Working with data in your research and paper

Ioannis Konstantinidis

Sr. Researcher, Dept. of Computer Science

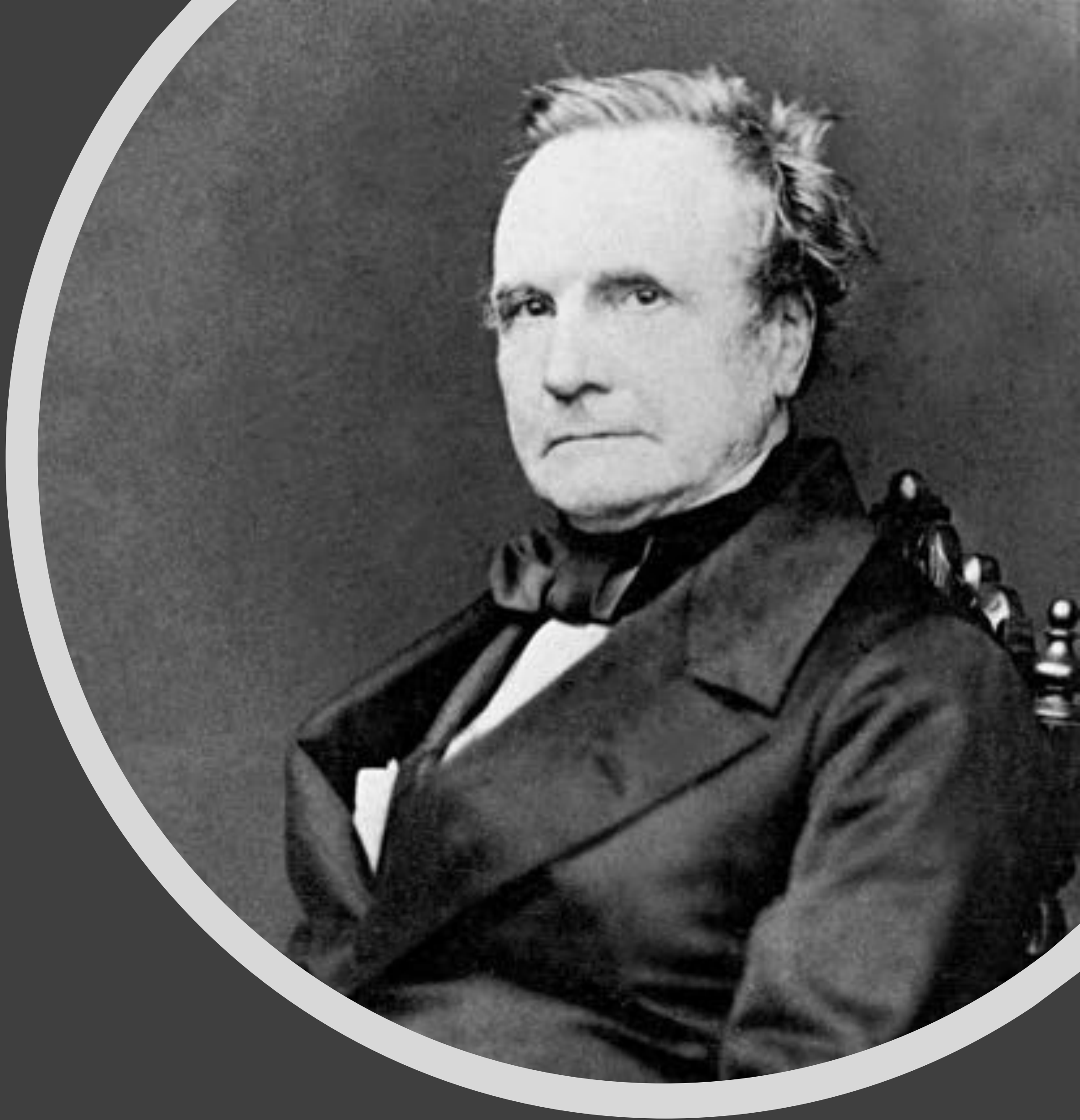
ikonstantinidis@uh.edu

Nutrition Facts
Serving Size
Servings Per Container

These slides were manufactured on equipment that processes words. May contain typos, mistakes, or omissions.

On two occasions I have been asked,—”Pray, Mr. Babbage, if you put into the machine wrong figures, will the right answers come out?” ... I am not able rightly to apprehend the kind of confusion of ideas that could provoke such a question.

Charles Babbage (1791-1871) *Passages from the Life of a Philosopher*, ch. 5 “Difference Engine No. 1” (1864)



Does

- the statistical summary say what you *think* it says?
- the statistical summary give the *full* picture?
- the statistical test ask the *right* question?
- the statistical test say what you *think* it says?

STATISTICAL SUMMARIES

Congratulations!

Your dataset summaries look right

But does your dataset contain “wrong figures”?

Does

- the statistical summary say what you *think* it says?
- the statistical summary give the *full* picture?
- the statistical test ask the *right* question?
- the statistical test say what you *think* it says?

If your weight is **average**, then

A. You are as likely to run into someone that weighs more than you as you are to run into someone that weighs less than you
B. If everyone else’s weight changed to match yours exactly, elevator capacity signs could stay the same; but if everyone’s weight changed to be double your weight, then elevator capacities would need to be cut in half
C. None of the above

If your weight is **average**, then

A. Median

vs.

B. Mean

Text-based summary (by threshold)

Centrality
What **value** splits the observations in half? (half the values are above, the other half are below) MEDIAN

The median describes RELATIVE POSITION for a SINGLE individual within an ENSEMBLE of peers

Text-based summary (by threshold)

Centrality
What **value** splits the observations in half? (half the values are above, the other half are below) MEDIAN

The median describes RELATIVE POSITION for a SINGLE individual within an ENSEMBLE of peers

We need to reorder the column of observations to compute (they must be in ascending / descending order) – standard LeetCode question!

Text-based summary (in aggregate)

Centrality
How does the sum total of all **values** compare[1]? MEAN

The mean compares CUMULATIVE VALUES for a POOLED ENSEMBLE of peers to a STANDARDIZED MEASURE (sum/#)

[1] to the number of observations

Text-based summary (in aggregate)

Centrality
How does the sum total of all **values** compare[1]?   MEAN

Simple to compute, even on paper – no need to reorder the column of observations

The mean compares CUMULATIVE VALUES for a POOLED ENSEMBLE of peers to a STANDARDIZED MEASURE (sum/#)

[1] to the number of observations

MEAN as a stand-in for MEDIAN

If the histogram is symmetric,

i.e., for each value above the median,

there is a value at equal distance below the median

and vice versa

then all these differences will cancel each other out when we compute the sum total of all the values,

so the MEAN will be equal to the MEDIAN

Cautions

If the histogram is not symmetric (we call that skew)

then the MEDIAN and MEAN might be very different from each other

Cautions

If the histogram is not symmetric (we call that skew)

then the MEDIAN and MEAN might be very different from each other

Why does this matter?

MEAN is the flip-side of the MEDIAN

The mean is the POV of the house

- Q: How <u>much</u> profit did the house *realize* (*per gambler)?*
- A: The mean is equal to the profit per gambler

Note: This is not saying how <u>many</u> people profited/lost

MEAN is the flip-side of the MEDIAN

The mean is the POV of the house

Q: How much profit did the house *realize* (*per gambler)?*

A: The mean is equal to the profit per gambler

Note: This is not saying how many people profited/lost

The median is the POV of the gambler

Q: How many gamblers in a group *realized a* profit?

A: If median > 0, then more than half profited; If median < 0, then less than half did

Note: This is not saying how much the profit/loss would be per gambler

If your weight is average, then

A. You are as likely to run into someone that weighs more than you as you are to run into someone that weighs less than you
B. If everyone else’s weight changed to match yours exactly, elevator capacity signs could stay the same; but if everyone’s weight changed to be double your weight, then elevator capacities would need to be cut in half
C. Clothes fitted in your size are the most popular size option
D. All of the above
E. None of the above

Text-based summaries: three ways

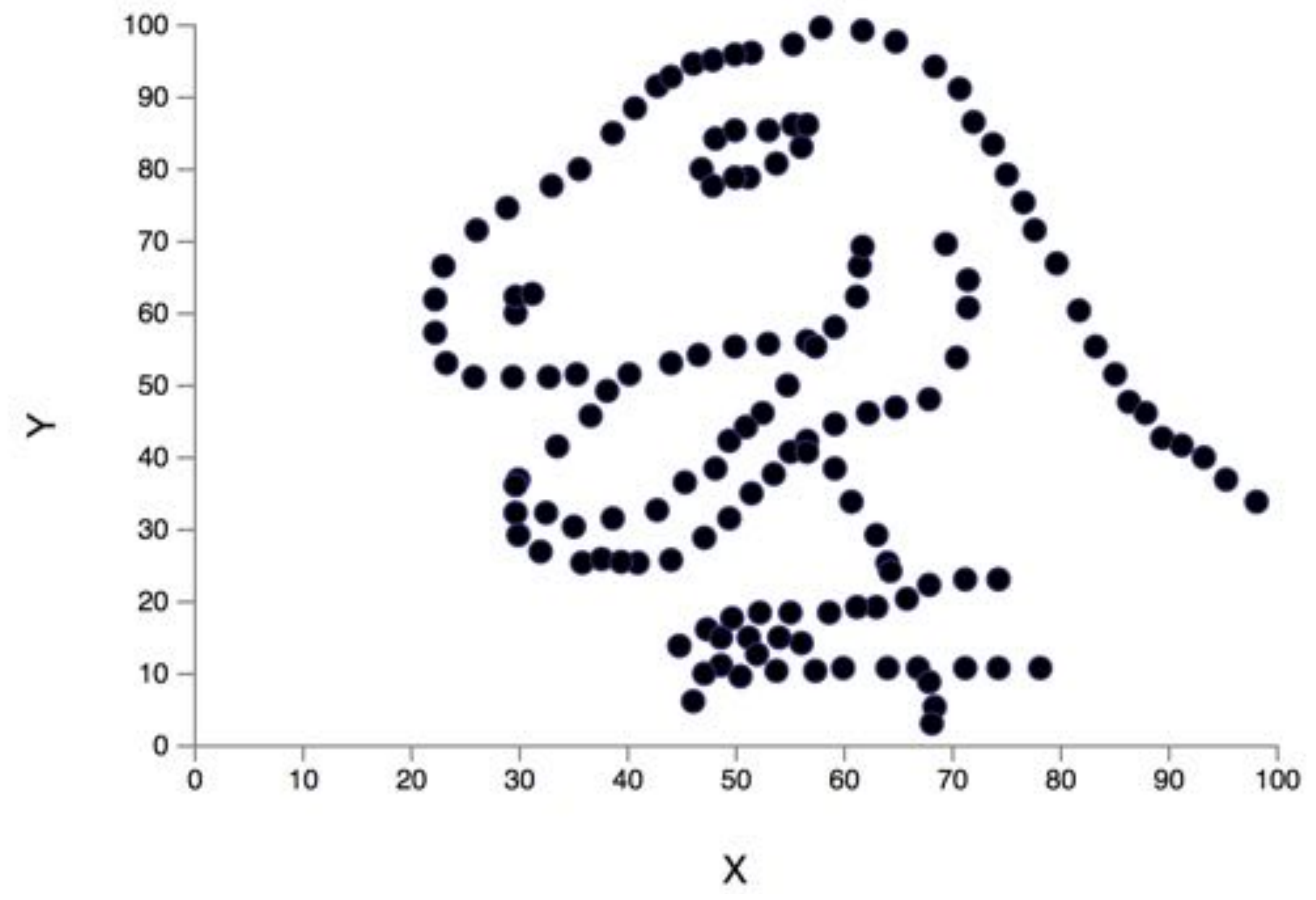
	Centrality	Dispersion
vote	What **value** is the most popular? MODE	How **many values** are very popular? Modality
threshold	What **value** splits the observations in half? (half the values are above, the other half are below) MEDIAN	What **band of values** splits the observations in half? (half the values are inside, the other half are outside) IQR
aggregate	How does the sum total of all **values** compare[1]? MEAN	How does the sum total of all **deviations[2]** compare[1]? Variance = (standard deviation)^2

