

Statistická analýza dat

Kontakt na vyučující:

Mgr. Helena Klimusová
Psychologický ústav FF MU
A. Nováka 1, Brno
budova C – 5. patro

- tel. 549494035
- e-mail klimusov@fss.muni.cz
- konzultační hodiny v JS 2005:
středa 16.00-17.00

Statistická analýza dat

- cíl kurzu
- organizace výuky
- osnova kurzu
- doporučená literatura
- hodnocení

Cíl kurzu

- obeznámit se s metodami a základními postupy statistické analýzy
- naučit se používat statistický software SPSS
- získat znalosti potřebné pro popis, analýzu a interpretaci výsledků kvantitativního výzkumu
- tyto znalosti aplikovat při analýze dat bakalářských a diplomových prací
- naučit se číst a porozumět metodám a výsledkům statistické analýzy ve výzkumných zprávách publikovaných v odborném tisku

Osnova

- **1/3 Úvod**
podstata statistiky, základní pojmy; úloha statistiky v psychologickém výzkumu, popisná a inferenční statistika, diskrétní a spojité proměnné, úroveň měření
- **8/3 Popisná statistika**
základní pojmy, četnostní tabulky, grafická zobrazení, bivariační tabulky, míry centrální tendence, míry rozptýlení, normální rozdělení
- **15/3 Inferenční statistika, odhady parametrů**
úvod do inferenční statistiky, pravděpodobnost, výběrové soubory, metody získávání reprezentativních vzorků, bodové a intervalové odhady parametrů, požadovaný rozsah výběru
- **22/3 Testování hypotéz**
princip, formulace hypotéz, pětistupňový model testování hypotéz, chyby I. a II. druhu, jednovýběrové testy hypotéz o průměru a podílu
- **29/3 Testování hypotéz**
dvouvýběrové testy hypotéz o průměru a podílu
- **5/4 Testování hypotéz**
analýza rozptylu
- **12/4 Testování hypotéz**
neparametrické pořadové testy, test Chí kvadrát
- **19/4 Vztahy mezi proměnnými**
základní pojmy, vztahy mezi proměnnými na nominální a ordinální úrovni měření - míry asociace založené na Chí kvadrátu, PRE míry, Spearmanův koeficient pořadové korelace
- **26/4 Vztahy mezi proměnnými, korelační a regresní analýza**
vztahy mezi proměnnými na intervalové a poměrové úrovni, grafické zobrazení, Pearsonův koeficient korelace, jednoduchá lineární regresní analýza
- **3/5 Multivariační techniky**
bivariační tabulky - kontrola pro další faktor; parciální korelace, mnohonásobná lineární regrese
- **10/5 Multivariační techniky; analýza reliability**
nástin dalších technik (faktorová analýza, shluková analýza); analýza reliability
- **17/5 Presentace výsledků statistické analýzy ve výzkumné zprávě**

Doporučená literatura

- Hendl, Jan: *Přehled statistických metod zpracování dat*. Praha: Portál 2004.
- Healey, Joseph F. *Statistics: A Tool for Social Research*. 3. vyd. Belmont, California: Wadsworth, 1993.
- Swoboda, Helmut. *Moderní statistika*. 1. vyd. Praha: Svoboda, 1977.
- Gregor, Miroslav, Přadková, Svatomíra, Spěšná, Daniela. *Statistika pro sociology*. 2. vyd. Brno MU 1993.

