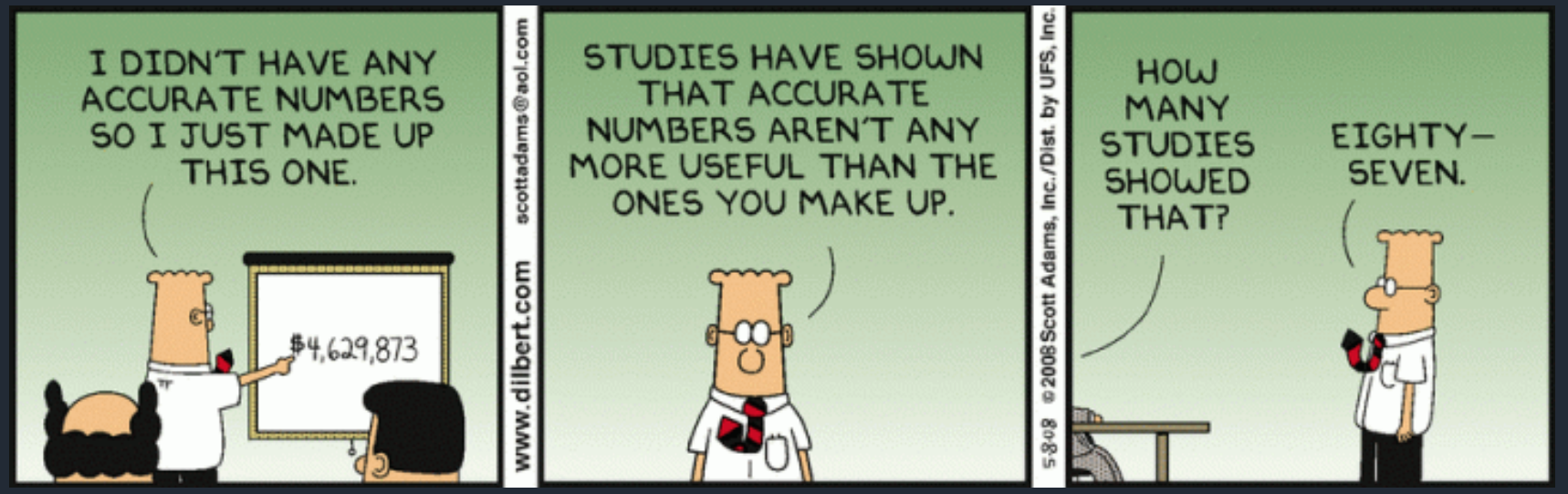
Inferential Statistics

Tyson S. Barrett, PhD

PSY 3500



Descriptive vs. Inferential

Descriptive: describes the data

How many in the sample? What is the average in the sample?

Inferential: infers about the population

How many are there in the population? What is the average in the population?

Descriptive vs. Inferential

We are now talking about inferential

Inferential: infers about the population

How many are there in the population? What is the average in the population?

Null Hypothesis Significance Testing

The most common approach in Psych

Helps researchers decide if an effect in a sample is in the population

Tests whether the result fits the null hypothesis or the alternative hypothesis

Null hypothesis = no effect, no relationship, no nothing
Alternative hypothesis = effect, relationship, something happening

Relies on p-values

P-value is the probability of obtaining a result as extreme or more extreme if the null is indeed true

Based on P-Values

The probability of obtaining results as extreme or more extreme if the null hypothesis is true

