FIT2086 Lecture 1 Summary, Part II
Introduction/Probability and Random Variables

Dr. Daniel F. Schmidt*

July 22, 2021

1 Introduction

Random variables, and the calculus of probability, play a central role in the science of data analysis. They are the language we use to describe uncertainty and the fact that we do not expect the outcomes of our experiments to take on values in a fixed, deterministic fashion. The primary use of random variables in analytic data science is in modelling the values of the observations that form our sample of data, and it is the calculus of probability that is the underlying tool that we use to define our random variables, and describe their behaviour. There are, in general, two ways in which randomness can manifest itself in our data:

1. **Measurement/Experiment Error**. In this case, our measurements themselves are random quantities that are not determined until we measure them. An example of such a measurement is the outcome of rolling a six-sided dice; each time we repeat the experiment, we produce an integer from one through to six. This form of randomness can be used to explain variation in observations due to imprecise measurement devices. For example, if we use a volt-meter to record the voltage of circuit, then the recorded voltages will vary from measurement to measurement based on the vagaries of the environment and inherent randomness in the both the apparatus and the physical process. This type of randomness is often called "experimental error" or "measurement error".

2. **Random Sampling from the Population**. What if our data measurements were instead the recorded heights of people in a population? We do not expect a person's height to vary substantially from measurement to measurement – so why should we be modelling these observations as random? The answer is that modelling our observations as random variables does not necessarily imply that we believe the data to be truly random in a generative sense. In this context, we are instead modelling the fact that from our finite, but very large population, we can only draw a sample consisting of a (relatively) small number individuals. The randomness here is not the measurement of the person's height *per se*, but rather the fact that the particular sample of people we draw is random, so that from sample to sample we will end up with different collections of recorded heights.

*Copyright (C) Daniel F. Schmidt, 2020

2 Random Variables

We will begin by first examining the concept of random variables and their associated probability distributions. This material is to a large extent, intended as a refresher, and as such is neither extensive nor particularly deep in its coverage. We examine both discrete and continuous random variables.

2.1 Sets and Subsets

Before we can begin our examination of random variables we will need to introduce some notation that we will use throughout this entire course. The study of random variables is the study of quantities that take on values from some specific set of possible events or outcomes. Therefore, we need a brief refresher on sets. In particular:

- to define a set we frequently use the notation $\mathcal{X} = \{a, b, c\}$, which defines a set $\mathcal{X}$ which contains the elements a, b and c.
- Given a set $\mathcal{X}$, we use $x \in \mathcal{X}$ to denote that the variable x takes on a value from the set $\mathcal{X}$.
- We use the notation $\mathcal{A} \subseteq \mathcal{X}$ to denote that the set $\mathcal{A}$ is subset of $\mathcal{X}$, i.e., that all elements in $\mathcal{A}$ appear in $\mathcal{X}$.
- The set union $\mathcal{A} = \mathcal{X} \cup \mathcal{Y}$ creates a new set $\mathcal{A}$ that contains all of the elements in both $\mathcal{X}$ and $\mathcal{Y}$
- The set intersection $\mathcal{A} = \mathcal{X} \cap \mathcal{Y}$ creates a new set $\mathcal{A}$ that contains only those elements that appear in both $\mathcal{X}$ and $\mathcal{Y}$
- The notation $\mathcal{A} = (a, b)$ denotes the set of numbers from a to b, exclusive, with $[a, b]$ denote an inclusive set.

Additionally, there several sets that we will use throughout this subject:

- $\mathbb{Z}$ is the set of all integers, and $\mathbb{Z}_+$ is the set of non-negative integers;
- $\mathbb{R}$ is the set of all real numbers, and $\mathbb{R}_+$ is the set of non-negative real numbers;
- $\emptyset$ denotes the empty set (i.e., a set containing no elements).

Armed with these definitions we can begin a revision of random variables and probability.

