Lecture Notes for the Introduction to Probability Course

Vladislav Kargin

June 9, 2021

Contents

Chapter 1

Combinatorial Probability and Basic Laws of Probabilistic Inference

1.1 What is randomness and what is probability?

"Probability is the most important concept in modern science, especially as nobody has the slightest notion what it means." Bertrand Russell

1.1.1 Setup and the classical definition of probability

Modern probability theory is a result of the scientific revolution which happened during the last 100 years. The basic setup can be seen in Figure 1.1. We have a data generation mechanism that produces outcomes ω. This mechanism is often incredibly complicated and the consensus is that it is difficult or maybe impossible to study it in all details. Instead, we study the *probability model*. This model should tell us what is the probability that the outcome ω belongs to a specific class of outcomes A which we call an event. This is simpler than finding all the details of the data generation mechanism and this is the reason why the theory of probability models is important.

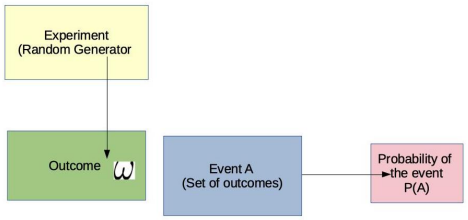


Figure 1.1: Basic setup: Experiment, outcome, event, probability

A very simple example of data generating mechanism (or *experiment*, – we use this two words interchangeably) is tossing a die. The possible outcomes form a set $\Omega = \{1, 2, 3, 4, 5, 6\}$. A possible event is $A = \{1, 4, 5\}$. We say that the event *occurred* in the experiment if the toss of the die resulted in one of 1, 4, or 5.

Another possible "experiment" results in the vector of the stock price data on a particular date. In this case, $\omega = (p_{Apple}, \ldots, p_{Zoom})$. A possible event is that an average of these price vector is greater than the average on the previous day ("stock market is up").

Yet another example is the top image that a search engine provides for the word "Picasso". In this case, the outcome is an image which we can represent as a three-dimensional array of pixel intensities (the third dimension is for color information).

You can see that a random experiment and its outcomes can be very complicated. Our task is simpler. This simply want to establish the basic properties of the probability function P.

The probability theory grew up from the study of simple experiments similar to tossing dice, and the definition of probability function in times of

French mathematician Laplace (who taught mathematics to Napoleon) was very simple. We start by considering this definition and move to a more modern definition due to Kolmogorov.

Definition 1.1.1 (Laplace's definition)**.** If each outcome is equally probable, then the probability of a random event comprised of several outcomes is equal to the number of favorable cases divided by the number of all possible cases.

Essentially, this means that if a random experiment has N *equally-probable* outcomes and the event $A = \{\omega_1, \omega_2, \ldots, \omega_k\}$ consists of k outcomes, then

$$\mathbb{P}(A) = \frac{k}{N}.$$

From the point of view of Laplace's definition, probability is the art of counting all possible and all favorable cases.

Example 1.1.2. Roll a fair die. What is the probability that you get an even number?

Example 1.1.3. Suppose we toss two fair coins. What is the probability that we observe a head and a tail?

The coins are distinguishable and the list of possible outcomes is in fact HH = "H on first coin, H on the second coin", HT, TH an TT. If we assume that these 4 outcomes are equally probable, then there are two favorable outcomes: HT and TH and, by Laplace's definition, the probability is 2/4 = 1/2. This agrees with experiment eventually justifying our assumption.

The assumption that the coins are distinguishable and the four possible outcomes are equiprobable works in the real world which we observe every day. However, if we consider less usual objects, for example, objects in quantum mechanics, then we encounter objects for which this assumption is violated.

In particular, elementary particles called bosons, – for example, photons are bosons, – are indistinguishable. So if we setup an experiment in which we use polarizations of two photons as coins, then possible outcomes in this

experiment will be HH, HT and TT, since we are not able to say which photon is which. Moreover, these three outcomes are equiprobable. For this experiment, Leibnitz' prediction is correct and the probability of HT is 1/3. The probability rules based on this set of assumptions are called the Bose-Einstein statistics. Note that Laplace's definition is OK in this situation. It simply happens that the set of equiprobable outcomes is different.

Another set of probability rules that arise in quantum mechanics is Fermi-Dirac statistics for fermion particles. If the coins behaved like electron spins, then only HT is a possible outcome and the probability to observe it is 1.

However suppose that we can identify the equally probable outcomes. Why is even in this case Laplace's definition not quite sufficient?

The problem starts when we have a very big number of outcomes. In this case we simply cannot find out what is the total number of possibilities. It is essentially infinite. We cannot count the number of possibilities for the images that a search engine gives us. And the probability of each particular image is essentially zero. Yet we want to build a probability model for this experiment, and we want to be able to calculate the probabilities for events such as "an image contain a person's face".

For this reason we allow the probability function to for every event from a large collection of events. And we only require that it satisfies several simple axioms. Whether it is a useful probability function should be defined by comparing the probability with frequencies of the event occurrences observed in data.

1.1.2 Kolmogorov's Axiomatic Probability

We study in this chapter Kolmogorov's axioms for probability function. The second important contribution of Kolmogorov is the concept of conditional probability of two events and and independence of two events. These is something new relative to Laplace's counting methods. However, we will learn about this component later.

The main idea of Kolmogorov's approach is that the probability function can be defined arbitrarily as long as it satisfies a set of axioms. The justi-

fication for a particular choice of the probability function eventually comes from its agreement with experimental data. (Also, it is useful to keep in mind that the axioms can be satisfied by "non-random" models so they are not meant to describe what is "true randomness".)

The standard setup is as follows. We have a set Ω which is the set of possible outcomes of an experiment. A *random event* is a collection of some outcomes, that is, a subset of the set Ω. (Sometimes not all sets of outcomes are allowed to be valid events but only those in a some special class $\mathcal{A}$. So "event" = "allowed collection of outcomes".)

Example 1.1.4. We roll a die. The set of all possible outcomes is $\Omega = \{1, 2, 3, 4, 5, 6\}$. The event "the outcome of a roll is even" is the subset $A = \{2, 4, 6\}$.

The *probability* $\mathbb{P}$ is a function, which assigns a real positive number to every *event* A in a collection of allowed events $\mathcal{A}$.

It is important to assign probability to every event, not only to every outcome, since it is possible that every particular outcome has zero probability but the probability of some events is not zero.

The set Ω with the collection of all possible events $\mathcal{A}$ with probability function $\mathbb{P}$ is called the *probability space*.

The function $\mathbb{P}$ should satisfy several axioms.

Axioms: Let Ω be the set of outcomes. We assign $\mathbb{P}(A)$, the probability of A, to each (possible) event $A \subset S$, in such a way that the following axioms are satisfied:

1. $\mathbb{P}(A) \geq 0$.
2. $\mathbb{P}(\Omega) = 1$.
3. If $A_1, A_2, A_3, \ldots$ is a countable sequence of *pairwise mutually exclusive* events, then
$$\mathbb{P}(A_1 \cup A_2 \cup A_3 \cup \ldots) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

Pairwise mutually exclusive events means that no outcome can be in two or more of A_i simultaneously. These events are also often called disjoint.

The third axiom is formulated for the infinite number of events but it is not very difficult to prove that it also holds for a finite number of disjoint events. So if we have disjoint events $A_1, \dots A_n$, then we have:

$$\mathbb{P}(A_1 \cup A_2 \dots \cup A_n) = \sum_{i=1}^{n} \mathbb{P}(A_i).$$

Example 1.1.5 (Relation to Laplace's definition). From the previous formula, it is easy to see that if a set Ω is finite and consists of N outcomes that have equal probability, then the probability of each outcome must be equal to $p(\omega) = 1/N$. Then the probability of an event is just the number of outcomes in this event multiplied by the probability of each outcome, and we recover Laplace definition:

$$\mathbb{P}(A) \equiv \text{Probability of event } A = \frac{|A| \equiv \text{the number of outcomes in } A}{N \equiv |\Omega| \equiv \text{ the number of possible outcomes}}.$$

Hence, Kolmogorov's definition includes Laplace's definition as a particular case.

The power of the axiomatic approach is that we can use more complicated definitions for the probability function. The only thing is to check that they satisfy axioms.

For example, there is an advantage even if the space of outcomes Ω is finite, since we can deal with non-fair dices and coins by assigning arbitrary probability to outcomes. Of course, the probability of each outcome should be non-negative and they all should add up to 1.

Outcomes are usually denoted ω in probability theory. So if the probability of an outcome ω is $p(\omega)$ then we only need to check that $p(\omega) \geq 0$ for all $\omega \in \Omega$, and that $\sum_{\omega \in \Omega} p(\omega) = 1$.

Then the probability of an event A is

$$\mathbb{P}(A) = \sum_{\omega \in A} p(\omega). \tag{1.1}$$

This definition can be extended to countable spaces Ω, since the sum makes sense even is we sum a countable sequence. In this case, it is a sum of a series that you studied in calculus.

In the uncountable case formula (1.1) does not make sense as a series. For example you cannot sum $p(\omega)$ if ω are all points of the interval $[0, 1]$. You will run into a problem of enumerating the points ω and defining the limit of the sum. However, even in this case there is a next level of sophistication, which allows us to define the probability that generalize the area definition.

It is natural define a probability of an event A by the formula:

$$\mathbb{P}(A) = \int_{\omega \in A} p(\omega) d\omega. \tag{1.2}$$

In this formula $p(\omega)$ is not a probability of outcome ω but something which is called a *probability density*. This is by analogy with physics: we cannot talk about a mass of a point of a continuous material but we can talk about a mass density at the point.

The mathematical content of this is that we can define probability of events using formula (1.2) provided that it satisfies all axioms and so it is a valid probability function. This will impose some conditions on the density $p(\omega)$, which we will study later.

1.1.3 More about set theory and indicator functions

We have seen that the probability is a function $\mathbb{P}$ on events, and events are subsets of a set Ω. This set Ω consists of all possible outcomes of an experiment and often called probability space.

Example 1.1.6. Example: $\Omega = \{1, 2, 3, 4, 5, 6\}$ for a toss of a die.

The statement about events are often formulated in the language of the set theory language because it is very concise and unambiguous. For calculations, it is also very helpful to learn about the indicator functions of events.

As we defined it above, an event is a collection of some outcomes. For example, $A = \{1, 3, 5\}$ is an event in the die rolling experiment. We say that an event A occurred if the experiment resulted in an outcome ω which belongs to the event A. For example, if the die rolling experiment gave 1 as an outcome then the event $A = \{1, 3, 5\}$ occurred.

Example 1.1.7 (Die Rolling). $\Omega = \{1, 2, 3, 4, 5, 6\}$

Let $A = \{\text{roll number is odd}\} = \{1, 3, 5\}$.

Let $B = \{\text{roll number is } \leq 2\} = \{1, 2\}$.

Let $C = \{1, 3\}$.

By definition, B is a *subset* of A (denoted $B \subset A$) if every element of B is also in A. Intuitively, if event B occurred in an experiment then necessarily also event A occurred. In our example $C \subset A$ but $B \not\subset A$.

The *union* of A and B (denoted $A \cup B$) is the set of all points which ar in A or in B or in both. Intuitively, the event $A \cup B$ means that either event A, or event B, or both occurred in the experiment. In our example $A \cup B = \{1, 2, 3, 5\}$ and $A \cup C = A$ (because $C \subset A$ so C cannot contribute any new elements to A).

The *intersection* of A and B (denoted $A \cap B$) is the set of all outcomes that are both in A and in B. That is, the event $A \cap B$ means that both A and B occurred in the experiment. In our example, $A \cap B = \{1\}$, $A \cap C = C$ (because $C \subset A$).

The *complement* of A (denoted either as $\overline{A}$ or as A^c) is the set of points that are in Ω but *not* in A. Probabilistically, the event A^c means that event A did NOT occur in the experiment. In our example $A^c = \{2, 4, 6\}$.

Sets A and B are called *mutually exclusive* (or *disjoint*) if $A \cap B = \emptyset$, where $\emptyset$ denotes the empty set (a set with no elements). Probabilistically, events A and B consist of different outcomes and cannot happen together. Disjoint sets have no elements in common. In our example, there are no disjoint sets since all of them have the element 1 in common.

However, A and A^c are disjoint. This is always true: a set is disjoint from its complement.

This is very flexible terminology. We can express many events using it. For example, suppose that we have three events, A, B, and C, and we are interested in the event that exactly two of them occur in the experiment. We can write this event as

$$(A \cap B \cap C^c) \cup (A \cap B^c \cap C) \cup (A^c \cap B \cap C).$$

Note that the events in this union are disjoint so we can use the additivity axiom and write that the probability of this event is

$$\mathbb{P}(A \cap B \cap C^c) + \mathbb{P}(A \cap B^c \cap C) + \mathbb{P}(A^c \cap B \cap C).$$

Now let us define another convenient tool.

The *indicator function* of an event A is a function on outcomes that takes value 1 if the outcome is in the event and 0 otherwise. The indicator function of an event A is denoted I_A (or sometimes $\mathbb{1}_A$). Formally,

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

For event $A = \{1, 3, 5\}$, the indicator function works as follows: $I_A(1) = 1$, $I_A(2) = 0$, $I_A(3) = 1$, $\mathbb{1}_A(4) = 0$, $I_A(5) = 1$, $I_A(6) = 0$. You can think about I_A as a vector of 0-s and 1-s where coordinates are outcomes. So in our example we have vector $I_A = (1, 0, 1, 0, 1, 0)$ and the 1 in position 3 means that $\omega = 3$ belongs to the event A. Similarly $I_B = (1, 1, 0, 0, 0, 0)$ and $I_C = (1, 0, 1, 0, 0, 0)$.

This tool is very convenient both for calculations and if one wants to represent events on a computer.

In terms of indicator functions, $B \subset A$ if and only if $I_B \leq I_A$ in the sense of functions, that is for every outcome ω, $I_B(\omega) \leq I_A(\omega)$.

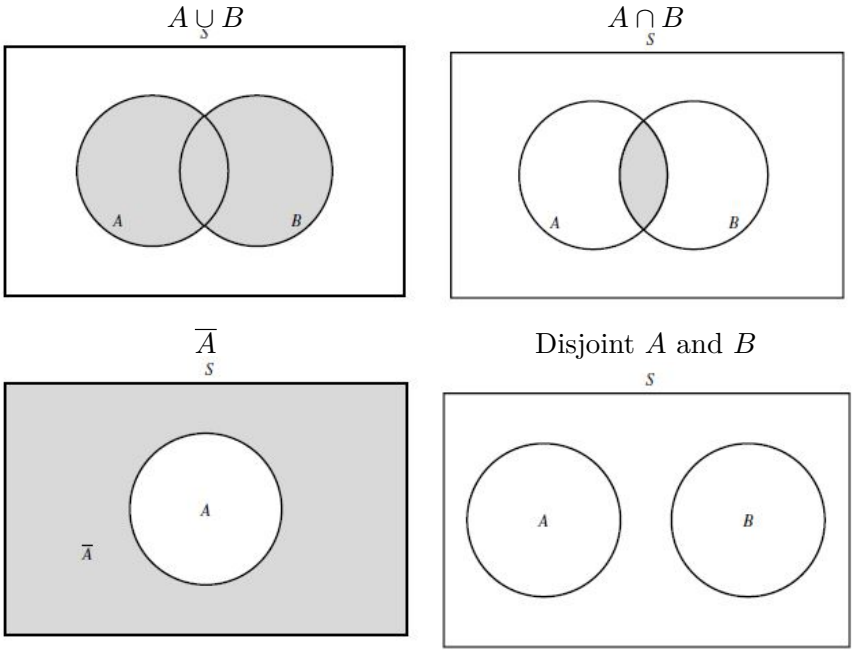
Also, in terms of indicator functions: $I_{A \cap B} = I_A I_B$, That is, we simply take the product of these two functions.

The formula for the union is a bit more complicated: $I_{A \cup B} = I_A + I_B - I_A I_B$. You can easily check it by verifying all possible cases. For example if both $\omega \in A$ and $\omega \in B$ then $I_{A \cup B}(\omega) = 1$ by definition of the union. At the same time $I_A(\omega) + I_B(\omega) - I_A(\omega) I_B(\omega) = 1 + 1 - 1 = 1$, which shows that the formula is valid in this particular case.

For the complement, $I_{A^c} = I - I_A$ where we took the liberty to write I for I_Ω. (This is simply the function that takes value 1 on every outcome.)

Finally, in terms of indicator functions A and B are disjoint if and only if $I_A I_B = 0$.

A Venn Diagram is a way to visualize simple sets.



Laws of the set theory The operations of the set theory satisfy a bunch of rules, which are sometimes useful in derivations of various probability formulas.

One set of rules is De Morgan's Laws:

1.
$$(A \cap B)^c = A^c \cup B^c$$

2.
$$(A \cup B)^c = A^c \cap B^c$$

They can be checked by considering all possibilities for the outcome ω (the first possibility $\omega \in A$ and $\omega \in B$, the second is $\omega \in A$ and $\omega \notin B$ and two other possibilities) and verifying that in each case ω is an element of the set on the left-hand side of the equality if and only if it is an element of the set on the right-hand side.

Alternatively, it can be checked by using the indicator functions. (It is helpful here that we know the distributive law for functions.)

For example for the first rule, we have

$$
\begin{aligned}
I_{(A\cap B)^c} &= I - I_{A\cap B} = I - I_A I_B \\
I_{A^c\cup B^c} &= I_{A^c} + I_{B^c} - I_{A^c} I_{B^c} \\
&= I - I_A + I - I_B - (I - I_A)(I - I_B) \\
&= I - I_A + I - I_B - (I - I_A - I_B + I_A I_B) \\
&= I - I_A I_B.
\end{aligned}
$$

Another set of rules is Distributive Laws:

1.
$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

2.
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

This formulas can be checked by considering all possibilities for an outcome ω. (That is, first we check if the law is satisfied if $\omega \in A$, $\omega \in B$, $\omega \in C$, then we check if it is satisfied if $\omega \in A$, $\omega \in B$, $\omega \notin C$ and so on. Eight possibilities to check.)

Alternatively these identities can be checked by using indicator functions, through a straightforward if lengthy calculation. For example, for the second distributive rule, we can use the expression for the indicator functions of a union and an intersection and calculate the left hand side of the formula as

$$
\begin{aligned}
\mathbb{1}_{A\cup(B\cap C)} &= \mathbb{1}_A + \mathbb{1}_{B\cap C} - \mathbb{1}_A \mathbb{1}_{B\cap C} \\
&= \mathbb{1}_A + \mathbb{1}_B \mathbb{1}_C - \mathbb{1}_A \mathbb{1}_B \mathbb{1}_C
\end{aligned}
$$

At the same time, the right hand side is

$$
\begin{aligned}
\mathbb{1}_{(A\cup B)\cap(A\cup C)} &= \mathbb{1}_{(A\cup B)} \mathbb{1}_{(A\cup C)} \\
&= (\mathbb{1}_A + \mathbb{1}_B - \mathbb{1}_A \mathbb{1}_B)(\mathbb{1}_A + \mathbb{1}_C - \mathbb{1}_A \mathbb{1}_C) \\
&= \mathbb{1}_A + \mathbb{1}_A \mathbb{1}_C - \mathbb{1}_A \mathbb{1}_C + \mathbb{1}_B \mathbb{1}_A + \mathbb{1}_B \mathbb{1}_C - \mathbb{1}_B \mathbb{1}_A \mathbb{1}_C \\
&- \mathbb{1}_A \mathbb{1}_B - \mathbb{1}_A \mathbb{1}_B \mathbb{1}_C + \mathbb{1}_A \mathbb{1}_B \mathbb{1}_C \\
&= \mathbb{1}_A + \mathbb{1}_B \mathbb{1}_C - \mathbb{1}_A \mathbb{1}_B \mathbb{1}_C,
\end{aligned}
$$

which is the same. In this calculation in the third equality we used repeatedly that the square of an indicator function equals to itself: $(1_A)^2 = 1_A$, $(1_B)^2 = 1_B$, $(1_C)^2 = 1_C$.

1.1.4 What is randomness?

Laplace's or Kolmogorov's definitions of probability do not say anything about what is randomness and what is the true nature of probability function. It is a very difficult question and mathematicians still struggle to find a good answer. An Austrian mathematician Richard von Mises (do not confuse with his older brother, economist Ludwig von Mises) suggested that randomness can only be defined for a random generator (experiment) that can produce an infinite sequence of outcomes and that the probability of an outcome is the limiting frequency of the outcome in this infinite sequence.

The part about the probability as a frequency is very useful. If we have a model for a function $\mathbb{P}(A)$ then we can perform the experiment many times and check whether the frequency of the event A is close to the probability predicted by the model.

But what about true nature of randomness?

Let an experiment produce only outcomes 0 and 1. Is the sequence of outcomes $S = 0101010101\ldots$ random? It does not look so, although the limiting frequency of the outcome 1 is well defined and equal 1/2. Von Mises suggested that the sequence is random only if it is not possible to select a subsequence that will have a different frequency than the original. (In the case of sequence S we could select all outcomes in odd places and the frequency of 1 would become 0.) Crucially, the requirement is that the selection should be done with a "computable" function. This lead to a study of computable functions and complexity of computability.

The problem in this case is that this definition that it is not possible to construct a truly random generator. We could only rely on physical processes which are postulated to be random. For example, we can rely on quantum-mechanical processes.

Recently, computer scientists started to study pseudo-randomness. Very vaguely, a pseudo-random generator depends on a seed s, the original state

of the generator. For example, we can define a rather bad generator

$$f(N, s) = sN \mod q,$$

where q is a big integer and s is a seed. At time N it outputs $\omega_N = f(N, s)$.

A pseudo-random generator is good (close to the true physical random generator) if from an observation of a short portion of the sequence of outcomes $S = \{\omega_1, \omega_2, \ldots\}$, it is computationally hard to recover the seed s. It is an area of research in computer science, and it is a tough area of research because it is difficult to prove that a particular problem is computationally hard. (It is a $1,000,000$ dollar problem to show that a particular class of problems (NP-complete problems) is computationally difficult. And it is not clear how hard is the problem of factorization of integer numbers, although most of cryptosystems used in financial and military applications rely on the hardness of this problem.)

We will not study this problem in this course.

1.2 Discrete Probability Models

A *discrete probability space* Ω is a probability space that contains a finite or countable number of points.

Example 1.2.1. Rolling a die. Here $\Omega = \{1, 2, 3, 4, 5, 6\}$

Example 1.2.2. Flip a coin until the first "head" occurs and record the number of "tails" that occurred before that flip. Here $\Omega = \{0, 1, 2, \ldots\}$, where, say, $\omega = 2$ means that the outcome was TTH - two tails before the first head.

For discrete probability spaces, if we know the probability of each outcome, then we can compute the probability of all events.

Definition 1.2.3. The *probability mass function* ("pmf") $p(\omega)$ is the function on the discrete probability space Ω which equals to the probability of outcome ω.

Then, for any event A we have:

$$\mathbb{P}(A) = \sum_{\omega \in A} p(\omega).$$

(This follows from Axiom 3 since the event A is a countable disjoint union of the sets that consist of a single outcome.)

Remark: often, the name "probability mass function" (or simply "pmf") refers to a distribution of a discrete random variable. We will talk about the pmf of a discrete random variable in the next section.

For the probability mass function $p(\omega)$, we should ensure that it is non-negative for every outcome, and that the sum over all outcomes equals to one: $\sum_{\omega \in \Omega} p(\omega) = 1$. Otherwise, Kolmogorov's Axioms would be violated.

In practice, this is often achieved by assigning non-negative weights to outcomes and then dividing each weight by the sum of all weights.

Example 1.2.4 (Laplace's pmf). If we assign weight 1 to each outcome and the total number of outcomes is N, then every outcome will be assigned the same weight $p(\omega) = 1/N$. This is the case when all outcomes are equally probable. For example if we have a "fair" die then the probability of the event $A = \{1, 3, 5\}$ is

$$\mathbb{P}(A) = p(1) + p(2) + p(3) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}.$$

Example 1.2.5. Suppose we have a "loaded" die and suppose that according to our model the probability of each outcome should proportional to the number that the die shows. Then we assign weights to each outcome in $\Omega = \{1, 2, 3, 4, 5, 6\}$ as follows $w(1) = 1$, $w(2) = 2$, ..., $w(6) = 6$. The sum of these weights is $1 + 2 + \ldots + 6 = 21$ and so we can define the probability mass function as

$$
\begin{aligned}
p(1) &= \frac{w(1)}{\sum_\omega w(\omega)} = \frac{1}{21}, \\
p(2) &= \frac{w(2)}{21} = \frac{2}{21}, \\
&\dots, \\
p(6) &= \frac{6}{21}.
\end{aligned}
$$

Then, the probability of the event $A = \{1, 3, 5\}$ is

$$
\mathbb{P}(A) = \frac{1}{21} + \frac{3}{21} + \frac{5}{21} = \frac{9}{21} = \frac{3}{7}.
$$

Here are some simple probability rules that can be easily derived from Kolmogorov's Axioms:

1. *Complement Rule:* $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.
2. *Monotonicity:* If $A \subset B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$.
3. *Additive rule:* $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

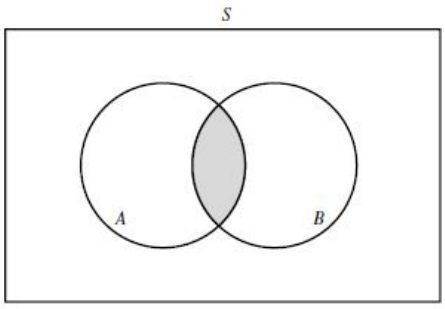
These rules are true for all probability spaces, but their proof is especially simple in the case of discrete spaces. For example, for the complement rule we have:

$$
\mathbb{P}(A) + \mathbb{P}(A^c) = \sum_{\omega \in A} p(\omega) + \sum_{\omega \notin A} p(\omega) = \sum_{\omega \in \Omega} p(\omega) = 1.
$$

For the additive rule we have:

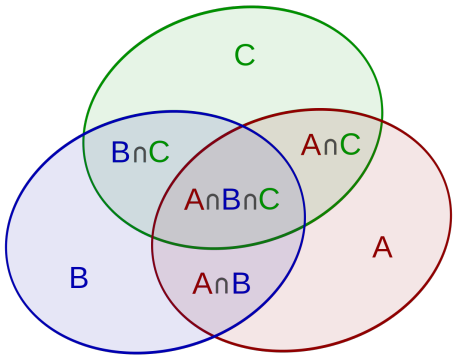
$$
\begin{aligned}
\mathbb{P}(A) + \mathbb{P}(B) &= \sum_{\omega \in A} p(\omega) + \sum_{\omega \in B} p(\omega) \\
&= \sum_{\omega \in A \cup B} p(\omega) + \sum_{\omega \in A \cap B} p(\omega) \\
&= \mathbb{P}(A \cup B) + \mathbb{P}(A \cap B),
\end{aligned}
$$

where to write the second equality we noted that in the sum in the first line the outcomes in the intersection of A and B were counted twice.



It is possible to write more general Inclusion-Exclusion formulas for unions of more than 2 events. For example

$$\mathbb{P}(A \cup B \cup C) = \mathbb{P}(A)+\mathbb{P}(B)+\mathbb{P}(C)-\mathbb{P}(A \cap B)-\mathbb{P}(A \cap C)-\mathbb{P}(B \cap C)+\mathbb{P}(A \cap B \cap C).$$



In general, an extended additive law holds for any n-events A_1, A_2, ...,A_n:

$$\begin{aligned}\mathbb{P}(A_1 \cup A_2 \cup \ldots \cup A_n) &= \mathbb{P}(A_1) + \mathbb{P}(A_2) + \ldots + \mathbb{P}(A_n) \\ &\quad - \mathbb{P}(A_1 \cap A_2) - \ldots - \mathbb{P}(A_{n-1} \cap A_n) \\ &\quad + \mathbb{P}(A_1 \cap A_2 \cap A_3) + \ldots \\ &\quad + (-1)^{n-1}\mathbb{P}(A_1 \cap A_2 \cap \ldots \cap A_n).\end{aligned}$$

1.3 Sample-point method (Combinatorial probability)

To summarize, in the case of discrete spaces one method to calculate the probability of an event A is as follows:

- Determine the probability of each outcome ω in event A.
- Sum these probabilities over all $\omega \in A$.

This will give the probability of the event A.

Let us return to the situation when all outcomes are equally probable. In this case, the probability of each outcome equals to 1 divided by the total number of possible outcomes, and we get the Laplace rule, from which we started this course:

$$P(A) = \frac{\text{Number of outcomes in } A}{\text{Total number of possible outcomes}}$$

Example 1.3.1. Toss three "fair" coins. What is the probability of the event $A = \{$We got at least 2 "heads"$\}$

The set of all outcomes $\Omega = \{HHH, HHT, HTH, \ldots, TTT\}$. The first letter can be chosen in two ways: H or T, then the second letter – also in two ways, and then the third one – also in two ways. So the total number of possibilities $2 \times 2 \times 2 = 2^3 = 8$. How many favorable outcomes? We can list all of them: $A = \{HHT, HTH, THH, HHH\}$. So there are 4 of them and the probability of A is $\mathbb{P}(A) = 4/8 = 1/2$.

Example 1.3.2. Roll two fair dice. Consider the event "The sum of the two dice is 7". What is the probability of this event?

The probability space is
$\Omega = \{(1,1), (1,2), \ldots, (6,5), (6,6)\}$
– 36 ordered pairs in total. The pairs that sum to 7 are $(1,6), (2,5), \ldots, (6,1)$. Six of them in total. So the probability is $6/36 = 1/6$.

Example 1.3.3 (Chevalier de Méré's problem of points). Two players A and B play a fair game such that the player who wins a total of 6 rounds first wins a prize. Suppose the game unexpectedly stops when A has won a total of 5 rounds (so have 5 points) and B has won a total of 3 rounds (so have 3 points). How should the prize be divided between A and B?

This is, of course, a more difficult problem. In a sense, the history of probability theory starts from the solution of this problem given by two great mathematicians.

This problem was suggested to a French mathematician and philosopher Pascal by his friend mathematician Roberval, who was asked for advice by a writer, courtier and avid gambler Chevalier de Méré. This was in 1654, in the time of Louis XIV, the King-Sun of France, in the time when D'Artagnan was already 50-year old. Pascal was also an inventor, a scientist, and a religious dissident. He built a first-ever mechanical calculator when he was around 20. For this reason, one of the modern computer languages was named after him. Pascal has also a unit of pressure named after him because Pascal discovered atmospheric pressure. (He implemented an idea due to Evangelista Torricelli.)

Roberval gave the following solution to the problem of points. Assume that the prize is divided according to the relative probabilities of A and B to win the game, had they continued to play.

Consider how the game could develop further. The possibilities are $\Omega = \{A_1, B_1A_2, B_1B_2A_3, B_1B_2B_3\}$, where A_1 means, for example, that A wins next round and since he won 6 rounds at that time, the game is over and he is the winner.

Only one case is favorable for player B: $B_1B_2B_3$. So the probability of B to win the game is $1/4$ and he should get 25% of the prize.

Pascal objected to this solution because it is clear that these outcomes

Figure 1.3: Blaise Pascal and Pierre Fermat

are not equally probable: it is clear that the outcome A_1 is more probable than $B_1B_2B_3$. So he developed a different solution, and decided to check if it makes sense by sending a letter to Fermat, who lived in the south of France, in Toulouse. Fermat was a judge and did his mathematical research in his spare time.

Pascal was indeed correct but Pascal used a method different from the sample point method, something like a backward induction: start with the situation when there is only one game that will definitely decide the winner (this happens when both won 5 games), then consider the case when there are at most two games that decide who are the winner, and so on.

Here we are more interested in Fermat's method of solution (which Fermat sent to Pascal a couple of weeks later).

The equally probable outcomes could be obtained if the players played 3 rounds of the game after interruption:

$$\begin{aligned}\Omega = \{&A_1A_2A_3, A_1A_2B_3, A_1B_2A_3, A_1B_2B_3,\\ &B_1A_2A_3, B_1A_2B_3, B_1B_2A_3, B_1B_2B_3\}\end{aligned}$$

Intuitively, these outcomes are all equally probable since the game is fair. Player B would win the prize in only one of these cases: $B_1B_2B_3$.

Hence the probability of his win is only $1/8 = 12.5\%$.

This solution is correct although it is a bit controversial: in the reality, the players have no need to play 3 games to determine the result of the game. However, there is no other way to introduce the cases with equal probabilities.

The upshot of all these examples is that in principle, it is easy to calculate the probability of an event if we know that the experiment outcomes are equally probable. We only need to calculate the number of the outcomes in the event.

However, in mathematics exact counting is often difficult.

Example 1.3.4. How many ways are to tile a $2m \times 2n$ chessboard with dominoes?

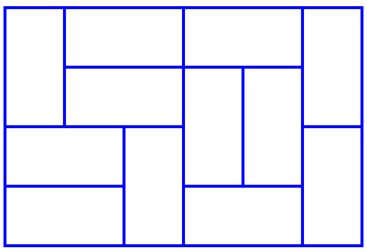


Figure 1.4: A domino tiling of a 4×6 board.

The number of the domino tilings of the $2m \times 2n$ chessboard is given by a strange formula:

$$4^{mn} \prod_{j=1}^{m} \prod_{k=1}^{n} \left(\cos^2 \frac{j\pi}{2m+1} + \cos^2 \frac{k\pi}{2n+1} \right)$$

For example for $m = 2$ and $n = 3$,

$$\begin{aligned} &4^6(\cos^2 36^\circ + \cos^2 25.71^\circ)(\cos^2 72^\circ + \cos^2 25.71^\circ) \\ &\times(\cos^2 36^\circ + \cos^2 51.43^\circ)(\cos^2 72^\circ + \cos^2 51.43^\circ) \\ &\times(\cos^2 36^\circ + \cos^2 77.14^\circ)(\cos^2 72^\circ + \cos^2 77.14^\circ) \\ &= 4^6(1.4662)(.9072)(1.0432)(.4842)(.7040)(.1450) \\ &= \mathbf{281} \end{aligned}$$

As another example, consider "meander curves": closed curves that cross a straight line in a specific number of places. An exact simple formula for the number

of these curves is not known. They can be enumerated only by an exhaustive computer search.

Fortunately, in many cases it is possible to count number of outcomes in an event by using relatively easy formulas. In other cases, when exact formulas are unavailable or unknown, it is still possible to count approximately.

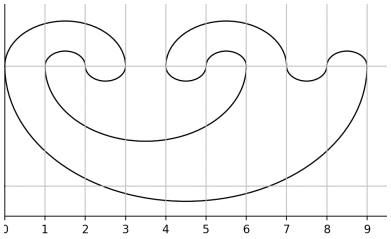


Figure 1.5: A meander curve with 10 crossings.

Here we just remind you about some tools from combinatorics and establish some notations for useful numbers: binomial and multinomial coefficients.

Basics of counting:

1. Multiplication Rule

If we select an element from a set having m elements and another element from a set having n elements, then there are mn possible pairs that could be selected. *Note*: This multiplication rule extends to any number of sets.

Example 1.3.5. We roll a 6-sided die, a 4-sided die, and a 20-sided die. The total number of possible outcomes is $6 \times 4 \times 20$.

2. Permutations

A *permutation* is an *ordered* arrangement of a specific number of distinct objects. (Order matters!) For example, if our objects are numbers from 1 to 6 then (3146) is a permutation of four of these objects, and it is different from (1346).

We denote the *number* of possible permutations of n distinct objects, taken r at a time, as P_r^n. The first object can be chose in n ways, the second, – in $n-1$ ways and the r-th, – in $n-r+1$ ways. So, by the multiplication rule, the total number of permutations is

$$P_r^n = n \times (n-1) \times (n-2) \times \ldots \times (n-r+1) = \frac{n!}{(n-r)!},$$

where $n!$ denotes the factorial of n: $n! = n \times (n-1) \times (n-2) \ldots \times 2 \times 1$. By convention, $0! = 1$.

In the example above we have $P_4^6 = 6 \times 5 \times 4 \times 3 = 360$.

In particular, if we are interested in the number of permutations of all n objects, then we have $P_n^n = n!$ permutations.

Example: There are 24 members of a student club. Four members will be chosen to be president, VP, secretary, and treasurer. How many different arrangements of officers are possible?

$$P_4^{24} = 24 \times 23 \times 22 \times 21.$$

3. Permutations when not all objects are distinct

In the previous example all objects were distinct. What happens if it is not the case?

Example 1.3.6. How many distinct words can be formed by permuting the letters of the name "BOBBY"?

We have total 5! permutations of all letters. But permutations of letters B among themselves do not bring anything new. So in our original set of 5! permutation there are some repetitions. For example if we will look at B-s as distinct and denote them B_1, B_2, B_3, then the permutations $B_1OB_2B_3Y$ and $B_2OB_1B_3Y$ represent the same word. Since there would be $3! = 6$ ways to permute letters B if they were distinguishable, hence every distinct word corresponds in fact to 6 permutations in our set of 5! permutations. Therefore, the true number of distinct words is

$$\frac{5!}{3!} = 5 \times 4 = 20.$$

By extending this kind of argument to a more general situation one can show the following theorem.

Theorem 1.3.7. *Consider permutations of n objects, in which n_1 are of the first type, n_2 are of the second type, …, n_s are of the s-th type, and $n_1 + n_2 + \ldots + n_s = n$ (that is, there are no objects of other types). The number of these permutations is:*

$$\frac{n!}{n_1!n_2!\ldots n_s!}$$

In Example 1.3.6 about BOBBY,

$$\binom{5}{3,1,1} = \frac{5!}{3!1!1!} = \frac{5!}{3!}$$

Definition 1.3.8. A *multinomial coefficient* is defined as:

$$\binom{n}{n_1, n_2, \ldots, n_s} = \frac{n!}{n_1! n_2! \ldots n_s!}$$

In the following we will often encounter one specific case of the multinomial coefficient.

Definition 1.3.9. In the case of only two types: $s = 2$, a multinomial coefficient is called a *binomial coefficient*. It is denoted

$$\binom{n}{k} := \binom{n}{k, n-k} := \frac{n!}{k!(n-k)!} = \frac{n(n-1)\ldots(n-k+1)}{k(k-1)\ldots 1}. \qquad (1.3)$$

In older books, the binomial coefficient if often denoted C_k^n and is called the *number of combinations* of k from n. So $C_k^n \equiv \binom{n}{k}$.

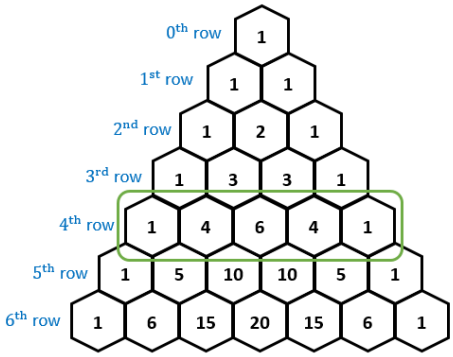
In calculations, a very useful property of binomial coefficients is that

$$\binom{n}{k} = \binom{n}{n-k},$$

which is obvious from definition (1.3). So, for example,

$$\binom{100}{98} = \binom{100}{2} = \frac{100 \times 99}{2 \times 1} = 4950.$$

Binomial coefficients also form the Pascal triangle. The numbers $\binom{n}{k}$ form the n-th row of this triangle, and each number in the row equals the sum of two numbers directly above.



This is also easy to prove by using definition (1.3).

A couple words on calculation. Some numerical software packages contain specialized functions for multinomial coefficients. For example in Mathematica, the coefficient

$$\binom{n}{n_1, \ldots, n_k}$$

can be computed as

$$\text{Multinomial}[n_1, \ldots, n_k].$$

For instance,

```
In= Multinomial[2, 4, 1]
Out= 105
```

Other packages do not have a built-in function for the multinomial coefficient but have a function for the factorial or for the binomial coefficients.

For example, in Matlab one can calculate the multinomial coefficient int the above example as

```
factorial(7)/(factorial(2)*factorial(4)*factorial(1)).
```

And for the binomial coefficients Matlab has function $nchoosek(n, k)$.

In Python 3.8 the binomial coefficient can be computed using its standard library math, by using function math.comb(n,k). For earlier versions of Python, it is not difficult to calculate multinomial coefficients using function math.factorial():

```
from math import factorial
factorial(7)/(factorial(2)*factorial(4)*factorial(1))
```

which results in 105.0.

4. Partitioning and combinations So far we used multinomial coefficients to count the number of permutations when not all objects are distinct.

Multinomial coefficients also appear in a closely related problem when we want to count in how many ways wee can partition objects into distinct subgroups but the order within the subgroups does not matter.

Example 1.3.10. Suppose 10 children are to be divided into 3 groups: 3 children will be in classes A and B, and 4 children in class C. How many ways to do it?

Here is a specific procedure for the division: line up the children and take the first 3 children in the line to class 3, the next 3 children to class B, and the remaining 4 children to class C. The total number of permutations in the line is 10!. However, the order of the first 3 children does not matter for the purposes of partitioning. The order of the next 3 children also does not matter, and the order of the remaining 4 does not matter. If we apply any permutation within these subgroups, they will result in the same division of children into the required 3 groups. Hence, the total number of divisions is

$$\frac{10!}{3!3!4!} = \binom{10}{3,3,4}$$

Note that this problem is different from the problem of permutations when some objects are identical. However, it also results in an answer that involves multinomial coefficients.

If a similar argument is applied in the general situation, then we get the following result:

Theorem 1.3.11. *The number of ways to partition n objects in k distinct groups, which have to contain n_1, n_2, …, n_k objects, respectively (with $n_1 + \ldots + n_k = n$), is*

$$\binom{n}{n_1, n_2, \ldots, n_k} = \frac{n!}{n_1! n_2! \ldots n_k!}.$$

Proof. Arrange objects in a specific order and let the first n_1 of them will be in the first group, next n_2 will be in the second group and so on. There are $n!$ different arrangements of objects in total. However, some of them results in the same partition.

Namely, if an arrangement can be obtained from another one by permuting the elements in the first group, then the elements in the second group, and so on, then these two arrangements both correspond to the same partition of objects.

Hence, the number of arrangements that result in a particular partition equals $n_1!n_2!\dots n_k!$. and therefore the number of possible partitions is $\frac{n!}{n_1!n_2!\dots n_k!}$. □

Again, it is important particular case, when there are only two groups.

Example 1.3.12. How many ways to select a 3-person commission from 13 faculty members?

Note that here we assume that the people within the commission are indistinguishable, for example, we do not care about whether one of the committee members is a chair. So it is easy to see that this is simply a particular case of the partitioning. We need to divide faculty into two groups: a group of 3 people in the committee and a group of other 10 people not in the committee. So, the answer is

$$\binom{13}{3,10} = \binom{13}{3}.$$

It is a frequent problem when we need to select r objects from the total group of n objects and we do not care about the order in the group of r objects. For example if we have objects $\{1,2,3,4,5,6\}$ and want to select 4 of them, then (1346) is the same selection as (3146). This is different from counting the number of permutations and often called the number of "combinations". The general argument is the same as in the example above and we get the following result.

Theorem 1.3.13. *The number of subsets of size r chosen from n objects is*

$$C_r^n \equiv \binom{n}{r} \equiv \binom{n}{r, n-r} \equiv \frac{n!}{r!(n-r)!}.$$

This is the reason why the binomial coefficient $\binom{n}{r}$ is also called the *number of combinations* of r elements out of n.

This coefficient arises in many combinatorial counting problems. For example, suppose we want to count the number of lattice paths from $(0,0)$ to (n,k) where the path can go only to the right (in the East

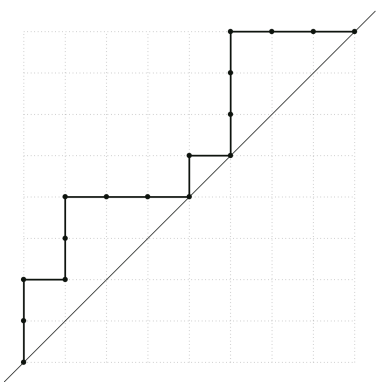


Figure 1.6: A lattice path from $(0,0)$ to $(8,8)$ which is above the

direction) or up (in the North direction). Then it is clear that we need to make $n+k$ steps to reach the point (n,k). Also, it is clear that n of them should be to the East and k, - to the North, and so to specify the path we only need to choose which n of these $n+k$ steps are to the east. So the total number of these paths is $\binom{n+k}{n}$.

Example 1.3.14. What is the probability that a random lattice path from $(0,0)$ to (n,n) will be always above the main diagonal.

We just calculated that the total number of possible paths from from $(0,0)$ to (n,n) is $\binom{2n}{n}$. How many of these paths stay above the diagonal?

To answer this question we can use a trick. Let us look at the first point that the path went below the diagonal and reflect the remaining part of the path relative to the line $y = x - 1$. It turns out that this gives a bijections of all the paths which go below the main diagonal with the paths that go from $(0,0)$ to $(n+1, n-1)$, and the number of such paths is $\binom{2n}{n+1}$. So, we find that the number of paths that always stay above the diagonal $y = x$ is

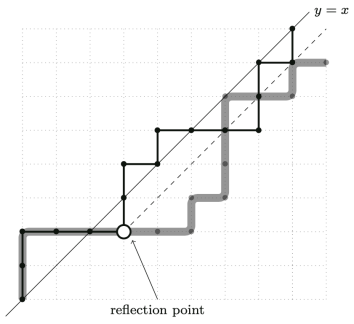


Figure 1.7: A reflection of a path that goes below the main diagonal. (Picture from Petersen "Eulerian numbers")

$$\binom{2n}{n} - \binom{2n}{n+1},$$

and this can be simplified as

$$\frac{1}{n+1}\binom{2n}{n},$$

which is called the Catalan number C_n.

So, by the Laplace definition, the probability that a random lattice path

from $(0,0)$ to (n,n) does not go below the diagonal $y=x$ is

$$\frac{\frac{1}{n+1}\binom{2n}{n}}{\binom{2n}{n}} = \frac{1}{n+1}.$$

Figure 1.8: Poker hands

In examples of elementary combinatorial probability, a deck of cards consists of 52 cards, which have 4 suits: diamonds (♢), clubs (♣)[1], hearts (♡) and spades (♠).

Every suit consist of 13 cards, which are ranked ($2 < 3 < \ldots < 10 < J < D < K < A$), where J is a Jack, D is a Dame, K is a King, and A is an Ace.

A hand consists of 5 randomly chosen cards.

Example 1.3.15. In poker, a full house is a hand that contains 3 cards of one rank and 2 cards of another rank. (For example 10♠10♣10♡2♡2♢) What is the probability of being dealt a full house?

The total number of hands is

$$\binom{52}{5} = \frac{52\times 51\times 50\times 49\times 48}{5\times 4\times 3\times 2\times 1} = 2598960 \approx 2.6\times 10^6.$$

(Note that we do not care about the order of cares in a hand.)

All hands have equal probability. We calculate the number of different full houses as follows. Select a rank and then select 3 cards from this rank. Then select a different rank and select 2 cards from that rank. This forms a hand and all these hands are different from each other. So the number of

[1]Note that clubs have somewhat unsuitable symbol. In French, this suite is named "Tréfle" which means "clover" and the card symbol depicts a three-leafed clover leaf. The Italian name is Fiori ("flower").

full houses is

$$13 \times \binom{4}{3} \times 12 \times \binom{4}{2} = 13 \times 4 \times 12 \times \frac{4 \times 3}{2 \times 1} = 3744.$$

So the probability is

$$\frac{3744}{2598960} \approx 0.0014 = 0.14\%.$$

Example 1.3.16. A flush is a hand that contains 5 cards of the same suit. For example $A\clubsuit K\clubsuit 10\clubsuit 3\clubsuit 2\clubsuit$. What is the probability of being dealt a flush?

We choose a suit and then 5 cards out of this suit. So the total number of flushes is

$$4 \times \binom{13}{5} = 4 \times \frac{13 \times 12 \times 11 \times 10 \times 9}{5 \times 4 \times 3 \times 2 \times 1} = 5148$$

So, the probability is

$$\frac{5148}{2598960} \approx 0.2\%.$$

This probability is larger than the probability of a full house, which is consistent with the convention that a full house is a stronger hand than a simple flush.

Example 1.3.17. A fair coin is tossed 4 times. What is the probability that we get exactly two heads?

The total number of outcomes is $2 \times 2 \times 2 \times 2 = 2^4 = 16$.

Now we need to count all favorable sequences that have exactly two heads. This is a partitioning problem. We partition 4 tosses into two groups: those which resulted in heads and those that resulted in tails. In this problem we are interested in sequences that have exactly 2 tosses which resulted in a head. There are $\binom{4}{2} = 6$ ways to select these two tosses among the four possible. So the probability is $\frac{6}{16} = 37.5\%$.

Problems of this kind will be important in later parts of the course. If we look at more general situation with n tosses, the solution easily generalizes and we see that the probability to get k heads in a sequence of n tosses is

$$p_{k,n} = \binom{n}{k} 2^{-n}.$$

Example 1.3.18. A group of people consists of 3 girls and 4 boys. In how many ways, they can select a committee that consists of 2 girls and 2 boys?

The total number to select a committee of 4 persons out of 7 people is $\binom{7}{4}$, the number of ways to select 2 committee girls out of 3 girls is $\binom{3}{2}$ and the total number to select 2 committee boys out of 4 boys is $\binom{4}{2}$. So the probability is

$$\frac{\binom{3}{2} \times \binom{4}{2}}{\binom{7}{4}} = \frac{3 \times 6}{35} \approx 0.5143 = 51.43\%$$

1.4 Conditional Probability and Independence

In many cases, the sample point method does not work. It is either difficult to find equally likely cases or difficult to count favorable cases.

The remedy is that we can postulate the probabilities of some events and require that the system of probabilities have some additional properties, which will allow us to calculate all other interesting probabilities. The most important of these properties are independence, the Markov property, and the martingale property. In this course we will study only independence. However, both of these properties are based on the concept of conditional probability.

1.4.1 Conditional probability

Figure 1.9: Conditional Probability

Suppose that we know that event B realized in an experiment. What is the probability of the event A given this information? For example, our calculations for the probability of a "full house" hand in a game of cards are valid only before we look at our own cards. Once we know our cards the probability that one of the other players has a "full house" changes

since the set of possible card distributions among players is now different from the situation when we have not known our cards. If we have $K\spadesuit$, then it cannot possibly be among cards of our opponents.

We can calculate the probabilities in this new reality by defining a new probability space, in which it is not possible that other players have the cards in your hand. The conditional probability is a technique to do it systematically.

We will simplify the situation a bit by making an additional assumption. To motivated it, suppose originally there were 3 outcomes ω_1, ω_2, ω_3 with equal probabilities 1/3. Suppose that we learned the information that w_3 is not a possible outcome. (Technically, the event $\{\omega_3\}^c$ has occurred.) How we should update the probabilities of ω_1 and ω_2? In principle, it depends on the details of the information obtained. It is quite possible that the information was of such nature that after we update our model we have probabilities $p(\omega_1) = 1/3$ and $p(\omega_2) = 2/3$. While the probability theory can handle this situation, however, we will adopt a simplifying assumption that the probability of each individual outcome is updated proportionally, by multiplying it by a constant which is the same for each outcome. So in this example the updating results in probabilities $p(\omega_1) = 1/2$ and $p(\omega_2) = 1/2$.

More generally, consider the simplest case when all outcomes in Ω have the same probability $p(\omega) = 1/N$ where N is the number of all possible outcomes. And assume that after we learned that event B realized, all outcomes in B are equally probable. Let the total number of outcomes be N, and let number of outcomes in A, B, and $A \cap B$ be denoted $|A|$, $|B|$, and $|A \cap B|$.

After it became known that event B realized the total number of possibilities became smaller. It is now not N but $|B|$.

What is the new number of favorable outcomes, that is, the outcomes in A? It is $|A \cap B|$ since we now know (with the new information) that the outcomes which are in A but not in B are no longer possible.

So the probability in question is

$$\mathbb{P}(A|B) = \frac{|A \cap B|}{|B|}. \tag{1.4}$$

We can write this expression differently and relate it to the probabilities of the events A, B, and $A \cap B$ before the information about B became known. Namely,

$$\mathbb{P}(A|B) = \frac{|A \cap B|/N}{|B|/N} = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

This formula is derived under the assumptions about the equal probability of outcomes both before and after observing event B. In general, we will assume that even if the assumption about equal probabilities not true, the information from observing B is absorbed in such a way that this formula is valid.

In other words, we define the conditional probability by this formula.

Definition 1.4.1. Assume $\mathbb{P}(B) > 0$. The conditional probability of event A given event B is defined as

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \tag{1.5}$$

Example 1.4.2. We roll a fair die. Define A="roll a 6" and B="roll an even number". What is $\mathbb{P}(A|B)$? What is $P(B|A)$?

We have $P(A \cap B) = P(A) = 1/6$ and $P(B) = 3/6 = 1/2$. Hence

$$P(A|B) = \frac{1/6}{1/2} = \frac{1}{3}.$$

For the second question $P(A \cap B)$ is still $\frac{1}{6}$ but now we have to divide by $P(A) = \frac{1}{6}$ and we find

$$P(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} = \frac{1/6}{1/6} = 1.$$

Conditional on rolling 6 we are certain that we rolled an even number.

Example 1.4.3. Suppose you roll two fair dice. What is the probability that at least one die is a 1, given that the sum of the two dice is 7?

Here all outcomes are equally probable and it is easier to use formula (1.4). The number of all combinations of dices that add up to 7 is 6: $(1,6),(2,5),\ldots,(6,1)$. Among them the number of outcomes at which at least one die is 1 is 2. These are $(1,6)$ and $(6,1)$. So the answer is $2/6 = 1/3$.

Properties of conditional probability

The conditional probability satisfy the three Kolmogorov's axioms with the sample space $\Omega = B$. Hence it satisfies all probability laws previously derived.

For example, the complement law and the union law now become:

$$\mathbb{P}(A^c|B) = 1 - \mathbb{P}(A|B),$$
$$\mathbb{P}(A \cup B|C) = \mathbb{P}(A|C) + \mathbb{P}(B|C) - \mathbb{P}(A \cap B|C).$$

1.4.2 Independence

Definition 1.4.4. Two events A and B are *independent* if (and only if) $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

If this equality does not hold, then A and B are called *dependent*.

Note that if $\mathbb{P}(B) \neq 0$, then independence of A and B implies that

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

Similarly, if $\mathbb{P}(A) \neq 0$ then independence imply that $\mathbb{P}(B|A) = \mathbb{P}(A)$. In other words, if two events are independent and we observed one of them, then it does not affect the probability of the other. Its conditional probability is the same as its unconditional probability, that is, as it was before the other event was observed.

Example 1.4.5. We roll a fair die. Define A="roll a 6", B="roll an even number", and C="roll a number greater than 4". Are A and C independent? Are B and C independent?

It is easy to see that $\mathbb{P}(A) = 1/6$, $\mathbb{P}(B) = 1/2$, and $\mathbb{P}(C) = 1/3$. Then event $A \cap C = \{6\} = A$ so $\mathbb{P}(A \cap C) = 1/6 \neq \mathbb{P}(A)\mathbb{P}(C)$ so these two events are not independent.

Event $B \cap C = \{6\}$ so $\mathbb{P}(B \cap C) = 1/6 = \mathbb{P}(B)\mathbb{P}(C)$ and so the events B and C are independent.

Example 1.4.6. Suppose you roll two fair dice. Consider the events: A = "the first die is a 1" and B = "the sum of the two dice is 7". Are they independent?

The total number of outcomes is 36. Event A contains 6 outcomes, $(1,1), (1,2), \ldots, (1,6)$ so its probability is $1/6$. Similarly, the probability of B is $1/6$. Event $A \cap B$ consist of only one outcome $(1,6)$ so its probability is $1/36$. Since $\frac{1}{36} = \frac{1}{6} \times \frac{1}{6}$, these events are independent.

Example 1.4.7. Suppose we toss an *unfair* coin 5 times. The probability that the coin lands heads up is $p \neq 1/2$. What is the probability that we observe exactly two heads in this sequence of tosses?

This example should be compared with the Example 1.3.17 for the *fair* coin where we used the equiprobability of all outcomes to find that the probability to observe k heads in n tosses is $\binom{n}{k}2^{-n}$.

Here the argument of Example 1.3.17 does not work, and we need an additional assumption to calculate the probability. The assumption that we use is that the result of the first toss is independent from the result of the second toss, from the result of the third toss and so on.

Let H_k and T_k denote the events that the coin lands heads or tails, respectively, in the k-th toss. For example, event H_1 consists of $2^4 = 16$ outcomes $HHHHH$, $HTHHH$, $HHTHH$, $HTTHH$, ..., $HTTTT$ that had H as the result of the first coin toss.

The assumption of independence allows us to calculate the probabilities of various intersections of these events. For example, the outcome $\{HHTTH\}$ corresponds to the intersection of events H_1, H_2, T_3, T_4, H_5,

and we can calculate its probability as

$$\begin{aligned}\mathbb{P}(HHTTH) &= \mathbb{P}(H_1 \cap H_2 \cap T_3 \cap T_4 \cap H_5) \\ &= \mathbb{P}(H_1)\mathbb{P}(H_2)\mathbb{P}(T_3)\mathbb{P}(T_4)\mathbb{P}(H_5) \\ &= p \times p \times (1-p) \times (1-p) \times p \\ &= p^3(1-p)^2.\end{aligned}$$

It is easy to note that while the probabilities of all sequences are in general different, these probabilities depend only on the number of heads in the sequence.

We are interested in sequences of coin tosses that resulted in 2 heads and by the previous argument the probability of each of these sequences is $p^2(1-p)^3$. So, to calculate the total probability of the event "two heads in a sequence of five tosses", we only need to calculate the number of all sequences with two heads. Fortunately, we already know how to do this due to Example 1.3.17, and in this case we get $\binom{5}{2}$ sequences. So the desired probability is

$$\binom{5}{2}p^2(1-p)^3.$$

In general, the probability to observe k heads in a sequence of n independent coin tosses is

$$\binom{n}{k}p^k(1-p)^{n-k},$$

where p is the probability to get heads in a single toss. Note that the result for the fair coin is recovered if we set $p = 1/2$. The probability becomes $\binom{n}{k}2^{-n}$.

1.4.3 Law of total probability

From the definition of the conditional probability, the probability of intersection of two events, A and B, is

$$\begin{aligned}\mathbb{P}(A \cap B) &= \mathbb{P}(A|B)\mathbb{P}(B), \text{ and} \\ &= \mathbb{P}(B|A)\mathbb{P}(A).\end{aligned} \tag{1.6}$$

This formula is sometimes called the *multiplicative law of probability*, however it is simply a restatement of the definition (1.5) for the conditional probability, and it holds for all events, whether they are independent or not. (For independent events A and B, this formula implies that $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ because $\mathbb{P}(A|B) = P(B)$ for independent events.)

Since this formula is a consequence of the definition of the conditional probability, it is only useful if we can find the conditional probability from some other sources, not from its definition. For example, this happens if $\mathbb{P}(A|B)$ can be estimated experimentally by observing how many times event A occurred in those situations when B happened. Now when we perform a different experiment we can still assume that $P(A|B)$ is the same although other probabilities (like $\mathbb{P}(B)$ or $\mathbb{P}(A \cap B)$) might have changed.

In some other models such as Markov chains, conditional probabilities are simply postulated from the beginning and then the results of the model can be compared to data.

Example 1.4.8. Suppose a basketball player makes 70% of his initial free throws. On his second attempt, he has an 80% success rate if he made the first and a 60% success rate if he missed the first.

What is the probability he makes both of a pair of free throws?

What is the probability he misses both free throws?

Let S_1 be the event that the player made his first free throw, and S_2 be the event that he made the second. Then, we know that $P(S_1) = 0.7$, $\mathbb{P}(S_2|S_1) = 0.8$, and $\mathbb{P}(S_2|S_1^c) = 0.6$. So, the probability that he makes both free throws is $\mathbb{P}(S_2 \cap S_1) = \mathbb{P}(S_2|S_1)\mathbb{P}(S_1) = 0.8 \times 0.7 = 0.56$ and the probability that he misses both is

$$\begin{aligned}\mathbb{P}(S_2^c \cap S_1^c) &= \mathbb{P}(S_2^c|S_1^c)\mathbb{P}(S_1^c) = (1 - \mathbb{P}(S_2|S_1^c))\mathbb{P}(S_1^c) \\ &= (1 - 0.6) \times (1 - 0.7) = 0.4 \times 0.3 = 0.12.\end{aligned}$$

Note that we used the complement law twice, in order to compute $\mathbb{P}(S_1^c)$ as $1 - \mathbb{P}(S_1)$ and $\mathbb{P}(S_2^c|S_1^c)$ as $(1 - \mathbb{P}(S_2|S_1^c))$.

Example 1.4.9 (Birthday Problem). What is the probability that in a set of n randomly chosen people some pair of them will have the same birthday?

What is the number of people that is needed to make the probability greater than $p = 90\%$.

Let us assume that the probability to be born on every day is the same and equals $1/365$. We start asking people one-by-one and we want to find the probability of the *complementary* event A_n that the first n people have been all born on different days. We can do by using induction and multiplicative formula for probability. Clearly $A_1 = 1$. Then we can simply write:

$$\mathbb{P}(A_n) = \mathbb{P}(A_n|A_{n-1})\mathbb{P}(A_{n-1}),$$

and for the conditional probability, if we know that the first $n-1$ people were born on different days then for person n there remains $365 - n + 1$ possibilities. Since all days are equally probable to be his or her birthday, so $P(A_n|A_{n-1}) = (365 - n + 1)/365$. So, calculating recursively we get

$$\mathbb{P}(A_n) = 1 \times \frac{364}{365} \times \ldots \times \frac{365 - n + 1}{365}$$

With a computer it is easy to evaluate this product. Here is how this can be done in Python.

```
import numpy as np
n = 41
prob = np.prod([1 - i/365 for i in range(n)])
print(prob)
Out: 0.09684838851826459
```

For $n = 40$ this code gives slightly more than 10%.

Alternatively, one can use a bit of calculus trickery and write:

$$\begin{aligned}
\log \mathbb{P}(A_n) &= \sum_{x=0}^{n-1} \log\left(1 - \frac{x}{365}\right) \\
&\approx -\sum_{x=0}^{n-1} \frac{x}{365} \approx -\frac{1}{365}\int_0^{n-1} x dx \\
&= -\frac{(n-1)^2}{2 \times 365},
\end{aligned}$$

where the approximation in the second line is by the Taylor expansion of $\log(1-x)$. It is good only if x is small relative to 365.

So if we want this probability $\approx 10\%$, then we should set

$$n = 1 + \sqrt{-2 \times 365 \log(0.1)} \approx 42.$$

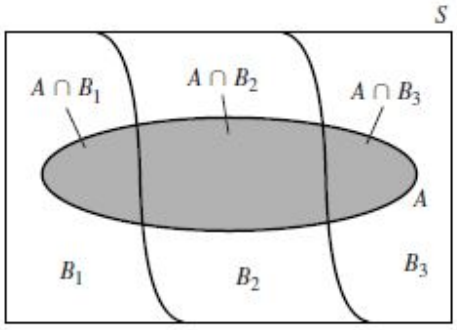
Now let us turn to the law of total probability. Intuitively, it is simply a method to calculate the probability of an event by splitting it into particular cases and then applying formula (1.6) to each case.

For some positive integer k, a collection of sets $\{B_1, B_2, \ldots, B_k\}$ forms a *partition* of S if:

1. $B_i \cap B_j = \emptyset$ if $i \neq j$.
2. $S = B_1 \cup B_2 \cup \ldots \cup B_k$.

Theorem 1.4.10 (Law of Total Probability)**.** *If $\{B_1, B_2, \ldots, B_k\}$ is a partition of S such that $\mathbb{P}(B_i) > 0$, $i = 1, \ldots, k$, then for any event A,*

$$\mathbb{P}(A) = \sum_{i=1}^{k} \mathbb{P}(A|B_i)\mathbb{P}(B_i).$$



Proof. $A = (A \cap B_1) \cup (A \cap B_2) \cup \ldots (A \cap B_k)$ and the terms in this union are disjoint.

By additive law, $\mathbb{P}(A) = \mathbb{P}(A \cap B_1) + \mathbb{P}(A \cap B_2) + \ldots + \mathbb{P}(A \cap B_k)$.

By applying the multiplicative law to each term in this sum, we get: $\mathbb{P}(A) = \mathbb{P}(A|B_1)\mathbb{P}(B_1) + \mathbb{P}(A|B_2)\mathbb{P}(B_2) + \ldots + \mathbb{P}(A|B_k)\mathbb{P}(B_k)$. □

Example 1.4.11. A box contains 1 green and 4 white balls. A ball is taken out of the box and its color is noted. Then it is returned to the box together with a new ball of the same color. Then the procedure is repeated.

Let G_k denotes the event that the ball drawn in the k-th round of this game is green. What is the probability of G_2? What is $\mathbb{P}(G_3)$?

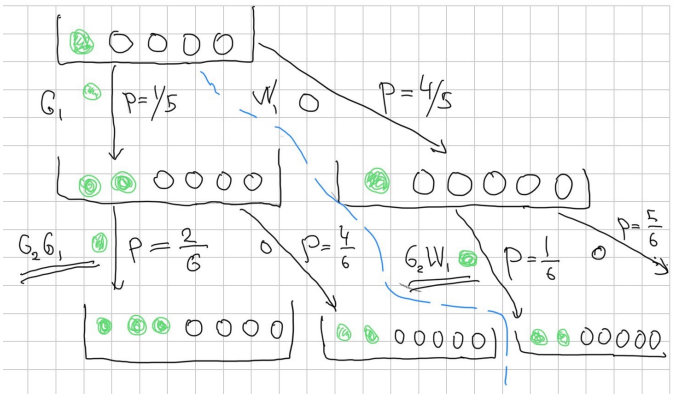


Figure 1.10: Possible histories of the box with balls

In order to know what happens in the 2nd round, we need to know what happened in the first. We have already denoted the event "a green ball is drawn in the first round" as G_1. Let the complementary event "a white ball is drawn in the first round" be denoted W_1. (So $W_1 = (G_1)^c$.)

The probability of G_1 is $1/5$, the number of green balls divided by the total number of balls initially in the box. Similarly, $\mathbb{P}(W_1) = 4/5$.

Then, if G_1 happened, then at the beginning of the second round we have 2 green balls and 4 white balls in be box, so the probability to get a green ball again is $\mathbb{P}(G_2|G_1) = 2/6$.

On the other hand, if W_1 happened in the first round then at the beginning of the second round we have 1 green and 5 white balls. Hence, the probability to get a green ball in the second round in this case is $\mathbb{P}(G_2|W_1) = 1/6$.

Now, at this moment we are ready to apply the total probability law:

$$\begin{aligned}\mathbb{P}(G_2) &= \mathbb{P}(G_2|G_1)\mathbb{P}(G_1) + \mathbb{P}(G_2|W_1)\mathbb{P}(W_1) = \\ &= \frac{2}{6}\times\frac{1}{5} + \frac{1}{6}\times\frac{4}{5} = \frac{2\times 1 + 1\times 4}{6\times 5} = \frac{1}{5}.\end{aligned}$$

So the probability that we get a green ball in the second round is the same as the probability that we get it in the first round.

It is interesting because we could expect that will become smaller because the white balls proliferate with larger probability and so it appears likely that at the second round the green balls will be even more rare than initially.

Now, what about $\mathbb{P}(G_3)$? This can be obtained by a similar calculation. Except that the word explanations would be very lengthy. So, let us just say that G_2G_1 is a shortcut notation for the event $G_2 \cap G_1$, that is, for the event that we got a green ball both in the first and in the second round. Other events like W_2G_1 are defined similarly. Then, in order to calculate $P(G_3)$ we need to consider all possible histories of how we draw the balls from the box. Hence, by the total probability law we have:

$$\begin{aligned}\mathbb{P}(G_3) = \mathbb{P}(G_3|G_2G_1)\mathbb{P}(G_2G_1) + \mathbb{P}(G_3|W_2G_1)\mathbb{P}(W_2G_1) \\ + \mathbb{P}(G_3|G_2W_1)\mathbb{P}(G_2W_1) + \mathbb{P}(G_3|W_2W_1).\end{aligned}$$

Consider, for example the first term in this sum. After we get two green balls in the first and the second round the ball has 3 green and 4 white balls. Hence $\mathbb{P}(G_3|G_2G_1) = 3/7$. In addition we can calculate $\mathbb{P}(G_2G_1)$ by using the multiplicative law:

$$\mathbb{P}(G_2G_1) = \mathbb{P}(G_2|G_1)\mathbb{P}(G_1) = \frac{2}{6} \times \frac{1}{5}.$$

Altogether, we get

$$\mathbb{P}(G_3|G_2G_1)\mathbb{P}(G_2G_1) = \frac{3}{7} \times \frac{2}{6} \times \frac{1}{5}.$$

All other terms are computed similarly and we get that

$$\begin{aligned}\mathbb{P}(G_3) &= \mathbb{P}(G_3|G_2G_1)\mathbb{P}(G_2|G_1)\mathbb{P}(G_1) + \mathbb{P}(G_3|W_2G_1)\mathbb{P}(W_2|G_1)\mathbb{P}(G_1) \\ &\quad + \mathbb{P}(G_3|G_2W_1)\mathbb{P}(G_2|W_1)\mathbb{P}(W_1) + \mathbb{P}(G_3|W_2W_1)\mathbb{P}(W_2|W_1)\mathbb{P}(W_1) \\ &= \frac{3 \times 2 \times 1 + 2 \times 4 \times 1 + 2 \times 1 \times 4 + 1 \times 5 \times 4}{7 \times 6 \times 5} \\ &= \frac{42}{210} = \frac{1}{5}.\end{aligned}$$

So $\mathbb{P}(G_3)$ is again the same. It looks like a pattern, so you can try to prove directly that the probability is always the same. However, the best way would be probably by using the theory of Markov chains.

1.5 Bayes' formula

One important thing to keep in mind is that $\mathbb{P}(A|B)$ and $\mathbb{P}(B|A)$ are not equal to each other.

Example 1.5.1. Toss a coin 5 times. Let A = "first toss is heads" and let B = "all 5 tosses are heads". Then $\mathbb{P}(A|B) = 1$ but $\mathbb{P}(B|A) = (\frac{1}{2})^4$. (The latter equality holds because if we know that the first toss is heads then B occurs only if the four other tosses are also heads and this happens with probability $1/2^4$.)

The simplest way to get from $\mathbb{P}(A|B)$ to $\mathbb{P}(B|A)$ is by using the Bayes formula.

Theorem 1.5.2 (Bayes' formula I)**.** *Suppose* $\mathbb{P}(A) \neq 0$ *and* $\mathbb{P}(B) \neq 0$. *Then,*

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)}$$

In the example above, we have $\mathbb{P}(A) = 1/2$ and $\mathbb{P}(B) = 1/2^5$, and so

$$\mathbb{P}(B|A) = \frac{1 \times \frac{1}{2^5}}{\frac{1}{2}} = \frac{1}{2^4}.$$

Proof. The claim simply follows from the definition of the conditioning probability and the multiplicative law.

$$\mathbb{P}(B|A) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)}$$

□

Very often the Bayes' formula is used in a specific situation, when we know that there are k mutually exclusive cases $B_1, \ldots B_n$, know the *prior* probability $P(B_i)$ for each of them, and know the conditional probabilities

of some other event A conditional on the realization of B_i. That is, we know the probabilities of $\mathbb{P}(A|B_i)$. These are the "direct inference" probabilities: if case B_i holds, then the probability of A is $\mathbb{P}(A|B_i)$.

Now, suppose that we performed the random experiment and learned that A has indeed occurred. We want to make an "inverse inference": how does the fact that A has occurred changes the probabilities of B_i. That is, what are *posterior* probabilities of B_i, $\mathbb{P}(B_i|A)$?

Theorem 1.5.3 (Bayes' formula II)**.** *If $\{B_1, B_2, \ldots, B_k\}$ is a partition of S such that $\mathbb{P}(B_i) > 0$, $i = 1, \ldots, k$, then for any $j = 1, \ldots, k$:*

$$\mathbb{P}(B_j|A) = \frac{\mathbb{P}(A|B_j)\mathbb{P}(B_j)}{\sum_{i=1}^{k} \mathbb{P}(A|B_i)\mathbb{P}(B_i)}.$$

Proof:

$$\mathbb{P}(B_j|A) = \frac{\mathbb{P}(A \cap B_j)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A|B_j)\mathbb{P}(B_j)}{\sum_{i=1}^{k} \mathbb{P}(A|B_i)\mathbb{P}(B_i)},$$

where the first equality is by definition, and the second equality is by the multiplicative and total probability laws. □

These two theorems are variants of the same idea. An especially simple case occurs when the partition B_i consists of only two elements B and B^c:

Corollary 1.5.4. *If 0<P(B)<1, then*

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c)} = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)}.$$

Proof. Take partition $\{B, B^c\}$ in Theorem 1.5.3. □

Example 1.5.5. A firm rents cars for its employees. It rents from 3 agencies: 60% from agency AAA, 30% from agency BBB, and 10% from agency CCC.

Suppose, typically 9% of cars from agency AAA need a tuneup, 20% from agency BBB need a tuneup, and 6% from agency CCC need a tuneup.

If an employee discovered that a particular rental car needs a tuneup, what is the chance that it came from agency BBB?

Let A, B, C denote the events that the car came from the agency AAA, BBB, and CCC, respectively, and let T denote the event that it needs a

tuneup. Then, we know that $\mathbb{P}(A) = 0.6$, $\mathbb{P}(B) = 0.3$, $\mathbb{P}(C) = 0.1$. We also know that $\mathbb{P}(T|A) = 0.09$, $\mathbb{P}(T|B) = 0.2$, and $\mathbb{P}(T|C) = 0.06$.

What we want to find is $\mathbb{P}(B|T)$. By applying Bayes formula as in Theorem 1.5.3, we get

$$\begin{aligned}\mathbb{P}(B|T) &= \frac{\mathbb{P}(T|B)\mathbb{P}(B)}{\mathbb{P}(T|A)\mathbb{P}(A) + \mathbb{P}(T|B)\mathbb{P}(B) + \mathbb{P}(T|C)\mathbb{P}(C)} \\ &= \frac{0.2 \times 0.3}{0.09 \times 0.6 + 0.2 \times 0.3 + 0.06 \times 0.1} = 0.5\end{aligned}$$

So the prior probability that the car came from the agency BBB was $\mathbb{P}(B) = 0.3$. After an employee discovered that the car needs a tune-up, the probability increased. The posterior probability $P(B|T)$ is 0.5.

The confusion between the conditional probabilities $\mathbb{P}(A|B)$ and $\mathbb{P}(B|A)$ often leads to an erroneous argument which is known as "The Base Rate Fallacy".

Example 1.5.6. Consider a routine screening test for a disease. Suppose the frequency of the disease in the population (base rate) is 0.5%. The test is highly accurate with a 5% false positive rate and a 10% false negative rate. (False positive rate is a percentage of times when a person a healthy but the test indicates that he or she is sick, and false negative is the percentage when the person is sick but the test indicates that he is not sick (healthy).)

You take the test and it comes back positive. What is the probability that you have the disease?

A typical reaction is that test erroneously gives a positive result in only 5% cases, so it must be that the probability that your are sick is the remaining 95%.

However 5% is the probability $\mathbb{P}(+|H)$, where "+" is the event that test came positive and H is the event that you are healthy. And 95% is $1 - \mathbb{P}(+|H) = \mathbb{P}(-|H)$, so it is the probability that you get negative result if you in fact healthy. What you really want to know is a very different probability, the probability that you are sick given that the test is positive: $\mathbb{P}(S|+)$.

So we apply Bayes' formula:

$$\begin{aligned}\mathbb{P}(S|+) &= \frac{\mathbb{P}(+|S)\mathbb{P}(S)}{\mathbb{P}(+|H)\mathbb{P}(H)+\mathbb{P}(+|S)\mathbb{P}(S)} \\ &= \frac{(1-0.1)\times 0.005}{0.05\times 0.995+(1-0.1)\times 0.005} \approx 0.083\end{aligned}$$

(Here $1-0.1$ is from the false negative rate: $\mathbb{P}(+|S) = 1-\mathbb{P}(-|S)$.)

So, the probability that you have a decease is actually only 8.3%. The intuition behind this is that while the test is quite accurate and gives strong evidence in favor of the decease, the prior probability of the decease (the base rate) is so small that it still outweighs the evidence given by the test.

Here is another example that shows how the Bayesian formula works.

Example 1.5.7. You start with a box that has 4 red and 1 blue ball in it. Your assistant draws one ball and puts that ball back in the urn along with another ball of the same color. She does it without telling you what was the color of the ball.

Now she draws another ball from the box and shows it to you.

Suppose this second ball is red. What is the probability the first ball was blue?

Let us use notation similar to the notation in Example 1.4.11. For instance, R_2 denotes the event that the ball drawn in the second round is red. Then we need to compute $\mathbb{P}(B_1|R_2)$. By using Bayes' formula, we get:

$$\begin{aligned}\mathbb{P}(B_1|R_2) &= \frac{\mathbb{P}(R_2|B_1)\mathbb{P}(B_1)}{\mathbb{P}(R_2|B_1)\mathbb{P}(B_1)+\mathbb{P}(R_2|R_1)\mathbb{P}(R_1)} \\ &= \frac{\frac{4}{6}\times\frac{1}{5}}{\frac{4}{6}\times\frac{1}{5}+\frac{5}{6}\times\frac{4}{5}} = \frac{4}{4+20} = \frac{1}{6}.\end{aligned}$$

Note that the original probability of the blue ball in the first round is $\mathbb{P}(B_1) = \frac{1}{5}$. So the observation of the red ball in the second round made the blue ball in the first round less likely.

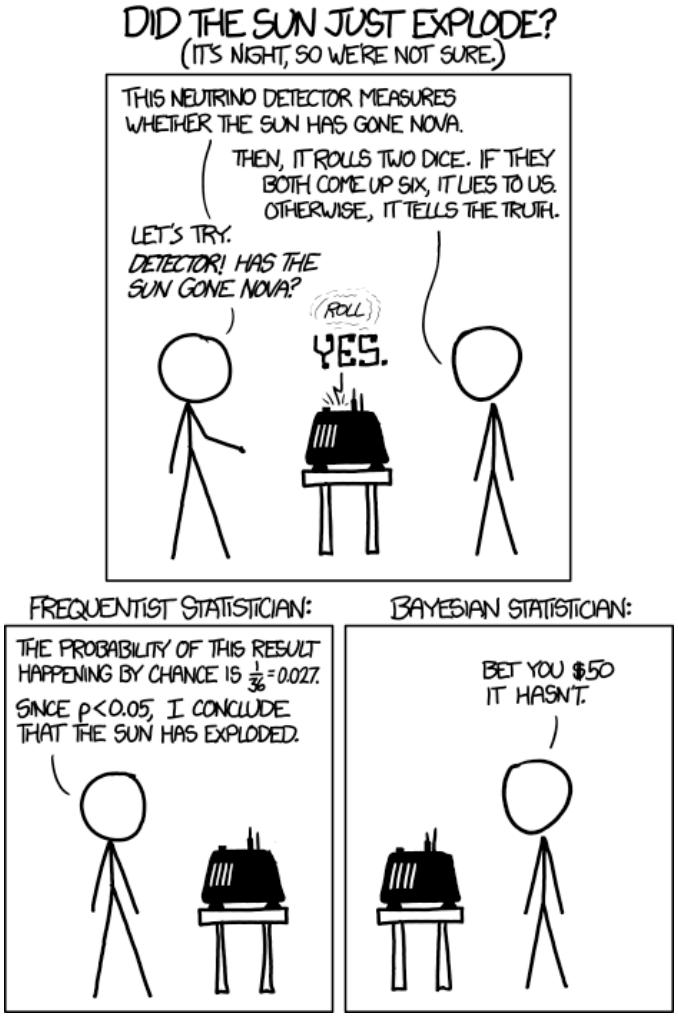


Figure 1.11: Another version of the base rate fallacy as an argument against frequentist statistics.

Chapter 2

Discrete random variables and probability distributions.

2.1 Pmf and cdf

Definition 2.1.1. A *random variable* (r.v.) X is a function on the probability space Ω.

So we perform an experiment, get an outcome ω, and then we calculate some quantity that depends on the outcome, $X(\omega)$. This is a random variable. We can think about one quantitative feature of a complicated random outcome.

Example 2.1.2. In an experiment, we roll two dice. A random variable $X =$ the sum of values on the two dice. For instance, if the experiment resulted in the outcome $(6, 5)$, then $X = 11$.

A random variable is a function of the outcome ω, so in principle, it should be written $X(\omega)$. But very often it is simply written X for shortness.

A *discrete random variable* is a r.v. that can take only a *finite* or *countably infinite* number of different values.

It is useful to think about values of a random variable X as outcomes of a related experiment. We perform the original experiment, get an outcome ω, calculate $X(\omega)$ and then forget about ω. So the value of X is the outcome

in this reduced experiment. In the previous example, it is as if your assistant rolled two dice and told you only the sum of the two dice.

The study of a single random variable is useful because it allows us to look at the simplest possible experiments, that is, on the experiments whose outcomes are integers or real numbers. The next step would be to look at the experiment where an outcome is a random vector. We will do it in later chapters.

For this reduced experiment, we can define the probability mass function of its outcomes. It is called the probability mass function of a random variable.

Definition 2.1.3. The probability mass function (pmf) of a discrete r.v X is a function that maps each possible value of X to its probability.

$$p_X(x) = \mathbb{P}(X = x),$$

where we use 'X=x' to denote the event $\{\omega : X(\omega) = x\}$.

Notation: We use capital letters (like X) to denote a random variable, and a lowercase letters (like x) to denote a particular value the r.v. may take.

Example 2.1.4. What is the pmf for the random variable X in Example 2.1.2?

$p_X(1) = 0$, since the sum can never be 1. Then, $X = 2$ can occur only if the outcome is $(1, 1)$. This has the probability $1/36$, so

$$p_X(2) = \frac{1}{36}.$$

$X = 3$ can happen for 2 outcomes: $(2, 1)$ and $(1, 2)$, so

$$p_X(3) = \frac{2}{36}.$$

Similarly, $p_X(4) = \frac{3}{36}$, $p_X(5) = \frac{4}{36}$, and so on, until $p_X(7) = \frac{6}{36}$

At $x = 8$ the pattern changes, because we do not have the outcomes $(1, 7)$ and $(7, 1)$ so there are only 5 outcomes that result in $X = 8$: $(2, 6)$, $(3, 5)$, $(4, 4)$, $(5, 3)$, $(6, 2)$. So $p_X(8) = \frac{5}{36}$.

Similarly, we can find that $p_X(9) = \frac{4}{36}$, $p_X(10) = \frac{3}{36}$, ..., $p_X(12) = \frac{1}{36}$.

For all other values of x, $p_X(x) = 0$, since these values are not possible. For example, $p_X(0) = 0$, $p_X(13) = 0$, $p_X(1.5) = 0$, and so on.

Example 2.1.5 (pmf of an indicator). Recall that we defined an indicator function for an event A as a function $I_A(\omega)$ that takes value 1 if $\omega \in A$, and value 0 if $\omega \notin A$. Since it is a function on Ω, it is a random variable and we can ask what is its pmf. Since I_A takes only two values: 0 and 1, its pmf is not zero only at 0 and 1 and

$$p_{I_A}(0) = \mathbb{P}(I_A = 0) = \mathbb{P}(\omega \notin A) = 1 - \mathbb{P}(A).$$

Similarly, $p_{I_A}(1) = \mathbb{P}(A)$. For all other x not equal to 0 or 1, $p_{I_A}(x) = 0$. In summary,

$$p_{I_A}(x) = \begin{cases} 1 - \mathbb{P}(A), & \text{if } x = 0, \\ \mathbb{P}(A), & \text{if } x = 1, \\ 0, & \text{otherwise.} \end{cases}$$

Example 2.1.6 (Number of heads in coin tossing experiment). Let a coin be tossed n times. The results of the tosses are independent and the probability to get heads is p. Let X be the number of heads we observed in the experiment. What is the pmf of X?

We recall from the calculations in the previous chapter, that the probability that we get exactly k heads in this experiment is $\binom{n}{k}p^k(1-p)^{n-k}$. By definition, the pmf of X is

$$p_X(k) = \binom{n}{k}p^k(1-p)^{n-k}, \text{ for } k \in \mathbb{Z}, \quad 0 \le k \le n,$$

and $p_X(k) = 0$, otherwise.

Theorem 2.1.7. *A probability mass function $p_X(x)$ for a discrete random variable X satisfies the following conditions.*

(i.) $0 \le p_X(x) \le 1$ *for all* $x \in \mathbb{R}$,

(ii.) $\sum_{x:p_X(x)>0} p_X(x) = 1$.

Proof. Condition (i.) follows from Kolmogorov's axioms because $p_X(x)$ is the probability of an event $\{\omega : X(\omega) = x\}$, so it is positive and does not exceed 1. Condition (ii.) holds by the third axiom because the events $\{\omega : X(\omega) = x\}$ form a countable disjoint family that covers all probability space Ω. □

Remark 1. In principle we could define pmf for arbitrary random variables, not only for discrete r.v. However, property (ii.) is often invalid for continuous r.v. (By continuous we mean a random variable which is not discrete.) Think about our example with shooting on a target, and let a random variable X be equal to the x coordinate of the outcome. The probability $p_X(x) = 0$ for every x, so the sum in (ii.) is empty and by convention empty sums are equal to 0 not 1. So pmf is not very useful for continuous random variables.

Remark 2. Every function that satisfies (i.) and (ii.) is a pmf of a discrete random variable. This is not difficult to prove by constructing an example of a required random variable.

Theorem 2.1.7 is a useful way to check if a given function is a valid pmf.

Another very important function is the cumulative distribution function. It will be more important for continuous random variables, but we define it right now so you could get used to it.

Definition 2.1.8. The cumulative distribution function (cdf) is

$$F_X(x) = \mathbb{P}(X \leq x).$$

If we know the pmf of a random variable then it is very easy to compute the cdf. It is simply the sum of $p_X(t)$ for those values of t which is smaller or equal to x:

$$F_X(x) = \sum_{t \leq x \text{ and } p_X(t) \neq 0} p_X(t). \tag{2.1}$$

Example 2.1.9. Consider the situation as in Example 2.1.2. What is $F_X(1)$? $F_X(2)$? $F_X(3)$? $F_X(3.5)$? $F_X(10.5)$? $F_X(11)$? $F_X(12)$? $F_X(13)$?

For $F_X(1)$ we note that the event $X \leq 1$ never happens, so its probability is zero. Alternatively, if we want to calculate $F_X(1)$ using formula (2.1) then

we note the smallest t when $p_X(t)$ is not zero is $t = 2$, so that the sum at the right hand side is empty.

For $F_X(2)$, formula (2.1) gives $F_X(2) = p_X(2) = 1/36$, since there is only one term in the sum.

Then,

$$F_X(3) = p_X(2) + p_X(3) = \frac{1}{36} + \frac{2}{36} = \frac{3}{36}.$$

For $F_X(3.5)$ it is actually the same sum as for $F_X(3)$:

$$F_X(3.5) = p_X(2) + p_X(3) = \frac{1}{36} + \frac{2}{36} = \frac{3}{36}.$$

Now we can calculate $F_X(10.5)$ by using formula (2.1):

$$F_X(10.5) = p_X(2) + p_X(3) + \ldots + p_X(10).$$

However, it is easier to note that the event $X \leq 10.5$ is complementary to the event $X > 10.5$. So by the formula for the complementary event:

$$\begin{aligned} F_X(10.5) &= \mathbb{P}(X \leq 10.5) = 1 - \mathbb{P}(X > 10.5) = 1 - (p_X(11) + p_X(12)) \\ &= 1 - \left(\frac{1}{36} + \frac{2}{36}\right) = \frac{33}{36}. \end{aligned}$$

For $F_X(11)$, we can write similarly,

$$\begin{aligned} F_X(11) &= \mathbb{P}(X \leq 11) = 1 - \mathbb{P}(X > 11) = 1 - p_X(12) \\ &= 1 - \frac{1}{36} = \frac{35}{36}. \end{aligned}$$

Note that the we calculate $\mathbb{P}(X > 11)$ we do not include $p_X(11)$ in the set of the terms that we subtract because the inequality in the description of the event is strict.

For $F_X(12)$, we have: $F_X(12) = 1 - \mathbb{P}(X > 12) = 1$, because $X > 12$ never happens, so $\mathbb{P}(X > 12) = 0$.

And for $F_X(13)$, we have the same calculation $F_X(13) = 1 - \mathbb{P}(X > 13) = 1$.

Now, we know how to calculate the cdf from the pmf of a random variable. Can we go in the other direction?

Yes. The recipe is that pmf $p_X(x)$ is not zero only at those points where the cdf $F_X(x)$ is discontinuous (that is, where it jumps) and the value of $p_X(x)$ equals to the size of this discontinuity (the size of the jump).

Example 2.1.10.

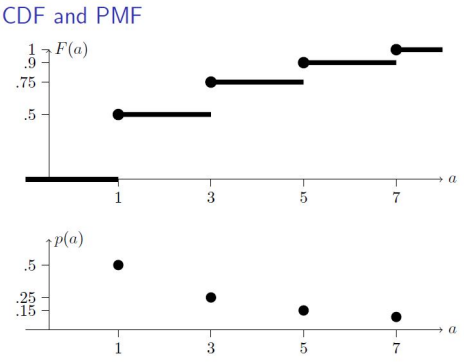


Figure 2.1: The cdf and pmf functions

Consider the cdf shown in the upper part of Figure 2.1 Recover the pmf (that is, recover the lower part of this picture).

Let us, for example compute $p_X(5)$. From the picture, we have $F_X(5) = 0.9$ and for x a little bit to the left of 5 we have $F_X(x) = 0.75$, so we have $p_X(5) = 0.9 - 0.75 = 0.15$. A little bit more rigorously,

$$\begin{aligned} p_X(5) &= F_X(5) - \lim_{x \uparrow 5} F_X(5) \\ &= 0.9 - 0.75 = 0.15. \end{aligned}$$

where $x \uparrow 5$ means that x tends to 5 from below.

You can check your understanding by playing more with this example.

Exercise 2.1.11. Let X a random variable with the cdf shown at Figure 2.1. So,

values of X:	1	3	5	7
cdf $F_X(x)$:	.5	.75	.9	1

In fact, this table is not quite correct description of the cdf, since it does not say what happens between the points in the table. We either have to say that the cdf does not change between these points, or simply give an explicit description like the following:

$$F_X(x) = \begin{cases} 0, & \text{if } x < 1, \\ 0.5, & \text{if } 1 \le x < 3, \\ 0.75, & \text{if } 3 \le x < 5, \\ 0.9, & \text{if } 5 \le x < 7, \\ 1, & \text{if } 7 \le x. \end{cases}$$

In the example above we calculated $p_X(5) = \mathbb{P}(X = 5)$. Do you know how to calculate $\mathbb{P}(X \le 3)$? $p_X(3) = \mathbb{P}(X = 3)$? $\mathbb{P}(X < 3)$?

2.2 Expected value

2.2.1 Definition

The expected value of a random variable is a weighted average of its values, where the weight of each value x is set to the probability of this value. Mathematically, we have the following definition.

Definition 2.2.1. Let X be a discrete r.v. with pmf $p_X(x)$. Then the *expected value* of X, denoted $\mathbb{E}X$, is:

$$\mathbb{E}X = \sum_{x:p_X(x)>0} x p_X(x).$$

The expected value can be defined also for continuous random variables but we will do it later.

If the random variable takes a countable number of values, the expected value does not always exist, since the sum $\sum_{x:p_X(x)>0} x p_X(x)$ may diverge.

The expected value $\mathbb{E}X$ is also often called the *expectation*, *mean*, or *average* of X, and it is often denoted μ in statistics.

Example 2.2.2. Consider the random variable X from Example 2.1.2. What is its expected value?

By using the pmf computed in 2.1.2, we can calculate:

$$\begin{aligned}\mathbb{E}X =& 2\times\frac{1}{36}+3\times\frac{2}{36}+4\times\frac{3}{36}+5\times\frac{4}{36}+6\times\frac{5}{36}+7\times\frac{6}{36}\\ &+8\times\frac{5}{36}+9\times\frac{4}{36}+10\times\frac{3}{36}+11\times\frac{2}{36}+12\times\frac{1}{36}\\ =&7.\end{aligned}$$

We will see soon that there is a simpler way to calculate this expected value.

Example 2.2.3. We roll two dice. You win \$1000 if the sum is 2 and lose \$100 otherwise. You play this game many times. How much do you expect to win on average per trial?

We will see later that the average win per trial for large number of trials approximately equal to the expected value of the random gain X. This gain has pmf $p_X(\$1000) = \frac{1}{36}$ and $p_X(-\$100) = \frac{35}{36}$. So,

$$\mathbb{E}X = \$1000\times\frac{1}{36}+(-\$100)\times\frac{35}{36}\approx -\$69.44$$

Example 2.2.4 (St. Petersburg Paradox). You are going to play a coin-tossing game with a fair coin in a casino. The casino will pay you \$1 if Head appears on the initial throw (and then the game stops). If on the first throw you get Tail, then the game continues and if Head appears on the second throw, then the casino will pay you \$2 (and stop the game). An so on with payouts doubling with each toss. In other words, the coin will be tossed until a Head appear and the casino will pay you $\$2^{n-1}$ dollars if the first Head appears on the n-th throw.

How much are you willing to pay the casino for a chance to participate in this game?

Professor of Economics: Any payment which is less than your expected earnings from the game gives you a positive expected net earning (at least if we ignore the time discount) and so an offer of the game with this up-front payment should be accepted.

So what are the expected earnings? Let us look at the various scenarios in this game. The probability that you win 1 is the probability of a very short scenario: H, so it is $\frac{1}{2}$. You win 2 if the development in the game is

TH, with probability $\frac{1}{4}$. The probability to win 4 is the probability of the sequence TTH, which is $\frac{1}{8}$, and so on. So, if X is your total payoff in the game, then

$$\begin{aligned}\mathbb{E}X &= \$1 \times \frac{1}{2} + \$2 \times \frac{1}{4} + \$4 \times \frac{1}{8} + \dots \\ &= \$\frac{1}{2} + \$\frac{1}{2} + \$\frac{1}{2} + \dots = \infty\end{aligned}$$

It appears that you should be willing to pay any amount for the chance to play this game. This does not seem to be reasonable.

A possible explanation of this paradox is that it is very risky to pay a million of dollars for this game. In most cases your gain will be small and you rely on a lucky sequence which will cover all your previous losses but unfortunately has a very small probability.

In principle, this risk can be removed if you can repeat the game many times, until you see this lucky sequence. However, you might have not enough time or money to reduce the risk by repetition of the game.

In Economics, this paradox is usually resolved by introducing "utility function" and time discounting. The criterion becomes that you should accept a payment for the game only if the expected time-discounted utility of the gain is larger than the reduction in utility due to payment. Then we need to know how to calculate the expectation of a function of a random variable, the utility function of the wealth after the game. We will discuss this in the next section after one additional example.

Example 2.2.5 (Expectation of an indicator). What is the expected value of the indicator of an event A?

We calculated the pmf of the indicator $\mathbb{1}_A$ in Example 2.1.5. Then,

$$\mathbb{E}\mathbb{1}_A = 1 \times \mathbb{P}(A) + 0 \times (1 - \mathbb{P}(A)) = \mathbb{P}(A).$$

So the expectation of the indicator of an event A equals the probability of this event.

2.2.2 Expected value for a function of a r.v.

Theorem 2.2.6. *If X is a discrete r.v. with pmf $p_X(x)$ and if $Y = g(X)$ is a real-valued function of X, then*

$$\mathbb{E}Y = \sum_{x \in \text{range}(X)} g(x)p_X(x),$$

where $x \in \text{range}(X)$ in the summation means that x runs over all possible values of r.v. X.

The key point of this theorem is that we sum not over all possible values of $Y = g(X)$ (as we would have to do if we used the definition) but over all possible values of X, and we use not the pmf of Y but the pmf of X.

Example 2.2.7. Let X has pmf $p_X(-2) = p_X(0) = p_X(2) = 1/3$. What is the expectation of $Y = X^2$.

We can do this problem in two different ways. First, if we want to calculate $\mathbb{E}Y$ by definition, we need the pmf of $Y = X^2$. This is easy to figure out in this particular example. The r.v. Y can take only values 0 and 4, and $p_Y(0) = \mathbb{P}(Y = 0) = \mathbb{P}(X = 0) = 1/3$ and $p_Y(4) = \mathbb{P}(Y = 4) = \mathbb{P}(X = -2) + \mathbb{P}(X = 2) = 2/3$.

Hence, by definition, $\mathbb{E}Y = 0 \times \frac{1}{3} + 4 \times \frac{2}{3} = 8/3$.

The second way is through our theorem: $\mathbb{E}Y = (-2)^2 \times \frac{1}{3} + 0^2 \times \frac{1}{3} + 2^2 \times \frac{1}{3} = 8/3$.

The point is that the calculation suggested by the theorem is typically easier than the calculation by the definition. (Although in this particular example they were of the same level of difficulty.)

Here is another example.

Example 2.2.8. Toss two fair coins. If X is the number of heads which we got in these two tosses, what is $\mathbb{E}(X^2)$?

Recall that the probability to get k heads in n tosses if the probability of a head is p is $\binom{n}{k}p^k(1-p)^{n-k}$ and for fair coin with $p = 1$ this simplifies to $\binom{n}{k}2^{-n}$. So, the pmf of X is $p_X(0) = 1/4$, $p_X(1) = 2/4$, $p_X(2) = 1/4$, and:

$$\mathbb{E}X^2 = 0^2 \times \frac{1}{4} + 1^2 \times \frac{2}{4} + 2^2 \times \frac{1}{4} = 1.5.$$

Proof of Theorem 2.2.6. Suppose $Y = g(X)$ take values $y_1, \dots, y_m$. Then by definition

$$\begin{aligned}
\mathbb{E}[g(X)] &= \sum_y y p_Y(y) \\
&= \sum_y y \mathbb{P}(g(X) = y) \\
&= \sum_y y \sum_{x: g(x)=y} p_X(x) \\
&= \sum_x \sum_{y=g(x)} y p_X(x) = \\
&= \sum_x g(x) p_X(x)
\end{aligned}$$

The second line is the definition of pmf $p_Y(y)$. In the third line we divide the event $\{g(X) = y\}$ into disjoint events $A_x = \{X = x\}$ where x runs over all possibilities such that $g(x) = y$. By additivity of probability for unions of disjoint events (Axiom 3), we have $\mathbb{P}(g(X) = y) = \sum_{x:g(x)=y} \mathbb{P}(X = x) = \sum_{x:g(x)=y} p_X(x)$.

The fourth line changes the order of summation by x and by y. And in the fifth line we recognize that the summation over y is actually trivial: for a given x, there is only one element y, such that $y = g(x)$. So we omit this summation sign and replace y with $g(x)$. This gives the desired formula. □

Example 2.2.9. Let us return to the setup of Example 2.2.4.

Professor of Economics (after some thought): A player should maximize her expected utility, for example $U(w) = \log(w)$, where $w =$ wealth. How do we calculate the expected utility? If the initial wealth is w_0, the player pays $\xi \le w_0$ for the game, and the game stops at round j, then the wealth of the player is $W = w_0 - \xi + 2^{j-1}$. This outcome happens with probability $\frac{1}{2^j}$. So the expected utility is

$$\mathbb{E}U = \mathbb{E}\log(W) = \sum_{j=1}^{\infty} \frac{1}{2^j} \log(w_0 - \xi + 2^{j-1}).$$

The utility if the payment is not made and the game is not played is $\log(w_0)$, and therefore the player should pay no more than $\overline{\xi}$, where $\overline{\xi}$ is the solution of the following equation:

$$\sum_{j=1}^{\infty} \frac{1}{2^j} \log(w_0 - \overline{\xi} + 2^{j-1}) = \log(w_0),$$

otherwise his expected utility is smaller than the utility from his original wealth.

2.2.3 Properties of the expected value

First of all note, that a constant b is a valid random variable. It is not random in the usual sense, since it is always the same for every outcome of the experiment. However, we can still think about it as function on the outcomes of the random experiment. This function maps every outcome to this constant b. The pmf of this random variable is $p(x) = 1$ for $x = b$ and 0 for all other values of x. Then by definition $\mathbb{E}(b) = bp(b) = b \times 1 = b$. So the expectation of a constant equals to the value of this constant. For example $\mathbb{E}(-1) = -1$, $\mathbb{E}(5) = 5$ and so on.

Theorem 2.2.10. *Let a and b be constant and X be a r.v. Then $\mathbb{E}(aX+b) = a\mathbb{E}X + b$.*

Proof.

$$\mathbb{E}(aX + b) = \sum_x (ax + b)p_X(x) = a\sum_x xp_X(x) + b\sum_x p_X(x) = a\mathbb{E}X + b.$$

□

Example 2.2.11. Suppose X is the number of gadgets that a factory makes in a month, $p = \$100$ is the price of each gadget and $c = 10,000$ is the fixed cost of production per month. If the average number of gadgets produced in a month is 250, what is the expected profit?

Profit is revenue minus cost $\pi = pX - c$, so the expected profit is $\mathbb{E}\pi = p\mathbb{E}X - c = 100 \times 250 - 10,000 = 15,000$.

Theorem 2.2.12. *Let $X_1, X_2, \dots X_k$ be r.v.'s, then*

$$\mathbb{E}(X_1 + X_2 + \ldots + X_k) = \mathbb{E}X_1 + \mathbb{E}X_2 + \ldots + \mathbb{E}X_k.$$

Proof. Here it is useful to think about each random variable X_j as a function of the experiment outcome ω, that has the pmf $p(\omega)$. Then, we can apply Theorem 2.2.6, thinking about $X_1 + X_2 + \ldots + X_k$ as a function of the random variable ω. [After some thought, we can see that it does not matter for this theorem that ω is not necessarily a real number.]

$$\begin{aligned}\mathbb{E}(X_1 + X_2 + \ldots + X_k) &= \sum_{\omega} \Big[\sum_{j=1}^{k} X_j(\omega)\Big] p(\omega) \\ &= \sum_{j=1}^{k} \sum_{\omega} X_j(\omega) p(\omega) = \sum_{j=1}^{k} \mathbb{E}X_j.\end{aligned}$$

□

Example 2.2.13. Let a coin be tossed n times. The results of the tosses are independent and the probability to get heads is p. Let X be the number of heads we observed in the experiment. What is $\mathbb{E}X$?

We can solve this example by two different methods. First, by definition and by the expression for the pmf in Example 2.1.6,

$$\begin{aligned}\mathbb{E}X &= \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} \\ &= \sum_{k=1}^{n} n \binom{n-1}{k-1} p^k (1-p)^{n-k} \\ &= pn \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} (1-p)^{n-1-(k-1)} \\ &= pn \sum_{l=0}^{n-1} \binom{n-1}{l} p^l (1-p)^{n-1-l} \\ &= pn.\end{aligned}$$

The equality in the second line is an identity which is easy to check from the definition of the binomial coefficient. In the third line we took pn outside

of the summation sign. In the fourth line we changed variable: $l = k - 1$. In the fifth line we used the fact that $\binom{n-1}{l} p^l (1-p)^{n-1-l}$ is the pmf for the random variable which is equal to the number of heads in the sequence of $n-1$ coin tosses. So by the property of pmf, the sum of this quantity over all values of the random variable equals 1.

You can see that this sum is not very difficult to calculate. However, the second method is much simpler.

Let I_k be the indicator function for the event H_k that in the k-th toss we obtain heads. Then, it is clear that our random variable $X = I_1 + I_2 + \ldots + I_n$, so

$$\begin{aligned} \mathbb{E}X &= \mathbb{E}(I_1 + I_2 + \ldots + I_n) = \mathbb{E}I_1 + \mathbb{E}I_2 + \ldots + \mathbb{E}I_n \\ &= np, \end{aligned}$$

where we used the fact from Example 2.2.5, that $\mathbb{E}I_j = \mathbb{P}(H_j) = p$.

Example 2.2.14. Suppose that n people are sitting around a table, and that everyone at the table gets up, runs around the room, and sits down randomly (i.e., all seating arrangements are equally likely).

What is the expected value of the number of people sitting in their original seat?

Let I_k be the indicator of the event A_k that a person k will be in the same sit after the permutation. Then, if X denotes the number of people sitting in their original seat, we have $X = I_1 + \ldots + I_n$, and

$$\mathbb{E}X = \mathbb{E}I_1 + \ldots + \mathbb{E}I_n = \mathbb{P}(A_1) + \ldots + \mathbb{P}(A_n)$$

But the total number of sits is n and after the permutation a person k is equally likely to sit in any of them. So the probability $\mathbb{P}(A_k) = 1/n$. So we have

$$\mathbb{E}X = n \times \frac{1}{n} = 1.$$

So, on average, one person will be sitting in his or her original sit.

2.3 Variance and standard deviation

2.3.1 Definition

Definition 2.3.1. The variance of a random variable X is defined by the following formula,

$$\mathbb{V}\text{ar}(X) = \mathbb{E}[(X - \mu)^2],$$

where $\mu = \mathbb{E}(X)$.

Intuitively, it is the average *squared* deviation of the random variable from its expectation.

It is often more convenient to calculate the deviation in the same units as the original variable. This means that we should take the square root of the variance.

Definition 2.3.2. The standard deviation of X is $\sqrt{\mathbb{V}\text{ar}(Y)}$.

It is often denoted $std(X)$ or simply σ. Then variance is σ^2.

In many cases, it is simpler to compute variance using a different formula.

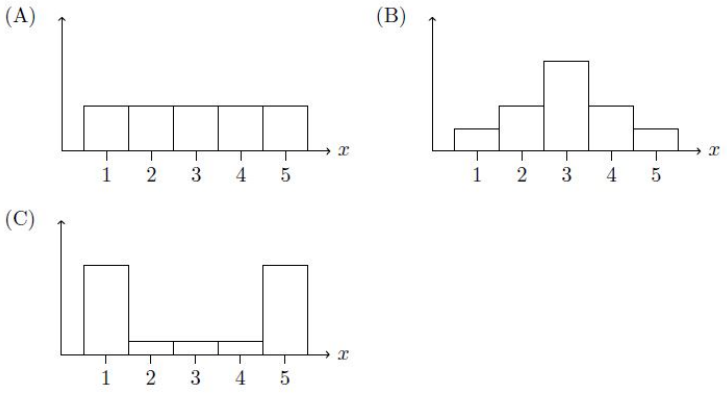
Theorem 2.3.3 (A different formula for variance)**.**

$$\mathbb{V}\text{ar}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2.$$

Proof. By expanding $(X - \mu)^2$ in the definition, we get

$$\begin{aligned}
\mathbb{V}\text{ar}(X) &= \mathbb{E}(X^2 - 2\mu X + \mu^2) \\
&= \mathbb{E}(X^2) - 2\mu\mathbb{E}(X) + \mu^2 \\
&= \mathbb{E}(X^2) - \mu^2 \\
&= \mathbb{E}(X^2) - (\mathbb{E}X)^2
\end{aligned}$$

□



Exercise 2.3.4.

These graphs show the pmf for 3 random variables. Order them by the size of the standard deviation from biggest to smallest.

Example 2.3.5. Suppose X has pmf:

x	1	2	3	4	5
$p(x)$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{4}{10}$	$\frac{2}{10}$	$\frac{1}{10}$

Find the expectation, variance and the standard deviation of X.

$$\mathbb{E}X = 1 \times \frac{1}{10} + 2 \times \frac{2}{10} + 3 \times \frac{4}{10} + 4 \times \frac{2}{10} + 5 \times \frac{1}{10} = 3$$
$$\mathbb{E}X^2 = 1 \times \frac{1}{10} + 4 \times \frac{2}{10} + 9 \times \frac{4}{10} + 16 \times \frac{2}{10} + 25 \times \frac{1}{10} = 10.2$$
$$\mathbb{V}\mathrm{ar}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 = 10.2 - 9 = 1.2$$
$$std(X) = \sqrt{1.2} \approx 1.0954$$

2.3.2 Properties

Theorem 2.3.6. *The variance and standard deviation have the following properties:*

1. $\mathbb{V}\mathrm{ar}(X) \geq 0$.
2. *If X is constant, $X = c$, then* $\mathbb{V}\mathrm{ar}(X) = 0$.
3. *If* $\mathbb{V}\mathrm{ar}(X) = 0$, *then $X = \mu$ with probability* 1.
4. *If a and b are constants, then*

$$\mathbb{V}\mathrm{ar}(aX + b) = a^2\mathbb{V}\mathrm{ar}(X),$$

5.

$$std(aX + b) = |a|std(X).$$

Proof. We will assume in the proof that the random variable X is discrete, although the result is valid in general.

(1) The variance is the expectation of a non-negative random variable, and so it is non-negative.

(2) This follows by direct calculation after we notice that $\mathbb{E}X = c$, and so $X - \mathbb{E}X = 0$ for all outcomes ω.

(3) If variance is 0 then all values of X that have positive probability must be equal to 0 otherwise the expectation of the squared difference would be positive.

(4) We can calculate:

$$\begin{aligned}
\mathbb{V}\mathrm{ar}(aX + b) &= \mathbb{E}\Big[(aX + b)^2\Big] - \Big[\mathbb{E}(aX + b)\Big]^2 \\
&= a^2\mathbb{E}X^2 + 2ab\mathbb{E}X + b^2 - \Big[a\mathbb{E}X + b\Big]^2 \\
&= a^2\Big[\mathbb{E}X^2 - (\mathbb{E}X)^2\Big] \\
&= a^2\mathbb{V}\mathrm{ar}(X).
\end{aligned}$$

(5) This follows from (4) and the definition of the standard deviation.

□

For expected values we have seen that $\mathbb{E}(X_1 + \ldots + X_n) = \mathbb{E}X_1 + \ldots + \mathbb{E}X_n$. Is a similar property holds for variance? In general, the answer is "no". This property holds if random variables $X_1, \ldots, X_n$ are independent or more generally, if they are uncorrelated. However, we postpone the discussion of the independence of several random variables to a later chapter.

Example 2.3.7. Let X be a non-constant random variable, so that $\mathbb{V}\mathrm{ar}(X) > 0$, and let $Y = -X$, that is, for every outcome ω, $Y(\omega) = -X(\omega)$. Then $\mathbb{V}\mathrm{ar}(Y) = \mathbb{V}\mathrm{ar}(X) > 0$. However, $X + Y = 0$ and therefore

$$\mathbb{V}\mathrm{ar}(X + Y) = 0 \neq \mathbb{V}\mathrm{ar}(X) + \mathbb{V}\mathrm{ar}(Y).$$

2.4 The zoo of discrete random variables

Here we will talk about some models for discrete random variables which frequently occur in practice.

- Binomial (and its simplest case, Bernoulli)
- Geometric
- Poisson
- Negative binomial
- Hypergeometric

2.4.1 Binomial

Recall Example 1.4.7, in which a coin was tossed n times. The results of different tosses are assumed to be independent and the probability to see heads is p in each toss. Then the random variable X is defined as the number of heads in these n tosses. Its pmf is

$$p_X(k) := \mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \text{ for } k = 0, 1, \ldots, n.$$

A random variable with this pmf is called a *binomial random variable* with parameters n and p.

(For $n = 1$ this random variable is called *Bernoulli* r.v. It takes values 1 or 0 with probability p and $1 - p$, respectively.)

Many random variables encountered in practice are binomial random variables.

Example 2.4.1. Suppose 2% of all items produced from an assembly line are defective. We randomly sample 50 items and count how many are defective. This count is a binomial random variable with parameters $n = 50$ and $p = 2\%$.

A binomial r.v. occurs in a binomial experiment, which is an experiment with the following characteristics:

1. There are a fixed number, denoted n, of identical trials.

2. Each trial results in one of two outcomes (which we can denote "S = Success" or "F = Failure").
3. The probability of "Success" is constant across trials and denoted p. Hence $\mathbb{P}$["Failure"]$= q = 1 - p$.
4. The trials are independent.

The total number of successes in a binomial experiment is a binomial r.v.

Here is another example.

Example 2.4.2. Suppose 40% of students at a college are male. We select 10 students at random and count how many are male. Is this a binomial experiment? What are parameters?

How do we calculate various probabilities for the binomial r.v.?

Example 2.4.3. Suppose 2% of items produced from an assembly line are defective. If we sample 10 items, what is the probability that 2 or more are defective?

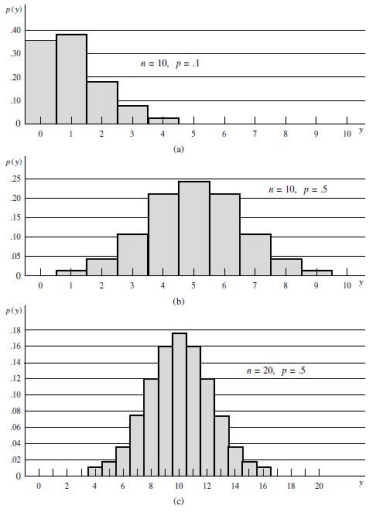


Figure 2.2: A historgram for binomial r.v. for various values of parameters.

Method 0.
Calculate with a calculator.

$$\begin{aligned}\mathbb{P}(Y \geq 2) &= 1 - \mathbb{P}(Y = 0) - \mathbb{P}(Y = 1) \\ &= 1 - \binom{10}{0}(0.2)^0(1-0.2)^{10} - \\ &- \binom{10}{1}(0.2)^1(1-0.2)^9 \\ &= 1 - 0.8^{10} - 10 \times 0.2 \times 0.8^9 = 0.6241903616\end{aligned}$$

Usually, we use tables or software.

Method 1, due to ancient Babylonians: lookup in the tables.[1]

[1] `https://www.theguardian.com/lifeandstyle/2014/may/17/ask-a-grown-up-who-invented-times-tables`

Table 1, Appendix 3 in the textbook gives cumulative probabilities $\mathbb{P}(Y \leq a)$. Looking it up in Table 1(b) for $n = 10$, we find that $\mathbb{P}(Y \leq 1) = .376$, hence $\mathbb{P}(Y \geq 2) = 1 - .376 = 0.624$.

(b) $n = 10$

	p													
a	0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95	0.99	a
0	.904	.599	.349	.107	.028	.006	.001	.000	.000	.000	.000	.000	.000	0
1	.996	.914	.736	.376	.149	.046	.011	.002	.000	.000	.000	.000	.000	1
2	1.000	.988	.930	.678	.383	.167	.055	.012	.002	.000	.000	.000	.000	2
3	1.000	.999	.987	.879	.650	.382	.172	.055	.011	.001	.000	.000	.000	3
4	1.000	1.000	.998	.967	.850	.633	.377	.166	.047	.006	.000	.000	.000	4
5	1.000	1.000	1.000	.994	.953	.834	.623	.367	.150	.033	.002	.000	.000	5
6	1.000	1.000	1.000	.999	.989	.945	.828	.618	.350	.121	.013	.001	.000	6
7	1.000	1.000	1.000	1.000	.998	.988	.945	.833	.617	.322	.070	.012	.000	7
8	1.000	1.000	1.000	1.000	1.000	.998	.989	.954	.851	.624	.264	.086	.004	8
9	1.000	1.000	1.000	1.000	1.000	1.000	.999	.994	.972	.893	.651	.401	.096	9

Figure 2.3: Table 1(b).

Method 2. Modern, with R software. R is both a computer language and a software tool which is popular among statisticians.

One can download and install R software from `https://cran.rstudio.com/`, and RStudio (an IDE, integrated development environment) from `https://www.rstudio.com/products/rstudio/download/`.

Then, we calculate $\mathbb{P}(Y \geq 2)$ by issuing the following command:

$$1 - \text{pbinom}(1, 10, 0.2)$$

or, alternatively,

$$\text{pbinom}(1, 10, 0.2, \text{lower.tail} = \text{F})$$

Answer: 0.6241904.

Method 3. Using Python

Python is popular among data scientists. It can be installed in many ways. The easiest is to download Anaconda from `https://www.anaconda.com/products/individual` and then install either Jupyter or Spyder from Anaconda.

Then one needs to start one of these programs, import package scipy.stats,

```
import scipy.stats as st
```

and then one can calculate the probabilities as

```
1 - st.binom.cdf(1, 10, 0.2)
```

Answer: 0.6241903616

Properties of the binomial r.v.

Theorem 2.4.4. *Let X be a binomial r.v. with n trials and success probability p. Then,*

1. $\mathbb{E}(X) = np$
2. $\mathbb{V}\mathrm{ar}(X) = npq$, *where* $q := 1 - p$.

Proof. (1) We have already proved that the expectation is np in the coin-tossing example.

(2) $\mathbb{V}\mathrm{ar}(X) = \mathbb{E}[X^2] - (\mathbb{E}(X))^2$.

It is not easy to find $\mathbb{E}(X^2)$ directly from the definition but we can use a trick. First we calculate

$$\begin{aligned}\mathbb{E}[X(X-1)] &= \sum_{k=0}^{n} k(k-1)\frac{n!}{k!(n-k)!}p^k q^{n-k} \\ &= \dots \\ &\text{(transformations similar to those done for the expected value)} \\ &= n(n-1)p^2.\end{aligned}$$

This allows us to calculate

$$\begin{aligned}\mathbb{E}(X^2) &= \mathbb{E}[X(X-1)] + \mathbb{E}(X) \\ &= n(n-1)p^2 + np\end{aligned}$$

Finally,

$$\begin{aligned}\mathbb{V}\mathrm{ar}(X) &= \mathbb{E}(X^2) - (\mathbb{E}X)^2 \\ &= n(n-1)p^2 + np - (np)^2 \\ &= -np^2 + np = np(1-p)\end{aligned}$$

□

There is alternative easier proof. However, it uses a fact that we will prove only later. First, we note that

$$X = I_1 + I_2 + \ldots + I_n,$$

where I_k is the indicator of the event that we have a success in the k-th trial.

It is easy to check that the variance of each I_k is $p(1-p)$. (We did it in one of the examples.) In addition, it turns out that the random variables I_k are *independent* from each other. We will study the independence of random variables later, and basically, it means that if know the value of one random variable, it does not change the probability that the other random variables takes some specific value. In our example, the independence of I_k is a consequence of the assumption that the results of different trials are independent.

It turns out that for *independent* random variables the variance of the sum equals the sum of the variances. So, in our particular case,

$$\mathbb{V}\mathrm{ar}(X) = \mathbb{V}\mathrm{ar}I_1 + \mathbb{V}\mathrm{ar}I_2 + \ldots + \mathbb{V}\mathrm{ar}I_n = np(1-p).$$

Example 2.4.5. 40% of students in a college are male. 10 students are selected. Let Y be the number of male students in the sample. Find $\mathbb{E}Y$ and $\mathbb{V}\mathrm{ar}(Y)$.

We calculate, $\mathbb{E}Y = np = 10 \times 0.4 = 4$ and $\mathbb{V}\mathrm{ar}(Y) = np(1-p) = 10 \times 0.4 \times 0.6 = 2.4$.

Exercise 2.4.6. Let $X \sim \mathrm{Binom}(n,p)$ and $Y \sim \mathrm{Binom}(m,p)$ be independent.

Then $X+Y$ distributed according to the following distribution:

(A) $\mathrm{Binom}(n+m,p)$

(B) $\mathrm{Binom}(nm,p)$

(C) $\mathrm{Binom}(n+m,2p)$

(D) other

Exercise 2.4.7. Let $X \sim \mathrm{Binom}(n,p_1)$ and $Y \sim \mathrm{Binom}(n,p_2)$ be independent.

Then $X+Y$ distributed according to the following distribution:

(A) Binom$(n, p_1 + p_2)$
(B) Binom$(2n, p_1 + p_2)$
(C) Binom$(n, p_1 p_2)$
(D) other

2.4.2 Geometric r.v.

Consider the experiment with a series of identical and independent trials, each resulting in either a Success or a Failure. Now, let us allow the number of the trials in the experiment be variable. Namely, the experiment is stopped after the first success.

Possible outcomes in this experiment are $S, FS, FFS, FFFS, \ldots$

Let the random variable X be the number of the trial, on which the first success occurs. This variable is called the geometric random variable and it is said that it is distributed according to the geometric probability distribution.

What is its pmf?

If $\mathbb{P}(\text{“}Success\text{”}) = p$ and $\mathbb{P}(\text{“}Failure\text{”}) = q = 1 - p$, then by the independence of trial results we can calculate:

$$\begin{aligned}
&\mathbb{P}(X = 1) = \mathbb{P}(S) = p,\\
&\mathbb{P}(X = 2) = \mathbb{P}(FS) = qp,\\
&\mathbb{P}(X = 3) = \mathbb{P}(FFS) = q^2 p,\\
&\qquad \ldots\\
&\mathbb{P}(Y = k) = \mathbb{P}(FF \ldots FS) = q^{k-1} p.
\end{aligned}$$

This is the pmf of the geometric r.v.

(Note this book uses a slightly non-standard definition. Often, the geometric r.v. is defined as the number of trials *before* the first success. So be careful when reading about geometric r.v. in other books or using software.)

We will indicate that X is a geometric r.v. with parameter p by writing $X \sim Geom(p)$ which reads “X is distributed according to the geometric distribution with parameter p.

Theorem 2.4.8. *If X is a geometric r.v., $X \sim Geom(p)$, then*

1. $\mathbb{E}(X) = \frac{1}{p}$
2. $\mathbb{V}\mathrm{ar}(X) = \frac{q}{p^2} = \frac{1-p}{p^2}$

Proof. (1) $\mathbb{E}(X) = \sum_{k=1}^{\infty} kq^{k-1}p$. How do we compute this sum?

Recall the formula for the geometric series: $\sum_{k=0}^{\infty} q^k = \frac{1}{1-q}$ and differentiate it over q:

$$\sum_{k=1}^{\infty} kq^{k-1} = \frac{d}{dq}\frac{1}{1-q} = \frac{1}{(1-q)^2}.$$

After we multiply the left-hand side of this equality by p, we will get exactly the expression for the expected value of the geometric variable. Hence,

$$\mathbb{E}(X) = \frac{p}{(1-q)^2} = \frac{1}{p}$$

(2) For the variance we can use a similar trick as for the binomial random variable, which means that we start by calculating $\mathbb{E}[X(X-1)]$:

$$\mathbb{E}[X(X-1)] := \sum_{k=2}^{\infty} k(k-1)q^{k-1}p,$$

and thee right hand side, after some small adjustment is exactly the second derivative of the series $1/(1-q) = \sum_{k=0}^{\infty} q^k$. So,

$$\begin{aligned}\mathbb{E}[X(X-1)] &= pq\frac{d^2}{dq^2}\Big[\frac{1}{1-q}\Big] \\ &= pq\frac{2}{(1-q)^3} = \frac{2q}{p^2}\end{aligned}$$

And then we can calculate the variance:

$$\begin{aligned}\mathbb{V}\mathrm{ar}(X) &= \mathbb{E}(X^2) - (\mathbb{E}X)^2 \\ &= \mathbb{E}[X(X-1)] + \mathbb{E}(X) - (\mathbb{E}X)^2 \\ &= \frac{2q}{p^2} + \frac{1}{p} - \frac{1}{p^2} = \frac{2q+p-1}{p^2} = \frac{q}{p^2}.\end{aligned}$$

□

2.4.3 Negative binomial

Negative binomial is a generalization of the geometric distribution.

Consider an experiment with a series of identical and independent trials, each resulting in either a Success or a Failure. Define a negative binomial r.v. Y as the number of the trial on which the *r-th* success occurs. The distribution of this random variable is negative binomial distribution with parameters p and r.

(If $r = 1$ then this Y is a *geometric* r.v. Also note that in many texts and software packages, the negative binomial r.v. is define as the number of trials *before* the r-th success.)

Compare the following descriptions of the binomial and the negative binomial random variables:

- For the binomial r.v., we fix the number of trials and count the number of *successes* obtained.
- For the negative binomial r.v., we fix the number of successes and count the number of *trials* needed to achieve this number of successes.

What is the pmf of a negative binomial r.v.? That is, what is the probability that the r-th success occurs exactly on trial k?

This event happens provided the first $k-1$ trials contain $r-1$ successes and the k-th trial is success. The probability of the former of these events is the probability that a binomial random variable with parameters $(k-1, p)$ takes value $r-1$, and the probability of the latter is p.

These two sub-events are independent and so the probability that $X = k$ is

$$p_X(k) = \binom{k-1}{r-1} p^{r-1} q^{k-r} \times p = \binom{k-1}{r-1} p^r q^{k-r}, \text{ where } k \geq r.$$

This is the pmf of the negative binomial distribution with parameters r and p. Note that it resembles the binomial distribution but somewhat different.

Theorem 2.4.9. *If X is a negative binomial distribution, $X \sim NB(r, p)$, then*

1. $\mathbb{E}(X) = \frac{r}{p}$
2. $\mathbb{V}\mathrm{ar}(X) = \frac{r(1-p)}{p^2}$

Proof. Negative binomial r.v. is a sum of r independent geometric r.v.'s with parameter p,

$$X = Y_1 + Y_2 + \ldots + Y_r.$$

Hence,

$$\mathbb{E}X = \mathbb{E}Y_1 + \ldots + \mathbb{E}Y_r = \frac{r}{p},$$
$$\mathbb{V}\mathrm{ar}(X) = \mathbb{V}\mathrm{ar}Y_1 + \ldots + \mathbb{V}\mathrm{ar}Y_r = \frac{rq}{p^2}$$

□

The argument in the proof of this theorem is also used in the solution of the following well-known problem.

Example 2.4.10 (Coupon Collector Problem). You collect coupons which comes with some items (like oatmeal). Every time you buy an item you get a coupon (like a sticker). There are N different types of coupons, they are all equally likely, so if you buy an item, then the probability to get a particular type of a coupon is $\frac{1}{N}$.

You buy one item every day. What is the expected number of days before you get a complete collection of coupons? (That is, how many items you need to buy?)

Let $Y_1, Y_2, \ldots, Y_N$ be the waiting times: namely, Y_j is the time between the moment when you collection got to the size $j-1$ and the moment when it got to the size j. ($Y_1 = 1$ and other variables are random.) We want to calculate is $\mathbb{E}X$, where X is the waiting time until all coupons are collected, and so:

$$X = Y_1 + Y_2 + \ldots + Y_N.$$

When you have $j-1$ different coupons, the probability that the next item contains a new coupon is

$$p_j = \frac{N-j+1}{N}.$$

The random variable Y_j is geometric with parameter p_j and expectation $\mathbb{E}Y_j = 1/p_j$.

Therefore,

$$\mathbb{E}X = \mathbb{E}Y_1 + \ldots \mathbb{E}Y_N = N\Big(\frac{1}{N} + \frac{1}{N-1} + \ldots + \frac{1}{2} + 1\Big) = NH_N,$$

where $H_N = 1 + \frac{1}{2} + \ldots \frac{1}{N}$ are called the harmonic numbers. The sum can be approximated by the integral, so

$$H_n \approx \int_1^N \frac{1}{x}\, dx = \log N.$$

So approximately, the expected number of items which you need to purchase to collect all N coupons is $N \log N$.

Example 2.4.11. Suppose 40% of employees at a firm have traces of asbestos in their lungs. The firm is asked to send 3 of such employees to a medical center for further testing.

1) Find the probability that exactly 10 employees must be checked to find 3 with asbestos traces.

We identify this distribution as the negative binomial with parameters $r = 3$ and $p = 40\%$.

$$\mathbb{P}(X = 10) = \binom{9}{3-1} 0.4^3 \times (1 - 0.4)^{10-3} = 0.0645$$

2) What is the expected number of employees that must be checked?

$$\mathbb{E}(X) = \frac{r}{p} = \frac{3}{0.4} = 7.5$$

3)If X is the number of employees that must be checked, what is $\mathbb{V}\mathrm{ar}(X)$ and σ?

$$\mathbb{V}\mathrm{ar}(X) = \frac{rq}{p^2} = \frac{3 \times 0.6}{0.4^2} = 11.25$$
$$Std(X) = \sqrt{11.25} \approx 3.354$$

Calculation of negative binomial and geometric variables in R and Python.

Example 2.4.12. Suppose that the probability of an applicant passing a driving test is 0.25 on any given attempt and that the attempts are independent. What is the probability that his initial pass is on his fourth try?

What is $\mathbb{P}(Y = 4)$? In R, this can be obtained by issuing the command:

```
dgeom(4-1, prob = 0.25)
```

Note that dgeom(...) stands for pmf and pgeom(...) stands for cdf of a geometric random variable in R. Also note that we used $4-1 = 3$ in the formula because in R, the geometric random variable is defined as the number of failures *before* the first success.

Example 2.4.13. Suppose 40% of employees at a firm have traces of asbestos in their lungs. The firm is asked to send 3 of such employees to a medical center for further testing.

Find the probability that *at most* 10 employees will be checked to find 3 with asbestos traces.

We need to calculate $\mathbb{P}(X \leq 10)$. Note that by definition this the cdf of X evaluated at 10.

```
pnbinom(10-3, size = 3, prob = 0.40)
```

The answer is 0.8327102.

Note that we use $10 - 3$ because the negative binomial random variable is defined as the number of *failures before* the given number of successes.

In Python, the code would be

```
import scipy.stats as st
st.nbinom.cdf(10 - 3, 3, 0.4)
```

The answer is 0.8327102464.

2.4.4 Hypergeometric

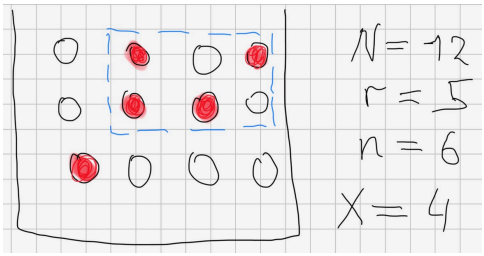


Figure 2.4: Hypergeometric random variable

Let an urn contain N balls and r of them are red (or "marked"). Others are white. Consider an experiment in which we take n balls without replacement. Let the random variable X be the number of red balls in the sample. The random variable X has the *hypergeometric* distribution with parameters N, r, n.

We can think about the experiment as taking n balls one-by-one from the box. If we take out a red ball, we mark the result of the trial as success. The random variable X is the number of successes.

This experiment is often called "sampling without replacement". The difference of this experiment from the binomial experiment is that the result of each trial affects results of the next trials. This implies that the trials are not independent. In particular, the probability to get a red ball in later trials depends on the results of earlier trials.

The probability mass function for X is

$$p_X(k) = \frac{\binom{r}{k}\binom{N-r}{n-k}}{\binom{N}{n}}$$

Proof. There are $\binom{N}{n}$ different choices of n balls out of N. This is the total number of outcomes in this experiment and all of them are equally probable.

If the sample contains k red balls then they could be chosen in $\binom{r}{k}$ different ways. At the same time the $n-k$ white balls could be chosen in $\binom{N-r}{n-k}$ different ways.

So the total number of experiment outcomes with k red balls is $\binom{r}{k}\binom{N-r}{n-k}$.

By the sample point method,

$$\mathbb{P}(X = k) = \frac{\binom{r}{k}\binom{N-r}{n-k}}{\binom{N}{n}}$$

□

Properties of the hypergeometric distribution

Theorem 2.4.14. *If $X \sim HyperGeom(N, r, n)$, then*

1. $\mathbb{E}(X) = n\left(\frac{r}{N}\right)$
2. $\mathbb{V}\text{ar}(X) = n\left(\frac{r}{N}\right)\left(1 - \frac{r}{N}\right)\left(\frac{N-n}{N-1}\right)$

We leave this theorem without proof.

Note the remarkable resemblance with the mean and the variance of the binomial distribution $Bin(p, n)$ if we set $p = r/N$.

The only difference is that the variance is multiplied by $\left(\frac{N-n}{N-1}\right)$.

This factor is sometimes called "finite population adjustment".

Example 2.4.15. From a set of 20 potential jurors (8 African-American and 12 white) 6 jurors were selected. If the jury selection was random, what is the probability of one or fewer African Americans on the jury? What is the expected number of African-American on the jury? What is the standard deviation?

We need to calculate $\mathbb{P}(X \leq 1)$. We can either do it by using a calculator:

$$\begin{aligned}\mathbb{P}(X \leq 1) &= \mathbb{P}(X = 0) + \mathbb{P}(X = 1) \\ &= \frac{\binom{8}{0}\binom{12}{6}}{\binom{20}{6}} + \frac{\binom{8}{1}\binom{12}{5}}{\binom{20}{6}} \\ &\approx 18.7\%,\end{aligned}$$

or by using a statistical software package. For example, in R:

```
phyper(1, 8, 12, 6)
```

gives the same answer 18.7%.

The expected value is $\mathbb{E}X = 8/20 \times 6 = 2.4$ and the standard deviation is

$$Std(X) = \sqrt{6 \times \frac{8}{20} \times \frac{12}{20} \times \frac{20-6}{20-1}} \approx 1.03$$

Exercise 2.4.16. An urn contains 20 marbles, of which 10 are green, 5 are blue, and 5 are red. 4 marbles are to be drawn from the urn, one at a time

without replacement. Let Y be the total number of green balls drawn in the sample.

$\mathrm{Var}(Y) =?$

2.4.5 Poisson r.v.

Consider an experiment in which a Geiger counter counts the number of radioactive particles that entered a gas camera in one second. The experiment has essentially a continuum of outcomes since a particle can enter a camera at any time, so the outcome of this experiment might be described by a set of times at which particles entered the camera. However, we are simply interested in the discrete random variable which is the size of this set. Given that the particles enter the camera independently this random variable is called the Poisson random variable. It characterized by a parameter λ which equals to its expected value.

Poisson random variable is useful to model

- Y = number of phone calls received per day.
- Y= number of accidents per week at an intersection.
- Y = number of spots per square inch of an orange.
- Y = number of galaxies in a particular spot of Universe,

and many-many similar phenomena.

How do we obtained its pmf? We do not yet learned how to describe continuous experiments especially so complex experiments as described above. However, it turns out that we can find the pmf of the Poisson random variable by approximating the continuous experiment by the binomial experiment.

Let us divide the interval of time in n subintervals and let us assume that a particle can enter during each subinterval with probability p. Assume that n so large that we can ignore the situation when two particles enter the camera during the same sub-interval.

Now we have a binomial experiment with n trials corresponding to subintervals and the probability of a success p. Here a success means that a

particle entered the tube during the sub-interval. The expected number of successes is np and we fix it at λ.

Let X_n be the number of successes in this binomial experiment. Then

$$\begin{aligned}\mathbb{P}(X_n = k) &= \binom{n}{k} p^k (1-p)^{n-k} \\ &= \frac{n(n-1)\dots(n-k+1)}{k!} \Big(\frac{\lambda}{n}\Big)^k \Big(1-\frac{\lambda}{n}\Big)^{n-k}\end{aligned}$$

It is easy to check that if k is fixed and $n \to \infty$ then the factor

$$\frac{n(n-1)\dots(n-k+1)}{n^k} \to 1.$$

The remaining part can be written as

$$\frac{\lambda^k}{k!} \left(1-\frac{\lambda}{n}\right)^n \left(1-\frac{\lambda}{n}\right)^{-k}$$

By using calculus, we find the limits $(1-\lambda/n)^n \to e^{-\lambda}$ and $(1-\lambda/n)^k \to 1$.

It follows that as $n \to \infty$,

$$\mathbb{P}(X_n = k) \to \frac{\lambda^k}{k!} e^{-\lambda}$$

We conclude that the pmf of the Poisson random variable is

$$p_X(k) = \mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \text{ where } k = 0, 1, 2, \dots$$

We can also verify directly that this is a valid pmf by checking that it adds up to 1:

$$\sum_{k=0}^{\infty} p_X(k) = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = 1,$$

because

$$e^{\lambda} = 1 + \lambda + \frac{1}{2}\lambda^2 + \frac{1}{3!}\lambda^3 + \dots = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!},$$

is the Taylor series expansion for the function $f(\lambda) = e^{\lambda}$.

The parameter λ is called intensity.

Theorem 2.4.17. *If $X \sim Pois(\lambda)$, then*

1. $\mathbb{E}X = \lambda$
2. $\mathbb{V}\text{ar}(X) = \lambda$

Intuitively, Poisson r.v. is a limit of binomial r.v.'s, hence its expectation and variance are also limits: $np \to \lambda$, $np(1-p) \to \lambda$. If we want a more rigorous proof, we simply use the definition and our usual collection of tricks.

Proof. (1)Expectation:

$$\begin{aligned}\mathbb{E}(X) &= \sum_{k=0}^{\infty} k\frac{\lambda^k}{k!}e^{-\lambda} \\ &= \lambda\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!}e^{-\lambda} \\ &\lambda\sum_{l=0}^{\infty}\frac{\lambda^l}{l!}e^{-\lambda} \\ &= \lambda,\end{aligned}$$

where in the last line we used the observation that the sum is the sum of the pmf over all possible values, and so it equals 1.

(2) For variance, we do a similar calculation as the above calculation, but for $\mathbb{E}[X(X-1)]$. We find that it equals λ^2. This implies that $\mathbb{E}(X^2) = \lambda^2+\lambda$, and therefore $\mathbb{V}\text{ar}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$. □

Example 2.4.18. Suppose the number of accidents per month at an industrial plant has a Poisson distribution with mean 2.6. Find the probability that there will be 4 accidents in the next month. Find the probability of having between 3 and 6 accidents in the next month. What is the probability of 10 accidents in the next *half-year*?

Directly by the formula for the pmf of a Poisson r.v.,

$$\mathbb{P}(X = 4) = e^{-2.6} \times \frac{2.6^4}{4!} \approx 14.14\%.$$

The cumulative Poisson probabilities can be evaluated either using Table 3 in Appendix 3 or by using R: $P(X \leq x) = ppois(x, \lambda)$. In this example, we have

$$\begin{aligned}\mathbb{P}(3 \leq Y \leq 6) &= \mathbb{P}(Y \leq 6) - \mathbb{P}(Y \leq 2) \\ &= 98.28\% - 51.84\% = 46.44\%\end{aligned}$$

The next question illustrates a very useful property of the Poisson random variable. Recall that we have seen that it two binomial random variables have parameters (n_1, p) and (n_2, p) respectively, and if they are independent, then their sum is also a binomial random variable with parameters $(n_1 + n_2, p)$. Since a Poisson random variable is a limit of binomial random variables, it is not surprising that a similar property holds for Poisson random variables.

Namely, if X_1 and X_2 are two Poisson random variables with parameters λ_1 and λ_2 respectively, and if they are independent then $X_1 + X_2$ is also a Poisson random variable with parameter $\lambda_1 + \lambda_2$.

In Example 2.4.18, the number of accidents in a half-year is the sum of the number of accidents in every of the 6 months that constitutes the half-year. So it is a sum of six Poisson random variables with parameter 2.6. Assuming that these random variables are independent, we conclude the number of the accidents in the half-year is a Poisson random variable Y with parameter 6×2.6.

Therefore,

$$\mathbb{P}(Y = 10) = e^{-6\times 2.6} \times \frac{(6 \times 2.6)^{10}}{10!} \approx 3.95\%.$$

Exercise 2.4.19. Suppose the number of accidents per month at an industrial plant has a Poisson distribution with mean 2.6.

What is the expected number of accidents in the next half-year?

Exercise 2.4.20. A parking lot has two entrances. Cars arrive at entrance I according to a Poisson distribution at an average of 5 per hour and at entrance II according to a Poisson distribution at an average of 3 per hour. (Assume that the numbers of cars arriving at the two entrances are independent.)

The total number of cars that arrive at parking lot is a Poisson random variable. What is its expected value?

Exercise 2.4.21. The daily probability of no accidents on a given piece of a highway is e^{-2}. The number of accidents follows the Poisson distribution.

What is the expected number of accidents per day?

Poisson approximation for binomial distribution

If n is large, p is small, and $\lambda = np$ is somewhat small (book: ≤ 7), then the $Bin(n, p)$ probabilities are approximately equal to the $Pois(\lambda)$ probabilities.

Example 2.4.22. Suppose there are $10,000$ students in a college. 3% of all students are vegetarians. Select 100 students at random. Find the probability that the sample contains at least 5 vegetarians.

Exact probability:

$$\begin{aligned}\mathbb{P}(X \geq 5) = 1 - \mathbb{P}(X \leq 4) &= 1 - \sum_{j=0}^{4} \frac{\binom{300}{j}\binom{10{,}000-300}{100-j}}{\binom{10{,}000}{100}} \\ &= 1 - phyper(4, 300, 10000 - 300, 100) = 0.1812679\end{aligned}$$

Binomial approximation:

$$\begin{aligned}\mathbb{P}(X \geq 5) = 1 - \mathbb{P}(X \leq 4) &= 1 - \sum_{j=0}^{4} \binom{100}{j} 0.03^j (1 - 0.03)^{100-j} \\ &= 1 - pbinom(4, size = 100, prob = 0.03) = 0.1821452\end{aligned}$$

This is different in the 3rd significant digit.

Poisson approximation: Here $\lambda = np = 100 \times 0.03 = 3$.

$$\begin{aligned}\mathbb{P}(X \geq 5) = 1 - \mathbb{P}(X \leq 4) &= 1 - e^{-\lambda} \sum_{j=0}^{4} \frac{\lambda^{-j}}{j!} \\ &= 1 - ppois(4, lambda = 3) = 0.1847368,\end{aligned}$$

This is also different only in the 3rd significant digit.

Exercises

Exercise 2.4.23. Let X be a random variable whose Probability Mass Function is given as

$$P(X = k) = \frac{5^k}{e^5 k!}$$

for $k = 0, 1, \ldots$

What is the variance of this random variable?

Exercise 2.4.24. Consider the probability given by the expression:

$$\mathbb{P}(Y = 3) = \frac{\binom{2,000,000}{3}\binom{98,000,000}{97}}{\binom{100,000,000}{100}}$$

What is the appropriate value for parameter p in the binomial approximation for this probability?

What is the appropriate value for parameter λ in the Poisson approximation for this probability?

Exercise 2.4.25. Let a r.v. Y have probability function $p(y) = \left(\frac{1}{2}\right)^y$ for $y = 1, 2, 3, \ldots$.

What is the distribution of Y?

2.5 Moments and moment-generating function

2.5.1 Moments

Definition 2.5.1. The k-th moment of a r.v. X is defined as

$$\mu_k(X) = \mathbb{E}(X^k)$$

Examples:

- The first moment, μ_1 is $\mathbb{E}X$, the *expected value* of X, .
- The second moment μ_2 is closely related to variance:

$$\mu_2(X) := \mathbb{E}(X^2) = \mathbb{V}\text{ar}(X) + (\mathbb{E}X)^2.$$

In applications, it is usually more useful to look at the *central* moments, which are defined as

$$\mu_k^c = \mathbb{E}[(X - \mu_1)^k]. \tag{2.2}$$

For example, variance is by its definition the second central moment of the random variable X. The variance is typically denoted σ^2, with σ denoting the standard deviation.

Example 2.5.2. The *skewness* of a random variable X is defined μ_3^c/σ^3, and the *kurtosis* X is defined as μ_4^c/σ^4. These quantities are often used to describe the shape of the pmf of a random variable X when the variability has already been taken into account. The division by powers of σ is introduced so that skewness and kurtosis are scale invariant: $skew(aX) = skew(X)$ for all positive a and $kurtosis(aX) = kurtosis(X)$ for all non-zero a.

Note that if we expand the expression on the right-hand side of (2.2) we can express the central moments in terms of usual moments, similar as we did if for variance. For example,

$$\begin{aligned}\mu_3^c &= \mathbb{E}[(X-\mu_1)^3] = \mathbb{E}(X^3) - 3\mu_1(\mathbb{E}X^2) + 3\mu_1^2\mathbb{E}X - \mu_1^3 \\ &= \mu_3 - 3\mu_1\mu_2 + 2\mu_1^3\end{aligned}$$

2.5.2 Moment-generating function

One way to calculate the moments of a random variable, especially high-order moments is through the moment-generating function.

The moment-generating function is a variant of an important mathematical tool called the Fourier transform. In probability theory, the Fourier transform of the probability mass function is called the *characteristic function* of a random variable, and it is defined as

$$\varphi_X(t) = \mathbb{E}(e^{itX}) = \sum_x e^{itx} p_X(x),$$

where i is the imaginary unit of complex numbers. In this course, we will not assume the knowledge of complex numbers and consider a closely related object, the Laplace transform of the probability mass function. This transform is called the *moment-generating function*, or *mgf*, $m_X(t)$ and defined as

$$m_X(t) = \mathbb{E}(e^{tX}) = \sum_x e^{tx} p_X(x).$$

Both characteristic and moment-generating functions determine the random variable X (except for some random variables which have very large variability and rarely occur in practice).

An advantage of characteristic functions over moment-generating functions is that they are well-defined for a larger class of functions and have better analytical properties. However, it will be not important for us here.

The following theorem explains that if we know the moment-generating function we can easily calculate the moments of a random variable X.

Theorem 2.5.3. *If $m_X(t)$ is the moment-generating function for a random variable X, then for any integer $k \geq 1$,*

$$\mu_k = \frac{d^k}{dt^k} m_X(t)\Big|_{t=0}.$$

The Taylor series for $m_X(t)$ is

$$m_X(t) = 1 + \sum_{k=1}^{\infty} \frac{\mu_k}{k!} t^k.$$

Proof. Recall the Taylor series for the exponential function:

$$\begin{aligned} e^x &= 1 + t + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + \dots \\ &= \sum_{k=0}^{\infty} \frac{1}{k!} x^k \end{aligned}$$

So by definition of the moment-generating function,

$$\begin{aligned} m_X(t) := \mathbb{E}(e^{tX}) &= 1 + \mathbb{E}(tX) + \frac{1}{2!}\mathbb{E}(tX)^2 + \frac{1}{3!}\mathbb{E}(tX)^3 + \dots \\ &= \sum_{k=0}^{\infty} \frac{1}{k!} \mathbb{E}(tX)^k \\ &= 1 + t\mu_1 + \frac{t^2}{2!}\mu_2 + \frac{t^3}{3!}\mu_3 + \dots = \sum_{k=0}^{\infty} \frac{t^k}{k!}\mu_k \end{aligned}$$

Hence, by differentiating this expression k times, we get

$$\frac{d^k}{dt^k} m_X(t) = \mu_k + t\mu_{k+1} + \frac{t^2}{2}\mu_{k+2} + \dots$$

After setting t to zero in the resulting expression, we get:

$$\frac{d^k}{dt^k} m_X(t)\Big|_{t=0} = \mu_k.$$

which is what we wanted to prove. □

Example 2.5.4. Suppose a random variable X has the following moment generating function

$$m_X(t) = \frac{3}{4}e^t + \frac{1}{4}e^{t^2}$$

for all t.

(Remark: This random variable is not discrete. However, it could be checked that it is a valid pmf.)

Find $\mathbb{E}(X)$, $\mathbb{E}(X^2)$ and $\mathbb{V}\mathrm{ar}(X)$.

We calculate the first and the second derivatives:

$$m'_X(t) = \frac{3}{4}e^t + \frac{1}{2}te^{t^2},$$
$$m''_X(t) = \frac{3}{4}e^t + \frac{1}{2}e^{t^2} + t^2e^{t^2}.$$

After substituting $t = 0$, we get

$$\mathbb{E}X \equiv \mu_1 = m'_X(0) = \frac{3}{4}$$
$$\mathbb{E}X^2 \equiv \mu_2 = m''_X(0) = \frac{3}{4} + \frac{1}{2} = \frac{5}{4},$$

and for variance, we have

$$\mathbb{V}\mathrm{ar}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \frac{5}{4} - \frac{9}{16} = \frac{11}{16}.$$

2.5.3 Mgf characterizes the distribution

For most random variables, if a mgf exists, it completely characterizes the distribution.

Sometimes, we can write down the pmf simply by closely observing the mgf function.

Example 2.5.5. Let the mgf of a discrete random variable X be

$$m_X(t) = \frac{1}{10}e^{-t} + \frac{2}{10} + \frac{3}{10}e^{t} + \frac{4}{10}e^{2t}.$$

Find the pmf of X.

Recall the definition of the mgf for discrete random variables:

$$m_X(t) = \sum_x e^{tx} p_X(x).$$

In our example we have

$$m_X(t) = \frac{1}{10}e^{-t} + \frac{2}{10}e^{0t} + \frac{3}{10}e^{t} + \frac{4}{10}e^{2t},$$

so we see from the arguments of the exponential functions that the possible values for random variable X are $\{-1, 0, 1, 2\}$ and we simply read off the pmf of X as the coefficients before the exponentials: $p_X(-1) = 1/10$, $p_X(0) = 2/10$, $p_X(1) = 3/10$, $p_X(2) = 4/10$.

2.5.4 Factorization property for mgf

An important property of the moment-generating function is that it factorizes for sums of independent random variables.

Theorem 2.5.6. *Suppose $X_1, X_2, \ldots, X_n$ are independent random variables and $S = X_1 + \ldots + X_n$, then the mgf of S is the product of the mgf of $X_1, \ldots, X_n$.*

$$m_S(t) = m_{X_1}(t) m_{X_2}(t) \ldots m_{X_n}(t).$$

(A similar property holds for characteristic functions). This property allows us to study sums of independent random variables. We will prove this theorem later, when we will study the independence of several random variables.

2.5.5 Mgf of named discrete r.v.

Now let us compute the mgf of all discrete distribution that we already know, except for the hypergeometric distribution which is a bit more difficult.

Example 2.5.7 (mgf for the indicator random variable, aka Bernoulli r.v.). Let $X = I_A$ be the indicator random variable for an event A. Recall that I_A takes value 1 with probability $p = \mathbb{P}(A)$ and value 0 with probability $1 - p$. What is its mgf?

$$m_{I_A}(t) = \mathbb{E}(e^{tI_A}) = e^{t\times 1}p + e^{t\times 0}(1-p) = pe^t + 1 - p.$$

Example 2.5.8 (Binomial distribution). Binomial r.v. $X \sim Bin(n,p)$ is the sum of n independent indicator functions $I_1, \ldots, I_n$, where the I_j is the indicator of the event that we had a success in the j-th trial. The probability of the success is p.

$$X = I_1 + I_2 + \ldots + I_n.$$

By using Theorem 2.5.6 and the result from Example 2.5.7, we get

$$m_X(t) = \Big(m_{I_1}(t)\Big)^n = (pe^t + q)^n,$$

where $q = 1 - p$.

Example 2.5.9 (Geometric distribution). For the geometric distribution with parameter p we have by definition

$$\begin{aligned} m_X(t) = \mathbb{E}\Big(e^{tX}\Big) &= e^t p + e^{2t}pq + e^{3t}pq^2 + \ldots \\ &= pe^t\Big(1 + qe^t + (qe^t)^2 + \ldots\Big) \\ &= \frac{pe^t}{1 - qe^t}. \end{aligned}$$

Example 2.5.10 (Negative binomial distribution). Recall that we can write the negative binomial distribution with parameters r and p as a sum of r independent geometric random variables with parameter p. Hence, by using Theorem 2.5.6,

$$m_X(t) = \Big(\frac{pe^t}{1 - qe^t}\Big)^r.$$

Example 2.5.11 (Poisson distribution). For a Poisson r.v. X with parameter λ, we have, by definition:

$$\begin{aligned} m_X(t) = \mathbb{E}\Big(e^{tX}\Big) &= \sum_{k=0}^{\infty} e^{tk} \frac{\lambda^k}{k!} e^{-\lambda} \\ &= e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda e^t)^k}{k!} \\ &= e^{-\lambda} e^{\lambda e^t}, \end{aligned}$$

where in the last line we used the Taylor series for the exponential function $e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$, with $x = \lambda e^t$.

So, the mgf for the Poisson random variable is

$$m_X(t) = e^{\lambda(e^t - 1)}.$$

As we mentioned before, the mgf completely characterizes the random variable in the sense that if two r.v.'s have the same mgf then they must have the same distribution.

Example 2.5.12. Suppose X is a r.v. with mgf

$$m_X(t) = e^{7.1e^t - 7.1}.$$

What is the distribution of X?

By looking at the function attentively we see that it is the mgf of a Poisson random variable with parameter $\lambda = 7.1$.

2.6 Markov's and Chebyshev's inequalities

If we know the pmf of a random variable, then it is easy to calculate probabilities. In particular, if we know the type of the distribution and several first moments then we can find the parameters of the distribution, learn the pmf, and calculate the desired probabilities.

Example 2.6.1. Suppose X has the Poisson distribution and $\mathbb{E}X = 2$. What is the probability $\mathbb{P}(X \geq 3)$?

We know that the parameter of the Poisson distribution λ equals the expectation of X. So, $\lambda = 2$, and by using the formula for the pmf of the Poisson distribution $p_X(k) = e^{-\lambda}\lambda^k/k!$, we have

$$\begin{aligned}\mathbb{P}(X \geq 3) &= 1 - \mathbb{P}(X \leq 2) = 1 - \mathbb{P}(X = 0) - \mathbb{P}(X = 1) - \mathbb{P}(X = 2) \\ &= 1 - e^{-2}\Big(1 + \frac{2^1}{1!} + \frac{2^2}{2!}\Big) = \dots \\ &\text{(or, using software)} \\ &= 1 - ppois(2, \lambda = 2) = 32.33\%\end{aligned}$$

Example 2.6.2. Suppose X has the Geometric distribution and $\mathbb{E}X = 2$. What is the probability $\mathbb{P}(X \geq 3)$?

In this case we know that $\mathbb{E}X = 1/p$, where p is the parameter of the geometric distribution. So in our case, $p = 1/2$, and we calculate:

$$\begin{aligned}\mathbb{P}(X \geq 3) &= 1 - \mathbb{P}(X \leq 2) \\ &= 1 - pgeom(2 - 1, prob = 2) = 25\%\end{aligned}$$

(Here $2 - 1$ is used instead of 2 because, the geometric r.v. in R is defined as the number of failures before the first success so it is less than our X by one.)

So we see that for this random variable, the probability to get a value $X \leq 3$ is smaller than the probability for a Poisson random variable with the same expectation.

Example 2.6.3. Suppose X has the binomial distribution with $n = 10$ and $\mathbb{E}X = 2$. What is the probability $\mathbb{P}(X \geq 3)$?

We know that for a binomial r.v. $\mathbb{E}X = np$ so we find that in our example, $p = 2/10 = 0.2$. So we can calculate:

$$\begin{aligned}\mathbb{P}(X \geq 3) &= 1 - \mathbb{P}(X \leq 2) \\ &= 1 - pbinom(2, size = 10, prob = 0.2) = 32.22\%\end{aligned}$$

This is close to what we found for the Poisson random variable, which is not surprising since we know that the Poisson distribution is a good approximation for the binomial distribution.

But what if we simply know that the expectation of X is 2 and do not know what is exactly the pmf of X? Can we say something about $\mathbb{P}(X \geq 3)$. In general, no. However, if the random variable X is *non-negative*, then the random variable with a fixed average value (expectation) cannot attain large values too frequently, and there is a way to make this argument precise.

More generally, if we only know a couple of the first moments, such as the expected value and variance of random variable X, then sometimes some information about probabilities is still available. This information comes in the form of inequalities: "probability is greater than some quantity" or "probability is smaller than some quantity".

These inequalities are especially useful in theoretical analysis, when nothing is known about a random variable except its first moments and one still wants to make a 100% valid statement.

The first result is about *non-negative* random variables.

Theorem 2.6.4 (Markov's inequality)**.** *Let $X \geq 0$ be a* non-negative *r.v. Then for every $t > 0$, we have*

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}X}{t}.$$

Intuitively, this theorem says that if X is non-negative and is average $\mathbb{E}X$ is relatively small, then the probability that the random variable exceeds a large threshold t must be small.

Example 2.6.5. Suppose X is non-negative and $\mathbb{E}X = 2$. What can be said about the probability $\mathbb{P}(X \geq 3)$?

By Markov inequality, $\mathbb{P}(X \geq 3) \leq 2/3 \approx 66.6\%$.

Note that the probabilities that we calculated in previous cases all satisfy this inequality. Note also, that the inequality is not very strong. In the cases we considered, the actual probability $\mathbb{P}(X \geq 3)$ is far below 66.6%. This is why these inequalities are not very helpful in practical applications despite its usefulness in theoretical analysis.

Proof of Markov's inequality. By definition,

$$\begin{aligned}\mathbb{E}X &= \sum_x x p_X(x) = \sum_{x<t} x p_X(x) + \sum_{x\geq t} x p_X(x) \\ &\geq \sum_{x\geq t} x p_X(x) \\ &\geq \sum_{x\geq t} t p_X(x) \\ &= t \sum_{x\geq t} p_X(x) = t\mathbb{P}(X \geq t).\end{aligned}$$

where in the second line we used the fact that X takes only non-negative values x, so if we drop the first sum, the result will decrease.

Hence, by dividing by t,

$$\frac{\mathbb{E}X}{t} \geq \mathbb{P}(X \geq t),$$

which is what we wanted to prove. □

What if we want to says something about the random variables which can take both positive and negative values. The second famous inequality holds for all random variables that have a finite variance.

Theorem 2.6.6 (Chebyshev's inequality)**.** *Let X be a r.v. with mean $\mathbb{E}X = \mu$ and variance $\mathbb{V}\mathrm{ar}(X) = \sigma^2$. Then for every $t > 0$, we have*

$$\mathbb{P}\Big(|X - \mu| \geq t\sigma\Big) \leq \frac{1}{t^2}.$$

Intuitively, this theorem says that the probability that a random variable X deviates more than t standard deviations from its mean is less than $1/t^2$. So if t is large than the probability of this deviation is small. Note, however, that this theorem becomes helpful only if $t > 1$ because otherwise the inequality simply says that this probability is smaller than 1 and we know it without applying any inequalities.

Another popular (and equivalent) form, in which this theorem is stated is

$$\mathbb{P}\Big(|X - \mu| \geq \varepsilon\Big) \leq \frac{\sigma^2}{\varepsilon^2}.$$

It is easy to see that it is equivalent by substituting $t = \varepsilon/\sigma$ in the original theorem.

Proof of Theorem 2.6.6. Let us define a new random variable $Y = (X-\mu)^2$. Note that this random variable is non-negative and that its expectation equals to the variance of the original random variable X. By applying Markov's inequality to Y with threshold $(t\sigma)^2$, we find:

$$\mathbb{P}\Big((X-\mu)^2 \geq (t\sigma)^2\Big) \leq \frac{\mathbb{E}(X-\mu)^2}{(t\sigma)^2} = \frac{\sigma^2}{(t\sigma)^2} = \frac{1}{t^2}.$$

However the event $(X-\mu)^2 \geq (t\sigma)^2$ is the same as the event $|X-\mu| \geq t\sigma$ provided that $t > 0$. Hence, we find

$$\mathbb{P}\Big(|X-\mu| \geq t\sigma\Big) \leq \frac{1}{t^2}.$$

□

Example 2.6.7. Let X have mean 20 and standard deviation 2. What can be said about the probability that a new realization of X will be between 16 and 24?

Note that 16 and 24 are 20 ± 4 and threshold 4 is twice the standard deviation. So by using the Chebyshev's inequality with $t = 2$, we can estimate the probability that X is outside of the interval $(16, 24)$:

$$\mathbb{P}\Big(X \notin (16,24)\Big) = \mathbb{P}\Big(|X-20| \geq 4\Big) \leq \frac{1}{2^2} = \frac{1}{4}.$$

Hence,

$$\mathbb{P}(16 < X < 24) \geq 1 - \frac{1}{2^2} = 3/4.$$

It is worth repeating that Markov's and Chebyshev's inequalities give us only bounds on the probabilities and are useful only if we do not know what is the actual distribution but know the first moments of this distribution.

If we do know the distribution, we can calculate the probability explicitly.

Example 2.6.8. Let $X \sim \text{Binom}(n = 4, p = 1/2)$. What is the upper bound on $\mathbb{P}(X \geq 4)$ given by Markov's inequality? What is the upper bound on

$\mathbb{P}(X \geq 4)$ given by Chebyshev's inequality? What is the exact value of $\mathbb{P}(X \geq 4)$?

By Markov's inequality, we have

$$\mathbb{P}(X \geq 4) \leq \frac{\mathbb{E}(X)}{4} = \frac{np}{4} = \frac{2}{4} = 50\%.$$

By Chebyshev's inequality, we can write (by using the lucky fact that the interval $[0, 4]$ has its center at the $\mathbb{E}X = 2$):

$$\begin{aligned}\mathbb{P}(X \geq 4) &= \mathbb{P}(X \notin [0, 4]) = \mathbb{P}(|X - 2| \geq 2) \\ &\leq \frac{\mathbb{V}\text{ar}(X)}{2^2} = \frac{npq}{4} = \frac{4 \times 1/2 \times 1/2}{4} \\ &= \frac{1}{4} = 25\%\end{aligned}$$

The exact probability is

$$\begin{aligned}\mathbb{P}(X \geq 4) &= 1 - \mathbb{P}(X \leq 3) \\ &= 1 - pbinom(3, size = 4, prob = 1/2) \\ &= 6.25\%\end{aligned}$$

(Alternatively $\mathbb{P}(X \geq 4) = \mathbb{P}(X = 4)$ since the binomial r.v. cannot take values large than 4, and we can calculate $\mathbb{P}(X = 4) = \binom{4}{4}(1/2)^4(1 - 1/2)^0 = 1/16 = 0.0625$.)

Obviously the last result is much better than the bounds given by Markov's and Chebyshev's inequalities. The power of inequalities is that they work even if we have no idea about the exact distribution except for its expected value and variance.

Chapter 3

Chapter 4: Continuous random variables.

3.1 Cdf and pdf

Continuous r.v.'s take an uncountably infinite number of possible values. The probability of every specific number is zero and, for this reason, pmf (probability mass function) is not useful for continuous r.v.'s. However, the cdf (cumulative distribution function) is still well defined.

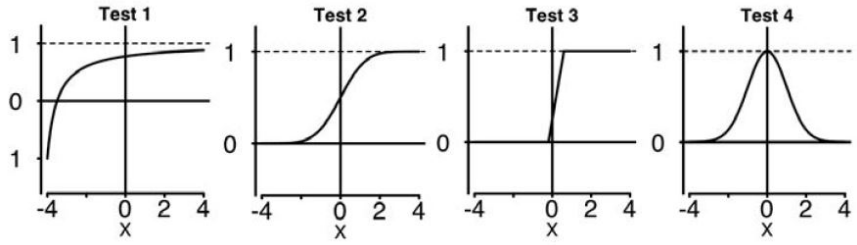
Recall that the *cumulative distribution function* (or cdf) of a r.v. X is defined as

$$F_X(x) = \mathbb{P}(X \leq x)$$

Recall also that we noted that every cdf function has the following properties:

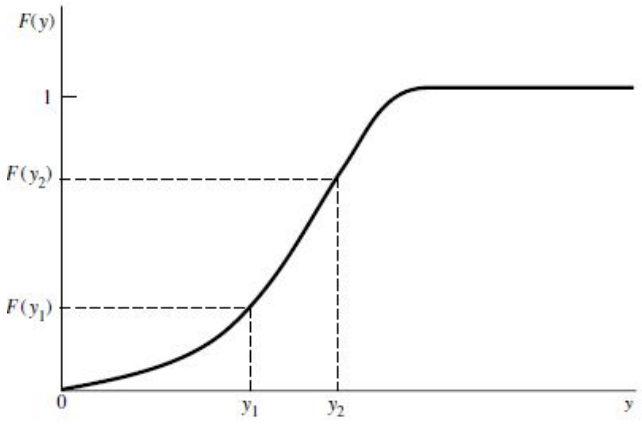
1. $\lim_{x\to-\infty} F_X(x) = 0$.
2. $\lim_{x\to\infty} F_X(x) = 1$.
3. $F_X(x)$ is a nondecreasing function of x.
4. $F_X(x)$ is a right-continuous function of x.

Exercise 3.1.1. Which of the following are graphs of valid cumulative distribution functions?



In contrast to the discrete random variables, the cdf of a continuous random variable X never jumps: there is no values x which have positive probability and which would cause a jump when we pass from $\mathbb{P}(X \leq x - \varepsilon$ to $\mathbb{P}(X \leq x)$. In other words, the cdf of a continuous random variable is continuous not only on the right but also on the left. In fact, we will use it as the definition of a continuous random variable.

Definition 3.1.2. A r.v. is said to be *continuous*, if its cumulative distribution function is continuous everywhere.



Example 3.1.3 (Uniform r.v.). Suppose that we know that there is exactly one message which is going to arrive in a one-second interval of time and that the exact time of its arrival is not known. Suppose also that every arrival time is equally probable.

Then the probability to arrive in a specific sub-interval of time is proportional to its length. In particular, the probability $P(X \leq x) = t$ if $0 \leq x \leq 1$.

So, the cumulative distribution function of a uniformly distributed random variable is

$$F_X(x) = \begin{cases} 0, & \text{if } x \leq 0, \\ x, & 0 < x \leq 1, \\ 1 & 1 < x. \end{cases}$$

It is a bit inconvenient that we do not have the probability mass function for continuous random variables, since all our formulas, – for expectation, for variance, for mgf, – were written in terms of the probability muss function.

In fact, there is an analogue of the probability mass function, which is called the probability density function. In order to define it, we assume that cdf $F_X(x)$ not only continuous but also differentiable everywhere, except possibly at a finite number of points in every finite interval. Then we have the following definition.

Definition 3.1.4. The probability density function (pdf) of a continuous r.v. X is defined as

$$f_X(x) = \frac{d}{dx} F_X(x).$$

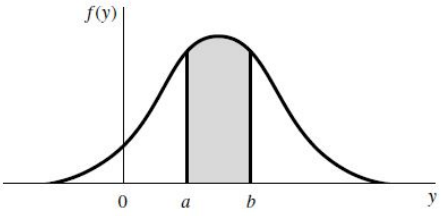
Example 3.1.5. For the uniform random variable that we considered in the previous example $F_X(x) = x$ for $x \in [0, 1]$ and constant (0 or 1), otherwise. Therefore,

$$f_X(x) = F_X'(x) = \begin{cases} 1, & \text{if } x \in [0, 1], \\ 0, & \text{otherwise.} \end{cases}$$

It is important to emphasize that the probability density $f_X(x)$ is not the probability that the random variable X takes value x. The probability density has the same relation to probability, as the mass density in physics has to mass. While it is not a probability, it is a very convenient function used in various calculation.

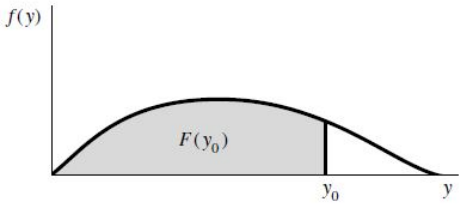
In particular, since the density is simply the derivative of the cdf, we can recover the cdf from the density by integrating:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt. \tag{3.1}$$



Compare this formula with the formula that we had for the discrete random variables:

$$F_X(x) = \sum_{t \leq x} p_X(t).$$



With a continuous r.v. X, we often want to find the probability that X falls in a specific interval $[a, b]$; that is, we want $\mathbb{P}(a \leq Y \leq b)$ for numbers $a < b$.

Theorem 3.1.6. *If X is a continuous r.v. with the cdf $F_X(x)$ and pdf $f_X(X)$ and $a < b$, then*

$$\begin{aligned} \mathbb{P}(a \leq X \leq b) &= F_X(b) - F_X(a) \\ &= \int_a^b f_X(x)\, dx. \end{aligned}$$

The integral over the density function represents the "area under the density curve" between $x = a$ and $x = b$:

Note that this integral formula is the analogue of the formula

$$\mathbb{P}(a \leq X \leq b) = \sum_{a \leq x \leq b} p_X(x),$$

for the discrete random variables, so the probability density function plays the role of the probability mass function for continuous random variables.

Proof.

$$\begin{aligned}\mathbb{P}(a \le X \le b) &= \mathbb{P}(X \le b) - \mathbb{P}(X < a) \\ &= F_X(b) - F_X(a) = \int_{-\infty}^{b} f_X(x)\,dx - \int_{-\infty}^{a} f_X(x)dx \\ &= \int_a^b f_X(x)\,dx.\end{aligned}$$

In the second line we used the fact that the cdf of a continuous random variable is continuous and therefore $\mathbb{P}(X < a) = \mathbb{P}(X \le a)$. In the third line we used the formula (3.1). □

Example 3.1.7. Let X be a continuous r.v. with pdf:

$$f(x) = \begin{cases} 0, & \text{for } y < 0, \\ 3x^2, & \text{for } 0 \le x \le 1, \\ 0, & \text{for } x > 1. \end{cases}$$

Find $\mathbb{P}(0.4 \le X \le 0.8)$.

$$\mathbb{P}(0.4 \le X \le 0.8) = \int_{0.4}^{0.8} 3x^2\,dx = x^3\Big|_{0.4}^{0.8} = 0.8^3 - 0.4^3 = 0.448$$

The pdf has the following properties:

1. $f_X(x) \ge 0$ for all x
   (because $F_X(x)$ is non-decreasing.)
2. $\int_{-\infty}^{\infty} f(t)dt = 1$.
   (because $F_X(\infty) = 1$.)

Example 3.1.8. Suppose X has range $[0, 2]$ and pdf $f(x) = cx$. What is the value of c? Compute $\mathbb{P}(1 \le X \le 2)$.

When we say that a random variable X has range $[a, b]$ this means that its pdf equals 0 outside of this interval. By the property of pdf, the integral of the pdf over the entire real line equals 1 and so its integral over the range

also equals 1. So,

$$\int_0^2 cx\,dx = 1,$$
$$c = \frac{1}{\int_0^2 x\,dx}$$

We calculate the integral as

$$\int_0^2 x\,dx = \frac{x^2}{2}\Big|_0^2 = 2.$$

Hence $c = \frac{1}{2}$. Then,

$$\mathbb{P}(1 \le X \le 2) = \int_1^2 \frac{x}{2} = \frac{x^2}{4}\Big|_1^2 = 1 - \frac{1}{4} = \frac{3}{4}.$$

Here is a couple of additional examples that show how to calculate the density from the cdf and other way around.

Example 3.1.9. Let X be a continuous r.v. with cdf:

$$F_X(x) = \begin{cases} 0, & \text{for } x < 0, \\ x^3, & \text{for } 0 \le x \le 1, \\ 1, & \text{for } x > 1. \end{cases}$$

Find the pdf $f_X(x)$ and graph both $F_X(x)$ and $f_X(x)$. Find $\mathbb{P}(X \le 0.5)$.

By differentiating, we find:

$$f_X(x) = \begin{cases} 3x^2, & \text{for } 0 \le x \le 1, \\ 0, & \text{otherwize.} \end{cases}$$

We can find $\mathbb{P}(X \le 0.5)$ in two ways. First of all, simply by definition $\mathbb{P}(X \le 0.5) = F_X(0.5) = 0.5^3 = 0.125$.

Second, if we did not know the cdf, we could recover it from pdf and write:

$$\mathbb{P}(X \le 0.5) = \int_0^{0.5} 3t^2\,dt = 0.5^3 = 0.125.$$

Note that the lower bound in the integration is set to 0 because for $x < 0$, $f_X(x) = 0$.

Example 3.1.10. Let X be a continuous r.v. with pdf

$$f_X(x) = \begin{cases} cx^{-1/2}, & \text{for } 0 < x \leq 4, \\ 0, & \text{elsewhere,} \end{cases}$$

for some constant c. Find c, find $F_X(x)$, graph $f_X(x)$ and $F_X(x)$, and find $\mathbb{P}(X < 1)$.

We should have

$$\int_0^4 cx^{-1/2}\,dx = 1,$$

so, by calculating the integral, we have

$$c2x^{1/2}\Big|_0^4 = 4c = 1,$$

and $c = 1/4$.

Then we calculate cdf:

$$\begin{aligned} F_X(x) &= \int_0^x \frac{1}{4}t^{-1/2}\,dt = \frac{1}{2}t^{1/2}\Big|_0^x \\ &= \frac{1}{2}x^{1/2}. \end{aligned}$$

More accurately,

$$F_X(x) = \begin{cases} 0, & \text{if } x \leq 0, \\ \frac{1}{2}x^{1/2}, & \text{if } 0 < x \leq 4, \\ 1, & \text{if } 4 < x. \end{cases}$$

For $\mathbb{P}(X < 1)$ we can used the cdf to calculate

$$\mathbb{P}(X < 1) = \frac{1}{2}1^{1/2} = \frac{1}{2}.$$

Exercise 3.1.11. Suppose X has range $[0, b]$ and pdf $f_X(x) = x^2/9$. What is b?

3.2 Expected value and variance

Definition 3.2.1. The *expected value* of a continuous r.v. X is

$$\mathbb{E}X = \int_{-\infty}^{\infty} x f_X(x)\, dx$$

provided that the integral exists.

Compare this with the definition of the expected value for the discrete random variable:

$$\mathbb{E}X = \sum_x x p_X(x).$$

We have a formula for the expected value of a function of a random variable, which we record here without proof.

Theorem 3.2.2. *If $g(X)$ is a function of a continuous r.v. X, then*

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(X) f_X(x)\, dx.$$

Again, this is useful to compare with the similar formula for the discrete random variables:

$$\mathbb{E}g(X) = \sum_x g(x) p_X(x).$$

Remark. While the integral above are for the whole real line, from $-\infty$ to ∞, typically the density $f_X(x)$ is zero outside some finite interval (a, b) which is called the support of the random variable X. In this cases it is enough to integrate over the support, from a to b. For example, in this case:

$$\mathbb{E}X = \int_a^b x f_X(x)\, dx$$

The expected value has the same properties as in the case of discrete random variables:

1. $\mathbb{E}[aX + b] = a\mathbb{E}X + b$.
2. $\mathbb{E}[X_1 + X_2 + \ldots + X_n] = \mathbb{E}X_1 + \mathbb{E}X_2 + \ldots + \mathbb{E}X_n$.

The variance and the standard deviation are defined in the same way as for the discrete r.v.:

$$\mathbb{V}\mathrm{ar}(X) = \mathbb{E}[(X-\mu)^2] = \mathbb{E}[X^2] - \mu^2,$$

where $\mu = \mathbb{E}X$.

And

$$Std(X) = \sqrt{\mathbb{V}\mathrm{ar}(Y)}.$$

Example 3.2.3. Let the pdf of a random variable X be

$$f(y) = \begin{cases} cx^{-1/2}, & \text{for } 0 < x \le 4, \\ 0, & \text{elsewhere,} \end{cases}$$

for some constant c. Find the expectation and variance of X.

We found that $c = 1/4$ in Example 3.1.10. So we write:

$$\begin{aligned} \mathbb{E}X &= \int_0^4 x \times \frac{1}{4} x^{-1/2}\, dx \\ &= \frac{1}{4}\int_0^4 x^{1/2}\, dx \\ &= \frac{1}{4} \times \frac{2}{3} x^{3/2}\Big|_0^4 \\ &= \frac{8}{6} = \frac{4}{3}. \end{aligned}$$

Similarly, we calculate

$$\begin{aligned} \mathbb{E}X^2 &= \int_0^4 x^2 \times \frac{1}{4} x^{-1/2}\, dx \\ &= \frac{1}{4}\int_0^4 x^{3/2}\, dx \\ &= \frac{1}{4} \times \frac{2}{5} x^{5/2}\Big|_0^4 \\ &= \frac{32}{10} = \frac{16}{5}. \end{aligned}$$

Finally,

$$\begin{aligned} &\mathbb{V}\mathrm{ar}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \frac{16}{5} - \left(\frac{4}{3}\right)^2 = \frac{64}{45} \approx 1.422 \\ &Std(X) = \sqrt{\mathbb{V}\mathrm{ar}(X)} \approx 1.193 \end{aligned}$$

3.3 Quantiles

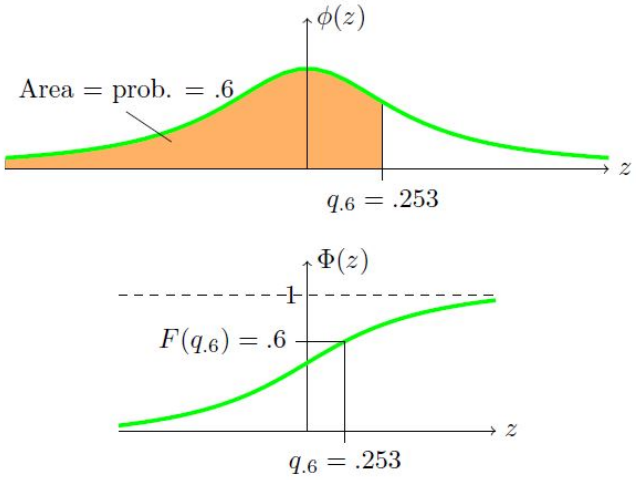
Quantiles can be defined both for discrete and continuous random variables.

Definition 3.3.1. Let $0 < \alpha < 1$. The α-quantile of a r.v. X, q_α, is the smallest value x such that $F_X(x) \geq \alpha$.

If X is a continuous r.v., q_α is the smallest value x such that $F_X(x) = \alpha$. (In case of discrete random variables, this definition can fail since for some α there is no value x that would satisfy this equality: the cdf can "jump over" this α.)

If, in addition, the cdf is strictly increasing, then the quantile function is simply the inverse of the cdf, that is, the α-quantile of X is the value x such that $F_X(x) = \alpha$.

Example: $q_{60\%} =?$. Let us see what it means graphically:



Definition 3.3.2. The 50% quantile of a random variable Y is called the median of Y.

Remark 1. The median of Y should be distinguished from the empirical median which is taught in elementary statistics courses and which is calculated when we have a bunch of realizations of Y.

Remark 2. Note that in Example 3.1.10 we found that the median of the random variable X in that example is 1 (since $\mathbb{P}(X < 1) = 0.5$) and

in Example 3.2.3 we calculated that $E(X) = 4/3$. This illustrates that the expected value and the median are often different.

In R, quantiles can be computed by functions like *qbinom* or *qpois*.

Example 3.3.3. What is the median of the binomial distribution with parameters $n = 15, p = 0.3$?

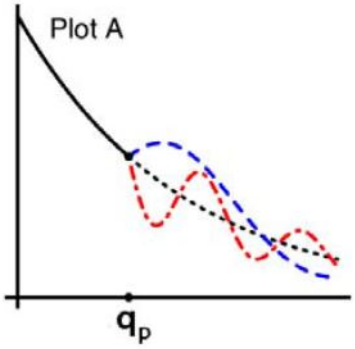
```
qbinom(0.5, size = 15, prob = 0.3) =  4.
```

Note that

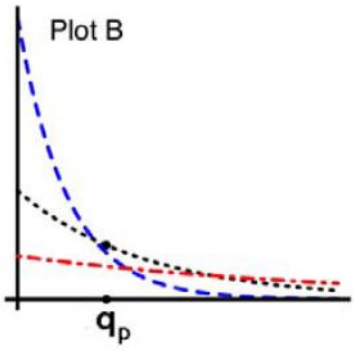
```
pbinom(4, size = 15, prob = 0.3) = 0.515
pbinom(3, size = 15, prob = 0.3) = 0.296
```

so that $x = 4$ is indeed the smallest value for which the binomial cdf is greater than 50%.

Exercise 3.3.4. In the plot below some densities are shown. The median of the black plot is always at q_p. Which density has the greatest median?



Exercise 3.3.5. In the plot below some densities are shown. The median of the black plot is always at q_p. Which density has the greatest median?



3.4 The Uniform random variable

Consider first a *discrete uniform* random variable on the interval $[a, b)$, where a and b are integers. This r.v. takes integer values $a, a+1, a+2, \ldots, b-1$ with equal probability. Its pmf is

$$p_X(x) = \frac{1}{b-a} \text{ if } a \le x < b,$$

and zero if x is either $< a$ or $\ge b$. Note that this pmf does not depend on x.

For example if $a = 2$ and $b = 5$, then X takes values $\{2, 3, 4\}$ with equal probability: $p_X(2) = p_X(3) = p_x(4) = 1/3$.

A *continuous uniform* random variable on the interval $[a, b)$ can take any value between a and b and its density is constant throughout this interval.

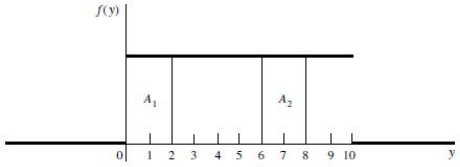
Definition 3.4.1 (Continuous uniform r.v.)**.** A r.v. X has a uniform probability distribution on the interval $[a, b]$, where $a < b$ are real numbers, if its pdf is

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a \le x \le b, \\ 0, & \text{otherwise.} \end{cases}$$

(If parameters $a = 0$ and $b = 1$, the distribution is called the standard uniform distribution.)

Example 3.4.2. Suppose a bus is supposed to arrive at a bus stop at 8 : 00 AM. In reality, it is equally likely to arrive at any time between 8 : 00 AM and 8 : 10 AM.

Let X be the amount of time between 8 : 00 AM and the moment when the bus arrives. This is a simple example of a continuous uniform random variable with parameters $a = 0$ and $b = 10$.



What is the probability the bus will arrive between 8 : 02 and 8 : 06 AM?

$$\mathbb{P}(2 \leq X \leq 6) = \int_2^6 \frac{1}{10}\, dx = \frac{6-2}{10} = 0.4.$$

Theorem 3.4.3. *If $X \sim \text{Unif}(a, b)$, then*

$$\mathbb{E}X = \frac{a+b}{2},$$
$$\mathbb{V}\text{ar}(Y) = \frac{(b-a)^2}{12}.$$

Proof.

$$\begin{aligned}
\mathbb{E}X &= \int_a^b x \frac{1}{b-a}\, dx = \frac{1}{b-a}\frac{b^2-a^2}{2} = \frac{a+b}{2}, \\
\mathbb{E}X^2 &= \int_a^b x^2 \frac{1}{b-a}\, dx = \frac{1}{b-a}\frac{b^3-a^3}{3} = \frac{a^2+ab+b^2}{3}, \\
\mathbb{V}\text{ar}(X) &= \mathbb{E}X^2 - (\mathbb{E}X)^2 = \frac{4(a^2+ab+b^2) - 3(a^2+2ab+b^2)}{12} \\
&= \frac{(b-a)^2}{12}
\end{aligned}$$

□

Example 3.4.4. Find the variance of $X \sim \text{Unif}(0, 4)$.

$$\mathbb{V}\text{ar}X = \frac{4^2}{12} = \frac{4}{3}$$

Example 3.4.5. Find the 70%-quantile of $X \sim \text{Unif}(2, 32)$.

The cdf of X is

$$F_X(x) = (x-2)/(32-2) \text{ if } x \in [2, 32],$$

so we need to solve the equation

$$\frac{x-2}{30} = 0.7,$$

which results in

$$x_{0.7} = 2 + 0.7 = 23.$$

The uniform distribution is the simplest possible continuous probability distribution. As it is simple, it is important. Now we are going to discuss two other most important distributions: the normal (or Gaussian) distribution and the exponential distribution. They are continuous analogues of the binomial and geometric distributions respectively.

3.5 The normal (or Gaussian) random variable

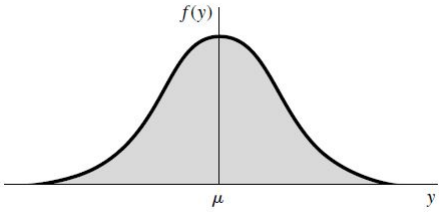
The Gaussian distribution is a continuous analogue of the binomial distribution. Suppose that we have a binomial experiment with large n. We assume that p is not small so that both the expected number of successes np and the standard deviation $\sqrt{np(1-p)}$ are large. In this case it is natural to look at probabilities of events

$$X \in [np - a\sqrt{np(1-p)}, np + b\sqrt{np(1-p)}],$$

This suggest defining a new random variable $Y = (X - np)/\sqrt{np(1-p)}$, so that these probabilities can be written as

$$\mathbb{P}(Y \in [-a, b]).$$

It turns out that for large n the cdf of this discrete random variable approaches the cdf of a continuous random variable which is called a *normal random variable* or a *Gaussian random variable*.



The distribution of this r.v., called the *normal distribution* or *Gaussian distribution*, is important because it is a good description of the distribution for many real-world random variables.

The fact that the most useful of all continuous distributions has a very unusual density is fascinating.

Definition 3.5.1. A r.v. X has the standard normal (standard Gaussian) distribution if its density function is

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

Let us check that this is a valid density function. We need to prove that the integral of this density equals to 1. Let us calculate the integral $I(t) = \int_{-\infty}^{\infty} e^{-tx^2} dx$.

$$\begin{aligned} I(t)^2 &= \int_{-\infty}^{\infty} e^{-tx^2}\, dx \int_{-\infty}^{\infty} e^{-ty^2}\, dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} dx\, dy\, e^{-t(x^2+y^2)} \\ &= \int_0^{2\pi} d\theta \int_0^{\infty} r e^{-tr^2} dr \\ &= -2\pi \frac{1}{2t} e^{-tr^2}\Big|_0^{\infty} = \frac{\pi}{t}. \end{aligned}$$

Hence,

$$\int_{-\infty}^{\infty} e^{-tx^2} dx = I(t) = \sqrt{\frac{\pi}{t}}. \tag{3.2}$$

In particular, if we use $t = \frac{1}{2}$, then we have

$$\int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2}\, dx = \sqrt{2\pi},$$

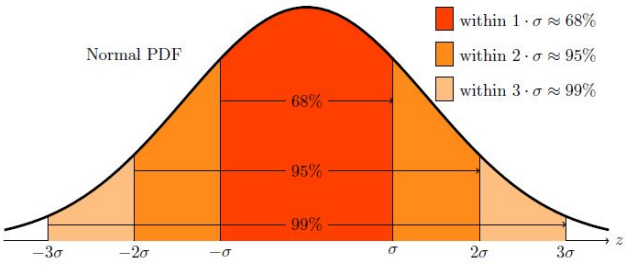
and so

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}\,dx = \frac{\sqrt{2\pi}}{\sqrt{2\pi}} = 1.$$

So the function $\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$ is indeed a valid probability density.

Note that the Gaussian pdf is defined over entire real line which means that this r.v. can take arbitrarily small or large values. Also note that this pdf is symmetric relative to the origin. In particular $\mathbb{P}(X < -x) = \mathbb{P}(X > x)$.

The behavior of the standard normal r.v. can be see from these examples
$\mathbb{P}(-1 \le Z \le 1) \approx .68$,
$\mathbb{P}(-2 \le Z \le 2) \approx .95$,
$\mathbb{P}(-3 \le Z \le 3) \approx .997$.



Theorem 3.5.2. *Let X has the standard normal distribution. Then $\mathbb{E}X = 0$ and $\mathbb{V}\mathrm{ar}(X) = 1$.*

Proof. We have:

$$\begin{aligned}\mathbb{E}X &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} xe^{-\frac{x^2}{2}}\,dx \\ &= \lim_{A\to\infty}\frac{1}{\sqrt{2\pi}}\int_{-A}^{A} xe^{-\frac{x^2}{2}}\,dx \\ &= -\lim_{A\to\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\Big|_{-A}^{A} = 0.\end{aligned}$$

For variance we differentiate equation (3.2) with respect to t and find that

$$\int_{-\infty}^{\infty} x^2 e^{-tx^2}dx = \frac{\sqrt{\pi}}{2}t^{-3/2}.$$

In particular, if we set $t = 1/2$, then

$$\mathbb{V}\text{ar}X = \mathbb{E}X^2 = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-x^2/2} dx = \frac{1}{2^{3/2}} 2^{3/2} = 1.$$

□

If we set $Y = \sigma X + \mu$, then as we will show later the density of the random variable Y is

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}.$$

This scaled random variable is also called a normal r.v., this time with parameters σ^2 and μ

Definition 3.5.3. A r.v. X has a normal distribution with parameters μ and σ^2, if its pdf is

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Notation: $X \sim \mathcal{N}(\mu, \sigma^2)$. The standard normal distribution has $\mu = 0$ and $\sigma = 1$.

Theorem 3.5.4. *Let Y be a normal random variable with parameters μ and σ^2. Then,*

$$\mathbb{E}Y = \mu, \text{ and}$$
$$\mathbb{V}\text{ar}Y = \sigma^2$$

Proof. We can write $Y = \mu + \sigma X$ where X is the standard normal r.v. Then by the properties of variance and expectation, and by results in Theorem 3.5.2, we have:

$$\mathbb{E}Y = \mu + \sigma \mathbb{E}X = \mu,$$
$$\mathbb{V}\text{ar}(Y) = \sigma^2 \mathbb{V}\text{ar}(X) = \sigma^2.$$

□

We have seen that if X is the standard normal random variable then $Y = \mu + \sigma X$ is the normal r.v. with parameters μ and σ^2. We have the converse result.

Theorem 3.5.5. *If $Y \sim N(\mu, \sigma^2)$, then the* standardized *version of Y,*

$$Z = \frac{Y - \mu}{\sigma},$$

has the standard normal distribution $N(0, 1)$.

We will prove this result later, when we study how to calculate the pdf of a function of a random variable.

How do we calculate the interval probabilities for the normal random variables?

If $X \sim N(\mu, \sigma^2)$, the cdf for X is

$$F_X(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} exp^{-\frac{(t-\mu)^2}{2\sigma^2}} dt.$$

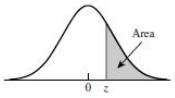
However, this integral cannot be expressed in elementary functions and has to be calculated via numerical integration. In particular, we are forced to use software or tables to calculate such normal probabilities.

In R, we can use function "pnorm", which takes parameters μ and σ as "mean" and "sd":

```
pnorm(x, mean = ..., sd = ...)
```

If the parameters is omitted, then the function will calculate the cdf of the standard normal variable: $pnorm(x)$ is the same as pnorm(x, mean = 0, sd = 1).

Table 4 Normal Curve Areas
Standard normal probability in right-hand tail
(for negative values of z, areas are found by symmetry)



	Second decimal place of z									
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0722	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0352	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019

Alternatively, we can use "standartization" and look-up necessary values in a table. (It is still a valuable skill at some professional exams.)

Example 3.5.6. A graduating class has GPAs that follow a normal distribution with mean 2.70 and variance 0.16.

1. What is the probability a randomly chosen student has a GPA greater than 2.50?
2. What is the probability that this random GPA is between 3.00 and 3.50?
3. Exactly 5% of students have GPA above what number?

Let Y be the GPA of a randomly chosen student. Then $Y \sim N(2.70, 0.16)$.

$$\begin{aligned}\mathbb{P}(Y \geq 2.50) &= 1 - \mathbb{P}(Y \leq 2.50) \\ &= 1 - pnorm(2.50, mean = 2.70, sd = sqrt(0.16)) \\ &= 0.6914625\end{aligned}$$

Alternatively, we can standardize the random variable Y and use tables:

$$\begin{aligned}\mathbb{P}(Y \geq 2.50) &= \mathbb{P}\Big(\frac{Y - 2.70}{\sqrt{0.16}} \geq \frac{2.50 - 2.70}{\sqrt{0.16}}\Big) \\ &= \mathbb{P}(Z \geq -0.5) \\ &= \mathbb{P}(Z \leq 0.5) = 1 - \mathbb{P}(Z \geq 0.5) \\ &= 1 - 0.3085 = 0.6915\end{aligned}$$

In the third line we used the symmetry of the distribution and wrote the necessary probability as $1-\mathbb{P}(Z \geq 0.5)$ since the table gives the probabilities $\mathbb{P}(Z \geq a)$

The probability that $3.00 \leq Y \leq 3.50$ can be computed similarly.

For the third question, we need to find y such that $\mathbb{P}(Y \geq y) = 5\%$, which is the same as $\mathbb{P}(Y \leq y) = 95\%$. This means that we want to calculate the 95% quantile of this distribution. In R, this is simply

```
qnorm(0.95, mean = 2.70, sd = sqrt(0.16))
= 3.357941
```

In order to use tables, we again use standartization:

$$\mathbb{P}(Y \leq y) = \mathbb{P}\Big(Z \leq \frac{y - 2.70}{\sqrt{0.16}}\Big)$$

We want to make this probability equal 95%. For the standard normal distribution, we can easily find the 95% quantile from the table: $q_{0.95} = 1.645$, so

$$\begin{aligned}\frac{y - 2.70}{\sqrt{0.16}} &= 1.645, \\ y = 2.70 + 1.645 \times \sqrt{0.16} &= 3.358\end{aligned}$$

Exercise 3.5.7. Let Y be a normal r.v. with mean μ and standard deviation $\sigma = 10$. Find μ such that $\mathbb{P}(Y \leq 10) = 0.75$.

Now, what about the moment generating function of a normal distribution?

Theorem 3.5.8. *If $X \sim N(0,1)$, then its moment-generating function is*

$$m_X(t) = e^{t^2/2}.$$

Proof. We calculate:

$$\begin{aligned} m_X(t) = \mathbb{E}(e^{tX}) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} e^{-\frac{x^2}{2}}\, dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(x-t)^2 - t^2}{2}}\, dx \\ &= e^{t^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(x-t)^2}{2}}\, dx \\ &= e^{t^2/2}, \end{aligned}$$

where the last integral (multiplied by the factor of $\frac{1}{\sqrt{2\pi}}$) is calculated as equal to 1 by the change of variable $y = x - t$.

□

By a similar calculation, we can get a more general result that the moment generating function of $Y \sim N(\mu, \sigma^2)$ is $m_Y(t) = e^{\mu t + \frac{\sigma^2}{2} t^2}$.

3.6 Exponential and Gamma distributions

3.6.1 Exponential

The exponential distribution is a continuous analogue of the geometric distribution. Recall that the geometric random variable is the number of trials in binomial experiment until the first success. Similarly, one can think about the exponential random variable as the amount the time until some event happens.

For example, exponential distribution can be used to model such situations as

- Lifelengths of manufactured parts
- Lengths of time between arrivals at a restorant
- Survival times for severely ill patients

Definition 3.6.1. A r.v. X has an exponential distribution with parameter $\beta > 0$ if its pdf is

$$f_X(x) = \frac{1}{\beta} e^{-x/\beta}$$

for $x \geq 0$.

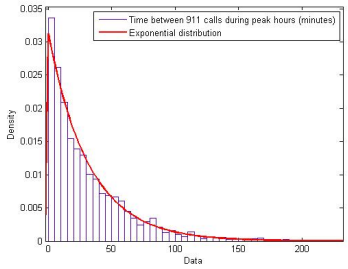


Figure 3.1: Histogram of the time intervals between 911 calls.

Warning: often the exponential distribution is used with a different choice of the parameter, namely $\lambda = 1/\beta$. As we will see in the next theorem, the meaning of β is the expected value of the distribution, or its *mean*. The parameter λ is useful since exponential random variables are often used for times between random events. If the events happen at the *rate* of λ per unit of time then the time intervals are distributed as exponential r.v. with mean $\beta = 1/\lambda$.

Theorem 3.6.2. *If X is an exponential r.v. with parameter β, then*

$$\mathbb{E}X = \beta,$$
$$\mathbb{V}\mathrm{ar}(Y) = \beta^2.$$

The mgf of X is

$$m_X(t) = \frac{1}{1 - \beta t}$$

for $t < \beta$.

Proof. Let us calculate the mgf first.

$$\begin{aligned} m_X(t) = \mathbb{E}e^{tX} &= \frac{1}{\beta} \int_0^\infty e^{tx} e^{-x/\beta}\, dx \\ &= \frac{1}{\beta} \int_0^\infty \beta e^{(\beta t - 1)y}\, dy, \end{aligned}$$

after we use the change of variables $x = \beta y$. The integral is

$$\begin{aligned}\int_0^\infty e^{(\beta t-1)y}\,dy &= -\frac{1}{1-\beta t}e^{(\beta t-1)y}\Big|_0^\infty \\ &= \frac{1}{1-\beta t}.\end{aligned}$$

Now, when we know the mgf, we can calculate the expectation and variance:

$$\begin{aligned}\mathbb{E}X &= m_X'(t)\Big|_{t=0} = \beta\frac{1}{(1-\beta t)^2}\Big|_{t=0} = \beta. \\ \mathbb{E}X^2 &= m_X''(t)\Big|_{t=0} = 2\beta^2\frac{1}{(1-\beta t)^3}\Big|_{t=0} = 2\beta^2. \\ \mathbb{V}\mathrm{ar}(X) &= \mathbb{E}X^2 - (\mathbb{E}X)^2 = 2\beta^2 - \beta^2 = \beta^2.\end{aligned}$$

□

For calculations, a very useful property of the exponential distribution is that it has a very simple survival function, which is by definition $S_X(x) := \mathbb{P}(X > x)$. That is, $S_X(x) = 1 - F_X(x)$, where $F_X(x)$ is the cdf of X.

For the exponential distribution we have:

$$S_X(x) = \frac{1}{\beta}\int_x^\infty e^{-t/\beta}\,dt = e^{-t/\beta}\Big|_x^\infty = e^{-x/\beta}.$$

Example 3.6.3. Let the lifetime of a part (in thousands of hours) follow an exponential distribution with mean lifetime 2000 hours.

Find the probability the part lasts between 2000 and 2500 hours.

Here $\beta = 2000$ and we calculate:

$$\begin{aligned}\mathbb{P}(2000 < X < 2500) &= \mathbb{P}(X > 2000) - \mathbb{P}(X > 2500) \\ &= S_X(2000) - S_X(2500) \\ &= e^{-2000/2000} - e^{-2500/2000} \\ &= e^{-1} - e^{-1.25} = 0.08137464\end{aligned}$$

The same answer could be obtained by using R:

```
pexp(2500, rate = 1/2000) - pexp(2000, rate = 1/2000)
```

Note that we used here rate $\lambda = 1/\beta$.

Example 3.6.4 (Memoryless property of the exponential distribution). The exponential distribution is very convenient to model lifetime of an equipment. However, it has some properties that make it not very realistic for this purpose. Intuitively, we believe that if an equipment served for certain amount of time than its future lifetime becomes somewhat shorter. However, with exponential distribution, this intuition does not work.

Let a lifetime X follow an exponential distribution. Suppose the part has lasted a units of time already. Then the conditional probability of it lasting at least b *additional* units of time is $\mathbb{P}(X > a + b | X > a)$, and it turns out that it is the same as the probability of lasting at least b units of time when the piece of equipment was new: $\mathbb{P}(X > b)$. This is called the memoryless property of the exponential distribution.

Proof. We have

$$\begin{aligned}
\mathbb{P}(X > a + b | X > a) &= \frac{\mathbb{P}(X > a + b, \text{ and } X > a)}{\mathbb{P}(X > a)} \\
&= \frac{\mathbb{P}(X > a + b)}{\mathbb{P}(X > a)} \\
&= \frac{e^{-(a+b)/\beta}}{e^{-a/\beta}} = e^{-b/\beta} \\
&= \mathbb{P}(X > b).
\end{aligned}$$

□

The same property holds also for the geometric distribution, and these are the only probability distributions with the memoryless property.

3.6.2 Gamma

The Gamma distribution plays approximately the same role with respect to the exponential distribution as the negative binomial with respect to the geometric distribution. It is also used to model the time until some event happens. For example, the Gamma distributed random variable with

parameter $\alpha = n$ can be used to model the time until the n-th breakdown of a device.

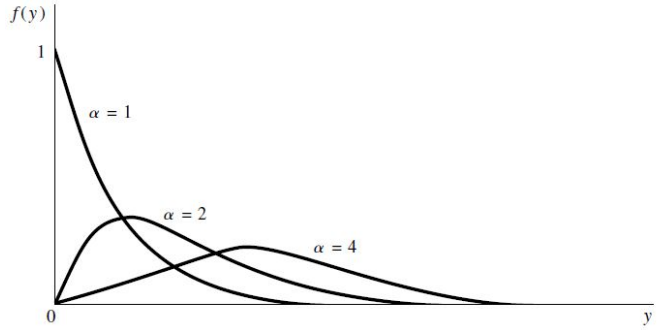
Definition 3.6.5. A continuous r.v. X has a Gamma distribution if its pdf is

$$f_X(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta},$$

where $\alpha > 0$, $\beta > 0$, and $x \in [0, \infty)$. Here $\Gamma(\alpha)$ is defined as

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx.$$

Note that the definition of $\Gamma(\alpha)$ ensures that $f_X(x)$ is a valid density. A useful property of the Gamma function is that $\Gamma(n) = (n-1)!$ for positive integers n. This can be proved from the definition of the Gamma function by repeated integration by parts.



The parameter α and β are called the *shape* and scale parameters, respectively. The choice of β is equivalent to the choice of units. The picture shows how the shape of the distribution depends on the parameter α. Note that the exponential distribution corresponds to the parameter $\alpha = 1$.

Theorem 3.6.6. *If X has a gamma distribution with parameters α and β, then*

$$\mathbb{E}X = \alpha\beta,$$
$$\mathbb{V}\mathrm{ar}(X) = \alpha\beta^2.$$

and the mgf of X is

$$m_X(t) = \frac{1}{(1-\beta t)^\alpha}.$$

Proof. Let us just prove the formula for the mgf. The expectation and variance can be easily obtained by differentiation, as we did it for the exponential r.v.

$$\begin{aligned} m_X(t) = \mathbb{E}(e^{tX}) &= \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^\infty e^{tx} x^{\alpha-1} e^{-x/\beta}\, dx \\ &= \frac{1}{\Gamma(\alpha)} \int_0^\infty y^{\alpha-1} e^{-(1-\beta t)y}\, dy \\ &= \frac{1}{(1-\beta t)^\alpha} \frac{1}{\Gamma(\alpha)} \int_0^\infty z^{\alpha-1} e^{-z}\, dz \\ &= \frac{1}{(1-\beta t)^\alpha}, \end{aligned}$$

where we used the change of variable $x = \beta y$ for the second equality, the change of variable $y = z/(1-\beta t)$ for the third equality, and used the definition of the Gamma function for the last equality. □

Example 3.6.7. Let $f_X(x) = cx^4 e^{-x/3}$ if $x \geq 0$ and zero elsewhere.

What value of c makes $f_X(x)$ a valid density? What is $\mathbb{E}X$? What is $\mathbb{V}\mathrm{ar}(X)$

We recognize the distribution as the Gamma distribution with parameters $\alpha = 5$ and $\beta = 3$. Hence $c = 1/(3^5\Gamma(5)) = 1/(3^5 \times 4!)$. Then, $\mathbb{E}X = 5 \times 3 = 15$ and $\mathbb{V}\mathrm{ar}(X) = 5 \times 3^2 = 45$.

Example 3.6.8 (More on Chebyshev's Inequality). Suppose $X \sim \Gamma(5, 3)$. By using Chebyshev's inequality, find the interval such that the probability that X is inside this interval is 75%. Find the actual probability that X is inside this interval.

Chebyshev's inequality says that

$$\mathbb{P}(|X - \mu| > \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}.$$

In our example we want to ensure that the probability to be inside the interval is at least 75%. This means that we want to make sure that the probability to be outside the interval is at most 25%.

The variance $\sigma^2 = 5 \times 3^2 = 45$ by the formula for the Gamma distribution with parameters $\alpha = 5$ and $\beta = 3$, so from Chebyshev's inequality, we get the following equation:

$$\frac{45}{\varepsilon^2} = 0.25 \text{ or } \varepsilon = \sqrt{\frac{45}{0.25}} = 13.4164.$$

Since $\mu = 5 \times 3 = 15$ we find that

$$\mathbb{P}(|X - 15| \leq 13.4164) \geq 0.75,$$

or that the required interval is $(15-13.4164, 15+13.4164) = (1.5836, 28.4164)$.

(It is perhaps worthwhile to note that if we got a negative number for the lower limit, e.g. - 1.5836, then we could use our knowledge that Gamma is always non-negative and could use 0 as a lower end-point of the interval.)

In order to calculate the actual probability for X to be in this interval, $\mathbb{P}\big(X \in (1.5836, 28.4164)\big)$, we use R:

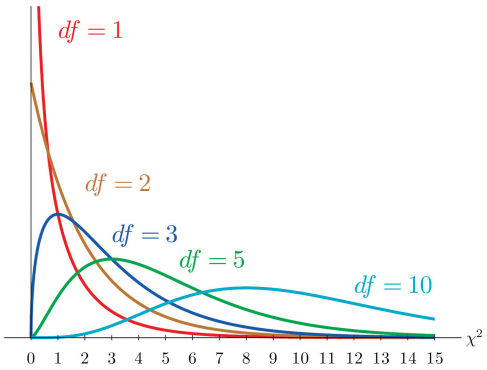
```
pgamma(28.4164, shape = 5, scale = 3)
                        - pgamma(1.5836, shape = 5, scale = 3)
```

which gives the probability 0.9588031.

An important special case of the Gamma distribution is the χ^2 distribution.

Definition 3.6.9. For any integer $\nu \geq 1$, a r.v. X has a chi-square (χ^2) distribution with ν *degrees of freedom* [$X \sim \chi^2(\nu)$] if X is a Gamma r.v. with

$$\alpha = \nu/2,$$
$$\beta = 2.$$



The chi-square distribution is important in statistics because of the following result.

Theorem 3.6.10. *If X_1, X_2, …X_ν are independent standard normal r.v.'s, then $Y = (X_1)^2 + (X_2)^2 + \ldots (X_\nu)^2$ has the χ^2 distribution with ν degrees of freedom.*

We will prove this result later.

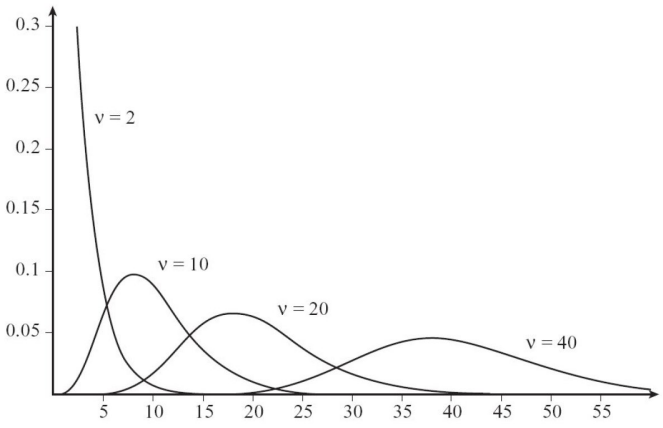
A direct consequence of the results for the Gamma distribution is the following theorem.

Theorem 3.6.11. *If $X \sim \chi^2(\nu)$, then*

$$\mathbb{E}X = \nu,$$
$$\mathbb{V}\text{ar}X = 2\nu,$$

The mgf of X is

$$m_X(t) = \frac{1}{(1-2t)^{\nu/2}}.$$



3.7 Beta distribution

The beta distribution does not have a discrete analogue. It is a generalization of the uniform distribution meant to give more possibilities to model a random variable that takes its value in a finite interval.

Definition 3.7.1. A r.v. X has a beta distribution with parameters $\alpha > 0$ and $\beta > 0$ [$X \sim Beta(\alpha, \beta)$] if its pdf is:

$$f_X(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1},$$

where

$$B(\alpha, \beta) = \int_0^1 y^{\alpha-1}(1-y)^{\beta-1} dy.$$

It can be proved that

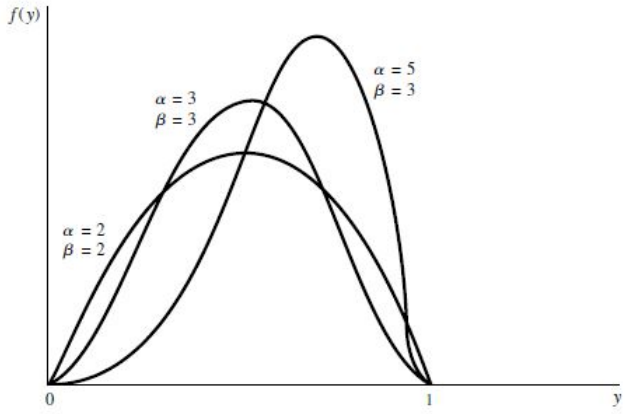
$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

Note 1: The beta pdf has support on $[0, 1]$.

Note 2: For integer α and β,

$$\frac{1}{B(\alpha, \beta)} = \frac{(\alpha+\beta-1)!}{(\alpha-1)!(\beta-1)!}.$$

Note 3: The uniform distribution is the Beta distribution with parameters $\alpha = 1$ and $\beta = 1$.



As the figure shows, the Beta distribution is quite flexible. Note also that α and β grow, the distribution becomes more and more concentrated around its expected value.

Beta-distributed random variables are used to model random quantity that takes values between 0 and 1. More generally, if a random quantity Y has its support on interval $[c, d]$ then we can model the standardized quantity $Y^* = \frac{Y-c}{d-c}$ as a random variable with beta distribution.

Examples:

- The proportion of a chemical product that is pure.
- The proportion of a hospital's patients infected with a certain virus.
- The parameter p of a binomial distribution.

Theorem 3.7.2. *If* $X \sim Beta(\alpha, \beta)$, *then*

$$\mathbb{E}(X) = \frac{\alpha}{\alpha + \beta},$$
$$\mathbb{V}\text{ar}(X) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

For the proof, see the textbook.

Example 3.7.3. Define the downtime rate as the proportion of time a machine is under repair. Suppose a factory produces machines whose downtime rate follows a $Beta(3, 18)$ distribution.

For a randomly selected machine, what is the expected downtime rate? What is the probability that a randomly selected machine is under repair less than 5% of the time? If machines act independently, in a shipment of 25 machines, what is the probability that at least 3 have downtime rates greater than 0.20?

For the first question, we have by the formula in the theorem:

$$\mathbb{E}(X) = \frac{3}{3+18} = \frac{1}{7} \approx 14.28\%$$

For the second one:

$$\mathbb{P}(X < 0.05) = pbeta(0.05, alpha = 3, beta = 18) = 0.07548367$$

The third sub-question needs to steps and two distributions: Beta and Binomial. First, we calculate the probability that a machine have downtime ratio greater than 20%

$$\begin{aligned}\mathbb{P}(X > 0.20) &= 1 - \mathbb{P}(X \leq 0.20) \\ &= 1 - pbeta(0.20, alpha = 3, beta = 18) = 0.2060847\end{aligned}$$

In the second step, we note that in a sample of 25 machines, the number of machines that have the downtime greater than 0.20 is a binomial r.v. Y with parameters $n = 25$ and $p = 0.2060847$. So, we have

$$\begin{aligned}\mathbb{P}(Y \geq 3) &= 1 - \mathbb{P}(Y \leq 2) \\ &= 1 - pbinom(2, size = 25, prob = 0.2060847) \\ &\approx 91.35\%\end{aligned}$$

3.8 Other distribution

The uniform, normal, and exponential probability distributions are fundamental for probability theory. The Beta and Gamma distributions are generalizations of the uniform and the exponential distributions, which are useful for modeling purposes.

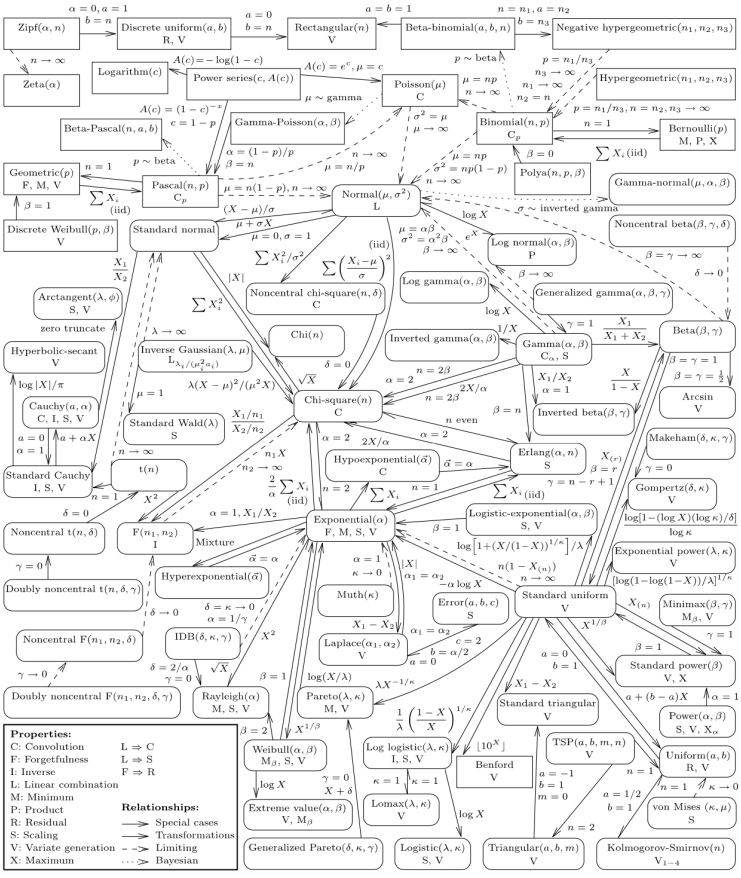


Figure 3.2: From "Univariate Distribution Relationships" by Lawrence M. Leemis and Jacquelyn T. McQueston.

There are some other important distributions, such as Cauchy, Weibull, and logistic distributions which are also commonly used for modeling various random phenomena.

A documentation for these distributions can be found in most statistical software packages.

Chapter 4

Multivariate Distributions

In many applications we measure not one but several random variables in an experiment.

Examples:

- A random household is sampled: X = income and Y = the number of children.
- A random animal: X = height and Y = weight.
- A random school-child: the IQ and birth weight.
- A random adult human: the frequency of exercise and the rate of heart disease
- A random city: the level of air pollution and rate of respiratory illness.
- A random Titanic's passenger: the gender and survival indicators.
- A Geiger camera records: the random variables X_1, X_2, X_3, ...are times of the particle detections.

Random variables $(X_1(\omega), \ldots, X_n(\omega))$ measured in an experiment with a random outcome ω form a random vector. How can we describe the properties of this vector?

For convenience, we will focus on vectors with two components although everything can be generalized to vectors with a larger (but still finite) number of components.

We will also first discuss a simpler case when the vectors can take only finite or countable number of values, so that everything can be described in terms of probabilities of individual values.

4.1 Discrete random variables

4.1.1 Joint and marginal pmf

Consider a *bivariate* random vector (X, Y).

Definition 4.1.1. Let X and Y are two discrete random variables. Their *joint probability mass function* ("joint pmf") is defined as

$$p_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y).$$

We will sometime omit the subscript and write simply $p(x, y)$ instead of $p_{X,Y}(x, y)$ to lighten the notation.

Definition 4.1.2. Let X and Y are two discrete random variables. The *marginal probability mass functions* of X and Y are the probability mass functions of X and Y, respectively, seen as single random variables. That is,

$$p_X(x) = \mathbb{P}(X = x),$$
$$p_Y(y) = \mathbb{P}(Y = y).$$

The reason for why we use a new name for the familiar concept, that is, for why we add the adjective "marginal", is that we sometimes have a joint pmf for X and Y and want to calculate the pmf for one of these variables.

For this purpose, we can use the following formula:

$$p_X(x) = \sum_{y : p_{X,Y}(x,y) > 0} p_{X,Y}(x, y). \tag{4.1}$$

Here the summation is over all y such that the joint pmf $p_{X,Y}(x, y)$ is positive for a given x.

Similarly,

$$p_Y(y) = \sum_x p_{X,Y}(x, y).$$

Example 4.1.3. Roll two dice: $X = \#$ on the first die, $Y = \#$ on the second die.

The joint pmf of X and Y can be represented with a two-way table

$X \backslash Y$	1	2	3	4	5	6
1	1/36	1/36	1/36	1/36	1/36	1/36
2	1/36	1/36	1/36	1/36	1/36	1/36
3	1/36	1/36	1/36	1/36	1/36	1/36
4	1/36	1/36	1/36	1/36	1/36	1/36
5	1/36	1/36	1/36	1/36	1/36	1/36
6	1/36	1/36	1/36	1/36	1/36	1/36

$p_{X,Y}(x, y) = \frac{1}{36}$ for all integer x and y between 1 and 6. By summing the values in the table over rows and columns, respectively, we find that $p_X(x) = 1/6$ for all integer x between 1 and 6, and $p_Y(y) = 1/6$ for all y between 1 and 6.

Example 4.1.4. Roll two dice: $X = \#$ on the first die, $T =$ total on both dice.

The joint pmf of X and T:

$X \backslash T$	2	3	4	5	6	7	8	9	10	11	12
1	1/36	1/36	1/36	1/36	1/36	1/36	0	0	0	0	0
2	0	1/36	1/36	1/36	1/36	1/36	1/36	0	0	0	0
3	0	0	1/36	1/36	1/36	1/36	1/36	1/36	0	0	0
4	0	0	0	1/36	1/36	1/36	1/36	1/36	1/36	0	0
5	0	0	0	0	1/36	1/36	1/36	1/36	1/36	1/36	0
6	0	0	0	0	0	1/36	1/36	1/36	1/36	1/36	1/36

Now by summing over rows we find the same marginal pmf for the random variable X, which is not surprising since this is the same random variable as in the previous example. For r.v. T, we sum the values in the table

over columns and find the marginal pmf $p_T(t)$ summarized in the table below.

t	2	3	4	5	6	7	8	9	10	11	12
$p_T(t)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Example 4.1.5. Researches at three universities (in New York, California, and Florida) are applying for two separate grants. Suppose all proposals are equally good, and so which university get the contracts can be seen as a random selection.

Let X = number of grants awarded to New York and Y= number of grants awarded to California.

What is the joint pmf of X and Y? What are the marginal pmf's?

Let us list all possible outcomes in this experiment. Let us use letters N, C, and F to denote New York, California and Florida respectively, and let (NC) denote the outcome that the first grant went to New York, and the second went to California, with all other outcomes denoted similarly. Then the list of possible outcomes consists of 9 elements:

$$NN, NC, NF, CN, CC, CF, FN, FC, FF,$$

and by assumption they are all equally probable.

So we get the joint pmf for X and Y summarized in the following table:

$x\backslash y$	0	1	2
0	1/9	2/9	1/9
1	2/9	2/9	0
2	1/9	0	0

For, example the probability $p_{X,Y}(0,1) = 2/9$ corresponds to the outcomes FC and CF in which New York gets zero grants and California gets 1 grant. By summing values in each row we get the marginal pmf for r.v. X, namely, $p_X(0) = 4/9$, $p_X(1) = 4/9$ and $p_X(2) = 1/9$. The random variable Y has the same pmf.

Example 4.1.6. For the situation described in the previous example, calculate the probability that New York obtains at least as many grants as California.

This is the sum of values in the table on or under the main diagonal:

$$\begin{aligned}\mathbb{P}(X \geq Y) &= p(0,0) + p(1,0) + p(1,1) + p(2,0) \\ &= \frac{1}{9} + \frac{2}{9} + \frac{2}{9} + \frac{1}{9} = \frac{6}{9} = \frac{2}{3}.\end{aligned}$$

Example 4.1.7. Roll two dice: $X = \#$ on the first die, $Y = \#$ on the second die.

What is the probability of the event that $Y - X \geq 2$?

We can visualize the event by shading the cells.

$X \backslash Y$	1	2	3	4	5	6
1	1/36	1/36	1/36	1/36	1/36	1/36
2	1/36	1/36	1/36	1/36	1/36	1/36
3	1/36	1/36	1/36	1/36	1/36	1/36
4	1/36	1/36	1/36	1/36	1/36	1/36
5	1/36	1/36	1/36	1/36	1/36	1/36
6	1/36	1/36	1/36	1/36	1/36	1/36

$\mathbb{P}(Y - X \geq 2) =$ sum of the numbers in the shaded cells $= \frac{10}{36}$

4.1.2 Conditional pmf and conditional expectation

Conditional pmf

Intuitively, the conditional pmf of Y given $X = x$ describes the probability distribution of the r.v. Y after we have learned that $X = x$.

Definition 4.1.8. Let X and Y be two discrete random variables. The conditional probability mass function is defined as

$$p_{Y|X}(y|x) = \mathbb{P}(Y = y | X = x)$$

We can calculate the conditional pmf in terms of the joint and marginal pmf's:

$$\begin{aligned} p_{Y|X}(y|x) &:= \frac{\mathbb{P}(Y=y, X=x)}{\mathbb{P}(X=x)} \\ &= \frac{p_{X,Y}(x,y)}{p_X(x)} \\ &= \frac{\text{joint pmf of } X \text{ and } Y \text{ at } (x,y)}{\text{marginal pmf of } X \text{ at } x}. \end{aligned}$$

Similarly, we can define the conditional pmf of X given Y:

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y).}$$

Example 4.1.9. Again, let us use the setup in Example 4.1.5.

Find the conditional pmf of X (number of grants awarded to New York) given $Y = 0$ and find $\mathbb{P}(X \geq 1|Y = 0)$.

Note that $p_Y(0) = 4/9$ so we get by the formula and the values for the joint pmf that we derived in Example 4.1.5:

$$\begin{aligned} p_{X|Y}(0|0) &= \frac{1/9}{4/9} = \frac{1}{4} \\ p_{X|Y}(1|0) &= \frac{2/9}{4/9} = \frac{1}{2} \\ p_{X|Y}(2|0) &= \frac{1/9}{4/9} = \frac{1}{4} \end{aligned}$$

Conditional expectation

Definition 4.1.10. Let X and Y be two discrete random variables. The *conditional expectation* of Y given $X = x$ is defined as the expectation of Y with respect to the conditional pmf of Y:

$$\mathbb{E}(Y|X = x) := \sum_y y p_{Y|X}(y|x).$$

Intuitively, the conditional expectation $\mathbb{E}(Y|X = x)$ is the average of the random variable Y after we learned that $X = x$. It is often used as a

predictor of random variable Y based on the information that we learned from X.

It is sometimes useful that by using the definition of the conditional pmf we can write the conditional expectation slightly differently, by using only the joint and marginal pmf:

$$\mathbb{E}(Y|X=x) := \frac{\sum_y y p_{X,Y}(x,y)}{p_X(x)}. \tag{4.2}$$

Note that for different realizations of X the value of the conditional expectation of Y may be different: $\mathbb{E}(Y|X=x)$ in general depends on x. In other words, we can think about the conditional expectation of Y as a random variable which is a function of the random variable X, $g(X)$. Every realization x of r.v. X is mapped to the number $E(Y|X=x)$. It is customary to write this function as $\mathbb{E}(Y|X)$. Note that despite appearance, this is a function of X only.

Similarly, the conditional expectation of X given Y, $\mathbb{E}(X|Y)$ is a function of Y.

Example 4.1.11. In the setting of Example 4.1.5, calculate the conditional expectation of X given different realizations of Y. Compare it with the unconditional expectation of X.

First, let us calculate the unconditional expectation. This is simply the expectation with respect to the marginal pmf of X. The marginal pmf was computed in Example 4.1.5 and we calculate:

$$\mathbb{E}X = \sum_x x p_X(x) = 0 \times \frac{4}{9} + 1 \times \frac{4}{9} + 2 \times \frac{1}{9} = \frac{2}{3}$$

By using the conditional pmf of X given $Y = 0$ from the previous example we can calculate by definition:

$$\mathbb{E}(X|Y=0) = 0 \times \frac{1}{4} + 1 \times \frac{1}{2} + 2 \times \frac{1}{4} = 1.$$

Note that given the information that California gets 0 grants, the conditional expectation of X is larger than the unconditional expectation.

For other values of Y we can either proceed similarly by calculating the conditional pmf first, or alternatively use the formula that gives the conditional expectation in terms of the joint and marginal pmf's:

$$\mathbb{E}(X|Y=1) = \frac{\sum_x x p_{X,Y}(x,1)}{p_Y(1)} = \frac{0 \times \frac{2}{9} + 1 \times \frac{2}{9}}{4/9} = \frac{1}{2}$$

In this case, the information that California gets 1 grant made the conditional expectation of X smaller than the unconditional expectation. The situation is even worse in the last remaining case.

$$\mathbb{E}(X|Y=2) = \frac{\sum_x x p_{X,Y}(x,2)}{p_Y(2)} = \frac{0 \times \frac{1}{9}}{1/9} = 0.$$

Now note that since the conditional expectation of X given Y is a function of Y, we can calculate the expectation of the conditional expectation (!) by using the pmf of Y. In this example we have:

$$\begin{aligned}\mathbb{E}\Big[\mathbb{E}(X|Y)\Big] &= \mathbb{E}(X|Y=0)p_Y(0) + \mathbb{E}(X|Y=1)p_Y(1) + \mathbb{E}(X|Y=2)p_Y(2) \\ &= 1 \times \frac{4}{9} + \frac{1}{2} \times \frac{4}{9} + 0 \times \frac{1}{9} = \frac{6}{9} = \frac{2}{3}.\end{aligned}$$

Note that this quantity equals the unconditional expectation $\mathbb{E}X$. This is not coincidence but a consequence of a general result.

Theorem 4.1.12 (Law of Iterated Expectation)**.** *Let X and Y be two discrete random variables such that $\mathbb{E}X < \infty$ and $\mathbb{E}(X|Y=y) < \infty$ for all y in the range of Y. Then,*

$$\mathbb{E}\Big[\mathbb{E}(X|Y)\Big] = \mathbb{E}X.$$

The result holds not only for discrete random variables but in a more general situation but we will prove it only for the discrete r.v.'s

Proof. First, by the formula for the expected value of a function of Y, we have:

$$\mathbb{E}\Big[\mathbb{E}(X|Y)\Big] = \sum_y \mathbb{E}(X|Y=y)p_Y(y).$$

We can rewrite this expression by using formula (4.2) for conditional expectation (modified here since we are calculating $\mathbb{E}(X|Y = y)$ and not $\mathbb{E}(Y|X = x)$).

$$\begin{aligned}\mathbb{E}\Big[\mathbb{E}(X|Y)\Big] &= \sum_y \frac{\sum_x x p_{X,Y}(x,y)}{p_Y(y)} p_Y(y) \\ &= \sum_x x \sum_y p_{X,Y}(x,y) \\ &= \sum_x x p_X(x) = \mathbb{E}X.\end{aligned}$$

where in the second line we changed the order of summation and for the third equality used formula (4.1) for the marginal pmf of X. □

Conditional Variance

By using the conditional pmf, we can also define the conditional expectation of every function of X given a realization of Y:

$$\mathbb{E}(g(X)|Y = y) = \sum_x g(x) p_{X|Y}(x|y).$$

In particular we can define the conditional variance of r.v. X given Y:

$$\mathbb{V}\mathrm{ar}(X|Y = y) = \mathbb{E}\Big[\big(X - \mathbb{E}(X|Y = y)\big)^2 \big| Y = y\Big].$$

We will not study the conditional variance in detail here but it is useful and we encounter it later.

Since the conditional variance is simply the variance with respect to the conditional pmf, we still have the useful formula in terms of the conditional second and first moments:

$$\mathbb{V}\mathrm{ar}(X|Y = y) = \mathbb{E}(X^2|Y = y) - \Big(\mathbb{E}(X|Y = y)\Big)^2.$$

Of course we can also define the conditional standard deviation as the square root of conditional variance.

Example 4.1.13. Consider the situation of Example 4.1.5. Calculate the conditional variance $\mathbb{V}\mathrm{ar}(X|Y = 1)$.

The conditional pmf for $Y = 1$ is $p_{X|Y}(0|1) = \frac{1}{2}$, $p_{X|Y}(1|1) = \frac{1}{2}$, and $p_{X|Y}(2|1) = 0$, and we calculated the conditional expectation $\mathbb{E}(X|Y = 1) = \frac{1}{2}$. In order to calculate the conditional variance, we calculate the conditional second moment:

$$\begin{aligned}\mathbb{E}(X^2|Y=1) &= \sum_x x^2 p_{X|Y}(x|1) \\ &= 0^2 \times \frac{1}{2} + 1^2 \times \frac{1}{2} + 2^2 \times 0 = \frac{1}{2}\end{aligned}$$

Hence,

$$\mathbb{V}\mathrm{ar}(X|Y=1) = \frac{1}{2} - \left(\frac{1}{2}\right)^2 = \frac{1}{4}.$$

4.2 Dependence between random variables

4.2.1 Independent random variables

Definition 4.2.1. Two discrete random variables X and Y are called *independent* if their joint pmf equal the product of their marginal pmf's for every value x and y:

$$p_{X,Y}(x,y) = p_X(x)p_Y(y).$$

Note that this simply means that the events $\{X = x\}$ and $\{Y = y\}$ are independent for every choice of x and y.

If X and Y are independent then we can calculate that the conditional pmf of X given $Y = y$ equals the marginal pmf of X:

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)} = \frac{p_X(x)p_Y(y)}{p_Y(y)} = p_X(x).$$

So the intuitive meaning of the independence is that if we observe any particular value of r.v. Y then it does not affect the probability that we observe any particular value of r.v. X. It remains equal to the probability that we had before this observation. Of course, it also works in the other direction: if we observe that $X = x$ then it does not affect the probability that $Y = y$. The conditional probability $p_{Y|X}(y|x)$ equals the marginal probability $p_Y(y)$.

Exercise 4.2.2. Convince yourself that the random variables in Example 4.1.3 are independent, and that the random variables in Examples 4.1.4 and 4.1.5 are not independent.

Independent random variables are building blocks for many multivariate distributions that occur in practice.

Example 4.2.3. Let X and Y be independent Poisson random variables with parameters μ and λ, respectively. What is their joint pmf?

From the definition of independence we have:

$$p_{X,Y}(x,y) = p_X(x)p_Y(y) = e^{-\mu}\frac{\mu^x}{x!}e^{-\lambda}\frac{\lambda^y}{y!}, \text{ for } x = 0,1,\ldots, y = 0,1,\ldots.$$

For example if $\mu = \lambda = 1$, then

$$p_{X,Y}(x,y) = e^{-2}\frac{1}{x!y!}.$$

An important fact about independent random variables is that their transformations are also independent.

Theorem 4.2.4. *Let X and Y be independent discrete random variables and f and g are function from $\mathbb{R}$ to $\mathbb{R}$. Then $f(X)$ and $g(Y)$ are also independent.*

Proof. Consider the joint pmf for $f(X)$ and $g(Y)$ evaluated at a and b:

$$\begin{aligned}
\mathbb{P}\Big(f(X) = a, g(Y) = b\Big) &= \sum_{\substack{x:f(x)=a \\ y:g(y)=b}} \mathbb{P}(X = x, Y = y) \\
&= \sum_{\substack{x:f(x)=a \\ y:g(y)=b}} \mathbb{P}(X = x)\mathbb{P}(Y = y) \\
&= \Big(\sum_{x:f(x)=a} \mathbb{P}(X = x)\Big)\Big(\sum_{y:g(y)=b} \mathbb{P}(Y = y)\Big) \\
&= \mathbb{P}(f(X) = a)\mathbb{P}(g(Y) = b),
\end{aligned}$$

which is the product of marginal pmf of $f(X)$ and $g(Y)$ evaluated at a and b. □

(To see how the proof works, suppose that x_1 and x_2 are the only possible values of X that are mapped by function f to a, and y_1, y_2 are the only possible values of Y mapped by g to b. Then,

$$\begin{aligned} P\Big(f(X)=a, g(Y)=b\Big) &= p_X(x_1)p_Y(y_1) + p_X(x_1)p_Y(y_2) \\ &\quad + p_X(x_2)p_Y(y_1) + p_X(x_2)p_X(y_2) \\ &= \Big(p_X(x_1)+p_X(x_2)\Big)\Big(p_Y(y_1)+p_Y(y_2)\Big) \\ &= \mathbb{P}(f(X)=a)\mathbb{P}(g(Y)=b), \end{aligned}$$

as it should be.)

One of the most important theorems about independent random variables is that the expectation of their product equals the product of their expectations.

Before stating this theorem, let us explain how we calculate the expectations of functions of several random variables. If $Z = f(X,Y)$ then to calculate the expectation of Z by definition, we need to calculate the pmf of Z first and then use the definition $\mathbb{E}Z = \sum_z z p_Z(z)$. However, it turns out that a simpler method can be used:

$$\mathbb{E}\big[f(X,Y)\big] = \sum_{x,y} f(x,y)p_{X,Y}(x,y).$$

We will skip the proof of this result.

Theorem 4.2.5. *If X and Y are independent random variable with finite expectation, then*

$$\mathbb{E}(XY) = (\mathbb{E}X)(\mathbb{E}Y)$$

Proof. The random variable XY is the function of random variables X and Y, so

$$\begin{aligned}\mathbb{E}(XY) &= \sum_{x,y} xy\, p_{X,Y}(x,y) \\ &= \sum_{x,y} xy\, p_X(x) p_Y(y) \\ &= \Big(\sum_x x\, p_X(x)\Big)\Big(\sum_y y\, p_Y(y)\Big) \\ &= \mathbb{E}(X)\mathbb{E}(Y).\end{aligned}$$

□

If we combine this with the previous theorem we get a more general result:

Theorem 4.2.6. *Let X and Y be independent discrete random variables and f and g are function from $\mathbb{R}$ to $\mathbb{R}$. Then*

$$\mathbb{E}\big(f(X)g(Y)\big) = \mathbb{E}\big(f(X)\big)\mathbb{E}\big(g(Y)\big).$$

Proof. This follows directly from the previous two theorems. □

Now we can prove two theorems that we announced without proof before.

Theorem 4.2.7. *Let X and Y be independent discrete random variables. Then, for the moment generating function of their sum, we have:*

$$m_{X+Y}(t) = m_X(t) m_Y(t).$$

Proof. We have:

$$\begin{aligned}m_{X+Y}(t) := \mathbb{E}e^{t(X+Y)} &= \mathbb{E}\Big(e^{tX}e^{tY}\Big) \\ &= \mathbb{E}e^{tX}\mathbb{E}e^{tY} = m_X(t)m_Y(t),\end{aligned}$$

where the first equality in the second line follows by the previous theorem. □

Theorem 4.2.8. *Let X and Y be independent discrete random variables. Then*

$$\mathbb{V}\mathrm{ar}(X+Y) = \mathbb{V}\mathrm{ar}(X) + \mathbb{V}\mathrm{ar}(Y).$$

Proof. We have

$$\begin{aligned}
\mathbb{V}\mathrm{ar}(X+Y) &= \mathbb{E}(X+Y)^2 - \Big(\mathbb{E}(X+Y)\Big)^2 \\
&= \Big(\mathbb{E}X^2 + 2\mathbb{E}(XY) + \mathbb{E}Y^2\Big) - \big(\mathbb{E}X\big)^2 + 2(\mathbb{E}X)(\mathbb{E}Y) - (\mathbb{E}Y)^2 \\
&= \mathbb{V}\mathrm{ar}(X) + \mathbb{V}\mathrm{ar}(Y) + 2\Big[\mathbb{E}(XY) - \big(\mathbb{E}X\big)\big(\mathbb{E}Y\big)\Big] \\
&= \mathbb{V}\mathrm{ar}(X) + \mathbb{V}\mathrm{ar}(Y),
\end{aligned}$$

where the last line follows because $\mathbb{E}(XY) - \big(\mathbb{E}X\big)\big(\mathbb{E}Y\big)$ by Theorem 4.2.5. □

4.2.2 Covariance

Sometimes we want to measure the degree to which the random variables X and Y are related to each other. A useful measure is called the covariance of X and Y.

Definition 4.2.9. The *covariance* of random variables X and Y is defined as

$$\mathrm{Cov}(X,Y) = \mathbb{E}(XY) - \big(\mathbb{E}X\big)\big(\mathbb{E}Y\big).$$

Note in particular, that the covariance is the generalization of the concept of variance. The variance of the random variable X is simply the covariance of X with itself:

$$\mathrm{Cov}(X,X) = \mathbb{E}(X^2) - (\mathbb{E}X)(\mathbb{E}X) = \mathbb{V}\mathrm{ar}(X).$$

For independent variables the covariance is zero by Theorem 4.2.5. However, the converse is not true: there are some random variables that have zero covariance but are not independent.

The covariance can be defined in equivalent way:

$$\mathrm{Cov}(X,Y) = \mathbb{E}\Big(\big(X-\mathbb{E}X\big)\big(Y-\mathbb{E}Y\big)\Big).$$

(It is an easy exercise to check that this is equivalent to the original definition.)

From this definition it can be seen that the covariance is positive if the deviations of random variables X and Y from their means tend to have the same sign in a realization of the random experiment, and it is negative if they tend to have the opposite sign. So, roughly speaking, the covariance measures the degree to which two random variables vary together.

Example 4.2.10. Let us again consider the situation in Example 4.1.5: 2 grants randomly awarded to 3 universities. Random variable X is the number of grants awarded to University N, and r.v. Y is the number of grants awarded to University C. What is the covariance of X and Y?

Recall that the joint probability is given by the following table:

$x\backslash y$	0	1	2
0	1/9	2/9	1/9
1	2/9	2/9	0
2	1/9	0	0

So we can calculate:

$$\begin{aligned}\mathbb{E}(XY) &= \sum_{x,y} xy\, p_{X,Y}(x,y)\\ &= 1\times 1\times\frac{2}{9} = \frac{2}{9}.\end{aligned}$$

The marginal pmf for X assigns probabilities $\frac{4}{9}$, $\frac{4}{9}$, $\frac{1}{9}$ to values 0, 1, and 2, so we can calculate $\mathbb{E}X = 0\times\frac{4}{9}+1\times\frac{4}{9}+2\times\frac{1}{9} = \frac{2}{3}$. By symmetry, $\mathbb{E}Y = \frac{2}{3}$, as well. So we get:

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathbb{E}(XY) - \big(\mathbb{E}X\big)\big(\mathbb{E}Y\big)\\ &= \frac{2}{9} - \frac{2}{3}\times\frac{2}{3} = -\frac{2}{9}.\end{aligned}$$

This is reasonable since the number of grants awarded to Universities N and C tend to move in opposite directions. The more grants is awarded to University C, the less grants is available for University N.

$Y \backslash X$	-1	0	1	$p(y_j)$
0	0	1/2	0	1/2
1	1/4	0	1/4	1/2
$p(x_i)$	1/4	1/2	1/4	1

Example 4.2.11.

The random variables X and Y with pmf as in the picture have zero covariance, however, they are not independent.

By repeating the proof of Theorem 4.2.8, we can derive an important identity that holds for all random variables:

$$\mathbb{V}\mathrm{ar}(X+Y) = \mathbb{V}\mathrm{ar}(X) + \mathbb{V}\mathrm{ar}(Y) + 2\mathrm{Cov}(X,Y).$$

Here are some useful properties of covariance, which is easy to check.

1. Covariance is symmetric $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$.
2. The covariance of a constant (that is, a non-random quantity) with every random variable is zero:

$$\mathrm{Cov}(c,X) = \mathrm{Cov}(X,c) = 0,$$

3. Covariance is linear in every argument:

$$\mathrm{Cov}(aX+bY,Z) = a\mathrm{Cov}(X,Z) + b\mathrm{Cov}(Y,Z) \text{ and}$$
$$\mathrm{Cov}(Z,aX+bY) = a\mathrm{Cov}(Z,X) + b\mathrm{Cov}(Z,Y)$$

Note in particular that properties (2) and (3) imply that

$$\mathrm{Cov}(aX+b,cY+d) = ac\mathrm{Cov}(X,Y).$$

By using these properties it is easy to calculate the covariances of random variables which are sums of other random variables.

Example 4.2.12.

$$\begin{aligned}\mathbb{V}\mathrm{ar}(aX+bY+c) &= \mathrm{Cov}(aX+bY,aX+bY)\\ &= a^2\mathbb{V}\mathrm{ar}(X) + b^2\mathbb{V}\mathrm{ar}(Y) + 2ab\mathrm{Cov}(X,Y).\end{aligned}$$

Here is another example.

Example 4.2.13. Let $U = aX + bY$ and $W = cX + dY$. Calculate $\mathrm{Cov}(U, W)$.

Then,

$$\begin{aligned}\mathrm{Cov}(U,W) &= \mathrm{Cov}(aX+bY, cX+dY)\\ &= \mathrm{Cov}(aX,cX) + \mathrm{Cov}(aX,dY) + \mathrm{Cov}(bY,cX) + \mathrm{Cov}(bY,dY)\\ &= ac\mathbb{V}\mathrm{ar}(X) + (ad+bc)\mathrm{Cov}(X,Y) + bd\mathbb{V}\mathrm{ar}(Y).\end{aligned}$$

In general we have the following formula:

$$\mathrm{Cov}\Big(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{n} b_j X_j\Big) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i b_j \mathrm{Cov}(X_i, X_j).$$

and for variance, it can be rewritten as

$$\begin{aligned}\mathbb{V}\mathrm{ar}\Big(\sum_{i=1}^{n} a_i X_i\Big) &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \mathrm{Cov}(X_i, X_j)\\ &= \sum_{i=1}^{n} a_i^2 \mathbb{V}\mathrm{ar}(X_i) + 2\sum_{i<j} a_i a_j \mathrm{Cov}(X_i, X_j),\end{aligned}$$

where the second sum is a sum over all pairs of i and j so that $i < j$.

Example 4.2.14. Suppose $\mathbb{V}\mathrm{ar}(X_1) = \mathbb{V}\mathrm{ar}(X_2) = 1$, $\mathrm{Cov}(X_1, X_2) = -1/2$. Calculate

$$\mathrm{Cov}(3X_1 + X_2 + 5, X_1 - 2X_2 - 10),$$
$$\text{and } \mathbb{V}\mathrm{ar}(X_1 - 8X_2 + 1)$$

We can write

$$\begin{aligned}&\mathrm{Cov}(3X_1 + X_2 + 5, X_1 - 2X_2 - 10) = \mathrm{Cov}(3X_1 + X_2, X_1 - 2X_2)\\ &= 3\mathrm{Cov}(X_1,X_1) - 6\mathrm{Cov}(X_1,X_2) + \mathrm{Cov}(X_2,X_1) - 2\mathrm{Cov}(X_2,X_2)\\ &= 3\mathbb{V}\mathrm{ar}(X_1) - 5\mathrm{Cov}(X_1,X_2) - 2\mathbb{V}\mathrm{ar}(X_2)\\ &= 3 - 5\times(-\frac{1}{2}) - 2 = 3.5\end{aligned}$$

For variance:

$$\begin{aligned}\mathbb{V}\mathrm{ar}(X_1 - 8X_2 + 1) &= \mathbb{V}\mathrm{ar}(X_1 - 8X_2) \\ &= \mathbb{V}\mathrm{ar}(X_1) + 8^2 \times \mathbb{V}\mathrm{ar}(X_2) - 2 \times 8 \times \mathrm{Cov}(X_1, X_2) \\ &= 1 + 64 - 2 \times 8 \times (-\frac{1}{2}) = 73.\end{aligned}$$

The calculations become especially easy for sums of independent random variables, since the covariance of two independent random variables is zero.

Example 4.2.15. Toss a coin 3 times. Assume that the tosses are independent and the probability to get a head is p. Let X be the number of heads in the first 2 tosses, and Y = number of heads in the last 2 tosses. Compute $\mathrm{Cov}(X, Y)$.

Let I_k be the indicator random variable for the event that we get a head in k-th toss. That is $I_k = 1$ if the k-th toss resulted in a head and $I_k = 0$ if it is resulted in a tail. The random variables I_k are independent and we have

$$\begin{aligned}X &= I_1 + I_2, \\ Y &= I_2 + I_3.\end{aligned}$$

Then,

$$\begin{aligned}\mathrm{Cov}(X, Y) &= \mathrm{Cov}(I_1 + I_2, I_2 + I_3) \\ &= \mathrm{Cov}(I_1, I_2) + \mathrm{Cov}(I_1, I_3) + \mathrm{Cov}(I_2, I_2) + \mathrm{Cov}(I_2, I_3) \\ &= \mathbb{V}\mathrm{ar}(I_2) = p(1-p).\end{aligned}$$

Example 4.2.16. Suppose $X_1, \ldots, X_n$ are n independent identically distributed ("i.i.d.") random variables with variance σ^2. Define their average as

$$\overline{X} = \frac{1}{n}(X_1 + \ldots + X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Find $\mathbb{V}\mathrm{ar}(\overline{X})$ and $\mathrm{Cov}(X_i - \overline{X}, \overline{X})$.

For variance, we have

$$\mathbb{V}\mathrm{ar}\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) = \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}\mathrm{ar}(X_i) = \frac{1}{n^2} \times n\sigma^2 = \frac{\sigma^2}{n}.$$

For covariance:

$$\begin{aligned}\mathrm{Cov}(X_i - \overline{X}, \overline{X}) &= \mathrm{Cov}\Big(X_i, \frac{1}{n}\sum_{j=1}^{n} X_j\Big) - \mathbb{V}\mathrm{ar}(\overline{X}) \\ &= \frac{1}{n}\mathrm{Cov}(X_i, X_i) - \frac{\sigma^2}{n} = 0.\end{aligned}$$

The covariance of two random variables always satisfies an important inequality.

Theorem 4.2.17. *Let X and Y be two random variables with finite variance. Then*

$$\Big|\mathrm{Cov}(X, Y)\Big| \leq \sqrt{\mathbb{V}\mathrm{ar}(X)\mathbb{V}\mathrm{ar}(Y)}.$$

Proof. Consider the variance of $X + tY$ where t is an arbitrary non-random number.

$$\begin{aligned}\mathbb{V}\mathrm{ar}(X + tY) &= \mathrm{Cov}(X + tY, X + tY) \\ &= \mathrm{Cov}(X, X) + 2t\mathrm{Cov}(X, Y) + t^2\mathrm{Cov}(Y, Y) \\ &= \mathbb{V}\mathrm{ar}(X) + 2t\mathrm{Cov}(X, Y) + t^2\mathbb{V}\mathrm{ar}(Y).\end{aligned}$$

This is a quadratic polynomial in t. Since it represents the variance of a random variable, it is non-negative for all real t. Hence, by properties of quadratic polynomials, the discriminant of this polynomial must be non-positive. (Otherwise it would have two real roots and it would take negative values between these roots.) Hence,

$$D = \mathrm{Cov}(X, Y)^2 - \mathbb{V}\mathrm{ar}(X)\mathbb{V}\mathrm{ar}(Y) \leq 0.$$

This inequality is equivalent to the inequality in the statement of the theorem. □

4.2.3 Correlation coefficient

One deficiency of the covariance as a measure of deviation from independence is that it depends on the scale of the random variables. If we multiply one of the variables X or Y by a constant factor a, then $\mathrm{Cov}(aX, Y) = a\mathrm{Cov}(X, Y)$.

The measure that removes this dependence on the scale (and on the choice of units) is the correlation coefficient.

Definition 4.2.18. For random variables X, Y with finite variances, the *correlation coefficient* is defined as

$$\mathrm{Cor}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathbb{V}\mathrm{ar}(X)\mathbb{V}\mathrm{ar}(Y)}}.$$

From the previous theorem we have that $-1 \leq \mathrm{Cor}(X,Y) \leq 1$. It is possible to show that the correlation is 1 if and only if $Y = aX + b$ with $a > 0$, and it is -1 if and only if $Y = aX + b$ with $a < 0$.

Two other properties of the correlation coefficient:

1. Correlation between X and Y is the covariance of the standardized versions of X and Y:

$$\mathrm{Cor}(X,Y) = \mathrm{Cov}\Big(\frac{X - \mathbb{E}X}{Std(X)}, \frac{Y - \mathbb{E}Y}{Std(Y)}\Big)$$

2. $\mathrm{Cor}(aX + b, cY + d)$ equals $\mathrm{Cor}(X,Y)$ if $ac > 0$,
and it equals $-\mathrm{Cor}(X,Y)$ if $ac < 0$.

Example 4.2.19. Consider the example with 2 grants awarded to 3 universities that we introduced earlier. What is the correlation between X and Y?

We calculated that $\mathrm{Cov}(X,Y) = -\frac{2}{9}$. For the variances, we use the marginal pmf in order to calculate:

$$\mathbb{E}(X^2) = 0^2 \times \frac{4}{9} + 1^2 \times \frac{4}{9} + 2^2 \times \frac{1}{9} = \frac{8}{9}.$$

Since we already calculated that $\mathbb{E}(X) = \frac{2}{3}$, we find that

$$\mathbb{V}\mathrm{ar}(X) = \frac{8}{9} - \Big(\frac{2}{3}\Big)^2 = \frac{4}{9}.$$

By symmetry, $\mathbb{V}\mathrm{ar}(Y) = \frac{4}{9}$ and the standard deviations of both variables are $\frac{2}{3}$. Hence,

$$\mathrm{Cor}(X,Y) = \frac{-\frac{2}{9}}{\frac{2}{3} \times \frac{2}{3}} = -\frac{1}{2}.$$

4.3 Continuous random variables

4.3.1 Joint cdf and pdf

For random variables that take arbitrary real values, the joint probability mass function is not sufficient to describe the properties of these variables since it is often zero everywhere. We need to define an analogue of the probability density, which we used previously to describe a single continuous random variable. As a first step to this goal, we define the joint cumulative distribution function, the joint cdf.

Definition 4.3.1. The *joint cumulative distribution function (cdf)* of the random variables X and Y is

$$F_{X,Y}(x, y) = \mathbb{P}(X \leq x, Y \leq y).$$

For discrete random variables, we can calculate the joint cdf by using the joint pmf:

$$F_{X,Y}(x, y) = \sum_{\substack{s \leq x, \\ t \leq y}} p(s, t).$$

Example 4.3.2. What is $F_{X,Y}(1, 1)$ in the example about grants and universities?

$$\begin{aligned} F_{X,Y}(1, 1) &= p_{X,Y}(0, 0) + p_{X,Y}(1, 0) + p_{X,Y}(0, 1) + p_{X,Y}(1, 1) \\ &= \frac{1}{9} + \frac{2}{9} + \frac{2}{9} + \frac{2}{9} = \frac{7}{9}. \end{aligned}$$

The benefit of the joint cdf is that it is defined for arbitrary random variables, not only for discrete random variables.

Random variables X and Y are called *jointly continuous* if their joint cdf $F_{X,Y}(x, y)$ is continuous in both arguments. (Curiously, there are variables X and Y that are both continuous as single random variables but are not jointly continuous.)

Now, we assume in addition that the joint cdf has at least two derivatives everywhere except possible some lines and points.

Then we can define the joint probability density function (pdf).

Definition 4.3.3. The *joint probability density (pdf)* of jointly continuous random variables X and Y is defined as

$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y).$$

The joint pdf is typically the main tool to describe the jointly continuous random variables.

In particular, if we are given the joint pdf, we can recover the joint cdf:

$$F_{X,Y}(x,y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f(s,t)\, ds dt.$$

The joint pdf is a non-negative function and its important property is that

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(s,t)\, ds dt = 1.$$

The intuitive meaning of the joint density value at a point (x,y) is that it is a normalized probability to find X and Y in a small square around (x,y). Namely, it allows us to calculate the probability that the pair of random variables X and Y takes value in a small rectangle around (x,y) as follows:

$$\mathbb{P}\big(X \in (x, x+dx), Y \in (y, y+dy)\big) \approx f_{X,Y}(x,y)\, dx\, dy,$$

assuming dx and dy are very small.

4.3.2 Calculating probabilities using joint pdf

If a probability event E can be formulated in terms of random variables X and Y,

$$E = \Big\{\omega : \big(X(\omega), Y(\omega)\big) \in A \subset \mathbb{R}^2\Big\}$$

then finding the probability of the event E amounts to integrating the joint pdf over the corresponding region A:

$$\mathbb{P}(E) = \int_A f_{X,Y}(x,y)\, dxdy.$$

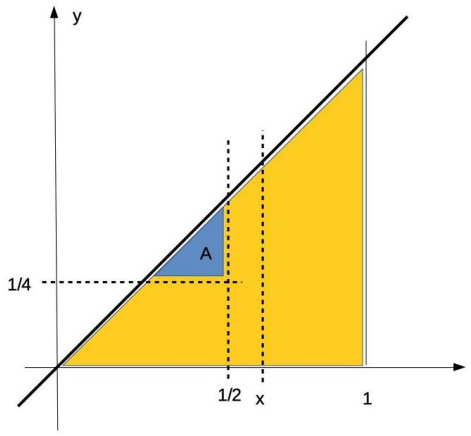


Figure 4.1

The tricky part here is that $f_{X,Y}$ is zero outside of its range, so we have to integrate over the intersection of the region A and the region where the density $f_{X,Y}(x,y)$ is not zero (the range of X, Y).

Example 4.3.4. Suppose we model two random variables X and Y with the joint pdf

$$f_{X,Y}(x,y) = \begin{cases} cx, & \text{for } 0 < y < x < 1, \\ 0, & \text{elsewhere.} \end{cases}$$

Find the constant c. Find the probability that $X < 1/2$ and $Y > 1/4$.

The range of the vector X, Y is the triangle R defined by the inequalities $y > 0$, $y < x$, and $x < 1$. In order to find the density we integrate the function x over this triangle. We do it by writing the double integral as a repeated integral We calculate:

$$\begin{aligned} \int\int_R x\,dxdy &= \int_0^1 \int_0^x x\,dy\,dx = \int_0^1 \left(xy\Big|_{y=0}^{y=x} \right) dx \\ &= \int_0^1 x^2\,dx = \frac{1}{3}. \end{aligned}$$

So, in order to make sure that the integral of the joint density over the triangle R equals 1, we have to set $c = 3$.

In order to calculate the probability $\mathbb{P}(Y > 1/4, X < 1/2)$, we integrate the joint pdf over different triangle A, which is the intersection of the region given by the inequalities $\{Y > 1/4, X < 1/2\}$ and the range R.

$$\begin{aligned}
\mathbb{P}(X < 1/2, Y > 1/4) &= \int_{1/4}^{1/2} \int_{1/4}^{x} 3x \, dy \, dx \\
&= \int_{1/4}^{1/2} \Big[3xy\Big]_{y=1/4}^{y=x} dx \\
&= \int_{1/4}^{1/2} \Big(3x^2 - \frac{3x}{4}\Big) dx \\
&= \Big[x^3 - \frac{3x^2}{8}\Big]_{x=1/4}^{x=1/2} \\
&= \Big[\frac{1}{8} - \frac{3}{32} - (\frac{1}{4^3} - \frac{3}{4^2 \times 8})\Big] \\
&= \frac{5}{128} = 3.90625\%
\end{aligned}$$

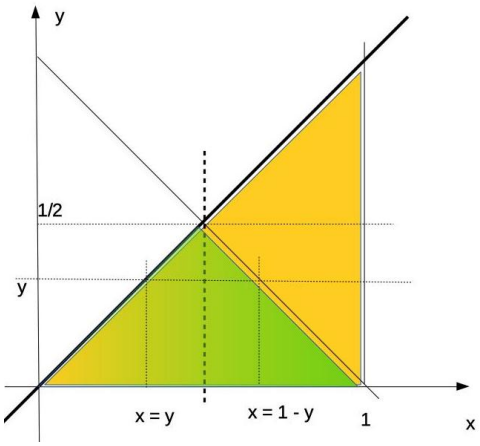


Figure 4.2

Example 4.3.5. In the previous example, find the probability that $X+Y < 1$.

We need to integrate over the green triangle which is the intersection of the range R of random variables X and Y and the half-plane given by the inequality $x + y < 1$. There are two ways to setup the repeated integral.

The first possibility is to have the inner integration over y and the outer integration over x. In this case, we are forced to divide the integration into two pieces (over two different triangles) depending on whether x is smaller or greater than $\frac{1}{2}$:

$$\mathbb{P}(X+Y<1)=\int_0^{1/2}\int_0^x 3x\,dy\,dx+\int_{1/2}^1\int_0^{1-x}3x\,dy\,dx$$

The second method (shown in Figure 4.2) is to have the outer integration over y and inner integration over x. Then we can write:

$$\mathbb{P}(X+Y<1)=\int_0^{1/2}\int_y^{1-y}3x\,dx\,dy.$$

Let us calculate the second integral:

$$\begin{aligned}\int_0^{1/2}\int_y^{1-y}3x\,dx\,dy&=\int_0^{1/2}\frac{3}{2}x^2\Big|_{x=y}^{x=1-y}dy\\&=\int_0^{1/2}\frac{3}{2}(1-2y)\,dy=\frac{3}{2}(y-y^2)\Big|_0^{1/2}\\&=\frac{3}{8}.\end{aligned}$$

(The first method gives:

$$\begin{aligned}\int_0^{1/2}\int_0^x 3x\,dy\,dx&=\int_0^{1/2}3x^2\,dx=\frac{1}{8},\\\int_{1/2}^1\int_0^{1-x}3x\,dy\,dx&=\int_{1/2}^1 3x(1-x)\,dx\\&=(\frac{3}{2}x^2-x^3)\Big|_{1/2}^1=\frac{3}{2}\times\frac{3}{4}-\frac{7}{8}=\frac{2}{8},\end{aligned}$$

so it gives the same result $\mathbb{P}(X+Y<1)=\frac{3}{8}$ but the calculations are a bit more cumbersome.)

Exercise 4.3.6. Suppose the time (in hours) to complete task 1 and task 2 for a random employee has the joint pdf:

$$f(X,Y) = \begin{cases} e^{-(x+y)}, & \text{for } 0 < x, 0 < y, \\ 0, & \text{elsewhere.} \end{cases}$$

Find the probability that a random employee takes less than 2 hours on task 1 and between 1 and 3 hours on task 2. What is the probability the employee takes longer on task 2 than on task 1?

4.3.3 Marginal densities

The marginal density is the analogue of marginal pmf for continuous random variables.

Definition 4.3.7. For a jointly continuous r.v.'s X and Y, the marginal pdf $f_X(x)$ is the probability density of the random variable X considered as a single random variable.

The marginal density of X can be found from the joint density by integrating over all values y in the range of Y.

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dy.$$

While the integral above is written from $-\infty$ to ∞, in fact we integrate over the range of Y, when $X = x$, since outside this range the joint density is 0. Again, this requires to be attentive to the range of the integration. It is highly advisable to draw a picture.

Similarly for $f_Y(y)$, we have

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dx.$$

Example 4.3.8. Suppose we model two random variables X and Y with the joint pdf

$$f_{X,Y}(x,y) = \begin{cases} 3x, & \text{for } 0 < y < x < 1, \\ 0, & \text{elsewhere.} \end{cases}$$

Find the marginal pdf's of X and Y and find $\mathbb{P}(Y > 1/2)$. Find $\mathbb{P}(X + Y < 1 | Y > 1/4)$. Find $\mathbb{E}Y$.

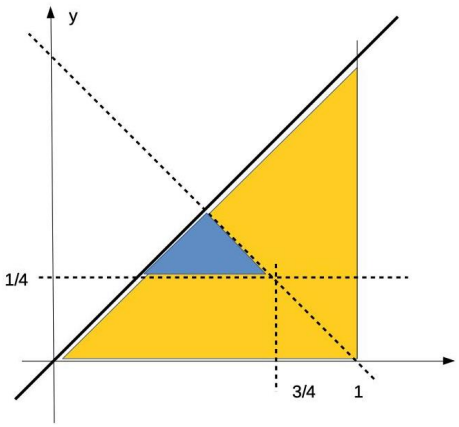


Figure 4.3

We calculate the marginal pdf's of X and Y by integrating the joint density $f_{X,Y}(x,y) = 3x$ over the suitable intervals. The important thing here is correctly set the limits of the integrations. They should correspond to the range of the variable which we integrating out. When we calculate $f_X(x)$ we are looking at the case when $X = x$ and evaluating the density at this point. In this case Y can only be between 0 and x. So these are our limits of integration.

In contrast if we evaluate $f_Y(y)$, then we consider $Y = y$, and then X can be only between y and 1. These facts are especially clear in the picture.

So we get the following two integrals:

$$f_X(x) = \int_0^x 3x\,dy = 3xy\big|_{y=0}^{y=x} = 3x^2,$$
$$f_Y(y) = \int_y^1 3x\,dx = \frac{3}{2}x^2\big|_{x=y}^{x=1} = \frac{3}{2}(1-y^2)$$

Now, in order to calculate $\mathbb{P}(Y > 1/2)$, we can integrate the marginal density:

$$\mathbb{P}(Y > 1/2) = \int_{1/2}^1 \frac{3}{2}(1-y^2)\,dy = \frac{3}{2}\big(y - \frac{y^3}{3}\big)\Big|_{y=1/2}^{y=1}$$
$$= \frac{3}{2}\Big(1 - \frac{1}{3} - \big(\frac{1}{2} - \frac{1/8}{3}\big)\Big) = \frac{5}{16}.$$

In order to calculate $\mathbb{P}(X+Y<1|Y>1/4)$, we need to calculate

$$\frac{\mathbb{P}(X+Y<1, Y>1/4)}{\mathbb{P}(Y>1/4)}$$

For the denominator, we have, similarly to the previous calculation,

$$\begin{aligned}\mathbb{P}(Y>1/4) &= \int_{1/4}^{1} \frac{3}{2}(1-y^2)\,dy = \frac{3}{2}\big(y-\frac{y^3}{3}\big)\Big|_{y=1/4}^{y=1} = 1-\frac{3}{2}\big(\frac{1}{4}-\frac{1/64}{3}\big)\\ &= 81/128.\end{aligned}$$

In order to calculate the numerator we need to integrate the joint density over the triangle shown in picture. We use the iterated integral with the outer integral over y and inner integral over x:

$$\begin{aligned}\mathbb{P}(X+Y<1, Y>1/4) &= \int_{1/4}^{1/2}\int_{y}^{1-y} 3x\,dx\,dy\\ &= \int_{1/4}^{1/2} \frac{3x^2}{2}\Big|_{x=y}^{x=(1-y)}\,dy\\ &= \int_{1/4}^{1/2} \frac{3}{2}(1-2y)\,dy\\ &= \frac{3}{2}(y-y^2)\Big|_{1/4}^{1/2}\\ &= \frac{3}{2}\Big(\frac{1}{2}-\frac{1}{4}-(\frac{1}{4}-\frac{1}{16})\Big) = \frac{3}{32}.\end{aligned}$$

Therefore

$$\mathbb{P}(X+Y<1|Y>1/4) = \frac{3/32}{81/128} = \frac{4}{27}.$$

For $\mathbb{E}Y$, we have:

$$\begin{aligned}\mathbb{E}Y &= \int_{-\infty}^{\infty} y f_Y(y)\,dy = \int_0^1 y\frac{3}{2}(1-y^2)\,dy\\ &= \Big[\frac{3}{4}y^2-\frac{3}{8}y^4\Big]_{y=0}^{y=1}\\ &= \frac{3}{8}.\end{aligned}$$

Exercise 4.3.9. Suppose for two proportions X and Y, we know that $X < Y$. We assume that the joint pdf is

$$f(x,y) = \begin{cases} cx, & \text{for } 0 < x < y < 1, \\ 0, & \text{elsewhere.} \end{cases}$$

Find c. Find the marginal pdf of X and Y.

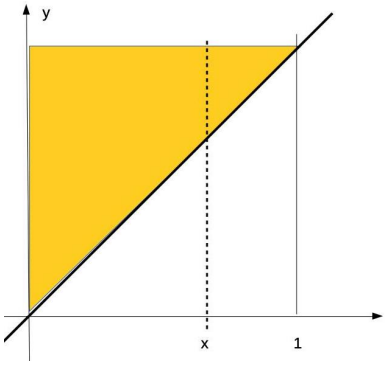


Figure 4.4

Note that the only difference of this exercise from the previous example is that the region at which the density is not zero and proportional to x is different. However, this will change both c and the marginal densities.

Exercise 4.3.10. Suppose X and Y are random variables and (X, Y) takes values in $[0,1] \times [0,1]$.

Suppose that the pdf is

$$k(x^2 + y^2).$$

Find k. Find the marginal pdf $f_X(x)$. Use this to find $\mathbb{P}(X < .5)$.

4.3.4 Conditional density (pdf)

Recall that for discrete random variables we defined the conditional pmf as

$$p_{X|Y}(x|y) := \mathbb{P}(X = x|Y = y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}.$$

This does not make sense for continuous random variables because both the numerator and denominator are zero. However we can generalize this formula by using joint and marginal densities instead of joint and marginal pmfs.

Definition 4.3.11. For jointly continuous random variables X and Y, the conditional probability density function (pdf) of X given Y is defined as

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{\text{joint pdf}}{\text{marginal pdf}}.$$

Similarly, the conditional pdf of Y given X is

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}.$$

Note that the conditional pdf $f_{X|Y}(x|y)$ is indeed a pdf as a function of x, (but not of y, which simply a parameter). In particular,

$$\int_{-\infty}^{\infty} f_{X|Y}(x|y)\,dx = 1$$

for every y in the range of r.v. Y. This is because

$$\int_{-\infty}^{\infty} f_{X|Y}(x|y)\,dx = \frac{1}{f_Y(y)} \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx = \frac{f_Y(y)}{f_Y(y)} = 1,$$

by the formula that says that the marginal density $f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx$.

The intuitive meaning of the conditional pdf $f_{X|Y}(x|y)$ is that it is the probability density of the r.v. X once we learned that r.v. Y equals y. We use it to calculate the probabilities of events under condition that $Y = y$, as illustrated in the next example.

Example 4.3.12.

$$f_{X,Y}(x,y) = \begin{cases} 3x, & \text{for } 0 < y < x < 1, \\ 0, & \text{elsewhere.} \end{cases}$$

Calculate the conditional densities $f_{X|Y}(x|y)$ and $f_{Y|X}(y|x)$. What is the probability that $X + Y < 1$ given that $Y = 1/4$?

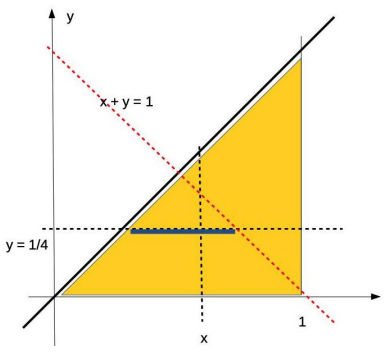


Figure 4.5

The marginal pdf's have been calculated in Example 4.3.8. So for the conditional density $f_{X|Y}(x|y)$ we have:

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{3x}{\frac{3}{2}(1-y^2)} = \frac{2x}{1-y^2}.$$

It is important to specify the region for which this formula holds. For a fixed y, the

joint density is not zero only if $y < x < 1$. So the final answer here is

$$f_{X|Y}(x|y) = \begin{cases} \frac{2x}{1-y^2}, & \text{if } x \in (y,1), \\ 0, & \text{otherwise.} \end{cases}$$

Similarly, for $f_{Y|X}(y|x)$, we calculate

$$\begin{aligned} f_{Y|X}(y|x) &= \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{3x}{3x^2} \\ &= \frac{1}{x}. \end{aligned}$$

For a fixed x the joint density is not zero only if $0 < y < x$. So the final answer is

$$f_{Y|X}(y|x) = \begin{cases} \frac{1}{x}, & \text{if } y \in (0,x), \\ 0, & \text{otherwise.} \end{cases}$$

Note that in this example the conditional density of Y given $X = x$ does not depend on y. This means that the conditional distribution is uniform on the interval $(0, x)$.

Finally, let us calculate the probability $\mathbb{P}(X + Y < 1 | Y = 1/4)$. Given that $Y = 1/4$, the event $X + Y < 1$ is the same as $X < 1 - \frac{1}{4} = \frac{3}{4}$. So we need to calculate $\mathbb{P}(X < \frac{3}{4} | Y = \frac{1}{4})$.

$$\begin{aligned} \mathbb{P}(X < \frac{3}{4} | Y = \frac{1}{4}) &= \int_{x<3/4} f_{X|Y}(x|\frac{1}{4})\, dx \\ &= \int_{1/4}^{3/4} \frac{2x}{1-(\frac{1}{4})^2}\, dx, \end{aligned}$$

where in the second line we used the formula for the conditional density $f_{X|Y}(x|y)$ and, in particular, the fact that this density is zero if $x < 1/4$. (Essentially, this is the integration of the conditional density over the blue

line in the picture.) Next, we calculate:

$$\begin{aligned}\int_{1/4}^{3/4} \frac{2x}{1-(\frac{1}{4})^2}\,dx &= \frac{16}{15}\int_{1/4}^{3/4} 2x\,dx \\ &= \frac{16}{15}\left(x^2\Big|_{1/4}^{3/4}\right) = \frac{16}{15}\left(\frac{9}{16}-\frac{1}{16}\right) \\ &= \frac{8}{15}.\end{aligned}$$

So the answer is

$$\mathbb{P}(X+Y<1|Y=1/4) = \frac{8}{15}.$$

4.3.5 Conditional Expectations

By using conditional densities, we can define conditional expectations.

Definition 4.3.13. The conditional expectations of a function $g(X)$ given $Y=y$ is defined as

$$\mathbb{E}[g(X)|Y=y] = \int g(x) f_{X|Y}(x|y)\,dx.$$

In particular,

$$\mathbb{E}[X|Y=y] = \int x f_{X|Y} f(x|y)\,dx.$$

Note that a conditional expectation of $g(X)$ given $Y=y$ is a function of y. In particular we can think about the conditional expectation as a random quantity, which is a function of the random variable Y.

Example 4.3.14.

$$f_{X,Y}(x,y) = \begin{cases} 3x, & \text{for } 0<y<x<1, \\ 0, & \text{elsewhere.}\end{cases}$$

Calculate the conditional expectations $\mathbb{E}(X|Y=y)$ and $\mathbb{E}(Y|X=x)$.

By using the conditional densities we calculate:

$$\begin{aligned}\mathbb{E}(X|Y=y) &= \int_y^1 x\frac{2x}{1-y^2}\,dx = \frac{2}{1-y^2}\left[\frac{x^3}{3}\right]_{x=y}^{x=1} \\ &= \frac{2}{3}\frac{1-y^3}{1-y^2} = \frac{2}{3}\frac{1+y+y^2}{1+y}.\end{aligned}$$

$$\begin{aligned}\mathbb{E}(Y|X=x) &= \int_0^x y\frac{1}{x}\,dy = \frac{1}{x}\Big[\frac{y^2}{2}\Big]_{y=0}^{y=x} \\ &= \frac{x^2}{2x} = \frac{x}{2}.\end{aligned}$$

The last result can also be seen from the fact that conditionally on $X = x$, the random variable Y is distributed uniformly at the interval $[0, x]$, hence its conditional expectation is exactly in the middle of this interval at $x/2$.

Example 4.3.15. Let us check that the law of repeated expectation: $\mathbb{E}\big[\mathbb{E}(X|Y)\big] = \mathbb{E}X$ holds in Example 4.3.14.

$$\begin{aligned}\mathbb{E}\big[\mathbb{E}(X|Y)\big] &= \int_0^1 \frac{2}{3}\frac{1-y^3}{1-y^2} \times \frac{3}{2}(1-y^2)\,dy \\ &= \int_0^1 (1-y^3)\,dy = \Big[y - \frac{y^4}{4}\Big]_0^1 \\ &= \frac{3}{4}.\end{aligned}$$

In the first line, we used that the marginal density of Y is $\frac{3}{2}(1-y^2)$.

On the other hand:

$$\mathbb{E}X = \int_0^1 x \times 3x^2\,dx = \frac{3}{4}x^4\Big|_0^1 = \frac{3}{4},$$

where we used the fact that $3x^2$ is the marginal density of x.

Exercise 4.3.16. Suppose for two proportions X and Y, we know that $X < Y$. We assume that the joint pdf is

$$f(x,y) = \begin{cases} 6x, & \text{for } 0 \le x < y \le 1, \\ 0, & \text{elsewhere.} \end{cases}$$

Find $f_{Y|X}(y|x)$, find $\mathbb{P}(Y < 0.8|X = 0.4)$, and find $\mathbb{E}(Y|X = x)$

4.3.6 Independence for continuous random variables

Recall that two discrete random variables are independent if their joint pmf is the product of their marginal pmf's:

$$p_{X,Y}(x,y) = p_X(x)p_Y(y).$$

Similarly, two jointly continuous random variables are independent if this property holds for their densities:

$$f_{X,Y}(x,y) = f_X(x)f_Y(y).$$

More generally (for arbitrary random variables), random variables X and Y are independent if their joint cdf can be written as a product of marginal cdfs of X and Y for all x and y.

$$F_{X,Y}(x,y) = F_X(x)F_Y(y).$$

If random variables are not independent, they are called dependent.

Immediately from the definition of the conditional density it follows that If two jointly continuous r.v.'s X and Y are independent, then

$$f_{X|Y}(x|y) = f_X(x),$$

for all x and y. Intuitively, if we learn that the random variable Y has value y it does not change the probability density of the random variable X.

All properties of the independent random variables that we formulated for discrete random variables hold also for continuous random variables, although the proofs are more cumbersome.

In particular, if X and Y are two independent random variables and g and h are two functions $\mathbb{R} \to \mathbb{R}$, then

$$\mathbb{E}\Big(g(X)h(Y)\Big) = \mathbb{E}g(X)\mathbb{E}h(Y).$$

Here is a useful criterion to check if two random variables are independent. Recall that support of a function is the region where the values of this function are different from zero.

Theorem 4.3.17. *If the joint pdf for X and Y can be factored into two non-negative functions,*

$$f_{X,Y}(x,y) = g(x)h(y),$$

and the support of the joint pdf is a rectangle, then X and Y are independent.

The point of this theorem is that in this factorization the functions $g(x)$ and $h(y)$ need not be valid pdf's.

Note also that if the support of the joint pdf of X and Y is not a rectangle, then these variables are automatically not independent.

Example 4.3.18. Let the joint density be

$$f_{X,Y}(x,y) = \begin{cases} 2y, & \text{for } 0 < x < 1, 0 < y < 1, \\ 0, & \text{elsewhere.} \end{cases}$$

Are these r.v.'s independent?

The support is the square $[0,1] \times (0,1]$, and the joint density can be factored as $2 \times y$, so the random variables are independent

Example 4.3.19. Let the joint density of X and Y be

$$f_{X,Y}(x,y) = \begin{cases} 6x, & \text{for } 0 < x < y < 1, \\ 0, & \text{elsewhere.} \end{cases}$$

Are these variable independent? Here the support of the joint density is a triangle so we can immediately conclude that these random variables are not independent.

Exercise 4.3.20. Which of the following pdf's corresponds to independent random variables X and Y?

(i) $f(x,y) = 4x^2y^3$.

(ii) $f(x,y) = \frac{1}{2}(x^3y + xy^3)$.

(iii) $f(x,y) = 6e^{-3x-2y}$.

(Assume that the support of the joint density of X and Y in all cases is a rectangle chosen in such a way that these are valid densities.)

The definition of the independence can be extended in a straightforward fashion to more than two variables.

Example 4.3.21. Life-lengths (in hundreds of hours) of a population of batteries follow an exponential distribution with parameter $\beta = 30$. We take a random sample of 5 batteries and observe their life-lengths. Find the joint pdf of these 5 measurements.

Let us denote these measurements $X_1, X_2, \dots X_5$. Their joint density depends of five variables, which we denote $x_1, x_2, \dots, x_5$. Then, by definition independence we have:

$$f_{X_1,\dots,X_5}(x_1, \dots, x_5) = f_{X_1}(x_1) f_{X_2}(x_2) \dots f_{X_5}(x_5),$$

where $f_{X_i}(x_i)$ is the marginal density for a random variable X_i, $i = 1, \dots, 5$. We are given that these variables are exponential, so for each i, the marginal density is

$$f_{X_i}(x_i) = \frac{1}{\beta} e^{-x_i/\beta} = \frac{1}{30} e^{-x_i/30}.$$

So, for the joint density we get the product of these quantities:

$$f_{X_1,\dots,X_5}(x_1, \dots, x_5) = \frac{1}{30^5} e^{-(x_1+\dots+x_5)/30}.$$

4.3.7 Expectations of functions and covariance for continuous r.v.'s

The definition of covariance for continuous random variables is the same as for the discrete random variables:

$$\mathrm{Cov}(X, Y) = \mathbb{E}(XY) - (\mathbb{E}X)(\mathbb{E}Y).$$

In particular for independent random variables the covariance is zero.

However, in order to calculate the covariance, we need to know how to calculate $\mathbb{E}(XY)$. More generally, we often need to calculate the expectation of a function of two random variables X and Y. This function can be a discrete or a continuous random variable, and it is not very easy to calculate it from the definition. Without a proof we formulate the following useful result.

Theorem 4.3.22. *Let X and Y be two jointly continuous r.v.'s with joint density $f_{X,Y}(x, y)$, and let h be a function $\mathbb{R}^2 \to \mathbb{R}$. Then,*

$$\mathbb{E}h(X, Y) = \iint_{\mathbb{R}^2} h(x, y) f_{X,Y}(x, y)\, dx\, dy.$$

While the double integral in the theorem is written over the whole plane, in practice it is taken over the support of the joint density $f_{X,Y}(x,y)$.

Example 4.3.23.

$$f_{X,Y}(x,y) = \begin{cases} 3x, & \text{for } 0 < y < x < 1, \\ 0, & \text{elsewhere.} \end{cases}$$

Calculate the covariance and correlation of X and Y.

First, we calculate $\mathbb{E}(XY)$.

$$\begin{aligned} \mathbb{E}(XY) &= \iint_A xy f_{X,Y}(x,y)\,dx\,dy \\ &= \int_0^1 \int_0^x 3x^2 y\,dy\,dx \\ &= \int_0^1 \frac{3}{2}x^2 y^2\Big|_{y=0}^{y=x} dx \\ &= \int_0^1 \frac{3}{2}x^4\,dx = \frac{3}{10}x^5\Big|_0^1 \\ &= \frac{3}{10}. \end{aligned}$$

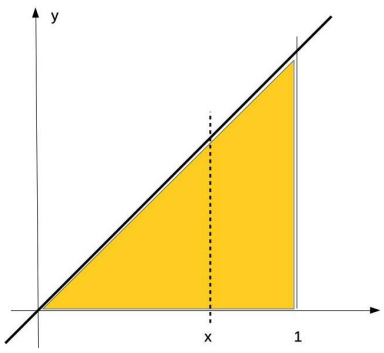


Figure 4.6

The expectation $\mathbb{E}Y$ was calculated in Example 4.3.8 as $\mathbb{E}Y = \frac{3}{8}$. For $\mathbb{E}X$, we calculate:

$$\begin{aligned} \mathbb{E}X &= \int_{-\infty}^{\infty} x f_X(x)\,dx = \int_0^1 3x^3\,dx \\ &= \frac{3}{4}x^4\Big|_0^1 = \frac{3}{4}. \end{aligned}$$

Therefore,

$$\begin{aligned} \mathrm{Cov}(X,Y) = \mathbb{E}(XY) - (\mathbb{E}X)(\mathbb{E}Y) &= \frac{3}{10} - \frac{3}{8} \times \frac{3}{4} \\ &= \frac{3}{160} = 0.01875. \end{aligned}$$

In order to calculate the correlation, we need the variances of X and Y. First we get the second moments:

$$\begin{aligned} \mathbb{E}(X^2) = \int_{-\infty}^{\infty} x^2 f_X(x)\,dx = \int_0^1 3x^4\,dx \\ = \frac{3}{5}x^5\Big|_0^1 = \frac{3}{5}, \end{aligned}$$

and

$$\begin{aligned}\mathbb{E}(Y^2) &= \int_{-\infty}^{\infty} y^2 f_Y(y)\,dy = \int_0^1 \frac{3}{2}y^2(1-y^2)\,dy \\ &= \frac{3}{2}\Big(\frac{1}{3} - \frac{1}{5}\Big) = \frac{1}{5}.\end{aligned}$$

So,

$$\begin{aligned}\mathbb{V}\text{ar}(X) &= \mathbb{E}X^2 - (\mathbb{E}X)^2 = \frac{3}{5} - \Big(\frac{3}{4}\Big)^2 = 0.0375 \\ \mathbb{V}\text{ar}(Y) &= \mathbb{E}Y^2 - (\mathbb{E}Y)^2 = \frac{1}{5} - \Big(\frac{3}{8}\Big)^2 = 0.059375\end{aligned}$$

Hence, the correlation is

$$\begin{aligned}\text{Cor}(X,Y) &= \frac{\text{Cov}(X,Y)}{\sqrt{\mathbb{V}\text{ar}(X)}\sqrt{\mathbb{V}\text{ar}(Y)}} = \frac{0.01875}{\sqrt{0.0375}\sqrt{0.059375}} \\ &= 0.3973597.\end{aligned}$$

4.3.8 Variance and conditional variance

We learned previously about the law of iterated expectations:

$$\mathbb{E}\Big[\mathbb{E}(X|Y)\Big] = \mathbb{E}X.$$

Is there a similar law for variance? The answer is "yes", although this law is more complicated.

Theorem 4.3.24 (Law of Iterated Variance)**.** *If X and Y are any r.v.'s, then*

$$\mathbb{V}\text{ar}(Y) = \mathbb{E}[\mathbb{V}\text{ar}(Y|X)] + \mathbb{V}\text{ar}[\mathbb{E}(Y|X)].$$

Proof.

$$\begin{aligned}\mathbb{E}[\mathbb{V}\text{ar}(Y|X)] + \mathbb{V}\text{ar}[\mathbb{E}(Y|X)] &= \mathbb{E}\big\{\mathbb{E}(Y^2|X) - (\mathbb{E}(Y|X))^2\big\} \\ &+ \mathbb{E}[(\mathbb{E}(Y|X))^2] - [\mathbb{E}(\mathbb{E}(Y|X))]^2 \\ &= \mathbb{E}\big\{\mathbb{E}(Y^2|X)\big\} - \Big\{\mathbb{E}(\mathbb{E}(Y|X))\Big\}^2 \\ &= \mathbb{E}(Y^2) - [\mathbb{E}(Y)]^2 \\ &= \mathbb{V}\text{ar}(Y).\end{aligned}$$

□

Example 4.3.25. Let X be the number of defective parts from an assembly line out of 20 produced per day.

Suppose the probability p of a part being defective is random. (Say, it changes from day to day.) The distribution of p is Beta$(1, 9)$.

Find $\mathbb{E}(X)$ and $\mathbb{V}\text{ar}(X)$.

The key to this problem is to recognize that we can easily calculate the expectation and variance of X *given that* p *is known* and that this is a conditional probability of X given that the random probability p takes a specific value. So we can use the law of iterated expectation and law of iterated variance.

First,

$$\mathbb{E}(X|p) = 20p \text{ and } \mathbb{V}\text{ar}(X|p) = 20p(1-p),$$

Then,

$$\mathbb{E}X = \mathbb{E}(\mathbb{E}(X|p)) = 20\mathbb{E}(p) = 20\frac{1}{1+9} = 2,$$

where we used the fact that the expectation of a beta-distributed random variable with parameters (α, β) is $\alpha/(\alpha+\beta)$.

For the variance we have the formula:

$$\mathbb{V}\text{ar}(X) = \mathbb{E}[\mathbb{V}\text{ar}(X|p)] + \mathbb{V}\text{ar}[\mathbb{E}(X|p)].$$

For the second part we can look up the formula for the variance of the beta distribution:

$$\mathbb{V}\text{ar}[\mathbb{E}(X|p)] = 20^2\mathbb{V}\text{ar}(p) = 20^2\frac{1\times 9}{(1+9)^2(1+9+1)} = \frac{36}{11}.$$

and the first part we can do by direct integration:

$$\mathbb{E}[\mathbb{V}\text{ar}(X|p)] = 20\times 9\int_0^1 p(1-p)\times(1-p)^8\,dp,$$

where we used the fact that the density of p is $9(1-p)^8$. Then we can use the formula for the beta integral and write:

$$\mathbb{E}[\mathbb{V}\text{ar}(X|p)] = 20\times 9\times\frac{\Gamma(2)\Gamma(10)}{\Gamma(10+2)} = 20\times 9\times\frac{9!}{11!} = \frac{20\times 9}{10\times 11} = \frac{18}{11}.$$

Altogether, we get

$$\mathbb{V}\mathrm{ar}(X) = \frac{18}{11} + \frac{36}{11} = \frac{54}{11} \approx 4.909...$$

(Note, by the way, that the most of the variation in X comes from the uncertainty about p. If we ignored this uncertainty and set p equal to its expected value: $p = 1/10$ then we would get the variance equal to $20 \times 0.1 \times 0.9 = 1.8$.)

4.4 The multinomial distribution

The multinomial distribution is a generalization of the binomial distribution.

Consider an experiment in which we look at a random text that contains $n = 10,000$ words and we count how many times each word occur in the text. These counts are random since they change from text to text. How we can model this situation?

Since the collection of words in one text can be different from the collection of words in another text, it is convenient to define a big dictionary of words that can occur in texts including one symbol to denote a word which is not in this dictionary: <UKN> = "Unknown word". So the dictionary is a big sequence of words. For example, we can have a dictionary of $K = 50,001$ words:

$$D = \{\text{a, aardwark, an, and, ..., zebra, <UKN>}\}.$$

Then we can identify the words with their places in this sequence. So any text of length n can be encoded by sequence of numbers $w_1, w_2, \ldots w_n$, where w_1 is the number of the first word, w_2 is the number of the second word, and so on. For example, a text that starts with: "An aardwark and a zebra ..." will be encoded as a vector $(3, 2, 4, 1, 50000, \ldots)$.

In the first, very rough approximation we can assume that the words of a text are independent. This model is called a multinomial experiment.

Often we are interested in the word counts, so let X_i be the number of times the word w_i occurs in a random text. For example, for our dictionary

D, the random variable X_1 denotes the number of times the word "a" occurs in a random text.

The joint distribution of X_i is called the multinomial distribution.

More generally we have the following description.

Multinomial experiment:

- There are n identical independent trials.
- The outcome of each trial falls into one of K categories.
- The probability that the outcome falls into category i is p_i (for $i = 1, 2, \ldots, K$), where $p_1 + p_2 + \ldots + p_K = 1$, and this set of probabilities is the same across trials.

The multinomial random vector is $X = (X_1, X_2, \ldots, X_K)$, where X_i is the number of trials resulting in category i.

Note that $X_1 + X_2 + \ldots + X_K = n$, the total number of trials.

In our example above we had $n = 10,000$ since our random texts contained $10,000$ words and $K = 50,001$ since our dictionary consisted of $50,001$ words (including <UKN>). The vector $(p_1, p_2, \ldots, p_{50,001})$ can be estimated from a data. For example, p_1 can be approximated by the frequency of the word "a" in a large number of texts.

Theorem 4.4.1. *The multinomial random vector $X = (X_1, X_2, \ldots, X_K)$ has the joint pmf*

$$p_X(x_1, x_2, \ldots, x_K) = \binom{n}{x_1, x_2, \ldots, x_K} p_1^{x_1} p_2^{x_2} \ldots p_K^{x_K},$$

where each x_i is an integer between 0 and n, and $\sum_{i=1}^{K} x_i = n$.

This probability distribution is called the multinomial distribution with parameters n and $p = (p_1, p_2, \ldots, p_K)$.

Note that binomial distribution case is a particular case of the multinomial. In this case the vector (X_1, X_2) consists of two components X_1 and X_2, that correspond to successes and failures, and the parameter vector is (p_1, p_2) where $p_1 = p$ and $p_2 = 1 - p$. We usually do not think about the

binomial random variable as multivariate random variable because the second component is completely determined by the first one: $X_2 = n - X_1$. Similarly, the multinomial random variable with k components essentially has $k-1$ "free" components. The last one is determined by the remaining: $X_k = n - (X_1 + \dots X_{k-1})$.

Example 4.4.2. Three card players play a series of matches. The probability that player A will win any game is 20%, the probability that player B will win is 30%, and the probability player C will win is 50%.

If they play 6 games, what is the probability that player A will win 1 game, player B will win 2 games, and player C will win 3?

If X_A, X_B, and X_C denote the number of wins by players A, B, and C, respectively, and if p_A, p_B, p_C are the probabilities to win a single game than we have:

$$\begin{aligned}\mathbb{P}(X_A = 1, X_B = 2, X_C = 3) &= \binom{6}{1,2,3}(p_A)(p_B)^2(p_C)^3 \\ &= \binom{6}{1,2,3}(0.2)(0.3)^2(0.5)^3 = 0.135\end{aligned}$$

Theorem 4.4.3. *If* $(X_1, \dots, X_n) \sim \text{Multinom}(n, p_1, \dots, p_k)$, *then*

1. $\mathbb{E}(X_i) = np_i$.
2. $\mathbb{V}\text{ar}(X_i) = np_i(1-p_i)$.
3. $\text{Cov}(X_i, X_j) = -np_ip_j$, *if* $i \neq j$.

Proof. Claims (1) and (2) follow because the marginal pmf of X_i is binomial with parameters n, p_i. (We can think about the reduced experiment: if we observe outcome i we count it as success and all other outcomes as failure. So the count of successes is X_i and this is obviously a binomial experiment with parameters n and p_i.)

In order to prove (3), assume that $i \neq j$ and define some indicator random variables. First, let $I_s = 1$ if the s-th trial results in outcome i, and $I_s = 0$ otherwise. Second, let $J_s = 1$ if the s-th trial results in outcome j, and $J_s = 0$ otherwise.

Then, as usual, we can use these indicator variables as counters an we have

$$X_i = I_1 + I_2 + \ldots + I_n.$$

$$X_j = J_1 + J_2 + \ldots + J_n.$$

In order to calculate covariance of X_i and X_j we compute:

$$\mathbb{E}(X_i X_j) = \mathbb{E}\Big[\Big(\sum_{s=1}^{n} I_s\Big)\Big(\sum_{t=1}^{n} J_t\Big)\Big] = \sum_{s,t} \mathbb{E}(I_s J_t),$$

where we expanded the product of the sums as one big sum over all possible combinations of s and t and used the linearity of the expectation.

First, suppose that $s \neq t$. Then, $\mathbb{E}(I_s J_t) = \mathbb{P}(I_s J_t = 1) = \mathbb{P}(I_s = 1, J_t = 1) = p_i p_j$. Second, if $s = t$ then $\mathbb{E}(I_s I_t) = 0$ since I_s and J_s cannot be both simultaneously 1.

Next, note that there are exactly $n(n-1)$ terms in the sum $\sum_{s,t} \mathbb{E}(I_s J_t)$ with $s \neq t$, and therefore,

$$\mathbb{E}(X_i X_j) = n(n-1) p_i p_j.$$

So we calculate:

$$\begin{aligned} \mathrm{Cov}(X_i, X_j) &= \mathbb{E}(X_i X_j) - \mathbb{E}(X_i)\mathbb{E}(X_j) \\ &= n(n-1) p_i p_j - n^2 p_i p_j = -n p_i p_j. \end{aligned}$$

□

One can also compute the conditional pmf for multinomial distribution. The main take-away from this calculation (which we skip) is that this conditional pmf is also a multinomial distribution with easily computable parameters.

Theorem 4.4.4. *Let* $(X_1, \ldots, X_K) \sim \mathrm{Multinom}(n, p_1, \ldots, p_K)$. *Suppose* $x_K \in \{0, \ldots, n\}$. *Then, the conditional joint pmf of random variables* $X_1, \ldots, X_{K-1}$ *given that* $X_K = x_K$ *is a multinomial pmf with parameters*

$$(n - x_K, \frac{p_1}{p_1 + \ldots + p_{K-1}}, \ldots, \frac{p_K}{p_1 + \ldots + p_{K-1}}).$$

So the sum of $X_1, \ldots, X_K$ is $n - x_K$ and the probabilities of X_i are simply rescaled from the original values so that their sum still equals 1.

Another useful observation is that the sum of a fixed set of X_i's have a binomial distribution with the parameter p equal to the sum of the parameters of these random variables.

For example, a random variable $Y = X_1 + X_3 + X_5$ has the binomial distribution with parameters n (same as for the multivariate distribution) and parameter $p = p_1 + p_3 + p_5$.

4.5 The multivariate normal distribution

Recall that the density of a normal random variable X with mean μ and variance σ^2 is

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(- \frac{(x-\mu)^2}{2\sigma^2}\Big)$$

If we have two *independent* normal random variables X and Y with parameters (μ_X, σ_X^2) and (μ_Y, σ_Y^2), then their joint density is simply the product of marginal densities:

$$\begin{aligned} f_{X,Y}(x,y) &= \frac{1}{\sqrt{2\pi\sigma_X^2}} \exp\Big(- \frac{(x-\mu_X)^2}{2\sigma_X^2}\Big) \times \frac{1}{\sqrt{2\pi\sigma_Y^2}} \exp\Big(- \frac{(y-\mu_Y)^2}{2\sigma_Y^2}\Big) \\ &= \frac{1}{2\pi\sigma_X\sigma_Y} \exp\Big(- \frac{(x-\mu_X)^2}{2\sigma_X^2} - \frac{(y-\mu_Y)^2}{2\sigma_Y^2}\Big). \end{aligned}$$

An extremely useful property of normal random variables is that if we form their linear combination:

$$U = aX + bY,$$

then it is also a normal random variable. We will prove it in the next chapter.

Moreover if we have two linear combinations U, and, say,

$$V = cX + dY,$$

then the joint density of these random variables also has a simple form:

$$f_{U,V}(u,v) = \frac{1}{2\pi c_1} \exp\Big(- \frac{(u-\mu_U)^2}{2c_2} - \frac{(v-\mu_V)^2}{2c_3} + 2\frac{(u-\mu_U)(v-\mu_V)}{2c_4}\Big),$$

where μ_U and μ_V are expected values of U and V and the constants c_1, c_2, c_3 and c_4 can be calculated in terms of variances of U and V and covariance of U and V.

This density function is called the density of the bivariate normal distribution. We skip the calculation of the constants $c_1, \ldots, c_4$ and present the final result. (We also call the random variable X and Y instead of U and V.)

Definition 4.5.1. The random variables X and Y have the bi-variate distribution with means μ_X and μ_Y, variances σ_X^2, σ_Y^2, and correlation ρ if its probability density is

$$f(x,y) = \frac{1}{(2\pi)\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\Big(- \frac{1}{2}Q(x,y)\Big),$$

where

$$Q = \frac{1}{1-\rho^2}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} - 2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right].$$

Basic properties of the bivariate Gaussian.

First, one can check that $f(x,y)$ in the definition is a valid density function. Second, the names for the parameters are justified, that is, one can verify by direct calculation the following result.

Theorem 4.5.2. *If X and Y have the bi-variate distribution with parameters μ_X, μ_Y, σ_X^2, σ_Y^2, and ρ, then*

- $\mathbb{E}X = \mu_X$ *and* $\mathbb{E}Y = \mu_Y$.
- $\mathbb{V}\mathrm{ar}(X) = \sigma_X^2$ *and* $\mathbb{V}\mathrm{ar}(Y) = \sigma_Y^2$.
- $\mathrm{Cov}(X,Y) = \rho\sigma_X\sigma_Y$, *and* $Corr(X,Y) = \rho$.

In addition we have the following results about marginal and conditional distributions.

Theorem 4.5.3. *If X and Y have the bi-variate distribution with parameters μ_X, μ_Y, σ_X^2, σ_Y^2, and ρ, then*

- *Marginal distributions: $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$.*
- *The conditional density of Y given that $X = x$ is normal:*

$$\mathcal{N}\left(\mu_Y + \rho\sigma_Y \frac{x - \mu_X}{\sigma_X}, (1-\rho^2)\sigma_Y^2\right).$$

Example 4.5.4. Suppose (X, Y) is a bi-variate normal vector with $\mu_X = \mu_Y = 0$, $\sigma_X = \sigma_Y = 2$, $\rho = 1/2$. Find $\mathbb{P}(Y > 0 | X = 1)$.

The conditional density of Y given that $X = 1$ is normal with mean

$$\mu_Y + \rho\sigma_Y \frac{x - \mu_X}{\sigma_X} = 0 + \frac{1}{2} \times 2 \times \frac{1-0}{2} = \frac{1}{2}$$

and variance

$$(1-\rho^2)\sigma_Y^2 = \left(1 - (\frac{1}{2})^2\right) \times 2^2 = 3$$

Hence,

$$\begin{aligned}\mathbb{P}(Y > 0 | X = 1) &= \mathbb{P}\Big(\frac{Y - 1/2}{\sqrt{3}} > \frac{-1/2}{\sqrt{3}} | X = 1\Big) \\ &= \mathbb{P}\Big(Z > \frac{-1/2}{\sqrt{3}}\Big) \approx 61.4\%,\end{aligned}$$

where Z denotes a standard normal random variable.

You can see that the formulas are rather lengthy. So, if we want to generalize these results to the case of more than two normal random variables, the language of linear algebra and matrices is indispensable. Let us reformulate the definition of the bi-variate random variables using matrices.

Let $\vec{\mu} = [\mu_X, \mu_Y]^t$ be a column vector with components μ_X and μ_Y. (The superscript t means that this is a transposed row vector $[\mu_X, \mu_Y]$.) Also let us define the *covariance matrix*:

$$\Sigma = \begin{bmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{bmatrix},$$

where $\sigma_X^2 > 0$, $\sigma_Y^2 > 0$, and $\sigma_{XY} = \rho\sigma_X\sigma_Y$, where the correlation coefficient $|\rho| \leq 1$.

Recall that for any two-by-two matrix

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix},$$

one can define the determinant:

$$\det(A) = ad - bc,$$

and the inverse matrix:

$$A^{-1} = \frac{1}{\det(A)} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix},$$

Using these notation we say that a random vector $\vec{W} = (X, Y)$ has the bi-variate normal (or Gaussian) distribution with parameters $\vec{\mu}$ and Σ, if its density function is

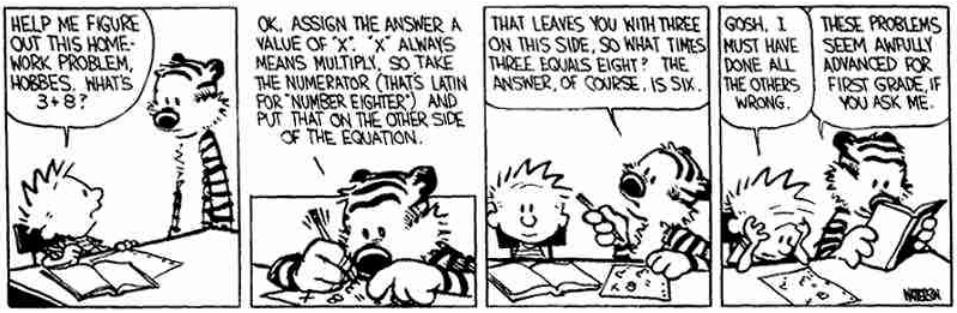
$$f_W(\vec{w}) = \frac{1}{(2\pi)\sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(\vec{w} - \vec{\mu})^t \Sigma^{-1} (\vec{w} - \vec{\mu})\right), \qquad (4.3)$$

where $\vec{w} = (x, y)^t$ and Σ^{-1} is the inverse of the matrix Σ.

Notation: $W \sim N(\mu, \Sigma)$.

The benefit of this definition is that it can be generalized in a straightforward fashion to the cases when the vector W has not two but much larger number of components. Say, if vector $W = (X_1, \ldots, X_{100})$ has 100 components, then the formula (4.3) is still valid except that we need to change the constant (2π) to $(2\pi)^{50}$, that the vector of expected values μ has 100 components and that the matrix Σ is 100-by-100 matrix.

This makes the manipulation of formulas much easier and the actual matrix calculations can be done by a computer.


HELP ME FIGURE OUT THIS HOME-WORK PROBLEM, HOBBES. WHAT'S 3+8?
OK, ASSIGN THE ANSWER A VALUE OF "X". "X" ALWAYS MEANS MULTIPLY, SO TAKE THE NUMERATOR (THAT'S LATIN FOR "NUMBER EIGHTER") AND PUT THAT ON THE OTHER SIDE OF THE EQUATION.
THAT LEAVES YOU WITH THREE ON THIS SIDE, SO WHAT TIMES THREE EQUALS EIGHT? THE ANSWER, OF COURSE, IS SIX.
GOSH, I MUST HAVE DONE ALL THE OTHERS WRONG.
THESE PROBLEMS SEEM AWFULLY ADVANCED FOR FIRST GRADE, IF YOU ASK ME.

Chapter 5

Functions of Random Variables

We are often interested in a function of one or several random variables, $Y = g(X_1, \ldots, X_n)$. For example, in statistics, one is often interested in the distribution of sample average and sample variance of data which are defined as

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

It is assumed that the observed random variables, $X_1, \ldots, X_n$ are independent and have the same distribution. Then $\overline{X}$ and s_X^2 are used to estimate the expectation and variance of thee distribution of X_i. Since $\overline{X}$ and s_X^2 are random, we are interested to know what is their distribution as random variables.

In general there are three methods for determining the distribution of a function of random variables:

1. The method of cdf's
2. The method of pdf's
3. The method of mgf's

We start by learning how these methods can be used when we have the function of only one random variable, $Y = g(X)$ and then study how they can be extended to functions of several random variables.

5.1 The method of cumulative distribution functions

5.1.1 Functions of one random variable

Let U be a function of a continuous r.v.'s X, $U = g(X)$ and let the density of X be $f_X(x)$. Then we can find the cdf of U by the definition of the cdf.

$$F_U(u) = \mathbb{P}(U \leq u) = \mathbb{P}(g(X) \leq u) = \int_A f_X(x)\,dx$$

where A is the region on the real line defined by the inequality $g(x) \leq u$.

Then, if desired, we can find the density of U by differentiating the cdf:

$$f_U(u) = F_U'(u).$$

Example 5.1.1. Let X be the amount of sugar produced per day (in tons). Suppose

$$f_X(x) = \begin{cases} 2x, & \text{for } 0 < x < 1, \\ 0, & \text{elsewhere.} \end{cases}$$

Let the profit $U = 3X - 1$ (in hundreds of \$). Find the probability density of U.

First, we calculate $F_U(u)$:

$$F_U(u) = \mathbb{P}(3X - 1 \leq u) = \mathbb{P}\big(X \leq (u+1)/3\big) = F_X\Big((u+1)/3\Big)$$

Note that the cdf of X is

$$F_X(x) = \begin{cases} 0, & \text{if } x < 0, \\ \int_0^x 2t\,dt = x^2 & \text{if } x \in [0,1], \\ 1, & \text{if } x > 1. \end{cases}$$

So if $(u+1)/3 < 0$ then $F_U(u) = 0$. This happens if $u < -1$.

If $(u+1)/3 > 1$ then $F_U(u) = 1$. This happens if $u > 2$.

In the intermediate cases, if $(u+1)/3$ between 0 and 1, we have $F_U(u) = \Big((u+1)/3\Big)^2$. So, altogether,

$$F_U(u) = \begin{cases} 0, & \text{if } u < -1, \\ \frac{(u+1)^2}{9} & \text{if } u \in [-1, 2], \\ 1, & \text{if } u > 2. \end{cases}$$

In order to get density, we simply differentiate this expression and get:

$$f_U(u) = \begin{cases} \frac{2(u+1)}{9} & \text{if } u \in [-1, 2], \\ 0, & \text{otherwise.} \end{cases}$$

Example 5.1.2. Let X have the pdf:

$$f_X(x) = \begin{cases} 6x(1-x), & \text{for } 0 < x < 1, \\ 0, & \text{elsewhere.} \end{cases}$$

Find the pdf of $U = X^3$.

Again we start with finding the cdf of U:

$$F_U(u) = \mathbb{P}(X^3 < u) = \mathbb{P}(X < u^{1/3}) = F_X(u^{1/3})$$

The cdf of X, $F_X(x)$ equals to 0 if $x < 0$ and 1 if $x > 1$. If x is between 0 and 1, then

$$F_X(x) = \int_0^x 6t(1-t)\,dt = 3x^2 - 2x^3.$$

So for $u^{1/3} \in [0, 1]$, which is the same as $u \in [0, 1]$, we have

$$F_U(u) = 3(u^{1/3})^2 - 2(u^{1/3})^3 = 3u^{2/3} - 2u.$$

So after the differentiating, we recover the density:

$$f_U(u) = \begin{cases} 2u^{-1/3} - 2, & \text{if } u \in [0, 1], \\ 0, & \text{otherwise.} \end{cases}$$

Exercise 5.1.3. Let X be an exponential variable with parameter β. Let $U = 1/X$. Find the density of U.

5.1.2 Functions of several random variables

Let $U = g(X, Y)$ be a function of r.v.'s X and Y that have the joint density $f_{X,Y}(x, y)$.

The formula for the cdf is almost the same as before:

$$F_U(u) = \mathbb{P}(U \le u) = \mathbb{P}(g(X, Y) \le u) = \int_A f_{X,Y}(x, y)\, dx\, dy,$$

where A is the region in $\mathbb{R}^2$ defined by the inequality $g(x, y) \le u$.

So the method is as follows.

- Find the region of values (x, y) such that $g(x, y) \le u$.
- Integrate $f_{X,Y}(x, y)$ over this region to obtain $\mathbb{P}(U \le u) = F_U(u)$.
- Differentiate $F_U(u)$ to obtain $f_U(u)$.

Example 5.1.4. Let X be the amount of gasoline stocked at the beginning of the week. Let Y be the amount of gasoline sold during the week. The joint density of X and Y is

$$f_{X,Y}(x, y) = \begin{cases} 3x, & \text{for } 0 \le y \le x \le 1, \\ 0, & \text{elsewhere.} \end{cases}$$

Find the density of $U = X - Y$, the amount of gasoline remaining at the end of the week.

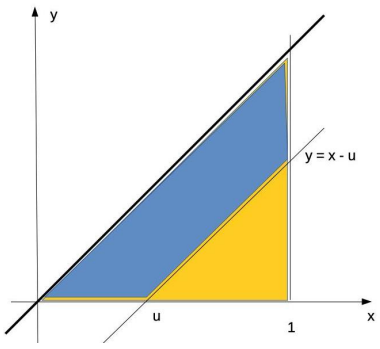


Figure 5.1

To calculate $F_U(u)$ we need to integrate the joint density over the region $\{(x, y) : x - y \le u\}$ or, in other terms $\{(x, y) : y \ge x - u\}$. It is shown as a blue trapeze in the picture. It is doable, but it is easier to setup the integral over the remaining yellow triangle, which gives $1 - F_U(u)$.

$$1 - F_U(u) = \int_u^1 \int_0^{x-u} 3x\, dy\, dx$$

We calculate this as

$$\begin{aligned}\int_u^1 \int_0^{x-u} 3x\,dy\,dx &= \int_u^1 3x(x-u)\,dx \\ &= \left(x^3 - \frac{3}{2}x^2u\right)\Big|_{x=u}^{x=1} \\ &= 1 - \frac{3}{2}u - \left(u^3 - \frac{3}{2}u^3\right) \\ &= 1 - \frac{3}{2}u + \frac{1}{2}u^3.\end{aligned}$$

Hence,

$$F_U(u) = \frac{3}{2}u - \frac{1}{2}u^3.$$

Hence $f_U(u) = F_U'(u) = \frac{3}{2}(1-u^2)$. Note that this calculation is only valid for u between 0 and 1. If u is outside of this interval then the density is zero. (It is impossible that U took value outside of the interval $[0,1]$.) So, the final answer is

$$f_U(u) = \begin{cases} \frac{3}{2}(1-u^2), & \text{if } u \in [0,1], \\ 0, & \text{otherwise.} \end{cases}$$

Example 5.1.5. Suppose the joint density of X and Y is

$$f_{X,Y}(x,y) = \begin{cases} 6e^{-3x-2y}, & \text{for } 0 \le x, 0 \le y, \\ 0, & \text{elsewhere.} \end{cases}$$

Find the pdf of $U = X + Y$.

Example 5.1.6. Suppose the joint density of X and Y is

$$f(x,y) = \begin{cases} 5xe^{-xy}, & \text{for } \frac{1}{5} \le x \le \frac{2}{5}, 0 \le y, \\ 0, & \text{elsewhere.} \end{cases}$$

Find the pdf of $U = XY$.

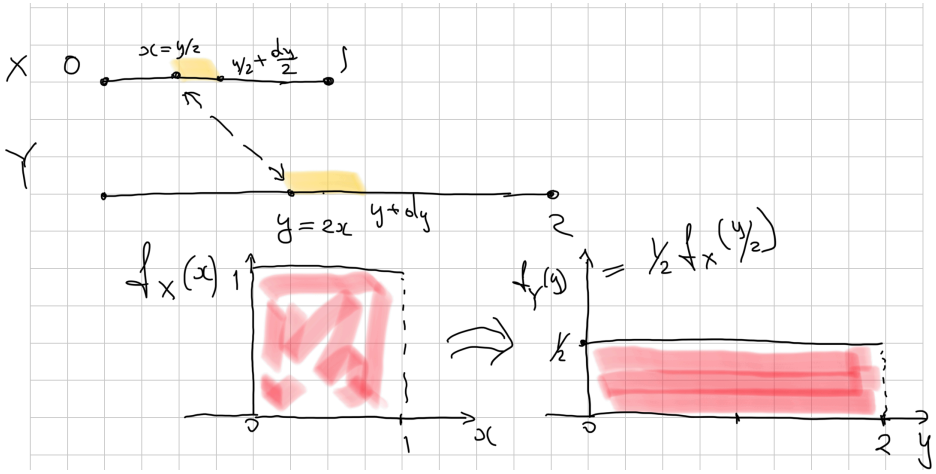


Figure 5.2: The idea of the pdf transformation method

5.2 The pdf transformation method

The pdf method is best to explain with a simple example.

Suppose that X is uniformly distributed on the interval $[0, 1]$ and we are interested in calculating the density of $Y = 2X$.

Then, intuitively, $f_Y(y)\,dy$ is the probability that Y takes its value in the interval $[y, y + dy]$. Since $Y = 2X$, this probability equals the probability that X takes its value in the interval $[y/2, y/2 + dy/2]$, which equals $f_X(y/2)\,(dy/2) = \frac{1}{2} f_X(y/2)dy$. So, we conclude that $f_Y(y) = \frac{1}{2} f_X(y/2)$, and the density of Y is constant and is equal to $1/2$ on the interval $[0, 2]$.

In other words, the density of Y at point y equals the density of X at the point $x = y/2$ (which corresponds to y under transformation $x \to y = 2x$) divided by the amount of the stretch that transformation $x \to 2x$ applies to the region around the point $x = y/2$.

This is the main idea of the pdf transformation method.

5.2.1 A function of one random variable

Let $U = g(X)$ where $g(x)$ is a one-to-one function (for example, a strictly increasing or a strictly decreasing function). And let $g^{(-1)}(u)$ is the inverse

of the function $g(x)$. That is,

$$g^{(-1)}\big(g(x)\big) = x.$$

Intuitively, $g(x)$ maps x to u and $g^{(-1)}(u)$ maps u back to x.

Assume that $g(x)$ is increasing and let us start as in the cdf method:

$$\begin{aligned} F_U(u) &= \mathbb{P}\big(g(X) \le u\big) = \mathbb{P}\big(X \le g^{(-1)}(u)\big) \\ &= F_X\big(g^{(-1)}(u)\big). \end{aligned}$$

Then by differentiating and using the chain rule we find that

$$f_U(u) = f_X\big(g^{-1}(u)\big)\frac{d}{du}g^{(-1)}(u).$$

If $g(x)$ is decreasing then the calculation is slightly different.

$$\begin{aligned} F_U(u) &= \mathbb{P}\big(g(X) \le u\big) = \mathbb{P}\big(X \ge g^{-1}(u)\big) \\ &= 1 - F_X\big(g^{-1}(u)\big), \end{aligned}$$

and by differentiating,

$$f_U(u) = -f_X\big(g^{-1}(u)\big)\frac{d}{du}g^{(-1)}(u)$$

These two formulas can be written as a single formula:

$$f_U(u) = f_X\big(g^{-1}(u)\big)\Big|\frac{d}{du}g^{(-1)}(u)\Big|$$

It is important also to determine the range of the random variable U. It is simply the image of the range of the random variable X under the transformation $g(x)$:

$$\text{Range}(U) = g\big(\text{Range}(X)\big).$$

Let us redo the examples that we considered earlier to convince ourselves that the pdf method is easier than the cdf method.

Example 5.2.1. Let X be the amount of sugar produced per day (in tons). Suppose

$$f_X(x) = \begin{cases} 2x, & \text{for } 0 < x < 1, \\ 0, & \text{elsewhere.} \end{cases}$$

Let the profit $U = 3X - 1$ (in hundreds of \$).

Find the density of U.

The transformation is $u = 3x - 1$, the inverse transformation is $x = (u+1)/3$. The transformation maps the range of X, $[0, 1]$, to the interval $[-1, 2]$. So by using the formula, we get

$$f_U(u) = 2\frac{u+1}{3}\frac{1}{3} = \frac{2}{9}(u+1), \text{ if } x \in [-1, 2].$$

Example 5.2.2. Let X have the pdf:

$$f_X(x) = \begin{cases} 6x(1-x), & \text{for } 0 < x < 1, \\ 0, & \text{elsewhere.} \end{cases}$$

Find the pdf of $U = X^3$.

The transformation is $u = x^3$ and the inverse transformation is $x = u^{1/3}$. So by using the formula, we get

$$f_U(u) = 6u^{1/3}(1 - u^{1/3}) \times \frac{1}{3}u^{-2/3} = 2\big(u^{-1/3} - 1\big).$$

The transformation $u = x^3$ sends the interval $[0, 1]$ (the range of X) to itself, so the final answer is

$$f_U(u) = \begin{cases} 2u^{-1/3} - 2, & \text{if } u \in [0, 1], \\ 0, & \text{otherwise.} \end{cases}$$

Example 5.2.3. Let X be a standard exponential r.v. Find the pdf of $U = \sqrt{X}$.

Here $X = U^2$, and we have

$$f_U(u) = f_X(u^2) \times 2u = 2ue^{-u^2}, \text{ if } u > 0.$$

Exercise 5.2.4 (Cauchy distribution). Let $U \sim unif(-\pi/2, \pi/2)$. Find the pdf of $X = tan(U)$.

One should be careful about the cases when the function $g : X \to Y$ is not one-to-one. Say, we want to calculate $f_Y(y)$ and the point y corresponds to two points x_1 and x_2. That is, we have both $g(x_1) = y$ and $g(x_2) = y$.

Then we have two inverse functions $g_1^{(-1)}$ and $g_2^{(-1)}$ such that $g_1^{(-1)}(y) = x_1$ and $g_2^{(-2)}(y) = x_2$ and they will both contribute to the density of Y. In this case, we have

$$f_Y(y) = f_X(g_1^{(-1)}(y)) \times \left|\frac{d}{dy} g_1^{(-1)}(y)\right| + f_X(g_2^{(-1)}(y)) \times \left|\frac{d}{dy} g_2^{(-1)}(y)\right|.$$

Example 5.2.5. Let X be uniformly distributed on $[-1, 1]$ and $Y = X^2$. Find the density of X.

Here we have $x_1 = \sqrt{y}$ and $x_2 = -\sqrt{y}$. Also note that $f_X(x) = 1/2$ on the interval $[-1, 1]$, and that the function g maps this interval to the interval $[0, 1]$. Therefore

$$f_Y(y) = \frac{1}{2} \times \left|\frac{d}{dy}\sqrt{y}\right| + \frac{1}{2} \times \left|\frac{d}{dy}(-\sqrt{y})\right| = \frac{1}{2} y^{-1/2} \text{ if } y \in [0, 1].$$

5.2.2 Functions of several variables

The density transformation method can also be extended to functions of several random variables. The difficulty is that this method is inherently designed for one-to-one functions. Indeed, if we have $U = g(X, Y)$, then to find the density of U at a point u by using the density transformation method, we need to find the pair (x, y) which is mapped to u, take the density $f_{X,Y}(x, y)$ at this point and adjust it for the stretching of the space introduced by transformation g. The difficulty is that there is an infinite number of these pairs (x, y) , so we should somehow integrate the density over all these pairs.

The way out of this difficulty is to introduce another function $V = g_2(X, Y)$, find the joint density of U and V and then find the marginal density of U by integrating out V.

So let us consider the case of two random variables X and Y which are mapped to two random variables U and V in a one-to-one fashion.

The transformation $g : (x, y) \to (u, v)$ is given by functions

$$u = g_1(x, y)$$
$$v = g_2(x, y),$$

where the transformation is one-to-one.

Let the inverse transformation be denoted $h : (u, v) \to (x, y)$ with components:

$$x = h_1(u, v)$$
$$y = h_2(u, v),$$

We assume that random variables X and Y have the joint density $f_{X,Y}(x, y)$, and the task is to find the joint pdf of (U, V).

We have the following theorem, which is complete analogue of the theorem for the univariate case: the joint density of (U, V) at a point (u, v) equals to the joint density of (X, Y) at the corresponding point (x, y) multiplied by a factor that measures the stretching. This factor equals to the Jacobian of the inverse transformation $h(u, v)$.

Theorem 5.2.6.

$$f_{U,V}(u, v) = f_{X,Y}(h_1(u, v), h_2(u, v)) \left| \frac{\partial h(u, v)}{\partial (u, v)} \right|$$

where

$$\frac{\partial h(u, v)}{\partial (u, v)} = \det \begin{bmatrix} \frac{\partial h_1(u,v)}{\partial u} & \frac{\partial h_1(u,v}{\partial v} \\ \frac{\partial h_2(u,v)}{\partial u} & \frac{\partial h_2(u,v)}{\partial v} \end{bmatrix}$$

is the Jacobian of the transformation h.

Example 5.2.7. Let X be the amount of gasoline stocked at the beginning of the week. Let Y be the amount of gasoline sold during the week. The joint density of X and Y is

$$f_{X,Y}(x, y) = \begin{cases} 3x, & \text{for } 0 \le y \le x \le 1, \\ 0, & \text{elsewhere.} \end{cases}$$

Find the density of $U = X - Y$, the amount of gasoline remaining at the end of the week.

In order to use the pdf method, we add another variable $V = Y$. Then the transformation g has components

$$u = x - y,$$
$$v = y,$$

and the inverse transformation h has components:

$$x = u + v,$$
$$y = v.$$

In particular the transformation g is one-to-one and so we can use the pdf method.

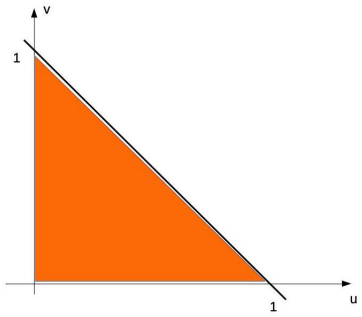


Figure 5.3

Next we need to determine what is the range of the random vector (U, V). The range of (X, Y) is the triangle with vertices $(0, 0)$, $(1, 0)$, $(1, 1)$. The transformation g is linear and therefore this triangle is mapped to a triangle with vertices $g(0, 0) = (0, 0)$, $g(1, 0) = (1, 0)$ and $g(1, 1) = (0, 1)$. The resulting triangle is the range of (U, V). On this range we can calculate the density by applying the formula above. The Jacobian of the inverse transformation is

$$J = \det \begin{bmatrix} \frac{\partial}{\partial u}(u+v) & \frac{\partial}{\partial v}(u+v) \\ \frac{\partial}{\partial u}v & \frac{\partial}{\partial v}v \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} = 1,$$

and therefore

$$f_{U,V}(u, v) = 3(u + v)|J| = 3(u + v).$$

We are interested not in the joint density of U and V but in the marginal density of U, which we calculate as

$$f_U(u) = \int_0^{1-u} 3(u+v)\,dv = 3\big(uv + \frac{1}{2}v^2\big)\Big|_{v=0}^{v=1-u}$$
$$= 3\Big(u(1-u) + \frac{1}{2}(1-2u+u^2)\Big) = \frac{3}{2}(1-u^2).$$

So the answer is

$$f_U(u) = \begin{cases} \frac{3}{2}(1-u^2), & \text{if } u \in [0,1], \\ 0, & \text{otherwise.} \end{cases}$$

Example 5.2.8. Suppose X and Y are r.v.'s with joint density

$$f_{X,Y}(x,y) = \begin{cases} e^{-(x+y)}, & \text{if } x>0, y>0, \\ 0, & \text{elsewhere.} \end{cases}$$

(These are two independent exponential random variables.) Find the joint density of $U = X+Y$ and $V = \frac{X}{X+Y}$. and then find the marginal densities of U and V.

One of the important issues is to find the range of the transformed random variables U and V. In order to do this, we look at the boundary of the range of r.v.'s X and Y. It consists of two lines: $x = 0, y \geq 0$ and $y = 0, x \geq 0$. It is useful to think about these lines as curves with a parametric representation. For the first line, the parametric representation is $x = 0, y = t$, where t changes from 0 to ∞. It is mapped to the line $u = x+y = t$ and $v = x/(x+y) = 0$, which we can identify as the horizontal line.

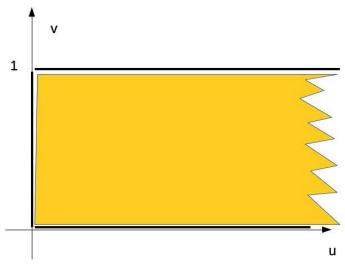


Figure 5.4

The second boundary line is $x = s, y = 0$, where s changes from 0 to ∞. It is mapped to the line $u = x+y = s$ and $v = x/(x+y) = 1$. This is another horizontal line. These two lines do not touch each other. It turns out that there is another piece on the boundary of the range of (U,V) which is the interval $u = 0, v \in [0,1]$.

So the range looks as in the picture.

The direct transformation is $u = x + y$, $v = x/(x + y)$, so for the inverse transformation we get

$$\begin{aligned} x &= uv, \\ y &= u - x = u - uv. \end{aligned}$$

The Jacobian of the inverse transformation is

$$J = \det \begin{bmatrix} v & u \\ 1 - v & -u \end{bmatrix} = -uv - u(1 - v) = -u.$$

Hence the joint density is

$$\begin{aligned} f_{U,V}(u, v) &= \begin{cases} ue^{-(uv+u-uv)}, & \text{if } u > 0, 0 < v < 1, \\ 0, & \text{elsewhere.} \end{cases} \\ &= \begin{cases} ue^{-u}, & \text{if } u > 0, 0 < v < 1, \\ 0, & \text{elsewhere.} \end{cases} \end{aligned}$$

For the marginal density of U, we have:

$$f_U(u) = \int_0^1 ue^{-u} dv = ue^{-u}, \text{ if } u \geq 0.$$

This means that the distribution of the sum of two standard exponential random variables has the gamma distribution with parameter $\alpha = 2$.

In order to find the marginal density of V, we integrate the joint density over u:

$$f_V(v) = \int_0^\infty ue^{-u}\, du = \Gamma(2) = 1! = 1,$$

and this is valid for every $v \in [0, 1]$. Hence,

$$f_V(v) = \begin{cases} 1, & \text{if } 0 < v < 1, \\ 0, & \text{otherwise.} \end{cases}$$

In other words, V has the uniform distribution between 0 and 1.

Exercise 5.2.9. Let the density of X and Y be

$$f_{X,Y}(x,y) = \begin{cases} 4x(1-y), & 0 < y < x < 1, \\ 0, \text{ otherwise.} \end{cases}$$

Let $U = Y/X$ and $V = 2X$.

Find the joint density of U and V.

Exercise 5.2.10. Suppose X and Y are r.v.'s with joint density

$$f_{X,Y}(x,y) = \begin{cases} 2(1-x), & \text{for } 0 < x < 1, 0 < y < 1, \\ 0, & \text{elsewhere.} \end{cases}$$

Find the density of $U = XY$. [Use an auxiliary variable $V = X$.]

5.2.3 Density for a sum of two random variables

A very frequent problem is to calculate the density for the sum of several independent random variables. In this section, we use the pdf transformation method to solve this problem for two random variables and in the next section we will see how to solve it by using the moment generating method.

In fact let us look at a more general problem. Suppose X and Y have the p.d.f. $f_{X,Y}(x,y)$ and we wish to calculate the p.d.f. of $U = X + Y$.

Let $V = Y$. Then $X = U - V$ and $Y = V$. Then we calculate the Jacobian of the inverse transformation as

$$\det \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} = 1$$

and so, the joint distribution of U and V is $f_{U,V}(u,v) = f_{X,Y}(u-v, v)$.

The marginal density of U is

$$f_U(u) = \int_{-\infty}^{\infty} f_{X,Y}(u-v, v)\, dv = \int_{-\infty}^{\infty} f_{X,Y}(t, u-t)\, dt.$$

(The last equality is by change of variable.)

If the random variables X and Y are independent with densities $f_X(x)$ and $f_Y(y)$, then the joint density is $f_X(x)f_Y(y)$ and so we get the following

formula for the marginal density of $U = X + Y$:

$$f_U(u) = \int_{-\infty}^{\infty} f_X(t) f_Y(u-t)\, dt.$$

This operation is called the *convolution* of functions f_X and f_Y.

Remark: In these formulas we wrote the integrals from $-\infty$ to ∞ and disregarded the ranges of X and Y. The reason is that the ranges are needed only if we need to calculate the integrals as explicit analytical formulas.

We can leave the integrals without specifying ranges if we do not intend to calculate the integrals explicitly, or if we do the calculation numerically. In the latter case, the functions $f_X(t)$ and $f_Y(t)$ will be calculated by numerical procedures and the integrals will be calculated approximately, for example by calculating Riemann sums.

Exercise 5.2.11. Let X and Y are two independent variables both distributed uniformly on the interval $[0, 1]$. Calculate the density of $U = X + Y$.

5.3 Method of moment-generating functions

Suppose $U = U(Y_1, \ldots, Y_n)$. The idea of the mgf method is that if the can calculate the mgf of U, $m_U(t)$ and recognize it as the mgf of a known distribution, then U has that distribution.

The method of mgf's is especially useful for deriving the distribution of the sum of several independent random variables

Recall that we proved earlier that if X_1, ..., X_n are independent and $U = X_1 + \ldots + X_n$, then

$$m_U(t) = \prod_{i=1}^{n} m_{X_i}(t).$$

So it is easy to compute the mgf of U. The main question is whether we can recognize the result as an mgf of a known random variable.

The following example is probably the most important for statistical applications.

Example 5.3.1. Suppose X_1, ..., X_n are independent normal r.v.'s with means μ_1, ..., μ_n and variances σ_1^2, ..., σ_n^2. What is the distribution of $U = X_1 + X_2 + \ldots + X_n$?

The mgf of X_i is

$$m_{X_i}(t) = \exp(\mu_i t + \frac{\sigma_i^2}{2} t^2),$$

and the mgf of U is

$$m_U(t) = \exp\Big((\mu_1 + \ldots \mu_n)t + \frac{(\sigma_1^2 + \ldots + \sigma_n^2)}{2} t^2\Big).$$

Hence, U is a normal r. v. with parameters $(\mu_1 + \ldots \mu_n)$ and $(\sigma_1^2 + \ldots + \sigma_n^2)$.

In fact, one can prove a more general fact: if several random variables $X_1, \ldots, X_n$ are jointly normal, then their sum is always normal (even if they are not independent). So one only needs to calculate the expectation and variance of $U = X_1 + \ldots + X_n$ in order to write down its density.

Here are some other examples, when we can calculate the density of the sum $X_1 + \ldots + X_n$ by using the mgf method.

Example 5.3.2. Suppose X_1, ..., X_n are independent exponential random variables with mean β. What is the distribution of $U = X_1 + \ldots + X_n$?

The mgf of an exponential r.v. is $m_X(t) = (1 - \beta t)^{-1}$. Hence the mgf of U is $m_U(t) = (1 - \beta t)^{-n}$. This can be recognized as the mgf of the Gamma distribution with parameters $\alpha = n$ and β. Hence the density of U is

$$f_U(t) = \begin{cases} \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{n-1} e^{-x/\beta}, & \text{if } x \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

Note that in this example it is important that all exponential variables have the same mean β.

Example 5.3.3. Suppose X_1, ..., X_n are independent Poisson r.v.'s with parameters λ_1, ..., λ_n. What is the distribution of $U = X_1 + \ldots + X_n$?

The Poisson r.v. X_i with parameter λ_i has the mgf

$$m_{X_i}(t) = \exp(\lambda_i(e^t - 1)).$$

Hence,

$$m_U(t) = \exp\Big((\lambda_1 + \ldots + \lambda_n)(e^t - 1)\Big),$$

and U is the Poisson random variable with parameter $(\lambda_1 + \ldots + \lambda_n)$.

Exercise 5.3.4. Suppose X_1, ..., X_n are independent r.v.'s with distributions $\Gamma(\alpha_1, \beta), \ldots, \Gamma(\alpha_n, \beta)$, respectively (note that β is the same for all random variables). What is the distribution of $U = X_1 + X_2 + \ldots + X_n$?

5.4 Order Statistics

Let X_1, ..., X_n be n random variables. The *order statistics* are:

$X_{(1)}$ = the smallest value in the sample,

$X_{(2)}$ = the second-smallest value in the sample,

$\vdots$

$X_{(n)}$ = the largest value in the sample.

So, $X_{(1)} \le X_{(2)} \le \ldots \le X_{(n)}$.

Order statistics are functions of the original random variables. For example,

$$X_{(1)} = \min\{X_1, \ldots, X_n\},$$
$$X_{(n)} = \max\{X_1, \ldots, X_n\}.$$

So, the order statistics are random variables and we can ask a question about their probability distribution.

The motivation behind this question is that order statistics frequently appear in applications. Here is a couple of examples.

- Let X_t denote the volume of network traffic at minute t. We are interested in behavior of $X_{(n)}$, the maximum traffic volume, that occurs in the interval of n minutes.
- We need a net to hold particles. If $X_1, \ldots, X_n$ are the sizes of particles, we are interested in $X_{(1)}$, their minimum size.

In addition, some common summary statistics are functions of order statistics.

If n is odd the *sample median* of the sample $X_1, X_2, \ldots, X_n$ is defined as $X_{\left((n+1)/2\right)}$. If n is even it is defined as $(X_{(n/2)} + X_{(n/2+1)})/2$.

A *sample range* is defined as $X_{(n)} - X_{(1)}$

In particular, the distribution of these statistics depends on the (joint) distribution of order statistics.

5.4.1 Distribution of order statistics via density transformation method

First let us see, how to calculate the joint density of order statistics. The vector of order statistics $(X_{(1)}, \ldots, X_{(n)})$ is a function of the vector $(X_1, \ldots X_n)$. The difficulty is that it is not one-to-one function. Consider, for example, $n = 3$, and a triple of numbers (a, b, c). Then, we can have

$$X_{(1)} = a, X_{(2)} = b, X_{(3)} = c$$

only if $a \leq b \leq c$. However this vector of values can arise for 6 different (assuming a, b, c are different) realizations of variables X_1, X_2, X_3. One possibility is that $X_1 = a, X_2 = b, X_3 = c$, another $X_1 = b, X_2 = a, X_3 = c$, and so on, with 6 possible permutations of set (a, b, c).

The good news is that the six inverse functions are simple and have the Jacobian equal to 1. For example, for the second case above, we have the inverse function $X_1 = X_{(2)}, X_2 = X_{(1)}, X_3 = X_{(3)}$, and the Jacobian of this transformation is

$$J = \left| \det \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right| = 1.$$

So, we can calculate the joint density as

$$f_{X_{(1)},X_{(2)},X_{(3)}}(a, b, c) = \begin{cases} f_{X_1,X_2,X_3}(a, b, c) + \ldots + f_{X_1,X_2,X_3}(c, b, a) & \text{if } a \leq b \leq c, \\ 0 & \text{otherwise,} \end{cases}$$

where summation is over all permutations of the set (a, b, c).

If random variables $X_1, \ldots, X_3$ are independent and identically distributed, then this formula simplifies. Let X_i have density $f(x)$ for every i, then we can rewrite the formula above as

$$f_{X_{(1)},X_{(2)},X_{(3)}}(a,b,c) = \begin{cases} 3!f(a)f(b)f(c) & \text{if } a \leq b \leq c, \\ 0 & \text{otherwise.} \end{cases}$$

Obviously, this formula generalizes to arbitrary $n \geq 1$ and we have the following theorem.

Theorem 5.4.1. *Suppose X_1, …, X_n, are i.i.d. with density $f(x)$. Then the joint density of the order statistic vector is*

$$f_{X_{(1)},\ldots,X_{(n)}}(x_1,\ldots,x_n) = \begin{cases} n!f(x_1)f(x_2)\ldots f(x_n) & \textit{if } x_1 \leq \ldots \leq x_n, \\ 0 & \textit{otherwise.} \end{cases}$$

5.4.2 Distribution of order statistics via cdf method

In order to find the probability distribution of individual order statistics, we can apply a variant of the cdf method. This is especially easy for the maximum and the minimum of the sample $X_1, \ldots, X_n$.

Theorem 5.4.2. *Suppose X_1, …, X_n, are i.i.d. with cdf $F_X(x)$. Then the cdf of the* maximum *statistic $X_{(n)}$ is*

$$F_{X_{(n)}}(x) = \Big[F_X(x)\Big]^n$$

Proof.

$$\begin{aligned} \mathbb{P}(X_{(n)} \leq x) &= \mathbb{P}\Big(\max\{X_1,\ldots,X_n\} \leq x\Big) \\ &= \mathbb{P}(X_1 \leq x, X_2 \leq x, \ldots, X_n \leq x) \\ &= \Big[\mathbb{P}(X_1 \leq x)\Big]^n \\ &= \Big[F_X(x)\Big]^n, \end{aligned}$$

where the equality in the third line follows by the independence of r.v.s X_1, …, X_n. □

In particular, if the variable X_i have density $f_X(x)$, then the order statistics also have a density, which can be calculated as the derivative fo the cdf.

$$f_{X_{(n)}}(x) = nf(x)F(x)^{n-1}.$$

A similar result holds for the minimum statistic.

Theorem 5.4.3. *Suppose X_1, …, X_n, are i.i.d. with cdf $F_X(x)$. Then the cdf of the* minimum *statistic $X_{(1)}$ is*

$$F_{X_{(1)}}(x) = 1 - \Big[1 - F_X(x)\Big]^n$$

Proof. Here it is more convenient to calculate the survival function:

$$\begin{aligned} 1 - F_{X_{(1)}}(x) &= \mathbb{P}(X_{(1)} > x) \\ &= \mathbb{P}\Big(\min\{X_1, \ldots, X_n\} > x\Big) \\ &= \mathbb{P}(X_1 > x, X_2 > x, \ldots, X_n > x) \\ &= \Big[\mathbb{P}(X_1 > x)\Big]^n \\ &= \Big[1 - F(x)\Big]^n. \end{aligned}$$

□

So, if random variables X_i have density then the minimum $X_{(1)}$ also has a density, which we calculate as

$$f_{X_{(1)}}(x) = nf(x)(1 - F(x))^{n-1}.$$

The application of these formulas is straightforward.

Example 5.4.4. Suppose wave heights have an exponential distribution with mean height 10 feet. If 200 waves crash during the night, what is the distribution of the highest wave? What is the probability that the highest wave is more than 50 feet? more than 100 feet?

The density and cdf of the wave height are

$$\begin{aligned} f_X(x) &= \frac{1}{10}e^{-x/10}, \\ F_X(x) &= 1 - e^{-x/10}. \end{aligned}$$

So we get

$$F_{X_{(200)}}(x) = \left(1 - e^{-x/10}\right)^{200}.$$

In particular,

$$\begin{aligned}\mathbb{P}(X_{(200)} > 50) &= 1 - F_{X_{(200)}}(50) = 1 - \left(1 - e^{-50/10}\right)^{200} \\ &= 0.7413165\end{aligned}$$

Similarly,

$$\begin{aligned}\mathbb{P}(X_{(200)} > 100) &= 1 - F_{X_{(200)}}(100) = 1 - \left(1 - e^{-100/10}\right)^{200} \\ &= 0.009039092\end{aligned}$$

Example 5.4.5. Suppose light bulbs' life-lengths have an exponential distribution with mean 1200 hours. Two bulbs are installed at the same time. What is the expected time until one bulb has burned out? What if we had installed 3 lamps? If X_1, X_2 are the life-lengths of the bulbs then the time until one bulb has burned out is $X_{(1)} = \min\{X_1, X_2\}$. The distribution of the exponential r.v. X_i has the survival function

$$\mathbb{P}(X_i > x) = e^{-x/1.2}$$

where we measure time in thousands of hours. Then, the survival function for $X_{(1)}$ is

$$\mathbb{P}(X_{(1)} > x) = \left[e^{-x/1.2}\right]^2 = e^{-x/0.6}$$

So we can conclude that $X_{(1)}$ has the exponential distribution with mean 600 hours. Hence, the expected time until one bulb has burned out is 600 hours.

Similarly, we can calculate for 3 bulbs, that $X_{(1)}$ has the exponential distribution with mean $1200/3 = 400$ and so the time until one bulb has burned out is 400 hours.

This the reflection of the general fact that if we have n i.i.d random variables $X_1, \ldots, X_n$ that have the exponential distribution with mean β,

then its minimum $X_{(1)}$ also have the exponential distribution, only with mean β/n.

However, the distribution of the maximum of exponential random variables is complicated.

We can also find the distribution of other order statistics. Consider, for example, the third smallest statistic $X_{(3)}$ (and assume that $n > 3$).

The event $X_{(3)} \leq x$ occurs when at least 3 of variables $X_1, \ldots X_n$ are smaller than or equal to x. The probability that exactly $k \geq 3$ specific variables $X_{i_1}, X_{i_2}, \ldots, X_{i_k}$ is $\leq x$ and all others X_{j_1}, ..., $X_{j_{n-k}}$ are $> x$ is

$$\mathbb{P}(X_{i_1} \leq x, \ldots, X_{i_k} \leq x, X_{j_1} > x, \ldots, X_{j_{n-k}} > x) = \Big[F_X(x)\Big]^k \Big[1 - F_X(x)\Big]^{n-k}$$

Now, we can choose the set of indices $i_1, \ldots, i_k$ in $\binom{n}{k}$ ways. So the probability that exactly k of n variables $X_1, \ldots, X_n$ are $\leq x$ and all others are $> x$ is

$$\binom{n}{k} \Big[F_X(x)\Big]^k \Big[1 - F_X(x)\Big]^{n-k}.$$

The probability that $X_{(3)} \leq x$ is the sum of all these probabilities over $k \geq 3$:

$$F_{X_{(3)}}(x) = \mathbb{P}(X_{(3)} \leq x) = \sum_{k=3}^{n} \binom{n}{k} F_X(x)^k (1 - F_X(x))^{n-k}.$$

More generally, for every s between 1 and n,

$$F_{X_{(s)}}(x) = \mathbb{P}(X_{(s)} \leq x) = \sum_{k=s}^{n} \binom{n}{k} F_X(x)^k (1 - F_X(x))^{n-k}.$$

If the random variables X_i have a density, then the density of $X_{(s)}$ can be obtained by the differentiation of this expression. Rather magically, the formula simplifies.

Theorem 5.4.6. *Let $X_1, \ldots, X_n$ be i.i.d continuous random variables with density $f_X(x)$ and cdf $F_X(x)$. Then, $X_{(s)}$ is a continuous random variable with the density given by*

$$f_{X_{(s)}}(x) = \binom{n}{s-1, 1, n-s} f_X(x) \Big[F_X(x)\Big]^{s-1} \Big[1 - F_X(x)\Big]^{n-s}$$

Proof. We differentiate the expression for the cdf of $X_{(s)}$ and find that

$$f_{X_{(s)}}(x) = f_X(x) \sum_{k=s}^{n} \binom{n}{k} \Big(kF_X(x)^{k-1}(1-F_X(x))^{n-k} \\ - (n-k)F_X(x)^k(1-F_X(x))^{n-k-1} \Big).$$

Let us write several terms of the sum explicitly. For shortness, we write F instead of $F_X(x)$.

$$\binom{n}{s} sF^{s-1}(1-F)^{n-s} - \binom{n}{s}(n-s)F^s(1-F)^{n-s-1} \\ + \binom{n}{s+1}(s+1)F^s(1-F)^{n-s-1} - \binom{n}{s+1}(n-s-1)F^{s+1}(1-F)^{n-s-2} \\ + \dots \\ + \binom{n}{n} nF^{n-1}(1-F)^0 - 0$$

By noting that $\binom{n}{k}(n-k) = \binom{n}{k+1}k+1$, we can observe that the sum telescopes to

$$\binom{n}{s} sF^{s-1}(1-F)^{n-s},$$

and since

$$\binom{n}{s} s = \frac{n!}{(s-1)!(n-s)!} = \binom{n}{s-1,1,n-s}$$

the claim of the theorem is proved. □

The formula have some intuitive meaning. We partition the variables $X_1, \dots, X_n$ into $s-1$ of those that have value $< x$, one that have value in a small interval of length dx around x and $n-s$ of those that are larger than x. The number of these partitions is given by the multinomial coefficient $\binom{n}{s-1,1,n-s}$ and the probability is $F_X(x)^{s-1} f_X(x)(1-F_X(x))^{n-s}dx$. This allows us to write the formula in the theorem, however this argument is not so easy to turn into a proof.

Example 5.4.7. Ten numbers are generated uniformly at random between 0 and 1. What is the distribution of the 3rd-smallest of these? What is the expected value of the 3rd smallest?

The cdf of the uniform distribution is $F_X(x) = x$ if $x \in [0,1]$ and 0, otherwise. So by our formula we have:

$$f_{X_{(3)}}(x) = \frac{10 \times 9 \times 8}{2} x^2 (1-x)^7.$$

This is the beta distribution with parameters $\alpha = 3$ and $\beta = 8$, and for the beta distribution we know how to calculate the expectation. So,

$$\mathbb{E}X_{(3)} = \frac{\alpha}{\alpha + \beta} = \frac{3}{3+8} = \frac{3}{11}.$$

More generally if $X_1, \ldots, X_n$ are i.i.d from the uniform distribution on $[0,1]$, then the order statistic $X_{(s)}$ has the beta distribution with parameters $\alpha = s$ and $\beta = n + 1 - s$, and the expectation of this statistic is

$$\mathbb{E}X_{(s)} = \frac{s}{n+1}.$$

It is interesting that sometimes one does not need the formulas for cdf or pdf to answer questions about order statistics.

Example 5.4.8. Let $X_1, \ldots, X_n$ are independent and identically distributed (i.i.d) observations having a *continuous* cdf $F(t)$. Let $X_{(k)}$ denote the k-th order statistic from this collection of random variables. Also let $Y_1, \ldots, Y_5$ are i. i. d. with the same distribution as X_i and independent of X_i.

1. If $n = 15$, what is $\mathbb{P}(X_{(7)} = X_{13})$?

   Note that the random variables $X_1, \ldots, X_n$ is a random permutation of the order statistics $X_{(1)}, \ldots, X_{(n)}$. The probability that after this permutation $X_{(7)}$ will be exactly in the place number 13 is equal to $1/15$ since all possible positions are equally probable after the permutation.

2. If $n = 15$, what is the probability $\mathbb{P}(X_{(3)} < Y_1 < X_{(10)})$?

Add Y_1 to $X_1, \ldots, X_{15}$ and order them. The possible outcomes are

$$\begin{aligned} Y_1 < X_{(1)} < \ldots < X_{(15)}, \\ X_{(1)} < Y_1 < X_{(2)} < \ldots < X_{(15)} \\ \ldots \\ X_{(1)} < \ldots < X_{(14)} < Y_1 < X_{(15)} \\ X_{(1)} < \ldots < X_{(15)} < Y_1 \end{aligned}$$

These are 16 outcomes which are equally probable. The event $\{X_{(3)} < Y_1 < X_{(10)}\}$ corresponds to those of the outcomes where Y_1 is immediately after $X_{(k)}$ with $k = 3, 4, \ldots, 9$. There are $(9-3)+1 = 7$ of these outcomes. Therefore the probability of the event is $7/16$.

3. What is the smallest n such that $\mathbb{P}(Y_1 < X_{(n)}) \geq .9$?

   By argument as in previous item, we have $\mathbb{P}(Y_1 < X_{(n)}) = n/n+1$. Hence, we need $1/(n+1) \leq 0.1$ or $n \geq 10 - 1 = 9$.

4. What is the probability that $Y_1, \ldots, Y_5$ all fall in the interval $[X_{(1)}, X_{(n)}]$?

   The probability that Y_1 falls in the interval $[X_{(1)}, X_{(n)}]$ is $(n-1)/(n+1)$. The random variables $Y_1, \ldots, Y_5$ are independent, so the probability that they all fall in the interval $[X_{(1)}, X_{(n)}]$ is

$$\left(\frac{n-1}{n+1}\right)^5.$$

Chapter 6

Sampling distributions, Law of Large Numbers, and Central Limit Theorem

6.1 Sampling distributions

Definition 6.1.1. A *statistic* is a function of several random variables.

A statistic is usually used to estimate an unknown parameter of a distribution from an observed sample taken from this distribution. For example, the *sample average* (or *sample mean*) $\overline{X}$ for a collection (or "sample") of random variables $X_1, X_2, \ldots, X_n$ is defined as

$$\overline{X} = \frac{1}{n}(X_1 + \ldots + X_n),$$

and is often used to estimate the expectation of the distribution of X_i.

6.1.1 Sample average of i.i.d normal random variables

For normal random variables, it is easy to calculate the distribution of the average.

Theorem 6.1.2. *Let independent random variables $X_1, \ldots, X_n$ have a normal distribution with mean μ and variance σ^2. Then the sample average $\overline{X}$*

has the normal distribution with mean μ and variance σ^2/n.

Proof. First, by Example 5.3.1, the sum $S = X_1 + \ldots + X_n$ is the normal r.v. with mean $n\mu$ and variance $n\sigma^2$. That means that its density is

$$f_S(s) = \frac{1}{\sqrt{2\pi n\sigma^2}} \exp\Big(- \frac{(s-n\mu)^2}{2n\sigma^2}\Big)$$

Next we apply the pdf transformation method to the function $\overline{X} = S/n$. The inverse function is $S = n\overline{X}$, so the Jacobian is n and we can write

$$\begin{aligned} f_{\overline{X}}(x) = f_S(nx) \times n &= \frac{n}{\sqrt{2\pi n\sigma^2}} \exp\Big(- \frac{(nx-n\mu)^2}{2n\sigma^2}\Big) \\ &= \frac{1}{\sqrt{2\pi(\sigma^2/n)}} \exp\Big(- \frac{(x-\mu)^2}{2(\sigma^2/n)}\Big), \end{aligned}$$

and we can identify this density as the density of a normal r.v. with mean μ and variance σ^2/n. □

As a consequence we can see that

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

has the standard normal distribution.

The point of this theorem is that the sample average $\overline{X}$ has the same mean as the original random variables $X_1, \ldots, X_n$ but smaller variance, so it can be used to estimate the mean μ from data.

Example 6.1.3. A bottling machine can be regulated so that it discharges an average of μ ounces per bottle. It has been observed that the amount of fill dispensed by the machine is normally distributed with $\sigma = 1.0$ ounce. A sample of $n = 9$ filled bottles is randomly selected from the output of the machine on a given day (all bottled with the same machine setting), and the ounces of fill are measured for each. Find the probability that the sample mean will be within 0.3 ounce of the true mean μ for the chosen machine setting. How many observations should be included in the sample if we wish $\overline{X}$ to be within .3 ounce of μ with probability 0.95?

Essentially we have $X_1, \ldots, X_9 \sim N(\mu, 1)$ and we want to calculate $\mathbb{P}(|\overline{X} - \mu| \leq 0.3)$. This is the same as

$$\begin{aligned} &\mathbb{P}\Big(\Big|\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}\Big| \leq \frac{0.3}{\sigma/\sqrt{n}}\Big) \\ = &\mathbb{P}\Big(-\frac{0.3}{1/3} \leq Z \leq \frac{0.3}{1/3}\Big) \end{aligned}$$

which can be calculated as

$$1-2\mathbb{P}(Z > .9) = 1-2 \times .1841 = .6318.$$

For the second question we observe that if we wish

$$1-2\mathbb{P}(Z > z) = 0.95,$$

then we need to have $\mathbb{P}(Z > z) = 0.025$, which corresponds to $z = 1.96$. Then we have equation

$$\frac{0.3}{1/\sqrt{n}} = 1.96,$$

which can be solved to give $n = (1.96/0.3)^2 = 42.68$. Since it is not possible to use fractional bottles, we should set $n = 43$.

6.1.2 Sum of the squares of standardized normal random variables

In the previous section we looked at the sample average $\overline{X}$ which is used to estimate the expected value μ of i. i. d. random variables $X_1, \ldots, X_n$.

What if we are instead interested in estimating variance of these random variables?

Recall that by definition the variance of a r.v. X is

$$\mathbb{V}\mathrm{ar}(X) := \mathbb{E}(X - \mu)^2,$$

where $\mu = \mathbb{E}X$. So $\mathbb{V}\mathrm{ar}(X)$ is the expected value of random variable $(X-\mu)^2$, and it is natural to estimate it by using the following statistic:

$$Y = \overline{(X - \mu)^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2.$$

The difficulty with this approach is that we typically do not know μ and it also should be estimated.

However, in this section we will pretend that we know μ, and postpone the question of unknown μ to the next section.

The statistic Y varies around the true value of variance σ^2 and it turns out that it is convenient to quantify this variation by looking at the distribution of nY/σ^2. (The reason for why we do this transformation is that we want that the resulting distribution belongs to the list of the distributions that we already defined.)

In the case of normal random variables, we can calculate the distribution of nY/σ^2 exactly.

Theorem 6.1.4. *Let $X_1, X_2, \ldots, X_n$ be independent random variables distributed according to $N(\mu, \sigma^2)$. Then $Z_i = (X_i - \mu)/\sigma$ are independent, standard normal random variables, $i = 1, 2, \ldots, n$, and*

$$\frac{nY}{\sigma^2} = \sum_{i=1}^{n} \Big(\frac{X_i - \mu}{\sigma}\Big)^2 = \sum_{i=1}^{n} Z_i^2$$

has a χ^2 distribution with n degrees of freedom, that is, the Gamma distribution with $\alpha = n/2$ and $\beta = 2$.

By using this theorem and some facts about the Gamma distribution, it is easy to figure out that $\mathbb{E}(nY/\sigma^2) = \alpha\beta = n$, so $\mathbb{E}Y = \sigma^2$. The variance of nY/σ^2 is $\alpha\beta^2 = 2n$, so the variance of Y is $2\sigma^4/n$. In particular, we see that the expected value of Y is the same as the variance of r.v.s. X_i, and the variance of Y declines as n grows.

Proof of Theorem 6.1.4. The first statement, that $Z_i = (X_i - \mu)/\sigma$ are standard normal random variables easily follows by either the pdf transformation method or by the mgf method. Also these transformed variables are still independent as functions of independent random variables.

To give full details for the pdf transformation method, we write the inverse function as $X_i = \mu + \sigma Z_i$, so the Jacobian is σ and the density of Z_i

is

$$f_{Z_i}(z) = \sigma f_{X_i}(\mu + \sigma z) = \frac{\sigma}{\sqrt{2\pi\sigma^2}} \exp\Big(- \frac{(\mu + \sigma z - \mu)^2}{2\sigma^2}\Big)$$
$$= \frac{1}{\sqrt{2\pi}} \exp\Big(- \frac{z^2}{2}\Big),$$

so Z_i is a standard normal random variable.

For the second statement, we first calculate the distribution of the Z^2, where Z is a standard normal random variable. One way to do it is by using the mgf method.

$$m_{Z^2}(t) = \mathbb{E}e^{tZ^2} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tz^2} e^{-z^2/2}\, dz$$
$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{z^2(t-1/2)}\, dz$$
$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\Big(- \frac{z^2}{2\frac{1}{1-2t}}\Big)\, dz$$
$$= \frac{1}{(1-2t)^{1/2}},$$

where the last step follows because we know that the density of a normal random variable with mean 0 and variance $1/(1-2t)$ has the integral equal to 1, so

$$\frac{1}{\sqrt{2\pi \times \frac{1}{1-2t}}} \int_{-\infty}^{\infty} \exp\Big(- \frac{z^2}{2\frac{1}{1-2t}}\Big)\, dz = 1, \text{ and therefore}$$
$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\Big(- \frac{z^2}{2\frac{1}{1-2t}}\Big)\, dz = \sqrt{\frac{1}{1-2t}},$$

We can recognize this mgf as the mgf of a gamma random variable with parameters $\alpha = 1/2$ and $\beta = 2$.

Next, we use the fact from Exercise 5.3.4 which implies that the sum of n gamma random variables with parameters α and $\beta = 2$ is a gamma distributed random variable with parameters $n\alpha = n/2$ and $\beta = 2$. □

Example 6.1.5. If $Z_1, Z_2, \ldots, Z_6$ denotes a random sample from the standard normal distribution, find a number b such that $\mathbb{P}(Z_1^2 + \ldots + Z_6^2 \leq b) = 0.95)$.

By the previous theorem, the sum is distributed as χ^2 with $df = 6$. Hence we can calculate this by using the R quantile function for the χ^2 distribution:

```
qchisq(0.95, 6)
```

which gives $b = 12.59159$.

6.1.3 Distribution of the sample variance for normal r.v.s

In this section and in the following we will consider functions of random variables which are important for statistical questions when the sample sizes are small (the number of observations is smaller than ≈ 30). Since in modern applications this is not a very typical situation, we only overview these results.

The result in the previous theorem is not satisfactory because in practice μ is not known. In statistics, the estimator for the parameter σ^2 of the normal distribution is called the *sample variance* and it is defined by the formula

$$S^2 = \frac{1}{n-1}\sum_{1}^{n}(X_i - \overline{X})^2.$$

Note that in difference from the previous statistics, we divide not by n but by $n-1$.

Theorem 6.1.6. *Let $X_1, X_2, \ldots, X_n$ be independent random variables distributed according to $N(\mu, \sigma^2)$. Then,*

$$\frac{(n-1)S^2}{\sigma^2} = \sum \Big(\frac{X_i - \overline{X}}{\sigma}\Big)^2$$

has a χ^2 distribution with $(n-1)$ df. Also, $\overline{X}$ and S^2 are independent random variables.

Note that in difference from the previous theorem χ^2 distribution has $n-1$ degrees of freedom, not n.

Unfortunately, the proof of this important theorem is not easy and we will skip it.

This theorem can be used to make some inferences about the variance of underlying distribution:

Example 6.1.7. In Example 6.1.3, the ounces of fill from the bottling machine are assumed to have a normal distribution with $\sigma^2 = 1$. Suppose that we plan to select a random sample of ten bottles and measure the amount of fill in each bottle. If these ten observations are used to calculate S^2, it might be useful to specify an interval of values around $\sigma^2 = 1$ that should include S^2 with a high probability provided our assumption about σ^2 is correct. If the resulting S^2 happen to be outside of this interval, it might be that something is wrong with our assumption about σ (or with our assumption about the normality of the disribution). So, the task is to find numbers b_1 and b_2 such that

$$\mathbb{P}(b_1 \leq S^2 \leq b_2) = 0.90.$$

We can rewrite the inequalities and write the probability as

$$\mathbb{P}\Big(\frac{(n-1)b_1}{\sigma^2} \leq \frac{(n-1)S^2}{\sigma^2} \leq \frac{(n-1)b_2}{\sigma^2}\Big) = 0.90.$$

The random variable $Y = (n-1)S^2/\sigma^2$ is distributed according to χ^2 distribution with $n-1 = 9$ degrees of freedom. So we are looking for numbers y_1 and y_2 such that

$$\mathbb{P}(y_1 \leq Y \leq y_2) = 0.90.$$

These numbers can be chosen in several different ways. One conventional choice is to choose them so that

$$\mathbb{P}(Y \geq y_2) = 0.05 \text{ and } \mathbb{P}(Y \leq y_1) = 0.05$$

By using R (or tables), we find

```
y_1 = qchisq(0.05, 9) = 3.325113,
y_2 = qchisq(0.95, 9) = 16.91898
```

Then we can calculate

$$b_1 = \frac{\sigma^2}{n-1} y_1 = \frac{1}{9} \times 3.325113 = 0.369457$$
$$b_2 = \frac{\sigma^2}{n-1} y_2 = \frac{1}{9} \times 16.91898 = 1.879886$$

So if we observe the sample variance S^2 outside of the interval $(0.369457, 1.879886)$, then we might suspect that something is wrong about our assumptions. Either σ^2 is different from the assumed value of 1 or the random variables are not normal.

6.1.4 Student's t distribution

For the problems of statistical inference one needs another distribution, which is called a Student's t distribution (or simply t distribution).

Definition 6.1.8. Let Z be a standard normal random variable and let W be a χ^2-distributed variable with $df = \nu$. Assume that Z and W are independent. Then, the random variable

$$T = \frac{Z}{\sqrt{W}/\nu}$$

is said to have a t *distribution* with ν degrees of freedom.

The density of the t-distribution with ν degrees of freedom is as follows:

$$f_T(t) = \frac{\Gamma\big((\nu+1)/2\big)}{\sqrt{\pi}\nu\Gamma\big(\nu/2\big)} \Big(1 + \frac{t^2}{\nu}\Big)^{-(\nu+1)/2},$$

for all real t. This density is symmetric, similar to the density of the normal distribution. However note that when t grows, the density declines much slower compared to the density of the normal distribution. (It declines as $1/t^\nu$ while the density of the normal declines as $\approx 1/e^{t^2}$ which is much faster.) For this reason, the t distribution is sometimes used to model the situation where big outliers are possible.

On the other hand, when ν is large it is easy to see that the density is close to the density of the normal distribution. Indeed, for every t,

$$\Big(1 + \frac{t^2}{\nu}\Big)^{-(\nu+1)/2} \to e^{-t^2/2},$$

as $\nu \to \infty$.

The density of the t distribution is relevant for our topic, – the sampling distributions, for the following reason.

We can write:

$$\frac{\overline{X}-\mu}{S/\sqrt{n}} = \frac{(\overline{X}-\mu)/(\sigma/\sqrt{n})}{\sqrt{\big[(n-1)S^2/\sigma^2\big]/(n-1)}}$$

If we apply Theorem 6.1.6 to the right hand side, this random variable has the t-distribution with $n-1$ degree of freedom.

The benefit of this result is that the quantity $\frac{\overline{X}-\mu}{S/\sqrt{n}}$ includes only one unknown quantity, μ, and not two, μ and σ, as, for example, the quantity $(\overline{X}-\mu)/(\sigma/\sqrt{n})$. In particular, it allows us to test our assumptions about μ only, even if we do not know σ.

Example 6.1.9. The tensile strength for a type of wire is normally distributed with unknown mean μ and unknown variance σ^2. Six pieces of wire were randomly selected from a large roll. The tensile strength X_i was measured for each $i = 1, 2, \ldots, 6$. Find the approximate probability that $\overline{X}$ will be within $2S/\sqrt{n}$ of the true population mean μ.

We have

$$\begin{aligned}\mathbb{P}(-2S/\sqrt{n} \le X-\mu \le 2S/\sqrt{n}) &= \mathbb{P}\Big(-2 \le \frac{X-\mu}{S/\sqrt{n}} \le 2\Big)\\ &= 1 - 2\mathbb{P}(T>2)\\ &= 2\mathbb{P}(T \le 2) - 1,\end{aligned}$$

where T has the t distribution with $5 = 6-1$ degrees of freedom. Then by using R, we calculate

```
2 * pt(2, 5) - 1 = 0.8980605 \approx 90%.
```

Note that if T had the standard normal distribution, then we would get

```
2 * pnorm(2) - 1 = 0.9544997 \approx 95%.
```

6.1.5 F distribution

One more distribution that sometimes occur in statistics is F distribution.

Definition 6.1.10. Let W_1 and W_2 be independent χ^2 distributed random variables with ν_1 and ν_2 degrees of freedom, respectively. Then the random variable

$$F = \frac{W_1/\nu_1}{W_2/\nu_2}$$

has an F distribution with μ_1 numerator degrees of freedom and μ_2 denominator degrees of freedom.

The density of F distribution is

$$f_F(x) = c(\nu_1, \nu_2) x^{(\nu_1/2)-1} \Big(1 + \frac{\nu_1}{\nu_2} x\Big)^{-(\nu_1+\nu_2)/2},$$

for $x \geq 0$.

It is useful for statistics because it allows to test assumptions about variances in two samples. if we have two independent samples of normally distributed random variable that have n_1 and n_2 elements, respectively, and if we assume that the variance of random variables in the first sample is σ_1^2 and that it is σ_2^2 in the second sample, then we can form the ratio

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{\sigma_2^2}{\sigma_1^2} \times \frac{S_1^2}{S_2^2}.$$

By Theorem 6.1.6, this ratio has the F distribution with parameters $(n_1 - 1, n_2 - 1)$, and we can use this to test our assumptions about σ_1 and σ_2.

Example 6.1.11. If we take independent samples of size $n_1 = 6$ and $n_2 = 10$ from two normal populations with equal population variances, find the number b such that

$$\mathbb{P}\Big(\frac{S_1^2}{S_2^2} \leq b\Big) = 0.95$$

(The idea is that if the ratio is greater than b we will reject the assumption that the variances are equal)

Since we assumed that $\sigma_1 = \sigma_2$, the ratio S_1^2/S_2^2 has the F distribution with parameters $(n_1 - 1, n_2 - 1) = (5, 9)$.

So we calculate b using R:

`qf(0.95, 5, 9) = 3.481659.`

6.2 Law of Large Numbers (LLN)

We have seen that for normal random variables $X_1, \ldots, X_n$ the distribution of the sample average $\overline{X}$ is normal with mean μ and variance σ^2/n. So the variance becomes smaller and smaller as n grows while the expectation remains equal to μ.

This is also true if the random variables are not necessarily normal and we can prove the following result.

Theorem 6.2.1 (Law of Large Numbers (LLN))**.** *Let $X_1, X_2, \ldots$ be an infinite sequence of i.i.d. random variables with finite expectation μ and variance σ^2. Let*

$$\overline{X}_n = \frac{X_1 + X_2 + ... + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Then for every $\epsilon > 0$, we have

$$\lim_{n\to\infty} \mathbb{P}(|\overline{X}_n - \mu| > \varepsilon) = 0.$$

In other words, the probability that $\overline{X}_n$ deviates from μ by more than ε converges to zero as n grows.

One says that $\overline{X}_n$ *converges to μ in probability.*

Proof. We use Chebyshev's inequality.

$$\mathbb{P}(|\overline{X}_n - \mu| > \varepsilon) \leq \frac{\mathbb{V}\mathrm{ar}(\overline{X}_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}$$

The right-hand side converges to zero. Therefore $\mathbb{P}(|\overline{X}_n - \mu| > \varepsilon)$ also converges to zero. □

6.3 The Central Limit Theorem

As in the previous section $X_1, X_2, \ldots$ is an infinite sequence of i.i.d. random variables with mean μ and standard deviation σ. As before

$$\overline{X}_n = \frac{X_1 + X_2 + ... + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

It is easy to calculate the mean and the variance of $\overline{X}$ by using formulas for the expectation and variance of linear combinations of random variables. $\mathbb{E}(X) = \mu$, $\mathbb{V}\text{ar}(\overline{X}) = \sigma^2/n$.

For normal random variables X_i, we have seen that $\overline{X}$ is normal with mean μ and variance σ^2/n. In particular this means that

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$$

has the standard normal distribution for all n.

What can be said about Z_n if X_i are i.i.d but not normal? It turns out that in this case Z_n is *approximately* normal for large n. More precisely, the following theorem holds.

Theorem 6.3.1 (The central limit theorem (CLT))**.** *Let $X_1, X_2, \ldots$ be i.i.d random variables with finite mean μ and variance σ^2. Define*

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}},$$

where $\overline{X}_n = (X_1 + \ldots X_n)/n$. Then the cdf of Z_n converges to the cdf of a standard normal variable. That is for all z,

$$F_{Z_n}(z) = \mathbb{P}(Z_n \leq z) \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2} dt$$

when $n \to \infty$.

Remark: Convergence of this type is called the *convergence in distribution*. So, the normalized sample average Z_n converges to a standard normal r.v. Z in distribution.

Proof. We will prove the theorem only for the case when the moment-generating function $m_{X_i}(t)$ exists for all t. The more general case can be proved similarly by using the characteristic function $\varphi_X(t) = \mathbb{E}(e^{itX})$ (where $i = \sqrt{-1}$ is the imaginary unit of complex numbers) which exists for all t.

The proof is based on an important statement which we accept without proof. Namely, if for a sequence of random variables $Y_1, Y_2, \ldots$, the moment generating functions of these variables converge to the mgf of a random variable Y, with the convergence for all real t, then the cdf of Y_n converges to the cdf of Y.

We will apply this statement to the random variables Z_n.

Note that

$$Z_n = \frac{(X_1 + \ldots + X_n)/n - \mu}{\sigma/\sqrt{n}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma}$$

Then if we use the notation $U_i = (X_i - \mu)/\sigma$, we have

$$m_{\sum U_i}(t) = \Big[m_{U_1}(t)\Big]^n,$$

and

$$m_{Z_n}(t) = m_{\frac{1}{\sqrt{n}}\sum U_i}(t) = m_{\sum U_i}(t/\sqrt{n}) = \Big[m_{U_1}(t/\sqrt{n})\Big]^n,$$

where in the second equality we used the identity $m_{aX}(t) = m_X(at)$ (obvious from the definition of the moment generating function).

Now, the random variable U_1 has mean zero and the variance 1, which means that $m'_{U_1}(0) = 0$ and $m''_{U_1}(0) = 1$. We can use the Taylor's formula with the remainder to write

$$m_{U_1}(t) = 1 + \frac{1}{2} m''_{U_1}(\xi) t^2,$$

where $\xi \in [0, t]$. And so

$$m_{Z_n}(t) = \Big[m_{U_1}(t/\sqrt{n})\Big]^n = \left[1 + \frac{m''_{U_1}(\xi_n)}{2}\Big(\frac{t}{\sqrt{n}}\Big)^2\right]^n$$

where $\xi_n \in [0, t/\sqrt{n}]$. As $n \to \infty$, $\xi_n \to 0$ and so $m''_{U_1}(\xi_n) \to m''_{U_1}(0) = 1$. (Here one needs the continuity of $m''_{U_1}(t)$ at $t = 0$ which we also assume without proof.) This implies $\frac{m''_{U_1}(\xi_n)}{2} t^2 \to t^2/2$.

Finally we use the fact that if $b_n \to b$ then

$$\left(1 + \frac{b_n}{n}\right)^n \to e^b.$$

(This can be considered as a good calculus exercise on limits.) If we apply it to $b_n = \frac{m''_{U_1}(\xi_n)}{2} t^2$ and $b = t^2/2$, then we get that

$$m_{Z_n}(t) = \left[1 + \frac{m''_{U_1}(\xi_n)}{2} \frac{t^2}{n}\right]^n \to e^{t^2/2} \text{ as } n \to \infty,$$

and it remains to recall that $e^{t^2/2}$ is the mgf of a standard normal random variable Z, in order to conclude that both the mgf and cdf of Z_n converge to the mgf and cdf of Z, respectively.

□

Note that we can write Z_n differently,

$$Z_n = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{(X_1 + \ldots + X_n) - \mu n}{\sigma\sqrt{n}}.$$

This gives another form of the Central Limit Theorem. Namely if $X_1, X_2, \ldots$ are i.i.d random variables with mean μ and variance σ^2, and if $S_n = X_1 + \ldots X_n$, then

$$\frac{S_n - \mu n}{\sigma\sqrt{n}} \to Z,$$

where Z is the standard normal random variable and the convergence is in distribution.

The main impact of these theorems is that if the number n of independent random variables $X_1, \ldots, X_n$ is large, then the sums and averages of these random variables behave as if $X_1, \ldots, X_n$ were normal random variables.

Example 6.3.2. The service times for customers at a cashier are i.i.d random variables with mean 2.5 minutes and standard deviation 2 minutes. Approximately, what is the probability that the cashier will take more than 4 hours to serve 100 people?

If the service times are denoted $X_1, X_2, \ldots$, then the question is about the probability

$$\mathbb{P}(\sum_{i=1}^{100} X_i > 4 \times 60) = \mathbb{P}\Big(\frac{\sum_{i=1}^{100} X_i - 2.5 \times 100}{2 \times \sqrt{100}} > \frac{240 - 2.5 \times 100}{2 \times \sqrt{100}}\Big)$$
$$\approx \mathbb{P}\Big(Z > -\frac{1}{2}\Big) = \mathbb{P}(Z \leq 1/2),$$

where Z is the normal random variable.

We can calculate using *R*:

```
pnorm(0.5) = 0.6914625
```

So the probability is approximately 69%.

6.3.1 Normal approximation to the binomial distribution

Note that a binomial random variable Y with parameter n and p can be seen as a sum of indicator random variables X_i, where $X_i = 1$ if there was a success in trial i and $X_i = 0$ if there was a zero. So for large n, the CLT is applicable and the binomial distribution can be approximated by the normal distribution. Since $\mu = \mathbb{E}X_i = p$ and $\sigma^2 = \mathbb{V}\mathrm{ar}(X_i) = p(1-p)$, for large n the random variable

$$\frac{Y - np}{\sqrt{np(1-p)}}$$

is distributed as approximately standard normal.

This approximation is often used in various estimates related to public polls.

Example 6.3.3. General elections of a president are going to be conducted in Freedonia. Suppose that 48% of the population support Martha, 48% supports Rose, and 2% support Dr. Who.

A poll asks 400 random people who they support. What is the probability that at least 50% of those polled prefer Rose?

Let Y be the number of those in the poll who prefer Rose. Then we want to find

$$\begin{aligned}\mathbb{P}(\frac{Y}{400} > 0.5) &= \mathbb{P}(Y > 200) \\ &= \mathbb{P}\Big(\frac{Y - 0.48 \times 400}{\sqrt{400 \times 0.48 \times (1 - 0.48)}} > \frac{200 - 0.48 \times 400}{\sqrt{400 \times 0.48 \times (1 - 0.48)}}\Big) \\ &\approx \mathbb{P}(Z > 0.8006408)\end{aligned}$$

By using R we find that this probability is

```
1 - pnorm(0.8005408) = 0.2116988
```

So the probability is around 21%.

Continuity correction

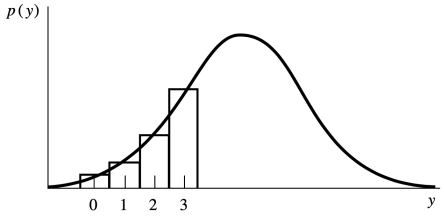


Figure 6.1: Normal approximation for binomial distribution with $n = 10$ and $p = 0.5$

Sometimes when n is not large it is useful to adjust the argument of the normal cdf when approximating binomial. For example, if we Y is binomial with $n = 10$ and $p = 0.5$, then direct calculation gives

$$\mathbb{P}(Y \leq 3) = pbinom(3, size = 10, p = 0.5) = 0.171875,$$

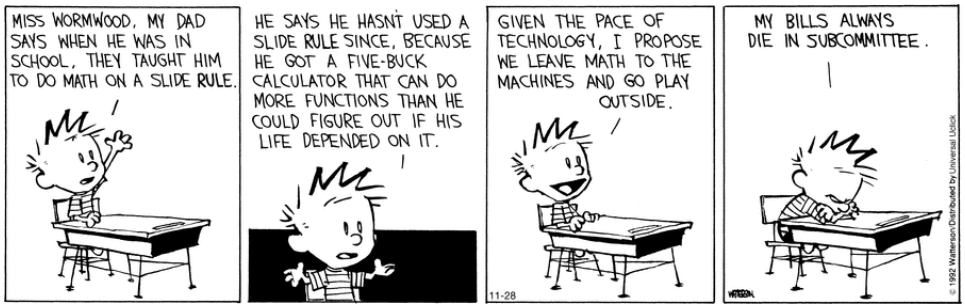
The normal approximation without correction gives

$$\begin{aligned}\mathbb{P}(Y \leq 3) &\approx \mathbb{P}\Big(Z \leq \frac{3 - 10 \times 0.5}{\sqrt{10 \times 0.5 \times (1 - 0.5)}}\Big) \\ &= pnorm(-1.264911) = 0.1029516,\end{aligned}$$

while the approximation with correction (note 3.5 instead of 3, with justification in Figure 6.1: for $x = 3.5$, the area under the curve gives a better approximation to the area of rectangles) gives

$$\begin{aligned}\mathbb{P}(Y \leq 3) &\approx \mathbb{P}\Big(Z \leq \frac{3.5 - 10 \times 0.5}{\sqrt{10 \times 0.5 \times (1 - 0.5)}}\Big) \\ &= pnorm(-0.9486833) = 0.1713909,\end{aligned}$$

which is very close to the true value.


MISS WORMWOOD, MY DAD SAYS WHEN HE WAS IN SCHOOL, THEY TAUGHT HIM TO DO MATH ON A SLIDE RULE.
HE SAYS HE HASN'T USED A SLIDE RULE SINCE, BECAUSE HE GOT A FIVE-BUCK CALCULATOR THAT CAN DO MORE FUNCTIONS THAN HE COULD FIGURE OUT IF HIS LIFE DEPENDED ON IT.
GIVEN THE PACE OF TECHNOLOGY, I PROPOSE WE LEAVE MATH TO THE MACHINES AND GO PLAY OUTSIDE.
MY BILLS ALWAYS DIE IN SUBCOMMITTEE.
11-28
© 1992 Watterson/Distributed by Universal Uclick