STAT 515 - Chapter 4 Supplement

Brian Habing - University of South Carolina
Last Updated: October 2, 2000

S4 - More on Sampling Distributions: The t, χ^2, and F Distributions

As we saw in Section 4.9, the normal distribution plays a pivotal roll in describing how the sample mean $\bar{x}$ will behave when you have a random sample $x_1, x_2, \ldots x_n$. Unfortunately the central limit theorem only applies when the sample size is large. Additionally, it only tells us about the sampling distribution of the sample mean, and not about the sampling distribution of the sample variance s^2. These limitations can be overcome if we can believe that the sample was taken from a population that was normal to begin with. That is, if we apply the methods in Section 4.6 and verify the data is normal, we can get the sampling distribution for $\bar{x}$ when n is small, and can also get the sampling distribution for s^2.

S4.1 - $\bar{x}$ and the Normal Distribution

A fact that is proved in STAT 512 is: if the random sample is drawn from a population that follows a normal distribution, then $Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ is exactly standard normal. In other words, if the base population is already normal, the central limit theorem result applies even when $n = 1$! The only difficulty in this is that we rarely, if ever, know the value of the parameter σ. Because of this we can't use this fact directly.

S4.2 - s^2 and the χ^2 (chi-squared) Distribution

The χ^2 distribution can be defined as follows. If $z_1, z_2, \ldots z_{(n-1)}$ are independent and each follows the standard normal distribution , then

$$X^2 = z_1^2 + z_2^2 + \cdots z_{(n-1)}^2$$

follows the χ^2 distribution with $(n-1)$ degrees of freedom. The table for this distribution (and a typical picture of it) can be found in TABLE XI on page 521. This distribution is skewed to the right, has mean $(n-1)$, variance $2(n-1)$, and takes all values 0 and higher. (The normal on the other hand takes all positive and negative values.)

The usefulness of this distribution becomes a bit clearer if we again consider the random sample $x_1, x_2, \dots x_n$ from a normal distribution. Looking at the formula for s^2:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

we can see that we are squaring a bunch of independent normal random variables (the x_i) and summing them up. The only reason that this isn't a χ^2 random variable is that they aren't standard normal, and we are dividing by n-1.

By multiplying both sides of the above equation by $n-1$ and dividing by σ^2 we get the following:

$$\frac{(n-1)s^2}{\sigma^2} = \sum_{i=1}^{n}(\frac{x_i - \bar{x}}{\sigma})^2$$

If the $\bar{x}$ on the right side of the equation were replaced by μ, then we would be summing up a bunch of $z = \frac{x_i - \bar{x}}{\sigma}$ and it would be the sum of n standard normals, making a χ^2 random variable with n degrees of freedom. Because we are using $\bar{x}$ instead of μ we lose one degree of freedom, and so:

$$\chi^2_{df=n-1} = \frac{(n-1)s^2}{\sigma^2} \qquad (S1)$$

where the *df* in the subscript is the number of degrees of freedom.

If the data come from a sample that is normal, we know that the left hand side of equation S1 behaves as a χ^2 random variable, we known $n-1$ because it is based on the sample size, and we know s^2 because we can calculate it from the data. If we solve this equation for σ^2 we could then get information about this unknown population parameter. We will discuss this more in Chapters 5 and 6.

Table XI on pages 521-522 of the text give several of the values of the χ^2 random variable for a variety of degrees of freedom. Each row of this table corresponds to a different number of degrees of freedom (*df*). Remember that you need to look at $df = n-1$ if you are using the sample variance to investigate the population variance. The rest of the table is set up the opposite of the normal table. The values along the top are the probabilities (the areas that are shaded in on the figure) and the body of the table contains the t values that go with those probabilities. Because each *df* needs its own row, no table can possibly contain all of the possible values. (A normal table can because we can change any normal to a standard normal, there is no such simplification for the χ^2.)

Say we wanted to know $P(T_{df=8} \leq 1.344)$. Looking at the $df = 8$ row of Table XI, we see that this corresponds to a probability of 0.995. The probability is for greater than 1.344 though, so we have to take 1-0.995 and get a value of 0.005. Many of the values are not in the table however. The example on page 395 is seeking $P(T_{df=2} \geq 6.52)$. The closest two values are 5.99 and 7.37. Because of this, all we can say about the probability is that it is between 0.05 and 0.025. If an exact value is needed then a computer package such as SAS would be used instead. We could also use the table in the reverse direction. The t_0 such that $P(T_{df=10} \geq t_0) = 0.010$ is 23.2093. The t_0 such that $P(T_{df=10} \leq t_0) = 0.010$ is 2.55821.

