Estimating Population Variance: theoretical approach and using Monte Carlo simulation

1. The mean, standard deviation and variance of a sample of readings

(See Chapter 5, section 5.1 in 'Introduction to Uncertainty in Measurement', by Les Kirkup and Bob Frenkel)

In statistics the term 'population' is used in a more general sense than in ordinary English, where it refers to a large number of living creatures, often humans. In the more general statistical sense, a population is often a very large number, or an infinite number, of possible readings or measurements. For example, a factory may have produced ten thousand steel ball bearings of a particular size. Because of unavoidable variability in the manufacturing process, the precise sizes of individual bearings will vary slightly, even though they are all intended to be 'nominally' the same size. We therefore have a population of ten thousand sizes. As another example, a high-quality digital multimeter (DMM) may be measuring the voltage of a battery and displaying it to six or seven decimal places. Because the voltage is not perfectly stable, and because there may be electrical 'pick-up' or interference from surrounding equipment or TV and radio transmissions, the displayed voltage fluctuates and may show a drift. In this second example, where we can in principle continue endlessly taking measurements, the population of voltages is evidently infinite.

How do we describe or characterise a population? The two obvious descriptions that immediately come to mind are: the average value and the range of values. The average value is more commonly given the technical term *mean* value and is often denoted by the Greek symbol μ ('mu'). The range of values is the difference between the maximum and minimum values, but the practically more useful and more common measure of the 'spread' of results is the *standard deviation* of the population, and this is often given the Greek symbol σ ('sigma'). The standard deviation is not the same as the range of values; in fact the standard deviation is less than the range by a factor that is generally between (roughly) 3 and 4.

If the population, of size N, contains N readings x_1, x_2,...x_N, the mean reading μ is defined as

$$\mu = \frac{x_1 + x_2 + ... x_N}{N} = \frac{\sum_{i=1}^{N} x_i}{N}. \tag{1}$$

The symbol Σ denotes summation and is a very commonly used shorthand expression in mathematics.

The standard deviation σ of the N readings is defined as

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}. \tag{2}$$

For example, consider the (absurdly small!) population of size $N = 4$ and comprising the readings $x_1 = 1.0$, $x_2 = 1.1$, $x_3 = 0.9$ and $x_4 = 1.2$. Then we have $\mu = 1.05$ and $\sigma = 0.112$ for this population. The range of values is $1.2 - 0.9 = 0.3$.

Consider another population, also with $N = 4$, but with the values $x_1 = 0.7$, $x_2 = 1.3$, $x_3 = 1.6$, $x_4 = 0.6$. This population also has mean $\mu = 1.05$, but its standard deviation σ is $\sigma = 0.415$. The standard deviation is larger than for the first population, and this is evidently how it should be, since although the second population has the same mean, its range of values is $1.6 - 0.6 = 1.0$, more than three times as large as for the first population.

With a large or infinite population, we evidently cannot afford the time nor the resources to measure every single member of the population. We therefore have to make do with a relatively much smaller *sample* from the population. We denote by n the size of the sample, with $n << N$. An immediate and rather obvious question arises. Unless we are fortunate in our choice of sample, the mean $\bar{x}$ of our sample will not be exactly equal to μ (although we expect them to be fairly close to each other). So if we take a large number of samples, will the average of the resulting large number of sample means $\bar{x}$ *tend* towards the 'true' population mean μ, or will this average be 'biased' too high or too low relative to μ, no matter how many samples we take? If the average of the large number of sample means does actually tend towards μ, then we say that the mean $\bar{x}$ of *a single sample* is an *unbiased* estimate of μ. We obviously prefer unbiased to biased estimates of population quantities. A similar question arises regarding the standard deviation of our sample of size n; is this, or is this not, an unbiased estimate of the population standard deviation?

