MATLAB

Class 5: statistics basics

Statistics aims to understand your data by describing it and making predictions

Descriptive stats aims to describe data's distribution by **central tendency** (location in a plane) and **dispersion** (spread) around the location

Inferential stats aims to predict future outcomes or observations based on your data

Useful descriptive stats and associated functions

Correlations

Correlations is the relationship between any two variables

They can be described in different ways:

Pearson—linear relationship between continuous variables

Spearman Rho—nonparametric rank correlation, describing two variables as a monotonic function

Correlations in MATLAB

corr — returns a matrix of pairwise correlations between columns

corrcoef — Returns the correlation between vectorized matrices

corr2 — returns correlation coefficient for matrices (i.e., one value for its 2-d inputs)

Reminder: Reshape can help you

Reshape is a useful function to transform any sized matrix into a different shape

Reshape(X, [new dimensions])

Note that the new dimensions need to be consistent with the previous ones

i.e., dim1*dim2*dim3 == newdim1*newdim2 etc..

T-tests

Student t-test allows you to test mean differences between normal distributions

There are many different **flavours** of t's

- one-sample vs two samples
- paired vs unpaired
- one tail vs two

T-test

T-tests assume that your data come from a **normal** distribution and the observations are sampled **independently** from one another

These assumptions apply for both paired and unpaired test

T-tests

One-sample t-tests test the hypothesis that mean is different than a prespecified u0

Independent-sample t-tests test the hypothesis that the mean difference between both samples is not 0

T-tests

Paired t-tests are used when the observations are repeated, i.e., each person is sampled twice thus each observation comes in a pair

Unpaired t-tests are used when the observations are independent

T-tests

One tail tests are used when you have **directional** hypotheses (i.e., group A is bigger than group B)

Two tail tests are used when you have **non-directional** hypotheses (i.e., group A is different than group B, but you don't care if it's bigger or smaller)

T-tests

In MATLAB there are two functions for the student t-test

ttest() is used for one-sample and paired tests while **ttest2()** is used for independent sample (i.e., two-sample) tests

Ttest()

One sample and paired t-tests

Ttest(x) runs a one-sampled t-test against zero

Ttest(x, y) paired t-test

'Alpha'

'Tail' can specify 'left', 'right', 'both'

Ttest2()

Ttest2() runs a two independent samples t-test.

Has the same specifiers as ttest() including:

‘Dim’ to specify a dimension along which to run the test

and ‘Vartype’ for ‘equal’ and ’unequal’ variances

T-tests

Given the same number of data which type of t-test is more stringent? Which one requires a bigger mean diff for the same value of t?

T-tests

Let us assume we collect data from 20 people (paired)

and observe a mean diff of 3.75, and a sd of 1

Then the t value for a paired test would be:

$$t = \frac{mean\ diff}{\frac{sd}{\sqrt{n}}} \text{ with df n} - 1$$

T-tests

Let us assume we collect data from 20 people (paired)

and observe a mean diff of 3.75, and a sd of 1

Then the t value for a paired test would be:

$$t = \frac{3.75}{\frac{1}{\sqrt{20}}} = \text{ with df } 19$$

T-tests

Let us assume we collect data from 20 people (independent groups) and observe a mean diff of 3.75, and a sd of 1 (assuming equal variance)

Then the t value for an unpaired test would be:

$$t = \frac{mean\ diff}{sp * \sqrt{\frac{1}{n1} + \frac{1}{n2}}} \text{ with sp} = \sqrt{\frac{(n1-1) * std1^2 + (n2-1) * std2^2}{n1 + n2 - 2}} \text{ df n1 + n2}$$

T-tests

Let us assume we collect data from 20 people (independent groups) and observe a mean diff of 3.75, and a sd of 1 (assuming equal variance)

Then the t value for an unpaired test would be:

$$\mathrm{sp} = \sqrt{\frac{(20-1)*1+(20-1)*1}{20+20-2}}$$

$$t = \frac{3.75}{sp * \sqrt{\frac{1}{20}+\frac{1}{20}}}. \quad \mathrm{df}\ 38$$

T-tests

T value for paired is larger, in comparison to t values of unpaired or independent tests

This is because **within-person** designs have more **power** as they control for more noise by observing the same person twice

Non-parametric tests

In stats there are some tests that are **non-parametric,** by this statisticians mean that the test does not require assuming a specific model or distribution for your data

These are sometimes referred to as **non-distributional** tests

These tests are used when you do not want to assume a specific distribution (e.g., violation), or do not know the distribution of your data

Non-parametric t-tests

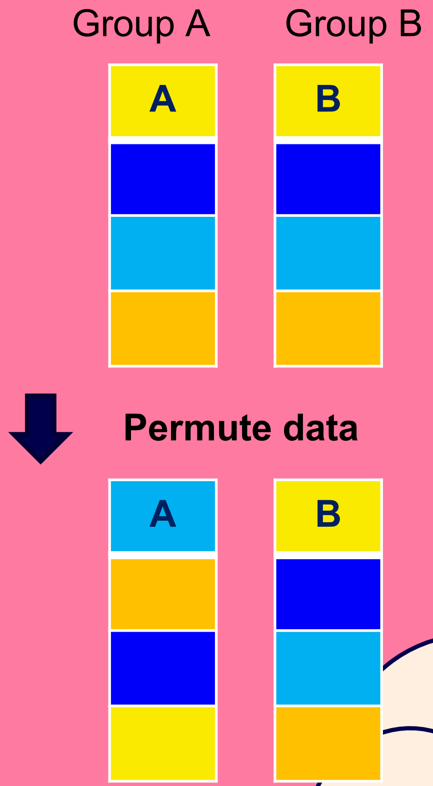
Regular t-tests assume your data is normally distributed

There is a *non-parametric equivalent of a t-test* called the **permutation t-test**. This test works on the premise that you can observe a *null distribution* from your own data by randomly permuting groups.

