MAE4003: Análise de dados amostrais complexos

Tamy H. M. Tsujimoto
tamy.tsujimoto@gmail.com

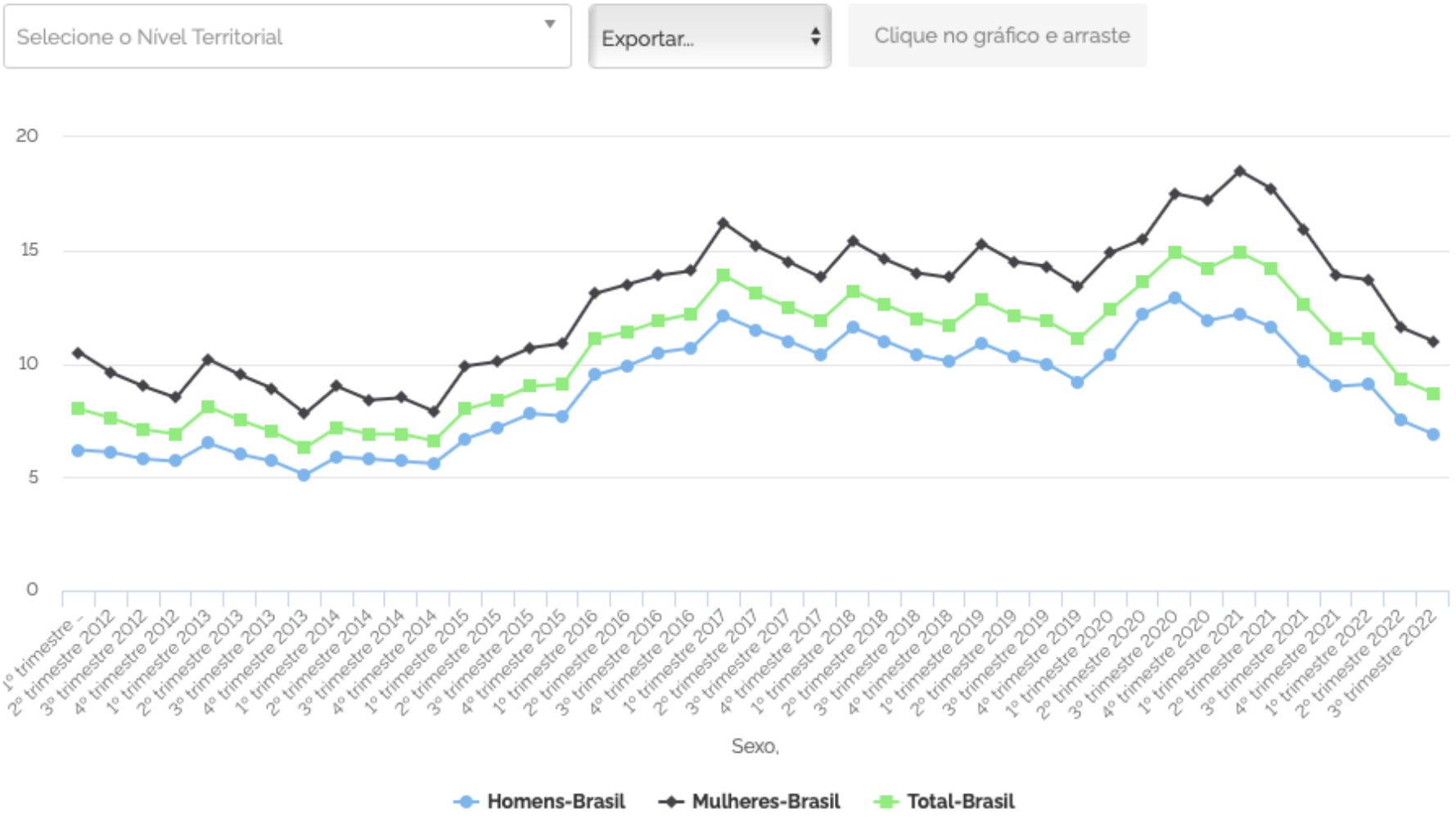
IME-USP
Curso de verão 2023

Aula 1: Introdução e conceitos básicos

Pesquisas amostrais na sociedade

- Papel crítico no fornecimento de estatísticas para diversas áreas
  - Distribuição de faixa etária, renda, gênero, raça
  - Índices de saúde, educação, indústria, desemprego
  - Pesquisas eleitorais, marketing, audiência
- Exemplos no Brasil:
  - Pesquisa Nacional de Saúde (PNS) - IBGE
  - Pesquisa Nacional por Amostra de Domicílio (PNAD) - IBGE
  - Sistema Nacional de Avaliação da Educação Básica (SAEB) - INEP

Taxa de desocupação, por sexo, 1º trimestre 2012 - 3º trimestre 2022



[0]Fonte: www.ibge.gov.br

Por que conduzir uma pesquisa por amostragem?

- Por que não medir todos os elementos de uma população de interesse (censo)?

Por que conduzir uma pesquisa por amostragem?

- Por que não medir todos os elementos de uma população de interesse (censo)?
  - Fornecer informações confiáveis a um custo muito menor

Por que conduzir uma pesquisa por amostragem?

- Por que não medir todos os elementos de uma população de interesse (censo)?
  - Fornecer informações confiáveis a um custo muito menor
  - Dados coletados mais rapidamente $\rightarrow$ estimativas publicadas em tempo hábil/mais frequentemente

Por que conduzir uma pesquisa por amostragem?

