Data Basics

Variables & Cases

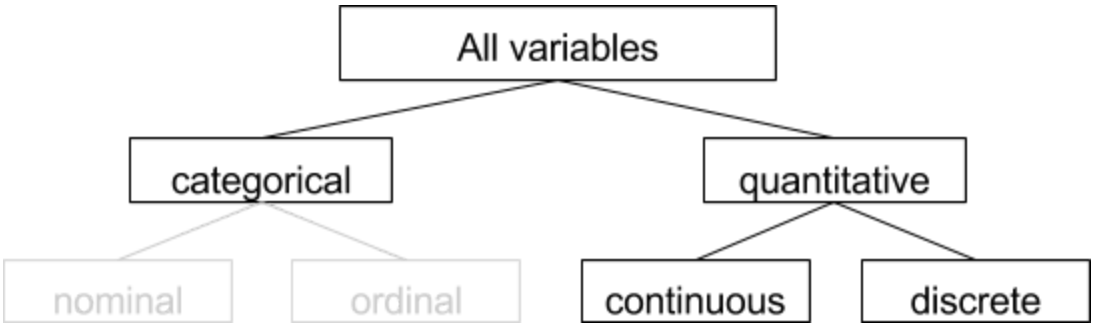
When conducting a study, information is collected regarding **cases** or **observational units**. **Variables** are measured characteristics of the units that can take on different values.

Example: In a hypothetical statistics class, students have the option of attending a tutorial before their term test to help them revise concepts. They are also provided with 10 homework questions to help them prepare but these are not marked. All students must attempt the quiz and write the term test. Data on the following **variables** were collected for each student (**case**):
Whether or not they attended the tutorial
How many homework questions were attempted
How many hours of sleep did they get before the test
What was their quiz mark
What was their test mark

Student #	Attended Tutorial	Homework Questions	Hours of sleep	Quiz Mark %	Test Mark %
1	Yes	none	5	78	70
2	No	some	7	82	86
3	No	all	6	78	79
4	Yes	some	4	70	72

In the table above, each row is called an **observation**.

We notice that not all the information collected for each variable in the table is the same i.e. some variables are numeric, others are not. Variables can be classified into different categories as follows:



Categorical variables: take on one of a limited and usually fixed number of values e.g. the variable “Homework Questions” where the values are: “some”, “none”, or “all”.
Quantitative variables: numerical variables e.g. the variables “Quiz Mark %”, “Test Mark %” and “Hours of Sleep”.
Quantitative variables can be continuous or discrete.
Continuous quantitative variables: can take on any value from between and including the minimum and maximum values eg. “Test Mark %” could be 93.3, 80.2, 75.0 etc.
Discrete quantitative variables: can take values from a specific set of numbers of that variable e.g. you flip a coin thrice and count the number of tails - the number you get must be an integer value i.e. 1, 2, or 3. You cannot get 2.4 heads - hence the number of tails is a discrete variable.

Response and Explanatory Variables

Suppose we wanted to investigate how the number of hours of sleep students got before the test affected their test mark - in this case, ‘hours of sleep’ would be our **explanatory variable** and ‘test mark’ would be our **response variable**. The explanatory and response variables are also known as “independent” and “dependent” variables respectively.

Samples and Populations

Let's now suppose that we wanted to investigate how the number of hours sleep before a test affected the test marks of every student at a university, or even better - every student around the world! It wouldn't be practical (or possible) to get data from each and every student so instead we would have to take a **sample** of students and make an *inference* about the entire **population** of students.

Population: The entire set of possible observations in which we are interested.
Sample: A subset of the population from which information is actually collected.

