Základy kvantitativní analýzy dat (statistiky)

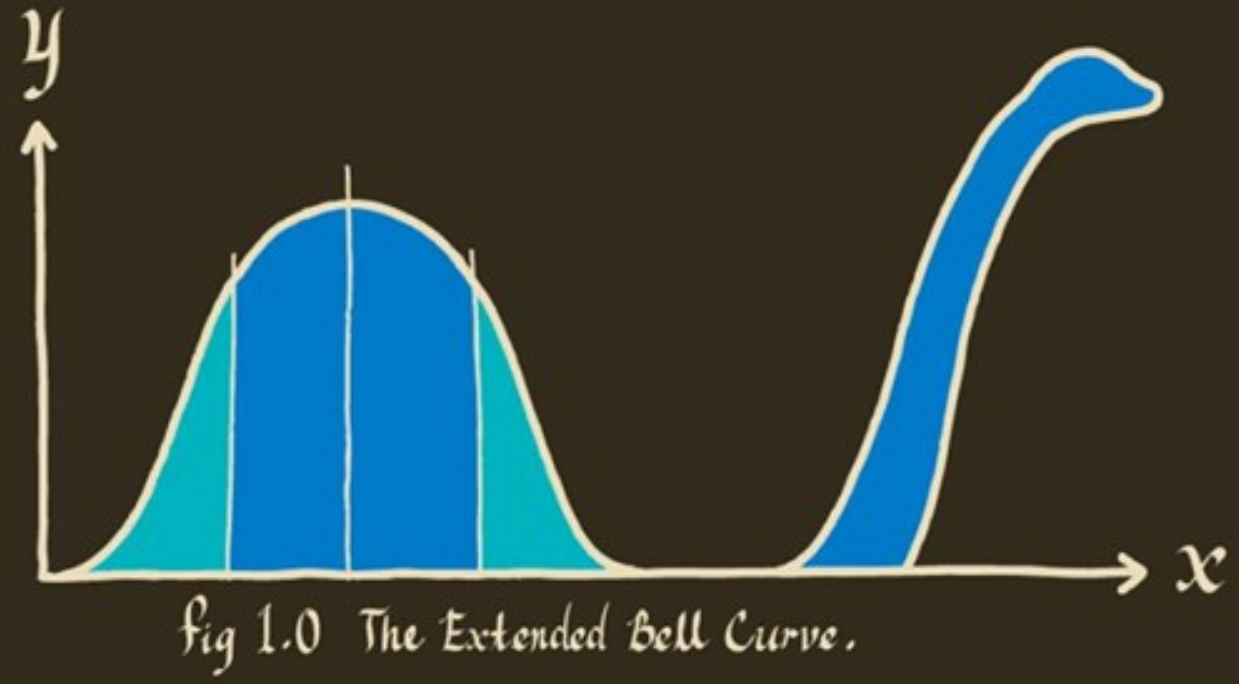


fig 1.0 The Extended Bell Curve.

Jan Kleiner

BSSn4405

jkleiner@mail.muni.cz

Co je cílem úvodu do statistiky v tomto kurzu?

- Prvotní seznámení se statistikou.
- Zdokonalení studentů v kvantitativním a smíšeném výzkumu.
- Nová možnost výzkumného směřování.
- Bourání mýtů (matematika v sociálních vědách).
- Základní představení statistických operací a modelů **nezbytných** pro kvantitativní design výzkumu.

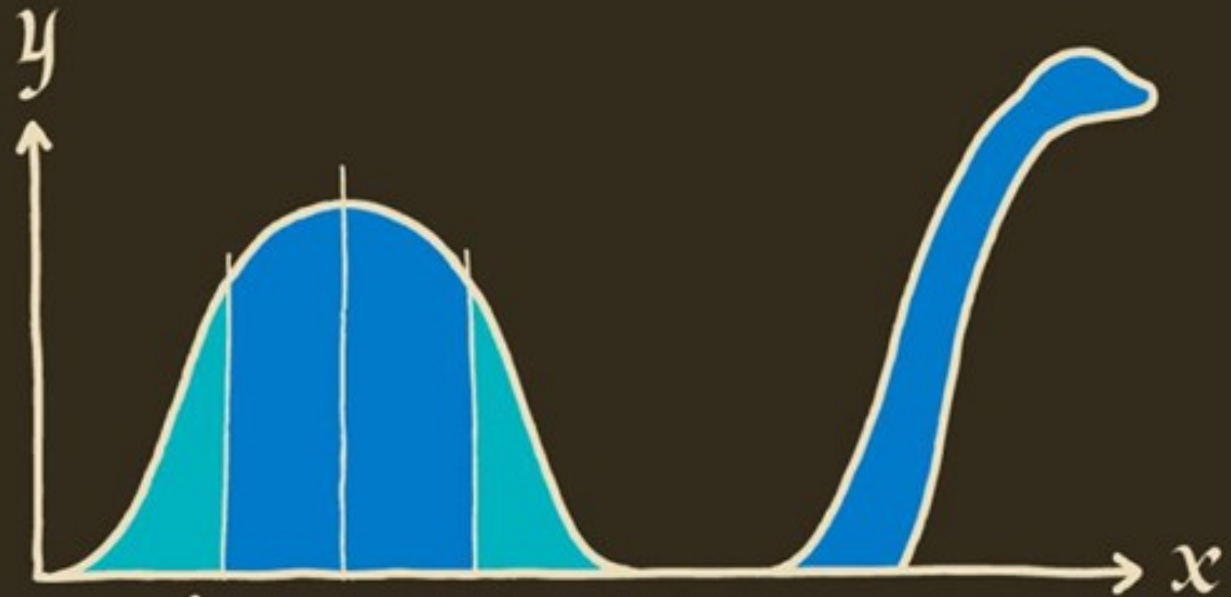


fig 1.0 The Extended Bell Curve.

