Hypothesis Testing

Ryan Miller

Outline

1. Hypotheses and null distributions
   - Null vs. alternative hypotheses, null distributions via simulation
2. Measuring evidence
   - p-values, guidelines
3. Common misconceptions
   - Large p-values, effect size, ignoring study design

Introduction

An experiment published in *Nature* explored whether infants have preference towards friendly behavior. 16 infants repeatedly watched demonstrations of two scenarios:

- A "helper" toy assisting the main character
- A "hinderer" toy blocking the main character

After watching these demonstrations, 14 of 16 infants chose the "helper" toy. The researchers were careful to randomize the color and shape of each character. Do the results of this study suggest that the infants can understand friendly behavior?

Questions

1) Did the design of this study do enough to prevent biases and confounding variables?

Questions

1) Did the design of this study do enough to prevent biases and confounding variables?
   - Probably, as factors like color or shape cannot explain the choices
2) What is a *falsifiable hypothesis* these researchers should evaluate?

Questions

1) Did the design of this study do enough to prevent biases and confounding variables?
   - Probably, as factors like color or shape cannot explain the choices
2) What is a *falsifiable hypothesis* these researchers should evaluate?
   - $H : p = 0.5$, or the infants chose randomly between the two options
3) What is a *model* that could simulate the outcomes we'd expect *if* $H : p = 0.5$ *is true?*

Questions

1) Did the design of this study do enough to prevent biases and confounding variables?
   - Probably, as factors like color or shape cannot explain the choices
2) What is a *falsifiable hypothesis* these researchers should evaluate?
   - $H : p = 0.5$, or the infants chose randomly between the two options
3) What is a *model* that could simulate the outcomes we'd expect *if* $H : p = 0.5$ *is true?*
   - Flipping a coin 16 times is akin to "replicating" this experiment under the hypothesis that $p = 0.5$

Hypothesis testing

Statistical tests involve two major components:

1) Proposing a **null hypothesis**, H_0, and an **alternative hypothesis**, H_a
   - The null hypothesis is falsifiable and the researchers want to disprove it
   - The alternative hypothesis represents the conclusion the researchers would like to establish

Hypothesis testing

Statistical tests involve two major components:

1) Proposing a **null hypothesis**, H_0, and an **alternative hypothesis**, H_a
   - The null hypothesis is falsifiable and the researchers want to disprove it
   - The alternative hypothesis represents the conclusion the researchers would like to establish
2) Deciding whether the sample data provide *sufficient evidence* to falsify the null hypothesis
   - A **null distribution** describes the outcomes that could have occurred *had the null hypothesis been true*
   - Evidence against H_0 comes from comparing the outcome observed from the *real data* versus the null distribution

Practice

1) Using the Randomization test for a single proportion menu of StatKey, input the results of the infant toy choice experiment and verify the null hypothesis $H_0 : p = 0.5$ is being used.
2) Click "generate 1000 samples" to create 1000 simulated study outcomes (under the assumption that the null hypothesis is true). How compatible is the observed outcome, $\hat{p} = 14/16 = 0.875$, with these simulated outcomes?

Practice (solution)

1) These steps are performed on StatKey
2) A proportion of 0.875 or larger happens in less than 1% of simulations, which suggests that such an outcome would be *very unlikely* if the null hypothesis were true. This small probability should be considered *strong evidence* against the null hypothesis.

P-values

A **p-value** is the probability of observing an outcome at least as unusual as the one observed in the real data under the assumption that the null hypothesis is true.

- A p-value is found by looking at either one tail (1-sided test) or both tails (2-sided test) of the null distribution
- A small p-value can be used to falsify the null hypothesis, but a large p-value should be considered "inconclusive"

P-values as a measure of evidence

Below are the original guidelines put forth by Ronald Fisher (creator of the p-value):

p-value	Evidence against the null
0.100	Borderline
0.050	Moderate
0.025	Substantial
0.010	Strong
0.001	Overwhelming

Fisher intended the p-value to be a quantitative measurement describing the strength of the evidence the sample data provide against a null hypothesis.

