I. ZÁKLADNÍ POJMY

Častá rozdělení ve statistice: χ^2_n-rozdělení

- nepoužívá se jako model nějaké situace v praxi, nýbrž jako pomocné rozdělení při testování statistických hypotéz

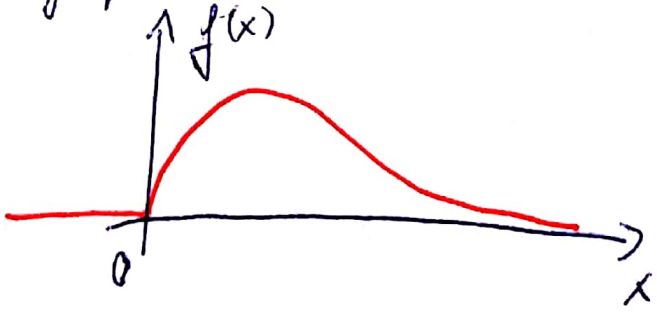
Nechť $X_1, X_2 \ldots, X_n$ jsou nezávislé stejně rozdělené náhodné veličiny s rozděleními $N(0,1)$. Pak náhodná veličina

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

má tzv. χ^2_n-rozdělení (čteno "chí-kvadrát rozdělení o n stupních volnosti").

- součet druhých mocnin ⇒ nemůže nabývat záporných hodnot

- podstatný pro nás bude tvar grafu hustoty



Skenováno pomo

Častá rozdělení ve statistice: Studentovo t-rozdělení

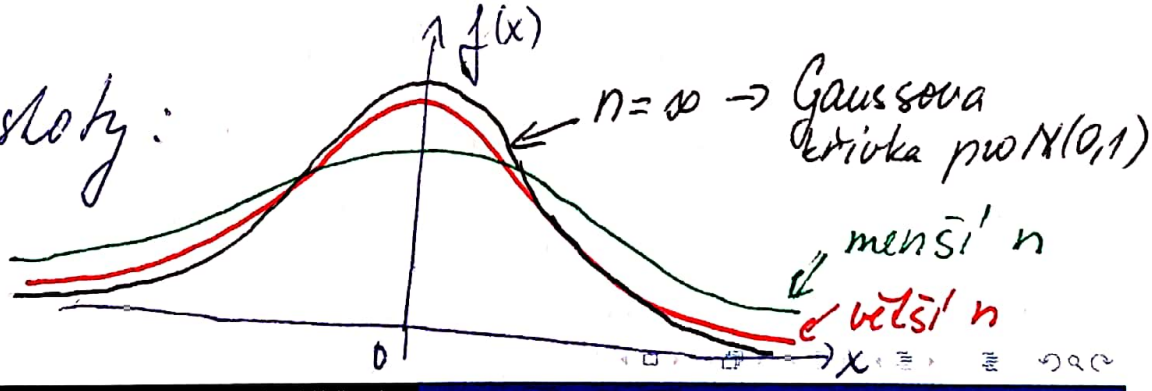
- rovněž se používá jen jako pomocné rozdělení, nikoliv jako model konkrétní situace

Nechť X je náhodná veličina s rozdělením $N(0,1)$ a Y je náhodná veličina s rozdělením χ^2_n. Pak náhodná veličina

$$Z = \frac{X}{\sqrt{Y}}\sqrt{n}$$

má tzv. t_n-rozdělení nazváno také Studentovo t-rozdělení o n stupních volnosti.

- tvar hustoty:



↳ je podobný Gaussově křivce pro rozdělení $N(0,1)$, ale pro všechna $n < \infty$ je „placatější" než $N(0,1)$ je až pro $n = \infty$, nicméně pro dostatečně velké n (> 20) jsou odchylky minimální

Skenováno pomo

Náhodný výběr

↳ je náhodným vektorem, neboť musíme brát v úvahu fakt, že při různém sběru n-tic hodnot je budeme dostávat různé, např. hojice hodu kostkou

Definice

Náhodnému vektoru $\mathbb{X} = (X_1, X_2 \ldots, X_n)^T$ *nezávislých stejně rozdělených náhodných veličin s distribuční funkcí* F_θ *závisející na parametru* θ *ve statistice říkáme náhodný výběr.* ~~*Číslu n říkáme rozsah výběru.*~~ *n-prvkový*

mohou být (1,2,3), (2,6,1), (1,2,2) atd.

V praxi, v níž provádíme statistickou analýzu, již za náh. výběr považujeme konkrétní n-tici čísel,

avšak při budování teorie musíme tuto n-tici uvažovat coby náhodný n-dimenzionální vektor. Konkrétní n-tici čísel nasbíraných v praxi říkáme realizace náhodného výběru, značíme $(x_1, \ldots, x_n)$
(malá písmena)

Skenováno pomo

(*) pozn.: pokud známe aritm. průměr $\bar{x}_n$ a $n-1$ hodnot realizace náh. výběru, můžeme n-tou hodnotu dopočítat $\Rightarrow$ není náhodné, např. $\bar{x} = 3$, $(x_1, x_2, x_3, x_4) = (1, 3, 3, ?) \Rightarrow$ $\Rightarrow$ víme, že $? = 5$

Výběrový průměr a výběrový rozptyl

Definice

Funkce náhodného výběru

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

jsou to náhodné veličiny, neboť X_i jsou náhodné!

se nazývá výběrový průměr a funkce

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

se nazývá výběrový rozptyl. $S_n = \sqrt{S_n^2}$ *je pak výběrová směrodatná odchylka.*

Po dosazení konkrétní realizace $(x_1, \dots, x_n)$ dostáváme aritmetický průměr (coby nejlepší odhad teoretické střední hodnoty), resp. realizaci výběrového rozptylu (coby nejlepší odhad teoretického rozptylu).

Souvislost $\mathbb{E}X \approx \bar{X}_n = \frac{1}{n}\cdot\sum_{i=1}^{n} X_i$ ← náhodné hodnoty

váhy pro

$$\text{var } X = \mathbb{E}(X - \mathbb{E}X)^2 = \frac{1}{n-1}\cdot\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

váhy pro $n-1$ náhodných hodnot (*)

Skenováno pomo

Výběrový průměr a výběrový rozptyl pro normální rozdělení

Norm. rozdělení má ve statistice speciální význam proto se pro něj odvozuje spousta vlastností zvlášť

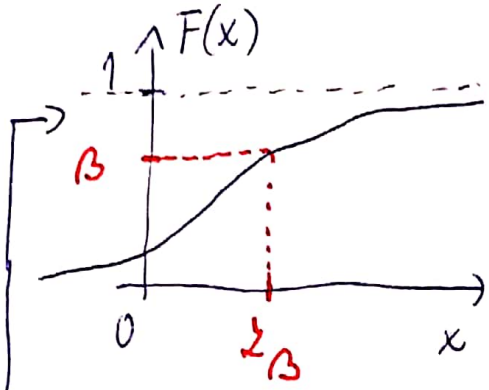
Věta

Nechť $\mathbb{X} = (X_1, X_2 \ldots, X_n)^T$ je náhodný výběr z rozdělení $N(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0$. Pak

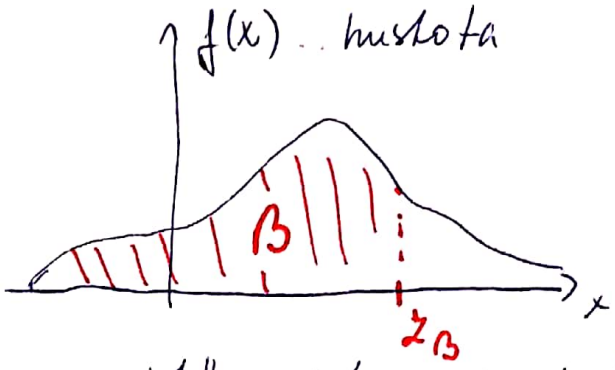
1. *výběrový průměr $\bar{X}_n$ a výběrový rozptyl S_n^2 jsou nezávislé náhodné veličiny,*
2. *rozdělení výběrového průměru $\bar{X}_n$ je $N(\mu, \sigma^2/n)$,*
3. *náhodná veličina $(n-1)S_n^2/\sigma^2$ má χ^2-rozdělení o $(n-1)$ stupních volnosti,*
4. *náhodná veličina $T = \frac{\bar{X}_n - \mu}{S_n}\sqrt{n}$ má t-rozdělení o $(n-1)$ stupních volnosti.*

tato jediná bude v přednášce využívána (je důsledkem předešlých tří vlastností a definice t_n-rozdělení), a to v sekci „testování hypotéz"!