[1] to the number of observations, i.e., sum/#

[2] squared distances from the mean, i.e., (value-MEAN)^2

Does

- ✓ the statistical summary say what you *think* it says?
- ➢ the statistical summary give the *full* picture?
- the statistical test ask the *right* question?
- the statistical test say what you *think* it says?



https://www.autodesk.com/research/publications/same-stats-different-graphs

The Datasaurus

STATISTICAL TESTS: meaningful differences

Congratulations! Your experiment found a difference in performance

STATISTICAL TESTS: meaningful differences

Congratulations! Your experiment found a difference in performance

But should you be measuring this difference to begin with?

Does

- ✓the statistical summary say what you *think* it says?
- ✓the statistical summary give the *full* picture?
- ➢the statistical test ask the *right* question?
- the statistical test say what you *think* it says?

SAT scores over time

GPA	SAT (1992)	SAT (2002)
A+	619	607
A	575	565
A−	546	538
B	486	479
C	434	424
All grades	**501**	**516**

Rinott, Yosef and Michael Tam, 2003, “Monotone Regrouping, Regression, and Simpson’s Paradox”, *The American Statistician*, 57(2): 139–141. doi:10.1198/0003130031397

SAT scores over time

GPA	SAT (1992)	SAT (2002)	% change
A+	619	607	-2%
A	575	565	-2%
A−	546	538	-1%
B	486	479	-1%
C	434	424	-2%
All grades	**501**	**516**	

Across ALL grades, an average DROP between 1% and 2%

Rinott, Yosef and Michael Tam, 2003, “Monotone Regrouping, Regression, and Simpson’s Paradox”, *The American Statistician*, 57(2): 139–141. doi:10.1198/0003130031397

SAT scores over time

GPA	SAT (1992)	SAT (2002)	% change
A+	619	607	
A	575	565	
A−	546	538	
B	486	479	
C	434	424	
All grades	**501**	**516**	**3%**

Among ALL students, an average INCREASE of 3%

Rinott, Yosef and Michael Tam, 2003, “Monotone Regrouping, Regression, and Simpson’s Paradox”, *The American Statistician*, 57(2): 139–141. doi:10.1198/0003130031397

SAT scores over time

GPA	SAT (1992)	SAT (2002)	% change
A+	619	607	-2%
A	575	565	-2%
A−	546	538	-1%
B	486	479	-1%
C	434	424	-2%
All grades	**501**	**516**	**3%**

Rinott, Yosef and Michael Tam, 2003, “Monotone Regrouping, Regression, and Simpson’s Paradox”, *The American Statistician*, 57(2): 139–141. doi:10.1198/0003130031397

Suppose grading curves change over time (“grade inflation”), so ALL students get slightly better grades.

- Now the high scorers in one letter grade will be classified among the low scorers in the next higher letter grade,
  - ➢ This would lower the SAT average *per group*.
- At the same time, the *overall* SAT average could rise from 501 to 516.

A conclusion from the **stratified** data that “students scores are falling” would be mistaken

Far better an approximate answer to the *right* question, which is often vague, than an *exact* answer to the wrong question, which can always be made precise.

- John Tukey, “The future of data analysis,” *Annals of Mathematical Statistics* 33 (1) (1962)
- https://projecteuclid.org/download/pdf_1/euclid.aoms/1177704711

STATISTICAL TESTS: meaningful differences

Congratulations! Your experiment found a difference in performance

But is this difference real or random?

Does

- ✓the statistical summary say what you *think* it says?
- ✓the statistical summary give the *full* picture?
- ✓the statistical test ask the *right* question?
- ➢the statistical test say what you *think* it says?

R. Fisher

- Thinks like a detective
- Tries to identify suspects
- Wants to doubt everyone

Fisher thinks like a detective

Null hypothesis (H_0) is the default position (claims innocence):

- This person's actions **are NOT** incriminating
- The difference in population means is ZERO

Fisher thinks like a detective

Null hypothesis (H_0) is the default position (claims innocence):

- This person's actions **are NOT** incriminating
- The difference in population means is ZERO

There is **no** specific alternative hypothesis; H_0 can be rejected for any reason

- A person may be declared suspect for **any** incriminating reason (e.g., obstruction)
- There is no minimum level for the difference in population (arbitrary precision)

Fisher thinks like a detective

Null hypothesis (H_0) is the default position (claims innocence):

- This person's actions **are NOT** incriminating
- The difference in population means is ZERO

There is **no** specific alternative hypothesis; H_0 can be rejected for any reason

- A person may be declared suspect for **any** incriminating reason (e.g., obstruction)
- There is no minimum level for the difference in population (arbitrary precision)

The test computes a *p*-value, which measures this likelihood:

Prob(evidence | H_0 is true)

i.e., what percentage of innocent people behave this way?

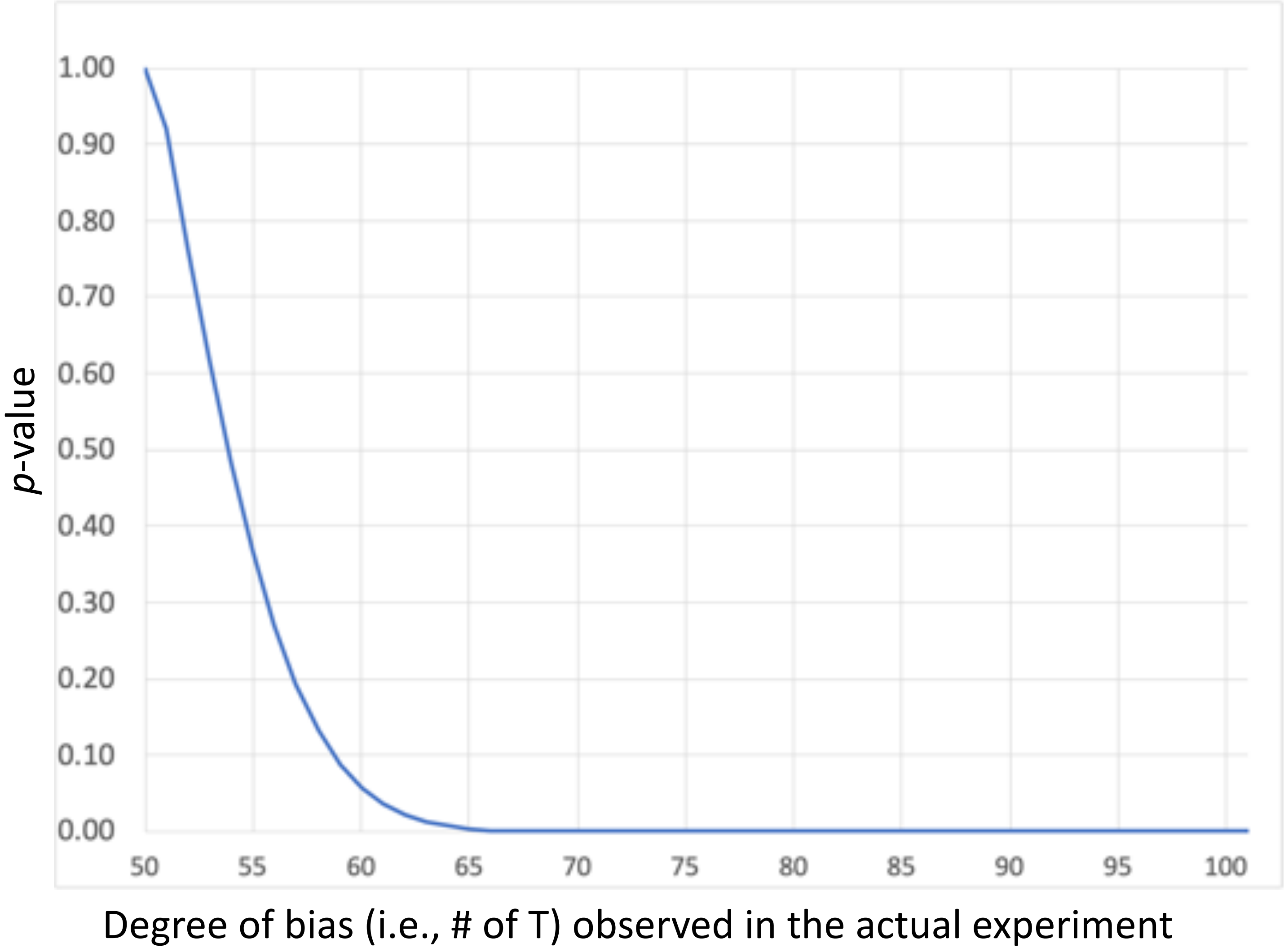
Time for a thought experiment

H_0: Coin is fair, meaning Prob(H) = Prob(T)

Experiment: 100 flips

p-value is the proportion of experiments that would produce a specific degree of bias (i.e., # of T), or more

