ZURN6311 DOTAZNÍKOVÝ VÝZKUM: SAMPLING 1

Lenka Dědková

ZÁVĚREČNÝ PROJEKT

- Mediální projekt: překryv v popisu projektu pro většinu žádný
- Výzkumný projekt: **první krok** před konkrétním plánováním
  - Zaměření na opodstatnění VO, hypotéz, což je zcela zásadní pro adekvátní výběr výzkumného designu a jeho plánování
  - V případě dotazníkového šetření je popis metody v projektu jen malá část (ne klíčová)

- Práce během semestru v našem kurzu: z velké části bude obecnější (nezaměřená na váš projekt), je ok pokud se k reálné VO pro vaši DP dostanete až koncem semestru

SAMPLING

- Jaké jsou typy samplingu
- Jak se realizují různé typy samplingu
- Jak se počítá výběrová chyba
- Velikost vzorku
- Jaké jsou zdroje chyb u samplingu

PROCES DOTAZNÍKOVÉHO ŠETŘENÍ

PROCES DOTAZNÍKOVÉHO ŠETŘENÍ

- Dá se dělit i jinak:
  - na část měření a odpovídání (dotazník jako nástroj)
  - na část vzorku (kdo reprezentuje cílovou populaci)

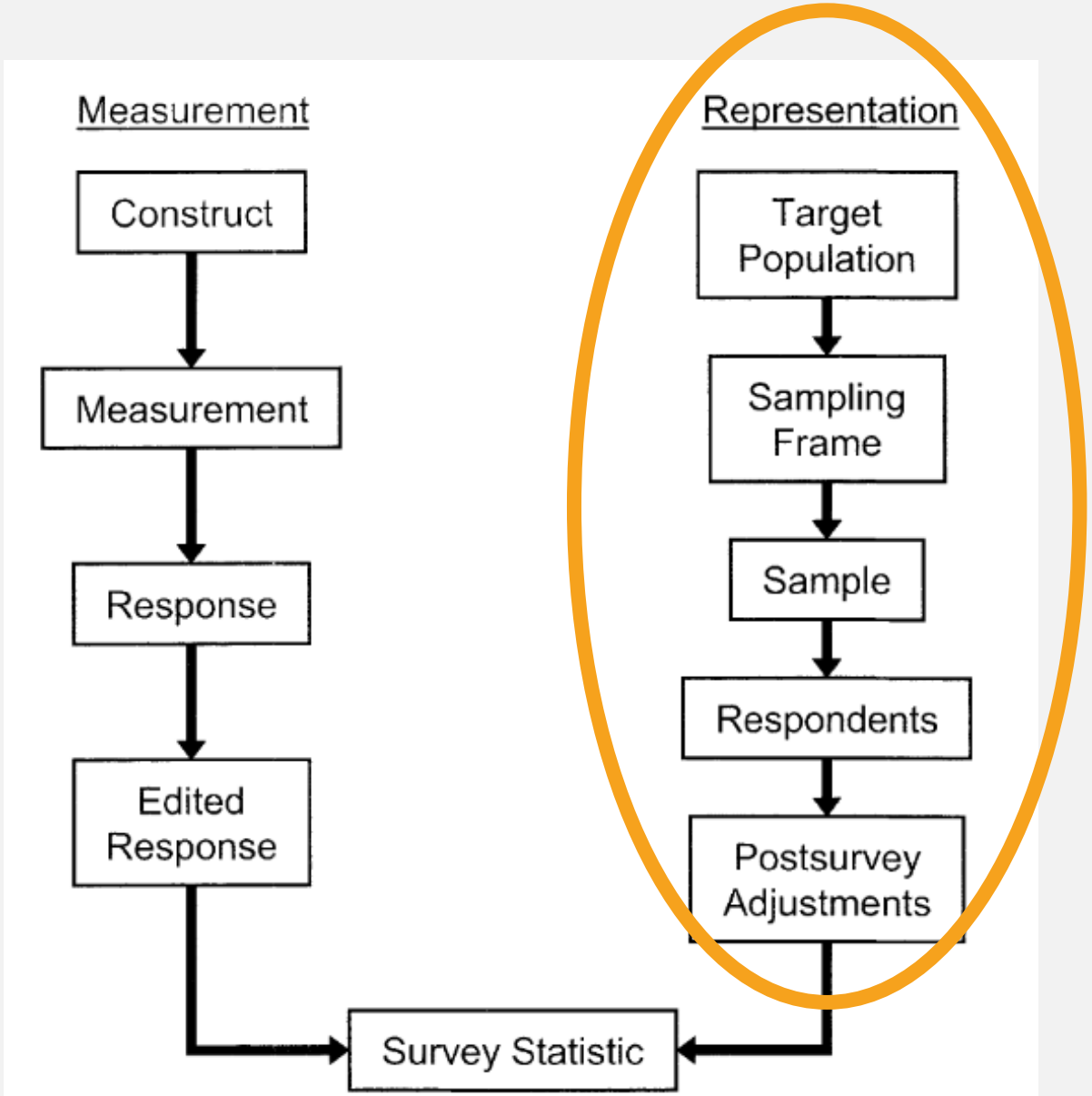


Figure 2.2 Survey lifecycle from a design perspective.

