Základy kvantitativní analýzy dat

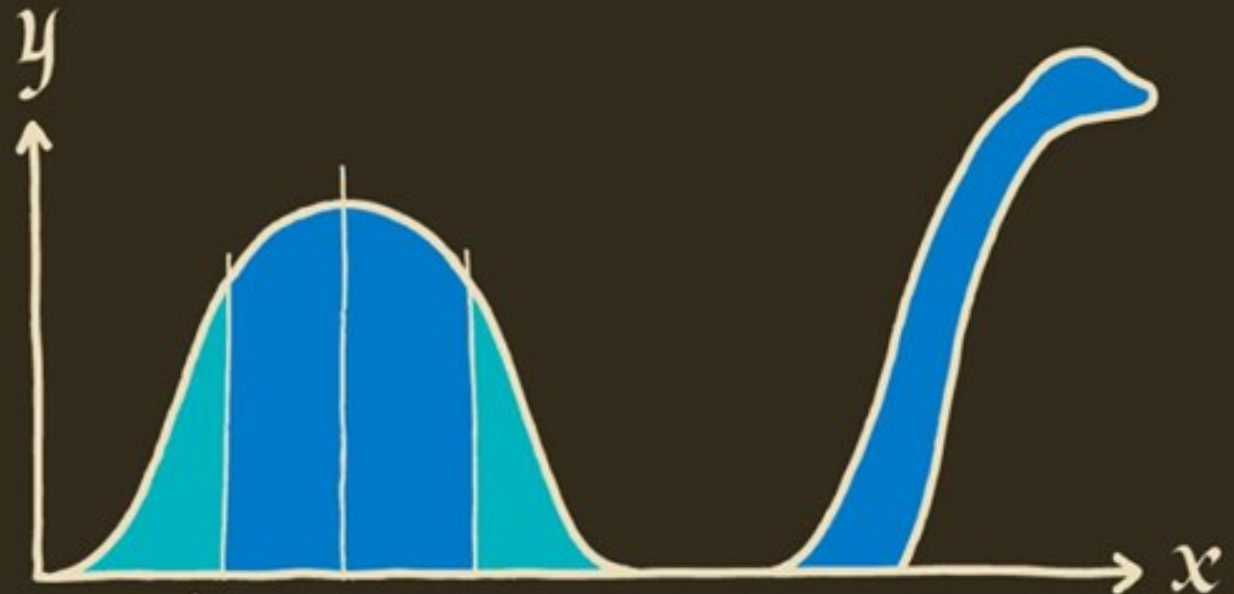


fig 1.0 The Extended Bell Curve.

