Does Sampling Matter in Psychological Science?

PsicoStat - 21/03/2025

Alberto Arletti
University of Padova
March 21, 2025

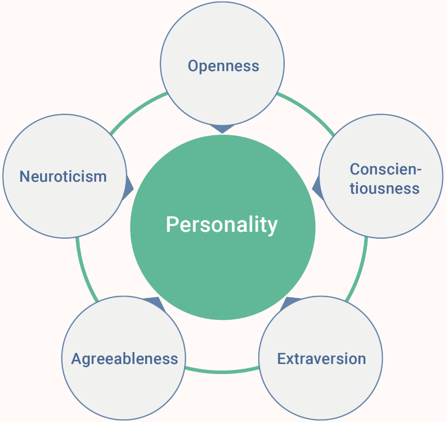
Outline

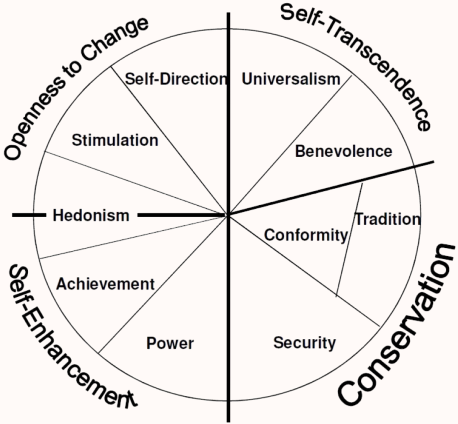
Introduction

Are we all the same?

Psychological phenomena are sometimes considered universal among **all human beings**.

Examples:





Example Study

I am measuring the relationship between choice in university course and depression diagnosis in university students. I send an e-mail on the University mailing list asking for volunteers.

Example Study

I am measuring the relationship between choice in university course and depression diagnosis in university students. I send an e-mail on the University mailing list asking for volunteers.

X are the predictors (age, gender, course etc..)
Y the target variable (depression diagnosis)
S is the selection mechanism

Example Study

I am measuring the relationship between choice in university course and depression diagnosis in university students. I send an e-mail on the University mailing list asking for volunteers.

X are the predictors (age, gender, course etc..)
Y the target variable (depression diagnosis)
S is the selection mechanism

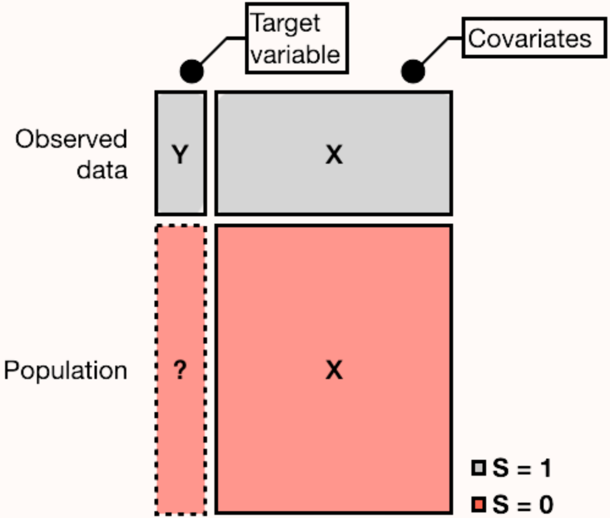
Selection Mechanism

A variable that indicates who is included in the study. If $S_i = 1$ the i-th individual in the population is included in the study (answer my email).

Basic Setting

X predictors
Y target variable
$S \in \{0, 1\}$ selection, if $S_i = 1$ then observation i is the dataset

Usually $S = 1$ is a very small part of the population.

