Introduction to statistics

Wenda Zhou

June 4, 2017

1 Introduction

1.1 What is statistics?

The Merriam-Webster dictionary defines statistics as a branch of mathematics dealing with the collection, analysis, intepretation and presentation of masses of numerical data. In particular, we will be most interested in understanding how to work with and analyze data in the presence of uncertainty and unknowns. Indeed, it is most often the case that the problems we encounter have inherent variability (e.g. noise in the measurements) and that we do not know or understand the complete process being studied. It is thus important to still be able to derive conclusions we can be confident abount despite the uncertainty.

The questions we will be interested in statistics can be roughly categorised into three broad types: prediction, estimation and testing.

Prediction Prediction is concerned about predicting quantities that are unknown to us by making use of related information that is known to us. For example, a weather forecaster predicts the weather tomorrow using meteorogical measures today and past information. Amazon predicts which products its customers are most likely to buy.

The problem of prediction has enjoyed an important revival in the past decade with the increase in computing power and the increasing amount of data being collected throughout the world, and is the central problem of machine learning.

Estimation Estimation is concerned about assigning values and uncertainty to quantities that are unknown and most often cannot be observed directly. For example, an economist may be interested in estimating the average effect (for example in lifetime earning increase) of obtaining a university degree. A pharmacologist may be interested in estimating the average effect (for example, in increase life expectancy) of a cancer treatment.

The problem of estimation is historically the central problem that motivated statistics. Although it shares many common aspects with prediction, it also differs in subtle but important ways.

Testing Testing is concerned about making decisions in the face of uncertainy, and is a problem that is closely related to that of estimation. For example, a drug company may be interested to understand whether a drug is more effective than the placebo. An advertiser may be interested in whether their advertisement is effective.

The problem of testing is very closely related to that of estimation, as it will often be the case that the test we wish to understand can be phrased in the sense of “is the effect zero"?. However, the language the notion of testing provides will prove valuable, and the fact that it is such a common problem warrants a separate mention.

1.2 Data

As the amount of data collected in the world increases, the diversity and variety of the data collected also increases. However, the vast majority of the data can still be understood in a simple rectangular fashion we describe below.

1.2.1 Rectangular data

It is often the case that we may think of a dataset as a collection of *observations* each having a collection of characteristics (often called *variables*). Most often, we are interested in how some variables (often called dependent or response variables) change or vary as a function of some other variables (often called *independent* or *explanatory* variables).

Example 1 (Clinical trial) Suppose a drug company did a clinical trial with a 100 patients for a drug designed to lower blood cholesterol. For each of them, they recorded the weight, age, sex, the blood cholesterol before and after taking the drug.

In this example, each patient corresponds to an observation, and the variables are the weight, age, sex, and blood cholesterol. In the context of a drug trial, we are interested in how effective the drug is, so the response variable could be the blood cholesterol after taking the drug, or maybe the difference in blood cholesterol before and after taking the drug. If we suspect that not everyone will respond similarly to the drug (e.g. the drug might be more effective for women than men), then we may be interested to consider the other variables as explanatory variables.

One interesting aspect in this case is that we may also want to consider the cholesterol before taking the drug as an explanatory variable. Indeed, if we believe that the effect of the drug depends on the initial amount of cholesterol (e.g. the drug works particularly well for people with very high levels of cholesterol, but not for others), then we certainly would want to understand it.

It will often be convenient to collect all the variables for all the observations into a *data matrix* or *data frame*. This is a rectangular table with each row corresponding to an observation and each column corresponding to a variable. It is usual to let n be the number of observations in the data frame.

Example 2 (Clinical trial (continued)) Suppose that we consider the same study as in example 1. We may collect all of the information into a data frame as below.

1.2.2 Data types

A given variable in a dataset will often have some restrictions on the values it can take, which we will refer to as the type of the variables. In example 1, the variable "sex" can only take two values (M / F), neither of which are numbers. On the other hand, the

Age	Sex	Weight (kg)	Cholesterol before (mg / dL)	Cholesterol after (mg / dL)
41	M	95	245	235
50	F	85	250	230
⋮	⋮	⋮	⋮	⋮

variable “weight” could (potentially) take any non-negative value. Most of the data we will study can be classified as either *numerical* or *categorical*.

Numerical variables

A numerical variable is a variable that represents a quantity and can take a range of numerical values. The quantity represented can be discrete such as a count (which may only take values 0, 1, . . .), or continuous (e.g. the concentration of blood lipids). In addition, the range of a numerical variable may often be restricted: some quantities are restricted to be non-negative (such as weight), and other quantities may be restricted to be between 0 and 1 (for example, the proportion of patient experiencing a side effect).

Be careful that not all numbers are numerical variables! For example, phone number or a zip code is not a numerical variable. Indeed, although they are numbers, they do not represent a quantity and we cannot add, subtract or average them.

A somewhat special yet oft encountered numerical variable type, *circular* variables represent quantities with somewhat unusual arithmetic properties, as they are values that “wrap around”. Common examples of circular data include time of day (e.g. 12.59pm is intuitively “close” to 1.01pm) or day of the year. We should take special care when operating on such variables and computing quantities such as averages.

Categorical variables

A categorical variable is a variable that represents groups or categories. A categorical variable usually takes on a finite number of possibilities known in advance (for exampled, male / female, or one of the 50 states of the U.S.). Each such category is called a *level*. In addition, all these examples display no particular order between the levels, and so are said to be *nominal*.

On the other hand, some categories may have a natural ordering. For example, suppose we asked the following question on a survey: “how often do you go to church”, and offered the following possible answers:

1. Less than once a year,
2. A few times per year,
3. A few times per month,
4. Every week,
5. Every day,

then it is clear that these categories have a natural ordering to them. We call such categories *ordinal*.

Likert scales

A very common type of variables in surveys and other similar datasets are the so called *Likert* scales (name after its inventor, psychologist Rensis Likert). The typical five-point likert item is of the form: “strongly disagree”, “disagree”, “neither agree nor disagree”, “agree”, “strongly agree”, although similar types of answers are also usually referred to as a Likert item.

From our previous discussion, Likert scales fall squarely into the ordinal categorical variable categories, with each category being ordered with respect to each other, but the variable not representing a specific numeric quantity. However, it is very common in practice to treat such scales as numeric, assigning for example a value from 1 to 5 to each item. Indeed, this simplification often works well in practice, especially when summarising large ensembles of Likert items.

Other types of data

Although the most common types of data fall within the numeric or categorical types, an increasing portion of the data collected today does not necessarilly belong to either of those types, or displays subtle differences and will require special treatment. We mention a couple of such data types for completeness, although we will not have the opportunity to study them in this course.

Graphs and networks An increasingly important type of data today emanates from the relationship between different entities. For example, social networks present a rich structure by examining for example each user’s friend or contact list, and the declared interests of each user. Such data is usually best summarised into a *graph*, which is often used to capture relationships between various entities (e.g. users).

Images, sound and other signals In recent years, there has been huge progress in the analysis of images, sounds and other types of signal (scientific imaging, neural data etc.). Google announced in May 2017 that their image recognition now outperforms humans on a benchmark dataset. Although the simplest models simply treat such signals as high-dimensional numeric variables (e.g. treating each pixel of the image as a number), the best models attempt to make use of the specific structure of those signals.

Text Parallel to the improvement in our ability to learn and process signals, the last five years has seen rapid improvement in our ability to understand text, most notably in fields such as machine translation and sentiment analysis. Although text is inherently a discrete structure, it is neither categorical nor numeric. The best models attempt to learn an efficient numeric representation of text.

1.3 Descriptive stastistics

Datasets are potentially complex objects with numerous variables, and so it is often useful for us to be able to synthetise the information of the entire dataset into some numbers, or descriptive statistics. These values that we compute are also called sample statistics, and hence will be referred to as *sample* (name-of-the-measure). In chapter 2, we will be instead looking at their theoretical counterparts, which will sometimes be called *population* (name-of-the-measure).

1.3.1 Measures of centrality

A *measure of centrality* is a number that attempts to summarise the location of the data in bulk, the two most common and well known being the (arithmetic) *mean* and the *median*.

The mean of a sample of n real-valued observations $x_1, \ldots, x_n$ is often written $\hat{x}$, and is defined as:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i. \tag{1.1}$$

The median of a sample of n real-valued observations $x_1, \ldots, x_n$ is the value of the observation such that half of the observation are smaller, and half of the observations are larger. This notion can be generalized to the notion of a *percentile*. The p^{th} percentile is the value such that $p\%$ of the observations are *below* the given value. The median is then the 50^{th} percentile. The often used first and third *quartiles* can also be defined as the 25^{th} and 75^{th} quartile.

1.3.2 Measures of dispersion

A *measure of dispersion* is a number that attempts to summarise the spread of the data. The most common measures are the variance (and its cousin the standard deviation), and the interquartile range.

The variance of a sample of n real-valued observations $x_1, \ldots, x_n$ is often written σ^2, and is defined as

$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2, \tag{1.2}$$

where $\bar{x}$ is the mean defined as in eq. (1.1). The standard deviation is then simply σ, the square root of the variance. The variance and standard deviation are also always non-negative quantities.

The *interquartile range*, often written *IQR*, is defined as the difference from the third quartile to the first quartile.

1.3.3 Measures of association

A *measure of association* is a number that attempts to summarise the association of two variables – i.e. how related they are. The most common such measure is the *covariance*, and its normalized version the *correlation*.

The covariance of two samples of n real-valued observations each, $x_1, \dots, x_n$, and $y_1, \dots, y_n$, is sometimes written σ_{xy}, and is defined by:

$$\sigma_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}). \tag{1.3}$$

Note that we may interpret the variance of x_i as the covariance of x_i with itself. The correlation is a normalized version of the covariance, and is a unit-free quantity, defined by:

$$\mathrm{corr}_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}, \tag{1.4}$$

where σ_x, σ_y are the standard deviations of the x_i and y_i respectively. Note that the correlation is always between -1 and 1, and has the same sign as the covariance. When the correlation is positive, we say that x and y are *positively correlated*. When the correlation is negative, we say that x and y are *negatively correlated*. When x and y are positively correlated, larger values of x tend to lead to larger values of y, whereas when x and y are negatively correlated, larger values of x tend to lead to smaller values of y.

1.3.4 Order statistics

For any sample of n real-valued observation $x_1, \dots, x_n$, we may be interested in the largest (or smallest) value. More generally, we may be interested in the p^{th} smallest (or largest) value. These values are called order statistics, and are usually written $x_{(p)}$ (note the parentheses, and pronounce "x order p"). By definition, $x_{(1)}$ is the first smallest – or simply smallest – value of the sample, whereas $x_{(n)}$ is the n^{th} smallest – or largest – value of the sample.

1.3.5 Descriptive statistics for categorical data

The descriptive statistics we have seen so far are inherently adapted to describing numerical data. However, they have no meaning when we consider categorical data. The most common summary for purely categorical data is called the *contigency table*, which collects the count of occurences of each category or combination of categories.

Indeed, suppose that we collect information concerning the hair colour (blonde, red, brown or black) and eye colour (blue, green, brown or black) of 20 individuals. In the rectangular data format that we are used to, that would correspond to 20 observations of 2 variables each, as in table 1.1. However, as the order of the observations does not matter, a way of summarising the data is the two-way contingency table, which records the number of individuals for each combination of eye colour and hair colour, as in table 1.2

In addition, we may also choose to ignore one or the other characteristic. Suppose for example that we only look at eye colour, ignoring hair colour. That is, we count the number of people with the given eye colour no matter what their hair colour is. We

Observation #	Hair Colour	Eye Colour
1	Brown	Blue
2	Blonde	Brown
⋮	⋮	⋮

Table 1.1: Sample of data collected from 20 individuals

	Eye colour			
Hair colour	Blue	Green	Brown	Black
Blonde	2	1	2	1
Red	1	1	2	0
Brown	1	0	4	2
Black	1	0	2	0

Table 1.2: contingency table of eye and hair colour for 20 individuals

Eye colour	Blue	Green	Brown	Black
	5	2	10	3

Table 1.3: Marginal distribution of eye colour

would then obtain a one-dimensional table, called the *marginal* table or distribution of hair colour. For example, see table 1.3.

On the other hand, instead of ignoring the hair colour, we may choose to only look at people with the given hair colour. This corresponds to looking at a single row of the two-way table, and is called the *conditional* table or distribution.

The notion of contigency table can be extended to more than two variables, but we are effectively *adding* a dimension for each variable. For example, suppose that we had also recorded the gender of the person in the previous example. We could then have a $4 \times 4 \times 2$ table of all the possibilities, but this can be difficult to present. A possibility is to present two 4×4 tables, one corresponding to men, the other to women. However, as the number of categories and variables increases, this unavoidably becomes more complex as the data itself becomes more complex.

1.3.6 Perils of descriptive statistics

Descriptive statistics are a convenient way of summarising often complex datasets. Due to their simplicity, they may however fail to capture the full complexity of the dataset. A common example is Anscombe's quartet, a collection of four samples plotted in fig. 1.1 that have the same mean and variance of both x and y, and the same covariance, but inherently different properties.

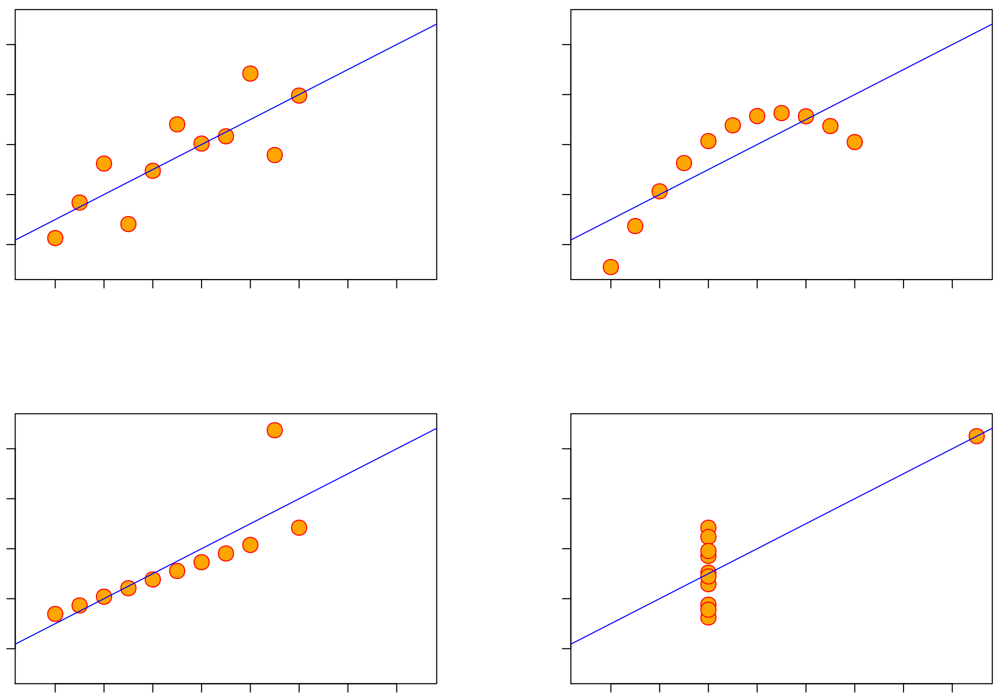

Figure 1.1: Anscombe's quartet, four datasets with the same mean, variance and covariance. Source: Schutz

Other common features that descriptive statistics often fail to capture are for example multi-modal datasets, where the data can be divided in several groups, or data with a lot of structure in the outliers. It is therefore often advantageous to not only look at summary statistics but also graphical representations of the data to better understand its structure.

1.4 Visualizing data

The first step in any analysis of the data is understanding the large structures that may exist in the data. The best way to achieve this is through graphing and visualizing data. By making use of the appropriate visualization, we will be able to understand better the data collected for a single variable, or how two (or more) variables relate to each other. As variables have different types, we will need different methods to visualize such variables.

