



**BALL STATE**  
**UNIVERSITY**

---

# **Module 1 Lecture - Sampling and Data**

Introduction to Statistical Methods

---

Quinton Quagliano, M.S., C.S.P

Department of Educational Psychology

## Table of Contents

<b>1</b>	<b>Overview and Introduction</b>	<b>2</b>
1.1	Textbook Learning Objectives . . . . .	2
1.2	Instructor Learning Objectives . . . . .	2
1.3	Introduction . . . . .	2
<b>2</b>	<b>Definitions of Statistics, Probability, and Key Terms</b>	<b>3</b>
2.1	Introduction . . . . .	3
2.2	Probability . . . . .	4
2.3	Key Terms . . . . .	4
<b>3</b>	<b>Data, Sampling, and Variation</b>	<b>6</b>
3.1	Introduction . . . . .	6
3.2	More on Qualitative Data . . . . .	6
3.3	Better Sampling . . . . .	7
3.4	Natural Variation . . . . .	8
<b>4</b>	<b>Frequency, Frequency Tables, and Levels of Measurement</b>	<b>8</b>
4.1	Introduction . . . . .	8
4.2	Levels of Measurement . . . . .	9
4.3	Frequency . . . . .	9
<b>5</b>	<b>Conclusion</b>	<b>10</b>
5.1	Recap . . . . .	10
5.2	Lecture Check-in . . . . .	10

# 1 Overview and Introduction

## 1.1 Textbook Learning Objectives

- Recognize and differentiate between key terms.
- Apply various types of sampling methods to data collection.
- Create and interpret frequency tables.

## 1.2 Instructor Learning Objectives

- Understand how variables can be described and represented in several different ways
- Appreciate the basic of how data comes to be, and how the collection method implies the scale of measurement

## 1.3 Introduction

- “Statistics” comprises various quantitative methods to \_\_\_\_\_ and analyze numeric data
  - I like to think of statistics as a \_\_\_\_\_ methods to various sciences - biology, psychology, education, economics, etc.
- At the core of statistics is \_\_\_\_\_, or how we describe the likelihood of certain events occurring
  - We’ll briefly dive into the \_\_\_\_\_ of probability in module 3
- In the modern age, statistics often tends to be closely associated with fields like \_\_\_\_\_ science and computer science
  - It’s important to understand that how we gather and transform data has a \_\_\_\_\_ impact on the end result or our analyses
- Statistics are just one part of the total research process, but are affected greatly by how we \_\_\_\_\_ our studies
  - In fact, statisticians are often involved with the entire research process, not just analysis - because they need to advise on the “best” way to gather data so that it is useful

### ! Important

While this is not a research design class, we will talk about design as it is relevant to analysis.


- As we progress through this class, we’ll also touch on how to \_\_\_\_\_ and show-off data and results

- Especially as we use more complicated methods, using the right graphical techniques can help us make our results \_\_\_\_\_ to non-statistical audiences.

## 2 Definitions of Statistics, Probability, and Key Terms

### 2.1 Introduction


- Statistics starts with data, which is some \_\_\_\_\_ amount of a phenomenon
  - In \_\_\_\_\_ research, our data may be grades, measures of student well being, measures of student engagement, etc.
  - At a more macro-scale, universities use measures like enrollment, class size, financial cost, etc for analysis

 Discuss: What are other examples of common data we would gather in education?

- We can treat this data via \_\_\_\_\_ statistics, where we summarize, describe, and demonstrate various components and structure of the data
  - E.g., \_\_\_\_\_, mode, standard deviation, etc. - we'll describe all of these in more detail later
- Usually in addition to descriptives, we would also do some type of \_\_\_\_\_ statistics, which helps draw probabilistic conclusions from the data
  - E.g., T-tests, ANOVA, anything that gives us \_\_\_\_\_ - once again, we'll be revisiting what exactly these are later in the course
- Crucially, statistics should always be \_\_\_\_\_ by a desire theoretically understand and investigate data
  - As researchers, our goal is to not \_\_\_\_\_ for conclusions from data, it is to carefully and responsibly treat it for analysis
  - Metaphorical example: we are not casting a wide net, we are casting a single line towards an area we are interested in

