Journal of CHEMOMETRICS

Published online in Wiley Online Library: 5 March 2015

(wileyonlinelibrary.com) DOI: 10.1002/cem.2695

Populations and samples

by Richard G. Brereton

Journal of Chemometrics Columnist

Let us imagine we are interested in the sugar content of potatoes grown in the English county of Kent and harvested in June 2012. We might be compiling a report about sugar in potatoes in specific geographical areas, or looking at seasonal trends, perhaps related to temperature and rainfall and soil conditions. There will be many millions of potatoes harvested during this period. The underlying distribution of sugar content in all these potatoes cannot be measured with complete certainty unless all potatoes harvested are analysed, which is clearly impracticable.

It is essential before any significant study is undertaken to ask relevant scientific questions, which then are used to define the population of individuals or measurements and so to determine how to obtain an adequate sample.

SAMPLES

So what we would normally do is select a few potatoes, maybe 100 harvested in Kent in June 2012 and analyse these instead. From these, we try then to estimate the underlying distribution of sugar content in the entire county that month. The 100 or so potatoes we take from farms and analyse are called a sample. Note the statistical and chemical definition varies. A chemist would define every bit of each potato as a sample in its own right, but statistical terminology would normally define all 100 potatoes as a single sample. A rival study might come and pick a separate set of 100 potatoes; hence, two samples (in chemical terms, two sets of samples) are analysed. We will use the statistical term for sample unless otherwise stated, consisting of n individual observations (our 100 potatoes). Our sample is a relatively small proportion of the total potato harvest, and of course, there is no guarantee that the distribution of sugar content from this sample will match that of the overall population: we will come back to this important issue in later articles but will assume that the sampling was adequately representative (which would involve obtaining different varieties and sizes from different farms and is by no means easy).

HOW MANY POTATOES?

- The entire potato crop harvested in July 2012 in Kent is said to be a population.
- There are estimated to be 3×10^{14} g of potatoes harvested a year worldwide; at an average size of 150 g, this implies around 2×10^{12} potatoes or an average of 285 per person.
- The population of Kent is 1.5 million and so represents around 0.02% of the world population.
- If potatoes were harvested in proportion to population this would represent a harvest of nearly 500 million potatoes per year, but because some countries are more productive, the UK being big potato eaters, we might estimate that 1 billion potatoes are harvested a year in Kent.
- If the potato harvest is over 5 months, then approximately 200 million potatoes would be harvested in a given month.
- Obviously, we cannot analyse all these potatoes, but these 200 million potatoes harvested in Kent in June 2012 are our population.
- Our sample of 100 potatoes represents around 0.00001% of the harvest.
- Note that there can be populations within populations: for example, the potatoes grown in Kent are only a small fraction of the total number of potatoes harvested in England in 2012 (which in turn is only a fraction of all potatoes harvested worldwide that year and so on). We can define any subgroup as a population; however, once defined, it is important to be able to obtain representative samples from this group. This is illustrated in Figure 1.

POPULATIONS WITHIN POPULATIONS

Whether we reduce the number of observations in a population to create another population or to create a sample depends how we do it. For example, take a class of 40 students. This could be our original population – we may for example be interested in their sporting habits. If there are 20 girls and 20 boys, just selecting all 20 boys creates a new population – all boys in the class. Selecting an approximately equal number of 10 boys and 10 girls for our study allows us to sample the class, as no group is sampled completely. Note that 15 boys and 5 girls is still a sample, even though it is a biased one. Of course, in most situations, the size of a sample is very much smaller than its parent population. However, there can be interesting dilemmas as to what we define as a population. For example, we may be studying the characteristics of all the rats in a given laboratory. Is this our population? It may be if the rats have been treated in a specific and

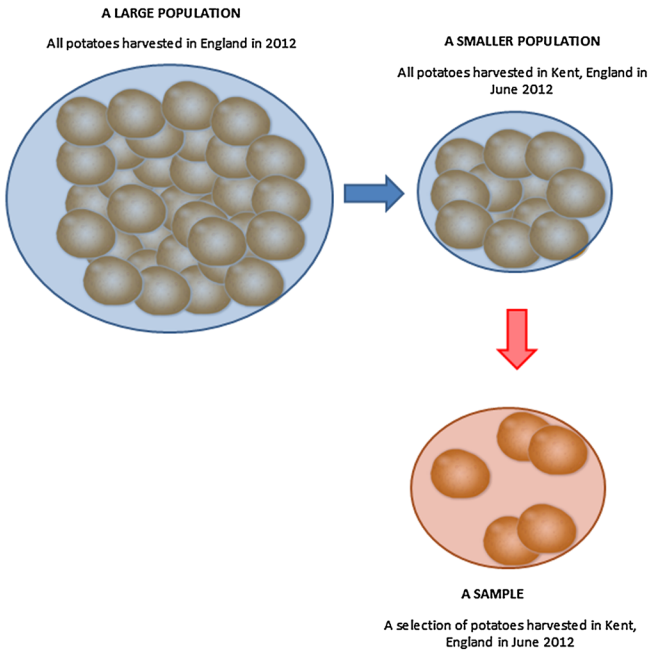

Figure 1. Populations and samples.

