INTRODUCTION TO DATA

Jose Toledo Luna

2026-09-29

INTRODUCTION

Statistics focuses on addressing the following three questions:

1. How can data best be collected ?
2. How should data be analyzed ?
3. What can be inferred from data ?

We will explore brief a discussion

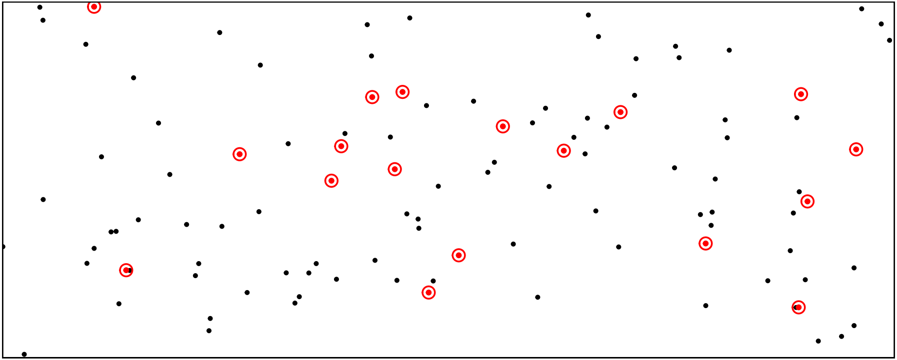
- Principles of data collection
- Summarizing and exploring data

DATA COLLECTION PRINCIPLES

SAMPLING

SIMPLE RANDOM SAMPLING

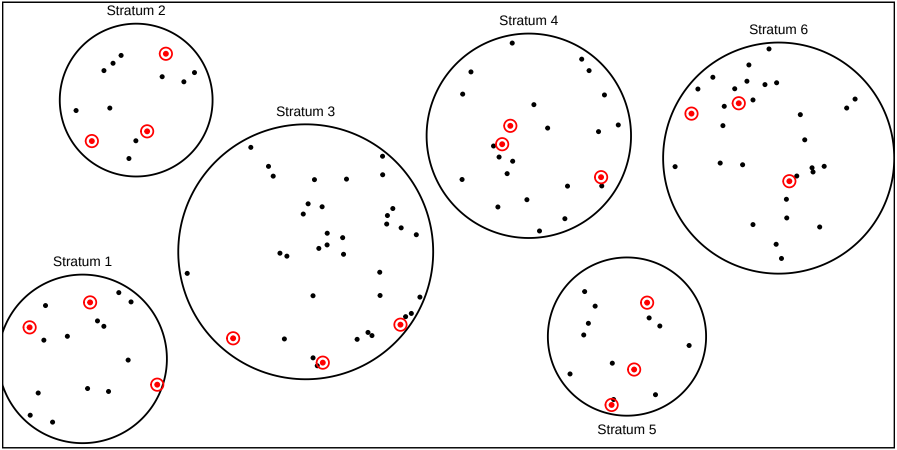
In a simple random sample, each case in the population has an equal chance of being included in the sample



Each case is sampled independently of the other cases

STRATIFIED

In stratified sampling, the population is first divided into groups called strata before cases are selected within each stratum (typically through simple random sample)

