Základy kvantitativní analýzy dat

Jan Kleiner

3. 4. 2024

BSSn4405 – online přednáška

jkleiner@mail.muni.cz

Statistika v sociálněvědním výzkumu

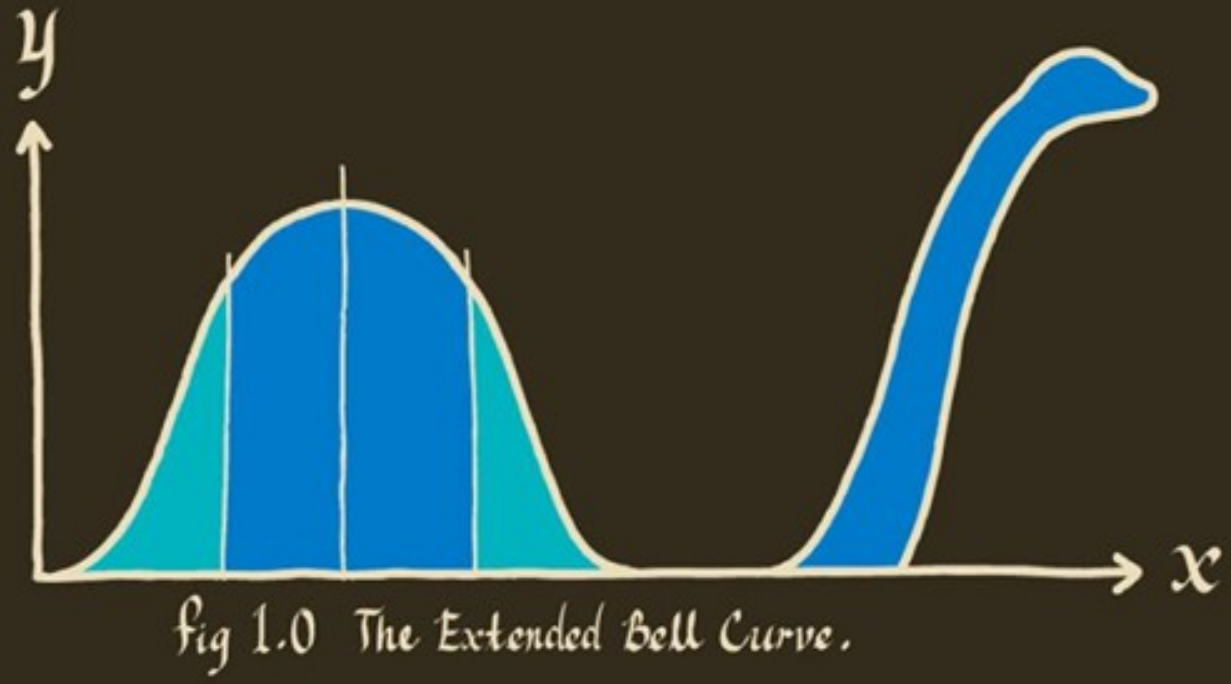


fig 1.0 The Extended Bell Curve.

Jan Kleiner
3. 4. 2024
BSSn4405
jkleiner@mail.muni.cz

Roadmap kurzu

Plánování a strategie výzkumu.

Vybrané metody sběru dat.

Základní přístupy k analýze dat.

Zajímavosti a nadstavba.

Roadmap kurzu

Plánování a strategie výzkumu.

Vybrané metody sběru dat.

Základní přístupy k analýze dat.

Zajímavosti a nadstavba.

Na této přednášce:

- se seznámíte se statistikou;
- zjistíte, že je nepostradatelná;
- zároveň ale zjistíte, že je zábavná, zajímavá a odhaluje skryté vzorce ve změti dat;
- budete vědět, kam dál se studiem statistiky;
- získáte nová výzkumná směřování.

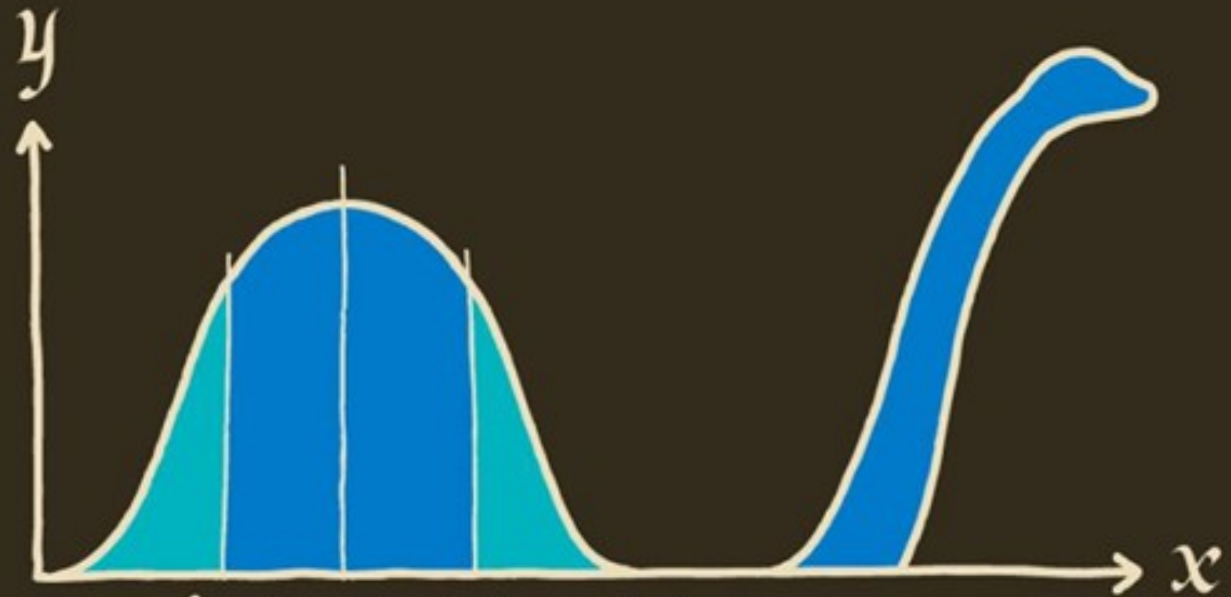
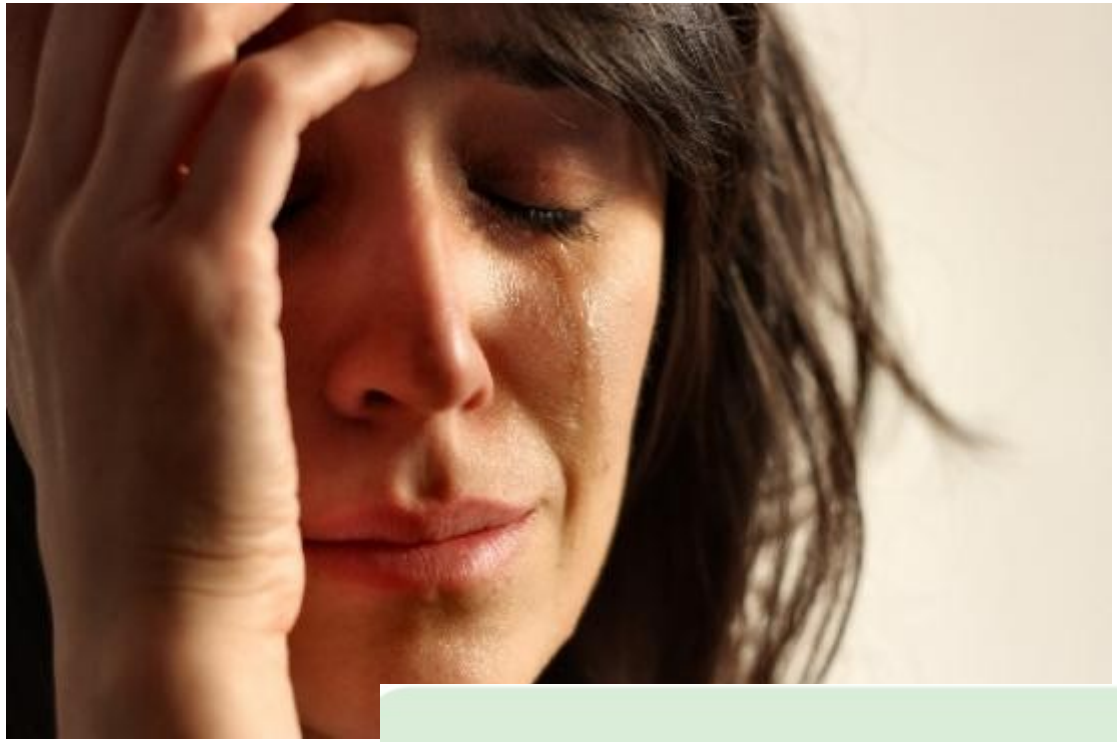


