Psy 216 Your Name _____
Dr. Crutcher's Class

Estimating the Variability of a Population from a Sample

Most of the time, when we compute the variability of a sample, we begin by doing the same thing as we did for a population. We first compute the mean for the sample. We then compute the deviations or differences between each of the scores and the mean. Squaring these differences and adding them up gives us the sum of squares. Dividing by the number of scores gives us the average variability or in other words, the **variance**.

However, when we select only a subset of the scores from a population, we are very unlikely to have any of the most extreme scores in our sample. The extreme scores remember are very unlikely to occur and are part of the extreme tails of the population distribution. This means that our estimate of the population variability or variance using our sample will tend to be somewhat smaller than the ACTUAL variability of the population from which the sample came. In other words, the estimate of the population variability is **biased,** because it consistently under-estimates the true variability in the population. We call our sample estimate of the population parameter, a **biased estimator.**

Is there some way we can fix this problem of consistently under-estimating the true variability of the population from which the sample came? Yes, when we compute the variability from the sample, instead of dividing the sum of the squared deviations by **n,** we can divide this sum by **n-1**. Why? Dividing by n-1 we are dividing by a slightly smaller number which causes the result of our dividing (the variability estimate) to be slightly larger than it would have been. In effect, we have boosted our estimate of the variability slightly--exactly what we want given that our sample estimate of the variability is too small.

Notice that there is a really nice feature to this procedure. What happens as the size of the sample increases? The estimate of the variability using **n-1** gets closer and closer to the estimate using **n**. If the sample is large enough, for all practical purposes the two estimates become indistinguishable. But that makes sense. What's happening as we select a larger and larger sample? Our sample more and more closely approximates the population in its characteristics. Therefore, the variability of the sample is closer and closer to the variability of the population. Hence the variability of the sample (using **n**) becomes indistinguishable from the estimate of the population variability (using **n-1**).

There is another way to see that using n-1 when estimating variability with a sample is a closer estimate to the true population variability: We can do a simulation.

Imagine the following situation: We have a population consisting of only 3 scores 1, 2, & 3. First compute the mean and variability for that population. The mean, $\mu = 2$ (Note that (1+2+3)/3 = 2) and the variance is .67.

x	x-μ	$(x-\mu)^2$
1	1-2	1
2	2-2	0
3	3-2	1

$$\sum (x-\mu)^2 = 2$$

$$\sigma^2 = \frac{\sum (x-\mu)^2}{N} = \frac{2}{3} = .666$$

Now let's try to estimate the variance using a sample. We will take all possible samples of size 2 and estimate the variance of the population:

Sample	$\bar{x}$	$(x-\bar{x})^2 / n$	$(x-\bar{x})^2 / n-1$
1,1	1.0	0.00	0.00
1,2	1.5	0.25	0.50
1,3	2.0	1.00	2.00
2,1	1.5	0.25	0.50
2,2	2.0	0.00	0.00
2,3	2.5	0.25	0.50
3,1	2.0	1.00	2.00
3,2	2.5	0.25	0.50
3,3	3.0	0.00	0.00
Overall Avg of Samples	2.0	0.33	0.67

Notice that if you take the average of the sample means you get the mean of the population (which is 2). Notice also that if you take the average of the sample variances, the estimate using **n-1** (.67) corresponds to the variance of the population (.67) whereas the estimate using **n** (.33) dramatically underestimates the variance in the population. In other words, on average, the estimate of the population variance from a sample is going to be closer to the actual value of the population variance when you use **n-l** rather than **n** to compute the variance. Especially when the sample size is small. As the sample size grows the variance of the sample will become closer to the true value of the population variance and it really won't matter whether you use **n** or **n-1**.