$$\sqrt{n}\{f(t) - f(\mu)\} \xrightarrow{\mathcal{L}} N_q(0, \mathcal{D}^\top \Sigma \mathcal{D}) \qquad \text{for} \quad n \longrightarrow \infty,$$

kde

$$\mathcal{D} = \left(\frac{\partial f_j}{\partial t_i} \right)(t) \Big|_{t=\mu}$$

$(p \times q)$ je matice parciálních derivací.

Pomocí této věty můžeme také nalézt transformace "stabilizující rozptyl".

Příklady

- Mnohorozměrné normální rozdělení:
  - marginální a podmíněná rozdělení,
  - nezávislost,
  - lineární transformace.
- Standardizace.
- Mahalanobisova transformace.

Příklad: T-test zapsaný pomocí „náhodných vektorů“ ($X \sim N_n(\mu, \mathrm{diag}(\sigma^2))$, $\overline{X}_n = 1_n^\top X/n$, $S^2 = \ldots$).

Týden 8

Téma:

- mnohorozměrná data,
- standardizace a Mahalanobisova transformace,
- projekce a lineární kombinace,
- hlavní komponenty.

Mnohorozměrná data

Opakování z minulého týdne:

$\mathcal{X}$ datová matice (n pozorování p-tice náhodných veličin, tzv. náhodného vektoru)

$$\mathcal{X} = \begin{pmatrix} x_{11} & \dots & x_{1p} \\ x_{21} & \dots & x_{2p} \\ \vdots & \dots & \vdots \\ x_{n1} & \dots & x_{np} \end{pmatrix}$$

Příklad: Grafické znázornění švýcarských bankovek (6D) na obrazovce počítače.

Průměr a výběrová varianční matice:

$$\overline{x} = \begin{pmatrix} \overline{x}_1 \\ \vdots \\ \overline{x}_p \end{pmatrix} = n^{-1}\mathcal{X}^\top \mathbf{1}_n$$

$$\begin{aligned} \mathcal{S} &= n^{-1}\mathcal{X}^\top\mathcal{X} - \overline{x}\,\overline{x}^\top \\ &= n^{-1}(\mathcal{X}^\top\mathcal{X} - n^{-1}\mathcal{X}^\top\mathbf{1}_n\mathbf{1}_n^\top\mathcal{X}) = n^{-1}\mathcal{X}^\top\mathcal{H}\mathcal{X} \end{aligned}$$

Centrovací matice $\mathcal{H} = \mathcal{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n^\top$.

$\mathcal{D} = \mathrm{diag}(s_{X_jX_j})$, kde X_j, $j = 1, \dots, p$ jsou sloupce matice $\mathcal{X}$

Centrovaná data: $\mathcal{X}_C = \mathcal{X} - n^{-1}\mathbf{1}_n\mathbf{1}_n^\top\mathcal{X} = \mathcal{H}\mathcal{X}$

Standardizovaná data: $\mathcal{X}_S = \mathcal{X}_C\mathcal{D}^{-1/2} = \mathcal{H}\mathcal{X}\mathcal{D}^{-1/2}$

Korelační matice $\mathcal{R} = \mathcal{D}^{-1/2}\mathcal{S}\mathcal{D}^{-1/2}$.

Lineární transformace:

$\mathcal{A}$ $(q \times p)$ matice konstant:

$$\mathcal{Y} = \mathcal{X}\mathcal{A}^\top = (y_1, \dots, y_n)^\top$$

$$\overline{y} = \mathcal{A}\overline{x}$$

$$\mathcal{S}_\mathcal{Y} = \mathcal{A}\mathcal{S}_\mathcal{X}\mathcal{A}^\top$$

Standardizací získáme centrovaná data s jednotkovými rozptyly.

Mahalanobisova transformace:

$$z_i = \mathcal{S}^{-1/2}(x_i - \overline{x}), \quad i = 1, \dots, n,$$

$$\mathcal{S}_\mathcal{Z} = n^{-1}\mathcal{Z}^\top\mathcal{H}\mathcal{Z} = \mathcal{I}_p, \ \overline{\mathcal{Z}} = 0.$$

Mahalanobisova transformace (sphering) vede na centrovaná data s jednotkovou varianční maticí (tj. nekorelované sloupce).

Hlavní komponenty (principal components)

Při grafickém znázornění mnohorozměrných dat se chceme soustředit na ty nejdůležitější projekce. Nejjednodušší metoda „hledání zajímavých projekcí“ je metoda hlavních komponent.

Cíl: nalézt standardizovanou lineární kombinaci s maximálním rozptylem.

$$\delta^\top X = \sum_{j=1}^p \delta_j X_j \quad \text{a přitom} \quad \underset{\text{standardizovaná}}{||\delta|| = 1}$$

$$\max_{\{\delta:\|\delta\|=1\}} Var(\delta^\top X) = \max_{\{\delta:\|\delta\|=1\}} \delta^\top Var(X)\delta.$$

Řešení pomocí lineární algebry (spektrální rozklad matice):

$$\delta = \gamma_1 = \text{ první vlastní vektor } Var(X)$$

Příklad:

Dvourozměrné normální rozdělení $N(0, \Sigma)$, $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \rho > 0.$

Vlastní čísla varianční matice jsou $\lambda_1 = 1 + \rho$ a $\lambda_2 = 1 - \rho$ s vlastními vektory

$$\gamma_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \gamma_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

Hlavní komponenty tedy jsou

$$Y = \Gamma^\top (X - \mu) = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} X$$

or

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} X_1 + X_2 \\ X_1 - X_2 \end{pmatrix}.$$

Příklad: (pokračování)

První hlavní komponenta:

$$Y_1 = \frac{1}{\sqrt{2}}(X_1 + X_2)$$

a druhá hlavní komponenta:

$$Y_2 = \frac{1}{\sqrt{2}}(X_1 - X_2).$$

Rozptyl první hlavní komponenty je:

$$\begin{aligned} Var(Y_1) &= Var\left\{\frac{1}{\sqrt{2}}(X_1 + X_2)\right\} = \frac{1}{2} Var(X_1 + X_2) \\ &= \frac{1}{2}\{Var(X_1) + Var(X_2) + 2\, Cov(X_1, X_2)\} \\ &= \frac{1}{2}(1 + 1 + 2\rho) = 1 + \rho = \lambda_1. \end{aligned}$$

Obdobně: $Var(Y_2) = \lambda_2$.

Vlastnosti hlavních komponent

Nechť $X \sim (\mu, \Sigma)$ a Y je transformace metodou hlavních komponent, tj. $Y = \Gamma^\top (X - \mu)$.

(Spektrální rozklad: $\Sigma = \Gamma \Lambda \Gamma^\top$, kde Γ je ortogonální a Λ diagonální.)

Pak platí:

$$\begin{aligned} &EY_j = 0 \\ &Var(Y_j) = \lambda_j \\ &Cov(Y_i, Y_j) = 0, \text{ for } i \neq j \\ &Var(Y_1) \geq \cdots \geq Var(Y_p) \geq 0 \\ &\textstyle\sum_j Var(Y_j) = \text{tr}(\Sigma) \\ &\textstyle\prod Var(Y_j) = |\Sigma|. \end{aligned}$$

Interpretace: $\lambda_j / \text{tr}(\Sigma)$ se považuje za podíl celkového rozptylu X vysvětlený j-tou hlavní komponentou.

Nechť $Y = a^\top X$ je standardizovaná lineární kombinace nekorelovaná s prvními k hlavními komponentami X. Pak $Var(Y)$ je největší pro $a = \gamma_{k+1}$

Kovariance a korelace mezi PC a X

$$\begin{aligned} Cov(X, Y) &= E(XY^\top) - EXEY^\top = E(XY^\top) \\ &= E(XX^\top\Gamma) - \mu\mu^\top\Gamma = Var(X)\Gamma \\ &= \Sigma\Gamma = \Gamma\Lambda\Gamma^\top\Gamma = \Gamma\Lambda \end{aligned}$$

$$\rho_{X_iY_j} = \gamma_{ij}\left(\frac{\lambda_j}{\sigma_{X_iX_i}}\right)^{1/2}$$

Lze jednoduše spočítat, že $\sum_i \rho^2_{X_iY_j} = 1$ (v grafu bude bod „Y_j“ se souřadnicemi $\rho_{X_iY_j}$ ležet na povrchu koule), $\sum_{i=1}^r \rho^2_{X_iY_j}$ můžeme interpretovat jako část variability X_i vysvětlenou prvními r HK.

Odhady, interpretace, příklady

V praxi se hlavní komponenty počítají pomocí spektrálního rozkladu výběrové varianční matice ($\mathcal{S} = \mathcal{G}\mathcal{L}\mathcal{G}^\top$), problémem může být závislost na měřítku. V R: `prcomp()`, `princomp()`.

Pokud jsou jednotlivé proměnné měřené v různých jednotkách, tak se doporučuje analyzovat standardizovaná data (to je prakticky totéž jako spektrální rozklad korelační matice). V R: `prcomp( , scale.=TRUE)`.

Volba počtu hlavních komponent (dimension reduction) je založena na velikosti vlastních čísel (grafické znázornění: screeplot).

Interpretace hlavních komponent je obvykle založena na korelacích hlavních komponent s původními proměnnými a na hodnotách hlavních komponent pro jednotlivá pozorování (grafické znázornění: biplot).

Hlavní komponenty v praxi

Algoritmus:

1. Rozhodnutí, jestli budeme analyzovat původní nebo standardizované proměnné.
2. Volba počtu hlavních komponent, kterými můžeme nahradit původní proměnné:
   1. rozumně malý počet HK, které přitom vysvětlují podstatnou část celkové variability,
   2. HK, které vysvětlují vyšší, než průměrné množství variability (= 1 při standardizované analýze).
3. Interpretace hlavních komponent (korelace s původními proměnnými, hodnoty hlavních komponent pro jednotlivá pozorování).

Příklad: Car Marks (carmean2), Timebudget (timebudget).

Týden 9

Téma:

- lineární model,
- matice modelu,
- kvadratická regrese.
- residua,
- residuální součet čtverců (RSS),
- koeficient determinace,
- odhady parametrů v lineárním modelu,

Bylo:

Měli jsme náhodné vektory (X), EX, $Var(X)$, ...

V praxi máme náhodný výběr a datovou matici $\mathcal{X}$, $\bar{x} = \mathcal{X}^\top 1_n/n$, $\mathcal{S} = \mathcal{X}^\top \mathcal{H}\mathcal{X}/n$

Užitečné nástroje:

derivace $f(x) = f(x_1, \ldots, x_p)$,

$$\frac{\partial f(x)}{\partial x} = \begin{pmatrix} \frac{\partial f(x_1,\ldots,x_p)}{\partial x_1} \\ \frac{\partial f(x_1,\ldots,x_p)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x_1,\ldots,x_p)}{\partial x_p} \end{pmatrix}$$

Příklad:

Lineární kombinace $f(x) = a^\top x = a_1 x_1 + \cdots + a_p x_p$.

Zřejmě platí:

$$\frac{\partial f(x)}{\partial x} = \frac{a^\top x}{\partial x} = \frac{x^\top a}{\partial x} = a$$

Příklad:

Kvadratická forma $f(x) = x^\top A x = \sum_{i,j} a_{ij} x_i x_j$, kde A je symetrická matice.

Jednoduše lze ověřit, že:

$$\frac{\partial x^\top A x}{\partial x} = 2Ax$$

Lineární model

$Y \sim (\mathcal{X}\beta, \sigma^2 \mathcal{I}_n)$

Pro jednoduchost předpokládáme, že $\mathcal{X}$ má plnou hodnost a $p < n$.

- Y závisle proměnná (náhodný vektor),
- β neznámé regresní koeficienty (neznámé parametry),
- $\mathcal{X}$ matice modelu = matice vysvětlujících proměnných (matice konstant),
- σ^2 rozptyl (nekorelovaných) náhodných chyb kolem středních hodnot $\mathcal{X}\beta$.

Lineární model se často zapisuje také ve tvaru:

$$Y_i = \beta_1 X_{1,i} + \cdots + \beta_p X_{p,i} + \varepsilon_i,$$

kde ε_i jsou nekorelované náhodné chyby (měření).

Tvar závislosti

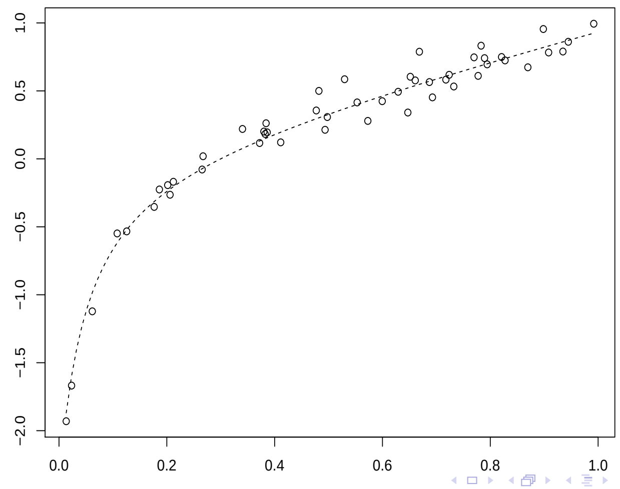
Volba matice modelu $\mathcal{X}$ určuje tvar závislosti (pomocí lineárního modelu lze jednoduše odhadnout i některé nelineární závislosti).

Regresní koeficienty β se obvykle odhadují metodou nejmenších čtverců a odhad se obvykle značí $\widehat{\beta}$.

