STA 580 — Spring 2011 — Dr. Charnigo

Lecture 6

Hypothesis test for a population proportion

Introduction. In Lecture 5 we discussed how to perform hypothesis tests concerning a population mean μ. However, we may wish to make inferences about other population parameters such as p, the population proportion of individuals for whom a statement of interest is true. For example, referring to the second "Obstetrics" on page 286, we may be interested in the proportion of pregnant women taking erythromycin (this is our population: pregnant women who take erythomycin) who experience nausea between the 24th and 28th weeks (this is the statement of interest: experience nausea between the 24th and 28th weeks). In partiuclar, we may wish to assert that this proportion is larger than 0.30, the proportion of pregnant women in general who experience nausea between the 24th and 28th weeks.

Formulation. Suppose that we have null and alternative hypotheses $H_0 : p = p_0$ and $H_1 : p > p_0$, where p_0 is a specified number between 0 and 1. Let $\hat{p}$ denote the sample proportion of individuals for whom the statement of interest is true, and let $\hat{P}$ denote the random conceptualization of $\hat{p}$.

If H_0 is true and n is large enough so that $np_0(1 - p_0) \geq 5$ (≥ 10 is even better)[1], then (as noted in Lectures 3 and 4) the Central Limit Theorem implies that $\hat{P}$ is approximately normal with expected value p_0 and standard

[1] See Equation 7.44 for a procedure that can be used with small n. However, you will not be asked to use this procedure in STA 580.

deviation $\sqrt{p_0(1-p_0)/n}$. Therefore, the quantity

$$Z := \frac{\hat{P} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

is approximately standard normal. In particular, $P(Z > z_{1-\alpha}) \approx \alpha$. So, if we define

$$z := \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

and reject H_0 when $z > z_{1-\alpha}$, in the long run (i.e., after a lifetime of conducting such hypothesis tests) the fraction of occasions on which we will end up rejecting a true null hypothesis will be approximately α.[2] That is, rejecting H_0 when $z > z_{1-\alpha}$ yields a testing procedure with approximate significance level α. Moreover, since z is large only when $\hat{p}$ exceeds p_0, the testing procedure entails rejection of H_0 (i.e., conclusion that the population proportion p exceeds p_0) only when the sample proportion $\hat{p}$ exceeds p_0.

As in Lecture 5, we refer to both random Z and numerical z as the "test statistic" and to $z_{1-\alpha}$ as the "critical value". The (approximate) p-value for this scenario is $P(Z > z)$, where Z is standard normal and z is the numerical test statistic.

Example (formulation). Let p denote the proportion of pregnant women taking erythromycin who experience nausea between the 24th and 28th weeks. Let p_0 denote the proportion of pregnant women in general who experience nausea between the 24th and 28th weeks; this proportion is known to equal 0.30. Since we wish to assert that $p > 0.30$, we take this as our

[2]You may wonder why we do not define z to be $(\hat{p} - p_0)/\sqrt{\hat{p}(1-\hat{p})/n}$, so that there is a stronger resemblance to the interval estimation procedure presented in Lecture 4. In fact, the definition $(\hat{p} - p_0)/\sqrt{\hat{p}(1-\hat{p})/n}$ leads to a valid testing procedure called the "Wald test". The procedure presented in these notes, based on $(\hat{p} - p_0)/\sqrt{p_0(1-p_0)/n}$, is more commonly used because its theoretical justification is easier. The procedure presented in these notes is called the "score test".

alternative hypothesis. So the null hypothesis becomes $p \leq 0.30$, which (for the reason mentioned in Lecture 5) we just write as $p = 0.30$.

We have $\hat{p} = 110/200 = 0.55$, $n = 200$, and $z_{1-\alpha} = z_{0.95} = 1.645$. Thus, to test $H_0 : p = 0.30$ against $H_1 : p > 0.30$ at significance level $\alpha = 0.05$, we compute

$$z = \frac{0.55 - 0.30}{\sqrt{0.30(1 - 0.30)/200}} = 7.72.$$

Since the test statistic $z = 7.72$ is far greater than the critical value $z_{0.95} = 1.645$, we easily reject H_0 and conclude that the proportion of pregnant women taking erythromycin who experience nausea between the 24th and 28th weeks is greater than 0.30. Moreover, the (approximate) p-value is

$$P(Z > 7.72) = 1 - P(Z \leq 7.72) = 1 - \Phi(7.72) \approx 1 - 1 = 0$$

since $\Phi(x) \approx 1$ for $x \geq 4$.

Reversal and two-sided test. If our null and alternative hypotheses were $H_0 : p = p_0$ and $H_1 : p < p_0$, we would proceed in the same fashion except that H_0 would be rejected if $z < -z_{1-\alpha}$. The p-value would be $P(Z \leq z)$.

