

## Supplement to Chapter 23: CHI-SQUARED DISTRIBUTIONS, T-DISTRIBUTIONS, AND DEGREES OF FREEDOM

To understand t-distributions, we first need to look at another family of distributions, the *chi-squared distributions*. These will also appear in Chapter 26 in studying categorical variables.

*Notation:*

- $N(\mu, \sigma)$  will stand for the normal distribution with mean  $\mu$  and standard deviation  $\sigma$ .
- The symbol  $\sim$  will indicate that a random variable has a certain distribution. For example,  $Y \sim N(4, 3)$  is short for “Y has a normal distribution with mean 4 and standard deviation 3”.

### ***I. Chi-squared Distributions***

*Definition:* The chi-squared distribution with  $k$  degrees of freedom is the distribution of a random variable that is the sum of the squares of  $k$  independent standard normal random variables. We’ll call this distribution  $\chi^2(k)$ .

Thus, if  $Z_1, \dots, Z_k$  are all standard normal random variables (i.e., each  $Z_i \sim N(0,1)$ ), and if they are independent, then

$$Z_1^2 + \dots + Z_k^2 \sim \chi^2(k).$$

For example, if we consider taking simple random samples (with replacement)  $y_1, \dots, y_k$  from some  $N(\mu, \sigma)$  distribution, and let  $Y_i$  denote the random variable whose value is  $y_i$ , then each  $\frac{Y_i - \mu}{\sigma}$  is standard normal, and  $\frac{Y_1 - \mu}{\sigma}, \dots, \frac{Y_k - \mu}{\sigma}$  are independent, so

$$\left(\frac{Y_1 - \mu}{\sigma}\right)^2 + \dots + \left(\frac{Y_k - \mu}{\sigma}\right)^2 \sim \chi^2(k).$$

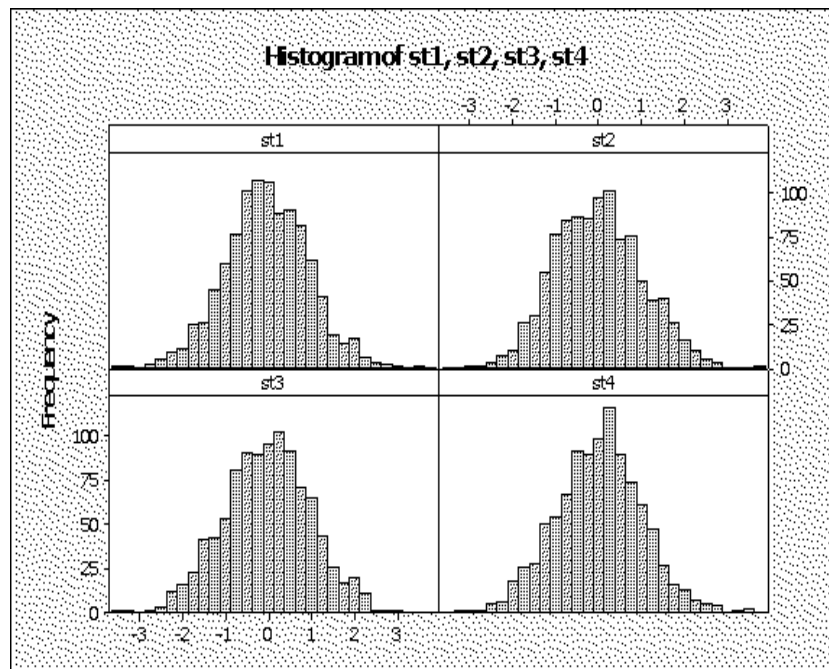
Notice that the phrase “degrees of freedom” refers to the number of independent standard normal variables involved. The idea is that since these  $k$  variables are independent, we can choose them “freely” (i.e., independently).

The following exercise should help you assimilate the definition of chi-squared distribution, as well as get a feel for the  $\chi^2(1)$  distribution.