An alternative expression of unbiasedness uses the term 'expectation'. The expectation of a quantity is the mean value of that quantity over an entire population. Then the mean of a sample is an unbiased estimate of the population mean if the expectation of a sample mean equals the population mean. It is shown in section 5.1.2 of the book that the expectation of the sample mean is in fact equal to the population mean. So $\bar{x}$ is an unbiased estimate of μ. We examine below in some detail the corresponding question regarding the standard deviation. It is convenient to work with the *square* of the standard deviation, known as the *variance*. The variance σ^2 of the population is, then,

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}. \tag{3}$$

2. The unbiased estimate s^2 of the variance σ^2 of a population

Let a sample consist of n independent readings x_1, x_2,...x_n, drawn from a population which is not necessarily Gaussian. We know that the mean $\bar{x}$ of our sample is given by

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{4}$$

and that $\bar{x}$ is an unbiased estimate of the population mean μ. We express this unbiasedness as:

$$E(\bar{x}) = \mu \tag{5}$$

where E denotes: 'expectation of'.

The expectation of the sum of quantities is the sum of the expectations of the quantities:

$$E(y_1 + y_2 + y_3 ...) = E(y_1) + E(y_2) + E(y_3) ... \tag{6}$$

A similar rule applies to the product of quantities, as long as they are mutually uncorrelated (this will be satisfied if they are independent of one another):

$$E(y_1 y_2 y_3 ...) = E(y_1)E(y_2)E(y_3) ... \tag{7}$$

Just as $\bar{x}$ is an unbiased estimate of μ, the following quantity s^2 is an unbiased estimate of the variance σ^2 of the population:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}. \tag{8}$$

The unbiasedness of s^2 is expressed, similarly to (5) above, as:

$$E(s^2) = \sigma^2. \tag{9}$$

As well as (5) and (9), we have the following results:

$$E(x_i) = \mu \tag{10}$$

and

$$E\left[(x_i - \mu)^2\right] = \sigma^2, \tag{11}$$

which can be used as alternative definitions of μ and σ^2.

We note that in (8), the sum is over all squared differences $(x_i - \bar{x})^2$ between the readings and the sample mean, but this sum is divided not by n but by $n - 1$. This can be understood intuitively as reasonable, because $\bar{x}$, being the mean of the x_i in the sample, tends to 'follow' the sample. If, for example, the sample that we pick happens to contain several fairly large values, then obviously their mean will also be rather large. The mean of the sample, in other words, is positively correlated with the sample values. Moreover, as might be expected, the smaller the sample size n, the larger will be the correlation. So the differences $(x_i - \bar{x})^2$ will not be precise measures of the variability of the x_i, but will be shrunken slightly. Dividing the sum of these squared differences by the smaller number $n - 1$, rather than by n, exactly compensates for this shrinking: dividing by a smaller number gives a bigger result. Naturally, if n is large, the shrinking may be negligible because of the smaller correlation, and $n - 1$ is then very close to n anyway.

To show that s in (8) satisfies $E(s^2) = \sigma^2$, we first establish the result:

$$E\left[(\bar{x} - \mu)^2\right] = \sigma^2/n \tag{12}$$

It is, incidentally, worth comparing (12) with (11). Equation (12) states that the variance of the mean of a sample is less by a factor of n than the variance of any reading in that sample, the latter being expressed by (11). This result, which applies only to uncorrelated readings, is well known as the theoretical underpinning of the notion that taking the average of several readings from a population generally gives a more reliable result than a single reading.

Expanding $(\bar{x}-\mu)^2$ in the left-hand side of (12) gives:

$$(\bar{x}-\mu)^2 = (\bar{x})^2 + \mu^2 - 2\bar{x}\mu \tag{13}$$

so

$$E\left[(\bar{x}-\mu)^2\right] = E\left[(\bar{x})^2\right] + \mu^2 - 2\mu E(\bar{x}), \tag{14}$$

since $E(\mu) = \mu$ (μ being the constant population mean) and $E(\bar{x}\mu) = \mu E(\bar{x})$.