Groves et al.,

POJMY

- **(Target) population, (cílová) populace** – ti, o které nám reálně jde, o kom se výzkum snaží něco říct (generalizovat výsledky na ně)
- **Sampling** = způsob výběru respondentů, zahrnuje vlastně vše
- **Sample, vzorek** – koho měříme; menší část cílové populace, která má vést k tomu, že díky ní budeme schopni něco říct o cílové populaci
  - Ve statistice: **N** = velikost vzorku, **n** = velikost nějaké části vzorku (N = 521, ženy n = 125)
  - V literatuře o samplingu ale často **N** = velikost cílové populace a **n** = velikost vašeho vzorku
  - A jako sample se někdy označuje „wishful“ sample – lidi, které oslovím a ne jen lidi, kteří pak reálně dotazník vyplní (u pravděpodobnostního výběru)
- **Elements, prvky** – jednotliví respondenti (jevy, případy), kteří vás zajímají
  - Cílová populace je složena z **N** prvků, váš vzorek z **n** prvků
  - Někdy „sampling unit“
- **Sampling frame, výběrový rámec populace** – elementy populace, ze kterých vybíráme vzorek

POJMY

- **Reprezentativní vzorek** – takový vzorek, který je podobný cílové populaci ve všech aspektech (jen je menší)
- **Parametr populace** – hodnota měřeného jevu (nebo vztahu) v cílové populaci, označují se typicky řeckými písmeny
- **Statistika vzorku** – hodnota měřeného jevu (nebo vztahu) ve vzorku
  - V datech máme statistiku, ze které odhadujeme parametry populace (estimations)
  - Průměrná úroveň a směrodatná odchylka mediální gramotnosti, korelace mezi věkem a občanskou participací, regresní koeficient…
  - **Rozdíl mezi parametrem a statistikou** – error, chyba
    - chyby vznikající u samplingu – sampling error, coverage error, selection error

CÍLOVÁ POPULACE

- Kdo nás vlastně zajímá? O kom se chceme našim výzkumem něco dozvědět?
  - Obecná populace dospělých v ČR
  - Studenti FSS
  - Zaměstnanci firmy *We‘re all happy*
  - Účastníci demonstrace
  - Klienti sociálních služeb
  - (ne vždy jen lidi, ale jsme v survey)

- **Každý: představte, na co se zaměřuje vaše DP a kdo je vaší cílovou populací**

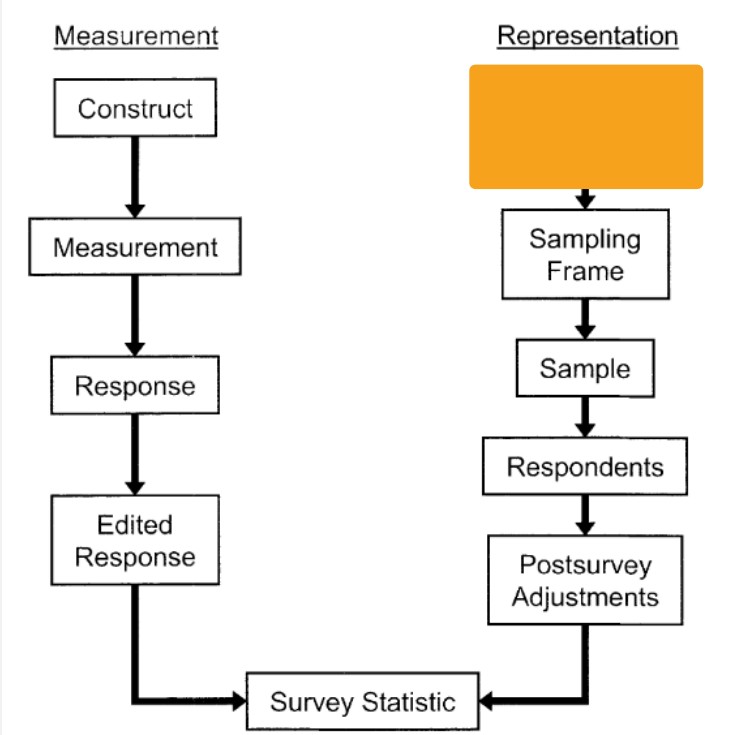


Figure 2.2 Survey lifecycle from a design perspective.

