INTRODUCTION TO PROBABILITY THEORY AND STATISTICS

HEINRICH MATZINGER
Georgia Tech
E-mail: matzi@math.gatech.edu

December 4, 2003

Contents

1 Definition and basic properties

1.1 Events

Imagine that we throw a die which has 4 sides. The outcome, of this experiment will be one of the four numbers: 1,2,3 or 4. The set of all possible outcomes in this case is:

$$\Omega = \{1, 2, 3, 4\}.$$

Ω is called the *outcome space* or *sample space*. Before doing the experiment we don't know what the outcome will be. Each possible outcome has a certain probability to occur. This die-experiment is a random experiment.
We can use our die to make bets. Somebody might bet that the number will be even. We throw the die: if the number we see is 2 or 4 we say that the event "even" has occurred or has been observed. We can identify the event "even" with the set: $\{2, 4\}$. This might seem a little bit abstract, but by identifying the event with a set, events become easier to handle: Sets are well known mathematical objects, whilst the events as we know them from every day language are not.
In a similar way one might bet that the outcome is a number greater-equal to 3. This event is realized when we observe a 3 or a 4. The event greater or equal 3 can thus be viewed as the set $\{3, 4\}$.
Another example, is the event "odd". This is the set: $\{1, 3\}$.
With this way of looking at things, events are simply subsets of Ω. Take another example: a coin with a side 0 and a side 1. The outcome space or sample space in that case is:

$$\Omega = \{0, 1\}.$$

The events are the subsets of Ω, in this case there are 4 of them:

$$\emptyset, \{0\}, \{1\}, \{0, 1\}.$$

Example 1.1 *It might at first seem very surprising that events can be viewed as sets. Consider for example the following sets:*
the set of bicycles which belong to a Ga tech student, the set of all sky-scrapers in Atlanta, the set of all one dollar bills which are currently in the US.

Let us give a couple of events:
the event that after X-mas the unemployment rate is lower than now, the event that our favorite pet dies from hart attack, the event that I go down with flue next week.
At first, it seems that events are something very different from sets. Let us see in a real world example how mathematicians view events as sets:
Assume that we are interested in where the American economy is going to stand in exactly one year from now. More specifically, we look at unemployment and inflation and wonder if they will be above or below their current level. To describe the situation which we encounter in a year from now, we introduce a two digit variable $Z = XY$. Let X be equal to one if unemployment is higher in a year than its current level. If it is lower, let X be equal to 0. Similarly, let Y be equal to one if inflation is higher in a year from now. If it is lower, let Y be equal to zero. The possible outcomes for Z are:

$$00, 01, 10, 11.$$

This is the situation of a random experiment, where the outcome is one of the four possible numbers: 00,01,10,11. We don't know what the outcome will be. But each possibility can occur with a certain probability. Let A be the event that unemployment is higher in a year. This corresponds to the outcomes 10 and 11. We thus identify the event A with the set:

$$\{10, 11\}.$$

Let B be the event that inflation is higher in a year form now. This corresponds to the outcomes 01 and 11. We thus view the event B as the set:

$$\{01, 11\}.$$

Recall that the intersection $A \cap B$ of two sets A and B, is the set consisting of all elements contained in both A and B. In our example, the intersection of A and B is equal to $A \cap B = \{11\}$. Let C designate the event that unemployment goes up and that inflation goes up at the same time. This corresponds to the outcome 11. Thus, C is identified with the set: $\{11\}$. In other words, $C = A \cap B$. The general rule which we must remember is:
For any events A and B, if C designates the event that A and B both occur at the same time, then $C = A \cap B$.
Let D be the event that unemployment or inflation will be up in a year from now. (By "or" we mean that at least one of them is up.) This corresponds to the outcomes: 01,10,11. Thus D gets identified with the set:

$$D = \{01, 10, 11\}.$$

Recall that the union of two sets A and B is defined to be the set consisting of all elements which are in A or in B. We see in our example that $D = A \cup B$. This is true in general. We must thus remember the following rule:
For any events A and B, if D designates the event that A or B occur, then $D = A \cup B$.

1.2 Frequencies

Assume that we have a six sided die. In this case the outcome space is

$$\Omega = \{1, 2, 3, 4, 5, 6\}.$$

The event "even" in this case is the set:

$$\{2, 4, 6\}$$

whilst the event "odd" is equal to

$$\{1, 3, 5\}.$$

Instead of throwing the die only once, we throw it several times. As a result, instead of just a number, we get a sequence of numbers. When throwing the six-sided die I obtained the sequence:

$$1, 4, 3, 5, 2, 6, 3, 4, 5, 3, \ldots$$

When repeating the same experiment which consists in throwing the die a couple of times, we are likely to obtain another sequence. The sequence we observe is a random sequence. In this example we observe one 3 within the first 5 trials and three 3's occurring within the first 10 trials. We write:

$$n_{\{3\}}$$

for the number of times we observe a 3 among the first n trials. In our example thus: for $n = 5$ we have $n_{\{3\}} = 1$ whilst for $n = 10$ we find $n_{\{3\}} = 3$.
Let A be an event. We denote by n_A the number of times A occurred up to time n.
Take for example A to be the event "even". In the above sequence within the first 5 trials we obtained 2 even numbers. Thus for $n = 5$ we have that $n_A = 2$. Within the first 10 trials we found 4 even numbers. Thus, for $n = 10$ we have $n_A = 4$. The proportion of even numbers n_A/n for the first 5 trials is equal to $2/5 = 40\%$. For the first 10 trials, this proportion is $4/10 = 40\%$.

1.3 Definition of probability

The basic definition of probability which we use is based on frequencies. For our definition of probability we need an "assumption" about the world surrounding us:
Let A designate an event. When we repeat the same random experiment independently many times we observe that on the long run the proportion of times A occurs tends to stabilize. Whenever we repeat this experiment, the proportion n_A/n on the long run tends to be the same number. A more mathematical way of formulating this, is to stay that n_A/n converges to a number only depending on A, as n tends to infinity. This is our basic assumption.

Assumption As we keep repeating the same random experiment under the same conditions and such that each trial is independent of the previous ones, we find that: the proportion n_A/n tends to a number which only depends on A, as $n \to \infty$.

We are now ready to give our definition of probability:

Definition 1.1 *Let A be an event. Assume that we repeat the same random experiment under exactly the same conditions independently many times. Let n_A designate the number of times the event A occurred within the n first repeats of the experiment. We define the probability of the event A to be the real number:*

$$P(A) =: \lim_{n\to\infty} \frac{n_A}{n}.$$

Thus, $P(A)$ designates the probability of the event A. Take for example a four-sided perfectly symmetric die. Because, of symmetry each side must have same probability. On the long run we will see a forth of the times a 1, a forth of the times a 2, a forth of the times a 3 and a forth of the times a 4. Thus, for the symmetric die the probability of each side is 0.25.

1.4 Direct consequences

From our definition of probability there are several useful facts, which follow immediately:

1. For any event A, we have that:
$$P(A) \geq 0.$$
2. For any event A, we have that:
$$P(A) \leq 1.$$
3. Let Ω designate the state space. Then:
$$P(\Omega) = 1.$$

Let us prove these elementary facts:

1. By definition $n_a/n \geq 0$. However, the limit of a sequence which is ≥ 0 is also ≥ 0. Since $P(A)$ is by definition equal to the limit of the sequence n_A/n we find that $P(A) \geq 0$.
2. By definition $n_A \leq n$. It follows that $n_a/n \leq 1$. The limit of a sequence which is always less or equal to one must also be less or equal to one. Thus, $P(A) = \lim_{n\to\infty} n_a/n \leq 1$.
3. By definition $n_\Omega = n$. Thus:
$$P(\Omega) = \lim_{n\to\infty} n_\Omega/n = \lim_{n\to\infty} n/n = \lim_{n\to\infty} 1 = 1.$$

The next two theorems are essential for solving many problems:

Theorem 1.1 *Let A and B be disjoint events. Then:*

$$P(A \cup B) = P(A) + P(B).$$

Proof. Let C be the event $C = A \cup B$. C is the event that A or B has occurred. Because A and B are disjoint, we have that A and B can not occur at the same time. Thus, when we count up to time n how many times C has occurred, we find that this is exactly equal to the number of times A has occurred plus the number or times B has occurred. In other words,

$$n_C = n_A + n_B. \tag{1.1}$$

From this it follows that:

$$P(C) = \lim_{n\to\infty} \frac{n_C}{n} = \lim_{n\to\infty} \frac{n_A + n_B}{n} = \lim_{n\to\infty} \left(\frac{n_A}{n} + \frac{n_B}{n} \right).$$

We know that the sum of limits is equal to the limit of the sum. Applying this to the right side of the last equality above, yields:

$$\lim_{n\to\infty} \left(\frac{n_A}{n} + \frac{n_B}{n} \right) = \lim_{n\to\infty} \frac{n_A}{n} + \lim_{n\to\infty} \frac{n_B}{n} = P(A) + P(B).$$

This finishes to prove that

$$P(C) = P(A \cup B) = P(A) + P(B).$$

■

Let us give an example which might help us to understand why equation 1.1 holds. Imagine we are using a 6-sided die. Let A be the event that we observe a 2 or a 3. Thus $A = \{2, 3\}$. Let B be the event that we observe a 1 or a 5. Thus, $B = \{1, 5\}$. The two events A and B are disjoint: it is not possible to observe at the same time A and B since $A \cap B = \emptyset$. Assume that we throw the die 10 times and obtain the sequence of numbers:

$$1, 3, 4, 6, 3, 4, 2, 5, 1, 2.$$

We have seen the event A four times: at the second, fifth, seventh and tenth trial. The event B is observed at the first trial, at the eight and ninth trials. $C = A \cup B = \{2, 3, 1, 5\}$ is observed at the trials number: 2,5,7,10 and 1,8,9. We thus find in this case that $n_A = 4$, $n_B = 3$ and $n_C = 7$ which confirms equation 1.1.