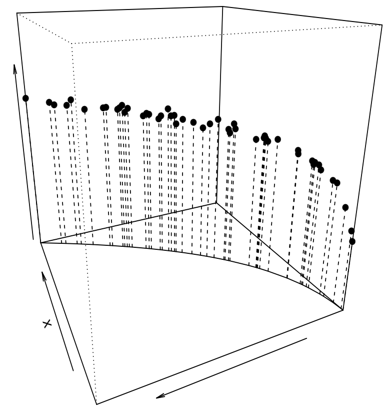
Příklad: Obecná přímka, přímka procházející počátkem.

Příklad: Kvadratická regrese: $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i$ (tj. lineární závislost Y_1 na X_i a X_i^2).

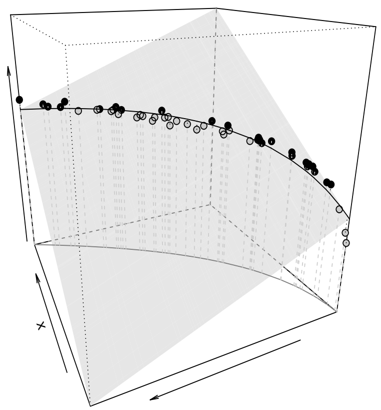
Příklad: simulovaná data
Skutečná závislost je zde výrazně nelineární.



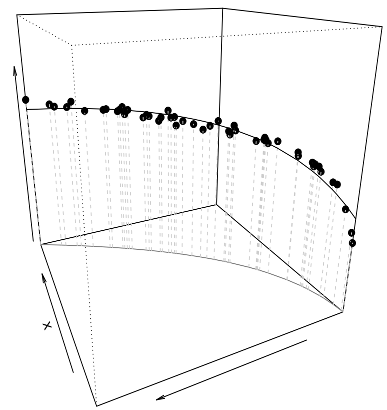
Příklad: simulovaná data
Kvadratická regrese: $EY_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2$, $i = 1, \ldots, n$.



Příklad: simulovaná data
Kvadratická regrese: $EY_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2$, $i = 1, \ldots, n$.

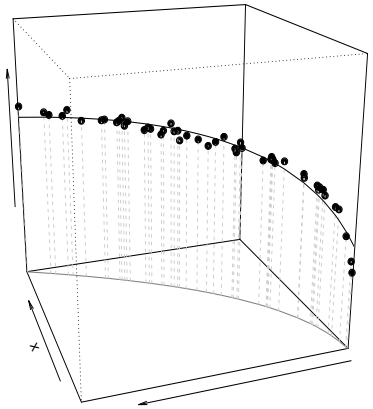


Příklad: simulovaná data
Kvadratická regrese: $EY_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2$, $i = 1, \ldots, n$.



Příklad: simulovaná data

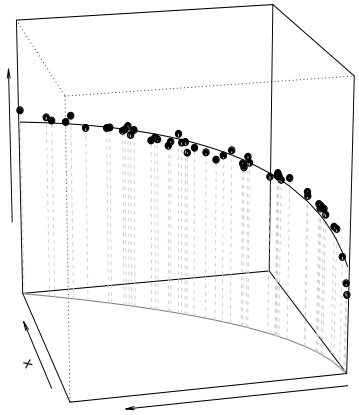
Kvadratická regrese: $EY_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2$, $i = 1, \ldots, n$.



Týden 9 Lineární model

Příklad: simulovaná data

Kvadratická regrese: $EY_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2$, $i = 1, \ldots, n$.

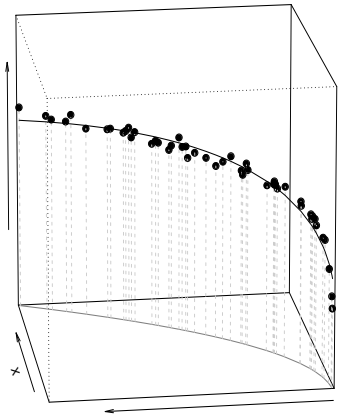


Zdeněk Hlávka (KPMS) NMAI 061 130 / 232

Týden 9 Lineární model

Příklad: simulovaná data

Kvadratická regrese: $EY_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2$, $i = 1, \ldots, n$.

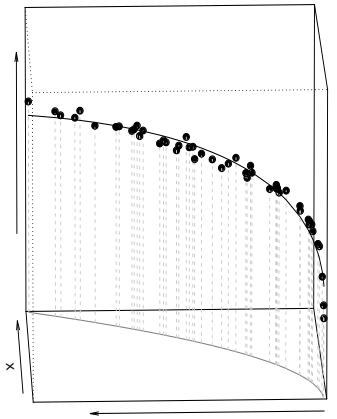


Zdeněk Hlávka (KPMS) NMAI 061 130 / 232

Týden 9 Lineární model

Příklad: simulovaná data

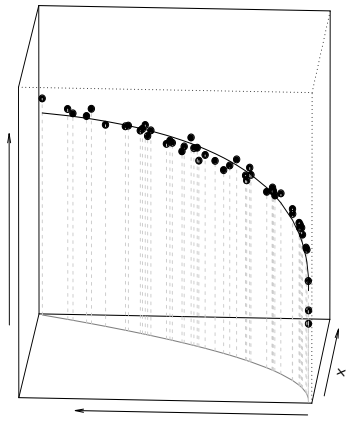
Kvadratická regrese: $EY_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2$, $i = 1, \ldots, n$.



Zdeněk Hlávka (KPMS) NMAI 061 130 / 232

Příklad: simulovaná data

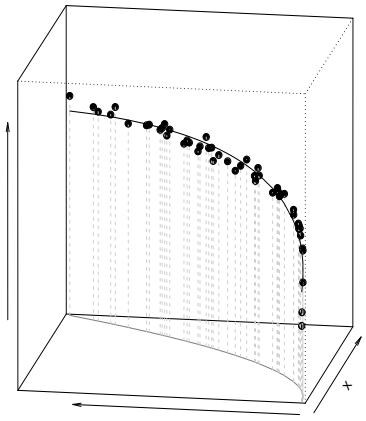
Kvadratická regrese: $EY_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2$, $i = 1, \ldots, n$.



Týden 9 Lineární model

Příklad: simulovaná data

Kvadratická regrese: $EY_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2$, $i = 1, \ldots, n$.

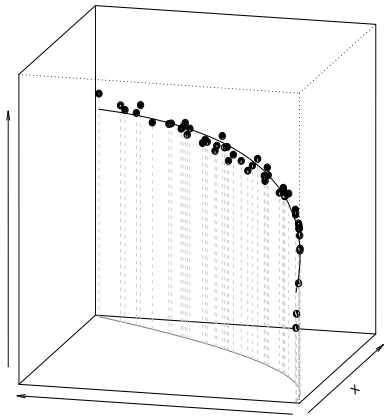


Zdeněk Hlávka (KPMS) NMAI 061 130 / 232

Týden 9 Lineární model

Příklad: simulovaná data

Kvadratická regrese: $EY_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2$, $i = 1, \ldots, n$.

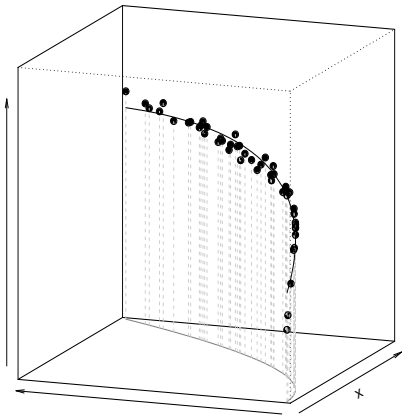


Zdeněk Hlávka (KPMS) NMAI 061 130 / 232

Týden 9 Lineární model

Příklad: simulovaná data

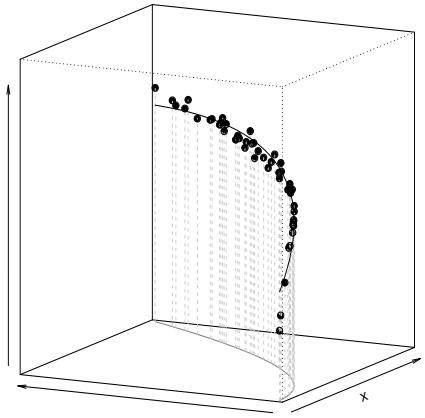
Kvadratická regrese: $EY_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2$, $i = 1, \ldots, n$.



Zdeněk Hlávka (KPMS) NMAI 061 130 / 232

Příklad: simulovaná data

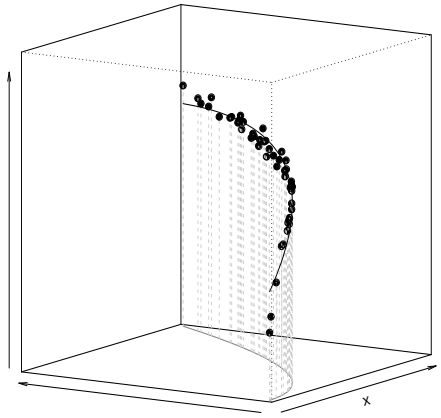
Kvadratická regrese: $EY_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2$, $i = 1, \ldots, n$.



Týden 9 Lineární model

Příklad: simulovaná data

Kvadratická regrese: $EY_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2$, $i = 1, \ldots, n$.

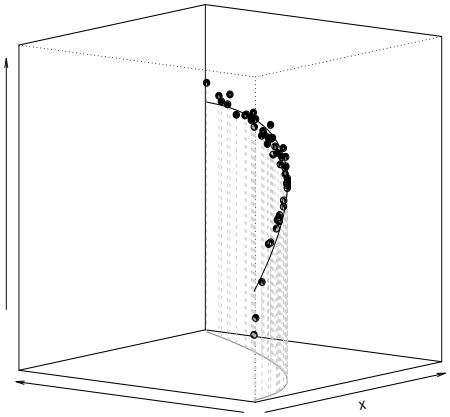


Zdeněk Hlávka (KPMS) NMAI 061 130 / 232

Týden 9 Lineární model

Příklad: simulovaná data

Kvadratická regrese: $EY_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2$, $i = 1, \ldots, n$.

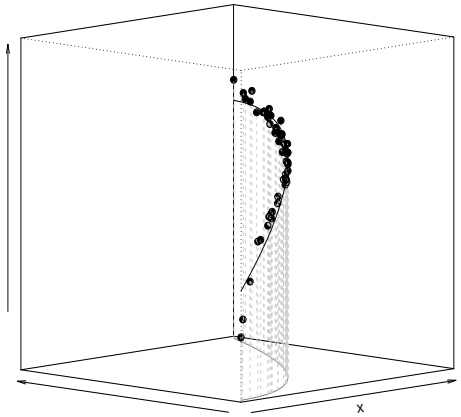


Zdeněk Hlávka (KPMS) NMAI 061 130 / 232

Týden 9 Lineární model

Příklad: simulovaná data

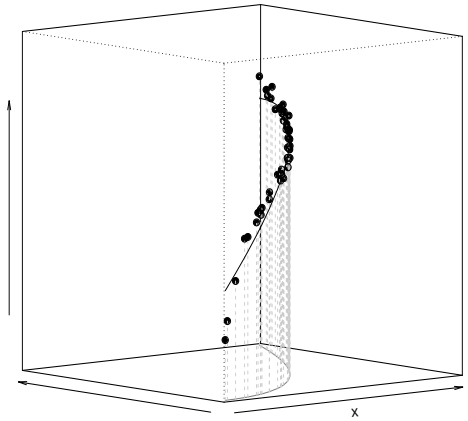
Kvadratická regrese: $EY_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2$, $i = 1, \ldots, n$.



Zdeněk Hlávka (KPMS) NMAI 061 130 / 232

Příklad: simulovaná data

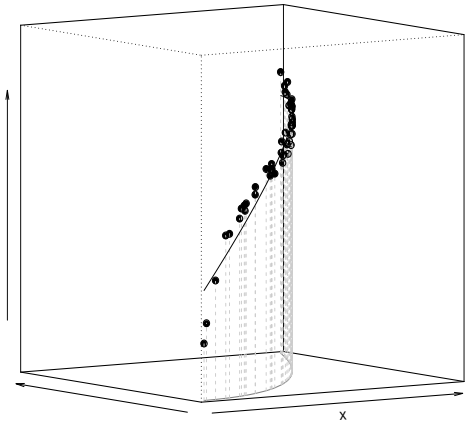
Kvadratická regrese: $EY_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2$, $i = 1, \ldots, n$.



Týden 9 Lineární model

Příklad: simulovaná data

Kvadratická regrese: $EY_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2$, $i = 1, \ldots, n$.

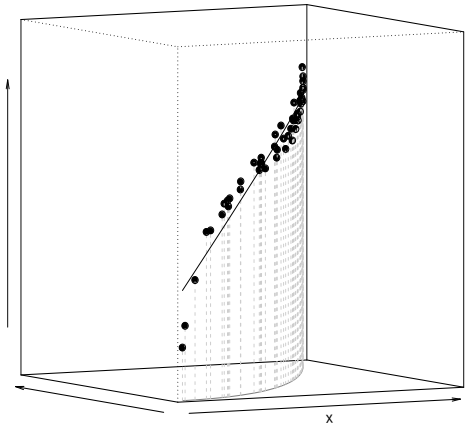


Zdeněk Hlávka (KPMS) NMAI 061 130 / 232

Týden 9 Lineární model

Příklad: simulovaná data

Kvadratická regrese: $EY_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2$, $i = 1, \ldots, n$.

