

A Short Introduction to Probability and Related Concepts

Harald Goldstein

Økonomisk institutt

August 2003

Contents

1	Introduction	1
2	Events and Probability	4
2.1	Mathematical description of events	4
2.2	Probability	6
2.3	Motivation for the axioms of probability	9
2.4	Exercises	10
3	More about probability	11
3.1	Some laws for probability	11
3.2	Conditional probability and independence	15
3.3	Motivation for conditional probability	19
3.4	Exercises	21
4	Stochastic (or Random) Variables, Expectation and Variance	22
4.1	Random variables and probability distribution	22
4.2	Expectation and variance	24
4.3	Jensen's inequality	30
4.4	Motivation of expectation	32
4.5	Exercises	33
5	Answers to the exercises	34

1 Introduction

We all have been exposed to probability and risk. We likely have a personal perception about what is probable and what is improbable. For example, we calculate the chance of being in an accident as small when we drive a car. Or, we believe that the probability of getting a job with a given level of education to be high or low. This can influence our decision to get an education or not. We also have an idea about events that are equally probable. We can, for example, believe that it is equally probable that our next child will be a boy or a girl. Before a football game, a coin is tossed to determine which end of the field from which the teams will start the game. This seems fair since the two possible outcomes are considered to have the same probability. A die is used in a game as an attempt to make it so that the six possible outcomes ("one", "two", "three", ..., "six"), with the highest degree of confidence, can be regarded as equally probable. A die that satisfies the requirement of equally probable outcomes, is often called *fair*.

What lies behind our intuition about probability is likely a perception about how frequent the actual event occurs. The more often the event occurs, the more probability one assigns to that event. This idea also lies behind classic probability calculations. Here, one goes a step further than merely dividing probability into rough categories like high, low, equally probable, etc. One imagines that probability can be measured on a scale between 0 and 1 where the end points 0 and 1 are included. An event that is certain to occur would have probability 1, while an impossible event, i.e. an event that never happens, would have probability 0.

Sometimes one prefers to express probability in percent, i.e. on a scale between 0 and 100. In order to get percent, one multiplies the probability by 100. For example, a probability of 25% equates to $1/4 = .25$ on the 0 - 1 scale.

Suppose that we toss two fair dice one time, and we are interested in the probability of getting at least one six (i.e. one or two sixes). Below (see example 2.2) we will show that it is reasonable to set the probability equal to $11/36 = 0.306$ (or 30.6% if one wishes), where 0.306 is rounded to 3 decimal places. What does this mean? The number $11/36$ is a measure of how often the event "at least one six" will occur when one repeats the experiment (tossing the two dice) many times. It does not mean that the event "at least two sixes" happens *exactly* 11 times out of 36 attempts, or *exactly* 110 times out of 360 attempts, but that we will get something in the neighborhood, on average, and in the long run "at least one six" in 11 out of 36 attempts. This will be further clarified below.

Let us imagine that we actually toss two dice many times. In the table below we have given a typical result that we would expect to see. The numbers were selected by a computer which simulates the results of 2000 tosses. Let A represent the event "at least one six". The table shows that A occurs 13 times in 36 attempts. We call the number of times A occurs the *absolute frequency* of A . The absolute frequency of A was therefore 13 out of 36 attempts, 33 out of 100 attempts, etc. When we divide the absolute frequency by the number of attempts, we get the *relative frequency* of A . Generally, if A occurs H times in n attempts, then H is the absolute frequency and $h = H/n$ is the relative frequency of A . The relative frequency of A in 36 attempts is thus $13/36 = .361$ (or 36.1% if one wishes) and, in 100 attempts, it is $33/100 = .330$ (or 33%), etc.

Number of attempts, n	36	100	400	1000	2000
Absolute frequency, H	13	33	127	291	616
Relative frequency, h	0.361	0.330	0.318	0.291	0.308

We see that the relative frequency seems to stabilize in the neighborhood of 0.306 as the number of attempts increases. This impression becomes more evident in figure 1, which shows the relative frequency plot against the number of attempts, n .

The tendency that the relative frequency seems to stabilize around a certain number as n increases has long been known as an empirical phenomenon in science and

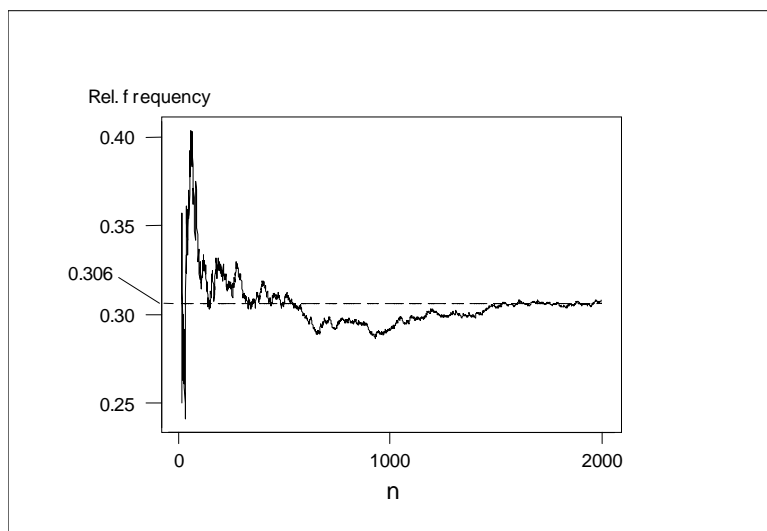


Figure 1: Plot of the relative frequency of "at least one six" when the number of tosses with two dice (attempts) increases from 10 to 2000 (simulated numbers).

gambling. One has therefore been tempted to assume that the relative frequencies, based on a series of equally likely events, actually converge towards certain numbers, called probabilities, when the number of events approaches infinity. If, for example, the relative frequency of $A =$ "at least one six" actually approaches 0.306 when the number of attempts, n , approaches infinity, 0.306 is interpreted as the probability that A will occur in an arbitrary attempt (tossing two dice).

This convergence is of course not possible to prove empirically (it is not possible to repeat an event infinitely many times), but it can be taken as a working hypothesis. Early attempts to build a theory of probability based on the idea of convergence of relative frequency proved, however, to lead to theoretical difficulties. Instead, it is common in modern probability theory to build the theory axiomatically, based on simple unproven fundamental assumptions, called axioms. In this theory, one does not claim to say what probability "really is" other than to postulate that it is a number between 0 and 1 that satisfies certain properties (axioms). It then turns out to be possible to prove mathematically, from the axioms, that under certain ideal conditions, the relative frequencies actually converge towards corresponding probabilities when the number of attempts approaches infinity. This is a fundamental result that is called the *law of large numbers*. We will not go through any proof of the law of large numbers in this introductory course but merely point at the importance of the results.

Probabilities are thus ideal quantities that can not be directly observed in reality. They can, however be estimated through different methods. For example:

- Sometimes they arise from the ideal assumptions which can be more or less realistic. For example, if it is reasonable to assume that a given die is fair, it

follows that the probability of getting one "one" in one toss must be $1/6$. (See example 2.1). The ideal assumption here is that the die is fair, i.e. that all six possible outcomes are equally probable.

- When one has access to a set of data, one can often resort to the law of large numbers and estimate a probability with the relative frequency. For example, a public opinion poll found that in a random sample of 1000 voters, 347 would vote for the Labor Party if there was an election soon thereafter. If the sample is picked purely at random from the population of Norwegian voters, then one can estimate the probability that an arbitrary voter will vote for the Labor Party, at that point in time, using the relative frequency, $h = 0.347$ (or 34.7%). Since the relative frequency is, in general, not equal to the probability, only approximately equal, an uncertainty problem arises. How uncertain is the estimate? In statistics, one has methods for estimating uncertainty, in this instance it could be expressed by a margin of error, calculated as ± 0.030 . This means, imprecisely, that one is quite certain that the probability lies between 0.317 and 0.377 which, in this case, turns out to be a quite large uncertainty. In accordance with the law of large numbers, the uncertainty becomes less with a larger sample. Larger samples, however, are too expensive and time consuming for the needs of public opinion poll authorities, and there have been methods developed that make it possible to reduce the uncertainty without increasing the sample size. We will not, however, go into uncertainty calculations in this course.
- Subjective estimation: "Having heard the forecast, I do not believe that the probability that it will rain tomorrow is more than 0.10, subjectively estimated according to my experience with the forecasts accuracy and the weather for this time of year".

2 Events and Probability

2.1 Mathematical description of events

A major principle in probability calculations is to find the probability of a complicated event by separating the event into smaller parts, which we know the probability of and subsequently add up. It turns out that the mathematical theory of sets is a good helping tool to precisely define and handle events.

We call the framework around the events we wish to analyze an *experiment*. An experiment could be that we toss a die and observe the result (the outcome), for example, a three. It can be a toss of two dice where we record the outcome of the two dice, for example, a two and a six, or more complicated, where we record the consumption of candy of each student in a class during one week.

