Journal of CHEMOMETRICS

Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/cem.2655

The normal distribution

by Richard G. Brereton

Journal of Chemometrics Columnist

Many chemometrics methods assume that the underlying population of experimental measurements is normally distributed. Whether using significance tests in soft independent modeling by class analogy, or confidence limits in regression, or interpreting the significance of factors using ANOVA, the assumption of normality is often buried deep inside. Often, decisions about, for example, whether a process is under control in manufacturing, or what the origins are of a food as measured by near infrared, or what the confidence limits are in the measurement of an additive in a fuel often depend on one or more underlying populations falling into a normal distribution.

CARL GAUSS

The normal (or Gaussian) distribution was first described by Carl Friedrich Gauss in 1809 [1] in the context of measurement errors in astronomy. During the 19th century, this distribution was applied extensively in the developing area of applied probability and statistics.

> *"Often decisions in chemometrics depend on one or more underlying populations falling into a normal distribution."*

MEN'S HEIGHTS

An example involves the measurement of the heights of several men. The measurements can be characterised by

- the mean or average height and
- the standard deviation representing the spread of heights.

In the UK, the mean male height is 1.778 m, and the standard deviation 0.076 m. For a homogenous population, the distribution of heights should fall approximately into a normal distribution. In fact, underlying distributions are often not in themselves normal, but by the central limit theorem, the sum of symmetric independent distributions often approximates the normal distribution. In later articles, we will explore the idea of independence, but it should suffice in this article to consider such populations first. It is primarily because of the central limit theorem that the concept of a normal distribution gained great popularity in early statistics [2].

REPRESENTATION AND CHARACTERISATION

We usually represent the normal distribution with the horizontal axis representing a measurement, such as men's heights, and the vertical axis representing a probability or frequency as in Figure 1 (a). This type of representation is often called a probability density function (often abbreviated pdf). Sometimes, it can also be represented as a cumulative distribution function (cdf): this is a graph of the proportion of samples below a certain value against the value itself as in Figure 1(b). For a pdf, the maximum should be at the mean (in our case 1.778 m), whereas for a cdf, the mean represents the halfway point.

A Gaussian (or normal) distribution is characterised by three parameters:

- the mean, which is the centre of the distribution;
- the area: for a pdf, this is 1 but otherwise may depend on the measurement scale (intensity) and number of samples (or sampling interval);
- the width: often measured at half height but sometimes expressed as a standard deviation, which is linearly related to the half height.

This is illustrated in Figure 2.

Measurements can consist of several different Gaussian distributions. In Figure 3, two partially overlapping normal distributions are illustrated. They may represent two groups of samples, for example, the length of adult mice from two subspecies. It is important to remember that the shape of both distributions illustrated is the same. The left-hand distribution is wider and corresponds on the whole to lower measurement values. But there is a region of overlap, so just by measuring the length, for example, of an adult mouse, we cannot perfectly distinguish between the two populations. In later articles, we will discuss how adding additional measurements may allow us to better distinguish samples from two groups even if there is overlap in the distributions for single measurements.

STANDARD NORMAL DISTRIBUTION

Usually, normal distributions are compared by putting them on the same scale to obtain the standard normal distribution. To do this,

- subtract the experimental mean, so that the new mean is shifted to 0 and
- divide the shifted data by the experimental standard deviation so that the standard deviation of the new data is now 1.

