Some Problems With *p*-values and Null Hypothesis Significance Testing

Tim Frasier

Prior Homework

Prior Homework

- Have students read the following:

1. Gigerenzer (2004) Mindless statistics. *The Journal of Socio-Economics* **33**: 587-606.

2. Halsey *et al*. (2015) The fickle *P* value generates irreproducible results. *Nature Methods* **12**: 179-185.

3. Hoekstra *et al*. (2014) Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review* **21**: 1157-1164.

4. Hauer (2004) The harm done by tests of significance. *Accident Analysis and Prevention* **36**: 495-500.

Is Confusing:

Do you know what you're actually calculating?

What Does a *p*-value Mean?

What Does a *p*-value Mean?

- The probability of obtaining a difference as great, or greater, between observed and expected results *if the null hypothesis is true*, and the experiment repeated many times

What Does a *p*-value Mean?

- The probability of obtaining a difference as great, or greater, between observed and expected results *if the null hypothesis is true*, and the experiment repeated many times

`p = 0.05`

Would expect a difference as big (or bigger) only ~5% of the time *if the null hypothesis is true, and the experiment repeated in the same way many times*

What Does a *p*-value Mean?

- The probability of obtaining a difference as great, or greater, between observed and expected results *if the null hypothesis is true*, and the experiment repeated many times

Not:

- The probability that the null hypothesis is true
- The probability that you are wrong (or right)

What Does a *p*-value Mean?

- The probability of obtaining a difference as great, or greater, between observed and expected results *if the null hypothesis is true*, and the experiment repeated many times

Mathematically:

$$P(\mathrm{D}|\mathrm{H_o}) \qquad \textbf{Not} \qquad P(\mathrm{H}|\mathrm{D})$$

What Does a *p*-value Mean?

- The probability of obtaining a difference as great, or greater, between observed and expected results *if the null hypothesis is true*, and the experiment repeated many times

Mathematically:

$P(\mathrm{D}|\mathrm{H_o})$ **Not** $P(\mathrm{H}|\mathrm{D})$

This is what we're interested in, no?

What Does a *p*-value Mean?

Rejecting one hypothesis (the null) does not mean that your alternative hypothesis is correct!

- In fact, it tells you *little* about the probability of your true hypothesis

There are *an infinite number* of other hypotheses that have been *no more or less supported* based on this single test

Example

$$P(\mathrm{D}|\mathrm{H_o}) \qquad \textbf{Not} \qquad P(\mathrm{H}|\mathrm{D})$$

Fallacy: Affirming the consequent

1. If P then Q
2. Q
3. Therefore P

Example

$P(\mathrm{D}|\mathrm{H_o})$ **Not** $P(\mathrm{H}|\mathrm{D})$

Fallacy: Affirming the consequent

1. If Bill Gates owns Fort Knox, then he is rich
2. Bill Gates is rich
3. Therefore, Bill Gates must own Fort Knox

Example

$P(\mathrm{D}|\mathrm{H_o})$ **Not** $P(\mathrm{H}|\mathrm{D})$

Fallacy: Affirming the consequent

Rejecting the null hypothesis does not mean your specific alternative (in mind) is true (there could be many alternative explanations)

Example

$$P(\mathrm{D}|\mathrm{H_o}) \qquad \textbf{Not} \qquad P(\mathrm{H}|\mathrm{D})$$

- Hypothesis: Tim is a king
- Data: Tim is a male

Example

$$P(\mathrm{D}|\mathrm{H_o}) \qquad \textbf{Not} \qquad P(\mathrm{H}|\mathrm{D})$$

- Hypothesis: Tim is a king
- Data: Tim is a male

$$P(\mathrm{D}|\mathrm{H_o}) = 1$$

All kings are males.
If Tim is a king, he is definitely a male

Example

$$P(\mathrm{D}|\mathrm{H_o}) \qquad \textbf{Not} \qquad P(\mathrm{H}|\mathrm{D})$$

- Hypothesis: Tim is a king
- Data: Tim is a male

$$P(\mathrm{D}|\mathrm{H_o}) = 1$$

All kings are males.
If Tim is a king, he is definitely a male

$$P(\mathrm{H}|\mathrm{D}) \neq 1$$

The fact that Tim is a male tells us almost nothing about the probability that he is a king

Example

Practice:

If H_0 is true, then this result would probably not occur.

This result has occurred.

Then H_0 is probably not true.

The usual logic, no?

From Cohen (1994) *American Psychologist* **49**: 997-1003.

Example

Practice:

If H_0 is true, then this result would probably not occur.

This result has occurred.

Then H_0 is probably not true.

In Real Life:

If a person is an American, then they are probably not a member of congress.

This person is a member of congress.

Then they are probably not an American.

Huh?

From Cohen (1994) *American Psychologist* **49**: 997-1003.

Confidence Intervals Not Much Better

- What does a 95% confidence interval tell you?

Confidence Intervals Not Much Better

- What does a 95% confidence interval tell you?

- If we were to repeat the experiment over and over, then 95% of the time the confidence intervals would contain the true mean[1,2]

Huh?

- Estimated values from a single experiment mean very little
- **No** information on where in that range the true value likely is!

1. Hoekstra et al. (2014) *Psychon. Bull. Rev.* **21**: 1157-1164
2. Only if data are good representation of underlying patterns, which who knows?

Confidence Intervals Not Much Better

- What does a 95% confidence interval tell you?

- If we were to repeat the experiment over and over, then 95% of the time the confidence intervals would contain the true mean[1,2]

Huh?

- Estimated values from a single experiment mean very little
- **No** information on where in that range the true value likely is!

Not:

- You are 95% confident that the truth lies within this range

1. Hoekstra et al. (2014) *Psychon. Bull. Rev.* **21**: 1157-1164
2. Only if data are good representation of underlying patterns, which who knows?

