

Lecture 10

The T distribution and Introduction to Sampling

10.1 Using the T distribution

You may have noticed a hole in the reasoning we used in reasoning about the husbands' heights in section 9.1. Our computations depended on the fact that

$$Z := \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

has standard normal distribution, when μ is the population mean and σ is the population standard deviation. But *we don't know σ* ! We did a little slight of hand, and substituted the sample standard deviation

$$S := \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$$

for the (unknown) value of σ . But S is only an estimate for σ : it's a random variable that might be too high, and might be too low. So, what we were calling Z is not really Z , but a quantity that we should give another name to:

$$T := \frac{\bar{X} - \mu}{S/\sqrt{n}}. \quad (10.1)$$

If S is too big, then $Z > T$, and if S is too small then $Z < T$. On average, you might suppose, Z and T would be about the same — and, in this you would be right. Does the distinction matter then?

Since T has an extra source of error in the denominator, you would expect it to be more widely scattered than Z . That means that if you compute T from the data, but look it up on a table computed from the distribution of Z — the standard normal distribution — you would underestimate the probability of a large value. The probability of rejecting a true null hypothesis (Type I error) will be larger than you thought it was, and the confidence intervals that you compute will be too narrow. This is very bad! If we make an error, we always want it to be on the side of underestimating our confidence.

Fortunately, we can compute the distribution of T (sometimes called “Student’s t ”, after the pseudonym under which statistician William Gossett published his first paper on the subject, in 1908). While the mathematics behind this is beyond the scope of this course, the results can be found in tables. These are a bit more complicated than the normal tables, because there is an extra parameter: Not surprisingly, the distribution depends on the number of samples. When the estimate is based on very few samples (so that the estimate of SD is particularly uncertain) we have a distribution which is far more spread out than the normal. When the number of samples is very large, the estimate s varies hardly at all from σ , and the corresponding t distribution is very close to normal. As with the χ^2 distribution, this parameter is called “degrees of freedom”. For the T statistic, the number of degrees of freedom is just $n - 1$, where n is the number of samples being averaged. Figure 10.1 shows the density of the t distribution for different degrees of freedom, together with that of the normal. Note that the t distribution is symmetric around 0, just like the normal distribution.

Table 10.1 gives the critical values for a level 0.05 hypothesis test when Z is replaced by t with different numbers of degrees of freedom. In other words, if we define $t_\alpha(d)$ to be the number such that $\mathbb{P}\{T < t_\alpha\} = \alpha$ when T has the Student distribution with d degrees of freedom, Table 10.1(a) gives values of $t_{0.95}$, and Table 10.1(b) gives values of $t_{0.975}$. Note that the values of $t_\alpha(d)$ decrease as d increases, approaching a maximum, which is $z_\alpha = t_\alpha(\infty)$.

10.1.1 Using t for confidence intervals: Single sample

Suppose we have observations x_1, \dots, x_n from a normal distribution, where the mean and the SD are both unknown. To compute confidence intervals

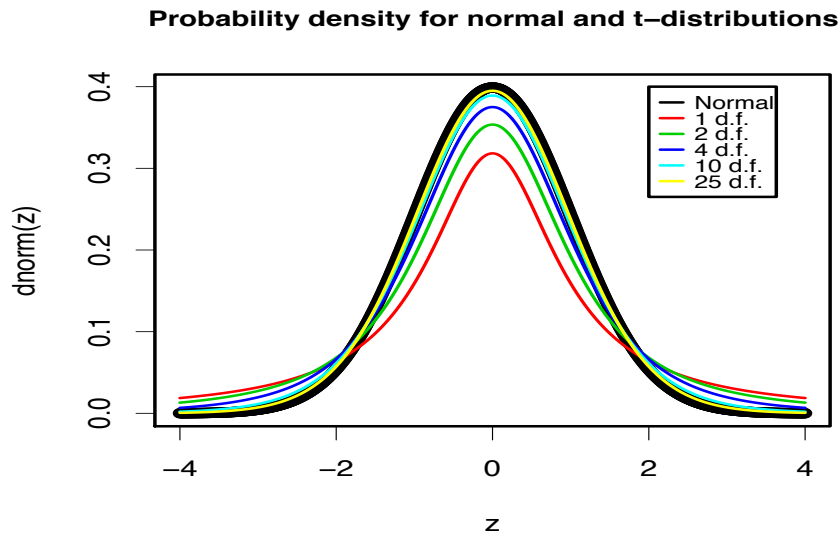


Figure 10.1: The standard normal density together with densities for the t distribution with different degrees of freedom.

