RAFAEL PEÑALOZA

PROBABILITY THEORY AND STATISTICS

FREE UNIVERSITY OF BOZEN-BOLZANO

Disclaimer

THIS MANUSCRIPT CONTAINS the lecture notes for the course on Probability Theory and Statistics at the Free University of Bozen-Bolzano. Most of the content is based on the Sheldon M. Ross's book.[1] There are currently five editions of that book; any of them should suffice for the needs of the course.[2]

This manuscript **does not** replace the material presented on the blackboard during the course. It may (in fact, it does) contain errors, typos, and miss relevant details discussed at the classroom.[3] Use it with care. The best advice is for you to attend the lecture and take notes by yourself, and use this text as supporting material.

Background information and other supplemental material like handouts, exercises, and code are provided at the OLE for the course. If you have not yet registered, **do it now**.

[1]

[2] The third edition is available for free at this location.

[3] Needless to say, all errors found here are my fault, and my fault only. If you notice any mistake, **please let me know** so that I can fix it.

Contents

Introduction to Probability Theory

PROBABILITY THEORY IS a means for measuring and speaking about the uncertainty about occurrences of events. The event will happen or not, but we do not know which will be the case. Suppose for example that I roll a die. The die will necessarily show one face, but we do not know which one; the event "the die will show a 6" is uncertain.

Probabilities *quantify* this uncertainty. In this case, we would say something like "the probability of the die showing a 6 is $1/6$," since there are six possible outcomes that we consider equally likely. There are two ways to understand this statement:

- if we repeat the experiment a large enough number of times, then in $1/6$ of them the outcome will be a 6;
- we believe that there is a 1 in 6 chance of this specific roll landing on 6.

The first is the *frequentist* view. This view suggests that a probability is intrinsic to the event, and it can be studied and determined by repeating an experiment. The second is the *subjective* (or *Bayesian*) view, where the probability refers to the belief of the agent stating the probability. Although these two views have some deep philosophical differences, they do not affect the study of the mathematical properties of probabilities. So we often disregard these differences when studying the theory of probabilities formally.[4]

Events

WE WILL DEAL with *experiments* whose outcome is uncertain, but where the set of all possible outcomes is known. This set of all possible outcomes is called the *sample space*, denoted by $\mathscr{S}$. When rolling a die, the sample space is $\mathscr{S} = \{1, \dots, 6\}$. Sample spaces may be very simple, or very complex. For example, in a race the sample space may describe all the possible orderings in which the participants can finish. Indeed, $\mathscr{S}$ may be infinite.

Any subset of the state space is called an *event*. If the outcome of the experiment is contained in a given event $\mathscr{E}$, then we say that $\mathscr{E}$ *occurred*. For example, the event of the die rolling an even number is $\mathscr{E} = \{2, 4, 6\}$.[5]

Since events are sets, we can define new events through set operations over other events. Given two events $\mathscr{E}$ and $\mathscr{F}$, their *union* ($\mathscr{E} \cup \mathscr{F}$), *intersection* ($\mathscr{E}\mathscr{F}$), and *complement* ($\overline{\mathscr{E}}$) are also events.[6] In our dice example,

[4] An accessible discussion on the Bayesian view, and its importance in modern statistics is available here.

[5] Obviously, more complex events are possible for other spaces.

[6] For a brief introduction to set theory and Venn diagrams, see the relevant handout.

if $\mathscr{E} = \{2,4,6\}$ is the event of rolling an even number, and $\mathscr{F} = \{2,3,5\}$ refers to the event of rolling a prime number, then

- $\mathscr{E} \cup \mathscr{F} = \{2,3,4,5,6\}$: rolling an even or a prime number;
- $\mathscr{E}\mathscr{F} = \{2\}$: rolling a prime even number; and
- $\overline{\mathscr{E}} = \{1,3,5\}$: rolling an odd number.

It is possible that an event is empty, for example by taking the intersection of two events that do not share any outcomes like $\{2\}$ and $\{3\}$. This event is denoted by $\emptyset$; e.g. $\{2\}\{3\} = \emptyset$. If $\mathscr{E}\mathscr{F} = \emptyset$, then $\mathscr{E}$ and $\mathscr{F}$ are *mutually exclusive*. Notice that the whole sample space is also an event, and that $\overline{\mathscr{S}} = \emptyset$.[7] Just as with sets, we can consider inclusions between events, too. $\mathscr{E}$ is *contained* in $\mathscr{F}$ if all outcomes in $\mathscr{E}$ are also in $\mathscr{F}$. In this case, we write $\mathscr{E} \subseteq \mathscr{F}$ or $\mathscr{F} \supseteq \mathscr{E}$. If $\mathscr{E} \subseteq \mathscr{F}$ and $\mathscr{F} \subseteq \mathscr{E}$, then $\mathscr{E}$ and $\mathscr{F}$ are *equivalent* ($\mathscr{E} \equiv \mathscr{F}$). For example, rolling a two is contained in rolling an even number, and is equivalent to rolling a prime even number.

Axioms of Probability

RECALL THAT THE FREQUENTIST view states that if we repeat an experiment repeatedly under the exact same conditions, then the proportion of times in which the outcome belongs to a given event $\mathscr{E}$ will converge to a constant, which reflects the probability of $\mathscr{E}$.

From a mathematical point of view, for every event $\mathscr{E}$ over a sample space $\mathscr{S}$ there is a number, called the *probability of $\mathscr{E}$* and denoted as $P(\mathscr{E})$ that satisfies the following three axioms:

Axiom 1 $0 \le P(\mathscr{E}) \le 1$;

Axiom 2 $P(\mathscr{S}) = 1$; and

Axiom 3 For any sequence of mutually exclusive events $\mathscr{E}_1, \mathscr{E}_2, \ldots,$[8] and for any $n \in \mathbb{N}$ it holds that $P(\bigcup_{i=1}^{n} \mathscr{E}_i) = \sum_{i=1}^{n} P(\mathscr{E}_i)$.

The first axiom says that the probability of any event is always a number between 0 and 1. Axiom 2 gives a completeness statement on probabilities: the probability of observing any outcome from the sample space is always 1. The last axiom shows how to aggregate the probabilities of mutually exclusive events: the probability of their union is always the sum of their individual probabilities.

The frequentist view of probabilities clearly satisfies these three axioms: for any event $\mathscr{E}$, the proportion (or frequency) of times in which a repeated experiment falls in an outcome from $\mathscr{E}$ is necessarily between 0 and 1; in every repetition of the experiment the outcome belongs to the sample space $\mathscr{S}$; and if two events $\mathscr{E}$ and $\mathscr{F}$ do not have any outcome in common, then the proportion of the time in which the outcome is in $\mathscr{E}$ or in $\mathscr{F}$ is the sum of their respective frequencies. As an example of this last axiom, suppose that $\mathscr{E}$ is the event where a roll of a die falls in an even number, and $\mathscr{F}$ that it falls on the number 1. Then, we expect $\mathscr{E}$ to hold half (50 percent) of the time, and $\mathscr{F}$ $\frac{1}{6}$-th of the time. Then $\mathscr{E} \cup \mathscr{F}$ should be observed approximately 66% of the time.

[7] In terms of language, we can see the union as a (non-exclusive) *or*, the intersection as an *and*, and the complement as a *no*.

[8] That is, events such that $\mathscr{E}_i\mathscr{E}_j = \emptyset$ whenever $i \neq j$.

These simple axioms allow us derive many properties of probabilities, and understand them formally. We start with two simple propositions that already showcase some interesting characteristics of the theory.

Proposition 1. *For every event* $\mathscr{E}$, $P(\overline{\mathscr{E}}) = 1 - P(\mathscr{E})$.

Proof. By definition, $\mathscr{E}$ and $\overline{\mathscr{E}}$ are disjoint, and $\mathscr{S} = \mathscr{E} \cup \overline{\mathscr{E}}$. Using Axiom 3, it follows that $P(\mathscr{S}) = P(\mathscr{E} \cup \overline{\mathscr{E}}) = P(\mathscr{E}) + P(\overline{\mathscr{E}})$. By Axiom 2 it also follows that $P(\mathscr{S}) = 1$, and hence $P(\overline{\mathscr{E}}) = 1 - P(\mathscr{E})$. □

In other words, the probability of an event *not* occurring is always one minus the probability of it occurring. For example, if the probability of rolling an even number in a die is 0.6, then the probability of rolling an odd number must be 0.4.

Notice that Axiom 3 provides a means for computing the probability of the union of two or more events, but only if they are mutually exclusive. We often need to compute such probabilities for events that are not necessarily disjoint. We can do that, as longs as we know the probability of the intersection.

Proposition 2. *For every two events* $\mathscr{E}, \mathscr{F}$, $P(\mathscr{E} \cup \mathscr{F}) = P(\mathscr{E}) + P(\mathscr{F}) - P(\mathscr{E}\mathscr{F})$.

Proof. To help in this proof, consider the Venn diagram in Figure 1. The three regions I, II, and III are mutually exclusive. Hence, it follows that

$$P(\mathscr{E} \cup \mathscr{F}) = P(\text{I}) + P(\text{II}) + P(\text{III});$$
$$P(\mathscr{E}) = P(\text{I}) + P(\text{II});$$
$$P(\mathscr{F}) = P(\text{II}) + P(\text{III}).$$

Thus, $P(\mathscr{E} \cup \mathscr{F}) = P(\mathscr{E}) + P(\mathscr{F}) - P(\text{II})$. Since $\text{II} \equiv \mathscr{E}\mathscr{F}$, this finishes the proof. □

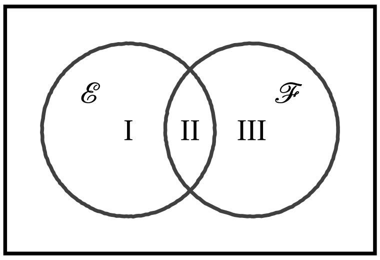


Figure 1: Three relevant regions in a Venn diagram.

Example 3. 23 percent of adults drink beer; 7 percent drink wine, and 2 percent drink both, beer and wine. What percentage of people drink neither beer nor wine?

Solution. To know how many people drink neither beer nor wine, it suffices to know how many people drink at least one of these beverages (the desired event is the complement of this one). Let $\mathscr{E}$ be the event of a person drinking beer, and $\mathscr{F}$ the event of drinking wine. The probability of drinking one of them is[9]

$$P(\mathscr{E} \cup \mathscr{F}) = P(\mathscr{E}) + P(\mathscr{F}) - P(\mathscr{E}\mathscr{F}) = 0.23 + 0.07 - 0.02 = 0.28.$$

The probability of a person being abstemious is $1 - 0.28 = 0.72$. △

ANOTHER IMPORTANT NOTION when speaking about uncertainty and probability is that of odds.

Definition 4. The *odds* of an event $\mathscr{E}$ is defined by $\frac{P(\mathscr{E})}{P(\overline{\mathscr{E}})} = \frac{P(\mathscr{E})}{1-P(\mathscr{E})}$.

Intuitively, the odds of an event $\mathscr{E}$ tells us how much more likely it is that $\mathscr{E}$ occurs than that it does not occur. If the odds of $\mathscr{E}$ is greater than 1, then it is more likely to see $\mathscr{E}$ than not; and *vice versa* if it is smaller than 1.

[9] To see this in a different way, think: how many people drink *only* beer? These are those that drink beer, minus those that drink beer and wine: $P(\mathscr{E}) - P(\mathscr{E}\mathscr{F})$. Those that drink *only* wine are $P(\mathscr{F}) - P(\mathscr{E}\mathscr{F})$. So, the people that drink beer or wine are those that drink only beer, plus those that drink only wine, plus those that drink both: $P(\mathscr{E}) - P(\mathscr{E}\mathscr{F}) + P(\mathscr{F}) - P(\mathscr{E}\mathscr{F}) + P(\mathscr{E}\mathscr{F})$.

For example, if rolling a 2 in a die has probability $P(\mathscr{E}) = \frac{1}{6}$, then the odds of this event is $\frac{1}{6}/\frac{5}{6} = \frac{1}{5}$. Consequently, it is five times more likely not to see a 2 than it is to roll the 2. In this case, we say that "the odds are 1 to 5 against the event $\mathscr{E}$."[10]

Uniformity

THERE ARE MANY CASES where it is natural to assume that every point in the sample space $\mathscr{S}$ is equally likely to occur. If the sample space is *finite*, (for example, if $\mathscr{S} = \{1, \dots, n\}$ for some $n \in \mathbb{N}$), then this means that

$$P(\{1\}) = \dots = P(\{n\}) = p.$$

Using Axioms 2 and 3, it follows that

$$1 = P(\mathscr{S}) = P(\{1\}) + \dots + P(\{n\}) = np,$$

and hence for every i, $P(\{i\}) = p = \frac{1}{n}$. That is, every outcome has the same probability of occurring determined by the total size of the sample space.

Axioms of probability:

1. $0 \leq P(\mathscr{E}) \leq 1$
2. $P(\mathscr{S}) = 1$
3. $P(\bigcup_{i=1}^{n} \mathscr{E}_i) = \sum_{i=1}^{n} P(\mathscr{E}_i)$ if $\mathscr{E}_i$s are mutually exclusive.

We can use Axiom 3 again to generalise this result to arbitrary events. For an event $\mathscr{E}$, let $\#\mathscr{E}$ denote the number of elements (outcomes) in $\mathscr{E}$. If all outcomes are equally likely to occur in a sample space with n elements,[11] then

$$P(\mathscr{E}) = \frac{\#\mathscr{E}}{n}.$$

That is, the probability of $\mathscr{E}$ is the proportion of elements in the sample space that belong to $\mathscr{E}$.

For that reason, it is often important to be able to *count* the number of different ways in which an event may occur. One very helpful rule for counting is the basic principle of counting.

Principle (Basic Principle of Counting). If two experiments are performed, such that Experiment 1 can result in any of m different outcomes, and for each of them Experiment 2 may result in any of n outcomes, then together there are mn different possible outcomes for the two experiments.

To show that this is the case, it suffices to enumerate all possible outcomes of the two experiments, arranging them in a matrix where every row corresponds to the outcome of Experiment 1, and every column to the outcome of Experiment 2. This matrix has overall mn elements (see Figure 2).

$$\begin{matrix} (1,1) & (1,2) & \dots & (1,n) \\ (2,1) & (2,2) & \dots & (2,n) \\ \vdots & \vdots & \ddots & \vdots \\ (m,1) & (m,2) & \dots & (m,n) \end{matrix}$$

Figure 2: Possible outcomes for the Basic Principle of Counting

Example 5. A drawer contains 8 black socks and 6 white socks. We "randomly" take two socks from the drawer. What is the probability of them *not* forming a pair?[12]

Solution. We first measure the size of the sample space: the possible outcomes of removing two socks from the drawer. There are 14 socks originally inside. First we can choose any of those 14 socks, and then any of the 13 remaining ones. Overall, there are $14 \cdot 13 = 182$ possible combinations. In order to *not* have a pair, we have to extract first a black one followed by a white one ($8 \cdot 6 = 48$ combinations) or first a white one followed by a black

[10] Or, dually, *in favour of* $\overline{\mathscr{E}}$.

[11] We often refer to this case as having a *uniform* sample space.

[12] That is, we want the probability of the two selected socks having different color.

one $6 \cdot 8 = 48$ combinations). Hence, assuming that each combination of the sample space is equally likely to occur,[13] we see that the probability is $\frac{48+48}{182} = \frac{96}{182} \approx 0.53$.

Obviously, the basic principle of counting can be generalised to cases where more than two experiments are performed. If r experiments are performed, such that Experiment 1 can result in any of n_1 different outcomes, and for each of them Experiment 2 may result in any of n_2 outcomes, and so on, then together there are $n_1 \cdot n_2 \cdots n_r$ different possible outcomes for the r experiments.

Consider for example the question of counting the number of ways in which a collection of n different objects can be organised over a line. For a small enough n, we can find this number through enumeration: all the linear ordering of the letters a, b, and c are abc, acb, bac, bca, cab, and cba; that is, there are 6 such orderings. We often call each of these orderings a *permutation*. There are six permutations on any set of three objects. Rather than enumerating them all, which becomes impossible (or at least tedious) as the number of objects grows, one can also count the number of permutations using the basic principle: the first object in the permutation may be any of 3 choices, the second any of the remaining 2 choices, and the last one is determined by the previous outcomes (that is, only one choice remains). In other words, there are $3 \cdot 2 \cdot 1 = 6$ permutations.

If instead of 3 we have n objects, then the same argument can be used to conclude that there are $n(n-1)(n-2)\cdots 3 \cdot 2 \cdot 1$ different permutations of them. This expression is known as *n factorial* and denoted by $n!$.[14]

Example 6. We want to organize in a bookshelf 10 books, which are divided in the following subjects: 4 are computer science, 3 are mathematics, 2 are statistics, and 1 is history. If we want all the books from the same subject to appear together, how many different arrangements are possible?

Solution. First we consider the number of possible different permutations of the subject matters: since there are 4 different subjects, we can choose any of 4! different orderings of them. Once that we have chosen the order of the subjects, we must choose the order of the books within the subject. The CS books allow 4! different ordering, the mathematics ones have 3! orders, statistics 2!, and the history book has only 1 = 1! ordering. Thus, overall there are $4!4!3!2!1! = 6{,}912$ possible ways to arrange the bookshelf keeping books from the same subject together.

Example 7. A course has 5 male and 3 female students. After an exam, all students are ranked according to their performance. Assuming that the scores of the exams are all different, (a) how many different rankings are possible? (b) If all rankings are equally likely, what is the probability of women getting the 3 highest scores?

Solution. (a) There are 8 people taking the test, so there are $8! = 40320$ possible rankings. (b) There are 3! rankings between the female students, and 5! rankings among males; thus through the basic principle of counting, there are 5!3! rankings where the first ranked are females. The desired probability is then $\frac{5!3!}{8!} = \frac{3 \cdot 2 \cdot 1}{8 \cdot 7 \cdot 6} = 1/56$.

[13] That is what we often mean when saying that an experiment is done "randomly."

[14] That is, $n! = n(n-1)(n-2)\cdots 3 \cdot 2 \cdot 1$. For the sake of completeness, we define $0! := 1$.

RATHER THAN ORDERING the objects in a collection with n elements, we sometimes want to count the number of groups with r objects that can be formed. For instance: how many groups of 3 elements can be formed with the five items A, B, C, D, and E? One way to solve this problem is the following. We have 5 ways to select the first element, 4 to select the second, and 3 to select the third. Thus, there are $5 \cdot 4 \cdot 3 = 60$ ways of selecting the group of 3 objects when the order in which they are selected is relevant. Notice, however, that if we are only interested in the groups of objects, then this method counts each group 6 times.[15] Thus, the total number of groups that can be formed (ignoring the ordering) is $\frac{5 \cdot 4 \cdot 3}{3 \cdot 2 \cdot 1} = 10$.

In general, there are $n(n-1)\cdots(n-r+1)$ different ways to obtain groups of r elements from a collection of n objects, if the order in which they are extracted is relevant. Since each group is counted $r!$ times through this method, when the order is not relevant there are

$$\frac{n(n-1)\cdots(n-r+1)}{r!} = \frac{n!}{(n-r)!r!}$$

groups of r elements from a collection of n objects.

Definition 8. Let $r \leq n$. The *combinations* of r objects in n is

$$\binom{n}{r} := \frac{n!}{(n-r)!r!}.$$

Again, $\binom{n}{r}$ is the number of groups of size r than can be extracted from a collection of n elements, when their ordering is not important; e.g., there are $\binom{7}{2} = \frac{7 \cdot 6}{2 \cdot 1} = 21$ different pairs that can be selected from a group of 7 people. Recall that $0! = 1$. Then $\binom{n}{0} = \binom{n}{n} = 1$.

Example 9. Five people are selected randomly from a group containing 5 men and 8 women. What is the probability that 3 women and 2 men are chosen?

Solution. By *random selection* we mean that each of the $\binom{13}{5}$ possible combinations is equally likely. There are $\binom{5}{2}$ possible combinations of 2 men, and $\binom{8}{3}$ of three women. From the basic principle of counting, the probability of this selection is then

$$\frac{\binom{5}{2}\binom{8}{3}}{\binom{13}{5}} = \frac{5 \cdot 4 \cdot 8 \cdot 7 \cdot 6 \cdot 5!}{13 \cdot 12 \cdot 11 \cdot 10 \cdot 9 \cdot 2 \cdot 3 \cdot 2} = \frac{8 \cdot 7 \cdot 5 \cdot 2}{13 \cdot 11 \cdot 9} = \frac{560}{1287}. \qquad \triangle$$

Example 10. If we randomly select k elements from a collection of n objects, what is the probability that a given object is among the k selected?

Solution. There is one way of choosing the selected item, and $\binom{n-1}{k-1}$ ways of selecting $k-1$ elements from the remaining objects in the collection. From the basic principle of counting, there are $1 \cdot \binom{n-1}{k-1}$ different subsets of k of the n elements that include the selected one. Since the total possible choices are $\binom{n}{k}$, the probability that a particular object is among the k selected is

$$\frac{\binom{n-1}{k-1}}{\binom{n}{k}} = \frac{(n-1)!(n-k)!k!}{(n-k)!(k-1)!n!} = \frac{k}{n}. \qquad \triangle$$

[15] For example, if the group selected is A, B, C, then it will count as different the cases where we extract ABC, and where we extract CBA.

Conditional Probability

ONE OF THE FUNDAMENTAL notions in probability theory is that of *conditional probabilities.* Conditional probabilities are a way of measuring uncertainty when some additional information (or evidence) is available: it allows us to update our beliefs based on some previous observations from the world. However, they are also useful by themselves, as one can often simplify a problem of computing a probability by considering the condition of a secondary event occurring or not.

To illustrate conditional probabilities, suppose that we roll a pair of dice. The sample space of this experiment is the set containing 36 outcomes $\mathscr{S} = \{(i,j) \mid 1 \leq i,j \leq 6\}$, where (i,j) refers to the outcome where the first die lands on i and the second on j. If each outcome is equally likely to occur, each of them has probability $1/36$.[16] Suppose that we observe that the first die landed on 3. What is the probability that the sum of the two dice is exactly 8?

Since we already know that the first die landed on 3, there are 6 possible outcomes remaining for our experiment: $(3,1),(3,2),(3,3),(3,4),(3,5)$, and $(3,6)$. In addition, as originally each of these outcomes was equally likely, the same should hold for this restricted setting; that is, given that the first die landed on 3, each of these outcomes now has probability $\frac{1}{6}$.[17] To conclude, the probability of the sum being 8 is then the probability of $(3,5)$, which is $\frac{1}{6}$.

Let $\mathscr{E}$ and $\mathscr{F}$ denote the event that the sum of the dice is 8, and the event that the first die lands on 3, respectively. Then, what we have just computed is called the *conditional probability* of $\mathscr{E}$ given $\mathscr{F}$, which is denoted by $P(\mathscr{E} \mid \mathscr{F})$.

To define the general notion of conditional probability, we follow the same intuition showcased by this example. Given that the event $\mathscr{F}$ holds, in order to observe $\mathscr{E}$ we in fact need to observe both events simultaneously; that is $\mathscr{E}\mathscr{F}$. However, since $\mathscr{F}$ is already known, we can consider it as the new sample space (all outcomes out of it are now impossible) and hence the probability of observing $\mathscr{E}$ becomes the probability of $\mathscr{E}\mathscr{F}$ relative to that of $\mathscr{F}$.

Definition 11. Let $\mathscr{E}$ and $\mathscr{F}$ be two events over the same sample space, such that $P(\mathscr{F}) > 0$. The *conditional probability of $\mathscr{E}$ given $\mathscr{F}$* is

$$P(\mathscr{E} \mid \mathscr{F}) = \frac{P(\mathscr{E}\mathscr{F})}{P(\mathscr{F})}$$

Notice that this definition only makes sense if $P(\mathscr{F}) > 0$. In fact, given our intuition that $\mathscr{F}$ is an observed evidence, it is reasonable to make such an assumption, as we do not expect to observe an event of probability 0.

Although we originally motivated conditional probabilities as a way to update beliefs in the presence of evidence,[18] Definition 11 is also consistent with the frequentist view. Suppose that an experiment is repeated a large number n of times. Since $P(\mathscr{F})$ expresses the proportion of times where $\mathscr{F}$ is observed, this event will occur approximately $nP(\mathscr{F})$ times. Likewise, $nP(\mathscr{E}\mathscr{F})$ times $\mathscr{E}$ and $\mathscr{F}$ will both happen. Thus, out of the ap-

[16] We say that the dice are *fair*.

[17] In fact, the (conditional) probability of all the other 30 events is now 0, since they cannot happen, given our evidence.

[18] Following the subjective view

proximately $nP(\mathscr{F})$ experiments whose outcome is in $\mathscr{F}$, $nP(\mathscr{E}\mathscr{F})$ also belong to $\mathscr{E}$. In other words, for the experiments in $\mathscr{F}$, the proportion whose outcome is also in $\mathscr{E}$ approximates $\frac{nP(\mathscr{E}\mathscr{F})}{nP(\mathscr{F})} = \frac{P(\mathscr{E}\mathscr{F})}{P(\mathscr{F})}$. This approximation gets tighter as the number of repetitions of the experiment increases.

Example 12. Suppose that a box contains 4 defective (do not work), 8 partially defective (work only briefly), and 20 working transistors. We choose one of them randomly, and use it. If it does not fail immediately, what is the probability that it is a working transistor?

Solution. Since the transistor did not fail immediately, we know that it is not one of the 4 defective ones. So we need to compute

$$P(\text{correct} \mid \overline{\text{defective}}) = \frac{P(\text{correct}, \overline{\text{defective}})}{P(\overline{\text{defective}})} = \frac{P(\text{correct})}{P(\overline{\text{defective}})},$$

where the last equation holds because every correct transistor is necessarily non-defective. Since the transistor was chosen randomly,[19] we have

$$P(\text{correct} \mid \overline{\text{defective}}) = \frac{\frac{20}{32}}{\frac{28}{32}} = \frac{5}{7}. \quad [20] \qquad \triangle$$

Example 13. You toss a coin twice, and on each throw you bet 5€ that it will fall on 'heads.' Knowing that you won at least one of the bets, what is the probability that you won 10€? (Assume that the coin is fair).

Solution. The sample space for the two coin tosses can be represented as $\mathscr{S} = \{(h,h),(h,t),(t,h),(t,t)\}$.[21] You win 10€ if both tosses land on heads; call this event $\mathscr{E}$. If $\mathscr{F}$ denotes the event that at least one toss is heads, then the desired probability is

$$P(\mathscr{E} \mid \mathscr{F}) = \frac{P(\mathscr{E}\mathscr{F})}{P(\mathscr{F})} = \frac{P(\{(h,h)\})}{P(\{(h,h),(h,t),(t,h)\})} = \frac{1/4}{3/4} = \frac{1}{3}. \qquad \triangle$$

The equation for conditional probabilities from Definition 11 can be rewritten as $P(\mathscr{E}\mathscr{F}) = P(\mathscr{E} \mid \mathscr{F})P(\mathscr{F})$.[22] This variation is helpful for computing probabilities of intersections of events, if the conditional probabilities are known.

Example 14. Your favourite football team has a 20% chance of reaching the final game, and if it does, then it has a 50% chance of winning. What is the probability of the team becoming the champion?

Solution. Let $\mathscr{F}$ be the event that the team reaches the final game, and $\mathscr{E}$ the event that it wins it. Then the probability of it being the champion is $P(\mathscr{E}\mathscr{F}) = P(\mathscr{E} \mid \mathscr{F})P(\mathscr{F}) = 0.5 \cdot 0.2 = 0.1$. $\triangle$

Bayes' Formula

FOR ANY TWO EVENTS $\mathscr{E}$ and $\mathscr{F}$, $\mathscr{E}$ can be equivalently expressed as a union $\mathscr{E}\mathscr{F} \cup \mathscr{E}\overline{\mathscr{F}}$ (see Figure 3). Indeed, any outcome in $\mathscr{E}$ must either be in $\mathscr{F}$ or in $\overline{\mathscr{F}}$, and hence also in the intersection of those events with $\mathscr{E}$. Since $\mathscr{E}\mathscr{F}$ and $\mathscr{E}\overline{\mathscr{F}}$ are mutually exclusive, Axiom 3 entails that

$$P(E) = P(\mathscr{E}\mathscr{F}) + P(\mathscr{E}\overline{\mathscr{F}}) = P(\mathscr{E} \mid \mathscr{F})P(\mathscr{F}) + P(\mathscr{E}\overline{\mathscr{F}})P(\overline{\mathscr{F}}). \qquad (1)$$

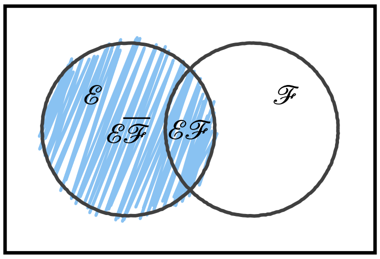


Figure 3: $\mathscr{E} = \mathscr{E}\mathscr{F} \cup \mathscr{E}\overline{\mathscr{F}}$.

[19] Each transistor is equally likely to be chosen.

[20] Notice that the same probability can be computed through counting: since the transistor is not defective, the problem reduces to computing the probability that a transistor, chosen at random from a box with 20 correct and 8 partially defective transistors, is correct. This is, of course, $\frac{20}{28}$.

[21] Where (h,t) represents that the first toss was heads, and the second tails.

[22] In fact, to avoid the need to require $P(\mathscr{F}) > 0$, conditional probabilities are often defined just using this variant.

This means that the probability of $\mathscr{E}$ is a weighted sum of the conditional probabilities over any other event $\mathscr{F}$ and its complement $\overline{\mathscr{F}}$, where the weights are given by the relative probabilities of the conditioning events to happen. This is very useful, because it allows to find probabilities by first conditioning relative to another event happening or not. There are many instances where this approach allows us to compute probabilities that would, otherwise, be impossible.

Example 15. There exist two kinds of people: cautious, and risk-takers. Within a one-year period, cautious people have an accident with probability of 0.2, while for risk-takers this probability is 0.4. If 30% of the population is risk-taker, what is the probability that a randomly chosen individual has an accident in this one-year period?

Solution. We condition over the chosen individual being a risk-taker or not. Let $\mathscr{E}$ be the event that the individual has an accident in the one-year period, and $\mathscr{F}$ the event that they are risk-takers.[23] Then,

$$P(\mathscr{E}) = P(\mathscr{E} \mid \mathscr{F})P(\mathscr{F}) + P(\mathscr{E} \mid \overline{\mathscr{F}})P(\overline{\mathscr{F}}) = 0.4 \cdot 0.3 + 0.2 \cdot 0.7 = 0.26 \qquad \triangle$$

THE FORMULA IN EQUATION (1) is useful not only for computing probabilities of events by themselves, but also to update initial probability assessments in the presence of additional or new information. We show how this works with an example.

Example 16. Consider again Example 15, and suppose that the randomly chosen individual has an accident. What is the probability that they are a risk-taker?°

Solution. Remember that, originally (lacking any further evidence) the individual has a 30% probability of being a risk-taker ($P(\mathscr{F}) = 0.3$). However, given the new information that they had an accident, we can re-evaluate this probability as follows:

$$P(\mathscr{F} \mid \mathscr{E}) = \frac{P(\mathscr{E}\mathscr{F})}{P(\mathscr{E})} = \frac{P(\mathscr{E} \mid \mathscr{F})P(\mathscr{F})}{P(\mathscr{E})} = \frac{0.3 \cdot 0.4}{0.26} = \frac{6}{13} \approx 0.4615. \qquad \triangle$$

THIS SAME FORMULA[24] can be generalised by allowing for a more fine-grained partition of the sample space as follows. Let $\mathscr{F}_1, \ldots, \mathscr{F}_n$ be a *partition* of the sample space $\mathscr{S}$.[25] In terms of events, this means that exactly one of these $\mathscr{F}_i$s must occur and, as before, $\mathscr{E} \equiv \bigcup_{i=1}^{n} \mathscr{E}\mathscr{F}_i$. Since these events are mutually exclusive, we get

$$P(\mathscr{E}) = \sum_{i=1}^{n} P(\mathscr{E}\mathscr{F}_i) = \sum_{i=1}^{n} P(\mathscr{E} \mid \mathscr{F}_i)P(\mathscr{F}_i). \tag{2}$$

That is, $P(\mathscr{E})$ can be computed by first conditioning over a partition of the sample space, and computing a weighted average of the results, where the weight corresponds to the probability of each of the events $\mathscr{F}_i$. Notice that the original formula is just a special case of this: $\mathscr{G}$ and $\overline{\mathscr{G}}$ form a partition.

Suppose that we know that $\mathscr{E}$ happened, and we are interested in knowing which of the events in the partition was observed. Using equation (2)

[23] $\overline{\mathscr{F}}$ means that they are cautious.

°Make the example of the medical test that seems very precise, but is not.

[24] In equation (1)

[25] $\mathscr{F}_1, \ldots, \mathscr{F}_n$ is a partition of the space $\mathscr{S}$ if they are mutually exclusive events and $\bigcup_{i=1}^{n} \mathscr{F}_i = \mathscr{S}$.

we get

$$P(\mathscr{F}_j \mid \mathscr{E}) = \frac{P(\mathscr{E}\mathscr{F}_j)}{P(\mathscr{E})} = \frac{P(\mathscr{E} \mid \mathscr{F}_j)P(\mathscr{F}_j)}{\sum_{i=1}^{n} P(\mathscr{E} \mid \mathscr{F}_j)P(\mathscr{F}_j)}. \tag{3}$$

This equation is known as *Bayes' formula*.[26] If we think of the events $\mathscr{F}_j$ as possible *hypotheses* about some subject of study, Bayes' formula can be interpreted as showing how opinions held about these hypotheses (i.e., $P(\mathscr{F}_j)$) should change on basis of the evidence of the experiment.[27]

Example 17. Consider a bug in a program, which may be in any of three different pieces of code with equal probability. For each of the three pieces i, let $1 - \alpha_i$ be the probability of finding the bug when searching in that piece given that it is, in fact, there.[28] If we search in the first piece, and we do not find any bug, what is the probability that the bug is in each of the three pieces $i \in \{1, 2, 3\}$?