Example 1.2 *Assume that we are throwing a fair coin with sides 0 and 1. Let X_i designate the number which we obtain when we flip the coin for the i-th time. Let A be the event that we observe right at the beginning the number 111. In other words:*

$$A = \{X_1 = 1, X_2 = 1, X_3 = 1\}.$$

Let B designate the event that we observe the number 101 when we read our random sequence starting from the second trial. Thus:

$$B = \{X_2 = 1, X_3 = 0, X_4 = 1\}.$$

Assume that we want to calculate the probability to observe that at least one of the two events A or B holds. In other words we want to calculate the probability of the event $C = A \cup B$.
Note that A and B can not occur both at the same time. The reason is that for A to hold it is necessary that $X_3 = 1$ and for B to hold it is necessary that $X_3 = 0$. X_3 however can not be equal at the same time to 0 ant to 1. Thus, A and B are disjoint events, so we are allowed to use theorem 1.1.We find, applying theorem 1.1 that:

$$P(A \cup B) = P(A) + P(B).$$

With a fair coin, each 3-digit number has same probability. There are 8, 3-digit numbers so each one has probability $1/8$. It follows that $P(A) = 1/8$ and $P(B) = 1/8$. Thus

$$P(A \cup B) = \frac{1}{8} + \frac{1}{8} = \frac{1}{4} = 25\%$$

The next theorem is useful for any pair of events A and B and not just disjoint events:

Theorem 1.2 *Let A and B be two events. Then:*

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

Proof. Let $C = A \cup B$. Let $D = B - A$, that is D consists of all the elements that are in B, but not in A. We have by definition that $C = D \cup A$ and that D and A are disjoint. Thus we can apply theorem 1.1 and find:

$$P(C) = P(A) + P(D) \tag{1.2}$$

Furthermore $(A \cap B)$ and D are disjoint, and we have $B = (A \cap B) \cup D$. We can thus apply theorem 1.1 and find that:

$$P(B) = P(A \cap B) + P(D) \tag{1.3}$$

Subtracting equation 1.3 from equation 1.2 yields:

$$P(C) - P(B) = P(A) - P(A \cap B).$$

By adding $P(B)$ on both sides of the last equation, we find:

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

This finishes this proof. ■

Problem 1.1 *Let a and b designate two genes. Let the probability that a randomly picked person in the US, has gene a be 20%. Let the probability for gene b be 30%. And eventually, let the probability that he has both genes at the same time be 10%. What is the probability to have at least one of the two genes?*

Let us explain how we solve the above problem: Let A, resp. B designate the event that the randomly picked person has gene a, resp. b. We know that:

- $P(A) = 20\%$
- $P(B) = 30\%$
- $P(A \cap B) = 10\%$

The event to have at least one gene is the event $A \cup B$. By theorem 1.2 we have that: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. Thus in our case: $P(A \cup B) = 20\% + 30\% - 10\% = 40\%$. This finishes to solve the above problem.
Often it is easier to calculate the probability of a complement than the probability of the event itself. In such a situation, the following theorem is useful:

Theorem 1.3 *Let A be an event and let A^c denote its complement. Then:*

$$P(A) = 1 - P(A^c)$$

Proof. Note that the events A and A^c are disjoint. Furthermore by definition $A \cup A^c = \Omega$. Recall that for the sample space Ω, we have that $P(\Omega) = 1$. We can thus apply theorem 1.1 and find that:

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

This implies that:

$$P(A) = 1 - P(A^c)$$

which finishes this proof. ■

1.5 Some inequalities

Theorem 1.4 *Let A and B be two events. Then:*

$$P(A \cup B) \leq P(A) + P(B)$$

Proof. We know by theorem 1.2 that

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

Since $P(A \cap B) \geq 0$ we have that

$$P(A) + P(B) - P(A \cap B) \leq P(A) + P(B).$$

It follows that

$$P(A \cup B) \leq P(A) + P(B).$$

■ For several events a similar theorem holds:

Theorem 1.5 *Let $A_1, \ldots, A_n$ be a collection of n events. Then*

$$P\left(A_1 \cup A_2 \cup \ldots A_n \cup\right) \leq A_1 + \ldots + A_n$$

Proof. By induction.

■ Another often used inequality is:

Theorem 1.6 *Let $A \subset B$. Then:*

$$P(A) \leq P(B).$$

Proof. If $A \subset B$, then for every n we have that:

$$n_A \leq n_B$$

hence also

$$\frac{n_A}{n} \leq \frac{n_B}{n}.$$

Thus:

$$\lim_{n\to\infty} \frac{n_A}{n} \leq \lim_{n\to\infty} \frac{n_B}{n}.$$

Hence

$$P(A) \leq P(B).$$

■

2 Expectation

In general if X denotes the outcome of a random experiment, then we call X a random variable.

Definition 2.1 *Let us consider a random experiment with a finite number of possible outcomes, where the state space is*

$$\Omega = \{x_1, x_2, \ldots, x_s\}.$$

Let X denote the outcome of this random experiment. For $x \in \Omega$, let p_x denote the probability that the outcome of our random experiment is is x. That is:

$$p_x := P(X = x).$$

We define the expected value $E[X]$:

$$E[X] := \sum_{x\in\Omega} x p_x.$$

Example 2.1 *Let X denote the value which we obtain when we throw a fair coin with side 0 and side 1. Then we find that:*

$$E[X] = 0.5 \times 1 + 0.5 \times 0 = 0.5$$

When we keep repeating the same random experiment independently and under the same conditions. Then, on the long run, we will see that the average value which we observe converges to the expectation. This is the content of the next theorem:

Theorem 2.1 *Assume we repeat the same random experiment under the same conditions independently many times. Let X_i denote the (random variable) which is the outcome of the i-th experiment. Then:*

$$\lim_{n\to\infty} \frac{(X_1 + X_2 + \ldots + X_n)}{n} = E[X_1] \tag{2.1}$$

This simply means that on the long run, the average is going to be equal to to the expectation.

Proof. Let Ω denote the state space of the random variables X_i:

$$\Omega = \{x_1, x_2, \ldots, x_s\}.$$

by regrouping the same terms together, we find:

$$X_1 + X_2 + \ldots + X_n = x_1 n_{x_1} + x_2 n_{x_2} + \ldots + x_s n_{x_s}.$$

(Remember that n_{x_i} denotes the number of times we observe the value x_i in the finite sequence: $X_1, X_2, \ldots, X_n$.) Thus:

$$\lim_{n\to\infty} \frac{(X_1 + X_2 + \ldots + X_n)}{n} = \lim_{n\to\infty} \left(x_1 \frac{n_{x_1}}{n} + \ldots + x_s \frac{n_{x_s}}{n} \right).$$

By definition

$$P(X_1 = x_i) = \lim_{n\to\infty} \frac{n_{x_i}}{n}.$$

Since the limit of a sum is the sum of the limits we find,

$$\begin{aligned} \lim_{n\to\infty} \left(x_1 \frac{n_{x_1}}{n} + \ldots + x_s \frac{n_{x_s}}{n} \right) &= x_1 \lim_{n\to\infty} \frac{n_{x_1}}{n} + \ldots + x_s \lim_{n\to\infty} \frac{n_{x_s}}{n} = \\ &= x_1 P(X = x_1) + \ldots + x_s P(X = x_s) = E[X_1]. \end{aligned}$$

■

The last theorem is called Law of Large Numbers.

Example 2.2 *Imagine that the profit a firm makes every month is random. Imagine also that the earnings from month to month are independent of each other and also have the same "probabilities". In this case we can view the sequence of earnings month for month, as a sequence of repeats of the same random experiment. Because of theorem 2.1, on the long run the monthly income will be equal to the expectation.*

Let us next give a few useful lemmas.