2.2 Probability distribution over a random variable

We begin by introducing the concept of a *random variable*. We say X is a *random variable (RV)* if it takes on values from a set of possible values $\mathcal{X}$ with specified *probabilities*. This is in contrast to a non-random variable, which takes on a value in a fixed, deterministic manner. We sometimes call $\mathcal{X}$ the *event space*, and seeing any x from $\mathcal{X}$ as observing the *event* $X = x$. We use the notation

$$\mathbb{P}(X = x),\ x \in \mathcal{X}$$

to describe the probability that the RV X will take on the value x from $\mathcal{X}$. The mapping from events X to probabilities is called a probability distribution.

Definition 1. A probability distribution is any mapping from the event space $\mathcal{X}$ to $\mathbb{R}$ that satisfies the following two properties:

$$\mathbb{P}(X = x) \in [0, 1] \text{ for all } x \in \mathcal{X},$$

and

$$\sum_x \mathbb{P}(X = x) = 1.$$

The first property tells us that the probability of any event must lie strictly between zero and one, inclusive. The second property tells us that the total probability of all the possible values $\mathcal{X}$ of x is always equal to one, i.e., that $\mathbb{P}(X \in \mathcal{X}) = 1$. We will use these properties frequently throughout this course. Given a probability distribution over a set of values $\mathcal{X}$, we can find the probability that X takes a value from a subset $\mathcal{A} \subset \mathcal{X}$ by adding up the probabilities assigned to all events in $\mathcal{A}$, i.e.,

$$\mathbb{P}(X \in \mathcal{A}) = \sum_{x \in \mathcal{A}} \mathbb{P}(X = x).$$

Another very important property of random variables that extends the above rule is the additivity of the probability of mutually exclusive sets of events. What this means is that the probability of (X taking on values from some set $\mathcal{A}$) or (X taking on values from another set $\mathcal{B}$) is equal to

$$\mathbb{P}(X \in \mathcal{A} \text{ or } X \in \mathcal{B}) = \mathbb{P}(X \in \mathcal{A}) + \mathbb{P}(X \in \mathcal{B}) \tag{1}$$

if $\mathcal{A}$ and $\mathcal{B}$ have no values in common, i.e., if $\mathcal{A} \cap \mathcal{B} = \emptyset$. More generally, the probability of X taking on a value in either the set $\mathcal{A}$ or set $\mathcal{B}$, where $\mathcal{A}$ and $\mathcal{B}$ may *overlap* (contain elements in common) is

$$\mathbb{P}(X \in \mathcal{A} \text{ or } X \in \mathcal{B}) = \mathbb{P}(X \in \mathcal{A}) + \mathbb{P}(X \in \mathcal{B}) - \mathbb{P}(X \in (\mathcal{A} \cap \mathcal{B})).$$

The term that we subtract accounts for the probability of events that were "counted twice" when adding the first two terms together; clearly in the specific case that $\mathcal{A}$ and $\mathcal{B}$ have no elements in common, the above formula reduces to (1).

Example 1: Simple Probability Distribution

As an example, let X be a random variable over the event space $\mathcal{X} = \{1, 2, 3\}$, with probability distribution

$$X = \begin{cases} 1 & \text{with probability } 1/2 \\ 2 & \text{with probability } 1/4 \\ 3 & \text{with probability } 1/4 \end{cases},$$

From this definition, we see that X is twice as likely to take on the value $X = 1$ than either $X = 2$ or $X = 3$. What does this mean in real terms? Our definition of probability is based on *frequencies*; that is, how often particular events occur in long runs of realisations of a random variable. A *realisation* of a random variable X is a particular value drawn at random from $\mathcal{X}$, with probabilities determined by the associated probability distribution $\mathbb{P}(X = x)$. For example, for our random variable defined above, a sample of twenty-two realisations might be

$$\mathbf{x} = (3, 3, 1, 3, 2, 1, 1, 1, 2, 3, 3, 2, 1, 3, 3, 2, 1, 2, 1, 2, 1, 1).$$

In this sequence, we can count nine "1"s as compared to six "2"s and seven "3"s. This is not unexpected, as we would expect "1"s to appear more frequently than either "2"s or "3"s in realisations

from such a distribution, and for very long realisations, we would expect to about twice as many "1"s as "2"s or "3"s.