S4.3 - $\bar{x}$, s, and Student's t-Distribution

As noted in S4.1, the difficulty with the central limit theorem is that it requires us to know σ. The tool we use to do this is the t distribution that was discovered in 1908 by chemist William Gosset at the Guinness brewery in Ireland. Because he didn't want employees at other breweries to know that he found statistics useful, he published his results under the pseudonym *Student*. Hence, the distribution is often known as Student's t distribution. A t-distribution is formed by dividing a standard normal by a χ^2 over its degrees of freedom, where the normal and the χ^2 are independent.

$$t_{df=n-1} = \frac{Z}{\sqrt{\frac{\chi^2_{df=n-1}}{n-1}}} \qquad (S2)$$

At first this seems to be more than a little bit out of nowhere. A fact proved in STAT 714 sheds some light on why it is useful however. If the sample $x_1, x_2, \ldots x_n$ is independent, then its sample mean $\bar{x}$ and sample variance s^2 are independent! If the sample comes from a normal distribution, S4.1 showed that $\bar{x}$ is related to a standard normal distribution, and S4.2 showed that s^2 is related to a χ^2 distribution. Combining these previous results gives:

$$t_{df=n-1} = \frac{Z}{\sqrt{\frac{\chi^2_{df=n-1}}{n-1}}} = \frac{\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{\frac{(n-1)s^2}{\sigma^2}}{n-1}}}$$

By cancelling the $n-1$ terms in the denominator, applying the square root, multiplying both the numerator and denominator by one over the denominator, and cancelling, we get

$$t_{df=n-1} = \frac{\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{s^2}{\sigma^2}}} = \frac{\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}}{\frac{s}{\sigma}} = \frac{\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\frac{\sigma}{s}}{\frac{s}{\sigma}\frac{\sigma}{s}} = \frac{\bar{x}-\mu}{s/\sqrt{n}} \qquad (S3)$$

Just as we could solve equation S1 to find out information about the population variance, we can solve equation S3 to find out information about the population mean, if the sample comes from a population that follows the normal distribution.

This usage is discussed more in Section 5.2. One useful fact about the t distribution is that it becomes very similar to the standard normal distribution as the sample size n increases.

Many tables for the t distribution stop at 30 degrees of freedom and simply refer the user to a standard normal table. Our table IV continues on past 30, but does not give all the values. Notice that the values change very little from one row to the next after about row eighteen. If you go to the bottom row, all of the values should be recognizable from the normal table.

S4.4 - Two variances and the F-Distribution

The final sampling distribution we will be concerned with is the F-distribution. The F-distribution is defined by:

$$F_{df_x=n_x-1,df_y=n_y-1} = \frac{\frac{X_x^2}{n_x-1}}{\frac{X_y^2}{n_y-1}} \tag{$S4$}$$

where X_x^2 and X_y^2 are independent χ^2 random variables with $n_x - 1$ and $n_y - 1$ degrees of freedom respectively. Because it is formed by using two χ^2 random variables, the F-distribution has two separate degrees of freedom, one for the numerator and one for the denominator. This makes the F tables even more complicated than the χ^2 or t tables.

The formula for the F-distribution again looks out of nowhere, until you recognize that we could get this formula by comparing two variances. Say we have independent random samples from two populations, call them $x_1, x_2, \ldots x_{n_x}$ and $y_1, y_2, \ldots y_{n_y}$. We could then write:

$$F_{df_x=n_x-1,df_y=n_y-1} = \frac{\frac{\frac{(n_x-1)s_x^2}{\sigma_x^2}}{n_x-1}}{\frac{\frac{(n_y-1)s_y^2}{\sigma_y^2}}{n_y-1}}$$

Cancelling and then inverting the fractions, we get:

$$F_{df_x=n_x-1,df_y=n_y-1} = \frac{\frac{s_x^2}{\sigma_x^2}}{\frac{s_y^2}{\sigma_y^2}} = \frac{\frac{s_x^2}{s_y^2}}{\frac{\sigma_x^2}{\sigma_y^2}} \qquad (S5)$$

Equation number S5 thus allows us to compare the variances of two different populations, if we can assume both populations are normally distributed. Tables VII, VIII, IX, and X give the values of the F-distribution for various combinations of degrees of freedoms and areas under the curve. The F-distribution is useful not only for comparing two variances, but in Section 7.6 we will see that it is useful for comparing more than two means, and in Chapter 9 that it is useful for predicting one variable from another.

One final fact that we will encounter later concerns the relationship between the t-distribution and the F-distribution. Turn back to formula S2, and square the numerator and the denominator. The denominator becomes the same as the denominator in S5. The numerator becomes a Z^2, which is just a χ^2 with one degree of freedom. We thus get the following result.

$$(t_{df=n-1})^2 = F_{df_x=1,df_y=n-1}$$

This result means that in some cases in Chapter 9 we will be able to work with either a t-distribution or an F-distribution and get the same result.