Confidence Intervals Not Much Better

- Effects on viability of heart muscle after a heart attack

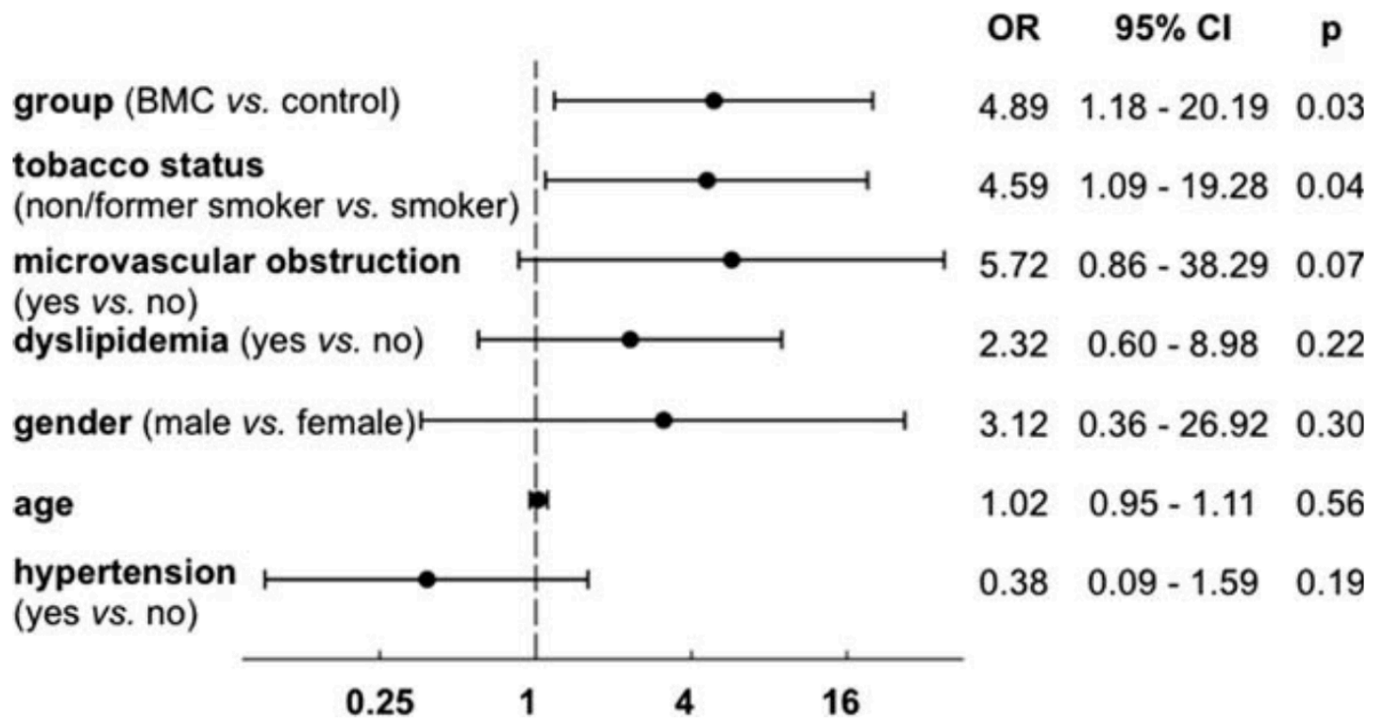


Figure 3 Multivariate logistic regression analysis for improvement of at least two non-viable segments becoming viable ($n = 77$). OR, odds ratio; CI, confidence interval; p: *P*-value.

From Roncalli et al. (2011) *European Heart Journal* **32**: 1748-1757.

Confidence Intervals Not Much Better

- Effects on viability of heart muscle a

Can't rule out 1 (no effect), so not significant

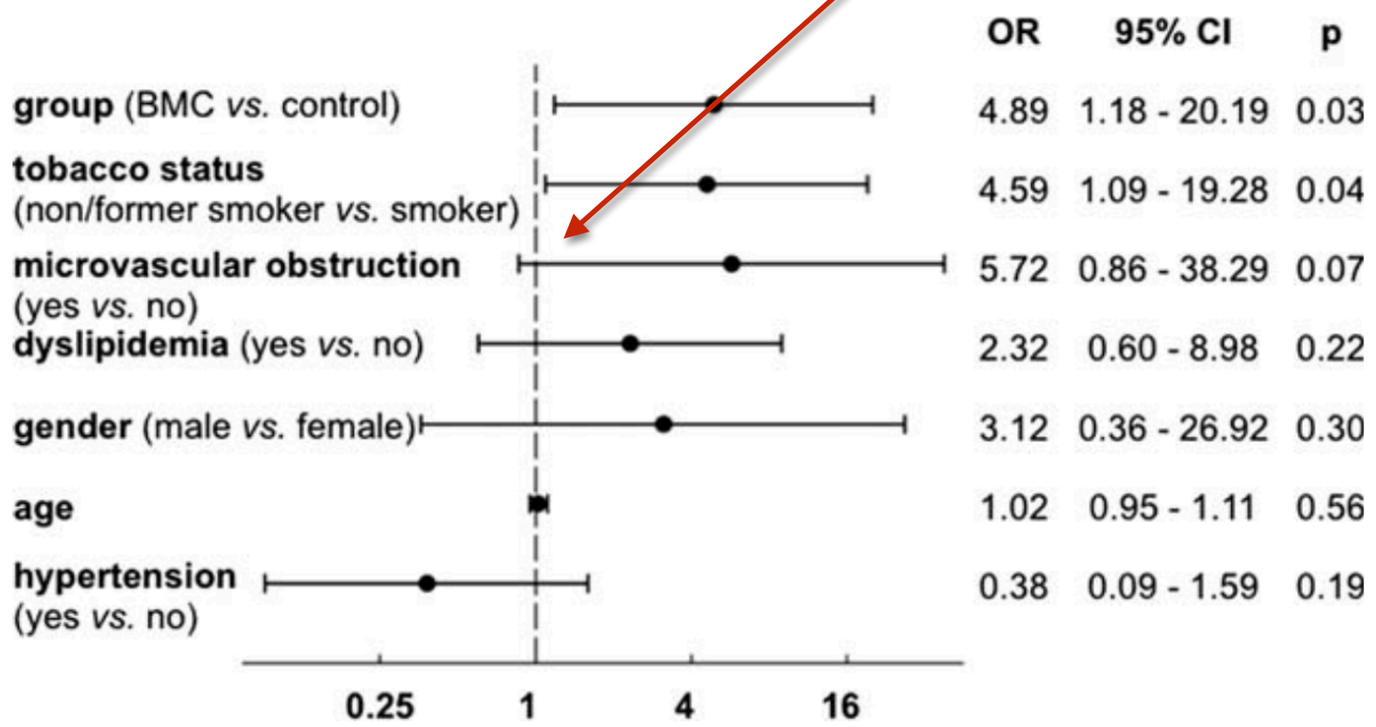


Figure 3 Multivariate logistic regression analysis for improvement of at least two non-viable segments becoming viable ($n = 77$). OR, odds ratio; CI, confidence interval; p: *P*-value.

From Roncalli et al. (2011) *European Heart Journal* **32**: 1748-1757.

Confidence Intervals Not

- Effects on viability of heart muscle a

Also can't rule out very large effects, and here's the kicker: no way to know which is more likely!!! True value is somewhere within this range...hopefully.

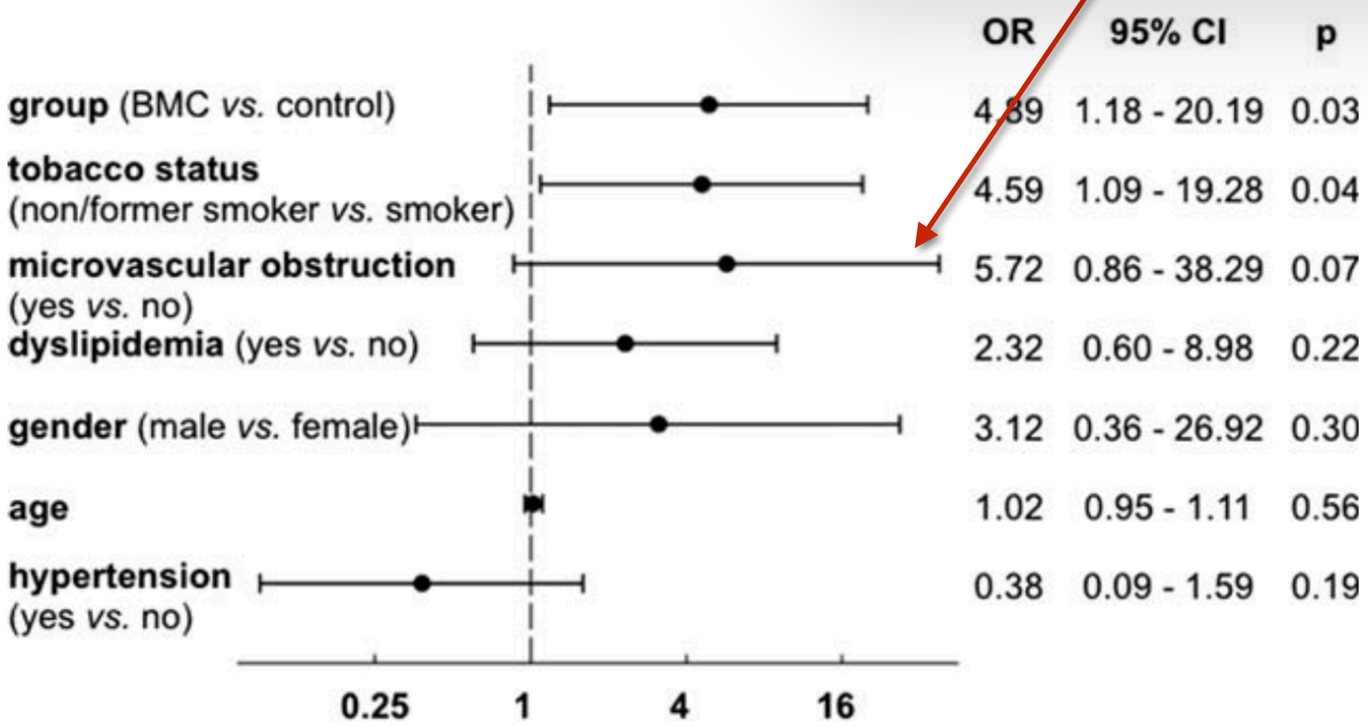


Figure 3 Multivariate logistic regression analysis for improvement of at least two non-viable segments becoming viable ($n = 77$). OR, odds ratio; CI, confidence interval; p: *P*-value.