The point of departure for the analysis is a list of all the possible *outcomes* of the experiment, which we can symbolize as: e_1, e_2, \dots, e_k . When we throw one die, the experiment is made up of $k = 6$ possible outcomes (if we disregard the possibility that it comes to rest on an edge, disappears, etc.), and we can set $e_1 = \text{"one"}$, $e_2 = \text{"two"}$, ..., $e_6 = \text{"six"}$. The sample space, S , is defined as a set of possible outcomes

$$S = \{e_1, e_2, \dots, e_k\}$$

where the brackets indicate that S should be interpreted as a set. Events can now be interpreted as a sub set of the sample space.

We will now look at the simple experiment of tossing one die. The sample space is defined as

$$S = \{e_1, e_2, \dots, e_6\} = \{\text{"one"}, \text{"two"}, \dots, \text{"six"}\}$$

An example of an event that may or may not occur, during this experiment, is $A = \text{"at least a four"}$. A can occur in three different ways, namely that the outcome of the toss is a four, a five or a six. We can therefore describe A as the set of outcomes in which A occurs

$$A = \text{"at least a four"} = \{e_4, e_5, e_6\}$$

and we see that A appears as a sub set of S . We say that A *occurs* if the outcome of the experiment (to toss the die one time) is among the elements in A .

Examples of other events within the scope of the experiment are

$$\begin{aligned} B &= \text{"An even number"} = \{e_2, e_4, e_6\} \\ C &= \text{"Not a one"} = \{e_2, e_3, e_4, e_5, e_6\} \\ D &= \text{"A two, three or five"} = \{e_2, e_3, e_5\} \\ E &= \text{"A six"} = \{e_6\} \end{aligned}$$

Assume that we actually toss a die and get a three as the outcome. Then, C and D have occurred, while A, B and E have not occurred.

Individual outcomes, e_1, e_2, \dots are also events, sometimes called elementary events. Since events are defined as sets, one, to be strict, should use brackets, $\{e_1\}, \{e_2\}, \dots$ for individual outcomes, but this gives rise to unnecessarily complicated notation. Therefore, it is common to ignore the parenthesis for individual outcomes when there is little danger for misunderstanding.

Two other extreme events are the *certain event*, namely the sample space itself, S , which, by definition, must always occur, and the *impossible event*, \emptyset , which never occurs. The symbol \emptyset is named according to set notation and means a set that is empty, that is to say does not contain any elements. In the example, the event

$$F = \text{“Something other than 1,2,3,4,5 or 6”} = \emptyset$$

is impossible since we have ruled out any other possible outcomes than 1, 2, ...6. If we in practice still got something else (the die on its edge, etc.) we consider the experiment invalid or not completed. What is regarded as possible or impossible is therefore, to a certain degree, a question of definition, which is a part of our precise definition of the experiment. If we had wished, there would be nothing stopping us from including $e_7 =$ "die on its edge" among the possible outcomes. The experiment would, in this case, have a sample space $S = \{e_1, e_2, \dots, e_6, e_7\}$, and the event F is no longer impossible.

2.2 Probability

The next step is to construct probabilities for the events within an experiment. One starts by stating probabilities for the elementary events in $S = \{e_1, e_2, \dots, e_k\}$, in other words, k numbers between 0 and 1 that we symbolize with $P(e_1), P(e_2), \dots, P(e_k)$, where $P(e_1)$ represents the probability that e_1 is the outcome of the experiment, $P(e_2)$ is the probability that e_2 occurs, etc. These so-called elementary probabilities are often stated by employing one of three methods outlined at the end of section 1.

For an arbitrary event, for example $A = \{e_4, e_5, e_6\}$, we find, in accordance with the axioms for probability, the probability for A , which we write as $P(A)$, by summing the elementary probabilities for all the outcomes that are included in A

$$P(A) = P(e_4) + P(e_5) + P(e_6)$$

Some of the most important axioms for probability follow below. As mentioned above, these cannot be proven directly but they can be motivated in various ways. In section 2.3, they are motivated by the law of large numbers. In short, the motivation is derived by showing that all the properties described by the axioms apply to the relative frequencies that we get by repeating the experiment many times. If it is true that the relative frequency approaches its respective probability as the number of attempts goes to infinity (the law of large numbers), the probability must necessarily inherit the same properties that apply to the relative frequencies.

Axioms:

Let the experiment have sample space, $S = \{e_1, e_2, \dots, e_k\}$.

For each event, A , the probability for A , $P(A)$ is ((2.2))
a number between 0 and 1 (0 and 1 included)

The elementary probabilities must sum to 1:

$$P(e_1) + P(e_2) + \dots + P(e_k) = 1 \quad ((2.2))$$

The certain event, S , and the impossible event, \emptyset , have probabilities:

$$\begin{aligned} P(S) &= 1 \\ P(\emptyset) &= 0 \end{aligned} \tag{2.3}$$

If A is an arbitrary event (contained in S), $P(A)$ is the summation of the elementary probabilities for all individual outcomes that are included in A :

$$P(A) = \sum_{\text{All } e \text{ included in } A} P(e) \tag{2.4}$$

Notice that axiom (2.1), strictly speaking is redundant since it follows from axioms (2.3) and (2.4) (Why?).

Example 2.1

What is $P(A)$, where A is the event "at least a four", when one fair die is tossed?

Solution: The sample space is as before, $S = \{e_1, e_2, \dots, e_6\}$, and $A = \{e_4, e_5, e_6\}$. Since the die are assumed to be fair, meaning that all individual outcomes are equally probable

$$P(e_1) = P(e_2) = \dots = P(e_6) = p$$

where we have named the *common* value p . From axiom (2.2) we get

$$1 = P(e_1) + P(e_2) + \dots + P(e_6) = p + p + \dots + p = 6p$$

p must also fulfill the equation, $6p = 1$, which gives $p = 1/6$. Therefore, all the elementary probabilities are equal to $1/6$ and we get from axiom (2.4)

$$P(A) = P(e_4) + P(e_5) + P(e_6) = 1/6 + 1/6 + 1/6 = 3/6$$

Consequently,

$$P(A) = 1/2$$

■

Example 2.1 is an example of a common situation in which the sample space is defined such that it is reasonable to assume that all of the elementary events, e_1, e_2, \dots, e_k , are equally probable. This is called a uniform probability model and is a situation where calculating probabilities is especially simple. By using the reasoning in example 2.1 (check!), we get the following rule:

Proposition 1 *Let A be an arbitrary event in an experiment where all of the individual outcomes, e_1, e_2, \dots, e_k , are equally probable (with probability $= 1/k$). Then,*

$$P(A) = \frac{r}{k} \quad ((2.5))$$

where k is the number of possible outcomes of the experiment, and r is the number of outcomes that are included in A .

Example 2.2

Let the experiment consist of a single toss of two fair dice. What is the probability of getting at least one six?

Solution: An outcome of this experiment consists of two numbers and can be described as (x, y) where x is the result of die 1 and y is the result of die 2. Since there are 36 combinations of x and y , the sample space consists of 36 possible outcomes

$$\begin{aligned} S &= \{e_1, e_2, \dots, e_{36}\} \\ &= \{(1, 1), (1, 2), (1, 3), \dots, (1, 6), \\ &\quad (2, 1), (2, 2), (2, 3), \dots, (2, 6), \\ &\quad (3, 1), (3, 2), (3, 3), \dots, (3, 6), \\ &\quad \vdots \\ &\quad (6, 1), (6, 2), (6, 3), \dots, (6, 6)\} \end{aligned}$$

Since the dice are fair, we can assume that the 36 outcomes are equally probable. The event $A =$ "at least one six" consists of 11 outcomes

$$A = \{(6, 1), (6, 2), (6, 3), (6, 4); (6, 5), (6, 6), (5, 6), (4, 6), (3, 6), (2, 6), (1, 6)\}$$

>From (2.5) it follows that we get

$$P(A) = \frac{11}{36} = 0.306 \quad (\text{with three decimal places})$$



Example 2.3

In a commercial building consisting of 7 floors, it is rare that one experiences taking the elevator from the first floor to the 7th without stopping at any floors in between. An employee has set up the following probability model

(subjective estimates) for the number of stops at intermediate floors in the middle of the day in one trip from the 1st to the 7th floor:

<i>Number of Stops</i>	0	1	2	3	4	5
<i>Probability</i>	0.01	0.05	0.25	0.34	0.25	0.10

The sample space here is $S = \{e_1, e_2, \dots, e_6\} = \{0, 1, 2, 3, 4, 5\}$ where $e_1 =$ “0 stops”, $e_2 =$ “1 stop”, etc. Since the elementary probabilities are not equal, we have an example of a non-uniform probability model. We find, for example, that the probability of at least three stops by:

$$P(\text{“At least 3 stops”}) = P(e_4) + P(e_5) + P(e_6) = 0.34 + 0.25 + 0.10 = 0.69$$

■

2.3 Motivation for the axioms of probability

As mentioned, we do not prove the axioms but justify them in other ways, for example by demonstrating phenomena that one can observe in reality. A highly used motivation for the probability axioms is to demonstrate the phenomenon that relative frequencies, from an experiment repeated many times, seem to stabilize as the number of attempts increases. We can look again at the example where we toss one die and the sample space is

$$S = \{e_1, e_2, \dots, e_6\} = \{1, 2, 3, 4, 5, 6\}$$

Assume that we toss the die 100 times (in general n times) and get

Outcome		100 attempts		n attempts	
		Abs. frequency	Rel. frequency	Abs. frequency	Rel. frequency
e_1	1	12	0.12	H_1	h_1
e_2	2	20	0.20	H_2	h_2
e_3	3	16	0.16	H_3	h_3
e_4	4	19	0.19	H_4	h_4
e_5	5	20	0.20	H_5	h_5
e_6	6	13	0.13	H_6	h_6
Sum		100	1	n	1

We will see that the relative frequencies fulfill the properties in (2.1)-(2.4). Consider for example the event $A =$ “at least a four” $= \{e_4, e_5, e_6\}$. The absolute frequency of A in 100 attempts is $H_A = 19 + 20 + 13 = 52$, and the relative frequency is $h_A = 52/100 = 0.52$. We see that we have got the same result as if we had added the relative frequencies of the outcomes in A : $0.52 = 0.19 + 0.20 + 0.13$.

