Chapter 3
Probability Theory

3.1 Introduction

Statistics is a science that is concerned with principles, methods, and techniques for collecting, processing, analyzing, presenting, and interpreting (numerical) data. Statistics can be divided roughly into descriptive statistics (Chap. 1) and inferential statistics (Chap. 2), as we have already suggested. Descriptive statistics summarizes and visualizes the observed data. It is usually not very difficult, but it forms an essential part of reporting (scientific) results. Inferential statistics tries to draw conclusions from the data that would hold true for part or the whole of the population from which the data is collected. The theory of probability, which is the topic of the next two theoretical chapters, makes it possible to connect the two disciplines of descriptive and inferential statistics.

We have already encountered some ideas from probability theory in the previous chapter. To start with, we discussed the probability of selecting a specific sample π_k and we briefly defined the notion of probability based on the throwing of a dice. In this chapter we work out these ideas more formally and discuss the probabilities of events; we define probabilities and discuss how to calculate with probabilities. In the previous chapter, when discussing bias, we have also encountered the expected population parameter $\mathbb{E}(T)$, but we have not yet detailed what expectations are exactly; this is something we cover in Chap. 4.

To summarize, descriptive statistics only get us so far. If we want to do more interesting things we need to have a formal, theoretical, understanding of probability. This is exactly what we cover in the next two chapters. However, despite being primarily theoretical, we introduce practical examples of each of the concepts we introduce throughout the chapters.

In this chapter we will study:

- Basic principles and terminology of probability theory
- Calculation rules for probability; the probability axioms
- Conditional probability
- Measures of risk, and their association with study designs
- Simpson's Paradox

M. Kaptein and E. van den Heuvel, *Statistics for Data Scientists*, Undergraduate Topics in Computer Science, https://doi.org/10.1007/978-3-030-10531-0_3

3.2 Definitions of Probability

In daily life, probability has a subjective interpretation, because everyone may have his or her own intuition or ideas about the likelihood of particular events occuring. An *event* is defined as something that happens or it is seen as a result of something. For instance, airplane crashes around the world or congenital anomalies in newborn babies can be considered events. These two types of events are considered rare because their frequency of occurrence is considered small with respect to the possible number of opportunities, but nobody knows exactly what the probability of such an event is. Using information or empirical data, the probability of an event can be made more quantitative. One could assign the ratio of the frequency of an event and the frequency of opportunities for the event to occur as the probability of this event: for instance, the yearly number of newborn babies with a congenital anomaly as a ratio of all yearly newborns or the number of airplane crashes in the last decade as a ratio of the number of flights in the same period.

The possible opportunities for an event to occur can also be viewed as a population of units (e.g., newborns or flights in a particular period of time) and the events can be seen as the population units with a specific characteristic (e.g., congenital anomalies or airplane crashes). In this context the definition of the probability of an event A for a finite population can be given by

$$\Pr(A) = \frac{N_A}{N}, \tag{3.1}$$

with N_A the number of units with characteristic A and N the size of the population.

The definition in Eq. (3.1) is only correct if each opportunity for the event to occur is as likely to produce the event as any other opportunity. Indeed, if for instance, congenital anomalies may occur more frequently for older women than for younger women or airplane crashes might occur more frequently for intercontinental flights than for continental flights, the definition in Eq. (3.1) is inappropriate. Of course, in such cases we may reduce the population into a smaller set of units or divide it into several subsets and then apply the definition in Eq. (3.1) to these subsets, but this can only be performed if the number of units in the subsets does not become too small. If we have to create very many subsets, it may happen that the probabilities for these subsets are only equal to 1 or 0, which would make the definition less useful. Another limitation of this definition is that it is defined for finite populations only. In the case of tossing a coin until a head appears, the sequence of tosses can in theory be infinitely long and the definition in Eq. (3.1) seems unsuitable.

This brings us to an alternative, and more theoretical, approach to the definition of probability of an event, which assigns to this probability the proportion of the occurrence of an event obtained from infinitely many repeated and identical trials or experiments under similar conditions. This definition has its origin in gambling; thus we will explain it by considering dice throwing once again: if a die is thrown n times and the event A is the single dot facing up, then the probability $\Pr(A)$ of the

event A can be *approximated* by the ratio of the number of throws n_A with a single dot facing up and the total number of throws n, i.e.,

$$\Pr(A) \approx \frac{n_A}{n}.$$

When the number of repeated trials n is increased it is expected that the proportion n_A/n converges to some value p (which would be equal to $1/6$ if the die is *fair*). Thus an alternative definition of probability can now be given by

$$\Pr(A) = \lim_{n\to\infty} \frac{n_A}{n} = p. \tag{3.2}$$

Clearly, this definition is only appropriate when repeated and identical trials can be conducted under almost similar conditions (e.g., gaming and gambling). Thus definition Eq. (3.2) is more conceptual for real situations, since it is impossible to conduct infinitely many or even many of these trials in practice. Indeed, in the case of congenital anomalies, it would be extreme to have each mother deliver (infinitely) many babies just to be able to approximate or apply the definition of probability in Eq. (3.2).

This fact, however, does not imply that the definition in Eq. (3.2) is useless. On the contrary, it merely shows where particular assumptions about probabilities are introduced. For instance, for each pregnant woman we could assign an individual probability parameter of reproducing a newborn baby with a congenital anomaly. If we are willing to assume that the probabilities for these women are all equal, the probability in Eq. (3.1) is an approximation to the probability in Eq. (3.2), but alternatively we could also assume that equal probabilities only exist for pregnant women of certain age. In the Additional material at the end of this chapter we provide a brief overview of the history of probability, highlighting different ways in which people have thought about probabilities and probability theory over the years.

Definition of probability. A formal mathematical framework of probability can be constructed (see the Additional material at the end of this chapter). In this textbook and in context with the above-mentioned definitions Eq. (3.1) and Eq. (3.2), we simply define the probability $\Pr(A)$ of an event A as an (unknown) value between zero and one, $0 \leq \Pr(A) \leq 1$, where both boundaries are allowed, which could either be approximated by collecting appropriate and real data or by the limit of a proportion of repeated and identical trials.[1] To operationalize probability we also need some calculation rules; we discuss these in the next section.

[1] It should be noted here that a probability of zero does not necessarily mean that the event will never occur. This seems contradictory, but we will explain this later. On the other hand, if the event can never occur, the probability is zero.

3.3 Probability Axioms

There are several calculation rules for probabilities, but before we discuss some of them we need to introduce some standard notation on events.

- The *complement* of event A is denoted by A^c and it indicates that event A does not occur. Thus the complement of having a female baby is having a male baby (although there exists literature that suggests that gender or sex is much more fluent).
- The occurrence of two events A and B at the same time is denoted by $A \cap B$. This is often referred to as the *joint* or *mutual* event. If event A represents a congenital anomaly in a newborn baby and event B represents the gender male of the baby, then $A \cap B$ represents the event that the baby is both male and has a congenital anomaly.
- The event that either A or B (or both) occurs is denoted by $A \cup B$. Thus in the example of newborn babies, $A \cup B$ means that the baby is either male (with or without an anomaly) or is female with a congenital anomaly. This is the complement of the event of having a female baby without an anomaly (i.e., $(A \cup B)^c = A^c \cap B^c$).

