Chapter 8: Inference for Proportions

MATH 560-01
Statistical Data Analysis

March 8, 2021

These slides are based on material from *Introduction to the Practice of Statistics* by David S. Moore, George P. McCabe, and Bruce A. Craig, 9th edition.

Sections

8.1 Inference for a Single Proportion

8.2 Comparing Two Proportions

8.1 Inference for a Single Proportion

After completing this section, students should be able to:

- compute *approximate large-sample confidence intervals for the population proportion* p in the Binomial setting when the sample size is large

8.1 Inference for a Single Proportion

After completing this section, students should be able to:

- compute *approximate large-sample confidence intervals for the population proportion* p in the Binomial setting when the sample size is large
- in the Binomial setting when the sample size is large, determine the *sample size* needed so that the margin of error for the approximate confidence interval is no more than a specified value m

8.1 Inference for a Single Proportion

After completing this section, students should be able to:

- compute *approximate large-sample confidence intervals for the population proportion* p in the Binomial setting when the sample size is large
- in the Binomial setting when the sample size is large, determine the *sample size* needed so that the margin of error for the approximate confidence interval is no more than a specified value m
- perform approximate *large-sample one-sided* or *two-sided* significance tests involving the population proportion p in the Binomial setting when the sample size is large

8.1 Inference for a Single Proportion

After completing this section, students should be able to:

- compute *approximate large-sample confidence intervals for the population proportion* p in the Binomial setting when the sample size is large
- in the Binomial setting when the sample size is large, determine the *sample size* needed so that the margin of error for the approximate confidence interval is no more than a specified value m
- perform approximate *large-sample one-sided* or *two-sided* significance tests involving the population proportion p in the Binomial setting when the sample size is large
- determine whether the large-sample confidence interval and large-sample significance test is appropriate based on the guidelines described in this section

8.1 Inference for a Single Proportion

- In this section, we discuss how to make inferences about p, the proportion of some characteristic in a large population.

8.1 Inference for a Single Proportion

- In this section, we discuss how to make inferences about p, the proportion of some characteristic in a large population.
- We select a random sample of size n from the population and count the number of "successes" X in the sample.

8.1 Inference for a Single Proportion

- In this section, we discuss how to make inferences about p, the proportion of some characteristic in a large population.
- We select a random sample of size n from the population and count the number of “successes” X in the sample.
- Then, we use the sample proportion $\hat{p} = \dfrac{X}{n}$ to estimate p.

8.1 Inference for a Single Proportion

- In this section, we discuss how to make inferences about p, the proportion of some characteristic in a large population.
- We select a random sample of size n from the population and count the number of "successes" X in the sample.
- Then, we use the sample proportion $\hat{p} = \dfrac{X}{n}$ to estimate p.
- If the sample is taken with replacement, then X follows a $B(n, p)$ distribution.

8.1 Inference for a Single Proportion

- In this section, we discuss how to make inferences about p, the proportion of some characteristic in a large population.
- We select a random sample of size n from the population and count the number of "successes" X in the sample.
- Then, we use the sample proportion $\hat{p} = \dfrac{X}{n}$ to estimate p.
- If the sample is taken with replacement, then X follows a $B(n, p)$ distribution.
- If we take a SRS and the population size is much larger than n (at least 20 times as large), then X approximately follows a $B(n, p)$ distribution.

8.1 Inference for a Single Proportion

- If n is small, we must use tests based on the Binomial distribution. These are awkward because of the discrete nature of Binomial distributions.

8.1 Inference for a Single Proportion

- If n is small, we must use tests based on the Binomial distribution. These are awkward because of the discrete nature of Binomial distributions.
- However, when n is large, $\hat{p}$ is approximately $N\left(p, \sqrt{\dfrac{p(1-p)}{n}}\right)$ (see Chapter 5) so we can use tests based on the Normal distribution.

8.1 Inference for a Single Proportion

- If n is small, we must use tests based on the Binomial distribution. These are awkward because of the discrete nature of Binomial distributions.
- However, when n is large, $\hat{p}$ is approximately $N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$ (see Chapter 5) so we can use tests based on the Normal distribution.
- In this chapter, we only consider inference procedures for large sample sizes based on the Normal distribution.