*Exercise 1:* Use the definition of a  $\chi^2(1)$  distribution and the 66-95-99.7 rule for the standard normal distribution (and/or anything else you know about the standard normal distribution) to help sketch the graph of the probability density

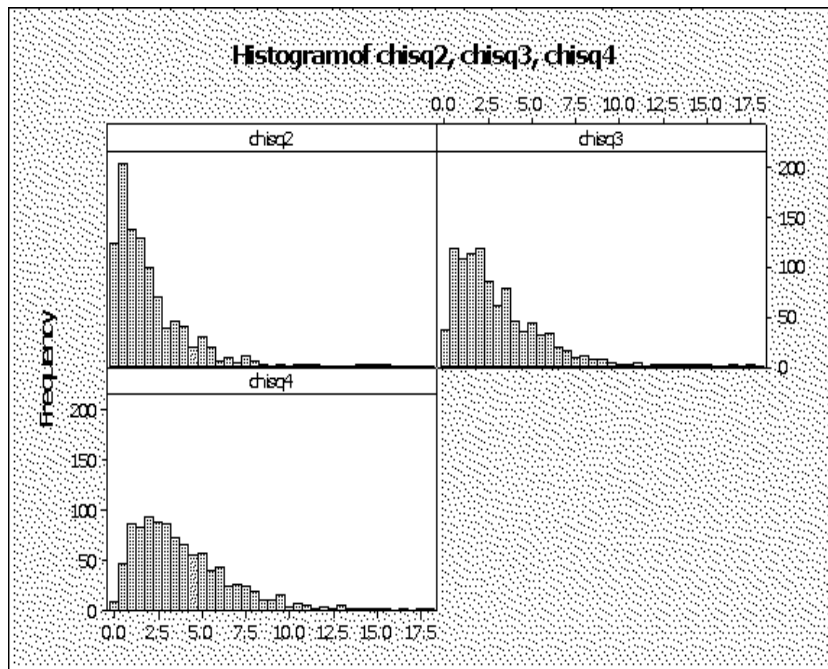
function of a  $\chi^2(1)$  distribution. (For example, what can you conclude about the  $\chi^2(1)$  curve from the fact that about 68% of the area under the standard normal curve lies between -1 and 1? What can you conclude about the  $\chi^2(1)$  curve from the fact that about 5% of the area under the standard normal lies beyond  $\pm 2$ ?)

For  $k > 1$ , it's harder to figure out what the  $\chi^2(k)$  distribution looks like just using the definition, but simulations using the definition can help. The following diagram shows histograms of four random samples of size 1000 from an  $N(0,1)$  distribution:



These four samples were put in columns labeled st1, st2, st3, st4. Taking the sum of the squares of the first two of these columns then gives (using the definition of a chi-squared distribution with two degrees of freedom) a random sample of size 1000 from a  $\chi^2(2)$  distribution. Similarly, adding the squares of the first three columns gives a random sample from a  $\chi^2(3)$  distribution, and forming the column  $(st1)^2 + (st2)^2 + (st3)^2 + (st4)^2$  yields a random sample from a  $\chi^2(4)$  distribution. Histograms of these three samples from chi-squared distributions are shown below, with the sample from the  $\chi^2(2)$  distribution in the upper left, the sample from the  $\chi^2(3)$  distribution in the upper right, and the sample from the  $\chi^2(4)$  distribution in the lower left.

The histograms show the shapes of the three distributions: the  $\chi^2(2)$  has a sharp peak at the left; the  $\chi^2(3)$  distribution has a less sharp peak not quite as far left; and the  $\chi^2(4)$  distribution has a still lower peak still a little further to the right. All three distributions are noticeably skewed to the right.



There is a picture of a typical chi-squared distribution on p. A-113 of the text.

*Thought question:* As  $k$  gets bigger and bigger, what type of distribution would you expect the  $\chi^2(k)$  distribution to look more and more like? [Hint: A chi-squared distribution is the sum of independent random variables.]

*Theorem:* A  $\chi^2(1)$  random variable has mean 1 and variance 2.

The proof of the theorem is beyond the scope of this course. It requires using a (rather messy) formula for the probability density function of a  $\chi^2(1)$  variable. Some courses in mathematical statistics include the proof.

