

Princípy dátovej vedy - doplnky k téme 16

Samuel Rosa

Fakulta matematiky, fyziky a informatiky
Univerzity Komenského v Bratislave

PRESS RELEASE



From/ Od:	Martin Slosiarik
Kontakt:	info@focus-research.sk
Date/Dátum:	17.11.2024
Page/Strany:	1
Subject/Vec:	Volebné preferencie politických strán – november 2024

Agentúra **FOCUS pre portál 360tka.sk** uskutočnila v dňoch **6.11. – 13.11.2024** prieskum verejnej mienky formou osobných rozhovorov. Výberovú vzorku tvorilo **1018 respondentov**, ktorí reprezentujú populáciu SR vo veku nad 18 rokov z hľadiska pohlavia, veku, vzdelania, národnosti, veľkostných kategórií sídiel a krajského členenia.

Respondentom bola položená nasledujúca otázka: *"Predstavte si, prosím, že by sa parlamentné voľby na Slovensku konali na budúci víkend. Zúčastnili by ste sa ich? Ak áno, ktorej strane alebo hnutiu by ste dali svoj hlas?"*¹

21,4% respondentov by nešlo voliť a 17,9% odpovedalo "neviem". Ostatných 60,7%² rozhodnutých respondentov rozdelilo svoje preferencie nasledovne:

Politická strana	% rozhodnutých voličov	95% interval spoľahlivosti	% voličov
			november 2024 september 2024
Progresívne Slovensko	22,6%	19,3% - 25,9%	22,1%
SMER – SD	21,9%	18,6% - 25,2%	23,3%
HLAS – SD	13,3%	10,6% - 16,0%	12,8%
REPUBLIKA	8,6%	6,4% - 10,8%	7,9%
SaS	6,7%	4,7% - 8,7%	5,6%
KDH	6,2%	4,3% - 8,1%	6,0%
Maďarská aliancia – Magyar Szövetség	4,8%	3,1% - 6,5%	5,2%
Demokrati	4,6%	2,9% - 6,3%	5,0%
Slovensko	3,6%	2,1% - 5,1%	4,1%
SNS	3,6%	2,1% - 5,1%	4,7%

One-sample t-test [edit]

A **one-sample Student's t-test** is a [location test](#) of whether the mean of a population has a value specified in a [null hypothesis](#). In testing the null hypothesis that the population mean is equal to a specified value μ_0 , one uses the statistic

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}},$$

where $\bar{x}$ is the sample mean, s is the [sample standard deviation](#) and n is the sample size. The [degrees of freedom](#) used in this test are $n - 1$. Although the parent population does not need to be normally distributed, the distribution of the population of sample means $\bar{x}$ is assumed to be normal.

By the [central limit theorem](#), if the observations are independent and the second moment exists, then t will be approximately normal $\mathcal{N}(0, 1)$.

zdroj: wikipedia

O rozdelení

Kolmogorov-Smirnov

- dáta $X_1, \dots, X_n$
- H_0 : X_i majú rozdelenie R
vs. H_1 : nemajú
- napr. H_1 : $X_i \not\sim N(\mu, \sigma^2)$ pre zvolené μ a σ^2
- typicky pre $\mu = \bar{x}$ a $\sigma^2 = s^2$

O rozdelení

Kolmogorov-Smirnov

- dáta $X_1, \dots, X_n$
- H_0 : X_i majú rozdelenie R
vs. H_1 : nemajú
- napr. H_1 : $X_i \not\sim N(\mu, \sigma^2)$ pre zvolené μ a σ^2
- typicky pre $\mu = \bar{x}$ a $\sigma^2 = s^2$

- alebo dáta $X_1, \dots, X_n; Y_1, \dots, Y_m$
- H_0 : X_i majú rovnaké rozdelenie ako Y_j
vs. H_1 : nemajú

O rozdelení

Kolmogorov-Smirnov

- dáta $X_1, \dots, X_n$
- H_0 : X_i majú rozdelenie R
vs. H_1 : nemajú
- napr. H_1 : $X_i \not\sim N(\mu, \sigma^2)$ pre zvolené μ a σ^2
- typicky pre $\mu = \bar{x}$ a $\sigma^2 = s^2$

- alebo dáta $X_1, \dots, X_n; Y_1, \dots, Y_m$
- H_0 : X_i majú rovnaké rozdelenie ako Y_j
vs. H_1 : nemajú
- napr. H_1 : muži majú iné rozdelenie výšok ako ženy