Reminder: a null distribution is the distribution when the *null hypothesis* is true

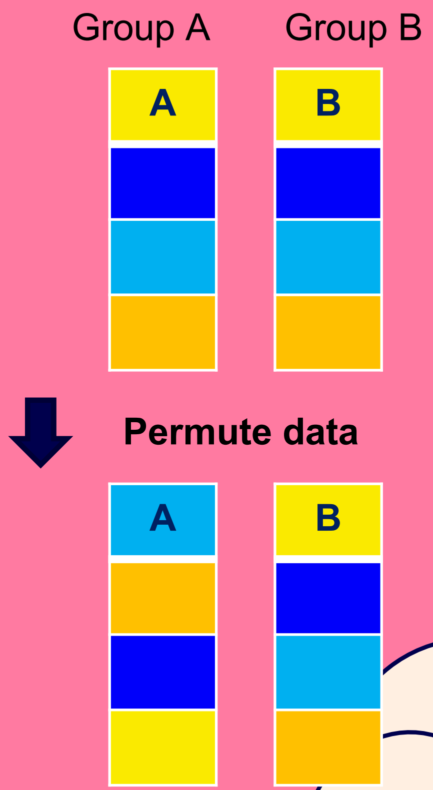
permutations

Permutations build a *null distribution* of data based on your observations under the assumption that randomly shuffling your data will void the effect of interest. Thus, you can measure how surprising the effect you observe is given your data based on the computed null distribution

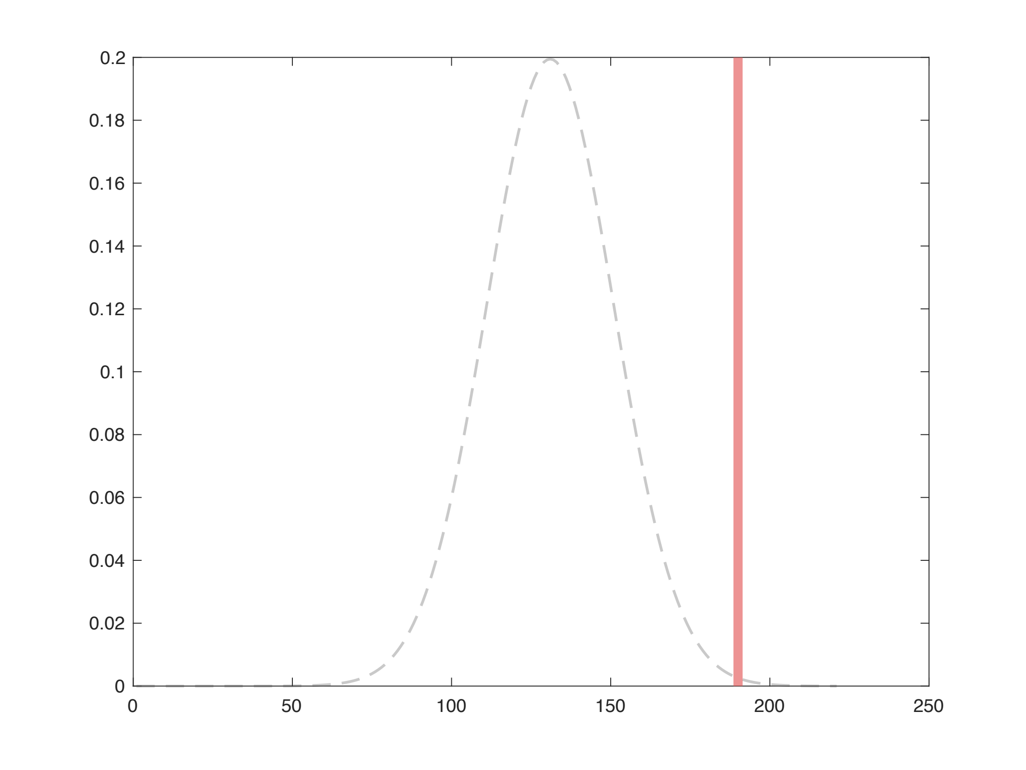


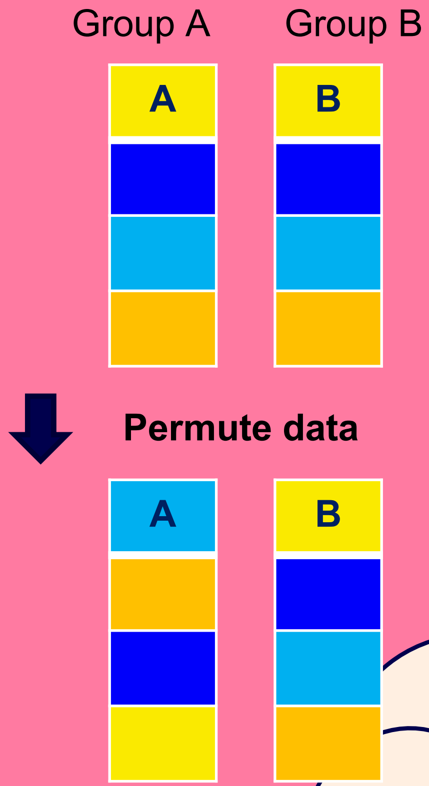
permutations

This technique is very similar to bootstrapping without replacement (i.e., no duplicate observations).