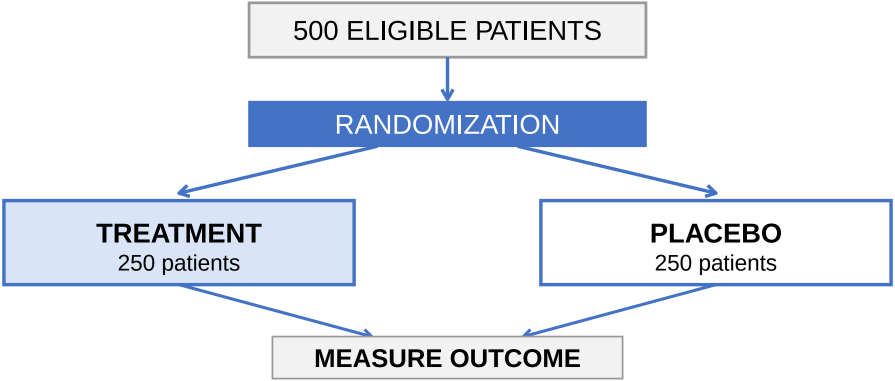


EXPERIMENTS AND OBSERVATIONAL STUDIES

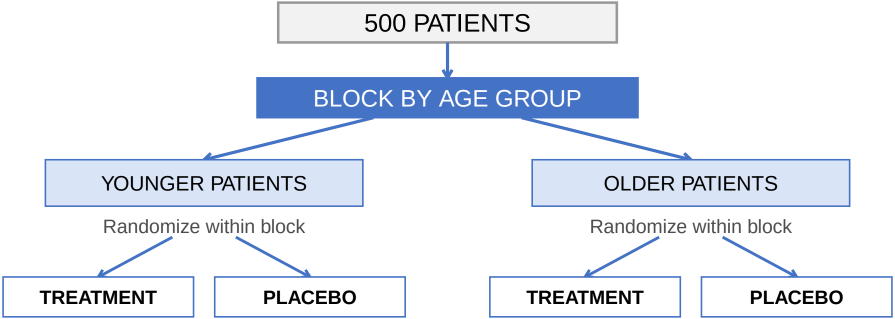
PRINCIPLES OF EXPERIMENTAL DESIGN

1. Control: Control for the *(potential)* effect of variables other than the ones directly studied
2. Randomize: Randomly assigning patients to treatment groups ensures that groups are balanced with respect to both variables that can and cannot be controlled.
3. Replicate: Within a study, replicate by collecting a sufficiently large sample. Or replicate the entire study
4. Block: If there are variables that are known or suspected to affect the response variable, first group subjects into blocks based on these variables, and then randomize cases within each block to treatment groups

RANDOMIZED EXPERIMENT



RANDOMIZED EXPERIMENT WITH BLOCKING



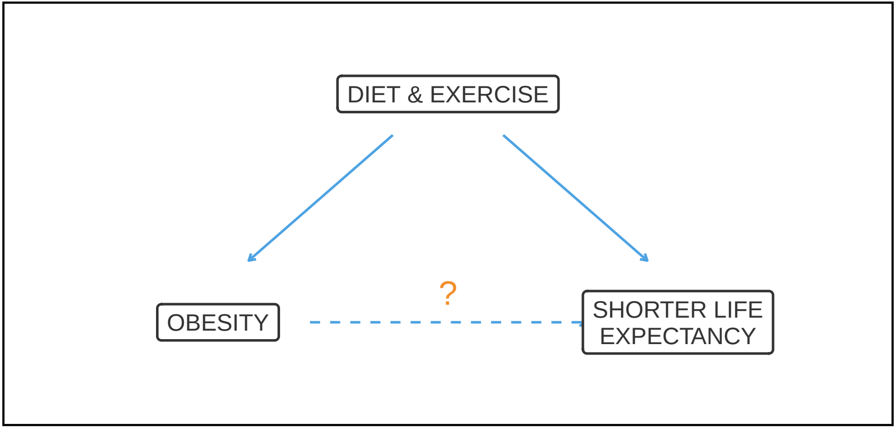
Compare treatment and placebo outcomes while accounting for age group

OBSERVATIONAL STUDIES

In observational studies, researchers simply observe selected potential explanatory and response variables

Only randomized controlled experiments should make causal conclusions about potential relationship between explanatory and response variables

Observational studies of obesity have shown that obese individuals tend to die sooner than individuals with normal weight



It would be misleading to conclude that obesity causes shorter life expectancy. Here diet and exercise are known as confounding variables, they are associated with both explanatory and response variables.