Solution. Let $\mathscr{F}_i, 1 \le i \le 3$ be the event that the bug is in the i-th piece of code, and $\mathscr{E}$ the event that the search in the first piece was unsuccessful. Then we have:

$$P(\mathscr{F}_1 \mid \mathscr{E}) = \frac{P(\mathscr{E} \mid \mathscr{F}_1)P(\mathscr{F}_1)}{\sum_{i=1}^{3} P(\mathscr{E} \mid \mathscr{F}_i)P(\mathscr{F}_i)} = \frac{\alpha_1 \cdot (1/3)}{\alpha_1 \cdot (1/3) + 1 \cdot (1/3) + 1 \cdot (1/3)} = \frac{\alpha_1}{\alpha_1 + 2},$$

and for $j, 2 \le i \le 3$,

$$P(\mathscr{F}_j \mid \mathscr{E}) = \frac{P(\mathscr{E} \mid \mathscr{F}_j)P(\mathscr{F}_j)}{\sum_{i=1}^{3} P(\mathscr{E} \mid \mathscr{F}_i)P(\mathscr{F}_i)} = \frac{1 \cdot (1/3)}{\alpha_1 \cdot (1/3) + 1 \cdot (1/3) + 1 \cdot (1/3)} = \frac{1}{\alpha_1 + 2}.$$

For instance, if $\alpha_1 = 0.4$, then the conditional probability that the bug is in the first piece, given that we did not find it there is $\frac{1}{6}$. △

Independent Events

AS IT CAN BE seen from the previous examples, in general the conditional probability of an event $\mathscr{E}$ given $\mathscr{F}$ is not equal to the (unconditional) probability $P(\mathscr{E})$. That is, having evidence about $\mathscr{F}$ generally changes the likelihood of observing $\mathscr{E}$. In the special case where $P(\mathscr{E} \mid \mathscr{F}) = P(\mathscr{E})$, we say that $\mathscr{E}$ and $\mathscr{F}$ are *independent*. In other words, $\mathscr{E}$ and $\mathscr{F}$ are independent if knowledge about $\mathscr{F}$ does not change the probability of $\mathscr{E}$ occurring.

Recall that $P(\mathscr{E}\mathscr{F}) = P(\mathscr{E} \mid \mathscr{F})P(\mathscr{F})$. Thus, if $\mathscr{E}$ and $\mathscr{F}$ are independent we have $P(\mathscr{E}\mathscr{F}) = P(\mathscr{E})P(\mathscr{F})$. The converse implication holds too.

Definition 18. Two events $\mathscr{E}, \mathscr{F}$ are *independent* if $P(\mathscr{E}\mathscr{F}) = P(\mathscr{E})P(\mathscr{F})$. If $\mathscr{E}$ and $\mathscr{F}$ are not independent, they are *dependent*.

Example 19. A card is selected at random from a standard deck of 52 cards.[29] Let $\mathscr{E}$ be the event of selecting an ace, and $\mathscr{F}$ the event of selecting a red card. Then, $\mathscr{E}$ and $\mathscr{F}$ are independent because $P(\mathscr{E}\mathscr{F}) = \frac{1}{26}$, $P(\mathscr{E}) = \frac{4}{52}$, and $P(\mathscr{F}) = \frac{26}{52}$. △

Proposition 20. *If $\mathscr{E}$ and $\mathscr{F}$ are independent, then so are $\mathscr{E}$ and $\overline{\mathscr{F}}$.*[30]

[26] After Thomas Bayes.

[27] The former is known as the *prior* probability, and the latter is the *posterior*.

[28] α_i is the *overlook probability*: how likely it is to oversee the bug when looking in the right place? Some bugs are difficult to catch.

[29] French playing cards.

[30] That is, if $\mathscr{E}$ and $\mathscr{F}$ are independent, the likelihood of $\mathscr{E}$ is unchanged by information about $\mathscr{F}$, whether it holds or not.

Proof. Assume that $\mathscr{E}$ and $\mathscr{F}$ are independent. Then we have

$$P(\mathscr{E}) = P(\mathscr{E}\mathscr{F}) + P(\mathscr{E}\overline{\mathscr{F}}) = P(\mathscr{E})P(\mathscr{F}) + P(\mathscr{E}\overline{\mathscr{F}}),$$

where the last equality follows from the independence of $\mathscr{E}$ and $\mathscr{F}$. Then

$$P(\mathscr{E}\overline{\mathscr{F}}) = P(\mathscr{E})(1 - P(\mathscr{F})) = P(\mathscr{E})P(\overline{\mathscr{F}}). \qquad \square$$

We know then that independence is symmetric and closed under complementation. However, surprisingly, it is *not* closed under intersections. That is, if $\mathscr{E}$ is independent of $\mathscr{F}$ and of $\mathscr{G}$, it is not necessarily the case that $\mathscr{E}$ and $\mathscr{F}\mathscr{G}$ are independent.

Example 21. You throw two fair dice. Let $\mathscr{E}$ be the event that the sum of the two dice is 7, $\mathscr{F}$ the event that the first die is 1, and $\mathscr{G}$ the event that the second die is 6. It can be shown that $\mathscr{E}$ is independent of both $\mathscr{F}$ and $\mathscr{G}$. But, obviously, $\mathscr{E}$ is not independent of $\mathscr{F}\mathscr{G}$; in fact, $P(\mathscr{E} \mid \mathscr{F}\mathscr{G}) = 1$.

This example shows that defining independence between more than two events is more complex than merely checking that all pairs of events are independent between them. We get the following definition.

Definition 22. Three events $\mathscr{E}$, $\mathscr{F}$, and $\mathscr{G}$ are *independent* if:[31]

$$\begin{aligned} P(\mathscr{E}\mathscr{F}\mathscr{G}) &= P(\mathscr{E})P(\mathscr{F})P(\mathscr{G}), \\ P(\mathscr{E}\mathscr{F}) &= P(\mathscr{E})P(\mathscr{F}), \\ P(\mathscr{E}\mathscr{G}) &= P(\mathscr{E})P(\mathscr{G}), \qquad \text{and} \\ P(\mathscr{F}\mathscr{G}) &= P(\mathscr{F})P(\mathscr{G}). \end{aligned}$$

If $\mathscr{E}$, $\mathscr{F}$, and $\mathscr{G}$ are independent, then $\mathscr{E}$ will necessarily be independent of any event formed by combinations of $\mathscr{F}$ and $\mathscr{G}$. For example, $\mathscr{E}$ is independent of $\mathscr{F} \cup \mathscr{G}$ because

$$\begin{aligned} P(\mathscr{E}(\mathscr{F} \cup \mathscr{G})) = P(\mathscr{E}\mathscr{F} \cup \mathscr{E}\mathscr{G}) &= P(\mathscr{E}\mathscr{F}) + P(\mathscr{E}\mathscr{G}) - P(\mathscr{E}\mathscr{F}\mathscr{G}) \\ &= P(\mathscr{E})P(\mathscr{F}) + P(\mathscr{E})P(\mathscr{G}) - P(\mathscr{E})P(\mathscr{F})P(\mathscr{G}) \\ &= P(\mathscr{E})\left(P(\mathscr{F}) + P(\mathscr{G}) - P(\mathscr{F}\mathscr{G})\right) = P(\mathscr{E})P(\mathscr{F} \cup \mathscr{G}). \end{aligned}$$

Definition 22 is generalized to more than three events in the obvious way. The events $\mathscr{E}_1, \ldots, \mathscr{E}_n$ are *independent* iff for every subset $\mathscr{F}_1, \ldots, \mathscr{F}_m$, $m \le n$ of these events, it holds that $P(\mathscr{F}_1 \cdots \mathscr{F}_m) = P(\mathscr{F}_1) \cdots P(\mathscr{F}_m)$.

Considering many different independent events becomes important when we deal with a sequence of repetitions of an experiment. For example, if an experiment consists in rolling a die several times, we may see each roll of the die as a subexperiment, and it makes sense to assume that the outcomes of previous (or future) rolls have no effect on the outcome of the current roll.

Example 23. Consider the parallel system in Figure 4, which works if at least one of the n components work. Each each component i, independently of all others, works with probability $p_i, 1 \le i \le n$, what is the probability that the system functions?

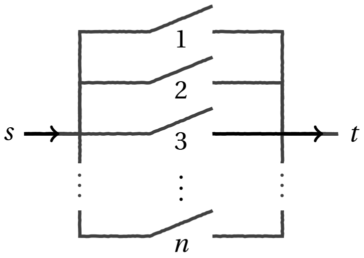


Figure 4: A parallel system with current flowing from s to t.

[31] Informally, this can be understood as every subset of events being mutually independent.

Solution. Let $\mathscr{E}$ be the event that the system functions, and $\mathscr{F}_i, 1 \leq i \leq n$ be the event that the component i functions. Then $P(\overline{\mathscr{E}})$ is exactly the probability that all components fail; that is

$$P(\mathscr{E}) = 1 - P(\overline{\mathscr{E}}) = 1 - P(\overline{\mathscr{F}_1} \cdots \overline{\mathscr{F}_n}) = 1 - \prod_{i=1}^{n} P(\overline{\mathscr{F}_i}),$$

where the last equality arises from the independence of the events.[32]

[32] The last expression can be rewritten to use the probabilities of the events, and not their complements, if needed.

Random Variables

Usually, when running experiments, we are only interested in some numerical values determined by the results. For example, after rolling two dice, we may be interested in their sum, but not on the specific individual die-values that led to that sum.[33] These quantities of interested are known as *random variables*.[34] Since the values of random variables are determined by the outcome of an experiment, they are associated with a probability degree.

Consider for example the random variable $\mathscr{X}$ given by the sum of two fair dice. $\mathscr{X}$ can take only values between 2 and 12, and the probability of each of these values is°

$$
\begin{aligned}
P([\mathscr{X}=2]) &= P(\{(1,1)\}) = \frac{1}{36}\\
P([\mathscr{X}=3]) &= P(\{(1,2),(2,1)\}) = \frac{2}{36}\\
P([\mathscr{X}=4]) &= P(\{(1,3),(2,2),(3,1)\}) = \frac{3}{36}\\
P([\mathscr{X}=5]) &= P(\{(1,4),(2,3),(3,2),(4,1)\}) = \frac{4}{36}\\
P([\mathscr{X}=6]) &= P(\{(1,5),(2,4),(3,3),(4,2),(5,1)\}) = \frac{5}{36}\\
P([\mathscr{X}=7]) &= P(\{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}) = \frac{6}{36}\\
P([\mathscr{X}=8]) &= P(\{(2,6),(3,5),(4,4),(5,3),(6,2)\}) = \frac{5}{36}\\
P([\mathscr{X}=9]) &= P(\{(3,6),(4,5),(5,4),(6,3)\}) = \frac{4}{36}\\
P([\mathscr{X}=10]) &= P(\{(4,6),(5,5),(6,4)\}) = \frac{3}{36}\\
P([\mathscr{X}=11]) &= P(\{(5,6),(6,5)\}) = \frac{2}{36}\\
P([\mathscr{X}=12]) &= P(\{(6,6)\}) = \frac{1}{36}.
\end{aligned}
$$

Since $\mathscr{X}$ must take some value, the sum of all these probabilities must be 1. This fact can be easily verified from this distribution.

Another random variable $\mathscr{Y}$ of interest can be the value of the first die. In this case, $\mathscr{Y}$ is equally likely to take any of the values from 1 to 6.°

These examples presented random variables taking finitely many different values. Random variables whose values can be described as a (finite or infinite) sequence $x_1, x_2, \ldots$ are called *discrete*. There are also random variables that can take a continuum of possible values; these are called *continuous*. For example, the weight of a person can take any value in

[33] If the sum is 7, we do not care which of the possible 6 ways to get this sum was actually observed.

[34] This is a bad choice of name, as they are not *random*, nor *variables*.

°Present only a few of these, and explain how they are obtained.

°Potential example to introduce indicator variables?

some interval (a,b).[35]

Definition 24. The *cumulative distribution function* (often called simply *distribution function*) of a random variable $\mathcal{X}$ is the function F defined for every real number x by:[36]

$$F(x) := P[\mathcal{X} \le x].$$

The notation $\mathcal{X} \sim F$ expresses that F is the distribution function of $\mathcal{X}$.

All probability questions about $\mathcal{X}$ can be answered in terms of its distribution function F. For example, to compute $P[a < \mathcal{X} \le b]$, we use the fact that $[\mathcal{X} \le b]$ can be decomposed into the two mutually exclusive events $[\mathcal{X} \le a]$ and $[a < \mathcal{X} \le b]$. Hence, $P[\mathcal{X} \le b] = P[\mathcal{X} \le a] + [a < \mathcal{X} \le b]$. From this, we can deduce

$$P[a < \mathcal{X} \le b] = F(b) - F(a).$$

Example 25. Consider the continuous random variable $\mathcal{X}$ with the distribution function

$$F(x) = \begin{cases} 0 & x \le 0 \\ 1 - \exp(-x^2) & x > 0 \end{cases}$$

The probability that $\mathcal{X}$ is greater than 1 is

$$P[\mathcal{X} > 1] = 1 - P[\mathcal{X} \le 1] = 1 - F(1) = \mathrm{e}^{-1} = 0.368. \qquad \triangle$$

Types of Random Variables

IF A RANDOM VARIABLE (RV) $\mathcal{X}$ is discrete, then we can define its *probability mass function* as $p(x) := P[\mathcal{X} = x]$. Since the variable is discrete, and $\mathcal{X}$ must take one value, we know that $p(x)$ is positive for at most countably many elements,[37] and that $\sum_{i=1}^{\infty} p(x_i) = 1$.

Example 26. Consider a RV $\mathcal{X}$ that can take values from $\{1,2,3\}$. If we know that $p(2) = \frac{1}{3}$ and $p(3) = \frac{1}{6}$, then we can deduce that

$$p(1) = 1 - p(2) - p(3) = 1 - \frac{1}{3} - \frac{1}{6} = \frac{1}{2}.$$

This function is shown in Figure 5. $\triangle$

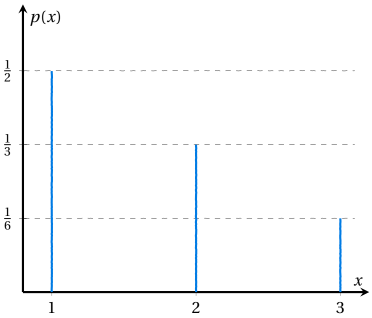


Figure 5: Probability mass function from Example 26.

The cumulative distribution function F can be computed in these cases from p by adding all the relevant values: $F(x) = \sum_{y \le x} p(y)$. Whenever $\mathcal{X}$ is a discrete RV whose possible values are $x_1 < x_2 < \ldots$, the distribution function F is a *step function*: it remains constant for the interval $[x_{i-1}, x_i)$, and then makes a jump of size $p(x_i)$ at x_i. For instance, if $\mathcal{X}$ has the probability mass function given in Example 26, then the cumulative distribution function of $\mathcal{X}$ is

$$F(x) = \begin{cases} 0 & x < 1 \\ \frac{1}{2} & 1 \le x < 2 \\ \frac{5}{6} & 2 \le x < 3 \\ 1 & 3 \le x, \end{cases}$$

as depicted in Figure 6.

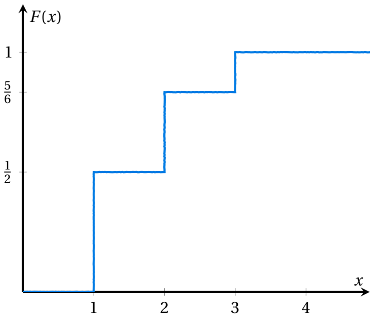


Figure 6: Cumulative distribution function from the RV of Example 26.

[35] There are also *mixed* random variables, but we will not consider them much during this lecture.

[36] For this course, we consider only random variables taking real numbers as values.

[37] At most those that can be taken by $\mathcal{X}$.

WE OFTEN NEED TO consider random variables that can take all the possible values from a (real) interval.

Definition 27. The RV $\mathscr{X}$ is *continuous* if there exists a non-negative function f defined over all real numbers $x \in \mathbb{R}$ such that for any set $B \subseteq \mathbb{R}$

$$P[\mathscr{X} \in B] = \int_B f(x)dx.$$

This function f is called the *probability density function* of $\mathscr{X}$.

Essentially, to compute the probability of an event defined by a set of values B, we find the integral of the probability density function, which is the intuitive generalization of the mass function for continuous variables. Since $\mathscr{X}$ must take some value in $(-\infty,\infty)$, $f(x)$ must satisfy

$$1 = P[\mathscr{X} \in (-\infty,\infty)] = \int_{-\infty}^{\infty} f(x)dx.$$

All probability statements can be answered integrating over $f(x)$; for instance, if $B = [a,b]$, then $P[a \le \mathscr{X} \le b] = \int_a^b f(x)dx$. In particular, if $a = b$ then we have $P[\mathscr{X} = a] = \int_a^a f(x)dx = 0$.[38] As an example, Figure 7 shows the probability density function

$$f(x) = \begin{cases} e^{-x} & x \ge 0 \\ 0 & x < 0, \end{cases}$$

and the area that represents $P[a < \mathscr{X} < b]$.

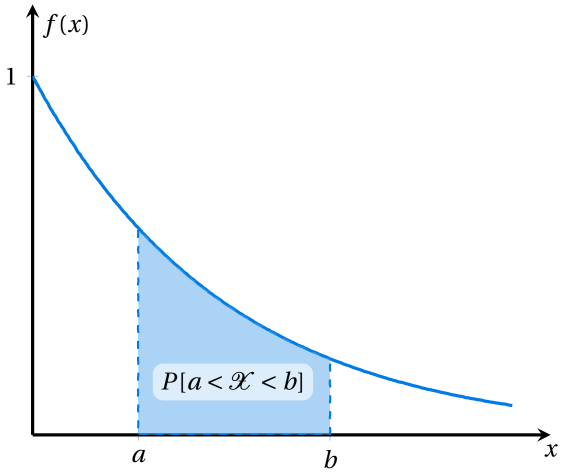


Figure 7: A probability density function, and the area representing $P[a < \mathscr{X} < b]$.

The probability density function f and the cumulative distribution F are related by the expression

$$F(a) = P[\mathscr{X} \le a] = \int_{-\infty}^{a} f(x)dx,$$

or equivalently, $\frac{d}{dx}F(x) = f(x)$. In other words, the density is the derivate of the cumulative distribution.

If the probability density function is continuous, then taking a small ε yields, from Definition 27,

$$P[a - \frac{\varepsilon}{2} \le \mathscr{X} \le a + \frac{\varepsilon}{2}] = \int_{a-\frac{\varepsilon}{2}}^{a+\frac{\varepsilon}{2}} f(x)dx \approx \varepsilon f(a).$$

[38] If $\mathscr{X}$ is continuous, then the probability of observing any specific value is always 0.

In other words, f tells us how likely it is to see a value that is close enough to a.

Example 28. Consider the continuous random variable $\mathscr{X}$ with the probability density function

$$f(x) = \begin{cases} c(4x - 2x^2) & 0 < x < 2 \\ 0 & \text{otherwise,} \end{cases}$$

for some constant c.

1. What is the value of c?

2. Compute $P[\mathscr{X} > 1]$.

Solution. Since f is a density function, it must hold that

$$1 = \int_{-\infty}^{\infty} f(x)dx = \int_0^2 c(4x - 2x^2)dx = c\left(2x^2 - \frac{2x^3}{3}\right)\Bigg|_0^2,$$

and hence $c = \frac{3}{8}$.

From this, we can conclude that

$$P[\mathscr{X} > 1] = \int_1^{\infty} f(x)dx = \frac{3}{8}\int_1^2 (4x - 2x^2)dx = \frac{1}{2}. \qquad \triangle$$

Joint Distributions

WHEN DEALING WITH EXPERIMENTS, we are often interested in the relationship between two or more random variables, rather than merely observing one. For example, for health policies, we may be interested in the relationship between hours spent sitting down and the incidence of back pain, or want to understand the relationship between working hours and productivity of employees.

To consider two random variables $\mathscr{X}$ and $\mathscr{Y}$ simultaneously, we need a *joint cumulative probability distribution function F* that specifies the probability of $\mathscr{X}$ and $\mathscr{Y}$ to be below a given value; that is,

$$F(x, y) = P[\mathscr{X} \le x, \mathscr{Y} \le y].$$

Knowing the joint probability distribution allows us to find the probabilities of different statements concerning with the variables $\mathscr{X}$ and $\mathscr{Y}$. For example, the distribution function $F_{\mathscr{X}}$ of $\mathscr{X}$ is obtained by not imposing any limit on the values of $\mathscr{Y}$; that is,

$$F_{\mathscr{X}}(x) = P[\mathscr{X} \le x] = P[\mathscr{X} \le x, \mathscr{Y} < \infty] = F(x, \infty).$$

Similarly, the cumulative distribution function of $\mathscr{Y}$ is $F_{\mathscr{Y}}(y) = F(\infty, y)$.[39]

For discrete random variables $\mathscr{X}$ and $\mathscr{Y}$ taking values $x_1, x_2, \ldots$, and $y_1, y_2, \ldots$, respectively, the *joint probability mass function p* of $\mathscr{X}$ and $\mathscr{Y}$ is defined in the obvious way: $p(x, y) = P[\mathscr{X} = x, \mathscr{Y} = y]$.

The individual probability mass functions of each of the variables can be easily obtained by eliminating the variable that is not of interest. E.g.,

[39] Formally, these would in fact be limits (i.e., $F_{\mathscr{X}}(x) = \lim_{y\to\infty} F(x, y)$, but this intuition serves to our purposes.

since $\mathscr{Y}$ must take some value y_j, the event $[\mathscr{X} = x]$ is equivalent to the union of the mutually exclusive events $[\mathscr{X} = x, \mathscr{Y} = y_j]$. Using Axiom 3 of probabilities, we have

$$P[\mathscr{X} = x] = P\left(\bigcup_j [\mathscr{X} = x, \mathscr{Y} = y_j]\right) = \sum_j P[\mathscr{X} = x, \mathscr{Y} = y_j] = \sum_j p(x, y_j).$$

Axioms of probability:

1. $0 \le P(\mathscr{E}) \le 1$
2. $P(\mathscr{S}) = 1$
3. $P(\bigcup_{i=1}^n \mathscr{E}_i) = \sum_{i=1}^n P(\mathscr{E}_i)$ if $\mathscr{E}_i$s are mutually exclusive.

Analogously, $P[\mathscr{Y} = y] = \sum_i p(x_i, y)$.

This says that the joint probability mass function fully determines the individual probability mass functions for each of the random variables involved. However, the converse does not hold: knowing $P[\mathscr{X} = x]$ and $P[\mathscr{Y} = y]$ is not sufficient for deriving $P[\mathscr{X} = x, \mathscr{Y} = y]$.

Example 29. Consider a box of batteries, where 2 are new, 3 are partially charged, and 4 are completely empty. We randomly select 3 batteries from this box. Let $\mathscr{X}$ and $\mathscr{Y}$ denote the number of new, and the number of partially charged batteries selected, respectively. The joint probability mass function $p(x, y) = P[\mathscr{X} = x, \mathscr{Y} = y]$ of $\mathscr{X}$ and $\mathscr{Y}$ is:

$$p(0,0) = \frac{\binom{4}{3}}{\binom{9}{3}} = \frac{4}{84} \qquad p(0,1) = \frac{\binom{3}{1}\binom{4}{2}}{\binom{9}{3}} = \frac{18}{84}$$

$$p(0,2) = \frac{\binom{3}{2}\binom{4}{1}}{\binom{9}{3}} = \frac{12}{84} \qquad p(0,3) = \frac{\binom{3}{3}}{\binom{9}{3}} = \frac{1}{84}$$

$$p(1,0) = \frac{\binom{2}{1}\binom{4}{2}}{\binom{9}{3}} = \frac{12}{84} \qquad p(1,1) = \frac{\binom{2}{1}\binom{3}{1}\binom{4}{1}}{\binom{9}{3}} = \frac{24}{84}$$

$$p(1,2) = \frac{\binom{2}{1}\binom{3}{2}}{\binom{9}{3}} = \frac{6}{84} \qquad p(2,0) = \frac{\binom{2}{2}\binom{4}{1}}{\binom{9}{3}} = \frac{4}{84}$$

$$p(2,1) = \frac{\binom{2}{2}\binom{3}{1}}{\binom{9}{3}} = \frac{3}{84}$$

A simpler way to express all these probabilities is shown in Table 1. △

Table 1: The joint mass function $p(x, y)$ from Example 29. The denominator 84 in the values is not included.

i \ j	0	1	2	3	Sum
0	4	18	12	1	35
1	12	24	6	0	42
2	4	3	0	0	7
Sum	20	45	18	1	

Notice that the probability mass function of $\mathscr{X}$ is obtained by summing the elements in a row, while the mass function of $\mathscr{Y}$ appears from summing through the columns. These probabilities are often known as the *marginal* probability mass functions of $\mathscr{X}$ and $\mathscr{Y}$, respectively.[40] To check the correctness of such a probability table, one should check that the sum of the marginal row and of the marginal column is 1.

Example 30. ° Consider a community where 15 percent of the families have no children, 20 percent have 1, 35 percent have 2, and 30 percent have 3. Suppose that each child is equally likely to be a boy or a girl. From a randomly chosen family, let $\mathscr{X}$ be the random variable measuring the number of boys, and $\mathscr{Y}$ the number of girls. The joint probability mass function is shown in Table 2. The cells in the first row are computed next.

°This example may be skipped.

$$\begin{aligned}
P[\mathscr{X} = 0, \mathscr{Y} = 0] &= P[\text{no children}] = 0.15 \\
P[\mathscr{X} = 0, \mathscr{Y} = 1] &= P[\text{1 child that is girl}] \\
&= P[\text{1 child}]P[\text{1 girl} \mid \text{1 child}] = 0.2 \cdot 0.5 = 0.1 \\
P[\mathscr{X} = 0, \mathscr{Y} = 2] &= P[\text{2 children that are girls}]
\end{aligned}$$

[40] To remember this name, think that they appear in the *margin* of the joint probability table.

$i \backslash j$	0	1	2	3	Sum
0	0.1500	0.1000	0.0875	0.0375	0.3750
1	0.1000	0.1750	0.1125	0	0.3875
2	0.0875	0.1125	0	0	0.2000
3	0.0375	0	0	0	0.0375
Sum	0.3750	0.3875	0.2000	0.0375	

Table 2: The joint mass $p(x,y)$ from Example 30.

$$= P[\text{2 children}]P[\text{2 girls} \mid \text{2 children}] = 0.35 \cdot (0.5)^2{=}0.0875$$

$$P[\mathscr{X} = 0, \mathscr{Y} = 3] = P[\text{3 children}]P[\text{3 girls} \mid \text{3 children}] = 0.3 \cdot (0.5)^3 = 0.0375.$$

From Table 2 it can be seen that, for example, the probability of having at least one girl is 0.625.[41] △

Definition 31. The RVs $\mathscr{X}$ and $\mathscr{Y}$ are *jointly continuous* if there exists a function $f(x,y)$ defined for all $x,y \in \mathbb{R}$ such that for every set $C \subseteq \mathbb{R} \times \mathbb{R}$[42]

$$P[(\mathscr{X},\mathscr{Y}) \in C] = \iint_{(x,y)\in C} f(x,y)dxdy.$$

The function $f(x,y)$ is called the *joint probability density function* of $\mathscr{X}$ and $\mathscr{Y}$.

If $A, B \subseteq \mathbb{R}$ are two sets of real numbers, then it follows from this definition that

$$P[\mathscr{X} \in A, \mathscr{Y} \in B] = \int_B \int_A f(x,y)dxdy.$$

From $F(a,b) = P[\mathscr{X} \le a, \mathscr{Y} \le b] = \int_{-\infty}^{b}\int_{-\infty}^{a} f(x,y)dxdy$, it follows (via differentiation) that whenever the partial derivatives are defined,

$$f(a,b) = \frac{\partial^2}{\partial a \partial b} F(a,b).$$

As in the case of single random variables, $f(a,b)$ is a measure of how likely it is for the random vector $(\mathscr{X},\mathscr{Y})$ be appear near (a,b); but the probability of being *exactly* (a,b) remains 0.

Notice that if $\mathscr{X}$ and $\mathscr{Y}$ are jointly continuous, then each of them is continuous individually. Moreover, the probability density function of $\mathscr{X}$ is $f_{\mathscr{X}}(x) := \int_{-\infty}^{\infty} f(x,y)dy$. Hence,

$$P[\mathscr{X} \in A] = P[\mathscr{X} \in A, \mathscr{Y} \in \mathbb{R}] = \int_A \int_{-\infty}^{\infty} f(x,y)dydx = \int_A f_{\mathscr{X}}(x)dx.$$

Example 32. Consider the joint density function for $\mathscr{X}$ and $\mathscr{Y}$

$$f(x,y) = \begin{cases} 2e^{-x}e^{-2y} & x,y > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Compute $P[\mathscr{X} > 1, \mathscr{Y} < 1]$, $P[\mathscr{X} < \mathscr{Y}]$, and $P[\mathscr{X} < a]$.[43]

Solution.

$$\begin{aligned} P[\mathscr{X} > 1, \mathscr{Y} < 1] &= \int_0^1 \int_1^{\infty} 2e^{-x}e^{-2y}dxdy \\ &= \int_0^1 2e^{-2y}(-e^{-x}|_1^{\infty})dy = e^{-1}\int_0^1 2e^{-2y}dy \end{aligned}$$

[41] Notice that the table is symmetric, and so the probability of having at least one boy is exactly the same in this case.

[42] That is, for every set in the two-dimensional plane.

[43] Recall that $\int e^{kx}dx = (1/k)e^{kx} + c$.

$$
\begin{aligned}
&= e^{-1}(-e^{-2y}|_0^1) = e^{-1}(1-e^{-2}) \\
P[\mathscr{X} < \mathscr{Y}] &= \iint_{(x,y)|x<y} 2e^{-x}e^{-2y}dxdy = \int_0^\infty \int_0^y 2e^{-x}e^{-2y}dxdy \\
&= \int_0^\infty 2e^{-2y}(1-e^{-y})dy = \int_0^\infty 2e^{-2y}dy - \int_0^\infty 2e^{-3y}dy \\
&= 1 - \frac{2}{3} = \frac{1}{3} \\
P[\mathscr{X} < a] &= \int_0^a \int_0^\infty 2e^{-x}e^{-2y}dydx = \int_0^a e^{-x}dx = 1-e^{-a} \qquad \triangle
\end{aligned}
$$

Independent Random Variables

THE RANDOM VARIABLES $\mathscr{X}$ and $\mathscr{Y}$ are *independent* if for any two sets $A, B \subseteq \mathbb{R}$ it holds that $P[\mathscr{X} \in A, \mathscr{Y} \in B] = P[\mathscr{X} \in A]P[\mathscr{Y} \in B]$. That is, they are independent if for all possible sets A, B, the events $\mathscr{E}_A = [\mathscr{X} \in A]$ and $\mathscr{F}_B = [\mathscr{Y} \in B]$ are independent. This definition of independence is equivalent to requiring that for all $a, b \in \mathbb{R}$

$$P[\mathscr{X} \le a, \mathscr{Y} \le b] = P[\mathscr{X} \le a]P[\mathscr{Y} \le b],$$

or, alternatively, that $F(a,b) = F_{\mathscr{X}}(a)F_{\mathscr{Y}}(b)$.[44]

If $\mathscr{X}$ and $\mathscr{Y}$ are discrete, independence is equivalent to requiring, for all $x, y \in \mathbb{R}$ that $p(x,y) = p_{\mathscr{X}}(x)p_{\mathscr{Y}}(y)$, where $p_{\mathscr{X}}$ and $p_{\mathscr{Y}}$ are the probability mass functions. In the continuous case, it refers to $f(x,y) = f_{\mathscr{X}}(x)f_{\mathscr{Y}}(y)$.[45]

Example 33. Let $\mathscr{X}$ and $\mathscr{Y}$ be two independent random variables, each having the density function

$$f(x) = \begin{cases} e^{-x} & x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Find the density function of $\mathscr{X}/\mathscr{Y}$.

Solution. We first compute the distribution function $F_{\mathscr{X}/\mathscr{Y}}$ of $\mathscr{X}/\mathscr{Y}$. Given $a > 0$,

$$
\begin{aligned}
F_{\mathscr{X}/\mathscr{Y}}(a) = P[\mathscr{X}/\mathscr{Y} \le a] &= \iint_{x/y \le a} f(x,y)dxdy \\
&= \iint_{x/y \le a} e^{-x}e^{-y}dxdy = \int_0^\infty \int_0^{ay} e^{-x}e^{-y}dxdy \\
&= \int_0^\infty (1-e^{-ay})e^{-y}dy = \left(-e^{-y} + \frac{e^{-(a+1)y}}{a+1}\right)\Bigg|_0^\infty = 1 - \frac{1}{a+1}.
\end{aligned}
$$

Differentiating this function over the variable a, we obtain the density function $f_{\mathscr{X}/\mathscr{Y}}(a) = 1/(a+1)^2$, for $0 < a < \infty$. $\triangle$

WE CAN GENERALIZE the notion of joint probability distributions from 2 to n random variables, defining the analogous notion of the joint cumulative probability distribution F, the joint probability mass function p if they are discrete, the joint probability density function f if they are continuous, and independence if every finite subcollection of random variables is independent.

[44] To prove this, notice that one direction corresponds directly to the definition. The other direction follows from the axioms of probability, with caveats of which sets are allowed.

[45] As for events, random variables that are not independent are called *dependent*.