In this section, we will be illustrating the techniques using a dataset of 1035 records of heigts and weights of MLB players obtained from the SOCR. For illustration, we have included the first 10 rows of the dataset in table 1.4. We note that the team and position variables are categorical, whereas the height, weight and age variables are numeric.

1.4.1 Visualizing one categorical variable

A single categorical variable can be summarised by simply the counts (or proportions) of each of its categories. Suppose in our example that we wish to understand if some positions are more represented than others in the dataset. The usual method for displaying such results is the *bar chart*, illustrated in fig. 1.2. The bar chart aggregates each categories by the number of responses in the given category, and plots each category side by side.

What not to do! Another unfortunately common visualization for such types of data is the infamous *pie chart*, as in fig. 1.3. However, it suffers from many problems, due to the fact that humans have difficulty comparing areas. For example, looking at fig. 1.3, it is difficult to compare the relative size of second basemen and shortstops, and a question such as whether there are twice as many relief pitchers as outfielders is extremely difficult to answer. Whenever you feel a pie chart would be an adequate visualization of the data, a bar chart is nearly always more appropriate.

1.4.2 Visualizing one numeric variable

Histograms Suppose we are now interested in visualizing the heights of players in the MLB. A common technique to visualize one numeric variable is the *histogram*, as seen in fig. 1.4. Such a graphic presents the count (or sometimes the proportion) of players having the height in the given bin. From the histogram, it is for example easy to see that nearly all players have a height between 70 and 80 inches. However, note that histograms can be very sensitive to the bin width, especially if the data is discrete. It is often a good idea to try a few different widths.

Table 1.4: First 10 row of SOCR MLB player data set

Name	Team	Position	Height (in)	Weight (lbs)	Age (yr)
Adam Donachie	BAL	Catcher	74	180	22.99
Paul Bako	BAL	Catcher	74	215	34.69
Ramon Hernandez	BAL	Catcher	72	210	30.78
Kevin Millar	BAL	First Baseman	72	210	35.43
Chris Gomez	BAL	First Baseman	73	188	35.71
Brian Roberts	BAL	Second Baseman	69	176	29.39
Miguel Tejada	BAL	Shortstop	69	209	30.77
Melvin Mora	BAL	Third Baseman	71	200	35.07
Aubrey Huff	BAL	Third Baseman	76	231	30.19
Adam Stern	BAL	Outfielder	71	180	27.05

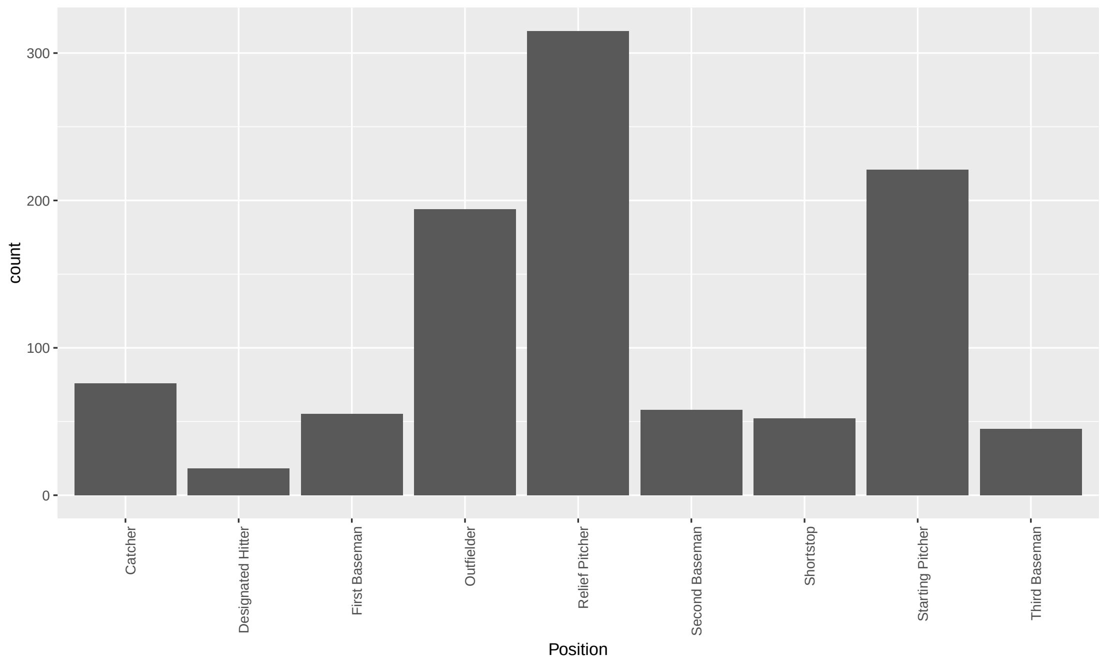

Figure 1.2: Bar chart of position

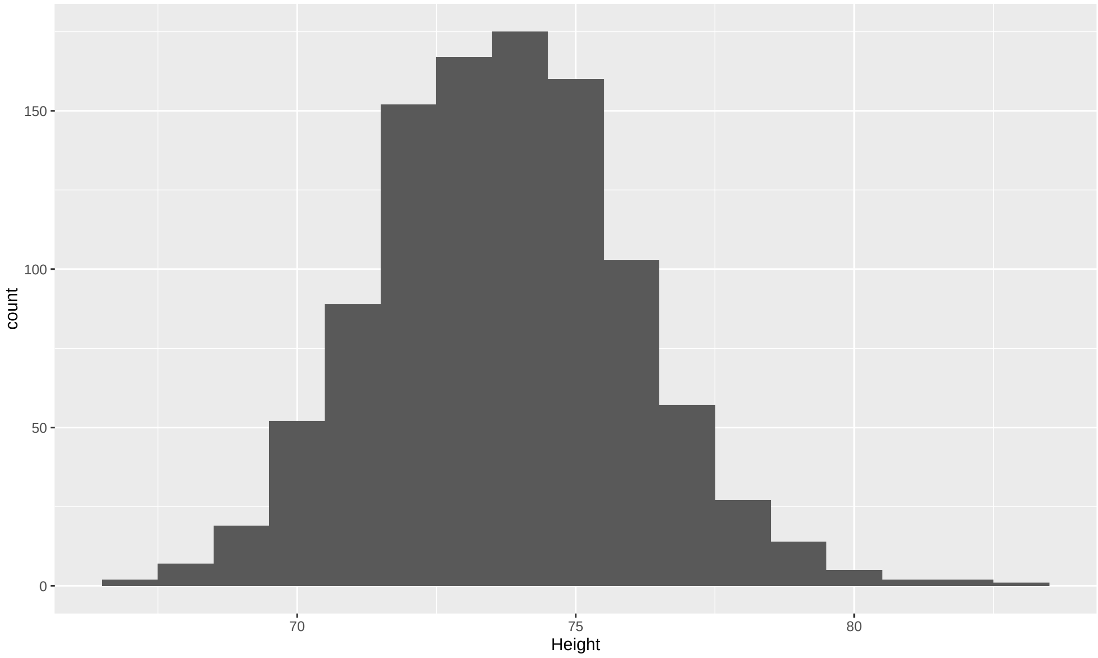

Figure 1.4: Histogram of heights

Density estimates In addition to the sensitivity to the bin width, histograms have an unfortunate characteristic of being quite sensitive to the placement of the bin edges. Kernel density estimators (often abbreviated *kde*) are an alternative way of displaying the same data without defining specific breakpoints. They can also be seen as estimators for a distribution's density, which we will discuss in the probability section. An example of a kde is given in fig. 1.5. KDEs feature a parameter similar to the bin width of a histogram, usually called the *bandwidth*.

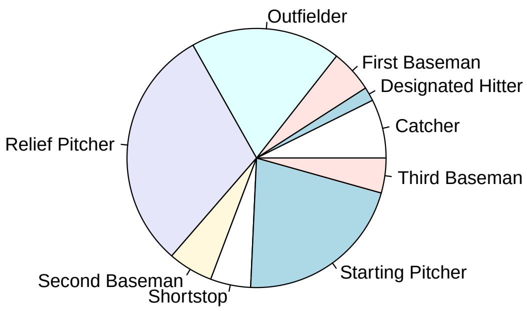

Figure 1.3: Pie chart of position

1.4.3 Visualizing two numeric variables

In this section and the next, we will be interested in visualizing how two variables relate. First, suppose that we wish to consider how the weight of a player is related to its height. In order to do so, we will use the *scatter plot*. The scatter plot (see fig. 1.6) displays one point for each observation (in this case each player), with the coordinate of the point determined by the two variables under consideration (in this case, height and weight).

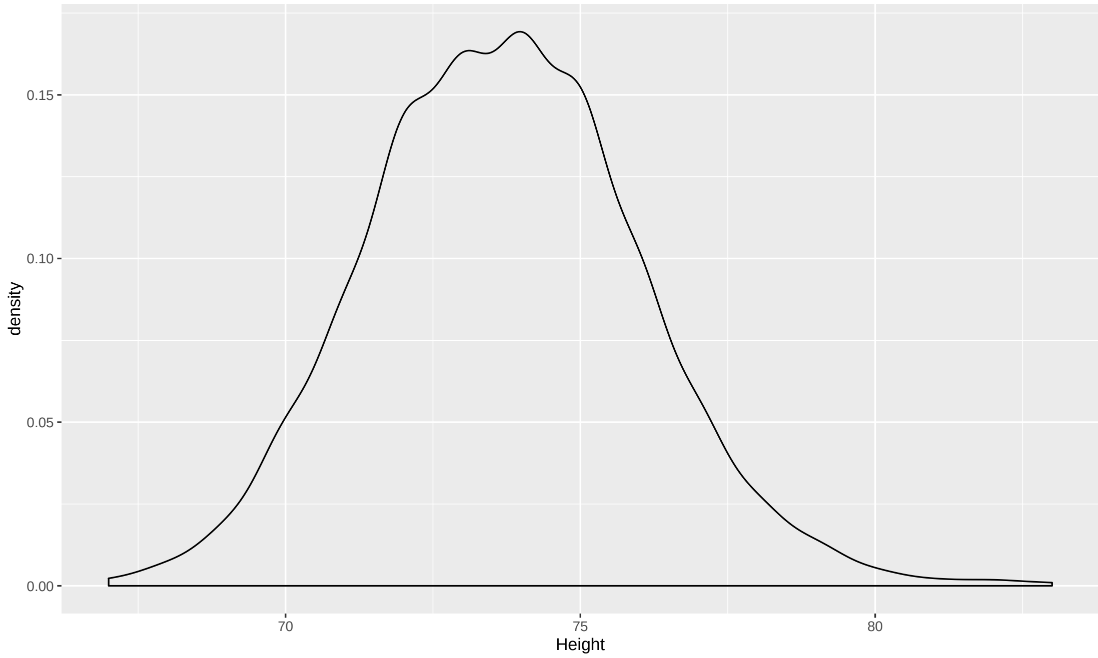

Figure 1.5: Kernel density estimate of heights

In this case, the scatter plot displays clearly the increasing relationship of height with weight.

Scatter plots are particularly adequate to represent numeric variables when the variables themselves are not related to the observations. For example, in this particular case, each observation is a different player, who could (in principle) have any height or weight. However, another common scenario is when one of the variables indexes the observations, for example in the case of time series. In this case, a line plot is a good choice to display trends and patterns in the data across time. For example, fig. 1.7 displays the usage of the Capital bikeshare program in Washington D.C. during the first week of June 2011.

1.4.4 Visualizing one numeric and one categorical variable

Boxplot Returning to the baseball player dataset, suppose now that we wish to understand how the weight of the players differ according to the position. One way to do so would be to group the players according to their position, and then plot some summary statistics for each the groups. The most common such plot is the well-known *boxplot*, which represents the median, quartiles and outlying observations of the data.

The standard boxplot contains three main parts (see fig. 1.8): a middle box with a line, which represents the first and third quartile (with the middle line representing the second quartile or median), whiskers on either side of the box (representing some deviation), and outlying points. In R, the convention is for the whiskers to extend to furthest away whilst still within $1.5 \times \text{IQR}$. Points beyond that distance are then plotted individually.

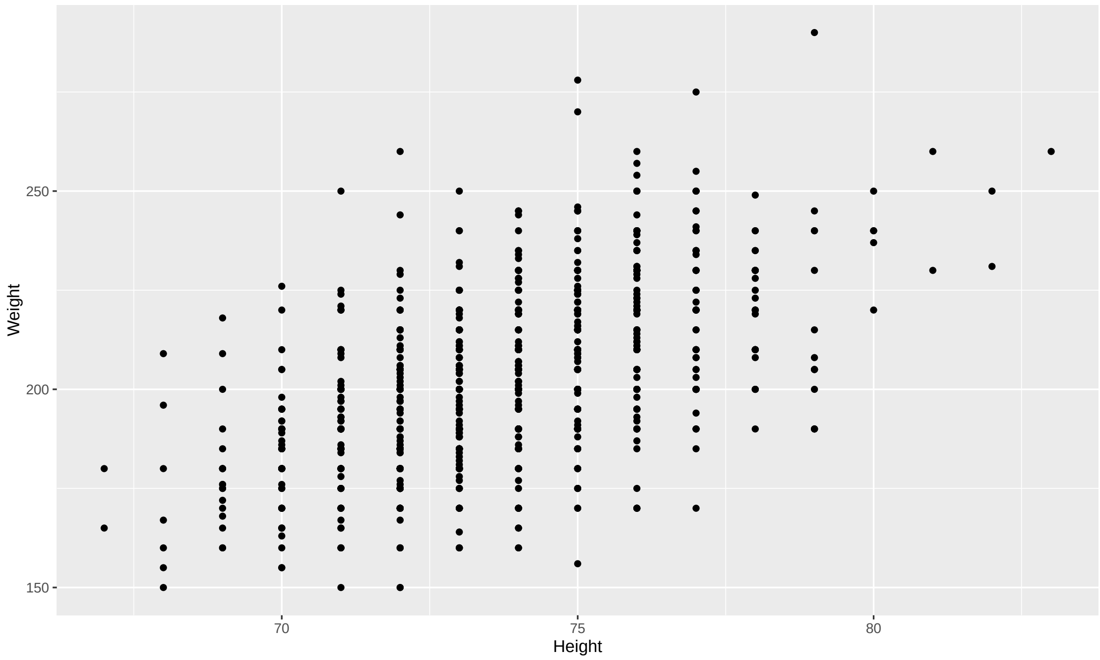

Figure 1.6: Scatter plot of heights and weight

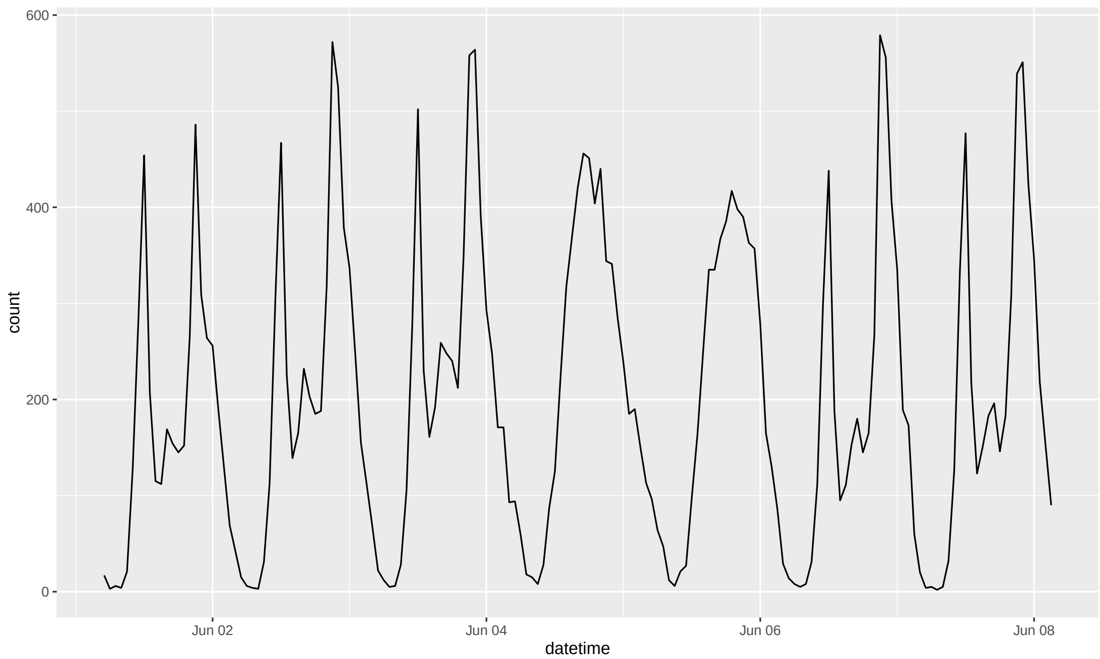

Figure 1.7: Bikeshare program usage

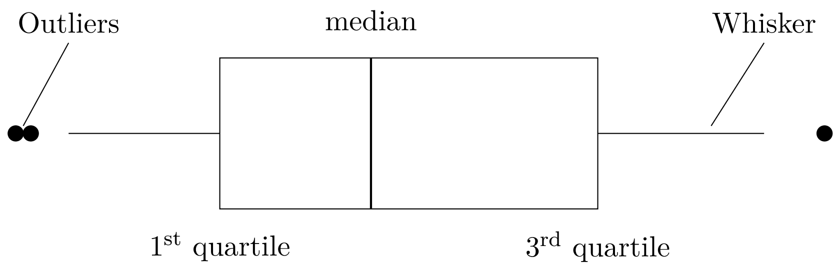

Figure 1.8: Anatomy of a boxplot
.