From Roncalli et al. (2011) *European Heart Journal* **32**: 1748-1757.

Confidence Intervals Not Much Better

- Effects on viability of heart muscle af

Mean estimate of effect is larger than others with “significant” effects!!

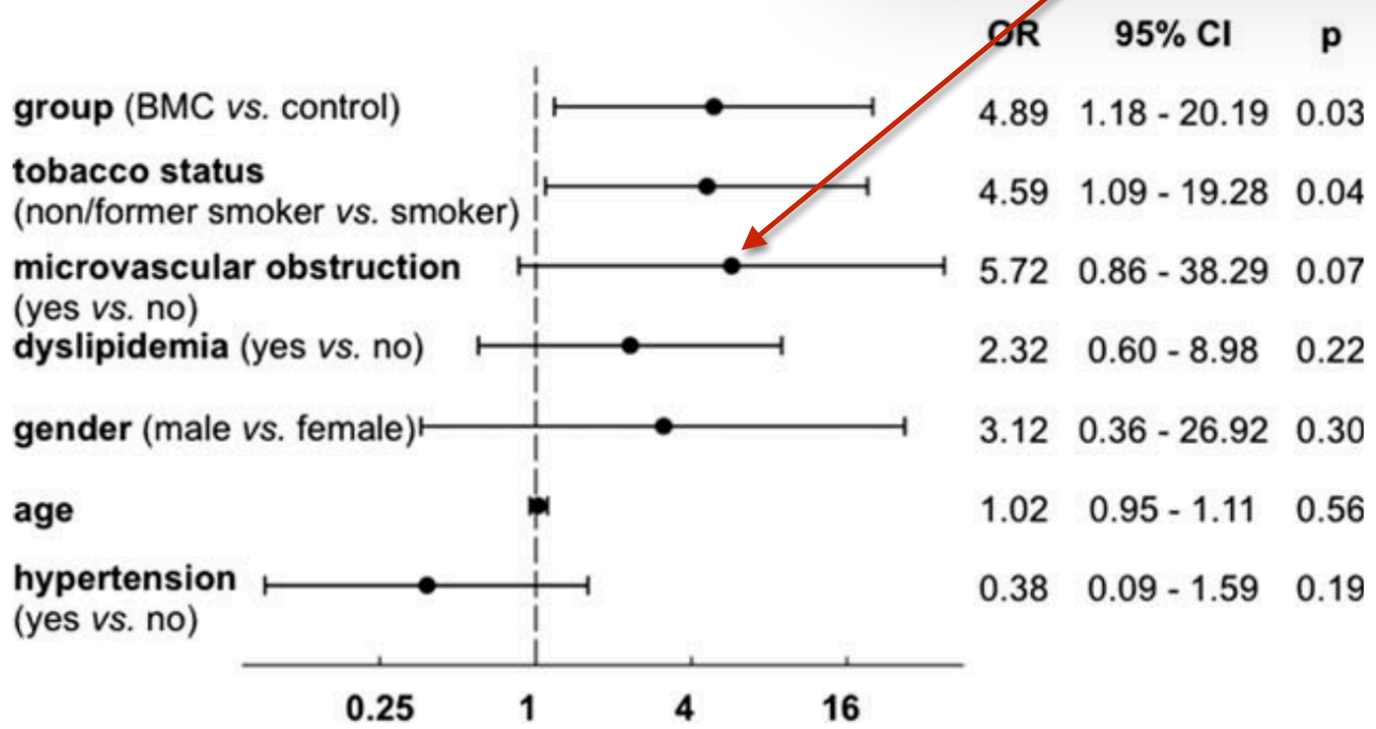
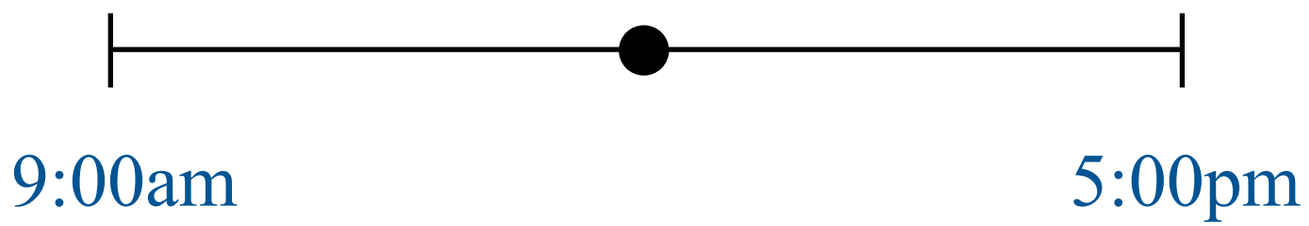


Figure 3 Multivariate logistic regression analysis for improvement of at least two non-viable segments becoming viable ($n = 77$). OR, odds ratio; CI, confidence interval; p: *P*-value.

From Roncalli et al. (2011) *European Heart Journal* **32**: 1748-1757.

Confidence Intervals Not Much Better

- When thinking of CIs, think of trying to book appointment with supervisor!
  - "What day and time can we meet to discuss my defence?"
  - "95% of the time I am in my office *sometime* between 9:00am and 5:00 pm"



How helpful is this when trying to come up with a specific time that would be best (with day or time)?

Background on Frequentist Approach*

- Parameters are fixed but unknown constants
  - The mean of a population

- Probabilities based on an infinite number of hypothetical repetitions of the experiment
  - Probabilities refer to the **data** under very specific conditions, **not** the parameter of interest

---

* Based on Bolstad (2004) *Introduction to Bayesian Statistics*. John Wiley & Sons

0.05 Criterion is Arbitrary

Does One Universal *p*-value Make Sense?

Suppose your studying a drug to treat cancer

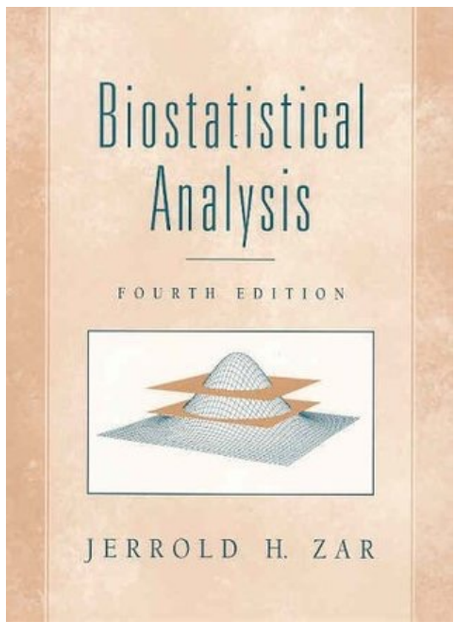
- Survival is noticeably higher in treatment group than in controls
- But $p = 0.051$

What do you do?

Does One Universal *p*-value Make Sense?

- The following section is based on how I, and likely you, were taught statistical hypothesis testing. However, this approach is incorrect, and an artificial hybridization of the ideas of Fisher and Neyman & Pearson[3]

- Regardless, it is useful to point out the logical flaws of a universal *p*-criterion within this commonly-used context



3. Gigerenzer (2004) *The Journal of Socio-Economics* **33**: 587-606.

Does One Universal *p*-value Make Sense?

Rationale for having a low *p* criterion (0.05) is to minimize the chance of incorrectly rejecting a true hypothesis (α)

But the more stringent we make this criterion, the more likely we are to incorrectly accept a false hypothesis (β)

Does One Universal *p*-value Make Sense?