We also have to provide some additional definitions relevant for probabilities

- The probability of no event must be zero. Not having events is indicated by the empty set $\emptyset$ and the probability is $\Pr(\emptyset) = 0$. For instance, if two events A and B can never occur together (mutually exclusive events), then it follows that $A \cap B = \emptyset$ and $\Pr(A \cap B) = \Pr(\emptyset) = 0$. The mutual event that a newborn baby has an anomaly in its uterus and is also a boy does not exist. This should have probability zero of occurring.
- The probability that event A occurs is one minus the probability that the event A does not occur; thus $\Pr(A) = 1 - \Pr(A^c)$. This rule is based on the assumption that either event A occurs or event A^c occurs. This means that $\Pr(A \cup A^c) = 1$, since we will see either A or A^c.
- We call two events A and B *independent* if and only if the probability of the mutual event is equal to the product of the probabilities of each event A and B separately. Thus the independence of events A and B (denoted by $A \perp B$) is equivalent with $\Pr(A \cap B) = \Pr(A) \cdot \Pr(B)$. Using products of probabilities when independence is given or assumed is applied frequently throughout the book. Note that any event A with the non-event $\emptyset$ is independent: $\Pr(A \cap \emptyset) = \Pr(\emptyset) = 0 = 0 \cdot \Pr(A) = \Pr(\emptyset)\Pr(A)$. Alternatively, if two events with a positive probability ($\Pr(A) > 0$ and $\Pr(B) > 0$) that are also mutually exclusive can never be independent: $0 = \Pr(\emptyset) = \Pr(A \cap B) < \Pr(A)\Pr(B)$. We will discuss dependencies in more detail in Chap. 6.

Using the above definitions we can define the following calculation rules:

1. If the events A and B are independent, then the events A and B^c, the events A^c and B, and the events A^c and B^c are also independent. Demonstrating this fact is left to the reader and part of the assignments.

2. The probability of the occurrence of either event A or B or both is equal to the sum of the probabilities of these events separately minus the probability that both events occur at the same time, i.e., $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$. Note that mutually exclusive events A and B imply that $\Pr(A \cup B) = \Pr(A) + \Pr(B)$.
3. The probability of an event A is the sum of the probability of both events A and B and the probability of both events A and B^c, thus $\Pr(A) = \Pr(A \cap B) + \Pr(A \cap B^c)$. This is sometimes referred to as the *law of total probability* and is frequently applied throughout the book. Note that this rule follows directly from the second rule.

Our definition of probability, combined with these calculation rules, jointly compose the core of our theoretical discussion of probabilities. In essence, all of the material in this chapter and the next can be derived from these simple rules. However, you will be surprised by the many interesting results we can find merely based on these simple rules!

3.3.1 Example: Using the Probability Axioms

To explain the calculation rules using a practical example we will make use of a deck of cards. A deck of cards contains 52 playing cards: 13 clubs, 13 diamonds, 13 hearts, and 13 spades. Diamonds and hearts have color red and clubs and spades are black. Each suit has the same 13 different values: 2, 3, 4, 5, 6, 7, 8, 9, 10, jack, queen, king, and ace. Now suppose one card is randomly collected from the deck, then we can answer the following questions:

- What is the probability that this card is a heart?
- What is the probability that this card is not a heart?
- What is the probability that it is a heart and a king?
- What is the probability that the card is a heart or a king?
- Are the events that the card is a heart and is a king independent?

Note that by virtue of our random selection of the card we are actually investigating the probabilities of a specific simple random sample containing a single unit (as discussed in Chap. 2).

The probability that the card is a heart is equal to $1/4$. This can be deduced in at least two ways. First of all, there are 52 cards in total (the population of cards) and each card is as likely to be drawn as any other card. The 13 hearts are all favorable for the event or outcome of drawing a heart. Thus using definition Eq. (3.1), the probability is given by $13/52 = 1/4$. An alternative approach is to define the population by the four suits, and only one suit would provide the appropriate event of drawing a card of hearts, leading to $1/4$ directly.

The probability that the card is not heart is now $3/4$, using the definition that the probability of the complementary event is one minus the probability of the event.

There is only one king of hearts, which leads to the probability of $1/52$ of the event that the card is both heart and a king, using definition Eq. (3.1).

The probability of drawing a king is $4/52 = 1/13$, since there are four kings among the 52 cards. The probability that the randomly collected card is hearts or a king is equal to $1/4 + 1/13 - 1/52 = 4/13$, using the second calculation rule.

Since the product of probabilities of hearts and a king is $(1/4) \cdot (1/13) = 1/52$, which is equal to the probability of drawing the king of hearts, the definition of independence implies that the events are independent.

3.4 Conditional Probability

In some situations probability statements are of interest for a particular subset of outcomes. For instance, what is the probability that the newborn baby has a congenital anomaly *given* that the baby is a boy. This question is of interest because it could be possible that congenital anomalies are more frequent for boys than for girls. This probability is generally not the same as the probability that both events congenital anomaly and gender male occur, because we have excluded events of the type female. If event A represents a congenital anomaly and B represents the event of a male newborn baby, then the probability of interest is the so-called *conditional probability* denoted by $\Pr(A|B)$. We refer to this conditional probability as the probability of event A given event B. It is defined by

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}, \text{ when } \Pr(B) > 0. \tag{3.3}$$

Clearly, if event B could never occur (i.e., $\Pr(B) = 0$), there is no reason to define the conditional probability in Eq. (3.3). However, for calculation purposes discussed hereafter we will need to define the conditional probability $\Pr(A|B) \equiv 0$ when $\Pr(B) = 0$.

In the case of newborn babies, one may think that the conditional probability is equal to the probability of a congenital anomaly, but this is only true when the two events congenital anomaly and gender male are independent. Indeed, if the events A and B are independent (and $\Pr(B) > 0$), then $\Pr(A \cap B) = \Pr(A)\Pr(B)$. Using the alternative formulation $\Pr(A \cap B) = \Pr(A|B)\Pr(B)$ of formula Eq. (3.3), the independence of the events A and B results in $\Pr(A|B) = \Pr(A)$. Thus the conditional probability of a congenital anomaly given the child is a boy is then equal to the probability of having a congenital anomaly irrespective of gender.

