Module 3E: Hypothesis Testing

Applied Epistemology: A Framework for how we know things scientifically

Agenda:

1. H_O/H_A: Our model of the test universe (the distribution of the variable)
2. Test & assumptions: are the assumptions met? Is the test valid?
3. Quantitative evidence: **p-value**, or critical value.
   - False positive =Type I (α), False Negative = Type II (β), Type III errors
   - Sensitivity, Specificity, Power → confusion matrix, ROC/AUC curve
   - Positive Predictive Power, Negative Predictive Power
   - Confusion Matrix
   - **ROC/AUC curve**
4. Conclusion & uncertainty/estimation

Errors in hypothesis testing:

Type I (α) = False Positive=0.05

P[type I]=P[rejecting Ho|Ho is true]

Type II (β) = False Negative

P[type II]=P[Fail-to-reject Ho|Ho is not true]

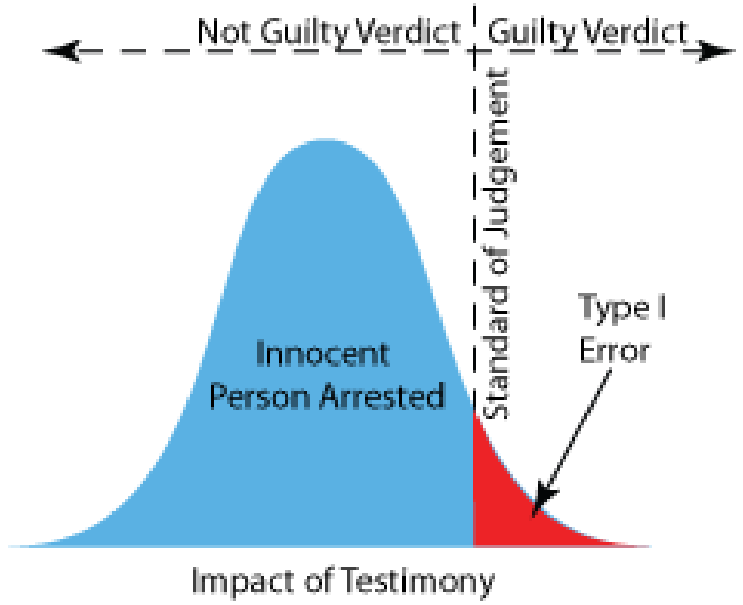
Type III*

- Less consistent definition but is usually correctly rejecting the null hypothesis for the wrong reason (i.e.. mistakenly using the wrong model).
- Right answer to the wrong problem

Type I (α) error:

Rejecting a true null hypothesis

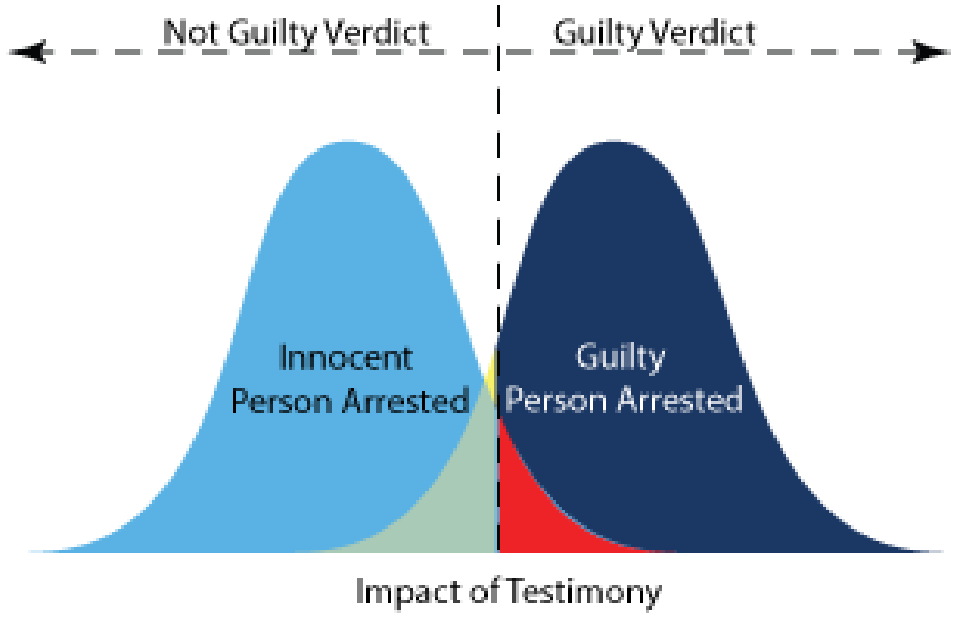
$P(\text{reject } H_0 | H_0 = \text{true}) = \alpha$