fig 1.0 The Extended Bell Curve.

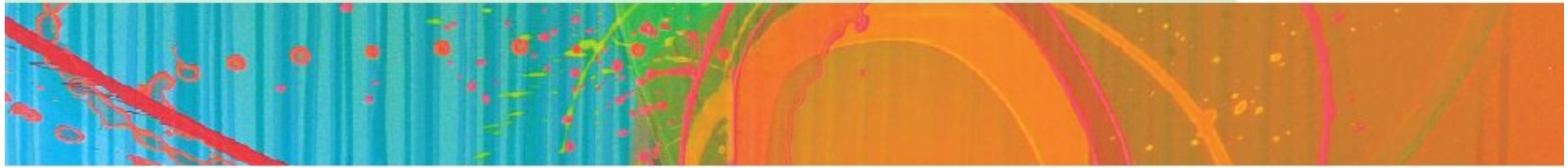
Literatura

- Povinná literatura a prezentace jsou komplementární.
- **Povinná literatura** obsahuje statistické základy.
  - **Andy Field** (2009: 31-60) – základy + všechny základní operace.
  - **Rabušic a kol.** (2019) – základy jinak a česky + některé operace.
- Další materiály
  - Magnellová a Van Loon – Seznamte se, statistika
  - Novotný, Svobodová - Jak pracuje věda?
  - Rumsey – Statistics for Dummies (spíše pro testování hypotéz).
  - Pennings a kol. – kvanti výzkum v politologii



Proč?
!

Why is my evil lecturer forcing me to learn statistics?



Zdroj: Field (2009)

1

Proč? I

- Pasivní i alespoň základní aktivní znalost statistiky je nezbytná.
- Podchycuje VEŠKERÉ hromadné jevy.
- V současné době je trendem smíšený výzkum (kvali+kvanti).
- Bez statistiky téměř nelze prokázat kauzalita.
  - No statistics, no experiment.
- Je objektivní (do jisté míry) a klade důraz na validitu a reliabilitu.

Proč? II

- Nové způsoby sběru dat vyžadují statistiku (kvůli velikosti) (viz např. *Everybody Lies* od Davidowitze, 2017).
  - Díky tomu navíc získáváme možnost zkoumat velmi zajímavá a neotřelá data (Google, Facebook apod.).
- Jen skok k AI a algoritmům rozhodování.
- Nutná, pokud chceme být „efektivními badateli“ (Rabušic a kol., 2019: 11).
- Má mě statistika nějak zajímat, i když nechci dělat „vědu“?
  - Ano – data science, vizualizace, efektivní evidence-based rozhodování apod.

„Statistické myšlení bude jednou pro efektivní občanství stejně nezbytné jako schopnost číst a psát“

(Wilks, 1950, cit. dle Rabušic a kol., 2019: 11)

→ Schopnost neskočit na špek ☺

„Student, který nezvládne metodologii kvant. výzkumu a analýzy dat, nebude ani dobrým výzkumníkem kvalitativním.“

(Rabušic a kol., 2019: 21)

Mně se taky zdá, že jsme tam všichni.

Statistika (Magnellová a Van Loon, 2010: 9-16)

- Původně politická aritmetika (*status*= státník).
- Nejprve vitální statistika.
  - Např. popisy a výčty sčítání lidu, sňatků.
  - Průměrné hodnoty.
- Později i matematická statistika.
  - „vědní obor zkoumající variabilitu, maticové počty. Zabývá se shromažďováním, klasifikací, popisem a interpretací dat získaných při sociálních průzkumech, vědeckých experimentech…“
  - Štěstí Skotů, Darwin, Malthus, Guiness, Florence Nightingale, hazard…
- Pro nás důležitá – deskriptivní a inferenční, popř. Bayesiánská statistika → tedy aplikovaná.

Větve a dělení statistiky

- Frekvenční statistika (fisherovská)
  - Deskriptivní
  - Inferenční
    - Univariační, multivariační modely.
- Bayesovská statistika
  - Apriori a aposteriori představa a jak se mění na základě našich dat (BF).
- Matematická vs. aplikovaná apod.
  → různé typy dělení a nejsou vyčerpávající

Inferenční statistika

Inference = odvození

- Odhad charakteristik populace ze vzorku.
- Lze vztah odhalený na vzorku očekávat v populaci?
- Testy významnosti (stojí na *p-value* a související hladině významnosti 95/99 %).
- Např. chí-kvadrát (nominální znaky), t-test (rozdíl dvou skupin – průměrů).
- Často používány špatně! – konzultujte s Rabušic a kol. (2019).

Neumí (Rabušic a kol., 2019: 23):

- Říct, zda je výsledek prakticky či vědecky důležitý.
- Testovat teoretické (odvozeny z teorie) hypotézy - pouze ty statistické (zobecňují výsledky z repre. Výběrového souboru na populaci) – ty testujeme JEN z pravděpodobnostního vzorku.

Na vše ostatní je tu deskriptivní statistika

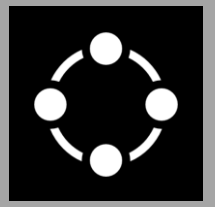
Popis vzorku.

Pozor! Asociace a korelace (vyšší míry deskripce) nelze použít na vysvětlení (explanaci):

Nemohou odpovědět na otázku proč – ta se týká příčin!

Statistika v rámci přístupů k účelům výzkumu

Casula a kol. (2020): Explorace, deskripce, explanace



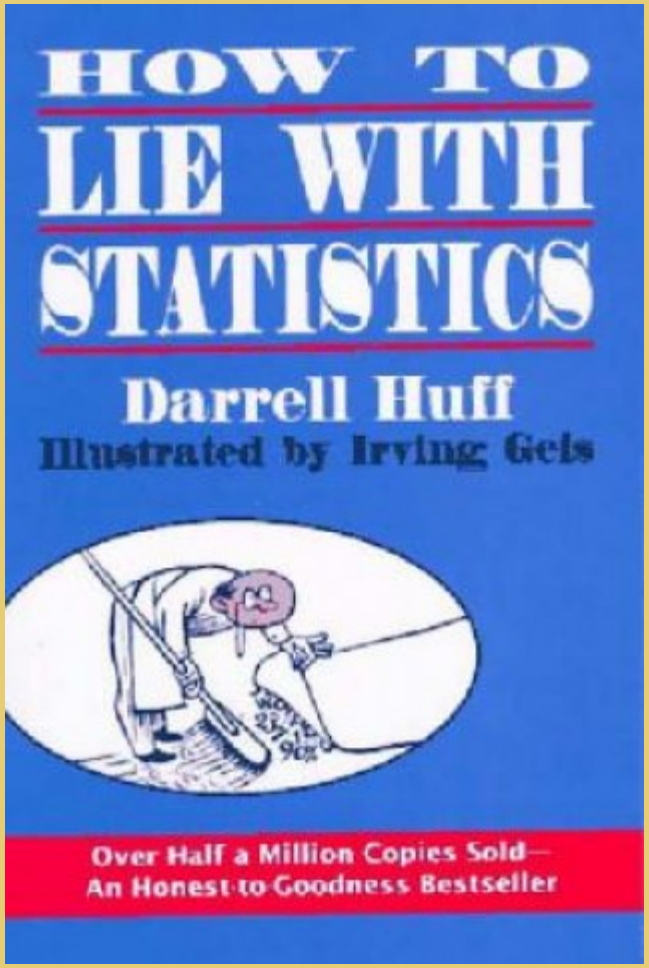
Rabušic a kol. (2019): Popis fenoménů (deskripce) → jejich vysvětlení (explanace) skrze nalezení pravděpodobnostních a kauzálních vztahů → predikce

Umí někdo statistiku?

