P-values are Random Variables

Duncan Murdoch

Department of Statistical and Actuarial Sciences
University of Western Ontario

October 4, 2007

Outline

This is joint work with Yu-Ling Tsai and James Adcock.

Outline

Teaching introductory statistics

- I've been teaching hypothesis testing in introductory statistics courses since 1988.
- Over time I have gradually changed the way I teach hypothesis testing and p-values; this talk describes my current ideas.
- A few recent events triggered the urge to write this up...

Teaching introductory statistics

- I've been teaching hypothesis testing in introductory statistics courses since 1988.
- Over time I have gradually changed the way I teach hypothesis testing and p-values; this talk describes my current ideas.
- A few recent events triggered the urge to write this up...

Teaching introductory statistics

- I've been teaching hypothesis testing in introductory statistics courses since 1988.
- Over time I have gradually changed the way I teach hypothesis testing and p-values; this talk describes my current ideas.
- A few recent events triggered the urge to write this up...

A trigger

On the R-help list in May 2006, regarding inconsistent results ($p = 0.7767, p = 0.9059, p = 0.1887$) when running a normality test on randomly generated data:

> *I mistakenly had thought the p-values would be more stable since I am artificially creating a random normal distribution. Is this expected for a normality test or is this an issue with how* `rnorm` *is producing random numbers? I guess if I run it many times, I would find that I would get many large values for the p-value?*
> *– Name withheld*

A trigger

On the R-help list in May 2006, regarding inconsistent results ($p = 0.7767, p = 0.9059, p = 0.1887$) when running a normality test on randomly generated data:

> *I mistakenly had thought the p-values would be more stable since I am artificially creating a random normal distribution. Is this expected for a normality test or is this an issue with how* `rnorm` *is producing random numbers? I guess if I run it many times, I would find that I would get many large values for the p-value?*
> *– Name withheld*

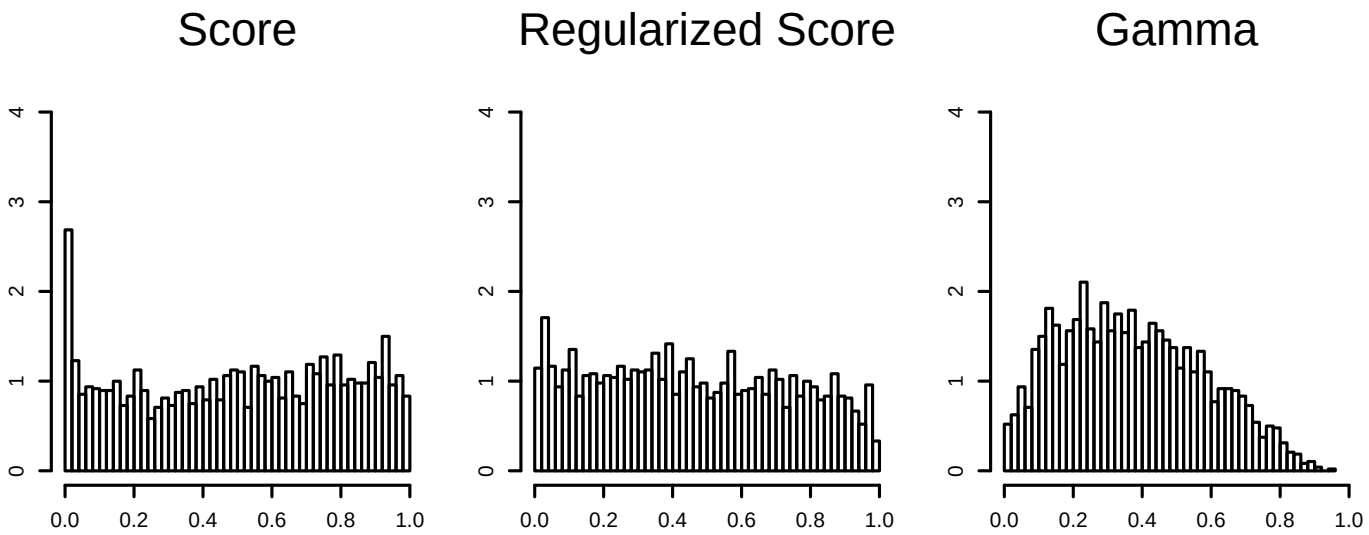
A response

Discussion followed on why this was not a reasonable expectation, including this:

> *We see this misunderstanding worryingly often. Worrying because it reveals that a fundamental aspect of statistical inference has not been grasped: that p-values are designed to be (approximately) uniformly distributed and fall below any given level with the stated probability, when the null hypothesis is true.*
> *– Peter Dalgaard*

A second trigger

At her thesis defence, Yu-Ling presented histograms of simulated p-values to illustrate deficiencies in some asymptotic approximations:



One of the examiners questioned this way of presenting the results.

Advice on the web

On a medical school research methods course web page:

> *The t-test value for the stress test indicates that the probability that the null hypothesis is true is smaller than one-in-twenty.*

Advice on the web

On a medical school research methods course web page:

> *The t-test value for the stress test indicates that the probability that the null hypothesis is true is smaller than one-in-twenty.*

I pointed out that this isn't correct,

Advice on the web

On a medical school research methods course web page:

> *The t-test value for the stress test indicates that the probability that the null hypothesis is true is smaller than one-in-twenty.*

I pointed out that this isn't correct, and received the response:

> *[This] is written the way it is to give students a way to make decisions about statistical results in journal articles. It is not for people learning about statistics. Thus, the interpretation of p-values is correct enough.*

Outline

The definition of a p-value

Given a null hypothesis H_0, an alternative H_1, and a test statistic T, the p-value is

> *the probability, computed assuming that H_0 is true, that the test statistic would take a value as extreme or more extreme than that actually observed.*
> *– Moore, D. S. (2007), The Basic Practice of Statistics*

In the typical case where large values of T are considered to be extreme, this is

$$p = P(T \geq t_{obs} | H_0)$$

Interpretation of a p-value

How should we interpret p?

