Sampling and missingness

Week 3: Sampling, Missingness, and Bias

PSTAT100 Winter 2025

Announcements

This week

Objective: Enable you to critically assess data quality based on how it was collected.

- **Sampling and statistical bias**
  - Sampling terminology
  - Common sampling scenarios
  - Sampling mechanisms
  - Statistical bias
- **The missing data problem**
  - Types of missingness: MCAR, MAR, and MNAR
  - Pitfalls and simple fixes
- **Principles of Measurement**

Warmup Example

> **Note**
>
> What is the typical family size (children only)?

Summarizing the Data

- Summary: c
- Data: $y_1, \ldots, y_n$
- Error: $y_1 - c, \ldots, y_n - c$
- Loss: l: $R \to R^+$

Summarizing the Data

Average Loss: $\frac{1}{n} \sum l(y_i, c)$ is also known as **Empirical Risk**

We can try to find the value c that minimizes the empirical risk for any loss function.

Minimize the Average Loss

$$\frac{1}{n}\sum l(y_i, c) = \frac{1}{n}\sum (y_i - c)^2$$

The Sample Average Minimizes the Empirical Risk

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 <= \frac{1}{n}\sum_{i=1}^{n}(y_i - c)^2$$

Considering the

Question more Carefully {.smaller}

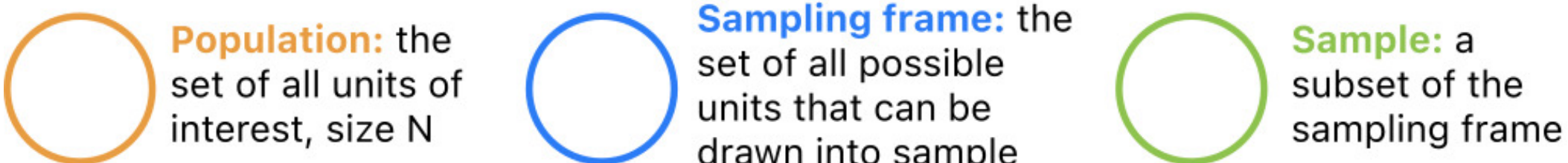
- What is the typical family size (children only)?
- What are we trying to measure?
- How can we measure this?

Sampling terminology

Here we'll introduce standard statistical terminology to describe data collection.

All data are collected somehow. A **sampling design** is a ***way of selecting observational units for measurement***. It can be construed as a particular relationship between:

- a **population** (all entities of interest);
- a **sampling frame** (all entities that are possible to measure); and
- a **sample** (a specific collection of entities).



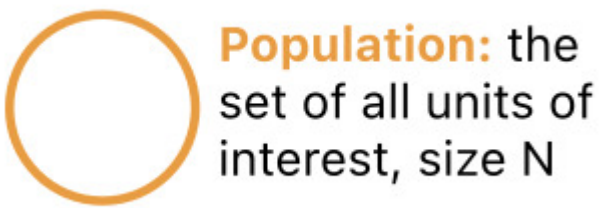
Population

Last week, we introduced the terminology **observational unit** to mean ***the entity measured for a study*** – datasets consist of observations made on observational units.

In less technical terms, all data are data *on* some kind of thing, such as countries, species, locations, and the like.

A statistical **population** is the ***collection of all units of interest***. For example:

- all countries (GDP data)
- all mammal species (Allison 1976)
- all babies born in the US (babynames data)
- all locations in a region (SB weather data)
- all adult U.S. residents (BRFSS data)



Sampling frame

There are usually some units in a population that can't be measured due to practical constraints – for instance, many adult U.S. residents don't have phones or addresses.

For this reason, it is useful to introduce the concept of a **sampling frame**, which refers to ***the collection of all units in a population that can be observed for a study***. For example:

Sampling frame: the set of all possible units that can be drawn into sample

- all countries reporting economic output between 1961 and 2019
- all babies with birth certificates from U.S. hospitals born between 1990 and 2018
- all adult U.S. residents with phone numbers in 2019

Sample

Finally, it's rarely feasible to measure every observable unit due to limited data collection resources – for instance, states don't have the time or money to call every phone number every year.

A **sample** is ***a subcollection of units in the sampling frame actually selected for study***. For instance:

- 234 countries;
- 62 mammal species;
- 13,684,689 babies born in CA;
- 1 weather station location at SB airport;
- 418,268 adult U.S. residents.

Sampling scenarios

We can now imagine a few common sampling scenarios by varying the relationship between population, frame, and sample.