If we deal with two events A and B, the relevant probabilities can be summarized in the a 2×2 *contingency table* given in Table 3.1. In column A and row B, the probability of the occurrence of both events A and B at the same time is given

Table 3.1 Conditional probabilities in a 2 × 2 contingency table

	A	A^c	
B	$\Pr(A \cap B) = \Pr(A\|B)\Pr(B)$	$\Pr(A^c \cap B) = \Pr(A^c\|B)\Pr(B)$	$\Pr(B)$
B^c	$\Pr(A \cap B^c) = \Pr(A\|B^c)\Pr(B^c)$	$\Pr(A^c \cap B^c) = \Pr(A^c\|B^c)\Pr(B^c)$	$\Pr(B^c)$
	$\Pr(A)$	$\Pr(A^c)$	1

by $\Pr(A \cap B)$. Using the conditional relation Eq. (3.3) it can also be expressed by $\Pr(A|B)\Pr(B)$.[2] For the other three cells the same type of probabilities are presented.

For row B in Table 3.1 the two probabilities in column A and A^c add up to the probability $\Pr(B)$ of event B, since $\Pr(A|B) + \Pr(A^c|B) = 1$. This last relation is due to the fact that if event B has already occurred, the probability for the occurrence of event A or event A^c are given by $\Pr(A|B)$ and $\Pr(A^c|B)$ (see also the second probability rule). Indeed, if we know that the new-born baby is a boy (event B), only the conditional probabilities of congenital anomalies $\Pr(A|B)$ and $\Pr(A^c|B)$ in row B can be observed. If on the other hand we know that the newborn baby is a girl, only the conditional probabilities $\Pr(A|B^c)$ and $\Pr(A^c|B^c)$ in row B^c can be observed. Also note that $\Pr(A|B)\Pr(B)$ and $\Pr(A|B^c)\Pr(B^c)$ add up to $\Pr(A)$, which also follows from the second probability rule after appyling definition Eq. (3.3).

3.4.1 Example: Using Conditional Probabilities

Conditional probabilities play an important role in medical testing where the medical test can produce false negative and false positive test results. In this context, the *sensitivity* of a medical test is the probability of a positive test results (disease indicated) when the patient truly has the disease and the *specificity* of the test is the probability of a negative test result (disease not indicated) for patients without this particular disease.

If event A represents a positive test result (and thus A^c represents an event with a negative test result) in a patient and event B represents the presence of the disease (and thus B^c represents the event with no disease), then the sensitivity and specificity of the medical test are given by the conditional probabilities $\Pr(A|B)$ and $\Pr(A^c|B^c)$, respectively (see Table 3.1).

Suppose that a diagnostic test for the detection of a particular disease has a sensitivity of 0.95 and a specificity of 0.9. Assume further that the proportion of patients with the disease (a priori probability) in the target population is equal to 0.7. Thus in this

[2] Using definition Eq. (3.3) we can write $\Pr(A \cap B)$ as $\Pr(A|B)\Pr(B)$, as we did in Table 3.1, but also as $\Pr(B|A)\Pr(A)$. Which one to use mostly depends on the practical situation. In Table 3.1 we could have used $\Pr(B|A)\Pr(A)$ as well.

example we have received the following information in Table 3.1: $\Pr(A|B) = 0.95$, $\Pr(A^c|B^c) = 0.9$, and $\Pr(B) = 0.7$. This particular example is for instance discussed by Veening et al. (2009, Chap. 13) for the detection of a hernia. Possible questions of interest are the size of the diagnostic probabilities:

- What is the probability of having the disease when the medical test has provided a positive test result: $\Pr(B|A)$?
- What is the probability of not having the disease when the medical test has provided a negative test result: $\Pr(B^c|A^c)$?

The calculation of these diagnostic probabilities can be obtained by the theorem of Bayes, or just *Bayes Theorem*, which is given by

$$\Pr(B|A) = \frac{\Pr(A \cap B)}{\Pr(A)} = \frac{\Pr(A|B)\Pr(B)}{\Pr(A)},$$

and is easily derived using relation Eq. (3.3).

Using relations $\Pr(A) = \Pr(A|B)\Pr(B) + \Pr(A|B^c)\Pr(B^c)$, see Table 3.1, $\Pr(B^c) = 1 - \Pr(B)$, and $\Pr(A|B^c) = 1 - \Pr(A^c|B^c)$, the conditional probability $\Pr(B|A)$ can be rewritten as

$$\Pr(B|A) = \frac{\Pr(A|B)\Pr(B)}{\Pr(A|B)\Pr(B) + [1 - \Pr(A^c|B^c)][1 - \Pr(B)]}. \tag{3.4}$$

The last expression in Eq. (3.4) contains only the terms $\Pr(A|B)$, $\Pr(A^c|B^c)$, and $\Pr(B)$ which were all provided in the example. Thus it becomes possible to calculate the conditional probability $\Pr(B|A)$. The answer is $0.9568 \approx (0.95 \cdot 0.7) / (0.95 \cdot 0.7 + 0.1 \cdot 0.3)$.

For the probability $\Pr(B^c|A^c)$ a similar formula can be established. It is the same formula Eq. (3.4), but with A replaced by A^c, B replaced by B^c, A^c replaced by A, and B^c replaced by B. If we again apply $\Pr(B^c) = 1 - \Pr(B)$, then we find for the conditional probability $\Pr(B^c|A^c)$ the following formula

$$\Pr\left(B^c|A^c\right) = \frac{\Pr(A^c|B^c)[1 - \Pr(B)]}{\Pr(A^c|B^c)[1 - \Pr(B)] + [1 - \Pr(A|B)]\Pr(B)}. \tag{3.5}$$

Thus the answer to the second question above is $0.8852 \approx (0.9 \cdot 0.3) / (0.9 \cdot 0.3 + 0.05 \cdot 0.7)$.

In the general public, there is little intuition about these conditional probabilities $\Pr(B|A)$ and $\Pr(B^c|A^c)$, when $\Pr(A|B)$, $\Pr(A^c|B^c)$, and $\Pr(B)$ are given, due to the role of $\Pr(B)$. If in our example above the probability of the disease is rare, say $\Pr(B) = 0.005$, the probability of having the disease given a positive test $\Pr(B|A)$ would become approximately equal to 0.046, which is still very low. Without the positive test, the probability for an arbitrary person to have the disease would be equal to $\Pr(B) = 0.005$, but if this person receives a positive test, this probability increases to 0.046. Clearly, there is a huge increase in probability (more than 9 times

higher), but since the a priori probability of the disease was very small, the probability on the disease after a positive test remains low.