Skenováno pomo



⇒



z definice

pro „lepší" představu: z_β je bod na ose x, který nechá plochu β pod sebou (tj. nalevo)

Kvantily

Definice

Nechť distribuční funkce F je spojitá, ryze monotonní a nechť $0 < \beta < 1$. Pak číslo z_β takové, že $F(z_\beta) = \beta$ se nazývá β-kvantil tohoto rozdělení.

Označení kvantilů používaných v této přednášce

Pozn.: indexy lze zaměnit, tj. psát $t_{n,\beta}$ a $\chi^2_{n,\beta}$

- u_β ... β-kvantil normovaného normálního rozdělení,
- $t_{\beta,n}$ ... β-kvantil Studentova $t-$rozdělení o n stupních volnosti,
- $\chi^2_{\beta,n}$ ... β-kvantil χ^2 rozdělení o n stupních volnosti.

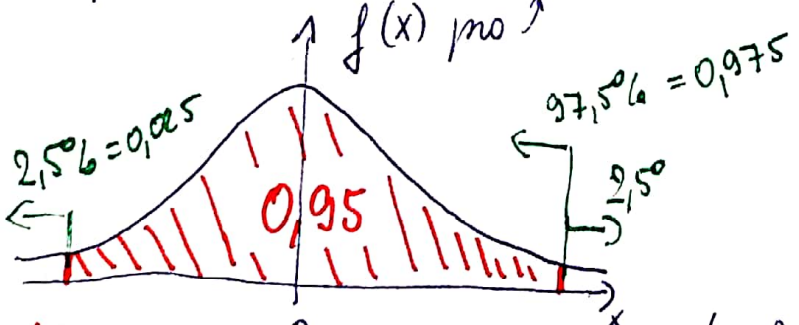
Poznámka

Je-li X náhodná veličina s distribuční funkcí F a kvantily z_β, pak

$$P(z_{\alpha/2} < X < z_{1-\alpha/2}) = F(z_{1-\alpha/2}) - F(z_{\alpha/2}) = 1 - \alpha.$$

využívá se pro sestrojení symetrického intervalu

o sudou pst, např. „prostředních 95% hodnot" pro rozdělení $N(0,1)$:



$u_{0,025}$ (⇐ 2,5% zůstává vlevo)

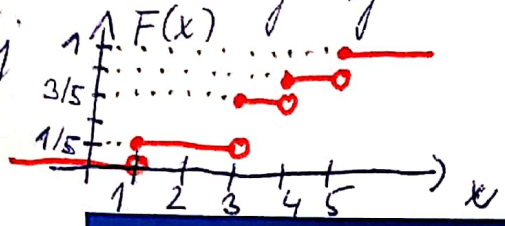
$u_{0,975}$ (⇐ 2,5% zůstává vpravo ⇒ 97,5% zůstává vlevo)

↑ odpovídá situaci, kdy $\alpha = 0,05 = 5\%$

Skenováno pomo

Připomenutí definice distr. fce F pro náh. vel. X:
$F(x) = P(X \leq x) \quad \forall x \in \mathbb{R}$

Pokud máme nasbírané hodnoty např. $(1,3,3,4,5)$ a náhodně je vybíráme, pak $P(X=1) = P(X=4) = P(X=5) = \frac{1}{5}$ a $P(X=3) = \frac{2}{5}$,

tj. 

, což přesně odpovídá

Empirická distribuční funkce a kvartily

Definice

Nechť $(x_1, x_2 \ldots, x_n)^T$ je realizace náhodného výběru $\mathbb{X} = (X_1, X_2 \ldots, X_n)^T$. Pak

$$F_{emp}(x) = \frac{\#\{x_i : x_i \leq x\}}{n},$$

odhadnuté z dat

kde # značí počet prvků, se nazývá empirická distribuční funkce.

Definice

Nechť $(x_1, x_2 \ldots, x_n)^T$ je realizace náhodného výběru $\mathbb{X} = (X_1, X_2 \ldots, X_n)^T$. Pak

- *$z = min(x_i : F_{emp}(x_i) \geq 1/4)$ se nazývá 1.kvartil,*
- *$z = min(x_i : F_{emp}(x_i) \geq 3/4)$ se nazývá 3.kvartil,*
- *$z = min(x_i : F_{emp}(x_i) \geq 1/2)$ se nazývá medián (2.kvartil),*
- *nejčastěji zastoupený prvek se nazývá modus.*

Pokus seřadíme hodnoty v datech (realizaci náh. výběru) od nejmenšího po největší, pak
1. kvartil = hodnota, které nechá 1/4 dat pod sebou a 3/4 nad sebou

3. kvartil = — " — — " — 3/4 — " — — " — 1/4

2. kvartil = medián = prostřední prvek

Skenováno pomo

Příklad

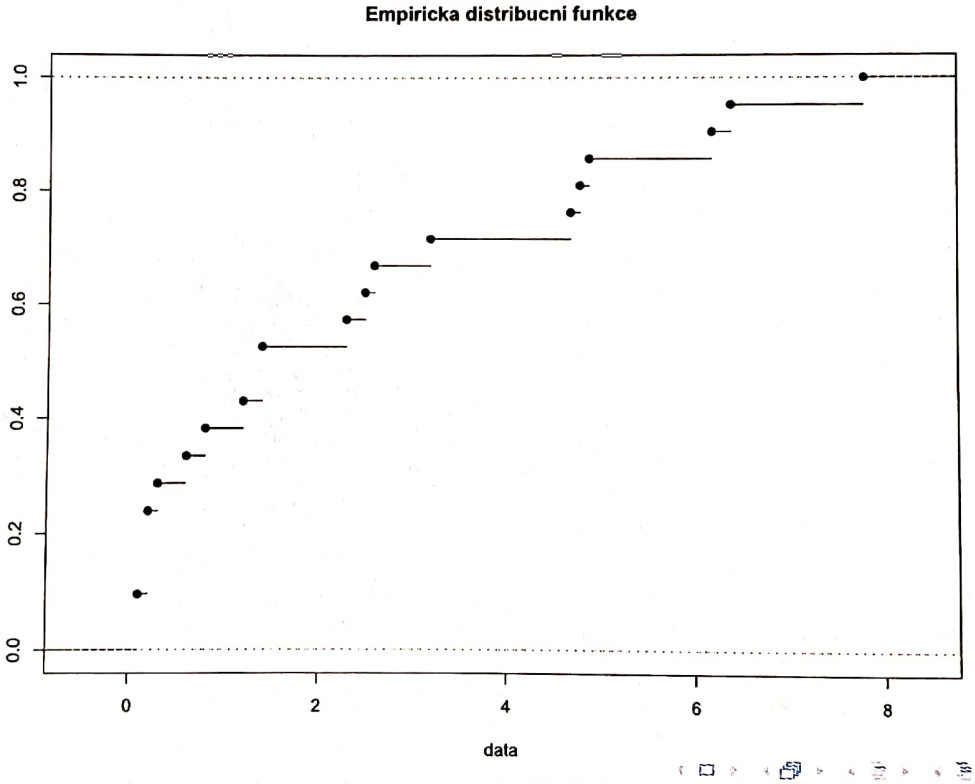
21× po sobě byly sledovány doby (v měsících) do poruchy daného přístroje. Naměřeny byly hodnoty:
4.9, 6.2, 2.6, 0.6, 0.3, 2.3, 3.2, 1.4, 6.4, 4.8, 1.2
2.5, 0.2, 0.2, 0.8, 0.1, 0.1, 1.4, 7.8, 0.2, 4.7.

