Inference for Proportions

- Inference for a single proportion
  - Sampling distribution of the proportion of successes
  - Wilson's estimate
  - Confidence interval for a population proportion
  - Significance test for a population proportion
  - Choosing an appropriate sample size
- Comparing two proportions

Introduction to Inference for Proportions

The data for many statistical studies are not measurements but are counts.

Example: A simple random sample of 267 college women aged 18 to 25 is chosen for a study of unhealthy eating behavior of college women. 184 individuals, or .689 percent of the sample, reported that they were on a diet sometime during the past year.

We will focus on procedures for statistical inference when the parameters that we want to do inference about are population proportions.

In the previous example, we are interested in the proportion of college women (the population) that were on a diet sometime during the past year. The sample of 267 is used to draw inference about the entire population.

Inference for a single proportion

Suppose we want to estimate the proportion p of some characteristic of a population, and we undertake the following procedure:

1. Draw an SRS of size n.
2. Record the number X of "successes" (those individuals having the characteristic).
3. What is the distribution of X?

   $X \sim B(n, p)$. Recall that the equation for binomial probabilities is

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

   with $E[X] = np$ and standard deviation $\mathrm{SD}(X) = \sqrt{np(1-p)}$,

1. From Chapter 5, we know that for large enough n, the Binomial distribution is very well approximated by the Normal distribution: X is approximately $N\left(\mu_X = np, \sigma_X = \sqrt{np(1-p)}\right)$.
2. What about $\hat{p} = \frac{X}{n}$ (the sample proportion of "successes")? $\hat{p}$ is approximately $N\left(\mu_X = p, \sigma_X = \sqrt{\frac{p(1-p)}{n}}\right)$
3. The above normal approximations can be used when n is sufficiently large – i.e. if $np \geq 10$ and $n(1-p) \geq 10$
4. $\hat{p}$ is an unbiased estimate of the unknown true population proportion p
5. An approximate $(1-\alpha)$ CI for the population proportion p is given by

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

   where z^* is chosen so that $P(Z > z^*) = \alpha/2$ for $Z \sim N(0,1)$.

What if we want to test whether $p = p_0$ for some fixed value p_0?

The null hypothesis is $p = p_0$, and under this hypothesis,

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim N(0,1)$$

Notice that we are using a different value for the SD of $\hat{p}$ than was used for the CI. Since H_0 specifies a true value for p, the SD of $\hat{p}$ under H_0 is given by

$$\sqrt{\frac{p_0(1-p_0)}{n}}$$

The p-values for this test are:

- $H_a : p > p_0$ $\quad P(Z \geq z)$
- $H_a : p < p_0$ $\quad P(Z \leq z)$
- $H_a : p \neq p_0$ $\quad 2P(Z \geq |z|)$

for $Z \sim N(0,1)$.

Example

It is believed that less than half of California lawers feel that the ethical standards of most lawyers are high. A random sample of 2700 California lawyers revealed only 1107 who felt that the ethical standards of most lawyers are high. (*AP, Nov. 12, 1994*).

1. Does this provide strong evidence for concluding that fewer than 50% of all California lawyers feel this way?
2. What is a 90% confidence interval for the true proportion of California lawyers who feel that ethical standards are high?

1. The large-sample Z significance test for a population proportion is recommended as long as the expected number of successes and failures under the null hypothesis, np_0 and $n(1-p_0)$, respectively, are both greater than 10.
2. Note that the standard error used for the confidence interval is estimated from the available data, whereas the denominator in the z statistic for a significance test uses the value given by the null hypothesis. As a results, the correspondence between the significance test and the confidence interval is no longer exact (but it is still very close).
3. Confidence intervals are more informative than significance tests in the context of inference for a single proportion. Confidence intervals allow us to determine values of p that are consistent with the observed results.

Wilson's Estimate for the Population Proportion

What if we observed no success among n trials? Then,

$$\hat{p} = 0 \qquad SE(\hat{p}) = 0$$

That is, we are *certain* that $p = 0$, which is implausible. (It is always possible that $p > 0$, but we just happened not to observe any successes.)

When $\hat{p}$ is close to 0 and 1, the confidence interval based on $\hat{p}$ is not quite accurate. Edwin Bidwell Wilson (1927) suggested the following adjustment, known as **Wilson's estimate**

$$\tilde{p} = \frac{X+2}{n+4}$$

where X is the number of successes among n trials. We simply add two "phantom" successes and two "phantom" failures to the data.

The approximate distribution of $\tilde{p}$ is

$$\tilde{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n+4}}\right)$$

CIs Using Wilson's Estimate

An approximate $(1-\alpha)$ CI for the population proportion p is given by

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

where z^* is chosen such that $P(Z < z^*) = \alpha/2$.

Example (cont.)

Suppose only 13 of the lawyers in our example believed that ethical standards are high. Then,

$$\hat{p} = \frac{13}{2700} = 0.0048$$

and a 90% CI for p is given by

$$0.0048 \pm 1.645\sqrt{\frac{0.0048(1-0.0048)}{2700}} = (0.0026,\ 0.0070)$$

Using Wilson's estimate,

$$\tilde{p} = \frac{13+2}{2700+4} = 0.0055$$

and a 90% CI for p is given by

$$0.0055 \pm 1.645\sqrt{\frac{0.0055(1-0.0055)}{2700+4}} = (0.0032,\ 0.0079)$$

Choosing an Appropriate Sample Size

Suppose we want a certain margin of error m for a $(1-\alpha)$ CI. What sample size should we use? If we knew p, we could solve for n in the margin of error formula, and get

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

Of course, we don't know p – and we can't even estimate it, because we haven't collected the data yet! What to do?

- Make a guess p^*. If the guess is good, the margin of error will be close to what we want.
- Be conservative: choose $p^* = 0.5$. This will yield the largest n of any p^*.

Example

A 1993 survey reported that 72.1% of freshmen responding to a national survey were attending the college of their first choice. Suppose that $n = 500$ students responded to the survey.

Find a 95% CI for the proportion p of college freshmen attending their first choice college.

What if we wanted a margin to know the proportion within 1% (i.e. ± 0.01) with 95% confidence? How many people should we interview?

Comparing two proportions

Suppose we have two populations A and B with unknown proportions p_1 and p_2 respectively. A SRS of size n_1 from A yields $\hat{p}_1$, and an independent SRS of size n_2 from B yields $\hat{p}_2$. Then,

$$(\hat{p}_1 - \hat{p}_2) \stackrel{\cdot}{\sim} N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$$

when n_1 and n_2 are large.

An approximate 95% CI for $p_1 - p_2$ is then given by

$$(\hat{p}_1 - \hat{p}_2) \pm 1.96\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$