9: Inference about a Proportion

Binary response

The past series of chapters have focused on quantitative outcomes. This chapter addresses categorical outcomes with two possible values (“binary variables”). For example, classifying someone as a smoker or non-smokers is a binary variable.

Whereas quantitative variable were summarized with sums and averages, categorical variables are summarized with counts and proportions.

The symbol \( \hat{p} \) (“p-hat”) is used to represent the sample proportion:

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

Illustrative example: Smoking survey. We select a SRS of 57 individuals. The sample has 17 smokers. Therefore, the sample proportion is \( \hat{p} = 17 / 57 = 0.298 \), or 29.8%. The goal of this chapter is to use this information to infer the proportion of people in the population who smoke.

Notes:

1. Proportions are a type of average in which “successes” are given a value of 1 and “failures” are given a value 0. For example, if we have 10 observations as follows \{0, 0, 0, 1, 0, 0, 0, 1, 0\}, \( n = 10 \), \( \sum x_i = 2 \), and sample mean \( \bar{x} = \frac{2}{10} = \hat{p} \).

   Principles applied in using \( \bar{x} \) to infer population mean \( \mu \) transfer to using sample proportion \( \hat{p} \) in inferring population proportion \( p \).

2. Sample proportions are used to estimate population prevalences and incidences. Prevalence \( \equiv \) the proportion in a cross-sectional sample and cumulative incidence (“risk”) \( \equiv \) the proportion of susceptible in a cohort who develop a condition over a fixed period of time.
Inferring population proportion $p$ (Normal approximation)

Let $p$ represent the proportion in the population. Sample proportion $\hat{p}$ is an unbiased estimator of parameter $p$. Keep in mind that sample proportion $\hat{p}$ in any given sample will not be an exact replica of population proportion $p$; some of the $\hat{p}$ s will be less than $p$, and some will be more. That is the nature of sampling. Over the long run, with repeated independent samples, $\hat{p}$ is an unbiased estimator of $p$.

Inferences about parameter $p$ rest on binomial distributions (Chapter 4). Binomial probabilities can be tedious to calculates so, when $n$ is large, a Normal approximation to the binomial is used. The Normal approximation to the binomial says that the number of successes in a sample will have a Normal distribution with $\mu = np$ with standard deviation $\sigma = \sqrt{npq}$ where $q = 1 - p$. Equivalent, when $n$ is large, the sample proportion $\hat{p}$ will vary according to a Normal distribution with expected value $p$ and standard error

$SE_{\hat{p}} = \sqrt{\frac{pq}{n}}$.

Here is the binomial sampling model of the number of successes for a binomial random variable $X$ with $n = 57$ and $p = 0.25$:

This distribution is nearly Normal. The random number of success $X \sim N(14.25, 3.27)$. It in addition, the sampling distribution of the proportion $\hat{p} \sim N(0.25, 0.0574)$. This is pretty advances stuff, but for now please note these Normal approximation hold when $npq \geq 5$ (so-called $npq$ rule). For the model above, $n = 57$ and $p = .25$, so $npq = (57)(0.25)(1-0.25) = 10.6875$. Since this exceeds 5, we can predict that the Normal approximation to the binomial can be trusted.
Confidence interval for $p$

A method called the “plus-four” method is used to calculate the confidence interval of $p$. This method is a modification of the standard Normal method, but is much more reliable, especially when $n$ is small, providing reliable results even when $n$ is as small as 10.

The general idea is to add two “successes” and two “failures” to the data before calculating the confidence interval. Then, the typical “estimate $\pm z \cdot$ standard error” formula is applied. Let $\tilde{x} \equiv$ the observed number of success plus two $= x + 2$, $\tilde{n} \equiv$ the sample size plus four $= n + 4$, and $\tilde{p} = \frac{\tilde{x}}{\tilde{n}}$. The $(1-\alpha)100\%$ confidence interval for $p$ is

\[
\tilde{p} + z_{1-\alpha} \cdot se_{\tilde{p}}
\]

where $se_{\tilde{p}} = \sqrt{\frac{\tilde{p}\tilde{q}}{\tilde{n}}}$.

Use $z = 1.645$ for 90% confidence, $z = 1.96$ for 95% confidence, and $z = 2.576$ for 99% confidence.