If our null and alternative hypotheses were $H_0 : p = p_0$ and $H_1 : p \neq p_0$, we would proceed in the same fashion except that H_0 would be rejected if $|z| > z_{1-\alpha/2}$ (Equation 7.42). The p-value would be $2P(Z > z)$ if $z > 0$ and $2P(Z \leq z)$ if $z \leq 0$ (Equation 7.43).

Power calculation. Consider testing $H_0 : p = p_0$ against $H_1 : p > p_0$. Once the Type I error probability (significance level) α and the sample size n are chosen, the Type II error probability β and the power $(1 - \beta)$ are automatically determined. The power can be calculated as (Equation 7.45)

$$\Phi\left[\sqrt{\frac{p_0(1-p_0)}{p_1(1-p_1)}}\left(-z_{1-\alpha}+\frac{\sqrt{n}|p_0-p_1|}{\sqrt{p_0(1-p_0)}}\right)\right],$$

where p_1 is a number greater than p_0 representing an investigator's best guess for p based on scientific reasoning and/or pilot data. Of course, if we want the Type II error probability, we can just subtract the power from 1,

$$1-\Phi\left[\sqrt{\frac{p_0(1-p_0)}{p_1(1-p_1)}}\left(-z_{1-\alpha}+\frac{\sqrt{n}|p_0-p_1|}{\sqrt{p_0(1-p_0)}}\right)\right].$$

The above formulas also apply if we are testing $H_0 : p = p_0$ against $H_1 : p < p_0$, provided that the guess p_1 is now a number less than p_0.

If we are testing $H_0 : p = p_0$ against $H_1 : p \neq p_0$, the above formulas are modified by changing $-z_{1-\alpha}$ to $-z_{1-\alpha/2}$. The guess p_1 can be any number different from p_0.

Example (power calculation). Continuing from the previous example, suppose that we want to know what the power would have been if the sample size were only 50. Taking $p_1 = 0.55$ (our best guess for p), $p_0 = 0.30$, $n = 50$, and $z_{1-\alpha} = z_{0.95} = 1.645$, we find that the power is

$$\begin{aligned}
\Phi\left[\sqrt{\frac{0.30(1-0.30)}{0.55(1-0.55)}}\left(-1.645+\frac{\sqrt{50}|0.30-0.55|}{\sqrt{0.30(1-0.30)}}\right)\right] &= \Phi\left[0.9211\left(-1.645+3.858\right)\right] \\
&= \Phi[2.038] \\
&= 0.979.
\end{aligned}$$

A typical target for power is 80% to 90%. Hence, if we had suspected p to be as high as 0.55, recruiting a sample as large as 50 would have been unnecessary for distinguishing between pregnant women taking erythromycin and pregnant women in general. On the other hand, larger sample sizes yield narrower interval estimates (Lecture 4), so having a large sample size would not be wasteful if we were interested in precisely estimating p (not just asserting that it is different from 0.30).

Sample size calculation. Consider testing $H_0 : p = p_0$ against $H_1 : p > p_0$. Suppose that the Type I error probability (significance level) α is specified but that we can make the sample size n as large as we need to attain the desired power. Then the sample size requirement to have power $1 - \beta$ is obtained by rounding up the result of (Equation 7.46)

$$n = \frac{p_0(1-p_0)\left(z_{1-\alpha} + z_{1-\beta}\sqrt{\frac{p_1(1-p_1)}{p_0(1-p_0)}}\right)^2}{(p_1 - p_0)^2},$$

where p_1 is as defined for the power calculation.

The above formula is unaltered if we are testing $H_0 : p = p_0$ against $H_1 : p < p_0$. If we are testing $H_0 : p = p_0$ against $H_1 : p \neq p_0$, the $z_{1-\alpha}$ in the above formula is changed to $z_{1-\alpha/2}$.

Example (sample size calculation). Continuing from the previous example, suppose that we want to know what sample size would have been sufficient for 90% power. Since $z_{1-\beta} = z_{0.90} = 1.282$, we have

$$n = \frac{0.30(1-0.30)\left(1.645 + 1.282\sqrt{\frac{0.55(1-0.55)}{0.30(1-0.30)}}\right)^2}{(0.55-0.30)^2} = 30.99 \approx 31.$$

Hypothesis test for a population variance

Introduction. We can also perform a hypothesis test concerning a population variance σ^2. For example, we may be interested in the variance of daily iron intake among 9- to 11-year-old males whose families live below the poverty level ("Nutrition", pages 287-288). We may suspect that this variance is less than $(5.56)^2 = 30.91$, the variance of daily iron intake among 9- to 11-year-old males in general.

