P-Values and NHPT

Jacob M. Montgomery

2017

P-Values and NHPT

What are they good for?

- Last class we:
  - Established what a test is
  - Discussed properties of a good test in terms of errors

What are they good for?

- Last class we:
  - Established what a test is
  - Discussed properties of a good test in terms of errors
- This class we:
  - Discuss p-values as a particular approach
  - Consider the Bayesian critique that maybe we should not be doing this at all

Key message

A p-values does not mean **any** of the things below:

- The probability that the alternative hypothesis is true

Key message

A p-values does not mean **any** of the things below:

- The probability that the alternative hypothesis is true
- The probability that the null hypothesis is false

Key message

A p-values does not mean **any** of the things below:

- The probability that the alternative hypothesis is true
- The probability that the null hypothesis is false
- The probability that your finding is a false positive

Key message

A p-values does not mean **any** of the things below:

- The probability that the alternative hypothesis is true
- The probability that the null hypothesis is false
- The probability that your finding is a false positive
- The probability that the finding was produced by random error

Key message

A p-values does not mean **any** of the things below:

- The probability that the alternative hypothesis is true
- The probability that the null hypothesis is false
- The probability that your finding is a false positive
- The probability that the finding was produced by random error
- The effect size was large,

Key message

A p-values does not mean **any** of the things below:

- The probability that the alternative hypothesis is true
- The probability that the null hypothesis is false
- The probability that your finding is a false positive
- The probability that the finding was produced by random error
- The effect size was large, small,

Key message

A p-values does not mean **any** of the things below:

- The probability that the alternative hypothesis is true
- The probability that the null hypothesis is false
- The probability that your finding is a false positive
- The probability that the finding was produced by random error
- The effect size was large, small, important, or unimportant

Key message

A p-values does not mean **any** of the things below:

- The probability that the alternative hypothesis is true
- The probability that the null hypothesis is false
- The probability that your finding is a false positive
- The probability that the finding was produced by random error
- The effect size was large, small, important, or unimportant

So what does it mean?

Defining a p-value broadly

Suppose that for every $\alpha \in (0, 1)$ we have a size α test with rejection region R_α. Then,

$$\text{p-value} = \mathit{inf}\,\{\alpha : t(\mathbf{x}) \in R_\alpha\}$$

- In words this means that the p-value is the smallest level at which we can reject H_0.

Getting down to business

- First we set up a null hypothesis.
- Then we assume that the null hypothesis is true.
- Then we consider a set of tests with size $\alpha \in (0, 1)$
- The p-value is the size of the largest test at which we can still reject H_0.

A critical value perspective

- Suppose that the size α test is of the form "reject H_0 iff $t(\mathbf{x}) \geq c_\alpha$"
- In this case the p-value is

$$\sup_{\theta \in \Theta_0} P(t(\mathbf{X}) \geq t(\mathbf{x})))$$

where $\mathbf{X}$ is the random variable and $\mathbf{x}$ is the observed value.

More intuitively

- A p-value is a measure of surprise

More intuitively

- A p-value is a measure of surprise
- It is the probability of observing a value of some statistic – ASSUMING THE NULL HYPOTHESIS IS TRUE – that is the same *or more extreme* than what was actually observed.

Let me explain this to you like a 19-year old

We are going to try and support our research hypothesis using a technique called *proof by contradiction*.

Let me explain this to you like a 19-year old

We are going to try and support our research hypothesis using a technique called *proof by contradiction*.

- First, we set up a null hypothesis (determined sceptic).

Let me explain this to you like a 19-year old

We are going to try and support our research hypothesis using a technique called *proof by contradiction*.

- First, we set up a null hypothesis (determined sceptic).
- Second, we use a sample statistic to show that the data we have observed would be very unlikely if the null hypothesis were true.

Let me explain this to you like a 19-year old

We are going to try and support our research hypothesis using a technique called *proof by contradiction*.

- First, we set up a null hypothesis (determined sceptic).
- Second, we use a sample statistic to show that the data we have observed would be very unlikely if the null hypothesis were true.
- Third, we (kind of) conclude that our research hypothesis is true. This is usually called the alternative hypothesis.

Let me explain this to you like a 19-year old

We are going to try and support our research hypothesis using a technique called *proof by contradiction*.

- First, we set up a null hypothesis (determined sceptic).
- Second, we use a sample statistic to show that the data we have observed would be very unlikely if the null hypothesis were true.
- Third, we (kind of) conclude that our research hypothesis is true. This is usually called the alternative hypothesis.
- To do this, we first need to specify our hypotheses.

