FIT2086 Lecture 2 Summary, Part I
Expectations

Dr. Daniel F. Schmidt*

August 18, 2020

1 Introduction

In this Lecture we look at a very important set of quantities that can be derived from the probability distribution of a random variable. These are known as expected values of (functions of) the random variable; in an informal sense, the expected value describes the average, or typical value of many realisations of a random variable. This concept is fundamental in statistical modelling and data science, as it allows us to assign a "best guess" value to a set of random variables. In this section, we will review the formal definition of an expected value, and see how this can be extended to describe not only the average value, but the average degree of variation of a random variable through variance and covariance. We conclude this section with an examination of some important rules and special cases of expectations that we will use frequently throughout the remainder of this unit.

2 The expected value of a random variable

We will begin this section with some definitions that play a very important role in data science.

Definition 1. The *expectation*, expected value, or *mean* of a discrete random variable X is given by

$$\mathbb{E}\left[X\right] = \sum_{x \in \mathcal{X}} x p(x) dx,$$

and given a probability density function $p(x)$ over a set $\mathcal{X}$, the expectation of the continuous random variable is given by

$$\mathbb{E}\left[X\right] = \int_{\mathcal{X}} x p(x) dx. \quad (1)$$

Both of the above formulas can be interpreted as weighted averages over the set of possible values X can take over, with each value in $x \in X$ being weighted by the probability $p(x)$ of observing it. In this sense the expected value can be interpreted as the "average" or "typical" realisation of the random variable X. The more likely the value is to occur under the given probability distribution, the more weight is assigned to that value in the weighted average. An interesting, but sometimes initially confusing property of expectations is that while the expected value can be interpreted as the value of a "typical" realisation of the random variable, it does not necessarily have to lie in the set $\mathcal{X}$ of *possible* values that X can assume, i.e., for $\mathbb{E}[X] \notin \mathcal{X}$, in general, if $\mathcal{X} \subset \mathbb{Z}$. This is more likely to be the case when $\mathcal{X}$ is discrete.

2.1 Expectations of Functions of RVs

More generally we can find the expectation of a *function* $f(X)$ of a random variable using

$$\mathbb{E}[f(X)] = \sum_{x \in \mathcal{X}} f(x)p(x)dx.$$

for discrete RVs, and

$$\mathbb{E}[f(X)] = \int_{\mathcal{X}} f(x)p(x)dx$$

for continuous RVs. As with an expectation of a random variable, the expectation of a function of a random variable is a weighted average of the function over all the values X can take on, with each value being weighted according to the probability of observing it. In general, it is the case that

$$\mathbb{E}[f(X)] \neq f(\mathbb{E}[X]), \tag{2}$$

that is, the expectation of a function is not equal to the function of the expectation; the exception is when $f(X) = a\,X + b$ (i.e., $f(X)$ is a linear function of X – we will see why this is the case in Section 2.2). Interestingly, finding the expectation of a function of a random variable is no different from defining a new random variable, say $X_f = f(X)$, and finding the expectation of this new random variable. Let us introduce the notation $\mathbb{E}_X[f(X)]$, where the sub-script on the expectation operator denotes the RV with which the expectation (weighted sum/integral, as appropriate) is being taken. Then, we can simply write

$$\mathbb{E}_X[f(X)] = \mathbb{E}_{X_f}[X_f].$$

so that there is no conceptual difference between expectations of a functions of RVs, and expectations of RVs. We will use this result frequently throughout the course.

2.1.1 Expectations of Multiple RVs

More generally, it is frequently the case that we have multiple random variables, and we wish to find the expectation of some function of these random variables. The concept of the expectation is easily generalized to this case.

Definition 2. Let X and Y be two random variables, and $f(X, Y)$ be a joint function of these two variables. Then, the expected value of $f(X, Y)$ is defined by

$$\mathbb{E}_{X,Y}[f(X,Y)] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} f(x,y)p(x,y),$$

with the obvious modification for continous RVs.

