Session 08: Experiments and hypothesis testing

wifi: GA-Guest, yellowpencil

GENERAL ASSEMBLY

```
cd ~/Documents/ga-ldn-ds37
git commit -am "your commit message here"
git pull
```

Today's session plan

1800-1830	Standup & Review
1830-1900	Linear algebra review
1900-1920	Break
1920-2000	Interpreting correlations
2000-2030	Handling missing data
2030-2100	Sampling methods and Type I/II errors

At the end of the session, you will be able to ...

Calculate covariance and correlation by hand and using numpy

Identify and handle different types of missing data

Understand different sampling methods

Identify Type I and Type II errors

Data Science Part Time

Review

Computers Out:
Review

Open the notebook ds37-08-01.ipynb and work through the exercises.

Data Science Part Time

Covariance

What's covariance?

Covariance is a measure of the **joint variability** of two random variables. We write the covariance between the variables **X** and **Y** as:

$$cov(X, Y) = \frac{\sum (x_i - \bar{X})(y_i - \bar{Y})}{n}$$

Group Exercise:

Covariance

Think about the numerator of this formula.

Under what **two** conditions will the numerator be **strongly positive**?

Under what **two** conditions will the numerator be **strongly negative**?

$$cov(X, Y) = \frac{\sum (x_i - \bar{X})(y_i - \bar{Y})}{n}$$

What's covariance?

Covariance is **strongly positive** when all values of (x_i-Xbar)(y_i-ybar) are positive. This happens when:

(x_i-Xbar) **and** (y_i-ybar) are positive, i.e. both values of x and y are greater than their respective means

(x_i-Xbar) **and** (y_i-ybar) are negative, i.e. both values of x and y are less than their respective means

$$cov(X, Y) = \frac{\sum (x_i - \bar{X})(y_i - \bar{Y})}{n}$$

What's covariance?

Covariance is **strongly negative** when all values of (x_i-Xbar)(y_i-ybar) are negative. This happens when:

(x_i-Xbar) is positive **and** (y_i-ybar) is negative.

(x_i-Xbar) is negative **and** (y_i-ybar) is positive.

$$cov(X, Y) = \frac{\sum (x_i - \bar{X})(y_i - \bar{Y})}{n}$$

Solo Exercise:

Calculating covariance

By hand, calculate the covariance between:

x = [2, 4, 6, 8, 10], y = [1, 2, 3, 4, 5]

...And the covariance between

x = [2, 4, 6, 8, 10], y = [10, 9, 8, 8, 6]

Before you calculate the covariances, make a prediction about whether the covariance will be strongly positive or negative.

You can do this by sketching out a plot of y vs x on paper.

Solo Exercise:

Covariance and variance

On paper, show that the covariance of a random variable X **with itself** is the same as the variance of X.

Hint: what's the formula for variance?

$$cov(X, X) = var(X)$$

Computers Out:

Covariance in Numpy

Let's open up ds37-08-01.ipynb

Problems with covariance

Covariance is a measure of whether two random variables behave in similar ways. But it's hard to compare the covariance of one set of variables with another, because it's not a **standardised** measure.

This is where **correlation** comes in!

Data Science Part Time

Correlation

Correlation

Correlation (or the correlation coefficient) is a standardised version of covariance. It shows us the same thing (i.e. whether there's a strong association between two variables) but it can only take values from -1 to 1.

