Chapter 1

Introduction

1.1 What are probability and statistics

Statistics. Statistics is the science of extracting reliable information from empirical data. The main idea behind statistics is that if we have enough evidence about a system (in the form of statistical data), then that evidence would contain enough information to allow us describe, predict and control the system.

Describing and predicting phenomena based on empirical evidence is the objective of all sciences; controlling them is the ambition of engineering. The common approach of physics (and other mathematical sciences) to these objectives is to build mathematical models. These models often provide quantitative mechanistic descriptions of natural phenomena in terms of their simpler constituents. Statistics merely takes a different approach to these same goals. Instead of seeking a mechanistic explanation of a system, statisticians describe a system by interpolating a large amount of data gathered from the same system.

Compared to the traditional approach of physics and other mathematical sciences, the approach of statistics has both advantages and disadvantages:

- *Advantages:*
  - Statistical analysis can be automated, and then performed in massive scales using computers.
  - Statistical analysis can be applied to complex systems (such as social and biological systems), where the approach of mathematical sciences still has a long way to go.
- *Disadvantages:*
  - Statistical analysis does not lead to a deep understanding of the phenomena the way models in physics do.
  - The conclusions of a statistical analysis of a data set apply only to the system from which the data comes from, and cannot be reliably extrapolated.

The data used in a statistical analysis often comes from a sample of the entities in a population, rather than the entire population itself. The information inferred from such a sampled data is inevitably subject to uncertainty. In order for it to be of any use, it is thus important that such information is accompanied by a quantitative measure of its uncertainty. The degree of uncertainty of the information naturally depends on the quantity (*how large is the sample*) and quality (*how representative is the sample*) of the sampled data.

Probability theory. Probability theory is the mathematics of reasoning about *chance* and *randomness*. As such, it is a branch of mathematics in the same way that calculus, number theory and geometry are.

But what is *chance*, and what do we mean by *randomness*? Most people have an intuitive conception of these two notions which is mirrored in everyday language. For instance, we understand the following statement:

- If we flip a coin, whether it comes up heads or tails is a matter of chance. In other words, the outcome of this experiment is random.

We even understand more quantified statements such as the following:

⊛
- If we flip a coin, then the chance of it coming up heads is the same as the chance of it coming up tails.

Still, upon further thought, we realize that it is not entirely clear what is exactly meant by such a statement. After all, the trajectory of the coin starting from when it is flipped by your thumb till it lands on the floor is governed by the laws of physics, and if we know the initial state of the coin (its position, its orientation, the momentum exerted by the thumb, ...), then at least *in principle* we should be able to write down the equations of its motion and solve them in order to exactly predict the face on which the coin lands.[1] If everything is predetermined by a set of equations, then what is meant by *chance*?

This is indeed a deep question which has kept philosophers, physicists and mathematicians puzzled for many years.[2] Two classical interpretations of a statement such as $\circledast$ are the following:

- *Subjective interpretation:*
  Chance is a way of describing our ignorance or uncertainty. For instance, when we flip a fair coin, we have no *practical* way of predicting the outcome. Assertion $\circledast$ merely reflects this uncertainty and the fact that we have no reason to presume one outcome is likelier than the other.

- *Frequentist interpretation:*
  Chances indicate idealized frequencies. For instance, assertion $\circledast$ simply expresses the idea that if we repeat flipping the same coin many many times, about half the time it should come up heads and the other half tails.

In this course, we will primarily rely on the intuition provided by the frequentist interpretation. However, you may notice that, in some examples we will encounter, this interpretation breaks down and becomes nonsensical.

Connection between probability theory and statistics. Although probability theory and statistics are two distinct disciplines, they are inherently intertwined. The link between the two goes in both directions:

- *Probability theory → Statistics*
  - Statistical analysis is often primarily based on probabilistic models of data.
  - Probability theory can be used to explain deep statistical phenomena.[3]
- *Statistics → Probability theory*
  - Statistics provides intuition and interpretation for probability theory.
  - Many of the problems studied in probability theory are originated from statistical applications.

1.2 Teaser: four examples

The following examples are chosen to give you a flavor of this course. The first two are within the domain of probability theory, while the last two concern typical questions in statistics.