permutations





Bootstrapping

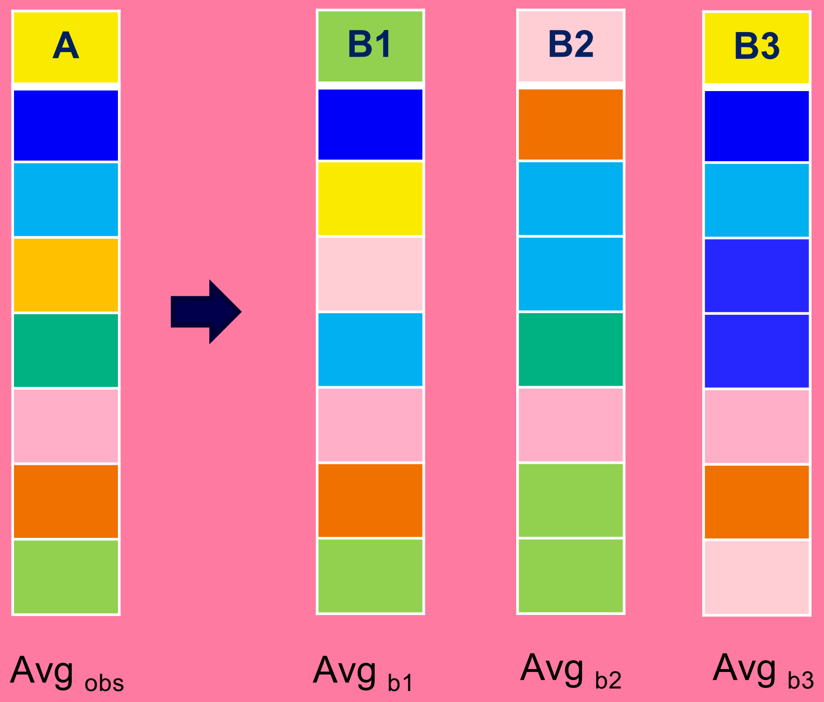
A very similar concept in statistics is the idea of **bootstrapping** to get a measure of uncertainty around an estimate (e.g., CI)

Bootstrapping is like permutations in that they resample your data, however bootstrapping requires **replacement**

It is often used to calculate the error associated to an estimate, effect, or performance of an algorithm and allows you to know if one given data point is driving the effect you see

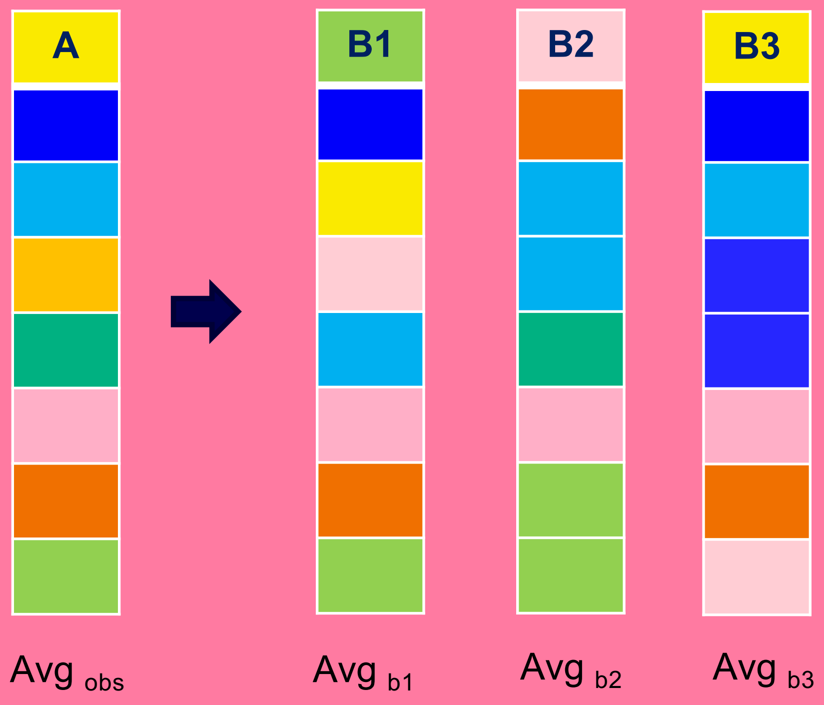
Bootstrapping

Bootstrapping is like permutations in that they resample your data, however bootstrapping requires **replacement**



Bootstrapping

The computed Avg for each **bootstrap** will allow you to get an idea of what **range** of values you could expect given your data



ANOVA

Analysis of Variance is a linear model to test if there is a difference between two or more means (i.e., groups sometimes called levels)

Let us start with a one-way ANOVA

ANOVA

One-way ANOVA (i.e., one FACTOR)

Test the difference between groups (levels). It answers the question:

Are the group means spread out much more than what you'd expect if you had sampled all the groups from the same distribution?

ANOVA

Are the group means spread out much more than what you'd expect if you had sampled all the groups from the same distribution?

$$H_0: a_1 = a_2 = a_3 = \ldots a_n \mid H_1: a_1 \neq a_2 \neq a_3 \neq \cdots a_n$$

ANOVA

Assumptions of a One-way ANOVA:

- Normality of data
- Homogeneity of variances (all variances are born equal)
- Independent samples

ANOVA

Normality of data

Can test this by visualizing your data / residuals (see regression) or by a Shapiro-Wilk test (see link)

ANOVA

Homogeneity of variances

Also called homoscedasticity can be assessed visually or with a Bartlett test (see code)

This is a more 'serious' assumption to break when running an ANOVA

ANOVA

Independence

We assume our observations are independent of one another. This may be broken when the observations are taken from the same subject (repeated measures) or when dependent variables are correlated (see regressions)

ANOVA

We test the significance of an ANOVA with an F test:

$$F = \frac{Between\ group\ variance}{Within\ group\ variance}$$

We asses each using Sum of squares (i.e., SS) weighted by the number of observations (i.e., n) in each group

ANOVA

Between-group SS

How spread apart are your group means

Within-group SS

How spread apart is each distribution

*** Note we always normalize these by their degrees of freedom

*** See code for visual depiction of an example with three means

ANOVA

How is a One-way ANOVA different from a t-test:

T-tests are used to test the difference between TWO means whereas ANOVAs test the difference between two or more means

