Statistics for Business and Economics

Marcelo Fernandes

bookboon
The eBook company

Marcelo Fernandes

Statistics for Business and Economics

Statistics for Business and Economics

ISBN 978-87-7681-481-6

Contents

Chapter 1

Introduction

This compendium aims at providing a comprehensive overview of the main topics that appear in any well-structured course sequence in statistics for business and economics at the undergraduate and MBA levels. The idea is to supplement either formal or informal statistic textbooks such as, e.g., "Basic Statistical Ideas for Managers" by D.K. Hildebrand and R.L. Ott and "The Practice of Business Statistics: Using Data for Decisions" by D.S. Moore, G.P. McCabe, W.M. Duckworth and S.L. Sclove, with a summary of theory as well as with a couple of extra examples. In what follows, we set the road map for this compendium by describing the main steps of statistical analysis.

Statistics is the science and art of making sense of both quantitative and qualitative data. Statistical thinking now dominates almost every field in science, including social sciences such as business, economics, management, and marketing. It is virtually impossible to avoid data analysis if we wish to monitor and improve the quality of products and processes within a business organization. This means that economists and managers have to deal almost daily with data gathering, management, and analysis.

1.1 Gathering data

Collecting data involves two key decisions. The first refers to what to measure. Unfortunately, it is not necessarily the case that the easiest-to-measure variable is the most relevant for the specific problem in hand. The second relates to how to obtain the data. Sometimes gathering data is costless, e.g., a simple matter of internet downloading. However, there are many situations in which one must take a more active approach and construct a data set from scratch.

Data gathering normally involves either sampling or experimentation. Albeit the latter is less common in social sciences, one should always have in mind that there is no need for a lab to run an experiment. There is pretty of room for experimentation within organizations. And we are not speaking exclusively about research and development. For instance, we could envision a sales competition to test how salespeople react to different levels of performance incentives. This is just one example of a key driver to improve quality of products and processes.

Sampling is a much more natural approach in social sciences. It is easy to appreciate that it is sometimes too costly, if not impossible, to gather universal data and hence it makes sense to restrict attention to a representative sample of the population. For instance, while census data are available only every 5 or 10 years due to the enormous cost/effort that it involves, there are several household and business surveys at the annual, quarterly, monthly, and sometimes even weekly frequency.

1.2 Data handling

Raw data are normally not very useful in that we must normally do some data manipulation before carrying out any piece of statistical analysis. Summarizing the data is the primary tool for this end. It allows us not only to assess how reliable the data are, but also to understand the main features of the data. Accordingly, it is the first step of any sensible data analysis.

Summarizing data is not only about number crunching. Actually, the first task to transform numbers into valuable information is invariably to graphically represent the data. A couple of simple graphs do wonders in describing the most salient features of the data. For example, pie charts are essential to answer questions relating to proportions and fractions. For instance, the riskiness of a portfolio typically depends on how much investment there is in the risk-free asset relative to the overall investment in risky assets such as those in the equity, commodities, and bond markets. Similarly, it is paramount to map the source of problems resulting in a warranty claim so as to ensure that design and production managers focus their improvement efforts on the right components of the product or production process.

The second step is to find the typical values of the data. It is important to know, for example, what is the average income of the households in a given residential neighborhood if you wish to open a high-end restaurant there. Averages are not sufficient though, for interest may sometimes lie on atypical values. It is very important to understand the probability of rare events in risk management. The insurance industry is much more concerned with extreme (rare) events than with averages.

The next step is to examine the variation in the data. For instance, one of the main tenets of modern finance relates to the risk-return tradeoff, where we normally gauge the riskiness of a portfolio by looking at how much the returns vary in magnitude relative to their average value. In quality control, we may improve the process by raising the average

quality of the final product as well as by reducing the quality variability. Understanding variability is also key to any statistical thinking in that it allows us to assess whether the variation we observe in the data is due to something other than random variation.

The final step is to assess whether there is any abnormal pattern in the data. For instance, it is interesting to examine nor only whether the data are symmetric around some value but also how likely it is to observe unusually high values that are relatively distant from the bulk of data.

1.3 Probability and statistical inference

It is very difficult to get data for the whole population. It is very often the case that it is too costly to gather a complete data set about a subset of characteristics in a population, either because of economic reasons or because of the computational burden. For instance, it is impossible for a firm that produces millions and millions of nails every day to check each one of their nails for quality control. This means that, in most instances, we will have to examine data coming from a sample of the population.

As a sample is just a glimpse of the entire population, it will entail some degree of uncertainty to the statistical problem. To ensure that we are able to deal with this uncertainty, it is very important to sample the data from its population in a **random** manner, otherwise some sort of selection bias might arise in the resulting data sample. For instance, if you wish to assess the performance of the hedge fund industry, it does not suffice to collect data about living hedge funds. We must also collect data on extinct funds for otherwise our database will be biased towards successful hedge funds. This sort of selection bias is also known as survivorship bias.

The random nature of a sample is what makes data variability so important. Probability theory essentially aims to study how this sampling variation affects statistical inference, improving our understanding how reliable our inference is. In addition, inference theory is one of the main quality-control tools in that it allows to assess whether a salient pattern in data is indeed genuine beyond reasonable random variation. For instance, some equity fund managers boast to have positive returns for a number of consecutive periods as if this would entail unrefutable evidence of genuine stock-picking ability. However, in a universe of thousands and thousands of equity funds, it is more than natural that, due to sheer luck, a few will enjoy several periods of positive returns even if the stock returns are symmetric around zero, taking positive and negative values with equal likelihood.

Chapter 2

Data description

The first step of data analysis is to summarize the data by drawing plots and charts as well as by computing some descriptive statistics. These tools essentially aim to provide a better understanding of how frequent the distinct data values are, and of how much variability there is around a typical value in the data.

2.1 Data distribution

It is well known that a picture tells more than a million words. The same applies to any serious data analysis for graphs are certainly among the best and most convenient data descriptors. We start with a very simple, though extremely useful, type of data plot that reveals the frequency at which any given data value (or interval) appears in the sample. A frequency table reports the number of times that a given observation occurs or, if based on relative terms, the frequency of that value divided by the number of observations in the sample.

Example A firm in the transformation industry classifies the individuals at managerial positions according to their university degree. There are currently 1 accountant, 3 administrators, 4 economists, 7 engineers, 2 lawyers, and 1 physicist. The corresponding frequency table is as follows.

degree	accounting	business	economics	engineering	law	physics
value	1	2	3	4	5	6
counts	1	3	4	7	2	1
relative frequency	1/18	1/6	2/9	7/18	1/9	1/18

Note that the degree subject that a manager holds is of a qualitative nature, and so it is not particularly meaningful if one associates a number to each one of these degrees. The above table does so in the row reading 'value' according to the alphabetical order, for instance.

The corresponding plot for this type of categorical data is the bar chart. Figure 2.1 plots a bar chart using the degrees data in the above example. This is the easiest way to identify particular shapes of the distribution of values, especially concerning data dispersion. Least data concentration occurs if the envelope of the bars forms a rectangle in that every data value appears at approximately the same frequency.

In statistical quality control, one very often employs bar charts to illustrate the reasons for quality failures (in order of importance, i.e., frequency). These bar charts (also known as Pareto charts in this particular case) are indeed very popular for highlighting the natural focus points for quality improvement.

Bar charts are clearly designed to describe the distribution of categorical data. In a similar vein, histograms are the easiest graphical tool for assessing the distribution of quantitative data. It is often the case that one must first group the data into intervals before plotting a histogram. In contrast to bar charts, histogram bins are contiguous, respecting some sort of scale.

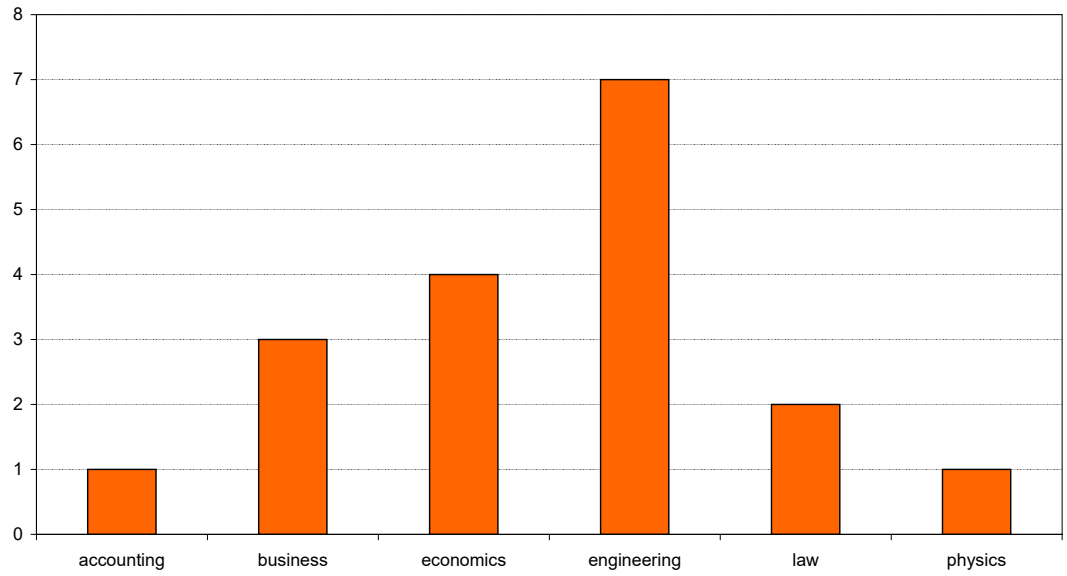

Figure 2.1: Bar chart of managers' degree subjects

2.2 Typical values

There are three popular measures of central tendency: mode, mean, and median. The mode refers to the most frequent observation in the sample. If a variable may take a large number of values, it is then convenient to group the data into intervals. In this instance, we define the

mode as the midpoint of the most frequent interval. Even though the mode is a very intuitive measure of central tendency, it is very sensitive to changes, even if only marginal, in data values or in the interval definition. The mean is the most commonly-used type of average and so it is often referred to simply as the average. The mean of a set of numbers is the sum of all of the elements in the set divided by the number of elements: i.e., $\bar{X}_N = \frac{1}{N}\sum_{i=1}^{N} X_i$. If the set is a statistical population, then we call it a population mean or expected value. If the data set is a sample of the population, we call the resulting statistic a sample mean. Finally, we define the median as the number separating the higher half of a sample/population from the lower half. We can compute the median of a finite set of numbers by sorting all the observations from lowest value to highest value and picking the middle one.

Example Consider a sample of MBA graduates, whose first salaries (in \$1,000 per annum) after graduating were as follows.

75	86	86	87	89	95	95	95	95	95
96	96	96	97	97	97	97	98	98	99
99	99	99	100	100	100	105	110	110	110
115	120	122	125	132	135	140	150	150	160
165	170	172	175	185	190	200	250	250	300

The mean salary is about \$126,140 per annum, whereas the median figure is exactly \$100,000 and the mode amounts to \$95,000. Now, if one groups the data into 8 evenly distributed bins between the minimum and maximum values, both the median and mode converge to same value of about \$91,000 (i.e., the midpoint of the second bin).

The mean value plays a major role in statistics. Although the median has several advantages over the mean, the latter is easier to manipulate for it involves a simple linear combination of the data rather than a non-differentiable function of the data as the median. In statistical quality control, for instance, it is very common to display a means chart (also known as x-bar chart), which essentially plots the mean of a variable through time. We

say that a process is in statistical control if the means vary randomly but in a stable fashion, whereas it is out of statistical control if the plot shows either a dramatic variation or systematic changes.

2.3 Measures of dispersion

While measures of central tendency are useful to understand what are the typical values of the data, measures of dispersion are important to describe the scatter of the data or, equivalently, data variability with respect to the central tendency. Two distinct samples may have the same mean or median, but different levels of variability, or vice-versa. A proper description of data set should always include both of these characteristics. There are various measures of dispersion, each with its own set of advantages and disadvantages.

We first define the sample range as the difference between the largest and smallest values in the sample. This is one of the simplest measures of variability to calculate. However, it depends only on the most extreme values of the sample, and hence it is very sensitive to outliers and atypical observations. In addition, it also provides no information whatsoever about the distribution of the remaining data points. To circumvent this problem, we may think of computing the interquartile range by taking the difference between the third and first quartiles of the distribution (i.e., subtracting the 25th percentile from the 75th percentile). This is not only a pretty good indicator of the spread in the center region of the data, but it is also much more resistant to extreme values than the sample range.

We now turn our attention to the median absolute deviation, which renders a more comprehensive alternative to the interquartile range by incorporating at least partially the information from all data points in the sample. We compute the median absolute deviation by means of $\mathrm{md}\,|X_i - \mathrm{md}(X)|$, where $\mathrm{md}(\cdot)$ denotes the median operator, yielding a very robust measure of dispersion to aberrant values in the sample. Finally, the most popular measure of dependence is the sample standard deviation as defined by the square root of the sample variance: i.e., $s_N = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \bar{X}_N\right)^2}$, where $\bar{X}_N$ is the sample mean.

The main advantage of variance-based measures of dispersion is that they are functions of a sample mean. In particular, the sample variance is the sample mean of the square of the deviations relative to the sample mean.

Example Consider the sample of MBA graduates from the previous example. The variance of their first salary after graduating is about \$2,288,400,000 per annum, whereas the standard deviation is \$47,837. The range is much larger, amounting to $300,000 - 75,000 = 225,000$ per annum. The huge difference between these two measures of dispersion suggests the presence of extreme values in the data. The fact that the interquartile range is $\frac{150{,}000+150{,}000}{2} - \frac{96{,}000+96{,}000}{2} = 54,000$—and hence closer the the standard deviation—seems to corroborate this interpretation. Finally, the median absolute deviation of the sample is only 10,000 indicating that the aberrant values of the sample are among the largest (rather than smallest) values.

In statistical quality control, it is also useful to plot some measures of dispersion over time. The most common are the R and S charts, which respectively depict how the range and the standard deviation vary over time. The standard deviation is also informative in a means chart for the interval [mean value ± two standard deviations] contains about 95% of the data if their histogram is approximately bell-shaped (symmetric with a single peak). An alternative is to plot control limits at the mean value ± three standard deviations, which should include all of the data inside. These procedures are very useful in that they reduce the likelihood of a manager to go fire-fighting every short-term variation in the means chart. Only variations that are very likely to reflect something out of control will fall outside the control limits.

A well-designed statistical quality-control system should take both means and dispersion charts into account for it is possible to improve on quality by reducing variability and/or by increasing average quality. For instance, a chef that reduces cooking time on average by 5 minutes, with 90% of the dishes arriving 10 minutes earlier and 10% arriving 40 minutes later, will probably not make the owner of the restaurant very happy.

Chapter 3

Basic principles of probability

3.1 Set theory

There are two fundamental sets, namely, the universe $\mathbb{U}$ and the empty set $\emptyset$. We say they are fundamental because $\emptyset \subseteq A \subseteq \mathbb{U}$ for every set A.

Taking the difference between sets A and B yields a set whose elements are in A but not in B: $A - B = \{x \,|\, x \in A \text{ and } x \notin B\}$. Note that $A - B$ is not necessarily the same as $B - A$. The union of A and B results in a set whose elements are in A or in B: $A \cup B = \{x \,|\, x \in A \text{ or } x \in B\}$. Naturally, if an element x belongs to both A and B, then it is also in the union $A \cup B$. In turn, the intersection of A and B individuates only the elements that both sets share in common: $A \cap B = \{x \,|\, x \in A \text{ and } x \in B\}$. Last but not least, the complement $\bar{A}$ of A defines a set with all elements in the universe that are not in A, that is to say, $\bar{A} = \mathbb{U} - A = \{x \,|\, x \notin A\}$.

Example Suppose that you roll a die and take note of the resulting value. The universe is the set with all possible values, namely, $\mathbb{U} = \{1, 2, 3, 4, 5, 6\}$. Consider the following two sets: $A = \{1, 2, 3, 4\}$ and $B = \{2, 4, 6\}$. It then follows that $A - B = \{1, 3\}$, $B - A = \{6\}$, $A \cup B = \{1, 2, 3, 4, 6\}$, and $A \cap B = \{2, 4\}$.

If A and B are complementing sets, i.e., $A = \bar{B}$, then $A - B = A$, $B - A = B$, $A \cup B = \mathbb{U}$, and $A \cap B = \emptyset$. Figure 3.1 illustrates how one may represent sets using a Venn diagram.

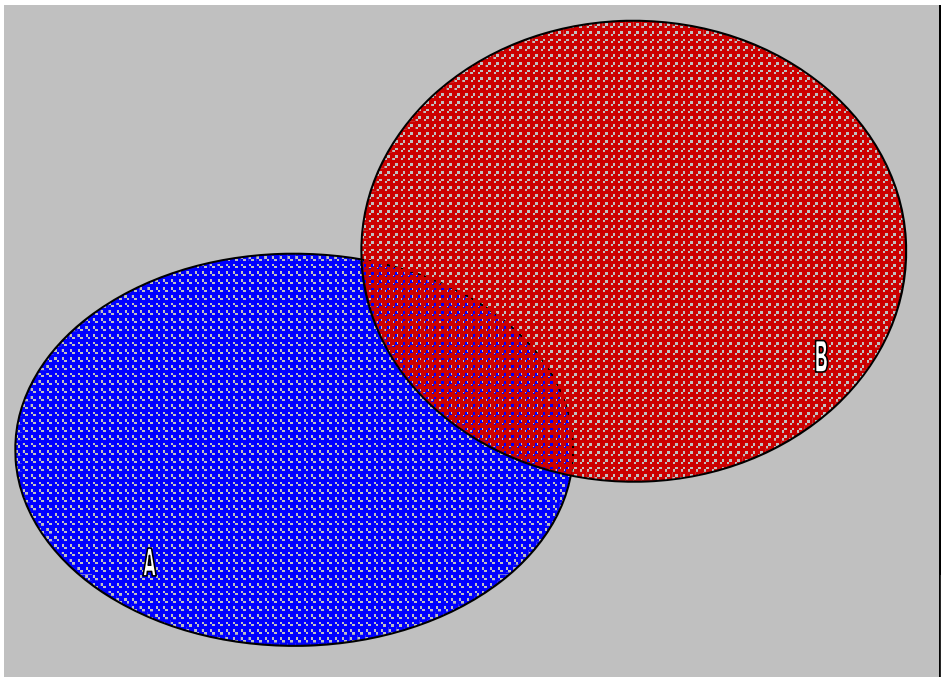

Figure 3.1: Venn diagram representing sets A (oval in blue and purple) and B (oval in red and purple) within the universe (rectangle box). The intersection $A \cap B$ of A and B is in purple, whereas the overall area in color (i.e., red, blue, and purple) corresponds to the union set $A \cup B$. The complement of A consists of the areas in grey and red, whereas the areas in grey and blue define the complement of B.

Properties The union and intersection operators are symmetric in that $A \cup B = B \cup A$ and $A \cap B = B \cap A$. They are also transitive in that $(A \cup B) \cup C = A \cup (B \cup C)$ and $(A \cap B) \cap C = A \cap (B \cap C)$.

From the above properties, it is straightforward to show that the following identities hold: (I1) $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$, (I2) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$, (I3) $A \cap \emptyset = \emptyset$, (I4) $A \cup \emptyset = A$, (I5) $\overline{A \cap B} = \bar{A} \cup \bar{B}$, (I6) $\overline{A \cup B} = \bar{A} \cap \bar{B}$, and (I7) $A = \overline{\bar{A}}$.

3.2 From set theory to probability

The probability counterpart for the universe in set theory is the sample space $\mathcal{S}$. Similarly, probability focus on events, which are subsets of possible outcomes in the sample space.

Example Suppose we wish to compute the probability of getting an even value in a die roll. The sample space is the universe of possible outcomes $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$, whereas the event of interest corresponds to the set $\{2, 4, 6\}$.

To combine events, we employ the same rules as for sets. Accordingly, the event $A \cup B$ occurs if and only if we observe an outcome that belongs to A **or** to B, whereas the event $A \cap B$ occurs if and only if both A **and** B happen. It is also straightforward to combine more than two events in that $\cup_{i=1}^{n} A_i$ occurs if and only if at least one of the events A_i happens, whereas $\cap_{i=1}^{n} A_i$ holds if and only if every event A_i occur for $i = 1, \ldots, n$. In the same vein, the event $\bar{A}$ occurs if and only if we do not observe any outcome that belongs to the event A. Finally, we say that two events are mutually exclusive if $A \cap B = \emptyset$, that is to say, they never occur at the same time. Mutually exclusive events are analogous to mutually exclusive sets in that their intersection is null.

3.2.1 Relative frequency

Suppose we repeat a given experiment n times and count how many times, say n_A and n_B, the events A and B occur, respectively. It then follows that the relative frequency of event A is $f_A = n_A/n$, whereas it is $f_B = n_B/n$ for event B. In addition, if events A and B are mutually exclusive (i.e., $A \cap B = \emptyset$), then the relative frequency of $C = A \cup B$ is $f_C = (n_A + n_B)/n = f_A + f_B$.

The relative frequency of any event is always between zero and one. Zero corresponds to an event that never occurs, whereas a relative frequency of one means that we always observe that particular event. The relative frequency is very important for the fundamental law of statistics (also known as the Glivenko-Cantelli theorem) says that, as the number of experiments n grows to infinity, it converges to the probability of the event: $f_A \to \Pr(A)$. Chapter 5 discusses this convergence in more details.

Example The Glivenko-Cantelli theorem is the principle underlying many sport competitions. The NBA play-offs are a good example. To ensure that the team with the best odds succeed, the playoffs are such that a team must win a given number of games against the same adversary before qualifying to the next round.

3.2.2 Event probability

It now remains to define what we exactly mean with the notion of probability. We associate a real number to the probability of observing the event A, denoted by $\Pr(A)$, satisfying the following properties:

P1 $0 \leq \Pr(A) \leq 1$;

P2 $\Pr(\mathcal{S}) = 1$;

P3 $\Pr(A \cup B) = \Pr(A) + \Pr(B)$ if $A \cap B = \emptyset$;

P4 $\Pr(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} \Pr(A_i)$ if the collection of events $\{A_i,\ i = 1, \ldots, n\}$ is pairwise mutually exclusive even if $n \to \infty$.

It is easy to see that **P4** follows immediately from **P3** if we restrict attention to a finite number of experiments ($n < \infty$). From properties **P1** to **P4**, it is possible to derive some important results concerning the different ways we may combine events.

Result It follows from **P1** to **P4** that

(a) $\Pr(\emptyset) = 0$,

(b) $\Pr(\bar{A}) = 1 - \Pr(A)$,

(c) $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$, and

(d) $\Pr(A) \leq \Pr(B)$ if $A \subseteq B$.

Proof: **(a)** By definition, the probability of event A is the same as the probability of the union of A and $\emptyset$, viz. $\Pr(A) = \Pr(A \cup \emptyset)$. However, A and $\emptyset$ are mutually exclusive events in that $A \cap \emptyset = \emptyset$, implying that $\Pr(A) = \Pr(A) + \Pr(\emptyset)$ by **P3**. **(b)** By definition, $A \cup \bar{A} = \mathcal{S}$ and $A \cap \bar{A} = \emptyset$, and so $\Pr(\mathcal{S}) = \Pr(A \cup \bar{A}) = \Pr(A) + \Pr(\bar{A}) = 1$ by **P2** and **P3**. **(c)** It is straightforward to observe that $A \cup B = A \cup (B \cap \bar{A})$ and that $A \cap (B \cap \bar{A}) = \emptyset$ for the event within parentheses consists of all outcomes in B that are not in A. It thus ensues that $\Pr(A \cup B) = \Pr\Big(A \cup (B \cap \bar{A})\Big) = \Pr(A) + \Pr(B \cap \bar{A})$. We now decompose the event B into outcomes that belong and not belong to A: $B = (A \cap B) \cup (B \cap \bar{A})$. There is no intersection between these two terms, hence $\Pr(B) - \Pr(A \cap B) = \Pr(B \cap \bar{A})$, yielding the result. **(d)** The previous decomposition reduces to $B = A \cup (B \cap \bar{A})$ given that $A \cap B = A$. It then follows that $\Pr(B) = \Pr(A) + \Pr(B \cap \bar{A}) \leq \Pr(A)$ in view that any probability is nonnegative. ■

3.2.3 Finite sample space

A finite sample space must have only a finite number of elements, say, $\{a_1, a_2, \ldots, a_n\}$. Let p_j denote the probability of observing the corresponding event $\{a_j\}$, for $j = 1, \ldots, n$. It is easy to appreciate that $0 \leq p_j \leq 1$ for all $j = 1, \ldots, n$ and that $\sum_{j=1}^{n} p_j = 1$ given that the

events $(a_1, \ldots, a_n)$ span the whole sample space. As the latter are also mutually exclusive, it follows that $\Pr(A) = p_{j_1} + \ldots, + p_{j_k} = \sum_{r=1}^{k} p_{j_r}$ for $A = \{a_{j_1}, \ldots, a_{j_k}\}$, with $1 \leq k \leq n$.

Example: The sample space corresponding to the value we obtain by throwing a die is $\{1, 2, 3, 4, 5, 6\}$ and the probability p_j of observing any value $j \in \{1, \ldots, 6\}$ is equal to $1/6$.

In general, if every element in the sample space is equiprobable, then the probability of observing a given event is equal to the ratio between the number of elements in the event and the number of elements in the sample space.

Examples

(1) Suppose the interest lies on the event of observing a value above 4 in a die throw. There are only two values in the sample space that satisfy this condition, namely, $\{5, 6\}$, and hence the probability of this event is $2/6 = 1/3$.

(2) Consider now flipping twice a coin and recording the heads and tails. The resulting sample space is $\{HH, HT, TH, TT\}$. As the elements of the sample space are equiprobable, the probability of observing only one head is $\frac{\#\{HT,TH\}}{\#\{HH,HT,TH,TT\}} = 2/4 = 1/2$.

These examples suggest that the most straightforward manner to compute the probability of a given event is to run experiments in which the elements of the sample space are equiprobable. Needless to say, it is not always very easy to contrive such experiments. We illustrate this issue with another example.

Example: Suppose one takes a nail from a box containing nails of three different sizes. It is typically easier to grab a larger nail than a small one and hence such an experiment would not yield equiprobable outcomes. However, the alternative experiment in which we first numerate the nails and then draw randomly a number to decide which nail to take would lead to equiprobable results.

3.2.4 Back to the basics: Learning how to count

The last example of the previous section illustrates a situation in which it is straightforward to redesign the experiment so as to induce equiprobable outcomes. Life is tough, though, and such an instance is the exception rather than the rule. For instance, a very common problem in quality control is to infer from a small random sample the probability of observing a given number of defective goods within a lot. This is evidently a situation that does not automatically lead to equiprobable outcomes given the sequential nature of the experiment. To deal with such a situation, we must first learn how to count the possible outcomes using some tools of combinatorics.

Multiplication Consider that an experiment consists of a sequence of two procedures, say, A and B. Let n_A and n_B denote the number of ways in which one can execute A and B, respectively. It then follows that there is $n = n_A\, n_B$ ways of executing such an experiment. In general, if the experiment consists of a sequence of k procedures, then one may run it in $n = \prod_{i=1}^{k} n_i$ different ways.

Addition Suppose now that the experiment involves k procedures in parallel (rather than in sequence). This means that we either execute the procedure 1 or the procedure 2 or ... or the procedure k. If n_i denotes the number of ways that one may carry out the procedure $i \in \{1, \ldots, k\}$, then there are $n = n_1 + \cdots + n_k = \sum_{i=1}^{k} n_i$ ways of running such an experiment.

Permutation Suppose now that we have a set of n different elements and we wish to know the number of sequences we can construct containing each element once, and only once. Note that the concept of sequence is distinct from that of a set, in that order of appearance matters. For instance, the sample space $\{a, b, c\}$ allows for the following permutations $(abc, acb, bac, bca, cab, cba)$. In general, there are $n! = \prod_{j=0}^{n-1}(n-j)$ possible permutations out of n elements because there are n options for the first element of the sequence, but only $n-1$ options for the second element, $n-2$ options for the third element and so on until we have only one remaining option for the last element of the sequence. There is also a more general meaning for permutation in combinatorics for which we form sequences of k different elements from a set of n elements. This means that we have n options for the first element of the sequence, but then $n-1$ options for the second element and so on until we have only $n-k+1$ options for the last element of the sequence. It thus follows that we have $n!/(n-k)!$ permutations of k out of n elements in this broader sense.

Combination This is a notion that only differs from permutation in that ordering does not matter. This means that we just wish to know how many subsets of k elements we can construct out of a set of n elements. For instance, it is possible to form the following subsets with two elements of $\{a, b, c, d\}$: $\{a, b\}$, $\{a, c\}$, $\{a, d\}$, $\{b, c\}$, $\{b, d\}$, and $\{c, d\}$. Note that $\{b, a\}$ does not count because it is exactly the same subset as $\{a, b\}$. This suggests that, in general, the number of combinations is inferior to the number of permutations because one must count only one of the sequences that employ the same elements but with a different ordering. In view that there are $n!/(n-r)!$ permutations of k out of n elements and $k!$ ways

to choose the ordering of these k elements, the number of possible combinations of k out of n elements is

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

Before we revisit the original quality control example, it is convenient to illustrate the use of the above combinatoric tools through another example.

Example: Suppose there is a syndicate with 5 engineers and 3 economists. How many committees of 3 people one can form with exactly 2 engineers? Well, we must form committees of 2 engineers and 1 economist. There are $\binom{5}{2}$ ways of choosing 2 out of 5 engineers, whereas there are $\binom{3}{1}$ ways of choosing 1 out of 3 economists. Altogether, this means that one can form $\binom{5}{2}\binom{3}{1} = 30$ committees with 2 engineers and 1 economist out of a group of 5 engineers and 3 economists.

We are now ready to reconsider the quality control problem of inferring the number of defective goods within a lot. Suppose, for instance, that a lot has n objects of which n_d are defective and that we draw a sample of k elements of which k_d are defective. We first note that there are $\binom{n}{k}$ ways of choosing k elements from a lot of n goods, whereas there are $\binom{n_d}{k_d}$ ways of combining k_d defective goods from a total of n_d defective goods within the lot as well as $\binom{n-n_d}{k-k_d}$ ways of choosing $(k-k_d)$ elements out of the $(n-n_d)$ non-defective goods within the lot. Accordingly, the probability of observing k_d defective goods within a sample of k goods is

$$\frac{\binom{k}{k_d}\binom{n-n_d}{k-k_d}}{\binom{n}{k}} \tag{3.1}$$

if there are n_d defective goods within a lot of n objects.

3.2.5 Conditional probability

We denote by $\Pr(A|B)$ the probability of event A given that we have already observed event B. Intuitively, conditioning on the realization of a given event has the effect of reducing the

sample space from $\mathcal{S}$ to the sample space spanned by B.

Examples

(1) Suppose that we throw a die twice. In the first throw, we observe a value equal to 6 and we wish to know what is the probability of observing a value of 2 in the second throw. In this instance, the fact that we have observed a value of 6 in the first throw has no impact in the value we will observe in the second throw for the two events are independent. This means that the first value brings about no information about the second throw and hence the probability of observing a value of 2 in the second throw given that we have observed a value of 6 in the first throw remains the same as before, that is to say, the probability of observing a value of 2: 1/6.

(2) Next, consider $A = \{(x_1, x_2)|x_1 + x_2 = 10\} = \{(5,5),(6,4),(4,6)\}$ and $B = \{(x_1, x_2)|x_1 > x_2\} = \{(2,1),(3,2),(3,1),\cdots,(6,5)\}$. The probability of A is $\Pr(A) = 3/36 = 1/12$, whereas the probability of B is $\Pr(B) = 15/36$. In addition, the probability of observing both A and B is $\Pr(A \cap B) = 1/36$. It thus turns out that the probability of observing A given B is $\Pr(A|B) = 1/15 = \Pr(A \cap B)/\Pr(B)$, whereas the probability of observing B given A is $\Pr(B|A) = 1/3 = \Pr(A \cap B)/\Pr(A)$.

It is obviously not by chance that, in general, $\Pr(A|B) = \Pr(A \cap B)/\Pr(B)$, for $\Pr(B) > 0$. By conditioning on event B we are restricting the sample space to B and hence we must consider the probability of observing both A and B and then normalize by the measure of event B. It is as if we were computing the relative frequency at which the event A occurs given the outcomes that are possible within event B. This notion makes sense even if we consider unconditional events. Indeed, the unconditional probability of A is the conditional probability of A given the sample space $\mathcal{S}$, i.e., $\Pr(A|\mathcal{S}) = \Pr(A \cap \mathcal{S})/\Pr(\mathcal{S}) = \Pr(A)$. Finally, it is also interesting to note that we may decompose the probability of $A \cap B$ into a conditional probability and a marginal probability, namely, $\Pr(A \cap B) = \Pr(A|B)\Pr(B) = \Pr(B|A)\Pr(A)$.

Example: Suppose that a computer lab has 4 new and 2 old desktops running Windows as well as 3 new and 1 old desktops running Linux. What is the probability of a student to randomly sit in front of a desktop running Windows? What is the likelihood that this particular desktop is new given that it runs Windows? Well, there are 10 computers in the lab of which 6 run Windows. This means that the answer of the first question is 3/5, whereas $\Pr(\text{new}|\text{Windows}) = \Pr(\text{new} \cap \text{Windows})/\Pr(\text{Windows}) = (4/10)/(6/10) = 2/3$.

Figure 3.2.5 illustrates a situation in which the events A and B are mutually exclusive and hence $A \cap B = \emptyset$. In this instance, the probability of both events occurring is obviously zero and so are both conditional probabilities, i.e., $\Pr(A \cap B) = 0 \Rightarrow \Pr(A|B) = \Pr(B|A) = 0$. In

contrast, Figure 3.3 depicts another polar case: $A \subset B$. Now, $\Pr(A \cap B) = \Pr(A)$, whereas $\Pr(A|B) = \Pr(A)/\Pr(B)$ and $\Pr(B|A) = 1$.

Decomposing a joint probability into the product of a conditional probability and of a marginal probability is a very useful tool, especially if one combines it with partitions of the sample space. Let $B_1, \ldots, B_k$ denote a partition of the sample space $\mathcal{S}$, that is to say,

(a) $B_i \cap B_j = \emptyset, \qquad 1 \leq i \neq j \leq k$

(b) $\cup_{i=1}^{k} B_i = \mathcal{S}$

(c) $\Pr(B_i) > 0, \qquad 1 \leq i \leq k.$

This partition yields the decomposition $A = (A \cap B_1) \cup (A \cap B_2) \cup \ldots \cup (A \cap B_k)$ for any event $A \in \mathcal{S}$. The nice thing about partitions is that they are mutually exclusive and hence $(A \cap B_i) \cap (A \cap B_j) = \emptyset$ for any $1 \leq i \neq j \leq k$. This means that $\Pr(A) = \sum_{i=1}^{k} \Pr(A \cap B_i) = \sum_{k}^{i=1} \Pr(A|B_i) \Pr(B_i)$.

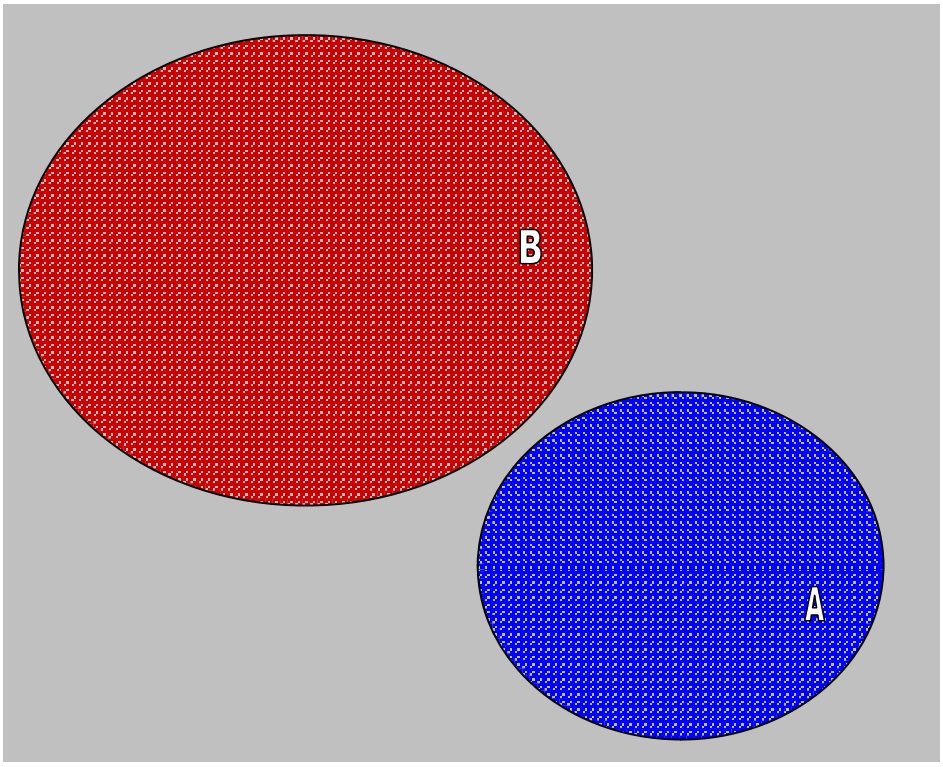

Figure 3.2: Venn diagram representing two mutually exclusive events A (oval in blue) and B (oval in red) within the sample space (rectangle box).

