1 One-Sided Testing

In a two-sided test, we hypothesize that $H_0 : \mu = \mu_0$ and look for evidence that it's wrong; such evidence would be a t-statistic too big in either the positive or negative direction, expressed as $H_1 : \mu \neq \mu_0$.

When we do a one-sided test, we are only concerned with whether the true mean is either below or above our guess, but not both. For instance, suppose we think that μ is greater than μ_0 and we want to test this guess. The claim being tested becomes the *alternative* hypothesis. So we test, say at 5 percent significance, the null and alternative hypotheses

$$H_0 : \mu \leq \mu_0, \tag{1}$$

$$H_1 : \mu > \mu_0. \tag{2}$$

We again assume that the null is true. We reject the null if we find strong enough evidence against the null, in favor of the alternative. Based on the specification, that evidence would be seen as a value of $\overline{x}$ that is "far enough" above μ_0, in other words, if $\overline{x} - \mu_0$ is very positive.

We quantify "far enough" by again using the t-statistic,

$$t \equiv \frac{\overline{x} - \mu_0}{s/\sqrt{n}} \sim T(n-1). \tag{3}$$

But again, we only reject the null if t is too far *positive*, and hence we only look at the right-tail of the distribution. Hence we put all 5 percent of the rejection region into the right-tail. Thus our critical value is $t_{.05,n-1} =$ `qt(.05, n-1, lower.tail=FALSE)` in R. We reject the null hypothesis if $t > t_{.05,n-1}$. In other words, the rejection region is $(t_{.05,n-1}, \infty)$. This is illustrated in Figure 1.

If instead we think that μ is less than μ_0, the test becomes

$$H_0 : \mu \geq \mu_0,$$

$$H_1 : \mu < \mu_0.$$

In this setup, evidence against the null is when $\overline{x}$ is "far enough" below μ_0. Thus, if the t-statistic is too far *negative*, then we reject the null. This means we are only considering the left-tail of the distribution, in which we put all 5 percent of the test significance. The critical value is therefore $-t_{.05,n-1} =$ `-qt(.05, n-1, lower.tail=FALSE)` in R. We reject

the null hypothesis if $t < -t_{.05,n-1}$. In other words, the rejection region is $(-\infty, -t_{.05,n-1})$.

Rule of Thumb: Put the hypothesis that contains the weak inequality as the null hypothesis, the strict inequality as the alternative.

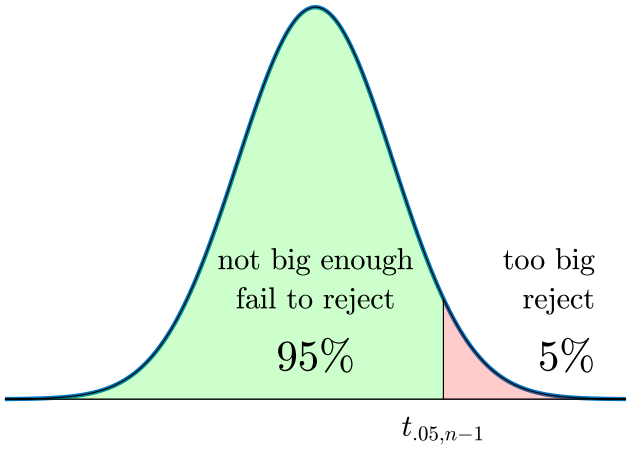

FIGURE 1: $t_{.05,n-1}$ is the number such that 5 percent of the mass of the $T(n-1)$ distribution falls above it. If $H_0 : \mu \leq \mu_0$ is true, then it is unlikely that our test statistic will fall above $t_{.05,n-1}$. But if it does, then we reject the null.

2 Difference in Means Testing

First let me say that there are two difference of means tests. One assumes that the two groups have equal variances; the other does not. Here I do the version where variances are assumed unequal (which is typically the case in reality but not necessarily in a classroom). You can find the case where variances are assumed equal in slides on Canvas, but I omit it.

