FIT2086 Lecture 2 Notes, Part II
Standard Probability Distributions

Dr. Daniel F. Schmidt[*]

August 21, 2020

1 Statistical Models as Probability Distributions

In Lecture 1 we examined the basic ideas underlying probability distributions, and the way in which the ideas of probability and randomness provide us with a natural language for describing observational data. An obvious question that arises is: how do we specify the probability distributions for these observations? In Lecture 1 we looked at the idea of directly assigning probabilities to each of the possible values, $\mathcal{X}$, that our random variable can assume. If the number of events is small, and we have some idea about the underlying physical phenomena (population) we are modelling, this approach can potentially be fruitful. However, if the number of possible values our observations can assume is large – even in the order of hundreds – it rapidly becomes infeasible. Furthermore, in practice, the data values that we record can generally take on an infinite number of different values; That is, the set $\mathcal{X}$ of values our RV can take may be infinitely large – consider, for example, the heights of people in a population, or the distances to a set of stars, etc. This means that direct assignment of probabilities to each event in the sample space $\mathcal{X}$, as done in the examples in Lecture 1, is impossible. To overcome this problem we instead employ parametric probability distributions.

Definition 1. Let $p(x \,|\, \boldsymbol{\theta})$ denote a probability distribution (density) over the event space $\mathcal{X}$ for all values of $\boldsymbol{\theta} \in \Theta$, with $\dim(\Theta) < |\mathcal{X}|$. Then, we call

- $p(x \,|\, \boldsymbol{\theta})$ a parametric probability distribution (model);
- $\boldsymbol{\theta}$ the parameters of the distribution (model); and
- the set Θ the parameter space.

The above specification can be interpreted as a probability distribution for which the assignment of probabilities to the values in the event space $\mathcal{X}$ is determined or controlled by the choice of the particular values of $\boldsymbol{\theta}$, which we call the *parameters* of the distribution. By varying these parameters we can produce a whole range (a "family") of different probability distributions. In general, the number of parameters is quite small – usually no more than two or three – and always less than the number of different values x that the RV X can assume. This means that we can build probability distributions over infinite event spaces by the specification of only a small handful of tunable parameters.

As the probability distribution function is controlled by the values of the model parameters $\boldsymbol{\theta}$, it also means that the properties of the RV X associated with the corresponding probability distribution, such as the mean and variance, are also implicitly determined by the values of the parameters $\boldsymbol{\theta}$. This is obvious if we consider the expected value of X:

$$\mathbb{E}\left[X\right] = \int_{\mathcal{X}} x\, p(x \,|\, \boldsymbol{\theta}) dx.$$

It is clear that by varying the values of the parameters $\boldsymbol{\theta}$ we will change the assignment of probabilities to the events $x \in \mathcal{X}$, and therefore change the expected value of X. The same applies to all the properties of X – the variance, the median, the cumulative distribution function, the quantiles, and so on. All these quantities are functions of the probability distribution $p(x \,|\, \boldsymbol{\theta})$, and therefore also functions of the parameters $\boldsymbol{\theta}$ that control the assignment of probabilities to the values of $\mathcal{X}$.

1.1 Parametric Distributions as Models of Reality

Why are parametric probability distributions so important to the fields of data science and statistics? The aim of quantitative data science is to build mathematical descriptions of natural phenomena. Whether such phenomena involves the behaviour of electrons, the flight paths of bees or the preferences of consumers on movie rental websites, the basic process remains the same: (i) we observe empirical data; (ii) we use this data to build ("infer") a mathematical description ("model") of the phenomenon that generated the data; and (iii) we use this model as a basis for either drawing conclusions about aspects of the phenomenon, or making predictions about its future behaviour.

Parametric probability distributions are ideally suited for use as *models* of reality, as they have a small number of tunable parameters whose behaviour can be readily understood through their relationship to basic quantities such as the mean and variance. By choosing these parameters appropriately we can often obtain a reasonable *approximation* to the true, unknown, underlying probability distribution that characterises a phenomenon of interest. While such models will clearly differ from reality to some smaller or larger extent, they provide simple, interpretable representations of reality. A quote by the famous statistician G.E.P.Box comes to mind:

"All models are wrong, but some are more useful than others".

Statisticians and data scientists acknowledge that the simple parametric probability distributions they use when modelling phenomena are unlikely to ever provide exact representations of reality. However, the reduction of a data sample from n observations down to a small number of simple, interpretable model parameters frequently allows data scientists and statistians to obtain great insights into the complex processes that generated the data they are analysing.

In this Lecture we will review a number of useful parametric probability distributions that are frequently employed as models of reality by data scientists. These distributions cover both continuous data, as as well as discrete data, including nominal categorical data. The list is far from exhaustive – there exists an enormous number of parametric distributions in the data science literature, many of which generalise the distributions we will examine. However, they do serve to introduce you to many of the most important parametric probability distributions, and offer at least one model for all three of the basic data types you are likely to encounter in practice (continuous, numerical discrete and categorical). The way in which we can choose the parameters so that these distributions provide reasonable approximations to the process that generated our observed data is deferred to Lecture 3. For the moment we will focus on the definitions and properties of these parametric probability distributions.

2 The Gaussian (normal) distribution

We begin our exploration of parametric probability distributions with what is perhaps the single most important distribution in data science. If we are interested in a distribution over the real (continuous) numbers, i.e., $\mathcal{X} \subseteq \mathbb{R}$, then there is clearly an infinite number of different values that our RV X can assume. In this setting, one of the most frequently employed parametric distributions is the *Gaussian distribution*, named after the German mathematician Carl Friedrich Gauss (1777–1855). This distribution is also known as the *normal distribution*, for reasons that will become clear in Lecture 4. The term "normal" suggests that the Gaussian distribution is some sort of default distribution for modelling continuous data, and in a sense this is true: the Gaussian distribution is undoubtedly the most commonly utilised continuous distribution in data analysis. The Gaussian distribution is characterised by two parameters: μ, which controls the mean of the distribution, and σ^2 which controls the variance of the distribution.

