Journal of CHEMOMETRICS

Published online in Wiley Online Library: 19 November 2015

(wileyonlinelibrary.com) DOI: 10.1002/cem.2763

Hotelling's *T* squared distribution, its relationship to the *F* distribution and its use in multivariate space

Hotelling's *T* squared statistic is used in multivariate datasets as a measure of distance from the centre of a distribution and follows the *F* distribution. Its use is compared with chi squared, and it is shown that for modest sample sizes, quite different conclusions may be drawn about whether observations are outliers. The low probability of finding observations in the centre of a multivariate distribution is also described.

In areas such as multivariate statistical process control where training sets may be relatively modest, it is important to understand that using chi square or Hotelling's T squared *can make a tremendous difference in decision-making.*

by Richard G. Brereton

Journal of Chemometrics Columnist

A major job for the chemometrician is to see how far an observation (such as a spectrum or chromatogram) is from the centre of an experimentally observed distribution (often called a training set) and then determine the probability that it is a member of the underlying population. By probability, we really mean how likely we estimate that an observation (represented by a datapoint) that is far away from the centre would be observed if we sampled the underlying population. If this probability is low, it is unlikely that it is a member of our population and may be an outlier or a member of a separate group, according to what we are testing. In different application areas, this probability can also be referred as confidence, significance or probability of rejecting the null hypothesis. These related concepts will be discussed in later articles and when what it is called may depend on who you are, for example, an analytical chemist, control engineer or statistician.

UNIVARIATE POPULATIONS

For a univariate population, we can use the normal (or *z*) distribution, or if sample sizes are small, the *t* distribution: just measure the distance from the centre (or mean) in units of standard deviations and use Excel or MATLAB or tables of critical values. When more than one variable is recorded, for multivariate datasets, the chi squared distribution is analogous to the normal distribution, and the number of Mahalanobis distances from the centre is used as the multivariate analogy to the standard deviation.

SMALL SAMPLE SIZES

If sample sizes are small, the approximation that the chi squared distribution represents a good estimation may break down, just as the normal distribution may not be an adequate model for small sample sizes. Remember that although the distributions may be similar in the centre (or close to the mean) for the same number of variables, they can differ considerably at the wings, where most important decisions are made. Although the *F* distribution is often thought of as the multivariate analogy to the *t* distribution, it is not quite so simple. During the development of modern multivariate methods in the early 20th century, new statistics were developed according to their most common application. In fact, the *F* statistic, as originally employed, was primarily used to assess the significance of the ratio between two variances, often via analysis of variance. Although we will discuss this application in later articles, our main concern at this juncture is how to estimate the probability that an observation, represented by a point at a given distance from the centre of distribution (as measured by the Mahalanobis distance), is a member of a parent distribution (or more precisely how frequently we expect a measurement that is far from the mean to turn up if it is a member of our candidate population).

HAROLD HOTELLING

Harold Hotelling was a US professor of statistics active mainly in the 1930s and 1940s and first described his measure in 1931 [1]. It is widely used in multivariate statistics to relate the distance from the mean of a dataset to a probability or confidence level.

UNIVARIATE STATISTICS

For univariate datasets, the squared distance from the mean in units of the sample standard deviation (or t^2) is the square of *t* and has the same distribution as $F(1,v_2)$ (see [2] for further definitions) for a sample size $n = v_2 + 1$. This can be checked in Excel.

- Consider the proportion of observations expected to be within two sample standard deviations of the mean.
- In Excel type (T.DIST(2,10,TRUE)-0.5)*2, the answer is 0.926. Remember we need to look at the proportion on either side of the mean.
- The same answer is obtained if we use F.DIST(4,1,10,TRUE), replacing 2 by its square.

If we have a very large sample size, for example, 1001, all the statistics (normal, chi squared (using z^2), *t* and *F* (using t^2)) give the same result for a univariate dataset.

- Try F.DIST(4,1,1000,TRUE), (NORM.S.DIST(2,TRUE)-0.5)*2, (T.DIST(2,1000,TRUE)-0.5)*2 and CHISQ.DIST(4,1,TRUE) and the result should be 0.954. Similar expressions can be obtained in MATLAB; this is a good check that you have got the right expressions.

MULTIVARIATE

By this argument, we expect the chi squared distribution and the *F* distribution to give the same results if we increase the number of variables, for large sample sizes. We will see this is not in fact so.