> *the smaller the p-value, the stronger the evidence against H_0 provided by the data.*
> *– Moore, D. S. (2007), The Basic Practice of Statistics*

How are p-values interpreted in the wild?

The definition is $p = P(T \geq t_{obs}|H_0)$. Some common misconceptions (from Wikipedia):

1. the probability that the null hypothesis is true, i.e. $P(H_0|\text{data})$.
2. the probability that a finding is "merely a fluke".
3. the probability of falsely rejecting the null hypothesis, i.e. $P[H_0 \cap (t_{obs} \geq t_{crit})]$.
4. the probability that a replicating experiment would not yield the same conclusion.

How are p-values interpreted in the wild?

The definition is $p = P(T \geq t_{obs}|H_0)$. Some common misconceptions (from Wikipedia):

1. the probability that the null hypothesis is true, i.e. $P(H_0|\text{data})$.
2. the probability that a finding is "merely a fluke".
3. the probability of falsely rejecting the null hypothesis, i.e. $P[H_0 \cap (t_{obs} \geq t_{crit})]$.
4. the probability that a replicating experiment would not yield the same conclusion.

How are p-values interpreted in the wild?

The definition is $p = P(T \geq t_{obs}|H_0)$. Some common misconceptions (from Wikipedia):

1. the probability that the null hypothesis is true, i.e. $P(H_0|\text{data})$.
2. the probability that a finding is "merely a fluke".
3. the probability of falsely rejecting the null hypothesis, i.e. $P[H_0 \cap (t_{obs} \geq t_{crit})]$.
4. the probability that a replicating experiment would not yield the same conclusion.

How are p-values interpreted in the wild?

The definition is $p = P(T \geq t_{obs}|H_0)$. Some common misconceptions (from Wikipedia):

1. the probability that the null hypothesis is true, i.e. $P(H_0|\text{data})$.
2. the probability that a finding is "merely a fluke".
3. the probability of falsely rejecting the null hypothesis, i.e. $P[H_0 \cap (t_{obs} \geq t_{crit})]$.
4. the probability that a replicating experiment would not yield the same conclusion.

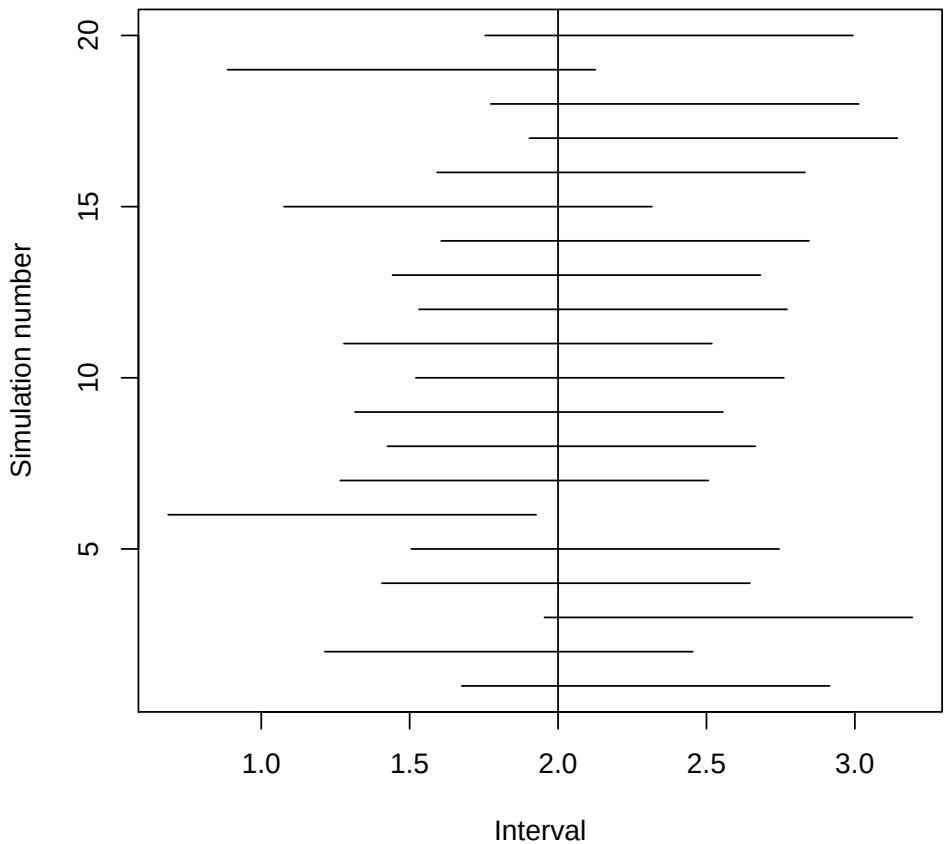
Outline

How do we teach confidence intervals?

Definitions are awkward:

> *A level C confidence interval for a parameter is an interval computed from sample data by a method that has probability C of producing an interval containing the true value of the parameter.*
> *–Moore and McCabe (2003), Introduction to the Practice of Statistics*

But a picture tells the story...