*Exercise 2:* Use the Theorem together with the definition of a  $\chi^2(k)$  distribution and properties of the mean and standard deviation to find the mean and variance of a  $\chi^2(k)$  distribution.

## II. *t* Distributions

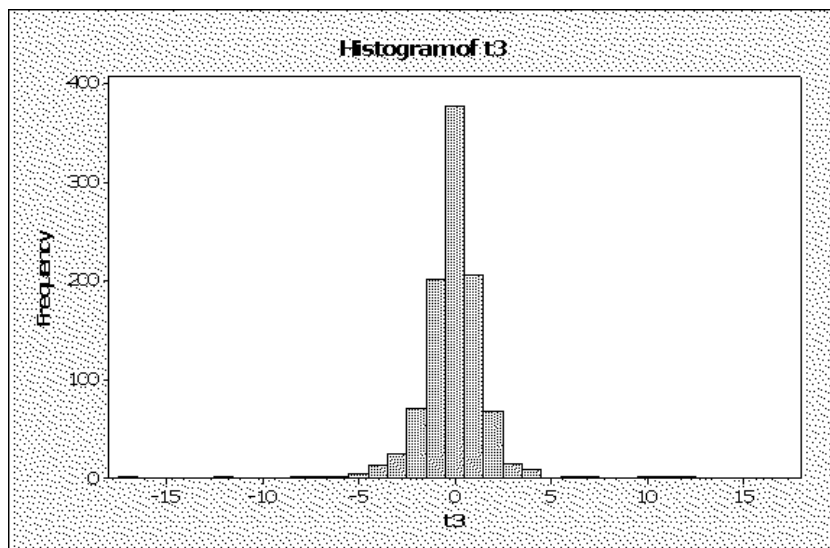
*Definition:* The *t distribution with  $k$  degrees of freedom* is the distribution of a random variable which is of the form  $\frac{Z}{\sqrt{U/k}}$  where

- i.  $Z \sim N(0,1)$
- ii.  $U \sim \chi^2(k)$ , and
- iii.  $Z$  and  $U$  are independent.

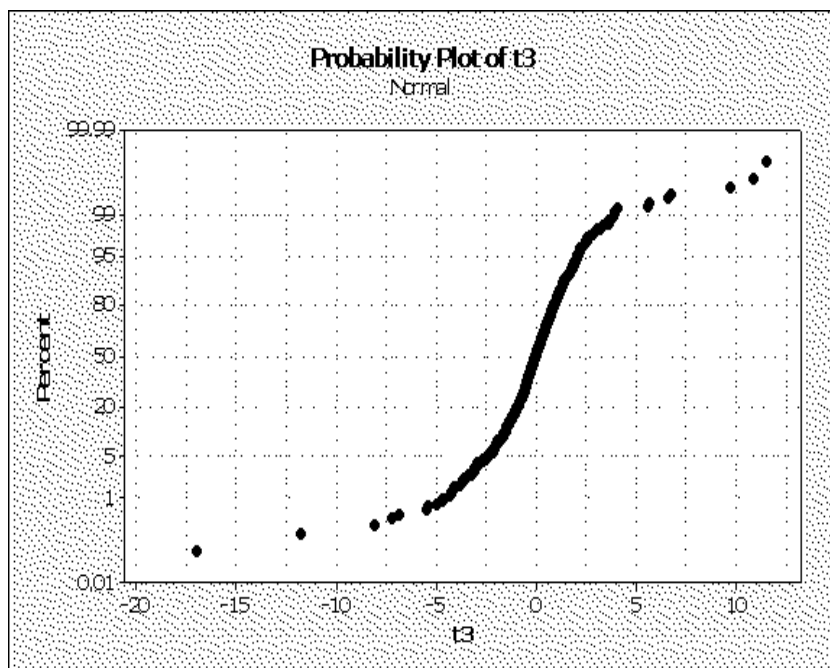
*Comment:* Notice that this definition says that the notion of “degrees of freedom” for a t-distribution comes from the notion of degrees of freedom of a chi-squared distribution: The degrees of freedom of a t-distribution are the number of squares of independent normal random variables that go into making up the chi-squared distribution occurring under the radical in the denominator of the t random variable  $\frac{Z}{\sqrt{U/k}}$ .

