OPAKOVÁNÍ SEZNÁMENÍ SE SPSS

PSYb252₀
Statistická analýza dat II
1. setkání

DNEŠNÍ PROGRAM

Představení kurzu a zdrojů

- Učebnice
- Data

Představení SPSS

Základní analytické postupy

CÍLE KURZU

Získat praktickou schopnost provádět statistické analýzy s více než 2 proměnnými

Rozumět prezentovaným výsledkům

Korektně komunikovat výsledky analýz

PŘEHLED TÉMAT

Seznámení se SPSS

Základní analytické postupy

Lineární regrese

Logistická regrese

Analýza rozptylu

Víceúrovňový lineární model

(Faktorová analýza v PSYb2590)

POŽADAVKY A ZKOUŠKA

Zpracování všech průběžných úkolů (trojice)

- Na každý seminář, nebodováno

Vstupní test – teorie statistické indukce (Field kap. 2)

Průběžný test

- Opakování, okruh č. 1
- Cvičné testy v ISu, budou rozšířeny

Zápočtový test

- Teoretické znalosti, termíny

Zkouška

- Během hodiny a půl vypracovat zprávu z analýzy na počítači
- Se všemi zdroji, vč. googlení

UČEBNICE

- Field, A.: ***Discovering statistics using SPSS***, 4th 5th
  - http://www.statisticshell.com/
  - http://www.uk.sagepub.com/field4e/study/default.htm
  - https://edge.sagepub.com/field5e

- Morgan et al (2002). ***From numbers to words. Reporting statistical results for the social sciences.*** Allyn & Bacon.

- American Psychological Association. (2001). ***Publication manual of the American Psychological Association*** *(6th ed.)*. Washington, DC: Author. **V říjnu 7. vydání!**

OPÁČKO S FIELDEM

Nelze všechno přečíst hned, projděte na test. Interaktivní sylabus se snaží ukázat, co jak číst.

Kap 1 – popisná statistika

Kap 2 – statistická indukce

Kap 3 – seznámení se SPSS

Kap 4 – Vyrábíme grafy

Kap 5 – Kontrolujeme předpoklady testů

Kap 6 – Neparametrické testy

Kap 7 – Korelace

Kap 9 – t-testy

Kap 18 – Chí-kvadrát

STATISTICKÝ SOFTWARE

- Umožňuje provádět analýzy rychle a ve velkém množství
- Nabízí možnosti správy dat – metadata, sdílení
- Ovlivňuje způsob práce – analýzy i jejich reportování

Excel – dostupnost, omezené možnosti pokročilejších analýz

IBM SPSS – sociální vědy, dinosaurus, garance – licence FSS

Statistica – všechny vědy, **-**skriptování – licence MU

Stata – i nejpokročilejší analýzy, nemáme multilicenci

R – možné je úplně vše, méně je garantováno

IBM SPSS

Nainstalovat z inet.muni.cz

Při instalaci je vhodné kývnout na nabídky ohledně python a R

SPSS - IMPORT A EXPORT DAT

.csv – obvyklý textový formát – hodnoty oddělené středníkem*, desetinná čárka, kromě názvů proměnných na 1. řádku žádná metadata

.xls(x) – MS Excel, metadata obvykle na samostatném listu

.sav – nativní formát dat SPSS, obsahuje hodně **metadat**

*V angličtině je standardním oddělovačem hodnot čárka a desetinný znak je tečka.
Problémy s importem dat za sebou často mají tuto prostou příčinu. SPSS i jiné programy si to nechají vysvětlit.

DATA: LONG2

Data zakladni.csv

Data zakladni.xls

Data zakladni.sav

SPSS – DATOVÁ MATICE

Datová matice, jak jsme se ji učili v PSY117 – *Data View*

- Názvy proměnných mohou být delší

Metadata zobrazena na samostatné záložce – *Variable View*

- **Typ proměnné** – numeric/string /date…
- **Label** – dlouhý název, popisek
- **Values** – popisky jednotlivých hodnot proměnné
- **Missing** – které hodnoty jsou kódy pro chybějící nebo neplatné odpovědi
- **Measure** – nominal/ordinal/scale

Třídění – pravým tl. myši, popř. *Data – Sort cases*

LONG2

Analyze > Descriptive statistics > Frequencies

Graphs > Chart builder

Graphs > Legacy dialogs > Bar

SPSS – OUTPUT, SYNTAX

Output – okno, kam se vypisují výstupy analýz, stromová hierarchie

Syntax – okno, jehož prostřednictvím se dají zadávat textové příkazy pro vykonání analýz

