Introduction to statistics

Wenda Zhou

May 23, 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, $x_1, \ldots, x_n$, and $y_1, \ldots, y_n$, 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:

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

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 $x_1, \ldots, x_n$, 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.

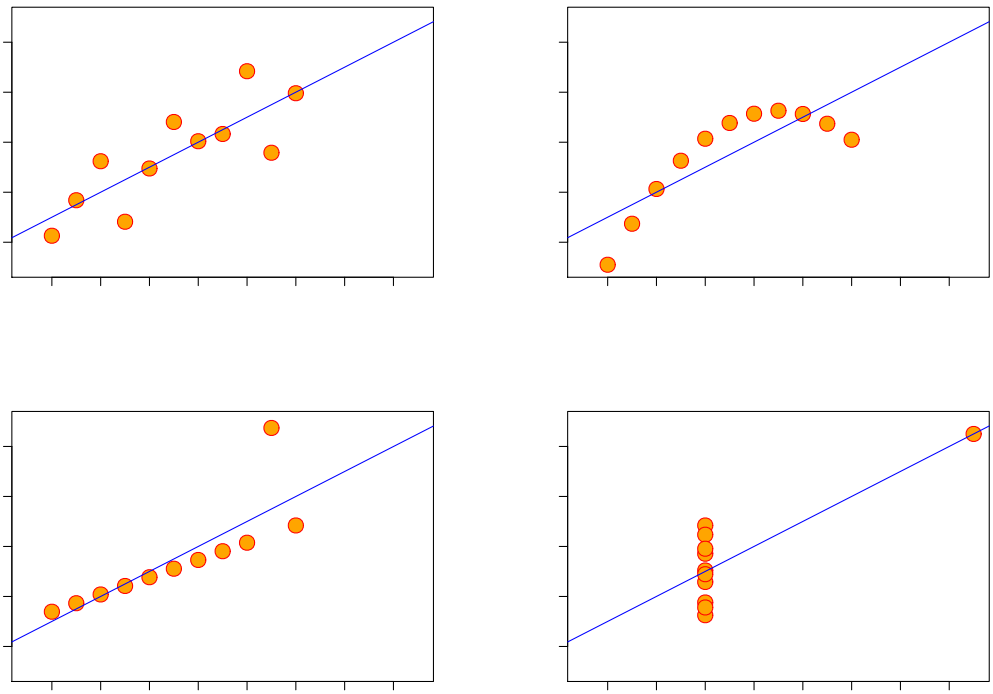

Figure 1.1: Anscombe's quartet, four datasets with the same mean, variance and covariance. Source: Schutz