Definition 2. The probability density function for a Gaussian distribution with mean μ and variance $\sigma^2 > 0$ is given by:

$$p(y \,|\, \mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{1}{2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right). \tag{1}$$

where we recall that $x^{1/2} = \sqrt{x}$.

At first glance, the formula (1) may appear somewhat intimidating, but at its core is actually quite straightforward. The first part is simply a normalising constant, which ensures that the distribution integrates to one, as required, for all values of μ and σ^2. The main component is the second term, which says that the probability density of values $y \in \mathcal{Y}$ grows smaller at an exponential rate as the square of the difference from the value and the mean μ of the distribution increases. This means that the normal distribution attains a maximum value of $1/\sqrt{2\pi\sigma^2}$ when $y = \mu$, and tails rapidly off towards zero as $|\mu - y| \to \infty$ in a symmetric fashion. The division by σ^2 controls the rate at which the decay occurs; the bigger the value of σ^2, the slower the probability tails off towards zero for values further away from the mean. Figure 1 shows four examples of the normal distribution; note that they all have the same essential form – a smooth bell shape that is symmetric around the mean μ and tails off to zero as $|y| \to \infty$. You could reproduce this graph by plugging in the appropriate values for μ and σ^2 in (1) and plotting the function from $y = -5$ to $y = 5$.

We will now introduce some special short-hand notation that is frequently used in statistics and probability to describe when a random variable follows (is distributed as per) one of the standard probability distributions.

Definition 3. If Y follows a normal distribution with mean μ and standard deviation σ, then we can write

$$Y \sim N(\mu, \sigma^2)$$

where the "$\sim$" should be read as "is distributed as per a", and the $N(\mu, \sigma^2)$ is shorthand that denotes a normal distribution with mean μ and variance σ^2.

As we explore further we will see that there exists a number of simple short-hand symbols for the common probability distributions. Using the above notation, we can say that if $Y \sim N(\mu, \sigma^2)$, then

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

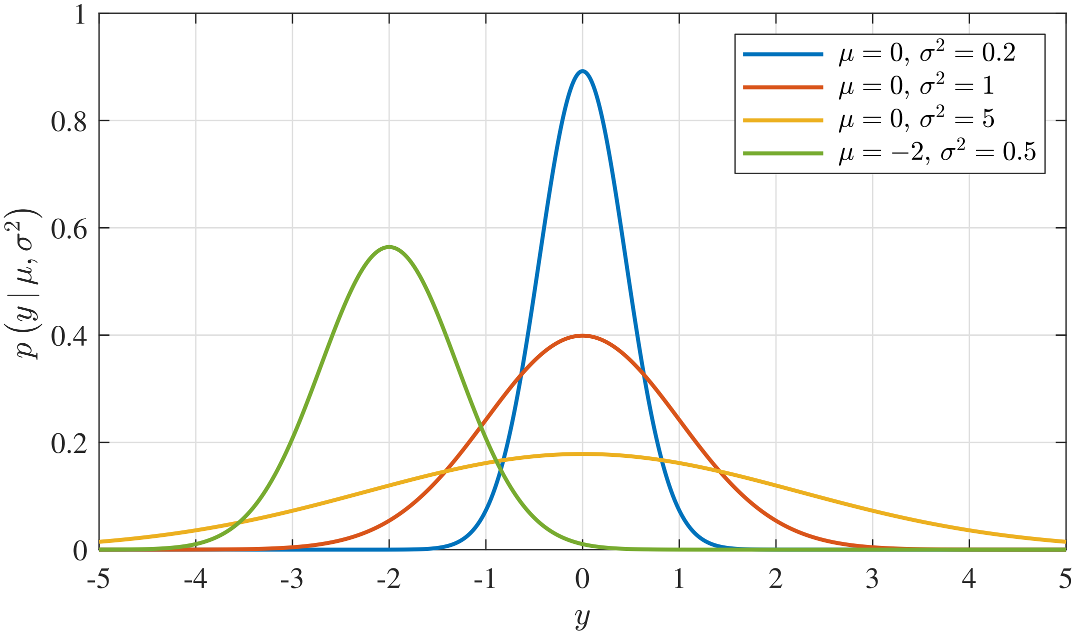

Figure 1: Probability density functions for several normal (Gaussian) distributions. The solid black curve is the *standard normal distribution*. Note that the normal distribution is symmetric and tails off to zero as $|y| \to \infty$.

This tells us that in the case of a random variable Y that is distributed as per a normal distribution, the mean of Y is equal to the parameter μ, the variance of Y is equal to the parameter σ^2. Therefore, by varying the parameters μ and σ^2, we can produce a symmetric distribution with any particular combination of mean and variance that we desire. This clearly demonstrates the way in which parametric probability distributions can be used as models of data – by setting parameters we can control certain properties of the random variables to match/model some real process. The Gaussian distribution is symmetric around its mean μ. This implies that the mean μ is also equal to the mode of the distribution (most common value, i.e., the value y that maximises (1)) as well as the median (i.e., the value m for which $\mathbb{P}(Y \leq m) = 0.5$) of the distribution. Therefore, in the case of a Gaussian distribution the parameter μ simultaneously controls the mean, mode and median of the resulting normal distribution.

2.1 The Gaussian CDF

As the normal distribution models continuous RVs, the cumulative distribution function (cdf) is required to determine the probability mass assigned to any interval of $\mathbb{R}$ (see Section X for a refresher on continuous RVs). In the case of the normal distribution (1) it turns out that the cumulative distribution function does not actually have a closed form – that is, there exists *no simple formula* that you can write down for the cdf! This may sound surprising, but such situations arise frequently in applied mathematics. However, while this may seem like a major impediment, all is not lost: even though there exists no simple, closed form expression for the cdf, there do exist a number of efficient *numerical* algorithms that allow us to compute the cdf accurately, and all standard statistical software packages, such as R, Python or MATLAB, have implementations these algorithms. Furthermore, at a coarse level, there are several well known rules regarding the cumulative distribution that are both important and highly useful; specifically, if $Y \sim N(\mu, \sigma^2)$, then

1. approximately 68.27% of probability falls within $(\mu - \sigma, \mu + \sigma)$, irrespective of the values of μ and σ. That is, 68.27% of samples from a normal distribution will fall within one standard deviation of the mean, i.e.,
$$\mathbb{P}(\mu - \sigma \leq Y \leq \mu + \sigma) \approx 0.6827;$$
2. approximately 95% of probability falls within $(\mu - 1.96\sigma, \mu + 1.96\sigma)$, i.e.,
$$\mathbb{P}(\mu - 1.96\sigma \leq Y \leq \mu + 1.96\sigma) \approx 0.95; \text{ and}$$
3. approximately 99.73% of probability falls within $(\mu - 3\sigma, \mu + 3\sigma)$.