- Program StatCheck replikuje analýzu, řada nesrovnalostí a chyb v polovině z testovaných 30 tis. psychologických článků (Chawla, 2017).
- Haller a Kraus (2002): Ne (Oaksův test a p-hodnota)

Klamání statistikou (viz např. Magnellová a Van Loon, 2010: 75)

- Je to „tupý“ nástroj. → Garbage in, garbage out.
- Např. průměrný měsíční příjem (cca 34 tis. CZK) vs. medián - cca 29 tis. CZK (ČSÚ, 2020).
- →Hraje zde důležitou roli etika!
- Problém durifikace dat – vnímáme chybně jako naprosto přesná čísla a bezmezně jim věříme.



Etika a statistika

- Lze zde poměrně jednoduše podvádět, ale stejně jednoduše na to zkušenější statistik přijde!
- HARKing, p-hacking, cherrypicking, selective omission.
- Reportovat tak, jak se má reportovat!
- Uchovávejte raw data – pro případ nutného přezkumu a veďte si důkladné poznámky!

6 principů vědecké metody (hypoteticko-deduktivní přístup)

1. Empiricky testovatelné (pozorování, data apod.).
2. Replikovatelné
3. Objektivní (intersubjektivní)
4. Transparentní
5. Falsifikovatelné (Karl Popper)
6. Logicky konzistentní

(Kvantitativní) výzkumný proces: jakou roli v něm zastává statistika?

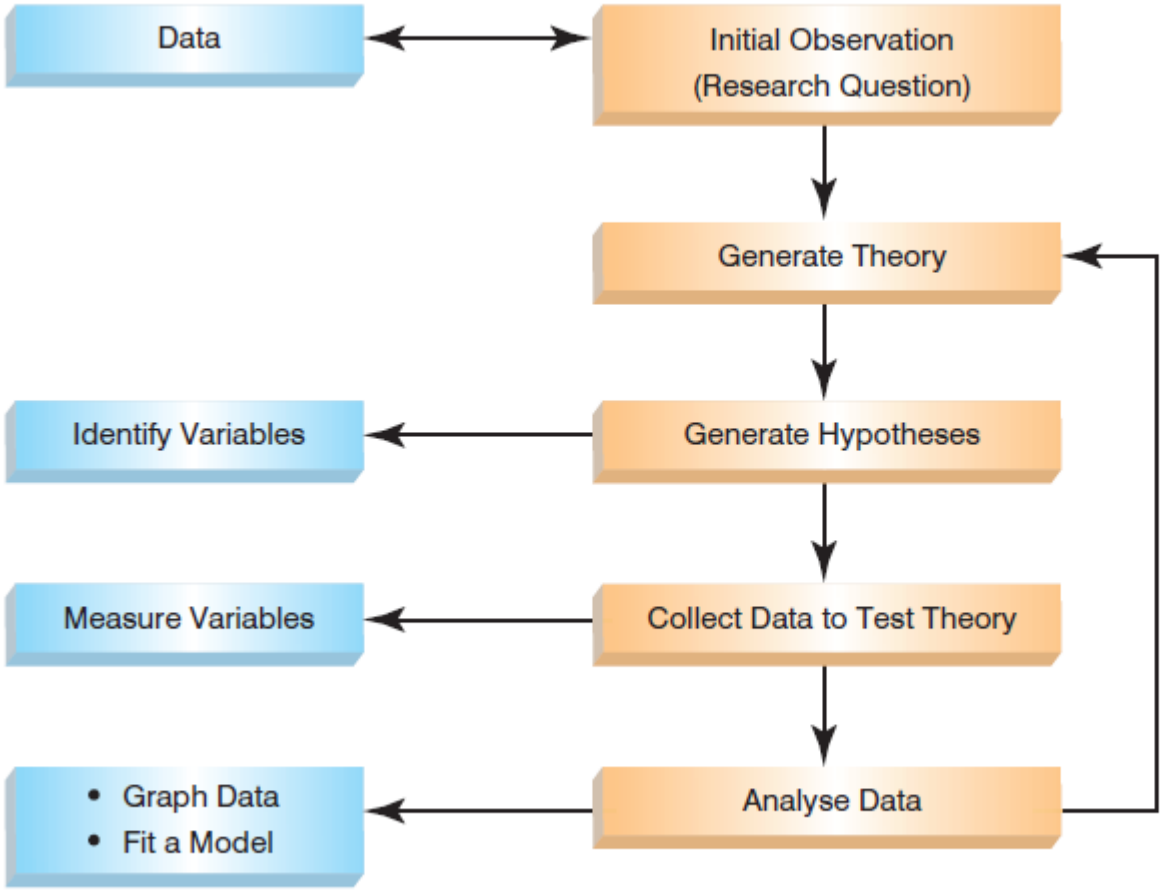
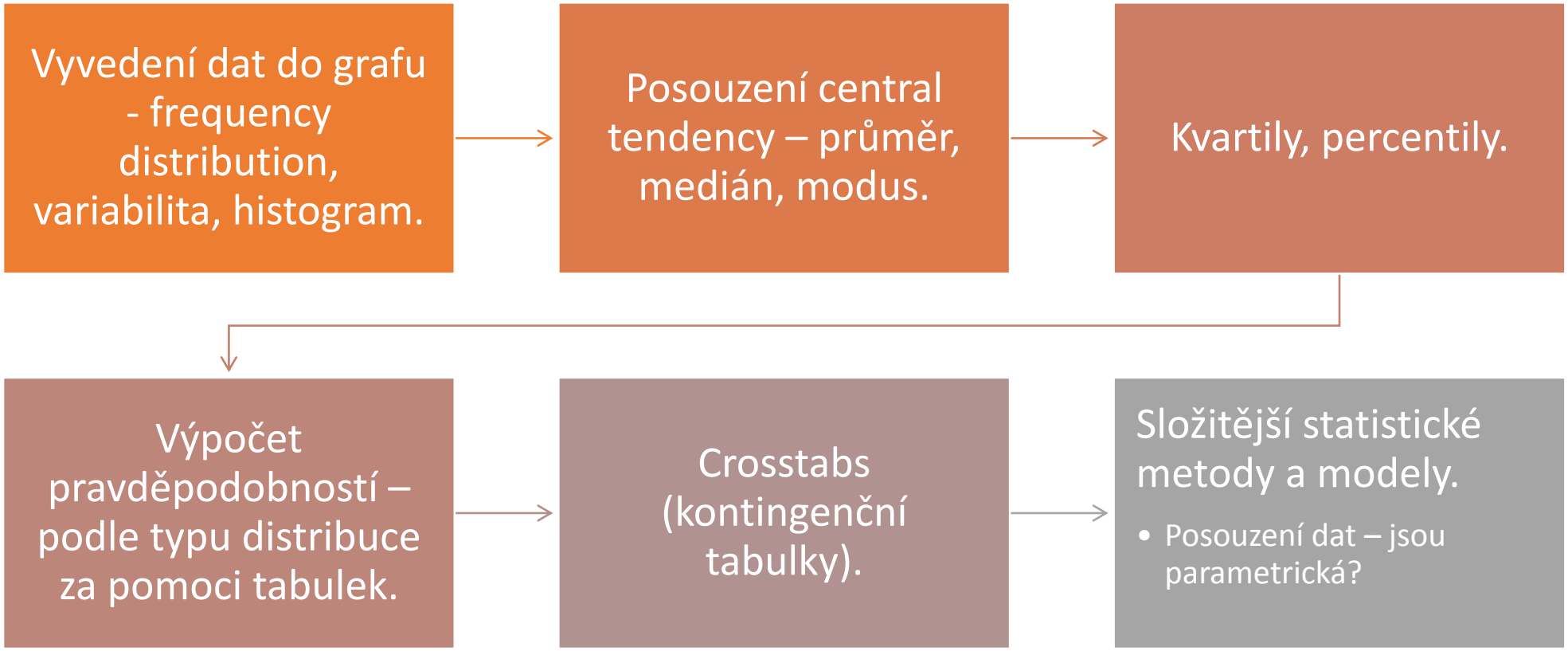