Type II (β) error:

Not rejecting a false null hypothesis

$P(\text{Fail to reject } H_0 | \text{Ho=Not True}) = \beta$



http://www.intuitor.com/statistics/T1T2Errors.html

Visualize Type I/II errors: One-sample Test of Means (Z test)

Choose Tail of the Test

- ○ One Tail, Upper Tail
- ○ One Tail, Lower Tail
- ◉ Two Tail

Choose Plot to Display

- ☑ Show Null Hypothesis Sampling Distribution
- ☑ Show Alternative Hypothesis Sampling Distribution

Choose alpha control via slider or menu

- ○ Choose among several fixed alpha levels
- ◉ Use a slider for alpha choice

Alpha, Type I Error rate

0.0005 | 0.05 | 0.15

0.0005 0.0155 0.0305 0.0455 0.0605 0.0755 0.0905 0.1055 0.1205 0.1355 0.15

Null Hypothesis Mean

100

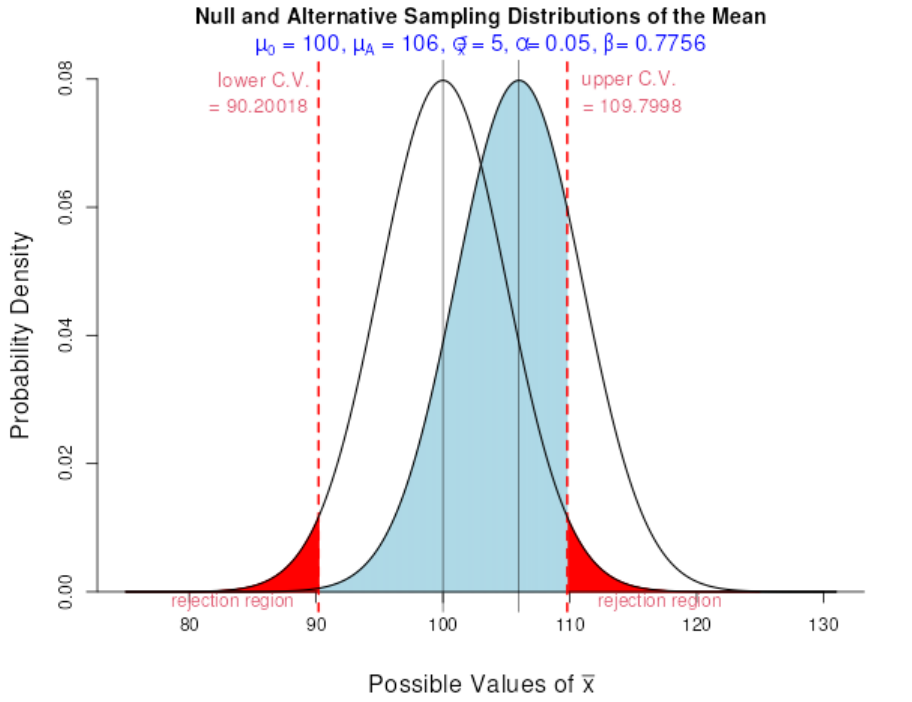
Alternative Hypothesis Mean

106

Standard Error of the Mean

5

Plot | About



Download plot as PDF

Download plot as png

https://shiny.rit.albany.edu/stat/betaprob/

	No Disease (H_0 true)	Disease (H_A true)
Fail To Reject H_0	**No Error** (specificity*)	**Type II**
Reject H_0	**Type I**	**No Error** (power, sensitivity*)

Definitions:

* (This is a rate) **Specificity** = P[FTR|Ho is True]/P[Total Ho is true] = True Negative Rate

