Výběr vzorku a bias

Lukáš Hejtmánek

lukas.hejtmanek@fhs.cuni.cz

Co si odnést

Co je to vzorkování/sampling/výběr vzorku?

Proč vzorkování děláme a k čemu primárně slouží?

Jaké metody máme k dispozici a v čem se liší?

Co to je výběrový bias?

Jakých systematických chyb se můžeme během vzorkování dopustit?

Kdy je výběr vzorku zásadní a kdy nehraje zas tak moc roli?

Co to je výběr vzorku/sampling

Výběr skupiny pro náš výzkum, kterou jsme schopni změřit

Jeden z nejdůležitějších kroků vybraného designu

Garbage in -> Garbage out

Cílem vzorkování je vybrat takovou skupinu, která

- **bude dobře representovat populaci**
- zúčastní se studie
- studii dokončí

Vzorek vs populace

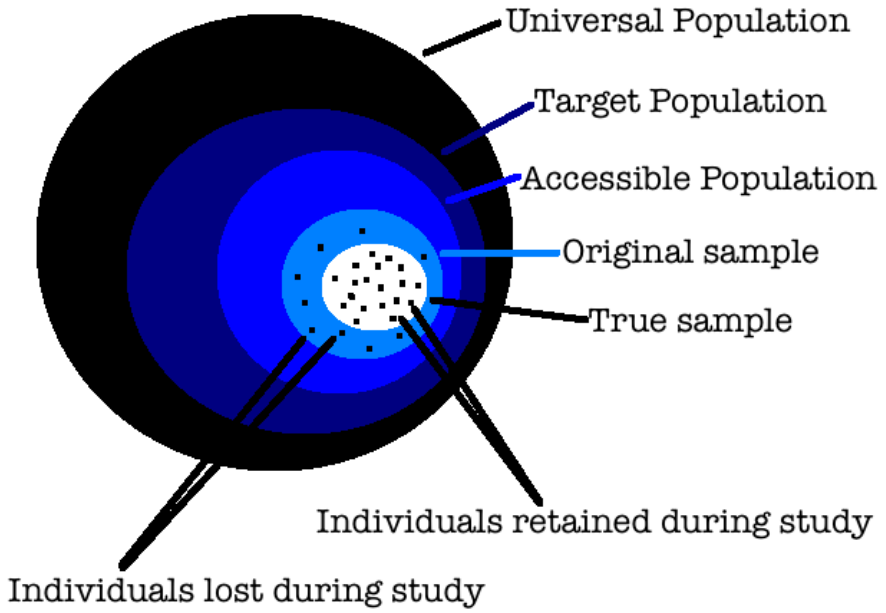
Populace = cílová skupina pro náš výzkum

Vzorek

skupina, která dostatečně přesně representuje populaci

Representativní vzorek pak umožňuje utvářet závěry pro populaci

Nerepresentativní vzorek vytváří zkreslení



Statistika nám umožňuje dělat závěry o populaci na základě menšího vzorku, ale aby fungovala, musí být vzorek dostatečně representativní celé populaci

Co je populace následujících otázkách?

https://journals.sagepub.com/action/showMostReadArticles?journalCode=PSS

A Large-Scale Test of the Goldilocks Hypothesis: Quantifying the Relations Between Digital-Screen Use and the Mental Well-Being of Adolescents

Andrew K. Przybylski ✉ and Netta Weinstein View all authors and affiliations

Volume 28, Issue 2 | https://doi.org/10.1177/0956797616678438

Thinking More or Feeling Less? Explaining the Foreign-Language Effect on Moral Judgment

Sayuri Hayakawa ✉, David Tannenbaum, [...], and Boaz Keysar +2 View all authors and affiliations

