INFO 1301

Prof. Michael Paul
Prof. William Aspray

Hypothesis Testing

21 October 2016

Research (Alternative) Hypotheses

In many cases, research takes the form of answering a question or testing a prediction, which is generally stated in the form of a hypothesis that can be tested. Two examples:

- Q: Does a training program in driver safety result in a decline in accident rate?
- H: People who take a driver safety course will have a lower accident rate than those who do not take the course.
- Q: What is the relationship between age and cell phone use?
- H: Cell phone use is higher for younger adults than for older adults.

Null and Alternative Hypotheses

- The way in which the research is carried out involves forming two hypotheses, the null hypothesis H_0 and the alternative (research, or working) hypothesis H_A.
- H_A is what I, as the researcher, predict will happen.
- We are going to pick a sample and do some statistical analysis, hoping to learn something about the entire population.
- In particular, we are trying to decide whether the results we get are due to the hypothesized reason or are simply due to chance (e.g. sampling error)
- H_0 states that the predictor variable does not make a difference and that any differences that show up in the statistical analysis of the sample are due to chance. [Note that H_0 is not the opposite of H_A.]
- We test H_0; and if we can reject H_0, we have reason to accept H_A (which is what we wanted all along). But I, as a good researcher, am initially skeptical and have to have good proof that allows me to reject H_0 and therefore accept H_A.
- [Note: Failing to reject H_0 does not mean that we accept it as true – only that our statistical test did not give us reason to reject H_0. Maybe some other test would.]

Examples of H_0 and H_A

- H_A: Exercise leads to weight loss. H_0: Exercise is unrelated to weight loss.
- H_A: Exposure to classical music increases IQ score. H_0: Exposure to classical music has no effect on IQ score. [$H_{opposite\text{-}A}$: Exposure to classical music decreases IQ score.]
- H_A: Extroverts are healthier than introverts. H_0: Extrovert and introverts are equally healthy.
- H_A: Sensitivity training reduces racial bias. H_0: People exposed to sensitivity training are no more tolerant than those not exposed to sensitivity training.

Type 1 and 2 Errors

- Remember that when we are doing our hypothesis testing, we are using statistics that only have a probability of being correct, e.g. using a confidence interval with only 95% confidence
- Thus we could get into a situation (called a Type 1 Error) in which H_0 is actually true, but our hypothesis testing leads us to reject H_0 in favor of H_A.
- Or we could get into a situation (called a Type 2 Error) in which H_A is true but we do not reject H_0.
- [The other two possibilities, when H_0 is true and we don’t reject it, or when H_A is true and we do reject H_0 in favor of H_A are fine.]

Examples of Type 1 and 2 Errors

- H_0: The defendant is innocent. (Remember, in the US court system, a defendant is assumed innocent until proven guilty beyond a reasonable doubt.)
- H_A: The defendant is guilty.
- Type 1 Error: Defendant is in fact innocent but is wrongly convicted.
- Type 2 Error: Defendant was in fact guilty but the court failed to convict them.

Relation between Type 1 and Type 2 Errors

- In the example above, if we changed “beyond a reasonable doubt” to “beyond any conceivable doubt”, fewer people would be wrongly convicted, so there would be fewer Type 1 Errors; however, it would make it harder to convict people who are actually guilty, so the number of Type 2 Errors would increase.
- If we changed “beyond a reasonable doubt” to “beyond a little doubt” would lower the Type 2 Error rate but would increase the Type 1 Error rate.
- This type of reverse interaction between Type 1 and Type 2 Error rates is common.

Testing hypotheses using confidence intervals

- Research question: were students in the YRBSS study doing weight training more or less often in 2013 than they were in the past?
- We have been studying the 2013 YRBSS results for the past several classes, and now we are going to compare it with the 2011 YRBSS study.
- H_0: The average number of days per week that YRBSS students lifted weights was the same for 2011 and 2013.
- H_A: The average number of days per week that YRBSS students lifted weights was different in 2013 from in 2011.
- We know from the text that $\mu_{2011} = 3.09$, so rewrite H_0 as $\mu_{2013} = 3.09$ (null value) and H_A as $\mu_{2013} \neq 3.09$.

Example cont.

- We go back to our sample yrbss.samp of 100 students from the 2013 survey that we discussed last class.
- The book tells us that for the weight training using yrbss.samp that $\overline{x}_{13}$ = 2.78 days with a standard deviation of s_{13} = 2.56 days.
- That $\overline{x}_{13}$ = 2.78 suggests there is less weight training in 2013 than in 2011; however, we need to consider the uncertainty introduced by our sampling.
- We therefore compute the 95% confidence interval for the average for all students for the 2013 survey.

