Probability and Statistics

Manjunath Krishnapur

Contents

Preface

About the book

What is statistics and what is probability?

Sometimes statistics is described as *the art or science of decision making in the face of uncertainty*. Here are some examples to illustrate what it means.

Example 1. Recall the apocryphal story of two women who go to King Solomon with a child, each claiming that it is her own daughter. The solution according to the story uses human psychology and is not relevant to recall here. But is this a reasonable question that the king can decide?

Daughters resemble mothers to varying degrees, and one cannot be absolutely sure of guessing correctly. On the other hand, by comparing various features of the child with those of the two women, there is certainly a decent chance to guess correctly.

If we could always get the right answer, or if we could never get it right, the question would not have been interesting. However, here we have uncertainty, but there is a decent chance of getting the right answer. That makes it interesting - particularly because there could be worse methods and better methods. For example, we can have a debate between *eyeists* and *nosists* and *earists* as to whether it is better to compare eyes or noses or ears in arriving at a decision.

The decision is rather easy if one women is Japanese and the other is Kenyan. It gets harder if the two women are close relatives, because they themselves are similar and their children may be expected to have similar features[1].

Example 2. The IISc cricket team meets the Basavanagudi cricket club for a match. Unfortunately, the Basavanagudi team forgot to bring a coin to toss. The IISc captain helpfully offers his coin, but can he be trusted? What if he spent the previous night doctoring the coin so that it falls on one side with probability 3/4 (or some other number)?

Instead of cricket, they could spend their time on the more interesting question of checking if the coin is *fair* or *biased*. Here is one way. If the coin is fair, in a large number of tosses, common sense suggests that we should get about equal number of heads and tails. So they toss the coin 100 times. If the number of heads is exactly 50, perhaps they will agree that it is fair. If the number of heads is 90, perhaps they will agree that it is biased. What if the number of heads is 60? Or 35? Where and on what basis to draw the line between fair and biased? Again we are faced with the question of making decision in the face of uncertainty.

[1] Genetic testing is of course an almost error-free way of deciding, but here we imagine a time before 1950 when we did not know the science or technology of DNA testing.

Example 3. A psychic claims to have divine visions unavailable to most of us. You are assigned the task of testing her claims. You take a standard deck of cards, shuffle it well and keep it face down on the table. The psychic writes down the list of cards in some order - whatever her vision tells her about how the deck is ordered. Then you count the number of correct guesses. If the number is 1 or 2, perhaps you can dismiss her claims. If it is 45, perhaps you ought to be take her seriously. Again, where to draw the line?

The logic is this. Roughly one may say that *surprise* is just the name for our reaction to an event that we *á priori* thought to have low chance of occurring. Thus, we approach the experiment with the belief that the psychic is just guessing at random, and if the results are such that under that random-guess-hypothesis they have very small probability, then we are willing to be surprised, that is willing to discard our preconception and accept that she is a psychic.

How low a probability is surprising? In the context of psychics, let us say, 1/10000. Once we fix that, we must find a number $m \leq 52$ such that by pure guessing, the probability to get more than m correct guesses is less that 1/10000. Then we tell the psychic that if she gets more than m correct guesses, we accept her claim, and otherwise, reject her claim. This raises the following question.

Question 4. For a deck of 52 cards, find the number m such that

$$\mathbb{P}(\text{by random guessing we get more than } m \text{ correct guesses}) < \frac{1}{10000}.$$

Summary: There are many situations in real life where one is required to make decisions under uncertainty. In particular, the psychic question shows us that we are required to calculate chances (under the assumption that the psychic is guessing at random). There are a great many other situations where we are required to compute probabilities under specific assumptions. The results of these computations allow us to decide whether or not something that has happened should be deemed extra-ordinary or was it to be expected anyway.

For the sake of clarity, we may divide the task into two parts. One of computing probabilities under precise mathematical assumptions. Another of deciding what mathematical assumptions are reasonable in a given situation. Once we agree on the second part (what assumptions/models are reasonable), then we can use probability computations from the first part to make decisions in real-life problems. The first aspect is essentially the subject of *probability*, the second is the subject of *statistics*[2]. In this course, we first discuss the concepts and techniques of probability and then come to a few of the kind of problems encountered in statistics.

Probability: Probability theory is a branch of pure mathematics, and forms the theoretical basis of statistics. In itself, probability theory has some basic objects and their relations (like real numbers, addition etc for analysis) and it makes no pretense of saying anything about

[2]These remarks are meant as a rough guide as we enter the course. Don't mistake it for an all-encompassing description of the two fields!

the real world (or makes the pretense of not saying anything about the real world). Axioms are given and theorems are then deduced about these objects, just as in any other part of mathematics.

But a very important aspect of probability is that it is *applicable*. In other words, there are many real-world situations in which it is reasonable to take a *model* in mathematical probability, and it turns out to reasonably replicate features of the real-world situation.

In the example above, to compute the probability one must make the assumption that the deck of cards was completely shuffled. In other words, all possible arrangements of the 52 cards are assumed to be equally likely. Note that there are $52! \approx 10^{68}$ such arrangements! Whether this assumption is reasonable or not depends on how well the card was shuffled, whether the psychic was able to get a peek at the cards, whether some insider is informing the psychic of the cards etc. All these are non-mathematical questions, and must be decided on other basis.

This relationship between a mathematical model and the real world is no different than inany other field of mathematics that applies to the real world. For example, Euclidean plane geometry is a branch of mathematics, based on clearly stated axioms. Triangles, rectangles, hexagons etc., are objects that are well-defined mathematical objects. In a real life situation, say when you are trying to figure out how much carpet is needed to cover a room, you may invoke one of these as a model for the room. For most rooms, a rectangle may be a good model but not a triangle.

However...: Probability and statistics are very relevant in many situations that do not involve any uncertainty on the face of it. Sometimes we introduce randomness, or assume randomness to do something or to analyse the situation. Here are some examples.

Example 5. *Sample survey.* A news organization wishes to predict the outcome of an election, a month before the election. Assuming that the voters have already decided who they will vote for, there is no randomness in the situation. It is just that no one knows the minds of the voters, hence the result is unknown and we loosely say things like "candidate A may win or candidate B may win", even though the result is already determined. To know the mind of the voters, the organization conducts an opinion poll - that is, they ask a small number of people ("sample") who they will vote for, and use that data to predict the percentage of voters for various candidates. It turns out that reliable results are obtained only if the samples are chosen *randomly* from the population. Asking your family members or friends (or for lack of it, facebook friends) is guaranteed to give unreliable results.

Example 6. *Compression of data.* Large files in a computer can be compressed to a .zip format and uncompressed when necessary. How is it possible to compress data like this? To give a very simple analogy, consider a long English word like *invertebrate*. If we take a novel and replace every occurrence of this word with "zqz", then it is certainly possible to recover the original novel (since "zqz" does not occur anywhere else). But the reduction in size by replacing the 12-letter word by the 3-letter word is not much, since the word *invertebrate*

does not occur often. Instead, if we replace the 4-letter word "then" by "zqz", then the total reduction obtained may be much higher, as the word "then" occurs quite often.

This suggests the following optimal way to represent words in English. The 26 most frequent words will be represented by single letters. The next 26×26 most frequent words will be represented by two letter words, the next $26 \times 26 \times 26$ most frequent words by three-letter words, etc. Assuming there are no errors in transcription, this is a good way to reduce the size of any text document! Now, this involves knowing what the frequencies of occurrences of various words in actual texts are. Such statistics of usage of words are therefore clearly relevant (and they could be different for biology textbooks as compared to 19th century novels).

EXAMPLE 7. There are situations where the question has no randomness, but we introduce randomness to solve something. This is true of many algorithms to search or sort or other tasks. This cannot be explained right now (and we may not go into it in this course), but let us give a simple reason to say why introducing randomness is a good idea in many situations. In the game of *rock-paper-scissors*, two people simultaneously shout one of the three words, rock, paper or scissors. The rule is that scissors beats paper, paper beats rock and rock beats scissors (if they both call the same word, they must repeat). In a game like this, although there is complete symmetry in the three items, it would be silly to have a fixed strategy. In other words, if you decide to always say rock, thinking that it doesn't matter which you choose, then your opponent can use that knowledge to always choose paper and thus win! In many games where the opponent gets to know your strategy (but not your move), the best strategy would involve randomly choosing your move.

Part 1

Probability

CHAPTER 1

The setting of discrete probability

1. Discrete probability spaces

DEFINITION 8. Let Ω be a finite or countable[1] set. Let $p : \Omega \to [0,1]$ be a function such that $\sum_{\omega\in\Omega} p(\omega) = 1$. Then (Ω, p) is called a *discrete probability space*.

The set Ω is called the *sample space* and $p(\omega)$ (often we write p_ω for simplicity of notation) are called *elementary probabilities*.

Any subset $A \subseteq \Omega$ is called an *event*. Its *probability* is defined as $\mathbb{P}(A) = \sum_{\omega\in A} p_\omega$.

Any function $X : \Omega \to \mathbb{R}$ is called a *random variable*. For a random variable we define its *expected value* or *mean* as $\mathbb{E}[X] = \sum_{\omega\in\Omega} X(\omega)p_\omega$.

All of probability in one line

Take an (interesting) probability space (Ω, p) and an (interesting) event $A \subseteq \Omega$. Find $\mathbb{P}(A)$.

This is the mathematical side of the picture. It is easy to make up any number of probability spaces. Just take a finite set and assign positive numbers to each element of the set so that the total is 1. The simplest way is to take a finite set Ω and define $p(\omega) = \frac{1}{\#\Omega}$ for each $\omega \in \Omega$.

EXAMPLE 9. Let $N \geq 1$ and let $\Omega = [N]$ (recall that $[N]$ denotes the set $\{1, 2, \ldots, N\}$). Let $p(\omega) = \frac{1}{N}$ for each $\omega \in [N]$. This is clearly a valid probability space.

But to have any content, we must understand the situations where a probability space can be used. In the following section, we start with "real-life" contexts and write down the probability spaces that could reasonably describe them. In the section after, we write down probability spaces like above, and then try to imagine situations where they could be applicable.

1.1. Probability in the real world. In real life, there are often situations where there are several possible outcomes but which one will occur is unpredictable in some way. For example, when we toss a coin, we may get heads or tails. In such cases we use words such as *probability or chance*, *event or happening*, *randomness* etc. What is the relationship between the intuitive and mathematical meanings of words such as probability or chance?

In a given physical situation, we choose one out of all possible probability spaces that we think captures best the chance happenings in the situation. The chosen probability space is

[1] For those unfamiliar with countable sets, it will be explained in some detail later.

then called a *model* or a *probability model* for the given situation. Once the model has been chosen, calculation of probabilities of events therein is a mathematical problem. Whether the model really captures the given situation, or whether the model is inadequate and over-simplified is a non-mathematical question. Nevertheless that is an important question, and can be answered by observing the real life situation and comparing the outcomes with predictions made using the model[2].

Now we describe several *random experiments* (a non-mathematical term to indicate a "real-life" phenomenon that is supposed to involve chance happenings) in which the previously given examples of probability spaces arise. Describing the probability space is the first step in any probability problem.

EXAMPLE 10. **Physical situation:** Toss a coin. Randomness enters because the coin may turn up head or tail and that it is inherently unpredictable. What is a good probability model for this situation?

Since there are two outcomes, the sample space $\Omega = \{0, 1\}$ (we use 1 for heads and 0 for tails, you may use any two symbols instead) is a clear choice. What about elementary probabilities? If the coin looks symmetrical, there is no reason to prefer one face over the other, so we are tempted to assign $p_0 = p_1 = \frac{1}{2}$. Then we have a probability model for the tossing of a fair coin.

If the coin does not look symmetrical, we cannot decide by pure thought what the probabilities should be. It must come either from some theory or experiment. In general, we shall have to take $p_1 = p$ and $p_0 = 1 - p$ for some $p \in [0, 1]$. This is a valid probability space.

There is always an approximation in going from the real-world to a mathematical model. For example, a real coin can land on its side. If the coin is very thick, then it might be closer to a cylinder which can land in three ways and then we would have to modify the model...

EXAMPLE 11. Toss n fair coins. Now the outcome of the experiment will tell us the result of the first toss, of the second toss and so on. Hence

$$\Omega = \{0, 1\}^n = \{\underline{\omega} : \underline{\omega} = (\omega_1, \ldots, \omega_n) \text{ with } \omega_i = 0 \text{ or } 1 \text{ for each } i \leq n\}.$$

Let $p_{\underline{\omega}} = 2^{-n}$ for each $\underline{\omega} \in \Omega$. Since Ω has 2^n elements, it follows that this is a valid assignment of elementary probabilities.

Can the probability space in Example 11 serve as a good model for the tossing of the same coin, n times in succession? The answer is yes, provided that the coin forgets the outcomes on the previous tosses. While that may seem obvious, it would be violated if our "coin" was a hollow lens filled with a semi-solid material like glue (then, depending on which way the

[2]Roughly speaking we may divide the course into two parts according to these two issues. In the probability part of the course, we shall take many such models for granted and learn how to calculate or approximately calculate probabilities. In the statistics part of the course we shall see some methods by which we can arrive at such models, or test the validity of a proposed model. Do not take this division too seriously.

coin fell on the first toss, the glue would settle more on the lower side and consequently the coin would be more likely to fall the same way again). This is a coin with memory!

EXAMPLE 12. Randomly throw r distinguishable balls into m labelled bins. The outcome of this experiment will tell us for each ball which urn/bin it went into. Hence

$$\Omega = \{\underline{\omega} : \underline{\omega} = (\omega_1, \ldots, \omega_r) \text{ with } 1 \leq \omega_i \leq m \text{ for each } i \leq r\}.$$

The cardinality of Ω is m^r (since each co-ordinate ω_i can take one of m values). Hence, if we set $p_{\underline{\omega}} = m^{-r}$ for each $\underline{\omega} \in \Omega$, we get a valid probability space.

The next example is more involved and interesting.

EXAMPLE 13. **Real-life situation:** Imagine a man-woman pair. Their first child is random, for example, the sex of the child, or the height to which the child will ultimately grow, etc cannot be predicted with certainty. How to make a probability model that captures the situation?

A possible probability model: Let there be n genes in each human, and each of the genes can take two possible values (Mendel's "factors"), which we denote as 0 or 1. Then, let $\Omega = \{0,1\}^n = \{\mathbf{x} = (x_1, \ldots, x_n) : x_i = 0 \text{ or } 1\}$. In this sense, each human being can be encoded as a vector in $\{0,1\}^n$.

To assign probabilities, one must know the parents. Let the two parents have gene sequences $\mathbf{a} = (a_1, \ldots, a_n)$ and $\mathbf{b} = (b_1, \ldots, b_n)$. Then the possible offsprings gene sequences are in the set $\Omega_0 := \{x \in \{0,1\}^n : x_i = a_i \text{ or } b_i, \text{ for each } i \leq n\}$. Let $L := \#\{i : a_i \neq b_i\}$. Then $\#\Omega_0 = 2^L$.

One possible assignment of probabilities is that each of these offsprings is equally likely. In that case we can capture the situation in the following probability models.

(1) Let Ω_0 be the sample space and let $p_x = 2^{-L}$ for each $x \in \Omega_0$.

(2) Let Ω be the sample space and let

$$p_x = \begin{cases} 2^{-L} & \text{if } x \in \Omega_0 \\ 0 & \text{if } x \notin \Omega_0. \end{cases}$$

The second one has the advantage that if we change the parent pair, we don't have to change the sample space, only the elementary probabilities. What are some interesting events? Hypothetically, the susceptibility to a disease X could be determined by the first ten genes, say the person is likely to get the disease if there are at-most four 1s among the first ten. This would correspond to the event that $A = \{x \in \Omega_0 : x_0 + \ldots + x_{10} \leq 4\}$. (Caution: As far as I know, reading the genetic sequence to infer about the phenotype is still an impractical task in general).

Reasonable model? There are many simplifications involved here. Firstly, genes are somewhat ill-defined concepts, better defined are nucleotides in the DNA (and even then there are two copies of each gene). Secondly, there are many "errors" in real DNA, even

the total number of genes can change, there can be big chunks missing, a whole extra chromosome etc. Thirdly, the assumption that all possible gene-sequences in Ω_0 are equally likely is incorrect - if two genes are physically close to each other in a chromosome, then they are likely to both come from the father or both from the mother. Lastly, if our interest originally was to guess the eventual height of the child or its intelligence, then it is not clear that these are determined by the genes alone (environmental factors such as availability of food etc. also matter). Finally, in case of the problem that Solomon faced, the information about genes of the parents was not available, the model as written would be useless.

REMARK 14. We have discussed at length the reasonability of the model in this example to indicate the significant effort needed to find a sufficiently accurate but also reasonably simple probability model for a real-world situation. Henceforth, we shall omit such caveats and simply switch back-and-forth between a real-world situation and a reasonable-looking probability model as if there is no difference between the two. However, thinking about the appropriateness of the chosen models is much encouraged.

2. Examples of discrete probability spaces

Let us start with two important examples of that include many other examples, although that may not be apparent on the surface.

EXAMPLE 15. **Sampling with replacement from a population**. Define $\Omega = \{\underline{\omega} \in [N]^k : \omega_i \in [N] \text{ for } 1 \leq i \leq k\}$ with $p_{\underline{\omega}} = 1/N^k$ for each $\underline{\omega} \in \Omega$. Here $[N]$ is the population (so the size of the population is N) and the size of the sample is k. Often the language used is of a box with N coupons from which k are drawn with replacement. That is, coupons are drawn one after another, and after each draw the coupon is returned to the box before the next one is drawn. Here are many examples that are special cases of this:

(1) Toss k fair coins (Example 11). This is sampling with replacement from $\{0, 1\}$, where 0 stands for tail and 1 stands for head.

(2) Roll a fair die k times. This is sampling with replacement from $\{1, 2, \ldots, 6\}$.

(3) Throw r balls uniformly at random into m bins (Example 12). Here $N = m$ and $k = r$.

(4) Record birthdays in a group of k people. Here $N = 365$. Of course, simplifications are involved, such as ignoring the leap years, ignoring the possibility of there being twins in the group, assuming that all days of the year are equally likely to be a birthday, and so on.

EXAMPLE 16. **Sampling without replacement from a population**. Now we take

$$\Omega = \left\{ \underline{\omega} \in [N]^k : \omega_i \text{ are distinct elements of } [N] \right\},$$

$$p_{\underline{\omega}} = \frac{1}{N(N-1)\dots(N-k+1)} \text{ for each } \underline{\omega} \in \Omega.$$

Here are some special cases of sampling without replacement.

(1) Deal 5 cards from a shuffled deck. Here $N = 52$ (population size) and $k = 5$ (sample size). Although you may feel that what is dealt is an unordered set, there is no harm in thinking of it as an ordered tuple, by drawing the cards one after another.

(2) Conduct a survey by sampling 100 people from a population of 10000 people. Here $N = 10000$ and $k = 100$.

(3) Students in a class of 25 stand in a line. Here $N = 25$ and $k = 25$.

(4) Throw k balls randomly into N bins, where each bin has a capacity of at most one.

EXAMPLE 17. **Non-uniform sampling**. In Example 15 and Example 16 we assumed that all coupons are equally likely to show up on any draw. We can allow for more general situation where there are numbers $a_1, \dots, a_N \geq 0$ such that $a_1 + \dots + a_N = 1$. This does not change the sample spaces, but the elementary probabilities do change.

- Sampling k times with replacement: $\Omega = [N]^k$ and $p(\underline{\omega}) = a_{\omega_1} a_{\omega_2} \dots a_{\omega_k}$ for $\underline{\omega} = (\omega_1, \dots, \omega_k)$. It can also be written as $a_1^{r_1(\omega)} a_2^{r_2(\omega)} \dots a_N^{r_N(\omega)}$ where $r_j(\omega) = \sum_{i=1}^N \mathbf{1}_{\omega_i = j}$ (number of times jth coupon is drawn).
- Sampling k times without replacement: $\Omega = \left\{ \underline{\omega} \in [N]^k : \omega_i \text{ are distinct elements of } [N] \right\}$ and $p(\underline{\omega}) = a_{\omega_1} \frac{a_{\omega_2}}{1-a_{\omega_1}} \frac{a_{\omega_3}}{1-(a_{\omega_1}+a_{\omega_2})} \cdots \frac{a_{\omega_k}}{1-(a_{\omega_1}+\dots+a_{\omega_{k-1}})}$.

In both cases, one can check that $p(\underline{\omega})$ do sum up to 1, by summing over ω_k, then over ω_{k-1} and so on.

EXAMPLE 18. **Shuffle a deck of 52 cards**. This is a special case of sampling without replacement with $N = k = 52$. Therefore, $\Omega = S_{52}$, the set of all permutations[3] of [52] and $p(\pi) = \frac{1}{52!}$ for $\pi \in S_{52}$.

EXAMPLE 19. **Place r indistinguishable balls in m distinguishable urns at random**. Now we come to the case when the balls are indistinguishable. Since the balls are indistinguishable, we can only count the number of balls in each urn. The sample space is

$$\Omega = \{(\ell_1, \dots, \ell_m) : \ell_i \geq 0,\ \ell_1 + \dots + \ell_m = r\}.$$

[3]We use the notation $[n]$ to denote the set $\{1, 2, \dots, n\}$. A permutation of $[n]$ is a vector $(i_1, i_2, \dots, i_n)$ where $i_1, \dots, i_n$ are distinct elements of $[n]$, in other words, they are $1, 2, \dots, n$ but in some order. Mathematically, we may define a permutation as a bijection $\pi : [n] \to [n]$. Indeed, for a bijection π, the numbers $\pi(1), \dots, \pi(n)$ are just $1, 2, \dots, n$ in some order.

We give two proposals for the elementary probabilities.

(1) (Maxwell-Boltzmann statistics): Let $p^{\text{MB}}_{(\ell_1,\ldots,\ell_m)} = \frac{r!}{\ell_1!\ell_2!\ldots\ell_m!}\frac{1}{m^r}$. These are the probabilities that result if we place r labelled balls in m labelled urns (which is a special case of sampling with replacement r times from a $[m]$), and then erase the labels on the balls.

(2) (Bose-Einstein statistics): Let $p^{\text{BE}}_{(\ell_1,\ldots,\ell_m)} = \frac{1}{\binom{m+r-1}{m-1}}$ for each $(\ell_1,\ldots,\ell_m) \in \Omega$. Elementary probabilities are chosen so that all distinguishable configurations are equally likely.

That these are legitimate probability spaces depend on two combinatorial facts.

Exercise 20. (1) Let $(\ell_1,\ldots,\ell_m) \in \Omega$. Show that the number of ways to place r distingushale balls into m bins so that there are exactly ℓ_j balls in the jth bin is equal to $\frac{r!}{\ell_1!\ell_2!\ldots\ell_m!}$. Hence or otherwise, show that $\sum_{\omega\in\Omega} p^{MB}_\omega = 1$.

(2) Show that $\#\Omega = \binom{m+r-1}{m-1}$. Hence, $\sum_{\omega\in\Omega} p^{BE}_\omega = 1$.

The two models are clearly different. Which one captures reality? We can arbitrarily label the balls for our convenience, and then erase the labels in the end. This clearly yields elementary probabilities p^{MB}. Or to put it another way, pick the balls one by one and assign them randomly to one of the urns. This suggests that p^{MB} is the "right one".

This leaves open the question of whether there is a natural mechanism of assigning balls to urns so that the probabilities p^{BE} shows up. No such mechanism has been found. But this probability space does occur in the physical world. If r photons ("indistinguishable balls") are to occupy m energy levels ("urns"), then empirically it has been verified that the correct probability space is the second one![4]

Example 21. **Toss a coin till a head turns up**. $\Omega = \{1, 01, 001, 0001, \ldots\} \cup \{\bar{0}\}$. Let us write $0^k1 = 0\ldots01$ as a short form for k zeros (tails) followed by 1 and $\bar{0}$ stands for the sequence of all tails. Let $p \in [0,1]$. Then, we set $p_{0^k1} = q^k p$ for each $k \in \mathbb{N}$. We also set $p_{\bar{0}} = 0$ if $p > 0$ and $p_{\bar{0}} = 1$ if $p = 0$. This is forced on us by the requirement that elementary probabilities add to 1.

Let $A = \{0^k1 : k \geq n\}$ be the event that at least n tails fall before a head turns up. Then $\mathbb{P}(A) = q^n p + q^{n+1}p + \ldots = q^n$.

[4]The probabilities p^{MB} and p^{BE} are called Maxwell-Boltzmann statistics and Bose-Einstein statistics. There is a third kind, called Fermi-Dirac statistics which is obeyed by electrons. For general $m \geq r$, the sample space is $\Omega_{\text{FD}} = \{(\ell_1,\ldots,\ell_m) : \ell_i = 0 \text{ or } 1 \text{ and } \ell_1 + \ldots + \ell_m = r\}$ with equal probabilities for each element. In words, all distinguishable configurations are equally likely, with the added constraint that at most one electron can occupy each energy level.

EXAMPLE 22. **Gibbs measures**. Let Ω be a finite set and let $\mathcal{H} : \Omega \to \mathbb{R}$ be a function. Fix $\beta \geq 0$. Define $Z_\beta = \sum_\omega e^{-\beta\mathcal{H}(\omega)}$ and then set $p_\omega = \frac{1}{Z_\beta} e^{-\beta\mathcal{H}(\omega)}$. This is clearly a valid assignment of probabilities.

This is a class of examples from statistical physics. In that context, Ω is the set of all possible states of a system and $\mathcal{H}(\omega)$ is the energy of the state ω. In mechanics a system settles down to the state with the lowest possible energy, but if there are thermal fluctuations (meaning the ambient temperature is not absolute zero), then the system may also be found in other states, but higher energies are less and less likely. In the above assignment, for two states ω and ω', we see that $p_\omega/p_{\omega'} = e^{\beta(\mathcal{H}(\omega')-\mathcal{H}(\omega))}$ showing that higher energy states are less probable. When $\beta = 0$, we get $p_\omega = 1/|\Omega|$, the uniform distribution on Ω. In statistical physics, β is equated to $1/\kappa T$ where T is the temperature and κ is Boltzmann's constant.

Different physical systems are defined by choosing Ω and $\mathcal{H}$ differently. Hence this provides a rich class of examples which are of great importance in probability.

It may seem that probability is trivial, since the only problem is to find the sum of p_ω for ω belonging to the event of interest. This is far from the case. The following example is an illustration.

EXAMPLE 23. **Percolation**. Fix m, n and consider a rectangle in $\mathbb{Z}^2$, $R = \{(i,j) \in \mathbb{Z}^2 : 0 \leq i \leq n,\ 0 \leq j \leq m\}$. Draw this on the plane along with the grid lines. We see $(m+1)n$ horizontal edges and $(n+1)m$ vertical edges. Let E be the set of $N = (m+1)n + (n+1)m$ edges and let Ω be the set of all subsets of E. Then $|\Omega| = 2^N$. Let $p_\omega = 2^{-N}$ for each $\omega \in \Omega$. An interesting event is

$$A = \{\omega \in \Omega : \text{ the subset of edges in } \omega \text{ connect the top side of } R \text{ to the bottom side of } R\}.$$

This may be thought of as follows. Imagine that each edge is a pipe through which water can flow. However each tube may be blocked or open. ω is the subset of pipes that are open. Now pour water at the top of the rectangle R. Will water trickle down to the bottom? The answer is yes if and only if ω belongs to A.

Finding $\mathbb{P}(A)$ is a very difficult problem. When n is large and $m = 2n$, it is expected that $\mathbb{P}(A)$ converges to a specific number, but proving it is an open problem as of today![5]

We now give two non-examples.

EXAMPLE 24. **A non-example - Pick a natural number uniformly at random**. The sample space is clearly $\Omega = \mathbb{N} = \{1, 2, 3, \ldots\}$. The phrase "uniformly at random" suggests that the elementary probabilities should be the same for all elements. That is $p_i = p$ for all $i \in \mathbb{N}$ for some p. If $p = 0$, then $\sum_{i\in\mathbb{N}} p_i = 0$ whereas if $p > 0$, then $\sum_{i\in\mathbb{N}} p_i = \infty$. This

[5] In a very similar problem on a triangular lattice, it was proved by Stanislav Smirnov (2001) for which he won a fields medal. Proof that computing probabilities is not always trivial!

means that there is no way to assign elementary probabilities so that each number has the same chance to be picked.

This appears obvious, but many folklore puzzles and paradoxes in probability are based on the faulty assumption that it is possible to pick a natural number at random. For example, when asked a question like "What is the probability that a random integer is odd?", many people answer 1/2. We want to emphasize that the probability space has to be defined first, and only then can probabilities of events be calculated. Thus, the question does not make sense to us and we do not have to answer it![6]

EXAMPLE 25. **Another non-example - Throwing darts**. A dart is thrown at a circular dart board. We assume that the dart does hit the board but were it hits is "random" in the same sense in which we say the a coin toss is random. Intuitively this appears to make sense. However our framework is not general enough to incorporate this example. Let us see why.

The dart board can be considered to be the disk $\Omega = \{(x,y) : x^2 + y^2 \leq r^2\}$ of given radius r. This is an uncountable set. We cannot assign elementary probabilities $p_{(x,y)}$ for each $(x,y) \in \Omega$ in any reasonable way. In fact the only reasonable assignment would be to set $p_{(x,y)} = 0$ for each (x,y) but then what is $\mathbb{P}(A)$ for a subset A? Uncountable sums are not well defined.

We need a branch of mathematics called *measure theory* to make proper sense of uncountable probability spaces. This will not be done in this course although we shall later say a bit about the difficulties involved. The same difficulty shows up in the following "random experiments" also.

(1) **Draw a number at random from the interval** $[0,1]$. $\Omega = [0,1]$ which is uncountable.

(2) **Toss a fair coin infinitely many times**. $\Omega = \{0,1\}^{\mathbb{N}} := \{\underline{\omega} = (\omega_1, \omega_2, \ldots) : \omega_i = 0 \text{ or } 1\}$. This is again an uncountable set.

REMARK 26. In one sense, the first non-example is almost irredeemable but the second non-example can be dealt with, except for technicalities beyond this course. We shall later give a set of working rules to work with such "continuous probabilities". Fully satisfactory development will have to wait for a course in measure theory.

[6]For those interested, there is one way to make sense of such questions. It is to consider a sequence of probability spaces $\Omega^{(n)} = \{1,2,\ldots,n\}$ with elementary probabilities $p_i^{(n)} = 1/n$ for each $i \in \Omega_n$. Then, for a subset $A \subseteq \mathbb{Z}$, we consider $\mathbb{P}_n(A \cap \Omega_n) = \#(A \cap [n])/n$. If these probabilities converge to a limit x as $n \to \infty$, then we could say that A has asymptotic probability x. In this sense, the set of odd numbers does have asymptotic probability 1/2, the set of numbers divisible by 7 has asymptotic probability 1/7 and the set of prime numbers has asymptotic probability 0. However, this notion of asymptotic probability has many shortcomings. Many subsets of natural numbers will not have an asymptotic probability, and even sets which do have asymptotic probability fail to satisfy basic rules of probability that we shall see later. Hence, we shall keep such examples out of our system.

3. Some probability calculations

PROBLEM 27. A coin is tossed n times. What is the probability that one gets exactly k heads?

This is a special case of sampling with replacement n times from $\{1,2\}$. If the coin has probability p of falling head, then $a_1 = p$ and $a_0 = q$ (where $q = 1-p$). Hence $\Omega = \{0,1\}^n$ and $p_{\underline{\omega}} = p^{\sum_i \omega_i} q^{n-\sum_i \omega_i}$ (observe that $\sum_i \omega_i$ is the number of heads and $\sum_i (1-\omega_i)$ is the number of tails).

The event of interest is $B_k = \{\underline{\omega} : \sum_{i=1}^n \omega_i = k\}$. For each $\underline{\omega} \in B$, we have $p(\underline{\omega}) = p^k q^{n-k}$ and there are $\binom{n}{k}$ elements in B_k. Therefore $\mathbb{P}(B_k) = \binom{n}{k} p^k q^{n-k}$. These numbers are called binomial probabilities, due to the appearance of the binomial coefficients. We shall encounter them repeatedly in the course.

PROBLEM 28. A die is rolled n times. What is the chance that the face j turns up n_j times, for $j = 1, 2, \ldots, 6$.

This is sampling with replacement n times from $[6]$. Hence $\Omega = [6]^n$. Let p_j denote the probability of the jth face turning up on a roll (so $p_1 + \ldots + p_6 = 1$). Then, $p(\underline{\omega}) = p_1^{r_1(\underline{\omega})} \ldots p_6^{r_6(\underline{\omega})}$, where $r_j(\omega) = \mathbf{1}_{\omega_1 = j} + \ldots + \mathbf{1}_{\omega_n = j}$ is the number of times the jth face turns up. The event of interest is $A = \{\underline{\omega} : r_j(\underline{\omega}) = n_j \text{ for } 1 \le j \le 6\}$.

Assume that $n_1 + \ldots + n_6 = n$ (otherwise $\mathbb{P}(A) = 0$). If $\underline{\omega} \in A$, then $p(\underline{\omega}) = p_{\omega_1} \ldots p_{\omega_n} = p_1^{n_1} \ldots p_6^{n_6}$. Although elements of Ω may have different elementary probabilities, all elements of A have exactly the same elementary probability! The number of elements in A is $\frac{n!}{n_1! \ldots n_6!}$ (why?). Therefore,

$$\mathbb{P}(A) = \frac{n!}{n_1! \ldots n_6!} p_1^{n_1} \ldots p_6^{n_6}.$$

More generally, consider the problem of throwing r distinguishable balls at random into m labelled bins where each ball can fall into the bins with probabilities $p_1, \ldots, p_m$. Then, the probability that we have exactly r_j balls in the jth bin, for $1 \le j \le m$, is given by

$$\frac{r!}{r_1! \ldots r_m!} p_1^{r_1} \ldots p_m^{r_m}.$$

These are called *multinomial probabilities*.

PROBLEM 29. In a party there are k people. Here are some events of interest.

- All the people are born in the same month.
- At least two people have the same birthday.
- Every month has someone's birthday.

The third one requires some new ideas, we deal with it after we talk about *inclusion-exclusion* (but you are encouraged to try it first!). We work out the first and second. The first is really a simple exercise, but the second one is a famous problem known as the *birthday problem* or *birthday paradox*, because of the surprising answer.

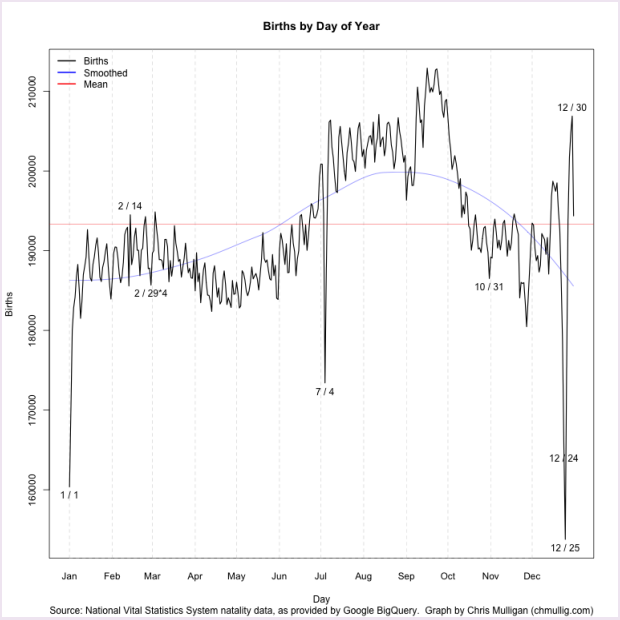

FIGURE 1. Frequencies of birthdays in the United States of America from 1969 to 1988. Data taken from Andrew Gelman.

First we write down the probability space, which is the one of sampling k times with replacement from $[N]$ where $N = 365$. We ignore leap year and assume that all days of the year are equally likely to be birthdays (see Figure 3). Thus, $\Omega = [N]^k$ and $p(\underline{\omega}) = N^{-k}$ for all $\underline{\omega} \in \Omega$.

▶ Let d_i be the number of days in the ith month, for $1 \leq i \leq 12$. The event A that all n people are born in the same month is the disjoint union of $A_1, \ldots, A_{12}$, where A_i is the event that all n people are born in the ith month. It is clear that $\#A_i = d_i^k$, and hence $\#A = d_1^k + \ldots + d_{12}^k$. Therefore,

$$\mathbb{P}(A) = \frac{1}{N^k}(d_1^k + \ldots + d_{12}^k).$$

To get an idea of its value, take $d_i = N/12$ for simplicity. Then $\mathbb{P}(A) = \frac{1}{12^{k-1}}$. Even for $k = 6$, this chance is smaller than 10^{-5}.

▶ The event of interest is that there is at least one day of the year that has at least two birthdays. In other words,

$$B = \{\underline{\omega} = (\omega_1, \ldots, \omega_k) : \omega_i = \omega_j \text{ for some } i \neq j\}.$$

It is difficult to count the number of elements in B (you are encouraged to try!). Turns out, it is easier to count the complementary event

$$B^c = \{\underline{\omega} = (\omega_1, \ldots, \omega_k) : \omega_1, \ldots, \omega_k \text{ are distinct}\}.$$

Indeed, the cardinality of B^c is $N(N-1)\dots(N-k+1)$, whence $\#B = \#\Omega - \#B^c = N^k - N(N-1)\dots(N-k+1)$. Thus,

$$\mathbb{P}(B) = 1 - \frac{N(N-1)\dots(N-k+1)}{N^k} = 1 - \prod_{j=1}^{k-1}\left(1 - \frac{j}{N}\right).$$

If $k > N$, the probability is obviously one. The reason this is called a "paradox" is that even for k much smaller than N, the probability becomes significantly large. Recalling that $N = 365$, here are the probabilities for a few special cases[7].

k	5	15	25	35	45
$\mathbb{P}(B)$	0.027	0.253	0.569	0.814	0.941

The surprise is that the probability is much higher than what most people expect. Most people compare the number of people, say 25, to the number of days which is 365. What matters is the number of pairs of people (since any pair can have the same birthday with a chance of 1/365), which is $\binom{25}{2} = 300$ which is comparable to 365.

► The third event of having a birthday in each month requires a new idea, that of *inclusion-exclusion* principle. We shall see it later.

EXAMPLE 30. **"Psychic" guesses a deck of cards**. The sample space is $\Omega = S_{52} \times S_{52}$ and if the person is guessing at random, then $p_{(\pi,\sigma)} = 1/(52!)^2$ for each pair (π, σ) of permutations. In a pair (π, σ), the permutation π denotes the actual order of cards in the shuffled deck, and σ denotes the order guessed by the psychic.

An interesting random variable is the number of correct guesses. This is the function $X : \Omega \to \mathbb{R}$ defined by $X(\pi, \sigma) = \sum_{i=1}^{52} \mathbf{1}_{\pi_i = \sigma_i}$. Correspondingly we have the events $A_k = \{(\pi, \sigma) : X(\pi, \sigma) \geq k\}$, which is the event of getting at least k guesses right.

[7] These computations can be done exactly on a computer. But here is a useful idea. For $|x| \leq \frac{1}{2}$ we have $e^{-x-x^2} \leq 1 - x \leq e^x$ (the right side inequality holds for all $x \in \mathbb{R}$). Therefore, $\mathbb{P}(B) \geq 1 - \prod_{j=1}^{k-1} e^{-\frac{j}{N}} = 1 - e^{-\frac{k(k-1)}{2N}}$. Similarly you can get an upper bound.

4. Countable and uncountable

DEFINITION 31. An set Ω is said to be *finite* if there is an $n \in \mathbb{N}$ and a bijection from Ω onto $[n]$. An infinite set Ω is said to be countable if there is a bijection from $\mathbb{N}$ onto Ω.

Generally, the word countable also includes finite sets. If Ω is an infinite countable set, then using any bijection $f : \mathbb{N} \to \Omega$, we can list the elements of Ω as a sequence

$$f(1), f(2), f(3) \dots$$

so that each element of Ω occurs exactly once in the sequence. Conversely, if you can write the elements of Ω as a sequence, it defines an injective function from natural numbers onto Ω (send 1 to the first element of the sequence, 2 to the second element etc).

EXAMPLE 32. The set of integers $\mathbb{Z}$ is countable. Define $f : \mathbb{N} \to \mathbb{Z}$ by

$$f(n) = \begin{cases} \frac{1}{2}n & \text{if } n \text{ is even.} \\ -\frac{1}{2}(n-1) & \text{if } n \text{ is odd.} \end{cases}$$

It is clear that f maps $\mathbb{N}$ into $\mathbb{Z}$. Check that it is one-one and onto. Thus, we have found a bijection from $\mathbb{N}$ onto $\mathbb{Z}$ which shows that $\mathbb{Z}$ is countable. This function is a formal way of saying the we can list the elements of $\mathbb{Z}$ as

$$0, +1, -1, +2, -2, +3, -3, \dots.$$

It is obvious, but good to realize there are wrong ways to try writing such a list. For example, if you list all the negative integers first, as $-1, -2, -3, \dots$, then you will never arrive at 0 or 1, and hence the list is incomplete!

EXAMPLE 33. The set $\mathbb{N} \times \mathbb{N}$ is countable. Rather than give a formula, we list the elements of $\mathbb{Z} \times \mathbb{Z}$ as follows.

$$(1,1),\ (1,2),(2,1),\ (1,3),(2,2),(3,1),\ (1,4),(2,3),(3,2),(4,1),\ \dots$$

The pattern should be clear. Use this list to define a bijection from $\mathbb{N}$ onto $\mathbb{N} \times \mathbb{N}$ and hence show that $\mathbb{N} \times \mathbb{N}$ is countable.

EXAMPLE 34. The set $\mathbb{Z} \times \mathbb{Z}$ is countable. This follows from the first two examples. Indeed, we have a bijection $f : \mathbb{N} \to \mathbb{Z}$ and a bijection $g : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$. Define a bijection $F : \mathbb{Z} \times \mathbb{Z} \to \mathbb{N} \times \mathbb{N}$ by $F(n,m) = f(g(f^{-1}(n), f^{-1}(m)))$. Then, F is one-one and onto. This shows that $\mathbb{Z} \times \mathbb{Z}$ is indeed countable.

EXAMPLE 35. The set of rational numbers $\mathbb{Q}$ is countable. Recall that rational numbers other than 0 can be written uniquely in the form p/q where p is a non-zero integer and q is a strictly positive integer, and there are no common factors of p and q (this is called the *lowest form* of the rational number r). Consider the map $f : \mathbb{Q} \to \mathbb{Z} \times \mathbb{Z}$ defined by

$$f(r) = \begin{cases} (0,1) & \text{if } r = 0 \\ (p,q) & \text{if } r = \frac{p}{q} \text{ in the lowest form.} \end{cases}$$

Clearly, f is injective and hence, it appears that $\mathbb{Z} \times \mathbb{Z}$ is a "bigger set" than $\mathbb{Q}$. Next define the function $g : \mathbb{Z} \to \mathbb{Q}$ by setting $g(n) = n$. This is also injective and hence we may say that "$\mathbb{Q}$ is a bigger set than $\mathbb{N}$.

But we have already seen that $\mathbb{N}$ and $\mathbb{Z} \times \mathbb{Z}$ are in bijection with each other, in that sense, they are of equal size. Since $\mathbb{Q}$ is sandwiched between the two it ought to be true that $\mathbb{Q}$ has the same size as $\mathbb{N}$, and thus countable.

This reasoning is not incorrect, but an argument is needed to make it an honest proof. This is indicated in the Schröder-Bernstein theorem stated later. Use that to fill the gap in the above argument, or alternately, try to directly find a bijection between $\mathbb{Q}$ and $\mathbb{N}$.

EXAMPLE 36. The set of real numbers $\mathbb{R}$ is not countable. The extraordinarily proof of this fact is due to Cantor, and the core idea, called the *diagonalization trick* is one that can be used in many other contexts.

Consider any function $f : \mathbb{N} \to [0, 1]$. We show that it is not onto, and hence not a bijection. Indeed, use the decimal expansion to write a number $x \in [0, 1]$ as $0.x_1x_2x_3\ldots$ where $x_i \in \{0, 1, \ldots, 9\}$. Write the decimal expansion for each of the numbers $f(1), f(2), f(3), \ldots$ as follows.

$$f(1) = 0.X_{1,1}X_{1,2}X_{1,3}\ldots$$

$$f(2) = 0.X_{2,1}X_{2,2}X_{2,3}\ldots$$

$$f(3) = 0.X_{3,1}X_{3,2}X_{3,3}\ldots$$

$$\ldots\ldots\ \ldots\ \ldots$$

Let $Y_1, Y_2, Y_3, \ldots$ be any numbers in $\{0, 1, \ldots, 9\}$ with the only condition that $Y_i \neq X_{i,i}$. Clearly it is possible to choose Y_i like this. Now consider the number $y = 0.Y_1Y_2Y_3\ldots$ which is a number in $[0, 1]$. However, it does not occur in the above list. Indeed, y disagrees with $f(1)$ in the first decimal place, disagrees with $f(2)$ in the second decimal place etc. Thus, $y \neq f(i)$ for any $i \in \mathbb{N}$ which means that f is not onto $[0, 1]$.

Thus, no function $f : \mathbb{N} \to [0, 1]$ is onto, and hence there is no bijection from $\mathbb{N}$ onto $[0, 1]$ and hence $[0, 1]$ is not countable. Obviously, if there is no onto function onto $[0, 1]$, there cannot be an onto function onto $\mathbb{R}$. Thus, $\mathbb{R}$ is also uncountable.

EXAMPLE 37. Let $A_1, A_2, \ldots$ be subsets of a set Ω. Suppose each A_i is countable (finite is allowed). Then $\cup_i A_i$ is also countable. We leave it as an exercise. [*Hint:* If each A_i is countably infinite and pairwise disjoint, then $\cup A_i$ can be thought of as $\mathbb{N} \times \mathbb{N}$].

LEMMA 38 (Schröder-Bernstein). *Let A, B be two sets and suppose there exist injective functions $f : A \to B$ and $g : B \to A$. Then, there exists a bijective function $h : A \to B$.*

This makes it much easier to show that a set is countable as it is generally easier to find injections in both directions than to find a bijection. Before sketching a proof (irrelevant to the rest of the course), let us make a general definition.

Definition 39. Let A and B be sets. We say that the cardinality of A is at most the cardinality of B and write $|A| \leq |B|$ if there is an injection from A into B. If there is a bijection from A onto B, then we say that A and B have the same cardinality and write $|A| = |B|$.

In these terms, the Schröder-Bernstien theorem can be rephrased as saying that if $|A| \leq |B|$ and $|B| \leq |A|$, then $|A| = |B|$. We omit the proof of the Schröder-Bernstien theoremas it is irrelevant to the rest of the course[8].

So far, we have seen many countable sets and some uncountable ones such as $\mathbb{R}$. Cantor's diagonalization method can be adapted to show that given any set A, there is a set of strictly larger cardinality. Recall that the power set of a set A is the set of all subsets of A. We denote it by $\mathcal{P}(A)$.

Proposition 40. *If A is a non-empty set, then there is no surjective function from A onto $\mathcal{P}(A)$.*

Proof. Let $f : A \to \mathcal{P}(A)$ be a function. Let $S = \{x \in A : x \notin f(x)\}$. Then $S \in \mathcal{P}(A)$, but there it is not in the range of f. On the contrary, if it were, say $S = f(y)$, then either $y \in S$ or $y \notin S$. But by the definition of S, if $y \in S$, then S should not contain y and if $y \notin S$, then S should contain y! Thus $S \neq f(y)$ for any y, showing that f is not a surjection. ■

On first sight this looks different from Cantor's diagonalization trick, but it is in fact the same. If $A = \mathbb{N}$, subsets of A can be identified with sequences of zeros and ones in the obvious fashion (one in the kth position means k belongs to the subset). Thus, we may write $f(n) = (x_{n,1}, x_{n,2}, \ldots)$ where $x_{n,j} \in \{0, 1\}$. The diagonalization procedure constructs a new sequence y such that $y_n \neq x_{n,n}$, which means that $y_n = 1$ if $x_{n,n} = 0$ and $y_n = x_{n,n} = 0$. The associated subset $\{n : y_n = 1\}$ is precisely the set S in the above proof.

5. On infinite sums

There were some subtleties in the definition of probabilities which we address now. The definition of $\mathbb{P}(A)$ for an event A and $\mathbb{E}[X]$ for a random variable X involve infinite sums (when Ω is countably infinite). In fact, in the very definition of probability space, we had the condition that $\sum_\omega p_\omega = 1$, but what is the meaning of this sum when Ω is infinite? In this

[8]For those interested, we describe the idea of the proof somewhat informally. Consider the two sets A and B (assumed to have no common elements) and draw a blue arrow from each $x \in A$ to $f(x) \in B$ and a red arrow from each $y \in B$ to $g(y) \in A$. Start at any $x \in A$ or $y \in B$ and follow the arrows in the forward and backward directions. There are only three possibilities (this uses the fact that f and g are injections)

(1) The search closes, and we discover a cycle of alternating blue and red arrows.

(2) The backward search ends after finitely many steps and the forward search continues forever.

(3) Both the backward and forward searches continue forever.

In the first and third case, just use the blue arrows to define the function h. In the second case, if the first element of the chain is in A, then use the blue arrows, and if the first element is in B, then use the red arrows (but in reverse direction) to define the function h. Check that the resulting function is a bijection!

section, we make precise the notion of infinite sums. In fact we shall give two methods of approach, it suffices to consider only the first.

5.1. First approach. Let Ω be a countable set, and let $f : \Omega \to \mathbb{R}$ be a function. We want to give a meaning to the infinite sum $\sum_{\omega\in\Omega} f(\omega)$. First we describe a natural attempt and then address the issues that it leaves open. To avoid discussion of trivialities, assume that Ω is an infinite set.

The idea: By definition of countability, there is a bijection $\varphi : \mathbb{N} \to \Omega$ which allows us to list the elements of Ω as $\omega_1 = \varphi(1), \omega_2 = \varphi(2), \ldots$. Consider the partial sums $x_n = f(\omega_1) + f(\omega_2) + \ldots + f(\omega_n)$. If $\lim_{n\to\infty} x_n$ exists and is equal to L, we could define L to be the infinite sum $\sum_{\omega\in\Omega} f(\omega)$.

The problem is that this may depend on the bijection φ chosen. For example, if $\psi : \mathbb{N} \to \Omega$ is a different bijection, we would write the elements of Ω in a different sequence $\omega'_1 = \psi(1), \omega'_2 = \psi(2), \ldots$, the partial sums $y_n = f(\omega'_1) + \ldots + f(\omega'_n)$ and then define $\sum_{\omega\in\Omega} f(\omega)$ as the limit $L' = \lim_{n\to\infty} (f(\omega'_1) + \ldots + f(\omega'_n))$.

Is it necessarily true that $L = L'$? Not in general, as an example given later shows. But if all arrangements do give the same limit, then there is no ambiguity.

Definition 41. Suppose that $\lim_{n\to\infty}(f(\varphi(1) + \ldots + \varphi(n)))$ exists and has the same value L for all bijections $\varphi : \mathbb{N} \to \Omega$, then we say that $\sum_{\omega\in\Omega} f(\omega)$ converges and has the value L.

While unambiguous, this appears impossible to check! How can we go over all bijections from $\mathbb{N}$ to Ω and check that they give the same answer. This is where the following theorem comes in handy.

Theorem 42. *Let $f : \Omega \to \mathbb{R}$ where Ω is countable. The following are equivalent:*

(1) $\sum_{\omega\in\Omega} f(\omega)$ converges according to Definition 41.

(2) For some bijection $\psi : \mathbb{N} \to \Omega$, the series $\sum_n |f(\psi(n))|$ converges.

The second condition is called *absolute convergence*. We only give a proof that (2) $\Longrightarrow$ (1), which is the useful part. The reason for (1) $\Longrightarrow$ (2) is indicated in Example 47.

Proof that (2) $\Longrightarrow$ (1). The proof is in two steps, first for non-negative f and then for general f.

Non-negative f: Let φ, ψ be two bijections from $\mathbb{N}$ to Ω and write $\omega_i = \varphi(i)$, $\underline{\omega}'_i = \psi(i)$ for simplicity of notation. Let $x_n = f(\omega_1) + \ldots + f(\omega_n)$ and $x'_n = f(\omega'_1) + \ldots + f(\omega'_n)$. Since $f \geq 0$, both (x_n) and (x'_n) are increasing sequences and hence converge in $[0, +\infty]$, to $L = \sup_n x_n$ and $L' = \sup_n x'_n$ respectively. As ψ is surjective, for any n, there is some m_n (possibly very large) such that $\{\omega_1, \ldots, \omega_n\} \subseteq \{\omega'_1, \ldots, \omega'_{m_n}\}$. But then $x_n \leq x'_{m_n}$. This shows that $L \leq L'$. Similarly, $L' \leq L$ and therefore, $L = L'$.

General $f : \Omega \to \mathbb{R}$: Fix bijection φ, ψ as before and define the partial sums x_n, x'_n in the same way. In addition, define $y_n = |f(\omega_1)| + \ldots + |f(\omega_n)|$ and $y'_n = |f(\omega'_1)| + \ldots + |f(\omega'_n)|$. By

assumption y_n and y_n' converge to the same finite number. Observe that $|f| - f$ is a non-negative functions whose partial sums under φ are $y_n - x_n$ while under ψ they are $y_n' - x_n'$. By the first part, $\lim(y_n - x_n)$ and $\lim(y_n' - x_n')$ exist and are equal (but possibly equal to $+\infty$). Therefore (note the use of finiteness of $\lim y_n$ here)

$$\lim x_n = \lim(x_n - y_n) + \lim y_n = \lim(x_n' - y_n') + \lim y_n' = \lim x_n'$$

completing the proof. ■

Remark 43. The proof shows that for a non-negative f,

$$\sum_{\omega\in\Omega} f(\omega) = \sup\left\{\sum_{\omega\in A} f(\omega) : A \subseteq \Omega \text{ is finite}\right\}$$

where both sides may be $+\infty$.

The usual properties of summation, without which life would not be worth living, remain valid.

Exercise 44. Let $f, g : \Omega \to \mathbb{R}_+$ and $a, b \in \mathbb{R}$. If $\sum f$ and $\sum g$ converge absolutely, then $\sum(af + bg)$ converges absolutely and $\sum(af + bg) = a\sum f + b\sum g$. Further, if $f(\omega) \le g(\omega)$ for all $\omega \in \Omega$, then $\sum f \le \sum g$.

Here is a more general and useful fact that we state without proof.

Theorem 45 (Fubini's theorem for sums). *Let $U = S \times T$, where S, T (and hence also U) are countable sets. Let $h : U \to \mathbb{R}$ be absolutely summable. Then,*

$$\sum_{u\in U} h(u) = \sum_{s\in S}\sum_{t\in T} h(s,t) = \sum_{t\in T}\sum_{s\in S} h(s,t). \tag{1}$$

The middle term in (1) is to be understood as follows: Fix s and form the function $h_s : T \to \mathbb{R}$ by $h_s(t) = h(s,t)$. Then h_s is absolutely summable for each s, and hence its sum L_s is well-defined. The function $s \mapsto L_s$ on S is also absolutely summable, and its sum is what is meant by $\sum_{s\in S}\sum_{t\in T} h(s,t)$. Similarly for the last term in (1).

As a corollary, we can prove an extension of Exercise 44 for a countable family of functions.

Corollary 46. *Let $f_n : \Omega \to \mathbb{R}$ for $n \in \mathbb{N}$. Assume that $\sum_{(\omega,n)\in\Omega\times\mathbb{N}} |f_n(\omega)| < \infty$. Then*

$$\sum_{n\in\mathbb{N}}\sum_{\omega\in\Omega} f_n(\omega) = \sum_{\omega\in\Omega} f(\omega)$$

where $f(\omega) = \sum_{n\in\mathbb{N}} f_n(\omega)$. If $f_n \ge 0$ for all n, the equality holds without any condition, except that both sides could be $+\infty$.

Proof. Define $h : \Omega \times \mathbb{N} \to \mathbb{R}$ by $h(\omega, n) = f_n(\omega)$. Then

$$\sum_{\omega\in\Omega}\sum_{n\in\mathbb{N}} h(\omega,n) = \sum_{\omega} f(\omega),$$

$$\sum_{n\in\mathbb{N}}\sum_{\omega\in\Omega} h(\omega,n) = \sum_{n\in\mathbb{N}}\sum_{\omega\in\Omega} f_n(\omega).$$

Fubbini's theorem (check the conditions!) says that the two iterated sums are equal. ■

EXAMPLE 47. This example will illustrate why we refuse to assign a value to $\sum_\omega f(\omega)$ in some cases. Let $\Omega = \mathbb{Z}$ and define $f(0) = 0$ and $f(n) = 1/n$ for $n \neq 0$. At first one may like to say that $\sum_{n\in\mathbb{Z}} f(n) = 0$, since we can cancel $f(n)$ and $f(-n)$ for each n. However, following our definitions

$$f_+(n) = \begin{cases} \frac{1}{n} & \text{if } n \geq 1 \\ 0 & \text{if } n \leq 0, \end{cases} \qquad f_-(n) = \begin{cases} \frac{1}{n} & \text{if } n \leq -1 \\ 0 & \text{if } n \geq 0. \end{cases}$$

Hence S_+ and S_- are both $+\infty$ which means our definition does not assign any value to the sum $\sum_\omega f(\omega)$.

Indeed, by ordering the numbers appropriately, we can get any value we like! For example, here is how to get 10. We know that $1 + \frac{1}{2} + \ldots + \frac{1}{n}$ grows without bound. Just keep adding these positive number till the sum exceeds 10 for the first time. Then start adding the negative numbers $-1 - \frac{1}{2} - \ldots - \frac{1}{m}$ till the sum comes below 10. Then add the positive numbers $\frac{1}{n+1} + \frac{1}{n+2} + \ldots + \frac{1}{n'}$ till the sum exceeds 10 again, and then negative numbers till the sum falls below 10 again, etc. Using the fact that the individual terms in the series are going to zero, it is easy to see that the partial sums then converge to 10. There is nothing special about 10, we can get any number we want!

One last remark on why we assumed Ω to be countable.

REMARK 48. What if Ω is uncountable? Take any $f : \Omega \to \mathbb{R}_+$. Define the sets $A_n = \{\omega : f(\omega) \geq 1/n\}$. For some n, if A_n has infinitely many elements, then clearly the only reasonable value that we can assign to $\sum f(\omega)$ is $+\infty$ (since the sum over elements of A_n itself is larger than any finite number). Therefore, for $\sum f(\omega)$ to be a finite number it is essential that A_n is a finite set for each set.

Now, a countable union of finite sets is countable (or finite). Therefore $A = \bigcup_n A_n$ is a countable set. But note that A is also the set $\{\omega : f(\omega) > 0\}$ (since, if $f(\omega) > 0$ it must belong to some A_n). Consequently, even if the underlying set Ω is uncountable, our function will have to be equal to zero except on a countable subset of Ω. In other words, we are reduced to the case of countable sums!

6. Exercises

Note: It is good practise to write the probability space before calculating the probability!

PROBLEM 1. From a box with 100 coupons labelled $1, 2, \ldots, 100$, two are sampled one after another without replacement. Assume that all possible results have the same probability. What is the probability that (a) the first number is odd? (b) the second number is odd? (c) both numbers are odd?

PROBLEM 2. Repeat the previous problem if the two coupons are sampled with replacement.

PROBLEM 3. On a chessboard, a white piece and the black king are placed on two distinct squares chosen at random (assume all possibilities are equally likely). What is the chance that the black king is in a position of *check* if the white piece placed is (a) a rook? (b) a bishop?

PROBLEM 4. A standard deck of cards is shuffled. What is the chance that the (a) the aces are all together (four consecutive cards)? (b) the spades occur in the natural order (i.e., A,1,2,...,10,J,Q,K)?

PROBLEM 5. A drunken man returns home and tries to open the lock of his house from a bunch of n keys by trying them at random till the door opens. Find the probability of the event that he opens it on the kth attempt in both the following cases: (1) He is so drunk that he may try the same key several times. (2) He is moderately drunk and remembers which keys he has already tried.

PROBLEM 6. On a book shelf, there are 52 books, and exactly 2 books have titles starting with any given letter in the English alphabet. If the books are arranged at random, what is the chance that the first letters of the title are ordered?

PROBLEM 7. A deck of 52 cards is shuffled well and 3 cards are dealt. Find the probability of the event that all three cards are from distinct suits.

PROBLEM 8. Place b indistinguishable blue balls and r indistinguishable red balls into m labelled bins, uniformly at random. Find the probability of the event that the first bin contains balls of both colors.

PROBLEM 9. A coin with probability p of turning up H (assume $0 < p < 1$) is tossed till we get a TH or a HT (i.e., two consecutive tosses must be different, eg., TTH or $HHHT$). Find the probability of the event that at least 5 tosses are required.

PROBLEM 10. In a sequence of coin tosses, show that the chance that m heads appear before n tails is equal to $\sum_{k=m}^{m+n-1} \binom{m+n-1}{k}$.

PROBLEM 11. Let $\mathbf{x} = (0,1,1,1,0,1)$ and $\mathbf{y} = (1,1,0,1,0,1)$. A new 6-tuple $\mathbf{z}$ is created at random by choosing each z_i to be x_i or y_i with equal chance, for $1 \leq i \leq 6$ (A toy model for how two DNA sequences can recombine to give a new one). Find the probability of the event that $\mathbf{z}$ is identical to $\mathbf{x}$.

PROBLEM 12. From a group of W women and M men, a team of L people is chosen at random (of course $L \leq W + M$). Find the probability of the event that the teams consists of exactly k women.

PROBLEM 13. Place r distinguishable balls in m labelled bins in such a way that each bin contains at most one ball. All *distinguishable* arrangements are deemed equally likely (this is known as Fermi-Dirac statistics). Find the probability that the first bin is empty.

PROBLEM 14. If r balls are placed in n bins at random, what is the probability that there is no empty bin if (a) $r = n$, (b) $r = n + 1$, (c) $r = n + 2$.

PROBLEM 15. A box contains $2N$ coupons labelled $1, 2, \ldots, 2N$. Draw k coupons (assume $k \leq N$) from the box one after another (1) with replacement, (2) without replacement. Find the probability of the event that no even numbered coupon is in the sample.

PROBLEM 16. A fair die is rolled repeatedly till a six shows up. What is the probability that no five appeared?

PROBLEM 17. Find the probability of the event that the total number of tosses of a coin is at least N if (a) the coin is tossed till we get two consecutive heads, (b) the coin is tossed till we get two (not necessarily consecutive) heads.

PROBLEM 18. A die is thrown till we see the number 6 turn up five times (not necessarily in succession). Find the probability that the number 1 is never seen.

PROBLEM 19. Three fair dice are rolled and the numbers that turn up are added. A probabilist reasoned that 11 and 12 are both equally likely, because there are six ways that either of them could arise:

$$11 = 6+4+1 = 6+3+2 = 5+5+1 = 5+4+2 = 5+3+3 = 4+4+3$$

$$12 = 6+5+1 = 6+4+2 = 6+3+3 = 5+5+2 = 5+4+3 = 4+4+4$$

Is this reasoning correct? In an experiment of 10000 trials, 11 showed up 1264 times while 12 showed up 1211 times.

PROBLEM 20. Suppose the probability for any person to be born on the kth day of the year is p_k, $1 \leq k \leq 365$, where $p_1 + \ldots + p_{365} = 1$. For two unrelated people, what is the chance that they have the same birthday? For what choice of $(p_1, \ldots, p_{365})$ is the probability maximized or minimized?

PROBLEM 21. In a class with 108 people, one student gets a joke by e-mail. He/she forwards it to one randomly chosen classmate. The recipient does the same - chooses a classmate at random (could be the sender too) and forwards it to him/her. The process goes on like this for 20 steps and stops. What is the probability that the first person to get the mail does not get it again? What is the chance that no one gets the e-mail more than once?

PROBLEM 22. Write the probability spaces for the following experiments. Coins and dice may not be fair!

(1) A coin is tossed till we get a head followed immediately by a tail. Find the probability of the event that the total number of tosses is at least N.

(2) A die is thrown till we see the number 6 turn up five times (not necessarily in succession). Find the probability that the number 1 is never seen.

(3) A coin is tossed till the first time when the number of heads (strictly) exceeds the number of tails. What is the probability that the number of tosses is at least 5.

(4) (Extra exercise for fun! Do not submit this part) In the previous experiment, find the probability that the number of tosses is more than N.

PROBLEM 23. (Feller, II.10.8) What is the probability that among k digits *(a)* 0 does not appear; *(b)* 1 does not appear; *(c)* neither 0 nor 1 appears; *(d)* at least one of the two digits 0 or 1 does not appear? Let A and B represent the events in *(a)* and *(b)*. Express the other events in terms of A and B.

PROBLEM 24. (Feller, II.10.20) From a population of N elements a sample of size k is taken. Find the probability that none of m prescribed elements will be included in the sample, assuming the sample to be *(a)* without replacement, *(b)* with replacement. Compare the numerical values for the two methods when *(i)* $N = 100$, $m = k = 3$, and *(ii)* $N = 100$, $m = k = 10$.

PROBLEM 25. (Feller, II.10.39) If r_1 indistinguishable red balls and r_2 indistinguishable blue balls are placed into n cells, find the number of distinguishable arrangements.

PROBLEM 26. (Feller, II.10.40) If r_1 dice and r_2 coins are thrown, how many results can be distinguished?

PROBLEM 27. A deck of n cards labelled $1, 2, \ldots, n$ is shuffled well. Find the probability that the digits *(a)* 1 and 2, *(b)* 1, 2, and 3, appear as neighbours in the order named. Find the probability that they occur in the order named, not necessarily consecutively.

PROBLEM 28. A deck of n cards labelled $1, 2, \ldots, n$ is shuffled well. Find the probability that the digits *(a)* 1 and 2, *(b)* 1, 2, and 3, appear as neighbours in the order named.

PROBLEM 29. (Feller, II.12.1) Prove the following identities for $n \geq 2$. [*Convention:* Let n be a positive integer. Then $\binom{n}{y} = 0$ if y is not an integer or if $y > n$].

$$1 - \binom{n}{1} + \binom{n}{2} - \ldots = 0$$

$$\binom{n}{1} + 2\binom{n}{2} + 3\binom{n}{3} + \ldots = n2^{n-1}$$

$$\binom{n}{1} - 2\binom{n}{2} + 3\binom{n}{3} - \ldots = 0$$

$$2.1\binom{n}{2} + 3.2\binom{n}{3} + 4.3\binom{n}{4} + \ldots = n(n-1)2^{n-2}$$

PROBLEM 30. (Feller, I.12.10) Prove that

$$\binom{n}{0}^2 + \binom{n}{1}^2 + \ldots + \binom{n}{n}^2 = \binom{2n}{n}.$$

Problem 31. (Feller, I.12.20) Using Stirling's formula, prove that $\frac{1}{2^{2n}}\binom{2n}{n} \sim \frac{1}{\sqrt{\pi n}}$. [*Convention:* $a_n \sim b_n$ is shorthand for $\lim_{n\to\infty} \frac{a_n}{b_n} = 1$].

Problem 32. Two positive integers $m \leq N$ are fixed. A box contains N coupons labelled $1, 2, \ldots, N$. A sample of m coupons is drawn.

(1) Write the probability space in the following two ways of drawing the sample.
   (a) (Sampling without replacement). A coupon is drawn uniformly at random, then a second coupon is drawn uniformly at random, and so on, till we have m coupons.
   (b) (Sampling with replacement). A coupon is drawn uniformly at random, its number is noted and the coupon is replaced in the box. Then a coupon is drawn at random from the box, the number is noted, and the coupon is returned to the box. This is done m times.

(2) Let $N = k + \ell$ where k, ℓ are positive integers. We think of $\{1, 2, \ldots, k\}$ as a sub-population of the whole population $\{1, 2, \ldots, N\}$. For each of the above two schemes of sampling (with and without replacement), calculate the probability that the sample of size m contains no elements from the sub-population $\{1, 2, \ldots, k\}$.

Problem 33. A particular segment of the DNA in a woman is *ATTAGCGG* and the corresponding segment in her husband is *CTAAGGCG*. Write the probability space for the same DNA segment in the future child of this man-woman pair. Assume that all possible combinations are equally likely, and ignore the possibility of mutation.

CHAPTER 2

Events and their probabilities

1. Basic rules of probability

So far we have defined the notion of probability space and probability of an event. But most often, we do not calculate probabilities from the definition. This is like in integration, where one defines the integral of a function as a limit of Riemann sums, but that definition is used only to find integrals of x^n, $\sin(x)$ and a few such functions. Instead, integrals of complicated expressions such as $x\sin(x) + 2\cos^2(x)\tan(x)$ are calculated using substitution rule, integration by parts and other such rules. In probability we need some similar rules relating probabilities of various combinations of events to the individual probabilities.

PROPOSITION 34. *Let $(\Omega, p.)$ be a discrete probability space.*

(1) For any event A, we have $0 \leq \mathbb{P}(A) \leq 1$. Also, $\mathbb{P}(\emptyset) = 0$ and $\mathbb{P}(\Omega) = 1$.

(2) Finite additivity of probability: *If $A_1, \ldots, A_n$ are pairwise disjoint events, then $\mathbb{P}(A_1 \cup \ldots \cup A_n) = \mathbb{P}(A_1) + \ldots + \mathbb{P}(A_n)$. In particular, $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$ for any event A.*

(3) Countable additivity of probability: *If $A_1, A_2, \ldots$ is a countable collection of pairwise disjoint events, then $\mathbb{P}(\cup A_i) = \sum_i \mathbb{P}(A_i)$.*

All of these may seem obvious, and indeed they would be totally obvious if we stuck to finite sample spaces. But the sample space could be countable, and then probability of events may involve infinite sums which need special care in manipulation. Therefore we must give a proof. In writing a proof, and in many future contexts, it is useful to introduce the following notation.

Notation: Let $A \subseteq \Omega$ be an event. Then, we define a function $\mathbf{1}_A : \Omega \to \mathbb{R}$, called the *indicator function of A*, as follows.

$$\mathbf{1}_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{if } \omega \notin A. \end{cases}$$

Since a function from Ω to $\mathbb{R}$ is called a random variable, the indicator of any event is a random variable. All information about the event A is in its indicator function (meaning, if we know the value of $\mathbf{1}_A(\omega)$, we know whether or not ω belongs to A). For example, we can write $\mathbb{P}(A) = \sum_{\omega\in\Omega} \mathbf{1}_A(\omega)p_\omega$.