The additivity rules from above allow us to compute the probability of more complex events occuring; for example, imagine we want to find $\mathbb{P}(X \geq 2)$. Using the additivity rule we could write

$$\begin{aligned} \mathbb{P}(X \in \{2\} \cup \{3\}) &= \mathbb{P}(X = 2) + \mathbb{P}(X = 3) \\ &= 1/4 + 1/4 \\ &= 1/2. \end{aligned}$$

Alternatively, for such a simple problem we could also obtain the same probability using the fact that probabilities must sum to one; i.e., $\mathbb{P}(X \geq 2) = 1 - \mathbb{P}(X = 1) = 1/2$. □

2.3 Probability distributions over multiple variables

If we are going to use random variables to represent, or model, the observations in our data sample, then we are clearly going to need a way of manipulating or dealing with more than a single random variable, as we clearly expect to obtain more than a single measurement or data point from our experiments. The calculus of probability is easily extended to the setting of more than one random variable. Let X and Y be random variables over some sets $\mathcal{X}$ and $\mathcal{Y}$. We can then define

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

as the *joint probability* of $X = x$ and $Y = y$; that is, the probability of both X taking on the specific value x and Y taking on the specific value y *at the same time*. The joint probability distribution completely describes the behaviour of the two random variables. From this joint distribution we can then derive two important quantities that we will use repeatedly throughout this course:

Definition 2. The marginal probability of $X = x$ is given by

$$\mathbb{P}(X = x) = \sum_{y \in \mathcal{Y}} \mathbb{P}(X = x, Y = y). \tag{2}$$

The marginal probability describes the probability of seeing $X = x$, regardless of which value the random variable Y takes.

Definition 3. The conditional probability is given by

$$\mathbb{P}(X = x \,|\, Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)}. \tag{3}$$

This describes the probability of observing $X = x$ if we know that the random variable takes on the value $Y = y$.

The conditional probability in particular sees substantial use in the later parts of this course when we look to build models that make predictions about one (unobserved) random variable, given that we have observed the values of number of other explanatory variables, which we call predictors.

Example 2: Simple Distribution over Two Random Variables

	$X=1$	$X=2$	$X=3$
$Y=1$	0.05	0.15	0.1
$Y=2$	0.25	0.15	0.3

Table 1: Example joint probability distribution of two discrete random variables.

As an example, we now consider a simple probability distribution over two random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$. The sets $\mathcal{X}$ and $\mathcal{Y}$ are the sets of values X and Y can take on, respectively, and the set $\mathcal{X} \times \mathcal{Y}$ is the set of values the pair can *jointly* assume. As an example, let $\mathcal{X} = \{1, 2, 3\}$ and $\mathcal{Y} = \{1, 2\}$; then the joint event space of (X, Y) is

$$\mathcal{X} \times \mathcal{Y} = \{\{1,1\}, \{2,1\}, \{3,1\}, \{1,2\}, \{2,2\}, \{3,2\}\}$$

which contains $|\mathcal{X}| \times |\mathcal{Y}| = 3 \times 2 = 6$ distinct elements. Now, consider the example joint probability distribution over $\mathcal{X} \times \mathcal{Y}$ described in Table 1.

Using the information in this table we can obtain the joint probability of X and Y for all possible values of $\mathcal{X} \times \mathcal{Y}$; for example, we see that $\mathbb{P}(X = 1, Y = 1) = 0.05$, $\mathbb{P}(X = 2, Y = 1) = 0.15$, and so on. This joint probability distribution completely specifies the frequency with which specific (x, y) pairs occur in realisations of these two random variables. From this information we can therefore also compute the marginal probabilities of various events; for example, using (2) we can find that

$$\begin{aligned} \mathbb{P}(Y = 1) &= 0.05 + 0.15 + 0.1 = 0.3, \\ \mathbb{P}(Y = 2) &= 0.25 + 0.15 + 0.3 = 0.7, \end{aligned}$$