α=type I= P[Reject|Ho is True] = False Positive

β=type II error = P[FTR|Ho is **not** True] = False Negative

Power= P[Reject|Ho is **not** True] = True Positive

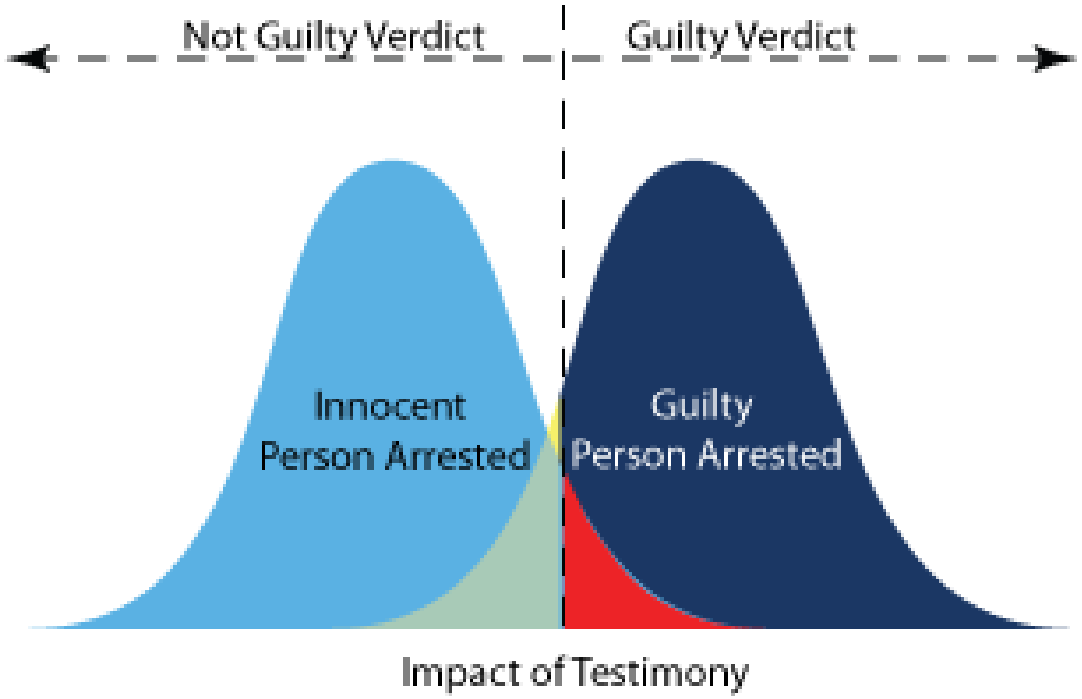
* (this is a rate) **Sensitivity** = P[Reject|Ho is **not** True]/P[Total Ho is NOT True] = True Positive Rate

Power can be increased by increasing the sample size (n)

Specificity = $\frac{TN}{TN+FN}$

	No Disease (H_0 true)	Disease (Ho is not true; H_A true)
Fail To Reject H_0	**No Error** P[FTR|Ho is true] True Negative	**Type II** P[FTR |Ho is not true] (False Negative)
Reject H_0	**Type I** P[reject|Ho] (False Positive)	**No Error** Power P[Reject|Ho is not true] (True Positive)

Sensitivity = $\frac{TP}{TP+FN}$



Generally, type I errors are the ones that we are concerned with in biology.
Although, there are circumstances when we are more concerned with type II errors (i.e.) Medicine

There is a trade-off between type I error and type II error

Scenario 1:

If you were designing a clinical trial for a life-saving but risky drug, which error would you minimize, Type I or Type II, and why?