These facts are interesting for two reasons: the first is that the rules hold regardless of the values of μ and σ, which is one of the reasons the normal distribution is so widely used. The second is that we can see that the probability drops off very rapidly as $|y - \mu|$ grows, with less than 0.003% of the probability being assigned to values for which $|y - \mu| > 3\sigma$. This tells us that if we choose a normal distribution to model our population, then we believe that almost all observations will lie within three standard deviations of the population mean, and large deviations from the mean are almost never expected to arise.

2.2 The Gaussian CDF and z-scores

We noted in our discussion above that the normal distributions in Figure 1 all look very similar. In fact, it turns out that normal random variates satisfy a very important *self similarity* property, that is frequently exploited in statistics. Essentially, every normal distribution is an appropriately scaled and shifted (translated) version of the so called standard unit normal $N(0, 1)$. An important implication of this self-similarity property is the following result.

Fact 1. If $Z \sim N(0, 1)$, then
$$Y = \sigma Z + \mu$$
is distributed as per a $N(\mu, \sigma^2)$; conversely, if $Y \sim N(\mu, \sigma^2)$ then
$$Z = (Y - \mu)/\sigma \tag{2}$$
is distributed as per a standard unit normal $N(0, 1)$.

Fact 1 is important becomes it tells us that if we take a random variable that follows a unit normal distribution, we can turn it into a new RV that follows an *arbitrary* normal distribution by multiplying it by the desired standard deviation and then adding the mean. Conversely, we can convert a RV $Y \sim N(\mu, \sigma^2)$ to a RV following a unit normal by reversing this transformation. This technique is used extensively in statistics, and the resulting RV derived by this approach is often referred to as a "z-score". A useful implication of this fact is that we only require the cumulative distribution function for the standard unit normal distribution to compute probabilities for any normal distribution. In particular, if $Y \sim N(\mu, \sigma^2)$, then from (2) we have the identity
$$\mathbb{P}(Y < y) = \mathbb{P}\left(Z < \frac{y - \mu}{\sigma}\right) \tag{3}$$
where $Z \sim N(0, 1)$.

2.3 Additivity of Normal Random Variables

The normal distribution has another interesting property: it is "closed under addition". More specifically, we have the following special result.

Fact 2. Let $Y_1 \sim N(\mu_1, \sigma_1^2)$ and $Y_2 \sim N(\mu_2, \sigma_2^2)$; then

$$Y_1 + Y_2 \sim N(\mu_1 + \mu_2,\, \sigma_1^2 + \sigma_2^2)$$

In words, this result tells us that if we have two normally distributed random variables, then their sum is also normally distributed. By simple induction, it also implies that the sum of any number of normally distributed random variables is normally distributed. This is very useful property, and greatly simplifies analysis of normally distributed data. In general, distributions that exhibit this property are called "infinitely divisible". To understand what this means, consider a random variable $Y \sim N(\mu, \sigma^2)$; then, by Fact 2, we can say that $Y = \sum_{i=1}^n Y_i$, where $Y_i \sim N(\mu_i, \sigma_i^2)$, with $n \geq 1$,

$$\sum_{i=1}^n \mu_i = \mu, \text{ and } \sum_{i=1}^n \sigma_i^2 = \sigma^2.$$

That is, *any* normally distributed random variable can be decomposed ("divided") into a sum of an arbitrary number of appropriate normally distributed RVs. The normal distribution is not the only infinitely divisible distribution, and in fact we will examine several more throughout this Lecture.

2.4 Implementation in R

The R statistical computing language has a number of functions associated with the normal distribution (and in fact, with most common distributions). These functions are

- `dnorm()`, which implements the PDF (1);
- `pnorm()`, which implements the CDF;
- `qnorm()`, which implements the quantile function (inverse CDF); and
- `rnorm()`, which lets us generate random realisations from a normal distribution.

As is standard in R, one may always use the syntax "`?dnorm`" to access help on a function (in this case, the `dnorm()`) function. Of course, in the case of Table 1, we are limited to computing probabilities for events involving z-scores less than four. One might expect that no such limit applies when using R, but in fact due to the finite precision of computer arithmetic there are limits on the size of $|z|$ for which cumulative probabilities can be accurately calculated. Perhaps even more surprising is the fact that, while theoretically

$$\mathbb{P}(Z < -|z|) = 1 - \mathbb{P}(Z < |z|)$$

is true for all z in the case of the normal distribution (due to its symmetric nature), the above equivalence does not necessarily hold for numerical computations. In fact, for the `pnorm()` function, $\mathbb{P}(Z < |z|)$ can only be calculated accurately for $|z| < 4$, while $\mathbb{P}(Z < -|z|)$ can be computed to a reasonable precision for $|z|$ as large as 30. The reasons for this discrepancy lie in the way in which digital computers encode and handle non-integer quantities. For the practitioner, it suffices to recognise that while certain statements may be mathematically equivalent, they are not necessarily equivalent when implemented on a finite precision digital computer, particularly when they involve very small (or very large) numbers. In these cases careful treatment is generally required to preserve accuracy when performing numerical computations.

