Statistics is the science of
i) Summarizing / describing / visualizing data
ii) Using collected data to draw inference about unknown outcomes
iii) Exploring the mathematical foundations which support the inferences we draw.

Anyone who collects data will have to use statistics at some point.
In this class, we go over the basic foundations and implementation of common statistical techniques.

Vocabulary

A subject: A person, place, or thing from which we collect data

A population: The collection of all subjects of interest. Usually unmeasurable.

A sample: A subset of the population, from which we have collected data.

A simple random sample (we refer to this as a random sample in this class): A sample where each subject is chosen randomly, with an equally likely chance of being selected.

A variable: A characteristic of a subject which varies in a non-random way.
Ex: Months vary, but in a known (non-random) way

A random variable: A variable whose outcome

is the result of some random process. Most things are random variables.

Types of Variables

In statistics, different types of variables are analyzed in different ways. The four main types are:

I) Categorical random variables: Variables which are naturally measured as labels. There are two types:
1) Nominal ~ Labels with no natural ordering/rank.
2) Ordinal ~ Labels with a natural ordering/rank

II) Numeric random variables - Variables that are naturally recorded/collected as numbers. They also have two types:
1) Continuous - Numbers which can (theoretically) take on any value in a certain interval.
2) Discrete - Numbers with natural gaps.

Ex: Us shoe size has natural gaps of 0.5, so is Discrete. Letter grades have a natural ranking, so is Ordinal. Shape of leaf has no natural ranking, so is Nominal. # of people has natural gaps, so is Discrete.

If I were to rank what types of variables are least difficult to most difficult, it would be: Continuous, Discrete, Nominal, Ordinal.

* I often will shorten random variable to r.v.

Why do we need Statistics?

Usually we draw a random sample, but want to make inferences about an unknown population. However, since every sample is random, any specific

numeric values / graphs we find from the sample is also random. This idea is called Sampling Variation.

This would make drawing conclusions difficult, which is why we use statistis. Statistics tryS to measure this random fluxuation, and use it and some assumptions to draw reasonable conclusions about the population, using a single sample.

Notationally, Y denotes all possible values of a r.v, and y denotes a specific value of Y that was measured.

Ex: My commute time in minutes home today is Y, a r.v. Once I'm home, I would write $y = 45$ mins since I have now collected that datum.

Similarly, Y_i is the i^{th} value of the r.v Y (and still itself random), and y_i is the i^{th} observed value of Y_i.

Summary Statistics (numeric data)

Statistics is the name of an academic discipline, but also a numerical value that summarizes a dataset in some way. The first statistics we will discuss are summarizing the center and spread.

Measures of Center

1) The sample mean, average, or expected value.

Let our sample be made of data points $y_1, y_2, \dots, y_n$ (n total data). The sample mean is denoted $\bar{y}$, and is:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}(y_1 + y_2 + \cdots + y_n) = y_1/n + y_2/n + \cdots + y_n/n$$

$$= (\text{sum of all data})/n.$$

Notice the mean uses $1/n^{Th}$ of all data to create a "typical value", which means every piece of data is weighted equally.

2) Median / Percentiles / Quartiles

The k^{th} percentile of a dataset is the value for which k% of the data lies below it, and (100-k)% of the data lies above it (for k between 0 to 100).

To calculate the k^{th} percentile (denoted $y^{(k)}$):

1) Order the data from smallest to largest

2) Calculate $(\frac{k}{100})(n+1)$ [n = sample size]

2) Use "the rounding rule": If 2) resulted in a whole number, the k^{th} percentile is the ordered number in the $[(k/100)(n+1)]^{Th}$ location.

If 2) resulted in a decimal, round $(\frac{k}{100})(n+1)$ up and down, and average the ordered numbers in those two locations.

The median is denoted $\tilde{y}$, and is the 50^{th} percentile. It is the literal center of the data.

Notice the median uses locations, and is calculated

based off of a maximum of two values (unlike $\bar{y}$).

Quartiles split the data into quarters, and
$Q_1 = 25^{th}$ percentile = first quartile
$Q_2 = 50^{th}$ percentile = second quartile = median
$Q_3 = 75^{th}$ percentile = third quartile

Ex: The survival time of patients with severe chronic heart disease (in months) was:
5, 7, 10, 15, 16, 17, 19, 20, 29, 32, 35 $(n = 11)$

a) Calculate the mean

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{11}(5+7+10+15+\cdots+35) = \frac{1}{11}(205) = 18.636$$

b) Calculate the median

$\tilde{y}$ is the 50^{th} percentile, so
1) Data is already ordered (yay)
2) $50/100(n+1) = 0.50(12) = 6^{th}$
3) 6 is a whole #, so $\tilde{y} = 6^{th}$ value $= 17$

c) Calculate Q_1, Q_2, Q_3, and the 90^{th} percentile
$Q_1 = 25^{th}$ percentile, $0.25(n+1) = 0.25(12) = 3^{rd}$ value
so $Q_1 = 10$
$Q_2 = \tilde{y} = 17$.
$Q_3 = 75^{th}$ percentile, $0.75(n+1) = 0.75(12) = 9^{th}$ value,
so $Q_3 = 29$
For $y^{(90)}$, $0.90(12) = 10.8^{th}$ value, so average 10^{th} and 11^{th}:
$y^{(90)} = (32+35)/2 = 33.5$

* Outliers

Outliers are unusually small or large observations. There are multiple definitions, but a common one uses a "boxplots" definition.

Boxplot outlier: Any observation that is

i) larger than $Q_3 + 1.5(Q_3 - Q_1)$ (upper cutoff)
ii) Smaller than $Q_1 - 1.5(Q_3 - Q_1)$ (lower cutoff)

Ex: From our previous example,
lower cutoff $= Q_1 - 1.5(Q_3 - Q_1) = 10 - 1.5(29-10) = -18.5$
upper cutoff $= Q_3 + 1.5(Q_3 - Q_1) = 29 + 1.5(29-10) = 57.5$

So there are no outliers.

✳ Measures of Spread (numeric data)
We may also be interested in the spread of the data - either overall or from its center.

1) The range of a data set: This is simply the maximum difference in values of the dataset.
I.e: range $= \max\{y_i\} - \min\{y_i\}$
= (maximum of dataset) - (minimum of dataset)

2) Sample Variance
The variance is the "typical squared deviation from the mean", and denoted s^2.
A deviation from the mean is: $(y_i - \bar{y})$, so a squared deviation is $(y_i - \bar{y})^2$. The total squared deviations are $\sum_{i=1}^{n}(y_i - \bar{y})^2$, and the variance is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right]$$

Fact: We don't use simply $\frac{1}{n-1}\Sigma(y_i - \bar{y})$ because $\Sigma(y_i - \bar{y}) = 0$ (this is also why $\bar{y}$ is considered the "center" of the data).

Fact: The denominator is $(n-1)$ because then s^2 more closely estimates the overall population variance.

Most people do not interpret the variance, since its units are (units of y)^2

3) Sample Standard Deviation, denoted s

$s = \sqrt{s^2}$ (deviation)

Which is interpreted as "The typical distance of a data point to its mean"

Ex: Continuing our example:

d) Calculate s and interpret it.

$\Sigma(y_i^2) = 5^2 + 7^2 + 10^2 + \cdots + 35^2 = 4795$

$s^2 = \frac{1}{11-1}(4795 - (11)[18.636)^2) = 97.469$

So

$s = \sqrt{97.469} = 9.873$

A typical deviation of survival time in months from the mean is 9.873