3.4.2 Computing Probabilities Using R

In Chaps. 1 and 2 we have seen how R can be used to perform descriptive analyses of data and simulate different sampling procedures. R is also a handy calculator for performing simple calculations such as those presented in the examples in Sects. 3.3.1, 3.4.1, and 3.5.4. For instance, consider the example of a medical test from Sect. 3.4.1, where the event A (A^c) represents a positive (negative) test result and the event B (B^c) represents the presence (absence) of a particular disease in a patient. The following probabilities are given: $\Pr(A|B) = 0.95$ (sensitivity of the test), $\Pr(A^c|B^c) = 0.9$ (specificity of the test), and $\Pr(B) = 0.7$ (proportion of patients in the target population having the disease). What we would like to know is the probability of having the disease given that the test is positive, $\Pr(B|A)$, and the probability of not having the disease given that the test is negative, $\Pr(B^c|A^c)$. These probabilities can be calculated using the formulas in Eqs. (3.4) and (3.5), respectively. We can use R to carry out these calculations as follows:

```
> # Specify known probabilities:
> P_A_given_B <- 0.95
> P_notA_given_notB <- 0.9
> P_B <- 0.7
>
> # Calculate Pr(B|A):
> P_A_given_B*P_B/(P_A_given_B*P_B+(1-P_notA_given_notB)*(1-P_B))
[1] 0.9568345
>
> # Calculate Pr(B^c|A^c):
> P_notA_given_notB*(1-P_B)/(P_notA_given_notB*(1-P_B)+
+   (1-P_A_given_B)*P_B)
[1] 0.8852459
```

Note that in the R code above `P_A_given_B` is the sensitivity $\Pr(A|B)$ and `P_notA_given_notB` is the specificity $\Pr(A^c|B^c)$.

3.5 Measures of Risk

The example of conditional probabilities for medical tests is a relevant example of probability theory, but it is of course not the only application of probability theory. Conditional probabilities also play a dominant role in the investigation of associations between events (see, for example Jewell 2003; Rothman et al. 2008; White et al. 2008). Two events or variables are considered *associated* when they are not independent. For instance, the probability of a congenital anomaly in newborn babies

may be different for boys and girls. If independence is violated, a comparison of the two conditional probabilities of congenital anomalies given boys and girls can be used to quantify the strength of the association. Different *association measures* or *measures of risk* exist for this purpose, such as the *risk difference*, *relative risk*, and *odds ratio*. We discuss each of these in turn.

To discuss these measures, it is convenient to change our notation slightly so that it is easier to see what is an *outcome* and what is an *explanatory variable*. Hence, instead of using A and B for events we will use D to denote the event of interest (outcome, result, *disease*) and E to denote the event that may affect the outcome (risk factor, explanatory event, *exposure*). In our example D is the event of the congenital anomaly and E is the event of being male, but one may consider many other examples:

- Lung cancer (D) with smoking (E) or non-smoking (E^c).
- Product failure (D) with automation (E) or manual processing (E^c).
- Passing the data science exam (D) with the use of our book (E) or the use of other books or no books at all (E^c).

3.5.1 Risk Difference

The *risk difference* or *excess risk* is an absolute measure of risk, since it is nothing more than the difference in the conditional probabilities, i.e.

$$ER = \Pr(D|E) - \Pr\left(D|E^c\right).$$

The risk difference is based on an additive model, i.e., $\Pr(D|E) = ER + \Pr(D|E^c)$. It always lies between -1 and 1.

If the risk difference is positive ($ER > 0$), there is a greater risk of the outcome when exposed (E) than when unexposed (E^c). A negative risk difference ($ER < 0$) implies that the exposure (E) is protective for the outcome. If the risk difference is equal to zero ($ER = 0$), the outcome (D) is independent of the exposure (E).

To see this last statement we will assume that $0 < \Pr(E) < 1$ and then use the definition of conditional probability and the calculation rules from Sect. 3.3.

$$\begin{aligned} ER = 0 &\iff \Pr(D|E) = \Pr(D|E^c) \\ &\iff \Pr(E^c)\Pr(D \cap E) = \Pr(E)\Pr(D \cap E^c) \\ &\iff [1 - \Pr(E)]\Pr(D \cap E) = \Pr(E)[\Pr(D) - \Pr(D \cap E)] \\ &\iff \Pr(D \cap E) = \Pr(D)\Pr(E). \end{aligned}$$

Many researchers feel that the risk difference is the most important measure, since it can be viewed as the excess number of cases (D) as a fraction of the population size. If the complete population (of size N) were to be exposed, the number of cases would be equal to $N \cdot \Pr(D) = N \cdot \Pr(D|E)$. If the complete population were unexposed the number of cases would be equal to $N \cdot \Pr(D) = N \cdot \Pr(D|E^c)$. Thus

the difference in these numbers of cases indicates how the number of cases were to change if a completely exposed population would change to a completely unexposed population.

3.5.2 Relative Risk

The *relative risk* would compare the two conditional probabilities $\Pr(D|E)$ and $\Pr(D|E^c)$ by taking the ratio, i.e.

$$RR = \frac{\Pr(D|E)}{\Pr(D|E^c)}.$$

It is common to take as denominator the risk of the outcome D for the unexposed group. Thus a relative risk larger than 1 ($RR > 1$) indicates that the exposed group has a higher probability of the outcome (D) than the unexposed one. A relative risk equal to one ($RR = 1$) implies that the outcome and exposure are independent. A relative risk less than one ($RR < 1$) indicates that the unexposed group has a higher probability of the outcome. The relative risk is based on a multiplicative model, i.e., $\Pr(D|E) = RR \cdot \Pr(D|E^c)$.

3.5.3 Odds Ratio

The third measure of risk is also a relative measure. The *odds ratio* compares the odds for the exposed group with the odds for the unexposed group. The *odds* is a measure of how likely the outcome occurs with respect to not observing this outcome. The odds comes from gambling, where profits of bets are expressed as 1 to x. For instance, the odds of 1 to 6 means that it is six times more likely to loose than to win. The odds can be defined mathematically by $O = p/(1-p)$, with p the probability of winning. The odds of the exposed group is $O_E = \Pr(D|E)/[1 - \Pr(D|E)]$ and the odds for the unexposed group is $O_{E^c} = \Pr(D|E^c)/[1 - \Pr(D|E^c)]$. The odds ratio is now given by

$$OR = \frac{O_E}{O_{E^c}} = \frac{\Pr(D|E)[1 - \Pr(D|E^c)]}{\Pr(D|E^c)[1 - \Pr(D|E)]} = \frac{\Pr(D^c|E^c)}{\Pr(D^c|E)} \times RR. \tag{3.6}$$

Similar to the relative risk, it is common to use the unexposed group as reference group, which implies that the odds of the unexposed group O_{E^c} is used in the denominator.

An odds ratio larger than one ($OR > 1$) indicates that the exposed group has a higher odds than the unexposed group, which implies that the exposed group has a higher probability of outcome D. An odds ratio of one ($OR = 1$) indicates that the

outcome is independent of the exposure, and an odds ratio smaller than one ($OR < 1$) indicates that the unexposed group has a higher probability of the outcome.

Note that the odds ratio and relative risk are always ordered. To be more precise, the odds ratio is always further away from 1 than the relative risk, i.e., $1 < RR < OR$ or $OR < RR < 1$. To see this, we will only demonstrate $1 < RR < OR$, since the other ordering $OR < RR < 1$ can be demonstrated in a similar way. If $RR > 1$, we have that $\Pr(D|E) > \Pr(D|E^c)$, using its definition. Since $\Pr(D^c|E) = 1 - \Pr(D|E)$ and $\Pr(D^c|E^c) = 1 - \Pr(D|E^c)$, we obtain that $\Pr(D^c|E) < \Pr(D^c|E^c)$. Combining this inequality with the relation in Eq. (3.6), we see that $OR > RR$. Note that the odds ratio and relative risk are equal to each other when $RR = 1$ (or $OR = 1$).