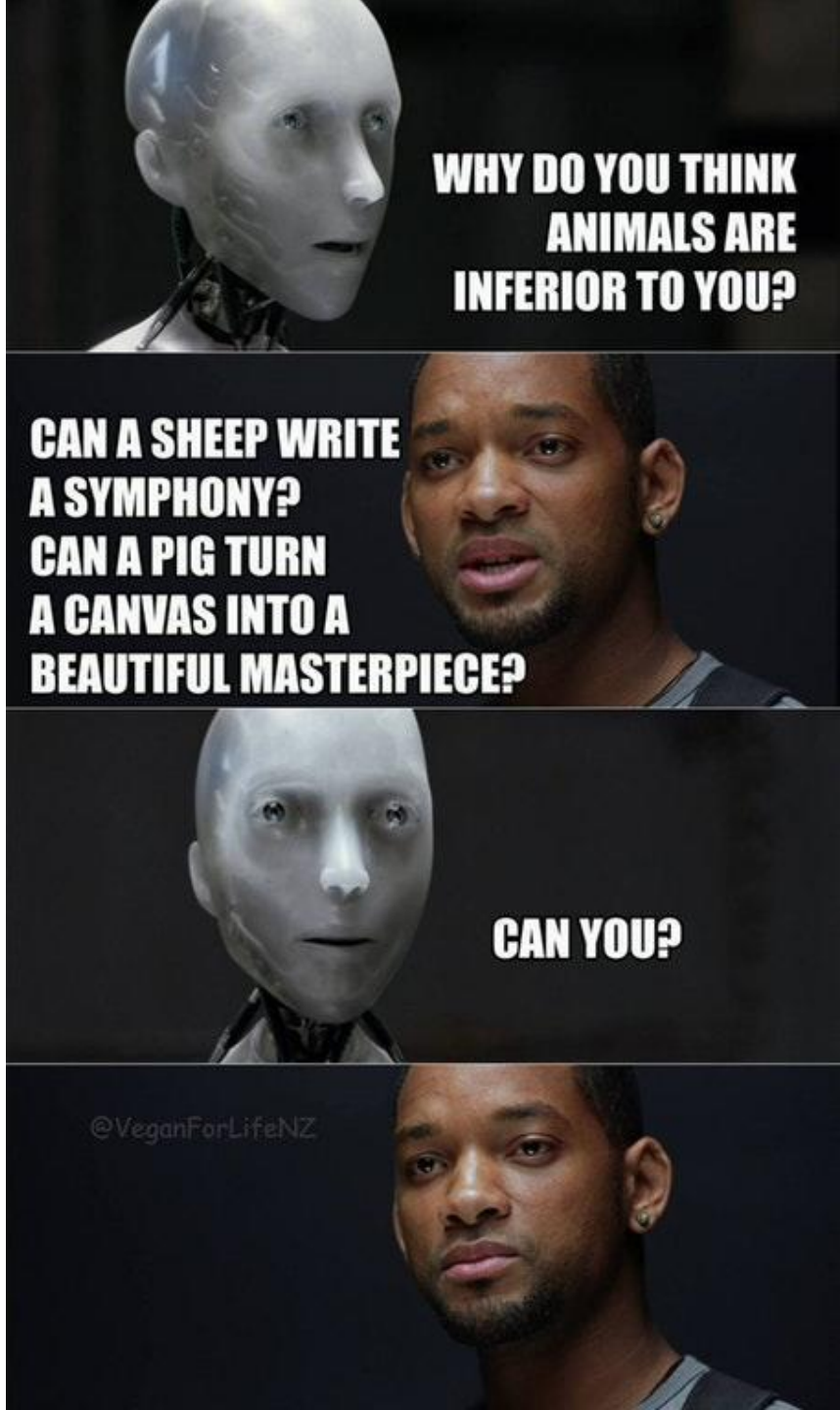
Volume 28, Issue 10 | https://doi.org/10.1177/0956797617720944

Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention

Gordon Pennycook ✉, Jonathon McPhetres, [...], and David G. Rand +2 View all authors and affiliations

Volume 31, Issue 7 | https://doi.org/10.1177/0956797620939054

Representativní vzorek?

