To P or not to P?

Knowing When the P-value is Less than Useful

Mike LaValley PhD

Dan White PT PhD

Disclosures

- None

Evidence-Based Medicine

- Chatfield C. Model uncertainty, data mining and statistical inference. *J R Statis Soc A*, 1985; 158:419-466.
- Dorey F. The P value: what is it and what does it tell you? *Clin Orthop Relat Res*, 2010; 468:2297-8.
- Du Prel J-B, Hommel G, Rohrig B, Blettner M. Confidence interval or P-value? *Dtsch Arztebl Int*, 2009; 106:335-9.
- Goodman SN. Toward evidence-based medical statistics. 1: the P value fallacy. *Ann Intern Med*, 1999; 130:995-1004
- Lesaffre E. Use and misuse of the P-value. *Bull NYU Hosp Jt Dis*, 2008; 66:146-9
- Stang A, Poole C, Kuss O. The ongoing tyranny of statistical significance testing in biomedical research. *Eur J Epidemiol*, 2010; 25: 225-30

Outline

- Definition of p-values
- Common uses for p-values when it may be of limited use
- Alternative approaches for when the p-value is of limited use
- Review examples from arthritis studies

What is a P-value?

- When we see ***p < 0.05*** in a paper what is it telling us?

What is a P-value?

- When we see ***p < 0.05*** in a paper what is it telling us?
  A. The authors obtained this result fewer than 5 times out of every 100 attempts

What is a P-value?

- When we see ***p < 0.05*** in a paper what is it telling us?
  A. The authors obtained this result fewer than 5 times out of every 100 attempts
  B. We should read this paper

What is a P-value?

- When we see ***p < 0.05*** in a paper what is it telling us?
  A. The authors obtained this result fewer than 5 times out of every 100 attempts
  B. We should read this paper
  C. If there truly was no effect then one would expect to see a result like this less than 5% of the time

What is a P-value?

- When we see ***p < 0.05*** in a paper what is it telling us?
  A. The authors obtained this result fewer than 5 times out of every 100 attempts
  B. We should read this paper
  C. If there truly was no effect then one would expect to see a result like this less than 5% of the time
  D. The probability that this result is false is less than 5%

What is a P-value?

- *When we see **p < 0.05*** in a paper what is it *telling us?*
  A. The authors obtained this result fewer than 5 times out of every 100 attempts
  B. We should read this paper
  C. If there truly was no effect then one would expect to see a result like this less than 5% of the time
  D. The probability that this result is false is less than 5%

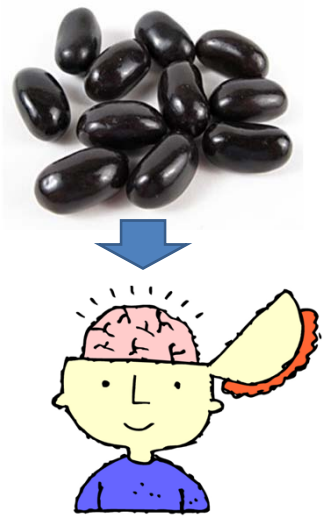
Vote

What is a P-value?

- When we see ***p < 0.05*** in a paper what is it telling us?
  A. The authors obtained this result fewer than 5 times out of every 100 attempts
  B. We should read this paper
  C. If there truly was no effect then one would expect to see a result like this less than 5% of the time
  D. The probability that this result is false is less than 5%

Definition of the P-value

- The usual set-up is that we have a hypothesis that we want to test
  - Consumption of black jelly beans increases a person's intelligence quotient (IQ)



Definition of the P-value

- The null hypothesis (H_0) would be
  - Consumption of black jelly beans does not change a person's IQ

Definition of the P-value

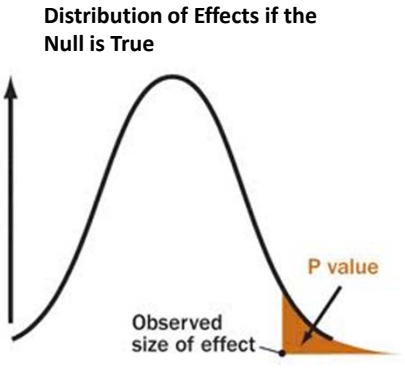
- The null hypothesis (H_0) would be
  - Consumption of black jelly beans does not change a person's IQ
- The null hypothesis states that there is no effect of the intervention or exposure
- What we are trying to disprove

Definition of the P-value

- We compare the experimental result to the sorts of results we would expect if the null hypothesis were true