We should emphasize that p-values are random variables

- Start by saying the p-value is simply a transformation of the test statistic.
- If the audience has enough mathematical sophistication, give a formula:

$$p = 1 - F(t_{obs})$$

  where $F(\cdot)$ is the CDF of T under H_0.
- Show (or state) that this results in $p \sim \text{Unif}(0, 1)$ under H_0.
- Mention that a good T will tend to be larger under H_1, so p will be smaller.
- THEN give Moore's statement, as one justification for this definition.

We should emphasize that p-values are random variables

- Start by saying the p-value is simply a transformation of the test statistic.
- If the audience has enough mathematical sophistication, give a formula:

$$p = 1 - F(t_{obs})$$

  where $F(\cdot)$ is the CDF of T under H_0.
- Show (or state) that this results in $p \sim \text{Unif}(0, 1)$ under H_0.
- Mention that a good T will tend to be larger under H_1, so p will be smaller.
- THEN give Moore's statement, as one justification for this definition.

We should emphasize that p-values are random variables

- Start by saying the p-value is simply a transformation of the test statistic.
- If the audience has enough mathematical sophistication, give a formula:

$$p = 1 - F(t_{obs})$$

  where $F(\cdot)$ is the CDF of T under H_0.
- Show (or state) that this results in $p \sim \text{Unif}(0, 1)$ under H_0.
- Mention that a good T will tend to be larger under H_1, so p will be smaller.
- THEN give Moore's statement, as one justification for this definition.

We should emphasize that p-values are random variables

- Start by saying the p-value is simply a transformation of the test statistic.
- If the audience has enough mathematical sophistication, give a formula:

$$p = 1 - F(t_{obs})$$

  where $F(\cdot)$ is the CDF of T under H_0.
- Show (or state) that this results in $p \sim \text{Unif}(0, 1)$ under H_0.
- Mention that a good T will tend to be larger under H_1, so p will be smaller.
- THEN give Moore's statement, as one justification for this definition.

We should emphasize that p-values are random variables

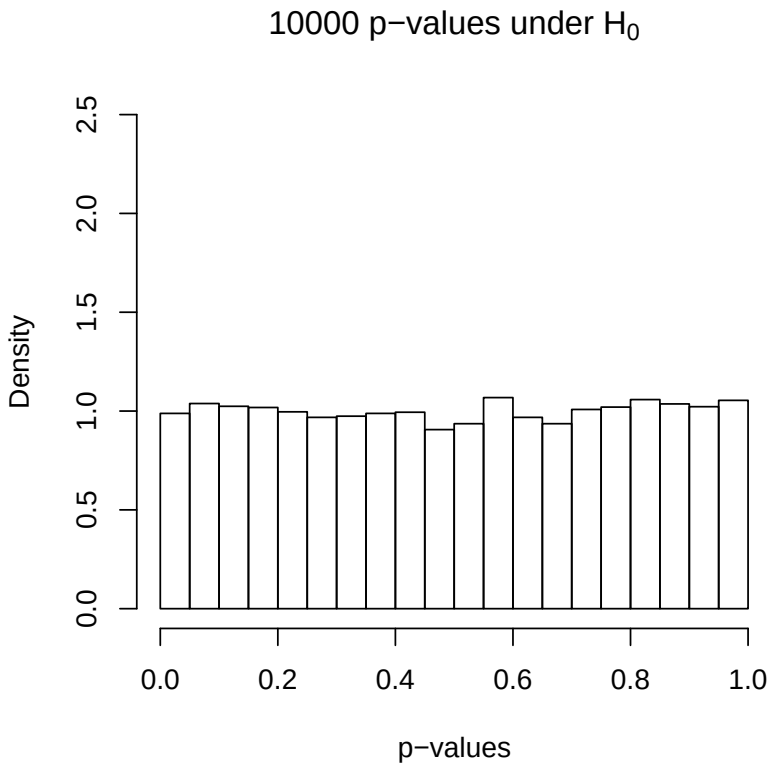
- Start by saying the p-value is simply a transformation of the test statistic.
- If the audience has enough mathematical sophistication, give a formula:

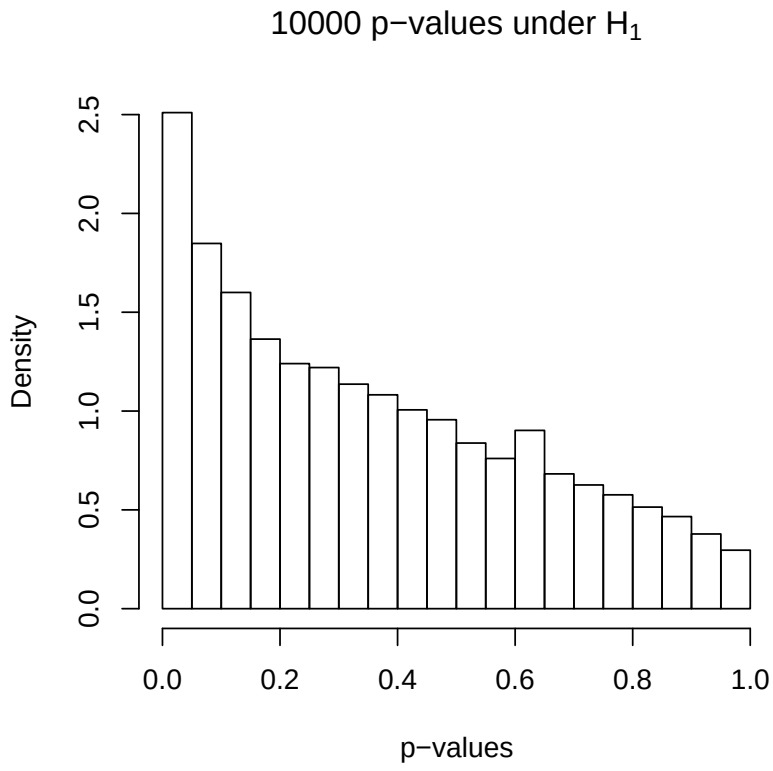
$$p = 1 - F(t_{obs})$$

  where $F(\cdot)$ is the CDF of T under H_0.
- Show (or state) that this results in $p \sim \text{Unif}(0, 1)$ under H_0.
- Mention that a good T will tend to be larger under H_1, so p will be smaller.
- THEN give Moore's statement, as one justification for this definition.