Describing Data

Five Number Summary

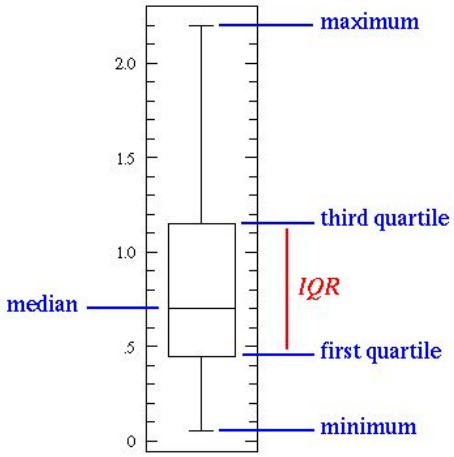
In order for us to make sense of the data we have, we have to be able to summarize it. This initial summary should give you a slight sense of what you expect your data to look like. It's not an in-depth analysis, but it's good to have an initial idea about your data. Note that 'data' here could mean any quantitative variable in your data set e.g. "Number of hours of sleep" or "Test Mark". (Summarizing categorical variables is a different process.)

Quantitative variables can be summarized using the "Five Number Summary":

- Minimum: the smallest observation
- Maximum: the largest observation
- Median: observation that is halfway from smallest to largest observation, when data are arranged in ascending order
- First Quartile: point that is a quarter of the way, when data are arranged in increasing order
- Third Quartile: point that is three quarters of the way, when data are arranged in increasing order

The Range is the difference between the maximum and minimum values.

The Interquartile Range (IQR) is simply the difference between the Third and First Quartiles. The five number summary above allows us to make a **boxplot**.



source: http://www.physics.csbsju.edu/stats/simple.box.defs.gif

Mean, Median, Mode

Oftentimes, you will need more than just the Five Number Summary to summarize your data.
Mean: Suppose we wanted to know what the average test mark was for the class. For this we would have to calculate the mean.
The mean is calculated as the sum of all the data values divided by the number of values. We use $\bar{x}$ ("x bar") to represent the mean.

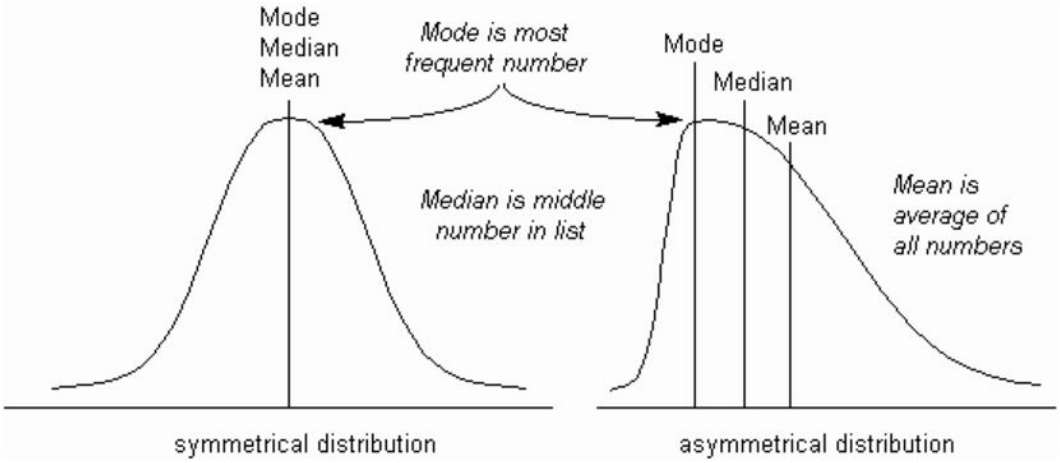
$$\bar{x} = \frac{\sum x}{N}$$

$$\sum x = \text{ the sum of } x$$

$$N = \text{ number of data}$$

Median: As before, the median is the middle value of a distribution that has been ordered from smallest to largest; for distributions with an even number of values, this is the mean of the two middle values.

Mode: The most frequently occurring value in the distribution e.g. if majority of students got 6 hours of sleep, the mode would be 6 (but the mode is NOT the same thing as the mean!)



Source: https://s3-eu-west-1.amazonaws.com/tutor2u-media/subjects/geography/studynoteimages/mean-median-mode.png

The figure above shows that if we have a symmetrical distribution of our data, the mean, median and mode could take on the same value. But we could also have an asymmetrical distribution which would mean that the mean, median and mode would be different values.