Jan Kleiner
21. 4. 2021
BSSn4405
jkleiner@mail.muni.cz

Statistika v sociálněvědním výzkumu

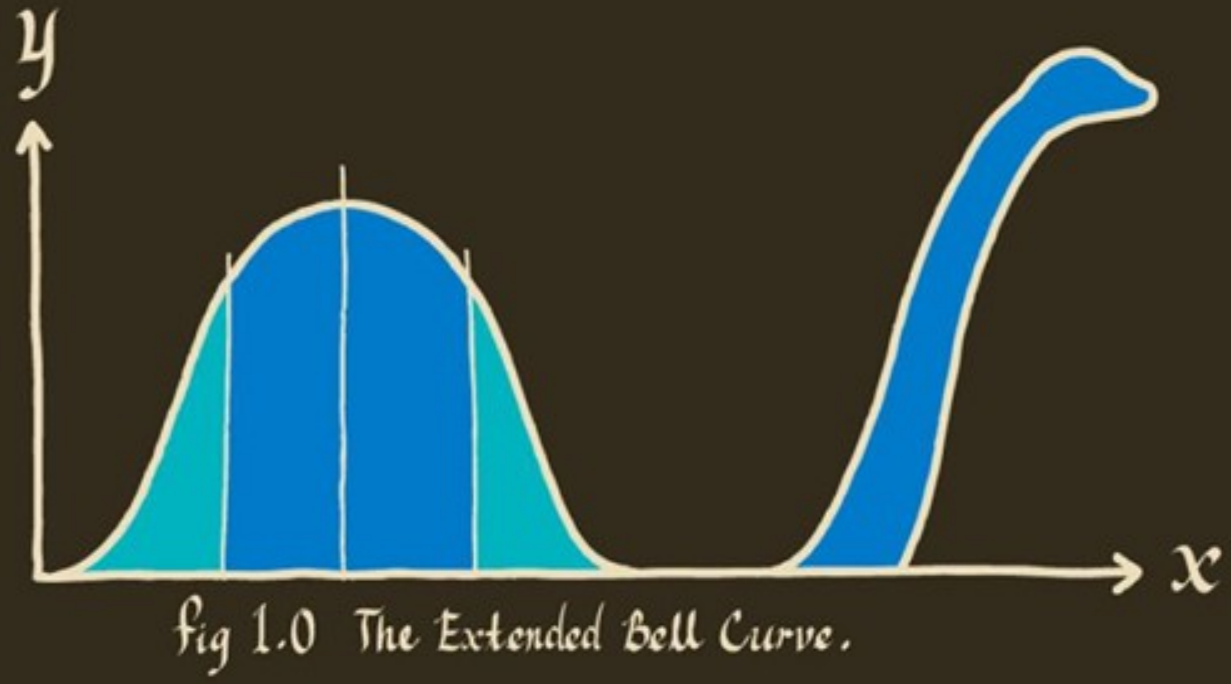


fig 1.0 The Extended Bell Curve.

Jan Kleiner
21. 4. 2021
BSSn4405
jkleiner@mail.muni.cz

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;
- získáte nové guilty pleasure;
- budete vědět, kam dál se studiem statistiky;
- získáte nová výzkumná směřování.

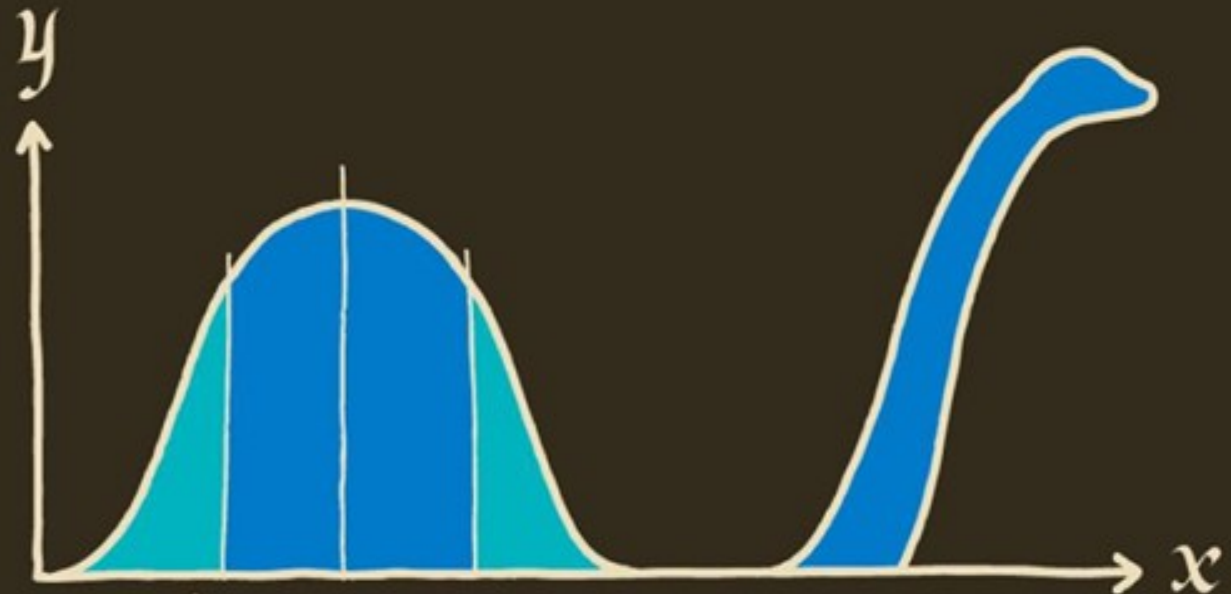
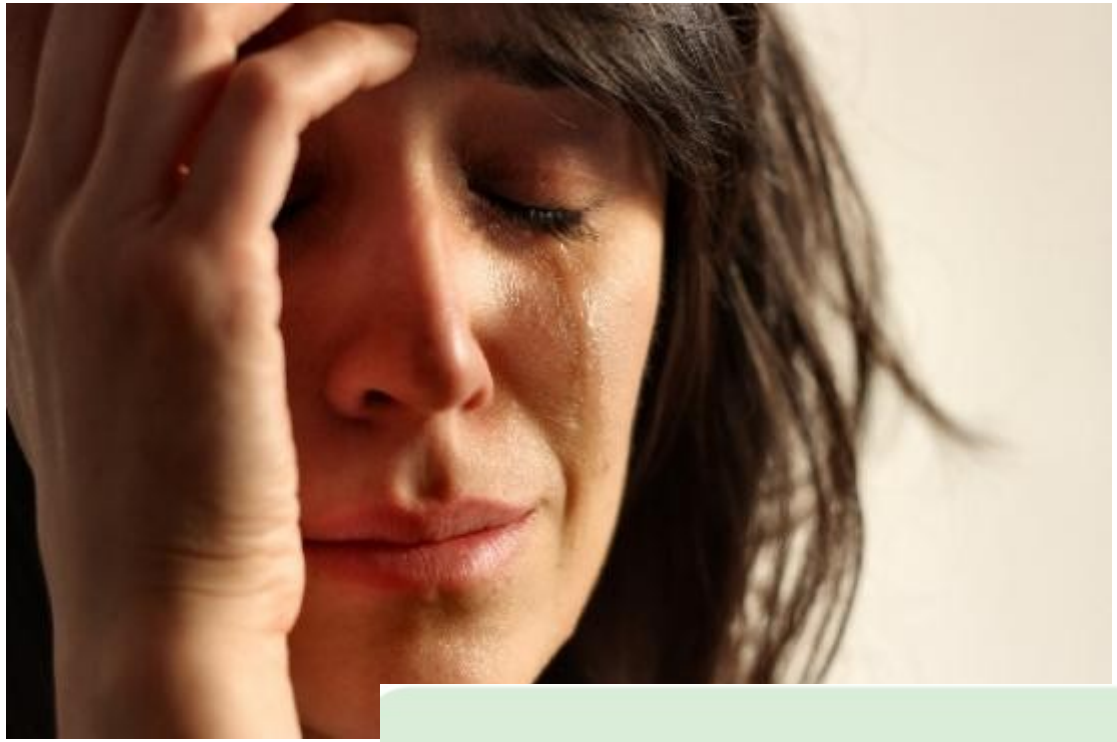


