Understanding dof in statistics

Marc Emanuel

March 18, 2015

When estimating the population mean from a sample of iid's it is easy to show that the sample mean an unbiased estimate of the population mean is. It is what you intuitively also expect, since the population mean is nothing but the average of that sample that is the whole population. And each member has the same mean expectation value. Thus the sample mean is unbiased when defined as

$$\hat{\mu} := (\sum_s x_i)/n \tag{1}$$

That it is unbiased is another way for saying that the expectation value (average under the population distribution) is equal to the population mean and does not depend on the sample size:

$$\langle \hat{\mu} \rangle = \frac{1}{n} \sum_{i=1}^{n} \langle x_i \rangle$$

(The brackets denote expectation value, old habit)

The sample estimate of the population variance on the other hand is defined as the accumulated sample variance divided by $n-1$. As reason it is often stated that the degrees of freedom are one less than the sample size. That sounds pretty vague and not really correct, while it is easy in words to explain why this is needed and pretty straightforward to show precisely that this is the case:

So the claim is : the right "unbiased" definition for the Estimate of the population variance from the measurement of a sample of size n is

$$\hat{v} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \hat{\mu})^2 \tag{2}$$

And now for the reason.: it is the use of the estimate of the population mean that does it. The number of degrees of freedom are in fact n but the sample estimate of the mean, although unbiased, is not an independent quantity. It is defined as a linear combination of an estimate calculated in terms of the same set of data, namely the sample$\{x_i\}$. Now this set consists of n stochastic variables that totally independently fluctuate hence the n-degrees of freedom. Lets say we keep $x_2 - x_n$ fixed and we just sample x_1 over and over again slowly x_1 explores the real line (in case of a normal distribution), you could say in the X_1 direction, but while exploring its direction, the sample mean is 100% correlated with it. You could also say that in the proposed setting x_1 has no freedom anymore, but

is determined by the rest of the sample and the mean That is the deeper reason of the loss of one degree of freedom. To see this mathematically we will have to compare the sample estimate with the true variance of the population:

$$\langle v \rangle = \left\langle (x - \langle x \rangle)^2 \right\rangle = \left\langle x^2 \right\rangle - \langle x \rangle^2 \tag{3}$$

The last expression will be our target Now we write the sample estimate (2) out in the sample elements only using Eq, (1)

$$\langle \hat{v} \rangle = \frac{1}{n-1} \sum_{i=1}^{n} \left\langle (x_i - \frac{1}{n} \sum_{j=1}^{n} x_j)^2 \right\rangle$$

expand the square and use linearity of integration

$$= \frac{1}{n-1} \sum_{i=1}^{n} \left[\left\langle x_i^2 \right\rangle - 2\frac{1}{n} \sum_{j=1}^{n} \langle x_i x_j \rangle + \frac{1}{n^2} \sum_{j=1}^{n} \sum_{k=1}^{n} \langle x_j x_k \rangle \right]$$

Since the x_i are idd

$$\langle x_i x_j \rangle = \begin{cases} \langle x_i \rangle \langle x_j \rangle = \langle x \rangle^2 & \text{if } i \neq j \\ \left\langle x_i^2 \right\rangle = \left\langle x^2 \right\rangle & \text{if } i = j \end{cases}$$

and thus from Eq. (3)

$$\langle \hat{v} \rangle = \frac{n}{n-1} \left[\left\langle x^2 \right\rangle - 2\frac{1}{n}(\left\langle x^2 \right\rangle + (n-1) \langle x \rangle^2) + \frac{1}{n}(\left\langle x^2 \right\rangle + (n-1) \langle x \rangle^2) \right] = \langle v \rangle$$

q.e.d. Note that if we want to estimate another function of the stochastic variable x that depends on the population mean and variance (or standard deviation) the sample estimate loses two dof's. And if you are still not convinced ,think about how to estimate the variance from a sample of size one (it is undefined).