# Proportions 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, $\pi$.

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

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

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

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

where $p^*$ is our hypothesized value for the true proportion of successful events.

A two-sided proportion test would be of the form

$$H_0 : \pi = p^*,$$

$$H_1 : \pi \neq p^*.$$

At 5% significance, we reject the null hypothesis when $|z| > z_{0.025}$, where $z_{0.025} =$ `qnorm(1-0.025)` in R or is otherwise found on a normal table.

# 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 $\mu$ is greater than $\mu_0$ and we want to test this guess. The claim being tested becomes the *alternative* hypothesis. So we test, say at 5% significance,

$$H_0 : \mu \leq \mu_0,$$

$$H_1 : \mu > \mu_0.$$

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 $\bar{x}$ that is "far enough" above $\mu_0$, in other words, if $\bar{x} - \mu_0$ is very positive.

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

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

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% of the rejection region into the right-tail. Thus our critical value is $t_{n-1,0.05} =$ `qt(1-0.05,n-1)` in R. We reject the null hypothesis if $t > t_{n-1,0.05}$. In other words, the rejection region is $(t_{n-1,0.05}, \infty)$.

If instead we think that $\mu$ is less than $\mu_0$, the test becomes

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

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

In this setup, evidence against the null is when $\bar{x}$ is "far enough" below $\mu_0$. 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% of the test significance. The critical value is therefore $-t_{n-1,0.05} =$ `qt(0.05,n-1)` in R. We reject the null hypothesis if $t < -t_{n-1,0.05}$. In other words, the rejection region is $(-\infty, -t_{n-1,0.05})$.

# Difference in Means

Suppose we are interested in two groups and how their means, $\mu_1$ and $\mu_2$, differ. We calculate sample means $\bar{x}_1$ and $\bar{x}_2$ as well as sample variances $s_1^2$ and $s_2^2$. We hypothesize that the difference in means is $\mu_0$. Thus our null hypothesis is $H_0 : \mu_1 - \mu_2 = \mu_0$. Hence we test

$$H_0 : \mu_1 - \mu_2 = \mu_0,$$

$$H_1 : \mu_1 - \mu_2 \neq \mu_0.$$

Suppose group 1 has sample size $n_1$ and group 2 has sample size $n_2$, not necessarily equal. The test statistic is

$$t \equiv \frac{(\bar{x}_1 - \bar{x}_2) - \mu_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim T(n_1 + n_2 - 2).$$

The reason we subtract 2 for degrees of freedom is because we are testing with respect to two variables, $\mu_1$ and $\mu_2$. From here, the testing procedure proceeds in the usual way.

# Chi-Squared Distribution

A **Chi-Squared** random variable, denoted $\chi^2$, is a sum of squared standard normal random variables. Many test statistics have chi-squared distribution, so we need to know about it. It has one parameter, the degrees of freedom $k$, and as such it is usually denoted $\chi^2(k)$. On quizzes and exams, we'll use a chi-squared table.

Suppose we have 20 degrees of freedom. We want to know critical value such that 5% of the area underneath the chi-squared curve lies to the right of it. Then we go to the table, look at the row with 20 degrees of freedom and the column with 0.05, which gives 31.410. Express this number as $\chi^2_{20,0.05}$ = `qchisq(1-0.05,20)` in R.

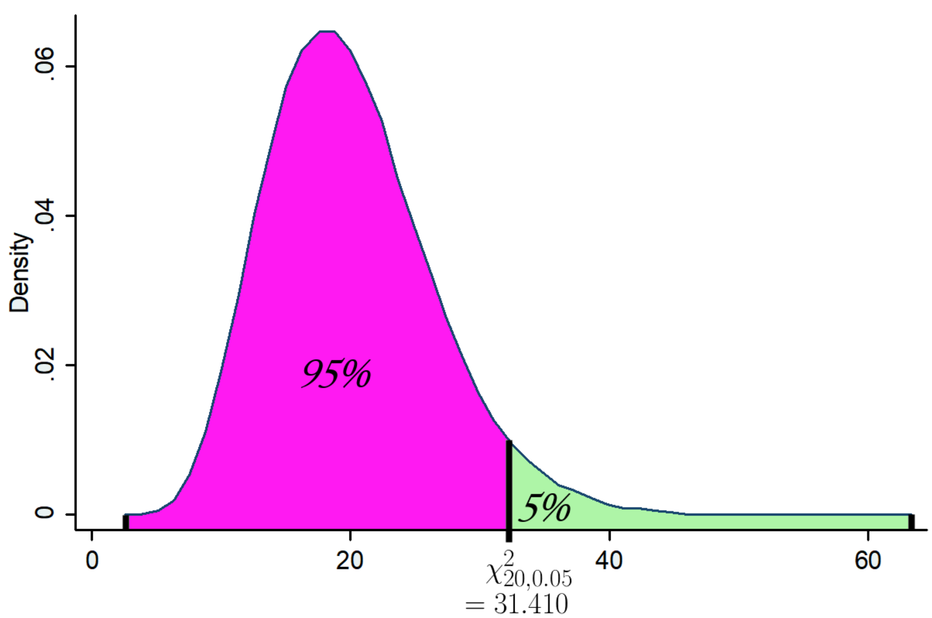


FIGURE 1: $\chi^2_{20,0.05}$ is the number such that 5% of the mass of the $\chi^2(20)$ distribution falls above it.

# ***F*-Distribution**

The ***F*-distribution** has 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 do regression analysis. Since we have to specify two degrees of freedom, it's difficult to condense a comprehensive $F$-distribution onto a single page. Hence our $F$ table contains only numbers for a 5% right-tail (therefore making it applicable to a one-sided test at 5% significance or a two-sided test with 10% significance, since in the latter we split the 10% into both tails).

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. Then we look at the row corresponding to $v_2 = 15$ degrees of freedom, and the column corresponding to $v_1 = 3$ degrees of freedom, which gives 3.287. Express this number as $F_{3,15,0.05} =$ `qf(1-0.05, 3, 15)` in R.