To see what a t-distribution looks like, we can use the four standard normal samples of 1000 obtained above to simulate a t distribution with 3 degrees of freedom: We use column s1 as our sample from Z and  $(st2)^2 + (st3)^2 + (st4)^2$  as our sample from U to calculate a sample from the t distribution  $\frac{Z}{\sqrt{U/3}}$  with 3 degrees

of freedom. The resulting histogram is:



Note that this histogram shows a distribution similar to the t-model with 2 degrees of freedom shown on p. 554 of the textbook: It's narrower in the middle than a normal curve would be, but has “heavier tails” – note in particular the outliers that would be very unusual in a normal distribution. The following normal probability plot of the simulated data draws attention to the outliers as well as the non-normality. (The plot is quite typical of a normal probability plot for a distribution with heavy tails on both sides.)



### ***III. Why the t-statistic introduced on p. 553 of the textbook has a t-distribution:***

1. *General set-up and notation:* Putting together the two parts of the definition of t-statistic in the box on p. 553 gives

$$t = \frac{\bar{y} - \mu}{s / \sqrt{n}},$$

where  $\bar{y}$  and  $s$  are, respectively, the mean and sample standard deviation calculated from the sample  $y_1, y_2, \dots, y_n$ .

To talk about the *distribution* of the t-statistic, we need to consider all possible random<sup>1</sup> samples of size  $n$  from the population for  $Y$ . We'll use the convention of using capital letters for random variables and small letters for their values for a particular sample. In this case, we have three statistics involved:  $\bar{Y}$ ,  $S$  and  $T$ . All three have the same associated random process: Choose a random sample from the population for  $Y$ . Their values are as follows:

The value of  $\bar{Y}$  is the sample mean  $\bar{y}$  of the sample chosen.

The value of  $S$  is the sample standard deviation  $s$  of the sample chosen.

The value of  $T$  is the t-statistic  $t = \frac{\bar{y} - \mu}{s / \sqrt{n}}$  calculated for the sample chosen.

The distributions of  $\bar{Y}$ ,  $S$  and  $T$  are called the sampling distributions of the mean, the sample standard deviation, and the t-statistic, respectively.

Note that the formula for calculating  $t$  from the data gives the formula  $T = \frac{\bar{Y} - \mu}{S/\sqrt{n}}$ , expressing the random variable  $T$  as a function of the random variables  $\bar{Y}$  and  $S$ .

We'll first discuss the  $t$ -statistic in the case where our underlying random variable  $Y$  is normal, then extend to the more general situation stated in Chapter 23.

2. The case of  $Y$  normal. For  $Y$  normal, we will use the following theorem:

*Theorem:* If  $Y$  is normal with mean  $\mu$  and standard deviation  $\sigma$ , and if we only consider simple random samples with replacement<sup>2</sup>, of fixed size  $n$ , then

a) The (sampling) distribution of  $\bar{Y}$  is normal with mean  $\mu$  and standard deviation  $\sigma/\sqrt{n}$ ,

b)  $\bar{Y}$  and  $S$  are independent random variables, and

c)  $(n-1) \frac{S^2}{\sigma^2} \sim \chi^2(n-1)$

The proof of this theorem is beyond the scope of this course, but may be found in most textbooks on mathematical statistics. Note that (a) is a special case of the Central Limit Theorem. We will give some discussion of the plausibility of parts (b) and (c) in the Comments section below.