This makes it much easier to compare the strength of association between different pairs of variables.

$$corr(X, Y) = \frac{cov(X, Y)}{\sigma_X \sigma_Y}$$

Solo Exercise:

Calculating correlation

By hand, calculate the correlation between:

x = [2, 4, 6, 8, 10], y = [1, 2, 3, 4, 5]

...And the covariance between

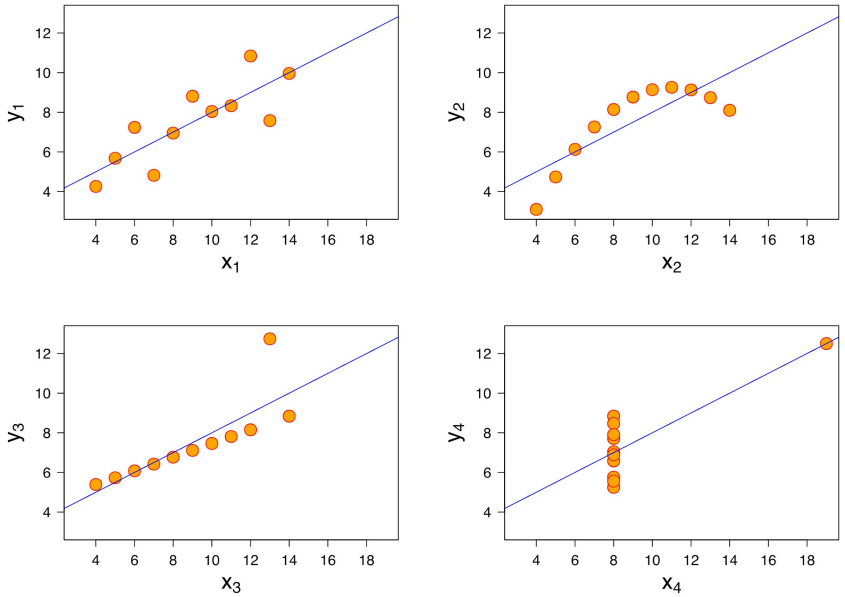
x = [2, 4, 6, 8, 10], y = [10, 9, 8, 8, 6]

Before you calculate the correlations, make a prediction about whether the correlation will be strongly positive or negative.

Visual sense checks are still important!

The four datasets in Anscombe's quartet all have the same correlation coefficient, despite looking very different when they're visualised.

It's important not to make assumptions about a dataset **solely** by looking at correlation coefficients.



The correlation matrix

In most data science tasks, we'll calculate correlations to:

- Understand which of our features are most strongly correlated with our target
- Which of our features are strongly correlated with each other

This helps give us an initial idea of which independent variables will be most useful in our modelling task.

How can we calculate the correlations between **all** possible pairs of variables in a large dataset?

The correlation matrix

The correlation matrix shows us the correlation coefficient between all pairs of variables in a dataframe.
Given **n** number of features from $\mathbf{X_1}$ to $\mathbf{X_n}$ the correlation matrix is:

$$\begin{bmatrix} var(X_1)/\sigma(X_1)^2 & cov(X_1, X_2)/\sigma(X_1)\sigma(X_2) & \dots & cov(X_1, X_n)/\sigma(X_1)\sigma(X_n) \\ cov(X_2, X_1)/\sigma(X_2)\sigma(X_1) & var(X_2)/\sigma(X_2)^2 & \dots & cov(X_2, X_n)\sigma(X_2)\sigma(X_n) \\ \dots & \dots & \dots & \dots \\ cov(X_n, X_1)/\sigma(X_n)\sigma(X_1) & cov(X_n, X_2)/\sigma(X_n)\sigma(X_2) & \dots & var(X_n)/\sigma(X_n)^2 \end{bmatrix}$$

The correlation coefficient between variable 2 and variable 1

The correlation coefficient between variable n and variable n

Data Science Part Time

Causation and correlation

Causation and correlation

Take a look at some of these examples of Spurious Correlations:
http://www.tylervigen.com/spurious-correlations

The variables are **correlated** but may or may not be **causal**.

Understanding this difference is critical for executing the data science workflow, especially when identifying and acquiring data.

Be careful not to say “caused” when you really mean “is associated with.”

Is a relationship causal?

One attempt that's commonly used in the medical field is based on work by Bradford Hill. He developed a list of “tests” that an analysis must pass in order to indicate a causal relationship:

Strength of association (effect size) A small association does not mean that there is not a causal effect, although the larger the association, the more likely the effect is to be causal.