Now we prove the proposition.

PROOF. (1) By definition of probability space $\mathbb{P}(\Omega) = 1$ and $\mathbb{P}(\emptyset) = 0$. If A is any event, then $\mathbf{1}_\emptyset(\omega)p_\omega \leq \mathbf{1}_A(\omega)p_\omega \leq \mathbf{1}_\Omega(\omega)p_\omega$. By Exercise 44, we get

$$\sum_{\omega\in\Omega} \mathbf{1}_\emptyset(\omega)p_\omega \leq \sum_{\omega\in\Omega} \mathbf{1}_A(\omega)p_\omega \leq \sum_{\omega\in\Omega} \mathbf{1}_\Omega(\omega)p_\omega.$$

As observed earlier, these sums are just $\mathbb{P}(\emptyset)$, $\mathbb{P}(A)$ and $\mathbb{P}(\Omega)$, respectively. Thus, $0 \leq \mathbb{P}(A) \leq 1$.

(2) It suffices to prove it for two sets (why?). Let A, B be two events such that $A \cap B = \emptyset$. Let $f(\omega) = p_\omega \mathbf{1}_A(\omega)$ and $g(\omega) = p_\omega \mathbf{1}_B(\omega)$ and $h(\omega) = p_\omega \mathbf{1}_{A\cup B}(\omega)$. Then, the disjointness of A and B implies that $f(\omega) + g(\omega) = h(\omega)$ for all $\omega \in \Omega$. Thus, by Exercise 44, we get

$$\sum_{\omega\in\Omega} f(\omega) + \sum_{\omega\in\Omega} g(\omega) = \sum_{\omega\in\Omega} h(\omega).$$

But the three sums here are precisely $\mathbb{P}(A)$, $\mathbb{P}(B)$ and $\mathbb{P}(A\cup B)$. Thus, we get $\mathbb{P}(A\cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

(3) This is similar to finite additivity but needs a more involved argument. We leave it as an exercise for the interested reader. ■

EXERCISE 35. Adapt the proof to prove that for a countable family of events A_k in a common probability space (no disjointness assumed), we have

$$\mathbb{P}\left(\bigcup_k A_k\right) \leq \sum_k \mathbb{P}(A_k).$$

DEFINITION 36 (Limsup and liminf of sets). If A_k, $k \geq 1$, is a sequence of subsets of Ω, we define

$$\limsup A_k = \bigcap_{N=1}^{\infty} \bigcup_{k=N}^{\infty} A_k, \qquad \text{and} \qquad \liminf A_k = \bigcup_{N=1}^{\infty} \bigcap_{k=N}^{\infty} A_k.$$

In words, $\limsup A_k$ is the set of all ω that belong to infinitely many of the A_ks, and $\liminf A_k$ is the set of all ω that belong to all but finitely many of the A_ks.

Two special cases are of increasing and decreasing sequences of events. This means $A_1 \subseteq A_2 \subseteq A_3 \subseteq \ldots$ and $A_1 \supseteq A_2 \supseteq A_3 \supseteq \ldots$. In these cases, the limsup and liminf are the same (so we refer to it as the limit of the sequence of sets). It is $\cup_k A_k$ in the case of increasing events and $\cap_k A_k$ in the case of decreasing events.

EXERCISE 37. Events below are all contained in a discrete probability space. Use countable additivity of probability to show that

(1) If A_k are increasing events with limit A, show that $\mathbb{P}(A)$ is the increasing limit of $\mathbb{P}(A_k)$.

(2) If A_k are decreasing events with limit A, show that $\mathbb{P}(A)$ is the decreasing limit of $\mathbb{P}(A_k)$.

Now we re-write the basic rules of probability as follows.

The basic rules of probability:

(1) $\mathbb{P}(\emptyset) = 0$, $\mathbb{P}(\Omega) = 1$ and $0 \leq \mathbb{P}(A) \leq 1$ for any event A.

(2) $\mathbb{P}\left(\bigcup_k A_k\right) \leq \sum_k \mathbb{P}(A_k)$ for any countable collection of events A_k.

(3) $\mathbb{P}\left(\bigcup_k A_k\right) = \sum_k \mathbb{P}(A_k)$ if A_k is a countable collection of pairwise disjoint events.

2. Inclusion-exclusion formula

In general, there is no simple rule for $\mathbb{P}(A \cup B)$ in terms of $\mathbb{P}(A)$ and $\mathbb{P}(B)$. Indeed, consider the probability space $\Omega = \{0, 1\}$ with $p_0 = p_1 = \frac{1}{2}$. If $A = \{0\}$ and $B = \{1\}$, then $\mathbb{P}(A) = \mathbb{P}(B) = \frac{1}{2}$ and $\mathbb{P}(A \cup B) = 1$. However, if $A = B = \{0\}$, then $\mathbb{P}(A) = \mathbb{P}(B) = \frac{1}{2}$ as before, but $\mathbb{P}(A \cup B) = \frac{1}{2}$. This shows that $\mathbb{P}(A \cup B)$ cannot be determined from $\mathbb{P}(A)$ and $\mathbb{P}(B)$. Similarly for $\mathbb{P}(A \cap B)$ or other set constructions.

However, it is easy to see that $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$. This formula is not entirely useless, because in special situations we shall later see that the probability of the intersection is easy to compute and hence we may compute the probability of the union. Generalizing this idea to more than two sets, we get the following surprisingly useful formula.

PROPOSITION 38 (Inclusion-Exclusion formula). *Let (Ω, p) be a probability space and let $A_1, \ldots, A_n$ be events. Then,*

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = S_1 - S_2 + S_3 - \ldots + (-1)^{n-1} S_n$$

where

$$S_k = \sum_{1 \leq i_1 < i_2 < \ldots < i_k \leq n} \mathbb{P}(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}).$$

We give two proofs, but the difference is only superficial. It is a good exercise to reason out why the two arguments are basically the same.

FIRST PROOF. For each $\omega \in \Omega$ we compute its contribution to the two sides. If $\omega \notin \bigcup_{i=1}^n A_i$, then p_ω is not counted on either side. Suppose $\omega \in \bigcup_{i=1}^n A_i$ so that p_ω is counted once on the left side. We count the number of times p_ω is counted on the right side by splitting into cases depending on the exact number of A_is that contain ω.

Suppose ω belongs to exactly one of the A_is. For simplicity let us suppose that $\omega \in A_1$ but $\omega \in A_i^c$ for $2 \leq i \leq n$. Then p_ω is counted once in S_1 but not counted in $S_2, \ldots, S_n$.

Suppose ω belongs to A_1 and A_2 but not any other A_i. Then p_ω is counted twice in S_1 (once for $\mathbb{P}(A_1)$ and once for $\mathbb{P}(A_2)$) and subtracted once in S_2 (in $\mathbb{P}(A_1 \cap A_2)$). Thus, it is effectively counted once on the right side. The same holds if ω belongs to A_i and A_j but not any other A_ks.

If ω belongs to $A_1, \ldots, A_k$ but not any other A_i, then on the right side, p_ω is added k times in S_1, subtracted $\binom{k}{2}$ times in S_2, added $\binom{k}{3}$ times in S_k and so on. Thus p_ω is effectively counted

$$\binom{k}{1} - \binom{k}{2} + \binom{k}{3} - \ldots + (-1)^{k-1}\binom{k}{k}$$

times. By the Binomial formula, this is just the expansion of $1 - (1-1)^k$ which is 1. ■

Second proof. Use the definition to write both sides of the statement. Let $A = \cup_{i=1}^n A_i$.

$$\text{LHS} = \sum_{\omega \in A} p_\omega = \sum_{\omega \in \Omega} \mathbf{1}_A(\omega) p_\omega.$$

Now we compute the right side. For any $i_1 < i_2 < \ldots < i_k$, we write

$$\mathbb{P}(A_{i_1} \cap \ldots \cap A_{i_k}) = \sum_{\omega \in \Omega} p_\omega \mathbf{1}_{A_{i_1} \cap \ldots \cap A_{i_k}}(\omega) = \sum_{\omega \in \Omega} p_\omega \prod_{\ell=1}^{k} \mathbf{1}_{A_{i_\ell}}(\omega).$$

Hence, the right hand side is given by adding over $i_1 < \ldots < i_k$, multiplying by $(-1)^{k-1}$ and then summing over k from 1 to n.

$$\begin{aligned}
\text{RHS} &= \sum_{k=1}^{n} (-1)^{k-1} \sum_{1 \le i_1 < \ldots < i_k \le n} \sum_{\omega \in \Omega} p_\omega \prod_{\ell=1}^{k} \mathbf{1}_{A_{i_\ell}}(\omega) \\
&= \sum_{\omega \in \Omega} \sum_{k=1}^{n} (-1)^{k-1} \sum_{1 \le i_1 < \ldots < i_k \le n} p_\omega \prod_{\ell=1}^{k} \mathbf{1}_{A_{i_\ell}}(\omega) \\
&= -\sum_{\omega \in \Omega} p_\omega \sum_{k=1}^{n} \sum_{1 \le i_1 < \ldots < i_k \le n} \prod_{\ell=1}^{k} (-\mathbf{1}_{A_{i_\ell}}(\omega)) \\
&= -\sum_{\omega \in \Omega} p_\omega \left(\prod_{j=1}^{n} (1 - \mathbf{1}_{A_j}(\omega)) \; - \; 1 \right) \\
&= \sum_{\omega \in \Omega} p_\omega \mathbf{1}_A(\omega).
\end{aligned}$$

because the quantity $\prod_{j=1}^{n} (1 - \mathbf{1}_{A_j}(\omega))$ equals -1 if ω belongs to at least one of the A_is, and is zero otherwise. Thus the claim follows. ■

As we remarked earlier, it turns out that in many settings it is possible to compute the probabilities of intersections. We give an example now.

EXAMPLE 39. Let $\Omega = S_{52} \times S_{52}$ with $p_\omega = \frac{1}{(52!)^2}$ for all $\omega \in \Omega$. Consider the event $A = \{(\pi, \sigma) : \pi(i) \neq \sigma(i)\ \forall i\}$. Informally, we imagine two shuffled decks of cards kept side by side (or perhaps one shuffled deck and another permutation denoting a "psychic's predictions" for the order in which the cards occur). Then A is the event that there are no matches (or correct guesses).

Let $A_i = \{(\pi, \sigma) : \pi(i) = \sigma(i)\}$ so that $A^c = A_1 \cup \ldots \cup A_{52}$. It is easy to see that $\mathbb{P}(A_{i_1} \cap A_{i_2} \ldots \cap A_{i_k}) = \frac{1}{52(52-1)\ldots(52-k+1)}$ for any $i_1 < i_2 < \ldots < i_k$ (why?). Therefore, by the inclusion-exclusion formula, we get

$$\begin{aligned}
\mathbb{P}(A^c) &= \binom{52}{1}\frac{1}{52} - \binom{52}{2}\frac{1}{52 \times 51} + \ldots + (-1)^{51}\binom{52}{52}\frac{1}{52 \times 51 \times \ldots \times 1} \\
&= 1 - \frac{1}{2!} + \frac{1}{3!} - \frac{1}{4!} + \ldots - \frac{1}{52!} \\
&\approx 1 - \frac{1}{e} \approx 0.6321
\end{aligned}$$

by the expansion $e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots$ at $x = -1$. Hence $\mathbb{P}(A) \approx e^{-1} \approx 0.3679$.

EXAMPLE 40. Place r distinguishable balls in m distinguishable urns at random. Let A be the event that some urn is empty. The probability space is $\Omega = \{\underline{\omega} = (\omega_1, \ldots, \omega_r) : 1 \leq \omega_i \leq m\}$ with $p_{\underline{\omega}} = m^{-r}$. Let $A_\ell = \{\underline{\omega} : \omega_i \neq \ell\}$ for $\ell = 1, 2 \ldots, m$. Then, $A = A_1 \cup \ldots \cup A_m$.

It is easy to see that $\mathbb{P}(A_{i_1} \cap \ldots \cap A_{i_k}) = (m-k)^r m^{-r}$ for any $i_1 < \ldots < i_k$. Therefore, in the inclusion-exclusion formula, $S_k = \binom{m}{k}\frac{(m-k)^r}{m^r}$. As $S_m = 0$ (meaning that all urns cannot be empty), the inclusion-exclusion formula becomes

$$\mathbb{P}(A) = \binom{m}{1}\left(1 - \frac{1}{m}\right)^r - \binom{m}{2}\left(1 - \frac{2}{m}\right)^r + \ldots + (-1)^{m-1}\binom{m}{m-1}\left(1 - \frac{m-1}{m}\right)^r.$$

The last term is zero (since all urns cannot be empty). I don't know if this expression can be simplified any more.

But here is a fun consequence. If $r < m$, then at least one bin must be empty, hence $\mathbb{P}(A) = 1$. This gives the identity

$$\sum_{k=1}^{m-1}(-1)^{k-1}\binom{m}{k}(m-k)^r = m^r \quad \text{for } 1 \leq r < m. \tag{2}$$

Try proving this directly!

Recall the problem that we asked earlier and postponed the solution of: To find the probability that every month has the birthday of someone in a group oof k people. If we assume that all the months are of equal size, then the above example with $m = 12$ and $r = k$ solves that question. We leave it as an exercise to solve it in general, if the ith month has d_i days.

2.1. Number of occurrences. We mention two useful formulas that can be proved on lines similar to the inclusion-exclusion principle. If we say "at least one of the events $A_1, A_2, \ldots, A_n$ occurs", we are talking about the union, $A_1 \cup A_2 \cup \ldots \cup A_n$. What about "at least m of the events

$A_1, A_2, \ldots, A_n$ occur", how to express it with set operations? It is not hard to see that this set is precisely

$$B_m = \bigcup_{1 \le i_1 < i_2 < \ldots < i_m \le n} (A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_m}).$$

The event that "exactly m of the events $A_1, A_2, \ldots, A_n$ occur" can be written as

$$C_m = B_m \setminus B_{m+1} = \bigcup_{\substack{S \subseteq [n] \\ |S|=m}} \left(\bigcap_{i \in S} A_i \right) \bigcap \left(\bigcap_{i \notin S} A_i^c \right).$$

We give two proofs of the following proposition, the first is combinatorial, similar to the proof we gave earlier for the inclusion-exclusion formula. Later we give a different kind of proof.

PROPOSITION 41. *Let $A_1, \ldots, A_n$ be events in a probability space (Ω, p) and let $m \le n$. Let B_m and C_m be as above. Then,*

$$\begin{aligned} \mathbb{P}(B_m) &= \sum_{k=m}^{n} (-1)^{k-m} \binom{k-1}{k-m} S_k \\ &= S_m - \binom{m}{1} S_{m+1} + \binom{m+1}{2} S_{m+2} - \binom{m+2}{3} S_{m+3} + \ldots \qquad (3) \\ \mathbb{P}(C_m) &= \sum_{k=m}^{n} (-1)^{k-m} \binom{k}{m} S_k \\ &= S_m - \binom{m+1}{1} S_{m+1} + \binom{m+2}{2} S_{m+2} - \binom{m+3}{3} S_{m+3} + \ldots \qquad (4) \end{aligned}$$

PROOF. If ω belongs to less than m of the A_is, then p_ω is not counted on either side of (4). If ω belongs to exactly ℓ of the A_is for some $\ell \ge m$, then p_ω is counted $\binom{\ell}{k}$ times in S_k, hence the total coefficient of p_ω on the right side of (4) is

$$\begin{aligned} &\binom{\ell}{m}\binom{m}{0} - \binom{\ell}{m+1}\binom{m+1}{1} + \binom{\ell}{m+2}\binom{m+2}{2} - \ldots \\ &= \frac{\ell!}{m!(\ell-m)!} \left(\binom{\ell-m}{0} - \binom{\ell-m}{1} + \binom{\ell-m}{2} - \ldots \right) = \binom{\ell}{m} (1-1)^{\ell-m} \end{aligned}$$

which is 1 if $\ell = m$ and 0 if $\ell > m$. This proves (4).

Next, B_m is the disjoint union of $C_m, C_{m+1}, \ldots$, hence using (4),

$$\begin{aligned} \mathbb{P}(B_m) &= \sum_{r=m}^{n} \sum_{k=r}^{n} (-1)^{k-r} \binom{k}{r} S_k = \sum_{k=m}^{n} S_k \sum_{r=m}^{k} (-1)^{k-r} \binom{k}{r} \\ &= \sum_{k=m}^{n} S_k (-1)^{k-m} \binom{k-1}{k-m} \end{aligned}$$

where we used the identity $\sum_{r=m}^{k} (-1)^r \binom{k}{r} = (-1)^m \frac{m}{k} \binom{k}{m}$. ∎

EXERCISE 42. Return to the setting of exercise 38 but with n cards in a deck, so that $\Omega = S_n \times S_n$ and $p_{(\pi,\sigma)} = \frac{1}{(n!)^2}$. Let A_m be the event that there are exactly m matches between the two decks.

(1) For fixed $m \geq 0$, show that $\mathbb{P}(A_m) \to e^{-1}\frac{1}{m!}$ as $n \to \infty$.

(2) Assume that the approximations above are valid for $n = 52$ and $m \leq 10$. Find the probability that there are at least 10 matches.

2.2. Alternate proof of the inclusion-exclusion formula. Here we give another proof of the inclusion-exclusion formula, using the beautiful and important notion of *generating functions*. Let $A_1, A_2, \ldots, A_n$ be events in a probability space (Ω, p). For any $k \geq 0$, let C_k denote the set of $\omega \in \Omega$ that belong to exactly k of the A_is. The generating function of the sequence $(\mathbb{P}(C_0), \ldots, \mathbb{P}(C_n))$ is defined as $\varphi(z) = \sum_{k\geq 0} \mathbb{P}(C_k) z^k$.

Observe that $\omega \in C_k$ if and only if $\mathbf{1}_{A_1}(\omega) + \ldots + \mathbf{1}_{A_n}(\omega) = k$. Hence, we may rewrite

$$\varphi(z) = \sum_{\omega\in\Omega} p_\omega z^{\mathbf{1}_{A_1}(\omega)+\ldots+\mathbf{1}_{A_n}(\omega)} = \sum_{\omega\in\Omega} p_\omega \prod_{j=1}^{n} z^{\mathbf{1}_{A_j}(\omega)}$$

$$= \sum_{\omega\in\Omega} p_\omega \prod_{j=1}^{n} (1 + (z-1)\mathbf{1}_{A_j}(\omega))$$

where we used the obvious fact that $z^\varepsilon = 1 + (z-1)\varepsilon$ if ε is 0 or 1. Now, $\prod_{j=1}^n (1 + a_j) = \sum_{k=0}^n \sum_{j_1<\ldots<j_k} a_{j_1} \ldots a_{j_k}$. Therefore,

$$\varphi(z) = \sum_{\omega\in\Omega} p_\omega \sum_{k=0}^{n} (z-1)^k \sum_{j_1<\ldots<j_k} \mathbf{1}_{A_{j_1}}(\omega) \ldots \mathbf{1}_{A_{j_k}}(\omega)$$

$$= \sum_{k=0}^{n} (z-1)^k \sum_{j_1<\ldots<j_k} \sum_{\omega} \mathbf{1}_{A_{j_1}}(\omega) \ldots \mathbf{1}_{A_{j_k}}(\omega)$$

$$= \sum_{k=0}^{n} (z-1)^k S_k.$$

Expanding $(z-1)^k$ by the binomial theorem and collecting the powers of z together gives

$$\varphi(z) = \sum_{r=0}^{n} z^r \sum_{k=r}^{n} \binom{k}{r} (-1)^{k-r} S_k.$$

Compare with the definition of $\varphi(z)$ and equate the coefficients of z^r to get

$$\mathbb{P}(C_r) = \sum_{k=r}^{n} \binom{k}{r} (-1)^{k-r} S_k.$$

This gives the second part of Proposition 40. The first part of Proposition 40 follows from this as before (or can you find a direct generating function proof starting from $\psi(z) = \sum_k \mathbb{P}(B_k) z^k$?).

2.3. An application to number theory*. If two numbers are picked at random, what is the chance that they are co-prime? As there is no way to pick a number uniformly at random from the infinite set $\mathbb{N}$, we interpret this problem as follows:

Let $\Omega = [N] \times [N]$ and let $p((i,j)) = \frac{1}{N^2}$ for all $(i,j) \in \Omega$. Let

$$\alpha_N = \mathbb{P}\{(a,b) \in \Omega : \text{g.c.d.}(a,b) = 1\}.$$

Find $\lim_{N\to\infty} \alpha_N$, if it exists.

Enumerate the primes below N as $p_1 < p_2 < \ldots, p_m$. Let $E_k = \{(a,b) : p_k \text{ divides both } a \text{ and } b\}$ be the event that (a,b) has a common factor of p_k. Then $E = \cup_{i=1}^m E_k$ is the event that (a,b) have a common factor other than 1. For $i_1 < \ldots < i_k$, the event $A_{i_1,\ldots,i_k}$ is the subset of all (a,b) such that both a and b are divisible by $p_{i_1} \ldots p_{i_k}$. Hence

$$\begin{aligned}\mathbb{P}\{A_{i_1} \cap \ldots \cap A_{i_k}\} &= \frac{1}{N^2}\left(\left\lfloor \frac{N}{p_{i_1}\ldots p_{i_k}} \right\rfloor\right)^2 \\ &\to \frac{1}{p_{i_1}^2 \ldots p_{i_k}^2} \quad \text{as } N \to \infty.\end{aligned}$$

Therefore, $S_k = \sum_{i_1<\ldots<i_k} \frac{1}{p_{i_1}^2 \ldots p_{i_k}^2}$. Hence (there is some lack of rigor here, can you spot it?)

$$\begin{aligned}\lim_{N\to\infty} \alpha_N &= 1 - \sum_i \frac{1}{p_i^2} - \sum_{i<j} \frac{1}{p_i^2 p_j^2} + \sum_{i<j<k} \frac{1}{p_i^2 p_j^2 p_k^2} + \ldots \\ &= \prod_k \left(1 - \frac{1}{p_k^2}\right).\end{aligned}$$

This product over all primes is a positive number between 0 and 1, and that is our answer. In fact, Euler showed that this number is $\frac{6}{\pi^2}$. Again ignoring some issues of infinite series (which can be justified), this can be seen as follows:

$$\begin{aligned}\prod_k \frac{1}{1-\frac{1}{p_k^2}} &= \prod_k \left(1 + \frac{1}{p_k^2} + \frac{1}{p_k^4} + \frac{1}{p_k^6} + \ldots\right) \\ &= \sum_{(r_j)} \frac{1}{p_1^{r_1} p_2^{r_2} \ldots} \quad \text{(sum is over } r_j \geq 0 \text{ satisfying } \sum_j r_j < \infty) \\ &= \sum_{n=1}^{\infty} \frac{1}{n^2}\end{aligned}$$

using the fundamental theorem of Arithmetic. It is well-known (also due to Euler) that this sum is $\frac{\pi^2}{6}$.

2.4. An application to counting trees*. A tree is a connected graph that has no cycles. Cayley showed that there are n^{n-2} trees with vertex set $[n]$ (two such trees are equal if their

edge sets are equal). For example, when $n = 3$, the trees are 1—2—3, 1—3—2 and 2—1—3. We prove Cayley's theorem by inclusion-exclusion[1].

Let $\mathcal{T}_n$ denote the set of trees with vertex et $[n]$. Let $b_n = \#\mathcal{T}_n$. Pick a tree uniformly at random from $\mathcal{T}_n$ and denote it T. Let A_k be the event that vertex k is a leaf of T (i.e., k appears in exactly one edge). Then $A_1 \cup \ldots \cup A_n$ has probability 1, since every tree must have a leaf. By inclusion-exclusion,

$$1 = S_1 - S_2 + S_3 + \ldots + (-1)^{n-1} S_n = 0.$$

Let us compute S_k. Fix $i_1 < \ldots < i_k$ and consider $A_{i_1} \cap \ldots A_{i_k}$. This means that vertices $i_1, \ldots, i_k$ are all leaves. To do that, first we must choose a tree whose vertex set is $[n] \setminus \{i_1, \ldots, i_k\}$, which can be done in b_{n-k} ways. Then the k leaves can be attached to any of the $n-k$ vertices, which can be done in $(n-k)^k$ ways. Therefore, $\mathbb{P}\{A_{i_1} \cap \ldots A_{i_k}\} = \frac{b_{n-k}}{b_n}(n-k)^k$. Therefore, $S_k = \binom{n}{k} \frac{b_{n-k}}{b_n}(n-k)^k$ and we arrive at

$$b_n = \sum_{k=1}^{n-1} (-1)^{k-1} b_{n-k} \binom{n}{k} (n-k)^k.$$

where we dropped $k = n$ term as $b_0 = 0$ (all vertices can't be leaves). Further, For $k = 1, 2, 3$, by direct enumeration it is easy to check that $b_k = k^{k-2}$ is valid. Inductively assume that $b_k = k^{k-2}$ for $k < n$. Then

$$b_n = \sum_{k=1}^{n-1} (-1)^{k-1} \binom{n}{k} (n-k)^{n-2}.$$

Now refer to the identity (2) with $r = n-2$ which shows that the right side sum is n^{n-2}, provided $n \geq 3$. Thus inductively we have proved that $b_n = n^{n-2}$.

3. Bonferroni's inequalities

Inclusion-exclusion formula is nice when we can calculate the probabilities of intersections of the events under consideration. Things are not always this nice, and sometimes that may be very difficult. Even if we could find them, summing them with signs according to the inclusion-exclusion formula may be difficult as the example 39 demonstrates. The *idea* behind the inclusion-exclusion formula can however be often used to compute *approximate values of probabilities*, which is very valuable in most applications. That is what we do next.

We know that $\mathbb{P}(A_1 \cup \ldots \cup A_n) \leq \mathbb{P}(A_1) + \ldots + \mathbb{P}(A_n)$ for any events $A_1, \ldots, A_n$. This is an extremely useful inequality, often called the *union bound*. Its usefulness is in the fact that there is no assumption made about the events A_is (such as whether they are disjoint or

[1]Thanks to Koushik Ramachandran for telling me about this proof. He pointed to the book by Santosh Venkatesh for this proof. On second thoughts, we used the identity (2) in this proof, and that identity was proved using inclusion-exclusion applied to balls in urns. It makes one wonder if one can "cancel" the two applications of inclusion-exclusion and write a more direct proof by mapping this problem of counting trees to that of balls in urns.

not). The following inequalities generalize the union bound, and gives both upper and lower bounds for the probability of the union of a bunch of events.

LEMMA 43 (Bonferroni's inequalities). *Let $A_1, \ldots, A_n$ be events in a probability space (Ω, p) and let $A = A_1 \cup \ldots \cup A_n$. We have the following upper and lower bounds for $\mathbb{P}(A)$.*

$$\mathbb{P}(A) \leq \sum_{k=1}^{m} (-1)^{k-1} S_k, \quad \textit{for any odd } m.$$

$$\mathbb{P}(A) \geq \sum_{k=1}^{m} (-1)^{k-1} S_k, \quad \textit{for any even } m.$$

PROOF. We shall write out the proof for the cases $m = 1$ and $m = 2$. When $m = 1$, the inequality is just the union bound

$$\mathbb{P}(A) \leq \mathbb{P}(A_1) + \ldots + \mathbb{P}(A_n)$$

which we know. When $m = 2$, the inequality to be proved is

$$\mathbb{P}(A) \geq \sum_{k} \mathbb{P}(A_k) - \sum_{k<\ell} \mathbb{P}(A_k \cap A_\ell)$$

To see this, fix $\omega \in \Omega$ and count the contribution of p_ω to both sides. Like in the proof of the inclusion-exclusion formula, for $\omega \notin A_1 \cup \ldots \cup A_n$, the contribution to both sides is zero. On the other hand, if ω belongs to exactly r of the sets for some $r \geq 1$, then it is counted once on the left side and $r - \binom{r}{2}$ times on the right side. Not that $r - \binom{r}{2} = \frac{1}{2} r(3 - r)$ which is always non-positive (one if $r = 1$, zero if $r = 2$ and non-positive if $r \geq 3$). Hence we get LHS $\geq$ RHS.

Similarly, one can prove the other inequalities in the series. We leave it as an exercise. The key point is that $r - \binom{r}{2} + \ldots + (-1)^{k-1}\binom{r}{k}$ is non-negative if k is odd and non-positive if k is even (prove this). Here as always $\binom{x}{y}$ is interpreted as zero if $y > x$. ■

Of particular importance are the first layer of inequalities,

$$\sum_{i=1}^{n} \mathbb{P}(A_i) - \sum_{1 \leq i < j \leq n} \mathbb{P}(A_i \cap A_j) \leq \mathbb{P}(A_1 \cup \ldots \cup A_n) \leq \sum_{i=1}^{n} \mathbb{P}(A_i).$$

The right side inequality is called the "union bound" and we use it all the time. The left side inequality is less common, but often useful. There is no guarantee that these inequalities are useful. It can happen that $S_1 \geq 1$ and $S_1 - S_2 \leq 0$, in which case we get the true but useless bounds for the probability!

Here is an application of these inequalities.

EXAMPLE 44. Return to Example 39. We obtained an exact expression for the answer, but that is rather complicated. For example, what is the probability of having at least one empty urn when $n = 40$ balls are placed at random in $r = 10$ urns? It would be complicated to sum the series. Instead, we could use Bonferroni's inequalities to get the following bounds.

$$m\left(1 - \frac{1}{m}\right)^r - \binom{m}{2}\left(1 - \frac{2}{m}\right)^r \leq \mathbb{P}(A) \leq m\left(1 - \frac{1}{m}\right)^r.$$

If we take $r = 40$ and $m = 10$, the bounds we get are $0.1418 \leq \mathbb{P}(A) \leq 0.1478$. Thus, we get a pretty decent approximation to the probability. By experimenting with other numbers you can check that the approximations are good when r is large compared to m but not otherwise. Can you reason why?

EXAMPLE 45. Consider the birthday problem with a a group of m people. Let A be the event that at least two people have the same birthday. Then $A = \bigcup_{i=1}^{m-1} \bigcup_{j=i+1}^{m} A_{i,j}$, where $A_{i,j}$ is the event that the ith person and the jth person have the same birthday. Then $\mathbb{P}(A_{i,j}) = \frac{1}{365}$ and $\mathbb{P}(A_{i,j} \cap A_{k,\ell}) = \frac{1}{365^2}$ if $\{i,j\} \neq \{k,\ell\}$ (consider both cases, where $\{i,j\}$ and $\{k,\ell\}$ are disjoint, and where they intersect in one element). Now the number of sets whose union is under consideration is $m(m-1)/2$. Therefore, in the inclusion-exclusion formula, $S_1 = \frac{m(m-1)}{2\times 365}$ and $S_2 < \frac{m^2(m-1)^2}{4\times 365}$.

For example, with $m = 23$, we get $S_1 = 0.69...$ and $S_2 = 0.48...$ so the bounds we get are $0.11 \leq \mathbb{P}(A) \leq 0.69$, which are not very good bounds. But they show that the probability is neither close to zero, nor to 1. Do the bounds get better or worse for larger m? Although worse than the exact answer, this method makes it more transparent why there is no paradox. What we should compare with the number of days in a year is not the number of people, but the number of *pairs* of people. When $m = 25$, there are 253 pairs of people, which is comparable to 365.

EXAMPLE 46 (Thomson's (Lord Kelvin's) version of the birthday paradox). Pour a glass of water (has about 10^{23} molecules) into the oceans of the world (have about 10^{45} molecules), stir the oceans well enough that it gets mixed up, then scoop out a glass of water from the oceans. What is the chance that at least one of the molecules that you poured in was scooped back? Naively one thinks the chance is abysmally close to zero, but it is in fact more than 90%! To see this, observe that the probability that none of the poured molecules was scooped back is

$$\frac{(10^{45}-1)\times(10^{45}-10^{23}-1)\ldots(10^{45}-10^{23}+1)}{(10^{45})^{10^{23}}} = \prod_{k=1}^{10^{23}-1}\left(1-\frac{k}{10^{45}}\right).$$

Using $1-x < e^{-x}$, this is seen to be less than $e^{-5} < 0.007$, which means that the complementary event has more than 0.993 probability.

Again, the correct way to think about the situation is not to compare the glass with the ocean, but that the number of pairs of molecules (first from first glass, second from second glass) is about $\frac{1}{2}10^{23} \times 10^{23} = 5 \times 10^{45}$, which is comparable to 10^{45}. What do Bonferroni's inequalities give in this problem?

4. Simulation

So far we have seen a few techniques to calculate probabilities of various interesting events. Important and useful as they are, there are a vast number of situations where they do

not allow one to compute probabilities. Here are some examples that are minor modifications of problems that we have solved.

▶ Birthday problem: Calculate the probability that in a group of n people there are at least k with a common birthday. When $k = 2$ we have solved this, but even for $k = 3$ it takes some effort to derive an exact formula, and it gets worse with larger k. But if we were in a situation where this probability was of practical use, it is much easier to get the answer using simulation. For example, with $n = 130$ and $k = 3$, a simple code conducting the experiment 10000 times gave the probability of 0.89 (and repeating this entire exercise gives consistent results).

▶ Balls in bins: When r balls are thrown into m bins, find the probability no bin has less than k balls. If $r < km$, then the probability is zero! But it is not easy to calculate it exactly. With $m = 10$, $r = 50$ and $k = 2$, a simulation with 10000 repetitions of the experiment yielded a probability of 0.59 approximately.

▶ Among all 5×5 matrices with entries that are 0 or 1, one is selected uniformly at random. What is the chance that it is singular? This can be calculated exactly on a computer, since the total number of such matrices if $2^{25} = 33554432$ which can be handled on a computer. But if the size of the matrix is increased to 10×10, it becomes impossible. Anyway, this is a difficult probability to calculate. By simulating the random experiment 10000 times, I got the answer of 0.63 approximately.

There are innumerable problems of probability that can be simulated on a computer, even very complicated ones, but are nearly impossible to analyse mathematically.

5. Exercises

PROBLEM 1. (Feller, III.6.3) Find the probability that in five tossings a coin falls head at least three times in succession.

PROBLEM 2. (Feller, III.6.1) Ten pairs of shoes are in a closet. Four shoes are selected at random. Find the probability that there will be at least one pair among the fours shoes selected.

PROBLEM 3. A deck of cards is shuffled and dealt to four players (13 cards each).

(1) Find the chance that at least one of the players has all cards of the same suite.

(2) Find the chance that at least one of the players has all four cards of the same type (i.e., all four aces or all sevens etc.).

PROBLEM 4. Let $A_1, A_2, A_3, \ldots$ be events in a probability space. Write the following events in terms of $A_1, A_2, \ldots$ using the usual set operations (union, intersection, complement).

(1) An infinite number of the events A_i occur.

(2) All except finitely many of the events A_i occur.

(3) Exactly k of the events A_i occur.

PROBLEM 5. (Feller, I.8.1) Let $A_1, \ldots, A_n$ be events in a probability space (Ω, p) and let $0 \leq m \leq n$. Let B_m be the event that at least m of the events $A_1, \ldots, A_n$ occur. Mathematically,

$$B_m = \bigcup_{1 \leq i_1 < i_2 < \ldots < i_m \leq n} (A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_m}).$$

Show that

$$\mathbb{P}(B_m) = S_m - \binom{m}{1} S_{m+1} + \binom{m+1}{2} S_{m+2} - \binom{m+2}{3} S_{m+3} + \ldots$$

where $S_k = \sum_{1 \leq i_1 < i_2 < \ldots < i_k \leq n} \mathbb{P}(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k})$.

PROBLEM 6. Recall the problem of matching two shuffled decks of cards, but with n cards in each deck, so that $\Omega_n = S_n \times S_n$ and $p_{(\pi,\sigma)} = \frac{1}{(n!)^2}$ for each $(\pi, \sigma) \in \Omega$. Let A_m be the event that there are exactly m matches between the two decks[2].

(1) For fixed $m \geq 0$, show that $\mathbb{P}(A_m) \to e^{-1} \frac{1}{m!}$ as $n \to \infty$.

(2) Assume that the approximations above are valid for $n = 52$ and $m \leq 10$. Find the probability that there are at least 10 matches.

[**Remark:** You may use the result of the previous problem to solve this one].

PROBLEM 7. Suppose n couples at a circular table. Show that the chance that no pair sits next to each other is

$$\frac{1}{n!} \sum_{k=0}^{n} (-1)^k \frac{2n}{2n-k} \binom{2n-k}{k} (n-k)!$$

PROBLEM 8. Place r_n distinguishable balls in n distinguishable urns. Let A_n be the event that at least one urn is empty[3].

(1) If $r_n = n^2$, show that $\mathbb{P}(A_n) \to 0$ as $n \to \infty$.

(2) If $r_n = Cn$ for some fixed constant C, show that $\mathbb{P}(A_n) \to 1$ as $n \to \infty$.

(3) Can you find an increasing function $f(\cdot)$ such that if $r_n = f(n)$, then $\mathbb{P}(A_n)$ does not converge to 0 or 1? [**Hint:** First try $r_n = n^\alpha$ for some α, not necessarily an integer].

PROBLEM 9. A box contains N coupons labelled $1, 2, \ldots, N$. Draw m_N coupons at random, with replacement, from the box. Let A_N be the event that every coupon from the box has appeared at least once in the sample.

(1) If $m_N = N^2$, show that $\mathbb{P}(A_N) \to 1$ as $N \to \infty$.

[2]Strictly speaking, we should write $A_{n,m}$, since the $A_{n,m} \subseteq \Omega_n$ but for ease of notation we omit the subscript n. Similarly, it would be appropriate to write p_n and $\mathbb{P}_n$ for the probabilities, but again, we simplify the notation when there is no risk of confusion.

[3]Similar to the previous comment, here it would be appropriate to write $\mathbb{P}_n(A_n)$ as the probability spaces are changing, but we keep the notation simple and simply write $\mathbb{P}(A_n)$.

(2) If $m_N = CN$ for some fixed constant C, show that $\mathbb{P}(A_N) \to 0$ as $N \to \infty$

(3) Can you find an increasing function $f(\cdot)$ such that if $m_N = f(N)$, then $\mathbb{P}(A_N)$ does not converge to 0 or 1? [**Hint:** See if you can relate this problem to the previous one].

PROBLEM 10. Let $A_1, \ldots, A_n$ be events in a common probability space. Let B be the event that *at least two* of the A_is occur. Prove that

$$\mathbb{P}(B) = S_2 - 2S_3 + 3S_4 - \ldots + (-1)^m(m-1)S_m$$

where $S_k = \sum_{1 \le i_1 < i_2 < \ldots < i_k \le n} \mathbb{P}\{A_{i_1} \cap \ldots \cap A_{i_k}\}$ for $1 \le k \le n$.**Not for submission:** More generally, you may show that the probability that at least ℓ of the A_is occur is equal to

$$S_\ell - \binom{\ell}{1} S_{\ell+1} + \binom{\ell+1}{2} S_{\ell+2} - \binom{\ell+2}{3} S_{\ell+3} + \ldots \tag{5}$$

PROBLEM 11. Continuing with the notation of the previous problem, assume the formula given in (5) holds. If C is the event that *exactly* ℓ of the events A_is occur, then show that

$$\mathbb{P}(C) = S_\ell - \binom{\ell+1}{\ell} S_{\ell+1} + \binom{\ell+2}{\ell} S_{\ell+2} - \ldots + (-1)^{n-\ell} S_n. \tag{6}$$

[**Hint:** If you want to prove this directly, without using (5), that is also okay.]

PROBLEM 12. Let $\vee A = \min\{a : a \in A\}$ and $\wedge A = \max\{a : a \in A\}$ for any finite subset A of $\mathbb{R}$, with the convention that $\wedge\emptyset = 0$. Show that for any finite non-empty set A,

$$\vee A = \sum_{B \subseteq A} (-1)^{|B|-1} \wedge B.$$

PROBLEM 13. If r distinguishable balls are placed at random into m labelled bins, write an expression for the probability that each bin contains at least two balls.

PROBLEM 14. A deck of cards is dealt to four players (13 cards each). Find the probability that at least one of the players has two or more *aces*.

PROBLEM 15. Let p be the probability that in a gathering of 2500 people, there is some day of the year that is not the birthday of anyone in the gathering. Make reasonable assumptions and argue that $0.3 \le p \le 0.4$.

PROBLEM 16. Two numbers are picked at random from $[N]$ where N is very large. What is the chance that the two numbers have no common factor (other than 1). The answer depends on N, but we are asking for the answer as $N \to \infty$.

PROBLEM 17. Consider the problem of a psychic guessing the order of a deck of shuffled cards. Assume complete randomness of the guesses. Use the formula in (6) to derive an expression for the probability that the number of guesses is exactly ℓ, for $0 \leq \ell \leq 52$. Use meaningful approximation to these probabilities and give numerical values (to 3 decimal places) of the probabilities for $\ell = 0, 1, 2, \ldots, 6$.

PROBLEM 18. Two tennis players serve alternately. The game stops when one of them loses on her own serve (a simplified rule! If you don't like this, you can then repeat the problem with the rule that the game stops when one of them is 2 points ahead for the first time). Let the probability of the first player (respectively second) losing on her serve be p_1 (respectively p_2). Find the probability that the first player wins. What is the number if both are expert players with $p_1 = 0.01$ and $p_2 = 0.02$ or vice versa? When p_i are small, does it matter significantly who starts first?

PROBLEM 19. Place r_m distinguishable balls in m distinguishable bins. Let A_m be the event that at least one bin is empty[4].

(1) If $r_n = m^2$, show that $\mathbb{P}(A_m) \to 0$ as $m \to \infty$.

(2) If $r_m = Cm$ for some fixed constant C, show that $\mathbb{P}(A_m) \to 1$ as $n \to \infty$.

(3) Can you find an increasing function $f(\cdot)$ such that if $r_m = f(m)$, then $\mathbb{P}(A_m)$ does not converge to 0 or 1? [**Hint:** First try $r_m = m^\alpha$ for some α, not necessarily an integer].

PROBLEM 20. Based on probability models and some experiments, a scientist finds that the chance for two given people to have the same fingerprint pattern (in reality only some features of the fingerprint pattern are compared, and that is what we mean) is 10^{-10}. Is this a good enough assurance that we can use this system to distinguish any two people in Karnataka (which has a population of 7 crore)?

[4]Here it would be appropriate to write $\mathbb{P}_m(A_n)$ as the probability spaces are changing, but we keep the notation simple and simply write $\mathbb{P}(A_n)$.

CHAPTER 3

Independence and conditioning of events

1. Independence - a first look

We remarked in the context of inclusion-exclusion formulas that often the probabilities of intersections of events is easy to find, and then we can use them to find probabilities of unions etc. In many contexts, this is related to one of the most important notions in probability.

DEFINITION 21. Let A, B be events in a common probability space. We say that A and B are *independent* is $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

EXAMPLE 22. Toss a fair coin n times. Then $\Omega = \{\underline{\omega} : \underline{\omega} = (\omega_1, \ldots, \omega_n),\ \omega_i \text{ is } 0 \text{ or } 1\}$ and $p_{\underline{\omega}} = 2^{-n}$ for each $\underline{\omega}$. Let $A = \{\underline{\omega} : \omega_1 = 0\}$ and let $B = \{\underline{\omega} : \omega_2 = 0\}$. Then, from the definition of probabilities, we can see that $\mathbb{P}(A) = 1/2$, $\mathbb{P}(B) = 1/2$ (because the elementary probabilities are equal, and both the sets A and B contain exactly 2^{n-1} elements). Further, $A \cap B = \{\underline{\omega} : \omega_1 = 1, \omega_2 = 0\}$ has 2^{n-2} elements, whence $\mathbb{P}(A \cap B) = 1/4$. Thus, $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ and hence A and B are independent.

If two events are independent, then the probability of their intersection can be found from the individual probabilities. How do we check if two events are independent? By checking if the probability of the event is equal to the product of the individual probabilities! It seems totally circular and useless! There are many reasons why it is not an empty notion as we shall see.

Firstly, in physical situationsdependence is related to a basic intuition we have about whether two events are related or not. For example, suppose you are thinking of betting Rs.1000 on a particular horse in a race. If you get the news that your cousin is getting married, it will perhaps not affect the amount you plan to bet. However, if you get the news that one of the other horses has been injected with undetectable drugs, it might affect the bet you want to place. In other words, certain events (like marriage of a cousin) have no bearing on the probability of the event of interest (the event that our horse wins) while other events (like the injection of drugs) do have an impact. This intuition is often put into the very definition of probability space that we have.

For example, in the above example of tossing a fair coin n times, it is our intuition that a coin does not remember how it fell previous times, and that chance of its falling head in any toss is just 1/2, irrespective of how many heads or tails occured before[1] And this intuition was

[1]It may be better to attribute this to experience rather than intuition. There have been reasonable people in history who believed that if a coin shows heads in ten tosses in a row, then on the next toss it is more likely to show tails (to 'compensate' for the overabundance of heads)! Clearly this is also someone's intuition, and different

used in defining the elementary probabilities as 2^{-n} each. Since we started with the intuitive notion of independence, and put that into the definition of the probability space, it is quite expected that the event that the first toss is a head should be independent of the event that the second toss is a tail. That is the calculation shown in above.

But how is independence useful mathematically if the conditions to check independence are the very conclusions we want?! The answer to this lies in the following fact (to be explained later). When certain events are independent, then many other collections of events that can be made out of them also turn out to be independent. For example, if A, B, C, D are independent (we have not yet defined what this means!), then $A \cup B$ and $C \cup D$ are also independent. Thus, starting from independence of certain events, we get independence of many other events. For example, any event depending on the first four tosses is independent of eny event depending on the next five tosses.

2. Conditional probability and independence

Deeinition 23. Let A, B be two events in the same probability space.

(1) If $\mathbb{P}(B) \neq 0$, we define the *conditional probability of A given B* as

$$\mathbb{P}\left(A \,\middle|\, B\right) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

(2) We say that A and B are *independent* if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. If $\mathbb{P}(B) \neq 0$, then A and B are independent if and only if $\mathbb{P}(A \,|\, B) = \mathbb{P}(A)$ (and similarly with the roles of A and B reversed). If $\mathbb{P}(B) = 0$, then A and B are necessarily independent since $\mathbb{P}(A \cap B)$ must also be 0.

What do these notions mean intuitively? In real life, we keep updating probabilities based on information that we get. For example, when playing cards, the chance that a randomly chosen card is an ace is 1/13, but having drawn a card, the probability for the next card may not be the same - if the first card was seen to be an ace, then the chance of the second being an ace falls to 3/51. This updated probability is called a conditional probability. Independence of two events A and B means that knowing whether or not A occured does not change the chance of occurrence of B. In other words, the conditional probability of A given B is the same as the unconditional (original) probability of A.

Example 24. Let $\Omega = \{(i, j) : 1 \leq i, j \leq 6\}$ with $p_{(i,j)} = \frac{1}{36}$. This is the probability space corresponding to a throw of two fair dice. Let $A = \{(i, j) : i \text{ is odd}\}$ and $B = \{(i, j) : j \text{ is 1 or 6}\}$ and $C = \{(i, j) : i + j = 4\}$. Then $A \cap B = \{(i, j) : i = 1, 3, \text{ or } 5, \text{ and } j = 1 \text{ or } 6\}$. Then, it is easy to see that

$$\mathbb{P}(A \cap B) = \frac{6}{36} = \frac{1}{6}, \;\; \mathbb{P}(A) = \frac{18}{36} = \frac{1}{2}, \;\; \mathbb{P}(B) = \frac{12}{36} = \frac{1}{3}.$$

---

from ours. Only experiment can decide which is correct, and any number of experiments with real coins show that our intuition is correct, and coins have no memory.

In this case, $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ and hence A and B are independent. On the other hand,

$$\mathbb{P}(A \cap C) = \mathbb{P}\{(1,3),(2,2)\} = \frac{1}{18}, \quad \mathbb{P}(C) = \mathbb{P}\{(1,3),(2,2),(3,1)\} = \frac{1}{12}.$$

Thus, $\mathbb{P}(A \cap C) \neq \mathbb{P}(A)\mathbb{P}(C)$ and hence A and C are not independent.

This agrees with the intuitive understanding of independence, since A is an event that depends only on the first toss and B is an event that depends only on the second toss. Therefore, A and B ought to be independent. However, C depends on both tosses, and hence cannot be expected to be independent of A. Indeed, it is easy to see that $\mathbb{P}(C \mid A) = \frac{1}{9}$.

EXAMPLE 25. Let $\Omega = S_{52}$ with $p_\pi = \frac{1}{52!}$. Define the events

$$A = \{\pi : \pi_1 \in \{10,20,30,40\}\}, \qquad A = \{\pi : \pi_2 \in \{10,20,30,40\}\}.$$

Then both $\mathbb{P}(A) = \mathbb{P}(B) = \frac{1}{13}$. However, $\mathbb{P}(B \mid A) = \frac{3}{51}$. One can also see that $\mathbb{P}(B \mid A^c) = \frac{4}{51}$.

In words, A (respectively B) could be the event that the first (respectively second) card is an ace. Then $\mathbb{P}(B) = 4/52$ to start with. When we see the first card, we update the probability. If the first card was not an ace, we update it to $\mathbb{P}(B \mid A^c)$ and if the first card was an ace, we update it to $\mathbb{P}(B \mid A)$.

Caution: Independence should not be confused with disjointness! In fact, they are poles apart. If A and B are disjoint, $\mathbb{P}(A \cap B) = 0$ and hence A and B can be independent if and only if one of $\mathbb{P}(A)$ or $\mathbb{P}(B)$ equals 0. Intuitively, if A and B are disjoint, then knowing that A occurred gives us a lot of information about B (that it did not occur!), so independence is not to be expected.

EXERCISE 26. If A and B are independent, show that the following pairs of events are also independent.

(1) A and B^c.

(2) A^c and B.

(3) A^c and B^c.

Total probability rule and Bayes' rule: Let $A_1, \ldots, A_n$ be pairwise disjoint and mutually exhaustive events in a probability space. Assume $\mathbb{P}(A_i) > 0$ for all i. This means that $A_i \cap A_j = \emptyset$ for any $i \neq j$ and $A_1 \cup A_2 \cup \ldots \cup A_n = \Omega$. We also refer to such a collection of events as a partition of the sample space.

PROPOSITION 27. *Let $A_1, \ldots, A_n$ partition the sample space as above. Let B be any event. Then,*

(1) (Total probability rule). $\mathbb{P}(B) = \mathbb{P}(A_1)\mathbb{P}(B \mid A_1) + \ldots + \mathbb{P}(A_n)\mathbb{P}(B \mid A_n)$.

(2) (Bayes' rule). Assume that $\mathbb{P}(B) > 0$. Then, for each $k = 1, 2, \ldots, n$, we have

$$\mathbb{P}(A_k \mid B) = \frac{\mathbb{P}(A_k)\mathbb{P}(B \mid A_k)}{\mathbb{P}(A_1)\mathbb{P}(B \mid A_1) + \ldots + \mathbb{P}(A_n)\mathbb{P}(B \mid A_n)}.$$

Proof. The proof is merely by following the definition.

(1) The right hand side is equal to

$$\mathbb{P}(A_1)\frac{\mathbb{P}(B \cap A_1)}{\mathbb{P}(A_1)} + \ldots + \mathbb{P}(A_n)\frac{\mathbb{P}(B \cap A_n)}{\mathbb{P}(A_n)} = \mathbb{P}(B \cap A_1) + \ldots + \mathbb{P}(B \cap A_n)$$

which is equal to $\mathbb{P}(B)$ since A_i are pairwise disjoint and exhaustive.

(2) Without loss of generality take $k = 1$. Note that $\mathbb{P}(A_1 \cap B) = \mathbb{P}(A_1)\mathbb{P}(B \mid A_1)$. Hence

$$\begin{aligned}\mathbb{P}(A_1 \mid B) &= \frac{\mathbb{P}(A_1 \cap B)}{\mathbb{P}(B)} \\ &= \frac{\mathbb{P}(A_1)\mathbb{P}(B \mid A_1)}{\mathbb{P}(A_1)\mathbb{P}(B \mid A_1) + \ldots + \mathbb{P}(A_n)\mathbb{P}(B \mid A_n)}\end{aligned}$$

where we used the total probability rule to get the denominator. ■

Exercise 28. Suppose A_i are events such that $\mathbb{P}(A_1 \cap \ldots \cap A_n) > 0$. Then show that

$$\mathbb{P}(A_1 \cap \ldots \cap A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2 \mid A_1)\mathbb{P}(A_3 \mid A_1 \cap A_2)\ldots\mathbb{P}(A_n \mid A_1 \cap \ldots \cap A_{n-1}).$$

Example 29. Consider a rare disease X that affects one in a million people. A medical test is used to test for the presence of the disease. The test is 99% accurate in the sense that if a person has no disease, the chance that the test shows positive is 1% and if the person has disease, the chance that the test shows negative is also 1%.

Suppose a person is tested for the disease and the test result is positive. What is the chance that the person has the disease X?

Let A be the event that the person has the disease X. Let B be the event that the test shows positive. The given data may be summarized as follows.

(1) $\mathbb{P}(A) = 10^{-6}$. Of course $\mathbb{P}(A^c) = 1 - 10^{-6}$.

(2) $\mathbb{P}(B \mid A) = 0.99$ and $\mathbb{P}(B \mid A^c) = 0.01$.

What we want to find is $\mathbb{P}(A \mid B)$. By Bayes' rule (the relevant partition is $A_1 = A$ and $A_2 = A^c$),

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\mathbb{P}(A)}{\mathbb{P}(B \mid A)\mathbb{P}(A) + \mathbb{P}(B \mid A^c)\mathbb{P}(A^c)} = \frac{0.99 \times 10^{-6}}{0.99 \times 10^{-6} + 0.01 \times (1 - 10^{-6})} = 0.000099.$$

The test is quite an accurate one, but the person tested positive has a really low chance of actually having the disease! Of course, one should observe that the chance of having disease is now approximately 10^{-4} which is considerably higher than 10^{-6}.

A calculation-free understanding of this surprising looking phenomenon can be achieved as follows: Let everyone in the population undergo the test. If there are 10^9 people in the

population, then there are only 10^3 people with the disease. The number of true positives is approximately $10^3 \times 0.99 \approx 10^3$ while the number of false positives is $(10^9 - 10^3) \times 0.01 \approx 10^7$. In other words, among all positives, the false positives are way more numerous than true positives.

The surprise here comes from not taking into account the relative sizes of the sub-populations with and without the disease. Here is another manifestation of exactly the same fallacious reasoning.

Question: Person X is introverted, very systematic in thinking, pedantic in talking and quite absent-minded. You are told that X is a doctor or a mathematician. What do you guess - doctor or mathematician?

As we saw in class, most people answer "mathematician". Even accepting the stereotype that a mathematician is more likely to have all these qualities than a doctor, this answer ignores the fact that there are perhaps a hundred times more doctors in the world than mathematicians! In fact, the situation is identical to the one in the example above, and the mistake is in confusing $\mathbb{P}(A\big|B)$ and $\mathbb{P}(B\big|A)$.

3. Independence of three or more events

Definition 30. Events $A_1, \ldots, A_n$ in a common probability space are said to be independent if $\mathbb{P}\left(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_m}\right) = \mathbb{P}(A_{i_1})\mathbb{P}(A_{i_2}) \ldots \mathbb{P}(A_{i_m})$ for every choice of $m \leq n$ and every choice of $1 \leq i_1 < i_2 < \ldots < i_m \leq n$.

The independence of n events requires us to check 2^n equations (that many choices of $i_1, i_2, \ldots$). Should it not suffice to check that each pair of A_i and A_j are independent? The following example shows that this is not the case!

Example 31. Let $\Omega = \{0,1\}^n$ with $p_{\underline{\omega}} = 2^{-n}$ for each $\underline{\omega} \in \Omega$. Define the events $A = \{\underline{\omega} : \omega_1 = 0\}$, $A = \{\underline{\omega} : \omega_2 = 0\}$ and $C = \{\underline{\omega} : \omega_1 + \omega_2 = 0 \text{ or } 2\}$. In words, we toss a fair coin n times and A denotes the event that the first toss is a tail, B denotes the event that the second toss is a tail and C denotes the event that the first two tosses are both heads or both tails. Then $\mathbb{P}(A) = \mathbb{P}(B) = \mathbb{P}(C) = \frac{1}{2}$. Further,

$$\mathbb{P}(A \cap B) = \frac{1}{4},\ \mathbb{P}(B \cap C) = \frac{1}{4},\ P(A \cap C) = \frac{1}{4},\ \mathbb{P}(A \cap B \cap C) = \frac{1}{4}.$$

Thus, A, B, C are independent *pairwise*, but not independent by our definition because $\mathbb{P}(A \cap B \cap C) \neq \frac{1}{8} = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)$.

Intuitively this is right. Knowing A does not given any information about C (similarly with A and B or B and C), but knowing A and B tells us completely whether or not C occurred! Thus it is right that the definition should not declare them to be independent.

Exercise 32. Let $A_1, \ldots, A_n$ be events in a common probability space. Then, $A_1, A_2, \ldots, A_n$ are independent if and only if the following equalities hold: For each i, define B_i as A_i and

A_i^c. Then

$$\mathbb{P}(B_1 \cap B_2 \cap \ldots \cap B_n) = \mathbb{P}(B_1)\mathbb{P}(B_2)\ldots\mathbb{P}(B_n).$$

Note: This should hold for any possible choice of B_is. In other words, the system of 2^n equalities in the definition of independence may be replaced by this new set of 2^n equalities. The latter system has the advantage that it immediately tells us that if $A_1, \ldots, A_n$ are independent, then $A_1, A_2^c, A_3, \ldots$ (for each i choose A_i or its complement) are independent.

4. Subtleties of conditional probability

Conditional probabilities are quite subtle. Apart from the common mistake of confusing $\mathbb{P}(A \mid B)$ for $\mathbb{P}(B \mid A)$, there are other points one sometimes overlooks. In fact, most of the paradoxical sounding puzzles in probability are based on confusing aspects of probability. Let us see one.

Question 33. A man says "I have two children, and one of them is a boy". What is the chance that the other one is a girl?

There are four possibilities BB, BG, GB, GG, of which GG has been eliminated. Of the remaining three, two are favourable, hence the chance is 2/3 that the other child is a girl. This is a possible solution. If you accept this as reasonable, here is another question.

Question 34. A man says "I have two children, and one of them is a boy born on a Sunday". What is the chance that the other one is a girl?

Does the addition of the information about the boy change the probability? One opinion is that it should not. The other is to follow the same solution pattern as before. Write down all the $2 \times 2 \times 7 \times 7$ possibilities: *BBss* (boy, boy, sunday, sunday), *BBsm*, etc. The given information that one is a boy who was born on Sunday eliminates many possibilities and what remain are 27 possibilities $BGt*$, $GB*t$, $BBt*$, $BB*t$ where $*$ is any day of the week. Take care to not double count *BBtt* to see that there are 27 possibilities. Of these, 14 are favourable (i.e., the other child is a girl), hence we conclude that the probability is 14/27.

Is the correct answer 14/27 as calculated here or is it 2/3 (since the information of the day of birth of the boy is irrelevant, why should we change our earlier answer of 2/3?)?

We leave it as food for thought. If you want a hint, the point is that to compute conditional probabilities, *it is not enough to know what the person said, but also what else he could have said*. Not realizing this point is the main source of confusion in many popular puzzles in probability.

5. Significance of conditional probability in life and the universe

There are two points we make in this section. One is that conditional probability is often the way we think of probabilistic situations. The second is on practical and philosophical significance of Bayes' rule.

5.1. Defining a probability space via conditional probabilities. So far all our probability spaces were defined by giving a sample space and specifying elementary probabilities. Even when we started with a word-description of a situation, translating it into a probability space was straightforward. But in some situations, conditional probability is in-built into the very description of the situation. Rather than talk in generalities, this is best explained by an example. This example is of considerable importance, and lends itself to many generalizations.

EXAMPLE 35 (Pólya's urn scheme). An urn has 1 blue and 1 red ball. A ball is drawn uniformly at random, its colour noted, and returned to the urn *along with an additional ball of the same colour*. Repeat the process, each time drawing a ball uniformly at random from the urn, and returning it with another ball of the same colour. To keep it specific, let us say we stop after the third ball is drawn.

What is the probability space? The outcome of the experiment is just the sequence of colours of the balls drawn. Hence the sample space is

$$\Omega = \{R, B\}^3 = \{(\omega_1, \omega_2, \omega_3) : \omega_i \in \{R, B\}\}.$$

What are the elementary probabilities? To take one example, $p(R, B, B)$ ought to be $\frac{1}{2} \times \frac{1}{3} \times \frac{2}{4}$. The reason is that there is a chance of $\frac{1}{2}$ of the first ball being red, then there are two red and one blue in the urn, so the chance of the second ball being blue is $\frac{1}{3}$ after which there are two red and two blue, so the chance of drawing a blue is $\frac{2}{4}$. In other words, we are computing the elementary probabilities *using* the rule in Exercise 28, writing $p(R, B, B) = p(R)p(B|R)p(B|R, B)$ (hopefully the notation is self-explanatory; by $p(B|R, B)$ we mean the chance of drawing a blue, given that the first two draws were red and blue, respectively). By similar reasoning, we can write the full list of elementary probabilities as

$$p(R,R,R) = p(B,B,B) = \frac{1}{2} \times \frac{2}{3} \times \frac{3}{4} \qquad p(R,R,B) = p(B,B,R) = \frac{1}{2} \times \frac{2}{3} \times \frac{1}{4}$$

$$p(R,B,R) = p(B,R,B) = \frac{1}{2} \times \frac{1}{3} \times \frac{2}{4} \qquad p(R,B,B) = p(B,R,R) = \frac{1}{2} \times \frac{1}{3} \times \frac{2}{4}$$

The problems lends itself to many generalizations, such as changing the initial composition of the urn, having more than two colours, different replacement rules (can put some balls of the same colour and/or some balls of the opposite colour) etc.

EXERCISE 36. If the initial urn has b blue and r red balls, and the experiment is carried out till n balls are drawn, show that $\Omega = \{0, 1\}^n$ (where $0 = R$ and $1 = B$) and

$$p(\omega_1, \ldots, \omega_n) = \frac{b(b+1)\ldots(b+k-1)r(r+1)\ldots(r+\ell-1)}{(b+r)(b+r+1)\ldots(b+r+n-1)}$$

where $k = \omega_1 + \ldots + \omega_n$ and $\ell = n - k$. The surprise is that the probability depends only on the number of blue and number of red drawn, not the order!

5.2. Some implications of the Bayes' rule. We have already seen some applications of Bayes' rule. In general, it is an important ingredient in sifting evidence. We explain it with two examples.

EXAMPLE 37. In a city there are two ethnic groups of people, a minority group A (5% of the population) and a majority group B (95%). One day, in a dark alley, an incident of theft occurs. When giving evidence to the police, the victim says that the perpetrator looked like he was from group A.

Since there are possibilities of making mistakes, the police check the ability of the victim to identify which group a person belongs to, by testing him against a large number of people. It turns out that he is right 90% of the time. Can they conclude that they might as well restrict their search to members of group A?

Assuming the same crime rate by members of either group, a standard application of Bayes' rule tell us that $\mathbb{P}(\text{criminal } \in A \mid \text{victim says } A)$ is equal to

$$\frac{\mathbb{P}(\text{victim says } A \mid \text{criminal } \in A)\mathbb{P}(\text{criminal } \in A)}{\mathbb{P}(\text{victim says } A \mid \text{criminal } \in A)\mathbb{P}(\text{criminal } \in A) + \mathbb{P}(\text{victim says A} \mid \text{criminal } \in B)\mathbb{P}(\text{criminal } \in B)}$$

$$= \frac{0.9 \times 0.05}{0.9 \times 0.05 + 0.1 \times 0.95} = 0.32\ldots$$

This is much smaller than 0.9, which is the answer that many are tempted to give right-away. Thus we see that not taking the base rate (i.e., that group A is a minority) into account gives a false impression.

Clearly the above situation is of relevance in society, when we pass judgements on groups of people. The next example is not about avoidance of mistakes, but places Bayes' rule as one of the fundamental modes of sifting evidence to arrive at conclusions about the world.

EXAMPLE 38. A standard application of Bayes' rule is to the following artificial looking problem: Urn-A has two red and one blue ball and Urn-B has two blue and one red ball. A fair die is thrown and if it turns up 6, then Urn-B is chosen and otherwise Urn-A is chosen. Balls are drawn uniformly at random, with replacement, from the chosen urn. You are not told the result of the die throw or which urn is chosen, but only that the first four balls drawn are B, R, B, B. What is the chance that the first urn was chosen?

We leave the solution of this to you, but talk about its meaning. Instead of the two urns, imagine two (or more) hypothesis. They could be two theories about some aspect of nature like Lamark's theory v/s Darwin's or that the space is flat as opposed to the space is curved etc. They could also be not-exactly-scientific hypotheses like "God exists v/s God does not exist", "Dosas in Vidyarthi bhavan are better than those in Janatha hotel v/s the reverse statement" etc. How do we come to any conclusion about such matters?

We imagine two competing hypotheses to be two different urns, and each observation of the world as a ball drawn from the urn that corresponds to the true hypothesis. With each draw, we update our conditional probabilities of the two hypotheses. Perhaps when it is overwhelmingly large (like 0.95) or small (like 0.05), we stop experimenting and decide

to accept the hypothesis that has higher conditional probability. Although in many real-life situations, there is no hope of having actual numbers to calculate the conditional probabilities, Bayes' rule may be taken as a caricature of the way we reason about the world.

5.3. When unlikely events happen... Imagine two disjoint events A and B in a probability space having probabilities 2^{-20} and 2^{-40} respectively. Then the event $A \cup B$ is also highly unlikely, with a probability of $2^{-20} + 2^{-40}$. If we know that $A \cup B$ did indeed occur, what is the chance that it happened through A rather than B? One can apply Bayes' rule, although direct calculation is easier in this case, to see that

$$\mathbb{P}(A \bigm| A \cup B) = \frac{1}{1 + 2^{-20}}, \qquad \mathbb{P}(B \bigm| A \cup B) = \frac{2^{-20}}{1 + 2^{-20}}.$$

In other words, given $A \cup B$, it is overwhelmingly likely that it happened through A, rather than through B. It is summarized by saying that *when an unlikely event happens, it happens in the least unlikely way.*

6. Exercises

PROBLEM 1. Let A, B be events with positive probability in a common probability space. We have seen in class that $\mathbb{P}(A|B)$ and $\mathbb{P}(B|A)$ are not to be confused.

(1) Show that $\mathbb{P}(A|B) = \mathbb{P}(B|A)$ if and only if $\mathbb{P}(A) = \mathbb{P}(B)$.

(2) Show that $\mathbb{P}(A|B) > \mathbb{P}(A)$ if and only if $\mathbb{P}(B|A) > \mathbb{P}(B)$. That is, if occurrence of B makes A more likely than it was before, then the occurrence of A makes B more likely than it was.

PROBLEM 2. There are 10 bins and the kth bin contains k black and $11 - k$ white balls. A bin is chosen uniformly at random. Then a ball is chosen uniformly at random from the chosen bin.

(1) Find the conditional probability that the chosen ball is black, given that the kth bin was chosen. Use this to compute the (unconditional) probability that the chosen ball is white.

(2) Given that the chosen ball is black, what is the probability that the kth bin was chosen?

PROBLEM 3. A fair die is thrown n times. For $1 \leq k \leq n-1$, let A_k be the event that the kth throw and the $(k+1)$st throw yield the same result. Are $A_1, \ldots, A_{n-1}$ independent? Are they pairwise independent?

PROBLEM 4. Two dice (not necessarily identical, and not necessarily fair) are thrown and let X be the total of the two numbers that turn up. Can you design the two dice so that X is equally likely to be any of the numbers $2, 3, \ldots, 12$?

PROBLEM 5. A fair die is thrown repeatedly (assume that the results are independent). Let $N_1 < N_2 < \ldots$ be the throw-numbers that show a six. Are N_1 and $N_2 - N_1$ independent? Are N_1 and N_2 independent?

PROBLEM 6. Find examples of discrete probability spaces and events A, B, C so that the following happen.

(1) The events A, B, C are pairwise independent but not mutually independent.

(2) $\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)$ but A, B, C are not independent.

PROBLEM 7. Let $A_1, \ldots, A_n$ be independent with $\mathbb{P}(A_i) = p$ for all i. Find the probability that (a) none of the events $A_1, \ldots, A_n$ occur, (b) all of the events $A_1, \ldots, A_n$ occur.

PROBLEM 8. A factory produces light bulbs and it is desired that the proportion of defective bulbs sold should be less than 2%. To ensure this, an inspector random chooses 100 bulbs from a batch and tests all of them. The whole batch is sent for sale if and only if there are at most 2 defective bulbs among the 100 tested. What is the chance that a batch with more than 2% defective bulbs is sent for sale?

PROBLEM 9. Same as previous problem, except that the inspector keeps testing bulbs one after another (assume with replacement) till he/she finds a defective one. The batch is passed (sent for sale) if the first 5 bulbs tested are not defective. What is the chance that a batch with more than 2% defective bulbs is sent for sale?

CHAPTER 4

Discrete random variables and their distributions

Let (Ω, p) be a probability space and $X : \Omega \to \mathbb{R}$ be a random variable. A random variable can be thought of as a measurement in a random experiment. For example, in playing many games, two dice are rolled, and the sum of the two numbers is taken. The sample space is $\{(i,j) : i,j \leq 6\}$, but we are only interested in the measurement $X((i,j)) = i + j$. If an event $A \subseteq \Omega$ is identified with its indicator $\mathbf{1}_A$, then we see that they correspond to special measurements of a "yes v/s no" kind. Random variables are useful for finer measurements.

A quick remark on terminology. One can allow a random variable is allowed to take values in arbitrary sets. But for definiteness we make them real-valued. By just renaming a set, we can always do this (for example, if $X : \Omega \to \{a, b, c\}$, then we can rename a, b, c as $1, 2, 3$ respectively, so X becomes real-valued). When it is more convenient to consider other co-domains, we will use words like *random vector* (if $X : \Omega \to \mathbb{R}^n$ for some n) or a *categorical variable* (if $X : \Omega \to S$, where S is some finite set with no structure, e.g., $\{\text{Tall}, \text{Short}\}$ or $\{\text{Wrinkled}, \text{Smooth}\}$), etc.

