ZURN6311 DOTAZNÍKOVÝ VÝZKUM: SAMPLING 2

Lenka Dědková

ÚKOL 1

- Co jste řešili – rozložení uživatelů FB z hlediska:
  - Genderu (rozdílné v různých zemích)
  - Věku
  - Zemí a regionů
  - SES a vzdělání
  - Přístupu na internet (pomocí jakého zařízení)
- Důležité: jak jsou jednotlivé kategorie (např. vzdělanostní skupiny, věkové rozložení, muži a ženy) zastoupené v obecné populaci, na kterou chcete generalizovat? V čem se uživatelé FB tedy liší od populace?
- Spousta dat dostupná pro americkou populaci, pro ČR např. data ohledně používání soc. sítí obecně
- Nezapomínejte jednotně citovat zdroje (např. podle APA) a to včetně reportů a statistik nebo neodborných zdrojů)
- Krátké shrnutí na konci textu

ÚKOL 1 – ZAJÍMAVÉ ZDROJE

- **Světové statistiky - Facebook:**
- Tankovska, H. (2021). *Statista*. https://www.statista.com/topics/751/facebook/
- World Population Review. (2021). *Facebook Users by Country 2021*. https://worldpopulationreview.com/country-rankings/facebook-users-by-country
- Facebook Investor Relations. (2021). *Annual reports*. https://investor.fb.com/financials/default.aspx

- **Světové statistiky – sociální média a internet:**
- Pew Research Center. (2019). Social Media Fact Sheet. https://www.pewresearch.org/internet/fact-sheet/social-media/
- Internet live stats. (2021). *Internet users*. https://www.internetlivestats.com/internet-users/
- Roser, M., Ritchie, H., & Ortiz-Ospina, E. (2021). Internet. Published online at OurWorldInData.org. https://ourworldindata.org/internet

ÚKOL 1 – ZAJÍMAVÉ ZDROJE

- **Česká republika – sociální sítě:**
- ČSÚ. 2020. *Osoby v ČR používající sociální sítě, 2020*. https://www.czso.cz/documents/10180/122362692/0620042051.pdf/a1a8dd54-2158-45bb-81ab-4953e1b2dd1e?version=1.1

- **Odborné studie:**
- Riberio, F. N., Benevenuto, F., & Zagheni, E. (2020). How Biased is the Population of Facebook Users? Comparing the Demographics of Facebook Users with Census Data to Generate Correction Factors. In *WebSci 20: 12th ACM Conference on Web Science*. https://dl.acm.org/doi/abs/10.1145/3394231.3397923
- Boas, T., Christenson, D., & Glick, D. (2020). Recruiting large online samples in the United States and India: Facebook, Mechanical Turk, and Qualtrics. *Political Science Research and Methods*, *8*(2), 232-250. https://doi.org/10.1017/psrm.2018.28
- Samuels, D., & Zucco, C. (2013). Using Facebook as a Subject Recruitment Tool for Survey-Experimental Research. Working Paper, *Social Science Research Network.* http://dx.doi.org/10.2139/ssrn.2101458

CÍLE SURVEYS

- **„Deskriptivní“ cíl** – popsat prevalence jevů v populaci, popsat průměry, mediány, mody, odchylky, popsat zastoupení typů respondentů v ohraničené populaci…
  - Často jen univariační analýzy (po jednotlivých proměnných,), ale mohou to být i vztahy mezi proměnnými
    - **To, co z nich dělá „desktriptivní“ cíl je snaha popsat nějak určenou, konečnou populaci**
  - Př. „8% dětí ve věku 10-16 let bylo v posledních 12 měsících terčem online útoků.“
  - „Lidé v ČR (1234) mají průměrný plat nižší než lidé v Německu (12345).“
  - „Průměrná síla vztahu mezi kritickým myšlením a dovedností poznat fake news je pro občany ČR rovna r = 0.35“
- **„Analytický“ cíl** (Koehler mluví o „causal inference“) – testujeme, co souvisí s čím, pro koho platí co apod. s **cílem popsat efekt**, který nějaké proměnná má na jinou proměnnou
  - Bivariační a multivariační analýzy
  - „Chlapci (OR = 2.58, $p < 0.001$) a starší děti (OR = 1.54, $p < 0.001$) jsou s vyšší pravděpodobností útočníky kyberšikany než dívky a mladší děti.“
- **Cíle a konkrétní VO ovlivňují, jaký vzorek a jak vybraný vzorek potřebuji pro zodpovězení VO**

TYPY VÝBĚRŮ

Pravděpodobnostní	Nepravděpodobnostní
Prostý náhodný	Příležitostný
Systematický náhodný + verze implicit stratification	Kvótní výběr
Stratifikovaný náhodný	Účelový výběr
Proporční stratifikovaný náhodný výběr	Referenční výběr
Neproporční stratifikovaný	Dobrovolný výběr
Klastrový náhodný	....

PRAVDĚPODOBNOSTNÍ VÝBĚRY

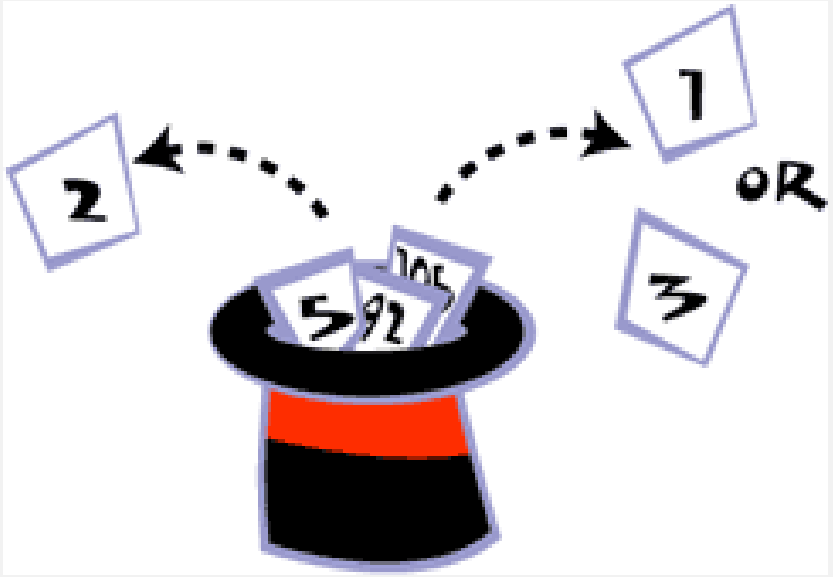
- Hlavní princip: elementy mají známou pravděpodobnost být vybrány do vzorku (osloveny)
- Nejlepší možný způsob výběru, pokud lze udělat (a je to finančně únosné)

- Předpokládá existenci/tvorbu výběrového rámce
  - Ať už je to seznam všech členů populace nebo na něj jdeme přes seznamy jiné (area frames, školy...)

PROSTÝ NÁHODNÝ

- Simple random sampling – základní náhodná technika
- Všechny elementy populace mají stejnou *p* na to být vybrány do vzorku (osloveny)
- **Equal probability of selection method (EPSEM)**
- Někdy se „vylosovaný“ element vrací zpět, ale doporučuje se to nedělat

- (elementy nemusejí být lidi – náhodný výběr platí i pro jiné než survey designy – např. náhodně vybereme komentáře pod statusem pro obsahovou analýzu apod.)