- Por que não medir todos os elementos de uma população de interesse (censo)?
  - Fornecer informações confiáveis a um custo muito menor
  - Dados coletados mais rapidamente $\rightarrow$ estimativas publicadas em tempo hábil/mais frequentemente
  - Maior cautela ao coletar os dados $\rightarrow$ estimativas mais precisas

Dados amostrais complexos pra quê?

- Inferência Clássica:

Dados amostrais complexos pra quê?

- Inferência Clássica:

Dados amostrais complexos pra quê?

- Inferência Clássica:
  - População (teoricamente) infinita

Dados amostrais complexos pra quê?

- Inferência Clássica:
  - População (teoricamente) infinita
  - Amostra aleatória simples $\rightarrow$ observações i.i.d

Dados amostrais complexos pra quê?

- Inferência Clássica:
  - População (teoricamente) infinita
  - Amostra aleatória simples $\rightarrow$ observações i.i.d
- Pesquisas amostrais: Dados amostrais complexos

Dados amostrais complexos pra quê?

- Inferência Clássica:
  - População (teoricamente) infinita
  - Amostra aleatória simples $\rightarrow$ observações i.i.d
- Pesquisas amostrais: Dados amostrais complexos
  - População finita

Dados amostrais complexos pra quê?

- Inferência Clássica:
  - População (teoricamente) infinita
  - Amostra aleatória simples $\rightarrow$ observações i.i.d
- Pesquisas amostrais: Dados amostrais complexos
  - População finita
  - Estratificação

Dados amostrais complexos pra quê?

- Inferência Clássica:
  - População (teoricamente) infinita
  - Amostra aleatória simples $\rightarrow$ observações i.i.d
- Pesquisas amostrais: Dados amostrais complexos
  - População finita
  - Estratificação
  - Conglomerados

Dados amostrais complexos pra quê?

- Inferência Clássica:
  - População (teoricamente) infinita
  - Amostra aleatória simples $\rightarrow$ observações i.i.d
- Pesquisas amostrais: Dados amostrais complexos
  - População finita
  - Estratificação
  - Conglomerados
  - Probabilidades de seleção desiguais

Dados amostrais complexos pra quê?

- Inferência Clássica:
  - População (teoricamente) infinita
  - Amostra aleatória simples $\rightarrow$ observações i.i.d
- Pesquisas amostrais: Dados amostrais complexos
  - População finita
  - Estratificação
  - Conglomerados
  - Probabilidades de seleção desiguais
  - Ajustes para não-resposta

Dados amostrais complexos pra quê?

- Inferência Clássica:
  - População (teoricamente) infinita
  - Amostra aleatória simples $\rightarrow$ observações i.i.d
- Pesquisas amostrais: Dados amostrais complexos
  - População finita
  - Estratificação
  - Conglomerados
  - Probabilidades de seleção desiguais
  - Ajustes para não-resposta
- Amostra representativa de maneira viável/eficiente (tempo e recurso)

Definições importantes[1]

[1] Bolfarine & de Oliveira Bussab (2005)

Definições importantes[1]

- **Unidade elementar**: entidade portadora das informações que pretende-se coletar (pessoa, família, domicílio, etc)

[1] Bolfarine & de Oliveira Bussab (2005)

Definições importantes[1]

- **Unidade elementar**: entidade portadora das informações que pretende-se coletar (pessoa, família, domicílio, etc)
- **População alvo**: População que se pretende atingir, usualmente estabelecida nos objetivos da pesquisa.

[1] Bolfarine & de Oliveira Bussab (2005)

Definições importantes[1]

- **Unidade elementar**: entidade portadora das informações que pretende-se coletar (pessoa, família, domicílio, etc)
- **População alvo**: População que se pretende atingir, usualmente estabelecida nos objetivos da pesquisa.
- **Sistema de referência**: Lista (em geral, imperfeita) das unidades da população alvo

[1] Bolfarine & de Oliveira Bussab (2005)

Definições importantes[1]

- **Unidade elementar**: entidade portadora das informações que pretende-se coletar (pessoa, família, domicílio, etc)
- **População alvo**: População que se pretende atingir, usualmente estabelecida nos objetivos da pesquisa.
- **Sistema de referência**: Lista (em geral, imperfeita) das unidades da população alvo
- **População referida**: População previamente disponível e descrita pelo sistema de referência

