Sample Variance (3 pages; 8/7/21)

(1) Strictly speaking, the variance of a sample is defined as

the average squared deviation from the sample mean;

ie $s^2 = \frac{1}{n}\sum(x_i - \overline{x})^2$

(and then the standard deviation is $s = \sqrt{\frac{1}{n}\sum(x_i - \overline{x})^2}$).

However, if the sample variance is intended to be an estimate for the population variance, then it can be shown that an unbiased estimate of the population variance is

$s^2 = \frac{1}{n-1}\sum(x_i - \overline{x})^2$

This means that, if we define S^2 to be the random variable

$\frac{1}{n-1}\sum(X_i - \overline{X})^2$, then $E(S^2) = \sigma^2$, the population variance.

[See the Appendix for a proof of this. The fact that we are using $\overline{x}$ (the sample mean) in the formula, instead of the population mean μ, means that the n deviations $x_i - \overline{x}$ are not independent (for example, $x_n - \overline{x}$ can be determined, if the other deviations are known). This suggests that an average of n deviations isn't appropriate, but doesn't constitute a proof that a divisor of $n-1$ gives the right value.]

For exam purposes, $s^2 = \frac{1}{n-1}\sum(x_i - \overline{x})^2$ is usually preferred, even when there is no mention of it being an estimate for the population variance.

(2) Alternative, and generally more convenient formulae are:

$s^2 = \frac{1}{n}\sum(x_i - \overline{x})^2 = \frac{1}{n}\{(\sum x_i{}^2) - n\overline{x}^2\}$

and $s^2 = \frac{1}{n-1}\sum(x_i - \overline{x})^2 = \frac{1}{n-1}\{(\sum x_i{}^2) - n\overline{x}^2\}$ when the divisor of $n-1$ is being used

Proof (for $s^2 = \frac{1}{n}\sum(x_i - \overline{x})^2$)

$s^2 = \frac{1}{n}\sum(x_i - \overline{x})^2$

$= \frac{1}{n}\{\sum(x_i{}^2 - 2x_i\overline{x} + \overline{x}^2)\}$

$= \frac{1}{n}\{(\sum x_i{}^2) - 2\overline{x}(\sum x_i) + n\overline{x}^2\}$

$= \frac{1}{n}\{(\sum x_i{}^2) - 2\overline{x}(n\overline{x}) + n\overline{x}^2\}$

$= \frac{1}{n}\{(\sum x_i{}^2) - n\overline{x}^2\}$

A useful check is as follows:

If all the n data items are the same, then each $x_i = \overline{x}$,

and $\sum x_i{}^2 = n\overline{x}^2$, so that $s^2 = 0$; as expected, since there is no variance amongst the x_i.

Notes

(i) $(\sum x_i{}^2) - n\overline{x}^2$ is often denoted S_{xx}

(ii) $(\sum x_i{}^2) - n\overline{x}^2$ can also be written as $(\sum x_i{}^2) - \frac{(\sum x_i)^2}{n}$

(iii) It is tempting to write $\frac{1}{n}\left\{(\sum x_i{}^2) - n\overline{x}^2\right\}$ as $\frac{1}{n}(\sum x_i{}^2) - \overline{x}^2$,

but $\frac{1}{n}\left\{(\sum x_i{}^2) - n\overline{x}^2\right\}$ has the advantage that it can easily be converted to $\frac{1}{n-1}\left\{(\sum x_i{}^2) - n\overline{x}^2\right\}$ if necessary.

Appendix

$S^2 = \frac{1}{n-1}([\sum X^2] - n\overline{X}^2)$ is an unbiased estimator for the population variance

Proof

$E(S^2) = \frac{1}{n-1}\{[\sum E(X^2)] - nE(\overline{X}^2)\}$

$= \frac{1}{n-1}\{nE(X^2) - nE\left(\overline{X}^2\right)\}$

$= \frac{n}{n-1}\{E(X^2) - E\left(\overline{X}^2\right)\}$

Now $\sigma^2 = E(X^2) - \mu^2$, so that $E(X^2) = \sigma^2 + \mu^2$

Also, $Var(\overline{X}) = E\left(\overline{X}^2\right) - \mu^2$,

and $Var(\overline{X}) = Var(\frac{X_1 + \cdots + X_n)}{n} = \frac{1}{n^2}Var(X_1 + \cdots + X_n)$

$= \frac{1}{n^2}\left(nVar(X_i)\right) = \frac{\sigma^2}{n}$,

so that $E\left(\overline{X}^2\right) = \frac{\sigma^2}{n} + \mu^2$

Then $E(S^2) = \frac{n}{n-1}([\sigma^2 + \mu^2] - \left[\frac{\sigma^2}{n} + \mu^2\right])$

$= \frac{n\sigma^2}{n-1}(1 - \frac{1}{n})$

$= \frac{n\sigma^2}{n-1}(\frac{n-1}{n}) = \sigma^2$