1. Probability distribution of a discrete random variable

The thing to know about a random variable is the set of values it can take and the probabilities with which it takes those values. This information is captured by the *probability mass function* of the random variable.

Probability mass function (pmf). The range of X is a countable subset of $\mathbb{R}$, denote it by $\text{Range}(X) = \{t_1, t_2, \ldots\}$. Then, define $f_X : \mathbb{R} \to [0,1]$ as the function

$$f_X(t) = \begin{cases} \mathbb{P}\{X = t\} & \text{if } t \in \text{Range}(X). \\ 0 & \text{if } t \notin \text{Range}(X). \end{cases}$$

Here $\{X = t\}$ is a short form for the event $\{\omega \in \Omega : X(\omega) = t\}$. We shall use such shortcuts throughout the course. One obvious property of the pmf is that $\sum_{t \in \mathbb{R}} f_X(t) = 1$. Conversely, any non-negative function $f : \mathbb{R} \to \mathbb{R}_+$ that is zero outside a countable set S and such that $\sum_{t \in \mathbb{R}} f(t) = 1$ is a pmf of some random variable.

We now introduce another gadget that has the same information as the probability mass function, in that either can be recovered from the other. Its real need will only become clear when we go beyond discrete random variables later.

Cumulative distribution function (CDF). Define $F_X : \mathbb{R} \to [0,1]$ by

$$F_X(t) = \mathbb{P}\{\omega : X(\omega) \leq t\}.$$

It is easy to see that one can recover the pmf from the CDF and vice versa. For example, given the pmf f, we can write the CDF as $F(t) = \sum_{u:u\le t} f(u)$. Conversely, given the CDF, by looking at the locations of the jumps and the sizes of the jumps, we can recover the pmf.

EXAMPLE 10. Let $\Omega = \{(i,j) : 1 \le i,j \le 6\}$ with $p_{(i,j)} = \frac{1}{36}$ for all $(i,j) \in \Omega$. Let $X : \Omega \to \mathbb{R}$ be the random variable defined by $X(i,j) = i + j$. Then, Range$(X) = \{2, 3, \ldots, 12\}$. The pmf of X is given by

k	2	3	4	5	6	7	8	9	10	11	12
$f_X(k)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

and the CDF is given by

t	< 2	$[2,3)$	$[3,4)$	$[4,5)$	$[5,6)$	$[6,7)$	$[7,8)$	$[8,9)$	$[9,10)$	$[10,11)$	$[11,12)$	≥ 12
$F_X(k)$	0	1/36	3/36	6/36	10/36	15/36	21/36	26/36	30/36	33/36	35/36	1

Basic properties of a CDF: The following observations are easy to make.

(1) F is an increasing function on $\mathbb{R}$.

(2) $\lim_{t\to+\infty} F(t) = 1$ and $\lim_{t\to-\infty} F(t) = 0$.

(3) F is right continuous, that is, $\lim_{h\searrow 0} F(t+h) = F(t)$ for all $t \in \mathbb{R}$.

(4) F increases only in jumps. This means that if F has no jump discontinuities (an increasing function has no other kind of discontinuity anyway) in an interval $[a,b]$, then $F(a) = F(b)$.

Since $F(t)$ is the probability of a certain event, these statements can be proved using the basic rules of probability that we saw earlier.

PROOF. Let $t < s$. Define two events, $A = \{\omega : X(\omega) \le t\}$ and $B = \{\omega : X(\omega) \le s\}$. Clearly $A \subseteq B$ and hence $F(t) = \mathbb{P}(A) \le \mathbb{P}(B) = F(s)$. This proves the first property.

To prove the second property, let $A_n = \{\omega : X(\omega) \le n\}$ for $n \ge 1$. Then, A_n are increasing in n and $\bigcup_{n=1}^{\infty} A_n = \Omega$. Hence, $F(n) = \mathbb{P}(A_n) \to \mathbb{P}(\Omega) = 1$ as $n \to \infty$. Since F is increasing, it follows that $\lim_{t\to+\infty} F(t) = 1$. Similarly one can prove that $\lim_{t\to-\infty} F(t) = 0$.

Right continuity of F is also proved the same way, by considering the events $B_n = \{\omega : X(\omega) \le t + \frac{1}{n}\}$. We omit details. ■

The point is that probabilistic questions about X can be answered by knowing its CDF F_X. Therefore, in a sense, the probability space becomes irrelevant. For example, the expected value of a random variable can be computed using its CDF only. Hence, we shall often make statements like "X is a random variable with pmf f" or "X is a random variable with CDF F", without bothering to indicate the probability space.

Some distributions (i.e., CDF or the associated pmf) occur frequently enough to merit a name.

2. Examples of discrete probability distributions

EXAMPLE 11. Let f and F be the pmf, CDF pair

$$f(t) = \begin{cases} p & \text{if } t = 1, \\ q & \text{if } t = 0, \end{cases} \qquad F_X(t) = \begin{cases} 1 & \text{if } t \geq 1, \\ q & \text{if } t \in [0,1), \\ 0 & \text{if } t < 0. \end{cases}$$

A random variable X having this pmf (or equivalently the CDF) is said to have *Bernoulli distribution* with parameter p and write $X \sim \text{Ber}(p)$. For example, if $\Omega = \{1, 2, \ldots, 10\}$ with $p_i = 1/10$, and $X(\omega) = \mathbf{1}_{\omega \leq 3}$, then $X \sim \text{Ber}(0.3)$. Any random variable taking only the values 0 and 1, has Bernoulli distribution.

EXAMPLE 12. Fix $n \geq 1$ and $p \in [0,1]$. The pmf defined by $f(k) = \binom{n}{k} p^k q^{n-k}$ for $0 \leq k \leq n$ is called the *Binomial distribution* with parameters n and p and is denoted $\text{Bin}(n,p)$. The CDF is as usual defined by $F(t) = \sum_{u:\mathbf{u} \leq t} f(u)$, but it does not have any particularly nice expression.

For example, if $\Omega = \{0,1\}^n$ with $p_{\underline{\omega}} = p^{\sum_i \omega_i} q^{n - \sum_i \omega_i}$, and $X(\underline{\omega}) = \omega_1 + \ldots + \omega_n$, then $X \sim \text{Bin}(n,p)$. In words, the number of heads in n tosses of a p-coin has $\text{Bin}(n,p)$ distribution.

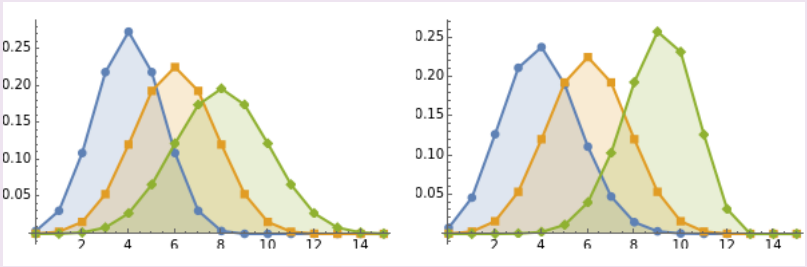

FIGURE 1. PMF of $\text{Bin}(n, 1/2)$ with $n = 8, 12, 16$ on the left. PMF of $\text{Bin}(12, p)$ with $p = \frac{1}{3}, \frac{1}{2}, \frac{3}{4}$ on the right. The lines joining the points have no meaning and are shown only for better visual effect.

EXAMPLE 13. Fix $p \in (0,1]$ and let $f(k) = q^k p$ for integer $k \geq 0$. This is called the *Geometric distribution* with parameter p and is denoted $\text{Geo}(p)$. The CDF is

$$F(t) = \begin{cases} 0 & \text{if } t < 0, \\ 1 - q^k & \text{if } k - 1 \leq t < k, \text{ for some } k \geq 1. \end{cases}$$

For example, if a p-coin is tossed till the first head turns up, then the number of tails is a random variable with $\text{Geo}(p)$ distribution. In some books, the random variable considered is the total number of tosses till the first head. This is one more than the number of tosses, and has pmf $g(k) = q^{k-1} p$ for $k = 1, 2 \ldots$. To prevent ambiguity, let us denote this distribution $\text{Geo}_+(p)$ in these notes (this is not standard notation).

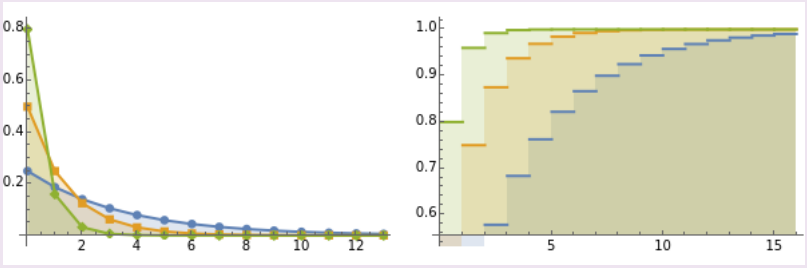

FIGURE 2. PMF and CDF of Geo(p) with $p = \frac{1}{4}, \frac{1}{2}, \frac{4}{5}$. Again, for the PMF, the lines joining the points are there for better visualization.

EXAMPLE 14. If we extend the previous experiment and toss a coin till m heads turn up, then the pmf of the number of tails is

$$g(k) = \binom{k+m-1}{k} q^k p^m, \qquad k \geq 0.$$

This is because the $k+m$ toss must be a head, and of the previous $k+m-1$ tosses, exactly k must be tails. This distribution is called Negative-Binomial distribution with parameters m and p, and denoted Neg-Bin$(m;p)$. When $m = 1$, this is the same as Geo(p). The pmf can also be written as[1]

$$g(k) = (-1)^k \binom{-m}{k} q^k p^m, \quad \text{for } k \geq 0,$$

which explains the term "negative binomial". Observe that written this way, we may also allow the parameter m to be any positive number, not necessarily an integer! Indeed, by the generalized Binomial theorem

$$\sum_{k=0}^{\infty} (-1)^k \binom{-m}{k} q^k p^m = p^m (1-q)^{-m} = 1,$$

for any m, showing that g is a valid pmf. For non-integer m, it does not make sense to talk of "waiting for the mth head", of course.

Like before, if we count the number of tosses till the mth head, then the pmf is

$$f(k) = \binom{k-1}{k-m} (1-p)^{k-m} p^m, \text{ for } k \geq m,$$

and we denote this distribution as Neg-Bin$_+(m;p)$.

EXAMPLE 15. Fix $\lambda > 0$ and define the pmf $f(k) = e^{-\lambda} \frac{\lambda^k}{k!}$. This is called the *Poisson distribution* with parameter λ and is denoted Pois(λ).

[1] For any complex number α and integer $k \geq 0$, recall that the binomial coefficient $\binom{\alpha}{k} = \frac{\alpha(\alpha-1)\ldots(\alpha-k+1)}{k!}$. When α is a positive integer, this agrees with the usual definition. When $\alpha = -m$ where m is a positive integer, we can rewrite this as $\binom{-m}{k} = (-1)^k \binom{m+k-1}{k}$. Thus $g(k) = (-1)^k \binom{-m}{k}$, explaining why the term "negative binomial" is used. The most important occurrence of the general binomial coefficients is in the generalized Binomial theorem that says that $(1+x)^\alpha = \sum_{k \geq 0} \binom{\alpha}{k} x^k$ valid for $|x| < 1$ and any $\alpha \in \mathbb{C}$.

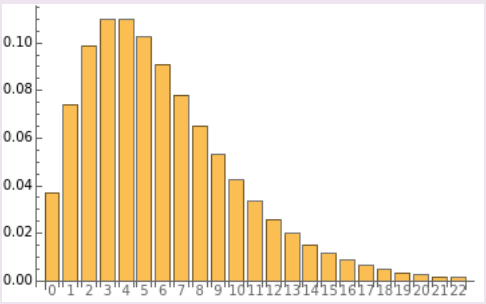

FIGURE 3. PDF of Neg-Bin(3; $\frac{1}{3}$).

Although Poisson is one of the most important of the distributions, it is hard to think of a simple natural experiment which gives a random variable having exactly the Poisson distribution (contrast with the Binomial and Geometric distributions that occur in coin-tossing experiments)! Mathematically it always arises as a limiting case. For example, in the problem of a psychic (randomly) guessing the cards in a deck, we saw that the number of matches (correct guesses) had an *approximately* Pois(1) distribution. Here is another example.

Let $X \sim \text{Bin}(n,p)$, where n is large, p is small, but $np = \lambda$ remains constant. Then,

$$\begin{aligned}\mathbb{P}\{X = k\} &= \binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k}\\ &= \frac{n(n-1)\dots(n-k+1)}{k!}\frac{\lambda^k}{n^k}\left(1-\frac{\lambda}{n}\right)^{n-k}.\end{aligned}$$

Note that $\frac{n(n-1)\dots(n-k+1)}{n^k} \to 1$ as $n \to \infty$ (since k is fixed). Also, $(1-\frac{\lambda}{n})^{n-k} \to e^{-\lambda}$ (if not clear, note that $(1-\frac{\lambda}{n})^n \to e^{-\lambda}$ and $(1-\frac{\lambda}{n})^{-k} \to 1$). Hence, the right hand side above converges to $e^{-\lambda}\frac{\lambda^k}{k!}$ which is the Pois(λ) pmf. In short, Bin($n, \frac{\lambda}{n}$) distribution is close to Pois(λ) distribution if n is large.

But why would we have the situation of large n and small p but balancing out so that $np = \lambda$? Here is an example to show that this is not so unnatural as it may appear at first.

EXAMPLE 16. (A physical example). A large amount of *payasa* is made in the hostel mess to serve 100 students. The intention is that each student get 2 raisins in their cup, so the cook adds 200 raisins and mixes the *payasa*, and then pours it into 100 cups. But the number of raisins that a given student gets is not necessarily 2 but random. Each raisin has a 0.01 chance to come into that particular cup, and there are 200 raisins, so the number of raisins in the cup must have Bin(200, 0.01). Observe that $np = 200 \times 0.01 = 2$. Thus the above calculations show that the number of raisins in a given cup has approximately Pois(2) distribution.

EXAMPLE 17. Fix positive integers b, w and $m \le b + w$. Define the pmf $f(k) = \frac{\binom{b}{k}\binom{w}{m-k}}{\binom{b+w}{m}}$ where the binomial coefficient $\binom{x}{y}$ is interpreted to be zero if $y > x$ (thus $f(k) > 0$ only for $\max\{m-w, 0\} \le k \le b$). This is called the *Hypergeometric distribution* with parameters b, w, m and we shall denote it by Hypergeo(b, w, m).

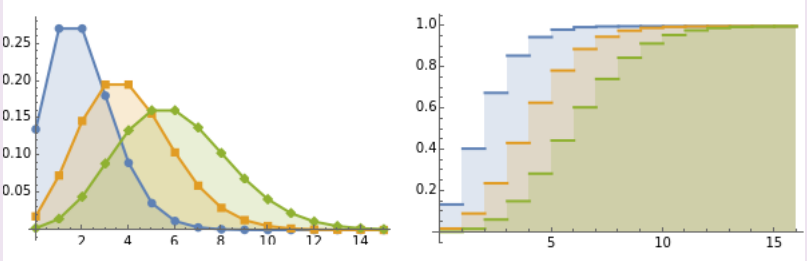

FIGURE 4. PMF and CDF of Pois(λ) with $\lambda = 2, 4, 6$.

Consider a population with b men and w women. The number of men in a random sample (without replacement) of size m, is a random variable with the Hypergeo(b, w, m) distribution.

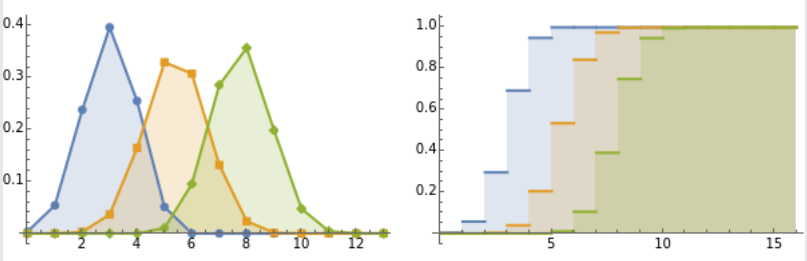

FIGURE 5. PMF and CDF of Hypergeo$(12, 8, m)$ with $m = 5, 9, 13$.

EXAMPLE 18. *Zipf's law* says that in natural languages, the most common word is twice as likely as the second most common word and three times more likely than the third most common word etc. If the words in the language are listed as $1, 2, \ldots, N$ in decreasing order of frequency, then Zipf's law is suggesting the pmf $f(k) = \frac{1/H_N}{k}$ for $k = 1, 2, \ldots, N$, where $H_N = 1 + \frac{1}{2} + \ldots + \frac{1}{N}$ is the normalizing constant. The CDF F is given by

$$\begin{aligned} F(x) &= \frac{1}{H_N}(1 + \frac{1}{2} + \ldots + \frac{1}{k}) \quad \text{if } k \leq x < k+1 \text{ for some } 1 \leq k \leq N-1 \\ &= \frac{H_k}{H_N}. \end{aligned}$$

It is well-know that $H_n = \log n - \gamma + o(1)$ as $n \to \infty$ where $\gamma = 0.5772\ldots$ is the Euler-Macheroni constant. Thus, if we take $k = N^b$ with $0 < b < 1$, then we see that $F(N^b) \approx b$. In other words, just $N^{0.99}$ words (which is a negligible fraction of N in the sense that $\frac{N^{0.99}}{N} \to 0$ as $N \to \infty$) comprise more than 99% of all written text in that language!

Similar observations hold for wealth distribution (it is said that the top 1% of rich people in India have about 40% of the wealth in the country). But it is not always Zipf's law that applies. A more general class of pmfs is given by $f(k) = \frac{1/C_{N,\alpha}}{k^\alpha}$ for $1 \leq k \leq N$. Here $\alpha > 0$ is a parameter and $C_{N,\alpha} = 1 + \frac{1}{2^\alpha} + \ldots + \frac{1}{N^\alpha}$.

EXAMPLE 19. *Discrete Pareto distribution* Fix $s > 1$ and let $f(k) = \frac{1/\zeta(s)}{k^s}$ for $k = 1, 2, \ldots$, where $\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}$ (it is called the *Riemann-zeta function*). Observe that this does not make sense for $s = 1$ as $\sum \frac{1}{k} = \infty$, which is why in Zipf's law we restricted to finite N.

If X has discrete Pareto distribution with parameter s, then $\mathbb{P}\{X \geq m\} = \sum_{k=m}^{\infty} \frac{1/\zeta(s)}{k^\alpha} \asymp \frac{1}{m^{\alpha-1}}$. The key point is that this "tail probability" decays slowly (unlike in Poisson, where it decays exponentially fast). This is the reason why Pareto is used to describe wealth distributions where the richest person has orders of magnitude more wealth than a typical person (but not, for example, heights of people: even the tallest person is less than twice the average height!).

3. Expectations of discrete random variables

While the pmf and CDF have all the information about a random variable, in many cases it is impossible or difficult to find them explicitly. In such cases, it is useful to have partial information. If one had to summarize the distribution by a single number, what should that number be? One of the best choices is the expectation.

DEFINITION 20. Let X be a random variable on (Ω, p) with pmf f. Its expected value is $\mathbb{E}[X] = \sum_{\omega \in \Omega} X(\omega) p_\omega$, provided the sum converges absolutely. Otherwise we say that the expectation does not exist.

3.1. Properties of expectation. Let X, Y be random variables on (Ω, p) and let $a, b \in \mathbb{R}$.

(1) Linearity: If X, Y have expectation, then so does $aX + bY$ and $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$. This is because

$$\begin{aligned}\mathbb{E}[aX + bY] &= \sum_{\omega} (aX(\omega) + bY(\omega))p(\omega) \\ &= a\sum_{\omega} X(\omega)p(\omega) + b\sum_{\omega} Y(\omega)p(\omega) = a\mathbb{E}[X] + b\mathbb{E}[Y].\end{aligned}$$

(2) Positivity/Monotonicity: If $X \geq 0$ (pointwise), then $\mathbb{E}[X] \geq 0$ with equality if and only if $\mathbb{P}\{X = 0\} = 1$. This is clear from the definition $\mathbb{E}[X] = \sum_\omega X(\omega)p(\omega)$. There can be no cancellation, and the sum is zero if and only if $X(\omega)p(\omega) = 0$ for all ω, which is the same as $\sum_{\omega: X(\omega)>0} p(\omega) = 0$, or equivalently $\mathbb{P}\{X > 0\} = 0$.

If $X \geq Y$ (pointwise) and expectations exist, then applying positivity to $X - Y$, we see that $\mathbb{E}[X] \geq \mathbb{E}[Y]$ with equality if and only if $\mathbb{P}\{X = Y\} = 1$.

(3) $\mathbb{E}[\mathbf{1}_A] = \mathbb{P}(A)$. In particular, $\mathbb{E}[\mathbf{1}] = 1$.

These three properties are crucial. In fact, one can forget the definition of expectation and simply work from these three properties.

REMARK 21. In fact, at least for finite Ω, these properties force the definition of expectation. Indeed, any random variable X can be written as a linear combination of indicators of singleton sets: $X(\cdot) = \sum_{\omega' \in \Omega} X(\omega')\mathbf{1}_{\{\omega'\}}$. The first and third property then force the definition of $\mathbb{E}[X]$. This argument does not quite work when Ω is countably infinite. To do that we need to observe

one more property which we omitted for simplicity: If $X_n, X \geq 0$ and $X_n \uparrow X$ pointwise, then $\mathbb{E}[X_n] \uparrow \mathbb{E}[X]$. This is known as MCT (monotone convergence). MCT is for random variables what countable additivity is for events (and linearity is analogous to finite additivity).

3.2. Computing expectations. In computing expectations, it helps to note that it depends only on the distribution of the random variable. Indeed, if X has pmf f, we claim that

$$\mathbb{E}[X] = \sum_{t \in \mathbb{R}} t f(t).$$

To show this, let Range$(X) = \{x_1, x_2, \ldots\}$. Let $A_k = \{\omega : X(\omega) = x_k\}$. By definition of pmf we have $\mathbb{P}(A_k) = f(x_k)$. Further, A_k are pairwise disjoint and exhaustive. Hence

$$\mathbb{E}[X] = \sum_{\omega \in \Omega} X(\omega) p_\omega = \sum_k \sum_{\omega \in A_k} X(\omega) p_\omega = \sum_k x_k \mathbb{P}(A_k) = \sum_k x_k f(x_k).$$

All the rearrangements of the sums here are justified by the assumption that expectation exists, i.e., absolute convergence of $\sum X(\omega)p(\omega)$.

► Going further, if $h : \mathbb{R} \to \mathbb{R}$ is any function, then the random variable $h(X)$ has expectation $\mathbb{E}[h(X)] = \sum_k h(x_k) f(x_k)$. For example, for $h(x) = x^2$ this says that $\mathbb{E}[X^2] = \sum_k x_k^2 f(x_k)$. Although this sounds trivial, there is a very useful point here. To calculate $\mathbb{E}[X^2]$ we do not have to compute the pmf of X^2 first, which can be done but would be more work. Instead, in the above formulas, $\mathbb{E}[h(X)]$ has been computed directly in terms of the pmf of X. For a proof, again partition Ω into the sets $A_k = \{\omega : X(\omega) = x_k\}$ and write

$$\mathbb{E}[h(X)] = \sum_{\omega \in \Omega} h(X(\omega)) p_\omega = \sum_k \sum_{\omega \in A_k} h(X(\omega)) p_\omega = \sum_k h(x_k) \mathbb{P}(A_k) = \sum_k h(x_k) f(x_k).$$

Example 22. Let $\Omega = \{0,1\}^n$ and $p(\omega) = p^{\omega_1 + \ldots + \omega_n} q^{n - (\omega_1 + \ldots + \omega_n)}$ be the probability space corresponding to n tosses of a coin. Let $X(\omega) = \omega_1 + \ldots + \omega_n$ be the number of heads. We can compute the expected value of X in three ways.

(1) Directly from the definition,

$$\mathbb{E}[X] = \sum_{\omega \in \{0,1\}^n} (\omega_1 + \ldots + \omega_n) p^{\omega_1 + \ldots + \omega_n} q^{n - \omega_1 + \ldots - \omega_n}$$

One way to compute this sum is to differentiate the identiity $(x+y)^n = \sum_\omega x^{\omega_1 + \ldots + \omega_n} y^{n - \omega_1 - \ldots - \omega_n}$ w.r.t. x, multiply by x, and then set $x = p$ and $y = q$. That gives $nx(x+y)^{n-1}|_{x=p, y=q} = np$.

(2) We know that $X \sim \text{Bin}(n,p)$. Hence using $k\binom{n}{k} = n\binom{n-1}{k-1}$, we see that

$$\mathbb{E}[X] = \sum_{k=0}^n k \binom{n}{k} p^k q^{n-k} = np \sum_{k=1}^n \binom{n-1}{k-1} p^{k-1} q^{n-k}$$

$$= np(p+q)^{n-1} \;=\; np.$$

(3) Write $X = X_1 + \ldots + X_n$ where $X_k(\omega) = \omega_k$ is the indicator that the kth toss is a head. Now $X_k \sim \text{Ber}(p)$, hence $\mathbb{E}[X_k] = p$, and therefore $\mathbb{E}[X] = np$ by linearity.

In this example, all three ways work, and it is difficult to appreciate the difference between the approaches. The value of the third method becomes clearer in cases where computing the pmf is way to complicated.

EXAMPLE 23. Consider the problem of throwing r distinguishable balls at random into m labelled bins. Let X denote the number of empty bins. Using the inclusion-exclusion formula and its extensions, we can write rather complicated formulas for the pmf of X. However it is much easier to realize that $X = X_1 + \ldots + X_m$, where $X_k = \mathbf{1}_{\text{bin } k \text{ is empty}}$. Then $\mathbb{E}[X_k] = \mathbb{P}\{k\text{th bin is empty}\} = (m-1)^r/m^r$. Summing over k, we see that $\mathbb{E}[X] = (m-1)^r/m^{r-1}$.

EXERCISE 24. Find $\mathbb{E}[X]$ and $\mathbb{E}[X^2]$ in each case. (1) $X \sim \text{Bin}(n,p)$. (2) $X \sim \text{Geo}(p)$. (3) $X \sim \text{Neg-Bin}(m;p)$. (4) $X \sim \text{Pois}(\lambda)$. (5) $X \sim \text{Hypergeo}(b,w;m)$.

EXERCISE 25. Consider the experiment of a psychic guessing a deck of n cards. Let X be the number of correct guesses. Without computing the pmf of X, show that $\mathbb{E}[X] = 1$ (yes, it does not depend on the size of the deck!)

4. Other quantities associated to distributions

Once we have the notion of expectation, we can define many other quantities of importance and interest. This is the case for all the quantities except for quantiles below.

Quantiles: For $0 < p < 1$, the pth quantile of X is any number x such that $\mathbb{P}\{X < x\} \le t \le \mathbb{P}\{X \le x\}$. The $\frac{1}{2}$-quantile is called the *median* and the $\frac{1}{4}$ and $\frac{3}{4}$ quantiles are called the lower and upper *quartiles*, respectively. They are not unique, for instance for $X \sim \text{Ber}(1/2)$, any number $x \in [0,1]$ is a median.

Moments: Let $k \ge 0$ be an integer. If $\mathbb{E}[X^k]$ exists, it is called the k^{th} *moment* of X. Existence means that $\mathbb{E}[|X|^k] < \infty$. While $|X|^k$ is defined for any real number k, note that X^k is in general well-defined only when k is an integer. For this reason, if $\mathbb{E}[|X|^p] < \infty$, we say that X has finite p-th moment, but unless $X \ge 0$ or p is an integer, there is no number called the pth moment of X.

Observe that if $0 < p_1 < p_2$, then $|X|^{p_1} \le |X|^{p_2} + 1$ pointwise. Therefore, by the monotonicity of expectation, if the p_2 moment is finite, then so is the p_1 moment. In words, if a moment exists, all lower moments also exist.

Variance: Let $\mu = \mathbb{E}[X]$ and define $\sigma^2 := \mathbb{E}\left[(X-\mu)^2\right]$. This is called the *variance* of X, also denoted by $\mathrm{Var}(X)$. It can be written in other forms. For example,

$$\begin{aligned}\sigma^2 &= \mathbb{E}[X^2 + \mu^2 - 2\mu X] && \text{(by expanding the square)}\\ &= \mathbb{E}[X^2] + \mu^2 - 2\mu\mathbb{E}[X] && \text{(by the linearity of expectation)}\\ &= \mathbb{E}[X^2] - \mu^2.\end{aligned}$$

That is $\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$. From the above calculations, it is also clear that the variance is finite if and only if the second moment is finite.

Standard deviation: The standard deviation of X is defined as $\mathrm{s.d.}(X) := \sqrt{\mathrm{Var}(X)}$. Note that it has the same dimensions as X. That is, if X is the height measured in meters, then $\mathrm{s.d.}(X)$ is also in meters (in contrast, the variance has units of squared-meters).

Mean absolute deviation: The mean absolute deviation of X is defined as the $\mathbb{E}[|X - \mathrm{med}(X)|]$.

Variance standard deviation, mean absolute deviation are all measures of spread or dispersion. We discuss it more detail below, after introducing a few other quantities of interest.

Coefficient of variation: The coefficient of variation of X is defined as $\mathrm{c.v.}(X) = \frac{\mathrm{s.d.}(X)}{|\mathbb{E}[X]|}$. This makes more sense for a positive random variable, for example, the height of a randomly chosen person in a population. If heights are converted from meters to centimeters, both mean and standard deviation scale up by a factor of 100, but the coefficient of variation remains the same. It is a dimension-free pure number.

Entropy: The *Shannon entropy* of a random variable X having pmf f is defined as

$$H(X) = -\sum_i f(t)\log_2 f(t_i)$$

Observe that unlike expectation, the entropy does not actually care about the values taken but only the probabilities of the distinct values. For example, if $X \sim \mathrm{Ber}(p)$ and $Y = 10X$, then $H(Y) = H(X) = -p\log_2 p - q\log_2 q$. In that sense, it is better to think of X as a *categorical random variables*, taking abstract values $a, b, c, \ldots$ that have no relationship such as closeness between them (unlike numbers). This allows us to use the same definition for a random variable taking values in any discrete set.

To see what entropy measures, see Figure 4. It vanishes at $p = 0$ and $p = 1$, is symmetric about $p = 1/2$, and maximum at $p = 1/2$. In some sense that matches how "predictable" a coin toss is. If $p = 0$ or $p = 1$, it is entirely predictable, and it is least predictable when $p = 1/2$. Thus, entropy is a measure of unpredictability.

Generating functions: A fruitful idea in mathematics is that of associating to a sequence (possibly finite) of numbers $(a_n)_{n\geq 0}$, a *generating function* $A(t) = \sum_{n\geq 0} a_n t^n$. We know from

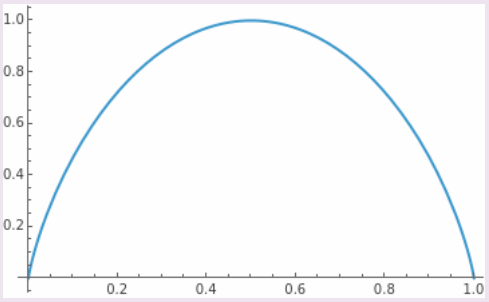

FIGURE 6. Shannon entropy of Ber(p) as a function of p.

analysis class that there is a number R (called radius of convergence) such that $A(t)$ converges for $|t| < R$ and diverges for $|t| > R$ (it can go either way when $|t| = R$). If $R = 0$ we are lost. But if $R > 0$, then A defines a smooth function on $(-R, R)$ and the power of Calculus can be brought to bear on the study of the sequence $(a_n)_{n\geq 0}$. This idea and its variants are useful to study the pmf of a random variable.

(1) Probability generating function (PGF): If X takes values in $\{0, 1, 2, \ldots\}$, we define its PGF as

$$G(s) = \mathbb{E}[s^X] = \sum_{n=0}^{\infty} f_X(n) s^n \tag{7}$$

provided the sum converges. Clearly it does for $|s| \leq 1$, as $\sum_n f_X(n) = 1$. If X takes fractional values, s^X does not make sense if $s < 0$, hence the restriction to $\mathbb{N}$-valued random variables.

(2) Moment generating function (MGF): If X is a real-valued random variable, we define its MGF as

$$M(t) = \mathbb{E}[e^{tX}] = \sum_x f_X(x) e^{tx},$$

provided the expectation exists. In a formal sense, MGF may be thought of as the PGF of X at e^t (since $e^t > 0$, the issue of ill-definedness does not arise).

(3) Characteristic functions (CF): If X is a real-valued random variable, we define its characteristic function as

$$\Psi(t) = \mathbb{E}[e^{itX}] := \mathbb{E}[\cos(tX)] + i\mathbb{E}[\sin(tX)].$$

For any $t \in \mathbb{R}$, the random variables $\sin(tX)$ and $\cos(tX)$ are bounded, hence the expectations exist. Thus unlike the PGF or MGF, the characteristic function is always well-defined. Its tremendous use will be seen when proving the central limit theorem (not in this course!).

In this course we shall only use the PGF to some extent.

EXAMPLE 26. Let $X \sim \text{Pois}(\lambda)$. Then its PGF is

$$G(s) = \sum_{k\geq 0} e^{-\lambda}\frac{\lambda^k}{k!}s^k = e^{-\lambda+\lambda s}.$$

In this case, the PGF is defined for all $s \in \mathbb{R}$.

Of what use is the PGF? If we differentiate the series (7) w.r.t. s repeatedly (term-by-term differentiation on the right has to be justified; we leave it to your analysis class, and it is believable in any case) we get $G'(s) = \sum_{n\geq 1} f(n)ns^{n-1}$, $G''(s) = \sum_{n\geq 2} n(n-1)f(n)s^{n-2}$, and more generally

$$G^{(r)}(s) = \sum_{n\geq r} n(n-1)\dots(n-r+1)f(n)s^{n-r}.$$

In particular, setting $s = 1$ (this also requires justification, as $s = 1$ may be on the boundary of convergence), we get $G^{(r)}(1) = \mathbb{E}[X(X-1)\dots(X-r+1)]$. The PGF thus provides a convenient short-cut to some calculations.

EXAMPLE 27. If $X \sim \text{Pois}(\lambda)$, the PGF is $G(s) = e^{-\lambda+\lambda s}$. Therefore, $\mathbb{E}[X(X-1)\dots(X-r+1)] = G^{(r)}(1) = \lambda^r$ for all $r \geq 1$. In particular,

$$\text{Var}(X) = \mathbb{E}[X(X-1)] + \mathbb{E}[X] - \mathbb{E}[X]^2 \;=\; G''(1) + G'(1) - G'(1)^2 \;=\; \lambda$$

and $\text{s.d.}(X) = \sqrt{\lambda}$.

EXERCISE 28. If $X \sim \text{Pois}(1)$, show that $\mathbb{E}[X^r]$ is the number of ways to partition the set $[r]$. For example, if $r = 3$, the partitions are $\{\{1\},\{2\},\{3\}\}$, $\{\{1,2\},\{3\}\}$, $\{\{1,3\},\{2\}\}$, $\{\{2,3\},\{1\}\}$, $\{\{1,2,3\}\}$, so $\mathbb{E}[X^3] = 4$.

Discussion: What do these quantities mean? Mean and median try to summarize the distribution of X by a single number. They are called *measures of central tendency*. We have already discussed their relative merits. The other quantities measure various other features of the distribution. Of particular importance is are measures of *spread* or *dispersion*.

The variance, the standard deviation and the mean absolute deviation are measures of dispersion. They measure how much a distribution is spread out. Suppose the average height of people in a city is 160 cm. This could be because everyone is 160 cm exactly or because half the people are 100 cm. while the other half are 220 cm., or alternately the heights could be uniformly spread over 150-170 cm., etc. To measure spread, one idea would be to fix a number a and consider $\mathbb{E}[|X-a|]$ (of course naively one might think $\mathbb{E}[X-a]$, but that is just $\mathbb{E}[X]-a$ and has no information other than the mean). It turns out (exercise in the homework) that this quantity is minimized when $a = \text{Med}(X)$, and the value for that a is precisely the mean absolute deviation.

But mathematically it is much better to consider $\mathbb{E}[|X-a|^2]$, which is minimized when $a = \mathbb{E}[X]$ and the value is the variance. Why is it better? Naive reason: Absolute value is a

pain, square can be expanded to see that $\mathbb{E}[|X-a|^2] = \mathbb{E}[X^2] - 2a\mathbb{E}[X] + a^2$. A more refined reason: Think of it as analogous to how we measure the distance between (p_1, q_1) and (p_2, q_2) by the Pythagorean expression $\sqrt{(p_1-q_1)^2+(p_2-q_2)^2}$ and not by $|p_1-q_1|+|p_2-q_2|$.

The standard deviation has the same units as the quantity. Fo example, if mean height is 160cm measured in centimeters with a standard deviation of 10cm, and the mean weight is 55kg with a standard deviation of 5kg, then we cannot say which of the two is less variable. To make such a comparison we need a dimension free quantity (a pure number). Coefficient of variation is such a quantity, as it measure the standard deviation per mean. For the height and weight data just described, the coefficients of variation are 1/16 and 1/11, respectively. Hence we may say that height is less variable than weight in this example.

5. Exercises

PROBLEM 1. A random experiment is described and a random variable observed. In each case, write the probability space, the random variable and the pmf of the random variable.

(1) Two fair dice are thrown. The sum of the two top faces is noted.

(2) Deal thirteen cards from a shuffled deck and count (a) the number of red cards (i.e., diamonds or hearts), (b) the number of kings, (c) the number of diamonds.

PROBLEM 2. Place r distinguishable balls in m distinguishable bins at random. Count the number of balls in the first bin.

(1) Write the probability space and the random variable described above.

(2) Find the probability mass function of the number of balls in the first bin.

(3) Find the expected value of the number of balls in the first bin.

PROBLEM 3. A number X is selected uniformly at random from $[N]$. Let U be the units digit of X^2. What is the probability distribution of U? The answer depends on N, but our interest is in what happens as $N \to \infty$. What if U denote the units deigit of X^3?

PROBLEM 4. Find $\mathbb{E}[X]$ and $\mathbb{E}[X^2]$ for the following random variables.

(1) $X \sim \text{Geo}(p)$.

(2) $X \sim \text{Hypergeo}(N_1, N_2, m)$.

PROBLEM 5. Let X be a non-negative integer-valued random variable with CDF $F(\cdot)$. Show that $\mathbb{E}[X] = \sum_{k=0}^{\infty}(1-F(k))$.

PROBLEM 6. A coin has probability p of falling head. Fix an integer $m \geq 1$. Toss the coin till the m^{th} head occurs. Let X be the number of tosses required.

(1) Show that X has pmf

$$f(k) = \binom{k-1}{m-1} p^m (1-p)^{k-m}, \quad k = m, m+1, m+2, \ldots.$$

(2) Find $\mathbb{E}[X]$ and $\mathbb{E}[X^2]$.

[**Note:** When $m = 1$, you should get the Geometric distribution with parameter p. We say that X has *negative-binomial distribution*. Some books define $Y := X - m$ (the number of tails till you get m heads) to be a negative binomial random variable. Then, Y takes values $0, 1, 2, \ldots$.]

PROBLEM 7. A coin is tossed n times. Let X be the number of heads minus the number of tails. Find the pmf of X. Hence or otherwise, find $\mathbb{E}[X]$.

PROBLEM 8. Suppose r balls are thrown into m bins, uniformly at random. Fix k and let X be the number of bins that have exactly k balls. Find $\mathbb{E}[X]$.

PROBLEM 9. A smoker has two matchboxes, one in each pocket of his jacket. Initially the boxes contain N matches each. Every time he wants to smoke, he puts his hand randomly in one of the two pockets and picks a matchstick from the corresponding box. There comes a first time when he finds the matchbox empty. Let X be the number of matches in the other box at that time. Find the pmf of X.

PROBLEM 10. A set is a triple of cards of the same kind (e.g., three aces) and a series is a triple of cards of the same suit that are consecutive (e.g., A,1,2 of spades, J,Q,K of clubs, but not Q,K,A).A hand of thirteen cards is dealt from a well-shuffled deck of 52 cards. Find the expected number of sets and series in the hand.

PROBLEM 11. A fair coin is tossed till two a HT pattern occurs for the first time (i.e., a head followed immediately by a tail). What is the pmf of the number of tosses? What is its expectation?Repeat the problem for the pattern HH (two consecutive heads).

PROBLEM 12. For a pmf $f(\cdot)$, the *mode* is defined as any point at which f attains its maximal value (i.e., t is a mode if $f(t) \geq f(s)$ for any s). For each of the following distributions, find the mode(s) of the distribution and the value of the pmf at the modes.

(1) Bin(n, p).

(2) Pois(λ).

(3) Geo(p).

PROBLEM 13. Use MATLAB for the following exercise.

(1) Plot the pmf of Binomial, Poisson and Geometric distributions for various values of the parameters. Observe the plots to say where the maximum is attained, how the shape changes with changes in parameter, etc.

(2) Simulate random numbers (number of samples can be 50 or 100 etc) from the same distributions and plot their histograms. Visually compare the histograms with the plots of the pmf.

(3) Consider the "real-life" data given in Feller's book (chapter 6) and plot their histograms. Compare with the plot of the pmf for the appropriate distribution with appropriate choice of parameters.

PROBLEM 14. Suppose r distinguishable balls are placed in m labelled bins at random. Each ball has probability p_k of going into the kth bin, where $p_1 + \ldots + p_m = 1$. Let X_k be the number of balls that go into the kth bin.

(1) Find the pmf of X_1.

(2) Find the pmf of the random variable $X_1 + X_2$.

PROBLEM 15. Two fair dice are thrown and let X be the total of the two numbers that show up. Find the pmf of X. What is the most likely value of X?

PROBLEM 16. In response to a job posting, a very large number N of candidates have applied. You have an interview procedure that can accurately rank candidates by their ability. Your goal is to hire someone of the top 10% of all applicants. If you allow yourself a chance of 0.001 to make a mistake, how many candidates do you need to interview? [*Remark:* If you don't allow any chance of mistake at all, you must interview at least $0.9N$ candidates and then pick the top one among them. Compare that to your answer.]

PROBLEM 17. A coin has probability p of falling head. Assume $0 < p < 1$ and fix an integer $m \geq 1$. Toss the coin till the m^{th} head occurs. Let X be the number of tosses required. Show that X has pmf

$$f(k) = \binom{k-1}{m-1} p^m (1-p)^{k-m}, \quad k = m, m+1, m+2, \ldots.$$

Find the CDF of X.[**Note:** When $m = 1$, this is the Geometric distribution with parameter p. We say that X has *negative-binomial distribution*. Some books define $Y := X - m$ (the number of tails till you get m heads) to be a negative binomial random variable. Then, Y takes values $0, 1, 2, \ldots$.]

PROBLEM 18. A box contains n coupons with one number on each coupon. We do not know the numbers but we know that they are distinct. Coupons are drawn one after another from the box, without replacement (i.e., after choosing a coupon at random, it is not put back into the box before drawing the next coupon). If the kth number drawn is larger than all the previous numbers, what is the chance that it is the largest of the n numbers?

CHAPTER 5

Probability distributions beyond the discrete

For a random variable X on a discrete probability space, we associated a CDF $F : \mathbb{R} \to [0,1]$ by defining $F(t) = \mathbb{P}\{X \leq t\}$. We saw that F is increasing ($F(t) \leq F(s)$ if $t \leq s$), right-continuous ($F(t+) = F(t)$ for all t), $F(t)$ goes to 0 or 1 as t tends to $-\infty$ or $+\infty$ respectively, and increases by jumps (i.e., $F(t) - F(s) = \sum_{s<u\leq t}(F(u) - F(u-))$). Let us drop the last property of increasing in jumps to make the following general definition of a CDF or distribution function.

Deeinition 19. A (cumulative) distribution function (or CDF for short) is any function $F : \mathbb{R} \to [0,1]$ that is increasing, right continuous, and satisfies $F(t) \to 0$ as $t \to -\infty$ and $F(t) \to 1$ as $t \to +\infty$.

Cumulative distribution functions of random variables on discrete probability spaces are distribution functions by this definition. However, there are others.

Example 20. Let

$$F(t) = \begin{cases} 0 & \text{if } t \leq 0, \\ t & \text{if } 0 < t < 1, \\ 1 & \text{if } t \geq 1. \end{cases}$$

Then it is easy to see that F is a distribution function. However, it has no jumps and hence it does not arise as the CDF of any random variable on a discrete probability space.

The question is, does this have any probability interpretation? Does it make sense to talk of a random variable X such that $\mathbb{P}\{X \leq t\} = F(t)$? Unless F increases by jumps, we cannot find such a random variable on any discrete probability space. We must go beyond. There are two ways to approach this.

(1) The first way is to learn the notion of uncountable probability spaces, which poses many subtleties. It requires a semester or so of real analysis and what is known as *measure theory* (a deep dive into the notions of length, area, volume). But after that one can define random variables on uncountable probability spaces and the above example will turn out to be the CDF of some random variable on some (uncountable) probability space.

(2) Just regard distribution functions such as the one in the above example as reasonable approximations to CDFs of some discrete random variables. For example, if $\Omega = \{0, 1, 2, \ldots, N\}$ and $p(k) = 1/(N+1)$ for all $0 \leq k \leq N$, and $X : \Omega \mapsto \mathbb{R}$ is defined by

$X(k) = k/n$, then it is easy to check that the CDF of X is given by

$$G(t) = \begin{cases} 0 & \text{if } t < 0, \\ \frac{k}{N+1} & \text{if } \frac{k-1}{N} \le t < \frac{k}{N} \text{ for some } k = 1, 2, \ldots, N \\ 1 & \text{if } t \ge 1. \end{cases}$$

Now, if N is very large, then the function G looks approximately like the function F, see Figure 5. Just as it is convenient to regard water as a continuous medium in some problems (although water is made up of molecules and is discrete at small scales), it is convenient to use the continuous function F as a reasonable approximation to the step function G.

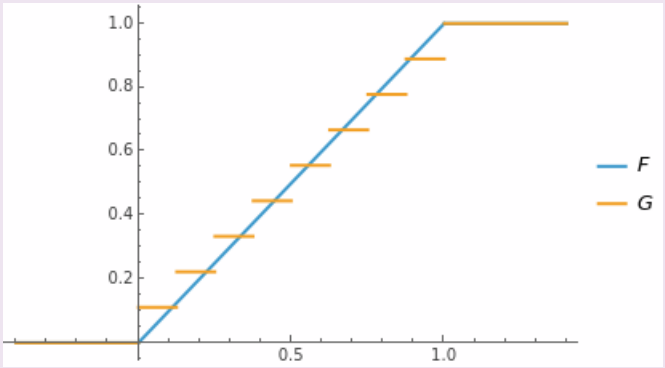

FIGURE 1. The distribution functions F and G (with $n = 8$).

We shall take the second option out. Whenever we write continuous distribution functions such as in the above example, at the back of our mind we have a discrete random variable taking a large number of closely placed values, whose CDF is approximated by our distribution function. The advantage of using continuous objects instead of discrete ones is that the powerful tools of Calculus become available to us.

1. Uncountable probability spaces - conceptual difficulties

The following two "random experiments" are easy to imagine, but difficult to fit into the framework of probability spaces[1].

(1) Toss a p-coin infinitely many times: Clearly the sample space is $\Omega = \{0,1\}^{\mathbb{N}}$. But what is $p_{\underline{\omega}}$ for any $\underline{\omega} \in \Omega$? The only reasonable answer is $p_{\underline{\omega}} = 0$ for all ω. But then how to define $\mathbb{P}(A)$ for any A? For example, if $A = \{\underline{\omega} : \omega_1 = 0, \omega_2 = 0, \omega_3 = 1\}$, then everyone agrees that $\mathbb{P}(A)$ "ought to be" q^2p, but how does that come about? The basic problem is that Ω is uncountable, and probabilities of events are not got by summing probabilities of singletons.

[1]This section should be omitted by everyone other than those who are keen to know what we meant by the conceptual difficulties of uncountable probability spaces

(2) Draw a number at random from $[0,1]$: Again, it is clear that $\Omega = [0,1]$, but it also seems reasonable that $p_x = 0$ for all x. Again, Ω is uncountable, and probabilities of events are not got by summing probabilities of singletons. It is "clear" that if $A = [0.1, 0.4]$, then $\mathbb{P}(A)$ "ought to be" 0.3, but it gets confusing when one tries to derive this from something more basic!

The resolution: Let Ω be uncountable. There is a class of *basic subsets* (usually not singletons) of Ω for which we take the probabilities as given. We also take the rules of probability, namely, countable additivity, as axioms. Then we use the rules to compute the probabilities of more complex events (subsets of Ω) by expressing those events in terms of the basic sets using countable intersections, unions and complements and applying the rules of probability.

Example 21. In the example of infinite sequence of tosses, $\Omega = \{0,1\}^{\mathbb{N}}$. Any set of the form $A = \{\underline{\omega} : \omega_1 = \varepsilon_1, \ldots, \omega_k = \varepsilon_k\}$ where $k \geq 1$ and $\varepsilon_i \in \{0,1\}$ will be called a basic set and its probability is defined to be $\mathbb{P}(A) = \prod_{j=1}^{k} p^{\varepsilon_j} q^{1-\varepsilon_j}$ where we assume that $p > 0$. Now consider a more complex event, for example, $B = \{\underline{\omega} : \omega_k = 1 \text{ for some } k\}$. We can write $B = A_1 \cup A_2 \cup A_3 \cup \ldots$ where $A_k = \{\underline{\omega} : \omega_1 = 0, \ldots, \omega_{k-1} = 0, \omega_k = 1\}$. Since A_k are pairwise disjoint, the rules of probability demand that $\mathbb{P}(B)$ should be $\sum_k \mathbb{P}(A_k) = \sum_k q^{k-1}p$ which is in fact equal to 1.

Example 22. In the example of drawing a number at random from $[0,1]$, $\Omega = [0,1]$. Any interval (a,b) with $0 \leq a < b \leq 1$ is called a basic set and its probability is defined as $\mathbb{P}(a,b) = b - a$. Now consider a non-basic event $B = [a,b]$. We can write $B = A_1 \cup A_2 \cup A_3 \ldots$ where $A_k = (a + (1/k), b - (1/k))$. Then A_k is an increasing sequence of events and the rules of probability say that $\mathbb{P}(B)$ must be equal to $\lim_{k\to\infty} \mathbb{P}(A_k) = \lim_{k\to\infty}(b - a - (2/k)) = b - a$. Another example could be $C = [0.1, 0.2) \cup (0.3, 0.7]$. Similarly argue that $\mathbb{P}(\{x\}) = 0$ for any $x \in [0,1]$. A more interesting one is $D = \mathbb{Q} \cap [0,1]$. Since it is a countable union of singletons, it must have zero probability! Even more interesting is the 1/3-Cantor set. Although uncountable, it has zero probability!

Consistency: Is this truly a solution to the question of uncountable spaces? Are we assured of never running into inconsistencies? Not always.

Example 23. Let $\Omega = [0,1]$ and let intervals (a,b) be open sets with their probabilities defined as $\mathbb{P}(a,b) = \sqrt{b-a}$. This quickly leads to problems. For example, $\mathbb{P}(0,1) = 1$ by definition. But $(0,1) = (0, 0.5) \cup (0.5, 1) \cup \{1/2\}$ from which the rules of probability would imply that $\mathbb{P}(0,1)$ must be at least $\mathbb{P}(0,1/2) + \mathbb{P}(1/2,1) = \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}} = \sqrt{2}$ which is greater than 1. Inconsistency!

Exercise 24. Show that we run into inconsistencies if we define $\mathbb{P}(a,b) = (b-a)^2$ for $0 \leq a < b \leq 1$.

Thus, one cannot arbitrarily assign probabilities to basic events. However, if we use the notion of distribution function to assign probabilities to intervals, then no inconsistencies arise.

THEOREM 25. *Let $\Omega = \mathbb{R}$ and let intervals of the form $(a, b]$ with $a < b$ be called basic sets. Let F be any distribution function.* Define *the probabilities of basic sets as $\mathbb{P}\{(a, b]\} = F(b) - F(a)$. Then, applying the rules of probability to compute probabilities of more complex sets (got by taking countable intersections, unions and complements) will never lead to inconsistency.*

Let F be any CDF. Then, the above consistency theorem really asserts that there exists (a possibly uncountable) probability space and a random variable such that $F(t) = \mathbb{P}\{X \leq t\}$ for all t. We say that X has distribution F. However, it takes a lot of technicalities to define what uncountable probability spaces look like and what random variables mean in this more general setting, we shall never define them.

The job of a probabilist consists in taking a CDF F (then the probabilities of intervals are already given to us as $F(b) - F(a)$ etc.) and find probabilities of more general subsets of $\mathbb{R}$. Here are the working rules. Instead we can use the following simple working rules to answer questions about the distribution of a random variable.

(1) For an $a < b$, we set $\mathbb{P}\{a < X \leq b\} := F(b) - F(a)$.

(2) If $I_j = (a_j, b_j]$ are countably many pairwise disjoint intervals, and $I = \bigcup_j I_j$, then we define $\mathbb{P}\{X \in I\} := \sum_j F(b_j) - F(a_j)$.

(3) For a general set $A \subseteq \mathbb{R}$, here is a general scheme: Find countably many pairwise disjoint intervals $I_j = (a_j, b_j]$ such that $A \subseteq \cup_j I_j$. Then we define $\mathbb{P}\{X \in A\}$ as the infimum (over all such coverings by intervals) of the quantity $\sum_j F(b_j) - F(a_j)$.

All of probability in another line: Take an (interesting) random variable X with a given CDF F and an (interesting) set $A \subseteq \mathbb{R}$. Find $\mathbb{P}\{X \in A\}$.

There are loose threads here but they can be safely ignored for this course. We just remark about them for those who are curious to know.

REMARK 26. The above method starts from a CDF F and defines $\mathbb{P}\{X \in A\}$ for all subsets $A \subseteq \mathbb{R}$. However, for most choices of F, the countable additivity property turns out to be violated! However, the sets which do violate them rarely arise in practice and hence we ignore them for the present.

EXERCISE 27. Let X be a random variable with distribution F. Use the working rules to find the following probabilities.

(1) Write $\mathbb{P}\{a < X < b\}$, $\mathbb{P}\{a \leq X < b\}$, $\mathbb{P}\{a \leq X \leq b\}$ in terms of F.

(2) Show that $\mathbb{P}\{X = a\} = F(a) - F(a-)$. In particular, this probability is zero unless F has a jump at a.

We now illustrate how to calculate the probabilities of rather non-trivial sets in a special case. It is not always possible to get an explicit answer as here.

EXAMPLE 28. Let F be the CDF defined in example 20. We calculate $\mathbb{P}\{X \in A\}$ for two sets A.

1. $A = \mathbb{Q} \cap [0,1]$. Since A is countable, we may write $A = \cup_n \{r_n\}$ and hence $A \subseteq \cup_n I_n$ where $I_n = (r_n, r_n + \delta 2^{-n}]$ for any fixed $\delta > 0$. Hence $\mathbb{P}\{X \in A\} \leq \sum_n F(r_n + \delta 2^{-n}) - F(r_n) \leq 2\delta$. Since this is true for every $\delta > 0$, we must have $\mathbb{P}\{X \in A\} = 0$. (We stuck to the letter of the recipe described earlier. It would have been simpler to say that any countable set is a countable union of singletons, and by the countable additivity of probability, must have probability zero. Here we used the fact that singletons have zero probability since F is continuous).

2. A = Cantor's set[2] How to find $\mathbb{P}\{X \in A\}$? Let A_n be the set of all $x \in [0,1]$ which do not have 1 in the first n digits of their ternary expansion. Then $A \subseteq A_n$. Further, it is not hard to see that $A_n = I_1 \cup I_2 \cup \ldots \cup I_{2^n}$ where each of the intervals I_j has length equal to 3^{-n}. Therefore, $\mathbb{P}\{X \in A\} \leq \mathbb{P}\{X \in A_n\} = 2^n 3^{-n}$ which goes to 0 as $n \to \infty$. Hence, $\mathbb{P}\{X \in A\} = 0$.

2. Examples of continuous distributions

Cumulative distributions will also be referred to as simply distribution functions or distributions. We start by giving two large classes of CDFs. There are CDFs that do not belong to either of these classes, but they may be safely ignored for the purposes of this course.

(1) (CDFs with pmf). Let f be a pmf, i.e., let $t_1, t_2, \ldots$ be a countable subset of reals and let $f(t_i)$ be non-negative numbers such that $\sum_i f(t_i) = 1$. Then, define $F : \mathbb{R} \to \mathbb{R}$ by

$$F(t) := \sum_{i: t_i \leq t} f(t_i).$$

Then, F is a CDF. Indeed, we have seen that it is the CDF of a discrete random variable. A special feature of this CDF is that it increases only in jumps (in more precise language, if F is continuous on an interval $[s,t]$, then $F(s) = F(t)$).

(2) (CDFs with pdf). Let $f : \mathbb{R} \to \mathbb{R}_+$ be a function (convenient to assume that it is a piece-wise continuous function) such that $\int_{-\infty}^{+\infty} f(u)du = 1$. Such a function is called a *probability density function* or pdf for short. Then, define $F : \mathbb{R} \to \mathbb{R}$ by

$$F(t) := \int_{-\infty}^{t} f(u)du.$$

[2]To define the Cantor set, recall that any $x \in [0,1]$ may be written in ternary expansion as $x = 0.u_1u_2\ldots := \sum_{n=1}^{\infty} u_n 3^{-n}$ where $u_n \in \{0,1,2\}$. This expansion is unique except if x is a rational number of the form $p/3^m$ for some integers p, m (these are called triadic rationals). For triadic rationals, there are two possible ternary expansions, a terminating one and a non-terminating one (for example, $x = 1/3$ can be written as $0.100\ldots$ or as $0.0222\ldots$). For definiteness, for triadic rationals we shall always take the non-terminating ternary expansion. With this preparation, the Cantor set is defined as the set of all x which do not have the digit 1 in their ternary expansion.

Again, F is a CDF. Indeed, it is clear that F has the increasing property (if $t > s$, then $F(t) - F(s) = \int_s^t f(u)du$ which is non-negative because $f(u)$ is non-negative for all u), and its limits at $\pm\infty$ are as they should be (why?). As for right-continuity, F is in-fact continuous. Actually F is differentiable except at points where f is discontinuous and $F'(t) = f(t)$.

REMARK 29. We understand the pmf. For example if X has pmf f, then $f(t_i)$ is just the probability that X takes the value t_i. How to interpret the pdf? If X has pdf f, then as we already remarked, the CDF is continuous and hence $\mathbb{P}\{X = t\} = 0$. Therefore $f(t)$ cannot be interpreted as $\mathbb{P}\{X = t\}$ (in fact, pdf can take values greater than 1, so it cannot be a probability!).

To interpret $f(a)$, take a small positive number δ and look at

$$F(a + \delta) - F(a) = \int_a^{a+\delta} f(u)du \approx \delta f(a).$$

In other words, $f(a)$ measures the chance of the random variable taking values near a. Higher the pdf, greater the chance of taking values near that point.

Among distributions with pmf, we have seen the Binomial, Poisson, Geometric and Hypergeometric families of distributions. Now we give many important examples of distributions (CDFs) with densities.

EXAMPLE 30. **Uniform distribution on the interval $[a, b]$:** Denoted Unif($[a, b]$) where $a < b$ is the distribution with density and distribution given by

$$\text{PDF: } f(t) = \begin{cases} \frac{1}{b-a} & \text{if } t \in (a, b) \\ 0 & \text{otherwise} \end{cases} \qquad \text{CDF: } F(t) = \begin{cases} 0 & \text{if } t \le a \\ \frac{t-a}{b-a} & \text{if } t \in (a, b) \\ 1 & \text{if } t \ge b. \end{cases}$$

EXAMPLE 31. **Exponential distribution with parameter λ:** Denoted Exp(λ) where $\lambda > 0$ is the distribution with density and distribution given by

$$\text{PDF: } f(t) = \begin{cases} \lambda e^{-\lambda t} & \text{if } t > 0 \\ 0 & \text{otherwise} \end{cases} \qquad \text{CDF: } F(t) = \begin{cases} 0 & \text{if } t \le 0 \\ 1 - e^{-\lambda t} & \text{if } t > 0. \end{cases}$$

EXAMPLE 32. **Normal distribution with parameters μ, σ^2:** Denoted N(μ, σ^2) where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ is the distribution with density and distribution given by

$$\text{PDF: } \varphi_{\mu,\sigma^2}(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(t-\mu)^2} \qquad \text{CDF: } \Phi_{\mu,\sigma^2}(t) = \int_{-\infty}^{t} \varphi_{\mu,\sigma^2}(u)du.$$

There is no closed form expression for the CDF. It is standard notation to write φ and Φ to denote the normal density and CDF when $\mu = 0$ and $\sigma^2 = 1$. $N(0,1)$ is called the standard normal distribution. By a change of variable one can check that $\Phi_{\mu,\sigma^2}(t) = \Phi(\frac{t-\mu}{\sigma})$.

We said that the normal CDF has no simple expression, but is it even clear that it is a CDF?! In other words, is the proposed density a true pdf? Clearly $\varphi(t) = \frac{1}{\sqrt{2\pi}}e^{-t^2/2}$ is non-negative. We need to check that its integral is 1.

LEMMA 33. *Fix $\mu \in \mathbb{R}$ and $\sigma > 0$ and let $\varphi(t) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2\sigma^2}(t-\mu)^2}$. Then, $\int_{-\infty}^{\infty} \varphi(t)dt = 1$.*

PROOF. It suffices to check the case $\mu = 0$ and $\sigma^2 = 1$ (why?). To find its integral is quite non-trivial. Let $I = \int_{-\infty}^{\infty} \varphi(t)dt$. We introduce the two-variable function $h(t,s) := \varphi(t)\varphi(s) = (2\pi)^{-1}e^{-(t^2+s^2)/2}$. On the one hand,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(t,s)dtds = \left(\int_{-\infty}^{+\infty} \varphi(t)dt\right)\left(\int_{-\infty}^{+\infty} \varphi(s)ds\right) = I^2.$$

On the other hand, using polar co-ordinates $t = r\cos\theta$, $s = r\sin\theta$, we see that

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(t,s)dtds = \int_0^{\infty}\int_0^{2\pi} (2\pi)^{-1}e^{-r^2/2}rd\theta dr = \int_0^{\infty} re^{-r^2/2}dr = 1$$

since $\frac{d}{dr}e^{-r^2/2} = -re^{-r^2/2}$. Thus $I^2 = 1$ and hence $I = 1$. ∎

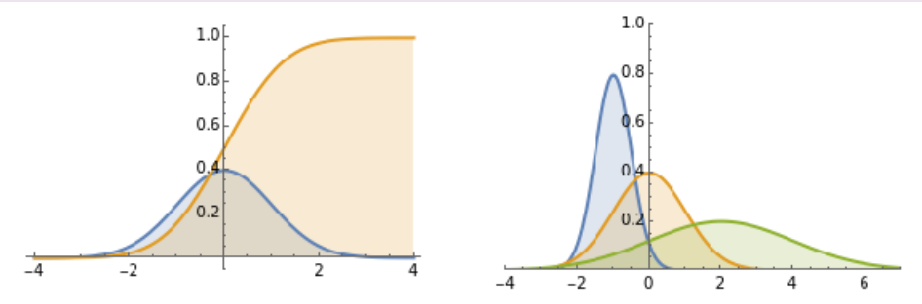

FIGURE 2. PDF and CDF of $N(0,1)$ on the left. PDF of $N(-1,1/2)$, $N(0,1)$ and $N(2,2)$ on the right.

EXAMPLE 34. **Gamma distribution with shape parameter ν and scaler parameter λ:**, where $\nu > 0$ and $\lambda > 0$, denoted Gamma(ν,λ) is the distribution with density and distribution given by -

$$\text{PDF: } f(t) = \begin{cases} \frac{1}{\Gamma(\nu)}\lambda^{\nu}t^{\nu-1}e^{-\lambda t} & \text{if } t > 0 \\ 0 & \text{otherwise} \end{cases} \qquad \text{CDF: } F(t) = \begin{cases} 0 & \text{if } t \le 0 \\ \int_0^t f(u)du & \text{if } t > 0. \end{cases}$$

Here $\Gamma(\nu) := \int_0^\infty t^{\nu-1}e^{-t}dt$. Firstly, f is a density, that is, that it integrates to 1. To see this, make the change of variable $\lambda t = u$ to see that

$$\int_0^\infty \lambda^\nu e^{-\lambda t}t^{\nu-1}dt = \int_0^\infty e^{-u}u^{\nu-1}dv = \Gamma(\nu).$$

Thus, $\int_0^\infty f(t)dt = 1$.

When $\nu = 1$, we get back the exponential distribution. Thus, the Gamma family subsumes the exponential distributions. For positive integer values of ν, one can actually write an expression for the CDF of Gamma(ν, λ) as (this is a homework problem)

$$F_{\nu,\lambda}(t) = 1 - e^{-\lambda t}\sum_{k=0}^{\nu-1}\frac{(\lambda t)^k}{k!}.$$

Once the expression is given, it is easy to check it by induction (and integration by parts). A curious observation is that the right hand side is exactly $\mathbb{P}(N \geq \nu)$ where $N \sim$ Pois(λt). This is in fact indicating a deep connection between Poisson distribution and the Gamma distributions. The function $\Gamma(\nu)$, also known as Euler's Gamma function, is an interesting and important one and occurs all over mathematics. [3]

---

[3] **The Gamma function:** The function $\Gamma : (0, \infty) \to \mathbb{R}$ defined by $\Gamma(\nu) = \int_0^\infty e^{-t}t^{\nu-1}dt$ is a very important function that often occurs in mathematics and physics. There is no simpler expression for it, although one can find it explicitly for special values of ν. One of its most important properties is that $\Gamma(\nu+1) = \nu\Gamma(\nu)$. To see this, consider

$$\Gamma(\nu+1) = \int_0^\infty e^{-t}t^\nu dt = -e^{-t}t^\nu\Big|_0^\infty + \nu\int_0^\infty e^{-t}t^{\nu-1}dt = \nu\Gamma(\nu).$$

Starting with $\Gamma(1) = 1$ (direct computation) and using the above relationship repeatedly one sees that $\Gamma(\nu) = (\nu-1)!$ for positive integer values of ν. Thus, the Gamma function interpolates the factorial function (which is defined only for positive integers). Can we compute it for any other ν? The answer is yes, but only for special values of ν. For example,

$$\Gamma(1/2) = \int_0^\infty x^{-1/2}e^{-x}dx = \sqrt{2}\int_0^\infty e^{-y^2/2}dy$$

by substituting $x = y^2/2$. The last integral was computed above in the context of the normal distribution and equal to $\sqrt{\pi/2}$. Hence we get $\Gamma(1/2) = \sqrt{\pi}$. From this, using again the relation $\Gamma(\nu+1) = \nu\Gamma(\nu)$, we can compute $\Gamma(3/2) = \frac{1}{2}\sqrt{\pi}$, $\Gamma(5/2) = \frac{3}{4}\sqrt{\pi}$, etc. Yet another useful fact about the Gamma function is its asymptotics as $\nu \to \infty$.

Stirling's approximation: $\frac{\Gamma(\nu+1)}{\nu^{\nu+\frac{1}{2}}e^{-\nu}\sqrt{2\pi}} \to 1$ as $\nu \to \infty$.

A small digression: It was Euler's idea to observe that $n! = \int_0^\infty x^n e^{-x}dx$ and that on the right side n could be replaced by any real number greater than -1. But this was his second approach to defining the Gamma function. His first approach was as follows. Fix a positive integer n. Then for any $\ell \geq 1$ (also a positive integer), we may write

$$n! = \frac{(n+\ell)!}{(n+1)(n+2)\dots(n+\ell)} = \frac{\ell!(\ell+1)\dots(\ell+n)}{(n+1)\dots(n+\ell)} = \frac{\ell!\,\ell^n}{(n+1)\dots(n+\ell)}\cdot\frac{(\ell+1)\dots(\ell+n)}{\ell^n}$$

The second factor approaches 1 as $\ell \to \infty$. Hence,

$$n! = \lim_{\mathbb{N}\ni\ell\to\infty}\frac{\ell!\,\ell^n}{(n+1)\dots(n+\ell)}.$$

Euler then showed (by a rather simple argument that we skip) that the limit on the right exists if we replace n by any complex number other than $\{-1, -2, -3, \dots\}$ (negative integers are a problem as they make the denominator zero). Thus, he extended the factorial function to all complex numbers except negative integers! It is a fun exercise

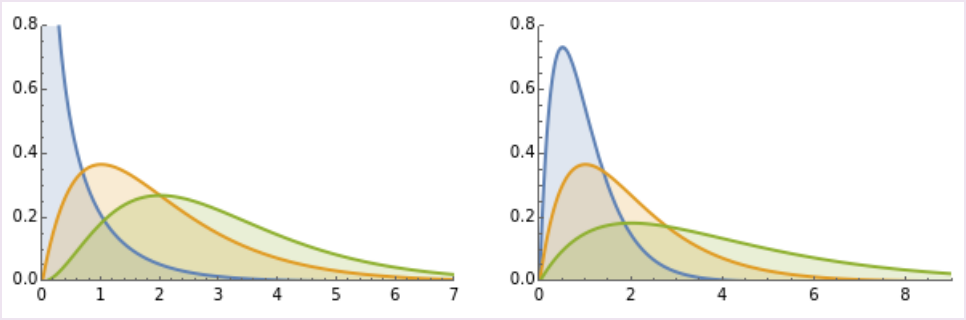

FIGURE 3. PDF of Gamma$(\nu, 1)$ for $\nu = \frac{1}{2}, 2, 3$ on the left. PDF of Gamma$(2, \lambda)$ for $\lambda = \frac{1}{2}, 1, 2$ on the right.