Again, the subscripts on the expectation operator can be used to denote which random variables the expectation is being taken with respect to in the case that it may be ambigious. This can be the case when we are interested in *conditional* expectations; that is, we have some function $f(X, Y)$ of two random variables and we wish to find the expectation of this quantity with respect to one of the variables, say X, while holding Y at some fixed value (i.e., we are conditioning on Y). We can write this in two ways:

$$\mathbb{E}_X\left[f(X,Y)\right] \quad \text{and} \quad \mathbb{E}\left[f(X,Y)\,|\,Y\right]$$

where the former implicitly defines the conditional expectation, as the subscript includes only X, while the latter makes it explicitly clear we are holding Y fixed by using the bar-notation to denote conditioning.

Example 1: Expectations of a Simple RV

It is always useful to examine a simple example to understand concepts, and we shall do this now in the context of the expectation rules we have learned so far. For simplicity, let X be a discrete RV defined on $\mathcal{X} = \{1, 2, 3\}$ with probability distribution

$$\mathbb{P}(X = 1) = 0.5,\ \mathbb{P}(X = 2) = 0.4,\ \mathbb{P}(X = 3) = 0.1.$$

It is clear that realisations of X following this distribution will more frequently be “1”s and “2”s than they will be “3”s. Let us find the expected value of X:

$$\mathbb{E}\left[X\right] = 1 \cdot 0.5 + 2 \cdot 0.4 + 3 \cdot 0.1 = 1.6.$$

This simple example demonstrates two of the ideas we have previously discussed: (i) that the expected value is representative of the “typical” values of realisations of X, in the sense that it is much closer to 1 and 2 than it is to 3; (ii) that the expected value of a realisation of X is not necessarily in the set of possible realisations, i.e., $1.6 \notin \{1, 2, 3\}$, as we previously discussed. Let us now examine the expected value of a function of X. Let $f(X) = \log X$; then

$$\mathbb{E}\left[\log X\right] = 0.5\,\log 1 + 0.4\,\log 2 + 0.1\,\log 0.1 \approx 0.3871$$

where $\log x$ denotes the natural logarithm (base-e) of x. We see that $\log \mathbb{E}\left[X\right] = \log 1.6$, and $\log 1.6 \neq 0.3871$, as stated by (2). □

2.2 The Linearity of Expectations

Given the definition of an expectation as a weighted sum (integral) of a function, which is a *linear operator*, it follows that expectations are themselves linear operators. This linearity property is very important, as it implies a number of useful properties that can be exploited when manipulating expressions containing expectations. For example,

$$\mathbb{E}\left[b\,X + c\right] = b\,\mathbb{E}\left[X\right] + c$$

where b, c are any constants with respect to X. The above result tells us that expectations of linear transformations of X are simply equivalent to the linear transformation of the expected value of X, which establishes the claim discussed after (2) in Section 2.1. This property implies the following useful identity:

$$\mathbb{E}_X\left[XY\right] = Y\,\mathbb{E}_X\left[X\right]$$

where the X subscript reinforces the fact that the expectation is taken with respect to the RV X, and not Y. In this case, Y is a constant with respect to X and may be taken outside of the expectation. We also have the property that

$$\mathbb{E}_{X,Y}\left[X + Y\right] = \mathbb{E}_X\left[X\right] + \mathbb{E}_Y\left[Y\right]$$

so that the expectation of a sum of two functions of random variables is the sum of the individual expectations. Note that this always holds, regardless of whether X and Y are independent, or the form of their distribution. This is one of the most important rules of expectations and is used frequently throughout data science.

3 Measures of Spread

The expected value, or mean, $\mathbb{E}[X]$ of a random variable is a representation of the typical value of a realisation of the RV. The *variance* allows us to quantify just how typical a value the mean actually is.

Definition 3. The variance is the expected squared-deviation of the RV X around its mean:

$$\mathbb{V}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \tag{3}$$

which extends in a straightforward fashion to functions:

$$\mathbb{V}[f(X)] = \mathbb{E}\left[(f(X) - \mathbb{E}[f(X)])^2\right]$$

The variance measures the average squared deviation between realisations of a random variable X and its mean $\mathbb{E}[X]$. The variance is one example of a *measure of spread*. In contrast to the mean, which indicates the average, or typical value of the random variable X, a measure of spread such as the variance indicates to us how close the realisations are on average to this typical value. In some sense they indicate just how "typical" the mean really is – if the variance is small, then the realisations of X will be tightly clustered around $\mathbb{E}[X]$. In contrast, when the variance is large, the less the realisations cluster around the mean and the more variability there is the values taken on by X. The variance is also given by the following alternative formula.