- Syntax je záznamem analýzy, podle kterého se dá znovu celá zopakovat

- I když příkazy nezadáváte prostřednictvím syntaxu/e, vše, co SPSS dělá, je zaznamenáno v *žurnálu*. Ten najdete Edit > Options > File Locations > Journal file
- Automatické zobrazování syntaxu v outputu: Edit > Options > Viewer > Display commands in the log

ZÁKLADNÍ POSTUP ANALÝZY DAT

1. Příprava, čištění a screening dat
2. Transformace, odvozené/vypočítané proměnné, rekódování
3. Popisné statistiky, vyjádření se k chybějícím datům
4. Plánované (konfirmační) analýzy
   a) ověření předpokladů
   b) testování plánovaných hypotéz / stanovení velikosti plánovaných efektů
5. Doplňkové, explorační analýzy

1. PŘÍPRAVA, ČIŠTĚNÍ A SCREENING DAT

Cílem je **mít** datovou matici podle pravidel z PSY117, **vědět**, co v ní je.

- Hrubá data je dobré mít uložena R/O a vždy pracovat s kopií.
- Tabulky četností, základní popisné statistiky – přípustné hodnoty
  - Kontingenční tabulky – mají data všechny skupiny účastníků?
- Používání kódů pro neplatná data. Opravování či mazání jen výjimečně, podle předem daných pravidel.
- Změny v datech dělat ideálně výhradně pomocí

PŘÍKLAD: LONG2 DATA

Data mají pocházet ze dvou kohort – šesťáků ZŠ a prváků SŠ

Jaké jsou přípustné věky v této populaci?

Analyze > Descriptive statistics > Explore

2. TRANSFORMACE PROMĚNNÝCH

Změna kódování proměnné, sloučení kódů/kategorií

- např. národnost můžeme chtít překódovat na dichotomii česká/cizí
- Transform – Recode into **Different** Variables… nebo v syntaxu RECODE … INTO…

Kategorizace spojité proměnné

- např. podle mediánu, či kvartilů

Vypočítání nové proměnné

- např. součet 5 položek do jednoho součtového skóru
- Transform – Compute variable… nebo v syntaxu COMPUTE nova=jedna+druha.

Transformacemi je dobré tvořit nové proměnné (nepřepisovat původní)

PŘÍKLAD: LONG2 DATA

Překódujme národnost

Vypočítejme proměnnou – počet dětí v rodině

```
po_deti = bratri_m + sestry_m + bratri_s + sestry_s + 1
```

3. POPISNÉ STATISTIKY, VYJÁDŘENÍ SE K CHYBĚJÍCÍM DATŮM

Popis rozložení hodnot relevantních (= použitých v analýze) proměnných – Je natolik souladu s očekáváním, aby byla následná analýza důvěryhodná?

Při analýze se díváme na momentové i pořadové statistiky a hlavně zobrazení rozložení jednotlivých proměnných.

Často se díváme i na bivariační vztahy mezi proměnnými, které jsou podkladem pro další analýzy

Reportujeme nejčastěji

(N),M,SD, min, max + komentář ke tvaru rozložení v textu pro spojité

četnosti pro kategorické

Univariační histogramy či boxploty výjimečně, jsou-li ústřední otázkou analýzy

3B. …VYJÁDŘENÍ SE K CHYBĚJÍCÍM DATŮM **MISSING DATA**

Data chybí z mnoha důvodů

- účastník se nakonec nezúčastnil, nebo poměrně brzy svou účast ukončil – UNIT NON-RESPONSE
- účastník využil svého práva na cokoli neodpovědět – ITEM NON-RESPONSE
- účastník odpověděl způsobem, který nelze považovat za platný, použitelný

Na důvodu chybění **záleží**

3B. …VYJÁDŘENÍ SE K CHYBĚJÍCÍM DATŮM **MISSING DATA**

Důvody chybění dat z hlediska statistiky

- missing data mechanism → missing data model
- MCAR – missing completely at random –
  - pro každého člověka je P chybění stejná, nijak to nesouvisí s tím, co měříme
  - kdybychom rozdělili účastníky na ty, kteří hodnotu mají, a ty, kteří ne, nenašli bychom u nich rozdíly v žádné proměnné
  - např. výpadek proudu, vypadlý list dotazníku, přeskočení položky v záznamovém archu (když nezkoumáme pozornost ;-)
  - Nedochází ke zkreslení statistik, jen k úbytku dat a přesnosti odhadů (CI)
  - Obvykle nerealistický předpoklad

