Aggregates

As we saw earlier, the probabilities associated with a (real-valued) random variable X are described by the distribution function F_X; where $F_X(t) = P(X \leq t)$. This is a non-decreasing function from $(-\infty, \infty)$ to $[0, 1]$ which tends to 0 on the left $(-\infty)$ and to 1 on the right $(+\infty)$; in other words, it has an "S" shape. Moreover, F_X is right continuous.

There are essentially three kinds of behaviours possible for such functions:

- *Discrete Jumps*: In this case, there is a discrete subset D of the real line so that the function is constant between two successive points of D. In other words, $P(X = t) \neq 0$ if and only if $t \in D$.
- *Density*: In this case there is a (non-negative) function $f_X(t)$ called the density function of X so that $F_X(t) = \int_{-\infty}^{t} f_X(s)ds$. In this case, $F_X(t)$ also called an absolutely continuous function.
- *Devil's Staircase*: These are somewhat strange *continuous* functions (e.g. Cantor's function, Minkowski's "?" function etc.) which only appear to increase "when we are not looking"! More precisely, there is a decreasing family of sets D_n so that $l(D_n) < 1/n$ and the derivative of the function exists and is zero *outside* D_n.

Due to the unusual nature of the third kind of functions, we will primarily focus on the first and second kind; we will call these the "discrete" and "continuous" case even though the second case should actually be called the "absolutely continuous" case. In fact, to study more general probability distributions we need to consider functions that exhibit a combination of the above behaviours.

Since there are still a large number of such functions, it is useful to find some "characteristics" of each distribution that give us some qualitative aspect of its behaviour. We shall consider such "aggregate" aspects in the current lecture.

Mean, Variance and Moments

The *mathematical expection* of X is defined as follows:

- Discrete Case: We define $E(X) = \sum_{d \in D} dP(X = d)$.
- Continuous Case: We define $E(X) = \int_{-infty}^{+\infty} sf_X(s)ds$.

Of course, in each case, there is an underlying assumption that the sum or integral exists! The mathematical expecation is also called the *mean* of the random variable X.

Suppose that the random variable X is "constructed" as follows. We have a (finite) population Ω of distinct entities w (say people, cars, physical objects, experiments and so on) and $X(w)$ is a numerical attribute associated with each entity (for example height of the person, the cost of a car, the volume of a physical object, the result of an experiment and so on). We then have some technique to pick elements of Ω "at random". Usually, this means that any entity in Ω is equally likely to be picked. In this case, it is not difficult to see that $E(X)$ is the same as the "average" of the numerical attribute over the population.

One of the powerful results in probability says that if we sample a population properly, the average of the sample is (with high probability) close the mathematical expectation $E(X)$ of the whole population. Thus, even when we do not know the actual probability distribution of X, we can estimate $E(X)$.

In order to give meaning to the expression "close to $E(X)$", we need to have some ways of measuring the deviation of X from $E(X)$. This is done by measuring the expectation of $(X - E(X))^k$ for various values of k.

More generally, if ϕ is a real valued function of a single variable, then we define:

- Discrete Case: We define $E(\phi(X)) = \sum_{d \in D} \phi(d) P(X = d)$.
- Continuous Case: We define $E(\phi(X)) = \int_{-infty}^{\infty} \phi(s) f_X(s) ds$.

The *Variance* of the random variable X is defined as $E((X - E(X))^2$ and is denoted by $\sigma^2(X)$. More generally, the k-th moment of X about its mean is $E((X - E(X))^k)$. The square root $\sigma(X)$ of $\sigma^2(X)$ is called its *standard deviation*. As we shall see later, it is a good measure (in a probabilistic sense) of how much X deviates from its mean.

One important identity that simplifies computations of the moments is $E(X + aY + b) = E(X) + aE(Y) + b$ where a and b are constants and X and Y are random variables. (Note that a constant c can be considered as "random" variable by declaring $P(c = c) = 1$ and $P(c = d) = 0$ for $c \neq d$!)

Using this identity, we see that:

$$\sigma^2(X) = E(X^2 - 2E(X)X + E(X)^2) = E(X^2) - 2E(X)^2 + E(X)^2 = E(X^2) - E(X)^2$$

Moreover, since $E(X + c) = E(X) + c$, we also see that $X + c - E(X + c) = X - E(X)$ so that:

$$\sigma^2(X + c) = E((X + c - E(X + c))^2) = E((X - E(X))^2) = \sigma^2(X)$$

More generally, the higher moments of X and $X + c$ are the same. This can be used to simplify a number of calculations by choosing a suitable constant c that reduces the size of the relevant values of $X + c$ that we need to take powers of.

Other Characteristics

In high school, we have learned the terms "median" and "mode". The latter is easier to define.

- Discrete case: mode is "the" value of the random variable which has the highest probability; i.e. it is m where $P(X = m)$ is the greatest among all $P(X = c)$ for all possible values of c. In other words, it is the "most likely".
- Continuous case: mode is "the" value for which the density function has the highest value; i.e. it is m where $f_X(m)$ is the the greatest among all possible $f_X(c)$ for all possible values of c.

The reason the "the" is in in quotes above is that such a value may not be unique. Moreover, in the continuous case it may not even exist since it is not too difficult (but a challenging exercise!) to write non-negative functions $f_X(t)$ so that $\int_{-\infty}^{+\infty} f_X(t)dt = 1$ but $f_X(t)$ is not bounded!

The median represents the smallest value of t so that $F_X(t) \geq 1/2$. (Recall that $F_X(t) = P(X \leq t)$ is the distribution function of X.) In other words, it is $m = \inf\{t|F_X(t) \geq 1/2\}$. Since F_X is right continuous (by the "infinite" law of probabilities), one can show that $F_X(m) \geq 1/2$.

More generally, for any number p between 0 and 1 we can look for the smallest value of t so that $F_X(t) \geq p$; this value of t is called the p-th quantile. This is specifically used to define "quartiles" for the case $p = 0.25$ (first quartile), $p = 0.5$ (median) and $p = 0.75$ (third quartile). It is also used to define percentile where the P-th percentile represents the smallest value of t so that $F_X(t) \geq P/100$.

Coarse Statistics

The mean, median, mode, variance, quartiles and percentiles represent the coarse (i.e. rough) information that we can gather about a population. Statistics is the science (or art!) of trying to determine the distribution of numerical attributes of a population by "sampling"; that is, measuring the numerical attributes of a sample of the entities from among the population.

If a random variable X represents the numerical attribute of an entity chosen randomly from the population (with all entities being equally likely), then the mean, median, mode, variance, and so on of the variable X is the same as the mean, median, mode, variance, and so on as computed by "high-school methods".

We will see later that when we use a suitable sampling technique we can (with high probability) get good estimates of these values from the same values for the *sample*. This idea is a key one in statistics.

Of course, it is unreasonable to ask for the p-quantile for *all* values of p unless we can determine the probability distribution precisely. For, if:

$$\phi(p) = \inf\{t | F_X(t) \geq p\}$$

then

$$F_X(t) = \sup\{p | \phi(p) \leq t\}$$

Hence, the *complete* quantile function determines the distribution function.