Friday, Feb 11

Why Do We Divide by $n-1$ When Computing s^2

Recall that the variance (s^2) for a sample of n observations is

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

Why divide by $n-1$ rather than n?

Consider the following population distribution.

x	$P(x)$
1	0.1
2	0.3
3	0.6

The mean of x is

$$\mu_x = 1 \times 0.1 + 2 \times 0.3 + 3 \times 0.6 = 2.5,$$

and the variance of x is

$$\sigma_x^2 = (1-2.5)^2 \times 0.1 + (2-2.5)^2 \times 0.3 + (3-2.5)^2 \times 0.6 = 0.45.$$

But in practice we would not know σ_x^2, but we could use the variance from a *sample* of observations (s^2) to estimate σ_x^2. Recall that s^2 is defined as

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

What is the sampling distribution of s^2 if $n = 2$?

Sample	Probability	s^2
1,1	0.01	0.0
2,1	0.03	0.5
3,1	0.06	2.0
1,2	0.03	0.5
2,2	0.09	0.0
3,2	0.18	0.5
1,3	0.06	2.0
2,3	0.18	0.5
3,3	0.36	0.0

The mean of s^2 is 0.45. Note that this equals σ_x^2. So s^2 is *unbiased* when we use it to estimate σ_x^2.

s^2	$P(s^2)$
0.0	0.46
0.5	0.42
2.0	0.12

But now suppose we compute the sample variance by dividing by n rather than $n-1$ so that

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

Now consider the sampling distribution of this version of s^2 when $n = 2$.

Sample	Probability	s^2
1,1	0.01	0.00
2,1	0.03	0.25
3,1	0.06	1.00
1,2	0.03	0.25
2,2	0.09	0.00
3,2	0.18	0.25
1,3	0.06	1.00
2,3	0.18	0.25
3,3	0.36	0.00

s^2	$P(s^2)$
0.00	0.46
0.25	0.42
1.00	0.12

The mean of s^2 is now 0.225. But this is less than σ_x^2 which is 0.45, so now s^2 is *biased* if we were to use it to estimate σ_x^2.

Example: Consider a survey of Hobbits in the Shire. Suppose that the mean height of all Hobbits is 107 cm and the standard deviation is 10 cm. Suppose we plan to select a sample of 100 Hobbits to estimate the mean height of all Hobbits in the Shire using the mean height from the sample. What do we know about the sampling distribution of the mean height from the sample?

Example: Suppose that 20% of Hobbits in the Shire have foot lice. We plan to select a sample of 100 Hobbits to estimate the proportion of Hobbits in the Shire that have foot lice by using the proportion of Hobbits in the sample that have foot lice. What do we know about the sampling distribution of the proportion from the sample?

Example: Consider again Darwin's fertilization study. Suppose we assume, for the sake of argument, that type of fertilization does not affect height. Then the mean of the population distribution of differences should be *zero*. Assume that this is true and that this population distribution as a standard deviation of 5. What do we know about the sampling distribution of the mean height difference in a sample of 15 observations?

Example: In one of Gregor Mendel's well know studies, he raised pea plants and observed the color of the offspring as green or yellow. Based on his ideas about inheritance, he thought that the probabilities of observing a green or yellow offspring are 0.25 and 0.75, respectively. In one study he observed that in a sample of 8023 observations, 6022 were yellow and 2001 were green. If Mendel is correct that the probability of a yellow offspring is 0.75, what do we know about the sampling distribution of the proportion of yellow offspring in a sample of 8023 observations?