

## Small sample tests

Sampling problems discussed so far dealt with means and proportions. Evaluation of their sampling errors was based on the normal distribution. In the case of the mean, the sampling distribution was normal because the variable was distributed normally in the population or because the Central Limit Theorem ensured normality for large samples. In the case of proportions, the normal distribution was used as an approximation for the underlying binomial distribution. In each case, we required a large sample ( $n \geq 30$ ). When samples are small,  $n < 30$ , when the population is normally distributed, and when the population *variance* has to be estimated from sample data, the distribution of the sample mean is no longer normal. A small sample distribution, known as the t-distribution, has to be used in this case. When samples are small and the distribution of the variable in the population is not normal, there is no readily available sampling distribution. When dealing with proportions coming from small samples, it is necessary to use the exact binomial distribution.

### 6.1 The t-distribution

Assume that the variable is distributed normally in the population with mean  $\mu$  and variance  $\sigma^2$ , i.e.  $X \sim N(\mu, \sigma^2)$ . If  $\sigma^2$  is known, then the sample mean is normally distributed, and we have no problem. However, in almost all cases we do not in fact know the population variance,  $\sigma^2$ , and must estimate it. We have seen that the estimator

$\hat{S}^2 = \sum_{i=1}^n (X_i - \bar{X})^2 / (n - 1)$  is an unbiased estimator of  $\sigma^2$ . We let

$\hat{S} = \sqrt{\hat{S}^2}$ . However, when we replace  $\sigma$  with  $\hat{S}$  in the usual formula for  $Z$ , we get:

$$t = \frac{\bar{X} - \mu}{\hat{S} / \sqrt{n}}$$

This does not have a normal distribution. It can be shown that this statistic, the t-statistic, has the t-distribution with n-1 degrees of freedom.

For large  $n$ , the t-distribution resembles the standard normal distribution, but we are interested here in small samples. The formula for the t-distribution is quite complicated, and depends on the number of degrees of freedom. However, it is symmetric about 0, so the same useful

shortcuts, such as  $P(t > -a) = P(t < a)$  can be used as for the standard normal. It can be shown that  $E(t) = 0$ , and  $\text{Var}(t) = k/(k-2)$ , where  $k$  is the number of degrees of freedom, so in this case,  $\text{Var}(t) = (n-1)/(n-3)$ .

Tables of the cumulative t-distribution for different numbers of degrees of freedom are available. There is also a t-distribution function in Excel: For  $x > 0$ , and  $k$  degrees of freedom, the function  $\text{TDIST}(x, n, 1)$  will return  $P(t > x)$ , while the function  $\text{TDIST}(x, n, 2)$  will return the 2-tailed test,  $P(t > x \text{ OR } t < -x)$ . There is also a function  $\text{TINV}(p, n)$  will return the critical value  $X_C$  for a 2-tailed t-distribution with  $n$  degrees of freedom, such that  $P(|t| > X_C) = p$ .

The distribution of  $X$  in the population has to be normal for the t-statistic to have the t-distribution. However, the t-distribution is quite robust, and small deviations from normality in the population will not invalidate it.

### Tables of the t-distribution

The t-distribution will depend on degrees of freedom. Typically, a table of the t-distribution will give the critical values corresponding to different probability levels for a 1-tailed test. (For a 2-tailed test, you must *halve* the probability level, since you are considering that probability in each 'tail'.) Part of a typical table by degrees of freedom ( $k$ ) and probability ( $\alpha$ ) is shown below.

$k/\alpha$	...	.05	.025	.01	...
1					
2					
3					
4		2.1318	2.7764	3.7469	
5		2.0150	2.5706	3.3649	

For example,  $P(t \geq 2.5706) = 0.025$  for the t-distribution with 5 degrees of freedom. We write  $2.5706 = t_{\alpha=0.05, k=5}$ , or  $t_{0.05, 5} = 2.5706$ .

### Uses of the t-distribution

As the t-distribution is a sampling distribution, it can be used to construct confidence intervals for the population mean  $\mu$  and to test hypotheses.