Decision thresholds

Many scientific fields have adopted $\alpha = 0.05$ as a threshold for "statistical significance"

- Data yielding a p-value smaller than $\alpha = 0.05$ are seen as sufficient evidence for rejecting H_0
- Data yielding a p-value larger than $\alpha = 0.05$ provide insufficient evidence and result in a "failure to reject H_0"

This black and white approach has it's flaws, but it's still very widely used.

Practice

We've previously discussed a study conducted by Johns Hopkins University that found 31 of 39 babies born 15 weeks early went on to survive. Do these data provide compelling evidence that a majority of babies born 15 weeks early survive?

1) Propose a null hypothesis and an alternative hypothesis
2) Use StatKey to create a null distribution and find the p-value measuring the evidence this study's observed outcome provides against the null hypothesis
3) Use a threshold of $\alpha = 0.05$ to make a decision regarding the null and alternative hypotheses

Practice (solution)

1) $H_0 : p = 0.5$, the null hypothesis is that 50% of babies born 15 weeks early survive vs. $H_a : p > 0.5$, the alternative that more than 50% survive
2) The 1-sided p-value should be zero (or very close to zero)
3) Because the p-value is so small, we reject H_0 in favor of H_a and conclude that these data provide *strong evidence* that more than 50% of babies born 15 weeks prematurely will survive

Practice - part 2

Wikipedia claims that 70% of babies born 15 weeks early will survive. Do the data in the Johns Hopkins University study (where 31 of 39 babies survived) provide compelling evidence against Wikipedia's claim?

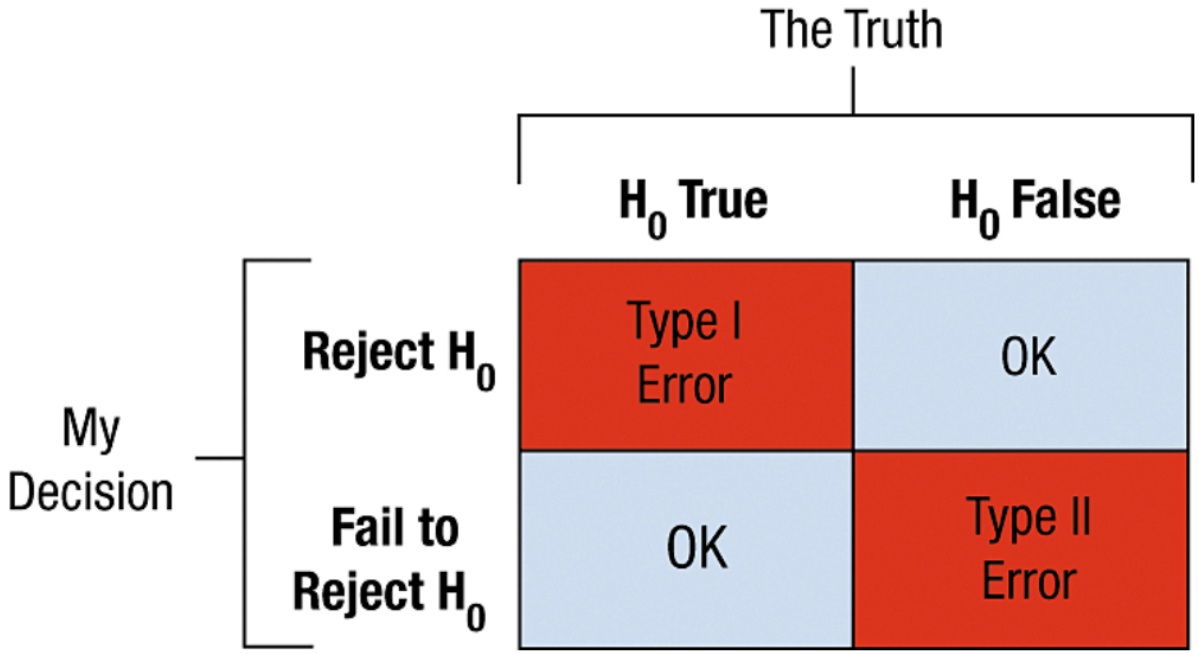
1) Propose a null hypothesis and an alternative hypothesis
2) Use StatKey to create a null distribution and find the *p*-value measuring the evidence this study's observed outcome provides against the null hypothesis
3) Use a threshold of $\alpha = 0.05$ to make a decision regarding the null and alternative hypotheses

Practice - part 2 (solution)

1) This time, $H_0 : p = 0.7$, and $H_a : p \neq 0.7$ (since either a larger than expected or smaller than expected result would disprove Wikipedia's claim)
2) This time, the 2-sided p-value is approximately 0.13
3) Because this p-value is larger than 0.05, there's *insufficient evidence* to reject H_0. It's unclear whether Wikipedia's claim is true, but these data are relatively compatible with it.