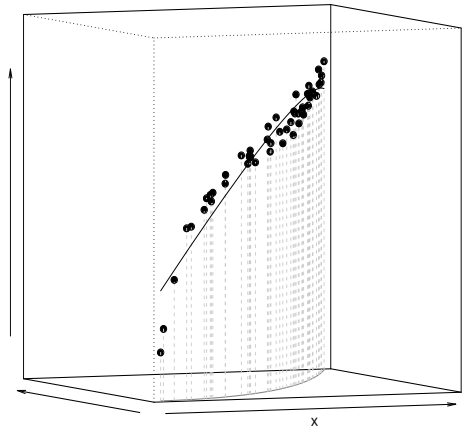


Zdeněk Hlávka (KPMS) NMAI 061 130 / 232

Týden 9 Lineární model

Příklad: simulovaná data

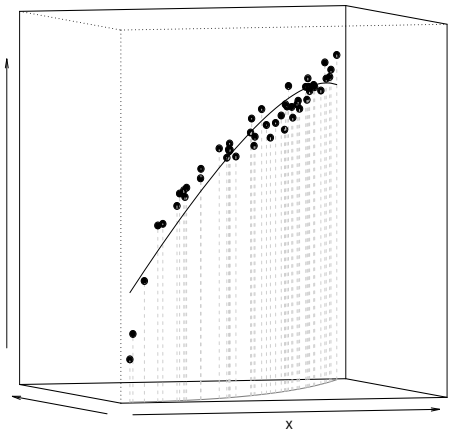
Kvadratická regrese: $EY_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2$, $i = 1, \ldots, n$.



Zdeněk Hlávka (KPMS) NMAI 061 130 / 232

Příklad: simulovaná data

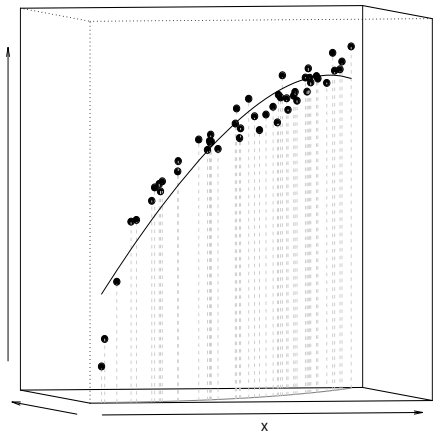
Kvadratická regrese: $EY_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2$, $i = 1, \ldots, n$.



Týden 9 Lineární model

Příklad: simulovaná data

Kvadratická regrese: $EY_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2$, $i = 1, \ldots, n$.

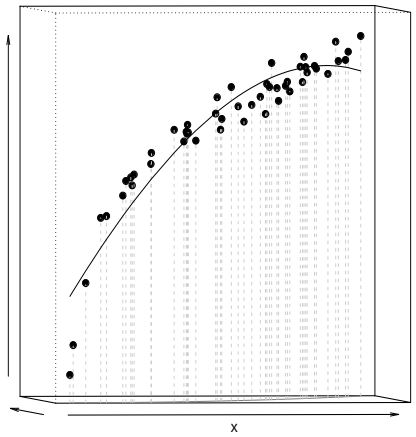


Zdeněk Hlávka (KPMS) NMAI 061 130 / 232

Týden 9 Lineární model

Příklad: simulovaná data

Kvadratická regrese: $EY_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2$, $i = 1, \ldots, n$.

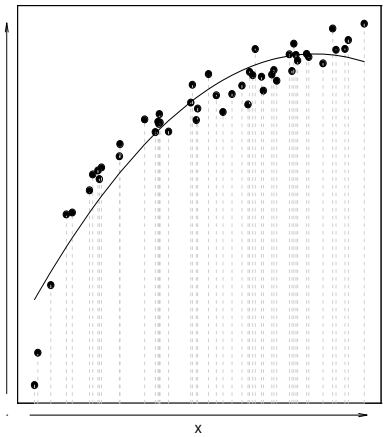


Zdeněk Hlávka (KPMS) NMAI 061 130 / 232

Týden 9 Lineární model

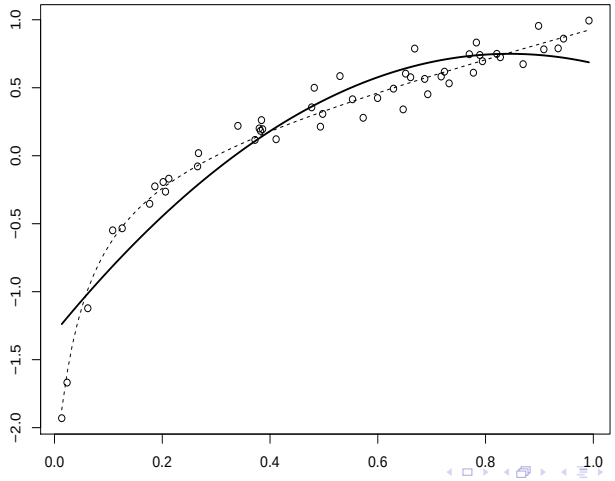
Příklad: simulovaná data

Kvadratická regrese: $EY_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2$, $i = 1, \ldots, n$.

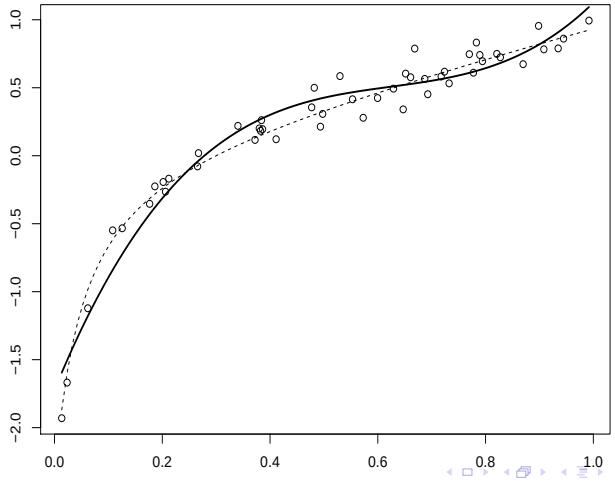


Zdeněk Hlávka (KPMS) NMAI 061 130 / 232

Příklad: simulovaná data
Kvadratická regrese: $EY_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2$, $i = 1, \ldots, n$.



Příklad: simulovaná data
Kubická regrese: $EY_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3$, $i = 1, \ldots, n$.



Příklady

Příklad: Polynomická regrese.

Příklad: Modelování periodicity (např. sezónní závislost):
$Y_i = \alpha \cos X_i + \beta \sin X_i + \varepsilon_i$ (pomocí pravidel pro počítání s goniometrickými funkcemi získáme $Y_i = \gamma \sin(X_i - \xi) + \varepsilon_i$).

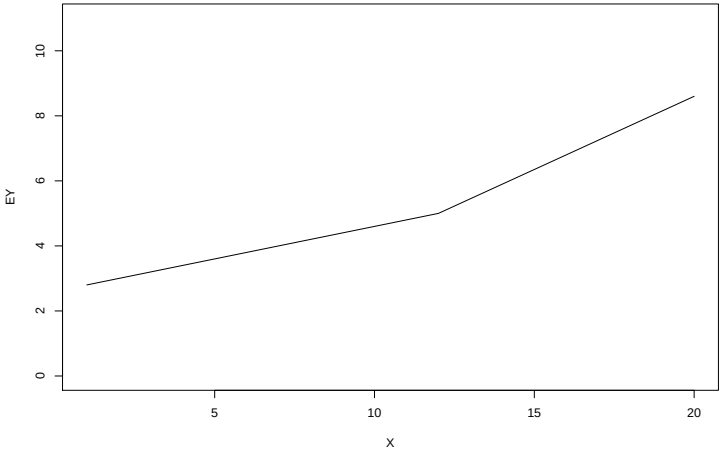
Příklad: Nezávislý náhodný výběr: $Y_i = \mu + \varepsilon_i$ (tj. $\beta = EY = \mu$).

Příklad: Dva nezávislé náhodné výběry (různé možnosti volby parametrů): data "swiss" - porovnání porodnosti v katolických a protestantských kantonech (1888).

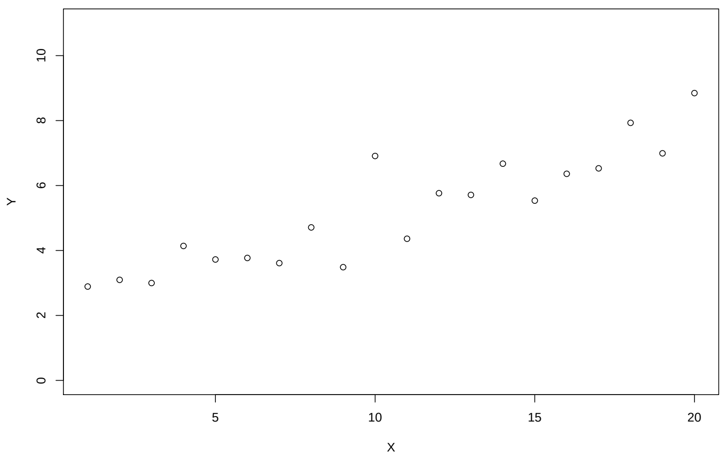
Příklad: Více vysvětlujících proměnných a interakce.

Příklad: Interakce spojité a faktorové vysvětlující proměnné (např. model dvou regresních přímek).

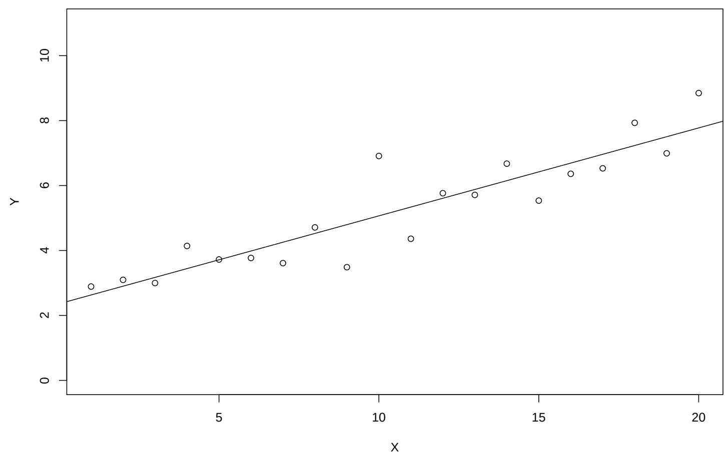
Příklad: Lineární model s „bodem změny sklonu regresní přímky".



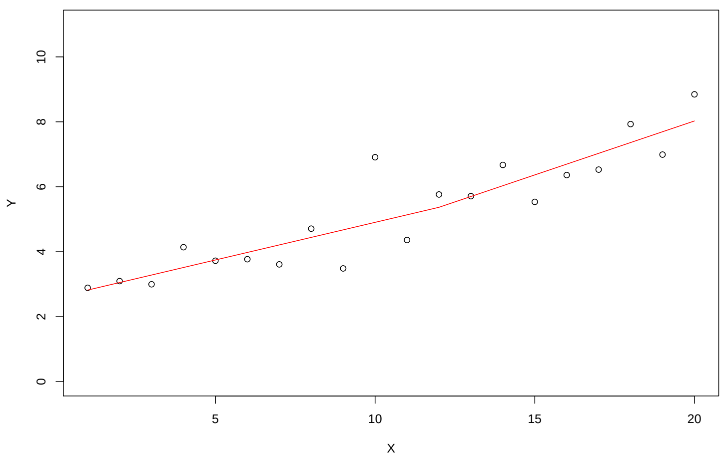
Příklad: Lineární model s „bodem změny sklonu regresní přímky".



Příklad: Lineární model s „bodem změny sklonu regresní přímky".



Příklad: Lineární model s „bodem změny sklonu regresní přímky".



Metoda nejmenších čtverců

Odhad regresních parametrů:

$$\begin{aligned}\widehat{\beta} &= \arg\min_{\beta\in R^p} \sum_{i=1}^{n} \{Y_i - (\beta_0 + \beta_1 X_{1,i} + \dots + \beta_p X_{p,i})\}^2 \\ &= \arg\min_{\beta\in R^p} \sum_{i=1}^{n} \{Y_i - \mathcal{X}_{[i]}\beta\}^2 \\ &= \arg\min_{\beta\in R^p} (Y - \mathcal{X}\beta)^\top (Y - \mathcal{X}\beta) = \arg\min_{\beta\in R^p} f(\beta)\end{aligned}$$

Minimum funkce $f(\beta)$ nalezneme vyřešením rovnice

$$\frac{\partial f(\beta)}{\partial \beta} = -2\mathcal{X}^\top Y + 2\mathcal{X}^\top \mathcal{X}\beta = 0,$$

tj. $\widehat{\beta} = (\mathcal{X}^\top \mathcal{X})^{-1}(\mathcal{X}^\top Y)$, pokud má matice $\mathcal{X}$ plnou hodnost.

Vlastnosti $\widehat{\beta}$

Velice jednoduše lze spočítat:

$E\widehat{\beta} = \ldots$

$Var\,\widehat{\beta} = \ldots$

Příklad: Kdy jsou odhady β_i nekorelované?

Pro odvození rozdělení odhadu $\widehat{\beta}$ potřebujeme navíc předpokládat znalost rozdělení náhodných chyb. To vede na *normální lineární model* a uvidíme, že náhodné veličiny $\widehat{\beta}_i$ mají za předpokladu normality t rozdělení, které lze použít pro testování hypotéz o parametrech β_i a ke konstrukci konfidenčních intervalů.

Proložené (vyrovnané) hodnoty $\widehat{Y} = \mathcal{X}\widehat{\beta} = \mathcal{X}(\mathcal{X}^\top\mathcal{X})^{-1}\mathcal{X}^\top Y = HY$, kde H je projekční matice do lineárního podprostoru generovaného sloupci matice $\mathcal{X}$.

$E\widehat{Y} = \ldots$

$Var\,\widehat{Y} = \ldots$

Residua $u = \widehat{\varepsilon} = Y - \mathcal{X}\widehat{\beta} = (\mathcal{I}_n - H)Y$ můžeme interpretovat jako odhad nepozorovaných náhodných chyb ε_i.

