Probability Theory

- SOR 1110 -

by

Prof. Lino Sant & Dr. Mark Anthony Caruana

Lecturer: Mark A. Caruana
Room: 513
e-mail: mark.caruana@um.edu.mt

Probability Theory

- SOR 1110 -

by
Prof. Lino Sant
&
Dr. Mark Anthony Caruana

L-Università ta' Malta
Faculty of Science

Department of
Statistics &
Operations Research

Table of Contents

Suggested Test Books

Chung , K.L., **Elementary Probability Theory**
Ross, S., **Introduction to Probability Models**
Feller, W., **An Introduction to Probability Theory & its Applications**
Freund, J.E., **Mathematical Statistics**
Grimmett, G.R., Stirzaker, D.R., **Probability and Random Processes**
Renji, A., **Probability Theory**

1. Historical Background

Historically, the interest in problems related to probability goes back to Biblical times. However it took a long time for the subject to become established as one of the main fundamental branches of mathematics. Formally a gambler's dispute in 1654 is considered to have been started off the creation of a mathematical theory of probability by Blaise Pascal and Pierre de Fermat.

There were two main areas which initially favored the development of probability:

- games of chance,
- insurance and demographic problems.

A brief overview of historical development:
Gambling: interest in dice and card games; fra Luca Pacioli Summa de Arithmetica, Geometrica, Proportioni et Proportinalita (1494); Girolamo Cardano Liber de Ludo Alea" (circa 1550, printed 1663 posthumously).

In 1654 Pascal in correspondence with Chevalier de Méré and later with Fermat is considered by some to have laid the foundations of probability theory.

Pascal dedicated a considerable amount of his research effort to a number of problems in probability or related areas. The study of what is called Pascal's triangle related to work on binomial coefficients. We mention also his original work on decision theory with reference to "proofs" of God's existence and emphasis on the principle of mathematical induction.

Huygens' De Rationcinis in Ludo Aleae (1657) in pamphlet form ended up being the only mathematical text available on probability for half a century. Expectation was its most notable new concept.

John Graunt's Natural and Political Observations Made upon Bills of Mortality (1662) dealt with demographic problems.

James Bernouilli produced a classic text in probability Ars Conjectandi (1713 posthumously). It was in 4 parts:

- Reproduction of De Rationcinis in Ludo Aleae,
- Permutations and combinations,
- Standard problems in games of chance,
- Applications to civil, moral and economic problems.

Bernoulli's Law of Large Numbers is enunciated. It is a result of great importance with mathematical and philosophical repercussions. Eventually it was to give an impetus to the frequentist interpretation in its controversy the versus subjective interpretation.

In De Moivre's Doctrine of Chances (1713) there are numerous problems on dice throwing, drawing balls and life annuities. Work on a number of problems in probability displays a high level of sophisticattion especially with reference to calculating binomial probabilities. This included Stirling's formula : $n! \approx \sqrt{2\pi n} n^n e^{-n}$

During the 18th century there was an increasing interest in solving problems of wider application: estimation of observation errors, demographic problems, political and social phenomena related to the duration and intensity of certain processes.

The preferred interpretation of probability became that of a calculus of beliefs. To fully comprehend the world we need infinite intelligence. Granted our incomplete knowledge we can only hope to work with degrees of reasonableness.

Daniel Bernoulli's study (1734) of the problem of calculating the probability that planets with random orbits would end up on roughly one plane of rotation introduce randomness in celestial mechanics.

Paradoxes become abundant in probability. Bernoulli worked on one of the more famous ones, the St. Petersburg's Paradox (1738).

Laplace was to systematize and extend previously known results into a modern theory. His monumental work Mécanique Céleste had taken up 26 years and earned him tremendous respect all over the Continent and in Britain.

Eventually Laplace spent a lot of time on probability as well. His work on observational errors, least squares method is found in his Théorie Analytique de Probabilités (1812).

It consists of 3 sections:

- Preface: Essai philosophique sur les probabilités,
- Book 1: Calculus of generating functions (analysis),
- Book 2: General theory of probability.

Laplace divided the subject into:

- probability proper,
- limit theorems,
- mathematical statistics.

He gave laws for addition and multiplication of probabilities, worked on problems of combinatorial nature with binomial setting featuring prominently, on geometric problems, sampling problems, expectations.

Laplace's work and views led his followers to the inappropriate and unjustified, and at times blatantly incorrect, use of the theory in numerous fields. This resulted in probability earning a reputation for lack of rigour and downright quackery.

Around the 1840's there was a return to a frequentist approach away from the classical rational belief interpretation.

Theories of errors and variation became numerous. Gauss' work is of special note. Major shift in thinking about the error law was mostly due to Quetelet in the 1820's. Statistics got off the ground.

Numerous problems in social physics brought a shift towards notions of regularities of a statistical nature. Statistics was providing lots of tools in economic and biological studies in the last century.

Various new applications relating to the properties of molecules and the mechanical theory of heat, later on to be called statistical mechanics, surfaced in physics.

Later on the notion of uncertainty in physics was posing intractable difficulties for which more heavy mathematical artillery was required.

The theory of stochastic processes made enormous progress during the late nineteenth century and the beginning of the 20^{th} especially on the continent.

The modern axiomatization of probability was first given by A.N. Komogorov in 1933 using the formalization provided by set theory.

2. Combinatorial Probability

In many situations involving probability we are usually faced with a number of outcomes each of which are possible, but out of which eventually only one will actually happen. The first natural thing to try to do when analyzing such a situation is to count all the possible outcomes. **Probability has to concern itself with what can happen before an event has actually occurred.**

Counting the number of possibilities available within finite settings is a very basic problem in many fields of knowledge. The finiteness requirement is clearly necessary because we cannot count uncountable sets! Over the centuries many mathematicians have devised quite a large number of counting techniques. The branch of mathematics dedicated to the study of such counting techniques is called combinatorics.

There are 3 main theorems which provide formulas for counting systematically:
• multiplication principle,
• principle of permutations,
• principle of combinations.

2.1 Theorem (Multiplication Principle)

The task of making k successive decisions (or choices) is to be considered as a global strategy. Let the 1^{st} decision be open to n_1 distinct choices, the 2^{nd} decision be open to n_2 distinct choices, ... , k^{th} decision be open to n_k distinct choices.

Then there are $n_1 \times n_2 \times \ldots \times n_k$ different strategies.

Proof

The proof follows by induction. (Try it as an exercise!)

■

2.2 Examples

2.2.1 In how many ways can n persons be placed in k rooms?

Answer: The 1^{st} person can be placed in k rooms.

The 2^{nd} person can be placed in k *rooms.*

⋮

The n^{th} person can be placed in k rooms.