CÍLOVÁ POPULACE

- **Pořád musí být dost specifický popis**
  - Kdo je „obecná populace dospělých v ČR“?
  - Patří sem jen lidé s ČR občanstvím? Nebo lidé s trvalým/přechodným pobytem v ČR?

- **Population specification error**: pokud špatně definujeme cílovou populaci
  - Např. vaším cílem je zjistit spokojenost studentů mediální výchovy s obsahem předmětu, ale jako populaci si zvolíte učitele
  - Je to víc otázka teorie a VO než samplingu (budete řešit ve Výzkumném projektu)

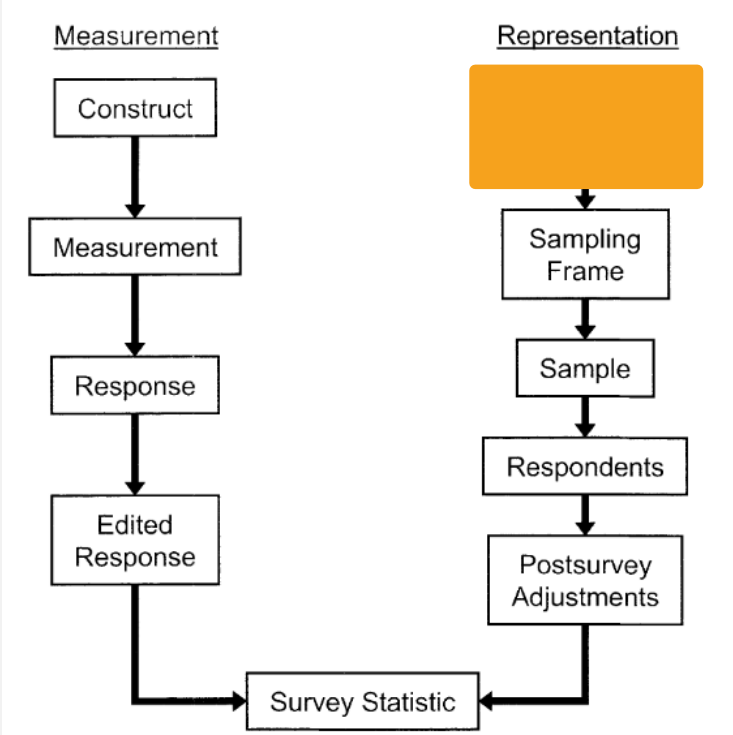


Figure 2.2 Survey lifecycle from a design perspective.

VÝBĚROVÝ RÁMEC

- Prvky cílové populace, které mají šanci být do výzkumu vybrány
- Ideálně máme seznam všech členů populace
  - Např. zaměstnanci České pošty
  - často ne
- **Coverage**: to, jak se výběrový rámec překrývá s cílovou populací, ideálně zahrnuje všechny právě jednou a nikoho navíc
  - Cílová populace = rámec

- **Zkuste se zamyslet nad tím, ke kterým cílovým populacím existuje kompletní výběrový rámec se seznamem**
  - A ke kterým byste se mohli dostat vy?

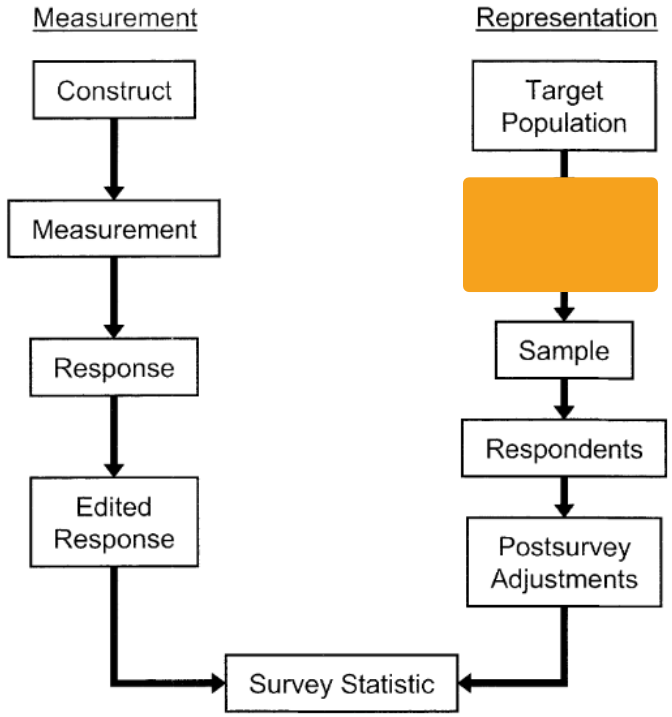


Figure 2.2 Survey lifecycle from a design perspective.