For an arbitrary n we have

$$H_A = H_4 + H_5 + H_6$$

from which

$$h_A = \frac{H_A}{n} = \frac{H_4 + H_5 + H_6}{n} = \frac{H_4}{n} + \frac{H_5}{n} + \frac{H_6}{n}$$

and we get that

$$h_A = h_4 + h_5 + h_6 \quad ((2.6))$$

must be true for all n . If we now believe the law of large numbers to be true (see the introduction), the relative frequencies should converge towards the probabilities as n increases. In other words, h_A converges towards $P(A)$ and h_i converges towards $P(e_i)$ for $i = 1, 2, \dots, 6$. Since (2.6) applies to all n , it must also be the case for the probabilities that

$$P(A) = P(e_4) + P(e_5) + P(e_6)$$

Equivalent reasoning applies to other events, and we have motivated axiom (2.4). Because the relative frequencies sum to 1, it follows that:

$$\begin{aligned} h_1 + h_2 + \dots + h_6 &= \frac{H_1}{n} + \frac{H_2}{n} + \dots + \frac{H_6}{n} = \\ &= \frac{H_1 + H_2 + \dots + H_6}{n} = \frac{n}{n} = 1 \end{aligned}$$

Since this is true for all n , it must also be true for the probabilities, which is axiom (2.2).

Since a relative frequency must always lie between 0 and 1, the same must be true for a probability (axiom (2.1)).

The certain event, S , must always occur and therefore have the absolute frequency n in n trials and relative frequency $h_S = n/n = 1$. This means that h_S also converges towards 1 as n increases, in which case $P(S) = 1$. Conversely, we have that the impossible event, \emptyset , never occurs and therefore has the absolute frequency $H_\emptyset = 0$ and relative frequency $h_\emptyset = 0/n = 0$ for all n . This means that the extreme value, $P(\emptyset)$, must also be 0. Axiom (2.3) has thus been motivated.

2.4 Exercises

2.1 A new economic's graduate is applying for a job at three places. She classifies the job interview as a success (S) or a failure (F) depending on whether or not the interview leads to a job offer. An outcome of this experiment can be described as (x, y, z) where x, y, z is the result of the job interviews (S or F) at place 1, 2, and 3 respectively.

- a. Set up the sample space for the experiment.
- b. Express the events "Exactly 2 successes" and "At least 2 successes" as a set of individual outcomes.

2.2 Nilsen drives the same way to work every day and passes 4 stoplights. After much experience, he decided to set up the following probabilities for the number of red lights he comes to on his way

Number of red lights	0	1	2	3	4
Probability	0.05	0.25	0.36	0.26	0.08

- a. What is the sample space for the "experiment" which is to record the number of red lights Nilsen comes to on his way to work?
 - b. What is the probability of coming to at least 2 red lights?
 - c. What is the probability of coming to at most 2 red lights ?
 - d. Is this a uniform probability model?
- 2.3** Find $P(\text{"At least a four"})$ for a die that is "fixed" such that the probability for a six is twice as much as the probability of the remaining outcomes. (Hint: Set $P(\text{"1"}) = P(\text{"2"}) = \dots = P(\text{"5"}) = p$ and $P(\text{"6"}) = 2p$. Use (2.2) to find p).
- 2.4** Find the probability that the sum of two fair dice equals 7 when tossed one time.
- 2.5** There are 5 girls and 10 boys in a class.
- a. One student is selected completely at random (i.e. such that all have the same chance to be selected) to go on an airplane ride. What is the probability that a girl is selected?
 - b. (*More demanding*). A selection of two students is made completely at random for the airplane ride. What is the probability that the selection consists of one boy and one girl?

3 More about probability

3.1 Some laws for probability

Events can be joined together with new events by the use of "or", "and" and "not". Let A and B be events. We then create new events, C, D, E, \dots :

- *Union*. $C = \text{"A or B"}$, which is written, $A \cup B$, (and is sometimes read as "A union B"). $A \cup B$ occurs if A or B or both occur.

- Intersection. $D = \text{“}A \text{ and } B \text{”}$, which is written, $A \cap B$, (and is sometimes read as “the intersection of A and B ”). $A \cap B$ occurs if both A and B occur.¹
- Compliment. $E = \text{“not } A \text{”}$, which is written, A^c , (and is sometimes read as “the compliment of A ”). A^c occurs if A does *not* occur.²

If, for example, $A = \{e_1, e_2, e_3, e_4\}$ and $B = \{e_3, e_4, e_5, e_6\}$ in an experiment with 8 outcomes, $S = \{e_1, e_2, \dots, e_8\}$, then

$$\begin{aligned} C &= A \cup B = \{e_1, e_2, e_3, e_4, e_5, e_6\} \\ D &= A \cap B = \{e_3, e_4\} \\ E &= A^c = \{e_5, e_6, e_7, e_8\} \end{aligned}$$

Note that the expression "or" is ambiguous in daily speech. “ A or B ” can mean “either A or B ”, i.e. that only one of A and B occur. The expression can also mean “at least one of A and B occur”, and it is this last meaning that is meant by $A \cup B$. In the example, the event is

$$\text{“ Either } A \text{ or } B \text{”} = \{e_1, e_2, e_5, e_6\}$$

If two events, G and H , cannot occur simultaneously (i.e. $G \cap H = \emptyset$), they are said to be *mutually exclusive* or *disjoint*. In the example, D and E are mutually exclusive ($D \cap E = \emptyset$), same as the two events D and “Either A or B ”. D and A are not mutually exclusive since $D \cap A = \{e_3, e_4\}$. In an experiment, the individual outcomes, e_1, e_2, \dots , by definition are mutually exclusive. When the experiment is carried out, one and only one, “ e ” will occur. Two different “ e ”s therefore can not occur simultaneously.

Other examples: For an arbitrary event, A , we have

$$A \cup A^c = S \tag{(3.7)}$$

$$A \cap A^c = \emptyset \tag{(3.8)}$$

A and A^c are consequently always mutually exclusive, while “ A or A^c ” will always occur. We say that A and A^c are complimentary events. We also have

$$S^c = \emptyset$$

$$\emptyset^c = S$$

Example 3.1

Let the experiment consist of recording children’s gender for an arbitrary three-child family. The outcome can be described as xyz where x, y, z are the genders (G or B) for

¹Some text books write AB rather than $A \cap B$.

²Some textbooks use the notation \bar{A} rather than A^c .

the oldest, middle and youngest, respectively. The sample space, $S = \{e_1, e_2, \dots, e_8\} = \{GGG, GGB, GBG, \dots, BBB\}$, consists of 8 outcomes. The events

$$\begin{aligned} A &= \text{“At least two boys”} = \{BBB, BBG, BGB, GBB\} \\ B &= \text{“At least two girls”} = \{GGG, GGB, GBG, BGG\} \end{aligned}$$

are complimentary, i.e. $B = A^c$. We also have

$$C = \text{“At least one of each gender”} = \{GGG, BBB\}^c$$

Exercise. Describe the events $D = \text{“The oldest is a boy”}$ and $E = \text{“No girl younger than any boy”}$ as sets. Find $A \cup D$ and $A \cap D$. Assume that the 8 possibilities in S are equally probable. Why is $P(D) = P(E) = 1/2$?