fig 1.0 The Extended Bell Curve.

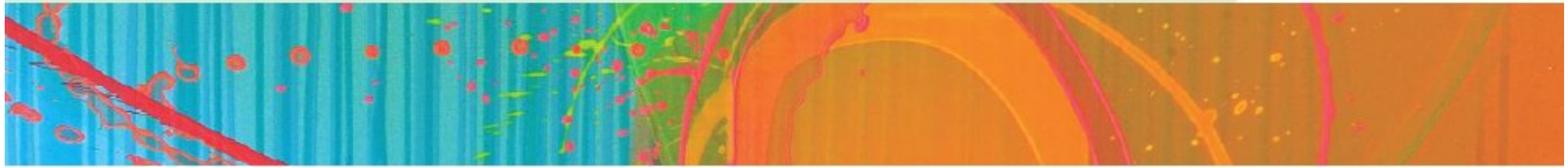
Literatura

- Povinná literatura a prezentace jsou komplementární.
- **Povinná literatura** obsahuje statistické základy.
  - **Andy Field** (2009: 31-60) - základy.
  - **Pennings** a kol. (2005: 55-69) zasazuje statistiku do metodologie politologie.
- Další materiály
  - Magnellová a Van Loon – Seznamte se, statistika
  - Rabušic, Soukup a Mareš – Statistická analýza sociálněvědních dat prostřednictvím SPSS
  - Novotný, Svobodová - Jak pracuje věda?



Proč?
!

Why is my evil lecturer forcing me to learn statistics?



Zdroj: Field (2009)

1

- Pasivní i alespoň základní aktivní znalost statistiky je nezbytná.
- 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.
- Staré způsoby sběru dat jsou mrtvé, nové vyžadují statistiku (kvůli velikosti) (viz např. *Everybody Lies* od S. S. Davidowitze, 2017).
  - Díky tomu navíc získáváme možnost zkoumat velmi zajímavá a neotřelá data (Google, Facebook apod.).
- Jen skok k AI a algoritmům rozhodování.
- Má mě statistika nějak zajímat, i když nechci dělat „vědu“?
  - Co je „věda“? ☺

Výzkum bez špetky statistiky vs. výzkum se znalostí statistiky.

Vyprávíme příběh (interpretujeme) na základě dat. Čím lepší nástroje máme, tím více je ten příběh blíže objektivní skutečnosti, realitě.

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

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!

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!

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?