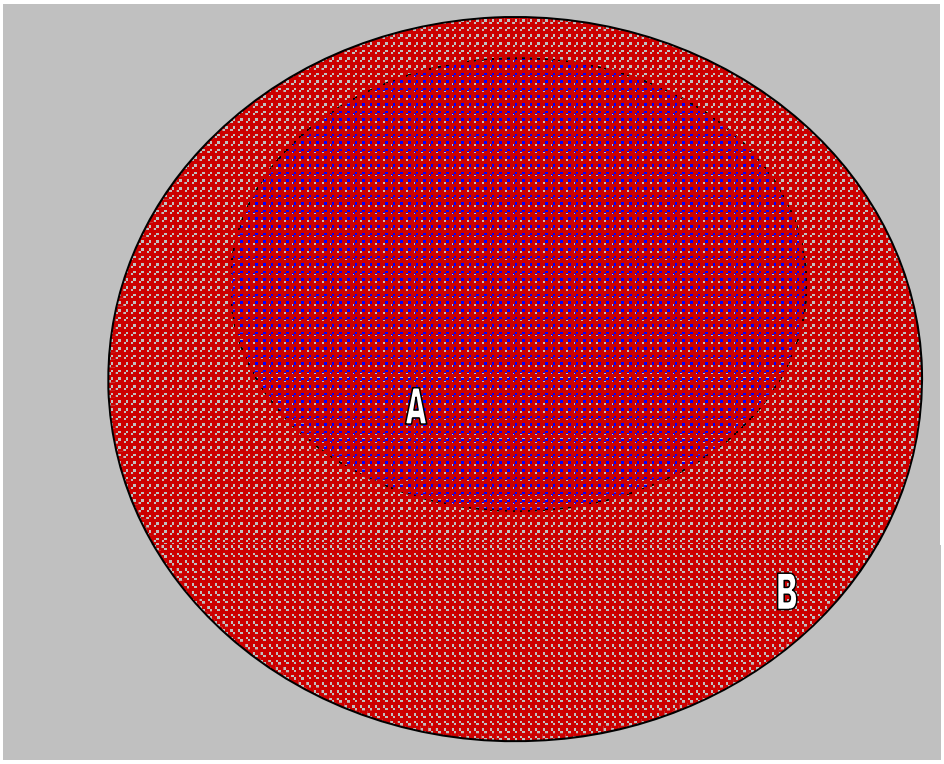

Figure 3.3: Venn diagram representing events A (oval in purple) and B (oval in red and purple) within the sample space (rectangle box) such that $A \cap B = A$.

For instance, if we define the sample space by the possible outcomes of a die throw, we may think of several distinct partitions as, for example,

(a) $B_i = \{i\}$ for $i = 1, \ldots, 6$

(b) $B_1 = \{1, 3, 5\}, B_2 = \{2, 4, 6\}$

(c) $B_1 = \{1, 2\}, B_2 = \{3, 4, 5\}, B_3 = \{6\}$.

Example: Consider a lot of 100 frying pans of which 20 are defective. Define the events $A = \{\text{first frying pan is defective}\}$ and $B = \{\text{second frying pan is defective}\}$ within a context of sequential sampling without reposition. The probability of observing event B naturally depends on whether the first frying pan is defective or not. Now, there are only two possible outcomes in that the first frying pan is either defective or not. This suggests a very simple partition of the sample space based on A and $\bar{A}$, giving way to $\Pr(B) = \Pr(B|A)\Pr(A) + \Pr(B|\bar{A})\Pr(\bar{A})$. In particular, $\Pr(B|A) = 19/99$ for there are only 19 defective frying pans left among the remaining 99 frying pans if A is true. Similarly, $\Pr(B|\bar{A}) = 20/99$, whereas $\Pr(A) = 1/5$ and $\Pr(\bar{A}) = 1 - \Pr(A) = 4/5$. We thus conclude that $\Pr(B) = \frac{19}{99}\frac{1}{5} + \frac{20}{99}\frac{4}{5} = \frac{1}{5}$.

In some instances, we cannot observe some events, and hence we must infer whether they are true or false given the available information. For instance, if you are in a building with no windows and someone arrives completely soaked with a broken umbrella, it sounds reasonable to infer that it is raining outside even if you cannot directly observe the weather. The Bayes rule formalizes how one should conduct such an inference based on conditional probabilities:

$$\Pr(B_i|A) = \frac{\Pr(A|B_i)\Pr(B_i)}{\sum_{j=1}^{k}\Pr(A|B_j)\Pr(B_j)} \qquad i = 1, \cdots, k$$

where $B_1, \ldots, B_k$ is a partition of the sample space. In the example above, we cannot observe whether it is raining, but we may partition the sample space (i.e., weather) into $B = \{\text{it is raining}\}$ and $\bar{B} = \{\text{it is not raining}\}$, and then calculate the probability of B given that we observe event $A = \{\text{someone arrives completely soaked with a broken umbrella}\}$.

The Bayes rule has innumerable applications in business, economics and finance. For instance, imagine you are the market maker for a given stock and that there are both informed and uninformed traders in the market. In contrast to informed traders, you do not know whether news are good or bad and hence you must infer it from the trades you observe in order to adjust your bid and ask quotes accordingly. If you observe much more traders buying than selling, then you will assign a higher probability to good news. If traders are selling much more than buying, then the likelihood of bad news rises. The Bayes rule is the mechanism at which you learn whether news are good or bad by looking at trades.

3.2.6 Independent events

Consider for a moment two mutually exclusive events A and B. Knowing about A gives loads of information about the likelihood of event B. In particular, if A occurs, we know for sure that event B did not occur. More formally, the conditional probability of B given that we observe A is $\Pr(B|A) = \Pr(A \cap B)/\Pr(A) = 0$ given that $A \cap B = \emptyset$ (see Figure 3.2.5). We thus conclude that A and B are dependent events given that knowing about one entails complete information about the other. Following this reasoning, it makes sense to associate independence with lack of information content. We thus say that A and B are independent events if and only if $\Pr(A|B) = \Pr(A)$. The latter condition means that $\Pr(A \cap B) = \Pr(A|B)\Pr(B) = \Pr(A)\Pr(B)$, which in turn is equivalent to say that $\Pr(B|A) = \Pr(B)$ given that $\Pr(A \cap B) = \Pr(B|A)\Pr(A)$ as well. Intuitively, if A and B are independent, the probability of observing A (or B) does not depend on whether B (or A has occurred) and hence conditioning on the sample space (i.e., looking at the unconditional distribution) or on the event B makes no difference.

Example: Consider a lot of 10,000 pipes of which 10% comes with some sort of indentation. Suppose we randomly draw two pipes from the lot and define the events $A_1 = \{\text{first pipe is in perfect conditions}\}$ and $A_2 = \{\text{second pipe is in perfect conditions}\}$. If sampling is with reposition, then events A_1 and A_2 are independent and so $\Pr(A_1 \cap A_2) =$

$\Pr(A_1)\Pr(A_2) = (0.9)^2 = 0.81$. However, if sampling is without reposition, then $\Pr(A_1 \cap A_2) = \Pr(A_2|A_1)\Pr(A_1) = 0.9\,\frac{8{,}999}{9{,}999}$, which is very marginally different from 0.81.

This example illustrates well a situation in which the events are not entirely independent, though assuming independence would simplify a lot the computation of the joint probability at the expenses of a very marginal cost due to the large sample. This is just to say that **sometimes** it pays off to assume independence between events even if we know that, in theory, they are not utterly independent.

Problem set

Exercise 1. Show that

$$\begin{aligned}\Pr(A \cup B \cup C) = \Pr(A) + \Pr(B) + \Pr(C) \\ - \Pr(A \cap B) - \Pr(A \cap C) - \Pr(B \cap C) \\ + \Pr(A \cap B \cap C).\end{aligned}$$

Solution We employ a similar decomposition to the one in the proof of **(c)**. In particular, $(A \cup B) \cup C = (A \cup B) \cup (C \cap \overline{A \cup B})$. As the intersection is null,

$$\Pr(A \cup B \cup C) = \Pr(A \cup B) + \Pr(C \cap \overline{A \cup B}).$$

We now decompose the event C into outcomes that belong and not belong to $A \cup B$:

$$C = \Big(C \cap (A \cup B)\Big) \cup \Big(C \cap \overline{A \cup B}\Big),$$

yielding $\Pr\left(C \cap \overline{A \cup B}\right) = \Pr(C) - \Pr\Big(C \cap (A \cup B)\Big)$. So far, we have that

$$\begin{aligned}\Pr(A \cup B \cup C) &= \Pr(A \cup B) + \Pr(C) - \Pr\Big(C \cap (A \cup B)\Big) \\ &= \Pr(A) + \Pr(B) + \Pr(C) - \Pr(A \cap B) - \Pr\Big(C \cap (A \cup B)\Big).\end{aligned}$$

It remains to show that the last term equals to $\Pr(A \cap C) + \Pr(B \cap C) - \Pr(A \cap B \cap C)$. To appreciate this, it suffices to see that $C \cap (A \cup B) = (A \cap C) \cup (B \cap C)$, which gives way

to $\Pr\Big(C \cap (A \cup B)\Big) = \Pr(A \cap C) + \Pr(B \cap C) - \Pr\Big((A \cap C) \cap (B \cap C)\Big)$ by **P3**. The last term is obviously equivalent to $\Pr(A \cap B \cap C)$, completing the proof. ■

Exercise 2. Consider two events A e B. Show that the probability that only one of these events occurs is $\Pr(A \cup B) - \Pr(A \cap B)$.

Solution Let C denote the event in which we observe only one event between A or B. It then consists of every possible outcome that it is in $A \cup B$ and not in $A \cap B$. It is straightforward to appreciate from a Venn diagram that $C = (A \cup B) - (A \cap B) = (A \cup B) \cap \overline{A \cap B} = (A \cap \bar{B}) \cup (\bar{A} \cap B)$. The last representation is the easiest to manipulate for it involves mutually exclusive events. In particular, it follows immediately from **(c)** that $\Pr(C) = \Pr(A) + \Pr(B) - 2\Pr(A \cup B) = \Pr(A \cup B) - \Pr(A \cap B)$. ■

Exercise 3. There are three plants that produce a given screw: A, B, and C. Plant A produces the double of screws than B and C, whose productions are at par. In addition, quality control is better at plants A and B in that only 2% of the screws they produce are defective as opposed to 4% in plant C. Suppose that we sample one screw from the warehouse that collects all screws produced by A, B, and C. What is the probability that the screw is defective? What is the probability that the defective screw is from plant A?

Solution: Let $A = \{\text{screw comes from plant } A\}$, $B = \{\text{screw comes from plant } B\}$, $C = \{\text{screw comes from plant } C\}$, and $D = \{\text{screw is defective}\}$. Given that A's production is twofold, it follows that $\Pr(A) = 1/2$ and that $\Pr(B) = \Pr(C) = 1/4$. We now decompose the event D according to whether the screw comes from A, B, or C. The latter forms a partition because if a screw comes from a given plant it cannot come from any other plant. In addition, there are only plants A, B, and C producing this particular screw. The decomposition yields

$$\begin{aligned}\Pr(D) &= \Pr(D|A)\Pr(A) + \Pr(D|B)\Pr(B) + \Pr(D|C)\Pr(C) \\ &= 0.02\,\frac{1}{2} + 0.02\,\frac{1}{4} + 0.04\,\frac{1}{4} = 0.025.\end{aligned}$$

To answer the second question, we must apply the Bayes rule for we do not observe whether the screw comes from a given plant, but we do know whether it is defective or not. So, the conditional probability that the screw is from A given that it is defective is $\Pr(A|D) = \frac{0.02\times 1/2}{0.02\times 1/2+0.02\times 1/4+0.04\times 1/4} = \frac{0.01}{0.01+0.005+0.01} = 2/5$. ■

Chapter 4

Probability distributions

4.1 Random variable

Dealing with events and sample spaces is very intuitive, but it is not very easy to keep track of things if the sample space is large. That is why we next introduce the notion of random variable, which entails a much easier approach to probability theory.

Definition: $X(s)$ is a random variable if $X(\cdot)$ is a function that assigns a real value to every element s in the sample space $\mathcal{S}$.

Example: Suppose we flip twice a coin and define the sample space as the sequence of heads and tails, that is to say, $\mathcal{S} = \{HH, HT, TH, TT\}$. Let X denote a random variable equal to the number of heads:

$$X(s) = \begin{cases} 0, & \text{if } s \in \{TT\} \\ 1, & \text{if } s \in \{HT, TH\} \\ 2, & \text{if } s \in \{HH\}. \end{cases}$$

Note that every element in the sample space corresponds to exactly one value of the random variable, though the latter may assume the same value for different elements of the sample space.

4.1.1 Discrete random variable

If X is a discrete random variable, then it takes only a countable number of values. This means that, in practice, we may consider a list of possible outcomes $x_1, \ldots, x_n$ (even if $n \rightarrow$

∞) for any discrete random variable X. Denoting the probability of observing a particular value by $p_i \equiv p(x_i) = \Pr(X = x_i)$, it follows that $p_i \geq 0$ for $i = 1 \ldots, n$ and that $\sum_{i=1}^{n} p_i = 1$. The function $p(\cdot)$ is known as the probability distribution function of the discrete random variable. For instance, a random variable following a discrete uniform probability function $p_i = 1/n$, with n finite, is the random-variable counterpart of equiprobable events. Figure 4.1 displays the probability distribution function of a discrete uniform random variable over the set $\{1, 2, \ldots, 10\}$.

Example: Suppose that a mutual fund buys and holds a given stock as long as price changes are nonnegative. Let stock prices follow a random walk such that the probability of observing a negative price change is 2/5. Define the sample space $\mathcal{S}$ and the random variable N according to the number of periods that are necessary to observe the mutual fund unwinding its position: $\mathcal{S} = \{1, 01, 001, 0001, \ldots\}$ and $N = \{1, 2, 3, 4, \ldots\}$. It is easy to see that $N = n$ if and only if we observe a negative price change in the nth period after $(n-1)$ periods of nonnegative returns. In addition, the random walk hypothesis implies that returns are independent over time, and so

$$\Pr(N = n) = \left(\frac{3}{5}\right)^{n-1} \frac{2}{5} \qquad n = 1, 2, \ldots$$

Just as a sanity check, let's test whether the above probability function sums up to one if we consider every possible outcome:

$$\sum_{n=1}^{\infty} \Pr(N = n) = \frac{2}{5}\left(1 + \frac{3}{5} + \frac{9}{25} + \cdots\right) = \frac{2}{5}\frac{1}{1 - 3/5} = 1.$$

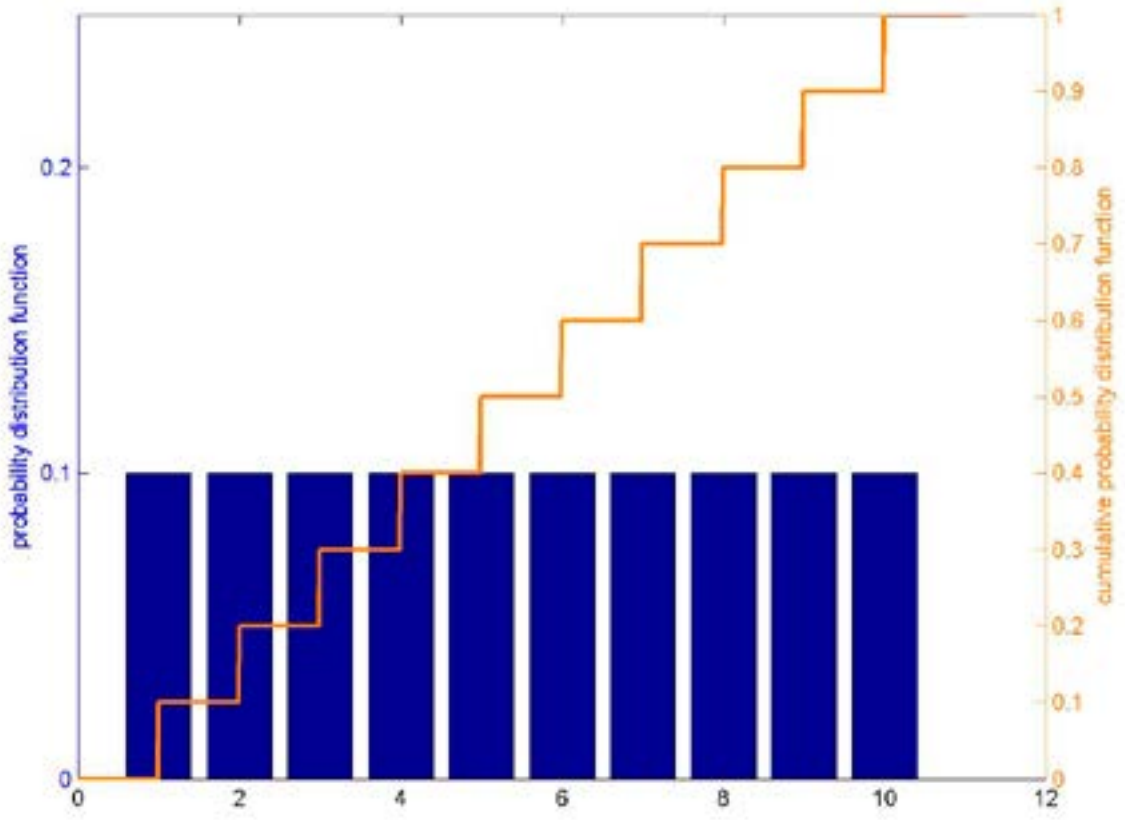

Figure 4.1: The left and right axes correspond to the probability distribution function and cumulative probability distribution function of a uniform distribution over $\{1, 2, \ldots, 10\}$, respectively.

Binomial distribution

A Bernoulli essay is the simplest and most intuitive of all probability distribution functions. It restricts attention to a binary random variable that takes value one with probability p, otherwise it is equal to zero (with probability $1-p$). Consider now a random variable that sums up the values of n independent Bernoulli essays. The probability distribution function of such a variable is by definition binomial.

Example: Suppose that a production line results in defective products with probability 0.20. A random draw of three products leads to a sample space given by

$$\mathcal{S} = \{DDD, DDN, DND, NDD, NND, NDN, DNN, NNN\},$$

where D and N refer to defective and non-defective products, respectively. The ordering does not matter much in most situations and so the typical random variable of interest is the number of defective goods $X \in \{0, 1, 2, 3\}$. The probability distribution function of X then is $p_0 = 0.8^3$, $p_1 = 3 \times 0.2 \times 0.8^2$, $p_2 = 3 \times 0.8 \times 0.2^2$, and $p_3 = 0.2^3$.

In the example above, it is readily seen from the sample space that there is only one manner to obtain either three defective goods or three non-defective goods. In contrast, there are three different ways to observe either one or two defective products due to the fact that the ordering does not matter. It is precisely the latter that explains why the binomial distribution function involves the combinatoric tool of combination.

Definition: Consider an experiment in which the event A occurs with probability $p = \Pr(A)$ and so $\Pr(\bar{A}) = 1 - p$. Run such an experiment independently n times. The resulting sample space is $\mathcal{S} = \{\text{all sequences } a_1, \ldots, a_n\}$, where a_i is either A or $\bar{A}$ for $i = 1, \ldots, n$. The random variable X that counts the number of times that the event A occurs has a binomial distribution function $\mathcal{B}(n, p)$ with parameters n (namely, the number of independent essays)

and p (namely, the probability of event A). The binomial distribution is such that

$$p_x = \Pr(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, \cdots, n. \tag{4.1}$$

To show that (4.2) is a probability distribution function, it suffices to confirm that it sums up to one given that $p_x > 0$. As expected,

$$\begin{aligned} \sum_{x=0}^{n} \Pr(X = x) &= \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} \\ &= [p + (1-p)]^n = 1, \end{aligned}$$

where the last equality comes from Newton's binomial expansion (hence the name of the distribution).

Problem set

Exercise 1. Paolo Maldini challenges Buffon for a series of 20 penalty kicks. In the first 10 penalty kicks, Maldini scores with probability 4/5. However, as from the 11th attempt, Maldini's age kicks in and the probability of scoring reduces to 1/2. Assuming that the outcomes are independent among themselves, compute the probability that Maldini scores exactly k goals.

Solution: Each penalty kick corresponds to a Bernoulli essay with probability $p_1 = 4/5$ of success in the first 10 attempts and $p_2 = 1/2$ from then on up to the 20th penalty kick. We thus split the problem into scoring k_1 goals in the first 10 attempts and $k - k_1$ goals in the second 10 penalty kicks. The former leads to a binomial distribution $\mathcal{B}(10, 4/5)$, whereas the latter to a binomial $\mathcal{B}(10, 1/2)$. Putting together gives way to

$$\begin{aligned} &\binom{10}{k_1} p_1^{k_1} (1-p_1)^{10-k_1} \times \binom{10}{k-k_1} p_2^{k-k_1} (1-p_1)^{10-k+k_1} \\ &\binom{10}{k_1} 0.8^{k_1} 0.2^{10-k_1} \times \binom{10}{k-k_1} 0.5^{10}. \end{aligned}$$

It now remains to sum up the ways at which Maldini can score k goals by scoring exactly k_1 goals in the first 10 attempts. To this end, we must first consider whether $k > 0$ or not,

yielding a probability of scoring k penalty kicks of

$$\Pr(X=k)=\sum_{k_1=\max\{0,k-10\}}^{\min\{k,10\}}\binom{10}{k_1}0.8^{k_1}\,0.2^{10-k_1}\binom{10}{k-k_1}0.5^{10}.$$

We sum from $\max\{0,k-10\}$ because if $k>10$ then Maldini should score at least $K-10$ in each series of 10 attempts. Similarly, we sum up to $\min\{k,10\}$ because if $k<10$, then Maldini cannot score more than k goals in each series of penalty kicks. ■

Exercise 2. Consider a random variable $X\in\{0,1,2,\ldots\}$ such that

$$\Pr(X=t)=(1-\alpha)\alpha^t,\qquad t=0,1,2,\cdots$$

(a) For which values of α (4.5) indeed is a probability distribution function?

(b) Show that for any positive integers s and t, $\Pr(X>s+t|X>s)=\Pr(X\geq t)$.

Solution:

(a) It follows from $\sum_{t=0}^{\infty}\Pr(X=t)=1$ that $(1-\alpha)\sum_{t=0}^{\infty}\alpha^t=1$. The latter involves an infinite sum of a geometric progression, which converges to $(1-\alpha)\sum_{t=0}^{\infty}\alpha^t=(1-\alpha)\frac{1}{1-\alpha}=1$ only if α belongs to the unit interval.

(b) It follows from the fact that $s+t\geq s>0$ that

$$\begin{aligned}\Pr(X>s+t|X>s)&=\frac{\Pr(X>s+t)}{\Pr(X>s)}=\frac{\sum_{r=s+t+1}^{\infty}(1-\alpha)\alpha^r}{\sum_{r=s+1}^{\infty}(1-\alpha)\alpha^r}\\&=\frac{\sum_{r=s+t+1}^{\infty}\alpha^r}{\sum_{r=s+1}^{\infty}\alpha^r}=\frac{(1-\alpha)^{-1}\alpha^{s+t+1}}{(1-\alpha)^{-1}\alpha^{s+1}}=\alpha^t.\end{aligned}$$

To complete the proof, it suffices to show that $\Pr(X\geq t)=\alpha^t$. That is indeed the case because

$$\begin{aligned}\Pr(X\geq t)&=\sum_{r=t}^{\infty}(1-\alpha)\alpha^r=(1-\alpha)\sum_{r=t}^{\infty}\alpha^r\\&=(1-\alpha)\,\frac{\alpha^t}{1-\alpha}=\alpha^t,\end{aligned}$$

where the penultimate equality comes from the fact that the infinite sum of a geometric progression is equal to the first term of the progression divided by one minus the quotient of the progression. ■

4.1.2 Cumulative probability distribution function

The goal not always lies on computing a pointwise probability. We are very often interested in understanding how likely it is to observe a range of values, e.g., the probability of observing $X \leq x$. This motivates us to define the cumulative distribution function as

$$F_X(x) = \Pr(X \leq x) = \sum_j p(x_j), \qquad \forall x_j \leq x.$$

Note that F_X is a nondecreasing step function in x given that $F_X(x_1) \leq F_X(x_2)$ if $x_1 \leq x_2$. In addition, the cumulative distribution function belongs to the unit interval given that $\lim_{x\to-\infty} F_X(x) = 0$ and $\lim_{x\to\infty} F_X(x) = 1$. Finally, if $X \in \{x_1, x_2, \ldots | x_1 < x_2 < \ldots\}$, then $\Pr(X = x_n) = \Pr(X \leq x_n) - \Pr(X \leq x_{n-1}) = F_X(x_n) - F_X(x_{n-1})$.

Example: Let $X \in \{x_1, x_2, x_3\}$ with $p(x_1) = \frac{1}{3}$, $p(x_2) = \frac{1}{6}$, and $p(x_3) = \frac{1}{2}$. The cumulative distribution function then reads

$$F_X(x) = \begin{cases} 0, & \text{if } -\infty < x < x_1 \\ 1/3 & \text{if } x_1 \leq x < x_2 \\ 1/2 & \text{if } x_2 \leq x < x_3 \\ 1 & \text{if } x_3 \leq x < \infty. \end{cases}$$

4.1.3 Continuous random variable

We say that a random variable is continuous if the probability of observing any particular value in the real line is zero. Accordingly, the notion of probability distribution function is meaningless and we have to come up with something a bit different, though with a similar interpretation. The analog of the probability distribution function for continuous random variables is the probability density function. To understand the latter, we first note that, within the context of continuous random variables, it only makes sense to talk about the probability of observing a value within a given interval. The probability density function then measures the mass of probability of an infinitesimal interval, that is to say, $\Pr(x < X < x + \Delta x)$ for a very small $\Delta x > 0$. In the following definition, we formalize such a notion.

Definition: The probability density function $f_X(\cdot)$ of a random variable X is such that

(a) $f_X(x) \geq 0, \qquad \forall x \in \mathbb{R}$

(b) $\int_{-\infty}^{+\infty} f_X(x)\,\mathrm{d}x = 1$

(c) $\Pr(a \leq X \leq b) = \int_a^b f_X(x)\,\mathrm{d}x$ for $-\infty < a < b < +\infty$.

Note that condition (a) corresponds to the restriction that the probability distribution function is positive for every element in the sample space, whereas (b) is analogous to the imposition that the probability distribution function sums up to one if evaluated at every element in the sample space. Finally, (c) reflects the fact that the probability of observing a particular value is zero given that $\Pr(X = x_0) = \int_{x_0}^{x_0} f_X(x)\,\mathrm{d}x = 0$. Of course, the fact that an event A has probability zero does not mean that it is impossible to observe it. It just means that it is improbable. For instance, imagine that we are measuring how much time one takes to run 100 meters. The probability of observing a value of precisely 10 seconds is zero ex-ante, for there is a continuum of values around 10 seconds, though it could well take exactly 10 seconds ex-post. A corollary is that

$$\Pr(a \leq X \leq b) = \Pr(a < X \leq b) = \Pr(a < X < b) = \Pr(a \leq X < b).$$

Examples

(1) Let a random variable satisfy the following probability density function

$$f_X(x) = \begin{cases} 2x, & \text{if } 0 < x < 1 \\ 0, & \text{otherwise.} \end{cases}$$

We first note that $f_X(x) \geq 0$ for any value of $x \in \mathbb{R}$ and that

$$\int_{-\infty}^{+\infty} f_X(x)\,\mathrm{d}x = \int_0^1 f_X(x)\,\mathrm{d}x = \int_0^1 2x\,\mathrm{d}x = x^2\big|_0^1 = 1.$$

In addition, if we wish to compute the probability of observing a given interval, say, $x \leq 1/2$, then it follows that

$$Pr\left(X \leq \frac{1}{2}\right) = \int_0^{1/2} 2x\,\mathrm{d}x = x^2\big|_0^{1/2} = \frac{1}{4}.$$

Finally, we may also compute the conditional probability of observing a value within an interval given that we know it belongs to a larger interval. For instance,

$$\Pr\left(X \leq \frac{1}{2} \,\middle|\, \frac{1}{3} \leq X \leq \frac{2}{3}\right) = \frac{\Pr\left(\frac{1}{3} \leq x \leq \frac{1}{2}\right)}{\Pr\left(\frac{1}{3} \leq x \leq \frac{2}{3}\right)} = \frac{x^2|_{1/3}^{1/2}}{x^2|_{1/3}^{2/3}} = \frac{5}{12}.$$

(2) Let X denote a random variable with density function $f_X(x) = \begin{cases} \alpha\, x^{-3}, & \text{if } 1 \leq x \leq 3 \\ 0 & \text{otherwise.} \end{cases}$

It is easy to appreciate that $f_X(x) \geq 0$ for all $x \in \mathbb{R}$ as long as $\alpha \geq 0$. In addition, to ensure that it integrates up to one over the real line, α must equal $9/4$ given that

$$\int_1^3 \alpha\, x^{-3}\,\mathrm{d}x = -\frac{\alpha}{2\,x^2}\bigg|_1^3 = \frac{4\,\alpha}{9}.$$

As before, the interest sometimes lies on calculating the probability of observing X within a given interval. The cumulative distribution function of a continuous random variable is

$$F_X(x) = \Pr(X \leq x) = \int_{-\infty}^{x} f_X(u)\,\mathrm{d}u.$$

The cumulative distribution function is a nondecreasing continuous function given that $F_X(x_1) \leq F_X(x_2)$ if $x_1 \leq x_2$. The continuity is in contrast with the step-like feature in the case of discrete random variables. It is as if the height of the steps shrink to zero

as the random variables moves from a discrete to a continuous nature, so that F_X becomes a continuous function. Given that we are now treating with a continuous random variable, the sum of the pointwise probabilities becomes the integral of the density function. As before, F_X is such that $\lim_{x\to-\infty} F_X(x) = 0$ and $\lim_{x\to\infty} F_X(x) = 1$. In addition, $\Pr(x_0 \leq X \leq x_1) = \Pr(X \leq x_1) - \Pr(X \leq x_0) = F_X(x_1) - F_X(x_0)$ for any $-\infty < x_0 \leq x_1 < \infty$. Finally, it also follows from the definition of an integral that $f_X(x) = \frac{\mathrm{d}}{\mathrm{d}x} F_X(x) = F'_X(x)$ for every $x \in \mathbb{R}$.

Examples

(1) Let X denote a random variable with cumulative distribution function

$$F_X(x) = \begin{cases} 1 - e^{-x}, & \text{if } x \geq 0 \\ 0, & \text{if } x < 0. \end{cases}$$

It is evident that $\lim_{x\to-\infty} F_X(x) = 0$ and that $\lim_{x\to\infty} F_X(x) = 1$, whereas differentiating the cumulative distribution function gives way to $f_X(x) = F'_X(x) = \begin{cases} e^{-x}, & \text{if } x \geq 0 \\ 0, & \text{if } x < 0. \end{cases}$

(2) Let X denote an exponential random variable with density function

$$f_X(x) = \begin{cases} \lambda\, e^{-\lambda\, x}, & \text{if } x \geq 0 \\ 0, & \text{if } x < 0. \end{cases}$$

Note that the previous example is a particular case of the exponential distribution with $\lambda = 1$. Now, consider the probability of observing X within the interval $(t, t+1)$:

$$\begin{aligned} \Pr(t < X < t+1) = F_X(t+1) - F_X(t) &= \int_t^{t+1} \lambda\, e^{-\lambda\, x}\, \mathrm{d}x \\ &= -\, e^{-\lambda\, x}\big|_t^{t+1} = e^{-\lambda\, t}(1 - e^{-\lambda}). \end{aligned}$$

Letting $\alpha = e^{-\lambda}$ then yields $\Pr(t < X < t+1) = (1-\alpha)\,\alpha^t$, which is the probability distribution function of the memoryless discrete random variable of the second exercise of the problem set in Section 4.1.

(3) Let X denote a random variable with density function

$$f_X(x) = \begin{cases} 6\, x\, (1-x), & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

As before, it is easy to see that $f_X(x) \geq 0$ for every $x \in \mathbb{R}$ and that $\int_0^1 6\, x\, (1-x)\, \mathrm{d}x = (3\, x^2 - 2\, x^3)\big|_0^1 = 3 - 2 = 1$. As for the cumulative distribution function, integrating the density function up to x yields

$$F_X(x) = \begin{cases} 0, & \text{if } x < 0 \\ x^2\, (3 - 2x) & \text{if } 0 \leq x \leq 1 \\ 1 & \text{if } x > 1. \end{cases}$$

We may employ the latter to compute the probability of observing a value within an interval as well as a conditional probability as, e.g.,

$$\Pr\left(X \leq \frac{1}{2} \,\middle|\, \frac{1}{3} < X < \frac{2}{3}\right) = \frac{(3x^2 - 2x^3)\big|_{1/3}^{1/2}}{(3x^2 - 2x^3)\big|_{1/3}^{2/3}} = \frac{3/4 - 1/4 - 1/3 + 2/27}{4/3 - 16/27 - 1/3 + 2/27} = \frac{10}{13}.$$

Uniform distribution

The simplest continuous distribution function is the uniform. It essentially dictates that intervals of same length are equiprobable within the support of the random variable, say

$[\alpha, \beta]$. The corresponding density function is

$$f_X(x) = \begin{cases} 1/(\beta - \alpha) & \text{if } \alpha \leq x \leq \beta \\ 0 & \text{otherwise,} \end{cases}$$

which implies that $\Pr(a \leq X \leq b) = \frac{b-a}{\beta-\alpha} = F_X(b) - F_X(a)$ with $-\infty < \alpha \leq a \leq b \leq \beta < \infty$. Finally, the cumulative distribution function reads

$$F_X(x) = \begin{cases} 0 & \text{if } -\infty < x < \alpha \\ x/(\beta - \alpha) & \text{if } \alpha \leq x \leq \beta \\ 1 & \text{if } \beta < x \leq \infty \end{cases}$$

Example: Let X denote a random variable that is uniformly distributed in the unit interval. The density function then is

$$f_X(x) = \begin{cases} \frac{1}{2} & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise,} \end{cases}$$

whereas the cumulative distribution function is

$$F_X(x) = \begin{cases} 0 & \text{if } -\infty < x < 0 \\ x & \text{if } 0 \leq x \leq 1 \\ 1 & \text{if } 1 < x < \infty. \end{cases}$$

There are several applications for the uniform distribution. For instance, if we wish to generate a random variable from a particular distribution, we may always start with a uniform distribution and then transform it to obtain the desired distribution. This is possible because, for any random variable X with cumulative distribution function F_X, $F_X(X)$ has a uniform distribution in the unit interval. More advanced applications include resampling techniques (e.g., bootstrap) and prior distributions in Bayesian analysis.

4.1.4 Functions of random variables

Consider a random variable $X : \ s \mapsto X(s) = x$ for any $s \in \mathcal{S}$ and a transformation $H : \ x \mapsto H(x) = y$ that maps a realization x of the random variable X into a real value

$y \in \mathbb{R}$. Instead of transforming the realization x, one may also transform the random variable X, giving way to another random variable $Y = H(X) : \ s \mapsto H[X(s)]$. Given that the randomness of Y is completely due to X, it is possible to compute the probability distribution of Y if we know the probability distribution function of X.

Example: Let $Y = H(X) = 2X + 1$ with X following a standard exponential distribution $f_X(x) = e^{-x}\,\mathbf{1}(x > 0)$, where $\mathbf{1}(A)$ is an indicator function that takes value one if A is true, zero otherwise. Given that X is positive, the support of Y is given by the interval $[1, \infty)$. It then follows that

$$\begin{aligned} F_Y(y) &= \Pr(Y \leq y) = \Pr(2X + 1 \leq y) = \Pr\left(X \leq \frac{y-1}{2}\right) \\ &= F_X\left(\frac{y-1}{2}\right) = \int_0^{(y-1)/2} e^{-x}\,\mathrm{d}x = \left(-e^{-x}\right)\Big|_0^{(y-1)/2} = 1 - e^{(1-y)/2}. \end{aligned}$$

In general, we deal with transformations of a random variable according to whether the latter is discrete or random. If X is a discrete random variable, then it is easy to appreciate that a random variable $Y = H(X)$ will also be discrete regardless of the transformation H. Letting $X \in \{x_1, \cdots, x_n, \cdots\}$ yields $Y \in \{y_1 = H(x_1), \cdots, y_n = H(x_n), \cdots\}$. In addition, $\Pr(Y = y_i) = \Pr(X = x_i)$ if the transformation H is such that each y corresponds to a unique value x.

Example: Let $X \in \{-1, 0, 1\}$ with $\Pr(X = -1) = 1/3$, $\Pr(X = 0) = 1/2$, and $\Pr(X = 1) = 1/6$. If $Y = X^2$, then $\Pr(Y = 0) = 1/2$ and $\Pr(Y = 1) = 1/2$.

Letting x_{i_k} denote the values of X such that $H(x_{i_k}) = y_i$ for every $k \in \{1, 2, \ldots\}$ leads to $\Pr(Y = y_i) = \sum_{k=1}^{\infty} \Pr(X = x_{i_k})$.

Example: Let $X \in \{1, 2, \ldots, n, \ldots\}$ with $\Pr(X = n) = 2^{-n}$ and let

$$Y = \begin{cases} 1 & \text{if } X \text{ is even,} \\ -1 & \text{if } X \text{ is odd.} \end{cases}$$

It thus follows that $\Pr(Y = 1) = \frac{1}{4} + \frac{1}{16} + \cdots = \frac{1/4}{1-1/4} = \frac{1}{3}$ and hence $\Pr(Y = -1) = \frac{2}{3}$.

If X is a continuous random variable, then $Y = H(X)$ is not necessarily continuous for discreteness can arise depending on the transformation H. Naturally, any continuous transformation preserves the continuity of the random variables.

Example: Let X denote a random variable in the real line and

$$Y = \begin{cases} -1 & \text{if } X < 0, \\ 1 & \text{if } X \geq 0. \end{cases}$$

In this instance, $\Pr(Y = y_i) = \int_A f_X(x)\,\mathrm{d}x$, where A denotes the event about X that corresponds to $\{Y = y_i\}$, that is to say, A is either the negative real line or the nonnegative real line.

In general, to derive the probabilistic structure of Y given X, we must first compute the

distribution function $F_Y(y) = \Pr(Y \leq y)$ by means of the events in X that correspond to $\{Y \leq y\}$ and then differentiate with respect to y to obtain the density function f_Y. Finally, we must also determine the support of Y by seeking the values of y for which $f_Y(y) > 0$.

Example: Let X denote a continuous random variable with density function

$$f_X(x) = \begin{cases} 2x & \text{if } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases} \tag{4.2}$$

Letting $Y = H(X) = 3X + 1$ then yields

$$\begin{aligned} F_Y(y) = \Pr(Y \leq y) &= \Pr(3X + 1 \leq y) = \Pr\left(X \leq \frac{y-1}{3}\right) \\ &= \int_0^{(y-1)/3} f_X(x)\,\mathrm{d}x = \int_0^{(y-1)/3} 2x\,\mathrm{d}x = x^2\big|_0^{(y-1)/3} \\ &= \left(\frac{y-1}{3}\right)^2 = (1-y)^2/9. \end{aligned}$$

The density function then is $f_Y(y) = F_Y'(y) = \frac{2}{9}(y-1)$, whereas the support of Y is given by the interval $(1, 4)$ for $y = 3x + 1$ with $0 < x < 1$ to ensure that both f_X and f_Y are bounded away from zero.