Example 1.2.1 (Flipping a coin)**.** This example is about the experiment of flipping a *fair* coin. By *fairness* we simply mean that the coin is not loaded one way or the other: each time we flip the coin, it is equally likely for it to come up heads or tails.

Suppose we flip our fair coin 10 times. Here is one possible outcome:

```
T H H T T T T H T H
```

Ⓠ What is the chance that in the first two flips we get heads?

$\boxed{\text{A}}$ $1/4$. In the first flip, we have $1/2$ chance of getting a head, and the same for the second flip.

Ⓠ What is the chance that all 10 flips show tails?

$\boxed{\text{A}}$ $1/2 \times 1/2 \times \cdots \times 1/2 = (1/2)^{10}$.
Alternatively, there are 2^{10} possible outcomes all of which are equally likely. Thus, the chance of each individual outcome is $1/2^{10}$.

Ⓠ What is the chance that we get *exactly* 4 heads out of 10 flips?

[1] This is purely hypothetical. In order for such a prediction to be accurate, we may need to include many many details, such as the atomic structure of the coin, the interactions between the coin and the air molecules, the perturbations in the gravity of the earth, the position of the moon, and so on and so forth. Moreover, we may need to know the initial state of the coin very very accurately.

[2] For more reading, see the entries “Chance versus Randomness” (by A. Eagle) and “Interpretations of Probability” (by A. Hájek) in *The Stanford Encyclopedia of Philosophy* (Fall 2019 Edition).

[3] Two such explanations are provided in the *law of large numbers* and the *central limit theorem*, which we will discuss later.

$\boxed{\text{A}}$ Let us list all the possible outcomes with exactly 4 heads:

$$\begin{gathered}\texttt{H H H H T T T T T T}\\ \texttt{H H H T H T T T T T}\\ \cdots\\ \texttt{T T T T T T H H H H}\end{gathered}$$

There are exactly $\binom{10}{4} = \frac{10!}{4!6!} = 210$ such outcomes,[4] each of which has probability $(1/2)^{10}$ of occurring. Thus, in total, the chance of getting exactly 4 heads is $\binom{10}{4}(1/2)^{10} \approx 0.205$.

Ⓠ What is the chance that we get more heads than tails?

$\boxed{\text{A1}}$ $\left[\binom{10}{6} + \binom{10}{7} + \binom{10}{8} + \binom{10}{9} + \binom{10}{10}\right](1/2)^{10}$.

Why? In order to have more heads than tails, the number of heads must either be 6, or 7, . . . or 10. Thus, the total chance is the sum of the chances of these possibilities.

Here is an alternative answer:

$\boxed{\text{A2}}$ $\frac{1}{2}\left(1 - \binom{10}{5}\frac{1}{2^{10}}\right) \approx 0.378$.

Why? There are three possibilities:

- more heads than tails,
- more tails than heads,
- equal number of heads and tails.

By symmetry, the first two possibilities are equally likely. Thus, the chance of having more heads than tails is simply half the chance of having unequal number of heads and tails. The chance of having equal number of heads and tails (i.e., 5 heads and 5 tails) is $\binom{10}{5}(1/2)^{10}$. Hence the chance of having unequal number of heads and tails is $1 - \binom{10}{5}(1/2)^{10}$. ○

Exercise. Using your knowledge of the binomial coefficients, verify that the above two answers are the same.

Example 1.2.2 (Genes and illnesses)**.** In this example, we consider a hypothetical scenario from medical research. Suppose that every person has a gene that can be either of type B (standing for blue) or R (standing for red). Scientists have pinpointed a possible connection between the presence of gene B and a certain disease X.

Based on statistical data, we know that

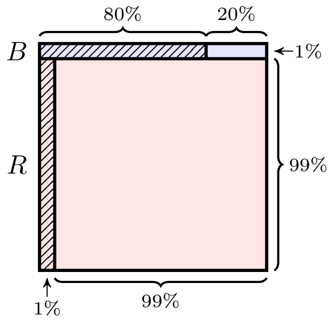
- About 1% of people have gene B.
- Among those with gene B, about 80% get disease X at some point throughout their lives.
- Among those with gene R, only 1% get disease X at some point throughout their lives.

This shows that the presence of gene B is a strong indicator of whether a person gets disease X or not: the chance of getting disease X is 80 times more for a person with gene B compared to a person with gene R!