8.1 Inference for a Single Proportion

Approximate large-sample confidence interval for p when n is large

- A level C confidence interval for p is

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

where z^* is the value for the $N(0,1)$ density with area C between $-z^*$ and z^* and $\hat{p} = \dfrac{X}{n}$.

8.1 Inference for a Single Proportion

Approximate large-sample confidence interval for p when n is large

- A level C confidence interval for p is

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

where z^* is the value for the $N(0,1)$ density with area C between $-z^*$ and z^* and $\hat{p} = \frac{X}{n}$.

- To use this approximation, the number of successes and the number of failures observed in the sample should be at least 10.

8.1 Inference for a Single Proportion

- **Example 8A**: The South African mathematician John Kerrich, while a prisoner of war in World War II, tossed a coin 10,000 times and obtained 5,067 heads.

  (a) Find a 95% confidence interval for the probability that the Kerrich's coin comes up heads.
  (b) Are the guidelines for when to use the large-sample confidence interval for a population proportion satisfied in this setting?

8.1 Inference for a Single Proportion

- *Answer*:

8.1 Inference for a Single Proportion

Choosing the sample size

- Suppose we are planning an experiment and want to choose the sample size so that the margin of error is at most m for a level C confidence interval.

8.1 Inference for a Single Proportion

Choosing the sample size

- Suppose we are planning an experiment and want to choose the sample size so that the margin of error is at most m for a level C confidence interval.
- Here, z^* is the critical value for confidence C and p^* is a guess for p.

$$z^* \sqrt{\frac{p^*(1-p^*)}{n}} \quad \leq \quad m$$

8.1 Inference for a Single Proportion

Choosing the sample size

- Suppose we are planning an experiment and want to choose the sample size so that the margin of error is at most m for a level C confidence interval.
- Here, z^* is the critical value for confidence C and p^* is a guess for p.

$$
\begin{aligned}
z^* \sqrt{\frac{p^*(1-p^*)}{n}} &\leq m \\
\frac{z^* \sqrt{p^*(1-p^*)}}{m} &\leq \sqrt{n}
\end{aligned}
$$

8.1 Inference for a Single Proportion

Choosing the sample size

- Suppose we are planning an experiment and want to choose the sample size so that the margin of error is at most m for a level C confidence interval.
- Here, z^* is the critical value for confidence C and p^* is a guess for p.

$$\begin{aligned} z^* \sqrt{\frac{p^*(1-p^*)}{n}} &\leq m \\ \frac{z^* \sqrt{p^*(1-p^*)}}{m} &\leq \sqrt{n} \\ \left(\frac{z^*}{m}\right)^2 p^*(1-p^*) &\leq n \end{aligned}$$

8.1 Inference for a Single Proportion

Choosing the sample size

- So, we need n to be at least $\left(\frac{z^*}{m}\right)^2 p^*(1-p^*)$.

8.1 Inference for a Single Proportion

Choosing the sample size

- So, we need n to be at least $\left(\frac{z^*}{m}\right)^2 p^*(1-p^*)$.
- A conservative choice for p^* is 0.5 since $p^*(1-p^*)$ is maximized when $p^* = 0.5$.

8.1 Inference for a Single Proportion

Choosing the sample size

- So, we need n to be at least $\left(\frac{z^*}{m}\right)^2 p^*(1-p^*)$.
- A conservative choice for p^* is 0.5 since $p^*(1-p^*)$ is maximized when $p^* = 0.5$.
- So choosing n to be at least $\frac{1}{4}\left(\frac{z^*}{m}\right)^2$ will work regardless of the true p.

8.1 Inference for a Single Proportion

- **Example 8B**: Suppose we plan to compute a 99% confidence interval for the probability that a coin comes up heads and want the margin of error to be no more than 0.01. How many times do we need to toss the coin to guarantee this margin of error?

