Handout 01: Sample Mean

In statistics, we are typically describing a repeated measurement of a random process. The goal is to use the observations of those measurements to better understand the process. As a way to model this process, we will characterize the repeated measurements by a set of independent and identically distributed random variables from some distribution $\mathcal{G}$. In symbols, we have $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{G}$. We call this a **random sample** of size n.[1] Each component X_i is called an **observation**. Often, we will be concerned with the expected value of the distribution, which is known as the **population mean**. Similarly, the **population variance** is the variance of the distribution $\mathcal{G}$. Typically, we will use μ_X and σ_X^2 to stand for these population quantities. We can drop the X subscript if it is clear which distribution we are working with.

[1] There is much more vocabulary on today's worksheet than we typically have. Make sure you review the material today a few times until you are confident with all of the terminology. We will be using these throughout the entire semester.

A **statistic** or **sample statistic** is any random variable defined as a function of a random sample. One of the most common statistics that we will use is the **sample mean**. It is denoted and defined by:

$$\bar{X} = \frac{1}{n} \times [X_1 + \cdots + X_n] = \frac{1}{n} \times \sum_{i=1}^{n} X_i$$

On today's worksheet, we will show that:

$$\mathbb{E}[\bar{X}] = \mu_X, \quad \text{Var}[\bar{X}] = \frac{\sigma_X^2}{n}.$$

From these quantities, we can quickly see that if $\mathcal{G}$ is a normal distribution, then $\bar{X}$ will also be a normal distribution with the corresponding mean and variance above. This also hold in the limit of large n for other distributions as a result of the central limit theorem.

A **point estimator** is a sample statistic used to estimate a population parameter. Common notation of a point estimator is to put a *hat* over the parameter of interest: $\hat{\theta}$.[2] The **bias** of a point estimator $\hat{\theta}$ of μ is given by $\mathbb{E}\hat{\theta} - \theta$, the difference between the expected value of the point estimator and the quantity of interest. A point estimator is said to be **unbiased** if the bias is equal to zero. Furthermore an estimator $\hat{\theta}$ of θ is said to be **consistent** if for all $\epsilon > 0$ we have:

[2] One of the key things to keep track of this semester is which quantities are constants (such as μ_X) and which are random variables ($\hat{\mu}_X$). Do not let the fact that these look similar hide the fact that these are very different quantities.

$$\lim_{n \to \infty} \mathbb{P}\left[|\hat{\theta} - \theta| > \epsilon\right] = 0.$$

Generally, we want to use estimators that have as small (ideally, zero) bias and that are consistent. We will check these properties of estimators that we introduce throughout the semester. On the worksheet, we will show that $\bar{X}$ is an unbiased and consistent estimator for the population mean θ.