We calculate a statistic that summarizes how much our data differs from what we would have expected to observe *if the null hypothesis were true*.

We calculate a statistic that summarizes how much our data differs from what we would have expected to observe *if the null hypothesis were true*. Usually this is something equivalent to a Z-statistic.

*A **P-Value** is a measure of surprise.*

*A **P-Value** is a measure of surprise. We ask, "If the null hypothesis is true, how likely is it that we would observe a test-statistic this extreme **or more**?"*

*A **P-Value** is a measure of surprise. We ask, "If the null hypothesis is true, how likely is it that we would observe a test-statistic this extreme **or more**?"*

- P-Values are very difficult for many people to understand.
- Smaller P-Values more strongly contradict the null.

How surprised would you have to be in order to conclude that the *null hypothesis* is false?

- Usually, $p \leq 0.05 \Rightarrow$ statistically significant result
- We would observe a test-statistic this extreme or more $1/20$ times if the null hypothesis was true.
- More generically, we want $p \leq \alpha$

How surprised would you have to be in order to conclude that the *null hypothesis* is false?

- Usually, $p \leq 0.05 \Rightarrow$ statistically significant result
- We would observe a test-statistic this extreme or more $1/20$ times if the null hypothesis was true.
- More generically, we want $p \leq \alpha$
- We "reject the null" and conclude that the evidence supports the alternative hypothesis.

Example

According to a union agreement, the mean income for all senior-level assembly-line workers in a large company equals \$525 per week. A representative of a women's group decides to analyze whether the mean income μ for female employees matches this norm. For a random sample of 36 female employees, $\bar{y} = \$495$ and $s = \$75$.

- Calculate the test using $\alpha = .01$.
- Draw a picture that represents the $p - value$.

Example

We flip a coin 20 times and observe 18 heads. Find the p-value for the null hypothesis that we have a fair coin.

Now let's put on our big kid pants

- The p-value you calculate is a function of the *null hypothesis you set up*.
- You can get a significant value because:
  - Your null hypothesis is false

Now let's put on our big kid pants

- The p-value you calculate is a function of the *null hypothesis you set up*.
- You can get a significant value because:
  - Your null hypothesis is false
  - You are underpowered

Now let's put on our big kid pants

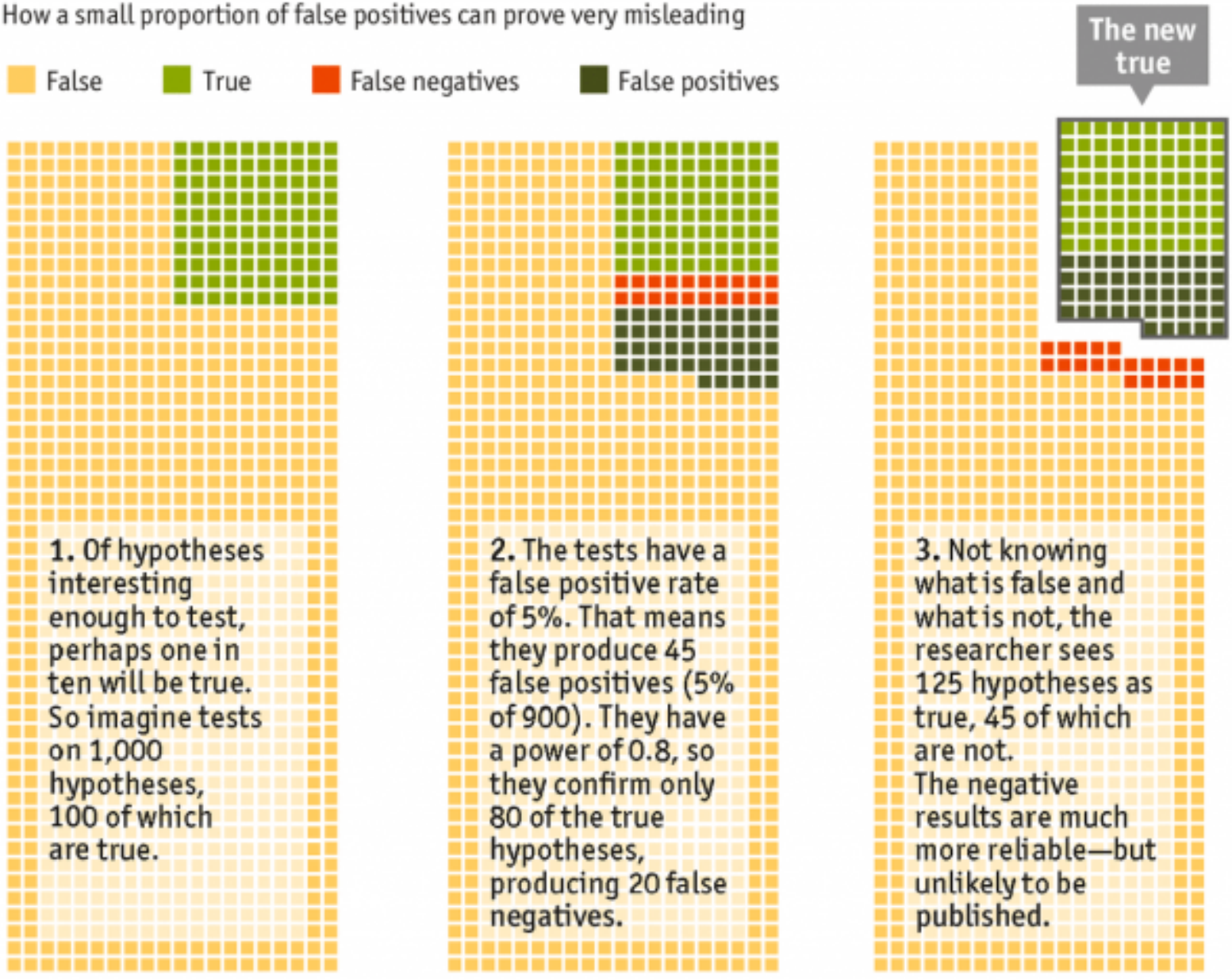
- The p-value you calculate is a function of the *null hypothesis you set up*.
- You can get a significant value because:
  - Your null hypothesis is false
  - You are underpowered (BEWARE SMALL SAMPLES)
  - You tested a lot of hypotheses