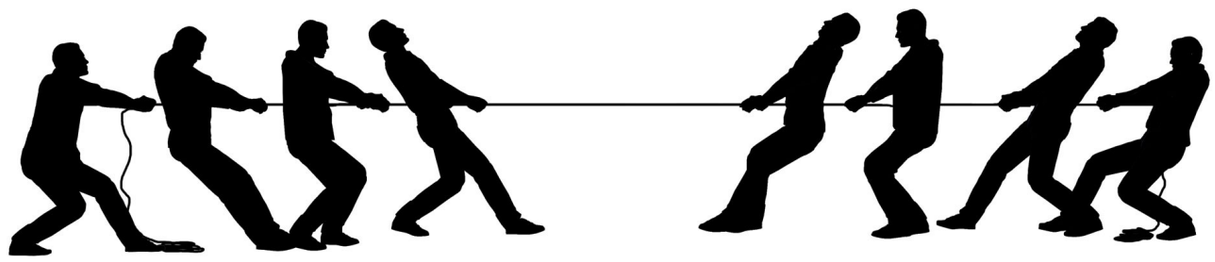
Rationale for having a low *p* criterion (0.05) is to minimize the chance of incorrectly rejecting a true hypothesis (α)

But the more stringent we make this criterion, the more likely we are to incorrectly accept a false hypothesis (β)

Are the consequences of α and β the same for all situations??

Does One Universal *p*-value Make Sense?

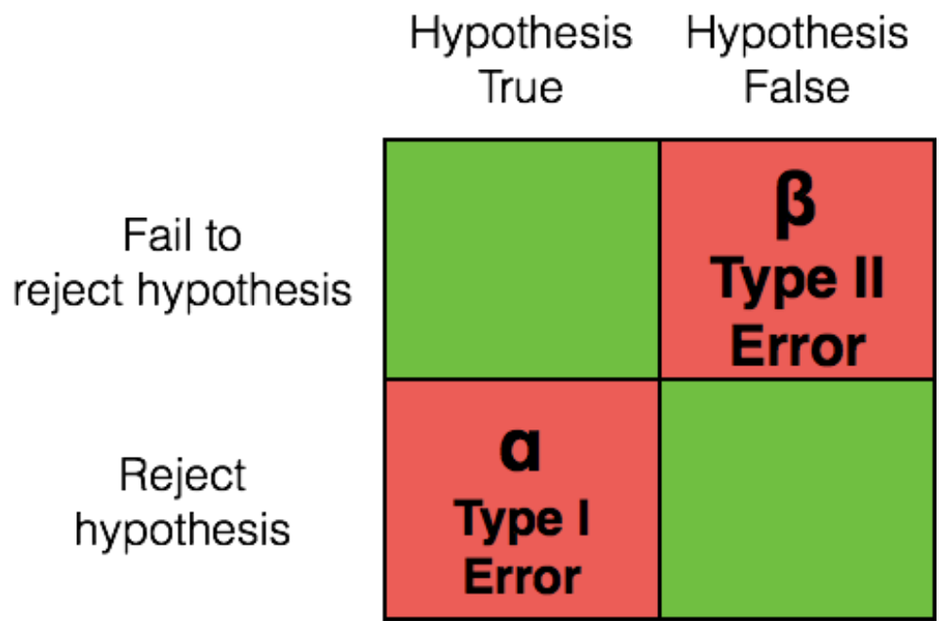
	Hypothesis True	Hypothesis False
Fail to reject hypothesis		**β Type II Error**
Reject hypothesis	**α Type I Error**	



Does One Universal *p*-value Make Sense?

Example: Criminal case

- Null Hypothesis: suspect is not guilty
- Would much rather let a guilty person go free (Type II error), than put an innocent person in jail (Type I error)

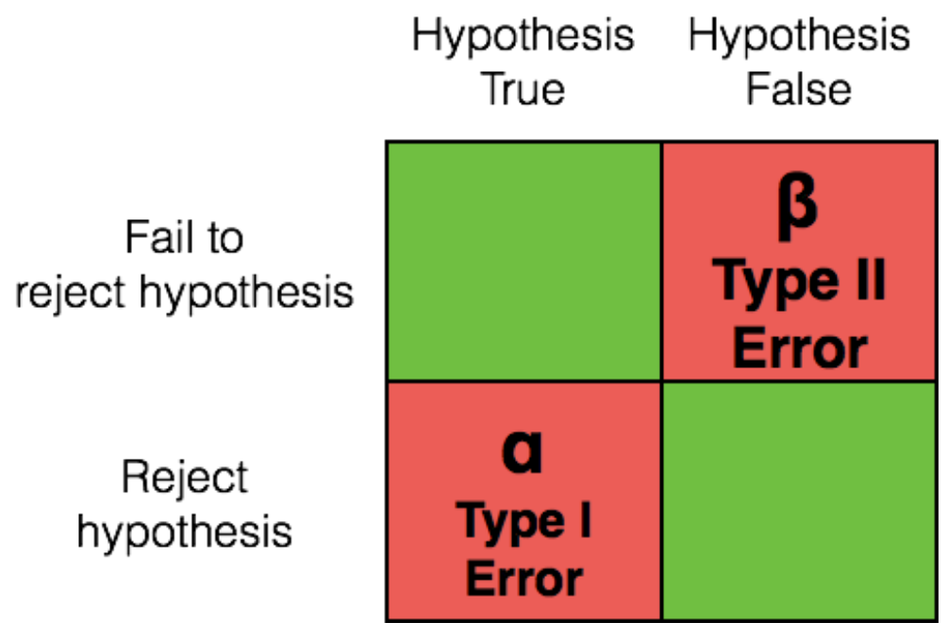


- Having low *p*-criterion (even lower than 0.05) is appropriate

Does One Universal *p*-value Make Sense?

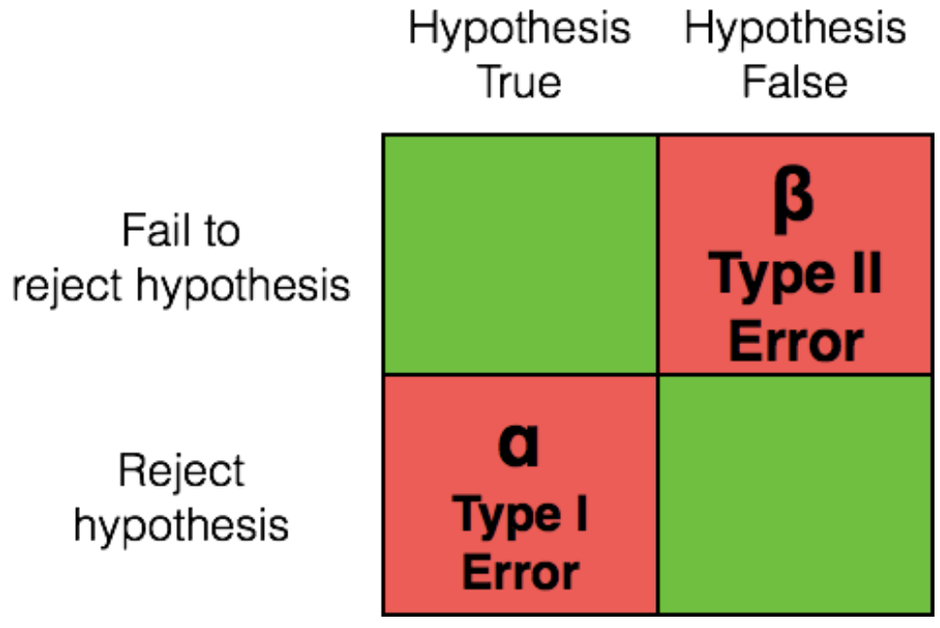
Example: Conservation

- Null Hypothesis: species is not going extinct
- Would much rather implement unnecessary conservation actions (Type I error) than allow a species to go extinct (Type II error)