Substituting (5) into (14) gives:

$$E\left[(\bar{x}-\mu)^2\right] = E\left[(\bar{x})^2\right] - \mu^2. \tag{15}$$

Squaring (4),

$$(\bar{x})^2 = \frac{\sum_{i=1}^{n} x_i^2 + \sum_{i\neq j}^{n} x_i x_j}{n^2}. \tag{16}$$

From (10) and (11), we have

$$E\left[(x_i-\mu)^2\right] = \sigma^2 = E(x_i^2) + \mu^2 - 2\mu E(x_i) = E(x_i^2) - \mu^2, \tag{17}$$

so

$$E(x_i^2) = \sigma^2 + \mu^2. \tag{18}$$

Now taking expectations of (16), and using (18),

$$E\left[(\bar{x})^2\right] = \frac{(n)(\sigma^2+\mu^2)}{n^2} + \frac{\sum_{i\neq j}^{n} E(x_i x_j)}{n^2}. \tag{19}$$

If x_i, x_j are uncorrelated for all i, j, then

$$E(x_i x_j) = E(x_i)E(x_j) = \mu^2, \tag{20}$$

using (10).

The second term on the right-hand size of (19) has $n(n-1)$ terms (since $i \neq j$ and the range of each of i and j is 1, 2,...n). Therefore (19) becomes:

$$E\left[(\bar{x})^2\right] = \frac{(n)(\sigma^2+\mu^2)}{n^2} + \frac{n(n-1)\mu^2}{n^2} \tag{21}$$

$$= \frac{\sigma^2}{n} + \mu^2. \tag{22}$$

Substituting (21) into (14) now gives

$$E\left[(\bar{x}-\mu)^2\right] = \frac{\sigma^2}{n} + \mu^2 - \mu^2 = \frac{\sigma^2}{n}, \tag{23}$$

which verifies (12).

From (8), we have

$$s^2 = \frac{\sum_{i=1}^n x_i^2 + n(\bar{x})^2 - 2n(\bar{x})^2}{n-1} \tag{24}$$

$$= \frac{\sum_{i=1}^n x_i^2 - (\bar{x})^2}{n-1}. \tag{25}$$

Using (18) and (22),

$$E(s^2) = \frac{n(\sigma^2+\mu^2)}{n-1} - \frac{n}{n-1}\left(\frac{\sigma^2}{n} + \mu^2\right) \tag{26}$$

$$= \sigma^2\left(\frac{n}{n-1} - \frac{1}{n-1}\right) + \mu^2\left(\frac{n}{n-1} - \frac{n}{n-1}\right) \tag{27}$$

$$= \sigma^2, \tag{28}$$

implying that s^2 as defined in (8) is an unbiased estimate of σ^2. We note that no assumption has been made about the distribution of the population — whether it is Gaussian, uniform or some other. (These distributions are discussed in Chapter 8 of the book).

We should note that $E(s^2) = \sigma^2$ does *not* imply that $E(s) = \sigma$. In other words, the standard deviation s of the sample, defined (from equation (8)) as

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}, \tag{29}$$

is not an unbiased estimate of σ. However, if the sample size n is large, then $E(s) \sim \sigma$ to a good approximation. For a small sample size like $n = 4$, it can be shown that $E(s) \sim 0.921\sigma$ if the population has a Gaussian distribution of readings, which is often the case. So for $n = 4$ the bias is such that s will, on the average, underestimate σ by about 8%. To estimate σ unbiasedly for $n = 4$ and a Gaussian distribution, we should use not s but $1.086s$, since $E(1.086s) = 1.086E(s) = 1.086 \times 0.921\sigma \sim \sigma$. This is discussed in Chapter 9 of the book (see in particular section 9.3).

3. Demonstration of (8) and (9) using Monte Carlo simulation

Equations (8) and (9) can be demonstrated using Monte Carlo simulation — a kind of 'experimental statistics'. To do so, we generate many, say 100 000, numbers distributed as a Gaussian distribution with mean 0 and standard deviation 1. (A very similar demonstration could use a different mean and standard deviation. Moreover, as will also be demonstrated, the distribution need not be Gaussian). We imagine a sample size of 4 ($n = 4$) and, accordingly, divide up these numbers into 25 000 samples each containing 4 numbers. For each sample, we calculate the variance using (8), and for comparison the variance using the divisor n instead of $n - 1$ in (8). We take the average of all 25 000 variances for the two cases (the correct unbiased case $n - 1 = 4 - 1 = 3$ and the biased case $n = 4$).

