Inference and p-values and statistical significance, oh my!

Julia L. Sharp

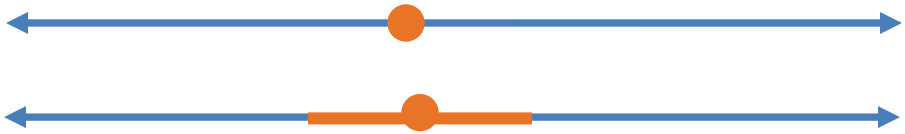
Overview

- Statistical inference and p-values
- Common misconceptions and misinterpretations of p-values
- Statistical significance
- Recommendations for good statistical practice
- Q&A

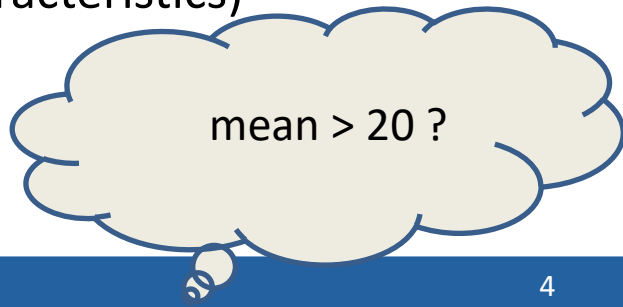
Statistical Inference and p-values

Statistical Inference

- Statistical inference is an informed guess about the population using sample information.
  - Estimate population parameters (characteristics)



  - Hypothesize about population parameters (characteristics)



What is a p-value, anyway?

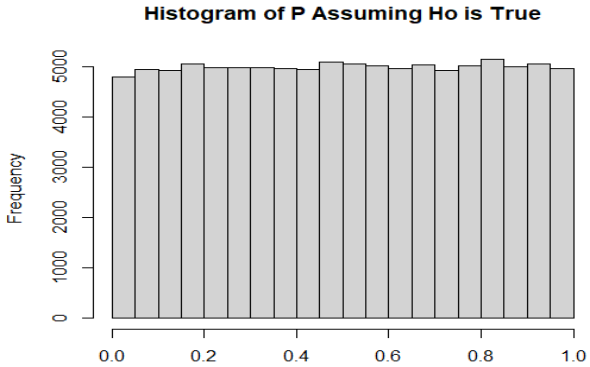
- A probability (but that's not all!)
- A value that is computed assuming that the null hypothesis (no effect) is true
- The definition we teach:
  - The probability, computed assuming the null hypothesis is true, that a new sample collected from the same population would produce a test statistic at least as extreme as the one calculated with the current sample.

ASA Statement Principle 1 About P-values (2016)

- "P-values can indicate how incompatible the data are with a specified statistical model."
  - A p-value is a descriptive summary measure.

An Important Note About the Distribution of the P-value

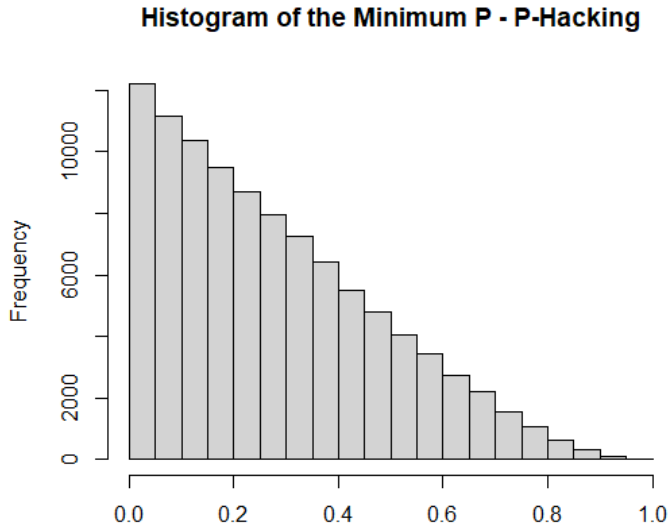
- What does the distribution of the p-value look like?
- Does a large p-value occur more often than a small p-value, since the null hypothesis is assumed to be true?
- Assume Ho is true, all p-values are equally likely.



More About the Distribution of the P-value

- Suppose we have three predictors, X1, X2, and X3 in a multiple regression analysis.
- What would happen if we only reported on the smallest p-value from this analysis?