$Eu = \ldots$

$Var\,u = \ldots$

$Cov(\widehat{Y}, u) = Cov\{HY, (\mathcal{I}_n - H)Y\} = \ldots$

Residuální součet čtverců: $RSS = u^\top u$.

Nevychýlený **odhad parametru** σ^2:

$$S^2 = \widehat{\sigma}^2 = \frac{RSS}{n-p} = \frac{u^\top u}{n-p} = \frac{\sum u_i^2}{n-p}.$$

Koeficient determinace R^2: jednoduché srovnání modelu s modelem bez vysvětlujících proměnných (odhad je pak průměr $\overline{Y}$):

$$R^2 = 1 - \frac{RSS}{\sum(Y_i - \overline{Y})^2} = \frac{\sum(Y_i - \overline{Y})^2 - RSS}{\sum(Y_i - \overline{Y})^2}.$$

Koeficient determinace říká, jak velká část variability kolem průměru $\sum(Y_i - \overline{Y})^2$ je vysvětlená závislostí na vysvětlujících proměnných.

Poznámky

- Prakticky zajímavé hypotézy lze často formulovat pomocí hypotéz o lineárních kombinacích vektoru parametrů β.
- Předpoklady (např. nezávislost náhodných chyb, konstantní rozptyl) se většinou ověřují pomocí různých typů residuí.
- Za dodatečného předpokladu normality získáme *normální lineární model*, ve kterém odhad MNČ odpovídá odhadu metodou maximální věrohodnosti a ve kterém umíme vypočítat přesné rozdělení odhadu β (konfidenční intervaly, hypotézy o *beta*).

Příští týden: testování jednotlivých parametrů, testování podmodelu, normální lineární model, residua.

Týden 10

Téma:

- normální lineární model,
- testování jednotlivých parametrů (za předpokladu normality),
- podmodel,
- testování podmodelu (za předpokladu normality),
- interakce,
- ověřování předpokladů lineárního modelu (rezidua).

Normální lineární model

$$Y \sim N_n(\mathcal{X}\beta, \sigma^2\mathcal{I}_n)$$

Význam všech symbolů je stejný jako u (obyčejného) lineárního modelu a navíc je jenom předpoklad normality.

Normální lineární model se často zapisuje ve tvaru:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \cdots + \beta_k X_{k,i} + \varepsilon_i,$$

kde $\varepsilon_i \sim N(0, \sigma^2)$ jsou nezávislé náhodné chyby (tj. budeme předpokládat, že matice modelu $\mathcal{X}$ má $k+1$ sloupců a v jejím prvním sloupci jsou samé jedničky).

Střední hodnotu i rozptyl odhadu $\widehat{\beta}$ jsme spočítali před týdnem (oba vzorečky samozřejmě platí i za dodatečného předpokladu normality).

Vlastnosti

Věta: Má-li matice $\mathcal{X}$ v normálním modelu $Y \sim N_n(\mathcal{X}\beta, \sigma^2\mathcal{I}_n)$ hodnost rovnou počtu jejích sloupců, potom:

- $\widehat{\beta} \sim N(\beta, \sigma^2 V)$, kde $V = (\mathcal{X}^\top \mathcal{X})^{-1}$,
- náhodné vektory $\widehat{\beta}$ a u (residua) jsou nezávislé,
- $$T_j = \frac{\widehat{\beta}_j - \beta_j}{S\sqrt{v_{jj}}} \sim t_{n-k-1}, \quad j = 0, \ldots, k,$$
- interval $(\widehat{\beta}_j - S\sqrt{v_{jj}}t_{n-k-1;1-\alpha/2}, \widehat{\beta}_j + S\sqrt{v_{jj}}t_{n-k-1;1-\alpha/2})$ je interval spolehlivosti pro β_j se spolehlivostí $1-\alpha$,
- množina $\{\beta \in R^{k+1} : (\beta - \widehat{\beta})^\top \mathcal{X}^\top \mathcal{X}(\beta - \widehat{\beta}) < (k+1)S^2 F_{k+1,n-k-1;1-\alpha}\}$ tvoří konfidenční množinu pro β se spolehlivostí $1-\alpha$

Testování jednotlivých parametrů

Podle předchozí věty (Zvára 2008, Věta 2.7) víme:

$$T_j = \frac{\widehat{\beta}_j - \beta_j}{S\sqrt{v_{jj}}} \sim t_{n-k-1}, \quad j = 0, \ldots, k,$$

Z toho plyne, že hypotézu $H_0 : \beta_j = 0$ zamítáme ve prospěch alternativní hypotézy $H_1 : \beta_j \neq 0$, pokud:

$$\frac{|\widehat{\beta}_j|}{S\sqrt{v_{jj}}} \geq t_{n-k-1;1-\alpha/2}.$$

Testování jednotlivých parametrů

Zvára 2008, str. 22: V případě mnohonásobné regrese při hodnocení odhadů $\widehat{\beta}_j$ a zejména t-statistik musíme být obezřetní. Odhad $\widehat{\beta}_j$ ukazuje odhad změny střední hodnoty Y při jednotkové změně j-té nezávisle proměnné, avšak při nezměněných hodnotách ostatních nezávisle proměnných. Testová statistika T_j testuje hypotézu $H_0 : \beta_j = 0$, takže ukazuje, nakolik vypovídá o chování střední hodnoty Y nad to, co o jejím chování víme z ostatních nezávisle proměnných. Testuje hypotézu, podle které j-tý regresor nesdělil o chování střední hodnoty Y nic nad to, co sdělily ostatní nezávisle proměnné v modelu již přítomné.

Příklad: Data Rubber: výstup funkce lm() v R: testové statistiky a p-hodnoty.

Testování lineární kombinace

Zajímavé hypotézy lze často formulovat jako hypotézy o hodnotě nějaké lineární kombinace parametrů modelu, např. $t^\top \beta$.

Konfidenční interval i test lze založit na statistice:

$$\frac{t^\top \widehat{\beta} - t^\top \beta}{S\sqrt{t^\top (\mathcal{X}^\top \mathcal{X})^{-1} t}} \sim t_{n-k-1}.$$

Příklad: Konfidenční interval pro hodnotu regresní přímky v pevně zvoleném bodě x.

Podmodel a testování podmodelu

Řekneme, že platí podmodel modelu $Y \sim N_n(\mathcal{X}\beta, \sigma^2 \mathcal{I}_n)$, pokud pro nějaký vektor β_0 platí $Y \sim N_n(\mathcal{X}_0\beta_0, \sigma^2 \mathcal{I}_n)$, kde lineární prostor generovaný sloupci matice $\mathcal{X}_0$ je podprostorem lineárního prostoru generovaného sloupci matice $\mathcal{X}$ a hodnost $\mathcal{X}_0$ je $r_0 < k + 1$.

Nové značení: RSS_0 je reziduální součet čtverců a $S_0^2 = RSS_0/(n - r_0)$ reziduální rozptyl v podmodelu.

Věta: Platí-li v normálním lineárním modelu podmodel, potom

$$F_0 = \frac{(RSS_0 - RSS)/(k + 1 - r_0)}{RSS/(n - k - 1)} \sim F_{k+1-r_0, n-k-1}.$$

Vytváření podmodelu: vypuštění sloupců matice $\mathcal{X}$ nebo lineární omezení na parametry.

Testování podmodelu

Z předchozí věty plyne, že hypotézu $H_0 : \beta_1 = \ldots \beta_k = 0$ zamítáme ve prospěch alternativní hypotézy $H_1 : \exists j, \beta_j \neq 0$, pokud:

$$F_0 \geq F_{k+1-r_0, n-k-1; 1-\alpha}.$$

Příklad: Data Rubber: výstup funkce lm() v R: test hypotézy $H_0 : \beta_2 = \beta_3 = 0$, funkce anova().

Ověřování předpokladů

Předpoklady kladené na náhodné chyby (nekorelovanost, konstantní rozptyl, případně normality) i na tvar modelu se nejčastěji ověřují pomocí residuí.

Už jsme měli, že v lineárním modelu $Y \sim (\mathcal{X}\beta, \sigma^2\mathcal{I}_n)$ platí $u = Y - \widehat{Y} = (\mathcal{I} - H)Y = MY \sim (0, \sigma^2 M)$, tj. střední hodnota residuí je vždy 0, ale residua mají různé rozptyly a jsou závislá.

Kromě obyčejných residuí se zavádí i residua normovaná, studentizovaná a další...

Residua: $u_i \sim (0, \sigma^2 m_{ii})$, kde m_{ii} je i-tý diagonální prvek $M = \mathcal{I} - H$, tj. $m_{ii} = 1 - h_{ii}$. V R: `resid()`.

Normovaná residua: $v_i = u_i/(S\sqrt{m_{ii}})$, v normálním modelu pro $m_{ii} > 0$ platí $Ev_i = 0$ a $Var\, v_i = 1$. V R: `rstandard()`.

Další druhy residuí lze definovat pomocí modelu bez i-tého pozorování:

$$Y_{[-i]} \sim (\mathcal{X}_{[-i]}\beta, \sigma^2\mathcal{I}_{n-1}),$$

kde $Y_{[-i]}$ a $\mathcal{X}_{[-i]}$ jsou Y a $\mathcal{X}$ bez i-tého řádku.

Studentizovaná residua: $v_i^* = u_i/(S_{[-i]}\sqrt{m_{ii}})$, za platnosti normálního modelu $v_i^* \sim t_{n-k-2}$ (Zvára 2008, str. 104). V R: `rstudent()`. Studentizovaná residua lze využít k testování „odlehlosti“ pozorování.

Další druhy residuí: nekorelovaná, parciální (s vyloučením vliv ostatních proměnných) viz Zvára (2008).

Pomocí grafů residuí lze odhalit:

- heteroskedasticitu,
- špatný tvar modelu (pokud např. jsou residua někde kladná a jinde záporná),
- důležité vynechané proměnné (pokud vidíme, že residua na nějaké vynechané proměnné závisí).

Residua lze využít k detekci odlehlých pozorování (hlavně ta studentizovaná), ale vlivná a zároveň odlehlá pozorování mohou být „maskovaná“ (viz cvičení).

V praxi vždy nezbytné:

- graf pozorování s proloženými hodnotami (ověřit, že nevychází nesmysly),
- grafy residuí (špatný tvar modelu, porušení předpokladů, odlehlá pozrování).

Testy o splnění předpokladů:

- tvar závislosti (např. test podmodelu),
- nezávislost pozorování (Durbin-Watson),
- konstantní rozptyl (homoskedasticita) - např. test shody rozptylů použitý na první a poslední třetinu residuí,
- normalita - test normality použitý na residua (Shapiro-Wilk).

Příště: problémy, které mohou nastávat i při splnění všech předpokladů (leverage—vlivná pozorování, multikolinearita).

Týden 11

Téma:

- vliv jednotlivých pozorování v lineární regresi:
  - leverage,
  - DFFITS,
  - DFBETAS,
  - Cookova vzdálenost,
- multikolinearita:
  - čísla podmíněnosti,
  - VIF.

Vlivná pozorování

$\widehat{Y} = HY$, kde $H = \mathcal{X}(\mathcal{X}^\top\mathcal{X})^{-1}\mathcal{X}^\top$

Tedy $\widehat{Y}_i = \sum_{j=1}^n h_{ij}Y_i$ a h_{ii} udává, nakolik důležité je i-té pozorování pro odhad Y_i. Takzvaná *vlivná pozorování* mají velkou hodnotu h_{ii} (a automaticky i malý rozptyl $Var\, u_i/\sigma = m_{ii} = 1 - h_{ii}$).

h_{ii} se označuje jako `hat.diag` nebo `leverage` (a závisí pouze na $\mathcal{X}$).

Víme, že $\text{rank}(H) = p = k + 1$. H je idempotentní a tedy její vlastní čísla λ_i jsou 0 nebo 1 (těch musí být p). Z toho plyne, že $\text{tr}(H) = \sum \lambda_i = p$.

Průměrná hodnota h_{ii} je p/n a i-té pozorování bude (v R) označené jako vlivné, pokud $h_{ii} > 3p/n$ (viz R funkce `influence.measures()`).

Jak se změní odhady při vynechání i-tého pozorování?

DFBETAS:

$$\Delta_i(\beta_j) = \frac{\widehat{\beta}_j - \widehat{\beta}_{[-i]j}}{S_{[-i]}\sqrt{v_{jj}}}$$

měří vliv i-tého pozorování na odhad j-tého regresního koeficientu.

Kritérium: $|\Delta_i(\beta_j)| > 1$.

DFFITS:

$$\Delta_i(EY_i) = \frac{\widehat{Y}_i - \widehat{Y}_{[-i]i}}{\sqrt{Var\,\widehat{Y}_i}} = \cdots = \sqrt{\frac{h_{ii}}{m_{ii}}}v_i^*$$

Kritérium: $|\Delta_i(EY_i)| > 3\sqrt{p/(n-p)}$.

Jak se změní celý vektor $\widehat{Y}$ při vynechání i-tého pozorování?

Cookova vzdálenost:

$$D_i = \frac{1}{pS^2}\|\widehat{Y} - \widehat{Y}_{[-i]}\|^2 = \cdots = v_i^2\frac{h_{ii}}{m_{ii}}\frac{1}{p},$$

kde normované residuum v_i měří „odlehlost“ a h_{ii}/m_{ii} „vlivnost“.

Kritérium: $F_{p,n-p}(D_i) > 0.5$, kde $F_{p,n-p}$ je distr. fce rozdělení $F_{p,n-p}$.

Jak se změní přesnost odhadů při vynechání i-tého pozorování?

