9: Inference About a Proportion

Introduction | Confidence Interval for *p* | Statistical Hypothesis Test

Introduction

Binary Data and Proportions

We consider categorical variables with only two possible values. Such data are referred to as binary data, Bernoulli variables, dichotomous data, and "0/1 data."

With continuous data we summarize data with sums and averages. With binary data, we rely on counts and proportions:

Illustrative example. A survey is conducted to learn about the interrelation of factors affecting smoking in teenagers. One objective of the survey is to determine the prevalence of teenage smoking in a particular population. Seventeen (17) of the 57 teenagers in the sample are smokers.

Let: p represent the proportion in the population (this is the same as binomial parameter p, the probability a person selected at random has the attribute being studied),

x represent the number of observations in the sample positive for the attribute, and

$\hat{p}$ represent the sample proportion, which is:

$$\hat{p} = \frac{x}{n} \tag{9.1}$$

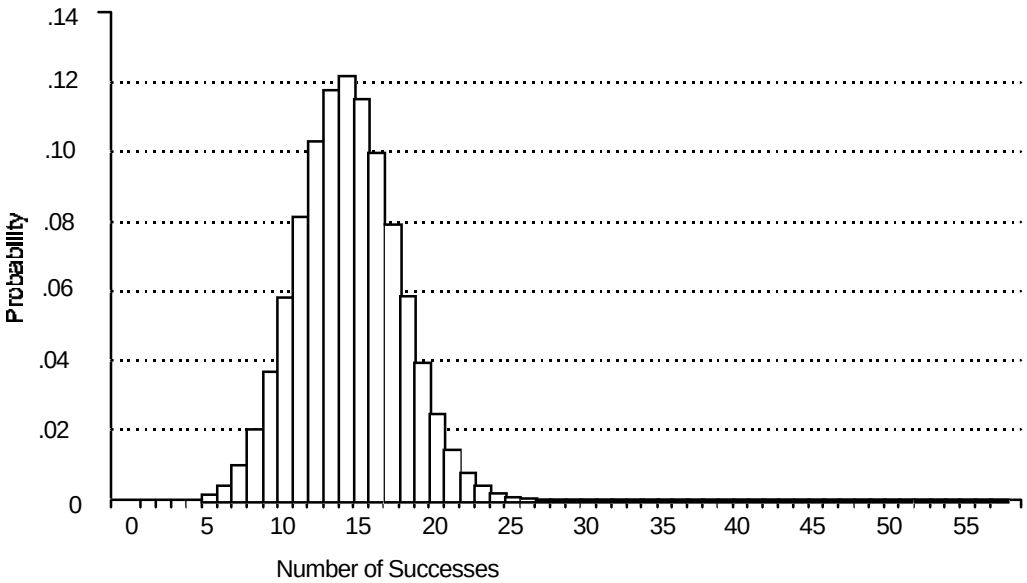
For the illustrative example, $\hat{p} = \frac{17}{57} = .298$.

What *inferences* can be made about the population from which the sample was taken? One inference is that the the sample proportion is the *point estimate* of the prevalence of smoking in this teenage population. This proportion when expressed as a percentage may be looked upon as an average of the number of teenagers per hundred who smoke. That is, from this survey it is inferred that the average number of smokers in all possible samples of 100 is 29.8; the prevalence of smoking in the population is 29.8%.

We also want an interval estimate of the prevalence of smoking in the population. That is, we want to know an interval within which the true prevalence is contained. This interval estimate is called the *confidence interval*. Confidence intervals for population proportions can be calculated in several ways. One methods takes advantage of the normal approximation to the binomial, which is described below.

Normal Approximation to the Binomial

We want to estimate parameter p with known degree of certainty. To do so, we rely on the binomial distribution. Recall from Chapter 4 that calculating binomial probabilities can be tedious. Fortunately, when n is large, a binomial distribution takes on properties of the normal distribution. For instance, a binomial distribution with $n = 57$ and $p = .25$ is:



This distribution is bell-shaped (approximately normal) with mean (μ) = np = (57)(.25) = 14.25 [by formula 4.1] and variance (σ^2) = npq = (57)(.25)(1–.25) = 10.6875 [by formula 4.2]. Thus, the number of observations positive for the attribute is normally distributed with a mean of 14.25 and variance of 10.6875.

The "npq rule." The normal approximation to the binomial is accurate when $npq \geq 5$, where n represents the sample size, p represents the population proportion, and $q = 1 - p$. We may substitute $\hat{p}$ for p and $\hat{q}$ for q when applying this rule based on empirical data.

Illustrative Example: When $n = 57$ and $x = 17$, $\hat{p} = .298$, and $n\hat{p}\hat{q}$ = (57)(.298)(1–.298) = 11.9. Therefore, the normal approximation can be used with these data.

Confidence Interval for *p*

When the normal approximation to the binomial holds (i.e., when $npq \geq 5$), a $(1-\alpha)100\%$ confidence given by:

$$\hat{p} \pm (z_{1-a/2})(se_{\hat{p}}) \qquad \textbf{(9.2)}$$

where

$z_{1-\alpha/2}$ represents the $1-\alpha/2$ percentile on a z distribution (e.g., for a 95% confidence interval use $z_{1-.05/2} = z_{.975} = 1.96$)

$se_{\hat{p}}$ represents the standard error of the proportion given by: $se_{\hat{p}} = \sqrt{\frac{\hat{p}\hat{q}}{n}}$

Illustrative example (95% confidence interval). For the illustrative data, $se_{\hat{p}} = \sqrt{\frac{(.298)(1-.298)}{57}} = .0606$

and a 95% confidence interval for $p = .298 \pm (1.96)(.0606) = .298 \pm .119 = (.179, .417) \cong (18\%, 42\%)$. Thus, we interpret are 95% confident that parameter p lies between 18% and 42%. Another good way to think about this is "the proportion is about 30% with a margin of error of ±12% at 95% confidence.