Consistency (reproducibility) Consistent findings observed by different persons in different places with different samples strengthens the likelihood of an effect.

Specificity Causation is likely if there is a very specific population at a specific site and a disease with no other likely explanation. The more specific an association between a factor and an effect, the greater the probability of a causal relationship.

Temporality The effect has to occur after the cause

Biological gradient Greater exposure should generally lead to greater incidence of the effect. However, in some cases, the mere presence of the factor can trigger the effect. In other cases, an inverse proportion is observed: greater exposure leads to lower incidence.

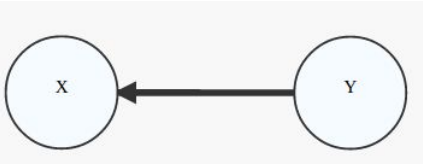
Plausibility A plausible mechanism between cause and effect is helpful (Hill noted that knowledge of the mechanism is limited by current knowledge).

Coherence Coherence between epidemiological and laboratory findings increases the likelihood of an effect.

Is a relationship causal?

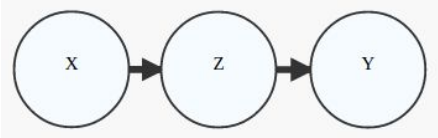
Let's imagine two variables, X and Y, are correlated. What are some of the causal structures that might produce this correlation?

X causes Y

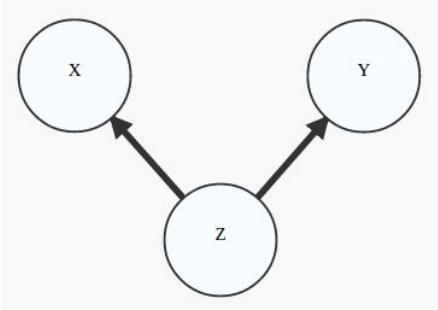


Y causes X

Correlation not significant



X causes Y indirectly through a third variable

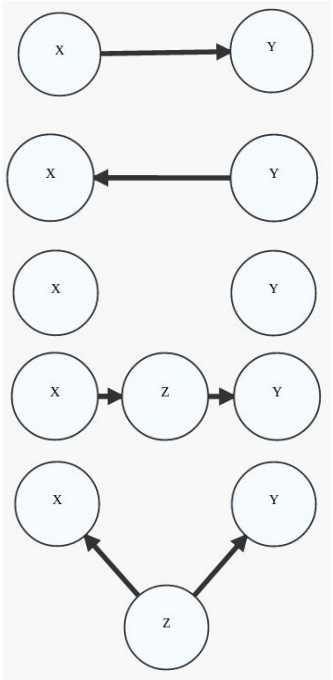


A third common factor (or **confounding variable**) causes both X and Y

Solo Exercise:

Is a relationship causal?

Match the text to the appropriate DAG (directed acyclic graph) that represents the likely causal structure:



Lung cancer is correlated with carrying a lighter

The period after 9/11 is correlated with cooler temperatures

Divorce rates are correlated with cheese consumption

Smoking is correlated with lung cancer

Sleeping more than 8hrs a night is correlated with risk of stroke*

STILL 5p CHEAPER THAN THE DAILY MAIL AND TEN TIMES BETTER!

DAILY EXPRESS

THE WORLD'S GREATEST NEWSPAPER express.co.uk

WEATHER: RAIN

THURSDAY FEBRUARY 26, 2015 55p

NEW AGONY FOR CLIFF AS POLICE EXPAND SEX ABUSE INQUIRY

SEE PAGE 7

VICTORIA CROSS FOR HERO WHO TOOK ON TWENTY TALIBAN KILLERS

SEE PAGE 5

TOO MUCH SLEEP COULD KILL YOU

More than 8 hours a night can double the risk of stroke

By David Pilditch

SLEEPING more than eight hours a night could dramatically increase the risk of suffering a stroke, scientists warned yesterday.

Researchers from Cambridge University found that people who regularly slept more than eight hours were twice as likely to suffer a stroke compared with average sleepers.

