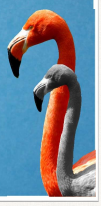
CHAPTER 10 Comparing Two Populations or Groups

10.1 Comparing Two Proportions

The Practice of Statistics, 5th Edition
Starnes, Tabor, Yates, Moore



Bedford Freeman Worth Publishers

Comparing Two Proportions

Learning Objectives

After this section, you should be able to:

- DESCRIBE the shape, center, and spread of the sampling distribution of the difference of two sample proportions.
- DETERMINE whether the conditions are met for doing inference about $p_1 - p_2$.
- CONSTRUCT and INTERPRET a confidence interval to compare two proportions.
- PERFORM a significance test to compare two proportions.

Introduction

Suppose we want to compare the proportions of individuals with a certain characteristic in Population 1 and Population 2. Let's call these parameters of interest p_1 and p_2. The ideal strategy is to take a separate random sample from each population and to compare the sample proportions with that characteristic.

What if we want to compare the effectiveness of Treatment 1 and Treatment 2 in a completely randomized experiment? This time, the parameters p_1 and p_2 that we want to compare are the true proportions of successful outcomes for each treatment. We use the proportions of successes in the two treatment groups to make the comparison.

Population or treatment	Parameter	Statistic	Sample size
1	p_1	$\hat{p}_1$	n_1
2	p_2	$\hat{p}_2$	n_2

Will I Snore When I'm 64?

In 2001, the National Sleep Foundation asked a random sample of adults about their sleep habits. One question was about snoring. Of the 995 respondents, 37% reported snoring a least a few nights a week in the past year. Split into age groups, 26.1% of the 184 people under 30 snored, while 39.2% of the 811 people 30 and over snored. What is the difference in the snoring rates for the two age groups?

We know the difference in the sample, but what is the difference in the general population?

Mar 1-2:50 PM

Can we say that the difference in the proportions is 39.2 - 26.1 = 13.1?

Why or why not?

Because we are using random samples to estimate, the proportion's we are using are just estimates of the true proportions. We expect the sample proportions to vary from sample to sample, so we need to include a margin of error.

Mar 2-1:26 PM

To answer this question, we can use our familiar formula to calculate the confidence interval:

statistic ± (critical value)•(SD of statistic)

In this case, the statistic is the difference in the proportions, $y_1 - y_2$

It doesn't matter which proportion is y_1 and which is y_2. For ease, we will put the larger proportion first.

Mar 2-8:29 AM

To answer this question, we need to know the standard deviation of the difference.

Recall that the variance of a sum or difference of two independent random variables is the sum of their variances:

Var (X - Y) = Var(X) + Var(Y)

and the standard deviation is the square root of the variance:

$SD_{(X-Y)} = \sqrt{Var(X) + Var(Y)} = \sqrt{(SD_X)^2 + (SD_Y)^2}$

Warning: X and Y must be independent

Mar 1-2:58 PM

In our example, let $\hat{p}_1$ = proportion of snorers 30 or older, and $\hat{p}_2$ = proportion of snorers under 30. n_1 and n_2 are the respective sample sizes.

$$\sigma_{\hat{p}_1} = \sqrt{\frac{p_1(1-p_1)}{n_1}} \quad and \quad \sigma_{\hat{p}_2} = \sqrt{\frac{p_2(1-p_2)}{n_2}}$$

Since we don't know p_1 and p_2, we'll use the sample proportions to calculate the standard error:

$$SE_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Mar 1-3:06 PM

$SE_{(\hat{p}_1-\hat{p}_2)}$

Mar 1-3:19 PM

Check Conditions: Same as before, but check for each sample or group.

Random: the data come from independent random samples or 2 groups in a randomized experiment

10%: $n_1 \le 0.10N_1$ and $n_2 \le 0.10N_2$

Large Counts: The counts of "successes" and "failures" in each sample or group must be at least 10.

In our snoring example:

Mar 2-8:28 AM

Construct the confidence interval:

Mar 2-8:50 AM

On the calculator:

Mar 2-9:12 AM

Conclusion:

Question: Is there a difference? How do we know?

Be sure your conclusion refers to the population, and not the sample!

Mar 2-9:14 AM

The Sampling Distribution of a Difference Between Two Proportions

To explore the sampling distribution of the difference between two proportions, let's start with two populations having a known proportion of successes.

At School 1, 70% of students did their homework last night

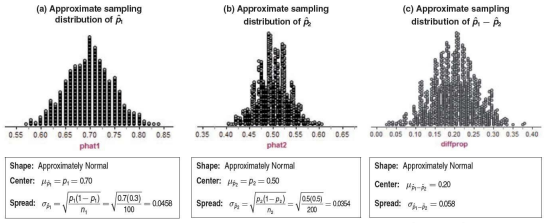
At School 2, 50% of students did their homework last night.

Suppose the counselor at School 1 takes an SRS of 100 students and records the sample proportion that did their homework.

School 2's counselor takes an SRS of 200 students and records the sample proportion that did their homework.

The Sampling Distribution of a Difference Between Two Proportions

Using Fathom software, we generated an SRS of 100 students from School 1 and a separate SRS of 200 students from School 2. The difference in sample proportions was then be calculated and plotted. We repeated this process 1000 times.



What do you notice about the shape, center, and spread of the sampling distribution of (c)?

The Sampling Distribution of a Difference Between Two Proportions

Shape: When np and n(1-p) are at least 10 for each population, the sampling distribution is approximately normal.

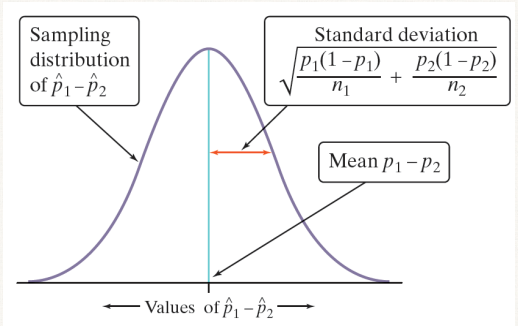
Center:

Spread:

(as long as each sample is no more than 10% of its population)

Mar 2-9:20 AM

The Sampling Distribution of a Difference Between Two Proportions



20. Anorexia The *Journal of the American Medical Association* reported on an experiment intended to see if the drug Prozac® could be used as a treatment for the eating disorder anorexia nervosa. The subjects, women being treated for anorexia, were randomly divided into two groups. Of the 49 who received Prozac, 35 were deemed healthy a year later, compared to 32 of the 44 who got the placebo.

a) Are the conditions for inference satisfied?

b) Find a 95% confidence interval for the difference in outcomes.

c) Use your confidence interval to explain whether you think Prozac is effective.

Mar 2-3:09 PM

Confidence Interval: *What* is the difference? Provides an interval of possible values for the difference?

Hypothesis Test: *Is there* a difference? Is one more than the other?

Mar 2-9:35 AM

Significance Tests for $p_1 - p_2$

An observed difference between two sample proportions can reflect an actual difference in the parameters, or it may just be due to chance variation in random sampling or random assignment. Significance tests help us decide which explanation makes more sense.

H_0: $p_1 - p_2 = 0$

H_a: $p_1 - p_2 \neq 0$

or $p_1 - p_2 > 0$

or $p_1 - p_2 < 0$

Conditions For Performing a Significance Test About A Difference In Proportions

Same as before:

Mar 2-11:38 AM

Standard Error: NOT the same!

Here's what's new:

Our null hypothesis is that the two proportions are equal. So why would we use different estimates for each? Think of our two samples as really just part of one bigger sample of the combined population.

