



Analýza dát

Prednáška 6

Základy testovania štatistických
hypotéz

Nulová a alternatívna hypotéza

Predstavujú dve protichodné tvrdenia, ktoré „súťažia“ medzi sebou o platnosť.

Nulová hypotéza H_0 reprezentuje „skeptický“ uhol pohľadu alebo tvrdenie, ktoré treba otestovať

Alternatívna hypotéza $H_1(H_A)$ reprezentuje alternatívne (opačné) tvrdenie a často obsahuje interval možných hodnôt parametra

Testovanie hypotéz pripomína americké súdnictvo...

Situácia: bola vykradnutá banka a „zmizlo“ 10 000 000 €. Vyšetrovatelia na základe informácií od obyvateľstva zadržali podozrivú osobu A.B. Vyšetrovatelia zbierali dôkazy, aby mohli osobu A.B. obžalovať a nakoniec odsúdiť. Ale: funguje prezumpcia nevinny, preto:

H_0 = obžalovaný A.B. je nevinný

H_1 = obžalovaný je A.B. vinný

Testovanie hypotéz pripomína americké súdnictvo...

Vyšetrovateľ/detektív sa snaží získať čo najviac dôkazov na to, aby mohol prehlásiť, že dôkazy sú také silné, že obžalovaný nemôže byť nevinný a teda je vinný. Málokedy môžeme obžalovaného prehlásiť za nevinného, iba sme nenazberali dostatok dôkazov na to, aby sme to mohli prehlásiť za vinného.

Príklad 1

Hans Rosling vypracoval 12 otázkový Gapminder test (s možnosťami A, B, C, kde práve jedna je správna) a zisťoval ako obyvatelia jednotlivých krajín vedia odpovedať na otázky ohľadom svetového zdravia. Keby obyvatelia nemali žiadne vedomosti o svetovom zdraví a odpovedali by náhodne (ako napr. šimpanzi), aký priemerný počet bodov (správnych odpovedí) by z Gapminder testu získali?

Príklad 1 - pokračovanie

Ako by mohla znieť nulová a alternatívna hypotéza?

H_0 = Američania nie sú zaťažení žiadnymi vedomosťami ani predsudkami a vyplňajú test náhodne (teda ako šimpanzi)

$$H_0: \mu = 4$$

H_1 = Američania nevypĺňajú test náhodne

$$H_1: \mu \neq 4$$

Príklad 1 - pokračovanie

Hans náhodne vybral 1200 Američanov, dal im robiť test a zistil, že priemer správnych odpovedí vo vzorke bol $\bar{x} = 3,6$ a smerodajná odchýlka $s = 4,52$.

Overme, či údaje získané zo vzorky 1200 Američanov predstavujú dostatočne silný dôkaz na to, aby sme mohli zamietnuť nulovú hypotézu (alebo, či je rozdiel medzi hodnotami 3,6 a 4 signifikantný).

Skúsme použiť 95% interval spoľahlivosti.

Čo ak by sme použili 99% IS?

Príklad 2

Hans zisťoval ako dopadli obyvatelia USA podľa otázok. V otázke č. 6 (OSN predpokladá, že do roku 2100 sa svetová populácia zvýši o 4 miliardy ľudí. Čo je hlavným dôvodom?) zisťoval percento Američanov, ktorí vedeli správne odpovedať.

Keby Američania odpovedali náhodne, aké percento správne odpovedajúcich by sme očakávali?

Príklad 2 - pokračovanie

Ako by mohla znieť nulová a alternatívna hypotéza?

H_0 = Američania nie sú zaťažení žiadnymi vedomosťami ani predsudkami a vyplňajú test náhodne (teda ako šimpanzi)

$$H_0: p = 33\% = 0,33$$

H_1 = Američania nevypĺňajú test náhodne

$$H_1: p \neq 33\% (0,33)$$

Príklad 2 - pokračovanie

Hans náhodne vybral 1200 Američanov, dal im robiť test a zistil, že Američanov správne odpovedajúcich na otázku č. 6 bolo 420.

Overme, či údaje získané zo vzorky 1200 Američanov predstavujú dostatočne silný dôkaz na to, aby sme mohli zamietnuť nulovú hypotézu. Skúsme použiť 95 % interval spoľahlivosti.

Čo ak by sme použili 99% IS?

Testovanie hypotéz pomocou intervalu spoľahlivosti

1. Formulácia nulovej a alternatívnej hypotézy

H_0 : parameter = nulovej hodnote

H_1 : parameter $\neq$ nulovej hodnote

2. Aký IS budeme používať?

3. Overenie CTL (pozor, rôzne pre priemer a relatívnu početnosť)

4. Výpočet IS pomocou SE (pozor, rôzne pre priemer a relatívnu početnosť)

5. Rozhodnutie:

- Ak nulová hodnota parametra **patrí** intervalu spoľahlivosti, tak nemôžem zamietnuť nulovú hypotézu (nenazbierali sme dostatok dôkazov)
- Ak nulová hodnota parametra **nepatrí** do intervalu spoľahlivosti, tak zamietnem nulovú hypotézu v prospech alternatívnej.