And those who went from sleeping less than six hours a night to more than eight hours were four times as likely to suffer the life-threatening condition, where the blood supply is cut off to part of the brain.

The scientists behind the new research say the results of their major study of 10,000 people could save the NHS millions of pounds every year.

Doctors have regularly extolled the virtues of a good night's sleep to recharge the batteries.

Previous studies have found that too little sleep may contribute to coronary heart disease

TURN TO PAGE 4

How the Palace blocked a BBC tribute to Diana

SEE PAGE 3

Data Science Part Time

Sampling bias

Sampling

When we're conducting a study or experiment, the aim will usually be to make conclusions about a **population** based on a **sample** or **subset** of that population.

For example:

Understanding the relationship between students' coffee consumption and grades, based on a sample of 100 students.

Modelling house price vs postcode using a sample of 1000 recently sold houses.

How do we select our **sample** in such a way that we can draw reliable conclusions about the whole **population**?

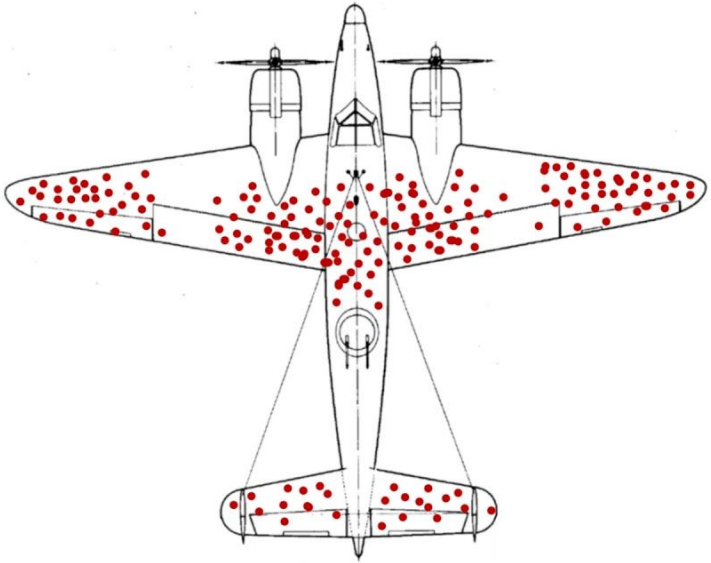
Group Exercise:

Case study

During WWII, the US Centre for Naval Analyses studied damage to aircraft returning from missions.

The visualisation shows the points where these aircraft had been damaged.

Based on this visualisation, where would you recommend extra reinforcements should be added to the body of an aircraft?



Sampling bias

Sampling bias occurs when a sample is collected in such a way that some members of the intended population are more or less likely to be included than others.

This can happen when a sample is taken non-randomly — either implicitly or explicitly.

When we have non-random sampling that results in sampling bias, it can affect the inferences or results of our analyses. We must be sure not to attribute our results to the process we observe when they could actually be because of non-random sampling.

When we have sampling bias, we aren't measuring what we think we are measuring.

Causes of sampling bias

Pre-screening: Purposely restricting the sample to a specific group or region.

This typically happens when people try to study priority areas to save costs and assume priority areas are the same as random areas.

Self-selection: When someone has the ability to non-randomly decide what is included in a sample.

This typically happens in surveys and polls but can also be an issue with other kinds of reporting.

Survivorship bias: When we select only surviving subjects in a sample over time.

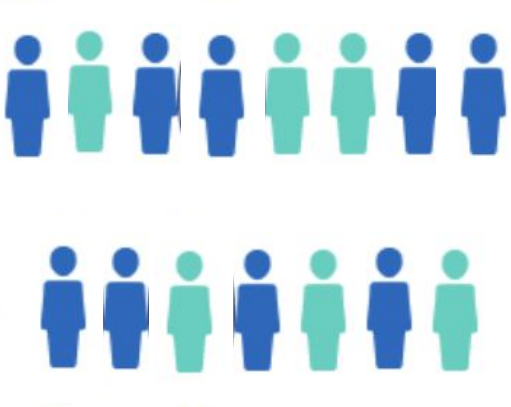
This might happen when we only look at existing customers and assume they have the same characteristics as new customers.

