Hypothesis Testing (concepts)

Ryan Miller

Outline

1. Hypothesis testing framework
   - Setting up hypotheses, null distributions, p-values
2. Decision errors
   - Statistical significance, types of errors
3. Common misconceptions

Introduction

Confidence intervals:

- Start with the sample data
  - Use probability to reliably estimate what you believe might be true of the population

Hypothesis testing:

- Start with a falsifiable hypothesis about the population
  - Use probability to determine if the sample data provide enough evidence to falsify the hypothesis

Falsifiable hypotheses

We can only make progress with hypotheses that are *testable* and are **falsifiable**, meaning you could observe evidence that *disproves it*.

1) H : There once was life on Mars
2) H : There's never been life on Mars

Which of the above hypotheses is *falsifiable*?

Statistical hypotheses

Below are some **statistical hypotheses** related to *population parameters*:

1) $H : \mu_1 - \mu_2 \neq 0$ is *not falsifiable* - even sample data where $\bar{x}_1$ is exactly equal to $\bar{x}_2$ doesn't eliminate the possibility of the population-level means being different (due to sampling variability)

Statistical hypotheses

Below are some **statistical hypotheses** related to *population parameters*:

1) $H : \mu_1 - \mu_2 \neq 0$ is *not falsifiable* - even sample data where $\bar{x}_1$ is exactly equal to $\bar{x}_2$ doesn't eliminate the possibility of the population-level means being different (due to sampling variability)
2) $H : \mu_1 - \mu_2 = 0$ is *falsifiable* - we could disprove it if the means in our *sample data* are so different that it would *extraordinary unlikely* for them to be equal *within the broader population*

Notice how a falsifiable statistical hypothesis must suggest a specific value (ie: zero)

Example - 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?

Example - hypothesis testing steps

1) What is a *falsifiable statistical hypothesis* that these researchers might be interested in disproving?

Example - hypothesis testing steps

1) What is a *falsifiable statistical hypothesis* that these researchers might be interested in disproving?
2) Assuming $H : p = 0.5$ is true, can you come up with a method for simulating the results of this experiment?

Example - hypothesis testing steps

1) What is a *falsifiable statistical hypothesis* that these researchers might be interested in disproving?
2) Assuming $H : p = 0.5$ is true, can you come up with a method for simulating the results of this experiment?
3) How does does the outcome that was observed in the real experiment compare with the distribution of simulated outcomes (which assumed $p = 0.5$)?

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 hope to disprove it
   - The alternative hypothesis is 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 hope to disprove it
   - The alternative hypothesis is 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 outcomes that could have been observed *had the null hypothesis been true*
   - The outcome observed in the *real data* should be compared against the null distribution

Null distributions

- In our first example, we built the null distribution via simulation, but a more consistent approach would use probability theory
- For $H_0 : p = 0.5$, and a sample size of $n = 16$, CLT suggests the following model for the null distribution:

$$\hat{p} \sim N\left(0.5, \sqrt{\frac{0.5(1-0.5)}{16}}\right)$$

We can then measure the likelihood of $\hat{p} = 14/16$, assuming H_0 is true, using `pnorm()`:

```
pnorm(14/16, mean = 0.5, sd = sqrt(0.5*0.5/16), lower.tail = FALSE)
```

```
## [1] 0.001349898
```

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.

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. According to Wikipedia, the survival rate for babies born this early is 70%. Does the the Johns Hopkins University study provide compelling evidence to refute Wikipedia's claim?

1) Propose a null hypothesis and an alternative hypothesis
2) Use Central Limit theorem to come up with a null distribution
3) Compare $\hat{p}$ against the null distribution to find the p-value and make a conclusion regarding the null and alternative hypotheses

Practice (solution)

