MAC-CPTM Situations Project

Situation 67: Sample and Population Variance

Prepared at University of Georgia
Center for Proficiency in Teaching Mathematics
15 July 2008 – Ken Montgomery

Prompt

A student in a statistics class has observed that the definition of sample variance $\left(s^2\right)$ is the average of the squares of the deviations from the mean of the data set.

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

Yet, to compute the average of these squared deviations, they are summed and then divided by (n-1). "Why," asked the student, "do you not divide by n? Is this an actual mean or some sort of 'pseudo-mean'?"

Commentary

This prompt addresses one of the prerequisite topics already discussed in the class's study of measures of central tendency: the formal definition of mean ($\bar{x}$).

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

The new definition of variance may seem contradictory to the definition of mean, since the mean of the data ($\bar{x}$) and the variance itself both characterize the data set consisting of n values, as evidenced by the indices of the summation in both formulae. Thus, there are n squared deviations being summed, not (n-1).

Mathematical Foci

Mathematical Focus 1

Since the deviations themselves (not the squared deviations) sum to zero, there is one deviation that is determined by all of the others, leaving (n-1) squared deviations to uniquely contribute to the overall variance.

Suppose that you only had two data points in your sample: 3 and 5. Then the sample mean would be 4, since this is the average of 3 and 5. We refer to the difference between a given datum and its mean as its deviation, so the deviation for 3 is (3-4) = (-1). This is the amount by which the datum deviates from the predicted value, which is the mean ($\bar{x}$). The sum of the deviations therefore must equal zero since the mean is the average of all of the data values. Doesn't this require that the second deviation equal (+1)? Certainly, and if we check, (5-4)=1.

This is true in general, because the deviations sum to zero. The last deviation is equal to the additive inverse of the sum of all of the other deviations, of which there was only one in the given example. Suppose, however that you have *n* data values. Since the deviations sum to 0, suppose that you are given the first (n-1) of the deviations, the last deviation (the n*th* one) could be determined, just as in the example above. Recall that a deviation is the difference between a given data point and the sample mean, in general $(x_i - \bar{x})$. Let us say that for n data points, you know that if you sum all of the deviations except one (in other words (n-1) of them) and get a sum of -7, then the last one (the n*th* one) must have a value of +7, in order for the sum of all n deviations to be 0. So, the last deviation is determined by the first (n-1) deviations. In averaging the squared deviations, we divide by (n-1) instead of n, because there are (n-1) unique pieces of information concerning variability (Agresti & Franklin, 2007).

Mathematical Focus 2

One degree of freedom is lost when the sample is used to estimate the mean, and so the variance is the average of the squared deviations.

Starting with the sum of the deviations, equal to zero, we can express the n*th* deviation in terms of the other (n-1) values:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = (x_1 - \bar{x}) + (x_2 - \bar{x}) + \ldots + (x_{n-1} - \bar{x}) + (x_n - \bar{x}) = 0$$

Solving this for the nth variance, we have:

$$(x_n - \bar{x}) = 0 - [(x_1 - \bar{x}) + (x_2 - \bar{x}) + \ldots + (x_{n-1} - \bar{x})] = 0 - \sum_{i=1}^{n-1}(x_i - \bar{x})$$

Thus, the nth variance is expressible as the negative sum of the (n-1) prior variances:

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

Thus, we are losing one degree of freedom, when we use the sample to estimate the mean and so the variance is the average of the squared deviations, which is the sum of the squared deviations, divided by (n-1).

One could divide by n, if the estimate of the mean was completely independent. Suppose that we were given the true value of the mean (not the value calculated from our data), and used this value to calculate our deviations, then we would divide by n. Also, suppose that we were given an estimate of the mean from a sample that is completely independent of the sample in which we are working then division by n would also be appropriate.

Mathematical Focus 3

The sample variance is an unbiased estimator of the population variance.

Two important concepts in this construction are the use of the expected value function and the notions of biased/unbiased estimators.

Freund (1992) pointed out that although statistics do not perfectly predict parameters, a statistic should at least do so on average. The expected value function is used to determine the average prediction of an estimator. Suppose X is a discrete random variable and the value of its probability distribution at x is given by *p(x)*, then the expected value of X is:

$$E(X) = \sum_x x \cdot p(x)$$

If an estimator has an expected value that is equal to the parameter that it intends to estimate then it is an unbiased estimator, otherwise it is biased.