Studijní materiály

- 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.
  - **Doporučená literatura?**
  - Dobré je učit se statistiku z různých zdrojů.

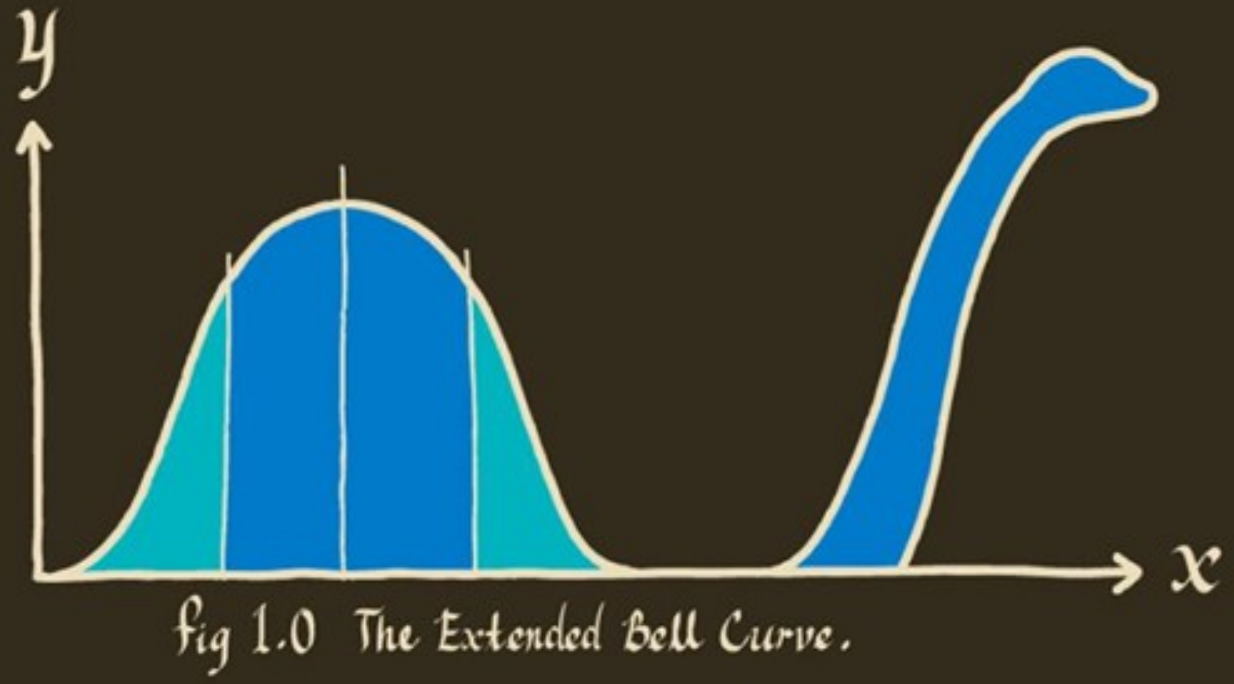


fig 1.0 The Extended Bell Curve.


Proč?
!

- Chystáte se ke sběru dat formou dotazníku nebo rozhovoru?
- Máte kvantitativně laděný výzkum?
- Máte rádi grafy a tabulky?
- Chcete tak, aby to bylo přijímáno širokou vědeckou komunitou prokazovat korelaci a kauzalitu?
- Chcete se zabývat hromadnými jevy v populaci, vytvářet vlastní teorie a rovnou si je i testovat (smíšený výzkum)?
- Chcete mít schopnost odborně posoudit kvantitativní výzkumy?

