PSY117
Statistická analýza dat v psychologii
Přednáška 9 2018

Statistické testování hypotéz

Země je kulatá ($p<0,05$).
Jacob Cohen

Od vzorku k populaci a zpět

Vzhledem k tomu, jaká nám na vzorku vyšla statistika, jaký je odpovídající populační parametr?

interval spolehlivosti

Pokud předpokládáme, že v populaci je hodnota parametru X, co si myslet o své hypotéze poté, co nám na vzorku vyšlo Y?

statistický test hypotézy

Hypotézy

- Příklady (statistických) hypotéz
  - H: $\mu = 100$: Populační průměr IQ je roven 100.
  - H: $\sigma = 10$: Populační směrodatná odchylka je 10.
  - H: $\mu_1 - \mu_2 = 0$: Populační průměry μ_1 (psychotici) a μ_2 (zdraví) jsou stejné.
  - H: $\rho_{xy} = 0$: Proměnné X (pití piva) a Y (dominance) spolu nekorelují
- Hypotézy?
  - Velké slovo pro (malá i velká) očekávání
  - Očekávání plynoucí ze zkušenosti a z předchozích empirických výzkumů
  - Očekávání plynoucí z teorie … s potenciálem teorii zpochybnit

AJ: statistical hypotheses testing, hypothesis, hypothesis supported by data

Na vzorku 100 náhodně vybraných dospělých jsme zjistili průměrné IQ (*m*) rovné 105 (*s* =14).
H: μ = 100

Možné interpretace:

A. *H* neplatí, vidíme přeci, že průměr není 100, ale 105.
B. 105 je dost blízko 100, to je jen výběrovou chybou. Není důvod považovat *H* za neplatnou.
C. … něco jiného je špatně

Statistický test hypotézy

Statistické testování založeno na p-nosti

- Známe-li pravděpodobnostní rozložení statistik můžeme usuzovat, **jak pravděpodobná je určitá výběrová statistika vzhledem k hypotéze: *P* (*D* |*H*)**
  - *Př. D* : $m=105$ nebo rozdíl mezi statistikou a hypotézou $|m-\mu| = 5$
    H : $\mu = 100$
  - *P* (*D* |*H*) je *P* ($m=105 \mid \mu = 100$) resp. *P* ($|m-\mu| \geq 5 \mid \mu = 100$)
- Je-li *P* (*D* |*H*) relativně vysoká, je tím hypotéza podpořena.
- Je-li *P* (*D* |*H*) relativně nízká, hypotéza je „činěna méně p-nou“

- Jak relativně „vysoká$_{\text{nízká}}$“ je vysoká$_{\text{nízká}}$ pravděpodobnost, abychom hypotézu podpořili$_{\text{zpochybnili}}$?

Jak vysoká $P(D|H)$ je nutná k podpoře H?

- Bayesovský přístup – otázka není relevantní
  - s H je spojena určitá p-nost a ta se díky $P(D|H)$ zvyšuje či snižuje
  - Bayesův teorém: $\boldsymbol{P(H|D)} = P(H) * P(D|H) / P(D)$
- Fisher – otázka je celkem relevantní
  - Princip falzifikace – H nelze potvrdit, pouze vyvrátit
  - Zamítnutí (zpochybnění) H_0: $P(D|H) <$ **0,05**; **0,01** podle oborových zvyků
  - Výsledek: Je-li $P(D|H)$ nízká, buď jsme měli štěstí/smůlu, nebo není H vhodným vysvětlením(modelem) dat. Další výzkum by měl prověřit tyto možnosti.
  - Flexibilní, dosti subjektivní přístup vhodný pro malé výzkumné programy

Jak vysoká $P(D|H)$ je nutná k podpoře H? (pokr.)

- Pearson, Neyman – otázka je naprosto relevantní
  - Frustrováni subjektivitou Fisherova přístupu
  - Jak často se budem mýlit, když budem při nízké $P(D|H_0)$ zamítat?
  - K odpovědi potřebujeme…
    - pevně stanovit hranici zamítání H
    - stanovit hypotézu, kterou budeme považovat za platnou, když zamítneme H – **alternativní hypotéza H_1**
    - mít představu o velikosti rozdílu mezi nulovou a alternativní hypotézou – **velikost účinku**
    - zajistit, aby pravděpodobnost zamítání H, pokud by H skutečně nebyla pravdivá, byla dostatečně vysoká – **síla testu**
  - Zavedli tedy princip vzájemně se doplňujících konkurenčních H
    - Vytvořme takovou H, kt. bude negací naší vědecké hypotézy a říkejme jí **nulová H**. Když se nám podaří nulovou H zamítnout, znamená to **podporu** pro naší vědeckou hypotézu.

But we may look at the purpose of tests from another view-point. Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behaviour with regard to them, in following which we insure that, in the long run of experience, we shall not be too often wrong. Here, for example, would be such a "rule of behaviour": to decide whether a hypothesis, H, of a given type be rejected or not, calculate a specified character, x, of the observed facts; if $x > x_0$ reject H, if $x \leqq x_0$ accept H. Such a rule tells us nothing as to whether in a particular case H is true when $x \leqq x_0$ or false when $x > x_0$. But it may often be proved that if we behave according to such a rule, then in the long run we shall reject H when it is true not more, say, than once in a hundred times, and in addition we may have evidence that we shall reject H sufficiently often when it is false.