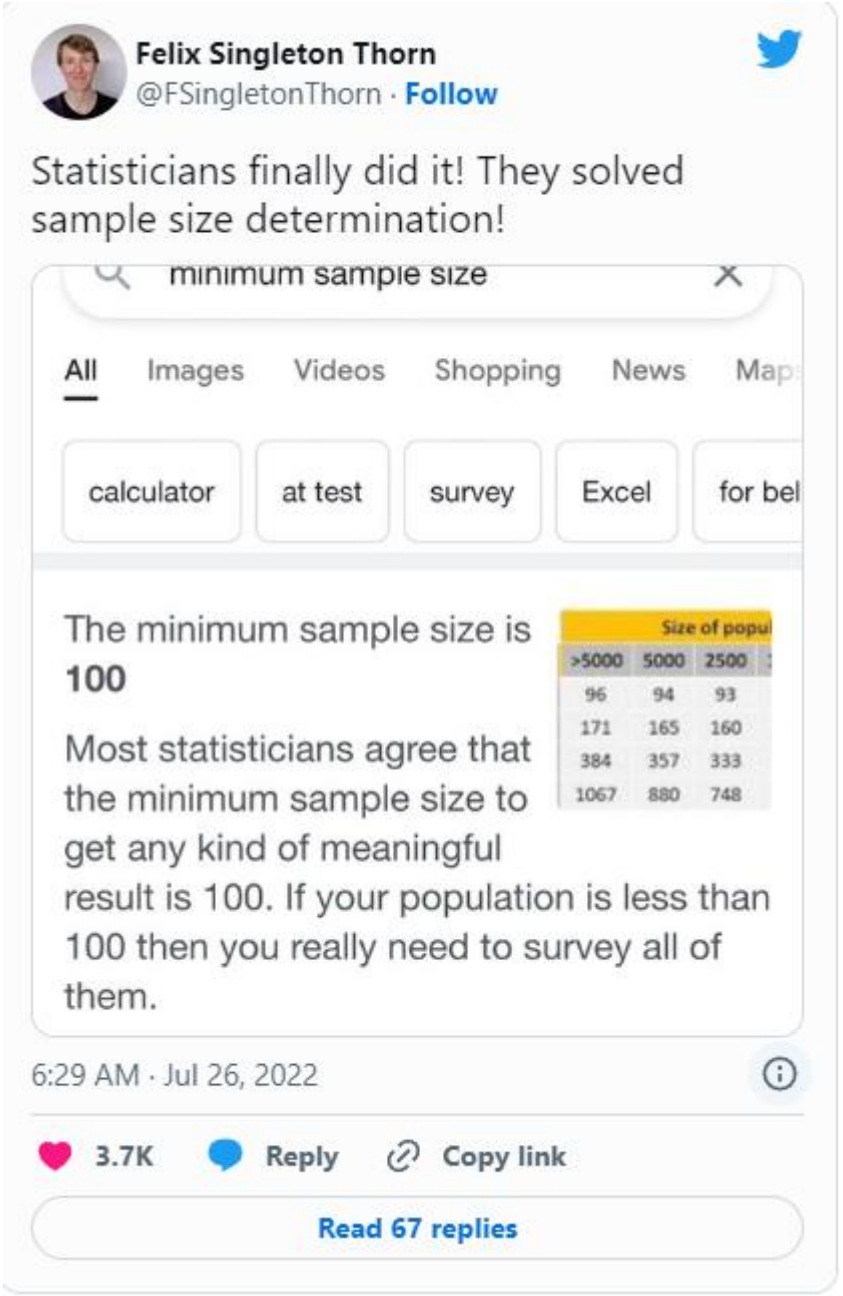


Velikost vzorku

Čím větší náhodný vzorek, tím representativnější

Číslo 30 vzniklo “humorně”, ale je považováno za typický standard

Čím víc, tím líp! Víc participantů nic nezkazí



Konkrétní příklad

Výzkumná otázka:

Chtěl bych vědět, kolik studentů UK má psychické obtíže, s jakými obtížemi studenti bojují, jaké jsou toho příčiny (interní, externí), co by jim pomohlo a co by UK mohla dělat?

Rád bych 100 osob. Jak to udělat?

Jak z populace vybrat

Náhodný vs nenáhodný výběr

Když nevím, tak je náhoda nejlepší rádce

Náhoda zabrání systematické chybě

Náhodný výběr

- Každý jedinec z populace má stejnou šanci se do vzorku dostat

Nenáhodný výběr

- Někteří jedinci nemají šanci, či mají šanci sníženou

Jak udělat náhodný vzorek studentů UK?

Representativnost

I náhodný vzorek nemusí být nutně representativní vůči všem, jen vůči cílové populaci

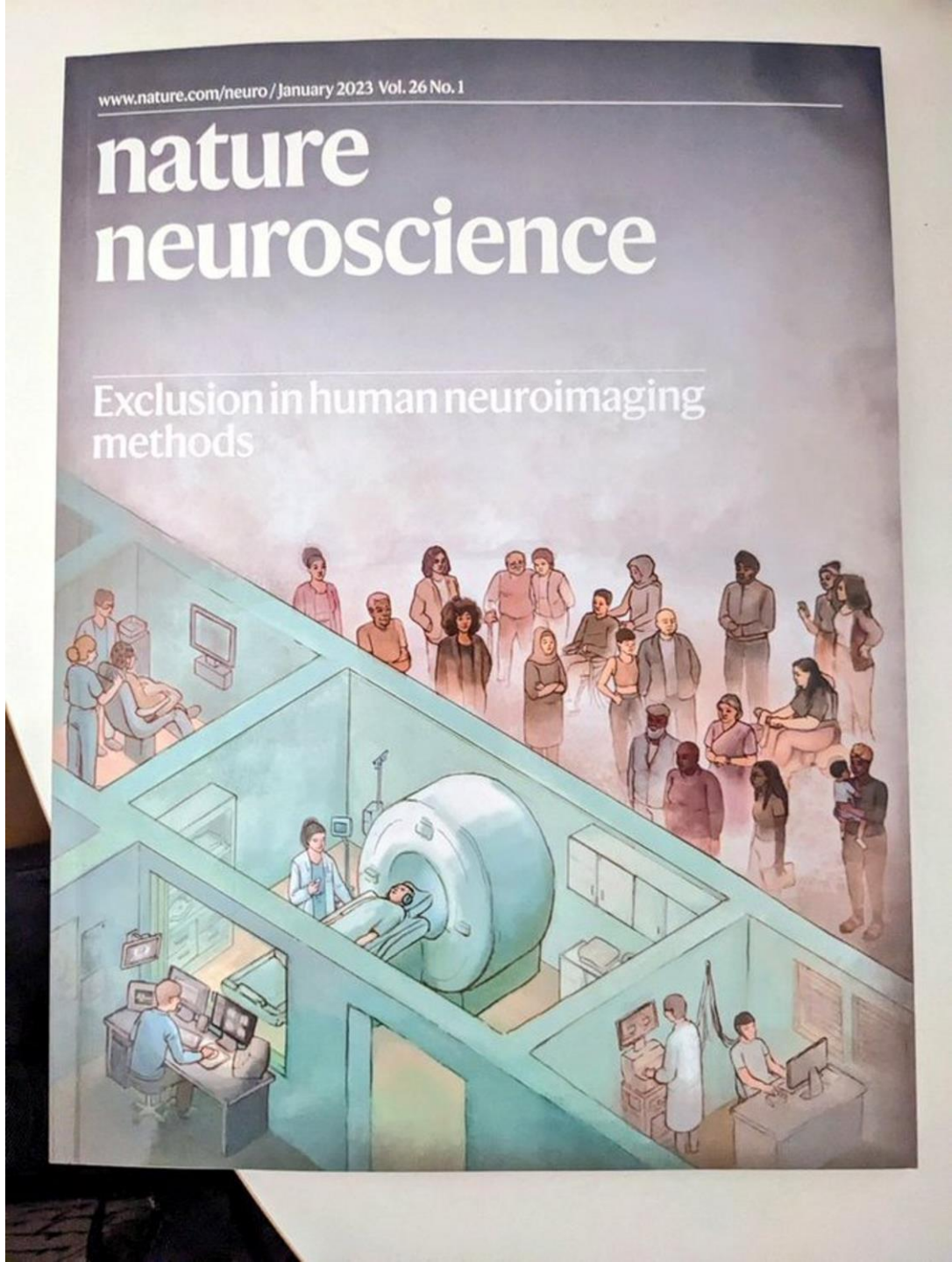
Perspective | Published: 23 December 2022