■

We have calculation rules to find the probability for complex events such as $A \cup B$, $A \cap B$ (taken up in the next section), and A^c :

Proposition 2 *i) If A and B are mutually exclusive events ($A \cap B = \emptyset$), then*

$$P(A \cup B) = P(A) + P(B) \quad ((3.3))$$

ii) If A is an arbitrary event, then

$$P(A^c) = 1 - P(A) \quad ((3.4))$$

iii) For arbitrary events, A and B ,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad ((3.5))$$

Proof. i): If for example, $A = \{e_1, e_2, e_3\}$ and $B = \{e_5, e_6\}$, are mutually exclusive, we find

$$A \cup B = \{e_1, e_2, e_3, e_5, e_6\}$$

and, consequently,

$$P(A \cup B) = P(e_1) + P(e_2) + P(e_3) + P(e_5) + P(e_6) = P(A) + P(B)$$

Equivalent reasoning applies to any arbitrary mutually exclusive events.

ii): Since $A \cup A^c = S$ and A, A^c are mutually exclusive, we can use (3.3) and get

$$1 = P(S) = P(A \cup A^c) = P(A) + P(A^c)$$

Therefore, it must be that $P(A^c) = 1 - P(A)$.

iii): For example, let $A = \{e_1, e_2, e_3, e_4\}$ and $B = \{e_3, e_4, e_5, e_6\}$. Then $A \cap B = \{e_3, e_4\}$ and

$$A \cup B = \{e_1, e_2, e_3, e_4, e_5, e_6\}$$

We find

$$\begin{aligned} P(A \cup B) &= P(e_1) + P(e_2) + P(e_3) + P(e_4) + P(e_5) + P(e_6) \\ P(A) &= P(e_1) + P(e_2) + P(e_3) + P(e_4) \\ P(B) &= P(e_3) + P(e_4) + P(e_5) + P(e_6) \end{aligned}$$

Combining $P(A)$ and $P(B)$, we get

$$\begin{aligned} P(A) + P(B) &= P(e_1) + P(e_2) + 2[P(e_3) + P(e_4)] + P(e_5) + P(e_6) = \\ &= P(A \cup B) + P(e_3) + P(e_4) = \\ &= P(A \cup B) + P(A \cap B) \end{aligned}$$

By pulling $P(A \cap B)$ on the other side of the equation we get (3.5). This reasoning can be generalized for all arbitrary events. ■

Example 3.2

Probability calculations are often complicated, even with the simple rule (2.5). Some times it is a good idea to try to calculate the probability for a complimentary event, $P(A^c)$, first, if it is easier, and then calculate $P(A)$ according to (3.4): $P(A) = 1 - P(A^c)$. For example, we could try to find the probability for $A =$ “At least two with the same birthday” in a group of 3 randomly selected people. The outcome of recording the birthdays for the three people can be described as (x, y, z) where x, y, z are all between 1 and 365. The sample space consists of $365^3 = 48\,627\,125$ possible outcomes which we assume are equally probable. (This assumption is not completely realistic given that we do not take account for leap year, twins and the fact that individual months (e.g. May, June) are perhaps more common birth-months. There is still reason to believe that the assumption gives a reasonable approximation.)

In order to calculate $P(A)$ using (2.5), we must find $r =$ number of possible combinations (x, y, z) where at least two of the numbers are the same, which is possible, but a little complicated (Try!). Instead we try to find $P(A^c) = P(\text{“All three have different birthdays”})$. k is still 365^3 , while $r =$ the number of combinations (x, y, z) where all three numbers are different. The answer is $365 \cdot 364 \cdot 363$, which can be seen as: There are 365 possibilities for x . For *each* of these there are 364 possibilities for y such that x and y are different, consequently, there are $365 \cdot 364$ possibilities for x, y . For *each*

of these there are 363 possibilities for z such that x, y, z are all different, in all $365 \cdot 364 \cdot 363$. Therefore, we have

$$P(A^c) = \frac{365 \cdot 364 \cdot 363}{365^3} = 0.992$$

which gives us $P(A) = 1 - P(A^c) = 0.008$.

This reasoning can be generalized to a group of m people and leads to the following formula for the probability that at least two in the group have the same birthday.

$$P(A) = 1 - \frac{365 \cdot 364 \cdot 363 \cdots (365 - m + 1)}{365^m}$$

In the table some examples have been calculated

m	$P(A^c)$	$P(A)$
3	0.992	0.008
20	0.589	0.411
30	0.294	0.706
50	0.030	0.970

In a lecture with 50 students it is nearly certain that at least two in the room have the same birthday. ■

3.2 Conditional probability and independence

Probability is a relative concept that depends on what we know about the outcome of the experiment. For example, the probability that a randomly selected taxpayer in Norway has low income (defined by a year's (taxable) income of less than 150,000kr.) is equal to 64% (based on numbers from 1996 given in the Statistical Yearbook 1998). However, if we get to know that the randomly selected taxpayer is a woman, the probability increases to 77%. A probability that is calculated with respect to additional information that one may have obtained about the outcome, is called conditional probability (conditional on what we know about the outcome of the experiment) and is defined by:

Definition (See motivation in section (3.3))

Let A and B be arbitrary events. Assume that B is possible, i.e. $P(B) > 0$. If we know that B has occurred, the probability for A may be altered. In that case, we write $P(A|B)$, which is read as *the conditional probability for A , given B* . It can be calculated with the formula

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad ((3.6))$$

where $P(B)$ and $P(A \cap B)$ refer to the original experiment (where we do not know whether B has occurred or not).

By multiplying both sides of (3.6) with $P(B)$, we get the important *multiplication theorem*.

$$P(A \cap B) = P(B) \cdot P(A|B) \quad ((3.7))$$

(3.7) proves that $P(A \cap B)$ can be decomposed into a product of probabilities. The theorem is useful because it often proves to be easier to derive $P(B)$ and $P(A|B)$ than $P(A \cap B)$ directly.

Example 3.3

We know that a certain family with two children has at least one girl. What is the probability that the other child is also a girl?

Solution. We describe the outcome of recording the children's genders for a arbitrary two child family as xy where x is the gender for the oldest child (G or B) and y for the youngest. The sample space is $S = \{GG, GB, BG, BB\}$. We assume that all individual outcomes are equally likely. Set

$$\begin{aligned} A &= \text{"Both children are girls"} = \{GG\} \\ B &= \text{"At least one of the children is a girl"} = \{GG, GB, BG\} \end{aligned}$$

We wish to find $P(A|B)$. B Has the probability $3/4$ and $A \cap B = \{GG\}$ has the probability $1/4$. Therefore,

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{1}{4}}{\frac{3}{4}} = \frac{1}{3}$$

Exercise: Find the probability that both children are girls if we know that the oldest is a girl. ■

Example 3.4

Assume we draw two cards from a deck of cards. Let E_1 and E_2 denote the event "draw an ace in the first draw" and "draw an ace in the second draw", respectively. The probability of pulling two aces can be found by the multiplication theorem

$$P(E_1 \cap E_2) = P(E_1) \cdot P(E_2|E_1) = \frac{4}{52} \cdot \frac{3}{51} = 0.00435$$

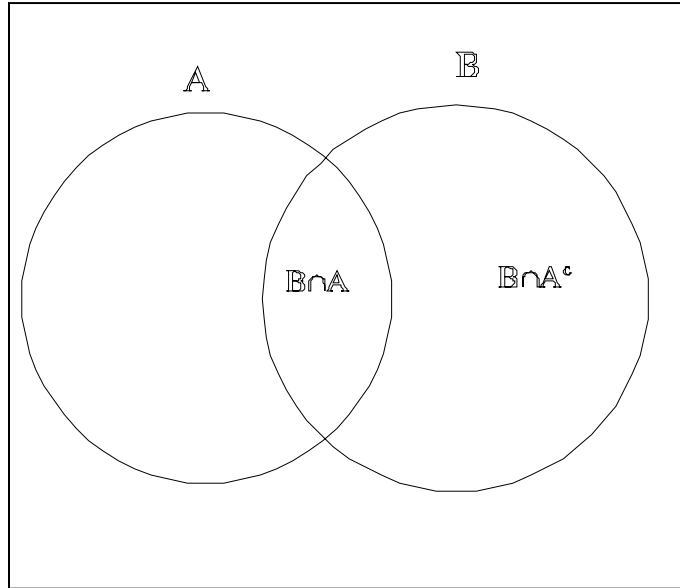


Figure 2: B can be divided into two mutually exclusive parts.

The first probability in the product, $P(E_1) = 4/52$, should be obvious. $P(E_2|E_1) = 3/51$ should also be clear. If we know that an ace has been drawn in the first trial, then there are 51 cards left, of which 3 are aces before the second draw.

A result that often causes confusion is that $P(E_2) = P(E_1) = 4/52$. In order to understand this, it is important to be clear about what $P(E_2)$ means, namely the probability of an ace when we do not know the result of the first draw. It is as if we lay the first card upside down on the table without looking at it before we pull the second card. Even though the deck of cards now has 51 cards, there are still 52 possible results, of which 4 are aces, when we draw the second card.

The result, $P(E_2) = 4/52$, can also be calculated by use of calculation rules for probability. Let A and B be two arbitrary events. B can then be divided up into two mutually exclusive parts, one part consisting of what B has in common with A , namely $B \cap A$, and the other part consisting of what B has in common with A^c , namely $B \cap A^c$ (see figure 2).

$$B = [B \cap A] \cup [B \cap A^c]$$

Since the union is mutually exclusive, the probability for B is

$$P(B) = P(B \cap A) + P(B \cap A^c)$$

Using this for E_1 and E_2 we get

$$P(E_2) = P(E_2 \cap E_1) + P(E_2 \cap E_1^c)$$

and subsequently by the multiplication theorem

$$\begin{aligned} P(E_2) &= P(E_1) \cdot P(E_2|E_1) + P(E_1^c) \cdot P(E_2|E_1^c) = \\ &= \frac{4}{52} \cdot \frac{3}{51} + \frac{48}{52} \cdot \frac{4}{51} = \\ &= \frac{4}{52 \cdot 51} (3 + 48) = \frac{4 \cdot 51}{52 \cdot 51} = \\ &= \frac{4}{52} \end{aligned}$$

■

We get an important special case of conditional probability when

$$P(A|B) = P(A) \tag{3.8}$$

i.e. the knowledge that B has occurred does not influence the probability of A . In this case we say that A and B are *independent* events. By definition of equation (3.6) we get

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

which can be written

$$P(A \cap B) = P(A) \cdot P(B) \tag{3.9}$$

It is equation (3.9) which is usually preferred as the definition of independence rather than (3.8). The reason is that equation (3.9) does not assume that $P(B) > 0$, upon which (3.8) is dependent.