Show pictures!

P-values are random variables, so it is natural to study their distribution by simulation.



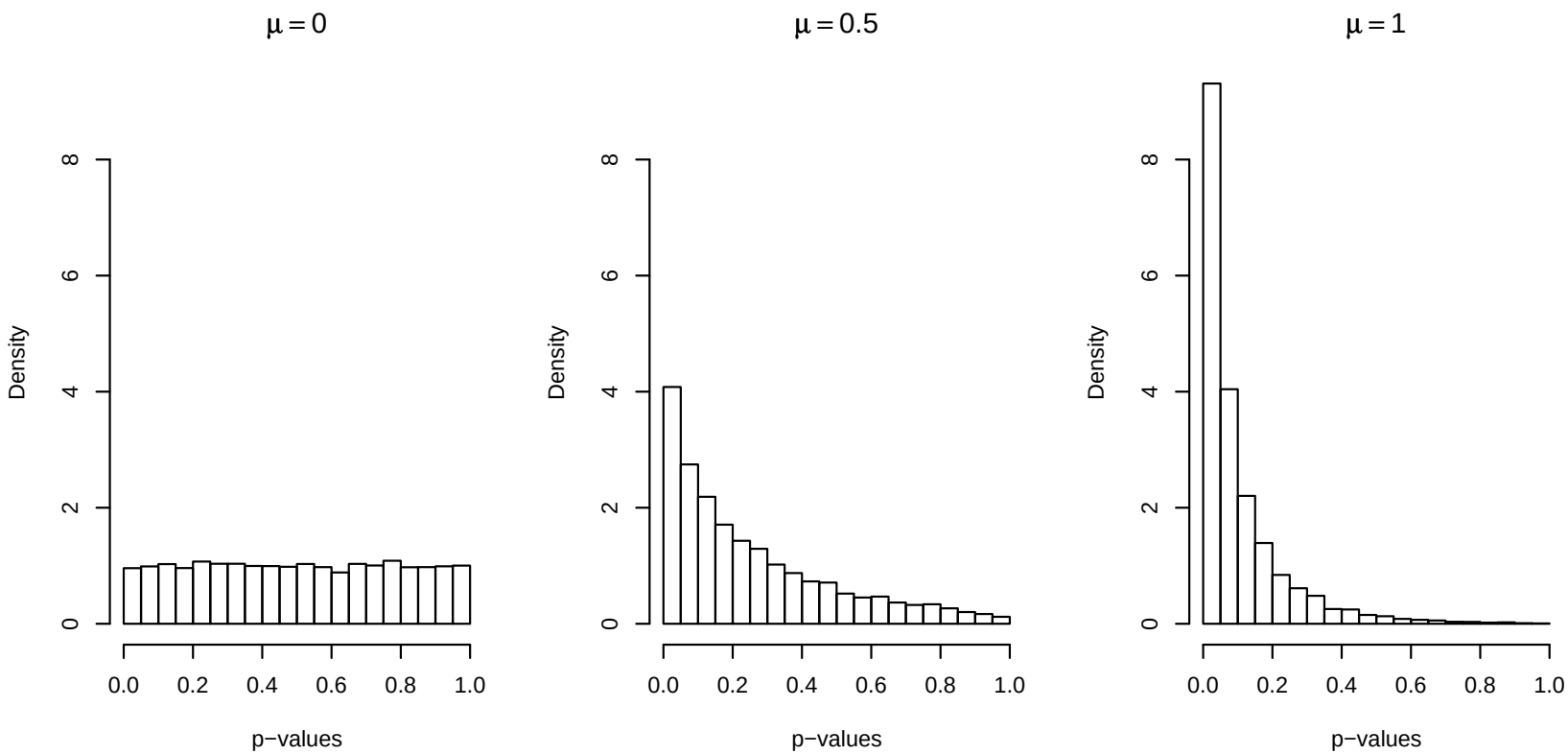


Histograms are easily understood.