- Having a low β value is far more important than having a low α value

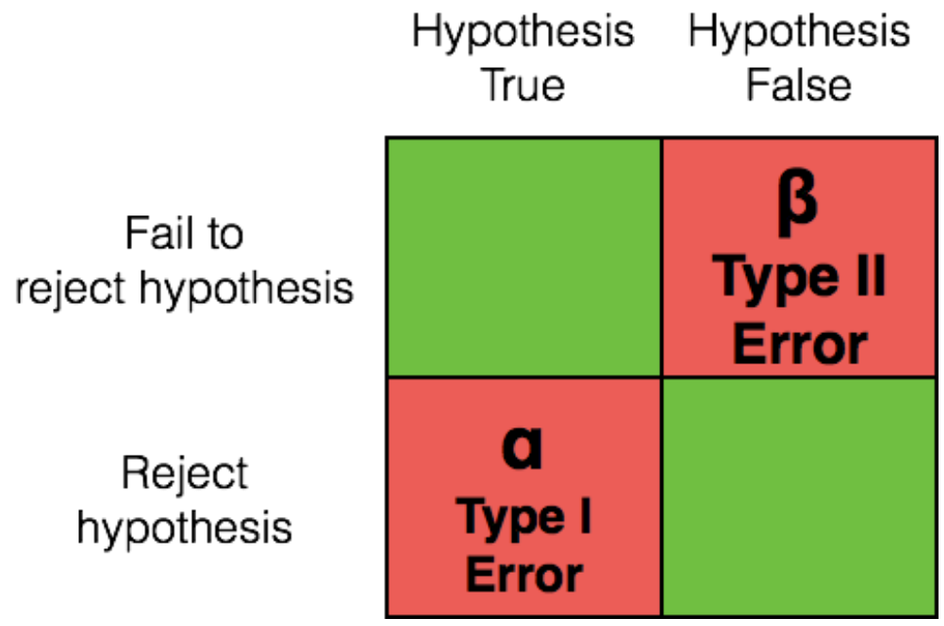
Does One Universal *p*-value Make Sense?



- The consequences of committing Type I and Type II errors are vastly different in different scenarios.
- One criterion across *every* test in *every* study is ridiculous!
- *p*-values over-emphasized, and β often ignored (when often it is just as, if not more, important)

Does One Universal *p*-value Make Sense?

- Canned single pre-set criterion has resulted in scientists not really thinking about their hypotheses and consequences of different errors



- Should have well-founded rationale for setting each error criterion at the values used

Often Equate “Not Significant” with “No Effect”

Not Significant ≠ No Effect

- Just because something was not found to have a “significant” effect does **not** mean that it has **no** effect!!!!
  - Binary decision throws away so much information
  - Leads us to incorrect conclusions

Not Significant ≠ No Effect

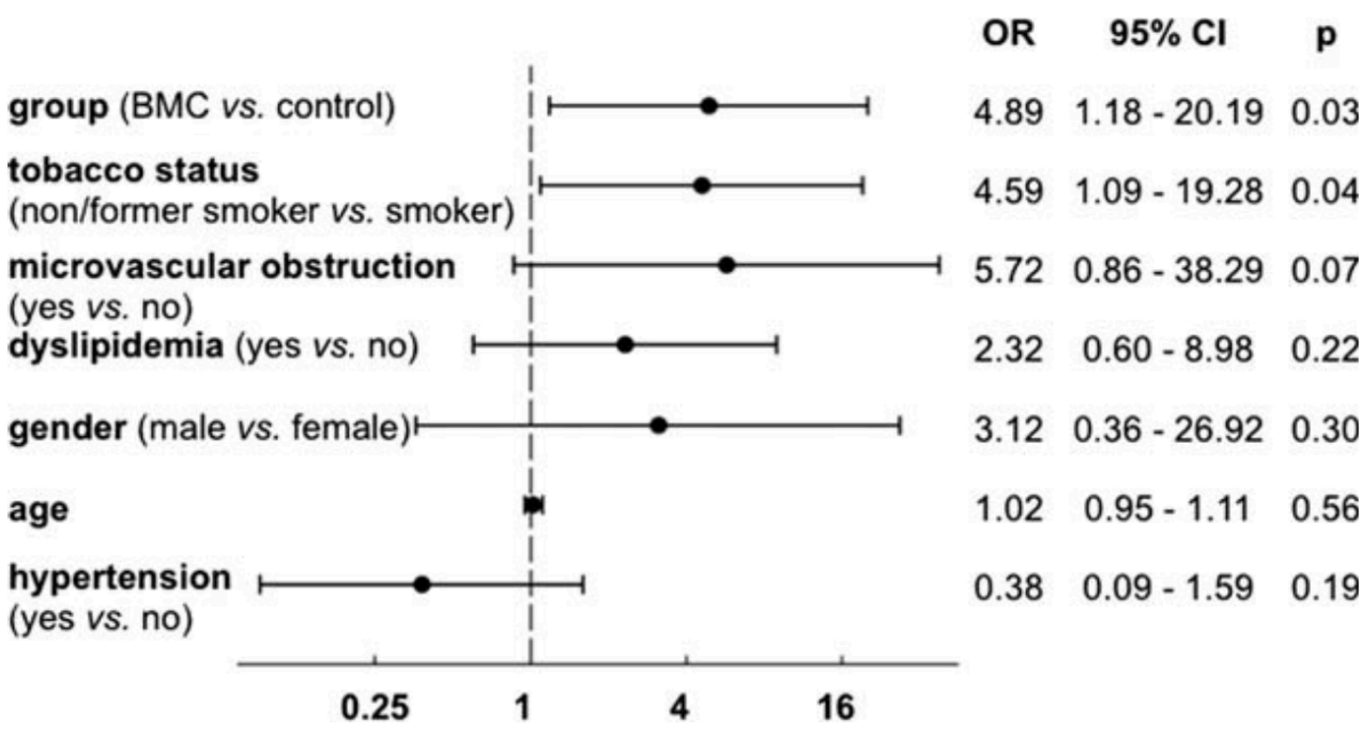


Figure 3 Multivariate logistic regression analysis for improvement of at least two non-viable segments becoming viable ($n = 77$). OR, odds ratio; CI, confidence interval; p: *P*-value.

From Roncalli et al. (2011) *European Heart Journal* **32**: 1748-1757.

Not Significant ≠ No Effect

ELSEVIER

Accident Analysis and Prevention 36 (2004) 495–500

ACCIDENT ANALYSIS & PREVENTION

www.elsevier.com/locate/aap

Viewpoint

The harm done by tests of significance

Ezra Hauer*

35 Merton Street, Apt. 1706, Toronto, Ont., Canada M4S 3G4

Abstract

Three historical episodes in which the application of null hypothesis significance testing (NHST) led to the mis-interpretation of data are described. It is argued that the pervasive use of this statistical ritual impedes the accumulation of knowledge and is unfit for use.

Keywords: Significance; Statistical hypothesis; Scientific method

Not Significant ≠ No Effect

Table 1
The Virginia RTOR study

	Before RTOR signing	After RTOR signing
Fatal crashes	0	0
Personal injury crashes	43	60
Persons injured	69	72
Property damage crashes	265	277
Property damage (US$)	161243	170807
Total crashes	308	337

Not Significant ≠ No Effect

> The problem is clear. Researchers obtain real data which, while noisy, time and again point in a certain direction. However, instead of saying: "here is my estimate of the safety effect, here is its precision, and this is how what I found relates to previous findings", the data is processed by NHST, and the researcher says, correctly but pointlessly: "I cannot be sure that the safety effect is not zero". Occasionally, the researcher adds, this time incorrectly and unjustifiably, a statement to the effect that: "since the result is not statistically significant, it is best to assume the safety effect to be zero". In this manner, good data are drained of real content, the direction of empirical conclusions reversed, and ordinary human and scientific reasoning is turned on its head for the sake of a venerable ritual. As to the habit of subjecting the data from each study to the NHST separately, as if no previous knowledge existed, Edwards (1976, p. 180) notes that "it is like trying to sink a battleship by firing lead shot at it for a long time".

Size Matters:

Diluting results to a simple yes/no decision removes most of the useful information

(Effect) Size Matters

- In most, if not all, scenarios we are more interested in the *size* of an effect, rather than a simple “yes”/“no” decision

- Interestingly, most obvious in topics that we really care about (don’t we care about our science?)

