CLARK UNIVERSITY

Inferences for proportions
Math 218, Mathematical Statistics

D Joyce, Spring 2016

Discuss chapter 9, inferences for proportions and count data.

This short chapter addresses the oldest question in statistics. We've got Bernoulli trials where there are only two possible outcomes, success with probability p, and failure with probability $q = 1-p$, and our job is to make inferences about these unknown probabilities.

Many polls are of this type. People are asked if agree or disagree with a position, or if they support one of two candidates. When there are only two possible answers, then each question is a Bernoulli trial. (Actual polls typically have a third response, "don't know," so they're actually multinomial distributions, but that can be turned into a two-response poll by ignoring the don't-know responses.)

Point estimator $\hat{p}$ for p. As usual, we'll let X_i be 1 if the i^{th} trial is a success, and 0 if a failure. Then $\sum X_i$ is the number of successes, and $\overline{X}$ is the proportion of successes.

The sample average $\overline{X}$ is the maximum likelihood estimator for p, and it's an unbiased estimator for p. Thus, $\overline{X}$ is usually taken to be the point estimator $\hat{p}$ for p, and $1 - \overline{X}$ as $\hat{q}$. Indeed, what else could you use?

But, perhaps it isn't a good estimator for very small values of n. If you flip a bent coin once and it comes out tails, is it reasonable to estimate that the probability of heads is 0? When we get to Bayesian statistics, something that's coming up pretty soon, we'll see that there are other point estimators that may make more sense.

Interval estimates for p. The number of successes $\sum X_i$ has a binomial distribution, and we can compute the probabilities of a binomial distribution exactly.

However, when n is large, we can approximate $\sum X_i$ by a normal distribution. Because statistical inferences for Bernoulli trials are so common, what counts as large n has been analyzed in detail in this case. So, rather then using a general rule of thumb that $n \geq 30$ or $n \geq 40$ is large enough, in this situation if both $n\overline{x}$ and $n(1-\overline{x})$, that is, both $n\hat{p}$ and $n\hat{q}$, are at least 10, then n can be considered large.

Using the normal approximation, the endpoints of a $(1-\alpha)$-confidence interval are

$$\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n}.$$

As mentioned in the text, there is a slightly better estimate based on further analysis of the binomial distribution.

Determining the sample size for a given margin of error. Recall that the margin of error E for a confidence interval is twice the width of the interval, that is, it's the value E so that that interval has endpoints $\overline{X} \pm E$. Thus, in this situation, the margin of error is

$$E = z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n}.$$

For given values of α and E, we can solve that equation for n to get $n = \left(\frac{z_{\alpha/2}}{E}\right)^2 \hat{p}\,\hat{q}$. Unfortunately, that doesn't help in determining n which has to be done before $\hat{p}$ and $\hat{q}$ are known. In all cases, however, $\hat{p}\hat{q} \leq \frac{1}{4}$, so a value of

$$n = \left(\frac{z_{\alpha/2}}{E}\right)^2 \frac{1}{4}$$

will do.

Hypothesis tests for proportions, large n. A Bernoulli trial X_i has mean p and variance pq, so

the sample mean $\overline{X}$ has mean p and variance pq/n. For large n, the standardized sample mean

$$Z = \frac{\overline{X} - p}{\sqrt{pq/n}} = \frac{\hat{p} - p}{\sqrt{pq/n}}$$

has an approximately standard normal distribution, where $\overline{X} = \hat{p}$ is a point estimator for the parameter p.

A two-sided hypothesis test for a proportion has $H_0 : p = p_0$ and $H_1 : p \neq p_0$, where p_0 is the hypothesized probability of success.

At the α significance level, that is, at the $1 - \alpha$ confidence level, we reject H_0 if

$$|z| > z_{\alpha/2},$$

which works out to equivalent to when

$$|\hat{p} - p_0| > z_{\alpha/2}\sqrt{p_0 q_0/n}$$

where $q_0 = 1 - p_0$.