As an alternative, if H is differentiable and strictly monotone, we could also determine F_Y by noting that $x = H^{-1}(y)$ and hence

$$\{H(X) \leq y\} \sim \begin{cases} \{X \leq H^{-1}(y)\} & \text{if } H \text{ is strictly increasing} \\ \{X \geq H^{-1}(y)\} & \text{if } H \text{ is strictly decreasing.} \end{cases}$$

It then suffices to appreciate that

$$\begin{aligned} F_Y(y) = \Pr(Y \leq y) &= \Pr\left[H(X) \leq y\right] = \Pr\left[H \gtrless H^{-1}(y)\right] \\ &= \begin{cases} 1 - F_X\left[H^{-1}(y)\right] & \text{if } H \text{ is strictly decreasing,} \\ F_X\left[H^{-1}(y)\right] & \text{if } H \text{ is strictly increasing.} \end{cases} \end{aligned}$$

As for the density function,

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y} F_Y(y) = \left|\frac{\mathrm{d}F_X}{\mathrm{d}x}\frac{\mathrm{d}x}{\mathrm{d}y}\right| = f_X(x)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right|,$$

where $x = H^{-1}(y)$.

Examples

(1) Let us revisit the previous example in which the density function of X is given by (4.2) and $Y = H(X) = 3X + 1$. The cumulative distribution function of Y is $F_Y(y) = \Pr(Y \leq y) = F\left[(y-1)/3\right]$, whereas the density function is $f_Y(y) = f_X(x)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right|$, with $x = (y-1)/3$. It then follows that $f_Y(y) = \frac{2}{9}(y-1)$. Just as a sanity check, we next check whether the latter integrates to one over the support of Y. If $x \in (0,1)$, then $Y = 3X + 1 \in (1,4)$ and so

$$\int_1^4 f_Y(y)\,\mathrm{d}y = \left(y^2/9 - 2y/9\right)\Big|_1^4 = \frac{16}{9} - \frac{8}{9} - \frac{1}{9} + \frac{2}{9} = 1.$$

(2) Letting now $Y = H(X) = e^{-X}$ yields

$$\begin{aligned} F_Y(y) &= \Pr(Y \leq y) = \Pr(e^{-X} \leq y) = \Pr(X \geq -\ln y) \\ &= \int_{-\ln y}^1 2x\,\mathrm{d}x = \left(x^2\right)\Big|_{-\ln y}^1 = 1 - (-\ln y)^2. \end{aligned}$$

Differentiating with respect to y then leads to $f_Y(y) = F_Y'(y) = -(2\ln y)/y$. As for the support, we confirm that $Y \in (1/e, 1)$ by showing that $\int_{1/e}^1 -\frac{2\ln y}{y}\,\mathrm{d}y = 1$. Needless to say, applying the alternative methods results in the same expressions for the cumulative distribution and density functions. Indeed,

$$F_Y(y) = \Pr(Y \leq y) = \Pr(X \geq -\ln y) = 1 - F_X(-\ln y),$$

whereas $f_Y(y) = f_X(x)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right| = -(2\ln y)/y$ given that $x = -\ln y$.

(3) Let X denote a random variable with density function

$$f_X(x) = \begin{cases} 1/2 & \text{if } -1 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

The cumulative distribution function of $Y = X^2$ then is

$$F_Y(y) = \Pr(Y \leq y) = \Pr(X^2 \leq y) = \Pr(-\sqrt{y} \leq X \leq \sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y}),$$

whereas the density function is

$$f_Y(y) = F'_Y(y) = \frac{f_X(\sqrt{y})}{2\sqrt{y})} - \frac{f_X(-\sqrt{y})}{-2\sqrt{y}} = \frac{1}{2\sqrt{y}}\left[f_X(\sqrt{y}) + f_X(-\sqrt{y})\right]$$
$$= \frac{1}{2\sqrt{y}}\left(\frac{1}{2} + \frac{1}{2}\right) = \frac{1}{2\sqrt{y}}$$

with a support given by the unit interval.

4.2 Random vectors and joint distributions

A random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ is a vector of, say n, random variables. There are three types of random vectors: continuous, discrete, and mixed. The latter is essentially a vector including both continuous and discrete random variables, and hence we will focus only on continuous and discrete random vectors. In what follows, we will consider a bivariate random vector (X, Y), though extending the discussion to n-dimensional random vectors is straightforward. The joint probability distribution function of a discrete bivariate random vector (X, Y) is $p(x, y) = \Pr(X = x, Y = y)$, where $X \in \{x_1, \ldots, x_n, \ldots\}$ and $Y = \{y_1, \ldots, y_n, \ldots\}$. Similarly, the joint density function of a continuous bivariate random vector (X, Y) is $f_{XY}(x, y) \sim \Pr(x \leq X \leq x + \Delta x, y \leq Y \leq y + \Delta y)$ for small enough Δx and Δy. As before, the density function is such that $f_{XY}(x, y) \geq 0$ for every $(x, y) \in \mathbb{R}^2$ and that $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y) \, \mathrm{d}x \, \mathrm{d}y = 1$.

Examples

(1) Suppose there are two shoemakers in a shop. The first does at best 5 shoes in a given month, whereas the second takes more time to make a shoe and hence does at most 3 shoes per month. Let $X \in \{0, 1, \ldots, 5\}$ and $Y \in \{0, 1, \ldots, 3\}$ denote the number of shoes by the first and second shoemakers in a given month, with probabilities given by

	$X = 0$	1	2	3	4	5
$Y = 0$	0	0.01	0.03	0.05	0.07	0.09
1	0.01	0.02	0.04	0.05	0.06	0.08
2	0.01	0.03	0.05	0.05	0.05	0.06
3	0.01	0.02	0.04	0.06	0.06	0.05

If the interest lies on the event $B = \{X > Y\}$, for instance, it then suffices to sum up the probabilities that appear in boldface, resulting in $\Pr(B) = 3/4$.

(2) The shop owner is a bit concerned with the amount of leather each shoemaker employs per month. Let X and Y now denote the quantity of leather the first and second shoemakers

spend, respectively. The shop owner is so miser that he gauges with infinite precision how much leather the shoemakers use, and so we may assume that (X, Y) is a continuous bivariate random variable. In addition, the joint density function is given by

$$f_{XY}(x,y) = \begin{cases} \alpha \text{ if } 5 \leq x \leq 10 \text{ and } 4 \leq y \leq 9 \\ 0 \text{ otherwise.} \end{cases}$$

We first compute α by integrating the density function over $\mathbb{R}^2$:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x,y)\,\mathrm{d}x\,\mathrm{d}y = 1$$
$$\int_4^9\int_5^{10} \alpha\,\mathrm{d}x\,\mathrm{d}y = \int_4^9 (\alpha x)|_5^{10}\,\mathrm{d}y = \int_4^9 5\alpha\,\mathrm{d}y$$
$$= (5\alpha y)|_4^9 = 25\alpha = 1,$$

implying that $\alpha = 1/25$. Next, let's compute how likely is the event $B = \{X > Y\}$ that the first shoemaker employs more leather than the second shoemaker:

$$\Pr(B) = 1 - \Pr(X \leq Y) = 1 - \int_5^9\int_5^y \frac{1}{25}\,\mathrm{d}x\,\mathrm{d}y$$
$$= 1 - \frac{1}{25}\int_5^9 (y-5)\,\mathrm{d}y = \frac{17}{25}.$$

The first integral is over the interval $[5, 9]$ because it is impossible to observe $Y \geq X$ if $4 \leq Y < 5$ given that $X \geq 5$, whereas the second integral is over the interval $[5, y]$ because X cannot exceed the value that we observe for Y given that $Y \geq X$.

(3) Let (X, Y) denote a bivariate random variable with joint density function

$$f_{XY}(x,y) = \begin{cases} x^2 + xy/3 & \text{if }) < x < 1 \text{ and } 0 < y < 2 \\ 0 & \text{otherwise.} \end{cases}$$

We first show that the above density function integrates to one:

$$\int_0^1\int_0^2 \left(x^2 + \frac{1}{3}\,xy\right)\,\mathrm{d}y\,\mathrm{d}x = \int_0^1 \left(x^2 y + \frac{1}{6}\,xy^2\right)\Big|_0^2\,\mathrm{d}x$$
$$= \int_0^1 \left(2x^2 + \frac{2}{3}\,x\right)\,\mathrm{d}x = \left(\frac{2}{3}\,x^3 + \frac{1}{3}\,x^2\right)\Big|_0^1$$
$$= \frac{2}{3} + \frac{1}{3} = 1.$$

We next illustrate how to compute the likelihood of an event that involves both X and Y, say, the probability of observing $\{X + Y \leq 1\}$:

$$\begin{aligned}
\Pr(X + Y \leq 1) &= \Pr(Y \leq 1 - X) \\
&= \int_0^1 \int_0^{1-x} \left(x^2 + \frac{1}{3}\, xy \right) \mathrm{d}y\, \mathrm{d}x = \int_0^1 \left(x^2 y + \frac{1}{6}\, xy^2 \right) \Big|_0^{1-x} \mathrm{d}x \\
&= \int_0^1 \left[x^2(1-x) + \frac{1}{6}\, x(1-x)^2 \right] \mathrm{d}x = \int_0^1 \left[x^2 - x^3 + \frac{1}{6}\, (x - 2x^2 + x^3) \right] \mathrm{d}x \\
&= \int_0^1 \left(\frac{2}{3}\, x^2 - \frac{5}{6}\, x^3 + \frac{1}{6}\, x \right) \mathrm{d}x = \left(\frac{2}{9}\, x^3 - \frac{5}{24}\, x^4 + \frac{1}{12}\, x \right) \Big|_0^1 \\
&= \frac{2}{9} - \frac{5}{24} + \frac{1}{12} = \frac{16 - 15 + 6}{72} = \frac{7}{72}.
\end{aligned}$$

In general, it follows that the joint probability distribution function of a continuous random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ is given by

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = \Pr(\boldsymbol{X} \leq \boldsymbol{x}) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_{\boldsymbol{X}}(x_1, \ldots, x_n)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n,$$

with $\boldsymbol{x} = (x_1, \ldots, x_n)$, whereas the joint density function is

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{\partial}{\partial \boldsymbol{x}'}\, F_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{\partial^n}{\partial x_1 \cdots \partial x_n}\, F_{\boldsymbol{X}}(x_1, \ldots, x_n).$$

4.3 Marginal distributions

Knowing the joint distribution function F_{XY} of (X, Y) also implies the knowledge of the marginal distributions F_X and F_Y of X and Y, respectively. After all, it contains all information about the probabilistic structure of both X and Y. To extract the marginal distributions of X and Y, it suffices to 'integrate' the other random variable out of the joint distribution. We employ quotation marks because integrating out a discrete random variable, say Y, corresponds to summing the joint probability for all possible values of Y given that any of these values may occur. Letting $B_j = \{X = x, Y = y_j\}$ for $j = 1, 2, \ldots$ then yields $p(x) = \Pr(X = x) = \Pr(B_1 \text{ or } \cdots \text{ or } B_n \text{ or } \cdots) = \sum_{j=1}^{\infty} p(x, y_j)$, given that these events are all mutually exclusive.

As for continuous random variables, the marginal density function of X is

$$\begin{aligned} f_X(x) &\sim \Pr(x \leq X \leq x + \Delta x) \qquad \text{for a very small } \Delta x > 0 \\ &= \Pr(x \leq X \leq x + \Delta x, -\infty < Y < \infty) = \int_{-\infty}^{\infty} f_{XY}(x, y)\, \mathrm{d}y, \end{aligned}$$

and hence

$$\Pr(a < X < b) = \Pr(a < X < b, -\infty < Y < \infty) = \int_a^b \int_{-\infty}^{\infty} f_{XY}(x, y)\, \mathrm{d}y\, \mathrm{d}x.$$

Examples

(1) Let (X, Y) denote a bivariate random vector with joint density given by

$$f_{XY}(x, y) = \begin{cases} 2(x + y - 2xy) & \text{if } 0 < x, y < 1 \\ 0 & \text{otherwise.} \end{cases}$$

We first confirm that (4.3) indeed is a density function by showing that it integrates to one:

$$\begin{aligned} \int_0^1 \int_0^1 2(x + y - 2xy)\, \mathrm{d}x\, \mathrm{d}y &= \int_0^1 \left(x^2 + 2xy - 2x^2 y\right)\Big|_0^1 \mathrm{d}y \\ &= \int_0^1 \mathrm{d}y = 1. \end{aligned}$$

It is evident from the above derivations that the marginal distributions of X and Y are both uniform in the unit interval, even though their joint distribution is not uniform.

(2) Let (X, Y) denote a uniform random variable in the rectangle $[\alpha_X, \beta_X] \times [\alpha_Y, \beta_Y]$. The joint density function then is

$$f_{XY}(x,y) = \begin{cases} \frac{1}{(\beta_X - \alpha_X)(\beta_Y - \alpha_Y)} & \text{if } \alpha_X < x < \beta_X, \alpha_Y < y < \beta_Y \\ 0 & \text{otherwise.} \end{cases}$$

Integrating X out yields a uniform density function in the interval $[\alpha_Y, \beta_Y]$ for Y, whereas integrating the latter out gives way to a uniform distribution function for X in the interval $[\alpha_X, \beta_X]$.

(3) Suppose now that (X, Y) is uniform in $B = \{(x, y) : \ 0 < x < 1, x^2 < y < x\}$. It follows from the fact that the area of B is $\int_0^1 (x - x^2)\, \mathrm{d}x = \frac{1}{6}$ that

$$f_{XY}(x,y) = \begin{cases} 6 & \text{if } (x,y) \in B \\ 0 & \text{if } (x,y) \notin B, \end{cases}$$

otherwise the joint density would integrate to something different from one. The marginal density of X then is $f_X(x) = \int_{x^2}^{x} 6\, \mathrm{d}y = 6\, x(1-x)$ for $0 \leq x \leq 1$, whereas the marginal of Y is $f_Y(y) = \int_y^{\sqrt{y}} 6\, \mathrm{d}x = 6(\sqrt{y} - y)$ for $0 \leq y \leq 1$.

The above examples illustrate that a uniform joint distribution does not ensure uniform marginals, just as uniform marginals do not imply a uniform joint distribution.

4.4 Conditional density function

In the case of discrete random variables, by definition, it suffices to compute the ratio between the probability of observing both events and the probability of observing the conditioning event, so that $p(x|y) = p(x,y)/p(y)$ if $p(y) > 0$ and $p(y|x) = p(x,y)/p(x)$ if $p(x) > 0$. Note that the conditional probability meets all of the conditions for a probability distribution function, namely, it is nonnegative for all values of x and sums up to one given that

$$\begin{aligned} \sum_{i=1}^{\infty} p(x_i|y) &= \sum_{i=1}^{\infty} \frac{p(x_i, y)}{p(y)} \\ &= \frac{1}{p(y)} \sum_{i=1}^{\infty} p(x_i, y) = \frac{p(y)}{p(y)} = 1. \end{aligned}$$

Example: Let us revisit the example of the two shoemakers and compute the probability of the first shoemaker to produce two shoes in a given month given that the second shoemaker has also produced two shoes. By the definition of conditional probability, it suffices to compute the ratio between the probability of observing both events and the probability of observing the conditioning event, so that

$$\Pr(X = 2|Y = 2) = \frac{\Pr(X = 2, Y = 2)}{\Pr(Y = 2)} = \frac{0.05}{0.25} = \frac{1}{5}.$$

As for continuous random variables, it is easy to see that the same applies to density functions given that they proxy for the probability of observing the random variable within an interval of infinitesimal length. It thus ensues that $f_{X|Y}(x|y) = f_{XY}(x, y)/f_Y(y)$ if $f_Y(y) > 0$ and that $f_{Y|X}(y|x) = f_{XY}(x, y)/f_X(x)$ if $f_X(x) > 0$. As before, the conditional density function indeed is a density function in that $f_{X|Y}(x|y) \geq 0$ for every $x \in \mathbb{R}$ and

$$\begin{aligned}\int_{-\infty}^{\infty} f_{X|Y}(x|y)\,\mathrm{d}x &= \int_{-\infty}^{\infty} \frac{f_{XY}(x, y)}{f_Y(y)}\,\mathrm{d}x \\ &= \frac{1}{f_Y(y)} \int_{-\infty}^{\infty} f_{XY}(x, y)\,\mathrm{d}x = \frac{f_Y(y)}{f_Y(y)} = 1.\end{aligned}$$

Example: Let (X, Y) denote a random vector with joint density function given by

$$f_{XY}(x, y) = \begin{cases} x^2 + \frac{1}{3}\,xy & \text{if } (x, y) \in [0, 1] \times [0, 2] \\ 0 & \text{otherwise.} \end{cases}$$

The conditional density of X given Y then is

$$\begin{aligned} f_{X|Y}(x|y) = \frac{f_{XY}(x, y)}{f_Y(y)} &= \frac{x^2 + \frac{1}{3}\,xy}{\int_0^1 \left(x^2 + \frac{1}{3}\,xy\right)\,\mathrm{d}x} = \frac{x^2 + \frac{1}{3}\,xy}{\left(\frac{1}{3}\,x^3 + \frac{1}{6}\,x^2 y\right)\Big|_0^1} \\ &= \frac{x^2 + \frac{1}{3}\,xy}{\frac{1}{3} + \frac{y}{6}} = \frac{6x^2 + 2xy}{2 + y}, \end{aligned}$$

for $0 \leq x \leq 1$ and $0 \leq y \leq 2$.

4.5 Independent random variables

There is a nice correspondence between independent events and independent random variables. The condition that $\Pr(A \cap B) = \Pr(A)\Pr(B)$ translates into $p(x, y) = p(x)p(y)$ if

X and Y are discrete random variables and into $f_{XY}(x,y) = f_X(x)f_Y(y)$ if X and Y are continuous random variables. That is to say, independence ensues if and only if the joint probability/density is the product of the marginals. In addition, we can also define independence by means of conditional probabilities/densities in that independence holds if and only if (a) $p(x|y) = p(x,y)/p(y) = p(x)$ and $p(y|x) = p(x,y)/p(x) = p(y)$ if X and Y are discrete; and (b) $f_{X|Y}(x|y) = f_{XY}(x,y)/f_Y(y) = f_X(x)$ and $f_{Y|X}(y|x) = f_{XY}(x,y)/f_X(x) = f_Y(y)$ if X and Y are continuous.

Examples

(1) Let X and Y denote the time it takes to observe a transaction for two stocks, with joint density function given by

$$f_{XY}(x,y) = \begin{cases} \exp\left[-(x+y)\right] & \text{if } x \geq 0, y \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

It is easy to see that $f_{XY}(x,y) = f_X(x)f_Y(y)$, with $f_X(x) = e^{-x}$ for $x > 0$ and $f_Y(y) = e^{-y}$ for $y > 0$, and hence X and Y are independent random variables.

(2) Let now X and Y denote random variables with joint density given by $f_{XY}(x,y) = 8xy$ for $0 \leq x \leq y \leq 1$. Although it is easy to decompose f_{XY} into the product of functions that depend exclusively on either X or Y, there is no way to get rid of the dependence in the support. The fact that X is always inferior to Y is what makes the two random variables dependent.

Finally, it is interesting to observe more closely how the link between independent events and independent random variables works. Let A and B denote independent events concerning the random variables X and Y, respectively. Equivalence follows from the fact that

$$\begin{aligned} \Pr(A \cap B) &= \int\int_{A\cap B} f_{XY}(x,y)\,\mathrm{d}x\,\mathrm{d}y = \int\int_{A\cap B} f_X(x)f_Y(y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_A f_X(x)\,\mathrm{d}x \int B f_Y(y)\,\mathrm{d}y = \Pr(A)\Pr(B). \end{aligned}$$

4.6 Expected value, moments, and co-moments

Another way to characterize a distribution is through its moments. The first moment refers to the expected value of the distribution, whereas the second moment relates to the dispersion of the distribution. There is also room for higher-order moments in that there are distributions with an infinite number of moments. For instance, the third moment tells us whether the distribution is asymmetric around the expected value, whilst the fourth moment determines how thick the tails of the distribution are, that is to say, how likely it is to observe an extreme realization regardless of whether to the right or to the left of the expected value.

4.6.1 Expected value

The first moment of a distribution is known as expected value or population mean. As the name implies, it entails a typical value for the distribution. Before formalizing the notion of expected value, it is convenient to start with an example to motivate the discussion.

Example: Gimli proposes a game to his friend Legolas with prizes in gold pieces as in the table below. In addition, if Legolas decides to play, he must pay beforehand one gold piece to Gimli.

die throw	1	2	3	4	5	6
Legolas' payoff	-3	-2	-1	1	2	5

Legolas' expected payoff then is the average payoff minus the entry costs, that is to say, $\mathbb{E}(\text{Legolas' payoff}) = \frac{1}{6}(5+2+1-1-2-3)-1 = -\frac{2}{3}$. This means that the game is not very fair to Legolas given that, on average, he will have to pay 2/3 of a gold piece to Gimli. To make the game fair, Legolas would have to bargain down the entry cost to 1/3.

In the above example, it is easy to compute the expected value because the outcomes of a die throw are equiprobable and hence it suffices to compute the arithmetic mean of the possible outcomes. In general, the values that the random variable can assume are not equiprobable and hence we must weigh by their mass of probability. We do that by applying the expectation operator $\mathbb{E}(\cdot)$ to the random variable of interest. In the case of discrete random variables, the expectation operator is such that $\mathbb{E}(X) = \sum_{i=1}^{\infty} x_i p(x_i)$. If the latter series does not converge to some finite value, then we say that the distribution has no expected value.

Examples

(1) Let $X \sim \mathcal{B}(n,p)$, so that $\Pr(X=x) = \binom{n}{x} p^x (1-p)^{n-x}$. The expected value of a

binomial distribution is

$$\begin{aligned}\mathbb{E}(X) &= \sum_{x=0}^{n} x \Pr(X = x) = \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x} \\ &= \sum_{x=1}^{n} x \binom{n}{x} p^x (1-p)^{n-x} = \sum_{x=1}^{n} x \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \\ &= \sum_{x=1}^{n} \frac{n!}{(x-1)!(n-x)!} p^x (1-p)^{n-x}.\end{aligned}$$

Letting $k = x - 1$ then yields

$$\begin{aligned}\mathbb{E}(X) &= n \sum_{k=0}^{n-1} \binom{n-1}{k} p^{k+1} (1-p)^{n-k-1} \\ &= np \sum_{k=0}^{n-1} \binom{n-1}{k} p^k (1-p)^{n-1-k} = np\end{aligned}$$

given that $\binom{n-1}{k} p^k (1-p)^{n-1-k}$ is the probability distribution function of a $\mathcal{B}(n-1, p)$. Note that the expected value of the binomial distribution corresponds to the absolute frequency at which we observe a value equal to one.

(2) Suppose that a given trading strategy normally entails a weekly return S, though it may profit less if there is a sequence of bad news. More precisely, the return reduces to $0 < R < S$ if there are up to two bad news, and to $L < 0$ if there are three or more bad news over the week. The probability of observing a bad news is $1/20$ in any given trading day of the week. Let now $X \in \{L, R, S\}$ and $B \in \{1, 2, 3, 4, 5\}$ denote the weekly return of this trading strategy and the number of bad news over the week, respectively. The latter is a binomial random variable with probability distribution function given by

$\Pr(B = k) = \binom{5}{k}(1/20)^k\,(19/20)^{n-k}$, and hence the expected value of X is

$$\begin{aligned}
\mathbb{E}(X) &= \Pr(B = 0)\,S + [\Pr(B = 1) + \Pr(B = 2)]\,R + \Pr(B \geq 3)\,L \\
&= (19/20)^5\,S + \left[5(1/20)(19/20)^4 + 10(1/20)^2(19/20)^3\right] R \\
&\quad + \left[10(1/20)^3(19/20)^2 + 5(1/20)^4(19/20) + (1/20)^5\right] L \\
&= (19/20)^5 S + 5(1/20)(19/20)^3\,[19/20 + 2(1/20)]\,R \\
&\quad + (1/20)^3\left[10(19/20)^2 + 5(1/20)(19/20) + (1/20)^2\right] L \\
&= 0.7737809\,S + 0.2250609\,R + 0.0011581\,L.
\end{aligned}$$

Note that the weights assigned to each outcome of the weekly return sum up to one for they correspond to their probability of occurring.

As for continuous random variables, the expectation operator is such that $\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x$. As before, if the function $g(x) = x f_X(x)$ is not integrable, then the distribution features no expected value.

Examples

(1) Let X denote a uniform random variable in the interval $[\alpha, \beta]$. In view that the density function is

$$f_X(x) = \begin{cases} \frac{1}{\beta-\alpha} & \text{if } \alpha < x < \beta \\ 0 & \text{otherwise} \end{cases}$$

it follows that

$$\mathbb{E}(X) = \int_\alpha^\beta \frac{x}{\beta-\alpha}\,\mathrm{d}x = \left(\frac{x^2/2}{\beta-\alpha}\right)\Bigg|_\alpha^\beta = \frac{\beta^2-\alpha^2}{2(\beta-\alpha)} = \frac{(\beta-\alpha)(\beta+\alpha)}{2(\beta-\alpha)} = \frac{\beta+\alpha}{2}.$$

This makes sense for the uniform continuous distribution is analogous to the case of equiprobable events and hence it suffices to take the arithmetic mean of the lower and upper limits of the support.

(2) Let X denote a random variable with density function given by

$$f_X(x) = \begin{cases} x/225 & \text{if } 0 \le x \le 15 \\ (30-x)/225 & \text{if } 15 \le x \le 30 \\ 0 & \text{otherwise.} \end{cases}$$

The expected value then is

$$\begin{aligned} \mathbb{E}(X) &= \int_0^{15} \frac{x^2}{225}\,\mathrm{d}x + \int_{15}^{30} \frac{(30-x)x}{225}\,\mathrm{d}x = \frac{1}{225}\left[\left(\frac{x^3}{3}\right)\Bigg|_0^{15} + \left(15x^2 - \frac{x^3}{3}\right)\Bigg|_{15}^{30}\right] \\ &= \frac{1}{15^2}\left(\frac{15^3}{3} + 15\times 30^2 - 15^3 - \frac{30^3}{3} + \frac{15^3}{3}\right) = 5 + 60 - 15 - 40 + 5 = 15. \end{aligned}$$

This result is pretty intuitive given that the density function looks like a symmetric triangle with peak at 15.

Suppose now we wish to derive the expected value of $Y = H(X)$. The expectation operator is such that

$$\mathbb{E}(Y) = \begin{cases} \sum_{i=1}^\infty y_i p(y_i) & \text{if discrete} \\ \int_{-\infty}^\infty y f_Y(y)\,\mathrm{d}y & \text{if continuous} \end{cases} = \begin{cases} \sum_{i=1}^\infty H(x_i) p(x_i) & \text{if discrete} \\ \int_{-\infty}^\infty H(x) f_X(x)\,\mathrm{d}x & \text{if continuous.} \end{cases}$$

This means that there is no need to derive the distribution of Y to compute its expected value. It suffices to compute the expectation of $H(X)$ given the density function of X.

Example: In some situations, the interest lies on the magnitude of the random variable regardless of the sign it takes. Suppose, for instance, that X has a double exponential density given by

$$f_X(x) = \begin{cases} \frac{1}{2}\, e^{x} & \text{if } x \leq 0 \\ \frac{1}{2}\, e^{-x} & \text{if } x \geq 0, \end{cases}$$

which is symmetric around zero. Now, the expected value of $Y = |X|$ is

$$\begin{aligned} \mathbb{E}(Y) &= \int_{-\infty}^{\infty} |x|\, f_X(x)\, \mathrm{d}x = \frac{1}{2}\left[\int_{-\infty}^{0} |x|\, e^{x}\, \mathrm{d}x + \int_{0}^{\infty} |x|\, e^{x}\, \mathrm{d}x\right] \\ &= \frac{1}{2}\left[\int_{-\infty}^{0} (-x) e^{x}\, \mathrm{d}x + \int_{0}^{\infty} x e^{-x}\, \mathrm{d}x\right] = \int_{0}^{\infty} x e^{-x}\, \mathrm{d}x = 1, \end{aligned}$$

where the last equality follows from integration by parts. Alternatively, one can compute the expected value of Y by first deriving its distribution. In particular,

$$\begin{aligned} F_Y(y) &= \Pr(Y \leq y) = \Pr(|X| \leq y) = \Pr(-y \leq X \leq y) = 2\Pr(0 \leq X \leq y) \\ &= 2\int_{0}^{y} f_X(x)\, \mathrm{d}x = 2\int_{0}^{y} e^{-x}\, \mathrm{d}x = \left(-e^{-x}\right)\Big|_{0}^{y} = 1 - e^{-y}, \end{aligned}$$

giving way to $f_Y(y) = e^{-y}$ for $y \geq 0$ and to $\mathbb{E}(Y) = \int_0^\infty y f_Y(y)\, \mathrm{d}y = \int_0^\infty y e^{-y}\, \mathrm{d}y = 1$.

It is possible to compute the expected value of a function of a random vector, say $Z = H(X, Y)$, along the same lines. More precisely,

$$\mathbb{E}(Z) = \begin{cases} \sum_{i=1}^{\infty} z_i p(z_i) & \text{if discrete} \\ \int_{-\infty}^{\infty} z f_Z(z)\, \mathrm{d}z & \text{if continuous} \end{cases} = \begin{cases} \sum_{i=1}^{\infty} H(x_i, y_i) p(x_i, y_i) & \text{if discrete} \\ \int_{-\infty}^{\infty} H(x, y) f_{XY}(x, y)\, \mathrm{d}x\, \mathrm{d}y & \text{if continuous,} \end{cases}$$

which avoids the derivation of the probability/density function of Z. Apart from that, the expectation operator has two other interesting properties. First, it is a linear operator in that $\mathbb{E}(aX + b) = a \sum_{i=1}^{n} \mathbb{E}(X_i) + b$ for any fixed constants a and b, if $X = \sum_{i=1}^{n} X_i$. Second, if the random variables are independent, then the expectation of their product is equal to the product of their expectations. This means that, if X and Y are independent, then $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$. The examples below employ these properties to derive expectations.

Examples

(1) A binomial distribution $\mathcal{B}(n, p)$ results from the sum of the outcomes of a sequence

of n independent Bernoulli essays with probability p. Let Y_i denote the outcome of each of these Bernoulli essays ($i = 1, \ldots, n$), taking value one with probability p, otherwise zero. The expected value of X then is $\mathbb{E}(X) = \mathbb{E}(Y_1 + Y_2 + \cdots + Y_n) = \sum_{i=1}^{n} \mathbb{E}(Y_i) = np$.

(2) Let D denote the weekly demand for apple crumbles, with probability distribution function $p_n = \Pr(D = n)$. Let C denote the cost of baking a unit of apple crumble and E the cost of keeping one apple crumble. Suppose that we sell each apple crumble at a price P and that our initial stock is of N apple crumbles. It then follows that our profit Π in a given week is a random variable given by

$$\Pi = \begin{cases} N(P-C) & \text{if } D \geq N \\ DP - NC - (N-D)E & \text{if } D < N \end{cases}$$

because in the latter case we produce N apple crumbles, sell D and then stock the ones that we are not able to sell during the week. This means that

$$\Pi = \begin{cases} N(P-C) & \text{with probability } \Pr(D \geq N) = 1 - \Pr(D < N) \\ D(P+E) - N(C+E) & \text{with probability } \Pr(D < N) \end{cases}$$

and hence

$$\begin{aligned}
\mathbb{E}(\Pi) &= N(P-C)\Pr(D \geq N) + \mathbb{E}\left[D(P+E) - N(C+E) | D < N\right]\Pr(D < N) \\
&= N(P-C)\sum_{n=N+1}^{\infty} p_n + (P+E)\sum_{n=0}^{N} n\,p_n - N(C+E)\sum_{n=0}^{N} p_n \\
&= N(P-C)\left[1 - \sum_{n=0}^{N} p_n\right] + (P+E)\sum_{n=0}^{N} n\,p_n - N(C+E)\sum_{n=0}^{N} p_n \\
&= N(P-C) + (P+E)\sum_{n=0}^{N} n\,p_n - \left[N(C+E) + N(P-C)\right]\sum_{n=0}^{N} p_n \\
&= N(P-C) + (P+E)\sum_{n=0}^{N} n\,p_n - N(P+E)\sum_{n=0}^{N} p_n \\
&= N(P-C) - (P+E)\sum_{n=0}^{N} (N-n)p_n.
\end{aligned}$$

If, for instance, $\Pr(D = n) = \frac{1}{10}$ for $n \in \{0, 1, \ldots, 9\}$, it then ensues that

$$\begin{aligned}\mathbb{E}(\Pi) &= N(P-C) - (P+E)\sum_{n=0}^{N}\frac{N-n}{10} = N(P-C) - \frac{P+E}{10}\left[N(N+1) - \sum_{n=0}^{N} n\right] \\ &= N(P-C) - \frac{P+E}{10}\left[N(N+1) - \frac{N(N+1)}{2}\right] = N(P-C) - \frac{P+E}{10}\,\frac{N(N+1)}{2} \\ &= N(P-C) - \frac{N(N+1)(P+E)}{20}.\end{aligned}$$

The above naturally only holds if $N \leq 9$, which makes sense given that it seems unreasonable to start the week with more than the maximum demand for apple crumbles.

The expectation operator is a linear operator and hence it is extremely easy to deal with affine functions of a random variable. Although we have already shown that it is also straightforward to handle non-affine functions of random variables, it is important to note that $\mathbb{E}[g(X)] \neq g[\mathbb{E}(X)]$ in general. For instance, the next example illustrates that $\mathbb{E}(X^2) \geq [\mathbb{E}(X)]^2$ for any random variable X.

Example: Let X denote a random variable that takes value either 1 or -1 with probability 1/2. Given that it is symmetric around zero, the expected value of X is obviously zero and hence $[\mathbb{E}(X)]^2 = 0$. In contrast, the expected value of its square exceeds zero given that $\mathbb{E}(X^2) = (1+1)1/2 = 1$.

4.6.2 Variance, covariance, and correlation

The variance measures the amount of variation and dispersion of a random variable around the mean value. This information is paramount to any problem of statistical inference. In finance, we are not only interested in the expected return of an investment, but also in how risky it is. Most people prefer an investment that entails a return of 10% for sure than one with a return of either 30% or -20% with probability 1/2 despite the fact that both investments have en expected return of 10%. The most basic measure of risk is given by the variance, which gauges the average magnitude of the deviations with respect to the mean value by means of a quadratic transformation: $\text{var}(X) = \mathbb{E}\left[X - \mathbb{E}(X)\right]^2$.

We could of course employ the absolute value rather than the square to measure the magnitude of the deviation. The advantage of using squares is that it is differentiable as opposed to the absolute value and that it is very easy to compute by means of the expectation operator. The drawback is that taking squares potentiates the extreme values, if any, in the data. Another disadvantage of the variance is that it is not in the same unit of the random variable. This is however easy to remedy in that we can always consider the standard deviation of the random variable, namely, the square root of the variance, so as to recover the original unit of measurement.

The variance is also known as the second centered moment of the distribution. We say it is centered because we are looking at deviations around the first moment (i.e., the expected value). It relates to the second moment because of the focus on the second power of the random variable. It is interesting to note that we can also write the variance as a function

of the first two (uncentered) moments of the distribution given that

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

The properties of the expectation operator are very helpful to compute the second uncentered moment of the distribution in view that it is not necessary to find the distribution of the square of the random variable. Indeed,

$$\mathbb{E}(X^2) = \begin{cases} \sum_{i=1}^{\infty} x_i^2 p(x_i) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} x^2 f_X(x)\,\mathrm{d}x & \text{if } X \text{ is continuous} \end{cases}$$

The properties of the expectation operator also imply a couple of properties for the variance. First, the variance of an affine function of X is proportional to the variance of X. In particular, $\operatorname{var}(aX + b) = a^2 \operatorname{var}(X)$ for any fixed constants a and b given that both the second moment and the square of the first moment will depend on the square of the slope coefficient a^2 and the fact that we take deviations with respect to the expected value will take care of the intercept b. Second, in the event that X and Y are independent random variables, the variance of the sum is equal to the sum of the variances, that is to say, $\operatorname{var}(X + Y) = \operatorname{var}(X) + \operatorname{var}(Y)$. In what follows, we demonstrate both properties by showing that $\operatorname{var}(aX + b + cY) = a^2 \operatorname{var}(X) + c^2 \operatorname{var}(Y)$ if X and Y are independent:

$$\begin{aligned}\operatorname{var}(aX + b + cY) &= \mathbb{E}\left[aX + b + cY - \mathbb{E}(aX + b + cY)\right]^2 \\ &= \mathbb{E}\left[aX + b + cY - a\,\mathbb{E}(X) - b - c\,\mathbb{E}(Y)\right]^2 \\ &= \mathbb{E}\left\{a[X - \mathbb{E}(X)] + c[Y - \mathbb{E}(Y)]\right\}^2 \\ &= \mathbb{E}\left\{a^2\left[X - \mathbb{E}(X)\right]^2 + 2ac\left[X - \mathbb{E}(X)\right][Y - \mathbb{E}(Y)] + c^2\left[Y - \mathbb{E}(Y)\right]^2\right\} \\ &= a^2 \operatorname{var}(X) + c^2 \operatorname{var}(Y) + 2ac\,\mathbb{E}[X - \mathbb{E}(X)][Y - \mathbb{E}(Y)].\end{aligned}$$

To show that the last term is equal to zero, it suffices to appreciate that

$$\mathbb{E}[X - \mathbb{E}(X)][Y - \mathbb{E}(Y)] = \mathbb{E}(XY) - 2\,\mathbb{E}(X)\mathbb{E}(Y) + \mathbb{E}(X)\mathbb{E}(Y) = 0 \tag{4.3}$$

given that independence between X and Y implies that the expectation of the product is the product of the expectations.

Equation (4.3) suggests a simple measure of dependence between two random variables based on how their deviations relative to the mean co-vary. Bearing that in mind, we define the covariance between X and Y as

$$\text{cov}(X, Y) = \mathbb{E}[X - \mathbb{E}(X)][Y - \mathbb{E}(Y)] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

The intuition is simple. If the deviations of $X - \mathbb{E}(X)$ tend to have the same sign as the deviations $Y - \mathbb{E}(Y)$, we then say that X and Y co-move together and hence their covariance is positive. In contrast, the covariance is negative if the deviations tend to have opposite signs. It is possible to show that the covariance is a measure of linear dependence and hence independence implies zero covariance (or, equivalently, orthogonality), though zero covariance does not necessarily imply independence.