Fact 1. Let X denote a random variable; then

$$\mathbb{V}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2, \tag{4}$$

that is, the variance of X is equal to the expected squared value of X, minus the square of the expected value of X.

Proof. It is straightforward to prove this relation using only the results we have covered so far:

$$\begin{aligned}
\mathbb{V}[X] &= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \\
&= \mathbb{E}\left[X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2\right] \\
&= \mathbb{E}\left[X^2\right] - 2\,\mathbb{E}[X]\,\mathbb{E}[X] + \mathbb{E}[X]^2
\end{aligned}$$

where the third step follows from the linearity of expectations. □

The linearity of expectations also yields another extremely useful identity involving the variance:

$$\mathbb{V}[c\,X] = c^2\,\mathbb{V}[X],$$

so that the variance of X times a constant c is equal to the variance of X times the square of c. An important quantity that is related to the variance is the standard deviation.

Definition 4. The standard deviation of a RV X is defined as

$$\begin{aligned} \mathrm{SD}\left[X\right] &= \sqrt{\mathbb{E}\left[\left(\mathbb{E}\left[X\right]-X\right)^2\right]}, \\ &= \sqrt{\mathbb{V}\left[X\right]} \end{aligned}$$

The standard deviation measures the average square-rooted squared deviation of a random variable X around its mean $\mathbb{E}\left[X\right]$. An important property of the standard deviation, in comparison to the variance, is that it has the same units of measurement as the RV X; that is, if X is measured in meters, then so is its standard deviation. This is in contrast to the variance, which is measured in the *square* of the units of measurement of X, i.e., if X is measured in meters, then the variance is in meters-squared. The variance is frequently easier to work with mathematically (when manipulating formulas, or proving results), but the standard deviation is a much more interpretable measure of spread and is usually presented in analyses of real data.

Example 2: Variance of a Discrete RV

Let us find the variance of the simple discrete RV discussed in Example 1. We can do this in two ways: the first is to note we have already found $\mathbb{E}\left[X\right] = 1.6$ and apply formula (3) directly, yielding

$$\mathbb{V}\left[X\right] = (1-1.6)^2 \cdot 0.5 + (2-1.6)^2 \cdot 0.4 + (3-1.6)^2 \cdot 0.1 = 0.44.$$

The second approach is to use the alternative formula (4) yielding:

$$\begin{aligned} \mathbb{V}\left[X\right] &= \mathbb{E}\left[X^2\right] - 1.6^2, \\ &= 1 \cdot 0.5 + 4 \cdot 0.4 + 9 \cdot 0.1 - 2.56, \\ &= 0.44, \end{aligned}$$

which we note gives (unsurpisingly) the same result. □

3.1 Covariance and Correlation

We are now in the position to define a pair of extremely important measures that capture the degree of variation between two random variables: the covariance and the correlation.

Definition 5. If we have two random variables X and Y we can define the covariance between them by:

$$\begin{aligned} \mathrm{cov}\left(X, Y\right) &= \mathbb{E}\left[(X-\mathbb{E}\left[X\right])(Y-\mathbb{E}\left[Y\right])\right], \\ &= \mathbb{E}\left[XY\right] - \mathbb{E}\left[X\right]\mathbb{E}\left[Y\right]. \end{aligned}$$

The covariance measures the average value of the product of the deviations $(X - \mathbb{E}\left[X\right])$ and $(Y - \mathbb{E}\left[Y\right])$. The covariance always satisfies

$$0 \leq \left|\mathrm{cov}\left(X, Y\right)\right| \leq \max\left\{\left|\mathbb{V}\left[X\right]\right|, \left|\mathbb{V}\left[Y\right]\right|\right\}.$$

The covariance measures similarity of the behaviour of the random variables X and Y, in the sense that a large covariance indicates that when X deviates positively (negatively) from $\mathbb{E}[X]$, then Y usually deviates positively (negatively) from $\mathbb{E}[Y]$. From the covariance, we can define the correlation between X and Y.

