Notes on probability

Student version

Last saved: 10/23/2026 11:10:00 AM

Index

1 General Information 4
2 Probability Theory 5
2.1 Sample spaces and events 5
2.2 Axioms of probability 8
2.3 Sample spaces having equally likely outcomes (uniform probability model) 9
2.3.1 Basic principle of counting 10
2.3.2 Permutations and subsets of known cardinality 12
2.4 Conditional probability 17
2.4.1 Total probability theorem and Bayes' formula 19
2.4.2 Independent events 23
2.4.3 Repeated trials 24
2.4.4 Parallel systems 28
3 Random variables 32
3.1 Cumulative Distribution Function of a random variable 34
3.2 Probability Mass Function of discrete RVs 36
3.3 Probability Density Function for continuous RVs 37
3.3.1 Exercises 39
3.4 Jointly distributed random variables 42
3.4.1 Joint PMF for discrete RV 42
3.4.2 Joint PDF for continuous RVs 43
3.4.3 Joint distributions of *n* random variables 46
3.5 Independent random variables 46
3.6 Mean value 49
3.6.1 Expectation of the sum of RVs 53
3.7 Variance 55
3.7.1 Variance of sums of RVs 58
3.7.2 Covariance and correlation 59
4 Special random variables 61
4.1 Discrete distributions 61
4.1.1 Bernoulli distribution 61
4.1.2 Binomial distribution 61
4.1.3 Poisson distribution 62

4.1.4 Geometric distribution 66
4.1.5 Probability-generating Functions 68
4.2 Continuous distributions 70
4.2.1 Uniform distribution 71
4.2.2 Exponential distribution 71
4.2.3 Laplace-Stieltjes Transform 75
4.2.4 Normal distribution 76
4.2.5 Central limit theorem 83
4.2.6 Percentiles 86
4.2.7 Chi-square distribution 87
4.2.8 Student's T distribution 87
4.3 Heavy-tailed distributions 89
4.3.1 Heavy-tail in action: load balancing in an Internet ASs 93
5 Review problems 95
5.1 Problem 1 – Roulette 95
5.1.1 Solution 95
5.2 Problem 2 – Voting 99
5.2.1 Solution 99
5.3 Problem 3 – Student tests 100
5.3.1 Solution 100
5.4 Problem 4 – Independent measures 102
5.4.1 Solution 103
5.5 Problem 5 – Switches 103
5.5.1 Solution 104
5.6 Problem 6 - Dice 105
5.6.1 Solution 105
5.7 Problem 7 – JPMF from JCDF 106
5.7.1 Solution 106
6 Appendix 108
6.1 Tables 108
6.2 Geometric RVs: capacity of data-link protocols 110

1 General Information

Prof. Ing. Giovanni Stea

Dipartimento di Ingegneria dell'Informazione

Largo L. Lazzarino 1, 56122 Pisa - Italy

Ph. : (+39) 050-2217.653 (direct) .599 (switch)

Fax : (+39) 050-2217.600

E-mail: giovanni.stea@unipi.it

Course book: whatever text on probability theory will probably be suitable. Most queuing theory books also have one or two chapters about probability theory, and they are often enough for a recap on the theory.

These notes are based on:

S. M. Ross, "Introduction to probability and statistics for engineers and scientists", Elsevier, cap. 2-6

Pre-requisites: algebra (factorials, permutations) and mathematic analysis (integrals, derivatives)

Module length: about 20 h, including exercises.

2 Probability Theory

2.1 Sample spaces and events

The concept of **probability** can be explained in several ways. The simplest way for engineers to visualize is that of **relative frequency**. If you repeat an experiment a very large number of times N, in independent conditions (meaning that the outcome of an experiment does not influence the outcome of the subsequent ones), and call k the number of times when a certain **event** E of interest is observed, you can define the probability of that event as:

$$P(E) = \lim_{N \to \infty} k/N$$

In practice, you never have time to perform an *infinite* number of experiments. Thus, the probability of an event is often determined by some other **knowledge**, often **a priori**, of the experiment we observe. For instance, symmetry reasons. We know that the probability that the **event "heads"** occurs in a coin flip is 50% if the coin is perfectly symmetric (*fair*). We can give it for granted, without repeating the experiment N times. If we did, we would just observe what we already supposed for a large N.

I have introduced some concepts without defining them formally.

We define a **random experiment** as one whose outcome is not predictable a priori. For instance:

1) The throw of a six-faced die
2) A horse race
3) An integrity test for an electronic device

In the first case, **I can define** the **outcome** (or **result**) as the number which is engraved on the face opposite the one on which the die rests. In the second case, the order of arrival for the horses. In the third case, the time at which the device stops working.

The concept of outcome lies in the mind of the observer. You can observe many things about an experiment. For instance, about the first one, I might be interested in the spatial position at which the die comes to rest, which is an entirely different thing. In the second experiment, I might be interested in the **inter-arrival time** of the horses.

Once you define what an outcome is, you can define the **sample space *S*.** This is, in fact, the set of all possible outcomes of the experiment. In the three above cases, we have:

1. $S = \{1,2,3,4,5,6\}$
2. Assuming we have seven horses, numbered in some order, $S = \{\text{all the permutations of } \{1,2,3,4,5,6,7\}\}$
3. $S = [0, +\infty)$

In the first two cases, the sample space has a finite number of elements. In the third case, it has an infinite number of elements.

We define an **event *E* as a subset of the sample space.**

1. $E = \{1,2,5\}$
2. $E = \{x \in S | \text{horse no. 5 arrived last}\}$
3. The device breaks between 100 and 200 hours

We say that an **event *E* has occurred** if the outcome of the experiment is included in *E*. For instance, if the order of arrival of the horses is 1, 2, 3, 4, 5, 6, 7, event *E* has not occurred.

Among the events, two peculiar ones are:

- The **null event**, which is represented by the **empty set**
- The **certain event**, which is represented by the sample space *S*

Events are **sets**. Hence, you can apply set algebra to events, using set algebra operators:

- Union ∪: $E \cup F$, a set which includes the outcomes that are in **either or both** *E* and *F*
- Intersection ∩: $E \cap F = EF$, a set that includes the outcomes that are **both** in *E* and *F*
- Complement: $E^c = S \backslash E$, the set of outcomes which are not in *E*.

Furthermore, you can use all the properties that you already know for set algebra (they are the same as for Boole algebra, with slightly different names)

- Union and intersection are **associative and commutative** operations
- Complement is **involutive**.
- De Morgan's laws:

$$\begin{cases} (A \cup B)^c = A^c \cap B^c \\ (A \cap B)^c = A^c \cup B^c \end{cases}$$

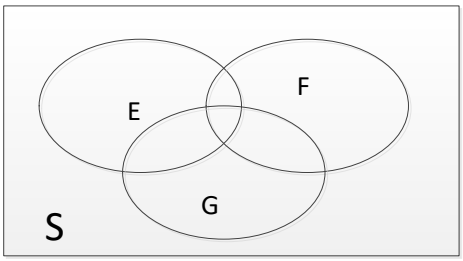
De Morgan's laws hold for an arbitrary number of sets, not just two. They can be readily proven using Venn's diagrams.

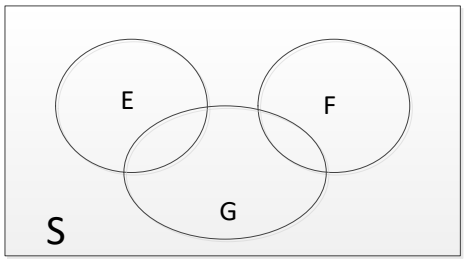
You should pay attention **not to confuse between an *event* and an *outcome***. An outcome is an element of the sample space. An event is a *subset* of the sample space. For **discrete sample spaces**, we obviously have subsets of just one element. Therefore you can associate an event to each outcome, i.e. a subset including only that outcome.

However, you should keep the two concepts separate.

Exercise

Let *E, F, G* be three arbitrary events. Find set-algebra expressions for the following occurrences. Two possible definitions of the sample space and events are the following:





A word of advice: the likeliest mistake in probability theory is *to misunderstand the text of an exercise.* Please take care when reading, because <u>every word matters</u>.

a) Only *E* occurs
   a. The outcome is in *E*, *but not in F and not in G*. Thus, the answer is EF^cG^c
b) Both *E* and *G*, but not *F*, occur
   a. EGF^c
c) At least one of the events occurs
   a. $E \cup F \cup G$
d) At least two of the events occur
   a. $EG \cup FG \cup EF$
e) All three occur
   a. EFG
f) None of the events occur
   a. It's the complement of c). Hence $(E \cup F \cup G)^c = E^cF^cG^c$
g) At most one of them occurs
   a. It's the complement of d) . Hence

   $(EF \cup FG \cup EG)^c =$
   $(EF)^c(FG)^c(EG)^c =$
   $(E^c \cup F^c)(F^c \cup G^c)(E^c \cup G^c)$

> Good hint on how to solve Probability Theory problems: when an event looks difficult to pinpoint: **try the complement**.

h) Exactly two of them occur
   a. From d), you have to exclude that all three occurs. Hence,

      $(EF \cup FG \cup EG) \cdot (EFG)^c$.

      or, you can exclude the third set manually from each intersection, i.e.:

      $EFG^c \cup FGE^c \cup EGF^c$. The two results equivalent, after some manipulations.
i) At most three of them occur
   a. This is certain, since there is no way more than three events out of three can occur.

2.2 Axioms of probability

Given the above definitions, you can lay down the **three axioms of probability**. Probability is **a number associated to an event**, which describes the relative frequency of that event. Thus:

1. $0 \leq P(E) \leq 1$.
2. $P(S) = 1$.
3. If E_i, E_j are events such that $E_i E_j = \emptyset$ if $i \neq j$ (**mutually exclusive events,** or **disjoint**), then $P(\bigcup_i E_i) = \sum_i P(E_i)$.

- The first axiom states that the probability is actually a relative frequency, hence it is a number between zero and one.
- The second axiom states that the sample space is the certain event, meaning that some outcome must occur when you run an experiment.
- The third axiom states that the probability of the union of **disjoint** events is the sum of the probabilities of the single events.

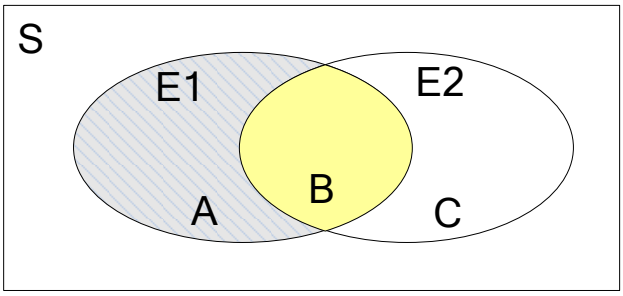
For instance, if you throw a die, the probability that event $E_1 = \{\text{outcomeisanevenno.}\}$ occurs is 1/2, and the probability that event $E_2 = \{1,5\}$ occurs is equal to 2/6. Given that the two events are disjoint (or mutually exclusive), then the probability that $E_1 \cup E_2$ occurs is equal to $P(E_1 \cup E_2) = P(E_1) + P(E_2) = 1/2 + 1/3 = 5/6$.

From the above axioms some useful properties can be derived, which I expect you to be able to apply from now on:

1. $P(E^c) = 1 - P(E)$. This is fairly obvious, given that the two events E, E^c are mutually exclusive (axiom 3) and their union is equal to the sample space (axiom 2).

2. $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 E_2)$

This can be readily proved with Venn diagrams.



$$
\begin{aligned}
&P(E_1 \cup E_2) = P(A \cup B \cup C) \\
&= P(A) + P(B) + P(C) \\
&= [P(A) + P(B)] + [P(B) + P(C)] - P(B) \\
&= P(A \cup B) + P(B \cup C) - P(B) \\
&= P(E_1) + P(E_2) - P(E_1 E_2)
\end{aligned}
$$

A, B, C are disjoint, hence I can sum their probabilities by the 3rd axiom

Add and subtract *P(B)*

Exercise

28% of the American males smoke cigarettes. 7% smoke cigars. 5% smoke both. What is the percentage of non-smokers?

Solution

NB: instead of trying to "guess" the correct answer, try **modeling the problem in terms of events**, and then using set algebra to get the answer. There will soon come a time where you won't be able to guess.

Call *E* the event "smokes cigarettes" and *F* the event "smokes cigars".

$\boldsymbol{P(E) = 0.28, P(F) = 0.07, P(EF) = 0.05}$

The required quantity is $P((E \cup F)^c)$

$$
\begin{aligned}
&\boldsymbol{P((E \cup F)^c) = 1 - P(E \cup F)} \\
&\boldsymbol{= 1 - \quad [P(E) + P(F) - P(EF)]} \\
&\boldsymbol{= 1 - 0.28 - 0.07 + 0.05 = 0.7}
\end{aligned}
$$

Thus, 70% of the Americans are non-smokers.

♦

2.3 Sample spaces having equally likely outcomes (uniform probability model)

In some cases (many, actually), the sample space of a random experiment:

a) Has a finite cardinality $N = |S|$,
b) includes ***equally likely*** **outcomes.**

For instance, the case of a dice throw or a coin flip, provided that they are *fair*. In this case, the probability of each outcome can only be equal to $p = 1/N$.

To be precise, each of the *N* outcomes is included in an event E_i, $1 \leq i \leq N$, these events are mutually exclusive, and their union is the sample space *S*, hence it must be that the sum of their probabilities is $N \cdot p = 1$.

In this case, we are in a **uniform probability model**, and we talk about **sample space with equally likely outcomes.** In these cases (and in **these only)**, the probability of an event *E* is:

$$P(E) = \frac{|E|}{N}$$

We already applied this property in the case of a dice throw, without defining it.

For random experiments with equally likely outcomes, then, in order to define the probability of an event, it is enough to **count the number of outcomes** included in the event, and divide it by the cardinality of the sample space. "It is enough" does not mean that it is easy. We have to reflect on **how to count**, and to do this we introduce the **basic principle of counting.**

2.3.1 Basic principle of counting

Basic principle of counting

Given an experiment C that is composed of two sub-experiments C_1 and C_2,having respectively N_1 and N_2 possible outcomes, the number of possible outcomes of experiment C is equal to $N_1 \cdot N_2$.

This can be obviously generalized from 2 to *k* experiments, for any *k*.

Given an experiment C that is composed of k *sub-experiments $C_1 \dots C_k$,each one having $N_1 \dots N_k$ possible outcomes, the number of possible outcomes of experiment C is equal to $\prod_{i=1}^{k} N_i$.*

Example:

Take an opaque urn with 6 black balls and 5 white balls. What is the probability that, extracting two **at random** (<u>without replacing them</u>), you get **a black and a white one** (whatever the order)?

The first question that one should ask is "what is an *outcome* for this experiment?" In this case, it is a *pair of balls*, of whichever color. Call $b_1, \dots b_{11}$ the balls, and assume that balls 1 to 6 are black, and balls 7 to 11 white. The sample space is:

$$S = \{(b_i, b_j) | 1 \leq i, j \leq 11, \quad i \neq j\}$$

What is its cardinality? It can be computed by observing that the random experiment is indeed composed of two subexperiments:

- A first one, that consists in extracting a ball at random from a set of 11
- A second one, that consists in extracting a ball at random from a set of 10

By the principle of counting, we have $11 \cdot 10 = 110$ possible outcomes. Each one of these is obviously **equally likely**, since nothing allows us to distinguish the spheres when we extract them (the meaning of "**at random**" is exactly this). Thus, we are in a UPM.

We need to define the event that interests us, and compute the number of outcomes that belong to that event. Our event is:

$$E = \{(b_i, b_j), 1 \leq i \leq 6, 7 \leq j \leq 11\} \cup \{(b_i, b_j), 7 \leq i \leq 11, 1 \leq j \leq 6\}$$

Recall that **order does not matter**.

Each subset has $6 \cdot 5 = 30$ elements. Hence, $|E| = 60$, and $P(E) = 6/11$. The last passage is licit since results are equally likely.

♦

Note: if we allowed for **replacement** of a ball after the first extraction, I would have $E' = E$ (the event that interests me has the same cardinality), but $N' = 11 \cdot 11 > N$. In this case, we would have $P(E') = 60/121$.

Exercise:

You can line up on a shelf 10 books: 4 are mathematics books, 3 are physics books, 2 are informatics books, and 1 is a chemistry book. Suppose you line them up *at random*. What is the probability that they end up sorted by subject?

Solution

For this experiment, the outcome is **a permutation of ten books.** We are again in a case of experiment having equally likely outcomes. The sample space is the set of all possible orderings of 10 books. Its cardinality can be computed by the basic principle of counting. There are 10 subexperiments, and the first one has 10 outcomes, the 2nd has 9 outcomes, etc. etc. Therefore, $N = 10 \cdot 9 \cdot \ldots \cdot 2 \cdot 1 = 10!$.

The expression $x!$ is called **"x factorial"**, and it is the product of all numbers from 1 up to and including x. That number counts all the **permutations** (sequences) of x objects. For convenience, it is $0! = 1$.

Now we have the denominator in the fraction. In order to get the numerator as well, we need to count the outcomes in the event whose probability we want to evaluate.

Let us start by considering **one particular** ordering of the subjects (e.g., MPIC). How many ways are there to sort books according to that subject order?

The answer is: all the possible permutations of 4 maths books, which are 4!, times all the permutations of physics books, which are 3!, etc.

Thus, for **that** ordering of subjects, we have $G = 4! \cdot 3! \cdot 2! \cdot 1!$ different orderings of books that belong to the event.

Now, how many ways are there to sort 4 subjects? There are 4! (again by the basic principle of counting). Therefore, $|E| = G \cdot 4!$, and:

$$P(E) = |E|/N = \frac{4! \cdot 4!\, 3!\, 2!}{10!} = \frac{1}{525}$$

♦

2.3.2 Permutations and subsets of known cardinality

In the previous exercise we have reasoned about counting the *permutations* of *n* objects. The basic principle of counting can also be used to solve the following problem:

How many *different permutations* of *k* elements can I extract from a set of *n* elements ($n \geq k$)?

- For the first element, there are *n* possible choices.
- For the 2nd element, there are *n*-1.
- …
- For the *k*th element, there are *n*-*k*+1=*n*-(*k*-1)

This is a composite experiment, with *k* subexperiments, etc.

Thus, I can select *k* elements in $n \cdot (n-1) \cdot \ldots \cdot (n-k+1) = n!/(n-k)! = S_{k,n}$ possible modes. This is the number of **permutations** of *k* elements. Observe that, if *k*=*n*, you obtain the same result as the previous exercise, since $0! = 1$, and there are *n!* permutations of *n* objects.

Let us now answer the following question, which is related to the former:

How many *different subsets* of *k* elements can I extract from a set of *n* elements ($n \geq k$)?

First of all: are the two questions any different? Yes, they are. In a permutation, the order of the elements matters. In a subset, it does not. Two different permutations of the same *k* elements are indeed the same subset. Therefore, the answer must be different as well.

The number $S_{k,n}$ that answer the *other* question is a good starting point in any case. I must keep into account the fact that **in a subset the order of the elements does not matter.** Hence, all the sequences that have the same elements permuted in a different order are the same subset. This means that expression $S_{k,n}$ counts the same subset several times, hence I have to **divide it** by some other number in order to get the result.

Assume that you want to count the number of 3-letter subsets of the alphabet, i.e. *n*=26, *k*=3. Using the above approach, the *same* subset $\{ABC\}$ can be obtained from the following permutations (mind the brackets: curly ones for sets, angular brackets for permutations):

$$\langle ABC \rangle, \langle ACB \rangle, \langle BAC \rangle, \langle BCA \rangle, \langle CAB \rangle, \langle CBA \rangle$$

The number of permutations (counted in $S_{k,n}$) that yield the same subset is itself the number of **permutations of *k* elements**, i.e., $k \cdot (k-1) \cdot \ldots \cdot 2 \cdot 1 = k!$, again by the basic principle of counting. Therefore, the answer to the last question is:

$$\frac{S_{k,n}}{k!} = \frac{n!}{k! \cdot (n-k)!} = \binom{n}{k}$$

The last expression is so important that it deserves a name of its own: it is called **binomial coefficient**, since it appears within Newton's binomial formula:

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} \cdot a^k \cdot b^{(n-k)}$$

Let us briefly recall some useful **properties of binomial coefficients,** that we will use in the following:

1. $\binom{n}{k} = 0 \quad \text{if } n < k$ (this is a definition, not a property)
2. $\binom{n}{k} = \binom{n}{n-k}, \quad n \geq k$ (obvious, given the definition). It also has an intuitive rationale: whenever you find a way to extract *k* elements from a set of *n*, you are also defining a way to extract the remaining *n-k.* Hence the two numbers must be equal .
3. $\binom{n}{0} = \binom{n}{n} = 1$ (it follows from the definition and from $0! = 1$)
4. $\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}, \quad n \geq k$ (recursive method for computing factorials; prove it as an exercise).

Exercise

A group of 5 boys and 10 girls are lined up in random order

a) What is the probability that the person in the 4th position is a boy?
b) What about the person in the 12th position?
c) What is the probability that Adam (a boy) is in the 3rd position?

Solution

What is the *outcome*? It's a **sequence of 15 people**. Are we in a UPM? Yes, since no sequence is more likely than another. So, we just need to count i) all the sequences and ii) all the ones that we like, in order to get the result.

First of all, we observe that, if the ordering is random, there is nothing special about the 4th or the 12th positions, so the answer to questions a) and b) must be the same.

There is a quick, intuitive answer, which can be found as follows: the probability must be the same as the one of selecting one person at random and finding a boy, which is 5/15.

Let's do things by the book, in any case, and **confirm the intuition** by the correct method (**never** trust intuition when you do probability exercises).

We are in a uniform probability model (all permutations are equally likely). The number of possible permutations of 15 people is 15!, which is the denominator of the fraction that we want to express.

For questions a) and b), the cardinality of the event that we are looking for can be computed as follows:

- There are two sub-experiments. The first one is "choose one boy to put in the 4th (12th) position from a set of 5 boys. The second one is "choose any possible permutation of the remaining 14 people".
- how many subsets of 1 boy can you extract from a set of 5 boys? The answer is $\binom{5}{1} = 5$.
- The number of permutations of 14 people (15 minus one boy) is 14!.
- Therefore, $|E| = 5 \cdot 14!$, E being the event "a boy is in the 4th (12th position)"

In a uniform probability model, the probability of E is therefore:

$$P(E) = \frac{|E|}{N} = \frac{5 \cdot 14!}{15!} = \frac{5}{15}$$

Again, there is no difference between two positions, hence the answer to a) and b) is the same.

As for c), the question is slightly different. We don't want just *any* boy in the 3rd position. Give the boys names, e.g. Adam, Bob, Charlie, Dan, Eric. We want (say) the probability that Adam is in a given position, the 3rd. Now, intuition says that:

i) Adam must be somewhere;

ii) there is nothing peculiar about the 3rd position, which would make it more/less likely for Adam to be there instead of anywhere else.

Thus, the answer can only be $P(F) = 1/15$. Obviously, it is the same for *any* boy and *any* position. Furthermore, it does *not* depend on the number of boys and girls, but only on the number of positions. If the boys were 13 and the girls 2, it would still be the same.

The (slightly longer) way to get to the same result is to observe that there are 15! permutations, 14! of which have Adam in the 3rd position, so the answer is $P(F) = 14!/15\,! = 1/15$.

♦

Exercise

Let's make a variation on the above exercise. Suppose that you have n boys/girls (sex does not matter here), and you want to make a sorted line of **only k** of them (with $k \leq n$), picking them at random from the set of n. Answer the following questions:

a) what is the probability that Adam is in the 1st position?

b) Is it the same probability that any particular boy is in any position (up to the kth)?

c) What is the probability that a particular boy ends up being part of the line?

Solution

a) there are $S_{k,n} = n!/(n-k)!$ sequences of k boys from a set of n. If you put Adam in the first position, then you have $S_{k-1,n-1} = (n-1)!/[(n-1)-(k-1)]! = (n-1)!/(n-k)!$ sequences of k-1 boys to be extracted from the remaining n-1 boys. Then, the answer is:

$$P(E) = \frac{S_{k-1,n-1}}{S_{k,n}} = \frac{(n-1)!/(n-k)!}{n!/(n-k)!} = \frac{(n-1)!}{n!} = \frac{1}{n}$$

This makes perfect sense intuitively, since Adam has the same chance as everyone else to be picked up as the first liner. Note that the probability **does not depend on k**, which makes perfect sense as well: the fact that Adam is picked up as the first in the line cannot depend on how long the line is.

b) of course it is the same. You just repeat the same argument, and you get to the same solution. The lines where Adam is (say) in the 2nd positions are the same number as those where he is in the 1st position. The same goes for any other boy.

c) Intuitively, this answer must depend on **both *k and n*.** For instance, if k=n, then we are certain that Adam will be somewhere along the line. If, instead k<n, he can be left out. Moreover, the closer k gets to n, the higher his chances will be.

To answer the question, consider that there are $S_{k,n} = n!/(n-k)!$ sequences of k boys from a set of n. The sequences where Adam is in the first position are $S_{k-1,n-1}$. So are those where Adam is in the *second, third, … k^{th}* position. These sequences are mutually exclusive (Adam cannot be in two positions at the same time). Hence, the answer is:

$$\frac{k \cdot S_{k-1,n-1}}{S_{k,n}} = \frac{k}{n}$$

The result makes sense, since it is coherent with our early intuition. Moreover, if k=1, it says that Adam has a 1/n chance of being selected, which is true. If $k = n - 1$, then we observe that Adam has a $1/n$ chance *not* to be selected, which is also true.

♦

Exercise

A basketball team is composed by 6 black and 6 white players. They have to sleep in a hotel for an away match, and they are paired *at random* in double rooms. What is the probability that in *every* room you only get players of the same color?

Solution

This is an exercise where intuition does not get us very far, hence we must go by the book.

What is the *outcome* of this experiment? It is **a <u>set</u> of six <u>subsets</u> of two elements**, e.g. $\{\{A,B\},\{C,D\}\ldots,\{K,L\}\}$, which means that A and B are in the same room (whichever the room number), since we **don't care about room numbers**. Furthermore, (A,B) and (B,A) are the same pair, i.e. the elements of a pair are **unordered** (hence we have *sets* and not *sequences* of 2 elements). The sample space is the set of all those outcomes, which are sets themselves.

We are in a uniform probability model, since people are paired at random. Hence the trick is to find:

a) how large the sample space is;

b) how many outcomes belong to the event whose probability I want to compute.

Let's get through point a). This is a composite experiment, with 6 subexperiments. There are $\binom{12}{2}$ ways to choose the first couple, $\binom{10}{2}$ ways to choose the 2nd, etc. The number of ways to select the six couples is:

$$\binom{12}{2}\cdot\binom{10}{2}\cdot\binom{8}{2}\cdot\binom{6}{2}\cdot\binom{4}{2}=\frac{12!}{2^6}$$

However, this is *not* the number of outcomes (once we have defined the outcomes as above). In fact, if we multiply the above numbers, we are implicitly saying that **order matters** in the room allocation: for each outcome where AB are in room 1 and CD are in room 2, there will be another one where CD and AB are swapped. Therefore, we must divide the above number for the number of different **permutations of six rooms**, i.e. by 6!. The correct number is:

$$|S|=\frac{12!}{2^6\cdot 6!}$$

Now, we need to count the outcomes within the event that we are interested in. These are all the possible ways to put **6 white players in 3 rooms, and 6 black players in 3 rooms.** These are exactly the same problems as above, with different numbers, hence the solution can be written readily:

$$C = \frac{6!}{2^3 \cdot 3!}$$

Therefore, the result is:

$$P(E) = \frac{C^2}{N} = \frac{5}{231}$$

♦

This can be seen as a composite experiment as well. The two subexperiments are:

1) Arrange six black players in 3 pairs
2) Arrange six white players in 3 pairs

And each subex. has C outcomes.

2.4 Conditional probability

In many cases it is useful to compute the probability of an event *E* **knowing that** some other event *F* has occurred. In this case, we talk about **conditional probability.**

$P(E|F)$ (probability of *E given F,* or *conditioned to F*)

Conditional probability allows you to **re-evaluate** the probability that an event *E* occurs, given that you have more information (i.e., you know that some other event *F* that may influence *E* has occurred).

For instance, when you throw **two dice**, the probability that event E "the sum of the two upward faces is larger than 9" is equal to:

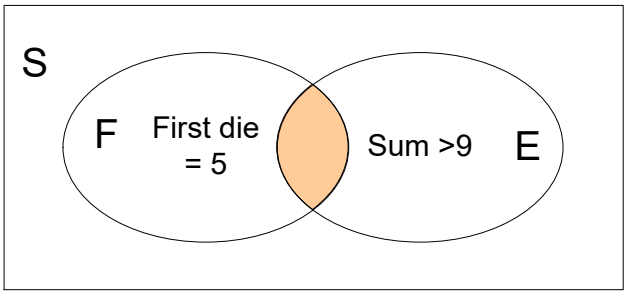
$$P(E) = \frac{|E|}{N} = \frac{|\{(5,5),(5,6),(6,5),(6,6),(6,4),(4,6)\}|}{|\{(x,y)|1 \le x \le 6, 1 \le y \le 6\}|} = \frac{6}{36} = \frac{1}{6}$$

Which was obtained by applying the basic principle of counting, given that each of the 36 results *(x,y)* is equally likely.

Suppose now that you know that **the first die is equal to 2**.

- What is the probability that the sum of the two dice is larger than 9, **given that <u>the first one</u> is 2?** The answer is obviously **zero**, seeing as there is no way that we can obtain more than 9 with a six-faced die. We will justify this more rigorously in a minute.
- What is the probability that the sum of the two dice is larger than 9, **given that <u>the first one</u> is 5?** Before giving a formal answer, we observe that, if we know that the first die is 5, then the probability is going to be higher. We have excluded many "low" values of the first die which would not allow us to get to the desired result.

To answer the 2nd question formally, let us define the event *F* as "the first die is equal to 5".



I will have to **count the outcomes within the intersection of the two events,** and relate them **not** to the number of outcomes in *S*, but to the **number of outcomes in *F***. This is because the phrase "given *F*" means that I have already excluded all the outcomes in F^c.

With this in mind, I can define:

$$P(E|F) = \frac{P(EF)}{P(F)}$$

Note that the conditional probability verifies the basic condition that we ask of probabilities, i.e. that $0 \leq P(E|F) \leq 1$. It does because $EF \subseteq F$, hence $P(EF) \leq P(F)$. Conditional probability is **only defined when** $P(F) > 0$ (it would make no sense to condition to an impossible event).

In the above example, $F = \{1^{st}\text{dieequalto5}\}$, $EF = \{(5,6),(5,5)\}$, and we straightforwardly obtain $P(F) = 1/6$, $P(EF) = 2/36 = 1/18$. Hence,

$$P(E|F) = \frac{1/18}{1/6} = \frac{1}{3}$$

Note that $P(E|F) > P(E)$, as we expected.

As for the **first question** (sum >9 | first die is 2), the intersection of the two events is the **null event**, whose probability is zero. Hence the conditional probability is zero as well, which matches our intuition.

Exercise

In a group of transistors there are 5 **defective** ones (that simply cannot be turned on), 10 **unreliable** (that stop working after a couple of hours), and 25 **working** ones. What is the probability that, having chosen one transistor at random, it is a *working* one, given that it has already worked for 5 minutes?

Solution

Call *E* the event "working transistor" and *F* the event "non defective transistor". We have to compute:

$$P(E|F) = \frac{P(EF)}{P(F)} = \frac{P(E)}{P(F)} = \frac{25/40}{(25+10)/40} = \frac{5}{7}$$

Note that $P(E) = 25/40 = 5/8$, hence we are a little more confident that the transistor will be working if it has already worked for 5 minutes. In fact, we have just excluded that it is defective.

♦

Exercise

- 52% of the students of one college are female

- 5% of the students are majoring in Computer Engineering
- 2% of the students are female majoring in Computer Engineering

Take a student at random, and compute the probability that:

a) It's female, given that it's majoring in Computer Engineering
b) It's majoring in Computer Engineering, given that it's female.