3B. …VYJÁDŘENÍ SE K CHYBĚJÍCÍM DATŮM **MISSING DATA**

Důvody chybění dat z hlediska statistiky

- missing data mechanism → missing data model
- MAR – missing at random –
  - P chybění je závislá na proměnné, kterou máme změřenou
  - kdybychom rozdělili účastníky na ty, kteří hodnotu mají, a ty, kteří ne, našli bychom u nich rozdíly jedné nebo více proměnných
  - např. ve třídě s horším klimatem je vyšší P přeskočení položky
  - Když se v analýze zohlední také ty proměnné, které souvisí s chyběním, nedochází ke zkreslení statistik, jen k úbytku dat a přesnosti odhadů (CI)
  - Realističtější předpoklad, ne vždy ale máme vše potřebné změřeno

3B. …VYJÁDŘENÍ SE K CHYBĚJÍCÍM DATŮM **MISSING DATA**

Důvody chybění dat z hlediska statistiky

- missing data mechanism → missing data model
- NMAR/MNAR – not missing at random –
  - P chybění ovlivňuje něco a my **nevíme** co
  - Neznámé vlivy nelze zohlednit

3B. …VYJÁDŘENÍ SE K CHYBĚJÍCÍM DATŮM **MISSING DATA**

MCAR, MAR, NMAR jsou **předpokládané** modely

Je těžké podpořit volbu předpokladu argumenty

Máme-li hodně dat, je dobré zjistit, zda chybění s nějakou proměnnou nesouvisí – a pak ji zahrnout

- např. Analyze > Missing Value Analysis
  - mj. počítá, zda se ti, komu hodnota proměnné chybí a komu ne, liší v nějaké spojité proměnné (t-testy)
- provedením většího množství analýz hledajících vztahy – může být velmi pracné

3B. …VYJÁDŘENÍ SE K CHYBĚJÍCÍM DATŮM **MISSING DATA – CO S NIMI?**

1. Komunikovat, kolik čeho kde chybí
2. Zakomponovat příčiny chybění do modelů (můžeme-li)
3. Použít obecné způsoby naložení s chybějícími daty

Vyřadit z analýz respondenty, kteří mají chybějící data LISTWISE DELETION – nejjednodušší & nejhorší volba

Počítat každou jednotlivou statistiku ze všech dostupných dat – PAIRWISE DELETION – zachová více informace, neodstraní zkreslení. N pro celou analýzu? – obvyklá bezpracná volba

IMPUTACE – doplnění chybějících dat

- Nesmírně záleží na P-nostním modelu chybění-doplňování – vyšší dívčí

PŘÍKLAD: LONG2 DATA

Analyze > Descriptive statistics > Frequencies

pro národnost a počet dětí i se sloupcovým grafem

4. PLÁNOVANÉ (KONFIRMAČNÍ) ANALÝZY OVĚŘENÍ PŘEDPOKLADŮ – FIELD 6

I když předpoklady předpokládáme, je dobré se ujistit, můžeme-li.

Nedodržení předpokladů má různé konsekvence – je dobré být pozorný.

Normalita

- primárně souvisí s přesností odhadu SE a p-hodnot
- zajímá nás u reziduí (resp. uvnitř skupin)
- histogram, Q-Q plot, testy normality mohou být zrádné

Homoskedascita

- primárně souvisí s přesností odhadu SE a p-hodnot
- scatterploty, boxploty pro kategorické (kategorizované) prediktory
- Test homoskedascity (Breusch–Pagan test) v SPSS jen pomocí R extension
- Leveneho test netřeba – lepší je korigovat – Welschův t-test

4. PLÁNOVANÉ (KONFIRMAČNÍ) ANALÝZY OVĚŘENÍ PŘEDPOKLADŮ – FIELD 6

Co s nedodrženými předpoklady?

BOOTSTRAPPING – odhadování SE hrubou silou

- v SPSS k dispozici

Využití korekce – např. Weschova korekce u t-testu

TRANSFORMACE do normality

- Problematičtější, než se zdá. Vhodnější je použít model, který počítá se zešikmeným rozložením – generalizované modely

Trimming, winsorizing

Neparametrické testy – jen pro jednodušší analýzy

4. PLÁNOVANÉ (KONFIRMAČNÍ) ANALÝZY TESTOVÁNÍ HYPOTÉZ, ODHAD MODELU

1. Spočítání statistik, které jsou odhadem parametrů, kterými operuje hypotéza.
   - od spočítání průměrů a jejich rozdílu, četností a jejich rozdílu, či korelací po složitější modely
   - vyjádření velikosti účinku
2. Zohlednění nejistoty dané tím, že máme jen VZOREK
   - vytvoření intervalu spolehlivosti pro rozdíl či korelaci
   - test (nulové) hypotézy

V SPSS obvykle dostaneme obojí v jednom kroku.

