STATISTIKA V KOSTCE

BODOVÉ A INTERVALOVÉ ODHADY

BODOVÉ ODHADY

- Bodový odhad = **jedno číslo**, které odhaduje neznámý populační parametr.
- Bodové odhady jsou jednoduché, ale neříkají nic o přesnosti -> proto existují **intervalové odhady**.

TYPICKÉ BODOVÉ ODHADY

- Populační **průměr μ ->** odhad je **x̄ (průměr)**
- Populační **rozptyl σ² ->** odhad je **S² (výběrový rozptyl)**
- Populační **podíl π ->** odhad je **p̂ = m/n (výběrový podíl)**
- Populační **úhrn N·π ->** odhad je **N·p̂ (výběrový úhrn)**

SMĚRODATNÁ/STANDARDNÍ CHYBA ODHADU

- Směrodatná chyba vyjadřuje přesnost bodového odhadu.
- Říká, jak moc se mohou odhadnuté hodnoty (např. průměr) lišit od skutečného populačního parametru, pokud opakuji výběr.

pro průměr: $\frac{S}{\sqrt{n}}$ **pro podíl:** $\sqrt{\frac{p(1-p)}{n}}$

VÝBĚROVÁ CHYBA

- Výběrová chyba je ta část intervalu, o kterou se **odhad liší** od skutečné hodnoty **při dané spolehlivosti**.
- Je to to, co se přičítá/odečítá k bodovému odhadu. Udává šířku intervalu spolehlivosti.
- Čím **větší spolehlivost -> větší výběrová chyba** čím **větší n -> menší výběrová chyba**.

kvantil: =T.INV(p; n-1)

výběrová chyba $t_{1-\alpha/2}(n-1)\frac{S}{\sqrt{n}}$

pro průměr, kde n ≤ 30

kvantil: =NORM.S.INV(p)

výběrová chyba $u_{1-\alpha/2}\frac{S}{\sqrt{n}}$

pro průměr, kde n > 30

výběrová chyba $u_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}$

pro podíl

INTERVALOVÉ ODHADY (INTERVAL SPOLEHLIVOSTI)

- Interval, ve kterém se očekává skutečná hodnota parametru s **danou spolehlivostí (1−α)**.

Nejčastěji:
95% interval -> α = 0,05
90% interval -> α = 0,10
99% interval -> α = 0,01

Interval má tvar: $\text{odhad} \pm \underbrace{(\text{kvantil } (u \text{ nebo } t)) \cdot (\text{standardní chyba})}_{\text{výběrová chyba}}$

STATISTIKA V KOSTCE

BODOVÉ A INTERVALOVÉ ODHADY

DVOUSTRANNÉ INTERVALY

- Odhaduje parametr z obou stran (dolní i horní mez).
- Používáme kvantil pro α/2. Rozděluje chybu na dvě strany, proto používá **kvantil pro α/2**.

$$\text{kritická hodnota} = u_{\alpha/2} \quad \text{nebo} \quad t_{\alpha/2}(n-1)$$

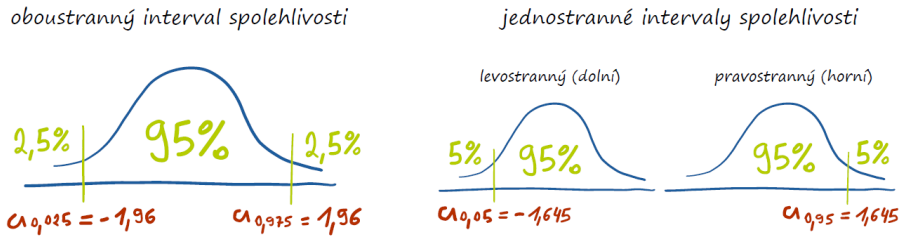
JEDNOSTRANNÉ INTERVALY

- Odhaduje jednu mez (jen horní nebo dolní).
- Má chybu jen na jedné straně, proto používá **kvantil pro α**.
- Jednostranný interval má menší kritickou hodnotu -> užší interval.

$$\text{kritická hodnota} = u_{\alpha} \quad \text{nebo} \quad t_{\alpha}(n-1)$$

POZOR – pokud chceme např. 95% oboustranný, tak použitý kvantil bude 97,5%.

grafické znázornění ->



KDY POUŽÍT T-ROZDĚLENÍ A KDY NORMOVANÉ NORMÁLNÍ ROZDĚLENÍ?