Solution

a) $P(F|I) = \frac{P(FI)}{P(I)} = \frac{2}{5}$,

b) $P(I|F) = \frac{P(FI)}{P(F)} = \frac{2}{52} = \frac{1}{26}$

♦

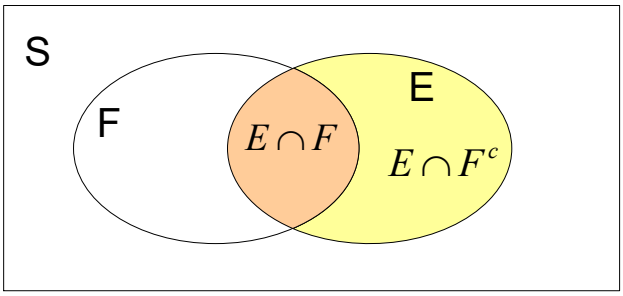
2.4.1 Total probability theorem and Bayes' formula

Conditional probability is an **extremely useful tool in practice**, because it allows you to compute unconditional probability in an easy way. Let us see how.

Event E can *always* be written in terms of another event F as follows

$$E = (E \cap F) \cup (E \cap F^c)$$

A simple Venn's diagram is enough to convince you of this.



The two events between parentheses are **mutually exclusive**, since F and F^c are mutually exclusive. Therefore, by the **third axiom of probability**:

$$P(E) = P(EF) + P(EF^c)$$

By substituting the expression of the two addenda in terms of conditional probability, assuming F and F^c as conditioning events, we obtain:

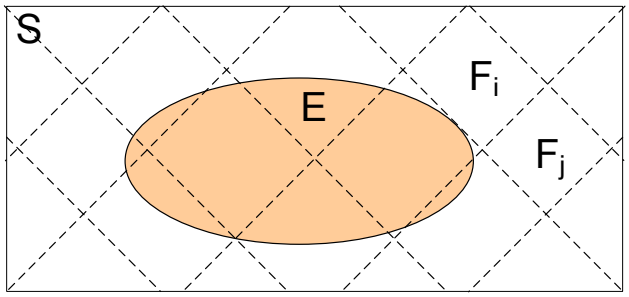
$$\begin{aligned} P(E) &= P(E|F) \cdot P(F) + P(E|F^c) \cdot P(F^c) \\ &= P(E|F) \cdot P(F) + P(E|F^c) \cdot [1 - P(F)] \end{aligned}$$

Where the last passage is given by the fact that $F \cup F^c$ covers the whole sample space.

This expression can be generalized to the case of ***N* mutually exclusive events**, whose union covers the whole sample space.

Given $F_1, \ldots, F_N$, such that $\cup_{i=1}^{N} F_i = S$, and $F_i \cap F_j = \emptyset$ if $i \neq j$, (i.e., given a way to **slice *S* in mutually disjoint events)**, we can compute the probability of an event E as:

$$P(E) = \sum_{i=1}^{N} P(EF_i) = \sum_{i=1}^{N} P(E|F_i) \cdot P(F_i)$$



The above formula is called **Theorem (or Law) of Total Probability**. It is interesting to ask oneself if this is of any **practical usefulness,** given that it seems to require more information to get to the same result. The truth is that it is often very hard to estimate $P(E)$, but it is often easy enough to estimate $P(E|F_i)$ for some events F_i whose probability is known. This makes the total probability law quite useful in practice.

Example

There are two classes of people, those that are accident-prone, and those that are not. An insurance company knows that accident-prone people have 40% probability of having an accident in a year, and non-accident-prone have 20% probability. They also know that 30% of the drivers are accident-prone. What is the probability that a new insurance policy subscriber will have an accident next year?

Call A the event "to have an accident in the next year", and B "to be accident-prone". We want to know $P(A)$, which is difficult to estimate. However, we know $P(B), P(A|B), P(A|B^c)$, hence we can readily apply the total probability theorem:

$$\begin{aligned} P(A) &= P(A|B) \cdot P(B) + P(A|B^c) \cdot P(B^c) \\ &= P(A|B) \cdot P(B) + P(A|B^c) \cdot [1 - P(B)] \\ &= 0.4 \cdot 0.3 + 0.2 \cdot 0.7 \\ &= 0.26 \end{aligned}$$

♦

Bayes' Theorem uses conditional probabilities, and is expressed as follows:

Given $F_1, \ldots, F_N$ (hypotheses), such that $\cup_{i=1}^{N} F_i = S$, and $F_i \cap F_j = \emptyset$ if $i \neq j$. Of these hypotheses, you know the a priori probability $P(F_j)$. Now, an event E occurs that may be due to some of the above hypotheses. The fact that event E has occurred modifies my knowledge about the hypotheses as follows (a posteriori probability):

$$P(F_j|E) = \frac{P(EF_j)}{P(E)} = \frac{P(E|F_j) \cdot P(F_j)}{\sum_{i=1}^{N} P(E|F_i) \cdot P(F_i)}$$

Bayes was a philosopher. This formula describes how the ***confidence* about a hypothesis** F_j is modified by the fact that an event E has occurred, given that E could be influenced by that hypothesis.

Exercise

A laboratory test to spot a particular blood disease is 99% accurate for those that are ill with that disease. However, the test has a meager 1% **false positive rate**, meaning that if you are *not* ill, in 1% of the cases the test will say that you are.

Given that 0.5% of the population has that disease, what is the probability that you will have the disease if the test finds you positive?

Solution

Call D the event "the subject is ill with the disease", and P the event "the test is positive". We have:

- $P(P|D) = 0.99$ (test accuracy)
- $P(P|D^c) = 0.01$ (false positive rate)
- $P(D) = 0.005$

$P(D)$ is the *a priori probability* that one has the disease, as given by statistics on the population. We want to compute $P(D|P)$, i.e. the *a posteriori* probability that one has the disease, as modified by the fact that that person has been detected positive by the blood test. We expect that $P(D|P) > P(D)$, since the test is aimed at giving a higher confidence.

By Bayes' formula, we have:

$$P(D|P) = \frac{P(DP)}{P(P)} = \frac{P(P|D) \cdot P(D)}{P(P)}$$
$$= \frac{P(P|D) \cdot P(D)}{P(P|D) \cdot P(D) + P(P|D^c) \cdot P(D^c)}$$
$$= \frac{P(P|D) \cdot P(D)}{P(P|D) \cdot P(D) + P(P|D^c) \cdot [1 - P(D)]}$$

Now we have all the numbers, and we can substitute them into the formula:

$$P(D|P) = \frac{0.99 \cdot 0.005}{0.99 \cdot 0.005 + 0.01 \cdot 0.995} \approx 0.332$$

The result seems **surprising**, given that the test is very accurate. However, it makes perfect sense if you reflect on it a bit. 0.5% means "1 every 200". Every 200 people:

- 1 person has the disease, and will be diagnosed as ill with 99% probability, i.e. almost certainly
- 199 people are perfectly healthy. Of these, $199 \cdot 0.01 \approx 2$ will be false positives.

Thus, on each three positives, two are false ones.

In Bayes' terms, knowing that an event P has occurred **modifies my opinion** on the hypothesis D. Before running the test, I could only think that there was $P(D) = 0.005$ that the next person had the disease. Now, I know that the person has $P(D|P) \approx 0.332 > P(D)$ probability of having the disease. The occurrence of the event has **increased my confidence on the hypothesis** (about 66 times).

For the same reason, **if the blood test comes out *negative***, I expect that $P(D|P^c) < P(D)$. Let's confirm this through computations:

$$\begin{aligned} P(D|P^c) &= \frac{P(DP^c)}{P(P^c)} = \frac{P(P^c|D) \cdot P(D)}{1 - P(P)} \\ &= \frac{\big(1 - P(P|D)\big) \cdot P(D)}{1 - P(P)} \\ &= \frac{0.01 \cdot 0.005}{1 - (0.99 \cdot 0.005 + 0.01 \cdot 0.995)} \\ &= 5.07 \cdot 10^{-5} \end{aligned}$$

Meaning that, if the test finds you negative, you have one chance in 20,000 of actually being ill (the probability *a posteriori* is 100 times smaller).

♦

Exercise

I ask my neighbor to water a plant while I am away. I believe that:

- If he doesn't water it, the plant has an 80% probability to die
- If he waters it, the plant has a 15% probability to die.

I also believe that the neighbor will remember to water the plant with 90% probability.

1) What is the probability that I will find the plant alive on my return?
2) Given that I return to find the plant dead, what is the probability that the neighbor forgot to water it?

Solution

[Before solving the exercise, we observe that this exercise introduces surreptitiously the concept of probability as *subjective measure on the confidence that an event will occur*, which is very different from the concept of "relative frequency". The first definition is somewhat more interesting for economists and philosophers, whereas the second one is interesting for scientists and engineers. We don't really care how the probabilities are defined right now, since we are interested in how to manipulate them. The good news is that the rules we devised for this hold regardless of the interpretation].

Let us define the two events of interest: *D={the plant dies}, W={the plant is watered}*. We have: $P(D|W) = 0.15, \quad P(D|W^c) = 0.8, \quad P(W) = 0.9$

1) We need to compute $P(D^c)$:

$$\begin{aligned} P(D^c) &= 1 - P(D) \\ &= 1 - [P(D|W) \cdot P(W) + P(D|W^c) \cdot P(W^c)] \\ &= 1 - \big[P(D|W) \cdot P(W) + P(D|W^c) \cdot \big(1 - P(W)\big)\big] \\ &= 1 - [0.15 \cdot 0.9 + 0.8 \cdot 0.1] \\ &= 0.785 \end{aligned}$$

2) We are now looking at $P(W^c|D)$, which we can find using Bayes' formula.

$$\begin{aligned} P(W^c|D) &= \frac{P(W^c D)}{P(D)} = \frac{P(D|W^c) \cdot P(W^c)}{P(D)} \\ &= \frac{P(D|W^c) \cdot [1 - P(W)]}{P(D|W) \cdot P(W) + P(D|W^c) \cdot [1 - P(W)]} \\ &\cong 0.372 \end{aligned}$$

The ***a priori* probability** that the neighbor will forget to water the plant is $P(W^c) = 0.1$. An event occurs, i.e. the plant dies, and the ***a posteriori* probability** that the neighbor forgot to water the plant can be reassessed. That probability is $P(W^c|D) = 0.372$, which is higher than the *a priori* one.

♦

2.4.2 Independent events

As we have seen, in general a **conditional probability** $P(E|F)$ is different from the **unconditional probability** $P(E)$. Knowing something about *F* changes my knowledge about *E*. In some cases, event *F* could be **irrelevant**, meaning that the occurrence of *F* does not give any information about *E*. In this case we say that *E* and *F* are **independent events.**

The formal definition of independent events is the following:

Two events *E* and *F* are independent if and only if:

$$P(EF) = P(E) \cdot P(F)$$

This definition is better understood if we write it down in terms of **conditional probabilities**:

$$P(E|F) = \frac{P(EF)}{P(F)} = \frac{P(E) \cdot P(F)}{P(F)} = P(E)$$

Where the last passage is due to independence. This formula says that conditioning E to F yields no information.

Some **useful properties**, that have to be kept in mind:

a) If E is independent of F, **it is independent of F^c as well**. This is fairly obvious: if knowing that F has occurred does not change things about E, even knowing that it **has not occurred** cannot.
b) Independence is **symmetric**: if E is independent of F, then F is independent of E. This one is obvious too, and can be proved by changing E and F in the above formula.
c) If E is independent of **both** F **and** G, **it is not necessarily independent of FG.** Independence between more than two events is tricky.

In practice, independence is seldom something that you have to verify yourself: rateher, it is implied by some ***a priori*** knowledge of your system. For more than a bunch of events, it is almost impossible to verify, in any case.

A very common case of independent events is the following.

2.4.3 Repeated trials

A typical case of independent events is given when an experiment consists in **repeating *n times* the same subexperiment** under "**independent conditions**" (e.g., the repeated flip of a coin). Independent conditions mean that the j-th trial is not influenced by the results of the previous trials. For instance, you can reasonably assume that this happens when you flip a coin several times.

Exercise

A coin is flipped five times, in independent conditions. Compute the probability that:

a) The first three flips yield the same outcome
b) Either the first three, or the last three flips, yield the same outcome
c) There are at least two heads in the first three flips, and two tails in the last three flips.

Solution

Each flip is an independent subexperiment. Thus, I can multiply the probabilities of events related to different flips. I can solve a) and b) using *either* this trick *or* the basic principle of counting. Let us use independence.

a) By independence:
   - You flip the coin once, and obtain an outcome with probability 1.
   - The second flip is equal to the first with probability ½
   - The third flip is equal to the first with probability ½.

   The three events are independent, hence the resulting probability is the product of the three above, which is exactly ¼.

   Taking into account that this is an UPM experiment, the possible combinations of three results are $N = 2^5$. Of these, eight are those that interest me, i.e. $E = \{(HHHxx), (TTTxx)\}$. Therefore, the requested result is $P = 8/32 = 1/4$.

   Note that the fact that this is a UPM is due to the fact that the coin is **fair.** If the coin is **unfair** (say, biased towards heads with 60% probability), then you **cannot use UPM**, but you **can still use independence.**

b) By independence, you have to compute the probability of event $F \cup L$, where $F = \{3\text{h/tatthebeginning}\}$, and $L = \{3\text{h/tattheend}\}$. It is $F \cap L = \{5\text{h/t}\}$. Thus $P(F) = 1 \cdot 1/2 \cdot 1/2 = P(L)$, and $P(FL) = 1 \cdot (1/2)^4$. Hence:

$$\begin{aligned} P(F \cup L) &= P(F) + P(L) - P(FL) \\ &= 1/4 + 1/4 - 1/16 \\ &= \frac{7}{16} \end{aligned}$$

   Using the UPM, we get to the same result through a longer route. The combinations of 5 outcomes are $N = 2^5$. We are interested in the cardinality of the event $E = E_{Hs} \cup E_{He} \cup E_{Ts} \cup E_{Te}$, with (*H*=heads, *s*=start, *T*=tails, *e*=end):

   $E_{Hs} = \{(HHHxx), x \in \{H, T\}\}$, and the same goes for the other three.

   Every subset has a cardinality of four elements, but subsets are **not disjoint**

   - $E_{Hs} \cap E_{He} = \{(HHHHH)\}$
   - $E_{Ts} \cap E_{Te} = \{(TTTTT)\}$
   - $E_{Hs} \cap E_{Te} = E_{Ts} \cap E_{He} = E_{Hs} \cap E_{Ts} = E_{Hs} \cap E_{Te} = \emptyset$

   Hence,

$$\begin{aligned} |E| &= |E_{Hs} \cup E_{He} \cup E_{Ts} \cup E_{Te}| \\ &= |E_{Hs}| + |E_{He}| + |E_{Ts}| + |E_{Te}| - |E_{Hs} \cap E_{He}| - |E_{Ts} \cap E_{Te}| \\ &= 4 \cdot 4 - 2 = 14 \end{aligned}$$

   The result is $P = 14/32 = 7/16$.

c) Here, it is preferable to use the UPM model. Of 32 outcomes, the following are in the event I want to observe:

$$xHHTT, HxHTT, HHxTT, HHTxT, HHTTx$$

Let us count how many outcomes we have overall, without including duplicates:

- For the first set, x can be H or T
- For the second one, only T (otherwise I am counting the same outcome twice)
- For the 3rd, only T
- For the 4th and 5th only H

We therefore have 6 outcomes in our event. The result is $P = 6/32 = 3/16$.

♦

Exercise

Mr. Rossi has a bunch of n keys, $n > 1$, one of which opens his door. What is the probability that i) choosing a key at random, and ii) discarding it if it is the wrong one,

a) He opens the door exactly on the kth attempt, $1 \leq k \leq n$
b) He opens the door *within k attempts*

Assume now that Mr. Rossi does not discard the wrong keys after a failed attempt.

c) Answer the previous questions again

Solution

There are at least two ways to answer question a). The first one is based on the principle of counting, and the second one is based on conditional probability. To introduce the second one, it is convenient to solve point b) first.

a1) the sample space is the **set of all permutations of *k* elements extracted from a set of *n*.** Note the difference between *permutations* and *subsets*, which I have already pointed out (in permutations, order matters. In subsets, it does not).

The permutations of k elements are:

$$N = S_{k,n} = n \cdot (n-1) \cdot \ldots \cdot (n-k+1) = n!/(n-k)!$$

We are in the case of *equally likely results* (every sequence has the same probability of every other, given that I am choosing the keys at random). I have to compute the cardinality of the event

$$E_k = \{k^{th}\text{keyistherightone}\}$$

Event E_k includes all the sequences of k-1 elements (the first k-1 wrong keys) taken from a set of n-1 keys (every key except the right one). Then:

$$|E_k| = S_{k-1,n-1} = (n-1)!/(n-k)!$$

Then you obtain:

$$P(E_k) = \frac{|E_k|}{N} = \frac{S_{k-1,n-1}}{S_{k,n}} = \frac{(n-1)!}{(n-k)!} \cdot \frac{(n-k)!}{n!} = \frac{1}{n}$$

And the result does not depend on *k*.

This is the same case we have already seen in a previous exercise. Question a) is equivalent to asking: "what is the probability that, sorting *n* keys at random, one particular key ends up in the kth position"? The answer is $1/n$, and does not depend on *k* (there is nothing special about the kth position).

b) The probability of opening the door *within k* attempts is the probability of event $F_k = \bigcup_{i=1}^{k} E_i$. The events$E_i$, $1 \leq i \leq n$ are **mutually exclusive**. Therefore, the probability I am looking for is:

$$P(F_k) = P\left(\bigcup_{i=1}^{k} E_i\right) = \sum_{i=1}^{k} P(E_i) = k \cdot \frac{1}{n} = \frac{k}{n}$$

The result is confirmed by the intuition: as *k* grows, the probability increases linearly, until it becomes *certainty* at the nth attempt.

a2) Let us exploit conditional probabilities and the theorem of total probability.

The probability of opening the door on the 1st attempt is $P(E_1) = 1/n$.

For the 2nd attempt, I can use **total probability**:

$$\begin{aligned} P(E_2) &= P(E_2|E_1) \cdot P(E_1) + P(E_2|{E_1}^c) \cdot P({E_1}^c) \\ &= 0 + \frac{1}{n-1} \cdot \frac{n-1}{n} \\ &= \frac{1}{n} \end{aligned}$$

In fact, $P(E_2|E_1) = 0$ since E_2E_1 is the null event. Furthermore, $P(E_2|{E_1}^c) = 1/(n-1)$, since the conditioning event implies that **one key is discarded**, hence the choice is reduced by one.

For a generic k^{th} attempt I have:

$$\begin{aligned} P(E_k) &= P(E_k|F_{k-1}) \cdot P(F_{k-1}) + P(E_k|{F_{k-1}}^c) \cdot P({F_{k-1}}^c) \\ &= 0 + \frac{1}{n-(k-1)} \cdot \left[1 - \frac{k-1}{n}\right] \\ &= \frac{1}{n} \end{aligned}$$

The first addendum is always null. For the second one, we apply the same reasoning: the conditional probability is to pick the right key from the bunch having discarded *k*-1 wrong keys. The probability of using the wrong key *k*-1 times is the complement of the one that has been computed at point b).

c) If Mr. Rossi is **not** discarding the wrong keys, he is executing *k* **repeated trials in independent conditions**, in each one of which the probability of getting the right key out of the bunch is $p = 1/n$. Therefore, the probability that the door opens ***on*** the kth attempt is:

$$P(E_k{}') = (1-p)^{k-1} \cdot p = \left(\frac{n-1}{n}\right)^{k-1} \cdot \frac{1}{n} = \frac{(n-1)^{k-1}}{n^k}$$

Note that $\lim_{k\to\infty} P(E_k) = 0$. This should not surprise us, since it is perfectly logical that, as k grows larger, the probability that Mr. Rossi gets the wrong key k-1 consecutive times goes to zero.

The event "Mr. Rossi opens the door <u>within</u> k attempts" is $F_k{}' = \bigcup_{i=1}^{k} E_i$. Moreover, events E_i are mutually disjoint. Therefore:

$$P(F_k{}') = \sum_{i=1}^{k} \left((1-p)^{i-1} \cdot p\right) = \frac{1}{n} \cdot \sum_{i=1}^{k} \left(1 - \frac{1}{n}\right)^{i-1}$$

The above one is a geometric series, that has a closed form. In order to save computations, it is easier to compute the probability of the **complementary event**, i.e. the event that Mr. Rossi is unable to open the door for k consecutive attempts. This one is clearly (by independence and repeated trials):

$$P(F_k{}'^c) = \left(\frac{n-1}{n}\right)^k = (1-p)^k$$

Hence we obtain:

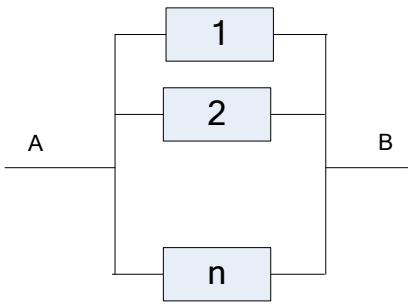
$$P(F_k{}') = 1 - P(F_k{}'^c) = 1 - \left(\frac{n-1}{n}\right)^k$$

This has an intuitive explanation. As k grows (i.e. you try again and again), the probability of opening the door tends to 1 *asymptotically*. However, for any *finite* number of attempts, there is still a residual probability that you get the wrong key every time.

♦

2.4.4 Parallel systems

A system is said to be *parallel* when it is composed of n subsystems, and it works if *at least one* of the subsystem is working. It is often the case that **the subsystems can be considered to be *independent***. In this case computing the probability that the system works is often simple enough.



The same model can be explained in terms of **switches.** In the above figure, subsystems are power switches, and current flows between A and B is **at least one of the switches** is closed (which is the same as saying that the system works if at least one subsystem is working).

Assume that every switch **is closed** with probability $\boldsymbol{p_i}$, $1 \leq i \leq n$, and that they are independent. The probability that the current flows is computed as follows:

Define A_i the event "switch i is closed". If $\boldsymbol{P(A_i) = p_i}$, then the probability of ${A_i}^c$ is $(1 - p_i)$.

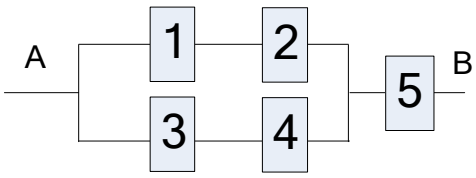
$$
\begin{aligned}
&P\{\text{systemworks}\} = 1 - P\{\text{currentdoesnotflow}\} \\
&= 1 - P\{\text{allswitchesareopen}\} \\
&= 1 - P({A_1}^c {A_2}^c \ldots {A_n}^c) \\
&= 1 - \textstyle\prod_{i=1}^{n} P({A_i}^c) \\
&= 1 - \textstyle\prod_{i=1}^{n} (1 - p_i)
\end{aligned}
$$

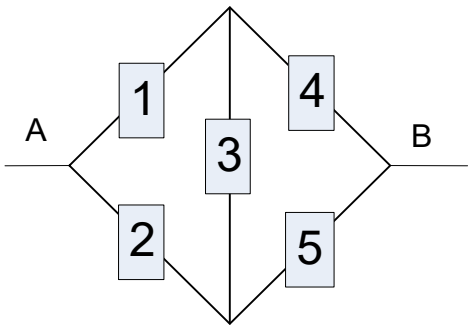
By independence. If A and B are independent, so are A^c e B^c

Here we exploit the trick of computing **the probability of the complementary** event. This is often important in many cases, so try to memorize it.

Exercise

Call p_i the probability that the i-th switch is closed. Compute the probability that current flows from A to B, assuming that switches are independent, in the two cases below.





Solution

Call E_i the event "the i-th switch is closed".

Consider the **system on the left.**

Call E_{sup} and E_{inf} the events for which the upper and lower branches are traversed by current. The event "current flows through" is the intersection of two independent events:

- The parallel system on the left allows current to flow through
- Switch 5 is closed

The parallel system on the left allows current through with the following probability:

$$
\begin{aligned}
p_{par} &= 1 - \left(1 - p_{\text{sup}}\right)(1 - p_{\text{inf}}) \\
&= 1 - (1 - p_1 \cdot p_2)(1 - p_3 \cdot p_4)
\end{aligned}
$$

Where the last passage is due to the fact that the switches (and, specifically, those on the same line) are independent. Therefore the probability is:

$$p_{par} \cdot p_5 = [1 - (1 - p_1 \cdot p_2)(1 - p_3 \cdot p_4)] \cdot p_5$$

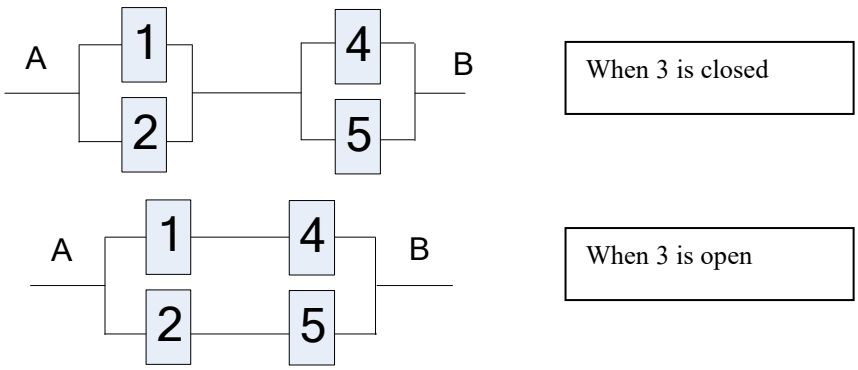
♦

Consider now the **system on the right.**

Call C the event "current flows". The exercise would be easy, were it not for switch 3. However, we can make two **alternative and complementary hypotheses** regarding switch 3, which is **either closed or open**. Then, we use the Theorem of Total Probability, which ensures us that:

$$P(C) = P(C|E_3) \cdot P(E_3) + P(C|E_3^c) \cdot P(E_3^c).$$

Computing the conditional probabilities in the two cases is a lot easier:



$P(C|E_3)$ can be computed by observing that the two parallel subsystems (on the left and on the right respectively) are **independent** by hypothesis. In fact, if all switches are independent, every combination of switches is independent of every other. Therefore, their probabilities can be computed as:

- *left parallel (1-2)*: $1 - (1 - p_1) \cdot (1 - p_2)$
- *right parallel (4-5)*: $1 - (1 - p_4) \cdot (1 - p_5)$

Hence: $P(C|E_3) = [1 - (1 - p_1) \cdot (1 - p_2)] \cdot [1 - (1 - p_4) \cdot (1 - p_5)]$

$P(C|E_3^c)$ can instead be computed directly, by observing that the probabilities of switches lying on the same line can be multiplied, since events are independent.

$$P(C|E_3^c) = 1 - (1 - p_1 p_4) \cdot (1 - p_2 p_5)$$

Putting it all together, we obtain:

$$\begin{aligned} P(C) &= P(C|E_3) \cdot P(E_3) + P(C|E_3^c) \cdot P(E_3^c) \\ &= [1 - (1 - p_1) \cdot (1 - p_2)] \cdot [1 - (1 - p_4) \cdot (1 - p_5)] \cdot p_3 + \\ &+ [1 - (1 - p_1 p_4) \cdot (1 - p_2 p_5)] \cdot (1 - p_3) \end{aligned}$$

♦

Exercise

We collect *k* coupons, each one of which can be independently of *n* different types, with probability p_j (such that, obviously, $\sum_{i=1}^{n} p_j = 1$).

What is the probability that a collection of *k* coupons contains at least *one* type-*i* or type-*j* coupon?

Solution

Define E_x={*a coupon is type-x*}. The probability that a coupon is of type *i* or *j* is $P(E_i \cup E_j)$, and it is equal to $P(E_i) + P(E_j) = p_i + p_j$ (the two events are mutually exclusive). Therefore, $P((E_i \cup E_j)^c) = 1 - (p_i + p_j)$ is the probability that a coupon is neither type-*i* nor type-*j*.

The probability that of *k* coupons **there aren't any** of type *i* or *j* (which is the complementary event of the one we are looking for) is therefore $P_c = [1 - (p_i + p_j)]^k$.

Therefore, the probability that I am looking for is $P = 1 - P_c = 1 - [1 - (p_i + p_j)]^k$.

♦

3 Random variables

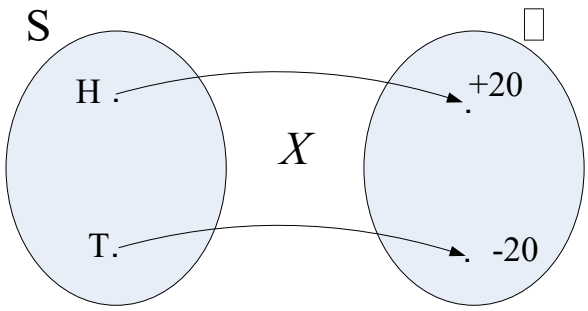
There are several ways to define random variables. The one I prefer is as **real-valued functions.** Given a **random experiment** whose sample space is S, we say that X is a **random variable** on S if it is a **real-valued function** $X: S \to \mathbb{R}$. Random variables are denoted with **uppercase letters**. Function X has **nothing random in itself**. It is instead a perfectly deterministic one. It is the outcome of the experiment that is random.

Example

Suppose that you have a coin flip, hence $S = \{H, T\}$, and that you bet 20 cents on "heads". You can define a **random variable *X*** as follow**:**

$$X(H) = +20, X(T) = -20$$

That random variable defines the net gain of your bet, given the outcome.



In this case I can (and henceforth will) take a shortcut and mention "**the probability that *X* is equal to +20",** $P\{X = +20\}$, meaning in fact the probability that **"the event occurs whose image through *X* is the real value +20"**, in this case event "heads".

That probability is $P\{X = +20\} = P\{X = -20\} = 0.5$

♦

Example

Take the random experiment consisting in the **throw of two dice**. The sample space is:

$$S = \{(d_1, d_2) | 1 \leq d_1, d_2 \leq 6\}$$

I can define the following random variables:

- X, **sum** of the values on each die: $X: S \to \mathbb{R}$, $X\big((d_1, d_2)\big) = d_1 + d_2$

- Y, **maximum** value on either die: $Y: S \to \mathbb{R}$, $Y\big((d_1, d_2)\big) = \max\{d_1, d_2\}$

X takes on values: $\{2,3,\ldots,11,12\}$, whereas Y takes on values: $\{1,2,\ldots,6\}$.

In both cases, I can associate a probability to each value using UPM. Let's do it for Y:

$$P\{Y=1\}=P\{(1,1)\}=1/36$$

$$P\{Y=2\}=P\{(2,1),(1,2),(2,2)\}=3/36$$

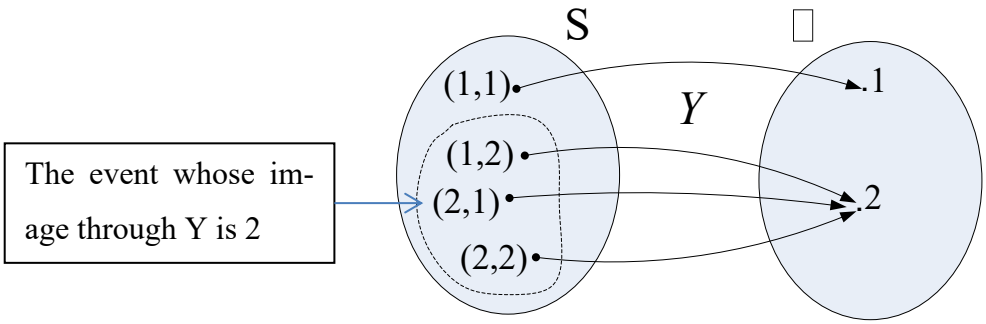
$$P\{Y=3\}=P\{(1,3),(2,3),(3,3),(3,2),(3,1)\}=5/36$$

$$P\{Y=4\}=P\{(1,4),(2,4),(3,4),(4,4),(4,3),(4,2),(4,1)\}=7/36$$

$$P\{Y=5\}=P\{(1,5),(2,5),(3,5),(4,5),(5,5),(5,4),(5,3),(5,2),(5,1)\}=9/36$$

$$P\{Y=6\}=1-\sum_{i=1}^{5}P\{Y=i\}=1-\frac{1+3+5+7+9}{36}=1-\frac{25}{36}=\frac{11}{36}$$

(The last one could be computed directly, but it's quicker this way).



♦

Example

We buy two electronic devices, each one of which can be either ***functioning*** **or** ***defective*** with some probability. The sample space for this random experiment is the set of the following outcomes:

S={ (f,f), (f,d), (d,f), (d,d) }

We have a probability for each outcome:

$$P(\{(f,f)\})=0.49,\ P(\{(f,d)\})=P(\{(d,f)\})=0.21, P(\{(d,d)\})=0.09$$

Given a random experiment, the definition of a random variable on that experiment is **in the mind of the observer**, same as the definition of an outcome.