To visualize the relationship between a numeric and categorical variable, we may then draw one box representing each group, and put them side by side. For the example with the baseball players, see fig. 1.9.

Violin plots An alternative to the boxplot takes inspiration from the kernel density plot (see section 1.4.2) by attempting to draw one kernel density estimate for each category and arrange them vertically. This can be particularly useful when one suspects that the data is multi-modal and cannot be adequately described by its quartiles. On the other hand, the plot is significantly more complex, and computing reasonable density estimates requires more data points, hence boxplots may be preferable when each of the category is small.

1.4.5 Visualizing more than two variables

It is often more difficult to visualize the relationship between more than two variables as our visual system is particularly suited to recognise patterns in 2 dimensions. The main ideas to visualize more than two variables rely on either super-imposing several plots, or putting them side by side.

Faceting The idea of *faceting* plots is quite similar to that of conditional contingency tables. Instead of visualizing three variables simultanously, we choose two variables and produce a plot using any of the techniques described above. However, instead of using all the observations in the plot, we instead split the observations according to the value of the third variable, and display those plots side by side.

For examples, in fig. 1.11, we have displayed side by side a boxplot for the number of users of the bike share program by hour of the day (where we are treating hour as a categorical variable). The plot is split between working days and non-working days, allowing us to easily visualize the different types of usage.

Other characteristics An alternative to creating several plots is to superimpose existing plots and represent the third variable by using a characteristic other than the position

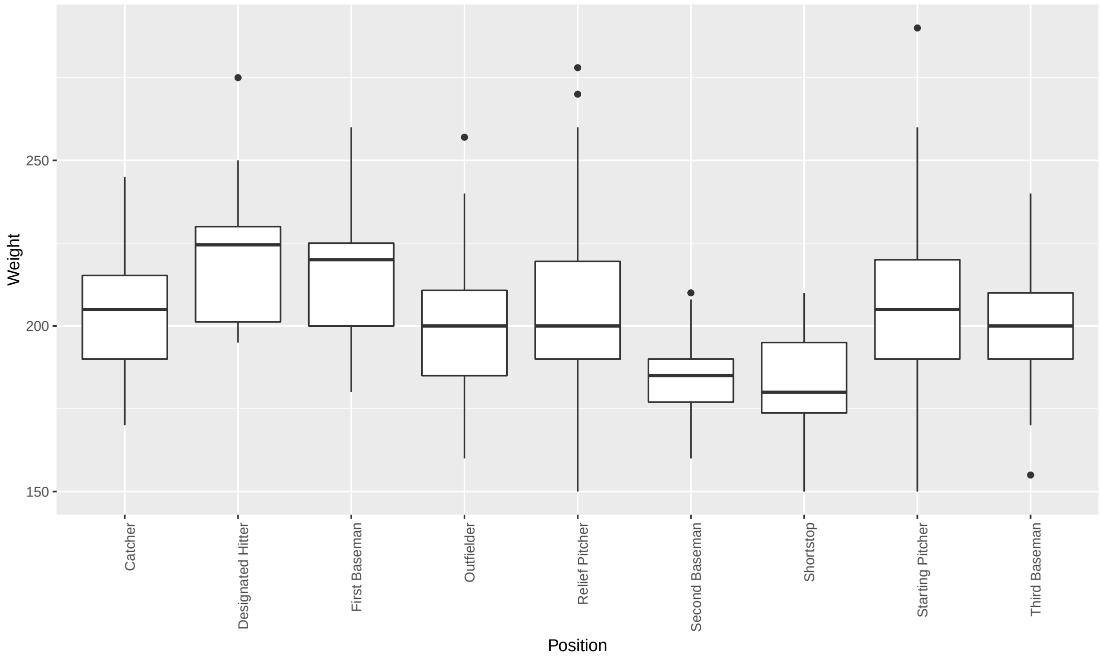

Figure 1.9: Boxplot of weight by position

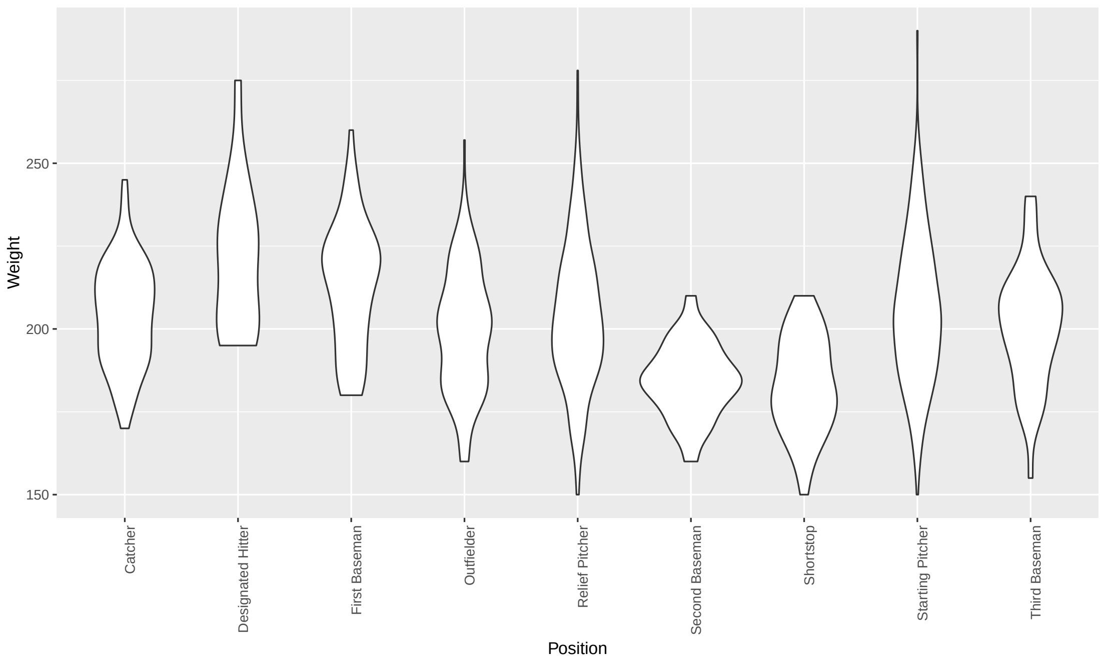

Figure 1.10: Violin plot of weight by position

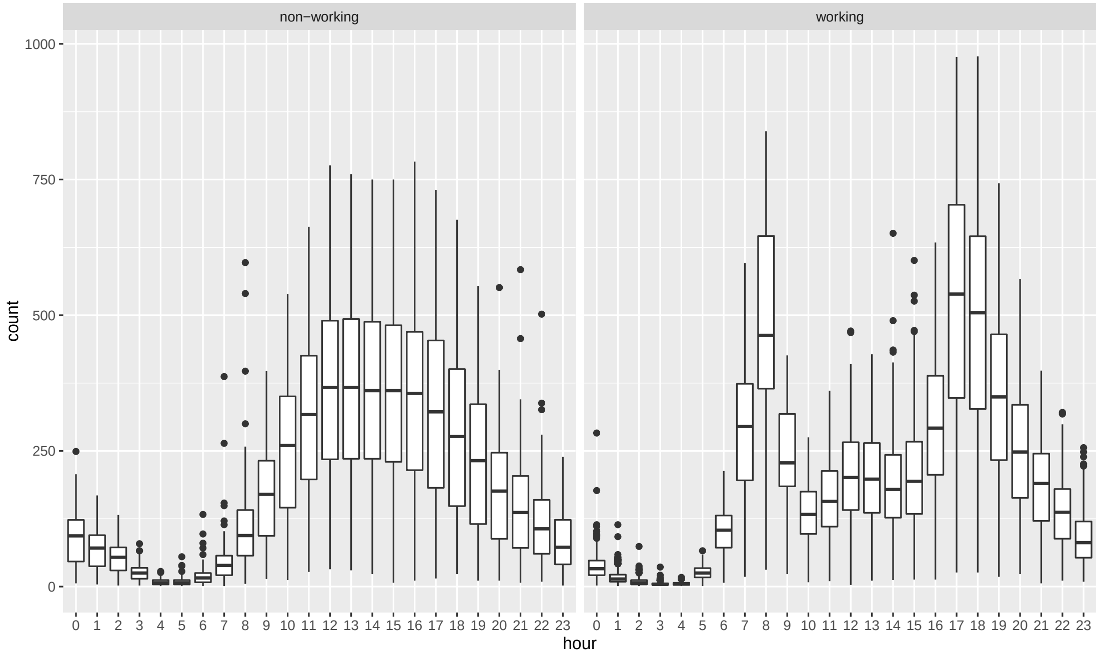

Figure 1.11: Usage of the bike sharing program by hour of day for working days and non-working days

of the graphic. For example, we may try to use a different colour between working and non-working days.

Different types of graphics will have different secondary characteristics that can be leveraged to represent further variables beyond the first two. For example, in addition to a points x and y coordinate, points in a scatter plot may also have different sizes and colours. In a line plot, we may use different colours and line types. In a bar chart, we may use different colours and fill patterns.

1.4.6 Visualizing other types of data

As we mentioned in section 1.2.2, there are numerous other types of data that are not easily represented in a simple rectangular format. Subsequently, these types of data are also difficult to visualize using the standard tools described above, and may require specially developed tools. For example, the problem of graph visualization is a rich field of its own right, with many existing tools. Similarly, the scientific communities have developed numerous visualization techniques adapted to the problems they are faced with.

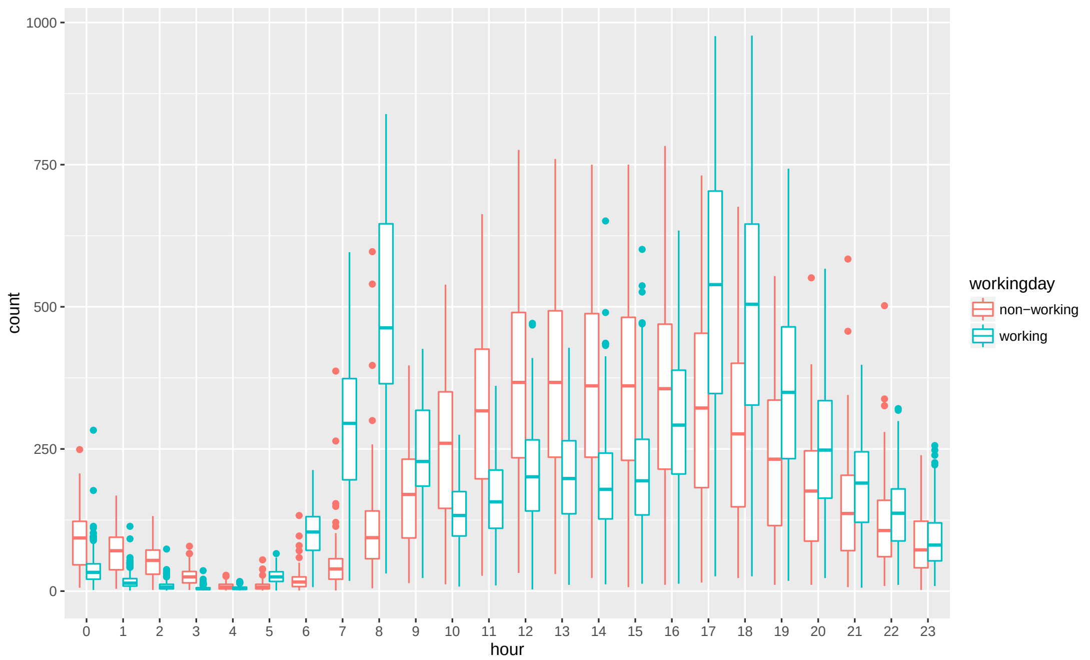

Figure 1.12: Usage of the bike sharing program by hour of day for working days and non-working days

2 Probability

Probability is the language that enables us to discuss uncertainty in a mathematical fashion. The modern theory of probability is axiomatized by the so-called Kolmogorov axioms of probability, which attempt to encode our intuitive notion of probability.

2.1 Probability axioms

2.1.1 Sample space and events

The basic element of probability theory is an event. Philosophically, an event is a contingent proposition, that is, an assertion that may or may not be true. For example, one could consider events: “the coin lands on head”, “the die lands on an even number”, “it will rain tomorrow”. The probability axioms then describe the rules any quantification of uncertainty (i.e. probability) of events must obey.

It is mathematically convenient to use the language of set theory to describe events in the following fashion. We will use the letter Ω to refer to the set (or “collection”) of *all* possible outcomes. We call Ω the *sample space*.

Example 3 Sample space for a coin flip. There are only two possible outcomes for a coin flip: heads (write H) and tails (write T). We may then collect all these outcomes into a set $\Omega = \{H, T\}$, the set with the two elements H and T.

Sample space for a die roll. For a standard 6-sided die, there are 6 possible outcomes corresponding to the numbers 1 to 6. Hence we may represent the sample space as the set $\Omega = \{1, 2, 3, 4, 5, 6\}$.

Sample space for the weather. The sample space for the weather is somewhat too large to describe as explicitly as we have for the previous two examples. However, we will see that our inability to formulate it in a precise fashion does not impede our ability to talk about uncertainty in quantitative way.

Given a sample space, we may then consider events in that space. In the case of a coin flip, an event could be that the coin comes up heads. In the case of a die roll, an event could be that the die lands on 3. However, an event could also be that the die lands on a 1 or a 3. This suggests that an event is a collection of outcomes – or more mathematically, a *subset* of the sample space. We will thus write $E \subseteq \Omega$ for an event E.

2.2 Calculus of probability

In order to quantify uncertainty of events, we will assign to each event E a number between 0 and 1, the probability of the event, and write $\mathrm{P}(E)$ for the probability of E.