ANOVA

How is a One-way ANOVA different from a t-test:

They both test slightly different questions: t-tests tells you if TWO means are different whereas an ANOVA will tell you if there are mean differences but NOT WHICH means are different

ANOVA

The ANOVA can be generalized to test multiple dependent variables like the Two-way ANOVA

With multiple factors / DV you can test interactions in what we call a 'crossed' model where every group A co-occurs with every group B

Crossed Two-way ANOVA

Factor / DV B	Factor / DV A		
	Group / Level 1A	Group / Level 2A	Group / Level 3A
Group / Level 1B	Mariah Halloween	Jlo Halloween	Lady Gaga Halloween
Group / Level 2B	Mariah Christmas	Jlo Christmas	Lady Gaga Christmas

ANOVA

Post-hoc testing:

After you are certain that the means differ, what next? Sometimes you need to do additional tests, and sometimes you don't. This all depends on your underlying question!!!

ANOVA

Post-hoc testing: NOT ALWAYS NECESSARY

In some cases, we just care that the means of our groups are different and not which differ from one another. When we have two groups, we know that they are different, no additional test is needed.

ANCOVA

There are many iterations of an ANOVA including the analysis of covariance!

Here we test the INDEPENDENT effect of dependent variables (factors) regardless of COVARIATES of no interest (referred to as nuisance variables)

*** we will cover this idea in more detail in the regression slides!

Signal Detection Theory

An analytic tool used to analyze data in it's ability to discriminate between two signals or signal and noise