Lemma 2.1 *Let X denote the outcome of a random experiment. (Thus X is a so-called random variable.) Let a be a real (non-random) number. Then:*

$$E[aX] = aE[X].$$

Proof. Let us repeat the same experiment independently many times. Let X_i denote the outcome of the i-th trial. Let Y_i be equal to $Y_i := aX_i$. Then by the law of large numbers, we have that

$$\lim_{n\to\infty} \frac{Y_1 + \ldots + Y_n}{n} = E[Y_1] = E[aX_1].$$

However:

$$\lim_{n\to\infty} \frac{Y_1 + \ldots + Y_n}{n} = \lim_{n\to\infty} \frac{aX_1 + \ldots + aX_n}{n} = \lim_{n\to\infty} a\left(\frac{X_1 + \ldots + X_n}{n}\right) =$$
$$= a \lim_{n\to\infty} \frac{X_1 + \ldots + X_n}{n} = aE[X_1].$$

This proves that $E[aX_1] = aE[X_1]$ and finishes this proof. ■

Lemma 2.2 *Let X, Y denote the outcomes of two random experiments.*
Then:

$$E[X + Y] = E[X] + E[Y].$$

Proof. Let us repeat the two random experiments independently many times. Let X_i denote the outcome of the i-th trial of the first random experiment. Let Y_i be equal to the outcome of the i-th trial of the second random experiment. For all $i \in \mathbb{N}$, let $Z_i := X_i + Y_i$. Then by the law of large numbers, we have that:

$$\lim_{n\to\infty} \frac{Z_1 + \ldots + Z_n}{n} = E[Z_1] = E[X_1 + Y_1].$$

However:

$$\lim_{n\to\infty} \frac{Z_1 + \ldots + Z_n}{n} = \lim_{n\to\infty} \frac{X_1 + Y_1 + X_2 + Y_2 + \ldots + X_n + Y_n}{n} =$$
$$= \lim_{n\to\infty} \left(\frac{(X_1 + \ldots + X_n) + (Y_1 + \ldots + Y_n)}{n}\right) =$$
$$= \lim_{n\to\infty} \frac{X_1 + \ldots + X_n}{n} + \lim_{n\to\infty} \frac{Y_1 + \ldots + Y_n}{n} + = E[X_1] + E[Y_1].$$

This proves that $E[X_1 + Y_1] = E[X_1] + E[Y_1]$ and finishes this proof. ■

Lemma 2.3 *Let X, Y denote the outcomes of two independent random experiments. Then:*

$$E[X \cdot Y] = E[X] \cdot E[Y].$$

Proof. Still to come. ■

3 Variance

Let X be the outcome of a random experiment. We define the variance of X to be equal to:

$$VAR[X] := E\left[\ (X - E[X])^2\ \right].$$

The square root of the variance is called standard deviation:

$$\sigma_X := \sqrt{VAR[X]}.$$

The standard deviation is a measure for the typical order of magnitude of how far away the value we get after doing the experiment once, is from $E[X]$.

Lemma 3.1 *Let a be a non-random number and X the outcome of a random experiment. Then:*

$$VAR[aX] = a^2 VAR[X].$$

Proof. We have:

$$\begin{aligned} VAR[aX] =& E[(aX - E[aX])^2] = E[(aX - aE[X])^2] = \\ =& E[a^2 \cdot (X - E[X])^2] = a^2 E[(X - E[X])^2] = a^2 \cdot VAR[X], \end{aligned}$$

which finishes to prove that: $VAR[aX] = a^2 VAR[X]$. ■

Lemma 3.2 *Let X be the outcome of a random experiment, (in other words a random variable). Then:*

$$VAR[X] = E[X^2] - (E[X])^2.$$

Proof. We have that

$$E[(X - E[X])^2] = E[X^2 - 2XE[X] + E[X]^2] = E[X^2] - 2E[XE[X]] + E[E[X]^2]. \quad (3.1)$$

Now $E[X]$ is a constant and constants can be taken out of the expectation. This implies that

$$E[XE[X]] = E[X]E[X] = E[X]^2. \quad (3.2)$$

On the other hand, the expectation of a constant is the constant itself. Thus, since $E[X]^2$ is a constant, we find:

$$E[E[X]^2] = E[X]^2. \quad (3.3)$$

Using equation 3.2 and 3.3 with 3.1 we find

$$E[(X - E[X])^2] = E[X^2] - 2E[X]^2 + E[X]^2 = E[X^2] - E[X]^2.$$

this finishes to prove that $VAR[X] = E[X^2] - E[X]^2$.

■

Lemma 3.3 *Let X and Y be the outcomes of two random experiments, which are independent of each other. Then:*

$$VAR[X+Y] = VAR[X] + VAR[Y].$$

Proof. We have:

$$\begin{aligned} VAR[X+Y] =& E[((X+Y) - E[X+Y])^2] = E[(X+Y-E[X]-E[Y])^2] = \\ & E[((X-E[X]) + (Y-E[Y]))^2] = \\ =& E[(X-E[X])^2 + 2(X-E[X])(Y-E[Y]) + (Y-E[Y])^2] = \\ =& E[(X-E[X])^2] + 2E[(X-E[X])(Y-E[Y])] + E[(Y-E[Y])^2] = \end{aligned}$$

Since X and Y are independent, we have that $(X - E[X])$ is also independent from $(Y - E[Y])$. Thus, we can use lemma 2.3, which says that the expectation of a product equals the product of the expectations in case the variables are independent. We find:

$$E[(X-E[X])(Y-E[Y])] = E[X-E[X]] \cdot E[Y-E[Y]].$$

Furthermore:

$$E[X-E[X]] = E[X] - E[E[X]] = E[X] - E[X] = 0$$

Thus

$$E[(X-E[X])(Y-E[Y])] = 0.$$

Applying this to the above formula for $VAR[X+Y]$, we get:

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

This finishes our proof. ■

4 Combinatorics

Theorem 4.1 *Let*

$$\Omega = \{x_1, x_2, \ldots, x_s\}.$$

denote the state space of a random experiment. Let each possible outcome have same probability. Let $E \subset \Omega$ be an event. Then,

$$P(E) = \frac{\text{number of outcomes in E}}{\text{total number of outcomes}} = \frac{|E|}{s}$$

Proof. We know that

$$P(\Omega) = 1$$

Now

$$P(\Omega) = P(X \in \{x_1, \ldots, x_s\}) = P(\{X = x_1\} \cup \ldots \cup \{X = x_s\}) = \sum_{t=1,\ldots,s} P(X = x_t)$$

since all the outcomes have equal probability, we have that

$$\sum_{t=1,\ldots,s} P(X = x_t) = sP(X = x_1).$$

Thus,

$$P(X = x_1) = \frac{1}{s}.$$

Now if:

$$E = \{y_1, \ldots, y_j\} \subset \Omega$$

We find that:

$$P(E) = P(X \in E) = P(\{X = y_1\} \cup \ldots \cup \{X_j = y_j\}) = \sum_{i=1}^{j} P(X = y_i) = \frac{j}{s}$$

which finishes the proof.

■ Next we present one of the main principles used in combinatorics:

Lemma 4.1 *Let $m_1, m_2, \ldots, m_r$ denote a given finite sequence of natural numbers. Assume that we have to make a sequence of r choices. At the s-th choice, assume that we have m_s possibilities to choose from. Then the total number of possibilities is:*

$$m_1 \cdot m_2 \cdot \ldots \cdot m_r$$

Why this lemma holds can best be understood when thinking of a tree, where at each knot which is s away from the root we have m_s new branches.

Example 4.1 *Assume we first throw a coin with a side 0 and a side 1. Then we throw a four sided die. Eventually we throw the coin again. For example we could get the number 031. How many differ numbers are there which we could get? The answer is: First we have to possibilities. For the second "choice" we have four, and eventually we have again two. Thus, $m_1 = 2, m_2 = 4, m_3 = 2$. This implies that the total number of possibilities is:*

$$m_1 \cdot m_2 \cdot m_3 = 2 \cdot 4 \cdot 2 = 16.$$

Recall that the product of all natural numbers which are less or equal to k, is denoted by $k!$. $k!$ is called k-factorial.

Lemma 4.2 *There are*

$$k!$$

possibilities to put k different objects in a linear order. Thus there are $k!$ permutations of k elements.

To realize why the last lemma above holds we use lemma 4.1. To place k different objects in a row we first choose the first object which we will place down. For this we have k possibilities. For the second object, there remain $k-1$ objects to choose from. For the third, there are $k-3$ possibilities to choose from. And so on and so forth. This then gives that the total number of possibilities is equal to $k \cdot (k-1) \ldots \cdot 2 \cdot 1$.

Lemma 4.3 *There are:*

$$\frac{n!}{(n-k)!}$$

possibilities to pick k out of n different objects, when the order in which we pick them matters.