Now, suppose that a random person is diagnosed with disease X.

Ⓠ What is the chance that the person has gene B?

$\boxed{\text{A}}$ Drawing a Venn diagram may help:



[4] Recall that $\binom{10}{4}$ denotes the number of ways one can choose 4 distinct objects from a collection of 10 distinguishable objects. The notation $\binom{10}{4}$ is read as 10 *choose* 4.

The shaded region represents the people with disease X. The area of each region is meant to be proportional to the ratio of people in the category represented by the region. However, note that the diagram is not to scale. Now, it should be clear that:

area of blue shaded region

area of shaded region

$$\begin{aligned}\langle\text{chance of } B \text{ if we know } X\rangle &= \frac{\langle\text{ratio of people with } B \text{ and } X\rangle}{\langle\text{ratio of people with } X\rangle} \\ &= \frac{\frac{1}{100}\times\frac{80}{100}}{\frac{1}{100}\times\frac{80}{100}+\frac{99}{100}\times\frac{1}{100}} = \frac{80}{179} \approx 0.447 < {}^1\!/_2\,.\end{aligned}$$

The result may come as a surprise: while gene B is a strong indicator of disease X, the occurrence of disease X is by no means an indicator of the presence of gene B! However, a little thought reveals no paradox: the overall ratio of people with gene B is so small that even if we know a person has disease X, it is still more likely that the person is among the larger group of people with gene R. ○

The distinction between what is known and what is uncertain is a common source of confusion in estimating chances in everyday life. Later on, we will talk about the so-called *Bayes' rule* which makes the above type of computations more streamlined.

Example 1.2.3 (Judging fairness)**.** Let us get back to flipping coins. For a *fair* coin, the two outcomes H and T are equally likely: each occurs with probability $^1\!/_2$. So, if we flip the same fair coin a large number of times, we expect to see heads *about* 50% of the times.

Suppose we have a coin, and we do not know whether it is fair or unfair. We flip the coin 1000 times and we get 550 heads.

Ⓠ Based on this evidence, can we judge whether the coin is fair or not?

One thing is clear: based on the evidence we cannot be 100% sure one way or the other. Indeed, getting 550 heads out of 1000 flips is a *possible* event, whether the coin is fair or not. Nevertheless, the result of the experiment contains some information about the coin. Is this information enough to judge the coin unfair with a reasonable degree of confidence?

To clarify the difficulty, consider the following hypothetical scenarios:

- We flip the coin 10 times and we get 6 heads.
- We flip the coin 100 times and we get 55 heads.
- We flip the coin 1000 times and we get 550 heads.
- We flip the coin 1000000 times and we get 550000 heads.

The first scenario contains hardly any indication of the unfairness of the coin. After all, a slight difference in the number of heads and tails is to be expected due to the random nature of the experiment. On the other hand, the last scenario is, for all practical purposes, a conclusive evidence that the coin is biased. So, where does "550-out-of-1000" stand? This illustrates the need for a quantitative reasoning to answer the question.

In order to quantify the strength of the evidence "550-out-of-1000" against the fairness of the coin, a natural idea is to ask how likely it is that a similar experiment with a *fair* coin leads to an outcome at least as extreme. The smaller the chance, the stronger the evidence against the fairness of the original coin.

One way to estimate the chance of obtaining at least 550 heads in 1000 flips of a fair coin is to use *computer simulation*.

Exercise. Write a computer program that simulates 1000 independent flips of a fair coin and returns the number of heads. Repeat the simulation 100000 times and find the number of simulations in which the number of heads is ≥ 550. Based on your simulation, how likely it is — approximately — to obtain at least 550 heads in 1000 independent flips of a fair coin?

Using computer simulation as outlined in the latter exercise is quite effective, and applicable to many other scenarios. However, the current example is simple enough that allows us to find the desired probability, without doing a simulation, by means of mathematical reasoning and some computation.