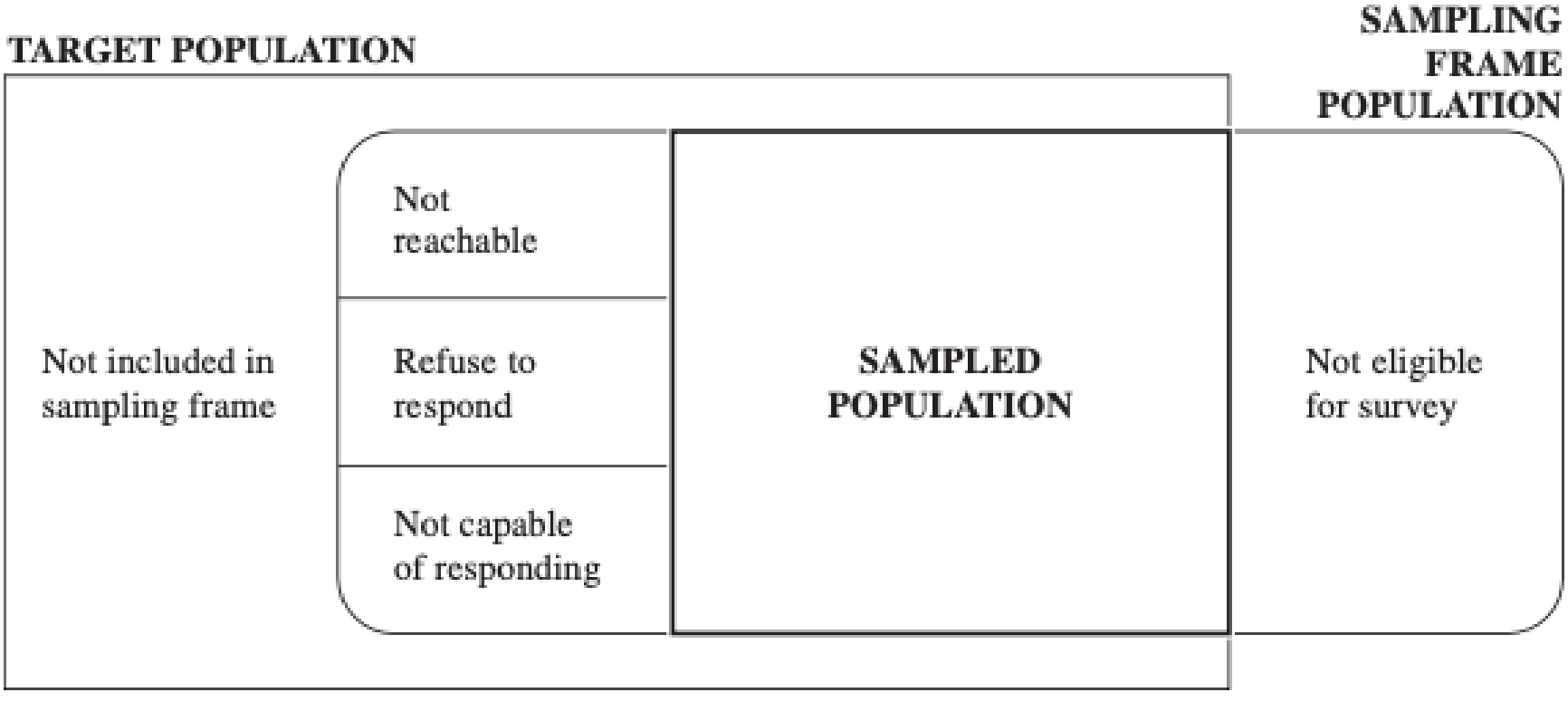
[1] Bolfarine & de Oliveira Bussab (2005)

Definições importantes[1]

- **Unidade elementar**: entidade portadora das informações que pretende-se coletar (pessoa, família, domicílio, etc)
- **População alvo**: População que se pretende atingir, usualmente estabelecida nos objetivos da pesquisa.
- **Sistema de referência**: Lista (em geral, imperfeita) das unidades da população alvo
- **População referida**: População previamente disponível e descrita pelo sistema de referência
- **População amostrada**: População da qual foi retirada a amostra.

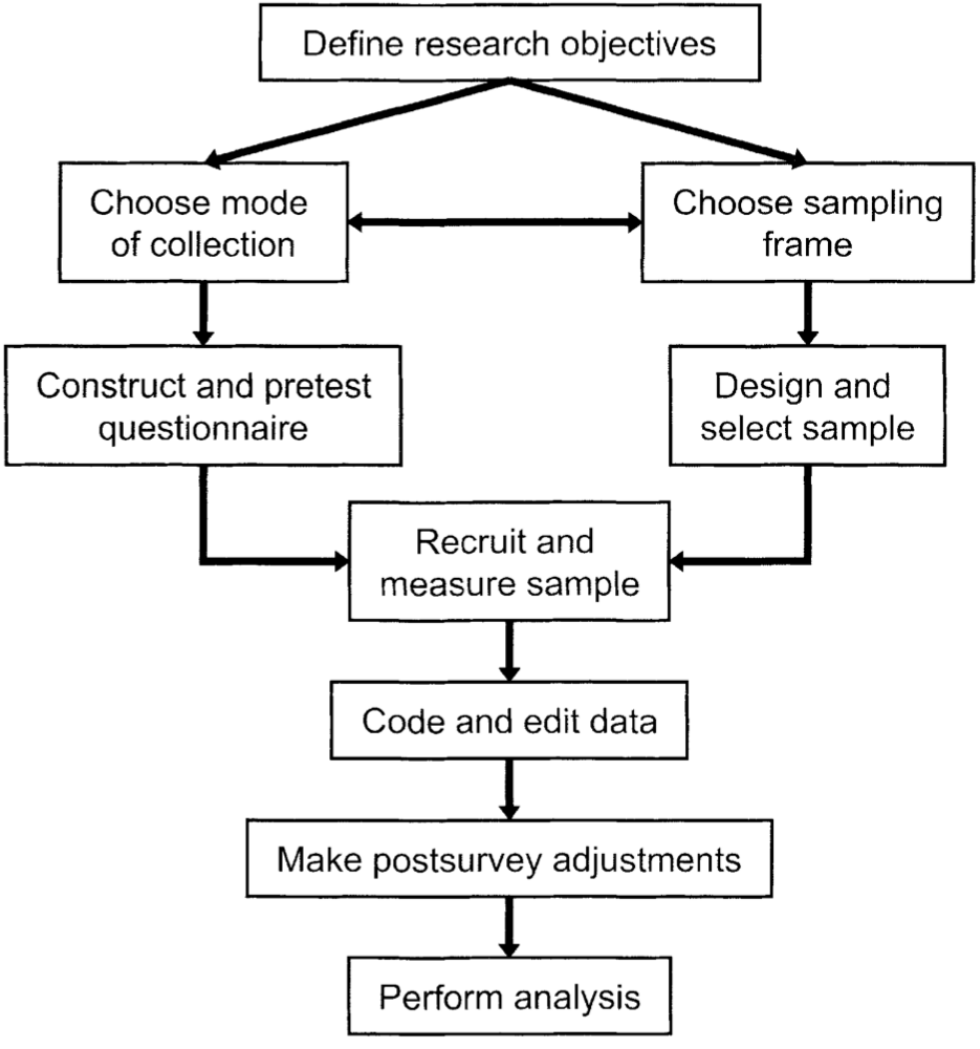
[1] Bolfarine & de Oliveira Bussab (2005)