Confronting racially exclusionary practices in the acquisition and analyses of neuroimaging data

J. A. Ricard, T. C. Parker, E. Dhamala, J. Kwasa, A. Allsop & A. J. Holmes

Nature Neuroscience **26**, 4–11 (2023) | Cite this article

8742 Accesses | **244** Altmetric | Metrics



Bias – systematická chyba

Systematická chyba ve výběru/měření

Western, educated, industrialized, rich, and democratic (WEIRD)

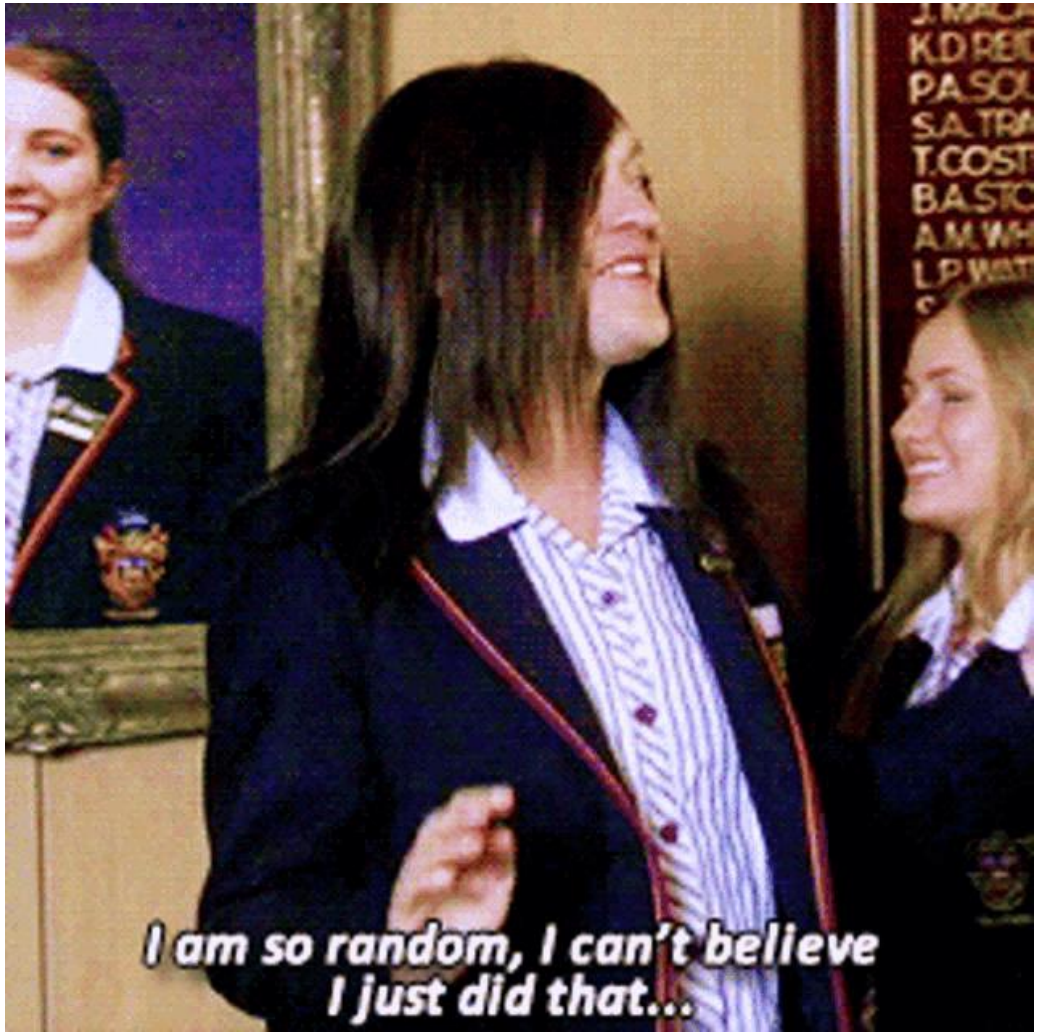
Beyond Western, Educated, Industrial, Rich, and Democratic (WEIRD) Psychology: Measuring and Mapping Scales of Cultural and Psychological Distance

Michael Muthukrishna, Adrian V. Bell, [...], and Braden Thue +4 View all authors and affiliations

Volume 31, Issue 6 | https://doi.org/10.1177/0956797620916782

Náhodný vzorek

1. Definice populace
2. Tvorba seznamu všech
   Opravdu všech!!!
3. Náhodný výběr
   Jak na to?
4. Oslovení
   Co když se neozvou?
5. Sběr dat



Jak na to?