3 Binary Data

3.1 The Bernoulli distribution

Let us now consider the simple, but extremely important, case of a discrete, binary RV Y; that is, a RV that can only take on one of two different values, say 0 or 1, i.e., $\mathcal{Y} = \{0, 1\}$. For example, the outcome of a toss of a coin is a binary RV, where a tail is treated as a zero and a head is treated as a one; or the diabetes status of an individual can be treated as a binary variable, with a one denoting the presence of diabetes and a zero denoting an absence of diabetes. Clearly binary data is everywhere. We can use the *Bernoulli distribution*, named after the Swiss scientist Jacob Bernoulli (1655–1705), to model these types of random variables.

Definition 4. If Y is a binary random variable following a Bernoulli distribution with probability of success $\theta \in [0, 1]$, then we say

$$Y \sim \mathrm{Be}(\theta),$$

and the probability mass function associated with Y is given by

$$p(y \,|\, \theta) = \theta^{y}(1-\theta)^{(1-y)} \tag{4}$$

for $y \in \{0, 1\}$.

The Bernoulli distribution says that the probability that $Y = 1$ is just the value of the parameter θ, and therefore by the properties of probability, that the probability that $Y = 0$ is simply $1 - \theta$. Plugging the values $y = 0$ and $y = 1$ into (4) will quickly verify that $p(y = 0 \,|\, \theta) = (1 - \theta)$ and $p(y = 1 \,|\, \theta) = \theta$. We often say that θ is the probability of observing a "success" – for example, a heads coming up in a coin toss. The usage of the word "success" to describe an outcome of $Y = 1$ is simply standard terminology, and it should obviously not be confused with any value judgement attached to the outcome associated with $Y = 1$!

3.1.1 Basic Properties of the Bernoulli Distribution

By using the formulas for expectation and variance presented in Lecture 1, and the fact that Y can only assume the two values $Y = 0$ and $Y = 1$, it is straightforward to verify that

$$\begin{aligned} \mathbb{E}\left[Y\right] &= \theta, \\ \mathbb{V}\left[Y\right] &= \theta(1-\theta). \end{aligned}$$

We see that both expectation and variance depend on the single parameter θ. The expected value of a Bernoulli trial is simply the probability of success; that is, if we toss a coin once we expect to see θ heads. The variance function for the Bernoulli is shown in Figure 2 and has an interesting shape. The variance approaches zero as θ approaches $\theta = 0$ or $\theta = 1$, and is largest when $\theta = 1/2$. This is due to the fact that the larger (smaller) the probability of success, the less chance there will be of observing a failure (success) in a series of realisations of the Bernoulli random variable, and therefore the less overall variability in the sequence as most observations will be equal to 1 (0). In the extreme case that $\theta = 0$ we will never see a success and there will be no variability in a sequence of realisations regardless of how many we observe, with a similar conclusion for the case that $\theta = 1$.

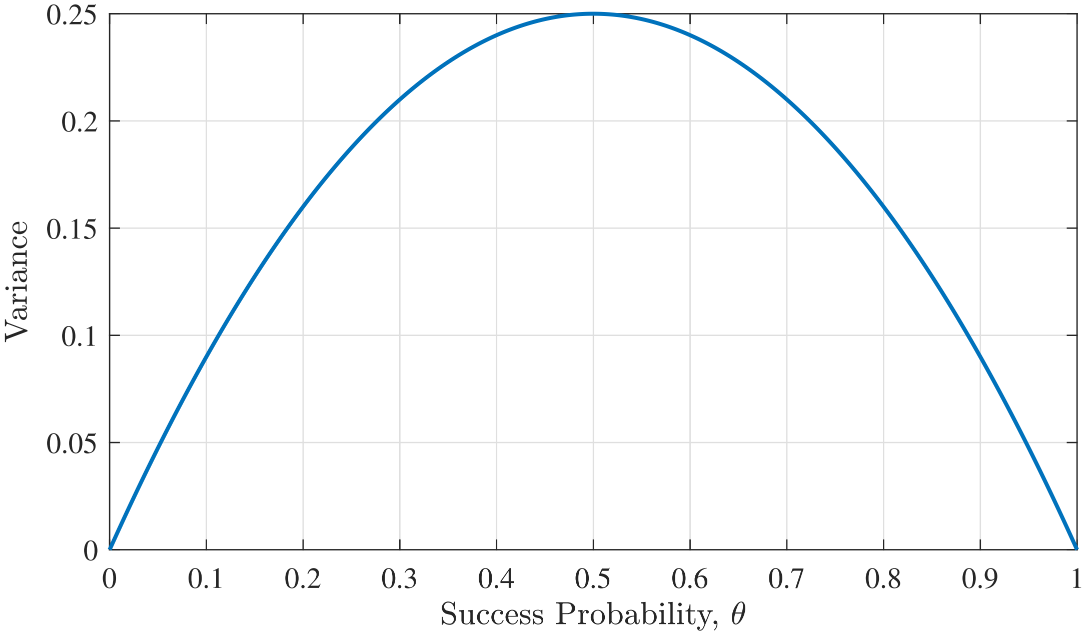

Figure 2: The variance of a Bernoulli random variable as a function of the success probability θ. Note that the variance is maximised when $\theta = 1/2$ (i.e., equal chance of zero or one) and minimised when $\theta = 0$ and $\theta = 1$.

3.2 The Binomial Distribution

Consider a series of realisations of n binary random variables, say $y_1, \ldots, y_n$. It is frequently the case that we wish to model the total number m of successes in this series of n realisations. If we assume that each of the binary random variables comes from the same Bernoulli distribution (i.e., each has the same probability of success), then number of successes follows what is known as a *binomial distribution*.

Definition 5. If Y follows a binomial distribution with probability of success $\theta \in [0, 1]$ and number of trials $n \geq 1$, then we say

$$M \sim \mathrm{Bin}(\theta, n),$$

and the probability mass function associated with M is given by

$$p(m \,|\, \theta) = \binom{n}{m} \theta^m (1 - \theta)^{(n-m)} \tag{5}$$

for $m \in \{0, \ldots, n\}$.