Denote an observational unit by U_i, and let:

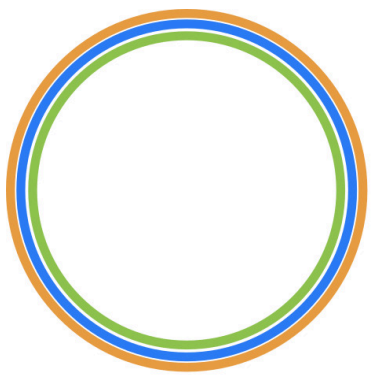
$$
\begin{aligned}
&\mathcal{U} = \{U_i\}_{i \in I} && (\text{universe}) \\
&P = \{U_1, \ldots, U_N\} \subseteq \mathcal{U} && (\text{population}) \\
&F = \{U_j : j \in J \subset I\} \subseteq P && (\text{frame}) \\
&S \subseteq F && (\text{sample})
\end{aligned}
$$

Census

The simplest scenario is a **population census**, where the entire population is observed.

For a census: $S = F = P$

*All properties of the population are definitevely **known** in a census.* So there is no need to model census data.



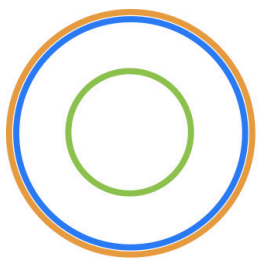
Census: sample covers entire population

Simple random sample

The statistical gold standard for inference, modeling, and prediction is the **simple random sample** in which units are selected at random from the population.

For a simple random sample: $S \subset F = P$

Sample properties are reflective of population properties in simple random samples. Population inference is straightforward.



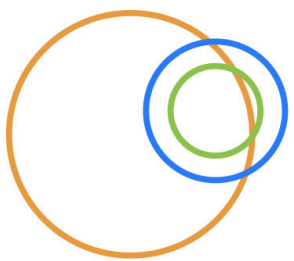
Ideal random sample: sampling frame is the population and the sample is drawn at random

‘Typical’ sample

More common in practice is a random sample from a sampling frame that overlaps but does not cover the population.

For a ‘typical’ sample: $S \subset F \quad \text{and} \quad F \cap P \neq \emptyset$

Sample properties are ***reflective of the frame*** *but not necessarily the study population.* Population inference gets more complicated and may not be possible.



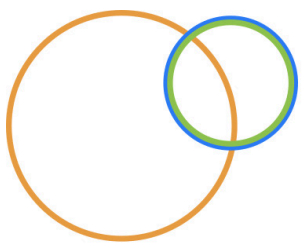
Typical sample in practice: sampling frame partially overlaps with population

‘Administrative’ data

Also common is **administrative data** in which all units are selected from a convenient frame that partly covers the population.

For administrative data: $S = F \quad \text{and} \quad F \cap P \neq \emptyset$

Administrative data are not really proper samples; they cannot be replicated and they do not represent any broader group. No inference is possible.



Administrative data: sample covers entire sampling frame, but sampling frame does not cover population

Scope of inference

The relationships among the population, frame, and sample determine the **scope of inference**: the ***extent to which conclusions based on the sample are generalizable***.

A good sampling design can ensure that the statistical properties of the sample are expected to match those of the population. If so, it is sound to generalize:

- the sample is said to be *representative* of the population
- the scope of inference is *broad*

A poor sampling design will produce samples that distort the statistical properties of the population. If so, it is not sound to generalize:

- sample statistics are subjet to bias
- the scope of inference is *narrow*

Characterizing sampling designs

The sampling scenarios above can be differentiated along two key attributes:

1. The overlap between the sampling frame and the population.
    - frame $=$ population
    - frame $\subset$ population
    - frame $\cap$ population $\neq \emptyset$
2. The mechanism of obtaining a sample from the sampling frame.
    - random sampling
    - convenience sampling

If you can articulate these two points, you have fully characterized the sampling design.

Sampling mechanisms

In order to describe sampling mechanisms precisely, we need a little terminology.

Each unit has some **inclusion probability** – ***the probability of being included in the sample***.

Let's suppose that the frame F comprises N units, and denote the inclusion probabilities by:

$$p_i = P(\text{unit } i \text{ is included in the sample}) \quad i = 1, \ldots, N$$

The inclusion probability of each unit depends on the physical procedure of collecting data.

Sampling mechanisms

Sampling mechanisms are ***methods of drawing samples*** and are categorized into four types based on inclusion probabilities.

- in a **census** every unit is included
  - $p_i = 1$ for every unit $i = 1, \ldots, N$
- in a **random sample** every unit is equally likely to be included
  - $p_i = p_j$ for every pair of units i, j
- in a **probability sample** units have different inclusion probabilities
  - $p_i \neq p_j$ for at least one $i \neq j$
- in a **nonrandom sample** there is no random mechanism
  - $p_i = 1$ for $i \in S$

Revisiting example datasets: GDP

Annual observations of GDP growth for 234 countries from 1961 - 2018.

- Population: all countries in existence between 1961-2019.
- Frame: all countries reporting economic output for at least one year between 1961 and 2019.
- Sample: equal to frame.

