Size Matters! & Other Statistical Concerns

Since this unit involves the collection, analysis, and reporting of data, it seems like a good idea to familiarize ourselves with some key terms and concepts related to data and statistics.

You do NOT have to use all these terms when writing your report. This information is helpful for us to know as we navigate all the information we see, hear, and sort through from various outlets on daily basis.

Data ► Information gathered during a study.

Statistics ► Collection and classification of data in number form.

Quantitative ► Objective results precisely measured in numbers.

Qualitative ► Subjective responses described in words.

While *data* in general can be both qualitative (words/text) or quantitative (numbers); *statistics* are a way to summarize quantitative data and are always reported as numbers (which can include percentages).

Population ► Largest group of similar things (which can include people).

Sample ► Smaller group from the population that is supposed to represent that population.

Data are (yes, the term *data* is plural) collected from a *population* about which the researcher is learning. Since in most situations data from an entire population are nearly impossible to collect, researchers collect a *sample* of data from that population.

In order for the data collected from a sample to be relevant or useful, the sample should be *representative* of the whole population, which means it should include approximately the same percentage of the various types of things as found in the whole population itself.

Representing a population becomes a problem when sample sizes are too small or too large. Size matters because…

► **A *too-small sample*** does not have the same proportion of all of the different types found in the population and thus fails to accurately represent the various types within that population. A too-small sample, then, produces inconclusive results that cannot be used to prove any claim.

For example, the survey we will conduct in this course will be a too-small sample that will not accurately represent the school's entire student population.

Further, if only one student in our sample reported living in a particular location (let's say, Montana), then we would need to understand that the responses of that one student would not, in fact, accurately represent the potential responses of all or most people living in the state of Montana.

► **A *too-large sample*** also produces inconclusive results because some samples are so large that statistical significance can be found for and between so many of the variables being measured that none of the results matter.

Since we just mentioned these terms…

Variables ► Characteristics/Values being measured in a sample.

Significance ► Measure of whether research results were due to chance.

When we talk about *variables* and *significance*, we are almost always also talking about…

Descriptive statistics ► Summary of a variable for a sample.

Procedures for organizing, summarizing, and displaying data.

Inferential statistics ► Ways to find relationships among variables.

Methods by which inferences are made about a larger group (population) on the basis of observations about a smaller group (sample).

Variables are the values being measured in a study, like age, truck preference, and environmental concern. When discussing variables in general—like how many Xers prefer Fords and how many Millennials prefer Chevys—we are talking about *descriptive statistics*.

But when we start investigating the relationships between variables, like how respondents' age and level of concern about the environment may impact their truck preference, we are talking about *inferential statistics*.

These two types of statistics should not be confused.

Presenting descriptive statistics as inferential statistics (that is, as numbers that imply a relationship between variables) creates many misunderstandings because descriptive statistics are often reported in ways that indicate a *significance* that may not actually exist among the variables in the data.

What *statistical significance* means is that the relationship between two variables is not due to chance or luck. Therefore, a *statistical significance* among variables

means those variables are, in fact, related in some way—a way that could or could not be proved from the data set. (See Example 1 on page 6.)

Descriptive statistics only look at the results of a data set at large and do not indicate relationships among any of the variables being reported, so the only significance they can show is a sizable difference between variables.

For example, a study may survey Americans about the most frequently consumed soft drink in all fifty states.

A statistical significant descriptive statistic based on this data might show a large difference in soft drink consumption between North Carolinians and Californians, but it would not indicate any relationship among the soft-drinking habits of the people in the states or that the soft-drinking habits of people in California are causing the soft-drinking habits of people in North Carolina. (Or even how the people in these states refer to these types of beverages, as soft drinks or sodas or pop or Coke!)

To explore this example further, let's say the data show that 3 million people in North Carolina consume soft drinks every day. We'll round the population of North Carolina to 10 million people, so that our statistic would be that 30% of North Carolinians drink soft drinks every day.