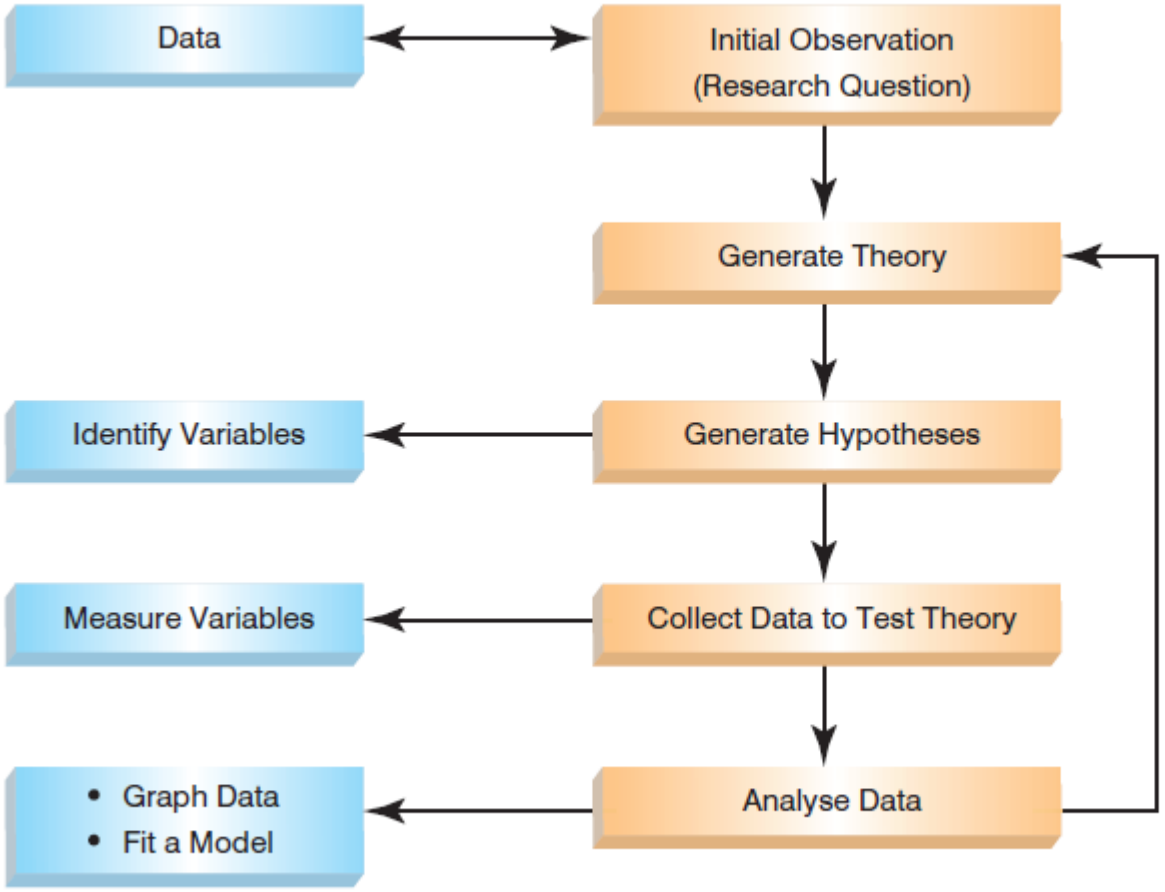


FIGURE 1.2
The research process

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.