SAMPLING ERROR

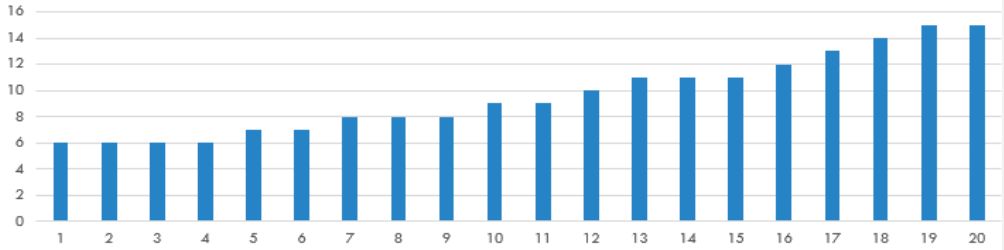
- Rozdíl mezi hodnotou vzorku a pravou hodnotou populace: **sampling error**
- I pokud vše měříme zcela přesně, statistiky z různých vzorků ze stejné populace budou přirozeně variovat (sampling variation) - přirozený důsledek toho, že máme vzorek a ne celou populaci

POPULACE

N = 20

$\bar{X}$ = 9.6, σ = 2.56

VAR = 6.55

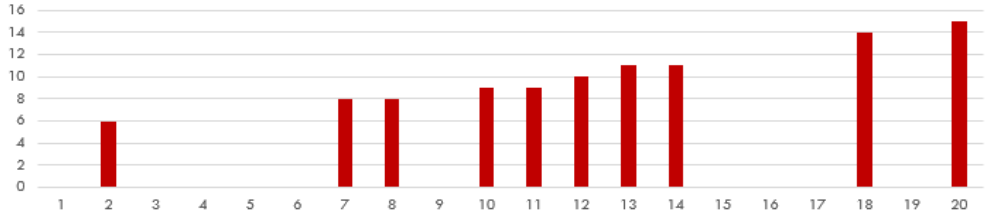


VZOREK 1

n = 10, M = 10.1

Biased VAR (děleno N) = 4.49, SD = 2.12

Unbiased VAR (děleno N-1) = 6.89, SD = 2.62

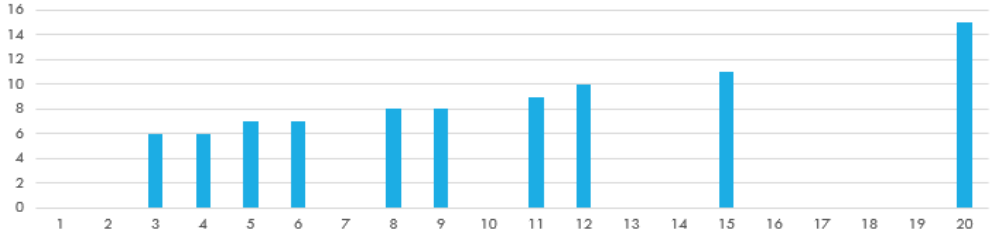


VZOREK 2

n = 10, M = 8.7

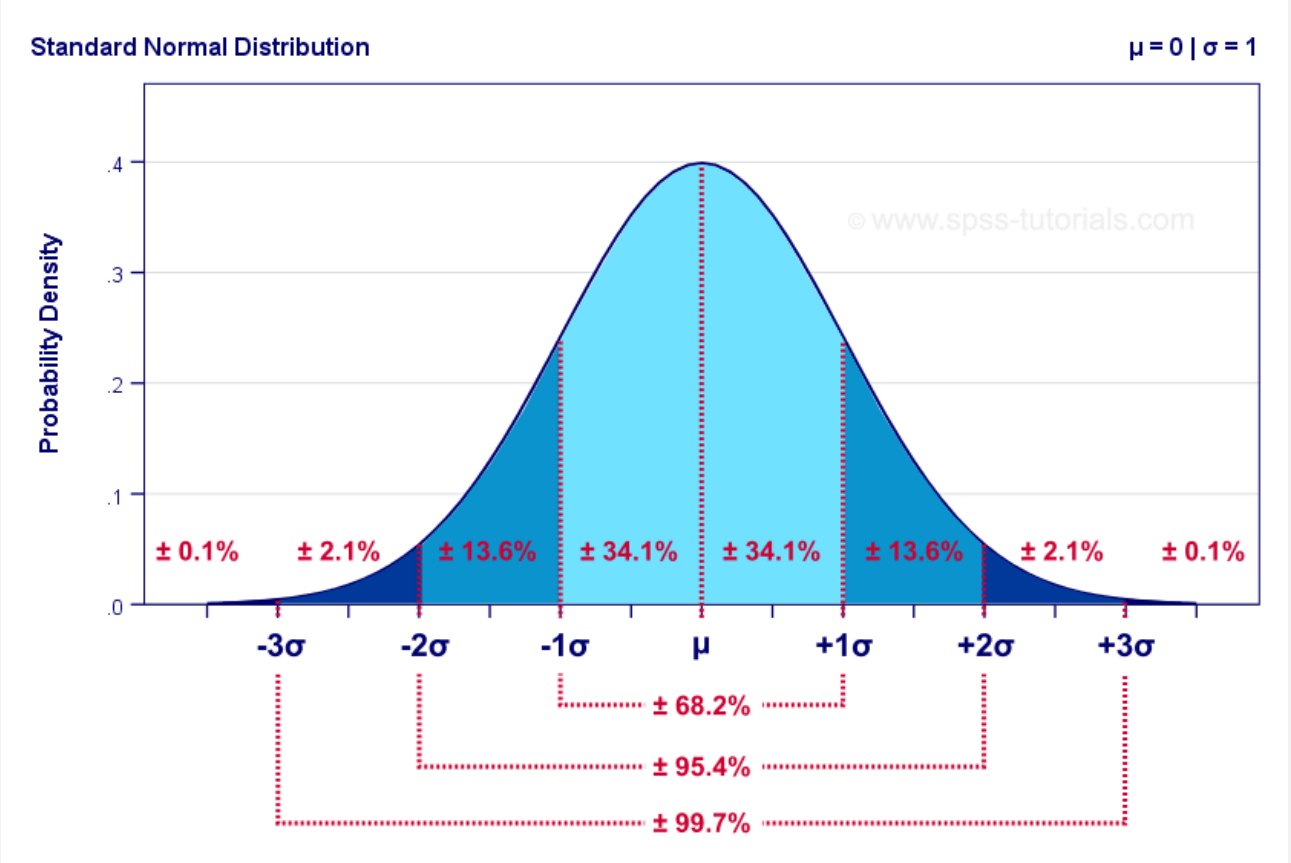
Biased VAR (děleno N) = 4.16, SD = 2.04

Unbiased VAR (děleno N-1) = 6.81, SD = 2.61

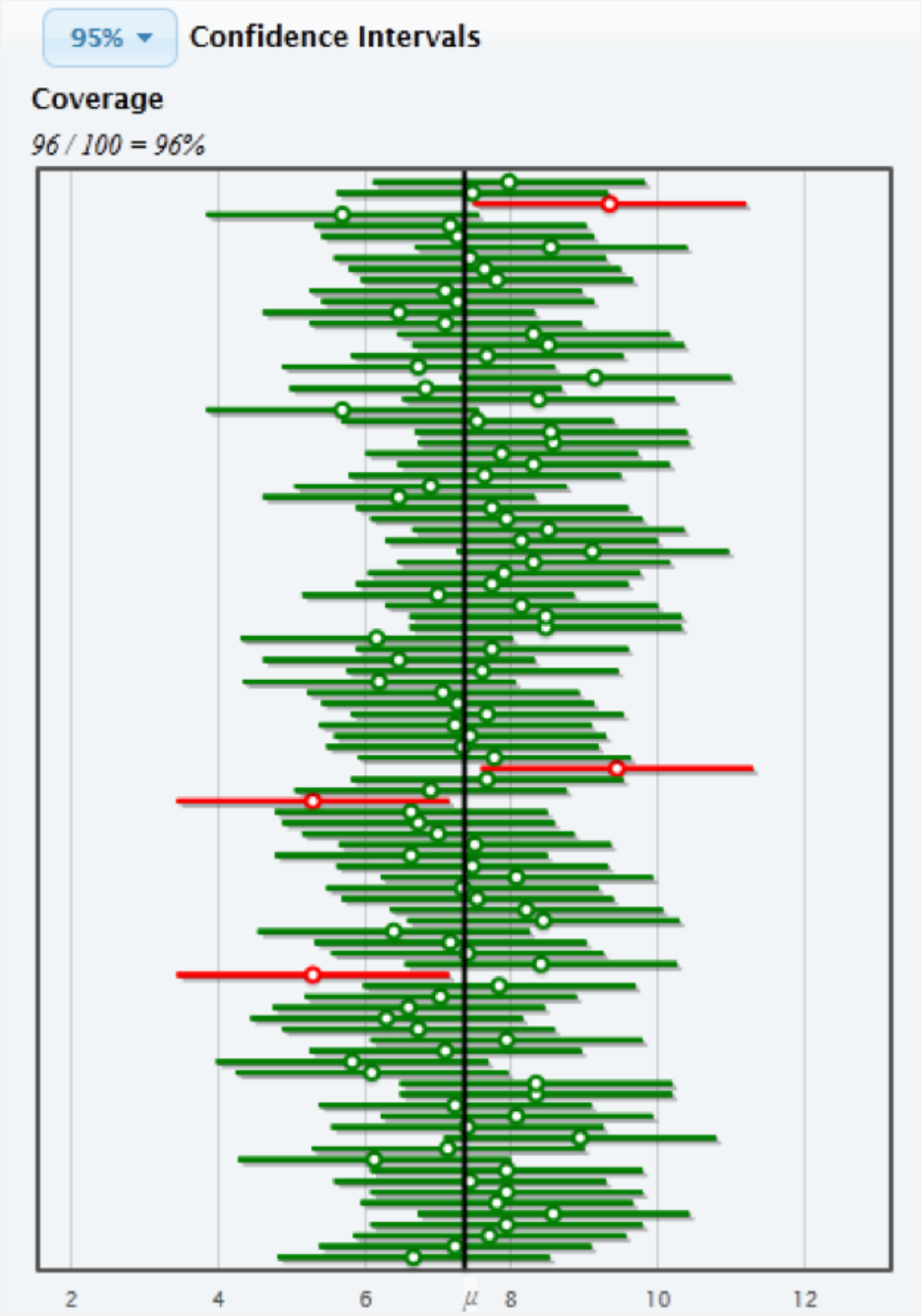


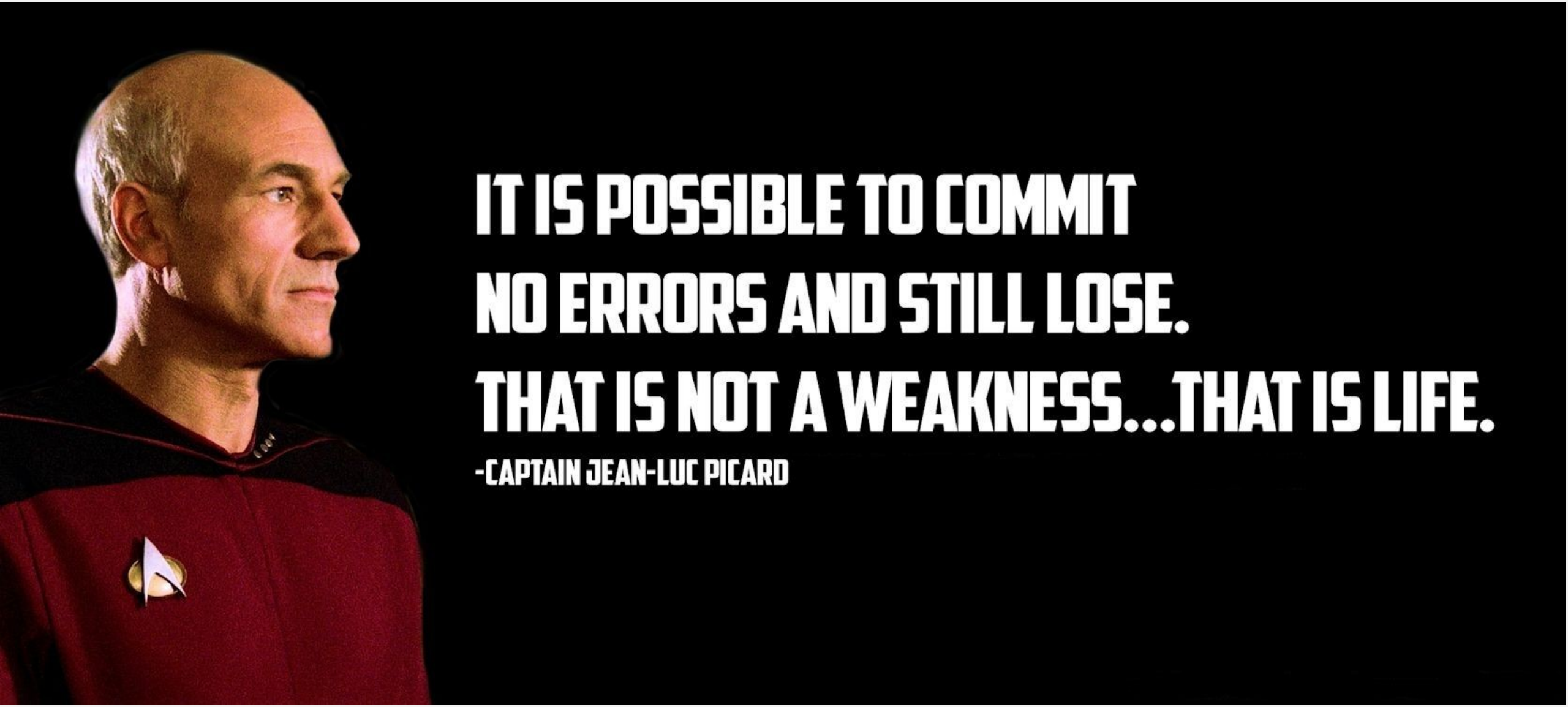
MARGIN OF SAMPLING ERROR

- Pokud bychom opakovaně vybírali stejně velké vzorky z populace a podívali se na rozložení statistiky, dostaneme **sampling distribution**, které při dostatečně velkých vzorcích nabývá tvaru normálního rozložení
- Pokud bychom dělali průměr z nekonečného opakování, pak sampling distribution of the mean
  - Průměr těchto průměrů by se rovnal průměru populace, jeho odchylka by pak ukazovala, jak moc jdou průměry jednotlivých vzorků vzdálené populačnímu průměru
  - Je to **teoretický koncept** – reálně nemáme data



- Protože nemáme data z teoretického nekonečného opakování, namísto „reálné“ odchylky průměru počítáme **standard error** of the mean (standard error, SE)
  - Používá se jiný výpočet s předpokladem normálního rozložení pro větší vzorky (>30) a t-rozložení pro menší vzorky, ale interpretace je stejná – jak moc průměry vzorků lítají kolem průměru populace
  - Velká SE (v porovnání s hodnotou průměru) znamená velkou variabilitu průměrů teoretických vzorků (náš vzorek neodpovídá dobře populaci)
  - Z SE se dále počítají intervaly spolehlivosti CI, confidence interval), typicky 95% nebo 99%
  - CI 95% = **pokud bychom dělali 100 náhodných vzorků z populace, pak v 95 z nich bude průměr populace ležet někde v jejich intervalu spolehlivosti**

