003 - Exploring Data - Part I

EPIB 607 - FALL 2020

Sahir Rai Bhatnagar
Department of Epidemiology, Biostatistics, and Occupational Health
McGill University

sahir.bhatnagar@mcgill.ca

slides compiled on March 3, 2021

What is statistics?

Data basics

Numerical data - single variable

Categorical data - single variable

Statistics in a Word

- Economics is about: *Money (and why it is good).*
- Psychology: *Why we think what we think (we think).*
- Biology: *Life.*
- Anthropology: *Who?*
- History: *What, where, and when?*
- Philosophy: *Why?*
- Engineering: *How?*
- Accounting: *How much?*

Statistics in a Word

- Economics is about: *Money (and why it is good).*
- Psychology: *Why we think what we think (we think).*
- Biology: *Life.*
- Anthropology: *Who?*
- History: *What, where, and when?*
- Philosophy: *Why?*
- Engineering: *How?*
- Accounting: *How much?*

Statistics is about: ***Variation***

Statistics is about quantifying uncertainty

- Data vary. People are different. We can’t see everything, let alone measure it all.
- Even what we do measure, we measure imperfectly.
- The data we wind up looking at and basing our decisions on provide, at best, an imperfect picture of the world.
- This fact lies at the heart of what Statistics is all about.
- How to make sense of it is a central challenge of Statistics.

Sampling from a population

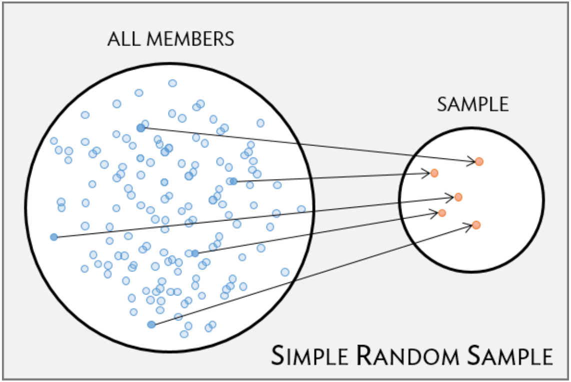


Figure: Random sampling is the best way to ensure that a sample reflects a population. In a simple random sample, each member of a population has the same chance of being sampled. For example, suppose we want to estimate the proportion of Montrealers who do not have a family doctor. We should randomly sample from different households in Montreal.

Selection bias

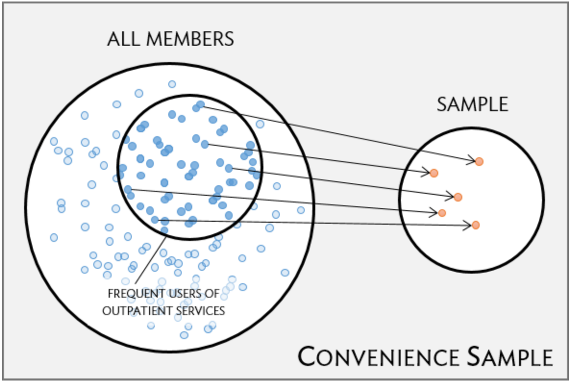


Figure: Instead of sampling from different households, we simply take all individuals from the same household.

The five “W”s and 1 “H”

Data values, no matter what kind, are useless without their context.

The five “W”s and 1 “H”

Data values, no matter what kind, are useless without their context.

- Who: Describe the individuals who were sampled (aka observations, subjects, biological samples, participants, experimental units, cases).
- What: Determine what is being measured. The characteristics recorded about each individual are called **variables**.

The five "W"s and 1 "H"

Data values, no matter what kind, are useless without their context.

- Who: Describe the individuals who were sampled (aka observations, subjects, biological samples, participants, experimental units, cases).
- What: Determine what is being measured. The characteristics recorded about each individual are called **variables**.
- Why: What was the purpose of the survey/experiment/study?
- When: When was the research conducted?
- Where: Where was the research conducted?

