PROBABILITY AND STATISTICS

MANJUNATH KRISHNAPUR

CONTENTS

Probability

1. 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 - for example, we can have a debate between *eyeists* and *nosists* as to whether it is better to compare the eyes or the noses in arriving at a decision.

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.

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 had low probability. 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 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 simple (and you can do it yourself)

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

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

Summary: There are many situations in real life where one is required to make decisions under uncertainty. A general template for the answer could be to fix a small number that we allow as the probability of error, and deduce thresholds based on it. This brings us to the question of computing probabilities in various situations.

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 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 situations in which it is reasonable to take a model in probability

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 52! orders of the 52 cards are assumed to be equally likely. 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.

However...: Probability and statistics are very relevant in many situations that do not involve any uncertainty on the face of it. Here are some examples.

Example 5. *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 6. Search algorithms such as Google, use many randomized procedures. This cannot be explained right now, 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.

2. Discrete probability spaces

Definition 7. 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*. Ω is called the *sample space* and p_ω are called *elementary probabilities*.

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

All of probability in one line: Take an (interesting) probability space (Ω, p) and an (interesting) event $A \subseteq \Omega$. Find $\mathbf{P}(A)$.

This is the mathematical side of the picture. It is easy to make up any number of probability spaces - simply take a finite set and assign non-negative numbers to each element of the set so that the total is 1.

Example 8. $\Omega = \{0,1\}$ and $p_0 = p_1 = \frac{1}{2}$. There are only four events here, $\emptyset, \{0\}, \{1\}$ and $\{0,1\}$. Their probabilities are, 0, $1/2$, $1/2$ and 1, respectively.

Example 9. $\Omega = \{0,1\}$. Fix a number $0 \le p \le 1$ and let $p_1 = p$ and $p_0 = 1-p$. The sample space is the same as before, but the probability space is different for each value of p. Again there are only four events, and their probabilities are $\mathbf{P}\{\emptyset\} = 0$, $\mathbf{P}\{0\} = 1-p$, $\mathbf{P}\{1\} = p$ and $\mathbf{P}\{0,1\} = 1$.

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

Example 10. Fix a positive integer n. Let

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

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.

There are $2^{\#\Omega} = 2^{2^n}$ events. One example is $A_k = \{\underline{\omega} : \underline{\omega} \in \Omega \text{ and } \omega_1 + \ldots + \omega_n = k\}$ where k is some fixed integer. In words, A_k consists of those n-tuples of zeros and ones that have a total of k many ones. Since there are $\binom{n}{k}$ ways to choose where to place these ones, we see that $\#A_k = \binom{n}{k}$. Consequently,

$$\mathbf{P}\{A_k\} = \sum_{\underline{\omega} \in A_k} p_{\underline{\omega}} = \frac{\#A_k}{2^n} = \begin{cases} \binom{n}{k} 2^{-n} & \text{if } 0 \leq k \leq n, \\ 0 & \text{otherwise.} \end{cases}$$

It will be convenient to adopt the notation that $\binom{a}{b} = 0$ if a, b are positive integers and if $b > a$ or if $b < 0$. Then we can simply write $\mathbf{P}\{A_k\} = \binom{n}{k} 2^{-n}$ without having to split the values of k into cases.

Example 11. Fix two positive integers r and m. Let

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

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.

Of course, there are 2^{m^r} many events, which is quite large even for small numbers like $m = 3$ and $r = 4$. Some interesting events are $A = \{\underline{\omega} : \omega_r = 1\}$, $B = \{\underline{\omega} : \omega_i \neq 1 \text{ for all } i\}$, $C = \{\underline{\omega} : \omega_i \neq \omega_j \text{ if } i \neq j\}$. The reason why these are interesting will be explained later. Because of equal elementary probabilities, the probability of an event S is just $\#S/m^r$.

- Counting A: We have m choices for each of $\omega_1, \ldots, \omega_{r-1}$. There is only one choice for ω_r. Hence $\#A = m^{r-1}$. Thus, $\mathbf{P}(A) = \frac{m^{r-1}}{m^r} = \frac{1}{m}$.
- Counting B: We have $m-1$ choices for each ω_i (since ω_i cannot be 1). Hence $\#B = (m-1)^r$ and thus $\mathbf{P}(B) = \frac{(m-1)^r}{m^r} = (1 - \frac{1}{m})^r$.
- Counting C: We must choose a distinct value for each $\omega_1, \ldots, \omega_r$. This is impossible if $m < r$. If $m \geq r$, then ω_1 can be chosen as any of m values. After ω_1 is chosen, there are $(m-1)$ possible values for ω_2, and then $(m-2)$ values for ω_3 etc., all the way till ω_r which has $(m-r+1)$ choices. Thus, $\#C = m(m-1)\ldots(m-r+1)$. Note that we get the same answer if we choose ω_i in a different order (it would be strange if we did not!).

  Thus, $\mathbf{P}(C) = \frac{m(m-1)\ldots(m-r+1)}{m^r}$. Note that this formula is also valid for $m < r$ since one of the factors on the right side is zero.

2.1. **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 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 12. Physical situation: Toss a coin. Randomness enters because we believe that the coin may turn up head or tail and that it is inherently unpredictable.

The corresponding probability model: Since there are two outcomes, the sample space $\Omega = \{0, 1\}$ (where we use 1 for heads and 2 for tails) is a clear choice. What about elementary probabilities? Under the equal chance hypothesis, we may take $p_0 = p_1 = \frac{1}{2}$. Then we have a probability model for the coin toss.

If the coin was not fair, we would change the model by keeping $\Omega = \{0, 1\}$ as before but letting $p_1 = p$ and $p_0 = 1 - p$ where the parameter $p \in [0, 1]$ is fixed.

Which model is correct? If the coin looks very symmetrical, then the two sides are equally likely to turn up, so the first model where $p_1 = p_0 = \frac{1}{2}$ is reasonable. However, if the coin looks irregular, then theoretical considerations are usually inadequate to arrive at the value of p. Experimenting with the coin (by tossing it a large number of times) is the only way.

There is always an approximation in going from the real-world to a mathematical model. For example, the model above ignores the possibility that the 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...

[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.

Thus we see that example 9 is a good model for a physical coin toss. What physical situations are captured by the probability spaces in example 10 and example 11?

Example 10: This probability space can be a model for tossing n fair coins. It is clear in what sense, so we omit details for you to fill in.

The same probability space can also be a model for the tossing of the same coin n times in succession. In this, we are implicitly assuming 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 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 11: There are several situations that can be captured by this probability space. We list some.

- There are r labelled balls and m labelled bins. One by one, we put the balls into bins "at random". Then, by letting ω_i be the bin-number into which the i^{th} ball goes, we can capture the full configuration by the vector $\underline{\omega} = (\omega_1, \ldots, \omega_n)$. If each ball is placed completely at random then the probabilities are m^{-r} for each configuration $\underline{\omega}$.

  In that example, A is the event that the last ball ends up in the first bin, B is the event that the first bin is empty and C is the event that no bin contains more than one ball.
- If $m = 6$, then this may also be the model for throwing a fair die r times. Then ω_i is the outcome on the i^{th} throw. Of course, it also models throwing r different (and distinguishable) fair dice.
- If $m = 2$ and $r = n$, this is same as Example 10, and thus models the tossing of n fair coins (or a fair coin n times).
- Let $m = 365$. Omitting the possibility of leap years, this is a model for choosing r people at random and noting their birthdays (which can be in any of 365 "bins"). If we assume that all days are equally likely as a birthday (is this really true?), then the same probability space is a model for this physical situation. In this example, C is the event that no two people have the same birthday.

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 $=(b_1, \ldots, b_n)$. Then the possible offsprings gene sequences are in the set $\Omega_0 := \{\mathbf{x} \in \{0,1\}^n : x_i = a_i \text{ or } b_i, \text{ for each } i \leq n\}$. Let $L := \#\{i : a_i \neq b_i\}$.

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 $p_{\mathbf{x}} = 2^{-L}$ for each $\mathbf{x} \in \Omega_0$.

(2) Let Ω be the sample space and let

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

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 $A = \{\mathbf{x} \in \Omega_0 : x_0 + \ldots + x_{10} \leq 4\}$. (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 use.

Remark 14. We have discussed at length the reasonability of the model in this example to indicate the enormous 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.

3. Examples of discrete probability spaces

Example 15. Toss n coins. We saw this before, but assumed that the coins are fair. Now we do not. The sample space is

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

Further we assign $p_{\underline{\omega}} = \alpha^{(1)}_{\omega_1} \ldots \alpha^{(n)}_{\omega_n}$. Here $\alpha^{(j)}_0$ and $\alpha^{(j)}_1$ are supposed to indicate the probabilities that the j^{th} coin falls tails up or heads up, respectively. Why did we take the product of $\alpha^{(j)}_{\cdot}$s and not some other combination? This is a non-mathematical question about what model is suited for the given real-life example. For now, the only justification is that empirically the above model seems to capture the real life situation accurately.

In particular, if the n coins are identical, we may write $p = \alpha^{(j)}_1$ (for any j) and the elementary probabilities become $p_{\underline{\omega}} = p^{\sum_i \omega_i} q^{n - \sum_i \omega_i}$ where $q = 1 - p$.

Fix $0 \leq k \leq n$ and let $B_k = \{\underline{\omega} : \sum_{i=1}^n \omega_i = k\}$ be the event that we see exactly k heads out of n tosses. Then $\mathbf{P}(B_k) = \binom{n}{k} p^k q^{n-k}$. If A_k is the event that there are at least k heads, then $\mathbf{P}(A_k) = \sum_{\ell=k}^{n} \binom{n}{\ell} p^\ell q^{n-\ell}$.

Example 16. Toss a coin n times. Again

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

$$p_{\underline{\omega}} = p^{\sum_i \omega_i} q^{n - \sum_i \omega_i}.$$

This is the same probability space that we got for the tossing of n identical looking coins. Implicit is the assumption that once a coin is tossed, for the next toss it is as good as a different coin but with the same p. It is possible to imagine a world where coins retain the memory of what happened before (or as explained before, we can make a "coin" that remembers previous tosses!), in which case this would not be a good model for the given situation. We don't believe that this is the case for coins in our world, and this can be verified empirically.

Example 17. Shuffle a deck of 52 cards. $\Omega = S_{52}$, the set of all permutations[3] of $[52]$ and $p_\pi = \frac{1}{52!}$ for each $\pi \in S_{52}$.

Example 18. "Psychic" guesses a deck of cards. The sample space is $\Omega = S_{52} \times S_{52}$ and $p_{(\pi,\sigma)} = 1/(52!)^2$ for each pair (π, σ) of permutations. In a pair (π, σ), the permutation π denotes the actual

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

order of cards in the shuffled deck, and σ denotes the order guessed by the psychic. If the guesses are purely random, then the probabilities are as we have written.

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 $A_k = \{(\pi, \sigma) : X(\pi, \sigma) \geq k\}$.

Example 19. Toss a coin till a head turns up. $\Omega = \{1, 01, 001, 0001, \ldots\} \cup \{\bar{0}\}$. Let us write $0^k 1 = 0 \ldots 01$ 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 $p_{0^k 1} = q^k p$ 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 $A = \{0^k 1 : k \geq n\}$ be the event that at least n tails fall before a head turns up. Then $\mathbf{P}(A) = q^n p + q^{n+1} p + \ldots = q^n$.

Example 20. Place r distinguishable balls in m distinguishable urns at random. We saw this before (the words "labelled" and "distinguishable" mean the same thing here). The sample space is $\Omega = [m]^r = \{\underline{\omega} = (\omega_1, \ldots, \omega_r) : 1 \leq \omega_i \leq m\}$ and $p_{\underline{\omega}} = m^{-r}$ for every $\underline{\omega} \in \Omega$. Here ω_i indicates the urn number into which the i^{th} ball goes.

Example 21. Place r indistinguishable balls in m distinguishable urns at random. Since the balls are indistinguishable, we can only count the number of balls in each urn. The sample space is

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

We give two proposals for the elementary probabilities.

(1) Let $p^{\text{MB}}_{(\ell_1, \ldots, \ell_m)} = \frac{m!}{\ell_1! \ell_2! \ldots \ell_m!} \frac{1}{m^r}$. These are the probabilities that result if we place r labelled balls in m labelled urns, and then erase the labels on the balls.

(2) Let $p^{\text{BE}}_{(\ell_1, \ldots, \ell_m)} = \frac{1}{\binom{m+r-1}{r-1}}$ 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 22. (1) Let $(\ell_1, \ldots, \ell_m) \in \Omega$. Show that $\#\{\underline{\omega} \in [m]^r : \sum_{j=1}^r \mathbf{1}_{\omega_j = i} = \ell_i \text{ for each } i \in [m]\} = \frac{n!}{\ell_1! \ell_2! \ldots \ell_m!}$. Hence or directly, show that $\sum_{\omega \in \Omega} p^{MB}_\omega = 1$.

(2) Show that $\#\Omega = \binom{m+r-1}{r-1}$. 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 23. 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.

Example 24. Sampling without replacement from a population. Now we take $\Omega = \{\underline{\omega} \in [N]^k : \omega_i$ are distinct ele with $p_{\underline{\omega}} = 1/N(N-1)\dots(N-k+1)$ for each $\underline{\omega} \in \Omega$.

Fix $m < N$ and define the random variable $X(\underline{\omega}) = \sum_{i=1}^{k} \mathbf{1}_{\omega_i \leq m}$. If the population $[N]$ contains a subset, say $[m]$, (could be the subset of people having a certain disease), then $X(\underline{\omega})$ counts the number of people in the sample who have the disease. Using X one can define events such as $A = \{\underline{\omega} : X(\underline{\omega}) = \ell\}$ for some $\ell \leq m$. If $\underline{\omega} \in A$, then ℓ of the ω_i must be in $[m]$ and the rest in $[N] \setminus [m]$. Hence

$$\#A = \binom{k}{\ell} m(m-1)\dots(m-\ell+1)(N-m)(N-m-1)\dots(N-m-(k-\ell)+1).$$

As the probabilities are equal for all sample points, we get

$$\begin{aligned}\mathbf{P}(A) &= \frac{\binom{k}{\ell} m(m-1)\dots(m-\ell+1)(N-m)(N-m-1)\dots(N-m-(k-\ell)+1)}{N(N-1)\dots(N-k+1)} \\ &= \frac{1}{\binom{N}{k}} \binom{m}{\ell} \binom{N-m}{k-\ell}.\end{aligned}$$

This expression arises whenever the population is subdivided into two parts and we count the number of samples that fall in one of the sub-populations.

[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 $\Omega_{\text{FD}} = \{(\ell_1, \dots, \ell_m) : \ell_i = 0 \text{ or } 1 \text{ and } \ell_1 + \dots + \ell_m = r\}$ 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 25. 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 event of interest. This is far from the case. The following example is an illustration.

Example 26. 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 $\mathbf{P}(A)$ is a very difficult problem. When n is large and $m = 2n$, it is expected that $\mathbf{P}(A)$ converges to a specific number, but proving it is an open problem as of today![5]

We now give two non-examples.

Example 27. 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 $\sum_{i \in \mathbb{N}} p_i = 0$

[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!

whereas if $p > 0$, then $\sum_{i \in \mathbb{N}} p_i = \infty$. This 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 28. A non-example. 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 $\mathbf{P}(A)$ 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 29. 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 $\mathbf{P}_n(A \cap \Omega_n) = \#(A \cap [n])/n$. 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.

4. Countable and uncountable

Definition 30. 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 31. 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 32. 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 33. 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 $F : \mathbb{N} \times \mathbb{N} \to \mathbb{Z} \times \mathbb{Z}$ by composing them, i.e., $F(n,m) = f(g(n))$. Then, F is one-one and onto. This shows that $\mathbb{Z} \times \mathbb{Z}$ is indeed countable.

Example 34. 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 35. 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.

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

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 36. 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 37 (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 $h : A \to B$.*

We omit the proof as it is irrelevant to the rest of the course[7].

5. On infinite sums

There were some subtleties in the definition of probabilities which we address now. The definition of $\mathbf{P}(A)$ for an event A and $\mathbf{E}[X]$ 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 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.

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 $x_n = f(\omega_1) + f(\omega_2) + \ldots + f(\omega_n)$. Since f is non-negative, these numbers are non-decreasing, i.e., $x_1 \le x_2 \le x_3 \le \ldots$. Hence, they converge to a finite number or to $+\infty$ (which is just another phrase for saying that the partial sums grow without bound). We would like to simply define the sum $\sum_{\omega\in\Omega} f(\omega)$ as the limit $L = \lim_{n\to\infty}(f(\omega_1) + \ldots + f(\omega_n)$, which may be finite or $+\infty$.

The problem is that this may depend on the bijection Ω 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 $L' = \lim_{n\to\infty}(f(\omega'_1) + \ldots + f(\omega'_n)$.

[7]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

(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.

The injectivity of f and g is used in checking that these are the only possibilities. 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, use the blue arrows, and if the first element is in B use the red arrows (but in reverse direction) to define the function h. Check that the resulting function is a bijection!

Is it necessarily true that $L = L'$?

Case I - Non-negative f: We claim that for any two bijections φ and ψ as above, the limits are the same (this means that the limits are $+\infty$ in both cases, or the same finite number in both cases). Indeed, fix any n and recall that $x_n = f(\omega_1) + \ldots + f(\omega_n)$. Now, ψ is surjective, hence there is some m (possibly very large) such that $\{\omega_1, \ldots, \omega_n\} \subseteq \{\omega'_1, \ldots, \omega'_m\}$. Now, we use the non-negativity of f to observe that

$$f(\omega_1) + \ldots + f(\omega_n) \leq f(\omega'_1) + \ldots + f(\omega'_m).$$

This is the same as $x_n \leq y_m$. Since y_k are non-decreasing, it follows that $x_n \leq y_m \leq y_{m+1} \leq y_{m+2} \ldots$, which implies that $x_n \leq L'$. Now let $n \to \infty$ and conclude that $L \leq L'$. Repeat the argument with the roles of φ and ψ reversed to conclude that $L' \leq L$. Hence $L = L'$, as desired to show.

In conclusion, for non-negative functions f, we can assign an unambiguous meaning to $\sum_\omega f(\omega)$ by setting it equal to $\lim_{n\to\infty}(f(\varphi(1) + \ldots + f(\varphi(n)))$, where $\varphi : \mathbb{N} \to \Omega$ is any bijection (the point being that the limit does not depend on the bijection chosen), and the limit here may be allowed to be $+\infty$ (in which case we say that the sum does not converge).

Case II - General $f : \Omega \to \mathbb{R}$: The above argument fails if f is allowed to take both positive and negative values (why?). In fact, the answers L and L' from different bijections may be completely different. An example is given later to illustrate this point. For now, here is how we deal with this problem.

For a real number x we introduce the notations, $x_+ = x \vee 0$ and $x_- = (-x) \vee 0$. Then $x = x_+ - x_-$ while $|x| = x_+ + x_-$. Define the non-negative functions $f_+, f_- : \Omega \to \mathbb{R}_+$ by $f_+(\omega) = (f(\omega))_+$ and $f_-(\omega) = (f(\omega))_-$. Observe that $f_+(\omega) - f_-(\omega) = f(\omega)$ while $f_+(\omega) + f_-(\omega) = |f(\omega)|$, for all $\omega \in \Omega$.

Example 38. Let $\Omega = \{a, b, c, d\}$ and let $f(a) = 1$, $f(b) = -1$, $f(c) = -3$, $f(4) = -0.3$. Then, $f_+(a) = 1$ and $f_+(b) = f_+(c) = f_+(d) = 0$ while $f_-(1) = 0$ and $f_-(b) = 1$, $f_-(c) = 3$, $f_-(d) = 0.3$.

Since f_+ and f_- are non-negative functions, we know how to define their sums. Let $S_+ = \sum_\omega f_+(\omega)$ and $S_- = \sum_\omega f_-(\omega)$. Recall that one or both of S_+, S_- could be equal to $+\infty$, in which case we say that $\sum_\omega f(\omega)$ *does not converge absolutely* and do not assign it any value. If both S_+ and S_- are finite, then we define $\sum_\omega f(\omega) = S_+ - S_-$. In this case we say that $\sum f$ *converges absolutely*.

This completes our definition of absolutely convergent sums. A few exercises to show that when working with absolutely convergent sums, the usual rules of addition remain valid. For example, we can add the numbers in any order.

Exercise 39. Show that $\sum_\omega f(\omega)$ converges absolutely if and only if $\sum_\omega |f(\omega)|$ is finite (since $|f(\omega)|$ is a non-negative function, this latter sum is always defined, and may equal $+\infty$).

For non-negative f, we can find the sum by using any particular bijection and then taking limits of partial sums. What about general f?

Exercise 40. Let $f : \Omega \to \mathbb{R}$. Suppose $\sum_{\omega\in\Omega} f(\omega)$ be summable and let the sum be S. Then, for any bijection $\varphi : \mathbb{N} \to \Omega$, we have $\lim_{n\to\infty}(f(\varphi(1)) + \ldots + f(\varphi(n))) = S$.

Conversely, if $\lim_{n\to\infty}(f(\varphi(1)) + \ldots + f(\varphi(n)))$ exists and is the same finite number for any bijection $\varphi : \mathbb{N} \to \mathbb{R}$, then f must be absolutely summable and $\sum_{\omega\in\Omega} f(\omega)$ is equal to this common limit.

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

Exercise 41. 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 $\sum f \le \sum g$.

Example 42. 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

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

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 43. What if Ω is uncountable? Take any $f : \Omega \to \mathbb{R}_+$. Define the sets $A_n = \{\omega : f(\omega) \ge 1/n\}$. 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!

5.2. **Second approach.** In the first approach, we assumed that you are already familiar with the notion of limits and series and used them to define countable sums. In the second approach, we start from scratch and define infinite sums. The end result is exactly the same. For the purposes of this course, you may ignore the rest of the section.

Definition 44. If Ω is a countable set and $f : \Omega \to \mathbb{R}_+$ is a non-negative function, then we define

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

where the supremum takes values in $\bar{\mathbb{R}}_+ = \mathbb{R}_+ \cup \{+\infty\}$. We say that $\sum f(\omega)$ converges if the supremum has a finite value.

Exercise 45. Show that if $f, g : \Omega \to \mathbb{R}_+$ and $a, b \in \mathbb{R}_+$, then $\sum(af + bg) = a \sum f + b \sum g$. Further, if $f(\omega) \leq g(\omega)$ for all $\omega \in \Omega$, then $\sum f \leq \sum g$.

Next, we would like to remove the condition of non-negativity. For a real number x we write $x_+ = x \vee 0$ and $x_- = (-x) \vee 0$. Then $x = x_+ - x_-$ while $|x| = x_+ + x_-$.

Definition 46. Now suppose $f : \Omega \to \mathbb{R}$ takes both positive and negative values. Then we first define the non-negative functions $f_+, f_- : \Omega \to \mathbb{R}_+$ by $f_+(\omega) = (f(\omega))_+$ and $f_-(\omega) = (f(\omega))_-$ and set $S_+ = \sum_\omega f_+(\omega)$ and $S_- = \sum_\omega f_-(\omega)$. If both S_+ and S_- are finite, then we define $\sum_\omega f(\omega) = S_+ - S_-$.

Remark 47. The condition that S_+ and S_- are both finite is the same as the condition that $\sum_\omega |f(\omega)|$ is finite. If these happen, we say that the sum $\sum f(\omega)$ *converges absolutely*.

Remark 48. Sometimes it is convenient to set $\sum f(\omega)$ to $+\infty$ if $S_+ = \infty$ and $S_- < \infty$ and set $\sum f(\omega)$ to $-\infty$ if $S_+ < \infty$ and $S_- = \infty$. But there is no reasonable value to assign if both the sums are infinite.

Exercise 49. Show that the two approaches give the same answers.

6. 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 defined 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 by various rules, such as substitution rule, integration by parts etc. In probability we need some similar rules relating probabilities of various combinations of events to the individual probabilities.

Proposition 50. *Let $(\Omega, p.)$ be a discrete probability space.*

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

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

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

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

Now we prove the proposition.

Proof. (1) By definition of probability space $\mathbf{P}(\Omega) = 1$ and $\mathbf{P}(\emptyset) = 0$. 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 41, 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 $\mathbf{P}(\emptyset)$, $\mathbf{P}(A)$ and $\mathbf{P}(\Omega)$, respectively. Thus, $0 \leq \mathbf{P}(A) \leq 1$.

(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 41, 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 $\mathbf{P}(A)$, $\mathbf{P}(B)$ and $\mathbf{P}(A \cup B)$. Thus, we get $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B)$.

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

Exercise 51. Adapt the proof to prove that for a countable family of events A_k in a common probability space (no disjointness assumed), we have

$$\mathbf{P}(\cup_k A_k) \leq \sum_k \mathbf{P}(A_k).$$

Definition 52 (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 53. 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 $\mathbf{P}(A)$ is the increasing limit of $\mathbf{P}(A_k)$.

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

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

The basic rules of probability:

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

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

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

7. Inclusion-exclusion formula

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

However, it is easy to see that $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$. 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 54 (Inclusion-Exclusion formula)**.** *Let (Ω, p) be a probability space and let $A_1, \ldots, A_n$ be events. Then,*

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

where

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

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 $\mathbf{P}(A_1)$ and once for $\mathbf{P}(A_2)$) and subtracted once in S_2 (in $\mathbf{P}(A_1 \cap A_2)$). 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

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

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 55. 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 $\mathbf{P}(A_{i_1} \cap A_{i_2} \ldots \cap A_{i_k}) = \frac{1}{52(52-1)\ldots(52-k+1)}$ for any $i_1 < i_2 < \ldots < i_k$ (why?). Therefore, by the inclusion-exclusion formula, we get

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

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

Example 56. Place n distinguishable balls in r distinguishable urns at random. Let A be the event that some urn is empty. The probability space is $\Omega = \{\underline{\omega} = (\omega_1, \ldots, \omega_n) : 1 \leq \omega_i \leq r\}$ with $p_{\underline{\omega}} = r^{-n}$. Let $A_\ell = \{\underline{\omega} : \omega_i \neq \ell\}$ for $\ell = 1, 2 \ldots, r$. Then, $A = A_1 \cup \ldots \cup A_{r-1}$ (as A_r is empty, we could include it or not, makes no difference).

It is easy to see that $\mathbf{P}(A_{i_1} \cap \ldots \cap A_{i_k}) = (r-k)^n r^{-n} = (1 - \frac{k}{r})^n$. We could use the inclusion-exclusion formula to write the expression

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

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

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

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

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

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

Exercise 57. Let $A_1, \ldots, A_n$ be events in a probability space (Ω, p) and let $m \leq n$. Let B_m and C_m be as above. Show that

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

Exercise 58. Return to the setting of exercise 55 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 $\mathbf{P}(A_m) \to e^{-1} \frac{1}{m!}$ 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.

8. 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 56 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 $\mathbf{P}(A_1 \cup \ldots \cup A_n) \leq \mathbf{P}(A_1) + \ldots + \mathbf{P}(A_n)$ 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 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 59 (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 $\mathbf{P}(A)$.*

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

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

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

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

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

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

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 $y > x$. ■

Here is an application of these inequalities.

Example 60. Return to Example 56. 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.

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

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

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

Example 62. 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 $\mathbf{P}(A) = 1/2$, $\mathbf{P}(B) = 1/2$ (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 $\mathbf{P}(A \cap B) = 1/4$. Thus, $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$ 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[8] And this intuition was 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).

[8]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 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.

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.

10. Conditional probability and independence

Definition 63. Let A, B be two events in the same probability space.

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

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

(2) We say that A and B are *independent* if $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$. If $\mathbf{P}(B) \neq 0$, then A and B are independent if and only if $\mathbf{P}(A \,\big|\, B) = \mathbf{P}(A)$ (and similarly with the roles of A and B reversed). If $\mathbf{P}(B) = 0$, then A and B are necessarily independent since $\mathbf{P}(A \cap B)$ 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 64. 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 $B = \{(i,j) : j \text{ is } 1 \text{ or } 6\}$ 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

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

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

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

Thus, $\mathbf{P}(A \cap C) \neq \mathbf{P}(A)\mathbf{P}(C)$ 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 $\mathbf{P}(C \bigm| A) = \frac{1}{9}$.

Example 65. 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 $\mathbf{P}(A) = \mathbf{P}(B) = \frac{1}{13}$. However, $\mathbf{P}(B \bigm| A) = \frac{3}{51}$. One can also see that $\mathbf{P}(B \bigm| A^c) = \frac{4}{51}$.

In words, A (respectively B) could be the event that the first (respectively second) card is an ace. Then $\mathbf{P}(B) = 4/52$ 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 $\mathbf{P}(B \bigm| A^c)$ and if the first card was an ace, we update it to $\mathbf{P}(B \bigm| A)$.

Caution: Independence should not be confused with disjointness! If A and B are disjoint, $\mathbf{P}(A \cap B) = 0$ and hence A and B can be independent if and only if one of $\mathbf{P}(A)$ or $\mathbf{P}(B)$ 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 66. 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 $\mathbf{P}(A_i) > 0$ 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.

Let B be any other event.

(1) (Total probability rule). $\mathbf{P}(B) = \mathbf{P}(A_1)\mathbf{P}(B \bigm| A_1) + \ldots + \mathbf{P}(A_n)\mathbf{P}(B \bigm| A_n)$.

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

$$\mathbf{P}(A_k \bigm| B) = \frac{\mathbf{P}(A_k)\mathbf{P}(B \bigm| A_k)}{\mathbf{P}(A_1)\mathbf{P}(B \bigm| A_1) + \ldots + \mathbf{P}(A_n)\mathbf{P}(B \bigm| A_n)}.$$

Proof. The proof is merely by following the definition.

(1) The right hand side is equal to

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

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

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

$$\begin{aligned}\mathbf{P}(A_1 \,\big|\, B) &= \frac{\mathbf{P}(A_1 \cap B)}{\mathbf{P}(B)} \\ &= \frac{\mathbf{P}(A_1)\mathbf{P}(B \,\big|\, A_1)}{\mathbf{P}(A_1)\mathbf{P}(B \,\big|\, A_1) + \ldots + \mathbf{P}(A_n)\mathbf{P}(B \,\big|\, A_n)}\end{aligned}$$

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

Exercise 67. Suppose A_i are events such that $\mathbf{P}(A_1 \cap \ldots \cap A_n) > 0$. Then show that

$$\mathbf{P}(A_1 \cap \ldots \cap A_n) = \mathbf{P}(A_1)\mathbf{P}(A_2 \,\big|\, A_1)\mathbf{P}(A_3 \,\big|\, A_1 \cap A_2) \ldots \mathbf{P}(A_n \,\big|\, A_1 \cap \ldots \cap A_{n-1}).$$

Example 68. 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) $\mathbf{P}(A) = 10^{-6}$. Of course $\mathbf{P}(A^c) = 1 - 10^{-6}$.

(2) $\mathbf{P}(B \,\big|\, A) = 0.99$ and $\mathbf{P}(B \,\big|\, A^c) = 0.01$.

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

$$\mathbf{P}(A \,\big|\, B) = \frac{\mathbf{P}(B \,\big|\, A)\mathbf{P}(A)}{\mathbf{P}(B \,\big|\, A)\mathbf{P}(A) + \mathbf{P}(B \,\big|\, A^c)\mathbf{P}(A^c)} = \frac{0.99 \times 10^{-6}}{0.99 \times 10^{-6} + 0.01 \times (1 - 10^{-6})} = 0.000099.$$

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: A person X is introverted, very systematic in thinking and somewhat absent-minded. You are told that he is a doctor or a mathematician. What would be your 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 $\mathbf{P}(A\big|B)$ and $\mathbf{P}(B\big|A)$.

11. Independence of three or more events

Definition 69. Events $A_1, \ldots, A_n$ in a common probability space are said to be independent if $\mathbf{P}\left(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_m}\right) = \mathbf{P}(A_{i_1})\mathbf{P}(A_{i_2}) \ldots \mathbf{P}(A_{i_m})$ 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 70. 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 out of the first two tosses are both heads or both tails. Then $\mathbf{P}(A) = \mathbf{P}(B) = \mathbf{P}(C) = \frac{1}{4}$. Further,

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

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

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 is is right that the definition should not declare them to be independent.

Exercise 71. 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

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

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.

12. Discrete probability distributions

Let (Ω, p) be a probability space and $X : \Omega \to \mathbb{R}$ be a random variable. We define two objects associated to X.

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

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

One obvious property is that $\sum_{t\in\mathbb{R}} f_X(t) = 1$. Conversely, any non-negative function f that is non-zero on a countable set S and such that $\sum_{t\in\mathbb{R}} f(t) = 1$ is a pmf of some random variable.

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

$$F_X(t) = \mathbf{P}\{\omega : X(\omega) \le t\}.$$

Example 72. 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, $\text{Range}(X) = \{2, 3, \ldots, 12\}$. The pmf and CDF of X are given by

$$f_X(k) = \begin{cases} 1/36 & \text{if } k = 2. \\ 2/36 & \text{if } k = 3. \\ 3/36 & \text{if } k = 4. \\ 4/36 & \text{if } k = 5. \\ 5/36 & \text{if } k = 6. \\ 6/36 & \text{if } k = 7. \\ 5/36 & \text{if } k = 8. \\ 4/36 & \text{if } k = 9. \\ 3/36 & \text{if } k = 10. \\ 2/36 & \text{if } k = 11. \\ 1/36 & \text{if } k = 12. \end{cases} \qquad F_X(t) = \begin{cases} 0 & \text{if } t < 2. \\ 1/36 & \text{if } t \in [2,3). \\ 3/36 & \text{if } t \in [3,4). \\ 6/36 & \text{if } t \in [4,5). \\ 10/36 & \text{if } t \in [5,6). \\ 15/36 & \text{if } t \in [6,7). \\ 21/36 & \text{if } t \in [7,8). \\ 26/36 & \text{if } t \in [8,9). \\ 30/36 & \text{if } t \in [9,10). \\ 33/36 & \text{if } t \in [10,11). \\ 35/36 & \text{if } t \in [11,12). \\ 1 & \text{if } t \ge 12. \end{cases}$$

A picture of the pmf and CDF for a Binomial distribution are shown in Figure **??**.

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 $F(t) = \mathbf{P}(A) \le \mathbf{P}(B) = F(s)$. 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, $F(n) = \mathbf{P}(A_n) \to \mathbf{P}(\Omega) = 1$ 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. ■

Remark 73. 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.

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.

Example 74. Let f and F be the pmf, CDF pair

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

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 $X(\omega) = \mathbf{1}_{\omega\le 3}$, then $X \sim \text{Ber}(0.3)$. Any random variable taking only the values 0 and 1, has Bernoulli distribution.

Example 75. 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 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 $X \sim$ Bin(n,p). In words, the number of heads in n tosses of a p-coin has Bin(n,p) distribution.

Example 76. Fix $p \in (0,1]$ and let $f(k) = q^{k-1}p$ for $k \in \mathbb{N}_+$. This is called the *Geometric distribution* with parameter p and is denoted Geo(p). The CDF is

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

For example, the number of tosses of a p-coin till the first head turns up, is a random variable with Geo(p) distribution.

Example 77. 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).

In the problem of a psychic (randomly) guessing the cards in a deck, we have seen that the number of matches (correct guesses) had an *approximately* Pois(1) distribution.

Example 78. Fix positive integers b, w and $m \leq b + w$. 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 $\max\{m - w, 0\} \leq k \leq b$). This is called the *Hypergeometric distribution* with parameters b, w, m and we shall denote it by Hypergeo(b,w,m).

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.

Computing expectations from the pmf Let X be a random variable on (Ω, p) with pmf f. Then we claim that

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

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

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

Similarly, $\mathbf{E}[X^2] = \sum_k x_k^2 f(x_k)$. More generally, if $h : \mathbb{R} \to \mathbb{R}$ is any function, then the random variable $h(X)$ has expectation $\mathbf{E}[h(X)] = \sum_k h(x_k) f(x_k)$. Although this sounds trivial, there is a very useful point here. To calculate $\mathbf{E}[X^2]$ we do not have to compute the pmf of X^2 first, which

can be done but would be more complicated. Instead, in the above formulas, $\mathbf{E}[h(X)]$ has been computed directly in terms of the pmf of X.

Exercise 79. Find $\mathbf{E}[X]$ and $\mathbf{E}[X^2]$ in each case.

(1) $X \sim \text{Bin}(n, p)$.

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

(3) $X \sim \text{Pois}(\lambda)$.

(4) $X \sim \text{Hypergeo}(b, w, m)$.

13. General probability distributions

We take the first three of the four properties of CDF proved in the previous section as the *definition* of a CDF or distribution function, in general.

Definition 80. A (cumulative) distribution function (or CDF for short) is any function $F : \mathbb{R} \to [0, 1]$ be a non-decreasing, right continuous function such that $F(t) \to 0$ as $t \to -\infty$ and $F(t) \to 1$ as $t \to +\infty$.

If (Ω, p) is a discrete probability space and $X : \Omega \mapsto \mathbb{R}$ is any random variable, then the function $F(t) = \mathbf{P}\{\omega : X(\omega) \leq t\}$ is a CDF, as discussed in the previous section. However, there are distribution functions that do not arise in this manner.

Example 81. 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.

There are two ways to rectify this issue.

(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 measure theory. 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 CDFs such as in the above example as reasonable approximations to CDFs of some discrete random variables. For example, if $\Omega = \{\omega_0, \omega_1, \ldots, \omega_N\}$ and $p(\omega_k) = 1/(N+1)$ for all $0 \leq k \leq N$, and $X : \Omega \mapsto \mathbb{R}$ is defined by $X(\omega_k) = k/n$, then it is easy to

check that the CDF of X is the function G given by

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

Now, if N is very large, then the function G looks approximately like the function F. 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.

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.

14. Uncountable probability spaces - conceptual difficulties

The following two "random experiments" are easy to imagine, but difficult to fit into the framework of probability spaces[9].

(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 $\mathbf{P}(A)$ for any A? For example, if $A = \{\underline{\omega} : \omega_1 = 0, \omega_2 = 0, \omega_3 = 1\}$, then everyone agrees that $\mathbf{P}(A)$ "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.

(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 $\mathbf{P}(A)$ "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.

[9]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

Example 82. In the example of infinite sequence of tosses, $\Omega = \{0,1\}^{\mathbb{N}}$. Any set of the form $A = \{\underline{\omega} : \omega_1 = \epsilon_1, \ldots, \omega_k = \epsilon_k\}$ where $k \geq 1$ and $\epsilon_i \in \{0,1\}$ will be called a basic set and its probability is defined to be $\mathbf{P}(A) = \prod_{j=1}^{k} p^{\epsilon_j} q^{1-\epsilon_j}$ 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 $\mathbf{P}(B)$ should be $\sum_k \mathbf{P}(A_k) = \sum_k q^{k-1}p$ which is in fact equal to 1.

Example 83. 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 $\mathbf{P}(a,b) = b - a$. 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 $\mathbf{P}(B)$ must be equal to $\lim_{k\to\infty} \mathbf{P}(A_k) = \lim_{k\to\infty}(b-a-(2/k)) = b-a$. Another example could be $C = [0.1, 0.2) \cup (0.3, 0.7]$. Similarly argue that $\mathbf{P}(\{x\}) = 0$ 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 84. Let $\Omega = [0,1]$ and let intervals (a,b) be open sets with their probabilities defined as $\mathbf{P}(a,b) = \sqrt{b-a}$. This quickly leads to problems. For example, $\mathbf{P}(0,1) = 1$ by definition. But $(0,1) = (0,0.5) \cup (0.5,1) \cup \{1/2\}$ from which the rules of probability would imply that $\mathbf{P}(0,1)$ must be at least $\mathbf{P}(0,1/2) + \mathbf{P}(1/2,1) = \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}} = \sqrt{2}$ which is greater than 1. Inconsistency!

Exercise 85. Show that we run into inconsistencies if we define $\mathbf{P}(a,b) = (b-a)^2$ 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 86. *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 $\mathbf{P}\{(a,b]\} = F(b) - F(a)$. 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 $F(t) = \mathbf{P}\{X \leq t\}$ 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 $\mathbf{P}\{a < X \leq b\} := F(b) - F(a)$.

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

(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 $\mathbf{P}\{X \in A\}$ 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 $\mathbf{P}\{X \in A\}$.

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 87. The above method starts from a CDF F and defines $\mathbf{P}\{X \in A\}$ 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 88. Let X be a random variable with distribution F. Use the working rules to find the following probabilities.

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

(2) Show that $\mathbf{P}\{X = a\} = F(a) - F(a-)$. 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 89. Let F be the CDF defined in example 81. We calculate $\mathbf{P}\{X \in A\}$ 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 $\mathbf{P}\{X \in A\} \leq \sum_n F(r_n + \delta 2^{-n}) - F(r_n) \leq 2\delta$. Since this is true for every $\delta > 0$, we must have $\mathbf{P}\{X \in A\} = 0$. (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[10] How to find $\mathbf{P}\{X \in A\}$? 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, $\mathbf{P}\{X \in A\} \leq \mathbf{P}\{X \in A_n\} = 2^n 3^{-n}$ which goes to 0 as $n \to \infty$. Hence, $\mathbf{P}\{X \in A\} = 0$.

15. 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 for practical purposes they may be ignored (for now).

(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 piecewise 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

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

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 90. 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 $\mathbf{P}\{X = t\} = 0$. Therefore $f(t)$ cannot be interpreted as $\mathbf{P}\{X = t\}$ (in fact, pdf can take values greater than 1, so it cannot be a probability!).

[10] 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.

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 91. 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 92. 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 93. 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

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

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 94. *Fix $\mu \in \mathbb{R}$ and $\sigma > 0$ and let $\varphi(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(t-\mu)^2}$. 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 $I = 1$. ■

Example 95. 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

$$\int_0^{\infty} \lambda^\nu e^{-\lambda t}t^{\nu-1}dt = \int_0^{\infty} e^{-u}u^{\nu-1}d\nu = \Gamma(\nu).$$

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 $\mathbf{P}(N \ge \nu)$ where $N \sim \text{Pois}(\lambda t)$. 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. [11]

[11] **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

Example 96. 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 -

$$\text{PDF: } f(t) = \begin{cases} \frac{1}{B(\alpha,\beta)} t^{\alpha-1}(1-t)^{\beta-1} & \text{if } t \in (0,1) \\ 0 & \text{otherwise} \end{cases} \qquad \text{CDF: } F(t) = \begin{cases} 0 & \text{if } t \leq 0 \\ \int_0^t f(u)du & \text{if } t \in (0,1) \\ 0 & \text{if } t \geq 1. \end{cases}$$

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 97. *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.

---

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,

$$n! = \lim_{\ell\to\infty} \frac{\ell!\,\ell^n}{(n+1)\dots(n+\ell)}.$$

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 to check that this agrees with the definition by the integral given earlier. In other words, for $\nu > -1$, we have

$$\lim_{\ell\to\infty} \frac{\ell!\,\ell^\nu}{(\nu+1)\dots(\nu+\ell)} = \int_0^\infty x^\nu e^{-x}dx.$$

Now for any other *positive integer* value of α and real $\beta > 0$ we can integrate by parts and get

$$\begin{aligned} B(\alpha,\beta) &= \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt \\ &= -\frac{1}{\beta}t^{\alpha-1}(1-t)^{\beta}\Big|_0^1 + \frac{\alpha-1}{\beta}\int_0^1 t^{\alpha-2}(1-t)^{\beta}dt \\ &= \frac{\alpha-1}{\beta}B(\alpha-1,\beta+1). \end{aligned}$$

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 98. 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

$$f(t) = \frac{\lambda}{\pi(\lambda^2+(t-a)^2)} \qquad F(t) = \frac{1}{2} + \frac{1}{\pi}\tan^{-1}\left(\frac{t-a}{\lambda}\right).$$

Remark 99. 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 $F(t) = \int_{-\infty}^t f(u)du$.

16. 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 100. 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

$$\mathbf{P}\{X = 1\} = \mathbf{P}\{U \le 0.4\} = 0.4, \qquad \mathbf{P}\{X = 0\} = \mathbf{P}\{U > 0.4\} = 0.6.$$

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

$$\mathbf{P}\{X = t_k\} = \mathbf{P}\left\{\sum_{j=1}^{k-1} f(t_j) < U \leq \sum_{j=1}^{k} f(t_j)\right\} = f(t_k).$$

Thus X has pmf f.

Exercise 101. 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 102. Suppose we want to draw from the Unif($[3, 7]$) distribution. Set $X = 4U + 3$. Clearly

$$\mathbf{P}\{X \leq t\} = \mathbf{P}\{U \leq \frac{t-3}{4}\} = \begin{cases} 0 & \text{if } t < 0 \\ (t-3)/4 & \text{if } 3 \leq t \leq 7 \\ 1 & \text{if } t > 7 \end{cases}.$$

This is precisely the CDF of Unif($[3, 7]$) distribution.

Example 103. 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,

$$F(t) := \mathbf{P}\{X \leq t\} = \mathbf{P}\{U \leq t^{1/3}\} = \begin{cases} 0 & \text{if } t < 0 \\ t^{1/3} & \text{if } 0 \leq t \leq 1 \\ 1 & \text{if } t > 1 \end{cases}.$$

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 $X \sim$ 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 104. *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,

$$F(t) = \mathbf{P}\{X \le t\} = \mathbf{P}\{U \le \psi^{-1}(t)\} = \begin{cases} 0 & \text{if } t \le a \\ \psi^{-1}(t) & \text{if } a < t < b \\ 1 & \text{if } t \ge b. \end{cases}$$

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 105. Consider the Exponential distribution with parameter λ whose CDF is

$$F(t) = \begin{cases} 0 & \text{if } t \le 0 \\ 1 - e^{-\lambda t} \text{ if } t > 0 \end{cases}$$

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 106. 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

$$\begin{matrix} t_0 & t_1 & t_2 & \ldots & \ldots & t_{1000} \\ w_0 & w_1 & w_2 & \ldots & \ldots & w_{1000} \end{matrix}.$$

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 107. 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 108. 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. For example,

17. 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 $\mathbf{P}(A)$ and $\mathbf{P}(B)$ is insufficient to calculate $\mathbf{P}(A \cap B)$ or $\mathbf{P}(A \cup B)$ etc.

Definition 109 (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

$$F(t_1, \ldots, t_m) = \mathbf{P}\{X_1 \leq t_1, \ldots, X_m \leq t_m\} = \mathbf{P}\left\{\bigcap_{i=1}^{m} \{X_i \leq t_i\}\right\}.$$

.

Example 110. 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

$$F(s,t) = \begin{cases} 0 & \text{if } s < 0 \text{ or } t < 0 \\ \mathbf{P}(A^c \cap B^c) & \text{if } s \geq 0,\ t < 1 \text{ or } t \geq 0,\ s < 1 \\ \mathbf{P}(A) & \text{if } 0 \leq s < 1 \text{ and } t \geq 1 \\ \mathbf{P}(B) & \text{if } 0 \leq t < 1 \text{ and } s \geq 1 \\ \mathbf{P}(A \cap B) & \text{if } s \geq 1, t \geq 1 \end{cases}$$

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 $s_1 \leq t_1, \ldots, s_m \leq t_m$, then $F(s_1, \ldots, s_m) \leq F(t_1, \ldots, t_m)$.

(2) $\lim F(t_1, \ldots, t_m) = 0$ if $\max\{t_1, \ldots, t_m\} \to -\infty$ (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 $F_1(t) := F(t, +\infty, \ldots, +\infty) := \lim F(t, s_2, \ldots, s_m)$ as $s_i \to +\infty$ for each $i = 2, \ldots, m$. This is true because if $A_n := \{X_1 \leq$

$t\} \cap \{X_2 \leq n\} \cap \ldots \cap \{X_m \leq n\}$, then as $n \to \infty$, the events A_n increase to the event $A = \{X_1 \leq t\}$. Hence $\mathbf{P}(A_n) \to \mathbf{P}(A)$. But $\mathbf{P}(A_n) = F(t, n, n, \ldots, n)$ and $\mathbf{P}(A) = F_1(t)$. 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

$$F(\mathbf{t}) = \sum_{i \,:\, \mathbf{t}_i \leq \mathbf{t}} w_i$$

where $\mathbf{s} \leq \mathbf{t}$ means that each co-ordinate of s is less than or equal to the corresponding co-ordinate of $\mathbf{t}$.

(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

$$F(\mathbf{t}) = \int_{-\infty}^{t_1} \ldots \int_{-\infty}^{t_m} f(u_1, \ldots, u_m) du_1 \ldots du_m.$$

We give two examples, one of each kind.

Example 111. (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[12]

In this case, the marginal distribution of X_k is 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

[12] 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...

to prove, but even easier to see from the balls in bins interpretation (just think of the last $n-\ell$ bins as one).

Example 112. (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

$$\int_{-\infty}^{+\infty} e^{-\frac{\tau}{2}(x-a)^2} dx = \sqrt{2\pi}\frac{1}{\sqrt{\tau}} \text{ for any } \tau > 0 \text{ and any } a \in \mathbb{R}. \tag{1}$$

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

$$ax^2 + by^2 + 2cxy = b\left(y + \frac{c}{b}\right)^2 + \left(a - \frac{c^2}{b}\right)x^2.$$

Therefore,

$$\begin{aligned}\int_{-\infty}^{\infty} e^{-\frac{1}{2}\left[ax^2+by^2+2cxy\right]} dy &= e^{-\frac{1}{2}(a-\frac{c^2}{b})x^2} \int_{-\infty}^{\infty} e^{-\frac{b}{2}(y+\frac{c}{b})^2} \\ &= e^{-\frac{1}{2}(a-\frac{c^2}{b})x^2} \frac{\sqrt{2\pi}}{\sqrt{b}}\end{aligned}$$

by (1) but ony if $b > 0$. Now we integrate over x and use (1) again (and the fact that $a - \frac{c^2}{b} > 0$) to get

$$\begin{aligned}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2}\left[a(x-\mu)^2+b(y-\nu)^2+2c(x-\mu)(y-\nu)\right]} dydx &= \frac{\sqrt{2\pi}}{\sqrt{b}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(a-\frac{c^2}{b})x^2} dx \\ &= \frac{\sqrt{2\pi}}{\sqrt{b}} \frac{\sqrt{2\pi}}{\sqrt{a-\frac{c^2}{b}}} = \frac{2\pi}{ab-c^2}.\end{aligned}$$

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.

Example 113. (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 $F_X(x_1, \ldots, x_m) = F_{X_1}(x_1) \ldots F_{X_m}(x_m)$.

18. 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 which is continuously differentiable (many exceptions can be made as remarked later).

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

$$g(\mathbf{y}) = f\left(T^{-1}\mathbf{y}\right) \Big| J[T^{-1}](\mathbf{y}) \Big|$$

where $J[T^{-1}](\mathbf{y})$ is the Jacobian determinant of the function T^{-1} at the point $\mathbf{y} = (y_1, \ldots, y_m)$.

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

$$g(\mathbf{y}) \approx \frac{1}{\text{area}(\mathcal{N})}\mathbf{P}\{\mathbf{Y} \in \mathcal{N}\}$$

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

$$\mathbf{P}\{\mathbf{Y} \in \mathcal{N}\} = \mathbf{P}\{\mathbf{X} \in T^{-1}(\mathcal{N})\}$$

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

$$\mathbf{P}\{\mathbf{X} \in T^{-1}(\mathcal{N})\} \approx \text{area}(T^{-1}(\mathcal{N}))f(\mathbf{x})$$

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 $\mathbb{R}^m$. It is sufficient if it is defined on the range of $\mathbf{X}$ (i.e., if $f(t_1,\ldots,t_m)=0$ for $(t_1,\ldots,t_m)=\mathbb{R}^m\setminus A$, then it is enough if T is defined on A.

(2) Even within the range of $\mathbf{X}$, we can allow T to be undefined, but $\mathbf{X}$ must have zero probability to fall in the set where it is undefined. For example, it can happen at finitely many points, or on a line (if $m\geq 2$) or on a plane (if $m\geq 3$) etc.

(3) Similarly, the differentiability of T is required only on a subset outside of which $\mathbf{X}$ has probability 0 of falling.

(4) One-one 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.

Example 114. Let X_1, X_2 be independent Exp(λ) random variables. Let $T(x_1,x_2)=(x_1+x_2, \frac{x_1}{x_1+x_2})$. This is well-defined on $\mathbb{R}^2_+$ (and note that $\mathbf{P}\{(X_1,X_2)\in\mathbb{R}^2_+\}=1$) and its range is $\mathbb{R}_+\times(0,1)$. The inverse function is $T^{-1}(y_1,y_2)=(y_1y_2, y_1(1-y_2))$. Its Jacobian determinant is

$$J[T^{-1}](y_1,y_2)=\det\begin{bmatrix} y_2 & y_1\\ 1-y_2 & -y_1\end{bmatrix}=-y_1.$$

(X_1,X_2) has density $f(x_1,x_2)=\lambda^2e^{-\lambda(x_1+x_2)}$ for $x_1,x_2>0$ (henceforth it will be a convention that the density is zero except where we specify it). Hence, the random variables $Y_1=X_1+X_2$ and $Y_2=\frac{X_1}{X_1+X_2}$ have joint density

$$g(y_1,y_2)=f(y_1y_2,y_1(1-y_2))|J[T^{-1}](y_1,y_2)|=\lambda^2e^{-\lambda(y_1y_2+y_1(1-y_2))}y_1=\lambda^2y_1e^{-\lambda y_1}$$

for $y_1>0$ and $y_2\in(0,1)$.

In particular, we see that $Y_1=X_1+X_2$ has density $h_1(t)=\int_0^1\lambda^2te^{-\lambda t}ds=\lambda^2te^{-\lambda t}$ (for $t>0$) which means that $Y_1\sim$ Gamma$(2,\lambda)$. Similarly, $Y_2=\frac{X_1}{X_1+X_2}$ has density $h_2(s)=\int_0^\infty\lambda^2te^{-\lambda t}dt=1$ (for $s\in(0,1)$) which means that Y_2 has Unif$(0,1)$ distribution. In fact, Y_1 and Y_2 are also independent since $g(u,v)=h_1(u)h_2(v)$.

Exercise 115. Let $X_1\sim$ Gamma(ν_1,λ) and $X_2\sim$ Gamma(ν_2,λ) (note that the shape parameter is the same) and assume that they are independent. Find the joint distribution of X_1+X_2 and $\frac{X_1}{X_1+X_2}$.

Example 116. 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

$$\begin{aligned} g(y_1, y_2) &= f(y_1 - y_2, y_2) \cdot 1 \\ &= \lambda^2 e^{-\lambda(y_1 - y_2)} e^{-\lambda y_2} \quad (\text{if } y_1 > y_2 > 0) \\ &= \lambda^2 e^{-\lambda y_1} \mathbf{1}_{y_1 > y_2 > 0}. \end{aligned}$$

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 117. Suppose $R \sim \text{Exp}(\lambda)$ and $\Theta \sim \text{Unif}(0, 2\pi)$ 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 118. 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).

19. Independence and conditioning of random variables

Definition 119. 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 $F_{\mathbf{X}}(t_1, \ldots, t_m) = F_1(t_1) \ldots F_m(t_m)$ for all $t_1, \ldots, t_m$.

Remark 120. Recalling the definition of independence of events, the equality $F_{\mathbf{X}}(t_1, \ldots, t_m) = F_1(t_1) \ldots F_m(t_m)$ is just saying that the events $\{X_1 \leq t_1\}, \ldots \{X_m \leq t_m\}$ are independent. More generally, it is true that $X_1, \ldots, X_m$ are independent if and only if $\{X_1 \in A_1\}, \ldots, \{X_m \in A_m\}$ are independent events for any $A_1, \ldots, A_m \subseteq \mathbb{R}$.

Remark 121. In case $X_1, \ldots, X_m$ have a joint pmf or a joint pdf (which we denote by $f(t_1, \ldots, t_m)$), the condition for independence is equivalent to

$$f(t_1, \ldots, t_m) = f_1(t_1) f_2(t_2) \ldots f_m(t_m)$$

where f_i is the marginal density (or pmf) of X_i. This fact can be derived from the definition easily. For example, in the case of densities, observe that

$$\begin{aligned} f(t_1,\dots,t_m) &= \frac{\partial^m}{\partial t_1 \dots \partial t_m} F(t_1,\dots,t_m) \quad \text{(true for any joint density)} \\ &= \frac{\partial^m}{\partial t_1 \dots \partial t_m} F_1(t_1)\dots F_m(t_m) \quad \text{(by independence)} \\ &= F_1'(t_1)\dots F_m'(t_m) \\ &= f_1(t_1)\dots f_m(t_m). \end{aligned}$$

When we turn it around, this gives us a quicker way to check independence.

Fact: Let $X_1,\dots,X_m$ be random variables with joint pdf $f(t_1,\dots,t_m)$. Suppose we can write this pdf as $f(t_1,\dots,t_m) = cg_1(t_1)g_2(t_2)\dots g_m(t_m)$ where c is a constant and g_i are some functions of one-variable. Then, $X_1,\dots,X_m$ are independent. Further, the marginal density of X_k is $c_k g_k(t)$ where $c_k = \frac{1}{\int_{-\infty}^{+\infty} g_k(s)ds}$. An analogous statement holds when $X_1,\dots,X_m$ have a joint pmf instead of pdf.

Example 122. 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 $X_1,\dots,X_n$ are independent. Indeed, the joint pmf of $X_1,\dots,X_n$ is

$$f(t_1,\dots,t_n) = p^{\sum t_k} q^{n-\sum t_k} \quad \text{where } t_i = 0 \text{ or } 1 \text{ for each } i \le n.$$

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 123. 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 124. 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 $\mathbf{P}\{(X, Y) \in [0.8, 1] \times [0.8, 1]\} = 0$ whereas $\mathbf{P}\{(X, Y) \in [0.8, 1]\}\mathbf{P}\{(X, Y) \in [0.8, 1]\} \neq 0$ (If X, Y were independent, we must have had $\mathbf{P}\{(X, Y) \in [a, b] \times [c, d]\} = \mathbf{P}\{X \in [a, b]\}\mathbf{P}\{Y \in [c, d]\}$ for any $a < b$ and $c < d$).

A very useful (and intuitively acceptable!) fact about independence is as follows.

Fact: Suppose $X_1, \ldots, X_n$ are independent random variables. Let $k_1 < k_2 < \ldots < k_m = n$. Let $Y_1 = h_1(X_1, \ldots, X_{k_1})$, $Y_2 = h_2(X_{k_1+1}, \ldots, X_{k_2}), \ldots Y_m = h_m(X_{k_{m-1}}, \ldots, X_{k_m})$. Then, $Y_1, \ldots, Y_m$ are also independent.

Remark 125. In the previous section we defined independence of events and now we have defined independence of random variables. How are they related? We leave it to you to check that events $A_1, \ldots, A_n$ are independent (according the definition of the previous section) if and only if the random variables $\mathbf{1}_{A_1}, \ldots, \mathbf{1}_{A_m}$ are independent (according the definition of this section)

The next part, about conditioning on random variables and conditional densities was not covered in class and is not included in syllabus.

Conditioning on random variables: Let $X_1, \ldots, X_{k+\ell}$ be random variables on a common probability space. Let $f(t_1, \ldots, t_{k+\ell})$ be the pmf of $(X_1, \ldots, X_{k+\ell})$ and let $g(t_1, \ldots, t_\ell)$ be the pmf of $(X_{k+1}, \ldots, X_{k+\ell})$ (of course we can compute g from f by summing over the first k indices). Then, for any $s_1, \ldots, s_\ell$ such that $\mathbf{P}\{X_{k+1} = s_1, \ldots X_m = s_\ell\} > 0$, we can define

(2)

$$h_{s_1,\ldots,s_\ell}(t_1, \ldots, t_k) = \mathbf{P}\{X_1 = t_1, \ldots, X_k = t_k \,\Big|\, X_{k+1} = s_1, \ldots X_m = s_\ell\} = \frac{f(t_1, \ldots, t_k, s_1, \ldots, s_\ell)}{g(s_1, \ldots, s_\ell)}.$$

It is easy to see that $h_{s_1,\ldots,s_\ell}(\cdot)$ is a pmf on $\mathbb{R}^k$. It is called the conditional pmf of $(X_1, \ldots, X_k)$ given that $X_{k+1} = s_1, \ldots X_m = s_\ell$.

Its interpretation is as follows. Originally we had random observables $X_1, \ldots, X_k$ which had a certain joint pmf. Then we observe the values of the random variables $X_{k+1}, \ldots, X_{k+\ell}$, say they turn out to be $s_1, \ldots, s_\ell$, respectively. Then we update the distribution (or pmf) of $X_1, \ldots, X_k$ according to the above recipe. The conditional pmf is the new function $h_{s_1,\ldots,s_\ell}(\cdot)$.

Exercise 126. 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

$$h_{s_1,\ldots,s_\ell}(t_1, \ldots, t_k) = \frac{f(t_1, \ldots, t_k, s_1, \ldots, s_\ell)}{g(s_1, \ldots, s_\ell)}. \tag{3}$$

This is well-defined whenever $g(s_1, \ldots, s_\ell) > 0$.

Remark 127. Note the difference between (2) and (3). In the latter we have left out the middle term because $\mathbf{P}\{X_{k+1} = s_1, \ldots, X_{k+\ell} = s_\ell\} = 0$. In (2) the definition of pmf comes from the definition of conditional probability of events but in (3) 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 128. 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 129. 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.

20. Mean and Variance

Let X be a random variable with distribution F. We shall assume that it has pmf or pdf denoted by f.

Definition 130. The *expected value* (also called *mean*) of X is defined as the quantity $\mathbf{E}[X] = \sum_t tf(t)$ if f is a pmf and $\mathbf{E}[X] = \int_{-\infty}^{+\infty} tf(t)dt$ if f is a pdf (provided the sum or the integral converges absolutely).

Note that this agrees with the definition we gave earlier for random variables with pmf. it is possible to define expect value for distributions without pmf or pdf, but we shall not do it here.

Properties of expectation: Let X, Y be random variables both having pmf f, g or pdf f, g, respectively.

(1) Then, $\mathbf{E}[aX + bY] = a\mathbf{E}[X] + b\mathbf{E}[Y]$ for any $a, b \in \mathbb{R}$. In particular, for a constant random variable (i.e., $X = a$ with probability 1 for some a, $\mathbf{E}[X] = a$). This is called *linearity* of expectation.

(2) If $X \geq Y$ (meaning, $X(\omega) \geq Y(\omega)$ for all ω), then $\mathbf{E}[X] \geq \mathbf{E}[Y]$

(3) If $\varphi : \mathbb{R} \to \mathbb{R}$, then

$$\mathbf{E}[\varphi(X)] = \begin{cases} \sum_t \varphi(t)f(t) & \text{if } f \text{ is a pmf.} \\ \int_{-\infty}^{+\infty} \varphi(t)f(t)dt & \text{if } f \text{ is a pdf.} \end{cases}$$

(4) More generally, if $(X_1, \ldots, X_n)$ has joint pdf $f(t_1, \ldots, t_n)$ and $V = T(X_1, \ldots, X_n)$ (here $T : \mathbb{R}^n \to \mathbb{R}$), then $\mathbf{E}[V] = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} T(x_1, \ldots, x_n)f(x_1, \ldots, x_n)dx_1 \ldots dx_n$.

For random variables on a discrete probability space (then they have pmf), we have essentially proved all these properties (or you can easily do so). For random variables with pmf, a proper proof require a bit of work. So we shall just take these for granted. We state one more property of expectations, its relationship to independence.

Lemma 131. *Let X, Y be random variables on a common probability space. If X and Y are independent, then $\mathbf{E}[H_1(X)H_2(Y)] = \mathbf{E}[H_1(X)]\mathbf{E}[H_2(Y)]$ for any functions $H_1, H_2 : \mathbb{R} \to \mathbb{R}$ (for which the expectations make sense). In particular, $\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]$.*

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,

$$\mathbf{E}[H_1(X)H_2(Y)] = \iint H_1(t)H_2(s)f(t, s)dtds = \left(\int_{-\infty}^{\infty} H_1(t)g(t)dt \right) \left(\int_{-\infty}^{\infty} H_2(s)g(s)ds \right)$$

which is precisely $\mathbf{E}[H_1(X)]\mathbf{E}[H_2(Y)]$. ∎

Expectation is a very important quantity. Using it, we can define several other quantities of interest.

Discussion: For simplicity let us take random variables to have densities in this discussion. You may adapt the remarks to the case of pmf easily. The density has all the information we need about a random variable. However, it is a function, which means that we have to know $f(t)$ for every t. In real life often we have random variables whose pdf is unknown or impossible to determine. It would be better to summarize the main features of the distribution (i.e., the density) in a few numbers. That is what the quantities defined below try to do.

Mean: Mean is another term for expected value.

Quantiles: Let us assume that the CDF F of X is strictly increasing and continuous. Then $F^{-1}(t)$ is well defined for every $t \in (0,1)$. For each $t \in (0,1)$, the number $Q_t = F^{-1}(t)$ is called the t-quantile. For example, the $1/2$-quantile, also called *median* is the number x such that $F(x) = \frac{1}{2}$ (unique when the CDF is strictly increasing and continuous). Similarly one defines $1/4$-quantile and $3/4$-quantile and these are sometimes called quartiles.[13]

Moments: The quantity $\mathbf{E}[X^k]$ (if it exists) is called the k^{th} *moment* of X.

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

$$\begin{aligned}
\sigma^2 &= \mathbf{E}[X^2 + \mu^2 - 2\mu X] \qquad \text{(by expanding the square)} \\
&= \mathbf{E}[X^2] + \mu^2 - 2\mu\mathbf{E}[X] \qquad \text{(by property (1) above)} \\
&= \mathbf{E}[X^2] - \mu^2.
\end{aligned}$$

That is $\text{Var}(X) = \mathbf{E}[X^2] - (\mathbf{E}[X])^2$.

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 $\mathbf{E}[|X - \text{med}(X)|]$.

Coefficient of variation: The coefficient of variation of X is defined as $\text{c.v.}(X) = \frac{\text{s.d.}(X)}{|\mathbf{E}[X]|}$.

[13] Another familiar quantity is the percentile, frequently used in reporting performance in competitive exams. For each x, the x-percentile is nothing but $F(x)$. For exam scores, it tells the proportion of exam-takers who scored less than or equal to x.

Covariance: Let X, Y be random variables on a common probability space. The *covariance* of X and Y is defined as $\text{Cov}(X,Y) = \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])]$. It can also be written as $\text{Cov}(X,Y) = \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y]$.

Correlation: Let X, Y be random variables on a common probability space. Their *correlation* is defined as $\text{Corr}(X,Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}}$.

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?

Measures of central tendency Mean and median try to summarize the distribution of X by a single number. Of course one number cannot capture the whole distribution, so there are many densities and mass functions that have the same mean or median. Which is better - mean or median? This question has no unambiguous answer. Mean has excellent mathematical properties (mainly linearity) which the median lacks (med$(X+Y)$ bears no general relationship to med$(X)+$ med(Y)). In contrast, mean is sensitive to outliers, while the median is far less so. For example, if the average income in a village of 50 people is 1000 Rs. per month, the immigration of multi-millionaire to the village will change the mean drastically but the median remains about the same. This is good, if by giving one number we are hoping to express the state of a typical individual in the population.

Measures of dispersion: 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. How widely the distribution is spread is measured by standard deviation and mean absolute deviation. Since we want deviation from mean, $\mathbf{E}[X - \mathbf{E}[X]]$ looks natural, but this is zero because of cancellation of positive and negative deviations. To prevent cancellation, we may put absolute values (getting to the m.a.d, but that is usually taken around the median) or we may square the deviations before taking expectation (giving the variance, and then the standard deviation). Variance and standard deviation have much better mathematical properties (as we shall see) and hence are usually preferred.

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.

Measures of association: The marginal distributions do not determine the joint distribution. For example, if (X, Y) is a point chosen at random from the unit square (with vertices $(0,0), (1,0), (0,1), (1,1)$) then X, Y both have marginal distribution that is uniform on $[0,1]$. If (U, V) is a point picked at random from the diagonal line (the line segment from $(0,0)$ to $(1,1)$, then again U and V have marginals that are uniform on $[0,1]$. But the two joint distributions are completely different. In particular, giving the means and standard deviations of X and Y does not tell anything about possible relationships between the two.

Covariance is the quantity that is used to measure the "association" of Y and X. Correlation is a dimension free quantity that measures the same. For example, we shall see that if $Y = X$, then Corr$(X, Y) = +1$, if $Y = -X$ then Corr$(X, Y) = -1$. Further, if X and Y are independent, then Corr$(X, Y) = 0$. In general, if an increase in X is likely to mean an increase in Y, then the correlation is positive and if an increase in X is likely to mean a decrease in Y then the correlation is negative.

Example 132. 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

$$\mathbf{E}[X] = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} x e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \mu.$$

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 133. Let $X \sim N(0,1)$. Show that $\mathbf{E}[X^n] = 0$ if n is odd and if n is even then $\mathbf{E}[X^n] = (n-1)(n-3)\dots(3)(1)$ (product of all odd numbers up to and including $n-1$). What happens if $X \sim N(0,\sigma^2)$?

Exercise 134. Calculate the mean and variance for the following distributions.

(1) $X \sim \text{Geo}(p)$. $\mathbf{E}[X] = \frac{1}{p}$ and $\text{Var}(X) = \frac{q}{p^2}$.

(2) $X \sim \text{Bin}(n,p)$. $\mathbf{E}[X] = np$ and $\text{Var}(X) = npq$.

(3) $X \sim \text{Pois}(\lambda)$. $\mathbf{E}[X] = \lambda$ and $\text{Var}(X) = \lambda$.

(4) $X \sim \text{Hypergeo}(N_1, N_2, m)$. $\mathbf{E}[X] = \frac{mN_1}{N_1+N_2}$ and $\text{Var}(X) = ??$.

Exercise 135. Calculate the mean and variance for the following distributions.

(1) $X \sim \text{Exp}(\lambda)$. $\mathbf{E}[X] = \frac{1}{\lambda}$ and $\text{Var}(X) = \frac{1}{\lambda^2}$.

(2) $X \sim \text{Gamma}(\nu,\lambda)$. $\mathbf{E}[X] = \frac{\nu}{\lambda}$ and $\text{Var}(X) = \frac{\nu}{\lambda^2}$.

(3) $X \sim \text{Unif}[0,1]$. $\mathbf{E}[X] = \frac{1}{2}$ and $\text{Var}(X) = \frac{1}{12}$.

(4) $X \sim \text{Beta}(p,q)$. $\mathbf{E}[X] = \frac{p}{p+q}$ and $\text{Var}(X) = \frac{pq}{(p+q)^2(p+q+1)}$.

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): $\text{Cov}(aX_1 + bX_2, Y) = a\text{Cov}(X_1, Y) + b\text{Cov}(X_2, Y)$ and $\text{Cov}(Y, aX_1 + bX_2) = a\text{Cov}(Y, X_1) + b\text{Cov}(Y, X_2)$

(2) (Symmetry): $\text{Cov}(X,Y) = \text{Cov}(Y,X)$.

(3) (Positivity): $\text{Cov}(X,X) \geq 0$ with equality if and only if X is a constant random variable. Indeed, $\text{Cov}(X,X) = \text{Var}(X)$.

Exercise 136. Show that $\text{Var}(cX) = c^2\text{Var}(X)$ (hence $\text{sd}(cX) = |c|\text{sd}(X)$). Further, if X and Y are independent, then $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$.

Note that the properties of ovariance are very much like properties of inner-products in vector spaces. In particular, we have the following analogue of the well-known inequality for vectors $(\mathbf{u}\cdot\mathbf{v})^2 \leq (\mathbf{u}\cdot\mathbf{u})(\mathbf{v}\cdot\mathbf{v})$.

Cauchy-Schwarz inequality: If X and Y are random variables with finite variances, then $(\text{Cov}(X,Y))^2 \leq \text{Var}(X)\text{Var}(Y)$ with equality if and only if $Y = aX + b$ for some scalars a, b.

If not convinced, follow the proof of Cauchy-Schwarz inequality that you have seen for vectors (basically, note that $\text{Var}(X + tY) \geq 0$ for any scalar t and choose an appropriate t to get the Cauchy-Schwarz's inequality.

21. Makov's and Chebyshev's inequalities

Let X be a non-negative integer valued random variable with pmf $f(k)$, $k = 0, 1, 2, \ldots$. Fix any number m, say $m = 10$. Then

$$\mathbf{E}[X] = \sum_{k=1}^{\infty} kf(k) \geq \sum_{k=10}^{\infty} kf(k) \geq \sum_{k=10}^{\infty} 10f(k) = 10\mathbf{P}\{X \geq 10\}.$$

More generally $m\mathbf{P}\{X \geq m\} \leq \mathbf{E}[X]$. 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 $\mathbf{P}\{X \geq t\} \leq \frac{1}{t}\mathbf{E}[X]$.

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 $\mathbf{E}[X] = \mathbf{E}[Y] + \mathbf{E}[Z] \geq \mathbf{E}[Z]$. On the other hand, $Z \geq t\mathbf{1}_{X\geq t}$ (why?). Therefore $\mathbf{E}[Z] \geq t\mathbf{E}[\mathbf{1}_{X\geq t}] = t\mathbf{P}\{X \geq t\}$. Putting these together we get $\mathbf{E}[X] \geq t\mathbf{P}\{X \geq t\}$ 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 $\mathbf{P}\{X \geq t\} \leq t^{-p}\mathbf{E}[X^p]$ for any $t > 0$.

(2) If X is a random variable with finite second moment, then $\mathbf{E}[|X - \mu| \geq t] \leq \frac{1}{t^2}\text{Var}(X)$. [*Chebyshev's inequality*]

(3) IF X is a random variable with finite exponential moments, then $\mathbf{P}(X > t) \leq e^{-\lambda t}\mathbf{E}[e^{\lambda X}]$ for any $\lambda > 0$.

Thus, if we only know that X has finite mean, the tail probability $\mathbf{P}(X > t)$ 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 $\mathbf{E}[e^{\lambda X}] < \infty$, 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

$$\mathbf{P}(|X-\mu| > r\sigma) \leq \frac{1}{r^2}$$

where $\sigma = \text{s.d.}(X)$. This measures the deviations in multiples of the standard deviation. This is a very general inequality. In specific cases we can get better bounds than $1/r^2$ (just like Markov inequality can be improved using higher moments, when they exist).

One more useful inequality we have already seen is the Cauchy-Schwarz inequality: $(\mathbf{E}[XY])^2 \leq \mathbf{E}[X^2]\mathbf{E}[Y^2]$ or $(\text{Cov}(X,Y))^2 \leq \text{Var}(X)\text{Var}(Y)$.

22. 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, $\mu = \mathbf{E}[X_1]$ and $\sigma^2 = \text{Var}(X_1)$ are well-defined.

Let $S_n = X_1 + \ldots + X_n$ (partial sums) and $\bar{X}_n = \frac{S_n}{n} = \frac{X_1+\ldots+X_n}{n}$ (*sample mean*). Then, by the properties of expectation and variance, we have

$$\mathbf{E}[S_n] = n\mu, \;\; \text{Var}(S_n) = n\sigma_1^2, \qquad \mathbf{E}[\bar{X}_n] = \mu, \;\; \text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}.$$

In particular, $\text{s.d.}(\bar{X}_n) = \sigma/\sqrt{n}$ decreases with n. If we apply Chebyshev's inequality to $\bar{X}_n$, we get for any $\delta > 0$ that

$$\mathbf{P}\{|\bar{X}_n - \mu| \geq \delta\} \leq \frac{\sigma^2}{\delta^2 n}.$$

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, we can get a better guess of what the value of p is.

Weak law of large numbers (Jacob Bernoulli): With the above notations, for any $\delta > 0$, we have

$$\mathbf{P}\{|\bar{X}_n - \mu| \geq \delta\} \leq \frac{\sigma^2}{\delta^2 n} \to 0 \quad \text{as } n \to \infty.$$

This is very general, in that we only assume the existence of variance. If X_k are assumed to have more moments, one can get better bounds. For example, when X_k are i.i.d. Ber(p), we have the following theorem.

Hoeffding's inequality: Let $X_1, \ldots, X_n$ be i.i.d. Ber(p). Then

$$\mathbf{P}\{|\bar{X}_n - p| \geq \delta\} \leq 2e^{-n\delta^2/2}.$$

23. 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

$$\mu = \int_0^1 \varphi(x)dx = I, \qquad \sigma^2 := \text{Var}(X_1) = \int_0^1 (\varphi(x) - I)^2 dx.$$

This gives the following method of finding I. Fix a large number N appropriately and pick N uniform random numbers U_k, $1 \le k \le N$. 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 $\mathbf{P}\{|\hat{I}_n - I| \ge \delta\} \to 0$ and hence we expect $\hat{I}_n$ to be close to I. How large should n be? For this, we fix two numbers $\epsilon = 0.01$ and $\delta = 0.001$ (you may change the numbers). By Chebyshev's inequality, observe that $\mathbf{P}\{|\hat{I}_n - I| \ge \delta\} \to \sigma^2/N\delta^2$.

First find N so that $\sigma^2/N\delta^2 < \epsilon$, i.e., $N = \lceil \frac{\sigma^2}{\delta^2} \rceil$. Then, the random variable $\hat{I}_N$ is within δ of I with probability greater than $1 - \epsilon$. This is a probabilistic method, hence there is a possibility of large error, but with a small probability. Observe that N grows proportional to *square* of $1/\delta$. To increase the accuracy by 10, you must increase the number of samples by a factor of 100.

One last point. To find N we need σ^2 which involves computing another integral involving φ which we do not know how to compute! Here we do not need the exact value of the integral. For example, if our functions satisfies $-M \le \varphi(x) \le M$ for all $x \in [0,1]$, then also $-M \le I \le M$ and hence $(\varphi(x) - \mathbf{1}_)^2 \le 4M^2$. This means that $\sigma^2 \le 4M^2$. Therefore, if we take $N = \lceil \frac{4M^2}{\delta^2} \rceil$ then the value of N is larger than required for the desired accuracy. We can work with this N. Note that the dependence of of N on δ does not change.

Exercise 137. We know that $\int_0^1 \frac{1}{1+x^2}dx = \frac{\pi}{4}$. Based on this, devise a method to find an approximate value of π. Use any software you like to implement your method and see how many sample you need to get an approximation to 1, 2 and 3 decimal places consistently (consistently means with a large enough probability, say 0.9).

Exercise 138. Devise a method to approximate e and π (there are many possible integrals).

This method can be used to evaluate integrals over any interval. For instance, how would you find $\int_a^b \varphi(t)dt$ 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? 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 $\{(x,y) : 0 \leq x, y, \leq 1,\ 2x^3y^2 \geq 1,\ x^2 + 2y^2 \leq 2.3\}$. 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.

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. Lagrange's method, Gauss quadrature etc are of this type. Under certain assumptions on φ, the accuracy of these integrals can be like $1/N$ as opposed to $1/\sqrt{N}$ in Monte-Carlo. 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!

24. 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 $\bar{X}_n$ has mean μ and standard deviation $\sigma/\sqrt{n}$.

This roughly means that $\bar{X}_n$ is close to μ, within a few multiples of $\sigma/\sqrt{n}$ (as shown by Chebyshev's inequality). Now we look at $\bar{X}_n$ with a finer microscope. In other words, we ask for the probability that $\bar{X}_n$ 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!

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

$$\mathbf{P}\left\{\mu + a\frac{\sigma}{\sqrt{n}} \leq \bar{X}_n \leq \mu + b\frac{\sigma}{\sqrt{n}}\right\} \to \Phi(b) - \Phi(a) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-t^2/2} dt.$$

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 $\mathbf{P}\{a < Z < b\} = \Phi(b) - \Phi(a)$.

(1) $\mathbf{P}\{a < \frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu) \leq b\} \to \Phi(b) - \Phi(a) = \mathbf{P}\{a < Z < b\}$. Put another way, this says that for large n, the random variable $\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$ has $N(0,1)$ distribution, approximately. Equivalently, $\sqrt{n}(\bar{X}_n - \mu)$ 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,

$$\mathbf{P}\{a \leq \frac{S_n - n\mu}{\sigma\sqrt{n}} \leq b\} \to \mathbf{P}\{a < Z < b\}.$$

The central limit theorem so deep and surprising and useful. The following example gives a hint as to why.

Example 139. 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

$$\mathbf{P}\{\frac{a}{\sqrt{3}} < Y_n < \frac{b}{\sqrt{3}}\} \to \Phi(b) - \Phi(a).$$

or equivalently, $\mathbf{P}\{a < Y_n < b\} \to \Phi(b\sqrt{3}) - \Phi(a\sqrt{3})$. 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 $\mathbf{P}\{a < Y_n < b\}$. 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!

We shall not prove the central limit theorem in general. But we indicate how it is done when X_k come from Exp(λ) distribution. This is optional and may be skipped.

CLT for Exponetials. 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 $\mathbf{P}\{a < Y_n < b\} = \int_a^b g_n(t)dt$. 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,

$$\mathbf{P}\{a < Y_n < b\} \to \int_a^b \varphi(t)dt = \Phi(b) - \Phi(a).$$

This proves CLT for the case of exponential random variables. ■

25. 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 140. *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 $k \geq 0$*

$$\mathbf{P}\{S_n = k\} \to e^{-\lambda}\frac{\lambda^k}{k!}.$$

Proof. Fix k and observe that

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

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 141. (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 $\text{Pois}(3)$ 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 142. 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 $\text{Pois}(1/2)$ distribution. Work out how this comes from the theorem.

The Poisson limit is a much more general phenomenon than what the theorem above captures. For example, consider the problem of a psychic guessing a deck of cards. If X is the number of correct guesses, we saw (by direct calculation and approximation) that $\mathbf{P}\{X = k\}$ is close to $e^{-1}/k!$. In other words X has approximately Pois(1) distribution. Does it follows from the theorem above. Let us try.

Set X_k to be the indicator of the event that the kth guess is correct. Then $X_k \sim$ Ber$(1/52)$ and $X = X_1 + \ldots + X_{52}$. It looks like the theorem tells us that X should have Pois(1) distribution (by taking $n = 52$ and $\lambda = 1$). But note that X_i are not independent random variables and hence the theorem does not strictly apply. 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"*.

26. Entropy, Gibbs distribution

Definition 143. Let X be a random variable that takes values in $\mathcal{A} = \{a_1, \ldots, a_k\}$ such that $\mathbf{P}(X = a_i) = p_i$. The entropy of X is defined as

$$H(X) := -\sum_{i=1}^{k} p_i \log p_i.$$

If X is a real-valued random variable with density f, its entropy is defined

$$H(X) := -\int f(t) \log f(t) dt.$$

Example 144. Let $X \sim$ Ber(p). Then $H(X) = p\log(1/p) + (1-p)\log(1/(1-p))$.

Example 145. Let $X \sim$ Geo(p). Then $H(X) = -\sum_{k=0}^{\infty} (\log p + k \log q) p q^k = -\log p - q^2 \log q$.

Example 146. Let $X \sim$ Exp(λ). Then $H(X) = \int_0^\infty (\log \lambda - t)\lambda e^{-\lambda t} dt = \log \lambda - \frac{1}{\lambda}$.

Example 147. Let $X \sim N(\mu, \sigma^2)$

Entropy is a measure of the randomness. For example, among the Ber(p) distributions, the entropy is maximized at $p = 1/2$ and minimized at $p = 0$ or 1. It quantifies the intuitive feeling that Ber$(1/2)$ is *more random* than Ber$(1/4)$.

Lemma 148. *(1) If $|\mathcal{A}| = k$, then $0 \le H(X) \le \log k$. $H(X) = 0$ if and only if X is degenerate and $H(X) = \log k$ if and only if $X \sim$ Unif$(\mathcal{A})$.*

(2) Let $f : \mathcal{A} \to \mathcal{B}$ and let $Y = f(X)$. Then $H(Y) \le H(X)$.

(3) Let X take values in $\mathcal{A}$ and Y take values in $\mathcal{B}$ and let $Z = (X, Y)$. Then $H(Z) \le H(X) + H(Y)$ with equality if and only if X and Y are independent.

Gibbs measures: Let $\mathcal{A}$ be a countable set and let $\mathcal{H} : \mathcal{A} \to \mathbb{R}$ be a given function. For any $E \in \mathbb{R}$, consider the set of $\mathcal{P}_E$ of all probability mass functions on Ω under which $\mathcal{H}$ has expected value E. In other words,

$$\mathcal{P}_E := \{\mathbf{p} = (p_i)_{i\in\mathcal{A}} : \sum_{i\in\mathcal{A}} p(i)\mathcal{H}(i) = E\}.$$

$\mathcal{P}_E$ is non-empty if and only if $\mathcal{H}_{\min} \le E \le \mathcal{H}_{\max}$.

Lemma 149. *Assume that $\mathcal{H}_{\min} \le E \le \mathcal{H}_{\max}$. Then, there is a unique pmf in $\mathcal{P}_E$ with maximal entropy and it is given by*

$$p_\beta(i) = \frac{1}{Z_\beta} e^{-\beta\mathcal{H}(i)}$$

where $Z_\beta = \sum_{i\in\mathcal{A}} e^{-\beta\mathcal{H}(i)}$ and the value of β is chosen to satisfy $\frac{1}{Z_\beta}\frac{\partial Z_\beta}{\partial \beta} = E$.

This minimizing pmf is called the *Boltzmann-Gibbs* distribution. An analogous theorem holds for densities.

Example 150. Let $\mathcal{A} = \{1, 2, \ldots, n\}$ and $\mathcal{H}(i) = 1$ for all i. Let $E = 1$ so that $\mathcal{P}_E$ is the same as all pmfs on $\mathcal{A}$. Clearly $p_\beta(i) = \frac{1}{n}$ for all $i \le n$. Indeed, we know that the maximal entropy is attained by the uniform distribution.

Example 151. Let $\mathcal{A} = \{0, 1, 2, \ldots\}$ and let $\mathcal{H}(i) = i$ for all i. Fix any $E > 0$. The Boltzmann-Gibbs distribution is given by $p_\beta(i) = \frac{1}{Z_\beta} e^{-\beta i}$. This is just the Geometric distribution with parameter chosen to have mean E.

Example 152. Let us blindly apply the lemma to densities.

(1) $\mathcal{A} = \mathbb{R}_+$ and $\mathcal{H}(x) = \lambda x$

(2) $\mathcal{A} = \mathbb{R}$ and $\mathcal{H}(x) = x^2$.

Statistics

1. INTRODUCTION

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 $\bar{X}_n$ (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 $[\bar{X}_n - \frac{3}{\sqrt{n}}\sqrt{\bar{X}_n(1-\bar{X}_n)}, \bar{X}_n + -\frac{3}{\sqrt{n}}\sqrt{\bar{X}_n(1-\bar{X}_n)}]$ where $\bar{X}_n$ 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!).

2. Estimation problems

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) 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.

(3) Can we guess the average height μ of all people in India by taking a random sample of 100 people and measuring their heights?

In such questions, there is an unknown parameter μ (there could be more than one unknown parameter too) whose value we are trying to guess based on the data. The data consists of i.i.d. random variables from a family of distributions. We assume that the family of distributions is known and the only unknown is (are) the value of the parameter(s). Rather than present the ideas in abstract let us see a few examples.

Example 153. Let $X_1, \ldots, X_n$ be i.i.d. random variables with Exponential density $f_\mu(x) = \frac{1}{\mu} e^{-x/\mu}$ (fro $x > 0$) where the value of $\mu > 0$ is unknown. How to *estimate* it using the data $X = (X_1, \ldots, X_n)$?

This is the framework in which we would study the second example above, namely the lie-time distribution of light bulbs. Observe that we have parameterized the exponential family of distributions differently from usual. We could equivalently have considered $g_\lambda(x) = \lambda e^{-\lambda x}$ but the interest is then in estimating $1/\lambda$ (which is the expected value) rather than λ. Here are two methods.

Method of moments: We observe that $\mu = \mathbf{E}_\mu[X_1]$, the mean of the distribution (also called *population mean*). Hence it seems reasonable to take the sample mean $\bar{X}_n$ as an estimate. On second thought, we realize that $\mathbf{E}_\mu[X_1^2] = 2\mu^2$ and hence $\mu = \sqrt{\frac{1}{2}\mathbf{E}_\mu[X_1^2]}$. Therefore it also seems reasonable to take the corresponding sample quantity, $T_n := \sqrt{\frac{1}{2n}(X_1^2 + \ldots + X_n^2)}$ as an estimate for μ. One can go further and write μ in various ways as $\mu = \sqrt{\text{Var}_\mu(X_1)}$, $\mu = \sqrt[3]{\frac{1}{6}\mathbf{E}_\mu[X_1^3]}$ etc. Each such expression motivates an estimate, just by substituting sample moments for population moments.

This is called estimating by the *method of moments* because we are equating the sample moments to population moments to obtain the estimate.

We can also use other features of the distribution, such as quantiles (we may call this the "method of quantiles"). In other words, obtain estimates by equating the sample quantiles to population quantiles. For example, the median of X_1 is $\mu \log 2$, hence a reasonable estimate for μ is $M_n/\log 2$, where M_n is a sample median. Alternately, the 25% quantile of Exponential$(1/\mu)$ distribution is $\mu \log(4/3)$ and hence another estimate for μ is $Q_n/\log(4/3)$ where Q_n is a 25% sample quantile.

Maximum likelihood method: The joint density of $X_1, \ldots, X_n$ is

$$g_\mu(x_1, \ldots, x_n) = \mu^{-n} e^{-\mu(x_1 + \ldots + x_n)} \qquad \text{if all } x_i > 0$$

(since X_i are independent, the joint density is a product). We evaluate the joint density at the observed data values. This is called the likelihood function. In other words, define,

$$L_X(\mu) := \mu^{-n} e^{-\frac{1}{\mu}\sum_{i=1}^n X_i}.$$

Two points: This is the joint density of $X_1, \ldots, X_n$, evaluated at the observed data. Further, we like to think of it as a function of μ with $X := (X_1, \ldots, X_n)$ being fixed.

When μ is the actual value, then $L_X(\mu)$ is the "likelihood" of seeing the data that we have actually observed. The *maximum likelihood estimate* is that value of μ that maximizes the likelihood

function. In our case, by differentiating and setting equal to zero we get,

$$0 = \frac{d}{d\mu} L_X(\mu) = -n\mu^{-n-1} e^{-\frac{1}{\mu}\sum_{i=1}^n X_i} + \mu^{-n} \left(\frac{1}{\mu^2} \sum_{i=1}^n X_i \right) e^{-\frac{1}{\mu}\sum_{i=1}^n X_i}$$

which is satisfied when $\mu = \frac{1}{n}\sum_{i=1}^n X_i = \bar{X}_n$. To distinguish this from the true value of μ which is unknown, it is customary to put a hat on the leter μ. We write $\hat{\mu}_{MLE} = \bar{X}_n$. We should really verify whether $L(\mu)$ is maximized or minimized (or neither) at this point, but we leave it to you to do the checking (eg., by looking at the second derivative).

Let us see the same methods at work in two more examples.

Example 154. Let $X_1, \ldots, X_n$ be i.i.d. Ber(p) random variables where the value of p is unknown. How to *estimate* it using the data $X = (X_1, \ldots, X_n)$?

Method of moments: We observe that $p = \mathbf{E}_p[X_1]$, the mean of the distribution (also called *population mean*). Hence, a method of moments estimator would be the sample mean $\bar{X}_n$. In this case, $\mathbf{E}_p[X_1^2] = p$ again but we don't get any new estimate because $X_k^2 = X_k$ (as X_k is 0 or 1)

Maximum likelihood method: Now we have a probability mass function instead of density. The joint pmf of of $X_1, \ldots, X_n$ is $f_p(x_1, \ldots, x_n = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}$ when each x_i is 0 or 1. The likelihood function is

$$L_X(p) := p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i} = p^{n\bar{X}_n}(1-p)^{n(1-\bar{X}_n)}.$$

We need to find the value of p that maximizes $L_X(p)$. Here is a trick that almost always simplifies calculations (try it in the previous example too!). Instead of maximizing $L_X(p)$, maximize $\ell_X(p) = \log L_X(p)$ (called the *log-likelihood function*). Since "log" is an increasing function, the maximizer will remain the same. In our case,

$$\ell_X(p) = \bar{X}_n \log p + n(1-\bar{X}_n)\log(1-p).$$

Differentiating and setting equal to 0, we get $\hat{p}_{MLE} = \bar{X}_n$. Again the sample mean is the maximum likelihood estimate.

A last example.

Example 155. Consider the two-parameter Laplace-density $f_{\theta,\alpha}(x) = \frac{1}{2\alpha} e^{-\frac{|x-\theta|}{\alpha}}$ for all $x \in \mathbb{R}$. Check that $f_{\theta,\alpha}$ is indeed a density for all $\theta \in \mathbb{R}$ and $\alpha > 0$.

Now suppose we have data $X_1, \ldots, X_n$ i.i.d. from $f_{\theta,\alpha}$ where we do not know the values of θ and α. How to estimate the parameters?

Method of moments: We compute

$$\mathbf{E}_{\theta,\alpha}[X_1] = \frac{1}{2\alpha}\int_{-\infty}^{+\infty} te^{-\frac{|t-\theta|}{\alpha}}dt = \frac{1}{2}\int_{-\infty}^{+\infty}(\alpha s+\theta)e^{-|s|}ds = \theta.$$

$$\mathbf{E}_{\theta,\alpha}[X_1^2] = \frac{1}{2\alpha}\int_{-\infty}^{+\infty} t^2e^{-\frac{|t-\theta|}{\alpha}}dt = \frac{1}{2}\int_{-\infty}^{+\infty}(\alpha s+\theta)^2e^{-|s|}ds = 2\alpha^2+\theta^2.$$

Thus the variance is $\text{Var}_{\theta,\alpha}(X_1) = 2\alpha^2$. Based on this, we can take the method of moments estimate to be $\hat{\theta}_n = \bar{X}_n$ (sample mean) and $\hat{\alpha}_n = \frac{1}{\sqrt{2}}s_n$ where $s_n^2 = \frac{1}{n-1}\sum_{i=1}^n(X_i-\bar{X}_n)^2$. At the moment the ideas of defining sample variance as s_n^2 may look strange and it might be more natural to take $V_n := \frac{1}{n}\sum_{i=1}^n(X_i-\bar{X}_n)^2$ as an estimate for the population variance. As we shall see later, s_n^2 has some desirable properties that V_n lacks. Whenever we say sample variance, we mean s_n^2, unless stated otherwise.

Maximum likelihood method: The likelihood function of the data is

$$L_X(\theta,\alpha) = \prod_{k=1}^n \frac{1}{2\alpha}\exp\left\{-\frac{|X_k-\theta|}{\alpha}\right\} = 2^{-n}\alpha^{-n}\exp\left\{-\sum_{k=1}^n\frac{|X_k-\theta|}{\alpha}\right\}.$$

The log-likelihood function is

$$\ell_X(\theta,\alpha) = \log L(\theta,\alpha) = -n\log 2 - n\log\alpha - \frac{1}{\alpha}\sum_{k=1}^n|X_k-\theta|.$$

We know that[14] for fixed $X_1,\ldots,X_n$, the value of $\sum_{k=1}^n|X_k-\theta|$ is minimized when $\theta = M_n$, the median of $X_1,\ldots,X_n$ (strictly speaking the median may have several choices, all of them are equally good). Thus we fix $\hat{\theta} = M_n$ and then we maximize $\ell(\hat{\theta},\alpha)$ over α by differentiating. We get $\hat{\alpha} = \frac{1}{n}\sum_{k=1}^n|X_k-\theta|$ (the sample mean-absolute deviation about the median). Thus the MLE of (θ,α) is $(\hat{\theta},\hat{\alpha})$.

In homeworks and tutorials you will see several other estimation problems which we list in the exercise below.

Exercise 156. Find an estimate for the unknown parameters by the method of moments and the maximum likelihood method.

[14] If you do not know here is an argument. Let $x_1 < x_2 < \ldots < x_n$ be n distinct real numbers and let $a \in \mathbb{R}$. Rewrite $\sum_{k=1}^n|x_k-a|$ as $(|x_1-a|+|x_n-a|)+(|x_2-a|+|x_{n-1}-a|)+\ldots$. By triangle inequality, we see that

$$|x_1-a|+|x_n-a| \geq x_n-x_1, \quad |x_2-a|+|x_{n-1}-a| \geq x_{n-1}-x_2, \quad |x_3-a|+|x_{n-2}-a| \geq x_{n-2}-x_3\ldots.$$

Further the first inequality is an equality if and only if $x_1 \leq a \leq x_n$, the second inequality is an equality if and only if $x_2 \leq a \leq x_{n-1}$ etc. In particular, if a is a median, then all these inequalities become equalities and shows that a median minimizes the given sum.

(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 157. $X_1, \ldots, X_n$ are i.i.d. Geo(p) Estimate $\mu = 1/p$.

Exercise 158. $X_1, \ldots, X_n$ are i.i.d. Pois(λ) Estimate λ.

Exercise 159. $X_1, \ldots, X_n$ are i.i.d. Beta(a, b) Estimate a, b.

The following exercise is approachable by the same methods but requires you to think a little.

Exercise 160. $X_1, \ldots, X_n$ are i.i.d. Uniform$[a, b]$ Estimate a, b.

3. Properties of estimates

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.

The setting: Let $X_1, \ldots, X_n$ be i.i.d random variables with a common density $f_\theta(x)$. The parameter θ is unknown and the goal is to estimate it. Let T_n be an estimator for θ, this just means that T_n is a function of $X_1, \ldots, X_n$ (in words, if we have the data at hand, we should be able to compute the value of T_n).

Bias: Define the *bias* of the estimator as $\text{bias}_{T_n}(\theta) := \mathbf{E}_\theta[T_n] - \theta$. 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 $\mathbf{E}_\theta$ to remind ourself that in computing the expectation we use the density f_θ. However we shall often omit the subscript for simplicity.

Mean-squared error: The *mean squared error* of T_n is defined as $\text{m.s.e.}_{T_n}(\theta) = \mathbf{E}_\theta[(T_n - \theta)^2]$. This is a function of θ. Smaller it is, better our estimate.

In computing mean squared error, it is useful to observe the formula

$$\text{m.s.e.}_{T_n}(\theta) = \text{Var}_{T_n}(\theta) + \left(\text{Bias}_{T_n}(\theta)\right)^2 .$$

To prove this, consider and random variable Y with mean μ and observe that for any real number a we have

$$\begin{aligned}\mathbf{E}[(Y-a)^2] &= \mathbf{E}[(Y-\mu+\mu-a)^2] = \mathbf{E}[(Y-\mu)^2] + (\mu-a)^2 + 2(\mu-a)\mathbf{E}[Y-\mu] \\ &= \mathbf{E}[(Y-\mu)^2] + (\mu-a)^2 = \text{Var}(Y) + (\mu-a)^2.\end{aligned}$$

Use this identity with T_n in place of Y and θ in place of a.

Example 161. Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$. Let $V_n = \frac{1}{n}\sum_{k=1}^n (X_k - \bar{X}_n)^2$ be an estimate for σ^2. By expanding the squares we get

$$V_n = \bar{X}_n^2 + \frac{1}{n}\sum_{k=1}^n X_k^2 - \frac{2}{n}\bar{X}_n \sum_{k=1}^n X_k = \left(\frac{1}{n}\sum_{k=1}^n X_k^2\right) - \bar{X}_n^2.$$

It is given that $\mathbf{E}[X_k] = \mu$ and $\text{Var}(X_k) = \sigma^2$. Hence $\mathbf{E}[X_k^2] = \mu^2 + \sigma^2$. We have seen before that $\text{Var}(\bar{X}_n) = \sigma^2$ and $\mathbf{E}[\bar{X}_n] = \mu$. Hence $\mathbf{E}[\bar{X}_n^2] = \mu^2 + \frac{\sigma^2}{n}$. Putting all this together, we get

$$\mathbf{E}\left[V_n\right] = \left(\frac{1}{n}\sum_{k=1}^n \mu^2 + \sigma^2\right) - \left(\mu^2 + \frac{\sigma^2}{n}\right) = \frac{n-1}{n}\sigma^2.$$

Thus, the bias of V_n is $\frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{1}{n}\sigma^2$.

Example 162. 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 $\mathbf{E}[W_n] = \sigma^2$. 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 $s_n^2 = \frac{1}{n-1}\sum_{k=1}^n (X_k - \bar{X}_n)^2$. Clearly $s_n^2 = \frac{n}{n-1}V_n$ and hence $\mathbf{E}[s_n^2] = \frac{n}{n-1}\mathbf{E}[V_n] = \sigma^2$. 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 $s_n^2 = \frac{1}{n-1}\sum_{k=1}^n (X_k - \bar{X}_n)^2$ is unbiased for the population variance. But $V_n = \frac{1}{n}\sum_{k=1}^n (X_k - \bar{X}_n)^2$ is not, in fact $\mathbf{E}[V_n] = \frac{n-1}{n}\sigma^2$.

It appears that s_n^2 is better, but the following remark says that one should be cautious in making such a statement.

Remark 163. 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 164. Let $X_1, \ldots, X_n$ be i.i.d. Ber(p). Then $\bar{X}_n$ is an estimate for p. It is unbiased since $\mathbf{E}[\bar{X}_n] = p$. Hence, the m.s.e of $\bar{X}_n$ 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 $\bar{X}_n$ which is unbiased and hence $\text{MSE}_{\bar{X}_n}(p) = \text{Var}(\bar{X}_n) = p(1-p)/n$. 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 $\text{MSE}_{\bar{Y}_n/2}(p) = \text{Var}(\bar{Y}_n) = 2p(1-2p)/4 = p(1-2p)/2$. It is not hard to see that for all $p < 1/2$, $\text{MSE}_{\bar{Y}_n/2}(p) < \text{MSE}_{\bar{X}_n}(p)$ 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 rest 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 $\bar{X}_n$ which has $\text{MSE}_{\bar{X}_n}(\mu) = \text{Var}(\bar{X}_n) = \mu^2/n$. 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 $X/2 \sim \text{Exp}(1/2\mu)$ 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 $\text{Exp}(1/2\mu)$ distribution. Conversely if we have data from $\text{Exp}(1/2\mu)$ 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.

4. Confidence intervals

So far, in estimating of an unknown parameter, we give a single number as our guess for the known parameter. It would be better to give an interval and say with what confidence we expect the true parameter to lie within it. As a very simple example, suppose we have one random variable X with $N(\mu, 1)$ distribution. How do we estimate μ? Suppose the observed value of X is 2.7. Going by any method, the guess for μ would be 2.7 itself. But of course μ is not equal to X, so we would like to give an interval in which μ lies. How about $[X-1, X+1]$? Or $[X-2, X+2]$? Using normal tables, we see that $\mathbf{P}(X-1 < \mu < X+1) = \mathbf{P}(-1 < (X-\mu) < 1) = \mathbf{P}(-1 < Z < 1) \approx 0.68$ and similarly $\mathbf{P}(X-2 < \mu < X+2) \approx 0.95$. Thus, by making the interval longer we can be more confident that the true parameter lies within. But the accuracy of our statement goes down (if you want to know the average height of people in India, and the answer you give is "between 100cm and 200cm", it is very probably correct, but of little use!). The probability with which our CI contains the unknown parameter is called the level of confidence. Usually we fix the level of confidence, say as 0.90 and find an interval *as short as possible* but subject to the condition that it should have a confidence level of 0.90.

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 $\bar{X}_n$ and enlarge it to an interval $[\bar{X}_n - a, \bar{X}_n + a]$. Similarly, in finding a CI for σ^2 we use the estimate $s_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ 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.

4.1. **Estimating μ when σ^2 is known.** We look for a confidence interval of the form $I_n = [\bar{X}_n - a, \bar{X}_n + a]$. Then,

$$\mathbf{P}\left(I_n \ni \mu\right) = \mathbf{P}\left(-a \le \bar{X}_n - \mu \le a\right) = \mathbf{P}\left(-\frac{a\sqrt{n}}{\sigma} \le \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \le \frac{a\sqrt{n}}{\sigma}\right)$$

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 $Y_2 \sim N(\nu, \tau^2)$ and they are independent, then $X + Y \sim N(\mu + \nu, \sigma^2 + \tau^2)$.

Consequently, $\bar{X}_n \sim N(0, \sigma^2/n)$ and $\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \sim N(0, 1)$. Therefore,

$$\mathbf{P}\left(I_n \ni \mu\right) = \mathbf{P}(-\frac{a\sqrt{n}}{\sigma} \le Z \le -\frac{a\sqrt{n}}{\sigma})$$

where $Z \sim N(0, 1)$. Fix any $0 < \alpha < 1$ and denote by z_α the number such that $\mathbf{P}(Z > z_\alpha) = \alpha$ (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

$$\mathbf{P}\left(\left[\bar{X}_n - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \bar{X}_n + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right] \ni \mu\right) = 1 - \alpha.$$

This is our confidence interval.

4.2. **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 165. Let $Z_1, \ldots, Z_n$ be i.i.d. $N(0, 1)$ random variables. Then, $Z_1^2 + \ldots + Z_n^2 \sim \text{Gamma}(n/2, 1/2)$.

Solution: For $t > 0$ we have

$$\mathbf{P}\{Z_1^2 \le t\} = \mathbf{P}\{-\sqrt{t} \le Z_1 \le \sqrt{t}\} = 2\int_0^{\sqrt{t}} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du = \frac{1}{\sqrt{2\pi}} \int_0^t e^{-s/2} s^{-1/2} ds.$$

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$(1/2, 1/2)$ is usually called the *chi-squared distribution with n degrees of freedom*. Let $\chi^2_n(\alpha)$ denote the $1-\alpha$ quantile of this distribution. Similarly, $\chi^2_n(1-\alpha)$ is the α quantile (i.e., the probability for the chi-squared random variable to fall below $\chi^2_n(1-\alpha)$ 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

$$\mathbf{P}\left\{\frac{nW_n}{\chi^2_n\left(\frac{\alpha}{2}\right)} \le \sigma^2 \le \frac{nW_n}{\chi^2_n\left(1-\frac{\alpha}{2}\right)}\right\} = \mathbf{P}\left\{\chi^2_n\left(1-\frac{\alpha}{2}\right) \le \frac{nW_n}{\sigma^2} \le \chi^2_n\left(\frac{\alpha}{2}\right)\right\} = 1-\alpha.$$

Thus, $\left[\frac{ns_n^2}{\chi^2_{n-1}\left(\frac{\alpha}{2}\right)}, \frac{ns_n^2}{\chi^2_{n-1}\left(1-\frac{\alpha}{2}\right)}\right]$ is a $(1-\alpha)$-confidence interval for σ^2.

An important result: 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 ($\sqrt{n}(\bar{X}_n - \mu)/\sigma$ and W_n/σ^2, respectively) which involved the data and the unknown parameter whose distributions we knew (standard normal and χ^2_n, 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., $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ 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 166. *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,*

$$\bar{Z}_n \sim N(\mu, \frac{\sigma^2}{n}), \quad \frac{(n-1)s_n^2}{\sigma^2} \sim \chi^2_{n-1},$$

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.

4.3. **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

$$\mathbf{P}\left\{\frac{(n-1)s_n^2}{\chi^2_{n-1}\left(\frac{\alpha}{2}\right)} \le \sigma^2 \le \frac{(n-1)s_n^2}{\chi^2_{n-1}\left(1-\frac{\alpha}{2}\right)}\right\} = \mathbf{P}\left\{\chi^2_{n-1}\left(1-\frac{\alpha}{2}\right) \le \frac{(n-1)s_n^2}{\sigma^2} \le \chi^2_{n-1}\left(\frac{\alpha}{2}\right)\right\}$$

$$= 1-\alpha.$$

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. 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.

4.4. **Estimating μ when σ^2 is unknown.** The earlier confidence interval We look for a confidence interval $[\bar{X}_n - \frac{\sigma}{\sqrt{n}}z_{\alpha/2}, \bar{X}_n + \frac{\sigma}{\sqrt{n}}z_{\alpha/2}]$ cannot be used as we do not know the value of σ.

A natural idea would be to use the estimate $s_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ 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

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \sim N(0,1).$$

Is it true if σ is replaced by s_n? Actually no, but we have a different distribution called *Student's t-distribution*.

Exercise 167. Let $Z \sim N(0,1)$ and $S^2 \sim \chi^2_n$ be independent. Then, the density of $\frac{Z}{S/\sqrt{n}}$ 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*.

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.

How does this help us? From Theorem 166 we know that $\frac{\sqrt{n}(\bar{X}_n-\mu)}{\sigma} \sim N(0,1)$, $\frac{(n-1)s_n^2}{\sigma^2} \sim \chi^2_{n-1}$, and the two are independent. Take these random variables in the above exercise to conclude that $\frac{\sqrt{n}(\bar{X}_n-\mu)}{s_n}$ 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).

$$\begin{aligned}
&\mathbf{P}\left(\bar{X}_n - \frac{s_n}{\sqrt{n}} t_{n-1}\left(\frac{\alpha}{2}\right) \leq \mu \leq \bar{X}_n + \frac{s_n}{\sqrt{n}} t_{n-1}\left(\frac{\alpha}{2}\right)\right) \\
&= \mathbf{P}\left(-t_{n-1}\left(\frac{\alpha}{2}\right) \leq \frac{\sqrt{n}(\bar{X}_n - \mu)}{s_n} \leq t_{n-1}\left(\frac{\alpha}{2}\right)\right) \\
&= \mathbf{P}\left(-t_{n-1}\left(\frac{\alpha}{2}\right) \leq T_{n-1} \leq t_{n-1}\left(\frac{\alpha}{2}\right)\right) \\
&= 1 - \alpha.
\end{aligned}$$

Hence, our $(1-\alpha)$-confidence interval is $\left[\bar{X}_n - \frac{s_n}{\sqrt{n}} t_{n-1}\left(\frac{\alpha}{2}\right), \bar{X}_n + \frac{s_n}{\sqrt{n}} t_{n-1}\left(\frac{\alpha}{2}\right)\right]$.

Remark 168. 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

$$\left[\bar{X}_n - \frac{s_n}{\sqrt{n}} z_{\alpha/2}, \bar{X}_n + \frac{s_n}{\sqrt{n}} z_{\alpha/2}\right].$$

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 $1-\alpha$.

5. 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

$$\left[\bar{X}_n - \frac{s_n}{\sqrt{n}} t_{n-1}\left(\frac{\alpha}{2}\right), \bar{X}_n + \frac{s_n}{\sqrt{n}} t_{n-1}\left(\frac{\alpha}{2}\right)\right] \text{ or } \left[\bar{X}_n - \frac{s_n}{\sqrt{n}} z_{\alpha/2}, \bar{X}_n + \frac{s_n}{\sqrt{n}} z_{\alpha/2}\right]$$

Is this a valid confidence interval in general? The answer is "No" for both. If X_i are from some general distribution then the distributions of $\sqrt{n}(\bar{X}_n - \mu)/s_n$ and $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ are very complicated to find. Even if X_i come from binomial or exponential family, these distributions will 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 $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ has approximately $N(0,1)$ distribution.