Decision errors

In reality, any conclusion drawn from a hypothesis test may or may not be correct:



- A **type I error** occurs when the null hypothesis is *rejected*, but in reality it is *true*
- A **type II error** occurs when the null hypothesis *cannot be rejected*, but in reality it is *false*

Trade-offs

- The likelihood of a statistical test resulting in a Type I error can be controlled by α, the threshold used for "statistical significance"
  - If $\alpha = 0.05$, we can expect a Type I error (false positive) in 5% of instances where H_0 is true

Trade-offs

- The likelihood of a statistical test resulting in a Type I error can be controlled by α, the threshold used for "statistical significance"
  - If $\alpha = 0.05$, we can expect a Type I error (false positive) in 5% of instances where H_0 is true
- Trivially, we could set $\alpha = 0$ and *never* make a Type I error, but how would this impact the likelihood we make a Type II error?

Trade-offs

- The likelihood of a statistical test resulting in a Type I error can be controlled by α, the threshold used for "statistical significance"
  - If $\alpha = 0.05$, we can expect a Type I error (false positive) in 5% of instances where H_0 is true
- Trivially, we could set $\alpha = 0$ and *never* make a Type I error, but how would this impact the likelihood we make a Type II error?
  - If we make it harder to reject H_0 (ie: we lower α), it becomes easier to make a Type II error

Practice

Jury trials in the US use the premise "innocent until proven guilty". Relating this to hypothesis testing, we can view a trial as a test of H_0 : Person A is not guilty vs. H_a : Person A is guilty

1) In words, what would a Type I and Type II error each represent in this scenario?
2) Which error would be worse? How might you choose α to be mindful of the trade-off between Type I and Type II errors?

Practice (solution)

1) A Type I error is convicting an innocent person. A Type II error is letting a guilty person go free.
2) A Type I error should be viewed as worse, so we might set a very strict decision threshold (ie: $\alpha = 0.01$ or even $\alpha = 0.001$). This is what courts actually do, as the standard of "beyond a reasonable doubt" is generally considered to be a very high bar.

Study replication and decision errors

- Part of the rationale for $\alpha = 0.05$ is that scientific research should always be replicated
  - Even if one study has a 5% chance of producing a false positive result (Type I error), the chances that three different studies each independently produce a Type I error is $0.05^3 = 0.000125$, or roughly 1 in 10,000
- Type II errors can be more insidious since findings that aren't statistically significant generally aren't published

