10 Inferences Involving Two Populations

10.4 Inferences Concerning the Difference between Proportions Using Two Independent Samples

Inferences Concerning the Difference between Proportions Using Two Independent Samples

We are often interested in making statistical comparisons between the **proportions, percentages**, or **probabilities** associated with two populations.

Examples:

These questions ask for such comparisons:

- Is the proportion of homeowners who favor a certain tax proposal different from the proportion of renters who favor it?

- Did a larger percentage of this semester's class than of last semester's class pass statistics?

More examples:

- Is the probability of a Democratic candidate winning in New York greater than the probability of a Republican candidate winning in Texas?

- Do students' opinions about the new code of conduct differ from those of the faculty? You have probably asked similar questions.

Note

- Now, we deal with two **binomial experiments** for two populations.
    - Population 1: We obtain x_1 success in a sample of size n_1
    - Population 2: We obtain x_2 success in a sample of size n_2

Inferences Concerning the Difference between Proportions Using Two Independent Samples

Notes

These are the properties of a **binomial experiment:**

1. The observed probability is $p' = x/n$, where x is the number of observed successes in n trials.
2. $q' = 1 - p'$.
3. p is the probability of success on an individual trial in a binomial probability experiment of n repeated independent trials.

In this section, we will compare two population proportions by using the difference between the observed proportions, $p'_1 - p'_2$, of two independent samples.

- $p'_1 = \frac{x_1}{n_1}$, and $p'_2 = \frac{x_2}{n_2}$

Inferences Concerning the Difference between Proportions Using Two Independent Samples

The observed difference, $p'_1 - p'_2$, belongs to a sampling distribution with the characteristics described in the following statement.

If independent samples of sizes n_1 and n_2 are drawn randomly from large with $p_1 = P_1(\text{success})$ and $p_2 = P_2(\text{success})$, respectively, then the sampling distribution of $p'_1 - p'_2$ has these properties:

1. mean $\mu_{p'_1 - p'_2} = p_1 - p_2$

2. standard error $$\sigma_{p'_1 - p'_2} = \sqrt{\left(\frac{p_1 q_1}{n_1}\right) + \left(\frac{p_2 q_2}{n_2}\right)} \qquad \textbf{(10.10)}$$

Inferences Concerning the Difference between Proportions Using Two Independent Samples

3. An approximately normal distribution if n_1 and n_2 are sufficiently large

In practice, we use the following *guidelines to ensure normality:*

1. The sample sizes are both larger than 20.

2. The products n_1p_1, n_1q_1, n_2p_2, and n_2q_2 are all larger than 5.

3. The samples consist of less than 10% of their respective populations.

Inferences Concerning the Difference between Proportions Using Two Independent Samples

Note

p_1 and p_2 are unknown; therefore, the products mentioned in guideline 2 will be estimated by $n_1p'_1, n_1q'_1, n_2p'_2,$ and $n_2q'_2.$

Inferences about the difference between two population proportions, $p_1 - p_2$, will be based on the following assumptions.

Assumptions for inferences about the difference between two proportions $p_1 - p_2$ The n_1 random observations and the n_2 random observations that form the two samples are selected independently from two populations that do not change during the sampling.

Confidence Interval Procedure

Confidence Interval Procedure

When we estimate the **difference between two proportions**, $p_1 - p_2$, we will base our estimates on the **unbiased sample statistic** $p'_1 - p'_2$. The point estimate, $p'_1 - p'_2$, becomes the center of the confidence interval and the confidence interval limits are found using the following formula:

Confidence Interval for the Difference between Two Proportions

$$(p'_1 - p'_2) - z(\alpha/2) \cdot \sqrt{\left(\frac{p'_1 q'_1}{n_1}\right) + \left(\frac{p'_2 q'_2}{n_2}\right)} \quad \text{to}$$

$$(p'_1 - p'_2) + z(\alpha/2) \cdot \sqrt{\left(\frac{p'_1 q'_1}{n_1}\right) + \left(\frac{p'_2 q'_2}{n_2}\right)} \qquad \textbf{(10.11)}$$

Example 12 – *Constructing a Confidence Interval for the Difference between Two Proportions*

In studying his campaign plans, Mr. Morris wishes to estimate the difference between men's and women's views regarding his appeal as a candidate.