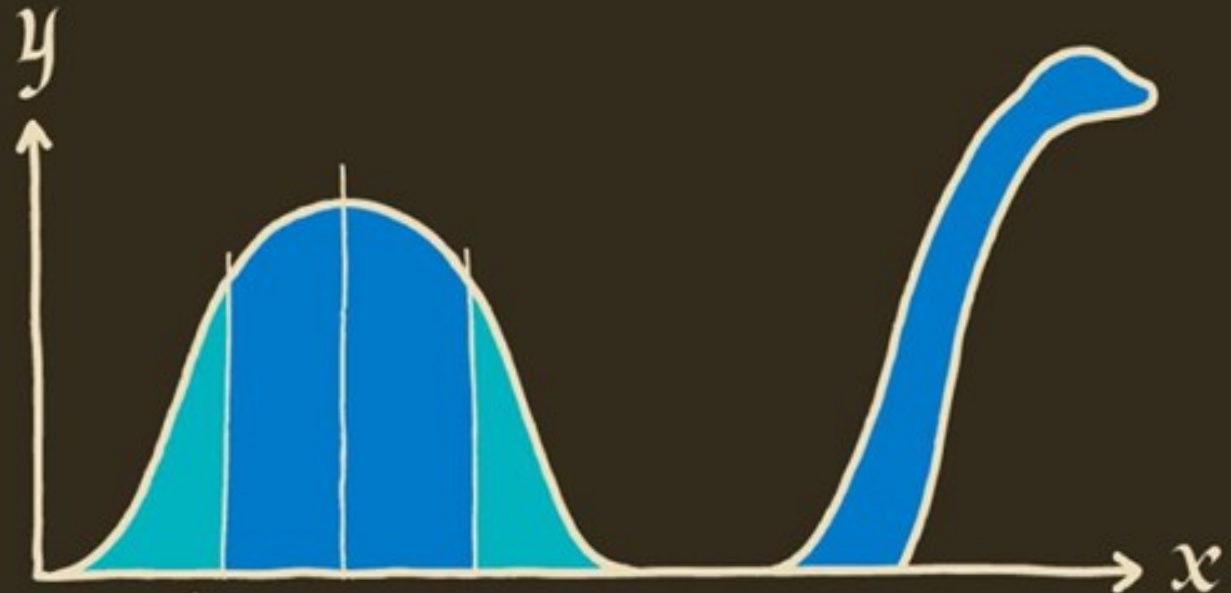
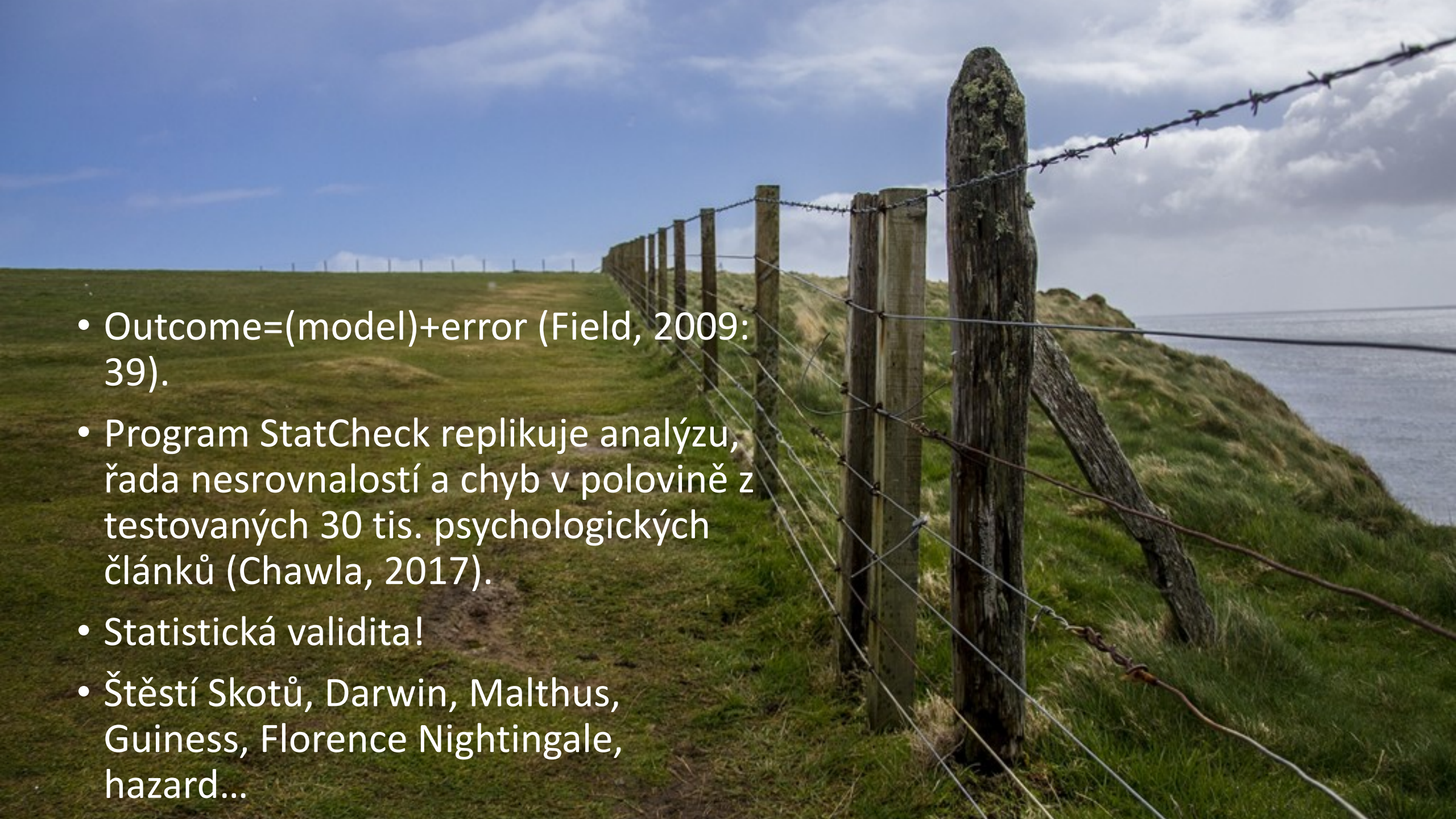


fig 1.0 The Extended Bell Curve.