Malý výběr (n ≤ 30)

- používá se kvantil **Studentova t-rozdělení**

Velký výběr (n > 30)

- používáme kvantil **Norm. normálního rozdělení.**

VZTAHY, KTERÉ JE NUTNÉ UMĚT

- **Vyšší jistota** = musíme „pokrýt víc možných hodnot“ -> přesnost **klesá**.

$$\uparrow \text{spolehlivost } (1-\alpha) \implies \downarrow \text{přesnost}$$

- Abychom měli **větší jistotu**, že skutečný parametr „spadne dovnitř intervalu“, interval se musí **rozšířit**.

$$\uparrow \text{spolehlivost } (1-\alpha) \implies \uparrow \text{šířka intervalu}$$

- **Větší výběr** = menší standardní chyba = **užší interval** = **přesnější odhad**.

$$\uparrow \text{velikost výběru } n \implies \uparrow \text{přesnost}$$

TYPICKÉ ÚLOHY

Výrobní inženýr měřil dobu seřízení stroje u 15 náhodných cyklů. Získal průměr 4,8 min a směrodatnou odchylku 0,9 min. Odhadněte 95% interval spolehlivosti skutečné průměrné doby seřízení.
($\bar{x}$=4,8 s=0,9 n=15 alfa = 0,05)

Průzkum mezi 320 zákazníky ukázal, že 102 preferuje prodlouženou záruku. Sestavte 90% interval spolehlivosti pro skutečný podíl zákazníků, kteří by si připlatili za záruku. **(n=320 p=102/302 alfa=0,1)**

INTERPRETACE VÝSLEDKŮ

95% interval spolehlivosti pro průměr je od 3.1 do 7.8 cm.
Odhadujeme, že podíl voličů se s 95% spolehlivosti nachází v intervalu od 2 do 4 %.

STATISTIKA V KOSTCE
TESTOVÁNÍ HYPOTÉZ

TESTOVÁNÍ

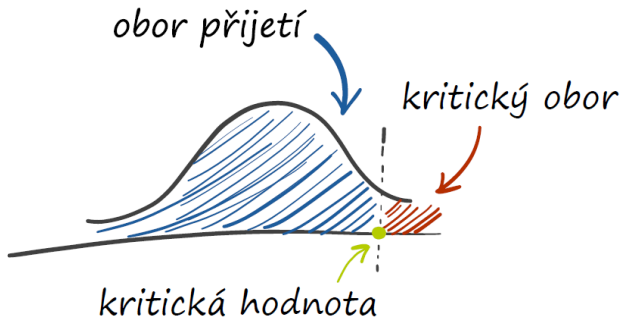
• Testování hypotéz je rozhodovací postup, který podle dat vyhodnocuje, jestli máme dost důkazů, abychom odmítli výchozí tvrzení o populaci. Vždy máme stanovenou H_0 a H_1.

• H_0 - vždy formulovaná jako **ROVNOST / SHODA / NULOVÝ ROZDÍL.**
• H_0 - je to, co chceme zpravidla vyvrátit.

• H_1 - popírá platnost nulové hypotézy, stojí v rozporu. Zpravidla značí rozdíl (diferenci), závislost mezi proměnnými.
• Alternativní hypotézy máme buď **jednostranné** (větší/menší než určité hodnota) nebo **oboustranné** (nerovnost určité hodnotě).

TESTOVÁ STATISTIKA –> POROVNÁNÍ S KRITICKÝM OBOREM

• Každý test má svou testovou statistiku (T, U, F, G…). Tu porovnáváme s kritickou hodnotou.
• Testové kritérium buď bude ležet v **oboru přijetí** nebo v **kritickém oboru.**



Pokud Testová statistika spadne (PRAVDA) do kritického oboru -> H_0 ZAMÍTÁME

p-hodnota ≤ α -> H_0 ZAMÍTÁME

Pokud Testová statistika NEspadne (NEPRAVDA) do kritického oboru -> H_0 NEZAMÍTÁME

p-hodnota > α -> H_0 NEZAMÍTÁME

P–HODNOTA

• Míra toho, jak moc data odporují H_0.
• Porovnává se s hladinou významnosti (alfa obvykle 1%, 5% nebo 10 %).

POUŽÍVANÉ TESTY

1 výběr -> test průměru / podílu (**pozor na velikost n**, stejné pravidlo jako u intervalů)
četnosti vs. očekávané hodnoty -> x^2 test dobré shody
3 a více skupin (porovnání průměrů) -> ANOVA
závislost dvou kategorických proměnných -> x^2 test nezávislosti. **POUŽIJTE DOPORUČENÉ SCHÉMA :)**

INTERPRETACE VÝSLEDKŮ

Když **H_0 zamítáme** - Testová statistika leží v kritickém oboru, proto zamítáme H_0 na hladině významnosti α. Máme dostatek důkazů pro alternativu H_1 (H_1 se prokázala). *Liší se, existuje rozdíl,...*

Když **H_0 nezamítáme** - Testová statistika neleží v kritickém oboru, proto H_0 nezamítáme. Nemáme dostatek důkazů k tvrzení H_1 (H_1 se neprokázala). *Neprokázalo se, že by existovat rozdíl, nelze prokázat,...*

NEPOUŽÍVAT: Zamítáme H_1, nezamítáme H_1, H_0 potvrzujeme/prokázalo se !!!