Jak udělat seznam všech obyvatel ČR?

Jak udělat seznam všech lidí s duševním onemocněním?

Jak udělat seznam všech lidí v CŘ, kterým někdo zemřel na Covid?

Jak udělat seznam psychologů/terapeutů?

Jak udělat seznam lidí, co žijí na Praze 1?

Proč jsou volby na P1 nerepresentativní?

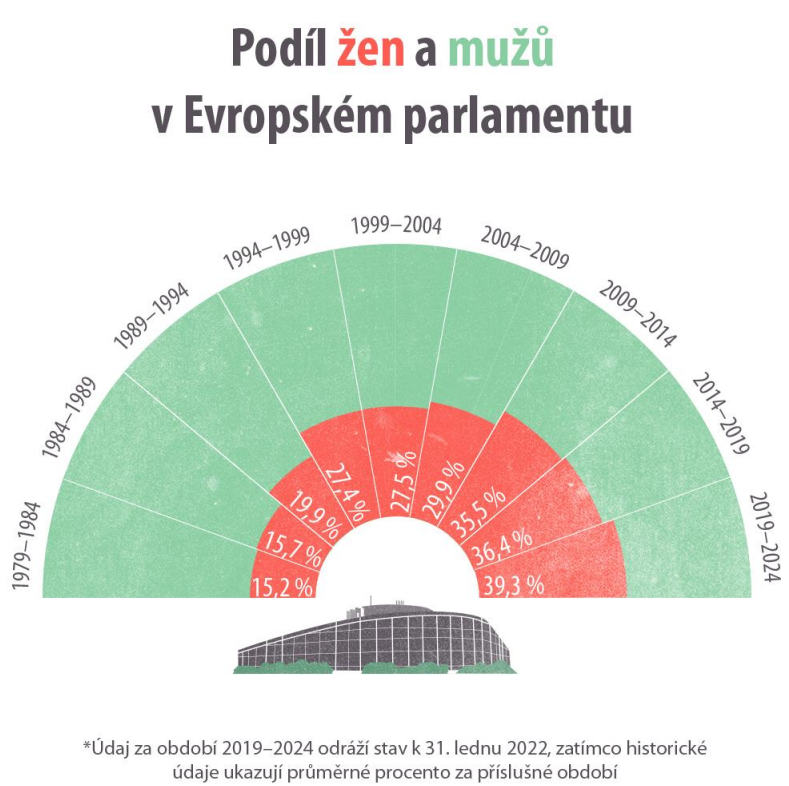
Stratifikované/stratifikační vzorkování

Strata
- odlišné kategorie, např. pohlaví, věkové kategorie atd.
- podíl strat ve vzorku odpovídá populačnímu podílu
- strata jsou relevantní k výzkumné otázce

Stratifikace je podobná odstiňování kontrolních proměnných, akorát přichází před samotným měřením

Klady
- vede k vyrovnanějšímu vzorku

Zápory
- Strata musíme znát dopředu
- Musíme znát velikost strat



Jak by bylo vhodné stratifikovat vzorek?

Jak udělat stratifikovaný vzorek studentů UK?

Je malý dárek dobrou motivací pro příspěvek na dobročinné účely?

Vede autoritativní leadership k lepším pracovním výsledkům?

Je hudba vhodný nástroj při učení se na zkoušku?

Nenáhodné výběry vzorku

Za nenáhodné považujeme ty výběry, které neumožnují každému členovi populace se výzkumu zúčastnit

Příležitostný (Oppportunity, convenience)

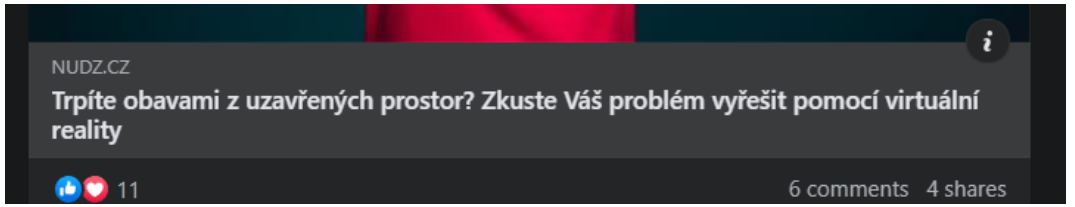
Lavinový (Snow-ball)

Účelový

Kvótový (stratifikační postup, ale ne náhodný)

Nenáhodné výběry vedou k systematické chybě (bias) v datech

Nenáhodné výběry vzorku



Příležitostný (Oppportunity, convenience)

Beru ty, ke kterým mám přístup a koho umím oslovit

Lavinový (Snow-ball)