Definition.

A and B are called independent events if

$$P(A \cap B) = P(A) \cdot P(B) \tag{3.10}$$

If (3.10) is not satisfied, we say that A and B are dependent.

Example 3.5

We toss a coin two times. The outcome is xy where x is the result of the first toss (H or T) and y of the second toss. The sample space is $S = \{TT, TH, HT, HH\}$. We assume that the four outcomes are equally probable. Let $K_1 = \text{“Heads in the first toss”} = \{HH, HT\}$ and $K_2 = \text{“Heads in the second toss”} = \{HH, TH\}$. We find that

$$\begin{aligned}P(K_1) &= P(K_2) = \frac{1}{2} \\P(K_1 \cap K_2) &= P(KK) = \frac{1}{4}\end{aligned}$$

We see that (3.10) is satisfied and we can therefore conclude that K_1 and K_2 are independent (which we expected). ■

In example 3.5 we deduced independence. More commonly, independence is used as a helping tool to build probability models. For example:

Example 3.6

Batteries come in two qualities, high (H) and low (L). Assume that the factory produces around 5% low-quality batteries. We record the quality of an arbitrary two-pack of batteries. The outcome is xy where x is the quality of the first battery and y of the second battery. $S = \{HH, HL, LH, LL\}$. It is reasonable to set $P(L) = 0.05$ and $P(H) = 0.95$ for a randomly chosen battery. If we in addition *assume* that the quality of the batteries in a two-pack is independent of each other, we have, for example, $P(HH) = P(H) \cdot P(H) = 0.95^2 = 0.9025$, $P(HL) = P(H) \cdot P(L) = 0.95 \cdot 0.05 = 0.0475$, etc. We therefore get the following probability model for S :

Outcome	HH	HL	LH	LL
Probability	0.9025	0.0475	0.0475	0.0025

which is a non-uniform probability model. The probability that the two-pack has an H-battery and a L-battery is for example $2 \cdot 0.0475 = 0.095$. ■

3.3 Motivation for conditional probability

Let us define low income as 0-150 000, middle income as 150 000 - 300 000, and high income as over 300 000 kr. The experiment consists of randomly selecting a taxpayer in Norway and recording the income category and gender of the taxpayer. The outcome is described by xy where x is gender (w for woman and m for man), and y is income category (l for low, m for middle and h for high). The sample space is

$$S = \{e_1, e_2, \dots, e_6\} = \{wl, wm, wh, ml, mm, mh\}$$

Assume that we repeat the experiment $n = 100$ times (i.e. select a completely random sample of 100 tax payers) and get the following frequencies

	Abs. frequency			Rel. frequency		
	Man	Woman	Total	Man	Woman	Total
Low	27	37	64	0.27	0.37	0.64
Middle	20	10	30	0.20	0.10	0.30
High	5	1	6	0.05	0.01	0.06
Total	52	48	100	0.52	0.48	1

((3.11))

Consider the events

$$\begin{aligned} L &= \text{“Low income”} = \{wl, ml\} \\ W &= \text{“Woman”} = \{wl, wm, wh\} \\ W \cap L &= \{wl\} \end{aligned}$$

We estimate the probabilities with relative frequencies calculated from the table (where h and H signify the relative and absolute frequencies, respectively)

$$\begin{aligned} P(L) &\approx h_L = \frac{H_L}{n} = \frac{64}{100} = 0.64 \\ P(W) &\approx h_W = \frac{H_W}{n} = \frac{48}{100} = 0.48 \\ P(W \cap L) &\approx h_{W \cap L} = \frac{H_{W \cap L}}{n} = \frac{37}{100} = 0.37 \end{aligned}$$

Suppose we find out that the randomly selected taxpayer is a woman. Then we would rather estimate the probability of L by

$$P(L|K) \approx \frac{37}{48} = \frac{H_{W \cap L}}{H_W} = 0.771$$

since 37 of 48 women selected have low incomes. Employing the law of large numbers (see introduction), we know that $h_{W \cap L}$ will converge towards $P(W \cap L)$ and h_W towards $P(W)$ as n increases. Therefore,

$$\frac{H_{W \cap L}}{H_W} = \frac{\frac{H_{W \cap L}}{n}}{\frac{H_W}{n}} = \frac{h_{W \cap L}}{h_W}$$

will converge towards $P(K \cap L)/P(K)$, which motivates the definition

$$P(L|W) = \frac{P(W \cap L)}{P(W)}$$

Equivalently, we find

$$P(L|M) = \frac{P(M \cap L)}{P(M)} \approx \frac{h_{M \cap L}}{h_M} = \frac{0.27}{0.52} = 0.519$$

In the table below we have calculated the conditional income distribution for woman and men together with the total distribution. This demonstrates how knowledge about the selected gender influences the probability for the three income categories. The numbers are representative for Norway in 1996 since the relative frequencies in table (3.11) are based on the Statistical Yearbook.

A	$P(A W)$	$P(A M)$	$P(A)$
$L = \text{“low”}$	0.771	0.519	0.64
$M = \text{“middle”}$	0.208	0.385	0.30
$H = \text{“high”}$	0.021	0.096	0.06
Total	1	1	1

3.4 Exercises

3.1 Jensen is interested in 3 stocks called 1,2 and 3 and wishes to invest in two of them. Stock 1 has the greatest growth potential, while stock 3 has the least growth potential. Jensen does not know this, and therefore chooses to gamble on two stocks selected completely at random from the three (i.e. so that all selections of two from 1,2,3 are equally probable). Define the events

- $A =$ “The two stocks with the greatest growth potential are selected”
- $B =$ “Stock 3 selected”
- $C =$ “Stock 1 selected”
- $D =$ “At least one of stocks 2 and 3 selected”

- a. Find $P(A^c)$, $P(B)$, $P(C)$, $P(D)$, $P(B \cap C)$, $P(B \cup C)$.
- b. Find $P(A|B)$ and $P(B|A)$. Are A and B independent? Are they mutually exclusive?
- c. Find $P(C|A)$ and $P(A|C)$. Are A and C independent? Are they mutually exclusive?

3.2 The economist in exercise 2.1 estimated the probability for success for the three job interviews as $1/4$, $3/4$ and $1/2$, respectively. Continue assuming that the outcome of the three interviews are independent of each other.

- a. Create a probability model for the sample space in exercise 2.1(Hint: Refer to example 3.6).
- b. Let S_0, S_1, S_2, S_3 represent the events “0 successes”, “1 success”, “2 successes”, “3 successes” respectively. Find the probability of these four events.
- c. Let $A =$ “At least 2 S’s” and $B =$ “1 or 2 F’s”. Find $P(S_2|A)$. Are A and B independent? Are A and B mutually exclusive?

3.3 Let A and B be two arbitrary and mutually exclusive events such that $P(A) > 0$ and $P(B) > 0$. Can they be independent?

3.4 A character, Mr. Hope, created by Sherlock Holmes, wants to commit two murders motivated by revenge. The plan consists of first presenting two indistinguishable pills, of which one contains a deadly poison, to the victim who is tricked into choosing one of the pills. Mr. Hope will then take the other pill. The plan is then to repeat the procedure for the other victim (Mr. Hope believes that he is protected by a higher power). What is the probability that he will succeed with his plan (provided that his belief is merely superstition)?

4 Stochastic (or Random) Variables, Expectation and Variance

4.1 Random variables and probability distribution

The word stochastic, which originates from greek, means essentially "random". A stochastic (i.e. random) variable is a numerical quantity for which the value is determined at random, i.e. by an outcome of an experiment. We often use uppercase letters for random variables, such as X, Y, Z, \dots etc. Examples:

$$\begin{aligned} X &= \text{Number of sixes in two tosses of one fair die.} \\ Y &= \text{Number of defective batteries in an arbitrary 3-pack.} \\ Z &= \text{Income of a randomly selected taxpayer.} \end{aligned} \tag{4.1}$$

Mathematically speaking, we can interpret any given numerical outcome of an experiment as a random variable.

The statistical properties of a random variable, X , is given by the probability distribution of X . This is determined by two things:

- The sample space of X , i.e. a set which includes all the possible values that X can assume: $S_X = \{x_1, x_2, x_3, \dots, x_k\}$
- A list of all associated probabilities: $P(X = x_1), P(X = x_2), P(X = x_3), \dots, P(X = x_k)$

The sample spaces of X and Y in (4.1) are, $S_X = \{0, 1, 2\}$ and $S_Y = \{0, 1, 2, 3\}$

The sample space of Z often takes the form of an interval, $S_Z = [0, \infty)$, i.e.. such that all numbers ≥ 0 are permitted as possible observations of Z (it turns out not to matter that S_Z also includes practically impossible values of Z . It is not a problem if S_Z is too large, just as long as all possible values are included.) This is an example of a

continuous sample space which is handled in a bit different manner and is not taken up in this course. Here we will only look at discrete sample spaces, i.e. where the possible values lie as isolated points on the number line.

