Chapter 19

Inference about a Population Proportion

Proportions

- The proportion of a population that has some outcome ("**success**") is *p*.
- The proportion of successes in a sample is measured by the **sample proportion**:

$$\hat{p} = \frac{\text{number of successes in the sample}}{\text{total number of observations in the sample}}$$

"p-hat"

Inference about a Proportion
Simple Conditions

SAMPLING DISTRIBUTION OF A SAMPLE PROPORTION

Choose an SRS of size n from a large population that contains population proportion p of "successes." Let $\hat{p}$ be the **sample proportion** of successes,

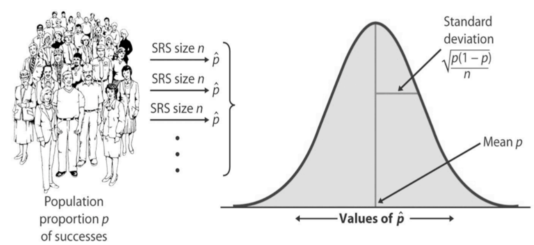
$$\hat{p} = \frac{\text{count of successes in the sample}}{n}$$

Then:

- As the sample size increases, the sampling distribution of $\hat{p}$ becomes **approximately Normal.**
- The **mean** of the sampling distribution is p.
- The **standard deviation** of the sampling distribution is

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

Inference about a Proportion
Sampling Distribution



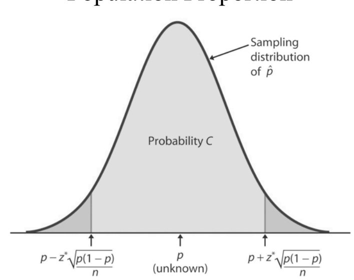
Standardized Sample Proportion

- Inference about a population proportion *p* is based on the *z* statistic that results from standardizing $\hat{p}$:

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

  – *z* has approximately the standard normal distribution as long as the sample is not too small ***and*** the sample is not a large part of the entire population.

Building a Confidence Interval
Population Proportion



Standard Error

Since the population proportion p is unknown, the standard deviation of the sample proportion will need to be estimated by substituting $\hat{p}$ for p.

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

Confidence Interval

LARGE-SAMPLE CONFIDENCE INTERVAL FOR A POPULATION PROPORTION

Draw an SRS of size n from a population with unknown proportion p of successes. An approximate level C confidence interval for p is

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

where z^* is the critical value for the standard Normal density curve with area C between $-z^*$ and z^*.

Use this interval only when the counts of successes and failures in the sample are both at least 15.

Case Study: Soft Drinks

A certain soft drink bottler wants to estimate the proportion of its customers that drink another brand of soft drink on a regular basis. A random sample of 100 customers yielded 18 who did in fact drink another brand of soft drink on a regular basis. Compute a 95% confidence interval ($z^* = 1.960$) to estimate the proportion of interest.

Case Study: Soft Drinks

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \frac{18}{100} \pm 1.960\sqrt{\frac{\frac{18}{100}\left(1-\frac{18}{100}\right)}{100}}$$

$$= 0.18 \pm 0.075$$

$$= 0.105 \quad \text{to} \quad 0.255$$

We are 95% confident that between 10.5% and 25.5% of the soft drink bottler's customers drink another brand of soft drink on a regular basis.

Adjustment to Confidence Interval

More Accurate Confidence Intervals for a Proportion

- The standard confidence interval approach yields unstable or erratic inferences.
- By adding four imaginary observations (two successes & two failures), the inferences can be stabilized.
- This leads to more accurate inference of a population proportion.

Adjustment to Confidence Interval

More Accurate Confidence Intervals for a Proportion

PLUS FOUR CONFIDENCE INTERVAL FOR A PROPORTION

Choose an SRS of size n from a large population that contains population proportion p of "successes." The **plus four estimate** of p is

$$\tilde{p} = \frac{\text{count of successes in the sample} + 2}{n+4}$$

An approximate level C confidence interval for p is

$$\tilde{p} \pm z^* \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}}$$

where z^* is the critical value for the standard Normal density curve with area C between $-z^*$ and z^*.

Use this interval when C is at least 90% and the sample size n is at least 10.

Case Study: Soft Drinks

"Plus Four" Confidence Interval

$$\tilde{p} = \frac{18+2}{100+4} = \frac{20}{104}$$

$$\tilde{p} \pm z^* \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}} = \frac{20}{104} \pm 1.960\sqrt{\frac{\frac{20}{104}\left(1-\frac{20}{104}\right)}{104}}$$

$$= 0.192 \pm 0.076$$

$$= 0.120 \text{ to } 0.272$$

We are 95% confident that between 12.0% and 27.2% of the soft drink bottler's customers drink another brand of soft drink on a regular basis. *(This is more accurate.)*

BPS - 5th Ed. Chapter 19 13

Choosing the Sample Size

SAMPLE SIZE FOR DESIRED MARGIN OF ERROR

The level C confidence interval for a population proportion p will have margin of error approximately equal to a specified value m when the sample size is

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

where p^* is a guessed value for the sample proportion. The margin of error will be less than or equal to m if you take the guess p^* to be 0.5.

Use this procedure even if you plan to use the "plus four" method.