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

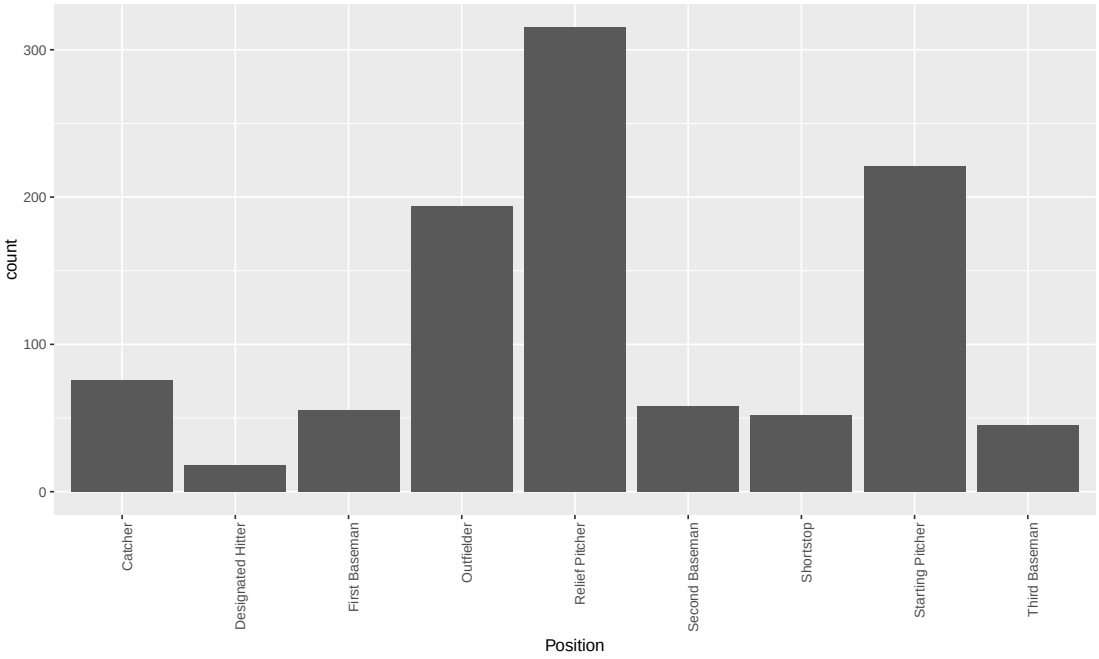

Figure 1.2: Bar chart of position

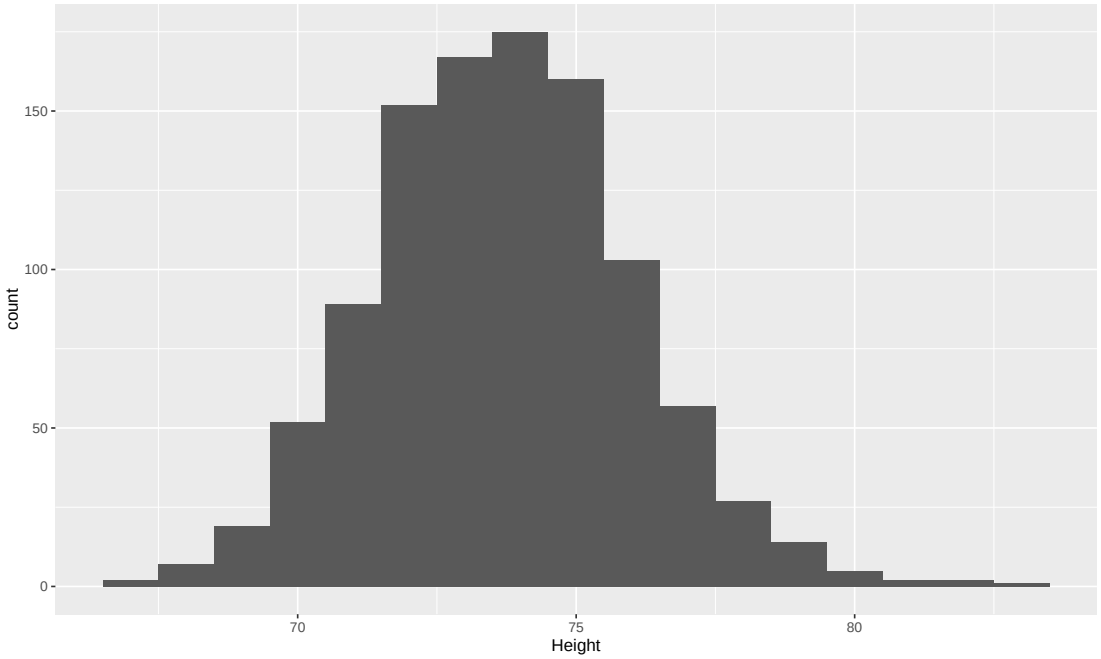

Figure 1.4: Histogram of heights

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*.

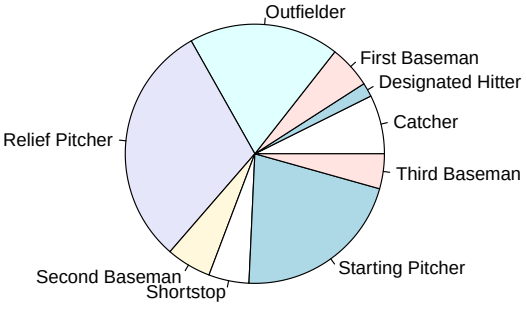

Figure 1.3: Pie chart of position

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).

