Module 3B : 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

What is “Statistical Thinking”?

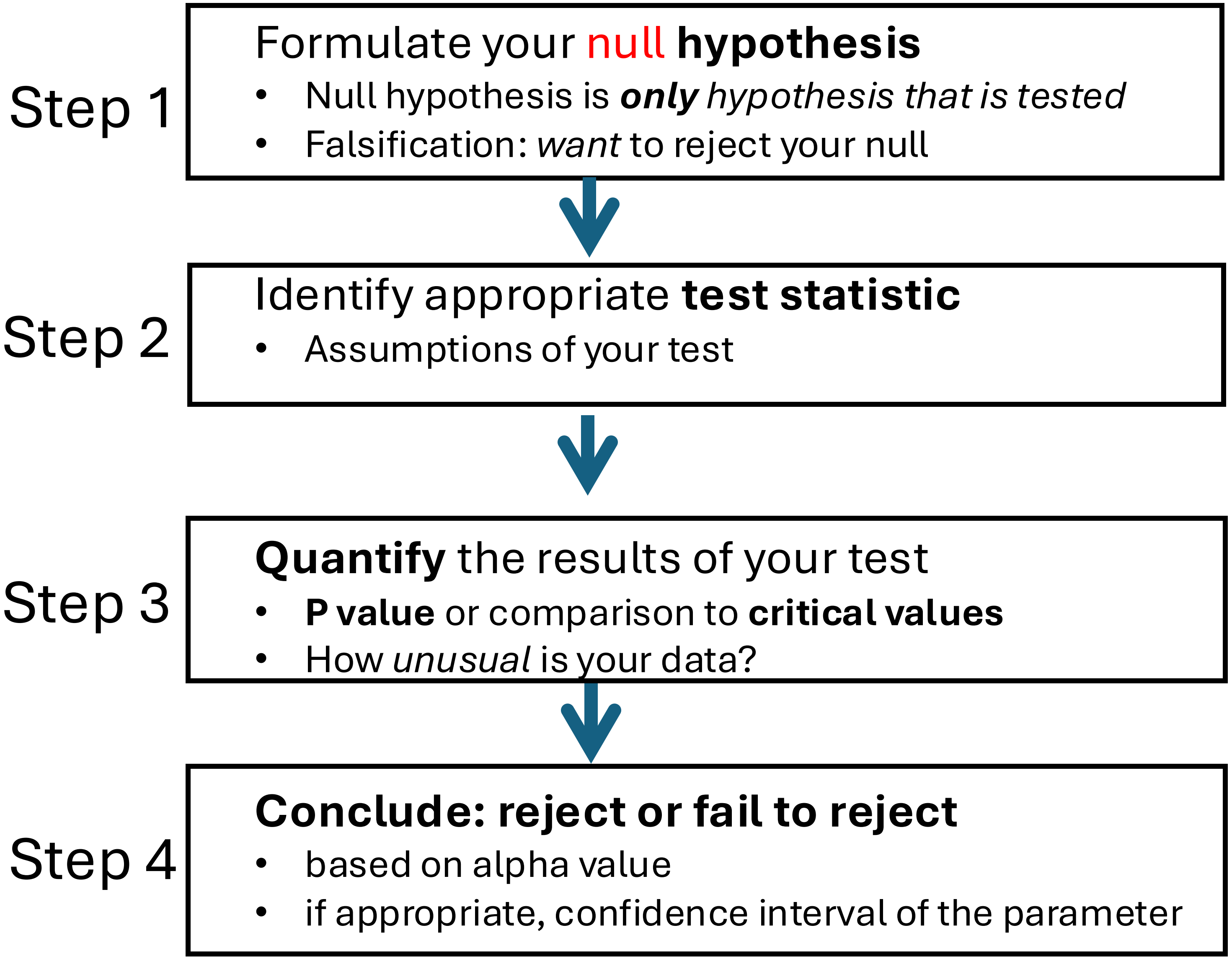
- Understanding complexity via:
  - Understanding Distributions;
  - Models and their assumptions;
  - Quantify uncertainty;
  - Thinking in probabilities;
  - Utilizing systematic criteria for decision making.
- Retraining our brains to not rely on heuristics/shortcuts and bias.

- Most of the work involved in statistics is clearly stating your hypothesis
    - What is your expectation? Can you quantify it? What is the sampling distribution?

- Hypothesis testing allows you to ask if a parameter ***significantly*** differs from the ***null*** expectation
    - It quantifies how unusual the data are *if you assume that the null hypothesis is true.*

- Hypotheses are about populations but are tested with data from samples
    - Assumes that the sampling is random.
    - (most common inferential statistics are parametric – they assume the sampling distribution follows a normal distribution)

Your pipeline for hypothesis testing in statistics



Step 1

Formulate your null **hypothesis**

- Null hypothesis is ***only*** *hypothesis that is tested*
- Falsification: *want* to reject your null

Step 2

Identify appropriate **test statistic**

- Assumptions of your test

Step 3

Quantify the results of your test

- **P value** or comparison to **critical values**
- How *unusual* is your data?

Step 4

Conclude: reject or fail to reject

- based on alpha value
- if appropriate, confidence interval of the parameter

Hypothesis testing automates binary decision making:

1. If p-value < **α** (also called significance level*) by convention, 0.05
   - ➢Reject null hypothesis

2. If p-value > **α**
   - ➢Fail to reject null hypothesis

- We can outline steps that help us make decisions

- **Remember: What is statistically significant is somewhat arbitrary:**

  p-value of 0.04999 is not so different from 0.050001

* Significance level is defined by the scientist before the experiment that quantifies acceptable levels of being wrong about the conclusion (usually the cut-off is 1 in 20 or 5% or 0.05).

Step 1: Making and using hypotheses:

The Null Hypothesis (H_0):

A specific statement about a population parameter made for the purpose of the argument. Usually carefully worded so that it can be rejected (falsified).

The Alternate Hypothesis (H_A):

Represents all other possible parameter values except that stated in H_0. It is often what the researcher hopes is true and remains after the Ho has been rejected.

H_0:

- The *only hypothesis actually tested by the data*

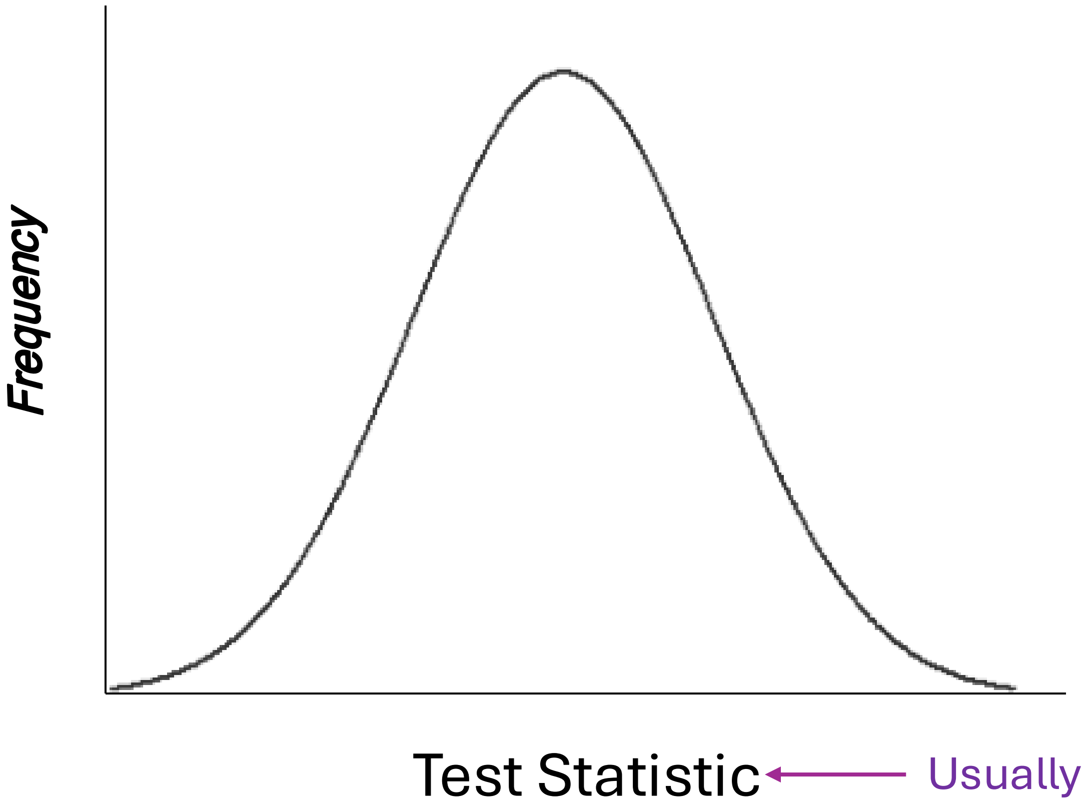
- *Usually, the skeptical POV*
  - *Claims **NO difference/effect***
  - *Observations are just due to chance*

- *Reject or Fail-To-Reject BUT **NEVER EVER** accept*

- *Rejecting H_0 reveals nothing about the magnitude of a parameter*

H_A:

- Usually, the statement that the researchers *hope* is true



Usually the mean

Step 2: Identify a Test Statistic:

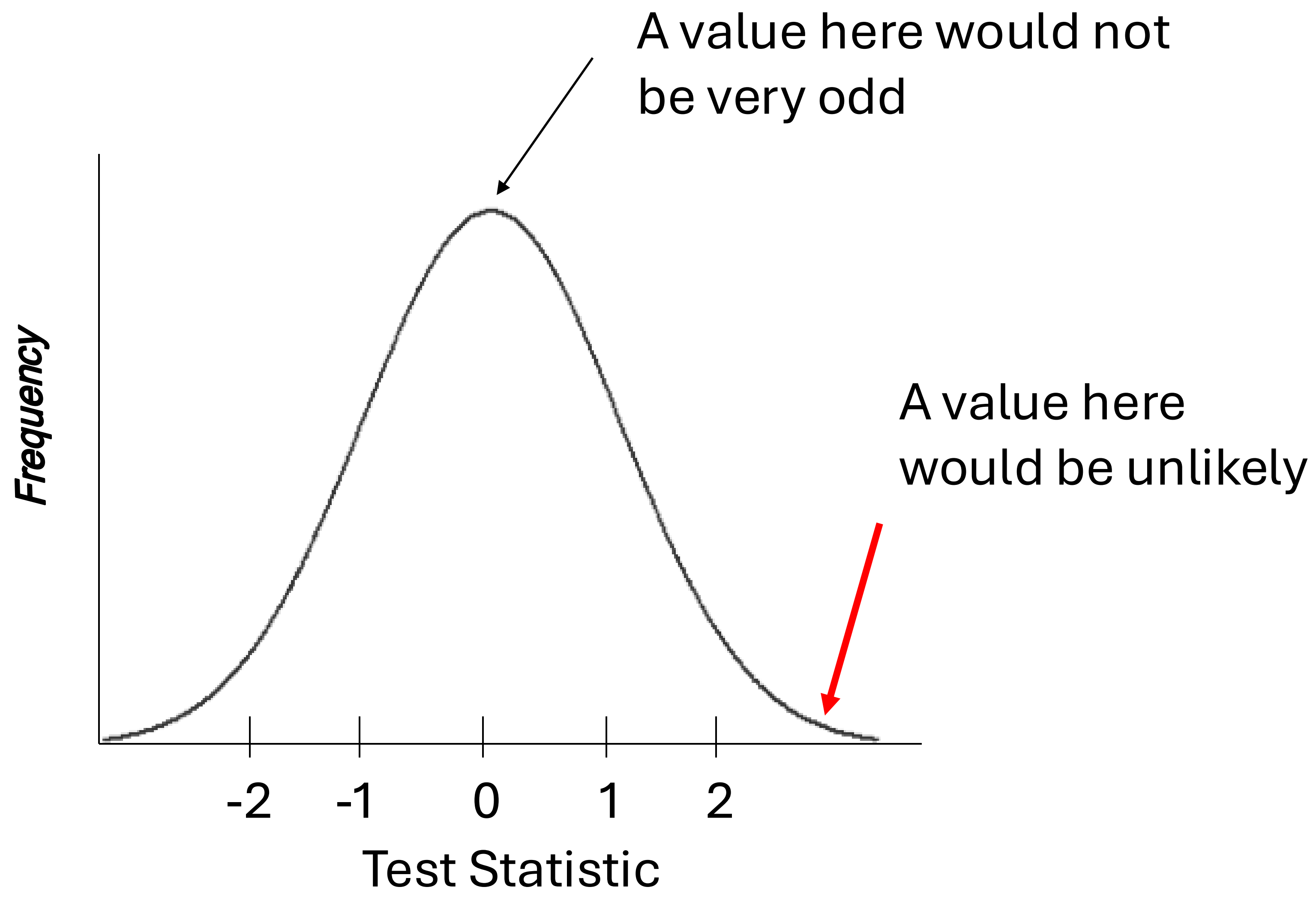
Quantity calculated from the data that is used to evaluate how compatible the results are with those expected the null hypothesis.

- How ‘weird’ are your results?
- Do your data support the assumptions of your test statistic?

Null Sampling Distribution:

Probability of the test statistic assuming the null hypothesis

- Usually assume Normal Distribution (for means, we can usually rely on CTL!)
- Null distribution can be acquired via computer simulations/modeling

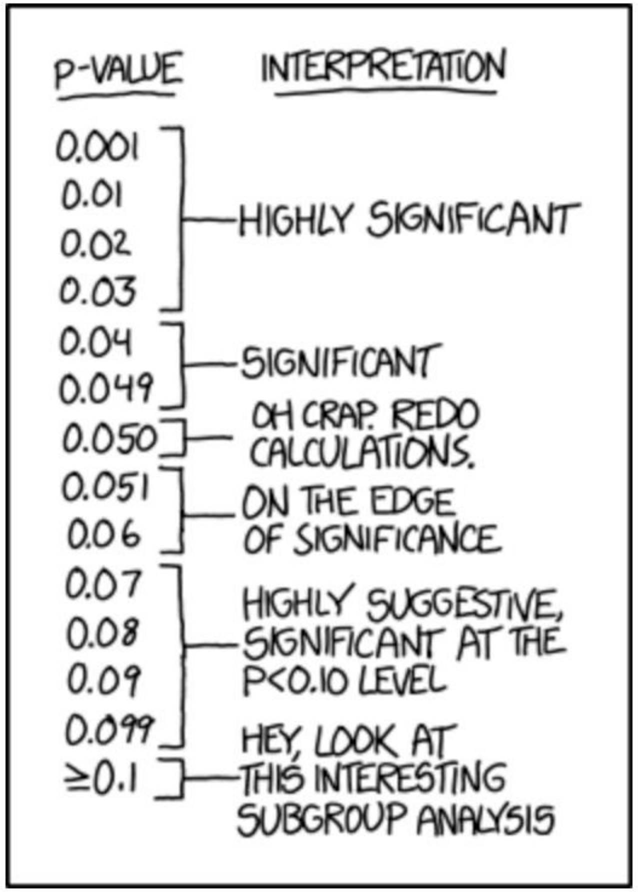


A value here would not be very odd

A value here would be unlikely

<u>P-Value:</u>

Probability of obtaining data that are <u>equal to or even more extreme</u> than the value assuming the null hypothesis is true