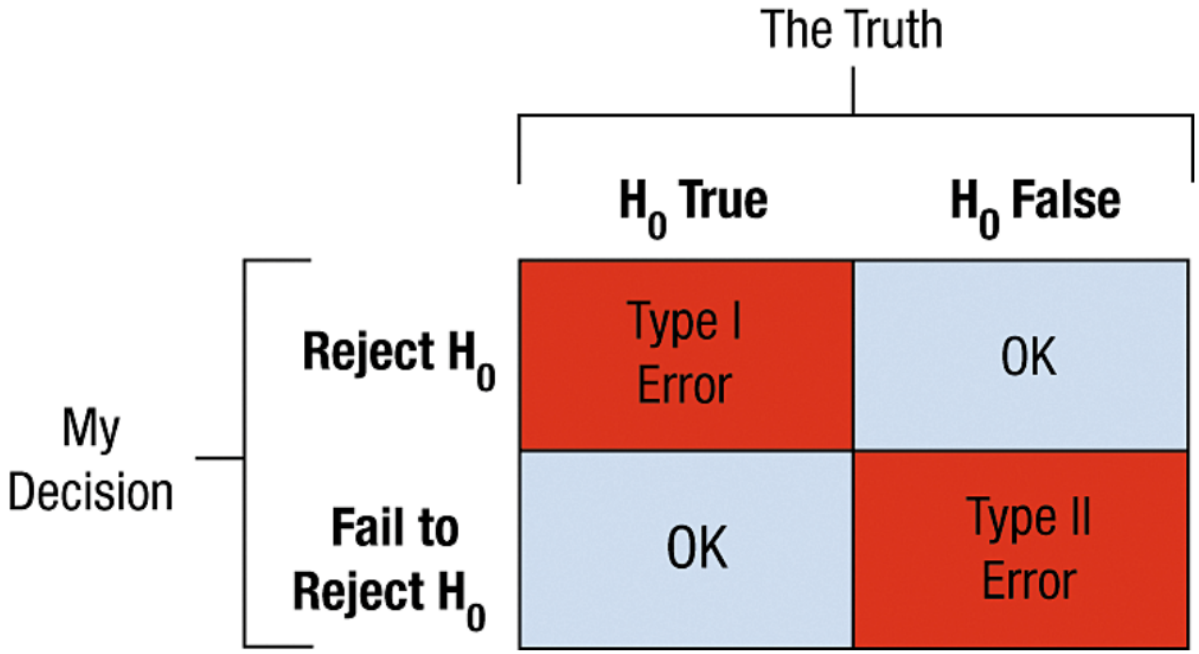
1) $H_0 : p = 0.7$ vs. $H_a : p \neq 0.7$
2) Under the assumption that $p = 0.7$, CLT suggests
   $\hat{p} \sim N\left(0.7, \sqrt{\frac{0.7(1-0.7)}{39}}\right)$
3) The two-sided p-value corresponding to $\hat{p} = 31/39$ is 0.196, which indicates these data provide insufficient evidence to disprove Wikipedia's claim. (Note: Wikipedia's claim may or may not be true, but these sample data are relatively compatible with it)

Decision thresholds

- Many scientific fields set a threshold that defines whether a finding is "statistically significant"
  - Under this paradigm, any *p*-value less than the decision threshold, α, is considered sufficient evidence to reject H_0
- A threshold of $\alpha = 0.05$ is the most widely used
  - So, if *p*-value $> \alpha = 0.05$, the finding is "statistically significant" (implying the null hypothesis was rejected)

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*

Decision errors

- As a data analyst, you *cannot* control whether H_0 is true or what the data look like, but you *can* control the value of α, the threshold used for rejecting H_0
  - If $\alpha = 0.05$, we can expect a Type I error (false positive) in 5% of instances where H_0 is true

Decision errors

- As a data analyst, you *cannot* control whether H_0 is true or what the data look like, but you *can* control the value of α, the threshold used for rejecting H_0
  - If $\alpha = 0.05$, we can expect a Type I error (false positive) in 5% of instances where H_0 is true
- How could you limit the Type I error rate to zero? What consequences would this have on the likelihood of making a Type II error?

Decision errors

- As a data analyst, you *cannot* control whether H_0 is true or what the data look like, but you *can* control the value of α, the threshold used for rejecting H_0
  - If $\alpha = 0.05$, we can expect a Type I error (false positive) in 5% of instances where H_0 is true
- How could you limit the Type I error rate to zero? What consequences would this have on the likelihood of making a Type II error?
  - The harder it is to reject H_0 (ie: the lower α is), the easier it is 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

