Lecture 2

1 Expectation and variance

Intuitively, the average of a data set is one way of describing the "centre" of the data set. It is not the only way; the median is another example. The average, or mean, is defined for a data set $x_1, \ldots, x_n$ as

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

There is a related quantity defined for random variables, called their expectation. It is defined as

$$E[X] = \sum_{n \in \mathbb{N}} x_n P(X = x_n),$$

if X is discrete and takes values in the set $\{x_1, x_2, \ldots\}$. It is defined as

$$E[X] = \int_{-\infty}^{\infty} x f(x) dx,$$

if X is continuous. (Both cases can be combined by defining it as $E[X] = \int_{-\infty}^{\infty} x dF(x)$, using the Riemann-Stieltjes integral.)

What is the connection between the mean of a data set and the expectation of a random variable? If the random variable were defined as a uniform random sample from the data set, they would be the same.

Alternatively, if we were to generate a large data set by considering repeated, independent realisations of the random variable, then the mean of this data set would be close to the expectation of the random variable. This statement is called the Law of Large Numbers.

Expectation of functions of a random variable: Let X be a random variable and g a function. Then, $Y = g(X)$ is another random variable (it is a function on the sample space defined by $Y(\omega) = (g \circ X)(\omega))$. To compute its expectation as above, we would first have to compute the distribution of the random variable Y. In fact, it turns out that there is an easier way:

$$E[Y] = \int_{-\infty}^{\infty} g(x)dF(x).$$

This should be intuitively obvious, at least in the discrete case.

Example: Let X be the outcome of the roll of a fair die, and define

$$Y = g(X) = \begin{cases} 1, & \text{if } X \text{ is even,} \\ 0, & \text{if } X \text{ is odd.} \end{cases}$$

Then, it is clear that $P(Y = 1) = 1/2$ and $P(Y = 0) = 1/2$, so $E[Y] = 0.5$. Alternatively, we have

$$E[Y] = \sum_{n=1}^{6} g(n)P(X = n) = (0 + 1 + 0 + 1 + 0 + 1)\frac{1}{6} = \frac{1}{2}.$$

Properties of the expectation:

1. Expectation is linear: For any two random variables X and Y defined on the same sample space, and any constants a and b, $E[aX + bY] = aE[X] + bE[Y]$. This is a very important property of expectations. Again, it is pretty easy to see in the discrete case. The result extends in the obvious way to sums of finitely many random variables.

2. The expectation of a constant is equal to that constant. (Think of a constant as a random variable which takes only one value, with probability 1).

Variance: Let X be a random variable and let us denote $E[X]$ by μ. The variance of X is defined as $\text{Var}(X) = E[(X - \mu)^2]$. Another way to write this is

$$\begin{aligned} \text{Var}(X) &= E[(X - \mu)^2] = E[X^2 - 2\mu X + \mu^2] \\ &= E[X^2] - 2\mu E[X] + \mu^2 = E[X^2] - (E[X])^2. \end{aligned}$$

Note that the variance has to be non-negative, because it is the expectation of a non-negative random variable. Thus, we have shown that $E[X^2] \geq (EX)^2$ for any random variable X. It is also clear from the definition that, for any real numbers a and b,

$$\text{Var}(aX + b) = a^2\text{Var}(X),$$

and that the variance of a constant is zero.

2 Joint and marginal distributions

Let us build up our intuition starting with discrete random variables. Roll two "independent" fair dice. (We haven't yet defined independence for random variables but just use your intuition.) Let X be the number shown on the first die, Y on the second and Z their sum. (Note that all three random variables are defined on the same sample space, $\{1,\ldots,6\}\times\{1,\ldots,6\}$.) The joint distribution of (X,Y) can be specified by writing down the probability of each of the 36 possible outcomes (i,j), $1\le i,j\le 6$. Likewise, the joint distribution of (X,Z) can be specified by specifying probabilities of 36 different events of the form (i,j), $1\le i\le 6$, $i+1\le j\le i+6$. In both cases, the function $(i,j)\mapsto p(i,j)$ is called the probability mass function.

