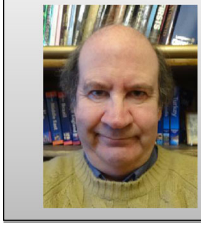
Published online in Wiley Online Library: 7 June 2015

(wileyonlinelibrary.com) DOI: 10.1002/cem.2713

The *t*-distribution and its relationship to the normal distribution



by Richard G. Brereton

Journal of Chemometrics Columnist

The *t*-distribution is used as an alternative to the normal distribution when sample sizes are small in order to estimate confidence or determine critical values that an observation is a given distance from the mean. It is a consequence of the sample standard deviation being a biased or underestimate (usually) of the population standard deviation.

> *Many early statistical studies were performed using limited sample sizes so it was necessary to develop approaches to deduce more information about the underlying population: even modern applications such as process control and metabolomics often suffer from modest sample sizes.*

Previously [1], we distinguished between samples and populations and looked at how the statistics obtained from a sample can be used to estimate parameters in a population.

DISTRIBUTIONS

In chemometrics, we are also concerned with distributions. We have already discussed the normal and chi-squared distributions [2,3]. These, however, are usually employed to describe populations. When sample sizes are relatively small, we find that these ideal distributions can be distorted. Many early statistical studies were performed using limited sample sizes so it was necessary to develop approaches, not just for estimating means and standard deviations, for using these datasets to deduce more information about the underlying population. Even in modern studies, applications such as process control and metabolomics often suffer from modest sample sizes.

WILLIAM GOSSETT

The *t*-distribution was developed for such situations. The first recognisable description in the literature was credited to Frederick R. Helmert in the late 19th century [4]. However, its current and more widely accepted description is due to William Gossett [5] in 1908. He published under the pseudonym 'Student' because he was working for the Guinness brewery and they did not, according to legend, allow him to use his own name. An alternative explanation is that they did not want rivals to know they were using this method. Although modern statistical ideas had many of their origins in the 19th century, our modern way of thinking was formulated primarily in the first few decades of the 20th century, especially due to the advocacy of R. A. Fisher and collaborators; hence, the terminology in common use is based primarily on Gossett's work, which was taken up by Fisher [6]. The *t*-distribution is also sometimes called Student's distribution or Student's *t*-distribution to reflect the pseudonymous authorship of this original paper.

BIASED ESTIMATOR

The basis of why we need to use this distribution is as follows.

- We will consider a univariate set of measurements.
- It assumes that the underlying population is normally distributed.
- When sampling, it is common to calculate the sample standard deviation, as described in the previous article, by $s = \sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2/(n-1)}$.
- The sample standard deviation is a biased estimator of the population standard deviation as on the whole it is more likely than not to be less than the population standard deviation.
- This bias depends on the sample size and underlying population distribution. There is no general expression for how much the bias is independent of the shape of the distribution, but we assume it is normal.
- Let us then imagine a situation in which the underlying population standard deviation is 2 units but the sample standard deviation is calculated to be 1.8. Hence, a measurement of 0.9 greater than the mean is incorrectly estimated to be 0.5 standard deviations greater than the mean, rather than 0.45.
- So there will appear a lower probability that a measurement is less than 0.5 units greater than the mean than for a normal distribution.
- This new distribution will have a lower probability density in the centre and bigger probability density in the wings than a normal distribution, and so is of different shapes to the normal distribution, and somewhat flatter.

DEGREES OF FREEDOM

- This new distribution is called the *t*-distribution. The smaller the sample size, the more it differs from the normal distribution.
- We usually talk about degrees of freedom, which are often denoted by ν, and equals $n-1$ where n is the sample size. So if we have a sample size of 8, there are 7 degrees of freedom.

- The shape of the *t*-distribution depends on ν.
- Although there may appear minor differences in the centre of the distribution compared with the normal distribution, the wings are often quite different in shape. Because many decisions, for example, in process control, are at 95% or 99% limits, and sample sizes can be quite small, there can be very different conclusions according to whether the normal or *t*-distribution is used.
- It is quite a dilemma that most observations, when sample sizes are small, are in the centre of a distribution as this is where most observations will lie, but most decisions are at the wings.
- It is important to understand that even though underlying measurements are still distributed normally, because of the problem in estimating the population standard deviations, small sample sizes will appear to be best modelled by *t*-distribution.
- It is also important to understand that while a normal distribution can often be modelled well when there are many measurements, an experimental *t*-distribution cannot easily be obtained, because we would need a large sample size to adequately model its shape. Hence, a *t*-distribution really is just an underlying statistical estimate of the probability of finding data at different distances from the mean relative to the sample standard deviation. This distance is called the *t*-statistic.

The *t*-distribution has several interesting properties.