- Try F.DIST(4,2,1000,TRUE) and CHISQ.DIST(4,2,TRUE), which is the case when there are two variables and we find that the two results now differ, the *F* distribution gives a probability of 0.981 and chi squared gives 0.864.
- Hence, when $v_1 > 1$ (or $k > 1$, i.e. there is more than one variable or more than one column in a datamatrix), the values of chi squared and *F* are different.
- This is because we are using the squared Mahalanobis distance as our statistic.

HOTELLING'S *T* SQUARED

- In order to obtain a similar result, we have to scale the squared Mahalanobis distance to give a new statistic called Hotelling's T^2, which we will denote by *H*.
- We define $H = (n-k)\,\Delta^2/(k\,(n-1))$, where *n* is the sample size and *k* the number of variables and Δ is the Mahalonobis distance. This involves scaling the squared Mahalanobis distance by a factor $(n-k)/(k\,(n-1))$ to get a new measure.
- This new statistic now follows the *F* distribution with $v_1 = k$ and $v_2 = n-k$.
- Note that when $k = 1$, the expression for *H* equals $(n-1)\,\Delta^2/(1\,(n-1)) = \Delta^2$, that is, we can use the squared Mahalanobis distance for univariate data.
- Now, try this out on our example previously and use the expression F.DIST(4*998/(2*999),2,1000,TRUE) in Excel and we find that the answer is 0.864 to three significant figures.

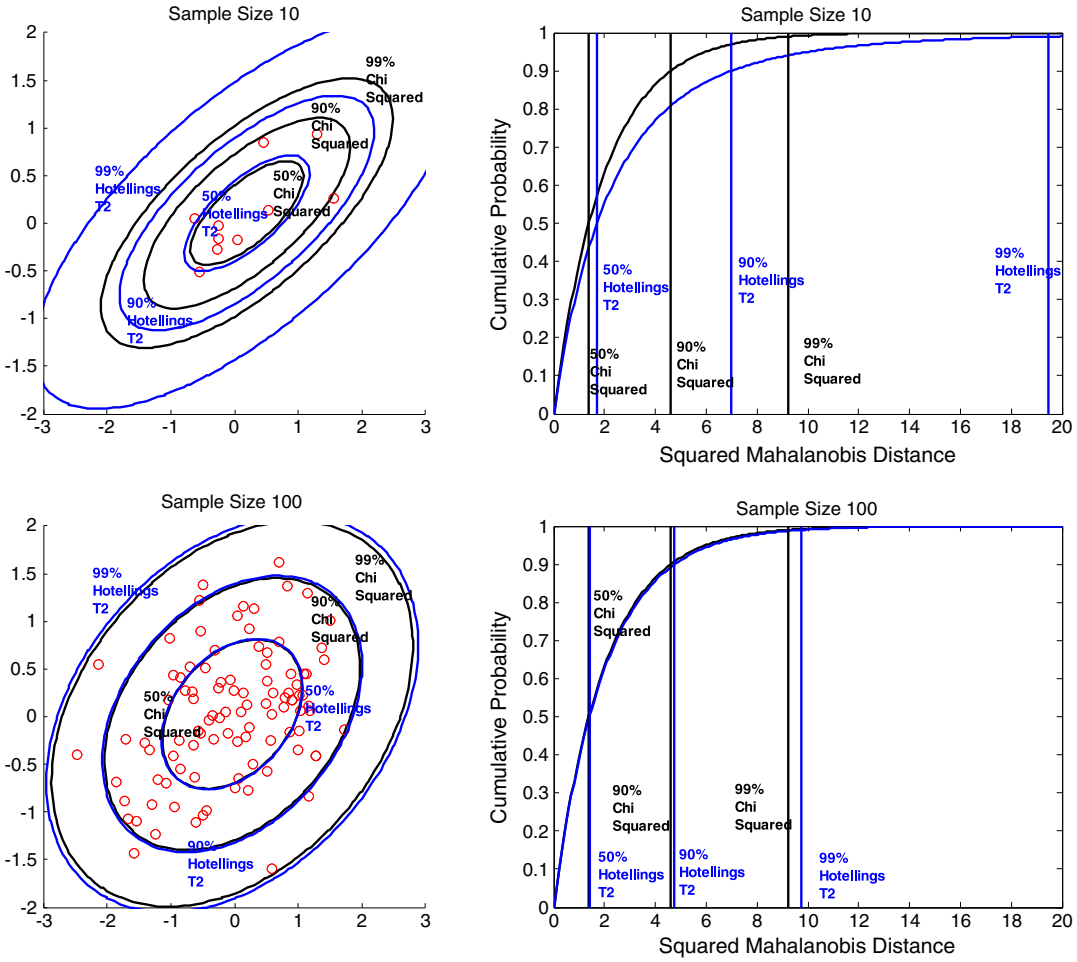

Figure 1. A total of 50%, 90% and 99% confidence limits (corresponding to $p = 0.5$, 0.1 and 0.01) for datasets consisting of sample sizes of 10, assuming an underlying normal distribution, using both chi squared and Hotelling's T^2 statistics.