8.1 Inference for a Single Proportion

- **Example 8B**: Suppose we plan to compute a 99% confidence interval for the probability that a coin comes up heads and want the margin of error to be no more than 0.01. How many times do we need to toss the coin to guarantee this margin of error?
- *Answer*:

8.1 Inference for a Single Proportion

Two-sided test for a population proportion when n is large

1. Test $H_0 : p = p_0$ vs. $H_a : p \neq p_0$

8.1 Inference for a Single Proportion

Two-sided test for a population proportion when n is large

1. Test $H_0 : p = p_0$ vs. $H_a : p \neq p_0$
2. Test statistic: $z = \dfrac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$

8.1 Inference for a Single Proportion

Two-sided test for a population proportion when n is large

1. Test $H_0 : p = p_0$ vs. $H_a : p \neq p_0$
2. Test statistic: $z = \dfrac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$
3. Use the Normal table to compute the P-value by finding the area in both tails: $P(|Z| > |z|) = 2P(Z > |z|)$

8.1 Inference for a Single Proportion

Two-sided test for a population proportion when n is large

1. Test $H_0 : p = p_0$ vs. $H_a : p \neq p_0$
2. Test statistic: $z = \dfrac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$
3. Use the Normal table to compute the P-value by finding the area in both tails: $P(|Z| > |z|) = 2P(Z > |z|)$
4. Reject H_0 if the P-value $\leq \alpha$

The book's guidelines recommend that this large-sample test should only be used when the expected number of successes np_0 and the expected number of failures $n(1 - p_0)$ are both greater than 10.

8.1 Inference for a Single Proportion

One-sided (left-sided) test for a population proportion when n is large

1. Test $H_0 : p = p_0$ vs. $H_a : p < p_0$
2. Test statistic: $z = \dfrac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$
3. Use the Normal table to compute the P-value by finding the area: $P(Z < z)$
4. Reject H_0 if the P-value $\leq \alpha$

The book's guidelines recommend that this large-sample test should only be used when the expected number of successes np_0 and the expected number of failures $n(1 - p_0)$ are both greater than 10.

8.1 Inference for a Single Proportion

One-sided (right-sided) test for a population proportion when n is large

1. Test $H_0 : p = p_0$ vs. $H_a : p > p_0$
2. Test statistic: $z = \dfrac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$
3. Use the Normal table to compute the P-value by finding the area: $P(Z > z)$
4. Reject H_0 if the P-value $\leq \alpha$

The book's guidelines recommend that this large-sample test should only be used when the expected number of successes np_0 and the expected number of failures $n(1 - p_0)$ are both greater than 10.

8.1 Inference for a Single Proportion

- **Example 8C**: The South African mathematician John Kerrich, while a prisoner of war in World War II, tossed a coin 10,000 times and obtained 5,067 heads.

  (a) Is there significant evidence at the 5% level that the probability that Kerrich's coin comes up heads is not 0.5?
  (b) Are the guidelines for when to use the large-sample significance test for a population proportion satisfied in this setting?

8.1 Inference for a Single Proportion

- *Answer*:

8.2 Comparing Two Proportions

After completing this section, students should be able to:

- compute *approximate large-sample confidence intervals for the difference* $p_1 - p_2$ *between population proportions* in the Binomial setting when the sample sizes are large and the samples are independent

8.2 Comparing Two Proportions

After completing this section, students should be able to:

- compute *approximate large-sample confidence intervals for the difference* $p_1 - p_2$ *between population proportions* in the Binomial setting when the sample sizes are large and the samples are independent
- perform *approximate large-sample one-sided* or *two-sided tests* for equality of two population proportions in the Binomial setting when the sample sizes are large and the samples are independent

8.2 Comparing Two Proportions

After completing this section, students should be able to:

- compute *approximate large-sample confidence intervals for the difference* $p_1 - p_2$ *between population proportions* in the Binomial setting when the sample sizes are large and the samples are independent
- perform *approximate large-sample one-sided* or *two-sided tests* for equality of two population proportions in the Binomial setting when the sample sizes are large and the samples are independent
- determine whether the large-sample confidence interval and large-sample significance test is appropriate based on the guidelines described in this section

8.2 Comparing Two Proportions

- In comparative studies of two populations, it is often of interest to compare proportions of some characteristics.