EXAMPLE 35. **Beta distributions:** Let $\alpha, \beta > 0$. The Beta distribution with parameters α, β, denoted Beta(α, β), is the distribution with density and distribution given by -

$$\text{PDF: } f(t) = \begin{cases} \frac{1}{B(\alpha,\beta)} t^{\alpha-1}(1-t)^{\beta-1} & \text{if } t \in (0,1) \\ 0 & \text{otherwise} \end{cases} \qquad \text{CDF: } F(t) = \begin{cases} 0 & \text{if } t \leq 0 \\ \int_0^t f(u)du & \text{if } t \in (0,1) \\ 1 & \text{if } t \geq 1. \end{cases}$$

Here $B(\alpha, \beta) := \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt$. Again, for special values of α, β (eg., positive integers), one can find the value of $B(\alpha, \beta)$, but in general there is no simple expression. However, it can be expressed in terms of the Gamma function!

PROPOSITION 36. *For any $\alpha, \beta > 0$, we have $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$.*

PROOF. For $\beta = 1$ we see that $B(\alpha, 1) = \int_0^1 t^{\alpha-1} = \frac{1}{\alpha}$ which is also equal to $\frac{\Gamma(\alpha)\Gamma(1)}{\Gamma(\alpha+1)}$ as required. Similarly (or by the symmetry relation $B(\alpha, \beta) = B(\beta, \alpha)$), we see that $B(1, \beta)$ also has the desired expression.

Now for any other *positive integer* value of α and real $\beta > 0$ we can integrate by parts and get

$$\begin{aligned} B(\alpha,\beta) &= \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt \\ &= -\frac{1}{\beta}t^{\alpha-1}(1-t)^{\beta}\Big|_0^1 + \frac{\alpha-1}{\beta}\int_0^1 t^{\alpha-2}(1-t)^{\beta}dt \\ &= \frac{\alpha-1}{\beta}B(\alpha-1,\beta+1). \end{aligned}$$

---

to check that this agrees with the definition by the integral given earlier. In other words, for $\nu > -1$, we have

$$\lim_{\mathbb{N}\ni\ell\to\infty} \frac{\ell!\,\ell^{\nu}}{(\nu+1)\ldots(\nu+\ell)} = \Gamma(\nu+1) = \int_0^{\infty} x^{\nu}e^{-x}dx.$$

Note that the first term vanishes because $\alpha > 1$ and $\beta > 0$. When α is an integer, we repeat this for α times and get

$$B(\alpha,\beta) = \frac{(\alpha-1)(\alpha-2)\dots 1}{\beta(\beta+1)\dots(\beta+\alpha-2)} B(1,\beta+\alpha-1).$$

But we already checked that $B(1,\beta+\alpha-1) = \frac{\Gamma(1)\Gamma(\alpha+\beta-1)}{\Gamma(\alpha+\beta)}$ from which we get

$$B(\alpha,\beta) = \frac{(\alpha-1)(\alpha-2)\dots 1}{\beta(\beta+1)\dots(\beta+\alpha-2)} \frac{\Gamma(1)\Gamma(\alpha+\beta-1)}{\Gamma(\alpha+\beta)} = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

by the recursion property of the Gamma function. Thus we have proved the proposition when α is a positive integer. By symmetry the same is true when β is a positive integer (and α can take any value). We do not bother to prove the proposition for general $\alpha,\beta > 0$ here. ■

EXAMPLE 37. **The standard Cauchy distribution:** is the distribution with density and distribution given by

$$\text{PDF: } f(t) = \frac{1}{\pi(1+t^2)} \qquad \text{CDF: } F(t) = \frac{1}{2} + \frac{1}{\pi}\tan^{-1} t.$$

One can also make a parametric family of Cauchy distributions with parameters $\lambda > 0$ and $a \in \mathbb{R}$ denoted Cauchy(a,λ) and having density and CDF

$$f(t) = \frac{\lambda}{\pi(\lambda^2+(t-a)^2)} \qquad F(t) = \frac{1}{2} + \frac{1}{\pi}\tan^{-1}\left(\frac{t-a}{\lambda}\right).$$

While the density function of the standard Cauchy looks qualitatively like that of the standard Gaussian, it decays much slower. That is, for large positive x, the Cauchy density looks like $\frac{1}{x^2}$ while the Gaussian density is more like $e^{-x^2/2}$ (omitting constant factors). The Cauchy distribution is said to be *heavy-tailed*, while the Gaussian is *light-tailed*. The kinds of situations modeled by Cauchy are quite different from those modeled by the Gaussian. In the Cauchy distribution, unusually large positive or negative values are not that uncommon.

EXAMPLE 38. **Pareto distributions:** The Pareto distribution with parameters (α, t) where $\alpha > 0$ and $t > 0$ is the CDF

$$F(x) = \begin{cases} 1 - \frac{t^\alpha}{x^\alpha} & \text{if } x \geq t, \\ 0 & \text{if } x \leq t. \end{cases}$$

It is clearly a valid CDF. Differentiating, we get the pdf $f(x) = \alpha t^\alpha x^{-\alpha-1}$ for $x \geq t$ (and $f(x) = 0$ for $x < t$). The *tail exponent* α dictates the tail decay of the density while the *scale parameter* t just gives a change of scale.

As the density decays only polynomially as $x \to \infty$, Pareto distributions are also heavy tailed, like the Cauchy distribution. They occur in many situations, for example in income distributions (so t is the minimum income, and $1 - F(x)$ is the proportion of people with

income more than x). Unlike in the light-tailed distributions, it is not uncommon to have individually with incomes that are phenomenally larger than the average.

REMARK 39. Does every CDF come from a pdf? Not necessarily. For example any CDF that is not continuous (for example, CDFs of discrete distributions such as Binomial, Poisson, Geometric etc.). In fact even continuous CDFs may not have densities (there is a good example manufactured out of the 1/3-Cantor set, but that would take us out of the topic now). However, suppose F is a *continuous* CDF and suppose F is differentiable except at finitely many points and that the derivative is a continuous function. Then $f(t) := F'(t)$ defines a pdf which by the fundamental theorm of Calculus satisfies $F(t) = \int_{-\infty}^{t} f(u)du$.

3. Simulation

As we have emphasized, probability is applicable to many situations in the real world. As such one may conduct experiments to verify the extent to which theorems are actually valid. For this we need to be able to draw numbers at random from any given distribution.

For example, take the case of Bernoulli(1/2) distribution. One experiment that can give this is that of physically tossing a coin. This is not entirely satisfactory for several reasons. Firstly, are real coins fair? Secondly, what if we change slightly and want to generate from Ber(0.45)? In this section, we describe how to draw random numbers from various distributions on a computer. We do not fully answer this question. Instead what we shall show is
If one can generate random numbers from Unif($[0,1]$) distribution, then one can draw random numbers from any other distribution. More precisely, suppose U is a random variable with Unif($[0,1]$) distribution. We want to simulate random numbers from a given distribution F. Then, we shall find a function $\psi : [0,1] \to \mathbb{R}$ so that the random variable $X := \psi(U)$ has the given distribution F.

The question of how to draw random numbers from Unif($[0,1]$) distribution is a very difficult one and we shall just make a few superficial remarks about that.

Drawing random numbers from a discrete pmf: First start with an example.

EXAMPLE 40. Suppose we want to draw random numbers from Ber(0.4) distribution. Let $\psi : [0,1] \to \mathbb{R}$ be defined as $\psi(t) = \mathbf{1}_{t \le 0.4}$. Let $X = \psi(U)$, i.e., $X = 1$ if $U \le 0.4$ and $X = 0$ otherwise. Then

$$\mathbb{P}\{X = 1\} = \mathbb{P}\{U \le 0.4\} = 0.4, \qquad \mathbb{P}\{X = 0\} = \mathbb{P}\{U > 0.4\} = 0.6.$$

Thus, X has Ber(0.4) distribution.

It is clear how to generalize this.

General rule: Suppose we are given a pmf f

$$\begin{pmatrix} t_1 & t_2 & t_3 & \dots \\ f(t_1) & f(t_2) & f(t_3) & \dots \end{pmatrix}.$$

Then, define $\psi : [0,1] \to \mathbb{R}$ as

$$\psi(u) = \begin{cases} t_1 & \text{if } u \in [0, f(t_1)] \\ t_2 & \text{if } u \in (f(t_1), f(t_1) + f(t_2)] \\ t_3 & \text{if } u \in (f(t_1) + f(t_2), f(t_1) + f(t_2) + f(t_3)] \\ \vdots & \vdots \end{cases}.$$

Then define $X = f(U)$. Clearly X takes the values $t_1, t_2, \ldots$ and

$$\mathbb{P}\{X = t_k\} = \mathbb{P}\left\{ \sum_{j=1}^{k-1} f(t_j) < U \leq \sum_{j=1}^{k} f(t_j) \right\} = f(t_k).$$

Thus X has pmf f.

Exercise 41. Draw 100 random numbers from each of the following distributions and draw the histograms. Compare with the pmf.

(1) Bin(n,p) for $n = 10, 20, 40$ and $p = 0.5, 0.3, 0.9$.

(2) Geo(p) for $p = 0.9, 0.5, 0.3$.

(3) Pois(λ) with $\lambda = 1, 4, 10$.

(4) Hypergeo(N_1, N_2, m) with $N_1 = 100, N_2 = 50, m = 20$, $N_1 = 1000, N_2 = 1000, m = 40$.

Drawing random numbers from a pdf: Clearly the procedure used for generating from a pmf is inapplicable here. First start with two examples. As before U is a Unif$([0,1])$ random variable.

Example 42. Suppose we want to draw from the Unif$([3,7])$ distribution. Set $X = 4U + 3$. Clearly

$$\mathbb{P}\{X \leq t\} = \mathbb{P}\{U \leq \frac{t-3}{4}\} = \begin{cases} 0 & \text{if } t < 0 \\ (t-3)/4 & \text{if } 3 \leq t \leq 7 \\ 1 & \text{if } t > 7 \end{cases}.$$

This is precisely the CDF of Unif$([3,7])$ distribution.

Example 43. Here let us do the opposite, just take some function of a uniform variable and see what CDF we get. Let $\psi(t) = t^3$ and let $X = \varphi(U) = U^3$. Then,

$$F(t) := \mathbb{P}\{X \leq t\} = \mathbb{P}\{U \leq t^{1/3}\} = \begin{cases} 0 & \text{if } t < 0 \\ t^{1/3} & \text{if } 0 \leq t \leq 1 \\ 1 & \text{if } t > 1 \end{cases}.$$

Differentiating the CDF, we get the density

$$f(t) = F'(t) = \begin{cases} \frac{1}{3}t^{-2/3} & \text{if } 0 < t < 1 \\ 0 & \text{otherwise.} \end{cases}$$

The derivative does not exist at 0 and 1, but as remarked earlier, it does not matter if we change the value of the density at finitely many points (as the integral over any interval will remain the same). Anyway, we notice that the density is that of Beta(1/3, 1). Hence $X \sim \text{Beta}(1/3, 1)$.

This gives us the idea that to generate random number from a CDF F, we should find a function $\psi : [0,1] \to \mathbb{R}$ such that $X := \psi(U)$ has the distribution F. How to find the distribution of X?

LEMMA 44. *Let $\psi : (0,1) \to \mathbb{R}$ be a strictly increasing function with $a = \psi(0+)$ and $b = \psi(1-)$. Let $X = \psi(U)$. Then X has CDF*

$$F(t) = \begin{cases} 0 & \text{if } t \le a \\ \psi^{-1}(t) & \text{if } a < t < b \\ 1 & \text{if } t \ge b. \end{cases}$$

If is ψ also differentiable and the derivative does not vanish anywhere (or vanishes at finitely many points only), then X has pdf

$$f(t) = \begin{cases} \left(\psi^{-1}\right)'(t) & \text{if } a < t < b \\ 0 & \text{if } t \notin (a,b). \end{cases}$$

PROOF. Since ψ is strictly increasing, $\psi(u) \le t$ if and only if $u \le \psi^{-1}(t)$. Hence,

$$F(t) = \mathbb{P}\{X \le t\} = \mathbb{P}\{U \le \psi^{-1}(t)\} = \begin{cases} 0 & \text{if } t \le a \\ \psi^{-1}(t) & \text{if } a < t < b \\ 1 & \text{if } t \ge b. \end{cases}$$

If ψ is differentiable at and $\psi(u) \neq 0$, then ψ^{-1} is differentiable at $t = \psi(u)$ (and indeed, $(\psi^{-1})'(t) = \frac{1}{\psi'(u)}$). Thus we get the formula for the density. ■

From this lemma, we immediately get the following rule for generating random numbers from a density.

How to simulate from a CDF: Let F be a CDF that is strictly increasing on an interval $[A, B]$ where $F(A) = 0$ and $F(B) = 1$ (it is allowed to take $A = -\infty$ and/or $B = +\infty$). Then define $\psi : (0,1) \to (A,B)$ as $\psi(u) = F^{-1}(u)$. Let $U \sim \text{Unif}([0,1])$ and let $X = \psi(U)$. Then X has CDF equal to F.

This follows from the lemma because ψ is define as the inverse of F and hence F (restricted to (A,B)) is the inverse of ψ. Further, as the inverse of a strictly increasing function, the function ψ is also strictly increasing.

EXAMPLE 45. Consider the Exponential distribution with parameter λ whose CDF is

$$F(t) = \begin{cases} 0 & \text{if } t \leq 0 \\ 1 - e^{-\lambda t} \text{ if } t > 0 \end{cases}$$

Take $A = 0$ and $B = +\infty$. Then F is increasing on $(0, \infty)$ and its inverse is the function $\psi(u) = -\frac{1}{\lambda}\log(1-u)$. Thus to simulate a random number from Exp(λ) distribution, we set $X = -\frac{1}{\lambda}\log(1-U)$.

When the CDF is not explicitly available as a function we can still adopt the above procedure but only numerically. Consider an example.

EXAMPLE 46. Suppose $F = \Phi$, the CDF of $N(0,1)$ distribution. Then we do not have an explicit form for either Φ or for its inverse Φ^{-1}. With a computer we can do the following. Pick a large number of closely placed points, for example divide the interval $[-5,5]$ into 1000 equal intervals of length 0.01 each. Let the endpoints of these intervals be labelled $t_0 < t_1 < \ldots < t_{1000}$. For each i, calculate $\Phi(t_i) = \int_{-\infty}^{t_i} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$ using numerical methods for integration, say the numerical value obtained is w_i. This is done only once and create the table of values

$$\begin{array}{cccccc} t_0 & t_1 & t_2 & \ldots & \ldots & t_{1000} \\ w_0 & w_1 & w_2 & \ldots & \ldots & w_{1000} \end{array}.$$

Now draw a uniform random number U. Look up the table and find the value of i for which $w_i < U < w_{i+1}$. Then set $X = t_i$. If it so happens that $U < w_0$, set $X = t_0 = -5$ and if $U > w_{1000}$ set $X = t_{1000} = 5$. But since $\Phi(-5) < 0.00001$ and $\Phi(5) > 0.99999$, it is highly unlikely that the last two cases will occur. The random variable X has a distribution close to $N(0,1)$.

EXERCISE 47. Give an explicit method to draw random numbers from the following densities.

(1) Cauchy distribution with density $\frac{1}{\pi(1+x^2)}$.

(2) Beta($\frac{1}{2}, \frac{1}{2}$) density $\frac{1}{\pi}\frac{1}{\sqrt{x(1-x)}}$ on $[0,1]$ (and zero elsewhere).

(3) Pareto(α) distribution which by definition has the density

$$f(t) = \begin{cases} \alpha t^{-\alpha-1} & \text{if } t \geq 1, \\ 0 & \text{if } t < 1. \end{cases}$$

We have described a general principle. When we do more computations with random variables and understand the relationships between different distributions, better tricks can be found. For example, we shall see later that we can generate two $N(0,1)$ random numbers as follows: Pick two uniform random numbers U, V and set $X = \sqrt{-2\log(1-U)}\cos(2\pi V)$ and

$Y = \sqrt{-2\log(1-U)}\sin(2\pi V)$. Then it turns out that X and Y have exactly $N(0,1)$ distribution! As another example, suppose we need to generate from Gamma$(3,1)$ distribution, we can first generate three uniforms U_1, U_2, U_3 and set $\xi_i = -\log(1-U_i)$ (so ξ_i have exponential distribution) and then define $X = \xi_1 + \xi_2 + \xi_3$. It turns out that X has Gamma$(3,1)$ distribution!

REMARK 48. We have conveniently skipped the question of how to draw random numbers from uniform distribution in the first place. This is a difficult topic and various results, proved and unproved, are used in generating such numbers.

4. Exercises

PROBLEM 1. Let X be a random variable with distribution (CDF) F and density f.

(1) Find the distribution and density of the random variable $2X$.

(2) Find the distribution and density of the random variable $X + 5$.

(3) Find the distribution and density of the random variable $-X$.

(4) Find the distribution and density of the random variable $1/X$.

PROBLEM 2. Let X be a random variable with Gamma(ν, λ) distribution. Let F be the CDF of X. When ν is a positive integer, show that for $t \geq 0$,

$$F(t) = 1 - e^{-\lambda t}\sum_{k=0}^{\nu-1}\frac{\lambda^k t^k}{k!}.$$

[**Note:** Observe that this quantity is the same as $\mathbb{P}(N \geq \nu)$ where N is a Poisson random variable with parameter λt. There is a connection here but we cannot discuss it now].

PROBLEM 3. Give explicit description of how you would simulate random variables from the following distributions.

(1) The standard Cauchy distribution with density $f(x) = \frac{1}{\pi(1+x^2)}$ for $x \in \mathbb{R}$.

(2) The Beta$(1/2, 1/2)$ distribution with density $\frac{1}{\pi\sqrt{x(1-x)}}$.

(3) (Do not need to submit this) Draw 100 random numbers from either of these densities (on MATLAB or any other program that gives uniform random numbers) using the above procedure and draw the histograms. Compare the histograms to the plot of the densities.

PROBLEM 4. In each of the following situations, the distribution of the random variable X is given. Find the distributionof Y (it is enough to find the density of Y).

(1) $X \sim$ Unif$[0,1]$ and $Y = \sin^{-1}(X)$.

(2) $X \sim$ Unif$[0,1]$ and $Y = \cos^{-1}(X)$.

(3) $X \sim N(0,1)$ and $Y = X^2$.

[**Note:** We define $\sin^{-1}$ to take values in $[-\pi/2, \pi/2]$ and $\cos^{-1}$ to take values in $[0, \pi]$. In the third part, observe that $f(x) = x^2$ is not a one-one function, so the formula given in the notes does not apply directly].

PROBLEM 5. (1) Let $f(k) = \frac{1}{k(k+1)}$ for integer $k \geq 1$. Show that f is a pmf and find the corresponding CDF.

(2) Let $\alpha > 0$ and set $F(x) = 1 - \frac{1}{x^\alpha}$ for $x \geq 1$ and $F(x) = 0$ for $x < 1$. Show that F is a CDF and find the corresponding density function. (This is known as the *Pareto* distribution).

PROBLEM 6. Give explicit description of how you would simulate random variables from the following distributions.

(1) The standard Cauchy distribution with density $f(x) = \frac{1}{\pi(1+x^2)}$ for $x \in \mathbb{R}$.

(2) The Beta(1/2, 1/2) distribution with density $\frac{1}{\pi\sqrt{x(1-x)}}$.

(3) (Do not need to submit this) Draw 100 random numbers from either of these densities (on MATLAB or any other program that gives uniform random numbers) using the above procedure and draw the histograms. Compare the histograms to the plot of the densities.

PROBLEM 7. Let X be a random variable with distribution function F. Let $a > 0$ and $b \in \mathbb{R}$ and define $Y = aX + b$.

(1) What is the CDF of Y?

(2) If X has a density f, find the density of Y.

PROBLEM 8. (1) Let $X \sim \text{Exp}(\lambda)$. Fix $s, t > 0$ and compute the conditional probability of the event $X > t + s$ given that $X > s$.

(2) Let ν be a positive integer. Show that the CDF of Gamma(ν, λ) distribution is given by

$$F(x) = 1 - e^{-\lambda x} \sum_{k=0}^{\nu-1} \frac{\lambda^k}{k!} x^k.$$

PROBLEM 9. Let $U \sim \text{Uniform}[0, 1]$. Find the density and distribution functions of (a) U^p (where $p > 0$), (b) $U/(1-U)$, (c) $\log(1/U)$, (d) $\frac{2}{\pi}\arcsin(U)$.

PROBLEM 10. Let $X \sim N(0, 1)$. Find the density of (a) $aX + b$ (where $a, b \in \mathbb{R}$), (b) X^2, (c) X^3, (d) e^X.

PROBLEM 11. If F is a CDF, show that it can have at most countably many discontinuity points.

PROBLEM 12. In these problems, use change of variable formula in one dimension to show that the families of distributions we have defined

(1) If $X \sim \text{Exp}(\lambda)$, show that $\lambda X \sim \text{Exp}(1)$. More generally, if $X \sim \text{Gamma}(\nu, \lambda)$, show that $\lambda X \sim \text{Gamma}(\nu, 1)$.

(2) If $X \sim N(\mu, \sigma^2)$, show that $\frac{X-\mu}{\sigma} \sim N(0, 1)$.

CHAPTER 6

Summary measures of univariate distributions

1. Expectation or mean

Let X be a random variable with distribution F. We wish to define the expected value of X. We have already done so for discrete random variables, and in fact in that case we have three ways to express its expected value (when it exists).

(1) $\mathbb{E}[X] = \sum_{\omega\in\Omega} X(\omega)p_\omega$. This was our definition of expectation.

(2) $\mathbb{E}[X] = \sum_t tf(t)$, where f is the pmf of X (the sum is a countable sum, write it as $\sum_i t_i f(t_i)$ where Range$(X) = \{t_i\}$, if you prefer). We showed this.

(3) One can show that $\mathbb{E}[X] = \int_0^\infty (1 - F(t))dt - \int_0^\infty F(-t)dt$. We leave this as an exercise to check (for ease, take X to be integer-valued).

If X is not discrete, which of these can we adopt as definition? The first is not possible, as we have not developed the notion of uncountable probability spaces (this was discussed in unwanted detail earlier). The second can be adapted to the case of random variables with density f in the most obvious way:

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} tf(t)dt \tag{8}$$

provided the integral converges absolutely (i.e., if $\int_{-\infty}^{+\infty} |t|f(t)dt < \infty$). The third definition is even more general, in that it does not require the existence of density, and in fact it can treat all distributions in a unified manner. But in this course, as we shall only work with random variables having pmf or pdf, we simply adopt the middle path and take (8) as the definition. It is more convenient to work with in standard examples.

Justifying (8): We reasoned by analogy and replaced the pmf by pdf and sum by integral. That is a reliable approach, but one may prefer a less formal[1] route to it. Recall that a random variable X with a pdf f may be thought of as a continuous approximation to the discrete random variable X_n taking values in $\frac{1}{n}\mathbb{Z}$ with pmf f_n defined by $f_n(k/n) = \int_{k/n}^{(k+1)/n} f(t)dt$

[1] In contrast to formal clothing which suggests a more 'proper' or strict dressing code, in mathematics the word formal means following the form of something rather than the content. For example, manipulating series without bothering about convergence. It is not accepted as a legitimate form of deduction in mathematics, but in expert hands (Euler and Ramanujan are famous examples) it can yield amazing results.

(equivalently, $\mathbb{P}\{X_n = \frac{k}{n}\} = \mathbb{P}\{\frac{k}{n} \le X < \frac{k+1}{n}\}$). As

$$\mathbb{E}[X_n] = \sum_{k\in\mathbb{Z}} \frac{k}{n} f_n(k/n) = \sum_{k\in\mathbb{Z}} \frac{k}{n} \int_{k/n}^{(k+1)/n} f(t)dt \approx \sum_{k\in\mathbb{Z}} \int_{k/n}^{(k+1)/n} tf(t)dt = \int_{-\infty}^{\infty} tf(t)dt,$$

we see that (8) is justified. With a bit more care, one can make the approximation "≈" above precise by showing that $\mathbb{E}[X_n] \to \int_{-\infty}^{\infty} tf(t)dt$ as $n \to \infty$ (whenever $\int |t|f(t)dt < \infty$), thus justifying the definition of $\mathbb{E}[X]$ as the limit. We summarize this in the following definition.

Definition 13. The *expected value* (also called *mean*) of X is defined as the quantity $\mathbb{E}[X] = \sum_t tf(t)$ if f is a pmf and $\mathbb{E}[X] = \int_{-\infty}^{+\infty} tf(t)dt$ if f is a pdf (provided the sum or the integral converges absolutely).

Properties of expectation: Let X, Y be random variables defined on the sample probability space and both having pmf f, g or pdf f, g, respectively. The following are the fundamental properties of expectation.

(1) *Linearity*: Then, $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$ for any $a, b \in \mathbb{R}$. In particular, for a constant random variable (i.e., $X = a$ with probability 1 for some a, $\mathbb{E}[X] = a$). This is called *linearity* of expectation.

(2) *Monotonicity/Positivity*: If $X \ge Y$ (meaning, $X(\omega) \ge Y(\omega)$ for all ω), then $\mathbb{E}[X] \ge \mathbb{E}[Y]$. Strict inequality holds unless $X = Y$ with probability 1.

(3) $\mathbb{E}[\mathbf{1}_A] = \mathbb{P}(A)$ for any even A.

For random variables on a discrete probability space (then they have pmf), we have proved these properties. For random variables with pmf, a proper proof requires definition of uncountable probability spaces, hence cannot be given. Alternatively, try to justify them by discretization (replacing X by X_n and letting $n \to \infty$ as above).

In addition, note the following useful short cuts to computing expectations.

(1) If $\varphi : \mathbb{R} \to \mathbb{R}$, then

$$\mathbb{E}[\varphi(X)] = \begin{cases} \sum_t \varphi(t)f(t) & \text{if } X \text{ has pmf } f. \\ \int_{-\infty}^{+\infty} \varphi(t)f(t)dt & \text{if } X \text{ has pdf } f. \end{cases}$$

(2) More generally, if $(X_1, \dots, X_n)$ has joint pdf $f(t_1, \dots, t_n)$ and $V = T(X_1, \dots, X_n)$ (here $T : \mathbb{R}^n \to \mathbb{R}$), then $\mathbb{E}[V] = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} T(x_1, \dots, x_n)f(x_1, \dots, x_n)dx_1 \dots dx_n$. A similar formula can be written when X_is have a joint pmf.

The convenience here comes from the fact that we don't need to find the pmf/pdf of $\varphi(X)$ or of $T(X_1, \dots, X_n)$. Lastly, we state one more property of expectations, its relationship to independence.

Lemma 14. *Let X, Y be random variables on a common probability space. If X and Y are independent, then $\mathbb{E}[H_1(X)H_2(Y)] = \mathbb{E}[H_1(X)]\mathbb{E}[H_2(Y)]$ for any functions $H_1, H_2 : \mathbb{R} \to \mathbb{R}$ (for which the expectations make sense). In particular, $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.*

PROOF. Independence means that the joint density (analogous statements for pmf omitted) of (X, Y) is f the form $f(t,s) = g(t)h(s)$ where $g(t)$ is the density of X and $h(s)$ is the density of Y. Hence,

$$\mathbb{E}[H_1(X)H_2(Y)] = \iint H_1(t)H_2(s)f(t,s)dtds = \left(\int_{-\infty}^{\infty} H_1(t)g(t)dt\right)\left(\int_{-\infty}^{\infty} H_2(s)g(s)ds\right)$$

which is precisely $\mathbb{E}[H_1(X)]\mathbb{E}[H_2(Y)]$. ■

Expectation gives a one-number-summary of the distribution. It is not the only choice, another is the *median*, defined as any number t such that $F(t-) \leq \frac{1}{2} \leq F(t)$. A simpler way to define it would have been to say that the median is the unique t such that $F(t) = \frac{1}{2}$, but neither the existence nor the uniqueness of such a t is guaranteed in general (it is, if F is strictly increasing a continuous, by the intermediate value theorem). More generally, for any $p \in (0,1)$, we define the p-quantile of X as any number t such $F(t-) \leq p \leq F(t)$. It is often denoted by $Q_p(X)$, and $Q_{\frac{1}{2}}(X)$ is just the median, denoted $\text{Med}(X)$.

REMARK 15. In high school or college, you might have seen the mean, median and quantiles defined for a collection of numbers $x_1, \ldots, x_n$. How do they relate to our definitions now? You did not have the notion of a random variable or its distribution then.

Consider a box with n coupons with the kth coupon carrying the label x_k. Now draw a coupon at random and note the number on the coupon as X. If x_ks are distinct, then this is a random variable whose range is $\{x_1, \ldots, x_n\}$ and it has pmf $f(x_k) = \frac{1}{n}$. If x_ks are not distinct, then we count with multiplicity (if there are exactly three coupons labelled 2.4, then $f(2.4) = \frac{3}{n}$). It is easy to see that the mean, median, quantiles of X according to the definitions we gave exactly coincide with the definitions you have seen before.

Which is better, mean or median?: There are situations where the median is preferred, such as in summarizing the income distribution of a group of people. The attractive property of the median is that it is robust, i.e., its value does not change easily due to outliers. For example, imagine a group of 50 employed youth from middle-income families having a median income of 25K and an average income of 24K. If a rich person with an income of 500K joins this group, the median will not change much (earlier it was the income of the 25th person, now it may be that of the 26th) but the average shoots up to 33K. Clearly the median better summarizes the economic status of this group of people. Using more than one quantile, for example $Q_{0.1}$ and $Q_{0.9}$, we get an idea of the range in which most of the numbers lie.

But the expectation has far better mathematical properties. For one, note how we calculated the new mean when a person was added, whereas to compute the median, we need the full set of numbers, not just the old median and the newcomer's income. This makes the median computation harder. The linearity and positivity properties of expectation make it much mroe amenable mathematically. In contrast $\text{Med}(X+Y)$ cannot be determined from

med(X) and Med(Y). Lastly, in situations where the density or mass function is symmetric about a point, the two coincide.

2. Other quantities associated to distributions

Once we have the notion of expectation, we can define many other quantities of importance and interest. In what follows, X is a random variable and the expectations of various random variables that occur are assumed to exist.

Moments: Let $k \geq 0$ be an integer. If $\mathbb{E}[X^k]$ exists, it is called the k^{th} *moment* of X. Existence means that $\mathbb{E}[|X|^k] < \infty$. While $|X|^k$ is defined for any real number k, note that X^k is in general well-defined only when k is an integer.

Variance: Let $\mu = \mathbb{E}[X]$ and define $\sigma^2 := \mathbb{E}\left[(X-\mu)^2\right]$. This is called the *variance* of X, also denoted by Var(X). It can be written in other forms. For example,

$$\begin{aligned}\sigma^2 &= \mathbb{E}[X^2 + \mu^2 - 2\mu X] \qquad \text{(by expanding the square)}\\ &= \mathbb{E}[X^2] + \mu^2 - 2\mu\mathbb{E}[X] \qquad \text{(by property (1) above)}\\ &= \mathbb{E}[X^2] - \mu^2.\end{aligned}$$

That is $\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$.

Standard deviation: The standard deviation of X is defined as $\text{s.d.}(X) := \sqrt{\text{Var}(X)}$.

Mean absolute deviation: The mean absolute deviation of X is defined as the $\mathbb{E}[|X - \text{med}(X)|]$.

Coefficient of variation: For a positive random variable X, its coefficient of variation is defined as $\text{c.v.}(X) = \frac{\text{s.d.}(X)}{\mathbb{E}[X]}$, which is pure number (dimension free).

Entropy: The entropy of a random variable X is defined as

$$\text{Ent}(X) = \begin{cases} -\sum_i f(t)\log(f(t_i)) & \text{if } X \text{ has pmf } f.\\ -\int f(t)\log(f(t)) & \text{if } X \text{ has pdf} f.\end{cases}$$

If $\mathbf{X} = (X_1, \ldots, X_n)$ is a random vector, we can define its entropy exactly by the same expressions, except that we use the joint pmf or pdf of $\mathbf{X}$ and the sum or integral is over points in $\mathbb{R}^n$.

Discussion: What do these quantities mean? Mean and median try to summarize the distribution of X by a single number. They are called *measures of central tendency*. We have already discussed their relative merits. The ones introduced here measure different features of the distribution.

Measures of dispersion: The variance, the standard deviation and the mean absolute deviation are *measures of dispersion*. They measure how much a distribution is spread out. Suppose the average height of people in a city is 160 cm. This could be because everyone is 160 cm exactly or because half the people are 100 cm. while the other half are 220 cm., or alternately the heights could be uniformly spread over 150-170 cm., etc. To measure spread, one idea would be to fix a number a and consider $\mathbb{E}[|X-a|]$ (of course naively one might think $\mathbb{E}[X-a]$, but that is just $\mathbb{E}[X]-a$ and has no information other than the mean). It turns out (exercise in the homework) that this quantity is minimized when $a = \text{Med}(X)$, and the value for that a is precisely the mean absolute deviation.

But mathematically it is much better to consider $\mathbb{E}[|X-a|^2]$, which is minimized when $a = \mathbb{E}[X]$ and the value is the variance. Why is it better? Naive reason: Absolute value is a pain, square can be expanded to see that $\mathbb{E}[|X-a|^2] = \mathbb{E}[X^2] - 2a\mathbb{E}[X] + a^2$. A more refined reason: Think of it as analogous to how we measure the distance between (p_1, q_1) and (p_2, q_2) by the Pythagorean expression $\sqrt{(p_1-q_1)^2+(p_2-q_2)^2}$ and not by $|p_1-q_1|+|p_2-q_2|$.

The standard deviation has the same units as the quantity. Fo example, if mean height is 160cm measured in centimeters with a standard deviation of 10cm, and the mean weight is 55kg with a standard deviation of 5kg, then we cannot say which of the two is less variable. To make such a comparison we need a dimension free quantity (a pure number). Coefficient of variation is such a quantity, as it measure the standard deviation per mean. For the height and weight data just described, the coefficients of variation are 1/16 and 1/11, respectively. Hence we may say that height is less variable than weight in this example.

EXAMPLE 16. Let $X \sim N(\mu, \sigma^2)$. Recall that its density is $\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. We can compute

$$\mathbb{E}[X] = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} x e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \mu.$$

On the other hand

$$\begin{aligned}
\text{Var}(X) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} (x-\mu)^2 e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx &= \sigma^2 \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} u^2 e^{-\frac{u^2}{2}} du \qquad \text{(substitute } x = \mu + \sigma u\text{)} \\
&= \sigma^2 \frac{2}{\sqrt{2\pi}} \int_{0}^{+\infty} u^2 e^{-\frac{u^2}{2}} du = \sigma^2 \frac{2\sqrt{2}}{\sqrt{2\pi}} \int_{0}^{+\infty} \sqrt{t} e^{-t} dt \qquad \text{(substitute } t = u^2/2\text{)} \\
&= \sigma^2 \frac{2\sqrt{2}}{\sqrt{2\pi}} \Gamma(3/2) = \sigma^2.
\end{aligned}$$

To get the last line, observe that $\Gamma(3/2) = \frac{1}{2}\Gamma(1/2)$ and $\Gamma(1/2) = \sqrt{\pi}$. Thus we now have a meaning for the parameters μ and σ^2 - they are the mean and variance of the $N(\mu, \sigma^2)$ distribution. Again note that the mean is the same for all $N(0, \sigma^2)$ distributions but the variances are different, capturing the spread of the distribution.

EXERCISE 17. Let $X \sim N(0,1)$. Show that $\mathbb{E}[X^n] = 0$ if n is odd and if n is even then $\mathbb{E}[X^n] = (n-1)(n-3)\dots(3)(1)$ (product of all odd numbers up to and including $n-1$). What happens if $X \sim N(0,\sigma^2)$?

EXERCISE 18. Calculate the mean and variance for the following distributions.

(1) $X \sim \text{Geo}(p)$. $\mathbb{E}[X] = \frac{1}{p}$ and $\text{Var}(X) = \frac{q}{p^2}$.

(2) $X \sim \text{Bin}(n,p)$. $\mathbb{E}[X] = np$ and $\text{Var}(X) = npq$.

(3) $X \sim \text{Pois}(\lambda)$. $\mathbb{E}[X] = \lambda$ and $\text{Var}(X) = \lambda$.

(4) $X \sim \text{Hypergeo}(N_1, N_2, m)$. $\mathbb{E}[X] = \frac{mN_1}{N_1+N_2}$ and $\text{Var}(X) =$??.

EXERCISE 19. Calculate the mean and variance for the following distributions.

(1) $X \sim \text{Exp}(\lambda)$. $\mathbb{E}[X] = \frac{1}{\lambda}$ and $\text{Var}(X) = \frac{1}{\lambda^2}$.

(2) $X \sim \text{Gamma}(\nu,\lambda)$. $\mathbb{E}[X] = \frac{\nu}{\lambda}$ and $\text{Var}(X) = \frac{\nu}{\lambda^2}$.

(3) $X \sim \text{Unif}[0,1]$. $\mathbb{E}[X] = \frac{1}{2}$ and $\text{Var}(X) = \frac{1}{12}$.

(4) $X \sim \text{Beta}(p,q)$. $\mathbb{E}[X] = \frac{p}{p+q}$ and $\text{Var}(X) = \frac{pq}{(p+q)^2(p+q+1)}$.

3. Markov's and Chebyshev's inequalities

Let X be a non-negative integer valued random variable with pmf $f(k)$, $k = 0,1,2,\dots$. Fix any number m, say $m = 10$. Then

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} kf(k) \geq \sum_{k=10}^{\infty} kf(k) \geq \sum_{k=10}^{\infty} 10f(k) = 10\mathbb{P}\{X \geq 10\}.$$

More generally $m\mathbb{P}\{X \geq m\} \leq \mathbb{E}[X]$. This shows that if the expected value is finite This idea is captured in general by the following inequality.

Markov's inequality: Let X be a non-negative random variable with finite expectation. Then, for any $t > 0$, we have $\mathbb{P}\{X \geq t\} \leq \frac{1}{t}\mathbb{E}[X]$.

PROOF. Fix $t > 0$ and let $Y = X\mathbf{1}_{X<t}$ and $Z = X\mathbf{1}_{X\geq t}$ so that $X = Y + Z$. Both Y and Z are non-negative random variable and hence $\mathbb{E}[X] = \mathbb{E}[Y] + \mathbb{E}[Z] \geq \mathbb{E}[Z]$. On the other hand, $Z \geq t\mathbf{1}_{X\geq t}$ (why?). Therefore $\mathbb{E}[Z] \geq t\mathbb{E}[\mathbf{1}_{X\geq t}] = t\mathbb{P}\{X \geq t\}$. Putting these together we get $\mathbb{E}[X] \geq t\mathbb{P}\{X \geq t\}$ as desired to show. ■

Markov's inequality is simple but surprisingly useful. Firstly, one can apply it to functions of our random variable and get many inequalities. Here are some.

Variants of Markov's inequality:

(1) If X is a non-negative random variable with finite p^{th} moment, then $\mathbb{P}\{X \geq t\} \leq t^{-p}\mathbb{E}[X^p]$ for any $t > 0$.

(2) If X is a random variable with finite second moment, then $\mathbb{P}\{|X-\mu| \geq t\} \leq \frac{1}{t^2}\text{Var}(X)$. [*Chebyshev's inequality*]

(3) If X is a random variable with finite exponential moments, then $\mathbb{P}\{X > t\} \leq e^{-\lambda t}\mathbb{E}[e^{\lambda X}]$ for any $\lambda > 0$. This is sometimes called *Chernoff's bound*.

Thus, if we only know that X has finite mean, the tail probability $\mathbb{P}(X > t)$ must decay at least as fast as $1/t$. But if we knew that the second moment was finite we could assert that the decay must be at least as fast as $1/t^2$, which is better. If $\mathbb{E}[e^{\lambda X}] < \infty$, then we get much faster decay of the tail, like $e^{-\lambda t}$.

Chebyshev's inequality captures again the intuitive notion that variance measures the spread of the distribution about the mean. The smaller the variance, lesser the spread. An alternate way to write Chebyshev's inequality is

$$\mathbb{P}\{|X-\mu| > r\sigma\} \leq \frac{1}{r^2}$$

where σ = s.d.(X). This measures the deviations in multiples of the standard deviation. The power of this inequality comes from its great generality - one needs to know nothing about the distribution other than its mean and variance. However, if we do have more information about the distribution, often we can get better bounds than $1/r^2$ (just like Markov inequality can be improved using higher moments, when they exist). We state one such inequality for the important case of symmetric Bernoulli random variables.

PROPOSITION 20 (Bernstein/Hoeffding inequality). *Let $X \sim Bin(n, \frac{1}{2})$. Then*

$$\mathbb{P}\left\{|X - \frac{1}{2}n| \geq t\right\} \leq 2e^{-\frac{2t^2}{n}}$$

To summarize, when we know very little about the distribution, we use Markov or Chebyshev inequalities. If we have some detailed knowledge of the distribution, the bounds on the probability can be greatly improved. The main idea in Chebyshev's inequality is often useful in analysis and probability. Let us illustrate how it can be used to get a bound for the tail of the Gaussian distribution.

3.1. An application to the tail of the Gaussian distribution. The tail of the standard Gaussian distribution is given by $\bar{\Phi}(t) = \frac{1}{\sqrt{2\pi}}\int_t^\infty e^{-x^2/2}dx = \mathbb{P}\{Z > t\}$ where $Z \sim N(0,1)$. There is no simple formula for this integral, but it is often often necessary to have a good upper bound for it. But based on the fact that we know how to integrate $xe^{-x^2/2}$, and using the Chebyshev idea, we can get a good estimate for $\bar{\Phi}(t)$ for $t > 0$.

Upper bound: As $\frac{x}{t} \geq 1$ for $x \geq t$, we have

$$\bar{\Phi}(t) = \frac{1}{\sqrt{2\pi}}\int_t^\infty e^{-\frac{1}{2}x^2}dx \;\leq\; \frac{1}{\sqrt{2\pi}}\int_t^\infty \frac{x}{t}e^{-\frac{1}{2}x^2}dx \;=\; \frac{1}{t\sqrt{2\pi}}e^{-\frac{1}{2}t^2}.$$

Lower bound: For $s > t > 0$, we use that $\frac{x}{s} \leq s$ for $x \in [t, s]$ to get

$$\bar{\Phi}(t) \geq \frac{1}{\sqrt{2\pi}} \int_t^s e^{-\frac{1}{2}x^2} dx \;\geq\; \frac{1}{\sqrt{2\pi}} \int_t^s \frac{x}{s} e^{-\frac{1}{2}x^2} dx \;=\; \frac{1}{s\sqrt{2\pi}} \left(e^{-\frac{1}{2}t^2} - e^{-\frac{1}{2}s^2} \right).$$

Take $s = t + \sqrt{t}$ and use that $s^2 \geq t^2 + 2t^{3/2}$ to get

$$\bar{\Phi}(t) \geq \frac{1}{t\sqrt{2\pi}} e^{-\frac{1}{2}t^2} \times \frac{1 - e^{-t^{\frac{3}{2}}}}{1 + \frac{1}{\sqrt{t}}}$$

From the upper and lower bounds, we see that

$$\bar{\Phi}(t) \sim \frac{1}{t\sqrt{2\pi}} e^{-\frac{1}{2}t^2} \text{ as } t \to \infty$$

meaning that the ratio of the two sides converges to 1.

4. Reasoning by averages

Many situations can be analysed by just considering averages (expectations). Analysis by expectation is not the final word in all cases, however it is often quite indicative and reliable. At a more basic level, it is a counter to more naive ways of reasoning. Sometimes people only consider the chance of something ignoring the cost if it does not happen. For example, most often one can get away with rushing across railway tracks before a train is due, but if one does not, the loss is great. Other times people consider only the amount of profit if something good happens, ignoring that the chance of it happening is tiny. For example, when buying lottery tickets. In contrast, expectation involves the product chance × profit, which is better.

► If you are offered x rupees if a die shows 6 and you have to pay 10 rupees if it does not fall 6, then your expected gain is $x \times \frac{1}{6} - 10 \times \frac{5}{6}$ which is positive if $x \geq 50$. In other words, if $x > 50$, it is profitable *on average* to bet, if $x < 50$ it will result in a loss *on average*.

If you ask whether you should bet or not based on this criterion, it is a different matter. If the game is to be played many times, then the law of large number that we shall see later tells us that yes, you should bet if $x > 50$ and that you should not bet if $x < 50$. If the game is to played once or a few times, then the average is not reliable, one must look at the full distribution of the profit random variable, or at least its variance.

► In an exam where each question is of multiple choice type, say 4 choices with a single correct answer, how should the negative marking be set to discourage random guessing? Suppose $+1$ is given for correct answer and $-x$ for a wrong answer. Then the expected marks gained on a question by guessing at random is $+1 \times \frac{1}{4} - x\frac{3}{4}$. To discourage random guessing, this should be negative, or $x \geq \frac{1}{3}$. That is wrong answers should get $-\frac{1}{3}$ or even less. Suppose the negative mark is set at $-\frac{1}{3}$. Then if you can eliminate one of the answers with certainty, then the expected marks by randomly guessing among the remaining three would be $+1 \times \frac{1}{3} - \frac{1}{3} \times \frac{2}{3} = \frac{1}{9}$. It is worth guessing. If the examiners want to discourage this eventuality, they should set the negative mark to be $-\frac{1}{2}$ or less.

► In BMTC buses, there are random inspections, and passengers without tickets are fined. There are two parameters that the BMTC can choose - frequency of inspections (say proportion p of all trips are inspected) and the quantum of fine (say R rupees). For simplicity if we assume that tickets cost T rupees, and all trips carry the same number of passengers, then the expected cost for a passenger not buying a ticket is $(1-p) \times 0 + p \times R = pR$. This should be more than the actual cost of the ticket to incentivize passengers to buy tickets, hence we want $pR > T$. There is the option to increase p or to increase R Increasing p involves a cost (to pay for the inspectors), but one cannot arbitrarily decrease p and increase R, as it does not make sense to ask for a fine of thousands of rupees for failing to purchase a ticket.

► R. A. Fisher, famous statistician and biologist, gave the following explanation for sex-ratios among many animals. In certain species of animals (elephant seals are given as an example in Dawkin's book *The selfish gene*), most males do not get to mate with any females. The few successful males get to mate with a large number of females. In such a situation, one may ask if it is worthwhile producing male offsprings? Still, the observed sex-ratio is very close to 1:1. To explain this with a calculation of averages, let us assume that there are M males and F females in the population (both are large numbers). And suppose successful males get to mate with K females, but each female mates with only one male. Also for simplicity, assume that each male-female pair results in one offspring.

If a prospective mother gets to have a male offspring, then he gets to mate with a probability of $\frac{F/K}{M}$, and if successful, will have K children. The expected number of grandchildren is $\frac{F}{M}$.

Is she gets to have a female offspring, then she gets to mate for sure, and will have one child. The expected number of grandchildren is 1.

Thus, if $F > M$, it is more advantageous to have a male child, and if $F < M$, it is more advantageous to have a female child (meaning that any heritable tendency to produce more females will spread more). In summary, any imbalance in the Male:Female ratio will get a push in the reverse direction, making the 1:1 ratio the unique stable equilibrium.

► Consider two options: Throw a die and if it turns up 6 you pay 300 rupees and otherwise you get 150 rupees. (A) No gambling, you get 50 rupees.The expected profit in the first case is $150 \times \frac{5}{6} - 300 \times \frac{1}{6} = 75$, which is higher than what you get in the second option. But in reality, most people would prefer the second option, as it does not involve any risk. The notion of risk is captured by standard deviation (and other measures), average alone is not always a good criterion.

► Another famous example where naive use of expectations fails is the *St. Petersburg paradox*. It is given in one of the homework sheets.

► Mendel's famous experiments with pea plants leading to his discovery of the laws of genetics is one of the great stories of Science. As it happens, some of the data on which he based his experiments appear to be fudged! This was discovered by Fisher himself. One of the data that does not stand up to scrutiny is as follows (we assume knowledge of basics of genetics here).One of the characteristics he studied was the shape of the pea pod, which

could be inflated (I) or constricted (C), where I is dominant and C is recessive. It so happens that all peas of a given plant are inflated (happens if the genotype is II, IC or CI) or all are constricted (only if the genotype is CC). If a plant that has IC genotype is crossed with another IC, it gives II, IC, CI, CC with equal probability. In particular, among the resulting plants with inflated pods, one third are expected to be pure II and the remaining are hybrid (IC or CI). In one experiment, Mendel reports that of the 600 plants of the second generation that were yellow, 201 were of genotype II. That is a remarkable agreement with experiment, or so it seems. In fact, it is a little too suspiciously close to the 1/3 ratio expected (toss a die 600 times and count how many ones or twos you get), but the issue cuts even deeper.

While we can see whether a plant has inflated pods, we do not directly see the genotype. The way Mendel could determine the genotype was to self-pollinate a plant many times and and check if all the resulting plants had inflated pods, as expected if the original plant had genotype II. But if the original plant had genotype IC or CI, some of the second generation plants should have genotype CC and hence have constricted pods. Mendel apparently produced 10 plants in the second generation (for each of the original 600 plants) for this purpose. However, even the original was an IC, we should expect $600 \times (3/4)^{10} \approx 22$ cases where all the second generation plants had inflated pods, leading to a mistaken conclusion that the original was an II. This is in addition to the expected 200 that are truly II. It looks as though Mendel's result is closer to what he thought should be expected rather than what should be expected according to his procedure! [*Caveat:* According to the above analysis, the number of plants that were deemed to be II should have been $\text{Bin}(600, 0.38)$ (as $\frac{1}{3} + \left(\frac{3}{4}\right)^{10} = 0.38$), which has a standard deviation of about 11. Hence, 201 is within 2 standard deviations of the expected 222, not such an unlikely possibility. So is Mendel being falsely accused?!]

5. Reasoning with mean and variance

Working with averages is an essential thing to learn, but the real role of probability comes to fore when we also consider *fluctuations*, i.e., deviations from average behaviour.

Example 21. Consider a cubical box of gas. How much of the air is in the top half and how much in the bottom half? Molecules of air move about at random, and if we regard each of them as deciding to go to the bottom half or the top half by an independent fair coin toss, then the number of molecules X in the top half has $\text{Bin}(N, \frac{1}{2})$ distribution, where we take the number of molecules to be $N = 10^{22}$ or some such very large number. Then $\mathbb{E}[X] = \frac{1}{2}N = 5 \times 10^{21}$ while $\text{sd}(X) = \sqrt{N/4} = 5 \times 10^{10}$. By Chebyshev's inequality, $|X - 5 \times 10^{21}| < 10^{15}$ with a probability of more than $1 - 10^{-8}$. As 10^{15} is tiny compared to 10^{21}, the proportion of molecules in the top half is exceedingly close to $\frac{1}{2}$ with exceedingly high probability. For all practical purposes, we may ignore fluctuations and say that half of the gas is in the top half.

The reason why fluctuations became negligible was the hugeness of N. In other situations, say tossing a coin 1000 times, betting on a lottery every day of one month, etc., you ignore fluctuations at your own peril. For example if a coin is fair, it still has a 5% chance of of

showing less than 475 heads in 1000 tosses. Tiny probabilities become significant if there are a lot of trials.

6. Exercises

PROBLEM 1. For a pdf $f(\cdot)$, the *mode* is defined as any point at which f attains its maximal value (i.e., t is a mode if $f(t) \geq f(s)$ for any s). For each of the following distributions, find the mode(s) of the distribution and the value of the pmf at the modes.

(1) $N(\mu, \sigma^2)$.

(2) Exp(λ).

(3) Gamma($\nu, 1$).

PROBLEM 2. Let F be a CDF. For each $0 < q < 1$, the *q-quantile(s)* of F is any number $t \in \mathbb{R}$ such that $F(s) \leq q$ if $s < t$ and $F(s) \geq q$ if $s > t$.

(1) If F is the CDF of Exp(λ) distribution, find its q-quantile(s).

(2) If F is the $N(0,1)$ distribution, use the normal tables to find the unique q-quantile for the following values of q: $0.01, 0.1, 0.25, 0.5, 0.75, 0.9, 0.99$.

(3) If F is the Geo(0.02) distribution, find a q-quantile for $q = 0.01, 0.25, 0.5, 0.75, 0.99$.

PROBLEM 3. Find $\mathbb{E}[X]$ and $\mathbb{E}[X^2]$ for the following cases.

(1) $X \sim N(\mu, \sigma^2)$.

(2) $X \sim$ Gamma(ν, λ). Note the answers for the particular case of Exp(λ).

(3) $X \sim$ Beta(p, q). Note the answers for the particular case of Unif$[0,1]$.

PROBLEM 4. What is the mode of the (a) Pois(λ) distribution? (b) Hpergeometric(M, W, K) distribution? (Mode means the point(s) where the pmf (or pdf) attains its maximal value).

PROBLEM 5. Find the means and variances of X in each of the following cases.

$$(a)\ X \sim \text{Bin}(n,p). \qquad (b)\ X \sim \text{Pois}(\lambda). \qquad (c) X \sim \text{Geo}(p).$$

PROBLEM 6. Find the means and variances of X in each of the following cases.

$$(a)\ X \sim \text{N}(\mu, \sigma^2). \qquad (b)\ X \sim \text{Gamma}(\nu, \lambda). \qquad (c)\ X \sim \text{Beta}(\nu_1, \nu_2). \qquad (d)\ X \sim \text{Unif}[a,b].$$

PROBLEM 7. (1) Let $\xi \sim \text{Exp}(\lambda)$. For any $t, s \geq 0$, show that $\mathbb{P}\{\xi > t+s \mid \xi > t\} = \mathbb{P}\{\xi > s\}$. (This is called the *memoryless property* of the exponential distribution).

(2) Show that if a non-negative random variable ξ has memoryless property (i.e., $\mathbb{P}\{\xi > t+s \mid \xi > t\} = \mathbb{P}\{\xi > s\}$), then ξ must have exponential distribution.

PROBLEM 8. Let X be a non-negative random variable with CDF $F(t)$.

(1) Show that $\mathbb{E}[X] = \int_0^\infty (1 - F(t))dt$ and more generally $\mathbb{E}[X^p] = \int_0^\infty pt^{p-1}(1 - F(t))dt$.

(2) If X is non-negative integer valued, then $\mathbb{E}[X] = \sum_{k=1}^\infty \mathbb{P}\{X \geq k\}$.

PROBLEM 9. (**) Let X be a random variable. Let $f(a) = \mathbb{E}[|X - a|]$ (makes sense if the first moment exists) and $g(a) = \mathbb{E}[(X - a)^2]$ (makes sense if the second moment exists).

(1) Show that g is minimized uniquely at $a = \mathbb{E}[X]$.

(2) Show that the minimizers of f are precisely the medians of X (recall that a number b is a median of X if $\mathbb{P}\{X \geq t\} \geq \frac{1}{2}$ and $\mathbb{P}\{X \leq t\} \geq \frac{1}{2}$).

PROBLEM 10. Find the expectation and variance for a random variable with the following distributions.

(1) (a) Bin(n, p), (b) Geo(p), (c) Pois(λ), (d) Hypergeo(N_1, N_2, m).

(2) (a) $N(\mu, \sigma^2)$, (b) Gamma(ν, λ), (c) Beta(p, q).

[**Note:** Although the computations are easy, the answers you get are worth remembering as they occur in various situations.]

PROBLEM 11. Find the expectation and variance for a random variable with the following distributions.(a) Bin(n, p), (b) Geo(p), (c) Pois(λ), (d) Hypergeo(N_1, N_2, m).[**Note:** Although the computations are easy, the answers you get are worth remembering as they occur in various situations.]

PROBLEM 12. Find the expectation and variance for a random variable with the following distributions.(a) $N(\mu, \sigma^2)$, (b) Gamma(ν, λ), (c) Beta(p, q).[**Note:** Although the computations are easy, the answers you get are worth remembering as they occur in various situations.]

CHAPTER 7

Joint distributions of random variables

1. Joint distributions

In many situations we study several random variables at once. In such a case, knowing the individual distributions is not sufficient to answer all relevant questions. This is like saying that knowing $\mathbb{P}(A)$ and $\mathbb{P}(B)$ is insufficient to calculate $\mathbb{P}(A \cap B)$ or $\mathbb{P}(A \cup B)$ etc.

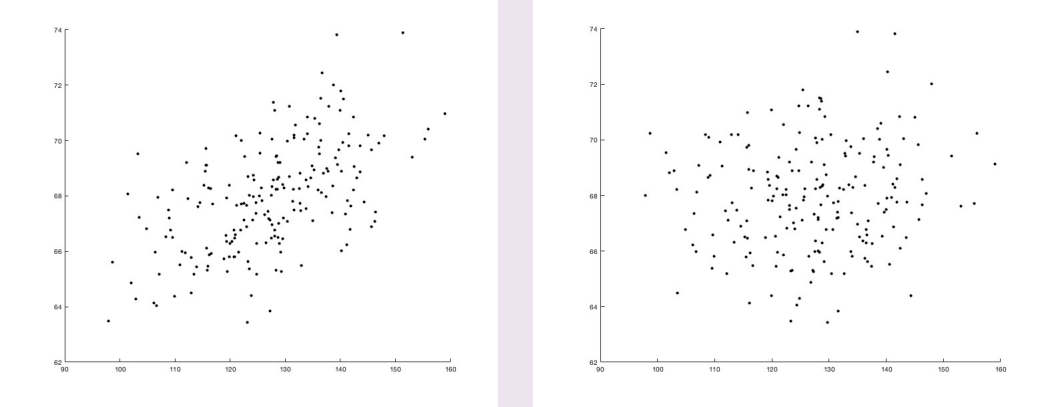

FIGURE 1. Left: Height and weight of a person. Right: Height of a person versus weight of an unrelated person. One can see that there is a tendency for weight to increase with height in the first picture, but no such relation in the second.

DEFINITION 13 (Joint distribution). Let $X_1, X_2, \ldots, X_m$ be random variables on the same probability space. We call $\mathbf{X} = (X_1, \ldots, X_m)$ a *random vector*, as it is just a vector of random variables. The CDF of $\mathbf{X}$, also called the joint CDF of $X_1, \ldots, X_m$ is the function $F : \mathbb{R}^m \to \mathbb{R}$ defined as

$$F(t_1, \ldots, t_m) = \mathbb{P}\{X_1 \leq t_1, \ldots, X_m \leq t_m\} = \mathbb{P}\left\{\bigcap_{i=1}^{m} \{X_i \leq t_i\}\right\}.$$

.

EXAMPLE 14. Consider two events A and B in the probability space and let $X = \mathbf{1}_A$ and $Y = \mathbf{1}_B$ be their indicator random variables. Their joint CDF is given by

$$F(s,t) = \begin{cases} 0 & \text{if } s < 0 \text{ or } t < 0 \\ \mathbb{P}(A^c \cap B^c) & \text{if } 0 \le s, t < 1 \\ \mathbb{P}(A^c) & \text{if } 0 \le s < 1 \text{ and } t \ge 1 \\ \mathbb{P}(B^c) & \text{if } 0 \le t < 1 \text{ and } s \ge 1 \\ 1 & \text{if } s \ge 1, t \ge 1 \end{cases}$$

Properties of joint CDFs: The following properties of the joint CDF $F : \mathbb{R}^m \to [0,1]$ are analogous to those of the 1-dimensional CDF and the proofs are similar.

(1) F is increasing in each co-ordinate. That is, if $s_1 \le t_1, \ldots, s_m \le t_m$, then $F(s_1, \ldots, s_m) \le F(t_1, \ldots, t_m)$.

(2) $\lim F(t_1, \ldots, t_m) = 0$ if $\min\{t_1, \ldots, t_m\} \to -\infty$ (i.e., one of the t_i goes to $-\infty$).

(3) $\lim F(t_1, \ldots, t_m) = 1$ if $\min\{t_1, \ldots, t_m\} \to +\infty$ (i.e., all of the t_i goes to $+\infty$).

(4) F is right continuous in each co-ordinate. That is $F(t_1+h_1, \ldots, t_m+h_m) \to F(t_1, \ldots, t_m)$ as $h_i \to 0+$.

Conversely any function having these four properties is the joint CDF of some random variables.

From the joint CDF, it is easy to recover the individual CDFs. Indeed, if $F : \mathbb{R}^m \to \mathbb{R}$ is the CDF of $\mathbf{X} = (X_1, \ldots, X_m)$, then the CDF of X_1 is given by

$$F_1(t) := F(t, +\infty, \ldots, +\infty) := \lim F(t, s_2, \ldots, s_m)$$

where the limit is taken as $s_i \to +\infty$ for each $i = 2, \ldots, m$. This is true because if $A_n := \{X_1 \le t\} \cap \{X_2 \le n\} \cap \ldots \cap \{X_m \le n\}$, then as $n \to \infty$, the events A_n increase to the event $A = \{X_1 \le t\}$. Hence $\mathbb{P}(A_n) \to \mathbb{P}(A)$. But $\mathbb{P}(A_n) = F(t, n, n, \ldots, n)$ and $\mathbb{P}(A) = F_1(t)$. Thus we see that $F_1(t) := F(t, +\infty, \ldots, +\infty)$.

More generally, we can recover the joint CDF of any subset of $X_1, \ldots, X_n$, for example, the joint CDF of $X_1, \ldots, X_k$ is just $F(t_1, \ldots, t_k, +\infty, \ldots, +\infty)$.

Joint pmf and pdf: Just like in the case of one random variable, we can consider the following two classes of random variables.

(1) Distributions with a pmf. These are CDFs for which there exist points $\mathbf{t}_1, \mathbf{t}_2, \ldots$ in $\mathbb{R}^m$ and non-negative numbers w_i such that $\sum_i w_i = 1$ (often we write $f(t_i)$ in place of w_i) and such that for every $\mathbf{t} \in \mathbb{R}^m$ we have

$$F(\mathbf{t}) = \sum_{i\,:\,\mathbf{t}_i \le \mathbf{t}} w_i$$

where $\mathbf{s} \leq \mathbf{t}$ means that each co-ordinate of s is less than or equal to the corresponding co-ordinate of $\mathbf{t}$.

Even within distributions with pmf, it is useful to think of the following simple situation: Consider two random variables X and Y, and suppose their ranges are $\{a_1, \ldots, a_n\}$ and $\{b_1, \ldots, b_m\}$ respectively. Then a convenient way to write their pmf is as a two-way table (called *contingency table*)

	b_1	b_2	$\ldots$	b_m	
a_1	$p_{1,1}$	$p_{1,2}$	$\ldots$	$p_{1,m}$	$p_{1,\bullet}$
a_2	$p_{2,1}$	$p_{2,2}$	$\ldots$	$p_{2,m}$	$p_{2,\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_n	$p_{n,1}$	$p_{n,2}$	$\ldots$	$p_{n,m}$	$p_{n,\bullet}$
	$p_{\bullet,1}$	$p_{\bullet,2}$	$\ldots$	$p_{\bullet,m}$	

Here $p_{i,j} = \mathbb{P}\{X = a_i, Y = b_j\}$. And the row and column sums are written as $p_{i,\bullet} = p_{i,1} + \ldots + p_{i,m}$ and $p_{\bullet,j} = p_{1,j} + \ldots + p_{n,j}$. It is clear that $p_{i,\bullet} = \mathbb{P}\{X = a_i\}$ and $p_{\bullet,j} = \mathbb{P}\{Y = b_j\}$. Thus, the individual distributions of X and of Y can be recovered from the joint distribution of X and Y.

(2) Distributions with a pdf. These are CDFs for which there is a non-negative function (may assume piecewise continuous for convenience) $f : \mathbb{R}^m \to \mathbb{R}_+$ such that for every $\mathbf{t} \in \mathbb{R}^m$ we have

$$F(\mathbf{t}) = \int_{-\infty}^{t_1} \ldots \int_{-\infty}^{t_m} f(u_1, \ldots, u_m) du_m \ldots du_1.$$

Again, it is useful to think of the case of two random variables, X and Y, having a joint pdf of $f(x, y)$. This can be thought in analogy to the contingency table above, except that the table has infinitely many rows (one for each x) and infinitely many columns (one for each y). The individual densities of X and of Y are got by integrating instead of summing:

$$f_1(u) = \int_{-\infty}^{\infty} f(u, y) dy, \qquad f_2(v) = \int_{-\infty}^{\infty} f(x, v) dx.$$

We give two examples, one of each kind.

EXAMPLE 15. (Multinomial distribution). Fix parameters r, m (two positive integers) and $p_1, \ldots, p_m$ (positive numbers that add to 1). The *multinomial pmf* with these parameters is given by

$$f(k_1, \ldots, k_{m-1}) = \frac{r!}{k_1! k_2! \ldots k_{m-1}! (r - \sum_{i=1}^{m-1} k_i)!} p_1^{k_1} \ldots p_{m-1}^{k_{m-1}} p_m^{r - \sum_{i=1}^{m-1} k_i},$$

if $k_i \geq 0$ are integers such that $k_1 + \ldots + k_{m-1} \leq r$. One situation where this distribution arises is when r balls are randomly placed in m bins, with each ball going into the jth bin

with probability p_j, and we look at the random vector $(X_1, \ldots, X_{m-1})$ where X_k is the number of balls that fell into the kth bin. This random vector has the multinomial pmf.[1]

In this case, the marginal distribution of X_k is $\text{Bin}(r, p_k)$. More generally, $(X_1, \ldots, X_\ell)$ has multinomial distribution with parameters $r, \ell, p_1, \ldots, p_\ell, p_0$ where $p_0 = 1 - (p_1 + \ldots + p_\ell)$. This is easy to prove, but even easier to see from the balls in bins interpretation (just think of the last $n - \ell$ bins as one).

EXAMPLE 16. (Bivariate normal distribution). This is the density on $\mathbb{R}^2$ given by

$$f(x, y) = \frac{\sqrt{ab - c^2}}{2\pi} e^{-\frac{1}{2}\left[a(x-\mu)^2 + b(y-\nu)^2 + 2c(x-\mu)(y-\nu)\right]},$$

where μ, ν, a, b, c are real parameters. We shall impose the conditions that $a > 0$, $b > 0$ and $ab - c^2 > 0$ (otherwise the above does not give a density, as we shall see).

The first thing is to check that this is indeed a density. We recall the one-dimensional Gaussian integral

$$\int_{-\infty}^{+\infty} e^{-\frac{\tau}{2}(x-a)^2} dx = \sqrt{2\pi}\frac{1}{\sqrt{\tau}} \text{ for any } \tau > 0 \text{ and any } a \in \mathbb{R}. \tag{9}$$

We shall take $\mu = \nu = 0$ (how do you compute the integral if they are not?). Then, the exponent in the density has the form

$$ax^2 + by^2 + 2cxy = b\left(y + \frac{c}{b}x\right)^2 + \left(a - \frac{c^2}{b}\right)x^2.$$

Therefore,

$$\begin{aligned}\int_{-\infty}^{\infty} e^{-\frac{1}{2}\left[ax^2 + by^2 + 2cxy\right]} dy &= e^{-\frac{1}{2}(a - \frac{c^2}{b})x^2} \int_{-\infty}^{\infty} e^{-\frac{b}{2}(y + \frac{c}{b}x)^2} dy \\ &= e^{-\frac{1}{2}(a - \frac{c^2}{b})x^2} \frac{\sqrt{2\pi}}{\sqrt{b}}\end{aligned}$$

by (9) but ony if $b > 0$. Now we integrate over x and use (9) again (and the fact that $a - \frac{c^2}{b} > 0$) to get

$$\begin{aligned}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2}\left[a(x-\mu)^2 + b(y-\nu)^2 + 2c(x-\mu)(y-\nu)\right]} dy dx &= \frac{\sqrt{2\pi}}{\sqrt{b}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(a - \frac{c^2}{b})x^2} dx \\ &= \frac{\sqrt{2\pi}}{\sqrt{b}} \frac{\sqrt{2\pi}}{\sqrt{a - \frac{c^2}{b}}} = \frac{2\pi}{\sqrt{ab - c^2}}.\end{aligned}$$

[1] In some books, the distribution of $(X_1, \ldots, X_m)$ is called the multinomial distribution. This has the pmf

$$g(k_1, \ldots, k_m) \frac{r!}{k_1! k_2! \ldots k_{m-1}! k_m!} p_1^{k_1} \ldots p_{m-1}^{k_{m-1}} p_m^{k_m}$$

where k_i are non-negative integers such that $k_1 + \ldots + k_m = r$. We have chosen our convention so that the binomial distribution is a special case of the multinomial...

This completes the proof that $f(x,y)$ is indeed a density. Note that $b > 0$ and $ab - c^2 > 0$ also implies that $a > 0$.

Matrix form of writing the density: Let $\Sigma^{-1} = \begin{bmatrix} a & c \\ c & b \end{bmatrix}$. Then, $\det(\Sigma) = \frac{1}{\det(\Sigma^{-1})} = \frac{1}{ab-c^2}$. Hence, we may re-write the density above as (let $\mathbf{u}$ be the column vector with co-ordinates x, y)

$$f(x,y) = \frac{1}{2\pi\sqrt{\det(\Sigma)}} e^{-\frac{1}{2}\mathbf{u}^t\Sigma^{-1}\mathbf{u}}.$$

This is precisely in the form in which we wrote for general n in the example earlier. The conditions $a > 0, b > 0, ab - c^2 > 0$ translate precisely to what is called positive-definiteness. One way to say it is that Σ is a symmetric matrix and all its eigenvalues are strictly positive.

Final form: We can now introduce an extra pair of parameters μ_1, μ_2 and define a density

$$f(x,y) = \frac{1}{2\pi\sqrt{\det(\Sigma)}} e^{-\frac{1}{2}(\mathbf{u}-\mu)^t\Sigma^{-1}(\mathbf{u}-\mu)}.$$

where μ is a column vector with co-ordinates μ_1, μ_2. This is the full bi-variate normal density.

An important class of joint distributions can be constructed as follows. We discuss independence of random variables in greater detail later.

EXAMPLE 17. (A class of examples). Let $f_1, f_2, \ldots, f_m$ be one-variable densities. In other words, $f_i : \mathbb{R} \to \mathbb{R}_+$ and $\int_{-\infty}^{\infty} f_i(x)dx = 1$. Then, we can make a multivariate density as follows. Define $f : \mathbb{R}^m \to \mathbb{R}_+^m$ by $f(x_1, \ldots, x_m) = f_1(x_1) \ldots f_m(x_m)$. Then f is a density.

If X_i are random variables on a common probability space and the joint density of $(X_1, \ldots, X_m)$ if $f(x_1, \ldots, x_m)$, then we say that X_i are *independent random variables*. It is easy to see that the marginal density of X_i if f_i. It is also the case that the joint CDF factors as $F_X(x_1, \ldots, x_m) = F_{X_1}(x_1) \ldots F_{X_m}(x_m)$.