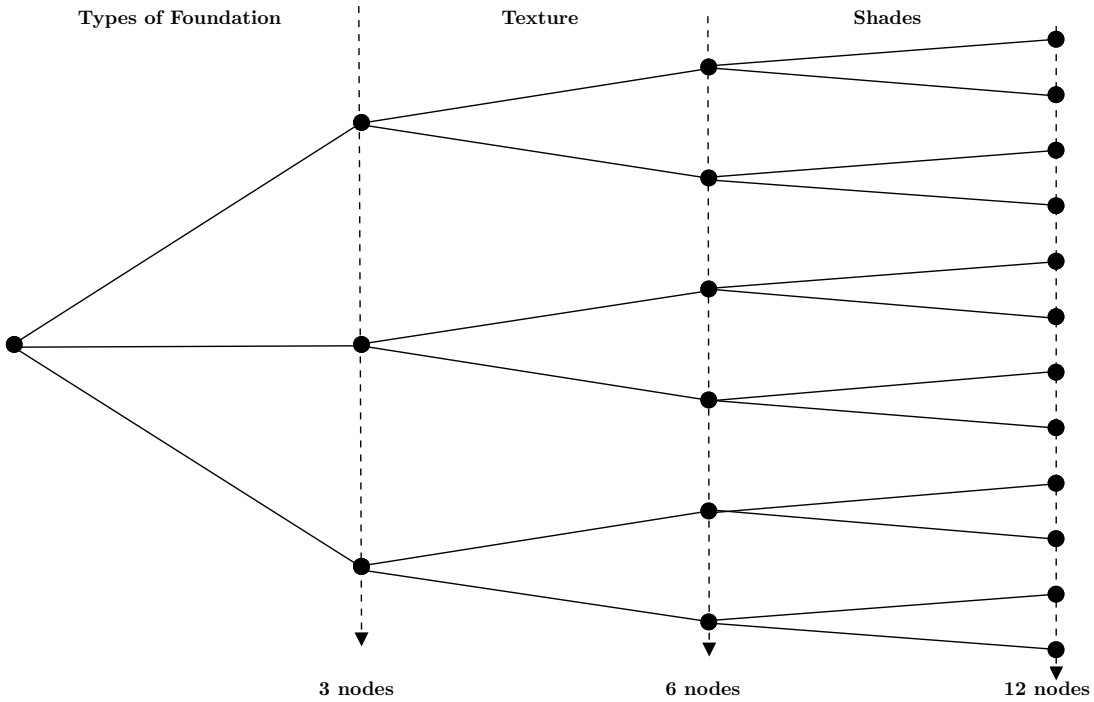
Thus, there are in all $\underbrace{k \times k \times \ldots \times k}_{\text{for n times}} = k^n$.

2.2.2 A menu with three courses having 10 choices for the first course, 20 for the second and 15 for the third course offers $10 \times 20 \times 15$ possible different meals in all.

2.2.3 Experiments are to be conducted on 6 different types of crops to see how their growth rate is affected by 5 different types of fertilizers during each of the 4 seasons. How many experiments have to be done?
Answer: $6 \times 5 \times 4 = 120$ experiments.

A much used aid in computing the number of possibilities in such situations is the tree diagram where successive branching points, or nodes, correspond to instances when decisions need to be taken and the number of forks correspond to the number of options available.

2.2.4 A company produces three types of make up foundation each specifically designed with a specific skin type in mind; oily, mature and dry. Each type of foundation comes in two textures, either liquid or compact. Furthermore, there are two shades: Pale Blush and English Rose. How many different types of products are there?



From the above we see that there are 12 nodes at the end of the tree diagram. Thus there are in all 12 different kinds of products.

2.3 Theorem (Principle of Permutations)

The number of ways of permuting (arranging) n distinct objects in k cells is given by:

$$n(n-1)...(n-k+1)=\frac{n!}{(n-k)!}={}^{n}P_{k}$$

We assuming that:

1. only one object can be placed in each cell,
2. order is important,
3. the whole procedure is done without replacement.

Proof:

We have k cells and we have n objects.

Thus:

In the first cell we can place one out of n objects.

In the second cell we can place one out of $(n-1)$ objects.

$\vdots$

In the k^{th} cell we can place one out of $n-(k-1)$ objects.

Finally we can apply Theorem 2.1 to obtain the following answer:

$$n\times(n-1)\times(n-2)\times\ldots\times(n-[k-1])$$

This can be written as $\frac{n!}{(n-k)!}$ which is ${}^{n}P_{k}$.

■

2.4 Theorem (Arranging Like Objects)

The number of permutations (arrangements) of n objects of which n_1 objects are identical, n_2 objects are identical, ... , n_k objects are identical is given by:

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

2.5 Examples

2.5.1 There are n! ways of putting n individuals in order and (n-1)! ways of putting n persons by a round table, taking only relative positions into consideration.

2.5.2 There are ${}^{n}\mathrm{P}_{k}$ ways of choosing samples of size k from a population of size n without replacement.

2.5.3 There are ${}^{n}\mathrm{P}_3$ ways of selecting the list of the top three out of n contestants.

2.5.4 Out of 100 football players we can choose ${}^{100}\mathrm{C}_{11}$ if we do not insist on each player having a fixed position. If we do this then the possibilities increase to ${}^{100}\mathrm{P}_{11}$.

2.5.6 In how many ways can the letters in the word ROCOCO?

Answer: $\frac{6!}{3!2!} = 60$ ways

2.6 Theorem (Principle of Combinations)

The number of ways of choosing k objects out n different objects is given by the formula:

$$ {}^{n}C_{k} = \frac{n!}{k!(n-k)!} $$

Here we assume that
1. order of selection is irrelevant,
2. selection of objects is being done without replacement.

Proof

If order were important ${}^{n}P_{k}$ possibilities would exist. Given k objects, we can order them in $k!$ ways. Hence we have $k!$ copies of each specific selection. Thus we divide ${}^{n}P_{k}$ by $k!$ giving us the required result.

■

2.6.1 Note

There is an intimate link between the formula ${}^{n}C_{k}$ and the binomial expansion for the algebraic expansion $(x+y)^n$.

In fact we have that $(x+y)^n = {}^{n}C_0 x^n + {}^{n}C_1 x^{n-1}y^1 + \ldots + {}^{n}C_k x^{n-k}y^k + {}^{n}C_n y^n$

2.7 Examples

2.7.1 There are ${}^{11}C_4$ ways of choosing 4 players from a team of 11 players.

2.7.2 How can n identical objects be distributed among k different individuals?
The solution is simple let's assume that each object is represented by a * (star).
Next we align them in a row. We then draw $(k-1)$ vertical lines which can divide the stars into k groups. (Here each group represents an individual.)
Let's have a simple example:
Assume we are going to distribute $n = 20$ one euro coins (€ 1) among $k = 6$ children.
Then using bars and stars we represent the case when 4 coins are given to the first child, 2 to the second child, 6 to the third child, none to the fourth, 5 to the fifth and three to the sixth child as follows:

* * * * | * * | * * * * * * | | * * * * * | * * *

In this case we have a total of 25 objects in a row.
Thus we have 25! possible arrangements. However the stars are identical and so are the bars. Thus we should divide 25! by 20! and by 5! i.e.

$\frac{25!}{20!5!}$ which can be re-written as $\frac{(20+6-1)!}{(\{20+6-1\}-\{6-1\})(6-1)!}$

Going back to the general case of n objects and k individuals the answer is:

$\frac{(n+k-1)!}{n!(k-1)!}$ which can be re-written as ${}^{n+k-1}C_{k-1}$.