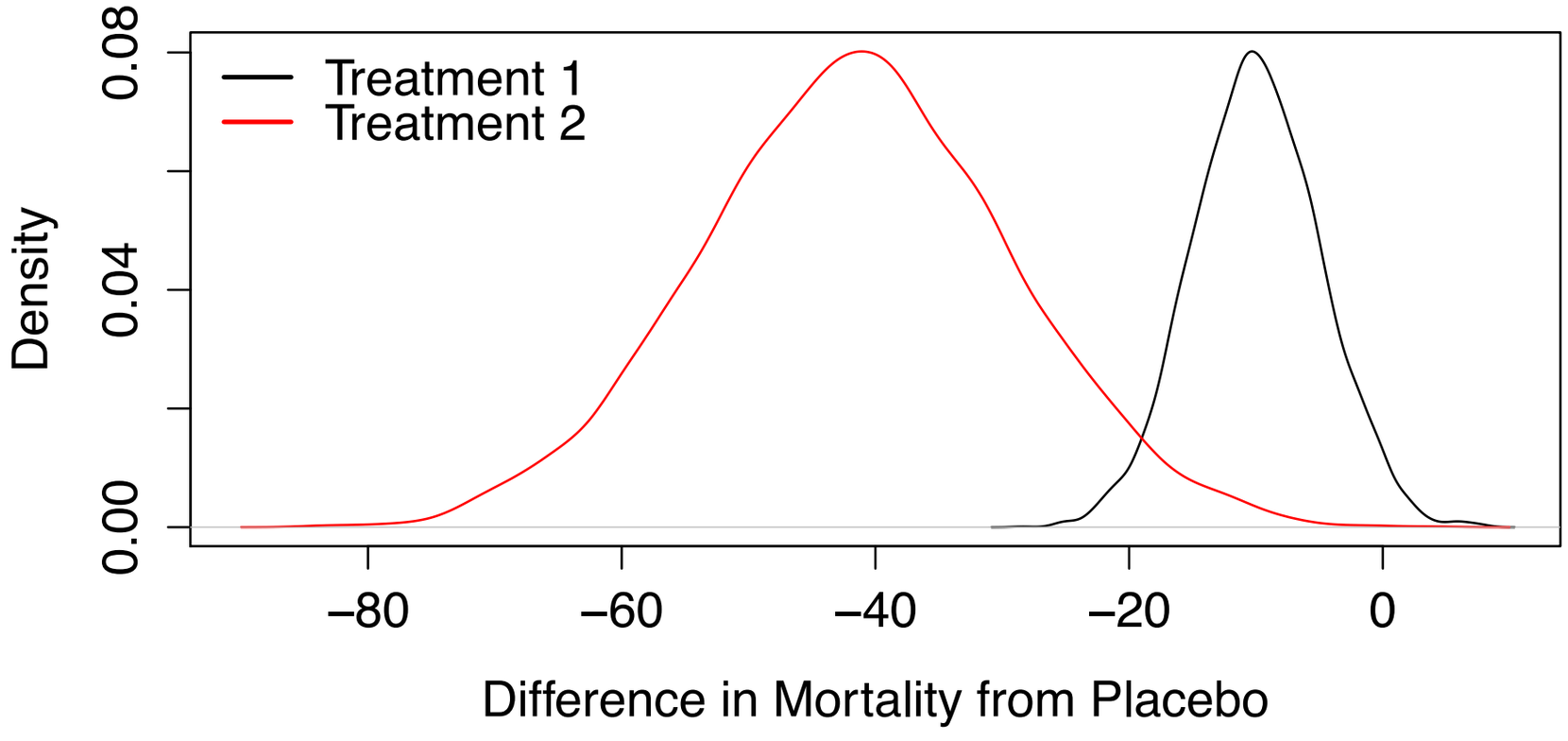
(Effect) Size Matters
Example: Cancer treatment

- Suppose you're given a choice between two cancer treatments, and you're only told that "both significantly reduce mortality risk from cancer, $p < 0.05$"
- Would that be enough for you, or would you want to know *by how much* each reduces mortality risk from cancer?

(Effect) Size Matters

Example: Cancer treatment

- Both are significant at $p < 0.05$, but effect size is far more important to us



(Effect) Size Matters

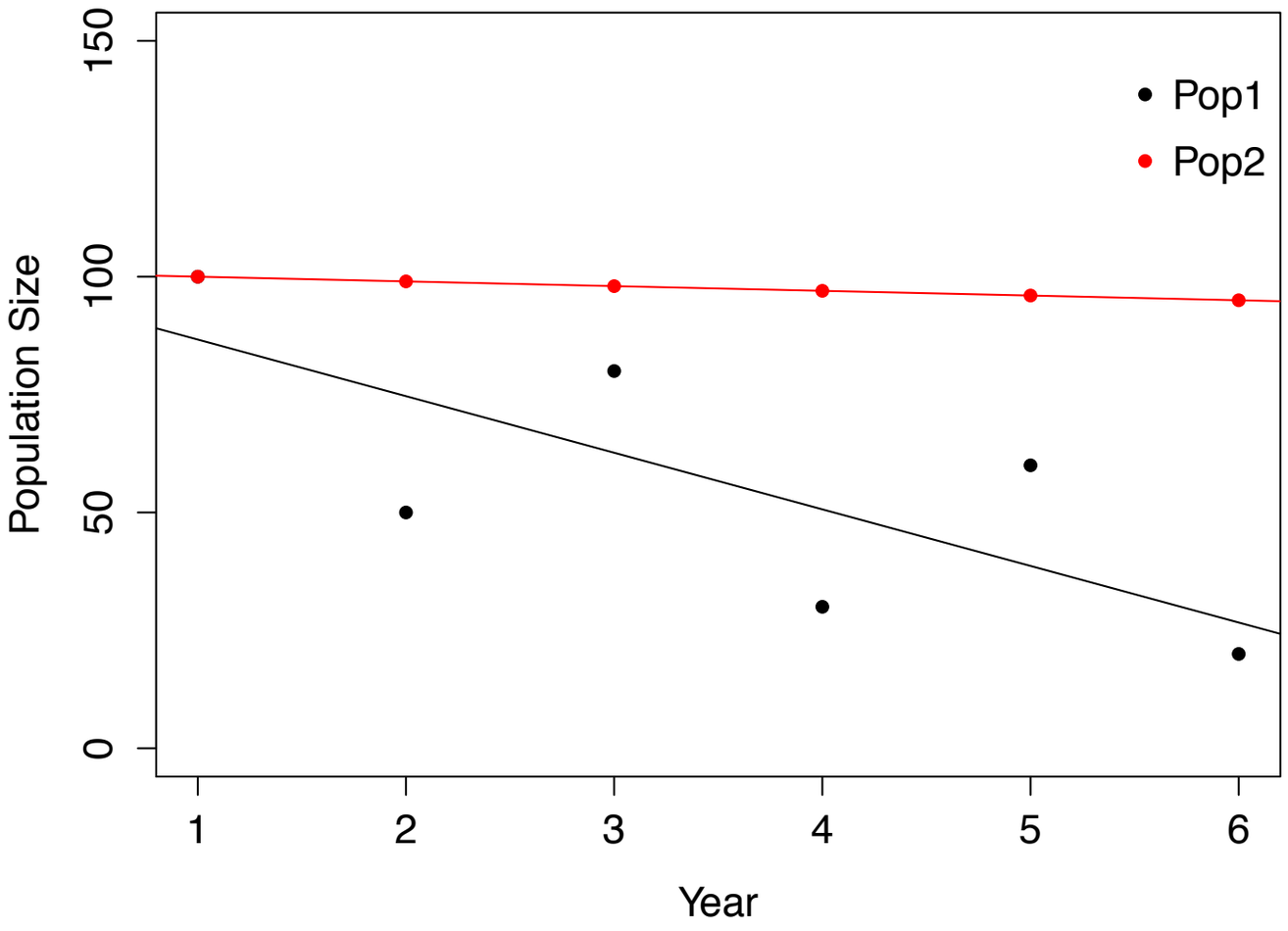
- The estimated size of an effect is/should often be more important to us than a simple yes/no conclusion
  - Yes/no only provides very crude understanding
- Allows for more appropriate decision making and evaluation of the data
- *p*-values tell us little about the data, and are a poor summary of effects

