Sampling Distributions

Motivation: Suppose we want to estimate the mean of some population, the current average salary μ of 2010 Rose graduates, for example. We get an IID sample by getting a small simple random sample of 2010 alumni. We compute the sample $\bar{x}$ and use this as our estimate of the unkown value of μ. How close will $\bar{x}$ be to μ? In order to answer this question we need the **sampling distribution** of $\bar{x}$. First, a couple of definitions:

Parameter: a parameter is a numerical constant associated with a population or process. The mean salary μ of 2010 Rose grads is a parameter.

Statistic: a function of the sample. Let $\{x_1, x_2, \ldots, x_n\}$ denote our simple random sample of alumni salaries. Then $\bar{x}$ is a statistic since

$$\begin{aligned} \bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ &= f(x_1, x_2, \ldots, x_n) \end{aligned}$$

Sampling Distribution of a Statistic: The sampling distribution of a statistic is just its distribution. Since a statistic is a function of the sample, and the sample is comprised of random variables, the statistic is a random variable. Although not practical, if we were to get multiple random samples of alumni, $\bar{x}$ would vary from sample to sample, randomly assuming different values.

Exercise 1: Suppose we toss a coin 10 times. We estimate the probability of heads, $p = P(H)$ by the sample proportion $\hat{p}$, the proportion of heads in our 10 tosses. Answer the following:

1. What is the parameter in this scenario?
2. What is the statistic?
3. Describe, as well as you can, the sampling distribution of the statistic.

Sampling Distribution of $\bar{x}$

Overview: Now we will derive the sampling distribution of the sample mean for IID samples. The key to doing this to recognize that the sample mean is a **linear combination** of the sample values, i.e.

$$\bar{x} = c_1X_1 + c_2X_2 + \cdots + c_nX_n$$

where $c_1, c_2, \ldots, c_n$ are constants.

Exercise 2: What are the values of c_1, c_2, $\ldots$, c_n above?

Properties of Linear Combination of RV's: Let L be a linear combination of the RV's X_1, X_2, ..., X_n:

$$L = c_1X_1 + c_2X_2 + \cdots c_nX_n$$

Note that L is an RV. The following results allow us to determine important properties of L from knowledge of the RV's X_1, X_2, ..., X_n.

1. The mean of L, μ_L:
$$\mu_L = c_1\mu_{X_1} + c_2\mu_{X_2} + \cdots + c_n\mu_{X_n}$$
2. The variance of L, σ_L^2, **IF $\mathbf{X_1}$, $\mathbf{X_2}$, ..., $\mathbf{X_n}$ are independent:**
$$\sigma_L^2 = c_1^2\sigma_{X_1}^2 + c_2^2\sigma_{X_2}^2 + \cdots + c_n^2\sigma_{X_n}^2$$
3. Normality of L: If X_1, X_2, ..., X_n are normal, then L is exactly normal.

Sampling Distribution of $\bar{x}$: If the sample $\{X_1, X_2, \ldots, X_n\}$ is IID then

1. $\mu_{\bar{x}} = \mu$ where μ is the mean of the population or process being sampled.
2. $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ where σ is the standard deviation of the population or process.
3. $\bar{x}$ is exactly normal if the population or process is normal.
4. **Central Limit Theorem:** $\bar{x}$ is approximately normal if the population or process is not normal **but the sample size n is large.** Typically $n \geq 30$ is large enough for $\bar{x}$ to be very nearly normal.

Proof: Properties 1-3 follow from the properties of linear combinations above and the fact that

$$\bar{x} = \frac{1}{n}X_1 + \frac{1}{n}X_2 + \cdots + \frac{1}{n}X_n$$

i.e., $c_1 = c_2 = \cdots = c_n = 1/n$. The proof of the Central Limit Theorem requires probability methods beyond the scope of this course.

Exercise 3: Explore the properties of $\bar{x}$'s sampling distribution by using the following applet:

`http://onlinestatbook.com/stat_sim/sampling_dist/`

Exercise 4: What do the above facts about the sampling distribution of $\bar{x}$ imply about its usefulness as an estimator of an unknown population or process mean μ?

Exercise 5: Suppose, unknown to use, the salaries of 2010 Rose grads are normally distributed with a mean salary of $\mu = \$60000$ and standard deviation of $\sigma = \$5000$. Do the following:

1. If we get a IID sample of size 25, what's the probability that $\bar{x}$, our estimate of μ, will be within \$1000 of μ?
2. If we want to increase the above probability, what should we do?