Another drawback of the covariance is that it has a strange unit. It is in units of X times units of Y, compromising a bit interpretability. The easiest remedy for that is to standardize the deviations of X and Y with respect to their means by their standard deviation, giving way to the correlation:

$$\operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var}(X)\operatorname{var}(Y)}}.$$

The latter has no unit in contrast to the covariance. In addition, standardizing by the standard deviation is also convenient for it makes the two deviations comparable. The correlation also has some very nice properties. First, it is at most one in magnitude, that is, $-1 \leq \operatorname{corr}(X,Y) \leq 1$. Second, as the covariance, the correlation is equal to zero if and only if X and Y are orthogonal (or, equivalently, linearly independent). Third, as in the variance and the covariance, the correlation is also based on expectations and hence we can employ all of the apparatus that comes with the expectation operator.

Examples

(1) A survey classify the degree of customers' satisfaction, say X, into a scale from 0 to 10. The answers to the survey indicate a symmetric probability distribution given by $p_0 = p_{10} = 0.05$, $p_1 = p_2 = p_8 = p_9 = 0.15$, $p_3 = \ldots = p_7 = 0.06$. The expected degree of satisfaction then is $\mathbb{E}(X) = (1+2+8+9)\times 0.15 + (3+4+5+6+7)\times 0.06 + 10\times 0.05 = 5$, which makes sense given that the distribution is symmetric around 5. We next compute the second moment of the customers' satisfaction, namely,

$$\mathbb{E}(X^2) = (1+4+64+81)\times 0.15 + (9+16+25+36+49)\times 0.06 + 100\times 0.05 = 35.6,$$

implying a variance of $\operatorname{var}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 = 35.6 - 25 = 10.6$ and a standard deviation of 3.25.

(2) Let X denote a binomial random variable $\mathcal{B}(n,p)$ with an expected value of $\mathbb{E}(X) = np$. Instead of computing the second moment directly from $\mathbb{E}(X^2) = \sum_{x=0}^{n} x^2 \binom{n}{p} p^x (1-p)^{n-x}$, we will take advantage of the definition of the binomial distribution as a sequence of n independent Bernoulli essays with probability p. Let Y_i, with $i \in \{1,\ldots,n\}$, denote these

Bernoulli essays. It then follows that $\text{var}(X) = \text{var}\left(\sum_{i=1}^{n} Y_i\right) = \sum_{i=1}^{n} \text{var}(Y_i)$ given that all the covariances are zero due to the independence between the Bernoulli essays. Now, the variance of Y_i is $\text{var}(Y_i) = \mathbb{E}(Y_i^2) - [\mathbb{E}(Y_i)]^2 = p - p^2 = p\,(1-p)$ and hence $\text{var}(X) = n\,p\,(1-p)$.

(3) Let X denote a uniform random variable in the interval $[\alpha, \beta]$: i.e., $X \sim \mathcal{U}(\alpha, \beta)$. We know that the expected value of X is $\mathbb{E}(X) = (\alpha + \beta)/2$, whereas the second moment is

$$\mathbb{E}(X^2) = \int_{\alpha}^{\beta} \frac{x^2}{\beta - \alpha}\,\mathrm{d}x = \left[\frac{x^3}{3(\beta - \alpha)}\right]\Bigg|_{\alpha}^{\beta} = \frac{\beta^3 - \alpha^3}{3(\beta - \alpha)}.$$

The variance of X then reads

$$\begin{aligned}\text{var}(X) &= \frac{\beta^3 - \alpha^3}{3(\beta - \alpha)} - \frac{(\alpha + \beta)^2}{4} = \frac{4(\beta^3 - \alpha^3) - 3(\beta - \alpha)(\alpha + \beta)^2}{12(\beta - \alpha)} \\ &= \frac{4\,\beta^3 - 4\,\alpha^3 - 3\,\beta^3 + 3\,\alpha^3 - 6\,\alpha\,\beta^2 + 6\,\alpha^2\,\beta - 3\,\alpha^2\,\beta + 3\,\alpha\,\beta^2}{12(\beta - \alpha)} \\ &= \frac{\beta^3 - \alpha^3 - 3\,\alpha\,\beta^2 + 3\,\alpha^2\,\beta}{12(\beta - \alpha)} = \frac{(\beta - \alpha)^3}{12(\beta - \alpha)} = \frac{(\beta - \alpha)^2}{12},\end{aligned}$$

which makes sense in that, as a measure of dispersion, it depends on the square (Euclidean) distance between the support bounds.

(4) Let (X, Y) denote a bivariate random variable with density function

$$f_{XY}(x, y) = \begin{cases} 2 & \text{if } 0 \leq x < y \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

The marginal densities then are $f_X(x) = \int_x^1 2\,\mathrm{d}y = 2(1 - x)$ for $0 \leq x \leq 1$ and $f_Y(y) = \int_0^y 2\,\mathrm{d}x = 2y$ for $0 \leq y \leq 1$, with expected values of $\mathbb{E}(X) = \int_0^1 2x(1 - x)\,\mathrm{d}x = \frac{1}{3}$ and $\mathbb{E}(Y) = \int_0^1 2y^2\,\mathrm{d}y = \frac{2}{3}$, respectively. As for the moments of second order, we start with the covariance, namely,

$$\text{cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\,\mathbb{E}(Y) = \int_0^1 \int_0^y 2xy\,\mathrm{d}x\,\mathrm{d}y - \frac{1}{3}\,\frac{2}{3} = \frac{1}{4} - \frac{2}{9} = \frac{1}{36}.$$

As for the variances, it follows that

$$\begin{aligned}\text{var}(X) &= \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 = \int_0^1 2x^2(1 - x)\,\mathrm{d}x - \frac{1}{9} = \frac{1}{6} - \frac{1}{9} = \frac{1}{18} \\ \text{var}(Y) &= \mathbb{E}(Y^2) - [\mathbb{E}(Y)]^2 = \int_0^1 2y^3\,\mathrm{d}y - \frac{4}{9} = \frac{1}{2} - \frac{4}{9} = \frac{1}{18},\end{aligned}$$

and hence the correlation between X and Y amounts to $1/2$.

The covariance is such that $\text{cov}(aX + b, cY + d) = ac\text{cov}(X, Y)$. The intercepts b and d have no impact in the covariance because it deals with deviations with respect to the mean value and the latter obviously shifts with b and d as much as X and Y, respectively. Further, the correlation is completely invariant to any affine transformation in that $\text{corr}(aX + b, cY + d) = \text{corr}(X, Y)$. The extra level of robustness stems from the fact that we standardize the deviations by the standard deviation, which changes with a and c in the same proportion as X and Y, respectively.

4.6.3 Higher-order moments

In general, we define the kth uncentered moment of a distribution as

$$\mu_k = \mathbb{E}(X^k) = \begin{cases} \sum_{i=1}^{\infty} x_i^k p(x_i) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} x^k f_X(x)\,\mathrm{d}x & \text{if } X \text{ is continuous.} \end{cases}$$

Similarly, we define the centered moments as $\bar{\mu}_k = \mathbb{E}[X - \mathbb{E}(X)]^k$. However, in most situations, we prefer to standardize the random variable not only by subtracting the mean, but also by dividing by the standard deviation, so as to obtain a quantity that is comparable across different random variables. For instance, we define skewness and kurtosis as the standardized third and fourth moments, respectively:

$$\text{sk}(X) = \mathbb{E}\left[\frac{X - \mathbb{E}(X)}{\sqrt{\text{var}(X)}}\right]^3 \qquad \text{and} \qquad \text{k}(X) = \mathbb{E}\left[\frac{X - \mathbb{E}(X)}{\sqrt{\text{var}(X)}}\right]^4.$$

The former gauges how asymmetric is the distribution relative to its mean value, whereas the latter measures how thick the left and right tails are. For instance, it is very well documented that stock returns display negative skewness and very high kurtosis, reflecting the fact that extreme negative returns are more frequent than extreme positive returns. In contrast, changes in exchange rates are typically symmetric around zero, though they also exhibit very high kurtosis implying thick tails.

4.7 Discrete distributions

In this section, we briefly review the discrete distributions we have seen in the previous sections and introduce a couple of other distribution functions that are often useful in practice.

4.7.1 Binomial

Consider a sequence of n independent experiments in which the event A may occur with probability $p = \Pr(A)$. The resulting sample space is $\mathcal{S} = \{\text{all sequences } a_1, \ldots, a_n\}$, where a_i is either A or $\bar{A}$ for $i = 1, \ldots, n$. The random variable X that counts the number of times that the event A occurs has a binomial distribution function $\mathcal{B}(n, p)$ with parameters n (namely, the number of independent essays) and p (namely, the probability of event A). The binomial distribution is such that

$$\Pr(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, \cdots, n. \tag{4.4}$$

The expected value of a binomial distribution is np, whereas the variance is $np(1-p)$. Figure 4.2 depicts the probability distribution function and cumulative distribution function of a binomial random variable with $n = 10$ and $p = 0.25$.

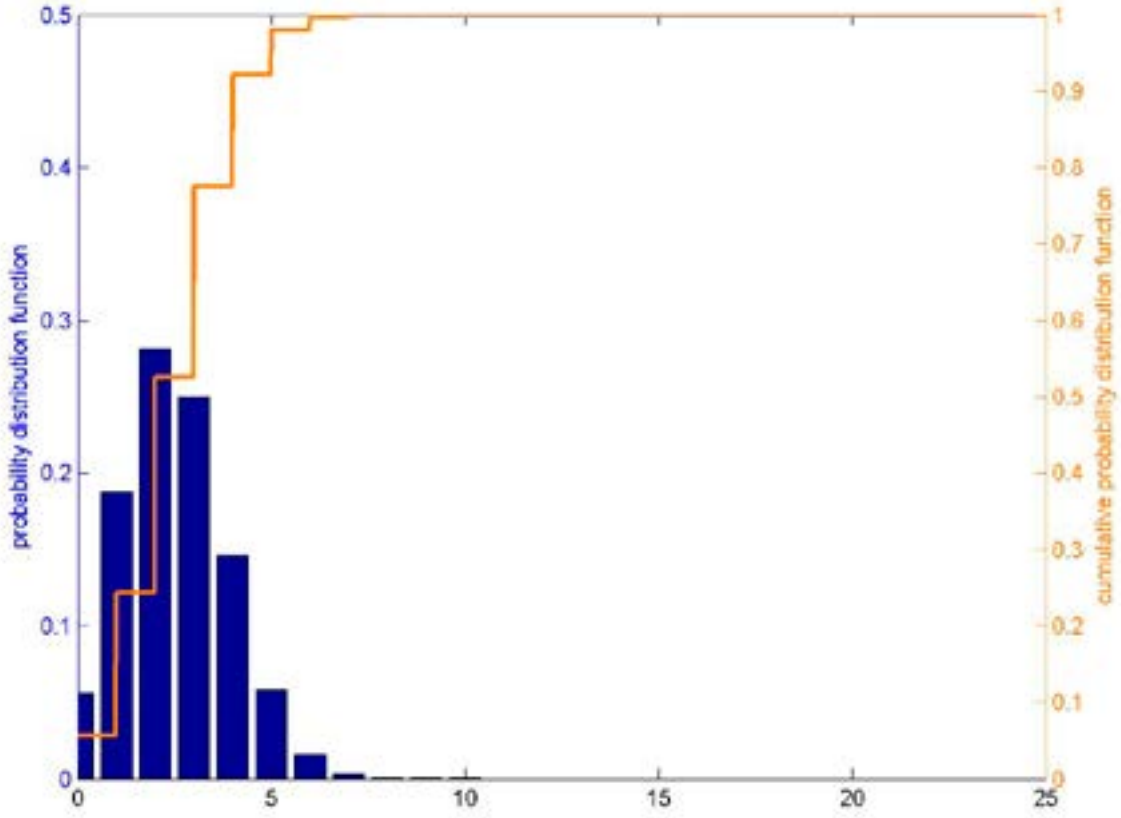

Figure 4.2: The left and right axes correspond to the probability distribution function and cumulative probability distribution function of a binomial random variable with $n = 10$ and $p = 0.25$, respectively.

There are two common violations of the binomial law. The first comes in the form of dependent Bernoulli essays, whereas the second stems from Bernoulli essays with different probabilities. An example of the latter is the Exercise 1 in Section 4.1.1. As for the former, dependence in the essays may entail a very strong impact in the probability distribution function, much more than changing probabilities. Suppose, for instance, the Bernoulli essays exhibit positive dependence. The sum of their values $X = Y_1 + Y_2 + \ldots + Y_n$ will then have a variance of $\text{var}(X) = \sum_{i=1}^{n} \text{var}(Y_i) + 2\sum_{i=2}^{n} Cov(Y_1, Y_i)$. The first term is equal to the variance of the binomial distribution $np(1-p)$, which will be dominated by the second term given that we are summing up n positive covariances. So, under positive/negative

dependence, the binomial distribution would under/overestimate the true variance of X. In contrast, changing probabilities will always overestimate the true variance of X as the first (very extreme!) example illustrates.

Examples

(1) Let $X = Y_1 + \ldots + Y_{20}$, where $Y_1, \ldots, Y_{10}$ are always equal to one and $Y_{11}, \ldots, Y_{20}$ are always equal to zero. The true variance of X is zero, though we would estimate a probability of 1/2 under the assumption that X is binomial and hence a variance of 5.

(2) Let $X = \sum_{i=1}^{50} Y_i + \sum_{i=51}^{100} Y_i$, where the first and second terms form binomial distributions $\mathcal{B}(50, 0.3)$ and $\mathcal{B}(50, 0.7)$, respectively. Assuming a binomial distribution with the average probability of $p = 0.5$ yields an expected value of 50 and a variance of 25. Now, we know from Exercise 1 in Section 4.1.1 that the true probability distribution function of X is

$$\begin{aligned}\Pr(X = k) &= \sum_{k_1=\max(0,k-50)}^{\min(k,50)} \binom{50}{k_1} 0.3^{k_1} 0.7^{50-k_1} \binom{50}{k-k_1} 0.7^{k-k_1} 0.3^{50-k+k_1} \\ &= \sum_{k_1=\max(0,k-50)}^{\min(k,50)} \binom{50}{k_1} \binom{50}{k-k_1} 0.3^{50-k+2k_1} 0.7^{50+k-2k_1},\end{aligned}$$

implying a variance of about 21.

The binomial distribution has a number of applications in practice. The most fruitful financial application is the binomial tree model of asset returns for derivatives pricing. The simplest binomial tree assumes that stock returns are independent over time taking value either Δ or $-\Delta$ with probability 1/2. Running such a model for a large number of periods yield the same solution for the price of a derivative as the Black-Scholes model. This is not surprising given that a binomial distribution $\mathcal{B}(n, p)$ converges to a normal distribution $\mathcal{N}\Big(np, np(1-p)\Big)$ as the number of periods n increases (see Figure 4.3). To make the model more realistic, the most advanced versions of the binomial tree may include time-varying probabilities and/or dependence over time, which of course contradict the assumptions of the binomial distribution.

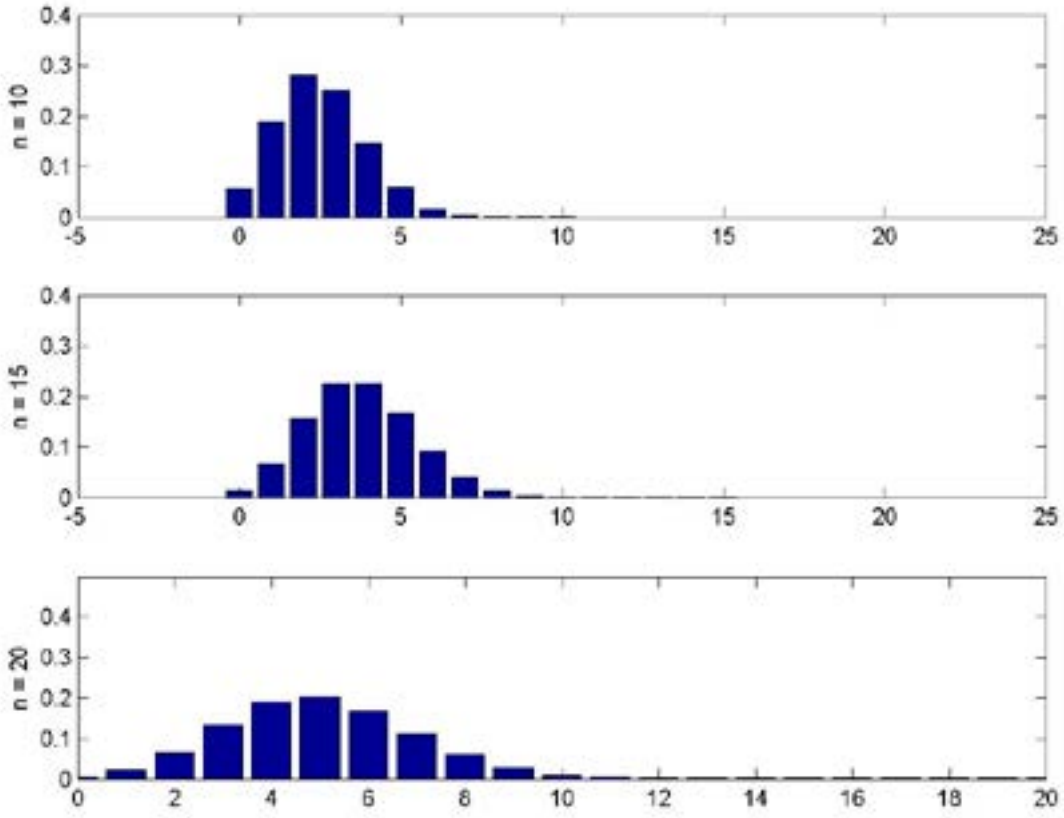

Figure 4.3: The probability distribution function of a binomial random variable resembles more and more the symmetric bell shape of a normal distribution as the number of essays increases. The plots refer to binomial distributions with $p = 0.25$ and $n \in \{10, 15, 20\}$.

Before applying the binomial distribution, we must make sure that the assumptions of constant probability and independence hold. For instance, if we perform a survey by phone between 16:00 and 20:00, the binomial distribution is probably not a good idea given that it is very likely that the audience changes with the time of the day, especially before and after working hours. Similarly, we cannot assume a binomial distribution if the interest lies on the number of stocks with negative returns in a given week. These are not independent events in that there are common factors that affect different stocks at the same time (e.g., the energy sector depends heavily on the price of oil).

4.7.2 Hypergeometric

The hypergeometric distribution arises in situations in which we draw a sample of n units from a population of size N consisting of two distinct groups of size N_1 and $N - N_1$, with $n < \min(N_1, N_2)$, and we define X as the number of units in either one of the groups, say the first group. We know from Section 3.2.4 that the probability of any event A is given by the ratio of the number of possible outcomes in A to the total number of possible outcomes. Accordingly,

$$\Pr(X = x) = \frac{\binom{N_1}{x}\binom{N-N_1}{n-x}}{\binom{N}{n}} \qquad \text{with } 0 \leq x \leq n$$

given that there are $\binom{N_1}{x}$ ways of choosing x from the N_1 units of the first group, $\binom{N-N_1}{n-x}$ ways of choosing the remaining $n - x$ units from the second group, and $\binom{N}{n}$ ways of choosing a sample of n units from a population of size N. Figure 4.4 displays the probability distribution function and cumulative distribution function of a hypergeometric random variable with $N = 50$, $N_1 = 25$, and $n = 10$.

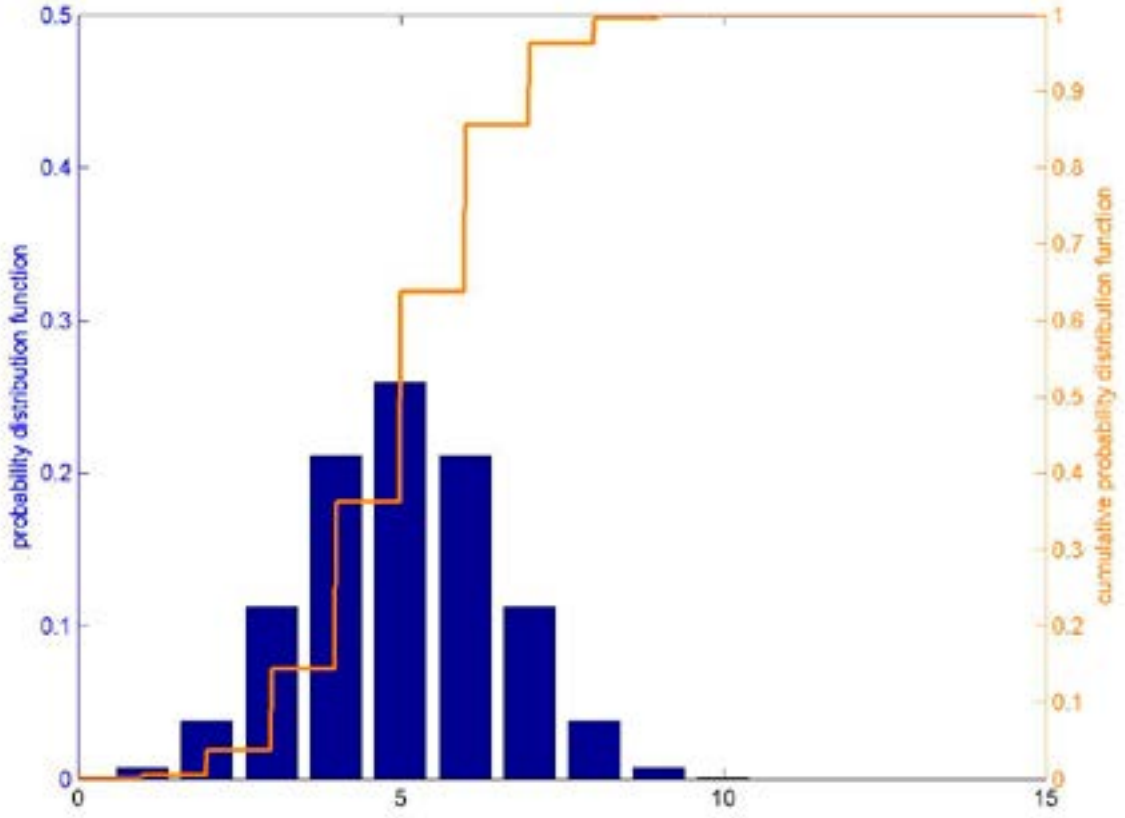

Figure 4.4: The left and right axes respectively correspond to the probability distribution function and cumulative probability distribution function of a hypergeometric random variable with $N = 50$, $N_1 = 25$, and $n = 10$.

The hypergeometric distribution has an expected value of $\mathbb{E}(X) = nN_1/N$ and variance of $\text{var}(X) = n\,\frac{N_1}{N}\left(1 - \frac{N_1}{N}\right)\frac{N-n}{N-1}$. If the population size is much larger than the sample size (i.e., $N \gg n$), the hypergeometric converges to a binomial distribution with probability $p = N_1/N$. Although the binomial approximation entails exactly the same expected value, the variance differs by a small-sample correction factor $\frac{N-n}{N-1}$. This happens because the only difference between the binomial and hypergeometric distributions is that the binomial samples with reposition, whereas the hypergeometric samples without reposition and hence the probability changes in a very particular way as we keep sampling the population. See the quality control problem in the end of Section 3.2.4, for instance.

Example: Consider a population of 87 financial analysts in which 13 are from one of the largest financial institutions in the world. Suppose we wish to form a committee with 10 financial analysts, but there is a concern that too many could come from this big institution. A rough estimate based on the binomial distribution for the probability of observing two financial analysts from the above financial institution is

$$\Pr(X=2) \overset{\mathcal{B}}{\sim} \binom{10}{2}\binom{13}{87}^{2}\binom{74}{87}^{8} \cong 0.275,$$

whereas the true probability given by the hypergeometric is

$$\Pr(X=2) = \frac{\binom{13}{2}\binom{74}{8}}{\binom{87}{10}} \cong 0.294.$$

4.7.3 Geometric

The setup of the geometric distribution is similar to that of the binomial. We conduct independent essays with probability p of success. In contrast to the binomial, the interest lies on how many essays are necessary to observe the next success. This means that we fix the number of success to one and let instead the number of essays to vary randomly. Letting X denote a geometric random variable yields

$$\Pr(X=x) = p(1-p)^{x-1} \qquad \text{with } x = 1, 2, 3, \cdots$$

The name of the distribution comes from the fact that the cumulative probability distribution function depends on the sum of a geometric progression. Figure 4.5 plots the probability distribution function and cumulative distribution function of a geometric random variable with $p = 0.25$.

Summing all possible outcomes yields

$$\sum_{x=1}^{\infty} \Pr(X=x) = \sum_{x=1}^{\infty} p(1-p)^{x-1} = \frac{p}{1-(1-p)} = 1.$$

Along the same lines, it is possible to show that the expected value and variance of the geometric distribution are respectively $1/p$ and $(1-p)/p^2$. In addition, Exercise 2 of Section

4.1.1 shows that a geometric random variable has no memory and hence the link with the exponential distribution in (4.8).

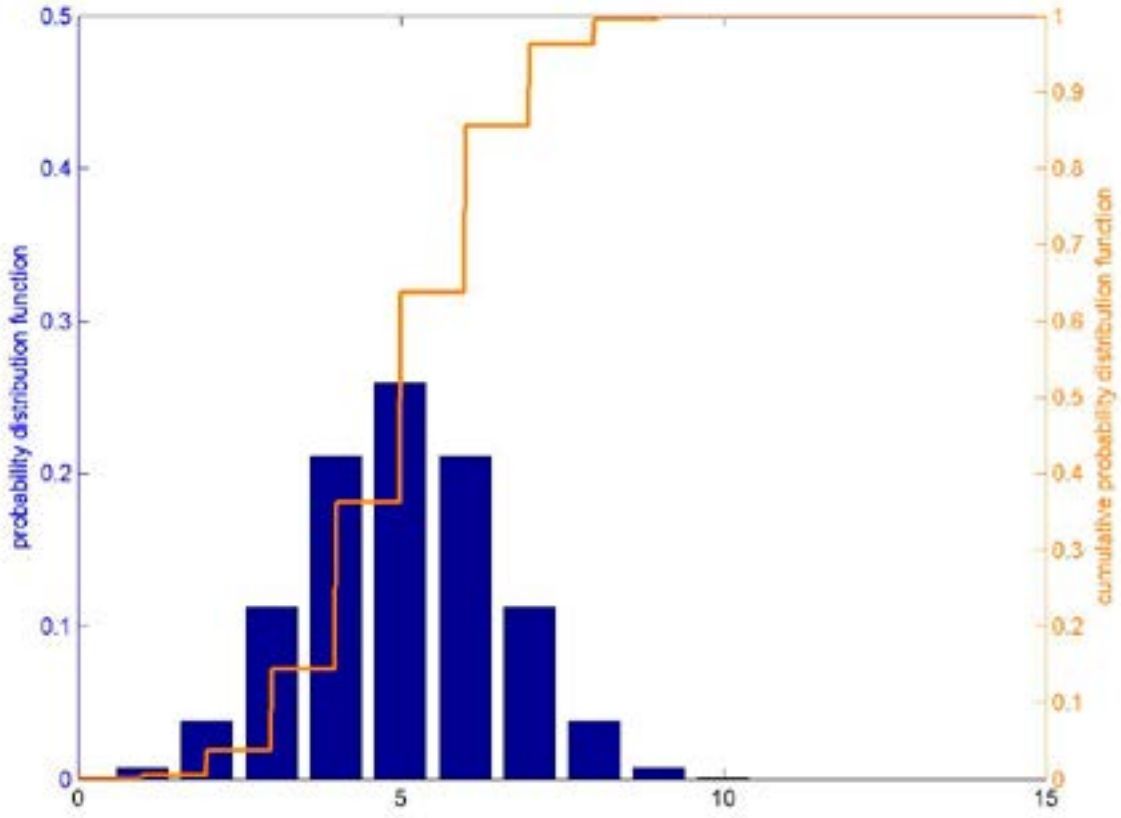

Figure 4.5: The left and right axes correspond to the probability distribution function and cumulative probability distribution function of a geometric random variable with $p = 0.25$, respectively.

4.7.4 Negative binomial

The last variation of the binomial distribution is the negative binomial. It has this name for it inverts the problem of the binomial distribution in that X denotes the number of Bernoulli essays that are necessary to observe k successes. If you wish, the negative binomial extends the geometric distribution in that we wait for k rather than only one success to occur. The

probability distribution function of the negative binomial is

$$\Pr(X = x)\binom{x-1}{k-1}p^k(1-p)^{x-k} \qquad \text{with } x \geq k.$$

Note that we employ the combination $\binom{x-1}{k-1}$ because we know that the last essay must result in a success. A negative binomial variate has an expected value of k/p, with variance of $k(1-p)/p^2$. Figure 4.6 displays the probability distribution function and cumulative distribution function of a negative binomial random variable with $p = 0.25$ and $k = 3$.

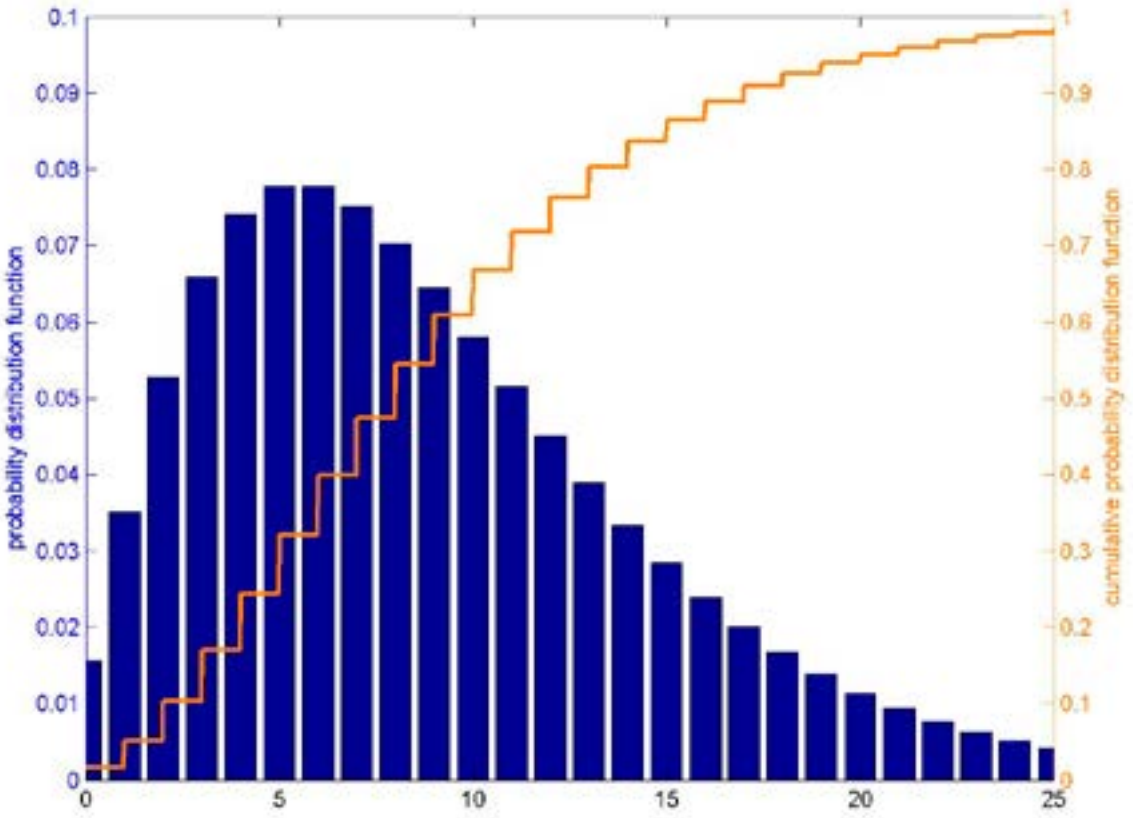

Figure 4.6: The left and right axes correspond to the probability distribution function and cumulative probability distribution function of a negative binomial random variable with $p = 0.25$ and $k = 3$, respectively.

4.7.5 Poisson

The Poisson distribution provides the simplest way to model events that occur at random over time. It assumes that there is a constant arrival rate within a time interval and that events are independent over time. There are a handful of applications for the Poisson distribution in practice. Electricity providers could well employ a Poisson distribution to model the occurrence of electrical tempests in the areas they serve (though storms could exhibit spatial dependence). Call centers typically assume a Poisson distribution for the number of phone calls they receive within a given interval of time (though there could exist an underlying event triggering many calls at approximately the same time, see example below). Commercial

banks normally employ a Poisson distribution to model how many clients will default on their loan payments within a given month (though it is hard to argue for independence in periods of financial distress).

Let X denote the number of events that occur within a given time interval. We say that X has a Poisson distribution with arrival rate λ if its probability distribution function is given by

$$\Pr(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \qquad \text{with } x = 0, 1, 2, \cdots$$

The arrival rate of λ is relative to the time interval of reference. For instance, if we are dealing with an arrival rate λ per minute, then we expect 5λ events within a 5-minute time interval. It turns out that both the expected value and variance of a Poisson are equal to λ. Figure 4.7 portraits the probability distribution function and cumulative distribution function of a Poisson random variable with an arrival rate of $lambda = 5$.

Example: To make the emergency call center more efficient, we must model the number of incoming telephone calls to the emergency number so as to better understand the likelihood of events such as more than 10 phone calls within a 5-minute time interval. We would presumably expect that the number of incoming calls between 18:00 and 19:00 is larger than the number of phone calls between 04:00 and 05:00. This implies that the arrival rate of emergency calls is different depending on the time of the day and hence we have to apply different Poisson distributions for different time periods.

Although the Poisson distribution seems very different from the binomial distribution, there is a close link between them. To see why, let's think about we would model the type of situation that calls for a Poisson distribution by means of a binomial distribution. The first step is to split the time interval of reference into n very short subintervals of equal length. The idea is to have small enough subintervals so as to ensure that the probability of observing more than one event within a subinterval is negligible in comparison with the probability of at most one occurrence. In other words, there is either 0 or 1 event in each

subinterval. This paves the way to the use of a binomial distribution given that we can now model this sequence of subintervals as a sequence of Bernoulli trials.

It remains to decide upon the probability p of observing the event, which should somehow relate to the arrival rate λ of the Poisson distribution. We know that, on average, there are $\lambda \Delta t$ events within a time interval of length Δt. Splitting Δt into n subintervals of time yields on average np events within Δt. It now suffices to equate the expected number of events under the binomial assumption with the arrival rate of the Poisson distribution to obtain $p = \lambda \frac{\Delta t}{n}$. Next, we consider a random variable X that counts the number of events within a time interval of length, say, $\Delta t = 1$.

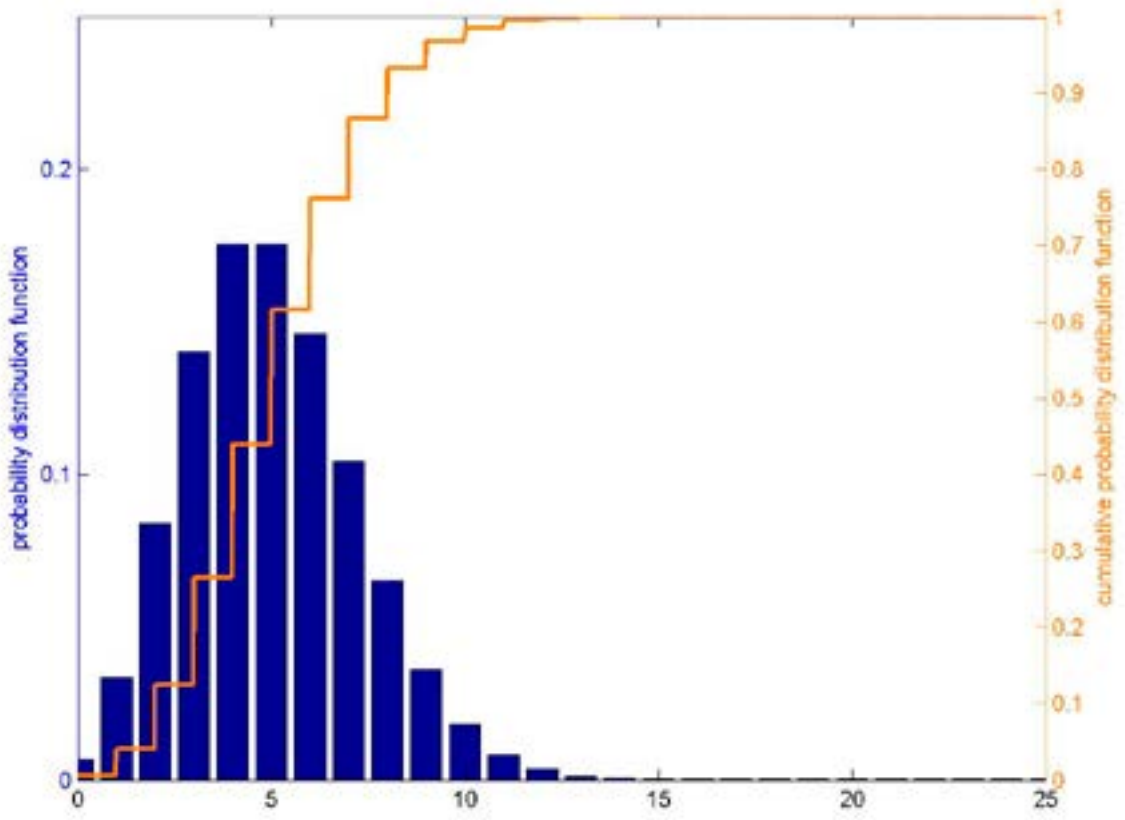

Figure 4.7: The left and right axes correspond to the probability distribution function and cumulative probability distribution function of a Poisson with an arrival rate of $\lambda = 5$, respectively.