- The mean of the distribution is equal to 0.
- The variance is equal to $\nu/(\nu - 2\ )$, if $\nu > 2$.
- The variance is always greater than 1, although it is close to 1 when there are many degrees of freedom.
- With infinite degrees of freedom, the *t*-distribution is the same as the standard normal distribution.

NORMAL DISTRIBUTION

The probability distribution function of the *t*-distributions for varying degrees of freedom is illustrated in Figure 1. Although the *t*-distributions do superficially appear quite similar to the normal distribution once ν is greater than 5, actually the proportion of measurements within 1.96 standard deviations on either side of the mean, which should be 95% for a normal distribution, differs noticeably even with there are 10 degrees of freedom.

Check this as follows.

- Calculate 2*(NORM.S.DIST(1.96,TRUE)) − 1 in Excel. This gives the proportion of data expected to be within 1.96 standard deviations of the mean using the normal distribution, which equals 0.950. It is two tailed that represent data on either side of the mean. This is represented in the blue area in Figure 2.
- Calculate 1 − (T.DIST.2T(1.96,10)), which equals 0.921. Note that perversely the syntax for the *t*-distribution in Excel is quite different to that for the normal distribution. The expression T.DIST.2T represents the two-tailed area of the *t*-distribution greater than 1.96 standard deviations from the mean for 10 degrees of freedom, in our case, or the white area in Figure 2. For the *t*-distribution, it is assumed that the standard deviation is 1 and mean is 0, whereas for the normal distribution functions, it is possible to specify these, or else in more recent versions of Excel that alternatively use the .S. in the middle of the expression.
- This implies that for the normal distribution, around 5% of measurements are expected to be more than 1.96 standard deviations from the mean, whereas for the corresponding *t*-distribution with 10 degrees of freedom, it is around 8% using the sample distribution.
- When the number of degrees of freedom reduces still further, the difference between the *t*-distribution and normal distribution is quite considerable. For 5 degrees of freedom, the corresponding value equals 0.893, so that it now implies that over 10% of measurements are expected to be further than 1.96 sample standard deviations from the mean.
- Hence, when there are small sample sizes, there can be considerable differences in, for example, 95% confidence limits between the normal and *t*-distributions.
- To check this, use the NORM.S.INV and the T.INV functions in Excel. Strangely for the inverse distribution (determining the number of standard deviations away from the mean corresponds to a given probability), both the normal and *t*-distributions have similar syntax, except for the .S., which specifies that the data are standardised for the normal distribution, which is assumed by default for the *t*-distribution. Use a probability of 0.975 rather than 0.95 as this represents a 0.025 area on either side of the mean (two tailed).
- For the normal distribution, the answer is 1.960 as expected. For the *t*-distribution and 2 degrees of freedom, it is 4.303, 5 degrees of freedom 2.571 and 10 degrees of freedom 2.228.
- When the number of degrees of freedom is large, then the *t*-distribution, of course, converges to the normal distribution.
- Note that if we knew the population standard deviation, even if the sample size was small, we would not need to use the *t*-distribution. However, this situation would be very unusual (it is possible if the mean were unknown but the standard deviation was known, for example, if we were to be interested in a shift in the mean).

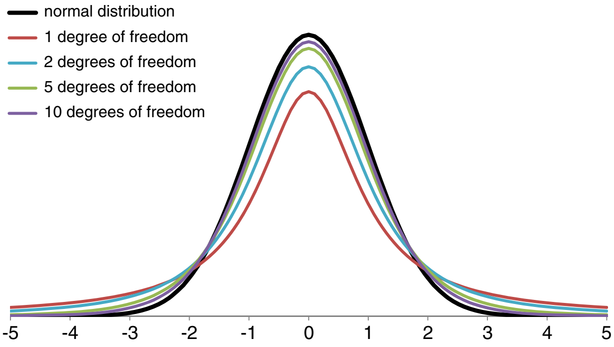


Figure 1. The probability density function for the normal distribution together with the *t*-distributions for different degrees of freedom. The horizontal axis represents standard deviations: for the normal distribution, this is the population standard deviation and for the *t*-distribution, the sample standard deviation.

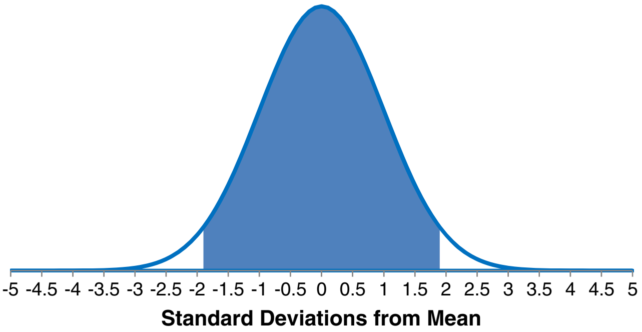


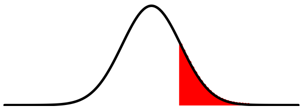
Figure 2. Two tailed probabilities: the blue area represents the probability that a measurement is within 1.96 standard deviations from the mean.