DATA FUNDAMENTALS

NORTH CAROLINA BIRTHS DATA

In 2004, the state of North Carolina released to the public a large data set containing information on births recorded in this state

This data set has been of interest to medical researchers who are studying the relation between habits and practices of expectant mothers and the birth of their children

This is a random sample of 1,000 cases from this data set, which can be accessed through openintro R package

```
births <- openintro::ncbirths
```

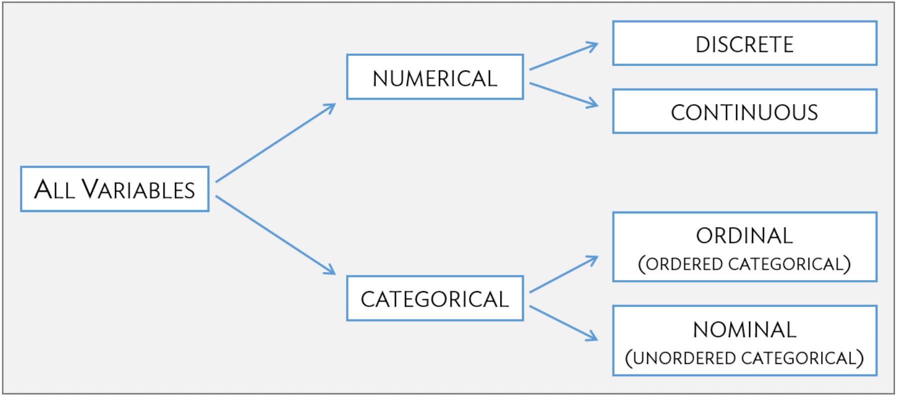
fage	mage	mature	weeks	premie	visits
NA	13	younger mom	39	full term	10
NA	14	younger mom	42	full term	15
19	15	younger mom	37	full term	11
21	15	younger mom	41	full term	6
NA	15	younger mom	39	full term	9

marital	gained	weight	lowbirthweight	gender	habit	whitemom
not married	38	7.63	not low	male	nonsmoker	not white
not married	20	7.88	not low	male	nonsmoker	not white
not married	38	6.63	not low	female	nonsmoker	white
not married	34	8.00	not low	male	nonsmoker	white
not married	27	6.38	not low	female	nonsmoker	not white

VARIABLE DESCRIPTIONS

Variable	Description
fage	Father's age in years.
mage	Mother's age in years.
mature	Maturity status of mother.
weeks	Length of pregnancy in weeks.
premie	Whether the birth was classified as premature (premie) or full-term.
visits	Number of hospital visits during pregnancy.
gained	Weight gained by mother during pregnancy in pounds.
weight	Weight of the baby at birth in pounds.
lowbirthweight	Whether baby was classified as low birthweight (low) or not (not low).
gender	Gender of the baby, female or male.
habit	Status of the mother as a nonsmoker or a smoker.
marital	Whether mother is married or not married at birth.
whitemom	Whether mom is white or not white.

TYPES OF VARIABLES



Numerical variables take on numerical values, such that numerical operations (sums, differences, etc.) are reasonable.

- *Discrete:* only take on integer values (i.e., # of family members)
- *Continuous:* can take on any value within a specified range (i.e., height)

Categorical variables take on values that are names or labels; the possible values are called the variable's levels.

- *Ordinal:* exists some natural ordering of levels (i.e., education)
- *Nominal:* no natural ordering of levels (i.e., gender)

NUMERICAL DATA

SUMMARIZING NUMERICAL DATA

Distribution: The collection of values for a numerical, continuous variable (i.e, weight, height)

We can describe and visualize summaries of a distribution using the following characteristics:

- Center: where is the middle of the distribution ?
- Shape: where and how many peaks does the distribution have ?
- Unusual Observations: are there any observations unusally low/high relative to the bulk of the observations ?
- Variability: how similar or varied are the values to each other ?

Remember the acronym CSUV

MEASURES OF CENTER

The sample mean of a numerical variable is the sum of all observations divided by the number of observations

$$\bar{x} = \frac{1}{n}\left(x_1 + x_2 + \cdots + x_n\right)$$

where $x_1, x_2, \ldots, x_n$ represent the n observed values in a sample.