[Neyman & Pearson, 1933]

převzato z D. Lakens http://www.educationandlearning.nl/uploads/cfeal/attachments/Presentation%20Daniel%20Lakens%20-%20morning_0.pdf

Dichotomizace výsledků výzkumu

- Výsledek výzkumu je v P-N přístupu zredukován na ano-ne

	H_0 podržena $P(D\|H_0) \geq \alpha$	**H_0 zamítnuta** $P(D\|H_0) \geq \alpha$
H_0 pravdivá (žádný efekt)	OK	chyba 1. typu α *(její pravděpodobnost)*
H_0 nepravdivá (efekt)	chyba 2. typu β	OK *P*: **Síla** (1-β)

Čím nižší je α, tím vyšší je β. Přesná podoba vztahu závisí na použitém testu. α i β mohou být nízké pouze při vysokých *n*.

AJ: type-I error, type-II error, (statistical) power

Terminologická vložka

H_0 : **nulová (statistická, testová, testovaná) hypotéza**

- obvykle logická negace (doplněk) vědecké hypotézy
- ve Fisherovském přístupu prostě hypotéza, jejíž testování pokládáme za přínosné

H_1 : **alternativní (vědecká, výzkumná) hypotéza**

- N-P (NHST): ta, o kterou nám primárně jde, doplněk nulové

$P(D|H_0)$, podle které rozhodujeme o víře v platnost H_0

- značí se ***p***, též p-value, p-hodnota (nebo v SPSS ***Sig.***, ale to je fuj)
- Je-li stanovena dopředu (N-P): **úroveň/hladina statistické významnosti** (průkaznosti), α, udává se často v procentech: 5%, 1%
  - p-nost chybného zamítnutí H_0 - **chyba prvního typu**

Jednostranné vs. **oboustranné** hypotézy

- jednostranné, směrové: H_0: $\mu \geq 23$, H_1: $\mu < 23$, z různých důvodů užíváme zdrženlivě
- oboustranné: H_0: $\mu = 23$, H_1: $\mu \neq 23$, připouští rozdíl oproti H_0 na obě strany

AJ: null hypothesis, scientific/alternative hypothesis, level of statistical significance, type I error, one-tailed, two-tailed, directional

http://rpsychologist.com/d3/NHST/

Pravděpodobnosti různých výsledků

	H_0 podržena $P(D\|H_0)\geq\alpha$	**H_0 zamítnuta** $P(D\|H_0)\geq\alpha$
H_0 pravdivá (žádný efekt)	OK	chyba 1. typu α *(její pravděpodobnost)*
H_0 nepravdivá (efekt)	chyba 2. typu β	OK *P*: **Síla** (1-β)

α = *P*(zamítnutí $\boldsymbol{H_0}$ | $\boldsymbol{H_0}$ pravdivá)
Nepodmíněná *P*(chyba I. typu)=α.*P*($\boldsymbol{H_0}$ pravdivá)

β = *P*(nezamítnutí $\boldsymbol{H_0}$ | $\boldsymbol{H_0}$ nepravdivá)
Nepodmíněná *P*(chyba II. typu)=β.*P*($\boldsymbol{H_0}$ nepravdivá)

Postup testování statistické hypotézy

1. Formulujte **testovou** (nulovou) **hypotézu**, kterou budete testovat (tj. vyvracet) (př. H_0: $\mu = 0$, nebo H_0: $\mu = 6$)
2. Zvolte **hladinu statistické významnosti**, tj. míru rizika, že dojde k chybě 1. typu (např. $\alpha = 0{,}05$) *(pro Fisheriány není nutno)*
3. Hledáme p-nost získání naší výběrové statistiky nebo extrémnější hodnoty, za předpokladu, že H_0 je pravdivá: $P(D|H_0)$, p
   - cesta vede přes znalost výběrového rozložení statistiky
   - např. $m = 0{,}5$. $P(|m| \geq 0{,}5|\mu=0)$
   - obvykle je nutný přepočet na tzv. *testovou statistiku*, např. *t, z…*
4. Zformulujeme závěr o H_0:
   - je-li $P(D|H_0) < \alpha$, pak H_0 *zamítáme* (P-N), zpochybňujeme (F)
   - je-li $P(D|H_0) \geq \alpha$, pak H_0 podpoříme

Příklad – jednovýběrový *t*-test

Terapie nevhodného chování.