3.5.4 Example: Using Risk Measures

To illustrate the calculations of the different risk measures we will consider the example of Dupuytren disease (outcome D) and discuss whether gender has an influence on this disease (E is male). Dupuytren disease causes the formation of nodules and strains in the palm of the hand that may lead to flexion contracture of the fingers. Based on a random sample of size $n = 763$ from the population of Groningen, a contingency table with Dupuytren disease and gender was obtained (Table 3.2). More can be found in Lanting et al. (2013).

The probabilities in the middle four cells in Table 3.2 (thus not in the bottom row nor in the last column) represent the probabilities that the two events occur together. Thus the probabilities $\Pr(D|E)$ and $\Pr(D|E^c)$ can be obtained by $\Pr(D|E) = \Pr(D \cap E) / \Pr(E) = 92/348 = 0.2644$ and $\Pr(D|E^c) = \Pr(D \cap E^c) / \Pr(E^c) = 77/415 = 0.1855$, respectively.

Given this information we can compute the risk difference, the relative risk, and the odds ratio:

- The risk difference is now $ER = 0.0788$, which implies that males have 7.88% absolute higher risk of Dupuytren disease than females.
- The relative risk is $RR = (92/348) / (77/415) = 1.4248$. This implies that males have a risk of Dupuytren disease that is almost 1.5 times larger than the risk for females.

Table 3.2 2×2 contingency table for Dupuytren disease and gender

Exposure	Disease outcome		Total
	Dupuytren (D)	No Dupuytren (D^c)	
Male (E)	$92/763 = 0.1206$	$256/763 = 0.3355$	$348/763 = 0.4561$
Female (E^c)	$77/763 = 0.1009$	$338/763 = 0.4430$	$415/763 = 0.5439$
Total	$169/763 = 0.2215$	$594/763 = 0.7785$	1

- The odds ratio of Dupuytren disease for males is $OR = (92 \times 338)/(77 \times 256) = 1.5775$, indicating that the odds for Dupuytren disease in males is more than 1.5 time larger than the odds for females. Thus males have a higher risk for Dupuytren disease.

In the next section we discuss what measures of association or risk we can calculate if we sample from the population in three different ways. Each of the sampling approaches will provide a 2×2 contingency table, just like the one in Table 3.2, but they may not provide estimates of the population proportions.

3.6 Sampling from Populations: Different Study Designs

The odds ratio is often considered more complex than the relative risk, in particular because of the simplicity of interpretation of the relative risk. The odds ratio is, however, more frequently used in practice than the relative risk. An important reason for this is that the odds ratio is symmetric in exposure E and outcome D. If the roles of the exposure and outcome are interchanged the odds ratio does not change, but the relative risk does.

To see this, we will again use the data presented in Table 3.2. Interchanging the roles of E and D results in a relative risk of $\Pr(E|D) / \Pr(E|D^c)$. This relative risk is equal to $(92/169)/(256/594) = 1.2631$ and it is quite different from the relative risk $\Pr(D|E) / \Pr(D|E^c) = 1.4248$. When the roles of E and D are interchanged, the odds ratio becomes $[\Pr(E|D) /(1 - \Pr(E|D))]/[\Pr(E|D^c) /(1 - \Pr(E|D^c)]$. Calculating this odds ratio results in the odds ratio of 1.5775 (as we already calculated in Sect. 3.5.4), since $[(92/169)/(77/169)]/[(256/594)/(338/594)] = (92 \times 338)/(77 \times 256)$. These results can be proven mathematically; we ask you to do so in the assignments.

As we will see hereafter, the symmetry of the odds ratio makes it possible to investigate the association between D and E irrespective of the way that the sample from the population was collected. This would not be the case for the risk difference and the relative risk. There are many ways in which we can select a sample from a population, but three of them are particularly common in medical research: *population-based* (cross-sectional), *exposure-based* (cohort study), and *disease-based* (case-control study). We discuss these three in turn, and also discuss their limitations with respect to calculating the different risk or association measures if there are any.

3.6.1 Cross-Sectional Study

In a cross-sectional study a simple random sample of size n is taken from the population (see Chap. 2). For each unit in the sample both the exposure and outcome are being observed and the units are then summarized into the four cells (E, D), (E, D^c),

(E^c, D), and (E^c, D^c). The 2×2 contingency table would then contain the number of units in each cell, just like we saw in Table 3.2 for Dupuytren disease.

This way of sampling implies that the proportions in the last row ($\Pr(D)$ and $\Pr(D^c)$) and the proportions in the last column ($\Pr(E)$ and $\Pr(E^c)$) of Table 3.1 would be unknown before sampling and they are being determined by the probability of outcome and exposure in the population. The example of Dupuytren disease in Table 3.2 was actually obtained with a population-based sample. Thus the observed probabilities in Table 3.1 obtained from the sample represent unbiased estimates of the population probabilities .

Here we have applied the theory of simple random sampling of Sect. 2.7 for estimation of a population proportion. For instance, if we define the binary variable x_i by 1 if unit i has both events E and D (thus $E \cap D$) and it is zero otherwise, the estimate of the population proportion $\Pr(E \cap D)$ would be the sample average of this binary variable. This sample average is equal to the number of units in cell (E, D) divided by the total sample size n; see also Table 3.2. This would also hold for any of the other cells, including the cells in the row and column totals ($\Pr(D)$, $\Pr(D^c)$, $\Pr(E)$, and $\Pr(E^c)$). Thus if we also want to express the mean squared error (MSE) for estimating any of the probabilities in Table 3.1, we could apply the MSE from Table 2.2 (see Sect. 2.2). Since the 2×2 contingency table with sample data provides proper estimates of the population proportions, the measures of risk that would use these estimates from the sampled contingency table are estimates of the population measures of risk. Thus calculation of the risk difference, the relative risk, and the odds ratio are all appropriate for cross-sectional studies.

3.6.2 Cohort Study

In a cohort study, a simple random sample is taken from the population of units who are exposed and another simple random sample is taken from the population of units who are unexposed. Thus this way of sampling relates directly to stratified sampling discussed in Chap. 2 with the strata being the group of exposed (E) and the group of unexposed (E^c). In each sample or stratum the outcome D is noted and the contingency table in Table 3.1 is filled. In this setting, the probabilities $\Pr(E)$ and $\Pr(E^c)$ are preselected before sampling and are fixed in the sample, whatever they are in the population. Thus the sample and the population may have very different probabilities.