One can also make the discrete version of this. In the contingency table example given earlier, set $p_{i,j} = \alpha_i\beta_j$ for some $\alpha_i, \beta_j \geq 0$ such that $\alpha_1 + \ldots + \alpha_n = 1$ and $\beta_1 + \ldots + \beta_m = 1$. Then the random variables are said to be independent. Note that in this case $p_{i,\bullet} = \alpha_i$ and $p_{\bullet,j} = \beta_j$.

2. Independence and conditioning of random variables

DEFINITION 18. Let $\mathbf{X} = (X_1, \ldots, X_m)$ be a random vector (this means that X_i are random variables on a common probability space). We say that X_i are *independent* if $\{X_1 \in A_1\}, \{X_2 \in A_2\}, \ldots, \{X_m \leq t_m\}$ are independent for any[2] $A_1, \ldots, A_m \subseteq \mathbb{R}$, i.e.,

$$\mathbb{P}\{X_1 \in A_1, \ldots, X_m \in A_m\} = \mathbb{P}\{X_1 \in A_1\} \ldots \mathbb{P}\{X_m \in A_m\}. \tag{10}$$

[2]Later when you learn measure theory, you will see that the sets have to be restricted for us to speak of their probabilities, but what you do not know cannot bother you, so we ignore that issue.

Independence is one of the fundamental concepts in probability. In applications to the real world, it means that knowing the value of some of the X_is don't give a clue about the values of the others. The following exercise shows that the independence of random variables is an extension of the notion of independence of events that we discussed earlier.

EXERCISE 19. Events $E_1, \ldots, E_n$ are independent (according to our earlier definition) if and only if the random variables $\mathbf{1}_{E_1}, \ldots, \mathbf{1}_{E_n}$ are independent (according to the above definition).

It is important to turn the concept around and look at it in different ways. Applying the definition to the events $A_k = \{X_k \leq t_k\}$ shows that

$$F_{\mathbf{X}}(t_1, \ldots, t_n) = F_{X_1}(t_1) \ldots F_{X_n}(t_n) \text{ foor all } t_1, \ldots, t_n \in \mathbb{R}. \tag{11}$$

In fact, the above condition is equivalent to independence, but we shall not need that. Since we shall only work with random variables that have pmf or pdf, let us see how independence is reflected in terms of pmf or pdf.

- Suppose $\mathbf{X}$ has joint pmf $f(t_1, \ldots, t_n)$ and let f_k be the marginal pmf of X_k. Then $X_1, \ldots, X_n$ are independent if and only if

$$f(t_1, \ldots, t_n) = f_1(t_1) f_2(t_2) \ldots f_n(t_n) \text{ for all } t_1, \ldots, t_n \in \mathbb{R}. \tag{12}$$

  Proof: If X_i are independent, fix $t_1, \ldots t_n \in \mathbb{R}$ and apply the definition with $A_k = \{X_k = t_k\}$ to see that $f(t_1, \ldots, t_n) = f_1(t_1) f_2(t_2) \ldots f_n(t_n)$.

  For the converse part, let $A_k \subseteq \mathbb{R}$ and consider

$$\begin{aligned} \mathbb{P}\{X_1 \in A_1, \ldots, X_n \in A_n\} &= \sum_{\mathbf{t} \in A_1 \times \ldots \times A_n} f(t_1, \ldots, t_n) \;=\; \sum_{t_j \in A_j,\ 1 \leq ej \leq n} f_1(t_1) f_2(t_2) \ldots f_n(t_n) \\ &= \prod_{k=1}^{n} \sum_{t_k \in A_k} f_k(t_k) \;=\; \prod_{k=1}^{m} \mathbb{P}\{X_k \in A_k\} \end{aligned}$$

  which shows the independence of $X_1, \ldots, X_n$.

- Assume that $\mathbf{X}$ has joint pdf $f(t_1, \ldots, t_m)$. Then X_is are independent if and only if

$$f(t_1, \ldots, t_m) = f_1(t_1) f_2(t_2) \ldots f_m(t_m). \tag{13}$$

  Proof: The proof of the converse part is almost identical to the one given above for pmfs, except that sums are replaced by integrals. That is, $\mathbb{P}\{X_1 \in A_1, \ldots, X_n \in A_n\}$ is equal to

$$\begin{aligned} \int_{A_1 \times \ldots \times A_n} f(t_1, \ldots, t_n) dt_1 \ldots dt_n &= \int_{A_1 \times \ldots \times A_n} f_1(t_1) f_2(t_2) \ldots f_n(t_n) dt_1 \ldots dt_n \\ &= \prod_{k=1}^{n} \int_{A_k} f_k(t_k) dt_k \;=\; \prod_{k=1}^{m} \mathbb{P}\{X_k \in A_k\}. \end{aligned}$$

  The forward part is a little more subtle, as we cannot take the events $\{X_k = t_k\}$ (which have probability zero). If we assume that the densities are continuous, then

$f(t_1, \ldots, t_m) = \frac{\partial^m}{\partial t_1 \ldots \partial t_m} F(t_1, \ldots, t_m)$. Independence implies (11), and upon differentiating w.r.t. $t_1, \ldots, t_n$ we see that

$$f(t_1, \ldots, t_m) = F_1'(t_1) \ldots F_m'(t_m) = f_1(t_1) \ldots f_m(t_m).$$

which is what we wanted to prove.

Here is yet another way to express independence.

PROPOSITION 20. *$X_1, \ldots, X_m$ are independent if and only if*

$$\mathbb{E}[\varphi_1(X_1) \ldots \varphi_m(X_m)] = \mathbb{E}[\varphi_1(X_1)] \ldots \mathbb{E}[\varphi_m(X_m)] \tag{14}$$

for any functions $\varphi_i : \mathbb{R} \to \mathbb{R}$ (assume expectations exist).

PROOF. Suppose (14) holds. Take any sets $A_k \subseteq \mathbb{R}$ and let $\varphi_k = \mathbf{1}_{A_k}$. The left and right sides of (14) are equal to the left and right sides of eq:independencervsdefn, hence $X_1, \ldots, X_n$ are independent.

To see the converse, let us assume that $\mathbf{X}$ has joint density (similar proof if it has joint pmf). By (13), we see that

$$\begin{aligned}
\mathbb{E}[\varphi_1(X_1) \ldots \varphi_m(X_m)] &= \int \ldots \int \varphi_1(t_1) \ldots \varphi_m(t_m) f_1(t_1) \ldots f_m(t_m) dt_1 \ldots dt_m \\
&= \prod_{k=1}^{m} \int \varphi_k(t_k) f_k(t_k) dt_k \\
&= \mathbb{E}[\varphi_1(X_1)] \ldots \mathbb{E}[\varphi_m(X_m)].
\end{aligned}$$

■

Each of (10)–(14) is a way of expressing independence of random variables. When we know independence and want to use it, (14) gives the strongest conclusion, but when we want to check independence, (12) or (13) are the easier ones to verify. The following exercise (slightly) simplifies the checking further.

EXERCISE 21. Suppose $f(t_1, \ldots, t_m) = c g_1(t_1) g_2(t_2) \ldots g_m(t_m)$ where c is a constant and g_i are some functions of one-variable (not necessarily densities). Then, $X_1, \ldots, X_m$ are independent. Further, the marginal density of X_k is $c_k g_k(t)$ for appropriate c_ks.

Now for some examples.

EXAMPLE 22. Let $\Omega = \{0,1\}^n$ with $p_{\underline{\omega}} = p^{\sum \omega_k} q^{n - \sum \omega_k}$. Define $X_k : \Omega \to \mathbb{R}$ by $X_k(\underline{\omega}) = \omega_k$. In words, we are considering the probability space corresponding to n tosses of a fair coin and X_k is the result of the kth toss. We claim that $X_1, \ldots, X_n$ are independent. Indeed, the joint pmf of $X_1, \ldots, X_n$ is

$$f(t_1, \ldots, t_n) = p^{\sum t_k} q^{n - \sum t_k} \quad \text{where } t_i = 0 \text{ or } 1 \text{ for each } i \leq n.$$

Clearly $f(t_1,\dots,t_m) = g(t_1)g(t_2)\dots g(t_n)$ where $g(s) = p^s q^{1-s}$ for $s = 0$ or 1 (this is just a terse way of saying that $g(s) = p$ if $s = 1$ and $g(s) = q$ if $s = 0$). Hence $X_1,\dots,X_n$ are independent and X_k has pmf g (i.e., $X_k \sim \text{Ber}(p)$).

EXAMPLE 23. Let (X,Y) have the bivariate normal density

$$f(x,y) = \frac{\sqrt{ab-c^2}}{\sqrt{2\pi}} e^{-\frac{1}{2}(a(x-\mu_1)^2+b(y-\mu_2)^2+2c(x-\mu_1)(y-\mu_2))}.$$

If $c = 0$, we observe that

$$f(x,y) = C_0 e^{-\frac{a(x-\mu_1)^2}{2}} e^{-\frac{b(y-\mu_2)^2}{2}} \qquad (C_0 \text{ is a constant, exact value unimportant})$$

from which we deduce that X and Y are independent and $X \sim N(\mu_1, \frac{1}{a})$ while $Y \sim N(\mu_2, \frac{1}{b})$.

Can you argue that if $c \neq 0$, then X and Y are not independent?

EXAMPLE 24. Let (X,Y) be a random vector with density $f(x,y) = \frac{1}{\pi}\mathbf{1}_{x^2+y^2\le 1}$ (i.e., it equals 1 if $x^2+y^2 \le 1$ and equals 0 otherwise). This corresponds to picking a point at random from the disk of radius 1 centered at $(0,0)$. We claim that X and Y are not independent. A quick way to see this is that if $I = [0.8,1]$, then $\mathbb{P}\{(X,Y) \in [0.8,1]\times[0.8,1]\} = 0$ whereas $\mathbb{P}\{(X,Y) \in [0.8,1]\}\mathbb{P}\{(X,Y) \in [0.8,1]\} \neq 0$ (If X,Y were independent, we must have had $\mathbb{P}\{(X,Y) \in [a,b]\times[c,d]\} = \mathbb{P}\{X \in [a,b]\}\mathbb{P}\{Y \in [c,d]\}$ for any $a<b$ and $c<d$).

A very useful (and intuitively acceptable!) fact about independence is as follows.

Fact: Suppose $X_1,\dots,X_n$ are independent random variables. Let $k_1 < k_2 < \dots < k_m = n$. Let $Y_1 = h_1(X_1,\dots,X_{k_1})$, $Y_2 = h_2(X_{k_1+1},\dots,X_{k_2}),\dots Y_m = h_m(X_{k_{m-1}+1},\dots,X_{k_m})$. Then, $Y_1,\dots,Y_m$ are also independent.

3. Conditioning on random variables

[3]Let $X_1,\dots,X_{k+\ell}$ be random variables on a common probability space. Let $f(t_1,\dots,t_{k+\ell})$ be the pmf of $(X_1,\dots,X_{k+\ell})$ and let $g(t_1,\dots,t_\ell)$ be the pmf of $(X_{k+1},\dots,X_{k+\ell})$ (of course we can compute g from f by summing over the first k indices). Then, for any $s_1,\dots,s_\ell$ such that $\mathbb{P}\{X_{k+1} = s_1,\dots X_m = s_\ell\} > 0$, we can define

(15)

$$h_{s_1,\dots,s_\ell}(t_1,\dots,t_k) = \mathbb{P}\{X_1 = t_1,\dots,X_k = t_k \,\big|\, X_{k+1} = s_1,\dots X_m = s_\ell\} = \frac{f(t_1,\dots,t_k,s_1,\dots,s_\ell)}{g(s_1,\dots,s_\ell)}.$$

It is easy to see that $h_{s_1,\dots,s_\ell}(\cdot)$ is a pmf on $\mathbb{R}^k$. It is called the conditional pmf of $(X_1,\dots,X_k)$ given that $X_{k+1} = s_1,\dots X_m = s_\ell$.

Its interpretation is as follows. Originally we had random observables $X_1,\dots,X_k$ which had a certain joint pmf. Then we observe the values of the random variables $X_{k+1},\dots,X_{k+\ell}$, say they turn out to be $s_1,\dots,s_\ell$, respectively. Then we update the distribution (or pmf) of $X_1,\dots,X_k$ according to the above recipe. The conditional pmf is the new function $h_{s_1,\dots,s_\ell}(\cdot)$.

[3]This part was not covered in class and may be safely omitted.

EXERCISE 25. Let $(X_1, \ldots, X_{n-1})$ be a random vector with multinomial distribution with parameters $r, n, p_1, \ldots, p_n$. Let $k < n-1$. Given that $X_{k+1} = s_1, \ldots, X_{n-1} = s_{n-k+1}$, show that the conditional distribution of $(X_1, \ldots, X_k)$ is multinomial with parameters $r', n', q_1, \ldots, q_{k+1}$ where $r' = r - (s_1 + \ldots + s_{n-k+1})$, $n' = k+1$, $q_j = p_j/(p_1 + \ldots + p_k + p_n)$ for $j \leq k$ and $q_{k+1} = p_n/(p_1 + \ldots + p_k + p_n)$.

This looks complicated, but is utterly obvious if you think in terms of assigning r balls into n urns by putting each ball into the urns with probabilities $p_1, \ldots, p_n$ and letting X_j denote the number of balls that end up in the j^{th} urn.

Conditional densities Now suppose $X_1, \ldots, X_{k+\ell}$ have joint density $f(t_1, \ldots, t_{k+\ell})$ and let $g(s_1, \ldots, s_\ell)$ by the density of $(X_{k+1}, \ldots, X_{k+\ell})$. Then, we define the conditional density of $(X_1, \ldots, X_k)$ given $X_{k+1} = s_1, \ldots, X_{k+\ell} = s_\ell$ as

$$h_{s_1,\ldots,s_\ell}(t_1, \ldots, t_k) = \frac{f(t_1, \ldots, t_k, s_1, \ldots, s_\ell)}{g(s_1, \ldots, s_\ell)}. \tag{16}$$

This is well-defined whenever $g(s_1, \ldots, s_\ell) > 0$.

REMARK 26. Note the difference between (15) and (16). In the latter we have left out the middle term because $\mathbb{P}\{X_{k+1} = s_1, \ldots, X_{k+\ell} = s_\ell\} = 0$. In (15) the definition of pmf comes from the definition of conditional probability of events but in (16) this is not so. We simply define the conditional density by analogy with the case of conditional pmf. This is similar to the difference between interpretation of pmf ($f(t)$ is actually the probability of an event) and pdf ($f(t)$ is not the probability of an event but the density of probability near t).

EXAMPLE 27. Let (X, Y) have bivariate normal density $f(x, y) = \frac{\sqrt{ab-c^2}}{2\pi} e^{-\frac{1}{2}(ax^2+by^2+2cxy)}$ (so we assume $a > 0, b > 0, ab - c^2 > 0$). In the mid-term you showed that the marginal distribution of Y is $N(0, \frac{a}{ab-c^2})$, that is it has density $g(y) = \frac{\sqrt{ab-c^2}}{\sqrt{2\pi a}} e^{-\frac{ab-c^2}{2a}y^2}$. Hence, the conditional density of X given $Y = y$ is

$$h_y(x) = \frac{f(x, y)}{g(y)} = \frac{\sqrt{a}}{\sqrt{2\pi}} e^{-\frac{a}{2}(x+\frac{c}{a}y)^2}.$$

Thus the conditional distribution of X given $Y = y$ is $N(-\frac{cy}{a}, \frac{1}{a})$. Compare this with marginal (unconditional) distribution of X which is $N(0, \frac{b}{ab-c^2})$.

In the special case when $c = 0$, we see that for any value of y, the conditional distribution of X given $Y = y$ is the same as the unconditional distribution of X. What does this mean? It is just another way of saying that X and Y are independent! Indeed, when $c = 0$, the joint density $f(x, y)$ splits into a product of two functions, one of x alone and one of y alone.

EXERCISE 28. Let (X, Y) have joint density $f(x, y)$. Let the marginal densities of X and Y be $g(x)$ and $h(y)$ respectively. Let $h_x(y)$ be the conditional density of Y given $X = x$.

(1) If X and Y are independent, show that for any x, we have $h_x(y) = h(y)$ for all y.

(2) If $h_x(y) = h(y)$ for all y and for all x, show that X and Y are independent.

Analogous statements hold for the case of pmf.

4. Change of variable formula

Let $\mathbf{X} = (X_1, \ldots, X_m)$ be a random vector with density $f(t_1, \ldots, t_m)$. Let $T : \mathbb{R}^m \to \mathbb{R}^m$ be a one-one function. Then it has an inverse $T^{-1} : \mathbb{R}^m \to \mathbb{R}^m$. Assume that T^{-1} is continuously differentiable.

Let $\mathbf{Y} = T(\mathbf{X})$. In co-ordinates we may write $\mathbf{Y} = (Y_1, \ldots, Y_m)$ and $Y_1 = T_1(X_1, \ldots, X_m) \ldots Y_m = T_m(X_1, \ldots, X_m)$ where $T_i : \mathbb{R}^m \to \mathbb{R}$ are the components of T.

Question: What is the joint density of $Y_1, \ldots, Y_m$?

The change of variable formula: In the setting described above, the joint density of $Y_1, \ldots, Y_m$ is given by

$$g(\mathbf{y}) = f\left(T^{-1}\mathbf{y}\right) \mid JT^{-1}(\mathbf{y}) \mid$$

where $JT^{-1}(\mathbf{y})$ is the Jacobian determinant of the function T^{-1} at the point $\mathbf{y} = (y_1, \ldots, y_m)$. That is, if we write $T^{-1} = S = (S_1, \ldots, S_m)$, with $S_i : \mathbb{R} \to \mathbb{R}$, then

$$JT^{-1}(\mathbf{y}) = \det\left(\frac{\partial S_i}{\partial y_j}(\mathbf{y})\right)_{1 \le i,j \le m}.$$

Justification: We shall not prove this formula, but give a imprecise but convincing justification that can be made into a proof. There are two factors on the right. The first one, $f(T^{-1}\mathbf{y})$ is easy to understand - if $\mathbf{Y}$ is to be close to $\mathbf{y}$, then $\mathbf{X}$ must be close to $T^{-1}\mathbf{y}$. The second factor involving the Jacobian determinant comes from the volume change. Let us explain with analogy with mass density which is a more familiar quantity.

Consider a solid cube with non-uniform density. If you rotate it, the density at any point now is the same as the original density, but at a different point (the one which came to the current position). Instead of rotating, suppose we uniformly expand the cube so that the center stays where it is and the side of the cube becomes twice what it is. What happens to the density at the center? It goes down by a factor of 8. This is simply because of volume change - the same mass spreads over a larger volume. More generally, we can have non-uniform expansion, we may cool some parts of the cube, heat some parts and to varying degrees. What happens to the density? At each point, the density changes by a factor given by the Jacobian determinant.

Now for a slightly more mathematical justification. We use the language for two variables ($m = 2$) but the same reasoning works for any m. Fix twp point $\mathbf{x} = (x_1, x_2)$ and $\mathbf{y} = (y_1, y_2)$ such that $\mathbf{y} = T(\mathbf{x})$ (and hence $\mathbf{x} = T^{-1}(\mathbf{y})$). The density of $\mathbf{Y}$ at $\mathbf{y}$ is given by

$$g(\mathbf{y}) \approx \frac{1}{\text{area}(\mathcal{N})}\mathbb{P}\{\mathbf{Y} \in \mathcal{N}\}$$

where $\mathcal{N}$ is a small neighbourhood of the point $\mathbf{y}$ (for example a disk of small radius δ centered at $\mathbf{y}$). By the one-one nature of T and the relationship $\mathbf{Y} = T(\mathbf{X})$, we see that

$$\mathbb{P}\{\mathbf{Y} \in \mathcal{N}\} = \mathbb{P}\{\mathbf{X} \in T^{-1}(\mathcal{N})\}$$

where $T^{-1}(\mathcal{N})$ is the image of $\mathcal{N}$ after mapping by T^{-1}. Now, $T^{-1}(\mathcal{N})$ is a small neighbourhood of $\mathbf{x}$ (if $\mathcal{N}$ is a disk, then $T^{-1}(\mathcal{N})$ would be an approximate ellipse) and hence, by the same interpretation of density we see that

$$\mathbb{P}\{\mathbf{X} \in T^{-1}(\mathcal{N})\} \approx \text{area}(T^{-1}(\mathcal{N}))f(\mathbf{x})$$

Putting the three displayed equations together, we arrive at the formula

$$g(\mathbf{y}) \approx f(\mathbf{x})\frac{\text{area}(T^{-1}(\mathcal{N}))}{\text{area}(\mathcal{N})}$$

Thus the problem boils down to how areas change under transformations. A linear map $S(\mathbf{y}) = A\mathbf{y}$ where A is a 2×2 matrix changes area of any region by a factor of $|\det(A)|$, i.e., $\text{area}(S(\mathcal{R})) = |\det(A)|\text{area}(\mathcal{R})$.

The differentiability of T means that in a small neighbourhood of $\mathbf{y}$, the mapping T^{-1} looks like a linear map, $T^{-1}(\mathbf{y} + \mathbf{h}) \approx \mathbf{x} + DT^{-1}(\mathbf{y})\mathbf{h}$. Therefore, the areas of small neighbourhoods of $\mathbf{y}$ change by a factor equal to $|\det(DT^{-1}(\mathbf{y}))|$ which is the Jacobian determinant. In other words, $\text{area}(T^{-1}(\mathcal{N})) \approx |JT^{-1}(\mathbf{y})|\text{area}(\mathcal{N})$. Consequently $g(\mathbf{y}) = f(T^{-1}\mathbf{y})|JT^{-1}(\mathbf{y})|$.

Enlarging the applicability of the change of variable formula: The change of variable formula is applicable in greater generality than we stated above.

(1) Firstly, T does not have to be defined on all of $\mathbb{R}^m$. If U, V are open subsets of $\mathbb{R}^m$ and $T : U \to V$ is a bijection so that $T^{-1} : V \to U$ is differentiable, then the change of variable formula is still applicable, provided that $\mathbf{X}$ takes values in U (i.e., $\mathbb{P}\{\mathbf{X} \in U\} = 1$) and $\mathbf{Y} = T(\mathbf{X})$.

(2) One may allow T^{-1} to be not differentiable at finitely many points. In fact, more generally, suppose $D \subseteq V$ is a closed subset so that $C = T^{-1}(D)$ is a closed subset of U. If $\mathbb{P}\{\mathbf{X} \in C\} = 0$ (equivalently if $\mathbb{P}\{\mathbf{Y} \in D\} = 0$), then we may replace U and V by $U' = U \setminus C$ and $V' = V \setminus D$ and apply the change of variable formula.

(3) Injectivity property of T is important, but there are special cases which can be dealt with by a slight modification. For example, if $T(x) = x^2$ or $T(x_1, x_2) = (x_1^2, x_2^2)$ where we can split the space into parts on each of which T is one-one. For concreteness, assume that T is a 3-to-1 map, i.e., every element of V has three pre-images. Then the change of variable formula modifies to

$$g(\mathbf{y}) = f(\mathbf{x}_1)JT_1^{-1}(\mathbf{y}) + f(\mathbf{x}_2)JT_2^{-1}(\mathbf{y}) + f(\mathbf{x}_3)JT_3^{-1}(\mathbf{y})$$

where $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ are the pre-images of $\mathbf{y}$, and T_i are one-one transformations of some neighbourhood of $\mathbf{x}_i$ to a some neighbourhood of $\mathbf{y}$.

EXAMPLE 29. Suppose we are given that X_1 and X_2 are independent and each has Exp(λ) distribution. What is the distribution of the random variable $X_1 + X_2$?

The change of variable formula works for transformations from $\mathbb{R}^m$ to $\mathbb{R}^m$ whereas here we have two random variables X_1, X_2 and our interest is in one random variable $X_1 + X_2$. To use the change of variable formula, we must introduce an *auxiliary* variable. For example, we take $Y_1 = X_1 + X_2$ and $Y_2 = X_1/(X_1 + X_2)$. Then as in the first example, we find the joint density of (Y_1, Y_2) using the change of variable formula and then integrate out the second variable to get the density of Y_1.

Let us emphasize the point that if our interest is only in Y_1, then we have a lot of freedom in choosing the auxiliary variable. The only condition is that from Y_1 and Y_2 we should be able to recover X_1 and X_2. Let us repeat the same using $Y_1 = X_1 + X_2$ and $Y_2 = X_2$. Then, $T(x_1, x_2) = (x_1 + x_2, x_2)$ maps $\mathbb{R}^2_+$ onto $Q := \{(y_1, y_2) : y_1 > y_2 > 0\}$ in a one-one manner. The inverse function is $T^{-1}(y_1, y_2) = (y_1 - y_2, y_2)$. It is easy to see that $JT^{-1}(y_1, y_2) = 1$ (check!). Hence, by the change of variable formula, the density of (Y_1, Y_2) is given by

$$\begin{aligned} g(y_1, y_2) &= f(y_1 - y_2, y_2) \cdot 1 \\ &= \lambda^2 e^{-\lambda(y_1 - y_2)} e^{-\lambda y_2} \text{ (if } y_1 > y_2 > 0) \\ &= \lambda^2 e^{-\lambda y_1} \mathbf{1}_{y_1 > y_2 > 0}. \end{aligned}$$

To get the density of Y_1, we integrate out the second variable. The density of Y_1 is

$$\begin{aligned} h(u) &= \int_{-\infty}^{\infty} \lambda^2 e^{-\lambda y_1} \mathbf{1}_{y_1 > y_2 > 0} dy_2 \\ &= \lambda^2 e^{-\lambda y_1} \int_0^{y_1} dy_2 \\ &= \lambda^2 y_1 e^{-\lambda y_1} \end{aligned}$$

which agrees with what we found before.

EXAMPLE 30. Suppose $R \sim$ Exp(λ) and $\Theta \sim$ Unif$(0, 2\pi)$ and the two are independent. Define $X = \sqrt{R}\cos(\Theta)$ and $Y = \sqrt{R}\sin(\Theta)$. We want to find the distribution of (X, Y). For this, we first write the joint density of (R, Θ) which is given by

$$f(r, \theta) = \frac{1}{2\pi} \lambda e^{-\lambda r} \quad \text{for } r > 0, \theta \in (0, 2\pi).$$

Define the transformation $T : \mathbb{R}_+ \times (0, 2\pi) \to \mathbb{R}^2$ by $T(r, \theta) = (\sqrt{r}\cos\theta, \sqrt{r}\sin\theta)$. The image of T consists of all $(x, y) \in \mathbb{R}^2$ with $y \neq 0$. The inverse is $T^{-1}(x, y) = (x^2 + y^2, \arctan(y/x))$ where $\arctan(y/x)$ is defined so as to take values in $(0, \pi)$ when $y > 0$ and to take values in $(\pi, 2\pi)$ when $y < 0$. Thus

$$JT^{-1}(x, y) = \det \begin{bmatrix} 2x & 2y \\ \frac{-y}{x^2+y^2} & \frac{x}{x^2+y^2} \end{bmatrix} = 2.$$

Therefore, (X, Y) has joint density

$$g(x,y) = 2f(x^2+y^2, \arctan(y/x)) = \frac{\lambda}{\pi} e^{-\lambda(x^2+y^2)}.$$

This is for $(x,y) \in \mathbb{R}^2$ with $y \neq 0$, but as we have remarked earlier, the value of a pdf in $\mathbb{R}^2$ on a line does not matter, we may define $g(x,y)$ as above for all (x,y) (main point is that the CDF does not change). Since $g(x,y)$ separates into a function of x and a function of y, X, Y are independent $N(0, \frac{1}{2\lambda})$.

REMARK 31. Relationships between random variables derived by the change of variable formulas can be used for simulation too. For instance, the CDF of $N(0,1)$ is not explicit and hence simulating from that distribution is difficult (must resort to numerical methods). However, we can easily simulate it as follows. Simulate an Exp(1/2) random variable R (easy, as the distribution function can be inverted) and simulate an independent Unif$(0, 2\pi)$ random variable Θ. Then set $X = \sqrt{R}\cos(\Theta)$ and $Y = \sqrt{R}\sin(\Theta)$. These are two independent $N(0,1)$ random numbers. Here it should be noted that the random numbers in $(0,1)$ given by a random number generator are supposed to be independent uniform random numbers (otherwise, it is not acceptable as a random number generator).

5. Applications of the change of variable formula

5.1. Beta integral. Let $X \sim \text{Gamma}(a,1)$ and $Y \sim \text{Gamma}(b,1)$ be independent random variables and let $U = X + Y$ and $V = \frac{X}{X+Y}$. The transformation $T(x,y) = (x+y, \frac{x}{x+y})$ is the same as the one we used in the example in the previous section, so by the Jacobian determinant calculation there, we know that the density of (U,V) is given by

$$g(u,v) = f(uv, u(1-v))u = \frac{1}{\Gamma(a)} e^{-uv}(uv)^{a-1} \times \frac{1}{\Gamma(b)} e^{-u(1-v)}(u(1-v))^{b-1} \times u$$

$$= \frac{1}{\Gamma(a+b)} e^{-u} u^{a+b-1} \times \frac{1}{B(a,b)} v^{a-1}(1-v)^{b-1} \times \frac{\Gamma(a+b)B(a,b)}{\Gamma(a)\Gamma(b)}$$

for $u > 0$ and $0 < v < 1$. The first factor is the Gamma$(a+b)$ density. The second factor is the Beta(a,b) density. This implies several things at once:

(1) The third factor must be 1, since $g(u,v)$ must be a density and hence give 1 when integrated over $u > 0$ and $0 < v < 1$, but the first two factors also integrate to 1 when integrated separately over u and v respectively. Therefore, $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ establishing the relationship between the Beta and Gamma functions.

(2) $U \sim \text{Gamma}(a+b, 1)$. Thus, the sum of independent Gamma variables with the same scale parameter 1 is also a Gamma distribution with scale parameter 1. The shape parameters just add. This remains valid if we change the scale parameter to λ (but it must be the same for both).

(3) $V \sim \text{Beta}(a,b)$. This is the most common way in which Beta distribution occurs.

(4) U and V are independent.

5.2. Simulating Gaussian random variables. Let X, Y be independent $N(0,1)$ random variables and let (R, Θ) be the polar co-ordinates of (X, Y). This means that $R \geq 0$ and $0 \leq \Theta < 2\pi$ and $X = R\cos\Theta$ and $Y = R\sin\Theta$. The Jacobian determinant of the transformation $(x, y) \mapsto (r, \theta)$ is r. The joint density of (X, Y) is $f(x, y) = \frac{1}{2\pi}e^{-(x^2+y^2)/2}$. From the change of variable formula we see that (R, Θ) has joint density

$$g(r, \theta) = \frac{1}{2\pi}e^{-(r^2\cos^2\theta + r^2\sin^2\theta)/2}r \;=\; \frac{1}{2\pi}e^{-\frac{1}{2}r^2}r$$

on $(r, \theta) \in (0, \infty) \times [0, 2\pi)$. Observe that $h(r) = e^{-r^2/2}r$ is a density on $(0, \infty)$ (in fact it is the density of the square-root of an Exp(1/2) distribution). If $k(\theta) = 1/2\pi$, then $g(r, \theta) = h(r)k(\theta)$, hence R and Θ are independent, $R^2 \sim \text{Exp}(1/2)$ and $\Theta \sim \text{Unif}(0, 2\pi)$.

This computation has a bonus consequence. Suppose we want to simulate a $N(0,1)$ random variable. We can do it by setting $Z = \Phi^{-1}(U)$ where $U \sim \text{Unif}[0,1]$. But as Φ and Φ^{-1} do not have explicit formulas, so this would have to be done numerically at best. However, using the above computation in reverse, we see that if $R^2 \sim \text{Exp}(1/2)$ and $\Theta \sim \text{Unif}(0, 2\pi)$ are independent, then $R\cos\Theta$ and $R\sin\Theta$ are independent $N(0,1)$ random variables. We can explicitly simulate R, Θ using two independent $U, V \sim \text{Unif}[0,1]$ variables by setting $R = \sqrt{-2\log U}$ and $\Theta = 2\pi V$. Using two uniforms, we get two Gaussians.

5.3. Multivariate Gaussian distribution. Let $Z_1, \ldots, Z_n$ be independent standard Gaussian random variables and let $Z = (Z_1, \ldots, Z_n)^t$ (a column vector). Let $B_{n\times n}$ be a non-singular (invertible) matrix and let μ be an $n \times 1$ column vector. Define $Y = \mu + BZ$, another random $n \times 1$ vector. What is the density of Y? We observe two things:

(1) As B is non-singular, the transformation $T(x) = \mu + Bx$ is a bijection from $\mathbb{R}^n$ to $\mathbb{R}^n$ with inverse $T^{-1}(y) = B^{-1}(y - \mu)$. The Jacobian determinant of $JT^{-1}(y) = \det(B^{-1}) = \frac{1}{\det(B)}$.

(2) The density of Z is $f(x) = \prod_{j=1}^n \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x_j^2} = \frac{1}{(2\pi)^{n/2}}e^{-\frac{1}{2}x^tx}$ as $x^tx = x_1^2 + \ldots + x_n^2$.

Putting these ingredients together, we see that the density of Y is

$$g(y) = \frac{1}{|\det(B)|}f(B^{-1}(y-\mu)) \;=\; \frac{1}{(2\pi)^{\frac{n}{2}}|\det(B)|}e^{-\frac{1}{2}(y-\mu)^t(B^t)^{-1}B^{-1}(y-\mu)}$$

$$= \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{\det(\Sigma)}}e^{-\frac{1}{2}(y-\mu)^t\Sigma^{-1}(y-\mu)}$$

where $\Sigma = BB^t$. Here we simply used the fact that $(B^t)^{-1}B^{-1} = \Sigma^{-1}$ and $\det(\Sigma) = \det(B)^2$.

The distribution of Y is called the n-dimensional multivariate Gaussian distribution with mean vector μ and covariance matrix Σ and the distribution is denoted $N_n(\mu, \Sigma)$.

Justify this terminology by checking that $\mathbb{E}[Y_j] = \mu_j$ and $\text{Cov}(Y_i, Y_j) = \sigma_{i,j}$ (the (i, j) entry of Σ). You can do this from the density g, or directly from the definition of Y as $\mu + BZ$.

Note that the mean can be any vector in $\mathbb{R}^n$, whereas the covariance matrix Σ is of the form BB^t for a non-singular $n \times n$ matrix B. Such matrices are said to be *positive definite*.

6. Summary measures of association

As with unitary distributions, quite often there is no hope of knowing the full joint distribution of a pair of random variables. Like mean and variance summarize a univariate distribution, we want summary measures that capture the dependence or association of two random variables.

Covariance and correlation: Let X, Y be random variables on a common probability space with expectations μ_X, μ_Y and variances σ_X^2, σ_Y^2. The *covariance* of X and Y is defined as $\text{Cov}(X, Y) = \mathbb{E}[(X-\mu_X)(Y-\mu_Y])]$ (this expectation will exist, for reasons given later). Expanding $(X - \mu_X)(Y - \mu_Y])$, check that it can also be written as

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

The *correlation* between X and Y is defined as

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}}.$$

What do they measure?: Note that $(X - \mu_X)(Y - \mu_Y)$ is positive when both X, Y are above average (i.e., $X \geq \mu_X$ and $Y \geq \mu_Y$) or both below average ($X \leq \mu_X$ and $Y \leq \mu_Y$). When one is above average and one is below average, then $(X-\mu_X)(Y-\mu_Y)$ is negative. Hence the positivity of the covariance means that X and Y go together and negativity of covariance means they tend to go in opposite directions (if one is high, the other will be low). Suppose X and Y are the height and weight of a randomly chosen person. If the chosen person is of above average height, we would guess that he/she is also of above average weight. Thus height and weight are positively correlated. On the other hand, if X denotes the lifetime of a person and Y denotes the amount of cigarette he/she smoked, we tend to think that if Y is on the higher side, X will be on the lower side. Smoking and lifetime are negatively correlated.

Observe that if X is height in meters and Y is the weight in kilograms, then $\text{Cov}(X, Y)$ is in units of kg.m. On the other hand, $\text{Corr}(X, Y)$ is dimension free, a pure number (since $\sqrt{Var(X)}$ is in meters and $\sqrt{Var(Y)}$ is in kilograms). Another way to think of it is this: The standardized version of a random variable X is the random variable $X' = \frac{X-\mu_X}{\sigma_X}$. Then X' has zero mean, unit variance, and most importantly, it is dimension free. Check that $\text{Cov}(X', Y') = \text{Corr}(X, Y)$. We shall later see that correlation is a number in $[-1, 1]$. Why do we care so much for a dimension free quantity like correlation?

For example, if you measure height in centimeters and weight in grams, the covariance between the height and weight goes up by a factor of 10^5, although in real terms nothing has changed. This makes the covariance hard to interpret, only its sign conveys something. Further, if we want to compare dependence across different pairs of random variables (for example, which is the preferable method to reduce lifetime? Smoking or Drinking alcohol in large quantities?), then correlation gives a standardized score in the range $[-1, 1]$ that can be interpreted and compared across experiments. A correlation of $+1$ denotes perfect

correlation (e.g., $\mathrm{Corr}(X, X) = 1$) while a correlation of -1 denotes perfect negative association (e.g., $\mathrm{Corr}(X, -X) = -1$).

Correlation is one of the most used and abused quantities in statistics. It is a primary tool that allows us to explore relationships between various quantities. In particular, as we shall see, independent random variables have zero correlation. Though the converse is false, it is the first check we make to see if two variables could be independent of each other. Of course, we don't stop with saying two variables are associated, we explore the causes. This is one of the frequent misuses of statistics. Even if $\mathrm{Corr}(X, Y) = 0.99$, it does not mean that X cause Y or Y causes X. It could be some other Z that causes both. Of course, sometimes the positive correlation could be due to one of the variables causing the other, or even both causing the other! But that conclusion will have to come from something more than just the correlation coefficient.

Properties of covariance and variance: Let X, Y, X_i, Y_i be random variables on a common probability space. Small letters a, b, c etc will denote scalars.

(1) (Bilinearity): $\mathrm{Cov}(aX_1 + bX_2, Y) = a\mathrm{Cov}(X_1, Y) + b\mathrm{Cov}(X_2, Y)$ (linearity in the first variable) and $\mathrm{Cov}(Y, aX_1 + bX_2) = a\mathrm{Cov}(Y, X_1) + b\mathrm{Cov}(Y, X_2)$ (linearity in the second variable).

(2) (Symmetry): $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$.

(3) (Positive semi-definitenss): $\mathrm{Cov}(X, X) \geq 0$ with equality if and only if X is a constant random variable. Indeed, $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$.

Exercise 32. Show that $\mathrm{Var}(cX) = c^2\mathrm{Var}(X)$ (hence $\mathrm{sd}(cX) = |c|\mathrm{sd}(X)$). Further, if X and Y are independent, then $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

Note that the properties of covariance are very much like properties of inner-products in vector spaces. The only difference is in that $\mathrm{Cov}(X, X) = 0$ does not mean that $X = 0$, but for inner products, $\mathbf{v} \cdot \mathbf{v} = 0$ implies $\mathbf{v} = 0$. For vectors, we know the Cauchy-Schwarz inequality: $(\mathbf{u} \cdot \mathbf{v})^2 \leq (\mathbf{u} \cdot \mathbf{u})(\mathbf{v} \cdot \mathbf{v})$, with equality if and only $\mathbf{u}, \mathbf{v}$ are linearly dependent, i.e., $a\mathbf{u} + b\mathbf{v} = 0$ for some $a, b \in \mathbb{R}$.

Cauchy-Schwarz inequality: If X and Y are random variables with finite variances, then $(\mathrm{Cov}(X, Y))^2 \leq \mathrm{Var}(X)\mathrm{Var}(Y)$ with equality if and only if $aX + bY + c = 0$ for some $a, b, c \in \mathbb{R}$.

One can of course repeat the proof of Cauchy-Schwarz inequality in this notation: For any $t \in \mathbb{R}$, as $X - tY$ has positive variance, we get

$$0 \leq \mathrm{Cov}(X - tY, X - tY) = \mathrm{Var}(X) - 2t\mathrm{Cov}(X, Y) + t^2\mathrm{Var}(Y).$$

As this is a quadratic expression in t, it follows that $\mathrm{Cov}(X, Y)^2 \leq \mathrm{Var}(X)\mathrm{Var}(Y)$ (as $at^2 + bt + c \geq 0$ for all t if and only if $b^2 \leq 4ac$). Taking square roots, we may rewrite this as

$$-1 \leq \mathrm{Corr}(X, Y) \leq 1.$$

The correlation coefficient is one of the great discoveries in statistics (less than two hundred years old!), one that is highly used and misused. One advantage of it compared to the covariance is that it is a pure number, and its range is $[-1, 1]$, with the extremes being attained when $Y = \pm X$. Independent variables have zero correlation. The converse is spectacularly false! Nevertheless, correlation coefficient is the first thing we measure to find the relationship between two variables. The greatest of troubles comes in what to interpret the result as! In particular, interpreting correlation as causation is the single biggest error one commits in statistics.

EXAMPLE 33. Suppose the exam scores (Y) and attendance in class (X) have a correlation of 0.8. The immediate conclusion one might come to is that poor attendance causes poor performance and hence recommend mandatory attendance of classes. As $\mathrm{Corr}(X, Y) = \mathrm{Corr}(Y, X)$, there is no way that the number itself can say whether X cause Y or Y causes X! Perhaps students have an idea of how well they are going to perform, and students who do not expect to perform well just decide to skip classes anyway. That would mean that Y is causing X. A third alternative is that there is yet another factor Z that is causing both X and Y. For example, the level of interest in a subject may determine attendance in class as well as the performance in exams. If so, there is no causal relationship between X and Y directly.

7. Exercises

PROBLEM 1. Let A, B be two events in a common probability space. Write the joint distributions (joint pmf) of the following random variables.

(1) $X = \mathbf{1}_A$ and $Y = \mathbf{1}_B$.

(2) $X = \mathbf{1}_{A\cap B}$ and $Y = \mathbf{1}_{A\cup B}$.

PROBLEM 2. Let $a > 0, b > 0$ and $ab > c^2$. Let (X, Y) have the bivariate normal distribution with density

$$f(x, y) = \frac{\sqrt{ab - c^2}}{2\pi} e^{-\frac{1}{2}\left[a(x-\mu)^2 + b(y-\nu)^2 + 2c(x-\mu)(y-\nu)\right]}.$$

Show that the marginal distributions are one-dimensional normal and find the parameters. For what values of the parameters are X and Y independent?

PROBLEM 3. Fix $r > 0$. Let (X, Y) be a random vector with density

$$f(x, y) = \begin{cases} \frac{1}{\pi r^2} & \text{if } x^2 + y^2 \le r^2, \\ 0 & \text{otherwise.} \end{cases}$$

This models the experiment of drawing a point at random from a disk of radius r centered at $(0, 0)$.

(1) Find the marginal densities of X and Y (i.e., find the density of X and find the density of Y separately).

(2) Can you solve the same problem if the point is drawn uniformly from the ellipse $\{(x, y) : \frac{x^2}{a^2} + \frac{y^2}{b^2} \le 1\}$?

PROBLEM 4. Let $\mathbf{X} = (X_1, \ldots, X_{n-1})$ be a Multinomial random variable with parameters $r, n, p_1, \ldots, p_n$ where r, n are positive integers and p_i are non-negative numbers that sum to 1. This means that $\mathbf{X}$ has pmf

$$f(k_1, \ldots, k_{n-1}) = \frac{n!}{k_1!k_2! \ldots k_{n-1}!(r - k_1 - \ldots - k_{n-1})!} p_1^{k_1} \ldots p_{n-1}^{k_{n-1}} p_n^{r-k_1-\ldots-k_{n-1}}$$

if $k_i \geq 0$ are integers that add to at most r.

(1) Let $m \leq n$. Show that the distribution of $(X_1, \ldots, X_{m-1})$ is Multinomial with parameters $r, m, \tilde{p}_1, \ldots, \tilde{p}_m$ where $\tilde{p}_i = p_i$ for $i \leq m-1$ and $\tilde{p}_m = p_m + \ldots + p_n$.

(2) The distribution of X_k is $\text{Bin}(r, p_k)$.

(3) (Do not need to submit this) Let $k_1 < k_2 < \ldots < k_m = n$. Define $Y_1 = X_1 + \ldots + X_{k_1-1}$, $Y_2 = X_{k_1} + \ldots + X_{k_2-1}, \ldots Y_m = X_{k_{m-1}} + \ldots + X_{k_m-1}$. What is the distribution of $(Y_1, \ldots, Y_m)$?

[**Note** Remember the balls-in-bins interpretation of Multinomial. Based on it, try to guess the answers before you start calculating anything!].

PROBLEM 5. Let r balls be placed in m bins at random. Let X_k be the number of balls in the k^{th} bin. Recall that $(X_1, \ldots, X_m)$ has a multinomial distribution. Find the joint distribution of (X_1, X_2) and the marginal distribution of X_1 and of X_2.

PROBLEM 6 (*Submit only parts (1) and (2)*). (1) Let X and Y be independent integer-valued random variables with pmf f and g respectively. That is, $\mathbb{P}\{X = k\} = f(k)$ and $\mathbb{P}\{Y = k\}g(k)$ for every $k \in \mathbb{Z}$. Then, show that $X + Y$ has the pmf h given by $h(k) = \sum_{n \in \mathbb{Z}} f(n)g(k-n)$ for each $k \in \mathbb{Z}$.

(2) Let $X \sim \text{Pois}(\lambda)$ and $Y \sim \text{Pois}(\mu)$ and assume that X and Y are independent. Show that $X + Y \sim \text{Pois}(\lambda + \mu)$.

(3) Let $X \sim \text{Bin}(n, p)$ and $Y \sim \text{Bin}(m, p)$ and assume that X and Y are independent. Show that $X + Y \sim \text{Bin}(n + m, p)$.

(4) Let $X \sim \text{Geo}(p)$ and $Y \sim \text{Geo}(p)$ and assume that X and Y are independent. Show that $X + Y$ has negative binomial distribution and find the parameters.

PROBLEM 7 (*Submit only parts (1) and (2)*). (1) Let X and Y be independent random variables with densities $f(x)$ and $g(y)$ respectively. Use the change of variable formula to show that $X + Y$ has the density $h(u)$ given by $h(u) = \int_{-\infty}^{\infty} f(s)g(u-s)ds$.

(2) Let X, Y be independent $\text{Unif}[-1, 1]$ random variables. Find the density of $X + Y$.

(3) Let $X \sim \text{Gamma}(\mu, \lambda)$ and $Y \sim \text{Gamma}(\nu, \lambda)$ and assume that X and Y are independent. Show that $X + Y \sim \text{Gamma}(\mu + \nu, \lambda)$.

(4) Let $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ and assume that X and Y are independent. Show that $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

PROBLEM 8. In each of the following cases, X and Y are independent random variables with the given distributions. You are asked to find the distribution of $X + Y$ using the convolution formula (when you encounter a named distribution, do identify it!).

(1) $X \sim \text{Gamma}(\nu, \lambda)$ and $Y \sim \text{Gamma}(\nu', \lambda)$.

(2) $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$.

(3) $X \sim \text{Pois}(\lambda)$ and $Y \sim \text{Pois}(\lambda')$.

(4) $X \sim \text{Geo}(p)$ and $Y \sim \text{Geo}(p)$.

[**Note:** Submit 2, 3, 4 only]

PROBLEM 9. Use the change of variable formula to solve the following problems.

(1) Let $X \sim \text{Pois}(\lambda)$ and $Y \sim \text{Pois}(\lambda')$ be independent. Let $Z = X + Y$. Show that the conditional distribution of X given $Z = m$ is $\text{Bin}(m, \frac{\lambda}{\lambda+\lambda'})$.

(2) Let $X \sim \text{Gamma}(\nu, \lambda)$ and $Y \sim \text{Gamma}(\nu', \lambda)$ be independent. Show that $X/(X + Y)$ has a Beta distribution and find its parameters.

(3) Let $X \sim N(0, 1)$ and $Y \sim N(0, 1)$ be independent. Show that X/Y has Cauchy distribution.

PROBLEM 10. Let (X, Y) be a bivariate normal with density

$$\frac{\sqrt{ab - c^2}}{2\pi} e^{-\frac{1}{2}(a(x-\mu)^2 + b(y-\nu)^2 + 2c(x-\mu)(y-\nu))}$$

where $a, b, ab - c^2$ are all positive and μ, ν are any real numbers.

(1) Show that $\mathbb{E}[X] = \mu$, $\mathbb{E}[Y] = \nu$, $\text{Var}(X) = \sigma_{1,1}$, $\text{Var}(Y) \sim \sigma_{2,2}$ and $\text{Cov}(X, Y) = \sigma_{1,2}$ where the matrix Σ (called the covariance matrix of (X, Y)) is defined as

$$\Sigma = \begin{bmatrix} \sigma_{1,1} & \sigma_{1,2} \\ \sigma_{2,1} & \sigma_{2,2} \end{bmatrix} := \begin{bmatrix} a & c \\ c & b \end{bmatrix}^{-1}.$$

(2) Find the conditional density of Y given X. When are X and Y independent? ("When" means under what conditions on the parameters a, b, c, μ, ν or in terms of $\sigma_{1,1}, \sigma_{2,2}, \sigma_{1,2}, \mu, \nu$?).

PROBLEM 11. Let X_1, X_2, X_3 be independent random variables, each having $\text{Ber}_{\pm}(1/2)$ distribution. This means $\mathbb{P}(X_1 = 1) = \mathbb{P}(X_1 = -1) = \frac{1}{2}$.

(1) Let $Y_1 = X_2X_3$, $Y_2 = X_1X_3$ and $Y_3 = X_1X_2$. Show that Y_1, Y_2, Y_3 are pairwise independent (i.e., any two of them are independent) but are not independent.

(2) Can you find three events A, B, C in some probability space such that they are pair-wise independent but not independent?

PROBLEM 12. (1) Let X_1, X_2 be independent random variables, both having $\text{Exp}(\lambda)$ distribution. Let $Z = \min\{X_1, X_2\}$. Show that $Z \sim \text{Exp}(2\lambda)$. What if we take the minimum of n independent exponential random variables?

Problem 13. Let A, B be two events in a common probability space. Write the joint distributions (joint pmf) of the following random variables.

(1) $X = \mathbf{1}_A$ and $Y = \mathbf{1}_B$.

(2) $X = \mathbf{1}_{A\cap B}$ and $Y = \mathbf{1}_{A\cup B}$.

Problem 14. (1) Let $X \sim \text{Exp}(\lambda)$. For any $t, s > 0$, show that $\mathbb{P}\{X > t+s \,\big|\, X > t\} = \mathbb{P}\{X > s\}$. (This is called the *memoryless property* of the exponential distribution).

(2) Show that if a non-negative random variable Y has memoryless property (i.e., $\mathbb{P}\{Y > t+s \,\big|\, Y > t\} = \mathbb{P}\{Y > s\}$ for all $s, t > 0$), then Y must have exponential distribution.

Problem 15. (1) Let X and Y be independent integer-valued random variables with pmf f and g respectively. That is, $\mathbb{P}\{X = k\} = f(k)$ and $\mathbb{P}\{Y = k\}g(k)$ for every $k \in \mathbb{Z}$. Then, show that $X + Y$ has the pmf h given by $h(k) = \sum_{n\in\mathbb{Z}} f(n)g(k-n)$ for each $k \in \mathbb{Z}$.

(2) Let $X \sim \text{Pois}(\lambda)$ and $Y \sim \text{Pois}(\mu)$ and assume that X and Y are independent. Show that $X + Y \sim \text{Pois}(\lambda + \mu)$.

Problem 16. Continuation of the previous problem.

(1) Let $X \sim \text{Bin}(n, p)$ and $Y \sim \text{Bin}(m, p)$ and assume that X and Y are independent. Show that $X + Y \sim \text{Bin}(n + m, p)$.

(2) Let $X \sim \text{Geo}(p)$ and $Y \sim \text{Geo}(p)$ and assume that X and Y are independent. Show that $X + Y$ has negative binomial distribution and find the parameters.

Problem 17. (1) Let X and Y be independent random variables with densities $f(x)$ and $g(y)$ respectively. Use the change of variable formula to show that $X + Y$ has the density $h(u)$ given by $h(u) = \int_{-\infty}^{\infty} f(s)g(u-s)ds$.

(2) Let X, Y be independent Unif$[-1, 1]$ random variables. Find the density of $X + Y$.

Problem 18. Continuation of the previous problem.

(1) Let $X \sim \text{Gamma}(\mu, \lambda)$ and $Y \sim \text{Gamma}(\nu, \lambda)$ and assume that X and Y are independent. Show that $X + Y \sim \text{Gamma}(\mu + \nu, \lambda)$.

(2) Let $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ and assume that X and Y are independent. Show that $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

Problem 19. Let (X, Y) have the bivariate normal distribution with density

$$f(x, y) = \frac{\sqrt{ab - c^2}}{2\pi} e^{-\frac{1}{2}[ax^2+by^2+2cxy]}.$$

Assume that $a > 0, c > 0, ab - c^2 > 0$ so that this is a valid density.

(1) Show that the marginal distributions are one-dimensional normal and find the parameters.

(2) For what values of the parameters are X and Y independent?

PROBLEM 20. A few more exercises in change of variable formula.

(1) If X, Y are independent $N(0,1)$ random variables, show that X/Y has the Cauchy distribution (with density $\frac{1}{\pi(1+x^2)}$).

(2) If $X \sim \text{Gamma}(\alpha, 1)$, $Y \sim \text{Gamma}(\beta, 1)$ are independent, then show that $X+Y$ and $X/(X+Y)$ are independent, $X+Y \sim \text{Gamma}(\alpha+\beta, 1)$ and $X/(X+Y) \sim \text{Beta}(\alpha, \beta)$.

(3) If X, Y are independent $N(0,1)$ random variables, show that X^2+Y^2 has Exp(1/2) distribution.

PROBLEM 21. Find all possible joint distributions of (X, Y) such that $X \sim \text{Ber}(1/2)$ and $Y \sim \text{Ber}(1/2)$. Find the correlation for each such joint distribution.

PROBLEM 22. Let (X, Y) have the bivariate normal distribution with density

$$f(x,y) = \frac{\sqrt{ab-c^2}}{2\pi} e^{-\frac{1}{2}\left[a(x-\mu)^2+b(y-\nu)^2+2c(x-\mu)(y-\nu)\right]}.$$

(1) Find the marginal distributions of X and of Y.

(2) Find means and variances of X and Y and the covariance and correlation of X with Y. Under what conditions on the parameters are X and Y independent?

[**Note:** It is very useful to introduce the matrix $\Sigma = \begin{bmatrix} a & c \\ c & b \end{bmatrix}^{-1} = \begin{bmatrix} \frac{b}{\Delta} & -\frac{c}{\Delta} \\ -\frac{c}{\Delta} & \frac{a}{\Delta} \end{bmatrix}$ which is called the *covariance matrix* of (X, Y). The answers can be written in terms of the entries of Σ.]

PROBLEM 23. Let r balls be placed in m bins at random. Let X_k be the number of balls in the k^{th} bin. Recall that $(X_1, \ldots, X_m)$ has a multinomial distribution.

(1) Find the joint distribution of (X_1, X_2) and the marginal distribution of X_1 and of X_2.

(2) Find the means, variances, covariance and correlation of X_1 and X_2.

(3) Let Y be the number of empty bins. Find the mean and variance of Y. [**Hint:** Write Y as $\mathbf{1}_{A_1} + \ldots + \mathbf{1}_{A_m}$ where A_k is the event that the k^{th} bin is empty].

PROBLEM 24. A box contains N coupons where the number w_k is written on the k^{th} coupon. Let $\mu = \frac{1}{N}\sum_{k=1}^{N} w_k$ be the "population mean" and let $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(w_i-\mu)^2$ be the "population variance". A sample of size m is drawn from the population, the values seen are $X_1, \ldots, X_m$. The sample mean $\bar{X}_m = (X_1 + \ldots + X_m)/m$ is formed. Find the mean and variance of $\bar{X}_m$ in the following two cases.

(1) The samples are drawn with replacement (i.e., draw a coupon, note the number, put the coupon back in the box, and draw again...).

(2) The samples are drawn without replacement.

PROBLEM 25. Place r balls in n bins uniformly at random. Let X_k be the number of balls in the k^{th} bin. Find $\mathbb{E}[X_k]$, $\text{Var}(X_k)$ and $\text{Cov}(X_k, X_\ell)$ for $1 \leq k, \ell \leq n$. [**Hint:** First do the case when $r = 1$. Then think how to use that to get the general case].

PROBLEM 26. Suppose X, Y, Z are i.i.d. random variables with each having marginal density $f(t)$.

(1) Find $\mathbb{E}\left[\frac{X}{X+Y+Z}\right]$ (assume that it exists).

(2) Find $\mathbb{P}(X < Y > Z)$.

PROBLEM 27. Recall the problem of a psychic guessing cards. Consider a shuffled deck of n cards and a psychic is supposed to guess the order of cards. Let M_n be the number of correct guesses.

(1) Assuming random guessing by the psychic, show that $\mathbb{E}[M_n] = 1$ and $\text{Var}(M_n) = 1$. [**Hint** Write M_n as $X_1 + \ldots + X_n$ where X_k is the indicator of the event that the k^{th} card is guessed correctly].

(2) Consider a variant of the game where the cards are dealt one by one and before each card is dealt, the psychic guesses what card it is going to be. In this case find $\mathbb{E}[M_n]$ and $\text{Var}(M_n)$.

PROBLEM 28. Place r balls in n bins uniformly at random. Let X_k be the number of balls in the k^{th} bin. Find $\mathbb{E}[X_k]$, $\text{Var}(X_k)$ and $\text{Cov}(X_k, X_\ell)$ for $1 \leq k, \ell \leq n$. [**Hint:** First do the case when $r = 1$. Then think how to use that to get the general case].

PROBLEM 29. Let X be a non-negative random variable with CDF $F(t)$.

(1) Show that $\mathbb{E}[X] = \int_0^\infty (1 - F(t))dt$ and more generally $\mathbb{E}[X^p] = \int_0^\infty pt^{p-1}(1 - F(t))dt$. [**Hint:** In showing this, you may assume that X has a density if you like, but it is not necessary for the above formulas to hold true]

(2) If X is non-negative integer valued, then $\mathbb{E}[X] = \sum_{k=1}^\infty \mathbb{P}\{X \geq k\}$.

PROBLEM 30. A deck consists of cards labelled $1, 2, \ldots, N$. The deck is shuffled well. Let X be the label on the first card and let Y be the label on the second card. Find the means and variances of X and Y and the covariance of X and Y.

PROBLEM 31. A box contains N coupons labelled $1, 2, \ldots, N$. A sample of size m is drawn from the population and the sample average $\bar{X}_m$ is computed. Find the mean and standard deviation of $\bar{X}_m$ in both the following cases.

(1) The m coupons are drawn with replacement.

(2) The m coupons are drawn without replacement (in this case, assume $m \leq N$).

PROBLEM 32. What is the maximum number n of random variables $X_1, \ldots, X_n$ that one can construct on some common probability space so that $\text{Var}(X_k) = 1$ for all k and $\text{Cov}(X_j, X_k) = -\alpha$ for all $j \neq k$? Here $\alpha > 0$ is given and the answer depends on α.

PROBLEM 33. Let $X \sim N(0,1)$. Although it is not possible to get an exact expression for the CDF of X, show that for any $t > 0$,

$$\mathbb{P}\{X \geq t\} \leq \frac{1}{\sqrt{2\pi}} \frac{e^{-t^2/2}}{t}$$

which shows that the tail of the CDF decays rapidly. [**Hint:** Use the idea used in the proof of Markov's inequality]

PROBLEM 34. Show that it is not possible to construct X, Y on the same probability space where X is a uniform random subset of $[n]$, Y is a uniform random subset of $[n+1]$, and $X \subseteq Y$.

PROBLEM 35. Show that it possible to construct X, Y on the same probability space where X is a uniform random subset of $[n]$, Y is a uniform random subset of $[n+1]$, and $X \Delta Y$ has at most two elements.

CHAPTER 8

Limit theorems

1. Weak law of large numbers

Let $X_1, X_2, \ldots$ be i.i.d random variables (independent random variables each having the same marginal distribution). Assume that the second moment of X_1 is finite. Then, $\mu = \mathbb{E}[X_1]$ and $\sigma^2 = \mathrm{Var}(X_1)$ are well-defined.

Let $S_n = X_1 + \ldots + X_n$ (partial sums) and $\overline{X}_n = \frac{S_n}{n} = \frac{X_1+\ldots+X_n}{n}$ (*sample mean*). Then, by the properties of expectation and variance, we have

$$\mathbb{E}[S_n] = n\mu, \;\; \mathrm{Var}(S_n) = n\sigma_1^2, \qquad \mathbb{E}[\overline{X}_n] = \mu, \;\; \mathrm{Var}(\overline{X}_n) = \frac{\sigma^2}{n}.$$

In particular, $\mathrm{s.d.}(\overline{X}_n) = \sigma/\sqrt{n}$ decreases with n, not inversely with n but inversely as the square root of n. If we apply Chebyshev's inequality to $\overline{X}_n$, we get for any $\delta > 0$ that

$$\mathbb{P}\{|\overline{X}_n - \mu| \geq \delta\} \leq \frac{\sigma^2}{\delta^2 n}. \tag{17}$$

This goes to zero as $n \to \infty$ (with $\delta > 0$ being fixed). This means that for large n the sample mean is unlikely to be far from μ (sometimes called "population mean"). This is consistent with our intuitive idea that if we toss a p-coin many times, the proportion of heads gives a good guess of what the value of p is. We summarize this below.

THEOREM 36 (Weak law of large numbers). *With the above notations, for any $\delta > 0$, we have*

$$\mathbb{P}\{|\overline{X}_n - \mu| \geq \delta\} \to 0 \quad \text{as } n \to \infty.$$

It suffices to assume that X_k have finite mean (no need for finite variance) for WLLN to hold. However, for such random variables, there is no analogue of the quantitative bound (17). Although the limiting statement has the virtue of connecting with our intuitive notion of probability and expectation, the quantitative bound is far more useful. In fact, it is worth making more assumptions and getting better bounds. We just mention one such result without proof.

PROPOSITION 37 (Hoeffding's inequality). *Assume that $X_1, \ldots, X_n$ are independent, $\mathbb{E}[X_k] = \mu$ for all k, and $|X_k - \mu| \leq B$ w.p.1. for each k. Then*

$$\mathbb{P}\{|\overline{X}_n - \mu| \geq \delta\} \leq 2e^{-\frac{n\delta^2}{2B^2}} \tag{18}$$

Compared to (17) which decays like $1/n$, the right side of (18) decays exponentially fast It gives better bounds for deviation probabilities, but under more restrictive assumptions.

EXAMPLE 38. If a fair coin is tossed 1000 times, then X_k are i.i.d. Ber(1/2). So $\mu = \frac{1}{2}$, $\sigma^2 = \frac{1}{4}$ and $B = \frac{1}{2}$. Let us compute the probability that the number of heads is between 450 and 550, which is the same as $0.45 \leq \overline{X}_n \leq 0.55$. According to Chebshev's inequality,

$$\mathbb{P}\{|\overline{X}_n - 0.5| > 0.05\} \leq \begin{cases} \frac{1/4}{(0.05)^2 \times 1000} = 0.1 & \text{if we apply Chebyshev's inequality,} \\ 2\exp\left\{-\frac{1000\times(0.05)^2}{2\times(1/2)^2}\right\} = 0.013 & \text{if we apply Hoeffding's inequality.} \end{cases}$$

The actual probability (one can use a computer to find it, or central limit theorem that we shall learn next) is in fact 0.0014.

Reversing this question, we can find out how many samples are needed to achieve a desired accuracy with a specified bound on the probability of error.

2. Application: Sample complexity

Let us ask how large n must be, so that we are fairly sure (say with probability $1-\gamma = 0.99$) that $\overline{X}_n$ lies close to μ (say within $\delta = 0.05$ of it)? This means that in (17) we take $\delta = 0.05$ and ask for the right side bound to be smaller than $1 - \gamma = 0.01$. This means that it suffices to take

$$n \geq \frac{\sigma^2}{\delta^2\gamma}. \tag{19}$$

Often we don't know σ^2 exactly, we know a bound for it like $\sigma^2 \leq M$, then we take $n \geq \frac{M}{\delta^2\gamma}$. This requirement of sample size is known as sample complexity. Okay, strictly speaking, what we have shown is that $\frac{M}{\delta^2\gamma}$ samples suffice. Is it the best we can do?

(1) In terms of the accuracy, we cannot do better, i.e., sample complexity depends on δ in proportion to $1/\delta^2$. In particular, to increase accuracy and decrease δ to $\delta/2$, the sample size must go up by a factor of 4.

To see this, take $X_k \sim N(0,1)$, then $\overline{X}_n \sim N(0, \frac{1}{n})$, hence $\sqrt{n}\overline{X}_n \sim N(0,1)$. Therefore, $\mathbb{P}\{\overline{X}_n \geq \delta\} = 1 - \Phi(\delta\sqrt{n})$. If $n < \frac{q_\gamma^2}{\delta^2}$ where q_γ is the $(1-\gamma)$-quantile of $N(0,1)$ distribution, then this probability is at least $\mathbb{P}\{|\overline{X}_n| \leq \delta\} \leq \mathbb{P}\{\overline{X}_n \leq \delta\} < 1 - \gamma$. Thus we need $n \geq \frac{q_\gamma^2}{\delta^2}$.

(2) The dependence on γ is not optimal in general. For example, if we assume that $|X_k - \mu| \leq B$. Then, Hoeffding's inequality (18), we see that $\mathbb{P}\{|\overline{X}_n - \mu| \geq \delta\} \leq \gamma$ provided $n \geq \frac{2B^2 \log\frac{2}{\gamma}}{\delta^2}$. As $\log\frac{2}{\gamma}$ grows much slower than $\frac{1}{\gamma}$ as $\gamma \to 0$, this bound has better dependence on γ.

2.1. A surprise about sample sizes. Agencies such as GallUp conduct surveys on many issues in many countries. Let us imagine a binary question such as "Do you believe humans are responsible for climate change?". If the agency sampled 100 people in Pakistan, how many people should it sample in India (population of 145 crores)? Since the population of India (145 crores in 2025) is about six times the population of Pakistan (25 crores), one might

expect that the sample size should be proportionally (or proportional to square or square root?) larger?

However, the sample complexity computation above shows that the sample size needed is exactly the same as in Pakistan! Indeed, if p is the proportion that answers "Yes" in a country, the sample may be treated as i.i.d. Ber(p). Hence, for a given accuracy $\delta > 0$ and an allowed probability of error γ, the sample required according to (19) is $n \geq \frac{p(1-p)}{\delta^2\gamma}$. This has no dependence on the size of the country's population at all!

3. Appplication: Monte-Carlo integration

In this section we give a simple application of WLLN. Let $\varphi : [0,1] \to \mathbb{R}$ be a continuous function. We would like to compute $I = \int_0^1 \varphi(x)dx$. Most often we cannot compute the integral explicitly and for an approximate value we resort to numerical methods. Here is an idea to use random numbers.

Let $U_1, U_2, \ldots, U_n$ be i.i.d. Unif$[0,1]$ random variables and let $X_1 = \varphi(U_1), \ldots, X_n = \varphi(U_n)$. Then, X_k are i.i.d. random variables with common mean and variance

$$\mu = \int_0^1 \varphi(x)dx = I, \qquad \sigma^2 := \mathrm{Var}(X_1) = \int_0^1 (\varphi(x) - I)^2 dx.$$

This gives the following method of finding I. Fix a large number N appropriately and pick N uniform random numbers U_k, $1 \leq k \leq N$. Then define $\hat{I}_N := \frac{1}{N}\sum_{k=1}^N \varphi(U_k)$. Present $\hat{I}_N$ as an approximate value of I.

In what sense is this an approximation of I and why? Indeed, by WLLN $\mathbb{P}\{|\hat{I}_n - I| \geq \delta\} \to 0$ and hence we expect $\hat{I}_n$ to be close to I.

As this is a practical problem, it is important to know how big n needs to be. Observe that if $|\varphi(x)| \leq B$ for all $x \in [0,1]$, then $|X_k| \leq B$. By the sample complexity analysis above, we see that if $n \geq \frac{2B^2 \log\frac{2}{\gamma}}{\delta^2}$, then $\hat{I}_n$ is within δ if I, with a probability of $1-\gamma$ or more.

Example 39. We know that $\int_0^1 \frac{4}{1+x^2}dx = \pi$. Thus, if U_k are i.i.d. Unif$[0,1]$ random variables, the $X_k = \frac{4}{1+U_k^2}$ and $\hat{I}_n = \frac{4}{n}\sum_{k=1}^n \frac{1}{1+U_k^2}$. The bound on the function is $B = 4$, so if we want a 99% guarantee (means $\gamma = 0.01$) that $\hat{I}_n$ agrees with I to two decimal places (means $\delta = 0.005$), then we need $n \geq \frac{8\log(200)}{(0.005)^2} \approx 17000$. This is certainly possible on a computer. But if we want three decimal places, then we need 1700000 samples, which starts getting to be too much. Here are the results in a few simulations (but you should do your own):

n	10	100	1000	10000	100000
Estimate	3.50569	3.20016	3.1182	3.14611	3.13994

This is not the only way to estimate π by simulation. Another method is to throw darts at the dartboard $[-1,1]^2$ and count the proportion of times the dart falls inside the unit disk. The

answer should be $\pi/4$. Here is the result of some simulations.

n	10	100	1000	10000	100000
Estimate	3.6	3.36	3.148	3.1432	3.13800

EXERCISE 40. Devise some method to find $e = 2.71828\ldots$ by simulation.

Extensions. This method can be used to evaluate integrals over any interval. For instance, our integrals could be of the form $\int_a^b \varphi(t)dt$ or $\int_0^\infty \varphi(t)e^{-t}dt$ or $\int_{-\infty}^\infty \varphi(t)e^{-t^2}dt$ where φ is a function on the appropriate interval. In such cases it makes sense to take advantage of the form of the integral.

