Section 2
Statistical Investigations

2.1 – Introduction

Designing an investigation

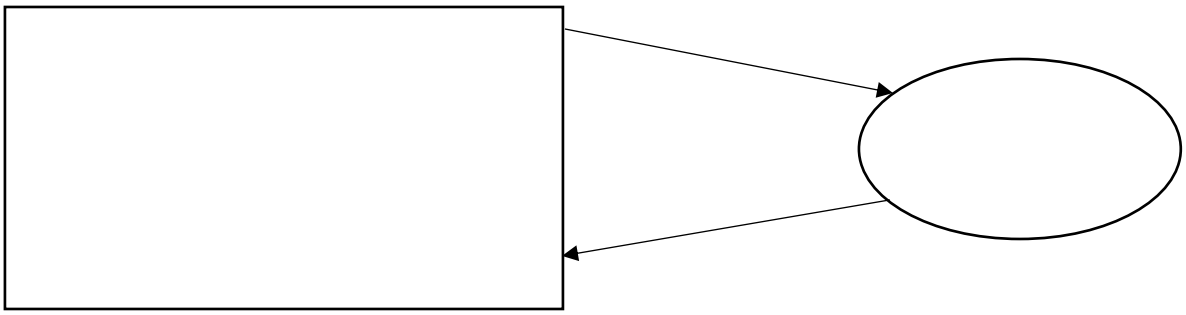
More and more professions rely on statistics to make informed decisions. Businesses collect data on consumers and products to project sales and inform marketing strategies. Clinical trials in medicine use statistics to determine if a new antibiotic or treatment is effective. Engineers often use statistics to measure and maintain the quality and reliability of systems. In all of these cases, statistics is leveraged to help quantify and understand the **uncertainty** of outcomes. In what ways can we measure this? Let's begin by examining an example of a statistical investigation.

> *Example:* You are interested in determining how much time undergraduate students at Illinois study in a given week. What kind of data would you collect? How would you collect it?

> What would you do with that data to address the goal of your investigation? What measures might you use?

As we can learn from this example, setting up a statistical investigation is often more challenging than doing the statistics once you have collected your data!

Probability and statistics



The _____ is the set of all subjects or objects relevant to the particular study.

A _____ is a subset of those subjects or objects that we selected with some method.

_____ uses information from a population to inform what samples might look like.

_____ uses information from samples to inform what the population might look like.

> *Example:* For the example on the number of hours studies by Illinois students, what would be the population of this study? What would a sample look like?

2.2 – Sampling and Replacement

Types of samples

A sample of size *n* selected from some population is called a _____ if it has the same chance of being selected as any other sample of size *n*.

> *Example:* A local manufacturing center produces 100 Vizio TVs per day. For quality control purposes, a simple random sample 15 of those TVs is selected and inspected for any possible defects.
>
> Let's number the TVs 1 to 100 based on when they are produced throughout the day. We can perform the following R Code to randomly select 100 of those TVs.
>
> ```
> sample(1:100, size=15)
> ```

Last week, we did an activity on cane toads where we used a sampler device to generate random samples of cane toads. This function in R works in a similar way to that sampler device – from the vector of items (or the physical bin of beads!) we can take a random selection from that vector. That sampler came to us pre-constructed with the inner workings of the sampler hidden to us, but soon, we will be constructing samplers to model real world random sampling processes ourselves.

The goal of taking a sample like this is to help ensure that the sample is _____ of the population. However, this is a surprisingly difficult task to do well in many situations.

Example: In our example about study hours with Illinois students, what might we need to conduct a sample like this? Is this feasible?

Collecting a sample in this way is often very challenging. There are other probability-based sampling methods that researchers can conduct that may be more practical for a given situation:

Systematic sample:

Cluster sample:

Stratified sample:

In practice, it may be difficult to use a probability-based sample for certain populations and study designs. This often results in researchers taking a _____, and can result in _____. Throughout the course, the tools we will learn all depend upon us selecting subjects randomly from the population using simple random samples. It's important to realize that in practice, this often does not happen, and we should be diligent to reduce bias in our results.

Examples: Determine the sampling method used in the two scenarios below.

1. A researcher was interested in determining the proportion of vegans in Urbana. They create a list of addresses from all residences in Urbana using data from Google Maps. They take a random selection of these addresses, and all of the residents at a selected address were asked if they were vegan or not.

2. A student at the University of Illinois is interested in determining whether all students supported the GEO bargaining process throughout the 2022-23 school year. To ensure that neither undergraduates nor graduates were overrepresented in their sample, they took random samples from each of these two groups. The University of Illinois enrolls 34,500 undergraduate students and 20,500 graduate students, so your sample of 110 students is composed of 69 undergraduates and 41 graduates.