For instance, I can define the random variable X **"number of functioning devices"** as follows:

$$X\big((f,f)\big)=2,\ X\big((f,d)\big)=1,\ X\big((d,f)\big)=1,\ X\big((d,d)\big)=0.$$

Hence I will have:

$$P\{X=2\}=0.49,\ P\{X=1\}=0.42,\ P\{X=0\}=0.09.$$

I can have something else in mind, and define a different random variable Y, that is equal to 1 if the number of functioning devices is **even**, and 0 if it is **odd**. In this case:

$$P\{Y=1\}=0.58,\ P\{Y=0\}=0.42,$$

Since the event whose image is 1 occurs with probability 0.58.

♦

The **values** of a random variable X are the subset **of real numbers** into which function X maps the sample space.

A **discrete** RV takes on a **discrete number of <u>real</u> values**, as in all the examples that we have just seen. Take care not to make a common mistake: the fact that a random variable is discrete does not have anything to do with **its values being integer** (as it was, by chance, in the previous examples). You just need to take "half the maximum of the two dice" to obtain a different random variable on the same experiment, which is still discrete, but whose values are real and not integer.

For a **continuous** RV, the set of possible values is **an interval of real numbers**. The typical case of a continuous RV is the **lifetime** of a device, and in general things that are connected with time or frequency.

Example

A random experiment consists in measuring the lifetime of a device. Its sample space is an interval $S = [0, \infty)$, and each outcome is a possible lifetime. We define the **continuous** RV $X: S \to \mathbb{R}$, such that $X(t) = t$.

If the example looks too trivial, take this one: measure the random lifetime of **two devices**. The sample space is $S = [0, \infty) \times [0, \infty) = \{(t_1, t_2) | t_1, t_2 \geq 0\}$, and define a continuous RV $X: S \to \mathbb{R}$, such that $X\big((t_1, t_2)\big) = \min(t_1, t_2)$.

♦

3.1 Cumulative Distribution Function of a random variable

A random variable X (either discrete or continuous) is **completely characterized** by its **Cumulative Distribution Function (CDF)**, or **distribution** for short. The latter is defined as follows:

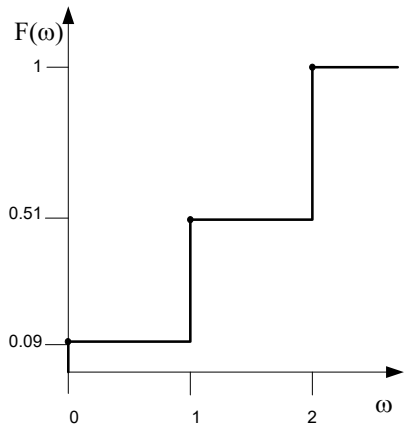
$$F(\omega) = P\{X \leq \omega\}$$

Omega is a **value**, and I am writing it in **lowercase**, whereas X is a random variable (uppercase). Notice the **weak inequality** (which is mandatory in the definition).

For instance, for RV X "number of functioning devices", we have:

$$F(0) = 0.09, F(1) = 0.51, F(2) = 1.$$

For a **discrete** RV, the CDF is a **staircase** function. A CDF (**any CDF**) is always **weakly monotonic,** and it has values between 0 and 1 on the y axis (which reports probabilities, in fact).



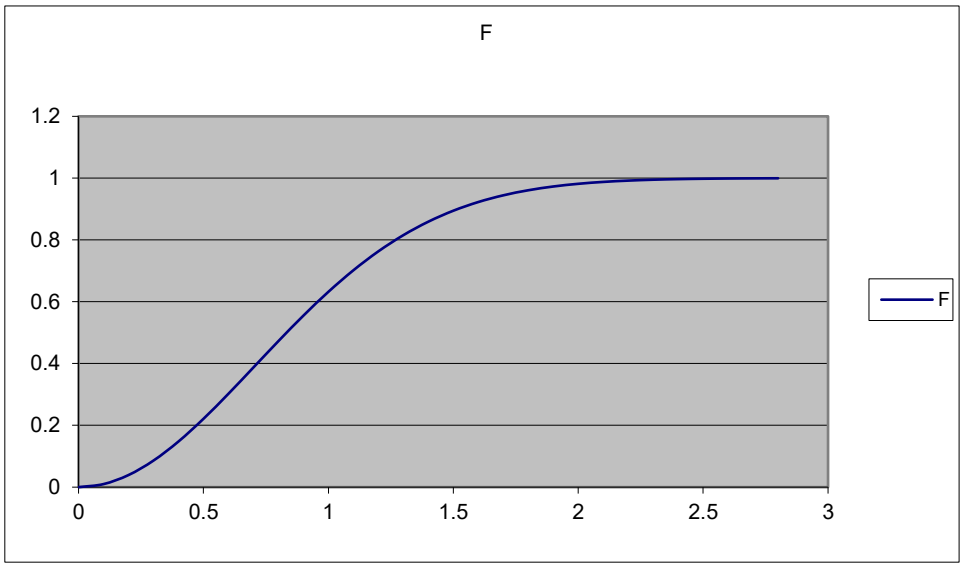
Moreover, it is quite obvious that $\lim_{\omega \to -\infty} F(\omega) = 0$, $\lim_{\omega \to \infty} F(\omega) = 1$. This holds for **any RV**.

Note that staircase CDFs are **right continuous**, meaning that the right values are those that are associated to a step. For instance, $F(1)$=0.51 (this is implied by the weak inequality in the definition, and that's why it is important to keep it in mind).

Exercise

Consider an RV X whose$F(\omega)$ CDF is the following:

$$F(\omega) = \begin{cases} 0 & \omega \leq 0 \\ 1 - e^{-\omega^2} & \omega > 0 \end{cases}$$



Compute $P\{X > 1\}$.

Solution:

$P\{X > 1\} = 1 - P\{X \leq 1\} = 1 - F(1) = 1 - (1 - e^{-1}) = 1/e.$

♦

Once you have the CDF of a RV, **you can answer any question related to the probability of that RV**. For instance, in the above example, I can ask myself what is the following probability: $P\{1 < X \leq 2\}$. The answer can be found as follows:

$$P\{X \leq 2\} = P\{X \leq 1\} + P\{1 < X \leq 2\}$$

I can sum the probabilities on the r.h.s. since the two events are **mutually exclusive**. From the definition of CDF; I obtain the following:

$$\begin{aligned} &P\{1 < X \leq 2\} = F(2) - F(1) \\ &= (1 - e^{-4}) - (1 - e^{-1}) \\ &= 1/e - 1/e^4 \end{aligned}$$

3.2 Probability Mass Function of discrete RVs

For **discrete RVs** (and for these **<u>only</u>**), a **Probability Mass Function** (PMF) can be defined. The definition is:

$$p(a) = P\{X = a\}$$

Note that we use **lowercase** to denote a PMF, and **uppercase** to denote a CDF.

For a discrete RV, $p(a)$ can be non-null only for a **numerable quantity of values**. This is because $\sum_{-\infty}^{+\infty} p(a) = 1$, an equality which is called **normalization condition**.

Furthermore, it is quite straightforward to observe that $F(a) = \sum_{x \leq a} p(x)$, hence we can find the CDF from the PMF. We can also do the reverse, since $p(a) = F(a) - F(a^-)$. Thus, knowing the PMF or the CDF of a discrete RV is pretty much the same thing.

Example

A discrete RV has 3 values: 1,2,3. We know $p(1) = 1/2$, $p(2) = 1/3$. Draw a graph of the PMF and CDF.

It is fairly obvious that $p(3) = 1 - \big(p(1) + p(2)\big) = 1/6$. The PMF has three spikes corresponding to the three values, each one as large as the related probability. The CDF is a staircase, with each step as large as the corresponding spike.

♦

As another example, we have already computed analytically the PMF for the RV *Y={maximum of two dice}*. From the latter, it is straightforward to compute the CDF for that RV.

3.3 Probability Density Function for continuous RVs

For continuous RVs **it makes no sense to define a PMF**. It is in fact impossible that a RV takes on exactly one value (with infinite precision) in a continuous space. The best that I can do in this case is to explore whether a RV is more likely to take on a value **in a certain interval than in another**.

For continuous RVs, we define the **Probability Density Function (PDF)** $\boldsymbol{f(x)}$ (mind the **lowercase**), which is a non negative function with the following property. $\boldsymbol{f(x)}$ is a PDF if, given any set B of real numbers,it is:

$$P\{X \in B\} = \int_B f(x)dx$$

From the above definition we quickly derive that:

$$P\{X \in (-\infty; +\infty)\} = \int_{-\infty}^{+\infty} f(x)dx = 1,$$

Which is again called **normalization condition**.

And, if B is an interval $[a, b]$, we obtain:

$$P\{a \leq X \leq b\} = \int_a^b f(x)dx = F(b) - F(a).$$

Note that, with continuous RVs, it does not really matter whether inequalities are strict or weak. From the above, we get a more intuitive "physical" explanation of the concept of PDF:

$$P\left\{X \in \left(a - \frac{\varepsilon}{2}; a + \frac{\varepsilon}{2}\right)\right\} = \int_{a-\frac{\varepsilon}{2}}^{a+\frac{\varepsilon}{2}} f(x)dx \approx \varepsilon \cdot f(a)$$

This means that $f(a)$ measures the probability that the RV takes on a value **around *a*.** If we wished to be more precise, and look for the probability that the RV takes on a value which is *exactly a*, we would get:

$$P\{X = a\} = P\{a \leq X \leq a\} = \int_a^a f(x)dx = F(a) - F(a) = 0,$$

Which is coherent with our earlier intuition.

Finally, we get the most important property:

$$F(a) = P\{X \leq a\} = P\{-\infty \leq X \leq a\} = \int_{-\infty}^{a} f(x)dx,$$

Which can be written in a derivative form, **differentiating with respect to *a***:

$$f(a) = \frac{\partial}{\partial a} F(a)$$

For continuous RVs, the PDF is the derivative of the CDF. Obviously enough, I can differentiate the CDF **only where it is differentiable**. The CDF **always exists**, whereas the PDF exists only where the CDF is differentiable.

This is not a practical concern, since in all the cases we will deal with the CDF will be differentiable everywhere.

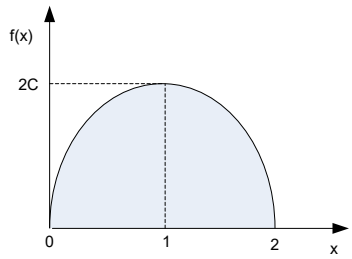
Knowing the PDF or the CDF of a continuous RV gives you the same information (except for the above *caveat*), in much the same way as knowing the PMF or the CDF gives you the same information for a discrete RV.

Example

Consider the following PDF of RV *X*:

$$f(x) = \begin{cases} C \cdot (4x - 2x^2) & 0 \le x \le 2 \\ 0 & otherwise \end{cases}$$

Where C is a real-valued constant.



Compute *C* and find probability $P\{X > 1\}$.

Solution

To compute *C* we just need to impose the **normalization condition**). The integral is the area of the figure bounded by the parabola and the *x* axis. Thus:

$$\int_0^2 C \cdot (4x - 2x^2)dx = 1$$

$$C \cdot \left[2x^2 - \frac{2}{3} \cdot x^3\right]_0^2 = 1$$

$$C \cdot \left[\left(8 - \frac{16}{3}\right)\right] = 1$$

$$C = \frac{3}{8}$$

As for the probability that $P\{X > 1\}$, obvious symmetry reasons confirm that it is 0.5. The figure is symmetric around the vertical line passing through 1, and its area is equal to 1, hence we are asking how large half the area is.

Let us confirm this using computations:

$$P\{X > 1\} = P\{1 \le X \le 2\} = \int_1^2 \frac{3}{8} \cdot (4x - 2x^2)dx$$
$$= \frac{3}{4} \cdot \left[x^2 - \frac{1}{3} \cdot x^3\right]_1^2$$
$$= \frac{3}{4} \cdot \left[\left(4 - \frac{8}{3}\right) - \left(1 - \frac{1}{3}\right)\right]_1^2$$
$$= \frac{3}{4} \cdot \left[\frac{4}{3} - \frac{2}{3}\right]$$
$$= \frac{1}{2}$$

♦

3.3.1 Exercises

Exercise

Five male and five female students are ranked according to their test grades. Assume each ranking is equally likely, and that there are no ex-aequo positions. Let *X* be a RV defined as *the <u>highest</u> position occupied by a female student* (meaning that the 1st is higher than the 2nd, etc.). Compute the PMF of *X*.

Solution

X is a discrete RV. We need to compute $p(j) = P\{X = j\}$ for $1 \le j \le 10$. Anyone can see that $p(j) = 0$ when $j \ge 7$, since it is impossible that no female students classify in the first 6 positions.

Let us start with $p(1)$, i.e. the probability that a female gets the best grades.

The female students that might get the best marks are 5, and for each one of these we have 9! possible rankings of the other students. The number of different rankings is 10!. Therefore:

$$p(1) = 5 \cdot \frac{9!}{10!} = \frac{1}{2}$$

Which is what the **intuition would suggest us**, i.e. the fact that – with equally likely outcomes and the same number of males and females – there is 50% probability that a female is in the first position.

Let's move on and compute $p(2)$. This is the probability that a *male* ranks first and a *female* ranks second. Since there are 5 males and 5 females, there are $5 \cdot 5 = 5^2$ couples (M, F) that can occupy the first two positions. For each one of these, we have 8! possible rankings of the other 8 students. Therefore:

$$p(2) = 5 \cdot 5 \cdot \frac{8!}{10!} = \frac{25}{10 \cdot 9} = \frac{5}{18}$$

We now move on to $p(3)$. The possible triplets (M, M, F) for the first three positions are $5 \cdot 4 \cdot 5$. For each one of these we have 7! rankings of the other 7 students. Hence:

$$p(3) = 5 \cdot (5 \cdot 4) \cdot \frac{7!}{10!} = \frac{5}{36}$$

Now we see quite well where we are heading. In order to compute the generic $p(j)$, I need to count the *permutations* of j-1 males from a set of 5, $S_{j-1,5} = 5!/[5-(j-1)]!$, and multiply it by the number of females that can occupy the jth position, which is again 5. As a last term, I have to compute the possible permutations of the remaining $(10-j)$ students, which are $(10-j)!$. In any case, I need to divide by 10!. The answer is therefore:

$$p(j) = 5 \cdot \frac{5!}{[5-(j-1)]!} \cdot \frac{(10-j)!}{10!},$$

Which holds for $1 \leq j \leq 6$ (i.e., for all the values for which the denominator is defined).

The values are shown in the table, and are obtained by instantiating the formula:

j	1	2	3	4	5	6	7-10
$p(j)$	$\frac{1}{2}$	$\frac{5}{18}$	$\frac{5}{36}$	$\frac{5}{84}$	$\frac{5}{252}$	$\frac{1}{252}$	0

One can (and should) **always test the normalization condition** *a posteriori*:

$$\sum_{j=1}^{10} p(j) = 1$$

Luckily enough, the condition holds, which confirms that I have done the computations correctly.

♦

Exercise

Let X be the difference between the number of *heads* and *tails* obtained when you flip a coin n times in independent conditions.

1) What are the values for X?
2) Compute the PMF of X when $n=3$

Solution

Point 1:

- with n=1 the values that X can take are +1 (1 heads - 0 tails) and -1 (0 heads - 1 tails)
- with n=2 the values for X are +2 (2 heads – 0 tails), 0, and -2
- with n=3 we have +3, +1, -1, -3

- with n=4 we have +4, +2, 0, -2, -4

In general, we have all the values from $-n$ to $+n$ included, at intervals of two, i.e. $S=\{-n+2j, 0 \leq j \leq n\}$.

Point 2:

If the coin if **fair**, the values with the same modulus (whichever their sign) **must** have the same probability. Therefore, we can limit ourselves to those with positive sign. Value +3 is obtained with 3 heads on 3 flips. Each heads has a probability of ½, and flips are independent (hence their probabilities can be multiplied). Thus: $p(+3) = p(-3) = (1/2)^3 = 1/8$.

In order to compute $p(+1) = p(-1)$ we can follow two methods:

- the quick one, that relies on symmetry and says that $p(+3) + p(+1) = 1/2$, hence $p(+1) = 3/8$.
- The slightly longer one, counting the number of favorable outcomes. These are: $\{(HHT), (HTH), (THH)\}$. These are 3, on 8 possibilities, hence the same result is obtained.

♦

Exercise

A radio uses 5 radio tubes. The lifetime of a radio tube is a continuous RV whose PDF is:

$$f(x) = \begin{cases} 0 & x \leq 100 \\ \dfrac{100}{x^2} & x > 100 \end{cases}$$

Compute the probability that *exactly* 2 tubes in 5 should be replaced within 150 hours of operation. Assume that the lifetimes of the tubes are independent.

Solution

It pays to be skeptical, hence we do a preliminary check that the above PDF really *is* a PDF. We do that by testing the normalization condition:

$$P\{X \in (-\infty; +\infty)\} = 100 \cdot \int_{100}^{+\infty} \frac{1}{x^2} dx = 100 \cdot \left[-\frac{1}{x}\right]_{100}^{+\infty} = 0 + \frac{100}{100} = 1$$

Now, the probability that **one tube** breaks within the first 150 hours is:

$$P\{X \leq 150\} = F(150) = \int_{-\infty}^{150} f(x)dx = 100 \cdot \left[-\frac{1}{x}\right]_{100}^{150} = \frac{1}{3} = p$$

The event of interest is "2 tubes on 5" are to be replaced, meaning that the other 3 actually work. Focus on two particular tubes (say, the 1st and the 2nd). The probability that these fail within 150 hours *and* the other three keep working is:

$$p^2 \cdot (1-p)^3$$

There are $\binom{5}{2}$ ways to take 2 elements from a set of 5. This means that there are $\binom{5}{2}$ outcomes $(R_1R_2R_3R_4R_5)$, having two faulty tubes, each one of which has the above probability. We can sum all these probabilities, since these outcomes are mutually exclusive. By doing this we obtain:

$$P = \binom{5}{2} p^2 \cdot (1-p)^3 = 10 \cdot \frac{1}{9} \cdot \frac{8}{27} = \frac{80}{243} \approx \frac{1}{3}$$

♦

3.4 Jointly distributed random variables

We are often interested in the *joint* distribution of two RV *X* and *Y*. Assume for instance that your experiment consists in shooting a target at random. The outcome of your experiment is **a couple of RVs**. Knowing the CDFs of each of the RVs gives you no information about where the points are located. You need the **joint CDFs** of both variables.

In some cases, you may want to know whether there is any **correlation** between two RVs. For instance, take the **number of cigarettes** smoked daily and the **age at which lung cancer is diagnosed**. Again, knowing each CDF alone is not overly informative, because you may want to know whether **large values** of the first are coupled with **small values of the second.**

Given two RVs (either discrete or continuous), their **Joint Cumulative Distribution Function (JCDF)** is defined as:

$$F(x,y) = P\{X \le x, Y \le y\}$$

Where the comma denotes a logical *and*, meaning the intersection of events $X \le x$ and $Y \le y$.

The **JCDF tells us everything there is to know** about the CDFs of the single RVs. In fact, to compute the CDF of variable *X*, I just need to observe that:

$$F_X(x) = P\{X \le x\} = P\{X \le x, Y \le +\infty\} = F(x, +\infty)$$

And the same goes for variable *Y*.

$$F_Y(y) = \ldots = F(+\infty, y)$$

It is **absolutely false** that the reverse holds. In general (once again, "in general") you cannot get information on the joint distribution from the **single** distributions.

3.4.1 Joint PMF for discrete RV

The JCDF exists for both discrete and continuous pairs of RVs. For **discrete RVs**, we can also define a **Joint Probability Mass Function (JPMF)**, as:

$$p(x,y) = P\{X = x, Y = y\}$$

In this case too we can get the PMFs of the single RVs from the JPMF. In fact:

$$\begin{aligned} &P\{X = x\} = P\{\cup_i (X = x, Y = y_i)\} \\ &= \textstyle\sum_i P\{X = x, Y = y_i\} \\ &= \textstyle\sum_i p(x, y_i) \end{aligned}$$

These are mutually disjoint events

And similarly, $P\{Y = y\} = \sum_j p(x_j, y)$.

Furthermore, the JCDF can be obtained from the JPMF as follows:

$$F(x, y) = P\{X \le x, Y \le y\} = \sum_{x_i \le x} \sum_{y_j \le y} p(x_i, y_j)$$

Given a pair of discrete RVs *X*,*Y*, whose JPMF is known, the two probabilities $P\{X = x\}$ and $P\{Y = y\}$ are often called **marginal probabilities**. The reason is that the JPMF is often given in a **table form** (with X, Y in row/column), hence the two above probabilities can be computed as row/column sums, and conveniently written on the **margin** of the table.

X \ Y	y1	y2	yk	
x1	p(x1,y1)	p(x1,y2)	p(x1,yk)	p(x1)
x2	p(x2,y1)	p(x2,y2)	p(x2,yk)	p(x2)
xh	p(xh,y1)	p(xh,y2)	p(xh,yk)	p(xh)
	p(y1)	p(y2)	p(yk)	

3.4.2 Joint PDF for continuous RVs

For two **continuous RVs** *X* and *Y*, if for every set C of pairs or real numbers (x, y), we have:

$$P\{(X, Y) \in C\} = \int_{(x,y) \in C} \int f(x, y) dx dy$$

Then $f(x, y)$ is called **Joint Probability Density Function (JPDF)** of *X* and *Y*.

More specifically, when *C* can be separated into two sets of real numbers $C = \{(x, y) | x \in A, y \in B\}$, we can rewrite the above integral as:

$$\begin{aligned} &P\{(X, Y) \in C\} = P\{X \in A, Y \in B\} \\ &= \int_A \left[\int_B f(x, y) dy \right] dx \end{aligned}$$

From the definition of JCDF, we obtain the following:

$$\begin{aligned} &F(a,b) = P\{X \le a, Y \le b\} \\ &= P\{X \in (-\infty, a], Y \in (-\infty, b]\} \\ &= \int_{-\infty}^{a} \int_{-\infty}^{b} f(x,y)dydx \end{aligned}$$

Hence, the JCDF can be obtained from the JPDF by integration. Furthermore, if we derive with respect to a and b (assuming that the JCDF is differentiable, which it normally is), we obtain the same relationship in the derivative form:

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

If the JPDF exists, then also the single PDFs of X and Y exist as well, and they can be computed from the JPDF quite easily.

$$\begin{aligned} &P\{X \in A\} = P\{X \in A, Y \in (-\infty, +\infty]\} \\ &= \int_{A} \left[\int_{-\infty}^{+\infty} f(x,y)dy \right] dx \end{aligned}$$

But we also know that $P\{X \in A\} = \int_A f_X(x)dx$, hence it is :

$$f_X(x) = \int_{-\infty}^{+\infty} f(x,y)dy$$

And the same reasoning holds for the other variable.

Exercise

The JPDF of two variables X and Y is:

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

Compute:

a) $P\{X > 1, Y < 1\}$, b) $P\{X < Y\}$, c) $P\{X < a\}$

Solution

a) the computations are quite straightforward:

$$
\begin{aligned}
P\{X > 1, Y < 1\} &= \int_0^1 \left[\int_1^{+\infty} f(x,y)dx\right] dy \\
&= 2\int_0^1 e^{-2y}\left[\int_1^{+\infty} e^{-x}dx\right] dy \\
&= 2 \cdot \frac{1}{e} \cdot \int_0^1 e^{-2y} dy \\
&= \frac{2}{e}\left[-\frac{1}{2}e^{-2y}\right]_0^1 \\
&= \frac{1}{e}(1 - e^{-2})
\end{aligned}
$$

b) Let's use intuition first. This JPDF decreases with *both* x and y, but it decreases *faster* with y. Therefore, given two values x, y taken at random, it is more likely that $x > y$ than the opposite. Hence, we expect $P\{X < Y\} < 1/2$. Let's see what the computations tell us:

$$
\begin{aligned}
P\{X < Y\} &= \int_0^{+\infty} \left[\int_0^y f(x,y)dx\right] dy \\
&= 2 \cdot \int_0^{+\infty} e^{-2y}[-e^{-x}]_0^y dy \\
&= 2 \cdot \int_0^{+\infty} e^{-2y}[1 - e^{-y}]dy \\
&= 2 \cdot \int_0^{+\infty} (e^{-2y} - e^{-3y})dy \\
&= 2 \cdot \left[-\frac{1}{2}e^{-2y} + \frac{1}{3}e^{-3y}\right]_0^{+\infty} \\
&= 2 \cdot \left[0 - \left(-\frac{1}{2} + \frac{1}{3}\right)\right] = \frac{1}{3}
\end{aligned}
$$

c) again using the same procedure:

$$
\begin{aligned}
P\{X < a\} &= \int_0^a \left[\int_0^{+\infty} f(x,y)dy\right] dx \\
&= 2 \cdot \int_0^a e^{-x}\left[-\frac{1}{2}e^{-2y}\right]_0^{+\infty} dx \\
&= \int_0^a e^{-x}dx \\
&= 1 - e^{-a}
\end{aligned}
$$

Note that the result is the CDF of X, i.e. $F_X(a)$ (whether the inequality is weak or strong in the definition does not really matter for continuous RVs, unless the F has discontinuities, which it has not).

♦

3.4.3 Joint distributions of *n* random variables

The above definitions, which we have introduced for systems of 2 RVs, can be extended to the case of n RVs $X_1, X_2, \ldots, X_n$. I can define:

$$F(x_1, \ldots, x_n) = P\{X_1 \leq x_1, \ldots, X_n \leq x_n\}$$

And readily obtain the CDF for the single RV X_i as:

$$F_{X_i}(x_i) = F(+\infty, \ldots, +\infty, x_i, +\infty, \ldots, +\infty)$$

Etc. etc.

3.5 Independent random variables

Two RVs *X, Y* are **independent if and only if:**

$$F(x, y) = F_X(x)F_Y(y)$$

Which means that their JCDFs is the product of the single CDFs. This is equivalent to saying that $\{X \leq x\}, \{Y \leq y\}$ are **independent events,** for every possible values x, y. In fact, consider that:

$$\begin{aligned} F(x, y) &= P\{X \leq x, Y \leq y\} \\ &= P\{X \leq x | Y \leq y\} \cdot P\{Y \leq y\} \end{aligned}$$

But, if $\{X \leq x\}, \{Y \leq y\}$ are independent events, then we have:

$$\begin{aligned} &P\{X \leq x | Y \leq y\} \cdot P\{Y \leq y\} \\ &= P\{X \leq x\} \cdot P\{Y \leq y\} = F_X(x) \cdot F_Y(y) \end{aligned}$$

If two RVs are independent, it follows that:

- If they are **discrete**, $p(x, y) = p_X(x) \cdot p_Y(y)$. The JPMF is the product of the single PMFs.
- If they are **continuous**, $f(x, y) = f_X(x) \cdot f_Y(y)$. The JPDF is the product of the single PDFs.

This obviously generalizes to n RVs, $X_1, X_2, \ldots, X_n$. These are independent **if and only if**:

$$F(x_1, \ldots, x_n) = \prod_{i=1}^{n} F_{X_i}(x_i)$$

Exercise

Let *X, Y* be two **independent** continuous RVs, whose PDFs are the same and the following:

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

Compute the PDF of $Z = X/Y$.

Solution

Since $X \in [0, +\infty)$ and $Y \in [0, +\infty)$, then $Z \in [0, +\infty)$. Define a subset of $\mathbb{R}^2$ $C_k = \{(x, y) | \, x/y \leq k\}$. We obtain that $F_Z(k)$=$P\{(X, Y) \in C_k\}$. This last can be computed as:

$$P\{(X,Y) \in C_k\} = \iint_{C_k} f(x,y)dxdy$$

By the **independence hypothesis** we observe that:

$$\begin{aligned}
F_Z(k) &= \int_0^{+\infty} \left[\int_0^{ky} (e^{-x}) \cdot (e^{-y})dx\right] dy \\
&= \int_0^{+\infty} (e^{-y}) \cdot \left[\int_0^{ky} (e^{-x})dx\right] dy \\
&= \int_0^{+\infty} (e^{-y}) \cdot [1 - e^{-ky}]dy \\
&= \int_0^{+\infty} \left(e^{-y} - e^{-(k+1)y}\right)dy \\
&= \left[-e^{-y} + \frac{e^{-(k+1)y}}{k+1}\right]_0^{+\infty} \\
&= 1 - \frac{1}{k+1}
\end{aligned}$$

We have just computed the CDF of the RV of interest (you can check that it verifies all the properties of a CDF). Since it is **continuous and differentiable** $\forall k \geq 0$, we can compute the PDF that is required by the exercise:

$$f_Z(k) = \frac{\partial}{\partial k} F_{X/Y}(k) = \frac{1}{(k+1)^2}$$

♦

Exercise

Given *n* RVs $X_1, \ldots, X_n$, **iid (independent and identically distributed)**, whose CDFs are $F(a)$, compute the CDF of the following two variables: $M = \max\{X_1, \ldots, X_n\}$ and $L = \min\{X_1, \ldots, X_n\}$.

Solution

For RV *M*, we can observe the following

$$\begin{aligned}
&F_M(a) = P\{\max\{X_1, \ldots, X_n\} \leq a\} \\
&= P\{X_1 \leq a, X_2 \leq a, \ldots, X_n \leq a\} \\
&= \prod_{i=1}^{n} P\{X_i \leq a\} \\
&= [F(a)]^n
\end{aligned}$$

Where the third passage is due to the assumption of independence.

For the *minimum L*, the reasoning is similar. It is easier if we go through the longer route:

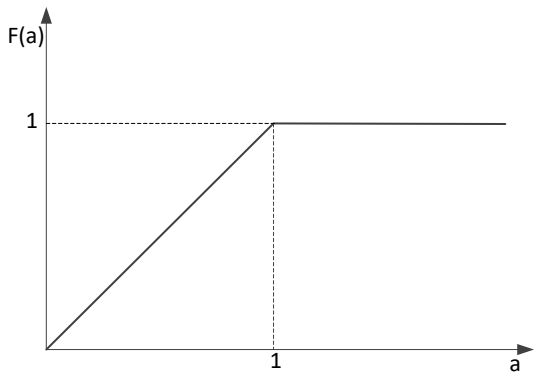
$$
\begin{aligned}
&1 - F_L(a) = P\{\min\{X_1, \dots, X_n\} > a\} \\
&P\{X_1 > a, X_2 > a, \dots, X_n > a\} \\
&= \prod_{i=1}^{n} P\{X_i > a\} \\
&= [1 - F(a)]^n
\end{aligned}
$$

From which we immediately obtain: $F_L(a) = 1 - [1 - F(a)]^n$.

Let's take a closer look at those formulas in order to find a **physical explanation**. Always recall that we are in the business of finding *explanations*, not formulas.

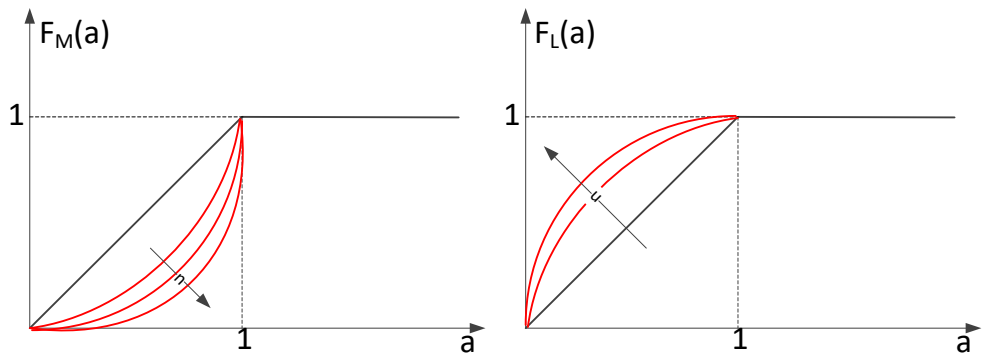
Assume that $F(a)$ is the following

$$
F(a) = \begin{cases} a & 0 \le a < 1 \\ 1 & a \ge 1 \\ 0 & a < 0 \end{cases}
$$



Then it is easy to see the following:

$$
F_M(a) = \begin{cases} a^n & 0 \le a < 1 \\ 1 & a \ge 1 \end{cases}, \quad F_L(a) = \begin{cases} 1 - [1 - a]^n & 0 \le a < 1 \\ 1 & a \ge 1 \end{cases}
$$



As n grows large, **the two distributions tend to a *step function*,** respectively in 1 (maximum) and 0 (minimum). There is a physical explanation for this. The n variables are *independent* and *uniformly distributed*. The fact that their PDF is uniform can be observed by deriving $F(a)$ (you get a constant function in $[0,1]$. If you take one sample at random, it can be anywhere in $[0,1]$. If you take n samples, there is an increasing probability that

- The **highest** sample will be near 1
- The **lowest** sample will be near 0.