One possible interpretation for this number is the long run proportion of events that are true: for example, we say that the probability of a fair coin landing on heads is 0.5 as if we would repeat the toin coss "infinitely" many times, half of them would land on heads – this is the so-called *frequentist* interpretation of probability. Another interpretation of this number is that it represents one's personal beliefs about the likelihood of success of a particular event – for example, we would say that the probability of a fair coin landing on heads is 0.5 as we would be willing to take a 1 for 1 bet on the coin landing on heads. This is the so-called subjective or *Bayesian* interpretation of probability. Both points of view will be useful for the statistician, but we note that the calculus of probability – that is, the rules we use to manipulate probabilities, remain the same.

In addition to assigning a number between 0 and 1 to each event, we will have some rules that relates the probability of related event – for example, we would like to codify the intuition that no matter which die we use, the event "the die lands on an even number" is more likely than the event "the die lands on the number 2 exactly" as the former includes the latter.

First, let us mention two special events, $\varnothing$ and Ω. Ω is simply the set of all possible outcomes, and we have that $\mathrm{P}(\Omega) = 1$, which we may interpret as the fact that of all the possible outcomes, one must happen. Its counterpart, $\varnothing$, is the empty set of outcomes, and we have that $\mathrm{P}(\varnothing) = 0$. We may also interpret these sets as events, in which case Ω is the event that always happens, whereas $\varnothing$ is the event that nothing happens.

2.2.1 Disjoint or mutually exclusive events

We now come to the rules of the calculus of probability, which relates the probability of related events. Consider two events A and B, what can we say about the probability of $A \cup B$: the event that either A or B (or both) happen? In general, this is a difficult question, as A and B may overlap a little (or a lot). However, in the special case that A and B do not overlap, that is, it is impossible that both A and B happen, we have the simple rule of additivity:

$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) \tag{2.1}$$

We say in this case that A and B are *disjoint* or *mutually exclusive*. We may also express the condition that A and B are mutually exclusive by the following mathematical assertion: $A \cap B = \varnothing$, which translates that there is no outcome where A and B both happen.

2.2.2 Complement of an event

Given an event A, we can consider the event that A does not happen. This event is called the *complement* of an event, and is written A^C. What can we say about the probability of A^C? Well, first, note that either A or A^C happens (that is, something either happens or does not happen).

Mathematically, we may codify this statement into the following equality:

$$A \cup A^C = \Omega, \tag{2.2}$$

which states that A, together with its complement A^C, are equal to the entire sample space.

Now, by definition, we also have, that A and A^C are disjoint (specifically, A^C was defined in this way). We may codify this statement mathematically as

$$A \cap A^C = \varnothing. \tag{2.3}$$

Now, this means that we may apply our previous rule for additivity of disjoint events, and write $P(A \cup A^C) = P(A) + P(A^C)$. However, by using eq. (2.2), we see that $P(A \cup A^C) = 1$. Hence, we deduce the following relationship between the probability of A and its complement:

$$\mathrm{P}(A) + \mathrm{P}(A^C) = 1, \tag{2.4}$$

which can also be written as $\mathrm{P}(A^C) = 1 - \mathrm{P}(A)$.

2.2.3 Inclusion-Exclusion

In some cases, we may be interested in calculating $A \cap B$ despite A and B not being disjoint. In that case, the additive formula does not directly apply. However, we may apply it with a slight modification, giving the so-called *inclusion-exclusion* formula.

To understand the inclusion-exclusion formula, it is important to understand why it is not true that $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B)$. By thinking for example about the case where $A = B$, we may see that it is due to the fact that we are counting the events that fall in both A and B twice on the right hand side.

In order to adjust for this, the inclusion formula has a correction on the right hand side, giving the formula:

$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B). \tag{2.5}$$

Example 4 Suppose that we still consider a die roll, and consider the events $A = \{1, 2\}$, and $B = \{2, 3\}$. We have that $\mathrm{P}(A) = \mathrm{P}(B) = 1/3$, and note that $\mathrm{P}(A \cup B) = 1/2$.

Now, the additive formula would claim $\mathrm{P}(A \cup B) = 2/3$, which is clearly wrong. Let's use the inclusion-exclusion formula instead. To do so, we need to compute $\mathrm{P}(A \cap B) = \mathrm{P}(\{2\}) = 1/6$, to finally obtain:

$$\mathrm{P}(A \cup B) = \frac{1}{3} + \frac{1}{3} - \frac{1}{6} = \frac{1}{2}, \tag{2.6}$$

as we expected.

In this case, we can see that without the correction term, we were counting the probability of the roll being 2 twice, making the calculation incorrect.

2.2.4 Independent events and multiplicative rule

Given an event A and an event B, we will often be interested in a very special relationship between the two events, called *independence*. Conceptually, independence denotes the fact that one event does not inform about the other. For example, consider a die roll, and the following two events $A = \{2, 4, 6\}$ (i.e. the result is even), and the event $B = \{1, 2\}$ (i.e. the roll is 1 or 2). Then knowing that the result was 1 or 2 does not give us any information about whether the result was even or odd.

Mathematically, we say that A and B are independent if we have that

$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B). \tag{2.7}$$

It is not immediately obvious how this mathematical statement relates to the fact that the event A does not given us any information about event B and vice versa, but we will explore this a bit more when defining conditional probabilities. If we apply the example to the A and B given above, we have that $\mathrm{P}(A) = 1/2$, $\mathrm{P}(B) = 1/3$, and $A \cap B = \{2\}$ thus $\mathrm{P}(A \cap B) = 1/6$, and everything works as expected.

We will be revisiting independence later for random variables, as it is one of the most important concepts in probability and underpins much of statistics.

2.2.5 Conditional probability

Conditional probability allows us to reason about random events about which we have received partial information. For example, suppose again that we are throwing a die, and suppose that we would like to know the probability of the roll being 4 or higher (i.e. 4, 5, 6). A priori, this happens half of the time. However, suppose now that we roll the die, and do not disclose the result, but only that the roll was odd. How should this change the probability of the roll being 4 or higher? Intuitively, this seems to make our chance worse, as there is only one odd number in 4, 5, 6, but two even numbers.

Let us then define the condition probability of an event A (in this case $A = \{4, 5, 6\}$) given event B (in this case $B = \{1, 3, 5\}$, the event that the roll was odd). Write the conditional probability $\mathrm{P}(A \mid B)$ (read “probability of A given B”), and define:

$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}. \tag{2.8}$$

Note that as we are dividing by the probability of B, we require B to have positive probability (i.e. $\mathrm{P}(B) = 0$).

Example 5 Let’s compute out our example in detail. We have that $A \cap B = \{5\}$, so that $\mathrm{P}(A \cap B) = 1/6$. We also have that $\mathrm{P}(B) = 1/2$. Hence we deduce that:

$$\mathrm{P}(A \mid B) = \frac{1/6}{1/2} = 1/3. \tag{2.9}$$

Note that this is indeed lower than the probability of A, which was initially $\mathrm{P}(A) = 1/2$.

2.2.6 Conditional probability and independence

As conditional probabilities allow us to reason about how probability change when we acquire knowledge about another event, let’s try to use them and quantify our intuition about independence. Our intuition about two events A and B being independent corresponded to the fact that knowing about one event did not affect the other. In particular, the probability of A should be the same as that of A given B. We will prove this fact.

From the definition of conditional probability (eq. (2.8)), we have that:

$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}. \tag{2.10}$$

Now, we have in addition by the definition of independence (eq. (2.7) that $\mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B)$. Substituting this in the equation above, we have that:

$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} = \frac{\mathrm{P}(A)\,\mathrm{P}(B)}{\mathrm{P}(B)} = \mathrm{P}(A), \tag{2.11}$$

where we have cancelled $\mathrm{P}(B)$ in the numerator and denominator.

2.2.7 Conditional probability and additivity

Conditional probabilities verify most of the properties of regular probabilities, and in particular the law of additivity for disjoint events and the inclusion exclusion rule. We prove the law of additivity of disjoint event for conditional probabilities.

Let A and B be disjoint events, that is, $A \cap B = \varnothing$. Let C be any event with $\mathrm{P}(C) > 0$. We would like to show that

$$\mathrm{P}(A \cup B \mid C) = \mathrm{P}(A \mid C) + \mathrm{P}(B \mid C). \tag{2.12}$$

We substitute the definition of conditional probabilities, and have to show that

$$\frac{\mathrm{P}((A \cup B) \cap C)}{\mathrm{P}(C)} = \frac{\mathrm{P}(A \cap C)}{\mathrm{P}(C)} + \frac{\mathrm{P}(B \cap C)}{\mathrm{P}(C)}. \tag{2.13}$$

Cancel the denominators to obtain

$$\mathrm{P}((A \cup B) \cap C) = \mathrm{P}(A \cap C) + \mathrm{P}(B \cap C). \tag{2.14}$$

Now, we would like to simplify the expression $(A \cup B) \cap C$. The following formulae, known as distributivity property, gives us the correct expressions:

Lemma 1 (Distributivity) *For any events A, B, C, we have that*

$$(A \cup B) \cap C = (A \cap C) \cup (A \cap B), \tag{2.15}$$
$$(A \cap B) \cup C = (A \cup C) \cap (A \cup B). \tag{2.16}$$

Hence we may simplify $(A \cup B) \cap C = (A \cap C) \cup (A \cap B)$. Now, we would like to apply the additivity rule. However, we first need to ensure that those two sets are disjoints, i.e. have empty intersection. We can check that

$$(A \cap C) \cap (A \cap B) = A \cap C \cap A \cap B = (A \cap B) \cap C = \varnothing \cap C = \varnothing, \tag{2.17}$$

hence we may apply the additivity rule, which gives us

$$\mathrm{P}((A \cap C) \cup (A \cap B)) = \mathrm{P}(A \cap C) + \mathrm{P}(A \cap B). \tag{2.18}$$

2.2.8 Law of total probability

In some cases, we may want to compute the probability of an event where we only know conditional probabilities of that event. For example, suppose that we wish to compute $\mathrm{P}(A)$ knowing $\mathrm{P}(A \mid B)$ and $\mathrm{P}(A \mid B^C)$. The *law of total probability* gives a formula that allows us to do so.

$$\mathrm{P}(A) = \mathrm{P}(A \mid B)\,\mathrm{P}(B) + \mathrm{P}(A \mid B^C)\,\mathrm{P}(B) \tag{2.19}$$

Intuitively, we know that either B or its complement B^C happen. So understanding what happens to A under both cases should allow us to reconstruct what happens to A, by weighting by the probability that each scenario happens.

Let us prove the formula. Note that we have by definition of the conditional probability that $\mathrm{P}(A \mid B)\,\mathrm{P}(B) = \mathrm{P}(A \cap B)$, and similarly $\mathrm{P}(A \mid B^C)\,\mathrm{P}(B^C) = \mathrm{P}(A \cap B^C)$. Now, we have again that $A \cap B$ and $A \cap B^C$ are disjoint, hence we may apply the additivity rule to obtain that the right hand side is equal to $\mathrm{P}((A \cap B) \cup (A \cap B^C))$. Now applying the distributivity rule in reverse, this is equal to $\mathrm{P}(A \cap (B \cup B^C)) = \mathrm{P}(A \cap \Omega) = \mathrm{P}(A)$, as required.

2.2.9 Conditional probabilities and bayes rule

Given two events A and B, we will often be interested in relating $\mathrm{P}(A \mid B)$ and $\mathrm{P}(B \mid A)$. For example, suppose that we are testing a medical diagnosis to screen for a disease, call it disease Z. We are given the following facts about the test: for a person that has disease Z, the test will return positive 99% of the time (in other words, we have a false negative rate of 1%). For a person that does not have disease Z, the test will return positive 1% of the time (in other words, we have a false positive rate of 1%).

Suppose that we administer the test to a patient that just came to the hospital, and the result is positive. What is the probability that the patient has the disease? A tempting answer is simply that the probability is 99%, after all, the test is only wrong 1% of the time. However, we will see that this answer is in fact quite wrong.

Let's set this problem up in a mathematical fashion, and let T be the event that the test is positive (so that T^C is the event that the test is negative), and let D be the event that the person has disease Z (and hence D^C the event that the person is healthy). Putting the given assertions about the test formally, we have that

$$\begin{aligned} \mathrm{P}(T \mid D) &= 0.99, \\ \mathrm{P}(T \mid D^C) &= 0.01. \end{aligned}$$

We are looking to answer the question: what is the probability that someone who tested positive has disease Z. Formally, we would like to compute $\mathrm{P}(D \mid T)$.

Baye's rule gives a formula to compute this quantity, by

$$\mathrm{P}(D \mid T) = \frac{\mathrm{P}(T \mid D)\,\mathrm{P}(D)}{\mathrm{P}(T)}. \tag{2.20}$$

Let's apply the formula to the example first. We are given $\mathrm{P}(T \mid D)$ from the problem statemente, but we are neither given $\mathrm{P}(D)$ nor $\mathrm{P}(T)$. Hence, suppose that $\mathrm{P}(D) = 0.01$, or that one in a hundred person has this disease Z. We may know compute $\mathrm{P}(T)$ according to the law of total probability (eq. (2.19)) to obtain

$$\begin{aligned} \mathrm{P}(T) &= \mathrm{P}(T \mid D)\,\mathrm{P}(D) + \mathrm{P}(T \mid D^C)\,\mathrm{P}(D^C) \\ &= 0.99 \times 0.01 + 0.01 \times 0.99 \\ &= 0.198 \end{aligned}$$

Plugging this quantity into the bayes rule (eq. (2.20)), we obtain that

$$\mathrm{P}(D \mid T) = \frac{0.99 \times 0.01}{0.198} = 0.5 \tag{2.21}$$

Hence a person testing positive only has about a 50-50 chance of actually having disease Z!

Let's finish by proving the Bayes rule. By definition of the conditional distribution, we have that $\mathrm{P}(T \mid D) = \mathrm{P}(T \cap D) / \mathrm{P}(D)$. Hence we have that the right hand side is:

$$\frac{\mathrm{P}(T \cap D)\,\mathrm{P}(D)}{\mathrm{P}(D)\,\mathrm{P}(T)} = \frac{\mathrm{P}(T \cap D)}{\mathrm{P}(T)} = \mathrm{P}(D \mid T). \tag{2.22}$$

2.3 Random variables

Random variables are the main probability tools we will be using to discuss data. They enable us to apply the notions of probability we learned on numeric data often encountered in statistics. Random variables represent numerical outcomes that are random or unknown. We will be interested in how to characterise the randomness of these numerical quantities.

In statistics, we will often use randomness to model processes that may not necessarily be inherently random, but are too complex to understand fully. For example, consider the problem of forecasting weather. One might argue that if we had an extremely powerful computer, and as many sensors as we desire, it would be possible to exactly predict the weather. However, within our knowledge and technology today, this is not possible. Hence we model the outcome as random. This is an example of *epistemic* randomness. Another example of such randomness is the coin toss: one could argue that with extremely precise sensors, and using Newton's law of motions, one could predict exactly how the coin lands. On the other hand, we believe that some processes a inherently random, which we refer to as *ontic* randomness. For example, the radioactive decay of uranium is understood to be inherently random according to the current theory of quantum mechanics.

We will be using the formalism of probability and random variables to model both the inherent ontic randomness, and (most often) epistemic randomness, whether due to incomplete information or complexity of the process. The calculus of probability is the same whether the randomness is ontic or epistemic.

2.3.1 What is a random variable?

A random variable is a numerical variable whose value is random. Random variables can be used to model the outcome of experiments or other phenomenon that have some randomness, and will usually be denote X, Y, Z, etc. For example, the value of the FTSE 100 Index tomrrow can be viewed as a random variable, as it is unknown today and has some randomness. Similarly, the temperature tomorrow could also be modelled as a random variable. Random variables will often also be appropriate to model information that is known, but merely hidden from us. For example, one could model a patient's blood pressure as a random variable. Although that information is not random – indeed we could measure it – it is unknown to us, and falls under epistemic randomness.