Hence, we see that

$$\mathbf{P}\left\{-z_{\alpha/2} \le \frac{\sqrt{n}(\bar{X}_n - \mu)}{s_n} \le z_{\alpha/2}\right\} \approx \Phi(z_{\alpha/2}) - \Phi(-z_{\alpha/2}) = 1 - \alpha.$$

Consequently, we may say that

$$\mathbf{P}\left\{\bar{X}_n - \frac{s_n}{\sqrt{n}} z_{\alpha/2} \le \mu \le \bar{X}_n + \frac{s_n}{\sqrt{n}} z_{\alpha/2}\right\} \approx 1 - \alpha.$$

Thus, $\left[\bar{X}_n - \frac{s_n}{\sqrt{n}} z_{\alpha/2}, \bar{X}_n + \frac{s_n}{\sqrt{n}} z_{\alpha/2}\right]$ is an approximate $(1-\alpha)$-confidence interval. Further, when n is large, the difference between $s_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ and $\hat{s}_n^2 := \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ is small (indeed, $s_n^2 = (n/(n-1))\hat{s}_n^2$). Hence it is also okay to use $\left[\bar{X}_n - \frac{\hat{s}_n}{\sqrt{n}} z_{\alpha/2}, \bar{X}_n + \frac{\hat{s}_n}{\sqrt{n}} z_{\alpha/2}\right]$ as an approximate $(1-\alpha)$-confidence interval.

Example 169. 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

$$\hat{s}_n^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}_n^2 = \bar{X}_n - (\bar{X}_n)^2 = \bar{X}_n(1 - \bar{X}_n).$$

Hence, an approximate $(1-\alpha)$-CI for p is given by

$$\left[\bar{X}_n - z_{\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}, \bar{X}_n + z_{\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}\right].$$

6. 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

$$\left[\bar{X}_n - z_{\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}, \bar{X}_n + z_{\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}\right].$$

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? 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 $\bar{X}_n$. Then we can find t (numerically) such that $[\bar{X}_n - t, \bar{X}_n + t]$ contains the true value of p in $(1-\alpha)$-proportion of the 10000 trials. Then, $[\bar{X}_n - t, \bar{X}_n + t]$ is a $(1-\alpha)$-CI for p. Alternately, we may try a CI of the form

$$\left[\bar{X}_n - t\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}, \bar{X}_n + t\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}\right].$$

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.

7. Testing problems - first example

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 170. 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 171. 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, $\alpha = \mathbf{P}_{H_0}\{\text{the test accepts } H_1\}$ where we write $\mathbf{P}_{H_0}$ to mean that the probability is calculated under the assumption that H_0 is true. Similarly one define the *power* of the test as $\beta = \mathbf{P}_{H_1}\{\text{the test accepts } H_1\}$. 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 both to be small, but that is not always achievable.

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 $p_k = \mathbf{P}\{X = k\}$ 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

$$\mathbf{P}_{H_0}\{\text{accept } H_1\} = \mathbf{P}_{H_0}\{X \geq k_0\} = \sum_{i=k_0}^{52} p_i.$$

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 $\mathbf{P}\{X \geq 4\} = p_4 + \ldots + p_{52} \approx 0.019$. 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.

8. 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 172. 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.

Example 173. 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 shall use the statistic $\mathcal{T} := \frac{\sqrt{n}(\bar{X}-\mu_0)}{s}$ where $\bar{X}$ and s are the sample mean and sample standard deviation.

(1) In the one-sided test, we accept the alternative hypothesis if $\mathcal{T} > t_{n-1}(\alpha)$.

(2) In the two sided-test, accept the alternative hypothesis if $\mathcal{T} > t_{n-1}(\alpha/2)$ or $\mathcal{T} < -t_{n-1}(\alpha/2)$.

The rationale behind the tests: 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. However, the magnitude depends on the standard deviation and hence we divide by s (if we knew σ we would divide by that). Another way to see that this

is reasonable is that $\mathcal{T}$ does not depend on the units in which you measure X_is (whether X_i are measured in meters or centimeters, the value of $\mathcal{T}$ does not change).

The significance level is α: The question is where to draw the threshold. We have seen before that *under the null hypothesis* $\mathcal{T}$ has a t_{n-1} distribution. Recall that this is because, if the null hypothesis is true, then $\frac{\sqrt{n}(\bar{X}-\mu_0)}{\sigma} \sim N(0,1)$, $(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$ and the two are independent. Thus, the given tests have significance level α for the two problems.

Remark 174. 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, we did not construct confidence interval for μ, but you should do so and check that it is closely related to the one-sided tests above.

9. 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 175. 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 176. 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 $\mathcal{T} = \sqrt{\frac{1}{n} + \frac{1}{m}} \left(\frac{\bar{X} - \bar{Y}}{S} \right)$. 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,

$$\bar{X} \sim N(\mu_1, \sigma_1^2/n),$$

$$\bar{Y} \sim N(\mu_2, \sigma_2^2/m),$$

$$\frac{(n-1)}{\sigma^2} s_X^2 \sim \chi_{n-1}^2,$$

$$\frac{(m-1)}{\sigma^2} s_Y^2 \sim \chi_{m-1}^2$$

and the four random variables are independent. From this, it follows that $(m+n-2)S^2$ has χ_{n+m-2}^2 distribution. *Under the null hypothesis* $\frac{1}{\sigma}\sqrt{\frac{1}{n} + \frac{1}{m}}(\bar{X} - \bar{Y})$ has $N(0,1)$ distribution and is independent of S. Taking ratios, we see that $\mathcal{T}$ has t_{m+n-2} distribution (under the null hypothesis).

10. 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

$$\left[\bar{X}_n - z_{\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}, \bar{X}_n + z_{\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}\right].$$

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

$$\bar{X}_n - z_{\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} \leq p_0 \leq \bar{X}_n + z_{\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}$$

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

$$\bar{X}_n - z_{\alpha/2}\frac{s_n}{\sqrt{n}} \leq \mu_0 \leq \bar{X}_n + z_{\alpha/2}\frac{s_n}{\sqrt{n}}.$$

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.

11. 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 H_0 : the die is fair, versus H_1 : the die is unfair[15].

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 $E_j = \mathbf{E}[O_j] = \frac{n}{6}$ 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.

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 http://galton.uchicago.edu/~lalley/Courses/312/PoissonProcesses.pdf.

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[16]

[15]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.

[16]When a new theory is proposed, it should prove itself and is put in the alterntive hypotheis, but here we take it as null.

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 $\hat{\lambda} = \bar{X}_n$ 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, 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.

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). 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 $\mathbf{P}\{T_n \leq x\} \to \int_0^x f_{k-1}(u)du$ 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 values (does not matter what the values are, say $t_1, t_2, \ldots, t_k$) with probabilities $p_1, \ldots, p_k$. 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.

12. 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 177. 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 178. 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.

Testing independence in bivariate normal: We shall not discuss this problem in detail but instead quickly give some indicators and move on. Here we have (X_i, Y_i) i.i.d bivariate normal random variables with $\mathbf{E}[X] = \mu_1$, $\mathbf{E}[Y] = \mu_2$, $\text{Var}(X) = \sigma_1^2$, $\text{Var}(Y) = \sigma_2^2$ and $\text{Corr}(X, Y) = \rho$. The testing problem is $H_0 : \ \rho = 0$ versus $H_1 : \ \rho \neq 0$. (Remember that if (X, Y) is bivariate normal, then X and Y are independent if and only if X and Y are uncorrelated.

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 away from 0. To decide the threshold we need the distribution of r_n under the null hypothesis.

Fisher: Under the null hypothesis, r_n^2 has Beta$(\frac{1}{2}, \frac{n-2}{2})$ distribution.

Using this result, we can draw the threshold for rejection using the Beta distribution (of course the explicit threshold can only be computed numerically). 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.

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 $\mu(i,j) = \mathbf{P}\{X = x_i, Y = y_j\}$ 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.

13. Regression and Linear regression

Let (X_i, Y_i) be i.i.d random variables. For example, we could pick people at random from a population and measure their height (X) and weight (Y). One question of interest is to predict the value of Y from the value of X. This may be useful if Y is difficult to measure directly. For instance, X could be the height of a person and Y could be the xxx

In other words, we assume that there is an underlying relationship $Y = f(X)$ for an unknown function f which we want to find. From a random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ we try to guess the function f.

If we allow 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 only if we assume that all X_i are distinct which happens if X has a continuous distribution). This is not a good predictor, because the next data point (U, V) will fall way off the curve. We have found a function that "predicts" well all the data we have, but not for a future observation!

Instead, we fix a class of functions, for example the collection of all linear functions $y = mx + c$ where $m, c \in \mathbb{R}$ and within this class, find the best fitting function.

Remark 179. 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 materiaal versus temperature).

(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 tantamount to fitting a quadratic polynomial curve for X, Y.

In short, multiple linear regression along with non-linear transformations of the individual variables, the class of functions f is greatly extended.

Finding the best linear fit: We need a criterion for deciding the "best". 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

$$\hat{\beta} = \frac{\sum_{k=1}^n (X_k - \bar{X}_n)(Y_k - \bar{Y}_n)}{\sum_{k=1}^n (X_k - \bar{X}_n)^2} = \frac{s_{X,Y}}{s_X^2}, \qquad \hat{\alpha} = \bar{Y}_n - \hat{\beta}\bar{X}_n$$

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

$$\sum_{k=1}^n \left((Y_k - \bar{Y}) - \beta(X_k - \bar{X})\right)^2$$

and then we simply set $\hat{\alpha} = \bar{Y} - \beta\bar{X}$. Let[17] $Z_k = \frac{Y_k - \bar{Y}}{X_k - \bar{X}}$ and $w_k = (X_k - \bar{X})^2 / s_X^2$. Then,

$$\sum_{k=1}^n \left((Y_k - \bar{Y}) - \beta(X_k - \bar{X})\right)^2 = s_X^2 \sum_{k=1}^n w_k \left(Z_k - \beta\right)^2.$$

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.

[17] We are dividing by $X_k - \bar{X}$. What if it is zero for some k? But note that in the expression $\sum \left((Y_k - \bar{Y}) - \beta(X_k - \bar{X})\right)^2$, 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).

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 $Y_i = \alpha + \beta X_i + \epsilon_i$ 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 180 (Hubble's 1929 experiment on the recession velocity of nebulae and their distance to earth)**.** Hubble collected the following data that I took from http://lib.stat.cmu.edu/DASL/Datafiles/Hubble.html. 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 $\sum_i (Y_i - Z_i)^2 = 1.1934$.

(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 https://www.cfa.harvard.edu/~dfabricant/huchra/hubble.plot.dat!).

Example 181. I have taken this example from the wonderful compilation of data sets by A.P.Gore, S.A.Paranjpe, M.B.Kulkarni, available at http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedataboo databook.html. 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 agains 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 $\bar{X} = 4.5$, $\bar{W} = -5.25$, std$(X) = 2.45$, std$(W) = 2.52$ and 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.

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 $\text{SS}_0 = \sum_{k=1}^{n}(Y_k - \bar{Y})^2 = (n-1)s_Y^2$). 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 $\text{SS}_1 = \sum_{k=1}^{n}(Y_k - \hat{\alpha} - \hat{\beta}X_k)^2$. Then, $R_1^2 = 1 - \frac{SS_1}{SS_0}$.

(3) Fit a quadratic function. The the residual sum of squares is $\text{SS}_2 = \sum_{k=1}^{n}(Y_k - \hat{\alpha} - \hat{\beta}X_k - \hat{\gamma}X_k^2)^2$ 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 $\bar{SS} = SS/(n-1-p)$. In particular, $\bar{SS}_0 = \frac{SS_0}{n-1} = s_Y^2$. Then define the adjusted R^2 as $\bar{R}^2 = 1 - \frac{\bar{SS}}{\bar{SS}_0}$.

In particular, $\bar{R}_0^2 = R_0^2$ as before. But $R_1^2 = 1 - \frac{SS_1/(n-2)}{SS_0/(n-1)}$. Note that $\bar{R}^2$ 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 $\bar{R}^2$ continues to increase and stop the moment it decreases.

Example 182. 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 $Y_i = 2X_i + \epsilon_i$. 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.

Plan of lectures

02/08	Intro. Probability space defns.
03/08	Examples of discrete prob spaces
06/08	Infinite sums
09/08	Basic rules of probability; Inclusion-exclusion
10/08	Bonferroni's inequalities. Combinatorial examples
13/08	–
16/08	–These three days, have them go over lots of combinatorial problems
17/08	–
20/08	Probability distributions. Binomial, Poisson, Geometric, Hypergeometric
23/08	Continuous CDFs and densities
24/08	Normal, Exponential and Gamma, Uniform and Beta, Cauchy
27/08	Padding
30/08	Expectation, variance, covariance
31/08	Inequalities - Cauchy-Schwarz, Jensen's, Markov, Chebyshev
03/09	Joint distributions. Change of variable formula.
06/09	Independence. Conditional probability.
07/09	Examples.
10/09	
13/09	
14/09	
17/09	
20/09	
21/09	

1	02	Intro. Prob spaces. Examples.
2	05	Infinite sums. Basic rules. Inclusion-exclusion
3	12	[Lost week]
4	19	Distributions with examples. CDF. Uncountable prob spaces. Examples of pdf.
5	26	Examples further. Simulation. Joint distributions. Independence.
6	02	Conditioning. Bayes' rule.
7	09	Expectation, variance, covariance. Inequalities.
8	16	WLLN.
9	23	Normal and Poisson convergence of Binomial.
10	30	Distribution of the sum. Whatever else.
11	07	Basic problems in statistics. Summarizing data.
12	14	Estimation problems.
13	21	Hypothesis testing problems.
14	28	Linear regression and least squares method.
15	04	Kolmogorov-Smirnov and Chi-squared tests.
16	11	Testing for independence. Contingency tables.
17	18	
18	25	
??	??	Random walks. Pólya's urn scheme. Branching processes.

Must include: Coupon collector problem. Banach's matchbox problem. Boltzmann-Gibbs, Fermi-Dirac and Bose-Einstein statistics. Sampling error in polls. Pólya's urn scheme. Ballot problem.

DATE	TARGET	ACTUAL	COMMENTS
02/Aug	Introductory lecture		
05/Aug	Probability space definition		
07/Aug	Examples of probability spaces		
09/Aug	—-		
12/Aug	Balls in bins, Nonexamples		
14/Aug	Countability, Infinite sums	Only countability	
16/Aug	Rules of probability	Infinite sums	
19/Aug	Inclusion exclusion	Countable probability spaces	
21/Aug	Bonferroni's inequalities	Rules of probability	
23/Aug	≈≈≈≈(Independence?)	Inclusion-exclusion	
26/Aug	Random variables, mean, pmf, cdf		
28/Aug	Binomial, Geometric, Poisson		
30/Aug	Simulation		
02/Sep	Conceptual difficulties of continuous distributions		
04/Sep	Continuous distributions		
06/Sep	Normal, exponential, Uniform		
09/Sep	Simulation		
11/Sep	Joint distributions, Independence		
13/Sep	Conditioning		
16/Sep	Conditioning		
18/Sep	Change of variable		
20/Sep	Change of variable		
23/Sep	Mean, variance, covariance		
25/Sep	Cauchy-Shwarz, Markov, Chebyshev		
27/Sep	Weak law of large numbers		
30/Sep	Monte Carlo integration		
02/Oct	Central limit theorem		
04/Oct	–End of probability–		
07/Sep	Statistics - introduction		
09/Oct	Estimation		
11/Oct	Estimation		
14/Sep	Confidence intervals		
16/Oct	Confidence intervals		
18/Oct	Wrap up estimation		
21/Oct	Testing		
23/Oct	Testing		
25/Oct	Testing		
28/Oct	Testing		
30/Oct	Testing		
01/Nov	Regression		
04/Nov	Regression		

Remark 183. Probably losing one week for midterm. But there must be two more weeks at the end. Assuming loss of a couple of classes to holidays, there may be just enough time. But schedule must be adhered to.

APPENDIX B. VARIOUS PIECES

There are many pieces that should be inserted in exercises if they cannot be covered in lectures or tutorials.

- Stirling's formula
- Poisson limit of Binomial
- Banach's matchbox problem
- Coupon collector problem
- Polya's urn scheme (definition)
- Random walk in one and two dimensions
- Gambler's ruin problem
- Ballot problem
- Catalan numbers
- Gamma function
- Beta function
- Branching process
- Integration of e^{-x^2}
- Multidimensional normal integral (at least bivariate)
- Hardy-Weinberg law
- Fisher's explanation of sex-ratios
- Mendel's actual data, falsification?
- Comparing literary styles, with example
- Sample surveys - actual examples?
-