STATISTIKA V KOSTCE

KONTINGENČNÍ TABULKY

KONTINGENČNÍ TABULKA

Pokud je **zadaný dataset** (obr.1) - nutné převést do tabulky (obr.2). Případně je rovnou **zadaná tabulka** (obr.2), či **textové zadání**, ze kterého ji vytvoříš.

ZÁVISLOST 2 SLOVNÍCH PROMĚNNÝCH

Respondent	Máte absolvovaný kurz?	Jaké jsou Vaše zkušenosti?
1	Ano	Začátečník
2	Ano	Středně pokročilý
3	Ne	Středně pokročilý
4	Ne	Středně pokročilý
5	Ne	Pokročilý
6	Ano	Pokročilý
7	Ano	Začátečník
...	...	...

obr. 1

POSTUP PŘEVODU V EXCELU

1. **Označ celá data** (obsahující dané proměnné).
2. Klikni **VLOŽIT -> Kontingenční tabulka** vyber umístění (nový list) -> OK.
3. V pravém panelu nastav pole:
   Proměnná 1 -> **ŘÁDKY**
   Proměnná 2 -> **SLOUPCE**
   Proměnná 1 nebo Proměnná 2 -> **HODNOTY**
   (Excel automaticky počítá četnosti – „Počet“)

Zkušenosti/Abs.kurz	Ano	Ne	Suma
Začátečník	12	21	**33**
Středně pokročilý	27	26	**53**
Pokročilý	36	23	**59**
Suma	**75**	**70**	**145**

obr. 2

VÝPOČET V EXCELU

První tabulka - napozorované četnosti
Druhá tabulka - očekávané četnosti (každá musí být **větší než 5**!!) (suma řádku * suma sloupce / celkovým počtem).
Nezapomenout na fixace (dolary)).
Třetí tabulka - zbytek výpočtu, viz. vzorec:
((napozorovaná - očekávaná)^2/očekávaná) a **všechny sečíst**.

	A	B	C	D	E	F
1	Vzdělání	A	B	C	Celkem	
2	ZŠ	75	75	50	200	
3	SŠ	40	70	40	150	
4	VŠ	35	5	10	50	
5	Celkem	150	150	100	400	
6		200 * 150 / 400				
7		$E2*B$5/$E$5	$E2*C$5/$E$5	$E2*D$5/$E$5		
8		75	75	50		
9		56,25	56,25	37,5		
10		18,75	18,75	12,5		
11						
12		((B2-B8)^2)/B8	((C2-C8)^2)/C8	((D2-D8)^2)/D8		
13		0	0	0		
14		4,694444444	3,361111111	0,166666667		
15		14,08333333	10,08333333	0,5		
16						
17		SUMA(B13:D15)	32,88888889			
18						

$\frac{(40-37,5)^2}{37,5}$

$$G = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(n_{ij}-n'_{ij})^2}{n'_{ij}}$$

UKÁZOVÁ ZADÁNÍ

Existuje významná závislost mezi typem fakulty a pohlavím studentů? (2 slovní).
Závisí preference politické strány na vzdělání respondenta? (2 slovní)

HYPOTÉZY

H0: znaky jsou nezávislé
H1: znaky jsou závislé

SÍLA ZÁVISLOSTI

Je vhodnější používat **Cramerův koeficient kontingence.** Udává sílu vztahu mezi kategoriálními proměnnými: čím je blíže 1, tím je vztah silnější; hodnoty blízko 0 znamenají slabý vztah.

DÁVEJ POZOR NA ČETNOSTI

Když jsou očekávané četnosti menší než 5, je potřeba sloučit kategorie s nízkými četnostmi tak, aby vznikly větší a stabilnější skupiny.
Např. sloučit zcela souhlasím a spíše souhlasím do souhlasím.

STATISTIKA V KOSTCE

ANOVA (ANALÝZA ROZPTYLU)

HYPOTÉZY

H0: Rovnost středních hodnot
H1: Alespoň jedna střední hodnota se významně liší

VZTAH ČÍSELNÉ A KATEGORICKÉ PROMĚNNÉ (ALESPOŇ 3 KAT.)