Table 10.1: Cutoffs for hypothesis tests at the 0.95 level, using the t statistic with different degrees of freedom. The ∞ level is the limit for a very large number of degrees of freedom, which is identical to the distribution of the Z statistic.

(a) one-tailed test		(b) two-tailed test	
degrees of freedom	critical value	degrees of freedom	critical value
1	6.31	1	12.7
2	2.92	2	4.30
4	2.13	4	2.78
10	1.81	10	2.23
50	1.68	50	2.01
∞	1.64	∞	1.96

with the t statistic, we follow the same procedures as in section 9.1, substituting s for σ , and the quantiles of the t distribution for the quantiles of the normal distribution: that is, where we looked up a number z on the normal table, such that $P(Z < z)$ was a certain probability, we substitute a number t such that $P(T < t)$ is that same probability, where T has the Student T distribution with the right number of degrees of freedom. Thus, if we want a 95% confidence interval, we take

$$\bar{X} \pm t \times \frac{s}{\sqrt{n}},$$

where t is found in the column marked “ $P = 0.05$ ” on the T-distribution table — 0.05 being the probability above t that we are excluding. It corresponds, of course, to $P(T < t) = 0.975$.

Example 10.1: Heights of British men

In section 9.1 we computed a confidence interval for the heights of married British men, based on a sample of size 200. Since we were using the sample SD to estimate the population SD, we should have used the t quantiles with 199 degrees of freedom, rather than the Z quantiles. If you look on a table of the t distribution you won't find a row corresponding to 199 degrees of freedom, though. Why not? The t distribution with 199 degrees of freedom is almost indistinguishable from the normal distribution. To give an example, the multiplier for a symmetric 90% normal confidence interval is $z_{0.95} = 1.645$; the corresponding t quantile is $t_{0.95}(199) = 1.653$, so the difference is less than 1%. There is no real application where you are likely to be able to notice an error of that magnitude. ■

Example 10.2: Kidney dialysis

A researcher measured the blood level of phosphate in the blood of dialysis patients on six consecutive clinical visits.¹ It is important to maintain the levels of various nutrients in appropriate bounds during dialysis treatment. The values are known to vary

¹This example is adapted from [MM98, p.529], where it was based on a Master's thesis of Joan M. Susic at Purdue University.

approximately according to a normal distribution. For one patient, the values (in mg/dl) were measured 5.6, 5.1, 4.6, 4.8, 5.7, 6.4. What is a symmetric 99% confidence interval for the patient's true phosphate level?

We compute

$$\begin{aligned}\bar{X} &= \frac{1}{6}(5.6 + 5.1 + 4.6 + 4.8 + 5.7 + 6.4) = 5.4\text{mg/dl} \\ s &= \sqrt{\frac{1}{5} \left((5.6 - 5.4)^2 + (5.1 - 5.4)^2 + (4.6 - 5.4)^2 \right. \\ &\quad \left. + (4.8 - 5.4)^2 + (5.7 - 5.4)^2 + (6.4 - 5.4)^2 \right)} \\ &= 0.67\text{mg/dl}.\end{aligned}$$

The number of degrees of freedom is 5. Thus, the symmetric confidence interval will be