Outline

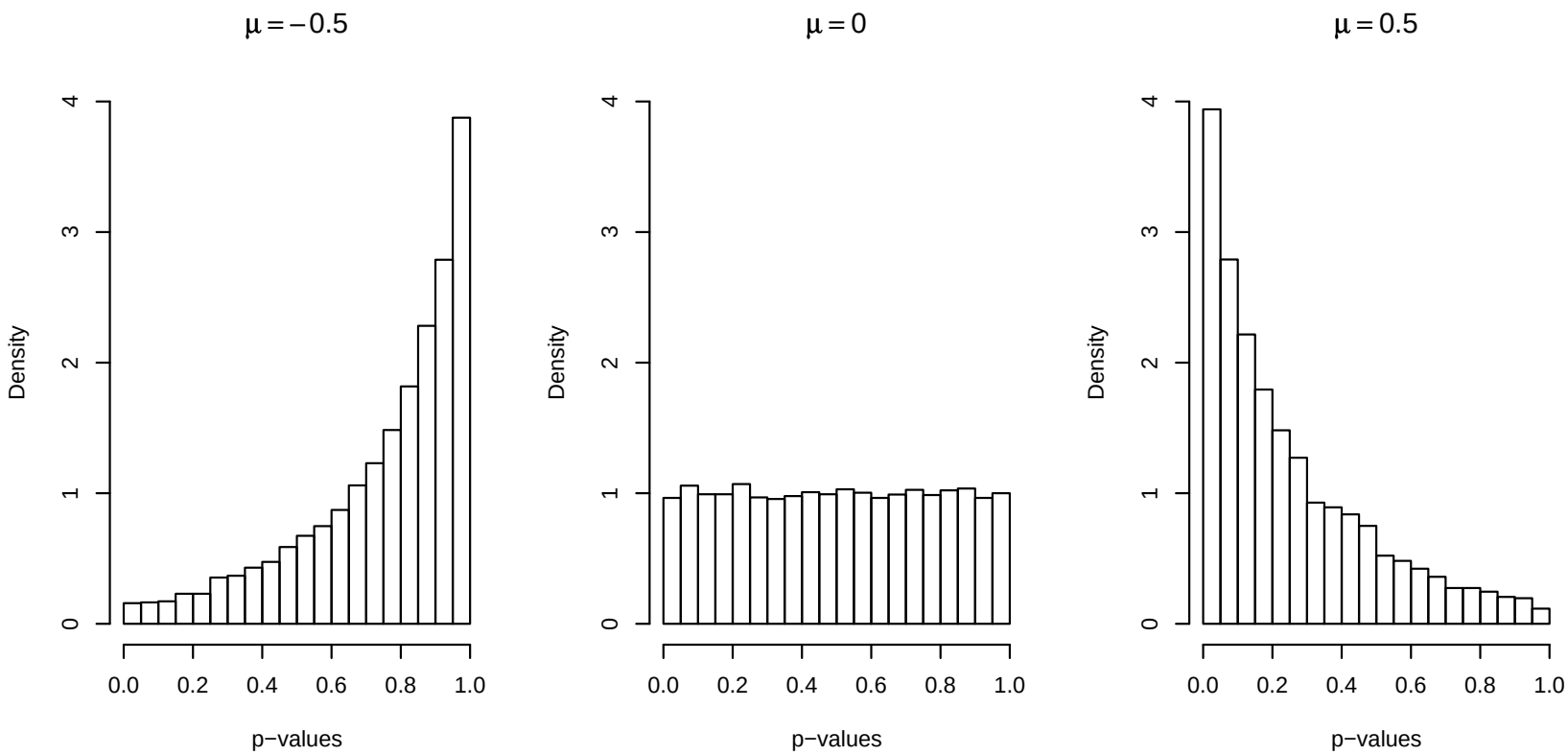
One-sample t-test

- Data $X_1, \ldots, X_4 \sim N(\mu, \sigma^2)$ i.i.d.
- Hypotheses $H_0 : \mu = 0$ versus $H_1 : \mu > 0$
- Test statistic $T = \frac{\bar{X}}{s/\sqrt{4}} \sim t_{(3)}$ under H_0



Composite null hypotheses

- When H_0 is composite, it may not uniquely determine the distribution.
- Hypotheses $H_0 : \mu \leq 0$ versus $H_1 : \mu > 0$



Violations of assumptions

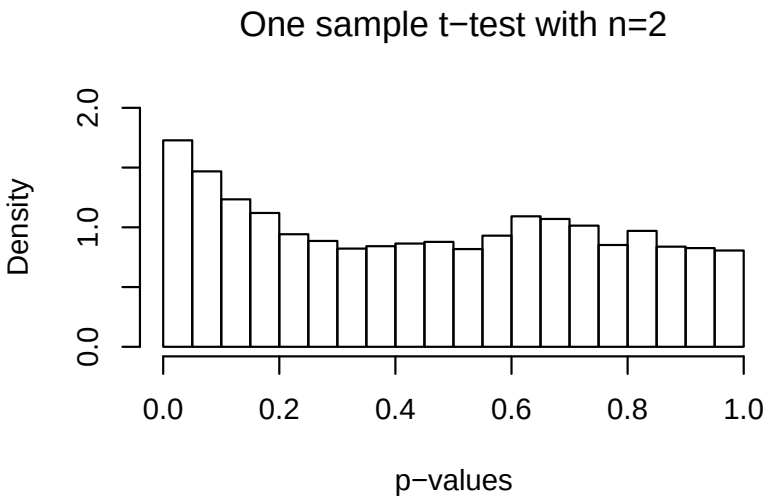
- If our assumptions are violated, the null distribution of p may be distorted, but larger samples often improve the approximations.
- Example: Assume data are $N(\mu, \sigma^2)$, but they really Exponential(1).
  1. One-sample t-test, $H_0 : \mu = 1$ versus $H_1 : \mu \neq 1$
  2. Two-sample t-test, $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$
- Note that H_0 is true in both cases. Let's look at the null distributions.