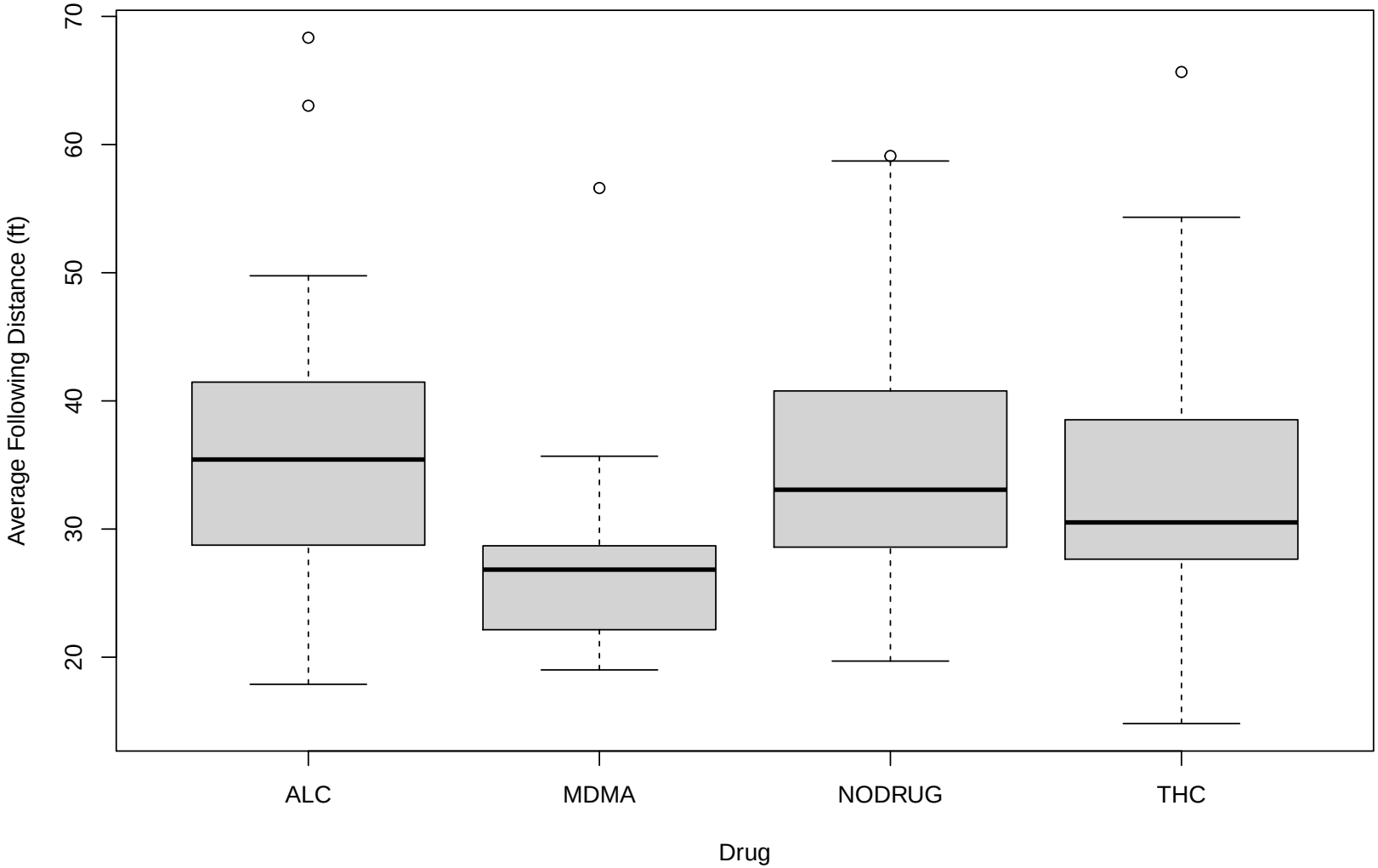
Family-wise error rates and false discovery rates

It is somewhat common for a single experiment to involve multiple hypotheses, an example is presented below:

- The NADS organization looked at the relationship between drug use and tailgating behavior while driving
- They classified participants into 4 groups according to the "hardest" substance they regularly used (No Drug, Alcohol, THC, or MDMA)
- These participants then drove a simulated route in an advanced driving simulator, and the researchers recorded their average following distance behind a lead vehicle as one of the study's outcomes

Family-wise error rates and false discovery rates

After removing a couple of outliers, here's what the data look like:



Family-wise error rates and false discovery rates

Since there are 4 different groups we'd like to compare, we might conduct 6 different hypothesis tests:

1. ALC vs NODRUG, p-value $= 0.5102$
2. ALC vs MDMA, p-value $= 0.00417$
3. ALC vs THC, p-value $= 0.8959$
4. THC vs NODRUG, p-value $= 0.4782$
5. THC vs MDMA, p-value $= 0.01383$
6. MDMA vs NODRUG, p-value $= 0.00216$

But if we compare each test's p-value against $\alpha = 0.05$, will the entire set of conclusions from this experiment (as a whole) still have a 5% Type I error rate?

