Discovering an Unbiased Normalization for Variation

1. The Problem

When estimating the "variation" in a set of N values by looking at only n of them, the following formulas are used for "variation":

(1) $$\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2 \quad \text{and}$$

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

where μ is the mean of the N objects and $\bar{x}$ bar is the mean of the n objects.

The two formulas are referred to as the "Population Variance" and the "Sample Variance". The two means are referred to as the Population mean and the Sample mean. The normalization by n-1 is counterintuitive since there are n summands in the Sample Variance formula. This normalization by n-1 is often justified with some statement that refers to "n-1 degrees of freedom". This justification is not often satisfying.

2. The Strategy

This note looks at the problem of choosing the appropriate normalization for variance without relying on any of the standard mumbo-jumbo of statistics. Specifically we seek measures of variance of the form:

(3) $$\frac{1}{f(N,N)}\sum_{i=1}^{N}(x_i-\mu)^2 \quad \text{and}$$

(4) $$\frac{1}{f(N,n)}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

where the function f is chosen in a meaningful way.

One strategy for "meaningfulness" is for the "average sample variance" (ASV) to be the same as the Population Variance. By this we mean to compute the sample variance over all possible samples of size n and average them. Specifically—

(5) $$\text{ASV} = \frac{1}{\binom{N}{n}}\sum_{\text{all samples}}\frac{1}{f(N,n)}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

where $\binom{N}{n}$ is the number of different subsets of size n that can be chosen from a Population of size N. Setting the average sample variance (5) equal to the population variance (3) and doing a little algebraic manipulation we find that the ratio of our normalizing factors should satisfy—

$$(6)\qquad \frac{f(N,n)}{f(N,N)} = \left(\frac{1}{\binom{N}{n}} \sum_{all\ samples} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right) \Big/ \left(\sum_{i=1}^{N} (x_i - \mu)^2 \right)$$

3. The Simplest Case

Consider the case with a Population of size three (N=3) and a sample size of two (n=2). Let the Population be designated by $\{x_1, x_2, x_3\}$. Since these variation formulas are independent under translations we can fix one of the x_i without losing generality. Let $x_3 = 0$. Now the set of all possible samples is

{
$\{x_1, x_2\}$,
$\{x_1, 0\}$,
$\{x_2, 0\}$
}

The Population mean is

$\mu = (x_1 + x_2 + 0)/3$

and the three sample means are

$\bar{x}_1 = (x_1 + x_2)/2$
$\bar{x}_2 = (x_1 + x_3)/2 = x_1/2$
$\bar{x}_3 = (x_2 + x_3)/2 = x_2/2$

and

$$\binom{N}{n} = \binom{3}{2} = 3$$

Plugging all this into (6) we find the ratio:

$$(7)\qquad \frac{f(3,2)}{f(3,3)} = \frac{\frac{1}{3}[(x_1 - \bar{x}_1)^2 + (x_2 - \bar{x}_1)^2 + (x_1 - \bar{x}_2)^2 + (x_3 - \bar{x}_2)^2 + (x_2 - \bar{x}_3)^2 + (x_3 - \bar{x}_3)^2]}{(x_1 - \mu)^2 + (x_2 - \mu)^2 + (x_3 - \mu)^2}$$

after some simplification we have

$$(8)\qquad \frac{f(3,2)}{f(3,3)} = \frac{\frac{1}{6}[(x_1 - x_2)^2 + x_1^2 + x_2^2]}{\frac{1}{9}[(2x_1 - x_2)^2 + (2x_2 - x_1)^2 + (x_1 + x_2)^2]}$$

which further simplifies to

$$(9)\qquad \frac{f(3,2)}{f(3,3)} = \frac{\frac{1}{3}[x_1^2 + x_1 x_2 + x_2^2]}{\frac{2}{3}[x_1^2 + x_1 x_2 + x_2^2]} = \frac{1}{2}$$

Thus for N=3 and n=2, to ensure an unbiased estimate of variation, our normalization terms should be in the ratio of one to two. Notice that the normalization terms from (1) and (2) actually have a ratio of one to three! This means that for N=3 and n=2 standard

formulas (1) and (2) will result in the sample variance consistently underestimating the population variance!

4. An Experimental Approach

For larger values of N the algebra gets complicated very rapidly. Notice that in the simplest case the ratio f(3,2)/f(3,3) did not depend on the actual x_i values. If that remains the case for other values of N and n then the ratio f(N,n)/f(N,N) can be found experimentally. An Excel worksheet should allow us to experimentally test the hypothesis that the ratio in (6) is independent of the data. If it is independent of the data, then our Excel worksheet will also give us the correct ratio.