2.8 Principle of Equiprobability

The simplicity with which the principle of equiprobability assigns probabilities to events might make it most natural:

Having counted the total number N of possibilities one counts the number n of possibilities making up event A and then $\mathbb{P}[A] \equiv n / N$ ***because possibilities are equally likely.***

This principle became very popular during Laplace's time. It was "justified" in many ways. One argument was that, lacking more precise knowledge, it makes sense for us to put each outcome on the same par by assuming that outcomes are equally likely. This seems perfectly acceptable in some cases. There are a good number of important examples where the principle of equiprobability works perfectly well. But it does not work for all.
Later on we shall consider circumstances where this principle doesn't apply. This indicates that probability models should be defined carefully if they are to be valid.

2.8.1 Coin tossing and dice throwing example

These 2 classic situations lend themselves very well to the equiprobability principle. There are two outcomes in the case of the toss of a coin; so heads gets probability $\frac{1}{2}$ like tails. The throw of a dice offers 6 possible outcomes 1, 2, 3, 4, 5, and 6, each of which gets probability of 1/6. Out of these 2 assumptions springs a huge collection of problems.
For instance:

- probability of 2 heads in 2 tosses of a coin = $\frac{1}{4}$, or is it?
- probability of heads and a tail = $\frac{1}{2}$; shouldn't it be 1/3?
- probability of an even number in a throw of a dice =3/6.
- most probable sum of numbers turning up in 2 throws of a dice is 7;
  probability = 6/36

The cases above immediately show that we have to be careful which "outcomes" or "events" are to be considered likely. Different assumptions give different probabilities for the same event even under equiprobability.
In the case of biased coins or dice, we would need to drop the equality argument right from the start.

2.8.2 Random Samples without replacement

Consider a population of size N. A sample of size n is selected, each selection being effected without replacement. This gives rise to ${}^{N}P_{n}$ samples each with a probability $1/\ {}^{N}P_{n}$ of being the sample chosen (call this assumption A1). This is so if we take order into consideration.

Suppose we are not interested in the order in which the sample was selected. Then, using the principle of combinations and the principle of equiprobability we have a probability of $1/\ {}^{N}C_{n}$ for each sample (call this assumption A2).

Consider the following results:

- the probability that a given individual be chosen in a sample of size n from a population of size N is given by: $n^{N-1}P_{n-1}\ /\ {}^{N}P_{n} = n/N$ ..... using A1.

Using A2 we get ${}^{N-1}C_{n-1} \big/ {}^{N}C_{n} = \dfrac{n}{N}$

- the probability that 2 given individuals be chosen in a sample of size n from a population of size N is given by: $n(n-1)^{N-2}P_{n-2} \big/ {}^{N}P_{n} = \dfrac{n(n-1)}{N(N-1)}$ using A1.

Using A2 we get ${}^{N-2}C_{n-2} \big/ {}^{N}C_{n} = \dfrac{n(n-1)}{N(N-1)}$.

There are equivalent arguments for deriving the formulas above using tree diagrams.

2.8.3 Random Samples with replacement

Consider a population of size N. A sample of size n is selected, each selection being effected with replacement – giving rise to N^{n} samples each with a probability $1/N^{n}$ of being the sample chosen under assumption A1 that order is taken into consideration.
If we make Assumption A2 allowing us to ignore order, the counting becomes much more complex.
Under assumption A1, we have the results:
The probability that a given individual be chosen in a sample of size n from a population of size N is: $1-\left(\dfrac{N-1}{N}\right)^{n}$ .We build the argument using complements.

The probability that 2 given individuals be chosen in a sample of size n from a population of size N is given by:

Here we use 2 interesting techniques: $1-2\left(\dfrac{N-1}{N}\right)^{n}+\left(\dfrac{N-2}{N}\right)^{n}$.

In 1 we compute the probability that the given individual is not selected at all, and then subtract it from 1.
In 2 we compute the probability that each given individual is excluded subtract it from 1 and then add the probability that both individuals are excluded.

2.8.4 The Bigamy Problem

A group of n males has to be paired with a group of n females not necessarily one to one. Thus a female might have more than one male. What is the probability that each person has only one partner of the opposite sex?

This is an example of a seemingly difficult problem which is reducible to what we have done already. There are n^n ways of writing a sequence of n different numbers with repetitions and n! ways of writing a sequence with n numbers without repetitions. The required probability is:

$$\frac{n(n-1)(n-2)....1}{n^n} = \frac{n}{n}\frac{n-1}{n}...\frac{2}{n}\frac{1}{n}$$
$$= \left(1-\frac{0}{n}\right)\left(1-\frac{1}{n}\right)...\left(1-\frac{n-2}{n}\right)\left(1-\frac{n-1}{n}\right)$$

which is a really small number indeed: as n increases indefinitely to infinity the probability tends to 0.

2.8.5 The Birthday Problem

How many people do we need to have in a room to make it a favorable bet (probability of success greater than 1/2) that two people in the room will have the same birthday?

Since there are 365 possible birthdays, it is tempting to guess that we would need about 1/2 this number, or 183. You would surely win this bet!

In fact, the number required for a favorable bet is only 23. To show this, we find the probability p_r that, in a room with r people, there is no duplication of birthdays; we will have a favorable bet if this probability is less than one half.

Assume that there are 365 possible birthdays for each person (we ignore leap years). Order the people from 1 to r. There are 365 possibilities for the first element of the sequence, and for each of these choices there are 365 for the second, and so forth, making 365^{r} possible sequences of birthdays. We must find the number of these sequences that have no duplication of birthdays. For such a sequence, we can choose any of the 365 days for the first element, then any of the remaining 364 for the second, 363 for the third, and so forth, until we make r choices. For the rth choice, there will be 365 − r + 1 possibilities. Hence, the total number of sequences with no duplications is

$$365 \times 364 \times \ldots \times (365 - r + 1)$$

Thus if we assume that each sequence is equally likely we have that

$$p_r = \frac{365 \times 364 \times \ldots \times (365 - r + 1)}{365^r}$$

Next we evaluate the above for different values of r. This can easily be accomplished via a simple Matlab program. Some of the results can be found in the table below:

Number of People (r)	Prob. that all birthdays are different (p_r)
20	0.588561616419420
21	0.556311664834794
22	0.524304692337450
23	0.492702765676014
24	0.461655742085471
25	0.431300296030536

```
clc
clear all
r=25
Z=[];
for i=1:r
    z=1;
    for j=0:i-1
        z=z*(365-j)/365;
    end
    Z=[Z,z];
end
format long
Results=[[1:r]' Z']
```

As we asserted above, the probability for no duplication changes from greater than one half to less than one half as we move from 22 to 23 people.

2.10 The Maxwell-Bolzmann Statistic

n balls (particles) are to be distributed randomly amongst k cells. We are here assuming that we are distinguishing between particles in different cells i.e. we care who is in each cell, not just how many are in each cell. There are k^n configurations in all.