P-VALUE	INTERPRETATION
0.001	HIGHLY SIGNIFICANT
0.01	
0.02	
0.03	
0.04	SIGNIFICANT
0.049	
0.050	OH CRAP. REDO CALCULATIONS.
0.051	ON THE EDGE OF SIGNIFICANCE
0.06	
0.07	HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
0.08	
0.09	
0.099	
≥0.1	HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS

http://xkcd.com/1478/

<u>How are P-values found?</u>

-Parametric tests: calculated in R or Python or use cut-off values in published tables.

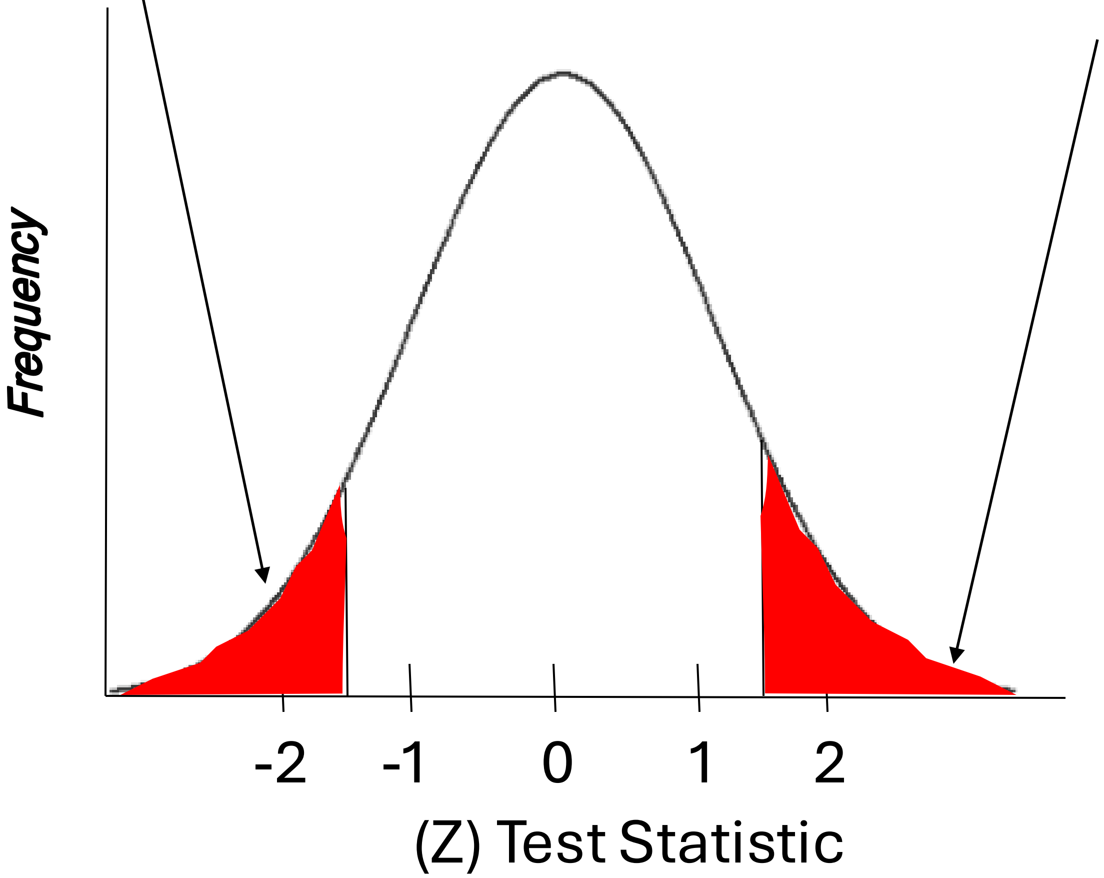
-Re-sampling

-Simulation

P-value

Probability that test statistic
is -1.5 or smaller

Probability that test statistic
is 1.5 or bigger



How do you use a P-value?

In hypothesis testing you can do one of two things:

Reject or Fail-to-Reject H_0

Statistical Significance:

α is used as the basis for rejecting the null hypothesis (α is set by the experimenter; p-values are calculated from the sample)

If p-value $\leq \alpha$, H_0 Rejected

If p-value $> \alpha$, FTR H_0

* α is often 0.05

Hacking p-values: getting the p-value you need to publish your results

- Nate Silver has a widget that demonstrate ‘p hacking’ here:

https://projects.fivethirtyeight.com/p-hacking/

- should we get rid of p-values? http://fivethirtyeight.com/features/statisticians-found-one-thing-they-can-agree-on-its-time-to-stop-misusing-p-values/

- Even well intentioned, honest researchers can accidentally “p-hack”
  - Stopping the study when p-value is significant (**n** individuals) but continuing other studies with more **n** when p-value isn’t yet significant (so you end up with a bias towards studies that have greater **n** and so are more likely to pick up smaller differences)
  - Play with outliers (include or exclude) until a significant p-value is achieved.

Which statement(s) is true about p-values?

----

a. p-value is the probability that the null hypothesis is true or false

b. **p-value reflects the weight of evidence against the null hypothesis**

c. p-value measures the size of the effect

d. if p value is less than or equal to the significance level, then the null hypothesis is not rejected.

Someone claims they make 90% of the shots they make on goal in soccer. If this is tested, what would be the null hypothesis?

----

a)You do not have enough information to make a null hypothesis because you don't know how they will do the test.

b) $\bar{x} = 0.9$ ←This is a sample value, and null hypothesis are about population parameter values μ =0.90

c)The person does not have a mean of making 90% of their soccer shots.

d)The person does have a mean of making 90% of their soccer shots.