2.3.2 Random variables and events

We would like to tie the notion of random variables back to the theory of probability we have developed in the previous section, which was centered around the notion of events. How do we obtain events from random variables? As a random variable is a random quantity, any assertion concerning that variable is an event.

Example 6 Weather forecasting Let X be a random variable representing the temperature tomorrow in degrees Celsius. Then, $X < 0$ is an event, and corresponds to the colloquial event that the temperature will be below freezing. In particular, we may form the question "what is the probability that the temperature will be below freezing tomorrow?". That probability would be written $\mathrm{P}(X < 0)$.

In principle, knowing the probability of all the events concerning a random variable would completely characterise the randomness of that quantity. However, there are many (usually infinitely many) such events, and describing them can be difficult. Instead, we will see that we can usually characterise the randomness completely with much less information.

2.3.3 Discrete random variables

We say a random variable is discrete if it takes finitely many (or at most countably many values). Most often, a discrete random variable will take integer values. For example, the number of heads in 10 coin tosses is a discrete random variable taking integer values between 0 and 10. Another example is the number of customers that visited a store in a given day.

Discrete random variables can be characterised completely by the so-called probability mass function, which describes the probability of each possible value of a discrete random variable. We define the p.m.f. of X to be the function f_X, where

$$f_X(x) = \mathrm{P}(X = x). \tag{2.23}$$

For example, if X is the number of heads in 10 coin tosses, then $f_X(1)$ is the probability that we obtain *exactly* one head in 10 tosses.

Example 7 Two coin tosses Let us compute the p.m.f. for X, where X is the number of heads in two coin tosses. For the case of two coins, there are four possible outcomes, being HH, HT, TH, TT, each equally likely. Hence we have that

$$\begin{aligned} f_X(2) &= \mathrm{P}(X=2) &&= \frac{1}{4}, \\ f_X(1) &= \mathrm{P}(X=1) &&= \frac{1}{2}, \\ f_X(0) &= \mathrm{P}(X=0) &&= \frac{1}{4}. \end{aligned}$$

The p.m.f. of a random variable allows us to compute the probability of all events concerning that random variable. For example, for the case of two tosses, consider the event $X \geq 1$ (at least one of the toss lands on heads). Then, note that $X \geq 1$ is equivalent to saying $X = 1$ or $X = 2$, and these two events are disjoint. We may thus apply the additivity rule to obtain

$$\mathrm{P}(X \geq 1) = \mathrm{P}(X=1) + \mathrm{P}(X=2) = f_X(1) + f_X(2) = \frac{3}{4}. \tag{2.24}$$

In this sense, to define a discrete random variable, it suffices to give its p.m.f.

As the p.m.f. encodes probabilities, it inherits some properties of probabalities. In particular, we have that $0 \leq f_X(x) \leq 1$ for all possible values of x. Furthermore, we have that:

$$\sum_x f_X(x) = 1, \tag{2.25}$$

where the sum is taken over all possible values of X. This encodes the fact that X has to take some value, and corresponds to the fact that $\mathrm{P}(\Omega) = 1$. We say that the p.m.f. is *normalized*.

2.3.4 Expectation

In addition to understanding the likelihood of certain outcomes, we will often be interested in capturing the average value of a random variable, also called its expectation. We define the expectation of a discrete random variable X with p.m.f. f_X to be:

$$\mathbb{E}\, X = \sum_x x\, \mathrm{P}(X = x) = \sum_x x f_X(x) \tag{2.26}$$

where the sum is over all possible values of X.

Example 8 Two coin tosses Let X be the number of heads in two coin tosses, and recall the p.m.f. of X from the previous section. We would like to compute $\mathbb{E}\, X$. X can take values 0, 1 or 2, hence plugging the values into the formula yields:

$$\mathbb{E}\, X = 0 \times \frac{1}{4} + 1 \times \frac{1}{2} + 2 \times \frac{1}{4} = 1. \tag{2.27}$$

Thus the average number of heads in two coin tosses is 1, as expected.

Now, in addition to expectation of random variables, we will also be interested in computing expectations of functions of random variables. We define the expectation of a function $g(X)$ as follows:

$$\mathbb{E}\, g(X) = \sum_x g(x)\, \mathrm{P}(X = x) = \sum_x g(x) f_X(x). \tag{2.28}$$

Example 9 Two coin tosses Let X be the number of heads in two coin tosses, and recall the p.m.f. of X from the previous section. Let $g(x) = x^2$, and suppose that we would like to compute $\mathbb{E}\, g(X)$. We write

$$\mathbb{E}\, g(X) = \mathbb{E}\, X^2 = 0^2 \times \frac{1}{4} + 1^2 \times \frac{1}{2} + 2^2 \times \frac{1}{4} = \frac{3}{2}. \tag{2.29}$$

Note that $\mathbb{E}\, X^2$ is not equal to $(\mathbb{E}\, X)^2$.

2.3.5 Population statistics

The notion of expectations allows us to define population statistics of a random variable. For example, we may define the mean, or the variance, of a random variable. To do so, we will often replace the average that appears in the statistic (i.e. the $n^{-1} \sum$) by the expectation sign.

For example, we have defined the sample mean as

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} X_i, \tag{2.30}$$

hence we replace the sum sign to obtain that the sample mean (also called expectation) should simply be

$$\mathbb{E}\, X. \tag{2.31}$$

Similarly, we have define the sample variance as

$$\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2. \tag{2.32}$$

Now, we replace the sample mean that appears above, and the sum sign, to obtain that the variance of a random variable is defined as

$$\sigma^2 = \mathbb{E}(X - \mathbb{E}\, X)^2 \tag{2.33}$$

Note that the population mean and variance are no longer random variables, but simple numbers instead.

Example 10 Two coin tosses Let X be the number of heads in two coin tosses. We recall that we had computed $\mathbb{E}\, X = 1$. Let's compute the variance of X.

$$\begin{aligned}\sigma^2 &= \mathbb{E}(X - \mathbb{E}\,X)^2 \\ &= \mathbb{E}(X-1)^2 \\ &= (0-1)^2 \times \frac{1}{4} + (1-1)^2 \times \frac{1}{2} + (2-1)^2 \times \frac{1}{4} \\ &= \frac{1}{2}.\end{aligned}$$

Hence the variance of the number of heads in two coin tosses is $\sigma^2 = 1/2$.

2.3.6 Continuous random variables

As opposed to discrete random variables, *continuous* random variables take values over an interval of the real line. For example, a continuous random variable is an appropriate model for the number of seconds (and fractions of a second) between two radioactive decay events of a sample of uranium. Another example of a continuous random variable might be the temperature tomorrow.

Although continuous random variables behave similarly to discrete random variables, and we will be using similar tools to understand both, they can be mathematically more delicate due to the following fact: a continuous random variable is equal to a given number with probability 0! Indeed, if X is a continuous random variable, then we must have that $\mathrm{P}(X = x) = 0$ for all x real. In particular, we won't be able to define a p.m.f. for a continuous random variable, as it would simply be 0 everywhere. However, we can still define a very similar object by making use of calculus.

For continuous random variables, the appropriate object we wish to define is the so-called *density*probability density function, also called *p.d.f.*, which encodes the *infintesimal* probability of the random variable being equal to a given value. For example, let X be the random variable that corresponds to a (uniformly) random number on the interval $[0, 1]$ (i.e. the set of all the real numbers from 0 to 1). We say X follows the uniform distribution on $[0, 1]$. Now, although we have that $\mathrm{P}(X = 0.5) = 0$, as it essentially never happens that the random number is *exactly* 0.5, we can certainly compute

$$\mathrm{P}(0 \le X \le 0.5) = \frac{1}{2}. \tag{2.34}$$

Indeed, as X is uniformly random, it certainly has a 50-50 chance of being in either the first or the second half of the interval.

Now, we can extend this idea to compute for example $\mathrm{P}(0 \le X \le 0.25) = 0.25$, as X has $1/4$ chance of being in the first quarter. In general, we have for any $0 < h < 1$ that:

$$\mathrm{P}(0 \le X \le h) = h. \tag{2.35}$$

Hence we could consider taking a limit

$$\lim_{h \to 0} \frac{1}{h}\,\mathrm{P}(0 \le X \le h) = 1, \tag{2.36}$$

which corresponds to the idea of a density function.

2.3.7 Distribution and density functions

In order to make the previous idea more rigorous and precise, we will take a detour to define the (cumulative) *distribution* function. If X is a random variable (continuous or discrete), we may define the distribution function F_X:

$$F_X(x) = \mathrm{P}(X \leq x). \tag{2.37}$$

For example, we have computed previously that for a uniform random variable on $[0, 1]$, we have

$$F_X(x) = x, \text{ for } 0 \leq x \leq 1. \tag{2.38}$$

The distribution function similarly fully characterises the distribution of a random variable, and allows us to compute probability of events involving the random variable.

Example 11 Let X be as above with the uniform distribution on $[0, 1]$. Suppose we wish to compute $\mathrm{P}(1/4 \leq X \leq 3/4)$. Note that we can see that this is $1/2$, as this interval occupies half of the space. Let us also obtain the result using the distribution function.

Now, note that $X \leq 3/4$ is the same as $X \leq 1/4$, or $1/4 \leq X \leq 3/4$, and the latter two are disjoint. Hence we may apply the additivity rule to obtain

$$\mathrm{P}(X \leq 3/4) = \mathrm{P}(X \leq 1/4) + \mathrm{P}(1/4 \leq X \leq 3/4), \tag{2.39}$$

which we may re-arrange to obtain:

$$\mathrm{P}(1/4 \leq X \leq 3/4) = \mathrm{P}(X \leq 3/4) - \mathrm{P}(X \leq 1/4) = F_X(3/4) - F_X(1/4) = 1/2. \tag{2.40}$$

Although the distribution function can perfectly encode the information of the distribution of a random variable, it is sometimes inconvenient, for example, it cannot be directly used to compute expectations.

Let us go back to the problem of defining a density for a continuous random variable. As we have remarked before, asking for X to be exactly equal to some value x always has probability 0. However, suppose instead that we just wished X to be very close, say within $h/2$, where $h > 0$. Then, we would have that

$$\mathrm{P}(x - h/2 \leq X \leq x + h/2) = F_X(x + h/2) - F_X(x - h/2). \tag{2.41}$$

As before, suppose that we normalize by h, and take the limit as $h \to 0$. We would then have that

$$\lim_{h \to 0} \frac{1}{h} \mathrm{P}(x - h/2 \leq X \leq x + h/2) = \lim_{h \to 0} \frac{1}{h}(F_X(x + h/2) - F_X(x - h/2)) = F'_X(x), \tag{2.42}$$

where we have recognised the expression for the derivative of F_X at x on the right hand side. We thus define the (probability) density function f_X of X to be:

$$f_X(x) = F'_X(x). \tag{2.43}$$

Example 12 As above, let X be the uniform random variable on $[0, 1]$. We had computed previously that $F_X(x) = x$, for $0 \leq x \leq 1$. Hence we can compute the density function to be:

$$f_X(x) = \frac{d}{dx}x = 1, \text{ for } 0 \leq x \leq 1. \tag{2.44}$$

We will most often only be given an expression for the density f_X, and not the distribution F_X, as the former tends to be mathematically simpler. How can we use f_X to compute probability of events? Let X be a random variable, and suppose that we only knew f_X but wished to compute $\mathrm{P}(0 \leq X \leq 1)$. From our previous discussion, we know that we have

$$\mathrm{P}(0 \leq X \leq 1) = F_X(1) - F_X(0), \tag{2.45}$$

but we do not have access to F. However, by the fundamental theorem of calculus, we know that

$$\int_0^1 f_X(x)dx = F_X(1) - F_X(0), \tag{2.46}$$

hence we deduce that we have

$$\mathrm{P}(0 \leq X \leq 1) = \int_0^1 f_X(x)dx \tag{2.47}$$

Example 13 As above, let X be the uniform random variable on $[0, 1]$. We had computed $f_X(x) = 1$, so we have that

$$\mathrm{P}(1/4 \leq X \leq 3/4) = \int_{1/4}^{3/4} f_X(x)dx = \frac{1}{2}. \tag{2.48}$$

Finally, as the p.d.f. f_X is related to probabilities, it must obey similar rules. In particular, we have that $f_X(x) \geq 0$ for all x. However, it is not necessary that $f_X(x) \leq 1$, as f_X itself is not a probability. We do have that:

$$\int_{-\infty}^{\infty} f_X(x)dx = 1, \tag{2.49}$$

and say that f_X must be normalized.

2.3.8 Expectation (bis)

We may know define expectations for continuous random variable. In general, the p.d.f. we play the role of the p.m.f. for a continuous random variable, and we will have to replace summation with integration. Thus, if X is a continuous random variable with p.d.f. f_X, we may define its expectation $\mathbb{E}\, X$ as:

$$\mathbb{E}\, X = \int x f_X(x)dx, \tag{2.50}$$

where the integral is over all possible values of X (usually, an interval or the real line).

Example 14 Uniform random variable Let us compute the expectation of a uniform random variable X on $[0,1]$. Recall that we have $f_X(x) = 1$ for $0 \leq x \leq 1$. Hence we have that

$$\mathbb{E}\,X = \int_0^1 x \cdot 1 dx = \frac{1}{2}. \tag{2.51}$$

Note that the bound of integrations correspond to the possible values of x.

We may similarly define the expectation of functions of continuous random variables. Let $g(x)$ be a function, we may then define $\mathbb{E}\,g(X)$ to be

$$\mathbb{E}\,g(X) = \int g(x) f_X(x) dx, \tag{2.52}$$

where again the integral is over all possible value of X.

Example 15 Let us compute the expectation of the function $g(x) = x^2$ of a uniform random variable X on $[0,1]$. Recall that we have $f_X(x) = 1$ for $0 \leq x \leq 1$. Hence we have that:

$$\mathbb{E}\,X^2 = \int_0^1 x^2 \cdot 1 dx = \frac{1}{3}. \tag{2.53}$$

Now that we have the expectation of a continuous random variable, this allows us to define the variance of continuous random variable using the exact same formula as eq. (2.33), namely:

$$\sigma^2 = \mathbb{E}(X - \mathbb{E}\,X)^2 \tag{2.54}$$

Example 16 Let us compute the variance of a uniform random variable X on $[0,1]$. Recall that we have $f_X(x) = 1$ for $0 \leq x \leq 1$, and that $\mathbb{E}\,X = 1/2$. Hence we have that:

$$\sigma^2 = \int_0^1 (x - 1/2)^2 \cdot 1 dx = \frac{1}{12} \tag{2.55}$$

2.4 Common distributions

It will often be useful to have distributions for which we understand the characteristics (such as p.m.f./p.d.f., mean, variance, etc.) well to model different types of variables. In this section, we describe some common distributions, their probabilistic properties, and some of the most common uses.

Most often, we will describe a *family* of distributions, that is, a set of distributions parametrised by some real parameters. These parameters will describe the exact behaviour of the distribution.

2.4.1 Bernoulli distribution

The *Bernoulli* distribution is a discrete distribution that represents a coin flip as a $0-1$ outcome. It is parametrised by p, the probability that the outcome is 1, where $0 \leq p \leq 1$. If X has the Bernoulli distribution with parameter p, we will write $X \sim \text{Bernoulli}(p)$.

The p.m.f. of the bernoulli distribution is given by:

$$\mathrm{P}(X = 0) = 1 - p \text{ and } \mathrm{P}(X = 1) = p \tag{2.56}$$

We have that it has expectation $\mathbb{E}\,X = p$ and variance $\sigma^2 = p(1-p)$.

The Bernoulli distribution is often used to model binary outcomes, for example, whether a patient was cured by a drug, or whether the customer purchased the product.