The five “W”s and 1 “H”

Data values, no matter what kind, are useless without their context.

- Who: Describe the individuals who were sampled (aka observations, subjects, biological samples, participants, experimental units, cases).
- What: Determine what is being measured. The characteristics recorded about each individual are called **variables**.
- Why: What was the purpose of the survey/experiment/study?
- When: When was the research conducted?
- Where: Where was the research conducted?
- How: Describe how the survey/experiment/study was conducted. Simple random sample (SRS), volunteers, select population, non-representative sample ?

Example: the FAMuSS study

- The Functional polymorphisms Associated with human Muscle Size and Strength study (FA- MuSS) measured a variety of demographic, phenotypic, and genetic characteristics for about 1,300 participants[1][2].

[1] Thompson PD, Moyna M, Seip, R, et al., 2004. Functional Polymorphisms Associated with Human Muscle Size and Strength. Medicine and Science in Sports and Exercise 36:1132 - 1139.

[2] see Harrington 1st edition, Section 1.2.2, for more details

Example: the FAMuSS study

- The Functional polymorphisms Associated with human Muscle Size and Strength study (FA- MuSS) measured a variety of demographic, phenotypic, and genetic characteristics for about 1,300 participants[1][2].
- One goal of the study—examine the association of demographic, physiological and genetic characteristics with muscle strength.
  - In simpler terms, study the "sports gene" ACTN3.

[1] Thompson PD, Moyna M, Seip, R, et al., 2004. Functional Polymorphisms Associated with Human Muscle Size and Strength. Medicine and Science in Sports and Exercise 36:1132 - 1139.

[2] see Harrington 1st edition, Section 1.2.2, for more details

Four rows from FAMuSS data matrix

```
# devtools::install_github("OI-Biostat/oi_biostat_data")
library(oibiostat)
data(famuss)
famuss %>%
  dplyr::glimpse()

## Rows: 595
## Columns: 9
## $ ndrm.ch     <dbl> 40.0, 25.0, 40.0, 125.0, 40.0, 75.0, 100.0, 57.1, 33.3,...
## $ drm.ch      <dbl> 40.0, 0.0, 0.0, 0.0, 20.0, 0.0, 0.0, -14.3, 0.0, 0.0, 2...
## $ sex         <fct> Female, Male, Female, Female, Female, Female, Female, F...
## $ age         <int> 27, 36, 24, 40, 32, 24, 30, 28, 27, 30, 20, 23, 24, 34,...
## $ race        <fct> Caucasian, Caucasian, Caucasian, Caucasian, Caucasian, ...
## $ height      <dbl> 65.0, 71.7, 65.0, 68.0, 61.0, 62.2, 65.0, 68.0, 68.2, 6...
## $ weight      <dbl> 199, 189, 134, 171, 118, 120, 134, 162, 189, 120, 131, ...
## $ actn3.r577x <fct> CC, CT, CT, CT, CC, CT, TT, CT, CC, CT, CT, CT, TT, CT,...
## $ bmi         <dbl> 33.112, 25.845, 22.296, 25.998, 22.293, 21.805, 22.296,...
```

FAMuSS Variables and their descriptions

Variable	Description
`ndrm.ch`	Percent change in strength in the non-dominant arm, comparing strength after to before training
`drm.ch`	Percent change in strength in a participant's dominant arm.
`sex`	Sex of the participant
`age`	Age in years
`race`	Recorded as African Am (African American), Caucasian, Asian, Hispanic, Other
`height`	Height in inches
`weight`	Weight in pounds
`actn3.r577x`	Genotype at the location r577x in the ACTN3 gene.
`bmi`	The participant's body mass index

Types of Variables

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

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

Types of Variables

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

- Discrete: only take on integer values (e.g., # of family members)
- Continuous: can take on any value within a specified range (e.g., 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 (e.g., education, likert scale)
- Nominal: no natural ordering of levels (e.g., gender)