modus

Uspořádáme-li pro lepší přehled tato data od nejmenší hodnoty k největší, dostaneme řadu hodnot:
1. kvartil
0.1, 0.1, 0.2, 0.2, 0.2, **0.3**, 0.6, 0.8, 1.2, 1.4, **1.4**, 2. kvartil = medián
2.3, 2.5, 2.6, 3.2, **4.7**, 4.8, 4.9, 6.2, 6.4, **7.8**.
3. kvartil maximum

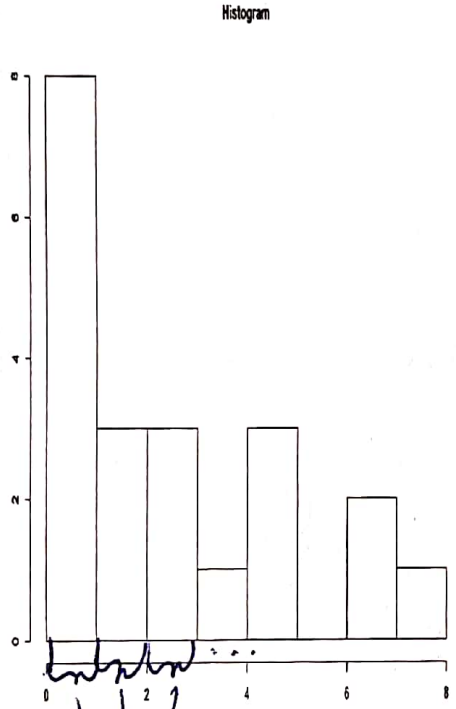
Z dat máme:
výběrový průměr $\bar{X}_{21} = 2.471$,
výběrový rozptyl $S^2_{21} = 5.81$,
výběrovou směrodatnou odchylku $S_{21} = \sqrt{5.81} = 2.21$,
1.kvartil = 0.3, medián (tj. 2.kvartil) = 1.4 a 3.kvartil = 4.7,
minimum = 0.1, maximum = 7.8, modus = 0.2.

Skenováno pomo

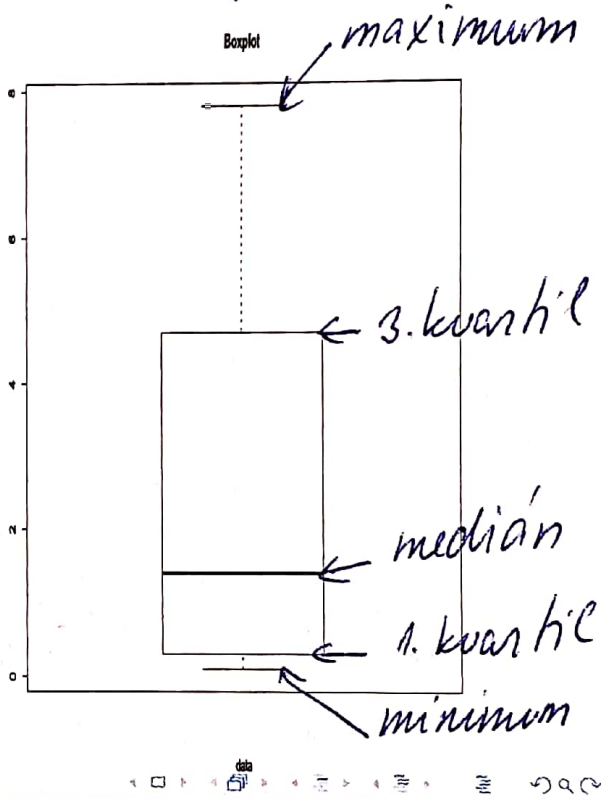


Skenováno pomo

Histogram



Krabicový graf = boxplot:



stejně dlouhé intervaly

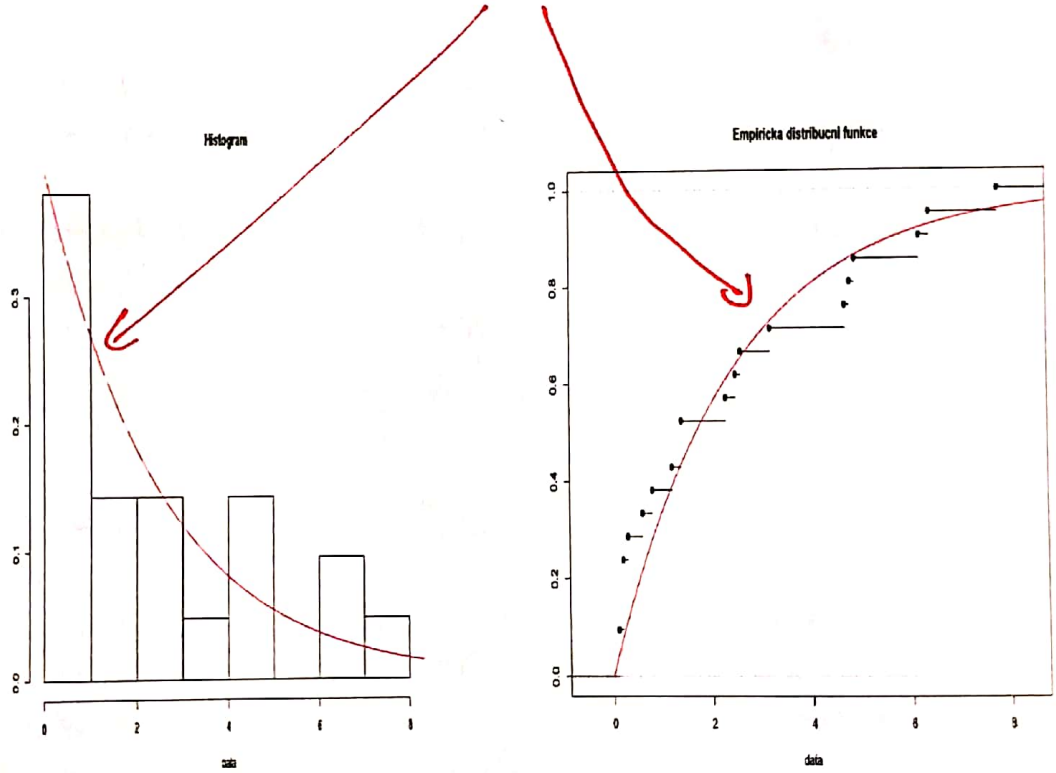
výška sloupce = počet hodnot v daném intervalu

Doporučený počet sloupců: byly existují kritéria, doporučují se počet sloupců takový, aby bylo rozdělení hodnot vidět „rozumně" (tj. 5-10 sl.)

Skenováno pomo

Empirická dist. fce je odhadem teoretické dist. fce.
Histogram (vynormovaný tak, aby součet ploch všech obdélníků byl roný 1) je odhadem hustoty.
Pokud si domníváme, že data pocházejí z nějakého rozdělení, lze emp. dist. fci, resp. histogram, porovnat s teoretickou dist. fcí, resp. hustotou.