In R, use the mean() function

```
> mean()
```

THe median is the value of the middle observation in a sample

If the number of observation (n) is

- Odd: The median is the middle observation
- Even: The median is the average of the two middle observations

In R, use the `median()` function

```
> median()
```

The median is the 50^{th} percentile, 50% of the observations lie below/above the median

MEASURES OF SPREAD

The sample variance s^2 is

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}$$

where $(x_i - \bar{x})$ is an observation's deviation, the distance between its value x and the same mean $\bar{x}$

The standard deviation measures *approximately* the distance between an observation and the mean

$$s = \sqrt{s^2} = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}}$$

It is obtained by taking the square root of the variance

To obtain the standard deviation (s) in R, use the sd() function

```r
sd(births$weight)
```

```
[1] 1.50886
```

use var() to obtain the sample variance,

```r
var(births$weight)
```

```
[1] 2.276659
```

we can verify $s = \sqrt{s^2}$

```r
sd(births$weight)^2
```

```
[1] 2.276659
```

GUIDED PRACTICE

Given the following observations $[x_1, x_2, \ldots, x_6] = [3, 2, 0, 5, 1, 4]$.

Obtain $\bar{x}$ and s^2.

FIVE NUMBER SUMMARY

A five-number summary is a set of descriptive statistics that provides a quick overview of a dataset's distribution

It consists of the following five values ordered from lowest to highest:

- Minimum: The smallest value in the dataset
- First Quartile (Q_1): The 25th percentile
- Median (Q_2): The middle value
- Third Quartile (Q_3): The 75th percentile
- Maximum: The largest value in the dataset

> **Note**
>
> The p^{th} percentile is the observation such that p% olf the remaining observations fall below this observation

To obtain the five number summary, use fivenum() function in R

```r
fivenum(births$weight)
```

```
[1]  1.00  6.38  7.31  8.06 11.75
```

QUARTILES

The interquartile range *(IQR)* is the distance between the third and first quartiles

$$IQR = Q_3 - Q_1$$

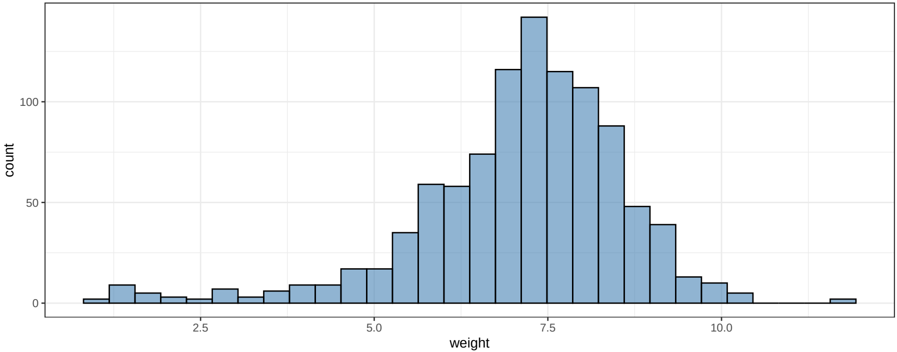
IQR measures the statistical spread of the middle 50% of a dataset. In R, use IQR() function

```
IQR(births$weight)
```

```
[1] 1.68
```

