Lesson 9

Estimation of a Proportion and One sample z-test of proportion

Lesson 9 Overview

- Lesson 8 covered statistical inference about population means
  - Estimation and confidence intervals of the population mean
  - Hypothesis tests of the mean from one sample
- Lesson 9 applies the same inference methods to a population proportion.

Lesson 9 Outline

- Review of binomial distribution
- Estimating the population proportion with the sample proportion
- Sampling distribution of the sample proportion
- Confidence interval for a population proportion
- One sample z-test of population proportion

Review: Binomial Distribution

- A binomial variable can only have one of two outcomes (for example, yes or no, + or -). Because there are only 2 possible outcomes binomial variables result in binary data.
- Select one of the two outcomes as a ‘success’
- For each trial (each repeat of an experiment) the probability of success (π) is constant.
- The binomial distribution gives the probability of the number of ‘successes’ in a specified number of trials (n).

Review: Binomial Distributions

- There are many binomial distributions – each are determined by the probability (π) of success and the number of specified trials (n).
- Notation for a binomial distribution is X~ B(n, π)
- For certain n and π, the binomial distribution can be approximated by the normal distribution with
  - **Mean = n π**
  - **Variance = n π $(1-\pi)$**
  - SD = $\sqrt{n\pi(1-\pi)}$

Normal Approximation of the Binomial Distribution

- The guideline for using the normal approximation to the binomial distribution is that
  - $n^* \pi > 5$ and $n^*(1 - \pi) > 5$
- For samples that do not meet this normal approximation criteria, the formula for the binomial distribution should be used (pg. 74 in text)
- Most often in health research, the normal approximation to the binomial distribution is appropriate because sample sizes are typically large enough.

Review: Proportions

- The Binomial distribution provides the probability of the total number of successes that occur over the ‘n’ trials
- Often we are interested in the proportion of successes instead of the total number of successes.
- The proportion is calculated by

$$\frac{\text{total number of successes}}{\text{number of trials}}$$

Sample proportion (p) as an estimate of π

To estimate the population proportion (π):

- The population proportion (π) is estimated by the sample proportion (p).
- The sample proportion is a statistic so it has a sampling distribution.
- Since the sample proportion can be thought of as a sample mean of binary data coded ‘1’ and ‘0’, the sampling distribution of the sample proportion is a normal distribution (by the Central Limit Theorem) when certain conditions are met.

Sampling distribution of p

If $X \sim B(n,\pi)$ and
$n^* p > 5$ and $n^*(1 - p) > 5$

Then,

$$p \sim N\left(\pi, \sqrt{\frac{\pi(1-\pi)}{n}}\right)$$

Sampling distribution of the sample proportion

- The sampling distribution of the sample proportion has
  - mean = π
  - variance = $\frac{\pi\,(1-\pi)}{n}$
- When we are estimating the population proportion, we don't know π so we use p (the sample proportion) to estimate the sample variance: $\frac{p\,(1-p)}{n}$
- The square root of the estimated sample variance is the standard error of the sample proportion:

$$SE(p) = \sqrt{\frac{p(1-p)}{n}}$$

Review: Standard Deviation vs. Standard Error

- *Standard deviation* measures the variability in the population and is based on measurements of individual observations
- *Standard error* is the standard deviation of the statistic and measures the variability of the statistic from repeated samples
  - The sampling distribution of ANY statistic has a standard error. SE(p) is the SE for the sampling distribution of the sample proportion

Confidence Interval for the population proportion (π)

- Use the general formula of the Confidence Interval for a CI of the population proportion

$$\text{Point Estimate} \pm \text{Confidence Coefficient} * \text{SE}$$

- The Point Estimate is the estimate of unknown population proportion from the sample. In this case, the sample proportion (p)
- Confidence Coefficient: the z-coefficient for the desired confidence level (i.e.1.96 for 95% CI)
- SE: Use the standard error of the sample proportion

95% Confidence Interval for the Population Proportion π

- From the sampling distribution of the sample proportion we can create a 95% confidence interval for the population proportion π.

$$p \pm 1.96 * SE(p)$$

$$\text{where} \quad SE(p) = \sqrt{\frac{p(1-p)}{n}}$$

- This is only applicable for samples that meet the normal distribution criteria:
  - n*p > 10 and n*(1-p) > 10

95% CI for population proportion when n*p ≤ 10 or n*(1-p) ≤ 10

- If the normal distribution criteria aren't met, confidence intervals of the population proportion can be calculated using *exact* binomial calculations.
- Confidence coefficients from the t-distribution are not used for confidence intervals of the population proportion.