População alvo x População amostrada



Fonte: Lohr (2021)

Processo de uma pesquisa



Fonte: Groves et al. (2011)

Fontes de erro

- **Aleatório ou amostral**: Erro resultante do processo amostral (medição de uma amostra em vez de toda a população) - diminui conforme aumentamos o tamanho amostral

Fontes de erro

- **Aleatório ou amostral**: Erro resultante do processo amostral (medição de uma amostra em vez de toda a população) - diminui conforme aumentamos o tamanho amostral
- **Não-aleatório ou sistemático**: Todos os erros que não podem ser contabilizados pela variabilidade amostral - **não** diminui conforme aumentamos o tamanho amostral

Fontes de erro

- **Aleatório ou amostral**: Erro resultante do processo amostral (medição de uma amostra em vez de toda a população) - diminui conforme aumentamos o tamanho amostral
- **Não-aleatório ou sistemático**: Todos os erros que não podem ser contabilizados pela variabilidade amostral - **não** diminui conforme aumentamos o tamanho amostral
  - Erro de seleção

Fontes de erro

- **Aleatório ou amostral**: Erro resultante do processo amostral (medição de uma amostra em vez de toda a população) - diminui conforme aumentamos o tamanho amostral
- **Não-aleatório ou sistemático**: Todos os erros que não podem ser contabilizados pela variabilidade amostral - **não** diminui conforme aumentamos o tamanho amostral
  - Erro de seleção
  - Erro de medida

Fontes de erro

Viés de seleção

O **viés de seleção** ocorre quando alguma parte da *população alvo* não está na *população amostrada*

Fontes de erro

Viés de seleção

O **viés de seleção** ocorre quando alguma parte da *população alvo* não está na *população amostrada*

- Procedimento de seleção da amostra é dependente com a variável de interesse

Fontes de erro

Viés de seleção

O **viés de seleção** ocorre quando alguma parte da *população alvo* não está na *população amostrada*

- Procedimento de seleção da amostra é dependente com a variável de interesse
- Falha na definição da população alvo

Fontes de erro

Viés de seleção

O **viés de seleção** ocorre quando alguma parte da *população alvo* não está na *população amostrada*

- Procedimento de seleção da amostra é dependente com a variável de interesse
- Falha na definição da população alvo
- Erro de omissão - parte da população alvo não está contida no sistema de referência

Fontes de erro

Viés de seleção

O **viés de seleção** ocorre quando alguma parte da *população alvo* não está na *população amostrada*

- Procedimento de seleção da amostra é dependente com a variável de interesse
- Falha na definição da população alvo
- Erro de omissão - parte da população alvo não está contida no sistema de referência
- Erro de comissão - inclusão de elementos fora da população alvo no sistema de referência

Fontes de erro

Viés de seleção

O **viés de seleção** ocorre quando alguma parte da *população alvo* não está na *população amostrada*

- Procedimento de seleção da amostra é dependente com a variável de interesse
- Falha na definição da população alvo
- Erro de omissão - parte da população alvo não está contida no sistema de referência
- Erro de comissão - inclusão de elementos fora da população alvo no sistema de referência
- Não-resposta

Fontes de erro

Viés de medida

O **viés de medição** ocorre quando a resposta tem uma tendência a diferir do valor verdadeiro em uma direção.

Fontes de erro

Viés de medida

O **viés de medição** ocorre quando a resposta tem uma tendência a diferir do valor verdadeiro em uma direção.

- Participantes não dizendo a verdade ou não ententendo perguntas

Fontes de erro

Viés de medida

O **viés de medição** ocorre quando a resposta tem uma tendência a diferir do valor verdadeiro em uma direção.

- Participantes não dizendo a verdade ou não ententendo perguntas
- Esquecimento dos fatos

Fontes de erro

Viés de medida

O **viés de medição** ocorre quando a resposta tem uma tendência a diferir do valor verdadeiro em uma direção.

- Participantes não dizendo a verdade ou não ententendo perguntas
- Esquecimento dos fatos
- Efeito do entrevistador

Fontes de erro

Viés de medida

O **viés de medição** ocorre quando a resposta tem uma tendência a diferir do valor verdadeiro em uma direção.