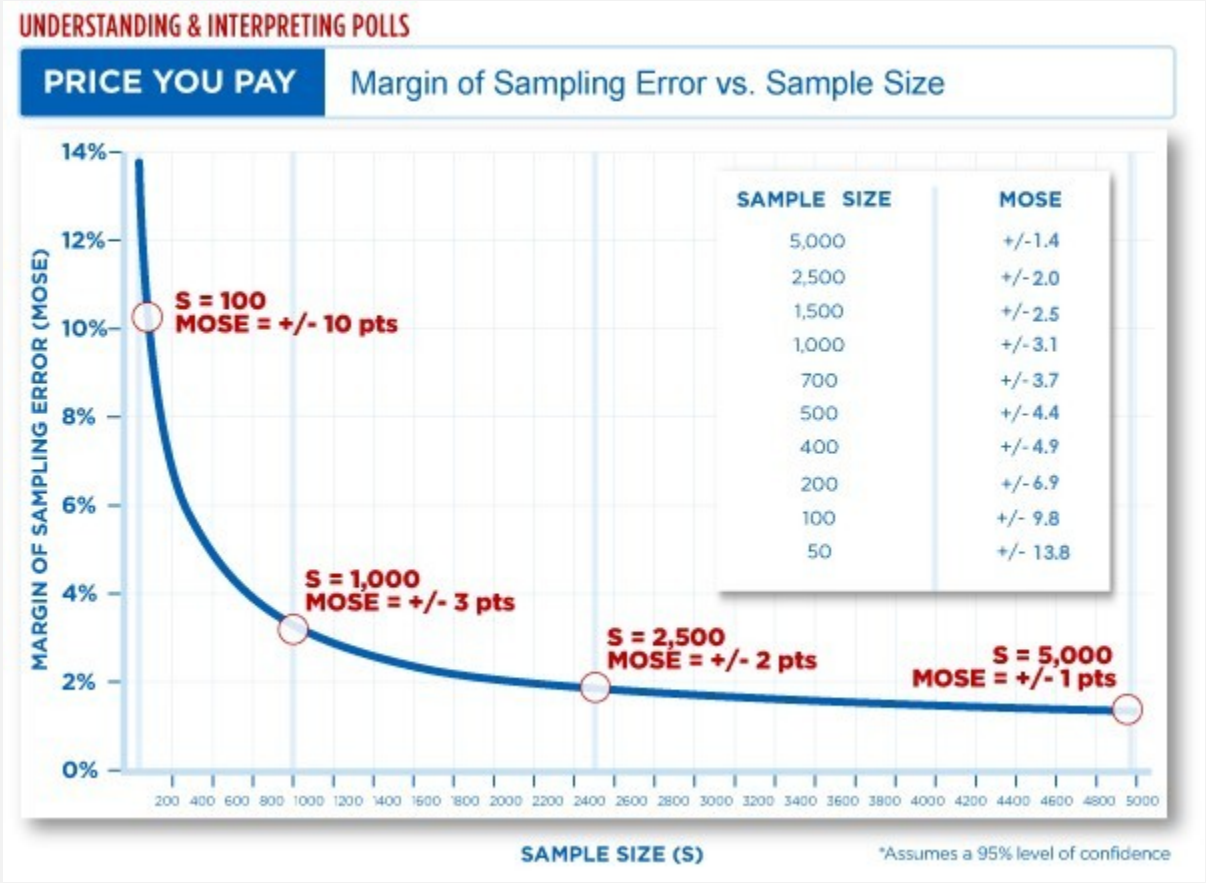



IT IS POSSIBLE TO COMMIT
NO ERRORS AND STILL LOSE.
THAT IS NOT A WEAKNESS...THAT IS LIFE.
-CAPTAIN JEAN-LUC PICARD

MARGIN OF SAMPLING ERROR

- Margin of (sampling) error: polovina intervalu spolehlivosti
- Míra preciznosti odhadu
- Point estimate: konkrétní hodnota statistiky vzorku (CI je +-MOSE kolem point estimate)
- S vyšším N užší interval

- Graf vpravo:
- **Předpoklad náhodného výběru vzorku**
  - U jiného výběru by se mu nemělo říkat „MOSE“
- Předpoklad vzorku, který není systematicky zkreslený
- CI 95%
- Proporce 0.5 (proporce: kolik lidí ve vzorku zadalo danou odpověď)
- Calculator: https://goodcalculators.com/margin-of-error-calculator/



VELIKOST VZORKU

- Za předpokladu pravděpodobnostního nezkresleného samplingu platí, že s vyšším počtem respondentů se zmenšují standard errors, zužují se intervaly spolehlivosti, máme vyšší sílu testu a méně chyb II typu u testování hypotéz (chyba II typu: nezamítnutí neplatné nulové hypotézy, čili když v datech nepodpoříme existující efekt)
  - → máme přesnější odhady efektů
- Benefit s nárustem vzorku do určité velikosti vzorku velký (viz předchozí graf)

- ALE: pro různé výzkumné cíle je potřeba brát v potaz mnohem víc faktorů, které ovlivňují počet respondentů, kterého bychom měli chtít ve vzorku dosáhnout

JAKOU VELIKOST VZORKU POTŘEBUJI?

- Čím přesnější chceme být v odhadu parametru populace, tím větší N
  - U některých výzkumů je vysoká přesnost zásadní
- Čím menší pravděpodobnost chyby II. typu chceme mít, tím vyšší N
- Čím vyšší statistickou sílu chceme a čím nižší hladinu významnosti pro zamítání hypotéz chceme, tím vyšší N
- Čím lepší měření máme, tím menší N stačí
  - Validita, reliabilita
- Čím slabší efekty očekáváme, tím vyšší N potřebujeme
  - Slabý efekt vs. silný efekt
  - V oblasti zkoumání mediálních účinků se obvykle setkáváme se slabými efekty – sledujte výsledky studií vašeho tématu, zapisujte si koeficienty ukazující na effect sizes (korelace, standardizované koeficienty v regresi, Cohenovo d, Cramerovo V,…)

JAKOU VELIKOST VZORKU POTŘEBUJI?

- Pokud je cílová populace hodně heterogenní (v tom, co nás výzkumně zajímá), potřebujeme vyšší N
  - Pokud jsou lidé v cílové populaci hodně podobní, stačí nám menší počet
- V různých typech samplingů je potřeba různě velká velikost vzorku
  - u stratifikovaného náhodného stačí menší N než u prostého náhodného
- Různé typy plánovaných analýz vyžadují různé N
  - K porovnání skupin je potřeba vyšší N než pokud nám jde jen o souhrnnou statistiku na všech respondentech
    - Pro smysluplné porovnání musí v každé skupině být alespoň nějaký počet lidí – pokud je některá skupina v populaci malička, musí se odhad potřebné velikosti vzorku odrazit od ní
  - Multivariační analýzy potřebují vyšší N než bivariační analýzy

Table 5.6 Number of research participants needed for small, medium, and large effect sizes at recommended power of 0.80 for $\alpha = 0.01$ and 0.05

	α					
	0.01			**0.05**		
Test	Small	Medium	Large	Small	Medium	Large
t test for two means[a]	586	95	38	393	64	26
Simple correlation (*r*)[b]	1,163	125	41	783	85	28
Analysis of variance[a]						
2 groups	586	95	38	393	64	26
3 groups	464	76	30	322	52	21
4 groups	388	63	25	274	45	18
5 groups	336	55	22	240	39	16
Multiple regression[b]						
2 predictors	698	97	45	481	67	30
3 predictors	780	108	50	547	76	34
4 predictors	841	118	55	599	84	38
5 predictors	901	126	59	645	91	42

Note: Effect size is the strength of relationship. Analysis of variance is used to compare two or more means for statistical significance. Multiple regression is used to predict or explain variance in a dependent variable using two or more independent variables (labeled "predictors" in table). Information from table was extracted from constructed by R. B. Johnson

[a] The sample size number is for each group. Multiply this number by the number of groups to determine the total sample size needed

[b] The sample size reported is the total sample size needed

Příklady potřebné velikosti vzorku pro různé analýzy a různě velké efekty

(Gideon, p. 71)

JAKOU VELIKOST VZORKU POTŘEBUJI?