Now let's put on our big kid pants

- The p-value you calculate is a function of the *null hypothesis you set up*.
- You can get a significant value because:
  - Your null hypothesis is false
  - You are underpowered (BEWARE SMALL SAMPLES)
  - You tested a lot of hypotheses
  - You estimated the sampling distribution wrong

Unlikely results

How a small proportion of false positives can prove very misleading



Source: *The Economist*

This is kind of a big deal

- We are increasingly aware that we are in a "replication crisis"
- P-values are at the heart of it (and there is no easy solution)
- As a practical matter:
  - Pre-registeration
  - Lower acceptable p-value standards for new findings
  - Wait for the meta analysis
  - DO NOT TRUST UNDERPOWERED STUDIES
  - Publish replications
  - Don't believe everything you read (but don't be a jerk about it)

This is kind of a big deal

- We are increasingly aware that we are in a "replication crisis"
- P-values are at the heart of it (and there is no easy solution)
- As a practical matter:
  - Pre-registeration
  - Lower acceptable p-value standards for new findings
  - Wait for the meta analysis
  - DO NOT TRUST UNDERPOWERED STUDIES
  - Publish replications
  - Don't believe everything you read (but don't be a jerk about it)
- Maybe don't do hypothesis testing? But what could we do instead?

Bayes Factors

- Consider the case where θ is a scalar and we are testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$.

Bayes Factors

- Consider the case where θ is a scalar and we are testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$.
- From a Bayesian perspective, we want to directly summarize the strength of the evidence for H_1 relative to H_0.

Bayes Factors

- Consider the case where θ is a scalar and we are testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$.
- From a Bayesian perspective, we want to directly summarize the strength of the evidence for H_1 relative to H_0.
- Using Bayes' rule, we get

$$\frac{\Pr(H_1|\mathbf{x})}{\Pr(H_0|\mathbf{x})} = \frac{f(\mathbf{x}|H_1)}{f(\mathbf{x}|H_0)} \times \frac{\pi(H_1)}{\pi(H_0)} \equiv \text{Bayes Factor} \times \text{Prior odds}$$

Bayes Factors

- Consider the case where θ is a scalar and we are testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$.
- From a Bayesian perspective, we want to directly summarize the strength of the evidence for H_1 relative to H_0.
- Using Bayes' rule, we get

$$\frac{\Pr(H_1|\mathbf{x})}{\Pr(H_0|\mathbf{x})} = \frac{f(\mathbf{x}|H_1)}{f(\mathbf{x}|H_0)} \times \frac{\pi(H_1)}{\pi(H_0)} \equiv \text{Bayes Factor} \times \text{Prior odds}$$

- This can be re-written as:

$$\frac{L(H_1)}{L(H_0)} \frac{\pi(H_1)}{\pi(H_0)}$$

- Note that if $H_1 = MLE$, the first part is the inverse of the likelihood ratio