Although it is quite easy to calculate the probability of a sample being at any specified number of standard deviations from the mean in Excel or Matlab or most computational packages, it is still common to use statistical tables.

STATISTICAL TABLES

Table I is a typical table. Unlike for the normal distribution, most tables just provide critical values of the test statistic. There could be a full table for different degrees of freedom of *t*, each degree of freedom generating a table analogous to the normal distribution, but this would fill up textbooks with superfluous information. Tables of critical values are either one tailed or two tailed. A one-tailed *t* table presents critical values greater than a certain distance (in units of sample standard deviations) from the mean. In Table I, we see that for 5 degrees of freedom and a one-tailed test and a critical value of 0.05, the *t* value is 3.747. This can be interpreted as follows.

Table I. One tailed *t*-distribution critical points: columns represent probabilities and rows represent degrees of freedom



	Probability		
Degree of freedom	0.05	0.01	0.005
1	6.314	31.821	63.656
2	2.920	6.965	9.925
3	2.353	4.541	5.841
4	2.132	3.747	4.604
5	2.015	3.365	4.032
6	1.943	3.143	3.707
7	1.895	2.998	3.499
8	1.860	2.896	3.355
9	1.833	2.821	3.250
10	1.812	2.764	3.169
11	1.796	2.718	3.106
12	1.782	2.681	3.055
13	1.771	2.650	3.012
14	1.761	2.624	2.977
15	1.753	2.602	2.947
16	1.746	2.583	2.921
17	1.740	2.567	2.898
18	1.734	2.552	2.878
19	1.729	2.539	2.861
20	1.725	2.528	2.845
25	1.708	2.485	2.787
30	1.697	2.457	2.750
35	1.690	2.438	2.724
40	1.684	2.423	2.704
45	1.679	2.412	2.690
50	1.676	2.403	2.678
55	1.673	2.396	2.668
60	1.671	2.390	2.660
65	1.669	2.385	2.654
70	1.667	2.381	2.648
80	1.664	2.374	2.639
90	1.662	2.368	2.632
100	1.660	2.364	2.626
∞	1.645	2.327	2.576

- Of measurements, 5% are expected to be greater than 3.747 sample standard deviations above the mean.
- Hence, we expect 95% of measurements to be less than this value.
- If we prefer a two-tailed *t* value, we simply double the proportion of measurements.
- Hence, we expect 10% of measurements to be 3.747 or more sample standard deviations away from the mean and 90% within 3.747 sample standard deviations of the mean.

Always check whether a table represents one-tailed or two-tailed probabilities. Note that as the number of degrees of freedom increases, the critical values start to resemble those for the normal distribution. There are several more comprehensive tables available on the Web of which we reference one [7].

CRITICAL LIMITS

Whereas the distinction between *t*-statistic and normal statistics may not appear very great at first, relatively modest samples sizes could make a substantial difference in decision-making, for example, in areas such as process control with decision limits (critical limits) at the 95% or 99% level. This is especially true in cases where there may be many other factors at play, such as whether samples do indeed fall into an underlying normal distribution or whether sampling is representative or random (both different concepts, which we will explore in later articles).

MULTIVARIATE *T*-DISTRIBUTION

Finally, it is important to note that there is a multivariate *t*-distribution in analogy to the multivariate normal distribution. We will not dwell further on this as it is rarely encountered in chemometrics, with the more widespread *F*-distribution being described in the next article.

Richard G. Brereton

School of Chemistry, University of Bristol, Cantock's Close, Bristol,
BS8 1TS, UK.
E-mail: richard.brereton1@bris.ac.uk

REFERENCES

1. Brereton RG. Populations and samples. *J. Chemom.* 2015. DOI: 10.1002/cem.2692.
2. Brereton RG. The normal distribution. *J. Chemom.* 2014; **28**: 789–792.
3. Brereton RG. The chi squared and multinormal distributions. *J. Chemom.* 2014; **29**: 9–12.
4. Helmert FR. Über die Bestimmung des wahrscheinlichen Fehlers aus einer endlichen Anzahl wahrer Beobachtungsfehler. *Z. Math. Phys.* 1875; **20**: 300–303.
5. Student. The probable error of the mean. *Biometrika* 1908; **6**: 1–25.
6. Fisher RA. Applications of "Student's" distribution. *Metron* 1925; **5**: 90–104.
7. Easycalculation.com. T distribution critical values table, http://easycalculation.com/statistics/t-distribution-critical-value-table.php