- Participantes não dizendo a verdade ou não ententendo perguntas
- Esquecimento dos fatos
- Efeito do entrevistador
- Efeito de desejabilidade social

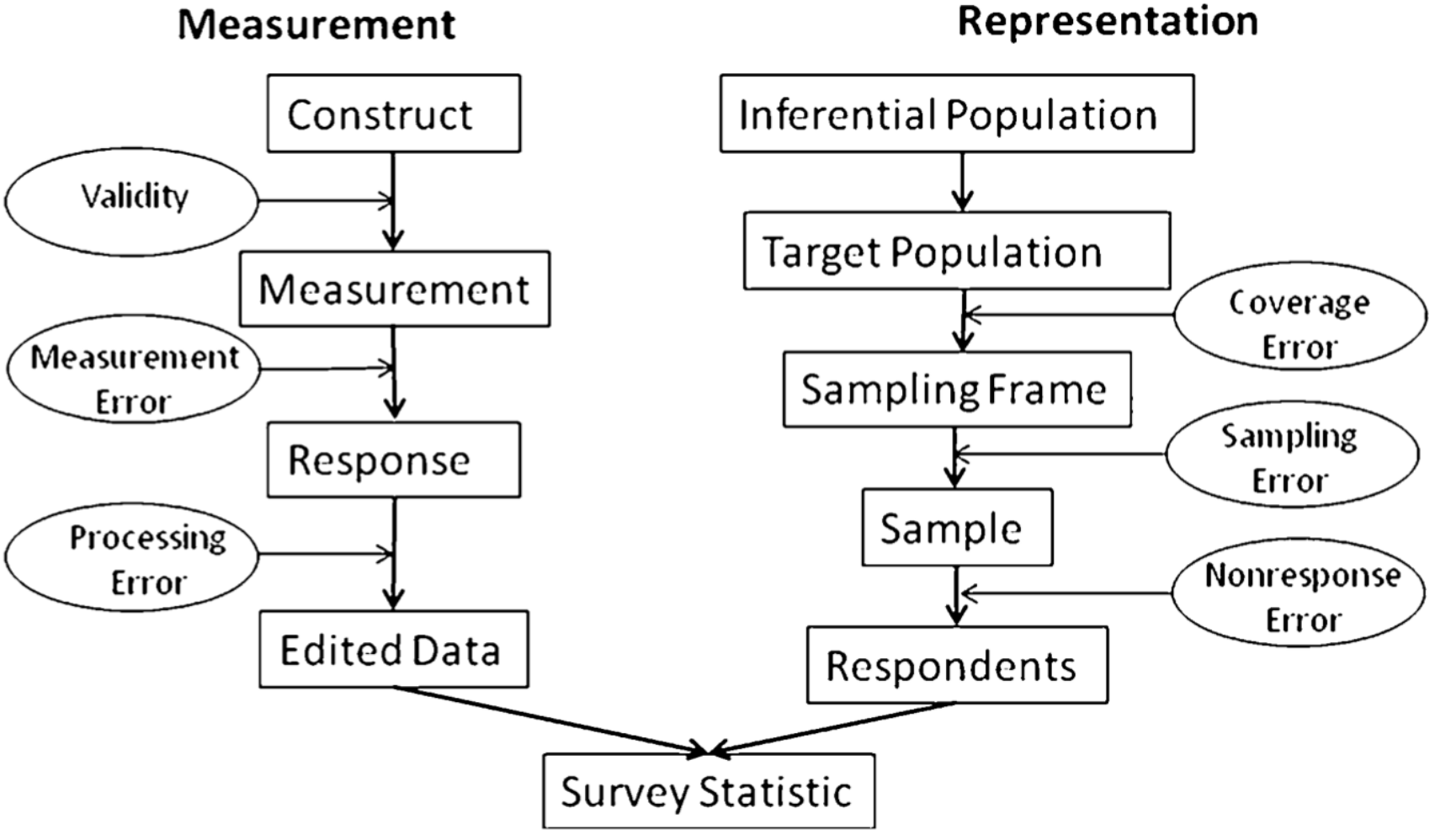
Fontes de erro

Viés de medida

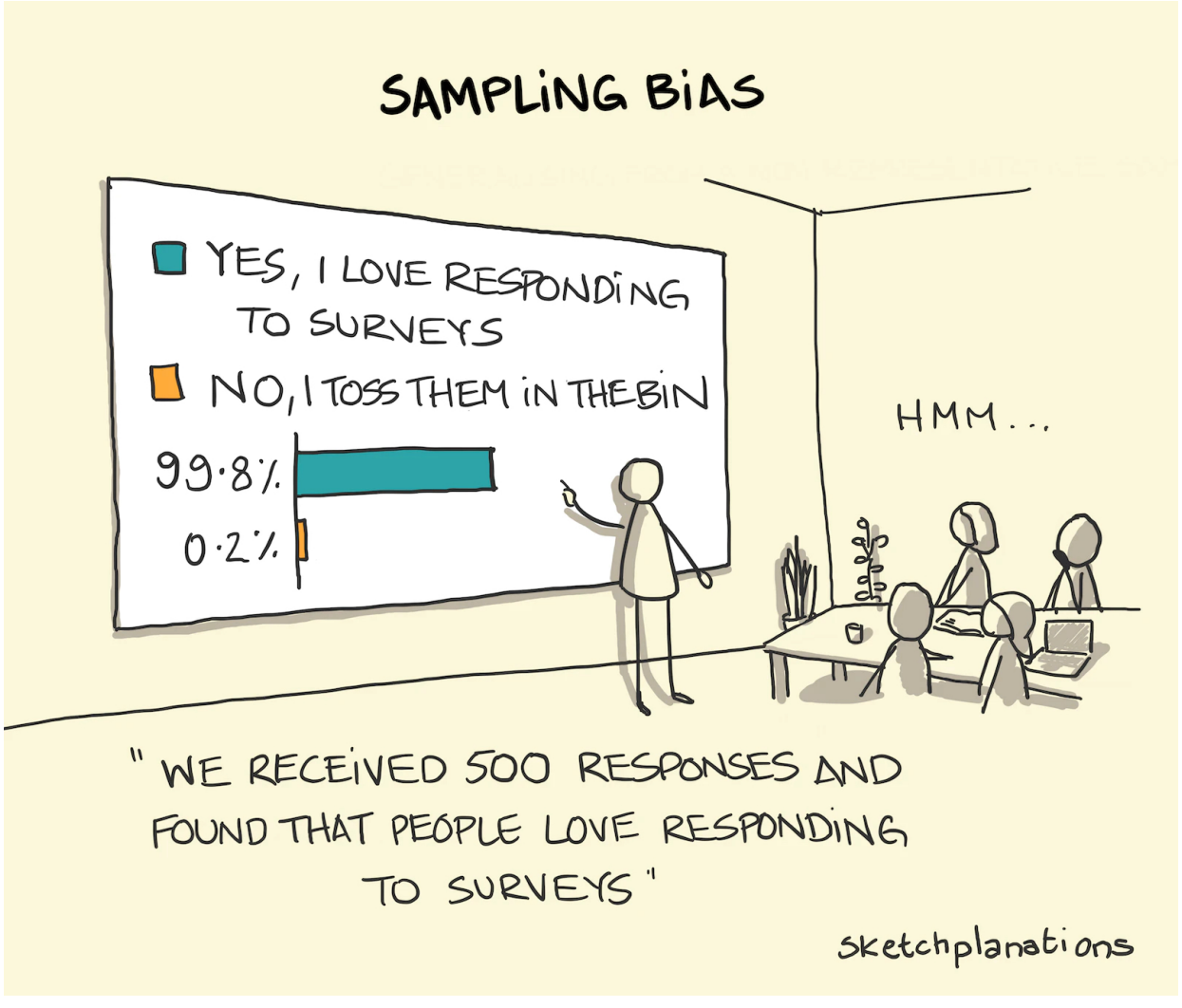
O **viés de medição** ocorre quando a resposta tem uma tendência a diferir do valor verdadeiro em uma direção.

- Participantes não dizendo a verdade ou não ententendo perguntas
- Esquecimento dos fatos
- Efeito do entrevistador
- Efeito de desejabilidade social
- Erro de processamento: erros decorrentes de codificação, transcrição, imputação, edição, tratamento de outliers e outros tipos de manipulação de dados antes da etapa de análise.

Fontes de erro



Fonte: Groves & Lyberg (2010)


SAMPLING BIAS
YES, I LOVE RESPONDING TO SURVEYS
NO, I TOSS THEM IN THE BIN
99·8%
0·2%
HMM...
" WE RECEIVED 500 RESPONSES AND FOUND THAT PEOPLE LOVE RESPONDING TO SURVEYS "
sketchplanations

Dúvidas?

Amostragem probabilística

Amostragem probabilística

- Amostra representativa

Amostragem probabilística

- Amostra representativa
- Cada unidade na população tem uma probabilidade conhecida e não nula de seleção

Amostragem probabilística

- Amostra representativa
- Cada unidade na população tem uma probabilidade conhecida e não nula de seleção
- Mecanismo de seleção é aleatório

