# FIT2086 Lecture 2 Summary
# Probability and Probability Distributions

Dr. Daniel F. Schmidt

August 14, 2017

## 1 Part I: Random Variables and Probability Distributions

**Probability distribution over 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. We sometimes call $\mathcal{X}$ the event space, and seeing any $x$ from $\mathcal{X}$ as observing the event $X = x$. We use the language

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

to describe the probability that the RV $X$ will take on the value $x$ from $\mathcal{X}$. A probability distribution satisfies two important properties:

$$\begin{aligned} \mathbb{P}(X = x) &\in [0, 1], \\ \sum_x \mathbb{P}(X = x) &= 1, \end{aligned}$$

which says that the probability of any event lies between zero and one, and the total probability over all the possible values $\mathcal{X}$ of $x$ is always equal to one. Another important property is the additivity of the probability of mutually exclusive events. What this means is that the probability of ($X$ taking on values from some set $A$) OR ($X$ taking on values from another set $B$) is equal to

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

if $A$ and $B$ have no values in common.

**Probability distributions over multiple variables**. Let $X$ and $Y$ be random variables over some sets $\mathcal{X}$ and $\mathcal{Y}$. We define

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

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 marginal probability of $X = x$ is given by

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

which is the probability of seeing $X = x$ irrespective of the value $Y$ takes on, and the conditional probability is given by

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

which is the probability of observing $X = x$ if we know that $Y = y$.

**Independent random variables**. Independent random variables play a very important role in probability, because they greatly simplify calculations. Two RVs $X$ and $Y$ are considered independent 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 $X = x$ and $Y = y$. An important implication of independence is that

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

In words this says that knowing the value of $Y$ tells us no new information about what value $X$ may take on. A particularly important sub-class of independent RVs are the independent and identically distributed (i.i.d.) RVs; $X_1$ and $X_2$ are i.i.d. if they are independent and also if

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

for all values of $x \in \mathcal{X}$. In words this says the two RVs have exactly the same marginal distribution over $\mathcal{X}$, and are independent.

**Probability density functions**. If the set of values $\mathcal{X}$ that a random variable $X$ can take is continuous (i.e., real numbers), then we say that $X$ follows a probability density function (pdf). A pdf satisfies

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

and

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

The term "density function" is used because it describes how densely the probability is distributed across the set $\mathcal{X}$. To find the probability of an interval such as $X \in (a, b)$ we integrate the pdf from $a$ to $b$:

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

One of the more confusing properties of continuous variables is that the probability of $X$ taking on any specific, exact real number is zero. Why is this? Consider the probability of $X \in (x_0 - \delta/2, x_0 + \delta/2)$, where $\delta > 0$ is the width of the region around a value $x_0$. Then, we can approximate the integral as the product of the width of the region $\delta$ times the height of the pdf at the point $p(x_0)$:

$$\begin{aligned} \mathbb{P}(x_0 - \delta/2 < X < x_0 + \delta/2) &= \int_{x_0-\delta/2}^{x_0+\delta/2} p(x)dx \\ &\approx p(x_0)\delta \end{aligned}$$

From this it is clear that the smaller the interval $\delta$ around a point $x_0$, the smaller the probability; taking $\delta \to 0$ clearly shows that $\mathbb{P}(X = x_0) = 0$.

**Cumulative distribution functions**. The cumulative distribution function (cdf) plays an important role in probability theory. Let us begin by introducing some shorthand notation for discrete RVs; namely, that

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

where $a \equiv b$ is read as "$a$ is equivalent to $b$". The cdf of a discrete random variable over the integers is then

$$\mathbb{P}(X \leq x) = \sum_{x' \leq x} p(x')$$

which can be interpreted as the probability of $X$ taking on any value less than or equal to $x$. For a continuous RV the cdf is

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

From the fact that the total probability of a distribution is one, we have the following useful properties:

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

With these two rules the cdf can be used to obtain probabilities over any interval or combination of intervals.

The inverse of the cdf is called the quantile function $Q(p)$. This 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 $p$. Formally we write

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

which we read 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 uses is to define quantities such as the median, $Q(p = 1/2)$, or the first and third quartiles, $Q(p = 1/4)$ and $Q(p = 3/4)$.

**Expectation of a random variable**. The expectation of a discrete random variable $X$ is

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

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

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

Both expectations are 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.