The data also show that 3 million people in California consume soft drinks every day. When we round the population of California to 40 million people, our statistic would be that just over 13% of Californians consume soft drinks every day.

The news report based on this data may claim, then, that Californians are healthier than North Carolinians because of the perceived significant difference in these percentages whereas the actual number of people are the same.

However, the only claim this descriptive statistic can actually make is that there is a higher percentage of people who consume soft drinks in North Carolina than in California.

No other inferences can be made based on this data observation.

Almost all of the statistics we learn about through media outlets are descriptive statistics that do not report significance in an accurate way (even though they claim to).

Similarly, almost all of the descriptive statistics we learn about through media outlets do not accurately report *causation*, which indicates that one variable has an effect upon another variable.

Inferential statistics, however, do the work of finding relationships between or among variables and do sometimes suggest causation.

But even then causation can only be determined if the research was conducted in the form of an experiment—which is also why most of the inferential statistics that reveal possible causations are reported almost exclusively in scientific journals and involve complicated mathematical models. (See Example 2 on page 7 of this document.)

We should not confuse causation with correlation.

Correlation ► Degree to which two factors seem to be connected.

Causation ► Relationship between cause and effect.
Action of one factor causing another factor.

We should not confuse causation with correlations. Just because two factors are related does NOT necessarily mean that one factor causes the other.

An examples of a correlational relationship is when two factors seem to move together in the same direction (like where one factor increases, another factor also increases) or in an inverse direction (like where one factor increases, another factor decreases).

In this way, correlational data do show relationships between factors but, again, do not show that one factor necessarily causes the other.

As an example, let's take a look at research done about the performance of the elite NBA athletes.

Data suggest that there is a significant correlational relationship between the arm-span-to-height ratio of NBA players and their success. While research does not indicate what that relationship might be (Does a larger arm span really make one a better basketball player?), it does indicate that NBA players with arm spans that exceed their height are often successful.

Here what researchers have found is that a *correlation* exists between the variables of player arm span and player success.

Here is also where we should be careful not to confuse correlation with causation:

> Just because successful NBA players tend to have long arm spans does NOT mean that long arm spans *guarantee* a player's successful career in the NBA.

Because in addition to arm-span-to-height ratios, researchers have also found many *other* correlations for successful NBA players that not all successful NBA players have.

In short, the data have not revealed the exact causes for success in the NBA because there are too many possibilities.

In the same way that we should be careful to interpret data in accurate ways, we should also be careful to formulate good questions that will lead to usable data. Two ways to ensure that research questions are posed in clear, non-leading ways are to define the question terms and to be mindful of the phrasing of the questions.

For more information about good and bad research questions, see the Reading Link: "5 Common Survey Question Mistakes that'll Ruin Your Data."

For an example of defining the terms of a question, see Example 3 on page 8 of this document.

By keeping in mind these statistical concerns, we can use the data we navigate every day to help us make decisions and become more informed citizens of our communities.

Even if we may not be able to confirm the odds of one day achieving our hoop dreams.

Example 1: Descriptive Statistics Reporting Findings that Are Not Significant

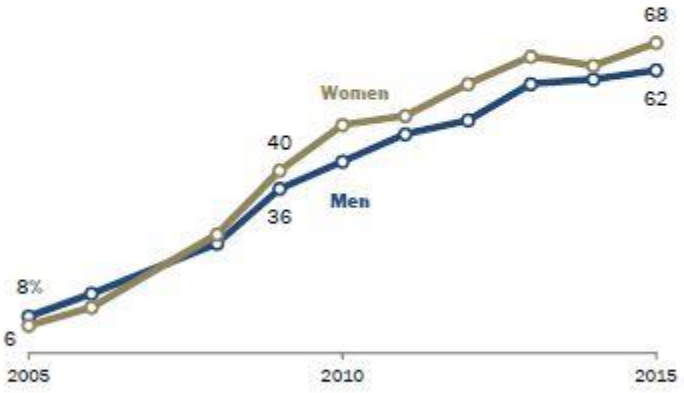
Social Media Usage by Gender: A Shifting Balance Over Time, With Parity Today

In 2005, 8% of men and 6% of women used social media.