2.4.2 Binomial distribution

The *Binomial* distribution is a discrete distribution that represents a series of coin flip by counting the number of heads. It is parametrised by p, the probability that the flip is a head, and by n, the number of flips. We write $X \sim \mathrm{Binom}(n, p)$.

The p.m.f. of the Binomial distribution is given by:

$$\mathrm{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{2.57}$$

We have that it has expectation $\mathbb{E}\,X = np$ and variance $\sigma^2 = p(1-p)$.

A binomial distribution is the sum of n independent Bernoulli distribution. It is most often used to model sums of binary outcomes. For example, the number of patient for which the drug trial was a success, or the number of match wins by a Basketball team in a season.

Note that if $X \sim \mathrm{Binom}(n_1, p)$, and $Y \sim \mathrm{Binom}(n_2, p)$, then $X+Y \sim \mathrm{Binom}(n_1+n_2, p)$.

2.4.3 Poisson distribution

The *Poisson* distribution is a discrete distribution that is usually used to represent a count. It is parametrised by a rate parameter λ, where $\lambda > 0$. We write $X \sim \mathrm{Poisson}(\lambda)$.

The p.m.f. of the Poisson distribution is given by:

$$\mathrm{P}(X = k) = \frac{1}{k!} \lambda^k e^{-\lambda}. \tag{2.58}$$

The Poisson distribution has mean $\mathbb{E}\,X = \lambda$ and variance $\sigma^2 = \lambda$.

The Poisson distribution is commonly used to model counts of occurences of events that are fairly independent. For example, it can be a good model for the number of customers in a store on a given day.

Note that if $X \sim \mathrm{Poisson}(\lambda_1)$ and $Y \sim \mathrm{Poisson}(\lambda_2)$, then $X + Y \sim \mathrm{Poisson}(\lambda_1 + \lambda_2)$.

2.4.4 Uniform distribution

The *uniform* distribution is a continuous distribution that represents values that are "equally likely" on an interval $[a, b]$ with $a < b$. We write $X \sim \mathcal{U}[a, b]$.

The p.d.f. of the uniform distribution is given by:

$$f_X(x) = \frac{1}{b-a}, \text{ for } a \leq x \leq b, \tag{2.59}$$

and $f_X(x) = 0$ everywhere else. The uniform distribution has mean $(a+b)/2$, and variance $(b-a)^2/12$.

The uniform distribution is often used as a simple probabilistic model for values that must be in a given interval. However, it is not particularly suitable for statistical models, as the assumption that every possible value is equally likely is usually unrealistic.

2.4.5 Normal distribution

The *Normal* distribution is a continuous distribution that is a standard model to represent a continuous quantity. It is parametrised by μ, the mean of the normal, and σ^2, the variance of the normal.

The p.d.f. of the normal distribution is given by:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}. \tag{2.60}$$

The normal distribution has mean $\mathbb{E}\, X = \mu$ and variance σ^2. The specific distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$ is called the *standard normal* distribution. Its p.d.f. is usually written $\phi(x)$.

The normal distribution is commonly used to model any continuous quantity. For example, it can be a good model for the weight of an animal, or the temperature of a sample.

Note that if $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$, then we have that $X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

2.4.6 Exponential distribution

The *exponential* distribution is a continuous distribution on the positive real numbers. It is parametrised by the rate $\lambda > 0$. We write $X \sim \mathrm{Exp}(\lambda)$.

The p.d.f. of the exponential distribution is given by:

$$f_X(x) = \lambda e^{-\lambda x}. \tag{2.61}$$

It has mean $\mathbb{E}\, X = 1/\lambda$ and variance $\sigma^2 = 1/\lambda^2$.

The exponential distribution is commonly used to model waiting times or survival times. For example, it might be a good distribution to model the time between two customers, or the time before a machine needs repairs. It is closely linked to the Poisson distribution: if e.g. customers arrive independently with a time gap distributed according to an exponential distribution, the number of customers in a unit of time is then Poisson.

2.5 Jointly distributed variables

We will often want to understand the behaviour of several related random variables at the same time. For example, we may be interested in understanding both the height and the weight of an individual. To do so, we may adapt the tools we have developed for a single random variable to the case of more than one random variable.

Die	1	2	3	4	5	6
X	0	1	0	1	0	1
Y	0	1	1	0	1	0

Table 2.1: List of outcomes for a single die roll

2.5.1 Joint distribution

We can express the random behaviour of two random variables using a *joint distribution*. For example, suppose X and Y are discrete random variables, the joint p.m.f. is then defined by

$$f_{XY}(x, y) = \mathrm{P}(X = x, Y = y). \tag{2.62}$$

If X and Y are continuous random variables, we may instead define their joint density f_{XY}.

Example 17 Suppose that we roll a single die, and let X be the random variable such that $X = 1$ if the roll is even, and $X = 0$ otherwise. Let Y be the random variable such that $Y = 1$ if the roll is prime, and $Y = 0$ otherwise. Let us collect all the possible outcomes of the die, and the corresponding values of X and Y into table 2.1.

By examining all the possible outcomes, we may compute the following joint p.m.f. for X and Y:

$$\begin{aligned}
f_{XY}(0, 0) &= \mathrm{P}(X = 0, Y = 0) = 1/6 \\
f_{XY}(1, 0) &= \mathrm{P}(X = 1, Y = 0) = 2/6 \\
f_{XY}(0, 1) &= \mathrm{P}(X = 0, Y = 1) = 2/6 \\
f_{XY}(1, 1) &= \mathrm{P}(X = 1, Y = 1) = 1/6
\end{aligned}$$

2.5.2 Joint distribution and events

Given two random variables X and Y, and their joint p.m.f. or p.d.f. f_{XY}, we may compute the probability of any events. For example, for X and Y given, we may be interested in computing $\mathrm{P}(X \leq 1, Y \geq 0)$. However, we may also desire to compute events of the type $\mathrm{P}(X \geq Y/2)$.

For X and Y discrete, we can simply sum over all the combinations of X and Y that verify the condition:

$$\mathrm{P}(X \geq Y/2) = \sum_{x,y:x\geq y/2} f_{XY}(x, y). \tag{2.63}$$

For X and Y continuous, we require double integration over the set of all x and y satisfying the condition, e.g.:

$$\mathrm{P}(X \geq Y/2) = \iint_{x,y:x\geq y/2} f_{XY}(x, y)\, dx\, dy \tag{2.64}$$

Example 18 Suppose X and Y are two continuous variables on $[0,1]$, with joint p.d.f. given by $f_{XY}(x,y) = 4xy$ for $0 \leq 1 \leq 1$. Let us compute $\mathrm{P}(X \leq Y/2)$.

We compute the double integral:

$$\begin{aligned}
\mathrm{P}(X \geq Y/2) &= \iint_{x,y:x \leq y/2} 4xy \, dx \, dy \\
&= 4 \int_0^1 \left(\int_0^{y/2} xy \, dx \right) dy \\
&= 4 \int_0^1 y \left(\int_0^{y/2} x \, dx \right) dy \\
&= 4 \int_0^1 yy^2/8 \, dy \\
&= \frac{1}{2} \int_0^1 y^3 \, dy \\
&= \frac{1}{8}
\end{aligned}$$

2.5.3 Joint distribution and expectations

Similarly, given two random variables X and Y, and their joint p.m.f. or p.d.f. f_{XY}, we may compute the expectation of a function $g(X,Y)$ of both variables.

For discrete random variables, we sum over all possible pairs of values of x and y:

$$\mathbb{E}\, g(X,Y) = \sum_{x,y} g(x,y) f_{XY}(x,y). \tag{2.65}$$

For continuous random variables, we integrate over all possible values of x and y:

$$\mathbb{E}\, g(X,Y) = \iint g(x,y) f_{XY}(x,y) \, dx \, dy \tag{2.66}$$

Example 19 Consider X, Y random variables on $[0,1]$ with joint p.d.f. $f_{XY}(x,y) = 6(x-y)^2$. Let us compute the expectation of $g(X,Y) = XY$.

$$\begin{aligned}
\mathbb{E}\, XY &= \iint xy6(x-y)^2 \, dx \, dy \\
&= 6 \int_0^1 \int_0^1 xy(x-y)^2 \, dx \, dy \\
&= 6 \int_0^1 \int_0^1 x^3 y - 2x^2 y^2 + xy^3 \, dx \, dy \\
&= 6 \int_0^1 \frac{1}{4} y - \frac{2}{3} y^2 + \frac{1}{2} y^3 \, dy \\
&= 6 \times [\frac{1}{8} - \frac{2}{9} + \frac{1}{8}] \\
&= 12(\frac{1}{8} - \frac{1}{9}) \\
&= 1/6.
\end{aligned}$$

2.5.4 Marginal distribution

Given two variables X and Y, with a joint p.m.f. or joint p.d.f. f_{XY}, suppose that we were only interested in the first variable X. Indeed, suppose that we "forget" about Y, what is the distribution of X?

If X, Y are discrete, we may compute the p.m.f. f_X of X by

$$f_X(x) = \sum_y f_{XY}(x, y), \tag{2.67}$$

where the sum is taken over all possible values of y.

Similarly, if X, Y are continuous, we may compute the p.d.f. f_X of X by:

$$f_X(x) = \int f_{XY}(x, y)\, dy, \tag{2.68}$$

where the integral is taken over all possible values of y.

Example 20 As in the previous section, consider a die roll, and X the random variable which is 1 if the roll is even, and 0 otherwise, and Y, the random variable which is 1 if the roll is prime, and 0 otherwise.

Let us compute the marginal p.m.f. of X. We have

$$\mathrm{P}(X = 0) = \mathrm{P}(X = 0, Y = 0) + \mathrm{P}(X = 0, Y = 1) = \frac{1}{6} + \frac{2}{6} = \frac{1}{2}. \tag{2.69}$$

2.5.5 Independent random variables

Let X and Y be two discrete random variables, and suppose that the events $X = x$ and $Y = y$ are independent for all x and y. Then, we would have by definition that:

$$\mathrm{P}(X = x, Y = y) = \mathrm{P}(X = x)\,\mathrm{P}(Y = y). \tag{2.70}$$

That is to say, we would have in terms of the p.m.f. that

$$f_{XY}(x, y) = f_X(x) f_Y(y). \tag{2.71}$$

When eq. (2.71) holds, for the p.m.f. if the variables are discrete, or the p.d.f. when the variables are continuous, we say that the two variables are *independent*.

Example 21 In the previous example of the die roll, we compute the joint p.m.f. and the marginal p.m.f. of X. It is not hard to see that the p.m.f. of Y is given by $\mathrm{P}(Y = 0) = \mathrm{P}(Y = 1) = 1/2$, as the event that a roll is odd happens with probability 0.5.

Now, note that $\mathrm{P}(X = 0, Y = 0) = 1/6$, whereas $\mathrm{P}(X = 0)\,\mathrm{P}(Y = 0) = 0.5 \times 0.5 = 1/4$, hence X and Y are *not* independent.

2.5.6 Conditional distributions

Let X and Y be two discrete random variables. We can consider the conditional distribution of X given $Y = y$, which is given by the definition of a conditional probability as:

$$\mathrm{P}(X = x \mid Y = y) = \frac{\mathrm{P}(X = x, Y = y)}{\mathrm{P}(Y = y)}. \tag{2.72}$$

Hence in terms of p.m.f., the conditional p.m.f. of X given Y, written $f_{X|Y}(x \mid y)$, is given by:

$$f_{X|Y}(x \mid y) = \frac{f_{XY}(x, y)}{f_Y(y)} \tag{2.73}$$

For continuous random variables, the event $\mathrm{P}(Y = y)$ always has probability 0, hence we may not directly define conditional distributions. However, we will define the conditional density according to eq. (2.73), replacing the p.m.f. by p.d.f.

2.6 Operation on random variables

As random variables represent numerical quantities, it is natural to operate on them as we would on numbers. The result of these operations define new random variables, for which we would like to understand their properties.

2.6.1 Transforming a random variable

Suppose that we would like to model the wealth of an individual as a random variable. This is a continuous quantity, hence a normal variable would be appropriate. However, empirical data and economic theories indicate that the more appropriate quantity to model is not the wealth, but the logarithm of the wealth. Suppose that we model the logarithm of the wealth X as a normal random variable. What is the distribution of the wealth $Y = e^X$?

In general, suppose that we have some continuous random variable X with density f_X, and some function g. We would like to compute the density of the variable $Y = g(X)$.

We have that

$$f_Y(y) = \frac{f_X(x)}{|g'(x)|}, \tag{2.74}$$

where $y = g(x)$.

Example 22 Suppose that X is standard normal, and $Y = e^X$. Then, the density of Y is given by:

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \frac{1}{e^x}. \tag{2.75}$$

Now, if $y = e^x$, we have $x = \log y$, hence replacing in the equation above gives:

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} \frac{1}{\log y} e^{-(\log y)^2/2}. \tag{2.76}$$

This is know as the *lognormal* distribution.

2.6.2 Sums of random variables

Given two random variables X and Y, we will often be interested in understanding the behaviour of the sum $X + Y$. For example, as we have seen previously, the binomial distribution can be understood as a sum of Bernoulli random variables.

Suppose X and Y are indepedent discrete random variables. Then $X + Y$ is also discrete random variable, and we may characterise its behaviour by its p.m.f. We thus wish to compute $\mathrm{P}(X + Y = k)$ for all k integers, say. Now, let us decompose the event $X+Y = k$ dependending on the value of X. Suppose $X = l$, then we must have $Y = k-l$. Conversely, if $X + Y = k$, then we must have $X = l$ and $Y = k - l$ for some l. Hence the event $X + Y = k$ is a disjoint union of the events $X = l, Y = k - l$. We thus deduce by the additivity rule that:

$$\mathrm{P}(X + Y = k) = \sum_l \mathrm{P}(X = l, Y = k - l), \tag{2.77}$$

where the sum is over all possible values of l. Writing this in terms of the joint p.m.f., we have that

$$f_{X+Y}(k) = \sum_l f_{X,Y}(l, k - l). \tag{2.78}$$

In the case that X and Y are independent, the joint p.m.f. factorizes, and we obtain that

$$f_{X+Y}(k) = \sum_l f_X(l) f_Y(k - l). \tag{2.79}$$

The right hand side is said to be a *convolution*.

If X and Y are continuous, we may compute the the p.d.f. of the sum $X + Y$ in a similar fashion as in eq. (2.78). Indeed, we have that the p.d.f. is given by

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_{X,Y}(t, z - t)\, dt. \tag{2.80}$$

If X and Y are independent, we may simplify the above expression to:

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(t) f_Y(z - t)\, dt. \tag{2.81}$$

Example 23 Let $X \sim \mathrm{Poisson}(\lambda_1)$ and $Y \sim \mathrm{Poisson}(\lambda_2)$ be independent Poisson random variables. What is the distribution of $X + Y$?

By eq. (2.79), we may write down the p.m.f. of the sum as

$$\mathrm{P}(X + Y = k) = \sum_{l=0}^{k} \mathrm{P}(X = l)\, \mathrm{P}(Y = k - l). \tag{2.82}$$

Use the formula for the Poisson p.m.f. to obtain:

$$\begin{aligned}
\mathrm{P}(X+Y=k) &= \sum_{l=0}^{k} \frac{\lambda_1^l e^{-\lambda_1}}{l!} \frac{\lambda_2^{k-l} e^{-\lambda_2}}{(k-l)!} \\
&= e^{-(\lambda_1+\lambda_2)} \sum_{l=0}^{k} \frac{1}{l!(k-l)!} \lambda_1^l \lambda_2^{k-l} \\
&= e^{-(\lambda_1+\lambda_2)} \frac{1}{k!} \sum_{l=0}^{k} \binom{n}{k} \lambda_1^l \lambda_2^{k-l} \\
&= e^{-(\lambda_1+\lambda_2)} \frac{(\lambda_1+\lambda_2)^k}{k!},
\end{aligned}$$

where we have used the binomial expansion of $(\lambda_1+\lambda_2)^k$. We now recognise that the computed p.m.f. corresponds to that of a Poisson$(\lambda_1+\lambda_2)$ random variable, hence we have that $X+Y \sim \text{Poisson}(\lambda_1+\lambda_2)$.