So:

1. Overlap: frame partly overlaps population.
2. Mechanism: sample is every country in the sampling frame.

This is administrative data with no scope of inference.

Revisiting example datasets: BRFSS data

Phone surveys of 418K U.S. residents in 2019.

- Population: all U.S. residents.
- Frame: all adult U.S. residents with phone numbers.
- Sample: 418K adult U.S. residents with phone numbers.

So:

1. Overlap: frame is a subset of the population.
2. Mechanism: probability sample.
    - Randomly selected phone numbers were dialed in each state, so individuals in less populous states or with multiple numbers are more likely to be included

This is a typical sample with narrow inference to adult residents with phone numbers.

Statistical bias

Statistical **bias** is the average difference between a sample property and a population property across all possible samples under a particular sampling design.

In less technical terms: the expected error of estimates.

Two possible sources of statistical bias:

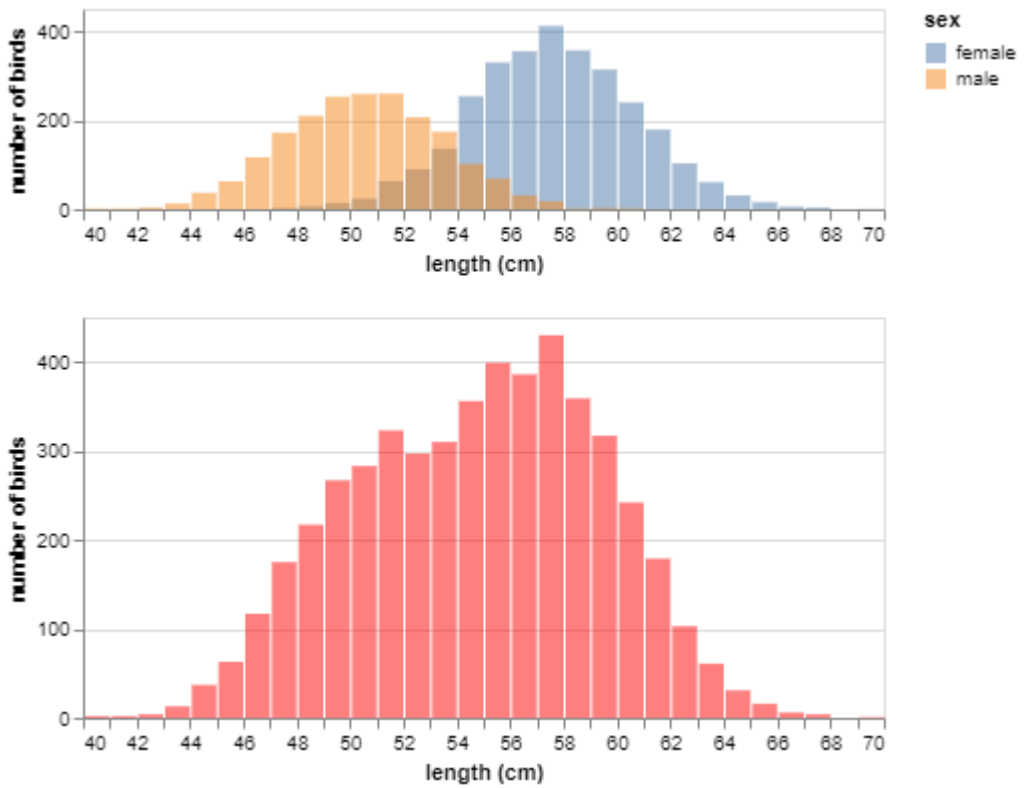
- An estimator systematically over- or under-estimates its target population property
  - *e.g.*, $\frac{1}{n}\sum_i (x_i - \bar{x})^2$ is biased for (underestimates) the population variance
- Sampling design systematically over- or under-represents certain observational units
  - *e.g.*, studies conducted on college campuses are biased towards (overrepresent) young adults

These are distinct from other kinds of bias that we are **not** discussing:

- Measurement bias: attributes or outcomes are measured unevenly across populations
- Experimenter bias: study design and/or outcomes favor an investigator's preconceptions

Sampling bias

In Lab 2 you'll explore sampling bias arising from sampling mechanisms. Here's a preview:



Distributions of body length by sex (top) and in aggregate (bottom) for a hypothetical population of 5K hawks.

Consider:

1. Are males or females generally longer?
2. How will the sample mean shift if disproportionately more males are sampled?
3. If disproportionately more females are sampled?

```
# A tibble: 2 × 2
  sex    `n()`
  <chr>  <int>
```

Bias corrections

If inclusion probabilities are known or estimable it is possible to apply bias corrections to estimates using inverse probability weighting.

