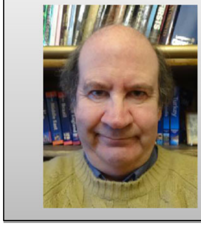
The chemometrics column

Journal of CHEMOMETRICS

Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/cem.2680

The chi squared and multinormal distributions



by Richard G. Brereton

Journal of Chemometrics Columnist

The first description of the chi squared (or χ^2) distribution was by the German mathematician Frederick R. Helmert in the late 19th century[1], although it was not clearly formulated in statistics until Karl Pearson developed the chi squared test in 1900, who independently rediscovered the distribution [2]. The modern name and statistical terminology arises primarily from Pearson's work.

> *A series of measurements may fall into a multivariate normal distribution but not a chi squared distribution, unless they are independent.*

NORMAL DISTRIBUTION

When there is one variable measured, the chi squared distribution can be very simply related to the normal distribution. Chi squared is simply the square of the distance to the mean. If the underlying observations are normally distributed, we obtain the chi squared distribution. Therefore, for example, if data are standardised (that is measurements are divided by the standard deviation after subtracting the mean), the proportion of observations less than four units of chi squared from the mean is the same as the proportion of observations, within two normalised units either side of the mean.

This can be illustrated by a simple example in Excel.

- Consider the proportion of observations within 1.5 standard deviations either side of the mean.
- Assuming an underlying normal distribution, this can be given by (NORMDIST.S(1.5.TRUE)-0.5) * 2 in Office 2010 (there are some variations in earlier versions of Excel, which we will not discuss for brevity).
- To understand this syntax, NORM.DIST.S(1,5,TRUE) means calculating the cumulative probability or expected proportion (as indicated by 'TRUE') less than 1.5 standard deviations above the mean, and 'S' stands for standardised, or mean of 0 and standard deviation of 1. Subtract 0.5 as we are interested in the observations between the mean and 1.5 standard deviations above it. It is multiplied by 2, as observations must be either side of the mean.
- This equals 0.866 or 86.6%
- Now calculate CHISQ.DIST(1.5^2,1,"TRUE")
- Rather irritatingly, the chi squared functions in Excel all assume a standard deviation of 1 and mean of 0, whereas for the normal distribution functions, you need the S to specify that the function is standardised (or else the mean and standard deviation could be entered explicitly). The 1 represents the number of degrees of freedom, which will be discussed in more detail in the succeeding text and TRUE indicates that this is a cumulative distribution function (cdf).
- You will find that they have the same answer.

This is illustrated in Figure 1.

MORE THAN ONE VARIABLE

The chi squared distribution becomes important in chemometrics when more than one variable is measured. For the normal distribution, we can look at how far a datapoint is from the mean, and this may be used to determine whether it is an outlier or whether it belongs to a predefined group. If the underlying distribution is assumed to be normal, a probability can be attached as discussed in *The normal distribution* [3]. We would like to perform a similar calculation in cases where there is more than one variable measured. Once the number of variables exceeds 1, there is no specific positive or negative direction to the mean or centre of the data, so instead, a squared

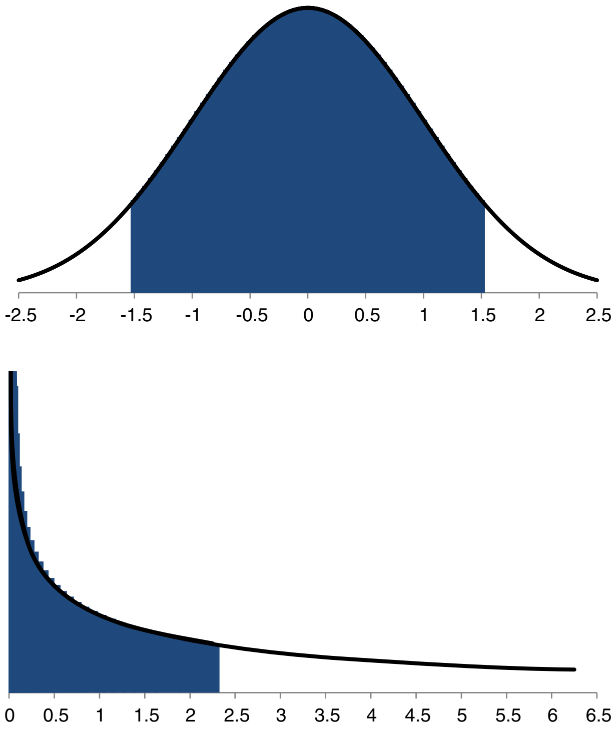


Figure 1. Normal distribution (top) and corresponding Chi squared distribution (bottom). The proportion area under the normal distribution curve between −1.5 and +1.5 standard deviations from the mean is the same as the area under the Chi squared distribution curve between 0 and 2.25 standard deviations from the mean (86.6%).