You can observe the same phenomenon **with discrete RVs as well**. If you throw a die, the probability that you get a 6 is 1/6. **If you throw a die 1000 times**, the probability that the *maximum* that you get in 1000 throws is a 6 is almost equal to 1.

The interesting thing is that this phenomenon does **not depend on the shape of** $F(a)$, provided that there exist two **finite values** $\boldsymbol{a_L, a_M}$ such that $F(a_L) = 0, F(a_M) = 1$ (otherwise the steps move infinitely to the left and to the right respectively).

♦

3.6 Mean value

The **mean value** (or **expectation**, or **expected value**) of a RV X is denoted with $E[X]$ and computed as follows:

- **Discrete RV**: $E[X] = \sum_i x_i \cdot p(x_i)$
- **Continuous RV**: $E[X] = \int_{-\infty}^{+\infty} x \cdot f(x)dx$

In the discrete case, it is the **weighted sum** of each value that the RV can take on, with the weights being given by the probability of that value. For the continuous case, it is slightly trickier to visualize, but a similar concept holds.

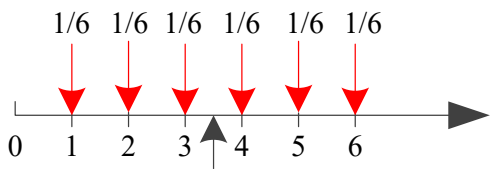
In the discrete case, the PMF owes its very name to the fact that the expression of the mean value is akin to that of the **center of mass**. Assume that you have an axis, where weights equal to $p(x_i)$ are set in position x_i. In that case, $E[X]$ represents the point where the axis is in equilibrium.

Example

Let us compute the mean value for a six-faced die:

$$E[X] = \frac{1}{6} \cdot (1 + 2 + 3 + 4 + 5 + 6) = \frac{21}{6} = \frac{7}{2}$$

Note that, for discrete RVs, $\boldsymbol{E[X]}$ **is not necessarily one of the values assumed by the RV**, as this case clearly shows



♦

Example

An interesting (and often useful) discrete RV is the **indicator variable** for an event A. This is defined as follows:

$$I_A = \begin{cases} 1 & \text{if } A \text{ occurs} \\ 0 & \text{if } A \text{ does not occur} \end{cases}$$

Therefore, $p(1) = P(A)$, $p(0) = 1 - P(A)$. For this variable, we have:

$$E[I_A] = 1 \cdot p(1) + 0 \cdot p(0) = p(1) = p(A)$$

The expected value of the indicator variable is the probability that event A occurs.

♦

Example

Compute the mean value of the RV having the following PDF (already seen in a previous lesson):

$$f(x) = \begin{cases} \frac{3}{8} \cdot (4x - 2x^2) & 0 \le x \le 2 \\ 0 & otherwise \end{cases}$$

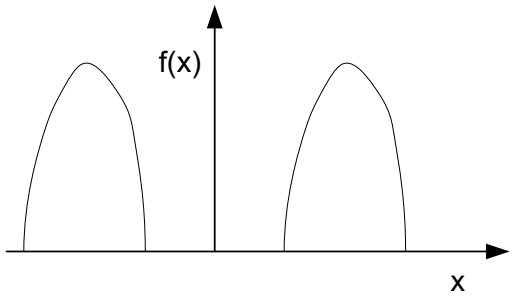
If you remember the shape of the PDF, you'll have few difficulties to observe that the mean value can only be **one**, by symmetry. Whenver the PDF is symmetric, its axis of symmetry is also the mean value. The computations should confirm this:

$$\begin{aligned} E[X] &= \int_0^2 x \cdot f(x) \ dx \\ &= \frac{3}{8} \cdot \int_0^2 x \cdot (4x - 2x^2) \ dx \\ &= \frac{3}{8} \cdot \int_0^2 (4x^2 - 2x^3) \ dx \\ &= \frac{3}{8} \cdot \left[\frac{4}{3}x^3 - \frac{2}{4}x^4\right]_0^2 \\ &= \frac{3}{8} \cdot \left[\frac{32}{3} - \frac{32}{4}\right] \\ &= \frac{3}{8} \cdot \frac{32}{12} = 1 \end{aligned}$$

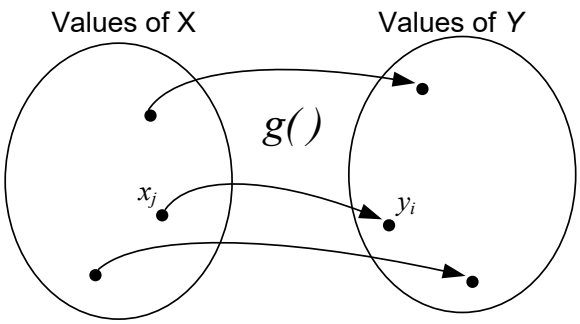
♦

In this case (a continuous RV), the mean value is the one for which you have the **maximum probability density.** In general, this is not the case. Worse yet, same as – for discrete RVs – not necessarily $E[X]$ is a value taken by the RV, **for continuous RVs it is not even true that $\boldsymbol{f(E[X]) > 0}$**, in general. In other words, $E[X]$ **is not necessarily one value that we are ever going to observe in practice** (or a value around which it is likely that observable values will coalesce).

For continuous RVs, it is enough to take an $f(\quad)$ which is symmetric w.r.t. the ordinate axis and null in the origin. Straightforward symmetry considerations are enough to convince anyone that $E[X]$ is in the origin, but $f(0) = 0$, so you will never see anything close to zero in an experiment.



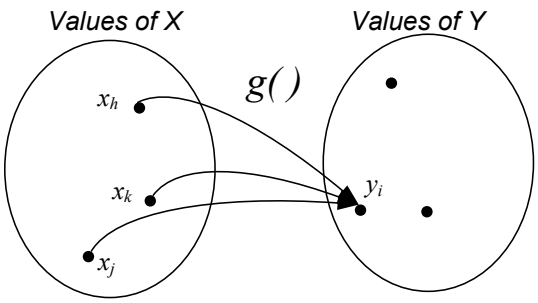
Take a discrete RV X (it works for continuous RVs as well), and assume that RV Y is a function of X, $Y = g(X)$. Assume initially that $g(\quad)$ is **injective**. This assumption will be removed later on, and it only serves the purpose of simplifying the exposition initially.



It is clear that $p_Y(y_i) = p_X(x_j)$ for any pair of values x_j,y_i such that $y_i = g(x_j)$. Hence, the mean value of Y can be computed as:

$$E[Y] = \sum_i y_i \cdot p_Y(y_i) = \sum_j g(x_j) p_X(x_j) = E[g(X)]$$

If we remove the initial hypothesis that the function is **injective**, we obtain the probability that RV Y takes on value y_i is the **sum of the probabilities of all the values x that function** $g(\quad)$ **maps into** y_i, i.e. $p_Y(y_i) = \sum_{j:g(x_j)=y_i} p_X(x_j)$



Hence, I can compute the mean value of Y again using the same formula as before:

$$E[Y] = \sum_i y_i \cdot p_Y(y_i) = \sum_i y_i \cdot \left[\sum_{j:g(x_j)=y_i} p_X(x_j)\right] = \sum_j g(x_j)p_X(x_j) = E[g(X)]$$

In other words, the mean value of **whatever function of a random variable** is computed by weighing the function values by the probability that they occur.

The same holds for **continuous RVs as well** (we are not going to prove it):

$$E[g(X)] = \int_{-\infty}^{+\infty} g(x) \cdot f(x)dx$$

Let us show some examples:

Example

Consider the discrete RV *X*, with $p(0) = 0.2$, $p(1) = 0.5$, $p(2) = 0.3$. Compute $E[X^2]$

RV X^2 has values 0, 1 e 4 with probability 0.2, 0.5 e 0.3 respectively (function $g(\quad)$ is in this case injective on the domain). Thus, we obtain:

$$E[X^2] = 0 \cdot 0.2 + 1 \cdot 0.5 + 4 \cdot 0.3 = 1.7$$

♦

Example

Continuous RV *X* is the time-to-repair (expressed in hours) of a failure in a power plant. Its PDF is:

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

This means that all the failures are repaired within one hour.

The cost of a failure is proportional to the *third power* of the time-to-repair. Compute the mean cost of a breakdown. We just need to apply the formula:

$$E[X^3] = \int_{-\infty}^{+\infty} x^3 \cdot f(x)dx = \int_0^1 x^3 dx = \left[\frac{1}{4}x^4\right]_0^1 = \frac{1}{4}$$

♦

Given that the way of computing the mean value of a function of a RV is true for **whatever function**, you readily obtain the following:

$$E[aX + b] = a \cdot E[X] + b$$

You just need to define $Y = g(X) = a \cdot X + b$ and apply a well-known property. In other words, given any RV (whether discrete or continuous), **scaling** its values by a constant corresponds to scaling analogously the mean value. Furthermore, **adding a constant offset** to the values offsets the mean value as well. The proof is due to the linearity of sum/integral operators, and it is left as an exercise.

The mean value of an RV is also called **first-order moment** of that variable. In general, the **nth-order moment** of a RV is $E[X^n]$. The 2nd-order moment is called **mean square value**.

Finally, note that the mean value is **dimensionally coherent** with the values of the RV. If the RV's values are – say – square meters, then so is the mean value.

3.6.1 Expectation of the sum of RVs

We have shown the former property for functions of **one** RV. It holds for functions of **two (*n*) random variables,** provided that we know their **joint** PMF/PDF.

In general:

$$E[g(X,Y)] = \begin{cases} \sum\sum g(x,y)\cdot p(x,y) & \text{discreteRVs} \\ \int\int g(x,y)f(x,y)dxdy & \text{continuousRVs} \end{cases}$$

A particular case is when $g(X,Y) = X + Y$ (**sum of two random variables**). In this case we have (we give the proof for the discrete case):

$$\begin{aligned} E[X+Y] &= \sum_x\sum_y (x+y)p(x,y) \\ &= \sum_x\sum_y x\cdot p(x,y) + \sum_x\sum_y y\cdot p(x,y) \\ &= \sum_x x\cdot\sum_y p(x,y) + \sum_y y\cdot\sum_x p(x,y) \\ &= \sum_x x\cdot p_X(x) + \sum_y y\cdot p_Y(y) \\ &= E[X] + E[Y] \end{aligned}$$

The proof is almost identical for the continuous case, *mutatis mutandis*. In general, we have the following property:

the mean value of the sum of *n* random variables (however distributed)
is the <u>sum</u> of the mean values of the single variables

In other words, $\sum$ **and** $\boldsymbol{E}[\quad]$ **commute**.

Note that we haven't required any hypothesis on the independence of the two (*n*) RVs. The result holds in the **most general case**, even if RVs are not independent.

Exercise

A secretary prints N letters, that have to be inserted into N sorted envelopes. A gust of wind blows them on the floor, in a random order, so that she does not know anymore which letter goes into which envelope.

Suppose that she inserts letters **randomly** into envelopes. What is the mean value of the number of letters correctly paired with their envelope?

Solution

Deifne events A_i as "the *i*-th letter goes into the *i*-th envelope". It is $P(A_i) = 1/N$, and it does not depend on the index *i*. Now define RV X_i, the indicator variable for event A_i. It is: $p_{X_i}(1) = p(A_i) = 1/N = E[X_i]$, by definition. Now, We need to count how many letters are in the correct envelope, hence we need to sum up all the indicator RVs, i.e. we need to define $Y = \sum_{i=1}^{N} X_i$, and compute $E[Y]$=$E[\sum_{i=1}^{N} X_i]$. However, we know that:

$$E[\sum_{i=1}^{N} X_i] = \sum_{i=1}^{N} E[X_i] = N \cdot E[X_i] = N \cdot 1/N = 1$$

On average, **whatever the number *N*, one letter only will find the right envelope.** If this seems surprising, note that:

- If $N = 1$, then the result is obvious (there is nothing random in this experiment)
- If $N = 2$, you get to the same result through another route: either *both* letters are in the right envelope, or *neither* are. In the first case, the number of letter correctly filed is equal to 2. In the second case, it is equal to 0. The two outcomes are equally likely, hence the mean value is again 1.

Similar reasoning can be used for higher values of N. Note, by the way, that events A_1, A_2 are *not* independent (when $N = 2$), hence RVs X_1, X_2 are not independent either. However, expectation and summation can commute even when RVs are not independent

♦

Exercise

Assume that coupons can be of *n* different types, with identical probability, and that you pick *k* of them. Compute the mean value of RV X, which is the number of <u>different</u> types included in a set of *k* coupons.

Solution

Again, we need to *count* the types of coupons. This means that it is useful to have indicator variables. Define the following RVs:

$$X_i = \begin{cases} 1 & \text{at least one type} - \text{i coupon in the set of k} \\ 0 & \text{otherwise} \end{cases}$$

It is not difficult to observe that $E[X] = E[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} E[X_i] = n \cdot E[X_i]$.

In order to compute $E[X_i]$, it is enough to write down the definition:

$$\begin{aligned} E[X_i] &= 1 \cdot P\{X_i = 1\} + 0P\{X_i = 0\} \\ &= P\{X_i = 1\} \\ &= 1 - P\{X_i = 0\} \\ &= 1 - \left(\frac{n-1}{n}\right)^k \end{aligned}$$

Each of the k coupons has independent probability of being of one type.

Then the correct result is:

$$E[X] = n \cdot \left[1 - \left(\frac{n-1}{n}\right)^k\right]$$

♦

Exercise

This exercise shows a property that will be useful later on.

Prove that, if X and Y are two *independent* RVs, then $E[XY] = E[X] \cdot E[Y]$

Solution

We give the proof for the discrete case (the one for the continuous case is left as an exercise).

$$\begin{aligned} E[XY] &= \sum_x \sum_y x \cdot y \cdot p(x, y) \\ &= \sum_x \sum_y x \cdot y \cdot p_X(x) \cdot p_Y(y) \\ &= \sum_x x \cdot p_X(x) \cdot \sum_y y \cdot p_Y(y) \\ &= E[X] \cdot E[Y] \end{aligned}$$

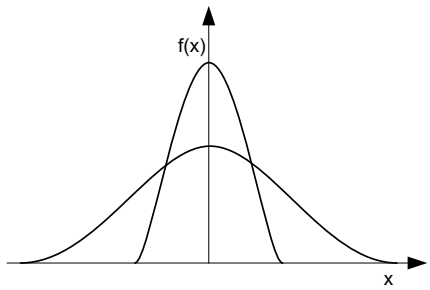
This is due to independence

♦

3.7 Variance

The **mean value** allows one to **summarize economically** as much information as possible on a given distribution. The mean value, alone, is a good indicator (it is the minimum-error predictor for the value of a RV). However, it is often insufficient.

We are interested in determining **how dispersed** the values of a RV are around their mean value. The two densities reported in the figure have the same mean (zero), but in the second case large-modulus value are more likely than in the first.



The measure of **how much a RV is dispersed around its mean value** is called **variance**, and it is defined as:

$$Var(X) = E[(X-\mu)^2]$$

A question which comes out naturally is **why the square**. First of all, we need something to **cancel out signs**, otherwise defects and surpluses would compensate. It is in fact obvious that:

$$E[X-\mu] = E[X] - \mu = 0.$$

In order to dispense with the sign you need **either a modulus or an even power**. The even power is preferable because it is analytically more tractable (it preserves differentiability, for instance).

Very often in practice, variance is computed as follows:

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

In other words, variance is the **mean square value** minus the **square of the mean value.**

Note that the variance (unlike the mean value) **is not dimensionally identical to the RV values.** For instance, if the RV – dimensionally speaking – is a time, its variance is a time squared. If needed, one can compute the **standard deviation**, defined as:

$$StDev(X) = \sqrt{Var(X)}$$

Let us compute some simple variances.

Example

Compute the variance for a six-faced die

$$E[X] = \frac{1}{6} \cdot (1+2+3+4+5+6) = \frac{21}{6} = \frac{7}{2}$$

Already computed

$$E[X^2] = \frac{1}{6} \cdot (1+4+9+16+25+36) = \frac{91}{6}$$

Hence, with the above formula, we get:

$$Var(X) = E[X^2] - E[X]^2 = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{91}{6} - \frac{49}{4} = \frac{182 - 147}{12} = \frac{35}{12}$$

♦

Example

Let us compute the variance for the **indicator RV.** For this, we remind that it is $E[I_A] = P(A)$. If we develop the computations, we obtain

$$\begin{aligned} Var(I_A) &= E\left[{I_A}^2\right] - E[I_A]^2 \\ &= E[I_A] - E[I_A]^2 \\ &= E[I_A] \cdot (1 - E[I_A]) \\ &= P(A) \cdot [1 - P(A)] \end{aligned}$$

Since ${I_A}^2 = I_A$

Which tells us that the variance is maximum when event A has a probability equal to ½.

♦

Example

Compute the variance for the RV characterized by the following PDF (already encountered before).

$$f(x) = \begin{cases} \frac{3}{8} \cdot (4x - 2x^2) & 0 \le x \le 2 \\ 0 & altrimenti \end{cases}$$

We remember that the mean value is **one** (by symmetry, and we also did the computations). For the variance, we have:

$$\begin{aligned} E[X^2] &= \int_0^2 x^2 \cdot f(x) \ dx \\ &= \frac{3}{8} \cdot \int_0^2 (4x^3 - 2x^4) \ dx \\ &= \frac{3}{8} \cdot \left[\frac{4}{4}x^4 - \frac{2}{5}x^5\right]_0^2 \\ &= \frac{3}{8} \cdot \left[16 - \frac{64}{5}\right] \\ &= \frac{3}{8} \cdot \frac{16}{5} = \frac{6}{5} \end{aligned}$$

Hence:

$$Var(X) = E[X^2] - E[X]^2 = \frac{6}{5} - (1)^2 = \frac{1}{5}$$

♦

3.7.1 Variance of sums of RVs

We have seen that, for the mean value, it is $E[aX + b] = a \cdot E[X] + b$. This is because the expectation is a **linear** operator. We cannot expect the same property to hold for the variance, which is instead **quadratic**.

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

This tells us that:

- **Offsets do not matter**. If you move the distribution on the horizontal axis, then the mean value does change, but the **dispersion** around the mean value does not.
- **Rescaling does matter**: if you rescale the RV by a constant factor, you are either
  - **Compressing** the distribution (if $a < 1$)
  - **Spreading out** the distribution (if $a > 1$)

In both cases, the multiplying constant is squared, because a square is inherent in the concept of variance.

We have also seen that $\sum$ **and** $\boldsymbol{E}[\quad]$ **commute**, i.e. the mean value of the sum is the sum of mean values. **This is not true of the variance**, however. A straightforward **counterexample** is the following: $Var(X + X) = Var(2X) = 4 \cdot Var(X)$ (remember that there's a **square**).

Therefore, we need to figure out what the formula is for the variance of the sum of two variables in general.

$$
\begin{aligned}
&Var(X + Y) = E\left[\left(X + Y - (\mu_X + \mu_Y)\right)^2\right] && \text{Permute the terms} \\
&= E[(X - \mu_X + Y - \mu_Y)^2] && \text{Develop the square binomial} \\
&= E[(X - \mu_X)^2 + (Y - \mu_Y)^2 + 2(X - \mu_X)(Y - \mu_Y)] \\
&= E[(X - \mu_X)^2] + E[(Y - \mu_Y)^2] + 2E[(X - \mu_X)(Y - \mu_Y)] && \text{Distribute E[] over the sum} \\
&= Var(X) + Var(Y) + 2E[(X - \mu_X)(Y - \mu_Y)]
\end{aligned}
$$

The last term is called **covariance** of *X, Y*. The definition is $Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$.

If you develop the computations for the covariance, you find the following:

$$
\begin{aligned}
&Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] \\
&= E[X \cdot Y + \mu_X \cdot \mu_Y - Y \cdot \mu_X - X \cdot \mu_Y] \\
&= E[X \cdot Y] + E[\mu_X \cdot \mu_Y] - E[Y \cdot \mu_X] - E[X \cdot \mu_Y] \\
&= E[X \cdot Y] + \mu_X \cdot \mu_Y - \mu_X \cdot \mu_Y - \mu_X \cdot \mu_Y \\
&= E[X \cdot Y] - \mu_X \cdot \mu_Y
\end{aligned}
$$

Therefore, the covariance is the **expectation of the product** minus the **product of the expectations**.

We can observe that:

- covariance is commutative: $Cov(X,Y) = Cov(Y,X)$.
- $Cov(X,X) = Var(X)$. This is straightforward, you just need to substitute in the above expression
- If X and Y are **independent RVs**, then it is $E[X \cdot Y] = \mu_X \cdot \mu_Y$ (we learned it from a previous exercise), hence $Cov(X,Y) = 0$

This means that:

For independent RVs, the variance of the sum is the sum of the variances.

Of course, this is not true in general, since for non-independent RVs **covariances are not null**. Note that **the reverse is generally false:** there are couples of RVs which are **not independent**, and whose covariance is null.

The above formula can be generalized for the case of the **sum of *n* variables.**

$$Var\left(\sum_i X_i\right) = \sum_i Var(X_i) + \sum_i \sum_{j \neq i} Cov(X_i, X_j)$$
$$= \sum_i \sum_j Cov(X_i, X_j)$$

3.7.2 Covariance and correlation

The covariance of two variables can be **positive, negative or null**. The covariance (and, more specifically, its **sign**) has an important physical interpretation. To show this, define two events A and B, and the related indicator RVs I_A, I_B:

$$I_A = \begin{cases} 1 & \text{if } A \text{ occurs} \\ 0 & \text{if } A \text{ does not occur} \end{cases}, \qquad I_B = \begin{cases} 1 & \text{if } B \text{ occurs} \\ 0 & \text{if } B \text{ does not occur} \end{cases}$$

RV $I_A \cdot I_B$ is equal to one if and only if event AB occurs. Let us compute now $Cov(I_A, I_B)$:

$$Cov(I_A, I_B) = E[I_A \cdot I_B] - E[I_A] \cdot E[I_B]$$
$$= P(AB) - P(A) \cdot P(B)$$

Now,

- if $Cov(I_A, I_B) > 0$, then $P(AB) > P(A) \cdot P(B)$, hence
$$P(A|B) = \frac{P(AB)}{P(B)} > P(A)$$
- Conversely, if $Cov(I_A, I_B) < 0$, then $P(AB) < P(A) \cdot P(B)$, hence
$$P(A|B) = \frac{P(AB)}{P(B)} < P(A)$$

The same is also true of $P(B|A)$ (recall that covariance is symmetric, and so is intersection). This means that:

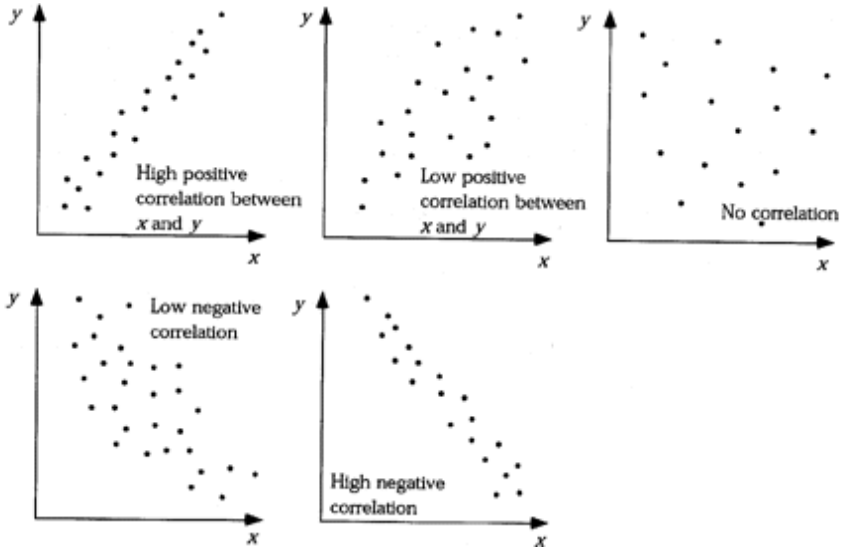
- A **positive** covariance indicates that A is **more likely** to occur if B occurs.
- A **negative** covariance indicates that A is **less likely** to occur if B occurs.

And it is fairly obvious that the two events influence each other. **If they were independent**, their covariance **would be null**.

In general, if you take 2 RVs *X, Y*, then a **positive covariance** means that large values of *X* often come together with large values of *Y*. A **negative** covariance instead means that large values of *X* often come together with small values of *Y* instead. A normalized measure of the above effect is called **correlation** (sometimes "Pearson's correlation coefficient"), and is the following:

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

It can be shown (you can do it as an exercise) that $-1 \leq Corr(X,Y) \leq +1$. When the correlation is positive, the higher it is the more likely it is that you find high values of *X* paired with high values of *Y*, etc. etc.



Basically, a positive correlation means that "high calls for high", i.e. the two RVs X, Y tend to co-occur on the same side of the respective mean value. A negative correlation means instead that "high calls for low", i.e. the two RVs X, Y tend to co-occur on the opposite sides of the respective mean value. The absolute magnitude of the correlation coefficient reveals how much one RV can be predicted from the other. If it is equal to one, then it means that $X = \mathrm{a} \cdot Y + b$, i.e., one is a linear scaling of the other.

Correlation is invariant under linear scaling of the RVs, i.e. $Corr(X,Y) = Corr(a \cdot X + b, \mathrm{c} \cdot Y + d)$, with $a, c > 0$.

4 Special random variables

Some RVs are often encountered in practice. This happens because **they are good models of some aspects of reality** (recall that this is a course about modeling). We now review them, and we observe that we have already met most of them during the previous lectures.

4.1 Discrete distributions

4.1.1 Bernoulli distribution

A Bernoullian RV is a **discrete** RV, which is equal to 1 with probability *p* and to 0 with probability (1-*p*). It is the same thing as an **indicator variable** for an event whose probability is *p*. Bernoullian experiments are those with a binary outcome, often referred to as **success/failure**.

As we have already seen, its parameters are:

- Mean value: $E[X] = p$
- Variance: $Var(X) = p \cdot (1-p)$

Which shows that the maximum uncertainty is achieved when $p = 1/2$.

4.1.2 Binomial distribution

This one is obtained from the Bernoullian, and it represents the **number of successes in *n* repeated trials in independent conditions**, where each trial has a probability of success equal to *p*.

Thus, a binomial distribution is a **discrete one**, characterized by **two parameters**, *n* and *p*. Computing its PMF is rather simple:

$$p(i) = P\{X = i\} = \binom{n}{i} p^i \cdot (1-p)^{n-i}$$

The fact that the one above is indeed a PMF is confirmed by the **normalization condition**, which is easy to test thanks to Newton's binomial formula (hence the name of the distribution):

$$\sum_{i=0}^{n} p(i) = \sum_{i=0}^{n} \binom{n}{i} p^i \cdot (1-p)^{n-i} = [p + (1-p)]^n = 1$$

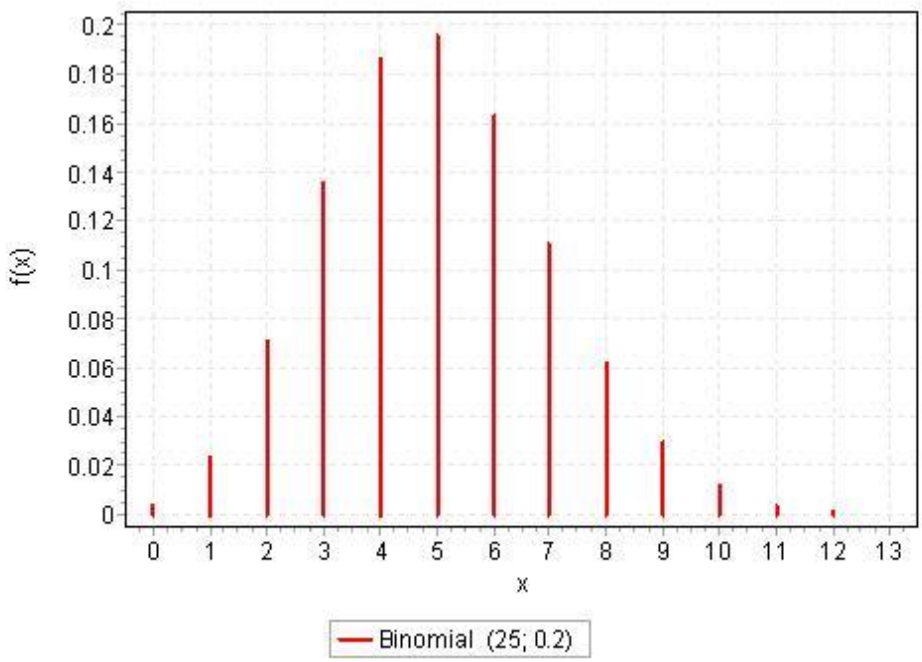
We have already met this RV in a previous exercise.

Let us compute the **mean value and the variance**. In order to do so, it is enough to observe that a binomial variable is the **sum of *n* iid Bernoullian RVs**, hence:

- $E[X] = E[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} E[X_i] = n \cdot p$
- Since they are independent, I can sum the variances of the single Bernoullian variables:

$$Var(X) = Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i) = n \cdot [p \cdot (1-p)]$$

Let us see what a binomial PMF looks like:



The following observations are in order

- It takes on positive values **between 0 and *n*** (and it is null outside that interval)
- It is **symmetric** *only* if p=0.5. In fact, in this case it can be easily shown that $p(i) = p(n-i)$ (since $\binom{n}{i} = \binom{n}{n-i}$). Otherwise it is skewed either to the left or to the right, depending on whether p<0.5 or vice versa.
- It has its maximum value ***around*** *np* (the latter number may not be an integer)

A fairly obvious property of binomial variables is the following:
If $X_1 \sim (n_1, p)$ *and* $X_2 \sim (n_2, p)$ *are two* ***independent*** *RVs,then* $X_1 + X_2 \sim (n_1 + n_2, p)$.

The property is obvious since the binomial counts the number of successes in the *sum of the trials* (if the two repeated trials have the same probability of success, that is, and RVs are independent).

4.1.3 Poisson distribution

A **discrete RV** is said to be **Poissonian** (of Poisson) with parameter $\lambda > 0$ if its PMF is the following:

$$p(i) = P\{X = i\} = e^{-\lambda} \cdot \frac{\lambda^i}{i!}, \text{ for each } i \geq 0$$

Therefore, a Poissonian variable can assume **arbitrarily large values**. The one written above *is* a PMF, which can be tested using the normalization condition:

$$\sum_{i=0}^{+\infty} p(i) = \sum_{i=0}^{+\infty} e^{-\lambda} \cdot \frac{\lambda^i}{i!} = e^{-\lambda} \cdot \sum_{i=0}^{+\infty} \frac{\lambda^i}{i!} = e^{-\lambda} \cdot e^{\lambda} = 1$$

Now, the above one is an *infinite* series. Therefore, it can converge only if $\lim_{i\to\infty} \frac{\lambda^i}{i!} = 0$. This means that the **Poissonian PMF goes to zero** as *i* grows large.

Let us compute its *mean value* and *variance*. Note that the book uses a different mathematical trick, which I find harder to understand, hence my computations will be slightly different (but will get you to the same result).

$$
\begin{aligned}
E[X] &= \sum_{i=0}^{+\infty} i \cdot \left(e^{-\lambda} \cdot \frac{\lambda^i}{i!}\right) \\
&= \sum_{i=1}^{+\infty} i \cdot \left(e^{-\lambda} \cdot \frac{\lambda^i}{i!}\right) \\
&= e^{-\lambda} \cdot \lambda \cdot \sum_{i=1}^{+\infty} \cancel{i} \cdot \left(\frac{\lambda^{i-1}}{\cancel{i} \cdot (i-1)!}\right) \\
&= e^{-\lambda} \cdot \lambda \cdot \sum_{i=0}^{+\infty} \left(\frac{\lambda^i}{i!}\right) \\
&= e^{-\lambda} \cdot \lambda \cdot e^{\lambda} \\
&= \lambda
\end{aligned}
$$

Therefore, parameter $\lambda > 0$ is the **mean value** of the Poissonian variable.