Suppose we are interested in two groups—call them Group A and Group B—and how their means, μ_A and μ_B, differ. We calculate sample means $\overline{x}_A$ and $\overline{x}_B$ as well as sample variances s_A^2 and s_B^2. We hypothesize that the difference in means is Δ_0; in practice we will often hypothesize that the difference is $\Delta_0 = 0$. Thus our null hypothesis is $H_0 : \mu_A - \mu_B = \Delta_0$. Hence we test

$$H_0 : \mu_A - \mu_B = \Delta_0, \tag{4}$$

$$H_1 : \mu_A - \mu_B \neq \Delta_0. \tag{5}$$

Suppose Group A has sample size n_A and Group B has sample size n_B, not necessarily equal. The test statistic is

$$t \equiv \frac{(\overline{x}_A - \overline{x}_B) - \Delta_0}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}} \sim T(n_A + n_B - 2). \tag{6}$$

The reason we subtract 2 for degrees of freedom is because we are testing with respect to two variables, μ_A and μ_B. From here, the testing procedure proceeds in the usual way.

I illustrated the case of a two-sided test, but you should be able to extend this to a one-sided test as well.

3 One Proportion Testing

For individual i, let $x_i = 1$ for a "successful" event and $x_i = 0$ for a "failure" event. For example, earning a degree might be the successful event, dropping out would therefore be the failure event. The sample proportion of individuals who succeeded is the typical mean, now denoted $p \equiv (\sum_{i=1}^{n} x_i)/n$. Think of p as being an estimate of the true population proportion of successes, π.

Because there are only two possibilities for x_i, we have to use special techniques and formulas. In particular, the standard error of estimate p is given by

$$\text{se}(p) = \sqrt{\frac{p(1-p)}{n}}. \tag{7}$$

Furthermore, sample sizes in proportions analysis are typically large. Large enough, in fact, that the standard normal distribution is typically used instead of $T(n-1)$. Thus we do not use a t-statistic but instead the z-statistic given by

$$z \equiv \frac{p - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} \sim \mathcal{N}(0,1)\,, \tag{8}$$

where π_0 is our hypothesized value for the true proportion of successful events.

A two-sided proportion test would be of the form

$$H_0 : \pi = \pi_0 \tag{9}$$

$$H_1 : \pi \neq \pi_0. \tag{10}$$

At 5 percent significance, we reject the null hypothesis when $|z| > z_{.025}$, where in R you

use $z_{.025}$ =`qnorm(.025, lower.tail=FALSE)`.

Note that for this analysis to be valid, we require that $n\pi_0 \geq 10$ and $n(1-\pi_0) \geq 10$. And again, I illustrated the case of a two-sided test, but you should be able to extend this to a one-sided test as well.

4 Two Proportions Testing

Suppose we have two different population proportions, π_A and π_B. We want to see whether the proportions are the same or not. We sample n_A times for Group A and find y_A successes; we sample n_B times for Group B and find y_B successes. Hence we find estimates

$$p_A = \frac{y_A}{n_A}, \quad p_B = \frac{y_B}{n_B},$$

and the total proportion of successes is

$$\overline{p} = \frac{y_A + y_B}{n_A + n_B}.$$

Our test is of the form

$$H_0 : \pi_A - \pi_B = \Delta_0,$$

$$H_1 : \pi_A - \pi_B \neq \Delta_0.$$

In practice, we will often have $\Delta_0 = 0$, that is, we'll test if there is any difference. We use test statistic

$$z \equiv \frac{(p_A - p_B) - \Delta_0}{\sqrt{\overline{p}(1-\overline{p})\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}} \sim \mathcal{N}(0,1)\,.$$

For this analysis to be valid, we require that $n\pi_0 \geq 5$ and $n(1-\pi_0) \geq 5$ for both n_A and n_B. I illustrated the case of a two-sided test, but you should be able to extend this to a one-sided test as well, deja vu, yes.

5 Chi-Square Distribution

A **chi-square** random variable, denoted χ^2, is a sum of squared standard normal random variables. Because it is a sum of squared objects, it is non-negative; and furthermore it

is right-skewed. Many test statistics have chi-square distribution, so we need to know about it. It has one parameter, *degrees of freedom* denoted k, and as such it is usually denoted $\chi^2(k)$.