### Confidence interval

If a random sample of size  $n$  comes from a normal population with mean  $\mu$  and variance  $\sigma^2$  (both  $\mu$  and  $\sigma^2$  being unknown) we can state

$$P\left[-t_{\alpha/2, n-1} < \frac{\bar{X} - \mu}{\hat{S} / \sqrt{n}} < t_{\alpha/2, n-1}\right] = 1 - \alpha.$$

Since there is a probability  $\alpha/2$  that the t-statistic will be higher than the  $\alpha/2$  critical value, and another  $\alpha/2$  that it will be below minus that value. This expression can be re-arrange to give the  $(1-\alpha)$  confidence interval for  $\mu$ ,

$$P\left[\bar{X} - t_{\alpha/2, n-1} \hat{S} / \sqrt{n} < \mu < \bar{X} + t_{\alpha/2, n-1} \hat{S} / \sqrt{n}\right] = 1 - \alpha.$$

For example, if we want a 95% confidence interval, then we choose  $\alpha = .05$ .

Compare this with the 95% confidence interval in the large sample case, and when we were assuming a known  $\sigma$ . Here we had

$$P\left[\bar{X} - 1.96\sigma / \sqrt{n} < \mu < \bar{X} + 1.96\sigma / \sqrt{n}\right] = 0.95.$$

Here, 1.96 is the critical value of the standard normal distribution, such that  $P(Z > 1.96) = 0.025$ . (Since this is a 2-tailed test). Thus, the  $\alpha/2$  critical value of the standard normal distribution is replaced with the  $\alpha/2$  critical value of the t-distribution. The population standard deviation  $\sigma$  is replaced by an unbiased estimate of the standard deviation,  $\hat{S}$ . In each case, the confidence interval is measured in standard errors of the sample mean; in the case of the known S.D.,  $SE(\bar{X}) = \sigma / \sqrt{n}$ , in the case of the unknown SD, it is  $\hat{S} / \sqrt{n}$ .

### Example

A random sample of 16 households is taken from a large block of flats, and shows that household expenditure on food is £42 per week, with a standard deviation of £10. Assuming that household expenditure on food is normally distributed, find the 95% confidence interval for the population mean.

$$\text{As } S^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2, \text{ we have}$$

$$\hat{S}^2 = S^2 * n / (n-1) = (16/15) * 10^2 = 106.67, \text{ so } \hat{S} = 10.33.$$

From tables,  $t_{.025,15} = 2.1314$ . Thus the confidence interval is

$$\mu = 42 \pm 2.1314(10.33/\sqrt{16}) = 42 \pm 5.50$$

or  $36.5 < \mu < 47.5$ . The confidence interval is quite wide, since  $n$  is small and  $\hat{S}$  is quite large.

### Test of hypothesis

The procedure for testing a hypothesis is similar to that used for large samples, i.e. based on the normal distribution, but instead of using the z-statistic, we now use the t-statistic.

Procedure: Set up the null hypothesis,  $H_0: \mu = \mu_0$  (say) and the alternative hypothesis,  $H_1: \mu \neq \mu_0$ . Choose the significance level  $\alpha$  at which  $H_0$  is to be tested. The test statistic is  $t = \frac{\bar{X} - \mu_0}{\hat{S} / \sqrt{n}}$ . The critical value of  $t$  is  $t_{\alpha/2, n-1}$  as

this is a 2-tailed test, and is found from tables. The decision rule is to reject  $H_0$  if  $|t| > t_{\alpha/2, n-1}$ , and accept  $H_0$  otherwise. If the alternative hypothesis were  $H_1: \mu > \mu_0$  or  $H_1: \mu < \mu_0$ , then we would use a 1-tailed test, with the critical value being  $t_{\alpha, n-1}$ .

Note again that our decision rule is based on measuring how many standard errors the sample mean is from the hypothesised population mean.