## 2.2 Probability

- **Probability** helps us contextualize inferential statistical \_\_\_\_\_ as having some amount of \_\_\_\_\_ that our results occurred not simply due to chance
- Probability underlies many \_\_\_\_\_ we make intuitively
  - E.g., What are the chances the car I buy will be low maintenance costs? How likely am I to really use this new device I buy? If I get this degree, I think I should have a higher likelihood of getting my desired job

 Discuss: Give another example of considering probability of a certain event occurring in your personal life

- Just like with our personal life, probability can be a useful way to understand a certain \_\_\_\_\_ in our data and research, and measure our confidence


## 2.3 Key Terms

- To start we need to establish many of the words we often use in statistics


### Important

Many terms in statistics have alternative names (which can be frustrating on occasion), I'll try to highlight when a term may sometimes be called something else.

- First, the group we desire to \_\_\_\_\_ and understand is our **population/population of interest**.
  - E.g., All high school students, all teachers
  - The population is a group we can't practically, \_\_\_\_\_ study (as they are usually too large or dispersed!)
  - In the case that we did somehow gather data about \_\_\_\_\_ members of a population, we would call this a **census**

 Discuss: Given this definition of a census, I just provided, explain why we would call the periodic counting of all individuals in the US a 'census'

- Instead of gather data on the population as a whole, we take a \_\_\_\_\_ subset of individuals that are meant to represent the population, which we would call a **sample**.
  - We get our sample via some method of **sampling**, which is exactly how we get our subset - it can be done in a “good” or “bad” way, more on this later
- The numeric values that we use to calculate and describe on our \_\_\_\_\_ is called a **statistic**, which, in turn, is meant to represent the **parameter** of the population
  - A mnemonic to remember this:
    - \* **P**oulation → Parameter
    - \* **S**ample → Statistic
- But, remember our goal is often to study the \_\_\_\_\_, not just the subset we gather data for!
  - Thus, we want to have great \_\_\_\_\_ that whatever statistics we have, accurately represent their respective parameters
  - One part of ensuring this accuracy is to use a sampling method that results in a **representative sample**
- **Variables** are some defined measure/observation with variance in the data
  - I.e., there needs to be \_\_\_\_\_ numbers in the data - otherwise it is a **constant**
  - Variables can be either **numeric** or **categorical**, that is, they produce data of a specified type
  - E.g., Age in years → \_\_\_\_\_
  - E.g., Job title → \_\_\_\_\_

 I gather information about all the individuals enrolled at a local college, and all live in the state of Indiana. Is state of residence a constant or variable?

- A) Constant
- B) Variable
- C) Neither
- D) Both

Explanation:

## 3 Data, Sampling, and Variation

### 3.1 Introduction

- Let's dive more into the different \_\_\_\_\_ of data
- **Qualitative data** come from a more descriptive (via words or \_\_\_\_\_)
  - E.g., Eye/hair color
  - Because of its nature, it is almost always \_\_\_\_\_ in nature
  - Qualitative data is often times represented in counts of occurrences of a certain description
- **Quantitative data** is something represented by an \_\_\_\_\_ or numeric measurement
  - However, within quantitative data, it can be **discrete** or **continuous**
  - Discrete data is that which is counted, or has no intervals between the integers, e.g., number of phone calls had → I can't have half a phone call
  - Continuous data does indeed have \_\_\_\_\_ between integers, e.g., Age → I can be half a year of age.

#### ! Important

As a general rule of thumb, quantitative data is easier and more versatile in analysis, but also is more reductive.

### 3.2 More on Qualitative Data

- As alluded to early, qualitative data can be represented by \_\_\_\_\_ or **proportions**.
  - These are often shown in tables or graphs that somehow \_\_\_\_\_ those proportions


### 3.3 Better Sampling

#### ! Important

This section is going to briefly veer into good research design, as it plays an important role in what conclusions we can make from the data. Our statistics are only as valuable as how well they represent their parameters!