Now, for the **variance**, we just apply the formula $Var(X) = E[X^2] - \mu^2 = E[X^2] - \lambda^2$. Let us compute the mean square value:

$$
\begin{aligned}
E[X^2] &= \sum_{i=0}^{+\infty} i^2 \cdot \left(e^{-\lambda} \cdot \frac{\lambda^i}{i!}\right) \\
&= e^{-\lambda} \cdot \lambda \cdot \sum_{i=1}^{+\infty} \cancel{i} \cdot i \cdot \frac{\lambda^{i-1}}{\cancel{i} \cdot (i-1)!} \\
&= e^{-\lambda} \cdot \lambda \cdot \sum_{j=0}^{+\infty} (j+1) \cdot \frac{\lambda^j}{j!} \\
&= e^{-\lambda} \cdot \lambda \cdot \left[\sum_{j=0}^{+\infty} j \cdot \frac{\lambda^j}{j!} + \sum_{j=0}^{+\infty} \frac{\lambda^j}{j!}\right] \\
&= e^{-\lambda} \cdot \lambda \cdot \left[e^{\lambda} \cdot \lambda + e^{\lambda}\right] \\
&= \lambda^2 + \lambda
\end{aligned}
$$

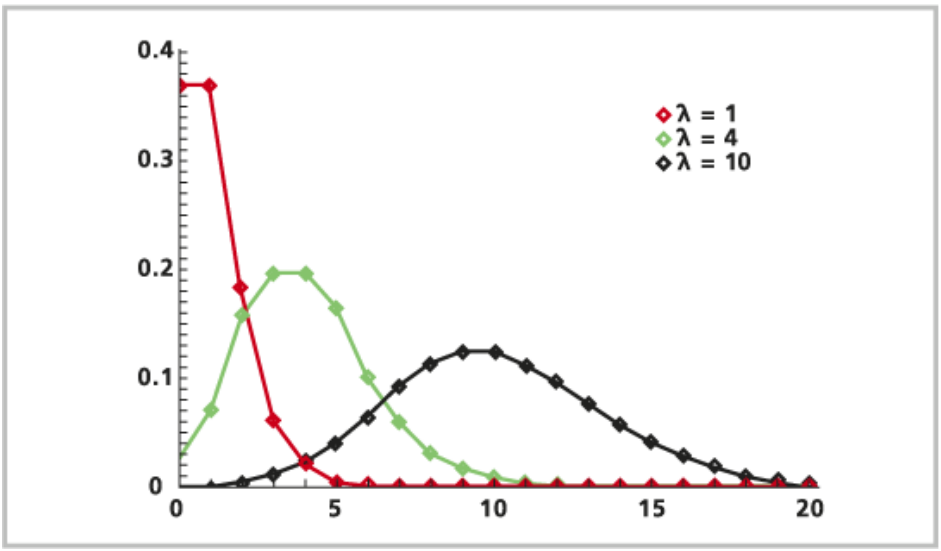
Therefore, the variance is: $Var(X) = (\lambda^2 + \lambda) - \lambda^2 = \lambda$. Hence, λ is also the **variance**.

Let us find out what a Poisson looks like. We know that $\lim_{i\to\infty} p(i) = 0$. Moreover, it is $p(0) = e^{-\lambda}$.

Now, if:

- $\lambda \leq 1$: the PMF is **decreasing** (at the numerator, we always multiply for something which is less than 1)
- $\lambda > 1$: the PMF **increases** while $i < \lambda$, and then it **decreases** when $i > \lambda$. Hence, in this case it will **peak around $\boldsymbol{\lambda}$** (which may not be an integer).

In any case, the distribution is **skewed**, and the right tail extends to infinity.



So, apart from an elegant mathematical formulation, has the **Poissonian variable any physical meaning?** The answer is **yes**, because:

The Poissonian variable approximates rather well a binomial variable with:
- **A large *n***
- **A small *p***
- $\boldsymbol{\lambda = n \cdot p}$

In other words, if we want to count the occurrence of **unlikely events in a large number of independent repeated trials**, then the Poissonian can help us. The Poissonian variable approximates rather well

- The number of typos in the page of a book
- The number of 100-year-old in a large population
- …

We could do it using a binomial, of course, but this is **more economical** from a computational standpoint. Recall that a binomial requires **large powers, binomial coefficients (with large factorials)**, etc.. If n is very large, then computing those coefficients is *hard* from a numerical standpoint (ill-conditioned numbers). Using a Poissonian can help you in this case.

The proof is easy. Take the **PMF of the binomial**, and observe that, by simply substituting $\lambda = n \cdot p$, we get:

$$P\{X = i\} = \binom{n}{i} p^i \cdot (1-p)^{n-i}$$
$$= \frac{n \cdot (n-1) \cdot \ldots \cdot (n-i+1)}{i!} \cdot \left(\frac{\lambda}{n}\right)^i \cdot \frac{(1-\lambda/n)^n}{(1-\lambda/n)^i}$$
$$= \frac{n \cdot (n-1) \cdot \ldots \cdot (n-i+1)}{n^i} \cdot \frac{\lambda^i}{i!} \cdot \frac{(1-\lambda/n)^n}{(1-\lambda/n)^i}$$

Now, if n is large, we have:

$$\frac{n \cdot (n-1) \cdot \ldots \cdot (n-i+1)}{n^i} \approx 1$$
$$(1-\lambda/n)^n \approx e^{-\lambda}$$
$$(1-\lambda/n)^i \approx 1$$

Hence we get $P\{X = i\} \approx \frac{\lambda^i}{i!} \cdot e^{-\lambda}$, which is a Poisson PMF.

Exercise

The average number of weekly accidents on a motorway segment is equal to 3. What is the probability that there will be **at least one** accident next week?

Solution

It is reasonable to think that

a) Very many cars travel on a motorway segment every week
b) The probability of an accident is quite low

If I *knew* both the number of cars and the probability of an accident, I could (theoretically) use a binomial RV (assuming that I can compute so large a factorial). The problem is that **I do not know these data**. However, I can reasonably model the problem using a Poissonian variable with a mean value $\lambda = 3$. Hence:

$$P\{X \geq 1\} = 1 - P\{X = 0\} = 1 - p(0) = 1 - e^{-3} \cdot \frac{3^0}{0!} = 1 - e^{-3} = 0.95$$

♦

An important property (which is useful in practice) of **poissonian variables** is the following:

*Take two **independent** Poissonian RVs $X_1 \sim poisson(\lambda_1), X_2 \sim poisson(\lambda_2)$. Then RV $X_1 + X_2$ is itself poissonian, with a mean $\lambda_1 + \lambda_2$.*

Proof

$$
\begin{aligned}
& p(k) = P\{X_1 + X_2 = k\} \\
& = \sum_{i=0}^{k} P\{X_1 = i, X_2 = k - i\} \\
& = \sum_{i=0}^{k} P\{X_1 = i\} \cdot P\{X_2 = k - i\} \text{ (by independence)} \\
& = \sum_{i=0}^{k} \left[\left(e^{-\lambda_1} \cdot \frac{{\lambda_1}^i}{i!} \right) \cdot \left(e^{-\lambda_2} \cdot \frac{{\lambda_2}^{k-i}}{(k-i)!} \right) \right] \left(\frac{\text{mul}}{\text{div}} \text{ by } k! \right) \\
& = \frac{e^{-(\lambda_1+\lambda_2)}}{k!} \cdot \sum_{i=0}^{k} \left[\frac{{\lambda_1}^i}{i!} \cdot \frac{{\lambda_2}^{k-i}}{(k-i)!} \cdot k! \right] \\
& = \frac{e^{-(\lambda_1+\lambda_2)}}{k!} \cdot \sum_{i=0}^{k} \left[\binom{k}{i} {\lambda_1}^i \cdot {\lambda_2}^{k-i} \right] \\
& = \frac{e^{-(\lambda_1+\lambda_2)}}{k!} \cdot (\lambda_1 + \lambda_2)^k
\end{aligned}
$$

♦

4.1.4 Geometric distribution[1]

The geometric distribution measures **number of failures before the first success** in a repeated trial experiment. Therefore, it takes values in $[0, +\infty)$, and its characteristic values (mean, variance) must depend on the trial success probability p.

Note: a different definition exists, whereby the geometric distribution counts the **of trials required to get the first success**, hence takes values in $[1, +\infty)$, and has slightly different characteristic values. Given the definition that we have assumed, it is easy to see that $P\{X = 0\} = p$. In fact, you need to have one successful trial in order to have 0 failures. Moving forward:

$$P\{X = 1\} = (1 - p) \cdot p$$

$$P\{X = 2\} = (1 - p)^2 \cdot p$$

$$\dots$$

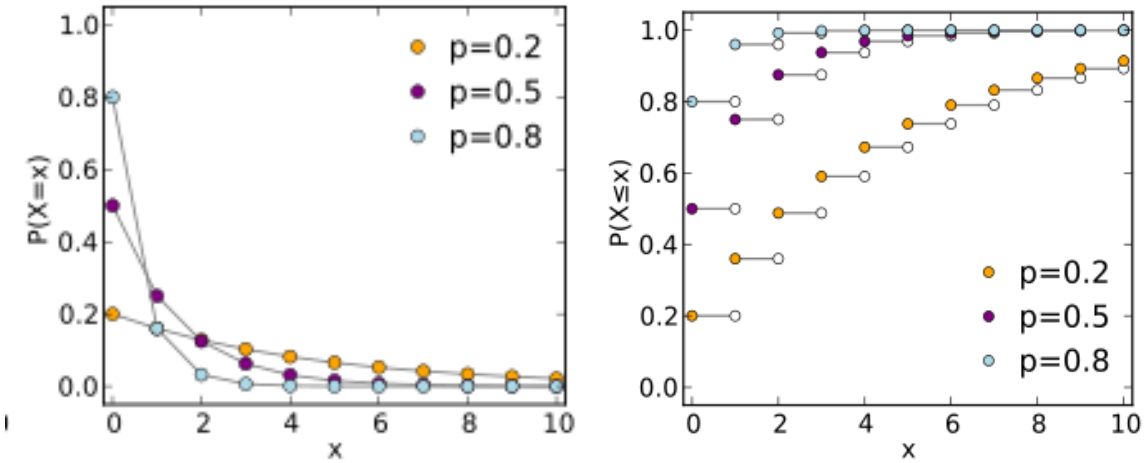
Therefore: $P\{X = k\} = (1 - p)^k \cdot p, \qquad k \geq 0$

Therefore, the PMF is a decreasing sequence, which **decays exponentially**. As far as the CDF is concerned, we can see that:

[1] Not on the Ross book

$$P(X \le k) = \sum_{i=0}^{k} (1-p)^i \cdot p$$
$$= p \cdot \frac{1-(1-p)^{k+1}}{1-(1-p)}$$
$$= 1-(1-p)^{k+1}$$

Therefore, the CDF grows exponentially, and lies more to the right if p is small.



If, instead, you use the version of the geometric that counts the number of **trials** to the first success, you have:

$$p(k) = P\{X = k\} = (1-p)^{k-1} \cdot p, \qquad k \ge 1$$
$$F(k) = P\{X \le k\} = 1-(1-p)^k, \qquad k \ge 1$$

And the plots are roughly the same.

Computing the moments is not too easy (whichever the version). For the one counting **failures**, we have:

$$E[X] = \sum_{k=0}^{+\infty} k \cdot (1-p)^k \cdot p$$
$$= p \cdot (1-p) \cdot \sum_{k=1}^{+\infty} \frac{\partial}{\partial p} - [(1-p)^k]$$
$$= -p \cdot (1-p) \cdot \frac{\partial}{\partial p} \sum_{k=1}^{+\infty} (1-p)^k$$
$$= -p \cdot (1-p) \cdot \frac{\partial}{\partial p} \left[\frac{1}{p} - 1\right]$$
$$= \frac{1-p}{p}$$

Start from 1

Bring out $p(1-p)$

Moreover, the variance can be shown to be $\sigma^2 = \frac{1-p}{p^2}$. In order to compute it, we will use a different technique, which we will introduce later.

The geometric distribution has a very important property. It is **memoryless**, and it is the only discrete distribution to have this property (we will see later on the continuous counterpart of the geometric distribution). Assuming the "failure" definition of the geometric, memoryless means that:

$$P\{X \geq n+m | X \geq n\} = P\{X \geq m\}.$$

Let us put this into context using a simple example. Assume that your experiment (e.g., getting "heads" out of a coin flip) has failed for n times, and you ask yourself what the probability is that it will fail for m **more** times before you get a success. Clearly, since coin flips are **independent trials**, you cannot gain anything by knowing the past. *The answer is independent of* ***how many times in the past*** *the experiment has failed.*

Let us prove it formally. Recall that in the "failures" version $F(k) = 1-(1-p)^{k+1}$.

$$\begin{aligned}
P\{X \geq n+m | X \geq n\} &= \frac{P\{X \geq n+m, X \geq n\}}{P\{X \geq n\}} \\
&= \frac{P\{X \geq n+m\}}{P\{X \geq n\}} = \frac{1-P\{X \leq n+m-1\}}{1-P\{X \leq n-1\}} \\
&= \frac{1-[1-(1-p)^{n+m}]}{1-[1-(1-p)^{n}]} \\
&= (1-p)^m \\
&= P\{X \geq m\}
\end{aligned}$$

Note that, if you use the "trials" version of the geometric RV, the memoryless properties still holds, but *strict* inequalitites are to be used instead of weak ones. This makes perfect sense. In fact, if you call X and Y the "failures" and "trials" version of a geometric RV, it is obviously $P\{X \geq n\} = P\{Y > n\}$.

A common mistake is sometimes made, which I now advise against. The memoryless property is sometimes mistaken for the following (wrong) statement:

$$P\{X \geq n+m | X \geq n\} = P\{X \geq n+m\}$$

This could only be true if $\{X \geq n+m\}$ and $\{X \geq n\}$ were independent, which they are not. In fact,

$$P\{X \geq n+m | X \geq n\} = \frac{P\{X \geq n+m\}}{P\{X \geq n\}} \neq P\{X \geq n+m\}.$$

For an example of usage of the geometric RV, see the Appendix at the end.

4.1.5 Probability-generating Functions

For **discrete and nonnegative** RVs (i.e., all the ones we have seen so far), we can exploit a *z-transform*, or **probability-generating function (PGF)**. For a RV X, the PGF is defined as:

$$G(z) = E[z^X] = \sum_{n=0}^{+\infty} p_n \cdot z^n,$$

where z is a complex number. The above sum converges if $|z| \leq 1$, i.e., in the unitary disc, hence G is only defined therein.

The PGF is useful for computing distribution moments. In fact, here are some of its properties:

- the **normalization condition** can be expressed as follows: $G(1) = \sum_{n=0}^{+\infty} p_n \cdot 1^n = 1$

- the **mean value** can be expressed as follows: $E[X] = \sum_{n=0}^{+\infty} p_n \cdot n = \frac{\partial}{\partial z}[\sum_{n=0}^{+\infty} p_n \cdot z^n]_{z=1} = G'(1)$
- for the same reason, we get

$$E[X(X-1)] = \sum_{n=0}^{+\infty} p_n \cdot n \cdot (n-1) = \frac{\partial^2}{\partial z^2}\left[\sum_{n=0}^{+\infty} p_n \cdot z^n\right]_{z=1} = G''(1)$$

This means that

- the **variance** can be computed as:

$$Var(X) = E[X^2] - E[X]^2 = E[X(X-1)] + E[X] - E[X]^2 = G''(1) + G'(1) - G'(1)^2$$

And so can other moments, if necessary.

Some other interesting properties are:

- **univocity**: if two RVs *X, Y* have the same PGFs, then they have the same PMFs and vice versa. This means that $\forall z, G_X(z) = G_Y(z) \Leftrightarrow \forall n, P\{X = n\} = P\{Y = n\}$. This implies that the PGF **has all the information of a PMF**.
- **convolution**: if you have *n* **independent** RVs (not necessarily identical) $X_1, \ldots, X_n$, and you know their PGFs $G_1(z), \ldots, G_n(z)$, then you can easily get the PGF of their **sum** $S = X_1 + \ldots + X_n$ as:

$$\begin{aligned} G_S(z) &= E[z^{X_1+\ldots+X_n}] = E[z^{X_1} \cdot z^{X_2} \cdot \ldots \cdot z^{X_n}] \\ &= E[z^{X_1}] \cdot E[z^{X_2}] \cdot \ldots \cdot E[z^{X_n}] \\ &= G_1(z) \cdot G_2(z) \cdot \ldots \cdot G_n(z) \end{aligned}$$

That is, you just **multiply their PGFs**. The middle passage is due to independence, of course.

One can always obtain the PMF by tweaking a PGF. In fact, it is:

$$p(k) = \frac{G^{(k)}(0)}{k!}$$

Which is useful in practice to compute the first few values of the PMF, which are often the most significant.

Here are some examples of PGFs for the discrete RVs that we have encountered so far:

- **Bernoulli**: $p(0) = 1 - p, \quad p(1) = p$. Thus, $G(z) = (1-p) \cdot z^0 + p \cdot z^1 = 1 - p + p \cdot z$
- **Binomial:** $p(i) = \binom{n}{i} p^i \cdot (1-p)^{n-i}$. Thus,

$$G(z) = \sum_{i=0}^{n} \binom{n}{i} p^i \cdot (1-p)^{n-i} \cdot z^i = [p \cdot z + (1-p)]^n$$

This one could have been found more quickly using convolution, by reasoning that the binomial is indeed the sum of *n* independent Bernoullians.

- **Poisson:** $p(i) = e^{-\lambda} \cdot \frac{\lambda^i}{i!}$. Thus, $G(z) = \sum_{i=0}^{+\infty} e^{-\lambda} \cdot \frac{\lambda^i}{i!} \cdot z^i = \frac{e^{-\lambda}}{e^{-\lambda z}} \cdot \cancel{\sum_{i=0}^{+\infty} e^{-\lambda z} \cdot \frac{(\lambda z)^i}{i!}} = e^{-\lambda+\lambda z}$
- **Geometric ("failures"):** $p(i) = (1-p)^i \cdot p$. Thus, $G(z) = \sum_{i=0}^{+\infty} (1-p)^i \cdot p \cdot z^i = \frac{p}{1-z\cdot(1-p)}$

We can use the above examples and theorems to compute **moments that we have already found to be difficult** to compute. Here are some examples:

For the **Poisson** distribution, computing the variance is **lengthy, at best**. Here is how to do it:

$$G(z) = e^{-\lambda+\lambda z} = e^{-\lambda} \cdot e^{\lambda z}$$

$G'(z) = \lambda \cdot e^{-\lambda} \cdot e^{\lambda z}$, hence $G'(1) = \lambda$

$G"(z) = \lambda^2 \cdot e^{-\lambda} \cdot e^{\lambda z}$, hence $G"(z) = \lambda^2$

Thus,

$$\sigma^2 = G"(1) + G'(1) - G'(1)^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

For the **geometric** distribution, we gave the mean for granted without computing it. The computation is instead very easy if you switch to the PGF:

$$G'(z) = \frac{\partial}{\partial z} \frac{p}{1 - z \cdot (1-p)} = p \cdot \{-[1 - z \cdot (1-p)]^{-2}\} \cdot [-(1-p)] = \frac{p(1-p)}{[1 - z \cdot (1-p)]^2}$$

$$G"(z) = \frac{\partial^2}{\partial z^2} \frac{p}{1 - z \cdot (1-p)} = \frac{\partial}{\partial z} \frac{p \cdot (1-p)}{[1 - z \cdot (1-p)]^2} = p \cdot (1-p) \cdot \{-2 \cdot [1 - z \cdot (1-p)]^{-3}\} \cdot [-(1-p)]$$
$$= \frac{2 \cdot p \cdot (1-p)^2}{[1 - z \cdot (1-p)]^3}$$

Hence $G'(1) = \frac{p(1-p)}{[1-(1-p)]^2} = \frac{1-p}{p}$, and $G"(1) = \frac{2\cdot p\cdot(1-p)^2}{[1-(1-p)]^3} = \frac{2\cdot(1-p)^2}{p^2}$

From these, we obtain:

$$\begin{aligned}
\sigma^2 &= G"(1) + G'(1) - G'(1)^2 \\
&= \frac{2 \cdot (1-p)^2}{p^2} + \frac{1-p}{p} - \left(\frac{1-p}{p}\right)^2 \\
&= \frac{2(1-p)^2 + p \cdot (1-p) - (1-p)^2}{p^2} \\
&= \frac{1-p}{p^2}
\end{aligned}$$

4.2 Continuous distributions

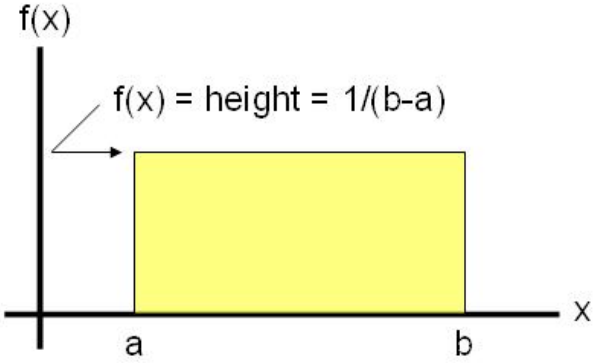
We now move to defining the most important *continuous* distributions.

4.2.1 Uniform distribution

A **continuous RV** is said to be **uniformly distributed** if its PDF is **constant over an interval** $[a, b]$, i.e.

$$f(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & otherwise \end{cases}$$

And we write $X \sim U(a, b)$.



Uniform RVs model the case when there is no particular preference for one value or another within an interval (or within a set, for the discrete version of the uniform variable). As far as the mean value and the variance are concerned, we have the following:

$$E[X] = \int_a^b \frac{1}{b-a} \cdot x \cdot dx = \frac{1}{b-a} \cdot \left[\frac{1}{2}x^2\right]_a^b = \frac{1}{b-a} \cdot \left(\frac{1}{2}b^2 - \frac{1}{2}a^2\right) = \frac{1}{2}(b+a)$$

In other words, the **mean value is the middle point of the interval**, which is what we could expect by symmetry. As for the variance, we have:

$$E[X^2] = \int_a^b \frac{1}{b-a} \cdot x^2 \cdot dx = \frac{1}{b-a} \cdot \left[\frac{1}{3}x^3\right]_a^b = \frac{1}{b-a} \cdot \left(\frac{1}{3}b^3 - \frac{1}{3}a^3\right) = \frac{1}{3}(b^2 + ab + a^2)$$

Hence:

$$\boldsymbol{Var(X) = E[X^2] - \mu^2 = \frac{1}{3}(b^2 + ab + a^2) - \frac{1}{4}(b+a)^2 = \frac{(b-a)^2}{12}}$$

It is also easy to compute the CDF of a uniform variable:

$$F(x) = \int_a^x \frac{1}{b-a} dx = \frac{x-a}{b-a} \text{ (if } a \le x \le b).$$

4.2.2 Exponential distribution

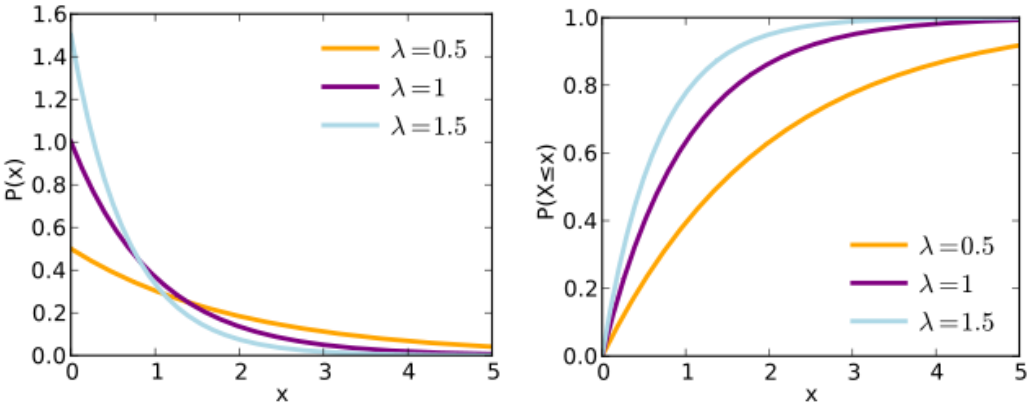
A continuous RV is said to be **(negative) exponential** with a **rate** $\lambda > 0$ if it has the following PDF:

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

One can easily get the CDF by integration:

$$F(x) = \int_0^x \lambda \cdot e^{-\lambda y} dy = \lambda \left[-\frac{1}{\lambda} \cdot e^{-\lambda y} \right]_0^x = 1 - e^{-\lambda x} \text{ (for } x \geq 0\text{, of course).}$$

The CDF approaches one **asymptotically**.



This means that the exponential variable can assume **arbitrarily large values** with some non-null probability. Exponential RVs are good models of the **amount of time** between events that occur at random (e.g., earthquakes, failures, etc.).

In order to compute the expectation and the variance, we need to solve **integrals by parts**. They are quite boring, so we will just skip the computations. There is a quicker way around them, which we will see later on.

$$E[X] = \int_0^{+\infty} x \cdot \lambda \cdot e^{-\lambda x} \cdot dx = \frac{1}{\lambda}, E[X^2] = \int_0^{+\infty} x^2 \cdot \lambda \cdot e^{-\lambda x} \cdot dx = \frac{2}{\lambda^2}$$

Hence

$$Var(X) = \frac{1}{\lambda^2}$$

Two important characteristics of the exponential RV are:

*Given n **independent** exponential RVs $X_1, \ldots, X_n$, whose rates are $\lambda_1, \ldots, \lambda_n$, the RV $Y = min\{X_1, X_2, \ldots X_n\}$ is itself exponential with a rate $\lambda = \sum_{i=1}^n \lambda_i$.*

Proof

Trivially,

$$\begin{aligned} &P\{Y > a\} = P\{\min\{X_1, X_2, \ldots X_n\} > a\} \\ &= P\{X_1 > a, \ldots, X_n > a\} \\ &= \prod_{i=1}^{n} e^{-\lambda_i a} \\ &= e^{-\sum_{i=1}^n \lambda_i a} \end{aligned}$$

Then, $F_Y(a) = 1 - e^{-\sum_{i=1}^n \lambda_i a}$.

Note that the same does not hold for the maximum.

♦

The above property mirrors something that we already learned from an exercise time ago. If we take a *large* number of independent exponential RVs, their minimum will be distributed more and more as a **step function in zero**. We know this to be true whatever the distribution, but in the case of the exponential we can appreciate it because this is exactly what happens when $\lambda \to +\infty$.

Example

Given the system below (in series), the lifetimes of its components are independent exponential RVs, with a rate λ_i. Compute the probability that the system globally works.

Solution

A series system works only if *all* its components work. Therefore, the first component to break also halts the system. Thus,

$$\begin{aligned} &P\{\text{system works at time } t\} = P\{X_1 > t, X_2 > t, X_3 > t\} \\ &= P\{\min(X_i) > t\} \\ &= 1 - F_{\min}(t) \\ &= e^{-(\lambda_1+\lambda_2+\lambda_3)t} \end{aligned}$$

♦

Perhaps the most important property (especially given that about half of the course relies on it) is the **memoryless property**, which can be stated as follows.

If X is an exponential RV,

$$P\{X > s + t | X > t\} = P\{X > s\}$$

Let us put this into context using a simple example. Assume that you want to know whether a device, that has already worked for *t* hours, will work for *s* more hours. Assume that its lifetime is an exponentially distributed variable. The memoryless property states that *the answer is independent of* ***how long in the past*** *the device has worked.*

Let us prove it formally:

$$P\{X > s + t | X > t\} = \frac{P\{X > s + t, X > t\}}{P\{X > t\}}$$
$$= \frac{P\{X > s + t\}}{P\{X > t\}}$$
$$= \frac{e^{-\lambda(s+t)}}{e^{-\lambda t}}$$
$$= e^{-\lambda s}$$

Example

The interarrival time of packets at a network switch is an exponential RV with an average $1/\lambda = 100$ ms. What is the probability that a packet will arrive later than 50 more ms, given that it hasn't arrived in the past *n* ms?

Solution

The solution is straightforward: $P\{X > 50\} = e^{-\frac{50}{100}} = e^{-\frac{1}{2}} = 0.604$.

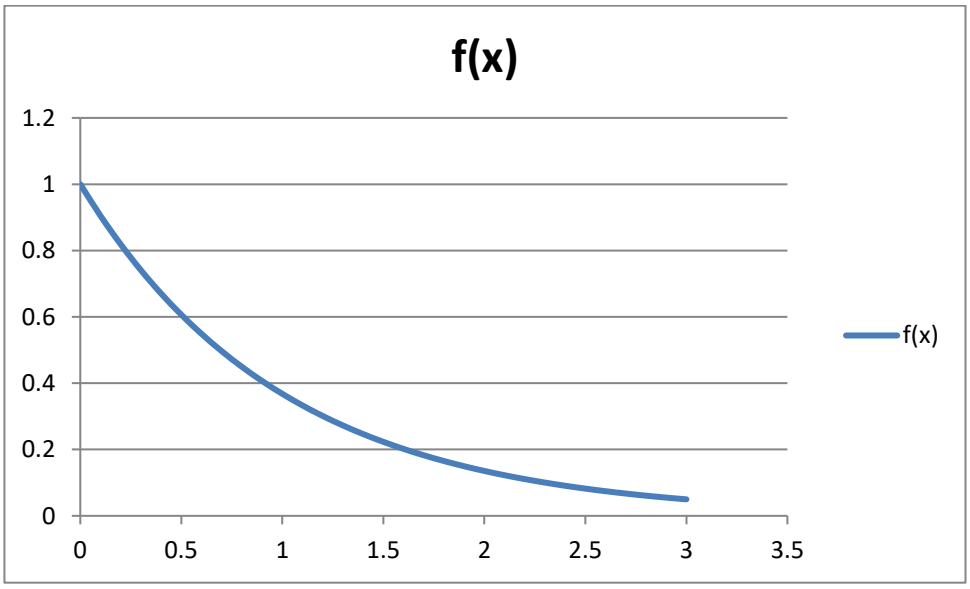
Obviously, if the RV was anything except exponential, then the answer would depend on **both** the **expired lifetime** and the **expected lifetime** (and I would need **more information** to solve the exercise, notably the value of *n*).

$$P\{X > 50 + n | X > n\} = \frac{P\{X > 50 + n\}}{P\{X > n\}} = \frac{1 - F(50 + n)}{1 - F(n)}$$

With exponential variables, I don't need that information.

♦

Let us give a **graphical interpretation** of the memoryless property.



The area below the blue curve is 1 (by the normalization condition). If you want to compute the conditional probability $P\{X > s + t | X > t\} = \frac{P\{X>s+t\}}{P\{X>t\}} = \frac{P\{X>s\}}{P\{X>0\}}$, then the ratio of

- the area to the right of $s+t$, to
- the area to the right of t

Is the same as the ratio of the area to the right of s to the total area below the curve. This means that shifting the origin to the right by t and rescaling the y axis accordingly preserves the shape of the exponential PDF. This is true for **any value of *t***: the further you move to the right, the smaller the area will be, hence the higher the normalization constant will be as well. This yields the memoryless property.

The exponential RV is the **only continuous distribution** having the memoryless property, much like the geometric RV is in the only discrete one. However, the similarity between the two extends further. *If X is an exponential RV with a rate* $\lambda > 0$, then $Y = \lfloor X \rfloor$ is geometric with a parameter $p = 1 - e^{-\lambda}$ (or, if you prefer, $\lambda = -\ln(1-p)$). This is evident from the figures as well.
In fact, we have that:

$$\begin{aligned} p_Y(k) &= P\{Y = k\} = P\{k \le X < k+1\} \\ &= F(k+1) - F(k) \\ &= \left(1 - e^{-\lambda(k+1)}\right) - \left(1 - e^{-\lambda k}\right) \\ &= \left(e^{-\lambda}\right)^k \cdot \left(1 - e^{-\lambda}\right) \\ &= (1-p)^k \cdot p \end{aligned}$$

And this is a geometric distribution.