Suppose we have $k = 20$ degrees of freedom. We want to know the critical value such that 5 percent of the area underneath the chi-square curve lies to the right of it, as illustrated in Figure 2. Express this number as $\chi^2_{.05,20}$, which can be found in R using the command `qchisq(.05, 20, lower.tail=FALSE)`.

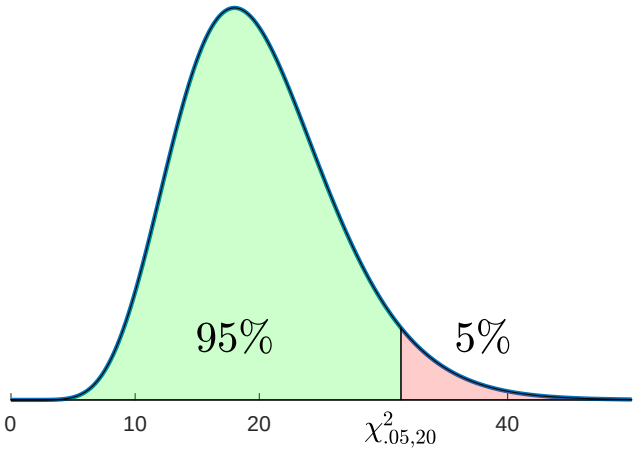

FIGURE 2: $\chi^2_{.05,20}$ is the number such that 5 percent of the $\chi^2(20)$ distribution falls above it.

6 Variance Testing

We usually do not know the true population variance σ^2, so we have to estimate it with

$$s^2 \equiv \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2. \tag{11}$$

But this is an estimation using a sample, and hence it has some uncertainty to it. We seek to quantify that uncertainty. For a two-sided test, we perform the test

$$H_0 : \sigma^2 = \sigma_0^2, \tag{12}$$

$$H_1 : \sigma^2 \neq \sigma_0^2, \tag{13}$$

where σ_0^2 is the hypothesized value for σ^2. The relevant test statistic is

$$\chi^2 \equiv \frac{(n-1)s^2}{\sigma_0^2} \sim \chi^2(n-1), \tag{14}$$

where the distribution is valid if either the population is normally distributed or if $n > 30$. We reject the null hypothesis if χ^2 is in the rejection region, which we now define.

Suppose we are testing at the 10 percent significance level. As usual, we chop the significance level in half for each tail. Problem is, the $\chi^2(n-1)$ distribution is not symmetric. Thus we must calculate two critical values to determine the rejection region. For instance, suppose that $n = 10$. Then we have $n - 1 = 9$ degrees of freedom. We want to find

$$\chi^2_{9,.05} = \text{the number such that 5 percent of the area is to the right of it,}$$
$$\chi^2_{9,.95} = \text{the number such that 95 percent of the area is to the right of it.}$$

These two numbers are visualized in Figure 3. In R, the lower critical value $\chi^2_{.05,9}$ can be found using command `qchisq(.05, 9, lower.tail=FALSE)` and the upper critical value $\chi^2_{.05,9}$ can be found with `qchisq(.95, 9, lower.tail=FALSE)`.

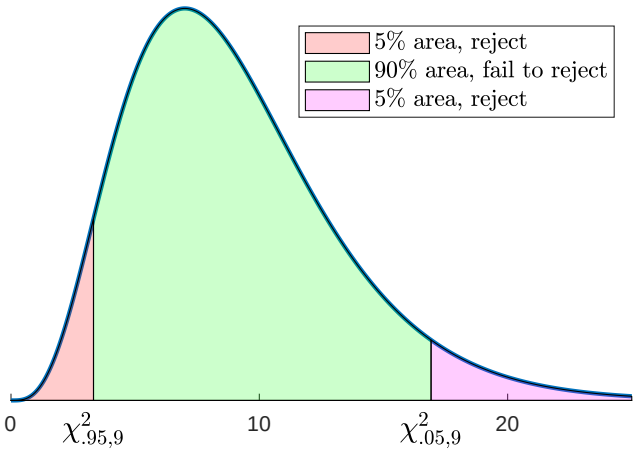

FIGURE 3: $\chi^2_{9,0.05}$ is the number such that 5 percent of the mass of the $\chi^2(9)$ distribution falls above it, and $\chi^2_{9,0.95}$ is the number such that 95 percent of the mass of the $\chi^2(9)$ distribution falls above it. Any χ^2 statistic in either of the 5 percent tails warrants rejecting the null at 10 percent significance.