Illustrative example (90% confidence interval). For a 90% confidence interval, use $\alpha = .10$. Thus, the 90% confidence interval for $p = \hat{p} \pm (z_{1-.10/2})(se_{\hat{p}}) = 298 \pm (1.64)(.0606) = .298 \pm .100 = (.198, .398) \cong$ (20%, 40%). Thus, the sample proportion is 30% with a margin of error of ±10% at 90% confidence.

Sample size requirements to achieve given precision. To estimate a proportion with 95% confidence so that the margin of error is no greater than d, the size of the sample should be:

$$n = \frac{(1.96)^2\, pq}{d^2} \qquad \textbf{(9.3)}$$

For example, to achieve a margin of error (d) of .05 for a proportion that is expected to be about 0.25, use $n = (1.96^2)(0.25)(1 - 0.25)/(0.05)^2 \cong 288$.

If you do not have a reasonable guestimate of the proportion you are trying to estimate, assume $p = 0.50$.* For example, to estimate a proportion with margin of error no greater than .03 where no reasonable estimate for p is available, use $n = (1.96^2)(.50)(.50) / (.03^2) \cong 1067$.

* This will maximize the sample size calculation and thus ensure sufficient precision.

Statistical Hypothesis Test

We want to test whether an observed proportion is different than a hypothetical expectation. Let p_0 represent the proportion parameter under the null hypothesis (the "null value"). We want to test H_0: $p = p_0$. The alternative hypothesis may be left-tailed (H_1: $p < p_0$), right-tailed (H_1: $p > p_0$), or two-tailed (H_1: $p \neq p_0$).

Illustrative example: We want to test whether the prevalence of smoking in a community is different than that of the United States as a whole. The prevalence of smoking in US is .25 (NCHS,1995, table 65). Thus, $p_0 = .25$ and H_0: $p = .25$. The illustrative test will be two-sided. Therefore, H_1: $p \neq .25$. Let $\alpha = .05$.

Before conducting this test, we check to see whether the normal approximation to the binomial can be used for testing purpose. For the illustrative example, $n = 57$, and $p_0 = .25$. Thus, $np_0q_0 = (57)(.25)(1 - .25) = 10.7$ and the normal approximation to the binomial can be used. The test statistic (based on the normal approximation) is:

$$z_{\text{stat}} = \frac{\hat{p} - p_0}{SE_{\hat{p}}} \qquad \textbf{(9.4)}$$

where

$\hat{p}$ represents the sample proportion,

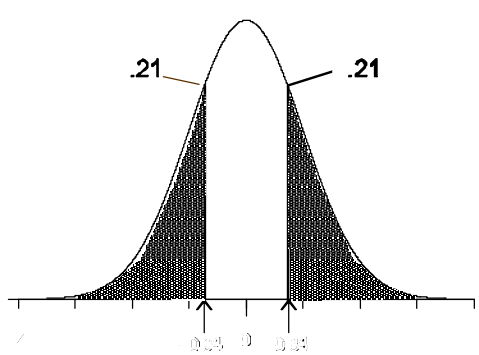
p_0 represents the value of the parameter under the null hypothesis (the null value), and

the standard error, $SE_{\hat{p}} = \sqrt{\dfrac{p_0 q_0}{n}}$.

For the illustrative data, $SE_{\hat{p}} = \dfrac{(.25)(1-.25)}{57} = .0574$ and $z_{stat} = \dfrac{.298-.25}{.0574} = 0.84$.

The two-tailed p value is twice the area under the normal curve to the right of 0.84. Thus, $p = 2 \times .21 = .42$ (see figure at bottom of page).

SPSS: Click `Statistics > Nonparametric Tests > Binomial`. Fill in the "Test Variable Dialogue Box" with the variable name containing the data and provided the null value in the box labeled "Test value."



OPTIONAL: EXACT BINOMIAL TEST

When the normal approximation cannot be used (i.e., when $npq < 5$), the number of successes in a given sample (X) follows a binomial distribution: X~b(n, p) and the exact p value is the binomial probability of observing at least X positives given n and p_0. Thus,

$$p \text{ value} = \Pr(X \geq x | X \sim b(n, p_0)) \qquad \textbf{(9.5)}$$

Illustrative example: Suppose we want to test whether there is a preference for a given procedure. Under the null hypothesis, half the patients will prefer procedure A and half will prefer procedure B. Let p represent the proportion preferring procedure A. Thus, under the null hypothesis (H_0): $p = 0.5$.

Suppose we observe 7 out of 8 patients prefer procedure A. Under H_0, the expected value of X is 4 (since we expect half the patients to choose A). Thus, the p value for the test = Pr(X $\geq$ 7 | X~b(n = 8, p = .5)) = Pr(X = 7) + Pr(X = 8). Using the formula 4.4 we calculate

$$\Pr(X = 7) = ({}_8C_7)(.5)^7(1-.5)^{8-7} = (8)(.0078)(.5) = .0313$$
$$\Pr(X = 8) = ({}_8C_8)(.5)^8(1-.5)^{8-8} = (1)(.0039)(1) = .0039$$

Thus, Pr(X $\geq$ 7) = .0313 + .0039 = .0352.