Exploring Data & Statistics

In this lesson, we'll cover another important piece of the machine-learning puzzle: statistics! Growing up, you probably noticed how many days were cold during the winter and how many days we warm during the summer, and without knowing when winter necessarily was, you could see that it was getting wintertime whenever it stopped being warm. Getting older, you could read the statistics on basketball scores, gas prices, and the proportion of episodes in a season that wasn't worth watching. I bet your teachers once made you all measure everyone's heights in your class, if only to see how tall the average boy and girl were. All of these measures cover the idea of ***statistics***.

To begin, we have the following definitions:

Definition:
Statistics is the study and science of data.

Pretty much, it's the topic that makes data useful. We do this by translating the information given to us into a language which can be perceived by the public.

Definition:
Experimental Units or **Observational Units** are the objects described by a set of data. These units may be people, animals, things, etc.

As stated, while we may use people most of the time, this isn't necessarily true. Folks in the social sciences (psychology, sociology, etc.) will often use people for their experimental units, mainly because those are often the subjects they care about.

One will take the topic of stress and test some number (what we will later denote by the letter ***n***) of individual – or more specifically, the clients referred to you – to record their signs or diagnoses of stress.

Other areas, such as construction and engineering, may instead look at a specific product to perform statistical tests. For example, perhaps we would like to test the durability to a Glue Brand X. We can use a number of statistical methods to determine how durable this brand is, ranging from basic descriptive statistics to more rigorous inference studies. Either way, our experimental units will end up as the items we apply the glue to.

Definition:
A **variable** is any characteristic of an individual or object under consideration.

This is our fun stuff. These characteristics are what allow us to classify what differentiates certain experimental units from others. Some examples of this include **height**, **age**, **eye color**, **number of movies you own**, and **proportion of nose hair compared to eyebrow hair**.

Of these variables, we can classify them into two specific types:

- **Categorical (or Qualitative) Variable:** *A variable which places an experimental unit into one of several groups of possible categories.*

These variables are typically ones that use words as opposed to numbers. Someone's **shirt color** would be an ideal example of a categorical variable, since there are only so many different types of shirt color that one could respond. (Also, when asking such a question, you would probably limit it to colors like **Blue, Red, Green, Yellow, Orange, Black,** and **Other**).

- **Quantitative Variable:** *Any variable measure a numerical quantity on each experimental unit.*

These types of variables deal with numbers over a specified range. For example, **height** would be considered a quantitative variable, as one's height may range from **0** to perhaps **10** ft (*anyone over 10ft may not perhaps be considered human – watch out!*).

Within our quantitative variables, we have two subtypes:

- **Discrete Variable:** *A quantitative variable that can only be a finite or a countably infinite number of values.*

For those who are a little fuzzy to all of these terms, **finite** means that there is an end. So, our range could be {0, 1, 2, 3, 4, …, 18, 19, 20} (number of fingers and toes one has ☺) and we could consider this variable to be discrete. To be **countably infinite** (or simply **countable**) means that we do have an infinite range (insanely huge!), however we can count them up. Usually this means some values stretching to infinity (like the number of offensive comments made by a newscaster – it could go on forever!). However, we are still able to count them off without missing anything in between.

- **Continuous Variable:** *A quantitative variable that can have infinitely many values corresponding to the points on a line interval.*

This may sound confusing, so the best way is to consider a number line from 0 to 1. On that number line, there are sooooo many possibilities. You could select something nice like ½, 0.246, or 0.0003. However, you could get something like $\pi/27$, something irrational and pretty darn hard to enumerate… but it's there! When a variable hit values that may be stuff like that, we typically have ourselves a continuous variable.

In the table such as the one below, we are given a series of values which relate to our experimental units. In this situation, we denote these as **cases,** and the **name** value relates to the **label**, also known as what we actually call these cases. As below, there are three cases, labeled under **Jodie**, **Mark**, and **Fred**. From this, we have our columns of variables, each describing our cases. (Note that Fred is female!)

Example:
[[1]] Below is a list of response values for possible roommates for an apartment. Describe which of the following variables are categorical and which are quantitative.