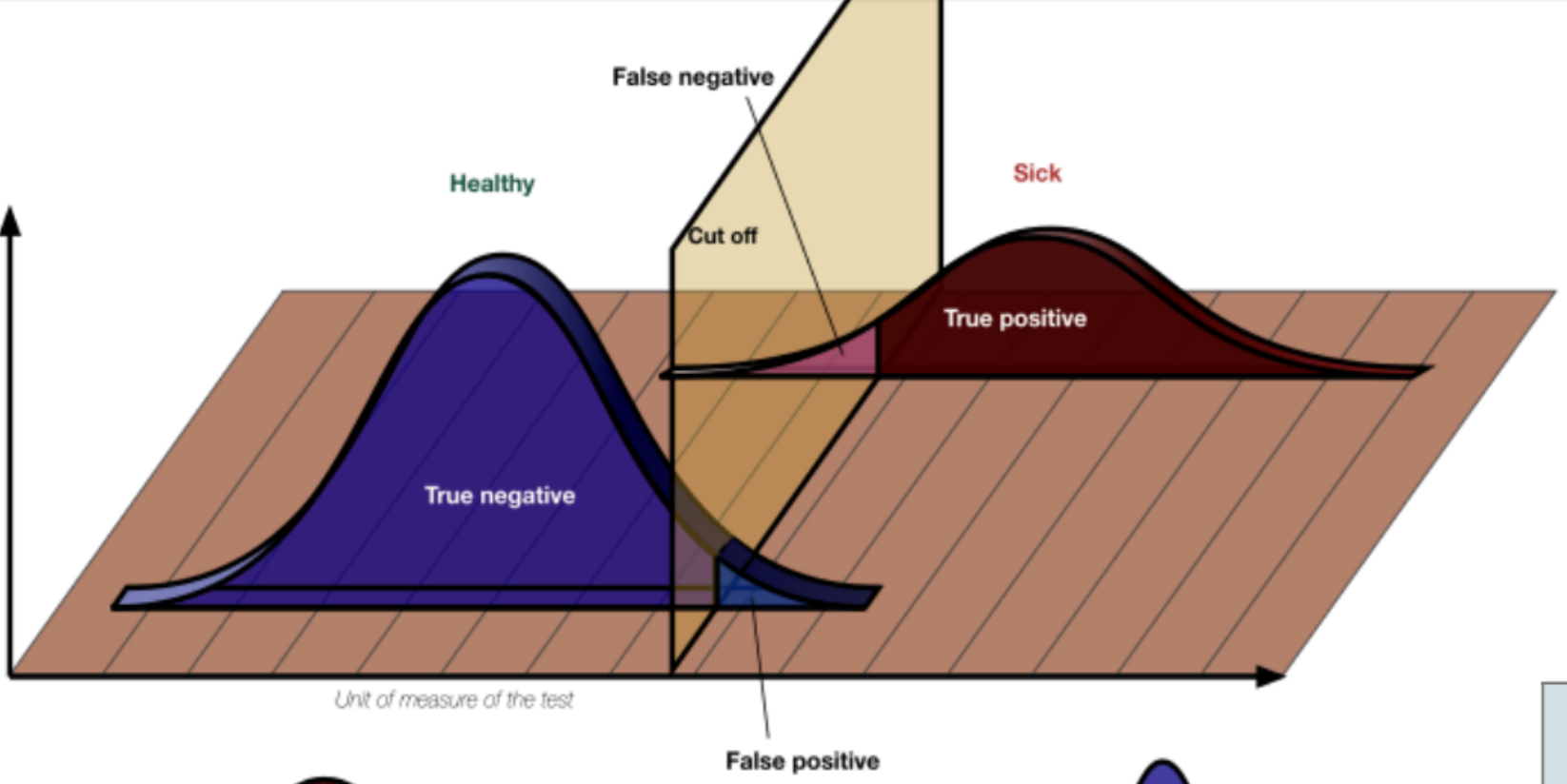
Scenario 2:

A child with an undiagnosed genetic condition undergoes whole-exome sequencing with the goal to identify *any plausible pathogenic variant* that might explain symptoms.

Context	Minimize α (False Positives)	Minimize β (False Negatives)
Genome-wide association discovery		
Variant prioritization in rare disease diagnosis		
Polygenic risk score development		
Early cancer or newborn screening		
Clinical trial confirmatory phase		
Preclinical safety or toxicity screens		

Context	Minimize α (False Positives)	Minimize β (False Negatives)
Genome-wide association discovery	☑ (avoid spurious hits)	
Variant prioritization in rare disease diagnosis		☑ (don’t miss causal variant)
Polygenic risk score development		☑ (capture weak effects)
Early cancer or newborn screening		☑ (catch every true case)
Clinical trial confirmatory phase	☑	
Preclinical safety or toxicity screens		☑

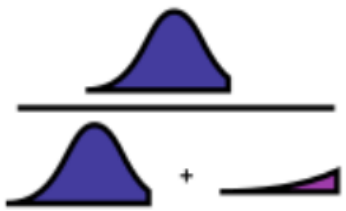
categorical variable prediction



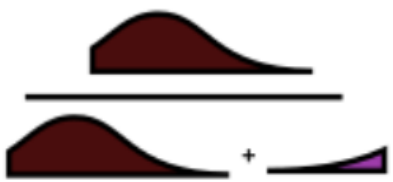
		Actual	
		Positive	Negative
Predicted	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