To illustrate this we consider the example of newborn babies again. The outcome will be the occurrence of congenital anomalies (D) and the exposure would represent the gender of the child, with male being the event (E). If we select a random sample of 500 male newborn babies and 1,000 female newborn babies, the observed contingency table would have twice as many girls as boys, but in practice this ratio is approximately one. Thus a consequence is that the probabilities in the cells of the contingency tables are no longer appropriate estimates for the population probabilities, since we have destroyed the ratio in probabilities for E and E^c. This also implies

that the probability $\Pr(D \cap E)$ in Table 3.1 does not reflect the true probability in the population either. This would become even more obvious if we assume that the exposure E is really rare in the population ($\Pr(E) \approx 0$) and all units with the exposure also have the outcome D. In the sample we would observe that $\Pr(D \cap E)$ is equal to one, while in the population this probability is close to zero since the exposure hardly occurs.[3]

Despite the fact that we cannot use the joint probabilities in the contingency table as estimates for the population probabilities, the risk difference, the relative risk, and the odds ratio in the sample are all appropriate estimates for the population when a cohort study is used. The reason is that these measures use the conditional probabilities only, where conditioning is done on the exposure. The $\Pr(D|E)$ and $\Pr(D|E^c)$ in the sample do represent the conditional population probabilities.

3.6.3 Case-Control Study

In a case-control study a simple random sample is taken from the population of units having the outcome and from the population of units without the outcome. Thus this way of sampling relates also directly to stratified sampling discussed in Chap. 2 with the strata being the group with outcome (D) and the group without outcome (D^c). In each sample or stratum the exposure of each unit is noted. Thus for disease-based sampling the probabilities $\Pr(D)$ and $\Pr(D^c)$ are known before sampling and are fixed in the sample. This means that the observed probabilities in the sample are inappropriate as estimates for the same probabilities in the population. Thus we cannot estimate how many units in the population have the outcome. Similar to the cohort study, we cannot estimate the joint probabilities $\Pr(D \cap E)$, $\Pr(D \cap E^c)$, $\Pr(D \cap E)$, and $\Pr(D \cap E)$ in the population from the sample. This is similar to the discussion in cohort studies.

The problem with case-control studies is that the conditional probabilities $\Pr(D|E)$ and $\Pr(D|E^c)$ cannot be determined either. To illustrate this, assume the following probabilities in the population $\Pr(D \cap E) = 0.08$, $\Pr(D \cap E^c) = 0.02$, $\Pr(D^c \cap E) = 0.12$, and $\Pr(D^c \cap E^c) = 0.78$. Thus in the population we have $\Pr(D|E) = 0.4$ and $\Pr(D|E^c) = 0.025$, which gives a relative risk of $RR = 16$. Now let us assume that the sample size in the outcome D group is equal to 900 and it is the same as in the D^c group. We would expect the following numbers in the 2×2 contingency Table 3.3, because $\Pr(E|D) = 0.8$ and $\Pr(E|D^c) = 0.133$.

The conditional probabilities $\Pr(D|E)$ and $\Pr(D|E^c)$ are now equal to $\Pr(D|E) = 720/840 = 0.8571$ and $\Pr(D|E^c) = 180/960 = 0.1875$. The relative risk for Table 3.3 is now given by $RR = 4.5714$, which is substantially lower than the relative risk in the population. The odds ratio, though, can still be properly estimated, due to the symmetry of the odds ratio, which does not change if the roles of D and E interchange.

[3] If, in this case, the population size(s) were known, we could calculate weighted averages to estimate the population parameters as we did in Chap. 2.

Table 3.3 2×2 contingency table for an artificial case-control study

Exposure	Disease outcome		Total
	D	D^c	
Male	720	120	840
Female	180	780	960
Total	900	900	1800

3.7 Simpson's Paradox

In Sect. 3.5 we discussed three different measures of risk for two types of events (outcome D and exposure E). In Sect. 3.6 we discussed three different observational study designs and demonstrated that not all three measures can be determined in each of these observational studies. However, in practice it is even more complicated, since we should always be aware of a third event C that may change the conclusions if the event data of D and E are split for C and C^c. This issue is best explained through an example. In Table 3.4 we report the numbers of successful removal of kidney stones with either percutaneous nephrolithotomy or open surgery (see Charig et al. 1986 for more details).

The relative risk of removal of kidney stones for percutaneous nephrolithotomy with respect to open surgery (which can be calculated from this dataset, as it is a cohort study) is determined by $RR = (289/350)/(273/350) = 1.0586$. This means that percutaneous nephrolithotomy increases the "risk" of successful removal of the kidney stones with respect to open surgery. However, if the data are split by size of kidney stone, a 2×2 contingency table for stones smaller and larger than 2 cm in diameter can be created. Let's assume that we obtain the two 2×2 contingency tables in Table 3.5. Note that if we combine the two tables into one 2×2 contingency table we obtain Table 3.4. The relative risks for the two sizes of kidney stones separately are determined at $RR_{\leq 2} = 0.9309$ and $RR_{>2} = 0.9417$. Thus it seems that open surgery has a higher success of kidney stone removal than percutaneous nephrolithotomy for both small and large stones. This seems to contradict the results from Table 3.4 and this contradiction is called Simpson's paradox (Simpson 1951), named after Edward Hugh Simpson.

Table 3.4 2×2 contingency table for removal of kidney stones and two surgical treatments

Exposure	Kidney stones outcome		Total
	Removal (D)	No removal (D^c)	
Nephrolithotomy (E)	289	61	350
Open surgery (E^c)	273	77	350
Total	562	138	700

Table 3.5 2×2 contingency table for removal of kidney stones and two surgical treatments by size of kidney stones

Exposure	Kidney Stones Outcome $\leq$ 2cm (C)		Total
	Removal (D)	No Removal (D^c)	
Nephrolithotomy (E)	234	36	270
Open Surgery(E^c)	81	6	87
Total	315	42	357

Exposure	Kidney Stones Outcome $>$ 2cm (C^c)		Total
	Removal (D)	No Removal (D^c)	
Nephrolithotomy (E)	55	25	80
Open Surgery(E^c)	192	71	263
Total	247	96	343

Simpson (1951) demonstrated that the association between D and E in the collapsed contingency table is preserved in the two separate contingency tables for C and C^c whenever one or both of the following restrictions hold true

$$\Pr(D \cap E \cap C) \Pr\left(D \cap E^c \cap C^c\right) = \Pr\left(D \cap E^c \cap C\right) \Pr\left(D \cap E \cap C^c\right)$$

$$\Pr(D \cap E \cap C) \Pr\left(D^c \cap E \cap C^c\right) = \Pr\left(D^c \cap E \cap C\right) \Pr\left(D \cap E \cap C^c\right)$$

The first equation implies that the odds ratio for having the exposure E for the presence or absence of C in the outcome group D is equal to one, i.e.

$$OR_{EC|D} = \frac{\Pr(E|C, D)\,[1 - \Pr(E|C^c, D)]}{\Pr(E|C^c, D)\,[1 - \Pr(E|C, D)]} = 1.$$

Thus E and C must be independent in the outcome group D, which means that $\Pr(E \cap C|D) = \Pr(E|D) \Pr(C|D)$. The second equation implies that the odds ratio for the outcome D in the presence or absence of C in the exposed group E is equal to one, i.e.

$$OR_{DC|E} = \frac{\Pr(D|C, E)\,[1 - \Pr(D|C^c, E)]}{\Pr(D|C^c, E)\,[1 - \Pr(D|C, E)]} = 1.$$

Thus this means that D and C are independent in the exposed group E, which means that $\Pr(D \cap C|E) = \Pr(D|E) \Pr(C|E)$.