Hodnocení

- kurz je ukončen zkouškou
- do celkového hodnocení (max. 70 bodů) se započítává i výsledek průběžného testu a body za domácí úkol
- **průběžný test** bude písemný, cca 10 min, možno získat max. 15 bodů; plánován na 19. dubna 2005
- **domácí úkol** – zpracování dat z projektu z Metodologie; možno získat max. 15 bodů; termín odevzdání 24.5. (vložit do odevzdávnice v IS)
- **zkouška** sestává z praktické části - vyřešení zadaných příkladů a sepsání výsledků; a ústní části - interpretace výsledků, vysvětlení statistických pojmů atd.; celkem max. 40 bodů
- hodnotí se především schopnost vybrat a správně použít adekvátní postup statistické analýzy (15 bodů) a schopnost interpretovat získané výsledky (20 bodů), mnohem méně podstatná je zdatnost v ovládní SPSS (5 bodů)
- **známka** se odvozuje z celkového počtu bodů (úkol + průběžný test + zkouška)
 - A: 61-70 bodů
 - B: 56-60 bodů
 - C: 51-55 bodů
 - D: 46-50 bodů
 - E: 41-45 bodů
 - F: 0-40 bodů

Statistika – úvod

Osnova:

- vymezení statistiky
- úloha statistiky v psychologickém výzkumu
- základní pojmy - měření, proměnné (diskrétní a spojitě proměnné; podle úrovně měření); popisná a induktivní statistika; populace a vzorek
- příprava dat před analýzou

Definice statistiky

- Statistika je naukou o tom, jak získat informace z numerických dat.
- Statistika je vědní obor zabývající se zkoumáním jevů, které mají hromadný charakter.
- Statistika je soubor metod, které nám umožňují rozhodnutí v případě nejistoty.
- Analýza dat náhodného charakteru, která předpokládá použití pokročilejšího aparátu matematiky a určitého matematického modelu (teorie pravděpodobnosti).
- Statistická analýza jsou záhadné, někdy až prapodivné manipulace s daty získanými pomocí experimentu, jejichž cílem je zastřít ten fakt, že výsledky nemají pro lidstvo žádný zobecnitelný význam. Obvykle jsou při ní používány počítače, což této proceduře propůjčuje další auru nereálnosti.*

Role statistiky ve výzkumu

- při výzkumu v sociálních vědách jsou střídavě používány deduktivní a induktivní postupy
- statistika se uplatňuje především u induktivních postupů
- (deduktivní postupy – z teorie jsou odvozeny hypotézy a z nich postup pozorování/měření; induktivní postupy: z pozorování jsou odvozeny obecné závěry, které mohou modifikovat teorii atd.)

Proměnné

- při výzkumu psychologové získávají, měří a analyzují **proměnné**
- proměnná je objekt, který může nabývat různých hodnot (na rozdíl od konstanty)
- jsou rozlišovány různé **typy proměnných**
- podle jejich role ve výzkumném plánu -závislé, nezávislé, intervenující
- diskrétní a spojité proměnné
- podle úrovně měření – nominální, pořadové, intervalové a poměrové

Závislé a nezávislé proměnné

- příklad experimentu: chceme zjistit účinnost prostředku na zlepšení paměti u seniorů ve srovnání s placebem
- **nezávislá** proměnná – podávaný prostředek (testovaný lék x placebo)
- **závislá** proměnná – výkon v testu paměti (např. počet vybavených slov)
- možné **intervenující** proměnné – ochota užívat léky, podmínky při testu, věk...

Závislé a nezávislé proměnné

- jiný příklad: Milgramův experiment s poslušností vůči autoritě

Závislé a nezávislé proměnné

- **nezávislá** proměnná – viditelnost a slyšitelnost žáka a přítomnost autority
- **závislá** proměnná – nejvyšší hodnota uděleného šoku
- možné **intervenující** proměnné –pohlaví učitele/žáka atd.