Replacement

When selecting people or objects from a finite population, there are two ways we can choose to do the sampling. If we sample _____, we can only sample each unit from a population once, that is, any unit selected from our population is removed from any further possible selections. If we sample _____, we will select from all units of the population regardless if they have been sampled before.

Example: Consider the familiar sampling devices below. What replacement method is being used when sampling from these devices?

Flipping a coin 10 times, recording whether the outcome was heads or tails on each flip.

Rolling a six-sided die 10 times, recording the resulting number on each roll.

Drawing a hand of 10 cards from a standard deck of playing cards.

Using the `sample()` function in R to draw 10 items from a vector.

Most methods that we will learn in this class assume that we are sampling _____, as it is much simpler mathematically to assume the same probability of selection for each sampling unit throughout the process. This is likely counterintuitive to how you would likely carry out a sampling method in practice; however, as you typically don't want to sample the same individual twice. However, most applications of statistics work with relatively large populations and relatively small samples. Thus, the difference between sampling with or without replacement is basically negligible, as you are unlikely to select the same individual twice anyway.

As mentioned before, we will soon be working on creating samplers in TinkerPlots. There will be options to choose whether you would like your device to sample with or without replacement. This choice is crucial to understand and make sure your device matches both the real-world data generating process that you are modeling and the statistical assumptions for that test!

2.3 – Variables and Measurement

Types of variables

When conducting a statistical investigation, we identify features or attributes of interest from the samples we are collecting. These features are called variables.

Example: What is the variable of interest from the example on Illinois students' study time?

When determining our variables in a statistical investigation, it is important to recognize whether the variables are categorical or quantitative in nature.

Categorical variables:

Quantitative variables:

These types of variables can be broken down further:

Type of Variable	Definition	Mathematical operations
Nominal		
Ordinal		
Discrete		
Continuous		

Examples: Identify the type of variable in each of the following scenarios:

1. Favorite flavor of ice cream.

2. Number of hours studied.

3. Number of days in a month someone goes to the gym.

4. A student's class (freshman, sophomore, etc.).

5. Zip code (e.g. 61801, 61820).

6. Support for GEO bargaining in the 2022-23 school year on a 1-5 scale (5 = strongly support, 1 = strongly oppose).

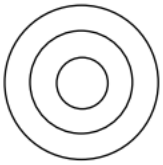
Measuring variables

Statistical investigations are filled with many types of uncertainties. Remember that our goal in using statistics is to use samples to make inferences about the larger population. But depending on what sample we obtained, our data may differ drastically. This type of variation is known as _____.

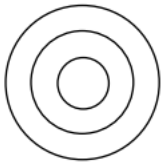
Thus, when we obtain a statistic, we know that this statistic will not exactly match the figure for the population. These numerical characteristics for the population are known as _____.

As previously alluded to, the sampling method we use will affect the accuracy of our results. Ideally, if we use a simple random sample, our results should be accurate. That doesn't mean individual samples will always match the population, but we don't have any reason to believe if we are missing "high" or "low" with the sample. Consider a dartboard: a darts player who is accurate is not going to hit the bullseye (or triple 20, for the savvy darts players) every time, but if we look at a "spray chart" of their darts, their darts should be centered around the bullseye. An inaccurate player would center the spray of their darts off the target. That player may be able to have a narrower spray of darts, but their aim is not centered on the target.

Accurate



Inaccurate



Taking this analogy back to statistics, we would say that the inaccurate darts player has a biased throw. If this were a statistical study, their sampling method would be biasing the results in some way based on how they were selecting participants. The difficult part with statistical studies is that we don't get a dartboard visual when we select a participant for a sample. Our statistics may be biased, but we don't know if we are missing too high or too low, or whether that statistic is just different from our parameter due to natural sampling variability.

> *Example*: Considering the previous example on study hours for Illinois students, what are some ways that we can construct a sample of Illinois students that may result in biased results?

2.4 – Summarizing Variables

Summarizing categorical variables

The simplest way to summarize categorical data is through percentages or proportions. We define and notate these as follows:

Sample proportion:

Population proportion:

Example: A soft drink company developed a new drink, and they want to gauge overall consumer preferences with this drink. They set up an open taste test at a shopping mall on a Friday afternoon, and are able to obtain 500 people to try their new drink as well as an existing soda to determine which they prefer. Of the 500 people they obtained, 317 preferred the new drink.

