Lecture 8: Probability & Estimation
Notation and Equations

Notation

P: probability. *P*(*event*) means the probability that *event* will occur, where *event* can be anything like $X = x$, or $X \leq x$, or $M > 0$. $P(X = x)$ can be abbreviated as $P(x)$, as we've been doing already.

R: a random variable (we won't use the generic name *R* after this lecture, because the random variables we care about in the future will have specific names)

s^2: sample variance; serves as an unbiased estimator of population variance, σ^2

s: sample standard deviation

Formula for expected value. The expected value of a random variable is the same as the mean of its probability distribution. The probability distribution is essentially an infinite population (the population of all times that we could sample the random variable), so to get its mean we use the same formula as for the mean of an infinite population.

$$E(R) = \sum_x x \cdot P(R = x) \qquad (1)$$

Another useful way to think of this formula is as a "weighted average." The formula averages all of the values *x* that could occur, with each value weighted by its probability, *P*(*x*). If *R* has a finite number of possible values, all of which are equally likely, then Equation 1 will become a simple average (mean) of those values. If some values are more probable than others, the expected value will be closer to the more probable values. If there is a value *x* for which *P*(*x*) is very close to 1, meaning that *x* is almost certain to be the value that is sampled, then Equation 1 gives an expected value that's very close to *x*.

It should be clear from the discussion above, but the mean of a random variable is the same as its expected value.

$$mean(R) = E(R) \qquad (2)$$

Properties of expected value. If you add a constant to a random variable, its expected value is changed by that constant. (Here *c* is any number.)

$$E(R + c) = E(R) + c \qquad (3)$$

If you multiply a random variable by a constant, its expected value is multiplied by that constant.

$$E(c \cdot R) = c \cdot E(R) \qquad (4)$$

The expected value of the sum of two random variables is the sum of their expected values. (For Equations 5 & 6, R_1 and R_2 are two different random variables.)

$$E(R_1 + R_2) = E(R_1) + E(R_2) \tag{5}$$

However, the expected value of the product of two random variables is normally **not** the product of their expected values. We'll learn more about this when we get to correlations.

$$E(R_1 \cdot R_2) \neq E(R_1) \cdot E(R_2) \tag{6}$$

<u>Relationship between sample mean and population mean</u>. The expected value of the sample mean, M, equals the population mean, μ.

$$E(M) = \mu \tag{7}$$

The easiest way to see this is to notice that the expected value of each individual observation, $E(X)$, equals μ. This follows directly from the definition of expected value. Since M is the average of the individual observations, its expected value is also μ. This follows from the basic properties of expected value:

$$\begin{aligned}
E(M) &= E\left(\frac{\sum_{sample} X}{n}\right) \\
&= \tfrac{1}{n} E\left(\sum_{sample} X\right) && \text{Multiplying a random variable by a number multiplies its expected value by the same number} \\
&= \tfrac{1}{n} \sum_{sample} E(X) && \text{The expected value of a sum of random variables is the sum of their expected values} \\
&= \tfrac{1}{n} \cdot n\mu && \text{Because we're adding } \mu \text{ } n \text{ times} \\
&= \mu
\end{aligned}$$

*<u>Variance as an expected value</u>. The population variance is the expected value of the squared deviation from the mean. This is obvious from the formula for variance: $\sigma^2 = \Sigma_{pop}(X - \mu)^2/N$. This is the average (mean) of $(X - \mu)^2$. Since mean and expected value are the same thing, σ^2 equals the expected value of $(X - \mu)^2$. This gives another way of thinking about variance.

$$\sigma^2 = E\left((X - \mu)^2\right) \tag{8}$$

*<u>Estimation of population variance</u>. We can use Equation 8 to create a candidate statistic for estimating the population variance. This statistic takes the squared deviations from μ and averages over all scores in the sample. The expected value of this statistic equals the population variance.

*Optional sections. These are here to help you understand the formulas below for sample variance and standard deviation.

$$E\left(\frac{\sum_{sample}(X-\mu)^2}{n}\right)=\sigma^2 \tag{9}$$

Equation 9 is an immediate consequence of Equation 8, using the same logic we used to prove Equation 1: The expected value of $(X-\mu)^2$ equals σ^2 for every individual score, so the expected value of their average is also σ^2.

$$\begin{aligned}E\left(\frac{\sum(X-\mu)^2}{n}\right)&=\tfrac{1}{n}\sum E(X-\mu)^2\\&=\tfrac{1}{n}\cdot n\sigma^2\\&=\sigma^2\end{aligned}$$

The problem with the statistic in Equation 9 is that we can't determine μ from the sample. What about using the sample mean, M, instead? Unfortunately, this statistic is always less than the statistic based on μ.

$$\frac{\sum_{sample}(X-M)^2}{n}<\frac{\sum_{sample}(X-\mu)^2}{n} \tag{10}$$

One way to see this is to realize that M always deviates from μ in the direction of the sample. That is, M is closer to the sample than μ is.

Equations 9 and 10 show that $\Sigma_{\text{sample}}(X-M)^2/n$ is a biased estimator of σ^2. This is too bad, because $\Sigma(X-M)^2/n$ is the most obvious choice for defining sample variance.

$$E\left(\frac{\sum_{sample}(X-M)^2}{n}\right)<\sigma^2 \tag{11}$$

Sample variance. The solution to the bias problem is to divide by $n-1$ instead of n. This finally gives us the formula for sample variance.

$$s^2=\frac{\sum_{sample}(X-M)^2}{n-1} \tag{12}$$

Dividing by $n-1$ makes the result slightly larger, and it turns out this gives a perfectly unbiased estimator of population variance.

$$E(s^2)=\sigma^2 \tag{13}$$

Sample standard deviation. Just as with the population, the sample standard deviation equals the square root of the sample variance.

$$s = \sqrt{\frac{\sum_{sample} (X - M)^2}{n - 1}} \tag{14}$$