Porovnání odhadů a skutečných funkcí 13



V příkladě: doby do poruchy ⇒ nejspíš Exp(λ) rozdělení,
tj. $F(x) = 0$ pro $x \leq 0$, $= 1 - e^{-\lambda x}$ pro $x \geq 0$
a $f(x) = 0$ pro $x \leq 0$, $= \lambda e^{-\lambda x}$ pro $x > 0$

Odhad λ: Pro $X \sim Exp(\lambda)$ máme $\mathbb{E}X = \frac{1}{\lambda} \Rightarrow \lambda = \frac{1}{\mathbb{E}X}$ ⇒
$\mathbb{E}X$ odhadneme jako $\bar{X} = 2{,}471$

stříška značí odhad → $\hat{\lambda}$

$\Rightarrow \hat{\lambda} = \frac{1}{2{,}471}$

Skenováno pomo

II. ODHADY PARAMETRŮ

Bodový odhad

Definice

Mějme náhodný výběr $\mathbb{X} = (X_1, X_2 \ldots, X_n)^T$, kde rozdělení (tj. distribuční funkce apod.) náhodné veličiny X_1 závisí na parametru θ. Bodový odhad parametru θ je jakákoliv funkce náhodného výběru $\theta^*(\mathbb{X})$, jejíž funkční předpis nezávisí na θ. Tento odhad se nazývá nestranný, jestliže $\mathrm{E}\theta^*(\mathbb{X}) = \theta$.

Poznámka

Výše zmíněná definice je sice matematicky přesná, avšak pro naše potřeby příliš obecná a nepřehledná. Budeme si proto pod bodovým odhadem představovat nikoliv funkci náhodného výběru $\theta^*(\mathbb{X})$, nýbrž samotné číslo $\hat{\theta}$, které získáme z realizace $(x_1, x_2 \ldots, x_n)^T$ náhodného výběru $\mathbb{X} = (X_1, X_2 \ldots, X_n)^T$ a které co nejpřesněji popisuje hledaný parametr θ.

Je dobra mít opět na paměti

skutečnost, že při různých realizacích náhodného výběru budou tyto odhady různé, tudíž ve své podstatě je odhad ~~také~~ náhodná veličina.

Skenováno pomo

Př.: V pěti posluchárnách po n studentech se píše test. Počty úspěšně složených testů jsou v jednotlivých posluchárnách: 28, 34, 30, 26, 32. Odhadněte (metodou momentů) počet studentů v jedné posluchárně, tj. n, a pst úspěchu u testu.

Metoda momentů 15

Mějme $(x_1, x_2 \dots, x_n)^T$ realizaci náh. výběru $\mathbb{X} = (X_1, X_2 \dots, X_n)^T$.

Rozdělení náhodné veličiny X_1 závisí na parametrech $\theta_1, \dots, \theta_k \in \Theta$, kde Θ je množina, z níž mohou parametry pocházet (např. nezáporná reálná čísla).

Předpoklad: $\mathrm{E}X_1^i < \infty \; \forall i = 1, \dots k$ a $\mathrm{E}X_1^i$ závisejí na $\theta_1, \dots, \theta_k$.

Postup: Položením

$$\mathrm{E}X_1^i = m_i,$$

kde m_i je $i-$tý výběrový moment získaný jako

$$m_i = \frac{1}{n} \sum_{j=1}^{n} x_j^i$$

pro všechna $i = 1, \dots k$, získáme soustavu k rovnic o k neznámých $\theta_1, \dots, \theta_k$, jejímž řešením jsou hledané odhady $\hat{\theta}_1, \dots, \hat{\theta}_k$.

Alternativa: Pokud $k = 2$, pak místo vztahů pro i-té momenty, $i = 1, 2$, můžeme uvažovat rovnice $\mathrm{E}X_1 = \bar{X}_n$ a $varX_1 = S_n^2$.

Nevýhoda: Tento odhad má velký rozptyl.

Řeš.: X ... počet úspěšných studentů v jedné posluchárně (tj. počet úspěchů z n pokusů)

$X \sim Binom(n, p) \Rightarrow \mathbb{E}X = np, \; varX = np(1-p)$

$\mathbb{E}X \approx \bar{X}_5 = \frac{1}{5}(28 + 34 + 30 + 26 + 32) = \frac{1}{5} \cdot 150 = \boxed{30}$

$varX \approx S_5^2 = \frac{1}{(4)} \left((28-30)^2 + (34-30)^2 + \dots + (32-30)^2 \right) = \frac{1}{4} \cdot 40 = 10$ $\Rightarrow$

↑ definice $\frac{1}{n-1}$

$\Rightarrow \begin{cases} np = 30 \\ np(1-p) = 10 \end{cases} \Rightarrow 1 - p = \frac{10}{30}$

$\hat{p} = \frac{2}{3}$

$np = n \cdot \frac{2}{3} = 30$ ← náš odhad

$\hat{n} = 45$

Skenováno pomo

→ zkráceně MMV nebo MLE (maximum likelihood estimate)

Idea: 1. Naměřená data $(x_1, \dots, x_n)$ jsou nejpravděpodobnější možná, tj. $P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n)$ je maximální

↑ 1. vybraná hodnota, ↑ 2. vybraná hodnota atd.

2. V náhodném výběru jsou $X_1, \dots, X_n$ nezávislé ⇒

$$P(X_1 = x_1, \dots, X_n = x_n) = P(X_1 = x_1) \cdot P(X_2 = x_2) \cdot \dots \cdot P(X_n = x_n)$$

Metoda maximální věrohodnosti

Předpokládejme, že máme $(x_1, x_2 \dots, x_n)^T$ realizaci náhodného výběru $\mathbb{X} = (X_1, X_2 \dots, X_n)^T$ z rozdělení s pravděpodobnostmi $P_\theta(X_1 = .)$ nebo s hustotou f_θ a nechť tyto pravděpodobnosti, resp. hustota, závisí na nějakém parametru $\theta \in \Theta$.

Definice

Odhad $\hat{\theta}$ je maximálně věrohodným odhadem, jestliže

$$\prod_{i=1}^{n} P_{\hat{\theta}}(X_1 = x_i) = \max_{\theta \in \Theta} \prod_{i=1}^{n} P_\theta(X_1 = x_i)),$$

resp.

$$\prod_{i=1}^{n} f_{\hat{\theta}}(x_i) = \max_{\theta \in \Theta} \prod_{i=1}^{n} f_\theta(x_i).$$

Poznámka

Většinou je však výhodnější pracovat s logaritmem součinu, neboť tak získáme součet logaritmů.

3. V náhodném výběru jsou $X_1, \dots, X_n$ stejně rozdělené (tj. $P(X_1 = k) = P(X_2 = k) = \dots = P(X_n = k)$ $\forall k \in \mathbb{R}$) ⇒

$$\Rightarrow P(X_1 = x_1) \cdot P(X_1 = x_2) \cdot \dots \cdot P(X_1 = x_n)$$

a to chceme maximalizovat.

Pozn.: Uvedený postup se týká diskrétních $X_1, \dots, X_n$, v spoj. případě $P(X_1 = \dots)$ nahrazujeme hustotou $f(x)$.

Skenováno pomo

Př.: Autosalon sledoval 10 týdnů po sobě počet prodaných vozů týdně a získal následující hodnoty: 2, 0, 2, 3, 5, 2, 3, 2, 2, 0
Určete, z jakého rozdělení pochází počet prodaných vozů za 1 týden, a odhadněte parametr tohoto rozdělení (metodou MLE).