The Bonferroni adjustment

The Type I error rate for this *family of tests* is inflated, suppose the null hypothesis is true for all 6 pairwise tests in the tailgating study (and the tests are independent); Then, using $\alpha = 0.05$:

$$\begin{aligned} Pr(\text{At least one type I error}) &= 1 - Pr(\text{No type I errors}) \\ &= 1 - (1 - 0.05)^6 = 26.5\% \end{aligned}$$

The Bonferroni adjustment

The Type I error rate for this *family of tests* is inflated, suppose the null hypothesis is true for all 6 pairwise tests in the tailgating study (and the tests are independent); Then, using $\alpha = 0.05$:

$$\begin{aligned} Pr(\text{At least one type I error}) &= 1 - Pr(\text{No type I errors}) \\ &= 1 - (1 - 0.05)^6 = 26.5\% \end{aligned}$$

This suggests a simple correction to significance threshold: $\alpha^* = \alpha/h$, where h is the number of hypothesis tests being performed. Then:

$$\begin{aligned} Pr(\text{At least one type I error}) &= 1 - Pr(\text{No type I errors}) \\ &= 1 - (1 - 0.05/6)^6 \approx 5\% \end{aligned}$$

The Bonferroni Adjustment

Setting $\alpha^* = \alpha/h$ is known as the **Bonferroni Adjustment**. If we apply this correction, how many of the 6 hypotheses can be rejected with a family-wise Type I error rate of 5%?

1. ALC vs NODRUG, p-value $= 0.5102$
2. ALC vs MDMA, p-value $= 0.00417$
3. ALC vs THC, p-value $= 0.8959$
4. THC vs NODRUG, p-value $= 0.4782$
5. THC vs MDMA, p-value $= 0.01383$
6. MDMA vs NODRUG, p-value $= 0.00216$

The Bonferroni Adjustment

Setting $\alpha^* = \alpha/h$ is known as the **Bonferroni Adjustment**. If we apply this correction, how many of the 6 hypotheses can be rejected with a family-wise Type I error rate of 5%?

1. ALC vs NODRUG, p-value $= 0.5102$
2. ALC vs MDMA, p-value $= 0.00417$
3. ALC vs THC, p-value $= 0.8959$
4. THC vs NODRUG, p-value $= 0.4782$
5. THC vs MDMA, p-value $= 0.01383$
6. MDMA vs NODRUG, p-value $= 0.00216$

Using $\alpha^* = 0.05/6 = 0.0083$ only 2 of 6 tests are now considered "statistically significant", but we've controlled the *family-wise* Type I error rate at 5%.

Bonferroni Adjusted p-values

- Occasionally you'll see a study report **adjusted p-values**
- For the Bonferroni adjustment, these are found by multiplying the original p-values by h (the number of tests)
- "Bonferroni Adjusted p-values" can then be compared directly to the target family-wise Type I error rate
  - For example, comparing the adjusted p-values against 0.05 will achieve a 5% family-wise Type I error rate

Practice

A genetic association study tested for differences in gene expression between two types of leukemia. The study tested 7129 genes.

1) If all 7129 tests were done using $\alpha = 0.01$, and there are no genetic differences between these two types of leukemia, how many "statistically significant" genes would be expected?
2) Suppose 783 genes had *p*-values less than 0.01, do you believe there is association between some genes and type of leukemia
3) Suppose you wanted to use the Bonferroni adjustment to ensure a Type I error rate no larger than 5%. What would your adjusted significance threshold be?
4) Suppose the "most significant" gene had a *p*-value of 0.000001, what is its *Bonferroni Adjusted p-value*?

Practice - Solution

1) You'd expect $7129 * 0.01 = 71$ Type I errors
2) Yes, there were over 10 times (712) more significant results than expected
3) $\alpha^* = 0.05/7129 = 0.000007$
4) The adjusted p-value is $0.000001 * 7129$, or $p^* = 0.007$

Common mistake #1 - interpretting high p-values

As a silly example, suppose Prof. Miller and Steph Curry compete in a 3-point shooting contest. Further, suppose that Prof. Miller makes 3 of 5 and Steph Curry makes 5 of 5.

- We might use these data to test the hypothesis that Steph Curry and Prof. Miller are equally good 3-pt shooters: $H_0 : p_1 - p_2 = 0$
- The result is a p-value of 0.17, but does that mean that Prof. Miller and Steph Curry are equally good?

