STATISTIKA V KOSTCE
DRUHÝ TEST

Pravděpodobnost

Nespojitá náhodná veličina:

Pravděpodobnostní funkce (P(x)) a Distribuční funkce (F(x)).

Spojitá náhodná veličina: (integrály)

Distribuční funkce (F(x)) a hustota pravděpodobnosti (f(x)).

Binomické rozdělení $Bi(n,\pi)$

$$P(x)=\binom{n}{x}\pi^x(1-\pi)^{n-x}$$

s vracením, hazardní hry (kostky, karty, ruleta)

Hypergeometrické rozdělení $Hg(M,N,n)$

$$P(x)=\frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

Bez vracení - kontrola, sázka.

dolní hranice, horní hranice

$$P(x_1<X\le x_2)=P\left(\frac{x_1-\mu}{\sigma}<\frac{X-\mu}{\sigma}\le\frac{x_2-\mu}{\sigma}\right)=P(u_1<U\le u_2)=\Phi(u_2)-\Phi(u_1)$$

Vzadání předpoklad normálního rozdělení

hledáme v tabulkách → Tabulka III. Distribuční funkce normálního rozdělení

Matematická statistika

a) σ^2 známý — kvantil, výběrová chyba

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

směrodatná chyba

Vzadání => Určete interval spolehlivosti (pr. 90%; 95%; 99%; ...)

známá odchylka (je zadaná; obecně daná)

b) σ^2 neznámý

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

neznámá odchylka (musí se spočítat; v zadání je že byla spočítaná.)

Studentovo rozdělení → Tabulka VI. **Kvantily rozdělení *t*** $T\sim t(n-1)$ - počet stupňů volnosti

$$P\left(P-u_{1-\alpha/2}\sqrt{\frac{P(1-P)}{n}}<\pi<P+u_{1-\alpha/2}\sqrt{\frac{P(1-P)}{n}}\right)=1-\alpha$$

→ zadáno relativně

normální rozdělení → Tabulka IV. **Kvantily normovaného normálního rozdělení (u_P)**

Testování statistických hypotéz

známá/neznámá - stejný princip jako u intervalů

σ^2 známý

$$U=\frac{\bar{X}-\mu_0}{\sigma}\sqrt{n}\qquad U\sim N(0,1)$$

σ^2 neznámý — neznáme rozptyl a $n\le 30$

$$T=\frac{\bar{X}-\mu_0}{S}\sqrt{n}\qquad T\sim t(n-1)$$

- Jestliže hodnota testové statistiky:
- **Patří (nerovnost platí)** do kritické oboru, potom na dané hladině významnosti **ZAMÍTÁME H0 a přijímáme H1.**
- **Nepatří (nerovnost neplatí)** do kritického oboru, potom na dané hladině významnosti **NEZAMÍTÁME H0. Výsledek je statisticky nevýznamný**. Pozor: Neznamená to, že zamítáme H1.

+ vždy uvádíme hladinu významnosti (na 5% hl. významnosti; na 1% hl. významnosti)

Testové kritérium

σ^2 neznámý ($n>30$)

$$U=\frac{\bar{X}-\mu_0}{S}\sqrt{n}\qquad U\approx N(0,1)$$

vždy rovnost (nezávislost)

H_0	H_1
$\mu=\mu_0$	$\mu>\mu_0$
	$\mu<\mu_0$
	$\mu\neq\mu_0$

dle zadání

Kritický obor

$W_\alpha=\{u;\ u\ge u_{1-\alpha}\}$

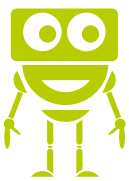
$W_\alpha=\{u;\ u\le -u_{1-\alpha}\}$

$W_\alpha=\{u;\ |u|\ge u_{1-\alpha/2}\}$

dle H1 vybereme kritický obor.

Testové kritérium — zadáno relativně (%)

$$U=\frac{P-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}\qquad U\approx N(0,1)$$



STATISTICKY NEKLASICKY

DOUČOVÁNÍ STATISTIKY S ADRIANOU

STATISTIKA V KOSTCE
DRUHÝ TEST

Rovnost středních hodnot dvou rozdělení (2 skupiny nebo 2 šetření) , vždy v Excelu.
velké nezávislé výběry

párový t-test: (stejná skupina 2. šetření)

Testové kritérium
σ_1^2 a σ_2^2 neznámé

$$U = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \qquad U \approx N(0,1)$$

– nezávislé výběry (ženy/muži; 1. skupina/2. skupina)
– ručně nebo Excel

EXCEL: Data - Analýza dat - Dvouvýběrový T-test s nerovností rozptylů

Testové kritérium

$$T = \frac{\sqrt{n}\bar{D}}{S_D} \qquad T \sim t(n-1)$$

závislé výběry (Před/Po; studenti psali 1. a 2. test;)

Chí-kvadrát test dobré shody

$$G = \sum_{j=1}^{k} \frac{(n_j - n\pi_{0,j})^2}{n\pi_{0,j}} \qquad G \approx \chi^2(k-1)$$

počet stupňů volnosti

DOMÁCÍ ÚKOL: Marketingový plán tvrdil, že záznam koncertu skupiny Kapička se prodá v poměru 70% CD, 20% DVD a 10% kazety. Za měsíc se skutečně prodá 2400 kusů CD, 1075 DVD a 384 kusů kazet tohoto koncertu. Ověřte, zda byl předpoklad marketingového plánu správný.