Example 4.1

Let us look at our first example, $X =$ the number of sixes in two tosses of one fair die. The possible values that X can take are 0, 1, 2, so that the sample space of X is $S_X = \{0, 1, 2\}$. What is left to determine are the three associated probabilities, $P(X = 0)$, $P(X = 1)$, $P(X = 2)$. The latter is the easiest. Let $A_1 =$ “Six in first toss” and $A_2 =$ “Six in the second toss”. It is reasonable to assume that the outcome of the two tosses are independent of each other such that A_1 and A_2 are independent events. In addition, $P(A_1) = P(A_2) = 1/6$. We therefore get

$$P(X = 2) = P(\text{“Two sixes”}) = P(A_1 \cap A_2) = P(A_1) \cdot P(A_2) = \frac{1}{36} \quad (4.2)$$

The sample space for the underlying experiment can, like in example 2.2, be described by

$$\begin{aligned} S &= \{e_1, e_2, \dots, e_{36}\} = \\ &= \{(1, 1), (1, 2), (1, 3), \dots, (1, 6), \\ &\quad (2, 1), (2, 2), (2, 3), \dots, (2, 6), \\ &\quad (3, 1), (3, 2), (3, 3), \dots, (3, 6), \\ &\quad \vdots \\ &\quad (6, 1), (6, 2), (6, 3), \dots, (6, 6)\} \end{aligned}$$

where, for example, the outcome (1, 3) means that the first toss gave a one and the second a three. For the same reason as (4.2), the probability of (1, 3) is 1/36. The same naturally applies for all of the other outcomes. We find

$$P(X = 1) = P(\{(6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (5, 6), (4, 6), (3, 6), (2, 6), (1, 6)\}) = \frac{10}{36}$$

Since $P(X = 0) + P(X = 1) + P(X = 2) = 1$ (why?), we have

$$P(X = 0) = 1 - P(X = 1) - P(X = 2) = 1 - \frac{10}{36} - \frac{1}{36} = \frac{25}{36}$$

The probability distribution for X can now be described by the following table

x	0	1	2	(4.3)
$P(X = x)$	$\frac{25}{36}$	$\frac{10}{36}$	$\frac{1}{36}$	

The expression $f(x) = P(X = x)$ can be seen as a function, defined for $x = 0, 1, 2$ by $f(0) = 25/36$, $f(1) = 10/36$, $f(2) = 1/36$. This is called the *elementary probability function for X*.

Now assume that we are offered the following bet: In return for paying a 4 kr. playing fee we get to toss the die 2 times and receive 10 kr. for each six we get. What is the probability distribution for the net gain, Y ?

The net gain can be expressed by help of X such that:

$$Y = 10X - 4$$

Since X is a random variable, Y must also be random (the value of Y is randomly determined). The possible values of Y are given by

$$\begin{array}{c|ccc} X & 0 & 1 & 2 \\ \hline Y & -4 & 6 & 16 \end{array}$$

We find $P(Y = -4) = P(X = 0) = 25/36$, etc. and the distribution of Y

$$\begin{array}{c|ccc} y & -4 & 6 & 16 \\ \hline P(Y = y) & \frac{25}{36} & \frac{10}{36} & \frac{1}{36} \end{array} \quad (4.4)$$

The elementary probability function for Y is $g(y) = P(Y = y)$, defined in (4.4) for $y = -4, 6, 16$.

Exercise: Explain why $Z =$ "the number of ones in two tosses of the die", has the same distribution as X .

■

4.2 Expectation and variance

The expected value of a random variable, X , is a number, written $E(X)$, which can be calculated from the probability distribution of X , **i.e.** *before* we have observed which value X takes when the experiment is carried out. The number $E(X)$ is constructed such that the average value of X when X is observed many times, approaches $E(X)$ as the number of observations increases. Assume that X is observed n times (i.e. the experiment is repeated n times under the same conditions). Label the observations we get as $a_1, a_2, a_3, \dots, a_n$ where a_1 is the value of X in the first experiment, a_2 is the value of X in the second experiment, etc. The average (mean) of the observations, often written as \bar{x} , is

$$\bar{x} = \frac{a_1 + a_2 + \dots + a_n}{n} = \frac{1}{n} \sum_{i=1}^n a_i$$

As n increases \bar{x} will, under certain condition, stabilize and approach the number $E(X)$. For this reason $E(X)$ is sometimes referred to as "the mean value of X in the

long run". This is a variant of the law of large numbers that can actually be proven mathematically from the axiom system for probability.

If $E(X)$ is unknown, the number can be estimated by \bar{x} if we have access to an observation set for X . In statistics, one has methods for calculating uncertainty by such an estimate.

So, how is $E(X)$ calculated? Let the probability distribution of X be given by $f(x) = P(X = x)$ for $x = x_1, x_2, \dots, x_k$, or

$$\frac{x}{f(x) = P(X = x)} \mid \begin{array}{cccccc} x_1 & x_2 & x_3 & \cdots & x_k \\ f(x_1) & f(x_2) & f(x_3) & \cdots & f(x_k) \end{array} \quad (4.5)$$

Definition (See motivation in section (4.4)).

If X has the distribution given by (4.5), the expected X is calculated by

$$\begin{aligned} E(X) &= x_1 P(X = x_1) + x_2 P(X = x_2) + \cdots + x_k P(X = x_k) \\ &= x_1 f(x_1) + x_2 f(x_2) + \cdots + x_k f(x_k) \\ &= \sum_{\text{All } x} x P(X = x) \end{aligned} \quad (4.6)$$

Example 4.2

Let us find the expected value of $X =$ "the number of sixes in two tosses of one die" from example 4.1. The distribution is

$$\frac{x}{P(X = x)} \mid \begin{array}{ccc} 0 & 1 & 2 \\ \frac{25}{36} & \frac{10}{36} & \frac{1}{36} \end{array}$$

By definition we get

$$E(X) = 0 \cdot \frac{25}{36} + 1 \cdot \frac{10}{36} + 2 \cdot \frac{1}{36} = \frac{12}{36} = \frac{1}{3}$$

It may seem a bit strange that the expected value is not among the possible values of X , but when one thinks of $E(X)$ as an average value of X (in the long run), it should nonetheless be understandable.

For the net gain, Y , in example 4.1 we get (from 4.4)

$$E(Y) = (-4) \cdot \frac{25}{36} + 6 \cdot \frac{10}{36} + 16 \cdot \frac{1}{36} = -\frac{24}{36} = -\frac{2}{3}$$

The expected net gain is negative, which means that we, on average (when we play many times), lose by participating in the game. ■

The expectations operator, E , fulfills some important algebraic properties which are brought together in the following proposition

Proposition 3 Let X, Y, \dots denote stochastic variables and a, b, \dots constants. It then holds that

i) $E(b) = b$

ii) $E(aX) = a E(X)$

iii) $E(aX + b) = a E(X) + b$

iv) If $y = g(x)$ is an arbitrary function of x , then $Y = g(X)$ is a stochastic variable with expectation

$$\begin{aligned} E(Y) &= E(g(X)) = \sum_{\text{All } x} g(x) P(X = x) = & (4.7) \\ &= g(x_1) P(X = x_1) + g(x_2) P(X = x_2) + \dots + g(x_k) P(X = x_k) \end{aligned}$$

v) If X_1 and X_2 are stochastic variables, $Y = X_1 + X_2$ is a stochastic variable with expectation

$$E(Y) = E(X_1 + X_2) = E(X_1) + E(X_2)$$

vi) If X_1, X_2, \dots, X_n are stochastic variables, $Y = X_1 + X_2 + \dots + X_n$ is a stochastic variable with expectation

$$E(Y) = E(X_1 + X_2 + \dots + X_n) = E(X_1) + E(X_2) + \dots + E(X_n)$$

We will not prove proposition 3 here but we will comment about some of the results. *i)* says that the expectation of a constant is equal to the constant itself, which intuitively seems reasonable. It does, however, follow from our definition. This is because the constant b can be seen as a trivial stochastic variable, Z , which always takes the value b . The sample space of Z is then $S_Z = \{b\}$ with the associated probability $P(Z = b) = 1$. The definition of expectation therefore gives $E(Z) = b P(Z = b) = b$.