Assumes there is an inherent uncertainty in the classification

We will look at the case where there are two classes to categorize

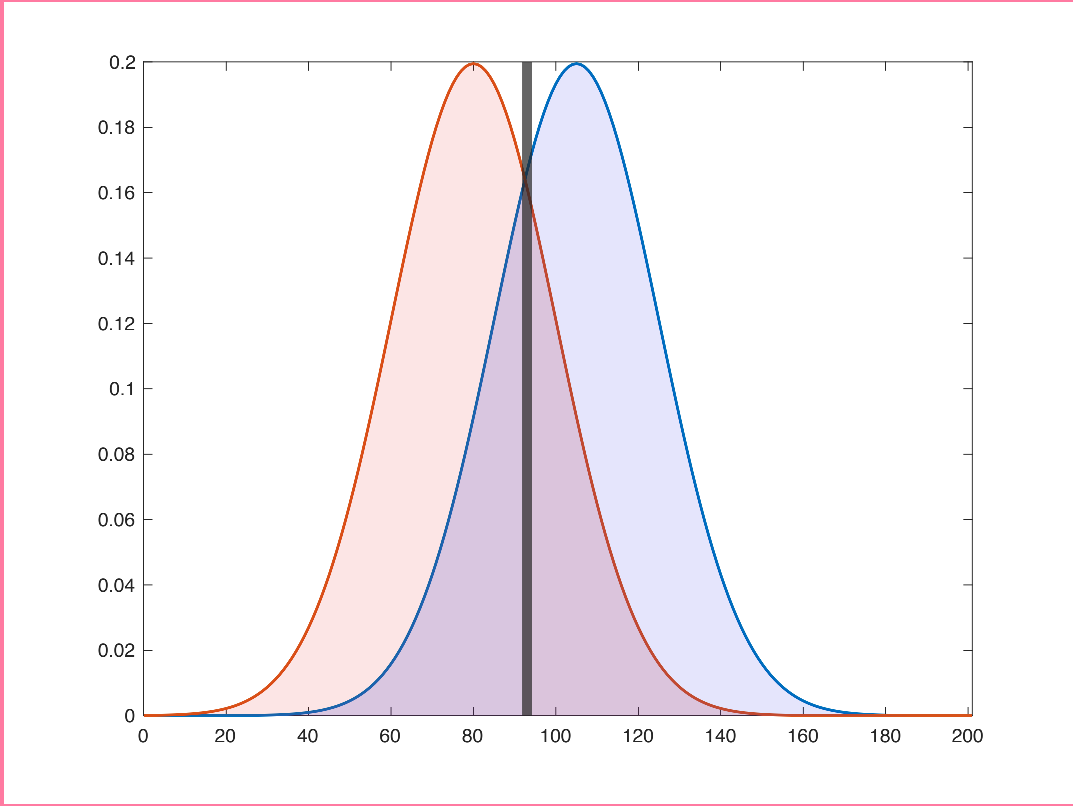
Signal Detection Theory

There are two parameters that describe signal detection theory

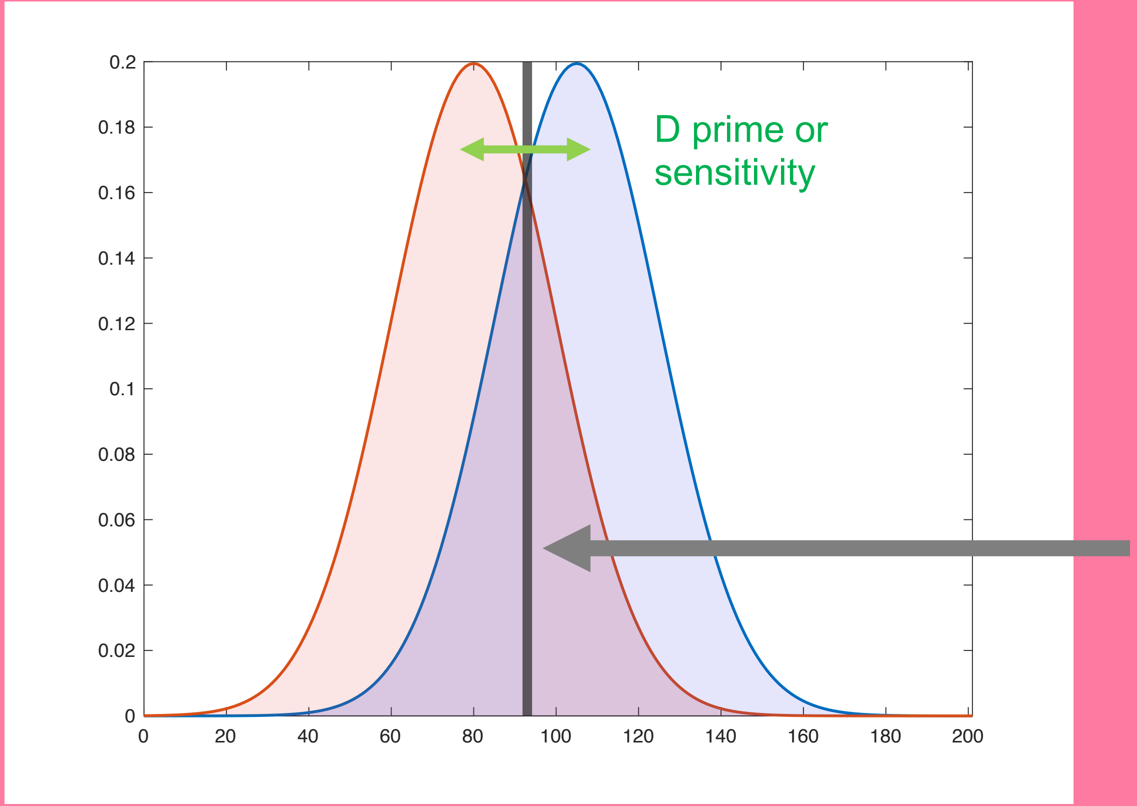
The criterion: where you draw the boundary between signal and noise

Sensitivity: one's ability to discriminate between signal and noise

Signal Detection Theory

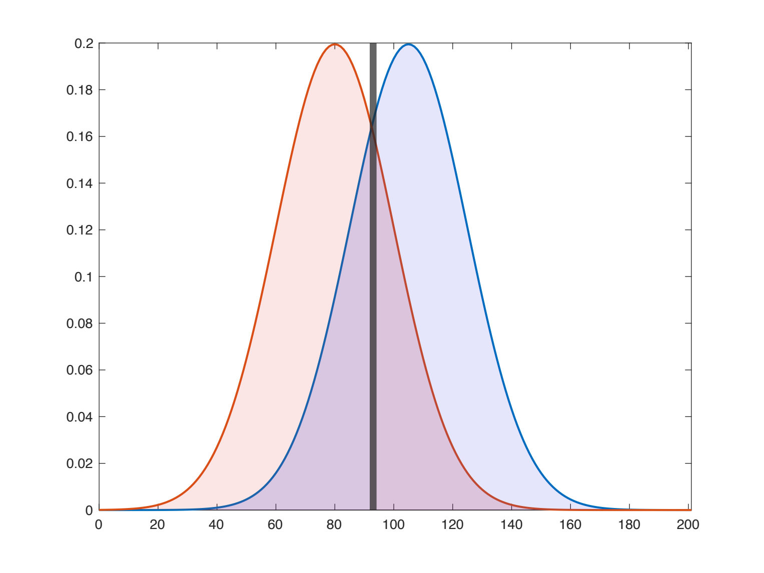


Signal Detection Theory



criterion

Signal Detection Theory



	Signal	Noise
Present	Hits	False Alarms
Absent	Misses	Correct Rejection

Signal Detection Theory

Sensitivity:

$$dprime = (z(Hits) - z(False\ Alarms))$$

Criterion:

$$c = -\frac{1}{2}(z(Hits) + z(False\ Alarms))$$

Where z() is the inverse of the cumulative normal distribution

Signal Detection Theory

- **What to do when you get values of 0 or 1 as probabilities?**
- You cannot take the inverse cumulative normal distribution of 0 or 1 as it returns infinite values.
- We therefore must apply a correction to our data

Signal Detection Theory

We assume that if we double the number of trials, by chance someone would have guessed the right answer i.e., if pHIT or pFA =0, then we correct to

$$\frac{1}{2 * Num\ Trials}$$

Signal Detection Theory

Similalry we assume that if we double the number of trials, someone would have made one mistake i.e., if pHIT or pFA =1, then correct with

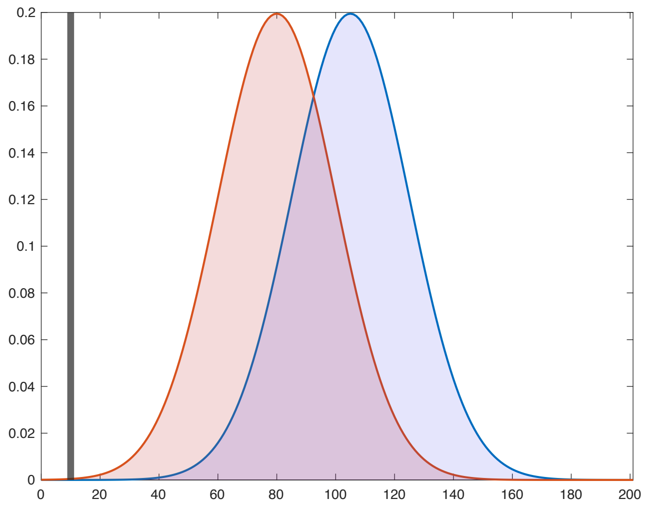
$$\frac{(2*Num\ Trials)-1}{2*Num\ Trials}$$