Formulation. Suppose that we have null and alternative hypotheses $H_0 : \sigma^2 = \sigma_0^2$ and $H_1 : \sigma^2 < \sigma_0^2$, where σ_0^2 is a specified positive number. Suppose, moreover, that we have a normal population. Then $(n-1)S^2/\sigma^2$ follows a chi-square distribution on $(n-1)$ degrees of freedom. Hence, rejecting H_0 if

$$\chi^2 := \frac{(n-1)s^2}{\sigma_0^2} < \chi^2_{n-1,\alpha}$$

provides a hypothesis test with significance level α for which H_0 is rejected when s^2 is small relative to σ_0^2.

Remarks. We cannot dispose of the normality assumption merely by increasing the sample size n. In this respect, making an inference about a population variance σ^2 is fundamentally more difficult than making an inference about a population mean μ. Also, note that the critical value is $\chi^2_{n-1,\alpha}$ rather than, say, $-\chi^2_{n-1,1-\alpha}$. This is because, unlike a T distribution, a chi-square distribution is not symmetric about 0. Indeed, a chi-square random variable can only assume positive values.

Example (formulation). Let σ^2 denote the variance of daily iron intake among 9- to 11-year-old males whose families live below the poverty level. We will test $H_0 : \sigma^2 = 30.91$ against $H_1 : \sigma^2 < 30.91$ at significance level $\alpha = 0.05$. We have $\sigma_0^2 = 30.91$, $s^2 = (4.75)^2 = 22.56$, $n = 51$, and $\chi^2_{50,0.05} = 34.76$. We compute $\chi^2 = 36.49$, which is not less than 34.76. We cannot reject H_0, so we cannot conclude that the variance of daily iron intake among 9- to 11-year-old males whose families live below the poverty level is less than the variance among 9- to 11-year-old males in general.

Reversal and two-sided test. If our null and alternative hypotheses were $H_0 : \sigma^2 = \sigma_0^2$ and $H_1 : \sigma^2 > \sigma_0^2$, we would proceed similarly except that H_0 would be rejected if $\chi^2 > \chi^2_{n-1,1-\alpha}$.

If our null and alternative hypotheses were $H_0 : \sigma^2 = \sigma_0^2$ and $H_1 : \sigma^2 \neq \sigma_0^2$, we would proceed similarly except that H_0 would be rejected if either $\chi^2 > \chi^2_{n-1,1-\alpha/2}$ or $\chi^2 < \chi^2_{n-1,\alpha/2}$ (Equation 7.40).

Confidence intervals and hypothesis tests

Relating the two approaches to inference. In the context of making inferences about a population mean μ, there is a nice equivalence between constructing a confidence interval and performing a hypothesis test (Equations 7.30 and 7.31).[3] Consider a test of $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$ at significance level α. Then H_0 is accepted if and only if μ_0 is included in the $100(1-\alpha)\%$ confidence interval for μ. This makes sense intuitively. If we cannot rule out that $\mu = \mu_0$ when we are performing a hypothesis test, then μ_0 is a plausible

[3] A similar equivalence holds for making inferences about a population proportion p if we use the Wald test mentioned in the previous footnote rather than the score test.

value for μ when we are constructing a confidence interval. On the other hand, if we can rule out that $\mu = \mu_0$ when we are performing a hypothesis test, then μ_0 is not a plausible value for μ when we are constructing a confidence interval.

Example (relating the two approaches to inference). In Lecture 4 we considered the scenario in "Infectious Disease" on page 222. Letting μ denote the population mean time to onset of AIDS following seroconversion, we found that a 95% confidence interval for μ was $[5.02, 5.46]$.

Now suppose that you are asked to test $H_0 : \mu = 5$ against $H_1 : \mu \neq 5$ at level $\alpha = 0.05$. Since 5 is not in the confidence interval $[5.02, 5.46]$, the equivalence between constructing a confidence interval and performing a hypothesis test automatically implies that H_0 must be rejected.

Is that hard to believe? Let's verify it directly. Recall that $n = 287$, $\bar{x} = 5.24$, and $s = 1.90$. The test statistic is

$$z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{5.24 - 5}{1.90/\sqrt{287}} = 2.140,$$

which is larger than the critical value $z_{1-\alpha/2} = z_{0.975} = 1.960$.

Now suppose that you are asked to identify the smallest μ_0 at which $H_0 : \mu = \mu_0$ would not be rejected in favor of $H_1 : \mu \neq \mu_0$. The equivalence between constructing a confidence interval and performing a hypothesis test automatically yields the answer 5.02.

Is that hard to believe? Let's verify it directly. The test statistic is

$$z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{5.24 - \mu_0}{1.90/\sqrt{287}}.$$

For us to be on the borderline between accepting H_0 and rejecting H_0 when $\bar{x} > \mu_0$, the test statistic must equal 1.960. Setting $\frac{5.24-\mu_0}{1.90/\sqrt{287}}$ equal to 1.960 and solving for μ_0 yields $\mu_0 = 5.24 - 1.960 \times 1.90/\sqrt{287} = 5.24 - 0.22 = 5.02$.