- SW pro a priori výpočet velikosti vzorku: Gpower: https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower.html
  - Není úplně user friendly

- Doporučujeme:
  - Statistikou sílu (power) – 0.80 (doporučuje se, někdy ale i striktnější)
  - Alfa – 0.05, 0.01 nebo 0.001
  - Pro odhad effect size – ideálně máte odhad z existujících článků, pokud ne, předpokládejte raději slabé efekty

- Pokud plánujete víc analýz – potřebujete takový vzorek, který umožní nejnáročnější analýzu
  - Do projektu v tomto kurzu není potřeba řešit velikost vzorku, vyžaduje to mít už jasnou představu o analýzách a efektech

JAKOU VELIKOST VZORKU POTŘEBUJI?

- Velikost je dobré navýšit o předpokládaný úbytek v důsledku (unit/item)nonresponse a v důsledku čištění dat

- Naopak pro velikost vzorku nezáleží na velikosti cílové populace – pokud není hodně malá (tisícovky)
  - Většinou je cílová populace mnohem větší: děti používající internet, diváci ČT, obecná populace,…

- To vše platilo pro pravděpodobnostní vzorky, u nepravděpodobnostních je to obtížnější o nutnost nějak „započítat“ zkreslení dané výběrem
  - Zvážit, kteří lidé spíš neodpoví nebo se k nim ani nedostanu a co to dělá s podobou dat
  - Znamená to větší/menší heterogenitu? Víc homogenní soubor znamená mmj. nižší variabilitu dat (→ slabší efekty)
  - Budu mít zastoupené všechny skupiny dostatečně?
  - Pro nepravděpodobnostních vzorky u surveys je a priori sample size hodně orientační, ale přesto fajn

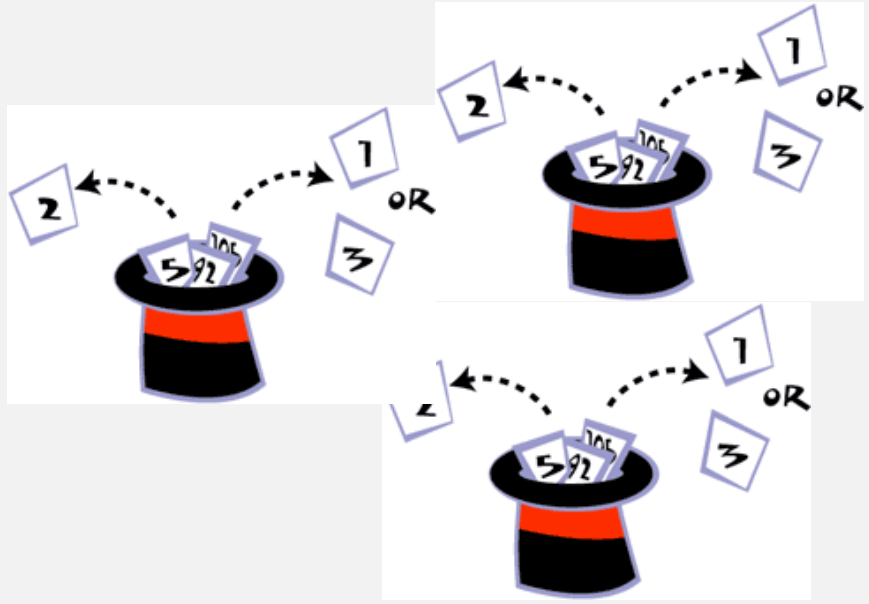
SYSTEMATICKÝ VÝBĚR

- Ze seznamu elementů populace se vybere náhodně počáteční bod a pak každý X-tý ve zvoleném intervalu
- Všichni mají stejnou pravděpodobnost, nicméně kombinace elementů stejnou *p* nemají

- Seznam může být řazený
  - Např. podle výše platu, podle pohlaví
  - V tom případě může být specificky systematický výběr lepší než prostý náhodný – méně hrozí varianta, kdy bychom náhodně zrovna někoho vynechali
    - Ale záleží na způsobu řazení – ne pokud je v seznamu nějaká periodicita (studenti ve stejně velkých třídách srovnáni podle prospěchu nebo podle genderu)
  - Implicit stratification nebo stratified-systematic sampling
- Pokud je seznam náhodně seřazený, pro odhad SE se používá stejná logika jako u náhodného prostého výběru

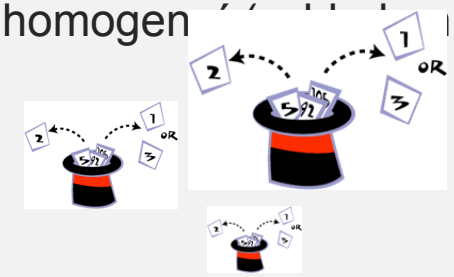
STRATIFIKOVANÝ NÁHODNÝ VÝBĚR

- Stratified (random) sampling
- Populaci rozdělíme do nepřekrývajících se skupin (vrstev, strat) a v rámci nich uděláme buď prostý náhodný nebo systematický náhodný výběr
- Stratification variable – proměnná, podle které jsou skupiny vytvořeny (nemusí být jen jedna)
  - Pohlaví, SES, volební preference, náboženské vyznání, vzdělání
  - Často to jsou základní sociodemografické proměnné



PROPORČNÍ STRATIFIKOVANÝ NÁHODNÝ VÝBĚR

- Proportionate stratification sampling, stratified sampling with proportional allocation
- Častější než stratifikovaný
- Při výběru v rámci strat udržujeme proporce – z každé vybíráme takový počet respondentů, aby celkový vzorek strukturou odpovídal populaci
- Pokud známe rozložení proměnné, podle kterých děláme straty, je to o něco efektivnější výběr než prostý náhodný – bude (častěji) dobře reprezentovat cílovou populaci v dané proměnné
  - A protože pořád využívá náhodný výběr, bude dobře reprezentovat i další proměnné
- Aby fungoval dobře, chceme mít straty co nejvíce homogen[illegible] k tomu, co nás zajímá)
- Je to pořád metoda EPSEM (equal probability)

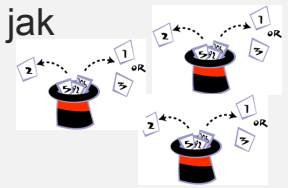


ČR, Stav k 31.12. 2019, ČSÚ

věk	N	%
10	122095	12.70
11	123576	12.85
12	119162	12.39
13	109443	11.38
14	103708	10.78
15	99048	10.30
16	95329	9.91
17	94886	9.87
18	94371	9.81