### Difference between two sample means

We may

If two small random samples are taken from two normal populations with the *same variance*, it can be shown that the statistic:

$t = \frac{[(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)]}{\hat{S}_p \sqrt{(1/n_1) + (1/n_2)}}$  has the t-distribution with  $(n_1 + n_2 - 2)$  degrees of freedom where

$$\hat{S}_p^2 = \frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}$$

is a pooled estimate of the common population variance, and  $n_1$  and  $n_2$  are the sample sizes. When the variables  $X_1$  and

$X_2$  are not normally distributed, or when the population variances are not equal, the test (sometimes called the student t-test) is not strictly valid. However the t-distribution is quite robust, so that small deviations from normality or small differences in the variances can be ignored in practice.

Very often, our null hypothesis will be that the two population means are equal, so that  $\mu_1 - \mu_2$  in the above formula will be equal to 0.

### Example

Continuing the last example, suppose that a random sample of 12 households taken from another large block of flats showed an average household food expenditure of £36 per week with a standard deviation of £9 per week. Assuming that household expenditure on food is normally distributed in each block, and that the population variances are equal, test the hypothesis that the two population means are the same.

$$H_0: \mu_1 - \mu_2 = 0 \quad H_1: \mu_1 \neq \mu_2. \quad \text{Assume } \alpha = 0.05.$$

We first calculate the estimated population variance,

$$\hat{S}_p^2 = [12(9^2) + 16(10^2)] / (12 + 16 - 2) = 98.92, \text{ hence } \hat{S}_p = 9.95$$

$$t = \frac{(42 - 36) - 0}{9.95 \sqrt{(1/16) + (1/12)}} = 1.58.$$

The critical value of t obtained from tables is  $t_{0.025, 26} = 2.0555$ .

(There are  $16 + 12 - 2 = 26$  d.f.).

As  $1.58 < 2.055$ ,  $H_0$  cannot be rejected at the 5% level of significance.

The t-statistic is also crucial in regression analysis, as the difference between an estimated regression parameter and the population parameter, divided by its standard error, has the t-distribution. We therefore use t-statistics to test hypotheses about regression parameters, for example the hypothesis that the parameter is equal to zero (i.e. no relationship between the variables).

## 6.2 The $\chi^2$ distribution

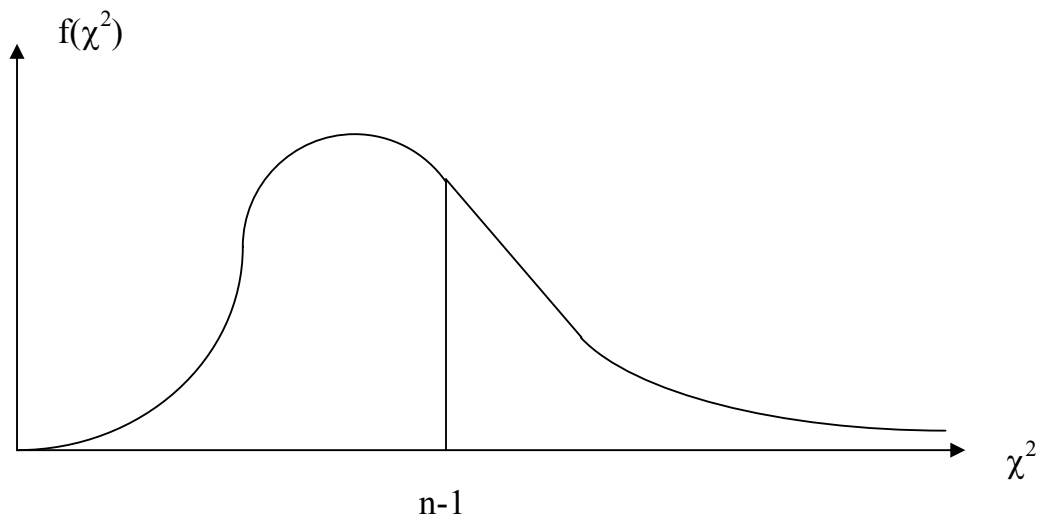
The  $\chi^2$  distribution has many applications; it can be used to test hypotheses about population variances, and about the distribution of two or more populations amongst different categories. (For example, are the distributions of different ethnic groups amongst different classes of job the same?) It also appears in many contexts in regression analysis. We introduce it initially in terms of population variances.