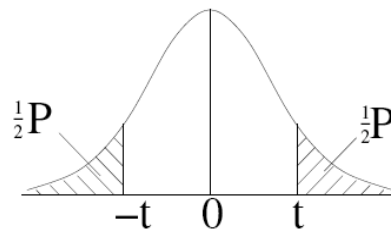
$$\left(5.4 - t \frac{0.67}{\sqrt{6}}, 5.4 + t \frac{0.67}{\sqrt{6}} \right) \text{mg/dl},$$

where t is chosen so that the T variable with 5 degrees of freedom has probability 0.01 of being bigger than t . ■

10.1.2 Using the T table

T tables are like the χ^2 table. For Z, the table in your official booklet allows you to choose your value of Z, and gives you the probability of finding Z below this value. Thus, if you were interested in finding z_α , you would look to find α inside the table, and then check which index Z corresponds to it. In principle, we could have a similar series of T tables, one for each number of degrees of freedom. To save space, though, and because people are usually not interested in the entire t distribution, but only in certain cutoffs, the T tables give much more restricted information. The rows of the T table represent degrees of freedom, and the columns represent cutoff probabilities. The values in the table are then the values of t that give the cutoffs at those probabilities. One peculiarity of these tables is that, whereas the Z table gives one-sided probabilities, the t table gives two-sided probabilities. This makes things a bit easier when you are computing symmetric confidence intervals, which is all that we will do here.

The probability we are looking for is 0.01, which is the last column of the table, so looking in the row for 5 d.f. (see Figure 10.2) we see that



Probability P of lying outside $\pm t$

d.f.	P=0.10	P=0.05	P=0.02	P=0.01
1	6.31	12.71	31.82	63.7
2	2.92	4.30	6.96	9.93
3	2.35	3.18	4.54	5.84
4	2.13	2.78	3.75	4.60
5	2.02	2.57	3.36	4.03
6	1.94	2.45	3.14	3.71

Figure 10.2: Excerpt from the official t table, p. 21.

the appropriate value of t is 4.03. Thus, we can be 99% confident that the patient's true average phosphate level is between 4.3mg/dl and 6.5mg/dl. Note that if we had known the SD for the measurements to be 0.67, instead of having estimated it from the observations, we would have used $z = 2.6$ (corresponding to a one-sided probability of 0.995) in place of $t = 4.03$, yielding a much narrower confidence interval.

Summary

If you want to compute an $\alpha \times 100\%$ confidence interval for the population mean of a normally distributed population based on n samples you do the following:

- (1). Compute the sample mean \bar{x} .
- (2). Compute the sample SD s .
- (3). Look on the table to find the number t in the row corresponding to $n - 1$ degrees of freedom and the column corresponding to α .
- (4). The confidence interval is from $\bar{x} - st/\sqrt{n}$ to $\bar{x} + st/\sqrt{n}$. In other words, we are $\alpha \times 100\%$ confident that μ is in this range.

10.1.3 Using t for Hypothesis tests

We continue Example 10.2. Suppose 4.0 mg/dl is a dangerous level of phosphate, and we want to be 99% sure that the patient is, on average, above that level. Of course, all of our measurements are above that level, but they are also quite variable. It could be that all six of our measurements were exceptionally high. How do we make a statistically precise test?

Let H_0 be the null hypothesis, that the patient's phosphate level is actually $\mu_0 = 4.0$ mg/dl. The alternative hypothesis is that it is a different value, so this is a two-sided test. Suppose we want to test, at the 0.01 level, whether the null hypothesis could be consistent with the observations. We compute

$$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = 5.15.$$

This statistic T has the t distribution with 5 degrees of freedom. The critical value is the value t such that the probability of $|T|$ being bigger than t is 0.01. This is the same value that we looked up in Example 10.2, which is 4.03. Since our T value is 5.15, we reject the null hypothesis. That is, T is much too big: the probability of such a high value is smaller than 0.01. (In fact, it is about 0.002.) Our conclusion is that the true value of μ is not 4.0.