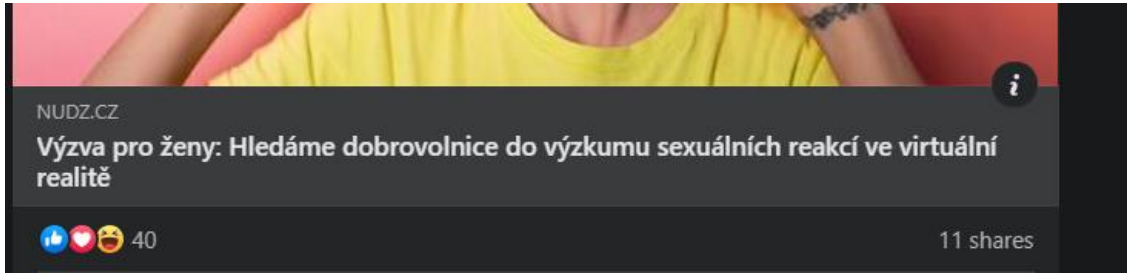
Účelový

Stanovení jasných kritérií a výběr konkrétních lidí, splňujících podmínky

Kvótový (stratifikační postup, ale ne náhodný)

Stanovení strat, pak nenáhodný výběr do chvíle, než se naplní kvóta

Už mám dost žen, potřebujeme muže



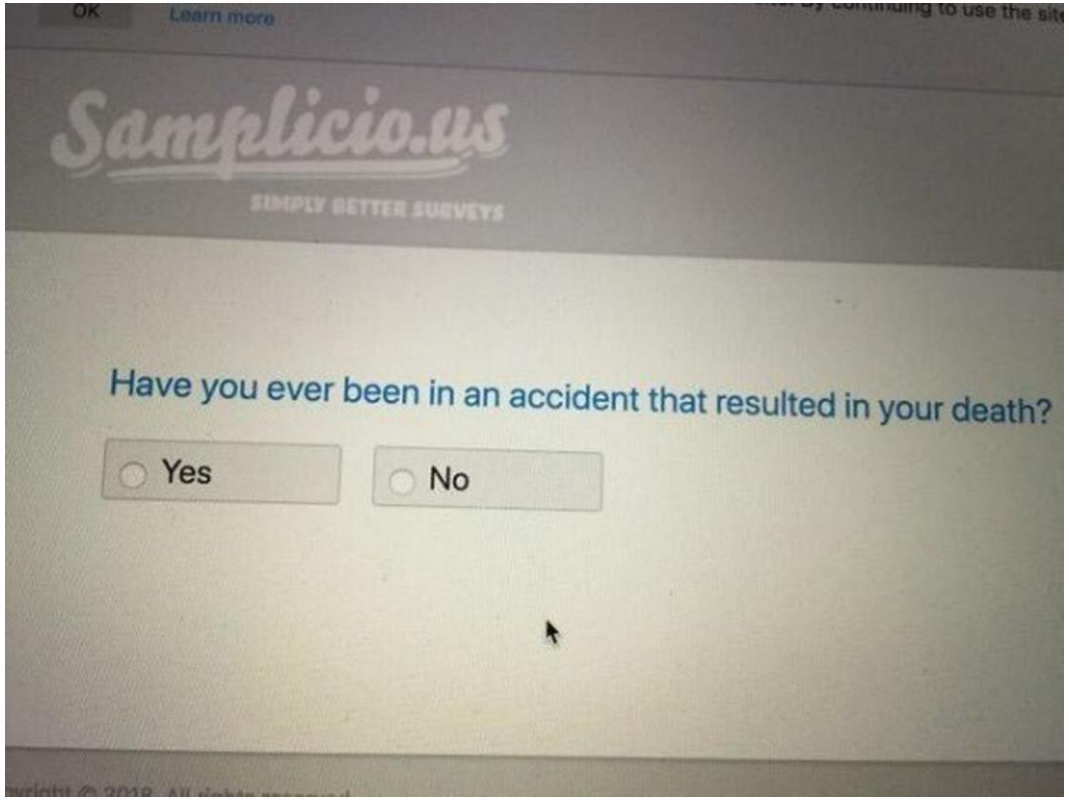
Náhodná vs systematická chyba

Náhodná chyba

- občas tak a občas jinak
- někdy zkrátka vyberu vzorek s nadanějšími studenty/depresivnějšími pacienty atd.
- Statistika modeluje a „odstraňuje“ náhodnou chybu výběru/měření

Systematická

- systematicky zkreslená na jednu stranu
- statistika jí vůbec nevidí (jsou zkrátka jiné, než by měla být)



Typy systematické chyby výběru

Sampling bias (obecný název pro systematicky chybný výběr)

Chyba podreprezentovanosti (undercoverage bias)

Chyba sebevýběru (self-selection bias)

Ne každý se do studie přihlásí

Chyba neodpovídání (non-response bias)