NEPROPORČNÍ STRATIFIKOVANÝ NÁHODNÝ VÝBĚR

- Disproportional stratified sampling
- Takový, kde záměrně volíme z každé straty počet respondentů, aby NEodpovídal rozložení v populaci
  - Např. u věku bychom chtěli rovnoměrné zastoupení, i když víme, že takové v populaci není
- Typické v případech, kdy chceme porovnávat skupiny respondentů a potřebujeme mít v každé skupině dostatečně velké N
  - Časté např. při výrazných disproporcích v populaci – počet otců, kterým je svěřena po rozvodu výhradní péče; ženy ve vedení velkých podniků; etnické minority,…
- Metoda není EPSEM (elementy nemají stejnou pravděpodobnost, některé mají vyšší, jiné nižší)
- Z takového vzorku pak nelze generalizovat na cílovou populaci bez použití vah
  - Pokud mám ve vzorku 50% lidí bez handicapu a 50% lidí s handicapem, jak vypovídající je průměrný příjem celého vzorku?
  - Lze ale generalizovat v rámci skupin



ČR, Stav k 31.12. 2019, ČSÚ

věk	N	%	Rovnoměrné %
10	122095	12.70	11.1
11	123576	12.85	11.1
12	119162	12.39	11.1
13	109443	11.38	11.1
14	103708	10.78	11.1
15	99048	10.30	11.1
16	95329	9.91	11.1
17	94886	9.87	11.1
18	94371	9.81	11.1

KLASTROVÝ SAMPLING

- Cluster sampling
- Náhodný výběr klastrů – skupiny obsahující víc elementů (doposud jsme náhodně vybírali přímo elementy, i když někdy stratifikovaně)
  - Školy nebo třídy, města, podniky, rodiny nebo domácnosti
- Vyžaduje větší N než prostý náhodný nebo stratifikovaný; při stejném N je náchylnější k většímu zkreslení (má větší výběrovou chybu)
- Výhodný v situacích, kdy jsou elementy (respondenti) např. geograficky vzdálení a sběr probíhá FtF – pak je výhodnější náhodně vybrat menší geografické celky a sběr provádět jen v nich
- Nebo, pokud nemáme existující rámec populace a musíme ho nejprve vytvořit, zatímco rámec pro klastry existuje (seznam škol na MŠMT)
  - Seznam elementů pak lze vytvořit jen pro náhodně vybrané klastry

KLASTROVÝ SAMPLING

- Jednoúrovňový (one-stage) – náhodný výběr klastrů a uvnitř nich vybrání všech elementů
- Dvou-úrovňový – náhodný výběr klastru a náhodný výběr uvnitř klastru
  - Pokud jsou klastry stejně velké, pak je tato metoda EPSEM
  - Pokud nejsou → probability proportional to size cluster sampling (PPS)
    - Pravděpodobnost výběru klastru je upravena, aby odpovídala jeho velikosti, tj. větší klastry mají větší pravděpodobnost být vybrány, což zaručí, že elementy uvnitř klastrů budu mít pravděpodobnost stejnou
- Multi-staged cluster sampling – více úrovní

CELKOVĚ K PRAVDĚPODOBNOSTNÍM VZORKŮM

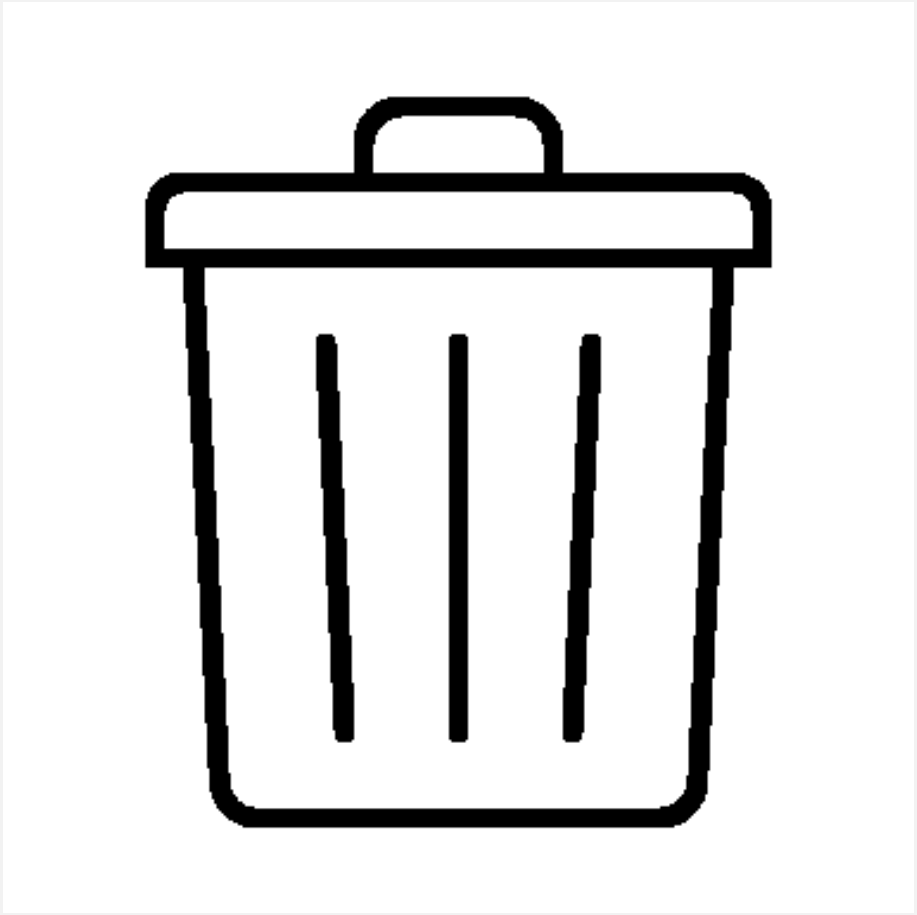
- Pravděpodobnostní vzorky jsou náročné a v realitě se velmi často sběry ideálu PS jen blíží
  - Často pravděpodobnost výběru elementů přesně neznáme, ale na základě jiných informací se odhaduje – pak jde o míru toho, kdy už daný odhad „za čarou“
  - I u pravděpodobnostních samplingů s dobrým rámcem se vyskytují jiné chyby (measurement, non-response), které mohou ve výsledku vést ke zkreslenému vzorku

- Nejčastěji se používají v různých kombinacích metod (straty, klastry, různé módy oslovování, vážení dat)
- Typicky u „velkých“ surveys: Current Population Survey (CPS), General Social Survey (GSS), EUKO IV (v ČR proportional stratified random clustered sampling), Eurobarometer, .. a pro výzkumy národních statistických úřadů

- U některých populací jsou de facto nemožné

- Pokud je naším cílem generalizovat nějakou hodnotu na populaci (deskriptivní cíl), pak to bez nich ale moc dobře neumíme
  - Můžeme mít kliku a trefit se
  - Ale nedokážeme odhadnout, jestli kliku máme nebo ne a jak daleko od pravé hodnoty populace můžeme být

NEPRAVDĚPODOBNOSTNÍ SAMPLINGY



KVÓTNÍ VÝBĚR

- Identifikace charakteristik, které chceme mít ve vzorku zastoupeny v nějakém poměru a následně hledání respondentů pomocí příležitostného výběru, dokud nejsou kvóty naplněny
  - Tím se snaží dosáhnout reprezentativity vůči dopředu daným proměnným
  - Kvóty – důležité charakteristiky vzhledem k VO, většinou proporční zastoupení i pro jejich kombinace
  - Typicky sociodemografické údaje: pohlaví, věk, velikost bydliště, ale i specifičtější – klienti předem vybraných bank, lidé s různou preferencí alternativních médií..
  - Kategorie mohou být velmi podobné (i stejné) jako u stratifikovaného výběru – ten ale využívá náhodný výběr
  - Vzpomeňte si na to, proč velké agentury mylně předpověděly prohru Trumana v 1948 (selection bias)