Thus the probability a particular configuration n_1 in cell 1, n_2 in cell 2, … , n_k in cell k is given by:

$$\frac{n!}{n_1!n_2!..n_k!}\frac{1}{k^n}$$

The probability that a particular cell has r particles is:

$$\frac{{}^nC_r(k-1)^{n-r}}{k^n} = {}^nC_r\frac{1}{k^r}\left(1-\frac{1}{k}\right)^{n-r}$$

This is called the Maxwell-Boltzmann distribution because the arguments above are equivalent to those used by Maxwell and Boltzmann in deriving the energy distribution of gases when considering each energy level as though it were a cell and atoms as distinguishable balls in the manner described above. One additional step that is usually done is to use Stirling's approximation $n! \approx \sqrt{2\pi n}\exp(-n)n^n$ to obtain limiting forms of the above expression as $n \to \infty$.

2.11 Bose-Einstein Statistics

In this context we assume that the n particles are completely indistinguishable. So that we consider only configurations with different allocations as being equiprobable. In this case for n particles in k cells we have ${}^{n+k-1}C_n = {}^{n+k-1}C_{k-1}$ configurations each of which has probability $1/{}^{n+k-1}C_n$.

This problem is equivalent to that of distributing k one euro coins amongst n persons, ignoring the coins' identity.

The derivation of this formula can be obtained by considering the problem of drawing a sequence of $n+k-1$ signs of which $k-1$ are cells and n are dots.

The probability that a particular cell has r particles is ${}^{n+k-r-2}C_{n-r} / {}^{n+k-1}C_n$.

2.12 The Fermi-Dirac Statistics

In this context we assume no cell can accommodate more than 1 particle. Usually $k > n$. Then the probability of a particular configuration is $1/{}^{k}C_n$. The probability that a particular cell is occupied is then ${}^{k-1}C_{n-1} / {}^{k}C_n = n/k$.

This problem is equivalent to that of assigning k jobs to n persons where one person can have at most one job.

The last two distributions serve as models for certain types of particle in atomic physics. They are refinements of the Maxwell-Boltzman distribution arguments and emphasize an important point about what outcomes to consider as equiprobable.

3. Geometric Probability

The combinatorial techniques above can be applied only in cases involving finitely many possibilities.
In cases where the set of possible outcomes, call it Ω, is infinite, we have to extend the idea. Equally probable does not make sense here. Also one cannot add the number of points on a curve or inside a closed curve.

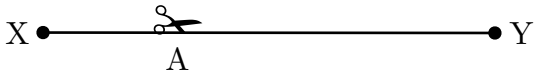
3.1 Infinite set of possible outcomes

To understand the concept of a set on infinite possible outcomes let's consider the following example:

A line XY of length l is cut at a point A selected randomly. What is the probability the left segment will be greater than the right?

The set Ω , contains all the possible points on the line XY . At once we notice that the number of points on the line XY is infinite. As a result we will have to measure sets as follows:

Length of line XY is l.



Let the set B contain all the points A on the line XY such that length $XA < l/2$.
Formally we write $B = \{A : XA < l/2\}$

Thus the required probability is $\dfrac{\text{length of set B}}{\text{lenght of line XY}} = \dfrac{l/2}{l} = 0.5$

As the above example suggests, in this chapter we shall have to measure sets of the type mentioned above differently:

(i) If we have one-dimensional curves, then we can use the concept of length.
(ii) If we have subsets of the xy-plane then the idea will be to use the concept of area (via integration).
(iii) If we have subsets in a 3-dimensional space then we can use the concept of volume (via double integration).
(iv) If we have subsets in an n-dimensional space then we can use the n-dimensional integrals.

The above idea suggests the following procedure:
Take some set Ω, representing all individual possibilities that can happen, and define on it some geometric measure, like length, area or volume.
Then for subsets of Ω whose measure exists we can establish an equivalence with events. The collection of all such subsets will form the σ-algebra F.

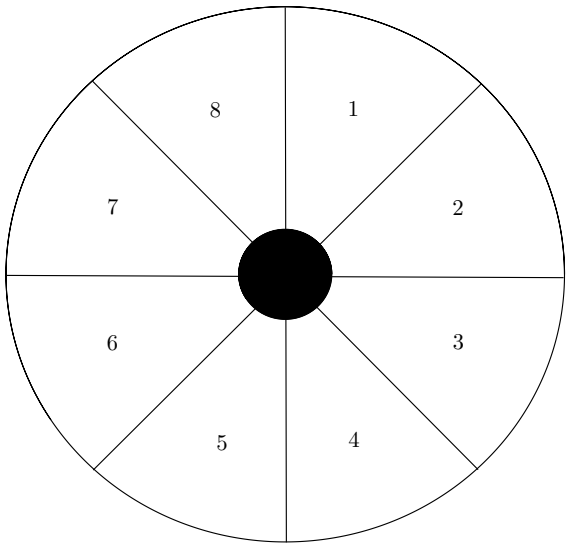
The probability of a given event is calculated first by selecting subset A corresponding to the event and then using the formula:

$$\mathbb{P}[A] = \frac{\text{measure of } A}{\text{measure of } \Omega}.$$

3.2 Examples

3.2.1 The Circular Target

The circular target shown here has radius R. Its bulls' eye (the dark circle in the middle) has an area of B units. Find the probability that an arrow hits sector 1, provided that the arrow hits the target.



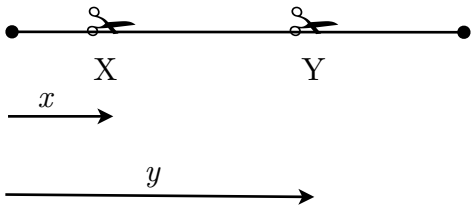
Solution

Let $\mathbf{\Omega}$ be the set which contains all the points which can be hit by the arrow. (In this case $\mathbf{\Omega}$ contains all the points in the target.)
Thus to find the required probability we find the area of sector 1 and divide it by the area of whole circle:

$$\mathbb{P}[\text{arrow hits sector 1}] = \frac{\frac{1}{8}\pi R^2 - \frac{1}{8}B}{\pi R^2} = \frac{\pi R^2 - B}{8\pi R^2}$$

3.2.2 The Rod of Length l

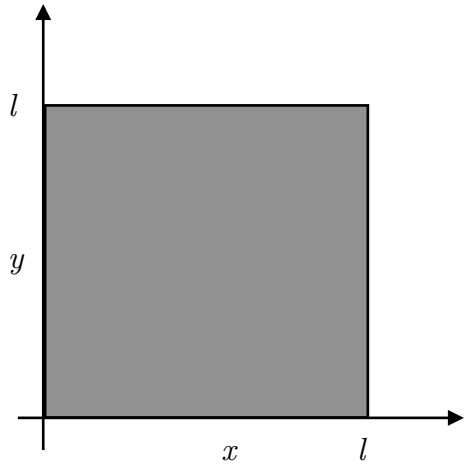
A rod of of length l is spilt randomly into 3 parts by making two cuts at random distances x and y from one end:



In this case the set which contains all the possible outcomes (the probability space) $\mathbf{\Omega}$ can be represented by the set :

$$\{(x,y) : 0 \le x \le l \text{ and } 0 \le y \le l\}$$

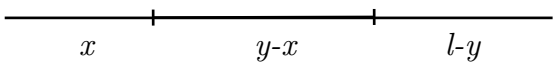
Each point (x,y) in the above set corresponds to exactly one possible outcome. In this case it is possible to draw a diagram of the probability space. In fact $\mathbf{\Omega}$ can be represented by a square of length l as shown below:



We would like to calculate the probability that a triangle can be formed by the resulting 3 lines.
A triangle can be formed only if the sum of the lengths of two parts of the line is greater than length of the remaining part. Here we have two scenarios:

Case 1

If $x < y$, then the line is divided as follows:



To form a triangle then:

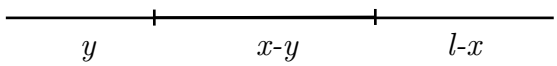
$$x + (l - y) > y - x$$

which simplifies to:

$$y - x < \frac{l}{2}$$

Case 2

If $y < x$, then the line is divided as follows:



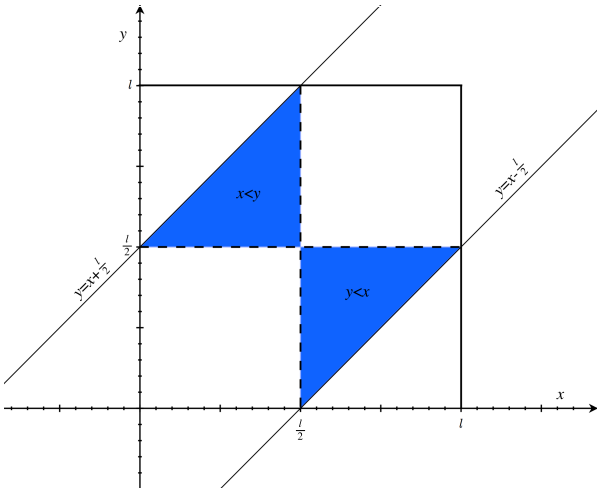
To form a triangle then:

$$y + (l - x) > x - y$$

which simplifies to:

$$x - y < \frac{l}{2}$$

If we draw the above two inequalities in the above diagram then the shaded region shown below represents all the points (x,y) which if chosen will allow a triangle to be formed.

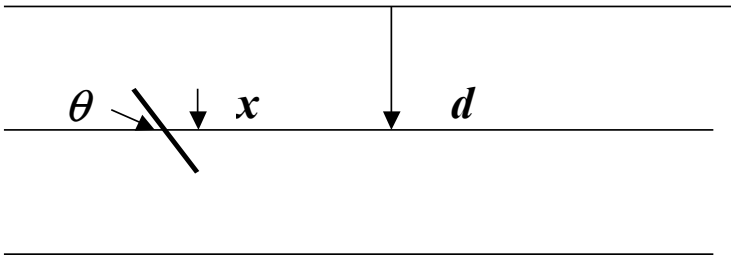


As a result we must now find the area of the shaded region and divide it by the area of the whole square.

$$\mathbb{P}[\text{triangle is formed}] = \frac{l^2 / 4}{l^2}$$
$$= \frac{1}{4}$$

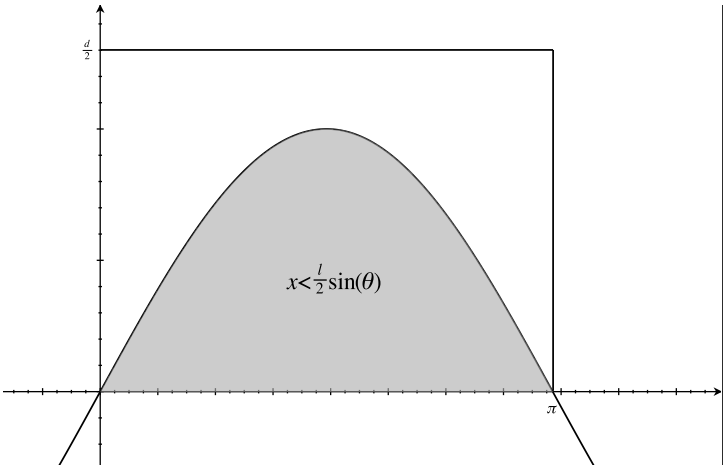
3.2.3 Buffon's Needle

A needle of length l is thrown randomly on a plane which is ruled by infinite parallel lines equally separated by a distance d, where $l < d$. We wish to find the probability that the needle intersects one line.



The set of all possibilities Ω can be "reduced" through equivalence of sets to the rectangle $\{(x,\theta): 0 \le x \le \frac{d}{2} \text{ and } 0 \le \theta \le \pi\}$ where x gives us the distance of the needle's mid-point from the nearest line and θ is the angle subtended by the needle to this line.

Consider the event that the needle intersects one line. This subset of possibilities can be shown geometrically to be the set of all points: $\{(x,\theta): 0 \le x \le \frac{l}{2}\sin(\theta)\}$:



The probability of this set is calculated by integrating the area under the curve $x = l\sin(\theta)$ and divide this result by the area of the set which contains all the points:

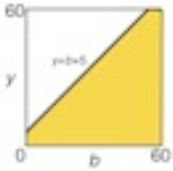
$$\mathbb{P}[\text{needle intersects one line}] = \frac{\int_0^{\pi} l\sin(\theta)\,d\theta}{d\pi} = \frac{2l}{d\pi}$$

3.3 The Bus Problem

Every day John either goes to work by bus or on foot. The bus does not have a fixed schedule and arrived at the bus stop at different times but always between 6am and 7am. John finds it hard to wake up in the morning and as such he too arrives at the bus stop at different times between 6am and 7am. The bus stop waits for 5 minutes at the stop before leaving. Furthermore, John waits at the stop for 20 minutes, afterwards he decides to walk to work it the bus doesn't turn up during this time. Find the probability that John catches the bus.

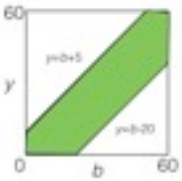
We have two continuous variables here: b the time in minutes past 6am that the bus arrives, and y the time in minutes past 6am that John arrives.
Since there are 2 independent variables, we will convert this into a 2-dimensional geometry problem. Specifically, we can think of the set of all outcomes Ω as the points in a square:

Next, we need to determine the region of "success"; that is, the points where one can catch the bus. Since the bus will wait for 5 minutes, John needs to arrive within 5 minutes of the bus' arrival, i.e. $y \leq b+5$. This translates to the following shaded region:



However, John only waits for 20 minutes, so he can't arrive more than 20 minutes before the bus, i.e. $y \geq b-20$. This translates to the following shaded region:

Combining our two conditions, we have a region of success as shown below:



Next we find the area of shaded region:

$$\mathbb{P}\left(\text{John catches the bus}\right) = \frac{\text{Area of shaded region}}{\text{Area of } \Omega} = \frac{60^2 - \frac{55^2}{2} - \frac{40^2}{2}}{60^2} = \frac{103}{288}$$

4. Probability Spaces

The need for a more precise mathematical formulation of the way in which probabilities are defined and manipulated was felt in the course of last century. The Russian A.N. Kolmogorov is credited with the modern mathematical foundations of probability

4.1 Axioms Defining Probability Spaces

Three major concepts characterize probabilistic situations at the level of simple probabilities:

1. All the possible single and individual outcomes which combinatorial principles allowed us to count and geometry to represent.
2. Certain collections of possibilities, grouped together, which we call "events" and for which we have rules of formation.
3. Probabilities of events for which we have rules of computation.