Definition 6. If we have two random variables X and Y, we can define the correlation between them by:

$$\operatorname{corr}(X, Y) = \frac{\operatorname{cov}(X, Y)}{\mathrm{SD}[X]\,\mathrm{SD}[Y]}.$$

The covariance depends on the units of X and Y. In contrast, the division by the product of standard deviations in the definition of correlation serves to *remove the unit of measurement* from the coefficient correlation. This standardises the measure so that the value always lies between -1 (perfect negative correlation) and 1 (perfect positive correlation), with a correlation of zero denoting no correlation. Such a standardisation helps dramatically in interpreting the correlation as a measure of assocation between X and Y, because it is free of the arbitrary choice of scale of measurement of both random variables. Correlation/covariance can be interpreted as follows:

- A positive correlation implies that if a X is greater than its expected value $\mathbb{E}[X]$, then the corresponding value Y will be more likely to be greater than its expected value $\mathbb{E}[Y]$.
- Conversely, a negative correlation implies that if a X is greater than its expected value $\mathbb{E}[X]$, then the corresponding value of Y will be more likely to be smaller than its expected value $\mathbb{E}[Y]$.

Correlation measures the *linear association* between the variables X and Y. A correlation of 1 is possible if and only if $X = aY + b$, i.e., that the random variable X is exactly equal to a linear transformation of Y. In contrast, if two RVs X and Y are independent, then $\operatorname{cov}(X, Y) = 0$ and the two random variables are uncorrelated. However, it is very important to understand that a covariance/correlation of zero does not imply that the opposite holds. As correlation measures the linear association between the two variables, it is very possible that two variables will be uncorrelated (a correlation of zero, or a small correlation) but strongly associated through a non-linear relationship.