4.2.3 Laplace-Stieltjes Transform

The LS transform mirrors the concept of the PGF for **continuous, non-negative** RVs. Given a PDF $f(t)$, it is:

$$L(s) = E[e^{-st}] = \int_0^{+\infty} e^{-st} \cdot f(t) \cdot dt,$$

With s being a complex variable. The integral converges as long as $\mathbb{R}(s) \ge 0$. The following properties are noteworthy:

- **Normalization**: $L(0) = 1$.
- **Central moments**: $E[X^k] = (-1)^k \cdot L^{(k)}(s)|_{s=0}$. Thus, the mean value is $-L'(0)$, the mean squared value is $+L''(0)$, etc.. Therefore, the variance is $\sigma^2 = L''(0) - [-L'(0)]^2$.

The same two properties already seen for the PGF also hold in this case:

- **univocity**: if two RVs *X, Y* have the same LSTs, then they have the same PDFs and vice versa. This means that $\forall x,\ \ f_X(x) = f_Y(x) \Leftrightarrow \forall s,\ \ L_X(s) = L_Y(s)$. The LST and the PDF retain the same amount of information.
- **convolution**: if you have *n* **independent** RVs (not necessarily identical) $X_1, \ldots, X_n$, and you know their LSTs $L_1(s), \ldots, L_n(s)$, then you can easily get the LST of their **sum** as:

$$\begin{aligned} L_S(s) &= E\left[e^{-s(X_1+\ldots+X_n)}\right] = E[e^{-s\cdot X_1} \cdot e^{-s\cdot X_2} \cdot \ldots \cdot e^{-s\cdot X_n}] \\ &= E[e^{-s\cdot X_1}] \cdot E[e^{-s\cdot X_2}] \cdot \ldots \cdot E[e^{-s\cdot X_n}] \\ &= L_1(s) \cdot L_2(s) \cdot \ldots \cdot L_n(s) \end{aligned}$$

That is, you just **multiply their LSTs**. The middle passage is due to independence, of course.

We compute the LST of the exponential:

$$\begin{aligned} L(s) &= E[e^{-st}] = \int_0^{+\infty} e^{-st} \cdot \lambda \cdot e^{-\lambda t} \cdot dt \\ &= \lambda \cdot \int_0^{+\infty} e^{-(s+\lambda)\cdot t} \cdot dt \\ &= \frac{\lambda}{-(s+\lambda)} \cdot \left[e^{-(s+\lambda)\cdot t}\right]_0^{+\infty} \\ &= \frac{\lambda}{s+\lambda} \end{aligned}$$

From this, you readily obtain the mean as:

$L'(s) = -1 \cdot \frac{\lambda}{(s+\lambda)^2}$, $L''(s) = +2 \cdot \frac{\lambda}{(s+\lambda)^3}$, hence

$E[X] = -L'(0) = +1 \cdot \frac{\lambda}{(\lambda)^2} = \frac{1}{\lambda}$, $E[X^2] = L''(0) = +2 \cdot \frac{\lambda}{(\lambda)^3} = \frac{2}{\lambda^2}$

$Var(X) = E[X^2] - E[X]^2 = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}$.

This is a lot quicker than integrating by parts, as we would have done otherwise. The LST is particularly useful when you work with exponential distributions and others derived from it (we will encounter some more later on in the course).

4.2.4 Normal distribution

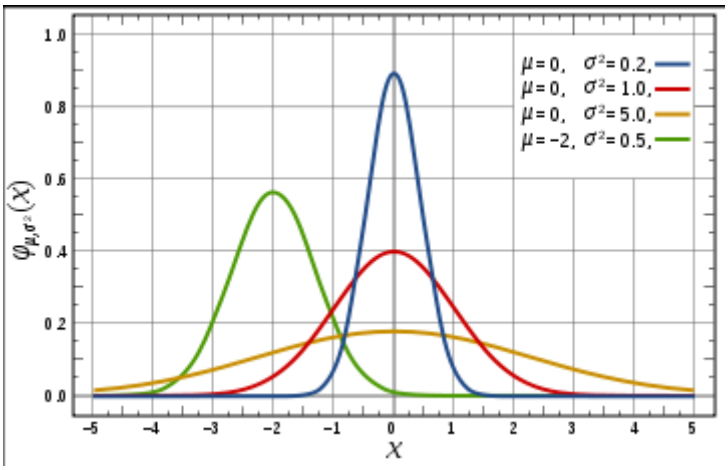
A **Normal** or **Gaussian** distribution is a **continuous** one, whose PDF is the following:

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

With σ being a positive number. This PDF takes on its **maximum value** when $x = \mu$, and that value is equal to: $1/\left(\sqrt{2\pi} \cdot \sigma\right) \sim 0.4/\sigma$. It is obviously **symmetric** around μ, hence the latter is also its **mean value**. Now, the higher σ, the smaller the peak *and* the heavier the tails (the area below must

be equal to one in any case). Hence, σ has something to do with how much values are *dispersed* around the mean. We give for granted (without proving it) that $\boldsymbol{\sigma^2}$ **is the variance.**

The PDF may take on values in $(-\infty; +\infty)$. The normal PDF is the typical **bell-shaped curve**.



More on the shape of the PDF: the PDF **changes concavity twice** (since it stretches to infinity in both directions and it has a maximum). It is quite interesting to see where the **inflection points** lie.

Exercise

Assume $\mu = 0$. Compute the inflection points for a Normal distribution.

Solution

These are the points where the **second derivative** is null. Hence, we solve the following equality:

$$\frac{\partial^2}{\partial x^2}\left(\frac{1}{\sqrt{2\pi}\cdot\sigma}e^{-\frac{x^2}{2\sigma^2}}\right) = 0$$

$$\frac{1}{\sqrt{2\pi}\cdot\sigma}\cdot\frac{\partial^2}{\partial x^2}\left(e^{f(x)}\right) =$$

$$\frac{1}{\sqrt{2\pi}\cdot\sigma}\cdot\left(f''(x)\cdot e^{f(x)} + f'(x)^2\cdot e^{f(x)}\right) =$$

$$\frac{1}{\sqrt{2\pi}\cdot\sigma}\cdot e^{-\frac{x^2}{2\sigma^2}}\cdot\left[\frac{-\cancel{2}}{\cancel{2}\sigma^2} + \left(\frac{-\cancel{2}x}{\cancel{2}\sigma^2}\right)^2\right] =$$

$$\frac{1}{\sqrt{2\pi}\cdot\sigma}\cdot e^{-\frac{x^2}{2\sigma^2}}\cdot\frac{1}{\sigma^2}\cdot\left[\frac{x^2}{\sigma^2} - 1\right]$$

The only possibility is that $x^2 = \sigma^2$, i.e. $x = \pm\sigma$.

Hence the inflection points are those that are σ units apart from the mean value.

♦

We have assumed that the normal PDF is indeed a PDF. We need to test the normalization condition in order to be sure.

$$\int_{-\infty}^{+\infty} f(x) \cdot dx = 1$$
$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi} \cdot \sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \cdot dx = 1 \qquad [z = (x-\mu)/\sigma\,, \quad dz = dx/\sigma]$$
$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{z^2}{2}} \cdot dz = 1$$
$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{z^2}{2}} \cdot dz \cdot \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{y^2}{2}} \cdot dy = 1$$
$$\frac{1}{2\pi} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} e^{-\frac{y^2+z^2}{2}} \cdot dy \cdot dz = 1$$

We switch to **polar coordinates** (ρ, θ), with $\rho[0; +\infty)$ $\theta[0; 2\pi)$ and $z = \rho\cos\theta; \quad y = \rho\sin\theta$.
Hence $\rho = y^2 + z^2$, and go on with the computation:

$$\frac{1}{2\pi} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} e^{-\frac{y^2+z^2}{2}} \cdot dy \cdot dz =$$
$$\frac{1}{2\pi} \int_{0}^{2\pi} \int_{0}^{+\infty} e^{-\frac{\rho^2}{2}} \cdot \begin{vmatrix} \frac{\partial z}{\partial \rho} & \frac{\partial z}{\partial \theta} \\ \frac{\partial y}{\partial \rho} & \frac{\partial y}{\partial \theta} \end{vmatrix} d\rho \cdot d\theta =$$
$$\frac{1}{2\pi} \int_{0}^{2\pi} \int_{0}^{+\infty} \rho \cdot e^{-\frac{\rho^2}{2}} \cdot d\rho \cdot d\theta =$$
$$\frac{1}{2\pi} \int_{0}^{2\pi} \int_{0}^{+\infty} -\left(\frac{\partial}{\partial \rho} e^{-\frac{\rho^2}{2}}\right) \cdot d\rho \cdot d\theta$$
$$= \frac{1}{2\pi} \cdot 2\pi \cdot \left[e^{-\frac{\rho^2}{2}}\right]_0^{+\infty} = 1$$

It took quite a while, but we are now convinced that the normalization holds.

We now want to address the following problem. Given $X \sim N(\mu, \sigma^2)$, can we compute $P\{X \in [a, b]\}$? The obvious answer is $F(b) - F(a)$, to compute which **we need the CDF** $F(x)$. Unfortunately, there is **no closed-form CDF for the normal distribution**. There is simply no known function whose derivative is $f(x)$. We can only do the computation **numerically**.
Now, this would imply that – theoretically speaking – we would need a **very large amount of information** (say, a lot of points on the *x* axis) **for any couple of values** $(\boldsymbol{\mu}, \boldsymbol{\sigma^2})$.
Thanks to some obvious symmetries, however, we can limit our computations to a **single couple of parameter values** $(\boldsymbol{\mu = 0, \sigma^2 = 1})$. A normal distribution with $\mu = 0, \sigma^2 = 1$ is called a **standard Normal variable**, denoted as $N(0,1)$.

Suppose I want to compute $P\{X \leq a\} = F(a)$, with $X \sim N(\mu, \sigma^2)$. I should be able to solve the following integral:

$$F(a) = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi} \cdot \sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \cdot dx \qquad \left[\text{set} z = \frac{x-\mu}{\sigma}\right]$$
$$= \frac{1}{\sqrt{2\pi} \cdot \cancel{\sigma}} \int_{-\infty}^{\frac{a-\mu}{\sigma}} \cancel{\sigma} \cdot e^{-\frac{z^2}{2}} \cdot dz$$
$$= \Phi\left(\frac{a-\mu}{\sigma}\right)$$

Where $\Phi(z)$ is the CDF of a **standard normal distribution**, **i.e.** $\boldsymbol{Z \sim N(0,1)}$**.**

This means that I can

a) Compute **numerically** (and write down on a table) a good, fine-grained set of values for $\Phi(z)$

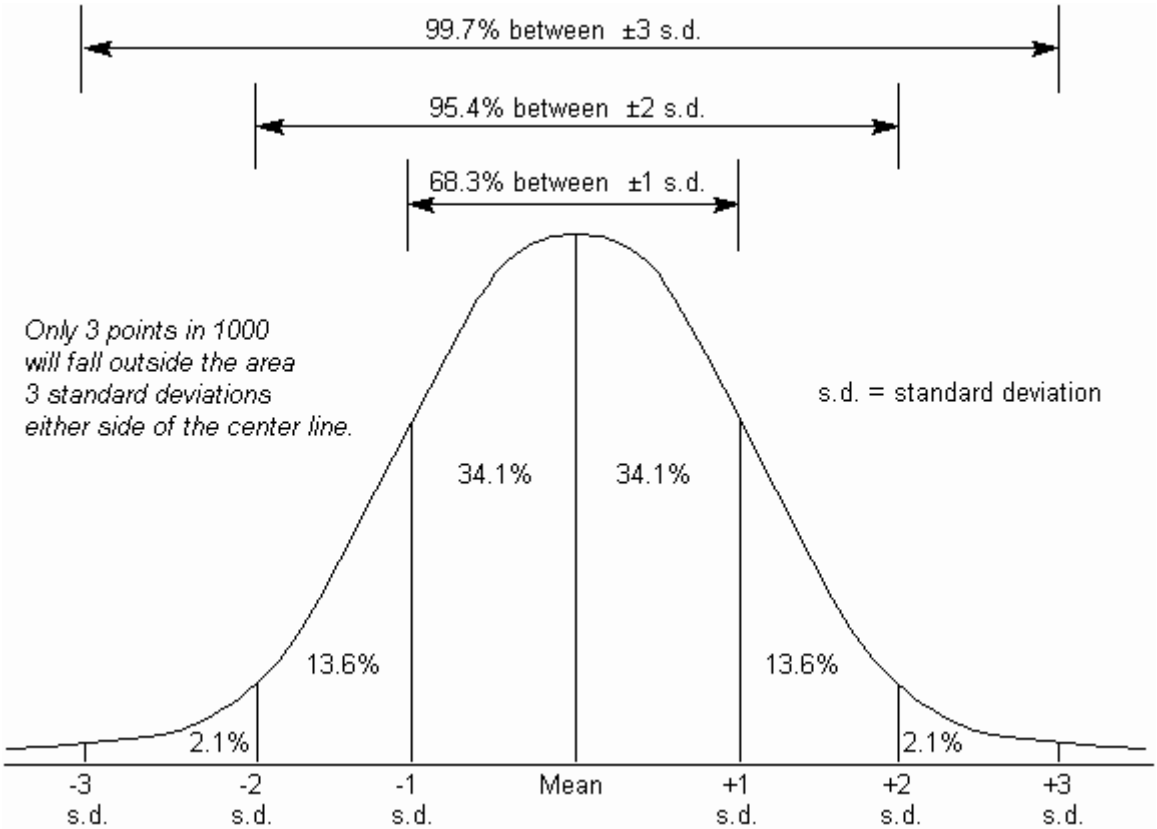
b) Every time I need to compute $P\{X \leq a\}$, with $X \sim N(\mu, \sigma^2)$, I just **look up** $\boldsymbol{\Phi\left(\frac{a-\mu}{\sigma}\right)}$.

Hence, I only need **one table** for every possible sets of values (μ, σ^2).

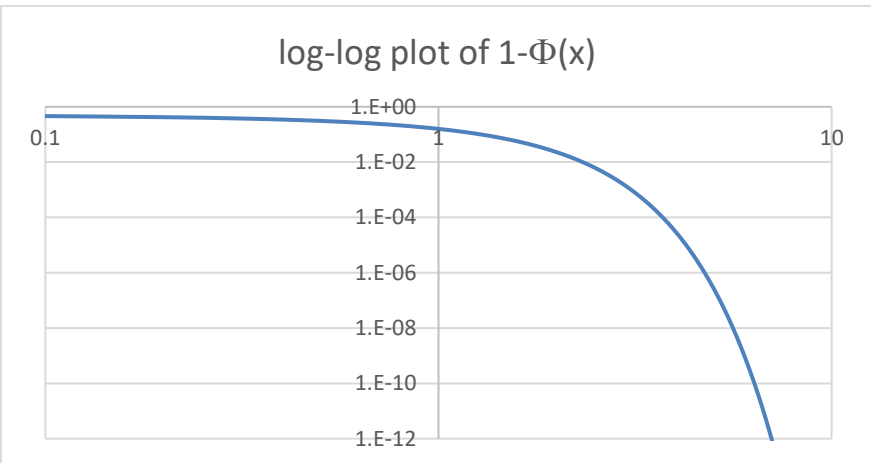
Moreover, function $\Phi(z)$ is such that:

- $\Phi(0) = 1/2$ (by obvious symmetry)
- $\Phi(-a) = P\{Z \leq -a\} = P\{Z \geq a\} = 1 - \Phi(a)$

Hence, you only need to tabulate the values of $\Phi(z)$ for **positive values of** $\boldsymbol{z}$**,** since we can obtain the missing values by symmetry. You can find $\Phi(z)$ in the appendix of every probability theory textbook (as well as on the Internet, of course, and at the end of these notes).



The interesting thing about the normal is that it **declines very steeply.** Only 4.5% of the values are more than 2 standard deviations away from the mean, and as few as 0.3% are more than 3 standard deviations away. This means that we can **limit ourselves to tabulating a $\boldsymbol{Z \sim N(0,1)}$ in $\boldsymbol{(0;5]}$ for most practical applications.** This observation can be substantiated by looking a log-log plot of the **right tail** of the standard normal. We plot function $1-\phi(x)$, which is the "complementary" CDF function, as a function of x, i.e. the number of standard deviations away from the mean. The graph below shows that the right tail decays very fast. For every order of magnitude of increase in x, the residual probability decreases by several orders of magnitude, and it decreases **faster** the more you go to the right.



This means that phenomena that are modeled as Normal RVs have a well-defined **scale**. You can measure things with a ruler whose length is the standard deviation σ, and the probability of encountering values which are more than a bunch of σs away from the mean is negligible. Out-of-scale values are **vanishingly unlikely** with Normal RVs.

Exercise

A binary message traverses a channel affected by a standard normal noise. To make the transmission more robust, we:

1) Transmit either +2 or -2 (instead of 1 and 0)
2) Decode "1" if the received signal is above 0.5, and "0" if below.

Compute the (conditional) error probabilities for either transmission.

Solution

Call R the noise. It is $R \sim N(0,1)$.

$$\begin{aligned} &P\{error|1\} = P\{R + 2 < 0.5\} \\ &= P\{R < -1.5\} \\ &= P\{R > 1.5\} \\ &= 1 - \Phi(1.5) \\ &= 0.0668 \end{aligned}$$

$$\begin{aligned} &P\{error|0\} = P\{R - 2 > 0.5\} \\ &= P\{R > 2.5\} \\ &= 1 - \Phi(2.5) \\ &= 0.0062 \end{aligned}$$

♦

Exercise

The power dissipated in a resistor is $W = cV^2$, with V being the tension. Assume that $c = 3$, $V \sim N(6,1)$. Compute $E[W]$ and $P\{W > 120\}$.

Solution

$$E[W] = E[3V^2] = 3E[V^2] = 3[Var(V) + {\mu_V}^2] = 3[\sigma^2 + {\mu_V}^2] = 3[1 + 36] = 111$$

$P\{W > 120\} = P\{3V^2 > 120\} = P\{V > 2\sqrt{10}\} + \cancel{P\{V < -2\sqrt{10}\}}$.

Note that $-2\sqrt{10}$ is around -6, i.e. more than 12 standard deviations away from the mean value. Hence, **we can safely ignore** that probability since it will be negligible.

However, $V \sim N(6,1)$, hence:

$P\{V > 2\sqrt{10}\} = P\left\{\frac{V-6}{1} > \frac{2\sqrt{10}-6}{1}\right\} = P\{Z > 0.3246\}$ (with $Z \sim N(0,1)$).

From the tables, I get $P\{Z > 0.3246\} = 1 - \Phi(0.3246) = 0.3727$

♦

A noticeable property of the normal distribution is the following:

Given n ***independent normal*** *RVs* $X_1, \ldots, X_n$, *RV* $Y = \sum_{i=1}^{n} \pm X_i$ *is itself normal with parameters* $\mu = \sum_{i=1}^{n} \pm\mu_i$, $\sigma^2 = \sum_{i=1}^{n} \sigma_i^2$

Note that the result for the **mean** is obvious, and does not require any hypothesis (let alone independence). The one on the **variance** is instead due to independence. Thus, the only non-trivial fact about this property is the fact that the sum of normals is normal itself.

Exercise

Data from the National Oceanic and Atmospheric Administration indicate that the yearly precipitation in Los Angeles is a normal random variable with a mean of 12.08 inches and a standard deviation of 3.1 inches.

a) Find the probability that the total precipitation during the next 2 years will exceed 25 inches.
b) Find the probability that next year's precipitation will exceed that of the following year by more than 3 inches.

Assume that the precipitation totals for the next 2 years are independent.

Solution

Call X_1, X_2 the precipitation for the next 2 years. Now, $X_1 + X_2$ is normal with mean 24.16 and variance $2 \cdot 3.1^2 = 19.22$. Hence,

$$P\{X_1 + X_2 > 25\} = P\left\{\frac{X_1 + X_2 - 24.16}{\sqrt{19.22}} > \frac{25 - 24.16}{\sqrt{19.22}}\right\}$$
$$= P\{Z > 0.1916\}$$
$$= 0.4240$$

b) RV $-X_2$ is normal with a mean equal to -12.08 and the same variance. It follows that $X_1 - X_2$ is a normal RV with mean 0 and variance 19.22. Hence,

$$P\{X_1 - X_2 > 3\} = P\left\{\frac{X_1 - X_2 - 0}{\sqrt{19.22}} > \frac{3-0}{\sqrt{19.22}}\right\}$$
$$= P\{Z > 0.6843\}$$
$$= 0.2469$$

Thus there is a 42.4% chance that the total precipitation in Los Angeles during the next 2 years will exceed 25 inches, and there is a 24.69% chance that next year's precipitation will exceed that of the following year by more than 3 inches.

♦

Why do people use **normal RVs?** Because these model phenomena which are indeed common in the real world, which occur with **large populations**. The motivation lies in the following, very important theorem (which is often believed to be *the* single most important result in probability theory).

4.2.5 Central limit theorem

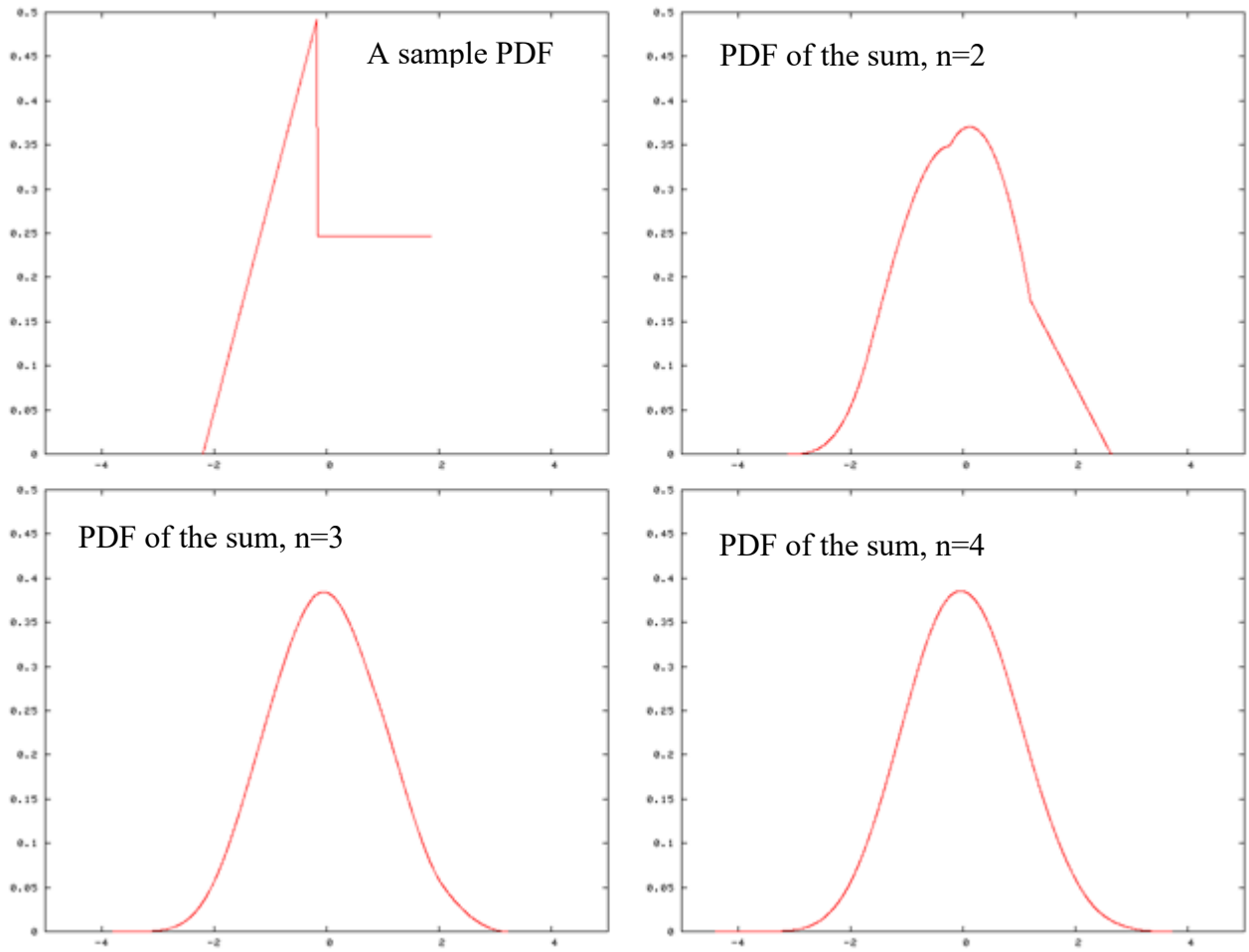
Given n **iid RVs** $\boldsymbol{X_i}$, having **finite** mean value μ and **finite** variance σ^2, whatever their distribution **(i.e. not necessarily normal), RV** $\boldsymbol{S = \sum_{i=1}^{n} X_i}$ has a mean $n \cdot \mu$, a variance $n \cdot \sigma^2$, and:

for large values of $\boldsymbol{n}$ ($\boldsymbol{n \geq 30}$) it is approximately normal

In other words:

$$\boldsymbol{P\left\{\frac{\sum_{i=1}^{n} X_i - n \cdot \mu}{\sqrt{n} \cdot \sigma} \leq x\right\} \cong \Phi(x)}$$

This means that the larger n is, the more similar to a normal RV that sum becomes, **whatever the distributions of the RVs** (as long as they have finite mean and variance)**.** Moreover, if the distributions are "tame" enough, the sum converges to a normal RV already **for small values of $\boldsymbol{n}$, e.g. 3-4.**



The figure above shows that, even with a **very skewed** PDF, the sum of 3-4 RVs is enough to obtain a perfectly normal distribution.

An alternative (but equivalent) formulation of the CLT is the following:
Define RV $M = 1/n \cdot \sum_{i=1}^{n} X_i$ *(sample mean). Then* $E[M] = \mu$, *and* $Var(M) = (n \cdot \sigma^2)/n^2 = \sigma^2/n$, *and* $M \sim N(\mu, \sigma^2/n)$.

This alternative formulation shows that the sample mean of *n* iid RVs converges to a normal, whose mean is the mean of the individual RVs. The variance goes to zero as $n \to \infty$.

The central limit theorem can be put to good use, for instance, to approximate **binomial distributions**. We know that a binomial RV is the sum of *n* iid bernoullian RVs. Hence we can apply the CLT when *n* is large.
In fact, a binomial has a mean value $\mu = n \cdot p$ and a variance $\sigma^2 = n \cdot p \cdot (1 - p)$, and it is approximately Gaussian ***if n is large***. In practice, this means if $n \cdot p \cdot (1 - p) > 10$. In this case, it is:

$$P\left\{\frac{X - n \cdot p}{\sqrt{n \cdot p \cdot (1-p)}} \leq x\right\} \cong \Phi(x)$$

Thus, there are two ways to approximate a binomial variable:

a) When n is large and p is small, you can use a Poisson RV with $\lambda = n \cdot p$
b) When $n \cdot p \cdot (1-p) > 10$, you can use a Gaussian with mean $\mu = n \cdot p$ and variance $\sigma^2 = n \cdot p \cdot (1-p)$.

Exercise

The ideal size of a first-year class at a college is 150 students. The college, knowing that, on average, 30% of those accepted will actually attend, uses a policy of approving the applications of 450 students. Compute the probability that more than 150 first-year students attend classes at the college.

Solution

Let X denote the number of students that attend; then assuming that each accepted applicant will independently attend, it follows that X is a binomial RV with $n = 450, p = 0.3$.

First of all, we cannot (in practice) use the binomial distribution to compute that property. In fact, $450! > 10^{1000}$, and p^{450} is also hard to compute numerically. Hence, we must settle for an approximated evaluation. Note that **we are not in a position to use the Poisson approximation**. In fact, p is <u>not small enough</u> (nor, probably, n large enough) to do this. However, it is $n \cdot p \cdot (1-p) = 94.5 \gg 10$, hence we expect the Gaussian approximation to be a good one. Call Y the normal that approximates the binomial. Since the binomial is a discrete distribution and the normal is a continuous one, **it is best to compute**:

$$P\{X = i\} = P\{i - 0.5 < Y < i + 0.5\}$$

when applying the normal approximation (this is called the **continuity correction**). Hence we must compute the required probability as:

$$P\{X \geq 151\} = P\{151 \leq X \leq 450\} = P\{150.5 < Y < 450.5\}$$

Which is equal to:

$$P\{150.5 < Y < 450.5\} = P\left\{\frac{150.5 - 450 \cdot 0.3}{\sqrt{450 \cdot 0.3 \cdot (1-0.3)}} < \frac{Y - 450 \cdot 0.3}{\sqrt{450 \cdot 0.3 \cdot (1-0.3)}} < \frac{450.5 - 450 \cdot 0.3}{\sqrt{450 \cdot 0.3 \cdot (1-0.3)}}\right\}$$
$$= P\{1.59 < Z < 32.5\} = 1 - \Phi(1.59)$$
$$= 0.06$$

Hence, only 6% of the time do more than 150 of the first 450 accepted actually attend. This may seem surprising. Note, however, that

- The mean value is $\mu = n \cdot p = 135$

- The standard deviation is $\sigma = \sqrt{n \cdot p \cdot (1-p)} \cong 9.72$.

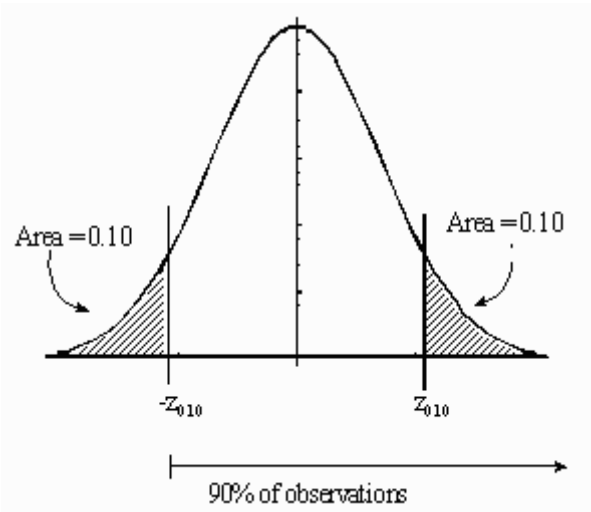
Thus, 150 students are **1.5 σ away from the mean**, which is quite a long way for a normal RV.
Finally, note that the *exact* binomial value is 0.0566, quite near the mark (barring numerical errors). If we had used a Poisson approximation, instead, we would have got $P\{X > 150\} = 0.093$, which is instead far off the mark, and possibly subject to numerical errors as well.

♦

Note, once again, that the assumption of **finite mean and variance** is paramount. There are distributions whose variance is infinite (Pareto, log-normal, sometimes Weibull, Cauchy are some), and the CLT does not work with them. These distributions are collectively known as **"heavy-tail",** and the sum of heavy-tail distributions **does not** converge to normal. It still converges, but to a different distribution ("stable" distributions). The normal is the (one and only) stable distribution for light-tailed RVs.

4.2.6 Percentiles

We define a $(1-\alpha)$ **percentile** z_α as the value for which $P\{X > z_\alpha\} = \alpha$ (or, if you prefer, $P\{X \leq z_\alpha\} = 1-\alpha$). For instance, $z_{0.01}$ is that value for which the residual **tail** of the normal distribution (i.e. the one to the right of $z_{0.01}$) has an area of 0.01.
The definition of percentile is not necessarily tied to the Normal random variable. However, you can find the percentiles for the **standard Normal** tabulated in the textbooks.



Percentiles are useful in several computations

- The 0.5 percentile is called **median**
- The 0.25 and 0.75 percentiles are called **first and third quartiles** respectively

- "High" percentiles are common ways to assess the **confidence** of a measure, as we will see later on.

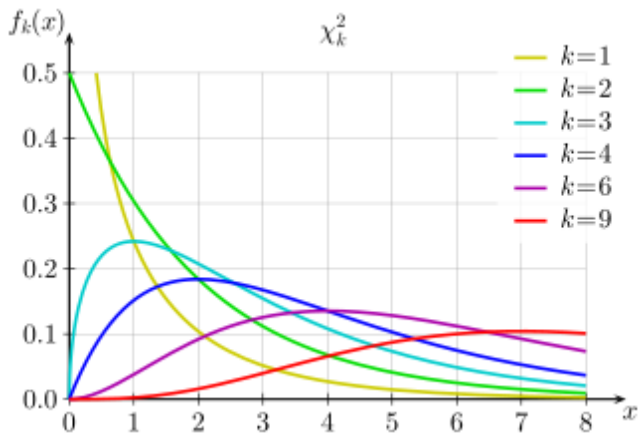
Note that the **notation** used to define percentiles varies among textbooks. In some cases, you might find $z_{0.01}$ indicated as $z_{0.99}$, to underline that the area to its ***left*** is 0.99. This is most confusing when confidence measures are to be computed. However, we will stick to a consistent notation throughout our notes.

