**Name: Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Classwork 3**

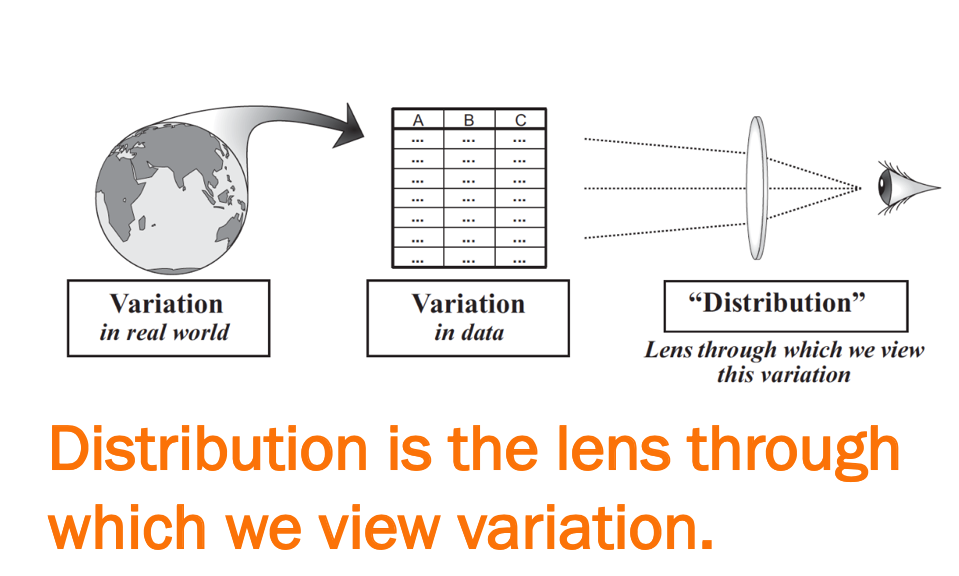
**Intro to Exploring Variation**

**Textbook Chapters: 1-2**

**Advanced Learning Objectives:**

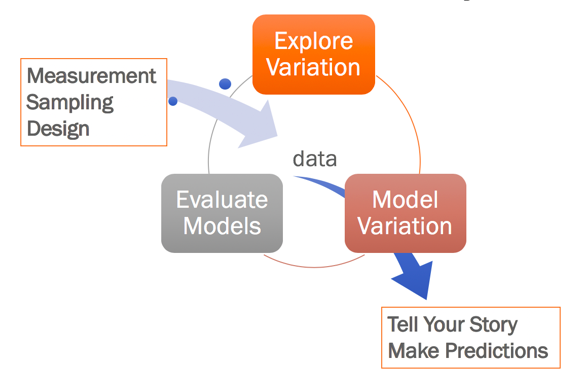
* Identify which parts of the Martin v. Westvaco simulation correspond to the Distribution Triad
* Discuss ways to measure and explore variation
* Define and practice explaining variation
* Classify different types of random variation

**Recap: The Core Ideas**

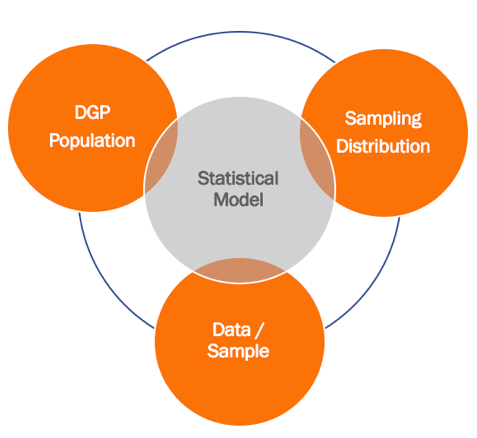


The goal of statistics is to explain and model \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ in the real world .

What is the opposite of variation?



If we can develop a good model of variation, we can make better \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ .



Statistical models can also help us make inferences about the \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ .

**Martin v. Westvaco & The Distribution Triad**

Identify which parts of the Martin v. Westvaco lab simulation correspond to the Distribution Triad (i.e., the Population, Sample, and Sampling distributions).

**Part I: Exploring Variation**

**Part II: Modeling Variation**

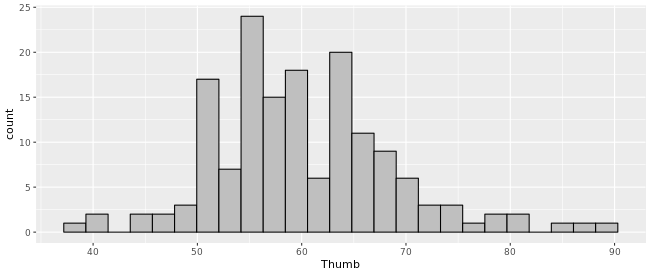
**Part III: Evaluating Models**

Our journey starts with **Part I: Exploring Variation.**

Once we have randomly and independently sampled and taken measurements of something, we can start exploring the variation (i.e., describing the data, and telling the story of the data) with **summary statistics** (e.g., the *mean*, the *median*, the *standard deviation*, etc.) and **visualizations** (e.g., *bar graphs, histograms, scatterplots*, etc.). This will help us identify patterns, and possible explanatory variables.

**favstats**(Fingers$Thumb)

**gf\_histogram**(~Thumb, data=Fingers, color = "black", fill = "grey")



|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **min** | **Q1** | **median** | **Q3** | **max** | **mean** | **sd** | **n** | **missing** |
| 39 | 55 | 60 | 65 | 90 | 60.10 | 8.73 | 157 | 0 |

How many variables are depicted in the histogram?

What is depicted along the x-axis?

What is depicted along the y-axis?

Thumb length in millimeters is a (circle one): **Quantitative / Categorical** variable.

On the histogram, find the mean (along the x-axis) and draw a vertical line going through the data. Then, circle the data in the histogram that represent the minimum and maximum values.

Based on this *sample*, what is the story about the *population* that is starting to emerge from these data?

**Measuring Fingers**

Psychology is hard because we are constantly trying to measure invisible stuff like self-efficacy, intelligence, memory, perception, ethnic identity… etc (How would YOU measure these things?). Then we want to figure out whether this stuff changes, improves, differs across people, etc. Later, we will learn about using the measurements we gather (the distribution of data) to create a model that will help us make predictions and determine likelihoods for particular outcomes. But it all starts with figuring out measurement -- how would we know how much of this thing someone has.

So let’s start by measuring something easy. Fingers.

Our goal today is to accurately measure each of your fingers to the nearest mm. Measure your non-dominant hand’s fingers (it will just make things easier for you) from the tip of your finger to that little fold/joint where your finger meets your palm.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Thumb | Pointer finger | Middle finger | Fourth finger | Pinkie finger |
|  |  |  |  |  |

1. What are some important characteristics about a person that affects how long their fingers are? What about some characteristics that probably do NOT affect how long their fingers are?

Get together with a partner. Don’t show them your finger measurements. **Silently** give them your hand and have them measure your fingers to the nearest mm. Write down their measurements of **your** fingers to the nearest mm. Try to keep a poker face. Take turns.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Thumb (1D) | Pointer finger (2D) | Middle finger (3D) | Fourth finger (4D) | Pinkie finger (5D) |
|  |  |  |  |  |

1. Did your partner come up with identical measurements for your fingers? Why do you think that is?
2. What are some things you could do if you wanted to make these measurements more similar? That is, how can you reduce the variability in your measurements? And would it make a difference if we used different units of measurement (e.g., centimeters, inches, nanometers, feet, etc.)?
3. Can we completely reduce the variability in measurement? Why or why not?

1. What do you think it means to “explain” variation?

1. Does explaining variation mean that the variation just goes away?

**Lab: Exploring Finger Measurements**

What are some reasons that researchers or businesses might be interested in finger measurements? What is some other information they may also want to know to go along with your finger measurements?

1. Go to Menti.com and enter your finger measurements.
2. Go to Canvas and find the link to the data file for the finger measurements in the Labwork 3 assignment.
3. We need to make sure our data is “tidy.”
   1. What are the cases, or observations? Do they go in columns or rows?
   2. What are the variables? Do they go in columns or rows? Which are categorical/quantitative?
4. Once our data is “tidy” you can copy the “publish to web as CSV” link and open the data frame in the R Sandbox using: **DataFrameName** <- **read.csv**("https://url.com", header = TRUE)

Next, we should consider any data manipulations we may need to do.

1. Are there any missing or unusual values? How can we find out?
2. Should we filter any variables? How can we do that?
3. How do we tell R which variables are categorical?
4. Should we recode any variables?
5. Should we transform any variables from a quantitative level of measurement to a categorical

level of measurement? How do we do that?

1. Should we aggregate any variables? How can we do that?
2. What summary statistics might we want to generate? What is the name of the function that can give us

several common descriptive statistics?

1. We’re going to start by exploring the general distribution of pinkie lengths from the class. Check your R Cheat Sheet for how to create a histogram: **gf\_histogram**(~ , data = )
   1. Roughly sketch the distribution below and make some observations. Also, get the

**favstats**( ).

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **min** | **Q1** | **median** | **Q3** | **max** | **mean** | **sd** | **n** | **missing** |
|  |  |  |  |  |  |  |  |  |

1. Next, we’re going to separate the data by the *categorical* variable **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**. Make a histogram like we did before but this time use the facet( ) function to create a different histogram for each group:

**gf\_histogram**(~\_\_\_\_\_\_\_\_\_\_\_\_\_\_ , data =\_\_\_\_\_\_\_\_\_\_\_\_\_\_)%>%**gf\_facet\_grid**(\_\_\_\_\_\_\_\_\_\_\_\_~ .)