JAK VZNIKÁ VÝBĚROVÝ RÁMEC

- Cílová populace se často mění
  - Lidé se stěhují, stárnou, přestávají/začínají být klienty, žáci přecházejí ze školy na jiné
  - Obvykle řešíme cílovou populaci v nějakém vymezeném čase, přestože z ní někdy chceme generalizovat nad ní
    - Klienti krizové linky v období 1.1. 2019-31.12. 2019
    - Studenti s nepřerušeným prezenčním studiem v JS 2020
    - Lidé s trvalým pobytem na území během celého roku 2020
    - Komentáře pod twitteroým účtem ke dni 9.3.2021

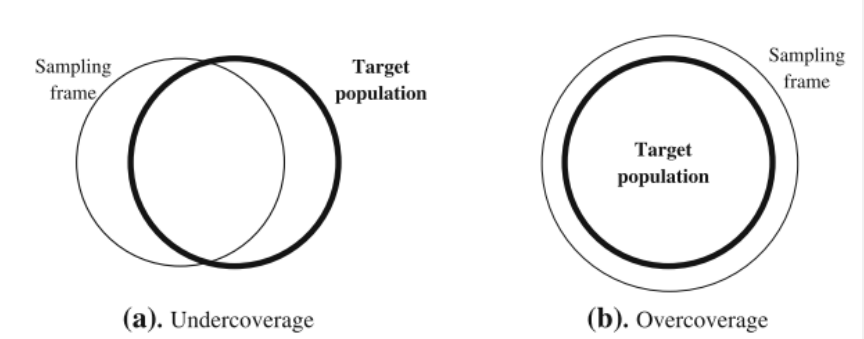
- **Hlavní cesty:**
  - Tvoříme (nebo získáme existující) kompletní seznam všech elementů populace (náhodný prostý a stratifikovaný výběr)
  - Tvoříme seznam jen náhodně vybraných sub-částí populace (náhodný klastrový výběr; vyžaduje seznam nadřazených celků)
  - Netvoříme seznam elementů jako takový, ale použijeme náhodné oslovování bez něj (znemožňuje spočítání coverage error)

COVERAGE ERROR

- **Undercoverage** – výběrový rámec nezahrnuje všechny elementy populace
- **Overcoverage** – zahrnuje i některé elementy mimo cílovou populaci

- Další nedokonalosti v seznamu
  - **Clustering**: jeden záznam ve výběrovém rámci vede k více elementům (např. samplujeme přes adresy – na jednom místě bydlí víc lidí)
  - **Duplication**: jeden element cílové populace je v rámci zastoupen vícekrát (někteří lidé mohou bydlet na víc místech)

- Tyto chyby dohromady = **coverage error** (část celkového rozdílu mezi statistikou vzorku a parametrem populace, který je způsoben výběrovým rámcem)



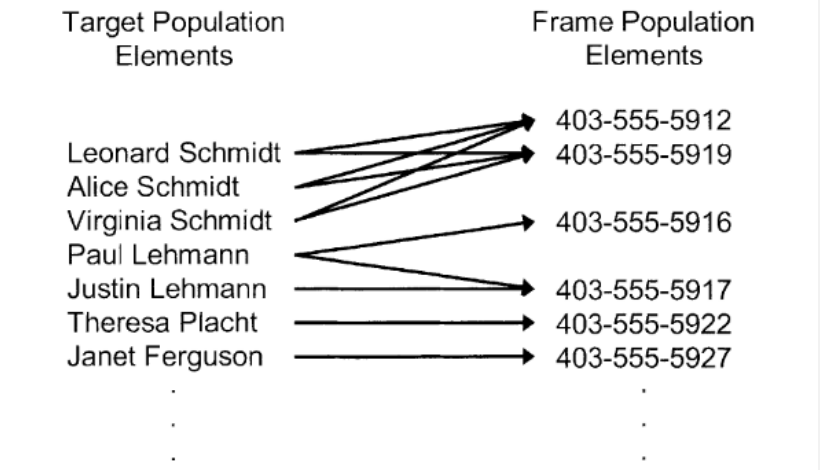


Figure 3.3 Clustering and duplication of target population elements relative to sampling frame elements.