- Ani z kvótního vzorku není možné jednoduše generalizovat, byť je lepší než příležitostný bez kvót
  - Pro různé VO různě ne/víme o proměnných, které by bylo dobré mít ve vzorku „správně“ zastoupené
  - V sociálních vědách často děláme více analýz na jednom dotazníkovém šetření (i DP) – potenciálních kombinací je příliš

ÚČELOVÝ/ZÁMĚRNÝ VÝBĚR

- Purposive sampling
- Nenáhodný výběr specifické populace
- Do vzorku hledáme respondenty podle nějaké klíčové charakteristiky
  - Časté u klinických (lidé s konkrétní diagnózou) studií
  - Nebo třeba účelově chceme jen voliče SPD, učitele informatiky nebo lidi s nějakou zkušeností (oběti kyberšikany)
- Opět nejde o EPSEM → není možná generalizace (ani na danou specifickou populaci)

REFERRAL SAMPLING

- Typické pro obtížně dosažitelné populace, které zároveň mají vzájemný kontakt

- **Snow-ball sampling** – dobrovolní účastníci studie jsou požádání o kontakty na další vhodné respondenty
  - Podle provázanosti skupin se může stát, že vzájemné kontakty po čase vedou jen na členy stejné sociální bubliny a nedostanou se z ní ven

- **Network sampling** – začíná s pravděpodobnostním vzorkem lidí, u kterých očekáváme, že mají napojení na naši chtěnou populaci nebo do ní sami mohou spadat
  - Screening, zda do ní spadají + žádost o kontakt, pokud takové lidi znají
  - Měl by vést k širšímu dosahu než snowball

CONVENIENCE SAMPLING

- **Příležitostný výběr** – výběr lidí, kteří jsou zrovna k dispozici
  - Oslovení spolužáků, lidí na vyplnto.cz, skupin na FB,…
- Často nevíme nic o tom, z jakého rámce populace byli respondenti vybráni, nelze spočítat response rate, elementy nemají známou šanci být do vzorku vybráni → nelze generalizovat na populaci
- Pro surveys je to podle mnohých akademiků zcela nevhodný typ samplingu
  - Pro jiné designy je častý, např. experimenty

DALŠÍ TYPY

- Volunteer sampling
  - Respondenti sami se hlásí do studie, o které se někde dozvědí, a účast je jim umožněna
  - Generalizace je vyloučena☺ , obří zkreslení vlivem toho, kteří lidé se o studii dozvědí + zájmu o téma

- River sampling
  - Při online rekrutaci – upozornění na studii na některých stránkách (k pozvánce se tak dostanou jen ti, kteří zrovna chodí na dané stránky)

- Router sampling
  - Respondenti jsou pozváni obecnou nabídkou zapojit se do výzkumu a až po poskytnutí základních screeningových údajů jsou navedeni k výzkumu, pro který splňují kritéria

DALŠÍ TYPY

- Online panely (AAPOR má skvělý report o online panelech: Baker et al., 2010)
  - Databáze lidí, kteří souhlasí s tím být oslovování k výzkumům
  - Velké rozdíly podle toho, jak daný online panel vzniká (rekrutace)
    - Opt-in panely (mTurk, některé panely profesionálních agentur)
    - Nabídka zapojení lidem rekrutovaným offline pro jiné výzkumy (profesionální agentury)
    - Účelové oslovování lidí ze segmentů chybějících v panelu pro doplnění panelu
    - Nejrůznější kombinace a ad hoc dohledávání respondentů podle konkrétních zakázek

NEPRAVDĚPODOBNOSTNÍ VÝBĚRY CELKOVĚ

- Jsou méně náročné (až supereasy), ale generalizace (zvlášť pro deksriptivní cíle) je problematická
- Ale:
- Lze najít doporučení, jak je sestavovat tak, aby co nejvíc reprezentovaly populaci (např. matching; typicky se týká kvótních výběrů), navržené úpravy pro výpočet standard error a články, které hodnotí míru úspěšnosti daného postupu
  - S použitím dobrých vah a minimalizací jiných chyb se pak výsledný vzorek může reprezentativnímu vzorku blížit a mít podobnou možnost generalizace jako PS (je to ale obtížné)
  - I pokud máme vzorek z PS zkreslený jinými chybami, vede pořád častěji k lepšímu odhadu parametru populace než convenience sample

- Over long-run – replikace na různých nepravděpodobnostních vzorcích poskytují pool výsledků, které nám o pravé hodnotě v populaci mohou vypovídat

CÍLE SURVEYS A KTERÉ VZORKY JSOU OK

- **„Deskriptivní“ cíl** – popsat prevalence jevů v populaci, popsat průměry, mediány, mody, odchylky, popsat zastoupení typů respondentů v ohraničené populaci…
  - To, co z nich dělá „desktriptivní“ cíl je snaha popsat nějak určenou, konečnou populaci

- Výběr vzorku a systematická zkreslení (vzhledem k tomu, co nás zajímá) jsou zde obrovsky důležité – čím preciznější odhad chceme, tím více
  - Bez PS (pravděpodobnostní sampling) to v podstatě není možné
  - Jedině, pokud je cílová populace ve zkoumané charakteristice velmi homogenní – pak by stačil i NPS, protože způsob výběru nemá co zkreslit☺

CÍLE SURVEYS A KTERÉ VZORKY JSOU OK

- **„Analytický“ cíl** – testujeme, co souvisí s čím, pro koho platí co apod. s cílem popsat efekt, který nějaké proměnná má na jinou proměnnou
  - I zde bychom ideálně chtěli PS; mnohdy je nemožný
- **Kdy stačí NPS:**
- Pokud máme hodně homogenní populaci vzhledem ke zkoumanému vztahu - např. předpokládáme, že efekt alkoholu na kognitivní schopnosti je pro všechny stejný
- Pokud zkoumaný vztah variuje mezi skupinami populace, můžeme zkoumat jen subpopulaci → tím si vytvoříme homogenní vzorek – např. jen muži se stejným BMI (role teorie)
  - Vytvořená sub-populace musí být pořád výzkumně dostatečně zajímavá sama o sobě
- Totéž můžeme ošetřit tím, že dané proměnné (pohlaví a BMI) přidáme do analýzy jako moderátory
  - Pokud by byl efekt v různých skupinách stejný, máme větší jistotu, že naše data je možné (pořád s nějakou opatrností) generalizovat

- **Pozor: v soc. vědách je ale proměnných, které ovlivňují efekty, celá řada**
  - Čím může být ovlivněna dovednost poznat fake news kromě kritického myšlení?
  - TEORIE!

- Opáčko: Systematická zkreslení nastávají, pokud proměnná, která ovlivňuje účast ve výzkumu, koreluje s proměnnými, které analyzuji
  - Pokud uděláme výzkum na studentkách VŠ, ovlivní to míru efektivity hormonální antikoncepce?
  - Pokud nám chybí respondenti, kteří nevyužívají internet, ovlivní to sílu vztahu mezi extraverzí a počtem přátel?
  - Jenže o hodně vlivech patrně ani nevíme, často se i obtížně odhaduje, které proměnné ovlivnily to, kdo se zapojil do výzkumu (nonresponse) a často také nevíme, jaké rozložení daná proměnná v naší populaci má
  - Proto se nejčastěji opíráme o základní sociodemografické proměnné (vč. SES), přičemž ale známe hodně jejich korelátů

CO TO ZNAMENÁ PRO DIPLOMKY?