The $\theta^m(1 - \theta)^{(n-m)}$ term in (5) is the probability of observing *any* sequence of n binary variables with exactly m successes. However, as the binomial distribution models the number of successes in a sequence, all sequences with m ones in them, regardless of the configuration of the sequence, are equivalent. For example, if we have $n = 4$ trials, the number of sequences that have exactly $m = 2$

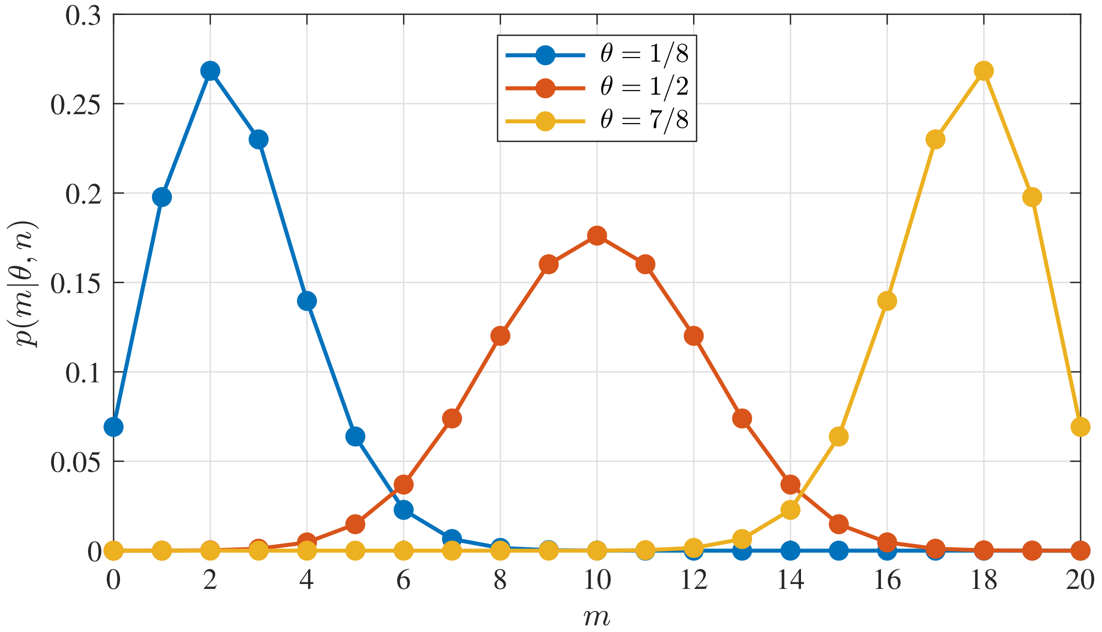

Figure 3: Probability mass functions for three binomial distributions. Note that the distribution is symmetric when $\theta = 1/2$ (a fair coin), and that the two distributions associated with $\theta = v$ and $\theta = 1 - v$ are mirrors of each other due to the arbitrary nature of labelling what is a "success" and a "failure".

successes is six; e.g.,

$$(1,1,0,0), (1,0,1,0), (1,0,0,1), (0,1,1,0), (0,1,0,1), (0,0,1,1)$$

all have exactly $m = 2$ successes. The $\binom{n}{m}$ component in (5) accounts specifically for this fact; it counts the number of equivalent sequences (i.e., those with m successes) among all binary sequences of length n and is given by

$$\binom{n}{m} = \frac{n!}{(n-m)!m!}.$$

This captures the number of ways of choosing m objects out of a total of n unique objects, and is called the binomial coefficient, from which the binomial distribution gets its name. For our example of $m = 2$ successes in $n = 4$ trials we can evaluate $\binom{4}{2}$ and see that it is indeed equal to six. The $m! = 1 \times 2 \times 3 \times \cdots \times m$ denotes the factorial function. The binomial distribution has two parameters: θ, the probability of seeing a success, and n, the number of binary trials we are interested in. Interestingly then, the set of values m over which the binomial distribution is defined depends on n; in fact, m can only take on the integer values $\{0, 1, 2, \ldots, n\}$.

3.2.1 Basic Properties of the Binomial Distribution

Let us assume, without any loss of generality, that we code our binary variables as a zero for a "failure" and a one for a "success". Then, given a sequence of n realisations of random variables $y_1, \ldots, y_n$ the number of successes in the sequence can be written as the sum

$$m = \sum_{i=1}^{n} y_i.$$

Therefore, we see that the binomial random variable M, which counts the number of successes in n identical trials, is defined as a sum of n independent Bernoulli random variables. This fact, coupled with the properties of expectations and variances of sums of independent random variables allows us to to easily verify that

$$\begin{aligned} \mathbb{E}[M] &= n\theta, \\ \mathbb{V}[M] &= n\theta(1-\theta), \end{aligned}$$

that is, the mean and variance of M is just n times the mean and variance of a single Bernoulli trial with success probability θ. Further, the fact that M is the sum of n indepenently and identically distributed Bernoulli random variables implies that binomial random variables are additive in a particular way.

Property 1. If $M_1 \sim \text{Bin}(\theta, n_1)$ and $M_2 \sim \text{Bin}(\theta, n_2)$ then

$$M_1 + M_2 \sim \text{Bin}(\theta, n_1 + n_2).$$

This fact is straightforward to verify; simply note that M_1 is the sum of n_1 Bernoulli RVs with success probability θ and M_2 is the sum of n_2 Bernoulli RVs with success probability θ. Therefore, $M_1 + M_2$ is just the sum of $n_1 + n_2$ Bernoulli RVs with success probability θ, which itself follows a binomial distribution with success probability θ and total number of trials $n = n_1 + n_2$, by definition.

3.2.2 Implementation in R

The R functions associated with the binomial distribution are `dbinom()` (the probability distribution function), `pbinom()` (the cdf function), `qbinom()` (the quantile function) and `rbinom()` (a function to generate random realisations from a binomial distribution). As a side note, you can generate a RV from a Bernoulli distribution by using a binomial distribution with $n = 1$.