We first compute from the binomial distribution the probability of not observing any event: $\Pr(X=0) = (1-p)^n = \left(1-\frac{\lambda}{n}\right)^n$. However, we may think of imposing $n \to \infty$ so as to ensure that the subintervals are small enough. This leads to $\Pr(X=0) = e^{-\lambda}$. We next compute a recursive relation for the probability distribution function under the binomial assumption, namely,

$$\begin{aligned}
\Pr(X=k) &= \binom{n}{k} p^k (1-p)^{n-k} \\
&= \frac{n!}{k!(n-k)!}\, p^{k-1}(1-p)^{n-k+1}\, \frac{p}{1-p} \\
&= \binom{n}{k-1} p^{k-1}(1-p)^{n-k+1}\, \frac{n-k+1}{k}\, \frac{p}{1-p} \\
&= \frac{(n-k+1)p}{k(1-p)}\, \Pr(X=k-1) \\
&= \frac{\lambda-(k-1)p}{k(1-p)}\, \Pr(X=k-1)
\end{aligned}$$

given that $p = \lambda/n$. Taking limits ($n \to \infty$ or, equivalently, $p \to 0$) then yields $\Pr(X=k) = \frac{\lambda}{k}\Pr(X=k-1)$. In particular, $\Pr(X=1) = \lambda\, e^{-\lambda}$ for $k=1$, $\Pr(X=2) = \frac{\lambda^2}{2}\, e^{-\lambda}$ for $k=2$, and so on. In general, $\Pr(X=k) = \frac{\lambda^k}{k!}\, e^{-\lambda}$ just as in the Poisson distribution.

The above discussion about the intimate link between the binomial and Poisson distributions is interesting because it motivates why so many credit institutions employ a Poisson distribution to model the arrival rate of credit defaults. Although the binomial distribution is theoretically more suitable, it is difficult to handle the probability distribution function of a binomial random variable if n is too large. As the probability of a credit default is typically very small (at least in a developed economy and in periods of normal activity), this is the ideal setup for a Poisson approximation of the binomial distribution.

4.8 Continuous distributions

In this section, we briefly review the few continuous distributions that we have seen in the previous sections. In addition, we introduce a series of other continuous distributions, most of them deriving from the normal distribution. Also known as the Gaussian distribution, the latter is the most important distribution in statistics not only because it is often a good assumption in practice, but also because it naturally arises in theory to approximate the distribution of the sample mean of any random variable (regardless of its distribution). This last result is known as the central limit theorem, which we will study in Section 5.2.

4.8.1 Uniform

This is the continuous counterpart of equiprobable events in that any interval of a given length within the support of the distribution will have exactly the same probability. More formally, let X denote a uniform random variable in the interval $[\alpha, \beta]$, with density function

$$f_X(x) = \begin{cases} \frac{1}{\beta-\alpha}, & \text{if } \alpha \leq x \leq \beta \\ 0, & \text{otherwise.} \end{cases}$$

We have already shown that the expected value of a uniform random variable is given by the average between the upper and lower limits of the support, i.e., $\mathbb{E}(X) = \frac{\alpha+\beta}{2}$, whereas the variance is $\text{var}(X) = \frac{(\beta-\alpha)^2}{12}$.

Example: Suppose that the delay of a given tram is uniformly distributed between 0 and

20 minutes during winter. This means that the probability of observing a delay of at least 8 minutes is

$$\begin{aligned}\Pr(X \geq 8) &= 1 - \Pr(X \leq 8) = 1 - F_X(8) \\ &= \int_8^{20} \frac{1}{20}\,\mathrm{d}x = 1 - \int_0^8 \frac{1}{20}\,\mathrm{d}x \\ &= 1 - \frac{8}{20} = \frac{12}{20} = 3/5.\end{aligned}$$

4.8.2 Exponential

The exponential distribution arises in a setting very similar to the one of the Poisson distribution. The difference is that the interest now lies on the time spell we must wait to observe the next event. As in the Poisson context, we assume that events are independent over time and the arrival rate is constant. Applications abound in quality control, including fatigue and reliability analysis. In finance, there is a new strand of the literature that aims to model the time between trades so as to better understand market activity. In addition, it is also interesting to observe how much time it takes to observe a change in prices for it conveys information about market volatility. Finally, labor economists are keen on carrying out duration analyses so as to study unemployment spells and time to promotions. In general, the exponential distribution plays a major role in duration analysis regardless of whether the duration has an economic, financial or quality-control interpretation.

The density function of an exponential variate is $f_X(x) = \frac{1}{\lambda}\, e^{-\lambda x}$ with $x > 0$ (zero otherwise), whereas the survival function $S_X(x) = 1 - F_X(x) = \Pr(X > x)$ is given by

$$\Pr(X > x) = \int_x^\infty f_X(t)\,\mathrm{d}t = 1 - \int_0^x \frac{1}{\lambda}\, e^{-t/\lambda}\,\mathrm{d}t = e^{-x/\lambda}.$$

This naturally means that the cumulative distribution function is $F_X(x) = 1 - e^{-x/\lambda}$, though it is more common in duration analysis to talk about the survival function. Both the expected value and the standard deviation of the exponential distribution are equal to λ, which reminds us again of the Poisson distribution whose expected value is equal to the variance. Figure

4.8 portraits the probability density and distribution functions of a standard exponential random variable (i.e., $\lambda = 1$).

As aforementioned, the exponential distribution somewhat resembles the geometric distribution in that it features no memory given that

$$\Pr(X > s + t | X > s) = \frac{\Pr(X > s + t)}{\Pr(X > s)} = \frac{e^{-(s+t)/\lambda}}{e^{-s/\lambda}} = e^{-t/\lambda} = \Pr(X > t).$$

The probability of waiting another interval of time of at least t is constant regardless of how much we have been waiting for.

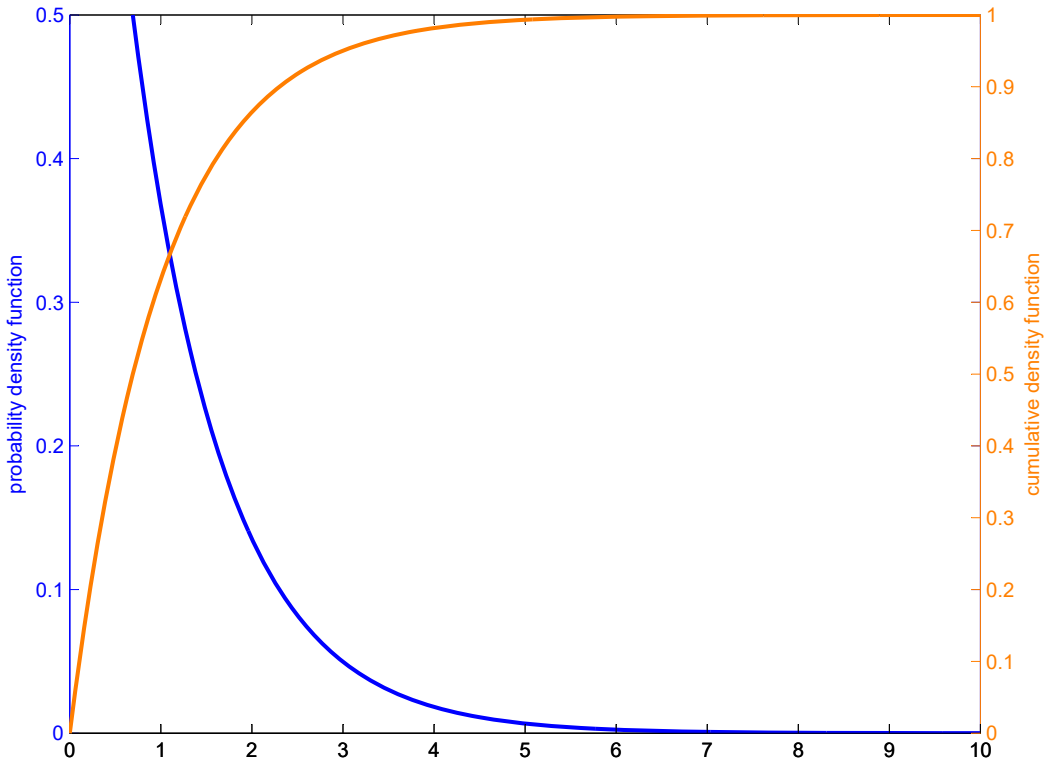

Figure 4.8: The left and right axes respectively correspond to the probability density and distribution functions of a standard exponential random variable, that is to say, an exponential with $\lambda = 1$.

4.8.3 Normal and related distributions

The normal (or Gaussian) is the fundamental distribution in statistics not only because it naturally appears in a wide array of situation, but also because of the central limit theorems that ensure it provides a good approximation in large samples for the sample mean of almost any random variable. The normal distribution is also easy to manipulate for a number of reasons. First, it is completely characterized by its mean and variance, which we denote by μ and σ^2, respectively. Second, in contrast to the distributions we have seen so far, there is no connection between the mean and variance of a Gaussian variate. This confers an extra level of flexibility to the normal distribution, explaining why it seems to work well in all sorts of situations. Third, the normal distribution is close under affine transformations in that a linear combination of Gaussian random variables is also Gaussian.

The density function of a normal random variable is given by

$$f_X(x) = \frac{1}{\sqrt{2\,\pi\,\sigma^2}}\,\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \qquad \text{with } -\infty < x < \infty.$$

We denote a normal random variable with mean μ and variance σ^2 by $X \sim \mathcal{N}(\mu, \sigma^2)$. There is no closed-form solution for the cumulative distribution function of a normal random variable and hence we must tabulate it. Naturally, it would be impossible to evaluate the distribution function of $X \sim \mathcal{N}(\mu, \sigma^2)$ for every value in the real line and for every mean-variance combination.

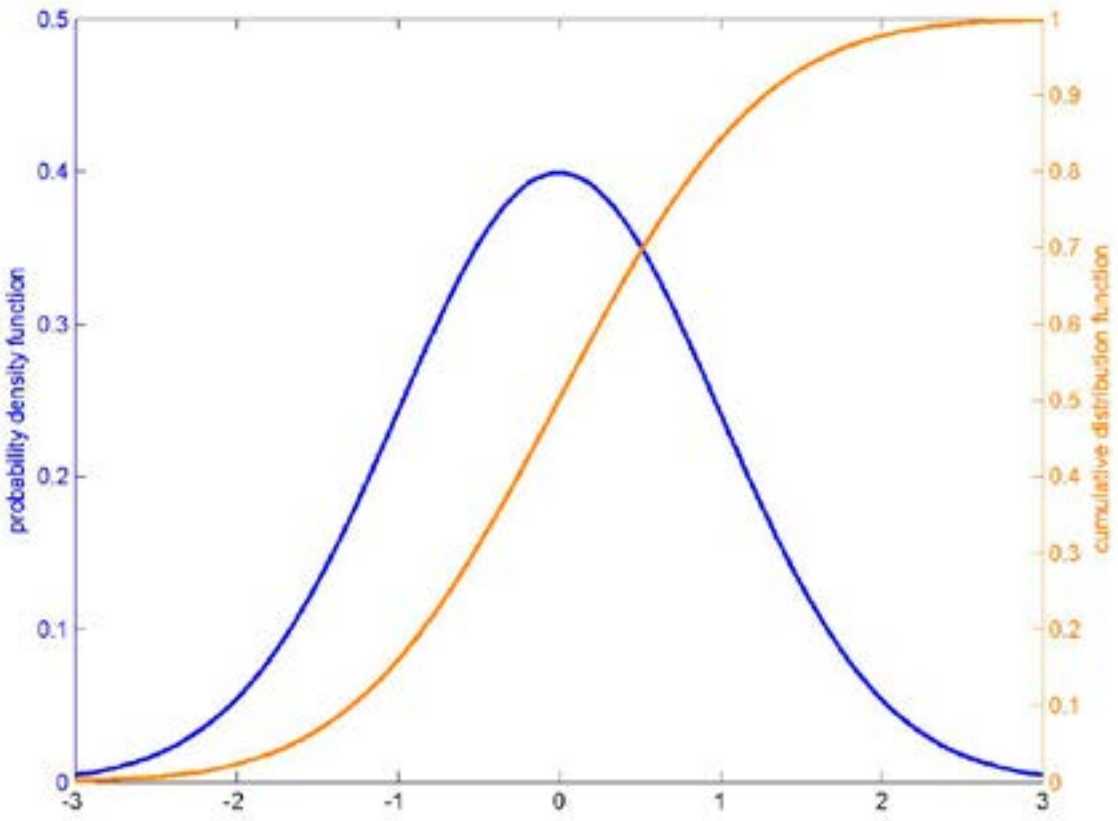

Figure 4.9: The left and right axes correspond to the probability density and distribution functions of a standard normal random variable, respectively.

To circumvent this problem, we tabulate only the standard normal distribution $\mathcal{N}(0, 1)$, which has zero mean and unit interval, given that we can obtain any other normal distribution

by means of a simple affine transformation: $Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$ if $X \sim \mathcal{N}(\mu, \sigma^2)$. The standard normal distribution and density play such a major role in statistics that we denote them by $\Phi(\cdot)$ and $\phi(\cdot)$, respectively. The latter is $\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} z^2\right)$ for $z \in \mathbb{R}$. Figure 4.9 displays the probability density and distribution functions of a standard normal random variable.

The normal distribution is symmetric and hence all odd centered moments are equal to zero. In contrast, the variance of the normal distribution determines the magnitude of every even centered moment. For instance, the kurtosis of the normal distribution is $\mathrm{k}(X) = \mathbb{E}\left[\frac{X-\mu}{\sigma}\right]^4 = \mathbb{E}(Z^4) = 3$ and that's why some people refer to $\mathrm{k}(X) - 3$ as excess kurtosis. Figure 4.10 shows precisely how increasing the dispersion affects the shape of a normal density function and hence the probability of observing extreme realizations.

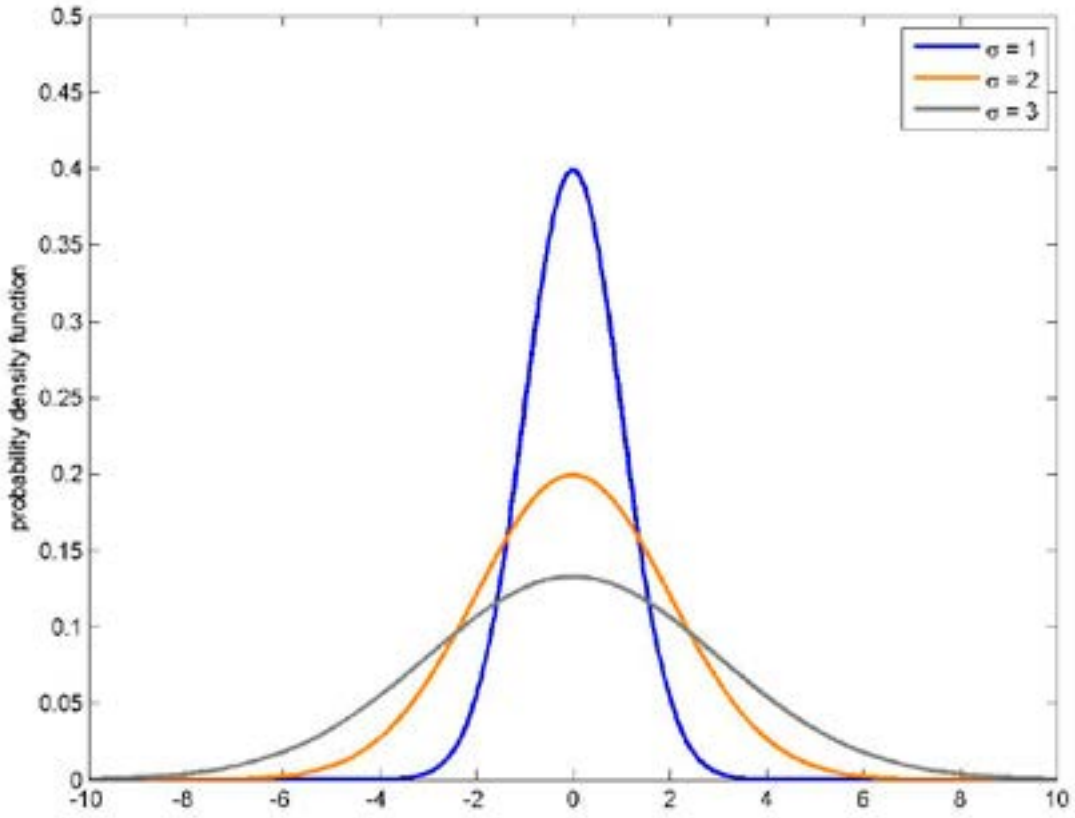

Figure 4.10: The probability density function of a normal random variable with zero mean and standard deviation $\sigma \in \{1, 2, 3\}$.

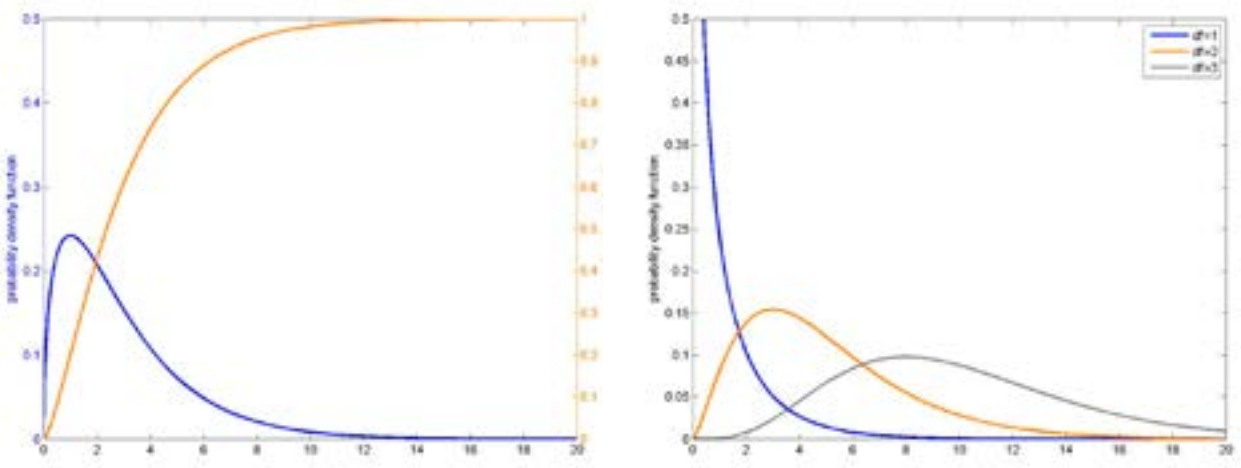

Figure 4.11: The first panel plots the probability density and distribution functions of a chi-square random variable in the left and right axes, respectively. The second panel shows how the shape of the chi-square density changes with the degrees of freedom.

The normal distribution gives way to a number of interesting distributions depending on how we transform it. In what follows, we discuss two distributions that derive from the Gaussian distribution and are of particular interest in the context of statistical inference. The first is the chi-square distribution, which consists of the sum of a number of squared independent standard normal random variables. Let $Z_i \sim \mathcal{N}(0,1)$ denote a sequence of mutually independent standard normal distributions for $i = 1, \ldots, N$. It then follows that $\chi^2_N = \sum_{i=1}^{N} Z_i^2$ is a chi-square distribution with N degrees of freedom. The mean of χ^2_N is N, whereas its variance is twofold amounting to $2N$. The chi-square is important in a number of situations. For instance, the chi-square distribution arises in a very natural manner if we wish to compute the probability that a standard normal random variable belongs to a symmetric interval around zero. Figure 4.11 not only plots the probability density and distribution functions of a chi-square random variable, but also shows how it changes with the degrees of freedom.

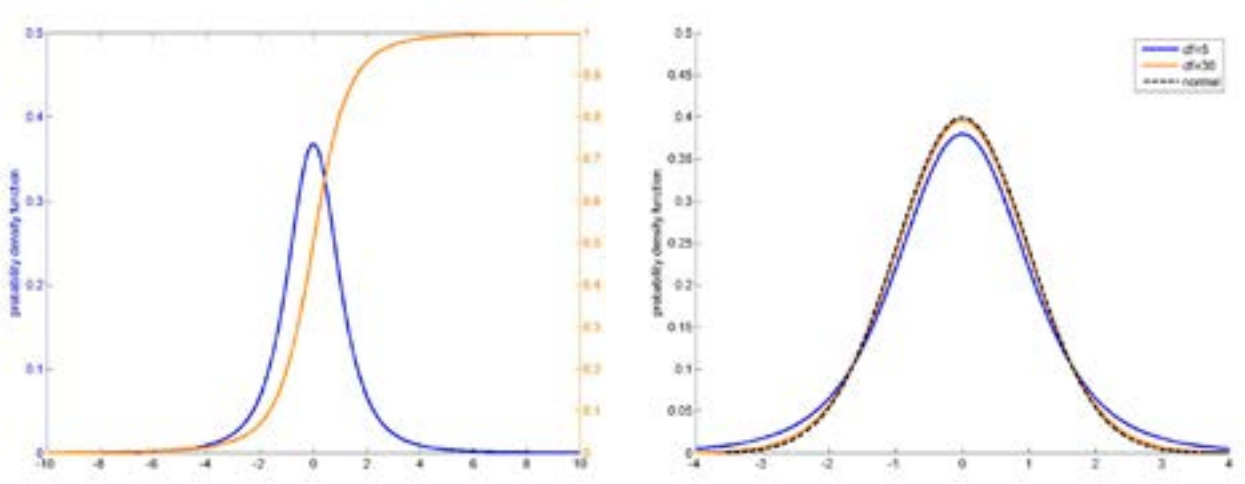

Figure 4.12: The first panel plots the probability density and distribution functions of a t-student random variable in the left and right axes, respectively. The second panel shows how the shape of t-student density varies with the degrees of freedom.

The second is the t-student distribution, which stems from a ratio of a standard normal distribution to the square root of an independent chi-square distribution divided by its degrees of freedom. In particular, we denote a t-student with N degrees of freedom by $t_N = \frac{Z_0}{\sqrt{\frac{1}{N}\sum_{i=1}^{N} Z_i^2}}$, where Z_i's are independent standard normal distributions for $i = 0, 1, \ldots, N$. The t-student is symmetric around the origin and hence has mean zero. In turn, the variance of a t-student random variable is $N/(N-2)$, with N denoting the degrees of freedom. It is easy to see that the variance is ill-defined if there are not enough degrees of freedom in that $N \leq 2$ implies a negative variance. This is true in general for the t-student distribution in that the kth moment exists if and only if the degrees of freedom exceed k. We will see later that the t-student distribution is paramount to hypothesis testing under the assumption of normality.

Figure 4.12 depicts the probability density and distribution functions of a t-student random variable with 3 degrees of freedom in the first panel, whereas the second panel illustrates how the shape of its density changes with the degrees of freedom. In particular, it is noticeable that the t-student converges to a standard normal distribution as the degrees of freedom increase.

Chapter 5

Random sampling

They way we collect the data is extremely important because, ideally, we would like to end up with a random sample. The latter consists of independent and identically distributed (iid) observations from the population. This is the ideal setup for two reasons. If data are independent, then it is easy to characterize the joint distribution because it is equivalent to the product of the marginals. In addition, if the data come from a common distribution, the marginals are all identical, thereby depending on the same vector $\boldsymbol{\theta}$ of parameters. We formalize these ideas using the joint density function of a random sample $\boldsymbol{X} = (X_1, \ldots, X_N)$, namely, $f_{\boldsymbol{X}}(x_1, \ldots, x_N) = \prod_{i=1}^{N} f_{X_i}(x_i) = \prod_{i=1}^{N} f_X(x_i; \boldsymbol{\theta})$. The first equality follows from independence, whereas the second ensues from the fact that the elements of the random sample are identically distributed. We typically represent a random sample by $X_i \sim \text{iid } f_X(\cdot; \boldsymbol{\theta})$ for $i = 1, \ldots, N$.

Data collection is not easy. It is indeed quite difficult to design a data sampling procedure that is free of any bias. The most common problems in business and economics relate to censorship, selection, survivorship, and no-answer biases. Censorship bias takes place whenever we cannot observe data within a given interval. For instance, if there are price limits in a stock exchange, we cannot observe stock prices either above the upper limit or below the lower limit given by the maximum daily oscillation. In such a context, censoring may occur for, even if the equilibrium price moves above or below the limits, we observe at most a price at one of the limits. The similar problem arises in exchange-rate target zones.

The only difference is that a successful speculative attack against the currency may cause the break of the target zone (i.e., price limits). A very different situation in which censorship plays a major role is in labor economics. For instance, if we are measuring how much time it takes for an individual to obtain a promotion, our data set will invariably include individuals who haven't been promoted yet. So, the most we can say is that the time to promotion of these individuals is larger than the number of periods we have been observing them in our sample.

Selection bias occurs in the event that the sampling procedure is such that the data tend to come from a specific group within the population. For instance, in most developing countries, women must decide whether they join the work force or stay at home full time taking care of the house chores, whereas men rarely have such an option. This means that we cannot directly compare the salary of male and female workers. They are different not only in gender, but also because female workers have taken a previous decision to join the labor market, whereas men didn't. Taking such a decision shows some degree of ambition and determination that is probably correlated with productivity and hence with salary. That is why a simple comparison of wage differentials will normally underestimate the discrimination against women in the labor market. The same reasoning applies to immigrants, as well. The simple fact that they have taken a previous decision to migrate, while others in the same situation didn't, indicates that they form a different group (perhaps more ambitious, focused, and determined). In finance, selection bias may also affect asset returns through a liquidity (rather than competition) channel. Illiquid assets by definition trade less frequently than liquid assets and hence they are much more likely to exhibit price staleness. If transaction prices do not change, then the return is zero. However, zero returns are not really reflecting the true change in the value of the asset in that they are merely an artifact due to illiquidity. This means that we cannot treat zero returns and nonzero returns in the same way.

Survivorship bias arises whenever we are looking at a sample of individuals/firms/units resulting from some sort of competition. Although it is much more natural to think about

survivorship bias in biology, where evolution establishes intense competition among different genes, examples abound in economics, finance, and management. For instance, hedge funds that perform poorly end up managing less funds that those doing well. As most indices are weighted by assets under management, they tend to reflect more the performance of the hedge funds that perform well over time, overestimating the overall performance in the industry. In fact, hedge funds that perform systematically poorly end up closing their doors, thereby disappearing from databases at a certain point in time. Survivorship bias then arises very strongly if we collect data by sampling the returns of all funds that currently exist since some date in the past. To avoid such a bias, we must first choose the starting data and then collect the data of all funds were operational since then. In this way, the data set would include not only the successful funds that are still in action, but also those that did do very well and ceased to exist.

The no-answer bias is very common in surveys. People who have strong opinions, especially negative, are typically much more inclined to answer a survey. That's why most lecturers are very keen to publicize teaching evaluation surveys to students. If they don't, it is very likely that the answers will have a negative bias for students that do not have many criticisms will presumably not bother to answer the survey as much as the students who are not happy. In work environments, we could well argue that workaholics tend not to respond work-unrelated surveys (e.g., menu of the eatery) for they prefer to dedicate their time to more productive tasks (it is also very likely that they bring their own sandwich from home to spend less time in lunch breaks!). This means that work-unrelated surveys will not represent entirely the views of the population in the firm due to the lack of answers from the workaholics.

Example: Suppose a commercial bank would like to work out how many days on average it takes to process a cheque. The cost of sampling all cheques is prohibitively high and hence the person in charge decides to draw a sample of 1,000 cheques. The snag is how to draw a

random sample given that we cannot simply tag all cheques with a number and then draw from a discrete uniform distribution. One solution is to draw every mth cheque that the bank processes until we observe 1,000 cheques. This type of sampling is not entirely random in that we will never observe two consecutive cheques (from the same firm, perhaps!) in the sample, but it should do the trick reasonably well.

The above example illustrates well the fact that random sampling is an abstract notion. In some situations, it is virtually impossible to draw a completely random sample, and hence we must do with samples that are " random enough". The next series of examples are more concrete in that they establish alternative procedures to draw a random sample in all sort of setups.

Examples

(1) Suppose a marketing firm wishes to interview 1,000 households in a town. One solution is to draw from a discrete uniform distribution random numbers that identify households given their addresses (or post codes). The interviewer then visit the address between 15:00 and 18:00 and, if no one answer, that we eliminate that address and replace by another drawn at random from the same discrete uniform distribution.

(2) Suppose a bookstore in the university campus wishes to evaluate the stock of textbooks before the beginning of the term. We may draw a random number that identifies a given location in the shelves and then check the textbook in that location as well as the 50 closest textbooks.

(3) A manager would like to assess the performance of the cleaning staff. Her assistant comes up with two assessment strategies. The first involves inspecting 15 offices completely at random, whereas the second strategy chooses one office at random from each of the three floors of the building and then inspect them as well as the 9 offices closest to each of them. The time to completion is the same for both strategies despite the fact the second strategy inspects the double of offices. The manager decides for the first strategy for it really entails a random sample. She rightly explains to her assistant that, even though the second strategy seems more efficient at first glance, it is less convenient for it would not generate a random sample. The reason is simple. The allocation of the cleaning staff is typically by wing and floor, and hence, by restricting attention to a given area of the floor (i.e., the neighborhood of the sampled office), she would risk to assess the work of a particular subset of janitors rather than the overall performance of the cleaning staff.

5.1 Sample statistics

Let $\boldsymbol{X} = (X_1, \ldots, X_N)$ denote a random sample with density function $f_X(\cdot; \boldsymbol{\theta})$. We don't know the true value of the vector $\boldsymbol{\theta}$ of population parameters and so the goal is to infer it from the realization of the random sample, say, $\boldsymbol{x} = (x_1, \ldots, x_n)$. Note that $\boldsymbol{x}$ is just

one possible realization for the random sample $\boldsymbol{X}$ with probability mass given by the joint density function evaluate at $\boldsymbol{x}$, namely, $f_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\theta}) = \prod_{i=1}^{N} f_X(x_i;\boldsymbol{\theta})$.

Example: Let the random variable X come from a discrete uniform distribution that takes integer values between 1 and 6. This means that the sample space is $\{1,2,3,4,5,6\}$, whose elements are drawn with probability 1/6. Suppose now that we take three random samples of three observations, say $(1,1,4)$, $(2,4,5)$, and $(2,3,2)$. Their sample means are respectively 2, 11/3, and 7/3, though the expected value of X is $(1+2+3+4+5+6)/6 = 7/2$.

What we wish to illustrate with the above example is that the sample mean is also a random variable, whose distribution depends on the distribution from which we draw the random sample. Accordingly, the value we observe for the sample mean varies with the sample we actually draw. Each different sample yields a distinct value for the sample mean. Needless to say, this holds for any function $g(\boldsymbol{X}) = g(X_1,\ldots,X_N)$ of the sample, which we call sample statistic. To conduct inference, we must always bear in mind that a sample statistic is random and hence we must determine its distribution, which we call sampling distribution.

Example: Let $\boldsymbol{X} = (X_1,\ldots,X_N)$ denote a random sample from a normal distribution with mean μ and variance σ^2, i.e., $X_i \sim \text{iid } \mathcal{N}(\mu,\sigma^2)$. The sample mean $\bar{X}_N = \frac{1}{N}\sum_{i=1}^{N} X_i$ is a random variable with expected value

$$\mathbb{E}(\bar{X}_N) = \mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}(X_i) = \frac{1}{N}\sum_{i=1}^{N}\mu = \frac{1}{N}\left(N\mu\right) = \mu$$

and variance

$$\text{var}(\bar{X}_N) = \text{var}\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right) = \frac{1}{N^2}\sum_{i=1}^{N}\text{var}(X_i) = \frac{1}{N^2}\sum_{i=1}^{N}\sigma^2 = \frac{1}{N^2}\left(N\sigma^2\right) = \sigma^2/N.$$

Finally, as the sample mean is a linear combination of normal random variables, it is also normal. We thus conclude that $\bar{X}_N \sim \mathcal{N}(\mu,\sigma^2/N)$, which differs from the normal distribution from which we draw the sample.

In the next section, we will show that, as long as the sample size is large enough, normality is often a very good approximation for most sampling distributions, even if the distribution from which we draw the sample is not normal (and even unknown). The next example concludes this section by depicting such a situation.

Example: The number of transactions on the Macau stock market is on average of 62,000 trades per week with a standard deviation of 7,000. Suppose now that we take note of the number of transactions per week within a year. The expected value of the sample mean is $\mathbb{E}(\bar{X}_N) = 62,000$ trades per week, with a variance of $\text{var}(\bar{X}_N) = 7,000^2/52 = 942,307.69$ given that there are 52 weeks in a year. The normal approximation for the sample mean distribution yields a probability of observing a sample mean at least two standard deviations away from its mean, i.e., $\Pr\left(\left|\frac{\bar{X}_N - 62,000}{942,307.69}\right| > 2\right)$, of about 5%.

5.2 Large-sample theory

In this section, we will discuss how to approximate the sampling distribution of a statistic in general. In particular, we will talk about asymptotic approximations in that we will let the sample size grow to infinity. Although these results hold only in the limit, we will see that they often provide a good guidance to the behavior of most statistics as long as the sample size is large enough. What is 'large enough' will of course depend on the task at hand. If we employ asymptotic results to approximate the distribution of the mean of a random sample coming from a uniform distribution, large enough could mean even 5 observations.

We first discuss what we mean by limit theory in the context of random variables. In particular, we establish several modes of convergence for sequences of random variables. Next, we establish some convergence results for sample means given that they play a major role in statistics. There are two asymptotic results that permeate almost any problem of statistical inference: Laws of large numbers say that sample means converge (in some sense that we will precise in the next section) to the population mean, whereas central limit theorems single out the conditions under which we can approximate the distribution of a sample mean with a normal distribution.

5.2.1 Modes of convergence

Let $X_1, X_2, \ldots$ denote a sequence of random variables, which we denote simply by X_N despite of the abuse of notation. We say that X_N converges in probability to a constant a, which we denote by $X_N \xrightarrow{p} a$, if $\lim_{N\to\infty} \Pr(|X_N - a| < \epsilon) = 1$ for any $\epsilon > 0$. We call a the probability limit of X_N, which we denote by $\operatorname{plim}_{N\to\infty} X_N = a$. The probability limit is a natural generalization of the mathematical notion of a limit. A stronger mode of convergence follows if we switch the order of the limit and probability operators. Indeed, it is much more stringent to impose that $\Pr\left(\lim_{N\to\infty} X_N = a\right) = 1$ for it constrains the function that the random variables in the sequence X_N use to map the sample space to the real line. This is what we call almost sure convergence or, equivalently, convergence with probability

one, which we denote by $X_N \xrightarrow{a.s.} a$. Mean squared convergence denotes a situation in which a sequence X_N of random variables converge in mean square to a constant a in that $\lim_{N\to\infty} \mathbb{E}(X_N - a)^2 = 0$. We then say that a is the mean square limit of the sequence X_N and write $X_N \xrightarrow{m.s.} a$.

These modes of convergence admit various extensions. For instance, we can consider the convergence to a random variable by writing that $X_N \xrightarrow{fmc} X$ if and only if $X_N - X \xrightarrow{fmc} 0$, where $fmc \in \{p, a.s., m.s\}$ denotes your favorite mode of convergence. In addition, we can also think of sequences of random vectors (or matrices) rather than sequences of scalar random variables. In this event, it suffices to apply your favorite mode of convergence element-wise.

The modes of convergence we have seen so far are pretty strong in that the sequence of random variables in the limit becomes degenerate given that it converges to a simple constant (without any form of randomness). In contrast, convergence in distribution deals with limiting distributions. We say that X_N converges in distribution to X if $\lim_{N\to\infty} F_{X_N}(x) = F_X(x)$ for any $x \in \mathbb{R}$. We call F_X the asymptotic (or limiting) distribution of X_N and denote this mode of convergence by $X_N \xrightarrow{d} X$ or $X_N \xrightarrow{d} F_X$.

Convergence in distribution is the weakest of the convergence definitions in that we constrain only the distribution of the random variable and not the values it take. In particular, it is possible to show that

$$\left.\begin{array}{l} X_N \xrightarrow{a.s.} X \\ X_N \xrightarrow{m.s.} X \end{array}\right\} \quad \Rightarrow \quad X_N \xrightarrow{p} X \quad \Rightarrow \quad X_N \xrightarrow{d} X.$$

As before, we can also extend the notion of convergence in distribution to a multivariate setting (i.e., to random vectors).

Finally, we conclude this section by showing how to manipulate and combine the different modes of convergence. We first note that $g(X_N) \xrightarrow{p} g(X)$ if $X_N \xrightarrow{p} X$ and that $g(X_N) \xrightarrow{d} g(X)$ if $X_N \xrightarrow{d} X$ provided that $g(\cdot)$ is a continuous function. This pair of results is known as the continuous mapping theorem. The Slutsky theorem is one of the various applications

of the continuous mapping theorem, ensuring not only that $X_N + Y_N \xrightarrow{d} X + a$ if $X_N \xrightarrow{d} X$ and $Y_N \xrightarrow{p} a$, but also that $X_N Y_N \xrightarrow{p} 0$ if $X_N \xrightarrow{d} X$ and $Y_N \xrightarrow{p} 0$. Finally, it is easy to see that $A_N X_N \xrightarrow{d} A X$ if $X_N \xrightarrow{d} X$ and $A_N \xrightarrow{p} A$, as well.

Example: Let $\boldsymbol{X}_N$ denote a k-dimensional random vector such that $\boldsymbol{X}_N \xrightarrow{p} \boldsymbol{\mu}$ and $\sqrt{N}(\boldsymbol{X}_N - \boldsymbol{\mu}) \xrightarrow{d} \boldsymbol{Z}$, where $\boldsymbol{\mu}$ is a vector of constants and $\boldsymbol{Z}$ is a random vector with some known distribution. Let now $\boldsymbol{\alpha}(\cdot) : \mathbb{R}^k \mapsto \mathbb{R}^r$ denote a multivariate function with continuous first-order derivatives. It then follows from a simple first-order Taylor expansion that $\sqrt{N}\Big(\boldsymbol{\alpha}(\boldsymbol{X}_N) - \boldsymbol{\alpha}(\boldsymbol{\mu})\Big) \xrightarrow{d} \boldsymbol{A}\,\boldsymbol{Z}$, where $\boldsymbol{A} = \frac{\partial \boldsymbol{\alpha}(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}'}$. To appreciate why, note that the first-order Taylor expansion is $\boldsymbol{\alpha}(\boldsymbol{X}_N) = \boldsymbol{\alpha}(\boldsymbol{\mu}) + \frac{\partial \boldsymbol{\alpha}(\boldsymbol{\mu}_*)}{\partial \boldsymbol{\mu}'}(\boldsymbol{X}_N - \boldsymbol{\mu})$ with $\boldsymbol{\mu}_* = \lambda\boldsymbol{\mu} + (1-\lambda)\boldsymbol{X}_N$ for some λ in the unit interval. Multiplying both sides by $\sqrt{N}$ then yields the result given that $\boldsymbol{\mu}_* \xrightarrow{p} \boldsymbol{\mu}$ for any λ provided that $\boldsymbol{X_N} \xrightarrow{p} \boldsymbol{\mu}$.