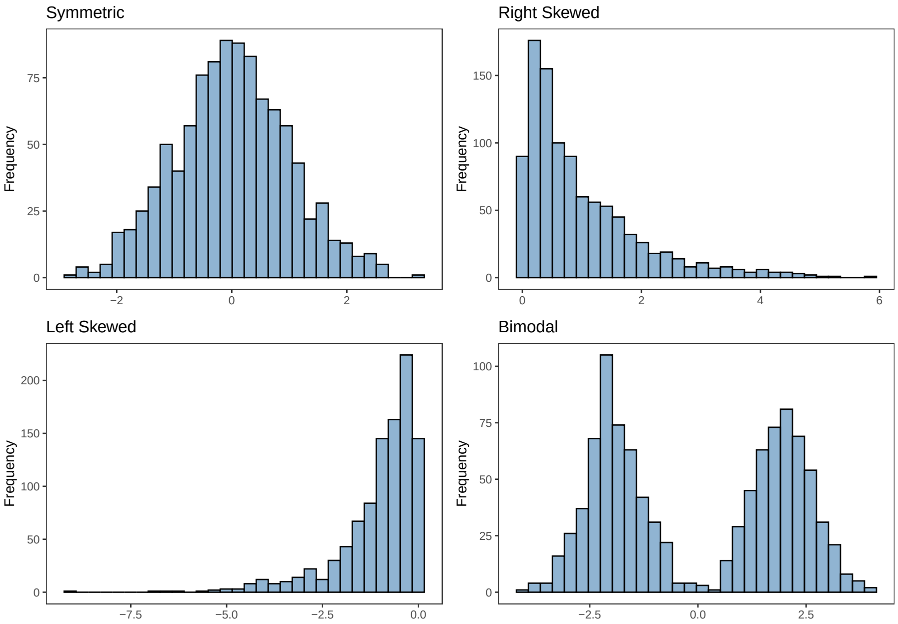
VISUALIZING NUMERICAL DATA

Histograms show important features of the shape of a distribution



Histograms provide a view of the data density

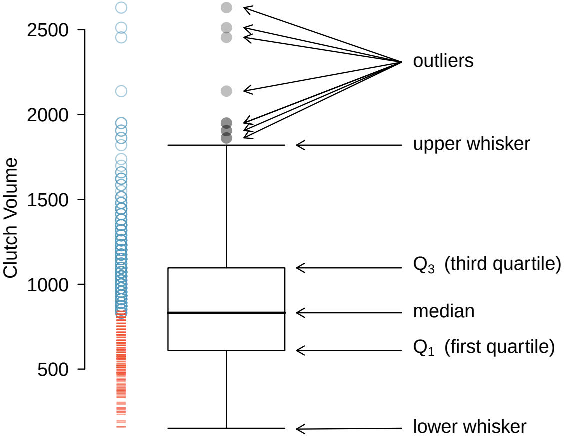
DESCRIBING A DISTRIBUTION



BOXPLOTS

A boxplot indicates the positions of the

- first, second, and third quartiles of a distributions
- potential outliers



OUTLIERS

Outliers are observations that are far from the center of a distribution

- Upper whisker (fence): $Q_3 + (1.5 \times IQR)$
- Lower whisker (fence): $Q_1 - (1.5 \times IQR)$

Any value below the lower fence or above the uppper fence are considered outliers

- Large outliers: values > upper fence
- Small outliers: values < lower fence

On a boxplot, potential outliers are shown with dots

GUIDED PRACTICE

Given the following observations, $[0, 10, 2, 2.5, 1.2, 0.5, 0.8]$.

Calculate lower/upper fences and determine if there are any outliers.

CATEGORICAL DATA

SUMMARIZING CATEGORICAL DATA

A table for a single variable, a frequency table or one-way table, summarize the distribution of observations among categories

```
table(births$marital)
```

```
not married     married
        386         613
```

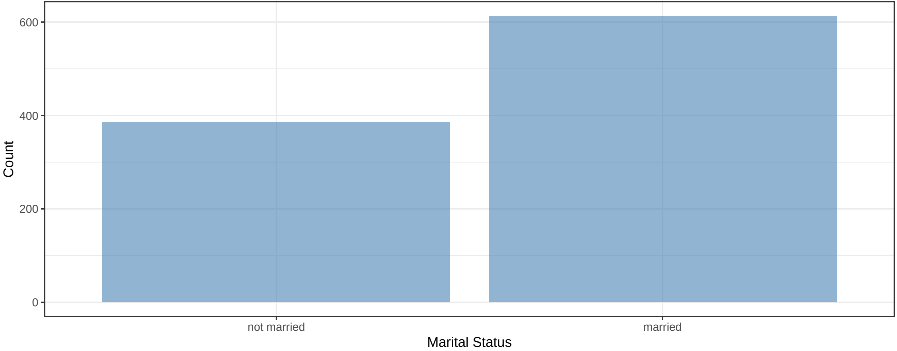
To express the frequency table in terms of proportions rather than counts, we can use prop.table(). Note, the input must be a table()

```
prop.table(table(births$marital))
```

```
not married     married
  0.3863864   0.6136136
```

VISUALIZING CATEGORICAL DATA

A barplot visually represents the frequenct or count of different categories withing a single variable



Each bar corresponds to a category level, with its height indicating the number of occurences