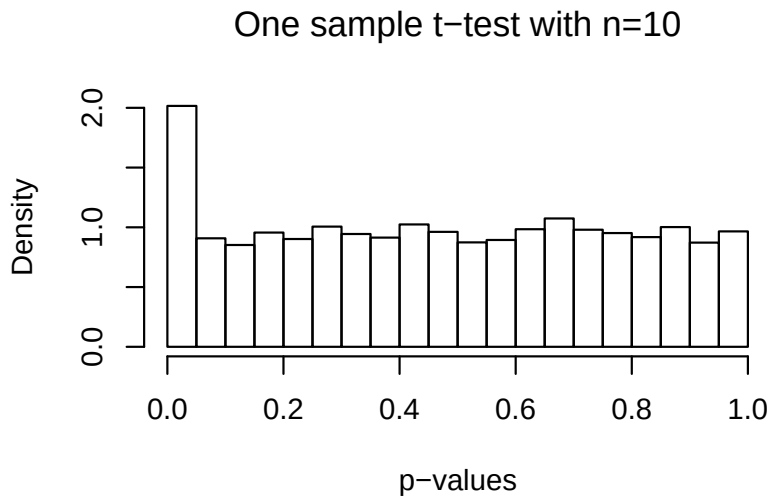
Violations of assumptions

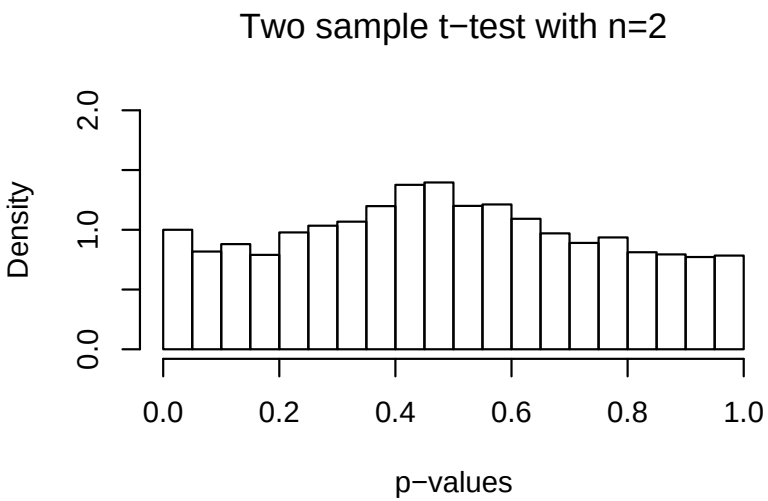
- If our assumptions are violated, the null distribution of p may be distorted, but larger samples often improve the approximations.
- Example: Assume data are $N(\mu, \sigma^2)$, but they really Exponential(1).
  1. One-sample t-test, $H_0 : \mu = 1$ versus $H_1 : \mu \neq 1$
  2. Two-sample t-test, $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$
- Note that H_0 is true in both cases. Let's look at the null distributions.

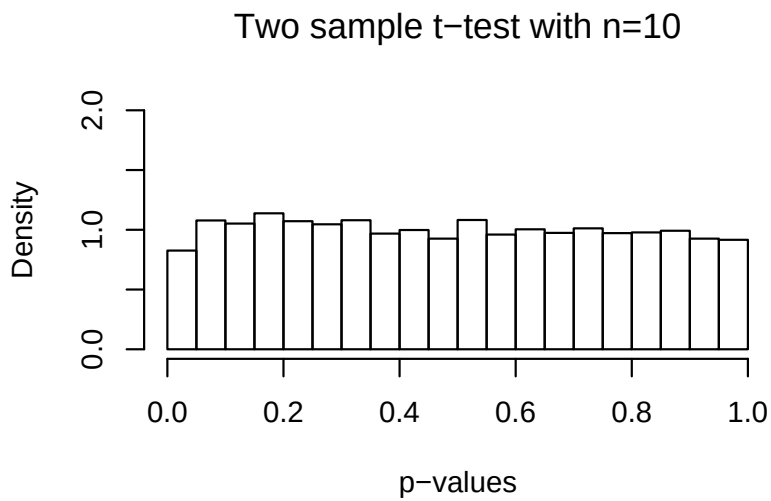
Violations of assumptions

- If our assumptions are violated, the null distribution of p may be distorted, but larger samples often improve the approximations.
- Example: Assume data are $N(\mu, \sigma^2)$, but they really Exponential(1).
  1. One-sample t-test, $H_0 : \mu = 1$ versus $H_1 : \mu \neq 1$
  2. Two-sample t-test, $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$
- Note that H_0 is true in both cases. Let's look at the null distributions.


One sample t−test with n=2
Density
0.0
1.0
2.0
0.0
0.2
0.4
0.6
0.8
1.0
p−values


One sample t−test with n=10
Density
0.0
1.0
2.0
0.0
0.2
0.4
0.6
0.8
1.0
p−values


Two sample t−test with n=2
Density
0.0
1.0
2.0
0.0
0.2
0.4
0.6
0.8
1.0
p−values


Two sample t−test with n=10
Density
0.0
1.0
2.0
0.0
0.2
0.4
0.6
0.8
1.0
p−values

Discrete data

- With discrete data, p-values inherit a discrete distribution. You won't see Unif$(0, 1)$ under the null.
- This makes display of simulated p-values harder, but the empirical CDF is not too bad.

Discrete data

- With discrete data, p-values inherit a discrete distribution. You won't see Unif$(0, 1)$ under the null.
- This makes display of simulated p-values harder, but the empirical CDF is not too bad.

Test for independence in a 2×2 table

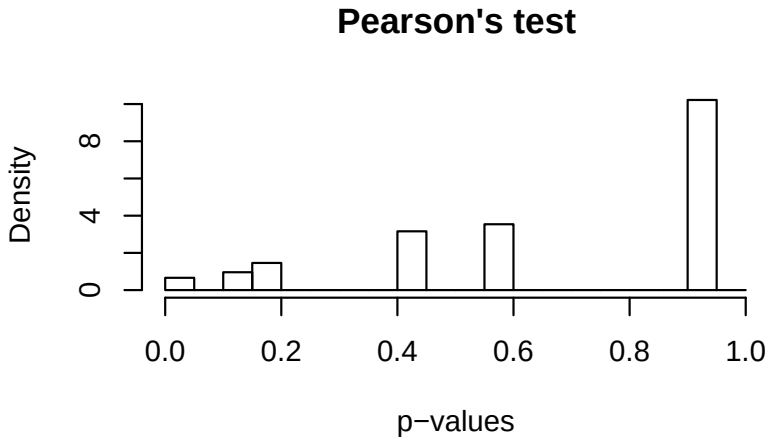
Example from Tamhane and Dunlop (2000), *Statistics and Data Analysis*