Ho: Your friend's proportion of shots on goal, $\mu \neq 0.90$ or , $\mu <= 0.90$

Ha: Your friend, $\mu = 0.90$

Errors in hypothesis testing:

Type I (α) = False Positive

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]

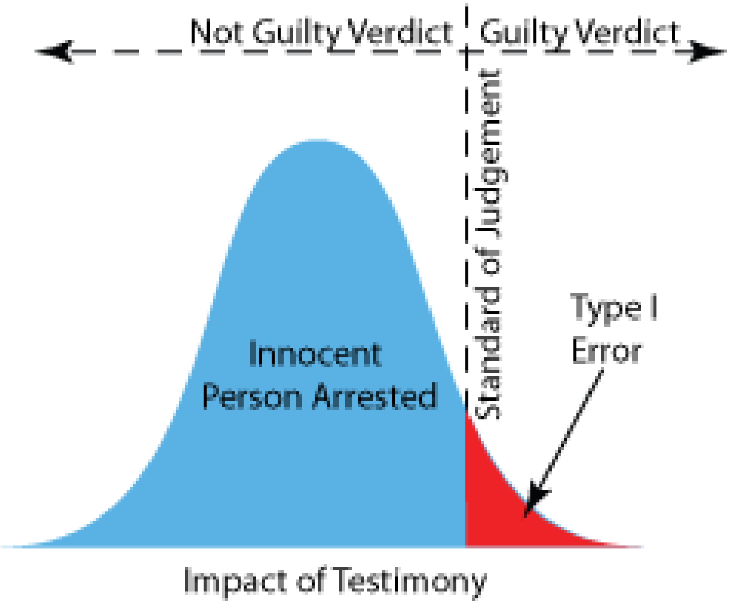
Type III*

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

Type I (α) error:

Rejecting a true null hypothesis

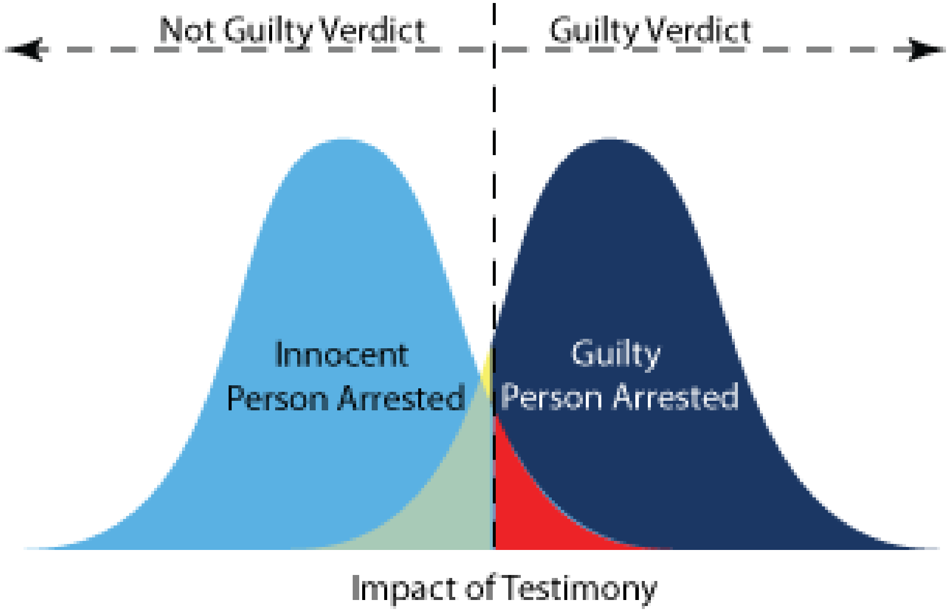
$P(\text{reject } H_0|H_0 = \text{true}) = \alpha$



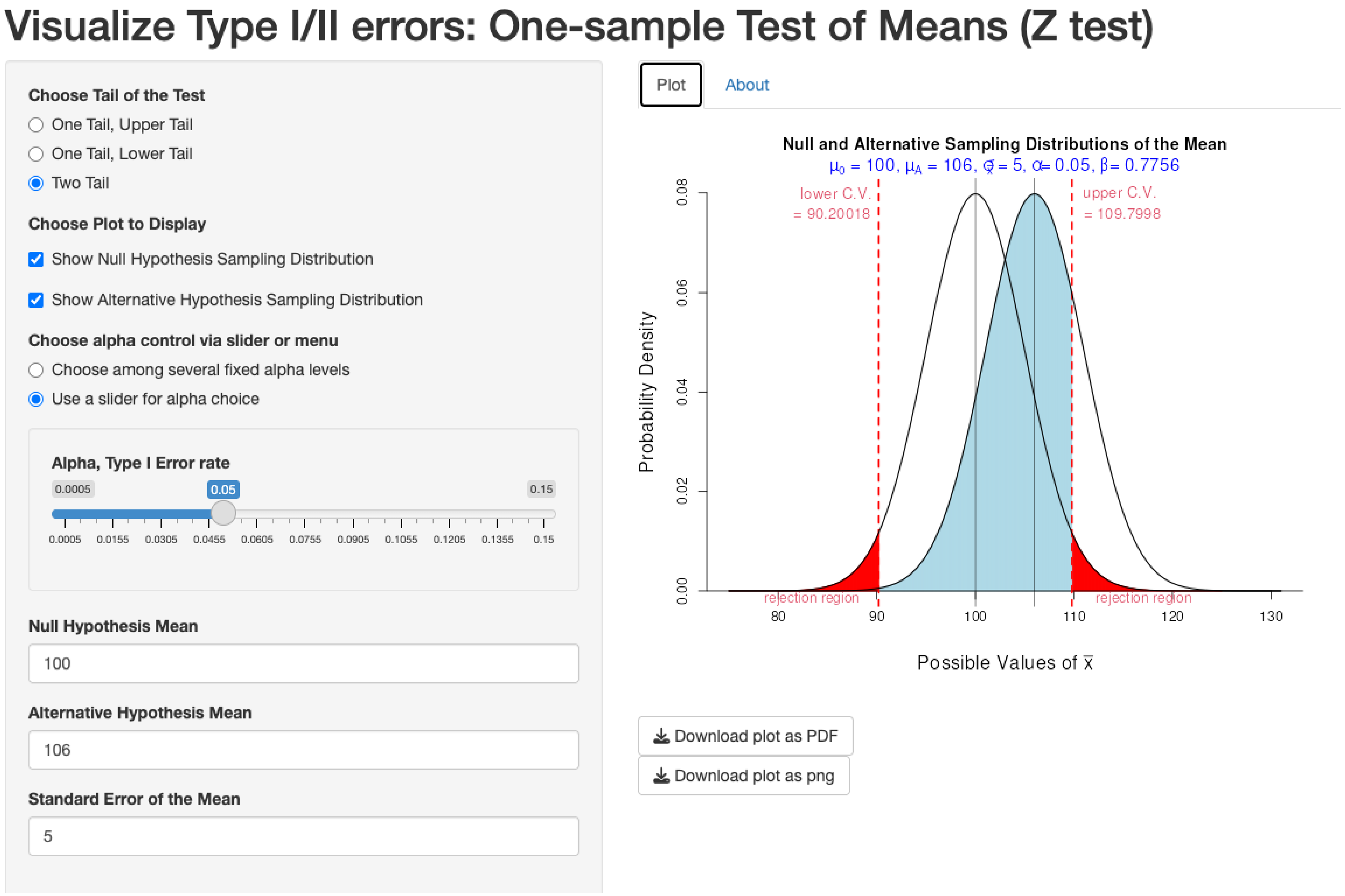
Type II (β) error:

Not rejecting a false null hypothesis

$P(\text{Fail to reject } H_0|H_0 \text{ is not true}) = \beta$



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

Null Hypothesis Mean

100

Alternative Hypothesis Mean

106

Standard Error of the Mean

5

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:

Specificity = P[FTR|Ho is True] = True Negative

$\boldsymbol{\alpha}$=type I= P[Reject|Ho is True] = False Positive

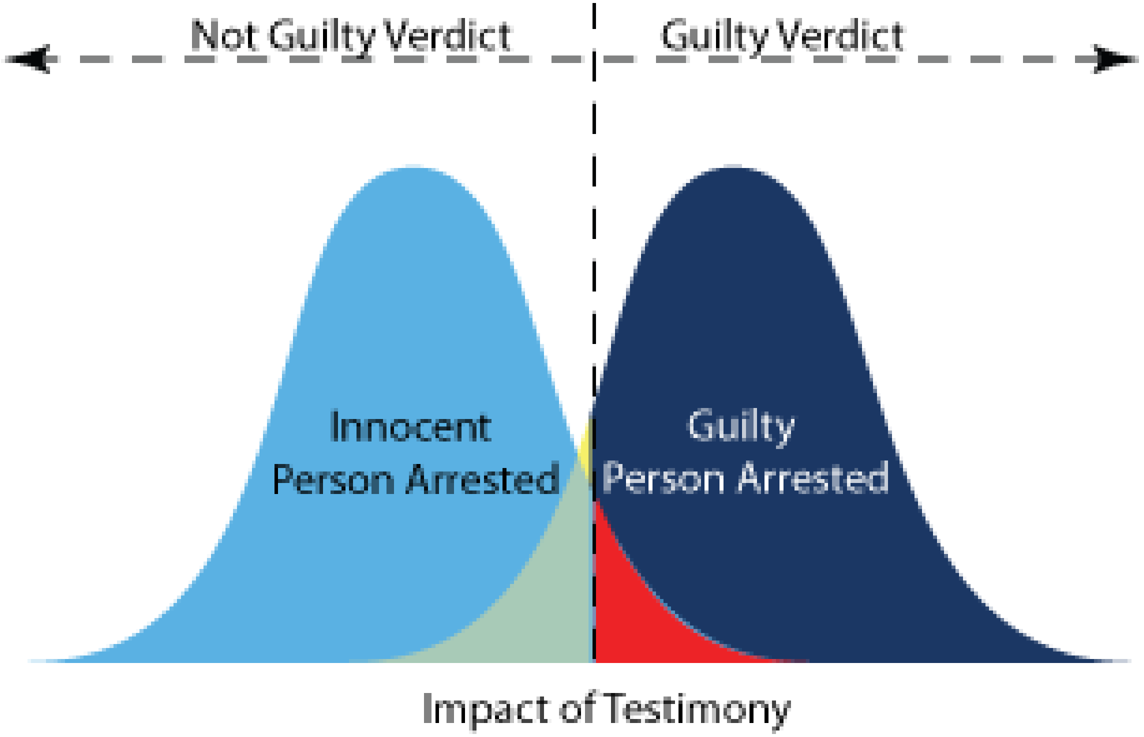
$\boldsymbol{\beta}$=type II error = P[FTR|Ho is **not** True] = False Negative

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

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

	No Disease **(H_0 true)**	**Disease** **(Ho is not true; H_A true)**
Fail To Reject H_0	**No Error** Specificity = 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/Sensitivity P[Reject|Ho is not true] (True Positive)

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



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

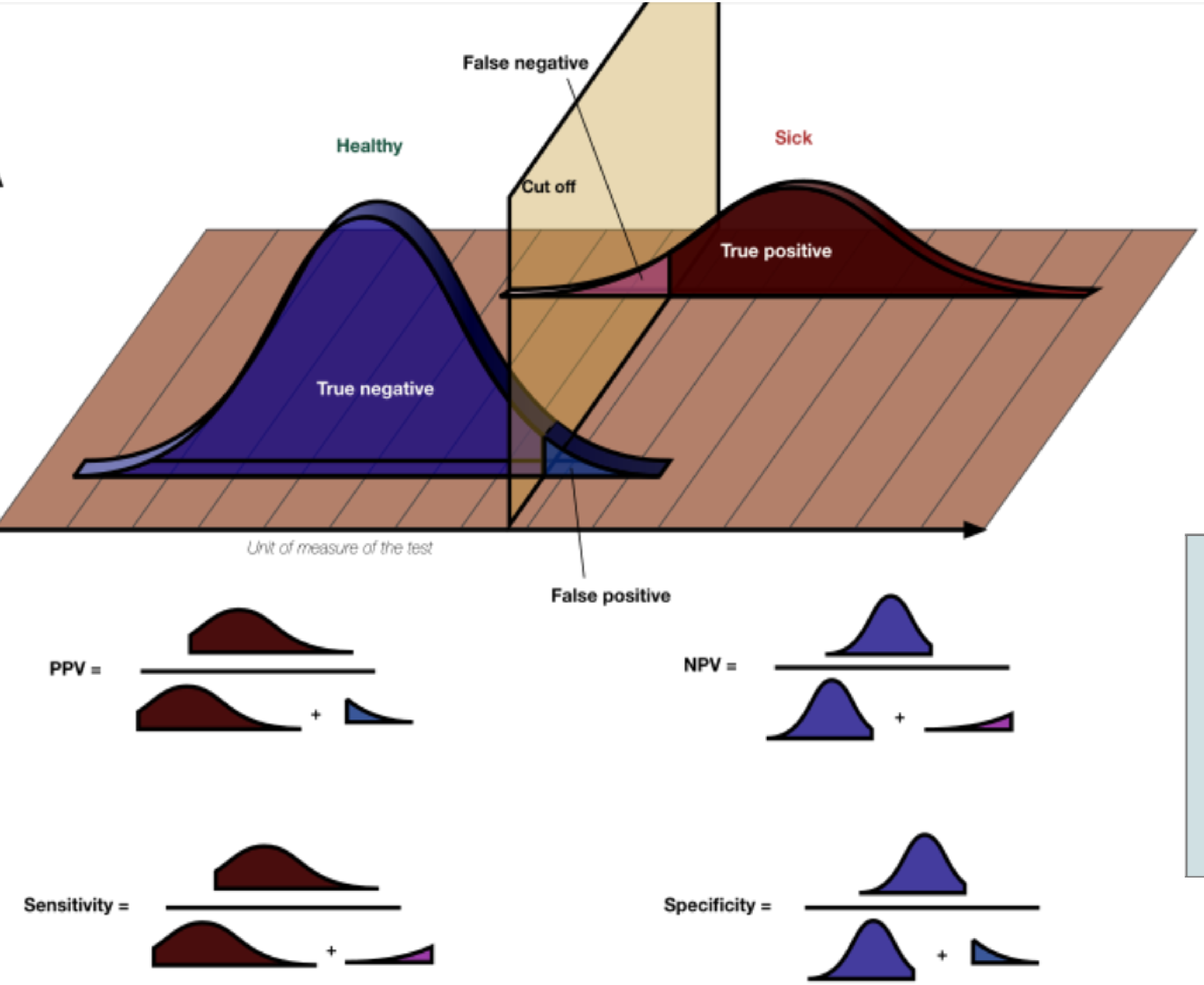
Scenario 1:

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.

categorical variable prediction



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

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 = $\dfrac{\textbf{TP}}{\textbf{(TP+FN)}}$