FIGURE 1.2
The research process

Zdroj: Field (2009: 3)

- Pracujeme s hromadnými daty, kterým přiřazujeme numerickou hodnotu (nominální/ordinální/kardinální proměnné).
- Ta jsme získali na základě designu odvíjejícího se od výzkumné otázky.
- Ta také určuje, co sledovat, jaké vlastnosti měřit atd.

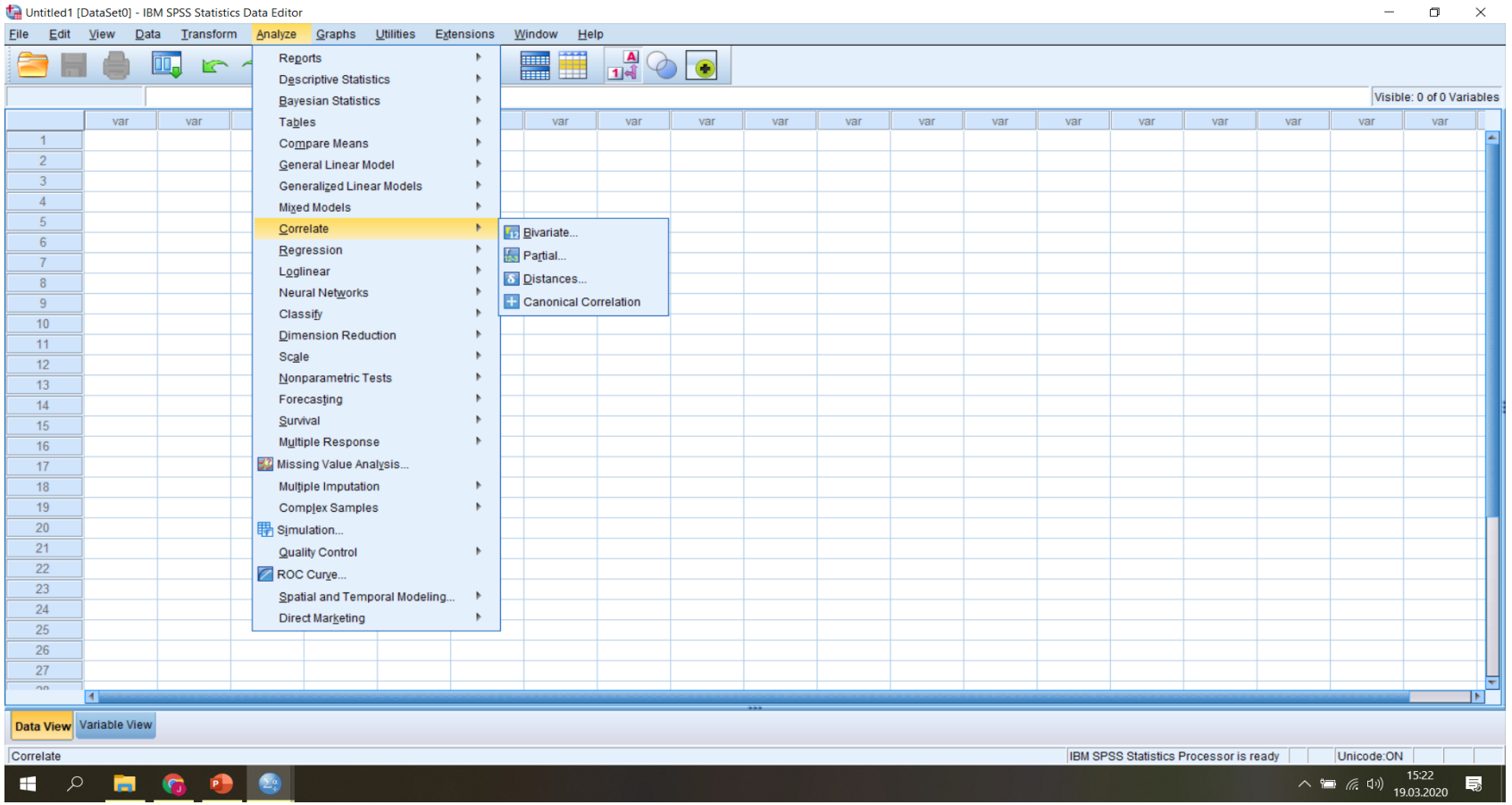
Analýza dat: první krůčky (Field, 2009: 1-30)



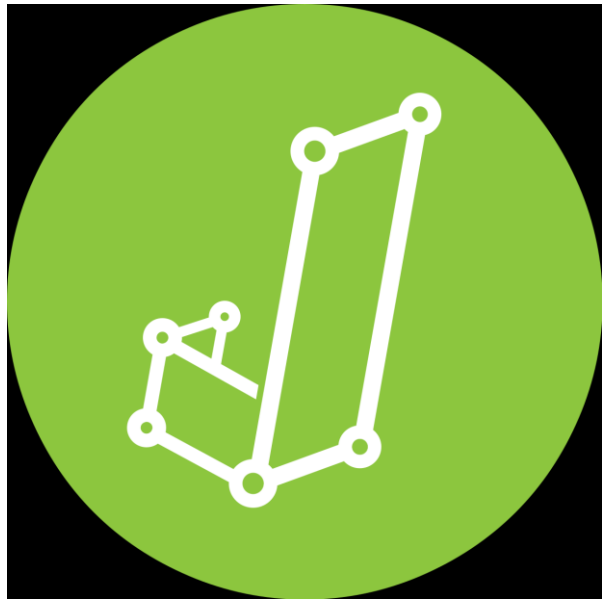
Statistický software

- SPSS, JASP, MS Excel, Rko

SPS
S



- Je statistický program od firmy IBM. Masarykova univerzita na něj má licenci a naleznete jej v aplikaci Inet (jako MS Office).
- Umožňuje tvorbu grafů, tabulek, histogramů, diagramů, scatterplotů aj.
- Počítá veškeré statistické výpočty k modelům – regresní analýza, korelace, kontingenční tabulky apod. a vyhazuje výsledky ve formě grafů, tabulek aj.
- Možnost exportu Excelových dat (jednoduchý přenos z dotazníku Google).