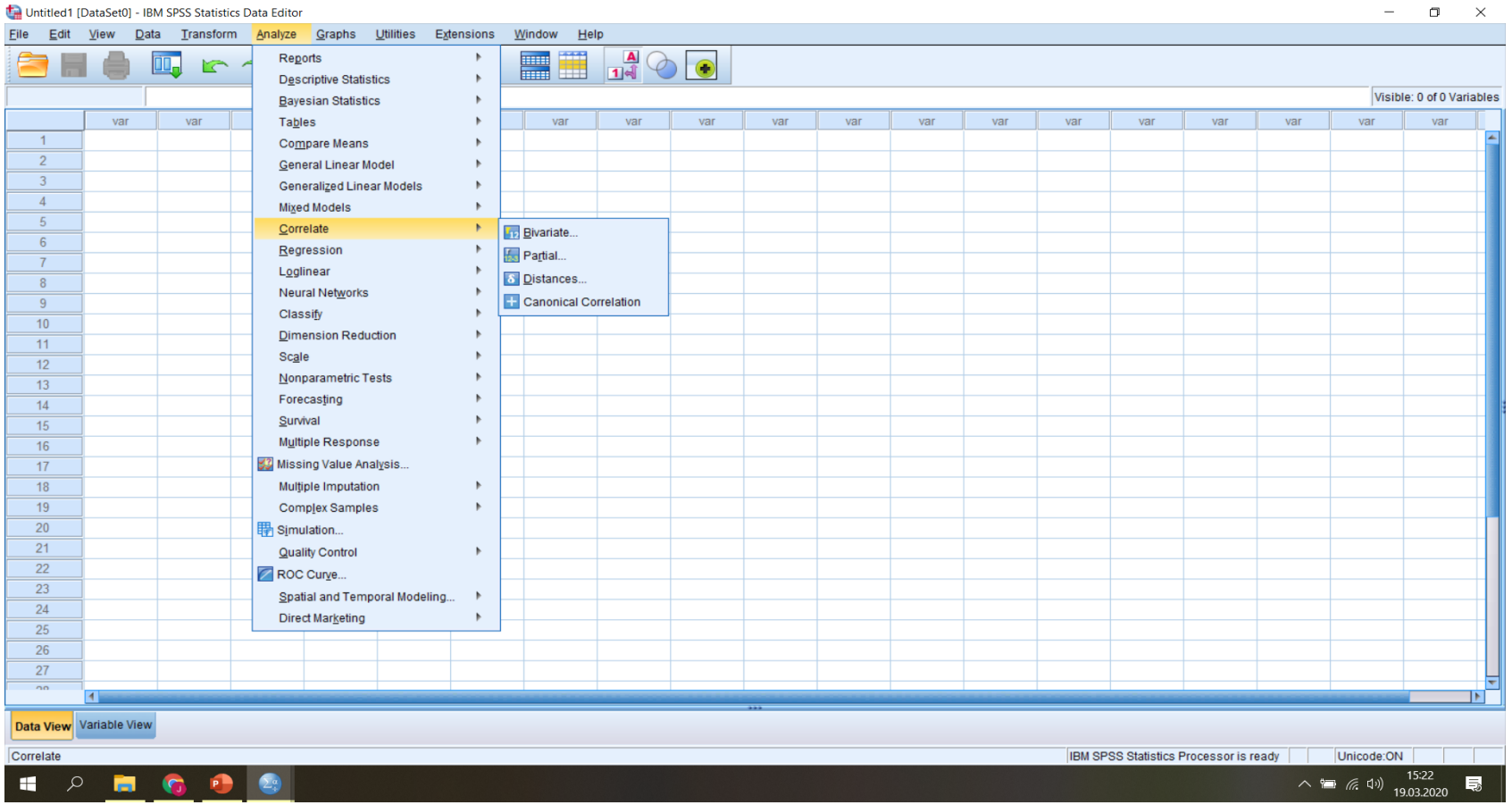
Analýza dat: první krůčky (Field, 2009: 1-30)

- Vyvedení dat do grafu - frequency distribution, variabilita, histogram.
- Posouzení central tendency – průměr, medián, modus.
- Kvartily, percentily.
- Výpočet pravděpodobností – podle typu distribuce za pomoci tabulek.
- Crosstabs (kontingenční tabulky).
- Složitější statistické metody a modely.
  - Posouzení dat – jsou parametrická?

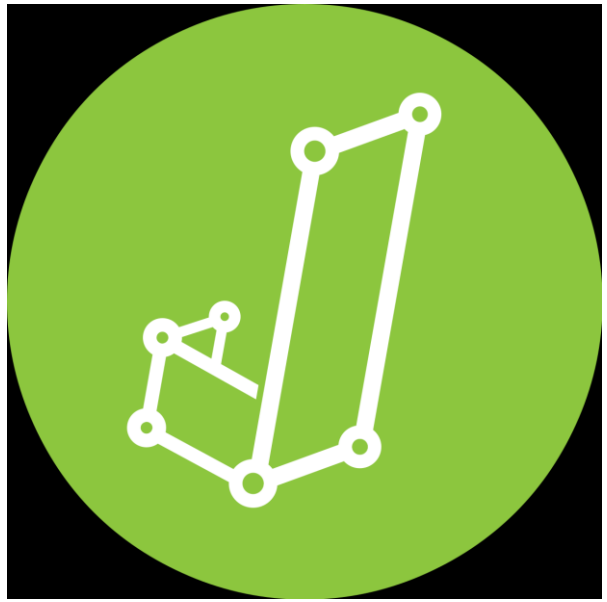
Statistický software

- 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).