COVRATIO:

$$\text{COVRATIO} = \frac{|\widehat{Var\,\beta}_{[-i]}|}{|\widehat{Var\,\beta}|} = \cdots = \frac{1}{m_{ii}} \left(\frac{n-p-v_i^2}{n-p-1} \right)^p$$

Kritérium: $|1 - \text{COVRATIO}| > 3p/(n-p)$.

Příklad

Příklad:

`Animals (MASS): lm(), plot.lm(), influence.measures().`

Jsou splněné předpoklady normálního lineárního modelu? Která pozorování jsou vlivná?

Jak se situace změní při použití logaritmické transformace?

Multikolinearita

Trocha teorie: v modelu $Y \sim (\mathcal{X}\beta, \sigma^2\mathcal{I})$ platí:

$$E\|\widehat{Y}\|^2 = \|\mathcal{X}\beta\|^2 + \sigma^2 \,\text{rank}(\mathcal{X}).$$

Má-li $\mathcal{X}$ plnou hodnost, pak:

$$E\|\widehat{\beta}\|^2 = \|\beta\|^2 + \sigma^2 \,\text{tr}(\mathcal{X}^\top\mathcal{X})^{-1}.$$

Střední hodnota délky odhadu $\widehat{Y}$ tedy závisí pouze na hodnosti $\mathcal{X}$, zatímco střední hodnota délky $\widehat{\beta}$ závisí na stopě matice $(\mathcal{X}^\top\mathcal{X})^{-1}$, prakticky tedy na součtu singulárních hodnot $\mathcal{X}$ umocněných na -2.

Detekce multikolinearity

Pokud má matice $\mathcal{X}$ hodně malou singulární hodnotu, tak odhad $\widehat{\beta}$ bude nabývat velkých hodnot (zároveň bude mít i velký rozptyl) a odhady regresních koeficientů nebudou mít rozumnou interpretaci.

Značení: $d_1 \geq d_2 \geq \cdots \geq d_k$ jsou singulární hodnoty $\mathcal{X}_0$ (matice modelu s centrovanými regresory bez absolutního členu).

Číslo podmíněnosti:

$$\kappa(\mathcal{X}) = (d_1/d_k)^2.$$

Kritérium: $\kappa(\mathcal{X}) > 25$.

Indexy podmíněnosti:

$$\eta_j = (d_1/d_j)^2.$$

Obě tyto charakteristiky ale (bohužel) závisí na měřítku.

Variance inflation factors

Ve standardizovaném regresním modelu lze odvodit:

$$\widehat{Var}\,\hat{\beta}_j^* = \frac{1-R^2}{n-k-1}\frac{1}{1-R_j^2},$$

kde $\hat{\beta}_j^*$ je odhad j-tého regresního koeficientu ve standardizovaném regresním modelu, R^2 je koeficient determinace a R_j^2 je koeficient determinace při regresi X_j na ostatní vysvětlující proměnné.

$\mathrm{VIF}_j = 1/1-R_j^2$ udává, kolikrát se zvětší rozptyl odhadu $\hat{\beta}_j^*$ kvůli korelacím mezi vysvětlujícími proměnnými.

Implementace v R: funkce `vif()` v knihovně `car`.

Poznámky

Problémy s porušením předpokladů lze řešit:

- transformací závisle proměnné (heteroskedasticita, nenormalita),
- úpravou regresního modelu,
- přidáním vysvětlující proměnné,
- použitím obecného lineárního modelu (známé korelace) nebo zobecněného lineárního modelu (nenormalita),
- vypuštěním odlehlých pozorování nebo použitím robustního odhadu.

Problémy s multikolinearitou lze řešit:

- naplánováním experimentu,
- použitím „stabilnějšího“ odhadu (hřebenová regrese, LASSO),
- regresí na hlavních komponentách (problémy s interpretací).

Problémy způsobené závislostí mezi náhodnými chybami lze řešit:

- metodami časových řad,
- modely s náhodnými efekty (Bayesovské hierarchické modely, GEE, apod.)

Týden 12

Téma:

- úvod do plánování experimentů,
- počítačové experimenty,
- regresní experimenty.

Navrhování experimentů

- Návrh experimentu $\times$ vyhodnocení experimentu.
- Bias $\times$ randomizace?
- Podíl statistika na návrhu experimentu (často je statistik přizván až po skončení sběru dat).
- Plánování rozsahu výběru (v praxi často *ex post*).
- Plán experimentu musí počítat s finančními a zákonnými omezeními i s fyzikálními zákony.

Např. v regresní analýze je rozptyl $\hat{\beta}$ určen reziduálním rozptylem σ^2 (který většinou neumíme ovlivnit) a maticí experimentu $\mathcal{X}$, která je určena tím, jak experiment naplánujeme.

Design of computer experiments

K řešení komplexních analytických problémů se čím dál častěji používají složité počítačové simulace.

Předpokládejme, že chování nějakého zařízení (nebo procesu) závisí na náhodném vektoru $\mathbf{X} = (X_1, \ldots, X_s)^\top$; FW93 [Fang, K. T., & Wang, Y. (1993). Number-theoretic methods in statistics. CRC Press] uvádí příklad elektrického obvodu, jehož chování závisí na charakteristikách, které se obvod od obvodu mírně (náhodně) liší (další příklady: umělé bytosti, rozmístění větrných elektráren, řízené střely).

Design of computer experiments

Pro příslušné zařízení je pak vyvinut matematický model (např. systém diferenciálních rovnic), který umožňuje naprogramovat odpovídající počítačovou simulaci. Výsledek (jako funkce zadaných vstupních parametrů) obvykle nebývá náhodný a nemá tedy smysl stejnou počítačovou simulaci opakovat; místo toho se snažíme najít tzv. „space-filling design“.

Často nás zajímá střední hodnota (nebo kvantil) nějaké charakteristiky, tj. např. $Eh(X)$ (spočítaná “přes náhodné vlivy např. počasí” pro pevné hodnoty “kontrolovaných parametrů”).

Odhad střední hodnoty

Jeden možný odhad, založený na výběrovém průměru nasimulovaných hodnot $h(\mathbf{X})$ lze snadno získat metodou Monte Carlo:

1. generujeme vektory $\mathbf{X}_i$ z rozdělení X,
2. $Eh(x)$ odhadneme výběrovým průměrem $\bar{h} = \sum_{i=1}^{n} h(\mathbf{X}_i)/n$.

Poznámka 1: podobné metody se používají i při numerickém integrování (věrohodnostní funkce pro GLM s náhodnými efekty) nebo při experimentech typu inventarizace lesa (odhad průměru nebo úhrnu ve vícerozměrném prostoru).

Poznámka 2: odhad metodou Monte Carlo je konzistentní, ale není příliš eficientní—různí autoři proto navrhli jiné způsoby generování $\mathbf{X}_i$, které vedou na odhady $\bar{h}$ s menším rozptylem (tyto metody se většinou snaží generovat hodnoty $\mathbf{X}_i$ víc “rovnoměrně”).

Latin hypercube sampling (LHS)

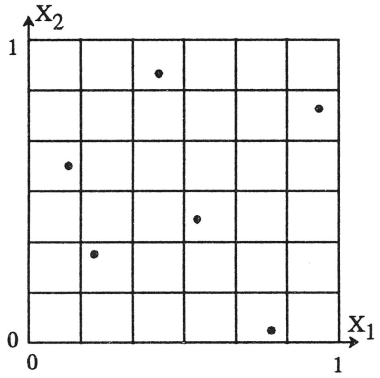
Princip metody “latinských hyperkostek” se trochu podobá tzv. latinským čtvercům. Cílem je zaručit, aby výsledný výběr $\mathbf{X}_i$ rovnoměrně pokrýval marginální rozdělení všech složek náhodného vektoru $\mathbf{X} = (X_1, \ldots, X_s)^\top$.

Předpokládejme, že distribuční funkce $\mathbf{X}$ je $F(x) = \prod_{i=1}^{s} F_k(x_k)$.

Jedna varianta postupu při generování LHS je:

1. Generujeme matici P dimenze $(n \times s)$, jejíž každý sloupec je nezávislá náhodná permutace $\{0, 1, \ldots, n-1\}$.
2. Generujeme matici U dimenze $(n \times s)$, jejíž prvky jsou iid $U(0,1)$ (nezávislé s P).
3. Zvolíme $\{\mathbf{x}_j = (x_{j1}, \ldots, x_{js})^\top, j = 1, \ldots, n\}$, kde
$$x_{jk} = F_k^{-1}\{(p_{jk} + u_{jk})/n\}$$
je výběr rozsahu n z rozdělení $F(x)$ získaný metodou LHS.

Příklad: LHS pro $s = 2$ a $n = 6$.



Uniform random design (URD)

Další možnost je použít metodu glp (good lattice points). Postup (pro generování rovnoměrně rozložených bodů uvnitř jednotkové krychle) je následující:

1. Generujeme glp množinu $\{a_k \in [0,1]^s, k = 1, \ldots, n\}$ pomocí generujícího vektoru $(n; h_1, \ldots h_s)$.
2. Generujeme n náhodných vektorů $u_i \in \mathbb{R}^s$ s rovnoměrným rozdělením na $(-1,1)^s$,
3. Výběr metodou URD je $\{x_k, k = 1, \ldots, n\}$, kde

$$x_k = a_k + u_k/2n.$$

FW93 [Theorem 5.3–5.4] ukazují, že odhad $\bar{h}_n$ je asymptoticky nevychýlený ($|E\bar{h}_n - E(h(\mathbf{X}))| = O(n^{-1}\log^s n)$) a $Var(\bar{h}_n) = O(n^{-2}\log^{2s} n)$.

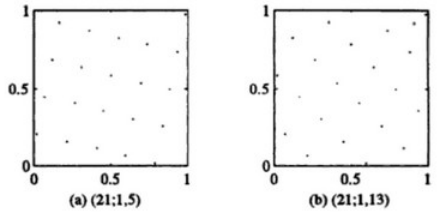
Good lattice points (glp)

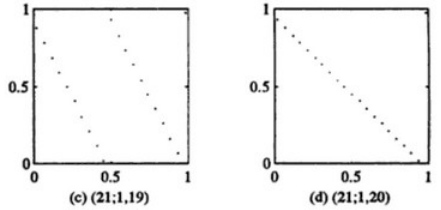
Pokud $\{x\}$ označuje desetinnou část čísla x, pak x_{ki} snadno vypočteme jako:

$$x_{ki} = \left\{ \frac{2kh_i - 1}{2n} \right\}.$$

Věta: Pro každé prvočíslo p existuje vektor přirozených čísel $h_p = (h_1, \ldots, h_s)$ takový, že množina síťových bodů generujícího vektoru $(p; h_1, \ldots, h_s)$ má diskrepanci $D(p) < c(s)p^{-1}(\log p)^s$.

Příklad: 21 bodů získaných pomocí různých generujících vektorů (FW93)





Generující vektory glp pro malé n a $s \in \{2, 3, 4\}$.

Table A.13 $s = 2, h_1 = 1$

n	5	7	9	11	13	15	17	19	21	23	25	27	29	31
h_2	2	3	4	7	5	11	5	14	13	9	7	16	23	21

Table A.14 $s = 3, h_1 = 1$

n	5	7	9	11	13	15	17	19	21	23	25	27	29	31
h_2	2	2	2	3	3	2	3	3	4	15	8	20	16	11
h_3	4	4	4	5	9	7	9	9	10	18	14	22	24	28

Table A.15 $s = 4, h_1 = 1$

n	7	9	11	13	15	17	19	21	23	25	27	29	31
h_2	2	2	2	6	2	2	5	2	2	4	5	4	15
h_3	3	4	5	8	4	4	7	10	5	6	17	6	19
h_4	6	7	7	10	8	8	9	17	10	9	25	16	22

Jednoduchá lineární regrese

Příklad:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad \varepsilon_i \text{ jsou iid}, E\varepsilon_i = 0, Var\, \varepsilon_i = \sigma^2.$$

Víme, že $Var\, \hat{\beta} = \sigma^2(\mathcal{X}^\top \mathcal{X})^{-1} = \dots$ tedy:

- $\hat{\beta}_0$ a $\hat{\beta}_1$ jsou korelované (korelaci lze „vynulovat“ vycentrováním vysvětlující proměnné),
- $$Var\, \hat{\beta}_1 = \frac{\sigma^2}{n} \frac{n}{\sum (x_i - \bar{x}_n)^2}.$$

Pokud tedy chceme získat co nejpřesnější odhad $\hat{\beta}_1$, musí být $\sum (x_i - \bar{x}_n)^2 / n$ co největší (pokud x_i můžeme volit v intervalu $[a, b]$, tak polovina měření $= a$ a druhá polovina měření $= b$).

Příklad: Na laboratorní váze potřebujeme co nejpřesněji určit váhu tří předmětů (θ_1, θ_2, θ_3). Předměty můžeme vážit jednotlivě i „po skupinách“.

Vyjdeme z následujících předpokladů:

- váhy nejsou zkalibrované a všechna měření tedy ovlivňuje systematická chyba měření θ_4.
- všechny ostatní vlivy považujeme za náhodné (s konstantním rozptylem).

Jednotlivé měření můžeme tedy zapsat jako:

$$y_i = \{F\}_{i1}\theta_1 + \{F\}_{i2}\theta_2 + \{F\}_{i3}\theta_3 + \theta_4 + \varepsilon_i,$$

kde $E\varepsilon_i = 0$ a $Var\, \varepsilon_i = \sigma^2$.