For the first object, we have n possibilities. For the second object we pick, we have $n-1$ remaining objects to choose from. For the last object which we pick, (that is the k-th which we pick), we have $n-k+1$ remaining objects to choose from. Thus the total number of possibilities is equal to:

$$n \cdot (n-1) \cdot \cdots \ldots \cdot (n-k+1)$$

which is equal to:

$$\frac{n!}{(n-k)!}.$$

The number $\frac{n!}{(n-k)!}$ is also equal to the number of words of length k written with a n-letter alphabet, when we require that the words never contain twice the same letter.

Lemma 4.4 *There are:*

$$\frac{n!}{k!(n-k)!}$$

subsets of size k in a set of size n.

The reason why the last lemma holds is the following: there are $k!$ ways of putting a given subset of size k into different orders. Thus, there are $k!$ times more ways to pick k elements, than there are subsets of size k.

Lemma 4.5 *There are:*

$$2^n$$

subsets of any size in a set of size n.

The reason why the last lemma above holds is the following: we can identify the subsets two binary vectors with n entries. For example, let $n = 5$. Let the set we consider be $\{1, 2, 3, 4, 5\}$. Take the binary vector:

$$(1, 1, 1, 0, 0).$$

This vector would correspond to the subset containing the first three elements of the set, thus to the subset:

$$\{1, 2, 3\}.$$

So, for every non zero entry in the vector we pick the corresponding element in the set. It is clear that this correspondence between subsets of a set of size n and binary vectors of dimension n is one to one. Thus, there is the same number of subsets as there is binary vectors of length n. The total number of binary vectors of dimension n however is 2^n.

5 Conditional probability

Definition 5.1 *Let A, B be two events. Then we define the probability of A conditional on the event B, and write $P(A|B)$ for the number:*

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

Definition 5.2 *Let A, B be two events. We say that A and B are independent of each other iff*

$$P(A \cap B) = P(A) \cdot P(B).$$

Note that A and B are independent of each other if and only if $P(A|B) = P(A)$. In other word, A and B are independent of each other if and only if the realization of one of the events does not affect the probability of the other.
Assume that we perform two random experiments independently of each other, in the sense that the two experiments do not interact. That is the experiments have no influence on each other. Le A denote an event related to the first experiment, and let B denote an event related to the second experiment. We saw in class that in this situation the equation $P(A \cap B) = P(A) \cdot P(B)$ must hold. And thus, A and B are independent in the sense of the above definition. To show this we used an argument where we simulated the two random experience by picking marbles from two bags.
There are also many cases, where events related to a same experiment are independent, in the sense of the above definition. For example for a fair die, the events $A = \{1, 2\}$ and $B = \{2, 4, 6\}$ are independent.
There can also be more than two independent events at a time:

Definition 5.3 *Let $A_1, A_2, \ldots, A_n$ be a finite collection of events. We say that $A_1, A_2, \ldots, A_n$ are all independent of each other iff*

$$P\left(\cap_{i \in I} A_i\right) = \Pi_{i \in I} P\left(A_i\right)$$

for every subset $I \subset \{1, 2, \ldots, n\}$.

The next example is very important for the test on Wednesday.

Example 5.1 *Assume we flip the same coin independently three times. Let the coin be biased, so that side 1 has probability* 60% *and side 0 has probability* 40%*. What is the probability to observe the number* 101*? (By this we mean: what is the probability to first get a 1, then a 0 and eventually, at the third trial, a 1 again?)*
To solve this problem let A_1*, resp.* A_3 *be the event that at the first, resp. third trial we get a one. Let* A_2 *be the event that at the second trial we get a zero. Observing a* 101 *is thus equal to the event* $A := A_1 \cap A_2 \cap A_3$*. Because, the trials are performed in an "independent" manner it follows that the events* A_1, A_2, A_3 *are independent of each other. Thus we have that:*

$$P(A_1 \cap A_2 \cap A_3) = P(A_1) \cdot P(A_2) \cdot P(A_3).$$

We have that:

$$P(A_1) = 60\%, P(A_2) = 40\%, P(A_3) = 60\%.$$

It follows that:

$$P(A_1 \cap A_2 \cap A_3) = 60\% \cdot 40\% \cdot 60\% = 0.144.$$

6 Important discrete random variables

6.1 Bernoulli variable

Let a coin have a side 0 and a side 1. Let p be the probability of side 1 and $1-p$ be the probability of side 0. Let X designate the random number we obtain when we flip this coin. Thus, with probability p the random variable X takes on the value 1 and with probability $1-p$ it takes on the value 0. The random variable X is called a *Bernoulli variable with parameter* p. It is named after the famous swiss mathematician Bernoulli.
For a Bernoulli variable X with parameter p we have:

- $E[X] = p$.
- $VAR[X] = p(1-p)$.

Let us show this:

$$E[X] = 1 \cdot p + 0 \cdot (1-p) = p.$$

For the variance we find:

$$VAR[X] = E[X^2] - (E[X])^2 = 1^2 \cdot p + 0^2 \cdot (1-p) - (E[X])^2 = p - p^2 = p(1-p).$$

6.2 Binomial random variable

Again, let a coin have a side 0 and a side 1. Let p be the probability of side 1 and $1-p$ be the probability of side 0. We toss this coin independently n times and count the numbers of 1's observed. The number Z of 1's observed after n coin-tosses is equal to

$$Z := X_1 + X_2 + \ldots + X_n$$

where X_i designates the result of the i-th toss. (Hence the X_i's are independent Bernoulli variables with parameter p.) The random variable Z is called a *binomial variable* with parameter p and n. For the binomial random variable with parameter p we find:

- $E[Z] = np$
- $VAR[Z] = np(1-p)$
- For $k \leq n$, we have: $P(Z = k) = \binom{n}{k} p^k (1-p)^{n-k}$.

Let us show the above statements:

$$E[Z] = E[X_1 + \ldots + X_n] = E[X_1] + \ldots + E[X_n] = n \cdot E[X_1] = n \cdot p.$$

Also:

$$VAR[Z] = VAR[X_1 + \ldots + X_n] = VAR[X_1] + \ldots + VAR[X_n] = nVAR[X_1] = np(1-p).$$

Let us calculate next the probability: $P(Z = k)$. We start with an example. Take $n = 3$ and $k = 2$. We want to calculate the probability to observe exactly to ones among the first three coin tosses. To observe exactly two ones out of three successive trials there are exactly three possibilities:

- Let A be the event: $X_1 = 1, X_2 = 1, X_3 = 0$
- Let B be the event: $X_1 = 1, X_2 = 0, X_3 = 1$
- Let C be the event: $X_1 = 0, X_2 = 1, X_3 = 1$.

Each of these possibilities has probability $p^2(1-p)$. As a matter of fact, since the trials are independent we have for example:

$$P\left(X_1 = 1, X_2 = 1, X_3 = 0\right) = P(X_1 = 1)P(X_2 = 1)P(X_3 = 0) = p^2(1-p).$$

The three different possibilities are disjoint of each other. Thus,

$$P(Z = 2) = P(A \cup B \cup C) = P(A) + P(B) + P(C) = 3p^2(1-p).$$

Here 3 is the number of realization where we have exactly two ones within the first three coin tosses. This is equal to the different number of ways, there is to choose two different objects out of three items. In other words the number three stand in our formula for "3 choose 2".
We can now generalize to n trials and a number $k \leq n$. There are "n choose k" possible outcomes for which among the first n coin tosses there appear exactly k ones. Each of these outcomes has probability:

$$p^k (1-p)^{(n-k)}.$$

This gives then:

$$P(Z = k) = \binom{n}{k} p^k (1-p)^{(n-k)}.$$

6.3 Geometric random variable

Again, let a coin have a side 0 and a side 1. Let p be the probability of side 1 and $1-p$ be the probability of side 0. We toss this coin independently n many times. Let X_i designate the result of the i-th coin-toss. Let T designate the number of trials it takes until we first observe a 1. For example, if we have:

$$X_1 = 0, X_2 = 0, X_3 = 1$$

we would have that $T = 3$. If we observe on the other hand:

$$X_1 = 0, X_2 = 1$$

we have that $T = 2$. T is a random variable. As we are going to show, we have:

- For $k > 0$, we have $P(T = k) = p(1-p)^{k-1}$.
- $E[T] = 1/p$
- $VAR[T] = (1-p)/p^2$

A random variable T for which $P(T = k) = p(1-p)^{k-1}$, $\forall k \in \mathbb{N}$, is called *geometric random variable with parameter p*. Let us next prove the above statements: For T to be equal to k we need to observe $k-1$ time a zero followed by a one. Thus:

$$P(T = k) = P(X_1 = 0, X_2 = 0, \ldots, X_{k-1} = 0, X_k = 1) = \\ P(X_1 = 0) \cdot P(X_2 = 0) \cdot \ldots \cdot P(X_{k-1} = 0) \cdot P(X_k = 1) = (1-p)^{k-1}p.$$