Metoda maximální věrohodnosti pro výběr z diskrétního rozdělení 17

Jedná se o klasické hledání extrému funkce, tj. zderivujeme, položíme derivaci rovnu 0 atd.

1. Vyjádřit věrohodnostní funkci $L(\theta) = \prod_{i=1}^{n} P_\theta(X_1 = x_i)$.
2. Zlogaritmovat a tím získat logaritmicko-věrohodnostní funkci $l(\theta) = \sum_{i=1}^{n} \log P_\theta(X_1 = x_i))$.
3. Položit $\frac{\partial l(\theta)}{\partial \theta} = 0$.
4. Řešení rovnice $\frac{\partial l(\theta)}{\partial \theta} = 0$ je hledaný maximálně věrohodný odhad $\hat{\theta}$.
   → Díky vlastnostem pdf již nebudu znovu derivovat, výsledek je automatický řešením.

Mezikrok, který není nutný, ale může být užitečný, neboť díky zlogaritmování derivujeme součet místo součinu n funkcí. Logaritmus přitom může být libovolný se základem > 1, většinou se používá ln.

Kateřina Staňková Helisová

Řeš.: X ... počet prodaných vozů týdně (nezávislé události)
$X \sim Po(\lambda)$ → Parametr λ chceme odhadnout.
tj. $P(X=k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda}$ pro $k = 0, 1, \ldots$

1.) $L(\lambda) = P(X=2) \cdot P(X=0) \cdot \ldots \cdot P(X=0) = \frac{\lambda^2}{2!} e^{-\lambda} \cdot \frac{\lambda^0}{0!} e^{-\lambda} \cdot \ldots \cdot \frac{\lambda^0}{0!} e^{-\lambda} =$
$= \frac{\lambda^{21} \cdot e^{-10\lambda}}{(2!)^5 \cdot (0!)^2 \cdot (3!)^2 \cdot (5!)^1}$

2.) $l(\lambda) = \ln L(\lambda) = \ln \ldots = 21 \ln \lambda - 10\lambda \ln e + \ln(\text{jmenovatel})$ (konstanta, $\ln e = 1$)

3.) $l'(\lambda) = \frac{21}{\lambda} - 10 = 0 \Rightarrow$ 4.) $\hat{\lambda} = \frac{21}{10} = 2{,}1$

Skenováno pomo

Př.: Škody (počítané v mil. Kč) po povodni na pěti náhodně vybraných historických budovách byly: 2; 3,5; 4; 2,5 a 8. Předpokládejme, že škody se řídí rozdělením s hustotou $f(x) = \beta^2 x e^{-\beta x}$ pro $x > 0$ (jinak $f(x) = 0$). Metodou MLE odhadněte parametr β.

Metoda maximální věrohodnosti pro výběr ze spojitého rozdělení

Úplně stejný postup jako v případě diskrétního rozdělení, pouze $P(X_i = x_i)$ nahradíme hustotou $f(x_i)$

1. Vyjádřit věrohodnostní funkci $L(\theta) = \prod_{i=1}^{n} f_\theta(x_i)$.
2. Zlogaritmovat a tím získat logaritmicko-věrohodnostní funkci $l(\theta) = \sum_{i=1}^{n} \log f_\theta(x_i)$.
3. Položit $\frac{\partial l(\theta)}{\partial \theta} = 0$.
4. Řešení rovnice $\frac{\partial l(\theta)}{\partial \theta} = 0$ je hledaný maximálně věrohodný odhad $\hat{\theta}$.

Řeš.: 1.) $L(\beta) = f(2) \cdot f(3,5) \cdot \ldots \cdot f(8) = \beta^2 \cdot 2 \cdot e^{-2\beta} \cdot \ldots \cdot \beta^2 \cdot 8 \cdot e^{-8\beta} =$
$= \beta^{10} \cdot 560 \cdot e^{-20\beta}$

2.) $l(\beta) = \ln L(\beta) = 10 \ln \beta + \underbrace{\ln 560}_{\text{konst.}} - 20\beta \cdot \underbrace{\ln e}_{=1}$

3.) $l'(\beta) = \frac{10}{\beta} - 20 = 0$

4.) $\hat{\beta} = 0,5$

Skenováno pomo

Připomenutí: Odhad parametru s dat je hodnota, která bere roli parametru v teoretických pstních rozděleních. Jelikož data jsou realizací náhodného výběru, je i tento odhad ve své podstatě náhodný.
Obdobně můžeme z dat sestrojit (náhodný) interval, do kterého nám skutečný parametr (tj. ten teoretický = ten pro celou populaci) spadne, díky náhodnosti však pouze s nějakou zadanou pstí (tzv. spolehlivostí)

Definice

Mějme náhodný výběr $\mathbb{X} = (X_1, X_2 \ldots, X_n)^T$ *a číslo* $\alpha \in (0, 1)$.

1. *Dvojice* $(\theta_L^*(X_1, \ldots, X_n), \theta_U^*(X_1, \ldots, X_n))$ *se nazývá intervalový odhad parametru* θ *o spolehlivosti* $1 - \alpha$, *jestliže*

$$P(\theta_L^*(X_1, \ldots, X_n) < \theta < \theta_U^*(X_1, \ldots, X_n)) = 1 - \alpha.$$

2. $(\theta_D^*(X_1, \ldots, X_n))$ *se nazývá dolní odhad parametru* θ *o spolehlivosti* $1 - \alpha$, *jestliže*

$$P(\theta_D^*(X_1, \ldots, X_n) < \theta) = 1 - \alpha.$$

3. $(\theta_H^*(X_1, \ldots, X_n))$ *se nazývá horní odhad parametru* θ *o spolehlivosti* $1 - \alpha$, *jestliže*

$$P(\theta_H^*(X_1, \ldots, X_n) > \theta) = 1 - \alpha.$$

Interpulace $\theta^*(X_1, \ldots, X_n)$: „Způsob, jak z dat získat konkrétní hranici hledaného intervalu"

Skenováno pomo

Připomenutí: CLV: $P\left(\frac{\sum_{i=1}^{n} X_i - n \cdot \mathbb{E}X_1}{\sqrt{n \cdot \mathrm{var} X_1}} \leq a\right) \xrightarrow[n\to\infty]{} \Phi(a)$

říká, že pro velké n konverguje $\sum_{i=1}^{n} X_i$ k normálnímu rozdělení, ať už je rozdělení samotných X_i jakékoliv.

Stř. hodnota $\mathbb{E}X$ se odhaduje výběr. průměrem $\bar{X}_n = \frac{\sum_{i=1}^{n} X_i}{n}$

Intervalové odhady založené na CLV 22

→ asymptotické = čím větší n, tím přesnější

Tj. lze použít pro dostatečně velké n (několik desítek)
... což ostatně ve statistice všechno :)

Věta

Nechť $\mathbb{X} = (X_1, X_2 \ldots, X_n)^T$ je náhodný výběr z libovolného rozdělení, pro které $0 < \sigma^2 < \infty$. Pak intervalovým odhadem $\mu = \mathrm{E}X$ o asymptotické spolehlivosti $1 - \alpha$ je

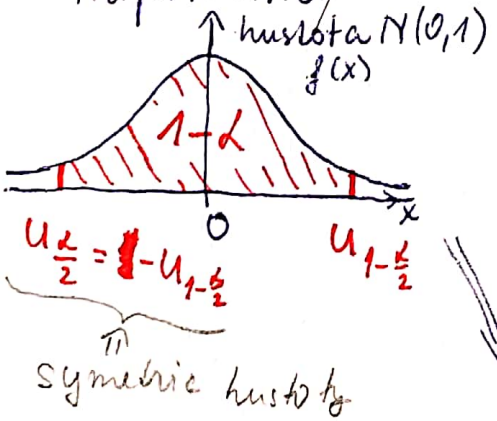
$$(\bar{X}_n - u_{1-\alpha/2}\frac{S_n}{\sqrt{n}}, \bar{X}_n + u_{1-\alpha/2}\frac{S_n}{\sqrt{n}}).$$

θ_L^* a θ_U^* z předešlého slidu

obě hranice jsou funkcemi $X_1, \ldots, X_n$, neboť $\bar{X}_n$ a S_n jsou získané z $X_1, \ldots, X_n$.