unique way, or is it considered a sample of all the rats in the world? This can be sorted out by posing the scientific question before our study. If we are only interested in reporting a specific treatment on a very controlled population of rats, then our entire laboratory is the population, but if we want to ask a more generalised question, it is just a sample and probably a very biased one.

DESCRIPTIVE AND INFERENTIAL STATISTICS

- From our sample, we calculate certain numbers, for example, the mean sugar content of potatoes and the standard deviation of the sugar content.
- These are called descriptive statistics and give us a numerical description of the sample.
- The population as a whole also has underlying properties that could be represented numerically, such as the mean and standard deviation of the sugar content of all potatoes harvested in Kent in June 2012.
- These underlying numerical properties are called population parameters and can be rarely measured directly unless the population is small and very well defined (e.g. all pupils in a given class).
- We try to estimate the population parameters values from our small, representative, sample via inferential statistics.
- Whether the population parameters that are inferred from a sample are reliable estimates depends on how representative and how large our sample is.

PARAMETERS AND STATISTICS

The simplest values of interest are the mean, standard deviation and variance, which we will later discuss. In other articles, we will discuss other properties of samples and populations. Notation and definitions can be rather confusing and inconsistent and is often misunderstood. Books on statistics, physics and engineering use slightly different terminologies, and chemometrics sits in the middle of all this. There are fundamentally three types of numbers of interest.

- Descriptive statistics obtained from the sample.
- Population parameters, which can normally not be measured directly.
- Inferential statistics, using observations from the sample to estimate population parameters. In most studies, these are the most interesting.

David Lane provides a good summary about estimation [1].

MEAN, STANDARD DEVIATION AND VARIANCE

There are lots of common ways of denoting the mean, standard deviation and variance as illustrated in Table I together with their computational formulae. Whereas, at first, the situation may seem simple, it is actually a little more complicated, and it is necessary to keep a cool head.

- We denote the sample size as n (e.g. the number of potatoes analysed – 100 in our case); then, the mean, variance and standard deviation can be defined very simply. The summation for the variance is divided by n as this is a simple description of the set of measurements.
- It is only if we use the sample statistics to estimate population parameters that it becomes more complicated. Under such circumstances, the variance used to describe the sample is usually an underestimate of the population variance and is sometimes called a biased estimator.
- To remove this bias, we divide the summation by $n-1$, to give what is sometimes called an unbiased estimator of the variance. By an unbiased estimator, we imply that if we took many independent samples, after a time, the average of all the variances should approximate to the population variances.
- The term $\sum_{i=1}^{n}(x_i-\bar{x})^2/(n-1)$ is often also called the sample variance. Note a bit of confusion; for descriptive statistics, it is usual to divide by the sample size n although this term is often called the population variance. There are two aims when taking a sample, the first being to describe it and the second to estimate the population parameters.
- The square root of the sample variance is called the sample standard deviation, given by $\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2/(n-1)}$. However, this is actually strictly speaking not the best estimate of the population standard deviation and is sometimes referred to as the 'slightly biased' estimate of the population standard deviation and is on average likely to be less than the estimate of the population standard deviation. Of course, we do not know the population standard deviation; sometimes it is lower and sometimes higher, but it is more likely to be lower. This discrepancy is the key to understanding why the t-distribution arises as discussed in the next article.
- The population standard deviation can never be known unless the entire population is sampled. There are theoretical relationships between the mean sample standard deviation and the population standard deviation, but these depend on the underlying population distribution and are rarely used. Some formulae are given in Table I but rarely used and should only be regarded as an approximation.