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…“



- Outcome=(model)+error (Field, 2009: 39).
- Program StatCheck replikuje analýzu, řada nesrovnalostí a chyb v polovině z testovaných 30 tis. psychologických článků (Chawla, 2017).
- Statistická validita!
- Štěstí Skotů, Darwin, Malthus, Guiness, Florence Nightingale, hazard…

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

- Je to „tupý“ nástroj.
- Např. průměrný měsíční příjem (cca 34 tis. CZK) vs. medián - cca 29 tis. CZK (ČSÚ, 2020).

- Jak neklamat – brát v úvahu veškeré informace a obzvláště variabilitu (vysvětlena dále) kolem průměrných hodnot.

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

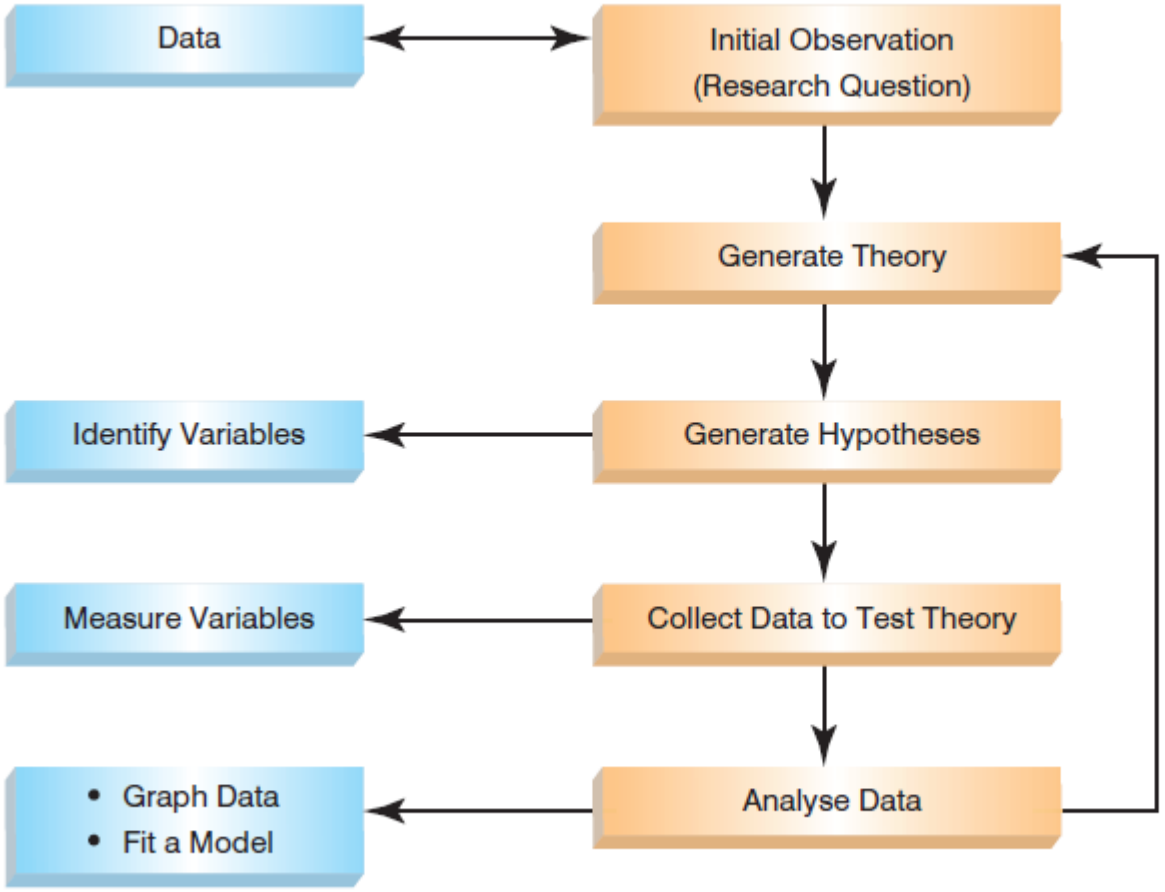


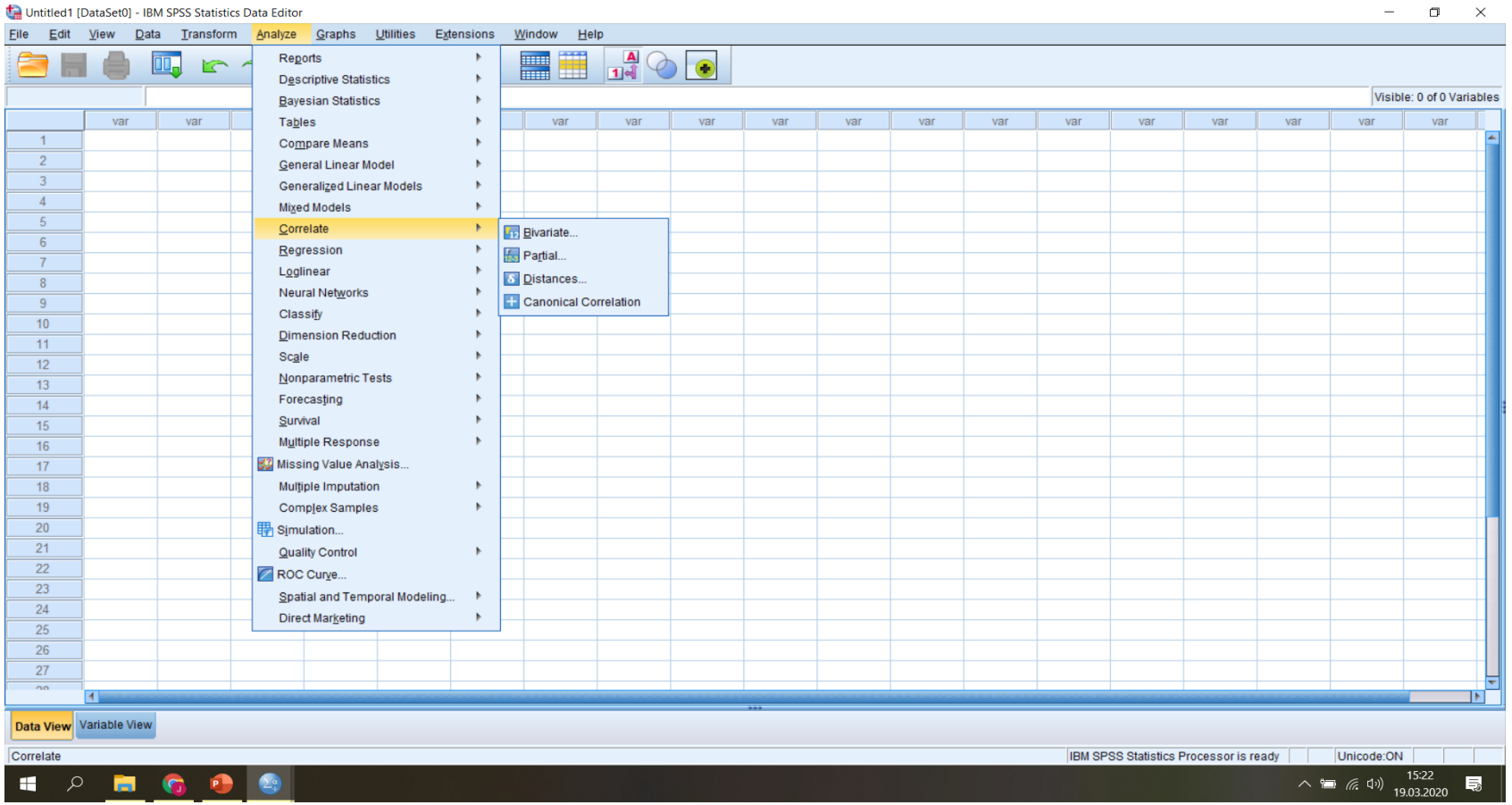
FIGURE 1.2
The research process

Zdroj: Field (2009: 3)

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.

SPSS



- 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.
- K jeho pochopení a k pochopení základních a často používaných statistických operací slouží celosemestrální (podzim) kurz na politologii „Kvantitativní přístupy v politologii“. Vše se učí prakticky a na příkladech → nejlepší způsob, jak tento guláš pochopit.
- Možnost exportu Excelových dat (jednoduchý přenos z dotazníku Google).

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.
4. Nenáhodný vzorek.
5. Neuvedená velikost vzorku.
6. Špatně interpretované korelace.
7. Intervenující proměnné.
8. Špatně uvedená čísla.
9. Selektivní reporting dat.
10. The Almighty Anecdote.