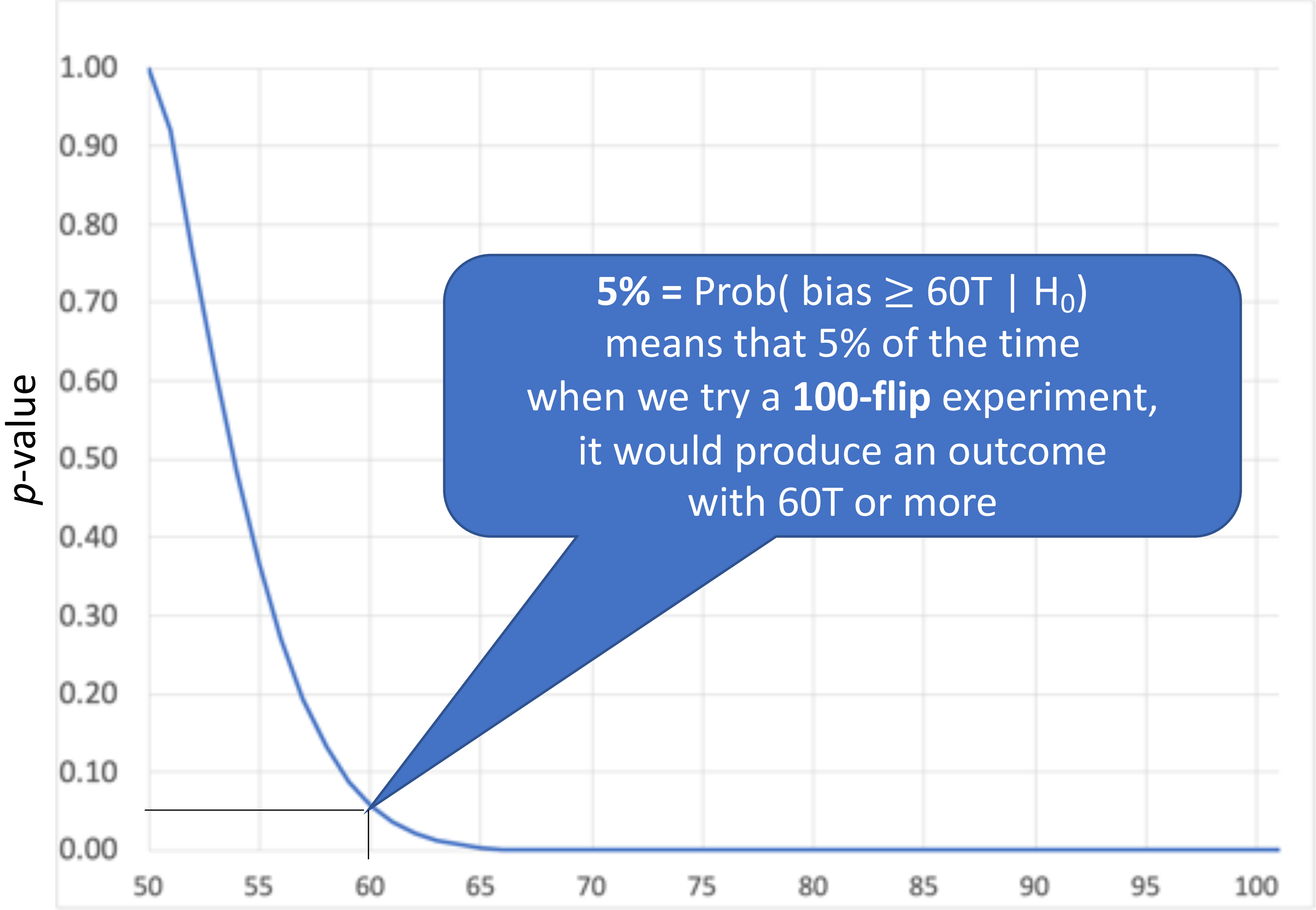
Also known as false alarms



H_0: Coin is fair, meaning Prob(H) = Prob(T)

Experiment: 100 flips

p-value is the proportion of experiments that would produce a specific degree of bias (i.e., # of T), or more



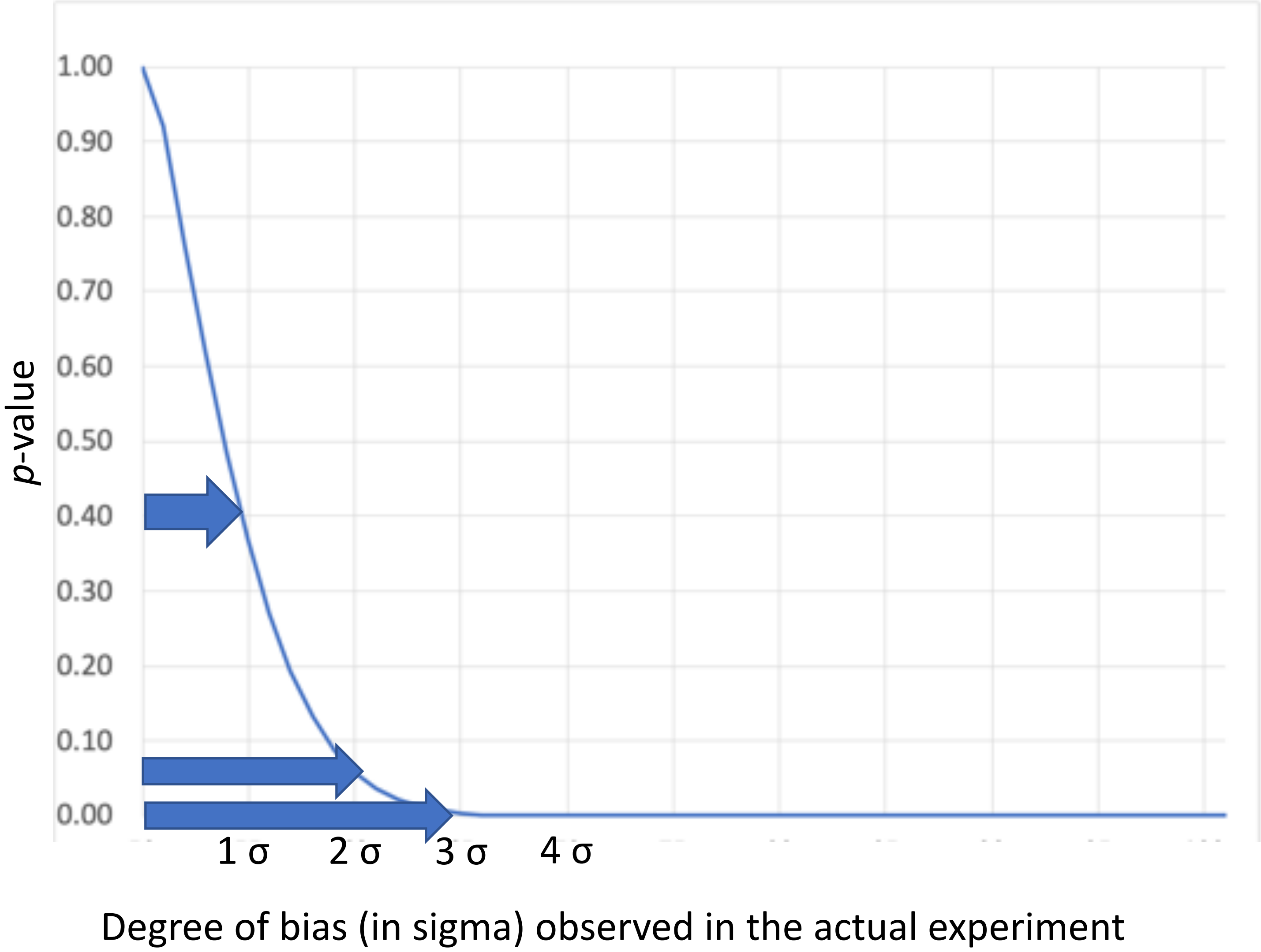
Degree of bias (i.e., # of T) observed in the actual experiment

H_0: Coin is fair, meaning Prob(H) = Prob(T)

Experiment: 100 flips

p-value is the proportion of experiments that would produce a specific degree of bias (i.e., # of T), or more

Significance is the # of standard deviations that correspond to that degree of bias (measured in *sigma*)

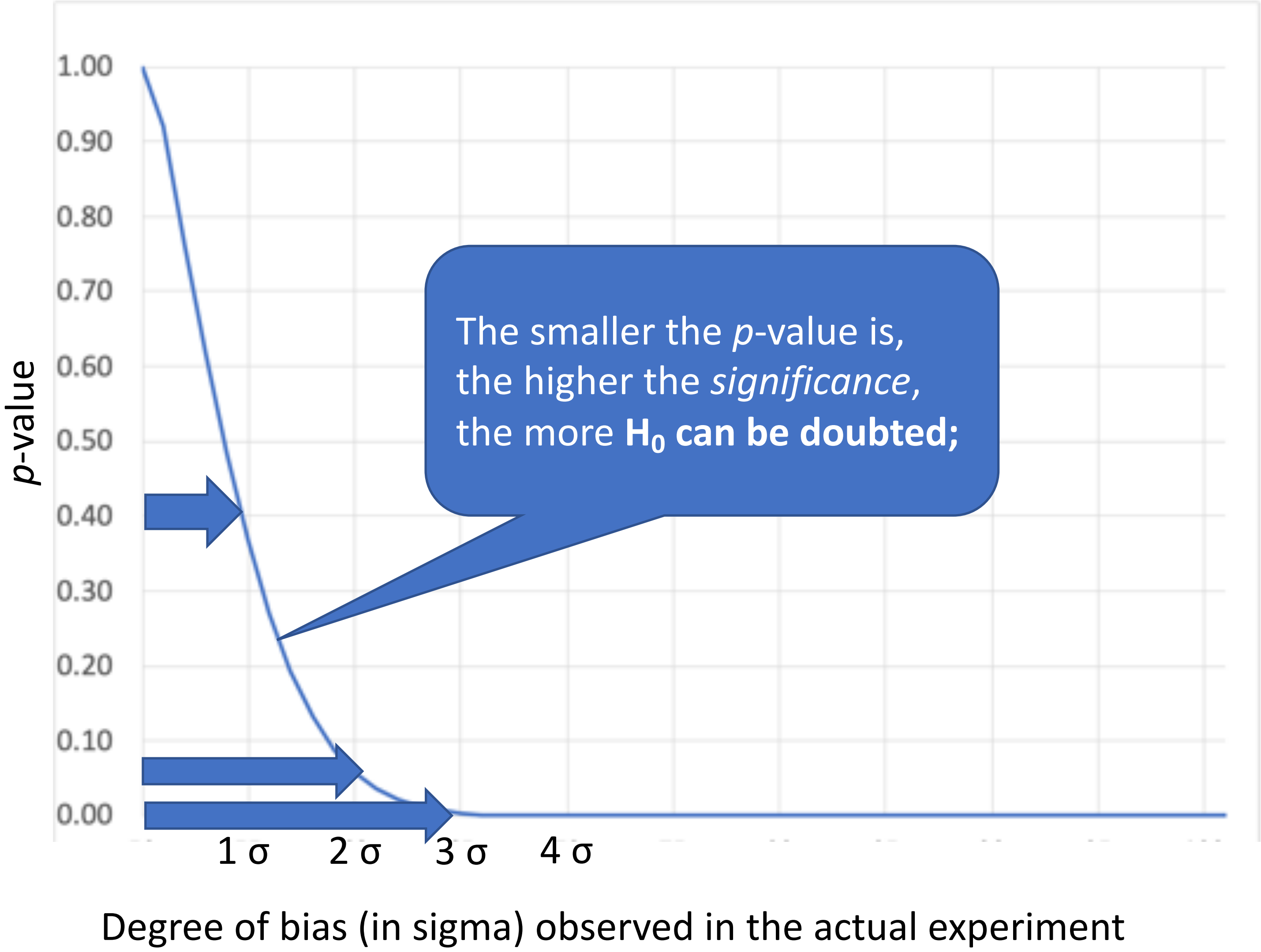


H_0: Coin is fair, meaning Prob(H) = Prob(T)