O rozdelení

Kolmogorov-Smirnov

- dáta $X_1, \dots, X_n$
- H_0 : X_i majú rozdelenie R
vs. H_1 : nemajú
- napr. H_1 : $X_i \not\sim N(\mu, \sigma^2)$ pre zvolené μ a σ^2
- typicky pre $\mu = \bar{x}$ a $\sigma^2 = s^2$

- alebo dáta $X_1, \dots, X_n; Y_1, \dots, Y_m$
- H_0 : X_i majú rovnaké rozdelenie ako Y_j
vs. H_1 : nemajú
- napr. H_1 : muži majú iné rozdelenie výšok ako ženy
- existuje veľa iných testov o rozdelení

O korelácii/závislosti

Test korelačného koeficientu

- dáta $(X_1, Y_1), \dots, (X_n, Y_n)$
- $\hat{\rho}_{X,Y} = 0.44 \Rightarrow \rho_{X,Y} \neq 0?$

O korelácii/závislosti

Test korelačného koeficientu

- dáta $(X_1, Y_1), \dots, (X_n, Y_n)$
- $\hat{\rho}_{X,Y} = 0.44 \Rightarrow \rho_{X,Y} \neq 0?$
- $H_0 : \rho_{X,Y} = 0$ vs. vs. $H_1 : \rho_{X,Y} \neq 0$

O závislosti

Pre kategorické dáta:

- 380 ľudí
 - krajina (SK, CZ, PL, HU)
 - Má blond vlasy? (áno / nie)
- Závisí “blond’avosť” od krajiny (v rámci V4)?

O závislosti

Pre kategorické dáta:

- 380 ľudí
 - krajina (SK, CZ, PL, HU)
 - Má blond vlasy? (áno / nie)
- Závisí “blond’avosť” od krajiny (v rámci V4)?
- H_1 : to, či má človek blond vlasy, závisí od krajiny V4
- H_0 : “blond’avosť” nezávisí od krajiny V4

O závislosti

Pre kategorické dáta:

- 380 ľudí
 - krajina (SK, CZ, PL, HU)
 - Má blond vlasy? (áno / nie)
- Závisí “blond’avosť” od krajiny (v rámci V4)?
- H_1 : to, či má človek blond vlasy, závisí od krajiny V4
- H_0 : “blond’avosť” nezávisí od krajiny V4
- Kontingenčná tabuľka (pivot table): počty

blond \ krajina	SK	CZ	PL	HU
áno	41	46	55	37
nie	59	35	51	55

O závislosti

Pre kategorické dáta:

- 380 ľudí
 - krajina (SK, CZ, PL, HU)
 - Má blond vlasy? (áno / nie)
- Závisí “blond’avosť” od krajiny (v rámci V4)?
- H_1 : to, či má človek blond vlasy, závisí od krajiny V4
- H_0 : “blond’avosť” nezávisí od krajiny V4
- Kontingenčná tabuľka (pivot table): počty

blond \ krajina	SK	CZ	PL	HU
áno	41	46	55	37
nie	59	35	51	55

- pre menšie počty dát: Fisher exact test
- pre väčšie: Personov χ^2 test nezávislosti

Upozornenia 2

- testy: rôzne schopné odhaliť, že H_0 neplatí
- rôzne predpoklady: preskúmať, nakoľko sú porušené
- pre veľmi veľa dát je “všetko signifikantné”
 - ⇒ treba pozerieť, ako veľmi je H_0 porušená

Upozornenia 2

- testy: rôzne schopné odhaliť, že H_0 neplatí
- rôzne predpoklady: preskúmať, nakoľko sú porušené
- pre veľmi veľa dát je “všetko signifikantné”
 - ⇒ treba pozeráť, ako veľmi je H_0 porušená
- $p < 0.05 \Rightarrow H_0$ zamietame, výsledky sú signifikantné
- $p > 0.05 \Rightarrow H_0$ nezamietame, výsledky nie sú signifikantné
 - málo dát
 - slabý test
 - ozaj H_0 platí

NIE: H_0 sme potvrdili

Upozornenia 2

- testy: rôzne schopné odhaliť, že H_0 neplatí
 - rôzne predpoklady: preskúmať, nakoľko sú porušené
 - pre veľmi veľa dát je “všetko signifikantné”
 - ⇒ treba pozeráť, ako veľmi je H_0 porušená
 - $p < 0.05 \Rightarrow H_0$ zamietame, výsledky sú signifikantné
 - $p > 0.05 \Rightarrow H_0$ nezamietame, výsledky nie sú signifikantné
 - málo dát
 - slabý test
 - ozaj H_0 platí
- NIE: H_0 sme potvrdili
- pozor na viacnásobné testovanie