Example 34. Consider a stock, whose price changes every day independently with the probability mass function $p(x)$ of the stock changing by x defined as

$$p(x) = \begin{cases} 0.05 & x \in \{-3,3\} \\ 0.10 & x \in \{-2,2\} \\ 0.15 & x \in \{-1,1\} \\ 0.40 & x = 0. \end{cases}$$

If $\mathscr{X}_i$ denotes the change in day i, then the probability that it increases in the next days by 1, 2, and 0 is

$$P[\mathscr{X}_1 = 1, \mathscr{X}_2 = 2, \mathscr{X}_3 = 0] = p(1)p(2)p(0) = 0.15 \cdot 0.1 \cdot 0.4 = 0.006. \qquad \triangle$$

Expectation

A FUNDAMENTAL NOTION IN probability theory is the expectation of a random variable. If the random variable $\mathscr{X}$ is discrete and takes values $x_1, x_2, \ldots$, then the *expectation* or *expected value* of $\mathscr{X}$ is

$$E[\mathscr{X}] := \sum_i x_i P[\mathscr{X} = x_i] = \sum_i x_i p(x_i).$$

That is, $E[\mathscr{X}]$ is the weighted average of the values of $\mathscr{X}$, where the weights are given by their probability of occurrence. For example, if the probability mass function of $\mathscr{X}$ is such that $p(0) = \frac{1}{2} = p(1)$, then $E[\mathscr{X}] = 0\frac{1}{2} + 1\frac{1}{2} = \frac{1}{2}$,[46] but if $p(0) = \frac{1}{3}$ and $p(1) = \frac{2}{3}$, then $E[\mathscr{X}] = 0\frac{1}{3} + 1\frac{2}{3} = \frac{2}{3}$. This follows from the fact that the value 1 is twice as likely to appear as the value 0.

An intuitive motivation for the notion of expectation arises from the frequentist view. If an experiment is repeated independently, then the proportion of times we observe a given event $\mathscr{E}$ is $P(\mathscr{E})$. Consider a random variable $\mathscr{X}$ taking values $x_1, x_2, \ldots$, which represents the winnings of one execution of a bet, and let p be its mass function.[47] Suppose that we repeat the bet n times, for n sufficiently large. Then, in approximately $np(x_i)$ of those times, we win x_i units. Since this is true for all i, overall we win $\sum_i x_i \cdot np(x_i)$. Thus, in average, on each bet, we will win

$$\sum_i \frac{x_i \cdot np(x_i)}{n} = \sum_i x_i p(x_i) = E[\mathscr{X}].$$

Example 35. If $\mathscr{X}$ is the outcome of rolling a fair die, then $p(i) = 1/6$ for all $i, 1 \le i \le 6$. Hence, we get

$$E[\mathscr{X}] = 1\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right) + 3\left(\frac{1}{6}\right) + 4\left(\frac{1}{6}\right) + 5\left(\frac{1}{6}\right) + 6\left(\frac{1}{6}\right) = \frac{21}{6} = \frac{7}{2}. \qquad \triangle$$

Notice from this example that the expectation of $\mathscr{X}$ might be a value that can never occur in $\mathscr{X}$. The expected value is thus not what we will likely observe, but rather the average of the results over a long run of repetitions of the experiment.[48]

The *indicator variable* for an event $\mathscr{E}$ is a random variable that takes value 1 if $\mathscr{E}$ occurs, and 0 otherwise. If $\mathscr{X}$ is the indicator variable of $\mathscr{E}$, then $E[\mathscr{X}] = 1P(\mathscr{E}) + 0P(\overline{\mathscr{E}}) = P(\mathscr{E})$. That is, the expected value of an indicator variable is exactly the probability of the event it indicates.

[46] The usual average of the values of $\mathscr{X}$.

[47] That is, with probability $p(x_i)$, we win x_i units.

[48] If we roll a die many times, the average of the results will converge to 3.5.

ONE CAN ALSO DEFINE the expectation for continuous random variables. Similarly to the discrete case, for a continuous random varable $\mathscr{X}$ with probability density function f, we have

$$E[\mathscr{X}] := \int_{-\infty}^{\infty} x f(x) dx.$$

Example 36. Your IKEA order will arrive at some point after 17:00. You know by experience that the number of hours $\mathscr{X}$ that you have to wait for the arrival is a random variable with the probability density function

$$f(x) = \begin{cases} \frac{1}{2} & 0 < x < 2 \\ 0 & \text{otherwise.} \end{cases}$$

The expected amount of time that you will wait is then

$$E[\mathscr{X}] = \int_{0}^{2} \frac{x}{2} dx = 1.$$

That is, on average, you have to wait one hour. △

The notion of expectation can be intuitively understood as that of the center of gravity. Suppose, for a discrete random variable $\mathscr{X}$, that the probability mass distribution is interpreted literally, with a physical object with mass proportional to $p(x_i)$ located at each point x_i. Then, $E[\mathscr{X}]$ is the point where such a structure would balance. See Figure 8. It is also important to notice that for every random variable $\mathscr{X}$, $E[\mathscr{X}]$ has the same units of measurement as $\mathscr{X}$.

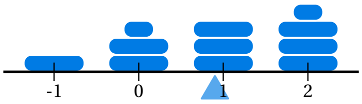


Figure 8: Expected value as the center of gravity.

Properties of Expectation

CONSIDER A GIVEN RANDOM variable $\mathscr{X}$ with its probability distribution. Often, we are not interested in $E[\mathscr{X}]$, but rather on a *function* of $\mathscr{X}$; say $g(\mathscr{X})$. Notice that $g(\mathscr{X})$ is itself a random variable, and hence has a probability distribution that must depend on the distribution of $\mathscr{X}$. If we can compute this distribution, then $E[g(\mathscr{X})]$ can be readily obtained from it.

Example 37. Let $\mathscr{X}$ be the random variable with probability mass function $p(0) = 0.3$, $p(1) = 0.5$, and $p(2) = 0.2$. Compute $E[\mathscr{X}^2]$.

Solution. If $\mathscr{Y} = \mathscr{X}^2$, then $\mathscr{Y}$ can take the values $0 = 0^2, 1 = 1^2$ and $4 = 2^2$ with probabilities $0.3, 0.5$, and 0.2, respectively. Then

$$E[\mathscr{X}^2] = E[\mathscr{Y}] = 0 \cdot 0.3 + 1 \cdot 0.5 + 4 \cdot 0.2 = 1.3. \qquad \triangle$$

Example 38. The time that it takes to correct a bug in a piece of software is a random variable $\mathscr{X}$ with density function

$$f(x) = \begin{cases} 1 & 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

If the cost of keeping a bug for time x is x^3, what is the expected cost of such a bug?

Solution. Let $\mathscr{Y} = \mathscr{X}^3$ be the random variable for the cost. The distribution $F_{\mathscr{Y}}$ of $\mathscr{Y}$ is given, for every $0 \le a \le 1$ by

$$F_{\mathscr{Y}}(a) = P[\mathscr{Y} \le a] = P[\mathscr{X}^3 \le a] = P[\mathscr{X} \le a^{1/3}] = \int_0^{a^{1/3}} 1dx = a^{1/3}.$$

Through differentiation, we get the density $f_{\mathscr{Y}}(a) = \frac{1}{3}a^{-2/3}$ for $0 \le a < 1$. Hence,

$$\begin{aligned} E[\mathscr{X}^3] = E[\mathscr{Y}] &= \int_{-\infty}^{\infty} a f_{\mathscr{Y}}(a) da \\ &= \int_0^1 a\frac{1}{3}a^{-2/3} da = \frac{1}{3}\int_0^1 a^{1/3} da \\ &= \frac{1}{3}\frac{3}{4}\, a^{4/3}\Big|_0^1 = 1/4 \end{aligned}$$

△

As we can see, this approach always allows us to compute the expected value of any function of $\mathscr{X}$, if we know the distribution of $\mathscr{X}$. However, the process is far from obvious. An easy way to do compute this expectation is based on the following intuition: since $g(\mathscr{X})$ takes the value $g(x)$ whenever $\mathscr{X} = x$, $E[g(\mathscr{X})]$ should be the weighted average of the values of $g(x)$ with the weights given by $p(x)$ for each possible value x.[49]

Proposition 39 (Expectation of Functions)**.** *Let $\mathscr{X}$ be a random variable, and g any real-valued function. Then:*

1. *if $\mathscr{X}$ is discrete with probability mass function $p(x)$, then*

$$E[g(\mathscr{X})] = \sum_x g(x)p(x);$$

2. *if $\mathscr{X}$ is continuous with probability density function $f(x)$, then*

$$E[g(\mathscr{X})] = \int_{-\infty}^{\infty} g(x)p(x)dx.$$

Corollary 40. *For any two constants $a, b \in \mathbb{R}$, $E[a\mathscr{X} + b] = aE[\mathscr{X}] + b$.*

Proof. We show here the discrete case.[50]

$$E[a\mathscr{X} + b] = \sum_x (ax+b)p(x) = a\sum_x xp(x) + b\sum_x p(x) = aE[\mathscr{X}] + b \qquad \square$$

Notice in particular that $E[b] = b$ and $E[a\mathscr{X}] = aE[\mathscr{X}]$.

The expected value of $\mathscr{X}$ is also known as the *mean* or the *first moment* of $\mathscr{X}$. In general, the *n-th moment* of $\mathscr{X}$, for any $n \in \mathbb{N}$, is defined as $E[\mathscr{X}^n]$, which we know how to compute by Proposition 39.

PROPOSITION 39 CAN BE EXTENDED to deal with several random variables. In the case of two random variables $\mathscr{X}$ and $\mathscr{Y}$, if g is a function over pairs of real numbers, then

$$E[g(\mathscr{X},\mathscr{Y})] = \begin{cases} \sum_y \sum_x g(x,y)p(x,y) & \mathscr{X},\mathscr{Y} \text{ discrete} \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f(x,y)dxdy & \mathscr{X},\mathscr{Y} \text{ jointly continuous.} \end{cases}$$

In particular, if $g(\mathscr{X},\mathscr{Y}) = \mathscr{X} + \mathscr{Y}$, where $\mathscr{X},\mathscr{Y}$ are jointly continuous, we get

$$E[\mathscr{X} + \mathscr{Y}] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x+y)f(x,y)dxdy$$

[49] This intuition holds in general. We can verify that the following proposition holds for the previous two examples.

[50] The continuous case is left as an exercise to the interested student.

$$
\begin{aligned}
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f(x,y)\,dx\,dy + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y f(x,y)\,dx\,dy \\
&= E[\mathscr{X}] + E[\mathscr{Y}], \qquad (\dagger)
\end{aligned}
$$

where (†) is a consequence of using $g_{\mathscr{X}}(\mathscr{X},\mathscr{Y}) = \mathscr{X}$; that is

$$E[\mathscr{X}] = E[g_{\mathscr{X}}(\mathscr{X},\mathscr{Y})] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f(x,y)\,dx\,dy,$$

and similarly for $g_{\mathscr{Y}}(\mathscr{X},\mathscr{Y}) = \mathscr{Y}$.

Since sums are associative, we can simply repeat this argument to obtain the general form for sums of random variables: for any $n \in \mathbb{N}$,

$$E\left[\sum_{i=1}^{n} \mathscr{X}_i\right] = \sum_{i=1}^{n} E[\mathscr{X}_i].$$

Example 41. Two fair dice are rolled. Find the expected value of their sum.

Solution. If $\mathscr{X}$ is the random variable representing the sum, we know that we can compute $E[\mathscr{X}] = \sum_{i=1}^{12} iP[\mathscr{X} = i]$. But a simpler approach is to consider two variables $\mathscr{Y}_1$ and $\mathscr{Y}_2$, which represent the result of each individual die. Then $\mathscr{X} = \mathscr{Y}_1 + \mathscr{Y}_2$, and $E[\mathscr{X}] = E[\mathscr{Y}_1] + E[\mathscr{Y}_2] = 7$.[51]

Example 42. ° You have n different pairs of shoes, which are all mixed up in your closet. If you pair them completely at random (selecting one left and one right shoe), what is the expected number of correct pairs you get?

°This example may be skipped.

Solution. If $\mathscr{X}$ is the number of correct pairs, we can compute $E[\mathscr{X}]$ via the random variables $\mathscr{X}_1, \ldots, \mathscr{X}_n$, where for each $i, 1 \le i \le n$,

Indicator variables.

$$\mathscr{X}_i = \begin{cases} 1 & i\text{-th pair is correct} \\ 0 & \text{otherwise.} \end{cases}$$

Then, $\mathscr{X} = \sum_{i=1}^{n} \mathscr{X}_i$. Since the i-th left shoe is equally likely to be matched with any of the n right shoes, it follows that $P[\mathscr{X}_i = 1] = 1/n$,[52] and hence $E[\mathscr{X}_i] = 1/n$. From this, we obtain $E[\mathscr{X}] = \sum_{i=1}^{n} E[\mathscr{X}_i] = n(1/n) = 1$.[53] △

AN IMPORTANT PROPERTY of the mean arises when trying to predict the value of a random variable $\mathscr{X}$. If we predict that $\mathscr{X}$ will take the value c (for example, when placing a bet), then the *square of the prediction error* is $(\mathscr{X} - c)^2$. Let $\mu = E[\mathscr{X}]$; then,

$$
\begin{aligned}
E[(\mathscr{X} - c)^2] = E[(\mathscr{X} - \mu + \mu - c)^2] &= E[(\mathscr{X} - \mu)^2 + 2(\mathscr{X} - \mu)(\mu - c) + (\mu - c)^2] \\
&= E[(\mathscr{X} - \mu)^2] + 2(\mu - 2)E[\mathscr{X} - \mu] + (\mu - c)^2 \\
&= E[(\mathscr{X} - \mu)^2] + (\mu - c)^2 \qquad (\dagger) \\
&\ge E[(\mathscr{X} - \mu)^2],
\end{aligned}
$$

where (†) follows from $E[\mathscr{X} - \mu] = E[\mathscr{X}] - \mu = 0$. Thus, the average squared error is minimised when we predict that $\mathscr{X}$ is equal to its mean μ.[54]

Variance

[51] See Example 35.

[52] Probability of the i-th left shoe matching its right one.

[53] This means that regardless of how many pairs we start with, in average only one pair will be correct.

[54] The best predictor of a random variable, in terms of minimizing the expected square of its error, is just the mean.

THE EXPECTED VALUE OF a random variable $\mathscr{X}$ provides the weighted average of the values it may take, but this information does not tell us much about the *distribution* of the values; e.g., how far apart are the values from each other.[55] For example, the random variables $\mathscr{X}$ and $\mathscr{Y}$ with $p_{\mathscr{X}}(0) = 1$ and $p_{\mathscr{Y}}(-100) = p_{\mathscr{Y}}(100) = 0.5$, respectively, have both the same expectation; namely 0. However, the values of $\mathscr{Y}$ are much more separated between themselves than those of $\mathscr{X}$, which is a constant.

In order to measure the variation of the values of $\mathscr{X}$, we can try to see how far is $\mathscr{X}$ from its mean μ in average; that is, $E[|\mathscr{X} - \mu|]$. However, dealing with absolute values is often problematic mathematically.[56] For that reason we consider the expected squared difference between them.

Definition 43. If $\mathscr{X}$ is a random variable with $E[\mathscr{X}] = \mu$, then the *variance* of $\mathscr{X}$ is $Var(\mathscr{X}) := E[(\mathscr{X} - \mu)^2]$.

Notice that

$$\begin{aligned} Var(\mathscr{X}) &= E[(\mathscr{X}-\mu)^2] = E[\mathscr{X}^2 - 2\mu\mathscr{X} + \mu^2] = E[\mathscr{X}^2] - 2\mu E[\mathscr{X}] + \mu^2 \\ &= E[\mathscr{X}^2] - \mu^2 = E[\mathscr{X}^2] - (E[\mathscr{X}])^2. \end{aligned}$$

This is often a simpler way to compute the variance.

Example 44. Let $\mathscr{X}$ represent the outcome of rolling a fair die. We know that $P[\mathscr{X} = i] = 1/6$ for $1 \le i \le 6$. Then, °

$$E[\mathscr{X}^2] = \frac{1}{6}\sum_{i=1}^{6} i^2 = \frac{1}{6}(1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2) = \frac{91}{6}.$$

We have also seen that $E[\mathscr{X}] = 7/2$. Thus,

$$Var(\mathscr{X}) = E[\mathscr{X}^2] - (E[\mathscr{X}])^2 = \frac{91}{6} - \frac{49}{4} = \frac{35}{12}. \qquad \triangle$$

The variance of a linear transformation of $\mathscr{X}$ is also easy to compute. Let $\mu = E[\mathscr{X}]$.[57] Then,

$$\begin{aligned} Var(a\mathscr{X} + b) &= E[(a\mathscr{X} + b - E[a\mathscr{X} + b])^2] = E[(a\mathscr{X} + b - a\mu - b)^2] \\ &= E[(a\mathscr{X} - a\mu)^2] = E[a^2(\mathscr{X} - \mu)^2] = a^2 E[(\mathscr{X} - \mu)^2] \\ &= a^2 Var(\mathscr{X}). \end{aligned}$$

In particular, this means that $Var(b) = 0$ and $Var(\mathscr{X} + b) = Var(\mathscr{X})$ for any constant b.[58] That is, constants have variance 0, and shifting the values of $\mathscr{X}$ by a constant does not affect its variance. However, scaling $\mathscr{X}$ by a constant scales quadratically the variance; i.e. $Var(a\mathscr{X}) = a^2 Var(\mathscr{X})$.

The value $\sqrt{Var(\mathscr{X})}$, called the *standard deviation* of $\mathscr{X}$, has the same units as the mean.

Covariance

WE HAVE SEEN THAT the mean of a sum of random variables is equal to the sum of their means. This result does not carry out to variances in general. In fact, we know already that

$$Var(\mathscr{X} + \mathscr{X}) = Var(2\mathscr{X}) = 4Var(\mathscr{X}) \neq Var(\mathscr{X}) + Var(\mathscr{X}).$$

[55] The *spread* or *variation* of the values.

[56] It requires a case analysis for when $\mathscr{X}$ is larger or smaller than μ.

°Should we explain the variance of an indicator variable? Exercise?

[57] Recall that $E[a\mathscr{X} + b] = aE[\mathscr{X}] + b$.

[58] By setting $a = 0$ and $a = 1$, respectively.

However, if two random variables are independent, then the variance of their sum corresponds to the sum of their variances. We will show this using the *covariance*.

Definition 45. Let $\mathscr{X}$ and $\mathscr{Y}$ be two random variables, and $\mu_{\mathscr{X}} = E[\mathscr{X}]$, $\mu_{\mathscr{Y}} = E[\mathscr{Y}]$. The *covariance* of $\mathscr{X}$ and $\mathscr{Y}$ is

$$Cov(\mathscr{X},\mathscr{Y}) = E[(\mathscr{X}-\mu_{\mathscr{X}})(\mathscr{Y}-\mu_{\mathscr{Y}})]$$

Expanding this definition, we obtain

$$\begin{aligned}Cov(\mathscr{X},\mathscr{Y}) &= E[\mathscr{X}\mathscr{Y}-\mu_{\mathscr{X}}\mathscr{Y}-\mu_{\mathscr{Y}}\mathscr{X}+\mu_{\mathscr{X}}\mu_{\mathscr{Y}}]\\ &= E[\mathscr{X}\mathscr{Y}]-\mu_{\mathscr{X}}E[\mathscr{Y}]-\mu_{\mathscr{Y}}E[\mathscr{X}]+\mu_{\mathscr{X}}\mu_{\mathscr{Y}}\\ &= E[\mathscr{X}\mathscr{Y}]-\mu_{\mathscr{X}}\mu_{\mathscr{Y}}-\mu_{\mathscr{Y}}\mu_{\mathscr{X}}+\mu_{\mathscr{X}}\mu_{\mathscr{Y}} = E[\mathscr{X}\mathscr{Y}]-E[\mathscr{X}]E[\mathscr{Y}].\end{aligned}$$

Notice that the covariance is symmetric, and $Cov(\mathscr{X},\mathscr{X}) = Var(\mathscr{X})$. In addition, for any constant a, $Cov(a\mathscr{X},\mathscr{Y}) = aCov(\mathscr{X},\mathscr{Y})$.° The covariance is also additive.

°Exercise!

Proposition 46. $Cov(\mathscr{X}_1+\mathscr{X}_2,\mathscr{Y}) = Cov(\mathscr{X}_1,\mathscr{Y})+Cov(\mathscr{X}_2,\mathscr{Y})$.

Proof.

$$\begin{aligned}Cov(\mathscr{X}_1+\mathscr{X}_2,\mathscr{Y}) &= E[(\mathscr{X}_1+\mathscr{X}_2)\mathscr{Y}]-E[\mathscr{X}_1+\mathscr{X}_2]E[\mathscr{Y}]\\ &= E[\mathscr{X}_1\mathscr{Y}]+E[\mathscr{X}_2\mathscr{Y}]-(E[\mathscr{X}_1]+E[\mathscr{X}_2])E[\mathscr{Y}]\\ &= E[\mathscr{X}_1\mathscr{Y}]-E[\mathscr{X}_1]E[\mathscr{Y}]+E[\mathscr{X}_2\mathscr{Y}]-E[\mathscr{X}_2]E[\mathscr{Y}]\\ &= Cov(\mathscr{X}_1,\mathscr{Y})+Cov(\mathscr{X}_2,\mathscr{Y}). \qquad \square\end{aligned}$$

This can be easily generalized to arbitrary sums and, using the symmetry of the covariance, we obtain the following theorem.

Theorem 47. $Cov(\sum_{i=1}^{n}\mathscr{X}_i,\sum_{j=1}^{m}\mathscr{Y}_j) = \sum_{i=1}^{n}\sum_{j=1}^{m}Cov(\mathscr{X}_i,\mathscr{Y}_j)$.

We can now compute the variance of the sum of random variables as:[59]

$$\begin{aligned}Var\left(\sum_{i=1}^{n}\mathscr{X}_i\right) &= Cov\left(\sum_{i=1}^{n}\mathscr{X}_i,\sum_{i=1}^{n}\mathscr{X}_i\right)\\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}Cov(\mathscr{X}_i,\mathscr{X}_j)\\ &= \sum_{i=1}^{n}\left[\sum_{j\neq i}Cov(\mathscr{X}_i,\mathscr{X}_j)+Cov(\mathscr{X}_i,\mathscr{X}_i)\right]\\ &= \sum_{i=1}^{n}\sum_{j\neq i}Cov(\mathscr{X}_i,\mathscr{X}_j)+\sum_{i=1}^{n}Cov(\mathscr{X}_i,\mathscr{X}_i)\\ &= \sum_{i=1}^{n}\sum_{j\neq i}Cov(\mathscr{X}_i,\mathscr{X}_j)+\sum_{i=1}^{n}Var(\mathscr{X}_i).\end{aligned}$$

If $n = 2$, then this means that

$$\begin{aligned}Var(\mathscr{X}+\mathscr{Y}) &= Var(\mathscr{X})+Var(\mathscr{Y})+Cov(\mathscr{X},\mathscr{Y})+Cov(\mathscr{Y},\mathscr{X})\\ &= Var(\mathscr{X})+Var(\mathscr{Y})+2Cov(\mathscr{X},\mathscr{Y}).\end{aligned}$$

Then we get to the result that we have hinted before.

[59] Recall that $Cov(\mathscr{X},\mathscr{X}) = Var(\mathscr{X})$.

Theorem 48. *If $\mathscr{X}$ and $\mathscr{Y}$ are independent, then $Cov(\mathscr{X},\mathscr{Y}) = 0$.*[60]

Proof. It suffices to show that $E[\mathscr{X}\mathscr{Y}] = E[\mathscr{X}]E[\mathscr{Y}]$. We show it for the discrete case only.[61]

$$\begin{aligned} E[\mathscr{X}\mathscr{Y}] &= \sum_j \sum_i x_i y_j P[\mathscr{X} = x_i, \mathscr{Y} = y_j] \\ &= \sum_j \sum_i x_i y_j P[\mathscr{X} = x_i] P[\mathscr{Y} = y_j] && (\dagger) \\ &= \sum_j y_j P[\mathscr{Y} = y_j] \sum_i x_i P[\mathscr{X} = x_i] = E[\mathscr{X}]E[\mathscr{Y}], \end{aligned}$$

where (†) follows by independence. □

In particular, if $\mathscr{X}_1, \ldots, \mathscr{X}_n$ are independent, then

$$Var(\sum_{i=1}^{n} \mathscr{X}_i) = \sum_{i=1}^{n} Var(\mathscr{X}_i).$$

Example 49. Compute the variance of the sum of 10 independent rolls of a fair die.

Solution. If $\mathscr{X}_i$ denotes the i-th roll, we get

$$Var(\sum_{i=1}^{10} \mathscr{X}_i) = \sum_{i=1}^{10} Var(\mathscr{X}_i) = 10(\frac{35}{12}) = \frac{175}{6}. \qquad \triangle$$

Intuitively, the covariance describes the relationship between two variables. Consider for example the indicator variables $\mathscr{X}$, $\mathscr{Y}$ for the events $\mathscr{E}$ and $\mathscr{F}$, respectively. It follows that

$$Cov(\mathscr{X},\mathscr{Y}) = E[\mathscr{X}\mathscr{Y}] - E[\mathscr{X}]E[\mathscr{Y}] = P[\mathscr{X} = 1, \mathscr{Y} = 1] - P[\mathscr{X} = 1]P[\mathscr{Y} = 1].$$

Then, $Cov(\mathscr{X},\mathscr{Y}) > 0$ if and only if $P[\mathscr{X} = 1, \mathscr{Y} = 1] > P[\mathscr{X} = 1]P[\mathscr{Y} = 1]$ or, equivalently, $P(\mathscr{Y} = 1 \mid \mathscr{X} = 1) > P[\mathscr{Y} = 1]$. In words, the covariance tells us whether it is more or less likely to observe $[\mathscr{Y} = 1]$ when we know that $\mathscr{X} = 1$.[62]

In general, a positive covariance between two RVs expresses that both variables grow together,[63] while a negative covariance says that one decreases while the other increases. The strength of the relationship between the two variables is indicated by their *correlation*, defined as

$$Corr(\mathscr{X},\mathscr{Y}) = \frac{Cov(\mathscr{X},\mathscr{Y})}{\sqrt{Var(\mathscr{X})Var(\mathscr{Y})}}.$$

This dimensionless value is always between -1 and 1.°

°Exercise!

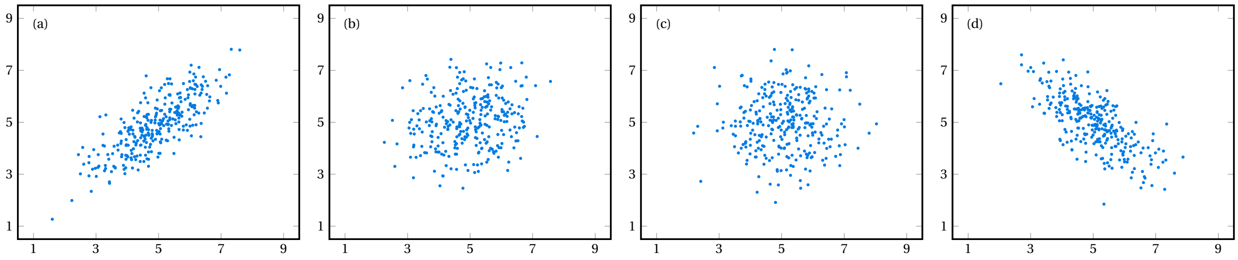


Figure 9: Random variables $\mathscr{X}$ and $\mathscr{Y}$ with correlations (a) 0.75; (b) 0.2; (c) 0; and (d) -0.75.

[60] Notice that this is not an equivalence: two variables may have covariance 0, and still not be independent.

[61] All other cases are analogous.

[62] And by symmetry, also the dual.

[63] That is, $\mathscr{Y}$ grows as $\mathscr{X}$ grows.

Moment Generating Functions

EVERY RANDOM VARIABLE $\mathscr{X}$ has a *moment generating function* (mgf) defined for every $t \in \mathbb{N}$ as

$$\phi(t) := E[e^{t\mathscr{X}}] = \begin{cases} \sum_x e^{tx} p(x) & \text{if } \mathscr{X} \text{ is discrete} \\ \int_x e^{tx} f(x) dx & \text{if } \mathscr{X} \text{ is continuous} \end{cases}$$

This function is called moment generating because all the moments of $\mathscr{X}$ can be obtained by differentiating it and evaluating at zero.[64] For example,

$$\phi'(t) = \frac{d}{dt} E\left[e^{t\mathscr{X}}\right] = E\left[\frac{d}{dt} e^{t\mathscr{X}}\right] = E\left[\mathscr{X} e^{t\mathscr{X}}\right].$$

Hence, $\phi'(0) = E[\mathscr{X}]$. Similarly, $\phi''(t) = E\left[\mathscr{X}^2 e^{t\mathscr{X}}\right]$ and so $\phi''(0) = E[\mathscr{X}^2]$. In general, the n-th moment of $\mathscr{X}$ is the n-th derivative of ϕ evaluated at $t = 0$.

An important property is that the mgf of the sum of *independent* random variables is the product of their individual moment generating functions. For two random variables, the mgf of $\mathscr{X} + \mathscr{Y}$ is

$$\begin{aligned} \phi_{\mathscr{X}+\mathscr{Y}}(t) &= E\left[e^{t(\mathscr{X}+\mathscr{Y})}\right] = E\left[e^{t\mathscr{X}} e^{t\mathscr{Y}}\right] \\ &= E\left[e^{t\mathscr{X}}\right] E\left[e^{t\mathscr{Y}}\right] = \phi_{\mathscr{X}}(t)\phi_{\mathscr{Y}}(t), \qquad (\dagger) \end{aligned}$$

where (†) follows because $\mathscr{X}$ and $\mathscr{Y}$, and hence also $e^{t\mathscr{X}}$ and $e^{t\mathscr{Y}}$ are independent.[65]

Interestingly, mgfs *uniquely* determine the distribution of random variables; that is, there is a one-to-one correspondence between distributions and their moment generating function.

The Weak Law of Large Numbers

WE NOW PROVE two important results.

Theorem 50 (Markov's Inequality). *Let $\mathscr{X}$ be a random variable that takes only non-negative values. Then, for every $a > 0$*

$$P[\mathscr{X} \ge a] \le \frac{E[\mathscr{X}]}{a}.$$

Proof. We prove it for a continuous variable with density f.

$$\begin{aligned} E[\mathscr{X}] &= \int_0^\infty x f(x) dx = \int_0^a x f(x) dx + \int_a^\infty x f(x) dx \\ &\ge \int_a^\infty x f(x) dx \ge \int_a^\infty a f(x) dx = a \int_a^\infty f(x) dx = aP[\mathscr{X} \ge a]. \quad \square \end{aligned}$$

Corollary 51 (Chebyshev's Inequality). *If $\mathscr{X}$ is a random variable with mean μ and variance σ^2, then for every $k > 0$ we have*

$$P[|\mathscr{X} - \mu| \ge k] \le \frac{\sigma^2}{k^2}.$$

[64] The n-th moment of $\mathscr{X}$ is $E[\mathscr{X}^n]$.

[65] In the proof of Theorem 48, we have shown that if two random variables are independent, then $E[\mathscr{X}\mathscr{Y}] = E[\mathscr{X}]E[\mathscr{Y}]$.

Proof. Notice that $(\mathcal{X}-\mu)^2$ is a non-negative random variable. So we can apply Markov's inequality, with $a=k^2$ to get

$$P[(\mathcal{X}-\mu)^2 \geq k^2] \leq \frac{E[(\mathcal{X}-\mu)^2]}{k^2}.$$

Since $(\mathcal{X}-mu)^2 \geq k^2$ holds if and only if $|\mathcal{X}-\mu| \geq k$, this implies

$$P[|\mathcal{X}-\mu| \geq k] \leq \frac{E[(\mathcal{X}-\mu)^2]}{k^2} \leq \frac{\sigma^2}{k^2}. \qquad \square$$

The Markov and Chevishev inequalities allow bounding probabilities even if the general distribution of $\mathcal{X}$ is unknown, as long as the mean, and potentially the variance, is known.[66]

Example 52. Suppose that the number of hours that a person works per week is a random variable with mean 40.

1. What can be said about the probability that this week they will work more than 60 hours?
2. If the variance of the working hours per week is 16, what is the probability that this week's work will be between 32 and 48 hours?

Solution. Let $\mathcal{X}$ be the number of working hours in a week.

1. Using Markov's inequality, $P[\mathcal{X} \geq 60] \leq \frac{E[\mathcal{X}]}{60} = \frac{40}{60} = 2/3$.
2. Using Chebyshev's inequality, $P[|\mathcal{X}-40| \geq 8] \leq \frac{16}{64} = 1/4$. So, we know that $P[|\mathcal{X}-40| \leq 8] = 1 - P[|\mathcal{X}-40| \geq 8] \geq 0.75$. △

If we use Chebyshev's inequality with distance $k\sigma$, we get that

$$P[|\mathcal{X}-\mu| \geq k\sigma] \leq \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}.$$

That is, the probability that $\mathcal{X}$ differs from its mean by at least k standard deviations is bounded by $\frac{1}{k^2}$.[67]

A consequence of this inequality is that if one takes several independent and identically distributed random variables, then their average will tend to their mean (with probability 1).[68]

Theorem 53 (Weak Law of Large Numbers)**.** *Let $\mathcal{X}_1, \mathcal{X}_2, \ldots$ be a sequence of independent, identically distributed RVs with mean $E[\mathcal{X}_i]=\mu$. Then, for any $\varepsilon > 0$*

$$\lim_{n\to\infty} P\left[\left|\frac{\sum_{i=1}^n \mathcal{X}_i}{n} - \mu\right| > \varepsilon\right] = 0.$$

Proof. Suppose for simplicity that the random variables have a finite variance σ^2. Since all the variables are independent, if $\mathcal{Y} = \frac{\sum_{i=1}^n \mathcal{X}_i}{n}$, then

$$E[\mathcal{Y}] = \mu \qquad\qquad Var(\mathcal{Y}) = \frac{\sigma^2}{n}.$$

Applying Chebyshev's inequality results in

$$P\left[\left|\frac{\sum_{i=1}^n \mathcal{X}_i}{n} - \mu\right| > \varepsilon\right] \leq \frac{\sigma^2}{n\varepsilon^2}. \qquad \square$$

[66] If the distribution is known, the probabilities can usually be computed precisely, without the need to approximate.