The above example illustrates the delta method, which consists of a useful tool to derive the asymptotic distribution of a known function of a statistic. It is interesting to note that, even though the example assumes that $\boldsymbol{X}_N$ converges in probability to a vector of constants, it suffices to multiply by $\sqrt{N}$ to avoid the convergence in probability and obtain convergence in distribution. The next section discusses more thoroughly the conditions under which this may happen in the context of sample means.

5.2.2 Limit theory for sample means

We can write most statistics of interest as sample means. For instance, the sample variance is the sample mean of the squared deviations with respect to the mean value in the sample. This means that we can learn a lot about the asymptotic behavior of most statistics if we understand well what happens with a generic sample mean in large samples. In what follows, we will discuss two sorts of asymptotic results: Laws of large numbers (LLN) handle convergence in probability and almost sure convergence, whereas central limit theorems (CLT) deal with convergence in distribution for sample means.

We start with Chebyshev's weak law of large numbers, which posits that $\bar{X}_N \equiv \frac{1}{N}\sum_{i=1}^{N} X_i \xrightarrow{p} \mu$ if $\lim_{N\to\infty} \mathbb{E}(\bar{X}_N) = \mu$ and $\lim_{N\to\infty} \text{var}(\bar{X}_N) = 0$. It is easy to see why this result hold by noting that these moment conditions essentially ensure convergence in mean square, which in turn implies convergence in probability. Also, the above result does not require a random sampling given that we are not necessarily imposing that $\mathbb{E}(X_i)$ is constant for $i = 1, \ldots, N$. If we compute the sample mean of a random sample, it then suffices to assume that $\mathbb{E}|X| < \infty$ to obtain $\bar{X}_N \xrightarrow{p} \mu \equiv \mathbb{E}(X)$. We will refer to this result as Khintchine's weak law of large numbers for random samples. By imposing a slightly more stringent condition, it is also possible to show that the sample mean almost surely converge to the true mean of a random sample (also known as Kolmogorov's strong law of large numbers).

Finally, we tackle the asymptotic distribution of a sample mean by means of Lindeberg-Lévy central limit theorem, which says that $\sqrt{N}(\bar{X}_N - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$ as long as X_i is iid with $\mathbb{E}|X_i|^2 < \infty$. To sum up, a sample mean $\bar{X}_N$ of iid random variables with finite mean and variance is such that $\bar{X}_N \xrightarrow{a.s.} \mu$ (due to LLN) and $\sqrt{N}(\bar{X}_N - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$ (due to CLT).

There is a whole bunch of different laws of large numbers and central limit theorems that deal with all sorts of settings. Extensions include, among others, LLN and CLT for random variables that are dependent and/or non-identically distributed as well as for random vectors and matrices. Although there is not much gain in showing/memorizing all the different conditions under which we can derive either a LLN or a CLT, it is important to know that such results exist for a wide array of situations.

Example: Let us revisit the last example of Section 5.1. Assuming that the number of transactions per week on the Macau stock exchange is iid over time, with some finite mean and variance, ensures that the normal approximation for the distribution of the sample mean will work well in practice. However, it is sort of risky to assume that the number of transactions per week is independent over time. A simple inspection of a time series plot

for such a series would typically indicate that market activity cluster over time. Although this does not affect the application of Chebyshev's weak LLN, we must consider a CLT that allows for dependence over time. There are indeed several central limit theorems that relax the assumption of independence (which we will not review here given that they impose primitive conditions that restrict the dependence between observations in quite complicated manners).

Chapter 6

Point and interval estimation

An estimator is a statistic that we employ to estimate an unknown population parameter such as, for example, a population mean. It is also a random variable in that its value depends on the particular realization of the sample. A parameter estimate then is the value that we observe for an estimator given the sample. The next example illustrates the fact that we can always think of many different estimators for a given quantity.

Example: Let X denote a Gaussian random variable with mean $1/3$ and unknown variance, that is, $X \sim \mathcal{N}(1/3, \sigma^2)$. Suppose that we draw a random sample of 5 observations with values: $x_1 = \frac{1}{3}$, $x_2 = \frac{1}{4}$, $x_3 = \frac{1}{2}$, $x_4 = \frac{1}{3}$, and $x_5 = \frac{2}{9}$. We could estimate the population mean by any of the following estimators:

1. $\widehat{\mu}_1 = X_1$, which produces an estimate of $\frac{1}{3}$;
2. $\widehat{\mu}_2 = \frac{X_1+X_4}{2}$, which entails an estimate of $\frac{1}{3}$;
3. $\widehat{\mu}_3 = X_2$, giving way to an estimate of $\frac{1}{4}$; and
4. $\widehat{\mu}_4 = \bar{X}_5 = \frac{1}{5}\sum_{i=1}^{5} X_i$, leading to an estimate of $\frac{59}{180}$.

Needless to say, the list above is not exhaustive at all and many other estimators exist for the population mean. That's exactly why we must come up with some criteria to choose between estimators. Sections 6.1.1 to 6.1.3 discuss some intuitive criteria, whereas Sections 6.1.4 and 6.1.5 describe the two most popular estimation methods in statistics. Regardless of

the method we employ, the usual estimation procedure involves three steps. First, we must draw/observe a random sample from the population of interest. Second, we must calculate a point estimate of the parameter. By point estimation, we mean assigning a unique value to the estimate as opposed to an interval of possible values as in interval estimation. Third, we must compute a measure of variability for the estimator that accounts for the sampling variation in the data. This often takes the form of computing confidence intervals.

In what follows, we first discuss point estimation and then turn attention to the more interesting problem of interval estimation. We say interval estimation is more interesting for it allows us to bridge the two main strands of statistical inference, namely, estimation and hypothesis testing.

6.1 Point estimation

We denote by $\widehat{\theta}_N$ a point estimator of θ based on a sample of N observations, though we will sometimes omit the dependence on the sample size to simplify notation. We next define in a more formal manner what we mean by point estimator.

Consider a sample $\boldsymbol{X}^{(N)} = (X_1, \ldots, X_N)$ of N random variables that we denote by X_i for $i = 1, \ldots, N$. We know that a statistic is a function of the sample $\boldsymbol{X}^{(N)}$ and we define a point estimator as a statistic $\widehat{\theta}_N \equiv \widehat{\theta}(\boldsymbol{X}^{(N)}) \equiv \widehat{\theta}(X_1, \ldots, X_N)$ that we employ to infer the value of a parameter θ of the joint distribution of $\boldsymbol{X}^{(N)} = (X_1, \ldots, X_N)$. The parameter estimate is the realization of the estimator, that is to say, $\widehat{\theta}_N \equiv \widehat{\theta}(x_1, \ldots, x_N)$. Note the abuse of notation in that we refer to both estimator and estimate as $\widehat{\theta}_N$.

There are several possible estimators for any parameter θ and hence we will discuss in what follows the sort of properties we would like our estimator to hold. In particular, we will start with the definition of the mean squared error of a given estimator so as to motivate the discussion about unbiasedness, consistency, and efficiency.

6.1.1 Mean squared error

As there are many candidate estimators for a given population parameter, it seems paramount to rank them through some measure of precision. The most popular measure is the mean squared error, which gauges the average distance of the estimator to the true parameter value by means of a quadratic distance. For a sample $\boldsymbol{X}^{(N)}$, the error of the estimator $\widehat{\theta}_N \equiv \widehat{\theta}(\boldsymbol{X}^{(N)})$ is given by $\widehat{\theta}_N - \theta$. Note that the estimation error depends not only on the estimator, but also on the particular sample we observe. Different samples will give distinct point estimates for the population parameter of interest.

Definition: The mean squared error of $\widehat{\theta}_N$ is $\text{MSE}(\widehat{\theta}_N, \theta) = \mathbb{E}(\widehat{\theta}_N - \theta)^2$.

The intuition for a measure such as the MSE is straightforward. We square the estimation error before taking averages in order to avoid negative values canceling out with positive values. It thus measure how far, on average, the set of estimates are from the population parameter of interest. The nicest thing about the mean squared error is that we can decompose it into two readily interpretable components:

$$\begin{aligned}
\text{MSE}(\widehat{\theta}_N, \theta) &= \mathbb{E}(\widehat{\theta}_N - \theta)^2 = \mathbb{E}\left[\widehat{\theta}_N - \mathbb{E}(\widehat{\theta}_N) + \mathbb{E}(\widehat{\theta}_N) - \theta\right]^2 \\
&= \mathbb{E}\left\{\left[\widehat{\theta}_N - \mathbb{E}(\widehat{\theta}_N)\right]^2 + \left[\mathbb{E}(\widehat{\theta}_N) - \theta\right]^2 - 2\left[\widehat{\theta}_N - \mathbb{E}(\widehat{\theta}_N)\right]\left[\mathbb{E}(\widehat{\theta}_N) - \theta\right]\right\} \\
&= \mathbb{E}\left[\widehat{\theta}_N - \mathbb{E}(\widehat{\theta}_N)\right]^2 + \mathbb{E}\left[\mathbb{E}(\widehat{\theta}_N) - \theta\right]^2 - 2\,\mathbb{E}\left[\widehat{\theta}_N - \mathbb{E}(\widehat{\theta}_N)\right]\left[\mathbb{E}(\widehat{\theta}_N) - \theta\right] \\
&= \text{var}(\widehat{\theta}_N) + \left[\mathbb{E}(\widehat{\theta}_N) - \theta\right]^2 - 2\,\mathbb{E}\left[\widehat{\theta}_N - \mathbb{E}(\widehat{\theta}_N)\right]\left[\mathbb{E}(\widehat{\theta}_N) - \theta\right] \\
&= \text{var}(\widehat{\theta}_N) + \left[\mathbb{E}(\widehat{\theta}_N) - \theta\right]^2.
\end{aligned}$$

This decomposition clarifies that the mean squared error is the sum of the variance of the estimator and the squared bias of the estimator. The variance indicates how far, on average, the set of estimates are from their expected value and hence it is a measure of precision of the estimator. We denote the square root of the variance of an estimator by standard error (rather than standard deviation). In contrast, we define the bias of $\widehat{\theta}_N$ as $\mathbb{E}(\widehat{\theta}_N) - \theta$, that

is to say, the difference between the average estimate and θ or, equivalently, the average estimation error. The bias thus measures how accurate is the estimator. We will later show that there is a bias-variance tradeoff, which translates into a trade-off between accuracy and precision.

We say that $\widehat{\theta}$ is an unbiased estimator of θ if and only if $\mathbb{E}(\widehat{\theta}_N) = \theta$. The bias is a property of the estimator, not of the estimate. Ideally, we would like to have the most accurate and precise estimator. The next definition formalizes this idea by setting up the MSE criterion for choosing between estimators.

Definition: Let Θ denote the parameter space that collects all possible values for the population parameter θ. Consider two estimators $\widehat{\theta}_1$ and $\widehat{\theta}_2$ for θ. We say that $\widehat{\theta}_1$ is more efficient than $\widehat{\theta}_2$ if $\text{MSE}(\widehat{\theta}_1, \theta) \leq \text{MSE}(\widehat{\theta}_2, \theta)$ for every $\theta \in \Theta$ and $\text{MSE}(\widehat{\theta}_1, \theta) < \text{MSE}(\widehat{\theta}_2, \theta)$ for at least one value of $\theta \in \Theta$.

Example: Let $\boldsymbol{X}^{(N)} = (X_1, \ldots, X_N)$ denote a random sample of iid $\mathcal{N}(\mu, \sigma^2)$, whose parameters we estimate by means of $\widehat{\mu}_N = \frac{1}{N}\sum_{i=1}^N X_i$ and $\tilde{\sigma}_N^2 = \frac{1}{N-1}\sum_{i=1}^N (X_i - \widehat{\mu}_N)^2$, respectively. Both estimators are unbiased in that $\mathbb{E}(\widehat{\mu}_N) = \frac{1}{N}\sum_{i=1}^N \mathbb{E}(X_i) = \mu$ and that

$$\begin{aligned}
\mathbb{E}(\tilde{\sigma}_N^2) &= \mathbb{E}\left[\frac{1}{N-1}\sum_{i=1}^N (X_i - \widehat{\mu}_N)^2\right] = \mathbb{E}\left[\frac{1}{N-1}\sum_{i=1}^N \left(X_i^2 - 2X_i\widehat{\mu}_N + \widehat{\mu}_N^2\right)\right] \\
&= \mathbb{E}\left[\frac{1}{N-1}\left(\sum_{i=1}^N X_i^2 - 2\,\widehat{\mu}_N \sum_{i=1}^N X_i + \sum_{i=1}^N \widehat{\mu}_N^2\right)\right] \\
&= \mathbb{E}\left[\frac{1}{N-1}\left(\sum_{i=1}^N X_i^2 - 2\,\widehat{\mu}_N\,(N\,\widehat{\mu}_N) + N\,\widehat{\mu}_N^2\right)\right] \\
&= \mathbb{E}\left[\frac{1}{N-1}\left(\sum_{i=1}^N X_i^2 - N\,\widehat{\mu}_N^2\right)\right] = \frac{1}{N-1}\left[\mathbb{E}\left(\sum_{i=1}^N X_i^2\right) - N\,\mathbb{E}(\widehat{\mu}_N^2)\right] \\
&= \frac{1}{N-1}\left[N\,\mathbb{E}(X_i^2) - N\,\mathbb{E}(\widehat{\mu}_N^2)\right] = \frac{N}{N-1}\left[(\mu^2 + \sigma^2) - \left(\mu^2 + \frac{\sigma^2}{N}\right)\right] \\
&= \frac{N}{N-1}\left(1 - \frac{1}{N}\right)\sigma^2 = \sigma^2
\end{aligned}$$

as the second uncentered moment of any random variable is the sum of the square of the first moment and the variance, i.e., $\mathbb{E}(X^2) = [\mathbb{E}(X)]^2 + \mathrm{var}(X)$. See below for the derivation of the variance of $\widehat{\mu}_N$. Unbiasedness means that the mean squared error of $\widehat{\mu}_N$ and $\tilde{\sigma}_N^2$ are just their variance. The variance of $\widehat{\mu}$ is

$$\begin{aligned}
\mathbb{E}\left(\widehat{\mu}_N - \mu\right)^2 &= \mathbb{E}\left(\frac{1}{N}\sum_{i=1}^N X_i - \mu\right)^2 = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N (X_i - \mu)\right]^2 \\
&= \mathbb{E}\left[\frac{1}{N^2}\sum_{1\le i,j\le N} (X_i - \mu)(X_j - \mu)\right] = \frac{1}{N^2}\sum_{1\le i,j\le N} \mathbb{E}[(X_i - \mu)(X_j - \mu)] \\
&= \frac{1}{N^2}\left[\sum_{i=1}^N \mathrm{var}(X_i) + 2\sum_{1\le i<j\le n} \mathrm{cov}(X_i, X_j)\right] = \frac{\sigma^2}{N},
\end{aligned}$$

given that independence ensures that $\mathrm{cov}(X_i, X_j) = 0$ for all $1 \le i \ne j \le N$. It is also possible to show that, under normality, the variance of $\tilde{\sigma}_N^2$ is given by

$$\mathbb{E}\left(\widehat{\sigma}_N^2 - \sigma^2\right)^2 = \frac{2}{N-1}\,\sigma^4.$$

We require normality for it ensures that the fourth moment depends only on σ^2. Consider now the following alternative variance estimator: $\widehat{\sigma}_N^2 = \frac{1}{N}\sum_{i=1}^N (X_i - \widehat{\mu}_N)^2$. This estimator

is obviously biased in that $\mathbb{E}(\widehat{\sigma}_N^2) = \mathbb{E}\left(\frac{N-1}{N}\,\tilde{\sigma}_N^2\right) = \frac{N-1}{N}\,\sigma^2$. As for the variance, we employ the same trick to show that

$$\begin{aligned}\operatorname{var}(\widehat{\sigma}_N^2) &= \operatorname{var}\left(\frac{N-1}{N}\,\tilde{\sigma}_N^2\right) = \left(\frac{N-1}{N}\right)^2 \operatorname{var}(\tilde{\sigma}_N^2) \\ &= \left(\frac{N-1}{N}\right)^2 \frac{2}{N-1}\,\sigma^4 = \frac{2(N-1)}{N^2}\,\sigma^4.\end{aligned}$$

This means that the mean squared error of $\widehat{\sigma}_N^2$ is

$$\begin{aligned}\operatorname{MSE}(\widehat{\sigma}_N^2, \sigma^2) &= \frac{2(N-1)}{N^2}\,\sigma^4 + \left(\frac{N-1}{N}\,\sigma^2 - \sigma^2\right)^2 = \left[\frac{2(N-1)}{N^2} + \frac{1}{N^2}\right]\sigma^4 \\ &= \left(\frac{2(N-1)+1}{N^2}\right)\sigma^4 = \frac{2N-1}{N^2}\,\sigma^4,\end{aligned}$$

which is strictly inferior to the mean squared error of $\tilde{\sigma}_N^2$. In the MSE sense, it then follows that $\widehat{\sigma}_N^2$ is superior to $\tilde{\sigma}_N^2$ as an estimator of σ^2.

6.1.2 Unbiasedness

Although it seems very reasonable to compare estimators purely on the basis of the mean squared error, it is very often the case that there exists no best estimator. The reason is that the class of all possible estimators is too large. One way to make the selection of estimators tractable is to restrict the search to a specific class of estimators. A natural choice is the collection of all unbiased estimators.

Definition: An estimator $\widehat{\theta}_N$ is a best unbiased estimator of θ if $\mathbb{E}(\widehat{\theta}_N) = \theta$ for all $\theta \in \Theta$ and $\operatorname{var}(\widehat{\theta}_N) \leq \operatorname{var}(\widetilde{\theta}_N)$ for any other unbiased estimator $\widetilde{\theta}_N$ such that $\mathbb{E}(\widetilde{\theta}_N) = \theta$.

The above definition does not help us much in the sense that it does not provides much information about the best unbiased estimator. The next result adds to this discussion by establishing a lower bound for the variance of any estimator. It is known as the Cramér-Rao inequality.

Theorem: Let $\boldsymbol{X}^{(N)} = (X_1, \ldots, X_N)$ denote a random vector with joint probability

density function $f(\boldsymbol{X}^{(N)};\theta)$. Consider an estimator $\widehat{\theta}_N$ of θ such that

$$\frac{\partial}{\partial\theta}\,\mathbb{E}(\widehat{\theta}_N) = \int_{-\infty}^{\infty} \frac{\partial}{\partial\theta}\left[\widehat{\theta}_N f(\boldsymbol{X}^{(N)};\theta)\right]\,\mathrm{d}\boldsymbol{X}^{(N)}$$

and $\mathrm{var}(\widehat{\theta}) < \infty$. It then follows that

$$\mathrm{var}(\widehat{\theta}) \geq \frac{\frac{\partial}{\partial\theta}\,\mathbb{E}(\widehat{\theta})}{\mathbb{E}\left[\frac{\partial}{\partial\theta}\,\ln f(\boldsymbol{X}^{(N)};\theta)\right]^2}.$$

It is easy to see that the numerator of the right-hand side of the above inequality is equal to one if the estimator is unbiased, whereas the denominator depends on the expected value of the first derivative of the logarithm of the joint density function. We call the latter the score function, which will play a key role in Section 6.1.5.

6.1.3 Consistency

Although it makes sense to impose unbiasedness, there are biased estimators that achieve better mean squared errors as we have seen in the example of Section 6.1.1. The most important is to nail the parameter θ down as the sample size increases. In that example, for instance, $\tilde{\sigma}_N^2$ is asymptotically unbiased in that the bias shrinks to zero as the sample size N goes to infinity. As the Nobel prize winner Clive Granger once said, "If you cannot get it right as the sample size grows to infinity, you shouldn't be in the business". The next definition formalizes this notion by introducing the concept of consistent estimators.

Definition: Let Θ denote the parameter space and $\{\widehat{\theta}_N;\, N \geq 1\}$ denote a sequence of estimators of the population parameter $\theta \in \Theta$ indexed by the sample size N. In particular, let $\widehat{\theta}_N$ denote an estimator based on the first N observations of a sample $(X_1, X_2, \ldots)$ from a given probability distribution $f(x;\theta)$. The sequence $\widehat{\theta}_N$ is (weakly) consistent on Θ if and only if, for all $\theta \in \Theta$ and for all $\varepsilon > 0$, it holds that $\lim_{N\to\infty} \Pr(|\widehat{\theta}_N - \theta| \geq \varepsilon) = 0$.

So, a consistent estimator converges in probability to the true value of the population parameter. There is no problem in using a different notion of convergence. If we are talking

about convergence in mean square then we say the estimator is consistent in mean square. Similarly, a strongly consistent estimator converges almost surely to the true value of the parameter.

Example: Suppose that $X_1, X_2, \ldots$ is a sequence of random variables drawn from a $\mathcal{N}(\mu, \sigma^2)$ distribution. To estimate μ based on the first N observations, we usually use the sample mean $\widehat{\theta}_N = (X_1 + \ldots + X_N)/N$, which is an unbiased estimator with variance of σ^2/N. Given that linear combinations of normal variates are also normal, $\widehat{\theta}_n$ is normal with mean μ and variance σ^2/N and hence $\sqrt{N}\,(\widehat{\theta}_N - \mu)/\sigma$ has a standard normal distribution. It then follows that

$$\Pr(\widehat{\theta}_N - \mu \geq \varepsilon) = \Pr\left(\sqrt{N}\,\frac{\widehat{\theta}_N - \mu}{\sigma} \geq \sqrt{N}\,\frac{\varepsilon}{\sigma}\right) = 1 - \Phi\left(\sqrt{N}\,\frac{\varepsilon}{\sigma}\right) \to 0$$

as N tends to infinity, for any fixed $\varepsilon > 0$. Similarly, $\Pr(\widehat{\theta}_N - \mu \leq -\varepsilon) \to 0$. Therefore, the sequence $\widehat{\theta}$ of sample mean is consistent for the population mean μ.

In the above example, it is easy to prove consistency due to the normality assumption. In general problem, it is not easy to find the exact distribution of $\widehat{\theta}_N$ and hence establishing consistency in a direct manner becomes intractable. That's why most proofs of consistency are based on the Chebychev inequality. Applying the latter to an estimator $\widehat{\theta}_N$ yields

$$\Pr(|\widehat{\theta}_N - \theta| \geq \varepsilon) \leq \frac{\mathbb{E}|\widehat{\theta}_N - \theta|^2}{\epsilon^2}.$$

As a consequence, it suffices to show that the mean squared error of $\widehat{\theta}_N$ converges to zero (or, equivalently, that $\widehat{\theta}_N$ is asymptotically unbiased and its variance shrinks to zero). Another useful result to prove consistency is the continuous mapping theorem, which dictates that, if $\widehat{\theta}_N$ is consistent for θ, then $g(\widehat{\theta}_N)$ is consistent for $g(\theta)$ if $g(\cdot)$ is a continuous real-valued function.

In the next sections, we discuss two methods for deriving consistent estimators of a population parameter. The first is the method of moments, which impose assumptions only on the moments of the distribution. This means that it does not require us to know the joint distribution of the data, only the moments. In contrast, the second method requires the specification of the joint distribution and, as such, it takes advantage of the whole probabilistic structure to derive efficient estimators for the population parameters.

6.1.4 Method of moments

The method of moments relies on the very simple idea of matching sample moments with their population counterparts. Let $(X_1, \ldots, X_N)$ denote a random sample with marginal probability density functions given by $f(X_i; \boldsymbol{\theta})$ with $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$. Assume that the first k moments of X_i exist, that is to say, $\mathbb{E}(X_i^k) \equiv \mu_k(\boldsymbol{\theta}) < \infty$. The moments naturally depend on the parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$, and so we can equate the first k sample moments to the corresponding k population moments to solve for $(\theta_1, \ldots, \theta_k)$:

$$\frac{1}{N}\sum_{i=1}^{N} X_i = \mu_1(\boldsymbol{\theta}),\ \frac{1}{N}\sum_{i=1}^{N} X_i^2 = \mu_2(\boldsymbol{\theta}),\ \ \ldots\ , \frac{1}{N}\sum_{i=1}^{N} X_i^k = \mu_k(\boldsymbol{\theta}).$$

The above forms a system of k equations with k incognita, and hence it suffices to solve for $(\theta_1, \ldots, \theta_k)$ as a function of the sample moments.

Examples

(1) Consider a random sample $(X_1, \ldots, X_N)$ from a normal distribution with mean μ and variance σ^2. Using the method of moments, we equate the first two sample and population moments giving way to $\widehat{\mu}_N = \frac{1}{N}\sum_{i=1}^N X_i$ and $\frac{1}{N}\sum_{i=1}^N X_i^2 = \widehat{\mu}_N^2 + \widehat{\sigma}_N^2$. The latter leads to $\widehat{\sigma}_N^2 = \frac{1}{N}\sum_{i=1}^N X_i^2 - \widehat{\mu}_N^2 = \frac{1}{N}\sum_{i=1}^N (X_i - \widehat{\mu}_N)^2$.

(2) Suppose that $(X_1, \ldots, X_N)$ is a random sample drawn from an exponential distribution with parameter λ. The methods of moments would suggest employing the sample mean as an estimator of λ.

(3) Suppose that $(X_1, \ldots, X_N)$ is a random sample from a negative binomial distribution with parameters k and p. The methods of moments lands us with $\frac{1}{N}\sum_{i=1}^N X_i = k/p$ and $\frac{1}{N}\sum_{i=1}^N X_i^2 = k(1-p)/p^2$. Solving for k and p then yields

$$\hat{k}_N = \frac{\left(\frac{1}{N}\sum_{i=1}^N X_i\right)^2}{\frac{1}{N}\sum_{i=1}^N X_i(1+X_i)} \qquad \text{and} \qquad \hat{p}_N = \frac{\frac{1}{N}\sum_{i=1}^N X_i}{\frac{1}{N}\sum_{i=1}^N X_i(1+X_i)}.$$

6.1.5 Maximum likelihood

The method of moments uses only a subset of the joint distribution moments and hence it cannot be as efficient as an estimator that exploits the whole information given by the joint distribution. The price to pay for the latter is that it is easier to misspecify the joint distribution than a couple of moments. For instance, economic theory speaks only about expectations (and hence moments), without much to say about distributions (unless as a simplifying assumption to make the model tractable).

Consider a ransom sample $(X_1, \ldots, X_N)$ drawn from a probability density function given by $f(X_i; \boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$. We define the likelihood function as

$$\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{X}) = \prod_{i=1}^N f(X_i; \boldsymbol{\theta}).$$

In other words, the likelihood function is equivalent to the joint density of the data but for the argument. While the joint density is a function of the random variables $(X_1, \ldots, X_N)$ given a parameter vector $\boldsymbol{\theta}$, the likelihood inverts the problem and considers a function of the parameter vector $\boldsymbol{\theta}$ given the sample we observe $(X_1, \ldots, X_N)$.

The maximum likelihood (ML) estimator $\widehat{\boldsymbol{\theta}}_N$ then searches for the parameter value that maximize the probability of observing the sample $(X_1, \ldots, X_N)$, resulting in

$$\widehat{\boldsymbol{\theta}}_N = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmax}}\ \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{X}) = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmax}} \prod_{i=1}^{N} f(X_i; \boldsymbol{\theta}).$$

In view that monotone transformation do not alter the maximization problem, we prefer to take the logarithm of the likelihood function so as to end up with a sum of log-densities for it is always easier to manipulate sums rather than products. This yields

$$\widehat{\boldsymbol{\theta}}_N = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmax}}\ \ln \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{X}) = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmax}} \sum_{i=1}^{N} \ln f(X_i; \boldsymbol{\theta}).$$

So, to find the ML estimator, we must equate the score vector $\frac{\partial}{\partial \boldsymbol{\theta}'} \ln f(\boldsymbol{X}; \boldsymbol{\theta})$ to zero.

The maximum likelihood method entails a number of interesting properties for the estimator. First, it is invariant to parameter transformations in that, if $\widehat{\boldsymbol{\theta}}$ is the ML estimator of $\boldsymbol{\theta}$, then the ML estimator of $\boldsymbol{g}(\boldsymbol{\theta})$ is $\boldsymbol{g}(\widehat{\boldsymbol{\theta}})$ for any function $\boldsymbol{g}$ of $\boldsymbol{\theta}$. Second, the ML estimator is very easy to deal with in large samples for it is asymptotically normal. Third, under certain weak regularity conditions, the ML estimator is asymptotically unbiased and efficient in that it achieves the lower bound given by the Cramér-Rao inequality as the sample size goes to infinity.

Examples

(1) Consider a random sample $(X_1, \ldots, X_N)$ from a binomial model with probability p. The likelihood is $\mathcal{L}(p; \boldsymbol{X}) = \sum_{i=1}^{N} \binom{n}{x_i} p^{x_i}(1-p)^{N-x_i}$ and hence the score function is

$$\frac{\partial}{\partial p} \ln \mathcal{L}(p; \boldsymbol{X}) = \frac{\sum_{i=1}^{N} X_i}{p} - \frac{N - \sum_{i=1}^{N} X_i}{1-p}.$$

Equating the score function to zero yields $\hat{p}_N = \frac{1}{N}\sum_{i=1}^{N} X_i$.

(2) Suppose that $(X_1, \ldots, X_N)$ is a random sample drawn from a Poisson distribution with

arrival rate λ. The first-order condition for the maximization of the log-likelihood function then is

$$\frac{\partial}{\partial\lambda}\ln\mathcal{L}(\lambda;\boldsymbol{X})\bigg|_{\lambda=\dot{\lambda}_N} = \frac{\partial}{\partial\lambda}\sum_{i=1}^{N}\ln\left(\frac{e^{-\lambda}\lambda^{X_i}}{X_i!}\right)\bigg|_{\lambda=\dot{\lambda}_N} = \frac{\partial}{\partial\lambda}\sum_{i=1}^{N}[-\lambda+X_i\ln\lambda-\ln(X_i!)]\bigg|_{\lambda=\dot{\lambda}_N}$$
$$= \frac{\partial}{\partial\lambda}\left[-N\lambda+\ln\lambda\sum_{i=1}^{N}X_i-\sum_{i=1}^{N}\ln(X_i!)\right]\bigg|_{\lambda=\dot{\lambda}_N} = \frac{\sum_{i=1}^{N}X_i}{\widehat{\lambda}_N}-N=0,$$

giving way to $\widehat{\lambda}_N=\frac{1}{N}\sum_{i=1}^{N}X_i$ as the maximum likelihood estimator.

(3) Suppose that $(X_1,\ldots,X_N)$ is a random sample from a normal distribution with mean μ and variance σ^2. The first-order conditions are $\frac{\partial}{\partial\mu}\ln\mathcal{L}(\mu,\sigma^2;\boldsymbol{X})=\frac{1}{\sigma^2}\sum_{i=1}^{N}(X_i-\mu)$ and $\frac{\partial}{\partial\sigma^2}\ln\mathcal{L}(\mu,\sigma^2;\boldsymbol{X})=-\frac{N}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{i=1}^{N}(X_i-\mu)^2$ yielding the following maximum likelihood estimators for the mean and variance: $\widehat{\mu}_N=\frac{1}{N}\sum_{i=1}^{N}X_i$ and $\widehat{\sigma}^2_N=\frac{1}{N}\sum_{i=1}^{N}(X_i-\widehat{\mu}_N)^2$, respectively.

To show that the maximum likelihood estimator is consistent and efficient, we must impose some regularity conditions. First, maximization of the likelihood function is over a compact parameter space $\Theta \subset \mathbb{R}^k$ and the true parameter vector $\boldsymbol{\theta}$ is in the interior of the parameter space. Second, the average log-likelihood function $s_N(\boldsymbol{\theta}) \equiv \frac{1}{N}\sum_{i=1}^{N} \ln \mathcal{L}(\boldsymbol{\theta};\, X_i)$ converges almost surely to its expected value for all possible parameter values, that is to say, $s_N(\boldsymbol{\theta}) \xrightarrow{a.s.} \mathbb{E}_{\boldsymbol{\theta}_0}[s_N(\boldsymbol{\theta})] \equiv s_\infty(\boldsymbol{\theta}, \boldsymbol{\theta_0})$ for every $\boldsymbol{\theta} \in \Theta$, where the expectation is taken over the joint distribution function evaluated at the true parameter value $\boldsymbol{\theta}_0 \in \Theta$. Third, $s_N(\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta} \in \Theta$, and hence the sample holds for $\mathbb{E}_{\boldsymbol{\theta}_0}[s_N(\boldsymbol{\theta})]$. Fourth, the latter has a unique maximum in $\boldsymbol{\theta} \in \Theta$.

We are now ready to show that $\widehat{\boldsymbol{\theta}}_N$ converges almost surely to the true value $\boldsymbol{\theta}_0$ of the population parameter vector. We start by noting that $\widehat{\boldsymbol{\theta}}_N$ for sure exists given that a continuous function always has a maximum in a compact set. Second, for any $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, it follows that $\mathbb{E}_{\boldsymbol{\theta}_0}\left[\ln \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{X}) - \ln \mathcal{L}(\boldsymbol{\theta}_0; \boldsymbol{X})\right] \leq \ln \mathbb{E}_{\boldsymbol{\theta}_0}\left[\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{X})/\mathcal{L}(\boldsymbol{\theta}_0; \boldsymbol{X})\right]$ due to the Jensen's inequality as the logarithmic function is concave. However,

$$\mathbb{E}_{\boldsymbol{\theta}_0}\left[\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{X})/\mathcal{L}(\boldsymbol{\theta}_0; \boldsymbol{X})\right] = \int_{-\infty}^{\infty} \frac{\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{x})}{\mathcal{L}(\boldsymbol{\theta}_0; \boldsymbol{x})}\, \mathcal{L}(\boldsymbol{\theta}_0; \boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = 1$$

and hence $\mathbb{E}_{\boldsymbol{\theta}_0}\left[\ln \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{X}) - \ln \mathcal{L}(\boldsymbol{\theta}_0; \boldsymbol{X})\right] \leq 0$. We now first divide both sides by N to yield $\mathbb{E}_{\boldsymbol{\theta}_0}\left[s_N(\boldsymbol{\theta}) - s_N(\boldsymbol{\theta}_0)\right] \leq 0$ and then take limits to obtain $s_\infty(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \leq s_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0)$ almost surely by the uniform convergence assumption. Further, the identification assumption ensures that the inequality is strict if $\theta \neq \theta_0$, whereas $s_\infty(\widehat{\theta}_N, \boldsymbol{\theta}_0) \geq s_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0)$ by construction given that $\widehat{\boldsymbol{\theta}}_N$ maximizes the average log-likelihood. Altogether, this means that $\widehat{\boldsymbol{\theta}}_N \xrightarrow{a.s.} \theta_0$, proving strong consistency.

The weak consistency of the maximum likelihood estimator is much easier in that it suffices to show that the mean of the ML estimator converges to the true value of the parameter, while its variance shrinks to zero. The regularity conditions we must impose to ensure weak consistency (i.e., convergence in probability) are indeed much milder than what we assume to achieve strong consistency (i.e., almost sure convergence). In what follows, we

derive the asymptotic mean and variance of the ML estimator and then derive its asymptotic normality under the assumption that $s_N(\boldsymbol{\theta})$ is twice continuously differentiable.

We first take the derivative of $s_N(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ to find the score function, which we then equate to zero to find maximum of the log-likelihood function, namely,

$$\frac{\partial}{\partial \boldsymbol{\theta}'} s_N(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^{N} \frac{\partial}{\partial \boldsymbol{\theta}'} \ln \mathcal{L}(\boldsymbol{\theta}; X_i) = 0.$$

Although we denote the score function by $\frac{\partial}{\partial \boldsymbol{\theta}'} s_N(\boldsymbol{\theta})$, note that it depends on $\boldsymbol{X}$ and hence it is also a random vector with the same dimension as the vector $\boldsymbol{\theta}$ of parameters. The first step of the proof is to show that the score function is on average zero for any $\boldsymbol{\theta} \in \Theta$. This is indeed the case as long as $\frac{\partial}{\partial \boldsymbol{\theta}'} \mathcal{L}(\boldsymbol{\theta}; X_i)$ is bounded, and so we can switch the order of differentiation and integration

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}} \left[\frac{\partial}{\partial \boldsymbol{\theta}'} \ln \mathcal{L}(\boldsymbol{\theta}; X_i) \right] &= \int_{-\infty}^{\infty} \frac{\partial}{\partial \boldsymbol{\theta}'} \ln \mathcal{L}(\boldsymbol{\theta}; x_i)\, \mathcal{L}(\boldsymbol{\theta}; x_i)\, \mathrm{d}x_i = \int_{-\infty}^{\infty} \frac{\frac{\partial}{\partial \boldsymbol{\theta}'} \mathcal{L}(\boldsymbol{\theta}; x_i)}{\mathcal{L}(\boldsymbol{\theta}; x_i)}\, \mathcal{L}(\boldsymbol{\theta}; x_i)\, \mathrm{d}x_i \\ &= \int_{-\infty}^{\infty} \frac{\partial}{\partial \boldsymbol{\theta}'} \mathcal{L}(\boldsymbol{\theta}; x_i)\, \mathrm{d}x_i = \frac{\partial}{\partial \boldsymbol{\theta}'} \int_{-\infty}^{\infty} \mathcal{L}(\boldsymbol{\theta}; x_i)\, \mathrm{d}x_i = \frac{\partial}{\partial \boldsymbol{\theta}'} 1 = 0.\end{aligned}$$

We next apply a Taylor expansion to the score function evaluated at the ML estimator, which is equal to zero given the first-order condition:

$$0 = \frac{\partial}{\partial \boldsymbol{\theta}'} \mathcal{L}(\widehat{\boldsymbol{\theta}}_N; \boldsymbol{X}) = \frac{\partial}{\partial \boldsymbol{\theta}'} \mathcal{L}(\boldsymbol{\theta}_0; \boldsymbol{X}) + \frac{\partial}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \mathcal{L}(\boldsymbol{\theta}_*; \boldsymbol{X})\, (\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0),$$

where $\boldsymbol{\theta}_* \in [\boldsymbol{\theta}_0, \widehat{\boldsymbol{\theta}}_N]$. It is straightforward to show that $\mathcal{H}_{\boldsymbol{\theta}_*} \equiv \frac{\partial}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \mathcal{L}(\boldsymbol{\theta}_*; \boldsymbol{X})$ is invertible and thus $\sqrt{N}\, (\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0) = -\mathcal{H}_{\boldsymbol{\theta}_*}^{-1} \sqrt{N}\, \frac{\partial}{\partial \boldsymbol{\theta}'} \mathcal{L}(\boldsymbol{\theta}_0; \boldsymbol{X})$. Note that $\mathcal{H}_{\boldsymbol{\theta}_*} = \frac{1}{N} \sum_{i=1}^{N} \frac{\partial}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \mathcal{L}(\boldsymbol{\theta}_*; X_i)$ and hence it does satisfy a strong law of large numbers provided that the variance of $\frac{\partial}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \mathcal{L}(\boldsymbol{\theta}_*; X_i)$ is finite. In addition, we know that the ML estimator is strongly consistent in that $\widehat{\boldsymbol{\theta}}$ converges almost surely to $\boldsymbol{\theta}_0$, implying that $\boldsymbol{\theta}_*$ converges as well to $\boldsymbol{\theta}_0$. Altogether, this means that the random matrix $\mathcal{H}_{\boldsymbol{\theta}_*}$ converges to $\mathbb{E}_{\boldsymbol{\theta}_0} \left[\frac{1}{N} \sum_{i=1}^{N} \frac{\partial}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \mathcal{L}(\boldsymbol{\theta}_0; X_i) \right] < \infty$ almost surely.