VÝSTUP Z ANALÝZY DAT

$\frac{S_{y.m}}{k-1} = \frac{1726,067}{3}$

$p \leq \alpha \rightarrow H_0$ zamítáme
$p > \alpha \rightarrow H_0$ nezamítáme

ANOVA

Zdroj variability		*SS*	*Rozdíl*	*MS*	*F*	*Hodnota P*	*F krit*
Mezi výběry	$S_{y.m}$	1726,067	$k-1$ 3	575,3556	0,738697	0,550702	3,587434
Všechny výběry	$S_{y.v}$	8567,667	$n-k$ 11	778,8788			
Celkem	S_y	10293,73	$n-1$ 14				

$F < F_{KRIT} \rightarrow H_0$ nezamítáme
$F \geq F_{KRIT} \rightarrow H_0$ zamítáme

Excel: Data - Analýza Dat - Anova: jeden faktor

UKÁZOVÁ ZADÁNÍ

Rozhodněte, zda se liší střední hodnoty indexu vzdělanosti dle typu VŠ.
(číselná - index vzdělanosti, kategorická - typ VŠ).

Liší se dosažené skóre v jazykových testech (AJ, NJ, FJ, ŠJ)
(číselná - skóre, kategorická - jazykový test).

POMĚR DETERMINACE

Poměr determinace ukazuje, jak velkou část celkové variability závislé proměnné, lze připsat rozdílům mezi skupinami. Hodnota se pohybuje mezi **0 a 1**, přičemž vyšší hodnota znamená silnější efekt faktoru.

$$P^2 = \frac{S_{y,m}}{S_y}$$

STATISTIKA V KOSTCE

REGRESE A KORELACE

MODELY:

Regresní přímka $Y=b_0+b_1x$ jedna Y (závislá) a jedna X (nezávislá)

Regresní parabola $Y=b_0+b_1x+b_2x^2$ jedna Y (závislá) a jedna X (nezávislá), přidat sloupec X^2 (do analýzy dat dát oba sloupce X)

Vícenásobná regrese $Y=b_0+b_1x+b_2x_2+....+b_kx_k$ jedna Y (závislá) a více X (nezávislých) (do analýzy dat dát všechny sloupce X)

VÝSTUP REGRESNÍ PŘÍMKY Z ANALÝZY DAT

Regresní statistika	
Násobné R	0,984743
Hodnota spolehlivosti R	0,969718
Nastavená hodnota spoleh	0,965933
Chyba stř. hodnoty	58,59154
Pozorování	10

Korelační koeficient R <-1;1> udává sílu a směr lineární závislosti (znamenko dle b_1 u přímky)

Koeficient determinace R^2 <0;1> int. Modelem se podařilo vysvětlit 96,9 % var. závisle proměnné.

Upravený koeficient determinace R^2_{ADJ}, menší než R^2, vyšší hodnota -> lepší model.

F-test (porovnat s krit.hodnotou)

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			

HYPOTÉZY F-TEST

H0: model je statistický nevýznamný
H1: model je statisticky významný

p-hodnota, pokud je p ≤ alfa, H_0 zamítáme.
pokud je p > alfa, H0 nezamítáme.

		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

t-test (porovnat s krit.hodnotou, nebo porovnat p-hodnotu s alfa)

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

b1 - **Směrnice:** Udává o kolik se v průměru změní Y, pokud x vzroste o 1.

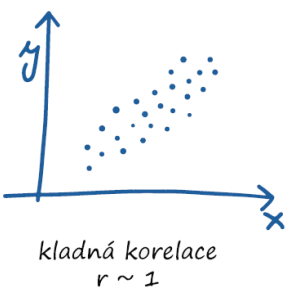
b0 - **Konstanta:** Hodnota Y pokud je x = 0.

HYPOTÉZY T-TEST REG. KOEFICIENTŮ

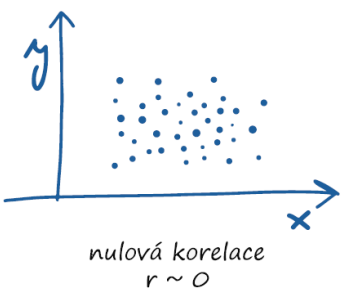
H0: regresní koeficient je nevýznamný
H1: regresní koeficient je významný

KORELACE

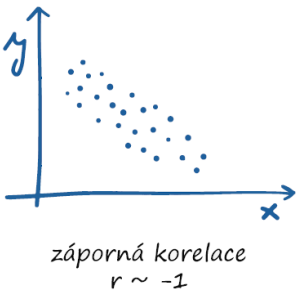
Pearsonova korelace - (viz. i výstup Násobné R z analýzy dat). Udává **sílu a směr lineární závislosti**. Vhodný pro 2 číselné proměnné. Lze udělat i test o korelačním koeficientu.



kladná korelace
r ~ 1



nulová korelace
r ~ 0



záporná korelace
r ~ -1

H_0	H_1
$\rho_{yx}=0$	$\rho_{yx}\neq 0$
Nevýznamný korelační koeficient	Významný korelační koeficient

Viz. vzorce

Speamanova korelace - Vhodná pro korelaci ordiálních (pořadových) proměnných. Leží rovněž v intervalu -1 až 1. **Udává sílu a směr monotónní závislosti** (nikoli lineární!).