Lecture 4 : Randomness and random variables

So far, we have talked about what is a distribution and its applications to obtaining insights into obtaining measure and spread as well as answering simple questions about the probability of a value being above, below or in between two. In this lecture, we discuss the notion of a *random variable* which is a sample drawn from a known distribution for which we can estimate the probability of occurrence. We will later discuss how can we go from given samples to parameters of the known or assumed distribution using various estimation techniques. Before we jump into these things, we need to discuss a few ideas of statistical dependence, independence, and the ideas of IID assumption that are common across many analysis or design rules we will use throughout the rest of the course.

Statistical (in)dependence

Suppose we are sampling from a distribution of heights y_1 and weights y_2 of a class. Each one of the components of our distribution has some distributions $p(y_1), p(y_2)$ these are called marginal distributions where the marginalization (simply taking out the effects of other variables) is over the rest of the components. It is common to ask if knowing that y_1 or y_2 takes a certain value would tell us anything about the other. These are called *conditional distributions*: within the samples, what is the probability of y_1 given that value of y_2 is set to y_0 and is represented using $p(y_2|y_1 = y_0)$. The conditional distribution helps us identify whether two components of the sample are statistically dependent or independent. Intuitively, two components are statistically dependent if the probability of one of the components changes for different values of other components and independent otherwise. Another common question two ask about two random variables is the join probability where we would like to obtain the likelihood of a sample to have certain values for the two components: $\Pr(y_1 = a, y_2 = b)$. The famous Bayes theorem says that the joint probability can be factorized:

$$p(y_1, y_2) = p(y_1) \times p(y_2|y_1) = p(y_2) \times p(y_1|y_2).$$

The statistical dependence between two variables can now be verified using the Bayes theorem: two variables are statistically independent iff:

$$p(y_1, y_2) = p(y_1) \times p(y_2)$$

Application of the above concepts to observations of different repeat experiments results in a really interesting idea called independently and identically distributed (iid) data that is very common in many of the assumptions behind the usability of statistical or machine learning techniques. If the repeat observations from an experiment are both statistically independent and have an identical measure of location and spread, then we can consider the data to be coming from a single observation. However, it is very easy to violate this assumption in our daily experiments: if our samples have some inherent dependencies that we have not accounted for.

Randomizing an experiment

Consider the simple case of running an experiment where each one of the samples we made is different in one or two design variables we would like to study. Given the above definition of a random variable, we would want our samples to randomized such that we have a known distribution of design variables that we both would like to study and do not want to *confound* with our goal of the experiment. Confounding is said to occur when the effect of one factor cannot be distinguished from that of another factor.

Assuming that we have the i.i.d data, we next turn our focus on the problem of estimation. Recall that our single and multi-variable Normal distributions are parameterized by two parameters that need to be estimated to obtain the probability distribution of the underlying population. From our previous lectures, we know that we only ever have access to a sample from a population. One approach to estimate the parameters required is to use the maximization approach of a likelihood function that represents the likelihood of observing the data under a probability distribution with given parameters μ, σ. Suppose our sampled data is represented as $[y_1, y_2, \ldots, y_n]$, we can write the likelihood function as:

$$\begin{aligned}\mathcal{L}(\mu, \sigma) &= p(y_1)p(y_2)\ldots p(y_n) \\ &= \prod_{i=1}^{n} p(y_i) \\ &= \sigma^{-n/2}(2\pi)^{-n/2} \exp\left(-\frac{1}{2\sigma}\sum_{i=1}^{n}(y_i - \mu)^2\right)\end{aligned}$$

The parameters we would need are the maximizers of the functional form we obtained for $\mathcal{L}(\mu, \sigma)$. It can be derived that the *maximum likelihood estimates* (MLE) of our sample data to be a Normal distribution are the following:

$$\begin{aligned}\hat{\mu} &= \frac{\sum_{i=1}^{n} y_i}{n} \\ \hat{\sigma} &= \frac{\sum_{i=1}^{n}(y_i - \hat{\mu})^2}{n}\end{aligned}$$

We can also extend the MLE approach to the multi-variable case and obtain the following formulae given n samples of k-dimensional random vectors $X = [x_1, x_2, \ldots, x_k]$ as $X_1, X_2, \ldots, X_n$:

$$\begin{aligned}\hat{\mu} &= \frac{1}{n}\sum_{i=1}^{n} X_i \\ \hat{\sigma} &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^\top (X_i - \hat{\mu})\end{aligned}$$

The derivations of these are a good exercise and you can follow these references:

- Normal distribution : `https://online.stat.psu.edu/stat415/lesson/1/1.2`
- Multivariate Normal distribution : `https://stats.stackexchange.com/a/351550`

Another approach for estimation is to use some prior knowledge about the parameters of the population distribution. For example, if we are dealing with an experiment where the responses are percentage yields of a product from a reaction or some mechanical properties of a material, we can use our prior information from literature about similar studies to improve our estimation capabilities. For a single variable Normal distribution we would do this by first rewriting the likelihood function using the Bayes theorem:

$$\begin{aligned}\mathcal{L}(\mu, \sigma) &= p(\mu)p(y|\mu) \\ &= \frac{1}{\sigma_0\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\mu - \mu_0}{\sigma_0}\right)^2\right) \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{y_i - \mu}{\sigma}\right)^2\right)\end{aligned}$$

which is equivalent to minimizing the following function of μ:

$$\sum_{i=1}^{n} \left(\frac{y_i - \mu}{\sigma} \right)^2 + \left(\frac{\mu - \mu_0}{\sigma_0} \right)^2$$

using the optimizing technique of setting the derivative to zero, we obtain:

$$\hat{\mu}_{\text{MAP}} = \frac{\sigma_0^2 \sum_{i=1}^{n} y_i + \sigma^2 \mu_0}{\sigma_0^2 n + \sigma^2}$$

In practice, however, obtaining analytical solutions for MAP is quite tricky and we instead rely on computational techniques to get approximate solutions as commonly done in other aspects of science.