which tells us that the probability of seeing $Y = 2$ is significantly higher (over twice) than the probability of seeing $Y = 1$, irrespective of the particular value that X assumes. We can also use the joint probability distribution to obtain the conditional probability; for example, using (3) we can find that

$$\begin{aligned} \mathbb{P}(X = 1 \,|\, Y = 1) &= \mathbb{P}(X = 1, Y = 1)/\mathbb{P}(Y = 1) \\ &= 0.05/0.3 \approx 0.1667 \end{aligned}$$

and

$$\begin{aligned} \mathbb{P}(X = 1 \,|\, Y = 2) &= \mathbb{P}(X = 1, Y = 2)/\mathbb{P}(Y = 2) \\ &= 0.25/0.7 \approx 0.3571. \end{aligned}$$

These results tell us that observing $X = 1$ is around twice as likely when $Y = 2$ as compared to the case that $Y = 1$. So knowing the value of Y can dramatically improve our information about the likely values of X. This captures the idea underlying a lot of data science models: if we imagine that Y is some attribute of an individual (say, whether they smoke or not) and X represents whether they will contract cancer, we can see that we can use the measured value of Y to build a simple model to predict the value of X for that individual. □

2.4 Independent random variables

The concept of independent random variables play a very important role in applied probability and data science, primarily because they greatly simplify calculations.

Definition 4. Two RVs X and Y are considered independent if and only if

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

for all values of x and y.

In words this says that if X and Y are independent, the joint probability of X taking on the value x and Y taking on the value y is equal to the product of the marginal probabilities of X and Y. An important implication of independence is that

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

The establishment of an "independent co-distribution" equation means that the group of random variables does not interact with each other, and that they come from the same probability distribution, providing a basis for subsequent probabilistic inference and statistical analysis

In words this says that knowing the value of Y tells us no useful information about what value the random variable X may take on. A particularly important sub-class of independent RVs are independent and identically distributed RVs.

Definition 5. Two RVs X_1 and X_2 are independent and identically distributed (iid) if they are (i) independent and (ii) if their probability distributions satisfy

$$\mathbb{P}(X_1 = x) = \mathbb{P}(X_2 = x)$$

for all values of $x \in \mathcal{X}$.

In words, if two RVs are iid, then the two RVs are both independent and possess exactly the same marginal distribution over the event space $\mathcal{X}$.

Independent random variables play a prominent role in data science, and modelling, because we usually assume the observations or measurements in our data sample to be independent. This makes particular sense if we treat our observations as being made on individuals sampled at random from our population; in general, we do not expect the height of the fourth person we sampled to provide additional information about the height of the fifth person we sampled. While one can obviously imagine situations in which exact independence may not hold, it is usually sufficient to assume this is the case; the advantages of this assumption are great simplifications in way in which we construct models of our population from our random data sample.

Applying these ideas to our example from Section 2.3, we see that the random variables X and Y in our example are clearly not independent as

$$\mathbb{P}(X = 1 \,|\, Y = 1) \neq \mathbb{P}(X = 1 \,|\, Y = 2)$$

i.e., knowing the value of Y changes the probability of X taking on the value "1".

3 Probability density and mass functions

3.1 Probability mass functions

So far we have confined our attention to discrete random variables that can take on a finite set of different values. We now introduce a device that will prove useful throughout the remainder of this book as we study probability distributions over infinite sets of discrete values: the probability mass function.

Definition 6. A probability mass function (PMF) is any function from a discrete event space $\mathcal{X} \subseteq \mathbb{Z}$ to $\mathbb{R}$ that satisfies the following two properties:

$$p(x) \geq 0 \text{ for all } x \in \mathcal{X},$$

and

$$\sum_{\mathcal{X}} p(x) = 1.$$

Given a PMF associated with a random variable, the probability assigned to an event $A \subset \mathcal{X}$ is

$$\mathbb{P}(X \in A) = \sum_{x \in A} p(x).$$

You can think about a probability mass function as being a function that describes in a formal sense how much probability is assigned to each event. While on the surface this might seem just another way of expressing what we have already been doing by directly specifying probabilities of events, it has the subtle advantage that it allows us to specify probabilities over infinite event spaces if a suitable functional form exists, which is clearly not possible by the simple tabulation of events and probabilities that we have seen so far. As an example, the probability mass function