For example, if $\varphi : (0,\infty) \to \mathbb{R}$, then $I := \int_0^\infty \varphi(t)dt = \mathbb{E}[\varphi(X)]$ where $X \sim \text{Exp}(1)$. Hence, if $X_1, \ldots, X_n$ are i.i.d. Exp(1) random variables, then

$$\hat{I}_n = \frac{\varphi(X_1) + \ldots + \varphi(X_n)}{n}$$

is a good approximation for I. We leave it as an exercise to figure out the precise statement of approximation using Chebyshev's inequality. Here are the results of a simulation for $\varphi(t) = \sqrt{t}$. Note that in this case $I = \Gamma(3/2) = \frac{1}{2}\sqrt{\pi} = 0.88622\ldots$. Simulations gave:

n	10	100	1000	10000	100000
Estimate	0.96536	0.827891	0.898006	0.883908	0.88761

It can also be used to evaluate multiple integrals (and consequently to find the areas and volumes of sets). The only condition is that it should be possible to evaluate the given function φ at a point x on the computer. To illustrate, consider the problem of finding the area of a region $\{(x,y) : 0 \le x, y, \le 1,\ 2x^3y^2 \ge 1,\ x^2 + 2y^2 \le 2.3\}$. It is complicated to work with such regions analytically, but given a point (x,y), it is easy to check on a computer whether all the constraints given are satisfied.

3.1. Random points versus fixed points? As a last remark, how do Monte-Carlo methods compare with the usual numerical methods? In the latter, usually a number N and a set of points $x_1, \ldots, x_N$ are fixed along with some weights $w_1, \ldots, w_N$ that sum to 1. Then one presents $\tilde{I} := \sum_{k=1}^N w_k \varphi(x_k)$ as the approximate value of I. The simplest one is when the integral is over $[0,1]$ and $x_k = \frac{k}{N}$, $1 \le k \le N$ and $w_k = \frac{1}{N}$ for all k.

There are other choices. For example, Lagrange's method, Gauss quadrature etc are of this type. Under certain assumptions on φ, the accuracy of these methods integrals can be far better than Monte-Carlo.

EXAMPLE 41. Let us again consider $I = \int_0^1 \frac{4}{1+x^2}dx = \int_{-1}^{+1} \frac{2}{1+x^2}dx$. The Gauss-Legendre quadrature with 5 points says that we should take points $x_1 = -x_5 = -0.90618$, $x_2 = -x_4 = -0.538469$, $x_3 = 0$ and weights $w_1 = w_5 = 0.236927$, $w_2 = w_4 = 0.478629$, $w_3 = 0.568889$.

This gives the numerical value of the integral as

$$\tilde{I} = \sum_{i=1}^{5} w_i \frac{4}{1+x_i^2} = 3.14234$$

Observe that it too 10000 points in the Monte-Carlo method to achieve this level of accuracy whereas here we used just 5 points and 5 weights! Even with equal weights $1/N$, we can check that about 200 points suffice to get 2 decimal places accuracy.

Given this, what is the point of Monte-Carlo method? The quadrature methods come with assumptions. But when those assumptions are not satisfied, $\tilde{I}$ can be way off I. One may regard this as a game of strategy as follows. I present a function φ (say bounded between -1 and 1) and you are expected to give an approximation to φ. Quadrature methods do a good job generically, but if I knew the procedure you use, then I can give a function for which your result is entirely wrong (for example, I pick a function φ which vanishes at each of the quadrature points!). However, with Monte-Carlo methods, even if I know the procedure, there is no way to prevent you from getting an approximation of accuracy $1/\sqrt{N}$. This is because neither of us know where the points U_k will fall!

4. Application: Weierstrass' approximation theorem

By the WLLN, if $X_1, \ldots, X_n$ are i.i.d. Ber(p), then $\frac{S_n}{n}$ is close to p with high probability (where $S_n = X_1 + \ldots + X_n$ as usual). Sergei Bernstein observed that this means that if $g : [0,1] \to \mathbb{R}$ is any continuous function, then $g(S_n/n)$ is close to $g(p)$ with high probability. Therefore, he reasoned, it must be the case that $\mathbb{E}[g(S_n/n)]$ must also be close to $g(p)$. But as $S_n \sim \text{Bin}(n,p)$, it follows that

$$\mathbb{E}[g(S_n/n)] = \sum_{k=0}^{n} g(k/n) \binom{n}{k} p^k (1-p)^{n-k}. \tag{20}$$

This is a polynomial in p. And it is close to $g(p)$. This leads us to the

THEOREM 42 (Weierstrass' approximation theorem). *Let $g : [0,1] \to \mathbb{R}$ be a continuous function. Then there exist polynomials B_n such that $\|B_n - g\|_{\sup} \to 0$ as $n \to \infty$.*

Of course, what we described was not quite a proof. For one, we had imprecise "close to", "must also be", etc. Further, the uniformity in Weierstrass' theorem was not addressed. But once the idea is there, it is easy to fix these issues and get an honest proof.

PROOF OF WEIERSTRASS' APPROXIMATION THEOREM. Let $B_n(p)$ be the polynomial on the right side of (20) (for any value of p). By (20) shows that for $p \in [0,1]$ we have $B_n(p) = \mathbb{E}[g(S_n/n)]$ where $S_n = X_1 + \ldots + X_n$ and X_k are i.i.d. Ber(p). Fix any $\delta > 0$ and write

$$\begin{aligned} |B_n(p) - g(p)| &\leq \mathbb{E}[|g(S_n/n) - p|] \\ &= \mathbb{E}\left[|g(S_n/n) - p| \mathbf{1}_{|\frac{S_n}{n} - p| \leq \delta}\right] + \mathbb{E}\left[|g(S_n/n) - p| \mathbf{1}_{|\frac{S_n}{n} - p| > \delta}\right] \end{aligned}$$

Let $\omega_g(\delta) = \sup\{|g(x) - g(y)| : x, y \in [0,1],\ |x-y| \le \delta\}$ and $\|g\| = \sup_{x\in[0,1]} |g(x)|$. Then the first expectation is bounded by $\omega_g(\delta)$. The second expectation is bounded by

$$\mathbb{E}\left[2\|g\|\mathbf{1}_{|\frac{S_n}{n}-p|>\delta}\right] = 2\|g\|\mathbb{P}\left\{|\frac{S_n}{n} - p| \ge \delta\right\} \le 2\|g\|\frac{p(1-p)}{n\delta^2}$$

by Chebyshev's inequality. Putting everything together and using $p(1-p) \le \frac{1}{4}$, we arrive at

$$|B_n(p) - g(p)| \le \omega_g(\delta) + \frac{\|g\|}{2n\delta^2}.$$

The uniform continuity of g on $[0,1]$ means that $\omega_g(\delta) \downarrow 0$ as $\delta \downarrow 0$. Thus given $\varepsilon > 0$, first choose $\delta > 0$ so that the first term is smaller than $\varepsilon/2$ and then choose n large enough so that the second term is smaller than $\varepsilon/2$. Hence $\|B_n - g\|_{\sup} < \varepsilon$. ■

5. Central limit theorem

Let $X_1, X_2, \ldots$ be i.i.d. random variables with expectation μ and variance σ^2. We saw that $\overline{X}_n$ has mean μ and standard deviation $\sigma/\sqrt{n}$.

This roughly means that $\overline{X}_n$ is close to μ, within a few multiples of $\sigma/\sqrt{n}$ (as shown by Chebyshev's inequality). Now we look at $\overline{X}_n$ with a finer microscope. In other words, we ask for the probability that $\overline{X}_n$ is within the tiny interval $[\mu + \frac{a}{\sqrt{n}}, \mu + \frac{b}{\sqrt{n}}]$ for any $a < b$. The answer turns out to be surprising and remarkable!

THEOREM 43 (Central limit theorem). *Let $X_1, X_2, \ldots$ be i.i.d. random variables with expectation μ and variance σ^2. We assume that $0 < \sigma^2 < \infty$. Then, for any $a < b$, we have*

$$\mathbb{P}\left\{\mu + a\frac{\sigma}{\sqrt{n}} \le \overline{X}_n \le \mu + b\frac{\sigma}{\sqrt{n}}\right\} \to \Phi(b) - \Phi(a) = \frac{1}{\sqrt{2\pi}}\int_a^b e^{-t^2/2}dt.$$

What is remarkable about this? The end result does not depend on the distribution of X_is at all! Only the mean and variance of the distribution were used! As this is one of the most important theorems in all of probability theory, we restate it in several forms, all equivalent to the above.

Restatements of central limit theorem: Let X_k be as above. Let $S_n = X_1 + \ldots + X_n$. Let Z be a $N(0,1)$ random variable. Then of course $\mathbb{P}\{a < Z < b\} = \Phi(b) - \Phi(a)$.

(1) $\mathbb{P}\{a < \frac{\sqrt{n}}{\sigma}(\overline{X}_n - \mu) \le b\} \to \Phi(b) - \Phi(a) = \mathbb{P}\{a < Z < b\}$. Put another way, this says that for large n, the random variable $\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}$ has $N(0,1)$ distribution, approximately. Equivalently, $\sqrt{n}(\overline{X}_n - \mu)$ has $N(0, \sigma^2)$ distribution, approximately.

(2) Yet another way to say the same is that S_n has approximately normal distribution with mean $n\mu$ and variance $n\sigma^2$. That is,

$$\mathbb{P}\left\{a \leq \frac{S_n - n\mu}{\sigma\sqrt{n}} \leq b\right\} \to \mathbb{P}\{a < Z < b\}.$$

The central limit theorem is deep and surprising and useful. The following example gives a hint as to why.

EXAMPLE 44. Let $U_1, \ldots, U_n$ be i.i.d. Uniform($[-1,1]$) random variables. Let $S_n = U_1 + \ldots + U_n$, let $\bar{U}_n = S_n/n$ (sample mean) and let $Y_n = S_n/\sqrt{n}$. Consider the problem of finding the distribution of any of these. Since they are got from each other by scaling, finding the distribution of one is the same as finding that of any other. For uniform $[-1,1]$, we know that $\mu = 0$ and $\sigma^2 = 1/3$. Hence, CLT tells us that

$$\mathbb{P}\left\{\frac{a}{\sqrt{3}} < Y_n < \frac{b}{\sqrt{3}}\right\} \to \Phi(b) - \Phi(a).$$

or equivalently, $\mathbb{P}\{a < Y_n < b\} \to \Phi(b\sqrt{3}) - \Phi(a\sqrt{3})$. For large n (practically, $n = 50$ is large enough) we may use this limit as a good aproximation to the probability we want.

Why is this surprising? The way to find the distribution of Y_n would be this. Using the convolution formula n times successively, one can find the density of $S_n = U_1 + \ldots + U_n$ (in principle! the actual integration may be intractable!). Then we can find the density of Y_n by another change of variable (in one dimension). Having got the density of Y_n, we integrate it from a to b to get $\mathbb{P}\{a < Y_n < b\}$. This is clearly a daunting task (if you don't feel so, just try it for $n = 5$).

The CLT cuts short all this and directly gives an approximate answer! And what is even more surprising is that the original distribution does not matter - we only need to know the mean and variance of the original distribution!

6. Reasons for CLT, short of a proof

In a first course like this, it is safe to skip the proof of CLT. Nevertheless, for the curious soul we give a rigorous proof (under third moment assumption) in Appendix **??**. Here we give some explanations that make it seem not too unnatural.

6.1. Fixed point and limit. Let E denote the set of all probability distributions on $\mathbb{R}$. For any distribution F, we define $T(F)$ to be the distribution of the random variable $(V_1 + V_2)/\sqrt{2}$, where V_1, V_2 are i.i.d. random variables with distribution F. Thus $T : E \to E$.

Now fix $X_1, X_2, \ldots$ i.i.d. with zero means and unit variances, and let F_n denote the distribution of $S_n/\sqrt{n}$. Observe that

$$\frac{S_{2n}}{\sqrt{2n}} = \frac{1}{\sqrt{2}}\left(\frac{X_1 + X_3 + \ldots + X_{2n-1}}{\sqrt{n}} + \frac{X_2 + X_4 + \ldots + X_{2n}}{\sqrt{n}}\right)$$

The two summands inside the bracket are i.i.d. with distribution F_n, while $S_{2n}/\sqrt{2n}$ has distribution F_{2n}. Therefore, $F_{2n} = T(F_n)$. Consequently, *if* F_n "converges" (we are being loose

about what it means) to some distribution G, then we must have $T(G) = G$. That is G must be a fixed point of T.

What are all the fixed points of T? If $X_1, X_2 \sim N(0, \sigma^2)$, then $\frac{X_1+X_2}{\sqrt{2}} \sim N(0, \sigma^2)$, hence $N(0, \sigma^2)$ is a fixed point of T for any σ^2. It is not trivial, but can be shown, that these are all the fixed points of T. In particular, if F_n (the distribution of $S_n/\sqrt{n}$) "converges" to some distribution, the limit must be $N(0, \sigma^2)$ for some σ^2. But since $S_n/\sqrt{n}$ has variance 1 for all n, it is reasonable to expect that the limit (if it exists) must be $N(0, 1)$.

6.2. CLT for special distributions. If the common distribution of X_ks is Bernoulli or Poisson or Exponential (and a few others), then we can exactly compute the distribution of S_n. In such cases, we can verify that the CLT holds by explicit computation. This is something of a reassurance, but not much as the usefulness of CLT is highest in those cases where we cannot find the distribution of S_n.

6.2.1. *CLT for Exponential random variables.* Let X_k be i.i.d. Exp(λ). Let X_k be i.i.d. Exp(1) random variables. They have mean $\mu = 1$ and variance $\sigma^2 = 1$. We know that (this was an exercise), $S_n = X_1 + \ldots + X_n$ has Gamma($n, 1$) distribution. Its density is given by $f_n(t) = e^{-t}t^{n-1}/(n-1)!$ for $t > 0$.

Now let $Y_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{S_n - n}{\sqrt{n}}$. By a change of variable (in one-dimension) we see that the density of Y_n is given by $g_n(t) = \sqrt{n} f_n(n + t\sqrt{n})$. Let us analyse this.

$$\begin{aligned}
g_n(t) &= \sqrt{n}\, \frac{1}{(n-1)!} e^{-(n+t\sqrt{n})} (n + t\sqrt{n})^{n-1} \\
&= \sqrt{n}\, \frac{n^{n-1}}{(n-1)!} e^{-n-t\sqrt{n}} \left(1 + \frac{t}{\sqrt{n}}\right)^{n-1} \\
&\approx \sqrt{n}\, \frac{n^{n-1}}{\sqrt{2\pi}(n-1)^{n-\frac{1}{2}} e^{-n+1}} e^{-n-t\sqrt{n}} \left(1 + \frac{t}{\sqrt{n}}\right)^{n-1} \quad \text{(by Stirling's formula)} \\
&= \frac{1}{\sqrt{2\pi}(1-\frac{1}{n})^{n-\frac{1}{2}} e^{1}} e^{-t\sqrt{n}} \left(1 + \frac{t}{\sqrt{n}}\right)^{n-1}.
\end{aligned}$$

To find the limit of this, first observe that $(1 - \frac{1}{n})^{n-\frac{1}{2}} \to e^{-1}$. It remains to find the limit of $w_n := e^{-t\sqrt{n}} \left(1 + \frac{t}{\sqrt{n}}\right)^{n-1}$. Easiest to do this by taking logarithms. Recall that $\log(1 + t) = t - \frac{t^2}{2} + \frac{t^3}{3} - \ldots$. Hence

$$\begin{aligned}
\log w_n &= -t\sqrt{n} + (n-1) \log \left(1 + \frac{t}{\sqrt{n}}\right) \\
&= -t\sqrt{n} + (n-1) \left[\frac{t}{\sqrt{n}} - \frac{t^2}{2n} + \frac{t^3}{3n^{3/2}} - \ldots\right] \\
&= -\frac{t^2}{2} + [\ldots]
\end{aligned}$$

where in $[\ldots]$ we have put all terms which go to zero as $n \to \infty$. Since there are infinitely many, we should argue that even after adding all of them, the total goes to zero as $n \to \infty$.

Let us skip this step and simply conclude that $\log w_n \to -t^2/2$. Therefore, $g_n(t) \to \varphi(t) := \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$ which is the standard normal density.

What we wanted was $\mathbb{P}\{a < Y_n < b\} = \int_a^b g_n(t)dt$. Since $g_n(t) \to \varphi(t)$ for each t, it is believable that $\int_a^b g_n(t)dt \to \int_a^b \varphi(t)dt$. This too needs justification but we skip it. Thus,

$$\mathbb{P}\{a < Y_n < b\} \to \int_a^b \varphi(t)dt = \Phi(b) - \Phi(a).$$

This proves CLT for the case of exponential random variables.

6.2.2. *CLT for Bernoulli random variables.* Let X_k be i.i.d. Ber(1/2) random variables. Then $\mu = \frac{1}{2}$ and $\sigma^2 = \frac{1}{4}$. Hence what we must show is that for $Y_n = \frac{S_n - \frac{n}{2}}{\sqrt{n}/2}$ and any $a < b$,

$$\mathbb{P}\{a \le Y_n \le b\} \to \Phi(b) - \Phi(a).$$

We know that $S_n \sim \text{Bin}(n, \frac{1}{2})$, hence, writing

$$\begin{aligned}\mathbb{P}\{a \le Y_n \le b\} &= \mathbb{P}\left\{\frac{n + a\sqrt{n}}{2} \le S_n \le \frac{n + b\sqrt{n}}{2}\right\} \\ &= \sum_{k=\frac{n+a\sqrt{n}}{2}}^{\frac{n+b\sqrt{n}}{2}} \binom{n}{k} \frac{1}{2^n}.\end{aligned}$$

Observe that if a, b are fixed, then for large n, both k and $n-k$ over which we are summing are also large. Hence we apply Stirlings' approximation to write

$$\begin{aligned}\binom{n}{k}\frac{1}{2^n} &\sim \frac{n^{n+\frac{1}{2}}e^{-n}\sqrt{2\pi}}{k^{k+\frac{1}{2}}e^{-k}\sqrt{2\pi} \times (n-k)^{n-k+\frac{1}{2}}e^{-(n-k)}\sqrt{2\pi} \times 2^n} \\ &= \frac{\sqrt{n}}{\sqrt{2\pi k(n-k)}} \times \left(\frac{2k}{n}\right)^{-k} \left(2(1 - \frac{k}{n}\right)^{-(n-k)} \\ &= \frac{\sqrt{n}}{\sqrt{2\pi k(n-k)}} \exp\left\{-k \log \frac{2k}{n} - (n-k)\log 2(1 - \frac{k}{n})\right\}\end{aligned}$$

In the range we want, $k = \frac{n + x\sqrt{n}}{2}$ for some $a \le x \le b$. Hence $\frac{2k}{n} = 1 + \frac{x}{\sqrt{n}}$ and $2(1 - \frac{k}{n}) = 1 - \frac{x}{\sqrt{n}}$. As x is fixed and n is large, we use the Taylor expansion of the logarithm to write

$$\begin{aligned}k \log \frac{2k}{n} &= \frac{n}{2}(1 + \frac{x}{\sqrt{n}})\left(\frac{x}{\sqrt{n}} - \frac{x^2}{2n} + \frac{x^3}{3n^{3/2}} - \dots\right), \\ (n-k)\log 2(1 - \frac{k}{n}) &= \frac{n}{2}(1 - \frac{x}{\sqrt{n}})\left(-\frac{x}{\sqrt{n}} - \frac{x^2}{2n} - \frac{x^3}{3n^{3/2}} - \dots\right).\end{aligned}$$

Adding these, we get $\frac{x^2}{2} + c\frac{x^4}{n} + \ldots$. Thus, ignoring higher order terms, we get

$$\mathbb{P}\{a \le Y_n \le b\} \approx \frac{1}{\sqrt{2\pi\frac{k}{n}(1-\frac{k}{n})}}\frac{1}{\sqrt{n}} \sum_{k=\frac{n+a\sqrt{n}}{2}}^{\frac{n+b\sqrt{n}}{2}} e^{-\frac{x^2}{2}}, \qquad x = \frac{2k-n}{\sqrt{n}},$$

$$\approx \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{x^2}{2}} dx$$

using the integral approximation to the sum. This is precisely what we wanted to prove.

7. A practical point in using central limit theorem

If X_k are i.i.d. Ber(1/2), then the central limit theorem says that

$$\mathbb{P}\{S_n \ge k_n\} = \mathbb{P}\left\{\frac{S_n - \frac{n}{2}}{\sqrt{n/4}} \ge \frac{k_n - \frac{n}{2}}{\sqrt{n/4}}\right\} \approx 1 - \Phi\left(\frac{k_n - \frac{n}{2}}{\sqrt{n/4}}\right)$$

provided $\frac{k_n - \frac{n}{2}}{\sqrt{n/4}}$ stays bounded. When n is large, instead of computing the Binomial coefficients and summing them, we may as well use the central limit approximation (the CDF Φ has to be evaluated on a computer, or looked up in a table).

As S_n is integer valued, the event $\{S_n \ge k_n\}$ is the same as $\{S_n > k_n - 1\}$ and $\{S_n \ge k_n - \frac{1}{2}\}$. But if we use these, the approximations for the probability we get as

$$1 - \Phi\left(\frac{k_n - 1 - \frac{n}{2}}{\sqrt{n/4}}\right) \qquad \text{and } 1 - \Phi\left(\frac{k_n - \frac{1}{2} - \frac{n}{2}}{\sqrt{n/4}}\right).$$

Which one should we use? The difference between the arguments of Φ in these cases are at most $\frac{1}{\sqrt{n/4}}$, and as Φ is continuous, the answers get close as $n \to \infty$. So it should not matter. However, in practical situations, in many statistical applications, we do apply CLT approximation with n as small as 50 or 100. For $n = 100$, note that $\frac{1}{\sqrt{n/4}} = 0.2$, so it can make a considerable difference in the probability. Hence the question has practical value.

Prescription: Whenever X_i are integer valued, always use the mid-point approximation. For example, write the event $\{35 \le S_{100} \le 42\}$ as $\{34.5 \le S_{100} \le 42.5\}$ and then apply CLT to get

$$\mathbb{P}\{35 \le S_{100} \le 42\} = \mathbb{P}\{34.5 \le S_{100} \le 42.5\} = \mathbb{P}\{-3.1 \le \frac{S_{100} - 50}{5} \le -1.5\}$$

$$\approx \Phi(-1.5) - \Phi(-3.1) \;=\; 0.933$$

Here is some numerical evidence that this gives better approximations for Binomial and Poisson distributions.

k	$\sum_{j=k}^{n} \binom{100}{j} 2^{-100}$	$1 - \Phi(\frac{k-51}{5})$	$1 - \Phi(\frac{k-50.5}{5})$	$1 - \Phi(\frac{k-50}{5})$
60	0.028444	0.0359303	0.0287166	0.0227501
45	0.864373	0.88493	0.864334	0.841345
68	0.000204389	0.000336929	0.000232629	0.000159109

k	$\sum_{j=0}^{k} e^{-49}\frac{49^j}{j!}$	$\Phi(\frac{k-49}{7})$	$1-\Phi(\frac{k-48.5}{7})$	$1-\Phi(\frac{k-48}{7})$
40	0.109803	0.0992714	0.112319	0.126549
57	0.885843,	0.873451	0.887681	0.900729
44	0.264727	0.237525	0.260158	0.283855

8. Poisson limit for rare events

Let $X_k \sim \text{Ber}(p)$ be independent random variables. Central limit theorem says that if p is fixed and n is large, the distribution of $(X_n - np)/\sqrt{np(1-p)}$ is close to the $N(0,1)$ distribution.

Now we consider a slightly different situation. Let $X_1, \ldots, X_n$ have $\text{Ber}(n, p_n)$ distribution where $p_n = \frac{\lambda}{n}$, where $\lambda > 0$ is fixed. Then, we shall show that the distribution of $X_1 + \ldots + X_n$ is close to that of $\text{Pois}(\lambda)$. Note that the distribution of X_1 changes with n and hence it would be more correct to write $X_{n,1}, \ldots, X_{n,n}$.

THEOREM 45. *Let $\lambda > 0$ be fixed and let $X_{n,1}, \ldots, X_{n,n}$ be i.i.d. Ber(λ/n). Let $S_n = X_{n,1} + \ldots + X_{n,n}$. Then, for every $k \geq 0$*

$$\mathbb{P}\{S_n = k\} \to e^{-\lambda}\frac{\lambda^k}{k!}.$$

PROOF. Fix k and observe that

$$\begin{aligned}\mathbb{P}\{S_n = k\} &= \binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k}\\ &= \frac{n(n-1)\ldots(n-k+1)}{k!}\frac{\lambda^k}{n^k}\left(1-\frac{\lambda}{n}\right)^{n-k}.\end{aligned}$$

Note that $\frac{n(n-1)\ldots(n-k+1)}{n^k} \to 1$ as $n \to \infty$ (since k is fixed). Also, $(1-\frac{\lambda}{n})^{n-k} \to e^{-\lambda}$ (if not clear, note that $(1-\frac{\lambda}{n})^n \to e^{-\lambda}$ and $(1-\frac{\lambda}{n})^{-k} \to 1$). Hence, the right hand side above converges to $e^{-\lambda}\frac{\lambda^k}{k!}$ which is what we wanted to show. ■

What is the meaning of this? Bernoulli random variables may be thought of as indicators of events, i.e., think of $X_{n,1}$ as $\mathbf{1}_{A_1}$ etc. The theorem considers n events which are independent and each of them is "rare" (since the probability of it occurring is λ/n which becomes small as n increases). The number of events increases but the chance of each events decreases in such a way that the expected number of events that occur stays constant. Then, the total number of events that actually occur has an approximately Poisson distribution.

EXAMPLE 46. (A physical example). A large amount of custard is made in the hostel mess to serve 100 students. The cook adds 300 raisins and mixes the custard so that on an average they get 3 raisins per student. But the number of raisins that a given student gets is random and the above theorem says that it has approximately Pois(3) distribution. How so? Let X_k be the indicator of the event that the kth raisin ends up in your cup. Since there are 100 cups, the chance of this happening is 1/100. The number of raisins in your cup is precisely $X_1 + X_2 + \ldots + X_{300}$. Appy the theorem (take $n = 100$ and $\lambda = 3$).

EXAMPLE 47. Place r balls in m bins at random. If $m = 1000$ and $r = 500$, then the number of balls in the first bin has approximately Pois(1/2) distribution. Work out how this comes from the theorem.

9. Some extensions of the limit theorems

Both the Gaussian and Poisson distributions arise as limits in numerous situations, even where the hypotheses made in the central limit theorem and the Poisson limit theorem do not exactly hold. There is no single theorem that covers all such occurrences, and such limit theorems are being proved to this day for various applications. Here we just mention one of each kind.

9.1. A Poisson limit without independence. Recall the problem of a psychic guessing a deck of n cards. If S_n is the number of correct guesses, we have seen by direct calculation and approximation that $\mathbb{P}\{S_n = k\}$ is close to $e^{-1}/k!$. More precisely, if k is fixed and $n \to \infty$, then $\mathbb{P}\{S_n = k\} \to \frac{1}{e\ k!}$. That is S_n has approximately Pois(1) distribution.

Does it follow from Theorem 45? We can write $S_n = X_{n,1} + \ldots + X_{n,n}$, where $X_{n,k}$ is the indicator of the event that the kth guess is correct. Then $X_{n,k} \sim \text{Ber}(1/n)$ for each k. It looks like the theorem tells us that S_n should have Pois(1) distribution approximately. But note that $X_{n,i}$ are not independent random variables and hence the theorem does not strictly apply. But for any $j < k$, it is easy to see that $\mathbb{P}\{X_{n,j} = 1, X_{n,k} = 1\} = \frac{1}{n(n-1)}$ which is very close (when n is large) to $\mathbb{P}\{X_{n,j} = 1\}\mathbb{P}\{X_{n,k} = 1\} = \frac{1}{n^2}$, so any pair of $X_{n,i}$ are nearly independent. Similarly, one can check that any fixed finite number of $X_{n,i}$s become nearly independent as $n \to \infty$.

The theorem should be thought of as one of many theorems that capture the theme *"in a large collection of rare events that are nearly independent, the actual number of events that occur is approximately Poisson"*.

9.2. A Central limit theorem without independence. Let $\mathbb{S}^{n-1}$ denote the sphere in n dimensions, that is

$$\mathbb{S}^{n-1} = \{x \in \mathbb{R}^n : x_1^+ \ldots + x_n^2\} = 1.$$

It is possible to define precisely what it means to pick a point at random from $\mathbb{S}^{n-1}$. Let $X^{(n)} = (X_1^{(n)}, \ldots, X_n^{(n)})$ be such a random vector. Then the claim is that any co-ordinate of $X^{(n)}$, say the first one, has approximately Gaussian distribution with mean 0 and variance $1/n$. More precisely, $\sqrt{n}X_1^{(n)}$ has approximately $N(0,1)$ distribution. Here is how the calculation goes. The significance of this theorem is explained at the end.

Let us write σ_{n-1} for the surface area of $\mathbb{S}^{n-1}$. One way to calculate it is inductively. If we fix the first co-ordinate as $x_1 = t$, then the remaining co-ordinates $(x_2, \ldots, x_{n-1}) \in \mathbb{R}^{n-1}$ and $x_2^+ \ldots + x_n^2 = 1 - t^2$. Thus, they form the sphere of radius $\sqrt{1-t^2}$ in $\mathbb{R}^{n-1}$, or in obvious notation it is basically $\sqrt{1-t^2}\mathbb{S}^{n-1}$. By the usual scaling behaviour, its surface area

is $(1-t^2)^{(n-2)/2}\sigma_{n-2}$. In other words,

$$\sigma_{n-1} = \sigma_{n-2}\int_{-1}^{1}(1-t^2)^{\frac{n}{2}-1}dt.$$

On the other hand, by the same logic, the surface area of the part of $\mathbb{S}^{n-1}$ with $a \le x_1 \le b$ (where $-1 \le a < b \le 1$) is

$$\sigma_{n-2}\int_{a}^{b}(1-t^2)^{\frac{n}{2}-1}dt$$

Taking ratios, we see that

$$\mathbb{P}\{a < X^{(n)} < b\} = \frac{\int_a^b (1-t^2)^{\frac{n}{2}-1}dt}{\int_{-1}^{1}(1-t^2)^{\frac{n}{2}-1}dt}.$$

This is just another way of saying that $X_1^{(n)}$ has density $C_n^{-1}(1-t^2)^{\frac{n}{2}-1}$ on $[-1,1]$ where $C_n = \int_{-1}^{1}(1-t^2)^{\frac{n}{2}-1}dt$. To make a link with familiar distributions, convince yourself that the density of $(X^{(n)}+1)/2$ is Beta$(\frac{n}{2},\frac{n}{2})$. In particular, check that $C_n = 2\text{Beta}(n/2, n/2)$.

The density of $\sqrt{n}X_1^{(n)}$ is then $\frac{1}{C_n\sqrt{n}}\left(1-\frac{u^2}{n}\right)^{\frac{n}{2}}-1$ for $u \in [-\sqrt{n},\sqrt{n}]$. As $n\to\infty$, this converges to $\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}$, which is the density of the $N(0,1)$ distribution.

Exercise 48. Show that if k is fixed, then as $n\to\infty$, the components $\sqrt{n}X_1^{(n)},\ldots,\sqrt{n}X_k^{(n)}$ are approximately *independent* standard Gaussian random variables.

Why did we do this calculation? It is essentially a calculation going back to Maxwell in his study of the velocity distribution of gases. Imagine a box containing a mono-atomic gas that is maintained at a fixed temperature. The temperature is nothing but the average kinetic energy of the molecules. If there are $N = 10^{23}$ molecules, then the velocity co-ordinates of all of them can be strung together as a vector $(X_1,\ldots,X_{3N})$. The constant temperature condition says that this vector lies on a sphere $R\mathbb{S}^{3N-1}$ for some $R>0$. What we proved above shows that the individual velocity components have approximately Gaussian distribution (mean 0, variance $R^2/3N$). This is Maxwell's velocity distribution of gases. One may be more interested in the kinetic energy of a single molecule. Using Exercise 48, it is seen to be the approximately the same as the distribution of the sum of squares of three independent Gaussians, which is Gamma$(3/2, R^2/3N)$.

10. Exercises

Problem 1. Consider the following integrals

$$\int_0^1 \frac{4}{1+x^2}dx = \pi, \qquad \int_0^1 \frac{1}{\sqrt{x(1-x)}}dx = \pi.$$

In either case, use Monte-Carlo integration with 100, 1000 and 10000 samples from uniform distribution to find approximations of π. Compare the approximations to the true value 3.1416....

PROBLEM 2. Recall the *coupon collector problem*. A box contains n coupons labelled $1, 2, \ldots, n$. Coupons are drawn at random from the box, repeatedly and with replacement. Let T_n be the number of draws needed till each of the coupons has appeared at least once.

(1) Show that $\mathbb{E}[T_n] \sim n \log n$ (this just means $\frac{1}{n \log n}\mathbb{E}[T_n] \to 1$).

(2) Show that $\mathrm{Var}(T_n) \leq 2n^2$.

(3) Show that $\mathbb{P}\left(\left|\frac{T_n}{n \log n} - 1\right| > \delta\right) \to 0$ for any $\delta > 0$.

[*Hint:* Consider the number of draws needed to get the first new coupon, the further number of draws needed to get the next coupon and so on].

PROBLEM 3. (**) Recall the coupon collector problem where coupons are drawn repeatedly (with replacement) from a box containing coupons labelled $1, 2, \ldots, N$. Let T_N be the number of draws made till all the coupons are seen.

(1) Find $\mathbb{E}[T_N]$ and $\mathrm{Var}(T_N)$.

(2) Use Chebyshev's inequality to show that for any $\delta > 0$, as $N \to \infty$ we have

$$\mathbb{P}\{(1-\delta)N \log N \leq T_N \leq (1+\delta)N \log N\} \to 1.$$

PROBLEM 4. (**) Let X be a non-negative random variable. Read the discussion following the problem to understand the significance of this problem.

(1) Suppose X_n takes the values n^2 and 0 with probabilities $1/n$ and $1 - (1/n)$, respectively. Compare $\mathbb{P}\{X_n > 0\}$ and $\mathbb{E}[X_n]$ for large n.

(2) Show the *second moment inequality* (aka *Paley-Zygmund inequality*): $\mathbb{P}\{X > 0\} \geq (\mathbb{E}[X])^2/\mathbb{E}[X^2]$.

[**Discussion:** Markov's inequality tells us that that the tail probability $\mathbb{P}\{X \geq t\}$ can be bounded from above using $\mathbb{E}[X]$. In particular, $\mathbb{P}\{X \geq r\mathbb{E}[X]\} \leq \frac{1}{r}$. A natural question is whether there is a lower bound for the tail probability in terms of the expected value. In other words, if the mean is large, must the random variable be large with significant probability?The first part shows that the answer is 'No' in general. The second part shows that the answer is 'Yes', provided we have control on the second moment $\mathbb{E}[X^2]$ from above. Notice why the inequality does not give any useful bound in the first part of the problem (what happens to the second moment of X_n?)]

Hoeffding's inequality: Using Chebyshev's inequality we got a bound of σ^2/nt^2 for the probability that the sample mean deviates from the population mean by more than t. This is very general. If we make more assumptions about our random variable, we can give better bounds. The following exercise is to illustrate this.

PROBLEM 5. (Optional!) Let $X_1, \ldots, X_n$ be i.i.d. $\mathrm{Ber}_{\pm}(1/2)$. That is, $\mathbb{P}\{X_k = +1\} = \mathbb{P}\{X_k = -1\} = \frac{1}{2}$. Let $\bar{X}_n = \frac{1}{n}(X_1 + \ldots + X_n)$. Show that

$$\mathbb{P}\{|\bar{X}_n| > t\} \leq 2e^{-nt^2/2}$$

by following these steps.

(1) Show that $\mathbb{P}\{\bar{X}_n > t\} \leq e^{-\theta t}\left(\frac{e^{\theta/n}+e^{-\theta/n}}{2}\right)^n$ for any $\theta > 0$.

(2) Prove the inequality $e^x + e^{-x} \leq 2e^{x^2/2}$ for any $x > 0$.

(3) Use the first two parts to show that $\mathbb{P}\{\bar{X}_n > t\} \leq e^{-nt^2/2}$ (you must make an appropriate choice of θ depending on t).

(4) Now consider $|\bar{X}_n|$ and break $\mathbb{P}\{|\bar{X}_n| > t\}$ into two summands to get the desired inequality.

[**Note:** Here $\mu = 0$ and $\sigma^2 = 1$, and hence Chebyshev's inequality only gives the bound $\mathbb{P}\{|\bar{X}_n| > t\} \leq \frac{1}{nt^2}$. Do you see that Hoeffding's inequality is better?]

PROBLEM 6. Let $X_1, X_2, \ldots$ be i.i.d. Uniform$[1,2]$ distribution. Let $S = X_1 + \ldots + X_{100}$. Give approximate quantiles at levels 0.01, 0.25, 0.5, 0.75, 0.99 for S. Use CLT and normal distribution tables.

PROBLEM 7. Let $X_1, \ldots, X_n$ be i.i.d. samples from a parametric family of discrete distributions. In each of the following cases, find the MLE for the unknown parameter(s) and find the bias.

(1) X_i are i.i.d. Ber(p) where p is unknown.

(2) X_i are i.i.d. $N(\mu, \sigma^2)$ where μ, σ^2 are unknown.

PROBLEM 8. Let $X_1, \ldots, X_n$ be i.i.d. samples from a parametric family of discrete distributions. In each of the following cases, find the MLE for the unknown parameter(s) and calculate the bias.

(1) X_i are i.i.d. Geo(p) where p is unknown.

(2) X_i are i.i.d. Unif$[a,b]$ where a, b are unknown.

PROBLEM 9. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. samples from a bivariate distribution. Let $\tau = \mathrm{Cov}(X_1, Y_1)$. Let $r_n = \frac{1}{n}\sum_{k=1}^n (X_k - \bar{X}_n)(Y_k - \bar{X}_n)$ be the sample covariance.

(1) Show that r_n is a biased estimate for τ and find the bias.

(2) Modify the estimate r_n to get an unbiased estimate of τ.

[**Remark:** It is often convenient, here and elsewhere, to realise that $\tau = \mathbb{E}[X_1 Y_1] - \mathbb{E}[X_1]\mathbb{E}[Y_1]$ and $r_n = (\frac{1}{n}\sum_{k=1}^n X_k Y_k) - \bar{X}_n \bar{Y}_n$.]

PROBLEM 10. Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables from a distribution F. Let M_n be a median of $X_1, \ldots, X_n$. Assume that the distribution F has a unique median, that is there is a unique number m such that $F(m) = \frac{1}{2}$. For any $\delta > 0$ show that $\mathbb{P}\{|M_n - m| \geq \delta\} \to 0$ as $n \to \infty$.[**Remark:** The above statement justifies using the sample median to estimate the population median, in the sense that at least for large sample sizes, the two are close. Similar justification for using sample mean to estimate expected value came from the law of large numbers]

The following problem is only for those mathematically minded.

PROBLEM 11. Let $X_1, X_2, \ldots$ be i.i.d. Pois(λ) random variables. Work out the exact distribution of $X_1 + \ldots + X_n$ and use it show the central limit theorem in this case. That is, show that for any $a < b$,

$$\mathbb{P}\left\{a \leq \frac{\sqrt{n}(\bar{X}_n - \lambda)}{\sqrt{\lambda}} \leq b\right\} \longrightarrow \mathbb{P}\{a \leq Z \leq b\}$$

where $Z \sim N(0,1)$.[**Remark:** This is analogous to the two cases of CLT that we showed in class, for exponential and for Bernoulli random variables].

The following problem shows that in certain situations, sums of random variables are approximately Poisson distributed. This gives a hint as to why Poisson distribution arises in many contexts. The question may be ignored safely from the exam point of view.

PROBLEM 12. Let $X_{n,1}, X_{n,2}, \ldots X_{n,n}$ be i.i.d. Ber(p_n) random variables. Let $S_n = X_{n,1} + \ldots + X_{n,n}$.If $np_n \to \lambda$ (a finite positive number), show that S_n has approximately Pois(λ) distribution in the sense that for any $k \in \mathbb{N}$,

$$\mathbb{P}\{S_n = k\} \to e^{-\lambda}\frac{\lambda^k}{k!}.$$

[**Remark:** In contrast, if $np_n \to \infty$, deduce from CLT that S_n has approximately a normal distribution, i.e.,

$$\mathbb{P}\left\{a \leq \frac{S_n - \mathbb{E}[S_n]}{\sqrt{\mathrm{Var}(S_n)}} \leq b\right\} \to \mathbb{P}\{a \leq Z \leq b\}$$

for any $a < b$.]

CHAPTER 9

Some interesting problems of probability

The problems we discuss here are important and used in answering many larger questions. Here we just take them as some interesting problems that can actually be solved.

1. The coupon collector problem

A box has N coupons labelled $1, 2, \ldots, N$. Coupons are drawn repeatedly and uniformly at random with replacement, till all the coupons have appeared at least once. Let T be the number of draws required. What can we say about T, particularly for large N, beyond the obvious statement that $N \leq T$?

It makes sense to consider the times $T_1 < T_2 < \ldots < T_N$ at which a new unseen coupon appears. So $T_1 = 1$ and $T_N = T$. The key to solving the problem is to consider the differences $\tau_1 = T_2 - T_1, \ldots, \tau_{N-1} = T_N - T_{N-1}$.

CLAIM 13. *$\tau_1, \ldots, \tau_{N-1}$ are independent and $\tau_k \sim Geo((N-k)/k)$.*

We shall leave an honest proof of the claim as an exercise, and just give the intuitive explanation that when k coupons have already turned up, we are waiting for one of the other $N - k$ coupons to show up. This is like waiting for a head in a sequence of tosses of a coin with $p = \frac{N-k}{N}$. Hence $\tau_k \sim \text{Geo}((N-k)/N)$. Since this also does not depend on what the values of $\tau_1, \ldots, \tau_{k-1}$ were, we should have the independence of $\tau_1, \ldots, \tau_{N-1}$.

Accepting the claim, we can easily solve the problem. First observe that $\mathbb{E}[\tau_k] = \frac{N}{N-k}$ and $\text{Var}(\tau_k) = \frac{kN}{(N-k)^2}$ by properties of Geometric distributions. Hence writing $T_N = 1 + \tau_1 + \ldots + \tau_{N-1}$ we see that

$$\mathbb{E}[T_N] = 1 + \sum_{k=1}^{N-1} \frac{N}{N-k} = N \sum_{k=1}^{N} \frac{1}{k} \sim N \log N$$

where $\sim$ means that the ratio of the two sides approaches 1 as $N \to \infty$. This only required linearity of Expectation. But using the independence of τ_ks and the fact that $\text{Var}(\tau_k) \leq N^2/(N-k)^2$, we see that

$$\text{Var}(T_N) = \sum_{k=1}^{N-1} \text{Var}(\tau_k) \;\leq\; N^2 \sum_{k=1}^{N-1} \frac{1}{(N-k)^2} \;\leq\; 2N^2$$

where we used the fact that $\sum_{j=1}^{\infty} \frac{1}{j^2} \leq 2$ (in fact the sum is $\pi^2/6$).

In conclusion, $\mathbb{E}[T_N] \sim N \log N$ and $\text{s.d.}(T_N) \leq \sqrt{2}N$. Therefore, by Chebyshev's inequality, T_N is in a window centered at $N \log N$ and of length a few multiples of N. The precise

statement is that

$$\mathbb{P}\{T_N \in [N\log N - Nh_N, N\log N + Nh_N]\} \to 1$$

for any h_N that goes to infinity. For example, we can take $h_N = \log\log N$ to see that $T_N/N\log N$ is very close to 1, with high probability.

2. The secretary problem

To fill the post of a secretary, N candidates have been called. Each of them has a true competence which will become known when you interview the candidate. Assume that no two are equally competent (so there are no ties and they can be ranked). You get to interview the candidates one by one, and immediately after interviewing a candidate, you must accept or reject the candidate. Once you accept a candidate, all the rest are sent home. Once rejected, a candidate cannot be called back.

Your goal is to get the best candidate (even getting the second best is considered a failure). With what chance can you achieve this?

It may seem that if N is large, the chance of stopping exactly at the best candidate must be small. Not so! Here is a surprising strategy: Interview $\lceil N/2\rceil$ of the candidates, and reject all of them (you will remember/note down their competencies, of course). Then continue interviewing the candidates and select the first candidate who is better than all the ones who came before. If no such candidate arrives, you select the last candidate.

Does that get the best candidate for you? Not necessarily, for example if the best candidate was in the first half, then you have already rejected her. However, *if* it happens that the best candidate is in the second half, and the second best candidate is in the first half, then this strategy does catch the best candidate! The chance of that happening (for large N) is approximately $\frac{1}{4}$. Thus you have a probability of $\frac{1}{4}$ of succeeding in the mission, irrespective of how large N is!

In fact, this strategy has a better than $\frac{1}{4}$ chance of succeeding. Suppose the top two candidates are in the second half and the third best candidate is in the first half (this event is disjoint from the one already considered above). Then the strategy stops at either the first or the second candidate, whichever comes first. Hence the chance of catching the best candidate has a chance of $\frac{1}{8}$. Going further, if the top k candidates are in the second half, and the $(k+1)$st best candidate is in the first half, then the chance of catching the top candidate is $\frac{1}{k}$, and the event we are asking has chance $\frac{1}{2^{k+1}}$. In conclusion, for any fixed ℓ, as $N \to \infty$, we have

$$\mathbb{P}\{\text{success}\} \geq (1 - o(1)) \sum_{k=1}^{\ell} \frac{1}{k2^{k+1}}$$

The reason for the $1 - o(1)$ is that the events of different candidates being in first half or second half are not independent, but only approximately so, as $N \to \infty$. Letting $\ell \to \infty$, we see that the success probability is arbitrarily close to $\frac{1}{2}\log 2 \approx 0.35$.

Generalize the above strategy by just fixing $\rho \in (0,1)$ and rejecting the first ρN candidates, and then continue the interviews and pick the first candidate who is better than all the ones before.

Exercise 14. Show that (as $N \to \infty$) the strategy has a chance of success at least $-\rho \log \rho$. What is the optimal choice of ρ?

3. A randomized algorithm

A box has a large number N of coupons with a number written on each (e.g., think of heights of people in a population of N people). Suppose it is desired to know the range of numbers in the box. How many coupons do we need to check?

The answer is obviously N, as even leaving out one coupon can give the wrong answer. But in many problems of importance, we may not be interested in the absolute minimum or maximum, but only make sure of finding an interval that has at least 80% of the numbers within (e.g., may be you are designing chairs or trousers; you may want to mainly stock up on sizes catering to the middle and leave out the extremes). How many coupons do we need to check? If you want to be absolutely sure, you must check more than $0.8N$ coupons, but this can be prohibitively expensive.

Although the problem has no randomness, it helps immensely to bring it in! Suppose you pick k coupons at random and take the minimum and maximum of those numbers to be the desired interval. Will you succeed? Not necessarily, you may be unlucky and end up with all numbers in the higher end. But what is the chance?

Let $a < b$ be the 10% and 90% quantiles (so there are 10% of the coupons below a and 10% above b). If one of our k numbers falls below a and one other falls above b, then our interval does contain 80% of the coupons. What is the chance of failing? Now we need to be more precise in the mode of sampling. Sampling without replacement makes more sense and in fact performs better, but for ease of calculation, let us do sampling with replacement. Then to fail, all k coupons must be above a or all k coupons must be below b. Thus

$$\mathbb{P}\{\text{success}\} \geq 1 - 2(0.9)^k.$$

For $k = 50$, this is $0.9896\ldots$ In other words, irrespective of how large the population is, by sampling just 50 of them, we can get the desired range with only 0.01 failure probability!

Remark 15. Observe that the choice of k has nothing to do with the population size N which is surprising! This is the same surprise as in the voting

4. Gambler's ruin problem

Consider two gamblers with capital amounts A and B. They play a sequence of games, in each of which the loser pays 1 rupee to the winner and each has a chance $\frac{1}{2}$ of winning. They play till one of them becomes bankrupt. What is the chance that the first gambler becomes bankrupt?

We can formulate this using fair coin tosses $X_1, X_2, \ldots$ (i.i.d $\text{Ber}_{\pm}(1/2)$), where $X_k = +1$ if the first gambler wins and -1 if the first gambler loses the kth game. Let $S_n = A + X_1 + \ldots + X_n$ for $n = 0, 1, 2 \ldots$. Then S_n is the money with the first gambler after n games. The play is stopped at the smallest n for which $S_n = 0$ (the first gambler has become bankrupt) or when $S_n = A + B$ (the second gambler has become bankrupt). Let $\tau = \min\{n : S_n = 0 \text{ or } S_n = A + B\}$.

CLAIM 16. *$\tau < \infty$ w.p.1. and $\mathbb{P}\{S_\tau = 0\} = \frac{B}{A+B}$ and $\mathbb{P}\{S_\tau = A + B\} = \frac{A}{A+B}$.*

In other words, the probability of winning is proportional to the initial capital.

PROOF THAT $\tau < \infty$. We know that if a p-coin with $p > 0$ is tossed repeatedly, eventually it will show a Head (the probability to not get a Head for k tosses is q^k which goes to 0 as $k \to \infty$). To utilize this, divide the steps into blocks of length $C := A + B$ and let E_k be the event that $X_j = +1$ for $kC < j \leq (k + 1)C$. As E_k are independent and $\mathbb{P}(E_k) = 1/2^C$ for each k, we know that one of them will occur, w.p.1. But if E_k occurs and $\tau > kC$, then $\tau < (k+1)C$ (because if $\tau > kC$, then $S_{kC} > 0$ and hence $S_{(k+1)C} = S_{kC} + C > C$). This shows that $\tau < \infty$ w.p.1. ■

PROOF THAT $S_\tau = 0$ W.P. $B/(A + B)$. As we have proved that $\tau < \infty$ w.p.1., we can talk of the random variable S_τ and it is equal to 0 or C. The trick is to fix C and consider $\mathbb{P}\{S_\tau = C\} =: f(A)$ as a function of A (and $B = C - A$). Obviously $f(0) = 0$ (first player is bankrupt to begin with!) and $f(C) = 1$(second player is bankrupt from the start). We need to find $f(A)$ for $1 \leq A \leq C - 1$. By the law of total probability

$$\begin{aligned} f(A) &= \mathbb{P}\{X_1 = +1\}\mathbb{P}\{S_\tau = C \;\big|\; X_1 = +1\} + \mathbb{P}\{X_1 = -1\}\mathbb{P}\{S_\tau = C \;\big|\; X_1 = -1\} \\ &= \frac{1}{2}f(A + 1) + \frac{1}{2}f(A - 1). \end{aligned}$$

Here we used the fact that if $X_1 = 1$, then the first player has $A + 1$ rupees, so it is like starting with a capital of $A + 1$, hence the probability if $f(A + 1)$. Similarly when $X_1 = -1$, the chance becomes $f(A - 1)$. The above recursions say that $f(A + 1) - f(A) = f(A) - f(A - 1)$, i.e., the differences are constant. As $f(C) - f(0) = 1$, we see that $f(A) = \frac{A}{C}$ for all A. ■

5. Recurrence of random walk on $\mathbb{Z}$

Let $X_1, X_2, \ldots$ be i.i.d. $\text{Ber}_{\pm}(1/2)$ and let $S_n = X_1 + \ldots + X_n$ with $S_0 = 0$. Then $\{S_n\}$ is called simple symmetric random walk (simple refers to steps being ± 1). Let $\tau_b = \min\{k \geq 1 : S_k = b\}$ for $b \in \mathbb{Z}$.

CLAIM 17. *$\mathbb{P}\{\tau_b < \infty\} = 1$ for all $b \in \mathbb{Z}$.*

The event $\{\tau_b < \infty\}$ is the same as $\bigcup_{k \geq 1}\{S_k = b\}$, the event that the random walk visits the location b (or if $b = 0$, the event that it returns to 0). According to the claim, this happens w.p.1. By the countable adidtivity of probability, it follows that $\bigcap_{b \in \mathbb{Z}}\{\tau_b < \infty\}$ also has probability 1. In other words, w.p.1., the random walk visits every location in $\mathbb{Z}$.

PROOF. Fix $b < 0$. We showed in the gambler's ruin problem that $\tau := \tau_b \wedge \tau_N < \infty$ w.p.1 (i.e., the probability to exit the interval $[b, N]$ is 1). But $\mathbb{P}\{\tau_b \neq \tau\} = \mathbb{P}\{\tau_N < \tau_b\} = \frac{b}{N}$, again by the gambler's ruin problem. Therefore, $\mathbb{P}\{\tau_b = \infty\} \leq \mathbb{P}\{\tau_b \neq \tau\} = \frac{b}{N}$ which goes to zero as $N \to \infty$. Hence $\mathbb{P}\{\tau_b < \infty\} = 1$.

An almost verbatim proof shows that $\tau_b < \infty$ w.p.1. for $b > 0$. But between τ_1 and τ_{-1}, the random walk must pass through the origin. Therefore, $\tau_0 < \infty$ w.p.1 too. ■

6. The ballot problem

There are k candidates $C_1, \ldots, C_k$ standing for an election, and $N_1 > N_2 > \ldots > N_k$ voters have cast votes in their favour respectively. All the votes are in a box from which they are drawn one after another uniformly at random and counted. What is the chance that throughout the counting process, the candidate C_1 is leading over C_2 who is leading over C_3 and so on (i.e., at all times the ordering is the same as at the end)?

This is called the *generalized ballot problem*, as the term *ballot problem* is usually applied to the case of two candidates and that is what we solve now.

CLAIM 18. *Let $k = 2$ with $N_1 > N_2$. Then the probability that the first candidate is leading over the second throughout the counting process is $\frac{N_1 - N_2}{N_1 + N_2}$.*

PROOF. Encode the counting process as a path in $\mathbb{Z}^2$ that starts at $(0, 0)$, and takes a vertical step for each vote to the first candidate and a horizontal step for each vote to the second candidate. The path ends at (N_2, N_1). Conversely, any path from $(0, 0)$ to (N_2, N_1) in the lattice $\mathbb{Z}^2$ that goes up or right in each step, corresponds to an order in which the votes are counted. In either side, it is easy to see that the cardinality is $\binom{N}{N_1}$ where $N = N_1 + N_2$.

Our problem is equivalent to asking for the number of paths of this kind that stay strictly above the diagonal $\{(j, j) : j \in \mathbb{Z}\}$ (except for the starting point which lies on the diagonal).

Obviously the first step has to be vertical, so the number is the same as the number of paths from $(0, 1)$ to (N_2, N_1) that do not hit the diagonal. Without that restriction, the number of paths is $\binom{N-1}{N_1-1}$. We need to subtract those paths from $(0, 1)$ to (N_2, N_1) that do hit the diagonal. We claim that such paths are in bijection with (unrestricted paths) from $(0, 1)$ to (N_2, N_1). If we accept this claim, then the number of paths we are looking for becomes

$$\binom{N-1}{N_1-1} - \binom{N-1}{N_1} = \frac{N_1 - N_2}{N} \binom{N}{N_1}.$$

Divide by the total number of paths (which is $\binom{N}{N_1}$) to see that the probability is $\frac{N_1 - N_2}{N}$. ■

Suppose we modify the question to demand that throughout the counting process, the first candidate should have *at least as many* votes as the second. The probability must increase, but to what?

EXERCISE 19. Use the result or proof of the ballot problem to prove that for the modified ballot problem, the probability is $\frac{N_1 - N_2 + 1}{N_1 + 1}$.

REMARK 20. For the k candidate ballot problem, if we modify "leading" to "weakly leading" as above (at all times, the first candidate has at least as many votes as the second, and so on), the answer is

$$\det\left(\frac{N_j!}{(N_j+i-j)!}\right)_{i,j\leq k}.$$

7. Group testing

A blood test for a contagious disease needs to be administered to a large group of people, to spot the infected cases and isolate them. Statistician and Economist Robert Dorfman came up with a clever idea in 1943 to reduce the costs of such large-scale testing.

The idea is to combine the blood samples of several people at a time and test the combined sample together. If the combined sample tests negative, we can dismiss all the people at once. If the combined sample tests positive, we test each of the individuals separately. It is assumed here that the sensitivity of the test does not go down by the grouping, which in practise can limit the number of people whose samples can be combined. But even if we combine two at a time, it gives a saving, as we shall see now!

Let p (assumed small) be the proportion of people in the population who have the disease. Let g be the number of people whose blood sample is combined. If we had to test all of them separately, we would need g tests. In the group-testing procedure, we may need $g+1$ tests if at least one of the g people is infected, which has probability at most pg (we assume $pg \ll 1$, so the union bound is good enough for us). In all other cases, we need only one test. Hence the expected number of tests is at most $1-pg+(g+1)pg = 1+pg^2$. This is only an expectation calculation, but if N is much larger than g, this is applied to a large number N/g of groups, and by the law of large numbers, the actual proportions will match the expectation. In conclusion, we only need $\frac{1}{g}+pg$ tests *per person*.

As an example, suppose $p = 0.001$ and $g = 10$. Then the expected number of tests per person is 0.11, which means that we save 89% of the cost of testing a large group of people! Even if grouping more than 3 is not desirable (say due to loss of sensitivity of the test), we get $\frac{1+0.001\times 9}{3} = 0.336$, which gives a cost-saving of 66%.

One can even try to optimize the group size. Since $\frac{1}{g}+pg$ is minimized when $g = \frac{1}{\sqrt{p}}$, that is the right group size, and plugging it in, we see that the number of tests per person is $2\sqrt{p}$. For $p = 0.001$, this becomes 0.063 with optimal group size of 31.6, so one ends up saving 93% of the cost! Of course, the caveat about decreasing sensitivity of the test due to dilution must be kept in mind before applying such calculations to a real situation.

Part 2

Statistics

CHAPTER 10

A bird's eye overview of Statistics

In statistics we are faced with data, which could be measurements in an experiment, responses in a survey etc. There will be some randomness, which may be inherent in the problem or due to errors in measurement etc. The problem in statistics is to make various kinds of inferences about the underlying distribution, from realizations of the random variables. We shall consider a few basic types of problems encountered in statistics. We shall mostly deal with examples, but sufficiently many that the general ideas should become clear too. It may be remarked that we stay with the simplest "textbook type problems" but we shall also see some real data. Unfortunately we shall not touch upon the problems of current interest, which typically involve very huge data sets etc. Here are the kinds of problems we study.

General setting: We shall have data (measurements perhaps), usually of the form $X_1, \ldots, X_n$ which are realizations of independent random variables from a common distribution. The underlying distribution is not known. In the problems we consider, typically the distribution is known, except for the values of a few parameters. Thus, we may write the data as $X_1, \ldots, X_n$ i.i.d. $f_\theta(x)$ where $f_\theta(x)$ is a pdf or pmf for each value of the parameter(s) θ. For example, the density could be of $N(\mu, \sigma^2)$ (two unknown parameters μ and σ^2) or of Pois(λ) (one unknown parameter λ).

(1) Estimation: Here, the question is to guess the value of the unknown θ from the sample $X_1, \ldots, X_n$. For example, if X_i are i.i.d. from Ber(p) distribution (p is unknown), then a reasonable guess for θ would be the sample mean $\overline{X}_n$ (an *estimator*). Is this the only one? Is it the "best" one? Such questions are addressed in estimation.

(2) Confidence intervals: Here again the problem is of estimating the value of a parameter, but instead of giving one value as a guess, we instead give an interval and quantify how sure we are that the interval will contain the unknown parameter. For example, a coin with unknown probability p of turning up head, is tossed n times. Then, a confidence interval for p could be of the form

$$\left[\overline{X}_n - \frac{3}{\sqrt{n}}\sqrt{\overline{X}_n(1-\overline{X}_n)}, \overline{X}_n + \frac{3}{\sqrt{n}}\sqrt{\overline{X}_n(1-\overline{X}_n)}\right]$$

where $\overline{X}_n$ is the proportion of heads in n tosses. The reason for such an interval will come later. It turns out that if n is large, one can say that with probability 0.99 ("confidence level"), this interval will contain the true value of the parameter.

(3) Hypothesis testing: In this type of problem we are required to decide between two competing choices ("hypotheses"). For example, it is claimed that one batch of students is better than a second batch of students in mathematics. One way to check this is to give the same exam to students in both exams and record the scores. Based on the scores, we have to decide whether the first batch is better than the second (one hypothesis) or whether there is not much difference between the two (the other hypothesis). One can imagine that this can be done by comparing the sample means etc., but that will come later.

A good analogy for testing problems is from law, where the judge has to decide whether an accused is guilty or not guilty. Evidence presented by lawyers take the role of data (but of course one does not really compute any probabilities quantitatively here!).

(4) Regression: Consider two measurements, such as height and weight. It is reasonable to say that weight and height are positively correlated (if the height is larger, the weight tends to be larger too), but is there a more quantitative relationship? Can we predict the weight (roughly) from the height? One could try to see if a linear function fits: wt. $= a$ ht. $+ b$ for some a, b. Or perhaps a more complicated fit such as wt. $= a$ ht. $+ b$ ht.$^2 + c$, etc. To see if this is a good fit, and to know what values of a, b, c to take, we need data. Thus, the problem is that we have some data (H_i, W_i), $i = 1, 2, \ldots, n$, and based on this data we try to find the best linear fit (or the best quadratic fit) etc.

As another example, consider the approximate law that the resistivity of a material is proportional to the temperature. What is the constant of proportionality (for a given material). Here we have a law that says $R = aT$ where a is not known. By taking many measurements at various temperatures we get data (T_i, R_i), $i = 1, 2, \ldots, n$. From this we must find the best possible a (if all the data points were to lie on a line $y = ax$, there would be no problem. In reality they never will, and that is why the choice is an issue!).

CHAPTER 11

Estimation

1. Introductory remarks

A general problem than encompasses many questions in statistics is that of finding the distribution from which a sample is drawn. What we get to see is the sample (also called data), what we do not know is the distribution. We can broadly consider two levels of ignorance of the distribution.

(1) *Non-parametric estimation problems*: Complete ignorance, or some minimal knowledge such as that the distribution is on positive real line or on the segment $[0,1]$, or on natural numbers, etc. For instance, if our data is on lifetimes of some electronic component, the distribution is on $(0,\infty)$. If the data is on family sizes, then it is a distribution on natural numbers.

(2) *Parametric estimation problems* In some situations, we know (by past studies, or by some scientific theory) that the distribution is of a certain kind, for example, Normal distribution or Hypergeometric distribution, etc. What we do not know are the values of the parameters. This narrows down our search to the values of the unknown parameters, which are few in number. For example, if our sample comprises the opinions of people on which of two candidates they will vote for in an election, then the underlying distribution is $\mathrm{Ber}(p)$, where p is the actual proportion in the population who will vote for the first candidate. We don't know p, and the goal is to estimate it.

Another way to state the difference is that in the non-parametric setting, there are infinitely many unknown parameters (namely $F(x)$, for $x \in \mathbb{R}$), whereas in the parametric setting there are finitely many parameters. In this course, we shall stick to parametric estimation problems.

2. Parametric estimation problems: the general setting

Consider the following examples.

(1) A coin has an unknown probability p of turning up head. We wish to determine the value of p. For this, we toss the coin 100 times and observe the outcomes. How to give a guess for the value of p based on the data?

(2) Can we guess the average height μ of all people in India by taking a random sample of 100 people and measuring their heights? We may assume that heights follow Normal distribution, but the mean and variance are unknown to begin with, and have to be estimated.

(3) A factory manufacture light bulbs whose lifetimes may be assumed to be exponential random variables with a mean life-time μ. We take a sample of 50 bulbs at random and measure their life-times $X_1, \ldots, X_{50}$. Based on this data, how can we present a reasonable guess for μ? We may want to do this so that the specifications can be printed on the product when sold.

The general setting: We have *data*, which we take to be[1] i.i.d. samples $X_1, \ldots, X_n$ from an unknown pmf/pdf $f_\theta(x)$. Here θ is a parameter whose value is fixed but *unknown to us*. All we know is that it ranges over some subset of real numbers (we may also have multiple unknown parameters, but for simplicity we have written just one). A *statistic* is a function of the data, in other words, it is of the form $T(X_1, \ldots, X_n)$. It is important to note that it can be computed from the data, i.e., it cannot depend on the unknown parameter(s). An *estimate* for θ is a statistic that we propose as a guess for θ. Usually an estimate for the parameter θ is denoted $\hat{\theta}$ (or $\hat{\theta}_1$, $\hat{\theta}_2$, etc., if we are considering multiple estimates).

In the three examples given above, the families of distributions are respectively

(1) Ber(p), $p \in [0,1]$. More explicitly, $f_p(x) = p^x(1-p)^{1-x}$ for $x \in \{0,1\}$.

(2) $N(\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma^2 > 0$. Here $f_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, $x \in \mathbb{R}$.

(3) Exp($1/\theta$), $\theta > 0$. Here $f_\theta(x) = \frac{1}{\theta} e^{-x/\theta}$.

Remark 21. Observe that there are many ways to parameterize the same collection of probability distributions. How we parameterize does not matter, provide we expand our goal to say that what we want to estimate is $g(\theta)$, for some function g. For instance, the third example above remains the same if we write the family of distributions as Exp(λ), $\lambda > 0$, and that we want to estimate $g(\lambda) = 1/\lambda$. But to keep things simple, we usually parameterize in such a way that θ is what we want to estimate.

Location and scale families: Let f be a probability density on $\mathbb{R}$. Then we can create parametric families of distributions from it as follows.

(1) *Location family* of probability distributions $f_\mu(x) = f(x-\mu)$, $\mu \in \mathbb{R}$. Observe that if $X \sim f$ then $X + \mu \sim f_\mu$.

(2) *Scale family* $f_\sigma(x) = \frac{1}{\sigma} f\left(\frac{x}{\sigma}\right)$, $\sigma > 0$. Observe that if $X \sim f$ then $\sigma X \sim f_\sigma$.

(3) *Location-scale family* $f_{\mu,\sigma}(x) = \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$, $\mu \in \mathbb{R}$, $\sigma > 0$. Observe that if $X \sim f$ then $\sigma X + \mu \sim f_\mu$.

Many natural examples are of this form. For example, Exp(λ) distributions form a scale family, $N(\mu, \sigma^2)$ form a location-scale family.

[1] In general, i.i.d. assumption is not made. In fact, abstractly we can treat the whole data as one point $X = (X_1, \ldots, X_n)$ having joint pmf/pdf h_θ on $\mathbb{R}^n$. In the i.i.d. case $h_\theta(x_1, \ldots, x_n) = f_\theta(x_1) \ldots f_\theta(x_n)$.

3. Methods of estimation

Assume that $X_1, \dots, X_n$ are i.i.d. from pmf/pdf f_θ. We present three general methods to estimate θ.

3.1. Method of moments or quantiles. We first figure out what feature of the the distribution is θ. Then we take the estimator $\hat{\theta}$ to be the same feature of the sample. For example, if θ is the mean of f_θ, then we take $\hat{\theta}$ to be the sample mean $\bar{X}_n$; if θ is the median of f_θ, we take $\hat{\theta}$ to be the sample median M_n (which is the middle point among the X_is if n is odd and the average of the middle two points if n is even). Since there may be many features that give θ, there is no unique choice.

Example 22. Let $X_1, \dots, X_n$ be i.i.d. Exp$(1/\theta)$, $\theta > 0$. Then $\theta = \mathbb{E}[X_1]$, hence one possible estimator is $\bar{X}_n$. But we may also notice that the median of Exp$(1/\theta)$ is $\theta \log 2$, therefore, another estimator for θ is $M_n/\log 2$, where M_n is the sample median. From the fat that s.d.$_\theta(X_1) = \theta$, we may take the sample variance $\sqrt{\frac{1}{n}\sum_{k=1}^n (X_k - \bar{X}_n)^2}$ as another estimator. Further, for any $0 < p < 1$, the pth quantile of Exp$(1/\theta)$ distribution is $\theta \log \frac{1}{1-p}$, which means that the sample pth quantile divided by $\log \frac{1}{1-p}$ is another estimator for θ (for example, if $p = 0.9$, it means we take the number Q such that 10% of the X_is are above Q, and set $\hat{\theta} = \frac{Q}{\log(1/(1-p))}$.

At the heart of it, the method of moments/quantiles is just common sense combined with the law of large numbers. The law of large numbers tells us that features of a large sample imitate the features of the underlying distribution. By equating a feature of the underlying distribution to the corresponding feature of the sample, we get an estimator for the unknown parameters. As we saw, there is no unique choice, but there are reasons to use lower moments instead of higher, and median instead of an quantiles on either extreme.

3.2. Maximum likelihood method. The joint density of $X_1, \dots, X_n$ is $h_\theta(x_1, \dots, x_n) = f_\theta(x_1) \dots f_\theta(x_n)$. We evaluate the joint density at the observed data values, and treat is as a function of θ. This is called the *likelihood function*. In other words

$$L_X(\theta) = f_\theta(X_1) \dots f_\theta(X_n)$$

where we have written X as a subsctipt to mean $X = (X_1, \dots, X_n)$. It usually simplifies calculations to take the logarithm and define the *log-likelihood function*

$$\ell_X(\theta) := \log L_X(\theta) = \sum_{k=1}^n \log f_\theta(X_k).$$

When θ is the actual value, then $L_X(\theta)$ is the "likelihood" of seeing the data that we have actually observed. The *maximum likelihood estimate* (usually called MLE) is that value of θ that maximizes the likelihood function or equivalently the log-likelihood function. In symbols

$$\hat{\theta}_{\text{MLE}} = \arg\max_\theta \ell_X(\theta).$$

In principle there may be multiple maximizers, in which case any choice can be called the MLE.

EXAMPLE 23. Again consider the case of i.i.d. Exp(λ) samples $X_1, \ldots, X_n$. Then $f_\lambda(x) = \lambda e^{-\lambda x}$ and hence $\ell_X(\lambda) = n \log \lambda - \lambda \sum_{k=1}^n X_k = n \log \lambda - n\lambda \bar{X}_n$. To maximize, we set the derivative to zero to get

$$0 = \frac{d}{d\mu} \ell_X(\lambda) = \frac{n}{\lambda} - n\bar{X}_n$$

which is satisfied uniquely when $\lambda = \frac{1}{\bar{X}_n}$. One can use the second derivative condition to check that this is a point of maximum. More easily, one may also see that $\ell_X(\lambda) \to -\infty$ as $\lambda \to 0$ or as $\lambda \to \infty$, which means that the unique point where the derivative vanishes must be the global maximum. In conclusion, $\hat{\lambda}_{\text{MLE}} = \frac{1}{\bar{X}_n}$.