	Estimate	Standard Error	P-value
Intercept	0.17	0.67	0.804
X1	0.15	0.12	0.218
X2	-0.05	0.06	0.428
X3	0.29	0.16	**0.087**



What is statistical significance?

- P-value is ‘small’, the data are not what we would expect if Ho is true.
- P-value is ‘small’ enough to our liking (specified before the study begins), we reject the null hypothesis and state that the result is ‘statistically significant.’
- ‘Significant’

How is ‘small’ defined?

- ‘Small’ is defined by the researcher before the study begins.
- A precedent has been set that the threshold to indicate ‘smallness’ is 0.05.

Where did the 0.05 significance level come from?

- "...If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty or one in a hundred. Personally, the writer prefers to set a low standard of significance at the 5 percent point, and ignore entirely all results which fail to reach this level. A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance..." (Fisher, 1926)

Common misconceptions and misinterpretations of p-values

3 Common Misconceptions and Misinterpretations of p-values

1.
    a. A p-value is the probability that the studied hypothesis is true.
    b. P tells us the probability that the rejection of the null hypothesis is due to chance. (Wasserstein and Lazar; ASA Statement Principle 2, 2016)

2. A p-value is a measure of the size of an effect or the importance of a result. (Wasserstein and Lazar; ASA Statement Principle 5, 2016)

3. A p-value provides a good measure of evidence regarding a model or hypothesis. (Wasserstein and Lazar; ASA Statement Principle 6, 2016)

Misconception 1a: A p-value is the probability that the studied hypothesis is true.

- The p-value is P(data | Ho is true)
- The p-value is NOT P(Ho is true | data)
- A p-value cannot be both the probability of Ho AND assume the probability of Ho is true.

Misconception 1a: A p-value is the probability that the studied hypothesis is true.

- This would be analogous to saying that the sensitivity of a test is the same as the positive predicted value:
  - Consider a rare disease so that P(disease +)=0.00006 and a very accurate test.
  - Suppose sensitivity = P(test +| disease +) = 0.98
  - False positive rate= P(test + | disease -) = 0.01
  - The positive predictive value =

    P(disease + | test +)

    = $\frac{P(\text{test} + \mid \text{disease} +) * P(\text{disease} +)}{P(\text{test} + \mid \text{disease} +)*P(\text{disease} +) + P(\text{test} + \mid \text{disease} -) * P(\text{disease} -)}$

    = 0.98 * 0.00006/(0.98*0.00006+0.01*0.99994) = 0.00585
  - 0.00585 is not equal to 0.98!

Misconception 1a: A p-value is the probability that the studied hypothesis is true.

- Another way to think about this misconception (Greenland et al., 2016):
  - Suppose p=0.02.
    - Misconception: The null hypothesis has a 2% chance of being true.
    - Truth: The data are NOT very close to the statistical model or null hypothesis explanation.
  - Suppose p=0.60.
    - Misconception: The null hypothesis has a 60% chance of being true.
    - Truth: The data are closer to the statistical model or null hypothesis explanation.

Misconception 1b: P tells us the probability that the rejection of the null hypothesis is due to chance.

- Suppose p=0.10. Misinterpretation: The probability that chance alone produced the association is 0.10.
  - Why is this wrong?
  - "To say that chance alone produced the observed association is logically equivalent to asserting that every assumption used to compute the p-value is correct, *including the null hypothesis*" (Greenland et al. 2016).

Misconception 2: A p-value is a measure of the size of an effect or the importance of a result.

- How can we get a small p-value?

1. Ha is true.
2. Ho is true and a rare event occurred by chance.
3. Ho is true and flexibility in the analysis drove the results

   $(P(p<0.05|Ho) > 0.05)$

4. Ho is false due to some possible research hypothesis other than the one examined.
5. Model assumptions are violated.

How else could we get a small p-value?

6. Sample size.
   - Larger samples from the same population will result in smaller p-values
   - Example:
     - "Women had a slightly higher percentage of transactions for which positive feedback had been given in the year preceding the current transaction (99.60% for women and 99.58% for men, $P < 0.05$)" (Kricheli-Katz and Regev, 2016).

Back to Misconception 2: A p-value is a measure of the size of an effect or the importance of a result.

- "Any effect, no matter how tiny, can produce a small p-value if the sample size or measurement precision is high enough, and large effects may produce unimpressive p-values if the sample size is small or measurements are imprecise (Wasserstein and Lazar; ASA Statement Principle 5, 2016)."