4.1.1 Sigma Algebras

We need to take the concept of events a bit deeper.
Consider:

1. When we talk about an odd number turning up in dice throwing,
2. When we talk about sequences starting with two heads in 5 successive coin tosses,
3. When we talk about some point in the lower half of a target being hit.

We are forming subsets of possibilities.
Clearly for each event we need to be able to deal with the eventuality of its nonoccurrence. In other words we need the complement of each event.

We need to be able to "join" events - if we have event A and event B we need to be able to talk about either some possibility in A or some possibility in B occurring.
All these ideas are captured succinctly in the definition of a σ-algebra.

4.1.1.1 Definition

A collection $\mathcal{F}$ of subsets of a set Ω is said to be a sigma-algebra (or σ-algebra) if it satisfies the following three conditions:

1. $\Omega \in \mathcal{F}$,
2. If $A \in \mathcal{F}$, then $A^C \in \mathcal{F}$, here A^C denotes the complement of the set A,
3. If $\{A_n\}_{n\in\mathbb{N}}$ is a sequence of elements of $\mathcal{F}$, then $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$.

Two extreme examples of σ-algebras are:

1. the power set $\wp(\Omega)$ the collection of all subsets of Ω,
2. the trivial σ-algebra $\{ \varnothing , \Omega \}$.

Usually σ-algebras are somewhere between these two extremities.
If we take $\Omega = \{1, 2, 3, 4\}$ then the following collection of subsets:

$$\{ \Omega, \varnothing, \{1\}, \{2\}, \{3, 4\}, \{1, 2\}, \{2, 3, 4\}, \{1, 3, 4\}\}$$

is indeed a σ-algebra, but it is not the power set.

In many cases specifying the relevant σ-algebra is avoided without diminishing the overall grasp excessively.

4.1.2 Probability Measures

There are a number of rules which we intuitively, and implicitly, assume to hold when computing probabilities. In effect probabilities are usually defined on subsets and hence can be looked at as functions on σ-algebra $\mathcal{F}$. They tell us how probable some events are relative to others. Probability assignments clearly have to satisfy some to qualify as probability measures.

Definition
The function $\mathbb{P} : \mathcal{F} \to [0,1]$ is said to be a probability measure if the following two conditions are satisfied:

1. $\mathbb{P}[\Omega] = 1$,
2. If $\{A_n\}_{n\in\mathbb{N}}$ is a DISJOINT sequence of sets in $\mathcal{F}$, then $\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}[A_n]$.

Stipulating probabilities to be positive, to add up to one are all intuitively clear. In addition we also expect them to add up over sets which have no “overlap” , that is which are disjoint. Let us give properties of probability measures which follow from the axioms above.

4.1.3 Theorem

Let $\mathbb{P}$ be a probability measure on $\mathcal{F}$, then:

1. $\mathbb{P}[A] = 1 - \mathbb{P}[A^C]$ for all $A \in \mathcal{F}$,
2. If $B \subset A$ then $\mathbb{P}[A] = \mathbb{P}[B] + \mathbb{P}[A \setminus B]$,
3. $\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B]$,
4. $\mathbb{P}[A \cup B \cup C] = \mathbb{P}[A] + \mathbb{P}[B] + \mathbb{P}[C] - \mathbb{P}[A \cap B] - \mathbb{P}[A \cap C] - \mathbb{P}[B \cap C] + \mathbb{P}[A \cap B \cap C]$.

Proof:

1. and 2. are trivially true.

To prove 3, we shall use the fact that $A \cup B$ can be written as the union of three disjoint sets as follows : $A \cup B = (A \setminus A \cap B) \cup (B \setminus A \cap B) \cup (A \cap B)$.

Hence, $\mathbb{P}[A \cup B] = \mathbb{P}[A \setminus A \cap B] + \mathbb{P}[B \setminus A \cap B] + \mathbb{P}[A \cap B]$.

Furthermore we have that $\mathbb{P}[A \setminus A \cap B] = \mathbb{P}[A] - \mathbb{P}[A \cap B]$,
$\mathbb{P}[B \setminus A \cap B] = \mathbb{P}[B] - \mathbb{P}[A \cap B]$.

Hence, $\mathbb{P}[A \cup B] = \mathbb{P}[A \setminus A \cap B] + \mathbb{P}[B \setminus A \cap B] + \mathbb{P}[A \cap B]$
$= \mathbb{P}[A] - \mathbb{P}[A \cap B] + \mathbb{P}[B] - \mathbb{P}[A \cap B] + \mathbb{P}[A \cap B]$
$= \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B]$

Proof of 4, as before we shall express $A \cup B \cup C$ as the sum of disjoint sets:

We start by writing $A \cup B \cup C$ as follows:

$$A \cup B \cup C = (A \cup B) \setminus ((A \cup B) \cap C) \cup C$$

Furthermore we have that $(A \cup B) \cap C = [((A \cap C) \cup (B \cap C)) \setminus (A \cap B \cap C)] \cup (A \cap B \cap C)$.

Thus: $\mathbb{P}[(A \cup B) \cap C] = \mathbb{P}[A \cap C] + \mathbb{P}[B \cap C] - \mathbb{P}[A \cap B \cap C]$ from which we can deduce that:

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

■

4.1.4 The Probability Space

A probability space is a triple $(\Omega, \mathcal{F}, \mathbb{P})$ where

- Ω is a set which contains all the possible outcomes,
- $\mathcal{F}$ is the sigma algebra which contains subsets of Ω ,
- $\mathbb{P}$ is the probability measure on $\mathcal{F}$.

The use of an abstract definition like this is twofold:

- It gives us the basic terminology and essential criteria for allowing us to decide whether a probabilistic model is suitable to model a particular situation.
- By simplifying the context to clear mathematical concepts, it allows us to derive a number of mathematical results which are true in virtue of the basic mathematical criteria underpinning the model.

One should mention that the above formulation is not the only definition available. A good number of other definitions, notably by von Mises, Savage, and de Finetti have been studied in depth. The Kolmogorov definitions are the ones which have gained widest acceptance, and seem to outperform the others in most applications.

4.2 Modelling Of Simple Situations

4.2.1 Coin Toss

Consider a coin being tossed twice.

$\Omega = \{(H,H),(H,T),(T,H),(T,T)\}$

$\mathcal{F} = \{(H,H),(H,T),(T,H),(T,T),$
$((H,H),(H,T)),((H,H),(T,H)),((H,H),(T,T)),$
$((H,T),(T,H)),((H,T),(T,T)),((T,H),(T,T)),$
$((H,H),(H,T),(T,H)),((H,H),(H,T),(T,T)),((H,T),(T,H),(T,T)),$
$\Omega,\phi\}$

$\mathbb{P}[(x,y)] = \frac{1}{4}$, $\forall x,y \in \{H,T\}$

4.2.2 A dice Being Thrown Once

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

$\mathcal{F} =$ set which contains all the subsets of Ω , 2^6 in all.