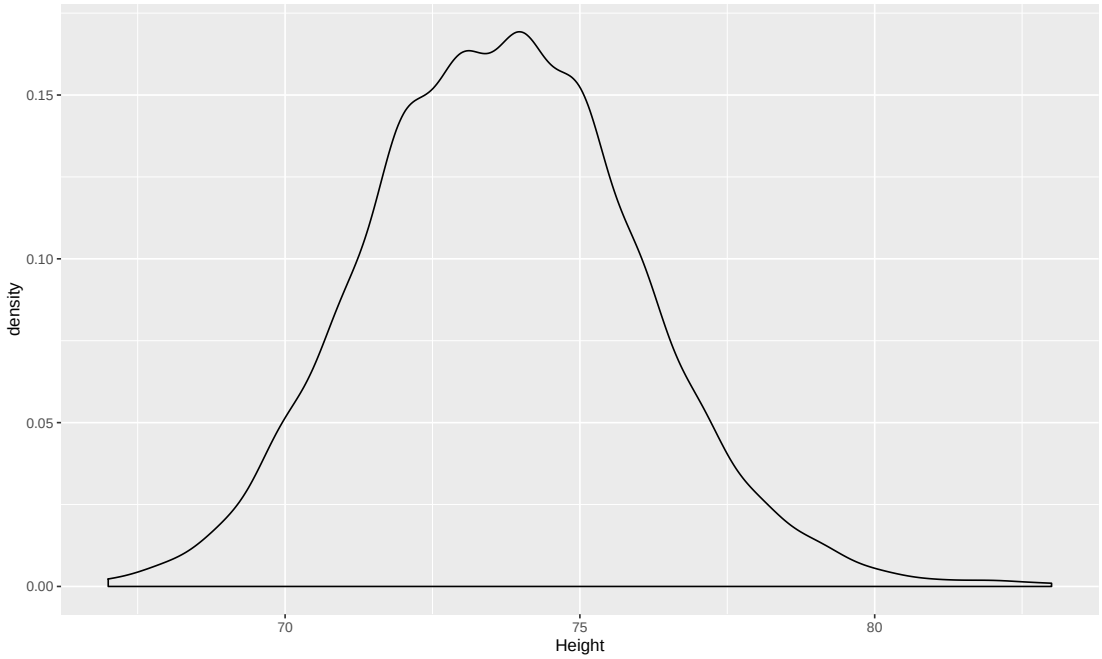

Figure 1.5: Kernel density estimate of heights

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 $1.5 \times$ IQR. Points beyond that distance are then plotted individually.