Create a world where the null is true

A world where there is no effect or no relationship

See what is common in that world

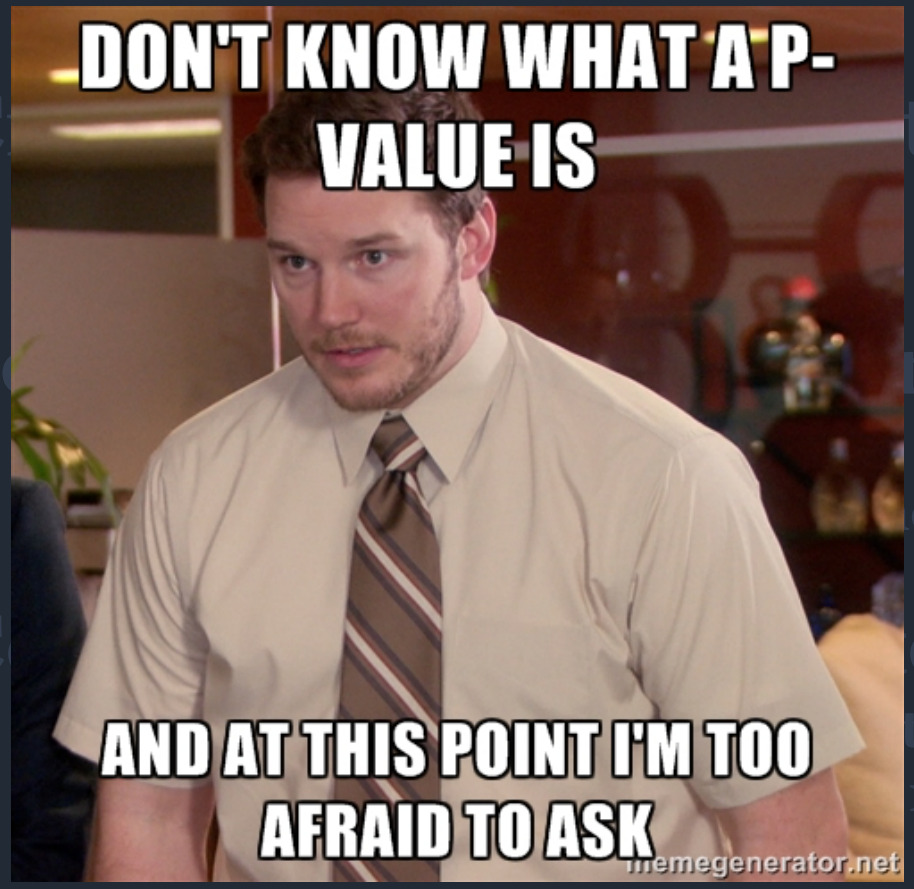
A world where there is no effect or no relationship can have results that look like there is an effect but there isn't

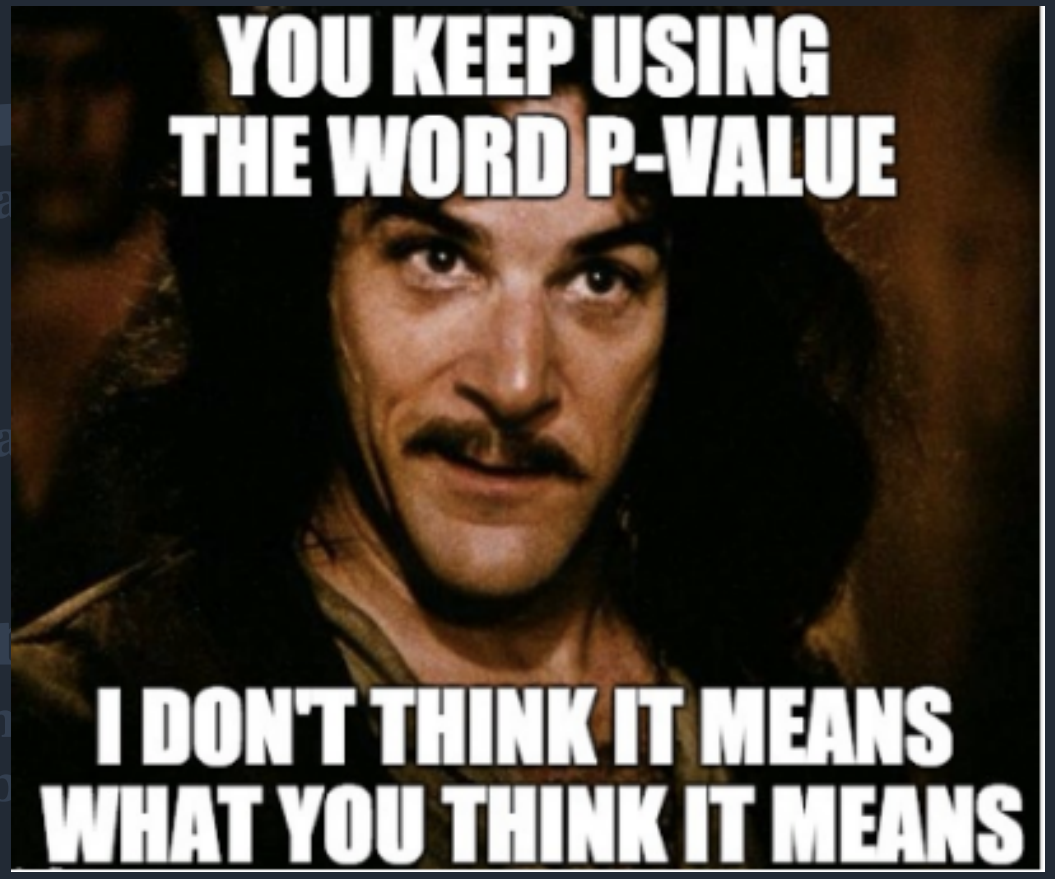
Compare your result to the null world ones