Confidence Interval for population proportion steps

- Estimate the population proportion (π) with the sample proportion (p)
- Check that the normal approximation to the binomial is valid
- Calculate the SE (p)
- Find the z-coefficient for the confidence level
- Calculate the upper and lower limits of the confidence interval
- Interpret the results

Confidence Interval: HIV+ Infants

- Background
  - 26% of infants born to HIV+ women test positive for the virus at, or shortly after, birth.
  - Researchers believe that infants born to mothers with advanced infection are at greater risk of being infected themselves.

- In a random sample of 150 babies born to women with high viral levels (suggesting advanced infection), 107 infants were HIV+.

- Construct a 95% confidence interval for the population proportion of HIV+ infants born to HIV+ women with advanced infection.

Confidence Interval: HIV+ Infants

- The sample proportion of HIV+ infants born to mothers with advanced infection =
- Check that the normal approximation to the binomial is valid
  - 150 * 0.713 = 106.95
  - 150* (1-0.713) = 43.05
  - Both are greater than 5 so the normal approximation is valid.

Confidence Interval: HIV+ Infants

- SE of the sample proportion =

- Confidence coefficient = 1.96
  - Use NORMSINV(0.975) or 1 – NORMSINV(0.025)
- 95% Confidence Interval:
  - Lower limit: 0.7133 – 1.96* 0.037 = 0.641
  - Upper limit: 0.7133 + 1.96* 0.037 = 0.786

FATAL BICYCLE ACCIDENTS AND ALCOHOL

- In the United States approximately 900 people die in bicycle accidents each year.

- One study examined a random sample of 1711 bicyclists aged 15 or older who died in bicycle accidents between 1987 and 1991. Of these, 542 had tested positive for alcohol (blood alcohol concentration of 0.01% or higher).

WHAT IS THE ESTIMATE OF THE PROPORTION OF PEOPLE INVOLVED IN A FATAL BIKE ACCIDENT, AGES 15 AND OLDER, WHO TEST POSITIVE FOR BLOOD ALCOHOL ?

1. 1711 * 0.01
2. 542/1711
3. 900 * 0.01
4. 542/ 900
5. Cannot be estimated from this type of study!

WHAT IS THE SHAPE OF THE SAMPLING DISTRIBUTION OF THIS ESTIMATE , *p* ?

1. Normal
2. Binomial
3. Student's t_{1710}
4. Uniform
5. None of the above

THE ESTIMATE AND 95% CI FOR π IS 0.317 (0.288, 0.346). THIS CAN BE INTERPRETED AS:

1. p is bigger than 0, so people who drink and ride their bike are more likely to die.
2. The true proportion of people who drink and are involved in a fatal bike accident is estimated as 31.7%.
3. Neither of the above.

One sample z-test of Population Proportion

- Another method of statistical inference about the population proportion is hypothesis testing
- The one-sample z-test can be used to compare the proportion of subjects with a certain characteristic to some standard proportion.
- The test is a z-test because the sampling distribution of the sample proportion (when the conditions are met) has a normal distribution.

One-sample z-test: the steps

1. State the Hypotheses
2. Calculate the test statistic
3. Calculate the p-value
4. State the conclusion of the test

Z-test: Youth smoking

- One of the U.S. National Health goals of 2000 is to reduce the smoking rate for youth (ages 13 – 18) from 35% (2000) to 15% by 2010.
- At a reduction rate of 2% per year, the projected 2005 rate (to meet the 2010 goal) = 25%.
- Minnesota Public Health researchers were interested in evaluating whether the 2005 smoking rate for Metro area (Mpls. and St. Paul) youth was different from the projected goal for 2005
- Set up a hypothesis test to test whether the 2005 smoking rate for metro area youth is significantly different from the projected 2005 rate of 25%.

Z-test Step 1: State Hypotheses

First identify the population parameter of interest and the hypothesized value

- Population parameter (π) :
- Hypothesized value (π_0):

1. State the hypotheses
   - H_0:
   - H_a:

Step 2. The test statistic

- Check that n*π and n*(1-π) are both > 5
  - n=900 and $\pi_0 = 0.25$
  - 900*0.25 = 225 and 900*(1-0.25) = 675
  - Both are > 5 so the normal approximation is valid
- The sampling distribution of the sample proportion is normal so the z-test is appropriate

$$z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

Test statistics for tests of the population mean and proportion

Hypothesis Test Parameter	Sample Statistic	Test Statistic
μ Population variance known	$\overline{x}$	$Z = \frac{\overline{x} - \mu_0}{\sigma / \sqrt{n}}$
μ Population variance estimated from sample	$\overline{x}$	$t_{n-1} = \frac{\overline{x} - \mu_0}{s / \sqrt{n}}$
π $n*\pi > 5$ and $n*(1-\pi) > 5$	p	$Z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$

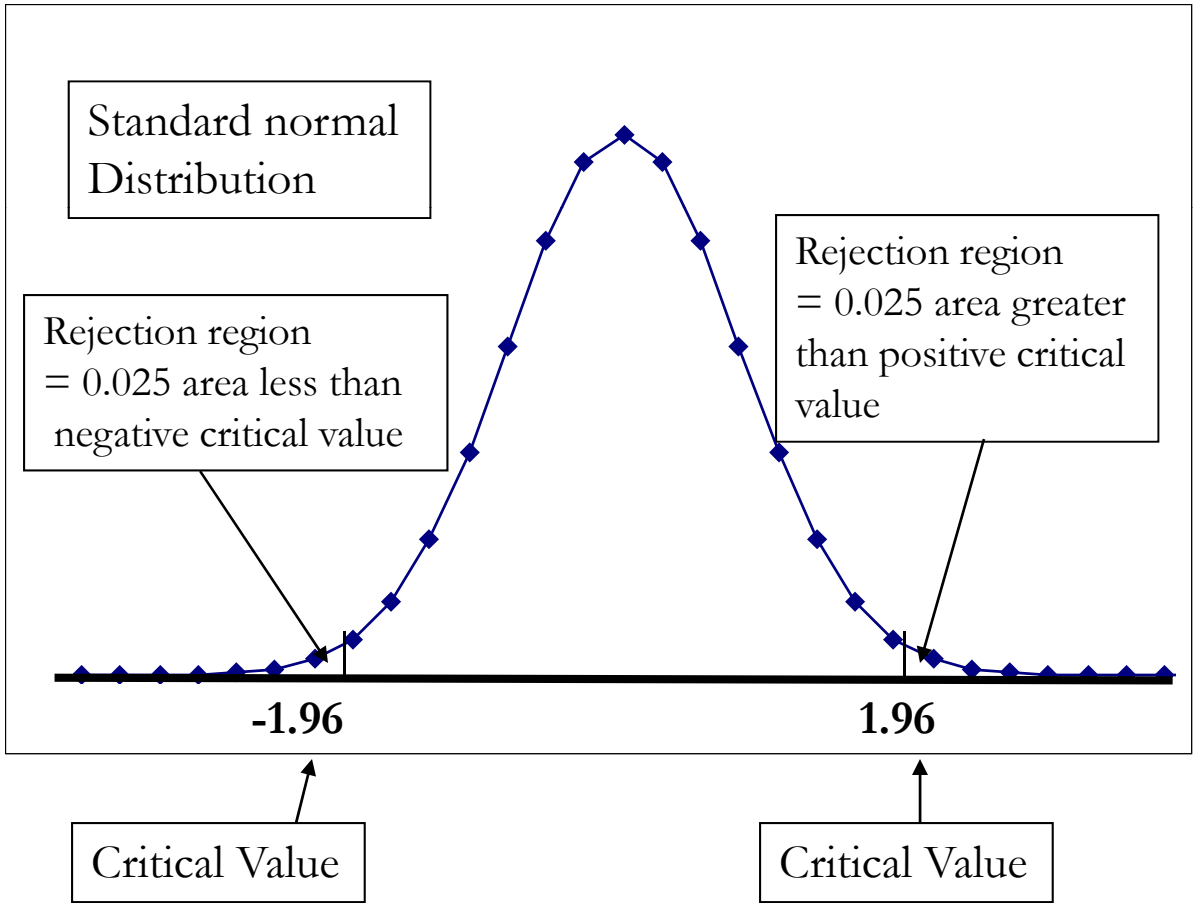
Set the significance level

- The conventional significance level for hypothesis tests is $\alpha = 0.05$
- The probability of a Type I error is 0.05 which means that the probability of incorrectly rejecting a true null hypothesis = 0.05.
- It's also possible to set up a hypothesis test with $\alpha = 0.01$ or $\alpha = 0.10$, but these aren't as common as $\alpha = 0.05$.

Determine the critical value(s) of the test

- The z-statistic used for this test has a standard normal distribution
- This is a two-tailed test so there are 2 critical values
- The rejection regions are the two tails of the standard normal distribution with 0.05/ 2 = 0.025 area in each tail
- The critical values for these rejection regions are -1.96 and 1.96
  - = NORMSINV(0.025) and =NORMSINV(0.975) in Excel

Critical values for z-test with significance level = 0.05



Collect the data

- A random sample of 900 metro area youth were interviewed in 2005 regarding their current smoking status
- 202 of the 900 youth reported that they smoked.
- The sample proportion = 202 / 900 = 0.224
- In 2005, 22.4% of the metro area youth smoked. This proportion is less than the projected proportion for 2005 of 25%.
- The hypothesis test results will determine if this difference is significant or not.

Calculate the test statistic

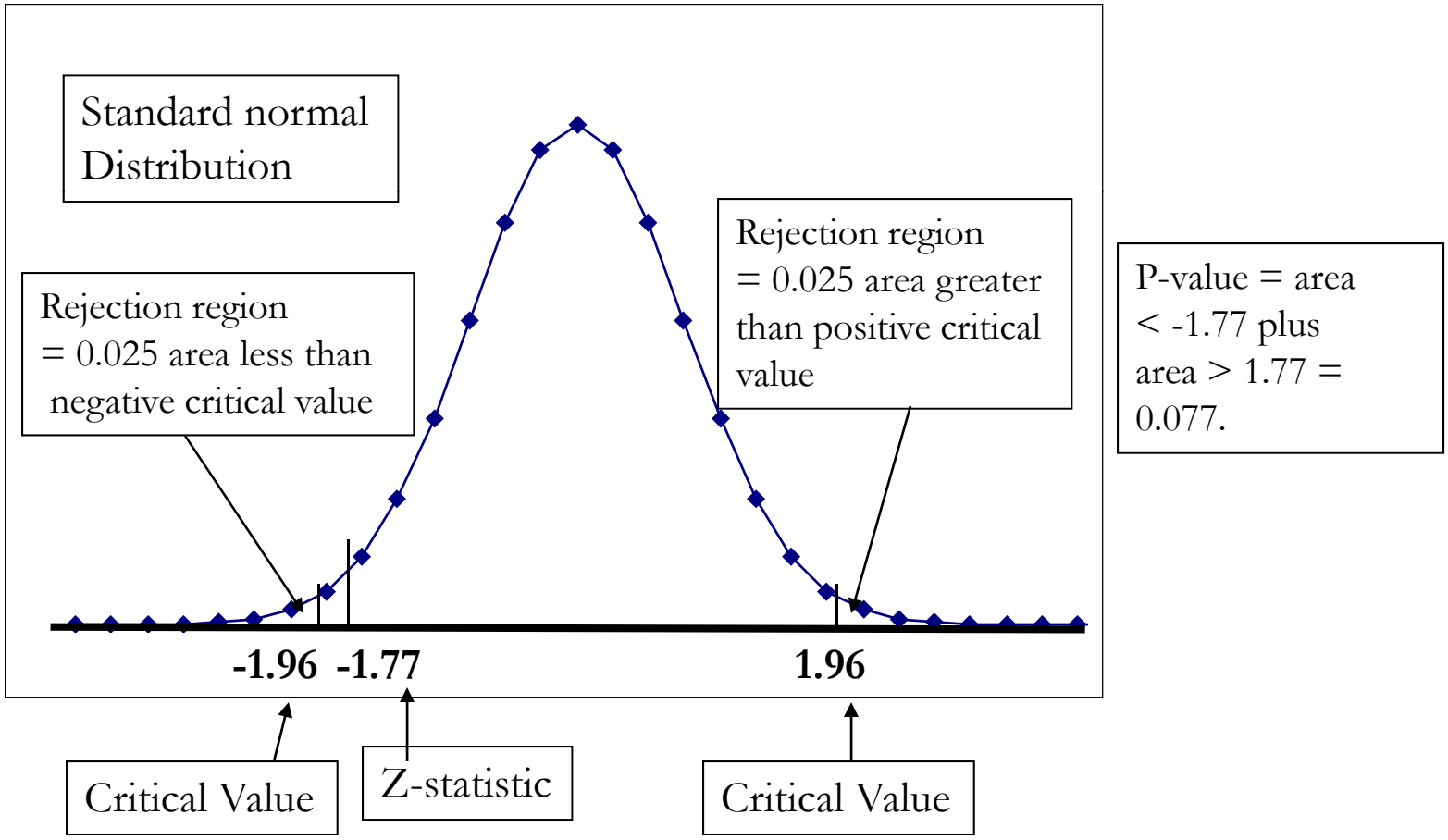
- The null hypothesis states the true proportion = 0.25
  - This value (π_o) is used to calculate the SE(p) since hypothesis testing is done with the assumption that the null hypothesis is true.
- The proportion (π) for the population of interest (metro area youth) is estimated from the sample data: p = 0.224
- Substitute these values into the formula for the z-statistic

Step 3. Calculate the p-value

- The p-value for the test statistic is the tail area of the sampling distribution beyond ± statistic
- The z-statistic = -1.77
- P-value = the area < -1.77 + area > 1.77
- Use the NORMSDIST function to find the p-value:

=NORMSDIST(-1.77) + 1-NORMDIST(1.77) = 0.077

Critical values, test statistic and p-value illustrated



Step 4. State the conclusion

- Test decision by the critical value method
  - The z-statistic is not in the rejection region so do not reject the null hypothesis
- Test Decision by the p-value method
  - The p-value (0.077) is greater than the significance level of 0.05 so do not reject the null hypothesis
- Both the critical value method and the p-value method result in the same decision: do not reject the null hypothesis that the 2005 smoking rate for Metro area youth = 0.25 (or 25%). Even though the metro area youth smoking rate is lower than the projected goal, the difference is not significant.

Step 4. State the conclusion (cont)

Either of these statements are valid descriptions of the test decision

- The 2005 Metro area youth smoking rate (22.4%) was slightly less than the U.S. projected 2005 youth smoking rate (25%) but this difference was not statistically significant (p = 0.077).
- The 2005 metro area youth smoking rate of 22.4% was marginally significantly (p = 0.077) less than the U.S. projected 2005 youth smoking rate (25%) to meet the 2010 goal of 15%.

Statistical Inference for Population Proportion

- Estimation and hypothesis tests are two methods of statistical inference for population proportions.
- Inference from the sample to the population is possible because the sampling distribution of the sample proportion is known when $n*\pi$ and $n*(1-\pi)$ are both greater than 10.
- When these conditions are met, the sample proportion has a normal distribution. Confidence intervals are constructed using z-coefficients and hypothesis tests are z-tests.

Readings and Assignments

- Reading
  - Chapter 5 pgs. 110-113
- Complete Lesson 9 Practice Exercises
- Excel Module 9 Examples
  - Confidence Interval of Proportion
  - Z-test of Proportion from one group
- Complete Homework 6 and submit by Due Date