PPV =

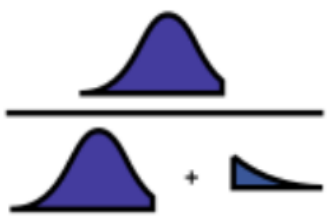
NPV =



Sensitivity =



Specificity =



Power is the ability of a test to reject a false null hypothesis

Power = 1 – Type II = 1- β Power = 1 - P(FTR H_0| H_A)= P(Reject H_0| H_A)

- Power can be increased by increasing the sample size, *n*

Other terms you will encounter:

Sensitivity = 1-Type II = Power = $\frac{\textbf{TP}}{\textbf{(TP+FN)}}$

Specificity = 1-Type I = $\frac{\textbf{TN}}{\textbf{(TN+FP)}}$

Precision = $\frac{\textbf{TP}}{\textbf{(TP+FP)}}$

Accuracy = $\frac{\textbf{(TN+TP)}}{\textbf{(TN+TP+FN+FP)}}$

Add more data points, n, and you are able to discriminate between smaller differences in the null and alternate hypotheses!

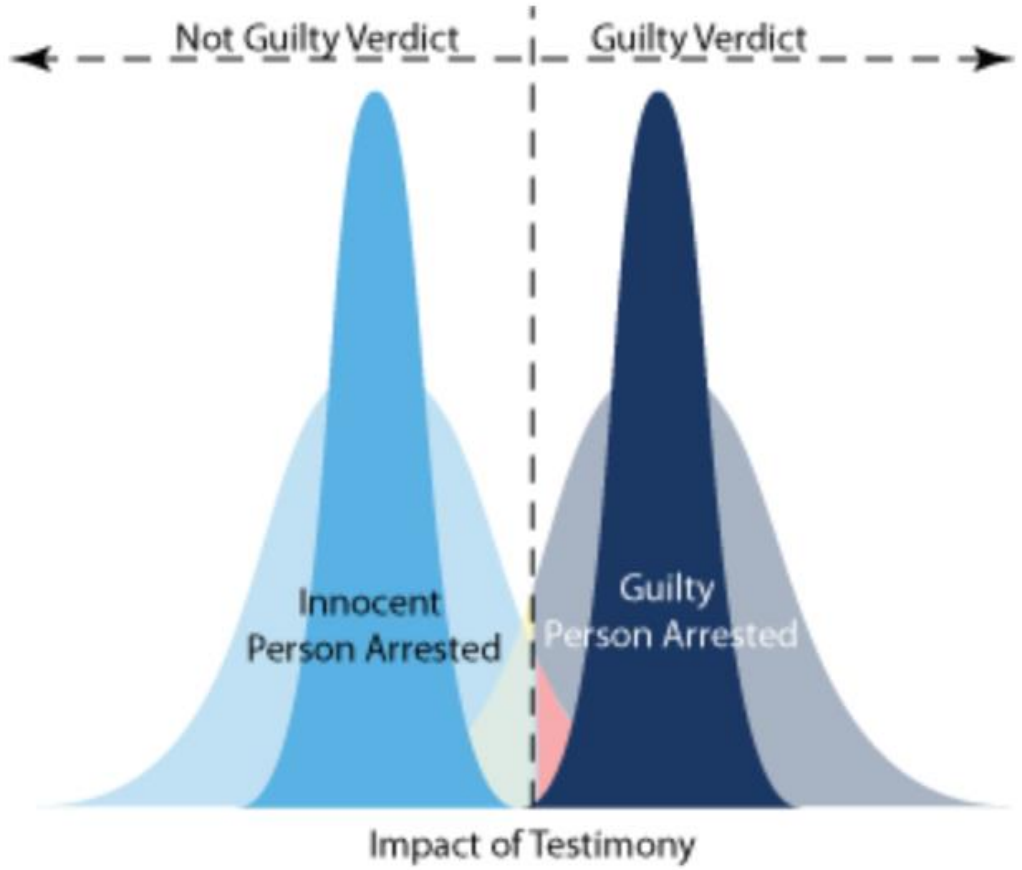


figure 5. The effects of increasing sample size or in other words, number of independent witnesses.

http://www.intuitor.com/statistics/T1T2Errors.html

Two clinical trials are carried out which both test the same null hypothesis under the same conditions with $\alpha = 0.05$. Trial A has 45 individuals and Trial B has 100 individuals.