So for now suppose  $Y$  is a normal random variable with mean  $\mu$  and standard deviation  $\sigma$ :

$$Y \sim N(\mu, \sigma).$$

By (a) of the Theorem, the sampling distribution of the sample mean  $\bar{Y}$  (for simple random samples with replacement, of fixed size  $n$ ) is normal with mean  $\mu$  and standard deviation  $\sigma/\sqrt{n}$ :

$$\bar{Y} \sim N(\mu, \sigma/\sqrt{n}).$$

Standardizing  $\bar{Y}$  then gives

$$\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1). \quad (*)$$

But we don't know  $\sigma$ , so we need to approximate it by the sample standard deviation  $s$ . It would be tempting to say that since  $s$  is approximately equal to  $\sigma$ ,

this substitution (in other words, considering  $\frac{\bar{Y} - \mu}{s/\sqrt{n}}$ ) should give us something

approximately normal. Unfortunately, there are two problems with this:

- First, using an approximation in the *denominator* of a fraction can sometimes make a big difference in what you're trying to approximate (See Footnote 3 for an example.)
- Second, we are using a different value of  $s$  for different samples (since  $s$  is calculated from the sample, just as the value of  $\bar{Y}$  is.) This is why we need to work with the random variable  $S$  rather than the individual sample standard deviation  $s$ . In other words, we need to work with the random

$$\text{variable } T = \frac{\bar{Y} - \mu}{S/\sqrt{n}}$$

To use the theorem, first apply a little algebra to see that

$$\frac{\bar{Y} - \mu}{S/\sqrt{n}} = \frac{\left(\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}}\right)}{S/\sigma} \quad (**)$$

Since  $Y$  is normal, the numerator in the right side of  $(**)$  is standard normal, as noted in equation  $(*)$  above. Also, by (c) of the theorem, the denominator of the right side of  $(**)$  is of the form  $\sqrt{U/(n-1)}$  where  $U = (n-1) \frac{S^2}{\sigma^2} \sim \chi^2(n-1)$ . Since

altering random variables by subtracting constants or dividing by constants does not affect independence, (b) of the theorem implies that the numerator and denominator of the right side of  $(**)$  are independent. Thus for  $Y$  normal, our test

statistic  $T = \frac{\bar{Y} - \mu}{S/\sqrt{n}}$  satisfies the definition of a  $t$  distribution with  $n-1$  degrees of

freedom.

**3. More generally:** The textbook states (pp. 555 – 556) assumptions and conditions that are needed to use the  $t$ -model:

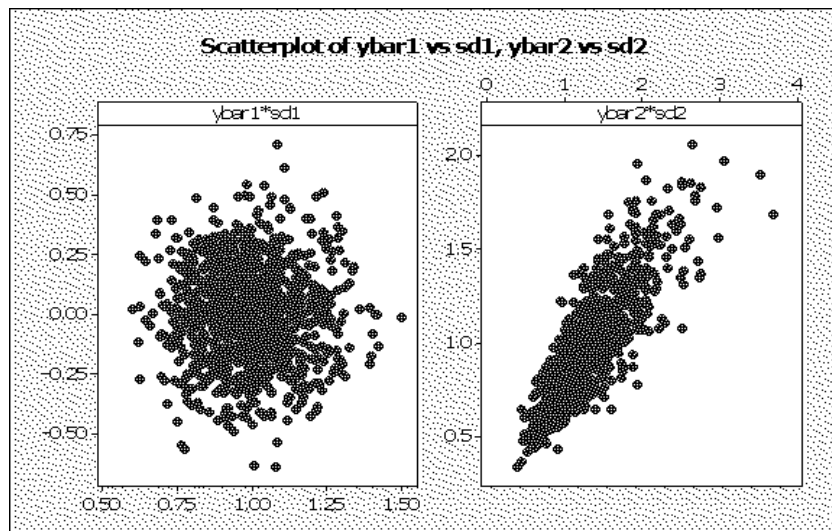
- The heading “Independence Assumption” on p. 555 includes an Independence Assumption, a Randomization Condition, and the 10% Condition. These three essentially say that the sample is close enough to a simple random with replacement to make the theorem close enough to true, still assuming normality of  $Y$ .
- The heading “Normal Population Assumption” on p. 556 consists of the “Nearly Normal Condition,” which essentially says that we can also weaken normality somewhat and still have the theorem close enough to true for most practical purposes. (The rough idea here is that, by the

central limit theorem,  $\bar{Y}$  will still be close enough to normal to make the theorem close enough to true.)