A synonym for percentile is **quantile**: the 90th percentile $z_{0.1}$ is also called the **0.9 quantile**.

4.2.7 Chi-square distribution

Normal RVs are so important that some RVs which are derived from the normal have been given names of their own (since they often arise in practical cases). Assume $Z_1, \ldots Z_k$ are independent **standard** Normal RVs. Then $X = {Z_1}^2 + {Z_2}^2 + \ldots + {Z_k}^2$ is a **chi-square with *k degrees of freedom*** $\boldsymbol{\chi_k^2}$. The expression of the PDF is rather complex and not particularly interesting (you can find it on textbooks). The CDF is only known **numerically** (there are k tables, one per number of degrees of freedom). Quite often, you find **high percentiles** as tabulated values, i.e. those numbers $\chi_{\alpha,k}^2$ such that $P\{X > \chi_{\alpha,k}^2\} = \alpha$.

The shape of the PDF is the one in the figure. When $k = 1,2$ it is monotonically decreasing, and then it starts exhibiting a peak, which grows to the right as k grows large. It is also clear that, for a suitably large k, the chi-square **approaches a normal RV**, because of the CLT (recall that $Z_1, \ldots Z_k$ are iid). For this distribution, it is $E[X] = k$, $Var(X) = 2k$.



4.2.8 Student's T distribution

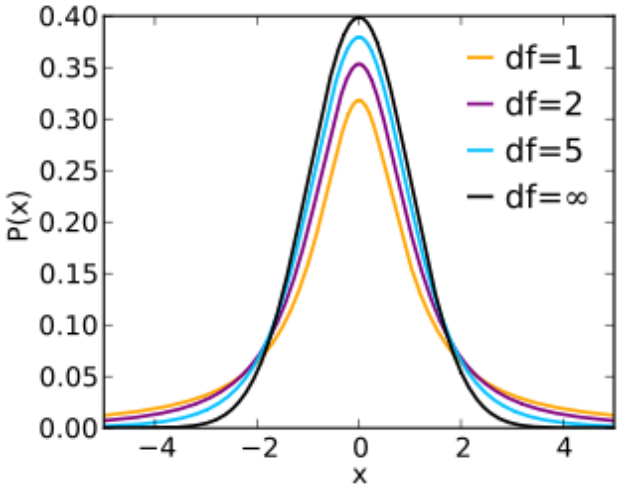
Assume that $Z \sim N(0,1)$, and that χ_n^2 is a chi-square with n degrees of freedom. Furthermore, assume that Z and χ_n^2 are **independent**. Then, we can define the following RV:

$$T_n = \frac{Z}{\sqrt{\chi_n^2/n}}$$

And call it a **Student's[2] T with n degrees of freedom.**

As n grows large, χ_n^2/n becomes a **constant**, equal to its mean value 1[3]. Therefore, the denominator of T_n tends to 1 as well. This means that, for **a very large n (e.g., n>30)**, it is $T_n \approx Z$, i.e. $T_n \sim N(0,1)$.

Student's T distribution tends to the standard Normal when n is large.



When n is above 30, there is really no difference between the two. When $n < 30$, the T distribution tends to have a **less pronounced peak** and **heavier tails** than a standard normal.

In any case, we have $E[T_n] = 0$, $Var(T_n) = n/(n-2)$, which is slightly larger than one for a large n. This last result confirms that $T_n \sim N(0,1)$ when n is large.

You can find a Student's T tabulated in the textbooks. Notably, you can find **high percentiles $t_{\alpha,n}$**, which are the values of practical use. In fact, the value $t_{\alpha,n}$ such that $P\{T_n \geq t_{\alpha,n}\} = \alpha$ is a number that **decreases with n,** and it is in general an **upper bound** on the percentile z_α of the standard normal. In fact, the area below the **tails** tend to reduce as n grows large.

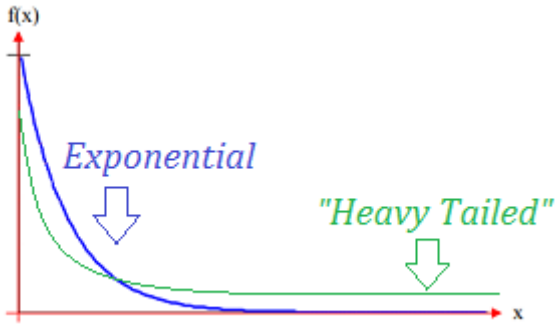
Student's T distribution is often used to assess the **confidence** of some measure. We will come back to this later on in the course.

[2] Student is not the name of the person who invented it. He was employed for a firm that would not allow him to publish scientific results, so he signed his scientific papers using "Student" as a pseudonym.

[3] This is a consequence of the CLT, which is evident from its second formulation.

4.3 Heavy-tailed distributions

What is a heavy tail? It is a (right) tail that **decays slower than an exponential**. The main consequence is that "very large" values (with respect to the mean) are still unlikely, but **not vanishingly unlikely**.



The most common definition of Heavy Tail (HT) is the following:

$$\forall \lambda > 0, \quad \lim_{x \to \infty} e^{\lambda x} \cdot \big(1 - F(x)\big) = \infty$$

That definition is not univocous, however. Some people intend slightly different things then they use HT, and there are similar classes of RVs which are almost (but not quite) overlapping. These are "long tail" and "subexponential", and it is $SubEx \subset LongT \subset HT$. However, most HT RVs that are of practical use are also SubEx, therefore, we will use the collective name HT and forget about the nuances.

Here is an example, which we have already encountered. This is a Pareto distribution:

$$f(x) = \begin{cases} 0 & x \leq 100 \\ \dfrac{100}{x^2} & x > 100 \end{cases}$$

In this case, we have:

$$F(x) = \int_{100}^{x} \frac{100}{y^2} dy = \left[\frac{-100}{y}\right]_{100}^{x} = \left[1 - \frac{100}{x}\right] \cdot 1_{\{x \geq 100\}}$$

However, $\lim_{x \to \infty} e^{\lambda x} \cdot \frac{100}{x} = \infty$, hence this distribution is **heavy-tailed**.

In fact, one can divide the world of distributions into two categories: light-tailed and heavy-tailed distributions. **The divide is in fact the exponential distribution**. Distributions that decay faster than the exponential (e.g., the Normal one) are light-tailed, and those that decay slower than the exponential are HT.

HT distributions are just **everywhere**, and they are **often misunderstood**. It is as if humans "want" things to be Normal[4], marvel that – more often than they expect – they are not, and keep thinking in Normal terms nonetheless, even when they know it to be wrong, sometimes to the price of huge, costly mistakes. A possible explanation for this is that the human brain evolved over millions of years the ability to guess the odds of physical fenomena, which are often Normal(ish). However, **informational** and **social** phenomena – which have appeared only very recently in evolutionary time – often give rise to heavy-tailed distributions.

One of the main reasons why HT RVs tend to appear quite often is that they are the natural outcome of **multiplicative processes**. If you sum up iid light-tailed RVs, you get a Normal. If instead you **multiply** iid light-tailed RVs, you get a HT RVs. You get a multiplicative process whenever something **grows proportionally to its size**, instead of growing by fixed increments (which would yield an additive process). Interest rates are multiplicative. Population size grows multiplicatively. Multiplicative processes are everywhere.
Intuitively, this statement makes sense. Take $X_1, X_2, \dots X_n$, IID RVs, and define $Y_i = logX_i$. Then RV $Z = \prod X_i$ is such that $logZ = \sum logX_i = \sum Y_i$. However, $\sum Y_i$ is Normal (by the CLT, as long as Y_i have finite variance), hence $logZ$ is Normal. This means that Z is a **lognormal distribution**. i.e. one whose log is a Normal. Lognormal distributions are HT.

Another common phenomenon that gives rise to HT is **preferential attachment** (or "rich get richer" principle). Consider a set of YouTube videos. Some may be more popular than others, which is reflected in their number of **views**. If a video is popular, those who watch it will recommend it to others. Therefore, the probability that a new YouTube user watches one video is skewed in favor or popular videos. New users **attach preferentially** to videos with high viewcounts. The probability that a new user will "attach to" (i.e., select) a video is proportional to its viewcount. This means that popular videos will get even more popular, and this generates a distribution of viewcounts which is HT. Some videos will have a viewcount which is **out of scale** w.r.t. the mean number of views. The number of views per YouTube video is in the thousands, whereas top videos have billions of views (i.e., 10^6 as many as the mean).

[4] This seems to be implicit in the name. If you name a distribution "Normal", you expect it to occur as a default case, and assume that every other distribution should be an exception, and as such require ad hoc explanations. This is plainly wrong, of course.

Preferential attachment explains the HT distribution of many phenomena, such as:

- **Internet connectivity**: the number of connections of an Autonomous System is typically HT. If you are a new Internet Service Provider, and you want to select a good backbone provider, you are more likely to choose a well-connected one.
- The **size of cities**: The more people live in a city, the better the opportunities and services are likely to be. This attracts even more people.
- The **income distribution** in a free-market society. The more money you have, the more you can invest in making more money. Since investment returns are multiplicative, you will get higher returns, and this phenomenon self-sustains.

Several quantities related to Computer Engineering follow HT distributions:

- The **lifetime of Unix processes**.
- The **number of packets in IP flows** (the so-called "mice and elephants" phenomenon)
- Again, Internet inter-AS connectivity.
- **Social-network** phenomena (e.g., X followers).

Example: Pareto distribution:

$$F(x) = 1 - \left(\frac{x_m}{x}\right)^{\alpha}, \quad f(x) = \frac{(x_m)^{\alpha} \cdot \alpha}{x^{\alpha+1}}$$

With $x_m > 0, \alpha > 0, x \in [x_m; +\infty]$.

When $\alpha \leq 2$, the variance is infinite. When $\alpha \leq 1$, the mean is infinite too.

HT distributions have some peculiar characteristics, which set them apart from light-tailed ones like the Normal.

Memory property. Assume that the RV X models the time at which some event is supposed to occur, and that t is a large constant: your event hasn't happened for quite a long time.

- For exponential RVs, it is $P\{X > s + t | X > t\} = P\{X > s\}$. This means that the fact that you have waited for a long time is completely irrelevant. Observing this phenomenon gives you no insight whatsoever as to when your event will occur.
- For light-tailed RVs (e.g., a Normal), it is $P\{X > s + t | X > t\} < P\{X > s\}$. The more you have waited, the less likely it is that you will have to wait more. Events that are over-due become *more likely* as time passes.

- For HT RVs, it is $P\{X > s + t | X > t\} > P\{X > s\}$. If you have waited for long, chances are *you will wait for longer still*. Events that are overdue become **less** likely as time passes.

Let us check the memory property with a Pareto distribution:

$$P\{X > s + t | X > t\} = \frac{P\{X > s + t\}}{P\{X > t\}} = \frac{\left(\frac{x_m}{s+t}\right)^\alpha}{\left(\frac{x_m}{t}\right)^\alpha} = \left(\frac{t}{s+t}\right)^\alpha$$

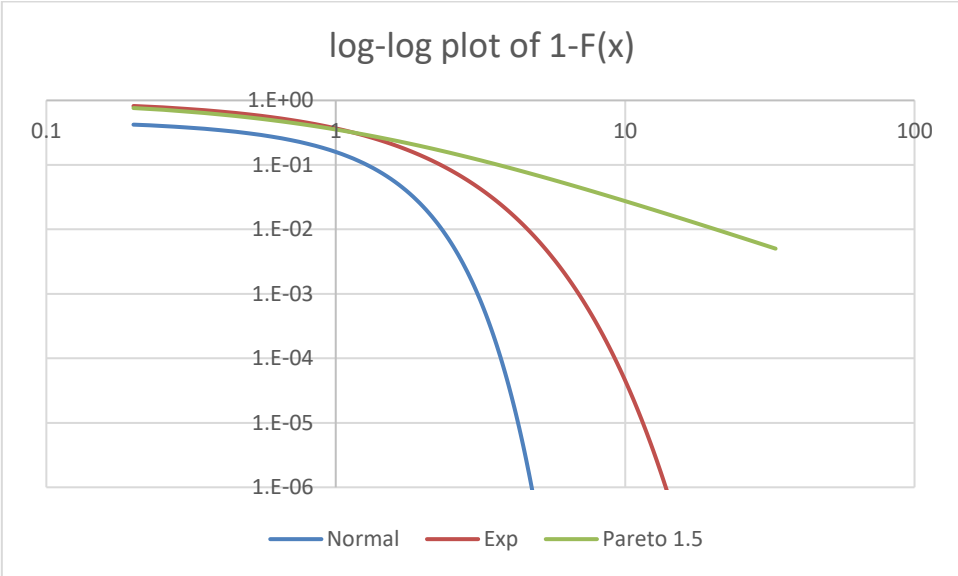
When t is large (i.e., you have waited for long), the above expression tends to 1 (hence the memory property will always hold for a large t). If something is out of scale, it will be so big time.
Human schedules sometimes follow HT distributions. If a project which was supposed to be completed in 2 years hasn't in 4, then it is likely to go on forever. Counting on it to be finished soon **just because it is overdue** is probably not the smartest thing to do. It would be in a Normal setting, but human schedules are not Normal.

Conspiracy vs. catastrophe. If you have a very large value for the **sum of IID RVs**, with light-tailed RVs, this is likely to be due to a "conspiracy": all the RV have a similar, large value. With heavy-tailed ones, it is probably due to a "catastrophe", i.e. one of the RVs has a value comparable to the total, and the other are irrelevant.
Take the following example: if I tell you that the total height of a group of 10 people is an exceptional 20 meters, you will probably think of a basketball team, where everyone is around 2 meters tall (i.e., on the right tail of a Normal distribution, since height is normally distributed). If, instead, I tell you that the collective **wealth** of 10 people is an exceptional two billion euros, chances are that there is **one** billionaire and nine small-time earners. The chances that you've hit on a group of ten people who earn around 200M each are negligible in comparison. In other words, in a **sample of many IID RVs**, most of the total will be given by **one or few very large values**, whereas the others (the vast majority) will be almost irrelevant.
In mathspeak, this property can be formulated as:

$$\lim_{x\to\infty} \frac{P\{X_1 + X_2 + \cdots, X_n > x\}}{P\{\max(X_1, X_2, \dots, X_{n)} > x\}} = 1$$

Scale-free property. While with light-tailed RVs the mean value is representative of a typical observation (at least by order of magnitude), it is quite hard to tell what the "typical" scale of a HT RV is, since the possible values are likely to span orders of magnitude. Very large values are uncommon, but **not vanishingly so**. It is **very** unwise to count on some out-of-scale event never to occur, if your

distribution is HT. The graph below shows what happens with the tails of light- and heavy-tailed RVs. It is quite clear that out-of-scale events are not impossible if your tail decays subexponentially.



Infinite variance. Not all HT RVs have infinite variance, but many do. For instance, a Pareto distribution is (always) HT, but its variance is infinite when $\alpha \leq 2$. Having infinite variance is a sufficient condition for being HT, but not a necessary one. Infinite variance prevents you from applying the CLT, which requires variance to be finite. There exists a "generalized" CLT that describes what happens when you sum up IID RVs with infinite variance (with a few extra hypotheses), and it does not yield a Normal distribution.

An intuitive rationale is the following:

- if you sum many IID RVs with a **finite** variance, their fluctuations **will be diluted** by the sum. The probability of extreme values will become vanishingly small, coherently with the Normal tails.
- If you sum many IID RVs with an **infinite variance**, instead, occasional out-of-scale fluctuations **will resist dilution** by the sum. The probability of extreme values will still be sizable, and this does not agree with Normal tails.

4.3.1 Leveraging heavy-tailedness: load balancing in an Internet AS

It has been observed that the size of TCP flows in an Internet AS is HT. This means that there are very many "mice", insignificant flows which transmit few bytes, and very few "elephants", huge flows that make the bulk of the traffic.

How can you solve this through load balancing? Given that re-routing a flow has a cost, is this going to be costly?

A cheap and effective load-balancing scheme would be **re-route some of the elephants**. This would be:

- **Effective**: by moving one elephant you move a **large portion** of the traffic in your domain. Balancing elephants is the only way to achieve load balancing at all.
- **Cheap**: there are very few elephants, so there will be very **few rerouting decisions**. Most of the traffic (i.e., mice) will not be affected.

How do you mark a flow for rerouting? i.e., how do you identify a (future) elephant? You can't, barring clairvoyance. That is because you cannot foresee **how many more packets** a flow will send before stopping.

However, you can **rely on the memory principle**: if something looks unusually large, chances are that it will be **very large**. Therefore, you can set (empirically) a threshold after which flows are likely to be elephants. If you set that threshold carefully, you will achieve near-perfect load balancing at a very low cost.

Could you have done the same if flow sizes were Normal, instead? The answer is no. You would need to move a lot of flows in order to balance your load, and the cost would have been too high.

5 Review problems

5.1 Problem 1 – Roulette

A roulette wheel has 38 slots, numbered 0, 00, and 1 through 36. If you bet 1 on a specified number, you either win 35 if the roulette ball lands on that number or lose 1 if it does not. If you continually make such bets, approximate the probability that

a) you are gaining money after 34 bets;
b) you are gaining money after 1,000 bets;
c) you are gaining money after 100,000 bets.

Assume that each roll of the roulette ball is equally likely to land on any of the 38 numbers.
Note that "gaining money after x bets" means that you have strictly more money than you started with.

5.1.1 Solution

First of all, what do you expect will happen as the number of bets increases? To answer this question, one should compare the odds of winning and the related payoff. Now, on every bet, your winning probability is 1/38, and the odds are 35:1. In the long run, you will win once every 38 games, and get 35 times your bet. This means that the more you play, the more you lose, on average. Thus, we would expect that the probability is going to be smaller from a) to c).
How do we solve this? This is an exercise that deals with repeated trials in independent conditions, which makes you think of binomial variables. However, although the three bullets require the same computation, they are very different settings, particularly as the number of trials is concerned. This means that we may want to try different approximations of a binomial distribution as we scale upward in the number of trials.

The success probability is always $p = 1/38$. Let's see what happens in the three cases:
a) since you win 35 times your bet, you are gaining if you have won *at least once* in the whole run of 34 bets. Therefore, we must compute the probability of at least one success in N=34 independent trials. This is equal to:

$$
\begin{aligned}
&p_{gain} = P\{k > 0\} \\
&= 1 - P\{k = 0\} \\
&= 1 - \left[\binom{N}{0} p^0 \cdot (1-p)^{N-0}\right] \\
&= 1 - \left(\frac{37}{38}\right)^{34} \\
&= 0.596
\end{aligned}
$$

Note that, for *any number of bets n* from 1 to 35, you have the following expression:

$$p_{gain}(n) = 1 - \left(\frac{37}{38}\right)^n \quad 1 \le n \le 34$$

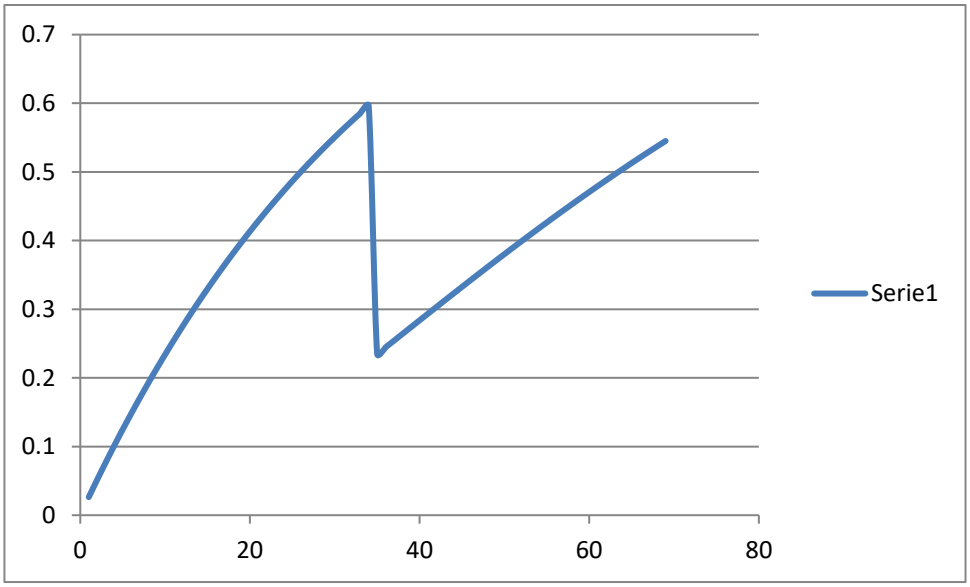
where $p_{gain}(n)$ is the probability of gaining after *j* bets. $p_{gain}(n)$ increases with *j*, which means that you should keep playing if you want to maximize your chance of winning. This is apparently in contrast with our earlier intuitive notion that the more you play, the more you lose. How do we reconcile the two?

What happens at the 36th bet? It happens that winning *once* is not enough anymore in order for you to be gaining. You need to win at least **twice**, until the 71th bet included. This means that, for $36 \le n \le 71$,

$$
\begin{aligned}
&p_{gain}(n) = P\{k > 1\} \\
&= 1 - P\{k = 0\} - P\{k = 1\} \\
&= 1 - \left[\binom{n}{0} p^0 \cdot (1-p)^{n-0}\right] - \left[\binom{n}{1} p^1 \cdot (1-p)^{n-1}\right] \\
&= 1 - \left(\frac{37}{38}\right)^n - n \cdot \frac{1}{38} \cdot \left(\frac{37}{38}\right)^{n-1} \\
&= 1 - \left[\left(\frac{37}{38}\right)^n + \frac{n}{37} \cdot \left(\frac{37}{38}\right)^n\right] \\
&= 1 - \left(1 + \frac{n}{37}\right) \cdot \left(\frac{37}{38}\right)^n
\end{aligned}
$$

And this is still an increasing function of n in $36 \le n \le 71$. It is $p_{gain}(36) = 0.234$, and $p_{gain}(71) = 0.545$.

This means that, again, you increase the chances to end up winning, but these chances are peaking towards a smaller peak. If you bet for the 72th time, you will need to win **at least three times**, and so on and so forth, and the chances to end up winning will increase towards a peak that gets smaller and smaller. This means that, in the end, you are just increasing the chances that you will lose, consistently with what we were suggesting based on intuition.



b) In N bets, you end up gaining if:

$$n_{win} > \frac{n_{loss}}{35} = \frac{N - n_{win}}{35}$$
$$n_{win} > \frac{N}{36}$$

This means that you need at least $n_{win} = \left\lfloor \frac{N}{36} \right\rfloor + 1$, which in our two cases is $n_{win} = 28$ and $n_{win} = 2778$ respectively. It is risky (at the very least) to try and compute $P\{k \geq n_{win}\}$ using binomial formulas, since you have high values of N (1000 and 100000, respectively), which imply huge factorials and numerical instability. Therefore, approximations are to be used. There are two approximations that one may try for binomial variables:

- the Poisson approximation, which holds if N is large and p is small. In this case $N = 1000,100000$ and $p = 1/38 = 0.0263$, which means that the conditions are met. A Poisson approximation is enforced by setting $\lambda = Np$
- the Gaussian approximation, which holds if $Np(1-p) > 10$. In our case, $Np(1-p) \approx 25.62,\ \ 2562$, respectively, so we are in a position to use this as well. In the Gaussian approximation, it is $\mu = Np$ and $\sigma^2 = Np(1-p)$

Among the two, the Gaussian is normally the quickest route, because it does not require summations. Let us try both for case b)

Using the Poisson approximation, we set $\lambda = Np \approx 26.32$ and we need to compute the probability that a Poisson variable with that mean value is greater than or equal to $N_{win} = 28$:

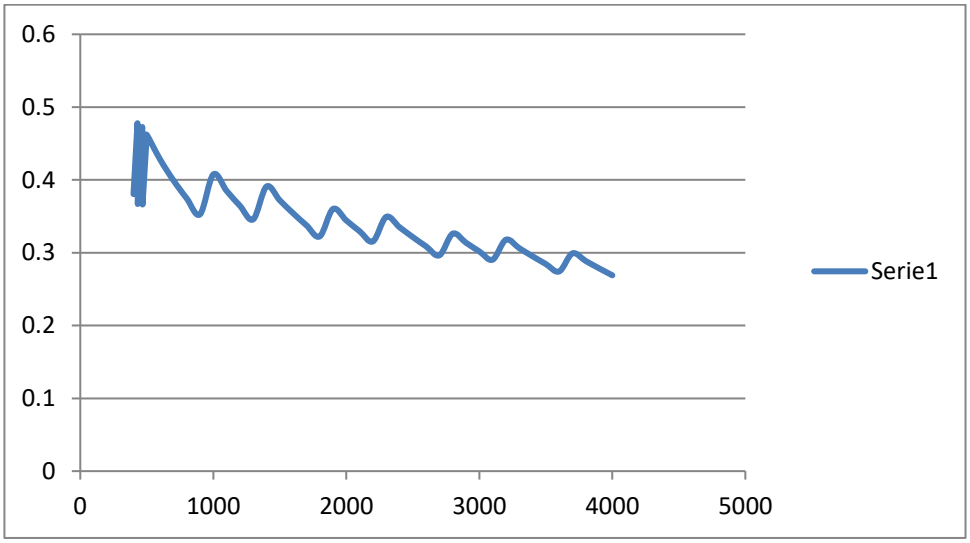
$$P\{N_{win} \geq 28\} = 1 - \sum_{j=0}^{27} e^{-\lambda} \cdot \frac{\lambda^j}{j!}$$

This can be easily done with Excel, and yields a result $p_{gain}(1000) \approx 0.397$.

To use a Gaussian approximation, which holds by the Central Limit Theorem, you need to set the following: $\mu = N \cdot p = 26.31, \quad \sigma^2 = N \cdot p \cdot (1-p) = 25.62$. With these numbers, you need to compute $P\{N_{win} \geq 27.5\}$, with $N_{win} \sim N(\mu, \sigma^2)$. Recall that it should be 27.5 (and not 28), because of the so-called "continuity correction". In this case, we have:

$$\begin{aligned} &P\{N_{win} \geq 27.5\} = 1 - P\{N_{win} \leq 27.5\} \\ &= 1 - P\left\{\frac{N_{win} - \mu}{\sigma} \leq \frac{27.5 - \mu}{\sigma}\right\} \\ &= 1 - \Phi(0.234) \\ &= 0.4052 \end{aligned}$$

Note that this result is entirely consistent with the one computed using the Poisson approximation.



c) $N = 100000$ In this case, we might in principle repeat the above procedure, only to observe that the Poisson approximation is itself numerically instable. Therefore, we are only left with the Gaussian approximation, which yields the following:

$$\mu = N \cdot p = 2631, \quad \sigma^2 = N \cdot p \cdot (1-p) = 2562$$

$P\{N_{win} \geq 2777.5\}$, with $N_{win} \sim N(\mu, \sigma^2)$. Note that $\sigma = \sqrt{2562} \approx 50.62$. This means that the required value is almost 3 standard deviations away from the mean, which means that it is very unlikely to come out. Putting numbers in, we get:

$$
\begin{aligned}
&P\{N_{win} \geq 2777.5\} = 1 - P\{N_{win} \leq 2777.5\} \\
&= 1 - P\left\{\frac{N_{win} - \mu}{\sigma} \leq \frac{2777.5 - \mu}{\sigma}\right\} \\
&= 1 - \Phi(2.88) \\
&= 0.0021
\end{aligned}
$$

This confirms that the probability of winning gets lower the more you play (though not monotonically so).

♦

5.2 Problem 2 – Voting

Consider a state with voter population N. There are two candidates in the state election for governor and the winner is chosen based on a simple majority. Let N_1 and N_2 be the total number of votes obtained by candidates 1 and 2, respectively, from voters other than Johnny. Johnny just voted for candidate 1, and he would like to know the probability that his vote affects the election results. Assume that each other voter (excluding Johnny) votes independently for candidates 1 and 2 with probabilities p_1, p_2, and also that $p_1 + p_2 < 1$ to allow for the case that a voter chooses not to vote for either candidate. Derive a formula for the probability that Johnny's vote affects the election results

5.2.1 Solution

First of all, what does it mean that Johnny's vote affects the elections? It may mean that either:

- $N_1 = N_2$ (Johnny's vote makes candidate 1 win the election)
- $N_1 = N_2 - 1$ (Johnny's vote doesn't allow candidate 2 to win the election, calling it a draw)

A first step in solving this problem is to compute the following JPDF:

$$P\{N_1 = j, N_2 = k\} = P\{N_1 = j | N_2 = k\} \cdot P\{N_2 = k\}$$

We have in fact:

$$P\{N_1 = j | N_2 = k\} = \binom{N-1-k}{j} {p_1}^j \cdot (1 - p_1 - p_2)^{N-1-k-j}$$

$$P\{N_2 = k\} = \binom{N-1}{k} {p_2}^k \cdot (1 - p_2)^{N-1-k}$$

Therefore

$$P\{N_1 = j, N_2 = k\} = \binom{N-1-k}{j} {p_1}^j \cdot (1 - p_1 - p_2)^{N-1-k-j} \cdot \binom{N-1}{k} {p_2}^k \cdot (1 - p_2)^{N-1-k}$$

With this result, we can compute the solution as:

$$\sum_{k=0}^{\left\lfloor \frac{N-1}{2} \right\rfloor} P\{N_1 = k, N_2 = k\} + \sum_{k=0}^{\left\lfloor \frac{N}{2} \right\rfloor - 1} P\{N_1 = k, N_2 = k+1\}$$

♦

5.3 Problem 3 – Student tests

A student test consists of N questions, each one of which has k possible answers, among which the student is requested to select the only correct one ($N, k > 0$).

Jack tries his luck by making blind, independent guesses at each question.

1) Find the expression for the PMF P_n of the number n of correct answers in the whole test
2) Discuss possible approximations of P_n when $N = 80$, $k = 5$, and compute an approximated value for the probability that Jack answers 40 or more questions correctly.
3) (For generic N and k values): assume that each correct answer earns Jack +1 points, and each wrong answer earns it -1 points. What is the mean value of Jack's score at the test s?
4) What is the score that one should get for a wrong answer, in order for the mean value of s to be equal to zero?

Now, assume that Jack has some knowledge in the subject matter of the test. He (thinks he) *knows* the answer to M questions ($0 \leq M \leq N$), which he sets apart in advance and answers to independently, and he *guesses* the remaining $N - M$ ones. When he *knows* the answer, he has 90% probability of actually getting the correct answer.

5) Compute the expression of the probability that Jack answers correctly to exactly M questions.

5.3.1 Solution

1) This is a repeated trial problem, with a probability of success equal to $p = 1/k$. The PMF is thus the well-known binomial PMF:

$$P_n = \binom{N}{n} p^n \cdot (1-p)^{N-n}$$

2) A binomial can be approximated with:
   a. A Poisson variable, if N is large and p is small. In this case, $p = 0.2$, hence this approximation incurs large errors.
   b. A Gaussian variable, if $N \cdot p \cdot (1-p) > 10$. With $N = 80$, $k = 5$, it is $N \cdot p \cdot (1-p) = 12.8$, hence the required condition holds.
   In this case, it is $P_n \approx P(n - 0.5 \leq X \leq n + 0.5)$, with $X \sim N\big(N \cdot p, N \cdot p \cdot (1-p)\big)$, i.e. $X \sim N(16, 12.8)$. Therefore, we have $P_n \approx P\left(\frac{n-0.5-16}{\sqrt{12.8}} \leq \frac{X-16}{\sqrt{12.8}} \leq \frac{n+0.5-16}{\sqrt{12.8}}\right)$.

With $n = 40$, it is

$$\begin{aligned} P\{n \geq 40\} &= 1 - \Phi\left(\frac{23.5}{\sqrt{12.8}}\right) \\ &\approx 1 - \Phi(6.57) \\ &\approx 0 \end{aligned}$$

Since notoriously $\Phi(x) \approx 0$ when $x \geq 3$. This also makes sense intuitively. Note that the answer obtained without approximations is $P\{n \geq 40\} \cong 2.06 \cdot 10^{-9}$.

3) The mean score for a question is $1 \cdot p + (-1) \cdot (1 - p) = 2p - 1$, and the sum of the means is equal to the mean of the sum. Therefore, the mean score for the whole test is $E[s] = N \cdot (2p - 1)$. This tells us that the mean test score is negative if the probability of success is less than 50% at each question, which makes sense intuitively.