Experiment: 100 flips

p-value is the proportion of experiments that would produce a specific degree of bias (i.e., # of T), or more

Significance is the # of standard deviations that correspond to that degree of bias (measured in *sigma*)

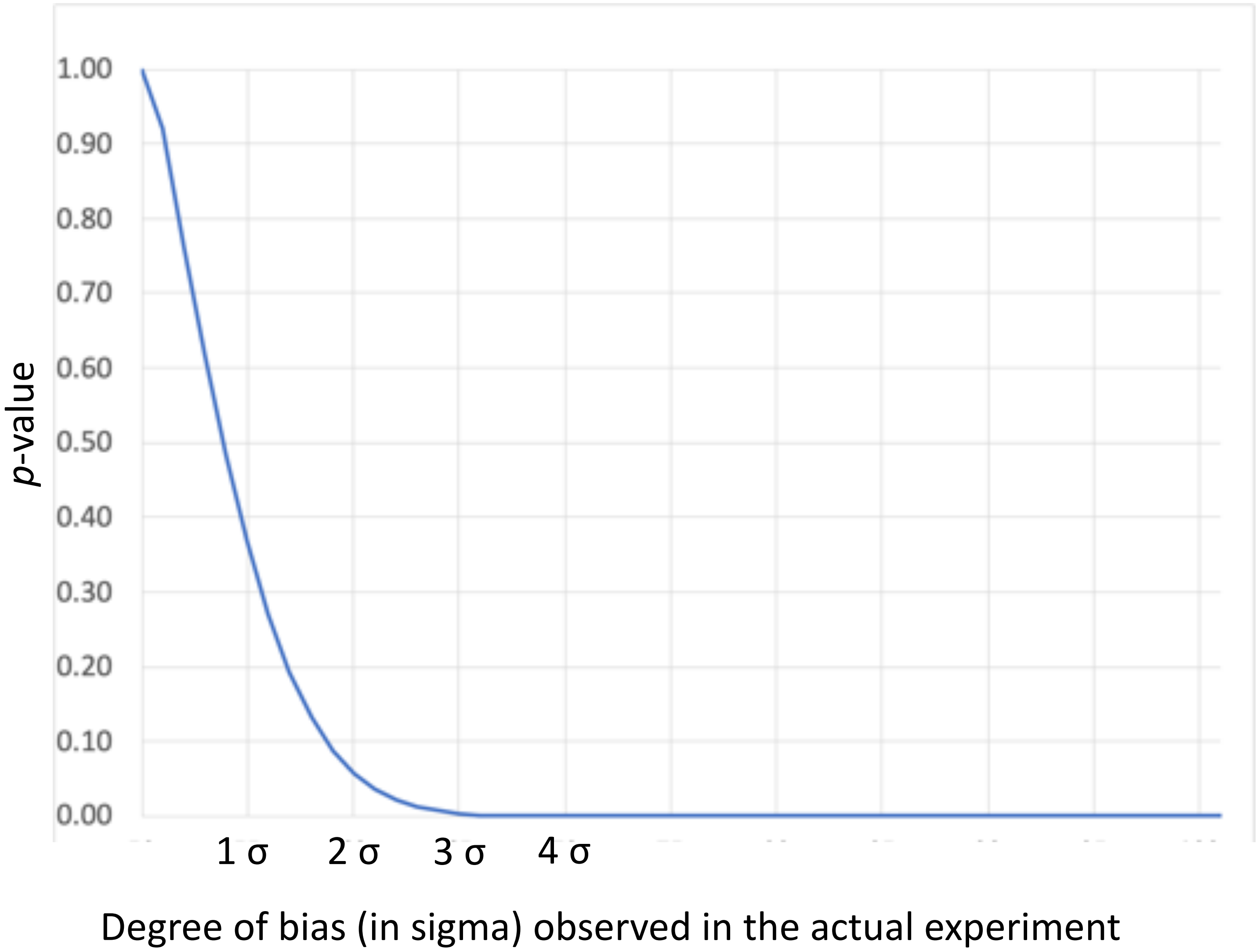


H_0: Coin is fair, meaning Prob(H) = Prob(T)

p-value and *significance* are an **inversely** related pair

The higher the *significance*, the lower the *p*-value that corresponds to it

They both depend on H_0 being true



Not all detectives think alike

A smaller *p*-value is always more significant (more cause for doubt, or less risk)
but different people/fields have different risk tolerance

- Opinion Polls are very risk tolerant: $p = 0.10$ means I've seen enough;
  it's well beyond the margin of error level (one sigma)

- Physics does not like risk: $p = 0.04$ means I'm not remotely convinced;
  it's barely past two sigma, not even close to five sigma

Given *tolerance* // e.g., 0.05 for (95% ↔ two sigma)

Compute *p*

IF *p* < *tolerance*

    // either guilty
    // or rare (based on significance_level) coincidence,
    // reject H_0

    declare suspect

ELSE

    seek more evidence OR close case

WRONG

Compute *p*

Compute *significance_level*

  // e.g., p = 0.05 means significance_level = 1.96 sigma

Print “H_0 may be rejected at significance level:”

Print *significance_level*

The real issue with Fisher's thinking

The real issue with Fisher's thinking

In this hypothetical example, the *p*-value is the ratio from the green row:

$$p = \frac{4}{96+4} = 0.04$$

	Acting suspiciously? NO	Acting suspiciously? YES
Guilty? NO	96	4

The *p*-value only measures this likelihood:

$$\mathit{Prob}(\text{ evidence } \mid H_0 \text{ is true })$$

The real issue with Fisher's thinking

In this hypothetical example, the *p*-value is the ratio from the green row:

$$p = \frac{4}{96+4} = 0.04$$

But we should be in the business of looking at the red row!

	Acting suspiciously? NO	Acting suspiciously? YES
Guilty? NO	96	4
Guilty? YES	?	?

Hypothesis testing should be a **binary classifier** algorithm

	Acting suspiciously? NO	Acting suspiciously? YES
Guilty? NO	True Negative	False Positive
Guilty? YES	False Negative	True Positive

J. Neyman and E. Pearson

- Think like lawyers
- Want to distinguish innocence from guilt
- Perform binary classification

A binary classification algorithm

	Acting suspiciously? NO	Acting suspiciously? YES
Guilty? NO	True Negative	False Positive
Guilty? YES	False Negative	True Positive

The p-value is the ratio: $\frac{false\ positive}{not\ guilty}$

AKA

- the probability of False Alarm,
- False Positive Rate (FPR)

We want this percentage to be **small**

(ideally it would be 0%)

A binary classification algorithm

	Acting suspiciously? NO	Acting suspiciously? YES
Guilty? NO	True Negative	False Positive
Guilty? YES	False Negative	True Positive

But we also need assumptions about the ratio:

$$\frac{true\ positive}{guilty}$$

AKA the probability of detection, or

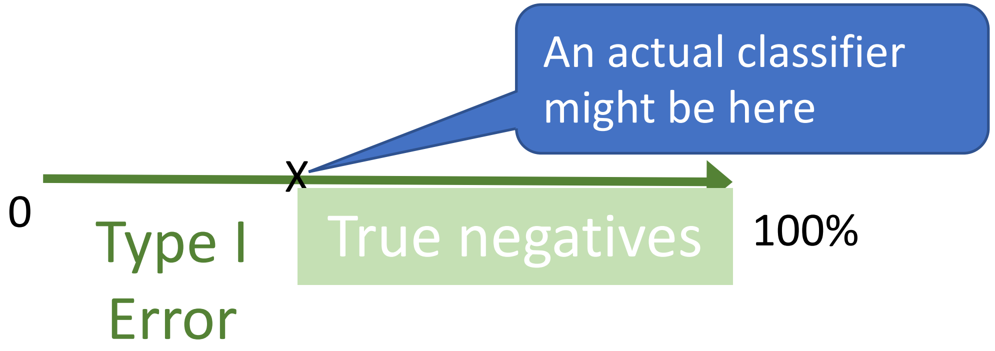
- True positive rate (TPR)
- recall
- sensitivity
- hit rate
- power

We want this percentage to be **large**

(ideally it would be 100%)

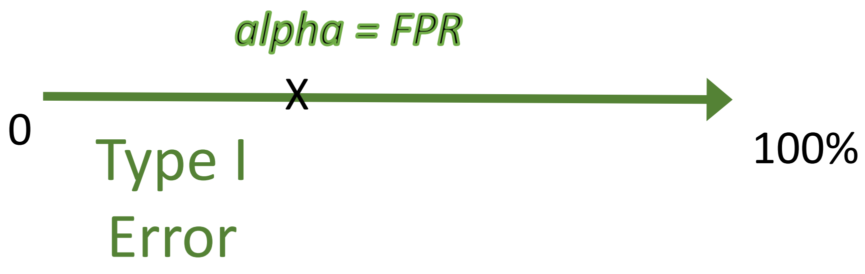
A binary classification algorithm

	Acting suspiciously? NO	Acting suspiciously? YES
Guilty? NO	True Negative	False Positive
Guilty? YES	False Negative	True Positive