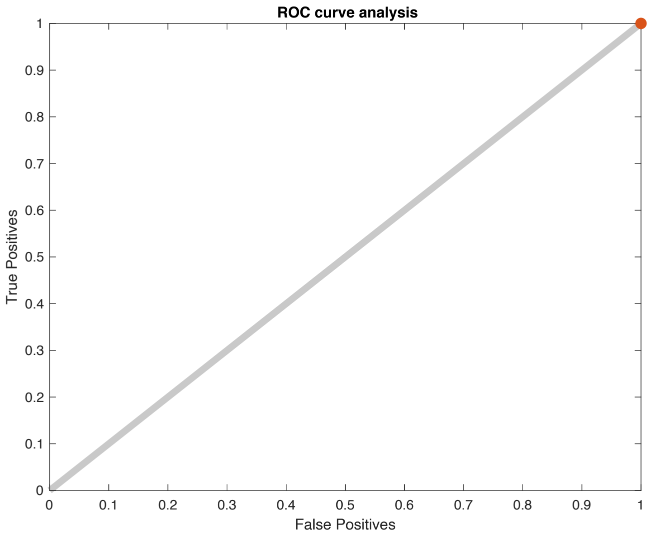
Receiver Operating Characteristic Curves

Is a diagnostic plot that helps visualize the ability of a binary classifier to separate two classes

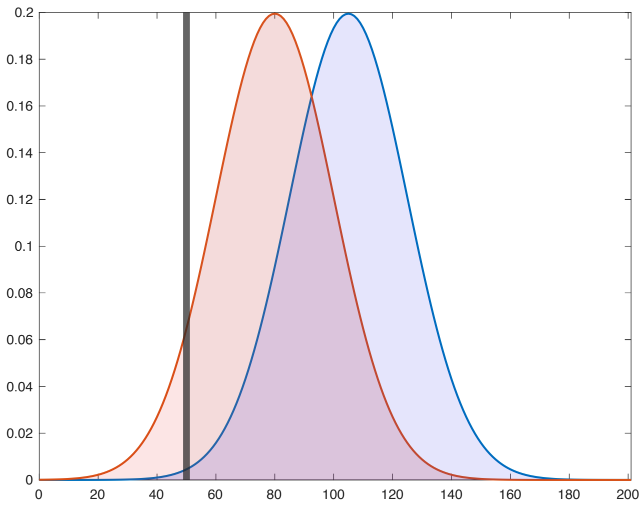
This is achieved by plotting the rate of **True Positives** against **false positives**

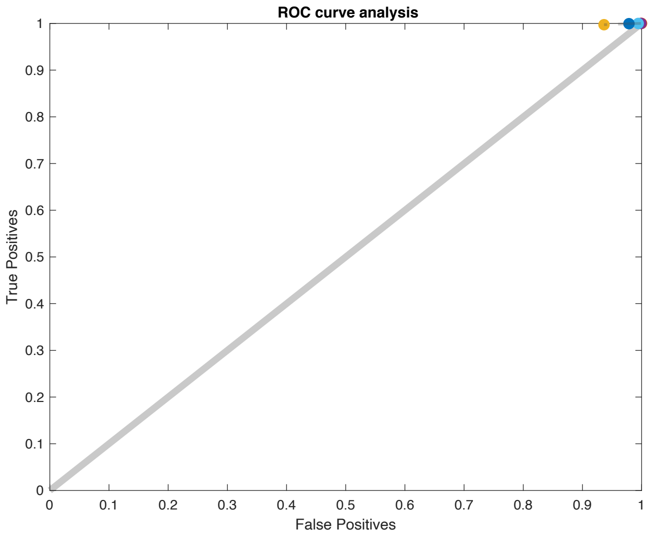
Receiver Operating Characteristic Curves