[67] Decreases quadratically.

[68] This is the basis of many statistical analyses, for repeated experiments

Suppose for example that we independently repeat a clinical trial. Let $\mathscr{E}$ be a fixed event that has probability $P(\mathscr{E})$ of occurring at each trial. If $\mathscr{X}_i$ is the indicator variable for $\mathscr{E}$ occurring at trial i, then $\sum_{i=1}^{n}$ is the number of times that $\mathscr{E}$ is observed in the first n trials. Since $E[\mathscr{X}_i] = P(\mathscr{E})$, it follows from the weak law of large numbers that the probability that the proportion of trials in which we observe $\mathscr{E}$ differs from $P(\mathscr{E})$ by more than ε tends to 0 as n grows.

Special Random Variables

WE INTRODUCE SOME OF the most commonly observed and used RVs.

Bernoulli and Binomial

CONSIDER AN EXPERIMENT WHOSE outcome may be a *success* or a *failure*, and $\mathscr{X}$ its indicator variable.[69] The probability mass function of $\mathscr{X}$ is defined by $P[\mathscr{X}=1]=p$; its *probability of success*. Such a random variable is called *Bernoulli*[70] and, as we know already, its expected value is equal to its probability of success p.

If we independently repeat the experiment n times, and $\mathscr{X}$ represents the number of successes observed, then $\mathscr{X}$ is a *binomial* random variable with parameters (n,p). The probability mass function of this random variable is defined for all $i, 0 \le i \le n$ as[71]

$$P[\mathscr{X}=i]=\binom{n}{i}p^i(1-p)^{n-i}.$$

The probability mass functions of three binomial random variables with parameters $(10,0.5)$, $(10,0.4)$, and $(10,0.75)$, respectively, are depicted in Figure 10. Notice that two of them lean (or *skew*) away from the center.

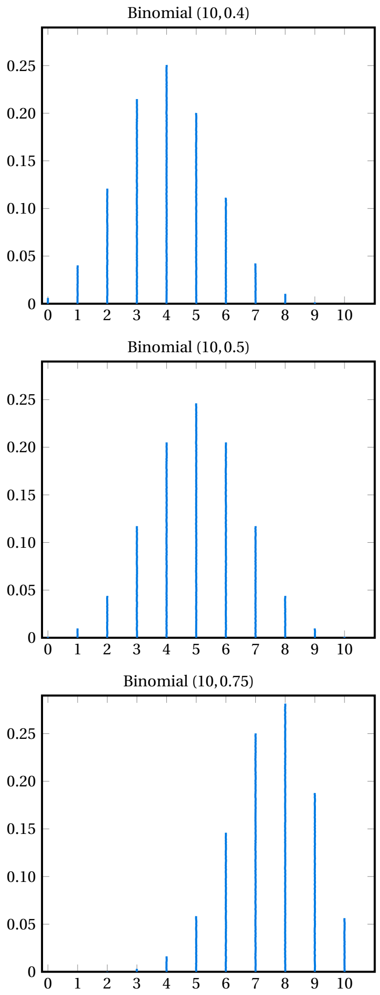


Figure 10: Three binomial probability mass functions.

Example 54. A system with n components works if at least half of its components function. If each component works, independently, with probability p,

1. for which values of p will a 5-component system be more reliable than a 3-component one?

2. In general, when is a $(2k+1)$-component system more reliable than a $(2k-1)$-component one?

Solution. For the first question, we know that a 5-component system and a 3-component system will work, respectively, with probability

$$\binom{5}{3}p^3(1-p)^2+\binom{5}{4}p^4(1-p)+p^5,$$
$$\binom{3}{2}p^2(1-p)+p^3.$$

The 5-component system is more reliable if

$$\binom{5}{3}p^3(1-p)^2+\binom{5}{4}p^4(1-p)+p^5 \ge \binom{3}{2}p^2(1-p)+p^3$$

[69] That is, $\mathscr{X}=1$ if the experiment succeeds, and $\mathscr{X}=0$ otherwise.

[70] After James Bernoulli.

[71] Recall that $\binom{n}{r}=\frac{n!}{r!(n-r)!}$.

or, equivalently, if $3(p-1)^2(2p-1) \geq 0$; that is if $p \geq 1/2$.

In general, consider a system with $2k+1$ components and let $\mathscr{X}$ denote the number of the first $2k-1$ that function. The system will work if any of the following three conditions hold:

1. $\mathscr{X} \geq k+1$;

2. $\mathscr{X} = k$ and at least one of the last two components functions; or

3. $\mathscr{X} = k-1$ and the last two components function.

The probability of this system working is

$$P[\mathscr{X} \geq k+1] + P[\mathscr{X} = k](1-(1-p)^2) + P[\mathscr{X} = k-1]p^2.$$

Similarly, the probability that the smaller system works is

$$P[\mathscr{X} = k] + P[\mathscr{X} \geq k+1].$$

This means that the larger system is more reliable if[72]

$$\begin{aligned}
0 &\leq P_{2k+1}(\text{works}) - P_{2k-1}(\text{works}) \\
&= P[\mathscr{X} = k-1]p^2 - P[\mathscr{X} = k](1-p)^2 \\
&= \binom{2k-1}{k-1} p^{k-1}(1-p)^k p^2 - \binom{2k-1}{k} p^k (1-p)^{k-1}(1-p)^2 \\
&= \binom{2k-1}{k} p^k (1-p)^k (p-(1-p)) = \binom{2k-1}{k} p^k (1-p)^k (2p-1), \qquad (\dagger)
\end{aligned}$$

where (†) follows from $\binom{2k-1}{k} = \binom{2k-1}{k-1}$. Thus, the larger system is more reliable if and only if $p \geq 1/2$.[73] △

Recall that a binomial random variable $\mathscr{X}$ represents the number of successes of n independent trials, where each trial has probability p of occurring. Then $\mathscr{X} = \sum_{i=1}^n \mathscr{X}_i$, where the $\mathscr{X}_i$s are independent Bernoulli random variables. In particular,[74]

$$\begin{aligned}
E[\mathscr{X}_i] &= P[\mathscr{X}_i = 1] = p \\
Var(\mathscr{X}_i) &= E[\mathscr{X}^2] - (E[\mathscr{X}])^2 = p - p^2 = p(1-p).
\end{aligned}$$

From this, it immediately follows that

$$\begin{aligned}
E[\mathscr{X}] &= \sum_{i=1}^n E[\mathscr{X}_i] = np \\
Var(\mathscr{X}) &= \sum_{i=1}^n Var(\mathscr{X}_i) = np(1-p).
\end{aligned}$$

In addition, if $\mathscr{X}_1$ and $\mathscr{X}_2$ are two independent binomial random variables with parameters (n_i, p) for $i = 1,2$, then their sum is also a binomial with parameters (n_1+n_2, p).[75]

TO COMPUTE THE DISTRIBUTION function of a binomial $\mathscr{X}$ with parameters (n, p) it is helpful to use the equation°

$$P[\mathscr{X} = k+1] = \frac{p}{1-p}\frac{n-k}{k+1}P[\mathscr{X} = k].$$

[72] The probability of the larger system to work is larger than the one of the smaller system.

[73] For any number N, $N(2p-1) \geq 0$ iff $2Np \geq N$.

[74] Notice that $\mathscr{X}^2 = \mathscr{X}$ because $\mathscr{X}$ can only take values 0 or 1.

[75] It is important that the probability in both binomials is the same. $\mathscr{X}_i$ denotes the number of successes in n_i independent trials. So their sum corresponds to the successes in the sum of their (independently made) trials.

°Exercise!

Example 55. Let $\mathscr{X}$ be a binomial with $n = 5$ and $p = 0.3$. Then, starting with $P[\mathscr{X} = 0] = (0.7)^5 = 0.168$ we obtain

$$\begin{aligned}
P[\mathscr{X} = 1] &= \frac{3}{7}\frac{5}{1}P[\mathscr{X} = 0] \\
P[\mathscr{X} = 2] &= \frac{3}{7}\frac{4}{2}P[\mathscr{X} = 1] \\
P[\mathscr{X} = 3] &= \frac{3}{7}\frac{3}{3}P[\mathscr{X} = 2] \\
P[\mathscr{X} = 4] &= \frac{3}{7}\frac{2}{4}P[\mathscr{X} = 3] \\
P[\mathscr{X} = 5] &= \frac{3}{7}\frac{1}{5}P[\mathscr{X} = 4].
\end{aligned}$$

△

It is obviously possible to generalize the Bernoulli and binomial random variables to allow more than two values to be taken. In this case, we would speak of a *multinomial* RV. Understanding the properties of such RVs is left as an exercise to the interested student.

Poisson

A POISSON RANDOM VARIABLE with parameter $\lambda > 0$ takes values in the natural numbers, and has the probability mass function, for every $i \in \mathbb{N}$,[76]

$$P[\mathscr{X} = i] = e^{-\lambda}\frac{\lambda^i}{i!};$$

see Figure 11.[77]

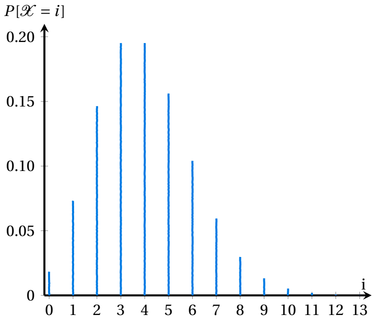


Figure 11: Poisson mass function for $\lambda = 4$.

To determine the mean and variance of a Poisson random variable, we compute its moment generating function, and its two first derivatives.

$$\begin{aligned}
\phi(t) = E[e^{t\mathscr{X}}] &= \sum_{i=0}^{\infty} e^{ti}e^{-\lambda}(\lambda^i/i!) \\
&= e^{-\lambda}\sum_{i=0}^{\infty}(\lambda e^t)^i/i! \\
&= e^{-\lambda}e^{\lambda e^t} = exp[\lambda(e^t - 1)] \\
\phi'(t) &= \lambda e^t exp[\lambda(e^t - 1)] \\
\phi''(t) &= (\lambda e^t)^2 exp[\lambda(e^t - 1)] + \lambda e^t exp[\lambda(e^t - 1)]
\end{aligned}$$

From these equations, we deduce

$$\begin{aligned}
E[\mathscr{X}] &= \phi'(0) = \lambda \\
Var(\mathscr{X}) &= \phi''(0) - (E[\mathscr{X}])^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.
\end{aligned}$$

That is, the mean and the variance of a Poisson random variable are both equal to the parameter λ.

Poisson random variables provide good approximations for binomial variables with parameters (n, p) if n is large and p is small. Suppose that $\mathscr{X}$ is such a binomial random variable and let $\lambda = np$. Then

$$\begin{aligned}
P[\mathscr{X} = i] &= \frac{n!}{(n-i)!i!}p^i(1-p)^{n-i} = \frac{n!}{(n-i)!i!}\left(\frac{\lambda}{n}\right)^i\left(1 - \frac{\lambda}{n}\right)^{n-i} \\
&= \frac{n(n-1)\cdots(n-i+1)}{n^i}\frac{\lambda^i}{i!}\frac{(1-\lambda/n)^n}{(1-\lambda/n)^i}.
\end{aligned}$$

Assuming that n is large and p is small, we can deduce that[78]

$$\left(1 - \frac{\lambda}{n}\right)^n \approx e^{-\lambda}; \qquad \left(1 - \frac{\lambda}{n}\right)^i \approx 1; \qquad \frac{n(n-1)\cdots(n-i+1)}{n^i} \approx 1.$$

And hence it follows that $P[\mathscr{X} = i] \approx e^{-\lambda}\frac{\lambda^i}{i!}$.[79]

[76] First defined by S. D. Poisson.

[77] To see that this is a mass function, recall that for all $x \in \mathbb{R}$, $\sum_{i=1}^{\infty}\frac{x^i}{i!} = e^x$.

[78] Recall that $\lim_{n\to\infty}(1 + x/n)^n = e^x$.

[79] If a very large number n of independent trials with a very low probability p of success are performed, then the number of successes is approximately a Poisson with $\lambda = np$.

Example 56. Suppose that, in average, there are three accidents in the highway between Trento and Bolzano every week. Compute the probability that there is at least one accident this week.

Solution. Let $\mathscr{X}$ denote the number of accidents this week. It is reasonable to assume that there is a large number of cars passing, each with a very low probability of having an accident. So $\mathscr{X}$ should be approximately Poisson distributed. Thus,

$$P[\mathscr{X} \geq 1] = 1 - P[\mathscr{X} = 0] \approx 1 - e^{-3}\frac{3^0}{0!} = 1 - e^{-3} \approx 0.95. \qquad \triangle$$

THE POISSON APPROXIMATION REMAINS valid in more general circumstances. If n independent trials are performed, each with a probability of success $p_i, 1 \leq i \leq n$, then if n is large and *each* p_i is small, the number of successful trials is approximately a Poisson with mean $\sum_{i=1}^n p_i$. This remains true without the independence assumption, as long as the dependence is not very strong.

Example 57. Consider n people that leave their umbrella at the entrance of a bar, and pick up one of the umbrellas randomly when leaving. If $\mathscr{X}$ denotes the number of people that take their own umbrella, then for large n $\mathscr{X}$ approximates a Poisson distribution with mean 1. Intuitively, this is the case because we can express $\mathscr{X} = \sum_{i=1}^n \mathscr{X}_i$, where each $\mathscr{X}_i$ is the indicator variable for person i picking their own umbrella. Since each person is equally likely to pick any umbrella, we have that $P[\mathscr{X} = i] = 1/n$.[80] Moreover $E[\mathscr{X}] = \sum_{i=1}^n E[\mathscr{X}_i] = n(1/n) = 1$.

However, if the j-th person picks up their own umbrella, then the i-th person ($i \neq j$) is equally likely to pick any of the remaining $n-1$ umbrellas; more precisely, $P[\mathscr{X}_i = 1 \mid \mathscr{X}_j = 1] = \frac{1}{n-1}$. This means that $\mathscr{X}_i$ and $\mathscr{X}_j$ are not independent, but their dependence is extremely weak.[81] $\triangle$

Poisson random variables are also *reproductive*; that is, the sum of two independent Poisson random variables $\mathscr{X}$ and $\mathscr{Y}$ is also Poisson. To prove this, consider the moment generating function of $\mathscr{X} + \mathscr{Y}$[82]

$$\begin{aligned} E[e^{t(\mathscr{X}+\mathscr{Y})}] &= E[e^{t\mathscr{X}} e^{t\mathscr{Y}}] = E[e^{t\mathscr{X}}]E[e^{t\mathscr{Y}}] \\ &= exp(\lambda_{\mathscr{X}}(e^t - 1))exp(\lambda_{\mathscr{Y}}(e^t - 1)) = exp((\lambda_{\mathscr{X}} + \lambda_{\mathscr{Y}})(e^t - 1)). \end{aligned}$$

Since this is the mgf of a Poisson with mean $\lambda_{\mathscr{X}} + \lambda_{\mathscr{Y}}$, it follows that $\mathscr{X} + \mathscr{Y}$ is indeed a Poisson.[83]

Example 58. The number of hourly customers at a bar is a Poisson with mean 4. What is the probability that over a 2-hour period there are no more than 3 customers?

Solution. Let $\mathscr{X}_i, i = 1,2$ be the number of customers during the i-th hour. Assuming that $\mathscr{X}_i$ and $\mathscr{X}_2$ are independent, then $\mathscr{X} + \mathscr{X}_2$ is Poisson with mean 8. Thus,

$$P[\mathscr{X}_1 + \mathscr{X}_2 \leq 3] = \sum_{n=0}^{3} e^{-8}\frac{8^i}{i!} = 0.423. \qquad \triangle$$

[80] That is, we can approximately simulate the experiment by a binomial with probability $1/n$.

[81] Specially when n tends to be large.

[82] Assume that the means of $\mathscr{X}$ and $\mathscr{Y}$ are $\lambda_{\mathscr{X}}$ and $\lambda_{\mathscr{Y}}$, respectively.

[83] Recall that there is a one-to-one correspondence between moment generating functions and distributions.

CONSIDER NOW A SCENARIO in which a random number N of events will occur, and each of them is independently of type 1 or type 2 with probabilities p and $1-p$, respectively.[84] Let $N_i, i = 1,2$ denote the number of events of type i observed. If N is a Poisson with mean λ then the joint mass function of N_1, N_2 is

$$\begin{aligned} P[N_1 = n, N_2 = m] &= P[N_1 = n, N_2 = m, N = n+m] \\ &= P[N_1 = n, N_2 = m \mid N = n+m] P[N = n+m] \\ &= P[N_1 = n, N_2 = m \mid N = n+m] e^{-\lambda} \frac{\lambda^{n+m}}{(n+m)!}. \end{aligned}$$

Since each of the $n+m$ events are independently of type 1 with probability p, the probability that there are *exactly* n events of type 1 is the probability that the binomial random variable $(n+m, p)$ takes value n. That is,

$$\begin{aligned} P[N_1 = n, N_2 = m] &= \frac{(n+m)!}{n!m!} p^n (1-p)^m e^{-\lambda} \frac{\lambda^{n+m}}{(n+m)!} \\ &= e^{-\lambda p} \frac{(\lambda p)^n}{n!} e^{-\lambda(1-p)} \frac{(\lambda(1-p))^m}{m!}. \end{aligned}$$

The probability mass function of N_1 is

$$\begin{aligned} P[N_1 = n] &= \sum_{m=0}^{\infty} P[N_1 = n, N_2 = m] \\ &= e^{-\lambda p} \frac{(\lambda p)^n}{n!} \sum_{m=0}^{\infty} e^{-\lambda(1-p)} \frac{(\lambda(1-p))^m}{m!} = e^{-\lambda p} \frac{(\lambda p)^n}{n!}. \end{aligned}$$

Thus, N_1 is a Poisson with mean λp, and similarly N_2 is Poisson with mean $\lambda(1-p)$. Moreover, these two variables are independent.

This result can be generalised to the case where each event can take any of r different categories with probabilities $p_1, \ldots, p_r$. That is, the numbers of type i events $(1 \le i \le r)$ are independent Poisson random variables with mean λp_i.

To compute the distribution of a Poisson $\mathscr{X}$ with mean λ we notice that

$$\frac{P[\mathscr{X} = i+1]}{P[\mathscr{X} = i]} = \frac{e^{-\lambda}\lambda^{i+1}/(i+1)!}{e^{-\lambda}\lambda^i/i!} = \frac{\lambda}{i+1}.$$

Starting from $P[\mathscr{X} = 0] = e^{-\lambda}$, we can use this equation to successively compute the probability of each successive value.[85]

Hypergeometric Random Variables

°CONSIDER A BIN WITH N working batteries, and M defective ones. We randomly pick up a sample of size n.[86] If $\mathscr{X}$ is the number of working batteries in the sample, then for all $i, 0 \le i \le \min(N, m)$[87]

$$P[\mathscr{X} = i] = \frac{\binom{N}{i}\binom{M}{n-i}}{\binom{N+M}{n}}.$$

A random variable with such a mass function is called *hypergeometric* with parameters N, M, n.

°Optional

[84] For example, the number of customers N in a bar is a random variable; each customer is male with probability p and female with probability $1-p$.

[85] $P[\mathscr{X} = i+1] = (\lambda/(i+1))P[\mathscr{X} = i]$.

[86] Any of the $\binom{N+M}{n}$ such samples is equally likely.

[87] To simplify, we assume $\binom{m}{r} = 0$ whenever $r > m$ or $r < 0$.

An intuitive way to see a hypergeometric random variable $\mathscr{X}$ is to think of the sample to be drawn sequentially and defining the random variables

$$\mathscr{X}_i = \begin{cases} 1 & i\text{-th selection is working} \\ 0 & \text{otherwise.} \end{cases}$$

Each selection is equally likely to be any of the $N+M$ batteries; hence $P[\mathscr{X}_i = 1] = \frac{N}{N+M}$. In addition, for any $i \neq j$,[88]

$$P[\mathscr{X}_i = 1, \mathscr{X}_j = 1] = P[\mathscr{X}_i = 1 \mid \mathscr{X}_j = 1] P[\mathscr{X}_j = 1] = \frac{N-1}{N+M-1} \frac{N}{N+M}.$$

Since $\mathscr{X} = \sum_{i=1}^{n} \mathscr{X}_i$, it follows that

$$\begin{aligned} E[\mathscr{X}] &= \sum_{i=1}^{n} E[\mathscr{X}_i] = \sum_{i=1}^{n} P[\mathscr{X}_i = 1] = \frac{nN}{N+M}, \\ Var(\mathscr{X}) &= \sum_{i=1}^{n} Var(\mathscr{X}_i) + 2 \sum_{1 \le i < j \le n} Cov(\mathscr{X}_i \mathscr{X}_j). \end{aligned}$$

Since each $\mathscr{X}_i$ is a Bernoulli, $Var(\mathscr{X}_i) = P[\mathscr{X}_i = 1](1 - P[\mathscr{X}_i = 1]) = \frac{NM}{(N+M)^2}$, and for $i < j$[89] $E[\mathscr{X}_i \mathscr{X}_j] = P[\mathscr{X}_i = 1, \mathscr{X}_j = 1] = \frac{N(N-1)}{(N+M)(N+M-1)}$. Thus,

$$\begin{aligned} Cov(\mathscr{X}_i, \mathscr{X}_j) &= E[\mathscr{X}_i \mathscr{X}_j] - E[\mathscr{X}_i] E[\mathscr{X}_j] \\ &= \frac{N(N-1)}{(N+M)(N+M-1)} - \left(\frac{N}{N+M} \right)^2 \\ &= \frac{-NM}{(N+M)^2 (N+M-1)}. \end{aligned}$$

Since there are $\binom{n}{2}$ terms in the sum of covariances, it follows that

$$\begin{aligned} Var(\mathscr{X}) &= \frac{nNM}{(N+M)^2} - \frac{n(n-1)NM}{(N+M)^2 (N+M-1)} \\ &= \frac{nNM}{(N+M)^2} \left[1 - \frac{n-1}{N+M-1} \right]. \end{aligned}$$

Let now $p = N/(N+M)$ be the proportion of batteries that are functioning. It then follows that $E[\mathscr{X}] = np$ and $Var(\mathscr{X}) = np(1-p)\left[1 - \frac{n-1}{N+M-1}\right]$.

If $N+M$ tends to infinity while preserving the same proportion p, then $Var(\mathscr{X})$ converges to $np(1-p)$, which is the variance of a binomial with parameters (n, p).

Let $\mathscr{X}$ and $\mathscr{Y}$ be two independent binomial random variables with parameters (n, p) and (m, p), respectively. The conditional mass function of $\mathscr{X}$ given that $\mathscr{X} + \mathscr{Y} = k$ is[90]

$$\begin{aligned} P[\mathscr{X} = i \mid \mathscr{X} + \mathscr{Y} = k] &= \frac{P[\mathscr{X} = i, \mathscr{Y} = k-i]}{P[\mathscr{X} + \mathscr{Y} = k]} = \frac{P[\mathscr{X} = i] P[\mathscr{Y} = k-i]}{P[\mathscr{X} + \mathscr{Y} = k]} \\ &= \frac{\binom{n}{i} p^i (1-p)^{n-i} \binom{m}{k-i} p^{k-i} (1-p)^{m-(k-i)}}{\binom{n+m}{k} p^k (1-p)^{n+m-k}} \\ &= \frac{\binom{n}{i}\binom{m}{k-1}}{\binom{n+m}{k}}; \end{aligned}$$

that is, this conditional distribution is a hypergeometric.

[88] Once that we have observed a working battery, another selection can be any of the remaining $N-1$ working batteries.

[89] $\mathscr{X}_i \mathscr{X}_j = 1$ iff $\mathscr{X}_i = 1 = \mathscr{X}_j$.

[90] Recall that the sum of two binomials (n, p) and (m, p) is a binomial $(n+m, p)$.

Uniform Random Variables

A RANDOM VARIABLE IS *uniformly distributed* over the interval $[\alpha, \beta]$ if its density function is (see Figure 13)

$$f(x) = \begin{cases} \frac{1}{\beta-\alpha} & \alpha \leq x \leq \beta \\ 0 & \text{otherwise.} \end{cases}$$

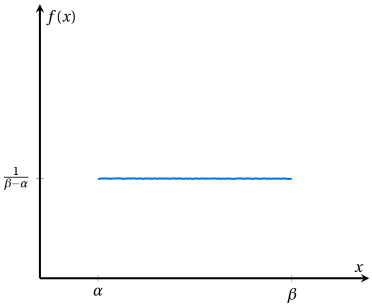


Figure 12: Uniform density function of a uniform distribution over $[\alpha, \beta]$.

The uniform distribution expresses that a random variable is equally likely to be near any value within $[\alpha, \beta]$. For a sub-interval of $I \subseteq [\alpha, \beta]$, the probability of $\mathscr{X}$ taking a value in I is the proportional size of I w.r.t. $\beta - \alpha$:

$$P[a < \mathscr{X} < b] = \frac{1}{\beta - \alpha} \int_a^b dx = \frac{b-a}{\beta - \alpha}.$$

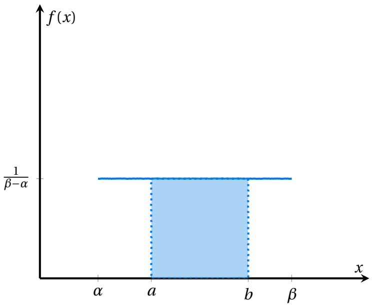


Figure 13: Probability of an interval in a uniform distribution.

For example, if $\mathscr{X}$ is uniformly distributed over the interval $[0, 10]$, then $P[\mathscr{X} > 6] = 0.4$ and $P[2 < \mathscr{X} < 5] = 0.3$.

Intuitively, the mean of a uniform random variable should appear at the middle of its interval. We confirm that this is the case:

$$E[\mathscr{X}] = \int_\alpha^\beta \frac{x}{\beta - \alpha} dx = \frac{\beta^2 - \alpha^2}{2(\beta - \alpha)} = \frac{(\beta + \alpha)(\beta - \alpha)}{2(\beta - \alpha)} = \frac{\alpha + \beta}{2}.$$

To compute the variance, we use[91]

$$E[\mathscr{X}^2] = \frac{1}{\beta - \alpha} \int_\alpha^\beta x^2 dx = \frac{\beta^3 - \alpha^3}{3(\beta - \alpha)} = \frac{\beta^2 + \alpha\beta + \alpha^2}{3},$$

which yields

$$Var(\mathscr{X}) = \frac{\beta^2 + \alpha\beta + \alpha^2}{3} - \left(\frac{\alpha + \beta}{2}\right)^2 = \frac{\alpha^2 + \beta^2 - 2\alpha\beta}{12} = \frac{(\beta - \alpha)^2}{12}.$$

The value of a uniform $(0, 1)$ random variable is called a *random number*. Modern computers use mathematical methods to generate sequences of independent (pseudo-)random numbers.[92] Random numbers are used often in clinical trials, for example in what are called *double-blind* tests.[93]

THE NOTION OF UNIFORM random variables can be extended to joint distributions. The joint probability distribution of $\mathscr{X}, \mathscr{Y}$ is uniform over a region R with area a if $f(x, y) = 1/a$ whenever $(x, y) \in R$, and 0 otherwise. For example, if R is the rectangular region between (α_1, α_2) and (β_1, β_2), then it is possible to show that $\mathscr{X}$ and $\mathscr{Y}$ are independent uniform distributions over $[\alpha_1, \beta_1]$ and $[\alpha_2, \beta_2]$, respectively.[94]

Exponential Random Variables

°A CONTINUOUS RANDOM VARIABLE with density function

°In the labs

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0, \end{cases}$$

[91] $x^3 - y^3 = (x^2 + xy + y^2)(x - y)$

[92] For a *true* random number generator, visit `http://www.random.org`.

[93] In a double blind test, a group of volunteers for a trial is divided in two subgroups (of the same size); one groups is given the treatment, and the other one is given a placebo. In this way, it is possible to determine if (and to what extent) the treatment is effective. Random numbers are used to guarantee that the division of the groups is really random, and not biased by some potentially hidden factor.

[94] To show this, simply compute the individual distributions of $\mathscr{X}$ and $\mathscr{Y}$.

for some constant λ is called *exponential* (or *exponentially distributed*), and λ is often called the *rate* of the distribution. The cumulative distribution function of such a variable is, for every $x \geq 0$,

$$F(x) = P[\mathscr{X} \leq x] = \int_0^x \lambda e^{-\lambda y} dy = 1 - e^{-\lambda x}.$$

This function often arises as the distribution of the amount of time until some event occurs.[95]

The moment generating function of an exponential is

$$\phi(t) = E[e^{t\mathscr{X}}] = \int_0^\infty e^{tx} \lambda e^{-\lambda x} dx = \lambda \int_0^\infty e^{-(\lambda - t)x} dx = \frac{\lambda}{\lambda - t}. \qquad t < \lambda$$

Differentiating, we get

$$\phi'(t) = \frac{\lambda}{(\lambda - t)^2}$$
$$\phi''(t) = \frac{2\lambda}{(\lambda - t)^3},$$

which implies $E[\mathscr{X}] = \phi'(0) = 1/\lambda$, and $Var(\mathscr{X}) = \phi''(0) - 1/\lambda^2 = 1/\lambda^2$.

The key property of the exponential distribution is that it is *memoryless*; that is, for all $s, t > 0$ it holds that $P[\mathscr{X} > s + t \mid \mathscr{X} > t] = P[\mathscr{X} > s]$.[96] To show this, notice that the notion of memory less is equivalent to stating that $\frac{P[\mathscr{X} > s+t, \mathscr{X} > t]}{P[\mathscr{X} > t]} = P[\mathscr{X} > s]$ or, equivalently, that

$$P[\mathscr{X} > s + t] = P[\mathscr{X} > t] P[\mathscr{X} > s].$$

If $\mathscr{X}$ is an exponential random variable, this equation is satisfied because $e^{-\lambda(s+t)} = e^{-\lambda s} e^{-\lambda t}$. Interestingly, the exponential is the only type of random variable that is memoryless.

Example 59. Consider 3 identical machines that work for an exponentially distributed amount of time with parameter λ. We use two of them until one breaks, and replace that one with the unused one (call it U). What is the probability that the next machine to break is U?

Solution. Due to the memoryless property, when U starts working the remaining lifetime of the originally working machine still available is equivalent to that of U. Hence, the probability of the next one to break to be U is exactly 0.5.

Proposition 60. *If $\mathscr{X}_1, \ldots, \mathscr{X}_n$ are independent exponential random variables with parameters $\lambda_1, \ldots, \lambda_n$, then $\min\{\mathscr{X}_1, \ldots, \mathscr{X}_n\}$ is an exponential with parameter $\sum_{i=1}^n \lambda_i$.*

Proof. The proposition follows because[97]

$$P[\min\{\mathscr{X}_1, \ldots, \mathscr{X}_n\} > x] = P[\mathscr{X}_1 > x, \ldots, \mathscr{X}_n > x] = \prod_{i=1}^n P[\mathscr{X}_i > n]$$
$$= \prod_{i=1}^n e^{-\lambda_i x} = e^{-\sum_{i=1}^n \lambda_i x}. \qquad \square$$

This property is useful, for example, if we are interested in understanding when will the first component of a system will fail. Another useful property is that if $\mathscr{X}$ is exponential with parameter λ, then $c\mathscr{X}$ is exponential with parameter λ/c. Indeed,

$$P[c\mathscr{X} \leq x] = P[\mathscr{X} \leq x/c] = 1 - e^{-\lambda x/c}.$$

[95] For example, the time until it rains, or until you receive a new email tend to be exponentially distributed.

[96] Intuitively, under the condition that the event has not been observed at time t, the probability of having to wait at least s more is exactly the same as having to wait s at the beginning. Conditioning restarts the clock.

[97] If $\mathscr{X}$ is exponential with parameter λ, then $F(x) = P[\mathscr{X} \leq x] = 1 - e^{-\lambda x}$.

Normal Random Variables

A random variable is *normally distributed* with parameters μ and σ^2 (denoted as $\mathscr{X} \sim \mathscr{N}(\mu, \sigma^2)$) if it has density

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}.$$

This is a bell-shaped symmetric curve that has its highest value ($1/\sqrt{2\pi}\sigma$) at μ (see Figure 14).

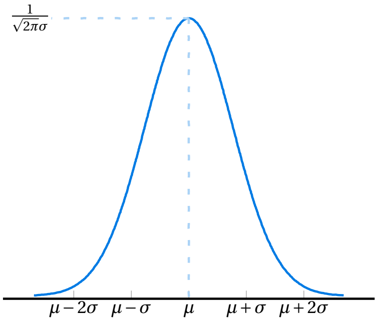


Figure 14: The normal density function.

The importance of the normal distribution and its use in statistics arises from the *central limit theorem* that, in a nutshell, states that when the number of observations is large enough, random phenomena tend to approximate a normal distribution.[98] Examples where this can be observed empirically are the heights of people, or the measuring error in physical quantities.

Intuitively, the mean and variance of a normal RV $\mathscr{X} \sim \mathscr{N}(\mu, \sigma^2)$ should be μ and σ^2, respectively. First, we see that $E[\mathscr{X} - \mu] = 0$,[99] by defining a new variable $y = (x-\mu)/\sigma$:[100]°

$$\begin{aligned} E[\mathscr{X} - \mu] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{(x-\mu)}{\sigma} e^{-(x-\mu)^2/2\sigma^2} dx \\ &= \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y e^{-y^2/2} dy \qquad (*) \\ &= \frac{\sigma}{\sqrt{2\pi}} \left(-e^{-y^2/2} \Big|_{-\infty}^{\infty} \right) = 0. \end{aligned}$$

To compute the variance, we use $u = y$ and $\frac{dv}{dy} = ye^{-y^2/2}$ to see that[101] °

$$\int_{-\infty}^{\infty} y^2 e^{-y^2/2} dy = -ye^{-y^2/2}\Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-y^2/2} dy = \int_{-\infty}^{\infty} e^{-y^2/2} dy.$$

Then, it follows that°

$$\begin{aligned} Var(\mathscr{X}) = E[(\mathscr{X} - \mu)^2] &= \int_{-\infty}^{\infty} \frac{(x-\mu)^2}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} dx \\ &= \sigma^2 \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} y^2 e^{-y^2/2} dy = \sigma^2 \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy = \sigma^2, \end{aligned}$$

where the last equation follows from $\frac{1}{\sqrt{2\pi}} e^{-y^2/2}$ being the density function of a normal (with parameters $\mu = 0$ and $\sigma = 1$); so the integral over all reals must be 1.