The appropriateness of these conditions as good rules of thumb has been established by a combination of mathematical theorems and simulations.

#### 4. Comments:

i. To help convince yourself of the plausibility of Part (b) of the theorem, try a simulation as follows: Take a number of simple random samples from a normal distribution and plot the resulting values of  $\bar{Y}$  vs S. Here is the result from one such simulation:



The left plot shows  $\bar{y}$  vs  $s$  for 1000 draws of a sample of size 25 from a standard normal distribution. The right plot shows  $\bar{y}$  vs  $s$  for 1000 draws of a sample of size 25 from a skewed distribution. The left plot is elliptical in nature, which is what is expected if the two variables plotted are indeed independent. On the other hand, the right plot shows a noticeable dependence between  $\bar{Y}$  and  $S$ :  $\bar{y}$  increases as  $s$  increases, and the conditional variance of  $\bar{Y}$  (as indicated by the scatter) also increases as  $S$  increases.

ii. To get a little insight into (c) of the Theorem, note first that

$$(n-1) \frac{S^2}{\sigma^2} = \sum_{i=1}^n \frac{(Y_i - \bar{Y})^2}{\sigma^2},$$

which is indeed a sum of squares, but of  $n$  squares, not  $n-1$ . However, the random variables being squared are not independent; the dependence arises from the relationship  $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$ . Using this relationship, it is possible to show



that  $(n-1)\frac{S^2}{\sigma^2}$  is indeed the sum of  $n-1$  *independent, standard normal* random variables.

Although the general proof is somewhat involved, the idea is fairly easy to see when  $n = 2$ :

First, a little algebra shows that (for  $n = 2$ )

$$Y_1 - \bar{Y} = \frac{Y_1 - Y_2}{2} \text{ and } Y_2 - \bar{Y} = \frac{Y_2 - Y_1}{2}.$$

Plugging these into the formula for  $S^2$  (with  $n = 2$ ) then gives

$$(n-1) \frac{S^2}{\sigma^2} = \frac{1}{\sigma^2} \left[ 2 \left( \frac{Y_1 - Y_2}{2} \right)^2 \right] = \left( \frac{Y_1 - Y_2}{\sqrt{2}\sigma} \right)^2 \quad (***)$$

Since  $Y_1$  and  $Y_2$  are independent and both are normal,  $Y_1 - Y_2$  is also normal (by a theorem from probability).

Since  $Y_1$  and  $Y_2$  have the same distribution,

$$E(Y_1 - Y_2) = E(Y_1) - E(Y_2) = 0$$

Using independence of  $Y_1$  and  $Y_2$ , we can also calculate

$$\begin{aligned} \text{Var}(Y_1 - Y_2) &= \text{Var}(Y_1) + \text{Var}(Y_2) \\ &= 2\sigma^2 \end{aligned}$$

Standardizing  $Y_1 - Y_2$  then shows that  $\frac{Y_1 - Y_2}{\sqrt{2}\sigma}$  is standard normal, so

equation (\*\*\*) shows that  $(n-1) \frac{S^2}{\sigma^2} \sim \chi^2(1)$  when  $n = 2$ .

### Footnotes

1. “Random” is admittedly a little vague here. In section 2, interpret it to mean “simple random sample with replacement.” (See also Footnote 2). In section 3, interpret random to mean “Fitting the conditions and assumptions for the t-model.”

2. Technically, the requirements are that the random variables  $Y_1, Y_2, \dots, Y_n$  representing the first, second, etc. values in the sample are “independent and identically distributed” (abbreviated as iid), which means they are independent and have the same distribution (i.e., the same probability density function).

3. Consider, for example, using 0.011 as an approximation of 0.01 when estimating  $1/0.01$ . Although 0.011 differs from 0.01 by only 0.001, when we use the approximation in the denominator, we get  $1/0.011 = 90.\overline{90}$ , which differs by more than 9 from  $1/0.01 = 100$  – a difference almost 3 orders of magnitude greater than the difference between 0.01 and 0.001.