In the example of kidney stones, we see that $\Pr(E \cap C|D) = 234/562$, $\Pr(E|D) = 289/562$, and $\Pr(C|D) = 315/562$. The product of probabilities $\Pr(E|D) \Pr(C|D) = 0.2882$, which is substantially lower than $\Pr(E \cap C|D) = 0.4164$. Additionally, we also obtain $\Pr(D \cap C|E) = 234/350$, $\Pr(D|E) = 289/350$, and $\Pr(C|E) = 270/350$. This shows that $\Pr(D|E) \Pr(C|E) = 0.6370$, which is lower than $\Pr(D \cap C|E) = 0.6686$.

If the two independence requirements are violated, the event C is called a *confounder*. In this case we should report the stratified analysis. Thus for the example

of kidney stone removal, the analysis should be conducted on the data in Table 3.5. This means that open surgery has a slightly higher success rate than percutaneous nephrolithotomy.[4]

Simpson's paradox also shows that data analysis is far from trivial and care should be taken when making bold statements about associations of events in populations.

3.8 Conclusion

In this chapter we started our exploration of the theory of probability. To do so, we defined probabilities, and we gave the basic computation rules to work with probabilities. We discussed probabilities (in terms of events) and several derived quantities that are used in practice to summarize data, such as distinct risk measures. Also, we discussed how sampling (or study design) and appropriate risk measures are closely related.

The probability rules we discuss in this chapter provide the foundation for discussing more probability theory; namely the theory of random variables. We will do so in the next chapter. In the additional materials for this chapter you will find a short history of probability. It is interesting to see that the same rules have originated multiple times, using different definitions of probabilities. For now we will continue using our definitions presented here. However, in Chap. 8 we will get back to some of the fundamental discussions and discuss the important role that Eq. (3.4) has in thinking about probabilities.

Problems

3.1 Two fair dice are thrown one by one.

1. What is the probability that the first die shows an odd number of eyes facing up?
2. What is the probability that the sum of the eyes of the two dice is eleven?

3.2 A card is randomly drawn from an incomplete deck of cards from which the ace of diamonds is missing.

1. What is the probability that the card is "clubs"?
2. What is the probability that the card is a "queen"?
3. Are the events "clubs" and "queen" independent?

3.3 In a group of children from primary school there are 18 girls and 15 boys. Of the girls, 9 have had measles. Of the boys, 6 have had measles.

[4] Note that Simpson's Paradox, and its solutions, are still heavily debated (see, Armistead 2014 for examples).

1. What is the probability that a randomly chosen child from this group has had measles?
2. If we randomly choose one person from the group of 18 girls, what is the probability that this girl has had measles?
3. Are the events "boy" and "measles" in this example independent?

3.4 In a Japanese cohort study, 5,322 male non-smokers and 7,019 male smokers were followed for four years. Of these men, 16 non-smokers and 77 smokers developed lung cancer.

1. What is the probability that a randomly chosen non-smoker from this group developed lung cancer?
2. What is the probability that a randomly chosen smoker from this group developed lung cancer?
3. Are the events "smoking" and "lung cancer" in this example independent?
4. What is the conditional probability that the patient is a smoker if he has developed lung cancer?

3.5 Prove mathematically that $A \perp B^c$, $A^c \perp B$, and $A^c \perp B^c$ if $A \perp B$.

3.6 In a life table the following probabilities are provided. Females can expect to live to an age of 50 years with a probability of 0.898. The probability drops to 0.571 for females with a life expectancy of 70 years. Given that a woman is 50 years old, what is the probability that she lives to an age of 70 years?

3.7 Suppose a particular disease is prevalent in a population with 60%. The sensitivity and specificity of the medical test for this disease are both 0.9. A patient from this population is visiting the physician and is tested for the disease.

1. What is the probability that the patient has the disease when the patient is tested positively?
2. If the sensitivity is 0.9, what is the minimum required specificity of the medical test to know with at least 95% certainty that the patient has the disease when tested positively?
3. If the specificity is 0.9, what is the minimum required sensitivity of the medical test to know with at least 95% certainty that the patient does not have the disease when tested negatively?

3.8 Use R to carry out the calculations presented in Sect. 3.5.4. First, use the `matrix()` function to store the numbers presented in Table 3.2 in a 3×3 matrix. Use the `dimnames` argument of the `matrix()` function to give the matrix meaningful row and column names. If you do not know how to use the `dimnames` argument try running `?matrix()` or search the Internet. Use the numbers stored in the matrix to calculate the risk difference, the relative risk, and the odds ratio. Give an interpretation of the results.

3.9 Consider the following 2×2 contingency table for the removal of kidney stones using two different treatments:

Treatment	Removal of kidney stones		Total
	Successful	Not successful	
Open surgery	273	77	350
Small incision	289	61	350
Total	552	148	700

1. What do you think is the study design that the researchers of the removal of kidney stones have selected?
2. Calculate the risk difference, the relative risk, and the odds ratio for a successful removal of kidney stones for a small incision with respect to open surgery. Based on these results, formulate your conclusion.
3. Prove mathematically that the odds ratio for outcome D with and without exposure E is the same as the odds ratio for the exposure E with and without the outcome D.

Additional Material I: A Historical Background of Probability

The theory of probability was inspired by and has its origin in gaming and gambling. In the 16th century, the Italian mathematician Girolamo Gardano (1501–1576) is considered the first to have calculated probabilities by theoretical arguments and possibly started the development of modern probability (see David (1955)). Also Galileo-Galilei (1564–1642) discussed probabilities, in particular for throwing three dice, but he may have thought that the problem was of little interest. The two French mathematicians Blaise Pascal (1623–1662) and Pierre de Fermat (1601–1665) discussed more complex calculations of popular dice games in a set of letter correspondences. They are often credited with the development of the first fundamental principles for probability theory. Their correspondence was probably initiated by a seeming contradiction presented by the French nobleman Chevalier de Méré, who believed that scoring a six once in four throws of a die is equal to scoring a double six simultaneously in 24 throws of two dice (see Sheynin (1977)). As we now know, these probabilities are approximately 0.5177 and 0.4914, respectively.