Amostragem probabilística

- Amostra representativa
- Cada unidade na população tem uma probabilidade conhecida e não nula de seleção
- Mecanismo de seleção é aleatório
- Tipos de amostragem probabilística:

Amostragem probabilística

- Amostra representativa
- Cada unidade na população tem uma probabilidade conhecida e não nula de seleção
- Mecanismo de seleção é aleatório
- Tipos de amostragem probabilística:
  - Amostra Aleatória Simples

Amostragem probabilística

- Amostra representativa
- Cada unidade na população tem uma probabilidade conhecida e não nula de seleção
- Mecanismo de seleção é aleatório
- Tipos de amostragem probabilística:
  - Amostra Aleatória Simples
  - Amostragem Estratificada

Amostragem probabilística

- Amostra representativa
- Cada unidade na população tem uma probabilidade conhecida e não nula de seleção
- Mecanismo de seleção é aleatório
- Tipos de amostragem probabilística:
  - Amostra Aleatória Simples
  - Amostragem Estratificada
  - Amostragem por Conglomerado

Amostragem probabilística

- Amostra representativa
- Cada unidade na população tem uma probabilidade conhecida e não nula de seleção
- Mecanismo de seleção é aleatório
- Tipos de amostragem probabilística:
  - Amostra Aleatória Simples
  - Amostragem Estratificada
  - Amostragem por Conglomerado
  - Amostragem Sistemática

