PUAD 7130: Sampling and Inference

David A. Hughes, Ph.D.

Auburn University at Montgomery

david.hughes@aum.edu

September 9, 2021

Introduction

By the time students complete this unit, they should be able to:

- Explain the best practices of sampling data,
- Use uncertainty in sampling to draw causal inferences, and
- Attach probabilistic statements to the likelihood their inferences are accurate.

Gathering data

- Why do we gather data?
- How *should* we gather data?

Sampling

- Drawing inferences about *populations of interest*.
- Sampling frame
- Probability vs. nonprobability sampling

Evaluating the accuracy of our samples

- The sampling distribution
- The central limit theorem

The standard error

- The standard error: $\hat{\sigma} = \frac{\sigma}{\sqrt{n}}$, where σ is the standard deviation of a sample, and n is the sample size
- What are each of the moving parts here up to?
- How big of a sample is big enough?

Making sense of uncertainty

- Suppose we wanted to know which grows taller: loblolly or long-leaf pine trees.
- If we measured the height of every single tree, this would be quite easy.
- But we can't (or at the very least, we shouldn't).

Making sense of uncertainty (contd.)

- We don't know the average height of *all* pines (μ).
- But we have the average height of a *sample* of them ($\bar{x}$).
- So we marshal our uncertainty from our samples and make probabilistic statements about the *likelihood* some pines are taller than others.
- We call this "hypothesis-testing."

What is hypothesis-testing?

- We start with a hypothesis: e.g., Loblolly $>$ Long-Leaf
- We then specify our "null" and "alternative" hypotheses (H_0 and H_a, respectively).
- H_a is the hypothesis you posited in your theoretical argument. H_0 says we're wrong—we assume we are.
- At the end of the day, we either "reject" or "fail to reject" the null hypothesis.

How does it work? An example

- Suppose we hypothesize that the average loblolly is *at least* 100 ft.
- A simple random sample of 9 trees yields:
  - $\bar{x} = 110$
  - $\sigma = 15$
- Can we claim with confidence that $\mu > 100$?

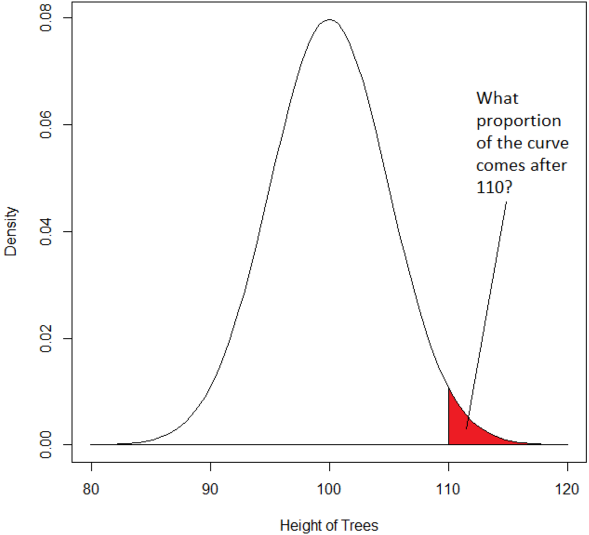
Hacking the sampling distribution

- We want to know the likelihood that we observed $\bar{x} = 110$, if the "real" mean was actually 100.
- If μ is really 100, then the sampling distribution tells us that increasingly larger values of $\bar{x}$ will be very unlikely.
- For example, we know that only 0.025% of observations fall greater than two standard errors to the right of μ.
- Therefore, we want to calculate the number of standard errors $\bar{x}$ is from H_0 and calculate the area under the curve.

Calculate the standard error for our sample of trees

- We found: $\bar{x} = 110$ and $\sigma = 15$
- And $\hat{\sigma} = \frac{\sigma}{\sqrt{n}}$.
- Therefore, $\hat{\sigma} = \frac{15}{\sqrt{9}} = 5$.

What's the probability we drew $\bar{x} = 110$ by chance?



The z-distribution

- The z-distribution is a standard normal distribution.
- A z-score is the number of standard errors an observation is from H_0.
- It is calculated with the following formula:

$$z = \frac{\bar{x} - \mu}{\hat{\sigma}_{\bar{x}}}$$

Making probabilistic statements with distributions

- We can use z-scores to make probabilistic statements.
- The proportion of the z-distribution above/below our z-score is the probability we observed that figure by chance.
- This proportion is known as a p-value. Every z-score has a corresponding p-value.

So when do we know to reject the null?

- "Critical values" help us evaluate our hypotheses, α.
- Your α specifies how small of a p-value you demand before you reject H_0.
- Therefore, α is a measure of your willingness to accept risk.

Hypothesis-testing with pine trees

- Two ways to hypothesis-test using the z-distribution
  - Compare your p-value to α.
  - Compare the absolute value of your z-score to a relevant threshold.
- Let's try this with our sample of 9 trees.

One or two-tailed tests?

- Is it enough that a sample mean be an outlier with respect to the null, or is it necessary that it is the "right" outlier?
- Resolving which is important can be critical in how we interpret our p-values *vis-à-vis* our specified α level.
- Arguably, our H_a should give us an idea about what strategy is appropriate here.

Review: The steps for hypothesis-testing

1. State the null and alternative hypotheses.
2. Choose the α level.
3. Choose a one or two-tailed test.
4. Find the z-score (the test statistic).
5. Compare this to the critical value you established.
6. "Reject" or "fail to reject" the null.

When good hypotheses go bad...

- A "Type I Error" occurs when we reject the null hypothesis, but we should not have.
- A "Type II Error" occurs when we fail to reject the null hypothesis, but we should not have.
- The probability we commit a Type I Error is increasing in α, vice versa Type II.

The confidence interval

- Confidence intervals offer another “look” at the hypothesis-test.
- For whatever your α level is, find the critical z-score associated with it.
- Calculate a 95% confidence interval (two-tailed) by: $\bar{x} \pm 1.96 \times \hat{\sigma}_{\bar{x}}$.

Comparing two groups

- Suppose:
  - H_a: $\bar{X}_1 > \bar{X}_2$
  - H_0: $\bar{X}_1 = \bar{X}_2$
- We hypothesis-test using a "difference-of-means" test
- The test:

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

Comparing two groups: An example

- We hypothesize that loblollies are taller than long-leafs.
- We gather data from 10 trees (5 loblollies, 5 long-leafs).
- We find:
  - $\bar{X}_{\text{Lob}} = 115$; $\bar{X}_{\text{Long}} = 110$
  - $\sigma^2_{\text{Lob}} = 10$; $\sigma^2_{\text{Long}} = 20$
- What's the likelihood that loblollies are, in fact, taller than long-leafs?

Conclusion

- Statistics allows us to make generalizations about populations of interest using their subsets.
- By understanding something about the underlying distributions that generate phenomena, we can test the likelihood of having observed certain data and draw causal inferences about the world around us.