Vysvětlení:

Připomeňme:



- $\mathbb{E}X$ budeme hledat někde kolem průměru $\bar{X}_n$ → střed intervalu
- $S_n = \sqrt{S_n^2}$ ... směrodatná odchylka interval rozšiřuje, proložit větší rozptyl dat vnáší víc „nejistoty"
- n.. počet dat naopak interval zužuje, neboť víc hodnot = víc přesnosti
- $u_{1-\frac{\alpha}{2}}$ ... udává prostřední $(1-\alpha)$ pst

Skenováno pomo

odvození intervalu z předešlého slidu

Důkaz.

Víme, že pro velká n platí $\frac{S_n}{\sigma} \to 1$, tj. S_n se asymptoticky blíží hodnotě σ. Z CLV víme, že $\frac{\sum X_i - n\mu}{\sqrt{n\sigma^2}}$ má přibližně normální rozdělení. To znamená, že

$$P(u_{\frac{\alpha}{2}} \leq \frac{\sum X_i - n\mu}{\sqrt{n\sigma^2}} \leq u_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

$$P(u_{\frac{\alpha}{2}} \leq \frac{\sum X_i - n\mu}{\sqrt{n}S_n} \leq u_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

$$P(\frac{\sum X_i}{n} + u_{1-\frac{\alpha}{2}}\frac{S_n}{\sqrt{n}} \geq \mu \geq \frac{\sum X_i}{n} + u_{\frac{\alpha}{2}}\frac{S_n}{\sqrt{n}}) = 1 - \alpha$$

$$P(\bar{X}_n + u_{1-\frac{\alpha}{2}}\frac{S_n}{\sqrt{n}} \geq \mu \geq \bar{X}_n - u_{1-\frac{\alpha}{2}}\frac{S_n}{\sqrt{n}}) = 1 - \alpha,$$

což je definice intervalového odhadu o spolehlivosti $1 - \alpha$. □

Skenováno pomo

Ukázali jsme si odvození intervalového odhadu pro střední hodnotu, což samozřejmě není obecný návod, jak intervalový odhad libovolného parametru odvodit, nicméně jako ukázka nám to slouží. Navíc střední hodnota je u základních spojitých rozdělení s parametrem přímo spjatá (např. $X \sim Po(\lambda) \Rightarrow \mathbb{E}X = \lambda$, $X \sim Exp(\lambda) \Rightarrow \mathbb{E}X = \frac{1}{\lambda}$, $X \sim Alt(p) \Rightarrow \mathbb{E}X = p \ldots$)

Intervalové odhady založené na CLV 24

Z toho plynou odhady parametru v těchto případech:

Věta

1. *Nechť $\mathbb{X} = (X_1, X_2 \ldots, X_n)^T$ je náhodný výběr z alternativního rozdělení s parametrem $0 < p < 1$. Pak intervalovým odhadem p o asymptotické spolehlivosti $1-\alpha$ je* ($= \mathbb{E}X_1$)

$$(\bar{X}_n - u_{1-\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}, \bar{X}_n + u_{1-\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}).$$

2. *Nechť $\mathbb{X} = (X_1, X_2 \ldots, X_n)^T$ je náhodný výběr z Poissonova rozdělení s parametrem $0 < \lambda < \infty$. Pak intervalovým odhadem λ o asymptotické spolehlivosti $1-\alpha$ je* ($= \mathbb{E}X_1$)

$$(\bar{X}_n - u_{1-\alpha/2}\sqrt{\frac{\bar{X}_n}{n}}, \bar{X}_n + u_{1-\alpha/2}\sqrt{\frac{\bar{X}_n}{n}}).$$

Kromě toho, že se jedná o odhad střední hodnoty (které odpovídá přímo parametrům), je využito toho, že

1.) Pro $X \sim Alt(p)$ je $\mathbb{E}X = p$, $\text{var}\, X = p(1-p)$, takže $S_n^2 \approx \bar{X}_n(1-\bar{X}_n)$ (neboť $p = \mathbb{E}X \approx \bar{X}_n$)

2.) Pro $X \sim Po(\lambda)$ je $\mathbb{E}X = \text{var}\, X = \lambda$, takže $S_n^2 \approx \bar{X}_n$

Skenováno pomo

III. TESTOVÁNÍ HYPOTÉZ

Zjednodušeně: 1.) Naměříme data.

2.) Prohlédneme o nich nějaké tvrzení (tzv. nulová hypotéza H_0) a stanovíme si, co nám „vadí" (tzv. alternativní hypotéza H_A), např. (*)

Princip testování hypotéz

- Nechť $\mathbb{X} = (X_1, X_2 \ldots, X_n)^T$ je náhodný výběr z rozdělení, které závisí na parametru $\theta \in \Theta$.
- Chceme-li zjistit, jestli θ patří do nějaké užší podmnožiny parametrů Θ_0, nazveme tuto domněnku nulovou hypotézou, (označme $H_0 : \theta \in \Theta_0$).
- Na základě výběru $\mathbb{X} = (X_1, X_2 \ldots, X_n)^T$ testujeme tuto hypotézu proti alternativě $H_A : \theta \in \Theta \setminus \Theta_0$. To se zpravidla provede tak, že se určí množina W (tzv. kritický obor) taková, že pokud $\mathbb{X} \in W$, pak H_0 zamítáme.

Poznámka

Většinou testujeme $H_0 : \theta = \theta_0$, kde θ_0 je nějaká konkrétní hodnota. Přirozenou alternativou je pak $H_A : \theta \neq \theta_0$. Někdy je však smysluplnější testovat např. proti alternativě $H_A : \theta > \theta_0$. I když totiž teoreticky může být $\theta < \theta_0$, nemá smysl se tou situací zabývat, např. pokud při testování střední hodnoty $H_0 : \mu = 3$ z dat víme, že $\bar{X}_n = 5$, pak nemá smysl uvažovat situaci $\mu < 3$.

(*) H_0: hmotnost oříšků v balení je 200g (myšleno v průměru),
H_A: — " — — — je menší než 200g,

tj. bavíme se pouze o těchto možnostech, pokud dostaneme v balíčku oříšky navíc, nestěžujeme si :)

3.) Pomocí doposud probrané teorie si stanovíme, o jakém rozmezí se data mohou pochybovat, abychom akceptovali H_0, a v jakém rozmezí už H_0 zamítáme a přiklíníme se k H_A.

Skenováno pomo

Jelikož máme k dispozici data, která jsou pouze vzorkem nějaké populace, závěry testování budou platit jen s nějakou pravděpodobností, tj. vždy musíme počítat s tím, že se dopouštíme chyby. K té dochází v důsledku toho, že „ideálnímu chování" z hypotézy Ho tolerujeme jisté odchýlení. Pokud je tolerance velká, necháváme „projít" i „špatné" situace, naopak pokud je příliš malá, odmítneme i „správné" situace:

Chyba 1. a 2.druhu a testovací hladina

Mohou nastat následující situace:

- H_0 ve skutečnosti platí a test ji nezamítá. √
- H_0 ve skutečnosti neplatí a test ji zamítá. √
- H_0 ve skutečnosti platí, ale test ji zamítá → nastala chyba 1.druhu.
- H_0 ve skutečnosti neplatí, ale test ji nezamítá → nastala chyba 2.druhu.