When a random sample of size  $n$  is taken from a population in which a variable  $X$  follows the normal distribution, it can be shown that the statistic

$$\chi^2 = nS^2/\sigma^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} \quad \text{where } \sigma^2 \text{ is the population variance has the } \chi^2$$

distribution with  $(n-1)$  degrees of freedom. The distribution depends on the number of degrees of freedom, it has a complicated formula and is positively skewed.

The variable  $\chi^2$  lies between zero and  $\infty$ ,  $E(\chi^2) = (n-1)$  and  $\text{Var}(\chi^2) = 2(n-1)$ . As  $n$  increases, the distribution slowly approaches the normal distribution. When  $n \geq 100$ , the approximation is quite close. A typical  $\chi^2$  distribution is shown below:



Tables show the area ( $\alpha$ ) under the  $\chi^2$  curve to the right of a particular value of  $\chi^2$  for a given number of degrees of freedom,  $k$ . For example, the entry in the table for  $k=4$  and  $\alpha=0.95$  is 0.7107. This means that

$P(\chi^2_4) > 0.7107 = 0.95$ . The entry for  $k=4$  and  $\alpha=0.05$  is 9.488, so  $P(\chi^2_4) > 9.488 = 0.05$ .

### Confidence interval for the population variance ( $\sigma^2$ )

As the statistic  $\chi^2 = nS^2/\sigma^2$  has the  $\chi^2$  distribution with  $n-1$  d.f., we can write

$$P[\chi^2_{.975, n-1} < (nS^2/\sigma^2) < \chi^2_{.025, n-1}] = 0.95.$$

Rearranging, we get a 95% confidence interval:

$$P[nS^2/\chi^2_{.975, n-1} < \sigma^2 < nS^2/\chi^2_{.025, n-1}] = 0.95.$$

Similarly, we may test hypotheses about  $\sigma^2$  using the  $\chi^2$  statistic.

### 6.3 The F-distribution

The F-distribution can be used to test equality of two population variances. It also occurs frequently in regression analysis. It is used to test whether a set of regression results as a whole is significant, and it can be used to test whether a more complicated model is to be preferred to a simpler model. We introduce it in terms of population variances.

If samples of size  $n_1$  and  $n_2$  respectively are taken from two normal populations with variances  $\sigma_1^2$  and  $\sigma_2^2$ , it can be shown that the statistic

$$F = \frac{(\hat{S}_1^2 / \sigma_1^2)}{(\hat{S}_2^2 / \sigma_2^2)}$$
 has the F-distribution with  $k_1 = n_1 - 1$  and  $k_2 = n_2 - 1$  d.f.,

where  $\hat{S}_1^2$  and  $\hat{S}_2^2$  are the unbiased estimates of the population variances, that is

$$\hat{S}_1^2 = \frac{\sum_{i=1}^n (X_{1i} - \bar{X}_1)^2}{n_1 - 1}$$
 and similarly for  $\hat{S}_2^2$ .

The F-distribution has a complicated formula, and depends on *two* degrees of freedom,  $k_1$  and  $k_2$ . It is positively skewed, taking values between 0 and  $\infty$ . It can be shown that  $E(F) = k_2 / (k_2 - 2)$  for  $k_2 > 2$ .

## Tables

The F-tables show critical values of F corresponding to different values of  $\alpha$  (tail probabilities) and different combinations of degrees of freedom,  $k_1=n_1-1$  in the numerator and  $k_2=n_2-1$  in the denominator. The table entry will show  $F_{k_1,k_2,\alpha}$  s.t.  $P(F > F_{k_1,k_2,\alpha}) = \alpha$ . An extract from a table is shown below:

$k_2/k_1$	1	2	3	4
1 $\alpha=.05$ $\alpha=.025$				
2 $\alpha=.05$ $\alpha=.025$				19.25 39.25
3 $\alpha=.05$ $\alpha=.025$				

For example,  $P(F > F_{.05,4,2} = 19.25) = .05$ ,  $P(F > F_{.025,4,2} = 39.25) = .025$ .