Závislé a nezávislé proměnné

- **nezávislá** proměnná – ta, se kterou experimentátor manipuluje (independent variable)
- **závislá** proměnná – proměnná měřená experimentátorem

(dependent variable)

- **intervenující** proměnná – e. se snaží její vliv eliminovat (znáhodňováním, vyrovnáváním) (intervening variable)

Kvalitativní a kvantitativní data

- **kvalitativní (kategorální)** – zjišťujeme hodnotu znaku – kategorií (např. pohlaví, náboženské vyznání, rodinný stav, barva očí, vzdělání); často fungují jako nezávislé proměnné (faktory); (categorical data)
- **kvantitativní (metrická) data** – zjištěná měřením pomocí nějakého nástroje (v širším smyslu); skóre v testu, tělesná váha...; obvykle závislé proměnné (measurement data)

Úroveň měření

- kategoriální proměnné jsou měřeny buď na nominální nebo pořadové úrovni
- **nominální** – čísla jsou hodnotám proměnné přiřazena náhodně; pouze hodnoty označují, ale nelze s nimi jako s čísly zacházet (jediné operace jsou = a \neq) (nominal scales)
- **pořadová** (ordinální) – hodnoty je možno uspořádat podle velikosti (např. pořadí v závodě, postojové škály), ale není možno stanovit, o kolik se liší (ordinal scales); možné početní operace jsou stále = a \neq , ale také $< a >$
- **intervalová** – hodnoty je možno uspořádat a vzdálenosti mezi nimi (intervaly) jsou shodné (např. rok narození, teplota) (interval scales); možné početní operace: = a \neq , $< a >$, + a -
- **poměrová** – stejné vlastnosti jako intervalová a navíc se mezi hodnotami vyskytuje přirozená 0 (indikující absenci znaku) – tj. má smysl se ptát, kolikrát je hodnota větší než jiná hodnota (ratio scales); např. věk, počet správně řešených úloh, počet dětí; kromě předchozích početních operací je možné také násobení a dělení

Úroveň měření

- kvantitativní data je možno zredukovat na kategoriální (např. místo hodnoty krevního tlaku údaj normotenze x

hypertenze)

- někdy je obtížné stanovit hraniční hodnoty (cut-off scores)

Účel typologie proměnných

- má pomoci při rozhodování o tom, jak nejlépe zobrazit, shrnout či analyzovat data

Populace a výběr

- data mohou být získávána na populaci (cenzus) nebo na vzorku populace (výběrové šetření)

Populace a výběr

- **populace** (základní soubor) - úplný souhrn lidí, objektů nebo věcí, které jsou předmětem našeho zkoumání
- je dána přesným stanovením jeho prvků
- prvky mohou být určeny buď jejich výčtem nebo vymezením některých společných vlastností (např. územní příslušnost, věk atd. – např. všechny děti s diagnostikovanou poruchou pozornosti a trvalým bydlištěm v JM kraji)

Populace a výběr

- **výběr** (výběrový soubor, vzorek) – výzkumníci se snaží, aby byl reprezentativní vůči příslušné populaci (pomocí postupů náhodného výběru)
- stejný soubor objektů může za určitých okolností představovat jednu populaci a jindy výběr

Parametry a statistiky

- pokud provádíme měření na celé populaci, jsou výsledky nazývány **parametry** populace; bývají označovány řeckými písmeny (např. průměr μ)
- výsledky z měření na vzorku se nazývají (výběrové) **statistiky**

- většinou jsou užívány k odhadu populačních parametrů pomocí postupů **statistického usuzování (inference)**

Popisná a inferenční statistika

- **popisná (deskriptivní, explorační) statistika** – třída technik, které slouží k popisu proměnných (např. typická hodnota proměnné, distribuce hodnot atd.)
- **inferenční statistika** – umožňuje rozhodnutí, zda zjištění ze vzorku platí i na populaci (a s jakou pravděpodobností) – odhady intervalu spolehlivosti, testování hypotéz
- inferenční statistika se dále dělí na **parametrické a neparametrické** postupy
- parametrickými testy ověřujeme platnost hypotézy o parametrech základního souboru (např. průměr) – vyžadují měření na intervalové úrovni

Příprava dat před analýzou

Osnova:

- uspořádání dat
- kontrola přesnosti dat
- vytvoření struktury databáze
- vkládání dat do PC
- transformace dat