JAK VZNIKÁ VÝBĚROVÝ RÁMEC

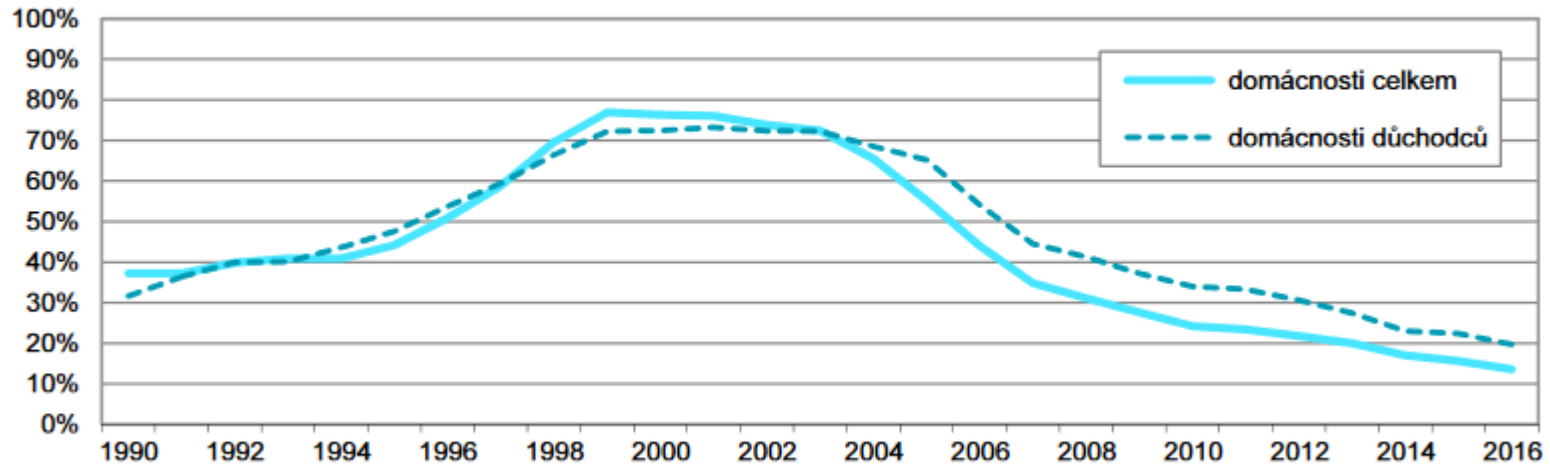
- Časté rámce pro obecnou populaci nebo domácnosti
- **Area frames** – geografické jednotky
  - Lze udělat seznam adres bytových prostor v každé (nebo jen náhodně vybrané) jednotce, z nichž se následně vybírá domácnost/jednotlivec v domácnosti
  - Pošta – papírový dotazník, obvykle i možnost vyplnit online
  - Random walk – náhodná procházka – tazatel dostane přiřazenou geografickou jednotku, počáteční bod (seed) a náhodně generovaný postup, na jaké adrese má oslovit respondenta
  - Potíže: clustering (víc osob v jedné domácnosti), undercoverage (pokud nějakou adresu mineme), duplication (pokud má osoba víc bydlišť)

- **Telephone frames** – typicky pro pevné linky
  - Zprvu telefonní seznam, rychle ale random digit dialing (RDD)
  - Jaká zkreslení tady budou?

- https://www.czso.cz/documents/10180/94876554/06202618b.pdf/34ed3250-c757-4a89-bbc4-a89e3f245d73?version=1.2

Domácnosti s pevnou telefonní linkou

Graf B1 Podíl domácností v Česku s pevnou telefonní linkou (%)



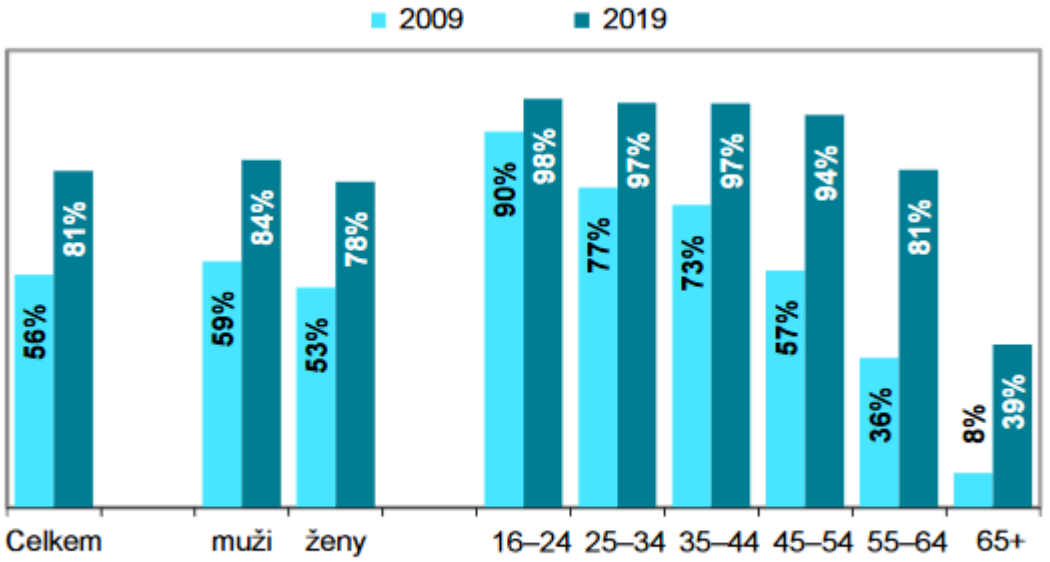
Zdroj: ČSÚ 2018, Statistika rodinných účtů

JAK VZNIKÁ VÝBĚROVÝ RÁMEC

- **Internet jako rámec pro obecnou populaci**
  - V případě, že bychom mohli náhodně oslovit uživatele internetu přímo přes internet (např. Netmonitor?)
  - Jaké potíže?