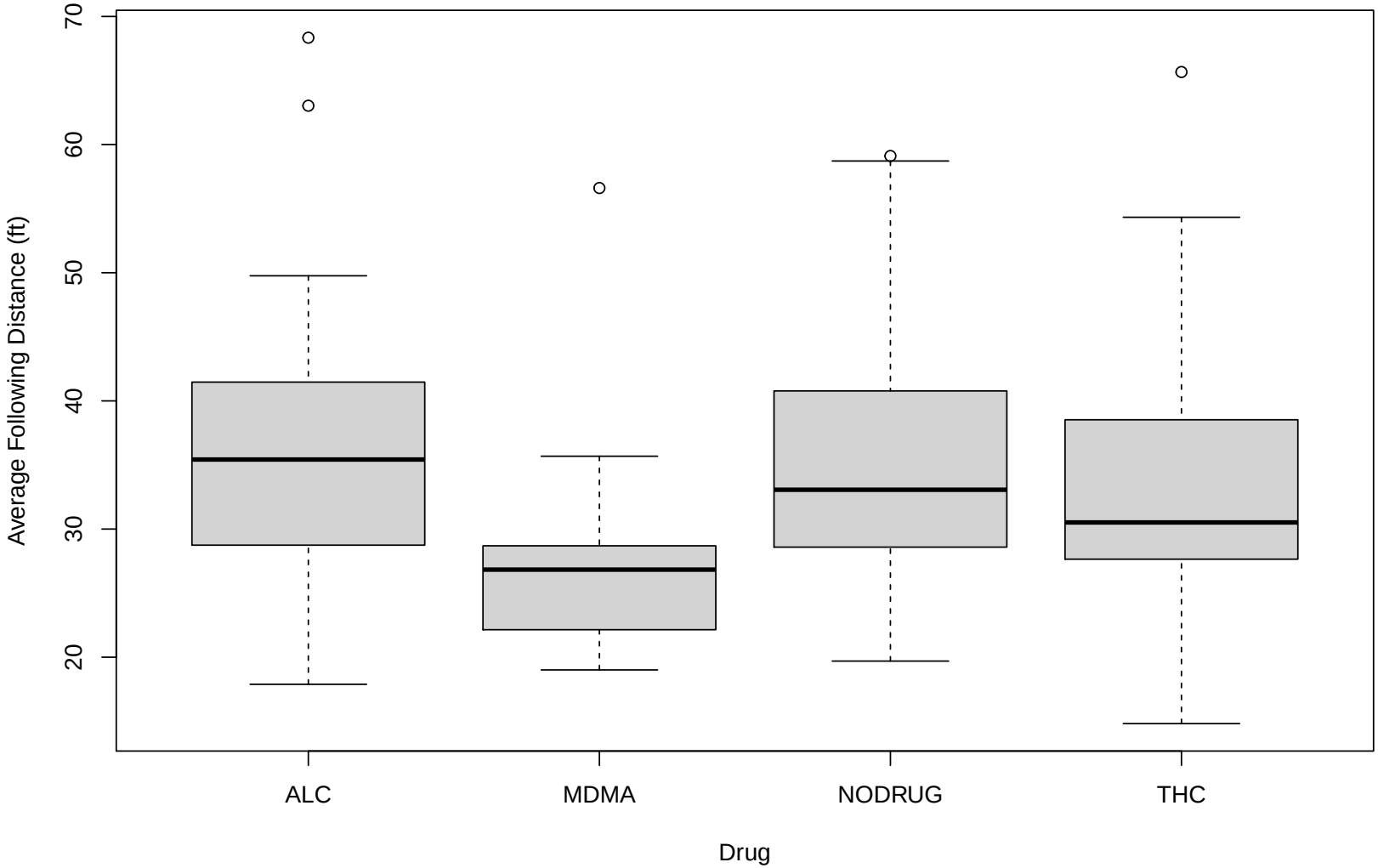
Family-wise error rates and false discovery rates

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%.

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, can we reject H_0 for this gene while controlling the experiment's family-wise Type I error rate at 5%?

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) Yes, a *p*-value of 0.000001 is less than $\alpha^* = 0.000007$, so it indicates sufficient evidence to reject H_0 while still facilitating the desired Type I error rate

Hypothesis testing vs. confidence interval estimation

- Hypothesis tests and confidence intervals are both used to evaluate the role random chance (sampling variability)
- Consider $H_0 : \mu_1 - \mu_2 = 0$. If a sample produces a 95% confidence interval estimate of (3.2, 10.1), do you think the p-value is more likely to be large or small?

Hypothesis testing vs. confidence interval estimation

- Hypothesis tests and confidence intervals are both used to evaluate the role random chance (sampling variability)
- Consider $H_0 : \mu_1 - \mu_2 = 0$. If a sample produces a 95% confidence interval estimate of (3.2, 10.1), do you think the *p*-value is more likely to be large or small?
  - Since 0 isn't in the 95% CI, the value specified in the null hypothesis is *implausible*
  - In fact, the two-sided *p*-value must be < 0.05 (the mathematical complement of the confidence level)

Practice

Suppose the two-sided p-value for a test of the hypothesis $H_0 : \mu_1 - \mu_2 = 0$ is 0.13

1) Do you think 0 is contained in the 95% confidence interval estimate of $\mu_1 - \mu_2$?
2) What about the 80% confidence interval estimate?

Practice (solution)

1) Yes, because the two-sided p-value is larger than 0.05 (the mathematical complement of a 95% confidence level) the value specified in the null hypothesis should be considered plausible by both the CI and the hypothesis testing conclusion
2) Since the two-sided p-value is smaller than 0.20, zero will not be contained in the 80% confidence interval

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 $\alpha = 0.05$
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)

- 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
- Worst is #1, which doesn't provide any meaningful insight into what the study suggests

Summary

Steps of a hypothesis test:

1) State the null and alternative hypothesis
2) Determine 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

Additionally, you should always be mindful of the possibility of Type I and Type II errors, and avoid the common mistakes described in this presentation whenever performing a hypothesis test.