Hypothesis Testing (part 3, misconceptions)

Ryan Miller

Introduction

- The previous presentations have introduced the general framework for hypothesis testing, as well as the p-value as a measure of evidence against the null hypothesis
- Unfortunately, p-values are very commonly misunderstood and are frequently used incorrectly
  - The misuse of p-values has become such a problem that *Basic and Applied Social Psychology* has banned their use (source)

Introduction

- The previous presentations have introduced the general framework for hypothesis testing, as well as the p-value as a measure of evidence against the null hypothesis
- Unfortunately, p-values are very commonly misunderstood and are frequently used incorrectly
  - The misuse of p-values has become such a problem that *Basic and Applied Social Psychology* has banned their use (source)
- It is my belief that p-values, if used properly, are a meaningful and important statistical tool
  - This presentation will cover common mistakes in interpreting p-values

Mistake #1 - "Proving" the Null Model

- Let's consider a silly example where the NBA's Steph Curry and Professor Miller compete by each shooting 5 three-point shots
  - I make 2 of 5, and Steph makes 5 of 5

Mistake #1 - "Proving" the Null Model

- Let's consider a silly example where the NBA's Steph Curry and Professor Miller compete by each shooting 5 three-point shots
  - I make 2 of 5, and Steph makes 5 of 5
- We might use a hypothesis test to evaluate the null hypothesis that we're both equally good three-point shooters (ie: $H_0 : p_{\text{Miller}} = p_{\text{Curry}}$)
  - The p-value for this scenario is 0.17
  - Does that mean we are equally good 3-pt shooters?

Mistake #1 - "Proving" the Null Model

- The answer is a resounding "no", Steph Curry and I are not equally good three-point shooters!

Mistake #1 - "Proving" the Null Model

- The answer is a resounding "no", Steph Curry and I are not equally good three-point shooters!
- The p-value measures the strength of evidence against the null hypothesis
  - In a sample involving only 5 shots, there isn't enough data to provide sufficient evidence against the null hypothesis
  - A lack of evidence does not mean that the null hypothesis is likely true

Mistake #1 - A Non-hypothetical Example

- It might seem professionals would easily avoid the mistake highlighted in that silly Steph Curry example, but unfortunately it happens quite often

Mistake #1 - A Non-hypothetical Example

- It might seem professionals would easily avoid the mistake highlighted in that silly Steph Curry example, but unfortunately it happens quite often
- In 2006, the Woman's Health Initiative evaluated the relationship between low-fat diets and reduced risk of breast cancer risk and found a p-value of 0.07

Mistake #1 - A Non-hypothetical Example

- It might seem professionals would easily avoid the mistake highlighted in that silly Steph Curry example, but unfortunately it happens quite often
- In 2006, the Woman's Health Initiative evaluated the relationship between low-fat diets and reduced risk of breast cancer risk and found a p-value of 0.07
  - The NY Times ran the headline: "Study Finds Lowfat Diets Won't Stop Cancer or Heart Disease"
  - The article described the study's results as: "The death knell for the belief that reducing the percentage of fat in the diet is important for health"
- In reality, these results simply indicates insufficient evidence linking dietary fat and breast cancer, it's very possible there is a small benefit but we cannot rule out random chance

Comments - "Proving" the Null Hypothesis

- Hypothesis testing is not designed to "prove" a null hypothesis, so you should never use it to try and do so
  - The null hypothesis is intended to be a "straw man" that researchers want to "knock down"

Comments - "Proving" the Null Hypothesis

- Hypothesis testing is not designed to "prove" a null hypothesis, so you should never use it to try and do so
  - The null hypothesis is intended to be a "straw man" that researchers want to "knock down"
- The closest thing to "proving" a null hypothesis is finding a very narrow confidence interval around the null value
  - This interval estimate would suggest the only plausible values for the parameter of interest are extremely close to those the null hypothesis suggests

Comments - Confidence Intervals vs. Hypothesis Tests

- Confidence intervals and hypothesis tests are two complementary tools for evaluating the variability in sample data
  - A confidence interval provides a range of plausible estimates for a population characteristic
  - A hypothesis test considers a null model for the population characteristic and measures how compatible the sample data are with this model

Comments - Confidence Intervals vs. Hypothesis Tests

- Confidence intervals and hypothesis tests are two complementary tools for evaluating the variability in sample data
  - A confidence interval provides a range of plausible estimates for a population characteristic
  - A hypothesis test considers a null model for the population characteristic and measures how compatible the sample data are with this model
- Consider $H_0 : p = 0.5$, and suppose our sample produces a 95% CI estimate for p of (0.53, 0.63)
  - This interval says that it is *not plausible* that $p = 0.5$, so we expect the hypothesis test to have a $p - value < 0.05$ (based upon the 95% confidence level)