Types of variables

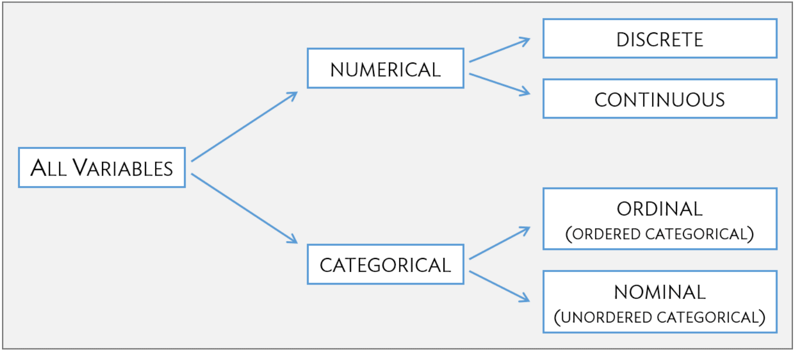


Figure: Types of variables

Exploring data with simple tools

- Techniques for exploring and summarizing data differ for numerical versus categorical variables.
- Numerical and graphical summaries are useful for examining variables one at a time, but also for exploring the relationships between variables.

Distributions and summary measures

- The collection of values for a numerical, continuous variable (e.g., `weight`) is the distribution for that variable.

Distributions and summary measures

- The collection of values for a numerical, continuous variable (e.g., `weight`) is the distribution for that variable.
- Numerical and graphical summaries convey characteristics of a distribution without listing all the values.
- Important characteristics include:
  - Center: where is the middle of the distribution?
    - Measures of center: mean, median

Distributions and summary measures

- The collection of values for a numerical, continuous variable (e.g., `weight`) is the distribution for that variable.
- Numerical and graphical summaries convey characteristics of a distribution without listing all the values.
- Important characteristics include:
  - Center: where is the middle of the distribution?
    - Measures of center: mean, median
  - Spread: how similar or varied are the values to each other?
    - Measures of spread: standard deviation, interquartile range

Measures of center: mean

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

$$\overline{y} = \frac{y_1 + y_2 + \cdots + y_n}{n}$$

where $y_1, y_2, \ldots, y_n$ represent the n observed values in a sample.
The mean weight in `famuss` is 155.65 pounds:

```
mean(famuss$weight)

## [1] 155.6479
```

Measures of center: median

The median is the value of the middle observation in a sample.
If the number of observations is

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

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

```r
median(famuss$weight)

## [1] 150

quantile(famuss$weight, probs = 0.50)

## 50%
## 150
```

Measures of spread

The standard deviation measures (approximately) the distance between a typical observation and the mean.

- An observation's deviation is the distance between its value y and the sample mean $\overline{y}$: $y - \overline{y}$.
- The sample variance s^2 is the sum of squared deviations divided by the number of observations minus 1.

$$s^2 = \frac{(y_1 - \overline{y})^2 + (y_2 - \overline{y})^2 + \cdots + (y_n - \overline{y})^2}{n-1},$$

  where $y_1, y_2, \ldots, y_n$ represent the n observed values.
- The standard deviation s is the square root of the variance.

$$s = \sqrt{\frac{(y_1 - \overline{y})^2 + (y_2 - \overline{y})^2 + \cdots + (y_n - \overline{y})^2}{n-1}}$$

```
sd(famuss$weight)

## [1] 34.58999
```

Measures of Spread: Percentiles/Quartiles

The p^{th} percentile is the observation such that $p\%$ of the remaining observations fall below this observation.

- The first quartile (Q_1) is the 25^{th} percentile.
- The second quartile (Q_2), i.e., the median, is the 50^{th} percentile.
- The third quartile (Q_3) is the 75^{th} percentile.