Důležité pojmy

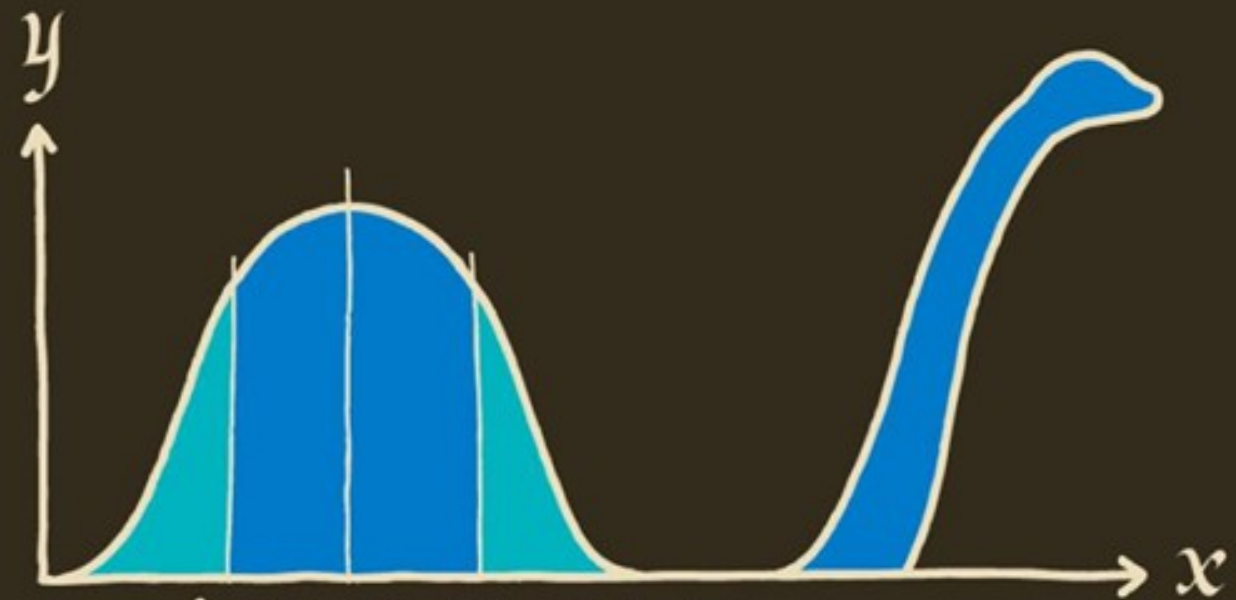


fig 1.0 The Extended Bell Curve.

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).

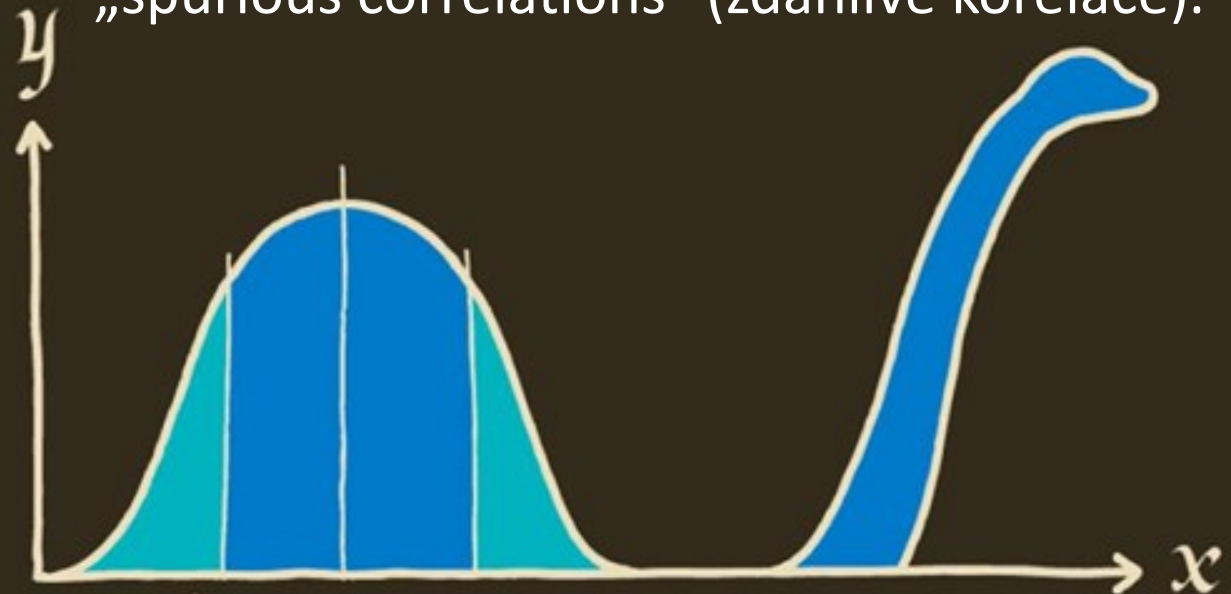


fig 1.0 The Extended Bell Curve.