- Rozdíl před-po: m=2,7; s=3,5; N=10
- H : Terapie má efekt. ($\mu \neq 0$) – oboustranná hypotéza

1. H_0 : Terapie nemá efekt: $\mu = 0$
2. V sociálních vědách běžně α=0,05
3. $P\ (|m| \geq 2{,}7 | \mu = 0) = ?$
   - $s_m = 3{,}5/\sqrt{10} = 1{,}1$
   - $t = (m-\mu)/s_m = 2{,}7/1{,}1 = 2{,}45$
   - $P\ (|t| \geq 2{,}45 | \tau = 0) = 2*(1–\text{T.DIST}(2{,}45;9;1)) = 0{,}04$ (nebo TDIST(2,45;9;2))
4. $P\ (|m| \geq 2{,}7 | \mu = 0) < 0{,}05$ >> zpochybníme H_0 - rozdíl mezi D a H_0 je **statisticky významný(průkazný, signifikantní)**

Protože m =2,7 je velmi málo pravděpodobný, kdyby byl rozdíl byl 0, tak nalézáme nepřímou podporu pro přesvědčení, že $\mu > 0$.

Příklad – jednovýběrový *t*-test

Terapie nevhodného chování.

- Rozdíl před-po: m=2,7; s=3,5; N=10
- H : Terapie má efekt. (μ >0) – ***jednostranná*** hypotéza

1. H_0 : Terapie nemá efekt: μ = 0 (Technicky je to $\mu \leq 0$, ale očekávání budujeme od toho =)
2. V sociálních vědách běžně α=0,05
3. $P\ (m \geq 2{,}7 | \mu = 0)$ = ?
   - $s_m = 3{,}5/\sqrt{10} = 1{,}1$
   - $t = (m-\mu)/s_m = 2{,}7/1{,}1 = 2{,}45$
   - $P\ (t \geq 2{,}45 | \tau = 0)$ = 1–T.DIST(2,45;9;1) = 0,02 (nebo TDIST(2,45;9))
4. $P\ (m \geq 2{,}7 | \mu = 0) < 0{,}05$ >> zamítáme H_0 - rozdíl mezi D a H0 je **statisticky významný(průkazný, signifikantní)**

Protože při m =2,7 málo pravděpodobný, kdyby byl rozdíl 0 nebo menší, tak nalézáme nepřímou podporu pro μ >0.

Příklad – jednovýběrový *t*-test

Terapie nevhodného chování.

- Rozdíl před-po: m=–2,7; s=3,5; N=10
- H : Terapie má efekt. (μ >0) – ***jednostranná*** hypotéza

1. H_0 : Terapie nemá efekt: $\mu = 0$ (Technicky je to $\mu \leq 0$, ale očekávání budujeme od toho =)
2. V sociálních vědách běžně α=0,05
3. $P\ (m \geq -2{,}7 | \mu = 0) = ?$
   - $s_m = 3{,}5/\sqrt{10} = 1{,}1$
   - $t = (m-\mu)/s_m = -2{,}7/1{,}1 = 2{,}45$
   - $P\ (t \geq -2{,}45 | \tau = 0) = 1 - $T.DIST(–2,45;9;1) = 0,98 (nebo 1-TDIST(2,45;9;1))
4. $P\ (m \geq -2{,}7 | \mu = 0) > 0{,}05$ >> **ne**zamítáme H_0 - rozdíl mezi D a H0 není **statisticky významný(průkazný, signifikantní)**

Protože při m =–2,7 je pravděpodobnější, že je rozdíl 0, než že je pozitivní, ponecháváme nulovou hypotézu v platnosti.

Jednostranné testy

- Používáme pouze, pokud rozdíl, který by měl opačné znaménko, než čekáme, je bezvýznamný, neinterpretovatelný.
  - Specificky se dají využít, když si přejeme nalézt explicitní podporu pro neexistenci rozdílu/korelace – TOST (Two One-Sided Test, test ekvivalence)
- Obvykle uvažujeme v jednostranných hypotézách, ale testujeme je oboustranně.
- Oboustranné testování je „bezpečná“ volba. Jednostranné obvykle přitahuje žádost o zdůvodnění.

Problémy statistického testování *H*

- Dichotomizace rozhodnutí
  - stejná *velikost účinku* dává při různých *N* jiné rozhodnutí o H_0
  - komplikuje kumulativní budování znalostní báze
- Problém interpretace *p*-hodnot
  - $p = P(D \mid H_0)$ a nikoli $P(H \mid D)$
  - *p* udává překvapivost dat, může vést k vyvrácení *H*, ne však k přijetí *H*
- Problém nulové hypotézy
  - Test je smysluplný, jen když je nulová hypotéza smysluplná.

Největší problém je tedy formální, bezmyšlenkovité testování.

- Jak z problémů ven?
  - VŽDY se primárně zajímat o velikost účinku (Cohenovo *d*, *r*, R^2, η^2, ω^2)
  - používat intervalové odhady, kdy to jen lze
  - testování hypotéz používat pouze doplňkově

Doporučené čtení

- Cohen.
- ASA statement 2016:
  http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108

Lakensova prezentace pro mírně pokročilé:
http://www.educationandlearning.nl/uploads/cfeal/attachments/Presentation%20Daniel%20Lakens%20-%20morning_0.pdf

Shrnutí

- Statistické testování hypotéz vychází z konstrukce intervalu spolehlivosti pro hypotetizovaný parametr
- Může znamenat (ne)podporu pro hypotézu, nikoli striktně potvrzení/vyvrácení