Example (still continuing)

- Remember that the confidence interval (assuming we have a normal distribution) is given by $\overline{x}_{13} \pm Z(SE)$.
- $\overline{x}_{13}$ is given as 2.78 and we saw in a previous class that Z = 1.96 when we want a 95% confidence interval.
- Remember, $SE = s_{13}/(\sqrt{n}) = 2.56/(\sqrt{100}) = .256$
- So the 95% confidence interval is

$$(2.78 – (1.96)(.256)\ ,\ 2.78 + (1.96)(.256)) = (2.27, 3.29)$$

- Since $\mu_{2011} = 3.09$ falls within the interval, we cannot say that the null hypothesis is implausible. Thus, we fail to reject the null hypothesis and cannot say that the amount of weight training is different in 2011 from the amount in 2013.
- [The book gives another example – about the cost of student housing - in which the null hypothesis is rejected.]

p-values

- We would like to be able to say more about how strongly we are able to reject or not reject the null hypothesis, e.g.
  - The null value (the parameter value under the null hypothesis) is in the 95% confidence interval but just barely, so we would not reject H_0. However, we might like to somehow say, quantitatively, that it was a close decision.
  - The null value is very far outside of the interval, so we reject H_0. However, we want to communicate that, not only did we reject the null hypothesis, but it wasn't even close.
- We will use something called the p-value to communicate this information.
- The **p-value** is the probability of observing data at least as favorable to the alternative hypothesis as our current data set, assuming the null hypothesis is true.

More about p-values

- In practice, we only want to reject H_0 when we have strong evidence. Thus we want to limit the Type 1 Errors.
- Practically, we don't want to create a Type 1 Error more than 5% of the time. We write this as $\alpha = .05$, which is known as the significance level.
- There might be cases we want to raise or lower the significance level.
- The smaller the p-value, the stronger the data favor H_A over H_0.
- Typically, if $p<.05$, we have sufficient evidence to reject H_0 in favor of H_A.

Using p-vales to test hypotheses

- The null hypothesis represents a skeptic's position or a position of no difference.
- We reject this position only if the evidence strongly favors H_A.
- A small p-value means that if the null hypothesis is true, there is a low probability of seeing a point estimate at least as extreme as the one we saw. We interpret this as strong evidence in favor of the alternative.
- We reject the null hypothesis if the p-value is smaller than the significance level, which is usually 0.05. Otherwise, we fail to reject H_0.

Calculating p-values

- There is no simple formula to plug into to calculate the p-value.
- In the kinds of problems discussed in the book, you can calculate the p-value by using your Z-table and reasoning to figure out one or both tails of a normal distribution.

4.5 Hen eggs

The distribution of the number of eggs laid by a certain species of hen during their breeding period is 35 eggs with a standard deviation of 18.2. Suppose a group of researchers randomly samples 45 hens of this species, counts the number of eggs laid during their breeding period, and records the sample mean. They repeat this 1,000 times, and build a distribution of sample means.

(a) What is this distribution called?

(b) Would you expect the shape of this distribution to be symmetric, right skewed, or left skewed?

Explain your reasoning.

(c) Calculate the variability of this distribution and state the appropriate term used to refer to

this value.

(d) Suppose the researchers' budget is reduced and they are only able to collect random samples of 10 hens. The sample mean of the number of eggs is recorded, and we repeat this 1,000 times, and build a new distribution of sample means. How will the variability of this new distribution compare to the variability of the original distribution?

4.11 Relaxing after work. The 2010 General Social Survey asked the question: “After an average work day, about how many hours do you have to relax or pursue activities that you enjoy?” to a random sample of 1,155 Americans.41 A 95% confidence interval for the mean number of hours spent relaxing or pursuing activities they enjoy was (1.38, 1.92).

(a) Interpret this interval in context of the data.

(b) Suppose another set of researchers reported a confidence interval with a larger margin of error based on the same sample of 1,155 Americans. How does their confidence level compare to the confidence level of the interval stated above?

(c) Suppose next year a new survey asking the same question is conducted, and this time the sample size is 2,500. Assuming that the population characteristics, with respect to how much time people spend relaxing after work, have not changed much within a year. How will the margin of error of the 95% confidence interval constructed based on data from the new survey compare to the margin of error of the interval stated above?