2.7 Properties of the expectation

The expectation of a random variable will be the main quantity we will be discussing in order to characterise the average behaviour of our estimators in statistics. It will thus be useful to understand some of its algebraic properties.

Let us start by mentioning a property that you will probably find obvious: the expectation of a constant. If a is a real constant (i.e. not random), then we have that $\mathbb{E}\, a = a$.

2.7.1 Expectation of a sum

From the definition of the expectation, we see that the expectation is essentially a sum (or integral). In particular, it is *linear*. That is, for X and Y random variables, we have that:

$$\mathbb{E}[X+Y] = \mathbb{E}\, X + \mathbb{E}\, Y. \tag{2.83}$$

In addition, if a is a real constant, we have that

$$\mathbb{E}\, aX = a\, \mathbb{E}\, X. \tag{2.84}$$

Indeed, recall the definition of the expectation as a sum (for discrete random variables), we have that:

$$\begin{aligned}
\mathbb{E}[X+Y] &= \sum_{x,y} (x+y)\, \mathrm{P}(X=x, Y=y) \\
&= \sum_{x,y} x\, \mathrm{P}(X=x, Y=y) + \sum_{x,y} y\, \mathrm{P}(X=x, Y=y) \\
&= \sum_{x} x\, \mathrm{P}(X=x) + \sum_{y} y\, \mathrm{P}(Y=y) \\
&= \mathbb{E}\, X + \mathbb{E}\, Y.
\end{aligned}$$

Unlike many other properties we will cover, this property does not necessitate that X and Y be independent.

Example 24 Let us derive another formula for the variance using the linearity property. We have defined variance as:

$$\sigma^2 = \mathbb{E}(X - \mathbb{E}\, X)^2 \tag{2.85}$$

We can expand the square to obtain the following:

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

We have used the fact that $\mathbb{E}\,X$ is a constant, and hence has expectation $\mathbb{E}\,X$, and the linearity of the expectation.

2.7.2 Expectation of product

We now turn to the expectation of a product of random variables. Unlike the sum, this will require that the variables be independent. Let X and Y be independent random variables, then we have that:

$$\mathbb{E}[XY] = (\mathbb{E}\,X)(\mathbb{E}\,Y) \tag{2.86}$$

Indeed, if X and Y are independent, then the p.m.f. or p.d.f. factorizes. We then have:

$$\begin{aligned}\mathbb{E}[XY] &= \sum_{x,y} xy f_{X,Y}(x,y) \\ &= \sum_{x,y} xy f_X(x) f_Y(y) \\ &= \sum_{x,y} x f_X(x) y f_Y(y) \\ &= (\sum_x x f_X(x))(\sum_y y f_Y(y)) \\ &= \mathbb{E}\,X\,\mathbb{E}\,Y.\end{aligned}$$

Note that we cannot bypass the requirement for independence. For example, suppose that $X = \pm 1$ with probability half for each. Then $\mathbb{E}\,X = 0$. However, $\mathbb{E}\,XX = \mathbb{E}\,X^2 = 1$ whereas $(\mathbb{E}\,X)^2 = 0$.

2.7.3 Variance of a sum

Using the previous two examples allows us to compute the variance of a sum of random variables. Let X and Y be random variables, and let σ^2_{X+Y} be the variance. We have that:

$$\begin{aligned}
\sigma^2_{X+Y} &= \mathbb{E}(X + Y - \mathbb{E}(X + Y))^2 \\
&= \mathbb{E}(X - \mathbb{E}\,X + Y - \mathbb{E}\,Y)^2 \\
&= \mathbb{E}\left[(X - \mathbb{E}\,X)^2 + (Y - \mathbb{E}\,Y)^2 + 2\,\mathbb{E}[(X - \mathbb{E}\,X)(Y - \mathbb{E}\,Y)]\right] \\
&= \mathbb{E}(X - \mathbb{E}\,X)^2 + \mathbb{E}(Y - \mathbb{E}\,Y)^2 + 2\,\mathbb{E}[(X - \mathbb{E}\,X)(Y - \mathbb{E}\,Y)] \\
&= \sigma^2_X + \sigma^2_Y + 2\sigma_{XY},
\end{aligned}$$

where we have put $\sigma_{XY} = \mathbb{E}[(X - \mathbb{E}\,X)(Y - \mathbb{E}\,Y)]$, the *covariance* of X and Y.

If X and Y are independent, we may compute the covariance explicitly by the product rule for expectations. Indeed, we have that:

$$\begin{aligned}
\sigma_{XY} &= \mathbb{E}[(X - \mathbb{E}\,X)(Y - \mathbb{E}\,Y)] \\
&= \mathbb{E}[XY - X\,\mathbb{E}\,Y - Y\,\mathbb{E}\,X + \mathbb{E}\,X\,\mathbb{E}\,Y] \\
&= \mathbb{E}\,XY - \mathbb{E}\,X\,\mathbb{E}\,Y - \mathbb{E}\,Y\,\mathbb{E}\,X + \mathbb{E}\,X\,\mathbb{E}\,Y \\
&= 0,
\end{aligned}$$

as $\mathbb{E}\,XY = \mathbb{E}\,X\,\mathbb{E}\,Y$ for X and Y independent. If $\sigma_{XY} = 0$, we say that X and Y are *uncorrelated*. In that case, we note that we have $\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y$, that is, the variance of the sum is the sum of the variance. Note that the same relationship does *not* hold for the standard deviation.

2.8 Limit theorems

Two fundamental limit theorems of probability will allow us to justify the theory of statistics to data, by showing that at least, as the amount of data we gather goes to infinity, we can be increasingly confident about our estimates.

2.8.1 Law of large numbers

The law of large numbers describes the first-order behaviour of sums of independent and identically distributed random variables. Let $X_1, \ldots, X_n$, be independent, identically distributed (i.i.d.) random variables. The sample mean may be written as:

$$S_n = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{2.87}$$

As the number of samples increases, we expect this quantity to get closer to the "true value" $\mathbb{E}\,X$, and to be exactly equal in the limit $n \to \infty$. This is essentially the assertion of the law of large numbers.

Theorem 1 (Law of large numbers) *Let $X_1, \ldots, X_n$ be independent, identically distributed random variables. Then, (under mild regularity assumptions), we have that (with probability 1):*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = \mathbb{E}\,X \tag{2.88}$$

A couple of notes on the conditions in parentheses: there are some distributions for which the randomness is so high that we cannot define a meaningful expectation. One common example is the *cauchy* distribution. In these cases, the law of large numbers does not apply.

The limit we are taking is also a random quantity, as it depends on the random variables $X_i, i = 1, 2, \ldots$. However, we are claiming that in the limit, there is in fact no randomness, and the limit always takes the value $\mathbb{E}\, X$. In order for this claim to make sense, we have to exclude a couple of cases, in which we get "infinitely" unlucky. For example, in a series of toin cosses, it is not *impossible* to get infinitely many tails. However, this happens with probability 0.

2.8.2 Central limit theorem

We have seen in the previous section that the sample mean of i.i.d. random variables converges to the expectation in the limit. Can we characterise the speed of convergence, and the distribution of the errors?

First, let us compute the variance of the sum. The variance is an indication of how much variation around the expectation our estimate has on average. First, we note the following property of the variance: for a random variable X and a real number a, we have that:

$$\operatorname{Var} aX = a^2 \operatorname{Var} X. \tag{2.89}$$

Indeed, recall that the variance has the same units as X^2. Now, we may use the formula above with our result for the variance of a sum to compute:

$$\operatorname{Var} \frac{1}{n} \sum_{i=1}^{n} X_i = \frac{1}{n^2} \sum_{i=1}^{n} \operatorname{Var} X_i = \frac{1}{n^2} n \operatorname{Var} X = \frac{\operatorname{Var} X}{n}. \tag{2.90}$$

Hence we see that the variance decays at a rate of n^{-1}. Thus, we have that the following quantity has constant variance:

$$\sqrt{n}(\bar{x} - \mathbb{E}\, X). \tag{2.91}$$

The central limit theorem allows us to characterise the distribution of this quantity exactly in the limit.

Theorem 2 *Central limit theorem Let X_i be i.i.d. random variables (with finite variance), with mean μ and variance σ^2. Let $\bar{x}_n = n^{-1} \sum_{i=1}^{n} X_i$ be the sample mean of the n first samples. We have that:*

$$\sqrt{n}(\bar{x}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2). \tag{2.92}$$

In the case of the central limit theorem, our quantity is converging to a random distribution. We can interpret this as saying that, if we were to repeat the experiment, and compute our error $\sqrt{n}(\bar{x} - \mu)$ again, the errors would follow a normal distribution. Note that this is a universal behaviour, and does not depend on the distribution of X. This in part underlies the prevalence of the normal distribution in statistics, as it arises naturally as the distribution of errors.

Example 25 Suppose that a store has an average of 200 customers per day. We wish to compute the probability that the store has more than 6150 customers over a month (30 days).

Let X_i be the number of customers on day i, and suppose that $X_i \sim \text{Poisson}(200)$. We wish to compute:

$$\begin{aligned}
\mathrm{P}(\sum_i X_i > 6150) &= \mathrm{P}(\frac{1}{30}\sum_i X_i > 205) \\
&= \mathrm{P}(\frac{1}{30}\sum_i X_i - 200 > 5) \\
&= \mathrm{P}(\sqrt{30}\left(\frac{1}{30}\sum_i X_i - 200\right) > 5\sqrt{30})
\end{aligned}$$

The quantity:

$$Z = \sqrt{30}\left(\frac{1}{30}\sum_i X_i - 200\right) \tag{2.93}$$

has an approximately normal distribution with mean 0 and variance 200 by the central limit theorem.

We may thus compute:

$$\mathrm{P}(Z > 5\sqrt{30}) = 0.02640. \tag{2.94}$$

By using properties of the Poisson distribution, we may also compute the probability exactly in terms of the Poisson distribution: $\sum_i X_i \sim \text{Poisson}(30 * 200)$, and

$$\mathrm{P}(\sum_i X_i > 6150) = 0.02637. \tag{2.95}$$

3 Sampling

The first step in any statistical analysis is the collection of data. Although we will not always have full control over this step, it is a crucial part of the statistical analysis, and must be analyzed as such. Collecting data in a sub-optimal fashion can often produce misleading results and limit the scopes of possible conclusions.

Data collection can be broadly categorised into two types: observational studies and experiments.

3.1 Experiments

In an *experiment*, the statistician is interested in understanding the relationship between some dependent variables and independent variables. To do so, the statistician will artificially control a given variable, for example whether a patient was given a drug, and observe its average effect on the variable of interest, for example the blood pressure.

3.1.1 Randomized experiment

Most commonly, the statistician will seek to assign patients randomly to either a trial or control group. This random assignment ensures that on average, a person given the treatment is the "same" as a person not given the treatment. Mathematically, the person given a treatment has the same distribution as a person not given the treatment.

This is the most powerful type of experiment, as it ensures that the difference observed between the two groups is entirely due to the treatment assignment. There can be no *confounding* variable.

3.1.2 Control groups and placebo

Although the randomized experiment design ensures that the difference in outcome only comes from whether the patient was assigned to a control or trial group, this is usually not sufficient. Indeed, the only conclusion we would be able to rigorously obtain on that is that difference in the entire treatment (e.g. meeting with doctors, taking the drug), produced the difference in the outcome. However, we are more often interested in a more specific aspect, for example, whether the drug itself produced a difference in the outcome.

Although this difference may seem trivial, the now well-document *placebo* effect states that the simple act of meeting with doctors, and taking a pill (which could have no medical effect), is enough to affect the outcome of a patient. In fact, the knowledge by the doctor (not the patient!) that there is a difference in treatment is enough to produce a different result. To address this issue, the clinical community has adopted the *double* *blind* trial, where each patient is issued either the real drug or a placebo, and the doctors do not know which.

In general, this issue highlights the fact that one should attempt to isolate the effect to be tested for as best as possible. In order to make the experiment unlikely to be affected by other possibilities, it is best to keep control and trial groups as similar as possible.

3.1.3 Sub-populations and inductive inference

A properly conducted randomized trial will usually be effective to isolate the desired effect. However, this effect is only observed for the population that participated in the study, and it is not always correct to extend it to a more general population. For example, suppose that a drug trial was conducted with 100 participants, who are white of germanic descent. However, the effect may not be the same in a hispanic population. It will often be necessary to replicate the experiment with other populations to ensure that the conclusion can be generalized.

3.2 Observational studies

It will often be the case that it is impossible or unethical to conduct the desired experiment. In this case, the statistician can only rely on an *observational* study, that is, observing different patients who happen to fall in a control or trial group. However, as the assignment to each group (for example, smoking) is no longer random, there is no guarantee that patients in one group are "statistically similar" to those in the other group. For example, there are well document links between smoking status and socio-economic status, which has further implications towards other health-related aspects, such as diet or access to healthcare.

We will thus often need to take special care in order to ensure that the comparison across groups that we wish to compute is valid, and will often not be able to reach conclusions that are quite as strong as in the experimental case. On the other hand, observational studies represent the vast majority of the data collected today, and allows us to use existing datasets instead of specifically designing an experiment for the problem at hand.

3.2.1 Prospective and retrospective studies

Observational studies can be broadly categorized into two main forms: *prospective* and *retrospective* studies. In a prospective study, the experimenter would identify individuals to observe, and collect information as time unfolds. Although these studies are able to provide strong evidence, they tend to be expensive. A famous prospective observational study is the Framingham heart study, which began in 1948 and tracked about 5000 adults and their descendants (it is today in its third generation of participants).

Retrospective studies, on the other hand, collects the data after the event of interest has taken place. For example, a researcher may identify 100 patients affected by lung cancer, and investigate their medical history. Retrospective studies are often much cheaper than prospective studies, and are particularly adapted for studying rare events. For example, the lifetime risk of lung cancer is about 7%. A prospective study on a cohort of 100 people would thus only observe about 7 cases of lung cancer on average, which is not adequate if one wishes to study lung cancer. On the other hand, a retrospective study would be able to select post-hoc 100 people who have experienced lung cancer.

3.2.2 Confounding

Observational studies face significant problems from confounding, which is the situation that a third variable is correlated with both the dependent variable and the independent variable. For example, consider an observational study of the effectiveness of sunscreen to prevent skin cancer. If we were only to measure the usage of sunscreen and the prevalence of skin cancer, we could possibly obtain a result that those who use sunscreen tend to be more likely to obtain skin cancer. However, we did not account for exposure to the sun, the confounding variable in this case: those who are more exposed to the sun are both more likely to use sunscreen, and more likely to experience skin cancer.

3.2.3 Natural experiments

In some cases, it may be possible to interpret a mechanism outside of the experimenter's control as a randomized experiment. This idea is of particular importance in fields such as econometrics, where it is often know as an *instrumental variable*. For example, a number of current estimates in inheritability of various characteristics comes from studies of twins separated at birth.

Another famous example is that of lifetime earnings of veterans of the Vietnam war. An initial simple estimate (by comparing American adults who were deployed vs. those who were not deployed) of the effect of the war on their lifetime earning seemed to indicate that veterans experienced a *positive* effect on their lifetime earnings. However, this comparison omits a number of effects, due to people volunteering. In 1990, Angrist used the fact that the draft lottery as a natural experiment to provide an estimate that in fact, the earning of veterans was about 15% less than the non-veterans.