Name	**Age (Years)**	**Sex**	**1st Major**	**Scholarship**	**Cat?**	**Dog?**	**GPA**	**State of Residence**
Jodie	20	Female	Physics	Yes	No	No	3.12	Nebraska
Mark	21	Male	Education	No	Yes	No	2.86	California
Fred	19	Female	English	Yes	No	Yes	3.88	Missouri

As y'all may have noticed, these values tend to range and build up at certain points. The GPA for perhaps a longer list may end up having most if not all values between a 2.5 and 4.0. There is a bit of variation in the data, and this pattern of variation determines our distribution:

Definition:
The **population** is the set of all measurements of interest to the experimenter.

This means EVERYBODY… we could possibly measure ☺. Think of it like all of the possible people that have taken the ACT last year. We could measure ALL those people… but that would take forever. So, instead we might only take 100 people at random and measure those instead, which follows our last major definition:

Definition:
A **sample** is a subset of cases selected from the population of interest. It will always be less than the population (since we sample within that pool).

We use statistics for most types of **empirical research** (research based on experience / observation). A lot of y'all will be using statistics in whatever you do (no joke)!

One major concept that is used with statistics is the use of **sampling**. We use sampling when we take a chunk out of our **population** (the entire group of people that fit our research design) to try to make a smaller measure that can estimate characteristics of the population. This chunk is called the **sample**.

Descriptive statistics are the procedures used to summarize and describe the important characteristics of a set of measurements. A few of these statistics include:

• Proportion	• Range	• Mean	• Median	• Mode

Each of these people/places/things/stuff within a sample is called **observations**, and are denoted by a certain variable, often we call it simply ***x***.

➔ The **proportion** a portion or part in its relation to the whole

$$Proportion = \left(\frac{\#\ of\ those\ who\ fit\ the\ measurement}{\#\ total\ in\ the\ pop/sample}\right).$$

If we end up dealing with **quantitative variables**, then we can examine a few other measures that help explain our data ☺.

➔ The **range** is the set of that is defined as the difference between the largest (our *maximum*) and the smallest (our *minimum*) number in the data.

➔ The **mean** or **average** is the set of *n* measurements added up and divided by *n.*

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum x}{n}$$

Note: The Σ symbol (*sigma*) next to the x means "sum", which tells us to add them up.

➔ The **median** of a data set containing n measurements, denoted M, is the middle number when the measurements are arranged in order from the smallest value to the largest value.

Notes:
When looking for the median:

- Order the data from smallest to largest.
- (a) if n is odd, then the median is the middle number.
  (b) if n is even, then the median is the average of the two middle numbers.

Example:
Consider this scenario: the cost of televisions exhibits huge variation—from $100 to $200 for a standard TV to $8,000 to $10,000 or more for a large plasma screen TV. *Consumer Reports* gives the prices for the top 10 LCD high definition TVs (HDTVs) in the 30- to 40-inch category. The data for these televisions are on the next page. Find the mean price of these 10 HDTVs.

Brand	**Price ($)**
JVC LT-40FH96	2900
Sony Bravia KDL-V32XBR1	1800
Sony Bravia KDL-V40XBR1	2600
Toshiba 37HLX95	3000
Sharp Aquos LC-32DA5U	1300
Sony Bravia KLV-S32A10	1500
Panasonic Viera TC-32LX50	1350
JVC LT-37X776	2000
LG 37LP1D	2200
Samsung LN-R328W	1200

[[2]] Find the mean and median price of the 10 HDTVs listed above.

Note: The median is less sensitive to extremely large (or small) measurements than the mean. For this reason, we say the median is a **resistant measure** of center, while the mean is not. The mean gets "pulled" in the direction of the possible outliers.

- To check our what we mean, consider an example of test scores: 20 80 90 95 95

➔ The **mode** is the measurement of the category or value that occurs most frequently.
  - If 2 or more numbers (but less than all of them) occur at the same frequency, we list all numbers that occur the most often.
  - If all numbers in a data set occur at the same frequency, we say there is **no mode.**

Example:
[[3]] Find the mode (or modes) for the following set of data:

19	20	24	21	16	7	19	14	20	23	21	20	23

Example:
[[4]] The following is a set of data of the number of hours of sleep a BSF student receives during a Monday night. *(n=10)*

7	6	8	5	3	9	8	4	6	8

So, with this data, what is our range?
What is our mode?
What is our median?
What is our mean?

Most studies suggest that you receive 7 hours of sleep per night. What is the proportion of students who sleep less than seven hours during a Monday night?