7 F-Distribution

The ***F*-distribution** is a right-skewed distribution with two different arguments for two different degrees of freedom, so we denote it $F(v_1, v_2)$. What exactly v_1 and v_2 are will become clear once we start testing with it. Suppose $v_1 = 3$ and $v_2 = 15$, and we want to find the critical value of $F(3, 15)$ distribution such that 5% of the data falls to the right of it. Express this number as $F_{.05,3,15}$, which can be found using command `qf(.05, 3, 15, lower.tail=FALSE)` in R. Visually this number will be qualitatively indistinct from the critical value shown in Figure 2.

8 Difference in Variations Testing

Suppose we have two groups. Group A has sample size n_A and Group B has sample size n_B. We calculate two different sample variances for each group, s_A^2 and s_B^2. We want to test if one has true population variance greater than the other. Suppose that the samples give $s_A^2 > s_B^2$. Then let us test $H_0 : \sigma_A^2 \leq \sigma_B^2$ against $H_1 : \sigma_A^2 > \sigma_B^2$ at 5 percent significance.

We will formulate the question in such a way that the test statistic leads to rejection if it is too far in the right tail. To that end, reformulate the test as

$$H_0 : \frac{\sigma_A^2}{\sigma_B^2} \leq 1, \tag{15}$$

$$H_1 : \frac{\sigma_A^2}{\sigma_B^2} > 1, \tag{16}$$

where we use test statistic

$$F \equiv \frac{s_A^2}{s_B^2} \sim F(n_A - 1, n_B - 1). \tag{17}$$

Thus, if we have evidence in favor of the alternative, then F will be greater than 1. If F is sufficiently greater than 1, then we reject the null. The critical value we use for the rejection rule is found using the command `qf(.05, 30, 20, lower.tail=FALSE)` in R.

Rule of Thumb: Put the group with the larger sample variance in the numerator. This will ensure that we do a (more powerful) right-tailed test.

9 Errors in Conclusion

Since we are never 100 percent confident in our conclusions, it is possible that we reject a null hypothesis even when it is true; and also possible that we fail to reject a null hypothesis even when it is false. We employ the following terminology to discuss such scenarios.

- *Type I Error*: Rejection of a true null hypothesis (false positive)
- *Type II Error*: Failing to reject a false null hypothesis (false negative)

The *size* of a test is the probability of mistakenly rejecting a true null. The *power* of a test is the probability of correctly rejecting a false null. A test is said to have significance level α if its size is less than or equal to α. In many cases (and all of *our* cases), the size and significance level of a test are equal.

10 Examples

Example 1

Last Halloween, I ate 84 Starburst candies. However, not all econ grad students have an unquenchable need for Starburst. I don't know how many Starburst econ grad students ate on average, but I'm interested in finding out the variance in Starburst consumption last Halloween because I want to know just how out of hand my Starburst habit was.

I tracked down the Starburst consumption for $n_E = 31$ econ grad students. The average was $\overline{x} = 22$ and the variance was $s^2 = 14$. Someone told me that the true variance in Starburst consumption among econ grad students is actually $\sigma_0^2 = 8$. I think they're full of crap and I want to demonstrate how wrong they are with 95% confidence. Can I?

Solution. The test being performed is

$$H_0 : \sigma^2 = 8,$$
$$H_1 : \sigma^2 \neq 8.$$

The test statistic is

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(30)14}{8} = 52.5.$$

The lower critical value is `qchisq(.975, 30, lower.tail=FALSE)` $= 16.791$, the upper critical value `qchisq(.025, 30, lower.tail=FALSE)` $= 46.979$. Since the test statistic is beyond the interval $[16.791, 46.979]$, which means it is in the rejection region, we reject the null hypothesis. Thus, I can tell that person how full of crap they are at 5 percent significance[1]: "If your guess was true, then there's a less than 5 percent chance that I'd have actually calculated $s^2 = 14$. So you're probably wrong." This case is illustrated in Figure 4.

