Module B: Knowing Your Data

Slide Deck B3:

Measures of Point Values

The section in which we learn how to calculate positions in the data. Here, we learn about the three primary measures of center (mean, median, and mode) as well as the quantiles (comparing within a data set), the z-scores (comparing across data sets), and correlations (relationships between two variables).

Today's Objectives

By the end of this slidedeck, you should

- understand the importance of the formula for the mean:
  - know why we prefer to use the mean when we can
  - know when we should *not* use the mean to summarize the data
- know how to calculate the mean, median, and mode;
- know how to calculate the percentiles (quantiles):
  - quartiles, quintiles, deciles, etc.
- calculate the z-score (a.k.a. the z-transformation);
- calculate the correlation between two variables; and
- perform the calculations in R.

Measures of Center

Definition

A **measure of center** of data is a number that represents the "typical" value in the data.

Examples

- mean
- median
- mode
- weighted mean
- mid-range

Sample Mean

The **sample mean** is the center of gravity of the data. It is the most commonly used measure of center, whether it should be used or not. The Central Limit Theorem (covered at the start of Module C) does explain why the mean gives us important information, whether it is the optimal measure of center or not.

Its formula is $$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Written out, the formula is $$\bar{x} = \frac{1}{n} \left(x_1 + x_2 + x_3 + \cdots + x_n \right)$$

In words, this just reads as

Add all the data values ($\sum_{i=1}^{n} x_i$), then divide by the number of data values ($1/n$).

Sample Mean

Note the following things about the sample mean:

- Every data value is included in the calculation.
- Every data value counts the same in the calculation.
- The order does not matter in the calculation.

One thing to think about is how each data value affects the measure of center. If a single data value can change the measure to an infinitely large value, then the measure is not **robust** to outliers.

The mean is *not* robust. A single outlier can make the mean incredibly large (or small). The next example illustrates this.

Example I

Let us calculate the mean download speed of my Internet connection. To do this, I measured it seven times. Here are the speeds in Mb/s:

$$92.22;\ 94.14;\ 94.25;\ 94.10;\ 94.45;\ 94.48;\ 93.24$$

By hand:

$$\begin{aligned}\bar{x} &= \Big(92.22 + 94.14 + 94.25 + 94.10 + 94.45 + 94.48 + 93.24 \Big) \div 7 \\ &= 93.84\end{aligned}$$

By R:

```
speed = c(92.22, 94.14, 94.25, 94.10, 94.45, 94.48, 93.24)
mean(speed)
```

Example II

Let us calculate the mean download speed of my Internet connection. After measuring it seven times (as previously), I measured it one final time. This time, the reported speed was 100 milliard Mb/s (1×10^{11}).

Thus, all eight speed measurements are

$$92.22;\ 94.14;\ 94.25;\ 94.10;\ 94.45;\ 94.48;\ 93.24;\ 100,000,000,000$$

Now, the mean is "1.25e+10," which is $1.25 \times 10^{10} = 12,500,000,000$:

```
speed2 = c(92.22, 94.14, 94.25, 94.10, 94.45, 94.48, 93.24, 1e11)
mean(speed2)
```

And so, with the inclusion of that single outlier, the mean increases by a factor of 10^9. Because of this, we see that the mean is definitely *not* robust to outliers.

- Is this a good thing or a bad thing?

Sample Median

The **sample median** is the "middle number" of a set of data, ranked from smallest to largest. The key to the median is that *about* half of the data values are above it, and *about* half are below it.

- When would this information be useful?

Sample Median

While there is a 'formula' to calculate the median, we will rely on the following algorithm (for now):

1. Rank the values from lowest to highest.
2. If the sample size is odd, the median is the middle ranked value.
3. If the sample size is even, the median is the mean of the two middle ranked values.

Note: We will see the mathematical formula in Learning Module 3 where we look at populations.

Sample Median

Note the following things about the sample median:

1. Every data *rank* (not value) is included in the calculation.
2. Every data *rank* (not value) counts the same in the calculation.
3. The order *does* matter in the calculation (eventually).

What do these tell us about how the median compares to the mean?

Note: In *contrast* to the mean, **the median *is* robust to outliers**. The next examples illustrate this.