REMARK 24. Although we considered the same example of Exponential family of distributions to illustrate the method of moments/quantiles and the MLE, on the first case we parameterized by the mean parameter θ whereas in the second case we parameterized by $\lambda = \frac{1}{\theta}$. Why?

If we use θ as the parameter, the MLE does not change (or more precisely, $\hat{\theta}_{\text{MLE}}$ will come out to be equal to $\frac{1}{\hat{\lambda}_{\text{MLE}}}$. Changing the parametrization does not affect the MLE result.

If we use λ as the parameter, the method of quantile estimates will be fine, but the method of moments estimators as stated will be problematic. According to our definition, we would need some function T so that $\mathbb{E}_\lambda[T(X_1)] = \lambda$ for all $\lambda > 0$, and then the method of moments estimator would be $\frac{1}{n}(T(X_1) + \ldots + T(X_n))$. But the problem is that no such T exists![2]

Two more examples.

EXAMPLE 25. Let $X_1, \ldots, X_n$ be i.i.d. Ber(p) random variables Estimate p.

As $p = \mathbb{E}_p[X_1]$, a method of moments estimator would be the sample mean $\overline{X}_n$. In this case, $\mathbb{E}_p[X_1^2] = p$ again but we don't get any new estimate because $X_k^2 = X_k$ (as X_k is 0 or 1).

As the Bernoulli pmf can be written as $p^x(1-p)^{1-x}$ for $x \in \{0, 1\}$, the likelihood function is

$$\ell_X(p) := \log \prod_{k=1}^n p^{X_k}(1-p)^{1-X_k} = n\overline{X}_n \log p + n(1-\overline{X}_n) \log(1-p).$$

We need to find the value of p that maximizes $\ell_X(p)$. Using Calculus, one may show that $\hat{p}_{\text{MLE}} = \bar{X}_n$, but here is another instructive approach. Fix $a \in (0, 1)$ and use the (strict) concavity of logarithm to see that

$$a \log \frac{p}{a} + (1-a) \log \frac{1-p}{1-a} \leq \log \left(a\frac{p}{a} + (1-a)\frac{1-p}{1-a} \right) = 0$$

[2]To see this, observe that what we are asking for is that $\lambda = \int_0^\infty T(x)\lambda e^{-\lambda x} dx$ for all $\lambda > 0$, or equivalently that $\int_0^\infty T(x)e^{-\lambda x} dx = 1$ for all $\lambda > 0$. Perhaps the easiest way to convince yourself that this is not possible is to prove that the integral must go to 0 as $\lambda \to \infty$.

with equality if and only if $\frac{p}{a} = \frac{1-p}{1-a}$ (which is the same as $p = a$). Rearranging,

$$a \log p + (1-a) \log (1-p) \leq a \log a + (1-a) \log(1-a)$$

with equality if and only if $p = a$. Thus the $p \mapsto a \log p + (1-a) \log (1-p)$ is maximized uniquely at $p = a$. This was the maximization problem we wanted to solve with $\bar{X}_n$ in place of a (in the special cases when $a = 0$ or $a = 1$, the log-likelihood function is decreasing or increasing respectively and hence we get the maximizer to be a again).

EXAMPLE 26. Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$, where both $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ are unknown. From the Gaussian density, we see that the log-likelihood function is

$$\ell_X(\mu, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{k=1}^{n} (X_k - \mu)^2 - \frac{n}{2} \log(2\pi\sigma^2).$$

Irrespective of the value of σ^2, the quantity $\sum_{k=1}^{N} (X_k - \mu)^2$ is minimized over μ when $\mu = \bar{X}_n$. Therefore, $\hat{\mu}_{\text{MLE}} = \bar{X}_n$. Now plug this in and using Calculus or otherwise, minimize

$$-\frac{1}{2\sigma^2} \sum_{k=1}^{n} (X_k - \bar{X}_n)^2 - \frac{n}{2} \log(2\pi\sigma^2)$$

to see that $\hat{\sigma^2}_{\text{MLE}} = S_n^2 = \frac{1}{n} \sum_{k=1}^{n} (X_k - \bar{X}_n)^2$ (observe that this is *not* the unbiased estimator $s_n^2 = \frac{1}{n-1} \sum_{k=1}^{n} (X_k - \bar{X}_n)^2$).

In homeworks and tutorials you will see several other estimation problems. Let us just remark here that in general, there is no reason to expect that we can get a closed form expression for the MLE. Indeed, if data points are i.i.d. from the shifted Cauchy distribution with density $f_\theta(x) = \frac{1}{\pi(1+(x-\theta)^2)}$, then $\ell_X(\theta) = \sum_{k=1}^{n} \log(1 + (X_k - \theta)^2) - n \log \pi$. When we set derivative to zero, we get $0 = \sum_{k=1}^{n} \frac{2(X_k - \theta)}{1+(X_k-\theta)^2}$. There is no closed form solution. Quite often, the computation of MLE has to be done on a computer.

3.3. Bayes' estimator. As before, $X_1, \ldots, X_n$ are i.i.d. f_θ, and $\theta \in I$ is unknown. We take a probability distribution ρ on I, called the *prior distribution*. We pretend that (or really believe that) the unknown θ is random (so we write the random variable as Θ), and that nature has selected its value at random from the probability distribution ρ. Once chosen, it remains fixed, and conditional on its value $\Theta = \theta$, the X_is are i.i.d. from f_θ. Then the joint pmf/pdf of $(\Theta; X_1, \ldots, X_n)$ is

$$\rho(\theta) f_\theta(x_1) f_\theta(x_1) \ldots f_\theta(x_n).$$

What we see is the data $X_1, \ldots, X_n$ and we must guess what the value of Θ is likely to be. Bayes' rule tells us that the conditional pmf/pdf of Θ given $(X_1, \ldots, X_n)$, called the *posterior distribution*, is

$$\rho(\theta \mid X_1, \ldots, X_n) = \frac{1}{Z(X_1, \ldots, X_n)} \rho(\theta) f_\theta(X_1) f_\theta(X_1) \ldots f_\theta(X_n)$$

where

$$Z(x_1,\dots,x_n) = \begin{cases} \sum_\varphi \rho(\varphi) f_\theta(x_1) f_\theta(x_1)\dots f_\theta(x_n) & \text{in case of pmf,} \\ \int \rho(\varphi) f_\theta(x_1) f_\theta(x_1)\dots f_\theta(x_n) d\varphi & \text{in case of pdf.} \end{cases}$$

The conditional pmf has all we want. For example, its mean can be taken as an estimator of the unknown parameter:

$$\begin{aligned} \hat{\theta}_{\text{Bayes}} &= \mathbb{E}[\Theta \mid X_1,\dots,X_n] \\ &= \begin{cases} \frac{1}{Z(X_1,\dots,X_n)} \sum_\varphi \varphi\rho(\varphi) f_\varphi(X_1) f_\varphi(X_1)\dots f_\varphi(X_n) & \text{in case of pmf,} \\ \frac{1}{Z(X_1,\dots,X_n)} \int \varphi\rho(\varphi) f_\varphi(X_1) f_\varphi(X_1)\dots f_\varphi(X_n) d\varphi & \text{in case of pdf.} \end{cases} \end{aligned}$$

The standard deviation of the posterior distribution is a measure of our uncertainty in the knowledge of the true value of the parameter.

The procedure will become clear in the examples. The main issue, sometimes a contentious one, is how to choose the prior distribution? That does not come with the problem and is a choice made by the statistician, and is supposed to indicate our prior knowledge of θ.

Example 27. Take the case of the Ber(p) sample. Since $p \in [0,1]$, one natural choice for the prior distribution is Unif$[0,1]$ (which indicates that we are not prejudiced towards some specific values of p over others, but think that everything is equally likely). So $\rho(p) = 1$ for $p \in [0,1]$. The posterior pmf is

$$\rho(p \mid X_1,\dots,X_n) = \frac{1}{Z(X_1,\dots,X_n)} p^{n\bar{X}_n}(1-p)^{n(1-\bar{X}_n)}$$

where

$$Z(x_1,\dots,x_n) = \int p^{n\bar{X}_n}(1-p)^{n(1-\bar{X}_n)} dp \;=\; \text{Beta}(n\bar{X}_n + 1, n(1-\bar{X}_n) + 1).$$

Thus, the posterior distribution is precisely the Beta$(n\bar{X}_n+1, n(1-\bar{X}_n)+1)$ distribution on $[0,1]$. We know that the mean and variance of the Beta(a,b) distribution are $\frac{a}{a+b}$ and $\frac{ab}{(a+b)^2(a+b+1)}$. Therefore,

$$\hat{\theta}_{\text{Bayes}} = \frac{n\bar{X}_n + 1}{n+2} = \frac{n}{n+2} \times \bar{X}_n + \frac{2}{n+2} \times \frac{1}{2}.$$

We see that the Bayes' estimator is a convex combination of the sample mean and the mean of the prior distribution with weights $\frac{n}{n+2}$ and $\frac{2}{n+2}$ respectively. When we have no data ($n = 0$), our guess is the mean of the prior distribution, and when we have a large amount of data (large n), our guess is close to the sample mean. The uncertainty in the Bayes' estimator is the standard deviation of the posterior distribution, which is the square root of

$$\frac{(n\bar{X}_n + 1)(n(1-\bar{X}_n) + 1)}{(n+2)^2(n+3)} \sim \frac{\bar{X}_n(1-\bar{X}_n)}{n}$$

for large n.

EXAMPLE 28. Let $X_1, \ldots, X_n$ be i.i.d. Pois(λ) where $\lambda > 0$. Let us choose the prior distribution to be Exp(α) for some $\alpha > 0$. Then the posterior density of λ is

$$\begin{aligned}\rho(\lambda \mid X_1, \ldots, X_n) &= \frac{1}{Z(X_1, \ldots, X_n)} \alpha e^{-\alpha\lambda} \prod_{k=1}^{n} e^{-\lambda} \frac{\lambda^{X_k}}{X_k!} \\ &= \frac{\alpha}{Z(X_1, \ldots, X_n) X_1! \ldots X_n!} e^{-(n+\alpha)\lambda} \lambda^{n\bar{X}_n}.\end{aligned}$$

We can compute $Z(X_1, \ldots, X_n)$, but we don not have to. We see that the density has the form $e^{-(n+\alpha)\lambda}\lambda^{n\bar{X}_n}$, so it must be the Gamma($n\bar{X}_n + 1, n + \alpha$) density, which tells us what the normalization constant is. Recalling that the mean of the Gamma distribution, we see that

$$\hat{\theta}_{\text{Bayes}} = \frac{n\bar{X}_n + 1}{n + \alpha} = \frac{n}{n + \alpha} \times \bar{X}_n + \frac{\alpha}{n + \alpha} \times \frac{1}{\alpha}.$$

Again, this is a convex combination of the sample mean $\bar{X}_n$ and the mean of the prior distribution $\frac{1}{\alpha}$. The more the data we have, the more the weight on the sample mean.

REMARK 29. The last example illustrates our earlier point about choosing the prior distribution. Why did we choose it to be Exponential distribution? So that the posterior distribution comes out nice! If you take an arbitrary prior distribution, then the posterior distribution becomes intractable (we would not recognize its form, we would not be able to compute $Z(X_1, \ldots, X_n)$). But is there any deeper reason why our prior knowledge must be captured by the Exponential distribution? Depending on how staunch a "Bayesian" you are, you can make up reasons to convince yourself that it must be so (incidentally, it is not only Exponential, you may also take the prior to be a general Gamma distribution and find that the posterior is again a Gamma distribution).

3.4. Further examples and exercises.

EXERCISE 30. Find an estimate for the unknown parameters by the method of moments and the maximum likelihood method.

(1) $X_1, \ldots, X_n$ are i.i.d. $N(\mu, 1)$. Estimate μ. How do your estimates change if the distribution is $N(\mu, 2)$?

(2) $X_1, \ldots, X_n$ are i.i.d. $N(0, \sigma^2)$. Estimate σ^2. How do your estimates change if the distribution is $N(7, \sigma^2)$?

(3) $X_1, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$. Estimate μ and σ^2.

[**Note:** The first case is when σ^2 is known and μ is unknown. Then the known value of σ^2 may be used to estimate μ. In the second case it is similar, now μ is known and σ^2 is not known. In the third case, both are unknown].

EXERCISE 31. $X_1, \ldots, X_n$ are i.i.d. Geo(p) Estimate $\mu = 1/p$.

EXERCISE 32. $X_1, \ldots, X_n$ are i.i.d. Pois(λ) Estimate λ.

EXERCISE 33. $X_1, \ldots, X_n$ are i.i.d. Beta(a, b) Estimate a, b.

EXERCISE 34. $X_1, \ldots, X_n$ are i.i.d. Uniform$[a, b]$ Estimate a, b.

4. Quality of an estimate

We have seen that there may be several competing estimates that can be used to estimate a parameter. How can one choose between these estimates? In this section we present some properties that may be considered desirable in an estimator. However, having these properties does not lead to an unambiguous choice of one estimate as the best for a problem.

Let T_n be an estimator for θ. If the data consists of i.i.d. $X_1, \ldots, X_n$ distributed according to f_θ, then this just means that T_n is a function of $X_1, \ldots, X_n$. The *mean squared error* of T_n is defined as

$$\text{m.s.e.}_{T_n}(\theta) = \mathbb{E}_\theta[(T_n - \theta)^2].$$

This is a function of θ. Smaller it is, better our estimate. But observe that it is not a number but a function of θ, hence it may not unambiguously help us to choose between two estimates.

EXAMPLE 35. Let $X_1, \ldots, X_n$ be i.i.d. Ber(p). Let $T_n = \bar{X}_n$ and $R_n = \frac{X_1+X_2}{2}$ and $S_n = \frac{1}{2}$. All of them can be used as estimates for p. What are their mean squared errors?

First, $\text{m.s.e.}_{T_n}(p) = \mathbb{E}[(\bar{X}_n - p)^2] = \frac{p(1-p)}{n}$ (check!). For the same reason with $n = 2$, we see that $\text{m.s.e.}_{R_n}(p) = \frac{p(1-p)}{2}$. Lastly, $\text{m.s.e.}_{S_n}(p) = (\frac{1}{2} - p)^2$.

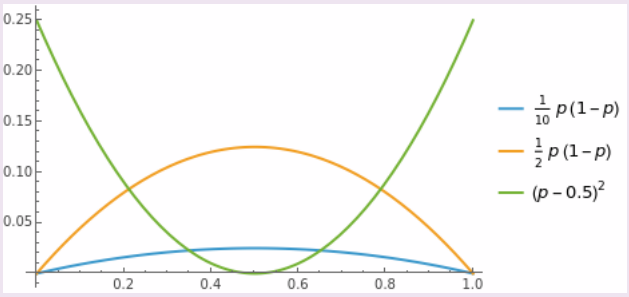

FIGURE 1. Mean squared errors of T_n, R_n and S_n with $n = 10$ in the Bernoullis example. While T_n unambiguously beats R_n, there are values of p for which S_n does better than T_n and other values of p for which T_n does better than S_n.

Two components of the mean squared error: Define the *bias* of the estimator T_n as $B_{T_n}(\theta) := \mathbb{E}_\theta[T_n] - \theta$ and write its variance as $V_{T_n}(\theta) = \text{Var}_\theta(T_n)$. If $\text{Bias}_{T_n}(\theta) = 0$ for all values of the parameter θ then we say that T_n is *unbiased* for θ. Here we write θ in the subscript of $\mathbb{E}_\theta$ to remind ourself that in computing the expectation we use the density f_θ. However we shall often omit the subscript for simplicity.

CLAIM 36. *For any estimator T_n, we have m.s.e.$_{T_n}(\theta) = V_{T_n}(\theta) + (B_{T_n}(\theta))^2$.*

PROOF. Consider any random variable Y with mean μ and observe that for any real number a and write $Y - a$ as $(Y - \mu) + (\mu - a)$ to get

$$\begin{aligned}\mathbb{E}[(Y-a)^2] &= \mathbb{E}[(Y-\mu)^2] + (\mu-a)^2 + 2(\mu-a)\mathbb{E}[Y-\mu] \\ &= \mathrm{Var}(Y) + (\mu-a)^2\end{aligned}$$

as the last term vanishes ($\mathbb{E}[Y - \mu] = 0$). Use this identity with T_n in place of Y and θ in place of a and evaluate expectations under the distribution f_θ to get the claim. ■

REMARK 37. An analogy. Consider shooting with a rifle having a telescopic sight. A given target can be missed for two reasons. One, the marksman may be unskilled and shoot all over the place, sometimes a meter to the right of the target, sometimes a meter to the left, etc. In this case, the shots have a large variance (but possibly right on spot on average!). Another person may consistently hit a point 20 cm. to the right of the target. Perhaps the telescopic sight is not set right, and this caused the systematic error. This is the bias. Both bias and variance contribute to missing the target.

EXAMPLE 38. Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$. Let $V_n = \frac{1}{n}\sum_{k=1}^n (X_k - \overline{X}_n)^2$ be an estimate for σ^2. By expanding the squares we get

$$V_n = \overline{X}_n^2 + \frac{1}{n}\sum_{k=1}^n X_k^2 - \frac{2}{n}\overline{X}_n \sum_{k=1}^n X_k = \left(\frac{1}{n}\sum_{k=1}^n X_k^2\right) - \overline{X}_n^2.$$

It is given that $\mathbb{E}[X_k] = \mu$ and $\mathrm{Var}(X_k) = \sigma^2$. Hence $\mathbb{E}[X_k^2] = \mu^2 + \sigma^2$. We have seen before that $\mathrm{Var}(\overline{X}_n) = \sigma^2$ and $\mathbb{E}[\overline{X}_n] = \mu$. Hence $\mathbb{E}[\overline{X}_n^2] = \mu^2 + \frac{\sigma^2}{n}$. Putting all this together, we get

$$\mathbb{E}[V_n] = \left(\frac{1}{n}\sum_{k=1}^n \mu^2 + \sigma^2\right) - \left(\mu^2 + \frac{\sigma^2}{n}\right) = \frac{n-1}{n}\sigma^2.$$

Thus, the bias of V_n is $\frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{1}{n}\sigma^2$.

EXAMPLE 39. For the same setting as the previous example, suppose $W_n = \frac{1}{n}\sum_{k=1}^n (X_k - \mu)^2$. Then it is easy to see that $\mathbb{E}[W_n] = \sigma^2$. Can we say that W_n is an unbiased estimate for σ^2? There is a hitch!

If the value of μ is unknown, then W_n is *not* an estimate (cannot compute it using $X_1, \ldots, X_n$!). However if μ is known, then it is an unbiased estimate. For example, if we knew that $\mu = 0$, then $W_n = \frac{1}{n}\sum_{k=1}^n X_k^2$ is an unbiased estimate for σ^2.

When μ is unknown, we define $s_n^2 = \frac{1}{n-1}\sum_{k=1}^n (X_k - \overline{X}_n)^2$. Clearly $s_n^2 = \frac{n}{n-1}V_n$ and hence $\mathbb{E}[s_n^2] = \frac{n}{n-1}\mathbb{E}[V_n] = \sigma^2$. Thus, s_n^2 is an unbiased estimate for σ^2. Note that s_n^2 depends only on the data and hence it is an estimate, whether μ is known or unknown.

All the remarks in the above two examples apply for any distribution, i.e.,

(1) The sample mean is unbiased for the population mean.

(2) The sample variance $s_n^2 = \frac{1}{n-1}\sum_{k=1}^n (X_k - \overline{X}_n)^2$ is unbiased for the population variance. But $V_n = \frac{1}{n}\sum_{k=1}^n (X_k - \overline{X}_n)^2$ is not, in fact $\mathbb{E}[V_n] = \frac{n-1}{n}\sigma^2$.

It appears that s_n^2 is better, but the following remark says that one should be cautious in making such a statement.

REMARK 40. In case of $N(\mu, \sigma^2)$ data, it turns out that although s_n^2 is unbiased and V_n is biased, the mean squared error of V_n is smaller! Further V_n is the maximum likelihood estimate of σ^2! Overall, unbiasedness is not so important as having smaller mean squared error, but for estimating variance (when the mean is not known), we always use s_n^2. The computation of the m.s.e is a bit tedious, so we skip it here.

EXAMPLE 41. Let $X_1, \ldots, X_n$ be i.i.d. Ber(p). Then $\overline{X}_n$ is an estimate for p. It is unbiased since $\mathbb{E}[\overline{X}_n] = p$. Hence, the m.s.e of $\overline{X}_n$ is just the variance which is equal to $p(1-p)/n$.

A puzzle: A coin C_1 has probability p of turning up head and a coin C_2 has probability $2p$ of turning up head. All we know is that $0 < p < \frac{1}{2}$. You are given 20 tosses. You can choose all tosses from C_1 or all tosses from C_2 or some tosses from each (the total is 20). If the objective is to estimate p, what do you do?

Solution: If we choose to have all $n = 20$ tosses from C_1, then we get $X_1, \ldots, X_n$ that are i.i.d. Ber(p). An estimate for p is $\overline{X}_n$ which is unbiased and hence $\text{MSE}_{\overline{X}_n}(p) = \text{Var}(\overline{X}_n) = p(1-p)/n$. On the other hand if we choose to have all 20 tosses from C_2, then we get $Y_1, \ldots, Y_n$ that are i.i.d. Ber($2p$). The estimate for p is now $\bar{Y}_n/2$ which is also unbiased and has $\text{MSE}_{\bar{Y}_n/2}(p) = \text{Var}(\bar{Y}_n)/4 = 2p(1-2p)/4n = p(1-2p)/2n$. As $1-2p < 1-p$, we see that $\text{MSE}_{\bar{Y}_n/2}(p) < \text{MSE}_{\overline{X}_n}(p)$ and hence choosing C_2 is better, at least by mean-squared criterion! It can be checked that if we choose to have k tosses from C_1 and the remaining $n-k$ from C_2, the MSE of the corresponding estimate will be between the two MSEs found above and hence not better than $\bar{Y}_n/2$.

Another puzzle: A factory produces light bulbs having an exponential distribution with mean μ. Another factory produces light bulbs having an exponential distribution with mean 2μ. Your goal is to estimate μ. You are allowed to choose a total of 50 light bulbs (all from the first or all from the second or some from each factory). What do you do?

Solution: If we pick all $n = 50$ bulbs from the first factory, we see $X_1, \ldots, X_n$ i.i.d. Exp($1/\mu$). The estimate for μ is $\overline{X}_n$ which has $\text{MSE}_{\overline{X}_n}(\mu) = \text{Var}(\overline{X}_n) = \mu^2/n$. If we choose all bulbs from factory 2 we get $Y_1, \ldots, Y_n$ i.i.d. Exp($1/2\mu$). The estimate for μ is $\bar{Y}_n/2$. But $\text{MSE}_{\bar{Y}_n/2}(\mu) = \text{Var}(\bar{Y}_n/2) = (2\mu)^2/4n = \mu^2/n$. The two mean-squared errors are exactly the same!

Probabilistic thinking: Is there any calculation-free explanation why the answers to the two puzzles are as above? Yes, and it is illustrative of what may be called probabilistic thinking. Take the second puzzle. Why are the two estimates same by mean-squared error? Is one better by some other criterion?

Recall that if $X \sim \text{Exp}(1/\mu)$ then $X/2 \sim \text{Exp}(2/\mu)$ and vice versa. Therefore, if we have data from $\text{Exp}(1/\mu)$ distribution, then we can divided all the numbers by 2 and convert it into data from $\text{Exp}(2/\mu)$ distribution. Conversely if we have data from $\text{Exp}(2/\mu)$ distribution, then we can convert it into data from $\text{Exp}(1/\mu)$ distribution by multiplying each number by 2. Hence there should be no advantage in choosing either factory. We leave it for you to think in analogous ways why in the first puzzle C_2 is better than C_1.

5. Best estimators?

Let S, T be two estimators of θ. If $\text{m.s.e.}_S(\theta) \leq \text{m.s.e.}_T(\theta)$ for all θ, then clearly we would choose S over T. But as we saw in Figure 35, the inequality may go either way for different values of θ, in which case we cannot choose one over another. The problem is that numbers are comparable, but functions are not necessarily comparable.

5.1. Minimax estimate. One way out is to reduce the function $\text{m.s.e.}_{T_n}(\theta)$ to a single number, and compare those numbers. But which number to reduce to? One choice is to take the worst case, i.e., define the *maximum risk* $R_{T_n} := \max_\theta \text{m.s.e.}_{T_n}(\theta)$ (another would be to integrate over θ, to get some form of "average risk"). Ignoring the possibility that the maximum is a supremum and that it can be infinite, this is a number associated to T_n. Hence we can compare any two estimators T_n and S_n using their maximum risks. In particular, if there is one estimator that minimizes the maximum risk, we call it a *minimax estimate*. Note that T_n^* is a minimax estimate if and only if $\max_\theta \text{m.s.e.}_{T_n^*}(\theta) \leq \max_\theta \text{m.s.e.}_{T_n}(\theta)$ for any estimator T_n.

Many of the estimates we have obtained (e.g., sample mean for Bernoulli and Normal samples) are minimax estimates. But we do not wish to go further in this line. One criticism of the minimax estimate is that taking the worst case is a pessimistic approach.

5.2. Best unbiased estimator. As we have seen, mean squared error has two components, bias and variance. Although intuitively it may seem that unbiasedness is desirable, it is not useful to have it at the cost of a much higher mean squared error. Further, unbiasedness is delicate in that if T_n is an unbiased estimator of θ, then $\frac{1}{T_n}$ is not an unbiased estimator of $\frac{1}{\theta}$. In fact, there are situations ($\text{Exp}(\lambda)$, $\lambda > 0$, and $\text{Geo}(p)$, $p \in [0,1]$), where unbiased estimators don't exist!

Nevertheless, insisting on unbiasedness rules out absurd estimators like the constant estimator $S_n = 0.5$ in the example of the Bernoulli sample. Further, it leads to some nice mathematics in which we indulge a little now! Keep in mind that for unbiased estimators, the variance and mean squared error are the same.

Consider a single parameter family f_θ, $\theta \in \Theta \subseteq \mathbb{R}$. Recall the log-likelihood function $\ell_X(\theta)$ which is $\sum_{k=1}^n \log f_\theta(X_k)$ in case of the i.i.d. sample. If $\theta \mapsto \ell_X(\theta)$ is differentiable, define

Fisher information as the function

$$I(\theta) = \mathrm{Var}_\theta\left(\frac{d}{d\theta}\ell_X(\theta)\right).$$

THEOREM 42 (Cramer-Rao bound). *Let T_n be any unbiased estimator of θ. Then*

$$Var_\theta(T_n) \geq \frac{1}{I(\theta)}.$$

To understand it, observe that if f_θ for nearby θ are very close to each other, then

(1) $I(\theta)$ must be small, because $\ell_X(\theta)$ changes slowly with θ.

(2) $\mathrm{Var}_\theta(T_n)$ must be large, because it becomes harder for T_n or any estimate to distinguish between f_θ for nearby values of θ.

Cramer-Rao bound quantifies this inverse relationship between the two quantities.

Before going into the proof of the Cramer-Rao inequality, let us see how it can be useful by an example.

EXAMPLE 43. Let $X_1, \ldots, X_n$ be i.i.d. Ber(p). Then $\ell_X(p) = n\bar{X}_n \log p + n(1-\bar{X})\log(1-p)$. Therefore, $\frac{d}{dp}\ell_X(p) = \frac{n\bar{X}_n}{p} - \frac{n(1-\bar{X}_n)}{1-p} = \frac{n}{p(1-p)}\bar{X}_n - \frac{n}{1-p}$. Therefore, the Fisher information is given by

$$I(p) = \frac{n^2}{p^2(1-p)^2}\mathrm{Var}_p(\bar{X}_n) = \frac{n}{p(1-p)}.$$

But (as used above) $\mathrm{Var}_p(\bar{X}_n) = \frac{p(1-p)}{n}$ which coincides with the Crame-Rao lower bound $\frac{1}{I(p)}$ for all p. This means that no unbiased estimator can do better than $\bar{X}_n$, for any value of the parameter. In other words, $\bar{X}_n$ is the[3] best unbiased estimator of p.

In short, if we can find an unbiased estimator that attains the Cramer-Rao lower bound, then it is the best unbiased estimator (according to mean squared error criterion). The technical name for such an estimator is *UMVUE* (Uniformly Minimum Variance Unbiased Estimator). In several examples, they do exist.

EXERCISE 44. Check if the Cramer-Rao lower bound is achieved by $\bar{X}_n$ in the following examples: (a) To estimate μ in $N(\mu, 1)$, $\mu > 0$. (b) To estimate μ in $\mathrm{Exp}(1/\mu)$, $\mu > 0$. (c) To estimate λ in $\mathrm{Pois}(\lambda)$, $\lambda > 0$.

PROOF OF CRAMER-RAO BOUND. Let T_n be an unbiased estimator of θ. Let $\dot{\ell}_X(\theta) = \frac{d}{d\theta}\ell_X(\theta)$ (this need not be an estimator as it could depend on θ). By Cauchy-Scwharz inequality

$$|\mathrm{Cov}_\theta(T_n, \dot{\ell}_X(\theta))|^2 \leq \mathrm{Var}_\theta(T_n)\mathrm{Var}_\theta(\dot{\ell}_X(\theta)) = \mathrm{Var}_\theta(T_n)I(\theta)$$

where the equality follows from the definition of Fisher information. Thus it suffices to show that the covariance on the left side is equal to 1. We compute it now. The joint pmf/pdf of

[3] One may question the use of "The" here. Could there not be another unbiased estimator which is also best? At this point, yes, but the proof of the Cramer-Rao bound shows that there is no such estimator other than $\bar{X}_n$.

$(X_1, \ldots, X_n)$ is $e^{\ell_x(\theta)}$ where $x = (x_1, \ldots, x_n)$. For definiteness, let us assume that we are dealing with pmf. Then

$$\mathbb{E}[\dot{\ell}_X(\theta)] = \sum_x e^{\ell_x(\theta)} \frac{d}{d\theta} \ell_x(\theta) = \sum_x \frac{d}{d\theta} e^{\ell_x(\theta)} = \frac{d}{d\theta} \sum_x e^{\ell_x(\theta)} = 0$$

because the sum is 1 (as $x \mapsto e^{\ell_x(\theta)}$ is a pmf for any θ). This last step of taking the derivative out of the sum is obviously fine if the sum is over a finite set A (which requires that $f_\theta(x) = 0$ for $x \notin A$, for all θ). But if it is an infinite sum, interchanging the derivative and sum needs a justification, which we do not bother with here. Therefore,

$$\begin{aligned}
\mathrm{Cov}_\theta(T_n, \dot{\ell}_X(\theta)) &= \mathbb{E}\left[T_n \dot{\ell}_X(\theta)\right] \quad \text{(the product of expectations vanishes)} \\
&= \sum_x e^{\ell_x(\theta)} \frac{d}{d\theta} \ell_x(\theta) T_n(x) \\
&= \frac{d}{d\theta} \sum_x T_n(x) e^{\ell_x(\theta)}.
\end{aligned}$$

Again, the interchange of derivative and sum needs a justification that we skip. But then, the sum is precisely $\mathbb{E}_\theta[T_n]$ which is equal to θ as T_n is unbiased. Differentiating, we see that the covariance is 1. The proof is complete. ■

Remark 45. The equality condition in Cauchy-Schwarz shows that the Cramer-Rao bound is achieved if and only if $\frac{d}{d\theta}\ell_x(\theta) = A_\theta T_n(x) + B_\theta$ for some θ dependent constants A_θ, B_θ.

This makes it easy to guess what T_n, if any, could attain the Cramer-Rao bound. Just compute $\frac{d}{d\theta}\ell_X(\theta)$ and see if it can be written as $A_\theta T_n(X) + B_\theta$, where A_θ, B_θ depend only on θ and $T_n(X)$ is an unbiased estimator of θ. If that can be done (in general there is no reason why that should be possible), then the estimate we are looking for is T_n.

Example 46. Let $X_1, \ldots, X_n$ be i.i.d. Geo($1/\theta$) where $\theta > 1$. The pmf of each X_i is $(1 - \frac{1}{\theta})^k \frac{1}{\theta}$ for $k \geq 0$. Therefore, $\ell_X(\theta) = n\bar{X}_n \log(1 - \frac{1}{\theta}) - n \log \theta$. Then $\frac{d}{d\theta}\ell_x(\theta) = -\frac{n}{\theta(\theta-1)}\bar{X}_n - \frac{n}{\theta}$. As $\mathbb{E}_\theta[X_1] = \theta$, we also have $\mathbb{E}_\theta[\bar{X}_n] = \theta$, hence $\bar{X}_n$ is an unbiased estimator of θ. As the derivative of the log-likelihood is linearly related to $\bar{X}_n$, we see that the sample mean attains the Cramer-Rao lower bound and is therefore the unbiased estimator with the lowest mean squared error.

Digression to be ignored: Let $\psi : \mathbb{R} \to \mathbb{R}$ be a function such that $f = \psi^2$ is a probability density (such a ψ is called an *amplitude function* in Quantum mechanics). Consider the location family $f_\theta(x) = f(x - \theta)$, $\theta \in \mathbb{R}$. Now assume that ψ is differentiable and decays fast at $\pm\infty$.

Let μ be the mean of the density f. Then $\mu + \theta$ is the mean of f_θ, hence $X - \mu$ is an unbiased estimate of θ.

If we have a single sample $X \sim f_\theta$, then $\ell_X(\theta) = \log f(x-\theta)$ and hence $\frac{d}{d\theta}\ell_X(\theta) = \frac{f'(X-\theta)}{f(X-\theta)} = \frac{2\psi'(X-\theta)}{\psi(X-\theta)}$. As we know that $\mathbb{E}_\theta[\frac{d}{d\theta}\ell_X(\theta)] = 0$ in general (see the proof of Cramer-Rao inequality),

$$I(\theta) = \mathrm{Var}_\theta\left(\frac{2\psi'(X-\theta)}{\psi(X-\theta)}\right) = 4\int_{\mathbb{R}} \frac{\psi'(x-\theta)^2}{\psi(x-\theta)^2} f(x-\theta)d\theta$$

$$= 4\int_{\mathbb{R}} \psi'(x-\theta)^2 dx.$$

Putting together the above observations, we apply Cramer-Rao bound at $\theta = 0$ to get

$$\frac{1}{4} \leq \mathrm{Var}_0(X-\mu) \times I(0)$$

$$= \int_{\mathbb{R}} (x-\mu)^2 |\psi(x)|^2 dx \times \int_{\mathbb{R}} |\psi'(x)|^2 dx.$$

This is known as Heisenberg's uncertainty principle, although the second factor is usually written as $\int_{\mathbb{R}} y^2 |\hat{\psi}(y)|^2 dy$, where $\hat{\psi}(y) = \int_{\mathbb{R}} \psi(x) e^{-2\pi i x y} dx$ is the Fourier transform of ψ. By properties of Fourier transform, it is known that $|\hat{\psi}|^2$ is also a probability density and $\int_{\mathbb{R}} |\psi'|^2 = \int y^2 |\hat{\psi}(y)|^2 dy$. One can also write the right hand side as the product of variances of the densities $|\psi|^2$ and $|\hat{\psi}|^2$.

6. Confidence intervals

So far, in estimating of an unknown parameter, we gave a single number as our guess for the known parameter. Since we certainly don't mean that we are guessing the exact value, it would be better to give an idea of how much we could be off. More precisely, given and interval and say with what confidence we expect the true parameter to lie within it.

A *confidence interval* is a random interval $I = [A, B]$ where A, B are statistics (i.e., computable from the data, without knowledge of the parameter values). The interval I is said to have confidence level $1-\alpha$ if $\mathbb{P}_\theta(I \ni \theta) \geq 1-\alpha$ for all values of θ.

Usually we fix the level of confidence, say as 0.90 and find a confidence interval *as short as possible* but subject to the condition that it should have a confidence level of 0.90. The higher the confidence we want, the longer the confidence interval. If we don't want the the length to increase, the only way is to get more data.

Example 47. Suppose we have one sample X from $N(\mu, 1)$ distribution where $\mu \in \mathbb{R}$ is unknown. How do we estimate μ? Suppose the observed value of X is 2.7, then our guess for μ is 2.7 itself. Hence $[X-a, X+a]$ for some $a > 0$ is a natural choice of a confidence interval. What is its confidence?

$$\mathbb{P}\{[X-a, X+a] \ni \mu\} = \mathbb{P}\{-a \leq X-\mu \leq a\} = \Phi(a) - \Phi(-a) = 2\Phi(a) - 1.$$

For $a = 0$, the confidence (the CI is the size interval $[X, X]$) level is zero and as $a \uparrow \infty$, the confidence level goes up to 1.

In this section we consider the problem of confidence intervals in Normal population. In the next we see a few other examples.

The setting: Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$ random variables. We consider four situations.

(1) Confidence interval for μ when σ^2 is known.

(2) Confidence interval for σ^2 when μ is known.

(3) Confidence interval for μ when σ^2 is unknown.

(4) Confidence interval for σ^2 when μ is unknown.

A starting point in finding a confidence interval for a parameter is to first start with an estimate for the parameter. For example, in finding a CI for μ, we may start with $\overline{X}_n$ and enlarge it to an interval $[\overline{X}_n - a, \overline{X}_n + a]$. Similarly, in finding a CI for σ^2 we use the estimate $s_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$ if μ is unknown and $W_n = \frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2$ if the value of μ is known.

6.1. Estimating μ when σ^2 is known. We look for a confidence interval of the form $I_n = [\overline{X}_n - a, \overline{X}_n + a]$. Then,

$$\mathbb{P}\left(I_n \ni \mu\right) = \mathbb{P}\left(-a \le \overline{X}_n - \mu \le a\right) = \mathbb{P}\left(-\frac{a\sqrt{n}}{\sigma} \le \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \le \frac{a\sqrt{n}}{\sigma}\right)$$

Now we use two facts about normal distribution that we have seen before.

(1) If $Y \sim N(\mu, \sigma^2)$ then $aX + b \sim N(a\mu + b, a^2\sigma^2)$.

(2) If $Y_1 \sim N(\mu, \sigma^2)$ and $Y_2 \sim N(\nu, \tau^2)$ are independent, then $Y_1 + Y_2 \sim N(\mu + \nu, \sigma^2 + \tau^2)$.

Consequently, $\overline{X}_n \sim N(0, \sigma^2/n)$ and $\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \sim N(0,1)$. Therefore,

$$\mathbb{P}\left(I_n \ni \mu\right) = \mathbb{P}(-\frac{a\sqrt{n}}{\sigma} \le Z \le -\frac{a\sqrt{n}}{\sigma})$$

where $Z \sim N(0,1)$. Fix any $0 < \alpha < 1$ and denote by z_α the number such that $\mathbb{P}(Z > z_\alpha) = \alpha$ (in other words, z_α is the $(1-\alpha)$-quantile of the standard normal distribution). For example, from normal tables we find that $z_{0.05} \approx 1.65$ and $z_{0.005} \approx 2.58$ etc.

If we set $a = z_{\alpha/2}\sigma/\sqrt{n}$, we get

$$\mathbb{P}\left(\left[\overline{X}_n - \frac{\sigma}{\sqrt{n}}z_{\alpha/2}, \overline{X}_n + \frac{\sigma}{\sqrt{n}}z_{\alpha/2}\right] \ni \mu\right) = 1 - \alpha.$$

This is our confidence interval.

6.2. Estimating σ^2 when μ is known. Since μ is known, we use $W_n = \frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2$ to estimate σ^2. Here is an exercise.

Exercise 48. Let $Z_1, \ldots, Z_n$ be i.i.d. $N(0,1)$ random variables. Then, $Z_1^2 + \ldots + Z_n^2 \sim$ Gamma$(n/2, 1/2)$.

Solution: For $t > 0$ we have

$$\mathbb{P}\{Z_1^2 \le t\} = \mathbb{P}\{-\sqrt{t} \le Z_1 \le \sqrt{t}\} = 2\int_0^{\sqrt{t}} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du = \frac{1}{\sqrt{2\pi}}\int_0^t e^{-s/2} s^{-1/2} ds.$$

Differentiate w.r.t t to see that the density of Z_1^2 is $h(t) = \frac{1}{\sqrt{\pi}} e^{-t/2} t^{-1/2} \sqrt{(1/2)}$, which is just the Gamma$(\frac{1}{2}, \frac{1}{2})$ density.

Now, each Z_k^2 has the same Gamma$(\frac{1}{2}, \frac{1}{2})$ density, and they are independent. Earlier we have seen that when we add independent Gamma random variables with the same scale parameter, the sum has a Gamma distribution with the same scale but whose shape parameter is the sum of the shape parameters of the individual summands. Therefore, $Z_1^2 + \ldots + Z_n^2$ has Gamma$(n/2, 1/2)$ distribution. This completes the solution to the exercise.

In statistics, the distribution Gamma$(\frac{n}{2}, \frac{1}{2})$ is usually called the *chi-squared distribution with n degrees of freedom*. Let $\chi_n^2(\alpha)$ denote the $1-\alpha$ quantile of this distribution. Similarly, $\chi_n^2(1-\alpha)$ is the α quantile (i.e., the probability for the chi-squared random variable to fall below $\chi_n^2(1-\alpha)$ is exactly α).

When X_i are i.i.d. $N(\mu, \sigma^2)$, we know that $(X_i - \mu)/\sigma$ are i.i.d. $N(0,1)$. Hence, by the above fact, we see that

$$\frac{nW_n}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2$$

has chi-squared distribution with n degrees of freedom. Hence

$$\mathbb{P}\left\{ \frac{nW_n}{\chi_n^2\left(\frac{\alpha}{2}\right)} \leq \sigma^2 \leq \frac{nW_n}{\chi_n^2\left(1-\frac{\alpha}{2}\right)} \right\} = \mathbb{P}\left\{ \chi_n^2\left(1-\frac{\alpha}{2}\right) \leq \frac{nW_n}{\sigma^2} \leq \chi_n^2\left(\frac{\alpha}{2}\right) \right\} = 1-\alpha.$$

Thus, $\left[\frac{nW_n}{\chi_n^2\left(\frac{\alpha}{2}\right)}, \frac{nW_n}{\chi_n^2\left(1-\frac{\alpha}{2}\right)} \right]$ is a $(1-\alpha)$-confidence interval for σ^2.

6.3. An important result on Gaussian samples. Before going to the next two confidence interval problems, let us try to understand the two examples already covered. In both cases, we came up with a random variable ($\sqrt{n}(\overline{X}_n - \mu)/\sigma$ and W_n/σ^2, respectively) which involved the data and the unknown parameter whose distributions we knew (standard normal and χ_n^2, respectively) and these distributions do not depend on any parameters. This is generally the key step in any confidence interval problem. For the next two problems, we cannot use the same two random variables as above as they depend on the other unknown parameter too (i.e., $\sqrt{n}(\overline{X}_n - \mu)/\sigma$ uses σ which will be unknown and W_n/σ^2 uses μ which will be unknown). Hence, we need a new result that we state without proof.

THEOREM 49. *Let $Z_1, \ldots, Z_n$ be i.i.d. $N(\mu, \sigma^2)$ random variables. Let $\bar{Z}_n$ and s_n^2 be the sample mean and the sample variance, respectively. Then,*

$$\bar{Z}_n \sim N(\mu, \frac{\sigma^2}{n}), \quad \frac{(n-1)s_n^2}{\sigma^2} \sim \chi_{n-1}^2,$$

and the two are independent.

This is not too hard to prove (a muscle-flexing exercise in change of variable formula) but we skip the proof. Note two important features. First, the surprising independence of the sample mean and the sample variance. Second, the sample variance (appropriately scaled)

has χ^2 distribution, just like W_n in the previous example, but the degree of freedom is reduced by 1. Now we use this theorem in computing confidence intervals.

6.4. Estimating σ^2 when μ is unknown. The estimate s_n^2 must be used as W_n depends on μ which is unknown. Theorem thm:indepofsamplemeanandvar tells us that $\frac{(n-1)s_n^2}{\sigma^2} \sim \chi^2_{n-1}$. Hence, by the same logic as before we get

$$\mathbb{P}\left\{\frac{(n-1)s_n^2}{\chi^2_{n-1}\left(\frac{\alpha}{2}\right)} \leq \sigma^2 \leq \frac{(n-1)s_n^2}{\chi^2_{n-1}\left(1-\frac{\alpha}{2}\right)}\right\} = \mathbb{P}\left\{\chi^2_{n-1}\left(1-\frac{\alpha}{2}\right) \leq \frac{(n-1)s_n^2}{\sigma^2} \leq \chi^2_{n-1}\left(\frac{\alpha}{2}\right)\right\}$$
$$= 1-\alpha.$$

Thus, $\left[\frac{(n-1)s_n^2}{\chi^2_{n-1}\left(\frac{\alpha}{2}\right)}, \frac{(n-1)s_n^2}{\chi^2_{n-1}\left(1-\frac{\alpha}{2}\right)}\right]$ is a $(1-\alpha)$-confidence interval for σ^2.

If μ is known, we could use the earlier confidence interval using W_n, or simply ignore the knowledge of μ and use the above confidence interval using s_n^2. What is the difference? The cost of ignoring the knowledge of μ is that the second confidence interval will be typically larger, although for large n the difference is slight. This is because the quantiles $\chi^2_n(\alpha/2)$ increase with n. The quantiles of $\chi^2_n(1-\alpha/2)$ also increase, but to see what happens, here is a table of the reciprocals (which is how the quantiles occur in the confidence intervals) for a few n:

n	1	3	5	7	9
$\frac{1}{\chi^2_n(0.9)}$	0.37	0.16	0.11	0.08	0.07
$\frac{1}{\chi^2_n(0.1)}$	63.33	1.71	0.62	0.35	0.24

On the other hand, if our knowledge of μ was inaccurate, then the first confidence interval is invalid (we have no idea what its level of confidence is!) which is more serious. In realistic situations it is unlikely that we will know one of the parameters but not the other - hence, most often one just uses the confidence interval based on s_n^2.

6.5. Estimating μ when σ^2 is unknown. The earlier confidence interval We look for a confidence interval $[\overline{X}_n - \frac{\sigma}{\sqrt{n}}z_{\alpha/2}, \overline{X}_n + \frac{\sigma}{\sqrt{n}}z_{\alpha/2}]$ cannot be used as we do not know the value of σ.

A natural idea would be to use the estimate $s_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$ in place of σ^2. However, recall that the earlier confidence interval (in particular, the cut-off values $z_{\alpha/2}$ in the CI) was an outcome of the fact that

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \sim N(0,1).$$

Is it true if σ is replaced by s_n? Actually no, but we have a different distribution called *Student's t-distribution*.

EXERCISE 50. Let $Z \sim N(0,1)$ and $S^2 \sim \chi^2_{n-1}$ be independent. Then, the density of $\frac{Z}{S/\sqrt{n-1}}$ is given by

$$\frac{1}{\sqrt{n-1}\text{Beta}(\frac{1}{2}, \frac{n-1}{2})} \frac{1}{\left(1 + \frac{t^2}{n-1}\right)^{\frac{n}{2}}}$$

for all $t \in \mathbb{R}$. This is known as *Student's t-distribution* with $n-1$ degrees of freedom. [*Note:* The reason for using χ^2_{n-1} rather than χ^2_n is that this is how it occurs in most situations, with n being the sample size. We use $(n-1)s_n^2$ which has χ^2_{n-1} distribution for Normal samples. It also reminds us (since the formula does not make sense when $n = 1$) that with 1 data point there is no way to estimate variance.]

The exact density of t-distribution is not important to remember, so the above exercise is optional. The point is that it can be computed from the change of variable formula and that by numerical integration its CDF can be tabulated.

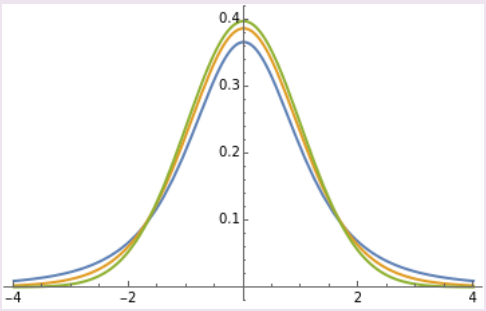

FIGURE 2. Densities of Student's t-distribution with 3 and 9 degrees of freedom (blue and orange, respectively) and the standard normal density function (green). The picture indicates that $t_n(\alpha)$ decreases to z_α as n increases to ∞, for small values of α.

How does this help us? From Theorem 49 we know that $\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \sim N(0,1)$, $\frac{(n-1)s_n^2}{\sigma^2} \sim \chi^2_{n-1}$, and the two are independent. Take these random variables in the above exercise to conclude that $\frac{\sqrt{n}(\overline{X}_n - \mu)}{s_n}$ has t_{n-1} distribution.

The t-distribution is symmetric about zero (the density at t and at $-t$ are the same). Further, as the number of degrees of freedom goes to infinity, the t-density converges to the standard normal density. What we need to know is that there are tables from which we can read off specific quantiles of the distribution. In particular, by $t_n(\alpha)$ we mean the $1-\alpha$ quantile of the t-distribution with n degrees of freedom. Then of course, the α quantile is $-t_n(\alpha)$.

Returning to the problem of the confidence interval, from the fact stated above, we see that (use T_n to indicate a random variable having t-distribution with n degrees of freedom).

$$\begin{aligned}
&\mathbb{P}\left(\overline{X}_n - \frac{s_n}{\sqrt{n}} t_{n-1}\left(\frac{\alpha}{2}\right) \le \mu \le \overline{X}_n + \frac{s_n}{\sqrt{n}} t_{n-1}\left(\frac{\alpha}{2}\right)\right) \\
&= \mathbb{P}\left(-t_{n-1}\left(\frac{\alpha}{2}\right) \le \frac{\sqrt{n}(\overline{X}_n - \mu)}{s_n} \le t_{n-1}\left(\frac{\alpha}{2}\right)\right) \\
&= \mathbb{P}\left(-t_{n-1}\left(\frac{\alpha}{2}\right) \le T_{n-1} \le t_{n-1}\left(\frac{\alpha}{2}\right)\right) \\
&= 1 - \alpha.
\end{aligned}$$

Hence, our $(1-\alpha)$-confidence interval is $\left[\overline{X}_n - \frac{s_n}{\sqrt{n}} t_{n-1}\left(\frac{\alpha}{2}\right), \overline{X}_n + \frac{s_n}{\sqrt{n}} t_{n-1}\left(\frac{\alpha}{2}\right)\right]$.

REMARK 51. We remarked earlier that as $n \to \infty$, the t_{n-1} density approaches the standard normal density. Hence, $t_{n-1}(\alpha)$ approaches z_α for any α (this can be seen by looking at the t-table for large degree of freedom). Therefore, when n is large, we may as well use

$$\left[\overline{X}_n - \frac{s_n}{\sqrt{n}} z_{\alpha/2}, \overline{X}_n + \frac{s_n}{\sqrt{n}} z_{\alpha/2}\right].$$

Strictly speaking the level of confidence is smaller than for the one with $t_{n-1}(\alpha/2)$. However for n large the level of confidence is quite close to $1-\alpha$.

Or to put it in another way, if we know σ^2 is it better to use the CI based on W_n or pretend to not know σ^2 and use the CI based on s_n? In the latter case we use the t-distribution quantiles, and it can be shown that if α is fixed, the sequence $t_n(\alpha)$ decreases to z_α. Therefore the quantile based on s_n^2 tends to be larger.

7. Confidence interval for the mean

Now suppose $X_1, \ldots, X_n$ are i.i.d. random variables from some distribution with mean μ and variance σ^2, both unknown. How can we construct a confidence interval for μ?

In case of normal distribution, recall that the $(1-\alpha)$-CI that we gave was

$$\left[\overline{X}_n - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \overline{X}_n + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right]$$

if σ^2 is known, and

$$\left[\overline{X}_n - \frac{s_n}{\sqrt{n}} t_{n-1}\left(\frac{\alpha}{2}\right), \overline{X}_n + \frac{s_n}{\sqrt{n}} t_{n-1}\left(\frac{\alpha}{2}\right)\right] \text{ or } \left[\overline{X}_n - \frac{s_n}{\sqrt{n}} z_{\alpha/2}, \overline{X}_n + \frac{s_n}{\sqrt{n}} z_{\alpha/2}\right]$$

if σ^2 is unknown.

Are this a valid confidence interval if X_i are i.i.d from a different distribution? The answer is "No" in both cases. If X_i are from some general distribution then the distributions of $\sqrt{n}(\overline{X}_n - \mu)/s_n$ and $\sqrt{n}(\overline{X}_n - \mu)/\sigma$ are very complicated to find. Even if X_i have Binomial or Exponential distributions, the distribution of $\sqrt{n}(\overline{X}_n - \mu)/s_n$ and $\sqrt{n}(\overline{X}_n - \mu)/\sigma$ depend on the parameters in a complex way (in particular, the distributions are not free from the parameters, which is important in constructing confidence intervals).

But suppose n is large. Then the sample variance is close to population variance and hence $s_n \approx \sigma$. Further, by CLT, we know that $\sqrt{n}(\overline{X}_n - \mu)/\sigma$ has approximately $N(0,1)$ distribution. Hence, we see that

$$\mathbb{P}\left\{-z_{\alpha/2} \le \frac{\sqrt{n}(\overline{X}_n - \mu)}{s_n} \le z_{\alpha/2}\right\} \approx \Phi(z_{\alpha/2}) - \Phi(-z_{\alpha/2}) = 1 - \alpha.$$

Consequently, we may say that

$$\mathbb{P}\left\{\overline{X}_n - \frac{s_n}{\sqrt{n}} z_{\alpha/2} \le \mu \le \overline{X}_n + \frac{s_n}{\sqrt{n}} z_{\alpha/2}\right\} \approx 1 - \alpha.$$

Thus, $\left[\overline{X}_n - \frac{s_n}{\sqrt{n}} z_{\alpha/2}, \overline{X}_n + \frac{s_n}{\sqrt{n}} z_{\alpha/2}\right]$ is an approximate $(1-\alpha)$-confidence interval. Further, when n is large, $s_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$ and $V_n := \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2$ are almost the same (indeed, $s_n^2 = \frac{n}{n-1} V_n$). Hence it is also okay to use $\left[\overline{X}_n - \frac{\sqrt{V_n}}{\sqrt{n}} z_{\alpha/2}, \overline{X}_n + \frac{\sqrt{V_n}}{\sqrt{n}} z_{\alpha/2}\right]$ as an approximate $(1-\alpha)$-confidence interval.

EXAMPLE 52. Let $X_1, \ldots, X_n$ be i.i.d. Ber(p). Consider the problem of finding a confidence interval for p. Since each X_i is 0 or 1, observe that

$$\hat{s}_n^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X}_n^2 = \overline{X}_n - (\overline{X}_n)^2 = \overline{X}_n(1 - \overline{X}_n).$$

Hence, an approximate $(1-\alpha)$-CI for p is given by

$$\left[\overline{X}_n - z_{\alpha/2}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}, \overline{X}_n + z_{\alpha/2}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}\right].$$

8. Actual confidence by simulation

Suppose we have a candidate confidence interval whose confidence we do not know. For example, let us take the confidence interval

$$\left[\overline{X}_n - z_{\alpha/2}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}, \overline{X}_n + z_{\alpha/2}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}\right].$$

for the parameter p of i.i.d. Ber(p) samples. We saw that for large n this has approximately $(1-\alpha)$ confidence. But how large is large? Or alternately, if n is not large, then what is the actual confidence level of this interval? One way to check this is by simulation. We explain how.

Take $p = 0.3$ and $n = 10$. Simulate $n = 10$ independent Ber(p) random variables and compute the confidence interval given above. Check whether it contains the true value of p (i.e., 0.3) or not. Repeat this exercise 10000 times and see what proportion of times it contains 0.3. That proportion is the true confidence, as opposed to $1-\alpha$ (which is valid only for large n). Repeat this experiment with $n = 20$, $n = 30$ etc. See how close the actual confidence is to

$1-\alpha$. Repeat this experiment with different value of p. The n you need to get close to $1-\alpha$ will depend on p (in particular, on how close p is to 1/2).

This was about checking the validity of a confidence interval that was specified. In a real situation, it may be that we can only get $n = 20$ samples. Then what can we do? If we have an idea of the approximate value of p, we can first simulate Ber(p) random numbers on a computer. We compute the sample mean each time, and repeat 10000 times to get so many values of the sample mean. Note that the histogram of these 10000 values tells us (approximately) the actual distribution of $\overline{X}_n$. Then we can find t (numerically) such that $[\overline{X}_n - t, \overline{X}_n + t]$ contains the true value of p in $(1-\alpha)$-proportion of the 10000 trials. Then, $[\overline{X}_n - t, \overline{X}_n + t]$ is a $(1-\alpha)$-CI for p. Alternately, we may try a CI of the form

$$\left[\overline{X}_n - t\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}, \overline{X}_n + t\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}\right].$$

where we choose t numerically to get $(1-\alpha)$ confidence.

Summary: The gist of this discussion is this. In the neatly worked out examples of the previous sections, we got explicit confidence intervals. But we assumed that we knew the data came from $N(\mu, \sigma^2)$ distribution. What if that is not quite right? What if it is not any of the nicely studied distributions? The results also become invalid in such cases. For large n, using law of large numbers and CLT we could overcome this issue. But for small n? The point is that using simulations we can calculate probabilities, distributions, etc, numerically and approximately. That is often better, since it is more robust to assumptions.

CHAPTER 12

Hypothesis testing

1. Hypothesis testing - first examples

Earlier in the course we discussed the problem of how to test whether a "psychic" can make predictions better than a random guesser. This is a prototype of what are called *testing problems*. We start with this simple example and introduce various general terms and notions in the context of this problem.

QUESTION 53. A "psychic" claims to guess the order of cards in a deck. We shuffle a deck of cards, ask her to guess and count the number of correct guesses, say X.

One hypotheses (we call it the *null hypothesis* and denote it by H_0) is that the psychic is guessing randomly. The *alternate hypothesis* (denoted H_1) is that his/her guesses are better than random guessing (in itself this does not imply existence of psychic powers. It could be that he/she has managed to see some of the cards etc.). Can we decide between the two hypotheses based on X?

What we need is a rule for deciding which hypothesis is true. A rule for deciding between the hypotheses is called a *test*. For example, the following are examples of rules (the only condition is that the rule must depend only on the data at hand).

EXAMPLE 54. We present three possible rules.

(1) If X is an even number declare that H_1 is true. Else declare that H_1 is false.

(2) If $X \geq 5$, then accept H_1, else reject H_1.

(3) If $X \geq 8$, then accept H_1, else reject H_1.

The first rule does not make much sense as the parity (evenness or oddness) has little to do with either hypothesis. On the other hand, the other two rules make some sense. They rely on the fact that if H_1 is true then we expect X to be larger than if H_0 is true. But the question still remains, should we draw the line at 5 or at 8 or somewhere else?

In testing problems there is only one objective, to avoid the following two possible types of mistakes.

Type-I error: H_0 is true but our rule concludes H_1.

Type-II error: H_1 is true but our rule concludes H_0.

The probability of type-I error is called the *significance level* of the test and usually denote by α. That is, $\alpha = \mathbb{P}_{H_0}\{\text{the test accepts } H_1\}$ where we write $\mathbb{P}_{H_0}$ to mean that the probability is calculated under the assumption that H_0 is true. Similarly one define the *power* of the test as

$\beta = \mathbb{P}_{H_1}\{\text{the test accepts } H_1\}$. Note that β is the probability of not making type-II error, and hence we would like it to be close to 1. Given two tests with the same level of significance, the one with higher power is better. Ideally we would like α to be close to 0 and β to be close to 1 simultaneously, but that is not always achievable. More precisely, it is achievable if we could increase sample size enormously, but we may not have the option).

We fix the desired level of significance, usually $\alpha = 0.05$ or 0.1 and only consider tests whose probability of type-I error is at most α. It may seem surprising that we take α to be so small. Indeed the two hypotheses are not treated equally. Usually H_0 is the default option, representing traditional belief and H_1 is a claim that must prove itself. As such, the burden of proof is on H_1.

To use analogy with law, when a person is convicted, there are two hypotheses, one that he is guilty and the other that he is not guilty. According to the maxim "innocent till proved guilty", one is not required to prove his/her innocence. On the other hand guilt must be proved. Thus the null hypothesis is "not guilty" and the alternative hypothesis is "guilty".

In our example of card-guessing, assuming random guessing, we have calculated the distribution of X long ago. Let $p_k = \mathbb{P}\{X = k\}$ for $k = 0, 1, \ldots, 52$. Now consider a test of the form "Accept H_1 if $X \geq k_0$ and reject otherwise". Its level of significance is

$$\mathbb{P}_{H_0}\{\text{accept } H_1\} = \mathbb{P}_{H_0}\{X \geq k_0\} = \sum_{i=k_0}^{52} p_i.$$

For $k_0 = 0$, the right side is 1 while for $k_0 = 52$ it is $1/52!$ which is tiny. As we increase k_0 there is a first time where it becomes less than or equal to α. We take that k_0 to be the threshold for cut-off.

In the same example of card-guessing, let $\alpha = 0.01$. Let us also assume that Poisson approximation holds. This means that $p_j \approx e^{-1}/j!$ for each j. Then, we are looking for the smallest k_0 such that $\sum_{j=k_0}^{\infty} e^{-1}/j! \leq 0.01$. For $k_0 = 4$, this sum is about 0.019 while for $k_0 = 5$ this sum is 0.004. Hence, we take $k_0 = 5$. In other words, accept H_1 if $X \geq 5$ and reject if $X < 5$. If we took $\alpha = 0.0001$ we would get $k_0 = 7$ and so on.

Strength of evidence: Rather than merely say that we accepted H_1 or rejected it would be better to say how strong the evidence is in favour of the alternative hypothesis. This is captured by the *p-value*, a central concept of decision making. It is defined as *the probability that data drawn from the null hypothesis would show closer agreement with the alternative hypothesis than the data we have at hand* (read it five times!).

Before we compute it in our example, let us return to the analogy with law. Suppose a man is convicted for murder. Recall that H_0 is that he is not guilty and H_1 is that he is guilty. Suppose his fingerprints were found in the house of the murdered person. Does it prove his guilt? It is some evidence in favour of it, but not necessarily strong. For example, if the convict was a friend of the murdered person, then he might be innocent but have left his fingerprints on his visits to his friend. However if the convict is a total stranger, then one wonders why, if he was innocent, his finger prints were found there. The evidence is

stronger for guilt. If bloodstains are found on his shirt, the evidence would be even stronger! In saying this, we are asking ourselves questions like "if he was innocent, how likely is it that his shirt is blood-stained?". That is p-value. Smaller the p-value, stronger the evidence for the alternate hypothesis.

Now we return to our example. Suppose the observed value is $X_{\text{obs}} = 4$. Then the p-value is $\mathbb{P}\{X \geq 4\} = p_4 + \ldots + p_{52} \approx 0.019$. If the observed value was $X_{\text{obs}} = 6$, then the p-value would be $p_6 + \ldots + p_{52} \approx 0.00059$. Note that the computation of p-value does not depend on the level of significance. It just depends on the given hypotheses and the chosen test.

2. The Likelihood ratio test

We consider the simplest hypothesis testing problem now. It is not in itself something that occurs in practical situations. But it is mathematically interesting and the solution is of relevance in more general problems.

Let f_0 and f_1 be two probability densities or probability mass functions. Our data consists of a single sample X, drawn from f_0 or from f_1. The simplest testing problem is one where we need to decide between

$$H_0 : X \sim f_0 \qquad \text{versus} \qquad H_1 : X \sim f_1.$$

If we fix a level of significance α, then there are many tests to choose from. All of them are as follows: Let $A \subseteq \mathbb{R}$ be a subset such that $\mathbb{P}_{H_0}(X \in A) \leq \alpha$, i.e., $\int_A f_0(x)dx \leq \alpha$ (if f_0 is a density, replace by sum if it is a mass function). Then consider the test T_A (it depends on the choice of A) which rejects the null and accepts the alternative if $X \in A$ and refuses to accept the alternative if $X \notin A$. Then the probability of type-I error is at most α, since we chose A to have probability less than or equal to α under f_0.

Example 55. Let f_0 be the $N(0,1)$ density and let f_1 be the $N(0.6,1)$ density. Let $\alpha = 0.10$. Here are a few sets that have probability (approximately) equal to 0.10 under f_0: (A) $A = (-\infty, -1.28]$, (B) $A = [1.28, \infty)$, (C) $A = [0, 0.26]$, (D) $A = [0.53, 0.85]$. Which of these should we pick as our *rejection region* (i.e., the test that rejects null when $X \in A$)? In this example, it seems most natural to choose the second one, as f_1 favours larger values of X and f_0 favours smaller values of X (relative to each other). But what is the basis for this choice?

If we have two unrelated densities, say $N(0,1)$ and standard Cauchy or $N(1,1)$ and Exp(1), then what would you do? Even if they have different means, it need not be true that the second kind of interval is best. Among all level-α tests, it makes sense to pick the one with the highest power (i.e., the smallest type-II error probability). This means, that we should *maximize*

$$\int_A f_1(x)dx \text{ over all } A \subseteq \mathbb{R} \text{ subject to } \int_A f_0(x)dx \leq \alpha.$$

The answer turns out to be clean!

LEMMA 56 (Neyman-Pearson). *Let f_0 and f_1 be densities or mass functions. Among all the α-level tests, the one with the highest power is the one whose rejection region is of the form $A_* = \{x : \frac{f_1(x)}{f_0(x)} \geq b_\alpha\}$, where b_α is chosen so that the type-I error of T_{A_*} is equal to α.*

REMARK 57. In the discrete setting, it may not be possible to find a b so that the level is equal to α. In this case, we may have to resort to *randomization*. We do not describe it in general now, but the Poisson example below explains it in that context and it should make it clear how the general situation goes.

The optimal test described in this lemma is called the *likelihood ratio test*.

EXAMPLE 58. Returning to the earlier example,

$$\frac{f_1(x)}{f_0(x)} = e^{\frac{1}{2}x^2 - \frac{1}{2}(x-0.6)^2} = e^{0.6x-0.18}.$$

Hence the rejection region is of the form $A_* = \{x : e^{0.6x-0.18} > b\}$ for some b, which is the same as $A_* = [c, \infty)$ for some other c. To achieve level α we take $c = z_\alpha$, hence the best test is the one that rejects the null hypothesis if and only if $X > z_\alpha$.

Let us do a discrete example.

EXAMPLE 59. Let $H_0 : X \sim \text{Pois}(3)$ and $H_1 : X \sim \text{Pois}(2)$. In this case

$$\frac{f_1(k)}{f_0(k)} = \frac{e^{-2}2^k/k!}{e^{-3}3^k/k!} = e\left(\frac{2}{3}\right)^k.$$

As this is a decreasing function of k, the rejection region A_* of the likelihood ratio test is of the form $\{0, 1, \ldots, \ell\}$ for some ℓ. The right choice of ℓ is such that the level is short of α, but if we include $\ell + 1$ the level would jump over α. In other words, we choose ℓ so that

$$\sum_{j=0}^{\ell} e^{-3}\frac{3^j}{j!} \leq \alpha < \sum_{j=0}^{\ell+1} e^{-3}\frac{3^j}{j!}.$$

Finally the test rejects the null if and only if $X \leq \ell$. In this case, the level of the test may not be exactly equal to α.

We can achieve that, provided we allow *randomization* as follows. Write $\alpha_1 = \sum_{j=0}^{\ell} e^{-3}\frac{3^j}{j!}$, so what is left-over is $\alpha - \alpha_1$. By our choice of ℓ, it is clear that $\alpha - \alpha_1 < f_1(\ell + 1)$. Now set $r = \frac{\alpha-\alpha_1}{f_1(\ell+1)}$ so that $r \in [0, 1)$. Now the test is as follows: If $X \leq \ell$, reject the null hypothesis. If $X = \ell + 1$, then draw a uniform random number $U \sim \text{Unif}[0, 1]$, and if $U \leq r$, then reject the null. In all other cases (either $X > \ell$ or $X = \ell + 1$ and $U > r$) we reject the alternative. It is easy to see that this test achieves level α.

Let us give an example to show that the region need not be a one sided interval.

EXAMPLE 60. Let $H_0 : X \sim N(0, 1)$ and $H_1 : X \sim N(0, 2)$. In this case

$$\frac{f_1(x)}{f_0(x)} = \frac{1}{\sqrt{2}}e^{\frac{x^2}{4}}$$

hence the rejection region of the likelihood ratio test is of the form $\{x : x^2 > c\}$ for some $c > 0$ or equivalently $\{x : |x| \geq \sqrt{c}\}$. To achieve level α we must take $\sqrt{c} = z_{\alpha/2}$. That is, reject the null hypothesis if and only if $|X| \geq z_{\alpha/2}$.

Proof of the Neyman-Pearson lemma. We write the proof when f_0 and f_1 are probability density functions. Let $A_* = \{x : f_1(x) \geq b_\alpha f_0(x)\}$ and let $A \subseteq \mathbb{R}$ be any set such that $\mathbb{P}_{H_0}(X \in A) = \alpha$ (recall that b_α is so chosen that $\mathbb{P}_{H_0}(X \in A_*) = \alpha$). Then,

$$\begin{aligned}
\mathbb{P}_{H_1}(X \in A) &= \int_{A \cap A_*} f_1(x)dx + \int_{A \setminus A_*} f_1(x)dx, \\
&\leq \int_{A \cap A_*} f_1(x)dx + \int_{A \setminus A_*} b_\alpha f_0(x)dx \\
&= \int_{A \cap A_*} f_1(x)dx + b_\alpha \int_{A \setminus A_*} f_0(x)dx. \qquad (21)
\end{aligned}$$

The inequality in the second line comes from the fact that $f_1(x) < b_\alpha f_0(x)$ for $x \in A \setminus A_*$. If we interchange the roles of A and A_*, we use that $f_1(x) \geq b_\alpha f_0(x)$ for $x \in A_* \setminus A$ to get

$$\begin{aligned}
\mathbb{P}_{H_1}(X \in A_*) &= \int_{A \cap A_*} f_1(x)dx + \int_{A_* \setminus A} f_1(x)dx, \\
&\geq \int_{A \cap A_*} f_1(x)dx + \int_{A_* \setminus A} b_\alpha f_0(x)dx \\
&= \int_{A \cap A_*} f_1(x)dx + b_\alpha \int_{A_* \setminus A} f_0(x)dx. \qquad (22)
\end{aligned}$$

Lastly, we claim that

$$\int_{A \setminus A_*} f_0(x)dx = \int_{A_* \setminus A} f_0(x)dx. \qquad (23)$$

To see this, we write the significance levels of the two tests as

$$\begin{aligned}
\alpha = \mathbb{P}_{H_0}(X \in A) &= \int_{A \cap A_*} f_0(x)dx + \int_{A \setminus A_*} f_0(x)dx, \\
\alpha = \mathbb{P}_{H_0}(X \in A_*) &= \int_{A \cap A_*} f_0(x)dx + \int_{A_* \setminus A} f_0(x)dx.
\end{aligned}$$

Subtracting the second from the first gives (23).