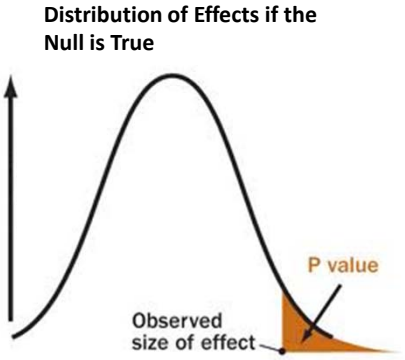
Definition of the P-value

- We compare the experimental result to the sorts of results we would expect if the null hypothesis were true
- The P-value is the probability of a larger effect than what was observed



Definition of the P-value

- A small P-value indicates that the experimental outcome is quite unlikely if the null hypothesis is true
- So, either
  - the null hypothesis is true and we have observed a very unlikely event
  - the null is not true



P-values

- Pros
  - Gives the strength of *evidence* against the null hypothesis
    - Small p-values indicate strong evidence against the null hypothesis
- Cons
  - Determined by both the size of the effect and the sample size. A small p-value could come from
    - Small effect and large sample
    - Large effect and small sample
  - Doesn't give the probability that the null hypothesis is wrong

Level of Significance

- The level of significance is the predetermined probability of **concluding** the null hypothesis is not true when it actually is true
- Like the P-value it is determined assuming that the null hypothesis is true
- Unlike the P-value this is fixed in advance by the researchers conducting the experiment
- Usually, the level of significance = 0.05

Level of Significance

- An experimental result is declared to be ***statistically significant*** if the P-value for the observed results is less than the level of significance

$$P-value < Significance\ Level = 0.05$$

- This is where the concern that the P-value < 0.05 comes from

Level of Significance

- But, remember that the P-value depends on the sample size

So the statistical significance depends on the sample size too!

- A big effect, or a big sample size, can lead to statistical significance

Common Uses of P-values Where They are of Limited Use

- Baseline characteristics table for a randomized controlled trials
- Describing the strength of an effect
- Double-dipping from the data
  - Post-hoc analyses
  - Multiple comparisons
  - Stepwise regression

Common Uses of P-values

- Baseline characteristics table for a randomized controlled trials

	Treatment Subjects	Placebo Subjects	P-values
Pain	5.6	6.1	0.10
BMI	28	27	0.25
Male Sex	14/30	10/28	0.31
. . .			
Age	62	63	0.82

Common Uses of P-values

- Baseline characteristics table for a ***randomized*** controlled trials
- This is a situation when the **null hypothesis** is actually **true**

	Treatment Subjects	Placebo Subjects	P-values
Pain	5.6	6.1	0.10
BMI	28	27	0.25
Male Sex	14/30	10/28	0.31
. . .			
Age	62	63	0.82

Common Uses of P-values

- Baseline characteristics table for a ***randomized*** controlled trials
- This is a situation when the **null hypothesis** is actually **true**
- Better to see if the groups differ in important ways than focus on the P-value

	Treatment Subjects	Placebo Subjects	P-values
Pain	5.6	6.1	0
BMI	28	27	0
Male Sex	14/30	10/28	0.
. . .			
Age	62	63	.2

Describing the Strength of Effect

- The interest in an experimental result may depend on the strength of the effect of the treatment or exposure under consideration
- Sometimes the number of 0's in the P-value is used to judge the strength of the effect
  - P-value = 0.06 – marginal strength
  - P-value = 0.01 – strong effect
  - P-value = 0.001 – stronger effect
  - P-value = 0.0001 – really strong effect

Describing the Strength of Effect

- The interest in an experimental result may depend on the strength of the effect of the treatment or exposure under consideration
- Sometimes the number of 0's in the P-value is used to judge the strength of the effect
  - P-valu… 0.06 – m…inal strength
  - P-value… …g effect
  - P-valu… …nger effect
  - P-valu… 0.0001 – …lly strong effect

Don't do this!!!

Describing the Strength of Effect

- The problem with using the number of 0's in the P-value to judge the strength of the effect is

The number of 0's in the p-value depends critically on the sample size

Describing the Strength of Effect

- Use the measure of treatment or exposure effect to gauge the strength of effect
- Dan will have some examples for this

Double-Dipping from the Data

- Basic statistical testing is designed for a 4 step approach
  1. Formulate a hypothesis
  2. Collect data to test the hypothesis
  3. Do the statistical analysis
  4. Calculate the P-value

Double-Dipping from the Data

- However, often something like this is done
  1. Formulate a hypothesis
  2. Collect data to test the hypothesis
  3. Use the data to determine the model to test
  4. Do the statistical analysis
  5. Calculate the P-value

Double-Dipping from the Data

- The problem is that the data are supposed to provide an independent test of the hypothesis
- If the data are used to determine the test to use, they no longer provide an independent test
  - We do this all the time!