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

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

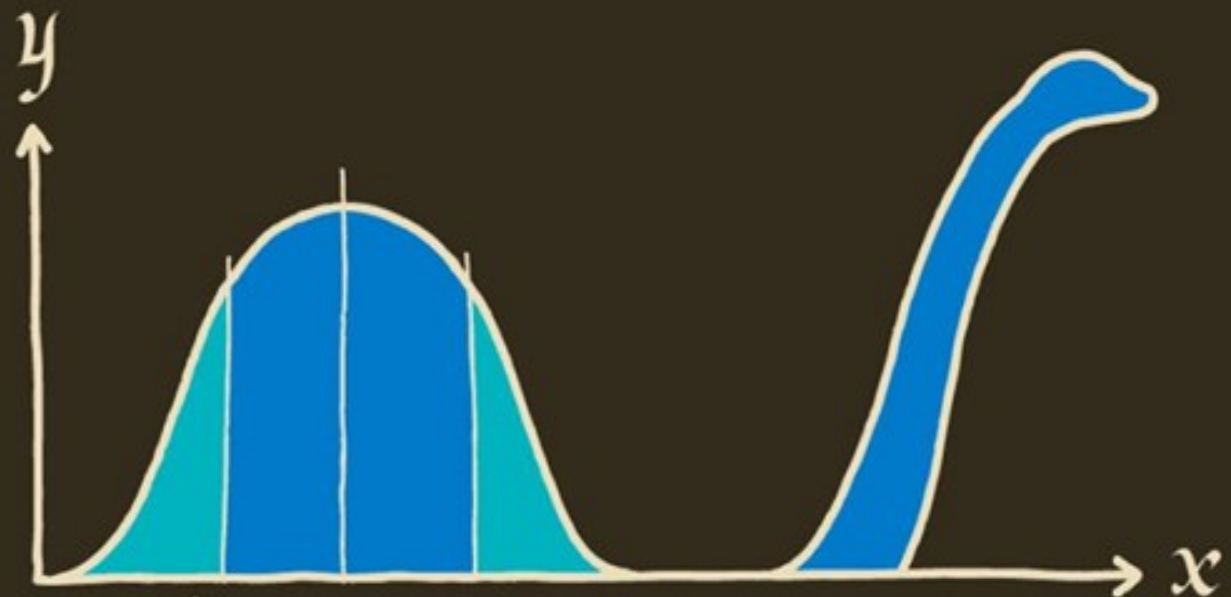


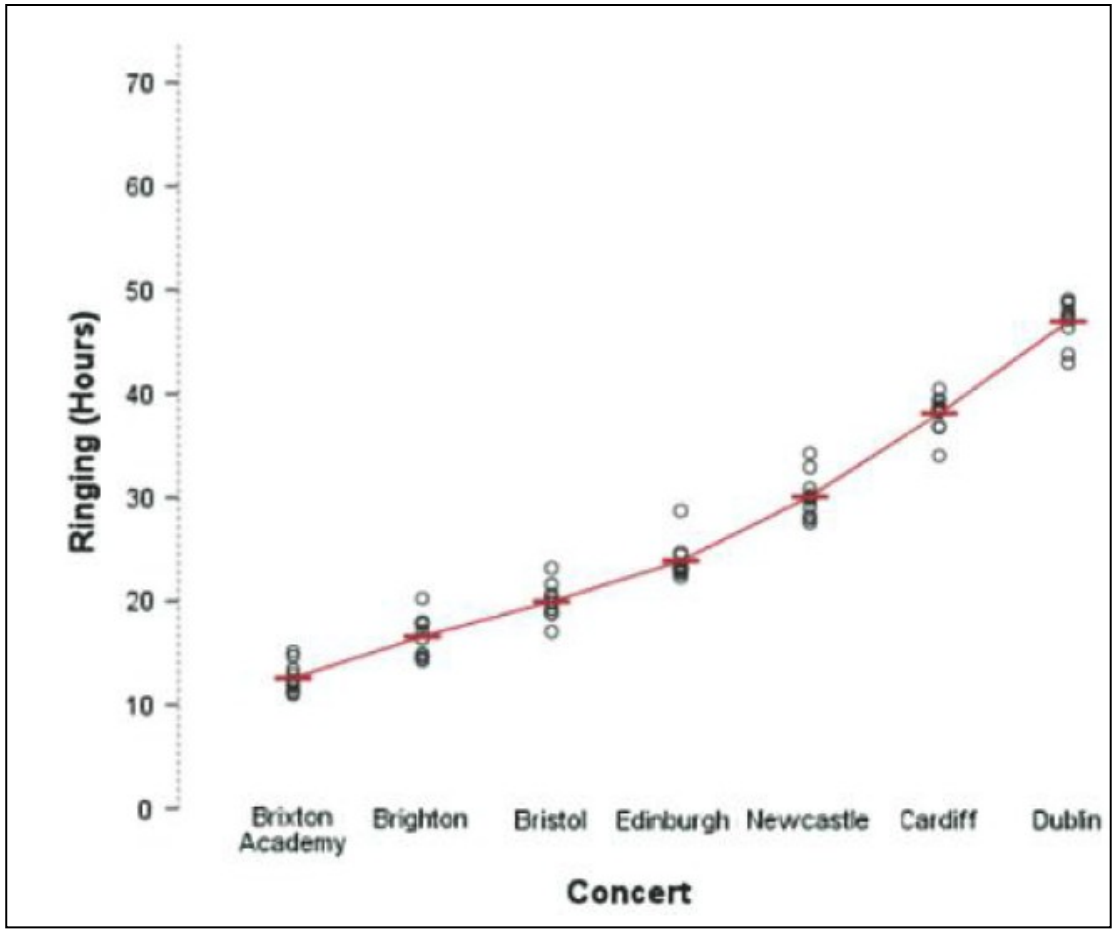
fig 1.0 The Extended Bell Curve.

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!

Variabilita

- Variabilita je kolísání, výkyv. Mírou variability je rozptyl (varience). Na grafu vidíme malý rozptyl (homogenní) na jednotlivých úsecích (v tomto případě koncertní lokace).



Eticky korektní příklad testování hypotézy

- Chceme otestovat H_1: Množství snězené zmrzliny má vliv na to, jestli se jedinec pozvrací (protože jí spápnul hodně).
- Závislá proměnná je nějaké vyjádření nevolnosti (např. dichotomická proměnná pozvrací se (1)/nepozvrací se(0)).
- Nezávislá proměnná je množství snězené zmrzliny (vyjádřeno v ml).
- Provedeme eticky korektní experiment, ve kterém budeme do spousty lidí (případů) cpát zmrzlinu a zaznamenávat, zda se pozvraceli, či nikoliv.
- Na výsledná data aplikujeme statistický model. V tomto případě je vhodná logistická regrese, která kauzálně určuje, zda daný *predictor* (nezávislá proměnná) má efekt na *outcome* – závislou proměnnou. Pokud pak vložíme do modelu další data (např. chci vědět, jestli se člověk pozvrací z 1 litru zmrzliny), pak je model schopen výsledek také predikovat.
- Jak by zněla nulová hypotéza (H_0) k H_1?