PŘÍKLAD: LONG2

Analyze > Descriptive statistics > Crosstabs

5. DOPLŇKOVÉ ANALÝZY

Stejně jako konfirmační. Jen je musíme jako reportovat jako doplňkové.

ZÁKLADNÍ POSTUP ANALÝZY DAT

1. Příprava, čištění a screening dat
2. Transformace, odvozené/vypočítané proměnné, rekódování
3. Popisné statistiky, vyjádření se k chybějícím datům
4. Plánované (konfirmační) analýzy
   a) ověření předpokladů
   b) testování plánovaných hypotéz / stanovení velikosti plánovaných efektů
5. Doplňkové, explorační analýzy

PLÁN ANALÝZY

Pro preregistraci i bez ní je dobré mít plán výše uvedeného ještě před získáváním dat.

Lépe se o něm mluví se zkušenostmi, a tak jej necháme na později.

OPEN SCIENCE DESIDERATA (osf.io)

TRANSPARENCE, otevřenost

Nad rámec standardního sdělování výsledků analýz ….

- Sdílení/komunikování všech kroků výzkumu, zejm. analytického postupu → analytické skripty, SPSS syntax
- Sdílení dat
- Preregistrace

→

REPRODUKOVATELNOST, DŮVĚRYHODNOST, KUMULATIVNOST

PREZENTACE STATISTICKÝCH ANALÝZ SEKCE METHOD-RESULTS

Představení dat (vzorek, metoda) – V APA samostatné sekce.

Popis kroků provedených při čištění a transformaci dat.

Popisné statistiky (popř. zobrazení rozložení, tabulky/grafy dle APA)

Formulace hypotéz.

Zdůvodnění volby testu, popř. analytického postupu

Rekapitulace splnění předpokladů zvoleného testu

Standardní prezentace testových statistik (u jednodušších testů v textu, u složitějších modelů v tabulkách) vč. velikosti účinku (ideálně i intervalu spolehlivosti)

Interpretace výsledků testu (modelu) vzhledem k hypotéze

ZÁKLADY APA-STYLU PREZENTACE VÝSLEDKŮ

Styl veden principem typografické jednoduchosti.

1. Pokud nechceme prezentovat velké množství čísel (<10), uvádíme je v textu jako součást věty.
   - Věty pro výsledky běžných analýz jsou do značné míry standardizované. Měníme v nich jen názvy proměnných a hodnoty statistik. Tyto věty najdete jak ve Fieldovi (např. kap. 10.10), tak v Morganové. Najdete je také v empirických článcích, které čtete.

2. Tabulky jsou jednoduché pouze s vodorovnými oddělujícími čarami.
   - Tabulky mají titulek, z něhož je patrné, co v tabulce je (nespoléhá se na vysvětlení v textu)
   - Pod tabulkou bývají poznámky vysvětlující zkratky a další info nutné k porozumění

3. Grafy šetříme a počítáme s jejich černobílým zobrazením.
   - Snažíme se do nich vtěsnat tolik informace, aby to stálo za to.
   - V PSYb2520 se přimlouvám spíše za jejich větší využívání, když už jsme století ovocného netopýra opustili.

Uvádění *p*-hodnot

- Preferujeme uvádění přesné hodnoty, např. $p = 0{,}013$, spíše než porovnání se zvolenou hladinou alfa (tedy $p < 0{,}05$).

Uvádění čísel

- Požíváme pouze tolik desetinných míst, kolik jich nese nějakou informační hodnotu. Při obvyklé přesnosti měření v psychologii to obvykle znamená 2-3 významné číslice, tj. řády - 1,23; 12,3; 123 apod. Neuvádíme tolik desetinných míst, kolik jich nám SPSS vypíše!

Tuzemské typografické konvence – odlišnosti od angličtiny

- desetinná čárka, středník jako oddělovač hodnot v seznamu
- nula před desetinnou čárkou u čísel <1
- mezera mezi číslem a znakem %, když ho čteme „procent“ – 12 % lidí
- absence mezery mezi číslem a znakem %, když ho čteme „procentní“ – 40% líh

ÚKOL: NASTARTOVAT

- Nainstalovat si SPSS a sehnat si Fielda, Morganovou a APA manuál
- Zopakovat si obsah PSY117 – nejprve indukci
- Najít si parťáky

DĚKUJI ZA POZORNOST