If

- p_i is the probability that individual i is included in the sample S
- Y_i are observations of a variable of interest

Then a bias-corrected estimate of the population mean is given by the weighted average:

$$\sum_{i \in S} \left(\frac{p_i^{-1}}{\sum_i p_i^{-1}} \right) Y_i$$

Bias correction example

Suppose we obtain a biased sample in which female hawks were 9 times as likely to be selected as males. This yields an overestimate:

```
cat('population mean: ', mean(pop$length), "\n")
```

```
population mean:  54.06243
```

```
cat('sample mean: ', mean(samp$length), "\n")
```

```
sample mean:  56.58709
```

But since we know the exact inclusion probabilities up to a proportionality constant, we can apply inverse probability weighting to adjust for bias:

```
# Create weights dataframe
weight_df <- tibble(
  sex = c("male", "female"),
  weight = c(1, 9)
)

# Join weights to sample and calculate weighted mean
samp_w <- samp %>%
  left_join(weight_df, by = "sex") %>%
  mutate(
    correction_factor = (1/weight) / sum(1/weight),
    weighted_length = length * correction_factor
  ) %>%
  summarise(weighted_mean = sum(weighted_length)) %>%
  pull(weighted_mean)
```

```
16
17 print(samp_w)
```

```
[1] 53.4902
```

Bias correction example

However, even if we *didn't* know the exact inclusion probabilities, we could estimate them from the sample:

```
table(samp$sex)
```

```
female   male
    87     13
```

And use the same approach:

```
# Calculate ratio of females to males in sample
ratio <- samp %>%
  count(sex) %>%
  pivot_wider(names_from = sex, values_from = n) %>%
  mutate(ratio = female/male) %>%
  pull(ratio)

# Create weights dataframe
weight_df <- tibble(
  sex = c("male", "female"),
  weight = c(1, ratio)
)

# Join weights and calculate weighted mean
samp_w <- samp %>%
  left_join(weight_df, by = "sex") %>%
  mutate(
    correction_factor = (1/weight) / sum(1/weight),
    weighted_length = length * correction_factor
```

```r
  ) %>%
  summarise(weighted_mean = sum(weighted_length)) %>%
  pull(weighted_mean)

print(samp_w)
```

```
[1] 54.00361
```

Remarks on IPW and bias correction

Inverse probability weighting can be applied to correct a wide range of estimators besides averages.

It is also applicable to adjust for bias due to missing data.

In principle, the technique is simple, but in practice, there are some common hurdles:

- usually inclusion probabilities are not known
- estimating inclusion probabilities can be difficult and messy

Missingness

Missing data arise when ***one or more variable measurements fail for a subset of observations***.

This can happen for a variety of reasons, but is very common in pratice due to, for instance:

- equipment failure;
- sample contamination or loss;
- respondents leaving questions blank;
- attrition (dropping out) of study participants.

Many researchers and data scientists ignore missingness by simply deleting affected observations, but this is bad practice! Missingness needs to be treated carefully.

Missing representations

It is standard practice to record observations with missingness but enter a special symbol (`..`, `–`, `NA`, etcetera) for missing values.

In R, missing values are mapped to a indicator denoted `NA`.

Here is some made-up data with two missing values:

```
1 library(tidyverse)
2 read_csv("data/some_data.csv")
```

```
# A tibble: 8 × 2
    obs value
  <dbl> <chr>
1     0 -0.9286936933427271
2     1 -0.3088381742999848
3     2 –
4     3 -1.4345064041945543
5     4 0.03958917896644836
6     5 –
7     6 -0.5316890502224456
8     7 1.4734842645335422
```

Notice that R thinks the `value` shoudl be a `chr` type because of the `–` symbols.

Missing representations

`read_csv` has the ability to map specified entries to `NA` when parsing data files:

```
1  library(tidyverse)
2  some_data <- read_csv("data/some_data.csv", na="-")
3  some_data
```

```
# A tibble: 8 × 2
    obs   value
  <dbl>   <dbl>
1     0 -0.929
2     1 -0.309
3     2 NA
4     3 -1.43
5     4  0.0396
6     5 NA
7     6 -0.532
8     7  1.47
```

Calculations with NA

Summarizing data with missing values typically returns `NA`:

```
mean(some_data$value)
```

```
[1] NA
```

Many functions can ignore missing values by setting the `na.rm` argument:

```
mean(some_data$value, na.rm=TRUE)
```

```
[1] -0.2817756
```

Those missing values could have been anything, and ignoring them changes the result from what it would have been!

```
# one counterfactual scenario
some_data$value[c(3, 6)] <- c(5, 6)
mean(some_data$value, na.rm=TRUE)
```

```
[1] 1.163668
```

The missing data problem

In a nutshell, the **missing data problem** is: ***how should missing values be handled in a data analysis?***

> Getting the software to run is one thing, but this alone does not address the challenges posed by the missing data. Unless the analyst, or the software vendor, provides some way to work around the missing values, the analysis cannot continue because calculations on missing values are not possible. There are many approaches to circumvent this problem. Each of these affects the end result in a different way. (Stef van Buuren, 2018)

There's no universal approach to the missing data problem. The choice of method depends on:

- the analysis objective;
- the *missing data mechanism*.