Given the joint pmf for (X,Z), we can compute the "marginal" pmf for one of them, say Z. For example, suppose we are given the following joint probabilities:

$$\begin{aligned} p_{X,Z}(1,4) = p_{X,Z}(2,4) = p_{X,Z}(3,4) &= \frac{1}{36}, \\ p_{X,Z}(4,4) = p_{X,Z}(5,4) = p_{X,Z}(6,4) &= 0, \end{aligned}$$

where $p_{X,Z}(i,j)$ denotes $P(X=i,Z=j)$. Then we can compute

$$p_Z(4) := P(Z=4) = \sum_{i=1}^{6} p_{X,Z}(i,4) = \frac{3}{36}.$$

Definitions: Let $X_1,\ldots,X_n$ be random variables defined on the same sample space. Their joint distribution function (or joint cdf) is defined as

$$\begin{aligned} &F_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = P(X_1\le x_1,\ldots,X_n\le x_n) \\ &= P(\{\omega\in\Omega : (X_1(\omega),\ldots,X_n(\omega))\in(-\infty,x_1]\times\cdots\times(-\infty,x_n]\}). \end{aligned}$$

(What properties should F satisfy?) Note that the cdf is defined for both discrete and continuous random variables.

We say that $X_1,\ldots,X_n$ have joint density f if f is a non-negative function such that

$$F_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \int_{-\infty}^{x_1}\cdots\int_{-\infty}^{x_n} f(u_1,\ldots,u_n)du_1\cdots du_n.$$

Likewise, we can find the probability that $(X_1,\ldots,X_n)$ lie in a (measurable) subset of $\mathbb{R}^n$ by integrating the joint density over this set.

Covariance: Let X and Y be random variables on the same sample space. Their covariance is defined as

$$\mathrm{Cov}(X,Y) = E[(X - EX)(Y - EY)].$$

Since EX and EY are numbers, $(X - EX)(Y - EY)$ is a function of (X, Y), i.e., it is also a random variable. Its expectation can be computed as

$$E[(X - EX)(Y - EY)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x - EX)(y - EY)f(x,y)dxdy,$$

assuming that (X, Y) have a joint density f.

The covariance can also be expressed as follows:

$$\begin{aligned}\mathrm{Cov}(X,Y) &= E[XY - X\cdot(EY) - Y\cdot(EX) + (EX)\cdot(EY)]\\ &= E[XY] - E[X\cdot(EY)] - E[Y\cdot(EX)] + E[(EX)\cdot(EY)]\\ &= E[XY] - (EY)\cdot(EX)] - (EX)\cdot(EY)] + (EX)\cdot(EY)\\ &= E[XY] - (EX)(EY).\end{aligned}$$

The covariance of any two random variables has the following property:

$$(\mathrm{Cov}(X,Y))^2 \leq \mathrm{Var}(X)\mathrm{Var}(Y).$$

This will follow from the fact that, for any two random variables,

$$(E[XY])^2 \leq (E[X^2])(E[Y^2]), \tag{1}$$

for which we now give a proof. Note that

$$0 \leq E[(X + aY)^2] = E[X^2] + 2aE[XY] + a^2E[Y^2] \tag{2}$$

for all $a \in \mathbb{R}$. The minimum of the RHS is attained at $a = -E[XY]/E[Y^2]$. Substituting this for a above, we get

$$0 \leq E[X^2] - 2\frac{(E[XY])^2}{E[Y^2]} + \frac{(E[XY])^2}{E[Y^2]}.$$

Re-arranging this yields (1) provided $E[Y^2] \neq 0$. If $E[Y^2] = 0$ but $E[X^2] \neq 0$, the proof still works with X and Y interchanged. If $E[X^2]$ and $E[Y^2]$ are both zero, then (2) tells us that $2aE[XY] \geq 0$ for all $a \in \mathbb{R}$, which is only possible if $E[XY] = 0$.

Independence: We discussed earlier what it means for events to be independent. What does it mean for two or more random variables defined on the same sample space to be independent? Loosely speaking, random variables $X_1, \ldots, X_n$ are mutually independent if *any* events involving each of them individually are independent. More precisely, they are mutually independent if, for any measurable subsets $B_1, \ldots, B_n$ of $\mathbb{R}$, we have

$$P(X_1 \in B_1, \ldots, X_n \in B_n) = \prod_{i=1}^{n} P(X_i \in B_i).$$

This is not an operationally useful definition (at least for continuous random variables) because it is not feasible to check this equality for all measurable subsets!

We have the following alternative characterisation of independence. Random variables $X_1, \ldots, X_n$ are mutually independent if, and only if,

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n), \quad \forall x_1, \ldots, x_n \in \mathbb{R}.$$

Here, F_{X_i} the marginal distribution of X_i. (Given the joint distribution, how do you compute the marginal distribution?) Equivalently, if the random variables possess a joint density, then they are mutually independent if, and only if,

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n), \quad \forall x_1, \ldots, x_n \in \mathbb{R}.$$

Note that the existence of marginal densities is guaranteed if there is a joint density. (How do you compute the marginal densities from the joint?)

