8.8: Variability- How Well Does the Mean Fit the Data?

Once we have described the central tendency of the data, we often also want to describe how variable the data are – this is sometimes also referred to as "dispersion", reflecting the fact that it describes how widely dispersed the data are.

We have already encountered the sum of squared errors above, which is the basis for the most commonly used measures of variability: the *variance* and the *standard deviation*. The variance for a population (referred to as σ^2) is simply the sum of squared errors divided by the number of observations - that is, it is exactly the same as the *mean squared error* that you encountered earlier:

$$\sigma^2 = \frac{SSE}{N} = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{N}$$

where μ is the population mean. The standard deviation is simply the square root of this – that is, the *root mean squared error* that we saw before. The standard deviation is useful because the errors are in the same units as the original data (undoing the squaring that we applied to the errors).

We usually don't have access to the entire population, so we have to compute the variance using a sample, which we refer to as $\hat{\sigma}^2$, with the "hat" representing the fact that this is an estimate based on a sample. The equation for $\hat{\sigma}^2$ is similar to the one for σ^2:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{X})^2}{n-1}$$

The only difference between the two equations is that we divide by n - 1 instead of N. This relates to a fundamental statistical concept: *degrees of freedom*. Remember that in order to compute the sample variance, we first had to estimate the sample mean $\bar{X}$. Having estimated this, one value in the data is no longer free to vary. For example, let's say we have the following data points for a variable x: [3, 5, 7, 9, 11], the mean of which is 7. Because we know that the mean of this dataset is 7, we can compute what any specific value would be if it were missing. For example, let's say we were to obscure the first value (3). Having done this, we still know that its value must be 3, because the mean of 7 implies that the sum of all of the values is $7 * n = 35$ and $35 - (5+7+9+11) = 3$.

So when we say that we have "lost" a degree of freedom, it means that there is a value that is not free to vary after fitting the model. In the context of the sample variance, if we don't account for the lost degree of freedom, then our estimate of the sample variance will be *biased*, causing us to underestimate the uncertainty of our estimate of the mean.