The table shows, for illustration, one hundred values from the Gaussian population of size 100 000 and mean 0 and standard deviation 1. For comparison, one hundred values are also shown from a uniform distribution extending from 0 to 1 (with mean therefore $\frac{1}{2}$). The results of the Monte Carlo simulation are illustrated next, after an introductory block diagram.

Gaussian population of 100 000
Mean 0, standard deviation 1
Samples of size 4: first 25 samples:

1	-2.585815	2	0.024374	3	-1.418745	4	0.674210
5	0.543145	6	-0.199996	7	1.583771	8	0.381309
9	-0.086750	10	0.148071	11	-1.118133	12	-0.886265
13	1.515258	14	-0.101430	15	1.741255	16	0.079619
17	-0.170757	18	-1.716129	19	-0.201561	20	-0.131813
21	-0.167282	22	-0.833529	23	0.853021	24	0.902268
25	-0.446727	26	-0.007423	27	-0.332461	28	0.037005
29	0.020367	30	-0.146662	31	1.375571	32	0.330749
33	0.980897	34	1.543196	35	1.788238	36	-0.080641
37	1.099143	38	-0.567788	39	-0.901259	40	0.587309
41	-0.100176	42	0.940348	43	0.591314	44	0.728907
45	0.894236	46	-0.237383	47	0.109399	48	-0.347329
49	1.048546	50	0.649478	51	-1.292128	52	-0.210678
53	-0.432180	54	0.407568	55	-0.638912	56	3.294370
57	0.294313	58	1.870753	59	-0.148265	60	1.447460
61	-0.566067	62	0.372851	63	-1.149085	64	0.169951
65	0.350986	66	-0.713021	67	-0.441229	68	0.380111
69	0.055844	70	-1.524628	71	0.007190	72	1.821286
73	1.117899	74	1.428137	75	-1.526690	76	-0.833840
77	0.834611	78	-0.111120	79	0.162576	80	0.266859
81	-1.245917	82	-1.682947	83	-0.376980	84	0.946597
85	-1.125063	86	0.000698	87	0.541020	88	0.210036
89	0.083045	90	1.083728	91	1.327011	92	1.312118
93	0.544686	94	-0.445188	95	1.929916	96	0.890427
97	1.217833	98	-0.932395	99	-0.126377	100	1.911818

Uniform population of 100 000
extending from 0 to 1
Samples of size 4: first 25 samples:

1	0.851861	2	0.234244	3	0.947130	4	0.419321
5	0.532818	6	0.078772	7	0.919987	8	0.235166
9	0.441439	10	0.276162	11	0.458671	12	0.887998
13	0.587625	14	0.217014	15	0.349702	16	0.451435
17	0.273201	18	0.691587	19	0.510050	20	0.415530
21	0.810426	22	0.829256	23	0.319412	24	0.357244
25	0.201223	26	0.953568	27	0.628917	28	0.216880
29	0.099747	30	0.441949	31	0.832755	32	0.117061
33	0.453570	34	0.152807	35	0.235978	36	0.084993
37	0.471476	38	0.104747	39	0.489978	40	0.057261
41	0.387356	42	0.288257	43	0.735057	44	0.111568
45	0.123223	46	0.011697	47	0.588628	48	0.078997
49	0.698097	50	0.918552	51	0.110281	52	0.499497
53	0.047627	54	0.464565	55	0.952701	56	0.049299
57	0.575952	58	0.034715	59	0.449590	60	0.270456
61	0.552198	62	0.790738	63	0.933588	64	0.830287
65	0.644230	66	0.587353	67	0.655884	68	0.442150
69	0.225713	70	0.562817	71	0.273296	72	0.285850
73	0.283646	74	0.242082	75	0.676899	76	0.646586
77	0.181504	78	0.533745	79	0.659507	80	0.348623
81	0.310592	82	0.124073	83	0.292402	84	0.406813
85	0.309958	86	0.465966	87	0.501753	88	0.966941
89	0.387302	90	0.386714	91	0.502628	92	0.667688
93	0.838576	94	0.955317	95	0.023078	96	0.868521
97	0.234767	98	0.727674	99	0.031519	100	0.744684