První návrh experimentu používá tato čtyři měření:

$$\begin{aligned} y_1 &= \theta_4 + \varepsilon_1 \\ y_2 &= \theta_1 + \theta_4 + \varepsilon_2 \\ y_3 &= \theta_2 + \theta_4 + \varepsilon_3 \\ y_4 &= \theta_3 + \theta_4 + \varepsilon_4 \end{aligned}$$

Pro matici experimentu máme

$$F = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}, \quad F^{-1} = \begin{pmatrix} -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}.$$

Ze soustavy normálních rovnic snadno plyne, že $\hat{\theta} = F^{-1}y$ a následně $Var\, \hat{\theta}_i = 2\sigma^2$ pro $i = 1, 2, 3$.

Druhý návrh experimentu používá tato čtyři měření:

$$\begin{aligned} y_1 &= \theta_1 + \theta_2 + \theta_3 + \theta_4 + \varepsilon_1 \\ y_2 &= \theta_1 + \theta_4 + \varepsilon_2 \\ y_3 &= \theta_2 + \theta_4 + \varepsilon_3 \\ y_4 &= \theta_3 + \theta_4 + \varepsilon_4 \end{aligned}$$

Pro matici experimentu máme

$$F = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}, \quad F^{-1} = \frac{1}{2} \begin{pmatrix} 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ -1 & 1 & 1 & 1 \end{pmatrix}.$$

Máme tedy $\hat{\theta}_1 = (y_1 + y_2 - y_3 - y_4)/2$, ..., $\hat{\theta}_3 = (y_1 - y_2 - y_3 + y_4)/2$ a $Var\,\hat{\theta}_i = \sigma^2$ pro $i = 1, 2, 3$.

Lineární model (více formálně)

Předpokládáme, že měřená výstupní veličina $y(x)$ v pokusu $x \in \mathcal{H}$ (kde $\mathcal{H}$ označuje množinu všech možných pokusů) má tvar:

$$y(x) = \theta_1 f_1(x) + \ldots \theta_m f_m(x) + \varepsilon(x),$$

tj. zkráceně

$$y(x) = \theta^\top f(x) + \varepsilon(x),$$

kde $E\varepsilon(x) = 0$ a $Var\,\varepsilon(x) = \sigma^2(x)$. Předpokládáme přitom, že rozptyly $\sigma^2(x)$ jsou známé nebo že alespoň $\sigma^2(x) = kw(x)$, kde $w(x)$ známe.

Nechť N označuje předepsaný (dovolený) počet pokusů, pak každou N-tici bodů $x^{(1)}, \ldots, x^{(N)}$ množiny $\mathcal{H}$ považujeme za *návrh experimentu se stanoveným rozsahem* N. Některé pokusy se přitom mohou několikrát opakovat. Předpokládáme, že pokusy se opakují nezávisle a nezáleží tedy na jejich pořadí.

Informační matice

Pokud $x^{(1)}, \ldots, x^{(N)}$ je návrh experimentu s rozsahem N, pak pro vektor naměřených hodnot $y = (y(x^{(1)}), \ldots, y(x^{(N)}))^\top$ máme

$$y = F\theta + \varepsilon,$$

kde $\{F\}_{ij} = f_j(x^{(i)})$, pro $i = 1, \ldots, N$ a $j = 1, \ldots, m$.

Informační matice je

$$M = F^\top \Sigma^{-1} F = \sum_{i=1}^{N} f(x^{(i)}) f^\top(x^{(i)}) \sigma^{-2}(x^{(i)}),$$

kde Σ^2 je diagonální matice s prvky $\sigma^2(x^{(i)})$ na diagonále.

Matice M^{-1} (případně $h^\top M^- h$) je kovarianční matice náhodného vektoru $\hat{\theta}$ (případně odhadu odhadnutelné funkce $h^\top\theta$). Informační matice M tedy lze využít při určování kvality návrhu experimentu.

Normovaná návrhová míra a normovaná informační matice

Nechť $x^{(1)}, \ldots, x^{(N)}$ je návrh experimentu s rozsahem N. *Normovanou návrhovou míru* ξ přidruženou k $x^{(1)}, \ldots, x^{(N)}$ definujeme jako:

$$\xi(x) = N(x)/N; \quad (x \in \mathcal{H}),$$

kde $N(x)$ je počet opakování pokusu x v návrhu $x^{(1)}, \ldots, x^{(N)}$.

Zřejmě platí $M = NM(\xi)$, kde

$$M(\xi) = \sum_{x \in \mathcal{H}} f(x) f^\top(x) \sigma^{-2}(x) \xi(x)$$

je tzv. *normovaná informační matice*.

Normovaná návrhová míra a normovaná informační matice

Pro normovanou návrhovou míru ξ (nebo stručněji *návrh* ξ) platí:

1. $\xi(x) \geq 0$ pro $x \in \mathcal{H}$,
2. $\sum_{x \in \mathcal{H}} \xi(x) = 1$,
3. množina $\{x; x \in \mathcal{H}, \xi(x) > 0\}$ je konečná.

V této kapitole budeme každou funkci ξ definovanou na $\mathcal{H}$ a splňující tyto tři vlastnosti považovat za normovanou návrhovou míru.

Návrh ξ interpretujeme tak, že měření probíhají pouze v těch pokusech, pro které je $\xi(x) > 0$ a číslo $\xi(x)$ je přímo úměrné počtu nezávislých opakování pokusu x.

Porovnávání návrhů regresních experimentů

Dva návrhy ξ a η považujeme za *ekvivalentní*, pokud $M(\xi) = M(\eta)$, tj. právě tehdy, když $Var_\xi(h^\top \hat{\theta}) = Var_\eta(h^\top \hat{\theta})$, $\forall h \in \mathbb{R}^m$.

Návrh ξ je *stejnoměrně lepší*, než návrh η, pokud

$$Var_\xi(h^\top \hat{\theta}) \leq Var_\eta(h^\top \hat{\theta}), \quad \forall h \in \mathbb{R}^m.$$

To platí právě když je matice $M(\xi) - M(\eta)$ je pozitivně semidefinitní ($u^\top [M(\xi) - M(\eta)] u \geq 0$, $\forall u \in \mathbb{R}^m$). Pokud parametr $h^\top \theta$ není při ξ odhadnutelný, definujeme $Var_\xi(h^\top \hat{\theta}) = \infty$.

Stejnoměrně nejlepší návrh ξ obecně neexistuje a místo toho se obvykle optimalizují jistá předem zvolená *kritéria optimality experimentu* (obvykle rozumné míry "velikosti" informační matice).

Elfvingova metoda

Elfvingova metoda minimalizace rozptylu odhadu $h^\top \hat{\theta}$ (*d-optimalita*) je následující:

Určíme množinu

$$T = \left\{ \frac{f(x)}{\sigma(x)} : x \in \mathcal{H} \right\} \bigcup \left\{ -\frac{f(x)}{\sigma(x)} : x \in \mathcal{H} \right\}$$

a její konvexní obal S.

Označme p přímku procházející počátkem a rovnoběžnou s vektorem h. Označme P bod, ve kterém přímka p protíná hranici množiny S. Bod P lze zapsat jako konvexní kombinaci bodů množiny T, tj. $P = \sum_{i=1}^n \lambda_i f(x^{(i)})$, která zároveň určuje d-optimální návrh $\xi^*(x^{(i)}) = |\lambda_i|$.

Příklad: Na zkušební dráze se ověřuje tah motoru při zrychlování automobilu. Zjednodušeně můžeme pohyb automobilu po testovací dráze popsat funkcí

$$s(x) = vx + zx^2/2,$$

kde $s(x)$ je poloha automobilu v čase x, v je rychlost v čase 0 a z je zrychlení. Parametry v a z neznáme. Použité měřící zařízení umožňuje změřit polohu automobilu 10× v časovém úseku 0–10 s (možná jsou i vícenásobná měření polohy ve stejném čase).

Použijeme model

$$Ey(x_i) = vx_i + \frac{z}{2}x_i^2 = \theta_1 x_i + \theta_2 x_i^2,$$

kde $Var\, y(x_i) = \sigma^2$ (a $f(x_i) = (x_i, x_i^2)^\top$) s informační maticí

$$M = \sum_{k=1}^{10} f(x_k) f^\top(x_k) \sigma^{-2} = \frac{1}{\sigma^2} \begin{pmatrix} \sum x_k^2 & \sum x_k^3 \\ \sum x_k^3 & \sum x_k^4 \end{pmatrix}.$$

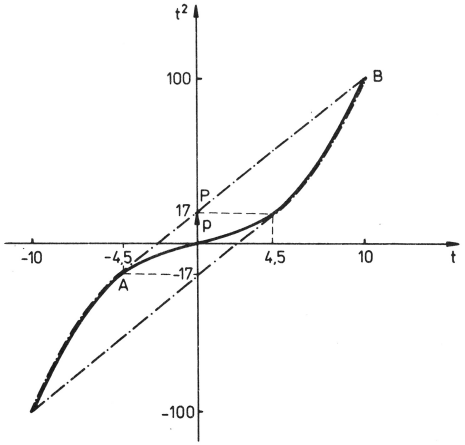
Nejdříve chceme nalézt optimální experiment pro určení zrychlení. V takovém případě je vhodné zvolit kritérium d-optimality pro vektor $h = (0,1)^\top$, tj. minimalizovat funkci

$$\Phi_1(M) = (0,1)M^{-1}\begin{pmatrix}0\\1\end{pmatrix}.$$

Na hledání d-optimálního návrhu můžeme použít Elfvingovu metodu.

Množina T je na obrázku vyznačena plnou čarou, její konvexní obal (množina S) je ohraničena čerchovanou čarou (v originále „bodkočiarkovane“).

Na obrázku je vyznačen bod P jako průsečík hranice množiny S a vektoru h. Bod P pak vyjádříme jako lineární kombinace bodů A a B z množiny T.



Bod P je definován jako průsečík hranice množiny S a vektoru h a lze jej zapsat jako konvexní kombinaci bodů A a B:

$$P = \begin{pmatrix}p_1\\p_2\end{pmatrix} = \begin{pmatrix}0\\17\end{pmatrix} = 0.7\begin{pmatrix}-4.5\\-4.5^2\end{pmatrix} + 0.3\begin{pmatrix}10\\10^2\end{pmatrix}$$

d-optimální návrh tedy je:

$$\begin{aligned}\xi^*(4.5) &= 0.7\\ \xi^*(10) &= 0.3\end{aligned}$$

Informační matice návrhu

$$M(\xi^*) = \sigma^{-2}\left[\begin{pmatrix}(4.5)^2 & (4.5)^3\\(4.5)^3 & (4.5)^4\end{pmatrix}0.7 + \begin{pmatrix}10^2 & 10^3\\10^3 & 10^4\end{pmatrix}0.3\right] = \dots$$

Varianční matice odhadu $\hat\theta$ je

$$M^{-1}(\xi^*) = \sigma^2\begin{pmatrix}0.2555 & 0.0283\\0.0283 & 0.0001\end{pmatrix}$$

a rozptyl odhadu parametru θ^2 je $10^{-4}\sigma^2$.

Doporučený postup při navrhování experimentu

- **Stanovení modelu experimentu:** co můžeme měřit; vztah mezi pozorováními a neznámými parametry; určení přesnosti měření (stačí až na multiplikativní konstantu); co chceme odhadovat (volba kritéria optimality).
- **Výpočet optimálního návrhu experimentu:** podle odborné literatury; porovnání s nějakým jednoduchým (rozumným) návrhem experimentu.
- **Ověření vypočteného návrhu experimentu:** je návrh „dobrý“ i podle jiných kritérií optimality?; zvážit drobné úpravy, které příliš nezhorší vlastnosti a přitom experiment zjednoduší; ověřit proveditelnost experimentu.

POZOR: optimální návrh experimentu nemusí umožňovat ověření předpokladů použitého modelu (např. o tvaru regresní přímky)!

Software a cvičení

Knihovna `AlgDesign` v R: funkce `optFederov()` a `optMonteCarlo()`.

Cvičení 1: Pomocí funkcí knihovny `AlgDesign` zkuste ověřit návrh experimentu pro odhad zrychlení (i pro jiné odhadované parametry).

Cvičení 2: Zkuste odvodit optimální návrh experimentu pro vážení předmětů na dvoumiskové laboratorní váze (tj. pokud se vážené předměty dají pokládat na obě misky).

Týden 13

Téma:

- úvod do logistické regrese,
- odhad a interpretace parametrů,
- deviance,
- test nezávislosti v kontingenční tabulce.

Kontingenční tabulka

V sociálních vědách se často vyskytují nominální (faktorové) proměnné.

Proměnná Z má I úrovní.

Proměnná Y má J úrovní.

Dohromady máme IJ kombinací úrovní faktorů Z a Y.

Kontingenční tabulka: sečteme výskyty jednotlivých kombinací (Z, Y) ve výběru a výsledek zaneseme do tabulky s I řádky a J sloupci.
V každém *políčku* je uveden počet měření s odpovídající kombinací hodnot Z a Y.

Příklad:

$$\mathcal{X} = \left(\begin{array}{ccc|c} 4 & 0 & 2 & 6 \\ 0 & 1 & 1 & 2 \\ 1 & 1 & 4 & 6 \\ \hline 5 & 2 & 7 & 14 \end{array}\right)$$

← Finance
← Energetika
← HiTech

↑ Frankfurt
↑ Berlín
↑ Mnichov

Sdružené rozdělení: $\pi_{ij} = P(Z = i, Y = j)$ je pravděpodobnost, že Z se rovná i a zároveň Y se rovná j.

Marginální rozdělení Z: $\pi_{i\cdot}$ je pravděpodobnost, že Z se rovná i.

Marginální rozdělení Y: $\pi_{\cdot j}$ je pravděpodobnost, že Y se rovná j.

Nezávislost

Vztah mezi Z a Y může být popsán jejich sdruženým rozdělením, podmíněným rozdělením Z za podmínky Y nebo podmíněným rozdělením Y za podmínky Z.