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).

A co dál? I

- Tento blok je pouze „vstupní branou“ do studia statistiky a ještě k tomu zdaleka nedokončenou.

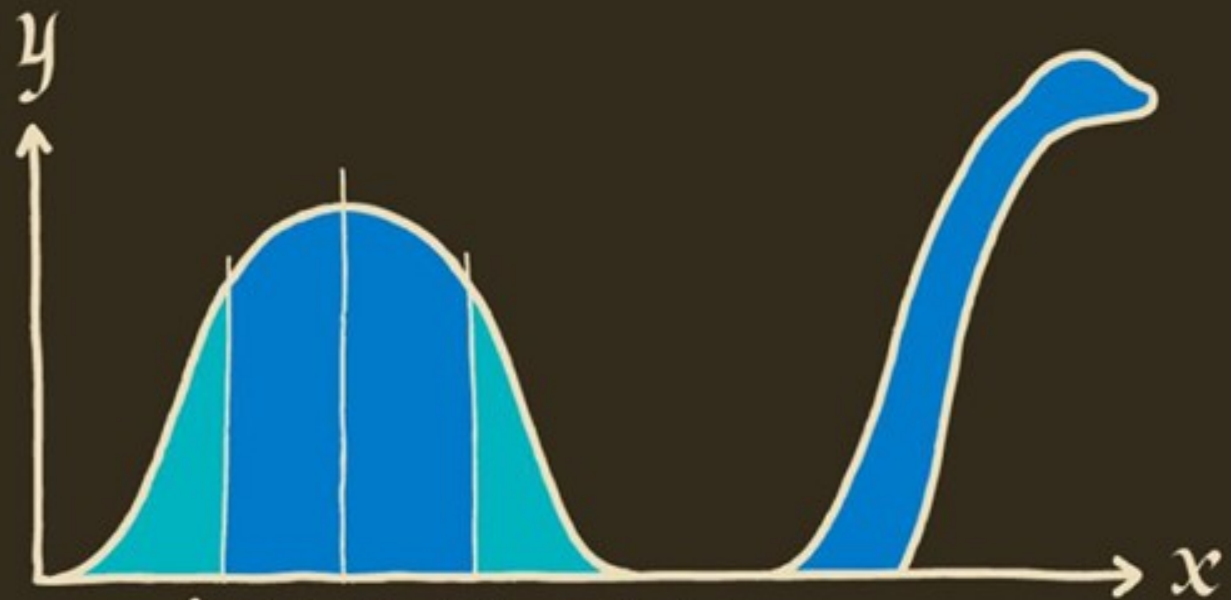


fig 1.0 The Extended Bell Curve.

A co dál? II

- 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).
- Studijní skupiny?

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.

Kontrolní otázky a příprava na zkoušku

- Anglické termíny a jejich české ekvivalenty mohou být matoucí (u zkoušky můžete používat libovolně oboje, pokud budou správně použity).
- Jedná se o složitou a pro spoustu studentů neznámou problematiku. Není pravděpodobné, abyste se hned orientovali ve všem! Otázky se týkají jak prezentace, tak i povinné literatury.

Základy 1z2

- K čemu slouží statistický model skutečnosti?
- Jaké jsou 3 typy modelů s ohledem na to, jak dobře odpovídají realitě?
- Jaké 3 nejjednodušší modely používáme k obecné sumarizaci kvantitativních dat? (mean, modus, medián)
- Co je odchylka (deviance)? Co vypovídá o daném modelu? A proč se používá v umocněné podobě (sum of squared errors; SS)?
- Jaké 3 parametry (measures of fit) určují jak vhodně průměr reprezentuje daná data? (SS, varience, std. deviation)
- Jaký je věcný rozdíl mezi směrodatnou odchylkou (std. deviance)a std. chybou (std. error)?
- Jaký význam má ve statistice Centrální limitní věta (central limit theorem)?

Základy 2z2

- Co jsou chyby 1. a 2. typu (type 1 and type 2 errors)?
- Co je interval spolehlivosti (konfidenční interval, confidence interval) a co znamená, pokud je jeho hodnota nízká, a pokud je vysoká?
- Vyjmenujte 5 fází kvantitativního výzkumu podle Fielda a stručně je popište.
- Co znamenají hladiny statistické významnosti 95 % ($p<.05$)a 99 % ($p<.01$)? Může být nulová hypotéza pravdivá? Stručně zdůvodněte.
- Co nám říká velikost efektu (effect size)?
- Lze vypočítat velikost potřebného vzorku, abychom detekovali požadovaný efekt? Pokud ano, co k tomu potřebujeme?
- V jakých případech využíváme vzorek populace a jak zajistíme, aby byl reprezentativní?

Děkuji za pozornost!