4 Uniform Distributions

The binomial distribution gives us a distribution over the integers $\{0, \ldots, n\}$. This is obviously useful, but it has the property that the probability is concentrated around the value $n\theta$. Frequently, however, we want to model situations in which no outcome is more likely than any other outcome. In this case can use the uniform distribution. The uniform distribution is also interesting as it is one of the few that has both a discrete and a continuous version, and both find frequent use in practical data science. The classic example of a discrete uniform distribution would be the outcome of a roll of a fair six-sided dice – each value from one to six is equally likely to occur. An example of a uniformly distributed continuous phenomena might be the location of rain drops falling onto a soccer pitch.

4.1 Discrete uniform distribution

We first examine the discrete case, in which we have a set of integers from a through to b, inclusive, and we believe that each one of these integers is equally likely.

Definition 6. If Y follows a discrete uniform distribution with parameters a and $b \leq a$ then we say

$$Y \sim \mathrm{U}(a, b)$$

and the probability mass function associated with Y is given by

$$\mathbb{P}(Y = y \,|\, a, b) = \begin{cases} \dfrac{1}{b - a + 1} & \text{if } Y \in \{a, a+1, \ldots, b\} \\ 0 & \text{otherwise} \end{cases} . \tag{6}$$

where $a, b, y \in \mathbb{Z}$.

It is trivial to establish that (6) sums to one. The name "uniform distribution" is obvious if one considers (4); every event is given the same probability of occuring, i.e., the probabilities are *uniform* (the same). The mean and variance of an RV X that follows a uniform distribution are straightforward to establish by using well known identities related to the sums of integers, and are given by:

$$\begin{aligned} \mathbb{E}\left[Y\right] &= \frac{a+b}{2}, \\ \mathbb{V}\left[Y\right] &= \frac{(b-a+1)^2 - 1}{12}. \end{aligned}$$

It is useful to note here that as with expectations in general, the expected value of X is not necessarily an integer; in fact, we can see that $\mathbb{E}\left[X\right]$ will be an integer if and only if a and b are both even integers. The cumulative distribution function has a closed form expression due to the simple form of the probability mass function, and is given by:

$$\mathbb{P}(Y \leq y) = \begin{cases} 0 & \text{if } y < a \\ \dfrac{y - a + 1}{b - a + 1} & \text{if } y \in \{a, \ldots, b\} \\ 1 & \text{if } y > b \end{cases} .$$

An important point to reflect on here is the impossibility of putting a uniform distribution over the *complete set* of integers $\mathbb{Z}$, i.e., to say that *every possible integer* is equally likely to occur. To see that this is the case, let $a = 0$ and take the limits of (6) as $b \to \infty$. In this case the probability of any single event in our infinite set $\{0, \ldots, \infty\}$ will be zero. We will examine one possible distribution that is appropriate for the set of positive integers later, and we will see that it has the property of assigning ever smaller probability to very large integers.

4.2 Continuous Uniform Distribution Over an Interval

As mentioned previously, the uniform distribution is interesting in that it is one of the few distributions which exists under the same name for both continuous and discrete RVs. We first examine the case in which the distribution is defined over an interval of the real numbers, i.e., over (a, b) with $b < a$.

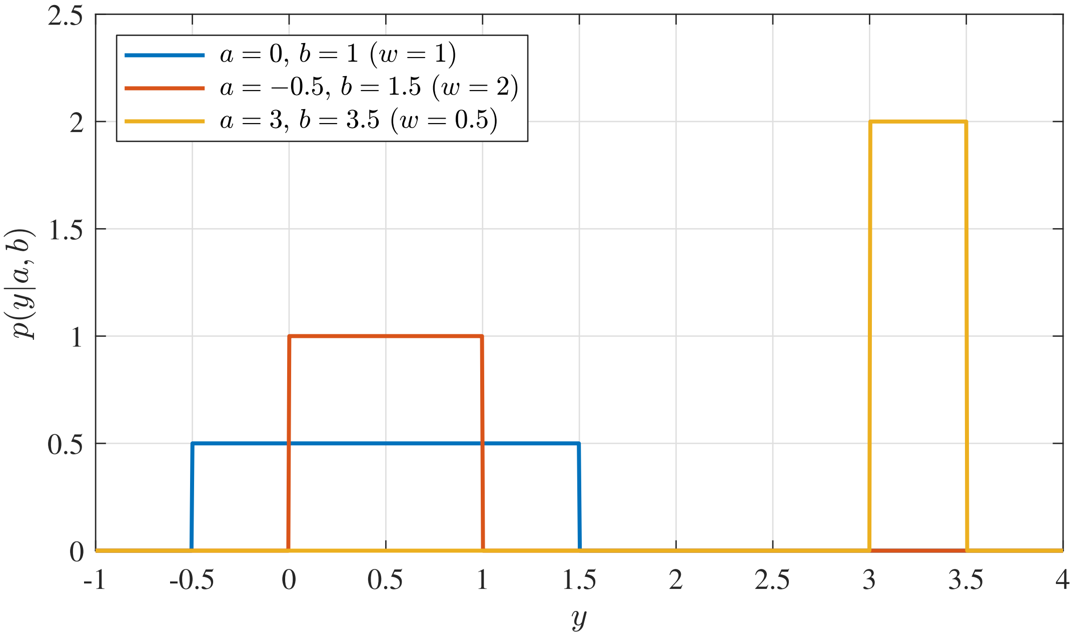

Figure 4: Uniform distributions for $a = 0$, $b = 1$ $(w = 1)$; $a = -0.5$, $b = 1.5$ $(w = 2)$, and $a = 3$, $b = 3.5$ $(w = 0.5)$. Note that the probability density is zero outside the interval over which the uniform distribution is defined. The connecting vertical lines are guides for the eye only.

Definition 7. If Y follows a continuous uniform distribution over the interval (a, b), with $b < a$, we say

$$Y \sim \mathrm{U}(a, b)$$

and the probability density function associated with Y is given by

$$p(y \,|\, a, b) = \begin{cases} 0 & \text{for } y < a \\ \dfrac{1}{b-a} & \text{for } a \leq y \leq b \\ 0 & \text{for } y > b \end{cases},$$

where $a, b, x \in \mathbb{R}$.