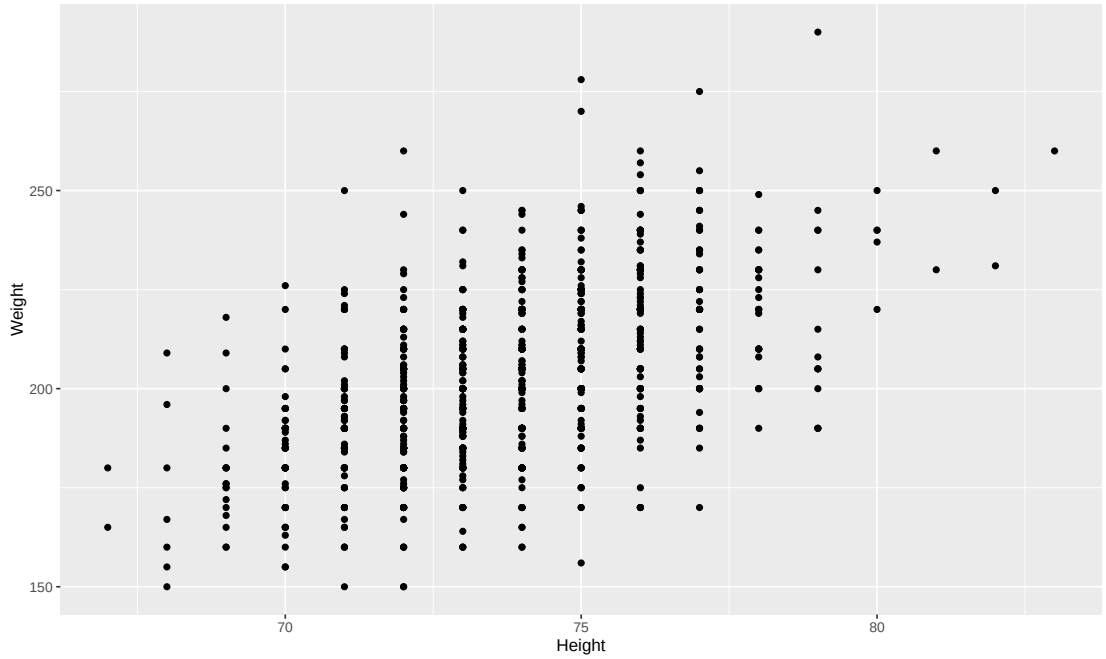

Figure 1.6: Scatter plot of heights and weight

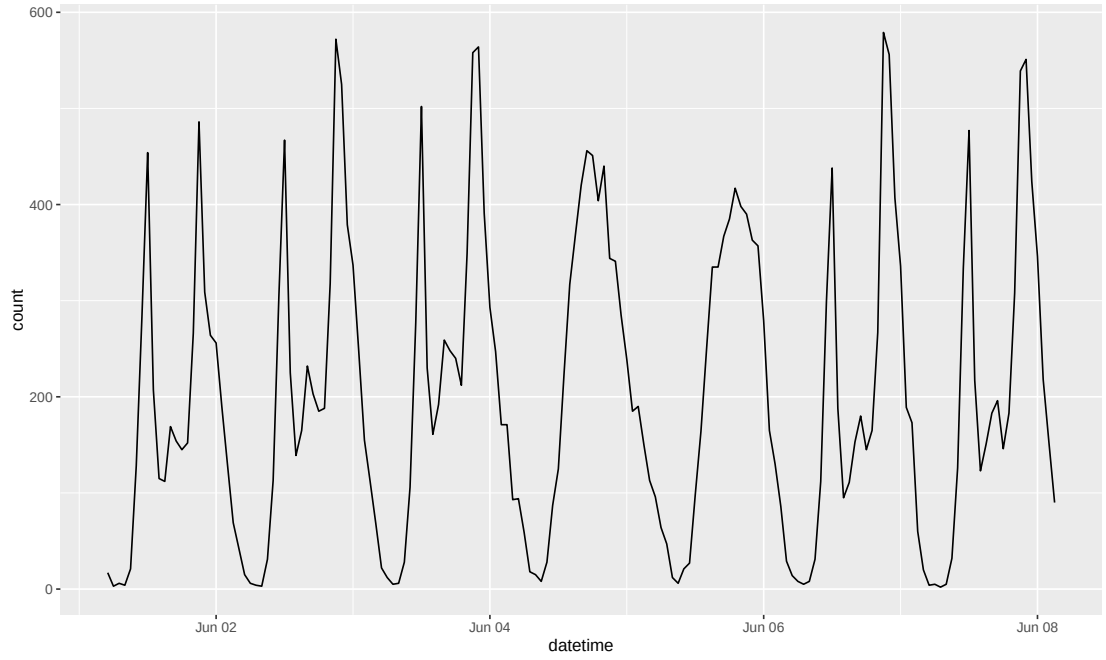

Figure 1.7: Bikeshare program usage

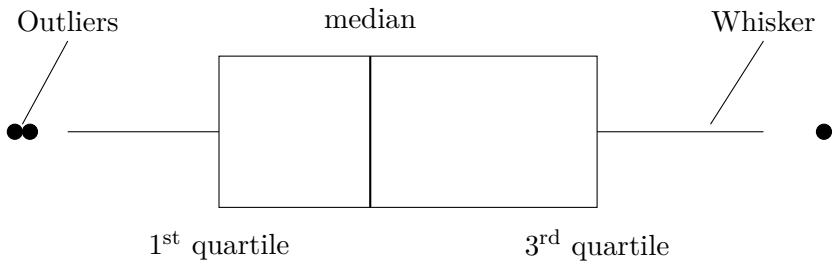

Figure 1.8: Anatomy of a boxplot

.

