WHY DOES THE SAMPLE VARIANCE HAVE N-1 IN THE DENOMINATOR?

In Chapter 4 (p. 59), the sample variance of a sample y_1, y_2, … , y_n was defined as

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1},$$

and described as "almost" the mean of the squared deviations $(y_i - \bar{y})^2$. It might seem more natural to use an n in the denominator, so that we really have the mean of the squared deviations (which we'll abbreviate as mosqd),

$$\text{mosqd } = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}$$

The reason we use n-1 rather than n is so that the sample variance will be what is called an *unbiased estimator* of the population variance σ^2. To explain what this means, we first define the term *estimator:*

> An *estimator* is a random variable whose underlying random process is choosing a sample, and whose value is a statistic (as defined on p. 285), based on that sample, that is used to estimate a population parameter.
>
> Examples:
>
> - $\hat{p}$ (considered as a random variable) is an estimator of p, the population proportion. (We'll use $\widehat{P}$ to denote the random variable and reserve $\hat{p}$ for the statistic that is the value of the random variable for a particular sample.)
> - $\bar{y}$ (considered as a random variable) is an estimator of μ, the population mean. (We'll use $\bar{Y}$ to denote the random variable and reserve $\bar{y}$ to denote the statistic that is the value of the random variable for a particular sample.)
>
> Note that the concepts of *estimate* and *estimator* are related but not the same: a particular value (calculated from a particular sample) of the estimator is an estimate.
>
> - If we use the notation introduced in parentheses above, then $\widehat{P}$ and $\bar{Y}$ are estimators, while $\hat{p}$ and $\bar{y}$ are estimates.
> - The distinction between estimator and estimate is similar to the distinction between a function and the value of a function when a particular number is plugged into the function.
>
> Note also that the distribution of an estimator is the sampling distribution of the related statistic. (e.g., the distribution of $\widehat{P}$ is the sampling distribution of $\hat{p}$.)

An *unbiased estimator* is an estimator whose expected value (i.e., the mean of the distribution of the estimator) is the parameter being estimated. (Intuitively, this seems like a desirable property for an estimator.)

Examples:

- The calculation E($\widehat{P}$) = p on p. 434 shows that the sample proportion $\widehat{P}$ is an unbiased estimator of the population proportion p.
- The sample mean $\bar{Y}$ is an unbiased estimator of the population mean μ: E($\bar{Y}) = \mu$. (This is not difficult to prove, using the definition of sample mean and properties of expected values.)

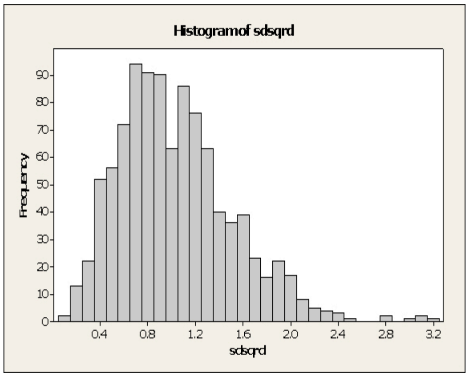
If we consider mosqd as an estimate of σ^2, we get a corresponding estimator, which we'll call MOSqD: The process for MOSqD is picking a random sample from the population for Y, and the value of MOSqD is mosqd calculated from that sample. (Note that the distribution of MOSqD is the sampling distribution of mosqd.)

We want to ask: Is MOSqD an unbiased estimator of the population variance σ^2?

- In other words, is σ^2 the expected value of MOSqD?
- In equation form: Is E(MOSqD) = σ^2?

To check this out informally, 1000 samples of size 10 from a standard normal distribution were generated. For each sample, mosqd was calculated. If MOSqD is an unbiased estimator of the population variance (which in this case is 1, since samples were from a standard normal distribution), the mean of the 1000 values of MOSqd should be pretty close to 1. This mean was in fact 0.9288 -- not very close to 1. But the mean of the values of the 1000 sample variances was 1.0320, which *is* pretty close to 1.