Group Exercise:

Types of sampling

We want to construct a **representative** sample from this population, i.e. the proportion of the two groups in the sample should be roughly the same as in the population as a whole (around 6:9 light blue:dark blue)

How would you devise a **random** sampling method to construct your sample?

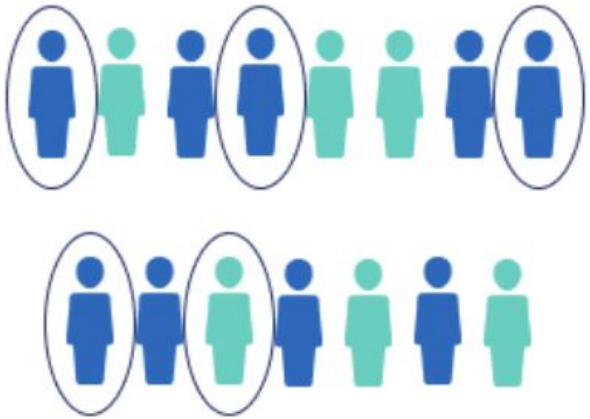


Types of sampling

Simple random sample

Every member of the population has an equal chance of being selected. Your sampling frame should include the whole population.

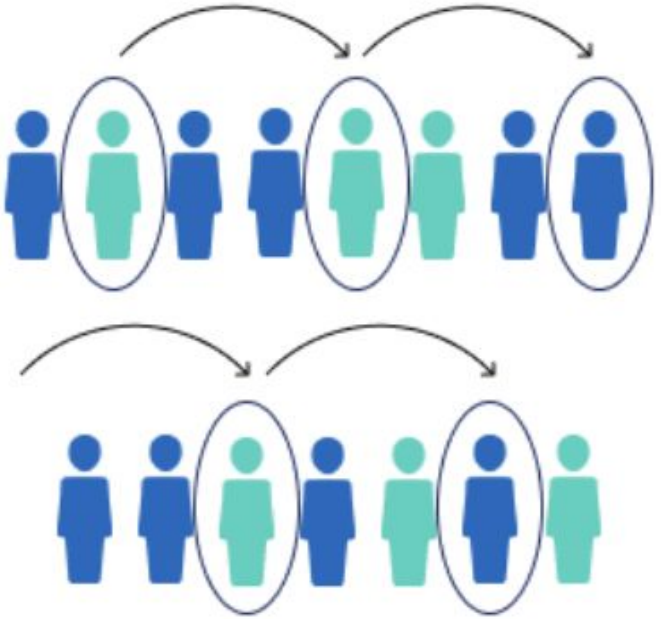
To conduct this type of sampling, you can use tools like random number generators or other techniques that are based entirely on chance.



Types of sampling

Systematic sample

Every member of the population is listed with a number, but instead of randomly generating numbers, individuals are chosen at regular intervals.

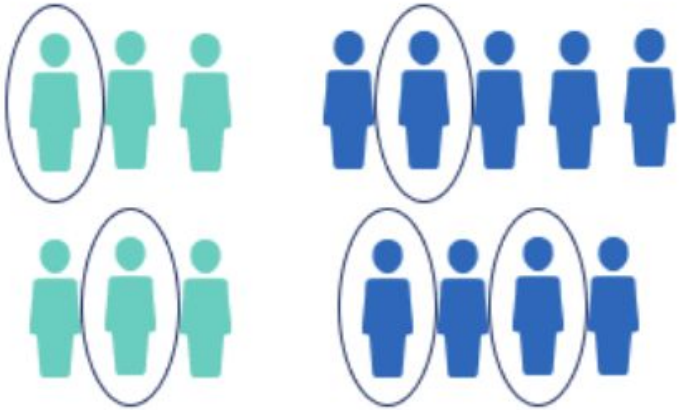


Types of sampling

Stratified Sampling

This sampling method is appropriate when the population has mixed characteristics, and you want to ensure that every characteristic is proportionally represented in the sample.