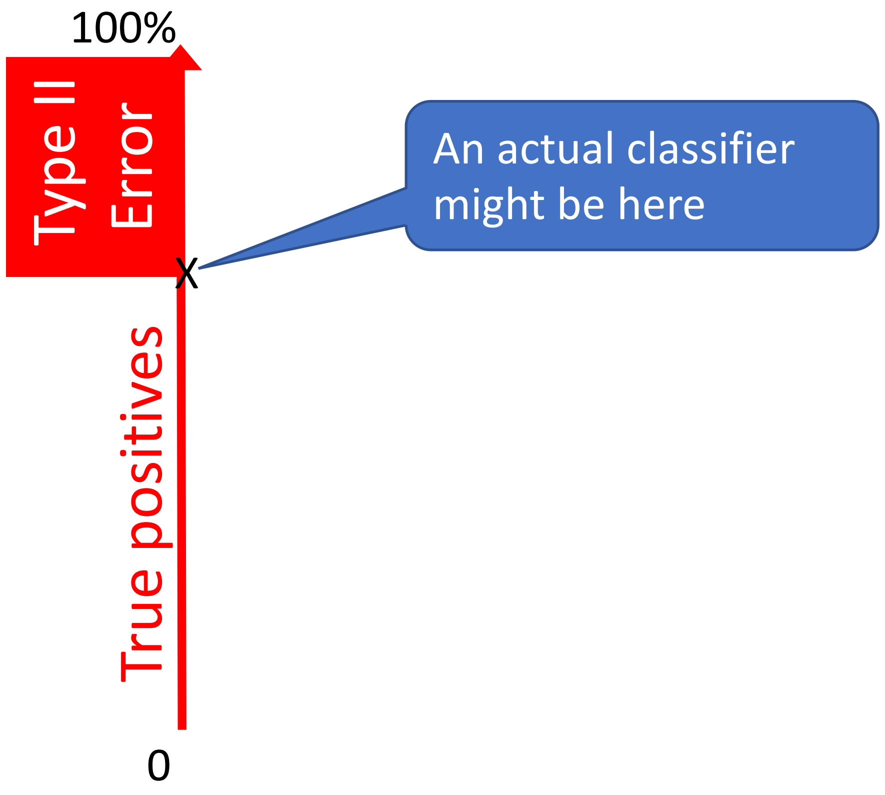
A binary classification algorithm

	Acting suspiciously? NO	Acting suspiciously? YES
Guilty? NO	True Negative	False Positive
Guilty? YES	False Negative	True Positive



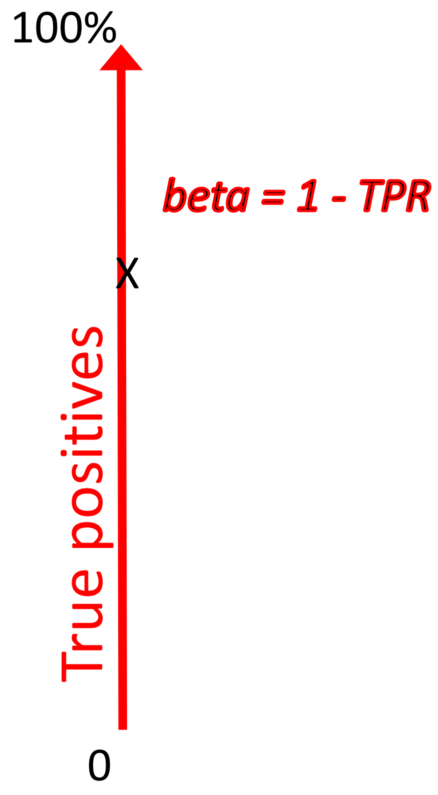
A binary classification algorithm

	Acting suspiciously? NO	Acting suspiciously? YES
Guilty? NO	True Negative	False Positive
Guilty? YES	False Negative	True Positive



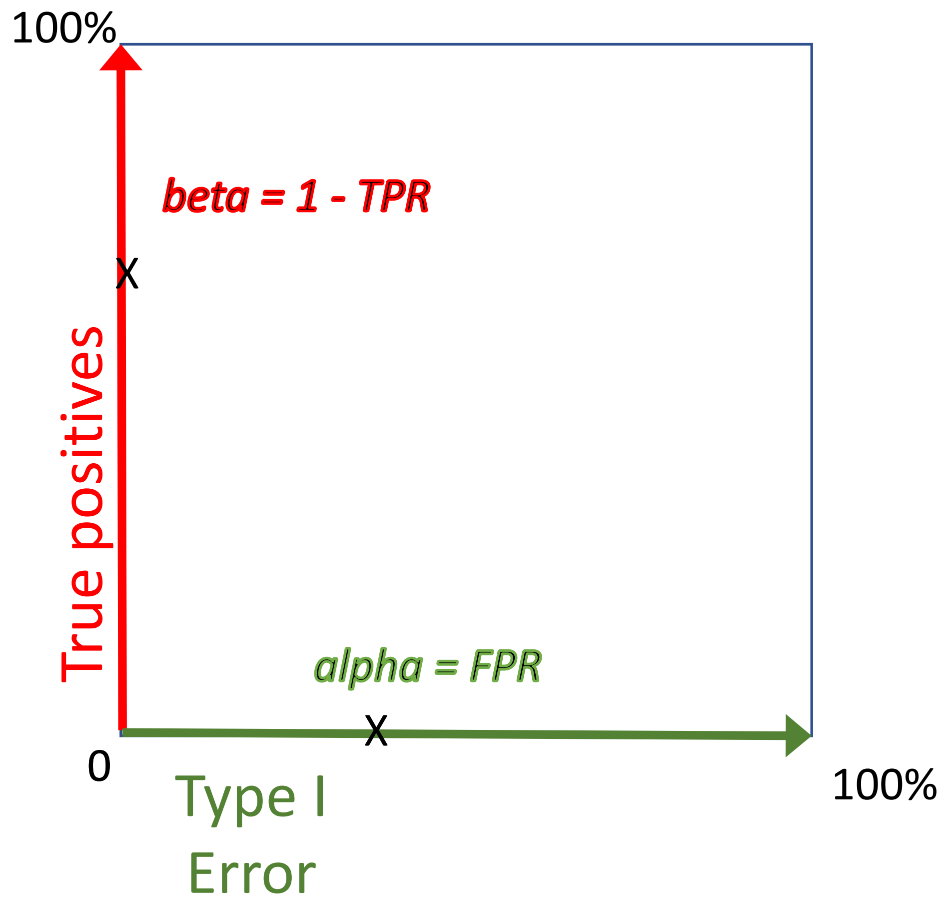
A binary classification algorithm

	Acting suspiciously? NO	Acting suspiciously? YES
Guilty? NO	True Negative	False Positive
Guilty? YES	False Negative	True Positive



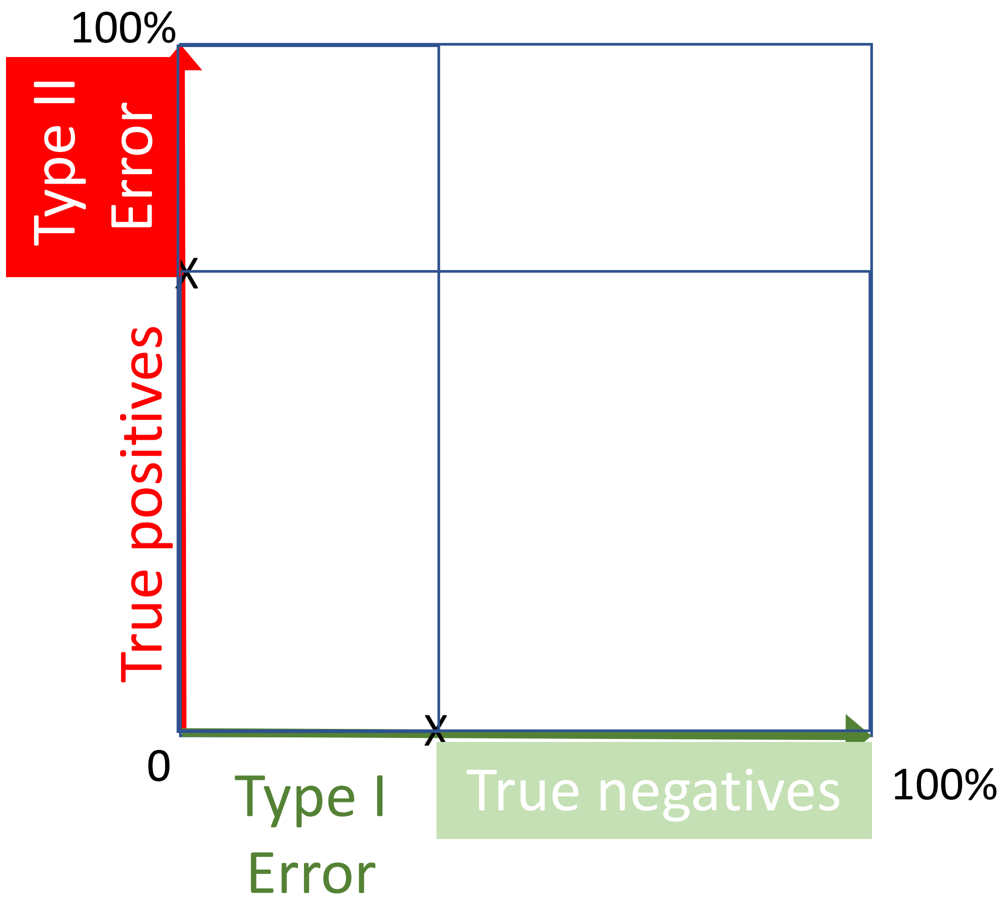
A binary classification algorithm

	Acting suspiciously? NO	Acting suspiciously? YES
Guilty? NO	True Negative	False Positive
Guilty? YES	False Negative	True Positive