Testovanie pomocou intervalov
spoľahlivosti je fajn, ale má háčiky...

Treba vymyslieť niečo iné

Proces testovania hypotéz

Vráťme sa k obžalovanému A.B.:

H_0 = obžalovaný A.B. je nevinný

H_1 = obžalovaný A.B. je vinný

Zbierame dôkazy (dáta) a usúdime, či sú tieto dôkazy dostatočne presvedčivé, aby sme mohli povedať, že obžalovaný nemôže byť nevinný a vtedy zamietneme nulovú hypotézu v prospech alternatívnej a obžalovaného vyhlásime za vinného.

Ak dôkazy nie sú dostatočne presvedčivé, musíme usúdiť, že nulovú hypotézu nie je možné zamietnuť. Pozor: netvrdíme, že obžalovaný je nevinný, iba nemáme dostatok dôkazov na to, aby sme ho prehlásili za vinného.

Chyby pri testovaní hypotéz

- V procese testovania hypotéz/v procese s obžalovaným A.B. môžu nastať 4 možnosti:

		Výsledok testu/Rozhodnutie súdu	
		H_0 nemôžeme zamietnuť/ obžalovaného neodsúdime	H_0 zamietame v prospech H_1 / obžalovaného odsúdime
Skutočnosť	H_0 je pravdivá obžalovaný je nevinný	OK	Chyba I. druhu
	H_1 je pravdivá obžalovaný je vinný	Chyba II. druhu	OK

Pokračovanie s US súdnictvom

- Čo predstavuje chyba I. druhu pre obžalovaného A.B.?
- Čo predstavuje chyba II. druhu pre obžalovaného A.B.?
- Ktorej chyby sme ochotní sa dopustiť? Ako veľmi? Aký je vzťah medzi oboma chybami?

Hladina významnosti α

- Predpokladajme, že je nulová hypotéza v skutočnosti pravdivá a my ju zamietneme (dopustili sme sa chyby I.druhu), ale neradi by sme túto chybu urobili vo viac ako 5% prípadov.
- Táto hodnota 5%=0,05 sa nazýva **hladina významnosti** a označuje sa α .
- Ak skonštruujeme 95%-ný interval, zodpovedá to 5% hladine významnosti.

P-hodnota

- nám umožňuje kvantifikovať silu dôkazov proti nulovej hypotéze;
- je pravdepodobnosť, že vzťah zistený z našich údajov je iba dôsledkom nešťastnej vzorky a ak by sme vybrali ďalšiu náhodnú vzorku, nemuseli by sme nájsť nič. P -hodnota je najnižšia hodnota hladiny významnosti, ktorá vedie k zamietnutiu nulovej hypotézy.
- P -hodnota je odhadovaná pravdepodobnosť zamietnutia pravdivej nulovej hypotézy. Čím menšia je P , tým viac sme presvedčení, že nulová hypotéza nie je pravdivá a mala by byť zamietnutá.

Príklad 1 – ešte raz

Hans náhodne vybral 1200 Američanov, dal im robiť test a zistil, že priemer správnych odpovedí vo vzorke bol $\bar{x} = 3,6$ a smerodajná odchýlka $s = 4,52$.

Overme, či údaje získané zo vzorky 1200 Američanov predstavujú dostatočne silný dôkaz na to, aby sme mohli zamietnuť nulovú hypotézu. Použime hladinu významnosti $\alpha = 0,05$.

Čo ak by sme použili hladinu významnosti $\alpha = 0,01$?

Príklad 2 – ešte raz

Hans náhodne vybral 1200 Američanov, dal im robiť test a zistil, že Američanov správne odpovedajúcich na otázku č. 6 bolo 420.

Overme, či údaje získané zo vzorky 1200 Američanov predstavujú dostatočne silný dôkaz na to, aby sme mohli zamietnuť nulovú hypotézu. Použime hladinu významnosti $\alpha = 0,05$.

Čo ak by sme použili hladinu významnosti $\alpha = 0,01$?

Testovanie hypotéz pomocou p -hodnoty

1. Formulovanie nulovej a alternatívnej hypotézy.
2. Stanovenie hladiny významnosti α .
3. Overenie CLT (pozor pre relatívnu početnosť používam hodnotu p)
4. Výpočet p-hodnoty pomocou SE a z-skóre.
5. Rozhodnutie
 - ak je p-hodnota menšia ako hladina významnosti, tak zamietame nulovú hypotézu v prospech alternatívnej hypotézy
 - ak je p-hodnota väčšia prípadne rovná hladine významnosti, nulovú hypotézu nemožno zamietnuť