BPS - 5th Ed. Chapter 19 14

Case Study: Soft Drinks

Suppose a certain soft drink bottler wants to estimate the proportion of its customers that drink another brand of soft drink on a regular basis using a 99% confidence interval, and we are instructed to do so such that the margin of error does not exceed 1 percent (0.01).

What sample size will be required to enable us to create such an interval?

BPS - 5th Ed. Chapter 19 15

Case Study: Soft Drinks

Since no preliminary results exist, use $p^* = 0.5$.

$$n = \left(\frac{z^*}{m}\right)^2 p^*(1-p^*) = \left(\frac{2.576}{0.01}\right)^2 (0.5)(1-0.5) = 16589.44$$

Thus, we will need to sample at least 16589.44 of the soft drink bottler's customers.

Note that since we cannot sample a fraction of an individual and using 16589 customers will yield a margin of error slightly more than 1% (0.01), our sample size should be $n = 16590$ customers .

BPS - 5th Ed. Chapter 19 16

The Hypotheses for Proportions

- Null: H_0: $p = p_0$
- One sided alternatives
  - H_a: $p > p_0$
  - H_a: $p < p_0$
- Two sided alternative
  - H_a: $p \neq p_0$

BPS - 5th Ed. Chapter 19 17

Test Statistic for Proportions

- Start with the z statistic that results from standardizing $\hat{p}$:

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

- Assuming that the null hypothesis is true (H_0: $p = p_0$), we use p_0 in the place of p:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

BPS - 5th Ed. Chapter 19 18

P-value for Testing Proportions

- H_a: $p > p_0$
  - *P*-value is the probability of getting a value as large or larger than the observed test statistic (*z*) value.
- H_a: $p < p_0$
  - *P*-value is the probability of getting a value as small or smaller than the observed test statistic (*z*) value.
- H_a: $p \neq p_0$
  - *P*-value is *two times* the probability of getting a value as large or larger than the absolute value of the observed test statistic (*z*) value.

SIGNIFICANCE TESTS FOR A PROPORTION

To test the hypothesis H_0: $p = p_0$, compute the z statistic

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}$$

In terms of a variable Z having the standard Normal distribution, the approximate P-value for a test of H_0 against

H_a: $p > p_0$ is $P(Z \geq z)$

H_a: $p < p_0$ is $P(Z \leq z)$

H_a: $p \neq p_0$ is $2P(Z \geq |z|)$

Use this test when the sample size n is so large that both np_0 and $n(1 - p_0)$ are 10 or more.

Case Study

Parental Discipline

Brown, C. S., (1994) "To spank or not to spank." *USA Weekend*, April 22-24, pp. 4-7.

What are parents' attitudes and practices on discipline?

Case Study: Discipline

Scenario

- Nationwide random telephone survey of 1,250 adults that covered many topics
- 474 respondents had children under 18 living at home
  - results on parental discipline are based on the smaller sample
- reported margin of error
  - 5% for this smaller sample

Case Study: Discipline

Reported Results

"The 1994 survey marks the first time a majority of parents reported *not* having physically disciplined their children in the previous year. Figures over the past six years show a steady decline in physical punishment, from a peak of 64 percent in 1988."

- The 1994 sample proportion who ***did not*** spank or hit was 51% !
- *Is this evidence that a majority of the population did not spank or hit? (Perform a test of significance.)*

Case Study: Discipline

The Hypotheses

- Null: The proportion of parents who physically disciplined their children in 1993 is the same as the proportion [*p*] of parents who *did not* physically discipline their children. [H_0: $p = 0.50$]
- Alt: A majority (more than 50%) of parents *did not* physically discipline their children in 1993. [H_a: $p > 0.50$]

Case Study: Discipline
Test Statistic

Based on the sample

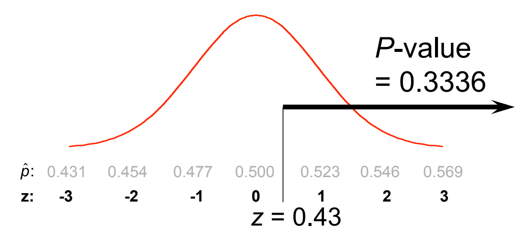
- n = 474 (large, so proportions follow Normal distribution)
- no physical discipline: 51%
  - $\hat{p} = 0.51$
  - standard error of p-hat: $\sqrt{.50(1-.50)/474} = 0.023$

    (where ***.50*** is p_0 from the null hypothesis)
- standardized score (test statistic)

  $z = (0.51 - 0.50) / 0.023 = 0.43$

BPS - 5th Ed. Chapter 19 25

Case Study: Discipline
P-value



From Table A, z = 0.43 is the 66.64th percentile.

BPS - 5th Ed. Chapter 19 26

Case Study: Discipline

1. ***Hypotheses:*** H_0: p = 0.50
   H_a: p > 0.50
2. ***Test Statistic:***

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.51 - 0.50}{\sqrt{\frac{(0.50)(1-0.50)}{474}}} = \frac{0.01}{0.023} = 0.43$$

3. ***P-value:*** P-value = $P(Z > 0.43) = 1 - 0.6664 = 0.3336$
4. ***Conclusion:***
   Since the *P*-value is larger than α = 0.10, there is no strong evidence that a majority of parents did not physically discipline their children during 1993.

BPS - 5th Ed. Chapter 19 27