Z a Y jsou nezávislé právě tehdy, když pro všechna i a j platí:
$\pi_{i|j} = \pi_{ij}/\pi_{\cdot j} = \pi_{i\cdot}$,
$\pi_{j|i} = \pi_{ji}/\pi_{i\cdot} = \pi_{\cdot j}$
nebo $\pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}$.

π_{ij} označuje nepozorované (neznámé) skutečné pravděpodobnosti.

Pozorované relativní četnosti budeme značit $p_{ij} = x_{ij}/x_{\bullet\bullet}$, kde x_{ij} jsou absolutní četnosti a $x_{\bullet\bullet}$ je rozsah výběru.

Pro odvození testu může být důležité, jakým způsobem kontingenční tabulka vznikala.

- Poisson sampling
  (vše je náhodné)
- Multinomial sampling
  (celkový počet pozorování, tj. součet x_{ij} je pevně daný)
- Independent multinomial sampling
  (celkový počet pozorování v každém řádku nebo sloupci je pevně daný)

Věrohodnostní funkce sice závisí na tom, jak kontingenční tabulka vznikla, ale maximálně věrohodný odhad π_{ij} naštěstí vyjde pokaždé stejně
$\widehat{\pi}_{ij} = p_{ij} = x_{ij}/x_{\bullet\bullet}$,

Za předpokladu nezávislosti získáme maximálně věrohodné odhady:

$$\tilde{\pi}_{ij} = p_{i\bullet}p_{\bullet j} = (x_{i\bullet}x_{\bullet j})/x_{\bullet\bullet}^2.$$

Obvyklým způsobem lze nyní odvodit test nezávislosti poměrem věrohodností v kontingenční tabulce (Likelihood-Ratio Test of Independence). Testová statistika je:

$$G^2 = -2\log\frac{\prod_i\prod_j(x_{i\bullet}x_{\bullet j})^{x_{ij}}}{x_{\bullet\bullet}^{x_{\bullet\bullet}}\prod_i\prod_j x_{ij}^{x_{ij}}} = 2\sum\sum x_{ij}\log(x_{ij}/E_{ij}),$$

kde E_{ij} je odhad očekávaných četností za předpokladu nezávislosti ,
$E_{ij} = \tilde{\pi}_{ij}x_{\bullet\bullet} = (x_{i\bullet}x_{\bullet j})/x_{\bullet\bullet}$.

Za platnosti nulové hypotézy (nezávislost) má G^2 rozdělení $\chi^2_{(I-1)(J-1)}$.

Pearsonův χ^2-test nezávislosti

Nejčastěji používaný test nezávislosti v kontingenční tabulce je založen na testové statistice:

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(x_{ij}-E_{ij})^2}{E_{ij}}.$$

Za platnosti nulové hypotézy (nezávislost) má testová statistika X^2 rozdělení $\chi^2_{(I-1)(J-1)}$.

Jako předpoklad pro použití tohoto testu se obvykle požaduje, aby všechny očekávané četnosti byly alespoň 5. Pokud jsou četnosti nižší, lze použít např. Fisherův přesný test.

EYE/HAIR	black	brown	red	blond
d.brown	68	119	26	7
l.brown	15	54	14	10
green	5	29	14	16
blue	20	84	17	94

```
> chisq.test(eyehair)

Pearson's Chi-squared test

data:  eyehair
X-squared = 138.2898, df = 9, p-value < 2.2e-16
```

Zamítáme nezávislost a jako další krok bychom se mohli podívat, které pozorované četnosti jsou příliš malé nebo příliš velké.

Logistická regrese

Lineární model nelze dobře použít, pokud závisle proměnná může nabývat pouze dvou hodnot (0/1, úspěch/neúspěch, smrt/přežití).

Logistická regrese je určena pro situace, ve kterých má závisle proměnná Alternativní (nebo Binomické) rozdělení.

Příklad: `Birthwt(MASS)`, závisle proměnná = nízká porodní váha.

Alternativní (Bernoulliho) rozdělení $A(p_i)$:

$$P(Y_i = 1) = p_i, \qquad P(Y_i = 0) = q_i = 1 - p_i.$$

Alternativní rozdělení $A(p_i)$ je speciální případ Binomického rozdělení $B(n, p_i)$ s $n = 1$.

Model logistické regrese

Mějmě závisle proměnnou $Y_1, \ldots, Y_n$ (0/1) a k vysvětlujících proměnných x_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, k$.

Model logistické regrese: Předpokládáme, že $Y_i \sim A(p_i)$ pro $i = 1, \ldots, n$, kde

$$E(Y_i|x_i) = p_i = \frac{1}{1 + \exp\{-(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})\}},$$

tj. ekvivalentně

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}.$$

Pozorujeme Y_i a x_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, k$ a chceme odhadnout neznámé parametry $\beta_0, \beta_1, \ldots, \beta_k$.

Interpretace koeficientů

Binární vysvětlující proměnná
(Data lze zapsat jako kontingenční tabulku 2×2.)

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 x_{i1}$$

Pro $x_{i1} = 0$ máme:

$$p_i = \frac{1}{1 + \exp(-\beta_0)} = \frac{\exp\beta_0}{1 + \exp\beta_0} \quad \text{a} \quad 1 - p_i = \frac{1}{1 + \exp\beta_0}.$$

Tj. šance $p_i/(1 - p_i) = \exp(\beta_0)$ a koeficient

$$\beta_0 = \log\left(\frac{p_i}{1 - p_i}\right)$$

je logaritmická šance (log odds) na úspěch při $x_{i1} = 0$.

Interpretace koeficientů

Pro $x_{i1} = 1$ máme:

$$p_i = \frac{1}{1+\exp(-\beta_0-\beta_1)} = \frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)} \text{ a } 1-p_i = \frac{1}{1+\exp(\beta_0+\beta_1)}.$$

Tj. šance $p_i/(1-p_i) = \exp(\beta_0+\beta_1)$ a koeficient

$$\beta_0+\beta_1 = \log\left(\frac{p_i}{1-p_i}\right)$$

je logaritmická šance (log odds) na úspěch při $x_{i1} = 1$.

Koeficient β_1 tedy udává, o kolik se zvýší logaritmická šance na úspěch při $x_{i1} = 1$ (v porovnání se šancí na úspěch při $x_{i1} = 0$).

Interpretace koeficientů

Poměr šancí (odds ratio) a logaritmický poměr šancí (log odds ratio)

Poměr šancí při $x_{i1} = 1$ a $x_{i1} = 0$ je:

$$\frac{\exp(\beta_0+\beta_1)}{\exp(\beta_0)} = \exp(\beta_1)$$

Koeficient β_1 se označuje jako logaritmický poměr šancí (log odds ratio).

Příklad: Pokud máme β_1 dané, pak se šance na úspěch (při $x_{i1} = 1$) zvýší $\exp(\beta_1)$-krát. Podle Taylorova rozvoje můžeme aproximovat $\exp(\beta_1) \doteq 1+\beta_1 = 100\% + \beta_1 \times 100\%$ a $\beta_1 \times 100$ zhruba říká, o kolik procent se (při $x_{i1} = 1$) zvýší šance na úspěch.
Např. $\beta_1 = 0.1$ odpovídá zvýšení šance na úspěch asi o 10%.

Interpretace koeficientů

Spojitá vysvětlující proměnná

šance, když $x_i = x$: $\exp(\beta_0+\beta_1 x)$

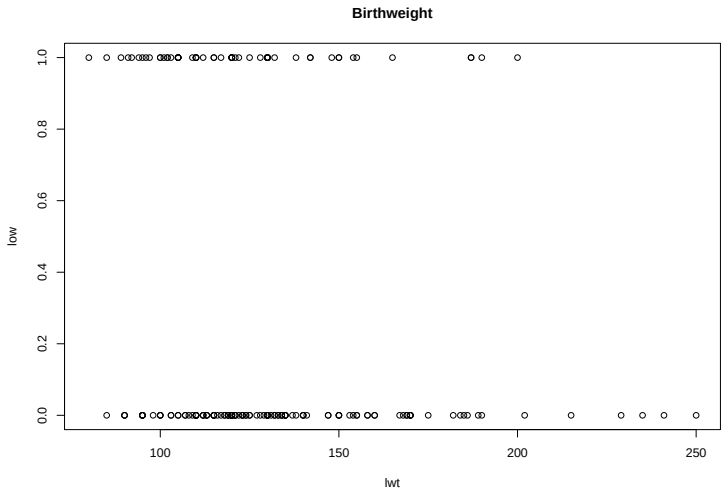
šance při $x_i = x+1$: $\exp\{\beta_0+\beta_1(x+1)\}$

podíl šancí při jednotkovém zvýšení x_i: $\exp(\beta_1)$

logaritmický podíl šancí (při jednotkovém zvýšení x_i): β_1

Více vysvětlujících proměnných

Koeficient β_j udává vliv j-té vysvětlující proměnné na pravděpodobnost (šanci) úspěchu za předpokladu, že hodnoty ostatních vysvětlujících proměnných se nemění.

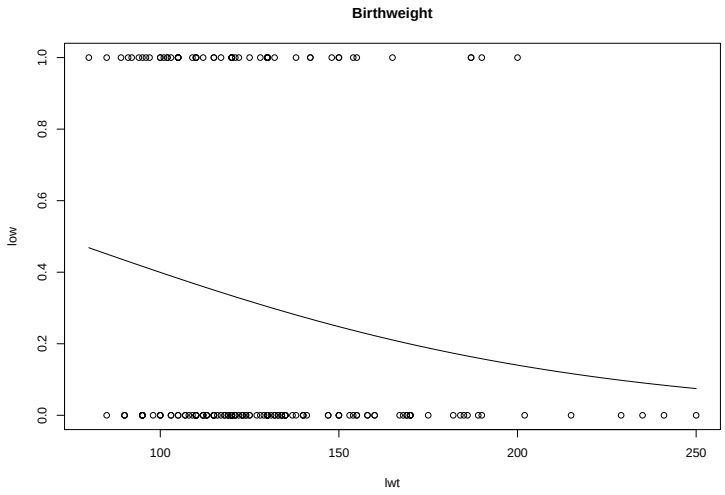


```
Call:  glm(formula = low ~ lwt, family = binomial)

Coefficients:
(Intercept)          lwt
    0.99831     -0.01406

Degrees of Freedom: 188 Total (i.e. Null);  187 Residual
Null Deviance:      234.7
Residual Deviance: 228.7  AIC: 232.7
```

$$p_i = \frac{1}{1 + \exp(-0.646733 + 0.002577 \times \text{lwt}_i)}$$



Odhad parametrů (MLE)

Příspěvek i-tého pozorování do věrohodnostní funkce:

$$L(p_i|Y_i) = p_i^{Y_i}(1-p_i)^{1-Y_i}.$$

Věrohodnost je $L(\beta|Y,x) = \prod p_i^{Y_i}(1-p_i)^{1-Y_i}$, kde p_i závisí na β_j a x_{ij}.

Logaritmická věrohodnost:

$$\begin{aligned} l(\beta|Y,x) &= \sum Y_i \log(p_i) + \sum (1-Y_i)\log(1-p_i) \\ &= \sum Y_i \log\{p_i/(1-p_i)\} + \sum \log(1-p_i) \\ &= \sum Y_i(\beta_0 + +\beta_1 x_{i1} + \cdots + \beta_k x_{ik}) + \sum \log\{1-p_i(\beta)\} \end{aligned}$$

Maximálně věrohodné odhady i odhad jejich asymptotické varianční matice získáme numericky (díky teorii maximální věrohodnosti).


```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0951  -0.9022  -0.8018   1.3609   1.9821

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.99831    0.78529   1.271   0.2036
lwt         -0.01406    0.00617  -2.279   0.0227 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 228.69  on 187  degrees of freedom
AIC: 232.69

Number of Fisher Scoring iterations: 4
```

Testování

AIC (Akaike's Information Criterion, $-2l_m + 2k$, kde l_m je log. věrohodnost modelu) se používá jako míra kvality modelu. Např. funkce `step()` automaticky hledá model s nejmenší hodnotou AIC.

Testování jednotlivých parametrů

Pomocí asymptotické normality jednotlivých odhadů parametrů (`summary.glm()`).

Testování podmodelu

Test poměrem věrohodností (likelihood ratio test): Jsou-li l_1 a l_0 maxima logaritmické věrohodnosti za platnosti H_1 a podmodelu H_0, pak za platnosti H_0 má statistika $-2(l_0 - l_1)$ asymptoticky rozdělení χ^2_r, kde r je rozdíl v dimenzi (počtu parametrů).

Deviance: $-2(l_m - l_s)$, kde l_m je log. věrohodnost modelu a l_s je log. věrohodnost saturovaného modelu (model s n parametry, který perfektně vysvětluje pozorovaná data).

```
Analysis of Deviance Table

Model 1: low ~ 1
Model 2: low ~ lwt
  Resid. Df Resid. Dev Df Deviance
1       188     234.67
2       187     228.69  1   5.9813
```

Null deviance: deviance modelu pouze s abs. členem.
Residual deviance: deviance zkoumaného modelu.

Test významnosti všech parametrů: testová statistika pro test podmodelu je rozdíl null deviance a residual deviance, za platnosti H_0 má asymptoticky rozdělení χ^2_k.

LRT = Likelihood Ratio Test

```
> anova(lr0,lr1,test="LRT")

Analysis of Deviance Table

Model 1: low ~ 1
Model 2: low ~ lwt
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       188     234.67
2       187     228.69  1   5.9813  0.01446 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1
```


Týden 14

Téma:

- shluková analýza,
- matice vzdáleností,
- základní hierarchické metody.

Shluková analýza

Shluková analýza (cluster analysis) je sada nástrojů a metod pro hledání skupin (shluků, clusterů) v mnohorozměrných datech.