**Expectations of functions of random variables**. More generally we can find the expectation of a function of a random variable as

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

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,

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

that is, the expectation of a function is not equal to the function of the expectation; an obvious exception is when $f(X) = X$ (i.e., the identity function).

**Expectation over multiple RVs**. Expectation over the RVs $X$ and $Y$ is defined by

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

with the obvious extension to continous RVs. 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. We can use this notation to show that if we define the random variable

$$X_f = f(X)$$

so that $X_f$ is a RV that is a function $f(\cdot)$ of the RV $X$, then we see that $\mathbb{E}_{X_f}\left[X_f\right] = \mathbb{E}_X\left[f(X)\right]$, and all the rules that apply to expectations of RVs immediately extend to functions of RVs.

**Linearity of expectation**. Expectations are so-called linear operators, which implies a lot of properties. For example,

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

where $c$ is any constant with respect to $X$, which implies

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

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

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

so that the expectation of a sum of two functions of random variables is the sum of the expectations.

**Variances**. The variance is the expected squared-deviation of the RV $X$ around its mean:

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

which extends in a straightforward fashion to functions:

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

The variance is also given by the alternative formula

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

that is, the variance of $X$ is equal to the expected squared value of $X$, minus the square of the expected value of $X$. The variance measures the average squared deviation of a random variable from its mean

$\mathbb{E}[X]$. The larger the variance, the more variability in the values taken on by $X$. By the linearity of expecation, the variance has the following useful property

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

so that the variance of $X$ times a constant $c$ is equal to the variance of $X$ times the square of $c$. An important related quantity is the standard deviation which is $\sqrt{\mathbb{V}[X]}$, and which measures the average deviation of a random variable $X$ from its mean $\mathbb{E}[X]$. An important property of the standard deviation is that it has the same units of measurement as the RV $X$; that is, if $X$ is measured in meters, so is its standard deviation.

**Covariances**. If we have two random variables $X$ and $Y$ we can define the covariance between them by:

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

and from this, we can define the correlation

$$\text{corr}(X,Y) = \frac{\text{cov}(X,Y)}{\sqrt{\mathbb{V}[X]\mathbb{V}[Y]}}.$$

Covariance depends on the units of $X$ and $Y$, while the division by the standard deviations in the definition of correlation removes the unit of measurement from the coefficient correlation ("standardises" it) so that the value always lies between -1 (perfect negative correlation) and 1 (perfect positive correlation), with a correlation of zero denoting no correlation. 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 we expect the corresponding value $Y$ will be more likely to be greater than its expected value $\mathbb{E}[Y]$.
- Conversely, negative correlation implies that if a $X$ is greater than its expected value $\mathbb{E}[X]$, then we expect the corresponding value $Y$ will be more likely to be smaller than its expected value $\mathbb{E}[Y]$.

If two RVs $X$ and $Y$ are independent, then $\text{cov}(X,Y) = 0$. However, a covariance of zero does not imply the opposite – as covariance/correlation measure linear dependence it is very possible to have zero/low correlation if the two variables are highly dependent in a nonlinear fashion.

**Expectations and Independent RVs.** 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 satisfy:

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

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 that is very useful is that if $\mathbb{E}_X[X] = 0$, or $\mathbb{E}_Y[Y] = 0$, then the expectation of the product will just be zero. If $X$ and $Y$ are independent, then the variance satisfies :

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

so that the variance of a sum of (functions of) the RVs $X$ and $Y$ is equal to the sum of the variances. This is a *very useful* property. Using the above property, and the property $\mathbb{V}[cX] = c^2\,\mathbb{V}[X]$ we can show that

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

by writing

$$\mathbb{V}[X - Y] = \mathbb{V}[X + (-1)Y] = \mathbb{V}[X] + (-1)^2\mathbb{V}[Y].$$

This result tells us that the variance of the difference of two RVs is equal to the *sum* of the variances of the RVs. 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 increasing the variability above the variability present in either of the RVs individually.

# 2 Part II: Statistical Models as Probability Distributions

**Parametric probability distributions**. In practice the distributions we work with are defined over a very large, or more commonly, infinite number of events. That is, the set $\mathcal{X}$ of values our RV can take is infinitely large. This means that direct specification of the probabilities of each event is impossible. To overcome this we use parametric probability distributions. What this means is that we specify the assignment of probabilities to different values in $\mathcal{X}$ using a function

$$p(x \,|\, \boldsymbol{\theta}),\ x \in \mathcal{X},\ \boldsymbol{\theta} \in \Theta$$

where we treat $p(x \,|\, \boldsymbol{\theta})$ as a probability density if the RV is continuous. The above should be thought of as a probability distribution controlled by the particular values of $\boldsymbol{\theta}$, which we call the parameters of the distribution. By changing or varying those parameters, we can produce a whole different range ("family") of probability distributions. In general, the number of parameters is quite small – usually no more than two or three – and in general less than the number of different values $x$ the RV $X$ can take on.

This means that the properties of the RV $X$, such as the mean and variance, are implicitly determined by the values of the parameters $\boldsymbol{\theta}$. This is obvious if we consider the expected value of $X$

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

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. The same applies to all the properties of $X$ – variance, cdf, quantiles, etc.

**The Gaussian (normal) distribution**. If we are interested in a distribution over the real (continuous) numbers then we clearly have an infinite number of different values our RV $X$ can take. In this case, one of the most common and important distributions is the Gaussian distribution. This is named after the German mathematician Carl Friedrich Gauss (1777–1855). This distribution is also often called the normal distribution, for reasons that will become clear in the next lecture. The Gaussian distribution is characterised by two parameters: $\mu$, which controls the mean of the distribution, and $\sigma^2$ which controls the variance of the distribution. The pdf for a Gaussian distribution is

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

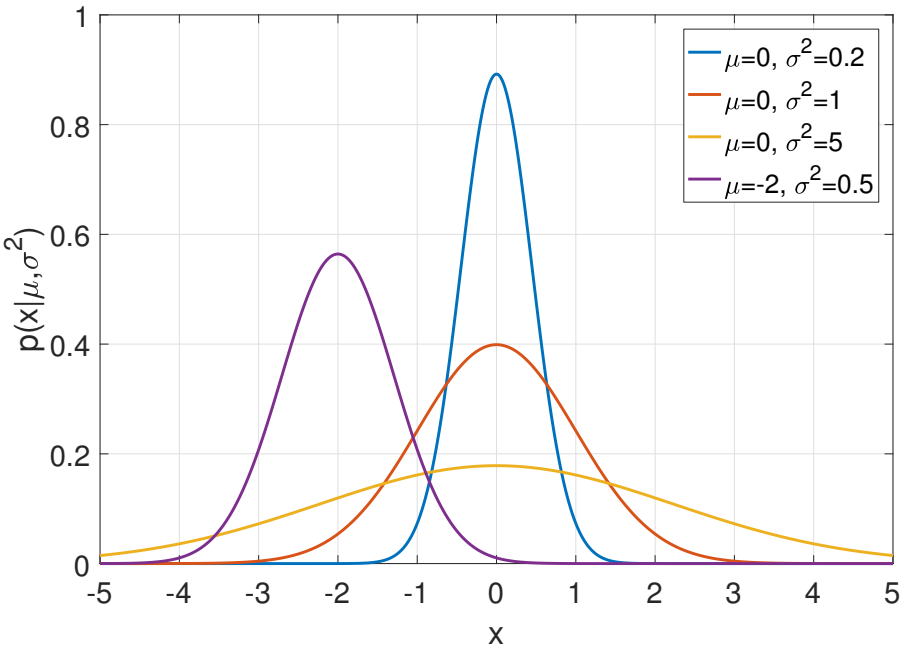


Figure 1: Probability density functions for several normal (Gaussian) distributions. The orange curve is the *standard normal distribution*. Note that the normal distribution is symmetric and tails off to zero as $|x| \to \infty$.

The formula (1) looks complicated, but at its core is quite straightforward. The first part is simply the normalising constant, which ensures that the distribution integrates to one, as required, for all values of $\mu$ and $\sigma^2$. The main component is the second part, which says that the probability of a value $x$ gets smaller at an exponential rate the further the value is away from the mean $\mu$. The division by $\sigma^2$ controls the rate at which the decay occurs; the bigger the value of $\sigma^2$, 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 bell shape that is symmetric around the mean $\mu$ and tails off to zero as $|x| \to \infty$. You could reproduce this graph by plugging in the appropriate values for $\mu$ and $\sigma^2$ in (1) and plotting the function from $x = -5$ to $x = 5$.

In statistics, we use some notation to describe when a RV is distributed following one of the common parametric distributions. For example, we say that if $X$ follows a normal then

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

where the " $\sim$" is read as "is distributed as per a". We noted above that normal distributions all look very similar – in fact, they satisfy a very important self similarity property; namely, that every normal distribution is a scaled and shifted (translated) version of the so called standard unit normal $N(0, 1)$. What this means is that if $Z \sim N(0, 1)$, then

$$X = \sigma Z + \mu$$

is distributed as per a $N(\mu, \sigma^2)$. So by taking a random variable that follows a unit normal, we can turn it into a RV that follows any normal by multiplying it by the desired standard deviation, and then adding the mean. Conversely, we can convert a RV $X \sim N(\mu, \sigma^2)$ to a RV following a unit normal by

reversing this transformation. This technique is used extensively in statistics. If $X \sim N(\mu, \sigma^2)$, then

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

which says that in the case of the normal distribution, the mean is equal to the parameter $\mu$ and the variance is equal to the parameter $\sigma^2$. Therefore, by varying $\mu$ and $\sigma^2$, we can make a distribution with any particular combination of mean and variance we desire. The Gaussian distribution is symmetric around the $\mu$; this implies that the mode (most common value, i.e., the value $x$ that maximises (1)) as well as the median are all equal to $\mu$. The cumulative distribution function (cdf) for the normal distribution does not actually have a closed form – that is, there is no simple formula you can write down for the cdf. However, all standard packages, such as R have efficient implementations of the cdf, and there are several well known rules that are important; namely, for any $N(\mu, \sigma^2)$

1. 68.27% of probability falls within $(\mu - \sigma, \mu + \sigma)$, irrespective of the values of $\mu$ and $\sigma$. That is, 68.27% of samples from a normal distribution will fall within one standard deviation of the mean.
2. 95% of probability falls within $(\mu - 1.96\sigma, \mu + 1.96\sigma)$;
3. 95.45% of probability falls within $(\mu - 2\sigma, \mu + 2\sigma)$; and
4. 99.73% of probability falls within $(\mu - 3\sigma, \mu + 3\sigma)$.

The R function associated with the normal distribution are `dnorm()` (the pdf function), `pnorm()` (the cdf function), `qnorm()` (the quantile function) and `rnorm`() (a function to generate random realisations from a normal distribution).

**The Bernoulli distribution**. Let us now consider the case of a discrete, binary RV $X$; that is, a RV that can only take on one of two different values, say 0 or 1, i.e., $\mathcal{X} = \{0, 1\}$. 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. We can use the Bernoulli distribution, named after the Swiss scientist Jacob Bernoulli (1655–1705), to model these types of random variables. The Bernoulli distribution says that the probability that $X = 1$ is

$$\mathbb{P}(X = 1 \,|\, \theta) = \theta, \theta \in [0, 1],$$

that is, the probability of $X = 1$ is just the value of the parameter $\theta$. This lets us write the parametric probability distribution over $x \in \{0, 1\}$ as

$$p(x \,|\, \theta) = \theta^x (1 - \theta)^{(1-x)}. \tag{2}$$

Try plugging in the values $x = 0$ and $x = 1$ into (2), and you will see that $p(x = 0 \,|\, \theta) = (1 - \theta)$ and $p(x = 1 \,|\, \theta) = \theta$. We often say that $\theta$ is the probability of observing a "success" – for example, a heads coming up in a coin toss. If a RV $X$ follows a Bernoulli distribution with success probability $\theta$, we write that $X \sim \text{Be}(\theta)$. By using our formulas for expectation and variance, it is very easy to calculate that

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

We see that both expectation and variance depend on the single parameter $\theta$. The variance function for the Bernoulli is very interesting, as it approaches zero as $\theta$ gets closer to 0 or 1, and is largest when $\theta = 1/2$. This is because the larger (smaller) the success probability, the less chance of a seeing

a failure (success) in a series of realisations of the Bernoulli random variable, and therefore the less variability in the random variable. In the extreme case of $\theta = 0$, we will never see a success and there is no variability, with a similar conclusion for when $\theta = 1$.

**The binomial distribution**. If we have a series of realisations of binary random variables, and we want to model the number of successes that occur in a number of binary trials, we can use the binomial distribution. For example, if we saw the sequence of binary variables

$$\mathbf{x} = (0, 1, 1, 1, 0, 1, 0, 0, 1, 1)$$

then the number of successes $m$ is equal to six (i.e., there are six 1s in the sequence). The probability of seeing $M = m$ successes in $n$ trials is given by

$$p(m \mid \theta) = \binom{n}{m} \theta^m (1 - \theta)^{(n-m)} \tag{3}$$

where

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

is the number of ways of choosing $m$ objects out of a total of $n$ identical objects, and is called the binomial coefficient, from which the binomial distribution gets its name. The $m! = 1 \times 2 \times 3 \times \cdots \times m$ denotes the factorial function. The binomial distribution has two parameters: $\theta$, 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 $\{0, 1, 2, \ldots, n\}$.

The $\theta^m(1 - \theta)^{(n-m)}$ term in (3) is the probability of observing any sequence of $n$ binary variables with exactly $m$ successes. The $\binom{n}{m}$ component in (3) accounts for the fact that if $0 < m < n$, there is more than one sequence of $n$ 0s and 1s that has exactly $m$ 1s in it. As we are interested only in the number of 1s in a sequence, all sequences with $m$ 1s in them, regardless of the configuration of the sequence, are equivalent. The $\binom{n}{m}$ term counts the number of equivalent sequences. For example, if we have $n = 4$ trials, the number of sequences that have exactly $m = 2$ 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. If we evaluate $\binom{4}{2}$ we see that it is equal to six. If a RV $M$ follows a binomial distribution with success probability $\theta$ and number of trials $n$, we say that $M \sim \text{Bin}(\theta, n)$. The number of 1s in a sequence $\mathbf{y}$ can be written as a sum:

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

Therefore, we see that the binomial *RV* $M$ is defined as a sum of independent Bernoulli RVs, and we can use the properties of expectations and variances of sums of RVs to easily find

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

that is, the mean and variance of $M$ is just $n$ times the mean and variance of a single Bernoulli trial with success probability $\theta$. 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

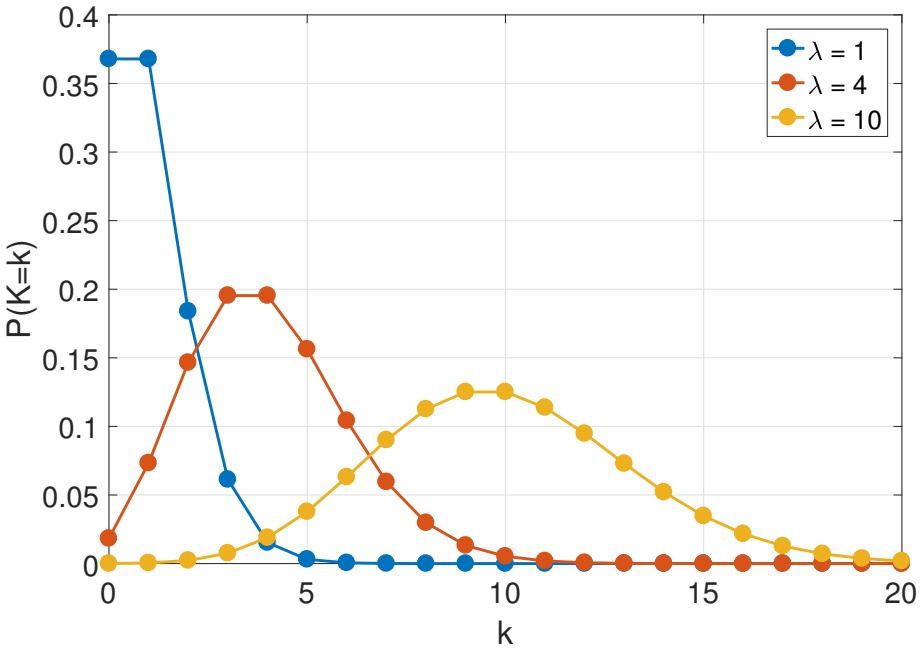


Figure 2: Poisson distribution for $\lambda = 1$, $\lambda = 4$ and $\lambda = 10$. The distribution is defined only on the integers – the connecting lines are only guides for the eye.

note, you can generate a RV from a Bernoulli distribution by using a binomial distribution with $n = 1$.

**The Poisson distribution**. Frequently we are interested in modelling counts of occurences 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. In this case the set of values our RV can take on is the set of non-negative integers, i.e., $\mathcal{X} = \{0, 1, 2, 3, \ldots\}$. A suitable distribution for these types of random variables is the Poisson distribution (named after the French scientist Simeon Poisson (1781–1840). This has the probability distribution

$$p(k \,|\, \lambda) = \frac{\lambda^k \exp(-k)}{k!} \tag{4}$$

where again we recall $k!$ denotes the factorial function. The Poisson has a single parameter, $\lambda > 0$, which we call the rate. If $X$ is distributed as per a Poisson distribution with rate $\lambda$, we write $X \sim \text{Poi}(\lambda)$. The mean and variance of a RV $X$ following a Poisson distribution is:

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

The Poisson distribution is an example of distributions in which the the variance grows with the mean; in fact, for the Poisson distribution the mean and variance are the same. This says that as the rate $\lambda$ grows, the average number of events increases, and the average spread around the mean also increases. The Poisson distribution, for several values of $\lambda$, is shown in Figure 2. Note that as $\lambda$ gets bigger, the distribution becomes more spread out around $k = \lambda$, showing that the variance is increasing. The Poisson is only one distribution for modelling counts, and it is therefore necessary to identify when it is appropriate for use. The following rules of thumb identify when a Poisson distribution will be (at least approximately) appropriate (taken from Wikipedia):

1. The occurence of one event does not affect the probability that a second event will occur; i.e., the events occur independently of each other. For example, telephone calls to a call centre are unlikely to influence each other.

2. The rate at which the events occur is constant – the rate cannot be higher in some intervals and lower in others. For example, the telephone call centre example may not exactly meet this criterion if calls are more frequently made during the weekends than on weekdays.

3. Two events cannot occur at exactly the same time instance.

4. The probability of an event in a small interval of time is proportional to the length of the interval; for our call centre example, this seems reasonable as the shorter a time interval we consider, the less chance of receiving a call in the time interval.

Even if all the conditions are not precisely met, the Poisson can still be a decent model. For example, our call centre example may violate condition 2, but if the difference between weekdays and weekends is not large, the Poisson may still do an adequate job of modelling our data. 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).

**The Weak Law of Large Numbers**. The last topic of importance in Lecture 2 is the so called Weak Law of Large Numbers (WLLN). This is a mathematical statement of something that probably seems intuitive to you already – that is, if you repeatedly toss a fair coin (probability of a head is 0.5), then for very large repetitions of the experiment, the proportion of heads you will have seen out of the total number of times you tossed the coin will get closer and closer to 0.5. Formally, the WLLN says that: if $X_1, \ldots, X_n$ are RVs with finite mean $\mathbb{E}[X_i] = \mu$, then for any small number $\varepsilon > 0$,

$$\mathbb{P}\left\{\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| > \varepsilon\right\} \to 0 \text{ as } n \to \infty.$$

This says that the probability that the sample mean $(X_1 + X_2 + \cdots + X_n)/n$ differs from the true mean $\mu$ by more than a distance of $\varepsilon$ goes to zero as the number of realisations $n$ grows to infinite, irrespective of how small we make $\varepsilon$. Informally, you can think of this result as saying that the (sample) mean of a realisation of random variables converges to the expected value as the number of realisations grows larger and larger. A useful outcome of this result is that we can use simulation to calculate the probabilities of events occuring (see Studio 2). We do this by noting that a probability is equivalent to the expected value of a Bernoulli variable, so that we can write any probability statement as an expectation using:

$$\mathbb{P}(X < x_0) = \mathbb{E}\left[I(X < x_0)\right],$$

where $I(\cdot)$ is the indicator function that takes on a value zero if the condition inside the function is not met, and a one if the condition is met. To see that the above equivalence holds:

$$\begin{aligned}
\mathbb{P}(X < x_0) &= \mathbb{E}\left[I(X < x_0)\right] \\
&= \int_{-\infty}^{x_0} 1 \cdot p(x)dx + \int_{x_0}^{\infty} 0 \cdot p(x) \\
&= \int_{-\infty}^{x_0} p(x)dx.
\end{aligned}$$

For example, we could estimate the probability that $\sqrt{|X|} > 1$ when $X \sim N(0,1)$ using R with the following code

```
x = rnorm(ns, 0, 1)
mean(sqrt(abs(x)) > 1)
```

This works because R evaluates conditional statements like `sqrt(abs(x))` to be a zero if the condition is false, and be a one if the condition is true. The `mean()` function then evaluates the proportion of times the statement evaluates to true, which in turn is an estimate of the probability that the statement is true. By making the value `ns` in the above code larger and larger, we can get a better and better estimate of the probability we are interested in, and this is guaranteed by the WLLN as described above.