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

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

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

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

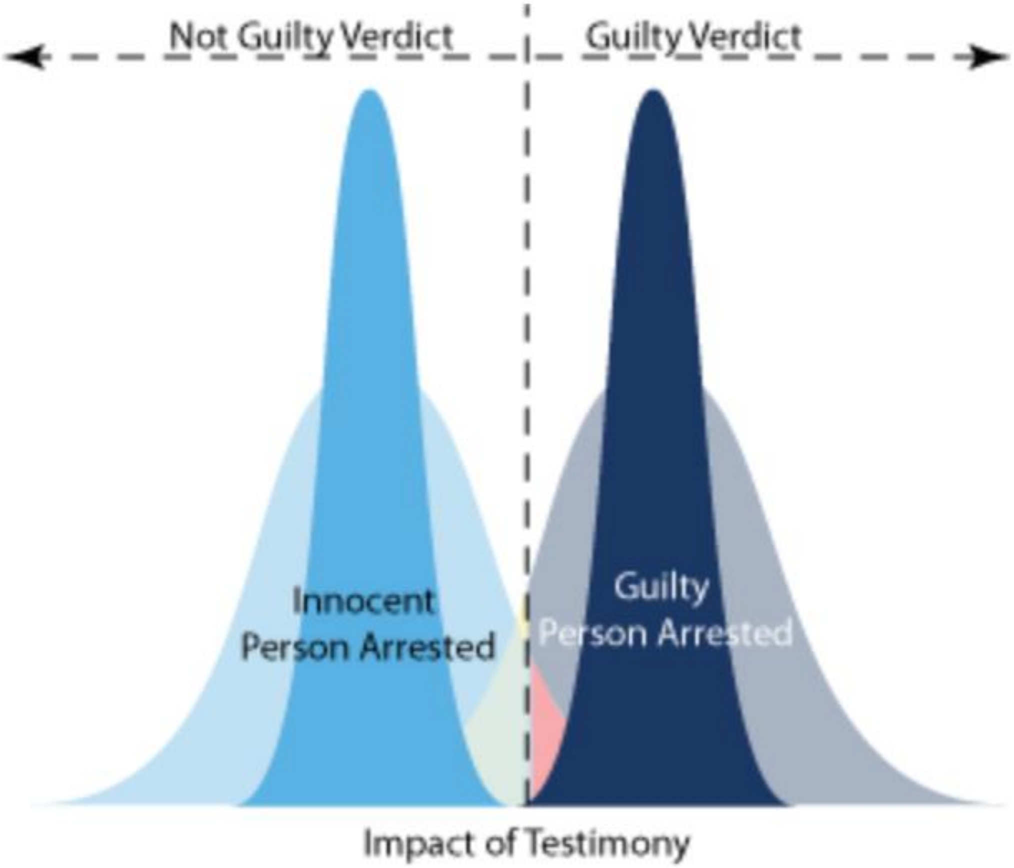


figure 5. The effects of increasing sample size or in other words, number of independent witnesses.

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 α = 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		

We will put TP, TN, FN, FP 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

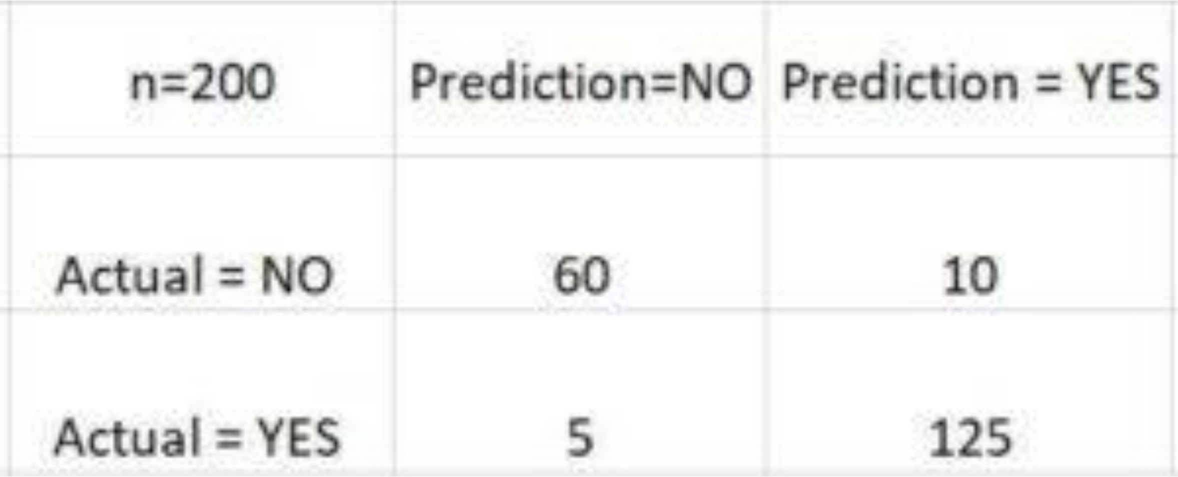
- Used in medical testing and diagnostic radiology etc.
- It is used for comparing test results across multiple thresholds
  - Measures how well a diagnostic test can distinguish between positive and negative cases
- TPR (y axis) versus FPR (x axis)

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/

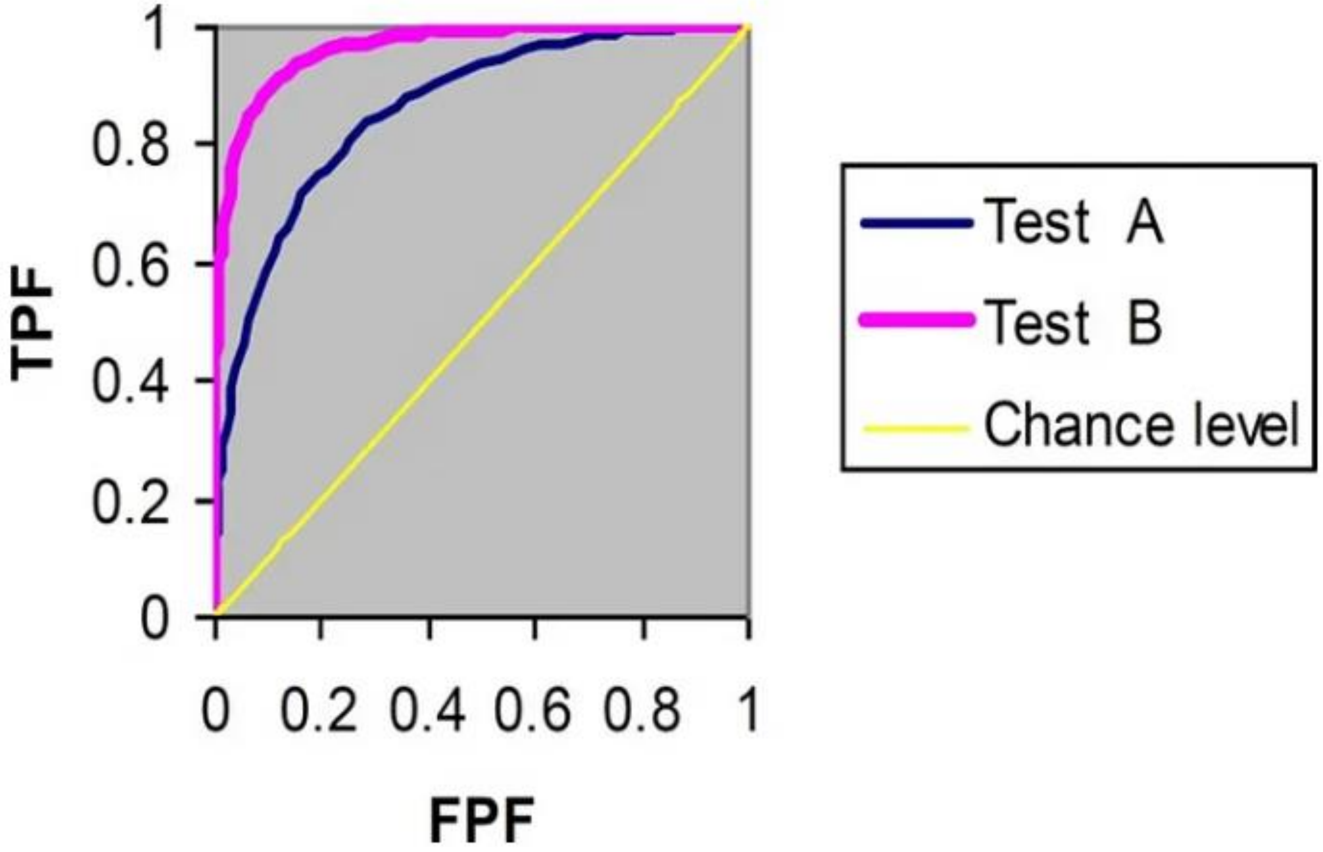
Related ideas:

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

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

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

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

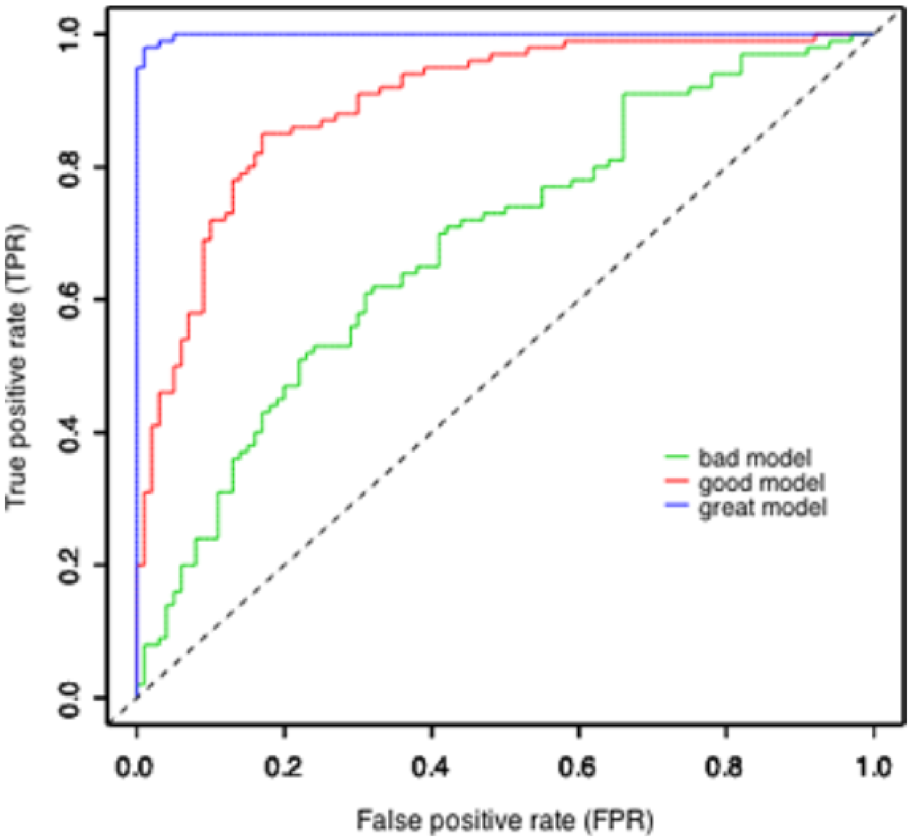


ROC curves of two diagnostic tasks (test A versus test B)(Image source)

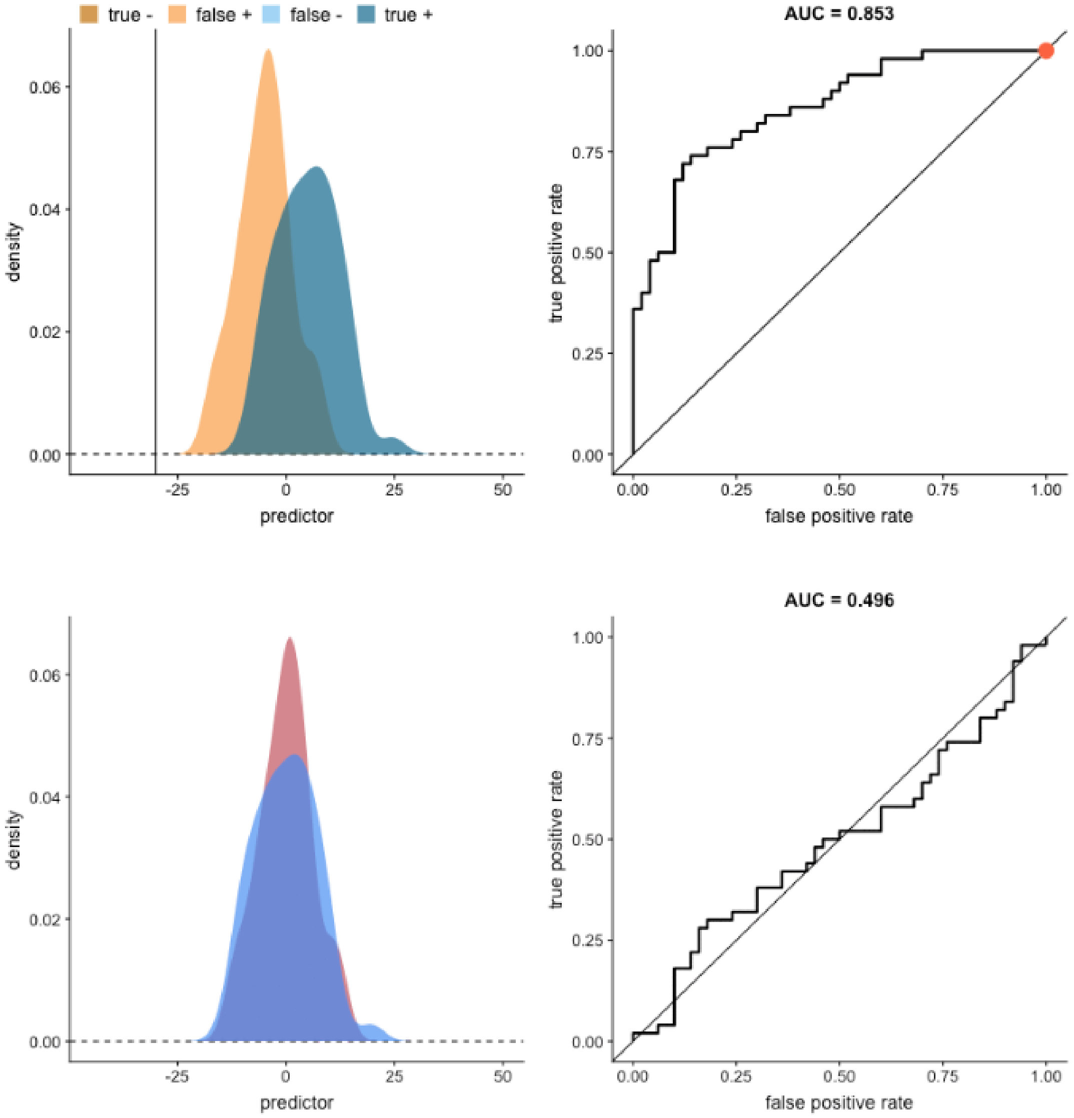
https://medium.com/@shaileydash/understanding-the-roc-and-auc-intuitively-31ca96445c02