The Dutch scientist Christiaan Huygens (1629–1695) is known to be the first to write a book solely on the subject of probability in 1657, in which a systematic manner of probability calculations was set out for gambling questions. Although he might not have met Pascal or Fermat, he was probably introduced to the theory of probability when he spent time in Paris in 1655. It is believed that the manuscript of Huygens initiated much more interest in the theory of probability. It profoundly influenced two other important contributors to this theory, namely the Swiss mathematician James Bernoulli (1654–1705) and the French mathematician Abraham de

Moivre (1667–1754). They both contributed to probability theory by introducing more complicated calculations in gambling questions, but Bernoulli also provided a philosophical foundation that would make probability suitable for broader applications.

The French mathematician and astronomer Pierre-Simon Laplace (1749–1827) took this development much further and applied probability theory to a host of new applications. It can be claimed that Laplace is responsible for the early development of mathematical statistics (see Stigler (1975)). In 1812 he published the first edition of a book on probability theory with a wide variety of analytical principles. For instance, he presents and applies the theorem of Bayes. This important calculation rule in probability, which plays a role in, for instance, diagnostic tests, as we discussed before, is credited to the English mathematician Thomas Bayes (1702–1761). It was published posthumously in 1764, but it did not receive much attention until Laplace published it in his book. It is unknown if Laplace was aware of the publication in 1764.

A more fundamental mathematical formulation of the definition of probability was developed by the Russian mathematician Andrey Nikolaevich Kolmogorov (1903–1987), who built upon theoretical results of other mathematical scientists. One could claim that the work of Kolmogorov ended the search for a precise mathematical definition of probability that is also comprehensive enough to be useful to describe a large set of practical phenomena.

Additional Material II: A Formal Definition of Probability

The first step is to introduce an outcome space Ω of elementary events (Grimmett et al. 2001). This is the set of outcomes that we may (theoretically) observe. For example, the outcome space Ω can be equal to $\Omega = \{1, 2, 3, 4, 5, 6\}$ if we through a die and each side of the die can finish on top. Then in the second step we need to define what is called a σ-field (or σ-algebra) $\mathscr{F}$. This is a set of subsets of the outcome space Ω. It needs to satisfy the following conditions:

1. The empty set $\emptyset$ must be an element of $\mathscr{F}$. Thus $\emptyset \in \mathscr{F}$.
2. The union of any number of subsets of $\mathscr{F}$ should be part of $\mathscr{F}$. If $A_1, A_2, \ldots\ldots \in \mathscr{F}$ then $\cup_{i=1}^{\infty} A_i \in \mathscr{F}$.
3. The complement of any subset in $\mathscr{F}$ is part of $\mathscr{F}$. If $A \in \mathscr{F}$ then $A^c \in \mathscr{F}$.

Note that we may define different σ-fields on the same outcome space. For instance, for the outcome space $\Omega = \{1, 2, 3, 4, 5, 6\}$ we could define $\mathscr{F} = \{\emptyset, \{6\}, \{1, 2, 3, 4, 5\}, \Omega\}$ as the σ-field. This σ-field shows that we are interested in the event of throwing a six. We may also be interested in the event of throwing an odd number, which would imply that the σ-field is equal to $\mathscr{F} = \{\emptyset, \{1, 3, 5\}, \{2, 4, 6\}, \Omega\}$. Alternatively, we may be interested in throwing any outcome, which means that the σ-field would contain any possible subset of Ω. Thus $\mathscr{F}$ would be equal to

$$\mathscr{F} = \{\emptyset, \{1\}, \{2\}, \ldots, \{6\}, \{1,2\}, \{1,3\}, \ldots, \{5,6\}, \{1,2,3\}, \{1,2,4\}, \ldots, \{4,5,6\}, \ldots, \{1,2,3,4,5,6\}\}.$$

Note that all three σ-fields satisfy the conditions listed above. Thus the σ-field determines what sort of events we are interested in. The σ-field relates to the question or to the probability of interest.

Now the final step is to define the probability measure Pr on $(\Omega, \mathscr{F})$ as a function from the σ-field $\mathscr{F}$ to the interval [0, 1], i.e., $\Pr : \mathscr{F} \to [0.1]$, that satisfies:

1. $\Pr(\emptyset) = 0$.
2. If $A_1, A_2, \ldots$ is a collection of disjoint members of $\mathscr{F}$, i.e., $A_i \in \mathscr{F}$ and $A_i \cap A_j = \emptyset$ for $i \neq j$, then $\Pr\left(\cup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \Pr(A_i)$.

The triplet $(\Omega, \mathscr{F}, \Pr)$ is called a *probability space* and Pr is called the *probability measure*.

Thus if we believe that throwing a six with one die is equal to $1/6$, the probability space may be written as $(\{1,2,3,4,5,6\}, \{\emptyset, \{6\}, \{1,2,3,4,5\}, \Omega\}, \Pr)$ with $\Pr(\{6\}) = 1/6$. Alternatively, if we do not know the probability of throwing a 6, we may introduce an unknown probability θ for the probability of throwing a 6, i.e., $\Pr(\{6\}) = \theta$. The probability space remains what it is, but Pr is changed now.

References

T.W. Armistead, Resurrecting the third variable: a critique of pearl's causal analysis of Simpson's paradox. Am. Stat. **68**(1), 1–7 (2014)

C.R. Charig, D.R. Webb, S.R. Payne, J.E. Wickham, Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy. Br. Med. J. (Clin. Res. Ed.) **292**(6524), 879–882 (1986)

G. Grimmett, D. Stirzaker et al., *Probability and Random Processes* (Oxford University Press, Oxford, 2001)

N.P. Jewell, *Statistics for Epidemiology* (Chapman and Hall/CRC, Boca Raton, 2003)

R. Lanting, E.R. Van Den Heuvel, B. Westerink, P.M. Werker, Prevalence of dupuytren disease in the Netherlands. Plast. Reconstr. Surg. **132**(2), 394–403 (2013)

K.J. Rothman, S. Greenland, T.L. Lash et al., *Modern Epidemiology*, vol. 3 (Wolters Kluwer Health/Lippincott Williams & Wilkins, Philadelphia, 2008)

E.H. Simpson, The interpretation of interaction in contingency tables. J. Roy. Stat. Soc.: Ser. B (Methodol.) **13**(2), 238–241 (1951)

E.P. Veening, R.O.B. Gans, J.B.M. Kuks, *Medische Consultvoering* (Bohn Stafleu van Loghum, Houten, 2009)

E. White, B.K. Armstrong, R. Saracci, *Principles of Exposure Measurement in Epidemiology: Collecting, Evaluating and Improving Measures of Disease Risk Factors* (OUP, Oxford, 2008)

F.N. David, Studies in the History of Probability and Statistics I. Dicing and Gaming (A Note on the History of Probability). *Biometrika*, **42**(1/2), 1–5 (1955)

O.B. Sheynin, Early history of the theory of probability. Archive for History of Exact Sciences, **17**(3), 201–259 (1977)

S.M. Stigler, Studies in the History of Probability and Statistics. XXXIV: Napoleonic statistics: The work of Laplace. *Biometrika*, **62**(2), 503–517 (1975)