4 Estimation

The problem of estimation is one of the central problems of statistics, and is the process through which we make inferences about a population from a collected sample.

4.1 Models and likelihood

In order to talk about a population in a mathematical fashion, we require a statistical *model* of the population. This model relates parameters of interest (for example, the effectiveness of a drug) to the data we observe (the blood pressure of individuals before and after taking the drug). Mathematically, we will make use of a family of probability distribution, most commonly one that we saw in section 2.4. This family of probability distribution thus relates the parameter to the data that is observed.

Example 26 Suppose we wish to understand the bias of a coin. In this case, the parameter of interest is the bias of the coin, which may be described as a number p, where p is the probability that the coin lands on heads.

Suppose we carry out the "experiment" of flipping the coin 10 times, then a statistical model for the experiment could be to say that the number of heads observed in the experiment follows a binomial distribution, with parameters $n = 10$ and p the bias.

We saw that the p.m.f. or p.d.f. was very useful in understanding the randomness of a quantity. Conversely in estimation, it will be useful to understand the relation between the parameter and the data. However, in probability, we think of the p.m.f. or p.d.f. as a function of the outcome (the value of the random variable) for a given parameter. In estimation, we are usually given the outcome as observed in the experiment we carried out, and instead wish to understand its implications on the parameter. We thus define the *likelihood*, which is the p.m.f. or p.d.f. viewed as a function of the parameter, and is usually writen $L(\theta)$, where θ is the parameter of interest.

Example 27 Continuing from the experiment of tossing 10 coins, the p.m.f. of the outcome is given by (for p fixed):

$$f_X(k) = \mathrm{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{4.1}$$

We will instead view this as a function of p. Suppose that we have k fixed (for example $k = 4$ if we observed four heads in the experiment), and consider the likelihood (written $L(p)$):

$$L(p) = \binom{n}{k} p^k (1-p)^{n-k} \tag{4.2}$$

We will usually use the *log-likelihood*, which is defined as the natural log of the likelihood, and written $\ell(\theta)$. This is mostly mathematically convenient, and will make the analysis much simpler.

4.2 Estimators

In this section, we will study *estimators*, which are functions of the data that attempt to "guess" the value of the parameter. For a given parameter θ, we will usually write $\hat{\theta}$ for an estimator of θ. Note that estimators depend on the data: they are thus *random* quantities. We will thus use the tools of probability to discuss estimators.

4.2.1 Unbiased estimators

For a general estimator, a natural question is to quantify how "good" an estimator is. A first attempt at capturing such a notion can be the very simple question: is our estimator right on average? That is, if we were to repeat the experiment many times, would the average outcome of the estimator be the true value?

We say an estimator $\hat{\theta}$ of θ is *unbiased* for θ if the following holds:

$$\mathbb{E}_\theta \hat{\theta} = \theta. \tag{4.3}$$

Here, the expectation is taken with respect to the data, supposing that the true parameter is given by θ.

Example 28 Consider again the case of the coin toss, and suppose that we consider the estimator:

$$\hat{p} = \frac{1}{n} X \tag{4.4}$$

Then, we claim that $\hat{p}$ is unbiased for X. Indeed, we may compute its expectation to obtain:

$$\mathbb{E}_p \hat{p} = \mathbb{E}_p \frac{1}{n} X = \frac{1}{n} \mathbb{E}_p X = \frac{1}{n} np = p. \tag{4.5}$$

However, an unbiased estimator only measures the fact that it is as likely to overestimate as underestimate the parameter of interest. In addition, we would hope that our estimator is close to the true value on average.

If $\hat{\theta}$ is unbiased for θ, then we have that the average squared-distance to the true value is given by:

$$\mathbb{E}_\theta (\hat{\theta} - \theta)^2 = \mathbb{E}_\theta (\hat{\theta} - \mathbb{E}\,\hat{\theta})^2 = \operatorname{Var} \hat{\theta}. \tag{4.6}$$

In particular, the notion of closeness for unbiased estimator may be captured by low variance of the estimator. To quantify this, we call the standard deviation of the estimator the *standard error*.

4.2.2 Mean-squared error

We will see that unbiasedness is not necessarily a desirable property of an estimator in general. Indeed, we will simply most of the time require that our estimator be close on average, or have low mean-squared error (often written *mse*):

$$\mathrm{mse}(\theta) = \mathbb{E}_\theta(\hat{\theta} - \theta)^2. \tag{4.7}$$

The mse quantifies the average distance of our estimate from the true value. It can be decomposed into two parts for which we will give individual interpretation:

$$\begin{aligned}
\mathrm{mse}(\theta) &= \mathbb{E}_\theta(\hat{\theta} - \theta)^2 \\
&= \mathbb{E}_\theta(\hat{\theta} - \mathbb{E}_\theta \hat{\theta} + \mathbb{E}_\theta \hat{\theta} - \theta)^2 \\
&= \mathbb{E}_\theta(\hat{\theta} - \mathbb{E}_\theta \hat{\theta})^2 + (\mathbb{E}_\theta \hat{\theta} - \theta)^2 \\
&= \mathrm{Var}_\theta \hat{\theta} + (\mathbb{E}_\theta \hat{\theta} - \theta)^2
\end{aligned}$$

We call $\mathrm{Var}\,\hat{\theta}$ the variance of the estimator, and $\mathbb{E}_\theta[\hat{\theta} - \theta]$ the bias of the estimator. We thus see that we may write:

$$\mathrm{mse} = \mathrm{variance} + \mathrm{bias}^2, \tag{4.8}$$

the *bias-variance* decomposition of the mse.

We will see that it can often be advantageous to trade off these two quantities, and in particular, it can sometimes be useful to incur a slight bias to reduce the variance greatly.

4.3 Maximum likelihodo estimation

Maximum likelihood estimation (or *mle*) is a general method of obtaining estimators for a given quantity. For a given statistical model and likelihood, the idea is to obtain the value of the parameter that maximizes the likelihood of the data. Conceptually, our guess for the data represents is the parameter that produces the observed data with the highest probability.

We illustrate this principle through two examples: a binomial example, and an exponential example.

4.3.1 Example: binomial model

Let X follow a binomial distribution with parameters n (known) and p (parameter of interest). The likelihood is then given by:

$$L(p) = \binom{n}{k} p^k (1-p)^{n-k} \tag{4.9}$$

We wish to compute the value of p that maximizes L, which is usually referred to as the arg-max and written $\arg\max_p L(p)$. However, as log is an increasing function, this is equivalent to computing the value of p that maximises $\log L(p) = \ell(p)$, the log-likelihood.

$$\ell(p) = k \log p + (n-k)\log(1-p) + \log \binom{n}{k} \tag{4.10}$$

Now, note that the last term $\log \binom{n}{k}$ does not depend on p, hence can be ignored for the purpose of maximization. We are thus left to maximize $k\log p + (n-k)\log(1-p)$. In order to do so, we will simply compute the derivative and set it to 0.

$$\frac{\partial \ell}{\partial p} = \frac{k}{p} - \frac{n-k}{1-p} = 0. \tag{4.11}$$

Solving the above equation in p gives $\hat{p} = k/n$, the mle for p in the binomial model.

4.3.2 Example: exponential model

Suppose now instead that we have $X_1, \dots, X_n$ i.i.d. random variables distributed with an exponential distribution of parameter λ. The likelihood is given by the joint p.d.f., which in this case is simply the product of the p.d.f. for each observation X_i as the observations are independent.

$$L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i} \tag{4.12}$$

We may similarly compute the log-likelihood to obtain:

$$\ell(\lambda) = n \log \lambda - \lambda \sum_{i=1}^{n} x_i. \tag{4.13}$$

To find the location of the maximum, we again differentiate and set the derivative to 0 to obtain:

$$\frac{\partial \ell}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i \tag{4.14}$$

and solving this equation in λ gives the mle for the exponential model:

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i}. \tag{4.15}$$

4.3.3 Theoretical properties for the mle

The mle has the advantage of having numerous theoretical properties for any reasonable model. In particular, it has two notable properties known as asymptotic *consistency* and asymptotic normality. We say that these properties are *asymptotic* as they only hold exactly as the sample size goes to infinity, but we will usually consider that they hold approximately for a finite sample.

Consistency denotes the fact that although the mle is not unbiased, its bias vanishes as the sample size goes to infinity. The mle is always right when we collect an infinite amount of data. This property can be seen as an analogue of the law of large number for the sample mean.

Normality denotes the fact that the mle follows an approximately normal distribution when the sample is large, with a standard error that can be computed. The exact experssion of this standard error is beyond the course, but most software tools we make use of are able to report it.

4.4 Method of moments

The method of moments is an alternative strategy to obtain estimators for a general estimation problem. Although it is often simpler than the mle, it is somewhat inferior and its theoretical properties can be difficult to analyze. It is also difficult to apply in complex models where the parameter is no longer a number but a more general mathematical object.

The idea of the method of moment is to match the population and sample moments, and solve the set of equations to obtain an estimator. First, let us define what a *moment* is.

Suppose we have a sample $X_1, \ldots, X_n$. We define the first population moment to be the population mean $\mu_1 = \mathbb{E}\, X$, and similarly, the first sample moment is the sample mean $M_1 = \frac{1}{n}\sum_{i=1}^n X_i$.

Now, for $j \geq 2$, we define the j^{th} centered population moment to be:

$$\mu_j = \mathbb{E}(X - \mathbb{E}\, X)^j, \tag{4.16}$$

and the j^{th} centered sample moment in a analogous fashion as:

$$M_j = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^j, \tag{4.17}$$

where $\bar{X}$ is the sample mean. The second centered moment is usually called the variance, the third one the *skewness*, and the fourth one the *kurtosis*.

4.4.1 Example: exponential distribution

Let us illustrate the method of moments by a simple example, the exponential example we also considered for the mle. Suppose that $X_1, \ldots, X_n$ follow an exponential distribution with parameter λ. The population mean is then given by $\mu_1 = \lambda^{-1}$. The sample mean is simply given by $M_1 = n^{-1}\sum_{i=1}^n X_i$.

To obtain an estimator of λ according to the method of moments, we simply equate $\mu_1 = M_1$, and solve for the parameter λ. We thus obtain:

$$\frac{1}{\lambda} = \frac{1}{n}\sum_{i=1}^n X_i, \tag{4.18}$$

which we may solve to obtain the method of moments estimator:

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^n X_i}. \tag{4.19}$$

4.4.2 Example: gamma distribution

Let us consider a slightly more complex example, the gamma distribution, with two parameters. Suppose that $X_1, \ldots, X_n$ follow a gamma distribution with shape k and scale θ. The p.d.f. is given by:

$$f_X(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-x/\theta}, \tag{4.20}$$

and the mean and variance of the gamma distribution are given by

$$\mathbb{E}\, X = k\theta \text{ and } \sigma^2 = k\theta^2. \tag{4.21}$$

Note that applying the mle methodology manually here would be quite difficult, as the p.d.f. is complex, although it is still quite easy with the help of a computer. However, we may apply the method of moments to obtain the system of equations:

$$\begin{cases} M_1 = k\theta, \\ M_2 = k\theta^2, \end{cases} \tag{4.22}$$

which we may solve in the parameters to obtain two estimators $\hat{\theta} = M_2/M_1$ and $\hat{k} = M_1^2/M_2$. In this case, we needed to match two moments as we had two parameters. In general, we will have to match as many moments as there are parameters.

4.5 Uncertainty in estimation

In addition to providing a "guess" for the parameter, a statistician is also interested in quantifying the certainty or uncertainty of the given guess. Indeed, although observing 5 heads in 10 tosses, and 500 heads in 1000 tosses gives the same estimate $\hat{p} = 0.5$, having a larger sample size indicates that the second estimate is more confident.

As we have seen before, one possibility to do so could be to report the standard error of the estimate, which gives some indication as to the variability of the estimate. However, in some cases the standard error may fail to capture the whole picture, and we will prefer to report a range of values, called a confidence interval.

4.5.1 Confidence intervals

The idea of a *confidence interval* is to produce an estimate as a range of possible values instead of a single value. We would like this range of values to capture the likely possibilities of the parameter. As this range is produced from the data, the range itself is a random quantity, and thus can be analyzed probabilistically.

Let us define formally a $(1 - \alpha)$ confidence interval (e.g. for a 95% confidence interval, $\alpha = 0.05$). We may write a confidence interval as $[a(X), b(X)]$, where a is the lower bound, and b the upper bound, and the dependence on the data X has been made explicit. Then, we say that $[a(X), b(X)]$ is a $(1 - \alpha)$ confidence interval for θ if:

$$\mathrm{P}_\theta(a(X) \leq \theta \leq b(X)) = 1 - \alpha, \tag{4.23}$$

where the probability is taken supposing that θ is the true value. We may interpret this as saying that if we repeat the experiment many times, and compute a 95% confidence interval in the same way every time, this interval will cover the true value of the parameter 95% of the experiments.

4.5.2 Confidence interval for a normal observation

We will now compute a confidence interval for a single normal observation. This is one of the most used forms of confidence intervals, as we have seen that mle are asymptotically normal, and hence this method may be used to produce confidence intervals for the mle.

Suppose that we observe a single observation $\hat{\theta} \sim \mathcal{N}(\theta, \sigma_\theta^2)$, where σ_θ is the standard error of our estimate (which is supposed known). We claim that the interval:

$$\left[\hat{\theta} - \sigma_\theta z_{1-\alpha/2}, \hat{\theta} + \sigma_\theta z_{1-\alpha/2}\right] \tag{4.24}$$

is a $(1-\alpha)$ confidence interval, where we have defined $z_{1-\alpha/2}$ to verify, where Z is a standard normal random variable:

$$\mathrm{P}(Z \leq z_{1-\alpha/2}) = 1 - \alpha/2. \tag{4.25}$$

Example 29 Suppose that after running an experiment, we obtain an estimate of $\hat{\theta} = 4.5$ with a standard error of 1.2. Let us compute a 95% confidence interval for θ. Using software, we may compute that $z_{1-\alpha/2} = 1.96$, and hence the interval is given by:

$$[4.5 - 1.96 \times 1.2, 4.5 + 1.96 \times 1.2] = [2.15, 6.85]. \tag{4.26}$$

4.5.3 Bootstrapping confidence intervals

The normal assumption is not always appropriate, and we will sometimes encounter cases in which we may not wish to use a normal approximation, or are not able to compute the standard error of the estimator. In these cases, the *bootstrap* is a general strategy that allows us to estimate a confidence interval from the data.

In general, we may compute a confidence interval by noting that if we knew a and b such that:

$$\mathrm{P}(a \leq \hat{\theta} - \theta \leq b) = 1 - \alpha. \tag{4.27}$$

Indeed, this would directly imply (by algebraic manipulation) that:

$$\mathrm{P}(\hat{\theta} - b \leq \theta \leq \hat{\theta} + a) = 1 - \alpha, \tag{4.28}$$

that is, the interval $[\hat{\theta} - b, \hat{\theta} + a]$ is a $(1-\alpha)$ confidence interval for θ.

Unfortunately, we do not have access to the distribution of $\hat{\theta} - \theta$, as the distribution of the estimator may be complex. Instead, we propose to estimate this distribution. We estimate the distribution by artificially creating new datasets by resampling from our existing data.

We create a new dataset of the same size as our existing dataset by picking observations from our dataset at random (we may pick the same observation several times). We then compute the value of our estimator on this resampled dataset, and call its value θ^{boot}. By simulating this artificial dataset many times, we may estimate the distribution of $\theta^{\text{boot}} - \hat{\theta}$.

Now, the idea is that the distribution of $\theta^{\text{boot}} - \hat{\theta}$ is close to that of $\hat{\theta} - \theta$, and so instead of using a and b from the original distribution, we may have a and b be sample quantiles from the bootstrap distribution.

Index