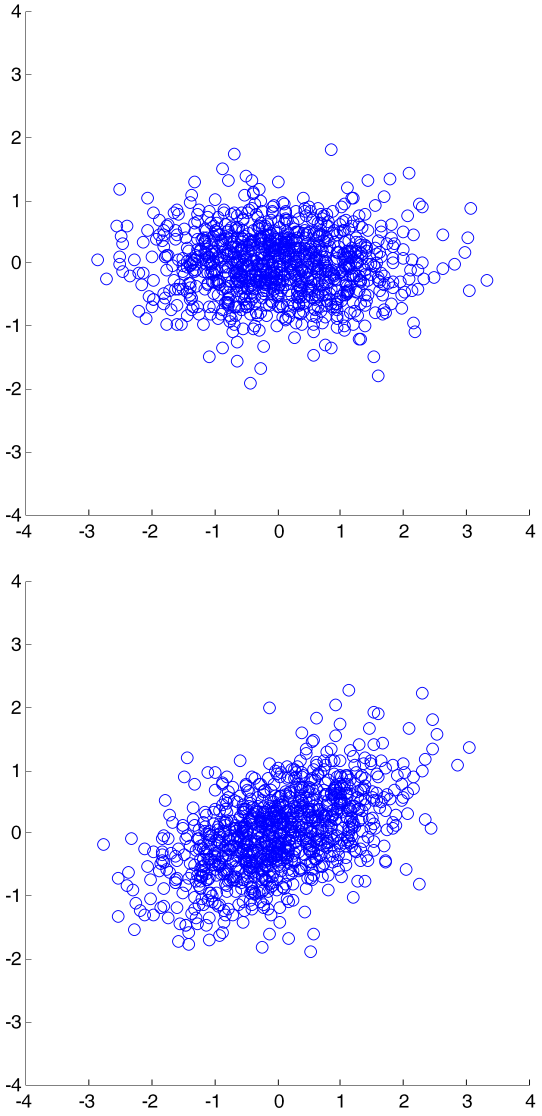


Figure 2. Two multinormal distributions. In the top one, both variables are independent, but in the bottom, they are not.

distance must be calculated. We can see that when one variable is recorded, if it is distributed normally, its square is distributed according to chi squared.

MULTIVARIATE NORMAL DISTRIBUTION

Before we return to chi squared using more than one variable, it is necessary first to understand the multivariate normal distribution.

If k variables are measured for a series of observations, then for the observations to form a multivariate normal distribution there must be the following conditions.

- Each variable in itself must be distributed normally.
- The sum of any combination of variables must also be distributed normally. In algebraic terms, if variable 1 is distributed normally, we call this z_1, then $az_1 + bz_2$ ....$+ nz_k$ must also form a normal distribution, where a, b and so on can be any number.
- In the vast majority of cases, if each variable on its own is normally distributed, the combination also forms a multivariate normal distribution, although there are a few specialised exceptions.

In the definition used above, it is not necessary for the variables to be independent or uncorrelated (we will discuss the difference between the two in a later article). Figure 2 illustrates two bivariate multinormal distributions. In the top one, both variables are independent, but in the bottom one, they are not. It so happens that the bottom one is simply the rotation of the top one. However, it is important to understand that both distributions satisfy the conditions of multivariate normality.

For the normal distribution, we can determine whether an observation is likely to belong to a predefined population, according to how many standard deviations it is away from the mean. If the probability is 5%, then we only expect to find an observation that far from the mean one 1 time in 20. Likewise for a multinormal distribution, the distance from the mean or centre of the measurements can be used to determine the likelihood of membership of a predefined population. Because distance in more than one dimension does not have a predefined sign, we have to use squared distances. Can we use the squared distance from the mean of a dataset to tell us how likely an observation is to be part of a predefined population, analogous to the one variable case?

CHI SQUARED DISTRIBUTION

Here is where the chi squared distribution comes into play. If more than one variable is recorded, then do the measurements fall into a chi squared distribution?

