Chapter 4
Random Variables and Distributions

4.1 Introduction

In the first chapter we discussed the calculation of some statistics that could be useful to summarize the observed data. In Chap. 2 we explained sampling approaches for the proper collection of data from populations. We demonstrated, using the appropriate statistics, how we may extend our conclusions beyond our sample to our population. Probability sampling required reasoning with probabilities, and we provided a more detailed description of this topic in Chap. 3. The topic of probability seems distant from the type of data that we looked at in the first chapter, but we did show how probability is related to measures of effect size for binary data. We will continue discussing real-world data in this chapter, but to do so we will need to make one more theoretical step. We will need to go from distinct events to dealing with more abstract *random variables*. This allows us to extend our theory on probability to other types of data without restricting it to specific events (i.e., binary data).

Thus, this chapter will introduce random variables so that we can talk about continuous and discrete data. Random variables are directly related to the data that we collect from the population; a relationship we explore in depth. Subsequently we will discuss the *distributions of random variables*. Distributions relate probabilities to outcomes of random variables. We will show that distributions may be considered "models" for describing variables from populations. We will discuss separately distributions for *discrete* random variables and for *continuous* random variables. In each case we will introduce several well-known distributions. In both cases we will also discuss properties of the random variables: we will explain their *expected value*, *variance*, and *moments*. These properties provide summaries of the population. They are closely related to the mean, variance, skewness, and kurtosis we discussed in Chaps. 1 and 2. However, we will only finish our circle—from data to theory to data—in the next chapter.

M. Kaptein and E. van den Heuvel, *Statistics for Data Scientists*, Undergraduate Topics in Computer Science, https://doi.org/10.1007/978-3-030-10531-0_4

In this chapter we will discuss:

- Populations and density functions
- The definition of random variables and probability distributions
- Probability distributions for continuous random variables
- Probability distributions for discrete random variables
- Formal definitions of means, variances, standard deviations, and other moments
- Examples of parametric probability distributions (Bernoulli, binomial, Poisson, normal, lognormal, uniform, and exponential)
- Using R to work with probability distributions.

4.2 Probability Density Functions

In Chap. 1 we introduced the histogram to visualize our data and we gave an example of a density plot, or in other words a *density function* (see Fig. 1.10). The density function may be viewed as a smooth version of the histogram if we standardize the frequencies on the vertical axis to proportions. It may be viewed as an approximation of the histogram on all units from the population if the population is very large (say million's and million's of units). The density function characterizes the occurrence of values for a specific variable (as depicted on the x-axis) on all units from the population. Since in practice all populations are finite, the density function is an abstract formulation of, or a "model" for, the "frequencies" of all population values. In statistics the density function is typically denoted by the small letter f and it is typically referred to as *probability density function* (PDF).

Since we have assumed that the PDF f is some kind of smooth approximation of the histogram, the PDF must satisfy two important conditions or properties. The first condition is that it cannot be negative, i.e., $f(x) \geq 0$ for every value x that is present in the population. Clearly, we cannot observe a negative frequency or proportion in histograms. We often extend the domain of this PDF to the whole real line $\mathbb{R}$, even though the values from the population may be restricted to a smaller domain. For values of x outside this domain, the PDF can then be defined equal to zero: $f(x) = 0$. For instance, measuring the amount of hours per week that school children watch television ranges theoretically from 0 to 168 hours. A PDF f would then be considered equal to zero for any negative value of x and for values larger than 168 hours (and possibly also for values inside this interval, but that depends on the behavior of all children in the population). The second condition for a PDF is that we assume that the "area under the curve" is equal to one, i.e.,

$$\int_{\mathbb{R}} f(x)dx = 1.$$

This essentially means that 100% of all unit values together form the population. If we were able to observe all values from the population we must have considered

or obtained all units from the population. This property makes it possible to relate PDFs to proportions of units in the population (or, as we will see later, to probability statements), as we have already indicated in Chap. 1. For instance, the proportion of school children that watches television for less than or equal to 2 hours per week can now be written as

$$0 \leq \int_{-\infty}^{2} f(x)dx = \int_{0}^{2} f(x)dx \leq \int_{\mathbb{R}} f(x)dx = 1.$$

Indeed, if all children watch television for less than two hours per week, then the integral on the left side would be equal to 1, since watching for less than two hours per week still represents the whole population of school children, but if all school children watch television for more than two hours per week, the integral would be equal to 0, since no child will watch less then two hours per week. In practice the proportion will be somewhere in between 0 and 1, since there will be children who hardly watch any television and those who watch a lot. Thus the integral indicates what proportion of school-children watch television for less than or equal to two hours a week. By studying these proportions (or integrals) for any type of interval, say $[a, b] \subset \mathbb{R}$, we would know or be able to retrieve the shape of the PDF f, or in other words, we would know exactly what proportion of the population would have what set of values. We will discuss this later in more detail.

Many different PDFs exist and they have been proposed over a period of more than two centuries to be able to describe populations and data in practical settings. These functions are often parametric functions, i.e., the PDF is known up to a set of parameters. The PDF is then often denoted by $f_{\boldsymbol{\theta}}$, where $\boldsymbol{\theta}$ represents the set or vector of m density parameters: $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)^T$.[1] Many books have been written on PDFs, so it would go too far to provide a long list here, but we do want to provide information about the *normal*, *log normal*, *uniform*, and *exponential* PDFs to give a flavor of the differences.

4.2.1 Normal Density Function

The normal PDF is very special within statistics, both for theoretical and for practical reasons. We will learn in Chaps. 4, 5, and 6 that it can be used to approximate other PDFs when the sample size or the size of the data is getting large. This has the advantage that important features of the normal density function can be transferred to other densities when the approximation is quite close. Beside these theoretical aspects, the normal PDF has been used often to analyze all kinds of datasets and it is

[1] Although the subscript notation that we introduce here is often used, in some contexts the notation $f(\cdot|\boldsymbol{\theta})$ is preferred to make explicit that the distribution function is *conditional* on the parameters (for example in Chap. 8 of this book). In the current chapter we will, however, use the subscript notation.

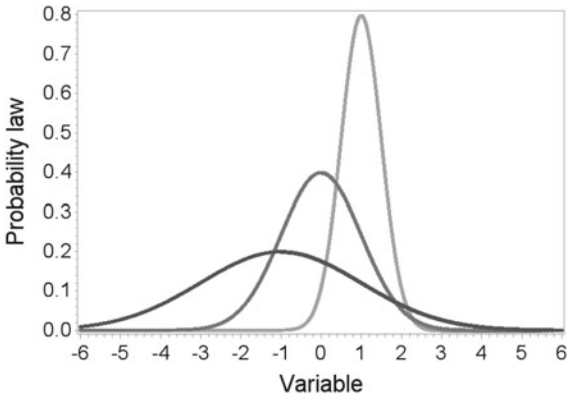

Fig. 4.1 Three normal density curves for different choices of the parameters

an underlying assumption of several of the modeling approaches that is outside the scope of this book.

The normal PDF was probably first introduced explicitly as a PDF by Carl Friedrich Gauss, and it is therefore often referred to as the *Gauss curve*. He used the normal PDF to describe random errors in measuring orbits (Sheynin 1979), in particular for the calculation of the orbit of the dwarf planet Ceres. In that period the topic was referred to as the "theory of errors". It was an important research area to determine how to deal with measurement errors in calculations. Today the normal PDF is still frequently used for describing data, since many types of measurements, like physical dimensions, are often properly described by the normal PDF.

The normal PDF has just two parameters: μ and σ.[2] The parameter μ indicates the mean value of the population of the variable of interest and the parameter σ indicates the standard deviation. These parameters represent the exact same two population parameters that we discussed in Chap. 2. The shape of the normal PDF is equal to the famous "bell-shape" curve that we have all seen somewhere before. Three different curves are represented in Fig. 4.1.

The normal PDF is mathematically formulated by

$$f_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \tag{4.1}$$

with $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. It is obvious that the normal PDF satisfies the first condition: $f_{\mu,\sigma}(x) > 0$ for all $x \in \mathbb{R}$, but it is not straightforward to show that the integral of this density is equal to one (but it really is).

When we choose $\mu = 0$ and $\sigma = 1$ in Eq. (4.1), we refer to this normal PDF as the *standard normal density function* or *standard normal PDF* and we rather prefer the

[2] Although in general we like to denote the parameters of a PDF with $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)^T$, for many specific PDFs other notation is used. For the normal PDF we should have used $\boldsymbol{\theta} = (\theta_1, \theta_2)^T$, with $\theta_1 = \mu$ and $\theta_2 = \sigma$, but μ and σ are more common in the literature.

notation ϕ instead of $f_{0,1}$, i.e., $\phi(x) = \exp\{-x^2/2\}/\sqrt{2\pi}$. This means that the normal PDF $f_{\mu,\sigma}(x)$ in Eq. (4.1) can now also be written as $f_{\mu,\sigma}(x) = \phi((x-\mu)/\sigma)/\sigma$.

Some well-known characteristics of the normal PDF are the areas under the curve for specific intervals. For instance, 95.45% of all the population values fall within the interval $[\mu - 2\sigma, \mu + 2\sigma]$, or formulated in terms of the integral:

$$\int_{\mu-2\sigma}^{\mu+2\sigma} \phi((x-\mu)/\sigma)dx = \int_{-2}^{2} \phi(x)dx = 0.9545.$$

Alternatively, 95% of the values fall within $[\mu - 1.96\sigma, \mu + 1.96\sigma]$ and 99.73% of all the population values fall within the interval $[\mu - 3\sigma, \mu + 3\sigma]$.

4.2.1.1 Normally Distributed Measurement Errors

Putting these characteristics in practice we follow the ideas of Gauss on measurement errors. We will assume that the normal PDF can be used to describe random errors in measuring some kind of quantity η (e.g., blood pressure of a human being, the diameter of a planet, the tensile strength of one plastic tube, etc.). We would expect that the population of all possible measurement errors, that we may obtain when we measure the quantity,[3] are on average equal to zero ($\mu = 0$), since it would be as likely to measure higher as lower values than the true value η that we are trying to capture. Based on the shape of the standard normal PDF, it is much more likely to obtain random errors closer to zero than random errors that will be far away from zero. Moreover, approximately 95.45% of all the possible random errors that we may obtain will fall within plus or minus twice the standard deviation away from zero, i.e., $[-2\sigma, +2\sigma]$, and 99.73% will fall within $[-3\sigma, +3\sigma]$. The standard deviation σ is here a measure of the precision of the measurement system.

For instance, the standard deviation of measuring blood pressure with oscillometric devices in human beings is approximately equal to 4.4 and 3.4 mmHg for systolic and diastolic blood pressure, respectively (Liu et al. 2015). Thus 95.45% of the random systolic blood pressure errors fall within $[-8.8, 8.8]$ mmHg and 99.73% will fall within $[-13.2, 13.2]$ mmHg. Thus if we measure a person with a systolic blood pressure $\eta = 120$ mmHg, our blood pressure reading will fall within 111.2 mmHg and 128.8 mmHg with 95.45% certainty and within 106.8 mmHg and 133.2 mmHg with 99.73% certainty. Something similar can be determined for diastolic blood pressure.

[3] Here we assume the existence of an infinite population of measurement errors having the normal PDF from which one error e is randomly sampled when we conduct one measurement of the quantity. This error is then added to the true value η to obtain a measurement $x = \eta + e$ of the quantity or a reading of the unit.

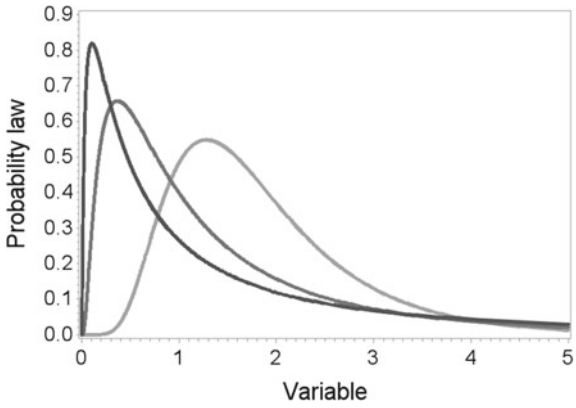

Fig. 4.2 Three lognormal population densities for different values of its parameters

4.2.2 Lognormal Density Function

Although the origin of the lognormal PDF comes from a more theoretical setting, the log normal PDF became very popular at the beginning of the 20th century, when the log normal PDF was being used for biological data (Kotz et al. 2004). It has been used in many different applications, ranging from agriculture to economics and from biology to the physical sciences. The lognormal PDF has some very nice properties, which makes it useful in many applications. As we will see, the lognormal PDF describes populations with positive values. This would make more sense than the normal PDF, which describes both positive and negative values, when quantities like particle size, economic growth, duration of games and activities, and measures of size are being studied. Furthermore, the relative standard deviation, which was formulated in Chap. 1, is constant for the lognormal PDF. This means that larger values demonstrate larger variability, but the ratio with variability is constant whether we observe smaller or larger values. Finally, the lognormal PDF is not symmetric like the normal PDF (see Fig. 4.2), which makes sense when values are limited from below, but not from above.

On the other hand, the lognormal PDF is closely related to the normal PDF. If the population values can be described by a lognormal PDF, the normal PDF would then describe the logarithmic transformation of the population values (using the natural logarithm). Thus, the relationship between the normal and lognormal PDFs is based on a log transformation. In practice, we often make use of this transformation, so that we can borrow the normal PDF characteristics in the log scale and then transform the results back to the original scale (using the inverse of the logarithm: $\exp\{x\}$).

The mathematical formulation of the lognormal density is given by

$$f_{\mu,\sigma}(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{-\frac{(\log(x)-\mu)^2}{2\sigma^2}\right\}, \tag{4.2}$$

with $x > 0$, $\log(x)$ the natural logarithm, $\mu \in \mathbb{R}$, and $\sigma^2 > 0$.[4] This PDF has been formulated only on positive values $x > 0$, while we have indicated that the domain of PDFs is typically formulated on the whole real line $\mathbb{R}$. For the part that is left out ($x \leq 0$), the density is then automatically assumed equal to zero. Thus the lognormal density is equal to zero ($f_{\mu,\sigma}(x) = 0$) for values of $x \leq 0$. In Fig. 4.2 a few examples of the log normal PDF are visualized.

Thus the lognormal PDF is also always non-negative for all values of $x \in \mathbb{R}$. Knowing that the integral of the normal PDF is equal to one, helps us to demonstrate that the integral of the lognormal PDF is also equal to one:

$$\int_0^\infty \frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{-\frac{(\log(x)-\mu)^2}{2\sigma^2}\right\} dx = \int_{-\infty}^\infty \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx = 1$$

Thus the lognormal density function also satisfies the two criteria for a PDF.

The parameters μ and σ have a different meaning in the lognormal PDF than in the normal PDF. They do represent the population mean and standard deviation, but only for the logarithmic transformed values of the population. Their meaning in relation to the mean and standard deviation of the population values in the original scale is now more complicated. The population mean and standard deviation in the original scale are now functions of both μ and σ. They are given by $\exp\{\mu + \sigma^2/2\}$ and $(\exp\{\sigma^2\} - 1)\exp\{2\mu + \sigma^2\}$, for the mean and standard deviation, respectively, see the additional material at the end of this chapter. The relative standard deviation (i.e., the standard deviation divided by the mean) is now a function of the parameter σ only: $\sqrt{\exp\{\sigma^2\} - 1}$. This property makes the lognormal density very useful for chemical measurements, where it is often assumed that the measurement error is a fixed percentage of the value that is being measured, i.e., they have a constant relative standard deviation. Indeed, in chemical analysis, the measurement error for higher concentrations is larger than for lower concentrations.

4.2.3 Uniform Density Function

We saw that a random measurement error that could be described by a normal PDF is more likely to be closer to zero than to be further away from zero (due to the bell shape of the density). For a uniform PDF this is different. If the random measurement error would be described by a uniform PDF, being close to or far away form zero would be equally likely. However, the uniform PDF has a finite domain, which means that the density is positive on an interval, say $[\theta_0, \theta_1]$, with $\theta_0 < \theta_1$ and $\theta_0, \theta_1 \in \mathbb{R}$, but zero everywhere else.

[4] Note that we use the same notation $f_{\mu,\sigma}$ for the normal PDF and lognormal PDF. This does not mean that the normal and lognormal PDFs are equal, but we did not want to use a different letter every time we introduce a new PDF. We believe that this does not lead to confusion, since we always mention which PDF we refer to.

The mathematical formulation of the uniform PDF is therefore given by

$$f_{\theta_0,\theta_1}(x) = \frac{1}{\theta_1 - \theta_0}, \quad x \in [\theta_0, \theta_1]. \tag{4.3}$$

It is obvious that the density is non-negative on the real line $\mathbb{R}$, since it is positive on $[\theta_0, \theta_1]$ and zero everywhere else. Furthermore, the area under the curve is equal to one, which can easily be determined using standard integral calculations:

$$\int_{\mathbb{R}} f_{\theta_0,\theta_1}(x)\, dx = \int_{\theta_0}^{\theta_1} \frac{1}{\theta_1 - \theta_0} dx = \frac{\theta_1 - \theta_0}{\theta_1 - \theta_0} = 1.$$

Thus the uniform PDF satisfies the two conditions for a PDF. The *standard uniform density* is given by the density in Eq. (4.3) with $\theta_0 = 0$ and $\theta_1 = 1$.

As can be seen from the density function in Eq. (4.3), the uniform PDF has two parameters (θ_0 and θ_1), similar to the normal and lognormal PDF, but the parameters θ_0 and θ_1 have a truly different interpretation. They indicate the lowest and highest values present in the population, or in other words, they represent the minimum and maximum values in the population. The mean and standard deviation for a population that is described by a uniform PDF are equal to $(\theta_0 + \theta_1)/2$ and $(\theta_1 - \theta_0)/\sqrt{12}$, respectively.

As an example, consider that the random measurement error for systolic blood pressure follows a uniform density symmetric around zero and has a population standard deviation of 4.4 mmHg. In this case, the parameters θ_0 and θ_1 would be equal to -7.62 and 7.62, respectively. A systolic blood pressure reading for a person with an actual systolic blood pressure of 120 mmHg would fall within the interval [112.38, 127.62] mmHg with 100% confidence and any value would be as likely as any other value in this interval. Although the uniform PDF is probably no longer used for measurement errors, it was suggested as possible PDFs for the theory of error before the normal density for measurement errors was introduced (Sheynin 1995).

The uniform PDF has also some nice theoretical properties that we will use later in this chapter. We would be able to simulate a population described by any density function through the use of the standard uniform density. If we draw a population using the standard uniform density, we would be able to make a proper transformation of these uniform values such that the transformed values would describe another density. Drawing values according to the uniform density with a computer using a pseudo-random number generator has been discussed in the additional material of Chap. 2.

4.2.4 Exponential Density Function

The exponential PDF has become a very popular density function in practice in the field of reliability, representing the failure times of complex equipment (Barlow

1984). For instance, it may describe the life time of a population of phones that were bought in 2019. Some phones may live for many years, while others may break-down or stop working within months. One important characteristic of the exponential PDF is its lack of memory of failure times. The occurrence of a failure of a machine (like a phone) in, say, week 52, assuming it survives week 51, is the same as the probability that this machine will fail this week (assuming it survived last week). Thus the exponential PDF describes populations with positive values, similar to the lognormal PDF.

The exponential PDF has only one parameter, which is different from the log-normal density, and it is mathematically described for positive x by the following function

$$f_\lambda(x) = \lambda \exp\{-x\lambda\}, \tag{4.4}$$

with $x > 0$ and $\lambda > 0$. For values of x less than or equal to zero, the PDF is equal to zero (as we already mentioned).

It is interesting to note that the exponential PDF is closely related to the double exponential PDF, which had been applied at least a century earlier in relation to the theory of errors (Hald 2008). The double exponential is symmetric, like the normal PDF, but the exponential PDF is skewed to the right like the lognormal PDF (see Fig. 4.3). The double exponential PDF can easily be determined using the exponential PDF. The double exponential PDF, for any value $x \in \mathbb{R}$, is defined by $0.5 f_\lambda(|x|)$, with $|\cdot|$ the absolute function.

Both the exponential and double exponential satisfy the conditions for a PDF. Clearly, they are non-negative on the real line $\mathbb{R}$ and it can be easily shown that the area under the PDF is equal to one. For the exponential PDF we have

$$\int_{\mathbb{R}} f_\lambda(x)dx = \int_0^\infty \lambda \exp\{-\lambda x\} = [-\exp\{-\lambda x\}]_0^\infty = 1.$$

Since the double exponential PDF is half the exponential PDF on positive values of x and half the exponential PDF on the negative values of x, the area under the double exponential PDF is also equal to one.

A few choices of the exponential PDF are visualized in Fig. 4.3. Visualizing the double exponential is just half the exponential PDF together with its mirror image on the left side of $x = 0$.

For the exponential PDF, the parameter λ is related to the population mean, since λ^{-1} represents the population mean. Thus the smaller the value of λ, the larger the population mean. If the life time of a phone (in years) is described by the exponential density with $\lambda = 0.25$, the mean life-time of a phone is then equal to 4 years. The population standard deviation is also equal to λ^{-1} and therefore depends on the mean of the population, but the relative standard deviation is then independent of the parameter and it is equal to 1. For the double exponential PDF, the mean is equal to zero, which makes it an attractive PDF for measurement errors. The standard deviation of the double exponential is equal to $\sqrt{2}/\lambda$. Thus, the larger the value λ, the closer the measurement errors will be.

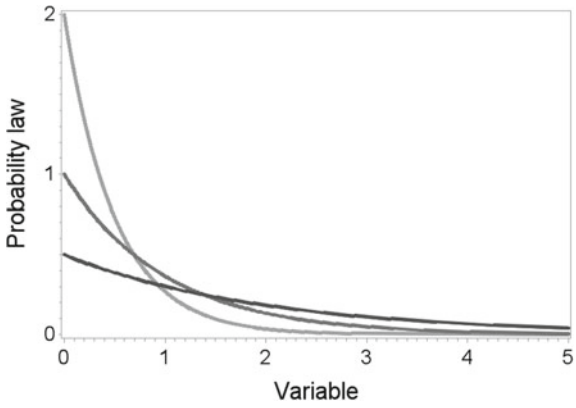

Fig. 4.3 Three exponential density curves

4.3 Distribution Functions and Continuous Random Variables

In the discussions on the theory of errors, PDFs were used to help describe the random measurement errors to be able to come to a proper calculation (often the arithmetic average or median) of the collected observations, like the calculation of the orbit of a celestial body using several measured positions. Whether a calculation of the observations would be better or more precise than just one of these observation was a topic of study.[5] The random errors were considered imperfections of the measurement process (often the human eye) that was trying to capture the true value of interest. These random errors were in that period not always seen as random variables (Sheynin 1995), while the concept of a random variable was already in use long before the theory of errors was discussed. Indeed, random variables were already used when the fundamentals of probability were developed many years earlier.

An intuitive definition of a *random variable* or *random quantity* is a variable for which the value or outcome is unknown and for which the outcome is influenced by some form of random phenomenon. There exists a more formal or mathematical definition, but it is outside the scope of this book.[6] A random variable is typically denoted by a capital letter, say X, Y, or T, to indicate that we do not know the value yet. A *realization* or an outcome of the random variable is then indicated by the same, but lower case, letter x, y, or t. This is in line with our definition of realization in Chap. 2. Indeed, a random variable may be seen as a variable that can in principle

[5] Now we know, with all our knowledge on probability and statistics, that a calculation of the observations like the arithmetic average is in most cases better than just selecting one of them.

[6] There is actually, and perhaps surprisingly, quite an active debate surrounding the definition of a random variable. A definition that is more mathematical but might still be accessible is the following: "A random variable is a mapping (i.e., a function) from events to the real number line". This definition allows us to mathematically link the material in Chap. 3—where we discussed events—to the material presented in this chapter. However, this definition is sometimes perceived as confusing as it does not contain any reference to random processes or outcomes.

be equal to any of the values in the population, and after probability sampling the outcome(s) will become known. Before sampling the outcomes are unknown, so we use capital X, and after sampling the outcomes would become known, so we use x. The probability sampling approach makes the variable of interest a random variable, as the sampling approach is here the random phenomenon. One of the earliest examples of a random variable, which is in a way unrelated to our sampling discussion, is, for instance, the age of death. Indeed, when somebody will die is unknown and in many ways random.

Random variables are very convenient to help quantify particular probabilities. John Graunt published in 1662 a *life table* for people in London. He provided probabilities of mortality at different age groups. For instance, he indicated that from 100 births, 36 of them would not grow older than 6 years of age, only one of them would reach an age of 76 years, and none of them would become 80 years or older (Glass 1950). In terms of mathematics, we can write such mortality probabilities in the form of $\Pr(X \leq x)$, where the Pr indicates probability, X is the random variable for age at death, and x is a specific age of interest.[7] For instance, in terms of the life table of John Graunt: $\Pr(X \leq 6) = 0.36$ represents the probability that a new born person would die before or at the age of 6 years old and it is equal to $0.36 = 36/100$.

The probability function $\Pr(X \leq x)$ is a general concept and can be used for any random variable. The random variable X can be the number of hours per week that a school child watches television and we may ask what is the probability that a school child watches less than or equal to two hours per week: $\Pr(X \leq 2)$. The probability $\Pr(X \leq x)$ is also referred to as the *distribution function* obtained in x and it is denoted by $F(x) = \Pr(X \leq x)$. Thus every random variable X has a distribution function F through $F(x) = \Pr(X \leq x)$, but also every distribution function F has a random variable X, namely the random variable X that makes $\Pr(X \leq x) = F(x)$. Thus the two concepts are directly related to each other and we then typically say that X is distributed according to F, i.e., $X \sim F$.

Each distribution function typically satisfies three conditions:

1. When the value x increases to infinity, the distribution function becomes equal to one, i.e., $\lim_{x\to\infty} F(x) = 1$. In terms of the examples of death and television watching this makes sense. When x is large, say larger than 168, every body has died before this age or watches this number of hours of television per week or less. Thus, in these examples $F(x) = 1$ for any $x > 168$.
2. When the value x decreases to minus infinity the distribution function becomes equal to zero, i.e., $\lim_{x\to-\infty} F(x) = 0$. Again this makes sense for the two examples, because no newborn baby would die before the age of zero nor does anybody watch less than zero hours of television per week. Thus, in these examples $F(x) = 0$ for any value $x < 0$.
3. The distribution function is a non-decreasing function, i.e., $F(x_1) \leq F(x_2)$ when $x_1 \leq x_2$. Indeed, the probability of dying within the age of x_1 cannot be larger

[7] In the analysis of life tables it is much more common to calculate probabilities of surviving after a specific age x, i.e., $\Pr(X > x)$, but this is of course equal to $\Pr(X > x) = 1 - \Pr(X \leq x)$, as we discussed in Chap. 3.

than the probability of dying at a higher age x_2. In theory, it is possible that the probability will stay at the same level between x_1 and x_2, indicating that in the interval $[x_1, x_2]$ no population values exist (e.g., no one would die between the ages of 19 and 20 years, say, or no one watches 6 to 7 hours per week television)

There is a direct relation between distribution functions and densities. If we start with a PDF, we can define a distribution function in the following way:

$$F(x) = \int_{-\infty}^{x} f(z)dz. \tag{4.5}$$

Clearly, this function F is a distribution function. When the value x increases to ∞ we obtain the full area under the density, which is by definition equal to one. The area under the density must become zero when x goes to $-\infty$ (there is no unit in the population with the value $-\infty$. And finally, when x_2 is larger than x_1, the area under the PDF from $-\infty$ up to x_2 is not smaller than the same area under the density up to x_1. The distribution function F is often referred to as the *cumulative density function* (CDF). It also shows that the distribution function F is defined as a function from the real line $\mathbb{R}$ to the interval $[0, 1]$. Note that a PDF now also defines a random variable, since it defines a distribution function and a distribution function defines a random variable.

For each of the PDFs we discussed in the previous subsection there exist an accompanied distribution function. This does not mean that we have a closed form expression of each distribution function, since we do not have this for the normal and lognormal distribution functions. However, for the uniform and exponential distribution function we do have an explicit form.

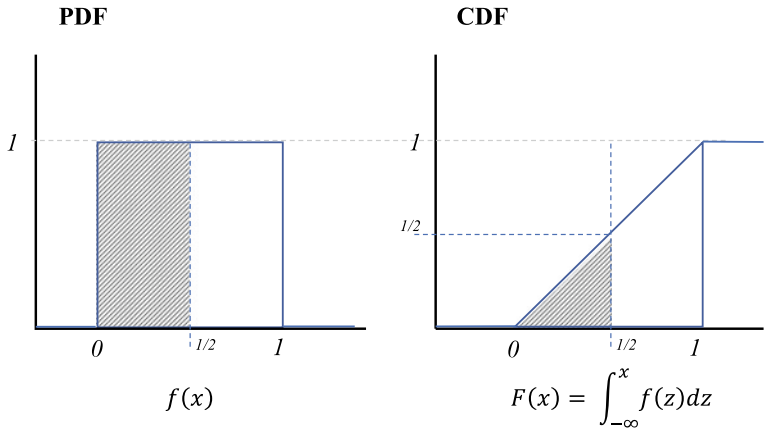

Fig. 4.4 Relationship between the PDF and CDF of a continuous random variable with a uniform distribution function

The uniform distribution function is given by

$$F_{\theta_0,\theta_1}(x) = \begin{cases} 0 & \text{for } x < \theta_0 \\ \frac{x}{\theta_1 - \theta_0} & \text{for } x \in [\theta_0, \theta_1] \\ 1 & \text{for } x > \theta_1 \end{cases}$$

This implies that the standard uniform distribution function is equal to $F_{0,1}(x) = x$ for $x \in [0, 1]$, $F_{0,1}(x) = 0$ for $x < 0$, and $F_{0,1}(x) = 1$ for $x > 1$. The relationship between the standard uniform PDF and CDF is illustrated in Fig. 4.4.

The exponential distribution function is given by

$$F_{\lambda}(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ 1 - \exp\{-\lambda x\} & \text{for } x > 0. \end{cases}$$

The standard exponential distribution function is $F_1(x) = 1 - \exp\{-x\}$ for $x > 0$ and zero everywhere else.

There are a few additional characteristics that we need to mention. First of all, there are theoretical (or exotic) examples where we can define a distribution function without having a density. This has to do with distribution functions that are not differentiable. We do not treat these distribution functions in this book, thus we always assume that there is a PDF that defines the distribution function. Under this assumption, we can obtain that the probability that a random variable X has its outcome in an interval $(x_1, x_2]$ is equal to

$$\begin{aligned} \Pr(X \in (x_1, x_2]) &= \Pr(x_1 < X \leq x_2) \\ &= \Pr(X \leq x_2) - \Pr(X \leq x_1) \\ &= F(x_2) - F(x_1) \\ &= \int_{-\infty}^{x_2} f(z)dz - \int_{-\infty}^{x_1} f(z)dz \\ &= \int_{x_1}^{x_2} f(z)dz. \end{aligned}$$

This equality formalizes our earlier discussion in Sect. 4.2 that the integral from x_1 to x_2 represents the proportion of units in the population having its values in this interval. Note that we have used probability rule $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$ from Chap. 3 for the second equality sign. If we define the two events A and B by $A = (X \leq x_2)$ and $B = (X > x_1)$, we have $A \cap B = (x_1 < X \leq x_2)$ and $A \cup B = (X \in \mathbb{R})$. Thus we obtain now: $1 = \Pr(X \in \mathbb{R}) = \Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B) = \Pr(X \leq x_2) + \Pr(X > x_1) - \Pr(x_1 < X \leq x_2)$. This implies that $\Pr(x_1 < X \leq x_2) = \Pr(X \leq x_2) + \Pr(X > x_1) - 1 = \Pr(X \leq x_2) - \Pr(X \leq x_1)$.

Finally, as a consequence of the second characteristic, the density value $f(x)$ is **not** equal to $\Pr(X = x)$. The probability $\Pr(X = x)$ is equal to zero for continuous random variables, since there is no surface area under $f(x)$. It also

implies that $\Pr(X < x) = \Pr(X \leq x)$ for continuous random variables. The fact that $f(x) \neq \Pr(X = x)$ is somewhat confusing and may be disappointing, since we started the introduction of a PDF as an approximation of the histogram for all values in a population. This emphasizes that the PDF is a mathematical abstractness or model for describing population values. The abstractness comes from the fact that densities are formulated for infinitely large populations. In terms of random measurement errors, it would make sense to assume that there is an infinite number of possible random errors that could influence the measurement.

4.4 Expected Values of Continuous Random Variables

In the subsection on probability densities we discussed the population mean and standard deviation. These population characteristics can now be more rigorously defined through the continuous random variables. The random variable represents a variable of the population without yet knowing its outcome. If we "average out" all the possible outcomes, where we weight each outcome with the PDF, we obtain a kind of "weighted average", similar to what we did in Chap. 2. However, we have many values to average out, in principle all values of $\mathbb{R}$, which we can not just average (there are far too many values). The averaging is then conducted by the use of integrals as a generalization of summation.

Let X be a continuous random variable with density f, then the *expected value of random variable* X is defined by

$$\mathbb{E}(X) = \int_{\mathbb{R}} xf\,(x)\,dx. \tag{4.6}$$

This expectation represent the population mean, which is typically denoted by the parameter μ as we used in Chap. 2. Thus, the population mean is $\mu = \mathbb{E}(X)$. Note that we have used the symbol $\mathbb{E}$ before in Chap. 2; at that point in the text we did not explain exactly what it meant, but now we know its formal definition. The population variance σ^2 can also be formulated in terms of an expected value of the random variable. The population variance σ^2 is now given by $\sigma^2 = \mathbb{E}(X - \mu)^2$, with

$$\mathbb{E}(X - \mu)^2 = \int_{\mathbb{R}} (x - \mu)^2 f\,(x)\,dx. \tag{4.7}$$

Clearly, we can generalize this concept. If we consider a (not necessarily continuous or differentiable) function $\psi : \mathbb{R} \to \mathbb{R}$, then the *expected value of the random variable* $\psi\,(X)$ is defined by

$$\mathbb{E}\psi\,(X) = \int_{\mathbb{R}} \psi\,(x)\,f\,(x)\,dx. \tag{4.8}$$

The population mean is now obtained by taking $\psi(x) = x$ and the population variance is obtained by taking $\psi(x) = (x - \mu)^2$. Thus the function ψ may depend on population parameters.

The mean μ is also called the *first moment* of the random variable X and the variance is called the *second central moment* of the random variable X, since it squares the random variable after the mean is subtracted. We can also investigate other moments of the random variable X. The *pth moment* of random variable X is obtained by Eq. (4.8) with $\psi(x) = x^p$ and the *pth central moment* of random variable X is obtained by choosing $\psi(x) = (x - \mu)^p$ in Eq. (4.8). The third and fourth central moments are related to the skewness and kurtosis of the population values. The skewness is equal to $\gamma_1 = \mathbb{E}(X - \mu)^3/\sigma^3$ and the kurtosis is $\gamma_2 = \mathbb{E}(X - \mu)^4/\sigma^4 - 3$. Note that the moments of a random variable X may not always exist: this depends on the density f.

In the following table we provide the mean, variance, skewness, and kurtosis of the five parametric distributions we introduced in Sects. 4.2 and 4.3. Section 4.3 has already provided the mean and variance, but not the skewness and kurtosis. Note that we have used the following notation $\tau^2 = \exp\{\sigma^2\}$ in the table.

	Mean	Variance	Skewness	Kurtosis
Normal distribution (μ, σ)	μ	σ^2	0	0
Lognormal distribution (μ, σ)	$\exp\{\mu\}\tau$	$(\tau^2 - 1)\tau^2 \exp\{2\mu\}$	$(\tau^2 + 2)\sqrt{\tau^2 - 1}$	$\tau^8 + 2\tau^6 + 3\tau^4 - 6$
Uniform distribution (θ_0, θ_1)	$[\theta_1 - \theta_0]/2$	$[\theta_1 - \theta_0]^2/12$	0	$-6/5$
Exponential distribution λ	λ^{-1}	λ^{-2}	2	6
Double exponential distribution λ	0	$2\lambda^{-2}$	0	3

In Eq. (4.8) we used the function ψ and therefore discussed the expected value of random variable $\psi(X)$. The random variable $\psi(X)$ can be seen as a mathematical transformation of the original random variable. Knowing the expected value of this transformed random variable provides us the mean of the population of transformed values. However, if we can establish the full distribution function of $\psi(X)$, this gives us much more information about the population of transformed values than just the mean.

The full distribution function of $\psi(X)$ can always be established, but it does not always have a simple workable form. To illustrate a case in which we can obtain the full distribution function of a transformed random variable in workable form, we will start with X being normally distributed with parameters μ and σ, and consider the function $\psi(x) = \exp\{x\}$. The distribution function of the normally distributed random variable X is given by

$$\Pr(X \leq x) = \int_{-\infty}^{x} \frac{1}{\sigma}\phi\left(\frac{z-\mu}{\sigma}\right) dz = \int_{-\infty}^{(x-\mu)/\sigma} \phi(z)dz,$$

and this normal distribution function is often denoted by $\Phi((x-\mu)/\sigma)$, which is defined as $\Phi(x) = \int_{-\infty}^{x} \phi(z)dz$. Then for any value $x > 0$, we can obtain the distribution function of $\exp\{X\}$ by

$$\begin{aligned}
\Pr(\exp\{X\} \leq x) &= \Pr(X \leq \log(x)) \\
&= \Phi\left(\frac{\log(x)-\mu}{\sigma}\right) \\
&= \int_{-\infty}^{\log(x)} \frac{1}{\sigma}\phi\left(\frac{z-\mu}{\sigma}\right) dz \\
&= \int_{0}^{x} \frac{1}{z\sigma}\phi\left(\frac{\log(z)-\mu}{\sigma}\right) dz. \qquad (4.9)
\end{aligned}$$

Since the last integral contains the lognormal PDF, we have obtained that the distribution of $\exp\{X\}$ is lognormally distributed with parameters μ and σ. Note that the integral does not start from $-\infty$, but we know that the lognormal density is zero for $x \leq 0$, thus the integral from $-\infty$ to 0 does not contribute. Thus we see that the lognormal PDF is related to the random variable $\exp\{X\}$ when X is normally distributed. We have now learned that the logarithm of a lognormally distributed random variable is normally distributed.

Now we can generalize this for any random variable X and any monotone differentiable function ψ. Let F be the distribution function of X and f the PDF, we then have

$$\begin{aligned}
\Pr(\psi^{-1}(X) \leq x) &= \Pr(X \leq \psi(x)) \\
&= F(\psi(x)) \\
&= \int_{-\infty}^{\psi(x)} f(z)dz \\
&= \int_{-\infty}^{x} \psi'(z) f(\psi(z))dz,
\end{aligned}$$

with ψ' the derivative of ψ. The calculations now show that the distribution function of random variable $\psi^{-1}(X)$ is equal to $F(\psi(x))$ and the PDF is $\psi'(x)f(\psi(x))$. An interesting consequence of this finding is that the distribution function of the random variable $F^{-1}(U)$, with U a standard uniform distributed random variable and F any invertible distribution function, is now given by F. Indeed, just applying the same procedure as above, we find that $\Pr(F^{-1}(U) \leq x) = \Pr(U \leq F(x)) = F(x)$. Thus the PDF of random variable $F^{-1}(U)$ must now be equal to f. This result for the standard uniform random variable U is very convenient if we want to simulate data from some kind of distribution function F, as we will explain in Sect. 4.8.3.

4.5 Distributions of Discrete Random Variables

Not all the data that we collect and observe are realizations of continuous random variables. Many applications provide us with discrete numerical data, e.g., the number of defective products, the number of microorganisms in a production environment, the presence or absence of a disease, the score on an intelligence test, etc. For these discrete numerical variables, we can also formulate random variables. A discrete random variable X is a random variable that takes its values in the set $\mathbb{N} = \{0, 1, 2, 3, ...\}$.[8]

For discrete random variables we can define $p_k = \Pr(X = k)$ as the probability of observing the outcome k. This is referred to as the *probability mass function* (PMF) if the probabilities p_k satisfy two conditions. First, all probabilities p_k should be nonnegative ($p_k \geq 0, \forall k$) and secondly, the probabilities need to add up to one, i.e., $\sum_{k=0}^{\infty} p_k = 1$. This second condition is related to the way we constructed probabilities: the probability that one of the events (in this case outcome k) happens—regardless of which one—is equal to 1.[9] Sometimes we would like to use the notation $f(x) = \Pr(X = x)$, with $x \in \mathbb{N}$, for the PMF, and we can then write the PMF out in full

$$f(x) = \begin{cases} 0 & \text{if } x < 0 \\ p_0 & \text{if } x = 0 \\ p_1 & \text{if } x = 1 \\ p_2 & \text{if } x = 2 \\ \vdots & \vdots \\ p_k & \text{if } x = k \\ \vdots & \vdots \end{cases}$$

If only a few numbers of discrete values are possible, like the outcomes for throwing a die, then most of the probabilities p_k will be equal to zero. For throwing a fair die, we may have the random variable X taking its values in the set $\{1, 2, 3, 4, 5, 6\}$ and the probabilities $p_0 = 0$, $p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = 1/6$, and $p_k = 0$ for $k > 6$. For binary events, like the occurrence of a disease, we can introduce the random variable X that takes its value in the set $\{0, 1\}$. The probabilities p_k may then be defined as $p_0 = 1 - p$, $p_1 = p$, and $p_k = 0$ for $k > 1$ and some value $p \in [0, 1]$.

The PMF for a discrete random variable is the equivalent of the PDF for a continuous random variable. This means that there is also a distribution function for a discrete random variable. The distribution function or cumulative density function (CDF) for a discrete random variable X is now given by $F(x) = \Pr(X \leq x) = \sum_{k=0}^{x} f(k)$.[10]

---

[8] Discrete does not always mean that we observe values in $\mathbb{N}$. For instance, grades on a data science test may take values in $\{1, 1.5, 2.0, 2.5, \ldots, 9.0, 9.5, 10\}$. Thus, it would be more rigorous to say that a discrete random variable X takes its values in the set $\{x_0, x_1, x_2, \ldots, x_k, \ldots\}$, with x_k an element of the real line ($x_k \in \mathbb{R}$) and with an ordering of the values $x_0 < x_1 < x_2 < \cdots$. However, in many practical settings we can map this set to a subset of $\mathbb{N}$ or to the whole set $\mathbb{N}$.

[9] In the more general setting, the probability can be defined as $P(X = x_k) = p_k$.

[10] If the set is $\{x_0, x_1, x_2, \ldots, x_k, \ldots\}$, with $x_0 < x_1 < x_2 < \cdots$, then the CDF is defined as $F(x) = \sum_{k=0}^{m_x} f(x_k)$, with m_x the largest value for k that satisfies $x_k \leq x$.

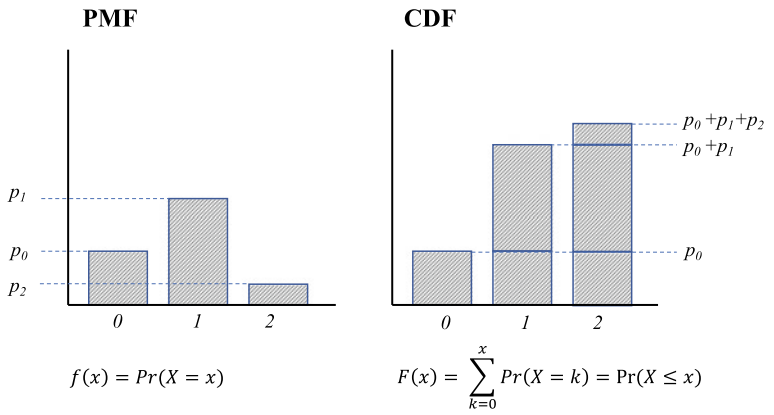

Fig. 4.5 Relationship between the PMF and CDF of a discrete random variable with three possible outcomes

The CDF represents the probability that the random variable X will have an outcome less than or equal to the value x, the same as for continuous random variables. We can write the CDF in full by stating:

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ p_0 & \text{if } x \leq 0 \\ p_0 + p_1 & \text{if } x \leq 1 \\ p_0 + p_1 + p_2 & \text{if } x \leq 2 \\ \vdots & \vdots \\ p_0 + p_1 + \cdots + p_k & \text{if } x \leq k \\ \vdots & \vdots \end{cases}$$

Knowing either the PMF or the CDF of a discrete random variable suffices to describe the probabilities associated with the values that the discrete random variable can take. Clearly, the PMF and CDF are closely related: Fig. 4.5 demonstrates the relationship between the PMF and the CDF.

The PMF and CDF for a discrete random variable that we introduced based on our coin-tossing and dice throwing scenarios presented earlier are:

$$f(x) = \begin{cases} \frac{1}{2} \text{ if } x = 0 \\ \frac{1}{2} \text{ if } x = 1 \end{cases}$$

and

$$F(x) = \begin{cases} 0 \text{ if } x < 0 \\ \frac{1}{2} \text{ if } 0 \leq x < 1 \\ 1 \text{ if } x \geq 1 \end{cases}$$

Note that we do not have to restrict the value x to values in $\mathbb{N}$ for the CDF. We can also use any x in $\mathbb{R}$. If we define $\lfloor x \rfloor$ as the highest integer value that is less than or equal to x, the CDF for any value $x \in \mathbb{R}$ is now given by $F(x) = F(\lfloor x \rfloor)$. This means that the CDF is a *step function*, see Fig. 4.5. For any value $x \in [\lfloor x \rfloor, \lfloor x + 1 \rfloor)$ the CDF is constant.

4.6 Expected Values of Discrete Random Variables

In this section we discuss expected values of a discrete random variable X, similar to what we did for continuous random variables. More specifically, we will discuss the expectation of the discrete random variable $\psi(X)$. The definition is similar to the definition for continuous random variables, but for discrete random variables we can use summation instead of using integrals. Thus, the expectation of a discrete random variable $\psi(X)$ is given by

$$\mathbb{E}(\psi(X)) = \sum_{k=0}^{\infty} \psi(k)\, p_k = \sum_{k=0}^{\infty} \psi(k) \Pr(X = k). \tag{4.10}$$

This definition is closely related to the definition of the expected population parameter for an estimator T as discussed in Chap. 2. If we would collect many realizations of the discrete random variable X, say N realizations, we expect to see value k with frequency $N \cdot p_k$. Thus, the mean value of the random variable $\psi(X)$ would be calculated with Eq. (4.10) when the number of realizations N becomes really large. This was the same argument used in Chap. 2 for an estimator T that was used on a sample of data that was collected with probability sampling.

If we choose $\psi(x) = x$ we obtain the expected value of X and this is again referred to as the *mean* of the random variable or the mean of the population, the same as for continuous random variables. We also use the same notation μ for this mean, i.e., $\mu = \mathbb{E}(X)$. By choosing $\psi(x)$ equal to $\psi(x) = (x - \mu)^2$ and using this in Eq. (4.10) we obtain the variance of a discrete random variable X, and denote this by $\sigma^2 = \mathbb{E}(X - \mu)^2$.

Similar as for the continuous random variables, we can investigate other moments of the discrete random variable X. The pth moment of a discrete random variable X is obtained by Eq. (4.10) with $\psi(x) = x^p$ and the pth central moment of a discrete random variable X is obtained by choosing $\psi(x) = (x - \mu)^p$ in Eq. (4.10). The skewness and kurtosis of a discrete random variable X (or equivalently the skewness and kurtosis for a population with discrete values), are equal to $\gamma_1 = \mathbb{E}(X - \mu)^3/\sigma^3$ and $\gamma_2 = \mathbb{E}(X - \mu)^4/\sigma^4 - 3$, respectively, using Eq. (4.10) for the expectation $\mathbb{E}$. Note that the moments of a discrete random variable X may not always exist: this depends on the choice of probabilities p_k.

4.7 Well-Known Discrete Distributions

Similar to the case of PDFs, the PMFs may typically have a particular form that is known up to a set of one or more parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)^T$. The PMF is then denoted by $f_{\boldsymbol{\theta}}(x)$, using the same notation as for continuous PDFs, where again we may use other symbols for the parameters $\boldsymbol{\theta}$, since this is more aligned with literature. People have studied different forms of the distribution of discrete random variables for a large number of applications. In this section we introduce four famous discrete distributions.[11] We do so by providing a story that motivates the random variable.

4.7.1 Bernoulli Probability Mass Function

The story of the Bernoulli random variable is simple: Bernoulli random variables are motivated by considering *binary* random variables: i.e., random variables that take on values 0 or 1. The simplest example of this is a single coin toss where we map tails to 0 and heads to 1. Now introduce the parameter p, with $0 \leq p \leq 1$, for the probability that the binary random variable X will be equal to 1. This gives rise to the following PMF:

$$\Pr(X = x) = f_p(x) = p^x(1-p)^{1-x},$$

with $x \in \{0, 1\}$ and $f_p(x) = 0$ for any other value of x. A binary random variable with the above PMF is said to be Bernoulli distributed. Also, note that we often write $X \sim \mathscr{B}(p)$ to denote that X is Bernoulli distributed with parameter p. We leave it to the reader to specify the Bernoulli CDF.

The mean and variance of a Bernoulli random variable are easily determined by using Eq. (4.10). The mean μ is equal to $\mu = \mathbb{E}(X) = \sum_{k=0}^{1} kp^k(1-p)^{1-k} = p$ and the variance σ^2 is equal to $\sigma^2 = \mathbb{E}(X-p)^2 = p^2(1-p) + (1-p)^2 p = p(1-p)$. Thus the mean and variance are just functions of the parameter p. The mean represents the average number of "ones" (or events), which is obviously equal to the probability p of observing the value 1.

4.7.2 Binomial Probability Mass Function

The binomial distribution follows from the idea that we might be interested in the total number of heads if we toss a coin multiple times or the total number of airplane accidents or crashes per year. The quantity of interest is the total number of ones S_n (e.g., heads for the coin and crashes for the airplanes), when n represents the total number of tosses or flights per year. The random variable S_n is then given by

[11] Many more distribution functions are known and often used and studied; we present only a small selection.

$S_n = X_1 + X_2 + \cdots + X_n$, with X_k the binary random variable for toss or flight k. Obviously, the random variable S_n can attain the outcome values $0, 1, 2, \ldots, n$. The small letter s is in this case used to indicate a possible value that S_n can take. If the outcome is equal to zero ($s = 0$) we have not seen a single head in n throws or any accident in n flights, while if $s = n$ all throws were heads or all airplane flights resulted in an accident.

The PMF of the binomial is given by:

$$P\left(S_n = s\right) = f_{n,p}(s) = \binom{n}{s} p^s \left(1 - p\right)^{n-s},$$

where

$$\binom{n}{s} = \frac{n!}{s!\left(n - s\right)!},$$

and $s!$ is the total number of permutations of a set of s (different) values. While already briefly introduced in Chap. 2, we discuss permutations in more detail in the additional materials at the end of this chapter. We have visualized three binomial PMFs in Fig. 4.6.

The binomial PMF has two parameters: probability $p \in [0, 1]$ and the number of trials n. In many settings the number of trials will be known, and only p is unknown. For instance, the probability p of passing a data science test would be unknown, but the number of students taking the exam is known upfront. However, in some settings the number of trials is not known, while the probability p is assumed known. For instance, the estimation of the number of microorganisms in a container solution (e.g., milk container) based on a set of binary test scores of small sample volumes from the container (Cochran 1950; van den Heuvel 2011).

To obtain the mean and variance for a binomial random variable with formula (4.10) is somewhat more work than for the Bernoulli. However, there exist closed form expressions. The mean is equal to $\mu = \mathbb{E}(S_n) = np$ and the variance is

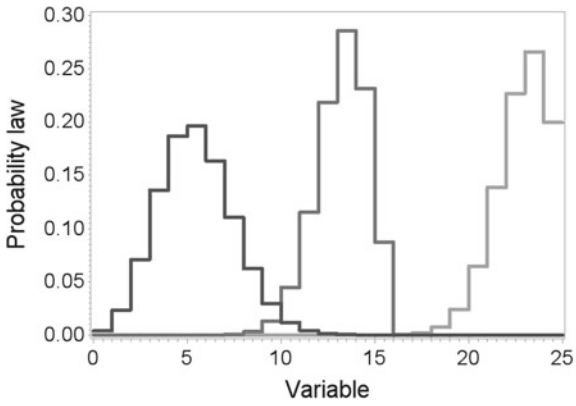

Fig. 4.6 PMFs for the binomial distribution

$\sigma^2 = \mathbb{E}(S_n - np)^2 = np(1-p)$. Since the binomial random variable S_n is the sum of n independent Bernoulli variables, it may be expected that the mean is just n times the mean of a Bernoulli random variable. However, this rule also seems to hold for the variance, which may be less expected. In Sect. 4.10 we will show that these rules hold true in general, irrespective of the underlying PMF.

4.7.3 Poisson Probability Mass Function

A disadvantage of a random variable with a binomial distribution function is that it is bounded by the number of trials n. There are, however, many applications where the number of events or count is not necessary bounded by a fixed number (or at least it is difficult to formulate this bound). In such cases we can use the Poisson distribution, which is often used to express the probability of a given number of events occurring in a fixed interval of time when these events occur with a known constant rate and independently of the time since the last event.

Let X be a random variable with outcome set $\{0, 1, 2, 3, \ldots\}$; then X has a Poisson PMF with parameter $\lambda > 0$ when the probability of observing k events is given by

$$P(X = k) = f_\lambda(k) = \frac{\lambda^k}{k!} \exp(-\lambda)$$

Figure 4.7 shows three different choices of the Poisson PMF.

The mean of a Poisson random variable is equal to $\mu = \mathbb{E}(X) = \lambda$. Thus the parameter λ represents the average number of counts. When the mean is larger than 5, the shape of the PMF looks very much like the normal PDF (see the rightmost PMF in Fig. 4.7). The normal PDF is then viewed as a smooth version of the discrete Poisson PMF. The variance of a Poisson random variable is equal to the mean, i.e.,

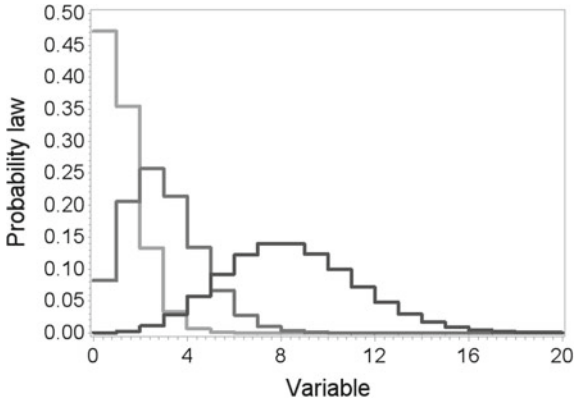

Fig. 4.7 PMFs for the Poisson distribution

$\sigma^2 = \mathbb{E}(X - \lambda)^2 = \lambda$. It requires some in-depth calculations to obtain the mean and variance using Eq. (4.10).

4.7.4 Negative Binomial Probability Mass Function

The negative binomial PMF is often considered a Poisson PMF with an extra amount of variation, even though it originated from a different type of application. In this original application, a random variable X has a negative binomial PMF when it represents the number of trials needed to obtain a fixed known number of binary events. For instance, how many products (e.g., light bulbs) X should be tested before we observe, say r, defective products (e.g., not working light bulbs), when each product has the same probability p of being defective. Thus the negative binomial has two parameters p and r, with r typically known in this application. However, in this form the connection with the Poisson is less obvious.

The negative binomial PMF that we will introduce has two parameters λ and δ, where λ still represents the mean, the same as for the Poisson random variable, but the δ represents an *overdispersion* parameter, indicating the extra amount of variation on top of the Poisson variation. The PMF for the original application is the same as the PMF we will introduce, but it is just a different way of parameterizing the PMF.

A negative binomial random variable X has its outcomes in the set {0, 1, 2, 3, ...}, like the Poisson random variable. The PMF is defined by

$$\Pr(X = k) = f_{\lambda,\delta}(k) = \frac{\Gamma(k + \delta^{-1})}{\Gamma(k+1)\Gamma(\delta^{-1})} \frac{(\delta\lambda)^k}{(1 + \delta\lambda)^{k+\delta^{-1}}},$$

with Γ the gamma function given by $\Gamma(z) = \int_0^\infty x^{z-1} \exp\{x\} dx$. Note that $\Gamma(k) = (k-1)!$, when k is an integer. A few choices for the negative binomial PMF are provided in Fig. 4.8. The mean and variance of the negative binomial random variable

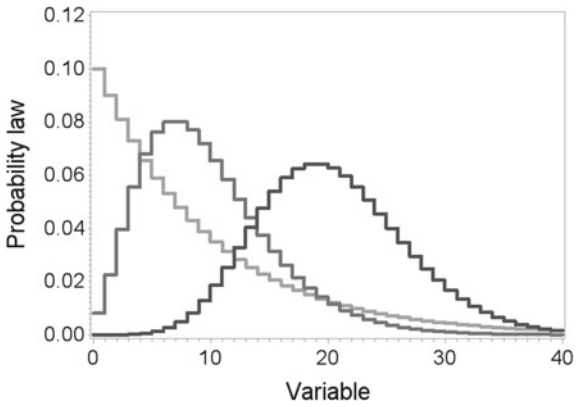

Fig. 4.8 PMFs for the negative binomial distribution

are given by $\mu = \mathbb{E}(X) = \lambda$ and $\sigma^2 = \mathbb{E}(X - \lambda)^2 = \lambda + \delta\lambda^2$. Clearly, in case the parameter δ converges to zero, the variance converges to the variance of a Poisson random variable. This is the reason that the parameter δ is called the overdispersion.

4.7.5 *Overview of Moments for Well-Known Discrete Distributions*

The following table provides the expected value, variance, skewness, and kurtosis for the four discrete distributions we introduced above. Note that we have already provided the means and variances.

	Mean	Variance	Skewness	Kurtosis
Bernoulli $f_p(x)$	p	$p(1-p)$	$\frac{1-2p}{\sqrt{p(1-p)}}$	$\frac{1-6p(1-p)}{p(1-p)}$
Binomial $f_{n,p}(x)$	np	$np(1-p)$	$\frac{1-2p}{\sqrt{np(1-p)}}$	$\frac{1-6p(1-p)}{np(1-p)}$
Poisson $f_\lambda(x)$	λ	λ	$1/\sqrt{\lambda}$	$1/\lambda$
Negative Binomial $f_{\lambda,\delta}(x)$	λ	$\lambda + \delta\lambda^2$	$\frac{1+2\delta\lambda}{\sqrt{\lambda(1+\delta\lambda)}}$	$6\delta + [\lambda(1+\delta\lambda)]^{-1}$

When the number of trials n for the binomial random variable is large, the skewness and kurtosis are close to zero. Actually, it is not the number of trials that is important, but either the number of events np or the number of non-events $n(1-p)$ that should be large to make the skewness and kurtosis close to zero. In that case the Binomial PMF looks very much like the normal PDF. An example of this situation is given by the most right PMF in Fig. 4.6. Here the number of trials is equal to $n = 25$ and the probability of an event is $p = 0.20$. This gives a skewness of 0.3 and a kurtosis of 0.01. See also Sect. 4.9.

Something similar is also true for the Poisson and Negative Binomial random variables. For the Poisson the mean λ should be relatively large to have a shape that is similar to a normal PDF. We already indicated this. The most right PMF in Fig. 4.7 is close to a normal PDF. The mean of this Poisson PMF was equal to $\lambda = 8$, which makes the skewness equal to 0.35 and the kurtosis equal to 0.125. For the Negative Binomial random variable the mean λ should also be large, but the overdispersion should not be too large. Indeed, a large mean will put the skewness close to zero, but the kurtosis may still be away from zero when delta is too large, due to the term 6δ in the kurtosis. In Fig. 4.8 the most right PMF is closest to a normal PDF, although it is still a little bit skewed and has a little bit thicker tails than the normal density. This PMF has a mean of $\lambda = 20$ and an overdispersion of $\delta = 0.05$, which makes the skewness equal to 0.47 and the kurtosis equal to 0.325.

4.8 Working with Distributions in R

The functions and packages in R can support us when working with random variables, PDFs, PMFs, and CDFs. First of all, R can help us calculate particular probabilities. As we have seen, not all CDFs have closed-form expressions. Thus, to determine a CDF value requires either calculation of integrals or otherwise summations of many terms. Numerical approaches have been implemented in R to help us do this with the computer. Secondly, R can help us create population values or samples from populations which are described by a PDF or PMF. We will demonstrate how you can use R to—by means of Monte-Carlo (MC) simulation—compute summaries of random variables with complex distribution functions. Summaries are sometimes obtained exact, like the means, variance, skewness, and kurtosis reported earlier, but not every type of summary can always be determined exactly or it may be more time-consuming than just using a MC simulation. Finally, we will also demonstrate a method that allows you to obtain realizations (draws) of a random variable with a specific distribution function that you may have created your self or that exists in the literature but not in R.

4.8.1 R Built-In Functions

R offers a number of well-known distribution functions. It uses a standardized naming scheme to name the functions that relate to probability distributions. The name always consists of (an abbreviation of) the mathematical name of the distribution function—for example `norm` for the normal distribution function—with one of the following prefixes:

- `d-` A distribution function with the prefix `d-`, for example `dnorm`, allows you to evaluate the PDF or PMF at a specific value. Thus, a call to `dnorm(x, mu, sigma)` evaluates the PDF of the normal distribution function with mean `mu` and standard deviation `sigma` at `x`.
- `p-` A distribution function with the prefix `p-`, for example `pnorm`, allows you to evaluate the CDF at a specific value. Thus, a call to `pnorm(x, mu, sigma)` evaluates the CDF in x of the normal distribution function with mean `mu` and standard deviation `sigma` at `x`.
- `q-` A distribution function with the prefix `q-`, for example `qnorm`, allows you to evaluate the so-called quantile or inverse function of the distribution functions. The quantile function specifies the value of the random variable that gives you the specific probability of the variable being less than or equal to that value. Thus the quantile function gives $x = F^{-1}(u)$ for probability or value $u \in (0, 1)$, with F the CDF.
- `r-` A distribution function with the prefix `r-`, for example `rnorm`, allows you to generate draws of a random variable with a specific distribution function. Thus `rnorm(1000, mu, sigma)` returns a vector of length 1,000 containing draws from a normal random variable with mean `mu` and standard deviation `sigma`.

The normal, uniform, and exponential distributions we covered in this chapter are called `-norm`, `-unif` and `-exp`. The lognormal distribution function does not exist in R, but it can be created through the normal distribution function. The binomial, Poisson, and negative binomial distributions are given by `-binom`, `-pois`, and `-nbinom`. The Bernoulli is just a special case of the binomial with a size (or number of trials) of one. It should be mentioned that the parametrization of the negative binomial in R is different from our formulation, which means that we need to study the difference between the parametrization in R and our formulation (when you investigate `-nbinom`, R gives info on this difference). A full overview of the built-in distribution functions can be found at https://stat.ethz.ch/R-manual/R-devel/library/stats/html/Distributions.html.

The code below gives an example in which we first evaluate the PDF and the CDF of the normal for a given mean and standard deviation. Next, we demonstrate the equality concerning the quantile function we highlighted above, and finally we obtain 10 draws from the same normal distribution function.

```
> mu <- 0        # Mean
> s2 <- 1        # Variance
> s <- sqrt(s2) # Standard deviation
>
> x <- 1
> dnorm(x, mean=mu, sd=s) # PDF of the normal distribution
    evaluated at x
[1] 0.2419707
> pnorm(x, mean=mu, sd=s) # CDF of the normal distribution
    evaluated at x
[1] 0.8413447
>
> p <- pnorm(x, mean=mu, sd=s)
> qnorm(p, mean=mu, sd=s) # The so-called quantile function Q(p)
    = x if and only if p = F(x)
[1] 1
>
> set.seed(982749)
> n <- 10
> rnorm(n, mean=mu, sd=s) # Get 10 draws / realizations from the
    distribution
 [1]  0.15958190 0.60671592 1.10638675 -1.03021164 0.14672386
      0.37733998
 [7]  0.55563879 0.77358142 0.61140111 -0.09188106
```

4.8.2 Using Monte-Carlo Methods

The ability to easily obtain draws from a distribution function allows us to approximate the properties of distribution functions by computing summaries of the draws obtained. This method is called Monte Carlo (or MC) simulation, and we can use it

to check our analytical results. For example, we can approximate the expected value of a random variable $X \sim \mathcal{N}(2, 9)$ using the following code[12]:

```
> draws <- rnorm(10^6, mean=2, sd=3)
> mean(draws)
[1] 1.996397
```

In this case we already knew that the mean was equal to 2. The simulation shows that we obtain a value very close to 2 and confirms our knowledge. This MC approach is closely related to simple random sampling in Chap. 2, but now we sample from an infinitely large population that is described by the normal distribution function. The example is somewhat simple, but it shows how simulation works. MC becomes more relevant when more complicated random variables are being studied. For instance, the expected value of $\exp\{\sqrt{X}\}$, with $X \sim \mathcal{N}(\mu, \sigma^2)$, is less easy to determine mathematically. You may think that this may be an exotic random variable to study, but practice often studies very interesting and complex random variables, often a combination of several random variables. Instead of evaluating the mean of the random variables, we could also study the variance and other moments (like skewness and kurtosis), which will be even more difficult to obtain mathematically.

To illustrate MC with multiple random variables, we can easily imagine a random variable Z whose distribution function is a combination of two normal distribution functions with different means and variances, which is something we call a *mixture* distribution: for instance, the distribution of body weight of women and men or the tensile strengths of plastic tubes produced from two production lines. One way of constructing such a variable is by imagining that we first throw a coin, $Y \sim \mathcal{B}(1/3)$, and subsequently we obtain a draw from one of two different normals: $Z_0 \sim \mathcal{N}(10, 1)$ if $Y = 0$ and $Z_1 \sim \mathcal{N}(20, 5)$ if $Y = 1$. Or, more generally, the random variable of interest Z is constructed as $Z = YZ_0 + (1 - Y)Z_1$ where $Y \sim \mathcal{B}(p)$, $Z_0 \sim \mathcal{N}(\mu_0, \sigma_0^2)$, and $Z_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ with $p = 1/3$, $\mu_0 = 10$, $\sigma_0^2 = 1$, $\mu_1 = 20$, and $\sigma_1^2 = 5$.

Since Z is a function of random variables, it is itself a random variable. However, it is not immediately clear (yet) how we could compute summaries such as its expectation, variance, moments, or even the percentage of values above a certain level (say 15). In Sect. 4.9 we will look at this variable more mathematically, but here we will study the variable through MC simulation:

```
> # Set number of draws, probability of coin, and mean and
    standard deviation of first normal distribution
> n <- 10^6
> p <- 1/3
> mu_1 <- 10
> s_1 <- 1
>
> # Flip a coin with probability p
> Y <- rbinom(n, size=1, prob=p)
>
```

[12] Note that `-norm` uses standard deviations instead of variances. You can always type `?rnorm` to see the exact arguments.

```
> # Generate the mixture draws (note that 20 = 1*10+10 and 5 = 1*
    4+1)
> Z <- rnorm(n, mean=(Y*10)+mu_1, sd=(Y*4)+s_1)
>
> # Plot the draws in a histogram
> hist(Z, freq=FALSE, breaks=50)
>
> # Compute mean and variance (increase the power to compute
    higher central moments)
> mean_Z <- mean(Z)
> var_Z <- mean((Z-mean(Z))^2)
> mean_Z
[1] 13.31672
> var_Z
[1] 31.09315

> # Compute percentage of values above 15
> P <- (Z>15)
> mean_P <- mean(P)
> mean_P
[1] 0.28147
```

Figure 4.9 shows the histogram of the draws generated using this code. The mixing of the two normals is clearly visible. Note that we use `freq=FALSE` to print a histogram that has probabilities on the y-axis instead of frequencies (which is the default we saw in Chap. 1).

The example above shows how we can better understand the properties of a distribution or a random variable, but it can also be used to evaluate estimators that are being used on samples from a population, as we discussed in Chap. 2. The only difference with Chap. 2 is that we are making certain assumptions on the population values in this simulation. Indeed, we have assumed that the population is described by a particular PDF or PMF or a particular set of random variables. If we study one large simulation of many draws (like 10^6 in the example above) we obtain a population that can give us better insight in the population characteristics, but if we

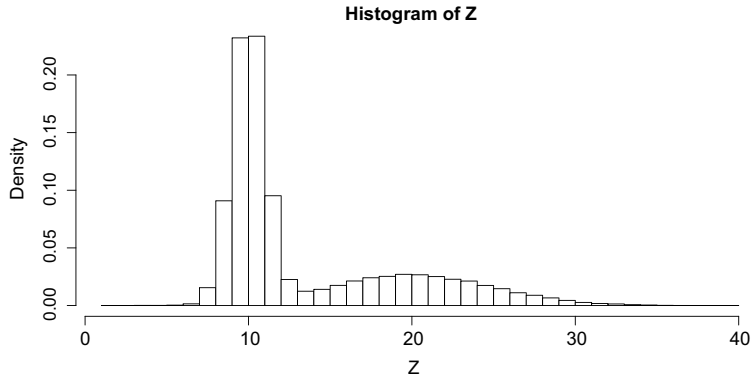

Fig. 4.9 The histogram of a mixture distribution

simulate repeatedly (say 10^3) a smaller number of draws (equal to an anticipated sample size n) and calculate our estimator on each simulation sample, we can evaluate the performance of the estimator. Using the (10^3) repeated simulations, we can approximate the mean, variance and other moments of the estimator, like we did in Chap. 2.

4.8.3 *Obtaining Draws from Distributions: Inverse Transform Sampling*

It is clear that R is very useful for working with probability distributions; we can evaluate PMFs, PDFs, and CDFs, and we can use MC simulation to compute expectations, variances, and moments of random variables and estimators—this is even possible when we might not be able to do so analytically. However, in the examples above we are inherently limited by Rs default functions; hence, we can only work with the well-known distributions that R supports. Although there are packages that implement more distributions, sometimes we might want to obtain draws from a distribution that is not well-known and implemented by others. If this is the case, we can sometimes use *inverse transform sampling* as we discussed in Sect. 4.4; we demonstrate it here in more detail to generate draws of an exponential distribution function—which actually *is* implemented in R—and than check our results.

Figure 4.10 shows the idea behind inverse transform sampling. As long as we know the CDF of a random variable, we can use draws from a uniform distribution function to generate draws from the variable of interest by evaluating the inverse of the CDF (obviously this is something we need to derive ourselves). For example, consider the exponential distribution function: the exponential distribution function with parameter λ has the following PDF and CDF:

$$f_\lambda(x) = \lambda \exp\{-\lambda x\} \qquad F_\lambda(x) = 1 - \exp\{-\lambda x\},$$

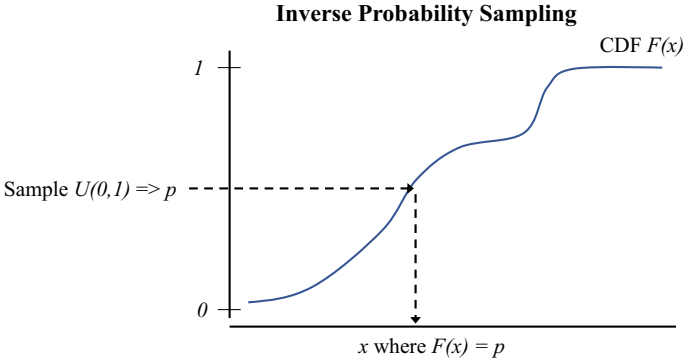

Fig. 4.10 The CDF of a complex continuous distribution function

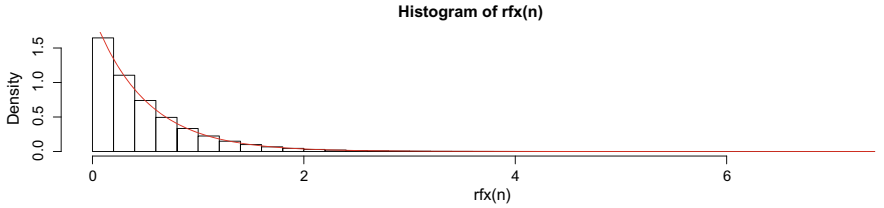

Fig. 4.11 Exponential approximated using draws obtained by inverse transform sampling

with $x > 0$. The inverse of the CDF is now equal to

$$F_{\lambda}^{-1}(u) = -\log(1-u)/\lambda,$$

with $u \in (0, 1)$. Now, we can implement the inverse CDF, $F^{-1}(x)$ in R, and use the inverse transform sampling trick:

```
> # lambda = 2
> # f(x)   = 2*exp(-2*x)
> # F(x)   = 1-exp(-2*x)
> # F(u)^-1 = -log(1-u)/2
>
> cdf_inverse <- function(u) {
+   -log(1-u)/2
+ }
>
> rfx <- function(n) {
+   u <- runif(n, min=0, max=1)
+   cdf_inverse(u)
+ }
>
> n <- 10^6
> draws <- rfx(n)
> hist(draws, freq=FALSE, breaks=40)
> curve(dexp(u, rate=2), col="red", add=TRUE)
```

The last two lines plot a histogram (see Fig. 4.11) of the obtained draws using our trick, and superimpose a curve using the built-in `dexp` function in R; it is clear that our sampling approach works very well! Thus using the uniform random variable we can in principle simulate any other random variable if we know the CDF and its inverse.

4.9 Relationships Between Distributions

We have only introduced a small number of well-known distribution functions above; many probability textbooks will provide many more examples of well-known distribution functions. However, our aim was just to introduce the main concepts; it's

Table 4.1 Comparisons of the probabilities of Poisson and binomial distribution functions

X	Binomial	Poisson	X	Binomial	Poisson	X	Binomial	Poisson
0	0.00000	0.00001	7	0.01456	0.04368	14	0.12441	0.09049
1	0.00000	0.00007	8	0.03550	0.06552	15	0.07465	0.07239
2	0.00000	0.00044	9	0.07099	0.08736	16	0.03499	0.05492
3	0.00004	0.00177	10	0.11714	0.10484	17	0.01235	0.03832
4	0.00027	0.00531	11	0.15974	0.11437	18	0.00309	0.02555
5	0.00129	0.01274	12	0.17971	0.11437	19	0.00049	0.01614
6	0.00485	0.02548	13	0.16588	0.10557	20	0.00004	0.00968

easy to look up the PDFs, CDFs, expectations and moments of specific distributions online (Wikipedia is actually a good source in this regard). Here we highlight a few well-known relationships between distribution functions. We have already seen the relationship between the Bernoulli and binomial distribution functions; we now discuss two more.

4.9.1 Binomial—Poisson

Although the Poisson and binomial distribution functions are different, as they have different supports, they can be close to each other. To demonstrate, Table 4.1 shows the Poisson probabilities next to the binomial probabilities when $\lambda = np$ is equal to 12 and $n = 20$. The probabilities are reasonably close although not extremely close.

The Poisson and binomial distribution functions are quite close whenever λ is equal to np and n is relatively large. It can be shown that the Poisson probabilities are the limit of the binomial probabilities when n converges to infinity under the condition that np converges to λ.

4.9.2 Binomial—Normal

Although the normal distribution function provides probabilities for continuous outcomes and the binomial distribution function provides probabilities for discrete outcomes, the normal distribution function may approximate the binomial distribution function (as we have already discussed in Sect. 4.7.5). The approximation is quite good under certain conditions, in particular when the skewness and kurtosis of the binomial distribution are close to zero.

Let S_n be a binomial random variable with parameters n and p. Probability calculation with the binomial distribution function can be adequately approximated with a normal distribution function when the mean np and the value $n(1-p)$ are both

larger than 5 and the sample size is at least 20 ($n \geq 20$). The binomial probability is then approximated by a normal probability as follows:

$$P\left(S_n \leq k\right) \approx P\left(Z \leq \left(k + 0.5 - np\right) / \sqrt{np\left(1-p\right)}\right),$$

with Z the random variable from a standard normal distribution function.

For large sample sizes this approximation can be very useful, since binomial probabilities may not be easily calculated with a computer, due to the complexity of calculating the number of permutations $n!$ when n is large.

4.10 Calculation Rules for Random Variables

In our discussion of the binomial random variable we saw that it can be formulated as the sum of n binary random variables. In the Monte Carlo simulation we created $Z = YZ_1 + (1-Y)Z_2$, with Y a binary variable and Z_k a normally distributed random variable having mean μ_k and variance σ_k^2. This shows that we are often interested in functions of random variables. In some cases we are able to determine the properties of these constructed random variables that are functions of the properties of the random variables that were used in the construction. This section will provide some calculation rules that apply in all cases (when certain independence conditions are satisfied), whatever the underlying PDF or PMF is used. Thus they are very generic rules. We will state the rules without giving proofs.

4.10.1 Rules for Single Random Variables

Here we assume that we have a random variable X, either discrete or continuous, and a constant c. The following rules hold true.

$$\begin{aligned}
\mathbb{E}\left(c\right) &= c \\
\mathbb{E}\left(cX\right) &= c\mathbb{E}\left(X\right) \\
\mathrm{Var}(X) &\geq 0 \\
\mathrm{Var}(X+c) &= \mathrm{Var}(X) \\
\mathrm{Var}(cX) &= c^2\mathrm{Var}(X) \\
\mathrm{Var}(X) &= \mathbb{E}(X^2) - [\mathbb{E}(X)]^2
\end{aligned}$$

4.10.2 Rules for Two Random Variables

Here we assume that we have a random variable X and a random variable Y. The following rules always hold true:

$$\begin{aligned} \mathbb{E}(X) &= \mathbb{E}(Y), \text{ when } X = Y \\ \mathbb{E}(X+Y) &= \mathbb{E}(X) + \mathbb{E}(Y) \end{aligned}$$

If we assume that the random variables X and Y are independent of each other (see also Chap. 6), we can provide a few other simple rules. Independence means that the outcome of X has nothing to do with the outcome of Y. For instance, when X represents the body weight of women and Y the body weight of men, and we draw randomly one woman and one man from the population, the weight of the woman will be unrelated to, or independent of, the weight of the man.[13]

More mathematically, independence can be defined through our definition of independent events in Chap. 3. If we introduce the events $A = \{X \leq x\}$ and $B = \{Y \leq y\}$, then independence of the two events is given by $\Pr(X \leq x, Y \leq y) = \Pr(A \cap B) = \Pr(A)\Pr(B) = \Pr(X \leq x)\Pr(Y \leq y) = F_X(x)F_Y(y)$, with F_X and F_Y the CDFs for X and Y, respectively. In Chap. 6 we will see that $\Pr(X \leq x, Y \leq y)$ is the joint CDF of X and Y, denoted by $F_{XY}(x, y)$. The two random variables X and Y are now considered independent, when this product of probabilities occurs for every x and y, i.e., when $F_{XY}(x, y) = F_X(x)F_Y(y)$ for all x and y. Note that, when X and Y are independent, the random variables $\varphi(X)$ and $\psi(Y)$ are independent, whatever functions φ and ψ are chosen.

The following rules will hold true when X and Y are independent.

$$\begin{aligned} \mathbb{E}(XY) &= \mathbb{E}(X)\mathbb{E}(Y) \\ \mathrm{Var}(X+Y) &= \mathrm{Var}(X) + \mathrm{Var}(Y) \\ \mathrm{Var}(X-Y) &= \mathrm{Var}(X) + \mathrm{Var}(Y) \\ \mathrm{Var}(XY) &= \mathrm{Var}(X)\mathrm{Var}(Y) + \mathrm{Var}(X)(\mathbb{E}(Y))^2 + \mathrm{Var}(Y)(\mathbb{E}(X))^2 \end{aligned}$$

The second rule shows why the variance of a binomial random variable is n times the variance of a Bernoulli random variable. This rule is applied (sequentially) to the random variable $S_n = X_1 + X_2 + \cdots + X_n$, with $X_1, X_2, \ldots, X_n$ independent random variables. Furthermore, the second and thrid rule tell us that variances of independent random variables always add up, even if we subtract random variables from each other.

[13] Yes, you are correct, practice is more complicated since a man and a woman may share a household and therefore their weights may be related.

Note that the first rule above shows that we could have calculated the mean of the mixture distribution discussed in Sect. 4.8.2 analytically:

$$\begin{aligned}\mathbb{E}(YZ_0 + (1-Y)Z_1) &= \mathbb{E}(YZ_0) + \mathbb{E}((1-Y)Z_1) \\ &= \mathbb{E}Y\mathbb{E}Z_0 + \mathbb{E}(1-Y)\mathbb{E}Z_1 \\ &= p\mu_0 + (1-p)\mu_1.\end{aligned}$$

The fourth rule helps us calculate the variance of YZ_0, which is $p(1-p)\sigma_0^2 + p(1-p)\mu_0^2 + \sigma_0^2 p^2 = p\sigma_0^2 + p(1-p)\mu_0^2$. However, computing the variance of $YZ_0 + (1-Y)Z_1$ is more difficult, since YZ_0 and $(1-Y)Z_1$ are not independent, due to the common random variable Y.

4.11 Conclusion

In this chapter we have introduced one more snippet of theory that we need to advance our analysis of data: we introduced random variables and distributions of random variables. In the next chapter we will, using all the theory that we have now developed, relate back to sample data. We will first discuss distribution functions of sample statistics over repeated random sampling, and we will find that these depend on the parameters of the population distributions that we assume. We will than discuss two methods of estimating these population parameters.

Problems

4.1 In a trial the patients ($n = 20$) are randomly assigned to the groups A and B. The randomization is done by throwing an unbiased die. When the number of dots is even, the patient will be in group A, otherwise in group B.

1. What is the probability that exactly 10 patients will be in group A?
2. What is the probability that at most 9 patients will be allocated to group A?

4.2 In Sect. 4.5 we discussed multiple discrete distribution functions by providing the PMF ($f(x)$) and discussing their means, variances, and central moments.

1. Derive the CDF of the Bernoulli distribution.
2. Determine mathematically that the mean of a binomially distributed random variable X with parameters p and n is equal to $\mathbb{E}X = np$.
3. Determine mathematically that the variance of a binomially distributed random variable X with parameters p and n is equal to $\mathbb{E}(X - np)^2 = np(1-p)$.
4. Determine mathematically that the mean of a Poisson distributed random variable X with parameter λ is equal to $\mathbb{E}X = \lambda$.

5. Determine mathematically that the variance of a Poisson distributed random variable X with parameter λ is equal to $\mathbb{E}(X-\lambda)^2 = \lambda$.
6. Use R to make a figure displaying the CDF of the Poisson distribution with $\lambda = 5$.
7. Determine mathematically that the mean of a uniform distributed random variable X with parameters a and b, with $a < b$, is equal to $\mathbb{E}X = (a+b)/2$.
8. Determine mathematically that the variance of a uniform distributed random variable X with parameters a and b, with $a < b$, is equal to $\mathbb{E}(X - (a+b)/2)^2 = \frac{1}{12}(b-a)^2$.

4.3 Let us assume that the probability of a person in the Netherlands being left-handed is 0.10. What is the probability that in a random group of 20 persons from the Netherlands you will find at least three left-handed persons?

4.4 A specific diagnostic test has a known sensitivity of 0.9 for the related disease. Five patients, all carriers of the disease, do the diagnostic test. Give the probability distribution function of the number of positive tests. This means that you need to calculate $P(S_5 = 0), P(S_5 = 1), \ldots, P(S_5 = 5)$, with S_5 the random variable that indicates the number of positive tests.

4.5 Consider the exponential CDF $F(x) = 1 - \exp(-\lambda x)$, for $x > 0$ and otherwise equal to zero. Now let X be distributed according to this exponential distribution.

1. Determine the mean and variance of X.
2. What is the median value of the exponential distribution function? Use the definition of the median we discussed in Chap. 1.

4.6 Consider the PDF $f(x) = 3x^2$ on the interval $(0, 1]$.

1. Demonstrate that the function is indeed a density.
2. What are the mean, variance and standard deviation?
3. How likely is it that the outcome will be in between 0.25 and 0.75?

4.7 The following questions concern the use of R to work with random variables

1. Use R to make a figure of both the PDF and the CDF of the normal distribution with parameters $\mu = 10$ and $\sigma^2 = 3$.
2. Compute the expected value and variance for the $\mathcal{N}(\mu = 10, \sigma^2 = 3)$ distribution using Monte Carlo simulation.

4.8 Implement inverse transform sampling for the PDF $f(x) = 1/2x$ defined from 0 to 2.

Additional Materials I: From Bernoulli to Binomial

In Chap. 2 we have already discussed permutations. Here we will repeat this for binary values and then discuss how the binomial distribution is generated from Bernoulli distributed random variables. Thus we will consider units (or subjects) i that have a Bernoulli random variable X_i with parameter p.

Consider a sample of three subjects ($n = 3$), and let x_1, x_2, and x_3 be the outcomes or realizations. These values can be ordered in six ($3! = 3 \cdot 2 \cdot 1 = 6$) different ways, or in other words there are six permutations, namely (x_1, x_2, x_3), (x_1, x_3, x_2), (x_2, x_1, x_3), (x_2, x_3, x_1), (x_3, x_1, x_2), and (x_3, x_2, x_1). To see this, we can see that for the first position there are three possibilities to choose from (x_1, x_2, or x_3), then for the second position there are only $2 = (3 - 1)$ possibilities left because the first position is already taken by one of the outcomes. Then for the third position there is only $1 = (3 - 2)$ possibility left, since the previous two positions are taken. Clearly, this can be generalized to k different values, leading to $k! = k \cdot (k - 1) \cdot (k - 2) \cdots 2 \cdot 1$ permutations.

For the binomial distribution function, the values or outcomes from the subjects are not all different, as they are either equal to zero or equal to one (they come from a binary random variable). For instance, when we consider again the three values x_1, x_2, and x_3, with the assumption that $x_1 = 1$, $x_2 = 0$, and $x_3 = 0$, there are still six permutations, but these permutations are not all unique in the sense that the sum over all the values is still the same. The permutation (x_1, x_2, x_3) is exactly the same as permutation (x_1, x_3, x_2), since they are both equal to $(1, 0, 0)$.

The number of unique permutations (also referred to as the number of combinations) is in this case thus three, since they are $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$. Clearly, for a given permutation we could permute all the zero's and all the ones without affecting the result. Thus the number of unique permutations is determined by the total number of permutations, divided by the number of permutations that can be made with the zero's and with the ones. Thus in the example we find $3!/(2! \cdot 1!) = 6/(2 \cdot 1) = 3$. More generally, when the outcomes consist of zero's and ones and the number of ones is for instance k, then the number of unique permutations is given by the binomial coefficient:

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

This so-called *binomial coefficient* is pronounced n over k or n choose k.

Each of the $n!/(k!(n-k)!)$ outcomes result in the exact same probability of occurrence, namely $p^k(1-p)^{n-k}$. To illustrate this with the three outcomes x_1, x_2, and x_3, the probability that $(1, 0, 0)$ occurs is $p(1-p)(1-p)$, that $(0, 1, 0)$ occurs is $(1-p)p(1-p)$, and that $(0, 0, 1)$ occurs is $(1-p)(1-p)p$. Thus all three outcomes have a probability of $p(1-p)^2$ of occurrence. Thus the probability that we see k events (or ones) is now equal to $[n!/(k!(n-k)!)]p^k(1-p)^{n-k}$, which results into the binomial distribution function.

Additional Materials II: The Log Normal Distribution

Determining the moments is a little bit of work, but can be determined by standard calculus methods and the knowledge that $\int_{\mathbb{R}} \sigma^{-1}\phi((x-\mu)/\sigma)dx = 1$. For instance, the first moment of a lognormal distributed random variable $X \sim \mathscr{LN}(\mu, \sigma^2)$ is given by

$$\begin{aligned}
\mathbb{E}(X) &= \int_{\mathbb{R}} x f_{\mu,\sigma}(x)dx \\
&= \int_0^\infty x \tfrac{1}{\sigma x}\phi\left(\tfrac{\log(x)-\mu}{\sigma}\right)dx \\
&= \int_{\mathbb{R}} \exp(z)\tfrac{1}{\sigma}\phi\left(\tfrac{z-\mu}{\sigma}\right)dz \\
&= \int_{\mathbb{R}} \tfrac{1}{\sigma\sqrt{2\pi}} \exp\left(-\tfrac{(z-\mu)^2}{2\sigma^2} + z\right)dz \\
&= \int_{\mathbb{R}} \tfrac{1}{\sigma\sqrt{2\pi}} \exp\left(-\tfrac{z^2-2z(\mu+\sigma^2)+\mu^2}{2\sigma^2}\right)dz \\
&= \exp\left(\tfrac{(\mu+\sigma^2)^2-\mu^2}{2\sigma^2}\right)\int_{\mathbb{R}} \tfrac{1}{\sigma\sqrt{2\pi}} \exp\left(-\tfrac{(z-\mu-\sigma^2)^2}{2\sigma^2}\right)dz \\
&= \exp\left(\mu + 0.5\sigma^2\right)\int_{\mathbb{R}} \tfrac{1}{\sigma}\phi\left(\tfrac{z-\mu-\sigma^2}{\sigma}\right)dz \\
&= \exp\left(\mu + 0.5\sigma^2\right)
\end{aligned}$$

Note that the population mean $\mathbb{E}(X)$ is a function of the density parameters μ and σ^2, which is typically different from the normal distribution. It also implies that the parameters μ and σ do not represent the mean and standard deviation of the random variable $X \sim \mathscr{LN}(\mu, \sigma^2)$.

Using similar calculus techniques, the variance is given by

$$\mathbb{E}(X - \exp(\mu + 0.5\sigma^2))^2 = \exp\left(2\mu + \sigma^2\right)\left(\exp\left(\sigma^2\right) - 1\right),$$

which again is a function of the density parameters μ and σ. This implies that the relative standard deviation is now equal to $RSD = 100\%\sqrt{\exp\left(\sigma^2\right) - 1}$. Thus the relative standard deviation is now only a function of σ^2 and does not depend on μ. The skewness and excess kurtosis are a little more elaborate. They are functions of just the parameter σ and do not depend on μ. To get some feeling about the values of the skewness and kurtosis, we visualized them as function of σ in Fig. 4.12. This figure also plots the relative standard deviation (not expressed as percentage).

The figure suggest that for larger values of σ, the skewness and kurtosis are deviating from the value zero. Thus for larger values of σ, the lognormal distribution function really deviates from the normal distribution function. Additionally, the relative standard deviation is also increasing with σ. For instance, a value of $\sigma = 0.5$ gives an $RSD = 53.29\%$.

The quantiles of the lognormal distribution function can be obtained by the quantiles z_p of the standard normal distribution function. Indeed, let x_p be the pth quantile of the lognormal distribution function: then we know that $F_{\mu,\sigma}(x_p) = p$, with $F_{\mu,\sigma}$ the CDF for $f_{\mu,\sigma}$ in Eq. (4.2) and hence, using relationship Eq. (4.9), we obtain that $x_p = \exp\left(\mu + \sigma z_p\right)$. It follows immediately that the median of the lognormal

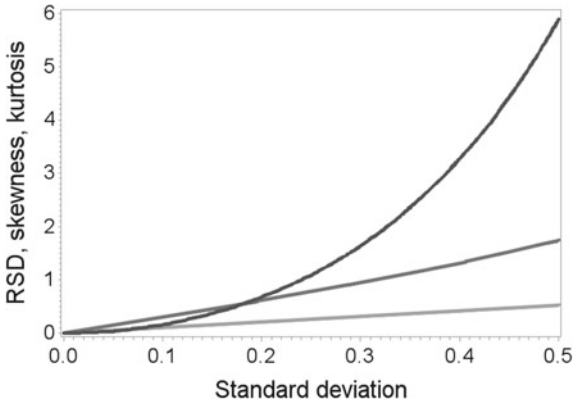

Fig. 4.12 RSD, skewness, and excess kurtosis of lognormal distribution: light gray curve: RSD; gray curve: γ_1, dark gray curve: γ_2

distribution function is equal to $\exp(\mu)$, since $z_{0.5} = 0$. The first and third quartiles are equal to $\exp(\mu - 0.67449\sigma)$ and $\exp(\mu + 0.67449\sigma)$, respectively, since $\Phi(-0.67449) = 1 - \Phi(0.67449) = 0.25$.

References

R.E. Barlow, Mathematical theory of reliability: a historical perspective. IEEE Trans. Reliab. **33**(1), 16–20 (1984)

W.G. Cochran, Estimation of bacterial densities by means of the "most probable number". Biometrics **6**(2), 105–116 (1950)

D. Glass, Graunt's life table. J. Inst. Actuar. **76**(1), 60–64 (1950)

A. Hald, *A History of Parametric Statistical Inference from Bernoulli to Fisher, 1713–1935*. (Springer Science & Business Media, 2008)

S. Kotz, N. Balakrishnan, N.L. Johnson, *Continuous Multivariate Distributions, Volume 1: Models and Applications* (Wiley, Hoboken, 2004)

C. Liu, D. Zheng, C. Griffiths, A. Murray, Comparison of repeatability of blood pressure measurements between oscillometric and auscultatory methods, in *2015 Computing in Cardiology Conference (CinC)* (IEEE, 2015), pp. 1073–1076

O.B. Sheynin, Cf gauss and the theory of errors. Arch. Hist. Exact Sci. **20**(1), 21–72 (1979)

O. Sheynin, Density curves in the theory of errors. Arch. Hist. Exact Sci. **49**(2), 163–196 (1995)

E. van den Heuvel, Estimation of the limit of detection for quantal response bioassays. Pharm. Stat. **10**(3), 203–212 (2011)