Combining (21), (22) and (23) shows that $\mathbb{P}_{H_1}(X \in A_*) \geq \mathbb{P}_{H_1}(X \in A)$. Thus, the test with rejection region A_* has the highest power among all level-α tests. ■

Summary: In this section we saw the simplest hypothesis testing problem. To be clear the simplicity does not refer to the fact that we have only one sample X, that can be generalized to multiple samples very easily. What we allude to is that both the null and alternate hypotheses specify one single distribution each. In most realistic examples of practical importance, at least the alternate hypothesis will consist not of one distribution but many. For example, we could have $H_0 : X \sim N(0, 1)$ versus $H_1 : X \sim N(\mu, 1)$ for some $\mu > 0$.

In such cases, the power (or the probability of type-II error) is not a number, but a function of μ, $\mu > 0$. So it does not make sense to ask for maximizing the power. In any case, we are not interested in optimality at this stage. We just want to consider a few interesting classes of testing problems, and come up with natural tests for them.

3. Testing for the mean of a normal population

Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$. We shall consider the following hypothesis testing problems.

(1) One sided test for the mean. $H_0 : \mu = \mu_0$ versus $H_1 : \mu > \mu_0$.

(2) Two sided test for the mean. $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$.

This kind of problem arises in many situations in comparing the effect of a treatment as follows.

Example 61. Consider a drug claimed to reduce blood pressure. How do we check if it actually does? We take a random sample of n patients, measure their blood pressures $Y_1, \ldots, Y_n$. We administer the drug to each of them and again measure the blood pressures $Y_1', \ldots, Y_n'$, respectively. Then, the question is whether the mean blood pressure decreases upon giving the treatment. To this effect, we define $X_i = Y_i - Y_i'$ and wish to test the hypothesis that the mean of X_is is strictly positive. If X_i are indeed normally distributed, this is exactly the one-sided test above.

Because of the kind of situation in the above example, this testing problem is also called *paired sample test*. We have pairs of samples (Y_i, Y_i'), where both are measurements on the same entity (before treatment and after treatment, for example). The same applies to test the efficacy of a fertilizer to increase yield, a proposed drug to decrease weight, a particular educational method to improve a skill, or a particular course such as the current *probability and statistics course* in increasing subject knowledge. To make a policy decision on such matters, we can conduct an experiment as in the above example.

For example, a bunch of students are tested on probability and statistics and their scores are noted. Then they are subjected to the course for a semester. They are tested again after the course (for the same marks, and at the same level of difficulty) and the scores are again noted. Take differences of the scores before and after, and test whether the mean of these differences is positive (or negative, depending on how you take the difference). This is a one-sided tests for the mean. Note that in these examples, we are taking the null hypothesis to be that there is no effect. In other words, the burden of proof is on the new drug or fertilizer or the instructor of the course.

The test: Now we present the test. We separate two cases, depending on whether we know σ^2.

(1) If we know σ^2 (then denote it σ_0^2), then the test statistic $\mathcal{Z} = \frac{\sqrt{n}(\overline{X}_n - \mu_0)}{\sigma_0}$ has $N(0,1)$ distribution, *under the null hypothesis*.

(2) If we don't know σ^2, then the statistic $\mathcal{T} := \frac{\sqrt{n}(\overline{X}_n - \mu_0)}{s_n}$ has Student's t-distribution with $n-1$ degrees of freedom, *under the null hypothesis*.

The tests based on these are as follows.

(1) One-sided test: If σ^2 is known, then accept the alternative if $\mathcal{Z} > z_\alpha$. If σ^2 is unknown, then accept the alternative hypothesis if $\mathcal{T} > t_{n-1}(\alpha)$.

(2) Two sided-test: If σ^2 is known, then accept the alternative hypothesis if $|\mathcal{Z}| > z_{\alpha/2}$. If σ^2 is unknown, then accept the alternative hypothesis if $|\mathcal{T}| > t_{n-1}(\alpha/2)|$.

It should be clear that these are level-α tests. But these are not the only ones. Even if we decide to base our test on $\mathcal{Z}$ or on $\mathcal{T}$, there are many choices of tests. For example, if σ^2 is not known, then we can choose any subset $A \subseteq \mathbb{R}$ that has probability α under the t-distribution with $n-1$ degrees of freedom, and reject the null when $\mathcal{T}$ falls inside A. We can provide two reasons for choosing the above test.

(1) If $\bar{X}$ is much larger than μ_0 then the greater is the evidence that the true mean μ is greater than μ_0. Hence the rejection region should be of the form $\mathcal{T} > b$ for some b (for the one-sided test when σ^2 is unknown).

(2) Another rationale comes from our analysis in the earlier section. Assume that σ^2 is known, and consider the one-sided test. If we replace the alternate hypothesis by the simpler hypothesis $H_1 : \mu = \mu_1$ for some fixed $\mu_1 > \mu_0$, then we know that the likelihood ratio test is the best, and that the LRT takes the form $\mathcal{Z} > z_\alpha$. Since this is the best test for any fixed μ_1, shouldn't it be the reasonable choice for the one-sided test too?

These give a rationale for the rejection region, once we accept that we want to base our test on $\mathcal{Z}$ or $\mathcal{T}$. But what is the reason for basing the test on these statistics?

(1) We want our test to be immune to change of units. For example, our conclusion regarding whether a fertilizer increases the growth of wheat should remain the same whether the height of the wheat grass is measured in centimeters or inches. This means that the test should be based on dimension-free quantities $(X_i - \mu_0)/\sigma_0$ or $(X_i - \mu_0)/s_n$.

(2) Why should the particular combinations ($\mathcal{Z}$ and $\mathcal{T}$) of these standardized observations be used? Why not another function? In a more advanced statistics class, one makes a point that what can be done using the whole data can also be done using $\mathcal{Z}$ (in case σ^2 is known).

REMARK 62. Earlier we considered the problem of constructing a $(1-\alpha)$-CI for μ when σ^2 is unknown. The two sided test abovecan be simply stated as follows: Accept the alternative at level α if the corresponding $(1-\alpha)$-CI does not contain μ_0. Conversely, if we had dealt with testing problems first, we could define a confidence interval as the set of all those μ_0 for which the corresponding test rejects the alternative.

Thus, confidence intervals and testing are closely related. This is true in some greater generality. For example, one-sided confidence interval for μ are closely related to the one-sided tests above.

4. Testing for the difference between means of two normal populations

Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu_1, \sigma_1^2)$ and let $Y_1, \ldots, Y_m$ be i.i.d. $N(\mu_2, \sigma_2^2)$. We shall consider the following hypothesis testing problems.

(1) One sided test for the difference in means. $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 > \mu_2$.

(2) Two sided test for the mean. $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$.

This kind of problem arises in many situations in comparing two different populations or the effect of two different treatments etc. Actual data sets of such questions can be found in the homework.

EXAMPLE 63. Suppose a new drug to reduce blood pressure is introduced by a pharmaceutical company. There is already an existing drug in the market which is working reasonably alright. But it is claimed by the company that the new drug is better. How to test this claim?

We take a random sample of $n + m$ patients and break them into two groups of n and of m patients. The first group is administered the new drug while the second group is administered the old drug. Let $X_1, \ldots, X_n$ be the *decrease in blood pressures* in the first group. Let $Y_1, \ldots, Y_m$ be the *decrease* in blood pressures in the second group. The claim is that one average X_is are larger than Y_is.

Note that it does not make sense to subtract $X_i - Y_i$ and reduce to a one sample test as in the previous section (here X_i is a measurement on one person and Y_i is a measurement on a completely different person! Even the number of persons in the two groups may differ). This is an example of a two-sample test as formulated above.

EXAMPLE 64. The same applies to many studies of comparision. If someone claims that Americans are taller than Indians on average, or if it is claimed that cycling a lot leads to increase in height, or if it is claimed that Chinese have higher IQ than Europeans, or if it is claimed that *Honda Activa* gives better mileage than *Suzuki Access*, etc., etc., the claims can be reduced to the two-sample testing problem as introduced above.

BIG ASSUMPTION: We shall assume that $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (yet unknown). This assumption is not made because it is natural or because it is often observed, but because it leads to mathematical simplification. Without this assumption, no exact level-α test has been found!

The test: Let $\bar{X}, \bar{Y}$ denote the sample means of X and Y and let s_X, s_Y denote the corresponding sample standard deviations. Since σ^2 is the assumed to be the same for both populations, s_X^2 and s_Y^2 can be combined to define

$$S^2 := \frac{(n-1)s_X^2 + (m-1)s_Y^2}{m+n-2}$$

which is a better estimate for σ^2 than just s_X^2 or s_Y^2 (this S^2 is better than simply taking $(s_X^2 + s_Y^2)/2$ because it gives greater weight to the larger sample).

Now define $\mathcal{T} = \frac{\bar{X}-\bar{Y}}{S\sqrt{\frac{1}{n}+\frac{1}{m}}}$. The following tests hav significance level α.

(1) For the one-sided test, accept the alternative if $\mathcal{T} > t_{n+m-2}(\alpha)$.

(2) For the one-sided test, accept the alternative if $\mathcal{T} > t_{n+m-2}(\alpha/2)$ or $\mathcal{T} < -t_{n+m-2}(\alpha/2)$.

The rationale behind the tests: If $\bar{X}$ is much larger than $\bar{Y}$ then the greater is the evidence that the true mean μ_1 is greater than μ_2. But again we need to standardize by dividing this by an estimate of σ, namely S. The resulting statistic $\mathcal{T}$ has a t_{m+n-2} distribution as explained below.

The significance level is α: The question is where to draw the threshold. From the facts we know,

$$\bar{X} \sim N(\mu_1, \sigma^2/n),$$

$$\bar{Y} \sim N(\mu_2, \sigma^2/m),$$

$$\frac{(n-1)}{\sigma^2} s_X^2 \sim \chi^2_{n-1},$$

$$\frac{(m-1)}{\sigma^2} s_Y^2 \sim \chi^2_{m-1}$$

and the four random variables are independent. From this, it follows that $\frac{(m+n-2)}{\sigma^2} S^2$ has χ^2_{n+m-2} distribution. Further, *Under the null hypothesis*

$$\frac{\bar{X}-\bar{Y}}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}} \sim N(0,1).$$

It is also clearly independent of S. Taking ratios, we see that (observe that σ cancels)

$$\mathcal{T} = \frac{\frac{\bar{X}-\bar{Y}}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}}}{\sqrt{\frac{S^2}{\sigma^2(m+n-2)}}}$$

has Student's t-distribution with $m + n - 2$ degrees of freedom (under the null hypothesis).

5. Testing for the mean in absence of normality

Suppose $X_1, \ldots, X_n$ are i.i.d. Ber(p). Consider the test

$$H_0 : p = p_0 \quad \text{versus} \quad H_1 : p \neq p_0.$$

One can also consider the one-sided test. Just as in the confidence interval problem, we can give a solution when n is large, using the approximation provided by the central limit

theorem. Recall that an approximate $(1-\alpha)$-CI is

$$\left[\overline{X}_n - z_{\alpha/2}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}, \overline{X}_n + z_{\alpha/2}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}\right].$$

Inverting this confidence interval, we see that a reasonable test is:

Reject the alternative if p_0 belongs to the above CI. That is, accept the alternative if

$$\overline{X}_n - z_{\alpha/2}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}} \leq p_0 \leq \overline{X}_n + z_{\alpha/2}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}$$

This test has (approximately) significance level α.

More generally, if we have data $X_1, \ldots, X_n$ from a population with mean μ and variance σ^2, then consider the test

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu \neq \mu_0.$$

A test with approximate significance level α is given by: Reject the alternative if

$$\overline{X}_n - z_{\alpha/2}\frac{s_n}{\sqrt{n}} \leq \mu_0 \leq \overline{X}_n + z_{\alpha/2}\frac{s_n}{\sqrt{n}}.$$

Just as with confidence intervals, we can find the actual level of significance (if n is not large enough) by simulating data on a computer.

5.1. An impossible testing problem. Suppose $X_1, \ldots, X_n$ are i.i.d. Ber(p) and consider the hypotheses

$$H_0 : p \neq \frac{1}{2} \qquad H_1 : p = \frac{1}{2}.$$

There are no tests at all for this problem. The null hypothesis is so broad that anything that can be explained by the alternative can also be explained by the null, hence there is never a reason to reject it. For example, suppose we toss the coin 100 times and get exactly 50 heads. That may indicate that the alternative is correct, but the null value of $p = 0.50001$ is fully consistent with the data. Since the null is our default belief, and we only reject it if there is strong evidence against it, we see that in this problem we never reject the null.

Another impossible to test problems are $H_0 : p \in \mathbb{Q}$ versus $H_1 : p \notin \mathbb{Q}$ (it is just as bad if we interchange the two hypotheses).

Remark 65. It is not just that the null hypothesis is a compound hypothesis (i.e., it has multiple values of the parameters). For example, the test $H_0 : p \leq \frac{1}{2}$ versus $H_1 : p > \frac{1}{2}$, is perfectly fine. In fact, the test for this is the same as if the null hypothesis simply said $H_0 : p = \frac{1}{2}$. In the impossible problems above, for every value of θ in the alternative hypothesis, there are values in the null hypothesis that are arbitrarily close to θ, hence the null is impossible to rule out.

Another remark is that even the impossible problems above become possible, if only we had *infinite data*! For example, if $X_1, X_2, \ldots$ are i.i.d. Ber(p), then the strong law of large

numbers (we only talked about the weak law of large numbers, this is slightly different) asserts that the sequence $\overline{X}_n$ converges to p with probability 1. Thus, by computing the limit, we can say for sure whether or not $p = \frac{1}{2}$. Then, not only can we decide between $H_0 : \ p \neq \frac{1}{2}$ versus $H_1 : \ p = \frac{1}{2}$, we can even decide with zero probabilities of type-1 and type-2 errors! But we never have infinite data in statistics, so this discussion is purely academic.

6. A few non-parametric testing problems

Although in the beginning we said that we stay with parametric problems in statistics, there are some pro problems of such importance that we wish to talk about them. Here we briefly describe three of them, and in successive sections we discuss them in some details.

► **Goodness of fit test**: We have said on various occasions that heights and weights follow Normal distribution, that lifetimes of certain electronic components follow Exponential distribution, that the number of traffic accidents in a junction follows Poisson distribution, etc. While there may be theoretical justifications for some of these, ultimately they have to be checked against data. How can we test that heights follow Normal distribution? We can formulate this as a hypothesis testing problem.

► **Test of independence**: Are musical and mathematical abilities independent of each other? Is academic performance independent of performance in sports? Is incidence of lung cancer independent of smoking? Is brain size independent of IQ score? Is the height of a pea plant independent of its seed-type (smooth/wrinkled)? There are many questions in the same vein that one can ask. In all these cases, we have two features X and Y of an individual entity (could be a person or a plant or something else), and we want to check if the features are independent of each other.

The natural idea is to take a sample of individuals and take the measurements $(X_1, Y_1), \ldots, (X_n, Y_n)$. For example X_k could be the score in a mathematics exam and Y_k could be the score of the same person in a music exam. A special case is when (X, Y) has jointly Normal distribution. In that case, independence is the same as the correlation $\rho = \mathrm{Cov}(X, Y)$ being equal to 0. Then the problem is clear, $H_0 : \ \rho = 0$ versus $H_1 : \ \rho \neq 0$. This is a parametric testing problem, and can be solved.

In the absence of Normality, independence is not the same as uncorrelatedness and we must take a different approach. We formulate the general problem of checking independence as a hypothesis testing problem.

► **Comparison of distributions**: Earlier we saw the two-sample test for comparison of means in a Normal population. The problem of comparison is a natural one. We can test the efficacy of two fertilizers by using them on two different mango orchards and measure the number of mangoes in each theights of plants in both fields. What if the heights are not normally distributed? We may be interested in comparing the IQ scores between two groups of people (the groups could be populations of two countries or students of two different schools. Again, what to do if we cannot assume normality of distributions?

We formulate and solve this as a very general non-parametric testing problem.

REMARK 66. Which is better, parametric or non-parametric? Of course this only matters in cases where there is a choice. For example, we can test for independence between X and Y assuming normality or without that assumption. Which is better? If the true distribution is Normal, then the parametric test will do much better (for the same level of significance, it wail have higher power). On the other hand, if our assumptions are on shaky foundation, it is better to do the non-parametric test. It is more conservative, but more reliable on the whole.

7. Chi-squared test for goodness of fit

At various times we have made statements such as "heights follow normal distribution", "lifetimes of bulbs follow exponential distribution" etc. Where do such claims come from? Over years of analysing data, of course. This leads to an interesting question. Can we test whether lifetimes of bulbs do follow exponential distribution?

We start with a simple example of testing whether a die is fair. The hypotheses are[1]

$$H_0: \text{ the die is fair} \qquad \text{versus} \qquad H_1: \text{ the die is unfair.}$$

We throw the die n times and record the observations $X_1, \ldots, X_n$. For $j \leq 6$, let O_j be the number of times we observe the face j turn up. In symbols $O_j = \sum_{i=1}^{n} \mathbf{1}_{X_i=j}$. Let $E_j = \mathbb{E}[O_j] = \frac{n}{6}$ be the expected number of times we see the face j (under the null hypothesis). Common sense says that if H_0 is true then O_j and E_j must be rather close for each j. How to measure the closeness? Karl Pearson introduced the test statistic

$$T := \sum_{j=1}^{6} \frac{(O_j - E_j)^2}{E_j}.$$

If the desired level of significance is α, then the Pearson χ^2-test says "Reject H_0 if $T \geq \chi^2_5(\alpha)$". The number of degrees of freedom is 5 here. In general, it is one less than the number of bins (i.e., how many terms you are summing to get T).

Some practical points: The χ^2 test is really an asymptotic statement. For large n, the level of significance is approximately $1-\alpha$. There is no assurance for small n. Further, in performing the test, it is recommended that each bin must have at least 5 observations (i.e., $O_j \geq 5$). Otherwise we club together bins with fewer entries. The number 5 is a rule of thumb, the more the better.

[1] You may feel that the null and alternative hypotheses are reversed. Is not independence a special property that should prove itself. Yes and no. Here we are imagining a situation where we have some reason to think that the die is fair. For example perhaps the die looks symmetric. In any case, if we reverse the hypotheses, then like in the examples in Section 5.1, it will become an impossible problem. The null will never be rejected (because there are unfair dice with probabilities arbitrarily close to that of the fair die).

Fitting the Poisson distribution: We consider the famous data collected by Rutherford, Chadwick and Ellis on the number of radioactive disintegrations. For details see the book of Feller's book (section VI.7) or this website.

The data consists of $X_1, \ldots, X_{2608}$ (where X_k is the number of particles detected by the counter in the k^{th} time interval. The hypotheses are

$$H_0: F \text{ is a Poisson distribution.} \qquad H_1: F \text{ is not Poisson.}$$

The physical theories predict that the distribution ought to be Poisson and hence we have taken it as the null hypothesis[2]

We define O_j as the number of time intervals in which we see exactly j particles. Thus $O_j = \sum_{i=1}^{2608} \mathbf{1}_{X_i=j}$. How do we find the expected numbers? If the null hypothesis had said that F has Poisson(1) distribution, we could use that to find the expected numbers. But H_0 only says Poisson(λ) for an unspecified λ? This brings in a new feature.

First estimate λ, for example $\hat{\lambda} = \overline{X}_n$ is an MLE as well as method of moments estimate. Then we use this to calculate Poisson probabilities and the expected numbers. In other words, $E_j = e^{-\hat{\lambda}} \frac{\hat{\lambda}^j}{j!}$. For the given data we find that $\hat{\lambda} = 3.87$. The table is as follows.

j	0	1	2	3	4	5	6	7	8	9	≥ 10
O_j	57	203	383	525	532	408	273	139	45	27	16
E_j	54.4	210.5	407.4	525.4	508.4	393.5	253.8	140.3	67.9	29.2	17.1

Two remarks: The original data would have consisted of several more bins for $j = 11, 12 \ldots$. These have been clubbed together to perform the χ^2 test (instead of a minimum of 5 per bin, they may have ensured that there are at least 10 per bin). Also, the estimate $\hat{\lambda} = 3.87$ was obtained before clubbing these bins. Indeed, if the data is merely presented as the above table, there will be some ambiguity in how to find $\hat{\lambda}$ as one of the bins says "≥ 10".

Then we compute

$$T = \sum_{j=0}^{10} \frac{(O_j - E_j)^2}{E_j} = 14.7.$$

Where should we look up in the χ^2 table? Earlier we said that the degrees of freedom is one less than the number of bins. Here we give the more general rule.

$$\text{Degrees of freedom of the } \chi^2 = \text{ No. of bins } - 1 - \text{No. of parameters estimated from data.}$$

In our case we estimated one parameter, namely λ. Hence the d.f. of the χ^2 is $11 - 1 - 1 = 9$. Looking at χ^2_9 table one can see that the p-value is 0.10. This is the probability that a χ^2_9 random variable is greater than 14.7. (Caution: Elsewhere I see that the p-value for this experiment is reported as 0.17, please check my calculations!). This means that at 5% level, we would not reject the null hypothesis. If the p-value was 0.17, we would not reject the null hypothesis even at 10% level.

[2]When a new theory is proposed, it should prove itself and is put in the alterntive hypotheis, but here we take it as null.

Fitting a continuous distribution: Chi-squared test can be used to test goodness of fit for continuous distributions too. We need some modifications. We must make bins of appropriate size, like $[a, a+h], [a+h, a+2h], \ldots, [a+h(k-1), a+hk]$ for a suitable h and k. Then we find the expected numbers in each bin using the null hypothesis (first estimating some parameters if necessary) and then proceed to compute T in the same way as before. Then check against the χ^2 table with the appropriate degrees of freedom. We omit details.

The probability theorem behind the χ^2-test for goodness of fit: Let $(W_1, \ldots, W_k)$ have multinomial distribution with parameters $n, m, (p_1, \ldots, p_k)$. In other words, place n balls at random into m bins, but each ball goes into the i^{th} bin with probability p_i, and distinct balls are assigned independently of each other. In the end, let W_i denote the number of balls in the i^{th} bin. The following proposition is the mathematics behind Pearson's test.

Proposition [Pearson]: Fix $k, p_1, \ldots, p_k$. Let $T_n = \sum_{i=1}^{k} \frac{(W_i - np_i)^2}{np_i}$. Then T_n converges to a χ^2_{k-1} distribution in the sense that $\mathbb{P}\{T_n \leq x\} \to \int_0^x f_{k-1}(u)du$ where f_{k-1} is the density of χ^2_{k-1} distribution.

How does this help? Suppose $X_1, \ldots, X_n$ are i.i.d. random variables taking k distinct values $t_1, t_2, \ldots, t_k$ (the actual values will not matter) with probabilities $p_1, \ldots, p_k$ respectively. Then, let W_i be the number of X_is whose value is t_i. Clearly, $(W_1, \ldots, W_k)$ has a multinomial distribution. Therefore, for large n, the random variable T_n defined above (which is in fact the χ^2-statistic of Pearson) has approximately χ^2_{k-1} distribution. This explains the test.

Sketch of proof of the proposition: Start with the case $k = 2$. Then, $W_1 \sim \text{Bin}(n, p_1)$ and $W_2 = r - W_1$. Thus, $T_n = \frac{(W_1 - np_1)^2}{np_1p_2}$ (recall that $p_1 + p_2 = 1$ and check this!). We know that $(W_1 - np_1)/\sqrt{np_1q_1}$ is approximately a $N(0,1)$ random variable, where $q_i = 1 - p_i$). Its square has (approximatelyχ^2_1 distribution. Thus the proposition is proved for $k = 2$.

When $k > 2$, what happens is that the random variables $\xi_i := (W_i - np_i)/\sqrt{np_iq_i}$ are approximately $N(0,1)$, but not independent. In fact the correlation between ξ_i and ξ_j is close to $-\sqrt{p_ip_j/q_iq_j}$. The sum of squares of ξ_is gives the χ^2 statistic. On the other hand, one can (with some clever linear algebra/matrix manipulation) write $\sum_{i=1}^{k} \xi_i^2$ as $\sum_{i=1}^{k-1} \eta_i^2$ where η_i are *independent* $N(0,1)$ random variables. Thus we get χ^2_{k-1} distribution.

8. Tests for independence

Suppose we have a bivariate sample $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ i.i.d. from a joint density (or joint pmf) $f(x, y)$. The question is to decide whether X_i is independent of Y_i.

Example 67. There are many situations in which such a problem arises. For example, suppose a bunch of students are given two exams, one testing mathematical skills and another testing verbal skills. The underlying goal may be to investigate whether the human brain has distinct centers for verbal and quantitative thinking.

EXAMPLE 68. As another example, say we want to investigate whether smoking causes lung cancer. In this case, for each person in the sample, we take two measurements - X (equals 1 if smoker and 0 if not) and Y (equal 1 if the person has lung cancer, 0 if not). The resulting data may be summarized in a two-way table as follows.

	$X=0$	$X=1$	
$Y=0$	$n_{0,0}$	$n_{0,1}$	$n_{0\cdot}$
$Y=1$	$n_{1,0}$	$n_{1,1}$	$n_{1\cdot}$
	$n_{\cdot 0}$	$n_{\cdot 1}$	n

Here the total sample is of n persons and $n_{i,j}$ denote the numbers in each of the four boxes. The numbers $n_{0\cdot}$ etc denote row or column sums. The statistical problem is to check if smoking (X) and incidence of lung cancer (Y) are positively correlated.

8.1. Special case of bivariate normal data. We shall not discuss this problem in detail but instead quickly give some indicators and move on. The point is that under the assumption that (X_i, Y_i) are jointly Gaussian, independence is equivalent to having no correlation, hence this becomes a parametric problem.

Here we have (X_i, Y_i) i.i.d bivariate normal random variables with $\mathbb{E}[X] = \mu_1$, $\mathbb{E}[Y] = \mu_2$, $\mathrm{Var}(X) = \sigma_1^2$, $\mathrm{Var}(Y) = \sigma_2^2$ and $\mathrm{Corr}(X, Y) = \rho$. The testing problem is $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$.

The natural statistic to consider is the sample correlation coefficient (*Pearson's r statistic*)

$$r_n := \frac{s_{X,Y}}{s_X \, s_Y}$$

where s_X^2, s_Y^2 are the sample variances of X and Y and $s_{X,Y} = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$ is the sample covariance. It is clear that the test should reject null hypothesis if r_n is far from 0. To decide the threshold we need the distribution of r_n under the null hypothesis.

THEOREM 69 (Fisher). *Under the null hypothesis, r_n^2 has Beta($\frac{1}{2}, \frac{n-2}{2}$) distribution. Alternately, under the null hypothesis, $\frac{r_n\sqrt{n-2}}{\sqrt{1-r_n^2}}$ has Student's-t distribution with $n-2$ degrees of freedom.*

Using this result, we can draw the threshold for rejection using the Beta distribution (of course the explicit threshold can only be computed numerically). That is, we reject the null hypothesis if and only if

$$\frac{|r_n|\sqrt{n-2}}{\sqrt{1-r_n^2}} \geq t_{n-2}(\alpha/2).$$

8.2. Without the normality assumption. If the assumption of normality of the data is not satisfied, then this test is invalid. However, for large n as usual we can obtain an asymptotically level-α test (for the hypothesis $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$, not for the question of independence) using the fact that $\frac{r_n\sqrt{n-2}}{\sqrt{1-r_n^2}}$ has approximately standard Normal distribution.

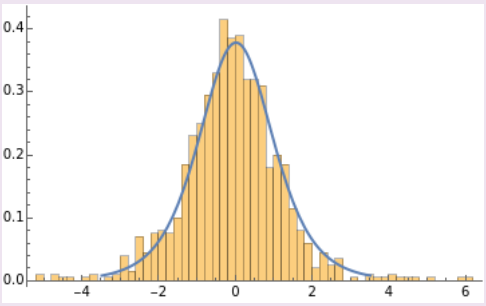

FIGURE 1. Histogram of 1000 samples of $\frac{r_n\sqrt{n-2}}{\sqrt{1-r_n^2}}$ with $n = 7$, overlaid with the density of student's t-distribution with 5 degrees of freedom.

Testing for independence in contingency tables: Here the measurements X and Y take values in $\{x_1, \ldots, x_k\}$ and $\{y_1, \ldots, y_\ell\}$, respectively. These x_i, y_j are categories, not numerical values (such as "smoking" and "non-smoking"). Let the total number of samples be n and let $N_{i,j}$ be the number of samples with values (x_i, y_j). Let $N_{i\cdot} = \sum_j N_{i,j}$ and let $N_{\cdot j} = \sum_i N_{i,j}$.

We want to test

$$H_0 : X \text{ and } Y \text{ are independent}$$

$$H_1 : X \text{ and } Y \text{ are not independent.}$$

Let $\mu(i,j) = \mathbb{P}\{X = x_i, Y = y_j\}$ be the joint pmf of (X, Y) and let $p(i)$, $q(j)$ be the marginal pmfs of X and Y respectively. From the sample, our estimates for these probabilities would be $\hat{\mu}(i,j) = N_{i,j}/n$ and $\hat{p}(i) = N_{i\cdot}/n$ and $\hat{q}(j) = N_{\cdot j}/n$ (which are consistent in the sense that $\sum_j \hat{\mu}(i,j) = \hat{p}(i)$ etc).

Under the null hypothesis we must have $\mu(i,j) = p(i)q(j)$. We test if these equalities hold (approximately) for the estimates. That is, define

$$T = \sum_{i=1}^{k}\sum_{j=1}^{\ell} \frac{(N_{i,j} - n\hat{p}(i)\hat{q}(j))^2}{n\hat{p}(i)\hat{q}(j)}.$$

Note that this is in the usual form of a χ^2 statistic (sum of (observed − expected)^2/expected).

The number of terms is $k\ell$. We lose one d.f. as usual but in addition we estimate $(k-1)$ parameters $p(i)$ (the last one $p(k)$ can be got from the others) and $(\ell - 1)$ parameters $q(j)$. Consequently, the total degress of freedom is $k\ell - 1 - (k-1) - (\ell - 1) = (k-1)(\ell - 1)$.

Hence, we reject the null hypothesis if $T > \chi^2_{(k-1)(\ell-1)}(\alpha)$ to get (an approximately) level α test.

9. Rank test for comparing two populations

We want to compare two populations on some feature. For example, IQ scores in two different groups or yields of rice in two fields treated with different fertilizers.

Let $X_1, \ldots, X_n$ be i.i.d from an unknown distribution F and let $Y_1, \ldots, Y_m$ be i.i.d. from a distribution G. Our test takes the form:

H_0: $F = G$.

H_1: G is larger than F in the sense that samples from G tend to be larger than samples from F.

The alternate hypothesis is not worded too precisely, but we leave it at that. If F and G we Normal distributions with means μ_1 and μ_2 (and the same variance σ^2), then the alternate hypothesis would be that $\mu_2 > \mu_1$ (which we have seen how to test using the t-statistic). For general distributions, it is not just the means that matter, so we have worded it as above.

Here is the beautiful idea on which we base the test. Combine all the X_is and the Y_is and arrange them in order (assume that there are no ties, which is true if the underlying F and G are continuous distributions). Let R_i denote the rank of X_i in the combined list (so $R_7 = 1$ is X_7 is the largest among all X_is and Y_is) and let R'_i denote the rank of Y_i in the combined list.

Under the null hypothesis, $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ are all i.i.d., hence all $(m+n)!$ orderings would be equally likely. In other words, the vector $(R_1, \ldots, R_n, R'_1, \ldots, R'_m)$ is uniformly distributed in the set S_n of all permutations of $\{1, 2, \ldots, m+n\}$.

Under the alternate hypothesis, X_i tend to be smaller than Y_is, and hence must have higher ranks. Therefore, a relevant statistic is Wilcoxon's rank statistic

$$W = \sum_{i=1}^{n} R_i - \frac{1}{2}n(n+m+1).$$

The reason for the second term is that under the null hypothesis, the for any i, the random variable R_i has uniform distribution in $\{1, 2, \ldots, n+m\}$ (but R_is are not independent of each other), hence $\mathbb{E}[R_i] = \frac{1}{2}(n+m+1)$. Therefore, $\mathbb{E}[W] = 0$. With a little more effort, one can also calculate that $\mathrm{Var}(W) = \frac{1}{12}mn(m+n+1)$ (under null hypothesis, of course).

A reasonable test is one that rejects the null if and only if W is larger than a threshold. There are three ways to come up with the cut-off values.

(1) There are theorems to the effect that if m, n are large, then under the null hypothesis, $\frac{W}{\sqrt{\mathrm{Var}(W)}}$ has approximately standard Gaussian distribution. Hence our test would be to reject the null if and only if $\frac{W-\mathbb{E}[W]}{\sqrt{\mathrm{Var}(W)}} > z_\alpha$.

(2) For very small m, n (so small that all $(m+n)!$ permutations of $\{1, 2, \ldots, m+n\}$ can be listed on a computer), we can find the exact probability mass function of W (under null hypothesis), by simply enumerating all permutations of $\{1, 2, \ldots, m+n\}$, and calculating the value of W for each. For example, when $n = 3$ and $m = 4$, here is the pmf of W.

k	0	± 1	± 2	± 3	± 4	± 5	± 6
$7!\mathbb{P}\{W = k\}$	720	576	576	432	288	144	144
$\mathbb{P}\{W = k\}$	0.142857	0.114286	0.114286	0.0857143	0.0571429	0.0285714	0.0285714

Once we have the exact distribution, we can of course draw the line at the $(1-\alpha)$-quantile. Keep in mind that in discrete situations, it may not be possible to find test of exact level α, unless randomization is allowed.

(3) If m and n are even moderately large, it becomes impossible to enumerate all the permutations and compute W for each. Instead we use simulations. Although we do not know the distribution from which the X_is and Y_is come (even under the null hypothesis), it is irrelevant to the rank statistic. We may assume that the common distribution is Unif$[0,1]$ (because random numbers from other distribution can be got by applying an increasing function to uniform random numbers, and ranks do not change under monotone transformations). We can just sample $m+n$ i.i.d. uniform random numbers, and compute W for it. Repeating this exercise a few hundred or a few thousand times, we get an idea of the distribution of W.

For example, here we have the results for $n=3$ and $m=4$, got by simulating 400 times.

k	0	± 1	± 2	± 3	± 4	± 5	± 6
Sample proportion (500 samples)	0.132	0.124	0.098	0.082	0.073	0.029	0.028
Sample proportion (1000 samples)	0.133	0.1145	0.1215	0.0845	0.0495	0.0325	0.031

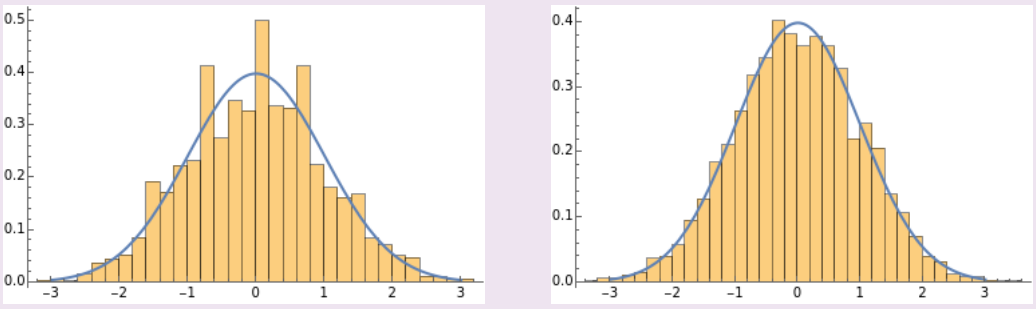

FIGURE 2. Histogram of 4000 samples of $\frac{W}{\sqrt{\text{Var}(W)}}$ with standard Normal density overlaid for comparison. Left: $n=10$ and $m=8$. Right: $n=20$ and $m=15$.

CHAPTER 13

Regression

1. Regression and Linear regression

1.1. Description of the problem. Consider two measurements in an experiment X, Y. They could be pressure and volume of a ballon, pressure and temperature of air in a chamber, height and weight of a person, body mass and brain mass of animals, etc. In these cases, we expect that there is a dependence between the variables. We think of the variable X as independent and Y as dependent, even if that is not always clear or true. We imagine that there is a functional relationship $Y = f(X)$ and it is the object of science to discover this function f.

The essential approach to this is to gather data, by experiment or by sampling or historical data, that we write as $(X_1, Y_1), \ldots, (X_n, Y_n)$. How to guess the function f from this data?

What is the difficulty, and where does statistics come into this? Even when the independent variable is under our control (like the pressure-volume experiment), there are errors in measurement. Hence, even if it is a truth that a functional dependence exists, Y_i is usually not exactly equal $f(X_i)$. In other situations (like the height-weight example), it is even worse as X_i, Y_i are realizations of random variables, and it may well happen that $X_1 = X_2$ but $Y_1 \neq Y_2$ (two people can have the same weight even if they have the same height), which means that can be no function such that $f(X_i) = Y_i$ for all i. The functional dependence is at best valid in some approximate, average sense, which can nevertheless be useful. Thus, statistics is essential.

Even so, if we allow ourselves to search among all possible functions, it is easy to find one that fits all the data points, i.e., there exists a function $f : \mathbb{R} \to \mathbb{R}$ (in fact we may take f to be a polynomials of degree n) such that $f(X_i) = Y_i$ for each $i \leq n$ (this is true if we assume that all X_i are distinct). This is not a good predictor, because the next data point (X_{n+1}, Y_{n+1}) will likely fall way off the curve. We have found a function that "predicts" well all the data we have, but not for a future observations! That is entirely useless.

To prevent this overfitting of data, the right approach is to restrict the class of functions. For example the collection of all linear functions $Y = \beta X + \alpha$ where $\alpha, \beta \in \mathbb{R}$. Here the nature of the function is assumed/imposed, and only the unknown parameters α, β are left to be determined from the data.

1.2. Why linear model? One may wonder if linearity is too restrictive. To some extent, but perhaps not as much as it sounds at first.

(1) Firstly, many relationships are linear in a reasonable range of the X variable (for example, resistance of a material versus temperature). The mathematical reason is

that any smooth functions is approximable by a linear function in a short enough interval.

(2) Secondly, we may sometimes transform the variables so that the relationship becomes linear. For example, if $Y = ae^{bX}$, then $\log(Y) = a' + b'X$ where $a' = \log(a)$ and $b' = \log(b)$ and hence in terms of the new variables X and $\log(Y)$, we have a linear relationship.

(3) Lastly, as a slight extension of linear regression, one can study *multiple linear regression*, where one has several independent variables $X^{(1)}, \ldots, X^{(p)}$ and try to fit a linear function $Y = \beta_1 X^{(1)} + \ldots + \beta_p X^{(p)}$. Once that is done, it increases the scope of curve fitting even more. For example, if we have two variable X, Y, then we can take $X^{(1)} = 1$, $X^{(2)} = X$, $X^{(3)} = X^2$. Then, linear regression of Y against $X^{(1)}, X^{(2)}, X^{(3)}$ is equivalent to fitting a quadratic polynomial curve $Y = a + bX + cX^2$.

(4) Combining the previous two points, we can also use linear regression to find fits such as $Y = a_0 + a_1 \cos(X) + a_2 \cos(2X) + b_1 \sin(X) + b_2 \sin(2X)$.

In short, multiple linear regression along with non-linear transformations of the individual variables allows the class of functions f is greatly extended.

Nevertheless, there are situations where one may consider other classes of functions with which to fit data. In the old style statistics, one would choose a simple, analytically tractable class of functions, possibly defined by a small number of parameters. (E.g., polynomials of degree at most 2). In modern ML (*machine learning*), the class of functions is simple to describe and very effective in practice, but mathematically ugly looking. One example is as follows: Consider the class of functions got by alternately composing (a) affine linear maps ($\mathbb{R}^k \ni u \mapsto Au + b \in \mathbb{R}^\ell$ for some $\ell \times k$ matrix A and vector $b \in \mathbb{R}^\ell$) and (b) co-ordinatewise ReLU function (that sends $(u_1, \ldots, u_k) \mapsto (u_1 \vee 0, \ldots, u_k \vee 0)$).The number of compositions allowed is usually fixed. While this class of functions is intractable mathematically, it is so large that very complex functions can be approximated by members of this class. These matters are outside the scope of this course.

2. One variable linear regression

We return to the simple linear regression model where X and Y are scalar variables, and we want to predict Y by a linear function $\alpha + \beta X$, where $\alpha, \beta \in \mathbb{R}$ are parameters to be chosen. To make a choice, we need a criterion for deciding the "best" fit. A basic one is the *method of least squares* which recommends finding α, β such that the error sum of squares $R^2 := \sum_{k=1}^n (Y_k - \alpha - \beta X_k)^2$ is minimized.

For fixed X_i, Y_i this is a simple problem in calculus. We get

$$\hat{\beta} = \frac{\sum_{k=1}^n (X_k - \overline{X}_n)(Y_k - \bar{Y}_n)}{\sum_{k=1}^n (X_k - \overline{X}_n)^2} = \frac{s_{X,Y}}{s_X^2}, \qquad \hat{\alpha} = \bar{Y}_n - \hat{\beta}\overline{X}_n$$

where $s_{X,Y}$ is the sample covariance of X, Y and s_X is the sample variance of X.

We leave the derivation of the least squares estimators by calculus to you. Instead we present another approach.

For a given choice of β, we know that the choice of α which minimizes R^2 is the sample mean of $Y_i - \beta X_i$ which is $\bar{Y} - \beta\bar{X}$. Thus, we only need to find $\hat{\beta}$ that minimizes

$$\sum_{k=1}^{n} \Big((Y_k - \bar{Y}) - \beta(X_k - \bar{X}) \Big)^2$$

and then we simply set $\hat{\alpha} = \bar{Y} - \beta\bar{X}$. Let[1] $Z_k = \frac{Y_k - \bar{Y}}{X_k - \bar{X}}$ and $w_k = (X_k - \bar{X})^2 / s_X^2$. Then,

$$\sum_{k=1}^{n} \Big((Y_k - \bar{Y}) - \beta(X_k - \bar{X}) \Big)^2 = s_X^2 \sum_{k=1}^{n} w_k \left(Z_k - \beta\right)^2 .$$

Since w_k are non-negative numbers that add to 1, we can intepret it as a probability mass function and hence we see that the minimizing β is given by the expectation with respect to this mass function. In other words,

$$\hat{\beta} = \sum_{k=1}^{n} w_k Z_k = \frac{s_{X,Y}}{s_X^2}.$$

Another way to write it is $\hat{\beta} = \frac{s_Y}{s_X} r_{X,Y}$ where $r_{X,Y}$ is the sample correlation coefficient.

A motivation for the least squares criterion: Suppose we make more detailed model assumptions as follows. Let X be a control variable (i.e., not random but we can tune it to any value, like temperature) and assume that $Y_i = \alpha + \beta X_i + \varepsilon_i$ where ε_i are i.i.d. $N(0, \sigma^2)$ "errors". Then, the data is essential Y_i that are independent $N(\alpha + \beta X_i, \sigma^2)$ random variables. Now we can extimate α, β by the maximum likelihood method.

EXAMPLE 70 (Hubble's 1929 experiment on the recession velocity of nebulae and their distance to earth). Hubble collected the following data that I took from this website. Here X is the number of megaparsecs from the nebula to earth and Y is the observed recession velocity in 10^3km/s.

X	0.032	0.034	0.214	0.263	0.275	0.275	0.45	0.5	0.5	0.63	0.8	2
Y	0.17	0.29	-0.13	-0.07	-0.185	-0.22	0.2	0.29	0.27	0.2	0.3	1.09
X	0.9	0.9	0.9	0.9	1	1.1	1.1	1.4	1.7	2	2	2
Y	-0.03	0.65	0.15	0.5	0.92	0.45	0.5	0.5	0.96	0.5	0.85	0.8

We fit two straight lines to this data.

(1) Fit the line $Y = \alpha + \beta X$. The least squares estimators (as derived earlier) turn out to be $\hat{\alpha} = -0.04078$ and $\hat{\beta} = 0.45416$. If $Z_i = \alpha + \beta X_i$ are the predicted values of Y_is, then one can see that the *residual sum of squares* is $\sum_i (Y_i - Z_i)^2 = 1.1934$.

[1] We are dividing by $X_k - \bar{X}$. What if it is zero for some k? But note that in the expression $\sum \Big((Y_k - \bar{Y}) - \beta(X_k - \bar{X}) \Big)^2$, all such terms do not involve β and hence can be safely left out of the summation. We leave the details for you to work out (the expressions at the end should involve all X_k, Y_k).

(2) Fit the line $Y = bX$. In this case we get $\hat{b}$ by minimizing $\sum_i (Y_i - bX_i)^2$. This is slightly different from before, but the same methods (calculus or the alternate argument we gave) work to give

$$\hat{b} = \frac{\sum_{i=1}^n Y_i X_i}{\sum_{i=1}^n X_i^2} = 0.42394.$$

The residual sum of squares $\sum_{i=1}^n (Y_i - bX_i)^2$ turns out to be 1.2064.

The residual sum of squares is smaller in the first, thus one may naively think that it is a better fit. However, note that the reduction is due to an extra parameter. Purely statistically, introducing extra parametrs will always reduce the residual sum of squares for obvious reasons. But the question is whether the extra parameter is worth the reduction. More precisely, if we fit the data too closely, then the next data point to be discovered (which may be nebula that is 10 megaparsecs away) may fall way off the curve.

More importantly, in this example, physics tells us that the line must pass through zero (that is, there is no recession velocity when two objects are very close). Therefore it is the second line that we consider, not the first. This gives the Hubble constant to be 423 km./s./megaparsec (the currently accepted values appear to be about 70, with data going up to distances of hundreds of megaparsecs...see this data!).

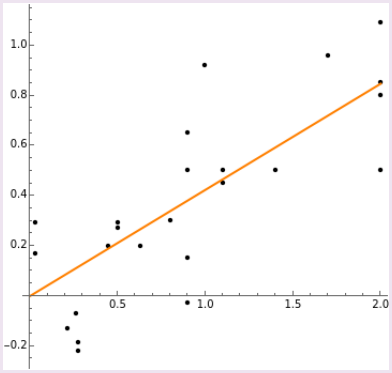

FIGURE 1. Regression (without intercept) of Y against X in Example 70

EXAMPLE 71. This example is from a wonderful compilation of data sets by A. P. Gore, S. A. Paranjpe, M. B. Kulkarni (the link appea. In this example, Y denotes the number of frogs of age X (in some delimited population).

X	1	2	3	4	5	6	7	8
Y	9093	35	30	28	12	8	5	2

A prediction about life-times says that the survival probability $P(t)$ (which is the chance that an individual survives up to age t or more) decays as $P(t) = Ae^{-bt}$ for some constants A and b. We would like to check this against the given data.

What we need are individuals that survive beyond age t. Taking Z to be the cumulative sums of Y, this gives us

X	1	2	3	4	5	6	7	8
Z	9213	120	85	55	27	15	7	2
$P = Z/n$	1.0000	0.0130	0.0092	0.0060	0.0029	0.0016	0.0008	0.0002
$W = \log P$	0	-4.3409	-4.6857	-5.1210	-5.8325	-6.4203	-7.1825	-8.4352

We compute that $\bar{X} = 4.5$, $\bar{W} = -5.25$, $\mathrm{std}(X) = 2.45$, $\mathrm{std}(W) = 2.52$ and $\mathrm{corr}(X, W) = -0.92$. Hence, in the linear regression $W = a + bX$, we see that $\hat{b} = 0.94$ and $\hat{a} = -9.49$. The residual sum of squares is 7.0.

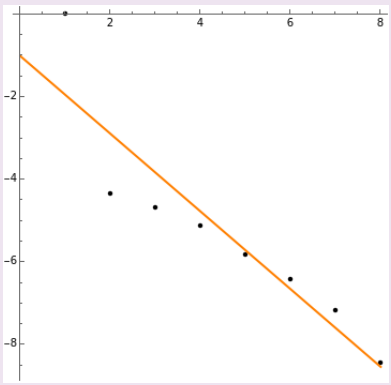

FIGURE 2. Regression of W against X in Example 71.

2.1. How good is the fit? For the same data $(X_1, Y_1), \ldots, (X_n, Y_n)$, suppose we have two candidates (a) $Y = f(X)$ and (b) $Y = g(X)$. How to decide which is better? Or how to say if a fit is good at all?

By the least-squares criterion, the answer is the one with smaller residual sum of squares $SS := \sum_{k=1}^{n}(Y_k - f(X_k))^2$. Usually one presents a closely related quantity $R^2 = 1 - \frac{SS}{SS_0}$ (where $\mathrm{SS}_0 = \sum_{k=1}^{n}(Y_k - \bar{Y})^2 = (n-1)s_Y^2$). Since SS_0 is (a multiple of) the total variance in Y, R^2 measures how much of it is "explained" by a particular fit. Note that $0 \leq R^2 \leq 1$. And higher (i.e., closer to 1) the R^2 is, the better the fit.

Thus, the first naive answer to the above question is to compute R^2 in the two situations (fitting by f and fitting by g) and see which is higher. But a more nuanced approach is preferable. Consider the same data and three situations.

(1) Fit a constant function. This means, choose α to minimize $\sum_{k=1}^{n}(Y_k - \alpha)^2$. The solution is $\hat{\alpha} = \bar{Y}$ and the residual sum of squares is SS_0 itself. Then, $R_0^2 = 0$.

(2) Fit a linear function. Then α, β are chosen as discussed earlier and the residual sum of squares is $\mathrm{SS}_1 = \sum_{k=1}^{n}(Y_k - \hat{\alpha} - \hat{\beta}X_k)^2$. Then, $R_1^2 = 1 - \frac{SS_1}{SS_0}$.

(3) Fit a quadratic function. The the residual sum of squares is $\mathrm{SS}_2 = \sum_{k=1}^{n}(Y_k - \hat{\alpha} - \hat{\beta}X_k - \hat{\gamma}X_k^2)^2$ where $\hat{\alpha}, \hat{\beta}, \hat{\gamma}$ are chosen so as to minimize $\sum_{k=1}^{n}(Y_k - \alpha - \beta X_k - \gamma X_k^2)^2$. Then $R_2^2 = 1 - \frac{SS_2}{SS_0}$.

Obviously we will have $R_2^2 \geq R_1^2 \geq R_0^2$ (since linear functions include constants and quadratic functions include linear ones). Does that mean that the third is better? If that were the conclusion, then we can continue to introduce more parameters as that will always reduce the residual sum of squares! But that comes at the cost of making the model more complicated (and having too many parameters means that it will fit the current data well, but not future data!). When to stop adding more parameters?

Qualitatively, a new parameter is desirable if it leads to a *significant increase* of the R^2. The question is, how big an increase is significant. For this, one introduces the notion of *adjusted* R^2, which is defined as follows:

If the model has p parameters, then define $\overline{SS} = SS/(n-1-p)$. In particular, $\overline{SS}_0 = \frac{SS_0}{n-1} = s_Y^2$. Then define the adjusted R^2 as $\overline{R}^2 = 1 - \frac{\overline{SS}}{\overline{SS}_0}$.

In particular, $\overline{R}_0^2 = R_0^2$ as before. But $R_1^2 = 1 - \frac{SS_1/(n-2)}{SS_0/(n-1)}$. Note that $\overline{R}^2$ does not necessarily increase upon adding an extra parameter. If we want a polynomial fit, then a rule of thumb is to keep adding more powers as long as $\overline{R}^2$ continues to increase and stop the moment it decreases.

EXAMPLE 72. To illustrate the point let us look at a simulated data set. I generated 25 i.i.d $N(0,1)$ variables X_i and then generated 25 i.i.d. $N(0,1/4)$ variables ε_i. And set $Y_i = 2X_i + \varepsilon_i$. The data set obtained was as follows.

X	-0.87	0.07	-1.22	-1.12	-0.01	1.53	-0.77	0.37	-0.23	1.11	-1.09	0.03	0.55
Y	-2.43	-0.56	-2.19	-2.32	-0.12	3.77	-1.4	0.84	0.34	1.83	-1.83	0.48	0.98
X	1.1	1.54	0.08	-1.5	-0.75	-1.07	2.35	-0.62	0.74	-0.2	0.88	-0.77	
Y	2.3	2.5	-0.41	-2.94	-1.13	-0.84	4.36	-1.14	1.45	-1.36	1.55	-2.43	

To this data set we fit two models (A) $Y = \beta X$ and (B) $Y = a + bX$. The results are as follows.

$$\text{SS}_0 = 96.20, \;\; R_0^2 = 0$$

$$\text{SS}_1 = 6.8651, \;\; R_1^2 = 0.9286, \;\; \bar{R}_1^2 = 0.9255$$

$$\text{SS}_2 = 6.8212, \;\; R_2^2 = 0.9291, \;\; \bar{R}_2^2 = 0.9227.$$

Note that the adjusted R^2 decreases (slightly) for the the second model. Thus, if we go by that, then the model with one parameter is chosen (correctly, as we generated from that model!). You can try various simulations yourself. Also note the high value of R_1^2 (and R_2^2) which indicates that it is not a bad fit at all.

3. Regression with more than one independent variable

Here we have a several variables $X_1, \ldots, X_p$ and Y. We want to predict Y based on a linear function of $X = (X_1, \ldots, X_p)$. In other words, we want to write

$$Y = \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon_p,$$

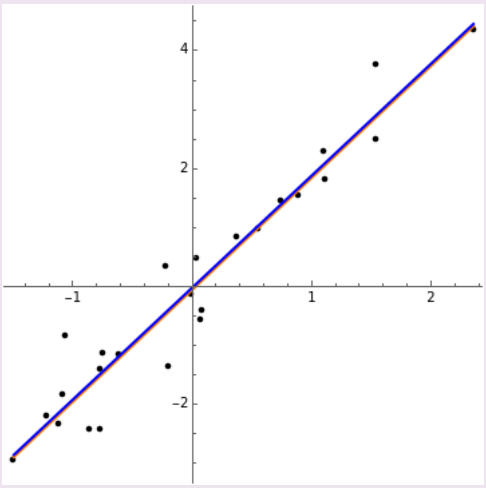

FIGURE 3. Regression of Y against X in Example 72. The blue line has one parameter (slope) and the orange line has two parameters (slope and intercept)

where ε_p is the prediction error. The problem is to find the best coefficients $\beta_1, \ldots, \beta_p$ so that the prediction errors are as small as possible. Observe that we did not allow a constant coefficient β_0. It is not needed because we may take $X_1 = 1$, the constant, in which case β_1 become the constant coefficient. Thus, the one variable linear regression we did corresponds to $p = 2$, with $X_1 = 1$ and $X_2 = X$.

The data we have is of the kind $(X_{i,1}, \ldots, X_{i,p}, Y_i)$ for $1 \leq i \leq n$. The least squares principles asks us to choose $\beta_1, \ldots, \beta_p$ so that

$$R^2 := \sum_{i=1}^{n} (Y_i - \beta_1 X_{i,1} - \beta_2 X_{i,2} - \ldots - \beta_p X_{i,p})^2 \tag{24}$$

is minimized.

3.1. First approach to the least squares minimization. First let us solve the coefficients in a simple way. Later we shall explain how a more sophisticated view with vectors and matrices is more illuminating. If we fix all coefficients except β_k, the right side of (24) is quadratic expression in β_k with positive coefficient for β_k^2. Therefore, there is a unique minimizer got by setting the derivative equal to 0 (or however you know to minimize quadratic expressions). That gives us the equations

$$0 = \frac{\partial R^2}{\partial \beta_k} = 2 \sum_{i=1}^{n} X_{i,k} (Y_i - \beta_1 X_{i,1} - \beta_2 X_{i,2} - \ldots - \beta_p X_{i,p})$$

which can be rewritten as

$$\sum_{i=1}^{k} Y_i X_{i,k} = \sum_{i=1}^{n} (\beta_1 X_{i,1} - \beta_2 X_{i,2} - \ldots - \beta_p X_{i,p}) X_{i,k}, \quad 1 \leq k \leq p.$$

It is useful to introduce the symbols $s_{U,V} = \sum_{i=1}^{n} u_i v_i$ for any two variables U, V. Then the above equations can be written more succinctly as

$$s_{Y,X_k} = s_{X_k,X_1}\beta_1 + \ldots s_{X_k,X_p}\beta_p, \;\; 1 \leq k \leq p.$$

Or writing the matrix $S_{X,X} := (s_{X_k,X_j})_{k,j\leq p}$ and the vector $S_{Y,X} = (s_{Y,X_k})_{k\leq p}$, we can write these as a single matrix equation $S_{Y,X} = S_{X,X}\beta$, where $\beta = (\beta_1, \ldots, \beta_p)^t$. The solution is of course

$$\beta^* = S_{X,X}^{-1} S_{Y,X}$$

provided we assume that $S_{X,X}$ is invertible. This is usually the case, so we do not worry about what to do when $S_{X,X}$ is singular.

3.2. Geometric approach to least squares minimization. Let us explain the same in a different way, assuming a basic knowledge of vectors and matrices. We introduce the $n \times 1$ column vector $Y = (Y_1, \ldots, Y_n)^t$, the $n \times p$ matrix $X = (X_{i,j})_{i\leq n, j\leq p}$ and the $p \times 1$ coefficient vector $\beta = (\beta_1, \ldots, \beta_p)^t$. Then

$$R^2 = \|Y - X\beta\|^2.$$

As β varies over $\mathbb{R}^p$, the vectors $X\beta$ span the column space of X, denoted $\mathcal{C}(X)$, a subspace of $\mathbb{R}^n$. Thus, minimizing R^2 is the problem of finding the closest point to Y in the subspace $\mathcal{C}(X)$. But that is precisely the problem of projections. If Q denotes the projection matrix onto $\mathcal{C}(X)$ (which means that $Qv = v$ for $v \in \mathcal{C}(X)$ and $Qv = 0$ for $v \perp \mathcal{C}(X)$), then R^2 is minimized when we choose β so that $X\beta = QY$, and the squared-error $R^2 = \|(I-Q)Y\|^2 = \|Y\|^2 - \|QY\|^2$.

The only remaining question is to find the matrix Q explicitly in terms of X.

LEMMA 73. *If rank$(X) = p$, then $Q = X(X^tX)^{-1}X^t$.*

PROOF. When X has rank p, so does X^tX (why?) and hence it is invertible. Therefore Q is well-defined and has rank at most p (when multiplying matrices, rank only decreases). Observe that $Q^t = Q$ and $Q^2 = Q$. Therefore Q is an orthogonal projection matrix. What is the space onto which it projects? Observe that $QX = X$, hence Q preserves the column space of X. In particular it has rank at least p. As rank$(Q) \leq p$, it follows that the range of Q is exactly $\mathcal{C}(X)$. In other words, Q is the orthogonal projection matrix onto $\mathcal{C}(X)$. ■

From the lemma and the discussion before it, we see that the solution to $Q\beta = Y$ is given by $\beta^* = X(X^tX)^{-1}XY$ and the error sum of squares is $R^2 = \|Y\|^2 - \|X\beta^*\|^2$.

CHAPTER 14

Sample surveys

Appendix: A proof of CLT under third moment assumption

We assume that $X_1, X_2, \ldots$ are i.i.d. with mean 0, variance 1 and finite third moment. The goal show that for any $a < b$

$$\mathbb{P}\left\{\frac{S_n}{\sqrt{n}} \in [a,b]\right\} \to \Phi(b) - \Phi(a). \tag{25}$$

We assume that on the same probability space we also have $Y_1, Y_2, \ldots$ that are i.i.d. $N(0,1)$ and that are also independent of the X_is. We write $S_n^Y = Y_1 + \ldots + Y_n$. Then $\frac{S_n^Y}{\sqrt{n}} \sim N(0,1)$ and hence $\Phi(b) - \Phi(a) = \mathbb{P}\{S_n^Y/\sqrt{n} \in [a,b]\}$. Thus, (**??**) can be rewritten as

$$\mathbb{E}\left[g\left(\frac{S_n^X}{\sqrt{n}}\right)\right] - \mathbb{E}\left[g\left(\frac{S_n^X}{\sqrt{n}}\right)\right] \to 0 \tag{26}$$

where $g = \mathbf{1}_{[a,b]}$. The proof proceeds in two steps:

(1) Prove (26) for smooth functions $g : \mathbb{R} \to \mathbb{R}$ with the property that g and its first three derivatives are uniformly bounded.

(2) Approximate $\mathbf{1}_{[a,b]}$ by smooth functions to deduce (26) for $g = \mathbf{1}_{[a,b]}$.

Step-1: Assume that g is smooth and that its first three derivatives are uniformly bounded. Write the telescoping sum

$$g\left(\frac{S_n^X}{\sqrt{n}}\right) - g\left(\frac{S_n^Y}{\sqrt{n}}\right) = \sum_{k=1}^{n} g(V_{k-1}) - g(V_k) \tag{27}$$

where $V_k = \frac{Y_1+\ldots+Y_k+X_{k+1}+\ldots+X_n}{\sqrt{n}}$ (thus $V_0 = S_n^X/\sqrt{n}$ and $V_n = S_n^Y/\sqrt{n}$). We further split the kth term as

$$g(V_{k-1}) - g(V_k) = (g(V_{k-1}) - g(W_k)) - (g(V_k) - g(W_k))$$

where $W_k = \frac{Y_1+\ldots+Y_{k-1}+0+X_{k+1}+\ldots+X_n}{\sqrt{n}}$ (key point is that W_k does not involve either X_k or Y_k). We write the Taylor series expansions

$$g(V_{k-1}) - g(W_k) = g'(W_k)\frac{X_k}{\sqrt{n}} + g''(W_k)\frac{X_k^2}{2n} + g'''(W_k^*)\frac{X_k^3}{6n^{\frac{3}{2}}},$$

$$g(V_k) - g(W_k) = g'(W_k)\frac{Y_k}{\sqrt{n}} + g''(W_k)\frac{Y_k^2}{2n} + g'''(W_k^\#)\frac{Y_k^3}{6n^{\frac{3}{2}}},$$

where W_k^* is some point between W_k and V_{k-1} and $W_k^\#$ is some point between W_k and V_k. Take expectations in the above pair of expressions, and use the independence of W_k with

X_k, Y_k to see that

$$\mathbb{E}[g(V_{k-1}) - g(W_k)] = \frac{1}{\sqrt{n}}\mathbb{E}[g'(W_k)]\mathbb{E}[X_k] + \frac{1}{2n}\mathbb{E}[g''(W_k)]\mathbb{E}[X_k^2] + \frac{1}{6n^{\frac{3}{2}}}\mathbb{E}[g'''(W_k^*)X_k^3],$$
$$\mathbb{E}[g(V_k) - g(W_k)] = \frac{1}{\sqrt{n}}\mathbb{E}[g'(W_k)]\mathbb{E}[Y_k] + \frac{1}{2n}\mathbb{E}[g''(W_k)]\mathbb{E}[Y_k^2] + \frac{1}{6n^{\frac{3}{2}}}\mathbb{E}[g'''(W_k^*)Y_k^3].$$

Take the difference of the two and recall that $\mathbb{E}[X_k] = 0 = \mathbb{E}[Y_k]$ and $\mathbb{E}[X_k^2] = 1 = \mathbb{E}[Y_k^2]$ to get

$$\mathbb{E}[g(V_{k-1})] - \mathbb{E}[g(V_k)] = \frac{1}{6n^{\frac{3}{2}}}\left(\mathbb{E}[g'''(W_k^*)X_k^3] + \mathbb{E}[g'''(W_k^{\#})Y_k^3]\right)$$

When we take absolute value, the term in the bracket is bounded by $\|g'''\|_{\sup}\gamma$ where $\gamma = \mathbb{E}[|X_1|^3] + \mathbb{E}[|Y_1|^3]$ (just some number we don't care about). Plug these bounds (which are the same for all k) into (27) to get

$$\left|\mathbb{E}\left[g\left(\frac{S_n^X}{\sqrt{n}}\right)\right] - \mathbb{E}\left[g\left(\frac{S_n^Y}{\sqrt{n}}\right)\right]\right| \leq \frac{\gamma\|g'''\|_{\sup}}{6\sqrt{n}}.$$

From this, (26) follows for smooth g with bounded derivatives up to order 3.

Step-2: Fix any $c < a < b < d$. One can find a smooth function $g : \mathbb{R} \to \mathbb{R}$ such that $\mathbf{1}_{[a,b]} \leq g \leq \mathbf{1}_{[c,d]}$ (this means that $0 \leq g \leq 1$ and $g = 1$ on $[a,b]$ and $g = 0$ outside $[c,d]$). As g and its derivatives are compactly supported and continuous, they are uniformly bounded, and hence (26) applies to this g. By the monotonicity of expectation, with $Z \sim N(0,1)$, we have

$$\mathbb{P}\left\{\frac{S_n}{\sqrt{n}} \in [a,b]\right\} \leq \mathbb{E}\left[g\left(\frac{S_n}{\sqrt{n}}\right)\right] \leq \mathbb{P}\left\{\frac{S_n}{\sqrt{n}} \in [c,d]\right\},$$
$$\mathbb{P}\{Z \in [a,b]\} \leq \mathbb{E}[g(Z)] \leq \mathbb{P}\{Z \in [c,d]\}$$

As $\mathbb{E}\left[g\left(\frac{S_n}{\sqrt{n}}\right)\right] \to \mathbb{E}[g(Z)]$ by Step-1, we see that

$$\limsup_{n\to\infty} \mathbb{P}\left\{\frac{S_n}{\sqrt{n}} \in [a,b]\right\} \leq \lim_{n\to\infty} \mathbb{E}\left[g\left(\frac{S_n}{\sqrt{n}}\right)\right] = \mathbb{E}[g(Z)] \leq \mathbb{P}\{Z \in [c,d]\},$$
$$\mathbb{P}\{Z \in [a,b]\} \leq \mathbb{E}[g(Z)] = \lim_{n\to\infty} \mathbb{E}\left[g\left(\frac{S_n}{\sqrt{n}}\right)\right] \leq \liminf_{n\to\infty} \mathbb{P}\left\{\frac{S_n}{\sqrt{n}} \in [c,d]\right\}.$$

Fixing $a = a_0, b = b_0$ and letting $c \uparrow a_0$ and $d \downarrow b_0$ or fixing $c = a_0, d = b_0$ and letting $a \downarrow a_0$ and $b \uparrow b_0$, we conclude that $\mathbb{P}\{S_n \in [a_0, b_0]\} \to \mathbb{P}\{Z \in [a_0, b_0]\}$ for any $a_0 < b_0$. The proof of CLT is complete. ■

Appendix: Stirlings' approximation

Stirling's approximation states that $\Gamma(\nu+1) \sim \nu^{\nu+\frac{1}{2}}e^{-\nu}\sqrt{2\pi}$ as $\nu \to \infty$. In particular, $n! \sim n^{n+\frac{1}{2}}e^{-n}\sqrt{2\pi}$ as $n \to \infty$ along integers.

To prove Stirlings' approximation, recall that

$$\Gamma(\nu+1) = \int_0^\infty e^{-x}x^\nu dx \;=\; \nu^{\nu+1}\int_0^\infty e^{-\nu t}t^\nu dt \;=\; \nu^{\nu+1}e^{-\nu}\int_0^\infty e^{-\nu(t-\log t-1)}dt. \tag{28}$$

The function $t \mapsto t - \log t - 1$ attains its minimum value of 0 uniquely at $t = 1$. Taking its Taylor expansion around 1, we can write $t - \log t - 1 = \frac{1}{2}(t-1)^2 + O(|t-1|^3)$. Therefore, given $\varepsilon > 0$, there is some $\delta > 0$ such that

$$(1-\varepsilon)\frac{1}{2}t^2 \le t - \log t - 1 \le (1+\varepsilon)\frac{1}{2}t^2 \quad \text{for } t \in J_\delta := [1-\delta, 1+\delta].$$

Therefore, $\int_{J_\delta} e^{-\nu(t-1-\log t)}dt$ is sandwiched between

$$\begin{aligned}\int_{J_\delta} e^{-\nu(1\pm\varepsilon)\frac{(t-1)^2}{2}}dt &= \frac{1}{\sqrt{\nu(1\pm\varepsilon)}}\int_{-\delta\sqrt{\nu(1\pm\varepsilon)}}^{\delta\sqrt{\nu(1\pm\varepsilon)}} e^{-\frac{1}{2}u^2}du \\ &= \frac{\sqrt{2\pi}}{\sqrt{\nu}} \times \frac{1-\bar{\Phi}(\delta\sqrt{\nu(1\pm\varepsilon)})}{\sqrt{1\pm\varepsilon}}.\end{aligned}$$

Now we consider the integral over J_δ^c. There, $\eta := \inf_{t\in J_\delta^c}(t-1-\log t)$ is strictly positive (as 1 is the unique minimizer and $t-1-\log t \to \infty$ as $t \to 0$ or $t \to \infty$). Therefore,

$$\int_{J_\delta^c} e^{-\nu(t-\log t-1)}dt \le e^{-(\nu-1)\eta}\int_{J_\delta^c} e^{-(t-\log t-1)}dt \le C_\delta e^{-\nu\eta}$$

Putting these together, and recalling that $\bar{\Phi}(u) \le e^{-u^2/2}$ for $u \ge 1$ and using that $\varepsilon > 0$ is arbitrary, we see that

$$\int_0^\infty e^{-\nu(t-1-\log t)}dt \sim \frac{\sqrt{2\pi}}{\sqrt{\nu}} \quad \text{as } \nu \to \infty.$$

Plugging this into (28), we arrive at Stirlings' approximation.