Chapter 20. Inference About a Population Proportion

The Sample Proportion \hat{p}

Definition. The statistic that estimates the parameter p, a proportion of a population that has some property, is the **sample proportion**

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

The sample proportion \hat{p} is called "*p*-hat."

S.20.1. Proportioned Stooges.

A random sample of size 20 of *Three Stooges* films includes 9 films which have Curly in the role of third stooge. What is the statistic which this sample yields and what parameter of the population does this estimate?

The Sampling Distribution of \hat{p}

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

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

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

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

Note. The condition for inference about a proportion include:

- We can regard our data as a **simple random sample** (SRS) from the population. This is, as usual, the most important condition.
- The sample size n is large enough to ensure that the distribution of \hat{p} is close to normal. We will see that

different inference procedures require different answers to the question "how large is large enough?"

Example. Exercise 20.3 page 495.

Large-Sample Confidence Intervals for a Proportion

Note. The "large-sample confidence interval for a population proportion" test consists of the following steps:

Draw a SRS of size n from a large population that contains an 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 numbers of successes and failures in the sample are both at least 15.

Example. Exercise 20.31 page 509.

Accurate Confidence Intervals for a Proportion

Note. The "plus four confidence interval for a proportion" consists of the following steps:

Draw a SRS of size n from a large population that contains an unknown proportion p of successes. To get the **plus four confidence interval for** p, add four imaginary observations, two successes and two failures. Then use the largesample confidence interval with the new sample size (n + 4)and count of successes (actual count +2). Use the interval when the confidence level is at least 90% and the sample size n is at least 10. (Dr. Bob says: "Hmmm...")

Choosing the Sample Size

Note. 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.

Example S.20.2. Curly Proportions.

A stoogeologist wants to know the proportion of Three Stooges fans who claim that their favorite "third stooge" is Curly. She wants to be C = 95% confident that the margin of error of her estimate is within m = 3% of the population proportion. How large a sample must she take?

Solution. We have C = 95% and m = 3% = 0.03. The critical value that corresponds to C = 95% = 0.95 is $z^* = 1.96$. To be conservative, we take $p^* = 0.5$ and we get from the above formula:

$$n = \left(\frac{z^*}{m}\right)^2 p^*(1-p^*) = \left(\frac{1.96}{0.03}\right)^2 (0.5)(1-0.5) = 1067.1.$$

We round up to get that the desired sample should be size n = 1068.

Significance Tests for a Proportion

Note. The significance test for a proportion consists of the following steps:

Draw a SRS of size n from a large population that contains an unknown proportion p of successes. 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 O-value for a test of H_0 against

$$H_a: p > p_0 \text{ is } P(Z \ge z)$$

$$H_a: p < p_0 \text{ is } P(Z \le z)$$

$$H_a: p \neq p_0 \text{ is } 2P(Z \ge |z|)$$

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

Example S.20.3. Testing Shemp.

The president of the Three Stooges Fan Club thinks that Shemp is underappreciated and that at least 20% of Three Stooges fans would consider Shemp the funniest stooge. He sends a questionnaire to a random sample of 100 of the members of the Fan Club and all of the questionnaires are returned. Only 15 of the fans replied that Shemp was their favorite Stooge. Perform the relevant hypothesis test for the president.

Solution. We have $p_0 = 0.20$, n = 100, and $\hat{p} = 15/100 = 0.15$. We compute the z statistic:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}} = \frac{0.15 - 0.20}{\sqrt{\frac{0.20(1 - 0.20)}{100}}} = -1.25.$$

We have $H_0: p = p_0 = 0.20$ and $H_a: p \ge 0.2$. From Table A we find that $P(Z \ge z) = P(Z \ge -1.25) = 1 - 0.1056 = 0.8944$. We would reject the null hypothesis if $P(Z \ge z)$ were near 0. It is not, so we fail to reject the null hypothesis and the sample is not supportive of H_a and so not supportive of the Fan Club president's opinion.

Example. Exercise 20.41 page 511.