Table I. Ways of denoting the mean, standard deviation and variance, together with common formulae

	Descriptive sample statistics	Inferential estimations of population parameters	Population parameters
Mean	$\bar{x}$ m m_0 $\sum_{i=1}^{n} x_i/n$	$\hat{\mu}$ k_1 $\hat{\mu}_0$ (unbiased estimate of the mean) $\sum_{i=1}^{n} x_i/n$	μ μ_0 $\sum_{i=1}^{N} x_i/N$
Standard deviation	S s_N $\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2/n}$	$\hat{\sigma}$ (unbiased estimate of the standard deviation) $\gamma\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2/(n-1)}, \gamma>1$ $\approx\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2/(n-1.5)}$ s s_{N-1} (slightly biased estimate of the standard deviation, also called sample standard deviation) $\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2/(n-1)}$	σ $\sqrt{\sum_{i=1}^{N}(x_i-\mu)^2/N}$
Variance	S^2 m_2 s_N^2 $\sum_{i=1}^{n}(x_i-\bar{x})^2/n$	s^2 $\hat{\mu}_2$ s_{N-1}^2 k_2 $\hat{\sigma^2}$ (unbiased estimate of the variance) $\sum_{i=1}^{n}(x_i-\bar{x})^2/(n-1)$	σ^2 μ_2 $\sum_{i=1}^{N}(x_i-\mu)^2/N$

Notes
1. It is assumed that n observations define a sample and there are N in a population.
2. The variable γ depends on the underlying distribution and the sample size; an approximate estimator is given assuming a normal distribution.
3. The sample standard deviation s or $s_{N\text{-}1}$ is commonly calculated but is not strictly an estimate of the population standard deviation.

- So remarkably, the best estimate of the population standard deviation is not the square root of the best estimate of the sample variance. This relates to the fact that the square-root of the mean sum of squares of a set of numbers is not the same as their mean.
- Note that the sample mean, in contrast, is divided by n and is an unbiased estimator of the population mean.

For a good summary of the unbiased estimation of the standard deviation please refer to a Wikipedia article [2]. Many books on process control often also discuss this in detail one of which is [3].

TERMINOLOGY

The following definitions follow from Table I although there is also quite a lot of different terminology in the literature. Whereas statisticians have their own terminology, sometimes chemometricians come across related concepts from engineering and physics.

- The jth moment around the mean is defined as m_j. This is the sum of the jth-order deviations from the mean divided by n. If $j=0$, this is defined as the 0th moment or the mean, if $j=1$ m_1 equals 0 (the deviations about the mean in any sample always sums to zero, as negative deviations cancel positive deviations), and for $j=2$, the m_2 is the variance.
- The k statistic is defined as the estimated value of the mean, variance, and so forth. Note that k_1 is the estimated value of the mean, whereas m_1 equals 0.
- In descriptive statistics, for the standard deviation and variance, we recommend using (upper case) S and for the sample standard deviation (lower case) s. This distinction is by no means universal. Also, the sample standard deviation involving the squared deviations from the mean divided by $n-1$ is also often denoted by s.
- It is, alternatively, possible to use the subscript N or $N-1$ to distinguish the method for calculation of the variance and standard deviation although this is by no means universally accepted.

Some of the statistical niceties distinguishing between the various methods for calculation of variances and standard deviations cannot in practice be told apart. However, for small sample sizes, there can be small differences in estimates using different formulae, but when sample sizes are small, it is also important to ensure that samples are representative. With the

increasing application of chemometrics to metabolomics, many papers report inadequate sample sizes, and only a small bias in sampling strategy can result in far greater errors in estimation of population parameters than could be distinguished by use of different computational formulae. In the univariate case where one measurement is made per object, once sample sizes are much larger than 20, there is no practical difference between the computational formulae.

SCIENTIFIC QUESTIONS

It is however fundamentally important to understand the difference between a sample and a population, and there are many misunderstandings in the literature that are a result of confusion on this point. A perfectly adequate result on a study of limited scope does not necessarily transfer to a larger and more varied parent population, and it is essential before any significant study is undertaken to ask relevant scientific questions, which then are used to define the population and so determine how to obtain an adequate sample. For example, if we are only interested in the potatoes grown in Kent, our sampling strategy would be quite different compared with a study of potatoes in general worldwide. In addition, if our scientific question were even more detailed, for example, to look at a particular supplier or variety of potato grown in Kent, we would need quite a different parent population and sampling strategy also.

A good summary about populations and samples is presented in the BMJ [4].

Richard G. Brereton

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

REFERENCES

1. Lane D.M. On-line Statistics Textbook, Characteristics of Estimators, http://onlinestatbook.com/2/estimation/characteristics.html.
2. Unbiassed Estimation of Standard Deviations, Wikipedia, http://en.wikipedia.org/wiki/Unbiased_estimation_of_standard_deviation.
3. Duncan A. J. *Quality Control and Industrial Statistics* (4th Ed.). Irwin, Homewood, 1974.
4. Swinscow T. D. V. Statistics at Square One, 3. Populations and samples. *BMJ*, http://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/3-populations-and-samples.