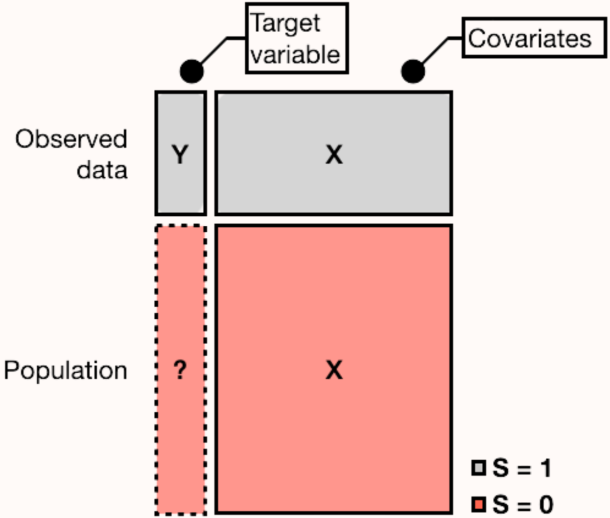


Basic Setting

X predictors
Y target variable
$S \in \{0, 1\}$ selection, if $S_i = 1$ then observation i is the dataset

Usually $S = 1$ is a very small part of the population.

What happens to all the i where $S = 0$?



Can we do this?

Just run:

```
glm(Y ~ X, family =
binomial(link = 'logit'))
```

Question

Can we ignore the selection mechanism S?

Can we do this?

Why Selection but not Missigness?

Would we ignore missing data?
If M is the missingness mechanism
$S_i = 1 \rightarrow M_i = 0$ and
$S_i = 0 \rightarrow M_i = 1$

Selection (Missingness) Mechanisms

Define: X predictors, Y target variable, $S \in \{0, 1\}$ selection, if $S_i = 1$ then observation i is the dataset, $f(\cdot)$ is a function.

MAR: Missing (Selection) at Random

The good kind

$$P(S = 1) \sim f(X), \quad \frac{P(S|X, Y)}{P(S|X)} = 1$$

Also called **Ignorable Selection**:We can focus on n.

Selection (Missingness) Mechanisms

Define: X predictors, Y target variable, $S \in \{0, 1\}$ selection, if $S_i = 1$ then observation i is the dataset, $f(\cdot)$ is a function.

MNAR: Missing **Not** at Random

The not so good kind

$$P(S = 1) \sim f(X, Y), \quad \frac{P(S|X, Y)}{P(S|X)} \neq 1$$

Also called **Non-Ignorable Selection**.

Example Study

X predictors such as: which course the student is in, age, gender etc..
Y target variable: diagnosis of depression

Examples of **Ignorable** selection

- Students of social sciences (X) might be more interested in the research topic.
- Female (X) students might be more agreeable and therefore willing to dedicate some time.

Examples of **Non-Ignorable** selection

- Depressed students (Y) might not reply to emails.

I'll just use the individuals who answered, as long as there are some with a diagnosis of depression!

MNAR in Population Totals Estimation

Population total estimation task

Estimate the total number of students with a diagnosis of depression in the University of Padua: $\bar{Y}_N = \sum_i^N Y_i$.

Population total estimation task

Estimate the total number of students with a diagnosis of depression in the University of Padua: $\bar{Y}_N = \sum_i^N Y_i$.

Question

What would be the bias of this estimate?

bias $= E(\overline{Y}_N - \overline{Y}_n)$

Population total estimation task

Estimate the total number of students with a diagnosis of depression in the University of Padua: $\bar{Y}_N = \sum_i^N Y_i$.

Question

What would be the bias of this estimate?

bias $= E(\overline{Y}_N - \overline{Y}_n)$

$$\text{bias} = \overline{Y}_N - \overline{Y}_n = \frac{E(YS)}{E(S)} - E(Y) = \frac{\text{Cov}(S,Y)}{E(S)} \qquad \text{(DDI)}$$

$$= \underbrace{\text{Corr}(S,Y)}_{\text{Data Quality}} \times \underbrace{\sqrt{\frac{1 - \sum_i^N S_i/N}{\sum_i^N S_i/N}}}_{\text{Data Quantity}} \times \underbrace{\sigma_Y}_{\text{Problem Difficulty}}$$

Meng (2018), pag. 680

Consequences of the lurking Monster N

Design Effect

How much more bias are we having in our study due to S not being optimal (Simple Random Sampling).

Consequences of the lurking Monster N

Design Effect

How much more bias are we having in our study due to S not being optimal (Simple Random Sampling).