That is, if our F distribution has 4 d.f. in the numerator, and 2 in the denominator, then the 95% critical value is 19.25, and the 97.5% critical value is 39.25. Note that only the upper-tailed values of the F-distribution are tabulated. This is because it is always possible to place the larger value of  $\hat{S}^2 / \sigma^2$  in the numerator of the F ratio, so that the observed values of F will always fall in the right-hand tail.

### Test of hypothesis

Set up hypotheses,  $H_0: \sigma_1^2 = \sigma_2^2$   $H_1: \sigma_1^2 \neq \sigma_2^2$ .

Select level of  $\alpha = 0.025$  (say, to get a 2-tailed test for significance level of 5%). The test statistic is

$$F = \frac{(\hat{S}_1^2 / \sigma_1^2)}{(\hat{S}_2^2 / \sigma_2^2)} = \frac{\hat{S}_1^2}{\hat{S}_2^2} \text{ since under } H_0, \sigma_1^2 = \sigma_2^2.$$

Convention: The larger estimate of the common population variance is placed in the numerator of the F-ratio, so if  $\hat{S}_2^2 > \hat{S}_1^2$ , we let  $F = \hat{S}_2^2 / \hat{S}_1^2$  in order to ensure that F falls in the upper tail of the F-distribution.

The critical value of F is obtained by looking at the F-table with  $k_1$  d.f. on the top of the table (horizontal) and  $k_2$  d.f. on the left hand side of the

table (vertical). Look up the appropriate box, and select the value of F for the appropriate value of  $\alpha$  in that table box.

Decision rule: For a one-tail test, if  $F > F_{\alpha, k_1, k_2}$ ,  $H_0$  can be rejected at the  $\alpha$  level of significance. For a 2-tailed test (usually the case),  $H_0$  can only be rejected at the  $2\alpha$  level of significance, e.g. if we want a 5% level of significance we must take  $\alpha = .025$ .

### Example

We want to test whether male and female students have different variances in their test scores on a certain course. The 25 male students have a sample variance of  $\sigma_m^2 = 225$ , and the 31 female students have a sample variance of  $\sigma_f^2 = 121$ . Test the hypotheses that the variances are equal at the 5% level of significance, using a 2-tailed test.

Our null hypothesis is  $H_0: \sigma_m^2 = \sigma_f^2$ .

First of all, we must calculate  $\hat{S}_1^2$  and  $\hat{S}_2^2$ .

We have that  $S_1^2 = 225$ , so  $\hat{S}_1^2 = S_1^2 * n_1 / (n_1 - 1) = 225 * 25 / 24 = 234.4$ , and  $\hat{S}_2^2 = S_2^2 * n_2 / (n_2 - 1) = 121 * 30 / 29 = 125.2$ . So the test statistic is

$F = 234.4 / 125.2 = 1.872$ . There are  $25 - 1 = 24$  d.f. in the numerator and  $30 - 1 = 29$  d.f. in the denominator. We use the FINV function in Excel, where  $\text{FINV}(p, k_1, k_2)$  gives the value of  $F^*$  s.t.  $P(F > F^*) = p$ , where the F-distribution has  $k_1$  and  $k_2$  d.f. Hence we want  $\text{FINV}(0.025, 24, 29) = 2.154$ . (Since we want a 2-tailed test). Since  $1.87 < 2.154$ , we cannot reject  $H_0$ , so we do not have sufficient evidence to conclude that male and female students have different variances.

(NB: it seems here that we have taken as our sample the whole class, so what is the difference between the sample variance and the 'population' variance? In this case, we would be taking our 'population' to be *male and female students in general*, or *hypothetical* future students on the course. Of course, we would need to consider carefully whether it is legitimate to extrapolate from our sample, this year's class, to the general case. This is a common problem in statistical and regression analysis; we might have quite a limited sample, and the question of whether we can extrapolate to future cases, or to say, different countries or different circumstances, is often quite uncertain.)