Population of interest:

Unit of observation:

Variable of interest:

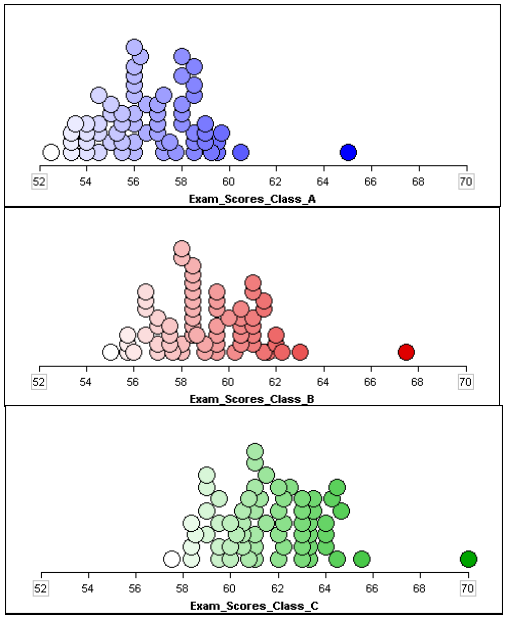
How was the sample taken? What kinds of biases might we observe?

Do we know the sample proportion in this investigation?

Do we know the population proportion in this investigation?

Summarizing quantitative variables: center

To understand what kind of values are typical for a quantitative variable, we typically look at measures of center. On your assignment that you turned in earlier this week, you were asked to think about this for the distribution of cell phone bills. How did you think about what was typical for a phone bill with the data distribution given? You also looked at distributions of exam scores – classes A, B, and C on that assignment were designed to have identical characteristics except for their measures of center, which shifted the distributions left or right along the x-axis.



The two most common measures used are the mean and median.

Sample mean:

Population mean:

Median:

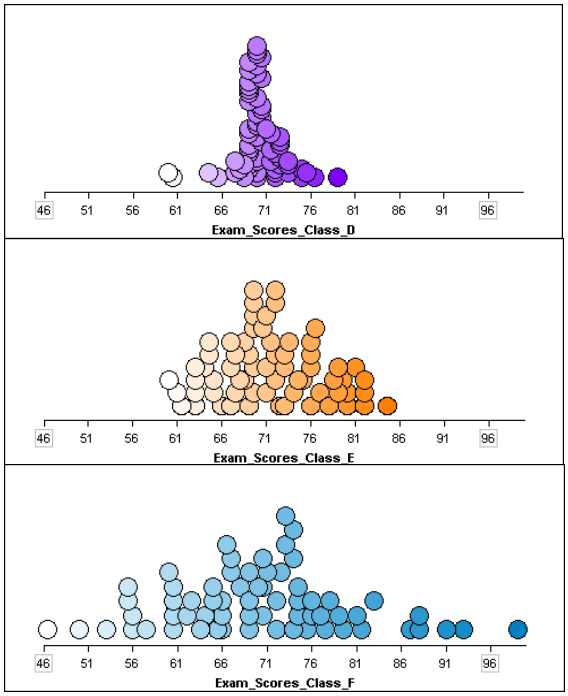
Whether you choose to use a mean or median to summarize the center of data often depends on the presence of outliers in the data. A mean considers the position of all data points equally, meaning it will react to any outliers in the data. The median only looks at the very central values of the data set, and ignores the values of the extreme values, so the median is not responsive to outlier values.

> *Example*: A senior statistics major was curious about the starting salaries for data analyst jobs in the USA. They obtained the following random sample of salaries from other recent graduates who started data analyst jobs in thousands of dollars:
>
> 72, 91, 80, 50, 90, 50, 55, 60, 100, 67
>
> What is our sample median? What is our sample mean?
>
> Suppose the senior statistics major made a typo and accidentally entered in 191 instead of 91. Would the median be affected by this change? How about the mean?

We compute the sample mean and median for this data in the previous lecture, but as a reminder, here is the R code for entering our data and computing the mean and median:

```
salary = c(72, 91, 80, 50, 90, 50, 55, 60, 100, 67)
mean(salary)          #mean/average of salaries
median(salary)        #median of salaries
```

Summarizing quantitative variables: variation
Another characteristic that came up during your assignment earlier this week was describing the overall **variation** in a distribution of data. Measures of variation describe how spread out or close together your data are. For the grade distribution examples on your assignment, distributions D, E, and F were designed to have the same measure of center, but their variation from that center differed.



What kinds of ways did you describe the variation quantitatively for the distribution of phone bills? I imagine this was challenging to do in comparison to measures of center, as you were likely familiar with the mean and median. One measure you are likely already been familiar with is the range.