- Note that if $H_1 = MLE$, the first part is the inverse of the likelihood ratio and if we give the same prior odds to both theories it is exactly the same.
- Let's unpack that a bit:

$$\frac{\Pr(H_1|\mathbf{x})}{\Pr(H_0|\mathbf{x})}$$

Jefferry's proposed the following rules for comparing two models
$BF(M_1 : M2) = \frac{\Pr(M_1|\mathbf{x})}{\Pr(M_2|\mathbf{x})}$

- $B(x) \geq 1 \rightarrow$ Model 1 supported
- $0.316 \leq B(x) < 1 \rightarrow$ Minimal evidence against Model 1 (Note $0.316 = 10^{-1/2}$)
- $0.1 \leq B(x) < 0.316 \rightarrow$ Substantial evidence against Model 1
- $0.01 \leq B(x) < 0.1 \rightarrow$ Strong evidence against Model 1
- $B(x) < 0.01 \rightarrow$ Decisive evidence against Model 1

Problems with Bayes factors

- The flaws with this paradigm open up when we consider the case where θ is a scalar and we are testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ where θ is a scalar that can take on many values.
- In these cases, improper (flat) priors can lead to non-defined Bayes factors

Problems with Bayes factors

- The flaws with this paradigm open up when we consider the case where θ is a scalar and we are testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ where θ is a scalar that can take on many values.
- In these cases, improper (flat) priors can lead to non-defined Bayes factors
- Bayes factors can become seriously sensitive to priors especially when they are "vague."
- We have to integrate over the entire parameter space

Problems with Bayes factors

- The flaws with this paradigm open up when we consider the case where θ is a scalar and we are testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ where θ is a scalar that can take on many values.
- In these cases, improper (flat) priors can lead to non-defined Bayes factors
- Bayes factors can become seriously sensitive to priors especially when they are "vague."
- We have to integrate over the entire parameter space which may not be tractable.
- For more, see: `http://www.stat.cmu.edu/~kass/papers/bayesfactors.pdf`

P-values and Bayes factors

- One of the nice things about this setup is that

$$P(M_0|\mathbf{x}) = \frac{BF(M_0 : M_1)}{1 + BF(M_0 : M_1)}$$

- Where the hell did that come from?

P-values and Bayes factors

- One of the nice things about this setup is that

$$P(M_0|\mathbf{x}) = \frac{BF(M_0 : M_1)}{1 + BF(M_0 : M_1)}$$

- Where the hell did that come from?
- And what does that mean anyways?

P-values and Bayes factors

- One of the nice things about this setup is that

$$P(M_0|\mathbf{x}) = \frac{BF(M_0 : M_1)}{1 + BF(M_0 : M_1)}$$

- Where the hell did that come from?
- And what does that mean anyways?
- This number will usually *not* correspond well with a p-value

Simulation 1

- Generate 100000 experiments with two variables x and y
- Both should be random normal samples of size $n = 100$, $mu = 100$, $\sigma = 20$.
- Perform a t-test and grab the p-value (difference of means)
- Plot a histogram of the p-values that result
- What does this mean? Are you surprised?

Simulation 2

- Generate 100000 experiments with two variables x and y
- X should be random normal samples of size $n = 100$, $mu = 100$, $\sigma = 20$.
- Y should be random normal samples of size $n = 100$, $mu = 103$, $\sigma = 20$.
- Perform a t-test and grab the p-value
- Plot a histogram of the p-values that result
- What does this mean? Are you surprised?

Simulation 3

- Let ϕ represent the proportion of cases where the null hypothesis is true
- Let α be the probability of rejecting the null hypothesis if it is false.
- Let $(1 - \beta)$ be the probability of rejecting given that the null hypothesis is false.
- If we test many hypotheses, the false positive rate is then:

$$\frac{\alpha\phi}{\alpha\phi(1-\beta)(1-\phi)}$$

- If $\frac{\phi}{1-\phi}$ is 1/5, what is the false positive rate when $\alpha = .05$ and $\beta = .75$?
- If $\frac{\phi}{1-\phi}$ is 1/20, what is the false positive rate when $\alpha = .05$ and $\beta = .75$?
- If $\frac{\phi}{1-\phi}$ is 1/40, what is the false positive rate when $\alpha = .05$ and $\beta = .6$?

Exercise

If I, Prof. Montgomery, flipped a coin 20 times and it came up heads 18 times, what is the probability that it is a fair coin? - Set the problem up as a complete Bayesian problem. - Set the problem up as a choice of two discrete problems and solve explicitly.