An important property is that any linear transformation $a\mathscr{X} + b$ of a normal random variable $\mathscr{X}$ is distributed as a normal too.[102] This means that the variable $Z = \frac{\mathscr{X}-\mu}{\sigma}$ is in fact a normal random variable with mean 0 and variance 1. This is called the *standard* or *unit* normal RV and its distribution is denoted by Φ.[103] We can simplify statements about $\mathscr{X}$ in terms of Z. For example, knowing that $\mathscr{X} < b$ iff $\frac{\mathscr{X}-\mu}{\sigma} < \frac{b-\mu}{\sigma}$ we get that

$$P[\mathscr{X} < b] = P\left[\frac{\mathscr{X}-\mu}{\sigma} < \frac{b-\mu}{\sigma}\right] = \Phi\left(\frac{b-\mu}{\sigma}\right).$$

In other words, for dealing with normal distributions, it suffices to know the values of Φ. There are many software systems and tables providing this

[98] This will be covered more extensively in the next chapter.

[99] Remember that $E[\mathscr{X} + b] = E[\mathscr{X}] + b$.

[100] Using u-substitutions, $\sigma dy = dx$.

°Since $y = (x-\mu)/\sigma$, differentiating on both sides, we get $dy = dx/\sigma$. At (*), we substitute into y, but then we need also to substitute into the differential. Notice the σ that appears outside! For the next one, just give the answer, but verify that it is correct by differentiation.

[101] Recall the *integration by parts* formula: $\int u \frac{dv}{dx} dx = uv - \int v \frac{du}{dx} dx$.

°We have already seen that the integral of $dv = ye^{-y^2/2}$ is $v = -e^{-y^2/2}$.

°Use the same u-substitution $y = (x-\mu)/\sigma$, with $\sigma dy = dx$. The second to last equality follows from the property derived before.

[102] As known already, with mean $a\mu + b$ and variance $a^2\sigma^2$.

[103] That is, $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2} dy$.

information.[104] These tables often provide values only for positive values of x, but since Z is symmetric, negative values can be handled easily. For example, $P[Z < -x] = P[X > x] = 1 - \Phi(x)$. Thus, if $x = 1$, we get that $P[Z < -1] = 1 - \Phi(x) = 1 - 0.8413 = 0.1587$.

Example 61. Data transmission is subject to channel noise disturbances. To reduce the the possibility of error, we encode binary messages using values $-2, 2$, which stand for the original 0 and 1, respectively. Due to noise, when we submit a message $x \in \{-2, 2\}$, it reaches its destination as $R = x + N$. The recipient then decodes the message by concluding, if $R \geq 0.5$ that the message sent was 1, and 0 otherwise. We consider that N has a standard normal distribution.

The probabilities of receiving a wrong signal, due to the noise are

$$P[R < 0.5 \mid signal = 1] = P[N < -1.5] = 1 - \Phi(1.5) = 0.0668$$
$$P[R \geq 0.5 \mid signal = 0] = P[N \geq 2.5] = 1 - \Phi(2.5) = 0.0062. \qquad \triangle$$

THE MOMENT GENERATING FUNCTION of a standard normal RV Z is°

$$E[e^{tZ}] = \int_{-\infty}^{\infty} e^{tx}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx = \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-(x^2-2tx)/2}dx$$
$$= e^{-t^2/2}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-(x-t)^2/2}dx = e^{-t^2/2}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-y^2/2}dy = e^{-t^2/2}.$$

°First, $(x-t)^2 = x^2 - 2tx + t^2$. Later, substitute $y = x - t$, and hence $dy = dx$.

A normal random variable $\mathscr{X}$ with mean μ and variance σ^2 can be expressed as $\sigma Z + \mu$. Thus, it moment generating function is

$$E[e^{t\mathscr{X}}] = E[e^{t\mu + t\sigma Z}] = e^{t\mu}E[e^{t\sigma Z}] = e^{t\mu}e^{-(t\sigma)^2/2} = e^{t\mu - \sigma^2 t^2/2}.$$

We can use this moment generating function to prove that the sum of independent normal random variables is a normal random variable.

Example 62. The height of European males is a normal RV with mean 177.6cm and standard deviation 4cm. If a person has two (adult) sons, find the probability that the older one is taller than the younger one by at least 2cm, assuming that the heights for each child are independent.

Solution. Let $\mathscr{X}_1$ and $\mathscr{X}_2$ be the heights of the first and second child, respectively. Since $-\mathscr{X}_2$ is a normal with mean -177.6 and variance $4^2 = 16$, $\mathscr{X}_1 - \mathscr{X}_2$ is a normal with mean 0 and variance 32. Then

$$P[\mathscr{X}_1 > \mathscr{X}_2 + 2] = P[\mathscr{X}_1 - \mathscr{X}_2 > 2] = P\left[\frac{\mathscr{X}_1 - \mathscr{X}_2}{\sqrt{32}} > \frac{2}{\sqrt{32}}\right]$$
$$= P[Z > 0.3536] \approx 1 - 0.6368 = 0.3632. \qquad \triangle$$

Given any number $\alpha \in (0, 1)$, let z_α be the value such that $P[Z \geq z_\alpha] = \alpha$ (see Figure 15). For example, we have that $z_{0.05} = 1.645$ and $z_{0.01} = 2.33$. The value z_α is the $100(1-\alpha)$ *percentile* of Z.[105]

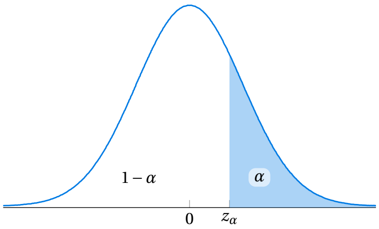


Figure 15: $P[Z > z_\alpha] = \alpha$.

Chi-Square

THE SUM OF n independent standard RVs $\mathscr{X} = \sum_{i=1}^{n} Z_i^2$ is a *chi-square RV with n degrees of freedom*. This is denoted by $\mathscr{X} \sim \chi_n^2$.

[104] See the OLE for one such table. We will see also how to obtain it from R.

[105] The name comes because $100(1-\alpha)$ percent of the time Z will fall below z_α.

The chi-square distribution is *additive* in the sense that if $\mathscr{X}$ and $\mathscr{Y}$ are independent chi-square with n and m degrees of freedom, respectively, then $\mathscr{X} + \mathscr{Y}$ is chi-square with $n + m$ degrees of freedom. This holds because $\mathscr{X} + \mathscr{Y}$ is the sum of $n + m$ squared independent standard RVs.

If $\mathscr{X}$ is chi-square with n degrees of freedom, then for any $\alpha \in (0,1)$ the value $\chi^2_{\alpha,n}$ is the value such that $P[\mathscr{X} \geq \chi^2_{\alpha,n}] = \alpha$.

One application where the chi-square distribution may be of interest is when dealing with measurement errors in multiple dimensions. If we have sensors measuring the position of an object in each dimension, then the square of the distance from the measured to the real value behaves as a chi-square, assuming that the measuring error is normal.[106]

The t-Distribution

IF A STANDARD NORMAL random variable Z and a chi-square random variable with n degrees of freedom χ^2_n are independent, then the RV

$$T_n := \frac{Z}{\sqrt{\chi^2_n / n}}$$

has a *t-distribution with n degrees of freedom*. Figure 16 shows the density function of T_n for different degrees of freedom.

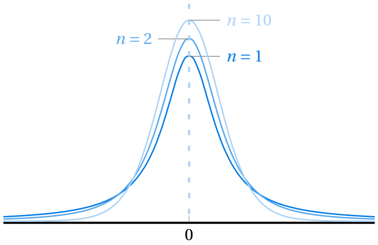


Figure 16: The density function of T_n for $n = 1, 2, 10$.

The t-density is symmetric on 0, and approximates the standard normal density as n grows. Recall that χ^2_n is the sum of the squares of n independent standard normal random variables. From the weak law of large numbers, it follows that as n grows, χ^2_n / n will tend with probability 1 to $E[Z_i^2] = 1$. Thus, for large enough n, $T_n = Z/\sqrt{\chi^2_n / n}$ will approximately have the same distribution as Z (see Figure 17).

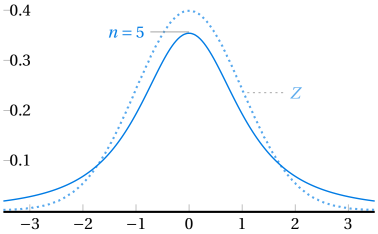


Figure 17: The density function of T_5 (solid) and Z (dotted).

The mean and variance of T_n are, for $n > 1$ and $m > 2$,

$$E[T_n] = 0$$
$$Var(T_n) = \frac{n}{n-2}.$$

Notice that the variance of T_n tends to 1 from above, as n goes to infinity. Given $\alpha \in (0,1)$, let $t_{\alpha,n}$ be such that $P[T_n \geq t_{\alpha,n}] = \alpha$. Since t is symmetric at 0, it follows that

$$\alpha = P[-T_n \geq t_{\alpha,n}] = P[T_n \leq -t_{\alpha,n}] = 1 - P[T_n > t_{\alpha,n}].$$

This means that $P[T_n \geq t_{\alpha,n}] = 1 - \alpha$ or, equivalently, $-t_{\alpha,n} = t_{1-\alpha,n}$. See Figure 18.

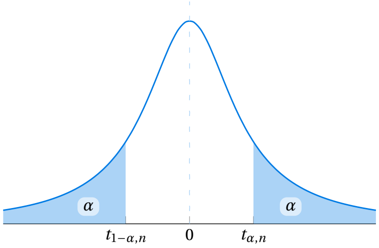


Figure 18: The areas before $-t_{\alpha,n}$ and after $t_{\alpha,n}$ have size α.

[106] We will use it later to estimate parameters of a distribution.

Sampling

THE GOAL OF STATISTICS is to draw conclusions about a large *population* by observing a suitable part (or a sample) of it. For the sample data to yield meaningful information about the whole population, one must make some assumptions about the relationship between the two of them. First, we assume that the measures on the individuals of the population are random variables having a shared distribution among the whole population. If the sample data is chosen randomly, it is reasonable to believe that they are independent random values from this distribution. Formally, a *sample* is a finite set of independent, identically distributed, random variables.[107]

Usually, the population distribution F is not known, and we try to use the data to draw conclusions about it. Sometimes, we may know the general shape of F, but not its precise parameters; for example, that F is normal or Poisson, but with unknown mean and variance. In this case, we speak about *parametric* inferences. If, on the contrary, nothing is assumed about F (except, perhaps, that it is continuous, or discrete), we deal with *non-parametric* inference problems.

From now on, a *statistic* is a random variable whose value is determined by the sample data. Two fundamental such statistics are the *sample mean* and the *sample variance*.

Sample Mean and the Central Limit Theorem

CONSIDER A POPULATION OF elements, each of whom is associated with a value derived from a random variable with mean μ and variance σ^2. These values are known as the *population mean* and *population variance*, respectively.[108] Suppose that we extract a sample from this population, and observe the values $\mathscr{X}_1, \ldots, \mathscr{X}_n$ associated to the elements of the sample. The *sample mean* is

$$\overline{\mathscr{X}} := \frac{\sum_{i=1}^{n} \mathscr{X}_i}{n}.$$

Notice that $\overline{\mathscr{X}}$ is also a random variable; so it has an associated expected value and variance. Since the RVs in a sample are always independent, these are given by

$$E[\overline{\mathscr{X}}] = E\left[\frac{\sum_{i=1}^{n} \mathscr{X}_i}{n}\right] = \frac{1}{n}\sum_{i=1}^{n} E[\mathscr{X}_i] = \mu,$$

$$Var(\overline{\mathscr{X}}) = Var\left(\frac{\sum_{i=1}^{n} \mathscr{X}_i}{n}\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(\mathscr{X}_i) = \frac{\sigma^2}{n}.$$

[107] Also called a *random sample.*

[108] Recall that all the elements in the sample are equally distributed.

In other words, the expected value of the sample mean is the population mean, while its variance is the population variance divided by the size of the sample. Thus, $\overline{\mathscr{X}}$ is centered around μ but its variance tends to 0 as the sample size grows. To understand how the RV $\overline{\mathscr{X}}$ is distributed, we consider the *central limit theorem* which in essence states that the sum of a large number of independent random variables has a distribution that approximates a normal.[109]

Theorem 63 (Central Limit Theorem)**.** *If $\mathscr{X}_1, \ldots, \mathscr{X}_n$ are independent identically distributed random variables with mean μ and variance σ^2, then for a large n the distribution of $\sum_{i=1}^{n} \mathscr{X}_i$ is approximately normal.*[110]

Example 64. An insurance company with 25,000 policy holders observes that the yearly claim of a holder is a RV with mean €320 and standard deviation €540. What is the probability that the total yearly claim is above €8.3 million?

Solution. Let $\mathscr{X}$ be the total yearly claim, and $\mathscr{X}_i$ the yearly claim of client i; then $\mathscr{X} = \sum_{i=1}^{25000} \mathscr{X}_i$. From the central limit theorem, $\mathscr{X}$ is approximately a normal with mean $320 \cdot 25000 = 8 \times 10^6$ and standard deviation $540 \cdot \sqrt{25000} = 8.53 \times 10^4$. Thus,

$$\begin{aligned} P[\mathscr{X} > 8.3 \times 10^6] &= P\left[\frac{\mathscr{X} - 8 \times 10^6}{8.53 \times 10^4} > \frac{8.3 \times 10^6 - 8 \times 10^6}{8.53 \times 10^4}\right] \\ &\approx P[Z > 3.51] \approx 0.00023. \end{aligned}$$

△

The central limit theorem has an important use with binomial random variables. Recall that a binomial $\mathscr{X}$ with parameters (n, p) is a sum of n independent Bernoulli random variables $\mathscr{X} = \sum_{i=1}^{n} \mathscr{X}_i$ with $E[\mathscr{X}_i] = p$ and $Var(\mathscr{X}_i) = p(1-p)$. Then, if n is large enough, $\frac{\mathscr{X} - np}{\sqrt{np(1-p)}}$ approximates a standard normal random variable (see Figure 19).[111]

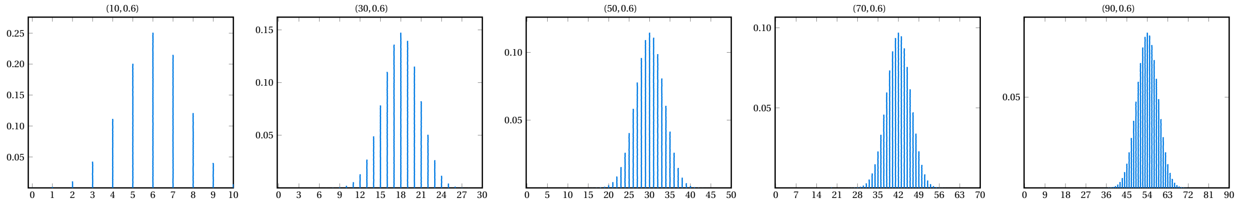


Figure 19: Binomial probability mass functions for $p = 0.6$ with n growing from 10 to 90. The mass converges to a normal density.

Example 65. An airplane fits 150 passengers. On a busy route, only 30% of the people that buy the ticket take the plane. If the airline starts selling 450 tickets per flight, what is the probability that the flight is overbooked?

Solution. Let $\mathscr{X}$ be the number of passengers in the flight. Assuming that each customer will independently decide to take the flight or not, $\mathscr{X}$ is a binomial $(450, 0.3)$. Since the binomial is discrete and the normal continuous, we compute $P[\mathscr{X} = i]$ as $P[i - 0.5 < \mathscr{X} < i + 0.5]$ when using the normal approximation.[112] Then,

$$P[\mathscr{X} > 150.5] = P\left[\frac{\mathscr{X} - 450 \cdot 0.3}{\sqrt{450 \cdot 0.3 \cdot 0.7}} > \frac{150.5 - 450 \cdot 0.3}{\sqrt{450 \cdot 0.3 \cdot 0.7}}\right] \approx P[Z > 1.59] = 0.06.$$

[109] The proof of this theorem is well beyond the scope of the lecture, but this is a fundamental result in statistics.

[110] A consequence of this theorem is that $(\sum_{i=1}^{n} \mathscr{X}_i - n\mu)/(\sigma\sqrt{n})$ approximates a standard normal random variable.

[111] Since the binomial is discrete, and the normal is continuous, sometimes it is necessary to make adjustments (called *continuity corrections*) when using this approximation. See footnote 112.

[112] This is the *continuity correction.*

That is, the probability of overbooking is 6%. △

We have seen two ways to approximate a binomial: the (discrete) Poisson that works well with n large and p small, and the (continuous) normal, which works when $np(1-p)$ is large.[113]

THE CENTRAL LIMIT THEOREM helps to approximate also the distribution of the sample mean. Indeed, $\overline{\mathscr{X}} = \sum_{i=1}^{n} \mathscr{X}_i / n$ will be approximately normal when n is large. In particular, $\frac{\overline{\mathscr{X}}-\mu}{\sigma/\sqrt{n}}$ approximates a standard normal RV.[114] Note that this approximation works, regardless of the distribution of the original RV $\mathscr{X}$.

Example 66. The measurements for the distance d to a star are affected by atmospheric disturbances. To get a precise reading, a measurement is repeated several times, and their average is used as an estimate for the actual distance. Assuming that each measurement is an independent random variable with mean d and standard deviation 2, how many measurements we need to have a 95% certainty of the estimate being within a value of 0.5 from the actual distance?

Solution. After n measurements, the sample mean $\overline{\mathscr{X}}$ of these measurements will be approximately a normal with mean d and standard deviation $2/\sqrt{n}$. Thus,[115]

$$\begin{aligned} P[-0.5 < \overline{\mathscr{X}} - d < 0.5] &= P\left[\frac{-0.5}{2/\sqrt{n}} < \frac{\overline{\mathscr{X}} - d}{2/\sqrt{n}} < \frac{0.5}{2/\sqrt{n}}\right] \\ &\approx P[-\sqrt{n}/4 < Z < \sqrt{n}/4] = 2P[Z < \sqrt{n}/4] - 1. \end{aligned}$$

To get a 95% certainty, we need an n such that $2P[Z < \sqrt{n}/4] - 1 \geq 0.95$; that is, $P[Z < \sqrt{n}/4] \geq 0.975$. Since $P[Z < 1.96] = 0.975$, we need n such that $\sqrt{n}/4 \geq 1.96$. Thus, at least 62 measurements are needed. △

While the central limit theorem guarantees that one will approximate a normal distribution as the sample size n grows, it is not explicit on how large should n be for the approximation to be good enough. In practical terms, a sample size of at least 30 will usually suffice, but depending on the population distribution, a much smaller sample might still work; in some cases, even 5 measurements would be good enough.[116]

Sample Variance

CONSIDER AGAIN A RANDOM sample from a distribution, and let $\overline{\mathscr{X}}$ be the sample mean. The *sample variance* is defined by

$$S^2 := \frac{\sum_{i=1}^{n} (\mathscr{X}_i - \overline{\mathscr{X}})^2}{n-1}.$$

The *sample standard deviation* is $S := \sqrt{S^2}$.

The sample standard deviation is also a random variable, and hence it is interesting to know its expected value. To find out $E[S^2]$ recall that for any sequence of numbers $x_1, \ldots, x_n$, and for $\overline{x} := \sum_{i=1}^{n} x_i / n$,

$$\sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2.$$

[113] In practice, the normal is a good approximation whenever $np(1-p) \geq 10$.

[114] Since $E[\overline{\mathscr{X}}] = \mu$ and $Var(\overline{\mathscr{X}}) = \sigma^2/n$.

[115] Recall that Z is symmetric around 0.

[116] Obviously, the larger the sample, the better the result will be.

It then follows that $(n-1)S^2 = \sum_{i=1}^n \mathscr{X}_i^2 - n\overline{\mathscr{X}}^2$. Taking expectations on both sides, and using the fact that all $\mathscr{X}_i$s are equally distributed,[117]

$$\begin{aligned}(n-1)E[S^2] &= E\left[\sum_{i=1}^n \mathscr{X}_i^2\right] - nE[\overline{\mathscr{X}}^2] \\ &= nE[\mathscr{X}_1^2] - nE[\overline{\mathscr{X}}^2] \\ &= nVar(\mathscr{X}_1) + n(E[\mathscr{X}_1])^2 - nVar(\overline{\mathscr{X}}) - n(E[\overline{\mathscr{X}}])^2 \\ &= n\sigma^2 + n\mu^2 - n(\sigma^2/n) - n\mu^2 = (n-1)\sigma^2.\end{aligned}$$

That is, the expected value of S^2 is the population variance σ^2.

Sampling from a Normal Population

SUPPOSE THAT WE SAMPLE from a normal distribution $\mathscr{X} \sim \mathscr{N}(\mu, \sigma^2)$. Since the sum of independent normal random variables is also normal, it follows that $\overline{\mathscr{X}}$ is also a normal, and in particular $\frac{\overline{\mathscr{X}}-\mu}{\sigma/\sqrt{n}}$ is a standard normal random variable.

Notice that $\sum_{i=1}^n (x_i - \overline{x})^2 = \sum_{i=1}^n (x_i - \mu)^2 - n(\overline{x}-\mu)^2$.[118] Then it follows that

$$\frac{\sum_{i=1}^n (\mathscr{X}_i - \mu)^2}{\sigma^2} = \frac{\sum_{i=1}^n (\mathscr{X}_i - \overline{\mathscr{X}})^2}{\sigma^2} + \frac{n(\overline{\mathscr{X}}-\mu)^2}{\sigma^2}.$$

That is,

$$\sum_{i=1}^n \left(\frac{\mathscr{X}_i - \mu}{\sigma}\right)^2 = \frac{\sum_{i=1}^n (\mathscr{X}_i - \overline{\mathscr{X}})^2}{\sigma^2} + \left(\frac{\sqrt{n}(\overline{\mathscr{X}}-\mu)}{\sigma}\right)^2.$$

Each $(\mathscr{X}_i - \mu)/\sigma$ is a standard normal distribution, and they are all independent. So, the left-hand side of this equation is a chi-square distribution with n degrees of freedom.[119] The second term on the right is also a chi-square with 1 degree of freedom. This suggests, given the additive property of chi-square distributions, that the missing term is also a chi-square with $n-1$ degrees of freedom. Indeed, we get the following important result.

Theorem 67. *If $\mathscr{X}_1, \ldots, \mathscr{X}_n$ is a sample from a normal population with mean μ and variance σ^2, then $\overline{\mathscr{X}}$ and S^2 are independent random variables with $\overline{\mathscr{X}}$ a normal and $(n-1)S^2/\sigma^2$ a chi-square with $n-1$ degrees of freedom.*

Importantly, this theorem not only establishes the distributions of the sample mean and sample variance, but also their independence. The latter only holds for normal distributions.

Corollary 68. $\sqrt{n}\frac{(\overline{\mathscr{X}}-\mu)}{S} \sim t_{n-1}$.[120]

Proof. Theorem 67 states that $\sqrt{n}\frac{(\overline{\mathscr{X}}-\mu)}{\sigma}$ is a standard normal independent from $(n-1)S^2/\sigma^2$, which is a chi-square with $n-1$ degrees of freedom. Hence,

$$\frac{\sqrt{n}(\overline{\mathscr{X}}-\mu)/\sigma}{\sqrt{(n-1)S^2/\sigma^2}} = \sqrt{n}\frac{(\overline{\mathscr{X}}-\mu)}{S}$$

is a t with $n-1$ degrees of freedom. □

[117] Recall that $E[\mathscr{Y}^2] = Var(\mathscr{Y}) + (E[\mathscr{Y}])^2$ for all random variables $\mathscr{Y}$.

[118] Let $y_i = x_i - \mu$. Then $\overline{y} = \overline{x} - \mu$ and

$$\begin{aligned}\sum((x_i-\mu)-(\overline{x}-\mu))^2 &= \sum(y_i - \overline{y})^2 \\ &= \sum y_i^2 - n\overline{y}^2 \\ &= \sum(x_i-\mu)^2 - n(\overline{x}-\mu)^2.\end{aligned}$$

[119] A chi-square with n degrees of freedom is the sum of the squares of n independent standard normal distributions.

[120] That is, $\sqrt{n}\frac{(\overline{\mathscr{X}}-\mu)}{S}$ has a t-distribution with $n-1$ degrees of freedom.

Sampling from a Finite Population

GIVEN A FINITE POPULATION of size N, a *random sample* of size n is such that any of the $\binom{N}{n}$ subsets of the population of size n is equally likely to be chosen.

Suppose that the proportion of the population with a characteristic of interest is p, and that we have a random sample of size n. For $i, 1 \le i \le n$ let $\mathscr{X}_i$ be the indicator variable for the i-th element in the sample having the characteristic. Then $\mathscr{X} = \sum_{i=1}^{n} \mathscr{X}_i$ is the count of elements in the sample having the characteristic, and the sample mean $\overline{\mathscr{X}} = \mathscr{X}/n$ is the proportion of the sample with the characteristic. We study $\mathscr{X}$ and $\overline{\mathscr{X}}$.

Notice that each $\mathscr{X}_i$ is a Bernoulli with parameter p.[121] However, they are not independent: we know already that $P[\mathscr{X}_2 = 1] = p$, but conditioning over the outcome of the first sample, we get $P[\mathscr{X}_2 = 1 \mid \mathscr{X}_1 = 1] = \frac{Np-1}{N-1}$. Knowing that we have already seen an element with the characteristic, the second-sampled element is equally likely to be any of the remaining $N-1$ elements, but only $Np-1$ of them have the property.[122]

When the size of the population N is very large in comparison to the sample size n, the difference between the unconditional and the conditional probabilities can be dismissed.[123] For example, if $N = 2000$ and $p = 0.5$, then $P[\mathscr{X}_2 = 1 \mid \mathscr{X}_1 = 1] = \frac{999}{1999} = 0.4997$; similarly $P[\mathscr{X}_2 = 1 \mid \mathscr{X}_1 = 0] = \frac{1000}{1999} = 0.5003$. Thus, assuming that N is large, we can think of each $\mathscr{X}_i$ as an independent Bernoulli, meaning that $\mathscr{X}$ is approximately a binomial with parameters (n, p).[124] Under this view it follows that°

$$\begin{aligned} E[\mathscr{X}] &= np & Var(\mathscr{X}) &= np(1-p), \\ E[\overline{\mathscr{X}}] &= E[\mathscr{X}]/n = p & Var(\overline{\mathscr{X}}) &= Var(\mathscr{X})/n^2 = p(1-p)/n. \end{aligned}$$

Example 69. Suppose that 40% of the population supports a given political candidate. Given a random sample of 150 individuals, find

1. the expected value and variance of the number of sampled individuals that favour the candidate;
2. the probability that more than half of the sample favours the candidate.[125]

Solution. The number of people $\mathscr{X}$ favouring the candidate in the sample is a binomial with parameters $(150, 0.4)$. So, $E[\mathscr{X}] = 150 \cdot 0.4 = 60$ and $Var(\mathscr{X}) = 150 \cdot 0.4 \cdot 0.6 = 36$.

To compute $P[\mathscr{X} \ge 76]$, we can use the normal approximation:

$$\begin{aligned} P[\mathscr{X} \ge 76] = P[\mathscr{X} \ge 75.5] &= P\left[\frac{\mathscr{X}-60}{6} \ge \frac{75.5-60}{6}\right] \\ &\approx P[Z \ge 2.5833] \approx 0.0049. \end{aligned}$$

△

It is important to notice that, although we considered it here only for the Bernoulli case, in general even if the random variables can take more than two values, when the population is large with respect to the sample size, we can always assume that the sample data are independent random variables with the population distribution.[126]

[121] Every $\mathscr{X}_i$ is an indicator variable for the i-th element from the sample to have the characteristic. But every element from the population is equally likely to be the i-th sampled.

[122] See Example 57.

[123] If we are sampling from the whole Italian population, this is a good assumption. If instead, we sample from the students in the classroom, it might not make much sense.

[124] Notice that $\mathscr{X}$ is in fact a hypergeometric random variable. Thus, we have shown that binomials approximate hypergeometrics when the number of elements is large w.r.t. to the number of selections.

°In the following, we assume that N is very large, and hence $\mathscr{X}$ is a binomial.

[125] When polling we are often trying to answer the dual questions: given the values of the sample, what is the probability that the full population supports the candidate. This is a topic for a different chapter.

[126] And hence, apply the central limit theorem.

Parameter Estimation

One of the main problems in statistics is to understand the distribution of a population given the data from a sample. Often, we can assume that the distribution is known up to a vector of unknown parameters. For example, we may say that it is a normal with unknown mean and variance,[127] or an exponential with unknown rate. In that case, we use the data to estimate the value of the missing parameters. These estimates may be precise (*point estimates*), or give a larger range (*interval estimates*).

Any statistic used to estimate the value of an unknown parameter θ is called an *estimator* of θ. The observed value of this estimator is an *estimate.*

Maximum Likelihood Estimators

Suppose that we observe a sequence $\mathscr{X}_1,\dots,\mathscr{X}_n$ of random variables whose joint distribution is known except for an unknown parameter θ. We want to use the observed values to understant θ. If, for example, the observations are all from independent exponential random variables with the same unknown rate θ, their joint density function is

$$\begin{aligned} f(x_1,\dots,x_n) &= f_{\mathscr{X}_1}(x_1)\cdots f_{\mathscr{X}_n}(x_n) \\ &= \prod_{i=1}^{n}\frac{1}{\theta}e^{-x_i/\theta} && (0 < x_i < \infty) \\ &= \frac{1}{\theta^n}exp\left(-\sum_{i=1}^{n}x_i/\theta\right). && (0 < x_i < \infty) \end{aligned}$$

One of the most used type of estimator is the *maximum likelihood* estimator. Let $f(x_1,\dots,x_n \mid \theta)$ be the joint probability mass or density function of the variables $\mathscr{X}_1,\dots,\mathscr{X}_n$;[128] this is a function of θ since this parameter is unknown. This function f represents the likelihood of observing $x_1,\dots,x_n$ when θ is the *true* value of the parameter. The maximum likelihood estimate $\hat{\theta}$ of θ is the value that maximises $f(x_1,\dots,x_n \mid \theta)$ given the observation $x_1,\dots,x_n$. One useful property for computing this estimator is that $f(x_1,\dots,x_n \mid \theta)$ and $\log\big(f(x_1,\dots,x_n \mid \theta)\big)$ are maximised at the same value of θ so it suffices to maximise the latter function.

Maximum Likelihood Estimator of a Bernoulli Parameter

[127] Recall the law of large numbers.

[128] Depending on whether they are discrete or jointly continuous.

CONSIDER A SEQUENCE OF n independent trials that have each a probability of success p, which we want to estimate.[129] The trial consists of n Bernoulli random variables $\mathscr{X}_i$ with an unknown parameter p. Then, for $x = 0, 1$ we have $P[\mathscr{X}_i = x] = p^x(1-p)^{1-x}$, and the joint probability mass function of the data is

$$\begin{aligned} f(x_1, \ldots, x_n \mid p) &= P[\mathscr{X}_1 = x_1, \ldots, \mathscr{X}_n = x_n \mid p] \\ &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}. \end{aligned}$$

We want to determine the value of p that maximises this function. This is easier to do by taking logarithms:

$$\log(f(x_1, \ldots, x_n \mid p)) = \sum_{i=1}^{n} x_i \log p + \left(n - \sum_{i=1}^{n} x_i\right) \log(1-p).$$

To optimise, we can find the point where the derivative (w.r.t. p) evaluates to 0; that is,

$$\frac{d}{dp}\log(f(x_1, \ldots, x_n \mid p)) = \frac{\sum_{i=1}^n x_i}{p} - \frac{n - \sum_{i=1}^n x_i}{1-p} = 0.$$

Solving this equation yields the estimator $\frac{\hat{p}}{\sum_{i=1}^n x_i} = \frac{1-\hat{p}}{n-\sum_{i=1}^n x_i}$, or equivalently

$$\hat{p} = \frac{\sum_{i=1}^n x_i}{n}.$$

That is, the maximum likelihood estimator of p is the proportion of successes observed in the trials.

Maximum Likelihood Estimator of a Poisson Parameter

°CONSIDER NOW INDEPENDENT POISSON random variables $\mathscr{X}_1, \ldots, \mathscr{X}_n$ all with the same (unknown) parameter λ. To estimate λ, we first compute the likelihood function

°Optional

$$f(x_1, \ldots, x_n \mid \lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}.$$

Taking the logarithm yields

$$\log\bigl(f(x_1, \ldots, x_n \mid \lambda)\bigr) = -n\lambda + \sum_{i=1}^{n} x_i \log(\lambda) - \log(\prod_{i=1}^{n} x_i).$$

To find the maximum, we evaluate the derivative (w.r.t. λ) at 0

$$\frac{d}{d\lambda}\log\bigl(f(x_1, \ldots, x_n \mid \lambda)\bigr) = -n + \frac{\sum_{i=1}^n x_i}{\lambda} = 0,$$

which yields $\hat{\lambda} = \frac{\sum_{i=1}^n x_i}{n}$.

For example, suppose that the number of customers at a bar in one hour is a Poisson with an unknown parameter λ that we want to estimate. If in a working day of 12 hours, the bar gets 180 customers, then the maximum likelihood estimate for λ is $156/12 = 15$.[130]

[129] This is the estimator of the parameter of a Bernoulli distribution.