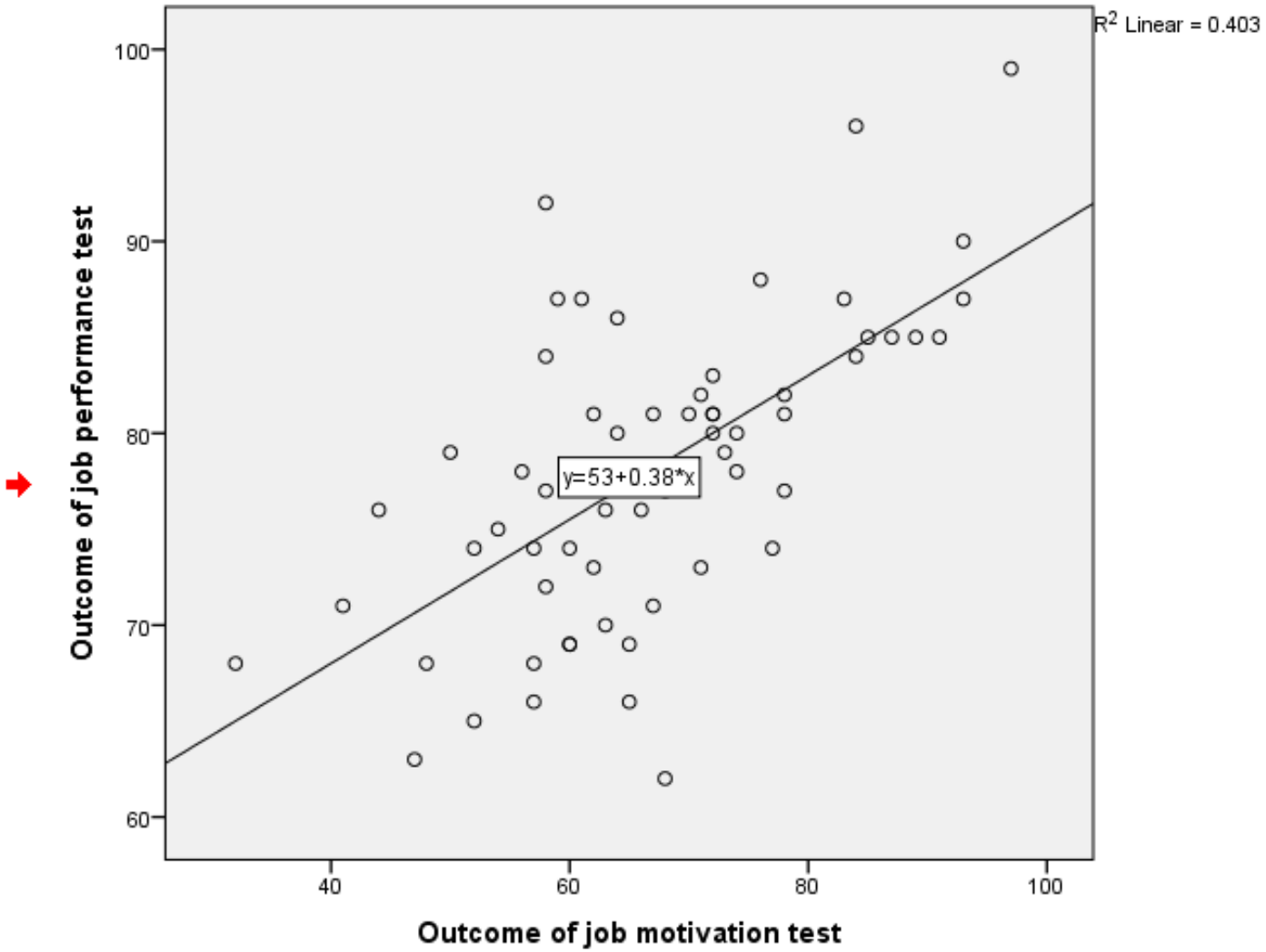


- JASP:
  - open source;
  - okamžitá odezva;
  - přehledné uživatelské prostředí;
  - Bayesiánská statistika;
  - běží na bázi Rka.

- Rko:
  - open source programovací jazyk;
  - uděláte v něm téměř vše;
  - složité;
  - nejrozšířenější;
  - R studio, R console.
  - Předpřipravený skript = kontinuální a cyklická analýza.

Příklady statistických operací a modelů

- T-test (parametrický).
- Mann-Whitney U test (neparametrický).
- OLS (lineární) regrese.



Checklist pro dotazník (Rumsey, 2010: 137-146)

- Garbage in, garbage out

1. Cílová populace je dobře definovaná.
2. Vzorek odpovídá cílové populaci;
3. a je náhodný;
4. a dostatečně velký (margin of error).
5. Non-response je minimalizovaná.
6. Typ dotazníku odpovídá potřebným datům.
7. Otázky jsou dobře strukturované a položené.
8. Správné načasování.
9. Personál je dobře trénovaný.
10. Na základě výsledků vytváříme adekvátní závěry.

Nejčastější statistické chyby (Rumsey, 2010: 155-162)

1. Zavádějící grafy.
2. Biased data.
3. Neuvedený margin of error (míra chyby inference na populaci).
4. Nenáhodný vzorek.
5. Neuvedená velikost vzorku.
6. Špatně interpretované korelace.
7. Intervenující proměnné.
8. Špatně uvedená čísla.
9. Selektivní reportování dat.
10. The Almighty Anecdote (sample size one).

Proměnné I

- Vlastnosti zkoumaných subjektů často nelze přímo pozorovat, musíme se spokojit s jejich indikátory (vzdělanost – dosažené vzdělání) → **operacionalizace** je „převodem abstraktních konstruktů do měřitelných znaků“ (Rabušic a kol., 2019: 18) → proměnných.
- Někdy proměnná jako méně abstraktní, ale ne ještě měřitelný jev – např. politická síla (konstrukt) → vliv v parlamentu (proměnná) → OP.: množství protlačených zákonů/léta v parlamentu/hodnocení veřejnosti apod. (měřitelné proměnné).
- Zásadní je zde konstruktová validita – jsou měřeny co nejpřesněji?

Proměnné II

- *Variables of interest*
  - Závislá (*dependent, outcome*)
  - Nezávislá (*independent, predictor)*
- *Variables of disinterest*
  - zavádějící proměnná – *confounder* - **zahrnuta** ve studii, ale nelze odlišit efekt od ostatních proměnných (např. cvičení a dieta – co z toho má vliv na → ztrátu hmotnosti?)
  - skrytá proměnná – *lurking variable* – **nezahrnutá** proměnná, která ovlivňuje obě proměnné např. prodej zmrzliny + počet utonutí (lurking je zde čas – roční období).
- + kontrolní proměnné (držíme konstantní), *background var.* (pro určování reprezentativnosti vzorku)

Proměnné II – úrovně měření

- *Levels of measurement.*
- Nominální (*nominal,* muž, žena).
- Ordinální (*ordinal,* pořadí v závodě).
- Intervalové (*interval,* žádná smysluplná nula – teplota).
- Poměrové (*ratio,* smysluplný bod nula – věk).

- Spojité (*continuous*) vs. nespojité (*discrete*).
  - Nominální a ordinální nespojité, ale i některé kvantitativní (např. počet mazlíčků).
  - Spojité jsou věk, hmotnost apod.
  - Odvíjí se od nich např. podoba grafů.

Korelace a kauzalita I (Magnellová a Van Loon, 2010: 117-120)

- Kauzalita je příčinný vztah mezi proměnnými, zatímco korelace znamená pouze to, že spolu dvě proměnné nějakým způsobem souvisí.
- K měření korelace se nejčastěji používá např. Pearsonův korelační koeficient (značí se R nebo r).
- Korelaci je nutné věcně vykládat. Existuje něco, co Pearson označuje jako „spurious correlations“ (zdánlivé korelace).

