Lecture 10: Sampling Distributions
Notation and Equations

Notation

$p(M)$: the distribution of sample means, i.e. the probability distribution for M over repeated samples from the same distribution

σ_M: standard error of the mean, which equals the standard deviation of $p(M)$

Formula for standard error of the mean. The standard error of the mean is the standard deviation of $p(M)$, i.e. of the distribution of sample means. It tells you the typical distance between the true population mean, μ, and the sample mean, M. In other words, it tells you how much sampling error to expect when using M to estimate μ.

The standard error depends on two things: the standard deviation of the raw data (σ), and the sample size (n). As σ increases, the data become less reliable, so M also becomes less reliable (σ_M increases). As n increases, the sample becomes more reliable because it contains more data, so M becomes more reliable too (σ_M decreases).

$$\sigma_M = \frac{\sigma}{\sqrt{n}} \qquad (1)$$

Central Limit Theorem. The CLT tells us what the distribution of sample means looks like. First, the mean of $p(M)$ is the same as the population mean (μ). Second, the standard deviation of $p(M)$ is the population standard deviation (σ) divided by $\sqrt{n}$. Third, the shape of $p(M)$ will be very close to Normal as long as n is large enough.

$$mean\left(p(M)\right) = \mu \qquad (2)$$

$$SD\left(p(M)\right) = \frac{\sigma}{\sqrt{n}} \qquad (3)$$

$$p(M) \approx Normal\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \qquad (4)$$

Notice that Equation 2 is just a restatement of the fact that M is an unbiased estimator of μ, because the mean of $p(M)$ is the same as $E(M)$. Equation 3 is the same as Equation 1 above, because the standard deviation of $p(M)$ is the same as the standard error of M. Both Equations 2 and 3 are true for all values of n (all sample sizes).

The real meat of the CLT, and the part that you basically need a PhD in statistics to prove, is the fact that $p(M)$ is approximately Normal. This is not necessarily true for small samples, but once $n \geq 30$, we can assume $p(M)$ is Normal and use Equation 4 to completely determine the mathematical form of the distribution $p(M)$.