V praxi jde především o následující problémy:

1. *Volba vzdálenosti (nebo míry podobnosti).*
2. *Volba algoritmu pro vytváření skupin.*

Matice vzdáleností

$\mathcal{X}(n \times p)$ je datová matice s n pozorováními p-rozměrného náhodného vektoru.

Matice vzdáleností mezi jednotlivými pozorováními je matice $\mathcal{D}(n \times n)$, kde

$$\mathcal{D} = \begin{pmatrix} d_{11} & d_{12} & \dots & \dots & \dots & d_{1n} \\ \vdots & d_{22} & & & & \vdots \\ \vdots & \vdots & \ddots & & & \vdots \\ \vdots & \vdots & & \ddots & & \vdots \\ \vdots & \vdots & & & \ddots & \vdots \\ d_{n1} & d_{n2} & \dots & \dots & \dots & d_{nn} \end{pmatrix}$$

Příklad:
L_2-norma: $d_{ij} = \|x_i - x_j\|_2$, kde x_i a x_j označují řádky matice $\mathcal{X}$.

Příklad:
L_1-norma: $d_{ij} = \sum_k |x_{i,k} - x_{j,k}|$ (Manhattan distance).

R: `dist()`

Pro nominální (nebo binární) proměnné bývá jednodušší definovat matici podobností (similarity matrix) $\mathcal{S}$. Z podobnosti s_{ij} ale můžeme jednoduše vyrobit vzdálenost d_{ij} např. jako $d_{ij} = \max_{i,j}\{s_{ij}\} - s_{ij}$.

Vzdálenosti mezi řádky kontingenční tabulky může měřit testová statistika X^2 pro test nezávislosti v příslušné „podtabulce“ $2 \times J$.

V praxi: pozor na různá měřítka!!

Algoritmy pro vytváření shluků

Dva základní typy shlukovacích algoritmů:

- hierarchické algoritmy
  dále se dělí na aglomerativní (agglomerative) a dělící (splitting, divisive).
- nehierarchické algoritmy (partioning algorithms).

Hlavní rozdíl je v tom, že hierarchické algoritmy konstruují posloupnost shluků postupným slučováním nebo rozdělováním shluků (v n krocích takto projdou všechny možné počty shluků), zatímco nehierarchické algoritmy postupně (iterativně) upravují pevně daný počet shluků.

Hierarchické aglomerativní techniky

Aglomerativní algoritmy se v praxi často používají kvůli své názornosti a výpočetní jednoduchosti:

1. Nalezneme nejjemnější rozdělení pozorování (n jednobodových shluků).
2. Spočteme matici vzdáleností $\mathcal{D}$.

OPAKUJ:

3. Nalezneme dva nejbližší shluky.
4. Tyto dva shluky sloučíme do jednoho.
5. Přepočítáme vzdálenosti mezi novými skupinami a získáme novou matici vzdáleností (mezi shluky) $\mathcal{D}$.

DOKUD nemáme pouze jeden shluk obsahující všechna pozorování.

Při slučování shluků P a Q spočítáme vzdálenost mezi sloučeným shlukem $P+Q$ a shlukem R pomocí vzorce:

$$d(R,P+Q) = \delta_1 d(R,P) + \delta_2 d(R,Q) + \delta_3 d(P,Q) + \delta_4 |d(R,P) - d(R,Q)|,$$

kde δ_j jsou váhy.

Nechť $n_P = \sum_{i=1}^{n} I(x_i \in P)$ označuje počet objektů ve skupině P.

	δ_1	δ_2	δ_3	δ_4
Single linkage	$1/2$	$1/2$	0	-1/2
Complete linkage	$1/2$	$1/2$	0	$1/2$
Average linkage (unweighted)	$1/2$	$1/2$	0	0
Average linkage (weighted)	$\frac{n_P}{n_P+n_Q}$	$\frac{n_Q}{n_P+n_Q}$	0	0
Centroid	$\frac{n_P}{n_P+n_Q}$	$\frac{n_Q}{n_P+n_Q}$	$-\frac{n_P n_Q}{(n_P+n_Q)^2}$	0
Median	$1/2$	$1/2$	-1/4	0
Ward	$\frac{n_R+n_P}{n_R+n_P+n_Q}$	$\frac{n_R+n_Q}{n_R+n_P+n_Q}$	$-\frac{n_R}{n_R+n_P+n_Q}$	0

Tabulka: Výpočet vzdáleností mezi shluky.

Příklad: Body $x_1 = (0,0), x_2 = (1,0), x_3 = (5,5)$, čtvercová Euklidovská vzdálenost a single linkage.
Začneme s $N = 3$ shluky $P = \{x_1\}, Q = \{x_2\}, R = \{x_3\}$.
Vzdálenost (single linkage) mezi zbývajícími dvěma shluky:

$$\begin{aligned} d(R,P+Q) &= \frac{1}{2}d(R,P) + \frac{1}{2}d(R,Q) - \frac{1}{2}|d(R,P) - d(R,Q)| \\ &= \frac{1}{2}d_{13} + \frac{1}{2}d_{23} - \frac{1}{2}\cdot|d_{13} - d_{23}| \\ &= \frac{50}{2} + \frac{41}{2} - \frac{1}{2}\cdot|50 - 41| \\ &= 41 \end{aligned}$$

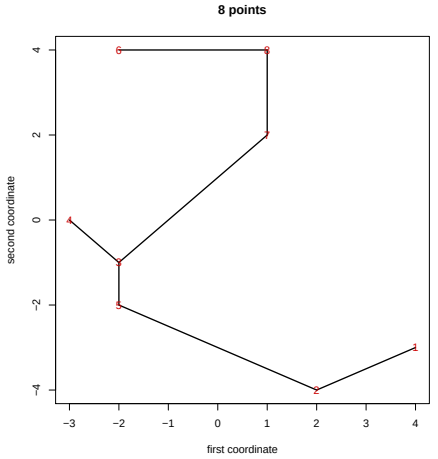
Nová matice vzdáleností je tedy $\left(\begin{pmatrix} 0 & 41 \\ 41 & 0 \end{pmatrix}\right)$.

Single linkage = nearest neighbor = nejbližší soused.

Dendrogram

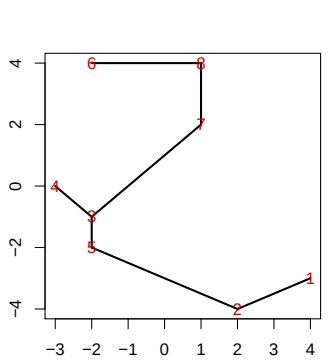
- grafické znázornění postupu při shlukování,
- obsahuje pozorování, posloupnost shluků a vzdálenosti mezi slučovanými shluky,
- na jedné ose jsou znázorněna pozorování,
- druhá osa určuje vzdálenosti mezi shluky.

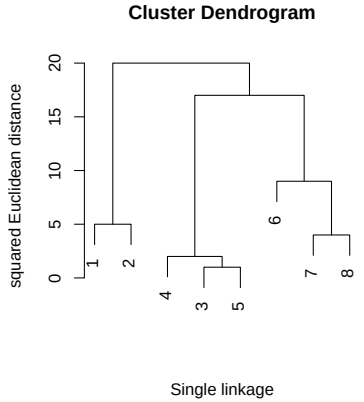
Příklad:



Matice vzdáleností (L_2) je

$$\mathcal{D} = \begin{pmatrix} 0 & 10 & 53 & 73 & 50 & 98 & 41 & 65 \\ & 0 & 25 & 41 & 20 & 80 & 37 & 65 \\ & & 0 & 2 & 1 & 25 & 18 & 34 \\ & & & 0 & 5 & 17 & 20 & 32 \\ & & & & 0 & 36 & 25 & 45 \\ & & & & & 0 & 13 & 9 \\ & & & & & & 0 & 4 \\ & & & & & & & 0 \end{pmatrix}$$





Uříznutím dendrogramu (`cutree()`) na úrovni 10 definujeme tři shluky: $\{1, 2\}$, $\{3, 4, 5\}$ and $\{6, 7, 8\}$.

Single linkage

definuje jako vzdálenost dvou shluků minimum individuálních vzdáleností.

$$d(R, P + Q) = \min\{d(R, P),\ d(R, Q)\}$$

Nazývá se také metoda *nejbližšího souseda*.

Tento algoritmus většinou vytváří velké a hodně roztáhlé shluky, které obvykle nevypadají příliš pěkně.

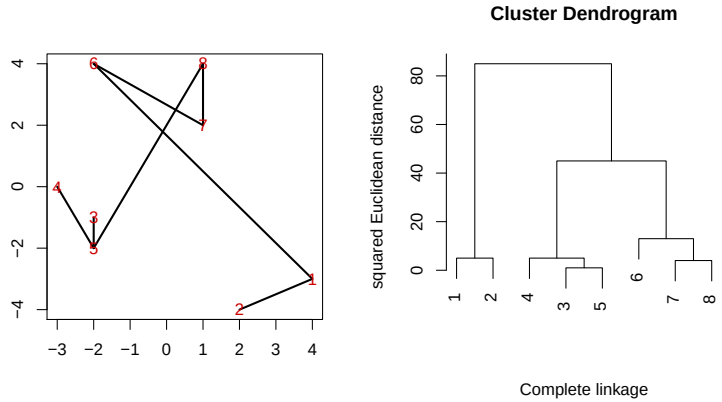
Complete linkage

definuje vzdálenost dvou shluků jako maximum individuálních vzdáleností.

$$d(R, P + Q) = \max\{d(R, P),\ d(R, Q)\}$$

Nazývá se také metoda *nejvzdálenějšího souseda*.

Tento algoritmus vytváří shluky podobných bodů, protože je založen na maximálních vzdálenostech.



Zde to vychází podobně jako single linkage, ale pro větší počet pozorování bývají shluky single a complete linkage hodně rozdílné.

Average & Centroid linkage

Další algoritmy, které počítají vzdálenost buď jako průměr vzdáleností nebo vzdálenost mezi středy shluků

Wardův algoritmus

- Wardův algoritmus měří vzdálenost pomocí toho, jak by se při sloučení dvou shluků zvýšila míra rozdílnosti pozorování uvnitř shluků (v podstatě: výběrový rozptyl vzdáleností od centra shluku).
- Cílem této metody je nalézt shluky, které budou působit velice homogenně.
- V praxi tento algoritmus většinou dává nejlépe vypadající shluky.

Algoritmus

Postup při shlukování:

1. volba vzdálenosti mezi pozorováními a výpočet matice vzdáleností (důležitá může být standardizace proměnných),
2. volba vzdálenosti mezi shluky (dobré výsledky dává Wardova metoda),
3. volba počtu shluků (pomocí dendrogramu),
4. popis výsledných shluků (grafy s označením shluků, tabulky průměrů).

R

Příklad:

```
data(iris)
sapply(iris[,1:4],sd)
iris.std=t(t(iris[,1:4])/sapply(iris[,1:4],sd)
apply(iris.std[,1:4],2,sd)
hc.iris.std = hclust(dist(iris.std[,1:4]),method="ward")
plot(hc.iris.std)
cl3=cutree(hc.iris.std,k=3)
table(cl3,iris[,5])
sapply(iris[,1:4],tapply,cl3,mean)
```


Zkouška

Písemná část:

- základní pojmy,
- odvození odhadu (momentová metoda, MLE) a jeho rozdělení,
- praktický příklad (t-testy, intervaly spolehlivosti, regrese).

Základní pojmy

Příklad: Mějme dvě nezávislé náhodné veličiny U_1 a U_2 s tzv. „trojúhelníkovým“ rozdělením na intervalu $(0, 3)$, tj. hustotu náhodných veličin U_1 a U_2 můžeme zapsat jako:

$$f(u) = \begin{cases} c \min(u, 3-u) & \text{pro } u \in (0,3), \\ 0 & \text{jinak.} \end{cases}$$

Určete hodnotu konstanty c tak, aby funkce $f(.)$ byla pravděpodobnostní hustota.

Příklad: Spočítejte střední hodnotu a rozptyl náhodných veličin U_1, U_2, $S = U_1 + U_2$ a $R = U_1 - U_2$.

Odvození odhadu (MLE, momentová metoda)

Příklad: Uvažujte náhodný výběr o rozsahu n z rozdělení s hustotou:

$$f(x) = \sqrt{\frac{\theta}{\pi}} \exp\left\{-\theta(x-1)^2\right\}.$$

Odvoďte odhad parametru θ ($\theta > 0$) metodou maximální věrohodnosti.

Příklad: Odvoďte asymptotické rozdělení odhadu parametru θ z předchozí úlohy. *Nápověda: využijte Fisherovu informaci* $\mathcal{F}_n = -E\partial^2 \ell(\theta)/\partial\theta^2$.

Praktický příklad (testování hypotéz, konfidenční intervaly, regresní analýza, logistická regrese, shlukování, . . .)

Příklad: Některé druhy včel potřebují k přežití vhodný úkryt. Při výzkumu včel druhu Osmia bylo na jaře nachystáno celkem 18 dřevěných desek, každá se šedesáti vyvrtanými otvory. Osm desek obsahovalo otvory o průměru 4 mm a deset desek otvory o průměru 6 mm. Na podzim bylo spočítáno, kolik otvorů bylo obsazeno včelami zkoumaného druhu:

4 mm	30	27	48	26	18	33	45	39		
6 mm	48	55	37	51	35	22	45	53	37	49

Zvolte vhodný test a rozhodněte, jestli obsazenost otvorů závisí na jejich průměru. Otestujte hypotézu, že průměrná obsazenost otvorů s průměrem 6 mm je přesně poloviční (v porovnání s obsazeností 4mm otvorů).