In fact, though, we're likely to be concerned, not with a particular value of μ , but just with whether μ is too big or too small. Suppose we are concerned to be sure that the average phosphate level μ is really *at least* $\mu_0 = 4.0$ mg/dl. In this case, we are performing a one-sided test, and we will reject T values that are too large (meaning that \bar{x} is too large to have plausibly resulted from sampling a distribution with mean μ_0). The computation of T proceeds as before, but now we have a different cutoff, corresponding to a probability twice as big as the level of the test, so 0.02. (This is because of the peculiar way the table is set up. We're now only interested in the probability in the upper tail of the t distribution, which is 0.01, but the table is indexed according to the total probability in both tails.) This is $t = 3.36$, meaning that we would have been more likely to reject the null hypothesis.

10.1.4 When do you use the Z or the T statistics?

When testing or making a confidence interval for the population mean,

- If you know the population variance, use Z .
- If you estimate the population variance from the sample, use T .
- Exception: Use Z when estimating a proportion.

- Another exception: If the number of samples is large there is no difference between Z and t. You may as well use Z, which is conceptually a bit simpler. For most purposes, $n = 50$ is large enough to do only Z tests.

10.1.5 Why do we divide by $n - 1$ in computing the sample SD?

This section is not examinable. The population variance is defined to be the average of the squared deviations from the mean (and the SD is defined to be the square root of that):

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

Why is it, then, that we estimate variance and SD by using a sample variance and sample SD in which n in the denominator is replaced by $n - 1$?

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

The answer is that, if x_1, \dots, x_n are random samples from a distribution with variance σ^2 , then s_x^2 is a better estimate for σ^2 than is σ_x^2 . Better in what sense? The technical word is “unbiased,” which simply means that over many trials it will turn out to be correct. In other words, σ_x^2 is, on average a bit too small, by exactly a factor of $(n - 1)/n$. It makes sense to expect it to be too small, since you would expect

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

to be just right, on average, if only we knew what μ was. Replacing μ by the estimate \bar{x} will make it smaller. (In fact, for any numbers x_1, \dots, x_n , the number a that makes $\sum (x_i - a)^2$ as small as possible is $a = \bar{x}$. Can you see why?)

As an example, consider the case $n = 2$, and let X_1, X_2 be two random

choices from the distribution. Then $\bar{X} = (X_1 + X_2)/2$, and

$$\begin{aligned}\sigma_X^2 &= \frac{1}{2} ((X_1 - \bar{X})^2 + (X_2 - \bar{X})^2) \\ &= \left(\frac{X_1 - X_2}{2} \right)^2 \\ &= \frac{1}{4} ((X_1 - \mu) + (\mu - X_2))^2 \\ &= \frac{1}{4} [(X_1 - \mu)^2 + (\mu - X_2)^2 + 2(X_1 - \mu)(\mu - X_2)].\end{aligned}$$

How big is this on average? The first two terms in the brackets will average to σ^2 (the technical term is, their *expectation* is σ^2), while the last term averages to 0. The total averages then to just $\sigma^2/2$.

10.2 Paired-sample t test

A study² was carried out to study the effect of cigarette smoking on blood clotting. Some health problems that smokers are prone to are a result of abnormal blood clotting. Blood was drawn from 11 individuals before and after they smoked a cigarette, and researchers measured the percentage of blood platelets — the factors responsible for initiating clot formation — that aggregated when exposed to a certain stimulus. The results are shown in Table 10.2.

We see that the “Before” numbers tend to be larger than the “After” numbers. But could this be simply a result of random variation? After all, there is quite a lot of natural variability in the numbers.

Imagine that we pick a random individual, who has a normally distributed Before score X_i . Smoking a cigarette adds a random effect (also normally distributed) D_i , to make the After score Y_i . It’s a mathematical fact that, if X and D are independent, and $Y = X + D$, then

$$\text{Var}(Y) = \text{Var}(X) + \text{Var}(D).$$

We are really interested to know whether D_i is positive on average, which we do by comparing the observed average value of d_i to the SD of d_i . But when we did the computation, we did not use the SD of d_i in the denominator; we used the SD of x_i and y_i , which is much bigger. That is, the average difference between Before and After numbers was found to be not large

²[Lev73], discussed in [Ric95].