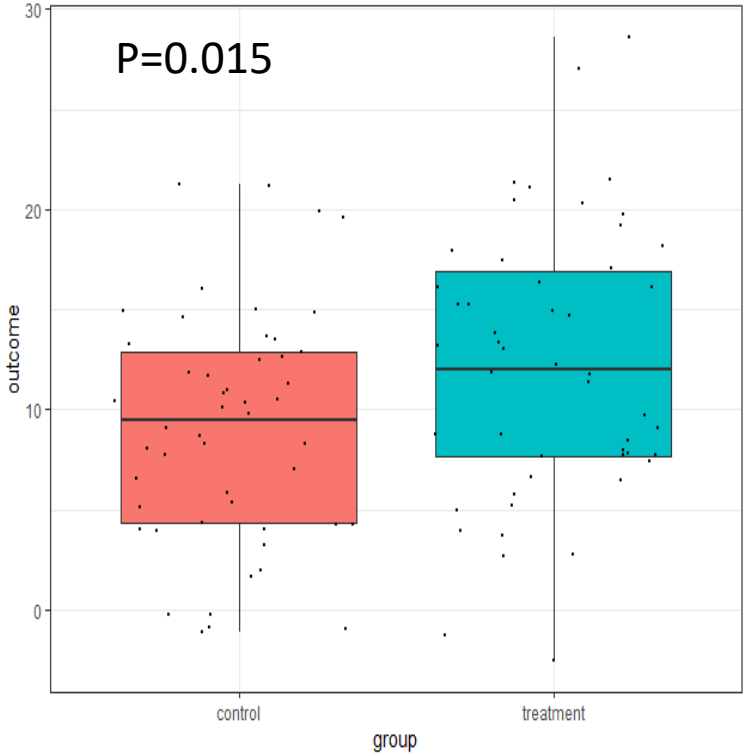
- Lack of statistical significance does not indicate that the effect size is small.

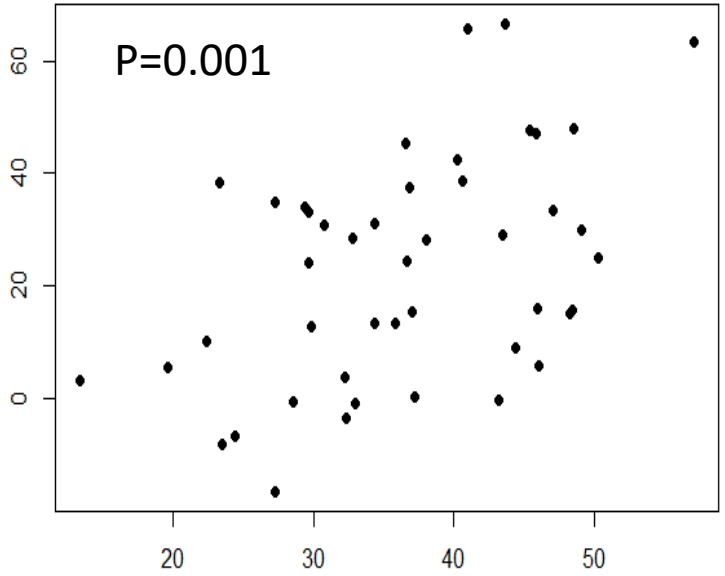
Misconception 2: A p-value is a measure of the size of an effect or the importance of a result.

- Hubbard (2016) summarizes an article by Freeman (1993) whom created a hypothetical data set on medical trials.
- All patients received both treatments A and B.
- The outcome is the patient's preference.

Trial	Number Preferring A	Number Preferring B	% Preferring A
1	15	5	75.0
2	114	86	57.0
3	1,046	954	52.3
4	1,001,455	998,555	50.1

Misconception 2: A p-value is a measure of the size of an effect or the importance of a result.





Misconception 3: A p-value provides a good measure of evidence regarding a model or hypothesis.

- Limited information in a single value without context (Wasserstein and Lazar; ASA Statement Principle 6, 2016).

- What's the difference between 0.049 and 0.051?