Comments - Confidence Intervals vs. Hypothesis Tests

- Confidence intervals and hypothesis tests are two complementary tools for evaluating the variability in sample data
  - A confidence interval provides a range of plausible estimates for a population characteristic
  - A hypothesis test considers a null model for the population characteristic and measures how compatible the sample data are with this model
- Consider $H_0 : p = 0.5$, and suppose our sample produces a 95% CI estimate for p of (0.53, 0.63)
  - This interval says that it is *not plausible* that $p = 0.5$, so we expect the hypothesis test to have a $p - value < 0.05$ (based upon the 95% confidence level)
- Again consider $H_0 : p = 0.5$, but now suppose a different sample leads to a sample proportion of $\hat{p} = 0.53$ and a p-value of 0.11, we'd expect the 95% confidence interval estimate from this sample to suggest that 0.5 *is* a plausible value (ie: the 95% CI would contain 0.5)

Mistake #2 - Clinical vs. Statistical Significance

- Confidence intervals and hypothesis tests lead to similar conclusions, but provide complementary information
- In the 1980s, *AstraZeneca* developed *Prilosec*, a very successful medication for healing erosive esophagitis (heart burn)
  - In the 2001, just before the company's patent on *Prilosec* was about to expire, *AstraZeneca* developed a new drug, *Nexium*

Mistake #2 - Clinical vs. Statistical Significance

- Confidence intervals and hypothesis tests lead to similar conclusions, but provide complementary information
- In the 1980s, *AstraZeneca* developed *Prilosec*, a very successful medication for healing erosive esophagitis (heart burn)
  - In the 2001, just before the company's patent on *Prilosec* was about to expire, *AstraZeneca* developed a new drug, *Nexium*
- To get *Nexium* approved by the FDA, *AstraZeneca* conducted a large randomized experiment comparing it to *Prilosec*
  - The experiment resulted in a p-value < 0.001, well below significance threshold of $\alpha = 0.05$ used by the FDA
- After its approval, *AstraZeneva* spent millions of dollars marketing *Nexium* and it soon became one of the top selling drugs in the world, leading to billions in profits

Mistake #2 - Clinical vs. Statistical Significance

- While p-value < 0.001, the observed healing rates were 87% for Prilosec and 90% for Nexium
  - The factor by which Nexium improved healing had a 95% CI of (1.02, 1.06)

Mistake #2 - Clinical vs. Statistical Significance

- While p-value < 0.001, the observed healing rates were 87% for Prilosec and 90% for Nexium
  - The factor by which Nexium improved healing had a 95% CI of (1.02, 1.06)
- Further, the active ingredients of these drugs are:
  - Omeprazole (Prilosec)
  - Esomeprazole (Nexium)
- Without getting too far into the chemistry (not my area of expertise), Omeprazole is a 50-50 mix of active and inactive isomers, while Esomeprazole only contains active "S" isomers

Mistake #2 - Clinical vs. Statistical Significance

- While p-value < 0.001, the observed healing rates were 87% for Prilosec and 90% for Nexium
  - The factor by which Nexium improved healing had a 95% CI of (1.02, 1.06)
- Further, the active ingredients of these drugs are:
  - Omeprazole (Prilosec)
  - Esomeprazole (Nexium)
- Without getting too far into the chemistry (not my area of expertise), Omeprazole is a 50-50 mix of active and inactive isomers, while Esomeprazole only contains active "S" isomers
- Critics of the pharmaceutical industry argue the results of the Nexium study were not **clinically significant**, meaning the differences in the two drugs aren't substantial enough to be influencing clinical practices

Mistake #2 - Clinical vs. Statistical Significance

- A very small p-value does not mean an observed relationship is large, meaningful, or important
  - The p-value is a tool for evaluating how plausible it is for an observed relationship to be explained by random chance

Mistake #2 - Clinical vs. Statistical Significance

- A very small p-value does not mean an observed relationship is large, meaningful, or important
  - The p-value is a tool for evaluating how plausible it is for an observed relationship to be explained by random chance
- With enough data, it is possible to show small/inconsequential relationships are unlikely to occur by chance alone
  - This doesn't mean those relationships have any real-world significance
  - Reporting confidence intervals along side hypothesis test results is one way to address this shortcoming

Summary

- A large or non-significant p-value does not mean that the null hypothesis is likely true
  - Instead, a large p-value only means there is insufficient evidence in the sample

Summary

- A large or non-significant p-value does not mean that the null hypothesis is likely true
  - Instead, a large p-value only means there is insufficient evidence in the sample
- A small or significant p-value does not mean the observed relationship is important or meaningful
  - Instead, a small p-value only means the sample data are unlikely to have occurred by random chance alone if the null model were true