$\mathbb{P}[\{i\}] = \frac{1}{6}$, $\forall i \in \{1,2,3,4,5,6\}$

4.2.3 Engine with Three Cylinders

Consider a car engine with three cylinders each with probability p of developing a fault.

$\Omega = \{(0,0,0),(1,0,0),(0,1,0),(0,0,1),(1,1,0),(1,0,1),(0,1,1),(1,1,1)\}$

where $1 \rightarrow$ fault , $0 \rightarrow$ no fault .

(The above set can also be written as follows: $\Omega = \{(x_1,x_2,x_3),\text{where},x_i \in \{0,1\},\forall i \in \{1,2,3\}$)

$\mathcal{F} =$ set which contains all the subsets of Ω , 2^8 in all.

$\mathbb{P}[(x_1,x_2,x_3)] = {}^3C_x p^x (1-p)^{3-x}$

4.2.4 Molecule Experiment

Consider an imaginary experiment where a molecule made up of one carbon atom, 3 hydrogen atoms and a hydroxyl ion (methyl alcohol CH_3OH if our chemistry serves us well) is bombarded with radiation sufficient to knock off not more than just one particle. Experimentation can only reveal whether it is a H atom ,C atom or OH ion which detaches with probabilities 0.4, 0.3 , 0.1 respectively.

$\Omega = \{0,1,2,3,4,5\}$, where $0 \rightarrow$ none, $1 \rightarrow$ first hydrogen atom, $2 \rightarrow$ second hydrogen atom, $3 \rightarrow$ third hydrogen atom, $4 \rightarrow$ carbon atom and $5 \rightarrow$ hydroxyl ion.

$\mathcal{F}$ = all subsets of Ω excluding those which distinguish between 1,2 and 3. In all there are 2^4 sets.

$\mathbb{P}[\{1,2,3\}] = 0.4$, $\mathbb{P}[\{4\}] = 0.3$, $\mathbb{P}[\{5\}] = 0.1$

$\mathbb{P}[\{0\}] = 0.2$

4.2.5 The Island

Consider an imaginary island shaped like an ellipse. Missiles are fired randomly at the island with a probability 0.2 of missing. We set:

$\Omega = \{\infty\} \cup \{(x,y) : ax^2 + by^2 \leq c\}$

$\mathcal{F}$ = all subsets of Ω whose area we can compute.

$\mathbb{P}[\{\text{miss}\}] = 0.2$

$\mathbb{P}[A] = 0.8 \times \text{Area of A / Area of ellipse}$

4.2.6 The Electronic Component

An electronic component is composed of 3 integrated circuits which are encased in 2 sealed compartments. There are 2 circuits in the first compartment and 1 in the second. If an IC malfunctions, the compartment containing it shuts itself off and the whole component fails to operate any further. While it is possible to know which of the compartments are shut off, it is not possible to identify which particular circuit is responsible. Tests on the IC's reveal that the probability of malfunctioning for each individual IC is 0.02, and we can assume that IC operate independently of one another.

$$\Omega = \{(x_1, x_2, x_3) : x_i = 0 \text{ or } 1 \text{ for } i = 1,2,3\}$$

Note: x_1 and x_2 refer to the ICs contained within first compartment while x_3 refers to the second compartment. Furthermore, 0 means failure, and 1 means IC is OK.

In this example, $\mathcal{F}$ cannot contain all the subsets of Ω . Since both compartments are sealed, and compartment one contains IC 1 and IC 2, then if either of these shuts down the entire component will stop working, however we cannot tell whether it is IC 1 or IC 2. In fact we can only tell whether the fault occurred in compartment 1 or compartment 2. Hence the way the sigma algebra is constructed must reflect the information given in this scenario.

As a result, these three possible events are indistinguishable:

$$(0,1,1),\ (1,0,1) \text{ and } (0,0,1)$$

This is because the fault always occurred within compartment 1 and we cannot tell which of the two ICs has failed.

The next we consider the eventuality that there is a fault in both compartments. In this case the following sets are indistinguishable:

$$(0,1,0),\ (1,0,0) \text{ and } (0,0,0)$$

In fact we are capable of noticing that both compartments have shut down, however we cannot tell which one of the two ICs contained within compartment 1 has failed.

There remains to consider only two cases:
(1,1,0) represents the eventuality that the fault developed in compartment 2
and

(1,1,1) no fault is recorded.

Thus $\mathcal{F} = \{\{(0,1,1),(1,0,1),(0,0,1)\},\{(1,1,0)\},\{(1,1,1)\},\{(0,1,0),(1,0,0),(0,0,0)\},$
all possible unions of these sets, all complements,
$\Omega, \phi\}$

$\mathbb{P}[\{(0,1,1),(1,0,1),(0,0,1)\}] = 0.98 \times 0.02 \times 2 + 0.98 \times 0.98$

$\mathbb{P}[\{(1,1,0)\}] = 0.98^2 \times 0.02$, $\mathbb{P}[\{(1,1,1)\}] = 0.98^3$

$\mathbb{P}[\{(0,1,0),(1,0,0),(0,0,0)\}] = 2 \times 0.02^2 \times 0.98 + 0.02^3$

4.2.7 A Coin Tossed Three Times.

A fair coin is tossed 3 times. We know only the tosses up to the second toss but not the third.

$\Omega = \{(x_1, x_2, x_3) : x_i = H \text{ or } T \text{ for } i = 1,2,3\}$

In this case:

$\mathcal{F} = \big\{\{(H,H,H),(H,H,T)\}, \{(H,T,H),(H,T,H)\},\{(T,H,H),(T,H,T)\},\{(T,T,H),(T,T,T)\}$
all possible unions of these sets, all complements,
$\Omega, \phi\}$

The probability of any of these sets is clearly equal to $\frac{1}{4}$.

4.2.8 Infinite Coin Toss

A fair coin is tossed infinitely often.

$$\Omega = \{(x_1, x_2, \ldots, x_n, \ldots) : x_i \in 0,1\}$$

$1 \rightarrow$ Heads, and $0 \rightarrow$ tails.

Ω may be visualized partly through the map φ from the set of all binary sequences to $[0,1]$ as follows:

$$\varphi((x_1, \ldots, x_n, \ldots)) = \sum_{i=1}^{\infty} \frac{x_i}{2}$$

This map however is not 1-to-1, nevertheless we can consider $[0,1]$ as a candidate for a probability space.

For $\mathcal{F}$ can we take just intervals from $[0,1]$?

Finding the probability of such sets is easy:
Given any interval $(a,b)\subseteq[0,1]$, we have that $\mathbb{P}[(a,b)]=b-a$, furthermore we have that $\mathbb{P}$ is positive and also $\mathbb{P}[(0,1)]=1$.

Is this definition sufficient?
What about complements?
What about unions?
What about countable unions?
How about taking the power set of [0,1] as its σ-algebra?