Design Effect

$$\text{Deff} = \frac{E(\bar{Y}_N - \bar{Y}_n)^2}{V_{\text{SRS}}(\bar{Y}_n)} = (N-1)E(\text{Corr}^2(S,Y))$$

Meng (2018), pag. 696

$$\text{Deff} = \frac{E(\bar{Y}_N - \bar{Y}_n)^2}{V_{\text{SRS}}(\bar{Y}_n)} = (\boxed{N} - 1)E(\text{Corr}^2(S, Y))$$

$$N?$$

How Statisticians Slew the Monster of Population Size: Random Sampling

Random Sampling: No matter the size of the plate (population, N) we can take a small bite (sample, n) and judge the whole meal.

Population total estimation task

Estimate the total number of students with a diagnosis of depression in the University of Padua: $\bar{Y}_N = \sum_i^N Y_i$.

Question

What would be the bias of this estimate?
bias $= \overline{Y}_N - \overline{Y}_n$

Without Random Sampling:

$$(\text{Expected bias})^2 \propto (\boxed{N} - 1)E(\mathrm{Corr}^2(S, Y))$$

Psychologists

"But I don't need to estimate population total, I am just interested in the coefficients!"

MNAR in Coefficients Estimation

Consequences of the lurking Monster N, pt. 2

Coefficient estimation task

Estimate the coefficient of regression β indicating the relationship between university course and depression diagnosis.

Consequences of the lurking Monster N, pt. 2

Coefficient estimation task

Estimate the coefficient of regression β indicating the relationship between university course and depression diagnosis.

Red Flag

Regression assumes i.i.d. (independence between observations).

Consequences of the lurking Monster N, pt. 2

Coefficient estimation task

Estimate the coefficient of regression β indicating the relationship between university course and depression diagnosis.

Red Flag

Regression assumes i.i.d. (independence between observations).

If observations are i.i.d.:

$$n = n_{\text{eff}}$$

Consequences of the lurking Monster N

Effective sample size

$$n_{\text{eff}} = \frac{n}{\text{Deff}}, \quad \text{Deff} = (N-1)E(\text{Corr}^2(S,Y))$$

Effective sample size

$$n_{\text{eff}} = \frac{n}{\text{Deff}}, \quad \text{Deff} = (N-1)E(\text{Corr}^2(S,Y))$$

We compare the MSE of $\bar{Y}_n$ with the MSE of Simple Random Sampling with sample size n. We set n^*_{eff} as the effective sample size for our sample:

$$n_{\text{eff}} \leq n^*_{\text{eff}} = \frac{n}{1-f}\frac{1}{\boxed{N}E(\text{Corr}^2(S,Y))}$$

Meng (2018), pag. 698

Consequences of the lurking Monster N

$$n_{\text{eff}} \leq n^*_{\text{eff}} = \frac{n}{1-f} \frac{1}{NE(\text{Corr}^2(S,Y))}$$

$n = 250$ sample size
$N = 65.000$ population size of UniPd students
$f = \frac{n}{N}$ sampling rate
$E(\text{Corr}^2(S,Y)) =$ expected correlation between outcome and selection mechanism.

Consequences of the lurking Monster N

$$n_{\text{eff}} \leq n^*_{\text{eff}} = \frac{n}{1-f} \frac{1}{NE(\text{Corr}^2(S,Y))}$$

$n = 250$ sample size
$N = 65.000$ population size of UniPd students
$f = \frac{n}{N}$ sampling rate
$E(\text{Corr}^2(S,Y))$ = expected correlation between outcome and selection mechanism.
We imagine a correlation similar to the one of the US election (0.00021).

Consequences of the lurking Monster N

$$n_{\text{eff}} \leq n^*_{\text{eff}} = \frac{n}{1-f} \frac{1}{NE(\text{Corr}^2(S,Y))}$$

$n = 250$ sample size
$N = 65.000$ population size of UniPd students
$f = \frac{n}{N}$ sampling rate
$E(\text{Corr}^2(S,Y))$ = expected correlation between outcome and selection mechanism.
We imagine a correlation similar to the one of the US election (0.00021).

$$n_{\text{eff}} \leq 18$$

Consequences of the lurking Monster N, pt. 2

Coefficient estimation task

Estimate the coefficient of regression β indicating the relationship between university course and depression diagnosis.

Red Flag

Regression assumes i.i.d. (independence between observations).