Which study, **A** or **B,** has higher power?

Confusion matrix – important for classifiers
- contingency table for outcomes
- at a certain level of significance (usually 0.5)

Reality / Model	H_0 true	H_A true
Fail To Reject H_0		
Reject H_0		

Put True Pos, True Neg, False Neg, False Pos. into the four quadrants

https://www.geeksforgeeks.org/confusion-matrix-machine-learning/
https://en.wikipedia.org/wiki/Confusion_matrix

Related ideas:

Accuracy = $\frac{TP + TN}{(TP+TN+FP+FN)}$

- Percentage of correct predictions

Precision = $\frac{TP}{(TP+FP)}$

* % of correct positive class

Recall, Sensitivity, TPR = $\frac{TP}{(TP+FN)}$

* % correct positive out of all correct

- These measurements each have trade-offs

Receiver-Operator Curve – Area Under the Curve

- Used in medical testing and diagnostic radiology etc.
- It is used for comparing test results across multiple thresholds
  - Measures how well a diagnostic test can distinguish between positive and negative cases
- TPR (y axis) versus FPR (x axis)

Specificity = $\frac{TN}{TN + FP}$

FPR = 1- specificity = $\frac{FP}{FP + TN}$

Here is a reasonable summary of the many different summary probabilities that are used (they are each sensitive to certain conditions and robust to others, so you will often use more than one) : https://www.cs.rpi.edu/~leen/misc-publications/SomeStatDefs.html

We have a data-set where we are predicting number of people who have more than $1000 in their bank account. Consider a data-set with 200 observations i.e., n=200

n=200	Prediction=NO	Prediction = YES
Actual = NO	60	10
Actual = YES	5	125

- ❑ Out of 200 cases, our classification model predicted “YES” 125 times, and “NO” 65 times.
- ❑ Out of 200 cases, our classification model predicted “YES” 135 times, and “NO” 5 times.
- ❑ Out of 200 cases, our classification model predicted “YES” 135 times, and “NO” 65 times.
- ❑ Out of 200 cases, our classification model predicted “YES” 135 times, and “NO” 60 times.

https://www.inabia.com/learning/quiz/confusion-matrix-quiz/

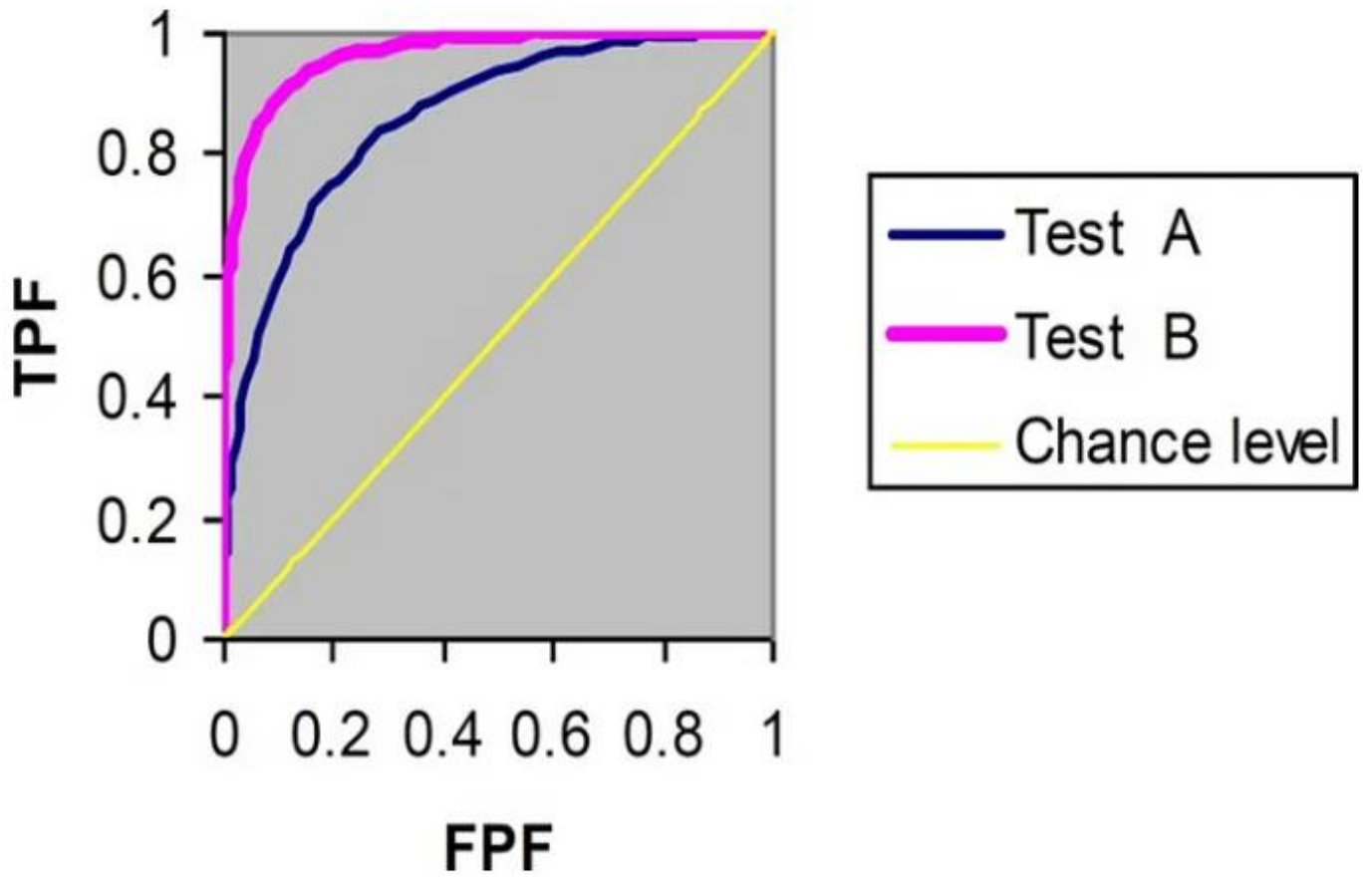
Related ideas:

Accuracy = $\frac{TP + TN}{(TP+TN+FP+FN)}$

Precision = $\frac{TP}{(TP+FP)}$

Recall, Sensitivity, TPR = $\frac{TP}{(TP+FN)}$

FPR = 1- specificity = $\frac{FP}{FP + TN}$

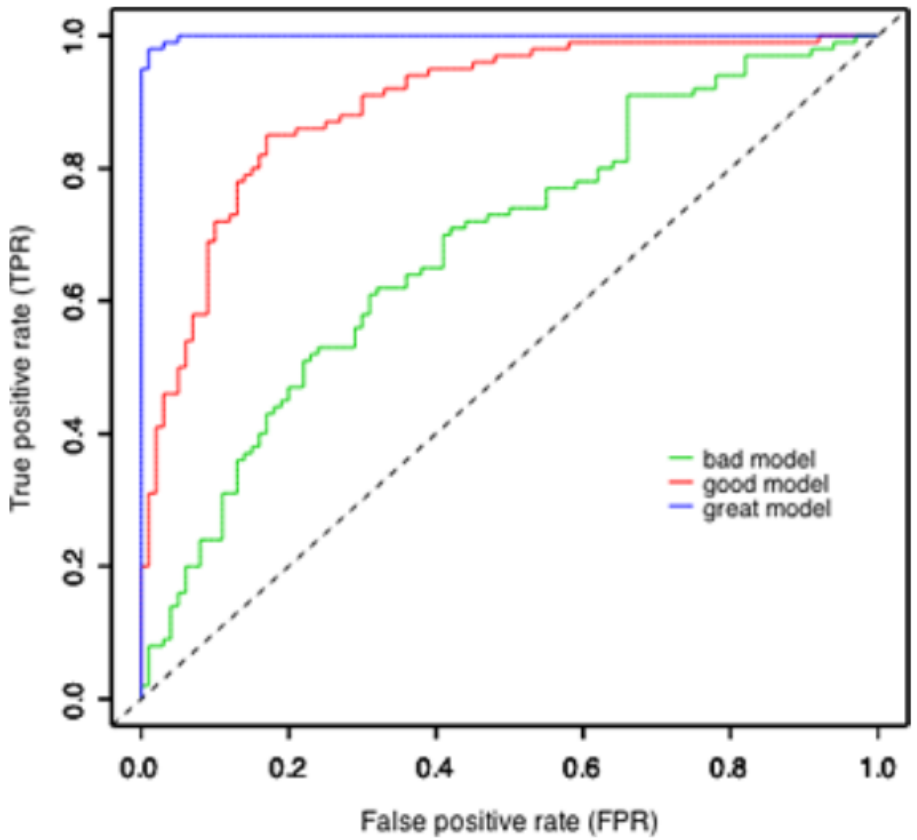


ROC curves of two diagnostic tasks (test A versus test B)(Image source)