- "0.051 is not significant" (https://mchankins.wordpress.com/2013/04/21/still-not-significant-2/):
  - "a robust trend toward significance (p=0.0503)"
  - "did not quite achieve significance (p=0.076)"

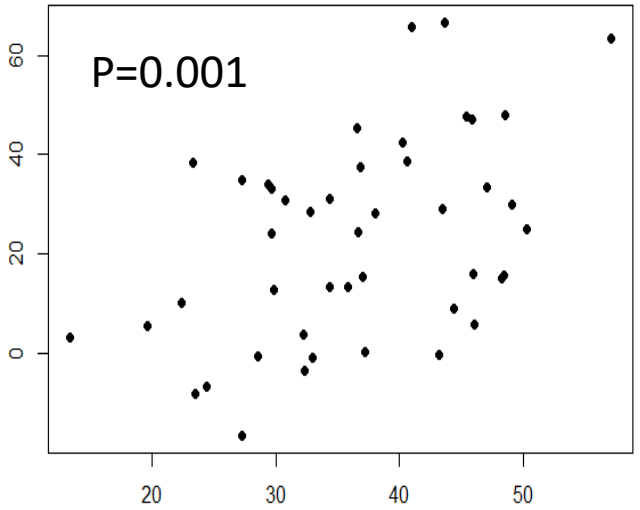
Misconception 3: A p-value provides a good measure of evidence regarding a model or hypothesis.

- A small p-value near 0.05 implies weak evidence against the null hypothesis.
- A small p-value does not mean that the null hypothesis is false.
  - The data are unusual if the assumptions used to compute it are correct (Greenland et al., 2016)

- A large p-value does not imply evidence that favors the null hypothesis.
- A large p-value does not mean that the null hypothesis is true.
  - The data are not unusual if the assumptions used to compute it are correct (Greenland et al., 2016)

Misconception 3: A p-value provides a good measure of evidence regarding a model or hypothesis.

- Statistical significance might be interpreted as a “real” effect.



Misconception 3: A p-value provides a good measure of evidence regarding a model or hypothesis.

- Goodman (2008) states:

This misconception "… is equivalent to saying that the magnitude of effect is not relevant, that only evidence relevant to a scientific conclusion is in the experiment at hand, and that both beliefs and actions flow directly from the statistical results."

The Phrase ‘Statistically Significant’

Wasserstein, Schirm, and Lazar (2019) Summary of *The American Statistician* Special Issue on P-values

- Wasserstein et al. (2019) state that the "…declaration of 'statistical significance' has today become meaningless."

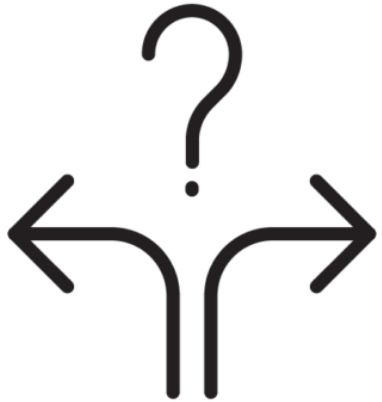
- ASA Statement Principle 3: "Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold."

- Gelman and Stern (2006) summarized this in a noteworthy title: "The Difference Between 'Significant' and 'Not Significant' is not Itself Statistically Significant"

But isn't a binary decision necessary in some situations?

- Yes! Sometimes a binary decision is required, but a p-value should not be the only thing that drives that decision.

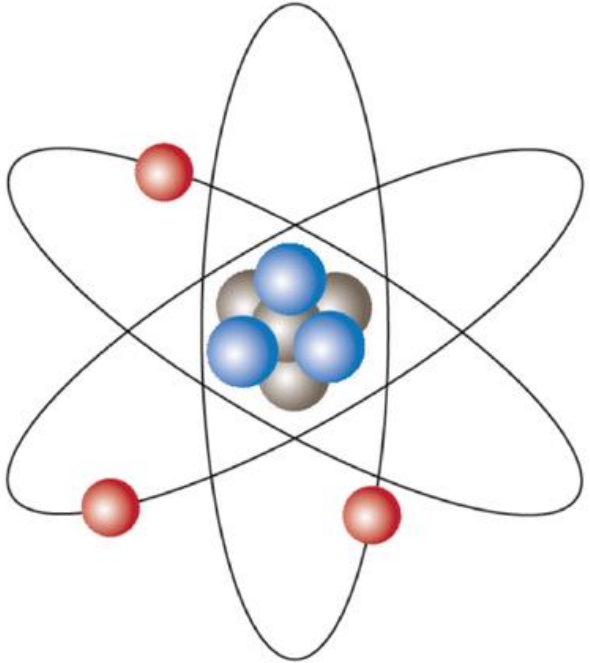
- Contextual factors should be a consideration.