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

$$IQR = Q_3 - Q_1$$

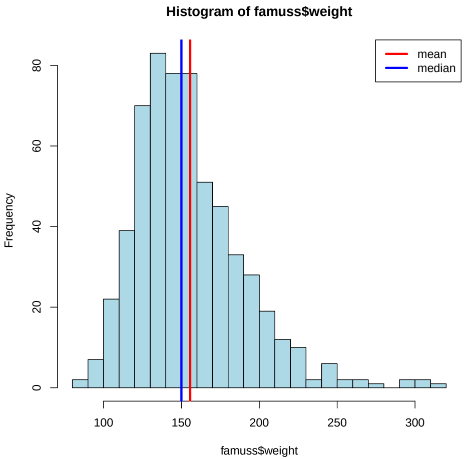
```r
IQR(famuss$weight)

## [1] 42

diff(quantile(famuss$weight, probs = c(0.25, 0.75)))

## 75%
##  42
```

Robust estimates

- The median and IQR are called robust estimates because they are less likely to be affected by extreme values than the mean and standard deviation.
- For distributions containing extreme observations, the median and IQR provide a more accurate sense of center and spread.

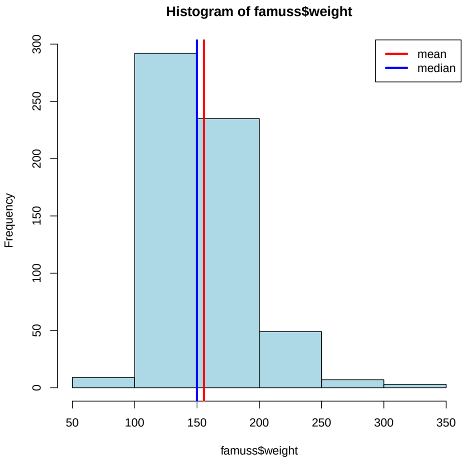
Histograms

```r
hist(famuss$weight, breaks = 30, col = 'lightblue')
abline(v = mean(famuss$weight), lwd = 3, col = "red")
abline(v = median(famuss$weight), lwd = 3, col = "blue")
legend("topright", legend = c("mean","median"), col = c("red","blue"), lwd = 3)
```



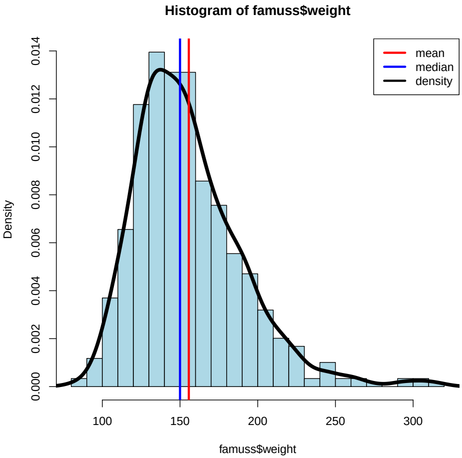
Histograms

```
hist(famuss$weight, breaks = 5, col = 'lightblue')
abline(v = mean(famuss$weight), lwd = 3, col = "red")
abline(v = median(famuss$weight), lwd = 3, col = "blue")
legend("topright", legend = c("mean","median"), col = c("red","blue"), lwd = 3)
```



Density plots

```
hist(famuss$weight, breaks = 30, col = 'lightblue', probability = TRUE)
openintro::densityPlot(famuss$weight, add = TRUE, lwd = 5)
abline(v = mean(famuss$weight), lwd = 3, col = "red")
abline(v = median(famuss$weight), lwd = 3, col = "blue")
legend("topright", legend = c("mean","median","density"),
       col = c("red","blue","black"), lwd = 3)
```



Histograms ...