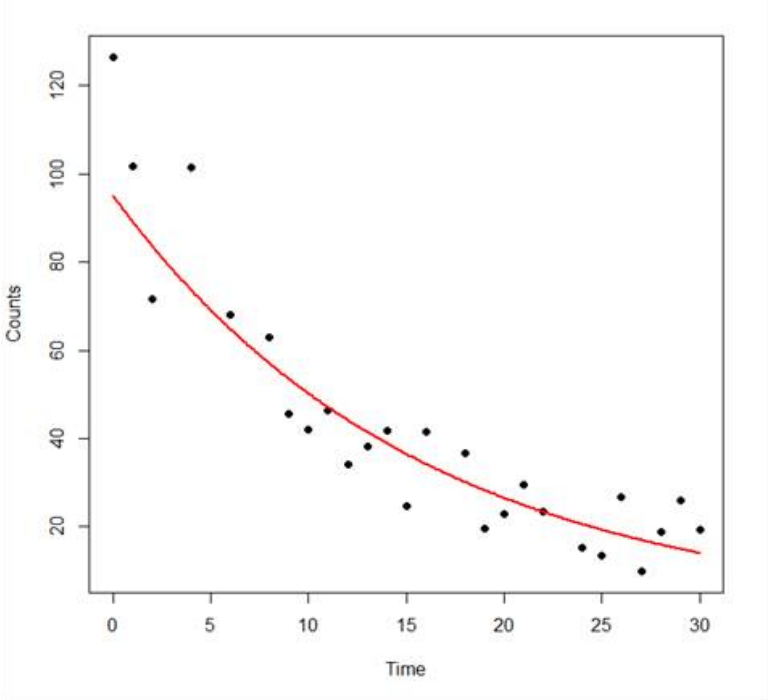


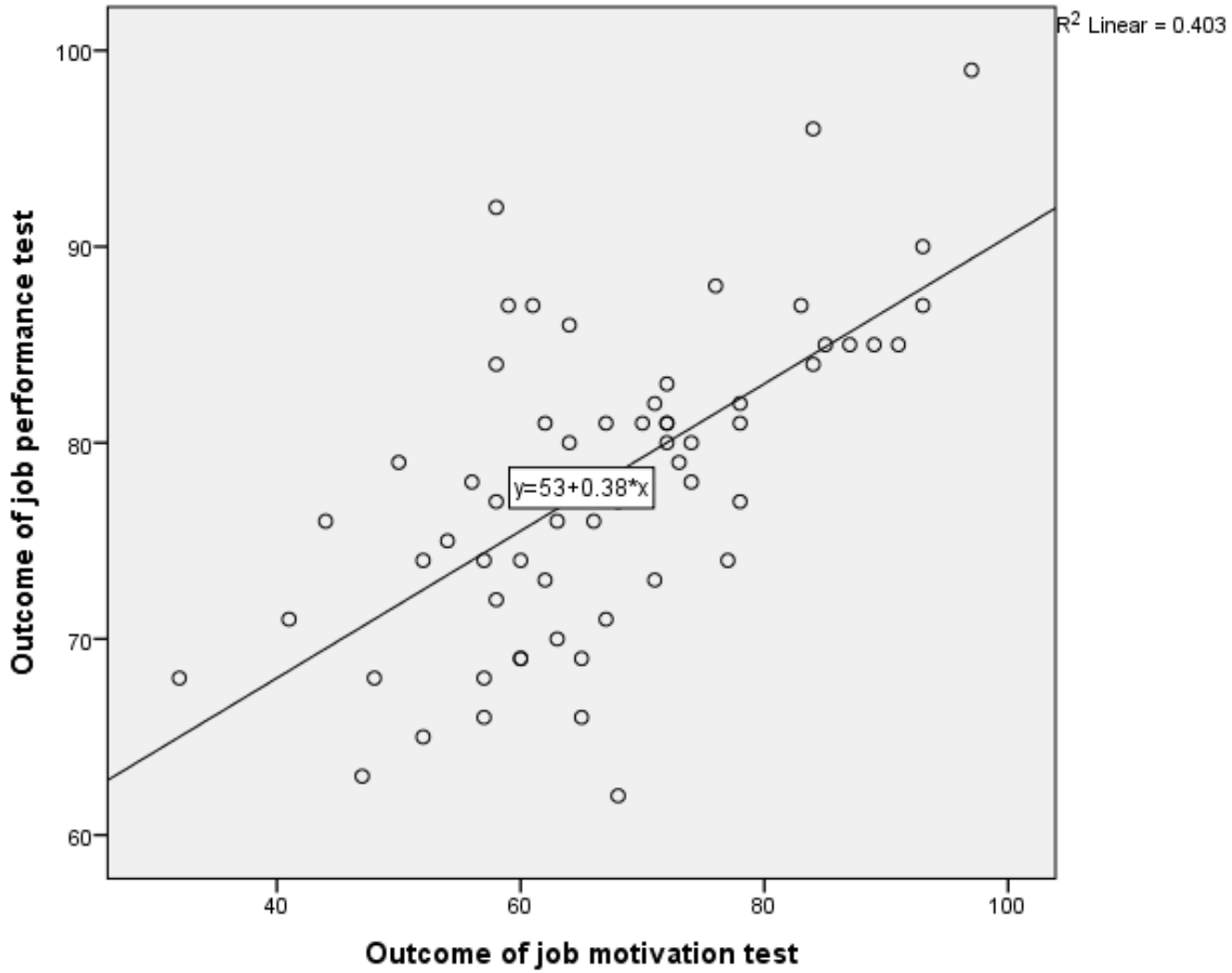
- JASP:
  - open source;
  - okamžitá odezva;
  - přehledné uživatelské prostředí;
  - Bayesiánská statistika;
  - běží na bázi Rka.