Example I

Let us calculate the median download speed of my Internet connection. To do this, I measured it seven times. Here are the speeds in Mb/s:

$$92.22;\ 94.14;\ 94.25;\ 94.10;\ 94.45;\ 94.48;\ 93.24$$

By hand:

$$\text{ordered}: \left\{\ 92.22;\ 93.24;\ 94.10;\ 94.14;\ 94.25;\ 94.45;\ 94.48\ \right\}$$

Since $n = 7$ is odd, the value at position $(n+1)/2 = 4$ is the median

$$\tilde{x} = 94.14$$

By R:

```
speed = c(92.22, 94.14, 94.25, 94.10, 94.45, 94.48, 93.24)
median(speed)
```


Example II

Let us calculate the median download speed of my Internet connection. After measuring it seven times (as previously), I measured it one last time. This time, the reported speed was 100 milliard Mb/s. Thus, all eight speed measurements are

$$92.22;\ 94.14;\ 94.25;\ 94.10;\ 94.45;\ 94.48;\ 93.24;\ 100,000,000,000$$

Now, the median is 94.195 Mb/s:

```
speed = c(92.22, 94.14, 94.25, 94.10, 94.45, 94.48, 93.24, 1e11)
median(speed)
```

And so, with the inclusion of that single outlier, the median changes very little, from 94.14 to 94.195.

- Is this a good thing or a bad thing?

Sample Mode

The **sample mode** is the most prevalent value in a set of data.

- When would *this* information be useful?

To calculate the mode by hand:

1. Determine which value occurs most often.

With this definition, do you think that the mode is robust to outliers? Please explain.

Sample Mode

Here is some terminology regarding the mode:

- If one value occurs most frequently, then the data are **unimodal**.
- If two values occurs most frequently, then the data are **bimodal**.
- If more than two values occurs most frequently, then the data are **multimodal**.
- If all values occur equally often, then the data have **no mode**.

Sample Mode

Note the following things about the sample mode:

- Every data *frequency* is included in the "calculation."
- Every data *value* counts the same in determining the frequency.
- The order does not matter in the calculation.
- Comparing data values (other than exclusion) is not needed.

How do these tell us about how the mode compares to the mean and median?

Example

Let us calculate the modal download speed of my Internet connection. To do this, I measured it seven times. Here are the speeds in Mb/s:

92.22; 94.14; 94.25; 94.10; 94.45; 94.48; 93.24

By hand: There is **no mode**, because all values happen once each.

By R:

```
source("http://rfs.kvasaheim.com/stat200.R")
speed = c(92.22, 94.14, 94.25, 94.10, 94.45, 94.48, 93.24)
modal(speed)
```

All of the Code

So, in summary, the above calculations can be done with this code:

```
source("http://rfs.kvasaheim.com/stat200.R")
speed  = c(92.22, 94.14, 94.25, 94.10, 94.45, 94.48, 93.24)
speed2 = c(92.22, 94.14, 94.25, 94.10, 94.45, 94.48, 93.24, 1e11)

mean(speed)
median(speed)
modal(speed)

mean(speed2)
median(speed2)
modal(speed2)
```

Note that you need to be able to determine what each line of code does. Doing so will help you be able to think through someone else's analysis and to better craft your own.

Measures of Position

If we care about describing the center of a variable, we would use a "measure of center." However, if we care about some other position in the variable, we would use a **measure of position**. Note that the measures of center are a *specific type* of measure of position. Also note that these are related to the median instead of the mean.

There are a couple of important ones to know for general research (!), a couple to know because of specific area research (†), and one important one to know to understand some mathematics later in the course (‡).

- ! Percentiles
- ! Quartiles
- † Quintiles
- † Deciles
- ‡ Z-score (Z-transformation)

Quantiles (a.k.a. Percentiles)

Where the median divides the data into two parts, **percentiles** divide the data into 100 parts. Note that the median is always a single value, whereas the percentiles can be written to indicate either a single value or a range (e.g., "a score is *in* the tenth percentile"). Context is key.

Quantiles are useful when we want to discuss things that are *not* the middle. For instance:

- In inequality research, we tend to focus on the 10th percentile to understand poverty.
- In basic statistics, we tend to focus on the 2.5th and 97.5th percentiles to better understand uncertainty in our estimates.
- Education research tends to focus on the 5th, 10th, or 90th percentile.

It all depends on your field of study. Ask around in your home department to see which percentile(s) you should focus on.

Quartiles

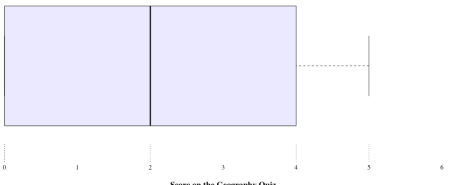
When calculating the median, we saw that it divides the ordered data into two equal parts. There may be times when we wish the divide the ordered data into *four* parts... into "quarters." The values that divide the *ordered* data into four parts are called the **quartiles**.

- The **first quartile**, Q_1:
  A value that separates the bottom 25% of the data from the top 75%.
- The **second quartile**, Q_2:
  A value that separates the bottom 50% of the data from the top 50%.
- The **third quartile**, Q_3:
  A value that separates the bottom 75% of the data from the top 25%.

Note that the **second quartile** is just the median... the value that separates the bottom 50% from the rest.

Quartiles

The quartiles are used in creating a typical box-and-whiskers plot (a.k.a. **boxplot**).