Recommendations for Good Statistical Practice

A.T.O.M.

- Accept uncertainty
- Be Thoughtful
- Be Open
- Be Modest



Accept Uncertainty

- Include a measure of uncertainty with every point estimate.
  - Standard deviation, standard error, confidence intervals.
- Think about the practicalities of the uncertainty boundaries – do they make sense in the context of the study?
- Including measures of uncertainty will remind us that we need scientific expertise and context for interpretation and to make conclusions.

Be Thoughtful

- Critical thinking
- In the disciplinary context, what are the
  - implications,
  - assumptions,
  - previous research,
  - potential outcomes in the context of the disciplinary research?

Be Thoughtful (Continued)

- P-value + context
  - the context for which the data were produced
  - relevant statistics and contextual information.
- Don’t use the binary threshold and call it done.

Be Open

- ASA Statement Principle 4: “Proper inference requires full reporting and transparency.”
- Encourage transparent and complete research practices.
- Integrate statistical and content expertise (Brownstein et al. (2019)).
- Report p-values as supporting continuous descriptive statistics.
- Report p-values for at least one other hypothesized value besides the usual ‘no difference’ hypothesis.
- Provide enough information so that other researchers can conduct relevant other analyses.

Be Modest

- Understand and state limitations of the study.
- Recognize that statistical inference is one component of scientific inference.
- Recognize that a single study is seldom definitive. Claims should not be overstated or over-generalized.

Be Modest (Continued)

- As stated in Amrhein, Trafimow, and Greenland (2019):

"….both scientists and the public confound statistics with reality. But statistical inference is a thought experiment, describing the predictive performance of models about reality. Of necessity, these models are extremely simplified relative to the complexities of actual study conduct and of the reality being studied. Statistical results must eventually mislead us when they are used and communicated as if they present this complex reality, rather than a model for it. This is not a problem of our statistical methods. It is a problem of interpretation and communication of results."

Conclusions

- P-value +
- Report p-values as continuous values.
- There is not a one-size fits all approach to interpreting statistical inference results.
  - The special issue of TAS has 43 different opinions!
  - Follow the ATOM guidance when interpreting statistical inference results.

- Wasserstein et al. (2019):

"…even when there is agreement on the destination, there is disagreement about what road to take."

References

- Brownstein, N. C., Louis, T. A., O'Hagan, A., and Pendergast, J. (2019). "The role of expert judgment in statistical inference and evidence-based decision-making," *The American Statistician*, 73(sup1): 56-68.
- Fisher, R. A. (1926). *Statistical Methods for Research Workers*, Edinburgh: Oliver & Boyd.
- Freeman, P. R. (1993). "The role of p-values in analysing trial results," *Statistics in Medicine*, 12: 1443-1452.
- Gelman, A. and Stern, H. (2006). "The difference between 'significant' and 'not significant' is not itself statistically significant," *The American Statistician*, 60: 328-331.
- Goodman, S. (2008). "Twelve p-value misconceptions," *Seminars in Hematology*, 135-140.
- Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., and Altman, D. G. (2016). "Statistical tests, p-values, confidence intervals, and power: a guide to misinterpretations," *European Journal of Epidemiology*, 31: 337-350.
- Hubbard, R. (2016). *Corrupt Research: The case for reconceptualizing empirical management and social science*, California: Sage.
- Kmetz, J. L. (2019). "Correcting corrupt research: recommendations for the profession to stop misuse of p-values," *The American Statistician*, 73(sup 1): 36-45.
- Kricheli-Katz, T., and Regev, T. (2016). "How many cents on the dollar? Women and men in produce markets," *Science Advances*, 2(2). https://advances.sciencemag.org/content/2/2/e1500599.full
- Wasserstein, R. and Lazar, N. A. (2016). "The ASA's Statement on *p*-values: context, process, and purpose," *The American Statistician*, 70: 129-133.
- Wasserstein, R., Schirm, A. L., and Lazar, N. A. (2019). "Moving to a world beyond '$p < 0.05$'," *The American Statistician*, 73(sup1): 1-19.

Thank You!