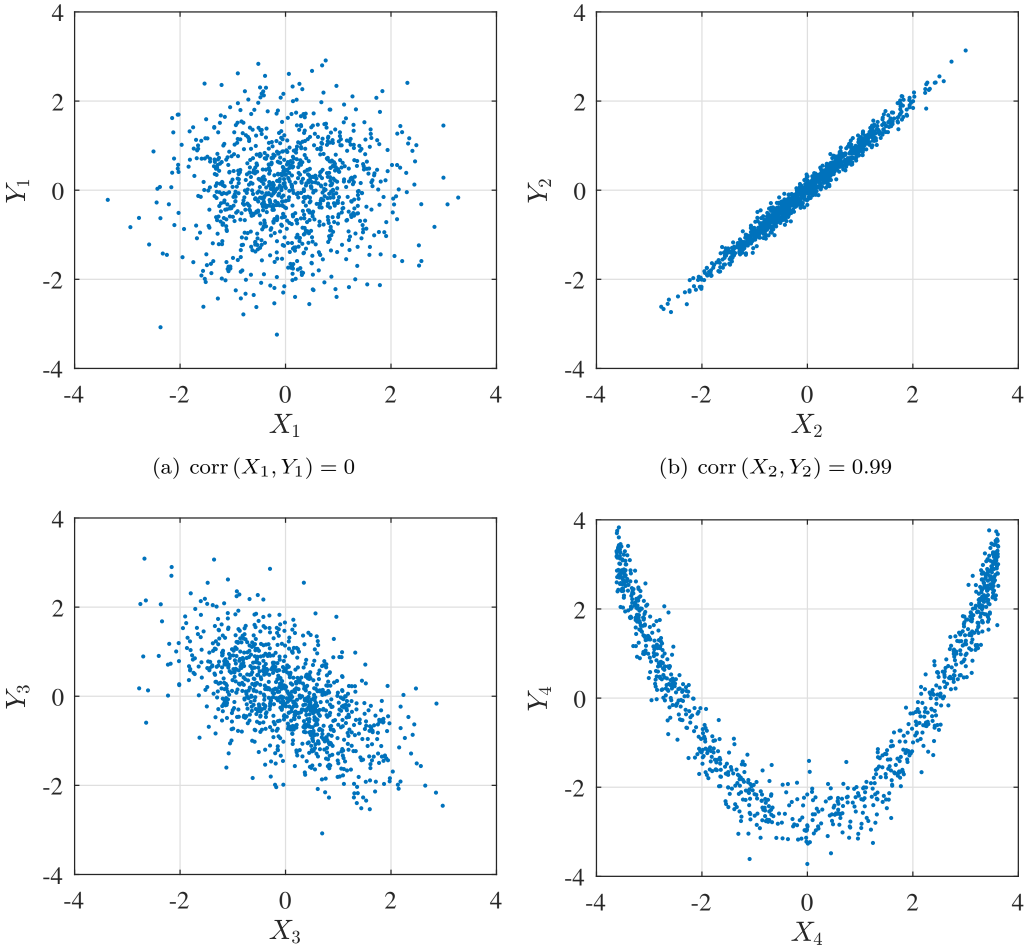
Example 3: Correlation of two RVs

To gain a some intuition into how to interpret different levels of correlation between random variables, Figure 1 displays scatter plots of realisations of four different pairs of random variables. Figure 1(a) shows the realisations of two variables that are virtually uncorrelated – there is no clear relationship between the values of X_1 and the values of Y_1. In contrast, Figure 1(b) shows two variables that are very highly positively correlated. The strong linear relationship between X_2 and Y_2 is very clear from the scatter plot; as X_2 increases, so does Y_2. Figure 1(c) shows two variables which are moderately negatively correlated; it is clear that as X_3 increases, the values of Y_3 tend to decrease.

Finally, Figure 1(d) shows two variables which are virtually uncorrelated; however, it is obvious that there exists a strong relationship between the variables. The problem is that the relationship is very nonlinear, and the correlation coefficient, which expresses linear association, will fail to pick up such an association. This example also demonstrates clearly how scatter plots are useful tools that should be used in conjunction with correlation coefficients to detect possible non-linear associations between realisations of random variables. □

3.2 Expectations and Independent RVs

As we previously discussed, independent random variables (see Lecture 1) play a particular important role in data science because they greatly simplify the mathematics underlying our models. Independent



(a) $\text{corr}\,(X_1, Y_1) = 0$

(b) $\text{corr}\,(X_2, Y_2) = 0.99$

(c) $\text{corr}\,(X_3, Y_3) = -0.6$

(d) $\text{corr}\,(X_4, Y_4) = 0$

Figure 1: Scatter plots of various random variables demonstrating different levels of correlation.

RVs also lead to a number of simplified results regarding expectations. For example, if the RVs X and Y are independent, i.e.

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y),$$

then the expectation of the product of (functions of) the variables satisfies

$$\mathbb{E}_{X,Y}\left[XY\right] = \mathbb{E}_X\left[X\right]\mathbb{E}_Y\left[Y\right],$$

so that the expected value of the product (of functions) of X and Y is equal to the product of the expected values (of the functions). An obvious corollary of this result that is incredibly useful is that if either $\mathbb{E}_X\left[X\right] = 0$ or $\mathbb{E}_Y\left[Y\right] = 0$, then the expectation of the product will simply be equal to zero. Independence of RVs also implies the following important property regarding the variance.

Fact 2. If X and Y are independent RVs, then the following identity holds

$$\mathbb{V}\left[X+Y\right]=\mathbb{V}\left[X\right]+\mathbb{V}\left[Y\right]$$

In words, Proposition 2 tells us that the variance of the sum of (functions of) the RVs X and Y is equal to the sum of the variances. This is obviously a *very useful* property; the proof of this result is left as an exercise for the reader (it can be completed using only the rules we have learned so far). If X and Y are independent, then we can also find the the variance of the difference of two random variables $\mathbb{V}\left[X-Y\right]$.

Fact 3. Let X and Y be independent random variables; then

$$\mathbb{V}\left[X-Y\right]=\mathbb{V}\left[X\right]+\mathbb{V}\left[Y\right]$$

that is, the variance of the difference $X-Y$ is equal to the sum of the variances of X and Y.