Just as the sample mean is an estimate of the population mean, the sample variance is an estimate of the population variance. However, division by (n-1) provides a better estimate of the population variance than division by n (Mendenhall, Beaver & Beaver, 2005). The bias for this estimate is equal to $\left(\frac{n-1}{n}\right)$. The product of this bias and the sample variance is equal to the population variance. Hence the use of the correction factor $\left(\frac{1}{n-1}\right)$ for this estimate (see the Post Commentary for a derivation).

Post Commentary

An exploration of Focus 3 requires a binomial expansion and substitution using the expected value function. The bias is equal to $\left(\frac{n-1}{n}\right)$. To understand why this is the case, we expand the first term in the sum of squared deviations.

$$(x_1 - \bar{x})^2 = x_1^2 - 2x_1\bar{x} + \bar{x}^2$$

Substituting from the definition of mean, we obtain,

$$x_1^2 - 2x_1\bar{x} + \bar{x}^2 = x_1^2 - 2x_1\left(\sum \frac{x_i}{n}\right) + \left(\sum \frac{x_i}{n}\right)^2$$

Taking the expected value, we obtain,

$$x_1^2 - 2x_1\left(\sum \frac{x_i}{n}\right) + \left(\sum \frac{x_i}{n}\right)^2 = x_1^2 - 2x_1\frac{x_1}{n} + \left(\frac{x_1}{n}\right)^2$$

$$= x_1^2 - 2\frac{x_1{}^2}{n} + \frac{x_1{}^2}{n^2} = \frac{n^2x_1^2 - 2nx_1^2 + x_1^2}{n^2} = \frac{(n^2 - 2n + 1)x_1^2}{n^2}$$

$$= \frac{(n-1)^2 x_1^2}{n^2} = \left(\frac{n-1}{n}\right)^2 x_1^2$$

Multiplying the sample variance by the bias we obtain the population variance,

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

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

Thus, the unbiased estimate of the variance is given by,

$$s^2 = \frac{\sigma^2}{\left(\frac{n-1}{n}\right)} = \left(\frac{n}{n-1}\right)\cdot\sigma^2 = \left(\frac{n}{n-1}\right)\cdot\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n} = \left(\frac{n}{n-1}\right)\cdot\left(\frac{1}{n}\right)\cdot\sum_{i=1}^{n}(x_i-\bar{x})^2$$

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

Furthermore, from the perspective of Mathematical Statistics, division by (n-1) makes S^2 an unbiased estimator of σ^2, the population variance. The following theorem is equivalent to Focus 3:

<u>Definition</u>: A statistic $\hat{\Theta}$ is an unbiased estimator of the parameter θ if and only if $E\left(\hat{\Theta}\right) = \theta$ (in other words if the expected value of the statistic equals the parameter).

<u>Theorem</u>: If S^2 is the variance of a random sample from an infinite population with the finite variance σ^2, then $E(S^2) = \sigma^2$.

<u>Proof</u>:

$$E(S^2) = E\left(\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}\right) = \frac{1}{n-1}\cdot E\left(\sum_{i=1}^{n}\left[(x_i-\mu)-(\bar{x}-\mu)\right]^2\right)$$

$$= \frac{1}{n-1}\cdot\left\{\sum_{i=1}^{n}E\left[(x_i-\mu)^2\right] - n\cdot E\left[(\bar{x}-\mu)^2\right]\right\}$$

Then, because $E\left[(x_i-\mu)^2\right] = \sigma^2$ and since $E\left[(\bar{x}-\mu)^2\right] = \frac{\sigma^2}{n}$, we have that

$$E\left(S^2\right)=\frac{1}{n-1}\cdot\left\{\sum_{i=1}^{n}\sigma^2-n\cdot\frac{\sigma^2}{n}\right\}=\sigma^2$$

and thus, by the definition, the sample variance is an unbiased estimator of the population variance.■

References

Agresti, A. & Franklin, C. (2007). *Statistics: The art and science of learning from data*. Upper Saddle River, NJ: Pearson/Prentice Hall.

Freund, J. E. (1992). *Mathematical Statistics* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Mendenhall, W., Beaver, R. J. & Beaver, B. M. (2005). *Introduction to Probability and Statistics (12th ed.)*. Pacific Grove, CA: Thomson Brooks/Cole.

http://arnoldkling.com/apstats/df.html

http://www.keithbower.com/pdflibrary/why_divide_by_n-1.pdf

http://courses.ncssm.edu/math/Stat_Inst2001/Section1/A03%20Why%20N%20Minus%20One.pdf

what critical information has been discovered
sum of deviations equal zero with a sense of why?

find one-line description for each focus and use as a guide to pair down (essence) focus.

Read through framework and work through other situations.
Work on non-final, modify.

Next meeting: July 15th Tuesday 10:00.