Starting in 2009, women started using social media at slightly higher rates than men, although this balance has shrunk yet again in recent years. Today, 68% of women and 62% of men report social media usage, a difference that is not statistically significant.

Women and Men Use Social Networking Sites at Comparable Rates

Among all American adults, % who use social networking sites, by gender



Source: Pew Research Center surveys, 2005-2006, 2008-2015. No data are available for 2007.

PEW RESEARCH CENTER

Source: Perrin, A. (2015, October). Social media usage: 2005-2015. *Pew Research Center*. Retrieved November 5, 2020, https://www.pewresearch.org/internet/2015/10/08/social-networking-usage-2005-2015/

Example 2: Statistics Results with Mathematical Models

Results

TPE in Fake News

Consistent with H1, the paired-sample t test revealed that participants perceived a greater influence of fake news on others ($M=5.57$, $SD=1.15$) than on themselves ($M=4.17$, $SD=1.47$), $t(334)=15.22$, $p<.001$. Such a result confirmed that there was, indeed, a TPE with respect to fake news.

Impact of TPE

Measurement Model. The full measurement model was first assessed utilizing a standard confirmatory factor analysis with maximum likelihood estimation procedure. The factor analysis yielded a moderated fit, $\chi^2=110.698$, degrees of freedom (df)$=39$, $p<.001$, root mean square error of approximation (RMSEA)$=.074$, comparative fit index (CFI)$=.972$, 95% confidence interval [CI]$=[0.058, 0.091]$. Table 1 reports the bivariate correlations among all variables.

The chi-square test is usually sensitive to sample size (Kline, 2011). Larger samples (e.g., $N>200$) will almost always result in a chi-square at $p<.05$. The chi-square

Source: Yang, F., & Horning, M. (2020). Reluctant to share: How third person perceptions of fake news discourage news readers from sharing "real news" on social media. *Social Media + Society*. doi: DOI: 10.1177/2056305120955173

Example 3: Defining Question Terms

Defining Social Media Users

The definition of a social media user in Pew Research Center surveys has changed when circumstances change. Our question about social networking use has varied over the course of these surveys, depending on the most common social networks at the time.

In 2005, social media users were defined as those who said "yes" to "Do you ever use online social or professional networking sites like Friendster or LinkedIn?"

In August 2006, social media users were defined as those who said "yes" to "Do you ever use an online social networking site like Myspace, Facebook or Friendster?"

From May 2008 to August 2011, social media users were defined as those who said "yes" to "Do you ever use the internet to use a social networking site like Myspace, Facebook[2] or LinkedIn?"

From February 2012 to January 2014, social media users were defined as those who said "yes" to "Do you ever use the internet to use a social networking site like Facebook, LinkedIn or Google Plus?"[3]

The most recent measure in July 2015 defined social media users as those who said "yes" to "Do you ever use a social networking site like Facebook, Twitter or LinkedIn?"

[2] In August 2008, the question wording was "Do you ever..." and did not include the phrase "use the internet to..." In December 2008, only Myspace and Facebook were in the question, not LinkedIn.
[3] In August 2012, the question wording was "Do you ever..." and did not include the phrase "use the internet to..."

Source: Perrin, A. (2015, October). Social media usage: 2005-2015. *Pew Research Center*. Retrieved November 5, 2020, https://www.pewresearch.org/internet/2015/10/08/social-networking-usage-2005-2015/

Finally, while we're at it, here are a few other helpful terms and definitions:

Mean ► Average result of test, survey, or experiment.

Median ► The middle value (the score that divides results in half).

Mode ► Most common (or frequent) result of test, survey, experiment.

Range ► Difference between the highest and lowest scores.