[70%; 20%; 10% => předpoklad
2400; 1075; 384 => konkrétní počet]

n = 2400 + 1075 + 384

• více % a více hodnot

Kvantily rozdělení χ^2 – vždy viz. 7

Analýza závislostí

Kontingenční tabulka (r x s) (počet řádků, počet sloupců)

=> více skupin a více hodnot:

$$G = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(n_{ij} - n_{i+}n_{+j}/n)^2}{n_{i+}n_{+j}/n} \qquad G \approx \chi^2((r-1)(s-1))$$

	A	B	C	D	E	Σ
muž	12	5	7	1	3	28
žena	3	8	4	2	1	18
Σ	15	13	11	3	4	46

$$G = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1+}n_{+1}n_{2+}n_{+2}}$$ – 2x2 tabulka

$$V = \sqrt{\frac{G}{n(m-1)}}, \; m = \min(r,s)$$ – Cramerův koeficient $\in \langle 0,1 \rangle$

$$C = \sqrt{\frac{G}{G+n}}$$ – Pearsonův koeficient (určení síly závislosti)

Analýza rozptylu

$$F = \frac{\frac{S_{y.m}}{k-1}}{\frac{S_{y.v}}{n-k}}$$ => Excel

dítě	1	5	3	8	
muž	12	5	2	1	3
žena	3	8	4	2	1

$$P^2 = \frac{S_{y.m}}{S_y}$$
(Poměr determinace) síla závislosti

více skupin, ale už se dál neřešení; jen kritérium skupin.

STATISTICKY NEKLASICKY

DOUČOVÁNÍ STATISTIKY S ADRIANOU

STATISTIKA V KOSTCE
DRUHÝ TEST

Regrese a korelace

regresní přímka $y = \beta_0 + \beta_1 x + \varepsilon,$

regresní parabola: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2$

vícenásobná regrese: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k$

Zabýváme se dvojicí veličin:
Y - vysvětlovaná, závisle proměnná – vždy 1!
X - vysvětlující, nezávisle proměnná – 1 nebo více

přímka nebo parabola → vícenásobná regrese

$R^2 = I^2 = \frac{S_T}{S_y}$ – koficient determinace $\in \langle 0,1 \rangle$

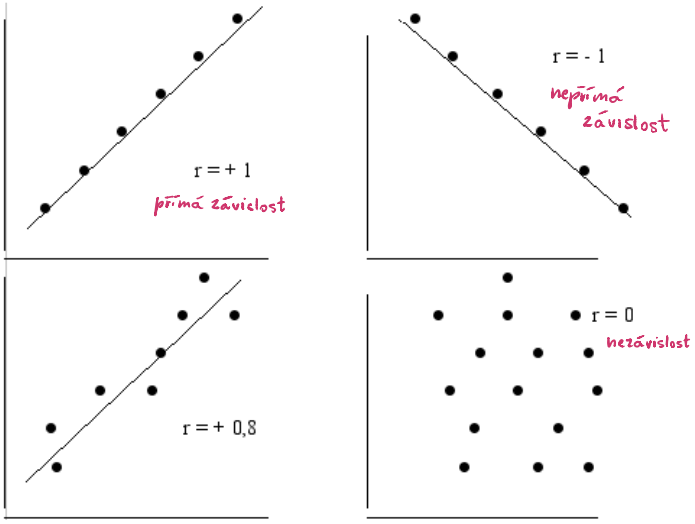
– používá se pro porovnání více modelů.
Je lepší přímka nebo parabola?

$I^2_{ADJ} = R^2_{ADJ} = 1-(1-I^2)\frac{n-1}{n-p}$ – upravený koficient determinace

korelační koeficient

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

$r_{yx} \in \langle -1, 1 \rangle$



Zdroj: absolventi.gymcheb.cz

Regresní statistika	
Násobné R	0,984743 – Korelační koeficient
Hodnota spolehlivosti R	0,969718 – Koeficient determinace
Nastavená hodnota spoleh	0,965933 – upravený koef. determinace
Chyba stř. hodnoty	58,59154
Pozorování	10

Test hypotézy o regresním parametru

		Koeficienty	ba stř. hodn	t Stat	Hodnota P	Dolní 95%	Horní 95%	Dolní 99,0%
Hranice	b_0	-160,347	41,00253	-3,91066	0,004477	-254,899	-65,7949	-297,926
Soubor X 1	b_1	7,573698	0,473188	16,00567	2,33E-07	6,482524	8,664873	5,985968

Excel: Data - Analýza Dat - Regrese (x - nezávislá, y-závislá)

významný

Test o modelu

ANOVA

	Rozdíl	SS	MS	F	významnost F
Regrese	1	879463,2	879463,2	256,1815	2,33E-07
Rezidua	8	27463,75	3432,969		
Celkem	9	906926,9			

STATISTICKY NEKLASICKY

DOUČOVÁNÍ STATISTIKY S ADRIANOU