   The same result could be obtained via a considerably longer route as follows: with n correct answers out of N, the test score will be $n - (N - n)$. Thus, the mean value of the score is $E[s] = \sum_{n=0}^{N} P_n \cdot [n - (N - n)] = 2\sum_{n=0}^{N} n \cdot P_n - N \cdot \sum_{n=0}^{N} P_n$. This can be further developed as:

$$\begin{aligned} E[s] &= 2\sum_{n=0}^{N} n \cdot P_n - N \cdot \sum_{n=0}^{N} P_n \\ &= 2\sum_{n=1}^{N} n \cdot \binom{N}{n} p^n \cdot (1-p)^{N-n} \quad - N \\ &= 2\sum_{n=1}^{N} N \cdot \binom{N-1}{n-1} p^n \cdot (1-p)^{N-n} \quad - N \\ &= 2N \cdot p \cdot \sum_{j=0}^{N-1} \binom{N-1}{j} p^j \cdot (1-p)^{(N-1)-j} \quad - N \\ &= 2N \cdot p - N \\ &= N \cdot (2p - 1) \end{aligned}$$

4) In order for the mean test score to be null, *some* test scores must be negative, and this can only be obtained if the score for a wrong answer is negative. Assume that a wrong answer gets $-\delta$, with $\delta > 0$. The mean test score is null if and only if the mean score of each question is null, and this happens if $1 \cdot p + (-\delta) \cdot (1 - p) = 0$, i.e. $\delta = p/(1 - p) = 1/(k - 1)$The same result could be obtained via the longer route, by observing that the mean test score is:

$$E[s] = \sum_{n=0}^{N} P_n \cdot [n - (N - n) \cdot \delta] = (1 + \delta)\sum_{n=0}^{N} n \cdot P_n - \delta N \cdot \sum_{n=0}^{N} P_n$$

   Based on the computations of the previous point, one clearly sees that the mean value for the test score is:

   $E[s] = (1 + \delta) \cdot N \cdot p - \delta N = N \cdot [(1 + \delta) \cdot p - \delta]$.

   Now, in order to be $E[s] = 0$, we need $\delta = p/(1 - p) = 1/(k - 1)$.

5) The probability that Jack gets n correct answers can be written as follows:

$$P_M{}' = \sum_{j=0}^{M} P\{(M-j)\text{correctguesses}, j\text{correct"known"answers}\}$$
$$= \sum_{j=0}^{M} (P_{M-j}|Q_j) \cdot Q_j$$

Where Q_j is the probability that Jack gets j correct answers among the M that he thinks he knows. Now, it is correct to assume that P_{M-j} and Q_j are independent of each other, hence the formula is:

$$P_M{}' = \sum_{j=0}^{M} P_{M-j} \cdot Q_j$$
$$= \sum_{j=0}^{M} \left[\binom{N-M}{M-j} p^{(M-j)} \cdot (1-p)^{(N-M)-(M-j)}\right] \cdot \left[\binom{M}{j} q^j \cdot (1-q)^{M-j}\right]$$
$$= p^M \cdot (1-p)^{(N-M)} \cdot \sum_{j=0}^{M} \binom{N-M}{M-j} \cdot \binom{M}{j} \cdot \left[\left(\frac{q}{p}\right)^j \cdot \left(\frac{1-q}{1-p}\right)^{M-j}\right]$$

Where $q = 0.9$ is the probability that Jack answers correctly to a question whose answer he thinks he knows. Note that the above summation can start from $j = \max(0, 2M - N)$, since the first binomial coefficient is null if $j < 2M - N$.

♦

5.4 Problem 4 – Independent measures

Two measurements X and Y are independently drawn from the same distribution with mean μ and variance σ^2, and a weighted sum $S = wX + (1-w) \cdot Y$ is computed, with $0 \le w \le 1$.

1) Find μ_S and σ_S^2.

2) Find the value of w that minimizes σ_S^2 and the minimum value for σ_S^2. Find an intuitive explanation for the findings.

3) Under the hypotheses of point 2), assuming that $\mu = 3, \sigma^2 = 8$ and that the distribution is symmetric around the mean value, compute $P\{S \le 7\}$.

4) Answer the above question again assuming that X and Y are normal.

5) Assume now that X and Y are *not* independent, and that $E[X \cdot Y] = \mu^2 + \Delta$. Answer again point 1). Is it possible that σ_S^2 decreases w.r.t. the previous case?

5.4.1 Solution

1) $\mu_S = E[S] = E[wX + (1-w)\cdot Y] = w\cdot E[X] + (1-w)\cdot E[Y] = w\cdot\mu + (1-w)\cdot\mu = \mu$

Since X and Y are i.i.d.,

$$Var(S) = Var(wX + (1-w)\cdot Y) = Var(wX) + Var\big((1-w)\cdot Y\big) = [w^2 + (1-w)^2]\cdot\sigma^2 = [2w^2 - 2w + 1]\cdot\sigma^2$$

2) The required values are the coordinate of the vertex of parabola $y = 2w^2 - 2w + 1$, i.e. $w' = 1/2$, and $y' = 1/2$. Hence, the minimum value for σ_S^2 is $\sigma^2/2$. The intuitive explanation is that, when $w' = 1/2$, S is the average of X and Y. By the central limit theorem, the average of *n i.i.d.* random variables has a smaller variance than the individual variables.

3) We have $7 = \mu_S + 2\sigma_S$. Hence, by Tchebishev's inequality, it is $P\{|S-\mu_S| \geq k\cdot\sigma_S\} \leq 1/k^2$, with $k = 2$. This means that $P\{S > 7\} + P\{S < -1\} = 1/4$. Therefore, since the RV is symmetric around the mean value, it is $P\{S > 7\} = 1/8$, hence $P\{S \leq 7\} = 7/8$.

4) If X and Y are normal, it is $P\{S \leq 7\}$=$P\{S \leq \mu_S + 2\sigma_S\} = P\left\{\frac{S-\mu_S}{\sigma_S} \leq 2\right\} = \Phi(2) = 0.9772$

5) μ_S stays the same, since the expectation is linear whether RVs are independent or not. The variance changes, and it incorporates the *covariance* between X and Y. More specifically, it is:

$$\begin{aligned}
&Var(S) = Var(wX + (1-w)\cdot Y)\\
&= Var(wX) + Var\big((1-w)\cdot Y\big) + 2\cdot Cov(wX, (1-w)\cdot Y)\\
&= w^2\cdot\sigma^2 + (1-w)^2\cdot\sigma^2 + 2\cdot w\cdot(1-w)\cdot[E[X\cdot Y] - \mu^2]\\
&= [2w^2 - 2w + 1]\cdot\sigma^2 + 2\cdot w(1-w)\cdot\Delta
\end{aligned}$$

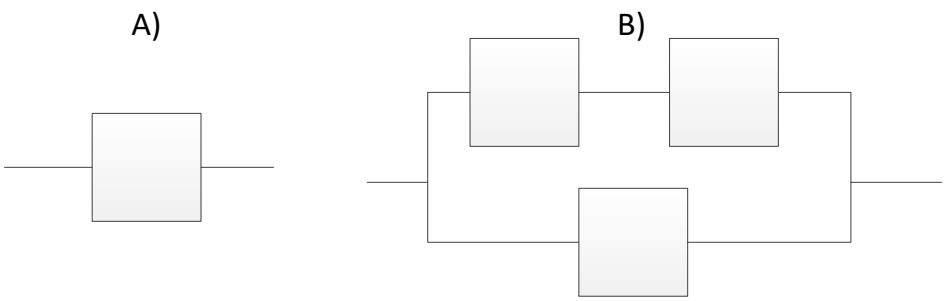
If $\Delta < 0$ (i.e. variables are negatively correlated), then the variance of S is actually smaller. This is because a large sample for X will be compensated by a smaller sample for Y and vice versa.

♦

5.5 Problem 5 – Switches

ACME components owns two switch production plants. In plant 1, each unit is defective with probability $p_1 = 10^{-5}$, independently from the others. In plant 2, the mean weekly number of defective units is equal to 5. The production of each plant is $n = 4\cdot 10^5$ units per week.

1) Compute mean and variance of the number of defective units produced by ACME in a week.
2) Draw a *qualitative* plot (with as many details are possible) of the PMF of the number of defective units in a week.
3) Compute the probability that the weekly number of defective units produced by ACME is equal to 5.
4) Compute the probability that the weekly number of defective units produced by ACME is less than 3.
5) Compute the probability that a randomly chosen unit is defective.



Suppose now that ACME units can be connected in series or in parallel as above, and that the resulting system works if there exists a way that connects both extremities traversing only non-defective systems.

6) Explain which of the two systems has a higher chance to be functioning. Jusitfy your findings.

5.5.1 Solution

1) Given that n is large and p is small, we can approximate the failure probability of each plant using a Poisson variable, whose average is $\lambda_i = n_i \cdot p_i$. Hence, it is $\lambda_1 = 4, \quad \lambda_2 = 5$. Thus, there are on average 9 defective units in a weekly production of $2n = 8 \cdot 10^5$ pieces. As for the variance, it is all the more reasonable to approximate the whole production using a Poisson variable, whose average and variance is equal to 9.
2) The Poisson variable has a bell shape, with an infinite right tail. It peaks around its mean value, which is equal to 9. Hence, the shape is the following:

3) The probability that 5 pieces are defective is equal to $p_5 = e^{-9} \cdot 9^5/5! = 0.060727$
4) The probability that less than 3 pieces are defective is equal to $p_0 + p_1 + p_2 = 1.23 \cdot 10^{-3} + 11.1 \cdot 10^{-3} + 49.98 \cdot 10^{-3} = 62.32 \cdot 10^{-3}$

5) The probability is the following:

$$\begin{aligned}
&p_d = P\{\text{defective}\} \\
&= P\{\text{defective}|\text{plant1}\} \cdot P\{\text{plant1}\} + P\{\text{defective}|\text{plant2}\} \cdot P\{\text{plant2}\} \\
&= 10^{-5} \cdot 0.5 + \frac{5}{4 \cdot 10^5} \cdot 0.5 \\
&= 1.125 \cdot 10^{-5}
\end{aligned}$$

6) System a) works with probability $p_a = 1 - p_d$. System b) works with probability

$$\begin{aligned}
&p_b = 1 - P\{\text{upperbranchfails}\} \cdot P\{\text{lowerbranchfails}\} \\
&= 1 - (1 - (1 - p_d)^2) \cdot p_d
\end{aligned}$$

Thus, $p_b > p_a$ if and only if

$$\begin{aligned}
&1 - (1 - (1 - p_d)^2) \cdot p_d > 1 - p_d \\
&p_d > (1 - (1 - p_d)^2) \cdot p_d \\
&1 > 1 - (1 - p_d)^2 \\
&p_d < 1
\end{aligned}$$

which is always true. System b) is always more reliable than system a), no matter what the failure probability of a single component is.

5.6 Problem 6 - Dice

In a dice game, you roll two dice. If you obtain a 2, 3, or 12, you immediately lose. If, instead, you obtain a 7 or 11, you immediately win. If you roll a 4, 5, 6, 8, 9, or 10, that becomes your "objective". In this case, you keep rolling the dice until either the "objective" comes up again – in which case you win – or until a 7 comes up, in which case you lose.

1. Calculate the probability that you win/lose at the first roll
2. Calculate the probability that you obtain a 4 at the first roll, and you win/lose at the second roll
3. Generalize the previous result to the case when you win/lose at the *n*-th roll (*n*>=2)
4. Using the previous two results, compute the probability that you win *at all*
5. Assume that you can bet 100$ on you winning. Compute the mean value of your payoff

5.6.1 Solution

1) Call $P_x = P\{x\}$

$P\{\text{win1}^{\text{st}}\} = P\{7,11\} = P_7 + P_{11} = \frac{6+2}{36} = \frac{8}{36}$;

$$P\{\text{lose1}^{\text{st}}\} = P\{2,3,12\} = P_2 + P_3 + P_{12} = \frac{1+2+1}{36} = \frac{4}{36}$$

2) $P\{41^{\text{st}}, \text{win}2^{\text{nd}}\} = P\{4,4\} = P_4 \cdot P\{4|4\} = P_4 \cdot P_4 = \frac{3}{36} \cdot \frac{3}{36} = \frac{9}{36^2}$

$$P\{41^{\text{st}}, \text{lose}2^{\text{nd}}\} = P\{4,7\} = P_4 \cdot P\{7|4\} = P_4 \cdot P_7 = \frac{3}{36} \cdot \frac{6}{36} = \frac{18}{36^2}$$

The last inequalities are due to the fact that subsequent rolls are independent experiments.

3) In order to win (lose) at the *n*-th roll, you have to obtain a result which is not in {4,7} for *n*-2 times before getting a 4 (7) again.

$$P\{41^{\text{st}}, \text{win}n^{th}\} = P\{4, [\sim(4,7)^{n-2}], 4\} = P_4 \cdot (1 - P_4 - P_7)^{n-2} \cdot P_4 = \frac{3}{36} \cdot \left(\frac{27}{36}\right)^{n-2} \cdot \frac{3}{36}$$

$$P\{41^{\text{st}}, \text{lose}n^{th}\} = P\{4, [\sim(4,7)^{n-2}], 7\} = P_4 \cdot (1 - P_4 - P_7)^{n-2} \cdot P_7 = \frac{3}{36} \cdot \left(\frac{27}{36}\right)^{n-2} \cdot \frac{6}{36}$$

4) We straightforwardly obtain:

$$\begin{aligned}
P_{win} &= P\{\text{win}1^{\text{st}}\} + \sum_{n=2}^{+\infty} \sum_{x \in \{4,5,6,8,9,10\}} {P_x}^2 \cdot (1 - P_x - P_7)^{n-2} \\
&= P\{\text{win}1^{\text{st}}\} + \sum_{x \in \{4,5,6,8,9,10\}} {P_x}^2 \cdot \sum_{n=2}^{+\infty} (1 - P_x - P_7)^{n-2} \\
&= P\{\text{win}1^{\text{st}}\} + \sum_{x \in \{4,5,6,8,9,10\}} \frac{{P_x}^2}{P_x + P_7} \\
&= \frac{8}{36} + 2 \cdot \left[\frac{\left(\frac{3}{36}\right)^2}{\left(\frac{3+6}{36}\right)} + \frac{\left(\frac{4}{36}\right)^2}{\left(\frac{4+6}{36}\right)} + \frac{\left(\frac{5}{36}\right)^2}{\left(\frac{5+6}{36}\right)} \right] \\
&= \frac{8}{36} + 2 \cdot \left[\frac{\left(\frac{9}{36}\right)}{9} + \frac{\left(\frac{16}{36}\right)}{10} + \frac{\left(\frac{25}{36}\right)}{11} \right] \\
&= \frac{2}{9} + \frac{1}{18} + \frac{4}{45} + \frac{25}{198} \\
&= \frac{244}{495} \simeq 0.493
\end{aligned}$$

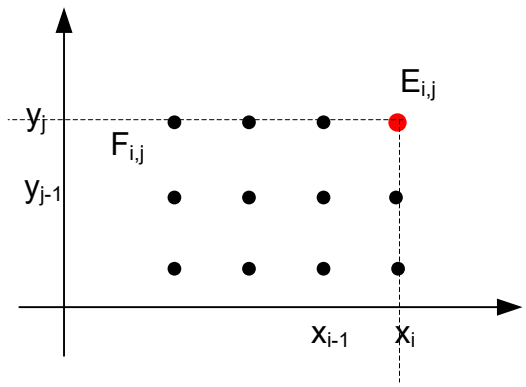
5) The expected payoff is $E[X] = 100 \cdot P_{win} - 100 \cdot (1 - P_{win}) = 100 \cdot (2P_{win} - 1) = -\frac{700}{495}$

5.7 Problem 7 – JPMF from JCDF

Given the JCDF $F(x_i, y_j)$ of two discrete RVs *X* and *Y*, compute the JPMF $p(x_i, y_j)$

5.7.1 Solution

Assume that (x, y) are points on a Cartesian plane, as in the figure.



Define $E_{i,j}$ the event $\{X = x_i, Y = y_j\}$, and $F_{i,j}$ the event $\{X \leq x_i, Y \leq y_j\}$. It is straightforward to observe that $P(E_{i,j}) = p(x_i, y_j)$, and $P(F_{i,j}) = F(x_i, y_j)$.

We straightforwardly obtain:

$$P(F_{i,j}) = P(F_{i-1,j} \cup F_{i,j-1} \cup E_{i,j}) = P(F_{i-1,j} \cup F_{i,j-1}) + P(E_{i,j})$$

Where $F_{i-1,j}$, $F_{i,j-1}$ are the rectangles that terminate one point to the left (or one point below) (x_i, y_j). The last passage is true because $E_{i,j}$ and $F_{i-1,j} \cup F_{i,j-1}$ are mutually exclusive, i.e. $E_{i,j} \cap (F_{i-1,j} \cup F_{i,j-1}) = \emptyset$.

In order to compute $P(F_{i-1,j} \cup F_{i,j-1})$ we need to point out their intersection (the two events are *not* mutually exclusive). Their intersection is in fact $F_{i-1,j-1}$, hence:

$$\begin{aligned} &P(F_{i-1,j} \cup F_{i,j-1}) = P(F_{i-1,j}) + (F_{i,j-1}) - P(F_{i-1,j}F_{i,j-1}) \\ &= P(F_{i-1,j}) + (F_{i,j-1}) - P(F_{i-1,j-1}) \end{aligned}$$

Now, we have everything we need to do the computations:

$$\begin{aligned} &p(x_i, y_j) = P(E_{i,j}) \\ &= P(F_{i,j}) - P(F_{i-1,j} \cup F_{i,j-1}) \\ &= P(F_{i,j}) - [P(F_{i,-1j}) + P(F_{i,j-1}) - P(F_{i-1,j-1})] \\ &= F(x_i, y_j) - [F(x_{i-1}, y_j) + F(x_i, y_{j-1}) - F(x_{i-1}, y_{j-1})] \end{aligned}$$

6 Appendix

6.1 Tables

TABLE 1 - Standard Normal Distribution Function $\Phi(x)$

x	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

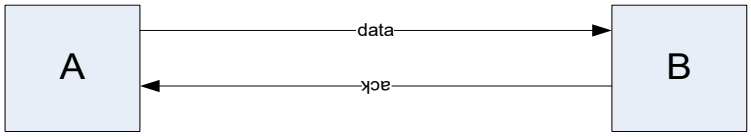
TABLE 2 - Values of $t_{\alpha,n}$

n	α =0.10	α =0.05	α =0.025	α =0.01	α =0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.474	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
∞	1.282	1.645	1.960	2.326	2.576

6.2 Geometric RVs: capacity of data-link protocols

We can exploit geometric RVs to derive insight about the **capacity of two data-link** network protocols.

A **data-link** network protocol transmit **frames** over a wire between two endpoints. The usual assumption is that the link is **full-duplex**, and that we consider only one side of the transmissions (i.e., A sending to B). A send **data frames** to B, which replies with **control (acknowledgment) frames**. Data frames have some **control bits**, l_c, and some **data bits**, l_d. Ack frames only have control bits l_c. Control information is required (back and forth) to understand what is going on, wheras data information is what A wants to transmit to B. It is safe to expect that $l_c \ll l_d$, otherwise the efficiency of the protocol will be low (hampered by the control overhead).



The **capacity of a network protocol** is **the percentage of time** it keeps the transmission line occupied, assuming that A always has something to transmit.

The following facts will guide our analysis (this is **networking 101**):

- Data frames are **numbered**, and ACK frames carry the number of the data frame they acknowledge.
- A frame (whether data or ACK) can be **corrupted** on the fly, with a certain **frame error probability** p that depends on the transmission medium.
- When a frame (data or control) is corrupted, its receiver **cannot decode it**, hence **discards it** and the protocol behaves as if no transmission had occurred at all.
- When B receives a frame correctly, it sends back an acknowledgement immediately
- When A receives an acknowledgemnt, it knows that the frame the ACK refers to has been received correctly.
- If **either the data or the ACK frame is corrupted**, A will not receive an ACK. The two cases are indeed different:
    - If the **data** frame is corrupted, B will **not** send the ACK back
    - If the data frame is correctly received, but the **ACK is corrupted**, B will not be able to decode the ACK.

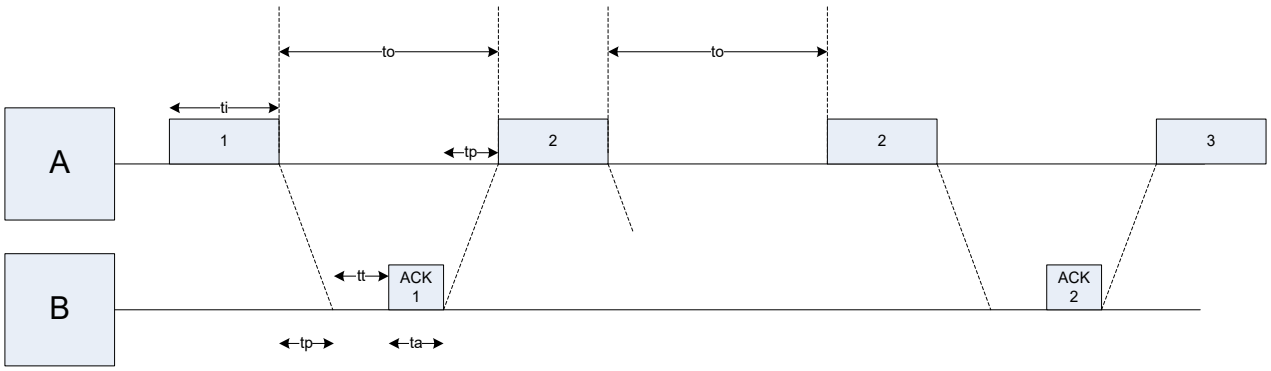
  In the first case, the data frame is at B, in the second it is not. However, A has no way of distinguishing the two cases, hence it will behave the same, and assume the worst.

- Since you only get **positive** feedback (i.e., when you receive an ACK), you can only rely on **timeout and retransmission** at the sender side. In fact, A retransmits frames when no ACK is received after a timeout. Unless he medium is irreparably damaged, an ACK will make it to A sooner or later, possibly after a certain number of retansmissions.

Call:
- t_p the propagation time on either direction (which is the link length divided by the speed of light in the medium).
- C the link speed (e.g., bits per second)
- $t_i = (l_c + l_d)/C$ the transmission time of a data frame
- $t_c = l_c/C$ the transmission time of an ACK frame
- t_t the maximum "think time" at the receiver B, before it actually sends an ACK back to A.

We want to see what capacity we get how using different protocol strategies.

The simplest strategy is called **<u>stop and wait</u>**. The sender sends one frame, waits for the ACK, and then goes on like this.



Since we know all the times involved, A can set the timeout equal to $t_o = 2 \cdot t_p + t_t + t_c$.

We can already foresee that this protocol keeps the medium busy **only when** $t_i \gg t_o$, since every time A transmits a frame, it has to keep idle for t_o **(at best)** before transmitting another. **In the absence of errors**, in fact, the capacity would be:

$$\rho = \frac{t_i}{t_i + t_o} < 1$$

After some straightforward computations, we would get:

$$\rho = \frac{l_c + l_d}{(2 \cdot C \cdot t_p + C \cdot t_t + l_c) + (l_c + l_d)} = \frac{1}{\frac{2 \cdot C \cdot t_p + C \cdot t_t + l_c}{l_c + l_d} + 1}$$

Hence, the obvious way to maximize the procotol capacity, if we can act on it, is to make **data frames infinitely long** ($l_d \to \infty$).

Since we **do have errors, instead**, and errors depend on the frame length (as we will see later on), this is not the right thing to do. Assume that l_d is fixed, and call p the probability that there is an errore (either in the data or in the ACK frame). Assume subsequent transmissions are **independent of each other**. In this case the right probability model are **repeated trials**, where $(1-p)$ is the success probability.

The transmission time of a frame is a **random variable** T, which can be computed as follows:

- $T = t_i + t_o$ with probability $(1-p)$
- $T = 2 \cdot (t_i + t_o)$ with probability $p \cdot (1-p)$
- $T = k \cdot (t_i + t_o)$ with probability $p^{k-1} \cdot (1-p)$ (for $k \geq 1$)

The capacity of the protocol in this case is:

$$\rho = \frac{t_i}{E[T]}.$$

Let us compute $E[T]$:

$$\begin{aligned}
E[T] &= \sum_{k=1}^{+\infty} k(t_i + t_o)p^{k-1}(1-p) \\
&= (t_i + t_o) \cdot (1-p) \cdot \left[\sum_{k=1}^{+\infty} k \cdot p^{k-1}\right] \\
&= (t_i + t_o) \cdot (1-p) \cdot \left[\frac{\partial}{\partial p} \sum_{k=1}^{+\infty} p^k\right] \\
&= (t_i + t_o) \cdot (1-p) \cdot \left[\frac{\partial}{\partial p} \frac{1}{1-p}\right] \\
&= (t_i + t_o) \cdot \cancel{(1-p)} \cdot \left[\frac{1}{(1-p)^{\cancel{2}}}\right] \\
&= \frac{t_i + t_o}{1-p}
\end{aligned}$$

Call $a = \frac{t_i + t_o}{t_i}$, then it is:

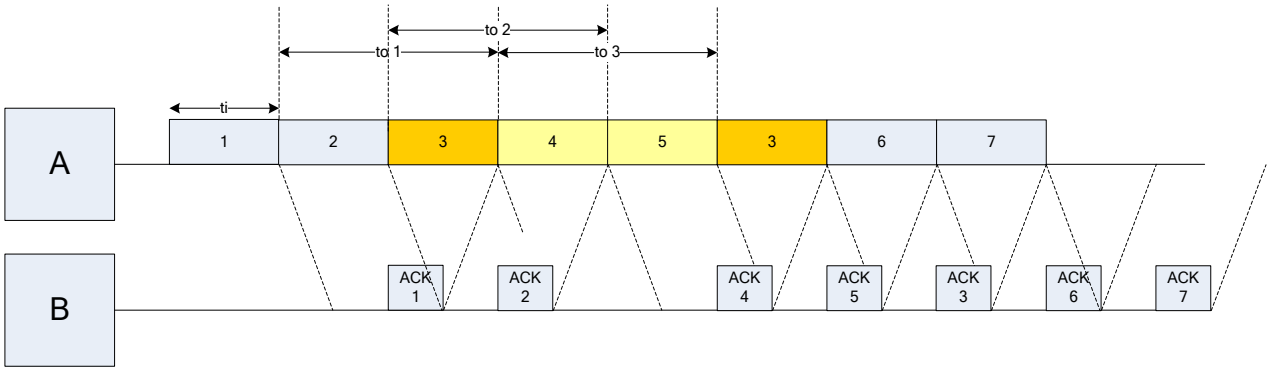
$$\rho = \frac{t_i}{a \cdot t_i} \cdot (1-p) = \frac{1-p}{a}$$

This tells us that **the capacity depends on:**

- the error probability: the higher the probability, the lower the capacity, which makes sense
- the **time overhead** a: with a null error probability, we get the same result as before.

The above computations were made assuming that the sender A is in **asymptotic conditions**, i.e., it always has something to transmit. This is the usual assumption when computing the capacity of a protocol.

Let us describe a different data link protocol, called **<u>selective repeat</u>**: in the latter, A transmits continuously, without waiting for the ACK of each frame. The receiver ACKs frames as they come. If a frame is missing (e.g., 1, 2, 4 arrive, but 3 is missing), then the receiver B either does nothing or sends a NACK3 (negative acknowledgement) to signal that 3 is missing. When a frame is in time out (or a NACK is received), the sender retransmits that frame only.



This requires a more complicated logic at the receiver, because frames may not arrive in sequence (with stop-and-wait, they always do).

We expect this protocol to have a **higher capacity** than the former. Let us see why:

The transmission time of a frame is a **random variable** T, which can be computed as follows:

- $T = t_i$ with probability $(1-p)$ **(compare with the former).**
- $T = t_i + (t_i + t_o)$ with probability $p \cdot (1-p)$
- $T = t_i + k \cdot (t_i + t_o)$ with probability $p^k \cdot (1-p)$ (for $k \geq 0$)

The capacity of the protocol is still:

$$\rho = \frac{t_i}{E[T]}.$$

Let us compute $E[T]$:

$$
\begin{aligned}
E[T] &= \sum_{k=0}^{+\infty} [t_i + k(t_i + t_o)] p^k (1-p) \\
&= \sum_{k=0}^{+\infty} t_i \cdot p^k (1-p) + \sum_{k=0}^{+\infty} k \cdot (t_i + t_o) p^k (1-p) \\
&= t_i \cdot (1-p) \sum_{k=0}^{+\infty} p^k + (t_i + t_o)(1-p) \sum_{k=0}^{+\infty} k \cdot p^k \\
&= t_i + \frac{p \cdot (t_i + t_o)}{1-p} \\
&= t_i + \frac{p \cdot a \cdot t_i}{1-p} \\
&= \frac{1 + p \cdot (a-1)}{1-p} t_i
\end{aligned}
$$

Thus, we have:

$$\rho = \frac{1-p}{1 + p \cdot (a-1)}$$

It is clear that the capacity still depends on the error probability. However, it does **not** depend anymore on a the way it did before. Even if a is large, if the error probability is *small*, then the capacity can get close to 1. With stop-and-wait, it could at most get to $1/a$.

So far we have assumed that the transmission time of a frame is **constant**. We have discussed thtat **long frames** are preferable, since they carry more bit and by doing so the timeout is amortized on more information.

However, the error probability must ultimately depend on the **length of the frame itself**. So far, both were assumed to be constant. Therefore, we can imagine that, the longer the frame (i.e., the higher l_d, since l_c is fixed and imposed by the protocol), the higher p will be.

A frequent model for the frame error probability is one of **independent bit errors**. Given a **bit error rate (BER)** p_b, i.e. the probability that one bit is corrupted, and assuming that all bits can be corrupted independently (which is not strictly true, anyway), the frame error probability is:

$$p = 1 - (1 - p_b)^{l_c + l_d}$$

There are **two contrasting effects at work** here, i.e.,

- a higher l_d brings benefits because it amortizes the control and timeout overhead
- a higher l_d increases the error probability, hence makes retransmission more frequent, hence it increases $E[T]$ and ultimately decreases the capacity.

As often happens in these cases, there is an **optimal frame length**, i.e. the one where you achieve the maximum throughput.

First let us define the throughput, which is the amount of information per unit of time. One frame is transmitted in $E[T]$ time, hence we get that

$$\lambda = \frac{l_d}{E[T]} = \frac{l_d \cdot (1-p)}{t_i + p \cdot t_o} = \frac{l_d \cdot (1-p)}{\frac{l_c + l_d}{C} + p \cdot t_o} = \frac{l_d \cdot (1-p_b)^{l_c+l_d}}{\frac{l_c + l_d}{C} + [1-(1-p_b)^{l_c+l_d}] \cdot t_o}$$

Of course, we can compute the derivative and equate it to zero to get the optimal data size, given the propagation times (which are embedded in t_o), the link speed C and the BER p_b.

Some computations can be done quickly, if we approximate $p = 1-(1-p_b)^{l_c+l_d} \approx (l_c + l_d) \cdot p_b$. In this case, we get:

$$\lambda = \frac{l_d \cdot [1-(l_c+l_d) \cdot p_b]}{\frac{l_c + l_d}{C} + [(l_c+l_d) \cdot p_b] \cdot t_o} = \left(\frac{l_d}{l_d + l_c} - l_d \cdot p_b\right) \cdot \frac{1}{\frac{1}{C} + p_b \cdot t_o}$$

The first term grows with l_d, whereas the second decreases with l_d. This shows the two contributions at work. By deriving w.r.t. l_d, we get the optimal length of the data part:

$$\frac{\partial}{\partial l_d}\left(\frac{l_d}{l_d + l_c} - l_d \cdot p_b\right) = \frac{\cancel{l_d} + l_c - \cancel{l_d}}{(l_d + l_c)^2} - p_b$$

And if we equate it to 0, we get:

$$l_d = \sqrt{\frac{l_c}{p_b}} - l_c.$$

It makes perfect sense that the optimal length decreases when the BER increases. For instance, considering $l_c = 16$, $p_b = 10^{-4}$, we get $l_d = \sqrt{16 \cdot 10^4} - 16 = 400 - 16 = 384$ bits.