Let us calculate the expectation of T. We find:

$$E[T] = \sum_{k=1}^{\infty} kp(1-p)^{k-1}$$

Let $f(x)$ be the function:

$$x \mapsto f(x) = \sum_{k=1}^{\infty} kx^{k-1}.$$

We have that

$$f(x) = \sum_{k=1}^{\infty} \frac{d(x^k)}{dx} = \frac{d\left(\sum_{k=1}^{\infty} x^k\right)}{dx} = \frac{d\left(x/(1-x)\right)}{dx} = \frac{1}{1-x} + \frac{x}{(1-x)^2} = \frac{1}{(1-x)^2} \quad (6.1)$$

This shows that:

$$\sum_{k=1}^{\infty} k(1-p)^{k-1} = f(1-p) = \frac{1}{(p)^2}$$

Thus,

$$E[T] = p \cdot \left(\sum_{k=1}^{\infty} k(1-p)^{k-1} \right) = p \cdot \frac{1}{(p)^2} = \frac{1}{p}.$$

Let us next calculate the variance of a geometric random variable. We find:

$$E[T^2] = \sum_{k=1}^{\infty} k^2 p(1-p)^{k-1}.$$

Let $g(.)$ be the map:

$$x \mapsto g(x) = \sum_{k=1}^{\infty} k^2 (x)^{k-1}$$

We find:

$$g(x) = \sum_{k=1}^{\infty} k \frac{d(x^k)}{dx} = \frac{d\left(x \sum_{k=1}^{\infty} k x^{k-1}\right)}{dx}$$

Using equation 6.1, we find:

$$g(x) = \frac{d\left(x/(1-x)^2\right)}{dx} = \frac{1+x}{(1-x)^3}.$$

This implies that

$$E[T^2] = pg(1-p) = \frac{2-p}{p^2}.$$

Now,

$$VAR[T] = E[T^2] - (E[T]^2) = \frac{2-p}{p^2} - \left(\frac{1}{p}\right)^2 = \frac{1-p}{p^2}.$$

7 Continuous random variables

So far we have only been studying discrete random variables. Let us see how continuous random variables are defined.

Definition 7.1 *Let X be a number "generated by a random experiment". (Such a random number is also called random variable). X is a continuous random variable if there exists a non-negative piecewise continuous function*

$$f : x \mapsto f(x) \quad \mathbb{R} \to \mathbb{R}^+$$

such that for any interval $I = [i_1, i_2] \subset \mathbb{R}$ we have that:

$$P(X \in I) = \int_I f(x)dx.$$

The function $f(.)$ is called the density function of X or simply the density of X.

Note that the notation $\int_I f(x)dx$ stands for:

$$\int_I f(x)dx = \int_{i_1}^{i_2} f(x)dx. \tag{7.1}$$

Recall also that integrals like the one appearing in equation 7.1 are defined to be equal to the air under the curve $f(.)$ and above the interval I.

Remark 7.1 *Let $f(.)$ be a piecewise continuous function from $\mathbb{R}$ into $\mathbb{R}$. Then, there exists a continuous random variable X such that $f(.)$ is the density of X, if and only if all of the following conditions are satisfied:*

1. *f is everywhere non-negative.*
2. *$\int_{\mathbb{R}} f(x)dx = 1$.*

Let us next give some important examples of continuous random variables:

- The uniform variable in the interval $I = [i_1, i_2]$, where $i_1 < i_2$. The density of $f(.)$ is equal to $1/|i_2 - i_2|$ everywhere in the interval I. Anywhere outside the interval I, $f(.)$ is equal to zero.
- The standard normal variable has density:

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

  A standard normal random variable is often denoted by $\mathcal{N}(0, 1)$.
- Let $\mu \in \mathbb{R}, \sigma > 0$ be given numbers. The density of the normal variable with expectation μ and standard deviation σ is defined to be equal to:

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

8 Distribution functions

Definition 8.1 *Let X be a random variable. The distribution function $F_X : \mathbb{R} \to \mathbb{R}$ of X is defined in the following way:*

$$F_X(s) := P(X \leq s)$$

for all $s \in \mathbb{R}$.

Let us next mention a few properties of the distribution function:

- F_X is an increasing function. This means that for any two numbers $s < t$ in $\mathbb{R}$, we have that $F_X(s) \leq F_X(t)$.
- $\lim_{s\to\infty} F_X(s) = 1$
- $\lim_{s\to-\infty} F_X(s) = 0$

We leave the proof of the three facts above to the reader.
Imagine next that X is a continuous random variable with density function f_X. Then, we have for all $s \in \mathbb{R}$, that:

$$F_X(s) = P(X \leq s) = \int_{-\infty}^{s} f_X(t)dt.$$

Taking the derivative on all sides of the above system of equations we find that:

$$\frac{dF_X(s)}{ds} = f_X(s).$$

In other words, for a continuous random variables X, the derivative of the distribution function is equal to the density of X. Hence, in this case, the distribution function is differentiable and thus also continuous. Another implication is: the distribution function uniquely determines the density function of f. This implies, that the distribution function determines uniquely all the probabilities of events which can be defined in terms of X.

Assume next that the random variable X has a finite state space:

$$\Omega_X = \{s_1, s_2, \ldots, s_r\}$$

such that $s_1 < s_2 < \ldots < s_r$. Then, the distribution function F_X is a step function. Left of s_1, we have that F_X is equal to zero. Right of s_r it is equal to one. Between s_i and s_{i+1}, that is on the interval $[s_i, s_{i+1}[$, the distribution function is constantly equal to:

$$\sum_{j\leq i} P(X = s_j).$$

(This holds for all i between 1 and $r-1$.)
To sum up: for continuous random variables the distribution functions are differentiable functions, whilst for discrete random variables the distribution functions are step functions. Let us next show how we can use the distribution function to simulate random variables. The situation is the following: our computer can generate a uniform random variable U in the interval $[0, 1]$. (This is a random variable with density equal to 1 in $[0, 1]$ and 0 everywhere else.) We want to generate a random variable with a given probability density function f_X, using U. We do this in the following manner: we plug the random number U into the map $invF_X$. (Here $invF_X$ designates the inverse map of $F_X(.)$.) The next lemma says that this method really produces a random variable with the desired density function.

Lemma 8.1 *Let f_X denote the density function of a continuous random variable and let F_X designate its distribution function. Let Y designate the random variable obtained by plugging the uniform random variable U into the inverse distribution function:*

$$Y := invF_X(U).$$

Then, the density of Y is equal to f_X.

Proof. Since, $F(.)$ is an increasing function. Thus for any number s we have:

$$Y \leq s.$$

is equivalent to

$$F_X(Y) \leq F_X(s).$$

Hence:

$$P(Y \leq s) = P(F_X(Y) \leq F_X(s)).$$

Now, $F_X(Y) = U$, thus

$$P(Y \leq s) = P(U \leq F_X(s)).$$

We know that $F_X(s) \in [0, 1]$. Using the fact that U has density function equal to one in the interval $[0, 1]$, we find:

$$P(U \leq F_X(s)) = \int_0^{F_X(s)} 1 \, dt = F_X(s).$$

Thus

$$P(Y \leq s) = F_X(s).$$

This shows that the distribution function F_Y of Y is equal to $F_X(s)$. Applying the derivative according to s to both $F_Y(s)$ and $F_X(s)$, yields:

$$f_Y(s) = f_X(s).$$

Hence, X and Y have same density function. This finishes the proof. ■

9 Expectation and variance for continuous random variables

Definition 9.1 *Let X be a continuous random variable with density function $f_X(.)$. Then, we define the expectation $E[X]$ of X to be:*

$$E[X] := \int_{-\infty}^{\infty} s f_X(s) ds.$$

Next we are going to prove that the law of large numbers also holds for continuous random variables.