- Ideally, a \_\_\_\_\_ sample will have similar characteristics in similar proportions to the population they are taken from
  - E.g., veterans in the United States skew more male than female, thus, a representative sample of that population will likely be more men than women
  - A “good” sampling method, will result in this \_\_\_\_\_ between the sample and population
- Good sampling methods are those that are **random**, which means that each member of the population has an equal, \_\_\_\_\_ chance of being selected into the sample
  - E.g., if my population was veterans, and there are 1000 total veterans, a random sampling technique would mean that each of those 1000 individuals has a 1 in 1000 chance of being in the sample
- Within the broader classification of \_\_\_\_\_ sampling methods, there are several subtypes
  - **Simple random** sampling is when we \_\_\_\_\_ all individuals in the population and simply use some random method to select numeric IDs of individuals to be included in the sample
  - **Systemic** sampling is similar, but uses a process of starting with a specified ID and \_\_\_\_\_ by a specific number to select the others
  - **Cluster** sampling is done when we have pre-existing clusters, or \_\_\_\_\_, in our population, and we randomly select from those clusters to include all individuals from the selected clusters
  - **Stratified** sampling is similar in concept to cluster sampling, but done with some individualistic \_\_\_\_\_ of the participants
- Any **non-random** method that does not follow the above philosophy can not ensure the \_\_\_\_\_ of the sample
  - Our \_\_\_\_\_ and interpretation may be swayed or hindered by this non-randomness, thus causing a sampling error
  - Another way to describe this is by calling this \_\_\_\_\_ process as a **sampling bias**



 Discuss: I put up a poster with a QR code in the physical office for the educational psychology department, and wait for individuals to complete my survey, does this seem like a random or non-random method of sampling?

### 3.4 Natural Variation

- It is normal and \_\_\_\_\_ that, if we take two different samples of a population, we can and will find that they have somewhat different characteristics
- However, if these are both believed to be representative samples, they should become \_\_\_\_\_ similar as they become larger
  - This is an idea we will revisit in our discussion on the Central Limit Theorem, a concept we'll go into much more later
- Proper use of statistics will help us \_\_\_\_\_ for the fact that our samples naturally vary

## 4 Frequency, Frequency Tables, and Levels of Measurement

### 4.1 Introduction


- Earlier, we gave some examples of how to \_\_\_\_\_ different variables as qualitative, quantitative, discrete, continuous, etc.
- We'll introduce some other terms to help us classify our variables, which will be \_\_\_\_\_ when we treat it

#### Important

It cannot be understated how important it is that you can properly describe and classify your variables, as it is crucial in determining what descriptive and inferential statistics can be used.

## 4.2 Levels of Measurement

- **Nominal scale** variables are qualitative and categorical, with classifying \_\_\_\_\_ and no defined “order”.
  - E.g., state/country of residence, hair/eye/skin color
- **Ordinal scale** variables are those that have an order, but there is not a clear \_\_\_\_\_ between each interval or place in the order
  - Thus, it sort of blurs the line between categorical and \_\_\_\_\_
  - E.g., place in a foot race, class rank
- **Interval scale** variables do have a clear, consistent interval in-between each integer, but no clear starting point
  - E.g., temperature, height (in inches)
- **Ratio scale** is the same as interval scale, but does have a clear zero point as well.
  - E.g., score on an exam
- Generally, the different levels/scales of measurement are different levels of restrictive for analysis, in this order: Nominal (most restrictive) to Ratio (least restrictive).
  - Of course, having mostly nominal scale data does not always spell doom, but it can involve more tricky \_\_\_\_\_

 Discuss: I put all the students in my class from shortest to tallest and assign them the number they are from being the shortest in the class, what scale of measurement would this data be and why?

## 4.3 Frequency

- When dealing especially with nominal or ordinal scale data, it can be useful to represent the data in terms of how often a certain data type \_\_\_\_\_ and how often it occurs relative to the other possibilities.

Gender	Frequency	Relative Frequency
Men	18	0.45
Women	22	0.55
<b>Total</b>	<b>40</b>	<b>1.00</b>

## 5 Conclusion

### 5.1 Recap

- Using statistics with our data first starts with properly understanding and identifying our variables, and their scales of measurements
- Sampling plays an important role in how representative our data is of the population of interest, which, in turn, helps establish whether our results are generalizable or not
- Probability is at the core of analyses, and helps us establish the likelihood of events occurring, which we will be revisiting later

### 5.2 Lecture Check-in

- Make sure to complete and submit the lecture check-in