8.2 Comparing Two Proportions

- In comparative studies of two populations, it is often of interest to compare proportions of some characteristics.
- In this section, we discuss how to make inferences about the difference between the two population proportions.

8.2 Comparing Two Proportions

- In comparative studies of two populations, it is often of interest to compare proportions of some characteristics.
- In this section, we discuss how to make inferences about the difference between the two population proportions.
- Here is notation used for this section:

Population	Population proportion	Sample size	Count of successes	Sample proportion
1	p_1	n_1	X_1	$\hat{p}_1$
2	p_2	n_2	X_2	$\hat{p}_2$

8.2 Comparing Two Proportions

- We use the difference $D = \hat{p}_1 - \hat{p}_2$ to compare the two populations.

8.2 Comparing Two Proportions

- We use the difference $D = \hat{p}_1 - \hat{p}_2$ to compare the two populations.
- If we have two independent random samples, then the sampling distribution of D is approximately $N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$ when both sample sizes are sufficiently large.

8.2 Comparing Two Proportions

- We use the difference $D = \hat{p}_1 - \hat{p}_2$ to compare the two populations.
- If we have two independent random samples, then the sampling distribution of D is approximately $N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$ when both sample sizes are sufficiently large.
  $\mu_{\hat{p}_1 - \hat{p}_2} = \mu_{\hat{p}_1} - \mu_{\hat{p}_2} = p_1 - p_2$

8.2 Comparing Two Proportions

- We use the difference $D = \hat{p}_1 - \hat{p}_2$ to compare the two populations.
- If we have two independent random samples, then the sampling distribution of D is approximately $N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$ when both sample sizes are sufficiently large.
  $\mu_{\hat{p}_1 - \hat{p}_2} = \mu_{\hat{p}_1} - \mu_{\hat{p}_2} = p_1 - p_2$
  $\sigma^2_{\hat{p}_1 - \hat{p}_2} = \sigma^2_{\hat{p}_1} + \sigma^2_{\hat{p}_2} = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$

8.2 Comparing Two Proportions

Approximate confidence interval for $p_1 - p_2$
when n_1 and n_2 are large

- A level C confidence interval for $p_1 - p_2$ is

$$D \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

where z^* is the value for the $N(0,1)$ density with area C between $-z^*$ and z^* and $D = \hat{p}_1 - \hat{p}_2$.

8.2 Comparing Two Proportions

Approximate confidence interval for $p_1 - p_2$
when n_1 and n_2 are large

- A level C confidence interval for $p_1 - p_2$ is

$$D \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

  where z^* is the value for the $N(0,1)$ density with area C between $-z^*$ and z^* and $D = \hat{p}_1 - \hat{p}_2$.
- To use this approximation, the number of successes and the number of failures observed in each sample should be at least 10.

8.2 Comparing Two Proportions

- **Example 8D**: A study examined the proportion of high school students who cheated on tests at least twice in 2002 and 2004. A reported 9054 out of 24142 students said they cheated at least twice in 2004. A reported 5794 out of 12121 said they cheated at least twice in 2002. Compute a 90% confidence interval for $p_{2004} - p_{2002}$ where p_{2004} is the proportion of students from the 2004 population who cheated at least twice and p_{2002} is the proportion of students from the 2002 population who cheated at least twice.

8.2 Comparing Two Proportions

- *Answer*:

8.2 Comparing Two Proportions

Significance Test for Difference in Proportions

- The standard error calculation is slightly different in the setting of a significance test with a null hypothesis $H_0 : p_1 = p_2$.

8.2 Comparing Two Proportions

Significance Test for Difference in Proportions

- The standard error calculation is slightly different in the setting of a significance test with a null hypothesis $H_0 : p_1 = p_2$.
- If H_0 is true, then both population have the same proportion (say p) and we use a pooled estimate $\hat{p} = \frac{X_1+X_2}{n_1+n_2}$ in the calculation of the standard error; that is, when H_0 is true, $\sigma_D = \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$ so $SE_{D_p} = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$.