Range:

The problem with range is that it is *very* sensitive to outlier values. Just one extreme observation will make the range much larger, even if it's just that data point that varies greatly from the rest of the data. To expand on this idea of the range, we can examine ideas of quantiles to look at inner ranges of the data.

Five-number summary:

Inter-quartile range (IQR):

This is a much more useful measure than the standard range for describing variation, but choosing to restrict your range to just the middle 50% can be a bit arbitrary. During your activity on hand spans this week, we examined ideas of measuring the typical variation from the center of a data set. As a class, we ended up deriving this idea of taking differences of each data value from the mean to understand how far each data value was from the mean. But doing this created both positive and negative differences, and taking the average of all these differences will always result in 0 due to the positive and negative values cancelling out perfectly.

Thus, we took the absolute values of these difference, and then took the average of these points. This measure is typically referred to as the mean absolute deviation, or MAD.

$$\text{MAD} = \frac{\sum |x_i - \bar{x}|}{n}$$

But at the very end of the activity, we looked at a measure called standard deviation. At that point, we just noted that this was "similar" to MAD. The formula for standard deviation can be written out as:

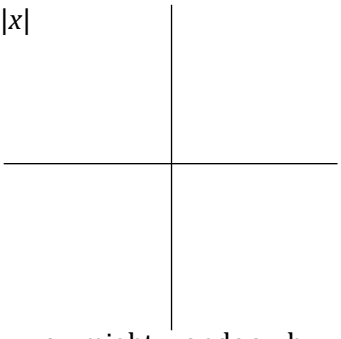
Sample standard deviation (and variance):

Population standard deviation (and variance):

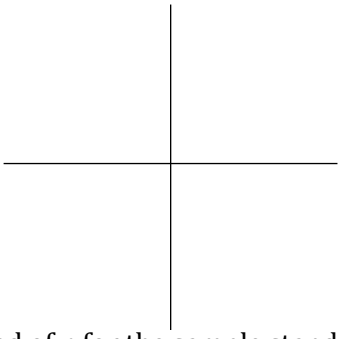
So there is some similarity between these values, where instead of getting the average distance from the mean, we have the square root of the (almost) average of the squared distances from the mean. Squaring values is similar to absolute value in that they are both functions to make differences positive, but why use the more complicated version?

The issue with MAD is that absolute values pose issues in trying to evaluate theoretical statistics concepts using calculus. As a result, standard deviation ends up playing nicely with many different statistical models that we will use. Since calculus is not a prerequisite for this course, I will not delve too deep into why this is the case, but the main idea is that squaring is a "smooth" function and absolute value is "pointy", and calculus doesn't like pointy things.

$y = |x|$



$y = x^2$



Additionally, you might wonder why we divide by $n - 1$ instead of n for the sample standard deviation, making it not a precise average of the squared deviations. The intuitive explanation for this is that the sample standard deviation is not only an estimate itself, but it uses the sample mean inside that estimate instead of the population mean. For a given data set, the sample mean is the single number that minimizes the total distance from each point to that number, and so in comparison to the population mean, using the sample mean systematically underestimates the variability of your data. Mathemagically, it turns out that the exact correction for this underestimation is to divide by a smaller $n - 1$ instead of n. If you find either of these answers completely unsatisfying, I encourage you to take Stat 400/410 to get a deeper understanding of this theory!

With all that said, we will interpret standard deviation as if it were like the mean absolute deviation or MAD in this class. That is, the standard deviation roughly gives us the average distance that a data point is from the mean or average value of the data set.

Example: Compute these measures of variation for our salary example from before.

We could compute range by hand, but it's simplest to let technology compute IQR and especially standard deviation!

```
max(salary) – min(salary)                    #sample range
quantile(salary)                             #five-number summary
quantile(salary)[4] - quantile(salary)[2]    #IQR
sd(salary)                                   #standard deviation
```

Example: Write out a sentence that interprets the standard deviation for the salary data set, using the data's context in your interpretation.

2.5 – Graphing and Plotting Variables

Graphics for categorical variables

Barplots are the standard choice for visualizing a single categorical variable. These plots show the names of the categories themselves on one axis with either the frequency or relative frequency (proportions) on the other axis.

Example: Draw out a bar plot for the new drink taste test study, where 317 people preferred the new drink, and 183 preferred the old drink.

We can also leverage R to create barplots for us.

```
drink = c(317, 183)              #vector to store the counts
barplot(drink,                   #referencing the vector
main = "Preferred Drink",        #title
xlab = "Drink",                  #x-axis label
ylab = "Frequency",              #y-axis label
names.arg = c("New", "Old"))     #category labels
```

To produce a graph with relative frequency (proportions), we just need to divide our vector by the sample size and update our *y*-axis label.

```
barplot(drink/500,                #now dividing by 500
main = "Preferred Drink",
xlab = "Drink",
ylab = "Proportion",              #updated y-axis label
names.arg = c("New", "Old"))
```

For variables with more than two categories, boxplots can be made in R by simply having a longer vector of counts and corresponding "names.arg" vector.