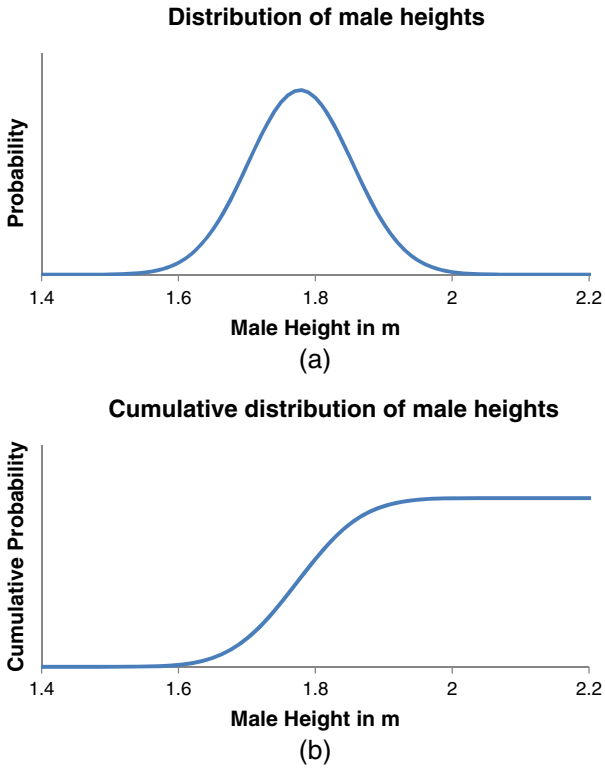


Figure 1. Normal distribution represented as (a) a probability density function (pdf) and (b) a cumulative distribution.

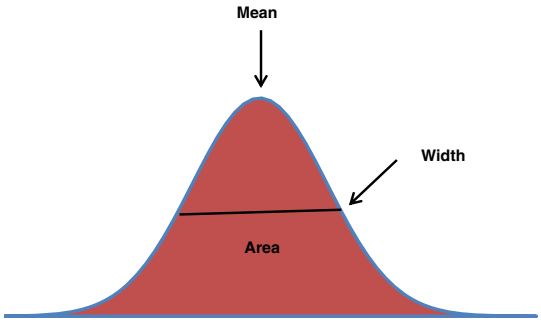


Figure 2. Parameters characterising a Gaussian or normal distribution.

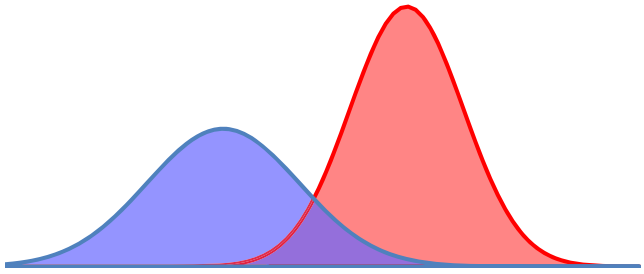


Figure 3. Two partially overlapping normal distributions.

Hence, a standard normal distribution has a mean of 0 and standard deviation of 1. Often, there is a further step of scaling the area under the distribution curve so it totals 1, so that the curve becomes a pdf rather than a frequency distribution.