- Rko:
  - open source programovací jazyk;
  - uděláte v něm téměř vše;
  - složité;
  - nejrozšířenější;
  - R studie, R console.

Příklady statistických operací a modelů

- T-test (parametrický) + ukázka.
- 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).

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).

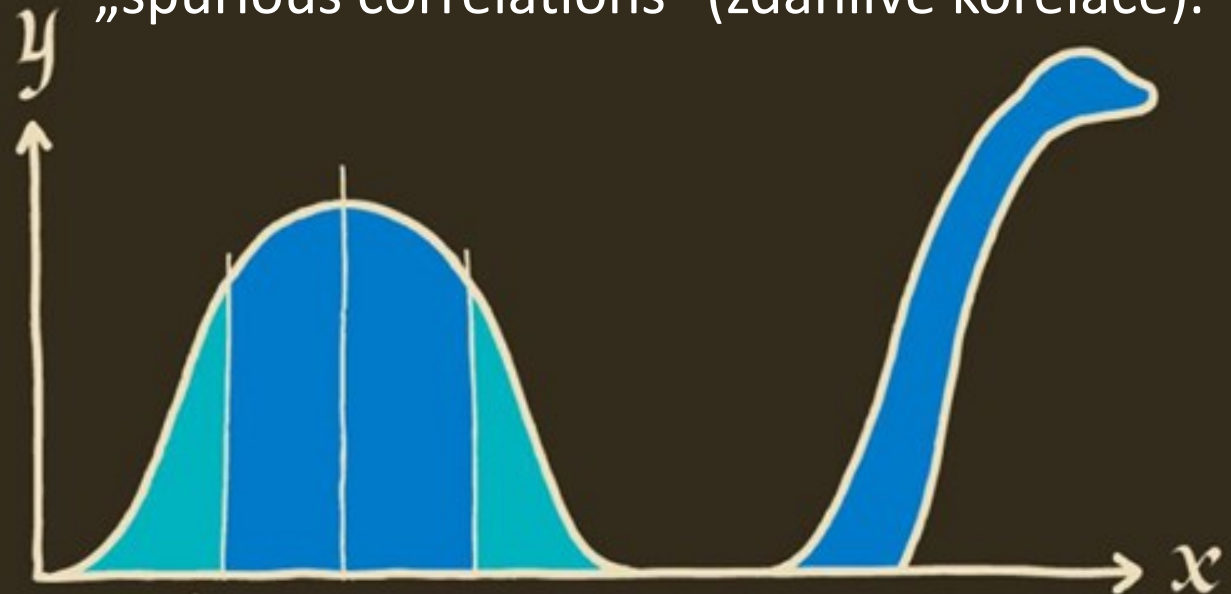


fig 1.0 The Extended Bell Curve.