Other -iles

All of the -iles can be calculated in terms of the percentiles.

- The median is the 50th percentile
- The values of Q_1 and Q_3 are the 25th and 75th percentiles
- The quintiles are the 0th, 20th, 40th, 60th, 80th, and 100th percentiles
- The deciles are the 0th, 10th, 20th, . . . , and 100th percentiles

The R function to calculate the percentiles is `quantile` function. The following calculates the 5th and 95th percentiles of the upload speeds data loaded earlier:

```
quantile( speed, c(0.05,0.95) )
```

Note that the first slot is the variable name and the second is the percentile(s) that needs to be calculated.

Examples

For the original upload speed data, calculate the following items:

1. mean
2. first quartile, Q_1
3. the 90th percentile
4. all quintiles
5. the difference between Q_1 and Q_3

Answers:

```
source("http://rfs.kvasaheim.com/stat200.R")
speed = c(92.22, 94.14, 94.25, 94.10, 94.45, 94.48, 93.24)

mean(speed)
quantile(speed, 0.25)
quantile(speed, 0.90)
quantile(speed, c(0.0, 0.2, 0.4, 0.6, 0.8, 1.0))
quantile(speed, 0.75) - quantile(speed,0.25)
```


Examples

For the new upload speed data, calculate the following items:

1. mean
2. first quartile, Q_1
3. the 90th percentile
4. all quintiles
5. the difference between Q_1 and Q_3

Answers:

```
source("http://rfs.kvasaheim.com/stat200.R")
speed2 = c(92.22, 94.14, 94.25, 94.10, 94.45, 94.48, 93.24, 1e11)

mean(speed)
quantile(speed2, 0.25)
quantile(speed2, 0.90)
quantile(speed2, c(0.0, 0.2, 0.4, 0.6, 0.8, 1.0))
quantile(speed2, 0.75) - quantile(speed2,0.25)
```

Z-Transform

Finally, let us look at a transformation that allows us to compare entities *across data sets* — the **z-transform**.

Definition

The **z-transform** is a function (transformation) applied to the data such that the transformed values have mean 0 and standard deviation 1.

Its formula is rather straight-forward:

$$z = \frac{x - \bar{x}}{s}$$

- *Why* would we want to do something like this?

Example

In a course, I gave two examinations. On the first, you scored 86; on the second, 88.

1. Did you do better on the second exam?
2. *Relative to others*, did you do better on the second exam?

Answers:

1. Yes, by 2 points.
2. I do not know. There is not enough information.

Example, Part II

In a course, I gave two examinations. On the first, you scored 86; on the second, 88. The class averaged 74 ± 4 on the first and 78 ± 10 on the second.

1. Did you do better on the second exam?
2. *Relative to others*, did you do better on the second exam?

Answers:

1. Still yes, and still by 2 points.
2. To answer this question, we need to calculate the z-scores for your two tests:

$$z_1 = \frac{x - \bar{x}}{s} = \frac{86 - 74}{4} = +3$$

$$z_2 = \frac{x - \bar{x}}{s} = \frac{88 - 78}{10} = +1$$

Example, Part II

Your z-score on the first test was +3. Your z-score on the second test was +1. So, taking into consideration how well the other class members did, you did worse *compared to them* on the second test.

The interpretation of the z-score is straight-forward. On the first test, you scored three standard deviations above average. On the second, you only scored *one* standard deviation above average.

Note that the z-score can help you compare yourself to the class across tests. It controls for the difficulty of the examination. Since the interpretation deals with how far above (or below) the mean is *in terms of standard deviations*, the z-score can be used regardless of the units used for the original data.

In the future (Learning Module 4), we will use the z-transformation to scale variables so that we can draw uniform conclusions about the mean.

Another Example

Goal: Determine how well Oregon did, in terms of the unemployment rate between 1990 and 2000, relative the the other states in the United States.

The usual analysis steps:

1. Decide the analysis steps
2. Load the data
3. Calculate the statistics (perform the analysis)
4. Interpret the results

Another Example

This is the R code I used for this analysis (along with a secondary analysis):

```
### Preamble
# More functionality
source("http://rfs.kvasaheim.com/stat200.R")

# The data
dt = read.csv("http://rfs.kvasaheim.com/data/crime.csv")
attach(dt)

### Analysis
# Absolute unemployment rates
unemp1990[state=="Oregon"]
unemp2000[state=="Oregon"]

# Relative unemployment rates
zscore(unemp1990)[state=="Oregon"]
zscore(unemp2000)[state=="Oregon"]
```

Another Example

The output from the previous code tells us:

In 1990 and 2000, Oregon's unemployment rates were 5.6% and 4.9%, respectively. The corresponding z-scores are 0.1138 and 0.9837.