He asks his campaign manager to take two random independent samples and find the 99% confidence interval for the difference between the proportions of women and men voters who plan to vote for him.

A sample of 1000 voters was taken from each population, with 388 men and 459 women favoring Mr. Morris.

Example 12 – *Solution*

Step 1 **Parameter of interest:** $p_w - p_m$, the difference between the proportion of women voters and the proportion of men voters who plan to vote for Mr. Morris

Step 2 **a. Assumptions:** The samples are randomly and independently selected.

b. Probability distribution: The standard normal distribution. The populations are large (all voters); the sample sizes are larger than 20; and the estimated values for $n_m p_m$, $n_m q_m$, $n_w p_w$, and $n_w q_w$ are all larger than 5.

Example 12 – *Solution*

cont'd

Therefore, the sampling distribution of $p'_w - p'_m$ should have an approximately normal distribution. The interval will be calculated using formula (10.11).

c. Level of confidence: $1 - \alpha = 0.99$

Step 3 Sample Information:

We have $n_m = 1000$, $x_m = 388$, $n_w = 1000$, and $x_w = 459$.

$$p'_m = \frac{x_m}{n_m} = \frac{388}{1000} = \mathbf{0.388} \qquad q'_m = 1 - 0.388 = \mathbf{0.612}$$

$$p'_w = \frac{x_w}{n_w} = \frac{459}{1000} = \mathbf{0.459} \qquad q'_w = 1 - 0.459 = \mathbf{0.541}$$

Example 12 – *Solution*

cont'd

Step 4

a. Confidence coefficient: This is a two-tailed situation, with $\alpha/2$ in each tail.

$z(\alpha/2) = z(0.005) = 2.58.$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
2.0	0.9773	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9865	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9980	0.9980	0.9981

b. Maximum error of estimate: Using the maximum error part of formula (10.11), we have

$$E = z(\alpha/2) \cdot \sqrt{\left(\frac{p'_w q'_w}{n_w}\right) + \left(\frac{p'_m q'_m}{n_m}\right)}$$

$$E = 2.58 \cdot \sqrt{\left(\frac{(0.459)(0.541)}{1000}\right) + \left(\frac{(0.388)(0.612)}{1000}\right)}$$

$$= 2.58\sqrt{0.000248 + 0.000237} = (2.58)(0.022) = \mathbf{0.057}$$

Example 12 – *Solution*

cont'd

c. Lower/upper confidence limits:

$$(p'_w - p'_m) \pm E$$

$$0.071 \pm 0.057$$

0.071 – 0.057 = **0.014** to 0.071 + 0.057 = **0.128**

Step 5 **a. Confidence interval:** 0.014 to 0.128 is the 99% confidence interval for $p_w - p_m$. With 99% confidence, we can say that there is a difference of 1.4% to 12.8% in Mr. Morris's voter appeal.

b. That is, a larger proportion of women than men favor Mr. Morris, and the difference in the proportions is between 1.4% and 12.8%.

Confidence Interval Procedure

Confidence intervals and hypothesis tests can sometimes be interchanged; that is, a confidence interval can be used in place of a hypothesis test.

For example, Example 12 called for a confidence interval. Now suppose that Mr. Morris asked, “Is there a difference in my voter appeal to men voters as opposed to women voters?”

To answer his question, you would not need to complete a hypothesis test if you chose to test at $\alpha = 0.01$ using a two-tailed test.

Confidence Interval Procedure

"No difference" would mean a difference of zero, which is not included in the interval from 0.014 to 0.128 (the interval determined in Example 12).

Therefore, a null hypothesis of "no difference" would be rejected, thereby substantiating the conclusion that a significant difference exists in voter appeal between the two groups.