$$p(x) = \frac{6}{\pi^2(1+x)^2}, \quad \mathcal{X} = \{0, 1, 2, \ldots\}, \tag{4}$$

assigns a non-zero probability to every non-negative integer. The fact that there is an infinite number of integers is not a problem because rather than using simple tabulation the probabilites are encoded into a function of x. It is (relatively) straightforward to verify that (4) satisfies the conditions of a PMF specified by Definition 6 (i.e., it sums to one and is non-negative for all values in $\mathcal{X}$).

The technique of using functions to map events to probabilities is central to data science. It allows us to build models of large, complex event spaces without needing to somehow tabulate large numbers of probabilities, which is potentially impossible. This idea is examined in greater detail in Lecture 2 where we are introduced to some very special functional forms for probability distributions. A further advantage of formally associated a function with the probabilities of events is that it provides us with a single unified language with which we can describe the probabilities of both discrete event spaces and continuous event spaces. This latter problem is handled by the extension of probability mass functions to probability density functions.

3.2 Probability density functions

We have confined our attention to discrete random variables; that is, random variables for which the event space $\mathcal{X}$ is (a subset of) the set of integers, i.e., $\mathcal{X} \subseteq \mathbb{Z}$. If the set of values $\mathcal{X}$ that a random variable X can take is instead a subset of the real numbers, then we say that X is a continous random variable and is described by a probability density function.

Definition 7. A probability density function (PDF) is any function from a continuous event space $\mathcal{X}$ to $\mathbb{R}$ that satisfies the following two properties:

$$p(x) \geq 0 \text{ for all } x \in \mathcal{X},$$

and

$$\int_{\mathcal{X}} p(x)dx = 1.$$

Note that in contrast to the case of discrete RVs, we now need to integrate our probability density function over the set $\mathcal{X}$ rather than summing. The term "density function" is used because a pdf describes how densely the probability is distributed across the set $\mathcal{X}$. To find the probability that a continuous RV X takes on a value in an interval (a, b) we simply integrate the pdf from a to b, i.e.,

$$\mathbb{P}(a < X < b) = \int_a^b p(x)dx. \tag{5}$$

More generally, we can find the probability of X taking on any value in a set $\mathcal{A} \subseteq \mathcal{X}$ by integrating the pdf over the set $\mathcal{A}$

$$\mathbb{P}(X \in \mathcal{A}) = \int_{\mathcal{A}} p(x)dx$$

with similar extensions to finding the probabilities of unions of sets to those discussed in Section 2.2 in regards to discrete RVs.

One of the more confusing properties of continuous random variables is that the probability of X taking on any specific, exact real number is zero. Why is this so? To see why this is the case, consider the probability that $X \in (x_0 - \delta, x_0 + \delta)$ with δ a "small number", i.e., the probability that X takes on real number in a small interval centered on the value x_0. From (5) we have

$$\begin{aligned}\mathbb{P}(x_0 - \delta < X < x_0 + \delta) &= \int_{x_0-\delta}^{x_0+\delta} p(x)dx \\ &= \left[\int p(x)dx\right]_{x=x_0+\delta} - \left[\int p(x)dx\right]_{x=x_0-\delta}\end{aligned}$$

where $\int p(x)dx$ denotes the indefinite integral of $p(x)$. As $\delta \to 0$, the interval $(x_0 - \delta, x_0 + \delta)$ shrinks and becomes closer and closer to containing only the single real value x_0. However, from the above equation we see that as $\delta \to 0$, the probability of the interval decreases and becomes equal to zero when $\delta = 0$ (i.e., when the interval contains only the value x_0).

3.3 The cumulative distribution function

The cumulative distribution function (CDF) plays a very important role in probability theory. To simplify our discussion, we first introduce some shorthand notation for discrete RVs; namely, we can use

$$\mathbb{P}(X = x) \equiv p(x)$$

where $a \equiv b$ should be read as "a is equivalent to b". Using this shorthand notation, we can define the CDF of discrete and continuous random variables.