- If the sample size is large, then the approximation is that $H = \Delta^2/k$ (and in practice, we are unlikely to notice the difference between the expressions even in the wings for modest sample sizes) so in our case, we can simplify the expression in Excel to F.DIST(4/2,2,1000,TRUE).
- This scaling is not necessary for chi squared. It is a consequence of historic origins of F and chi squared and has no fundamental statistic importance – it is a bit like measuring a distance in different units; we may have a recipe book that uses fluid ounces but a measuring cylinder that uses litres. Famously, a NASA Mars orbiter was lost precisely because of such a mishap, when Lockheed used imperial (English) units and NASA metric units [3]. Neither side was wrong in isolation, and it was only when the two measures were put together that the disaster happened.

Hence, Hotelling's T^2 follows the F distribution and can therefore be used as a means of converting the Mahalanobis distance from a centroid to a probability of belonging to a predefined multivariate distribution. As an approximation, this statistic equals the squared Mahalanobis distance from the mean divided by the number of variables unless sample sizes are small.

CHI SQUARED

The different conclusions that can be obtained using Hotelling's T^2 compared with chi squared can be visualised in Figure 1. Two datasets, one with sample size 10 and the other 100, are illustrated. It can be seen that the 50% confidence limits are not very different for either sample size. However, for the 99% confidence limits, there is a very considerable difference between chi squared and Hotelling's T^2 when the sample size equals 10. This can be shown by looking at the cumulative distribution function that approaches 1 considerably more slowly using the T^2 rather than the chi squared statistic when sample sizes are small.

MULTIVARIATE STATISTICAL PROCESS CONTROL

In areas such as multivariate statistical process control where training sets may be relatively modest, it is important to understand that using these two different measures can make a tremendous difference in decision-making. This difference is more pronounced, the more the variables. Low sample to variable ratios is often found in chemometrics, when it is particularly important to be aware of which type of statistics is employed. However, one word of warning is that often, we are not certain that a dataset arises from a population that is normally distributed and it is rare that enough observations are recorded to model the wings well yet most decisions are made at the wings, which can differ substantially between different distributions.

THE CENTRE

An interesting observation for multivariate datasets is that if there are three variables or more, the highest probability of finding observations is not at the centre. This can be illustrated by graphs of the F and chi squared distributions as presented in previous articles. This may be counterintuitive, but sometimes, statistical results are just that. A simple explanation is to envisage data being distributed on the (hyper)surface or a (hyper)sphere. Although the density is still the most in the centre, the surface of the sphere increases faster (initially); then, the density decreases, the two effects interacting. Hence, the most 'typical' observation is not one in which all variables are at their central values.

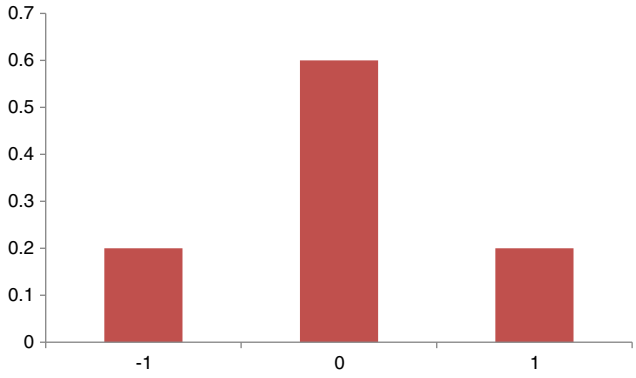

Figure 2. A simple distribution where a number can take one of the three possible values, with 0 being the central value.