Comment: Here is a histogram of the sample variances from these 1000 samples. Note that it does not look like it represents a normal distribution. In fact, the sampling distribution of variances is *not* normal – although if we used samples of size noticeably larger than 10, we would get a distribution that was closer to normal.



We will prove that the sample variance, S^2 (not MOSqD) is an unbiased estimator of the population variance σ^2.

- Note: To help distinguish between the estimator and an estimate for a particular sample, we are using S^2 to stand for the estimator (random variable) and s^2 to stand for a particular value of S^2 (i.e., s^2 stands for the sample variance of a particular sample.)

The proof will use the following two formulas:

(1) $\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} {y_i}^2$ - $n\bar{y}^2$

(Note that this gives an alternate formula for the numerator of the formula for the sample variance s^2)

> *Exercise*: Prove formula (1). [Hint: Multiply out $(y_i - \bar{y})^2$ and use properties of summations and the definition of $\bar{y}$.]

(2) For any random variable Y, $\mathrm{Var}(Y) = E(Y^2) - [E(Y)]^2$.

> You have probably seen the proof of this in M362K. If not, or you'd like to refresh your memory, you can prove it yourself, starting with the definition of $\mathrm{Var}(Y) = E([Y - E(Y)]^2)$, multiplying out $[Y - E(Y)]^2$, and using properties of E.

We will use formula (2) in the rearranged form

(3) $E(Y^2) = \mathrm{Var}(Y) + [E(Y)]^2$

Proof that S^2 is an unbiased estimator of the population variance σ^2:

(This proof depends on the assumption that sampling is done with replacement.)

Let Y_i denote the random variable whose process is "choose a random sample $y_1, y_2, \dots, y_n$ of size n" from the random variable Y, and whose value for that choice is y_i. With this notation, the formula for s^2 translates into a formula for S^2:

$$S^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} Y_i^2 - n\left(\frac{\sum_{i=1}^{n} Y_i}{n}\right)^2\right]$$

$$= \frac{1}{n-1}\sum_{i=1}^{n} Y_i^2 - \frac{1}{n(n-1)}\left(\sum_{i=1}^{n} Y_i\right)^2$$

So

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

$$= \frac{1}{n-1}\sum_{i=1}^{n} E(Y_i^2) - \frac{1}{n(n-1)} E\left(\left(\sum_{i=1}^{n} Y_i\right)^2\right) \quad \text{(using properties of E)}$$

Now apply formula (3) to each term $E(Y_i^2)$ and $E((\sum_{i=1}^{n} Y_i)^2)$ to get

$$E(S^2) = \frac{1}{n-1}\sum_{i=1}^{n}[Var(Y_i) + [E(Y_i)]^2] - \frac{1}{n(n-1)}[Var(\sum_{i=1}^{n} Y_i) + [E(\sum_{i=1}^{n} Y_i)]^2]$$

Since each Y_i represents a choice from the random variable Y, we know that each Y_i has the same distribution (hence the same mean and variance) as Y. So if μ and σ^2 are the mean and variance of Y, then $E(Y_i) = \mu$ and $\mathrm{Var}(Y_i) = \sigma^2$ for each i. Also, since the samples are chosen without replacement, we know that the Y_i's are independent. Using these (plus the additivity property of E), we can simplify the expression above to get

$$E(S^2) = \frac{1}{n-1}\sum_{i=1}^{n}[\sigma^2 + \mu^2] - \frac{1}{n(n-1)}[Var(\sum_{i=1}^{n} Y_i) + [\sum_{i=1}^{n} E(Y_i)]^2]$$

by independence of Y_i's

$$= \frac{n}{n-1}(\sigma^2 + \mu^2) - \frac{1}{n(n-1)}\left[\sum_{i=1}^{n} Var(Y_i) + (n\mu)^2\right]$$

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

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

$$= \frac{n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2}{n-1} = \sigma^2$$

(Be sure to check where each property is used in the calculations!)