[130] That is, on average, there will be 15 customers every hour.

Estimators for Normal Parameters

CONSIDER NOW THE MORE complex case of estimating both parameters from a normal distribution.[131] If $\mathcal{X}_1,\ldots,\mathcal{X}_n$ are independent identically distributed normal random variables, their joint density, and its logarithm are

$$\begin{aligned} f(x_1,\ldots,x_n \mid \mu,\sigma) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} exp\left(\frac{-(x_i-\mu)^2}{2\sigma^2}\right) \\ &= \left(\frac{1}{2\pi}\right)^{n/2} \frac{1}{\sigma^n} exp\left(\frac{-\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right) \\ \log\big(f(x_1,\ldots,x_n \mid \mu,\sigma)\big) &= -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{-\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2} \end{aligned}$$

As usual, the values of μ and σ that maximise this function can be found by differentiating and evaluating at 0.

$$\begin{aligned} \frac{\partial}{\partial\mu}\log\big(f(x_1,\ldots,x_n \mid \mu,\sigma)\big) &= \frac{\sum_{i=1}^{n}(x_i-\mu)}{\sigma^2} = 0, \\ \frac{\partial}{\partial\sigma}\log\big(f(x_1,\ldots,x_n \mid \mu,\sigma)\big) &= -\frac{n}{\sigma} + \frac{-\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^3} = 0. \end{aligned}$$

This yields, when solving,

$$\begin{aligned} \hat{\mu} &= \sum_{i=1}^{n} x_i/n \\ \hat{\sigma} &= \left(\sum_{i=1}^{n}(x_i-\hat{\mu})^2/n\right)^{1/2}. \end{aligned}$$

Note that the maximum likelihood estimator for the standard deviation is *not* the same as the sample standard deviation,[132] but their difference decreases as n grows.

IN ALL THE PREVIOUS cases, the maximum likelihood estimator for the mean was the sample mean. We show that this is not necessarily the case. If $\mathcal{X}_1,\ldots,\mathcal{X}_n$ are a sample from a uniform distribution over $(0,\theta)$, where θ is unknown, then we have

$$f(x_1,\ldots,x_n \mid \theta) = \begin{cases} \frac{1}{\theta^n} & 0 \le x_i \le \theta, 1 \le i \le n \\ 0 & \text{otherwise.} \end{cases}$$

To maximise this function, we need to choose θ to be as small as possible. Since θ cannot be smaller than any of the observed x_is, the maximum likelihood estimator of θ is $\hat{\theta} = \max\{x_1,\ldots,x_n\}$. Thus, the maximum likelihood estimator for the mean of the distribution is $\max\{x_1,\ldots,x_n\}/2$.[133]

Interval Estimates

SUPPOSE THAT WE HAVE a sample from a normal population with unknown mean μ and *known* variance σ^2. The maximum likelihood estimator for μ is $\overline{\mathcal{X}} = \sum_{i=1}^{n} \mathcal{X}_i/n$. However, we do not expect the value of $\overline{\mathcal{X}}$

[131] In this case, the mean μ and the standard deviation σ.

[132] The former divides by n while the latter divides by $(n-1)$.

[133] Remember that the mean of a uniform distribution over $(0,a)$ is $a/2$.

to be *exactly* μ. For this reason, we often prefer to compute an interval guaranteed to contain μ (with some degree of confidence). This interval is found through the probability distribution of the point estimator. In this example, $\overline{\mathscr{X}} \sim \mathcal{N}(\mu, \sigma^2/n)$ and hence $\frac{\overline{\mathscr{X}}-\mu}{\sigma/\sqrt{n}} \sim Z$. Thus, it follows that[134] °

$$\begin{aligned} 0.95 &= P\left[-1.96 < \frac{\sqrt{n}}{\sigma}(\overline{\mathscr{X}} - \mu) < 1.96\right] \\ &= P\left[-1.96\frac{\sigma}{\sqrt{n}} < \overline{\mathscr{X}} - \mu < 1.96\frac{\sigma}{\sqrt{n}}\right] \\ &= P\left[-1.96\frac{\sigma}{\sqrt{n}} < \mu - \overline{\mathscr{X}} < 1.96\frac{\sigma}{\sqrt{n}}\right] && \text{(symmetry)} \\ &= P\left[\overline{\mathscr{X}} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \overline{\mathscr{X}} + 1.96\frac{\sigma}{\sqrt{n}}\right]. \end{aligned}$$

In other words, there is a 95% chance that the *true mean* μ lies within a distance $1.96\frac{\sigma}{\sqrt{n}}$ from $\overline{\mathscr{X}}$. If we observe that $\overline{\mathscr{X}} = \overline{x}$, then we have a 95% confidence that μ is in the interval

$$\left(\overline{x} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{x} + 1.96\frac{\sigma}{\sqrt{n}}\right).$$

This is known as the *95 percent confidence interval* of μ.[135]

Example 70. When sending a signal μ over a noisy channel, it is received as $\mu + N$ where N is a normal with mean 0 and variance 4. To reduce the error, each signal is sent 9 times, and the receiver is tasked with estimating the original signal. Suppose that we receive the signals 5, 8.5, 12, 15, 7, 9, 7.5, 6.5, and 10.5. Then $\overline{x} = 81/9 = 9$. Through the maximum likelihood estimator, we conclude that the signal was 9. To have a better understanding of the signal, however, we decide to compute its 95% confidence interval, which is

$$(9 - 1.96\frac{2}{3}, 9 + 1.96\frac{2}{3}) = (7.69, 10.31).$$

We can guarantee with 95% confidence that the *true* original signal lies in this interval. △

An interval like the one from this example is known as a *two-sided confidence interval* since we allow for the real mean μ to be on either side of the point estimator. Alternatively, we may prefer a *one-sided confidence interval*; for example, finding a value x such that we can assert, with 95% confidence, that $\mu > x$. To find this value x, notice that $P[Z < 1.645] = 0.95$. This means that

$$0.95 = P\left[\sqrt{n}\frac{\overline{\mathscr{X}} - \mu}{\sigma} < 1.645\right] = P\left[\overline{\mathscr{X}} - 1.645\frac{\sigma}{\sqrt{n}} < \mu\right].$$

The 95% *one-sided upper confidence interval* for μ is $(\overline{x} - 1.645\sigma/\sqrt{n}, \infty)$, where $\overline{x}$ is the observed value of the sample mean. Similarly, the 95% *one-sided lower confidence interval* for μ is $(-\infty, \overline{x} + 1.645\sigma/\sqrt{n})$.

Example 71. The upper 95% confidence interval estimate of μ from Example 70 is $(9 - 1.645\frac{2}{3}, \infty) = (7.903, \infty)$. △

The use of 95% in the previous discussion is completely arbitrary; it is possible to compute the confidence intervals for any desired level of confidence.[136] Recall that for any $\alpha \in (0,1)$, $P[-z_{\alpha/2} < Z < z_{\alpha/2}] = 1 - a$;[137] see

[134] Recall Example 66.

°Make a drawing!

[135] One important note on confidence intervals: a "95% confidence interval" for μ does not mean that μ has a probability of 95% of belonging to this interval; in fact, there are no random variables involved to make a probabilistic statement. What it says is that when using this method, in 95% of the cases, the interval will in fact contain the *true mean* μ (but we will miss it 5% of the time).

[136] In practice, one would like the confidence to be high, but the interval not to be too wide.

[137] For $\alpha \in (0,1)$, z_α is the value such that $P[Z < z_\alpha] = \alpha$.

Figure 20. In the computation of the 95% confidence interval, we used the fact that for $\alpha = 0.05$, $z_{\alpha/2} = z_{0.025} = 1.96$. Following the same computations, we can generalise the idea to *any* confidence $1-\alpha$: the $100(1-\alpha)$ percent two-sided confidence interval for μ is

$$(\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}).$$

Similarly, the one-sided confidence intervals for the same level of confidence are $(\overline{x} - z_\alpha\frac{\sigma}{\sqrt{n}}, \infty)$, and $(-\infty, \overline{x} + z_\alpha\frac{\sigma}{\sqrt{n}})$, respectively, where $\overline{x}$ is the observed sample mean.

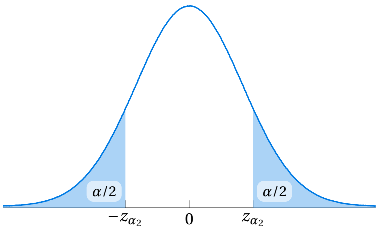


Figure 20: $P[-z_{\alpha/2} < Z < z_{\alpha/2}] = 1-\alpha$.

Example 72. Suppose that in Example 70° we are interested in the 99% two-sided, and one sided upper confidence interval estimates for μ. Then, knowing that $z_{0.005} = 2.58$ and $z_{0.01} = 2.33$, we get that these intervals are $9 \pm 2.58{}^2/_3 = (7.28, 10.72)$ and $(9 - 2.33{}^2/_3, \infty) = (7.447, \infty)$, respectively.

°$\overline{x} = 9, n = 9$

SOMETIMES, WE ARE INTERESTED in finding a small confidence interval, without reducing the level of confidence. For example, we may want to guarantee with 99% confidence, that μ is within an interval of length 1.[138] Since $z_{0.005} = 2.58$, the 99% confidence interval for μ over a sample of size n is $(\overline{x} - 2.58\frac{\sigma}{\sqrt{n}}, \overline{x} + 2.58\frac{\sigma}{\sqrt{n}})$, which has length $5.16\frac{\sigma}{\sqrt{n}}$. Thus, to obtain an interval of length 1, we need to choose an n such that $5.16\frac{\sigma}{\sqrt{n}} = 1$; that is, $n = (5.16\sigma)^2$.[139]

Example 73. A packaging machine fills a product to a random weight with unknown mean and standard deviation of 0.5kg. To be 95% certain that our estimate of the mean is correct to ±200g, then we need a sample size n such that $1.96\frac{\sigma}{\sqrt{n}} \le 0.2$.[140] That is, $n \ge (9.8 \cdot 0.5)^2 = 4.9^2 = 24.01$. △

SUPPOSE NOW THAT WE want to construct a confidence interval for μ from a sample of a normal distribution with unknown mean and variance. Since σ is unknown, we cannot simply build a standard normal distribution as before.[141] However, for the sample standard deviation S,[142] we know from Corollary 68 that $\sqrt{n}\frac{(\overline{\mathscr{X}}-\mu)}{S}$ is a t-random variable with $n-1$ degrees of freedom.° Thus, for every $\alpha \in (0,1)$,

$$P\left[-t_{\alpha/2,n-1} < \sqrt{n}\frac{(\overline{\mathscr{X}}-\mu)}{S} < t_{\alpha/2,n-1}\right] = 1-\alpha.$$

°We know that $\sqrt{n}(\overline{\mathscr{X}}-\mu)/\sigma \sim Z$ and that $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$. So

$$\frac{\sqrt{n}(\overline{\mathscr{X}}-\mu)/\sigma}{\sqrt{(n-1)S^2/(\sigma^2(n-1))}} \sim T_{n-1}.$$

Following an analogous argument as for the normal before, if $\overline{x}$ and s are the observed values for $\overline{\mathscr{X}}$ and S, respectively, we can say with $100(1-\alpha)$ confidence that $\mu \in (\overline{x} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}, \overline{x} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}})$.[143]

Example 74. °Consider again Example 70, but with an unknown noise variance. To compute a 95% confidence interval for the signal μ, we use again $\overline{x} = 9$, and compute $s^2 = \frac{\sum_{i=1}^n x_i^2 - 9(\overline{x})^2}{8} = 9.5$,[144] that is, $s = 3.802$. Knowing that $t_{0.025,8} = 2.306$, a 95% confidence interval for μ is

$$\left(9 - 2.306\frac{3.802}{3}, 9 + 2.306\frac{3.802}{3}\right) = (6.63, 11.37). \qquad \triangle$$

°Can be skipped if data is not available from before.

Similarly, we can compute the one-sided upper and lower confidence intervals, respectively as $(\overline{x} - t_{\alpha,n-1}\frac{S}{\sqrt{n}}, \infty)$ and $(-\infty, \overline{x} + t_{\alpha,n-1}\frac{S}{\sqrt{n}})$.

[138] In this case, the only parameter that we can manipulate is the sample size.

[139] This can be generalised to other levels of confidence in the obvious way.

[140] Remember that $z_{0.025} = 1.96$.

[141] By saying that $\sqrt{n}(\overline{\mathscr{X}}-\mu)/\sigma \sim \mathcal{N}(0,1)$.

[142] $S = \sum_{i=1}^n(\mathscr{X}_i - \overline{\mathscr{X}})/(n-1)$.

[143] Compare this for the case where the variance σ is known: the formula is essentially the same, but now we use an estimate for σ, and need to correct through a t distribution.

[144] Recall that we have previously shown that $(n-1)S^2 = \sum_{i=1}^n \mathscr{X}_i^2 - n\overline{\mathscr{X}}^2$.

To recall, when the standard deviation is known, the computation of the confidence interval is based on the fact that $\sqrt{n}(\overline{\mathscr{X}}-\mu)/\sigma \sim \mathscr{N}(0,1)$, while when σ is unknown we estimate S, and use that $\sqrt{n}(\overline{\mathscr{X}}-\mu)/S \sim t_{n-1}$. Notice that the confidence interval does not need to be larger when σ is unknown. In fact, it is completely possible that the sample variance turns out to be much smaller than the population variance. However, we can show that the *mean length* of the interval is larger when σ is unknown. More precisely, we can show that $t_{\alpha,n-1}E[S] \geq z_\alpha \sigma$.

For this section, we have assumed that the sample was taken from a normal distribution. Using the Central Limit Theorem, we can extend these results to any population distribution, provided that the sample is large enough. Indeed, in that case, $\sqrt{n}(\overline{\mathscr{X}}-\mu)/\sigma$ will be approximately normal, and hence $\sqrt{n}(\overline{\mathscr{X}}-\mu)/S$ will approximate a t.

CONSIDER NOW A SAMPLE from a normal distribution with unknown μ and σ, which we want to use to predict the value of a newly sampled element.[145] A natural point predictor for that value is the sample mean $\overline{\mathscr{X}}$. If we want an interval predictor, we can notice that $\overline{\mathscr{X}}$ is a normal with mean μ and variance σ^2/n, and is independent of $\mathscr{X}_{n+1} \sim \mathscr{N}(\mu,\sigma^2)$. Thus, $\mathscr{X}_{n+1}-\overline{\mathscr{X}} \sim \mathscr{N}(0,\sigma^2/n+\sigma^2)$ or, in other words,

$$\frac{\mathscr{X}_{n+1}-\overline{\mathscr{X}}}{\sigma\sqrt{1+1/n}} \sim \mathscr{N}(0,1).$$

Since σ remains unknown, we need to estimate it too. As before, using the sample variance $S^2 = \sum_{i=1}^n (\mathscr{X}_i - \overline{\mathscr{X}})^2/(n-1)$, which is independent of the previous random variable,[146] we can conclude that

$$\frac{\mathscr{X}_{n+1}-\overline{\mathscr{X}}}{S\sqrt{1+1/n}} \sim t_{n-1}.$$

Hence, we get that for any $\alpha \in (0,1)$

$$\begin{aligned}1-\alpha &= P\left[-t_{\alpha/2,n-1} < \frac{\mathscr{X}_{n+1}-\overline{\mathscr{X}}}{S\sqrt{1+1/n}} < t_{\alpha/2,n-1}\right] \\ &= P[\overline{\mathscr{X}} - t_{\alpha/2,n-1}S\sqrt{1+1/n} < \mathscr{X}_{n+1} < \overline{\mathscr{X}} + t_{\alpha/2,n-1}S\sqrt{1+1/n}].\end{aligned}$$

In other words, if $\overline{x}$ and s are the observed values of $\overline{\mathscr{X}}$ and S, respectively, we can predict with $100(1-\alpha)$ percent confidence that

$$\mathscr{X}_{n+1} \in \left(\overline{x} - t_{\alpha/2,n-1}s\sqrt{1+1/n}, \overline{x} + t_{\alpha/2,n-1}s\sqrt{1+1/n}\right).$$

JUST AS WE DID for the mean, one can also construct confidence intervals for an unknown variance σ^2. In this case, recall that $(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$.[147] Since chi-square distributions are not symmetric and are positive,[148] we need to be careful when building the confidence intervals. In this case,

$$\begin{aligned}1-\alpha &= P\left[\chi^2_{1-\alpha/2,n-1} \leq (n-1)\frac{S^2}{\sigma^2} \leq \chi^2_{\alpha/2,n-1}\right] \\ &= P\left[\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right].\end{aligned}$$

[145] That is, we observe $\mathscr{X}_1,\ldots,\mathscr{X}_n$, and we want to predict the value of $\mathscr{X}_{n+1}$.

[146] See Theorem 67.

[147] Theorem 67.

[148] A chi-square is a sum of squares, so it must take only positive values.

In other words, if s^2 is the observed sample variance, then a $100(1-\alpha)$ confidence interval for σ^2 is

$$\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right).$$

One sided intervals can be computed following the same ideas.

Estimating the Difference of Means

CONSIDER NOW TWO INDEPENDENT samples $\mathscr{X}_1,\ldots,\mathscr{X}_n$ from a normal population with mean μ_1 and variance σ_1^2, and $\mathscr{Y}_1,\ldots,\mathscr{Y}_m$ from a different normal population with mean μ_2 and variance σ_2^2. We want to estimate $\mu_1-\mu_2$.[149] It can be shown that $\overline{\mathscr{X}}-\overline{\mathscr{Y}}$ is the maximum likelihood estimator for $\mu_1-\mu_2$, where $\overline{\mathscr{X}}$ and $\overline{\mathscr{Y}}$ are the sample means.

To find confidence intervals, we need first to understand the distribution of $\overline{\mathscr{X}}-\overline{\mathscr{Y}}$. We know that $\overline{\mathscr{X}} \sim \mathscr{N}(\mu_1,\sigma_1^2/n)$ and $\overline{\mathscr{Y}} \sim \mathscr{N}(\mu_2,\sigma_2^2/m)$. This means that[150]

$$\overline{\mathscr{X}}-\overline{\mathscr{Y}} \sim \mathscr{N}\left(\mu_1-\mu_2, \frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}\right),$$

and hence

$$\frac{\overline{\mathscr{X}}-\overline{\mathscr{Y}}-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}} \sim \mathscr{N}(0,1).$$

If the two variances are known, then it easy to verify that the two-sided and one-sided $100(1-\alpha)$ confidence interval estimates for $\mu_1-\mu_2$ are, respectively,

$$\left(\overline{x}-\overline{y}-z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}, \overline{x}-\overline{y}+z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}\right);$$

$$\left(-\infty, \overline{x}-\overline{y}+z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}\right); \text{ and}$$

$$\left(\overline{x}-\overline{y}-z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}, \infty\right);$$

where $\overline{x}$ and $\overline{y}$ are the observed sample means.

If the variances are unknown, then we can try to use the same idea of estimating them through the sample variances

$$S_1^2 = \sum_{i=1}^{n}(\mathscr{X}_i-\overline{\mathscr{X}})/(n-1)$$
$$S_2^2 = \sum_{i=1}^{m}(\mathscr{Y}_i-\overline{\mathscr{X}})/(m-1),$$

and use $\frac{\overline{\mathscr{X}}-\overline{\mathscr{Y}}-(\mu_1+\mu_2)}{\sqrt{S_1^2/n+S_2^2/m}}$ to construct a confidence interval. However, we do not know how this random variable is distributed. Unfortunately, this distribution is very difficult to understand in general. So we focus on the special case where $\sigma_1=\sigma_2$.[151]

[149] For example, if we want to know if a system is faster than another.

[150] Remember that the sum of independent normals is a normal.

[151] This is, in fact, the only case that we can feasibly handle.

Suppose that the population variances are both equal to an unknown σ^2. In that case, we know[152]

$$(n-1)\frac{S_1^2}{\sigma^2} \sim \chi^2_{n-1},$$
$$(m-1)\frac{S_2^2}{\sigma^2} \sim \chi^2_{m-1}.$$

Independence of the samples implies that these two chi-square distributions are also independent, and hence[153]

$$(n-1)\frac{S_1^2}{\sigma^2} + (m-1)\frac{S_2^2}{\sigma^2} \sim \chi^2_{n+m-2}.$$

We have also already established that

$$\frac{\overline{\mathscr{X}} - \overline{\mathscr{Y}} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}} \sim \mathscr{N}(0,1),$$

and that $\overline{\mathscr{X}}, \overline{\mathscr{Y}}, S_1$, and S_2 are independent random variables. Define now

$$S_p^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}.$$

Then it follows that[154]

$$\frac{\overline{\mathscr{X}} - \overline{\mathscr{Y}} - (\mu_1 - \mu_2)}{\sqrt{S_p^2(1/n + 1/m)}} = \frac{\overline{\mathscr{X}} - \overline{\mathscr{Y}} - (\mu_1 - \mu_2)}{\sqrt{\sigma^2(1/n+1/m)}} \Bigg/ \sqrt{S_p^2/\sigma^2}$$

is a t-distribution with $n+m-2$ degrees of freedom. Thus, the $100(1-\alpha)$ confidence interval for $\mu_1 - \mu_2$ is[155]

$$\left(\overline{x} - \overline{y} - t_{a/2,n+m-2} s_p \sqrt{1/n + 1/m}, \overline{x} - \overline{y} - t_{a/2,n+m-2} s_p \sqrt{1/n+1/m}\right).$$

WE NEEDED TO ESTIMATE the unknown variance σ^2 to compute this confidence interval. In this case, we had two different samples that provided two estimates for this variance. The estimator S_p^2 used is the weighted average of the two sample variances where the weights are proportional to their degrees of freedom. In general, this is called the *pooled estimator*.

Approximate Confidence Interval for the Mean of a Bernoulli

SUPPOSE THAT WE WANT to estimate the parameter p of a Bernoulli. If $\mathscr{X}$ denotes the number of successes observed in a sample of size n, then $\mathscr{X}$ is a binomial (n,p). If the sample is large enough, we know that $\mathscr{X}$ is approximately a normal with mean np and variance $np(1-p)$. Thus, for $\alpha \in (0,1)$

$$P\left[-z_{\alpha/2} < \frac{\mathscr{X} - np}{\sqrt{np(1-p)}} < z_{\alpha/2}\right] \approx 1-\alpha.$$

Then an approximate confidence *region* for p is

$$\left\{p \,\middle|\, -z_{\alpha/2} < \frac{\mathscr{X} - np}{\sqrt{np(1-p)}} < z_{\alpha/2}\right\}.$$

[152] See again Theorem 67.

[153] Recall that the sum of two independent chi-squares is a chi-square with degrees of freedom equal to the sum of the degrees of freedom.

[154] $(n+m-2)S_p^2/\sigma^2 = (n-1)\frac{S_1^2}{\sigma^2} + (m-1)\frac{S_2^2}{\sigma^2}$.

[155] The one-sided intervals can be easily derived in the same way.

However, this is *not* an interval. In order to find an interval, consider $\hat{p} = \mathcal{X}/n$. Since $\hat{p}$ is the maximum likelihood estimator of p, it should be approximately equal to p, and $\sqrt{n\hat{p}(1-\hat{p})} \approx \sqrt{np(1-p)}$. Thus, $\frac{\mathcal{X}-np}{\sqrt{n\hat{p}(1-\hat{p})}}$ is still approximately a standard normal RV. This implies then that

$$P\left[\hat{p} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} < p < \hat{p} + z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}\right] \approx 1-\alpha,$$

which yields the (approximate) confidence interval for p.

Example 75. A poll result expresses that 52% of the population favours a candidate with a margin of error of ±4%. Knowing that polls are commonly expressed as 95% confidence intervals, what does this statement mean, and what can we say about the size of the poll?

Solution. A 95% confidence interval for the proportion p of people in favour of the candidate given a sample of size n is

$$\hat{p} \pm z_{0.025}\sqrt{\hat{p}(1-\hat{p})/n} = 0.52 \pm 1.96\sqrt{0.52 \cdot 0.48/n}.$$

Since the margin of error is ±4%, we can conclude that

$$1.96\sqrt{0.52 \cdot 0.48/n} = 0.04.$$

Which, solving for n yields $n = \frac{(1.96)^2(0.52)(0.48)}{(0.04)^2} = 599.29$. That is, approximately 599 people were polled. △

As before, we may be interested in finding a confidence interval of a bounded size b, and need to find out the size of the sample needed to achieve this. Notice that the length of this interval is $2z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$. Unfortunately, we do not known $\hat{p}$ (or even p) in advance, so we cannot use this bound to determine the right size n. The trick in this case is to start with a partial sample. With the first sample, we compute an estimate p^* of p. Then, using p^* as an approximation for p (and $\hat{p}$), we compute the desired sample size by solving

$$\begin{aligned} b &= 2z_{\alpha/2}\sqrt{p^*(1-p^*)/n} \\ b^2 &= (2z_{\alpha/2})^2 p^*(1-p^*)/n \\ n &= \frac{(2z_{\alpha/2})^2 p^*(1-p^*)}{b^2}. \end{aligned}$$

If the original partial sample had k elements, then we need to find an additional sample of $n-k$ elements.

Alternatively, notice that the $100(1-\alpha)\%$ confidence interval of p will have length b if n is at least $\frac{(2z_{\alpha/2})^2}{b^2}p(1-p)$, which is maximised when $p = 1/2$. Thus, any sample size $n \geq \frac{z^2_{\alpha/2}}{b^2}$ will be such that the confidence interval has length at most b.[156]

[156] This sample will often be too large in comparison to the bound found before, but it requires no preliminary sampling, and no additional computations.

Hypothesis Testing

HYPOTHESIS TESTING IS A close relative to parameter estimation. The difference is that, rather than trying to find out the unknown parameters of a distribution, hypothesis testing tries to verify—or refute—a statement (hypothesis) about these parameters.[157] For example, we can try to verify whether the mean height of unibz students is 200cm by taking a sample and computing their height. We want to develop a procedure that verifies whether the observed values are consistent with the hypothesis. If the value derived from the sample is inconsistent with our hypothesis (e.g. we get a sample mean of 170cm), then we can *reject* the hypothesis. Otherwise (e.g., if the sample mean is 198cm), we cannot.[158] Usually, we will want to test the opposite of our hypothesis, to be able to reject it.

Significance Levels

SUPPOSE THAT WE WANT to test a hypothesis H_0 about a parameter θ. We often call H_0 the *null hypothesis.*[159] For example, if θ is the mean of a normal distribution, one could generate hypotheses such as

$$H_0 : \theta = 1 \qquad \text{or} \qquad H_0 : \theta \leq 1.$$

Notice that the latter, when true, will not specify the distribution of the population. A hypothesis of this kind is called *composite* as it refers to a set of possible values. A hypothesis that fully specifies the distribution, like the former, is *simple*.

Given a sample of size n, we need to decide when to reject the hypothesis. Formally, a test for H_0 is defined through a region in $\mathbb{R}^n$, called the *critical region*. The idea is that C defines the region where H_0 will be rejected; i.e., we reject H_0 if the sample $(\mathscr{X}_1, \dots, \mathscr{X}_n) \in C$. One common test for the hypothesis that the mean θ of a normal with variance 1 is 1 is the critical region[160]

$$C = \{(\mathscr{X}_1, \dots, \mathscr{X}_n) \mid |\overline{\mathscr{X}} - 1| > 1.96\sqrt{n}\}.$$

That is, we reject the null hypothesis $\theta = 1$ if the sample mean differs from 1 by at least $1.96\sqrt{n}$.°

Recall that the idea is to check whether H_0 is consistent with the observations in the data. So we should only reject it if the data is very unlikely when H_0 is true. We do this by specifying a value α, called the *significance*

[157] It is a *hypothesis* because it is not known whether it is true or not.

[158] In the literature, people often speak of *accepting* a hypothesis. This is a confusing use of the word that lead to misinterpretations. When one *rejects* a hypothesis, it means that there is strong evidence against it. In that sense *accepting* a hypothesis would mean only that there is no evidence against it (but there may also be no strong evidence in its favour).

[159] As mentioned before, the null hypothesis is often the opposite of our claim and we want to reject it.

[160] Compare this with 95% confidence intervals.

°Say something about Type I and Type II errors: rejecting H_0 when it is correct, and not rejecting it when it is wrong.

level such that H_0 has probability at most α of being rejected when it is true.[161] It is good practice to decide this α in advance; in life sciences this is often set to $0.1, 0.05$, or 0.005.

The usual approach to hypothesis testing is the following. Given a hypothesis $H_0 : \theta \in \omega$, where ω is a set of parameter values, we find a point estimator $\breve{\theta}$ for θ, and reject H_0 if $\breve{\theta}$ is *far* from ω. Obviously, the specific distance required depends on the distribution of $\breve{\theta}$ under the assumption of H_0 being true, and on the desired significance level. For example, the critical region computed before rejects the hypothesis if the sample mean[162] is further than $1.96/\sqrt{n}$ from 1. This was chosen to meet the significance level $\alpha = 0.05$.

The Mean of a Normal Population

CONSIDER A SAMPLE FROM a normal population with an unknown mean μ and a known variance σ^2. Suppose that we are interested in testing the null hypothesis $H_0 : \mu = \mu_0$ against the alternative hypothesis $H_1 : \mu \neq \mu_0$, where μ_0 is a given constant. We would like to reject H_0 if the sample mean $\overline{\mathscr{X}}$ is far from μ; that is, we want to define a critical region of the form

$$C = \{\mathscr{X}_1, \ldots, \mathscr{X}_n \mid |\overline{\mathscr{X}} - \mu| > c\},$$

where c is some suitably chosen constant.

If we are interested in a given significance level α, we need to find the value of c that makes the probability of type I error to be α; that is, we want a c such that[163]

$$P[|\overline{\mathscr{X}} - \mu_0| > c \mid \mu = \mu_0] = \alpha.$$

If $\mu = \mu_0$, then $\overline{\mathscr{X}} \sim \mathcal{N}(\mu_0, \sigma^2/n)$; equivalently, $\frac{\sqrt{n}(\overline{\mathscr{X}} - \mu_0)}{\sigma^2}$ is a standard normal distribution. Thus, we are searching for a c such that

$$\alpha = P\left[|Z| > \frac{c\sqrt{n}}{\sigma}\right] = 2P\left[Z > \frac{c\sqrt{n}}{\sigma}\right],$$

or equivalently, $P\left[Z > \frac{c\sqrt{n}}{\sigma}\right] = \alpha/2$. This is solved by setting $\frac{c\sqrt{n}}{\sigma} = z_{\alpha/2}$; i.e., $c = \frac{z_{\alpha/2}\sigma}{\sqrt{n}}$. This means that the significance level α test is to reject H_0 iff $|\overline{\mathscr{X}} - \mu_0| > \frac{z_{\alpha/2}\sigma}{\sqrt{n}}$ (see Figure 21); or equivalently, if $\frac{\sqrt{n}}{\sigma}|\overline{\mathscr{X}} - \mu_0| > z_{\alpha/2}$.

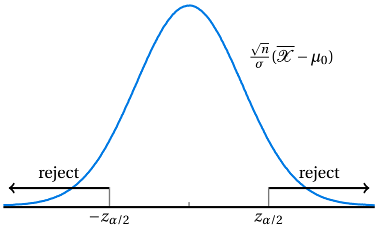


Figure 21: Critical region for a test.

Example 76. We want to test the hypothesis that a normal random variable $\mathscr{X}$ with variance 4 has mean 8.[164] From a sample of size 9, we obtain $\overline{\mathscr{X}} = 9.2$. Given a significance level $\alpha = 0.05$, we compute the test statistic

$$\frac{\sqrt{n}}{\sigma}|\overline{\mathscr{X}} - \mu_0| = \frac{3}{2}(1.2) = 1.8.$$

Since this number is not greater than $z_{0.025} = 1.96$, we cannot reject the hypothesis.[165] △

One important question is what is the *right* significance level to use. The specific choice depends on the application, but to be fair and correct it should be decided *before* the test is made, and not adapted to fit our desired conclusions. Importantly, a lower significance level makes it harder to reject the hypothesis, decreasing our chance of error in that case.

[161] That is, we bound the probability of error of *type I*—rejecting H_0 when it is correct—by α.

[162] That is, the point estimate for the mean.

[163] Assuming that $\mu = \mu_0$ (that is, H_0 is true), the probability of rejecting H_0 should be α.

[164] That is, $H_0 : \mu = 8$.

[165] Notice that if we had chosen a more liberal significance level (like 0.1) then we could have rejected H_0, but under the understanding that the probability of error of type I would increase (from 5 to 10% in this case).

The hypothesis test computes first the test statistic $v := \frac{\sqrt{n}}{\sigma}|\overline{\mathscr{X}} - \mu_0|$, and rejects the null hypothesis $H_0 : \mu = \mu_0$ if $\alpha \geq P[|Z| \geq v]$.[166] The probability $P[|Z| \geq v]$ is called the *p-value* of the test. It provides the critical significance level in the sense that H_0 will be rejected iff the significance level α is greater than or equal to the p-value.°

Example 77. In Example 76 we have computed the test statistic to be°

$$\frac{\sqrt{n}}{\sigma}|\overline{\mathscr{X}} - \mu_0| = 1.8.$$

Then, the p-value is $P[|Z| > 1.8] = 2P[Z > 1.8] = 2 \cdot 0.036 = 0.072$. This means that the null hypothesis $H_0 : \mu = 8$ will be rejected for any significance level $\alpha > 0.072$.[167]

Suppose that instead of 9.2 the sample yielded $\overline{\mathscr{X}} = 10.4$. Then, the test statistic is $\frac{\sqrt{n}}{\sigma}|\overline{\mathscr{X}} - \mu_0| = 3.6$, which gives the p-value $2P[Z > 3.6] = 0.0003$. The hypothesis would be rejected even for very low significance levels. △

ANOTHER TYPE OF ERROR is called *type II*: not rejecting a wrong hypothesis.[168] In the case of hypotheses for the mean μ of a normal, we define the function°

$$\begin{aligned}
\beta(\mu) &:= P[\text{not rejecting } H_0 \mid \mu] = P\left[\left|\frac{\overline{\mathscr{X}} - \mu_0}{\sigma/\sqrt{n}}\right| \leq z_{\alpha/2} \mid \mu\right] \\
&= P\left[-z_{\alpha/2} \leq \frac{\overline{\mathscr{X}} - \mu_0}{\sigma/\sqrt{n}} \leq z_{\alpha/2} \mid \mu\right]
\end{aligned}$$

The function $\beta(\mu)$, called the *operating characteristic (OC) curve*, describes the probability of *not* rejecting H_0 when the true mean is μ.