Variance & Standard Deviation

More often than not, five number summaries and measures of centre (mean, median, mode) are not enough to answer questions we have about our data. We must know the variability of the data set - how much does the data diverge or vary from the mean? For example, if 75% is the class average , it is obvious that not every student got 75%. Well, did most students get less than 75% or more than 75%? This is where variance and standard deviation come in use. If Student 1 got 78% on the test, their deviation from the mean is 78-75 = 3. Similarly, Student 2's deviation from the mean is 86-75 = 11. We can continue with this calculation for all students, and calculate the **standard deviation** which is *roughly* the average between individual data and the mean. The standard deviation is denoted by s and tells us how much on average individual scores vary from the mean. The formula for standard deviation is:

$$s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

n = The number of data points

$\bar{x}$ = The mean of the x_i

x_i = Each of the values of the data

(You'll notice that the formula is not exactly the average of the data points, but you won't have to worry about proofs in this context)

If we take the square of the standard deviation, we get the **variance** which is important in concepts such as covariance (which we won't be covering).

Confidence Intervals

Earlier, we talked about how, if we wanted to gain information about an entire population, we would have to take a sample to make inferences about the entire population of our study. (While we didn't mention this before, in order for a sample to be generalizable to the population, we must make sure that it is collected randomly, without any bias - but sampling methods are a bit of a different, longer story).
Values concerning a sample are referred to as **sample statistics** while values concerning a population are referred to as **population parameters**.

Statistic: A measure concerning a sample (e.g., sample mean) - this can be calculated
Parameter: A measure concerning a population (e.g., population mean) - this is fixed and unknown (remember the whole point of taking a sample is so that we can make inferences about the population and estimate its parameters)

Here are some parameters and statistics:

	Population parameter	Sample Statistic
Mean	μ	$\bar{x}$
Standard Deviation	σ	s
Proportions	p	$\hat{p}$

A good way to estimate the population parameter is by constructing confidence intervals. Suppose we wanted to investigate how many hours of sleep students were getting on average, during midterm season. In this case, our population parameter is μ (the mean), which we know nothing about. We would take a sample of 50 or 100 (the more the better) students and administer our study. We would take the mean of these 100 values and that would give us our

sample statistic $\bar{x}$ (the mean of the sample). In this case, the sample mean would be called the **point estimate** of the population mean.

Point estimate: A point estimate of a population parameter is a single value used to estimate the population parameter.

Let's say that for this sample, we got 5 hours as the mean. So now could we say that all students around the world get an average of 5 hours of sleep before a test? We probably shouldn't.
If we gave a range of values instead, we'd have a higher chance of estimating the population parameter correctly.
Think about it like this - which statement is more believable: "I can confidently say that, based on my sample of 100 students, the average number of hours of sleep for all students around the world is 5 hours" VS. "I can confidently say that, based on my sample of 100 students, the average number of hours of sleep for all students around the world is between 4.5 and 5.5 hours."? (hopefully the second statement is more believable).

In statistics, we have to be able to quantify our confidence - can we report the results of our study with 70% confidence or 99% confidence - it could be that we're only 1% confident! This is known as the **confidence level**. So if the 95% confidence interval for the population mean is 4.5 to 5.5, we would say that we are "95% confident that the average number of hours of sleep for all students around the world is between 4.5 and 6 hours" - this can be written as $4.5 < \mu < 5.5$, $[4.5, 5.5]$, although the general form of a confidence interval is:

sample statistic ± margin of error

The margin of error will depend on two factors: (1) the level of confidence ;and (2) the value of the standard error. (We won't go into this in great detail because we doubt you'd have to calculate standard errors and margin of errors in the context of this challenge - it is probably better to just understand the concept of confidence intervals). For clarity, if the standard error for our first sample was calculated to be 0.25, our confidence interval would be

$5 \pm 2(0.25) \rightarrow 5 \pm 0.5$, where 0.5 is the margin of error. This will change as the confidence level changes. A 99% confidence interval will be narrower than a 95% confidence interval.