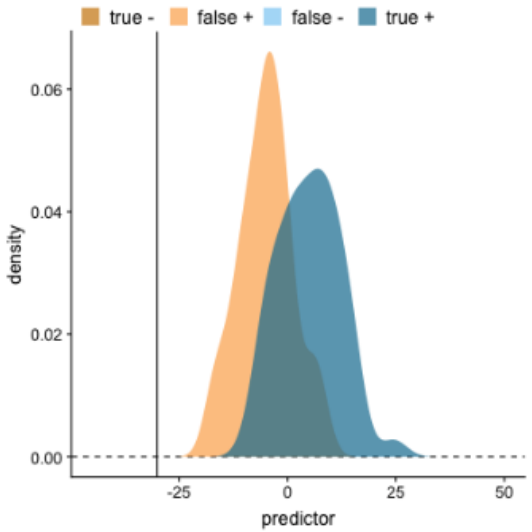
https://medium.com/@shaileydash/understanding-the-roc-and-auc-intuitively-31ca96445c02

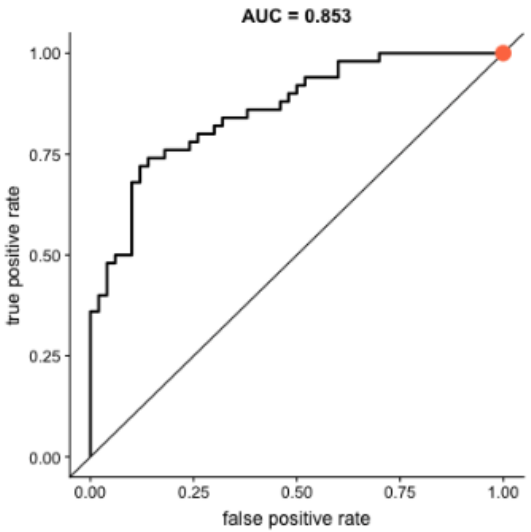
Receiver Operating Characteristic

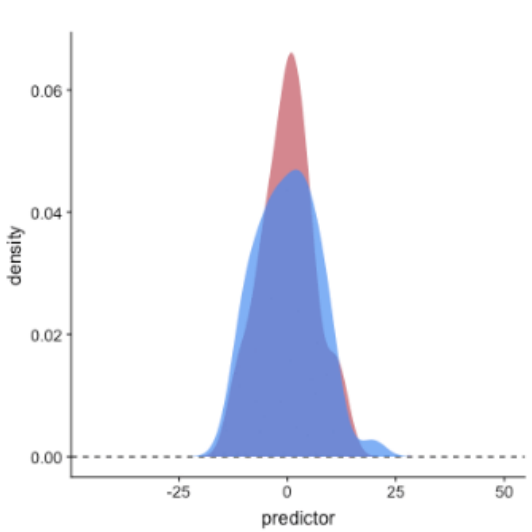
ROC-AUC

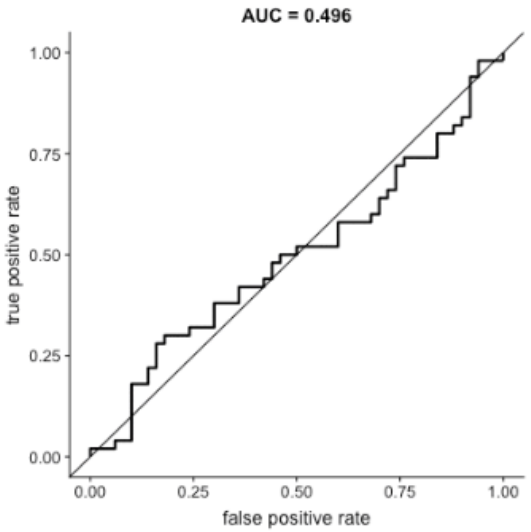


The value can range from 0 to 1. However AUC score of a random classifier for balanced data is 0.5









https://ekja.org/journal/view.php?doi=10.4097/kja.21209