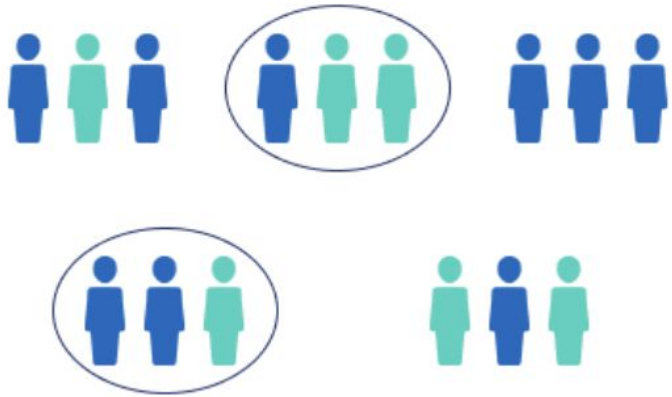
You divide the population into subgroups (called strata) based on the relevant characteristic (e.g. gender, age range, income bracket, job role).
From the overall proportions of the population, you calculate how many people should be sampled from each subgroup. Then you use random or systematic sampling to select a sample from each subgroup.



Types of sampling

Cluster sampling

Cluster sampling also involves dividing the population into subgroups, but each subgroup should have similar characteristics to the whole sample. Instead of sampling individuals from each subgroup, you randomly select entire subgroups.



Intro to Python

Missing data

Missing data

Sometimes, missing data in our dataset will introduce sampling bias. There are three main **types** of missing data:

Missing completely at random (MCAR)

The reason that the data are missing is completely random and introduces no sampling bias. In this case, it's safe to drop or impute rows with missing values. We can test for this by looking at other attributes for missing and non-missing groups to see if they match.

Missing at random (MAR)

The data are missing in a way that is related to one of the independent variables. This introduces sampling bias. Like other instances of sampling bias, we can fix this by modeling the selection process. This is done by building a model to impute the missing value based on other variables.

Missing not at random (MNAR)

The data are missing in a way that is related to the dependent variable. We can't test for this. We also can't fix this in a reasonable way.

Computers Out:
Missing data

We're modelling IQ as a function of age. What types of missing data are present in the examples below?

Complete data	
Age	IQ score
25	133
26	121
29	91
30	105
30	110
31	98
44	118
46	93
48	141
51	104
51	116
54	97

Incomplete data	
Age	IQ score
25	133
26	121
29	
30	
30	110
31	
44	118
46	
48	141
51	
51	116
54	

Incomplete data	
Age	IQ score
25	
26	
29	
30	
30	
31	
44	118
46	93
48	141
51	104
51	116
54	97

Incomplete data	
Age	IQ score
25	
26	121
29	91
30	
30	110
31	
44	118
46	93
48	
51	
51	116
54	

Intro to Python

Error types

Class imbalance

Sometimes a sample may include an overrepresentation of one type of class. E.g. airport security may have 990 X-ray scans showing the absence of a weapon. Due to natural scarcity, it may only provide 10 scans showing a weapon.

If our goal is to create a model that indicates whether or not a weapon is present, then we are at a disadvantage. **Ignoring** the class imbalance would lead to a model that always guesses that a weapon is not present!

A simple way to get around this is to **undersample** the majority class, deliberately leaving us with a balanced data set of 10 each. However, this is less than ideal, as it effectively ignores much of the available data.

Alternatively, we could **oversample** the minority class by duplicating examples. Again, this is not ideal. Because we have very little data, this will magnify small differences that may just be errors, leading to a model that overfits.