Check of *iii)*: In example 4.2 with $X =$ "the number of sixes in two tosses of one die", we found that $E(X) = 1/3$ and $E(Y) = -2/3$, where $Y = 10X - 4$, both found by the definition of expectation. Using *iii)*, we get

$$E(Y) = E(10X - 4) \stackrel{\text{iii)}}{=} 10 E(X) - 4 = \frac{10}{3} - 4 = \frac{10 - 12}{3} = -\frac{2}{3}$$

which is the same as we found from the definition and represents a small check of *iii)*.

ii) is a special case of *iii)* (for $b = 0$).

iv) is a very important result. It says that we do not need to first find the expectation of $Y = g(X)$ (which can be complicated) in order to find $E(Y)$. We can find it directly from the distribution of X which was reported in (4.7). Assume that we wish to find $E(Z)$ where $Z = (X - 2)^2$ from the distribution of X in example 4.2. (4.7) gives

$$\begin{aligned} E(Z) &= E[(X - 2)^2] \stackrel{\text{iv)}}{=} \sum_{\text{All } x} (x - 2)^2 P(X = x) = \\ &= (0 - 2)^2 \frac{25}{36} + (1 - 2)^2 \frac{10}{36} + (2 - 2)^2 \frac{1}{36} = \\ &= 4 \cdot \frac{25}{36} + 1 \cdot \frac{10}{36} = \frac{110}{36} = 3.056 \end{aligned}$$

A relevant example where *iv*) becomes useful is the calculation of expected utility: Let A be an uncertain (numerical) asset which assumes the following values a_1, a_2, \dots, a_k , with probability, $p_i = P(A = a_i)$, for $i = 1, 2, \dots, k$. If $U(a)$ is a utility function we can use *iv*) to calculate the expected utility of A :

$$E(U(A)) = \sum_{\text{All } a} U(a) P(A = a) = U(a_1)p_1 + U(a_2)p_2 + \dots + U(a_k)p_k \quad (4.8)$$

Example 4.3 (continuation of ex. 4.1)

We may call a gamble, where the expected net gain is 0, *fair*. Is there a playing fee for the game in example 4.1 which makes the game fair? Let the playing fee be a kr. The net gain is then $Y = 10X - a$. The game is fair if

$$0 = E(Y) \stackrel{\text{iii)}}{=} 10 E(X) - a = \frac{10}{3} - a = 3.333\dots - a$$

that is to say $a = \text{kr. } 3,333\dots$. A fair playing fee is therefore not possible in Norwegian currency. A fee of 3.50 gives a loss in the long run, while 3 kroner gives a positive net gain in the long run.

Assume the playing fee is 3.50. What is the expected total net gain for 1000 plays? For a simple game it is $E(Y) = \frac{10}{3} - \frac{7}{2} = -\frac{1}{6}$. Let Y_i be the net gain for game no. i , for $i = 1, 2, 3, \dots, 1000$. The total net gain is then $Y = Y_1 + Y_2 + \dots + Y_{1000}$. From proposition 3 *vi*) we get

$$E(Y) \stackrel{\text{vi)}}{=} E(Y_1) + \dots + E(Y_{1000}) = -\frac{1}{6} - \frac{1}{6} - \dots - \frac{1}{6} = -\frac{1000}{6} = -167 \text{ kroner}$$

■

We will now define the *variance* of a stochastic variable, X , which we write $\text{Var}(X)$. The variance of X is a measure of the degree of variation of X . The distribution of X and Y in example 4.1 have the same probabilities, but the possible values of Y are more spread out on the number line than the X values. Y varies with $\{-4, 6, 16\}$ as its possible values while X varies with $\{0, 1, 2\}$ as its possible values. Y thus has greater variation than X .

Now let X be an arbitrary stochastic variable. Set $E(X) = \mu$ (where μ is a greek letter for “m” and pronounced “mu”). The idea is that if the distance between X and μ , namely $X - \mu$, is mostly large, X has a large spread, but if $X - \mu$ is mostly small, the spread is also small. An initial suggestion for a measure of spread (variation) could therefore be the long run average value of $X - \mu$, i.e. $E(X - \mu)$. This is not a good measure since

we always have $E(X - \mu) = E(X) - \mu = \mu - \mu = 0$, and the reason for this is that the average negative distances will exactly cancel out the positive distances. Another measure could be to look at the absolute distance, $|X - \mu|$, instead and measure the long run average value, $E(|X - \mu|)$, as the measure of variation. This measure appears quite often in newer literature, but is mathematically more complex to handle than the classic measure which builds on the squared distance, $(X - \mu)^2$, which also does away with the sign of $X - \mu$. The variance of X is defined as the long run average value of the stochastic variable $(X - \mu)^2$, namely:

Definition

$$\text{Var}(X) = E[(X - \mu)^2] \tag{4.9}$$

By use of the rule in proposition 3 we can write (4.9) in another way, which often results in simpler calculations

$$\text{Var}(X) = E(X^2) - \mu^2 \tag{4.10}$$

Proof.

$$\begin{aligned} \text{Var}(X) &= E[(X - \mu)^2] = E[X^2 - 2\mu X + \mu^2] \stackrel{\text{vi)}}{=} \\ &= E(X^2) + E(-2\mu X) + E(\mu^2) \stackrel{\text{i),ii)}}{=} \\ &= E(X^2) - 2\mu E(X) + \mu^2 = E(X^2) - 2\mu^2 + \mu^2 = \\ &= E(X^2) - \mu^2 \end{aligned}$$

where *i*), *ii*), etc. refers to proposition 3. ■

Definition

The standard deviation of X is defined as $\sqrt{\text{Var}(X)}$. The standard deviation measures the same thing as $\text{Var}(X)$, but on a different scale. The standard deviation is also measured in the same units as X (if X is measured in "kroner", the standard deviation is also measured in kroner).

Example 4.4 (continuation of ex. 4.1)

We will now calculate the variance of X and Y in example 4.1. From proposition 3 we get

$$\begin{aligned} E(X^2) &= 0^2 \frac{25}{36} + 1^2 \frac{10}{36} + 2^2 \frac{1}{36} = \frac{7}{18} \\ E(Y^2) &= (-4)^2 \frac{25}{36} + 6^2 \frac{10}{36} + 16^2 \frac{1}{36} = \frac{254}{9} \end{aligned}$$

In example 4.2 we found $E(X) = 1/3$ and $E(Y) = -2/3$. We now use (4.10)

$$\begin{aligned}\text{Var}(X) &= \frac{7}{18} - \left(\frac{1}{3}\right)^2 = \frac{5}{18} \\ \text{Var}(Y) &= \frac{254}{9} - \left(-\frac{2}{3}\right)^2 = \frac{250}{9}\end{aligned}$$

Standard deviation is

$$\begin{aligned}\sqrt{\text{Var}(X)} &= \sqrt{\frac{5}{18}} = 0.527 \\ \sqrt{\text{Var}(Y)} &= \sqrt{\frac{250}{9}} = 5.270\end{aligned}$$

Our measure of variation therefore confirms that Y has considerably greater variation than X . ■

In addition, the variance fulfills certain algebraic laws. For example, a constant, a , has no variation, which is confirmed by $\text{Var}(a) = 0$:

$$\begin{aligned}\text{Var}(a) &= E[(a - E(a))^2] = E[(a - a)^2] = \\ &= E(0^2) = E(0) = 0\end{aligned}$$

We have

Proposition 4 *Let X, Y, \dots denote random variables and a, b, \dots constants. It then holds that*

- a) $\text{Var}(a) = 0$
- b) $\text{Var}(bX + a) = b^2 \text{Var}(X)$

a) we have already demonstrated. b) follows by the same method,

$$\begin{aligned}\text{Var}(bX + a) &= E[(bX + a - E(bX + a))^2] \\ &= E[(bX + a - bE(X) - a)^2] = E[(bX - bE(X))^2] \\ &= E[b^2(X - E(X))^2] = b^2 E[(X - E(X))^2] \\ &= b^2 \text{Var}(X)\end{aligned}$$

By using proposition 4 b), the calculation of $\text{Var}(Y)$ in example 4.4 will become a bit simpler:

$$\text{Var}(Y) = \text{Var}(10X - 4) = 100 \text{Var}(X) = 100 \frac{5}{18} = \frac{500}{18} = \frac{250}{9}$$

Example 4.5 (continuation of example 4.1)

Is the expected net gain a sufficient measure of how attractive a game of uncertainty is? In order to get some insight into this question, we will look at two games. Game 1 is similar to example 4.1, but with a playing fee of 2 kroner. The net gain then becomes $Y = 10X - 2$, where X , as before, is the number of sixes in two tosses of a die. The expected net gain is

$$E(Y) = \frac{10}{3} - 2 = \frac{4}{3}$$

Game 2 consists of receiving 10,000 kr. for every six one gets, in two tosses of the die. In order to participate in this game, however, one must pay 3,332 kr. The net gain, Z , is then $Z = 10,000X - 3,332$, and the expected net gain is

$$E(Z) = \frac{10,000}{3} - 3,332 = \frac{4}{3}$$

Games 1 and 2 therefore have the same expected net gain, but are they equally attractive? Most people would likely answer "no" since the risk is different. Y varies in the sample space $S_Y = \{-2, 8, 18\}$ with the probability of a loss of 2 kroner equal to $25/36 \approx 69\%$. The probability of losing is large, but the loss is so small that many people would be interested in playing. Z , on the other hand, varies between $S_Z = \{-3,332, 6,668, 16,668\}$ with associated probabilities, 69%, 28%, 3% respectively. The probability for a loss of 3,332 kr. is 69%. Because of widespread risk aversion, many would not wish to gamble in game 2. The difference in risk can be expressed by the difference in variation between Y and Z , for example measured by the standard deviations, $\sqrt{\text{Var}(Y)} = 5.27$ and $\sqrt{\text{Var}(Z)} = 5\,270$ (check!).