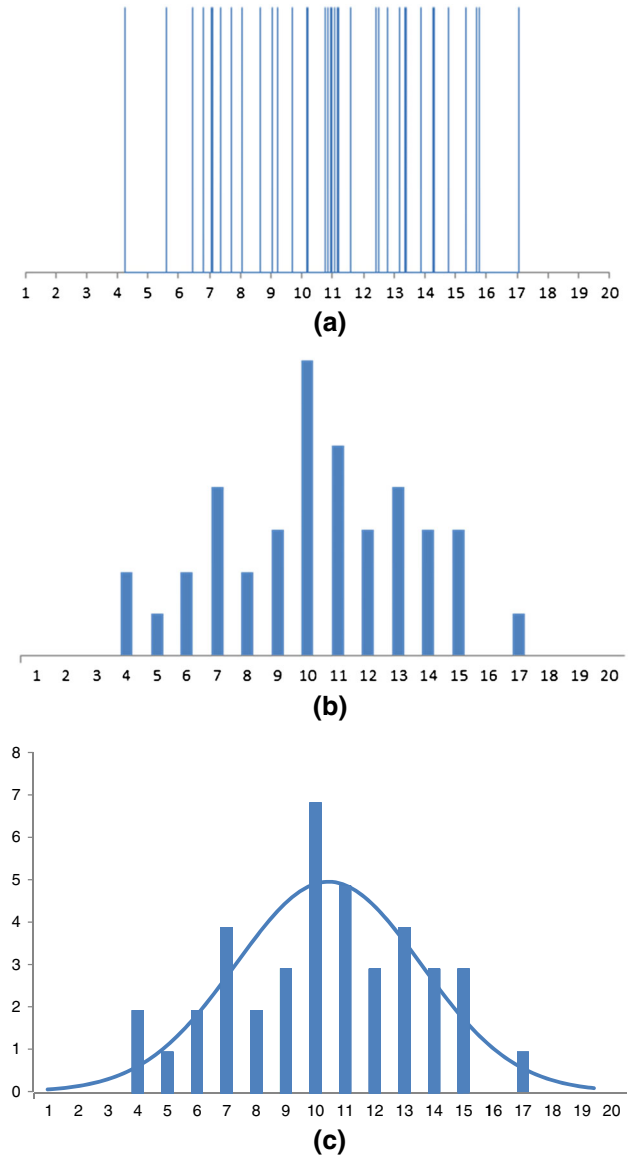


Figure 4. A set of 40 observations (a) as raw measurements, (b) as number of observations over intervals of 1 and (c) with underlying normal curve superimposed.

VISUALISING EXPERIMENTAL DATA

If you are analysing experimental data, to visualise approximately if they fall into a normal distribution, you can do as follows.

- Plot the values of the experimental observations as in Figure 4(a).
- Use an interval (or window) and count the number of observations within each interval. In the example of Figure 4(b), we chose an interval of 1 unit. Obviously, a sensible interval must be chosen. Too small and there may not be sufficient observations per interval, and if too large, there will not be sufficient detail.
- Work out the standard deviation and mean of the distribution. Visualise a Gaussian with the same shape, given by $\exp(-(x-m)^2/(2s^2))$, where m is the mean and s the standard deviation of the data; if you want to be very rigorous, multiply this by a factor $N/s\sqrt{(2\pi)}$, where N is the total number of samples: the vertical scale (or area) should now be the same as that of the original data (the

denominator is used to make the area under a normal curve equal to 1 in the absence of any numerator, so it becomes a pdf). Compare this curve to the original data, as in Figure 4(c).

We will look more at tests of normality in other articles, but it is often constructive to visualise information in a simple graphical manner before moving forward.

PREDICTION

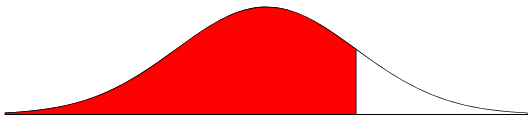
If we believe that a distribution is normal, it is then possible using the normal distribution to predict how many or what portion of samples we anticipate to correspond to a given range of measurements. In the example of Figure 4, we calculate the mean of the data to be 10.76 and the standard deviation 3.22 (there is a small difference between the population and sample standard deviation, which will be discussed in subsequent articles. We will use the latter, but the difference is very small for 40 samples and is unlikely to have any measurable experimental significance).

- Consider asking how many samples correspond to a measurement between 6 and 9.
- To do this, calculate the difference between the cdf at a value of 9, and at a value of 6.
- Using Excel, the value for 9 in our example is given by NORMDIST(9,10.76,3.22,TRUE) or 0.292; this represents the expected portion of readings less than 9. The 'TRUE' means that we are interested in the cumulative distribution. Subtracting this from the corresponding value for 6 (0.069), we have 0.223.
- This implies that 0.223 or 22.3% of the readings are expected to fall between 6 and 9, or there is a 22.3% probability of finding a sample within this range.
- As there are 40 samples in the original data, we would expect around 8.9 (or to round it up 9) samples are in this range.
- Looking at the data, two readings are between 6 and 7, four between 7 and 8, and two between 8 and 9 (Figure 4(c)) or a total of 8, which agrees very well. Because the sample size is limited, our experimental data will not in itself perfectly fit a normal distribution, so some differences are anticipated.