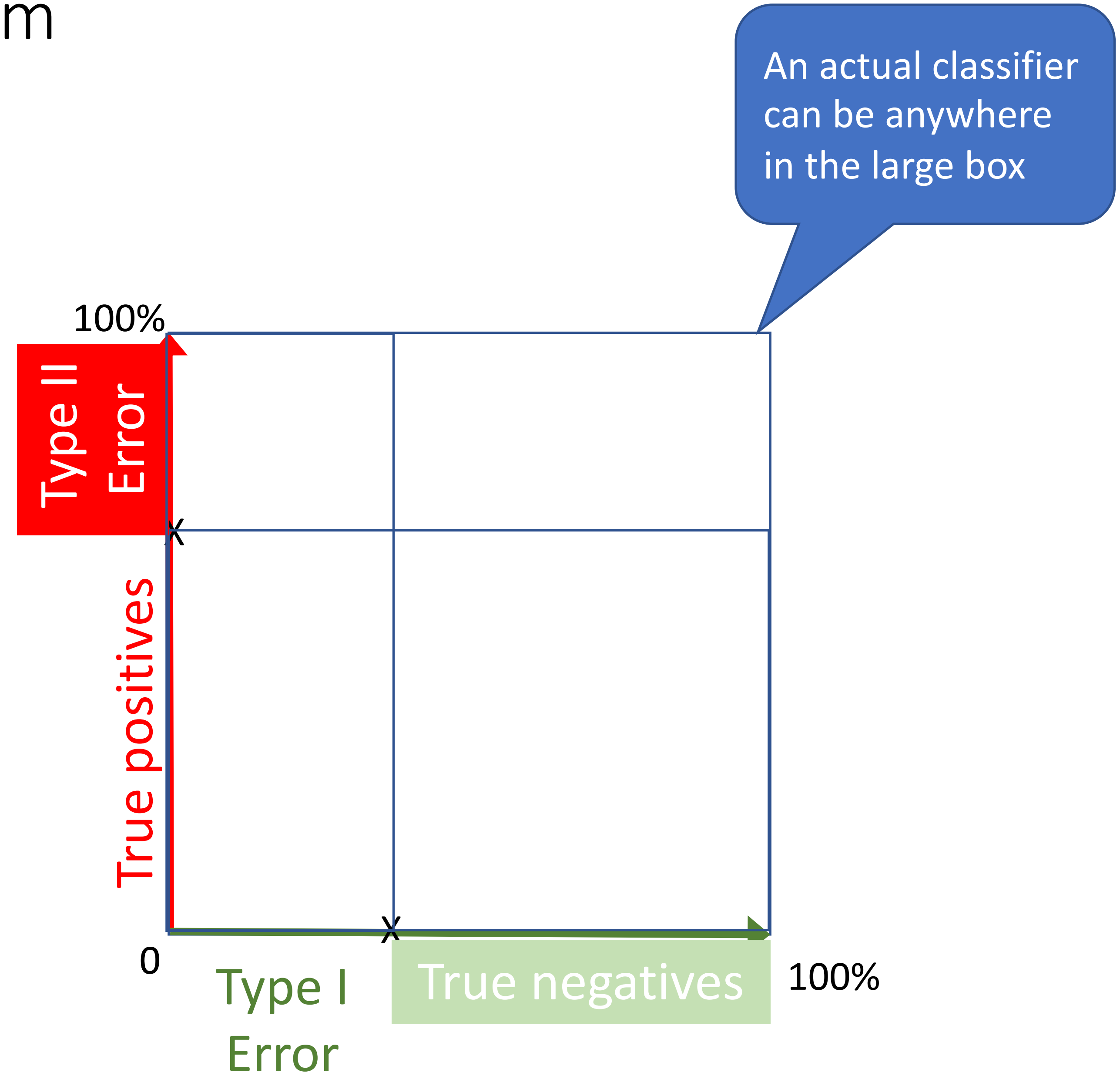
A binary classification algorithm

	Acting suspiciously? NO	Acting suspiciously? YES
Guilty? NO	True Negative	False Positive
Guilty? YES	False Negative	True Positive



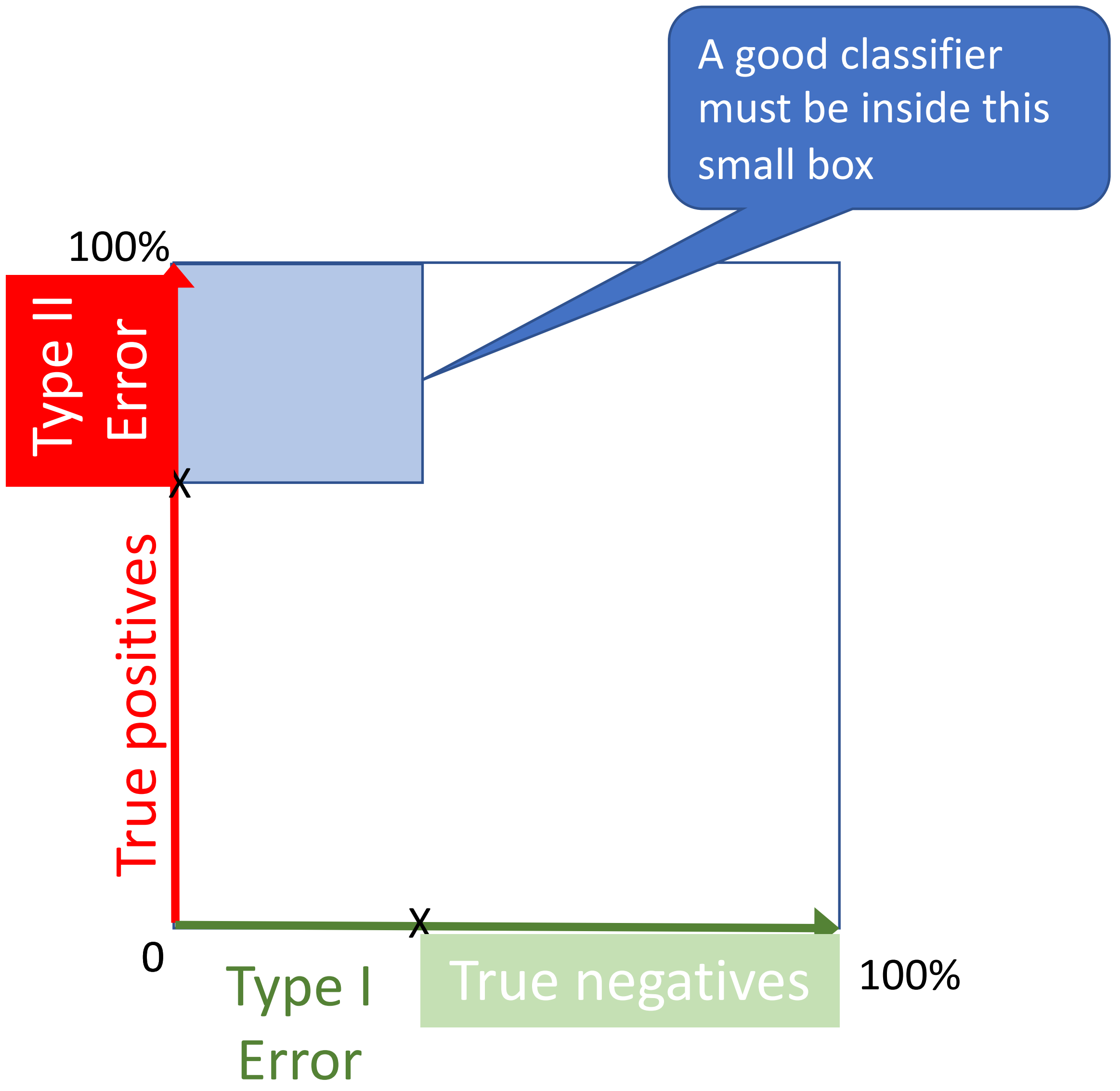
A binary classification algorithm

	Acting suspiciously? NO	Acting suspiciously? YES
Guilty? NO	True Negative	False Positive
Guilty? YES	False Negative	True Positive



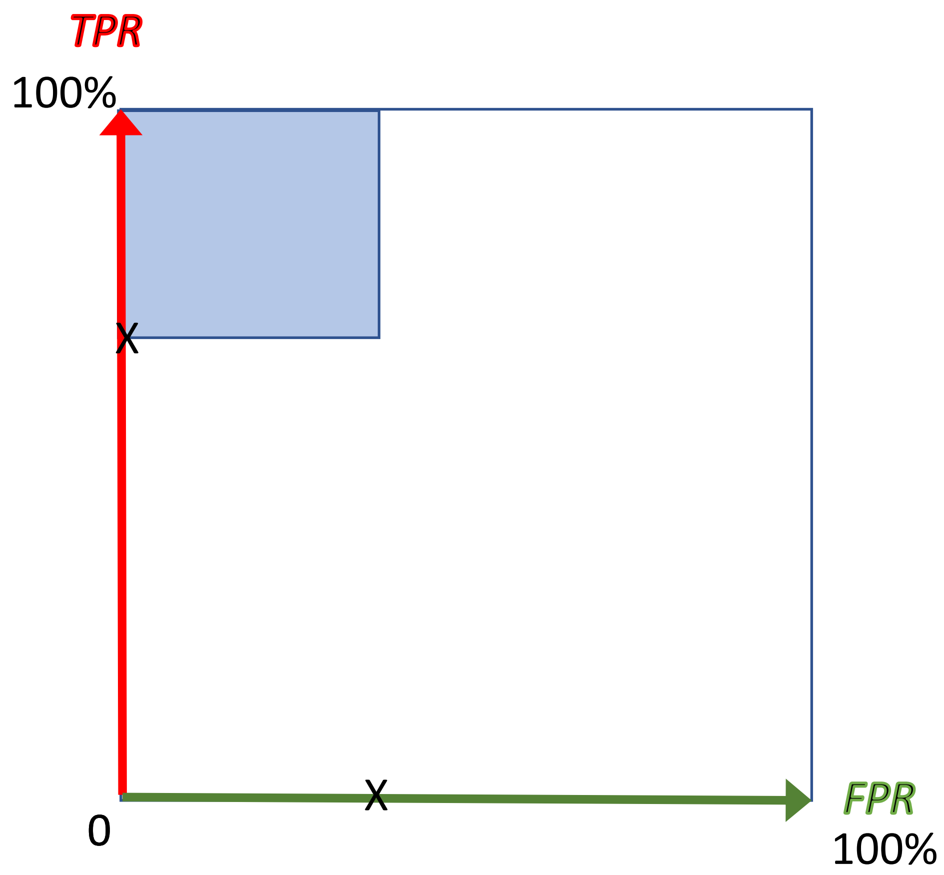
A binary classification algorithm

	Acting suspiciously? NO	Acting suspiciously? YES
Guilty? NO	True Negative	False Positive
Guilty? YES	False Negative	True Positive



A binary classification algorithm

	Acting suspiciously? NO	Acting suspiciously? YES
Guilty? NO	True Negative	False Positive
Guilty? YES	False Negative	True Positive



Recall: Fisher thinks like a detective

Null hypothesis (H_0) is the default position (claims innocence):

- This person’s actions **are NOT** incriminating
- There is no minimum level for the difference in population (arbitrary precision)

Contrast: Neyman and Pearson think like lawyers

Null hypothesis (H_M) is the main/default position (presumed innocent):

- The prosecution has **NOT** proved guilt beyond **reasonable doubt**
- The difference in population means (effect size) is **NOT** above a MINIMUM LEVEL (fixed precision)

Contrast: Neyman and Pearson think like lawyers

Null hypothesis (H_M) is the main/default position (presumed innocent):

- The prosecution has **NOT** proved guilt beyond **reasonable doubt**
- The difference in population means (effect size) is **NOT** above a MINIMUM LEVEL (fixed precision)
- The specific treatment being tested did **NOT** produce a **detectable** effect

Recall: Fisher thinks like a detective

There is **no** specific alternative hypothesis; H_0 can be rejected for any reason

- A person may be declared suspect for **any** incriminating reason (e.g., obstruction)
- The difference in population means can be arbitrarily small!