Examples:

1. Suppose U_1 and U_2 are uniform on $[0,1]$, and independent of each other. Let $X_1 = \min\{U_1, U_2\}$ and $X_2 = \max\{U_1, U_2\}$. Let us compute the joint distribution of (X_1, X_2), denoted F. First, it is clear that $F(x_1, x_2)$ is equal to zero if either x_1 or x_2 is negative, and equal to 1 if both are bigger than 1.

   Let us consider the case where both are between 0 and 1. (You might want to work out the other cases for yourself.) First, if $x_1 > x_2$, it is clear that $F(x_1, x_2) = F(x_2, x_2)$, so it suffices to consider $x_1 \le x_2$. In that case, we have

$$\begin{aligned}
&P(X_1 \le x_1, X_2 \le x_2) = P(U_1 \le x_1, U_2 \le x_1) \\
&\quad + P(U_1 \le x_1, x_1 < U_2 \le x_2) + P(U_2 \le x_1, x_1 < U_1 \le x_2) \\
&= x_1^2 + 2x_1(x_2 - x_1) = 2x_1x_2 - x_1^2).
\end{aligned}$$

From this, we can calculate the density. On the region $0 \leq x_1 \leq x_2 \leq 1$, the density is given by

$$f(x_1, x_2) = \frac{\partial^2 F(x, y)}{\partial x \partial y} = 2.$$

The density is zero outside this region.

2. Let us return to the example of a dart thrown at a circular dartboard of unit radius, and equally likely to fall anywhere on it. In this case, it is easy to see that the dart's position has density

$$f(x, y) = \frac{1}{\pi} 1(x^2 + y^2 \leq 1).$$

Are the co-ordinates X and Y independent? Can you compute the marginal densities of X and Y?

3. Multivariate normal distribution: The random variables $X_1, \ldots, X_n$ are said to be jointly normally distributed with mean vector $\mu = (\mu_1, \ldots, \mu_n)$ and covariance matrix C if C is a positive definite matrix, and they have the joint density function

$$f(\mathbf{x}) = \frac{1}{\sqrt{2\pi |\det(C)|}} \exp\Big(-\frac{1}{2}(\mathbf{x} - \mu)^T C^{-1} (\mathbf{x} - \mu)\Big),$$

where $\mathbf{x} = (x_1, \ldots, x_n)^T$.

3 Conditional distributions and conditional expectations

Let us go back to the example of rolling two fair dice. Let X and Y denote the number showing on the individual dice, and Z their sum. What do we mean by the conditional distribution of X given Z? Earlier, we defined conditional probability for events. What it means to specify the above is to specify the conditional probability of every event involving X given any event involving Z. How would this work in the above example?

Consider first the event $Z = 2$. Conditioning on this, we have

$$P(X = 1 | Z = 2) = \frac{P(X = 1, Z = 2)}{P(Z = 2)} = \frac{P(X = 1, Y = 1)}{P(X = 1, Y = 1)} = 1.$$

Also, it is clear that $P(X = j|Z = 2) = 0$ for all $j \in \{2, \ldots, 6\}$. Likewise, conditioning on $Z = 3$, we have $P(X = 1|Z = 3) = P(X = 2|Z = 3) = 1/2$ and $P(X = j|Z = 3) = 0$ for $j \notin \{1, 2\}$. Similarly, we can compute conditional probabilities conditioning on each of the "elementary" events $Z = k$. We shall use the notation $p_{X|Z}(\cdot|k)$ to denote the probability mass function of X conditional on the event $Z = k$. (Recall that a conditional probability is also a probability, hence this is a probability mass function.)

To complete the description, we also have to specify probabilities conditional on events of the form $Z \in \{1, 2\}$, $Z \in \{2, 3, 5\}$ etc. But this is not necessary, because we can compute all such conditional probabilities from the marginal distribution of Z, and the conditional pmf described above, using Bayes' formula. For example,

$$\begin{aligned} P(X = 1|Z \in \{2, 3\}) &= \frac{P(X = 1, Z \in \{2, 3\})}{P(Z \in \{2, 3\})} \\ &= \frac{P(X = 1, Z = 2) + P(X = 1, Z = 3)}{P(Z \in \{2, 3\})} \\ &= \frac{p_Z(2)p_{X|Z}(1|2) + p_Z(3)p_{X|Z}(1|3)}{p_Z(2) + p_Z(3)}. \end{aligned}$$

The last quantity above can be computed given the marginal pmf of Z and the conditional pmf of X conditioned on elementary events for Z.