Combine the counts to get an estimate of the proportion:

In our example: $$p_c = \frac{X_1 + X_2}{n_1 + n_2}$$

Mar 2-11:41 AM

Substitute this pooled proportion for both proportions in the SE formula:

$$SE_{pooled(p_1-p_2)} = \sqrt{\frac{p_c(1-p_c)}{n_1} + \frac{p_c(1-p_c)}{n_2}}$$

In our example:

Mar 2-1:39 PM

On the formula sheet:

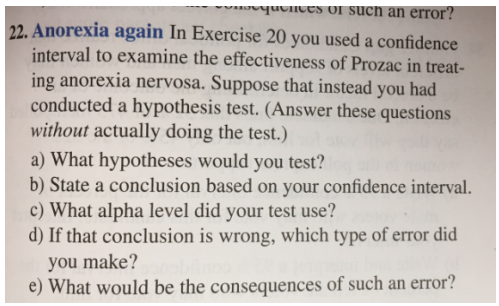
$$\sqrt{p(1-p)}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

You have to know which *p* to use. For hypothesis test of the difference of two proportions, use p_c

Mar 2-2:58 PM

Test Statistic: $z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{SE_{pooled}(\hat{p}_1 - \hat{p}_2)}$

Mar 2-1:53 PM

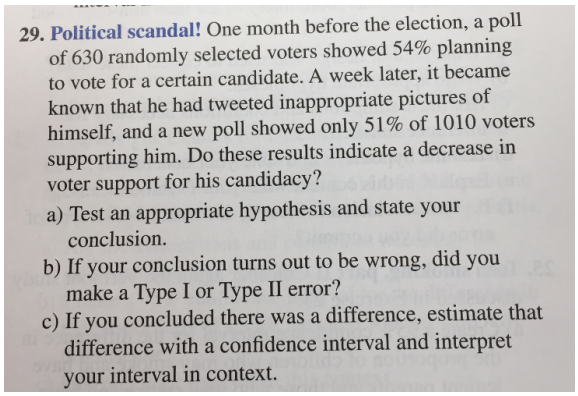


22. **Anorexia again** In Exercise 20 you used a confidence interval to examine the effectiveness of Prozac in treating anorexia nervosa. Suppose that instead you had conducted a hypothesis test. (Answer these questions *without* actually doing the test.)

a) What hypotheses would you test?
b) State a conclusion based on your confidence interval.
c) What alpha level did your test use?
d) If that conclusion is wrong, which type of error did you make?
e) What would be the consequences of such an error?

Mar 2-3:07 PM

Mar 2-3:23 PM



29. **Political scandal!** One month before the election, a poll of 630 randomly selected voters showed 54% planning to vote for a certain candidate. A week later, it became known that he had tweeted inappropriate pictures of himself, and a new poll showed only 51% of 1010 voters supporting him. Do these results indicate a decrease in voter support for his candidacy?

a) Test an appropriate hypothesis and state your conclusion.
b) If your conclusion turns out to be wrong, did you make a Type I or Type II error?
c) If you concluded there was a difference, estimate that difference with a confidence interval and interpret your interval in context.

Mar 2-3:20 PM

Mar 2-3:23 PM

Inference for Experiments

Many important statistical results come from randomized comparative experiments. Defining the parameters in experimental settings is more challenging.

•Most experiments on people use recruited volunteers as subjects.

•When subjects are not randomly selected, researchers cannot generalize the results of an experiment to some larger populations of interest.

•Researchers can draw cause-and-effect conclusions that apply to people like those who took part in the experiment.

•Unless the experimental units are randomly selected, we don't need to check the 10% condition when performing inference about an experiment.

Mar 2-3:06 PM

Comparing Two Proportions

Section Summary

In this section, we learned how to…

- DESCRIBE the shape, center, and spread of the sampling distribution of the difference of two sample proportions.
- DETERMINE whether the conditions are met for doing inference about $p_1 - p_2$
- CONSTRUCT and INTERPRET a confidence interval to compare two proportions.
- PERFORM a significance test to compare two proportions.

Mar 2-3:03 PM