Solo Exercise:

Types of errors

Imagine we built the airport security system in the previous slide. It **always** predicts that a weapon is not present.

When shown 100 passengers, 1 of whom is carrying a weapon:

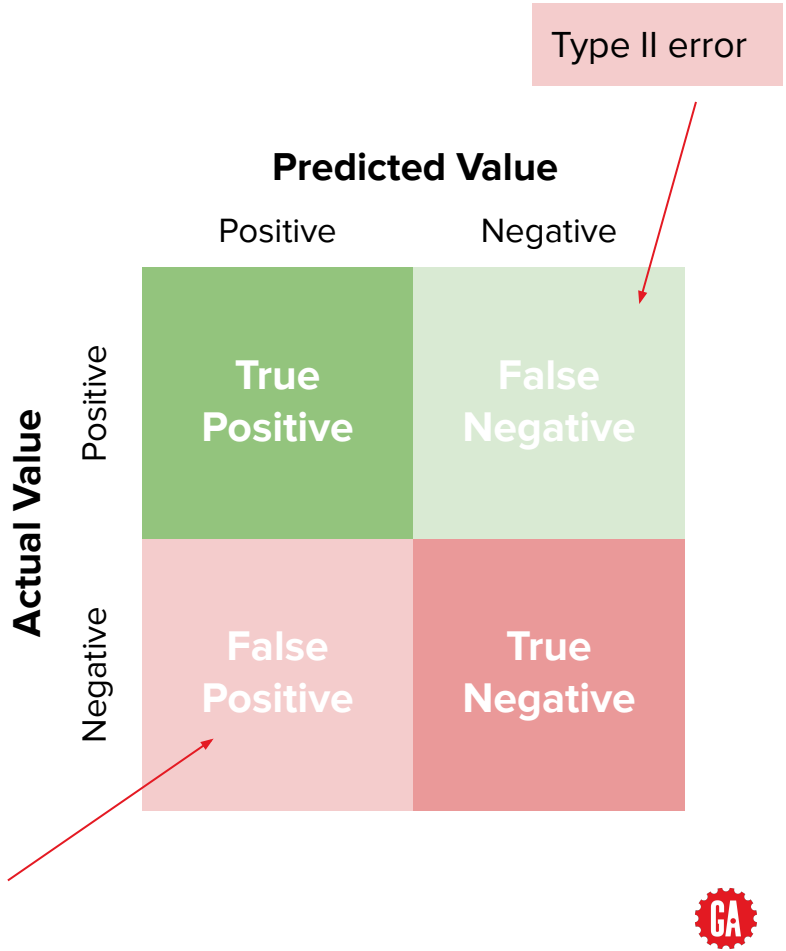
- What is the overall accuracy of the system
- What % of the time will it guess a weapon is not present when there's no weapon present?
- What % of the time will it guess a weapon is present when there's a weapon present?
- What % of the time will it guess a weapon is present when there's no weapon present?
- What % of the time will it guess no weapon is present when there's a weapon present?

Which of these metrics is the most important in this context?

How Good is Your Model?

For supervised models, it's important to be able to measure the **accuracy** of our model in predicting a value (for regression tasks) or classifying a data point (for classification tasks).

For **classifiers** there are two types of error, which we visualise using a **confusion matrix.**



Group Exercise:

Which Metric Matters?

10 minutes

For each example below:

1. Write a confusion matrix: true positive, false positive, true negative, false negative. Then decide what each square represents for that specific example.
2. Define the benefit of a true positive and true negative and the cost of a false positive and false negative.

Examples:

1. A test is developed for determining if a patient has cancer or not.
2. A newspaper company is targeting a marketing campaign for "at risk" users that may stop paying for the product soon.
3. You build a spam classifier for your email system.

Intro to Python

Let's Review

At the end of the session, you will be able to ...

Calculate covariance and correlation by hand and using numpy

Identify and handle different types of missing data

Understand different sampling methods

Identify Type I and Type II errors

Coming up next week...

- Natural language processing and linear regression


GA