Contrast: Neyman and Pearson think like lawyers

There is a **specific alternative** hypothesis H_A (guilty **as charged**):

- The prosecution **has** proved the charges beyond **reasonable doubt**
- The difference in population means is ABOVE a MINIMUM LEVEL (fixed precision)

Contrast: Neyman and Pearson think like lawyers

There is a **specific alternative** hypothesis H_A (guilty **as charged**):

- The prosecution **has** proved the charges beyond **reasonable doubt**
- The difference in population means is ABOVE a MINIMUM LEVEL (fixed precision)
- The specific treatment being tested **INDID** produced a **detectable** effect

Recall: Fisher thinks like a detective

The *p*-value only measures this likelihood:

Prob(evidence | H_0 is true)

Contrast: Neyman and Pearson think like lawyers

We must compute the unique *p*-value cutoff defined by the trade-off between

$Prob$(evidence | H_M is true)

and

$Prob$(evidence | H_A is true)

(the precision level)

Decision Time

Detection error tradeoff (DET)

Causes two different types of possible error

- **Mistaken detection**: the effect was above the minimum level, but it was not produced by the treatment / wrongful conviction (Type I error)
- **Missed detection:** the treatment produced an effect, but it was not above the minimum level / guilty yet acquitted (Type II error)

Detection error tradeoff (DET)

Causes two different types of possible error

And these two errors depend on each other

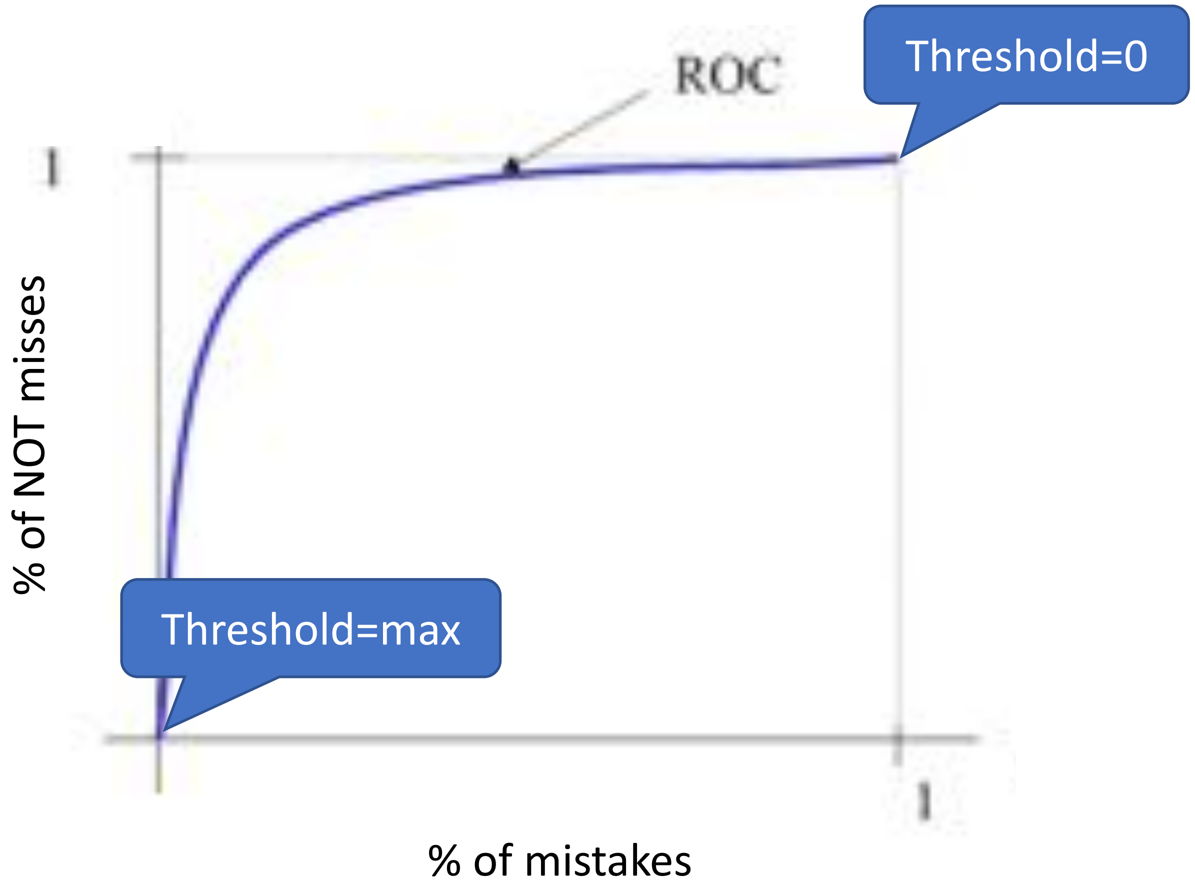
- Minimum precision = 0 means that everyone will be convicted
  - no missed detections (power=100%) AND
  - maximum mistaken detections
- As the minimum precision threshold increases,
  - more guilty people will walk scot-free (less power), but ALSO
  - fewer innocent people will be convicted (mistakes)
- If the minimum precision threshold is high enough,
  - all guilty people will be acquitted / no power, because no jury trial will result in a conviction (reasonable doubt becomes unreasonably lax)

Detection error tradeoff (DET)

Any given threshold corresponds to a specific pair of values for

- % of mistakes (p-value) and
- % of power (omissions)

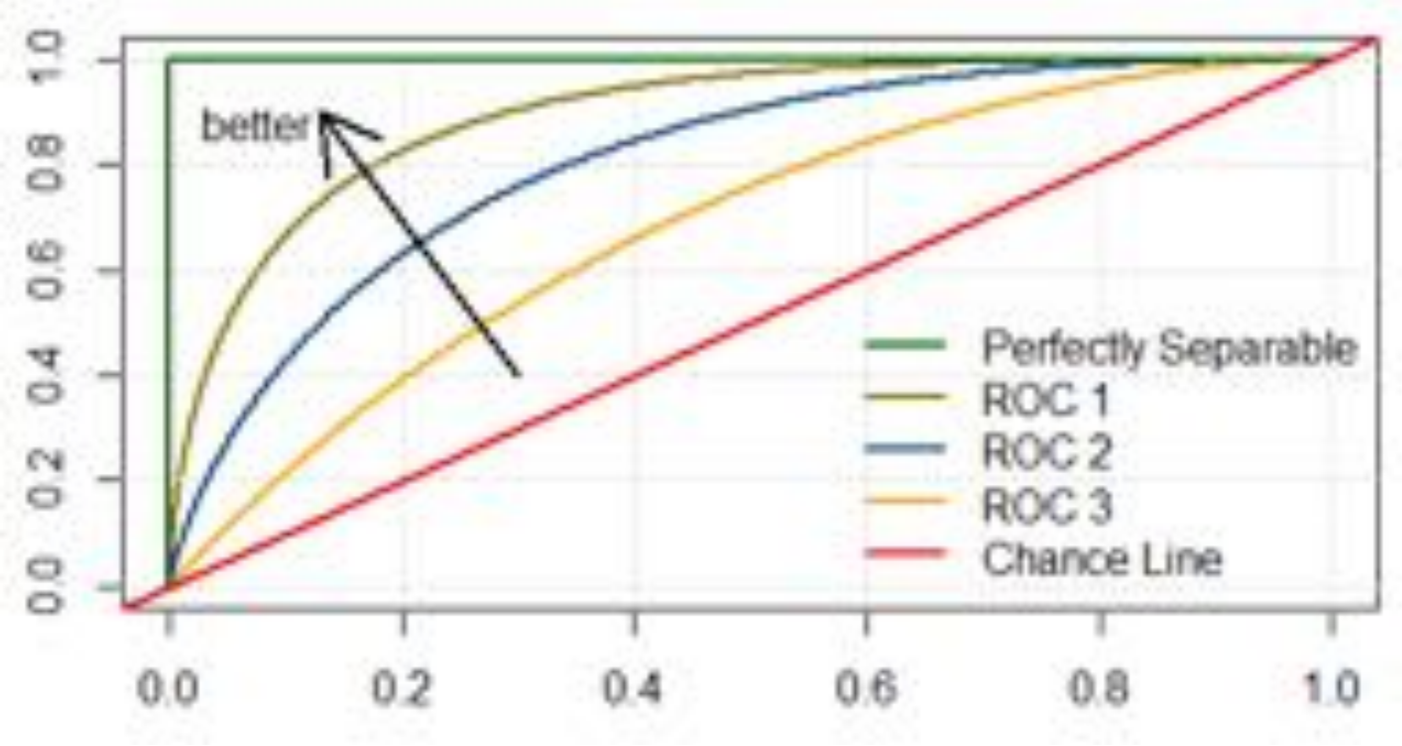
these two rates are not independent, but fall along a curve (sometimes called ROC)