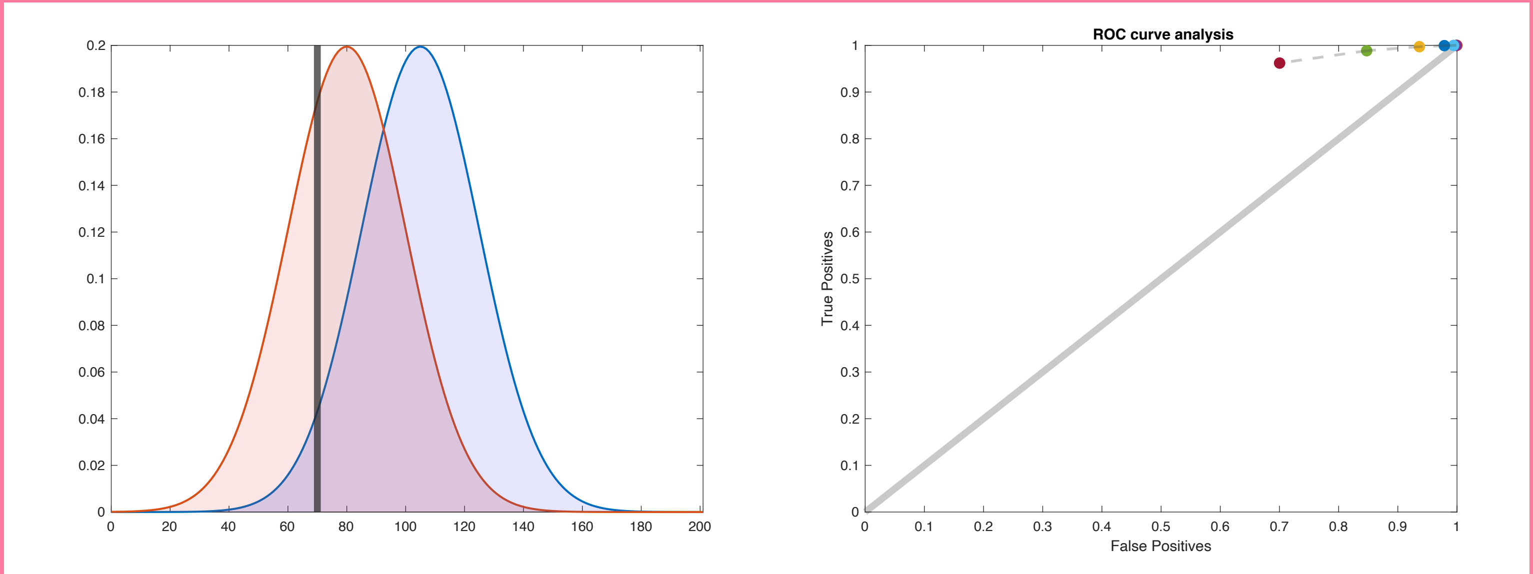


Receiver Operating Characteristic Curves

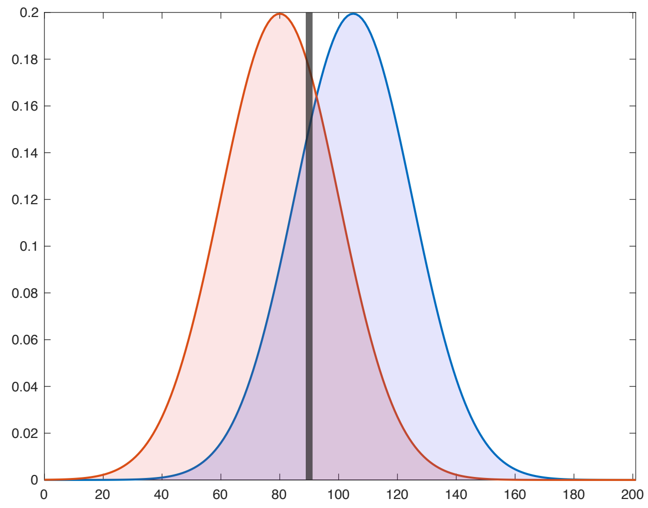


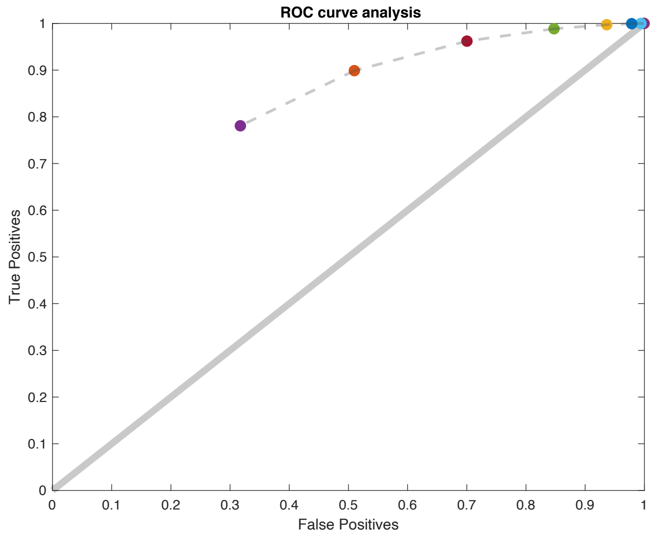


Receiver Operating Characteristic Curves

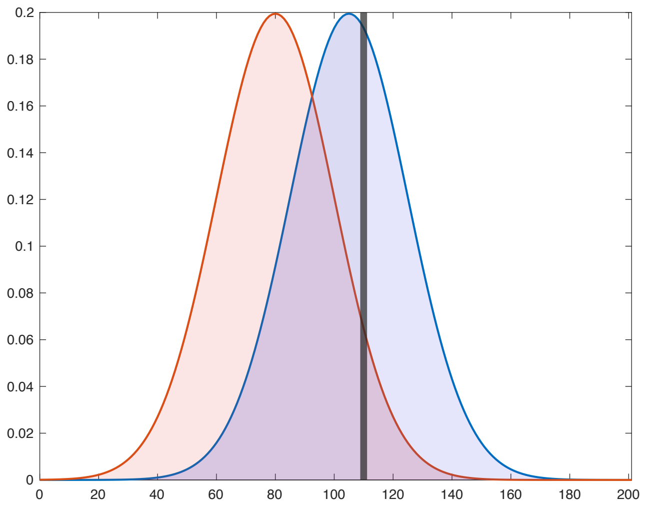


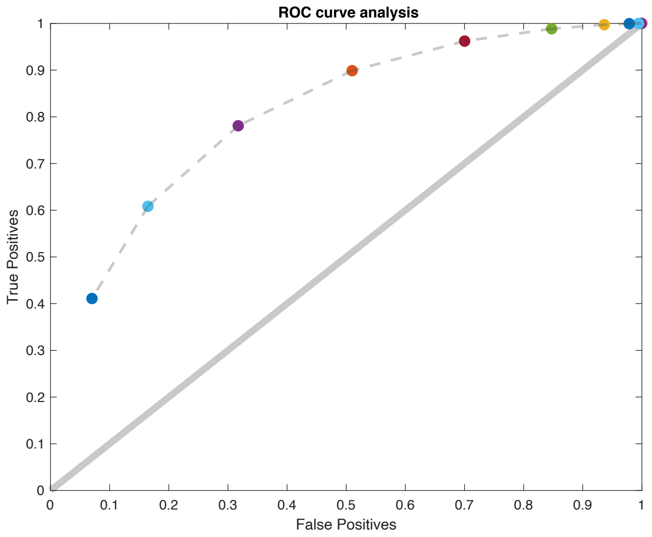
Receiver Operating Characteristic Curves