Observations are not i.i.d.!

$$n \neq n_{\text{eff}}$$

Consequences of the lurking Monster N, pt. 2

Coefficient estimation task

Estimate the coefficient of regression β indicating the relationship between university course and depression diagnosis.

Coefficients must be wrong. In **MNAR**:

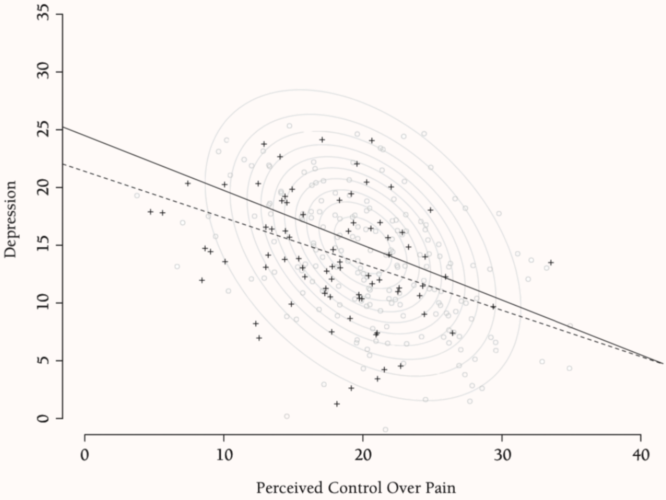
$$\text{if } \frac{P(S|X,Y)}{P(S|X)} \neq 1 \text{ then } \frac{P(Y|X,S=1)}{P(Y|X,S)} \neq 1$$

or, in other words: $\boxed{P(Y|X,S=1) \neq P(Y|X,S)}$
which means:

$$\hat{\beta}_{S=1} \neq \beta_S$$

Proof in Sahoo et al. (2022), Theorem 2.

Example

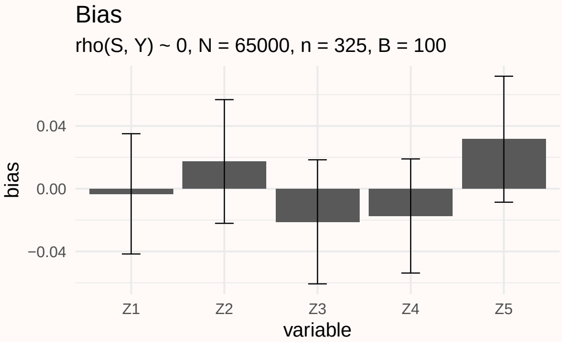


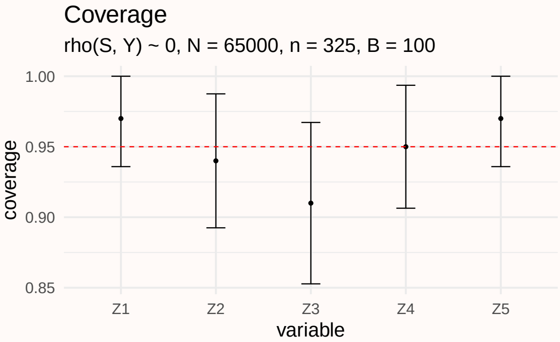
Enders (2022), Ch. 9

The Big Data Paradox

A Simple Simulation - MAR

$N = 65000, \quad f = 0.005, \quad p = 5$, MAR Sample

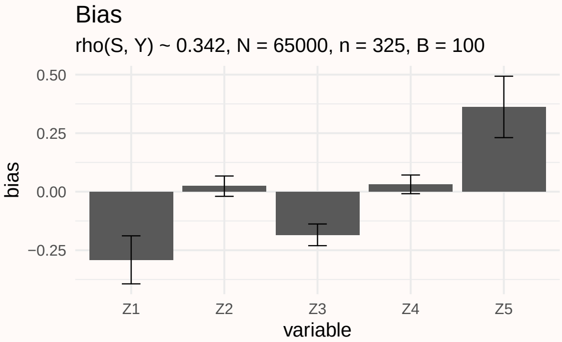


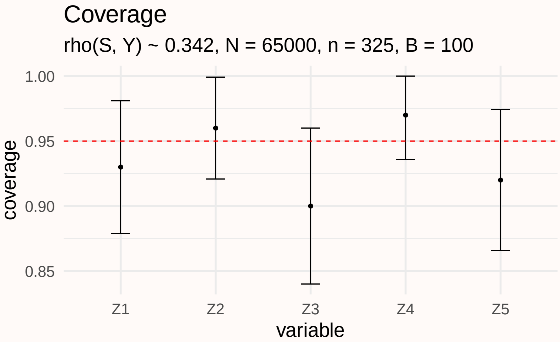


MCSE errorbars.