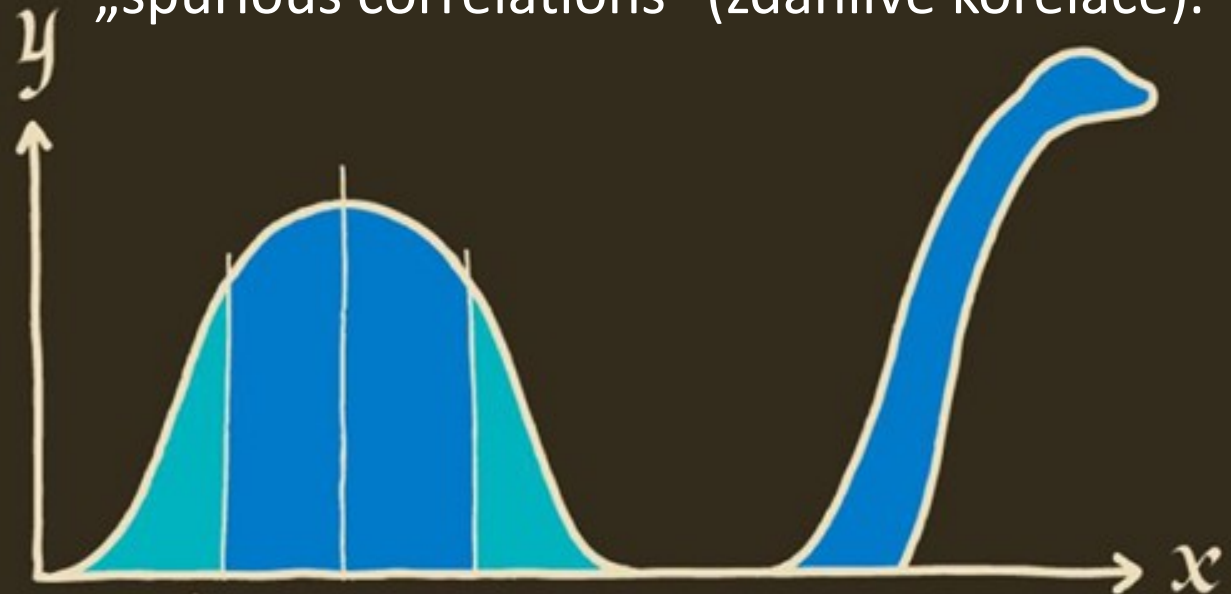


fig 1.0 The Extended Bell Curve.

- Příklady zdánlivé korelace (https://www.tylervigen.com/spurious-correlations)

  - G. Yule (1899): „Asociace“ – vztah mezi 2 a více nespojitými proměnnými

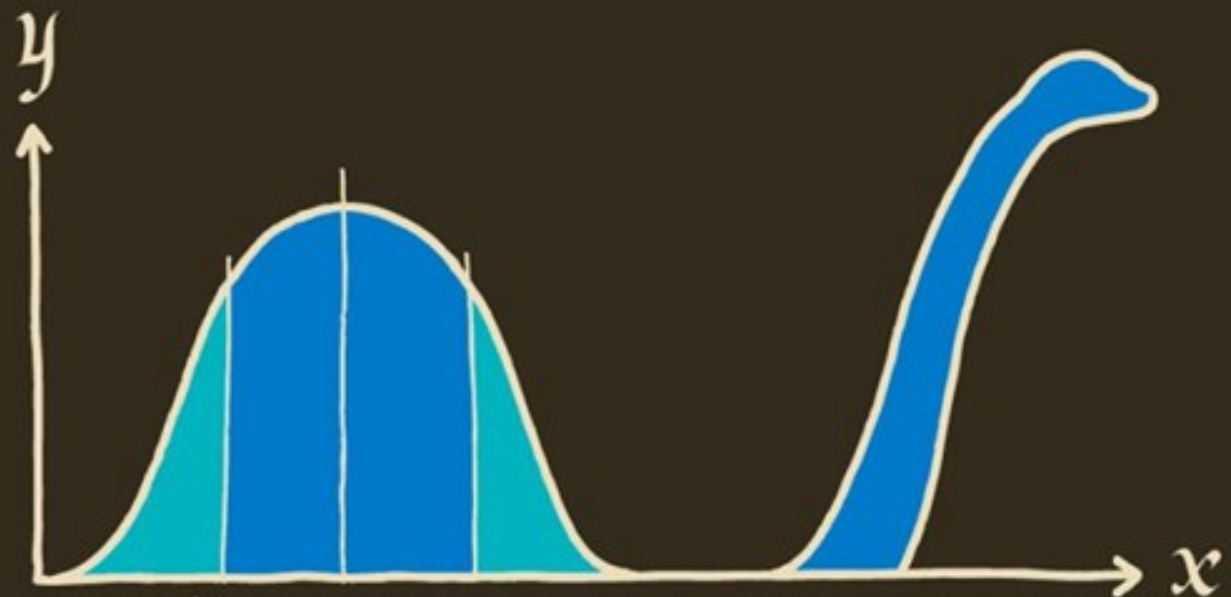


fig 1.0 The Extended Bell Curve.

Korelace a kauzalita II (Field, 2009:1-26)

- Kauzalita podle Humea (1748):
  - 1) Příčina a následek musí proběhnout časově blízko sebe.
  - 2) příčina musí proběhnout před následkem.
  - 3) Daný efekt nemůže proběhnout bez přítomnosti příčiny.
  - + J. Mill (1865) 4) všechna ostatní vysvětlení příčinného vztahu jsou vyloučena

  - Statistické podmínky pro regresi: lineární vzor(scatterplot) v datech a korelace (střední až silná).
- →Korelace neimplikuje kauzalitu, ale kauzalita korelaci potřebuje!
- 2 základná typy studií pro testování hypotéz:
  - Observační (korelační) – výsledkem je, že spolu dvě proměnné korelují.
  - Experimentální – může prokázat cause-and-effect relationship.
- Prokazování korelace a kauzality se neváže ke konkrétním statistickým postupům, ale výzkumnému designu (korelační/observační vs. experimentální)!

Validita a reliabilita

- Reliabilita spjatá s replikovatelností.
- Validita – akurátně vystihuje realitu.
- Reliabilita se bez validity obejde, ale validita bez reliability ne – pročpak? ☺

Typy validity

- Konstruktová *(construct, measurement)* – konstrukty jsou měřeny akurátně – klíčová role operacionalizace.
- Interní (vnitřní, *internal*) - hypotetická příčina skutečně zapříčiňuje pozorovaný efekt.
- Externí (vnější, *external*) – efekt se opakuje i u jiných skupin, časů – zobecnitelnost.
- Konvergentní

Typy reliability

- *Measurement*
  - *Test-retest* – 2 testy po sobě u jednoho subjektu – měly by být víceméně konzistentní.
  - *Internal consistency* – možno měřit korelaci otázek.
  - *Intra* a *inter observer/rater consistency* – jeden pozorovatel, 2x s odstupem času/ dva pozorovatelé
  - *Parallel forms* – různé verze testu, které mají být ekvivalentní.

Sampling (vzorkování) a populace

- Pro inferenční statistiku nezbytný!
- Populace → vzorek → element.
- Pravděpodobnostní je zlatý standard.
  - Prostý náhodný výběr.
    - Každý element má stejnou šanci dostat se do vzorku.
  - Prostý systematický výběr.
    - Např. 1. element vybrán náhodně a pak každý 5.
  - Komplexní stratifikovaný náhodný výběr.
    - Populace rozdělena do vyčerpávajících a výlučných strat a z nich je tvořen náhodným výběrem vzorek.
  - Kompl. *Multi-stage cluster sampling*
    - Populace je v několika fázích clusterována (např. kraj → okres → škola → třída → náhodný výběr).