Theorem 9.1 *Let $X_1, X_2, \ldots$ be a sequence of i.i.d. continuous random variables all with same density function $f_X(.)$. Then,*

$$\lim_{n\to\infty} \frac{X_1 + X_2 + \ldots + X_n}{n} = E[X_1].$$

Proof. Let $\Delta > 0$ be a fix number. Let us approximate the continuous variables X_i by a discrete variable X_i^Δ. For this we let X_i^Δ be the largest integer multiple of Δ which is still smaller equal to X_i. In this way, we always get that

$$|X_i^\Delta - X_i| < \Delta.$$

This implies that:

$$\left| \frac{X_1 + X_2 + \ldots + X_n}{n} - \frac{X_1^\Delta + X_2^\Delta + \ldots + X_n^\Delta}{n} \right| < \Delta$$

However the variables X_i^Δ are discrete. So for them the law of large number has already been proven and we find:

$$\lim_{n\to\infty} \frac{X_1^\Delta + X_2^\Delta + \ldots + X_n^\Delta}{n} = E\left[X_1^\Delta\right] \tag{9.1}$$

We have that

$$E[X_i^\Delta] = \sum_{z\in\mathbb{Z}} z\Delta \cdot P(X_i^\Delta = z\Delta)$$

However, by definition:

$$P(X_i^\Delta = z\Delta) = P(X_i \in [z\Delta, (z+1)\Delta[).$$

The expression on the right side of the last inequality is equal to

$$\int_{z\Delta}^{(z+1)\Delta} f_X(s)ds.$$

Thus

$$E[X_i^\Delta] = \sum_{z\in\mathbb{Z}} z\Delta \int_{z\Delta}^{(z+1)\Delta} f_X(s)ds.$$

As Δ tends to zero, the expression on the left side of the last equality above tends to:

$$\int_{-\infty}^{\infty} s f_X(s)ds$$

This implies that by taking Δ fix and sufficiently small, we have that, for large enough n , the fraction

$$\frac{X_1 + X_2 + \ldots + X_n}{n}$$

is as close as we want from

$$\int_{-\infty}^{\infty} s f_X(s) ds.$$

This implies that

$$\frac{X_1 + X_2 + \ldots + X_n}{n}$$

actually converges to

$$\int_{-\infty}^{\infty} s f_X(s) ds.$$

■ The linearity of expectation holds in the same way as for discrete random variables. This is the content of the next lemma.

Lemma 9.1 *Let X and Y be two continuous random variables and let a be a number. Then*

$$E[X + Y] = E[X] + E[Y]$$

and

$$E[aX] = aE[X]$$

Proof. The proof goes like in the discrete case: The only thing used for the proof in the discrete case is the law of large numbers. Since the central limit theorem also holds for the continuous case, the exactly same proof holds for the continuous case. ■

10 Central limit theorem

The Central Limit Theorem (CLT) is one of the most important theorems in probability. Roughly speaking it says that if we build the sum of many independent random variables, no matter what these little contributions are, we will always get approximately a normal distribution. This is very important in every day life, because often times you have situations where a lot of little independent "things" add up. So, you end up observing something which is approximately a normal random variable. For example, when you make a measurement you are most of the time in this situation. That is, when you don't make one big measurement error. In that case, you have a lot of little imprecisions which add up to give you your measurement error. Most of the time, these imprecisions can be seen as close to being independent of each other. This then implies: unless you make one big error, you will always end up having your measurement-error being close to a normal variable.

Let $X_1, X_2, X_3, \ldots$ be a sequence of independent, identically distributed random variables. (This means that they are the outcome of the same random experiment repeated several times independently.) Let μ denote the expectation $\mu := E[X_1]$ and let σ denote the standard deviation $\sigma := \sqrt{VAR[X_1]}$. Let Z denote the sum

$$Z := X_1 + X_2 + X_3 + \ldots + X_n.$$

Then, by the calculation rules we learned for expectation and variance it follows that:

$$E[Z] = n\mu$$

and the standard deviation σ_Z of Z is equal to:

$$\sigma_Z = \sqrt{n}\sigma.$$

When you subtract from a random variable its mean and divide by the standard deviation then you always get a new variable with zero expectation and variance equal to one. Thus the "standardized" sum:

$$\frac{Z - n\mu}{\sqrt{n}\sigma}$$

has expectation zero and standard deviation 1. The central limit theorem says that on top of this, for large n, the expression

$$\frac{Z - n\mu}{\sqrt{n}\sigma}$$

is close to being a standard normal variable. Let us now formulate the central limit theorem:

Theorem 10.1 *Let*

$$X_1, X_2, X_3, \ldots$$

be a sequence of independent, identically distributed random variables. Then we have that for large n, the normalized sum Y:

$$Y := \frac{X_1 + \ldots + X_n - n\mu}{\sigma\sqrt{n}}$$

is close to being a standard normal random variable.

The version of the Central Limit Theorem is not yet very precise. As a matter of fact, what means "close to being a standard normal random variable"? We certainly understand what means that two points are close to each other. But we have not yet discussed the concept of closeness for random variables. Let us do this by using the example of a six-sided die. Let us assume that we have a six-sided die which is not perfectly symmetric. For $i \in \{1, 2, \ldots, 6\}$, let p_i denote the probability of side i:

$$P(X = i) = p_i$$

where X denotes the number which we get when we through this die once. A perfectly symmetric die would have the probabilities p_i all equal to $1/6$. Say, our die is not exactly symmetric but close to a perfectly symmetric die. What does this mean? This means that for all $i \in \{1, 2, \ldots, 6\}$ we have that p_i is close to $1/6$.
For the die example with have a finite number of outcomes. For a continuous random

variable on the other hand we are interested in the probabilities of intervals. By this I means that we are interested for a given interval I, in the probabilities that the random experiment gives result in I. If X denotes our continuous random variable, this means that we are interested in the probabilities of type:

$$P(X \in I).$$

We are now ready to explain what we mean by: "two continuous random variables X and Y have there probability laws close to each other". By X and Y are close (have probability laws which are closed to each other) we mean: for each interval I we have that the real number $P(Y \in I)$ is close to the real number $P(X \in I)$. For the interval, $i = [i_1, i_2]$ with $i_1 < i_2$, we have that

$$P(X \in I) = P(X \leq i_2) - P(X < i_1).$$

It follows that if we know all the probabilities for semi-infinite intervals we can determine the probabilities of type $P(X \in I)$. Thus, for two continuous random variables X and Y to be close to each other (with respect to their probability law), it is enough to ask that for all $x \in \mathbb{R}$ we have that the real number $P(X \leq x)$ is close to the real number $P(Y \leq y)$.
Now that we have clarified the concept of closeness in distribution for continuous random variables, we are ready to formulate the CLT in a more precise way. Hence saying that

$$Z := \frac{X_1 + \ldots + X_n - n\mu}{\sigma\sqrt{n}}$$

is close to a standard normal random variable $\mathcal{N}(0,1)$ means that for every $z \in \mathbb{R}$ we have that:

$$P(Z \leq z)$$

is close to

$$P(\mathcal{N}(0,1) \leq z).$$

In other words, as n goes to infinity, $P(Z \leq z)$ converges to $P(\mathcal{N}(0,1) \leq z)$. Let us give a more precise version of the CLT then what we have done so far:

Theorem 10.2 *Let*

$$X_1, X_2, X_3, \ldots$$

be a sequence of independent, identically distributed random variables. Let $E[X_1] = \mu$ *and* $\sqrt{VAR[X_1]} = \sigma$. *Then, for any* $z \in \mathbb{Z}$, *we have that:*

$$\lim_{n\to\infty} P\left(\frac{X_1 + \ldots + X_n - n\mu}{\sigma\sqrt{n}} \leq z\right) = P(\mathcal{N}(0,1) \leq z).$$

11 Statistical testing

Let us first give an example:
Assume that you read in the newspaper that 50% of the population in Atlanta smokes. You don't believe that number, so you start a survey. You ask 100 randomly chosen people, and find that 70 out of the hundred smoke. Now, you want to know if the result of your survey constitutes strong evidence against the 50% claimed by the newspaper.
If the true percentage of the population of Atlanta which smokes would be 50%, you would expect to find in your survey a number closer to 50 people. However, it could be that although the true percentage is 50%, you still observe a figure as high as 70. Just by chance. So, the procedure is the following: determine the probability of getting 70 people or more in your survey who smoke, given that the percentage would really be 50%. If that probability is very small you decide to reject the idea that 50% of the population smoke in Atlanta. In general one takes a fix level $\alpha > 0$ and rejects the idea one wants to test if the probability is smaller than α. Most of the times statisticians work with α being equal to 0.05 or 0.1. So, if the probability of getting 70 people or more in our survey who smoke is smaller than $\alpha = 0.05$(the probability given that 50% of the population smokes), then statisticains will say: we reject the hypothesis that 50% of the population in Atlanta smokes. We do this on the confidence level $\alpha = 0.05$, based on the evidence of our survey.
How do we calculate the probability to observe 70 or more people in our survey who smoke if the percentage would really be 50% of the Atlanta population? For this it is important how we choose, the people for our survey. The correct way to choose them is the following: take a complete list of the inhabitants of Atlanta. Numerate them. Choose 100 of them with replacement and with equal probability. This means that a person could appear twice.
Let X_i be equal to one if the i-th person chosen is a smoker. Then, if we chose the people following the procedure above we find that the X_i's are i.i.d. and that $P(X_i = 1) = p$ where p designates the true percentage of people in Atlanta who smoke. Then also $E[X_i] = p$. The total number of people in our survey who smoke Z, can now be expressed as

$$Z := X_1 + X_2 + \ldots + X_{100}.$$

Let $P_{50\%}(.)$ designate the probability given that the true percentage which smoke is really 50%. Testing if 50% in Atlanta smoke can now be discribed as follows:

- Calculate the probability:

$$P_{50\%}(X_1 + \ldots + X_{100} \geq 70).$$

- If the above probability is smaller than $\alpha = 0.05$ we reject the hypothesis that 50% of the population smokes in Atlanta (we reject it on the $\alpha = 0.05$ level). Otherwise, we keep the hypothesis. When we keep the hypothesis, this means that the result of our survey does not constitute strong evidence against the hypothesis: the result of the survey does not "contradict" the hypothesis.