Common mistake #1 - interpretting high p-values

- A high p-value indicates the data provide insufficient evidence against the null hypothesis (not that the null hypothesis is likely true!)
  - Sample size was an important factor in the Steph Curry example, as 5 shot attempts isn't enough data to make a statistically justified decision

Common mistake #2 - statistical vs. practical significance

- In the 1980s, AstraZeneca developed *Prilosec*, a highly successful heartburn medication
  - The FDA patent for Prilosec expired in 2001, prompting AstraZeneca to try to replace Prilosec with a new drug, *Nexium*
- In a clinical trial comparing the two drugs, Prilosec had a healing rate of 87.5%, while Nexium had a healing rate of 90%

Common mistake #2 - statistical vs. practical significance

- In the 1980s, AstraZeneca developed *Prilosec*, a highly successful heartburn medication
  - The FDA patent for Prilosec expired in 2001, prompting AstraZeneca to try to replace Prilosec with a new drug, *Nexium*
- In a clinical trial comparing the two drugs, Prilosec had a healing rate of 87.5%, while Nexium had a healing rate of 90%
  - The sample size of the trial was very large (over 6000 participants) and the difference between the drugs was "highly significant" with a p-value ≤ 0.01

Common mistake #2 - statistical vs. practical significance

- *p*-values depend upon both *sample size* and **effect size**
  - It's possible for an effect size to be small enough to make *no practical difference*, but for the *p*-value to be very small (due to a large sample size)
- Avoid interpreting a very small *p*-value as being indicative of a very important scientific finding

Common mistake #3 - ignoring study design

- In the 1970s, UC-Berkeley was investigated for possible sex-discrimination in admissions to its graduate programs
  - For the fall semester of 1973, 3715 of 8442 male applicants were accepted, but only 1512 of 4321 female applicants were accepted (a difference in proportions of 0.09)
  - Using StatKey, how does the observed difference in proportions compare to the null distribution corresponding to $H_0 : p_m - p_f = 0$?

Common mistake #3 - ignoring study design

The *p*-value in the UC-Berkeley example is less than 0.0001, but here's what the data look like when *stratified* by department:

Department	All		Men		Women	
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
A	933	64%	***825***	62%	108	**82%**
B	585	63%	***560***	63%	25	**68%**
C	918	35%	325	**37%**	***593***	34%
D	792	34%	417	33%	375	**35%**
E	584	25%	191	**28%**	***393***	24%
F	714	6%	373	6%	341	**7%**

- Clearly the overall difference in male - female acceptance can be explained by the *confounding variable* of department
  - Males tended to disproportionately apply to programs with higher acceptance rates

Proper reporting

Proper scientific reporting of a statistical test should include the following:

1) An effect size, such as a point estimate and/or a confidence interval
2) The p-value itself, not just whether it's above or below 0.05
3) A practical conclusion, not just whether to reject H_0 or not reject H_0

Practice

Below are 4 different sentences that report the results of the same study. Rank them from best to worst.

1) The studied provided compelling evidence to reject the hypothesis that Nexium and Prilosec are equally good.
2) The study found that Nexium offered a statistically significant improvement over Prilosec, with a p-value less than 0.01
3) The study found that Nexium had significantly higher healing rate than Prilosec (90% vs. 87.5%, p $= 0.003$)
4) According to the study, Nexium was found to be significantly better at treating heartburn than Prilosec

Practice (solution)

- The best is #3, it provides an effect size, the exact p-value, and reasonable summary
- Next is #2, which provides some indication of the p-value and a reasonable summary
- Next is #4, which at least gives reasonable summary
- The worst is #1, it doesn't provide any meaningful insight regarding which treatment should be preferred

Summary

Steps of a hypothesis test:

1) State the null and alternative hypothesis
2) Find the null distribution
3) Compare the observed outcome against the null distribution to find the p-value
4) Use the p-value (and the effect size) to make a conclusion

Summary (continued)

Scenarios we've encountered and the proper hypotheses:

- A single proportion - $H_0 : p = \_$ (sometimes 0.5, sometimes depends on the context)
- A single mean - $H_0 : \mu = \_$ (depends on the context)
- A difference in proportions - $H_0 : p1 - p2 = \_$ (usually zero)
- A difference in means - $H_0 : \mu_1 - \mu_2 = \_$ (usually zero)

Summary (continued)

Mistakes to avoid:

1) Performing multiple tests in a single experiment without acknowledging the Type I error rate
2) Believing that a large p-value means the null hypothesis is likely to be true
3) Ignoring effect size when interpreting a p-value (p-values do not measure scientific or practical importance)
4) Ignoring study design when interpreting a p-value (confounding variables and biases are not considered in the p-value calculation)