Sketch these histograms (roughly) in the space below (or on the next page). Did we “explain” some of the variation we saw in the original histogram of pinkie length? That is, does knowing one person’s value on the variable \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ help us make a better guess at their pinkie length (if we are using the middle of the distribution as our first guess)?

1. Now, let’s try to separate the data by the *categorical* variable **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**. Make a histogram like we did before but this time use the facet( ) function to create a different histogram for each group of the new categorical variable. Sketch these histograms (roughly) in the space below. Did we “explain” some of the variation we saw in the original histogram of pinkie length?
2. Finally, let’s try to explore the relationship between pinkie length and the *continuous* variable **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**. Check your R cheat sheet for the **gf\_point** function to create a scatterplot. Sketch the scatterplot (roughly) in the space below. Does this continuous variable appear to “explain” some of the variation we see in pinkie length?

**Explaining Variation**

We start with variation in general. In this case, we started with everyone’s pinkie length (all the pinkie lengths lumped together). We first describe the general variation, but part of our job as statisticians will also be to “explain” that variation. This is where we bring in other variables to see if they can help explain *why* a variable varies the way it does.

When considering categorical variables to explain quantitative variables, we can see evidence of this if the separated data results in any distributions with LESS variation than the original, general variation, and a shift in the CENTER for each distribution (resulting in unique predictions for each category). Today, we found this with the variable(s) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

This can also be seen when considering two quantitative variables, if we find that the variables move in a linear way together with minimal scatter. Today we found this with the variable(s) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ .

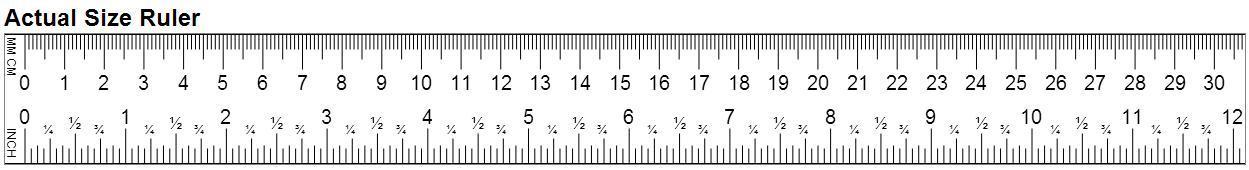
1. Once again, what does it mean to “explain variation”? How did we do it in the pinkie finger case?

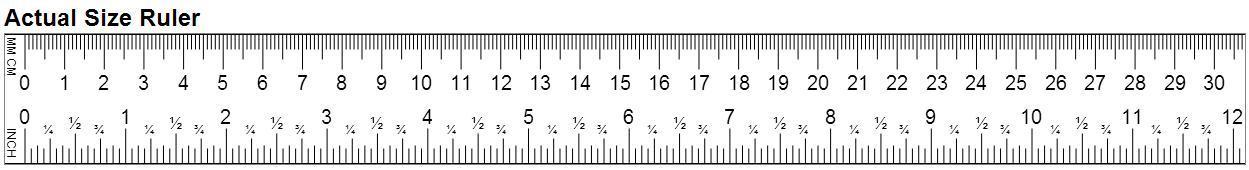
We won’t be able to explain all of the variation, so that will count as unexplained variation, sometimes called “random variation,” or “error.”

1. The **random variation** will include a few different kinds of variation such as the types listed below.
   1. Not-yet-explained Variation:
   2. Measurement error:
   3. Sampling variation (think back to the many samples we drew for the Martin v. Westvaco simulation...)

**\*Note.\* Make sure all questions are answered to receive full credit for classwork.**

Take Home Messages:

1. Distributions have variation (it’s likely that not all values will be the same when you measure something, even when two people measure the same thing using the same instrument you will get measurement variation)
2. Some of that variation can be explained by other variables: Gender, Height, Hormones, Injury, etc.
3. To “explain variation” is to use one variable (e.g., “Height”) to understand why another variable varies (e.g., finger length) (e.g., taller people tend to have longer finger lengths)
4. Some of that variation is “random variation” and may not be explainable (also known as “error”)
5. It’s important to examine where variation (or “error”) in our data comes from in order to minimize it and have a more accurate model of our variables (i.e., the less inherent variation there is, or the more variation that can be explained by other factors, the better our predictions will be. For example, if I wanted to predict your thumb length, I could make better predictions if I knew your height, or gender, or history of finger injuries, etc., than if I just guessed without knowing that information.



\*Note: 1 cm = 10 mm