Hladina testu:
Zvolíme si tzv. hladinu testu α (obvykle 0.05, někdy ale také 0.01 i méně) a W volíme tak, aby pravděpodobnost chyby 1.druhu nebyla větší než α (obvykle tak, aby byla rovna α).

→ určuje šířku „tolerantního pásma", a to pomocí α-kvantilu

Testování pak probíhá následovně:
1.) Stanovíme si H_0, H_A a hladinu α
2.) Z dat spočítáme číslo (tzv. testovou statistiku) o jehož pstním chování něco víme, většinou známe rozdělení
3.) Porovnáme spočtenou test. statistiku s kvantilem příslušného rozdělení a na základě toho H_0 zamítneme nebo nezamítneme (vždy vynášíme závěry o H_0 !)

Skenováno pomo

Př.: Přístav dostává od města dotace podle velikosti lodí, které v něm parkují. Správce přístavu uvedl do formuláře, že průměrná délka lodí je 12 m. Kontrola z úřadu náhodně vybrala 9 lodí a zjistila, že jejich délky jsou:
12,5 m; 11,2 m; 11,5 m; 12,5 m; 10,5 m; 12,1 m; 11,2 m; 10,5 m; 11,5 m.
Testujte, zda na základě těchto dat můžeme dát správci dotaci na lodě průměrné délky 12 m.

Řeš.: Nejprve si pro použití t-testu musíme spočítat

$$\bar{X} = \frac{1}{9}(12{,}5 + 11{,}2 + \ldots + 11{,}5) = 11{,}5 \; ; \; S^2 = \frac{1}{8}\left((12{,}5-11{,}5)^2 + \ldots (11{,}5-11{,}5)^2\right) = 0{,}5675$$

1.)

Testování střední hodnoty normálního rozdělení (t-testy): Jednovýběrový t-test

29

Nechť $(X_1, X_2 \ldots, X_n)^T$ je náhodný výběr z $N(\mu, \sigma^2)$, kde $\sigma^2 > 0$, a předpokládejme, že ani jeden parametr není znám. Testujeme-li $H_0 : \mu = \mu_0$ oproti $H_A : \mu \neq \mu_0$, je postup následující:

1. Vypočteme hodnotu $T_0 = \frac{\bar{X}_n - \mu_0}{S_n}\sqrt{n}$.
2. Pokud $|T_0| \geq t_{1-\alpha/2,n-1}$, zamítáme H_0. V opačném případě pak H_0 nezamítáme.

Jednostranný test $H_0 : \mu = \mu_0$ oproti $H_A : \mu > \mu_0$ pak probíhá podobně:

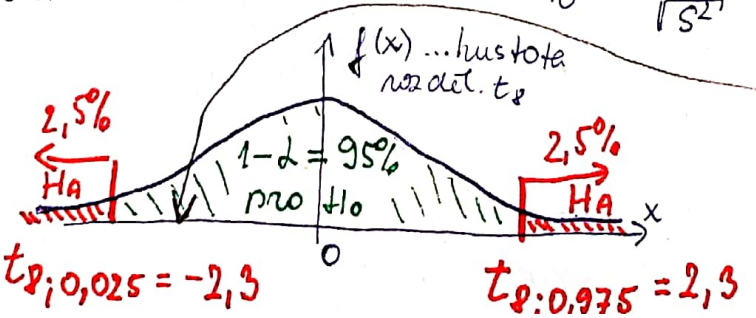
1. Vypočteme hodnotu $T_0 = \frac{\bar{X}_n - \mu_0}{S_n}\sqrt{n}$.
2. Pokud $T_0 \geq t_{1-\alpha,n-1}$, zamítáme H_0. V opačném případě pak H_0 nezamítáme.

Poznámka: $H_0 : \mu = \mu_0$ oproti $H_A : \mu < \mu_0$ pak zamítáme v případě, že $T_0 \leq t_{\alpha,n-1}$.

v „ideálním" případě by $\bar{X}_n = \mu_0$, takže čitatel = 0. Když dovolíme toleranci, bude čitatel kolísat kolem nuly a s menší pst bude od nuly dál. Tj. hodnoty daleko od nuly „nahrávají" alternativě H_A.

jelikož z 5. slidu víme, že $T \sim t_{n-1}$ (rozdělení)

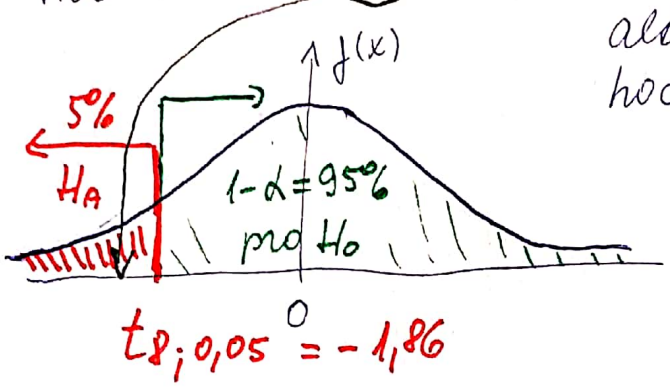
Přirozeně se nabízí testovat $H_0 : \mathbb{E}X = 12$
$H_A : \mathbb{E}X \neq 12$, kde X ... délka lodi [m]
zvolíme testovací hladinu $\alpha = 5\%$ (nejčastěji používaná hodnota) a počítáme test. statistiku $T_0 = \frac{\bar{X}-12}{\sqrt{S^2}} \cdot \sqrt{9} \doteq \frac{11{,}5-12}{0{,}75} \cdot 3 \doteq -2$ ← otázka: je to dost blízko 0?



jelikož $|-2| < t_{8;0,975}$ (tj. padla do „regionu pro H_0"), hypotézu H_0 nezamítáme při $H_A : \mathbb{E}X \neq 12$ na hladině $\alpha = 5\%$

Skenováno pomo

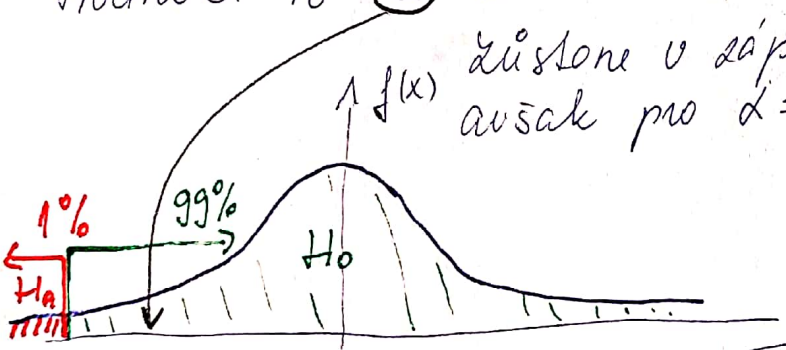
Nicméně předpokládá se, že správce přístavu bude mít tendenci délku lodí spíše nadhodnocovat, proto se nabízí testovat naději proti alternativní hypotéze $H_A: \mathbb{E}X < 12$.
Hodnota $T_0 = -2$ se nezmění, avšak $\alpha = 5\%$ připadající alternativě se přesune do záporných hodnot.



$t_{8;0,05} = -1,86$

Jelikož nyní $T_0 = -2 < -1,86 = t_{8;0,05}$ (tj. -2 spadá do „regionu proti H_0"), **hypotézu $H_0: \mathbb{E}X = 12$ zamítáme ve prospěch $H_A: \mathbb{E}X < 12$ na hladině $\alpha = 5\%$**

Správce namítá, že protitrend, že lodě jsou (průměrně) kratší než 12 m, je nemístě, nicméně může se bránit: „A proč testujete na hladině 5%? Dopouštíte se tak zamítnutí každého 20. požadavku, který je oprávněný! Testujte na hladině $\alpha = 1\%$!"
Hodnota $T_0 = -2$ se nám nezmění, „region proti H_0" také zůstane v záporných hodnotách, avšak pro $\alpha = 1\%$ bude menší.