One-sided hypothesis tests have similar criteria where $\alpha/2$ is replace by α and $|\hat{p} - p_0|$ is replaced by either $\hat{p} - p_0$ or $p_0 - \hat{p}$, depending on the side.

What is the P-value for this test? The P-value, also called the observed level of significance, is the smallest α where H_0 is rejected. So the P-value for the two-sided hypothesis test is

$$\begin{aligned} P - \text{value} &= P(\text{Test rejects } H_0 | p = p_0) \\ &= 2(1 - \Phi(|z|) \\ &= 2\left(1 - \Phi\left(\left|\hat{p} - p\sqrt{pq/n}\right|\right)\right) \end{aligned}$$

The power $\pi(p)$ is the probability of rejecting H_0 for any value of p. Thus, the P-value calculated above is just $\pi(p_0)$. For this z-test,

$$\begin{aligned} \pi(p) &= P(\text{Test rejects } H_0 | p) \\ &= P\left(|\hat{p} - p_0| > z_{\alpha/2}\sqrt{p_0 q_0/n}\right) \\ &= P\left(\hat{p} < p_0 - z_{\alpha/2}\sqrt{p_0 q_0/n}\right) \\ &\quad + P\left(\hat{p} > p_0 + z_{\alpha/2}\sqrt{p_0 q_0/n}\right) \\ &= P\left(\hat{p} - p < p_0 - p - z_{\alpha/2}\sqrt{p_0 q_0/n}\right) \\ &\quad + P\left(\hat{p} - p > p_0 - p + z_{\alpha/2}\sqrt{p_0 q_0/n}\right) \\ &= P\left(z < \frac{p_0 - p - z_{\alpha/2}\sqrt{p_0 q_0/n}}{\sqrt{pq/n}}\right) \\ &\quad + P\left(z > \frac{p_0 - p + z_{\alpha/2}\sqrt{p_0 q_0/n}}{\sqrt{pq/n}}\right) \\ &= \Phi\left(\frac{p - p_0\sqrt{n} - z_{\alpha/2}\sqrt{p_0 q_0}}{\sqrt{pq}}\right) \\ &\quad \Phi\left(\frac{p_0 - p\sqrt{n} - z_{\alpha/2}\sqrt{p_0 q_0}}{\sqrt{pq}}\right) \end{aligned}$$

Inferences for comparing two proportions. Here we have two Bernoulli populations with unknown parameters p_1 and p_2.

Suppose we have random samples $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ from these two populations. Sample means $\overline{X}$ and $\overline{Y}$ are point estimators $\hat{p}_1$ and $\hat{p}_2$, respectively. These estimators have means p_1 and p_2 and variances $p_1 q_1/n_1$ and $p_2 q_2/n_2$, respectively.

For large samples the difference $\hat{p}_1 - \hat{p}_2$ is approximately normally distributed with mean $p_1 - p_2$ and variance $p_1 q_1/n_1 + p_2 q_2/n_2$. An estimate for this variance is $\hat{p}_1\hat{q}_1/n_1 + \hat{p}_2\hat{q}_2/n_2$, so the statistic

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}_1\hat{q}_1/n_1 + \hat{p}_2\hat{q}_2/n_2}}$$

is approximately standard normal, so it can be used to make statistical inferences about $p_1 - p_2$.

A $(1-\alpha)$-confidence interval for $p_1 - p_2$ has endpoints

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}.$$

As in the last chapter, we can have hypothesis tests for the difference of the means where the null hypothese $H_0 : p_1 - p_2 = d$ is that the means differ by a constant d. Typically, however, the hypothesized difference is $d = 0$, and $H_0 : p_1 = p_2$. In that case, a pooled estimate of p, the overall probability of success, is just $\hat{p}$, the proportion of successes in all $n_1 + n_2$ trials, that's just a weighted sum of $\hat{p}_1$ and $\hat{p}_2$

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}.$$

Therefore, the statistic

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\,(\frac{1}{n_1} + \frac{1}{n_2})}}$$

is approximately standard normal and can be used for statistical inferences.

Math 218 Home Page at
`http://math.clarku.edu/~djoyce/ma218/`