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



Two clinical trials are carried out which both test the same null hypothesis under the same conditions with α = 0.05. Trial A has 45 individuals and Trial B has 100 individuals. Power=1-type II (Beta)

Which of the following is true about the two trials described above:

a. Study A has higher probability of type I error than Study B and Study B has a higher probability of type II error than Study A

b. Study A has a lower probability of type I error than Study B and Study B has a lower probability of type II error than Study A

c. Study A has the same type I error as Study B and Study A has a higher probability of type II error than Study B. B/c Power=1-P(type II error)

d. Study A has the same type I error as Study B and Study B also has a higher probability of type II error than Study A

Austin et al (2006): sifted through health care data for >10 million residents and **223** different reasons for admissions; **12 astrological signs.**

Conclusion: 72 conditions were significantly associated with a particular zodiac sign.

https://www.sciencedirect.com/science/article/abs/pii/S0895435606001247

Two independent studies are performed to test the same null hypothesis.

What is the probability that one or both of the studies obtains a significant result and rejects the null hypothesis ***even if the null hypothesis is true***? Assume that in each study there is a 0.05 probability of rejecting the null hypothesis.

a. **0.10**

b. **0.075**

c. **0.05**

d. **0.0975**

We can rephrase the above a bit: "If we repeat this experiment 100 times, what proportion of our 'significant' results would actually be false positives?"

You can consider the previous question in one of two equally valid ways:

P(at least 1 study obtains significant results)

= 1-P(neither study obtains significant results)

= 1 – $(1-0.05)^2$ = 0.0975

P[1st study significant OR 2nd study is significant]

= (0.05)+(0.05) –$(0.05)^2$ = 0.0975

The experimenter thinks that they are using an alpha=0.05, but they are actually using an alpha =0.0975

223 admissions reasons, 223*12 hypothesis

Austin et al (2006): sifted through health care data for >10 million residents and **223** different reasons for admissions; **12 astrological signs.**

Conclusion: 72 conditions were significantly associated with a particular zodiac sign.

- **This is actually 223*12 hypothesis being tested (~2500 hypothesis)**
- You expect 134 statistically significant associations just due to change so 72 is < 134 (calculated with alpha =0.05)

Bonferroni correction

alpha*=alpha/num of hypothesis = 0.05/(223*12) = 0.0000187

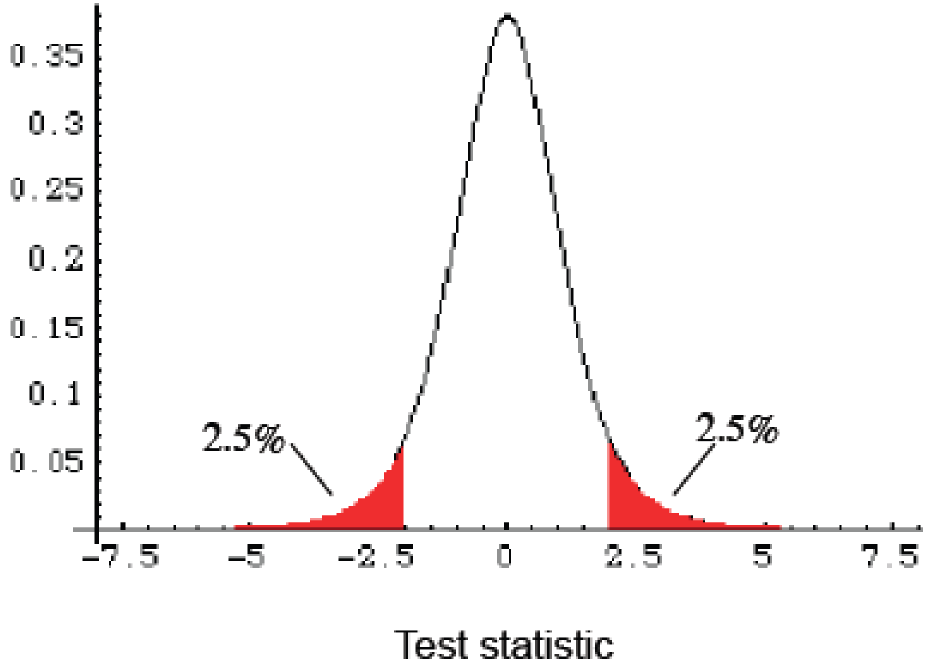
- (GWAS tests hundreds of thousands if not millions at a time)

https://www.sciencedirect.com/science/article/abs/pii/S0895435606001247

One tailed and two tailed tests:

- Most tests are two-tailed tests

- This means that a deviation in either direction would reject the null hypothesis
  - this means that α is divided into $\frac{\alpha}{2}$ on the one side and $\frac{\alpha}{2}$ on the other



One Tailed Tests:

Only used when the other tail is nonsensical

- **Example:**
  - Comparing grades on a multiple-choice test to random guessing
  - Do dogs resemble their owners?

- **Example:**
  - Do daughters resemble their biological fathers?
    - Experiment involves a subject who examines photo of one girl and two adult men and guesses the father
    - If subjects pick father correctly > 0.5 then the hypothesis being tested would FTR
    - Wouldn’t make sense that daughters would, on average, resemble their biological fathers less than other men.

- Some parting words & popular misconceptions:
    - FTR does not mean ACCEPT
        - We ***never*** accept the null hypothesis (more information could become available)
    - If FTR the null hypothesis, we can conclude that the data is compatible with the hypothesis
- If the result is statistically significant there is a temptation to believe that the effect is large. DO NOT GIVE IN THIS TO ERRONEOUS BELIEF.
    - Nor does it mean that the effect is interesting
    - If the sample size is large (and measurements have little variation) then even inconsequential differences will be significant
- **P-values are calculated from the data itself**. In contrast, the alpha value is set by the experimenter prior to conducting the experiment. P-values and alpha are related BUT THEY ARE NOT THE SAME!

- Why use hypothesis testing at all?
    - Why don’t we skip hypothesis testing since confidence intervals give us similar information **plus** gives us information about the actual magnitude of the parameter?
        - *Main purpose of hypothesis testing is to determine if sufficient evidence has been presented to support a scientific claim*
- Deploy these tools wisely
    - Just because something is statistically significant does not mean it is biologically important or interesting
    - **Almost any null hypothesis can be rejected with a large enough sample**

Upload to Canvas as part of your Certification of Completion

Module 3B Questions:

Finish up and hand in any of the questions in orange AND the confusion matrix.