Indeed, as in Example 1.2.1, we have

$$\langle\text{chance of } \geq 550 \text{ heads out of } 1000 \text{ flips of a fair coin}\rangle = \sum_{k=550}^{1000}\binom{1000}{k}\frac{1}{2^{1000}} \tag{$\clubsuit$}$$

Computing this horrendous sum by hand is hopeless. Fortunately, we can exploit computers for such lengthy computations.[5] Using the statistical software R, we get the approximate value

$$(\clubsuit) \approx 0.0008653 \qquad\qquad \text{(using R)}$$

Alternatively, there is a mathematical trick for approximating the sum in (♣) with an integral, which in turn can be numerically computed with a computer software such as Mathematica, Maple, or Sage:

$$\begin{aligned}(\clubsuit) &\approx \frac{1}{\sqrt{2\pi}}\int_{\frac{549.5-500}{\sqrt{1000}\times\frac{1}{2}}}^{\infty} \mathrm{e}^{-\frac{1}{2}x^2}\,\mathrm{d}x && \text{(using a mathematical trick)} \\ &\approx 0.0008721 && \text{(using WolframAlpha)}\end{aligned}$$

Let us now get back to the question of the fairness of the coin. The calculated probability 0.00087 is quite small: it is less than 0.1%. Therefore, the evidence against the fairness of the coin is rather strong. In other words, based on the available evidence, we can be fairly confident that the coin is unfair.[6] ○

The "mathematical trick" mentioned in the latter example is in fact a deep mathematical theorem known as the *central limit theorem*, which is one of the fundamental links between probability theory and statistics. We will talk about the central limit theorem later in the course. The strategy we used to judge about the fairness of the coin is a standard strategy for statistical hypothesis testing. We will explore this further towards the end of the course.

Example 1.2.4 (Estimating bias)**.** This is a continuation of the previous example. Suppose that based on the available evidence (550 heads out of 1000 flips), we want to estimate the amount of unfairness or bias in the coin.

To this end, we need to refine our model of a coin flip. Let us denote by p the chance that the coin comes up heads. The chance that the coin comes up tails is then $1-p$. The coin is fair if $p = 1/2$. If $p > 1/2$ or $p < 1/2$, then the coin is *biased*.

Ⓠ Based on the same evidence (550 heads out of 1000 flips), what is a good estimate for p?

$\boxed{\text{A1}}$ The true value of p is understood as the frequency of heads when we flip the coin many many times. Hence, a reasonable estimate for p based on the available evidence is

$$\hat{p} = \langle\text{frequency of heads observed in } 1000 \text{ flips}\rangle = \frac{550}{1000} = 0.55 > 1/2\,.$$

This is referred to as a *point estimate* for p. We will discuss the topic of point estimation later in the course.

The disadvantage of a point estimate is that it does not convey any information about the accuracy or reliability of the estimate. For instance, the three scenarios

- getting 55 heads out of 100 flips,
- getting 550 heads out of 1000 flips,
- getting 550000 heads out of 1000000 flips,

would all give the same estimate 0.55. However, the estimate would clearly be much more precise/reliable if it came from the third scenario than if it came from the first scenario.

$\boxed{\text{A2}}$ As we shall see, using the same evidence (550 heads out of 1000 flips), it is possible to provide a better type of answer such as:

$$p = 0.55 \pm 0.041 \qquad \text{with } 99\% \text{ confidence.} \qquad (♪)$$

Observe that this answer has more information compared to the point estimate 0.55: the value 0.041 is an indication of the *precision* of the estimate, while 99% represents its reliability or *confidence level*. There is a general trade-off between precision and confidence. In particular, the same evidence can be used to provide other estimates:

$$\begin{aligned} p &= 0.55 \pm 0.052 && \text{with } 99.9\% \text{ confidence,} && (\sharp♪) \\ p &= 0.55 \pm 0.026 && \text{with } 90\% \text{ confidence.} && (\flat♪) \end{aligned}$$

Compared to (♪), the estimate in (♯♪) has lower precision but is more reliable. On the other hand, the estimate in (♭♪) has higher precision but with lower confidence.

Such estimates are referred to as *confidence intervals* or *interval estimates*. The precise interpretation of confidence intervals as well as how to find them will be later discussed in the course. ○

[5] Even with a computer, calculating such a sum is somewhat tricky. The reason is that the sum contains a large number of very small values. Hence, if we are not careful, the rounding errors could add up leading to larger errors. Fortunately, the standard software such as R and Python have routines for careful computation of such sums.

[6] Let us emphasize that the final judgement about whether the coin should be ruled as unfair or not is up to us. Indeed, the questions of "how small a probability is very small" and "how strong an evidence is very strong" are subjective.