Precision Heavily Influences *p*-values

Precision Problems

Example: Two populations

Year	Pop1	Pop2
1	100	100
2	50	99
3	80	98
4	30	97
5	60	96
6	20	95

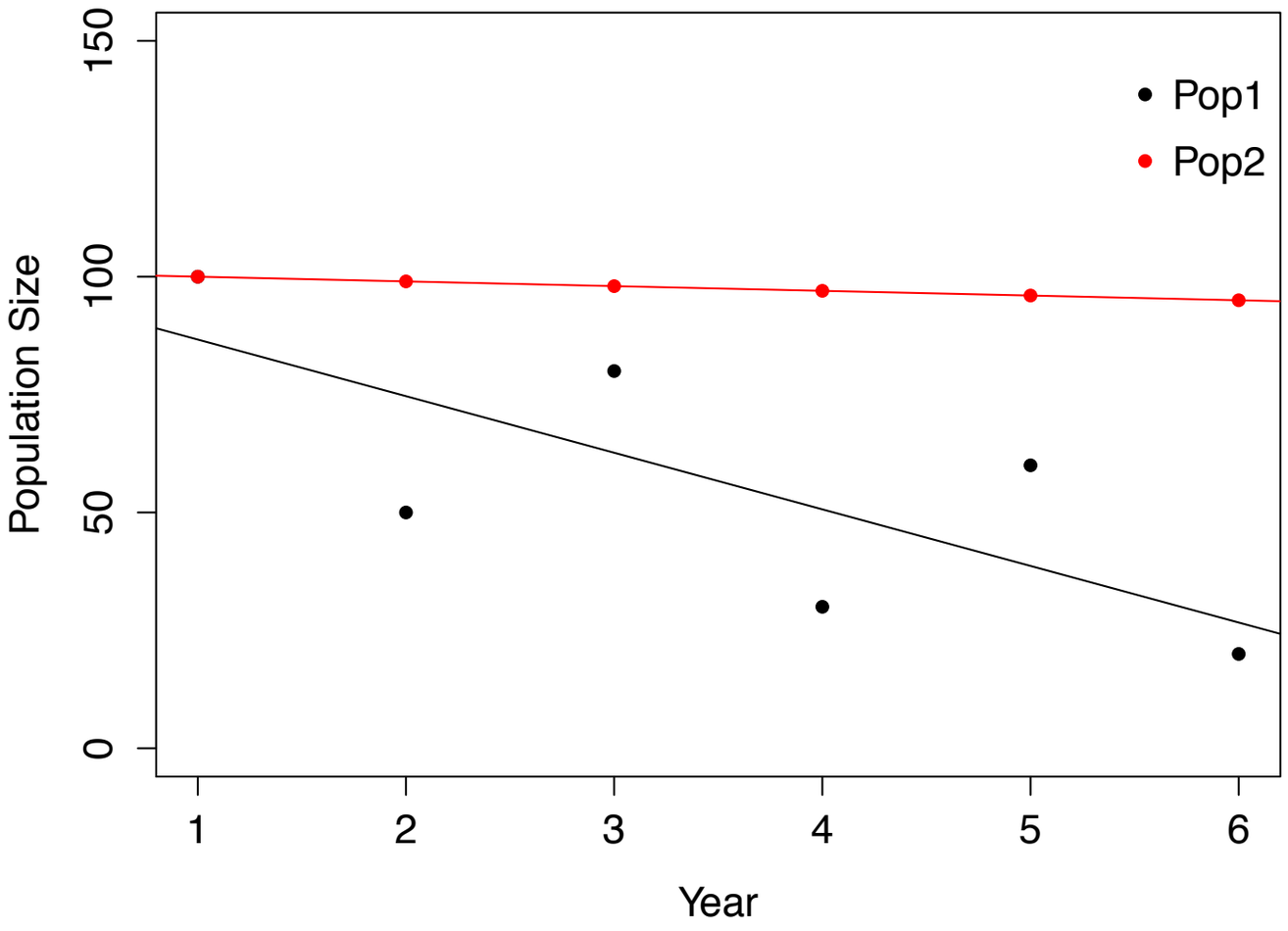


Which should we be most worried about?

Precision Problems

Example: Two populations

Year	Pop1	Pop2
1	100	100
2	50	99
3	80	98
4	30	97
5	60	96
6	20	95



Pop1 p = 0.08

Pop2 p = 2.2×10^{-16}

Huh?

p-values Don't Behave Well

p Roulett

Video

p-values Change Depending on Your Intent

p-values Change With Intent

- See:
  - Chapter 11, Kruschke (2015)
  - Berger & Berry (1988)

p-values Change With Intent

Example: Bonferroni correction for multiple tests

- Critical *p*-values become more strict with multiple tests to maintain the same probability of falsely rejecting a true hypothesis

# of Tests	*p* criterion
1	0.05
2	0.025
3	0.017
…	…

p-values Change With Intent

Example: Bonferroni correction for multiple tests

- Critical *p*-values become more strict with multiple tests to maintain the same probability of falsely rejecting a true hypothesis

Makes sense mathematically, but what about logically?

# of Tests	*p* criterion
1	0.05
2	0.025
3	0.017
…	…

p-values Change With Intent
Example: Bonferroni correction for multiple tests

- Suppose I performed one experiment, and compared two data sets
  - Obtain a moderate effect size (0.52), and a "significant" *p*-value (0.03)

- My interpretation would be that there is a significant difference between these groups
  - With subsequent implications on my interpretation of the world, future research path, etc.

p-values Change With Intent

Example: Bonferroni correction for multiple tests

- Suppose instead that I performed the exact same experiment, and got exactly the same results, but now it is one of two experiments that I performed
- My *p*-value no longer falls below the corrected criterion; my results not considered significant
  - Completely opposite implications on my interpretation of the world, future research path, etc. even though data were exactly the same

p-values Change With Intent

Example: Bonferroni correction for multiple tests

- Suppose instead that I performed the exact same experiment, and got exactly the same results, but now it is one of two experiments that I performed
- My *p*-value no longer falls below the corrected criterion; my results not considered significant
  - Completely opposite implications on my interpretation of the world, future research path, etc. even though data were exactly the same

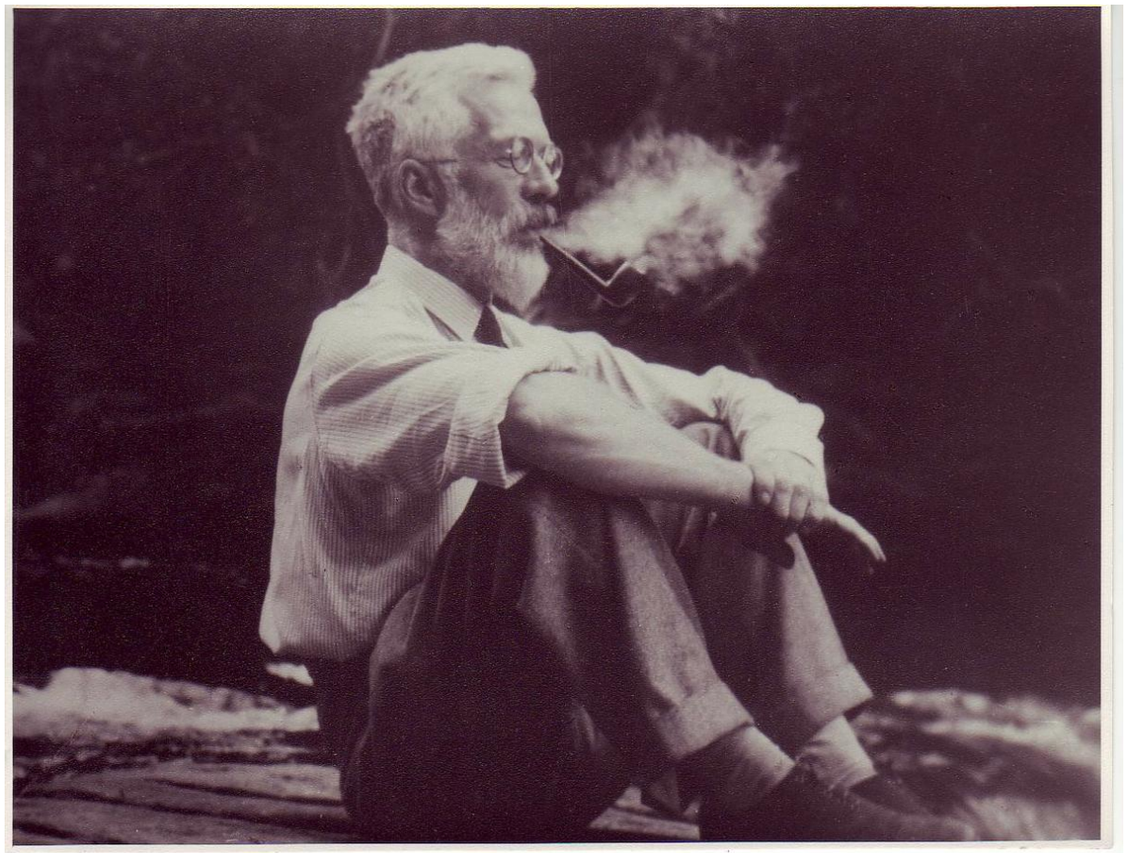
Does this make sense?