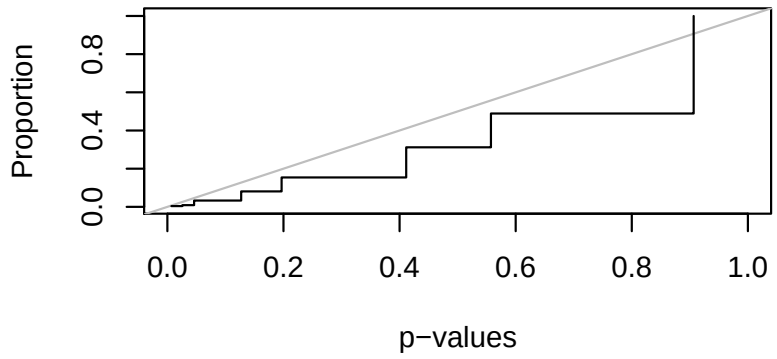
	Success	Failure	Total
Prednisone	14	7	21
Prednisone + VCR	38	4	42
Total	52	11	63

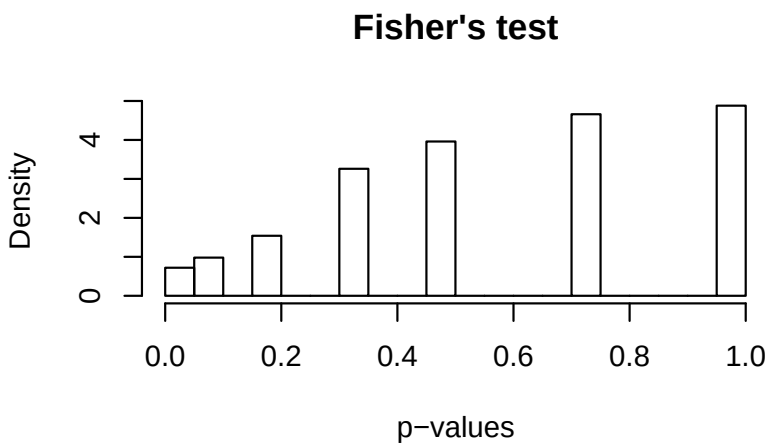
Pearson's chi-square p-value: 0.04608
Fisher's exact p-value: 0.03232

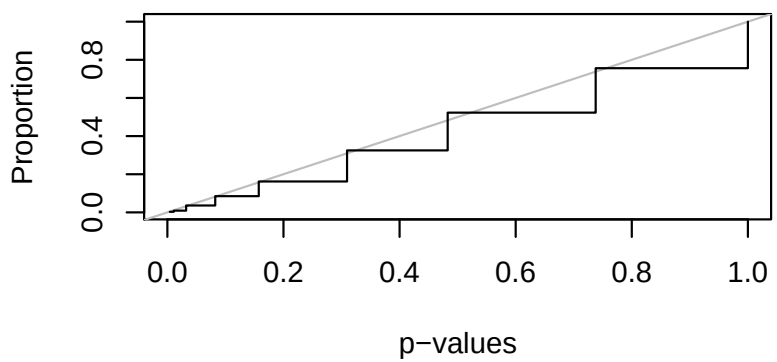
What are the null distributions like?

Null tables with fixed margins.

