Statistical Inference and the Scientific Method

Ryan Miller

Outline

1. The scientific method
   - framework and investigative steps
2. Falsifying a hypothesis
   - a conceptual framework for statistical testing

The scientific method

1. Propose a hypothesis
2. Collect data to evaluate the hypothesis
3. Assess the strength of evidence the data provide and reach a conclusion
4. Repeat steps #2 and #3 until a consensus is reached

Falsifiable hypotheses

In step #1 we focus on hypotheses that are *testable* and are **falsifiable**, meaning you could observe evidence that *disproves it*.

1) H : There once was life on Mars is *not falsifiable* - we could never disprove it, even if we extensively explored every aspect of Mars with the best available scientific tools

Falsifiable hypotheses

In step #1 we focus on hypotheses that are *testable* and are **falsifiable**, meaning you could observe evidence that *disproves it*.

1) H : There once was life on Mars is *not falsifiable* - we could never disprove it, even if we extensively explored every aspect of Mars with the best available scientific tools
2) H : There's never been life on Mars is *falsifiable* - we could disprove it by finding sufficiently convincing evidence of life

Notice that by disproving the second hypothesis we've effectively established the first hypothesis as true!

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*

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

The scientific method and statistical inference

In order to use statistical methods to establish a scientific relationship we must do the following:

1) Evaluate the possibility of *bias* and *confounding variables* (study design)
2) Propose a falsifiable hypothesis stating that the desired relationship *doesn't exist* and then use statistical methods (ie: confidence intervals) to establish sufficient evidence against that hypothesis.
3) Have others independently replicate our conclusions.

We'll spend a lot of time on #2 throughout the remainder of the semester, but #1 and #3 are just as important to keep in mind.

Practice

In 2012, researchers took 166 adults suffering from acute sinusitis and randomly split them into two groups: the first received a 10-day course of antibiotics and the second received 10-days of placebo pills that were similar in taste and appearance. Participants were then evaluated by staff who did not know their assigned treatment see if their condition had improved (yes or no)

	Yes	No	Total
Antibiotics	66	19	85
Placebo	65	16	81
Total	131	35	166

1) How concerned, if at all, should we be about the possibility of bias or confounding variables in this study?
2) What is the falsifiable hypothesis that these researchers should evaluate if they hope to establish antibiotics as an effective treatment?

Practice (continued)

In 2012, researchers took 166 adults suffering from acute sinusitis and randomly split them into two groups: the first received a 10-day course of antibiotics and the second received 10-days of placebo pills that were similar in taste and appearance. Participants were then evaluated by staff who did not know their assigned treatment see if their condition had improved (yes or no)

	Yes	No	Total
Antibiotics	66	19	85
Placebo	65	16	81
Total	131	35	166

3) Use statistical methods to construct a 95% CI estimate for the difference in the proportion who experienced an improvement in each group. Does this interval support the conclusion that antibiotics are more effective than placebo?
4) If we'd used a lower confidence level, is it possible that our interval might support a different conclusion?

Practice (solution)

1) This experiment was randomized, so confounding variables should not be a concern. They researchers also used a placebo and double-blinding, which should prevent many sources of possible bias.
2) We should propose the hypothesis that antibiotics are no different from placebo in terms of improvement in sinusitis. This is akin to saying $p_1 - p_2 = 0$ (where p_1 and p_2 are the population proportions receiving either treatment that improve).
3) We've covered two ways to construct this interval (bootstrapping and CLT formulas), using bootstrapping I got (-0.148, 0.096). This interval does not provide sufficient evidence to refute the hypothesis in 2)
4) Yes, for example the 30% CI is (-0.050, -0.001), which suggests that antibiotics are actually worse than placebo (but only with 30% confidence)

Next Steps

- Confidence intervals can be used to evaluate hypotheses, but they aren't the best statistical tool for doing so
  - It's easy for a confidence interval to under-sell the amount of evidence against a falsifiable hypothesis
  - For example, if the 95% CI in our last example had been (0.2, 0.3) in favor of antibiotics group we could actually be more than 95% confident of a difference (since the interval is so narrow and so far away from zero)

Next Steps

- Confidence intervals can be used to evaluate hypotheses, but they aren't the best statistical tool for doing so
  - It's easy for a confidence interval to under-sell the amount of evidence against a falsifiable hypothesis
  - For example, if the 95% CI in our last example had been (0.2, 0.3) in favor of antibiotics group we could actually be more than 95% confident of a difference (since the interval is so narrow and so far away from zero)
- We'll spend the remainder of the semester covering *hypothesis testing*, a broad area of statistics aimed at more precisely quantifying the degree of compatibility the sample data has with a null hypothesis