Proof. Noting that $a-b=a+(-1)b$ we can write:

$$\begin{aligned}\mathbb{V}\left[X-Y\right] &= \mathbb{V}\left[X+(-1)Y\right]\\ &= \mathbb{V}\left[X\right]+\mathbb{V}\left[(-1)Y\right]\\ &= \mathbb{V}\left[X\right]+(-1)^2\mathbb{V}\left[X\right]\end{aligned}$$

where the third step follows the fact $\mathbb{V}\left[cX\right]=c^2\,\mathbb{V}\left[X\right]$. □

This result tells us that the variance of the difference of two RVs is equal to the *sum* of the variances of the RVs. Initially, this might seem initially unintuitive, but it makes sense if you think of the variance as quantifying the amount of variability in a RV. Whether you are adding or subtracting two RVs, you are combining two variable quantities, and therefore increasing the overall variability above the variability that is present in either of the individual RVs.

4 Approximate Expectations of Functions of RVs

We are frequently interested in computing the expectations and variances of functions of RVs. We know that in general, if X is a RV and $f(x)$ a function then

$$\mathbb{E}\left[f(X)\right]\neq f(\mathbb{E}\left[X\right])$$

and there is no silver bullet for computing these expectations. However, if we make some assumptions about both X and $f(x)$ then we can make some statements about value of the expectation. The approach we consider is to *approximate* the function $f(X)$ by a simpler function through the Taylor series approach, and then compute expectations for this simpler function.

Fact 4. Let $f(x)$ be a twice differentiable function in x, and let X be a random variable satisfying $\mathbb{E}[X] = \mu < \infty$ and $\mathbb{V}[X] = \sigma^2 < \infty$; then we have

$$\begin{aligned} \mathbb{E}[f(X)] &\approx f(\mu) + \left[\left. \frac{d^2 f(x)}{dx^2} \right|_{x=\mu} \right] \frac{\sigma^2}{2} \\ \mathbb{V}[f(X)] &\approx \left[\left. \frac{df(x)}{dx} \right|_{x=\mu} \right]^2 \sigma^2 \end{aligned}$$

While only being approximate, Fact 4 provides us with a way to obtain rich information about the behaviour of $\mathbb{E}[f(X)]$ and $\mathbb{V}[f(X)]$ as μ_X and σ_X^2 vary. Further, it is useful to note that while exact bounds on how close the approximate values are to the exact values is not immediately available, it is a fact that the approximation becomes better as $\sigma_X^2 \to 0$, i.e., as the variance of X gets smaller and smaller

Example 4: Approximating Expectation/Variance

As an example, consider the transformation $f(X) = aX^2 + c$. Then, we need to find

$$\frac{df(x)}{dx} = 2ax, \quad \frac{d^2 f(x)}{dx^2} = 2a$$

which yields

$$\begin{aligned} \mathbb{E}[f(X)] &\approx a\mu^2 + c + a\sigma^2 \\ \mathbb{V}[f(X)] &\approx 4(a\mu)^2\sigma^2 \end{aligned}$$

Examining these results we see an interesting outcome; if we take a RV with variance $\mathbb{V}[X] = \sigma^2$ that is independent of the mean μ, then the new random variable $Z = X^2$ has a variance that is linked to the square of the mean of X. As an aside, we note that the above approximation for $\mathbb{E}[f(X)]$ is actually *exact*. To see this, recall the alternative formula for the variance $\mathbb{V}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$; from this we can write

$$\begin{aligned} \mathbb{E}\left[X^2\right] &= \mathbb{E}[X]^2 + \mathbb{V}[X] \\ &= \mu^2 + \sigma^2 \end{aligned}$$

and noting that $\mathbb{E}[f(X)] = a\,\mathbb{E}\left[X^2\right] + c$ by linearity we have

$$\mathbb{E}\left[a\,X^2 + c\right] = a\mu^2 + a\sigma^2 + c$$

which matches our approximation from above. This is because the appoximation is based on a second-order Taylor series expansion, which happens to be an exact representation of $f(X)$ when $f(X)$ is a quadratic. For a general $f(X)$, this will *not* be the case! We also note that that variance approximation is not exact because the Taylor series expansion is no longer precise enough when dealing with variances. □