Tipos de amostragem probabilística

- **Amostra Aleatória Simples (AAS)**: Todo subconjunto de n unidades da população possui a mesma chance de ser amostrado

Tipos de amostragem probabilística

- **Amostra Aleatória Simples (AAS)**: Todo subconjunto de n unidades da população possui a mesma chance de ser amostrado
- **Amostragem Estratificada**: População dividida em subgrupos (estratos) e uma AAS é selecionada de cada estrato de maneira independente

Tipos de amostragem probabilística

- **Amostra Aleatória Simples (AAS)**: Todo subconjunto de n unidades da população possui a mesma chance de ser amostrado
- **Amostragem Estratificada**: População dividida em subgrupos (estratos) e uma AAS é selecionada de cada estrato de maneira independente
- **Amostragem por Conglomerado**: Unidades da população agregadas em unidades amostrais maiores (clusters). As unidades de observação são selecionados a partir de clusters amostrados

Tipos de amostragem probabilística

- **Amostra Aleatória Simples (AAS)**: Todo subconjunto de n unidades da população possui a mesma chance de ser amostrado
- **Amostragem Estratificada**: População dividida em subgrupos (estratos) e uma AAS é selecionada de cada estrato de maneira independente
- **Amostragem por Conglomerado**: Unidades da população agregadas em unidades amostrais maiores (clusters). As unidades de observação são selecionados a partir de clusters amostrados
- **Amostragem Sistemática**: Ponto de partida escolhido aleatoriamente entre as k primeiras observações na população. Esta unidade e cada k-ésima unidade seguinte são selecionadas para a amostra

Analisando dados de amostragem complexa

Tipos de estudos

- Descritivo: Quantidades populacionais (ex: total, média, proporção)
- Analítico: Modelo estatístico gerador da população finita

Modos de Inferência

- Probabilística (Design-based): abordagem não-paramétrica
- Modelagem (Model-based): abordagem paramétrica
- Conjunta (Joint Inference): abordagem mista

População e amostra

- U: população finita de tamanho N com elementos indexados por $i \in \{1, \cdots, N\}$
- Cada índice i é associado ao vetor $(y_i, x_i, z_i) \in \mathbb{R}^p \times \mathbb{R}^k \times \mathbb{R}_+^q$
  - y_i: vetor $p \times 1$ de variáveis de interesse
  - x_i: vetor $k \times 1$ de variáveis auxiliares
  - z_i: vetor $q \times 1$ de variáveis da amostragem complexa (disponível no momento do planejamento da pesquisa para todas as unidades da população)
- s: amostra (subconjunto) de tamanho n coletada segundo um **plano amostral** da população finita U

Plano Amostral

- $\mathcal{S} = \{s : s \subset U\}$: coleção de subconjuntos da população finita U de acordo com um dado desenho amostral
- $\sigma(\mathcal{S})$: sigma-álgebra gerada por $\mathcal{S}$
- S: variável aleatória que assume valores em $\sigma(\mathcal{S})$

Plano Amostral

O mecanismo usado para selecionar a amostra s da população finita U é chamado **plano amostral**:

$$p(s) = P(S = s) = P(\text{amostra s ser selecionada}), \quad s \in \mathcal{S}$$

- $p(s) \geq 0$ para todo $s \in \mathcal{S}$
- $\sum_{s \in \mathcal{S}} p(s) = 1$

Probabilidade de Inclusão

Seja s amostra selecionada de acordo com um plano amostral $p(s)$. Definimos o $\xi_i(S) = I(S \ni i)$, $i = 1, \cdots, N$.

Probabilidade de Inclusão

Denotamos como π_i a probabilidade de inclusão da unidade i na amostra s:

$$\pi_i = \mathbb{P}(S \ni i) = \mathbb{P}(\xi_i = 1) = \sum_{s: i \in s} p(s)$$

e denotamos como π_{ij} a probabilidade conjunta de inclusão das unidades i e j na amostra s:

$$\pi_{ij} = \mathbb{P}(S \ni i, j) = \mathbb{P}(\xi_i = 1, \xi_j = 1) = \sum_{s: i,j \in s} p(s)$$

$p(s)$ é um plano amostral **probabilístico** se $\pi_i > 0 \ \forall i$

Probabilidade de Inclusão

Definimos o $\xi_i(S) = I(S \ni i)$, $i = 1, \cdots, N$.

Resultados

- $\mathrm{E}(\xi_i) = \pi_i$
- $\mathrm{Var}(\xi_i) = \pi_i(1 - \pi_i)$
- $\mathrm{Cov}(\xi_i, \xi_j) = \pi_{ij} - \pi_i\pi_j$

Peso amostral

Peso amostral

Dado um plano amostral $p(.)$, definimos o **peso amostral** da i-ésima observação ($i = 1, \cdots, N$) como o inverso da sua probabilidade de inclusão:

$$w_i = 1/\pi_i$$

Interpretação: w_i é o número de unidades populacionais representadas pela observação i.