Uspořádání dat

- výzkumná data mohou pocházet z různých zdrojů: dotazníky, rozhovory, měření při experimentech (pretest, posttest), pozorování
- je třeba všechny formuláře, dotazníky, záznamové archy atd. řádně označit, aby je vždy bylo možno i zpětně dohledat
- původní data se u profesionálních výzkumů po určitou dobu **archivují** (min. 5-7 let)
- je vhodné vytvořit si databázi – **přehled** všech použitých zdrojů dat (např. v MS Access nebo Excel, Statistica, SPSS)

Kontrola přesnosti dat

- někdy vhodné provést již při sběru dat – pak je možno příp. chybějící nebo jinak problematické odpovědi vyřešit na místě
- kontrola**: čitelnosti odpovědí, úplnosti odpovědí, důležitých informací (datum, jméno výzkumníka atp.)
- je vhodné z dalších kroků vyloučit nevalidní odpovědi (na všechny otázky stejné atd.)

Vytvoření struktury databáze

- data je možno vkládat přímo do statistických programů nebo do databázových programů (a potom je převést)
- rozsáhlé výzkumné projekty mají vytištěnu podrobnou kódovací příručku (*codebook*) s popisem proměnných

Vytvoření struktury databáze

pro každou proměnnou:

- jméno proměnné (např. *pohlaví*)
- popis proměnné (např. *celkový skór dotazníku deprese*)
- popis hodnot proměnné (např. *1=chlapci, 2=dívky*)
- způsob kódování chybějících hodnot (např. 9)
- formát proměnné (numerický, textový, datum...)

Vkládání dat do PC

- pro **kontrolu přesnosti vkládání** existují nejrůznější postupy
- např. dvojité vkládání (*double entry*) – speciální program, který při druhém vkládání porovnává data s původním údajem a upozorní na chyby; není příliš rozšířeno
- jednodušší je náhodně zkontrolovat určitý počet případů (při velkém procentu chyb kódovat znovu a porovnat)
- vždy je nutno alespoň zkontrolovat, zda rozsah zadaných hodnot odpovídá předpokladům

Transformace dat

- obrácení položek** – některé položky bývají formulovány v opačném směru než zbytek škály;
- jejich hodnoty je třeba překódovat (např. u 5 bodové škály se bude 1 rovnat 5, 2 bude 4 atd. až 5 bude 1)
- vzorec pro obrácení škály
$$\text{nová hodnota} = \text{nejvyšší hodnota} + 1 - \text{stará hodnota}$$
- v SPSS je možno provést pomocí příkazů COMPUTE nebo RECODE

Transformace dat

- výpočet celkových skóru** – většinou součtem hodnot několika proměnných (příp. průměr atp.)
- v příkazu COMPUTE se zadá příslušný vzorec, např. pro součet $\text{celk} = \text{p1} + \text{p2} + \text{p3}$

- nebo rovnou funkce (*sum, mean*)
celk = sum (p1 TO p3)

Transformace dat

- **sloučení hodnot proměnných** – např. ze 7 bodové škály chceme udělat 3 bodovou; nebo z věku zadaného v letech vytvořit 3 věkové skupiny
- příkaz RECODE

Kontrolní otázky

- klasifikace proměnných podle jejich role ve výzkumném plánu
- rozdíly mezi nominální, ordinální, intervalovou a poměrovou úrovní měření; možné početní operace
- vymezení populace a výběru, parametrů a statistik
- vymezení popisné a inferenční statistiky
- postup přípravy dat před analýzou

Zopakovat z kurzu metodologie

- výzkumný projekt
- měření, objektivita, reliabilita, validita
- výzkumné plány
- výběr (postupy)

Literatura

- Hendl – kapitoly 1 a 2
- doplňující:
 - Gaito, J. (1980). Measurement scales and statistics: Resurgence of an old misconception. *Psychological Bulletin*, 87, 564-567
 - Velleman, P. F. & Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading. *The American Statistician*, 47(1), 65-72.
 - Lord, F. (1953). The Statistical treatment of football numbers. *The American Psychologist*, 8, 750-751.