Missing data in PSTAT100

We won't go too far into this topic in PSTAT 100. Our goal will be awareness-raising, specifically:

- characterizing types of missingness (missing data mechanisms);
- understanding missingness as a potential source of bias;
- basic do's and don't's when it comes to missingness.

If you are interested in the topic, Stef van Buuren's *Flexible Imputation of Missing Data* (the source of one of your readings this week) provides an excellent introduction.

Missing data mechanisms

Missing data mechanisms (like sampling mechanisms) are characterized by the probabilities that observations go missing.

For dataset $X = \{x_{ij}\}$ comprising

- n rows/observations
- p columns/variables

denote the probability that a value goes missing as:

$$q_{ij} = P(x_{ij} \text{ is missing})$$

Missing completely at random

Data are **missing completely at random** (MCAR) if the ***probabilities of missing entries are uniformly equal***.

$$q_{ij} = q \quad \text{for all} \quad i = 1, \ldots, n \quad \text{and} \quad j = 1, \ldots, p$$

This implies that the cause of missingness is unrelated to the data: missing values can be ignored. This is the easiest scenario to handle.

Missing at random

Data are **missing at random** (MAR) if the ***probabilities of missing entries depend on observed data***.

$$q_{ij} = f(\mathbf{x}_i)$$

This implies that information about the cause of missingness is captured within the dataset. As a result:

- it is possible to estimate q_{ij}
- bias corrections using inverse probability weighting can be implemented

Missing not at random

Data are **missing not at random** (MNAR) if the ***probabilities of missing entries depend on unobserved data***.

$$q_{ij} = f(z_i, x_{ij}) \quad z_i \text{ unknown}$$

This implies that information about the cause of missingness is unavailable. This is the most complicated scenario.

Assessing the missing data mechanism

Importantly, there is no easy diagnostic check to distinguish MCAR, MAR, and MNAR without measuring some of the missing data.

So in practice, usually one has to make an informed assumption based on knowledge of the data collection process.

Example: GDP data

In the GDP growth data, growth measurements are missing for many countries before a certain year.

We might be able to hypothesize about why – perhaps a country didn't exist or didn't keep reliable records for a period of time.However, the data as they are contain no additional information that might explain the cause of missingness.

So these data are MNAR.

Handling missing data

Simple approaches:

1. **Drop observations with missing values** (`drop_na()`)
    - Works if data are MCAR
    - Induces bias if MAR or MNAR
    - Causes information loss
2. **Mean imputation** (`replace_na()`)
    - Only suitable for very small proportion of missing values
    - Distorts distributions
    - Induces bias if MAR or MNAR

Mean imputation

```r
gdp <- read_csv('data/annual_growth.csv')
gdp
```

```
# A tibble: 264 × 61
   `Country Name`       `Country Code` `1961` `1962` `1963` `1964` `1965` `1966`
   <chr>                <chr>           <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
 1 Aruba                ABW            NA    NA      NA      NA     NA   NA
 2 Afghanistan          AFG            NA    NA      NA      NA     NA   NA
 3 Angola               AGO            NA    NA      NA      NA     NA   NA
 4 Albania              ALB            NA    NA      NA      NA     NA   NA
 5 Andorra              AND            NA    NA      NA      NA     NA   NA
 6 Arab World           ARB            NA    NA      NA      NA     NA   NA
 7 United Arab Emirates ARE            NA    NA      NA      NA     NA   NA
 8 Argentina            ARG             5.43 -0.852  -5.31   10.1   10.6 -0.660
 9 Armenia              ARM            NA    NA      NA      NA     NA   NA
10 American Samoa       ASM            NA    NA      NA      NA     NA   NA
# ℹ 254 more rows
# ℹ 53 more variables: `1967` <dbl>, `1968` <dbl>, `1969` <dbl>, `1970` <dbl>,
#   `1971` <dbl>, `1972` <dbl>, `1973` <dbl>, `1974` <dbl>, `1975` <dbl>,
#   `1976` <dbl>, `1977` <dbl>, `1978` <dbl>, `1979` <dbl>, `1980` <dbl>,
#   `1981` <dbl>, `1982` <dbl>, `1983` <dbl>, `1984` <dbl>, `1985` <dbl>,
#   `1986` <dbl>, `1987` <dbl>, `1988` <dbl>, `1989` <dbl>, `1990` <dbl>,
#   `1991` <dbl>, `1992` <dbl>, `1993` <dbl>, `1994` <dbl>, `1995` <dbl>, …
```