Summary

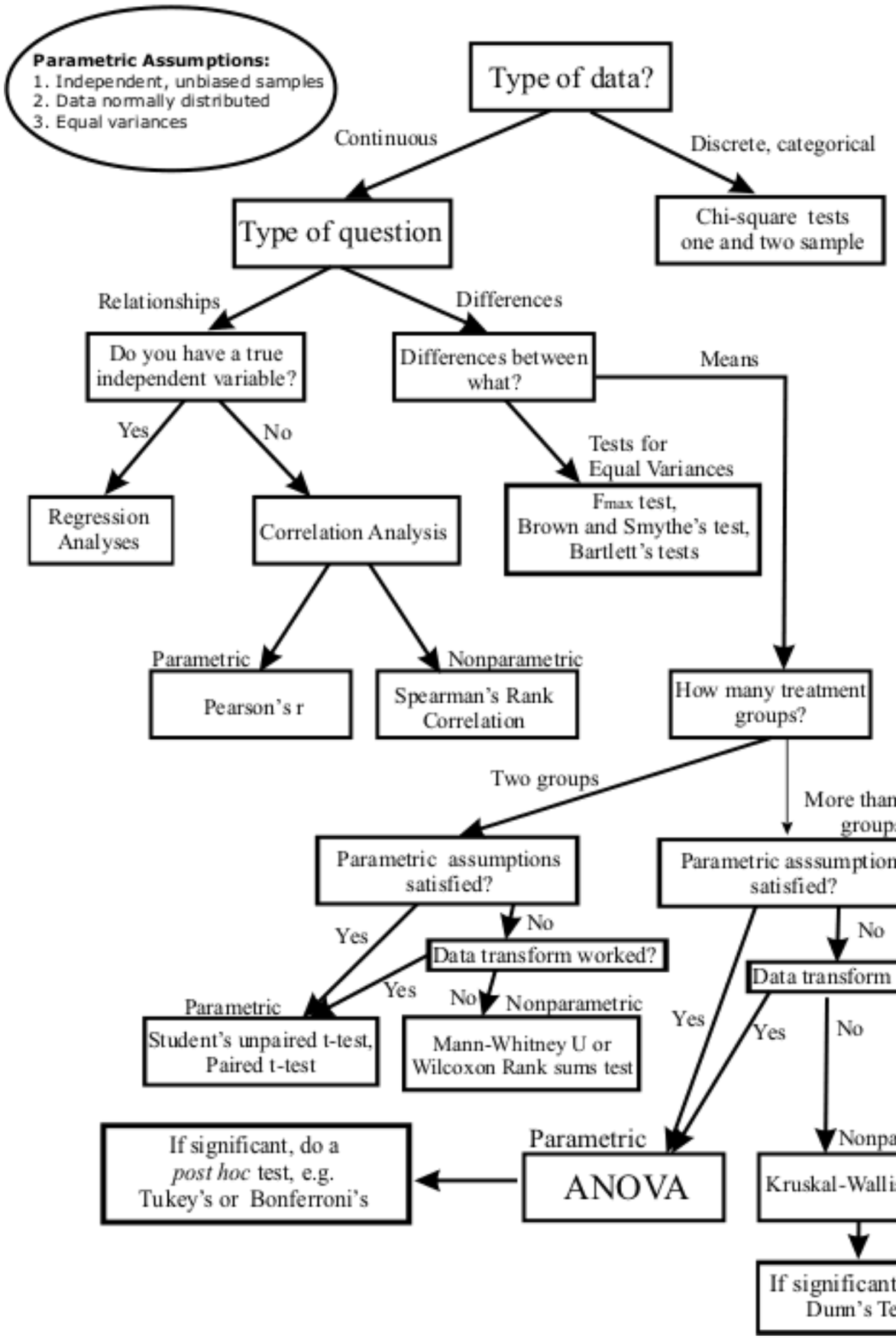
Summary

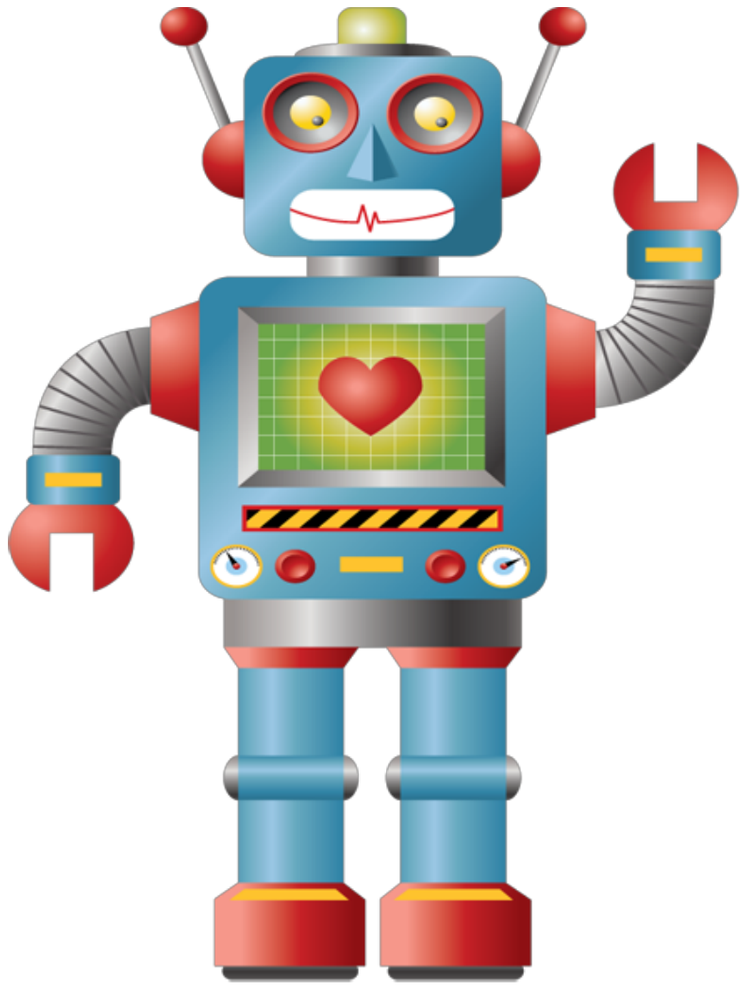
- *p*-values and confidence intervals have many undesirable characteristics for basing our understanding of the world on

- While they are useful and appropriate in some situations, they are not for many others, nor are they they only tool in the statistical toolbox



Flow Chart for Selecting Commonly Used Statistical Tests





Questions?