- The square of each individual variable must fall into an underlying chi squared distribution.
- So must $z_1^2 + z_2^2 + \ ....z_N^2$ using the same notation as in the preceding text.
- Note that there is no multiplier before the zs.
- Note, however, a further condition, that the variables must be independent: to form a multinormal distribution, this is not necessary.
- The math to prove this is quite complicated, but an intuitive explanation is the one the sum of $x^2 + y^2$ is not the same as $(x + y)^2$ unless the product xy is equal to 0 (which clearly would not be in the one dimensional case; but if these were vectors, this could happen).

Hence, a series of measurements may fall into a multivariate normal distribution but not a chi squared distribution, unless they are independent. So, how can we use chi squared to determine whether an observation belongs to a predefined population by looking at the distance from the centre? When only one variable is measured, it is of course simple.

MAHALANOBIS DISTANCE

A new measurement of distance comes into play, the Mahalanobis distance. This will be discussed in detail the next article, but it breaks the correlation and also puts both variables on the same scale. The Mahalanobis distance when there is more than one variable can be thought analogous to the standard deviation. Figure 3 is of the Mahalanobis distance of 2 (or a squared distance of 4) units from the centre of a bivariate normal distribution.

Now comes the trick. If an underlying distribution is multinormal, we expect the Mahalanobis distances to be characterised by a chi squared distribution. The number of degrees of freedom of the chi squared distribution equals the number of variables. So for two variables, it has 2 degrees of freedom.

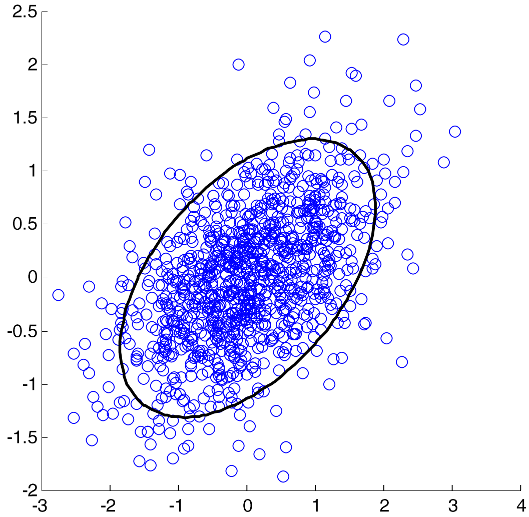


Figure 3. Contours at 2 Mahalanobis distance units from the centre of a two variable multinormal distribution.

- Using Excel, we can predict the number of observations that is outside a defined number of Mahalanobis distances from the centre.
- In the example of Figure 3, there are 1000 observations.
- The percentage predicted to be further than 2 Mahalanobis distance units from the centre is given by CHISQ.DIST.RT(4,2), where 4 is the square of 2 (the number of units, remember this is squared) and 2 is the number of degrees of freedom. Note that the extension RT represents the right-hand area of the chi squared curve, that is, the proportion greater than a given distance from the centre, 1-CHISQ.DIST(4,2,"TRUE") gives the same result, the TRUE meaning a cdf (see *The normal distribution* [3] for definition).
- This is 0.135.
- This means that we expect about 135 out of 1000 observations to be outside this distance.
- By visually inspecting the figure, this number is right (the relationship will not always be exact unless the sample size is very large).

We have used the chi squared distribution with 2 degrees of freedom.

DEGREES OF FREEDOM

As more variables are measured, the shape of the chi squared distribution changes. Figure 4 represents chi squared for several different values of k or number of variables or dimensionality of the variable space or degrees of freedom (sometimes denoted by ν). Remember that the definition of the χ^2 statistic is the squared Mahalanobis distance and that this distribution is only obeyed if the underlying distribution is multinormal (so, one could calculate the squared Mahalonobis distances of samples that are not multinormal, and they would not fall into a chi squared distribution).

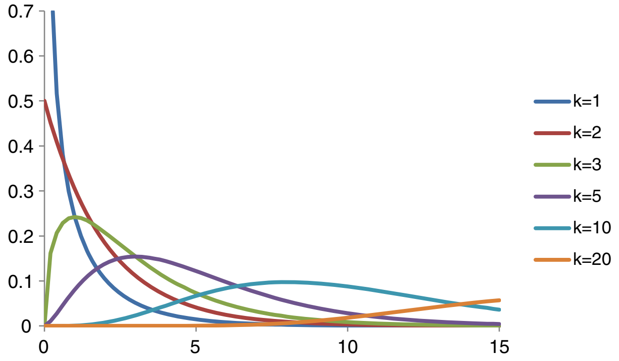


Figure 4. Chi squared distributions for different degrees of freedom (k).