Sampling (vzorkování) a populace

- Nepravděpodobnostní ale také nejsou k zahození.
  - *Convenient*.
    - Nejvýhodnější elementy, nenáročné, obrovský bias.
  - *Snowball*.
    - Typ *convinient samplingu*, ale relativně dobrý.
  - *Purpose*.
    - Hlavně kvali, na základě úsudku výzkumníka.
  - *Quota*.
    - Approximace vzorku na populaci na základě charakteristik → elementy do kategorií vybírány *convenient samp*.

Kde hledat data k analýze?

- Tipy viz Pennings, Keman a Kleinnijenhuis (2006: 56-60).
- Ucelené statistické soubory od státních i nestátních institucí (např. Český statistický úřad apod.).
- Vlastní sběr.
- Google Trends, Google AdWords apod.
- Různé databáze (některé jsou neplacené, některé placené).
- EU, ČR apod. – open data a open science iniciativy a datasety.
  - Eurobarometry.
- Facebook, Twitter apod.
- Supplementary data files u některých článků.

A co dál?

- Tímto to bohužel zdaleka nekončí.

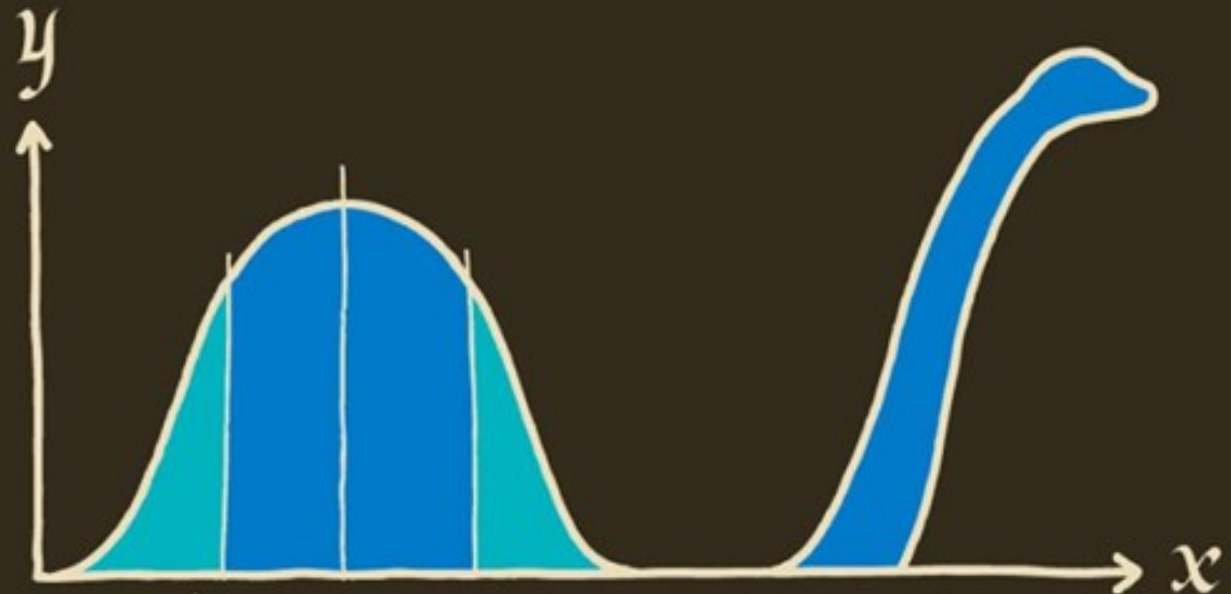


fig 1.0 The Extended Bell Curve.

Neprobrali jsme, ale je důležité:

- Hrozby validitě a reliabilitě a jak mitigovat jejich riziko.
- Experimentální, kvazi-experimentální a korelační designy výzkumu a jejich praktické provádění.
- Typy distribucí dat a jejich implikace.
- Potřeba nastudovat v případě potřeby!

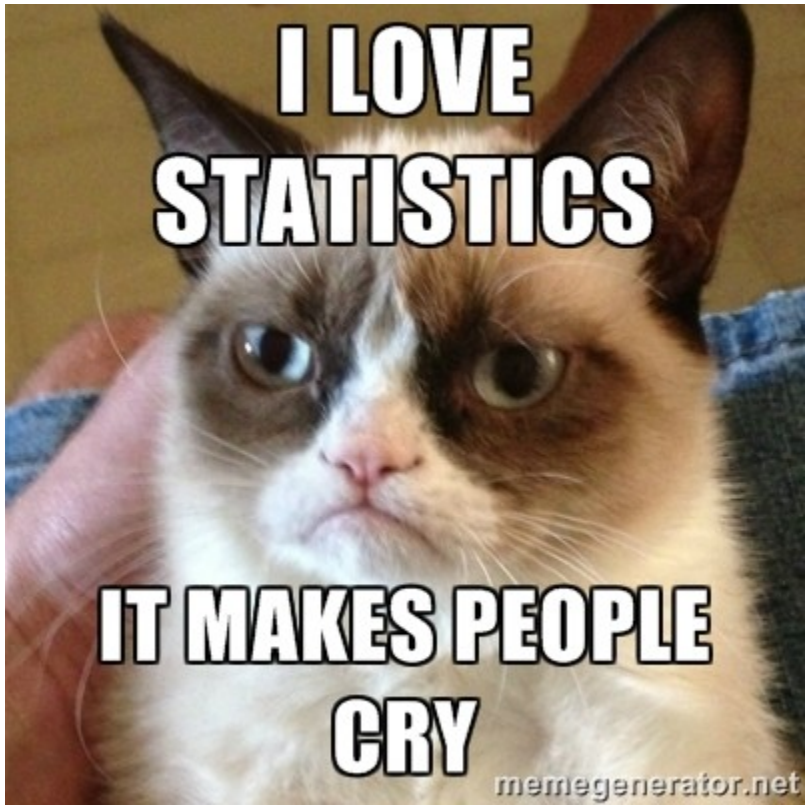
A co dál?

- Podzimní kurz „Kvantitativní přístupy v politologii“ – praktická aplikace statistiky v programu SPSS s doc. Spáčem a doc. Pinkem.
- Kniha **Seznamte se, statistika** od Van Loona a Magnellové (2009).
- Studium Big Data – aplikace statistiky na obrovské množství dat např. z vyhledávání Googlu (aplikace a Trends) → možnost zajímavých výzkumných výsledků a směřování. Ideální vstupní branou je kniha **Everybody Lies** (2017) od S. S. Davidowitze (od něj je na internetu i spousta zajímavých článků).
- Kurzy University of Amsterdam (Coursera).
- „Youtuber“ Petr Soukup aj.
- Další knihy a články v závislosti na konkrétních problémech.