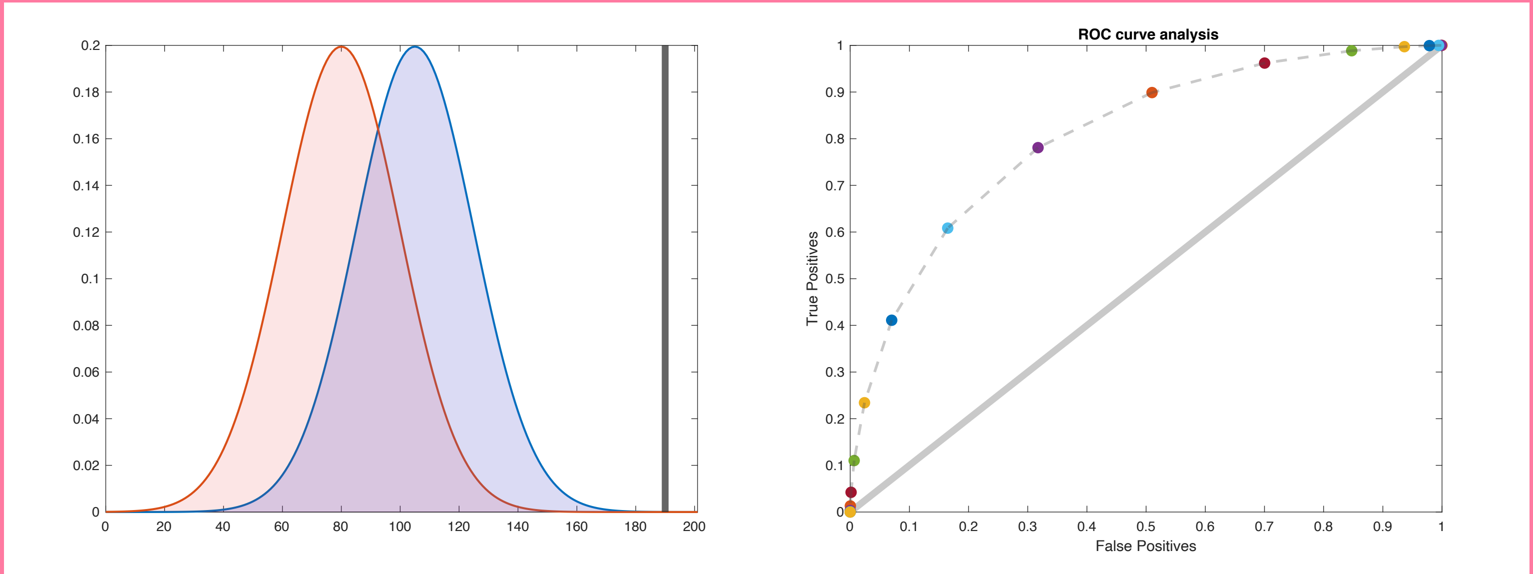


Receiver Operating Characteristic Curves





Receiver Operating Characteristic Curves

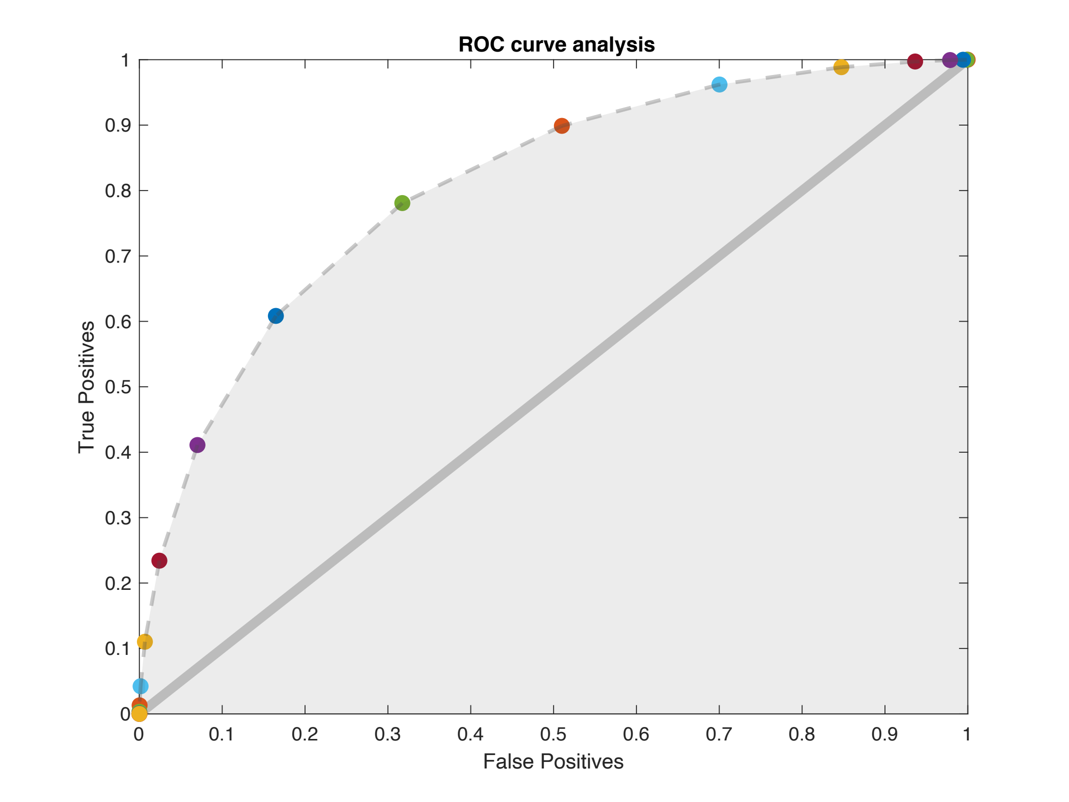


Area Under the Curve

Measures the area under the ROC curve and reflects one's ability to classify classes given a variable

AUC of 1 or 0 is perfect classification while 0.5 is chance level (i.e., along the diagonal)

Area Under the Curve



ROC in MATLAB

[X,Y,T,AUC] = perfcurve(labels, scores, posclass)

labels—the two classes to be discriminated

scores—the ‘x’ values used in discrimination

posclass—which class is larger of the two

ROC in MATLAB

[X,Y,T,AUC] = perfcurve(labels, scores, posclass)

X— x values of ROC curve

Y— y values of ROC curve

T—array of thresholds used

AUC—returns the value of AUC