**Illustrative example: confidence interval for proportion $p$.** In the smoking prevalence illustrative example $n = 57$ and $x = 17$. What is the 95% confidence interval for population prevalence $p$?

\[
\begin{align*}
\tilde{x} &= x + 2 = 17 + 2 = 19 \\
\tilde{n} &= n + 4 = 57 + 4 = 61 \\
\tilde{p} &= \frac{\tilde{x}}{\tilde{n}} = \frac{19}{61} = 0.3115 \\
\tilde{q} &= 1 - \tilde{p} = 1 - 0.3115 = 0.6885 \\
se_{\tilde{p}} &= \sqrt{\frac{(0.3115)(0.6885)}{61}} = 0.0593
\end{align*}
\]

The 95% confidence interval for $p$

\[
= 0.3115 \pm (1.96)(0.0593)
\]

\[
= 0.3115 \pm 0.1162
\]

\[
= 0.1953 \text{ to } 0.4277 \text{ or between } 20\% \text{ and } 43\%.
\]
Sample Size Requirements to Limit Margin of Error

In planning a study, we want to collect enough data to estimate population proportion \( p \) with adequate precision. In an earlier chapter we had determined the sample size to determine population mean \( \mu \) with margin of error \( d \). We apply a similar method in determining sample size requirements to estimate population proportion \( p \).

Let \( d \) represent the margin of error. This provides the “wiggle room” around \( \hat{p} \); it is half the confidence interval width. To achieve margin of error \( d \) use

\[
n = \frac{z_{\frac{1-\alpha}{2}}^2 \hat{p} q^*}{d^2}
\]

where \( p^* \) represent the an educated guess for the proportion and \( q^* = 1 - p^* \).

When no reasonable guess of \( p \) is available, use \( p^* = 0.50 \) to provide a “worst-case scenario” sample size (i.e., more than enough data).

**Illustrative example:** *Smoking survey, sample size requirements for confidence interval.* Recall the “Smoking survey” illustrative example presented earlier in the chapter. We want to re-sample the population and calculate a 95% confidence interval with greater precision. How large a sample is needed to shrink the margin of error in the “Smoking survey illustrative data” to 0.05? How large a sample is needed to shrink the margin of error to 0.03? The prior sample had \( \hat{p} = 0.30 \), so let’s use this for \( p^* \).

**Solutions:**

To achieve a margin of error of 0.05,

\[
 n = \frac{z_{\frac{1-0.95}{2}}^2 \hat{p} q^*}{d^2} = \frac{1.96^2 \cdot 0.30 \cdot 0.70}{0.05^2} = 322.7. \text{ Round this up to 323 to ensure adequate precision.}
\]

To achieve a margin of error of 0.03,

\[
 n = \frac{1.96^2 \cdot 0.30 \cdot 0.70}{0.03^2} = 896.4, \text{ so use 897 individuals.}
\]

The increased precision has the price of a larger sample size.
Hypothesis test (Normal approximation)

Let $p_0$ denote the value of population proportion $p$ under the null hypothesis. Before beginning the test, check to see whether a Normal approximation can be used by checking whether $np_0q_0 \geq 5$, where $q_0 = 1 - p_0$. If $np_0q_0 < 5$, an “exact” binomial test is required. We do not cover the exact binomial test.

(A) **Hypotheses:** The null hypothesis is $H_0: p = p_0$, where $p$ represents the population proportion and $p_0$ is its expectation under the null hypothesis. The alternative hypothesis is either $H_1: p \neq p_0$ (two-sided), $H_1: p < p_0$ (one-sided to the left), or $H_1: p > p_0$ (one-sided to the right).

(B) **Test statistic:** The test statistic is $Z_{stat} = \frac{\hat{p} - p_0}{SE_{\hat{p}}}$ where $\hat{p}$ represents the sample proportion, $p_0$ = the null value, and $SE_{\hat{p}} = \sqrt{\frac{p_0q_0}{n}}$.

(C) **$P$-value:** The $z_{stat}$ is converted to a $P$-value in the usual fashion. Small $P$-values provide strong evidence against $H_0$.

(D) **Significance statement (optional).** Reject $H_0$ when $P \leq \alpha$. in which case the difference is said to be significant.

**Illustrative example.** The prevalence of smoking in U. S. adults is approximately 25% (NCHS, 1995, Table 65). We observe 17 smokers in 57 individuals. Therefore, $\hat{p} = 29.8\%$. Does this provide significant evidence that the population from which the sample was drawn has a prevalence that exceeds the national average? Let’s do a two-sided test.

Under the null hypothesis $p_0 = 0.25$. Before conducting the test we check whether the Normal approximation to the binomial holds by calculating $np_0q_0 = (57)(.25)(1-.25) = 10.7$. We can proceed with a Normal approximation test.

(A) $H_0: p = .25$ versus $H_1: p \neq .25$.

(B) $SE_{\hat{p}} = \sqrt{\frac{(.25)(1-.25)}{57}} = .0574$ and $z_{stat} = \frac{.298-.25}{.0574} \approx 0.84$.

(C) $P = 0.4010$. This does not provide strong evidence against $H_0$.

(D) $P > \alpha$; $H_0$ is retained. The difference is not significant.