7.1 Inference for a Population Proportion

Definition. The statistic that estimates the parameter p is the *sample proportion*

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

Assumptions for Inference

Note. Standardize $\hat{p}$ by subtracting its mean and dividing by its standard deviation. The result is a z statistic:

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

The statistic z has approximately the standard normal distribution $N(0,1)$. Inference about p uses this z statistic and standard normal critical values. See Figure 7.1 (and TM-108).

Note. We need to deal with the fact that we don't know the standard deviation $\sqrt{p(1-p)/n}$ because we don't know p. Here's what to do:

- To test the null hypothesis $H_0 : p = p_0$ that the unknown p has a specific value p_0, just replace p by p_0 in the z statistic.
- In a confidence interval for p, we have no specific value to substitute. In large samples, $\hat{p}$ will be close to p, so replace the standard deviation by the *standard error of p*

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

The confidence interval has the form

$$\text{estimate } \pm z^* SE_{\text{estimate}}.$$

Note. Some assumptions for inference about a proportion are:

- The data are an SRS from the population of interest.
- The population is at least 10 times as large as the sample.
- For a test of $H_0 : p = p_0$, the sample size n is so large that both np_0 and $n(1-p_0)$ are 10 or more. For a confidence interval, n is so large that both the count of successes $n\hat{p}$ and the count of failures $n(1-\hat{p})$ are 10 or more.

Example 7.4. We want to use the National AIDS Behavioral Surveys data to give a confidence interval for the proportion of adult heterosexuals who have had multiple sexual partners. Does the sample meet the requirements for inference?

- The sampling design was in fact a complex stratified sample, and the survey used inference procedures for that design. The overall effect is to close to an SRS, however.
- The number of adult heterosexuals (the population) is much larger than 10 times the sample size, $n = 2673$.
- The counts of "Yes" and "No" responses are much greater than 10:

$$n\hat{p} = (2673)(.0636) = 170$$

$$n(1-\hat{p}) = (2673)(.9364) = 2503.$$

The second and third requirements are easily met. The first requirement, that the sample be an SRS, is only approximately met.

The z Procedures

Note. To perform a large-sample inference for a population proportion, do the following. Draw an SRS of size n from a large 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 upper $(1-C)/2$ standard normal critical value. 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 H_0 against

$$H_a : p > p_0 \quad \text{is} \quad P(Z \geq z)$$
$$H_a : p < p_0 \quad \text{is} \quad P(Z \leq z)$$
$$H_a : p \neq p_0 \quad \text{is} \quad P(Z \geq |z|)$$

Example 7.6. A coin that is balanced should come up heads half the time in the long run. The population for coin tossing contains the results of tossing the coin forever. The parameter p is the probability of a head, which is the proportion of all tosses that give a head. The tosses we actually make are an SRS from this population. The French naturalist Count Buffon (1707 - 1788) tossed a coin 4040 times. He got 2048 heads. The sample proportion of heads is

$$\hat{p} = \frac{2048}{4040} = .5069.$$

That's a bit more than one-half. Is this evidence that Buffon's coin was not balanced? This is a job for a significance test.

Step 1: Hypotheses. The null hypothesis says that the coin is balanced ($p = .5$). The alternative hypothesis is two-sided, because we did not suspect before seeing the data that the coin favored either heads or tails. We therefore test the hypotheses

$$H_0 : p = .5$$

$$H_a : p \neq .5.$$

The null hypothesis gives the value $p_0 = .5$.

Step 2: Test Statistic. The z test statistic is

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{.5069 - .5}{\frac{(.5)(.5)}{4040}} = .88.$$

Step 3: $P-$Value. Because the test is two-sided, the $P-$value is the area under the standard normal curve more than 0.88 away from 0 in either direction. Figure 7.2 (and TM-109) shows this area. From Table A (TM-139, TM-140) we find that the area below -.088 is 0.1894. The $P-$value is twice this area: $P = 2(.1894) = .3788$.

Conclusion. A proportion of heads as far from one-half as Buffon's would happen 38% of the time when a balanced coin is tossed 4040 times. Buffon's result doesn't show that his coin is unbalanced.

Note. In Example 7.6, we failed to find good evidence against $H_0 : p = .5$. We cannot conclude that H_0 is true, that is that the coin is perfectly balanced. No doubt p is not exactly 0.5. The test of significance only shows that the results of Buffon's 4040 tosses can't distinguish this coin from one that is perfectly balanced. To see what values of p are consistent with the sample results, use a confidence interval.

Example 7.7. The 95% confidence interval for the probability p that Buffon's coin gives a head is

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = .5069 \pm 1.960 \sqrt{\frac{(.5069)(.4931)}{4040}} = (.4915, .5223).$$

We are 95% confident that the probability of a head is between 0.4915 and 0.5223.

Choosing the Sample Size

Note. The level C confidence interval for a population proportion p will have a 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 7.8. Gloria Chavez and Ronald Flynn are candidates for mayor in a large city. You are planning a sample survey to determine what percent of the voters plan to vote for Chavez. This is a population proportion p. You will contact an SRS of registered voters in the city. You want to estimate p with 95% confidence and a margin of error no greater than 3%, or 0.03. How large a sample do you need? The winner's share in all but the most lopsided elections is between 30% and 70% of the vote. So use the guess $p^* = .5$. The sample size you need is

$$n \left(\frac{1.96}{.03}\right)^2 (.5)(1 - .5) = 1.067.1.$$

You should round the result up to $n = 1068$. If you want a 2.5% margin of error, we have (after rounding)

$$n = \left(\frac{1.96}{.025}\right)^2 (.5)(1 - .5) = 1537.$$

For a 2% margin of error the sample size you need is

$$n = \left(\frac{1.96}{.02}\right)^2 (.5)(1 - .5) = 2401.$$

As usual, smaller margins of error call for larger samples.