It now remains to study the asymptotic behavior of $\sqrt{N}\, \frac{\partial}{\partial \boldsymbol{\theta}'} \mathcal{L}(\boldsymbol{\theta}_0; \boldsymbol{X})$. In view that we have already seen that the latter has mean zero for any $\boldsymbol{\theta} \in \Theta$, it is reasonable to assume there

is a central limit theorem that applies. Letting $\mathcal{I}_\infty(\boldsymbol{\theta}_0) \equiv \lim_{N\to\infty} \text{var}\left(\sqrt{N}\,\frac{\partial}{\partial \boldsymbol{\theta}'}\,\mathcal{L}(\boldsymbol{\theta}_0; \boldsymbol{X})\right)$ then yields $\mathcal{I}_\infty^{-1}(\boldsymbol{\theta}_0)\,\sqrt{N}\,\frac{\partial}{\partial \boldsymbol{\theta}'}\,\mathcal{L}(\boldsymbol{\theta}_0; \boldsymbol{X}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{I}_k)$, where $\boldsymbol{I}_k$ is a k-dimensional identity matrix and $\mathcal{I}_\infty(\boldsymbol{\theta}_0)$ is known as the information matrix.[1] Combining all of the above ingredients gives way to

$$\sqrt{N}(\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathcal{H}_\infty^{-1}(\boldsymbol{\theta}_0)\mathcal{I}_\infty(\boldsymbol{\theta}_0)\mathcal{H}_\infty^{-1}(\boldsymbol{\theta}_0)\right).$$

To demonstrate that the variance of the ML estimator achieves the Cramér-Rao lower bound and hence it is efficient, it suffices to show that $\mathcal{I}_\infty(\boldsymbol{\theta}_0) = -\mathcal{H}_\infty^{-1}(\boldsymbol{\theta}_0)$. To do so, we first differentiate both sides of $\int_{-\infty}^{\infty} \frac{\partial}{\partial \boldsymbol{\theta}'}\,\ln \mathcal{L}(\boldsymbol{\theta}; x_i)\,\mathcal{L}(\boldsymbol{\theta}; x_i)\,\mathrm{d}x_i = 0$ with respect to $\boldsymbol{\theta}$, yielding

$$\begin{aligned}
&\int_{-\infty}^{\infty} \frac{\partial}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\,\ln \mathcal{L}(\boldsymbol{\theta}; x_i)\,\mathcal{L}(\boldsymbol{\theta}; x_i)\,\mathrm{d}x_i + \int_{-\infty}^{\infty} \frac{\partial}{\partial \boldsymbol{\theta}'}\,\ln \mathcal{L}(\boldsymbol{\theta}; x_i)\,\frac{\partial}{\partial \boldsymbol{\theta}}\,\mathcal{L}(\boldsymbol{\theta}; x_i)\,\mathrm{d}x_i = 0 \\
\Rightarrow\ &\mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\partial}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\,\ln \mathcal{L}(\boldsymbol{\theta}; x_i)\right] + \mathbb{E}_{\boldsymbol{\theta}}\left\{\left[\frac{\partial}{\partial \boldsymbol{\theta}'}\,\ln \mathcal{L}(\boldsymbol{\theta}; x_i)\right]\left[\frac{\partial}{\partial \boldsymbol{\theta}}\,\ln \mathcal{L}(\boldsymbol{\theta}; x_i)\right]\right\} = 0 \\
\Rightarrow\ &\mathbb{E}_{\boldsymbol{\theta}}\left[\frac{1}{N}\sum_{i=1}^{N} \frac{\partial}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\,\ln \mathcal{L}(\boldsymbol{\theta}; x_i)\right] = -\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\boldsymbol{\theta}}\left\{\left[\frac{\partial}{\partial \boldsymbol{\theta}'}\,\ln \mathcal{L}(\boldsymbol{\theta}; x_i)\right]\left[\frac{\partial}{\partial \boldsymbol{\theta}}\,\ln \mathcal{L}(\boldsymbol{\theta}; x_i)\right]\right\} \\
\Rightarrow\ &\mathbb{E}_{\boldsymbol{\theta}}\left[\mathcal{H}(\boldsymbol{\theta})\right] = -\mathbb{E}_{\boldsymbol{\theta}}\left\{N\left[\frac{\partial}{\partial \boldsymbol{\theta}'}\,\ln \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{X})\right]\left[\frac{\partial}{\partial \boldsymbol{\theta}}\,\ln \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{X})\right]\right\}
\end{aligned}$$

given that the scores $\frac{\partial}{\partial \boldsymbol{\theta}}\,\ln \mathcal{L}(\boldsymbol{\theta}; x_i)$ and $\frac{\partial}{\partial \boldsymbol{\theta}}\,\ln \mathcal{L}(\boldsymbol{\theta}; x_j)$ are uncorrelated for $1 \leq i \neq j \leq N$. Finally, taking limits yield the result.

6.2 Interval estimation

It is obviously useful to know the expected value of a random variable, but ideally we must also have some idea of variability. For instance, it is always good news to hear that one of our investments is giving on average 10% of return per month. However, it is even more comforting to know that it has an expected return of $10\% \pm 1\%$. The idea is to derive a

[1] Although we have spoken about joint distributions, we have not explicitly introduced any multivariate distribution. A multivariate normal distribution depends on a vector of means $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$. The elements of $\boldsymbol{\mu}$ correspond to the individuals means, whereas the elements in the main diagonal of $\boldsymbol{\Sigma}$ refer to the individual variances. The off-diagonal elements of $\boldsymbol{\Sigma}$ denote the covariance between the components of the random vector. So, if a random vector $\boldsymbol{X} = (X_1, \ldots, X_k)$ is multivariate normal with a covariance matrix given by the k-dimensional identity matrix $\boldsymbol{I}_k$, then X_i and X_j are independent for any $i \neq j$. In general, if $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then a vector of linear combinations of the elements of $\boldsymbol{X}$ is also multivariate normal, namely, $\boldsymbol{AX} \sim \mathcal{N}(\boldsymbol{A\mu}, \boldsymbol{A\Sigma}A')$.

measure of precision from the sampling distribution of the sample mean (or any other point estimator). In this way, we may provide a range of values within which we expect the mean of the distribution to belong rather than give only a point estimate.

Examples

(1) Let $X \sim \mathcal{N}(\mu, \sigma^2)$ with σ^2 known. We estimate μ by means of the sample mean $\bar{X}_N$, whose sampling distribution is $\bar{X}_N \sim \mathcal{N}(\mu, \sigma^2/N)$. To establish a confidence interval around the sample mean in which the true value of μ belongs with 95% of probability, we first note that $\sqrt{N}(\bar{X}_N - \mu)/\sigma$ is standard normal and hence $\Pr\left(\sqrt{N}\left|\bar{X}_N - \mu\right|/\sigma \leq 1.96\right) = 95\%$. It then follows that

$$\Pr\left(\sqrt{N}\left|\frac{\bar{X}_N - \mu}{\sigma}\right| \leq 1.96\right) = \Pr\left(\bar{X}_N - 1.96\,\frac{\sigma}{\sqrt{N}} \leq \mu \leq \bar{X}_N + 1,96\frac{\sigma}{\sqrt{N}}\right) = 0,95$$

This result holds exactly only because of normality and because we know the value of the variance. In large samples, the above confidence interval nevertheless provides a good approximation because the central limit theorem and the weak law of large numbers ensure that the sampling distribution converges to a normal distribution, whereas the sample variance converges in probability to the true variance as $N \to \infty$, respectively.

(2) Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ denote a random sample of Bernoulli essays that take value one with probability p, zero otherwise. To estimate the probability p, a natural estimator is the relative frequency $\widehat{p}_N = \frac{1}{N}\sum_{i=1}^{N} X_i$, which is unbiased given that

$$\mathbb{E}(\widehat{p}_N) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}(X_i) = \frac{1}{N}\,(Np) = p.$$

This means that the mean squared error of the relative frequency corresponds to its variance, which is given by

$$\mathbb{E}\,(\widehat{p}_N - p)^2 = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}(X_i - p)\right]^2 = \frac{1}{N}\,\mathbb{E}(X-p)^2 = \frac{1}{N}\,\mathrm{var}(X) = \frac{1}{N}\,p\,(1-p).$$

Note that the variance shrinks to zero as $N \to \infty$, confirming that the relative frequency is a consistent estimator in that it converges in probability to the probability p. The central

limit theorem says that the binomial converges to a normal distribution as the sample size grows. Given that $\sum_{i=1}^{N} X_i$ has by definition a binomial distribution with expected mean Np and variance $Np(1-p)$, it follows that $\widehat{p}_N$ weakly converges to a normal distribution with mean p and variance $p\,(1-p)/N$. Accordingly, the probability p belongs to the interval

$$\left[\widehat{p}_N - 1.96\sqrt{\widehat{p}_N(1-\widehat{p}_N)/N}, \widehat{p}_N + 1.96\sqrt{\widehat{p}_N(1-\widehat{p}_N)/N}\right]$$

with 95% of confidence. As before, we are using not only the central limit theorem to justify asymptotic normality, but also the law of large numbers to ensure that we estimate the variance consistently by plugging in $\widehat{p}_N$ in lieu of p.

It is interesting to review what happens under normality just to fix some ideas. So, let $X_i \sim$ iid $\mathcal{N}(\mu, \sigma^2)$ for $i = 1, \ldots, N$. We know that the sample mean $\bar{X}_N$ has expected value μ and variance σ^2/N. In addition, as it is a linear combination of normal variates, $\bar{X}_N \sim \mathcal{N}(\mu, \sigma^2/N)$ or, equivalently, $\sqrt{N}\,(\bar{X} - \mu)/\sigma \sim \mathcal{N}(0, 1)$. The snag is that, in general, the standard deviation σ is unknown and hence we can at best replace it with a consistent estimator. For instance, the maximum likelihood estimator of the variance of a normal distribution is

$$\begin{aligned}\widehat{\sigma}_N^2 &= \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X}_N)^2 = \frac{1}{N}\sum_{i=1}^{N}\left(X_i^2 - 2X_i\bar{X}_N + \bar{X}_N^2\right) \\ &= \frac{1}{N}\sum_{i=1}^{N}X_i^2 - 2\,\bar{X}_N\,\frac{1}{N}\sum_{i=1}^{N}X_i + \bar{X}_N^2 = \frac{1}{N}\sum_{i=1}^{N}X_i^2 - 2\bar{X}_N^2 + \bar{X}_N^2 \\ &= \frac{1}{N}\sum_{i=1}^{N}X_i^2 - \bar{X}_N^2,\end{aligned}$$

with an expected value of

$$\begin{aligned}\mathbb{E}(\widehat{\sigma}_N^2) &= \mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N}X_i^2 - \bar{X}_N^2\right) = \mathbb{E}(X^2) - \mathbb{E}(\bar{X}_N^2) \\ &= \mu^2 + \sigma^2 - \mu^2 - \sigma^2/N = \sigma^2(1 - 1/N) = \frac{N-1}{N}\,\sigma^2 \overset{N\uparrow\infty}{\longrightarrow} \sigma^2.\end{aligned}$$

Although it is asymptotically unbiased, the ML estimator is biased in small samples. As we know, to find an unbiased estimator for the variance of a normal distribution, it suffices to multiply the ML estimator by a factor of $N/(N-1)$, giving way to

$$\tilde{\sigma}_N^2 = \frac{N}{N-1}\,\widehat{\sigma}_N^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X}_N)^2.$$

Intuitively, we subtract one from the denominator because we have already lost one degree of freedom due to the estimation of the mean.

The nice thing about assuming normality is that it allows us to say something about the distribution of $\tilde{\sigma}_N^2$ and hence about the exact distribution of the sample mean. To appreciate

that, we first note that

$$\begin{aligned}
\tilde{\sigma}_N^2 &= \frac{1}{N-1}\sum_{i=1}^{N}(X_i-\bar{X}_N)^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left[(X_i-\mu)-(\bar{X}_N-\mu)\right]^2 \\
&= \frac{1}{N-1}\sum_{i=1}^{N}\left[(X_i-\mu)^2-2(X_i-\mu)(\bar{X}_N-\mu)+(\bar{X}_N-\mu)^2\right] \\
&= \frac{1}{N-1}\left[\sum_{i=1}^{N}(X_i-\mu)^2-2\sum_{i=1}^{N}(X_i-\mu)(\bar{X}_N-\mu)+\sum_{i=1}^{N}(\bar{X}_N-\mu)^2\right] \\
&= \frac{1}{N-1}\left[\sum_{i=1}^{N}(X_i-\mu)^2-2N(\bar{X}_N-\mu)^2+N(\bar{X}_N-\mu)^2\right] \\
&= \frac{1}{N-1}\left[\sum_{i=1}^{N}(X_i-\mu)^2-N(\bar{X}_N-\mu)^2\right] \\
&= \frac{1}{N-1}\sum_{i=1}^{N}(X_i-\mu)^2-\frac{N}{N-1}(\bar{X}_N-\mu)^2.
\end{aligned}$$

Dividing both sides by the variance then yields

$$\begin{aligned}
(N-1)\,\frac{\tilde{\sigma}_N^2}{\sigma^2} &= \sum_{i=1}^{N}\left(\frac{X_i-\mu}{\sigma}\right)^2-N\left(\frac{\bar{X}_N-\mu}{\sigma}\right)^2 \\
&= \sum_{i=1}^{N}\left(\frac{X_i-\mu}{\sigma}\right)^2-\left(\frac{\bar{X}_N-\mu}{\sigma/\sqrt{N}}\right)^2.
\end{aligned}$$

Both terms within brackets refer to squared standard normal random variables and hence have chi-square distributions. Moreover, it is possible to show that $(N-1)\,\tilde{\sigma}_N^2/\sigma^2$ is a chi-square random variable with $N-1$ degrees of freedom. Bearing that in mind, we now turn attention to

$$\sqrt{N}\,\frac{\bar{X}_N-\mu}{\tilde{\sigma}_N} = \frac{\sqrt{N}(\bar{X}_N-\mu)/\sigma}{\tilde{\sigma}_N/\sigma},$$

whose distribution is t-student with $N-1$ degrees of freedom given that the numerator is standard normal and the denominator is a chi-square divided by its degrees of freedom.[2] Altogether, this means that, under normality, we can construct an exact confidence interval for the mean value of the distribution. In particular, it follows that

$$\Pr\left(\left|\sqrt{N}\,\frac{\bar{X}_N-\mu}{\tilde{\sigma}_N}\right| \leq t_{N-1}(1-\alpha/2)\right) = 1-\alpha,$$

[2] It is also possible to show that, under normality, the random variables in the numerator and denominator are independent.

where $t_{N-1}(1-\alpha/2)$ is the $(1-\alpha/2)$ percentile of a t-student with $N-1$ degrees of freedom. For instance, to construct a 95% confidence interval, we set α to 5% and hence we must look at the 97.5% percentile of the t-student. As the latter distribution is symmetric, the 97.5% percentile is equal to the absolute value of the 2.5% percentile, so that μ will belong to the interval $\left[\bar{X}_N + t_{N-1}(\alpha/2)\tilde{\sigma}_N/\sqrt{N}, \bar{X}_N + t_{N-1}(1-\alpha/2)\tilde{\sigma}_N/\sqrt{N}\right]$ with a probability of 95%.

If normality does not hold, then we must in general employ the central limit theorem and the law of large numbers to justify the asymptotic approximation of the confidence interval based on the normal distribution. Let $X_i \sim$ iid $F_X(\mu, \sigma^2)$ for $i = 1, \ldots, N$. Standardizing the sample mean then yields

$$\sqrt{N}\,\frac{\bar{X}_N - \mu}{\tilde{\sigma}_N} = \sqrt{N}\,\frac{\bar{X}_N - \mu}{\sigma}\;\frac{\sigma}{\tilde{\sigma}_N} \xrightarrow{d} \mathcal{N}(0,1)$$

given that the central limit theorem dictates that the first term of the right-hand side of the equality is asymptotically standard normal and consistency ensures that the second term converges in probability to one as $N \to \infty$. Actually, the same result follows for any other consistent variance estimator. This means that $\mu \in [\bar{X}_N + z_{\alpha/2}\,\tilde{\sigma}_N/\sqrt{N}, \bar{X}_N + z_{1-\alpha/2}\,\tilde{\sigma}_N/\sqrt{N}]$ with probability $1-\alpha$, where $z_{\alpha/2}$ and $z_{1-\alpha/2}$ are the $\alpha/2$ and $(1-\alpha/2)$ percentiles of the standard normal distribution. As before, due to the symmetry of the normal distribution, it turns out that $z_\delta = -z_{1-\delta}$ for any $0 \leq \delta \leq 1$.

Chapter 7

Hypothesis testing

We know from previous chapters that, whenever we try to infer a quantity, the resulting estimate varies with the sample. The random nature of any sample statistic is such that we should always think twice before interpreting a result. For instance, we might wonder whether a sample mean of 11 confirms or not a hypothesized value of 10 for the population mean. The answer of course lies on the typical variation of the data. If the standard deviation is very small, say 0.01, then a sample mean of 11 is pretty far away from 10. We probably would conclude differently if the standard deviation were large, say 5. Building confidence intervals is just a first step to answer this type of questions. Significance (or hypothesis) testing provides a much more general tool for this sort of task. In particular, it allows us to check on how strong the statistical evidence is in favor of or against a hypothesis about the data (e.g., whether the true mean is 10 given that the sample mean is 11).

Significance testing starts with a partition of the probability space into two regions: the null hypothesis $\mathbb{H}_0$ and the alternative hypothesis $\mathbb{H}_1$. The former consists of all events for which the relation of interest holds, whereas the alternative hypothesis is simply the negation of the null. The idea is to observe the data seeking for evidence against the null hypothesis. Note that a statistical test can at best contradict the null hypothesis. However, failing to reject the null hypothesis does not necessarily mean that we should accept it, just that we do not have enough material to reject it. The testing strategy then is to develop a statistic that should reflect the relation of interest as hypothesized by the null. This means that we will

have to derive the sampling distribution of some test statistic conditioning on the fact that the null holds. To make this task easier, we normally define the null as the simplest case. For instance, it is much easier to compute the distribution for the sample mean by setting the population mean to 10 rather than considering any value different from 10. In general, the alternative hypothesis typically reflects a change/impact in the process/population, while the null hypothesis indicates the absence of a change or impact.

Examples

(1) Suppose we wish to test whether the true population mean μ exceeds 15. We can then define the null hypothesis as $\mathbb{H}_0 : \mu > 15$ and the alternative hypothesis as $\mathbb{H}_1 : \mu \leq 15$. This is a directional test given that we are attempting to evince deviations with respect to a particular direction. We could thus consider a statistical test that rejects the sample mean if it belongs to a given interval. Intuitively, to determine the latter, we should appreciate that it does not suffice to observe a sample mean below 15 to reject the null. The reason is simple. As a consistent estimator, the sample mean should provide us a value in the neighborhood of the true mean. If the latter is 15.1, for instance, then it is likely that we will observe a sample mean below 15 even though the null hypothesis hold.

(2) Suppose now the interest lies on testing whether $\mu = 15$. We then define the null hypothesis as $\mathbb{H}_0 : \mu = 15$ given that it is easier to derive the distribution of the sample mean if we know the true value of the population mean. In contrast to the previous example, testing $\mathbb{H}_0$ involves no direction in that we should observe both positive and negative deviations with respect to 15 (instead of only negative) in our attempt to reject the null hypothesis. As before, because of sampling variation, we should consider a test in which we reject the null if the sample mean is either too large or too small relative to 15. Needless to say, we should define 'too large' and 'too small' according to the distribution of the sample mean just as we have done for confidence intervals in the previous chapter.

To understand whether the value we observe for a sample statistic is plausible or not,

given the amount of randomness we specify in the null hypothesis, we must hinge our analysis on the distribution of that sample statistic under the null. For instance, most people would not reject the null hypotheses in the above example if the sample mean were 14.9999, though most would reject the null in the second example if observing a sample mean of 500. As we mention in the second example, to find the appropriate rejection region, we must follow a procedure very similar to that we have used in the previous chapter to obtain confidence intervals. Indeed, the first step in the derivation of a statistical testing procedure is to obtain a rejection region by fixing a significance level for the test. Just as a confidence interval of 95% will on average miss the true value of the parameter 5% of the times, a test with 95% significance level will incorrectly reject the null hypothesis at most 5% of the times.

The main difference between confidence intervals and tests is that there is only one type of error in the former, i.e., missing the true value of the parameter. In contrast, there are two sources of errors within hypothesis testing. We can either commit an error of type I or an error of type II. The first arises if we reject the null hypothesis even though it is true, whereas the second refers to the event of failing to reject a false null. To sum up, if we denote by R the event of rejecting the null hypothesis $\mathbb{H}_0$, then $\Pr(\text{type I error}) = \Pr(R \,|\, \mathbb{H}_0 \text{ is true})$ and $\Pr(\text{type II error}) = \Pr(\bar{R} \,|\, \mathbb{H}_0 \text{ is false})$ with $\bar{R}$ denoting the complement event of R and hence a non-rejection of the null.

Note that we treat these errors in a very asymmetric fashion in that we fix only the maximum tolerable probability of a type I error, e.g., $\Pr(\text{type I error}) \leq \alpha\%$ if the significance level is of $(1-\alpha)\%$. It turns out that it is impossible to control both errors at the same type and hence the best we can do is to find a statistical procedure that minimizes the chances of an error of type II given a fixed probability of an error of type I. Alternatively, we could think of doing the contrary, that is to say, fixing the probability of a type II error and then obtain the testing procedure that minimizes the likelihood of a type I error. Although there is no logical reason to do so (i.e., fixing the error of type I), there is a moral reason (who said moral does not play a role in science?) as advocated by two of the most eminent statisticians of all times, namely, Jerzy Neyman (1894-1981) and Egon Pearson (1895-1980). The motivation for their idea of fixing the type I error rests on a jury trial. Most people would agree with Neyman and Pearson that it is preferable to free a guilty criminal than to put an innocent in jail. We thus fix the probability of committing a type I error for it is more damaging than the type II error.

Example We may see a pregnancy test as a statistical procedure that decides whether there is enough evidence supporting pregnancy. As statistical procedure can at best reject the null hypothesis, this means that the null of a pregnancy test is actually the absence of pregnancy. Most people would agree that a false negative is potentially more damaging than

a false positive. This is in line with the Neyman-Pearson solution in that a false positive corresponds to a type I error (rejecting a true null), whereas a false negative refers to a type II error (failing to reject a false null).

In what follows, we first show how to derive the rejection region for sample means. We then introduce the concepts of size, level and power of a test, which derive from the type I and type II errors. Next, we introduce the notion of p-value, which somehow gauges the strength of the evidence against the null hypothesis. Computing p-values is an alternative to setting ex-ante the significance level of the test and hence a rejection region. Finally, in the lasts section, we discuss hypothesis testing in a more general fashion by assuming a likelihood approach.

7.1 Rejection region for sample means

We motivate this section with a simple example. A pharmaceutical lab is running clinic trials to assess whether a new medicine to control the levels of cholesterol indeed works better than the current medicine in the market. The clinical trials consider two groups of 100 patients. Group A takes the new medicine, whereas group B are subject to the standard treatment. To evaluate the relative performance of the new medicine, the lab measures the difference in the cholesterol decrease between groups A and B (in percentage points). The null hypothesis of interest is that, on average, there is no difference between the two treatments: $\mathbb{H}_0 : \mu_{A-B} = 0$.

The most natural estimator for the mean difference between the results of groups A and B is the difference between their sample means (or, equivalently, the sample mean difference). We must now ask ourselves whether a sample difference of, say, 1.06 is plausible given a standard deviation of, say, 2 under the null of zero mean difference. The idea is very similar to what we do if we wish to construct a confidence interval. The central limit theorem ensures that we can approximate the sampling distribution of the sample mean difference in

large samples by a standard Gaussian distribution if we subtract its mean and divide by its standard deviation. The null hypothesis says that the mean difference is zero, while saying nothing about the standard deviation. The standard error of the sample mean difference is the standard deviation divided by the square root of the sample size, so that it amounts to 1/5. Letting z_q denote the qth percentile of a standard normal distribution then yields

$$\Pr\left(\left|\frac{\widehat{\mu}_{A-B}-0}{1/5}\right| > z_{1-\alpha/2}\,\middle|\,\mathbb{H}_0\right) = \Pr\left(|\widehat{\mu}_{A-B}| > \frac{z_{1-\alpha/2}}{5}\,\middle|\,\mathbb{H}_0\right) \cong \alpha.$$

This suggests rejecting at the 5% significance level if $|\widehat{\mu}_{A-B}| > \frac{1}{5}\,z_{1-\alpha/2} = \frac{1.96}{5} = 0.392$. The sample mean difference of 1.06 is obviously superior to 0.392 and hence we conclude that the two treatments entail different performances.

The above test is two-sided in that it looks whether both positive and negative deviations are large. The fact that the sample mean difference is positive also indicates that the new medicine performs relatively better. To confirm that, we must test a directional null hypothesis by means of one-sided tests. So, if we change the null to $\mathbb{H}_0 : \mu_{A-B} > 0$, only large negative deviations will contradict the null hypothesis. It then follow from $\Pr\left(\widehat{\mu}_{A-B} < \frac{z_\alpha}{5} \,\middle|\, \mathbb{H}_0\right) \cong \alpha$ that the rejection region is $(-\infty, \frac{z_\alpha}{5}]$. Note that the αth percentile of the standard normal distribution is negative ($z_\alpha < 0$) for any $\alpha < 1/2$. In contrast, if we define the null hypothesis as $\mathbb{H}_0 : \mu_{A-B} < 0$, then we would have to worry only about large positive sample mean differences, yielding $[\frac{z_{1-\alpha}}{5}, \infty)$ as a rejection region given that $\Pr\left(\widehat{\mu}_{A-B} > \frac{z_{1-\alpha}}{5} \,\middle|\, \mathbb{H}_0\right) \cong \alpha$. Figure 7.1 illustrates the main standard normal percentiles for testing purposes.

Let us now consider the general two-sided case in which we wish to test the null that the population mean of a random sample $\boldsymbol{X} = (X_1, \ldots, X_N)$ is equal to μ_0. In what follows, we consider three different setups. The first and second settings respectively assume normality with known and unknown variances, whereas the third imposes no specific distribution for the data. Regardless of the data distribution, the sample mean is on average equal to μ_0 under the null hypothesis, with variance σ^2/N.

Normality, known variance: If $X_i \sim$ iid $\mathcal{N}(\mu, \sigma^2)$ with a known variance σ^2, it follows that $\frac{\bar{X}_N - \mu_0}{\sigma/\sqrt{N}}$ is standard normal and hence the rejection region at the α significance level is $(-\infty, \mu_0 + \frac{\sigma}{\sqrt{N}}\, z_{\alpha/2}] \cup [\mu_0 + \frac{\sigma}{\sqrt{N}}\, z_{1-\alpha/2}, \infty)$. Note that the above intervals are symmetric in that $z_{1-\alpha/2} = -z_{\alpha/2}$ for any $0 < \alpha < 1$.

Normality, unknown variance: If $X_i \sim$ iid $\mathcal{N}(\mu, \sigma^2)$ with a unknown variance σ^2, it follows that $\frac{\bar{X}_N - \mu_0}{\dot{\sigma}/\sqrt{N}}$ is t-student with $N-1$ degrees of freedom. The rejection region at the α significance level is $(-\infty, \mu_0 + \frac{\sigma}{\sqrt{N}}\, t_{N-1}^{(\alpha/2)}] \cup [\mu_0 + \frac{\sigma}{\sqrt{N}}\, t_{N-1}^{(\alpha/2)}, \infty)$, where $t_d^{(q)}$ denote the qth percentile of the t-student with d degrees of freedom. Note that the above intervals are symmetric given that $t_d^{(1-\alpha/2)} = -t_d^{(\alpha/2)}$ for any $0 < \alpha < 1$ regardless of the number of degrees of freedom.

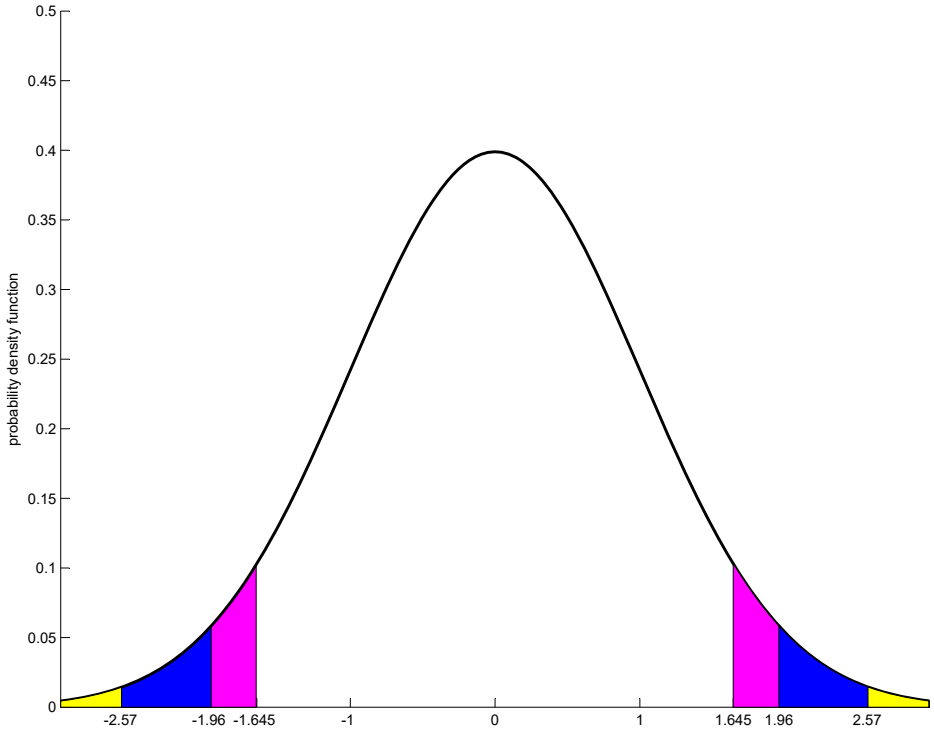

Figure 7.1: The area in yellow corresponds to 1% of the probability mass of a standard normal distribution (i.e., 0.5% in the lower tail plus 0.5% in the upper tail), whereas the area in yellow and blue responds for 5% of the probability mass (i.e., 2.5% in the lower tail plus 2.5% in the upper tail). Finally, the area in pink entails another 2.5% of the probability mass in each of the tails, so that the area in yellow, blue and pink amounts to 10% of the probability mass of a standard normal distribution.

Unknown distribution: If $X_i \sim$ iid $f_X(\mu, \sigma^2)$ with a unknown variance σ^2, it follows from the central limit theorem that $\frac{\bar{X}_N - \mu_0}{\dot{\sigma}/\sqrt{N}}$ converges in distribution to a standard normal as the sample size grows. The asymptotic rejection region at the α significance level then is $(-\infty, \mu_0 + \frac{\sigma}{\sqrt{N}}\, z_{\alpha/2}] \cup [\mu_0 + \frac{\sigma}{\sqrt{N}}\, z_{1-\alpha/2}, \infty)$.

Adapting the above rejection intervals to one-sided tests is pretty straightforward. It suffices to take the interval of interest and replace $\alpha/2$ by α in the percentile of the sampling distribution. For instance, under normality and unknown variance, the rejection region for the one-sided test for $\mathbb{H}_0 : \mu > \mu_0$ against $\mathbb{H}_1 : \mu \leq \mu_0$ is $[\mu_0 + \frac{\sigma}{\sqrt{N}}\, t_{N-1}^{(1-\alpha)}, \infty)$, whereas we would reject $\mathbb{H}_0 : \mu < \mu_0$ if the test statistic falls within $(-\infty, \mu_0 + \frac{\sigma}{\sqrt{N}}\, t_{N-1}^{(\alpha)}]$. The critical

value that defines the rejection region for the one-sided test depends on μ_0 just as the two-sided test, despite the fact we define the null hypothesis through an inequality rather than equality. By conditioning the test statistic on μ_0, we are essentially taking the *least favorable situation* within the alternative hypothesis, that is to say, the value of $\mu \in \mathbb{H}_1$ that is closest to the null hypothesis. It turns out that considering the least favorable situation within the alternative hypothesis alleviates the probability of committing a type II error.

Example: A barista prepares a sequence of 16 espressos, taking note of how much time it takes to pour the best possible expresso. As the sample mean time amounts to 26 seconds, the barista concludes that the expresso machine is too fast and decides to fine-tune it in order to increase the preparation time by 2 seconds. The quality-control manager disagrees with the barista for the following reasons. First, the sample size is too small. Second, the sample is not entirely random given that the barista prepares the espressos in a straight sequence. Third, the barista is not accounting for the randomness of the data. The sample standard deviation is indeed quite palpable, at 6 seconds. Fourth, it could be more interesting to fine-tune the expresso machine so as to reduce the variability of the preparation time rather than to increase the mean time. To substantiate her argument, the quality-control manager tests whether the mean time is equal to 28 within a random sample context. The difference between the hypothesized and sample means is equal to 2 seconds. The following large-sample approximation then holds

$$\begin{aligned}\Pr\left(\sqrt{16}\left|\frac{\bar{X}_{16}-28}{6}\right|>z_{1-\alpha/2}\,\middle|\,\mathbb{H}_0:\mu=28\right) &\cong 1-\Phi(z_{1-\alpha/2})+\Phi(-z_{1-\alpha/2})\\ &= 1-\Phi(z_{1-\alpha/2})+\Phi(z_{\alpha/2})\\ &= 1-(1-\alpha/2)+\alpha/2=\alpha,\end{aligned}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of a standard normal. Setting α to 5% then yields a test that rejects the null $\mathbb{H}_0:\mu=28$ if the sample mean does not belong to the interval $\left[28\pm\frac{3\times1.96}{2}\right]$. This is not the case here as the sample mean is 26 seconds and so there is no statistical reason to believe that the expresso machine requires adjustment.

7.2 Size, level, and power of a test

In this section, we extend the discussion to a more general setting in which we are interested in a parameter θ of the distribution (not necessarily the mean). As before, the derivation of a testing procedure involves two major steps. The first is to obtain a test statistic that is able to distinguish the null from the alternative hypothesis. For instance, if we are interested in the arrival rate of a Poisson distribution, it is then natural to focus either on the sample mean or on the sample variance.[1] The second is to derive the rejection region for the test statistic. The rejection region depends of course on the level of significance α, which denotes the upper limit for the probability of committing a type I error. A similar concept is given by the (exact/asymptotic) size of a test, which corresponds to the (exact/limiting) probability of observing a type I error.

In general, we are only able to compute the size of a test if both null and alternative hypotheses are simple, that is to say, they involve only one value for the parameter vector $\mathbb{H}_0 : \theta = \theta_0$ against $\mathbb{H}_1 : \theta = \theta_1$. Unfortunately, most situations refer to at least one composite hypothesis, e.g., $\mathbb{H}_0 : \theta = \theta_0$ against $\mathbb{H}_1 : \theta < \theta_0$ or $\mathbb{H}_0 : \theta = \theta_0$ against $\mathbb{H}_1 : \theta > \theta_0$ or $\mathbb{H}_0 : \theta = \theta_0$ against $\mathbb{H}_1 : \theta \neq \theta_1$ or $\mathbb{H}_0 : \theta \geq \theta_0$ against $\mathbb{H}_1 : \theta < \theta_0$ or $\mathbb{H}_0 : \theta \leq \theta_0$ against $\mathbb{H}_1 : \theta > \theta_0$. Note that it does not make much sense to think about a situation in which the null hypothesis is composite and the alternative is simple. It is always easier to derive the distribution of the test statistic for a given value of the parameter (rather than an interval), and so it would payoff to invert the hypotheses.

Well, both level and size relate to the type I error. To make it fair, we will now define a concept that derives from the probability of committing a type II error. The power of a test is the probability of correctly rejecting the null hypothesis, namely,

$$\Pr(R \,|\, \mathbb{H}_0 \text{ is false}) = 1 - \Pr(\bar{R} \,|\, \mathbb{H}_0 \text{ is false}) = 1 - \Pr(\text{type II error})$$

So, we should attempt to obtain the most powerful test as possible if we wish to minimize

[1] Recall that if X is a Poisson with arrival rate λ, then $\mathbb{E}(X) = \text{var}(X) = \lambda$.

the likelihood of having a type II error. In general, the power of a test is a function of the value of the parameter vector under the alternative. The power function degenerates to a constant only in the event of a simple alternative hypothesis, viz. $\mathbb{H}_1 : \theta = \theta_1$. To work out the logic of the derivation of the power, let's revisit the barista example from the previous section.

Example: Suppose that it actually takes on average 24 seconds for pouring a perfect expresso. In the previous section, we have computed a large-sample approximation under the null for the distribution of the sample mean. We now derive the asymptotic power of the means test at the α level of significance conditioning on $\mu = 24$.

The probability of falling into the rejection region is

$$\begin{aligned}
&\Pr\left(\sqrt{16}\left|\frac{\bar{X}_{16}-28}{6}\right|>1.96\,\middle|\,\mu=24\right)\\
&=1-\Pr\left(2\left|\frac{\bar{X}_{16}-28}{3}\right|\leq 1.96\,\middle|\,\mu=24\right)\\
&=1-\Pr\left(28-2.94\leq\bar{X}_{16}\leq 28+2.94\,\middle|\,\mu=24\right)\\
&=1-\Pr\left(25.06\leq\bar{X}_{16}\leq 30.94\,\middle|\,\mu=24\right)\\
&=1-\Pr\left(\sqrt{16}\,\frac{25.06-24}{6}\leq\sqrt{16}\,\frac{\bar{X}_{16}-24}{6}\leq\sqrt{16}\,\frac{30.94-24}{6}\,\middle|\,\mu=24\right)\\
&\cong 1-\left[\Phi\left(\frac{6.94}{3/2}\right)-\Phi\left(\frac{1.06}{3/2}\right)\right]\\
&=1-0.999998142+0.760113176=0.760115034.
\end{aligned}$$

Note that this power figure holds only asymptotically for we are taking the normal approximation for the unknown distribution of the sample mean.