In general, the risk of a game (investment) with an uncertain asset will not only depend on the variation of the net gain of the current game but also on the variation of other, related games in the market (see Varian, "Intermediate Microeconomics", section 13.2). ■

4.3 Jensen's inequality

A very famous inequality in probability calculations is Jensen's inequality:

Proposition 5 *If X is a stochastic variable with expectation, $E(X)$, and $g(x)$ is an arbitrary concave and differentiable function, it holds that*

a) $Eg(X) \leq g(E(X))$

b) *If $g(x)$ is strictly concave and $\text{Var}(X) > 0$, we have strict inequality in a): $Eg(X) < g(E(X))$.*

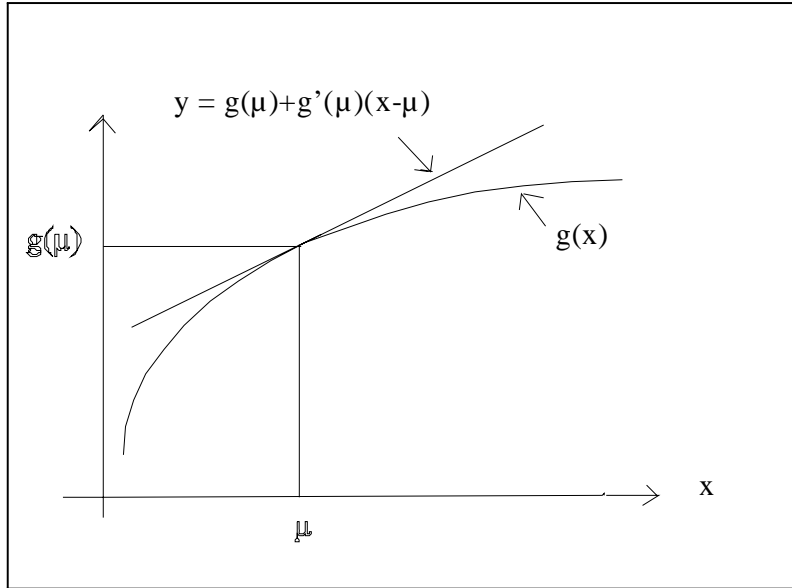


Figure 3: The tangent to the concave curve always lies above the curve

Proof. Set $\mu = E(X)$. The tangent to the curve $g(x)$ for $x = \mu$, has the equation of the following form

$$y = l(x) = g(\mu) + g'(\mu)(x - \mu)$$

where $g'(\mu)$ is the derivative of $g(x)$ for $x = \mu$. Since g is concave, the tangent, in its entirety, will lie above $g(x)$, i.e. $l(x) \geq g(x)$ for all x . (See figure 3). If g is strictly concave ($g''(x) < 0$ for all x), $l(x) > g(x)$ will be true for all $x \neq \mu$.

If we replace x with the stochastic variable X , we get two new stochastic variables, $Y = l(X)$, and $Z = g(X)$, which must fulfill $Y \geq Z$ irrespective of the outcome of the experiment which decides the values of Y and Z . Therefore, it must also be true that $E(Y) \geq E(Z)$ (think of the expectation as the average value). Since $g(\mu)$ and $g'(\mu)$ are constants, then according to proposition 3

$$\begin{aligned} E(Y) &= E[g(\mu) + g'(\mu)(X - \mu)] = g(\mu) + g'(\mu) E(X - \mu) = \\ &= g(\mu) + g'(\mu)(\mu - \mu) = g(\mu) = g(E(X)) \end{aligned}$$

We therefore have that $E(Z) = E g(X) \leq E(Y) = g(E(X))$ which is a). If g is strictly concave and $\text{Var}(X) > 0$, then $Y > Z$ with a positive probability, which leads to (omitting some details here) that $E(Y) > E(Z)$, which gives b). ■

Example 4.6 (Risk aversion)

Let A be an uncertain (numerical) asset which takes the following values, a_1, a_2, \dots, a_k , with probabilities, $p_i = P(A = a_i)$, for $i = 1, 2, \dots, k$. Assume

that an individual has a strictly concave utility function, $U(a)$, with respect to the assets a_1, a_2, \dots, a_k . Let the expected value of A be

$$\mu = E(A) = a_1p_1 + a_2p_2 + \dots + a_kp_k$$

By Jensen's inequality (b), it follows that the expected utility of A (see (4.8)) fulfills

$$EU(A) < U(\mu)$$

If the utility of the uncertain asset, A , can be represented by expected utility, $EU(A)$, it follows that such an individual will prefer a certain asset with value μ to an uncertain asset, A , with an expected value equal to μ .

■

4.4 Motivation of expectation

Let us try to motivate $E(X) = 1/3$ for $X =$ "number of sixes in two tosses of the die", from example 4.1. The probability distribution for X is given by

x	0	1	2
$P(X = x)$	$\frac{25}{36}$	$\frac{10}{36}$	$\frac{1}{36}$

Imagine that we observe X n times (i.e. we repeat two tosses of the die n times) and get data of the type 0, 0, 2, 1, 0, 0, 0, 0, 1, 0, ..., 0, 1, which we can describe as n numbers, $a_1, a_2, a_3, \dots, a_n$. We can collect the numbers in a frequency table.

x	H_x	$h_x = \frac{H_x}{n}$
0	H_0	h_0
1	H_1	h_1
2	H_2	h_2
Sum	n	1

where H_x, h_x for $x = 0, 1, 2$ are the absolute and relative frequencies, respectively, of the value x . We get the average value for X from the data

$$\bar{x} = \frac{1}{n}(a_1 + a_2 + a_3 + \dots + a_n)$$

Since the order in which we sum the a 's does not matter, we can first sum the 0's, then the 1's and, lastly the 2's:

$$\begin{aligned} \bar{x} &= \frac{1}{n}(\underbrace{0 + \dots + 0}_{H_0} + \underbrace{1 + \dots + 1}_{H_1} + \underbrace{2 + \dots + 2}_{H_2}) \\ &= \frac{1}{n}(0 \cdot H_0 + 1 \cdot H_1 + 2 \cdot H_2) \\ &= 0 \cdot \frac{H_0}{n} + 1 \cdot \frac{H_1}{n} + 2 \cdot \frac{H_2}{n} \\ &= 0 \cdot h_0 + 1 \cdot h_1 + 2 \cdot h_2 \end{aligned}$$

In accordance with the law of large numbers, h_x will (we believe) approach $P(X = x)$ when n increases. Examining the last expression for \bar{x} we see that \bar{x} will therefore approach the number

$$0 \cdot P(X = 0) + 1 \cdot P(X = 1) + 2 \cdot P(X = 2)$$

which is precisely the expression for the definition of $E(X)$. By plugging the numbers in, we see that \bar{x} approaches

$$0 \cdot \frac{25}{36} + 1 \cdot \frac{10}{36} + 2 \cdot \frac{1}{36} = \frac{1}{3}$$

which we found for $E(X)$ in example 4.1.

4.5 Exercises

4.1 (Continuation of exercises 2.1 and 3.2). Let X be the number of successful interviews attended by the economist. Set up the distribution of X and find the expectation and variance of X .

4.2 Consider the following game. You may toss a coin up to 4 times. If the coin is heads (H) in the first toss, the game stops and you do not get paid any prize. If the coin is tails (T) in the first toss, you are allowed to toss it another time. If the coin is H in the second toss, the game stops and you get paid 2 kroner. If it is T in the second toss, you are allowed to toss it again. If it is H in the third toss, the game stops and you get 4 kroner. If it is T the on third toss, you may toss the coin a fourth time. If it is H on the fourth toss, the game stops and you get 8 kroner. If it is T on the fourth toss, the game also stops, but you get nothing.

In order to participate in the game, you must pay a kroner. Determine a such that the game is fair. Is the variance of the net gain dependent on a ?

4.3 Let the stochastic variable, X , have expectation, μ , and standard deviation, σ , (σ is a greek “s” and is called “small sigma”). To standardize X means to subtract the expectation and divide by the standard deviation. Show (i.e. by using the propositions in the text) that the standardized variable Z

$$Z = \frac{X - \mu}{\sigma}$$

has expectation 0 and standard deviation 1.

4.4 Let X have the distribution given by (4.3). Find the probability distribution of $Y = (X - 1)^2$.

5 Answers to the exercises

2.1 a. $S = \{SSS, SSF, SFS, FSS, FFS, FSF, SFF, FFF\}$

b. $\{SSF, SFS, FSS\}$ and $\{SSS, SSF, SFS, FSS\}$

2.2 a. $S = \{0, 1, 2, 3, 4\}$

b. 0.7

c. 0.66

d. No

2.3 $4/7 = 0.571$

2.4 $1/6$

2.5 a. $1/3$

b. $10/21 = 0.476$

3.1 a. $2/3, 2/3, 2/3, 1, 1/3, 1.$

b. 0, 0, no, yes

c. 1, $1/2$, no, no

3.2 a.

Outcome	SSS	SSF	SFS	FSS	FFS	FSF	SFF	FFF
Probability	$3/32$	$3/32$	$1/32$	$9/32$	$3/32$	$9/32$	$1/32$	$3/32$

b. $3/32, 13/32, 13/32, 3/32$

c. $13/16$, yes, no

3.3 No

3.4 $1/4$

4.1

x	0	1	2	3
$P(X = x)$	$3/32$	$13/32$	$13/32$	$3/32$

, $3/2, 5/8$

4.2 1.50, no

4.4

y	0	1
$P(Y = y)$	$5/18$	$13/18$