Hypothesis-Testing Procedure

Hypothesis-Testing Procedure

When the null **hypothesis—there is no difference between two proportions**—is being tested,

$$H_0 : p_1 = p_2$$

the **test statistic** will be the difference between the observed proportions divided by the **standard error**; it is found with the following formula:

Test Statistic for the Difference between Two Proportions—Population Proportion Known

$$z\bigstar = \frac{p'_1 - p'_2}{\sqrt{pq\left[\left(\frac{1}{n_1}\right) + \left(\frac{1}{n_2}\right)\right]}} \qquad \textbf{(10.12)}$$

Hypothesis-Testing Procedure

Notes

1. The null hypothesis is $p_1 = p_2$ or $p_1 - p_2 = 0$ (the difference is zero).

2. Nonzero differences between proportions are not discussed in this section.

3. The numerator of formula (10.12) could be written as $(p'_1 - p'_2) - (p_1 - p_2)$, but since the null hypothesis is assumed to be true during the test, $p_1 - p_2 = 0$. By substitution, the numerator becomes simply $p'_1 - p'_2$.

Hypothesis-Testing Procedure

4. Since the null hypothesis is $p_1 = p_2$, the standard error of $p_1' - p_2'$, $\sqrt{\left(\frac{p_1 q_1}{n_1}\right) + \left(\frac{p_2 q_2}{n_2}\right)}$, can be written as $\sqrt{pq\left[\left(\frac{1}{n_1}\right) + \left(\frac{1}{n_2}\right)\right]}$, where $p = p_1 = p_2$ and $q = 1 - p$.

5. When the null hypothesis states $p_1 = p_2$ and does not specify the value of either p_1 or p_2, the two sets of sample data will be pooled to obtain the estimate for *p.*

Hypothesis-Testing Procedure

This pooled probability (known as p'_p) is the total number of successes divided by the total number of observations with the two samples combined; it is found using the next formula:

$$p'_p = \frac{x_1 + x_2}{n_1 + n_2} \tag{10.13}$$

and q'_p is its complement,

$$q'_p = 1 - p'_p \tag{10.14}$$

Hypothesis-Testing Procedure

When the pooled estimate, p'_p, is being used, formula (10.12) becomes formula (10.15):

Test Statistic for the Difference between Two Proportions—Population Proportion Unknown

$$z\star = \frac{p'_1 - p'_2}{\sqrt{(p'_p)(q'_p)\left[\left(\frac{1}{n_1}\right) + \left(\frac{1}{n_2}\right)\right]}} \quad (10.15)$$

Example 13 – *One-Tailed Hypothesis Test for the Difference between Two Proportions*

A salesperson for a new manufacturer of cellular phones claims not only that they cost the retailer less but also that the percentage of defective cellular phones found among her products will be no higher than the percentage of defectives found in a competitor's line.

To test this statement, a retailer took random samples of each manufacturer's product.

Product	Number Defective	Number Checked
Salesperson's	15	150
Competitor's	6	150

Can we reject the salesperson's claim at the 0.05 level of significance?

Can we reject the salesperson’s claim at the 0.05 level of significance?

Solution:

Step 1 **a. Parameter of interest:** $p_s - p_c$: the difference between the proportion of defectives in the salesperson’s product and the proportion of defectives in the competitor’s product

b. Statement of hypotheses: The concern of the retailer is that the salesperson’s less expensive product may be of a poorer quality, meaning a greater proportion of defectives.

Example 13 – *Solution*

cont'd

If we use the difference "suspected larger proportion – smaller proportion," then the alternative hypothesis is "The difference is positive (greater than zero)."

H_o: $p_s - p_c = 0$ ($\leq$) (salesperson's defective rate is no higher than competitor's)

H_a: $p_s - p_c > 0$ (salesperson's defective rate is higher than competitor's)

Example 13 – *Solution*

cont'd

Step 2 **a. Assumptions:** Random samples were selected from the products of two different manufacturers.

b. The test statistic to be used: The standard normal distribution. Populations are very large (all cellular phones produced); the samples are larger than 20; and the estimated products $n_s p'_s$, $n_s q'_s$, $n_c p'_c$, and $n_c q'_c$ are all larger than 5. Therefore, the sampling distribution should have an approximately normal distribution. $z\bigstar$ will be calculated using formula (10.15).

c. Level of significance: $\alpha = 0.05$

Example 13 – *Solution*

cont'd

Step 3 **a. Sample information:**

$$p'_s = \frac{x_s}{n_s} = \frac{15}{150} = \mathbf{0.10}$$

$$p'_c = \frac{x_c}{n_c} = \frac{6}{150} = \mathbf{0.04}$$

$$p'_p = \frac{x_1 + x_2}{n_1 + n_2} = \frac{15 + 6}{150 + 150} = \frac{21}{300} = \mathbf{0.07}$$

$$q'_p = 1 - p'_p = 1 - 0.07 = \mathbf{0.93}$$

b. Calculated test statistic:

$$z\bigstar = \frac{p'_s - p'_c}{\sqrt{(p'_p)(q'_p)\left[\left(\frac{1}{n_s}\right) + \left(\frac{1}{n_c}\right)\right]}}: \quad z\bigstar = \frac{0.10 - 0.04}{\sqrt{(0.07)(0.93)\left[\left(\frac{1}{150}\right) + \left(\frac{1}{150}\right)\right]}}$$

$$= \frac{0.06}{\sqrt{0.000868}} = \frac{0.06}{0.02946} = \mathbf{2.04}$$

Example 13 – *Solution*

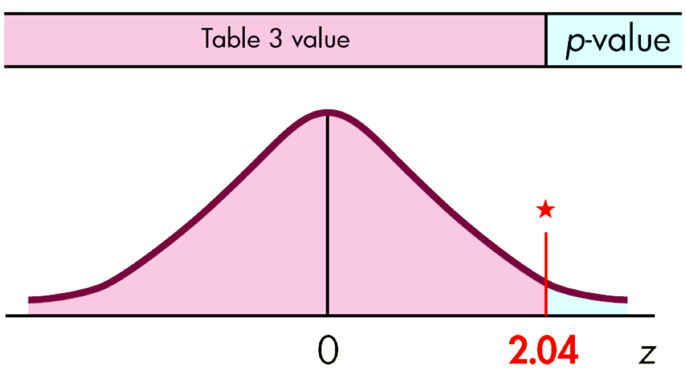
cont'd

Step 4 Probability Distribution:

***p*-Value approach:**

a. Use the right-hand tail because H_a expresses concern for values related to "higher than."

P = *p*-value = $P(z\bigstar > 2.04)$, as shown in the figure.



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
2.0	0.9773	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9865	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9980	0.9980	0.9981
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

p-value: **P** = 1.0000 – 0.9793 = **0.0207**

b. The *p*-value is smaller than α.

Example 13 – *Solution*

cont'd

Step 4 Probability Distribution:

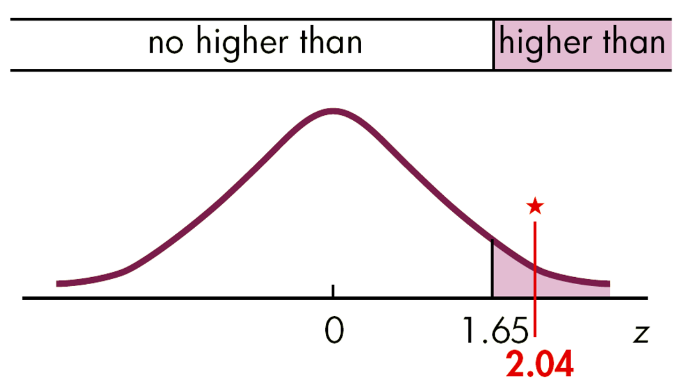
Classical approach:

a. The critical region is the right-hand tail because H_a expresses concern for values related to "higher than."
The critical value is:

$z(0.05) = \mathbf{1.65.}$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9430	0.9441
1.6	0.9452	0.9463	0.9474	0.9485	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9700	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9762	0.9767
2.0	0.9773	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817

b. $z^\bigstar$ is in the critical region, as shown in **red** in the figure.



Example 13 – *Solution*

cont'd

Step 5 **a. Decision:** Reject H_o.

b. Conclusion: At the 0.05 level of significance, there is sufficient evidence to reject the salesperson's claim; the proportion of her company's cellular phones that are defective is higher than the proportion of her competitor's cellular phones that are defective.