Now, a sample of just 100 students is not going to give us an accurate estimate of a population of millions of students. In order for our results to be reliable and accurate, we have to replicate the study, multiple times, each time using a different sample. So now we take another sample of 100 students and this time we get an average of 6.5 hours of sleep. This presents us with a problem. Is 5 or 6.5 hours a better estimate of the population mean? So we see that point estimates differ from one sample to the other. Point estimates also include error. We can construct a confidence interval for each sample and then we can be fairly certain that the actual parameter lies somewhere within all those values. Hopefully this also explains why having a bigger sample and replication is important.

Hypothesis Testing

Hypothesis testing is a more formal method of estimating a population parameter and is one that you will probably encounter multiple times (too many times?) as you progress through your education and career.
To start, we need to hypothesize a population parameter and compare our sample statistic with this hypothesized parameter (again, we will need to quantify this comparison - we will see how to do that in the later sections). For example, we are very optimistic and hypothesize that the mean number of hours of sleep for all students is 8. Now we need to test this hypothesis by comparing this value to what we observe in our sample.

Hypotheses - Null and Alternative

The first step in hypothesis testing is formally writing out our hypotheses. This helps us keep track of what we're measuring, what we expect the outcome to be etc. For each test, we need to have the **Null hypothesis (H_0)** and the **Alternative hypothesis (H_1 or H_A)**.

The null hypothesis says that we would observe no difference between the population parameter and the sample statistic and that any deviation in the sample statistic is due to random error/by chance.

The alternative hypothesis says that the population and sample means would be different, and this difference could not be just due to error- the deviation (of the observed sample statistic from the expected, hypothesized value) is too large to be explained by chance alone.

Continuing our sleep example, our hypotheses would be:

H_0: There is no difference between the mean number of hours of sleep of the population and the sample i.e. $\mu = \mu_0$ (where μ_0 is null value, our hypothesized value) $\rightarrow \mu = 8$
H_A: There is a difference between the mean number of hours of sleep of the population and the sample i.e. $\mu \neq 8.$

From our subsequent testing, we will either reject or not reject the null hypothesis. (Because we are only hypothesizing, remember that we cannot *prove* a hypothesis - we can only obtain evidence to support or not support it, so even if we reject the null hypothesis, we will never *prove* the alternative hypothesis).

P-values and Statistical Significance

When conducting a hypothesis, we always start by assuming that the null hypothesis is true - that there is no difference between our expected and observed values. The **p-value** is the probability of getting our observed value given that the null hypothesis is true i.e. how likely are our data, given that the null hypothesis is true. This can be difficult to understand by definition, but might be better understood with an example. Let's say that we carried out our study on a sample of 100 students, and observed a mean of 5 hours of sleep (as before). But we had hypothesized that the mean would be around 8 hours. So, if we assume that the true population mean is 8 hours, what's the chance that we got 5 hours as our sample mean for this one sample? This chance or probability is called the p-value.
A low p-value would basically tell us that our data are unlikely to be observed if the null is true i.e. if the true mean is really 8 hours, then it's very unlikely to observe a sample with a mean of 5 hours - something might be wrong with our hypothesis - the null is falsifiable and we have evidence to support the alternative hypothesis. Conversely, a high p-value would tell us that our data are likely to be observed if the null is true i.e. if the true mean is really 8 hours, we are likely

to observe a sample with a mean of 5 hours - the null is not falsifiable, and we don't have evidence to support the alternative hypothesis.

But how do we decide which p-value is low enough or high enough to make our decision about the null hypothesis? We use a threshold value called the **significance level** which is typically set to 0.05 and denoted by α (alpha).
If the p-value is less than the significance level we say that the result is statistically significant. We reject H_0, and we have strong evidence favoring H_A.
If the p-value is greater than the significance level, we say that the result is not statistically significant. We do not reject H_0, and we do not have strong evidence for H_A.