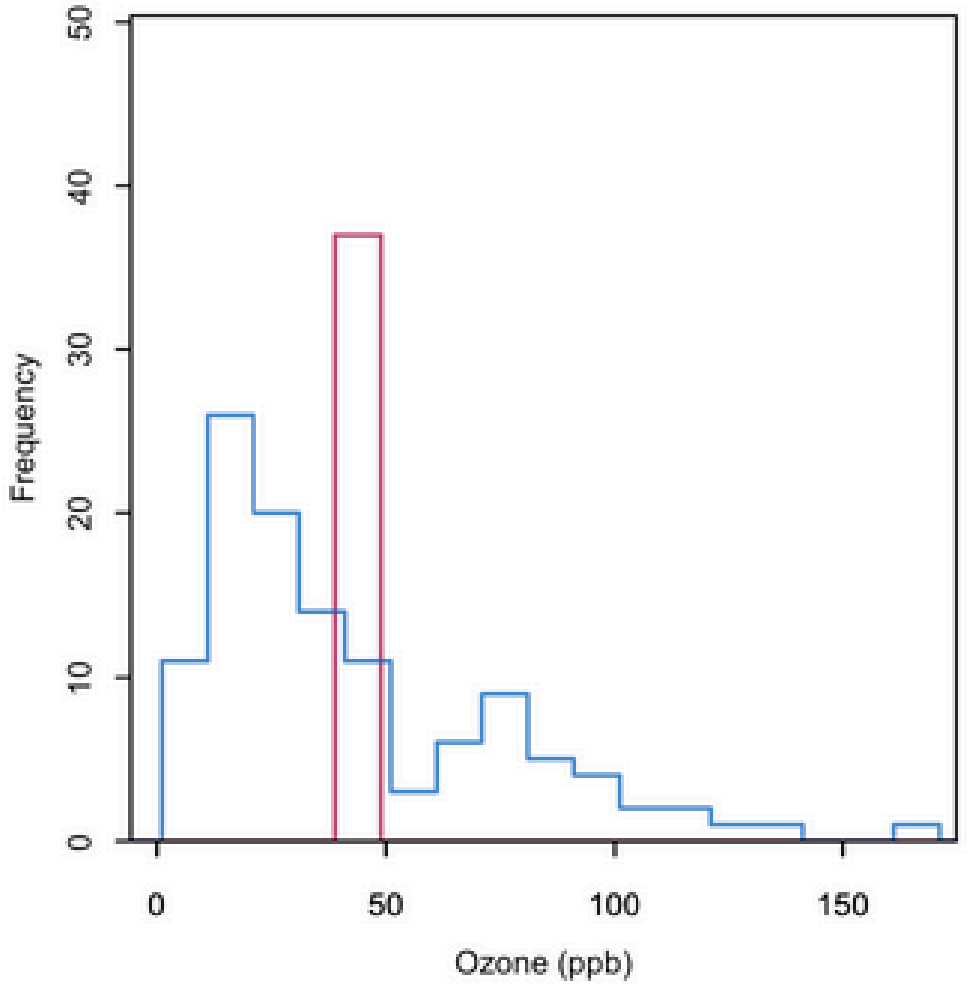
Mean imputation

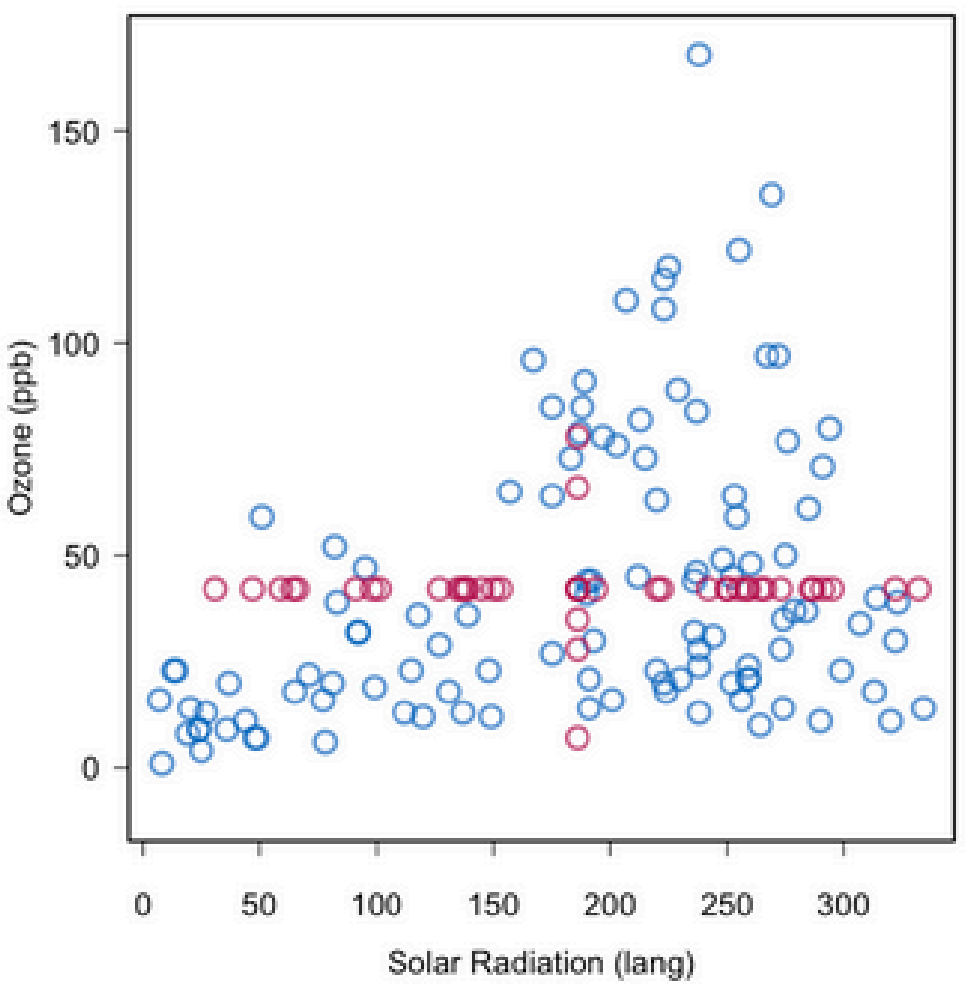
Replace each country GDP with the geometric mean

```
geometric_mean <- \(x) prod(x, na.rm=TRUE) ^ (1/sum(!is.na(x)))
gdp |> pivot_longer(cols=where(is.numeric), names_to="Year", values_to="gdp") |>
  group_by(`Country Name`) |>
  mutate(gdp = 1 + gdp/100) |>
  mutate(gdp = replace_na(gdp, replace = geometric_mean(gdp)))
```

```
# A tibble: 15,576 × 4
# Groups:   Country Name [264]
   `Country Name` `Country Code` Year    gdp
   <chr>          <chr>          <chr> <dbl>
 1 Aruba          ABW            1961   1.03
 2 Aruba          ABW            1962   1.03
 3 Aruba          ABW            1963   1.03
 4 Aruba          ABW            1964   1.03
 5 Aruba          ABW            1965   1.03
 6 Aruba          ABW            1966   1.03
 7 Aruba          ABW            1967   1.03
 8 Aruba          ABW            1968   1.03
 9 Aruba          ABW            1969   1.03
10 Aruba          ABW            1970   1.03
# ℹ 15,566 more rows
```

Perils of mean imputation





Imputing too many missing values distorts the distribution of sample values.

Other common approaches to missingness

When data are MCAR or MAR, one can:

- model the probability of missingness and apply bias corrections to estimated quantities using inverse probability weighting
- model the variables with missing observations as functions of the other variables and perform model-based imputation

Do’s and don’t’s

Do:

1. Always check for missing values *upon import*.
    - Tabulate the proportion of observations with missingness
    - Tabulate the proportion of values for each variable that are missing
2. Take time to find out the reasons data are missing.
    - Determine which outcomes are coded as missing.
    - Investigate the physical mechanisms involved.
3. Report missing data if they are present.

Don’t:

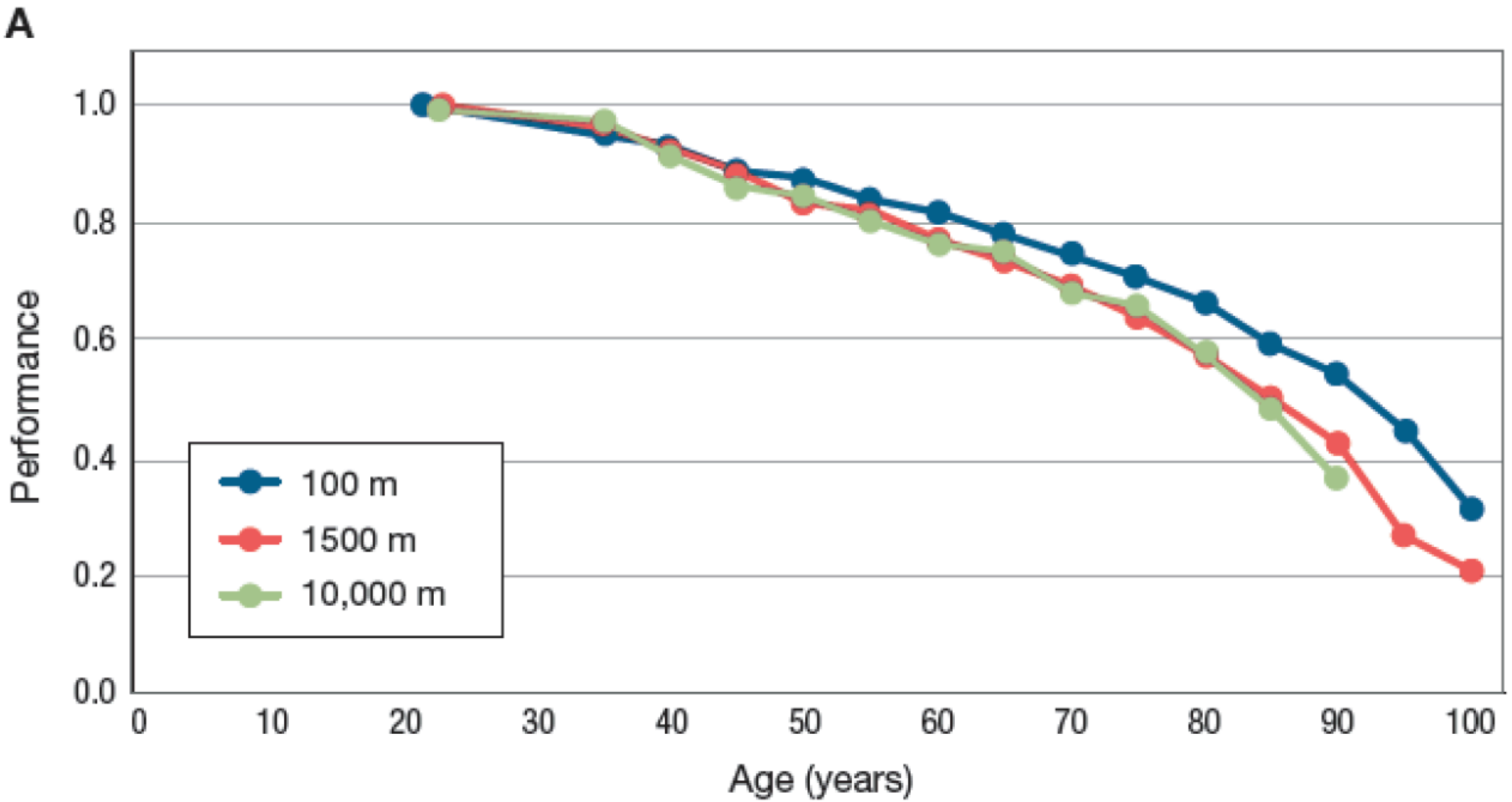
1. Rely on software defaults for handling missing values.
2. Drop missing values if data are not MCAR.

Principles of Measurement

Criteria for satisfactory measurement include:

- relevance: measurements are relevant to the question
- reliability: measurements are precise enough to answer the question
- non-distorting: measurement process does not distort the system under study
  - If it does, can you account for it?
- cost: cost can be monetary, social, ethical etc

Declining Performance with Age



Declining Performance with Age

Not a problem:

- We Likely have adequate precision
  - Track & fields records are very precise
- No measurement distortion
  - Athletes aren’t affected by observation

Declining Performance with Age

Possibly a problem:

- We aren’t measuring the relevant quantity
  - Each point on the curve comes from a different individual
  - Curve shows record performance, may not be indicative of typical decline
- Sample size and selection
  - What would happen if sample size decreases with age?

Declining Performance with Age

```
1 max(rnorm(100000))
```

```
[1] 4.48761
```

```
1 max(rnorm(1000))
```

```
[1] 2.853695
```

```
1 max(rnorm(10))
```

```
[1] 2.22357
```

Measurement distortion

Measurement distortion

Table 1. Behavior of marmosets in the hidden and visible observers conditions (* significant differences between conditions, Student t test, $p < 0.05$)

	Experimental Conditions	
Parameters	**Blind-hidden observers**	**Visible observers**
Latency to arrive at the feeder (min.)	102.3 ± 37.86	53.0 ± 20.03
Duration of the session (min.)	53.5 ± 5.47	51.16 ± 6.16
Distance from the observers (meters)	5.01 ± 1.56	2.21 ± 0.92
Alarm calls (frequency/10 min.)	0.77 ± 0.09	8.65 ± 2.42 *
Mobbing calls (frequency/10 min.)	0.58 ± 0.28	3.69 ± 2.60
Staring at the observers (frequency/10 min.)	1.28 ± 0.27	22.31 ± 6.31 *
Staring in other directions (frequency/10 min.)	5.14 ± 0.32	1.22 ± 0.58 *
Phee calls (frequency/10 min.)	5.49 ± 2.25	6.45 ± 3.19
Other vocalizations (frequency/10 min.)	4.16 ± 2.02	2.77 ± 0.80

http://pepsic.bvsalud.org/scielo.php?pid=S1517-28052006000200004&script=sci_abstract&tlng=en

Summary