In general, to compute the (asymptotic) power function of a two-sided means test, it suffices to appreciate that the probability of rejecting the null for $\mu=\mu_1\neq\mu_0$ is

$$\begin{aligned}
&\Pr\left(\sqrt{N}\left|\frac{\bar{X}_N-\mu_0}{\widehat{\sigma}_N}\right|>z_{1-\alpha/2}\,\middle|\,\mu=\mu_1\right)\\
&=1-\Pr\left(\sqrt{N}\left|\frac{\bar{X}_N-\mu_0}{\widehat{\sigma}_N}\right|\leq z_{1-\alpha/2}\,\middle|\,\mu=\mu_1\right)\\
&=1-\Pr\left(-z_{1-\alpha/2}\leq\sqrt{N}\,\frac{\bar{X}_N-\mu_0}{\widehat{\sigma}_N}\leq z_{1-\alpha/2}\,\middle|\,\mu=\mu_1\right)\\
&=1-\Pr\left(\mu_0-z_{1-\alpha/2}\,\frac{\widehat{\sigma}_N}{\sqrt{N}}\leq\bar{X}_N\leq\mu_0+z_{1-\alpha/2}\,\frac{\widehat{\sigma}_N}{\sqrt{N}}\,\middle|\,\mu=\mu_1\right)\\
&=1-\Pr\left(\mu_0-\mu_1-z_{1-\alpha/2}\,\frac{\widehat{\sigma}_N}{\sqrt{N}}\leq\bar{X}_N-\mu_1\leq\mu_0-\mu_1+z_{1-\alpha/2}\,\frac{\widehat{\sigma}_N}{\sqrt{N}}\,\middle|\,\mu=\mu_1\right)\\
&=1-\Pr\left(\sqrt{N}\,\frac{\mu_0-\mu_1}{\widehat{\sigma}_N}-z_{1-\alpha/2}\leq\sqrt{N}\,\frac{\bar{X}_N-\mu_1}{\widehat{\sigma}_N}\leq\sqrt{N}\,\frac{\mu_0-\mu_1}{\widehat{\sigma}_N}+z_{1-\alpha/2}\,\middle|\,\mu=\mu_1\right)\\
&\cong 1-\Phi\left(\sqrt{N}\,\frac{\mu_0-\mu_1}{\widehat{\sigma}_N}+z_{1-\alpha/2}\right)+\Phi\left(\sqrt{N}\,\frac{\mu_0-\mu_1}{\widehat{\sigma}_N}-z_{1-\alpha/2}\right).
\end{aligned}$$

Note that the power function converges to one as the sample size increases provided that $\mu\neq\mu_1$ because both cumulative distribution functions converge to the same value (namely, ± 1 depending on whether $\mu_0\gtrless\mu_1$).

It is straightforward to deal with one-sided tests as well. For instance, for a means test of $\mathbb{H}_0 : \mu = \mu_0$ against $\mathbb{H}_1 : \mu > \mu_0$, the test statistic is $\frac{\bar{X}_N - \mu_0}{\dot{\sigma}_N/\sqrt{N}}$ with an asymptotic critical value given by the $(1-\alpha)$th percentile of the standard normal distribution given that $\Pr\left(\frac{\bar{X}_N - \mu_0}{\dot{\sigma}_N/\sqrt{N}} > z_{1-\alpha}\right) \cong \alpha$ under the null hypothesis. Letting $\mu_1 > \mu_0$ denote a mean value under the alternative yields a power of

$$\begin{aligned}
\Pr\left(\sqrt{N}\,\frac{\bar{X}_N - \mu_0}{\widehat{\sigma}_N} > z_{1-\alpha} \,\middle|\, \mu = \mu_1\right) &= 1 - \Pr\left(\bar{X}_N \leq \mu_0 + z_{1-\alpha}\,\frac{\widehat{\sigma}_N}{\sqrt{N}} \,\middle|\, \mu = \mu_1\right) \\
&= 1 - \Pr\left(\bar{X}_N - \mu_1 \leq \mu_0 - \mu_1 + z_{1-\alpha}\,\frac{\widehat{\sigma}_N}{\sqrt{N}} \,\middle|\, \mu = \mu_1\right) \\
&= 1 - \Pr\left(\frac{\bar{X}_N - \mu_1}{\widehat{\sigma}_N/\sqrt{N}} \leq \frac{\mu_0 - \mu_1}{\widehat{\sigma}_N/\sqrt{N}} + z_{1-\alpha} \,\middle|\, \mu = \mu_1\right) \\
&\cong 1 - \Phi\left(\sqrt{N}\,\frac{\mu_0 - \mu_1}{\widehat{\sigma}_N} + z_{1-\alpha}\right).
\end{aligned}$$

As before, power converges to one as the sample size increases. This property is known as consistency. We say a test is consistent if it has asymptotic unit power for any fixed alternative.

In the previous chapter, we have seen that it is typically very difficult to obtain efficient estimators if we do not restrict attention to a specific class (e.g., class of unbiased estimators). The same problem arises if we wish to derive a uniformly most powerful test at a certain significance level. Unless we confine attention to simple null and alternative hypotheses, it is not possible to derive optimal tests without imposing further restrictions. To appreciate why, it suffices to imagine a situation in which we wish to test $\mathbb{H}_0 : \theta = \theta_0$ against $\mathbb{H}_1 : \theta \neq \theta_0$. It is easy to see that the one-sided test for $\mathbb{H}_0 : \theta = \theta_0$ against $\mathbb{H}_0 : \theta > \theta_0$ is more powerful than the two-sided test if $\theta = \theta_1 > \theta_0$, just as the one-sided test for $\mathbb{H}_0 : \theta = \theta_0$ against $\mathbb{H}_0 : \theta < \theta_0$ is more powerful than the two-sided test if $\theta = \theta_1 < \theta_0$.

Figure 7.2 illustrates this fact by plotting the power functions of one-sided tests for $\mathbb{H}_0 : \theta = \theta_0$ against either $\mathbb{H}_1 : \theta < \theta_0$ or $\mathbb{H}_1 : \theta > \theta_0$ at the α and $\alpha/2$ level of significance. The power of the one-sided tests are inferior to their levels of significance for values of θ that strongly contradict the alternative hypothesis (e.g., large positive values for $\mathbb{H}_1 : \theta < \theta_0$).

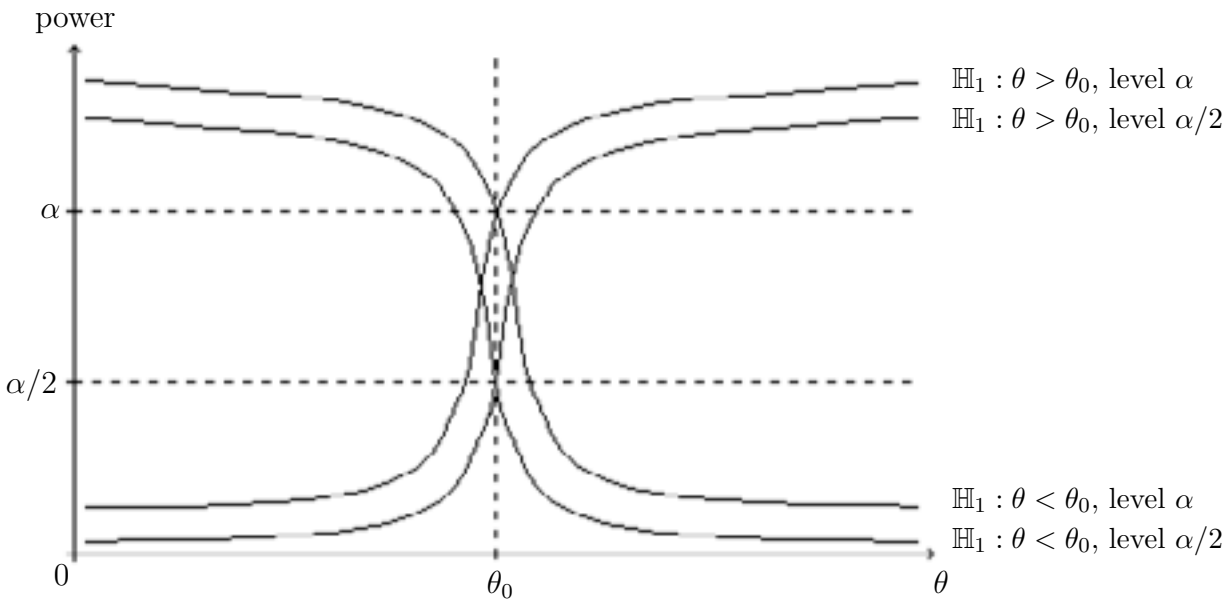

Figure 7.2: Power functions of one-sided tests for $\mathbb{H}_0 : \theta = \theta_0$

This is natural, though not acceptable for a test of $\mathbb{H}_0 : \theta = \theta_0$, because these tests are not designed to look at deviations from the null in both directions. That's exactly why we prefer to restrict attention to unbiased tests, that is to say, tests whose power are always above size. Applying such a criterion to the above situation clarifies why most people would prefer the two-sided test instead of one of the one-sided tests. To obtain the power function of a two-sided test of $\mathbb{H}_0 : \theta = \theta_0$, it suffices to sum up the power function of the one-sided tests at $\alpha/2$ significance level against $\mathbb{H}_1 : \theta > \theta_0$ and $\mathbb{H}_1 : \theta < \theta_0$.

7.3 Interpreting p-values

The Neyman-Pearson paradigm leads to a dichotomy in the context of hypothesis testing in that we can either reject or not the null hypothesis given a certain significance level. We would expect however that there are rejections and rejections. How far a test statistic extends into the rejection region should intuitively convey some information about the weight of the sample evidence against the null hypothesis. To measure how much evidence we have against the null, we employ the concept of p-value, which refers to the probability under the null that the value of the test statistic is at least as extreme as the one we actually observe in the sample. Smaller p-values correspond to more conclusive sample evidence given that we impose the null. In other words, the p-value is the smallest significance level at which we would reject the null hypothesis given the observed value of the test statistic.

Computing p-values is like taking the opposite route we take to derive a rejection region. To obtain the latter, we fix the level of significance α in the computation of the critical values. To find a p-value of an one-sided test, we compute the tail probability of the test statistic by evaluating the corresponding distribution at the sample statistic. As for two-sided tests, we must just multiply the one-sided p-value by two if the sampling distribution is symmetric. The main difference between the level of significance and the p-value is that the latter is a function of the sample, whereas we the former is a fixed probability that we choose ex-ante.

For instance, the p-value of an asymptotic means test is

$$\Pr\left(\sqrt{N}\,\frac{\bar{X}_N-\mu_0}{\widehat{\sigma}_N}>\sqrt{N}\,\frac{\bar{x}_N-\mu_0}{\widehat{\sigma}_N}\right)=1-\Phi\left(\sqrt{N}\,\frac{\bar{x}_N-\mu_0}{\widehat{\sigma}_N}\right)$$

if the alternative hypothesis is $\mathbb{H}_1:\mu>\mu_0$, whereas it is

$$\Pr\left(\sqrt{N}\,\frac{\bar{X}_N-\mu_0}{\widehat{\sigma}_N}<\sqrt{N}\,\frac{\bar{x}_N-\mu_0}{\widehat{\sigma}_N}\right)=\Phi\left(\sqrt{N}\,\frac{\bar{x}_N-\mu_0}{\widehat{\sigma}_N}\right)$$

for $\mathbb{H}_1:\mu<\mu_0$. As for two-sided tests, the p-value reads

$$2\Pr\left(\sqrt{N}\,\frac{\bar{X}_N-\mu_0}{\widehat{\sigma}_N}>\sqrt{N}\,\left|\frac{\bar{x}_N-\mu_0}{\widehat{\sigma}_N}\right|\right)=2\left[1-\Phi\left(\sqrt{N}\,\left|\frac{\bar{x}_N-\mu_0}{\widehat{\sigma}_N}\right|\right)\right]$$

for $\mathbb{H}_1 : \mu \neq \mu_0$. To better understand how we compute p-values, let's revisit the barista example one more time.

Example: Under the null distribution that it takes on average 28 seconds for pouring a perfect expresso, the asymptotic normal approximation for the distribution of the sample mean implies the following p-value for a sample mean of 26 seconds:

$$2\Pr\left(\sqrt{16}\,\frac{\bar{X}_{16}-28}{6} > \sqrt{16}\left|\frac{\bar{x}_{16}-28}{6}\right| \,\middle|\, \mathbb{H}_0 : \mu = 28\right) \cong 2\left[1-\Phi\left(4\left|\frac{26-28}{6}\right|\right)\right]$$
$$= 2[1-\Phi(4/3)] = 2(1-0.90878878)$$
$$= 0.18242244.$$

This means that we cannot reject the null hypothesis at the usual levels of significance (i.e., 1%, 5% and 10%). We must be ready to consider a level of significance of about 18.25% if we really wish to reject the null.

Before concluding this section, it is useful to talk about what p-value is not about. First, it is not about the probability that the null hypothesis is true. We could never produce such a probability. We compute the p-value under the null and hence it cannot say anything about how likely the null hypothesis is. In addition, it does not make any sense to compute the probability of a hypothesis given that the latter is not a random variable. Second, a large p-value does not necessarily imply that the null is true. It just means that we don't have enough evidence to reject it. Third, the p-value does not say anything about the magnitude of the deviation with respect to the null hypothesis. To sum up, the p-value entails the confidence that we may have in the null hypothesis to explain the result we actually observe in the sample.

7.4 Likelihood-based tests

The discussion in Section suggests that it is very often the case there is no uniformly most powerful test for a given set of null and alternative hypotheses. It turns nonetheless out that

likelihood-based tests typically yield very powerful tests in a wide array of situations. In particular, if it exists, a uniformly most powerful (unbiased) test is very often equivalent to a likelihood-based test. This means that likelihood methods entail not only efficient estimators, but also a framework to build satisfactory tests.

Let $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^k$ denote a k-dimensional parameter vector of which the likelihood $\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{X})$ is a function. Consider the problem of testing the composite null hypothesis $\mathbb{H}_0 : \boldsymbol{\theta} \in \Theta_0$ against the composite alternative hypothesis $\mathbb{H}_0 : \boldsymbol{\theta} \in \Theta - \Theta_0$. We now define the likelihood ratio as

$$\lambda(\boldsymbol{X}) \equiv \frac{\max_{\theta \in \Theta_0} \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{X})}{\max_{\theta \in \Theta} \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{X})} = \frac{\mathcal{L}(\widehat{\boldsymbol{\theta}}_N^{(0)}; \boldsymbol{X})}{\mathcal{L}(\widehat{\boldsymbol{\theta}}_N; \boldsymbol{X})},$$

where $\widehat{\boldsymbol{\theta}}_N^{(0)}$ and $\widehat{\boldsymbol{\theta}}_N$ are the restricted and unrestricted maximum likelihood estimators, respectively. The restricted optimization means that we search for the parameter vector that maximizes the log-likelihood function only within the null parameter space Θ_0, whereas the unrestricted optimization yields the usual ML estimator of $\boldsymbol{\theta}$.

The intuition for a likelihood-ratio test is very simple. In the event that the null hypothesis is true, the unrestricted optimization will (in the limit as $N \to \infty$) yield a value for the parameter vector within Θ_0 and hence the log-likelihood ratio will take a unit value. If the null is false, then the unrestricted optimization will yield a value for $\boldsymbol{\theta} \in \Theta - \Theta_0$ and hence the ratio will take a value below one. This suggests a rejection region of the form $\{\boldsymbol{X} : \lambda(\boldsymbol{X}) \leq C_\alpha\}$ for some constant $0 \leq C_\alpha \leq 1$ that depends on the significance level α.

Example: Let $\boldsymbol{X}$ denote a random sample from a normal distribution with mean μ and variance σ^2. Suppose that the interest lies on testing the null hypothesis $\mathbb{H}_0 : \mu = \mu_0$ against the alternative $\mathbb{H}_1 : \mu \neq \mu_0$ by means of likelihood methods. As the (unrestricted) likelihood function is $(2\pi\sigma^2)^{-N/2} \exp\left[\frac{1}{2\sigma^2}\sum_{i=1}^N (X_i - \mu)^2\right]$, the (unrestricted) maximum likelihood estimators for μ and σ^2 are the sample mean $\bar{X}_N$ and sample variance $\widehat{\sigma}_N^2$. In contrast, confining attention to the null hypothesis yields a restricted likelihood function of $(2\pi\sigma^2)^{-N/2} \exp\left[\frac{1}{2\sigma^2}\sum_{i=1}^N (X_i - \mu_0)^2\right]$ with restricted ML estimators given by μ_0 and $\tilde{\sigma}_N^2 = \frac{1}{N}\sum_{i=1}^N (X_i - \mu_0)^2$. It then follows that the likelihood ratio is

$$\begin{aligned}
\lambda(\boldsymbol{X}) &= \frac{(2\pi\tilde{\sigma}_N^2)^{-N/2} \exp\left[\frac{1}{2\tilde{\sigma}_N^2}\sum_{i=1}^N (X_i - \mu_0)^2\right]}{(2\pi\widehat{\sigma}_N^2)^{-N/2} \exp\left[\frac{1}{2\tilde{\sigma}_N^2}\sum_{i=1}^N (X_i - \bar{X}_N)^2\right]} \\
&= \frac{\left[\frac{1}{N}\sum_{i=1}^N (X_i - \mu_0)^2\right]^{-N/2} \exp\left[\frac{N\sum_{i=1}^N (X_i-\mu_0)^2}{2\sum_{i=1}^N (X_i-\mu_0)^2}\right]}{\left[\frac{1}{N}\sum_{i=1}^N (X_i - \bar{X}_N)^2\right]^{-N/2} \exp\left[\frac{N\sum_{i=1}^N (X_i-\bar{X}_N)^2}{2\sum_{i=1}^N (X_i-\bar{X}_N)^2}\right]} \\
&= \frac{\left[\sum_{i=1}^N (X_i - \mu_0)^2\right]^{-N/2} \exp(N/2)}{\left[\sum_{i=1}^N (X_i - \bar{X}_N)^2\right]^{-N/2} \exp(N/2)} = \left[\frac{\sum_{i=1}^N (X_i - \mu_0)^2}{\sum_{i=1}^N (X_i - \bar{X}_N)^2}\right]^{-N/2}.
\end{aligned}$$

To compute the critical value k_α of the rejection region, we must first derive the distribution of $\lambda(\boldsymbol{X})$ under the null distribution. This may look like a daunting task, but it is actually straightforward for we can write the numerator of the fraction as

$$\sum_{i=1}^N (X_i - \mu_0)^2 = \sum_{i=1}^N (X_i - \bar{X}_N + \bar{X}_N - \mu_0)^2 = \sum_{i=1}^N (X_i - \bar{X}_N)^2 - N(\bar{X}_N - \mu_0)^2,$$

which implies that

$$\lambda(\boldsymbol{X}) = \left[1 + \frac{N(\bar{X}_N - \mu_0)^2}{\sum_{i=1}^{N}(X_i - \bar{X}_N)^2}\right]^{-N/2}.$$

Well, now it suffices to appreciate that the likelihood ratio is a monotone decreasing function of $\sqrt{N}(\bar{X}_N - \mu_0)/s_N$ given that the fraction within brackets is the square of the latter divided by $N-1$. It then follows from $\sqrt{N}(\bar{X}_N - \mu_0)/s_N \sim t_{N-1}$ that a rejection region of the form $\{\boldsymbol{X} : \left|\sqrt{N}(\bar{X}_N - \mu_0)/s_N\right| \geq t_{N-1}(1-\alpha/2)\}$, where $t_{N-1}(1-\alpha/2)$ is the $(1-\alpha/2)$th percentile of a t-student distribution with $N-1$ degrees of freedom, yields a test with a significance level of α.

The above example shows that it is possible to compute the rejection rate of a likelihood ratio test by looking at whether it depends exclusively on a statistic with a known sampling distribution. In general, however, it is very difficult to derive the exact sampling distribution of the likelihood ratio and so we must employ asymptotic approximations. Assume, for instance, that $\boldsymbol{X} = (X_1, \ldots, X_N)$ is a random sample from a distribution F_θ and that we wish to test $\mathbb{H}_0 : \theta = \theta_0$ against $\mathbb{H}_0 : \theta \neq \theta_0$. The fact that the unrestricted ML estimator is consistent under both the null and alternative hypotheses ensures that $\ln \mathcal{L}(\theta_0; \boldsymbol{X})$ admits a Taylor expansion around $\widehat{\theta}_N$:

$$\begin{aligned}\ln \mathcal{L}(\theta_0; \boldsymbol{X}) = \ln \mathcal{L}(\widehat{\theta}_N; \boldsymbol{X}) + \frac{\partial}{\partial\theta} \ln \mathcal{L}(\widehat{\theta}_N; \boldsymbol{X})(\theta_0 - \widehat{\theta}_N) + \frac{1}{2}\frac{\partial^2}{\partial\theta^2} \ln \mathcal{L}(\widehat{\theta}_N; \boldsymbol{X})(\theta_0 - \widehat{\theta}_N)^2 \\ + \frac{1}{6}\frac{\partial^3}{\partial\theta^3} \ln \mathcal{L}(\theta_*; \boldsymbol{X})(\theta_0 - \widehat{\theta}_N)^3,\end{aligned}$$

where $\theta_* = \lambda\theta_0 + (1-\lambda)\widehat{\theta}_N$ for $0 \leq \lambda \leq 1$. The definition of the ML estimator is such that the first derivative of the log-likelihood function is zero, whereas the fact that $\widehat{\theta}_N$ is a $\sqrt{N}-$consistent estimator ensures that the last term of the expansion converges to zero at a very fast rate. It then follows that

$$-2\ln \lambda(\boldsymbol{X}) = 2\left(\ln \mathcal{L}(\widehat{\theta}_N; \boldsymbol{X}) - \ln \mathcal{L}(\theta_0; \boldsymbol{X}\right) \cong -\frac{\partial^2}{\partial\theta^2}\mathcal{L}(\widehat{\theta}_N; \boldsymbol{X})(\widehat{\theta}_N - \theta_0)^2. \qquad (7.1)$$

Now, we know that under the null $\sqrt{N}(\widehat{\theta}_N - \theta_0)$ weakly converges to a normal distribution

with mean zero and variance given by the inverse of the information matrix

$$\mathcal{I}_\infty(\theta_0) \equiv -\lim_{N\to\infty} \frac{1}{N} \frac{\partial^2}{\partial\theta^2} \mathcal{L}(\widehat{\theta}_N; \boldsymbol{X}).$$

This means that $LR = -2\ln\lambda(\boldsymbol{X})$ is asymptotically chi-square with one degree of freedom for the right-hand side of (7.1) is the square of a standard normal variate. This suggests that a test that rejects the null hypothesis if the likelihood ratio $LR \geq \chi_1^2(1-\alpha)$, where the latter denotes the $(1-\alpha)$th percentile of the chi-square distribution with one degree of freedom, is asymptotically of level α.

Example: Let $X_i \sim$ iid Poisson(λ) for $i = 1, \ldots, N$ and define the null and alternative hypotheses as $\mathbb{H}_0 : \lambda = \lambda_0$ and $\mathbb{H}_1 : \lambda \neq \lambda_0$, respectively. The likelihood ratio then is

$$\begin{aligned} LR = -2\ln\lambda(\boldsymbol{X}) &= -2\ln \frac{\exp(-N\lambda_0)\,\lambda_0^{\sum_{i=1}^N X_i}}{\exp(-N\widehat{\lambda}_N)\,\widehat{\lambda}_N^{\sum_{i=1}^N X_i}} \\ &= -2N\left[(\lambda_0 - \widehat{\lambda}_N) - \widehat{\lambda}_N \ln(\lambda_0/\widehat{\lambda}_N)\right] \xrightarrow{d} \chi_1^2, \end{aligned}$$

where $\widehat{\lambda}_N = \frac{1}{N}\sum_{i=1}^N X_i$ is the ML estimator of the Poisson arrival rate.

We next extend this result to a more general setting as well as derive two additional likelihood-based tests that are asymptotically equivalent to the likelihood ratio test. We start by establishing some notation. Let $\Theta_0 = \{\boldsymbol{\theta} : \boldsymbol{R}(\boldsymbol{\theta}) = \mathbf{0},\ \boldsymbol{\theta} \in \Theta\}$, where $\boldsymbol{R}(\boldsymbol{\theta}) = \mathbf{0}$ represents a system of r nonlinear equations concerning $\boldsymbol{\theta}$. For instance, we could think of testing whether $\theta_1 + \theta_2 = 1$ and $\theta_3 = \ldots = \theta_k = 0$, giving way to a system of $r = k - 1$ restrictions of the form $\boldsymbol{R}(\boldsymbol{\theta}) = (\theta_1 + \theta_2 - 1, \theta_3, \ldots, \theta_k)' = \mathbf{0}$. Recall that the unrestricted maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_N$ is such that $\sqrt{N}(\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathcal{I}_\infty^{-1}(\boldsymbol{\theta}))$ and that the score function is such that $\frac{1}{\sqrt{N}} \frac{\partial}{\partial \boldsymbol{\theta}'} \ln \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{X}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathcal{I}_\infty(\boldsymbol{\theta}))$, where $\mathcal{I}_\infty(\boldsymbol{\theta})$ is the information matrix. In contrast, the restricted maximum likelihood estimator $\tilde{\boldsymbol{\theta}}_N$ maximizes the log-likelihood function subject to $\boldsymbol{R}(\boldsymbol{\theta}) = \mathbf{0}$ (and so it does not equate the score function to zero for it has to account for the Lagrange multiplier term).

Along the same lines as before, the likelihood ratio is

$$\begin{aligned} LR = -2 \ln \lambda(\boldsymbol{X}) &= 2 \left(\ln \mathcal{L}(\widehat{\boldsymbol{\theta}}_N; \boldsymbol{X}) - \ln \mathcal{L}(\tilde{\boldsymbol{\theta}}_N; \boldsymbol{X} \right) \\ &\cong (\widehat{\boldsymbol{\theta}}_N - \tilde{\boldsymbol{\theta}}_N)' \left[-\frac{\partial}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \mathcal{L}(\widehat{\boldsymbol{\theta}}_N; \boldsymbol{X}) \right] (\widehat{\boldsymbol{\theta}}_N - \tilde{\boldsymbol{\theta}}_N) \end{aligned} \tag{7.2}$$

given that, under the null, a Taylor expansion is admissible for both estimators are consistent and hence close to each other. Now, it is possible to show that, under the null, the asymptotic variance of $\sqrt{N}(\widehat{\boldsymbol{\theta}}_N - \tilde{\boldsymbol{\theta}}_N)$ is

$$\lim_{N \to \infty} \left[-\frac{1}{N} \frac{\partial}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \mathcal{L}(\widehat{\boldsymbol{\theta}}_N; \boldsymbol{X}) \right]^{-1}.$$

This implies that the right-hand side of (7.2) converges in distribution to a chi-square with r degrees of freedom. To appreciate why, it suffices to observe that $\widehat{\boldsymbol{\theta}}_N$ and $\tilde{\boldsymbol{\theta}}_N$ respectively estimate k and $k - r$ free parameters, so that their difference concerns only r elements.

Figure 7.3 shows that the likelihood ratio test gauges the difference between the criterion function that we maximize either with or without constraints. It also illustrate two alternative routes to assess whether the data is consistent with the constraints in the parameter space. The first is to measure the difference between the restricted and unrestricted ML

estimators or, equivalently, to evaluate whether the unrestricted ML estimator satisfies the restriction in the null hypothesis. This testing strategy gives way to what we call Wald tests. The second route is to evaluate whether the score function of the constrained ML estimator is close to zero. The motivation lies on the fact that, in the limit, it is completely costless to impose a true null. This translates into a Lagrange multiplier in the vicinity of zero, so that the first-order condition reduces to equating the score function to zero. Lagrange multiplier tests rely then on measuring how different from zero is the score function evaluated at the constrained ML estimator.

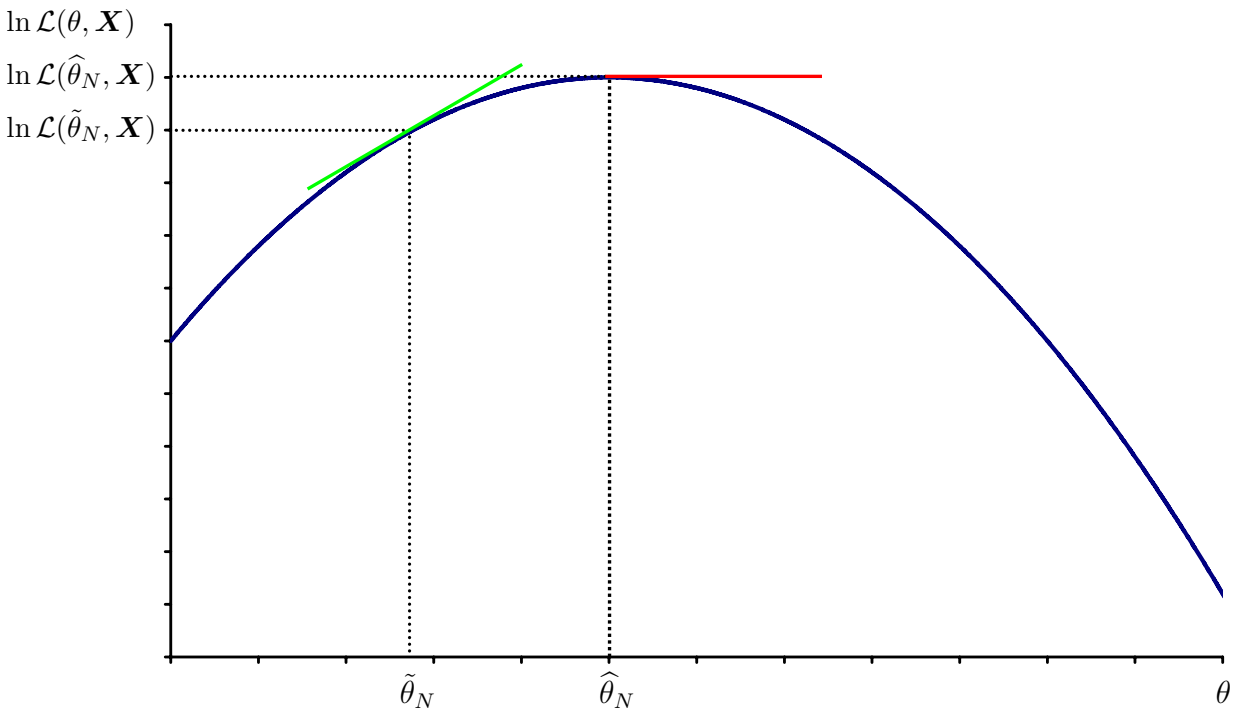

Figure 7.3: Likelihood-based tests based on unrestricted and restricted ML estimators ($\widehat{\theta}_N$ and $\tilde{\theta}_N$, respectively). The log-likelihood test measures the difference between the constrained and unconstrained log-likelihood functions, whereas the Wald test gauges the difference between the unrestricted and restricted ML estimators. The Lagrange multiplier test assesses the magnitude of the constrained score function by focusing on the slope of the green line. The zero slope of the red line reflects the fact that the unconstrained score function is equal to zero by definition.

We first show how to compute Wald tests and then discuss Lagrange multiplier tests. As usual, we will derive the necessary asymptotic theory by means of Taylor expansions. Wald tests are about whether the unconstrained ML estimator meets the restrictions in the null hypothesis and so we start with a Taylor expansion of $\boldsymbol{R}(\boldsymbol{\theta})$ around $\widehat{\boldsymbol{\theta}}_N$, namely,

$$\boldsymbol{R}(\boldsymbol{\theta}) \cong \boldsymbol{R}(\widehat{\boldsymbol{\theta}}_N) + \boldsymbol{R}_{\boldsymbol{\theta}}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_N) \qquad \text{with } \boldsymbol{R}_{\boldsymbol{\theta}} = \frac{\partial}{\partial \boldsymbol{\theta}'}\boldsymbol{R}(\boldsymbol{\theta}).$$

It is now evident that $\sqrt{N}[\boldsymbol{R}(\widehat{\boldsymbol{\theta}}_N) - \boldsymbol{R}(\boldsymbol{\theta})]$ will converge to a multivariate normal distribution with mean zero and covariance matrix given by $\boldsymbol{R}_{\boldsymbol{\theta}}\,\mathcal{I}_\infty^{-1}(\boldsymbol{\theta})\,\boldsymbol{R}_{\boldsymbol{\theta}}'$.[2] Well, if the null is true, we expect that the (unrestricted) ML estimator will approximately satisfy the system of nonlinear restrictions in that $\boldsymbol{R}(\widehat{\boldsymbol{\theta}}_N) \cong \mathbf{0}$. This suggests gauging whether the magnitude of $\boldsymbol{R}(\widehat{\boldsymbol{\theta}}_N)$ deviates from zero significantly as a way of testing $\mathbb{H}_0$ against $\mathbb{H}_1$. In particular, we know that $\sqrt{N}\boldsymbol{R}(\widehat{\boldsymbol{\theta}}_N) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{R}_{\boldsymbol{\theta}}\,\mathcal{I}_\infty^{-1}(\boldsymbol{\theta})\,\boldsymbol{R}_{\boldsymbol{\theta}}')$ under the null and hence it suffices to take a quadratic form of $\sqrt{N}\boldsymbol{R}(\widehat{\boldsymbol{\theta}}_N)$ normalized by its covariance matrix to end up with an asymptotically chi-square distribution with r degrees of freedom, namely, $W \equiv N\,\boldsymbol{R}(\widehat{\boldsymbol{\theta}}_N)'[\boldsymbol{R}_{\boldsymbol{\theta}}\,\mathcal{I}_\infty^{-1}(\boldsymbol{\theta})\,\boldsymbol{R}_{\boldsymbol{\theta}}']^{-1}\boldsymbol{R}(\widehat{\boldsymbol{\theta}}_N) \xrightarrow{d} \chi_r$. Note that by taking a quadratic form we automatically avoid negative and positive deviations from zero to cancel out. The asymptotic Wald test then rejects the null at the α significance level if $W \geq \chi_r^2(1-\alpha)$, where the latter denotes the $(1-\alpha)$th percentile of the chi-square distribution with r degrees of freedom.

Example: Let $X_i \sim$ iid $\mathcal{B}(1,p)$ for $i = 1, \ldots, N$. Define the null and alternative hypotheses as $\mathbb{H}_0 : p = p_0$ and $\mathbb{H}_1 : p \neq p_0$, respectively. The unconstrained maximum likelihood estimator of p is the sample mean $\widehat{p}_N = \sum_{i=1}^N X_i$, whose variance is $p(1-p)/N$. Applying a central limit theorem then yields

$$W = N\,\frac{(\widehat{p}_N - p_0)^2}{\widehat{p}_N(1-\widehat{p}_N)} \xrightarrow{d} \chi_1^2$$

suggesting us to reject the null at the α significance level if $W \geq \chi_1^2(1-\alpha)$.

[2] See Footnote 1 in Section 6.1.5 for a very brief discussion about the multivariate normal distribution.

We now turn our attention to the Lagrange multiplier test. The score function $\frac{\partial}{\partial \boldsymbol{\theta}'} \ln \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{X})$ is on average zero for any $\boldsymbol{\theta} \in \Theta$ and hence it is zero also for any $\boldsymbol{\theta} \in \Theta_0$. In addition, the variance of the score function is under the null equal to

$$\text{var}\left(\frac{\partial}{\partial \boldsymbol{\theta}'} \ln \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{X}) \middle| \boldsymbol{\theta} \in \Theta_0 \right) = -\mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \ln \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{X}) \middle| \boldsymbol{\theta} \in \Theta_0 \right] \equiv \mathcal{I}_N(\boldsymbol{\theta}),$$

which in the limit coincides with the information matrix $\mathcal{I}_\infty(\boldsymbol{\theta})$. It thus follows that

$$LM = \tfrac{\partial}{\partial \boldsymbol{\theta}'} \ln \mathcal{L}(\tilde{\boldsymbol{\theta}}_N; \boldsymbol{X})' \mathcal{I}_N^{-1}(\tilde{\boldsymbol{\theta}}_N) \tfrac{\partial}{\partial \boldsymbol{\theta}'} \ln \mathcal{L}(\tilde{\boldsymbol{\theta}}_N; \boldsymbol{X}) \xrightarrow{d} \chi^2_r$$

and hence we must reject the null hypothesis if $LM \geq \chi^2_r(1-\alpha)$ to obtain an asymptotic test of level α. Note that the chi-square distribution has r degrees of freedom even though $\tilde{\boldsymbol{\theta}}_N$ has $k-r$ free parameters. This is because the score of the $k-r$ free parameters must equate to zero, remaining only r dimensions for the score function to vary (i.e., those affected by the restrictions).

Example: Let's revisit the previous example in which $\boldsymbol{X} = (X_1, \ldots, X_N)$ with $X_i \sim$ iid $\mathcal{B}(1, p)$ for $i = 1, \ldots, N$. The LM test statistic for $\mathbb{H}_0 : p = p_0$ against $\mathbb{H}_1 : p \neq p_0$ then is

$$LM = N \frac{(\widehat{p}_N - p_0)^2}{p_0(1-p_0)} \xrightarrow{d} \chi^2_1$$

given that the score function evaluated at p_0 is $(\widehat{p}_N - p_0)/[p_0(1-p_0)/N]$ and the corresponding information matrix is $N/[p_0/(1-p_0)]$. We would thus reject the null if $LM \geq \chi^2_1(1-\alpha)$ to obtain an asymptotic test at the α level of significance.

In the above example, it is evident that the Wald and LM tests are asymptotically equivalent for the difference between their denominators shrink to zero under the null as the sample mean $\widehat{p}_N$ converges almost surely to p_0. This asymptotic equivalence actually holds in general, linking not only Wald and LM tests but also likelihood ratio tests. This should come with no surprise given that the three statistics intuitively carry the same information as it is easily seen in Figure 7.3.