$T_0 = -2 > -2,9 = t_{8;0,01}$

$t_{8;0,01} = -2,9$

Hypotézu $H_0: \mathbb{E}X = 12$ nezamítáme proti $H_A: \mathbb{E}X < 12$ na hladině $\alpha = 1\%$

ZÁVĚR:
Na základě těchto dat je těžko říct, zda dát za pravdu správci přístavu, ale zato můžeme dát za pravdu výroku (občas připisovanému Winstonu Churchillovi, byť asi mylně):
„Nevěřím žádné statistice, kterou jsem sám nezfalšoval." :)

Skenováno pomo

Př.: Při 600 hodech kostkou jsme získali hodnoty

1	2	3	4	5	6
110	80	100	90	100	120 krát.

Testujte, zda je kostka v pořádku (tj. zda jsou počty jednotlivých stran, které padly, přibližně stejné).

2) χ^2 test dobré shody 34

Používá se v situacích, kdy může nastat (řekněme v nějakém pokusu) k možných výsledků.

Chceme-li testovat H_0, že skutečné hodnoty dílčích pravděpodobností těchto výsledků jsou rovny předem zadaným číslům $p_1, \dots, p_k$, oproti alternativě, že aspoň jedna hodnota p_i je jiná, postupujeme následovně:

1. Vypočteme hodnotu $\chi^2 = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i}$, kde n počet opakování pokusu, X_i ... počty i-tého výsledku.
2. Pokud $\chi^2 > \chi^2_{1-\alpha,k-1}$, zamítáme H_0. V opačném případě pak H_0 nezamítáme.

Pozn.: Jelikož np_i je „ideální" počet i-tého výsledku (ideální ve smyslu, že pst tohoto výsledku je rovna p_i), je test založen na porovnání naměřeného počtu i-tého výsledku s ideálním.

Řeš.: $H_0: p_1 = p_2 = \dots = p_6 = \frac{1}{6}$, kde p_i = pst, že padlo i puntíků

H_A: aspoň jedno $p_i \neq \frac{1}{6}$

zvolme např. $\alpha = 5\%$

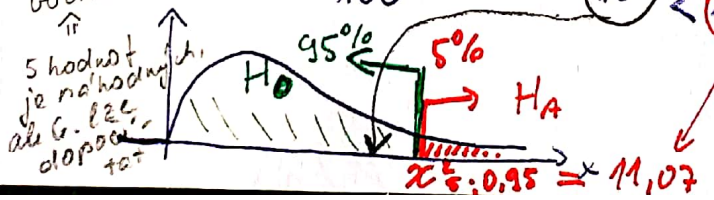
$$\chi^2 = \frac{(110 - 600\cdot\frac{1}{6})^2}{600\cdot\frac{1}{6}} + \frac{(80 - 600\cdot\frac{1}{6})^2}{600\cdot\frac{1}{6}} + \dots + \frac{(120 - 600\cdot\frac{1}{6})^2}{600\cdot\frac{1}{6}} =$$

$= 100$, což je ideální počet

$$= \frac{10^2 + 20^2 + \dots + 20^2}{100} = 10 < 11{,}07 = \chi^2_{5;0{,}95}$$

má χ^2-rozdělení s $(6-1)$ stupni volnosti

5 hodnot je náhodných, ale 6. lze dopočítat



H_0 nezamítáme na hladině 5%

Skenováno pom10

Př: Restaurace sledovala počty lahví klasického nezlívaného vína a vína na svařák během roku, a pozorovala následující:

	jaro	léto	podzim	zima	Σ
klasické víno	140	190	160	110	600
víno na svařák	60	10	140	190	400
Σ	200	200	300	300	1000

Je typ vína závislý na ročním období? Testujte.

Připomenutí: Náh. veličiny X a Y (diskrétní) jsou nezávislé, pokud $P(X=i, Y=j) = P(X=i) \cdot P(Y=j)$ $\forall i,j \in R$

$\frac{n_{ij}}{n}$ $\frac{n_{i\cdot}}{n}$ $\frac{n_{\cdot j}}{n}$

3.) Test nezávislosti v kontingenční tabulce 35

- Mějme výběr $(Y_1, Z_1), (Y_2, Z_2) \ldots, (Y_n, Z_n)$, kde Y_k může nabývat hodnot $1, \ldots, r$ a Z_k může nabývat hodnot $1, \ldots, c$.
- Testujeme H_0: "Y a Z jsou nezávislé" oproti H_A: "Y a Z nejsou nezávislé".
- Označíme-li n_{ij} počet výskytů dvojice $(Y_k = i, Z_k = j)$, pak matici o rozměru $r \times c$ s prvky n_{ij} říkáme kontingenční tabulka a prvkům n_{ij} říkáme sdružené četnosti. Marginální četnosti jsou pak dány vztahy

$$n_{i\cdot} = \sum_j n_{ij}, \quad n_{\cdot j} = \sum_i n_{ij}.$$

odhady psh' z tabulky

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

$n_{ij} - \frac{n_{i\cdot} \cdot n_{\cdot j}}{n} = 0$

$\left(n_{ij} - \frac{n_{i\cdot} \cdot n_{\cdot j}}{n}\right)^2 = 0$

Pozor: Vždy testujeme nezávislost, i když je v otázce závislost!

- Postup:
  1. Vypočteme hodnotu (tj. opět porovnáváme pozorovanou hodnotu s "ideální" hodnotou)

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(n_{ij} - \frac{n_{i\cdot} n_{\cdot j}}{n}\right)^2}{\frac{n_{i\cdot} n_{\cdot j}}{n}} \quad \Leftrightarrow \quad \sum_i \sum_j \left(n_{ij} - \frac{n_{i\cdot} \cdot n_{\cdot j}}{n}\right)^2 = 0$$

  2. Pokud $\chi^2 \geq \chi^2_{1-\alpha,(r-1)(c-1)}$, zamítáme H_0 ve prospěch H_A.

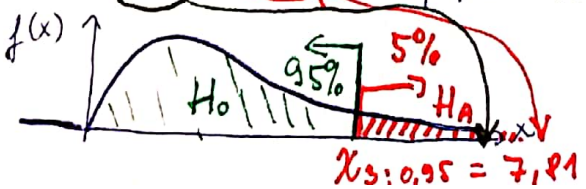
Řeš.: H_0: typ vína a roční období jsou nezávislé

H_A: — | | — nejsou —||— —

zvolme např. $\alpha = 5\%$

$$\chi^2 = \frac{\left(140 - \frac{600 \cdot 200}{1000}\right)^2}{\frac{600 \cdot 200}{1000}} + \frac{\left(190 - \frac{600 \cdot 200}{1000}\right)^2}{\frac{600 \cdot 200}{1000}} + \frac{\left(160 - \frac{600 \cdot 300}{1000}\right)^2}{\frac{600 \cdot 300}{1000}} + \ldots + \frac{\left(190 - \frac{400 \cdot 300}{1000}\right)^2}{\frac{400 \cdot 300}{1000}} =$$

$= (18{,}4) > (7{,}81) = \chi^2_{3;0,95} = \chi^2_{(2-1)(4-1);0,95}$



H_0: jsou nezávislé
$\Rightarrow$ zamítáme ve prospěch H_A: nejsou nezávislé na hladině $\alpha = 5\%$

Skenováno pomo