Ne každý se ozvě zpět či odpoví na otázku

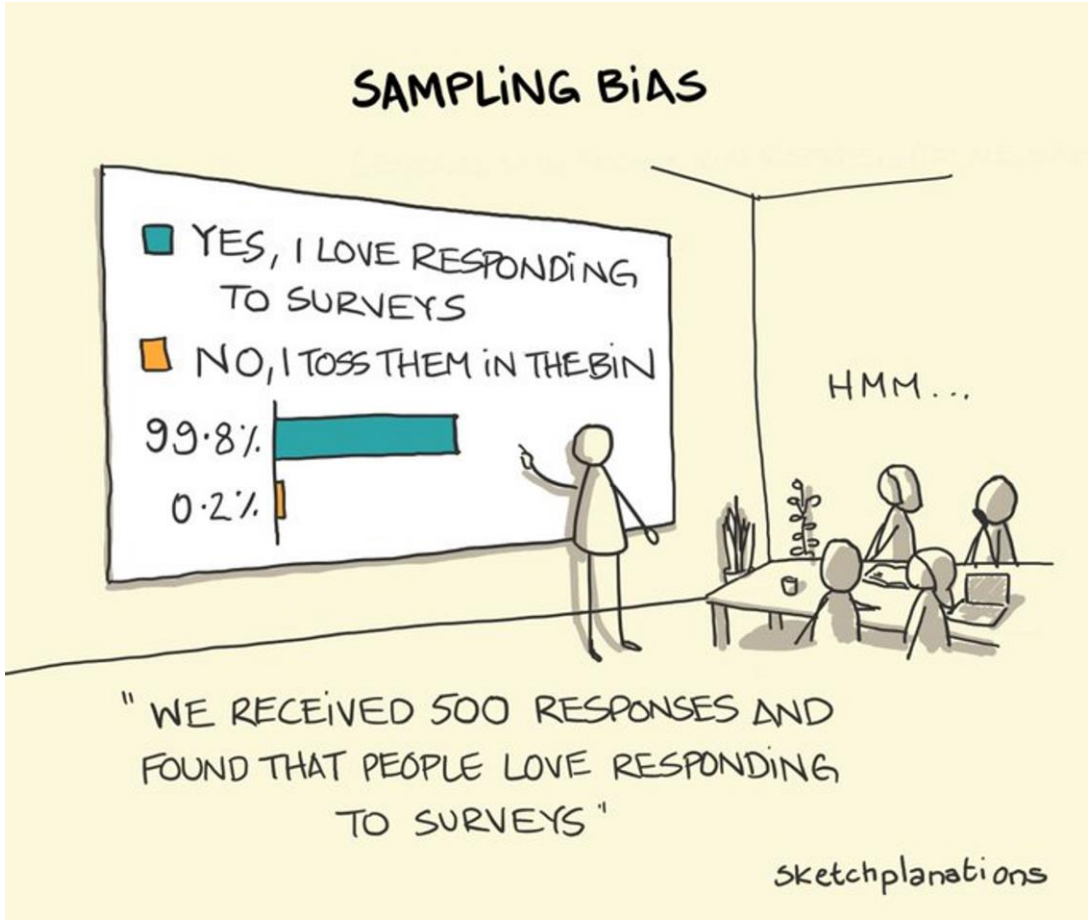
Chyba přežívání (Attrition bias, Survivorship bias)

Chyba sebevýběru (self-selection) a neodpovídání (non response)

Do studie se hlásí pouze osoby, kterým je téma blízké, které zajímá či ke kterému se chtějí vyjádřit

Přihlaste se do studie, kde budeme zkoumat pozitivní vliv počítačových her na kognitivní schopnosti dětí

Dejte nám vědět, jak se vám líbily přednášky profesora Nudného?



Chyba podreprezentovanosti (undercoverage)

Část populace není ve vzorku vůbec zastoupená či je zastoupená nedostatečně

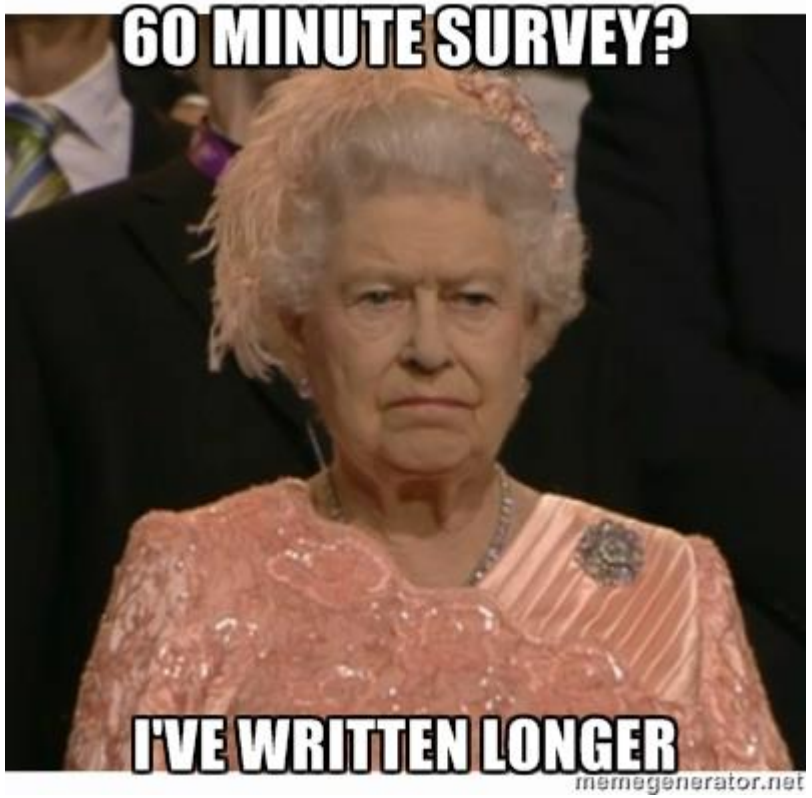
Typická pro convenience/opportunity sampling

