14. Inference on Proportions

In previous inference chapters, we concentrated on answering questions about a population mean (or means). Often, we are interested estimating the proportion, p, of a population having a certain characteristic. Here, we discuss methods for constructing tests and CIs for population proportions.

E.g. Abortion drug RU 486. A study in France examined the effectiveness of the drug RU 486 for terminating early pregnancies.

- n = 488 women given the drug
- x = 473 women terminated pregnancy

What's the "natural" point estimate for the proportion of women that terminate pregnancy after taking RU 486?

- $\hat{p} = \frac{x}{n} = \frac{473}{488} = 0.969$
- where the hat "^" denotes an estimate of the population parameter, in this case, p

As before, we're not satisfied with just a point estimate, we want to get an idea of the precision of this estimate. Again, we'll construct a CI for p for this purpose. We'll go on to do tests as well.

14.1. Normal approximation to the Binomial Distribution and 14.2. Sampling Distribution of a Proportion

In the above example, if it's reasonable to assume that the pregnant women were randomly selected from some population of pregnant women (or population of potentially pregnant women—a "conceptual" population), then it would be

reasonable to assume independence among outcomes between women (pregnancy terminated/not terminated). Notice this is a binary outcome and that there were a fixed number of "trials", n. Assuming that the probability, p, of terminating when given the drug is constant, we can model the number of women terminating pregnancy, X, as a binomial(n,p) random variable. How do we use this to answer questions about p? Recall some things:

- $X = \sum_{i=1}^{n} Y_i$
- where each $Y_i \overset{iid}{\sim} bin(1,p)$ (i.e., the Y_i are independent, identically distributed (iid) bin(1,p) or Bernoulli RVs)
- $\hat{p} = \frac{X}{n} = \frac{\sum_{i=1}^{n} Y_i}{n} = \overline{Y} \sim N(p, \frac{p(1-p)}{n})$, at least approximately, according to the CLT (remember the mean and variance of bin(n,p)?).

When is the approximation "good"?

- Rule of thumb for normal approx
  - o np>5
  - o n(1-p)>5
- Slight problem: don't know p…

14.3. Confidence Intervals (one population p)

E.g., RU 486, cont'd. Armed with the above theoretical result, we press on to calculate a confidence interval for the proportion, p, of women who terminate pregnancy after taking RU 486.

$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$ is, at least approximately, N(0,1), so

$P(-1.96 \le Z \le 1.96) = P(-1.96 \le \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \le 1.96) \approx 0.95$

or, after some algebra,

$P(\hat{p} - 1.96\sqrt{\frac{p(1-p)}{n}} \le p \le \hat{p} + 1.96\sqrt{\frac{p(1-p)}{n}}) \approx 0.95$ (before we plug in a value for p-hat (recall, p-hat is an RV and we feel somewhat uncomfortable saying "probability" when p-hat is fixed after we plug in a number, hence, again, our use of the "fudge" term "confidence"))

Still, we're stuck, sort of, since our confidence limits involve p (which is unknown!). That's okay, we'll make a second approximation and plug in p-hat for p and still call the result a 95% CI:

$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = (\hat{p} - 1.96\sqrt{\frac{p(1-p)}{n}}, \hat{p} + 1.96\sqrt{\frac{p(1-p)}{n}})$ is our (approximate) 95% CI for the population proportion, p.

$\frac{473}{488} \pm 1.96\sqrt{\frac{\frac{473}{488}(\frac{15}{488})}{488}} = (0.954, 0.985)$

Same mantra holds: "We are 95% confident that the true, unknown population proportion of terminated pregnancies among women taking RU 486 is between 0.954 and 0.985." Also, same interpretive statement may be made: "If we were to sample 488 pregnant women 1000 times, each time feeding them RU 486, and calculating 1000 CI in the above manner, then approximately 950 of the intervals would contain the true, unknown population proportion, p."

Generally,

$\hat{p} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = (\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}, \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}})$ is an approximate (1-α)*100% CI

One-sided CIs follow in the (hopefully) obvious way.

14.4. Hypothesis Testing (for one population, i.e., one sample hypothesis testing)

E.g., Drool Dribbler, a starting player for a major college basketball team, made only 38.4% of his free throws last season. After practicing hard in the off-season, and, after a noticeable improvement in his salivary control, DD sank 25 of 40 free throws in his first 8 games of the following season. Sure, he's much less embarrassing to hang out with now, but has his free throw percentage improved? (okay, not a biostat e.g.)

Hypothesis (one or two-sided?)

test statistic (under the null…)

p-value

conclusion

14.5. Sample Size Estimation

Reading is FUNdamental!

14.6. Comparison of Two Proportions (i.e., hypothesis testing for 2 population proportions or 2-sample hypothesis testing

E.g., Aspirin and cerebral ischemia (stroke). A clinical trial randomly assigned patients into treatment and control groups. This study was double-blind in the sense that neither the patient nor the patients' doctors knew who received the aspirin and who received the placebo tablet. After six months, the doctors classified each patients progress as "favorable" or "unfavorable"

- Aspirin Group:
  - $n_1 = 78$
  - $x_1 = 63$ (counts number of favorable outcomes)
  - p-hat$_1 = x_1/n_1 = 63/78 = 0.808$
- Control Group;
  - $n_2 = 77$
  - $x_2 = 43$
  - p-hat$_2 = x_2/n_2 = 43/77 = 0.558$

Is the proportion of favorable outcomes in the (conceptual) aspirin population different (should we say, higher?) than that of the control population? In other words, is the above difference in sample proportions due to chance alone?

Using the CLT we know

$$\hat{p}_1 = \frac{X_1}{n_1} = \frac{\sum_{i=1}^{n_1} Y_{1,i}}{n_1} = \overline{Y_1} \sim N(p_1, \frac{p_1(1-p_1)}{n_1})$$

and,

$$\hat{p}_2 = \frac{X_2}{n_2} = \frac{\sum_{i=1}^{n_2} Y_{2,i}}{n_2} = \overline{Y_2} \sim N(p_2, \frac{p_2(1-p_2)}{n_2})$$

So, using other elementary results we have

$(\hat{p}_1 - \hat{p}_2) \sim N(p_1 - p_2, (\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}))$, at least approximately if our rule of thumb is met for both populations.

E.g., cont'd. We wish to test the hypothesis that the proportion of favorable assessments is higher for the aspirin population verses the control population (1-sided alternative).

Hypothesis (α = ?):

test statistic (computed under the null as usual)
(common p and it's pooled estimate under the null)

p-value

conclusion

Two-sided tests are done with the (hopefully) obvious modification.

Of course, we can also construct a (two-sided or one-sided) CIs for the difference in two population proportions. This is left for you to figure out (see POB)—there should be no surprises here.