Double-Dipping from the Data

- When do we double dip?
  - Choosing what to adjust for based on the observed associations in the data
  - Post-hoc testing
  - Stepwise regression
  - Data mining
  - Multiple testing

Double-Dipping from the Data

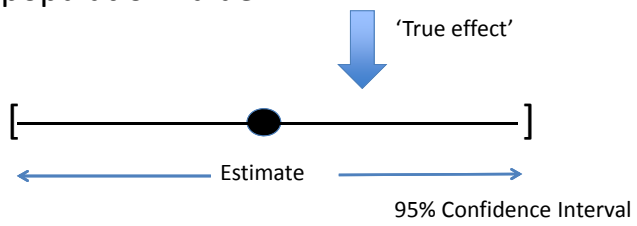
- How can we address double dipping?
  - The best way is to have an independent set of data to test the hypothesis after the first set of data is used to come up with the model
    - Model construction dataset and separate validation dataset
  - Acknowledge that the results are *exploratory* and should be tested with fresh data before being accepted as conclusive
  - P-value adjustment methods for multiple testing

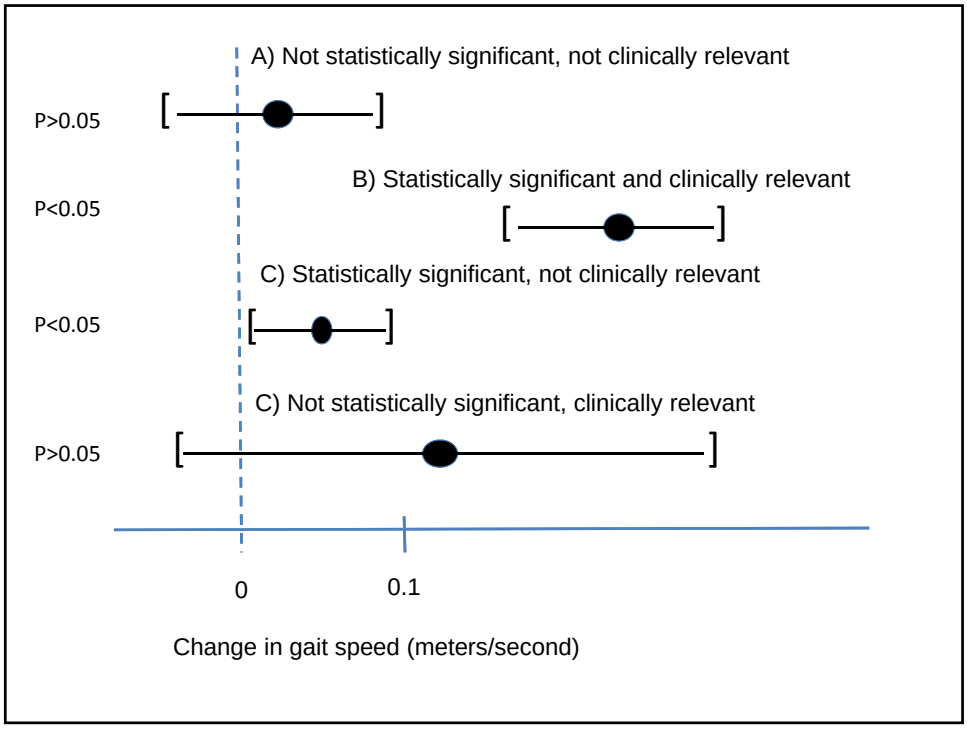
Alternative approaches

- What to do instead of using a p-value
  - Confidence Intervals
  - Effect size measures
  - Meaningful change (MCID, MDC)

Confidence Intervals

- 95% confidence interval
  - Estimate and range of values for 'true' effect
  - We are confident that 95% of the time the method will produce an interval that contains the true population value





Effect Size Measures

- Descriptive measure of the strength of effect in data
  - Distance between the null hypothesis and the observed data
- Use with confidence intervals