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

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

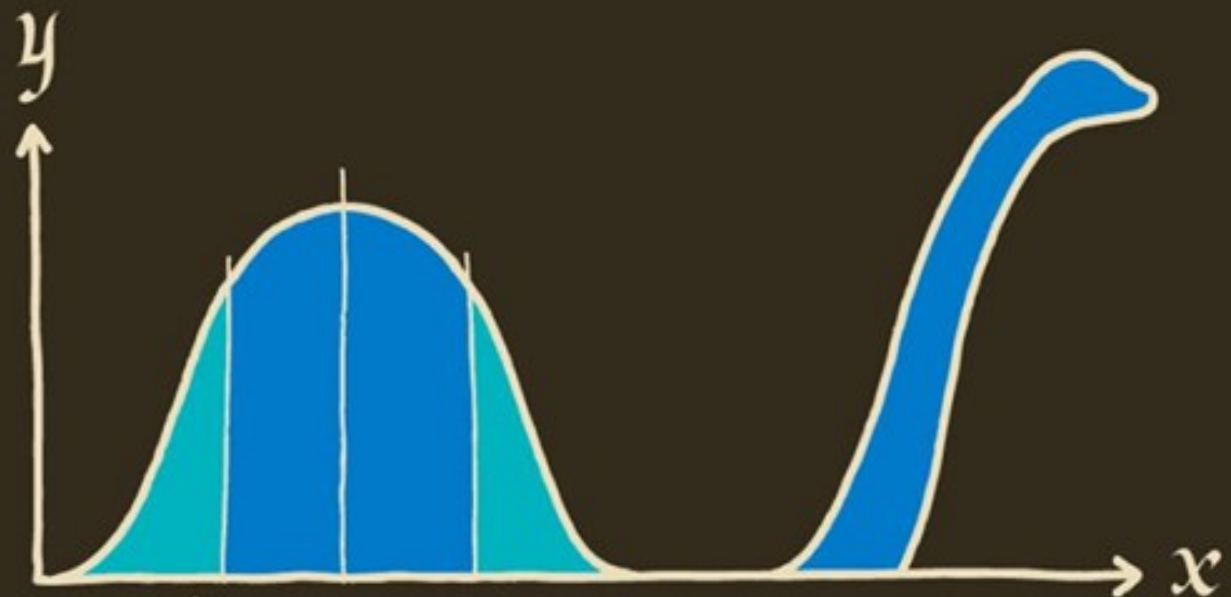


fig 1.0 The Extended Bell Curve.

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
- Dnes
  - 4 překážky kauzality
  - 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!

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).
- Facebook, Twitter apod.

A co dál?

- Tímto to bohužel zdaleka nekončí.

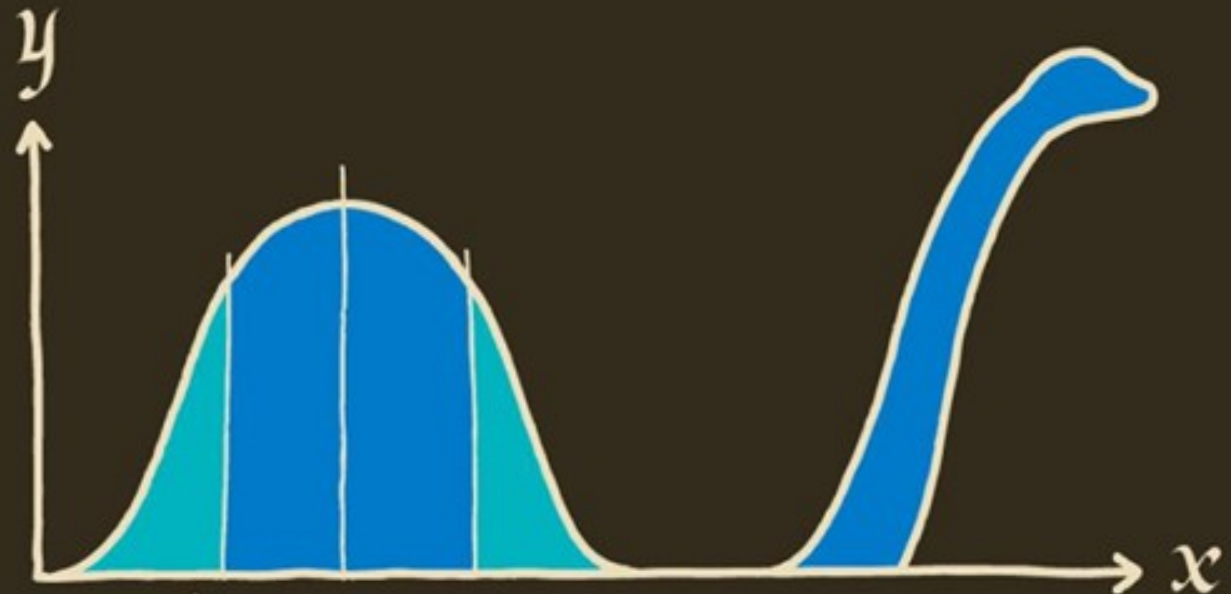


fig 1.0 The Extended Bell Curve.

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 – relativně nová 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.

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.

Děkuji za pozornost!