- Note that when k is large, the chi squared distribution resembles a normal distribution.
- Note that once k is greater than or equal to 3, it no longer has a maximum at 0. This means that the centre of the population is no longer the place where it is most likely to find the data. This is because although the density still is the greatest at the centre, the hypersurface of the hypersphere that represents equal Mahalanobis distance from the centre expands faster initially, so it is a trade-off between density and area.

There are other interesting properties.

- The mean of the distribution equals k.
- The variance equals $2\,k$.
- If $k \geq 2$, the maximum value occurs at a value of χ^2, or squared Mahalonobis distance, of $k-2$.

Remember that the shape of the multinormal distribution, for example, as characterised by the sum of the individual measurements for each observation, is the same as the normal distribution no matter how many variables are measured, unlike the chi squared distribution.

TABLES

In many traditional books, the chi squared distribution is often presented in tabular form. A typical table is presented in Table I, although there are other ways of presenting it. The rows represent the number of degrees of freedom or variables in our case and the columns the cumulative probability. The columns are often called critical p-values.

- Consider a case where three variables are measured.
- What is the critical value of χ^2 within which 19 out of 20 observations should be found?
- The value is 7.815 (or a Mahalanobis distance of 2.80) as can be verified in Table I.
- If we record the data on a sample size of 200, we expect 10 to have a χ^2 greater than 7.815.
- Some people call this a p-value of 0.05. A p-value is $1-$the cumulative probability.
- It can also be calculated in Excel by CHISQ.INV.RT(0.05,3), which gives the same answer.

Compare the numbers for 1 degree of freedom to those for the normal distribution (Table I of *The normal distribution* [3]).

- A value of 2.706 corresponds to a cumulative probability of 0.9 or a p-value of 0.1, that is, 90% of observations should lie within this value, and 10% above it.
- The square root is 1.645. This means that we anticipate 90% will be between -1.645 and $+1.645$ standard deviations of the mean for the normal distribution.
- Interpolating between 1.64 and 1.65 in the normal distribution table, we obtain a value of around 0.95.
- However, the CDF is a cumulative function, and this implies that 5% is more than 1.645 standard deviations above the mean, so an additional 5% is less than 1.645 below the mean.

Table I. Critical points of the Chi squared distribution

	0.99	0.975	0.95	0.9	0.1	0.05	0.025	0.01
1	0.00016	0.00098	0.00393	0.01579	2.706	3.841	5.024	6.635
2	0.0201	0.0506	0.1026	0.2107	4.605	5.991	7.378	9.21
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.61	9.236	11.07	12.832	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.69	2.167	2.833	12.017	14.067	16.013	18.475
8	1.647	2.18	2.733	3.49	13.362	15.507	17.535	20.09
9	2.088	2.7	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.94	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.92	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.041	19.812	22.362	24.736	27.688
14	4.66	5.629	6.571	7.79	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
20	8.26	9.591	10.851	12.443	28.412	31.41	34.17	37.566
25	11.524	13.12	14.611	16.473	34.382	37.652	40.646	44.314
30	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892
40	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691
50	29.707	32.357	34.764	37.689	63.167	67.505	71.42	76.154
100	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807

The rows represent the degrees of freedom and the columns the probability levels, the cells of the values of chi squared or the squared Mahalanobis distance. The probabilities are conventionally represented by *p*-values.

- Hence, 10% of observations for a normally distributed population should have a *z*-value between ±1.645 or a χ^2 greater than this.

There are numerous resources that describe further mathematical properties of the χ^2 distribution, two online ones which are referenced here [4,5].

Richard G. Brereton

School of Chemistry, University of Bristol, Cantock's Close, Bristol, BS8 1TS, UK

REFERENCES

1. Helmert FR. Die Genauigkeit der Formel von Peters zur Berechnung des wahrscheinliehen Beobachtungsfehlers direckter Beobachtungen gleicher Genauigket. *Astr. Nachr.* 1876; **88**: 113–132.
2. Pearson K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can reasonably be supposed to have arisen from random sampling. *Philos. Mag.* 1900; **50**: 157–175.
3. Brereton RG. The normal distribution. *J. Chemom.* 2014. DOI: 10.1002/cem.2655
4. NIST/SEMATECH e-handbook of statistical methods, April 2012, 1.3.6.6.6. Chi-Square Distribution http://www.itl.nist.gov/div898/handbook/eda/section3/eda3666.htm
5. Lane DM. Chi square distribution *Online Statistics Education: An Interactive Multimedia Course of Study* http://onlinestatbook.com/2/chi_square/distribution.html