Problémy s

- **Časem a penězmi** – nemám kapacitu účastnit se studie
- **Jazykovou bariérou** (zejména USA výzkumy)
- **Způsoby oslovení** (facebook, telefon, oslovení na ulici)

Propojená se sebevýběrem

- Metodika studie eliminuje některé skupiny

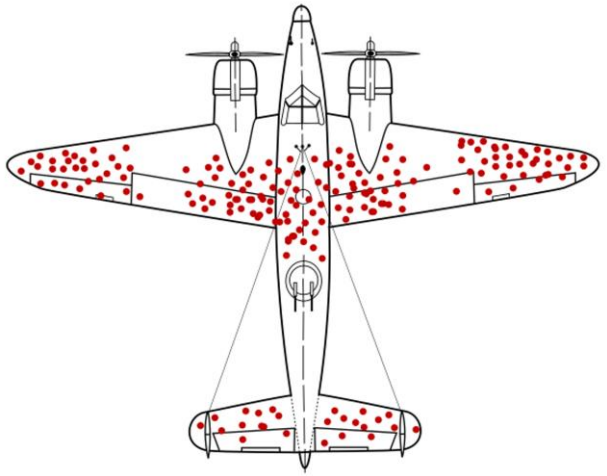


Problém peněžité odpovědi

Problém peněz za studie

odstraňuje undercoverage bias a vytváří self-selection bias

Chyba přežívání a přeživšího



Proč mají televizní seriály stoupající tendenci v hodnocení?

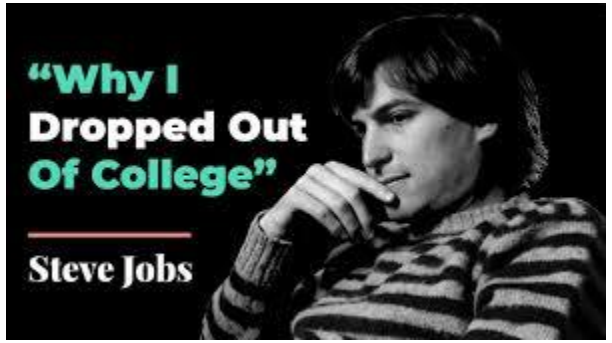
Naše léčba je efektivní, po 30 ti týdnech je 95 procent pacientů spokojených.

chyba přežívání / attrition bias

Ve studii zůstávají jen osoby, které baví/vyhovuje/pomáhá

chyba přeživšího / survivorship bias

Tvoření závěrů na základě těch, kteří ve studii zbyli bez reflexe toho, kdo jí nedokončil



Typy systematické chyby výběru

Jak udělat nenáhodné výběry studentů UK?
A k jakým chybám to povede?

Kde je bias

Cílem této studie je prozkoumat účinnost nového programu hubnutí u dospělých. Studie využívá k náboru účastníků online průzkum šířený pomocí Google Adds a Facebook Adds.

Cílem této studie je vyhodnotit účinnost nového programu na snížení stresu u pracovníků nemocnic. Studie provádí nábor účastníků zasláním e-mailu všem zaměstnancům nemocnice.

Cílem této studie je prozkoumat vztah mezi pěší dostupností sousedství a fyzickou aktivitou starších dospělých. Studie získává účastníky prostřednictvím letáků vyvěšených v centrech pro seniory a komunitách důchodců.

Cílem této studie je prozkoumat vliv nové výukové metody na výsledky středoškolských studentů v matematice. Studie získává účastníky zasláním dopisů všem rodičům středoškoláků v určitém okrese.

Co s biasem?

Pokud očekáváme možný bias, pak se minimálně zeptáme
- Pohlaví, věk, národnost, znalosti/zkušenosti atd.

Ne vždy bias vadí/ne vždy se jedná o bias
- O bias se jedná pouze, pokud ve vzorku vznikne systematická chyba, ale to se nestane vždy

Bias se dá odstínit, pokud je „variabilní“
- Nemusíme se v ČR ptát na národnost, pokud to všichni budou Češi
- Lepší Slováky vyhodit, než analyzovat 40 Čechů a 2 Slováky

Problém je **nevariabilní** stínící proměnná
- Pokud víme, že věk má vliv na pozornost, pak je vzorek 50ti lidí věku 25-27 k ničemu
- 40 žen na 5 mužů je také problém, potřebujeme alespoň nějaké minimum

Vzorek nemusí být vyrovnaný, pokud je **variabilní**
- Např. 100 žen na 30 mužů může být stále OK

Jak biasu předejít?

Zeptejte se v týmu - co jiného, než nezávislá proměnná, by mohlo ovlivňovat výsledek/závislou proměnnou?

Hledáte všechny možné typy stínících proměnných

Vadí, že jsem výzkum dělala v ČR?

Vadí, že jsem dělala výzkum jen na dvou školách?

Co by se mohlo stát jinak na jiných školách?

Mohu se na potenciální stínící proměnné zkrátka jen doptat?

V případě stínících proměnných chceme dostatečnou variabilitu!!!

I výzkum v psychologii je psychologie ☺

vcítění do situací, empatické představování si výstupů

Děkuji za pozornost

lukas.hejtmanek@fhs.cuni.cz