It is not obvious how to extend this idea to continuous distributions because, if X and Z are continuous random variables, then $P(Z = z)$ will be zero for any z. Hence, we can't compute conditional probabilities conditional on this event. But let's do it heuristically anyway. Suppose (X, Z) have a joint density $f_{X,Z}$ and marginals f_X and f_Z. Thus, for an infinitesimal dz, the probability that Z is in $(z, z + dz)$ is $f(z)dz$. Now, conditional on this, what is the probability that X lies in $(x, x + dx)$. We can compute this using Bayes' formula:

$$\begin{gathered} P(X \in (x, x + dx)|Z \in (z, z + dz)) = \\ \frac{P((X, Z) \in (x, x + dx) \times (z, z + dz))}{P(Z \in (z, z + dz))} = \frac{f(x, z)dxdz}{f(z)dz}. \end{gathered}$$

This motivates us to define the conditional density as follows:

$$f_{X|Z}(x|z) = \frac{f_{X,Z}(x, z)}{f_Z(z)}.$$

Note that, for each fixed z, this defines a density function. We can use it to define the conditional cdf

$$F_{X|Z}(x|z) = \int_{-\infty}^{x} f_{X|Z}(u|z)du.$$

Having defined conditional distributions, we can define the property of conditional independence. We say that random variables X and Y are conditionally independent given Z if

$$F_{X,Y|Z}(x,y|z) = F_{X|Z}(x|z)F_{Y|Z}(y|z) \quad \text{for all } x, y, z.$$

Next, we turn to conditional expectations. Their definition is very similar to that of conditional distributions. As usual, we start with the discrete space. Let us go back to the example of the two dice, where X and Y denote the outcomes on the individual dice, and Z their sum. Say we are interested in the expectation of X conditional on Z. As in the case of conditional distributions, it suffices to specify $E[X|Z = z]$ for each possible value of z. Thus, for example,

$$E[X|Z = 3] = 1 \cdot P(X = 1|Z = 3) + 2 \cdot P(X = 2|Z = 3) = \frac{1}{2} + \frac{2}{2} = \frac{3}{2}.$$

In general, in the discrete case,

$$E[X|Z = z] = \sum xP(X = x|Z = z) = \frac{\sum xP(X = x, Z = z)}{P(Z = z)}.$$

This is a number, one for each possible value of Z. We can think of these numbers as describing a function of Z. In other words, $E[X|Z] = g(Z)$, where g is the function defined by $g(z) = E[X|Z = z]$. Thus, $E[X|Z]$ is itself a random variable.

The definition of conditional expectations in the continuous case is analogous. We shall only be interested in cases where the joint (and hence, conditional) densities exist. If that is so, we can define

$$E[X|Y = y] = \int_{-\infty}^{\infty} xf_{X|Y}(x|y)dx = \int_{-\infty}^{\infty} x\frac{f_{X,Y}(x,y)}{f_Y(y)}dx,$$

and think of $E[X|Y]$ as a function of Y, whose value at y is specified by the equation above.

Conditional expectation satisfies the same linearity property as expectation, i.e., $E[aX+bY|Z] = aE[X]+ bE[Y]$. The analogue of the second property, namely that the expectation of a constant is constant, is somewhat different. It is that the expectation of any function of Z, conditional on Z, behaves like a constant. In particular, it is conditionally independent of every random variable. In other words,

$$E[Xh(Z)|Z] = E[X|Z]E[h(Z)|Z] = h(Z)E[X|Z],$$

for any random variables X and Z, and any measurable function h.

Conditional expectation satisfies one more property, which doesn't have an analogue for expectations. Recall that $E[X|Z]$ is itself a random variable. What is the expected value of this random variable? It turns out that

$$E[E[X|Z]] = E[X].$$

This is easy to prove, at least in the discrete case. It can also be extended in the form of a chain rule, as follows. Observe that $E[X|Y,Z]$ is a function of (Y,Z) and itself a random variable. If we compute its conditional expectation given Z, then we get another random variable, which is a function of Z. And we have,

$$E[E[E[X|Y,Z]|Z]] = E[X].$$