This density is a *piecewise* (i.e., broken into distinct "pieces") continuous function composed of three different pieces: one when $x < a$, one when $x > b$ and a piece for x between a and b. The parameter a determines the start of the uniform distribution, and the parameter b determines the end of the distribution; $w = b - a$ is an alternative parameter that instead determines the *width* of the distribution. Figure 4 shows three examples of uniform distributions with different a and b values. Notice how have the exact same shape – only the range on which they are defined, and the height are different.

4.2.1 Properties of the Uniform Distribution over an Interval

If $Y \sim U(a, b)$, then the mean and variance of Y are then given by

$$\begin{aligned} \mathbb{E}[Y] &= \frac{a+b}{2} = a + \frac{w}{2}, \\ \mathbb{V}[Y] &= \frac{(b-a)^2}{12} = \frac{w^2}{12}, \end{aligned}$$

where we see that the variance is more concisely described in terms of the width of the uniform distribution.The above quantities are straightforward to compute due to the simple form of the PDF. This simple form also leads to a particular (piecewise) simple form for the CDF:

$$\mathbb{P}(Y \leq y) = \begin{cases} 0 & \text{if } y < a \\ \dfrac{y-a}{b-a} & \text{if } y \in (a, b) \\ 1 & \text{if } y > b \end{cases}.$$

Due to the piecewise nature of the discrete uniform distribution, the CDF for the continuous uniform distribution is itself a piecewise function.

5 The Poisson Distribution

Frequently we are interested in modelling counts of occurrences of things over a time period; for example, the number of telephone calls made in an hour, or the number of people kicked to death by horses in a single year, etc. We have previously seen the binomial distribution, which models counts of successes in a fixed number of trials, and the uniform distribution over the integers. A potential problem with trying to use either of these distributions to model arbitrary counts of events is that the *support* of the distributions, i.e., the set of values which are assigned non-zero probabilities, is finite.

When modelling the number of occurrences of a phenomena we would potentially have no idea about the maximum possible number we might observe. In this case the set of values our RV can take on is the set of non-negative integers, i.e., $\mathcal{Y} = \{0, 1, 2, 3, \ldots\}$, and the binomial or discrete uniform distribution are no longer appropriate. A suitable distribution for these types of random variables is the *Poisson distribution* (named after the French scientist Simeon Poisson (1781–1840).

Definition 8. If Y follows a Poisson distribution with rate $\lambda > 0$, then we say

$$Y \sim \text{Poi}(\lambda).$$

The probability mass function associated with Y is given by

$$p(y \,|\, \lambda) = \frac{\lambda^y \exp(-\lambda)}{y!}, \tag{7}$$

where $y!$ denotes the factorial function.

The PMF of the Poisson distribution is quite simple; the $e^{-\lambda}$ term is a normalizing constant required to ensure the probability mass function sums to one, while the $\lambda^y/y!$ term controls the behaviour of the distribution as y grows. If $\lambda > 1$ is large enough, this term will first grow as y increases. However, the factorial in the denominator always grows sufficiently fast that at some point the probability mass will begin to decrease towards zero as y continues to grow.

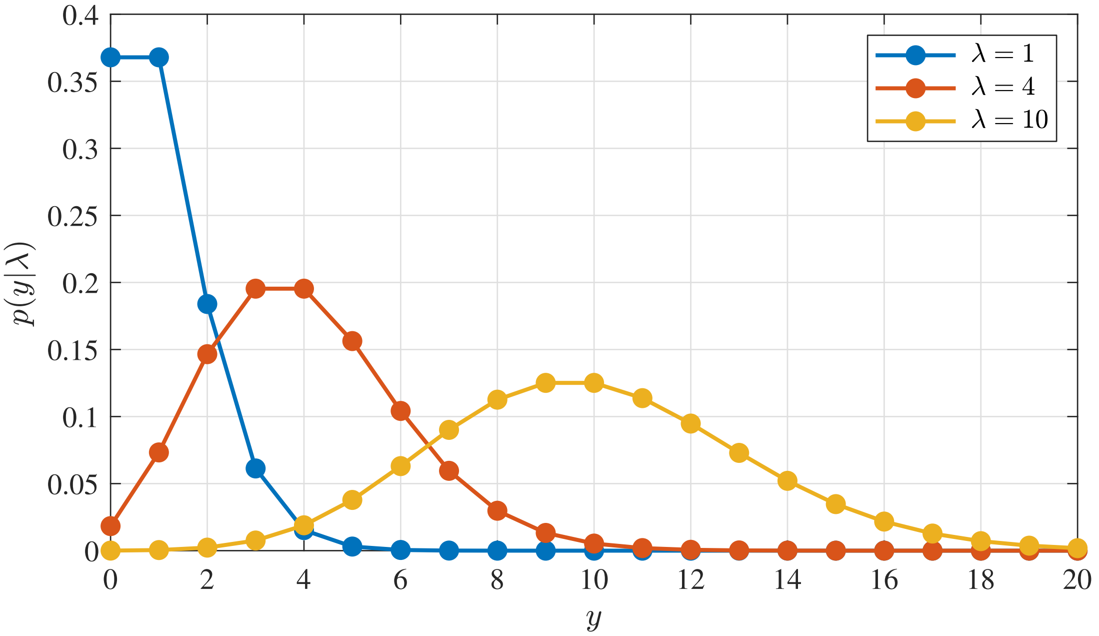

Figure 5: Poisson distribution for $\lambda = 1$, $\lambda = 4$ and $\lambda = 10$. The distribution is defined only on the integers – the connecting lines are only guides for the eye.

5.1 Basic Properties of the Poisson Distribution

The mean and variance of a RV Y that follows a Poisson distribution with rate λ are

$$\begin{aligned} \mathbb{E}\left[Y\right] &= \lambda, \\ \mathbb{V}\left[Y\right] &= \lambda. \end{aligned}$$

The the mean and variance of a Poisson distribution are linked, with the variance growing with increasing mean; in fact, for the Poisson distribution the relationship is particularly simple as the mean and variance are the same. This tells us that as the rate λ grows, not only does the average number of events increases, but that the average spread around the mean also increases. The Poisson distribution, for several values of λ, is shown in Figure 5. Note the several aspects of the Poisson distribution: (i) that as λ gets bigger, the distribution becomes more spread out around $k = \lambda$, showing that the variance is increasing; (ii) the probability mass decreases to zero very rapidly as $y \to \infty$, and (iii) the Poisson distribution exhibits a unimodal "bell-like" shape when $\lambda > 1$.