Exemplo

Suponha a seguinte população $U = \{1, 2, 3, 4\}$. As possíveis amostras de tamanho $n = 2$ são:

$$s_1 = \{1, 2\} \quad s_2 = \{1, 3\} \quad s_3 = \{1, 4\}$$
$$s_4 = \{2, 3\} \quad s_5 = \{2, 4\} \quad s_6 = \{3, 4\}$$

Suponha o seguinte plano amostral:

$$p(s_1) = 1/3 \quad p(s_2) = 1/6 \quad p(s_3) = p(s_4) = p(s_5) = 0 \quad p(s_6) = 1/2$$

Neste caso:

$$\pi_1 = p(s_1) + p(s_2) + p(s_3) = 1/2$$
$$\pi_2 = p(s_1) + p(s_4) + p(s_5) = 1/3$$
$$\pi_3 = p(s_2 + p(s_4) + p(s_6) = 2/3$$
$$\pi_4 = p(s_3) + p(s_5) + p(s_6) = 1/2$$

Estimadores

- $\theta_N = \theta(y_1, \cdots, y_N)$: Parâmetro populacional de interesse
- $\hat{\theta} = \hat{\theta}(S)$: estimador do parâmetro θ
- Esperança com relação à distribuição de aleatorização:

$$\mathrm{E}(\hat{\theta}) = \sum_{s \in \mathcal{S}} p(s)\hat{\theta}(s)$$

- Variância com relação à distribuição de aleatorização:

$$\mathrm{Var}(\hat{\theta}) = \sum_{s \in \mathcal{S}} p(s)\{\hat{\theta}(s) - \mathrm{E}(\hat{\theta})\}^2$$

- $\hat{\theta}$ é não enviesado com relação à distribuição de aleatorização se $\mathrm{E}(\hat{\theta}) = \theta_N$

Parâmetros Populacionais

- Total populacional:

$$t_N = \sum_{i=1}^{N} y_i$$

- Média populacional:

$$\bar{y}_N = N^{-1} \sum_{i=1}^{N} y_i$$

- Variância populacional:

$$S_N^2 = (N-1)^{-1} \sum_{i=1}^{N} (y_i - \bar{y}_N)^2$$

Estimador π-ponderado (Horvitz-Thompson)

Seja $t_N = \sum_{i=1}^{N} y_i$ o total populacional

Estimador Horvitz-Thompson

$$\hat{t}_\pi = \sum_{i \in S} \frac{y_i}{\pi_i} = \sum_{i=1}^{N} \frac{\xi_i}{\pi_i} y_i$$

- Base para outros estimadores
- $\mathrm{E}(\hat{t}_\pi) = \sum_{i=1}^{N} y_i$
- $\mathrm{Var}(\hat{t}_\pi) = \sum_{i=1}^{N} \sum_{j=1}^{N} (\pi_{ij} - \pi_i \pi_j) \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}$
- $\widehat{\mathrm{Var}}(\hat{t}_\pi) = \sum_{i=1}^{N} \sum_{j=1}^{N} \xi_i \xi_j \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \right) \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}$

Laboratório 1: Introdução

PNAD Contínua

- Pesquisa Nacional por Amostra de Domicílios Contínua - IBGE
- Força de trabalho, e outras informações necessárias para o estudo do desenvolvimento socioeconômico do País
- Plano amostral Complexo
  - Estratificação
  - Conglomeração em múltiplos estágios
  - Probabilidades desiguais de seleção
  - Calibração dos pesos amostrais
  - Descrição detalhada em Silva et al. (2002)
- Divulgação
  - Trimestral: Informações da pesquisa básica
  - Anual: Informações das pesquisas suplementares acumulados para compor a amostra do ano

PNAD Contínua: Variáveis

1. Identificação e Controle
2. Características Gerais dos Moradores
3. Características de educação para os moradores de 5 anos ou mais de idade
4. Características de trabalho das pessoas de 14 anos ou mais de idade
5. Variáveis Derivadas
6. Pesos Replicados

Software

- R: `survey`, `gtsummary`, `tidyverse`
  - Lumley, T. (2011). Complex surveys: a guide to analysis using R. John Wiley & Sons.
  - http://r-survey.r-forge.r-project.org/survey/index.html
- SAS: PROC SURVEY
  - **Referência**: Lewis, T. H. (2016). Complex survey data analysis with SAS. Chapman and Hall/CRC.
- STATA, SPSS, SUDAAN

PNAD Contínua: Obtendo os dados

- Site do microdados do PNADc no IBGE
- Utilizando a função `get_pnadc` do pacote `PNADcIBGE`
  1. Instalar o pacote: `install.packages("PNADcIBGE")`
  2. Carregar o pacote: `library(PNADcIBGE)`
  3. Baixar os dados brutos (4o trimestre de 2021):

```
pnad_data_20214 <- get_pnadc(year=2021,
                             quarter = 4,
                             design=FALSE)
```

- **Mais detalhes**: https://rpubs.com/gabriel-assuncao-ibge/pnadc

[plain, noframenumbering] # Referências

Bolfarine, H. & de Oliveira Bussab, W. (2005). *Elementos de amostragem*. Editora Blucher.

Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E. & Tourangeau, R. (2011). *Survey methodology*. John Wiley & Sons.

Groves, R. M. & Lyberg, L. (2010). Total survey error: Past, present, and future. *Public opinion quarterly* **74**, 849–879.

Lohr, S. L. (2021). *Sampling: design and analysis*. Chapman and Hall/CRC.

Silva, P. L. d. N., Pessoa, D. G. C. & Lila, M. F. (2002). Análise estatística de dados da pnad: incorporando a estrutura do plano amostral. *Ciência & Saúde Coletiva* **7**, 659–670.