- Ve většině případů nebudete mít vzorek, který by umožňoval adekvátně zodpovědět deskriptivní cíle
- NESTANOVUJTE SI DESKRIPTIVNÍ VÝZKUMNÉ OTÁZKY
  - (i když popisné statistiky samozřejmě použijete pro popis vašeho vzorku)

- Kdy se může generalizace na cílovou populaci podařit?
- Specifická cílová populace, která i vám umožňuje sampling, který povede k repre vzorku, tj. buď varianta náhodného výběru jednotek analýzy nebo cenzus
  - stáhnete všechny příspěvky z Twitterových účtů politiků v PS a z nich náhodně vyberete ty, které budete analyzovat
  - analyzujete všechny články, které obsahují nějaké klíčové pojmy
  - populace uživatelů konkrétní služby, k jejichž seznamu máte přístup
- Hodně homogenní populace vzhledem k VO
  - ….?

CO TO ZNAMENÁ PRO DIPLOMKY?

- Pokud budete mít nějakou verzi nepravděpodobnostního výběru, pak:
  - Je potřeba detailní popis toho, jak probíhala rekrutace respondentů – jen tak lze alespoň částečně zhodnotit, o jaké populaci nám výsledky (možná) něco vypovídají
    - Pro online výzkumy: cherries (Checklist for Reporting Results of Internet E-Surveys) https://geriatricare.files.wordpress.com/2009/12/cherries-lijst.pdf
  - Pečlivě zvažte (a změřte) kovariáty (proměnné, pro které kontrolujete – o které „čistíte“ vztah mezi zkoumanými proměnnými) a moderátory (podmínky – očekáváme, že daný vztah proměnných je stejný u všech nebo má být jiný pro různé skupiny lidí?)
  - Zvažte kvóty – Qualtrics je umí nastavovat a kontrolovat
  - Přemýšlejte o tom, kdo měl šanci se k vašemu výzkumu dostat (coverage bias, selection bias) a kdo na něj patrně častěji pozitivně reagoval a výzkum vyplnil (nonresponse bias)
  - **Věnujte se tomu v diskuzi!**
    - **Limity vzorku a jejich (pravděpodobné) dopady na výsledky nesmí chybět**
    - Diskutujte i to, jak by vaše výsledky asi vypadaly v případě „lepšího“ vzorku (jak by vypadaly, kdybyste měli dobrý repre vzorek s využitím PS?)

CO TO ZNAMENÁ PRO CHÁPÁNÍ VÝZKUMŮ/REPORTŮ?

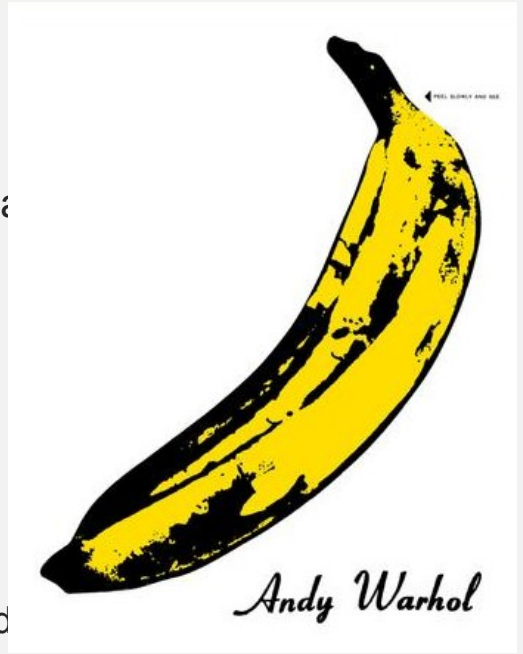
- **Nenechte se ohromit velkým N – samo o sobě vůbec nic neznamená**
- Pokud studie/report má hlavní cíl deskriptivní, pak convenience sample je v sociálních vědách prakticky vždy nevhodný
  - Procenta, rozložení jednotlivých proměnných – jsou extrémně citlivé na kvalitu vzorku
- U dobrých výzkumů je sampling vždy popsán – dohledávejte si informace, hodnoťte si sami, co z takových dat lze/nelze zobecnit na koho
  - Berte v potaz nejen sampling, ale všechny chyby (total survey error)
- „Reprezentativní“ vzorek „*vzhledem ke konkrétním indikátorům“* lze získat i jinak než pravděpodobnostním výběrem
  - Repre se ne vždy rovná PS

VĚTŠÍ CVIČENÍ

- Chcete zjistit, jaké % populace prodělalo COVID-19 a naplánujete studii, kam mohou chodit dobrovolníci a dozvědí se o ní z médií
  - Co je to za typ výběru vzorku?
  - Jaké bude selection bias? (chyba v důsledku toho, jak/koho oslovuji)
  - Jaké bude non-response bias? (chyba v důsledku toho, kdo se zúčastní)
  - Pro jakou jinou VO by takový vzorek mohl fungovat?
  - Jak byste naplánovali sampling jinak, abychom prevalenci odhadli s vyšší jistotou?

ÚKOL 2

- Pan Novák je populistický politik, jehož rétorika se často obrací na „staré zlaté časy“, kdy byl svět jednodušší, a místy má tendenci sklouzávat ke konspiračním teoriím. Na FB má cca 125 000 sledujících.
- Protože by se rád stal ještě populárnějším, chtěl by zjistit, co si voliči v ČR myslí o vybraných tématech.
  - Možnost manželství stejnopohlavních párů
  - Vnímaná závažnost COVID-19
  - Odkud správně loupat banán
- A protože je šetřivý, nechce za takový výzkum utratit ani korunu. Rozhodne se proto zveřejnit otázky na da témata na svém FB profilu.
- Váš úkol je popsat:
  - Kdo je cílová populace?
  - O jaký typ výběru vzorku se jedná?
  - K jakých chybám takový výběr vede?
  - Jak se tyto chyby patrně projeví u každého tématu (k jakým zkreslením vzhledem k tématu asi dojde)?
  - Navrhněte jinou strategii, jak by bylo možné získat vzorek voličů ČR, který by vedl k menším zkreslením (pořád ale nechcete utratit vůbec nic).
    - Pokud je potřeba, strategie může být různá pro různá témata
  - Navrhněte strategii, pokud by se pan Novák zajímal jen o názory svých voličů.



LITERATURA

- Baker, R., Blumberg, S. J., Brick, J. M., Couper, M. P., Courtright, M., Dennis, J. M., ... Zahs, D. (2010). Research synthesis: AAPOR report on online panels. Public Opinion Quarterly. https://doi.org/10.1093/poq/nfq048
- Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., ... & Tourangeau, R. (2013). Summary report of the AAPOR task force on non-probability sampling. *Journal of survey statistics and methodology*, *1*(2), 90-143.
- Gideon, L. (Ed.). (2012). *Handbook of survey methodology for the social sciences*. New York: Springer.
- Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey Methodology* (Vol. 561). John Wiley & Sons.
- Kohler, U. (2019). Possible uses of nonprobability sampling for the social sciences. *Survey Methods: Insights from the Field*, 1-12.
- Kohler, U., Kreuter, F., & Stuart, E. A. (2019). Nonprobability sampling and causal analysis. *Annual review of statistics and its application*, *6*, 149-172.