Graphics for quantitative variables
Two ways to show the overall distribution of data are through histograms and stem-and-leaf plots.

Example: Draw out a stem-and-leaf plot for the salary data

72, 91, 80, 50, 90, 50, 55, 60, 100, 67

Example: Use R to create a histogram of the salary data.

We can use the following R code to create a histogram.

```
hist(salary, breaks=5)
```

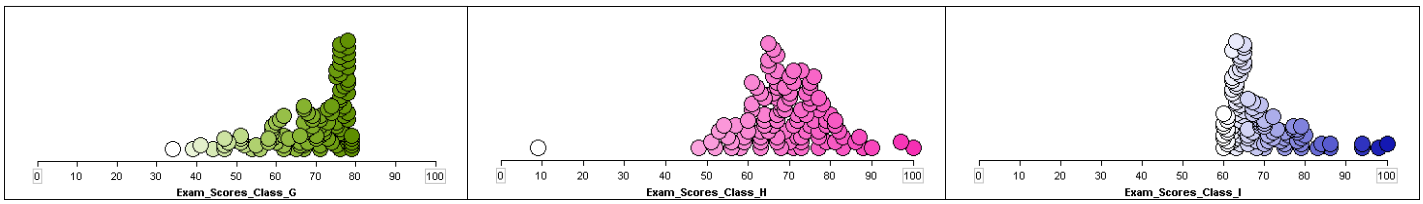
Histograms are made by dividing up the interval of data into equal sized bins, and then use vertical bars to indicate how many data points are within that bin. Stem-and-leaf plots give us more information regarding individual data points, but histograms give us more flexibility in choosing the width of our bins.

The histogram gives us information about the central tendency and the variability of a distribution, which we have talked about previously. Let's draw some shapes of histograms to illustrate this!

Center Variability

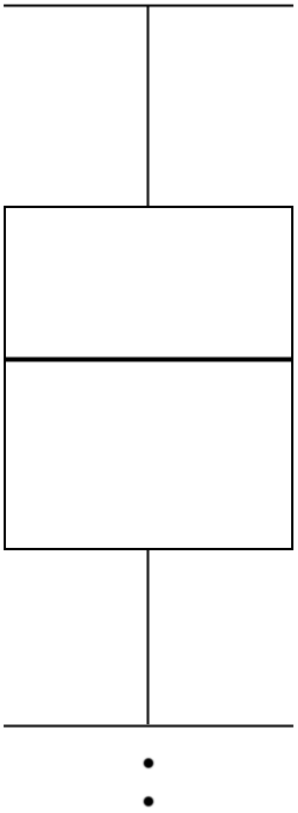
An additional piece of information that histograms can provide is the _____ of data.

To describe the shape of data, we will usually talk about its _____. We examined the idea of shape in our previous assignment with grade distributions. Distributions G, H, and I illustrated three different shapes as shown below.



Example: Describe the shape of the salary data from the previous exercises.

One final visualization used for quantitative data is a boxplot, also known as a box-and-whisker plot. These are effectively visualizations of the five number summary we described earlier. The visualization on the right shows how these numbers from the five number summary correspond with the boxplot.



However, it is important to note that boxplots display outlier values in the data set as dots, separate from the box and whiskers themselves. This occurs when a data value meets one of the two conditions below:

- Data value > Q_3 + 1.5*IQR
- Data value < Q_1 – 1.5*IQR

If this occurs, the outer whiskers would not actually represent the true max/min values, but the non-outlier max/min values. The other values represented on the boxplot still represent the appropriate quantiles.

Example: Draw a boxplot for the salary data based on the five-number summary found previously in R.

We can also use R to create boxplots for us.

```
boxplot(salary, main="Boxplot of salary", ylab="salary")
```

Unlike with the histograms, the graph does not come with a title or axis label, so additional arguments are added to the boxplot code above.

2.6 – Additional Practice

Example: Using the **bears.csv** data file we used in the additional practice in the last section, complete the following:

- Identify each of the variables in this data set as quantitative/categorical, and whether they are nominal, ordinal, discrete, or continuous.

- Create a histogram of the weights of the bears in R.

- Create a boxplot for the weights of the bears in R.

- Find the mean and standard deviation for the weights of the bears. Interpret this standard deviation in the context of the problem.