- https://www.czso.cz/csu/czso/vyuzivani-informacnich-a-komunikacnich-technologii-v-domacnostech-a-mezi-jednotlivci-2020

Graf C4 Používání internetu podle pohlaví a věku



podíl z celkového počtu osob v dané skupině

Zdroj: ČSÚ, Šetření o využívání ICT v domácnostech a mezi jednotlivci

JAK VELKÝ PROBLÉM TO JE?

- **Undercoverage** je velký problém, zvláště pokud nějaké skupiny lidí z populace (významné z hlediska VO) vynechává systematicky
- Bias = systematické zkreslení
- Existují různé mechanismy pro její omezení
  - Např. kombinace různých rámců (RDD + area)
  - Více detailů viz Groves
  - Abychom dokázali generalizovat z dat, potřebujeme co nejvíce vědět o tom, jací respondenti nejsou ve výběrovém rámci: teorie, různé existující statistiky, odborné články..
    - úkol

- **Over coverage** obvykle nebývá takový problém – pokud není příliš velká
  - Můžeme takové respondenty prostě během procesování dat vyřadit
  - Pokud dopředu víme alespoň % odhad kolik takových jednotek v rámci populace je, nadsadíme cílový počet respondentů

JAK VZNIKÁ VÝBĚROVÝ RÁMEC

- Pro specifickou cílovou populaci
  - Záleží na charakteristikách cílové populace a tom, co o ní vlastně víme

- Pokud je populace dostatečně velká a existuje výběrový rámec, který ji kompletně obsahuje (overcoverage)
  - Pomocí screeningových položek se lze dostat k naší populaci
  - Pokud je ale naše cílovka jen zlomek z rámce, pak je to velmi náročné a neefektivní (jak efektivní je dostat se k handicapovaným respondentům skrze seznam adres v ČR?)

- Obtížně získatelné populace: malé nebo obtížně dostupné
  - Můžeme zkusit vytvořit nový rámec (nový seznam prvků, ze kterého pak vybíráme respondenty)
  - Kombinací jiných seznamů (např. klienti různých neziskovek) – často jsou takové seznamy ale důvěrné, nedostaneme se k nim

CO KDYŽ TO NEJDE…

- V praxi velmi často ideální situace, kdy máme perfektní výběrový rámec, nenastává
- Vždy je potřeba hodnotit, jak moc se rámec od cílové populace liší a zda jsou odlišnosti vzhledem k naší VO podstatné více nebo méně

- Někdy prostě ani výběrový rámec není možné sestavit
  - Zkuste vymyslet případy, kdy to patrně nejde

- **Pokud je sampling frame příliš vzdálený od cílové populace:**
  - Úprava VO
  - Smířit se s coverage errors

VZOREK

- Je soubor respondentů vybraný z populace
- Obvykle jen malý zlomek
  - **Sampling error** = část rozdílu mezi statistikou vzorku a parametrem populace, která je způsobená prostě tím, že máme vzorek a ne celou populaci =
- Ne všichni pozvaní budou ale ochotní
  - **Nonresponse error** (část rozdílu způsobená tím, kdo odmítá účast na výzkumu)