Definition 8. If X is a discrete RV over (a subset of) the integers, then its cumulative distribution function (CDF) is given by

$$\mathbb{P}(X \leq x) = \sum_{x' \leq x} p(x'), \quad x \in \mathcal{X}.$$

If X is a continuous RV over (a subset of) the real numbers, then its CDF is given by

$$\mathbb{P}(X \leq x) = \int_{-\infty}^{x} p(x')dx'.$$

The CDF returns the probability that X takes on any value less than or equal to some value x. Note that for the CDF to make sense there is the implicit requirement for the notion of an "order" amongst the values of the set $\mathcal{X}$. The CDF is important because, in a similar fashion to the PDF, it also offers a complete description of the probabilistic behaviour of the random variable X. In fact, if X is a continuous RV, we can see that

$$\frac{d}{dx}\{\mathbb{P}(X \leq x)\} = p(x),$$

so that we can derive the PDF from the CDF, and vice versa. By making use of the fact that the total probability of any distribution is one we can derive the following useful relations regarding the CDF:

$$\begin{aligned} \mathbb{P}(X > x) &= 1 - \mathbb{P}(X \leq x) \\ \mathbb{P}(a \leq X \leq b) &= \mathbb{P}(X \leq b) - \mathbb{P}(X \leq a) \end{aligned}$$

These two rules can be used to obtain probabilities over any interval, or any combination of intervals, of the event space $\mathcal{X}$.

3.4 The quantile function

The inverse of the CDF (the "inverse CDF") is also sometimes called the *quantile function* $Q(p)$ as it returns the quantiles of the distribution. The p-th quantiles is the value $x \in \mathcal{X}$ for which the cumulative probability is equal to p. The quantile function takes as an argument a value from zero to one, say p, and returns the value x such that the probability of X being less than $Q(p)$ is equal to p.

Definition 9. Let X be a discrete or continuous RV over a set $\mathcal{X}$. Then, its quantile function $Q(p)$ is given by

$$Q(p) = \{x \in \mathcal{X} : \mathbb{P}(X \leq x) = p\} \tag{6}$$

where $p \in [0, 1]$.

We can read (6) as "find the value of x in the set $\mathcal{X}$ such that $\mathbb{P}(X \leq x)$ is equal to p". The quantile function is frequently used in statistics. One of its most well known uses is to the median of a probability distribution, i.e.,

$$\text{median}\,[X] = Q(p = 1/2).$$

The median is the value of $\mathcal{X}$ which divides the probability distribution in half in terms of probability mass. If one was to randomly generate values from a random variable with a median of m, then we would expect around half of our realisations to be smaller than m and half of realisations to be greater than m. Two other values which are commonly utilised in statistics are the first and third quartiles, $Q(p = 1/4)$ and $Q(p = 3/4)$. Knowledge of these quantities give us an immediate appreciation of the

basic manner in which probability is distributed over the event space, in much the same way knowledge of these quantities from a dataset gives us information on how the data values are distributed.

3.5 The mode of a distribution

The mode of a probability distribution is defined as the value in the set $\mathcal{X}$ for which the probability is largest.

Definition 10. Let X be a discrete (continuous) RV over a set $\mathcal{X}$, characterised by a probability distribution (density) $p(x)$; then, the mode of $p(x)$ is given by

$$\text{mode}\,[X] = \arg\max_{x \in \mathcal{X}} \{p(x)\}\,.$$

For a discrete probability distribution the mode is a well defined element from the probability space, and the mode itself is a reasonable measure of the most likely, and therefore, most "typical" value of a realisation of the random variable X. In contrast, if X is continuous then the mode, which defines the value of x which maximises the probability density function, has a much less clear interpretation – this is because the probability of an real number is exactly equal to zero (see Section 3). However, the mode of a continuous distribution is still representative of the typical, or average value of a distribution, as it captures a value for which the probability density is greatest – and given that most probability density functions are continuous, one would also expect the density to be high at least in a small neighbourhood of this value.