We can ask different types of questions; for example, if we measure the heights of 1000 men in the UK, how many do we expect will be more than 2 m high using the example of men's heights? The answer is 1.74; that is, typically, we would just encounter one or two people over this height (around 6 ft 7 in.).

We can say that the underlying population is likely to be normally distributed, but the observed data will not fit this exactly. Unless all possible samples are recorded, we can only measure a

Table I. Cumulative normal distribution



	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.5	0.504	0.508	0.512	0.516	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.591	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.648	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.67	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.695	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.719	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.758	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.791	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.834	0.8365	0.8389
1	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.877	0.879	0.881	0.883
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.898	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.937	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.975	0.9756	0.9761	0.9767
	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
2	0.9772	0.9821	0.9861	0.9893	0.9918	0.9938	0.9953	0.9965	0.9974	0.9981
3	0.9987	0.999	0.9993	0.9995	0.9997	0.9998	0.9998	0.9999	0.9999	1

fraction (usually a tiny one) of the underlying population; we try to use these to determine information about the population as a whole, but this depends on our sample being representative.

Most modern scientists have access to tools such as Excel or MATLAB for answering the preceding types of questions. However, most traditional statistical texts were written many decades ago, and texts evolve very slowly, so many people still encounter normal distribution tables. Table I is of the cumulative standard normal distribution. The rows and columns represent the number of standard deviations above the mean. It can be used to answer questions such as the ones described earlier.

- For the example of Figure 4, consider asking how many samples we expect to have readings between 6 and 9.
- 6 is (6–10.76) or 4.76 units below the mean of 10.76.
- This corresponds to 4.76/3.22 or 1.47 standard deviations.
- Similarly, 9 is 0.55 standard deviations below the mean.
- Looking at Table I, we find that we expect 0.7088 of the samples to exceed a value of 0.55 standard deviations above the mean; similarly, 0.9292 exceeds 1.47 standard deviations.
- Subtracting these two, we expect that 0.2204 samples (or 22.0%) will be between these limits (remember the normal distribution curve is symmetrical).
- Hence, we anticipate that around 0.22×40 or 8.8 samples out of 40 will be within these limits.
- Note that this is slightly different to the answer obtained by Excel: this is because normal distribution tables are only presented to a certain degree of precision, for example, 9 is actually 0.5466 rather than 0.55 standard deviations below the mean, but most tables only present the number of standard deviations to two decimal point accuracy.
- The small difference between the result from Excel and using the normal distribution table is most likely to be unmeasurable, and indeed normal distributions only approximate experimental measurements. Unlike, for example, physical or quantum chemistry where predictions from theory can often be obtained to a high degree of precision (primarily because sample sizes are in practice huge), in many applications of chemometrics, sample sizes are usually too small to worry about small discrepancies in the accuracies of predictions.

Hence, it is possible to answer numerical questions about samples by assuming that measurements are normally distributed. In the example of men's heights, we can ask what proportion of men are likely to exceed 1.8 m in height? Or in a population of 30 million people, how many do we anticipate are between 1.6 and 1.7 m in height? Or what is the chance that a man is less than 1.4 m (4 ft 7 in.) in height? By assuming normality, this can be done just by determining the mean and standard deviation of a small but representative sample.

The assumptions that a homogeneous population of samples from a single group or origin is normally distributed is often buried deeply within many chemometrics tests, and it is important to understand this as a fundamental basis of much statistical inference, as well as when the assumptions are not valid.

There are many more detailed discussions of the normal distribution. The NIST Engineering and Statistics Handbook [3] is particularly recommended as an online source for further reading.

REFERENCES

1. Gauss CF. *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium*. (available ABC Books, Lowfield Heath: Crawley, United Kingdom), 1809.
2. Stewart I. *17 Equations that Changed the World*, Chapter 7, Profile Books: London, 2013.
3. NIST/SEMATECH e-Handbook of Statistical Methods. April 2012, 1.3.6.6.1. Normal Distribution, http://www.itl.nist.gov/div898/handbook/eda/section3/eda3661.htm