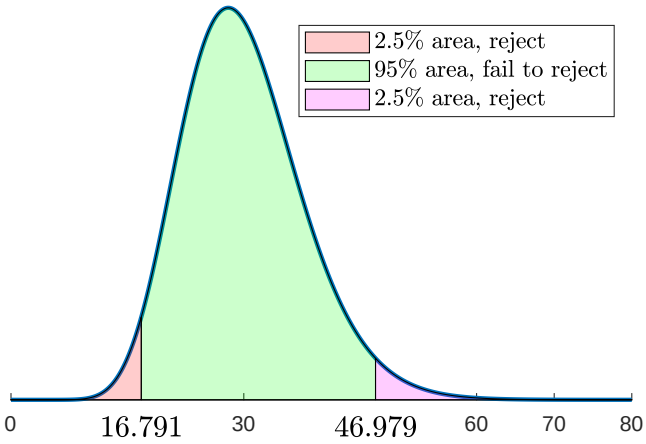

FIGURE 4: If the null is true, then there's a less than 5 percent chance of seeing a χ^2 statistic in the red regions. Since we found $\chi^2 = 52.5$, we reject the null.

Example 2

I also tracked down the Starburst consumption for $n_P = 21$ political science grad students. Their average was $\overline{x}_P = 28$ and the variance was $s_P^2 = 11$, compared to $\overline{x}_E = 22$ and $s_E^2 = 14$ for econ grad students. Someone told me that the true variance in Starburst consumption among political science grad students is lower than that among econ grad students. Test this claim at 5 percent significance.

Solution. This is a one-sided test, so the claim (with the strict inequality) becomes the alternative hypothesis. Rephrase "variance in Starburst consumption among political science grad students is lower" as "variance in Starburst consumption among econ grad

[1] "Full of crap at 5 percent significance" is not standard statistical jargon.

students is higher." The test is

$$H_0 : \frac{\sigma_E^2}{\sigma_P^2} \leq 1,$$

$$H_1 : \frac{\sigma_E^2}{\sigma_P^2} > 1,$$

where we use test statistic

$$F \equiv \frac{s_E^2}{s_P^2} \sim F(n_E - 1, n_P - 1),$$

such that $n_E - 1$ is the numerator (econ) degrees of freedom, and $n_P - 1$ is the denominator (polisci) degrees of freedom. Thus we reject the null in favor of the alternative if we find a test statistic sufficiently larger than 1 (which is consistent with the claim that $\sigma_E^2 > \sigma_P^2$).

Rule of Thumb: Put the group with the larger sample variance in the numerator; then we can do a (more powerful) one-sided test.

Our test statistic is $F = 14/11 \approx 1.273$. We have "numerator" degrees of freedom $v_1 = n_E - 1 = 30$ and "denominator" degrees of freedom $v_2 = n_P - 1 = 20$, so the critical value is given by `qf(.05, 30, 20, lower.tail=FALSE)` $= 2.039$. As shown in Figure 5, The test statistic is below the critical value, hence we fail to reject the null: we have insufficient evidence to claim with 95% confidence that $\sigma_E^2 > \sigma_P^2$.

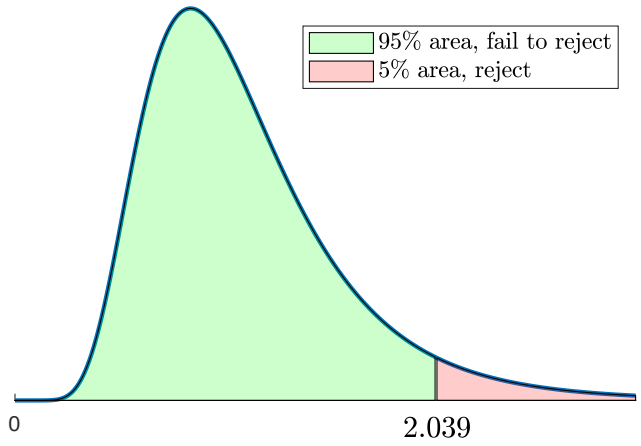

FIGURE 5: The test statistic $F = 1.273$ falls below the critical value $F_{30,20,0.05} = 2.039$, so we fail to reject the null.