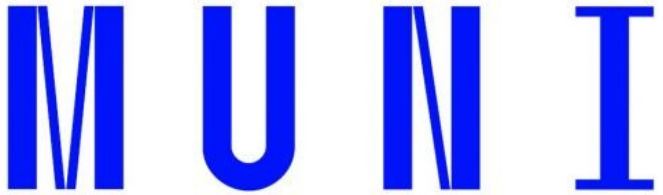
Reference

- Pennings, Paul; Keman, Hans a Kleinnijenhuis, Jan. (2006): *Doing Research in Political Science: An Introduction of Comparative Methods and Statistics*. 2nd Edition. Sage Publications, ISBN 978-1-4129-0377-6.
- Field, Andy (2009): *Discovering Statistics Using SPSS*. 3rd Edition. Sage Publications: London, ISBN 978-1-8478-7907-3.
- Davidowitz, S. S. (2017). Everybody Lies: Big Data, New Data, and What the Internet Can Tell Us About Who We Really Are. New York: Day Street Books.
- Rumsey, Deborah J. (2010): *Statistics Essentials for dummies*. Indianapolis: Wiley Publishing, Inc. ISBN 978-0-470-61839-4
- Magnello, Eileen a Van Loon, Borin. (2010). *Seznamte se, statistika.* Praha: Portál. ISBN 978-80-7367-753-4.
- Český statistický úřad. (2020). *Průměrné mzdy - 2. čtvrtletí 2019.* Dostupné z: https://www.czso.cz/csu/czso/cri/prumerne-mzdy-2-ctvrtleti-2019.
- Chawla, Dalmeet S. (2017). Controversial software is proving surprisingly accurate at spotting errors in psychology papers. *Science.* Dostupné z: https://www.sciencemag.org/news/2017/11/controversial-software-proving-surprisingly-accurate-spotting-errors-psychology-papers.
- Haller, H., & Kraus, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research, 7*(1), 1–20.
- Rabušic, Ladislav; Soukup, Petr; Mareš, Petr. 2019. *Statistická analýza sociálněvědních dat (prostřednictvím SPSS).* Brno: Masarykova univerzita, s. 11-13; 17-49.

Děkuji za pozornost!



Cvičení: Testování hypotéz a vybrané statistické operace v Rku



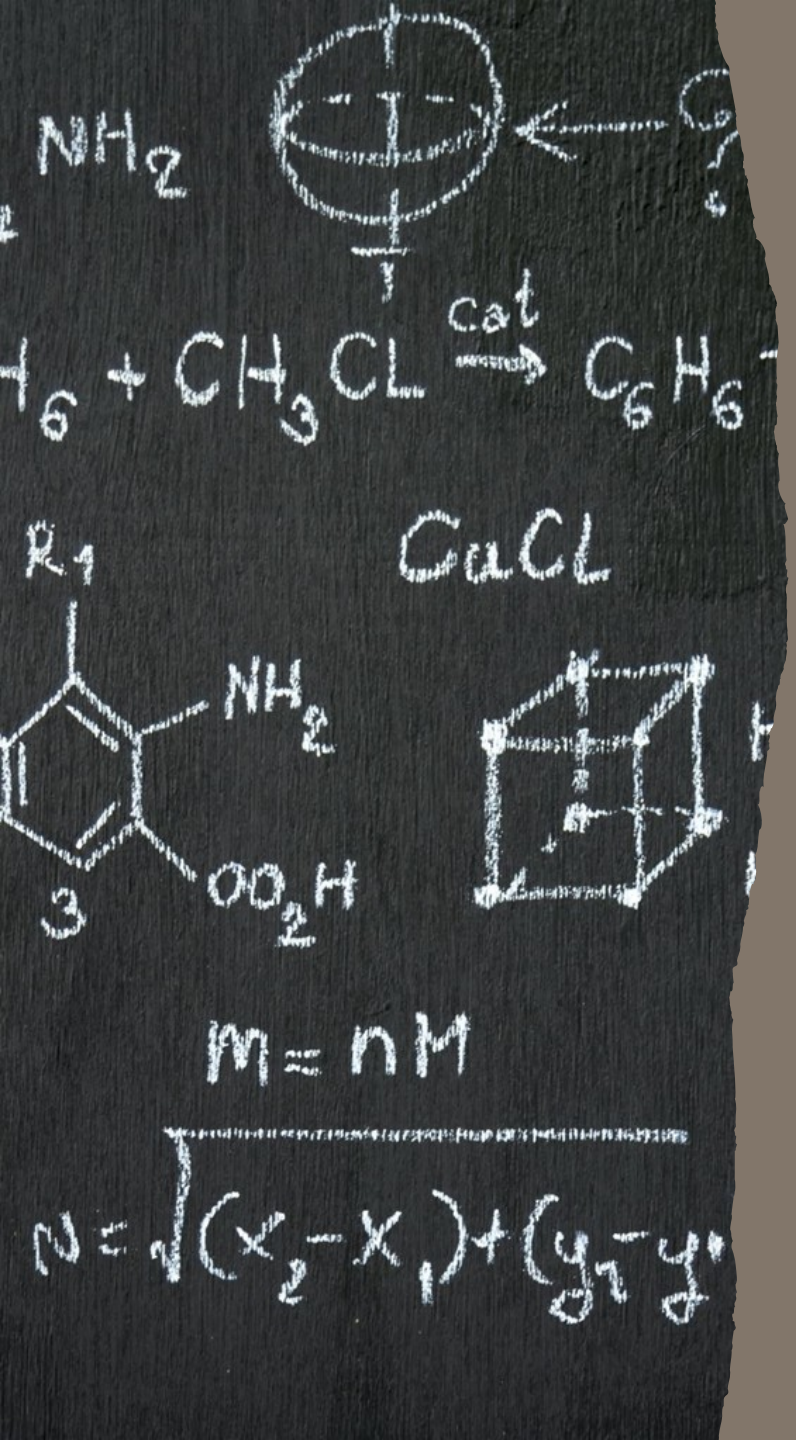
Přídavek nad rámec základů statistiky

Jan Kleiner

Online přednáška 3. 4. 2024

Testování hypotéz I

- Rozhodnutí, zda dostupná data dostatečně podporují konkrétní hypotézu.
- + inferenční statistika: existuje dostatečný důkaz ve vzorku dat, který by umožnil usuzovat, že závěry platí pro celou populaci?
- **<u>Typy chyb</u>**
  - **Chyba 1. typu**: Zamítnutí nulové hypotézy, když je pravdivá (falešně pozitivní).
  - **Chyba 2. typu**: Nezamítnutí nulové hypotézy, když je nepravdivá (falešně negativní).
- Vždy zvažujte kontext, velikost efektu a praktickou významnost, nejen p-hodnotu.



Testování hypotéz II

- **Nulová hypotéza (H_0):** Status quo; tvrzení, že není žádný efekt nebo rozdíl.
- **Alternativní hypotéza (H_1):** To, co se snažíme dokázat; tvrzení, že existuje efekt nebo rozdíl.
- **P-hodnota**: Pravděpodobnost pozorování dat, nebo něčeho extrémnějšího, pokud je nulová hypotéza pravdivá.
- **Hranice významnosti (α):** Práh pro významnost, obvykle stanovený na 0.05.
- Pokud p-hodnota ≤ α: Zamítnout H_0 (důkazy jsou proti H_0 a ve prospěch H_1).
- Pokud p-hodnota > α: Nezamítnout H_0 (nedostatek důkazů na podporu H_1).

Dělení statistických testů – parametrické vs. neparametrické

Vybrané parametrické testy (operace)

- 1. T-test – one-sample (dependent), independent
  - Rozdíl průměrů u jedné (změna v rámci experimentu typu within-subject design), nebo u dvou skupin (rozdíl mezi skupinami – např. muži vs. ženy).
- 2. ANOVA – one-way/repeated measures – průměry 3 a více skupiny, nebo repeated measures experimental design (v rámci jednoho člověka/elementu).
- 3. OLS Regrese (lineární), logistická regrese (2 skupiny), multinomiální regrese (závislá proměnná více jak 2 skupiny)
- 4. Korelace (Pearson´s r, Spearman´s rho, Kendall´s tau).

Vybrané neparametrické testy (operace)

- Independent t-test → Mann-Whitney U Test (Wilcoxon Rank-Sum Test).
- Dependent t-test → Wilcoxon Signed-Rank Test.
- One –way ANOVA → Kruskal-Wallis H Test.
- Spearman´s Rank Correlation → Pearson´s Correlation.
- Atd. – viz Field (2009, p. 822).