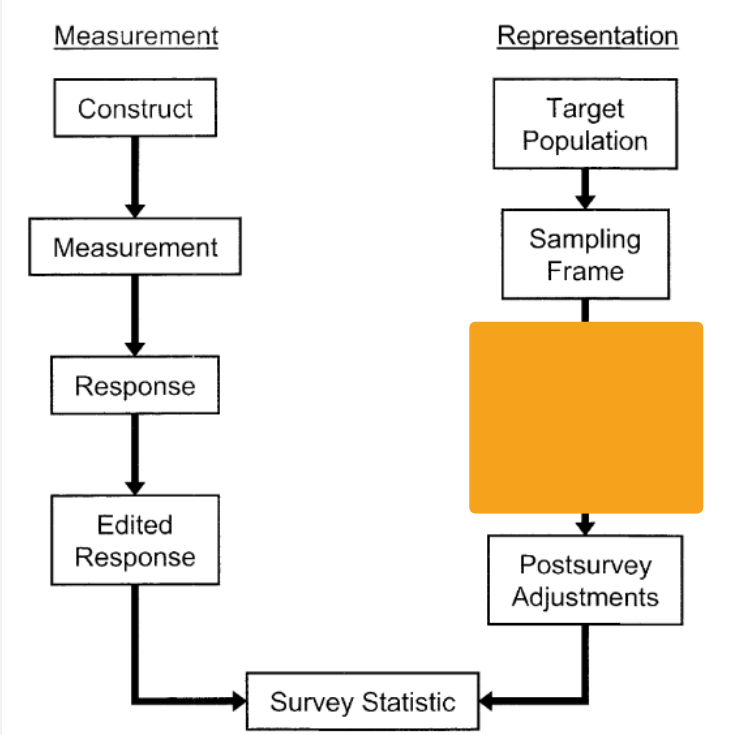


Figure 2.2 Survey lifecycle from a design perspective.

VZOREK

- **Selection error** – způsobená tím, koho a jak oslovujeme
  - Např. pozvánka na vašem FB profilu zasáhne jen vaši sociální bublinu
  - Tazatelé si zvolí oslovovat jen některé typy lidí
  - V pravděpodobnostním výběru by neměla být přítomná, ale je typická v nepravděpodobnostních vzorcích

- **Sample attrition error** – jen u longitudinálních designů
  - Způsobená úbytkem respondentů napříč vlnami sběru dat
  - Někdy se jako „attrition“ označuje i úbytek v rámci jednoho dotazníku (jak respondenti postupně přestávají vyplňovat dotazník)

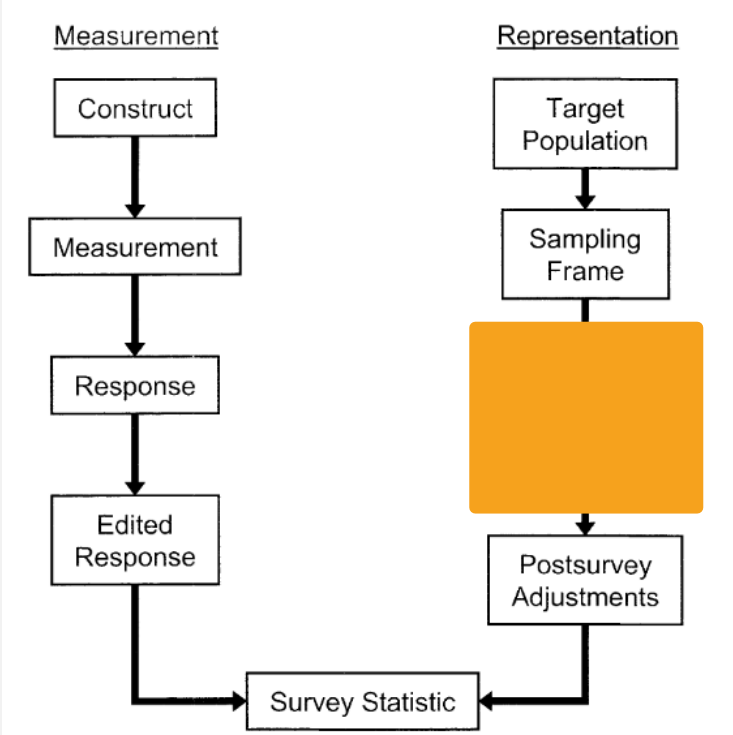


Figure 2.2 Survey lifecycle from a design perspective.

BIAS - ZKRESLENÍ

- Error = rozdíl mezi hodnotou ve vzorku a pravou hodnotu
- Bias = systematické zkreslení
- Tj. bias = pokud to, že někdo chybí (z různých důvodů) ve vzorku anebo v odpovědích na konkrétní položky ve výzkumu (item-response) koreluje s proměnnou, která nás výzkumně zajímá (ať už je závislá nebo nezávislá)
  - Pokud se lidé preferující D. Trumpa častěji neúčastní předvolebních výzkumů, pak jejich preference na prezidenta koreluje s tím, zda odpoví nebo neodpoví na výzkum
- **Stejný vzorek tak může být zkreslený s ohledem na jednu hodnotu/analýzu, ale nezkreslený s ohledem na jinou**
  - Vzorek výš by byl zkreslený s ohledem na odhad preference kandidátů, ale ne, pokud by nás zajímala