5.2 Further Properties of Poisson Random Variables

Poisson random variables have an important additivity property similar to that possessed by normally distributed random variables.

Property 2. If $Y_1 \sim \text{Poi}(\lambda_1)$ and $Y_2 \sim \text{Poi}(\lambda_2)$ then

$$Y_1 + Y_2 \sim \text{Poi}(\lambda_1 + \lambda_2),$$

so that the sum of two Poisson RVs Y_1 and Y_2 is also a Poisson RV with a rate equal to the sum of the rates of the two RVs Y_1 and Y_2.

Additionally, just as with the normal distribution, we have the more general result that if $Y \sim \text{Poi}(\lambda)$, then we can always write $Y = \sum_{i=1}^{n} Y_i$ where

$$Y_i \sim \text{Poi}(\lambda_i) \text{ and } \sum_{i=1}^{n} \lambda_i = \lambda.$$

That is, any Poisson RV can be decomposed into a sum of appropriate Poisson RVs (is infinitely divisible). This has a very important implication for Poisson RVs. Remember that a Poisson RV models the number of events recorded in a finite time period, say T; $X_T \sim \text{Poi}(\lambda)$ therefore determines the distribution of the number of events occurring in the time period T. Knowing this, one might then then ask what is the distribution of the number of events of the same phenomena, say $X_{T/k}$, occurring in a different time period T/k? Using the above divisibility result we have the result that

$$X_{T/k} \sim \text{Poi}(\lambda/k) \tag{8}$$

so that the number of events in time period T/k is Poisson RV with a rate of λ/k. This means that if our phenomena of interest follows a Poisson distribution we need only establish the rate for some (arbitrary) fixed time period. The result (8) shows us that it is then straightforward to determine the distribution of the same phenomena for any other time period; and further, that the resulting random variable will always follow a Poisson distribution.

5.3 The Poisson Distribution in R

The R functions associated with the Poisson distribution are `dpois()` (the probability distribution function), `ppois()` (the cdf function), `qpois()` (the quantile function) and `rpois()` (a function to generate random realisations from a Poisson distribution).

5.4 The Appropriateness of the Poisson Distribution

The Poisson is only one of a number of different distributions that can be used to model count data. We therefore conclude by identifying when it is likely to be appropriate to use a Poisson distribution as a model of a count process. To do so we can examine the assumptions about the underlying process which lead to a Poisson distribution; if the process we are interested in modelling (approximately) satisfies these assumptions then the Poisson may be appropriate. The assumptions are:

1. That the occurence of one event in a time period has no effect on the probability that a second event will occur in the same time period; i.e., that the events occur *independently* of each other. An example would be the number of telephone calls received by a call centre in an hour period on a regular day.

2. That the rate at which the events occur is constant over time. That is, the rate λ cannot be higher in some time intervals and lower in other time intervals. If this is the case then we say the process generating the count data is *stationary*. The telephone call centre example made above may not precisely meet this criterion; it is highly likeley that more calls are received during the weekends than on weekdays, or vice versa, depending on the nature of the call centre.

3. That it is not possible for two events to occur exactly simultaneously. This condition is quite unrestrictive.

4. That the probability of an event occuring within some small time interval t is proportional to t; that is, it is $k > 0$ times more likely for an event to occur over a time period of length kt than over a time period of length t. In the case of the call centre example, this assumption seems reasonable as the shorter the time interval we consider, the smaller the probability of receiving a call in the time interval.

Even if all the conditions are not precisely met, the Poisson can still be an adequate model. For example, our call centre example may violate assumption 2 (the stationarity assumption), but if the difference in the average rate of phone calls received on weekdays and weekends is not substantial, the Poisson may still provide an adequate model for our data.

$\lvert z \rvert$	$\mathbb{P}(Z < -\lvert z \rvert)$	$\mathbb{P}(Z < \lvert z \rvert)$	$\lvert z \rvert$	$\mathbb{P}(Z < -\lvert z \rvert)$	$\mathbb{P}(Z < \lvert z \rvert)$
0.000	0.500000	0.500000	2.047	0.020353	0.979647
0.093	0.462943	0.537057	2.140	0.016196	0.983804
0.186	0.426204	0.573796	2.233	0.012789	0.987211
0.279	0.390096	0.609904	2.326	0.010020	0.989980
0.372	0.354912	0.645088	2.419	0.007790	0.992210
0.465	0.320924	0.679076	2.512	0.006009	0.993991
0.558	0.288375	0.711625	2.605	0.004598	0.995402
0.651	0.257471	0.742529	2.698	0.003491	0.996509
0.744	0.228382	0.771618	2.791	0.002630	0.997370
0.837	0.201237	0.798763	2.884	0.001965	0.998035
0.930	0.176125	0.823875	2.977	0.001457	0.998543
1.023	0.153093	0.846907	3.070	0.001071	0.998929
1.116	0.132151	0.867849	3.163	0.000781	0.999219
1.209	0.113273	0.886727	3.256	0.000565	0.999435
1.302	0.096403	0.903597	3.349	0.000406	0.999594
1.395	0.081455	0.918545	3.442	0.000289	0.999711
1.488	0.068326	0.931674	3.535	0.000204	0.999796
1.581	0.056894	0.943106	3.628	0.000143	0.999857
1.674	0.047024	0.952976	3.721	0.000099	0.999901
1.767	0.038577	0.961423	3.814	0.000068	0.999932
1.860	0.031410	0.968590	3.907	0.000047	0.999953
1.953	0.025381	0.974619	> 4.000	< 0.000032	> 0.999968

Table 1: Cumulative Distribution Function for the Standard Normal Distribution $Z \sim N(0, 1)$