Measures of spread
Overview

Quick review of distributions
Outliers
The mean and median
The standard deviation
Z-scores
Percentiles
R Markdown and the next worksheet
Categorical variables: Sample vs. Population proportion

\[ \pi_{\text{red}} \]

\[ \hat{p}_{\text{red}} \]
Quantitative variables: Sample vs. Population means
Plato and shadows: distributions and histograms

(a) Skewed to the right

(b) Skewed to the left

(c) Symmetric and bell-shaped

(d) Symmetric but not bell-shaped

Income distribution
Neat facts - average NFL player is:

• 1. About 25 years old (age)
• 2. Just over 6'2" in height (height)
• 3. Weighs a little more than 244lbs (weight)
• 4. Makes slightly less than $1.5M in salary per year (salary)

Question: Can you tell which histogram goes with which trait?
Task is to add the labels: age, height, weight, and salary
  • Hint: There are a wide range of positions in football that have very different roles
    • E.g., placekickers only play for small factions of the game, while quarterbacks are essentially to a team’s success

First: what is the label for the y-axis?
  • A: Frequency or count
If you don’t want an ex, label you axes!
Back to the Gapminder data...

# get a data frame with information about the countries in the world
> source("/home/shared/intro_stats/cs206_functions.R")
> country_data <- get_cs206_data("gapminder_2007")

Can you plot a histogram of the population of each country with 20 bins?
> pop_vec <- country_data$pop  # first create a vector with the population of each country
> hist(pop_vec, n = 20)       # then create the histogram
What is missing from this histogram?

Axes labels could be more informative!
Labeling axes

Can you figure out how to label the axes?

• A: xlab and ylab!

\[
> \text{hist(pop_vec, n = 20,}
\]
\[
\hspace{1cm} \text{ylab = "Frequency",}
\]
\[
\hspace{1cm} \text{xlab = "Population",}
\]
\[
\hspace{1cm} \text{main = "World countries population in 2007"})
\]
Outliers

An **outlier** is an observed value that is notably distinct from the other values in a dataset by being much smaller or larger than the rest of the data.

Outliers can potentially have a large influence on the statistics you calculate

- One should examine outliers in more detail to understand what is causing them
Out, liar!

Your theory is wrong!
Descriptive statistics for the center of a distribution

Graphs are useful for visualizing data to get a sense of what the data look like.

We can also summarize data numerically.

Question: what is a numerical summary of a sample of data called?

A: a statistic!

Two important statistics that can be used to describe the center of the data are the mean and the median.
Sample and population mean

\( \mu \)

(d) Symmetric but not bell-shaped

Population

Data Collection

Statistical Inference

Sample

\( \bar{x} \)
The mean

\[
\text{Mean} = \frac{\text{Sum of all data values}}{\text{Number of data values}}
\]

\[
\text{Mean} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}
\]

R: `mean(x)`
R: `mean(x, na.rm = TRUE)`
Give the proper notation: $\mu$ vs. $\bar{x}$?

We measure the height of 50 randomly chosen Hampshire students.

We measure the height of all Hampshire students.

Can you calculate the mean of the countries life expectancy in R?

```r
> life_expectancy <- country_data$lifeExp
> mean(life_expectancy)
```
Histograms: a way to plot quantitative data

- Skewed to the right
- Skewed to the left
- Symmetric and bell-shaped
- Symmetric but not bell-shaped
The median

The **median** is a value that splits the data in half
  • i.e., half the values in the data are smaller than the median and half are larger

To calculate the median for a data sample of size $n$, sort the data and then:

  • If $n$ is odd: The middle value of the sorted data
  • If $n$ is even: The average of the middle two values of the sorted data
Example of calculating the mean and median

When a perspective Hampshire student visit a Hampshire webpage a ‘ping’ is generated

Below is a random sample of ping counts from 7 perspective students who pinged the site at least once:
12, 45, 6, 4, 158, 10, 59

Q: What is the mean and median ping count in this sample?

A: mean = 42
median = 12

Mean = \frac{\sum_{i=1}^{n} x_i}{n}
The median

R:
```r
median(v)
median(v, na.rm = TRUE)
```
Resistance

We say that a statistics is **resistant** if it is relatively unaffected by extreme values (outliers).

The median is resistant when the mean is not

Example:

- Mean US salary = $72,641
- Median US salary = $51,939
Measure of central tendency: mean and median

World countries 2007 GDP per capita

Is the red line the mean or the median?
Characterizing the spread

The mean and median are numbers that tell us about the center of a distribution.

We can also use numbers to characterize how data is spread.
Average monthly temperature: Des Moines vs. San Francisco

Data measured on April 14th from 1997 to 2010:

Mean temperature (°F): Des Moines = 54.49  San Fran = 54.01
The standard deviation

The **standard deviation** (for a quantitative variable) is a measure of the spread of the data

It gives a rough estimate for a typical distance a point is from the center
Notation

The standard deviation of a *sample* is denoted $s$
  • It measure the spread of the data from the sample mean

The standard deviation of the *population* is denoted $\sigma$
  • It measure the spread of the data from the population mean
Population and sample standard deviation

(c) Symmetric and bell-shaped

Population

Data Collection

Statistical Inference

Sample

$\sigma$

$S$
Which has the larger standard deviation?

\[ s_{DM} = 11.73 \, ^\circ\text{F} \quad s_{SF} = 3.38 \, ^\circ\text{F} \]
The standard deviation

The standard deviation can be computed using the following formula:

$$s = \sqrt{\frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$
Example: computing the standard deviation

Suppose we had a sample with \( n = 4 \) points:

\[
x_1 = 8, \quad x_2 = 2, \quad x_3 = 6, \quad x_4 = 4,
\]

We can compute the mean using the formula:

\[
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{4} \cdot (x_1 + x_2 + x_3 + x_4) = \frac{1}{4} \cdot (8 + 2 + 6 + 4)
\]

The standard deviation can be computed using the formula:

\[
s = \sqrt{\frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \bar{x})^2} \quad \text{(remember order of operations!)}
\]
Hot dogs!

Every 4\textsuperscript{th} of July, Nathan’s Famous in NYC holds a hot dog eating contest where contestants try to eat as many hot dogs as they can in 10 minutes.

Part 1: Calculate the mean and standard deviation for the number of hot dogs eaten!

\[ s = \sqrt{\frac{1}{(n - 1)} \sum_{i=1}^{n} (x_i - \bar{x})^2} \]
Worksheet 2: examining cars sold

Lock5 questions
Worksheet 2!

A. Accessing R with the Hampshire server:
   https://asterius.hampshire.edu

B. Go to the console and download a file using the following command:

   > source('/home/shared/intro_stats_2018/cs206_functions.R')
   > get_worksheet(2)