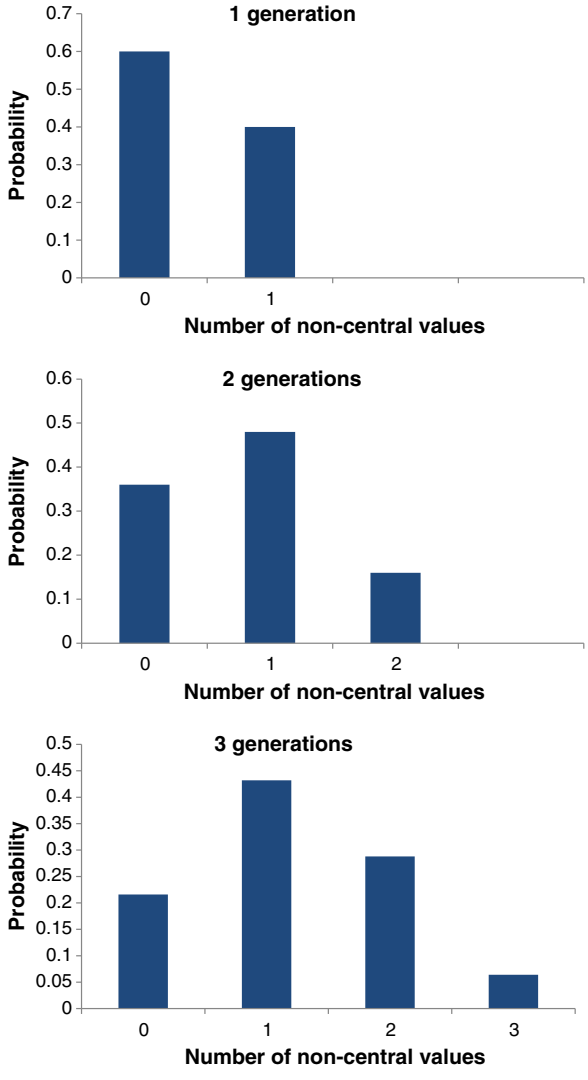

Figure 3. Change in distribution as a number with probability 0.6 having a central value of 0, and 0.2 having values of +1 or −1, is generated a different number of times.

As an example, we take three unrelated measurements. In England, the mean male weight is around 83 kg, the mean male earnings are £28000 per year and a man is caught in average on closed-circuit television cameras 300 times a day. Assuming distributions are symmetrical, does this mean a man with these three characteristics together is most 'typical'? In fact, this combination, providing the characteristics are not correlated, is very unlikely, and the 'typical' man will not be found in the centre of the distribution.

A simple numerical example illustrates the following:

- Consider a number that can take one of the three values, −1, 0 and +1 and that the probabilities of each value is 0.2, 0.6 and 0.2, respectively. Its pdf is illustrated in Figure 2.
- Generate this number twice. There are nine possible combinations, namely,
  {1 1} {1 0} {1 +1} {0 1} {0 0} {0 +1} {+1 1} {+1 0} and {+1 +1}.
- It is assumed that each time the number is generated is independently of the last time.
- Now, count the number of ways of generating 2 central readings (1), 1 central reading (4) and 0 central readings (4). Consider the number of 'non-central' readings as a distance from the centre so a result of {0 − 1} is considered to be 1 'reading' from the centre.
- The probabilities that the result is 0, 1 or 2 'readings' from the centre are 0.36, 0.48 and 0.16 in our case. For the interested readers, these can be obtained via the binomial distribution, but we will not present the calculation in this article.
- When we generate the number three times, the probabilities become 0.216, 0.432, 0.288 and 0.064, respectively, for four possible results.
- We can visualise the change in distribution as the number is generated one, two and three times, in Figure 3 and see that in this case, the majority of observations are found in the centre only if the number is generated once.
- Of course, the change in appearance depends on the underlying probability distribution and the number of times the number is generated, just like the *F* distribution.
- It is important to remember that this depends on each generator (equivalent to a variable) being independent, however the Mahalanobis distance measure is analogous to performing principal components analysis (and standardising) before measuring distances and will break any dependence but may reduce the dimensionality of the original dataset.

Richard G. Brereton

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

REFERENCES

1. Hotelling H. The generalization of Student's ratio. *Ann. Math. Stat.* 1931; **2**: 360–378.
2. Brereton RG. The F distribution and its relationship to the chi squared and t distributions. *J. Chemometr.* DOI:10.1002/cem.2734.
3. NASA's metric confusion caused Mars orbiter loss, CNN, September 30, 1999, http://edition.cnn.com/TECH/space/9909/30/mars.metric/index.html.