Interpretation:

> *While the unemployment rate in Oregon dropped from* 5.6% *to* 4.9% *between 1990 and 2000, relative to the other states in the United States, it actually increased by* 0.8699 *standard deviations. Thus, Oregon did worse than average in the 1990s in terms of the unemployment rate (the rate increased, relative to the typical state).*

Measures of Correlation

We now move beyond summary statistics about a single variable. Let us look at quantifying the *relationship* between variables.

Definition

All correlation measures indicate the strength of the relationship between two variables. The values range between -1 (perfect negative correlation) and +1 (perfect positive correlation).

There are three types of correlation I introduce here:

- r, the Pearson correlation coefficient
- ρ, the Spearman rank correlation coefficient
- τ, the Kendall rank correlation coefficient

Pearson's Correlation Coefficient

The Pearson correlation coefficient, r, is the "usual" measure of correlation between two numeric variables. It was developed by Karl Pearson in 1895 from a related idea introduced by Francis Galton in the 1880s, and for which the mathematical formula was derived and published by Auguste Bravais in 1844.

The Pearson correlation coefficient:

- is *not* robust to outliers, so its value can be misleading if outliers are present
- only quantifies the *linear* relationship between two variables
- is appropriate for variables that are at least interval level

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

The R function is

```
cor(x,y, method="pearson")
```


Spearman's Rank Correlation Coefficient

Spearman's rank correlation coefficient (Spearman's ρ) is a measure of the correlation between the *ranks* of two variables. Essentially, it measures how well the relationship between two variables can be described using a monotonic function. Spearman devised it in 1904 based on Pearson's earlier writings (1895).

The Spearman coefficient:

- *is* robust to outliers
- is appropriate for variables that are at least ordinal level

Spearman's ρ is the Pearson correlation coefficient on the ranks (as opposed to the values).

The R function is

```
cor(x,y, method="spearman")
```

Kendall's Rank Correlation Coefficient

The Kendall rank correlation coefficient, commonly referred to as Kendall's τ coefficient, measures the *ordinal* association between two measured quantities. Like Spearman's ρ, it is a measure of rank correlation. Maurice Kendall developed it in 1938, though Gustav Fechner had proposed a similar measure in the context of time series in 1897.

Kendall's τ coefficient:

- *is* robust to outliers
- is appropriate for variables that are at least ordinal level

$$\tau = \frac{2}{n(n-1)} \sum_{i<j} \text{sign}(x_i - x_j)\text{sign}(y_i - y_j)$$

The `R` function is

```
cor(x,y, method="kendall")
```


Correlation Example

Let us determine the correlation between the IQ of a person and how much television they watch per week. This is the `R` code I used for this analysis:

```
# The data
iq = c(106,100,86,101,99,103,97,113,112,110)
tv = c(7,27,2,50,8,29,20,12,6,17)

# The analysis
cor(iq,tv, method="pearson")
cor(iq,tv, method="spearman")
cor(iq,tv, method="kendall")
```

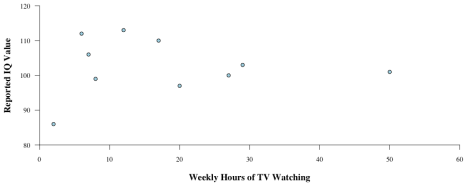
The output:

```
 0.000378362
-0.030303030
-0.022222222
```

So, which of the three correlations is "most" appropriate for these two variables?

Correlation Example

Here is a graphic of the data. It may help us determine which of the three measures would be "best."



Today's Objectives

Now that we have concluded this lecture, you should be able to

- understand the importance of the formula for the mean:
  - know why we prefer to use the mean when we can
  - know when we should *not* use the mean to summarize the data
- know how to calculate the mean, median, and mode;
- know how to calculate the percentiles (quantiles):
  - quartiles, quintiles, deciles, etc.
- calculate the z-score (a.k.a. the z-transformation);
- calculate the correlation between two variables; and
- perform the calculations in R.

Today's R Functions

In this slide deck, we saw the following R functions:

- `mean`
- `median`
- `modal`

- `quantile`
- `boxplot`
- `zscore`
- `cor`

- `source`
- `read.csv`
- [*some condition*]

Supplemental Activities

The following activities may be of interest to you in terms of today's topics:

- SCA 2a is for measures of **center**
- SCA 2b is for measures of **position**
- SCA 2c is for measures of spread (for next time)

Note that you can access all Statistical Computing Activities here:
`https://www.kvasaheim.com/courses/stat200/sca/`

For future interest: Laboratory Activity D explores *why* we prefer the mean to other measures of center. The reason has to do with precision (and not accuracy).

Supplemental Readings

The following are some readings that may be of interest to you in terms of measures of position (and calculating them in R):

- Hawkes Learning: Sections 3.1 and 3.3
- Intro to Modern Statistics: Chapter 5
- R for Starters: Section 4.2