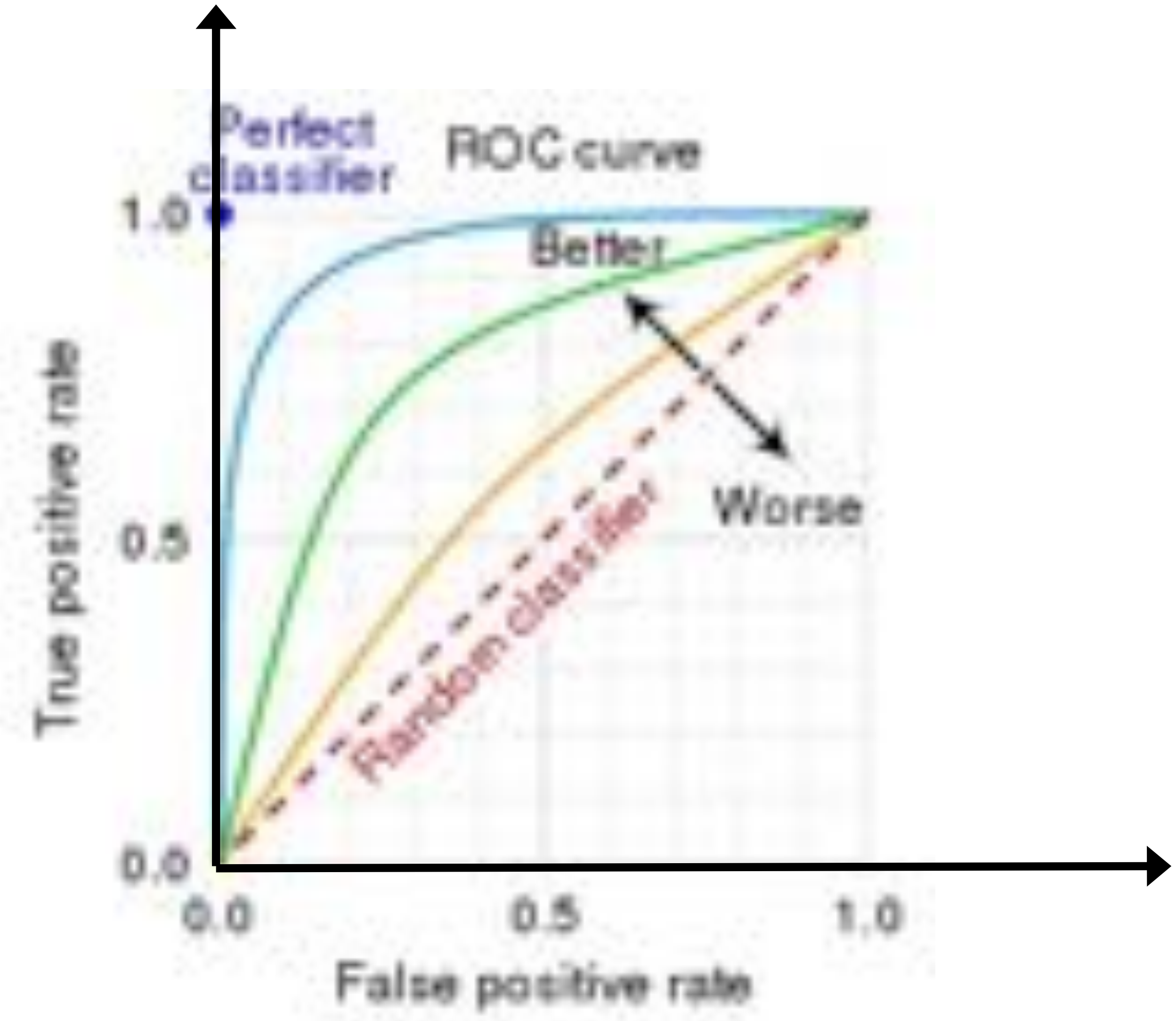
The power of sample size

More samples (higher N) lead to better DET/ROC curves

- Higher power for given *p*-value
- Lower *p*-value for given power



The power of sample size



$\frac{true\ positive}{positive}$ is the

- Positive Predictive Value (PPV)
- *Precision*

	Acting suspiciously? NO	Acting suspiciously? YES
Guilty? NO	True Negative	False Positive
Guilty? YES	False Negative	True Positive

Back to the jury

Contrast: Neyman and Pearson think like lawyers

Fix values for

- ***alpha,*** the long-term probability of **mistaken** detection (Type I error):
  - wrongful conviction, or
  - falsely accepting the alternative/**prosecution's** argument H_A
- ***beta,*** the long-term probability of **missed** detection (Type II error):
  - guilty yet acquitted, or
  - falsely accepting the main/**defense's** argument H_M
- keep *beta* > *alpha*

Contrast: Neyman and Pearson think like lawyers

With *alpha* and *beta* computed,

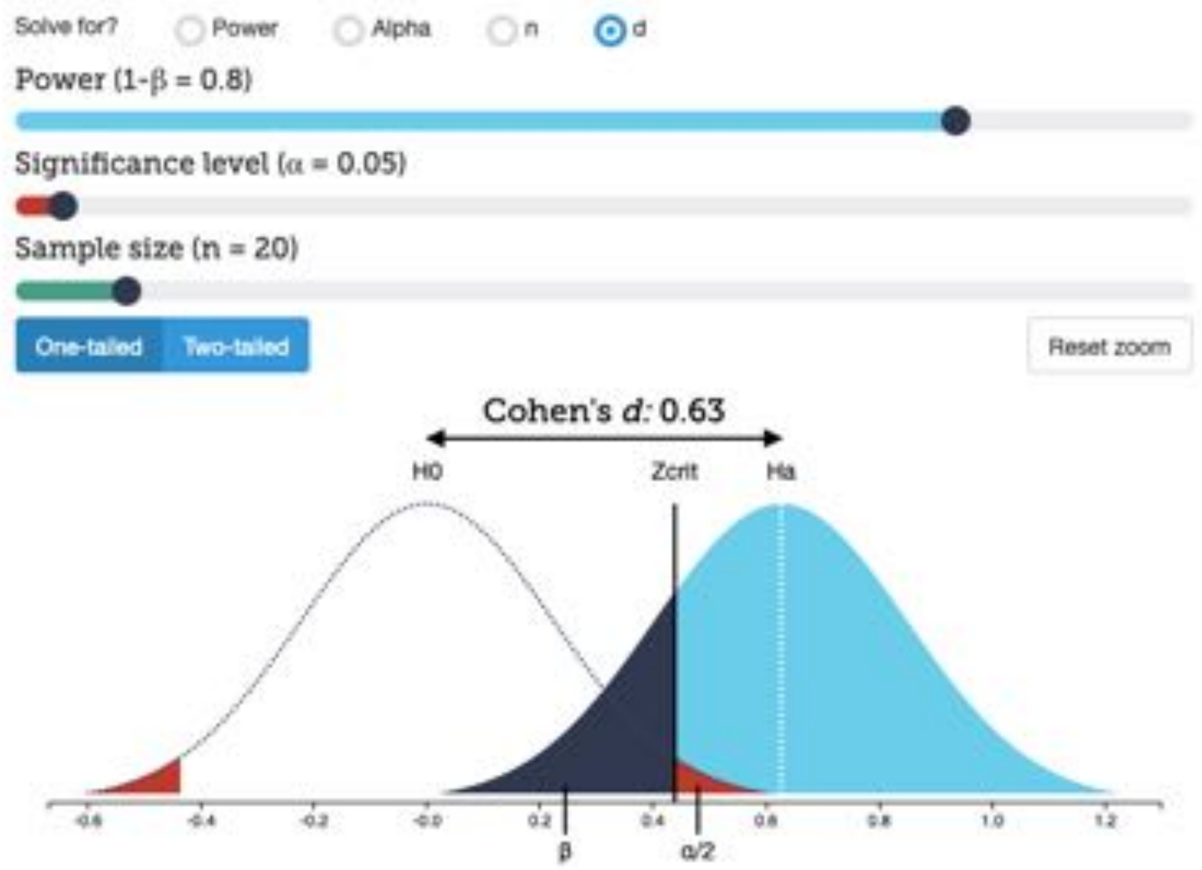
- Set the **fixed** threshold for p-value to be *alpha*
- Use *beta* to compute the **fixed** power value that reflects the sample size

(Note that *power* = 1 – *beta*)

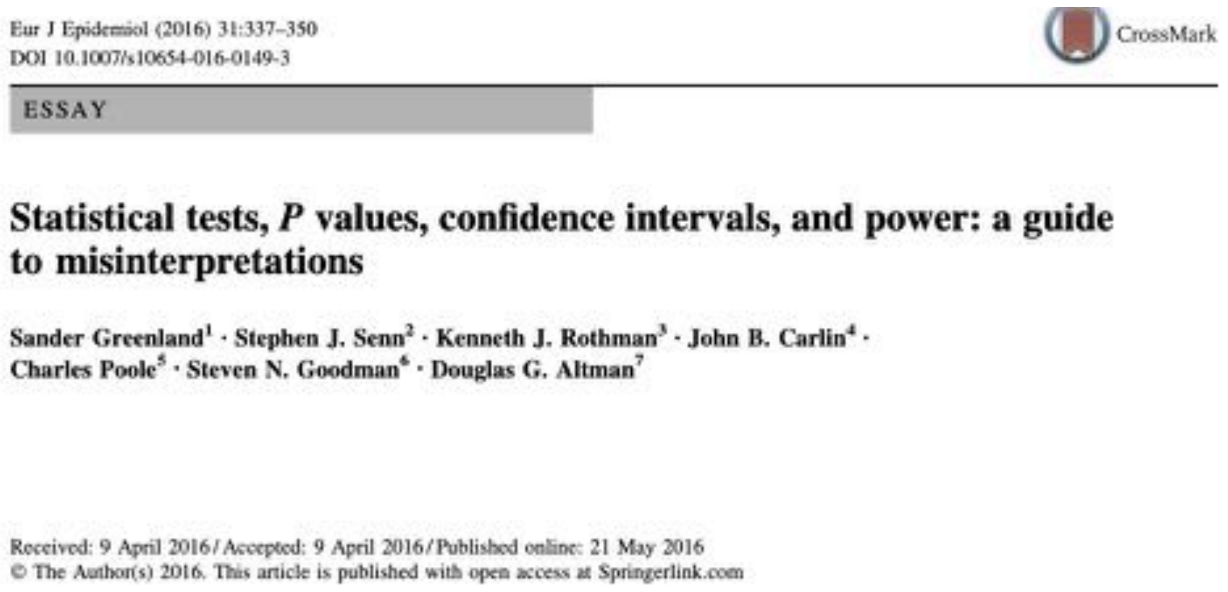
Neyman and Pearson think like lawyers

```
Given alpha < beta
Compute p
Compute power   // power is based on amount of evidence (sample size)
IF power < 1 – beta
   // not enough evidence was presented either way, so inconclusive
   warning “TEST LACKS SUFFICIENT POWER TO MAKE RELIABLE DECISIONS”
   // but prosecution has the burden of proof
   accept H_M
ELSE
   IF p < alpha
      accept H_A   // enough incriminating evidence was presented, so find guilty
   ELSE
      accept H_M   // enough exonerating evidence was presented, so find innocent
```

For more on this topic:



https://rpsychologist.com/d3/nhst/



Does

- ✓the statistical summary say what you *think* it says?
- ✓the statistical summary give the *full* picture?
- ✓the statistical test ask the *right* question?
- ✓the statistical test say what you *think* it says?


That's all Folks!