Note that we could also have done the test on the $\alpha = 0.1$ level. In that case we would reject the hypothesis if that probability is smaller that 0.1.
Next we are explaining how we can calculate approximately the probability $P_{50\%}(Z \geq 70)$, using the CLT. Simply note that, by basic algebra, the inequality

$$Z \geq 70$$

is equivalent to

$$Z - n\mu \geq 70 - n\mu$$

which is itself equivalent to:

$$\frac{Z - n\mu}{\sigma\sqrt{n}} \geq \frac{70 - n\mu}{\sigma\sqrt{n}}.$$

Equivalent inequalities must also have same probability. Hence:

$$P_{50\%}(Z \geq 70) = P_{50\%}(Z - n\mu \geq 70 - n\mu) = P_{50\%}\left(\frac{Z - n\mu}{\sigma\sqrt{n}} \geq \frac{70 - n\mu}{\sigma\sqrt{n}}\right) \tag{11.1}$$

By the CLT we have that

$$\frac{Z - n\mu}{\sigma\sqrt{n}}$$

is close to being a standard normal random variable $\mathcal{N}(0,1)$. Thus, the probability on the right side of inequality 11.1, is approximately equal to

$$P\left(\mathcal{N}(0,1) \geq \frac{70 - n\mu}{\sigma\sqrt{n}}\right). \tag{11.2}$$

If the probability in expression 11.2 is smaller than 0.05 then we reject the hypothesis that 50% of the Atlanta populations smokes. (on the $\alpha = 0.05$ level). We can look up the probability that the standard normal $\mathcal{N}(0,1)$ is smaller than the number $(70 - n\mu)/(\sigma\sqrt{n})$ in a table. We have tables, for the standard normal variable $\mathcal{N}(0,1)$.

11.1 Looking up probabilities for the standard normal in a table

Let $z \in \mathbb{R}$. Let $\phi(z)$ denote the probability that a standard normal variable is smaller equal than z. Thus:

$$\phi(z) := P(\mathcal{N}(0,1) \leq z) = \int_{-\infty}^{z} \frac{1}{2\pi} e^{-x^2/2} dx.$$

For example, let $z > 0$ be a number. Say wen want to find the probability

$$P(\mathcal{N}(0,1) \geq z). \tag{11.3}$$

The table for the standard normal gives the values of $\phi(z)$ for $z > 0$ thus we have to try to express probability 11.3 in terms of $\phi(z)$. For this note that:

$$P(\mathcal{N}(0,1) \geq z) = 1 - P(\mathcal{N}(0,1) < z).$$

Furthermore, $P(\mathcal{N}(0,1) < z)$ is equal to $P(\mathcal{N}(0,1) \leq z) = \phi(z)$. Thus we find that:

$$P(\mathcal{N}(0,1) \geq z) = 1 - \phi(z).$$

Let us next explain how, if $z < 0$, we can find the probability:

$$P(\mathcal{N}(0,1) \leq z).$$

Note that $\mathcal{N}(0,1)$ is symmetric around the origin. Thus,

$$P(\mathcal{N}(0,1) \leq z) = P(\mathcal{N}(0,1) \geq |z|).$$

This brings us back to the previously studied case. We find

$$P(\mathcal{N}(0,1) \leq z) = 1 - \phi(|z|).$$

Eventually let $z > 0$ again. What is the probability:

$$P(-z \leq \mathcal{N}(0,1) \leq z)$$

equal to? For this problem note that

$$P(-z \leq \mathcal{N}(0,1) \leq z) = 1 - P(\mathcal{N}(0,1) \geq z) - P(\mathcal{N}(0,1) \leq -z).$$

Thus, we find that:

$$P(-z \leq \mathcal{N}(0,1) \leq z) = 1 - (1 - \phi(z)) - (1 - \phi(z)) = 2\phi(z) - 1.$$

12 Statistical estimation

12.1 An example

Imagine that we want to measure the distance d between two points y and z. Every time we repeat the measurement we make a measurement error. In order to improve the precision we make several mesurements and then take the average value measured. Let X_i designate measurement number i and ϵ_i the error number i. We have that:

$$X_i = d + \epsilon_i.$$

We assume that the measurement errors are i.i.d. such that

$$E[\epsilon_i] = 0$$

and

$$Var[\epsilon_i] = \sigma^2.$$

The standard deviation σ of the measurement instrument is supposed to be know to us. Imagine that we make 4 measurements and find in meters the four values:

$$100, 102, 99, 101$$

We see that the distance d must be around 101 meters. However, the exact value of the distance d remains unknown to us, since each of the four measurements above contains an error. So, we can only estimate what the true distance is equal to. Typically we take the average of the measurements as estimate for d. We write $\hat{d}$ for our estimate of d. In the case we decide to use the average of our measurements as estimate for d, we have that:

$$\hat{d} = \frac{X_1 + X_2 + X_3 + X_4}{4}.$$

The advantage of taking four measurements of the same distance instead of only one, is that that probability to have a large error is reduced. The errors in the different measurements tend to even each other out when we compute the average. As a matter of fact, assume we make n measurements and then take the average. In this case:

$$\hat{d} := \frac{X_1 + \ldots + X_n}{n}.$$

We find:

$$E[\hat{d}] = (1/n)\left(E[X_1] + \ldots + E[X_n]\right) = \\ (1/n)\left(nE[X_1]\right) = E[X_1] = E[d + \epsilon_i] = E[d] + E[\epsilon_i] = d + 0 = d.$$

An estimator which has its expectation equal to the true value we want to estimate is called *unbiased* estimator.
Let us calculate:

$$VAR[\hat{d}] = VAR\left[\frac{X_1 + \ldots + X_n}{n}\right] = \frac{1}{n^2}\left(VAR[X_1] + \ldots + VAR[X_n]\right) = \\ \frac{1}{n^2}\left(nVAR[X_1]\right) = VAR[X_1]/n$$

Thus, the standart deviation of $\hat{d}$ is equal to

$$\sqrt{VAR[X_1]/n} = \sigma/\sqrt{n}.$$

The standard deviation of the average $\hat{d}$ is thus $\sqrt{n}$ times smaller than the standard deviation of the error when we make one measurement. This justifies taking several measurements and taking the average, since it reduces the size of a typical error by a factor $\sqrt{n}$.
When we make a measurement and give an estimate of what the distance is, it is important that when know the order of magnitude of the error. Imagine for example that the order of magnitude of the error is 100 meters. The situation would then be: " our estimate of

the distance is 101 meters, and the precision of this estimate is plus/ minus 100 meters". In this case our estimate our estimate of the distance is almost useless because of the hudge imprecision. This is why, we try to always give the "precision" of the estimate. Since the errors are random, theoretically even very large errors are always possible. Very large errors however have small probability. Hence one tries to be able to be able to give a upper bound on the size of the error which holds with a given probability. Typically one uses the probabilities 95% or 99%. The type of statement one whishes to make is for example: our estimate for the distance is 101 meters. Furthermore, with 95% probability the true distance is within 2 meters of our estimate. In this case the interval $[99, 103]$ is called the 95% *confidence interval* for d. With 95% probability, d should lie within this interval. More precisely, we look for a real number $a > 0$ such that:

$$P(\hat{d} - a \leq d \leq \hat{d} + a) = 95\%$$

or equivalently:

$$P(-a \leq \hat{d} - d \leq a) = 95\%$$

Hence we are looking for a number a such that:

$$95\% = P\left(-a \leq \frac{X_1 + \ldots + X_n}{n} - d \leq a\right) = P\left(-a \leq \frac{d + \epsilon_1 + \ldots + d + \epsilon_n - nd}{n} \leq a\right) =$$
$$= P\left(-a \leq \frac{\epsilon_1 + \ldots + \epsilon_n}{n} \leq a\right)$$

Now, either way we assume that the errors ϵ_I are normal or that n is big enough so that the sum $\epsilon_1 + \ldots + \epsilon_n$ is approximately normal due to the central limit theorem. Dividing the sum $\epsilon_1 + \ldots + \epsilon_n$ by $\sigma\sqrt{n}$, we get (approximately) a standard normal variable. This then gives:

$$95\% = P\left(-\frac{a\sqrt{n}}{\sigma} \leq \frac{\epsilon_1 + \ldots + \epsilon_n}{\sigma\sqrt{n}} \leq \frac{a\sqrt{n}}{\sigma}\right) \approx P\left(-\frac{a\sqrt{n}}{\sigma} \leq \mathcal{N}(0,1) \leq \frac{a\sqrt{n}}{\sigma}\right)$$

We thus find the number $b > 0$ from the table for standard normal random variable such that:

$$95\% = P(-b \leq \mathcal{N}(0,1) \leq b).$$

Hence:

$$95\% = \phi(b) - (1 - \phi(b)) = 2\phi(b) - 1$$

where $\phi(.)$ designates the distribution function of the standard normal variable. Then, we find $a > 0$ solving:

$$b = \frac{a\sqrt{n}}{\sigma}.$$

The confidence interval on the 95% level is then:

$$[\hat{d} - a, \hat{d} + a].$$

This means that although we don't know the exact value of d, we can say that with 95% probability d lies in the interval $[\hat{d} - a, \hat{d} + a]$.