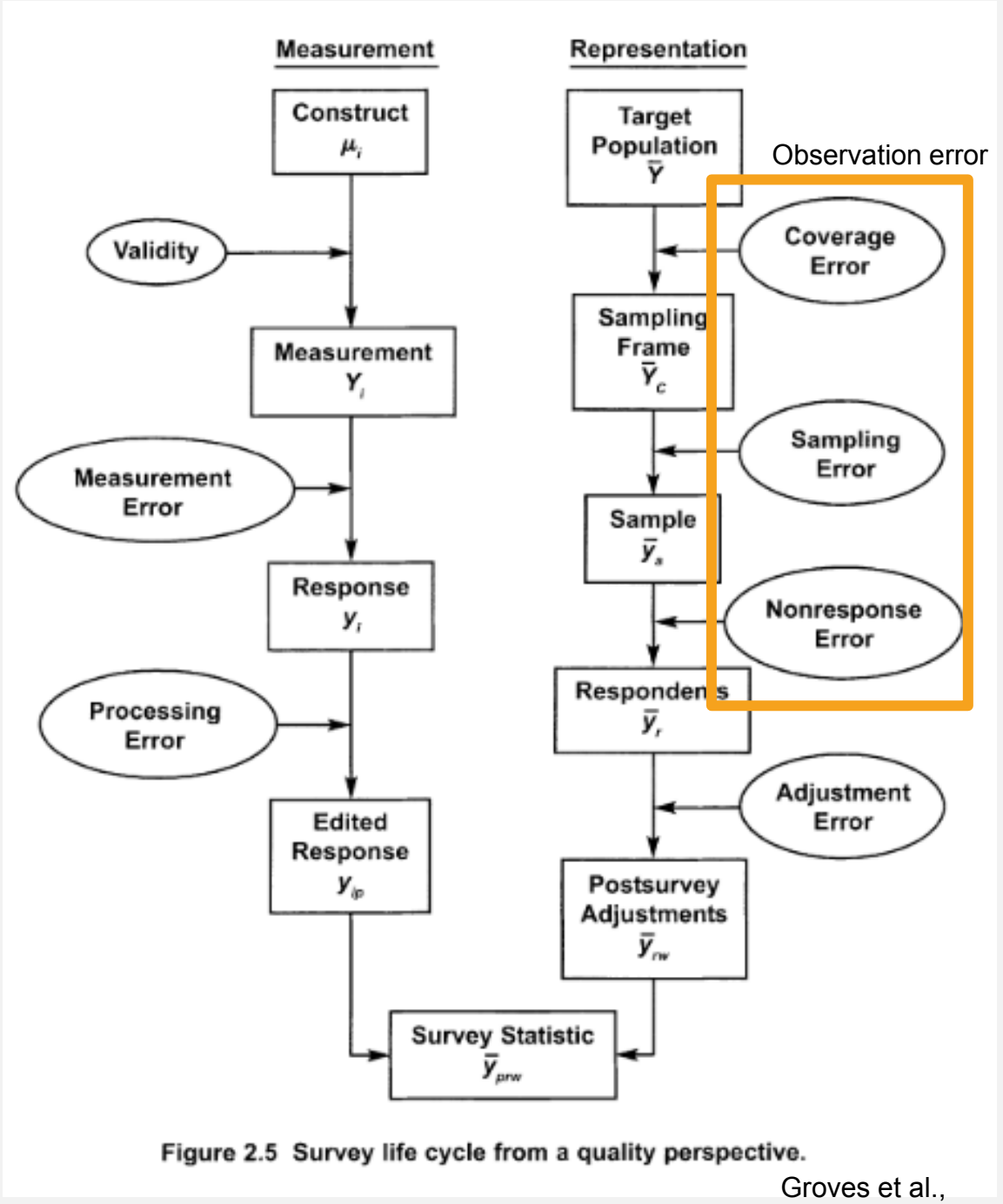


Figure 2.5 Survey life cycle from a quality perspective.

Groves et al.,

VĚTŠÍ CVIČENÍ

- Představte si, že chcete dělat výzkum na obecné populaci a pro vytvoření rámce a oslovení zvolíte adresy bytových prostor
  - Vynechává tento rámec systematicky někoho? Koho?
  - Co když takový sběr budeme chtít provést v červenci?

- Co když chceme stejným způsobem dělat výzkum, jehož cílem je zjistit:
  - Počet domácností s připojením k internetu?
  - Mediální praxe mladých dospělých?

- Jak byste sestavili (se pokusili sestavit) výběrový rámec nebo jaké jiné rámce byste využili, pokud by vaše cílová populace byla:
  - Čtenáři Deníku N?
  - Lidé bez příslušnosti k náboženství?
  - Lidé, kteří finančně podporují neziskovky?

ÚKOL

- Jste Mark Zuckeberg a chcete využít FB jako výběrový rámec pro výzkum, který byste chtěli zobecnit na obecnou populaci.
  - Jako CEO máte tím pádem dokonalý výběrový rámec pro uživatele FB a máte možnost provést pravděpodobnostní výběr.
- Popište, kdo v takovém rámci bude reprezentovaný dobře a kdo naopak špatně.

- (využijte k tomu zdroje – hledejte statistiky, hledejte články; nestačí jen se zamyslet)

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.