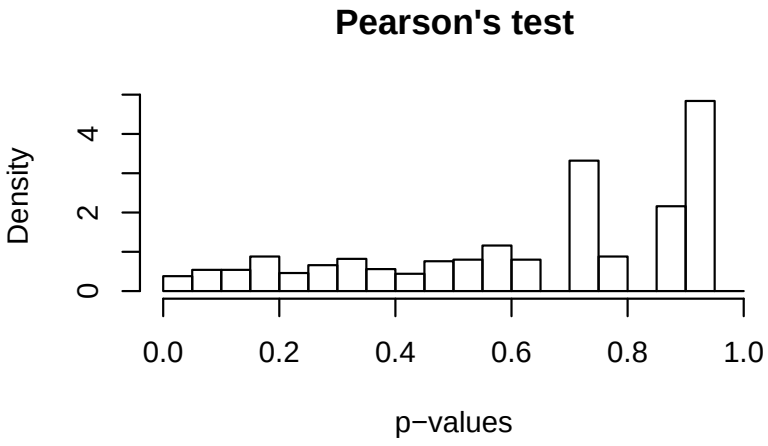


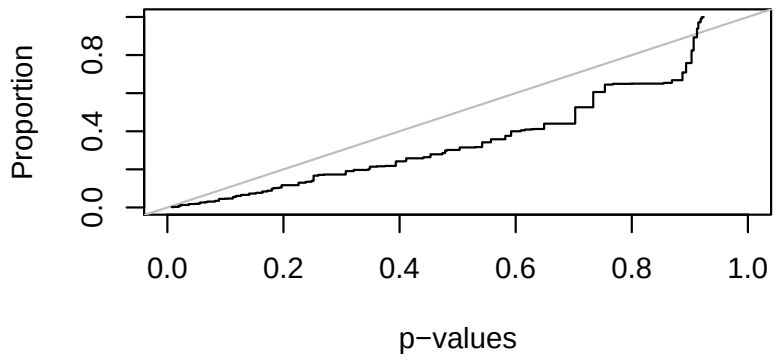


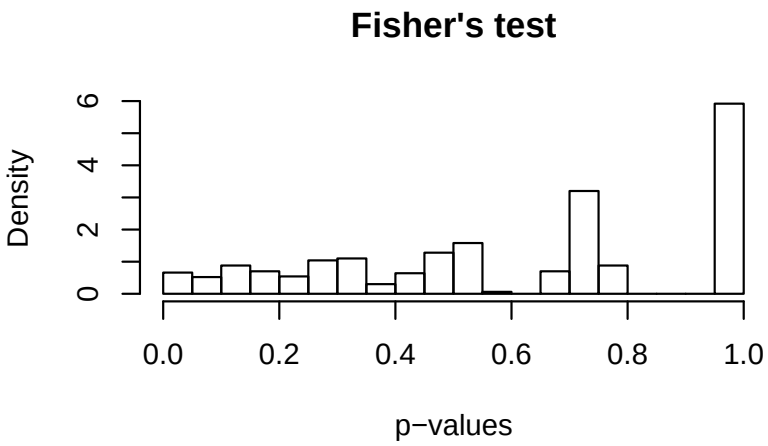


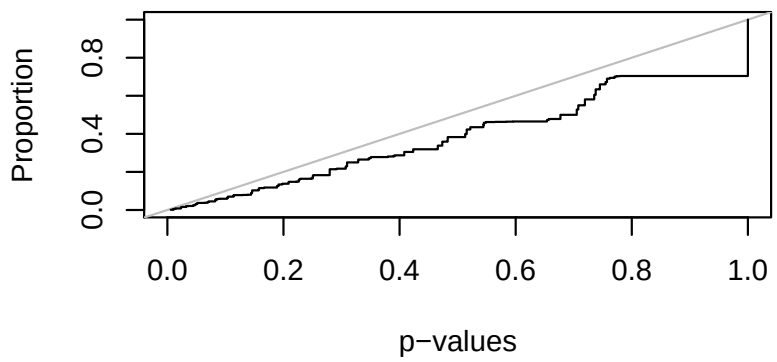


Null tables with independent rows, $P(\text{success}) = 52/63$









Other examples

- Explore robustness in other situations where the assumptions are violated. Look for the effect of violations on the power of the test.
- Study Welch's correction for unequal variances in a two-sample t-test. What happens when the variances are equal? What happens if we do not use it when we should?
- Show Monte Carlo *p*-values when the null distribution is only available by simulation. Explore bootstrap tests.
- Explore other asymptotic approximations by studying the distributions of nominal *p*-values.

Other examples

- Explore robustness in other situations where the assumptions are violated. Look for the effect of violations on the power of the test.
- Study Welch's correction for unequal variances in a two-sample t-test. What happens when the variances are equal? What happens if we do not use it when we should?
- Show Monte Carlo *p*-values when the null distribution is only available by simulation. Explore bootstrap tests.
- Explore other asymptotic approximations by studying the distributions of nominal *p*-values.

Other examples

- Explore robustness in other situations where the assumptions are violated. Look for the effect of violations on the power of the test.
- Study Welch's correction for unequal variances in a two-sample t-test. What happens when the variances are equal? What happens if we do not use it when we should?
- Show Monte Carlo *p*-values when the null distribution is only available by simulation. Explore bootstrap tests.
- Explore other asymptotic approximations by studying the distributions of nominal *p*-values.

Other examples

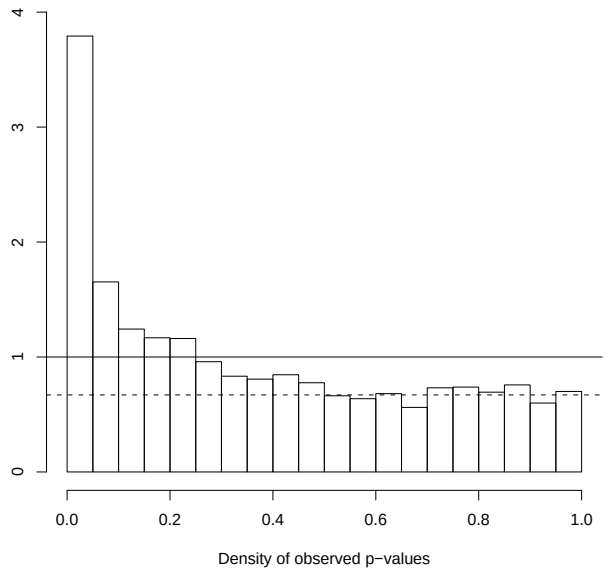
- Explore robustness in other situations where the assumptions are violated. Look for the effect of violations on the power of the test.
- Study Welch's correction for unequal variances in a two-sample t-test. What happens when the variances are equal? What happens if we do not use it when we should?
- Show Monte Carlo *p*-values when the null distribution is only available by simulation. Explore bootstrap tests.
- Explore other asymptotic approximations by studying the distributions of nominal *p*-values.

Still more examples

- In multiple testing, illustrate the distribution of the smallest of n p-values, and the distribution of Bonferroni-corrected p-values.
- Storey and Tibshirani (2003) used histograms of p-values in a collection of genomewide tests in order to illustrate false discovery rate calculations.

Still more examples

- In multiple testing, illustrate the distribution of the smallest of n p-values, and the distribution of Bonferroni-corrected p-values.
- Storey and Tibshirani (2003) used histograms of p-values in a collection of genomewide tests in order to illustrate false discovery rate calculations.



Conclusion

- Many students end up with fallacious interpretations of p-values, e.g. $P(H_0|\text{data})$.
- We should look at histograms (or ECDF plots) of p-values from simulations.
- P-values are random variables!