- Histograms show important features of the shape of a distribution:
  - Symmetry, or lack of it (skew)
  - Minimum and maximum values
  - Regions of high frequency (modes)

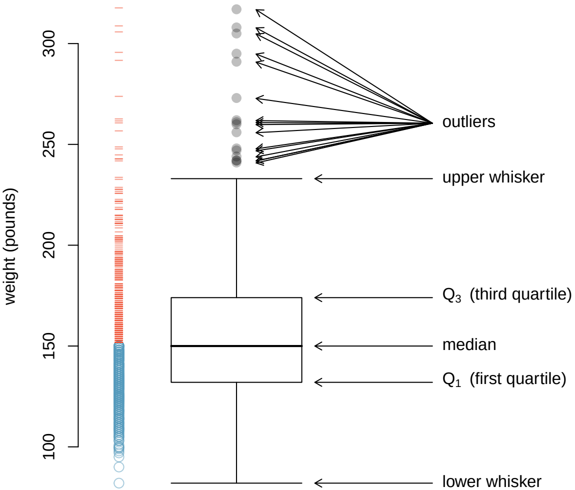
Histograms ...

- Histograms show important features of the shape of a distribution:
  - Symmetry, or lack of it (skew)
  - Minimum and maximum values
  - Regions of high frequency (modes)
- Histograms not so good for:
  - Displaying median, quartiles
  - Showing subtle skewing
  - Identifying extreme values

Histograms ...

- Histograms show important features of the shape of a distribution:
    - Symmetry, or lack of it (skew)
    - Minimum and maximum values
    - Regions of high frequency (modes)
- Histograms not so good for:
    - Displaying median, quartiles
    - Showing subtle skewing
    - Identifying extreme values
- Remember that histograms are sensitive to the number of bins!

Boxplot for weight



Boxplots

A boxplot indicates the positions of the first, second, and third quartiles of a distribution in addition to potential **outliers**, observations that are far from the center of a distribution.

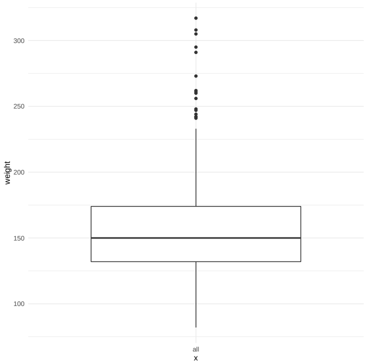
- Large outliers: values > $Q_3 + (1.5 \times IQR)$
- Small outliers: values < $Q_1 - (1.5 \times IQR)$

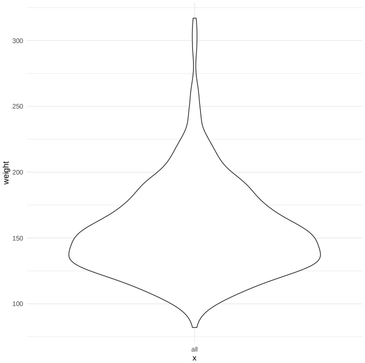
On a boxplot...

- The rectangle extends from the first quartile to the third quartile, with a line at the second quartile (median).
- Whiskers capture data between $Q_1 - (1.5 \times IQR)$ and $Q_3 + (1.5 \times IQR)$; whiskers must end at data points.
- Potential outliers shown with dots.

Boxplots vs. Violin plots

```
p1 <- ggplot(data = famuss, mapping = aes(x = "all", y = weight)) + theme_minimal()
p1 + geom_boxplot()
p1 + geom_violin()
```





What is statistics?

Data basics

Numerical data - single variable

Categorical data - single variable

Tables

A table for a single variable, a frequency table or one-way table, summarizes the distribution of observations among categories.
Based on the table, describe the distribution of genotype at the location actn3.r577x among the study participants.

```
table(famuss$actn3.r577x)

##
##  CC  CT  TT
## 173 261 161
```

Bar plots for categorical data

A bar plot is a common way to display a single categorical variable.

```
graphics::barplot(table(famuss$actn3.r577x))
sjPlot::plot_frq(famuss$actn3.r577x)
```