12.2 Estimation of variance

Assume that we are in the same situation as in the previous subsection. The only difference is that instead of trying to determine the distance we want to find out how precise our measurement instrument is. In other words, we try to determine the standard deviation $\sigma = \sqrt{VAR[\epsilon_i]}$. For this we make several measurements of the distance between to points y and z. We choose the point so that we know the distance d between them. Again, if X_i designates the i-th measurement we have $X_i = d + \epsilon_i$. Define the random variable Z_i in the following way:

$$Z_i := (X_i - d)^2 = \epsilon_i^2.$$

Thus:

$$E[Z_i] = VAR[\epsilon_i].$$

We have argued that if we have a number of independent copies of the same random variables, a good way to estimate the expectation is to take the average. Thus to estimate the expectation $E[Z_i]$, we take the average:

$$\hat{E[Z_i]} := \frac{Z_1 + \ldots + Z_n}{n}.$$

In other words, as an estimate for $VAR[\epsilon_i] = E[Z_i] = \sigma^2$, we take:

$$\frac{Z_1 + \ldots + Z_n}{n} = \frac{(X_1 - d)^2 + \ldots + (X_n - d)^2}{n}.$$

The estimate for σ is then simply the square root of the estimate for the variance. Thus, our estimator for $\sigma = \sqrt{VAR[\epsilon_i]}$ is:

$$\hat{\sigma} = \sqrt{\frac{(X_1 - d)^2 + \ldots + (X_n - d)^2}{n}}.$$

If the distance d, should not be known, we simply take and estimate for d instead of d. In that case our estimate for σ is

$$\hat{\sigma} = \sqrt{\frac{(X_1 - \hat{d})^2 + \ldots + (X_n - \hat{d})^2}{n-1}}$$

where

$$\hat{d} := \frac{X_1 + \ldots + X_n}{n}.$$

(Note that instead of dividing by n in the case that d in unknown, we divide usually by $n-1$. This is a little detail which I am not going to explain. For large d, it is not important since then $n/(n-1)$ is close to 1.)

12.3 Maximum Likelihood estimation

Imagine the following situation: we have two 6-sided dice. Let X designate the number we obtain when we throw the first die. Let Y designate the number we obtain when we throw the second one. Assume that the first die is regular whilst the second is skewed. We have:

$$(P(X=1), P(X=2), \ldots, P(X=6)) = (1/6, 1/6, 1/6, 1/6, 1/6, 1/6).$$

(Note that $1/6 = 0.1\bar{6}$) Assume furthermore that:

$$(P(Y=1), \ldots, P(Y=6)) = (0.01, 0.3, 0.2, 0.1, 0.1, 0.29).$$

Imagine that we are playing the following game: I choose from a bag one of the two dice. Then I throw it and get a number between 1 and 6. I don't tell you which die I used, but I tell you the number obtained. You have to guess which die I used based on the number which I tell you. (This guessing is what statisticians call estimating.) For example I tell you that obtained the number 1. With the first die, the probability to obtain a 1 is $0.1\bar{6}$, whilst with the second die it is 0.01. The probability to obtain a 1 is thus much smaller with the second die. Having obtained a one makes us thus think "that it is likelier" that the die used is the first die. Our guess will thus be the first die. Of course you could be wrong, but based on what you know the first die appears to be "likelier".
If on the other hand, after throwing the die we obtain a 2 we guess that it was the second die which got used. The reason is that with the second die a 2 has a probability of 0.3 which is larger than the probability to see a 2 with the first die. Again, our guess might be wrong, but when we observe a 2, the second die seem "likelier". The method of guessing described here is called *Maximum likelihood estimation*. It consist in guessing (estimating) the possibility which makes the observed result most likely. In other words, we choose the possibility, for which the probability of the observed outcome is highest.
Let us look at is in a slightly more abstract way. Let I designate the first die and II the second. For $x = 1, 2, \ldots, 6$, let $P(x, I)$ designate the probability that the number we obtain by throwing the first die equals to x. Thus:

$$P(x, I) := P(X = x).$$

Let $P(x, II)$ designate the probability that the number we obtain by throwing the second die equals to x. Thus:

$$P(x, II) := P(Y = x).$$

For example, $P(1, I)$ is the probability that the first die gives a 1 and $P(1, II)$ is the probability that the second die equals 1 whilst $P(2, II)$ designates the probability that the second die gives a 2.
Let θ be a (non-random) variable with can take one out of two values: I or II. Statisticians call θ the parameter. In this example guessing which die we are using, is the same as trying to figure out if θ equals I or II. We consider the probability function $P(.,.)$ with two entries:

$$(x, \theta) \mapsto P(x, \theta).$$

Formally what we did can be describe as follows: given that we observe an outcome x, we take the θ wich maximizes $P(x, \theta)$ as our guess for which die was used. . Our *maximum likelihood estimate* $\hat{\theta}$ of θ is the theta maximizing $P(x, \theta)$ where x is the observed outcome. This is a general method, and can be used in many different settings. Let us give another example of maximum likelihood estimation, based on the same principle.

12.4 Estimation of parameter for geometric random variables

Let $T_1, T_2, \ldots$ be a sequence of i.i.d. geometric random variables with parameter $p > 0$. Assume that $p > 0$ is unknown. We want to estimate p (in other words we want to try to guess what p is approximately equal to). Say we observe:

$$(T_1, T_2, T_3, T_4, T_5) = (6, 7, 5, 8, 8)$$

Based on this evidence, what should our estimate $\hat{p}$ for p be? (Hence what should our guess for the unknown p be?) We can use the Maximum Likelihood method. For this the estimate $\hat{p}$ is the $p \in [0, 1]$ for which the probability to observe

$$(6, 7, 5, 8, 8)$$

is maximal. Since we assumed the T_i's to be independent we find that:

$$P\left(\ (T_1, T_2, T_3, T_4, T_5) = (6, 7, 5, 8, 8)\ \right) \tag{12.1}$$

is equal to

$$P(T_1 = 6) \cdot P(T_2 = 7) \cdot \ldots \cdot P(T_5 = 8)$$

For a geometric random variable T with parameter p we have that:

$$P\left(T = k\right) = p(1-p)^{k-1}.$$

Thus the probability 12.1 is equal to:

$$p(1-p)^5 \cdot p(1-p)^6 \cdot \ldots \cdot p(1-p)^7 = \exp(\ln(p) + 5\ln(1-p) + \ldots + \ln(p) + 7\ln(1-p)). \tag{12.2}$$

We want to find p maximizing the last expression. This is the same as maximizing the expression:

$$\ln(p) + 5\ln(1-p) + \ldots + \ln(p) + 7\ln(1-p),$$

since exp(.) is an increasing function. To find the maximum, we take the derivative according to p and set it equal to 0. This gives:

$$0 = \frac{d\left(\ln(p) + 5\ln(1-p) + \ldots + \ln(p) + 7\ln(1-p)\right)}{dp} = \frac{1}{p} - \frac{5}{1-p} + \ldots + \frac{1}{p} - \frac{7}{1-p}.$$

The last equality leads to:

$$n(1-p) = (5 + \ldots + 7)p$$

where n designates the number or observations. (In the special example considered here $n = 5$.) We find:

$$n = (6 + \ldots + 8)p = p(T_1 + T_2 + \ldots + T_n)$$

and hence:

$$\frac{1}{p} = \frac{6 + 7 + 5 + 8 + 8}{5} = \frac{T_1 + T_2 + \ldots + T_n}{n} \tag{12.3}$$

Our estimate $\hat{p}$ of p is the p which maximizes expression 12.2. This is the p which satisfies equation 12.3. Thus our estimate:

$$\hat{p} := \left(\frac{6 + 7 + 5 + 8 + 8}{5}\right)^{-1} = \left(\frac{T_1 + T_2 + \ldots + T_n}{n}\right)^{-1}.$$