A Simple Simulation - MNAR

$N = 65000, \quad f = 0.005, \quad p = 5,$ MNAR Sample

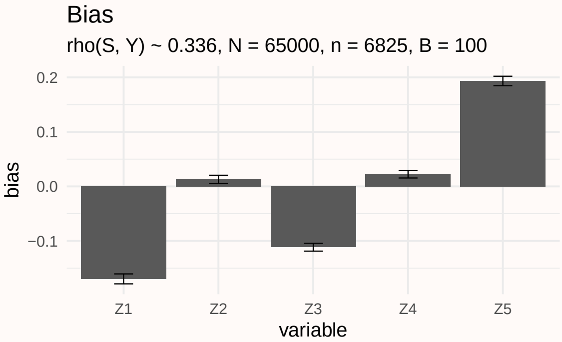


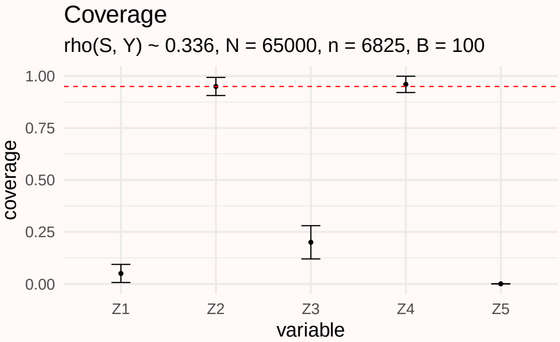


MCSE errorbars.

A Simple Simulation - MNAR and large n

$N = 65000, \quad f = 0.105, \quad p = 5$, MNAR Sample





MCSE errorbars.

A very broad problem

Prof. Xiao-Li Meng

Big Data Paradox:
The bigger the data, the surer we fool ourselves

Random Samples: An Endangered Species

Are representative samples possible in the social sciences?

"..the idea will rarely work in a complicated social problem because we always have additional variables that may have important consequences for the outcome." Kruskal and Mosteller (1979)

Random Samples: An Endangered Species

Are representative samples possible in the social sciences?

"..the idea will rarely work in a complicated social problem because we always have additional variables that may have important consequences for the outcome." Kruskal and Mosteller (1979)

Cost of single survey

internet panel	mail	phone	face to face
$< 10$$	48$	81$	192$

Diffusion of Online Surveys

In 2010, 31% of all surveys in Germany were online. Heen et al. (2014)

References

- A wide review on the statistical use of non-probability samples and polls across different disciplines:
  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.
- Exemplifies the Big Data Paradox and introduced the concept of d.d.i., similar to Γ:
  Meng, X. L. (2018). Statistical paradises and paradoxes in big data (i) law of large populations, big data paradox, and the 2016 us presidential election. The Annals of Applied Statistics, 12(2), 685-726.
- Manuscript (pre-print) presenting the Rockafellar-Uryasev regression:
  Sahoo, R., Lei, L., Wager, S. (2022). Learning from a biased sample. arXiv preprint arXiv:2209.01754.

References

Craig K Enders. *Applied missing data analysis.* Guilford Publications, 2022.

Miliaikeala Heen, Joel D Lieberman, and TD Meithe. A comparison of different online sampling approaches for generating national samples. 2014.

William Kruskal and Frederick Mosteller. Representative sampling, iii: The current statistical literature. *International Statistical Review/Revue Internationale de Statistique*, pages 245–265, 1979.

Xiao-Li Meng. Statistical paradises and paradoxes in big data (i) law of large populations, big data paradox, and the 2016 us presidential election. *The Annals of Applied Statistics*, 12(2):685–726, 2018.

Roshni Sahoo, Lihua Lei, and Stefan Wager. Learning from a biased sample. *arXiv preprint arXiv:2209.01754*, 2022.