Lecture 2 : Distributions, Statistics, and Probability

In this lecture, we cover some basics to refresh our memory of statistics which we need to make the most out of this course. We start by defining what we mean by a sample mean, variance, and median and provide clear definitions of each before discussing some advanced topics. As most of you would be aware, we typically use what is called a *distribution* to define key characteristics of the data we obtained from an experiment. These characteristics include *location* (that measures where our observed samples are roughly centered), *spread* (that measures how much the observed samples extend over the location), and of course the *outliers* that simply fall out of the combined description of the distribution using location and spread. These are in some sense the "extreme" points observations of your experimental run. When you have a large number of results, the first set way to analyze data is by constructing what you all might know is the frequency distribution represented by a histogram. Let us now make a key distinction between a *population* and a *sample.* A population is what we can conceptually obtain if we can generate the data by repeatedly performing the experimental run infinitely. But for our purposes, we will assume that the data generation was run for a large number N. A sample on the other hand is what we would have got by running our experiment–a small set of observations. Essentially, the sample is a draw from the population. Let us now discuss what we call a *random draw* or a *random sample.* A random sample is a draw where each member of the population has an equal chance of being chosen. Suppose you have drawn randomly from a population and the value observed is y, having access to the population frequency distribution we can define a probability–namely y being greater, lesser, or within a particular range of interest. For a frequency histogram i.e. we have made the width of the rectangular boxes to be the relative frequency of values occurring in that interval n_i/N, the probability would be the area–to the right for greater, to the left for lesser and in between for the range. Essentially, within the accuracies of the relative frequency of the interval, we can get all the probability values we need from the histogram thus we call this a *probability distribution.* Because we define everything based on area, we make one more definition wherein for an interval of width h, and the rectangular area P, a *probability density* is defined as $p = P/h$. The name density also hints to us that it is not the probability but rather needs to be multiplied by the interval width h much like how mass is obtained by multiplying density with volume. However, most of you might be familiar with the notion of *probability density function* which is a continuous approximation when the interval width h becomes infinitesimally small. There are several examples of this such as normal distribution, Binomial distribution, etc. which we will cover in the next set of lectures. For randomly sampled data, we usually represent it using summary statistics such as the average, variance, and median. Note that we distinguished between a sample and a population so we use different definitions to summarize them. For a sample of n observations, its average is defined as $\bar{y} = \frac{\sum y}{n}$ and when the samples are a large number like N, we can obtain the population mean defined similarily as $\mu = \frac{\sum y}{N}$. This is the same as the location we described earlier using our intuition. It is common to distinguish between the summary of the population and a sample by defining the former as a parameter and the latter as a statistic. The parameter μ is In the experimental design, it is common to consider a random sample from the population, but often we observe that this does not apply to many real situations. Consider a generation situation where you are collecting temperature data. It is challenging to obtain a random sample in this case because warmer days tend to follow one another or in other words, you have some form of autocorrelation via continuity. The applicability of the random sample hypothesis is key to the design of experiments and we will discuss many techniques to ensure this assumption is relevant.

Measures of location and spread

The mean of a population is also referred to as the expected value denoted using $\mathbb{E}(y)$ or the first moment of an underlying distribution. Think of it as a measure of balance on a horizontal line. But most of you would know that we typically also need a *variance* a measure of how far aby particular y is from the population mean μ. It is defined as follows:

$$\sigma^2 = \mathbb{E}(y-\mu)^2 = \frac{\sum(y-\mu)^2}{N}$$

We can make similar definitions for measures pertaining to a sample instead of a population. Given a sample of points y and its corresponding mean $\bar{y}$, the sample variation similar to the σ^2 for the population is defined as :

$$s^2 = \frac{\sum(y-\bar{y})^2}{n-1}$$

The denominator of $n-1$ instead of n is an interesting thing to ponder. One reason is that the knowledge of the mean of the sample $\bar{y}$ already takes out one degree of freedom of the sample– because we can get the value of n-th sample if know the mean and values of other $n-1$ samples. Another commonly used term in the context of a sample is the signal-to-noise ratio which in statistics terms is defined as the ratio $\bar{y}/s$. The median of any sample is obtained by sorting the data values and taking the middle value of the sample. Sometimes median value is used instead (for example in statistics of income of a geographical location such as the King county MFTE programs) because it is a more accurate summary measure of income: Median household income is a more robust and accurate measure for summarizing income at the geographic level as compared to average household income since it is not affected by a small number of extremely high or low-income outlier households.