If it is similar to the null world, then the null may be true. If it is very different from the null world, it may not be true (i.e., there is an effect).

Based on P-Values

The probability of obtaining results as extreme or more extreme if the null hypothesis is true





How NHST Works

1. Specify hypotheses

2. Calculate the estimate

3. Compare estimate to what is expected if the null was true

4. If estimate is highly unlikely if null is true, then reject the null and conclude there is an effect

How NHST Works

1. Specify hypotheses

2. Calculate the esti

3. Compare estimat
is expected if the nu
true

4. If estimate is
if
en
and
is

Null and Alternative hypotheses (not your general research hypotheses)

How NHST Works

1. Specify hypotheses

2. Calculate the estimate

3. Compare estimat
is expected if the nu
true

4. If estimate is highly unlikely if null is true, then
and
is

Usually a test statistic or a correlation

How NHST Works

1. Specify hypothese
2. Calculate the esti
3. Compare estimate to what is expected if the null was true

If null is true, we can expect results to be similar to what the null is

s
if
en
and
conclude there is an effect

How NHST Works

1. Specify hypotheses

If the estimate is very different than what we would expect if the null was true, then maybe the null isn't true

is expected if the null was true

4. If estimate is highly unlikely if null is true, then reject the null and conclude there is an effect

Rejecting and Failing to Reject

What is this all about?

Rejecting the Null

Says we believe the null is not true in the population
There is evidence that there is an effect or relationship

Failing to Reject the Null

Is basically saying there is not enough evidence to reject the null but we can't say there is not an effect

Rejecting and Failing to Reject

What is this all about?

Rejecting the Null

Says we believe the null is not
There is evidence that there is

Conclude:
Yes, there is an effect

Failing to Reject the Null

Is basically saying there is not enough evidence to reject the null but we can't say there is not an effect

Rejecting and Failing to Reject

What is this all about?

Rejecting the Null

Says we believe the null is not true in the population
There is evidence that there is an effect or relationship

Failing to Reject the Null

Is … gh e…
ca…

Conclude:
There is not enough evidence that there is an effect

Basically is a ???
(could be a small effect or no effect)

Errors in NHST

NHST requires a decision which means we can make a mistake

Statistical Significance vs. Meaningfulness

In larger samples, a small effect can be statistically significant

- But might be irrelevant (many gender studies are like this)