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

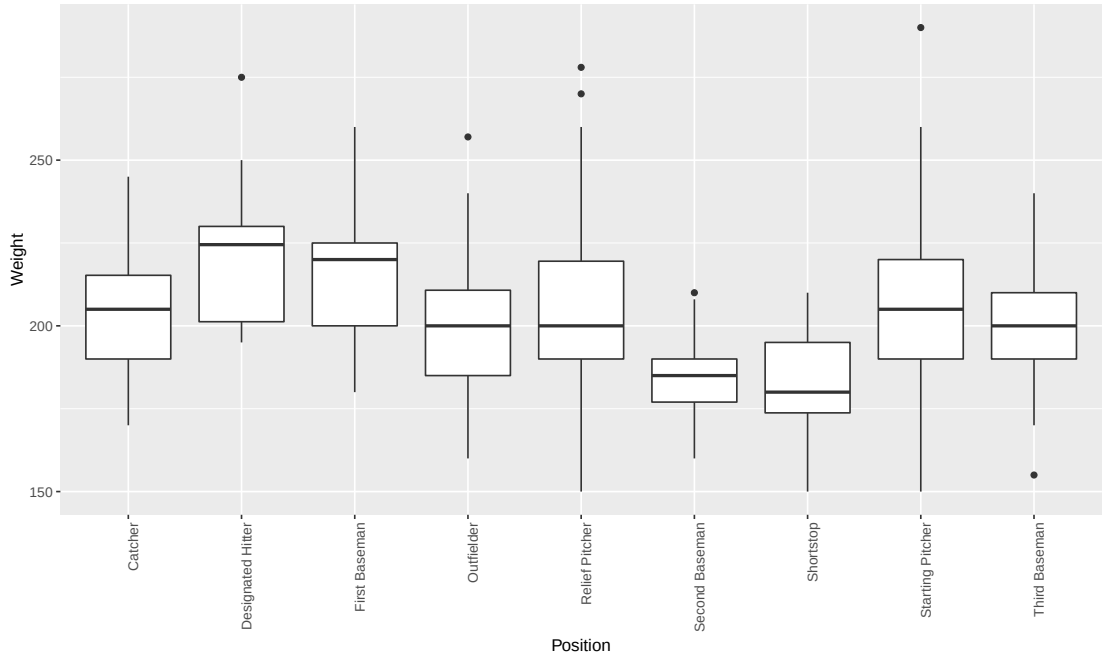

Figure 1.9: Boxplot of weight by position

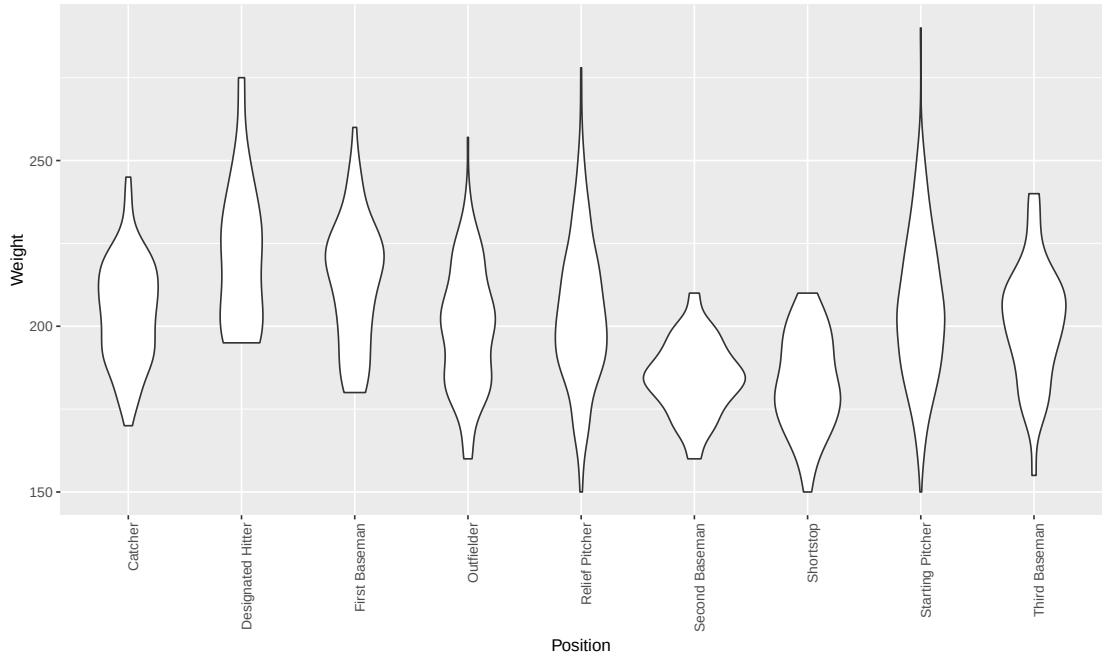

Figure 1.10: Violin plot of weight by position

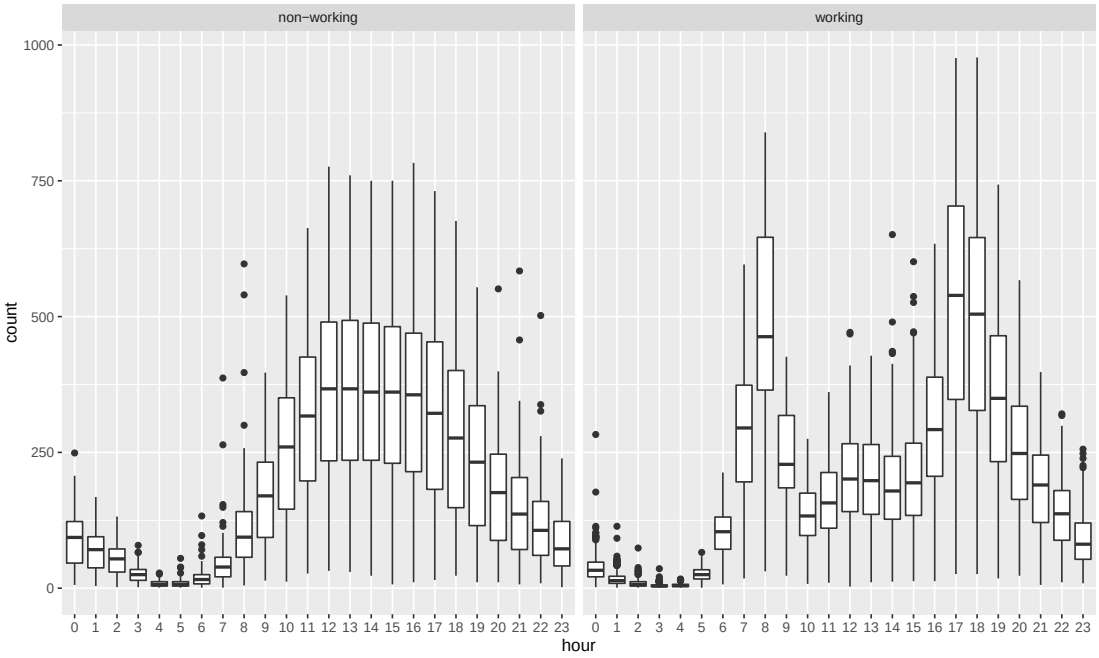

Figure 1.11: Usage of the bike sharing program by hour of day for working days and non-working days

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.

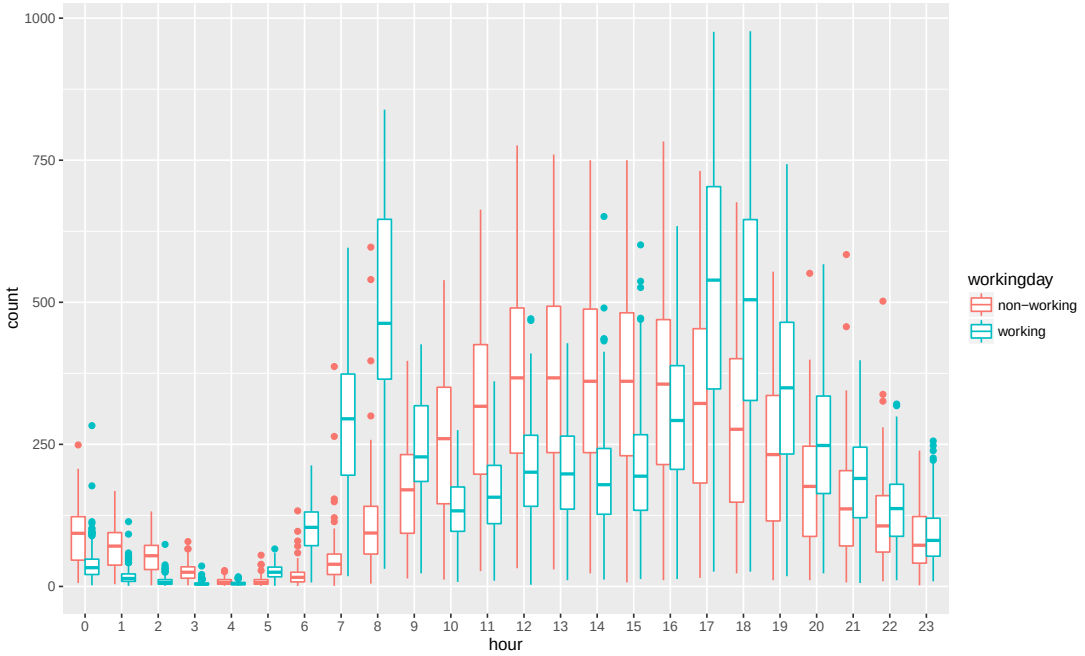

Figure 1.12: Usage of the bike sharing program by hour of day for working days and non-working days

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 dice 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 dice roll. For a standard 6-sided dice, 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 dice roll, an event could be that the dice lands on 3. However, an event could also be that the dice 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.

2.2.1 Disjoint or mutually exclusive events

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 dice we use, the event "the dice lands on an even number" is more likely than the event "the dice 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.

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*.

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 dice 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 dice, 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 dice, 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

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

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

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

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:

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

Index