Recall that $\overline{\mathscr{X}}$ is a normal with mean μ and variance σ^2/n. In other words, $\frac{\overline{\mathscr{X}} - \mu}{\sigma/\sqrt{n}} \sim \mathscr{N}(0,1)$. This means that[169]°

$$\begin{aligned}
\beta(\mu) &= P_\mu\left[-z_{\alpha/2} \leq \frac{\overline{\mathscr{X}} - \mu_0}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right] \\
&= P_\mu\left[-z_{\alpha/2} - \frac{\mu}{\sigma/\sqrt{n}} \leq \frac{\overline{\mathscr{X}} - \mu_0 - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2} - \frac{\mu}{\sigma/\sqrt{n}}\right] \\
&= P_\mu\left[-z_{\alpha/2} - \frac{\mu}{\sigma/\sqrt{n}} \leq Z - \frac{\mu_0}{\sigma/\sqrt{n}} \leq z_{\alpha/2} - \frac{\mu}{\sigma/\sqrt{n}}\right] \\
&= P_\mu\left[\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} - z_{\alpha/2} \leq Z \leq \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_{\alpha/2}\right] \\
&= P_\mu\left[Z \leq \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_{\alpha/2}\right] - P_\mu\left[Z \leq \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} - z_{\alpha/2}\right] \\
&= \Phi\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_{\alpha/2}\right) - \Phi\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} - z_{\alpha/2}\right).
\end{aligned}$$

The precise values of this function obviously depend on the significance level α. For a fixed significance level, the OC curve is symmetric around μ_0 and depends on μ, and its relation to the hypothesis μ_0.° It reaches its maximum $1 - \alpha$ at $\mu = \mu_0$ and decreases as μ gets farther away from this point.

Example 78. Continuing Example 76, suppose that we want to compute the probability of not rejecting the null hypothesis $H_0 : \mu = 8$ when the true

[166] The probability that a standard normal distribution is farther away than the test statistic.

°Emphasise this, and the definition.

°$\overline{\mathscr{X}} \sim \mathscr{N}(\mu, 4)$, $\overline{\mathscr{X}} = 9.2$, $n = 9$, $\alpha = 0.05$.

[167] Recall from Example 76 that the test was not rejected with $\alpha = 0.05$, but would have been rejected with $\alpha = 0.1$.

[168] In our case, failing to reject it when it is wrong.

°μ is the true mean.

[169] We use P_μ to say that this probability depends on the actual value of μ.

°We are just trying to solve for Z.

°Add figure of the OC: looks a bit like a normal.

value of the mean is 10.[170] Then we compute $\frac{\sqrt{n}}{\sigma}(\mu_0 - \mu) = \frac{3}{2}(8-10) = -3$. Using the significance level $\alpha = 0.05$, since $z_{0.025} = 1.96$ the probability is

$$\beta(10) = \Phi(-3+1.96) - \Phi(-3-1.96) = \Phi(-1.04) - \Phi(-4.96) = 0.145. \qquad \triangle$$

The complement of the OC curve, $1-\beta(\mu)$ is called the *power function* of the test. It expresses the probability of rejecting the hypothesis when the true value is μ.

THE OC FUNCTION IS used to determine the size of a sample needed to guarantee properties of type II errors. Suppose that we want to guarantee that the probability of not rejecting $H_0 : \mu = \mu_0$ is approximately β. For each constant μ_1, we want to find an n such that $\beta(\mu_1) \approx \beta$.[171] That is,

$$\beta \approx \Phi\left(\frac{\mu_0-\mu_1}{\sigma/\sqrt{n}} + z_{\alpha/2}\right) - \Phi\left(\frac{\mu_0-\mu_1}{\sigma/\sqrt{n}} - z_{\alpha/2}\right).$$

Unfortunately, we cannot solve this problem analytically. To approximate n, we can use the following reasoning. Suppose (w.l.o.g.) that $\mu_1 > \mu_0$. Then $\mu_0 - \mu_1 < 0$ and hence

$$\frac{\sqrt{n}(\mu_0-\mu_1)}{\sigma} - z_{\alpha/2} < -z_{\alpha/2}.$$

Since the cumulative density function Φ is always increasing, we get

$$\Phi\left(\frac{\sqrt{n}(\mu_0-\mu_1)}{\sigma} - z_{\alpha/2}\right) \le \Phi(-z_{\alpha/2}) = P[Z \le -z_{\alpha/2}] = P[Z \ge z_{\alpha/2}] = \alpha/2.$$

Assuming that α is small enough—as usually done when dealing with confidence levels—it follows that $\Phi\left(\frac{\sqrt{n}(\mu_0-\mu_1)}{\sigma} - z_{\alpha/2}\right) \approx 0$, which means that

$$\beta \approx \Phi\left(\frac{\sqrt{n}(\mu_0-\mu_1)}{\sigma} + z_{\alpha/2}\right). \qquad (\dagger)$$

Notice that $\beta \in (0,1)$, which means that

$$\beta = P[Z > z_\beta] = P[Z < -z_\beta] = \Phi(-z_\beta).$$

Together with (†), it follows that we want some n such that

$$\Phi(-z_\beta) \approx \Phi\left(\frac{\sqrt{n}(\mu_0-\mu_1)}{\sigma} + z_{\alpha/2}\right); \quad \text{i.e.,}$$
$$-z_\beta \approx \frac{\sqrt{n}(\mu_0-\mu_1)}{\sigma} + z_{\alpha/2}.$$

Solving for n then yields

$$n \approx \frac{(z_\beta + z_{\alpha/2})^2\sigma^2}{(\mu_0-\mu_1)^2}$$

Example 79. Continuing Example 76, if we want the 0.05 level test for the hypothesis $H_0 : \mu = 8$ to have at least a 75% probability of rejection when $\mu = 9.2$, then we need a sample of size[172]

$$n \approx \frac{(z_{0.25}+z_{0.025})^2 \cdot 4}{(8-9.2)^2} = \frac{4(1.96+0.67)^2}{1.2^2} = 19.21.$$

Thus we need a sample of size 20. Let us confirm this using the OC function. If the true mean is 9.2, and $\mu_0 = 8$ then[173]

[170] Recall from Example 76 that the sample size is 9 and the variance is 4.

[171] Recall that the OC function depends on μ.

[172] Recall that β is the probability of not rejecting; in this case $\beta = 0.25$. Recall also that $z_{0.025} = 1.96$ and $z_{0.25} = 0.67$.

[173] Recall that $\beta(\mu)$ is

$$\Phi\left(\frac{\mu_0-\mu}{\sigma/\sqrt{n}} + z_{\alpha/2}\right) - \Phi\left(\frac{\mu_0-\mu}{\sigma/\sqrt{n}} - z_{\alpha/2}\right).$$

$$\begin{aligned}\beta(9.2) &= \Phi\left(\frac{\sqrt{20}(8-9.2)}{2}+1.96\right)-\Phi\left(\frac{\sqrt{20}(8-9.2)}{2}-1.96\right)\\ &= \Phi(-0.723)-\Phi(-4.643)\approx\Phi(-0.723)=1-\Phi(0.723)=0.235.\end{aligned}$$

Thus, in this case, if the true mean is 9.2, there will be a 76.5% chance of rejecting the null hypothesis $H_0 : \mu = 8$. △

THROUGHOUT THE PREVIOUS ANALYSIS, we considered a test that would reject whenever the observed sample mean is far away from the hypothesis mean, in any direction. Sometimes, a more meaningful hypothesis is to *bound* the mean e.g. from above. In this case, our null hypothesis will be of the form $H_0 : \mu \le \mu_0$ and should be rejected only if the estimate of μ is much greater than μ_0, but not if it is much smaller.[174] In other words, the critical region becomes

$$C = \{(\mathscr{X}_1,\dots,\mathscr{X}_n) \mid \overline{\mathscr{X}} - \mu_0 > c\},$$

for a suitably chosen c.[175] For the test to have a significance level α, we need that[176]

$$P[\overline{\mathscr{X}} - \mu_0 > c \mid \mu = \mu_0] = \alpha.$$

As usual, we use the fact that $\overline{\mathscr{X}} \sim \mathscr{N}(\mu, \sigma^2/n)$; thus if the true mean is μ_0 then $\frac{\sqrt{n}(\overline{\mathscr{X}}-\mu_0)}{\sigma} \sim \mathscr{N}(0,1)$. We want a c such that $P\left[Z > \frac{\sqrt{n}c}{\sigma}\right] = \alpha$; that is $\frac{\sqrt{n}c}{\sigma} = z_\alpha$, or[177]

$$c = \frac{z_\alpha \sigma}{\sqrt{n}}.$$

Hence, the test is to reject H_0 iff $\overline{\mathscr{X}} - \mu_0 > z_\alpha\sigma/\sqrt{n}$. This is the *one-sided critical region.*

To compute p-values in a one-side test like this one, we proceed as before: first use the data to compute the statistic $\sqrt{n}(\overline{\mathscr{X}} - \mu_0)/\sigma$. The p-value is the probability that a standard normal distribution is at least as large as this number.[178]

Example 80. Consider again Example 76,° where we want to test the hypothesis $H_0 : \mu \le 8$. The test statistic is $\sqrt{n}(\overline{\mathscr{X}} - \mu_0)/\sigma = 3(1.2)/2 = 1.8$. Then, the p-value is[179] $1 - \Phi(1.8) = 0.036$. This means that the test will reject H_0 for any significance level above 0.036. △

The OC° function of the one-sided test $\beta(\mu) = P[\text{not rejecting } H_0 \mid \mu]$ can be computed accordingly, dependent on the significance level α, by[180]

$$\begin{aligned}\beta(\mu) &= P[\overline{\mathscr{X}} \le \mu_0 + z_\alpha\sigma/\sqrt{n}] = P\left[\frac{\sqrt{n}(\overline{\mathscr{X}}-\mu)}{\sigma} \le \frac{\sqrt{n}(\mu_0-\mu)}{\sigma} + z_\alpha\right]\\ &= P[Z \le \frac{\sqrt{n}(\mu_0-\mu)}{\sigma} + z_\alpha] = \Phi\left(\frac{\sqrt{n}(\mu_0-\mu)}{\sigma} + z_\alpha\right)\end{aligned}$$

Notice that β decreases as μ increases. Recall that $\Phi(z_\alpha) = 1 - \alpha$. Thus, $\beta(\mu_0) = 1 - \alpha$.

In the construction, we used the condition $\mu = \mu_0$ to test the hypothesis $H_0 : \mu \le \mu_0$. We need to verify that this test does preserve the significance level α; i.e., that when H_0 is true, the probability of rejecting is bounded by α; equivalently, we may verify that $1 - \beta(\mu) \le \alpha$ holds for all $\mu \le \mu_0$;[181] or

[174] The latter case would fall inside our hypothesis.

[175] Remember that the critical region is the area of the input variables where the null hypothesis will be rejected. Compare this to the two-sided critical region from Page 60.

[176] The significance level is the probability of error of type I: rejecting the hypothesis when it is true. Notice that the conditioning is still w.r.t. $\mu = \mu_0$. This refers to the extreme case; if the actual mean is smaller than μ_0 it will be harder to reject the hypothesis.

[177] Compare again with two-sided confidence intervals.

[178] If $v := \sqrt{n}(\overline{\mathscr{X}} - \mu_0)/\sigma$ is the test statistic, then the p-value is $P[Z \ge v]$.

°Normal random variable with variance 4. With a sample of size 9, $\overline{\mathscr{X}} = 9.2$.

[179] The p-value in this case is $P[Z > 1.8]$; see note 178.

°Operating Characteristic

[180] Under the assumption that the true mean is μ, $\frac{\sqrt{n}(\overline{\mathscr{X}}-\mu)}{\sigma} \sim \mathscr{N}(0,1)$.

[181] Remember that the OC function $\beta(\mu)$ expresses the probability of not rejecting H_0 if the true mean is μ.

alternatively $\beta(\mu) \geq 1-\alpha$. By the previous discussion, we know already that for all $\mu \leq \mu_0$, $\beta(\mu) \geq \beta(\mu_0) = 1-\alpha$. Thus, this test does have a significance level of α for the hypothesis $H_0 : \mu \leq \mu_0$.

JUST AS WE HAVE done for $\mu \leq \mu_0$, it is also possible to introduce a one-sided test that bounds the mean by below. In this case, we would reject the null hypothesis $H_0 : \mu \geq \mu_0$ with significance level α iff $\frac{\sqrt{n}}{\sigma}(\overline{\mathscr{X}} - \mu_0) < -z_\alpha$, and the p-value is equal to the probability that Z gets a value under this test statistic.

Example 81. A bottled water company claims that its water contains in average less than 8mg of sodium per litre. A sample of 25 1-litre bottles yields an average of 7.5mg of sodium. What can we conclude, with a significance level of 5%, if we know that the standard deviation of sodium content is 1.8mg?

Solution. Remember that the only relevant thing that can be done through hypothesis testing is to *reject* the null hypothesis. Thus, in order to support the claim that the average is less than 8mg, we try to reject the opposite claim; that is, we intend to reject $H_0 : \mu \geq 8$. The test statistic is

$$\sqrt{n}(\overline{\mathscr{X}} - \mu_0)/\sigma = 5(7.5-8)/1.8 = -1.389,$$

and hence the p-value is $P[Z < -1.389] = 0.082$. Since this number is greater than 0.05, we cannot reject at 5% significance level the null hypothesis: the evidence is not strong enough to support the claim. △

SO FAR WE HAVE considered a known population variance σ^2. If the variance is unknown, but we still want to test the hypothesis $H_0 : \mu = \mu_0$ for some constant μ_0,[182] we want to reject H_0 is the sample mean is too far from μ_0, but we need an adequate notion of being far. In the previous case, we used the fact that $\overline{\mathscr{X}}$ is a normal to reject whenever $\left|\frac{\sqrt{n}(\overline{\mathscr{X}}-\mu_0)}{\sigma}\right| > z_{\alpha/2}$. As σ is not known anymore, we use its maximum likelihood estimator S.

To achieve a significance level α, recall that if $\mu = \mu_0$, the RV $\frac{\sqrt{n}(\overline{\mathscr{X}}-\mu_0)}{S}$ has a t-distribution with $n-1$ degrees of freedom. In particular,

$$P\left[-t_{\alpha/2,n-1} < \frac{\sqrt{n}(\overline{\mathscr{X}}-\mu_0)}{S} < t_{\alpha/2,n-1}\right] = 1-\alpha.$$

Thus, we reject H_0 (with significance level α) iff $\left|\frac{\sqrt{n}(\overline{\mathscr{X}}-\mu_0)}{S}\right| > t_{\alpha/2,n-1}$.

If t is the observed value of the test statistic $T = \frac{\sqrt{n}(\overline{\mathscr{X}}-\mu_0)}{S}$, then the p-value of this test is the probability that $|T|$ exceeds $|t|$.[183] This p-value tells us the significance levels at which the null hypothesis will be rejected.

Example 82. A worried neighbour claims that students drink an average of 3 litres of beer every night. To investigate this claim, 25 randomly selected students are observed. The observations yield a sample mean of 2.91l and a sample standard deviation of 0.47l. To verify the claim, we test the hypothesis $H_0 : \mu = 3$. The test statistic is

$$T = \frac{\sqrt{n}(\overline{\mathscr{X}}-\mu_0)}{S} = -\frac{5 \cdot 0.09}{0.47} = 0.9574.$$

[182] This is not a simple hypothesis anymore, because it covers a large space depending on the values of σ.

[183] The probability that the absolute value of a t random variable with $n-1$ degrees of freedom is larger than $|t|$.

The p-value for this test data is

$$P[|T_{24}| > 0.9574] = 2P[T_{24} > 0.9574] = 0.3479.$$

In other words, the only way to reject this hypothesis is to incur in a very high significance level (that is, increase the probability of error of type I). This experiment is consistent with the hypothesis. △

One-sided hypothesis tests are built in the obvious way. Hence, the hypothesis $H_0 : \mu \le \mu_0$ will be rejected iff $\frac{\sqrt{n}(\overline{\mathscr{X}}-\mu_0)}{S} > t_{\alpha,n-1}$.

Equality of Means of Normal Populations

OFTEN, WE WANT TO compare two approaches; for example, whether two software systems are equally efficient. This is usually verified by testing whether two normal populations have the same mean.

Let $\mathscr{X}_1,\dots,\mathscr{X}_n$ and $\mathscr{Y}_1,\dots,\mathscr{Y}_m$ be two independent samples from normal RVs having unknown means $\mu_{\mathscr{X}}, \mu_{\mathscr{Y}}$, and known variances $\sigma^2_{\mathscr{X}}, \sigma^2_{\mathscr{Y}}$. We want to test the hypothesis $H_0 : \mu_{\mathscr{X}} = \mu_{\mathscr{Y}}$. Obviously, $\overline{\mathscr{X}} - \overline{\mathscr{Y}}$ is an estimator for $\mu_{\mathscr{X}} - \mu_{\mathscr{Y}}$. Rewriting the null hypothesis as $H_0 : \mu_{\mathscr{X}} - \mu_{\mathscr{Y}} = 0$, we would like to reject H_0 if $\overline{\mathscr{X}} - \overline{\mathscr{Y}}$ is far from zero; i.e., if $|\overline{\mathscr{X}} - \overline{\mathscr{Y}}| > c$ from some suitable c.

Since $\mathscr{X}$ and $\mathscr{Y}$ are independent normal distributions, we know that $\overline{\mathscr{X}} - \overline{\mathscr{Y}} \sim \mathscr{N}\left(\mu_{\mathscr{X}} - \mu_{\mathscr{Y}}, \frac{\sigma^2_{\mathscr{X}}}{n} + \frac{\sigma^2_{\mathscr{Y}}}{m}\right)$;[184] or equivalently,

$$\frac{\overline{\mathscr{X}} - \overline{\mathscr{Y}} - (\mu_{\mathscr{X}} - \mu_{\mathscr{Y}})}{\sqrt{\frac{\sigma^2_{\mathscr{X}}}{n} + \frac{\sigma^2_{\mathscr{Y}}}{m}}} \sim \mathscr{N}(0,1).$$

If H_0 is true, then $\mu_{\mathscr{X}} - \mu_{\mathscr{Y}} = 0$ and so $(\overline{\mathscr{X}} - \overline{\mathscr{Y}}) \Big/ \sqrt{\frac{\sigma^2_{\mathscr{X}}}{n} + \frac{\sigma^2_{\mathscr{Y}}}{m}} \sim Z$. Then,

$$P\left[-z_{\alpha/2} \le \frac{\overline{\mathscr{X}} - \overline{\mathscr{Y}}}{\sqrt{\frac{\sigma^2_{\mathscr{X}}}{n} + \frac{\sigma^2_{\mathscr{Y}}}{m}}} \le z_{\alpha/2}\right] = 1 - \alpha.$$

In other words, for a significance level α, we reject $H_0 : \mu_{\mathscr{X}} = \mu_{\mathscr{Y}}$ iff[185]

$$\frac{|\overline{\mathscr{X}} - \overline{\mathscr{Y}}|}{\sqrt{\frac{\sigma^2_{\mathscr{X}}}{n} + \frac{\sigma^2_{\mathscr{Y}}}{m}}} \ge z_{\alpha/2}.$$

Following the same idea, the hypothesis $H_0 : \mu_{\mathscr{X}} \le \mu_{\mathscr{Y}}$ will be rejected with significance level α iff

$$\overline{\mathscr{X}} - \overline{\mathscr{Y}} \ge z_\alpha \sqrt{\frac{\sigma^2_{\mathscr{X}}}{n} + \frac{\sigma^2_{\mathscr{Y}}}{m}}.$$

SUPPOSE NOW THAT THE variances of the two normals $\mathscr{X}$ and $\mathscr{Y}$ are unknown but equal to a value σ^2.[186] To find an adequate critical region, we

[184] Remember that the sample of $\mathscr{X}$ has size n and that for $\mathscr{Y}$ has size m.

[185] The test statistic (which is useful for computing the p-value) is the left-hand side of this inequality.

[186] That is, $\sigma^2 = \sigma^2_{\mathscr{X}} = \sigma^2_{\mathscr{Y}}$; see note 151.

first estimate the value of σ^2. In this case, we can use the sample variances

$$S_{\mathscr{X}}^2 = \frac{\sum_{i=1}^{n}(\mathscr{X}_i - \overline{\mathscr{X}})^2}{n-1}$$
$$S_{\mathscr{Y}}^2 = \frac{\sum_{i=1}^{m}(\mathscr{Y}_i - \overline{\mathscr{Y}})^2}{m-1},$$

and the fact that[187]

$$\frac{\overline{\mathscr{X}} - \overline{\mathscr{Y}} - (\mu_{\mathscr{X}} - \mu_{\mathscr{Y}})}{\sqrt{S_p^2(1/n + 1/m)}} \sim t_{n+m-2},$$

where S_p^2 is the pooled estimator of σ^2.[188] Then, if the null hypothesis $H_0 : \mu_{\mathscr{X}} = \mu_{\mathscr{Y}}$ is true, the statistic $\frac{\overline{\mathscr{X}} - \overline{\mathscr{Y}}}{\sqrt{S_p^2(1/n+1/m)}}$ has a t-distribution with $n + m - 2$ degrees of freedom. Thus, we reject this hypothesis iff

$$\frac{|\overline{\mathscr{X}} - \overline{\mathscr{Y}}|}{\sqrt{S_p^2(1/n + 1/m)}} \geq t_{\alpha/2,n+m-2}.$$

The p-values and the one-sided hypothesis tests can be derived in an analogous manner.

[187] See page 57.

[188] $S_p^2 = \frac{(n-1)S_{\mathscr{X}}^2 + (m-1)S_{\mathscr{Y}}^2}{n+m-2}$.

Regression

IN HYPOTHESIS TESTING AND parameter estimation, we are concerned with understanding a single random variable.[189] In practical applications, we often want to understand the relationship between two or more variables. For instance, the relationship between age, type of work, and the number of accidents they have in a year.

In regression, we consider a single *response* variable (also known as *dependent* variable) $\mathscr{Y}$ that depends on a set of *input* (or *independent*) variables $\mathscr{X}_1, \dots, \mathscr{X}_n$.[190] Our goal is to understand this dependency; e.g., to predict, given the age and job description of a person, their likelihood of having an accident. The simplest such relationship is linear. That is, where $\mathscr{Y} = \beta_0 + \beta_1 \mathscr{X}_1 + \cdots + \beta_n \mathscr{X}_n$ for some $\beta_i \in \mathbb{R}, 0 \leq i \leq n$. If this relationship was precise, then from $n+1$ data points we could compute the exact values of the β_is, and from them predict the precise response given the values of the input variables. In practice, we expect the response to be affected by a random error. That is, we have

$$\mathscr{Y} = \beta_0 + \beta_1 \mathscr{X}_1 + \cdots + \beta_n \mathscr{X}_n + \varepsilon,$$

where ε is a RV with mean 0. That is, $E[\mathscr{Y} \mid \vec{\mathscr{X}}] = \beta_0 + \beta_1 \mathscr{X}_1 + \cdots + \beta_n \mathscr{X}_n$.[191]

This defines the *linear regression* of $\mathscr{Y}$ on the independent variables $\mathscr{X}_i$. In this case, the values $\beta_i, 0 \leq i \leq n$ are called the *regression coefficients*.[192] The question now is how to obtain these coefficients.

Least Squares Method

WE ARE INTERESTED IN estimating the parameters of a linear regression model. For a simple regression problem with estimated coefficients $\alpha = a$, $\beta = b$, and an observation x_i of the input variable, the response should be $y = a + bx_i$. Due to the random error, the response will in fact be a value y_i that may differ from this prediction. Given a set of n data points, we want to choose the estimator that minimises the sum of the squared errors. That is, we want to minimise $SS = \sum_{i=1}^{n} (y_i - a - bx_i)^2$.[193] To achieve this, we equate the partial derivatives of SS w.r.t. the variables a and b to 0.

$$0 = \frac{\partial SS}{\partial a} = -2 \sum_{i=1}^{n} (y_i - a - bx_i)$$
$$0 = \frac{\partial SS}{\partial b} = -2 \sum_{i=1}^{n} x_i (y_i - a - bx_i),$$

[189] The only exception we have seen is in testing the equality of means of two RVs. However, even in that case, we only check whether the means are equal or not, without trying to understand the relation between the two.

[190] Notice the abuse on the naming. Independence in this case refers to the response, and has nothing to do with probabilistic independence.

[191] $\vec{\mathscr{X}}$ denotes the whole set $\mathscr{X}_1, \dots, \mathscr{X}_n$.

[192] If there is only one independent variable $\mathscr{Y} = \alpha + \beta\mathscr{X} + \varepsilon$, we speak of a *simple regression*.

[193] The name SS comes from “Sum of Squares.”

or equivalently,

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i, \qquad \sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2.$$

These are known as the *normal equations.* Substituting $\overline{y} = \sum_{i=1}^{n} y_i / n$, and $\overline{x} = \sum_{i=1}^{n} x_i / n$, the first equation becomes $a = \overline{y} - b\overline{x}$. We can now substitute this value in the second equation to get

$$\sum_{i=1}^{n} x_i y_i = (\overline{y} - b\overline{x})\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \overline{y}n\overline{x} - b\overline{x}n\overline{x} + b\sum_{i=1}^{n} x_i^2,$$

and hence

$$b = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}.$$

The line $a + bx$, where a, b are computed from data using these equations is called the *estimated regression line.*

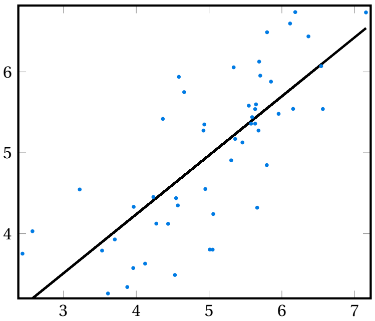


Figure 22: Some linearly related data, with its estimated regression line.

Estimator Distribution

SO FAR WE HAVE only assumed that the random error has mean 0. To study the properties of the estimators, we further assume that these errors are independent normal RVs with variance σ^2. This means that given the input values $\mathscr{X}_i$ the responses $\mathscr{Y}_i$ are independent normal RV with mean $\alpha + \beta\mathscr{X}_i$ and variance σ^2.[194]

The least squares estimator b of β can be expressed as

$$b = \frac{\sum_{i=1}^{n} (x_i - \overline{x}) y_i}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2},$$

which is a linear transformation of independent normal random variables $\mathscr{Y}_i$, and hence also a normal RV. We compute its mean and variance.[195]

$$\begin{aligned}
E[b] &= \frac{\sum_{i=1}^{n} (x_i - \overline{x}) E[y_i]}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(\alpha + \beta x_i)}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} \\
&= \frac{\alpha \sum_{i=1}^{n} (x_i - \overline{x}) + \beta \sum_{i=1}^{n} x_i (x_i - \overline{x})}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} \\
&= \beta \frac{\sum_{i=1}^{n} x_i^2 - \overline{x}\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} && (\dagger) \\
&= \beta \\
Var(b) &= \frac{Var\left(\sum_{i=1}^{n} (x_i - \overline{x}) y_i\right)}{\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right)^2} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2 Var(y_i)}{\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right)^2} \\
&= \frac{\sigma^2 \sum_{i=1}^{n} (x_i - \overline{x})^2}{\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right)^2} \\
&= \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} && (\ddagger)
\end{aligned}$$

In particular this means that b can serve as an estimator of β.[196]

[194] Notice that the variance σ^2 does not depend on the input variable, but is a constant associated to the error distribution. This variance is not necessarily known, and we must estimate it from the data.

[195] Recall that $\sum_{i=1}^{n} (x_i - \overline{x}) = 0$ (†). In addition, $\sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum_{i=1} x_i^2 - n\overline{x}^2$ (‡).

[196] An estimator whose expected value corresponds to true value being estimated is called *unbiased.*

Recall that $a = \sum_{i=1}^{n} y_i/n - b\overline{x}$. Then a is also a linear combination of independent normal random variables, and hence is also a normal. As we did with b, we analyse the parameters of this RV.

$$E[a] = \sum_{i=1}^{n} E[y_i]/n - \overline{x}E[b] = \sum_{i=1}^{n} (\alpha + \beta x_i)/n - \beta\overline{x} = \alpha + \beta\overline{x} - \beta\overline{x} = \alpha,$$
$$Var(a) = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right)}.$$

THE DIFFERENCES BETWEEN THE responses and their least square estimators $y_i - a - bx_i$, $1 \le i \le n$, are called *residuals*. To estimate the error variance σ^2 we use the sum of squares of the residuals

$$SS_R = \sum_{i=1}^{n} (y_i - a - bx_i)^2.$$

It can be shown that SS_R is independent from A and B and that SS_R/σ^2 is a chi-square with $n-2$ degrees of freedom.[197] Thus $E[SS_R/\sigma^2] = n-2$, or equivalently $E[SS_R/(n-2)] = \sigma^2$.

To summarise, if the responses $y_i, 1 \le i \le n$ are normally distributed with mean $\alpha + \beta x_i$ and (common) variance σ^2, then the least squares estimates for α and β are $a = \overline{y} - b\overline{x}$ and $b = \frac{\sum_{i=1}^{n}(x_i - \overline{x})y_i}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}$, respectively. These estimators are normally distributed as well:

$$a \sim \mathcal{N}\left(\alpha, \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right)}\right),$$
$$b \sim \mathcal{N}\left(\beta, \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}\right).$$

Moreover, the sum of squares of residuals is a chi-square with $n-2$ degrees of freedom, and is independent from a and b.

Statistical Inferences

WHEN CONSIDERING THE REGRESSION model $\mathscr{Y} = \alpha + \beta\mathscr{X} + \varepsilon$, an important question to ask is whether $\beta = 0$.[198] Thus, we want to test the null hypothesis $H_0 : \beta = 0$ vs. the alternative hypothesis $H_1 : \beta \neq 0$. Recall from our previous discussion that $(b-\beta)/\sqrt{\sigma^2/\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} \sim \mathcal{N}(0,1)$[199] and is indepedent from $SS_R/\sigma^2 \sim \chi^2_{n-2}$. Thus, we get that the variable

$$\frac{\sqrt{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}(b-\beta)/\sigma}{\sqrt{\frac{SS_R}{\sigma^2(n-2)}}} = \sqrt{\frac{(n-2)\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}{SS_R}}(b-\beta)$$

has a t-distribution with $n-2$ degrees of freedom.°[200]

If the null hypothesis is true (i.e., $\beta = 0$) then

$$\sqrt{\frac{(n-2)\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}{SS_R}}b \sim t_{n-2}.$$

Thus, we reject H_0 with significance γ[201] if $\sqrt{\frac{(n-2)S_{xx}}{SS_R}}|b| > t_{\gamma/2,n-2}$. The

[197] The actual proof of these facts is outside the scope of this course.

[198] If this is the case, then the response does not depend on the input variable.

[199] $b \sim \mathcal{N}\left(\beta, \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}\right)$.

°Define at this point $S_{xx} = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2$.

[200] Notice that we could not use the normal distribution observed before because it depends on the unknown σ^2. Instead, we estimate this variance through $SS_R/(n-2)$, but this requires the use of a t-distribution.

[201] We use here γ to avoid confusions with the regression parameter α.

p-value can be computed as usual: from the test statistic $v = \sqrt{\frac{(n-2)S_{xx}}{SS_R}}|b|$, the p-value is $P[|T_{n-2}| > v] = 2P[T_{n-2} > v]$.

To get a confidence interval estimator for β we proceed as usual. Given $0 < \gamma < 1$, we know that

$$P\left[-t_{\gamma/2,n-2} \leq \sqrt{\frac{(n-2)S_{xx}}{SS_R}}(b-\beta) \leq t_{\gamma/2,n-2}\right] = 1-\gamma.$$

Solving for β yields the $100(1-\gamma)\%$ confidence interval estimator of β

$$\left(B - \sqrt{\frac{SS_R}{(n-2)S_{xx}}}t_{\gamma/2,n-2}, B + \sqrt{\frac{SS_R}{(n-2)S_{xx}}}t_{\gamma/2,n-2}\right).$$

If one is instead interested in inferences concerning the parameter α, one can proceed in the exact same manner. Indeed, we observe that

$$\sqrt{\frac{n(n-2)S_{xx}}{SS_R \sum_{i=1}^n x_i^2}}(a-\alpha) \sim t_{n-2}.$$

From this fact, we can derive confidence intervals, and hypothesis tests, alongside their p-values.

Suppose now that we have a given input value x_0 and we want to produce a confidence interval or test a hypothesis about the mean response w.r.t. this value. Since we do not know the parameters α, β, we can estimate them through a and b as before. But to obtain a meaningful confidence interval, we must understand the distribution of $a + bx_0$.