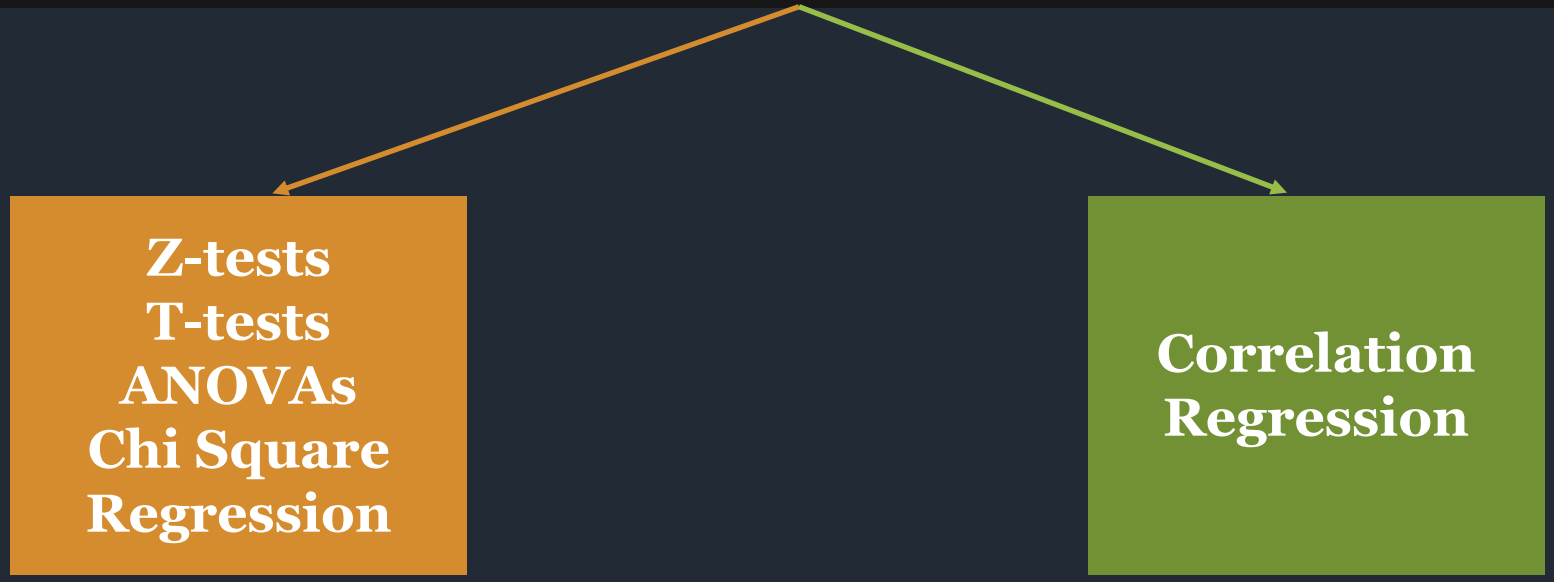
In small samples, the effect could be big but not statistically significant

- Could mean the effect is just noise or it could mean we need a bigger sample

Common Tests

These are the basic, common tests used in Psychology (there are MANY others)

Does research question have to do with looking at differences among groups or relationships among continuous variables?



Statistical Power

The ability to find an effect if it exists

Things that improve power

- Bigger samples
- Better measures
- Longitudinal designs
- Fewer confounders
- Bigger effects

Criticisms of NHST

It has a few important problems

Misunderstanding of what a p-value means

It is NOT the probability that the null hypothesis is true

The logic of NHST of using an arbitrary p-value cut-off is faulty

Why .05?

The null hypothesis may not be a very useful comparison

Potential Solutions

Several have been proposed and some are becoming more common

Report effect sizes and confidence intervals

Have journals publish all results (significant or not)

Use Bayesian Statistics

I think all should be used more (especially the first two)

Replicability Crisis

Recent work has shown issues with replicating results of previous studies

Not necessarily always a bad thing

We expect studies to not always be replicated but sometimes it is due to poor research practices (see page 380 in the book)

Replicability

Recent work has shown issues with replicati

Not necessarily always

We expect studies to not always
poor research practices (see pag

1. The selective deletion of outliers in order to influence (usually by artificially inflating) statistical relationships among the measured variables.
2. The selective reporting of results, cherry-picking only those findings that support one's hypotheses.
3. Mining the data without an a priori hypothesis, only to claim that a statistically significant result had been originally predicted, a practice referred to as "HARKing" or hypothesizing after the results are known (Kerr, 1998).
4. A practice colloquially known as "p-hacking" (briefly discussed in the previous section), in which a researcher might perform inferential statistical calculations to see if a result was significant before deciding whether to recruit additional participants and collect more data (Head et al., 2015). As you have learned, the probability of finding a statistically significant result is influenced by the number of participants in the study.
5. Outright fabrication of data (as in the case of Diederik Stapel, described at the start of Chapter 3), although this would be a case of fraud rather than a "research practice."

Replicability Crisis

Recent work has shown issues with replicating results of previous studies

Not necessarily always a bad thing

We expect studies to not always be replicated but sometimes it is due to poor research practices (see page 380 in the book)

Generally want to assume researchers are being honest and are doing their best

You can become too cynical otherwise…

Use Open Science!!!

Provide your data, your analyses, your research questions, everything that was done in the study (usually as supplementary material)