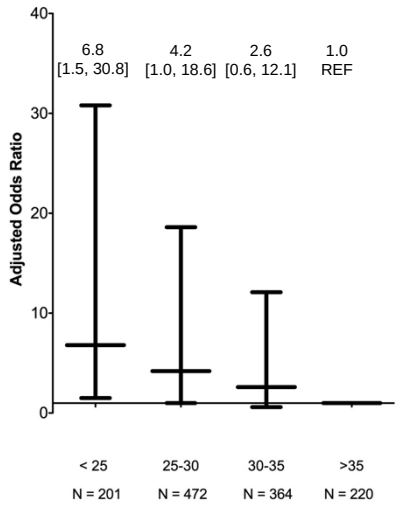
Effect Size Measure Examples

Type of Outcome	Effect Size

Effect Size Example

- Odds of meeting 2008 Physical Activity Guidelines for Americans
- By BMI categories
  - <25
  - 25-29
  - 30-34
  - >35

BMI



Meaningful Change

- Meaningful Clinical Important Difference (MCID)
  - Minimal amount of change anchored to self-report of minimal improvement or minimal decline.
- Minimal Detectable Change (MDC)
  - Minimal amount of change beyond statistical error. Anchored to a distribution

MCID Example

- Tubach et al 2005
- MCID for improvement in WOMAC pain
  - 1362 outpatients with knee OA (≥ 30/100 VAS pain)
  - 4 week Tx with NSAID

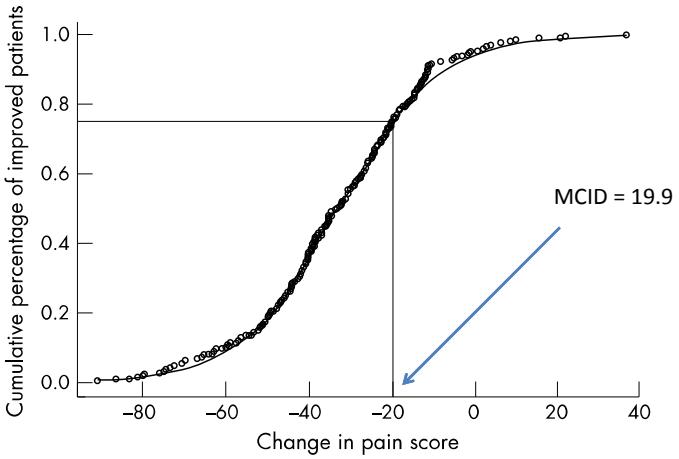


Figure 1 Aspects of the cumulative distribution function used to determine the MCII (changes in pain score in patients with knee OA; n = 265). Among patients considering their response to treatment as good on a five point Likert scale, 75% experienced a decrease in pain between baseline and final visit of >19.9 mm on a 0–100 mm VAS (a change between −100 mm and −19.9 mm).

Examples of less than ideal uses of p-values

BOSTON UNIVERSITY

Race disparities in discharge location after TJR

- 163,900 participants following Total Joint Replacement
- Institution vs home
  - E.g. Skilled nursing home
- Higher OR = Discharge to institution

Race disparities in discharge location after TJR

RACE	OR	P-value	95% CI
White/private	1.0		Reference

n=163,900

Race disparities in discharge location after TJR

RACE	OR	P-value	95% CI
White/private	1.0		Reference
Black /private	1.69	<0.0001	1.59, 1.80
Hispanic/private	1.28	<0.0001	1.20, 1.38
Black/Medicade	0.46	<0.0001	0.35, 0.60
Hispanic/Medicade	0.57	<0.0001	0.44, 0.73

n=163,900

RCT of inspiratory muscle training

- 20 subjects randomized into one of 4 groups
- Compare subject characteristics between groups

RCT of inspiratory muscle training

Measure	Group 1	Group 2	Group 3	Group 4	Between-Subjects p-value

RCT of inspiratory muscle training

Measure	Group 1	Group 2	Group 3	Group 4	Between-Subjects p-value
Sex	4/6	5/5	6/4	5/5	NS
Age	21.0	21.7	22.8	21.3	NS
BMI	23.0	23.6	22.1	23.7	NS

Convert to percentage and add 95% CI

Add 95% CI

List p-value

Exercise Trial

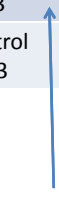
- Randomized trial n=26
- Change in $V0_2$ max
- Intervention and control arms
- Outcome
  - Baseline
  - Post-Intervention
  - Follow-up

Exercise vs Control RCT

Number of Subjects	Baseline	Post-intervention	Follow-up	P-value (Within-Group Changes over time)

Exercise vs Control RCT

Number of Subjects	Baseline	Post-intervention	Follow-up	P-value (Within-Group Changes over time)
Intervention n=13	32.15	36.31	34.67	0.009
Control N=13	25.99	25.55	26.86	0.24



Show change Intervention - Control

Show between group p-value

Thank you!

Questions?

BOSTON UNIVERSITY