8.2 Comparing Two Proportions

Two-sided test for equality of population proportions when n_1 and n_2 are large

1. Test $H_0 : p_1 = p_2$ vs. $H_a : p_1 \neq p_2$

8.2 Comparing Two Proportions

Two-sided test for equality of population proportions when n_1 and n_2 are large

1. Test $H_0 : p_1 = p_2$ vs. $H_a : p_1 \neq p_2$
2. Test statistic: $z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\,(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ where

$$\hat{p}_1 = \frac{X_1}{n_1},\ \hat{p}_2 = \frac{X_2}{n_2},\ \text{and } \hat{p} = \frac{X_1 + X_2}{n_1 + n_2}.$$

8.2 Comparing Two Proportions

Two-sided test for equality of population proportions when n_1 and n_2 are large

1. Test $H_0 : p_1 = p_2$ vs. $H_a : p_1 \neq p_2$
2. Test statistic: $z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\,(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ where

   $\hat{p}_1 = \dfrac{X_1}{n_1}$, $\hat{p}_2 = \dfrac{X_2}{n_2}$, and $\hat{p} = \dfrac{X_1 + X_2}{n_1 + n_2}$.
3. Use the Normal table to compute the P-value by finding the area in both tails: $P(|Z| > |z|) = 2P(Z > |z|)$

8.2 Comparing Two Proportions

Two-sided test for equality of population proportions when n_1 and n_2 are large

1. Test $H_0 : p_1 = p_2$ vs. $H_a : p_1 \neq p_2$
2. Test statistic: $z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\,(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$ where

   $$\hat{p}_1 = \frac{X_1}{n_1},\ \hat{p}_2 = \frac{X_2}{n_2},\ \text{and } \hat{p} = \frac{X_1 + X_2}{n_1 + n_2}.$$
3. Use the Normal table to compute the P-value by finding the area in both tails: $P(|Z| > |z|) = 2P(Z > |z|)$
4. Reject H_0 if the P-value $\leq \alpha$

The book's guidelines recommend that this large-sample test should only be used when the number of successes and the number of failures are both at least 5.

8.2 Comparing Two Proportions

One-sided (left-sided) test for equality of population proportions when n_1 and n_2 are large

1. Test $H_0 : p_1 = p_2$ vs. $H_a : p_1 < p_2$
2. Test statistic: $z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\,(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ where

   $$\hat{p}_1 = \frac{X_1}{n_1},\ \hat{p}_2 = \frac{X_2}{n_2}, \text{ and } \hat{p} = \frac{X_1 + X_2}{n_1 + n_2}.$$
3. Use the Normal table to compute the P-value by finding the area: $P(Z < z)$
4. Reject H_0 if the P-value $\leq \alpha$

The book's guidelines recommend that this large-sample test should only be used when the number of successes and the number of failures are both at least 5.

8.2 Comparing Two Proportions

One-sided (right-sided) test for equality of population proportions when n_1 and n_2 are large

1. Test $H_0 : p_1 = p_2$ vs. $H_a : p_1 > p_2$
2. Test statistic: $z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\,(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$ where

   $\hat{p}_1 = \dfrac{X_1}{n_1}$, $\hat{p}_2 = \dfrac{X_2}{n_2}$, and $\hat{p} = \dfrac{X_1 + X_2}{n_1 + n_2}$.
3. Use the Normal table to compute the P-value by finding the area: $P(Z > z)$
4. Reject H_0 if the P-value $\leq \alpha$

The book's guidelines recommend that this large-sample test should only be used when the number of successes and the number of failures are both at least 5.

8.2 Comparing Two Proportions

- **Example 8E**: An experiment is performed to compare the effectiveness of a new drug to a current drug typically used to treat headaches. A randomized controlled double-blind design is used in which 100 subjects who have headaches receive the new drug and 100 subjects who have headaches receive the standard drug. Among the subjects who receive the new drug, 75 experience relief within 2 hours. Among the subjects who receive the standard drug, 60 experience relief within 2 hours. Is this sufficient evidence to conclude that the new drug is more effective in quickly providing some relief to subjects with headaches? Perform an appropriate significance test at level $\alpha = .01$.

8.2 Comparing Two Proportions

- *Answer*: