6.1 and 6.3 Inference About Proportions

Eric Nordmoe

Math 261

Outline

- Formulas for Standard Errors
- Normal-based Inference for Proportions

Case Study: Hitchhiker Snails

Case Study: Hitchhiker Snails

- A type of small snail (*Tornatellides boeningi*) is very widespread in Japan, and colonies of the snails that are genetically very similar have been found very far apart.
- How could the snails travel such long distances?
- Biologist Shinichiro Wada fed 174 live snails to birds, and found that 26 were excreted live out the other end. (The snails are apparently able to seal their shells shut to keep the digestive fluids from getting in).

Yong, E. "The Scatological Hitchhiker Snail," Discover, October 2011, 13.

Research Question

What proportion of “hitchhiker” snails survive when ingested by a bird?

Interval for Proportion

1. Calculate sample statistic
2. Find $z*$ for desired level of confidence
3. Calculate standard error (see formula)
4. Use $\text{statistic} \pm z^* \times \text{SE}$
5. Interpret in context

From Last Time: Central Limit Theorem

For random samples with a *sufficiently large* sample size, the distribution of sample statistics for a mean or a proportion is approximately normal.

- For proportions, “sufficiently large” can be taken as at least 10 in each category.
- The larger the sample sizes, the more normal the distribution of the statistic will be.

Accuracy

- The accuracy of simulation methods depends on the number of simulations (more simulations $\Rightarrow$ more accurate).
- The accuracy of the normal distribution depends on the sample size (larger sample size $\Rightarrow$ more accurate).
- If the distribution of the statistic is normal and you have generated many simulated statistics, the answers should be very close.

Normal Confidence Interval Formula

$$\text{sample statistic} \pm z^* \times \text{SE}$$

- SE from bootstrap distribution.

Normal Formula for p-values

$$z = \frac{\text{sample statistic} - \text{null parameter}}{\text{SE}}$$

- Compare z to $N(0, 1)$ to get p-value.
- SE from randomization distribution.

Standard error

- The *classical approach* uses formulas rather than simulations to find standard errors.
- Methods should agree when classical conditions are met.

Standard Error Formulas

Parameter	Distribution	Standard Error
Proportion	Normal	$\sqrt{\frac{p(1-p)}{n}}$
Difference in Proportions	Normal	$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$
Mean	t, df$= n - 1$	$\sqrt{\frac{\sigma^2}{n}}$
Difference in Means	t, df$= \min(n_1, n_2) - 1$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

SE Formula Considerations

- n is always in the denominator (larger sample size gives smaller standard error).
- Standard error related to square root of $1/n$.
- Standard error formulas use population parameters (unknown!)
  - For intervals, plug in the sample statistic(s) as your best guess at the parameter(s)
  - For tests, plug in the null value for the parameter(s), because you want the distribution assuming H_0 true (if the parameters have anything to do with H_0)

Case Study: Hitchhiker Snails

- A type of small snail (*Tornatellides boeningi*) is very widespread in Japan, and colonies of the snails that are genetically very similar have been found very far apart.
- How could the snails travel such long distances?
- Biologist Shinichiro Wada fed 174 live snails to birds, and found that 26 were excreted live out the other end. (The snails are apparently able to seal their shells shut to keep the digestive fluids from getting in).

Yong, E. "The Scatological Hitchhiker Snail," Discover, October 2011, 13.

Research Question

What proportion of “hitchhiker” snails survive when ingested by a bird?

Interval for Proportion

1. Calculate sample statistic
2. Find $z*$ for desired level of confidence
3. Calculate standard error (see formula)
4. Use $\text{statistic} \pm z^* \times \text{SE}$
5. Interpret in context

Interval Calculation

- Study results: 26 of 174 snails survived
- Calculate a 90% confidence interval for p.

Conclusion

We are 90% confident that between 0.105 and 0.193 of hitchhiker snails survive being ingested by a bird.

Comparing Proportions: Bird Species Effect?

- The snails were fed to two different birds
- 14.3% of the 119 snails fed to Japanese white-eyes survived
- 16.4% of the 55 snails fed to brown-eared bulbuls survived

Japanese White-eye

Brown-eared bulbul

Photo by Laitche / CC BY-SA (https://creativecommons.org/licenses/by-sa/4.0)

Interval Calculation

- Use the study results and the standard error formula to calculate a 95% confidence interval for the difference in proportions, $p_W - p_B$.

Case Study: Hormone Replacement Therapy

Does hormone replacement therapy cause increased risk of breast cancer?

Hormone Replacement Therapy

- Until 2002, hormone replacement therapy (HRT), estrogen and/or progesterone, was commonly prescribed to post-menopausal women.
- This changed in 2002, when the results of a large clinical trial were published
  - 8506 women were randomized to take HRT and 8102 were randomized to placebo.
  - 166 HRT and 124 placebo women developed any kind of cancer.

Hypothesis Test

1. State hypotheses
2. Calculate sample statistic
3. Calculate SE
4. Calculate $z = (\text{statistic} - \text{null})/\text{SE}$
5. Compare z to standard normal distribution to find p-value.
6. State a conclusion in context.

Hypotheses

Does hormone replacement therapy cause increased risk of invasive breast cancer?

- Let p_1 = proportion of women taking HRT who get invasive breast cancer
- Let p_2 = proportion of women not taking HRT who get invasive breast cancer

What are the appropriate hypotheses here?

Calculate z-Statistic

Compute:

$$z = \frac{\text{sample statistic} - \text{null parameter}}{\text{SE}}$$

where

$$\text{SE} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

- But what to use for p_1 and p_2?
- How to assume H_0 true?

Sample Proportions

- HRT group: $\hat{p}_1 = \frac{166}{8506} = 0.0195$
- Placebo group: $\hat{p}_2 = \frac{124}{8102} = 0.0153$

Are these two significantly different?

Null Values

- Testing a difference in proportions: $H_0 : p_1 = p_2$
- Use the overall sample proportion from both groups (called the *pooled proportion*) to estimate the common value $p_1 = p_2 = p$:

$$\hat{p} = \frac{166 + 124}{8506 + 8102} = 0.017$$

- Note that this is in between the sample proportions for each group.

Standard Error Calculations

Substitute the pooled $\hat{p}$ into

$$\text{SE} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

to get

$$\text{SE} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}$$

Calculations give us the SE for the hypothesis test:

$$\text{SE} = \sqrt{\frac{0.017(1-0.017)}{8506} + \frac{0.017(1-0.017)}{8102}} = 0.002$$

Test Statistic

$$
\begin{aligned}
z &= \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}} \\
&= \frac{.0195 - .0153}{0.002} \\
&= \boxed{2.1}
\end{aligned}
$$

We can use StatKey to get the one-sided p-value of $P(Z > 2.1) = .018$

Conclusion

- If there were no difference between HRT and placebo regarding invasive breast cancer, we would only see results this extreme about 2 out of 100 times.
- We have evidence that HRT increases risk of invasive breast cancer.

Follow-up: The trial was terminated because of these findings, and hormone replacement therapy is now no longer routinely recommended.

More Practice

- Same trial, different variable of interest.
    - 8506 women were randomized to take HRT, 8102 were randomized to placebo.
    - 502 HRT and 458 placebo women developed any kind of cancer.

Does hormone replacement therapy cause increased risk of cancer in general?

Inference formulas for proportions

Parameter of Interest	Confidence Interval	Test of Significance
p	$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$
$p_1 - p_2$	$(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$	$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}}$

Margin of Error and Sample Sizes

Margin of Error

$$\hat{p} \pm z^* \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- Question: What is the margin of error?

Margin of Error for Choosing Sample Sizes

$$\text{MOE} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- You can choose your sample size in advance, depending on your desired margin of error.
- Given this formula for margin of error, solve for n to get.

$$n = \left(\frac{z^*}{\text{MOE}}\right)^2 \hat{p}(1-\hat{p})$$

- But what to use for $\hat{p}$?

Margin of Error

$$n = \left(\frac{z^*}{\text{MOE}} \right)^2 \tilde{p}(1 - \tilde{p})$$

- $\tilde{p}$ is our guessed estimate of p before collecting data.
- If no information about p is available, use $\tilde{p} = 0.5$ as the guessed values.
- For a 95% confidence interval, $z^* \approx 2$ so

$$n \approx \left(\frac{1}{\text{MOE}} \right)^2$$

Margin of Error Example

Suppose we want to estimate a proportion with a margin of error of 0.03 with 95% confidence.

- How large a sample size do we need?