3.14: Where Does the N - 1 Come from?

If we know μ and we have a set of N data points, the best estimate we can make of the variance is

$$\sigma^2 = \int_{u_{min}}^{u_{max}} (u-\mu)^2 \left(\frac{df}{du}\right) du \approx \sum_{i=1}^{N} (u_i-\mu)^2 \left(\frac{1}{N}\right)$$

We have said that if we must use $\overline{u}$ to approximate the mean, the best estimate of σ^2, usually denoted s^2, is

$$estimated\ \sigma^2 = s^2 = \sum_{i=1}^{N} (u_i-\overline{u})^2 \left(\frac{1}{N-1}\right)$$

The use of $N-1$, rather than N, in the denominator is distinctly non-intuitive; so much so that this equation often causes great irritation. Let us see how this equation comes about.

Suppose that we have a distribution whose mean is μ and variance is σ^2. Suppose that we draw N values of the random variable, u, from the distribution. We want to think about the expected value of $(u-\mu)^2$. Let us write $(u-\mu)$ as

$$(u-\mu) = (u-\overline{u}) + (\overline{u}-\mu).$$

Squaring this gives

$$(u-\mu)^2 = (u-\overline{u})^2 + (\overline{u}-\mu)^2 + 2(u-\overline{u})(\overline{u}-\mu).$$

From our definition of expected value, we can write:

$$\begin{aligned} &\text{Expected value of } (u-\mu)^2 = \\ &\quad = expected\ value\ of\ \ (u-\overline{u})^2 \\ &\quad + expected\ value\ of\ \ (\overline{u}-\mu)^2 \\ &\quad + expected\ value\ of\ 2(u-\overline{u})(\overline{u}-\mu) \end{aligned}$$

From our discussion above, we can recognize each of these expected values:

- The expected value of $(u-\mu)^2$ is the variance of the original distribution, which is σ^2. Since this is a definition, it is exact.
- The best possible *estimate* of the expected value of $(u-\overline{u})^2$ is

$$\sum_{i=1}^{N} (u_i-\overline{u})^2 \left(\frac{1}{N}\right)$$

- The expected value of $(\overline{u}-\mu)^2$ is the expected value of the variance of averages of N random variables drawn from the original distribution. That is, the expected value of $(\overline{u}-\mu)^2$ is what we would get if we repeatedly drew N values from the original distribution, computed the average of each set of N values, and then found the variance of this new distribution of average values. By the central limit theorem, this variance is σ^2/N. Thus, the expected value of $(\overline{u}-\mu)^2$ is exactly σ^2/N.
- Since $(\overline{u}-\mu)$ is constant, the expected value of $2(u-\overline{u})(\overline{u}-\mu)$ is

$$2(\overline{u}-\mu)\left[\frac{1}{N}\sum_{i=1}^{N}(u_i-\overline{u})\right]$$

  which is equal to zero, because

$$\sum_{i=1}^{N}(u_i-\overline{u}) = \left(\sum_{i=1}^{N} u_i\right) - N\overline{u} = 0$$

  by the definition of $\overline{u}$.

Substituting, our expression for the expected value of $(u-\mu)^2$ becomes:

$$\sigma^2 \approx \sum_{i=1}^{N} (u_i - \overline{u})^2 \left(\frac{1}{N}\right) + \frac{\sigma^2}{N}$$

so that

$$\sigma^2 \left(1 - \frac{1}{N}\right) = \sigma^2 \left(\frac{N-1}{N}\right) \approx \sum_{i=1}^{N} \frac{(u_i - \overline{u})^2}{N}$$

and

$$\sigma^2 \approx \sum_{i=1}^{N} \frac{(u_i - \overline{u})^2}{N-1}$$

That is, as originally stated, when we must use $\overline{u}$ rather than the true mean, μ, in the sum of squared differences, the best possible *estimate* of σ^2, usually denoted s^2, is obtained by dividing by $N-1$, rather than by N.