Note that $a + bx_0 = \overline{y} - b(\overline{x} - x_0) = \sum_{i=1}^n y_i \left(1/n - c(x_i - \overline{x})(\overline{x} - x_0)\right)$, where c is a constant.[202]° Each y_i is an independent observation of a normal random variable, and all the other parameters in the expression are constants given the data. Hence, $a + bx_0$ is a linear combination of independent normal random variables, which means that it is also a normal itself. Now we just need to know its mean and its variance.

$$\begin{aligned}
E[a+bx_0] &= E[a] + x_0E[b] = \alpha + \beta x_0 \\
Var(a+bx_0) &= \sum_{i=1}^n \left(1/n - c(x_i - \overline{x})(\overline{x} - x_0)\right)^2 Var(y_i) \\
&= \sigma^2 \sum_{i=1}^n \left(1/n^2 - 2c(x_i - \overline{x})(\overline{x} - x_0)/n + c^2(x_i - \overline{x})^2(\overline{x} - x_0)^2\right) \\
&= \sigma^2 \left(1/n - 2c(\overline{x} - x_0)/n \sum_{i=1}^n (x_i - \overline{x}) + c^2(\overline{x} - x_0)^2 \sum_{i=1}^n (x_i - \overline{x})^2\right).
\end{aligned}$$

Since $\sum_{i=1}^n (x_i - \overline{x}) = 0$ and $\sum_{i=1}^n (x_i - \overline{x})^2 = \sum_{i=1}^n x_i^2 - n\overline{x}^2 = 1/c$, it then follows that

$$Var(a+bx_0) = \sigma^2\left(1/n + c(\overline{x} - x_0)^2\right).$$

In brief, $a + bx_0 \sim \mathcal{N}\left(\alpha + \beta x_0, \sigma^2\left(1/n + (\overline{x} - x_0)^2/S_{xx}\right)\right)$. To handle the unknown σ^2, recall that $SS_R/\sigma^2 \sim \chi^2_{n-2}$. Thus

$$\frac{a + bx_0 - (\alpha + \beta x_0)}{\sqrt{1/n + (x_0 - \overline{x})^2/S_{xx}}\sqrt{SS_R/n-2}} \sim t_{n-2}.$$

Using this fact, we can then build confidence intervals and hypothesis tests for $\alpha + \beta x_0$ as usual.

[202] Recall that

$$\begin{aligned}
a &= \overline{y} + b\overline{x} \\
b &= \frac{\sum_{i=1}^n (x_i - \overline{x})y_i}{\sum_{i=1}^n x_i^2 - n\overline{x}^2}.
\end{aligned}$$

The denominator of b is a constant that we will call c in the following analysis.

°Can use $c = 1/S_{xx}$ as the latter has been defined already.

Quality of the Model

When trying to measure the variability of the response variable, one can use the squared error $S_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2$. In this case, however, there are two different factors that influence the variability of the response. On the one hand, each value y_i depends on the input value x_i that may be different for each i. On the other hand, the response is a random variable with variance σ^2.[203]

An important question is how much of this variation is caused by the different input values, and how much depends on the inherent uncertainty of the response. To answer this question, observe that the value $SS_R = \sum_{i=1}^{n}(y_i - a - bx_i)^2$ can be interpreted as the remaining squared variation once that the influence of the input values has been taken into account. Thus $S_{yy} - SS_R$ measures the amount of variation that is explained by the input values. We can then define the *coefficient of determination* $R^2 := \frac{S_{yy} - SS_R}{S_{yy}}$, which expresses the proportion of variation explained by the input values.

Obviously, R^2 is always between 0 and 1. Intuitively, a high coefficient of determination (close to 1) indicates that most of the variation results from the different input values, and the opposite holds for a low R^2. This value is used as an indicator of the fitness of the model: the higher R^2 is, the regression model is considered to fit well the data.

Notice that linear regression can be applied to any piece of data, regardless of its shape. However, the quality of the model requires a linear dependency between the variables. Sometimes a visual exploration suffices to rule out the possibility of a linear model (see Figure 23). However, a direct visualisation cannot deal with more veiled cases of non-linearity.

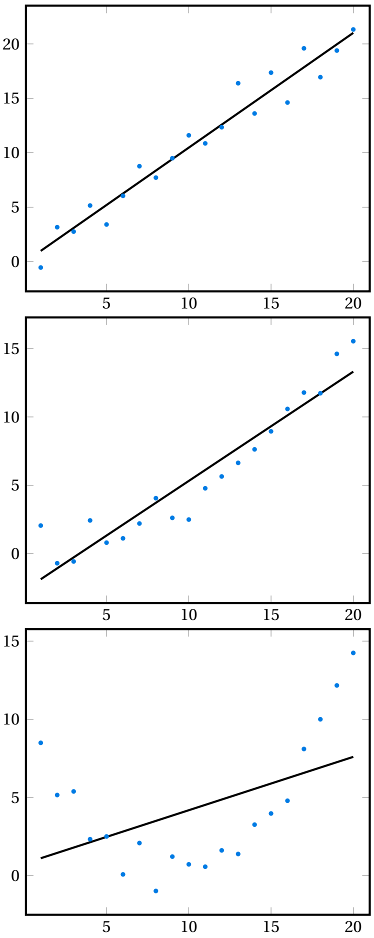


Figure 23: Three data samples with their linear regression fitting model. For the third plot, the linear model is obviously inadequate.

As a means to assess the quality of the model, one can analyse the residuals $y_i - (a + bx_i)$. Remember that each y_i is a normal; $a + bx_i$ estimates its mean, and $SS_R/(n-2)$ estimates its variance.[204] Then, the *standardised residuals* $\frac{y_i - (a+bx_i)}{\sqrt{SS_R/(n-2)}}$ should all be approximately distributed as a standard normal RV, and they are all independent. Then, approximately 95% of these residuals should lie in $(-2, 2)$,[205] and they should show a *random* behaviour without any obvious patterns. Figure 24 plots the standardised residuals from the first two data sets of Figure 23. The second plot shows a pattern that suggests that the linear model is not adequate for this data.

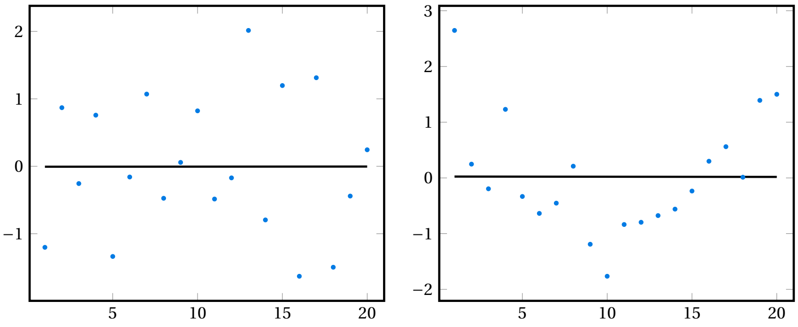


Figure 24: Residuals for the first two plots in Figure 23. The first one appears random, and hence is a good fit for the linear model. The second shows a pattern that suggests dependencies between the residuals.

[203] Even with identical values on the input variable, the responses may be different.

[204] Recall the assumption that the variance of the error is constant for all input values.

[205] $P[-1.96 < Z < 1.96] = 0.95$.

Non-linear Responses

EVEN WHEN THE RESPONSE does not depend linearly on the input variable, if the type of dependency is known one can sometimes transform the variables to obtain a linear model. The usual scenario is when the response variable grows exponentially on the input, that is, when the relationship is of the form $w = c \cdot d^x$, where c and d are two constants that we want to estimate.[206] By taking the logarithm in both sides, we get $\log(w) = \log(c) + x\log(d)$. This now looks like the linear model $y = \alpha + \beta x$.

At this point, we can use a standard linear regression process (that is, least squares) to estimate the parameters α and β with a and b, respectively. Then, $w = \exp\{a + bx\}$.

FOR OUR REGRESSION MODEL, we have assumed that the variance of the the response is fixed for any input value. Often, a more realistic assumption is that these variances are dependent on the input variable, but known up to a proportionality constant. More formally, we know that for each piece of data, $Var(y_i) = {}^{\sigma^2}/_{w_i}$. As before, to obtain a linear model, we want to minimise the sum of squares, but in this case the weight of each error should be proportional to the variance.[207] That is, we choose the estimators a, b that minimise the sum

$$\sum_{i=1}^{n} \frac{\left(y_i - (a + bx_i)\right)^2}{Var(y_i)} = \frac{1}{\sigma^2} \sum_{i=1}^{n} w_i (y_i - a - bx_i)^2.$$

To find these values, we differentiate with respect to a and b and equate the derivatives to 0, as usual. Hence, we obtain◦

$$\sum_{i=1}^{n} w_i y_i = a \sum_{i=1}^{n} w_i + b \sum_{i=1}^{n} w_i x_i,$$
$$\sum_{i=1}^{n} w_i x_i y_i = a \sum_{i=1}^{n} w_i x_i + b \sum_{i=1}^{n} w_i x_i^2.$$

Solving these equations, we obtain the *weighted* least square estimators.

In general, it is not always obvious to know how does the variance depend on the input variable. Still, there are many cases where this can be reasonably expected. A simple example is if one is trying to understand the relationship of water used in a house, dependent on the number of people living in it. Assuming that the amount of water used by each person is independent with a fixed variance, then the variance of use in a household will be proportional to the number of people in the house. Sometimes, the variance dependency can be understood by an analysis of the scatter plot of the data and its standardised residuals, as depicted in Figure 25.

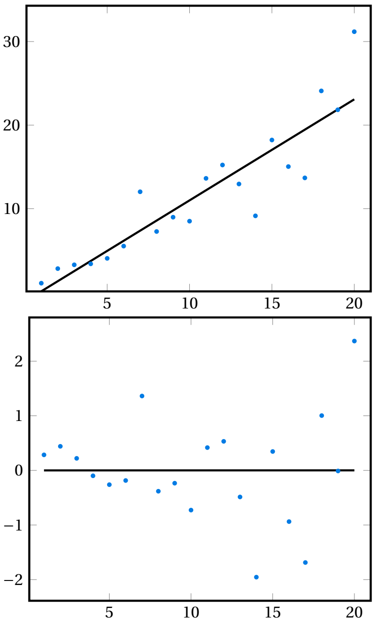


Figure 25: The variance of the data (observed by the margin of error) seems to increase with x.

AS WE HAVE ALREADY seen, the relationship between the independent and the response variables may not always be linear. Still, the dependency may follow a polynomial of the form $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_r x^r + \varepsilon$. In this case, we want to estimate the regression coefficients β_i. Generalising the idea of linear regression, given a data sample of size n, we compute the least square estimators $b_0, \ldots, b_r$ of $\beta_0, \ldots, \beta_r$, which are the values that

[206] Of course, one of them might be known, but the more general setting is where they are both unknown.

[207] It should be clear that for the data points where the variance of y is larger, we should expect to observe larger errors. To handle all the data uniformly, we weight each point proportionally to their variance.

◦

$$\frac{\partial}{\partial a} = \frac{-2}{\sigma^2} \sum_{i=1}^{n} w_i (y_i - a - bx_i),$$
$$\frac{\partial}{\partial b} = \frac{-2}{\sigma^2} \sum_{i=1}^{n} w_i (y_i - a - bx_i) x_i.$$

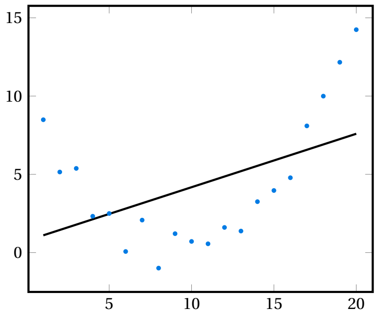
minimise

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x - \cdots - \beta_r x^r)^2.$$

Following the usual approach,[208] we obtain the so-called *normal equations*

$$\begin{aligned}
\sum_{i=1}^{n} y_i &= b_0 n + b_1 \sum_{i=1}^{n} x_i + b_2 \sum_{i=1}^{n} x_i^2 + \ldots + b_r \sum_{i=1}^{n} x_i^r \\
\sum_{i=1}^{n} x_i y_i &= b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 + b_2 \sum_{i=1}^{n} x_i^3 + \ldots + b_r \sum_{i=1}^{n} x_i^{r+1} \\
\sum_{i=1}^{n} x_i^2 y_i &= b_0 \sum_{i=1}^{n} x_i^2 + b_1 \sum_{i=1}^{n} x_i^3 + b_2 \sum_{i=1}^{n} x_i^4 + \ldots + b_r \sum_{i=1}^{n} x_i^{r+2} \\
&\vdots \qquad\qquad \vdots \\
\sum_{i=1}^{n} x_i^r y_i &= b_0 \sum_{i=1}^{n} x_i^r + b_1 \sum_{i=1}^{n} x_i^{r+1} + b_2 \sum_{i=1}^{n} x_i^{r+2} + \ldots + b_r \sum_{i=1}^{n} x_i^{2r}
\end{aligned}$$

The most important question to answer in polynomial regression is what degree of a polynomial to use. This requires a careful analysis regarding the trade-off between fitting and prediction. On the one hand, a polynomial with a higher degree will be able to fit the data better, producing smaller residuals.[209] However, recall that the values of y are subject to a random noise. A better fitness in this case will mean fitting the noise as well, which may reduce the predictive power of the model.[210] As a rule of thumb, you want to have the smallest degree that matches the general shape observed in the scatter plot of the data.



To understand this, consider again the third plot from Figure 23. A visual inspection suggests that the data follows a quadratic pattern. Thus, we may try to fit it through a quadratic regression method. The resulting curve, on the left of Figure 26, provides a good approximation to the observed data. A polynomial of degree 15 (Figure 26, right), fits the observed data much better, but makes some unreasonable predictions around the limit values.

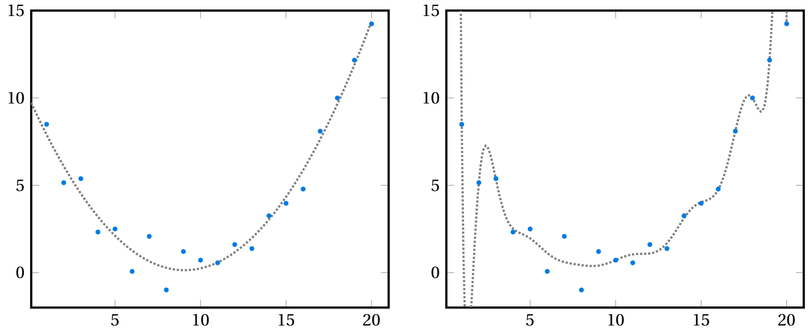


Figure 26: Data from Figure 23 fitted through polynomial regression of degree 2 (left) and 15 (right). The fitting curve is represented by a dashed line. Notice that the polynomial of degree 15 approximates the data better, but predicts the response for values of the independent variable around 1.5 and around 19.5 to be far away from all observations. This is a clear case of over-fitting.

[208] Equating the partial derivatives to 0.

[209] In principle, a polynomial of degree n can perfectly fit a data sample of size n if all the values of the independent variable are different.

[210] This is known as *overfitting*.

Non-Parametric Hypothesis Testing

Non-parametric hypothesis testing refers to cases where no assumption is made about the shape of the underlying probability distribution. When some information is known, one should try to use a parametric test as in the previous chapters, but in some scenarios the only feasible solution is a non-parametric test.

Sign Test

Suppose that we have a sample $\mathscr{X}_1, \ldots, \mathscr{X}_n$ from a continuous distribution F, and we are interested in testing a hypothesis about the *median* of F.[211] If m denotes the median of F, we want to test the hypothesis $H_0 : m = m_0$.

Consider the fact that F is a distribution function. This means that for every $i, 1 \leq i \leq n$, $P[\mathscr{X}_i \leq m_0] = F(m_0)$. Taking this into account, we can define the Bernoulli variables

$$I_i = \begin{cases} 1 & \mathscr{X}_i \leq m_0 \\ 0 & \text{otherwise.} \end{cases}$$

These variables are all independent, with parameter $p = F(m_0)$. Overall, our null hypothesis is equivalent to stating that $p = 0.5$. Let $v = \sum_{i=1}^n I_i$, and $\mathscr{Y}$ be a binomial random variable with parameters $(n, 0.5)$. Then the p-value for the test that $p = 0.5$ is

$$2\min\{P[\mathscr{Y} \leq v], P[\mathscr{Y} \geq v]\}.$$

In general,

$$P[Bin(n,p) \geq v] = P[n - Bin(n,p) \leq n - v] = P[Bin(n, 1-p) \leq n - v],$$

which means that the p-value is

$$\begin{cases} 2P[\mathscr{Y} \leq v] & v \leq \frac{n}{2} \\ 2P[\mathscr{Y} \leq n - v] & v \geq \frac{n}{2}. \end{cases}$$

What this test is doing is counting the number of successes or, in other words, the signs of the values $\mathscr{X}_i - m_0$. Hence, it is called the *sign test.*

Example 83. A sample of size 100 has 62 values smaller than a given m_0 and 38 that are greater. The p-value for the test that the median is m_0 is

[211] The *median* of F is the point x such that $F(x) = 0.5$.

$2P[Bin(100,0.5) \leq 38]$. By the central limit theorem, we know that

$$P[Bin(100,0.5) \leq 38] \approx P[Z \leq \frac{38.5-50}{\sqrt{25}}] = P[Z \leq -2.3] = 0.0107,$$

and hence the p-value is 0.0214. Thus, the hypothesis $H_0 : m = m_0$ will be rejected for a significance level of 0.05 but not for one of 0.01. △

One use of the sign test is to measure the effectiveness of a treatment. Suppose that we have a series of n patients with an ailment, and we apply a new treatment to them. Some of them will get better, while others may get worse. If the treatment has no effect, we will expect the median of the differences in the results to be 0. If k of them got better, then we would reject the hypothesis that the median change is 0, with significance level α, iff[212]

$$\sum_{i=0}^{k} \binom{n}{i} (0.5)^n \leq \frac{\alpha}{2}.$$

Of course, it is also possible to do one-sided tests for a population median. Suppose that we are interested in testing $H_0 : m \leq m_0$. If p denotes the probability that a population value is less than m_0, then H_0 is correct if and only if $p \geq 0.5$. To use the sign test, we obtain a sample of size n. If v of them have value less than m_0, then the p-value will be the probability of observing this value v (or something smaller) by chance if each element had probability 0.5 of being less than m_0; i.e. $P[Bin(n,0.5) \leq v]$. Dually, if we are interested in testing that the median is at least m_0, then the p-value is $P[Bin(n,0.5) \geq v]$.

Signed Rank Test

THE SIGNED RANK TEST is used to verify that the distribution of a population is symmetric around the value m_0; that is, it verifies the hypothesis $H_0 : P[\mathscr{X} < m_0 - a] = P[\mathscr{X} > m_0 + a]$ for all $a > 0$.

Suppose that $\mathscr{X}_i, \ldots, \mathscr{X}_n$ is a sample from a random variable $\mathscr{X}$ with unknown distribution, and let $\mathscr{Y}_i = \mathscr{X}_i - m_0, 1 \leq i \leq n$. Once we have observed these variables, we can rank them according to their absolute values. We can now define, for every $j, 1 \leq j \leq n$

$$I_j := \begin{cases} 1 & \text{if the } j\text{-th smallest value is from an observation } \mathscr{Y} < 0, \\ 0 & \text{otherwise.} \end{cases}$$

We then build the test statistic $T := \sum_{j=1}^{n} j I_j$.[213] This statistic counts, like the signed test, the number of values that are smaller than the hypothesised median m_0, but now weights them according to their relative distance to this value.[214]

Example 84. Consider a sample of size 4 with observed values $\mathscr{X}_1 = 4.1$, $\mathscr{X}_2 = 1.5$, $\mathscr{X}_3 = 3.1$, $\mathscr{X}_4 = 1$, and suppose that we want to test the hypothesis $H_0 : m = 2$.[215] The rankings of $|\mathscr{X}_i - 2|$ are $0.5, 1, 1.1, 2.1$, which come from $\mathscr{X}_2, \mathscr{X}_4, \mathscr{X}_3$, and $\mathscr{X}_1$, respectively. The first two come from values ≤ 2, and the last two from values ≥ 2. Hence $I_1 = I_2 = 1$, and $I_3 = I_4 = 0$. The test statistic is then $T = 1 + 2 = 3$. △

[212] Note that the sign test does not take into account the *strength* of the effect. If half of the results are extremely positive, and the other half only slightly negative, the null hypothesis will not be rejected. On the contrary, if all the effects are negative, but only by a tiny amount, the hypothesis will be rejected.

[213] The idea is that the ranks should be symmetric; approximately 0s and 1s should be interleaving.

[214] That is, elements that are farther away from m_0 get a larger weight.

[215] That is, $m_0 = 2$.

If the null hypothesis H_0 that the distribution is symmetric around m_0 is true, the mean and variance of the test statistic can be easily computed. Since the distribution of $\mathscr{Y} = \mathscr{X} - m_0$ is symmetric around 0, for any possible value of $|\mathscr{Y}|$ it is equally likely that $\mathscr{Y} > 0$ or $\mathscr{Y} < 0$. Thus, for each $j, 1 \leq j \leq n$ it follows that $P[I_j = 1] = 1/2 = P[I_j = 0]$. This means that[216]

$$E[T] = E\left[\sum_{j=1}^{n} jI_j\right] = \sum_{j=1}^{n} j/2 = \frac{n(n+1)}{4}$$
$$Var(T) = Var\left(\sum_{j=1}^{n} jI_j\right) = \sum_{j=1}^{n} j^2 Var(I_j) = \sum_{j=1}^{n} j^2/4 = \frac{n(n+1)(2n+1)}{24}.$$

Clearly, one can use the Central Limit Theorem (if the sample size is large enough) to approximate the distribution of T through a normal. Here we show how to make a precise computation of the p-value.

To obtain a test with significance level α for H_0, we should reject H_0 if the observed value t of T is such that $P[T \leq t] \leq \alpha/2$ or $P[T \geq t] \leq \alpha/2$. This means that the p-value for the test when the observed value is t is

$$2\min\{P[T \leq t], P[T \geq t]\}$$

To be able to compute this value, we take advantage of the equivalence

$$P[T \geq t] = P\left[T \leq \frac{n(n+1)}{2} - t\right],$$

which means that the p-value is

$$2\min\left\{P[T \leq t], P\left[T \leq \frac{n(n+1)}{2} - t\right]\right\} = 2P[T \leq t^*],$$

where $t^* := \min\{t, \frac{n(n+1)}{2} - t\}$.

To compute $P[T \leq t^*]$, let $P_k(i)$ denote the probability that the signed rank test statistic for a sample of size k is less than or equal to i, under the assumption that H_0 is true. We will compute $P_k(i)$ recursively starting with $k = 1$. If $k = 1$, we have a sample of size 1; that is, only one observation. If H_0 is true, then this observation is equally likely to be less or greater than m_0. So,

$$P_1(i) = \begin{cases} 0 & i < 0 \\ 1/2 & i = 1 \\ 1 & i \geq 1 \end{cases}$$

To compute $P_k(i)$ we condition over the value of the variable I_k:

$$\begin{aligned} P_k(i) &= P\left[\sum_{i=1}^{k} jI_j \leq i\right] \\ &= P\left[\sum_{i=1}^{k} jI_j \leq i \mid I_k = 1\right] P[I_k = 1] + P\left[\sum_{i=1}^{k} jI_j \leq i \mid I_k = 0\right] P[I_k = 0] \\ &= P\left[\sum_{i=1}^{k-1} jI_j \leq i - k \mid I_k = 1\right] P[I_k = 1] + P\left[\sum_{i=1}^{k-1} jI_j \leq i \mid I_k = 0\right] P[I_k = 0] \\ &= P\left[\sum_{i=1}^{k-1} jI_j \leq i - k\right] P[I_k = 1] + P\left[\sum_{i=1}^{k-1} jI_j \leq i\right] P[I_k = 0]. \end{aligned}$$

[216] Recall that the variance of a Bernoulli is $p(1-p)$, and the sum of the first n squares is $\frac{n(n+1)(2n+1)}{6}$.

As before, if H_0 holds, then $P[I_k = 0] = P[I_k = 1] = 1/2$. Moreover, $\sum_{i=1}^{k-1} jI_j$ is the signed rank statistic for a sample of size $k-1$. Hence,

$$P_k(i) = \frac{P_{k-1}(i-k)}{2} + \frac{P_{k-1}(i)}{2}.$$

This recursive construction can be used to compute the values of $P_n(t^*)$ in general.

The Two-sample Problem

SUPPOSE NOW THAT WE want to compare two methods, which cannot be applied to the same elements. Then we have two samples $\mathscr{X}_1, \ldots, \mathscr{X}_n$ and $\mathscr{Y}_1, \ldots, \mathscr{Y}_m$ of the results produced by method 1 and 2, which have continuous distribution functions F and G, respectively. Then, we are interested in testing the hypothesis $H_0 : F = G$.

One idea for testing this hypothesis is the *rank sum test*.[217] This test first orders all the $n+m$ values in the two samples, and ranks them according to the order; that is, the smallest value hast rank 1, the second smallest has rank 2, and so on, until the largest values that gets rank $n+m$. For each $i, 1 \le i \le n$, define the random variable R_i, which gets the rank of the value of $\mathscr{X}_i$. Then, we can define the test statistic

$$T := \sum_{i=1}^{n} R_i,$$

which sums the ranks of the values of the first sample.[218]

We should reject H_0 with significance level α if the observed value t of T is such that $P[T \le t] \le \alpha/2$ or $P[T \ge t] \le \alpha/2$.[219] For any integer t,

$$P[T \ge t] = 1 - P[T < t] = 1 - P[T \le t-1].$$

Hence, we should reject H_0 iff $P[T \le t] \le \alpha/2$ or $P[T \le t-1] \ge 1-\alpha/2$.

To compute these probabilities, we show how to recursively find the values $P(N, M, K)$ describing the probability $T \le K$ when the sample sizes are N and M, and the null hypothesis H_0 is true. To achieve this, we condition on whether the largest value from both samples was observed from the first, or from the second sample.

If it belongs to the first sample, then the value of T is $N+M$ plus the sum of the ranks of the other $N-1$ elements of the first sample. Under the assumption that H_0 is true, all the samples come from the same distribution, and hence $P(N, M, K) = P(N-1, M, K-N-M)$. Similarly, if the largest element belongs to the second sample, then $P(N, M, K) = P(N, M-1, K)$. Since the whole sample is independent, the largest value is equally likely to be any of the $N+M$ observed elements; that is, it will belong to the first sample with probability $N/(N+M)$. Overall, we get

$$P(N, M, K) = \frac{N \cdot P(N-1, M, K-N-M)}{N+M} + \frac{M \cdot P(N, M-1, K)}{N+M}.$$

To start the recursion, we notice that[220]

$$P(1,0,K) = \begin{cases} 0 & K \le 0 \\ 1 & K > 0, \end{cases} \qquad P(0,1,K) = \begin{cases} 0 & K < 0 \\ 1 & K \ge 0. \end{cases}$$

[217] Also known as *Mann-Whitney* or *Wilcoxon* test.

[218] Notice that the two samples may have different sizes. Still, the idea is that if the two distributions are the same, the ranks of $\mathscr{X}$ should be interleaved with those of $\mathscr{Y}$.

[219] We reject if the sum of the ranks is either too small or too large to be explained by chance.

[220] Notice that this test is agnostic to which is the first and which is the second sample. However, this recursion is more efficient if the first sample is that with the smallest rank sum.

Recall that the rank sum test rejects H_0 iff

$$2P(n,m,t) \le \alpha \quad \text{or} \quad 2(1-P(n,m,t-1)) \le \alpha.$$

Thus, the p-value of this statistic, when the observed value of T is t is $2\min\{P(n,m,t), 2(1-P(n,m,t-1))\}$.

SUPPOSE THAT, RATHER THAN two, we want to check that the distributions F_i of k populations are equal. That is, we have the null hypothesis $H_0 : F_1 = F_2 = \cdots = F_k$.

To test this hypothesis, we generate independent samples from each of the populations, having size $n_i, 1 \le i \le k$. Let $N = \sum_{i=1}^{k} n_i$ be the total number of data values obtained. As before, we rank these values from the smallest to the largest, and define R_i to be the sum of the ranks of the individuals from population i.

If H_0 is true, the rank of each value is equally likely to be any of number in $\{1,\dots,N\}$. In particular, the expected value of its rank is $\frac{\sum_{x=1}^{N} x}{N} = \frac{N+1}{2}$.[221] If we define $\overline{r} := \frac{N+1}{2}$, it follows that if H_0 is true, then the expected sum of the ranks of the population i is $n_i\overline{r}$;[222] i.e., $E[R_i] = n_i\overline{r}$.

If H_0 is true, the difference between $E[R_i]$ and $n_i\overline{r}$ should be small for each population $i, 1 \le i \le k$. To check the hypothesis, we measure the square proportional error via the test statistic

$$T = \sum_{i=1}^{k} \frac{(R_i - n_i\overline{r})^2}{n_i\overline{r}},$$

and reject the hypothesis is T is large. Since $\sum_{i=1}^{k} R_i$ is the sum of all the ranks of all the populations, it follows that $\sum_{i=1}^{k} R_i = \sum_{j=1}^{n} j = \frac{N(N+1)}{2} = N\overline{r}$. Hence,

$$\begin{aligned} T &= \frac{1}{\overline{r}} \sum_{i=1}^{k} \frac{R_i^2 - 2R_i n_i \overline{r} + n_i^2 \overline{r}^2}{n_i} \\ &= \frac{1}{\overline{r}} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 2\sum_{i=1}^{k} R_i + \overline{r}\sum_{i=1}^{k} n_i \\ &= \frac{1}{\overline{r}} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - N\overline{r}. \end{aligned}$$

Given the data, N and $\overline{r}$ are constant. Hence T is large iff $TS := \sum_{i=1}^{k} \frac{R_i^2}{n_i}$ is large. We use TS as a new test statistic. To find the significance level, it is important to know the distribution of TS when the null hypothesis is true. Although the distribution in itself is quite complex, it is possible to see that, under some minimal conditions,[223]

$$\frac{12}{N(N+1)} TS - 3(N+1) \sim \chi^2_{k-1}$$

Hence, the test for significance level α would be to reject the hypothesis H_0 iff $\frac{12}{N(N+1)} TS - 3(N+1) \ge \chi^2_{k-1,\alpha}$

Testing for Randomness

[221] Remember that $\sum_{x=1}^{n} x = n(n+1)/2$.

[222] The population has n_i elements, each of which has expected rank $\overline{r}$.

[223] Essentially, that each n_i is at least 5.

ONE BASIC ASSUMPTION IN statistics is that the input data is generated through a random sample from a population. We consider here a test that verifies whether this is indeed the case.

Consider first the case of a Bernoulli random variable; i.e., all the values are 0 or 1. Given a sample $\mathscr{X}_1, \dots, \mathscr{X}_N$, a sequence of consecutive equal values is called a *run*. For example, 111011 has 3 runs (first a run of 1s, then one of 0s and finally one of 1s).[224] Suppose that in the sample of size N, we observe 1 n times, and 0 m times ($N = n + m$), and let R be the number of runs. If the null hypothesis H_0 stating that this is a random sample is true, then the sequence observed would be equally likely to be any of the $\frac{N!}{n!m!} = \binom{n+m}{n}$ permutations of these values,[225] and hence the probability mass function of R is such that $P[R = k]$ is the proportion of permutations of n 1s and m 0s that have k runs.

It can be shown that, assuming H_0 is true,

$$\begin{aligned} P[R = 2k] &= 2\frac{\binom{m-1}{k-1}\binom{n-1}{k-1}}{\binom{n+m}{n}} \\ P[R = 2k+1] &= \frac{\binom{m-1}{k-1}\binom{n-1}{k} + \binom{m-1}{k}\binom{n-1}{k-1}}{\binom{n+m}{n}}. \end{aligned}$$

The hypothesis H_0 should be rejected if the number of runs observed is too large (the sequence is just changing between 0 and 1) or too small (there are long sequences of the same number). More precisely, the p-value for this *runs test*, when the observed number of runs is r is

$$2\min\{P[R \geq r], P[R \leq r]\}.$$

IF WE WANT TO test randomness of a sample from a variable $\mathscr{X}$ that is not a Bernoulli, we can simply check for runs of values above and below the sample median. More precisely, let m_s denote the sample median of a sample $\mathscr{X}_1, \dots, \mathscr{X}_N$, and let n, m be the number of values smaller or equal than m_s and greater than m_s, respectively.[226] We can then define new Bernoulli variables

$$I_j = \begin{cases} 1 & \mathscr{X}_j \leq m_s \\ 0 & \text{otherwise.} \end{cases}$$

If the original data is random, then the sample of the variables I_j will also be random. Hence, the runs test can be used to verify that the original data is random.

If n and m are both large and the null hypothesis is true, then R approximates a normal distribution with

$$\begin{aligned} \mu &= \frac{2nm}{n+m} + 1 \\ \sigma^2 &= \frac{2nm(2nm - n - m)}{(n+m)^2(n+m-1)}. \end{aligned}$$

Therefore,

$$\begin{aligned} P[R \leq r] &= P\left[\frac{R-\mu}{\sigma} \leq \frac{r-\mu}{\sigma}\right] \approx \Phi\left(\frac{r-\mu}{\sigma}\right) \\ P[R \geq r] &\approx 1 - \Phi\left(\frac{r-\mu}{\sigma}\right). \end{aligned}$$

[224] The number of runs in a sample is the number of times the value changes plus 1. It says how many times the results *switch*.

[225] If we make a random sample, the observations could have been obtained in any arbitrary order.

[226] That is, if N is even, then $n = m = N/2$; otherwise, $n = (N+1)/2$ and $m = (N-1)/2$.

This means that, in this case, the p-value for the runs test for randomness is approximately

$$2\min\left\{\Phi\left(\frac{r-\mu}{\sigma}\right), 1-\Phi\left(\frac{r-\mu}{\sigma}\right)\right\}.$$

Example 85. Suppose that a sequence of sixty 0s and sixty 1s results in 75 runs. Then, we can compute

$$\begin{aligned}\mu &= \frac{2nm}{n+m}+1 = \frac{2(60)(60)}{2(60)}+1 = 61,\\ \sigma^2 &= \frac{2nm(2nm-n-m)}{(n+m)^2(n+m-1)} = \frac{3540}{119},\\ \frac{r-\mu}{\sigma} &= \frac{75-61}{5.45} = 2.567.\end{aligned}$$

Then, the p-value is approximately

$$2\min\{\Phi(2.567), 1-\Phi(2.567)\} = 2(1-0.9949) = 0.0102.$$

For comparison, the actual p-value for this test is in fact 0.0130. The approximation is quite good. △