4.3 Multiple Events Probabilities

We end this chapter by considering an interesting and useful result about combining probabilities involving multiple events. We use this result in a number of important applications. We first give a generalization of the probability of the union of events formula:

4.3.1 Theorem

Given a sequence of events A_1,A_2, ..., A_n, we have that:

$$\mathbb{P}[A_1\cup A_2\ldots\cup A_n]=\sum_{i=1}^{n}\mathbb{P}[A_i]-\sum_{i_1=1}^{n}\sum_{i_2>i_1}^{n}\mathbb{P}[A_{i_1}\cap A_{i_2}]+\sum_{i_1=1}^{n}\sum_{i_2>i_1}^{n}\sum_{i_3>i_2}^{n}\mathbb{P}[A_{i_1}\cap A_{i_2}\cap A_{i_3}]-\ldots$$
$$\cdots+(-1)^{n+1}\sum_{i_1=1}^{n}\sum_{i_2>i_1}^{n}\cdots\sum_{i_n>i_{n-1}}^{n}\mathbb{P}[A_{i_1}\cap A_{i_2}\cap\cdots\cap A_{i_n}]$$

Proof

We shall only give a sketch for a proof. We need to make sure that when we add the probabilities of events we do not add any subsets more than once and we do not subtract it either.
Given sequence of events $A_1,\ldots,A_n$, there are points which are included simultaneously in events A_i and A_j say. Adding $\mathbb{P}[A_i]$ with $\mathbb{P}[A_j]$ might result in adding the intersection probability $\mathbb{P}[A_i\cap A_j]$ twice.

In such cases we need to subtract the parts which are included twice.
Having subtracted the probability of all possible intersections, we have to remember that if we take any 3 events, say A_i,A_j and A_k all points which they have in common will be present in one of the intersection of $A_i\cap A_j$, $A_k\cap A_i$, and $A_j\cap A_k$. We shall thus have subtracted the probability of the event $A_i\cap A_j\cap A_k$ once "too much". So we have to add it.

This argument, called inclusion-exclusion, can be repeated up to n events to obtain the general result stated in the theorem.

■

4.4 Examples

4.4.1 The Matching problem

n objects are associated uniquely with elements from another set of n objects. This could be a set of n couples, a set of n persons and their socks, two sets of cards. One of the two sets is shuffled and the two sets associated anew at random. A match occurs when a pair which existed in the original association is formed again. What is probability of having exactly at least match?

We shall treat the first set as though it were numbered and hence ordered from 1 to n. The second set will then be shuffled randomly. There are $n!$ ways of permuting the second set and associating each member with another member from the first set.
Let A_i be the event that we have a match at the i'th place. Its probability is given by:

$$\mathbb{P}[A_i] = \frac{(n-1)!}{n!} = \frac{1}{n}$$

Furthermore we have that

$$\mathbb{P}[A_i \cap A_j] = \frac{(n-2)!}{n!} = \frac{1}{n(n-1)} \qquad \mathbb{P}[A_i \cap A_j \cap A_k] = \frac{1}{n(n-1)(n-2)}$$

and

$$\mathbb{P}[A_i \cap A_j \cdots \cap A_k] = \frac{1}{n!}$$

Now there are n A_i's , ${}^nC_2$ ways of choosing 2 A_i's, ... , ${}^nC_k$ ways of choosing k A_i's . Using the result from the theorem 4.3.1, we have that:

$$\begin{aligned}\mathbb{P}[A_1 \cup A_2 \cdots \cup A_n] &= \sum_{i=1}^{n} \frac{1}{n} - \sum_{i_1=1}^{n}\sum_{i_2>i_1}^{n} \frac{1}{n(n-1)} + \ldots \\ &\quad + \sum_{i_1=1}^{n}\sum_{i_2>i_1}^{n}\sum_{i_3>i_2}^{n} \frac{1}{n(n-1)(n-2)} - \cdots + (-1)^{n+1} \sum_{i_1=1}^{n}\sum_{i_2>i_1}^{n} \cdots \sum_{i_n>i_{n-1}}^{n} \frac{1}{n!} \\ &= 1 - \frac{1}{2!} + \frac{1}{3!} - \ldots\ldots + (-1)^{n+1}\frac{1}{n!}\end{aligned}$$

As $n \to \infty$, the expression above tends to 1-e^{-1} .

4.4.2 The Occupancy Problem

k balls are distributed amongst n cells, each ball being equally likely to be housed in any of the available cells. We would like to compute the probability that there are m cells empty.

We assume the cells are numbered 1 to n. There are n^k possibilities in all, each being equally likely. Let A_i correspond to the event that the i'th cell is empty.

We start by working out the probability that no cell is empty:

$$1 - \mathbb{P}[A_1 \cup A_2 \cup \cdots \cup A_n]$$

Now we have that:

$$\mathbb{P}[A_i] = \frac{(n-1)^k}{n^k} = \left(1 - \frac{1}{n}\right)^k \qquad \mathbb{P}[A_i \cap A_j] = \left(1 - \frac{2}{n}\right)^k$$

$$\mathbb{P}[A_1 \cap A_2 \cap ... \cap A_n] = 0$$

The probability that at least one cell is empty is then given by theorem 4.3.1:

$$\begin{aligned}\mathbb{P}[A_1 \cup A_2 ... \cup A_n] &= \sum_{i=1}^{n}\left(1-\frac{1}{n}\right)^k - \sum_{i_1=1}^{n}\sum_{i_2>i_1}^{n}\left(1-\frac{2}{n}\right)^k + \dots \\ &\quad + \sum_{i_1=1}^{n}\sum_{i_2>i_1}^{n}\sum_{i_3>i_2}^{n}\left(1-\frac{3}{n}\right)^k - \dots + (-1)^{n+1}\sum_{i_1=1}^{n}\sum_{i_2>i_1}^{n}\cdots\sum_{i_n>i_{n-1}}^{n}\left(1-\frac{n}{n}\right)^k\end{aligned}$$

Hence probability that no cell is empty is given by :

$$\begin{aligned}1 - \mathbb{P}[A_1 \cup A_2 \cup \dots \cup A_n] &= 1 - \left[{}^nC_1\left(1-\frac{1}{n}\right)^k - {}^nC_2\left(1-\frac{2}{n}\right)^k + \dots + (-1)^{i\,n}C_i\left(1-\frac{i}{n}\right)^k + \dots\right] \\ &= 1 - {}^nC_1\left(1-\frac{1}{n}\right)^k + {}^nC_2\left(1-\frac{2}{n}\right)^k - \dots - (-1)^{i\,n}C_i\left(1-\frac{i}{n}\right)^k + \dots\end{aligned}$$

and the number of configurations without any empty cells is given by:

$$\left\{1 - {}^nC_1\left(1-\frac{1}{n}\right)^k + {}^nC_2\left(1-\frac{2}{n}\right)^k - \dots - (-1)^{i\,n}C_i\left(1-\frac{i}{n}\right)^k + \dots\right\} n^k$$

Next we compute the probability that exactly m cells remain empty.

This is equivalent to choosing m cells out of n to keep them empty and then putting all the k balls in the remaining n - m cells. None of these n - m cells must be empty. The number of configurations for m-n cells without any empty cells is given by:

$$\left\{1 - {}^{n-m}C_1\left(1-\frac{1}{n-m}\right)^k + \dots - (-1)^{i\ (n-m)}C_i\left(1-\frac{i}{n-m}\right)^k + \dots\right\}(n-m)^k$$

Assembling all the pieces, we find that the probability of exactly m empty cells is:

$${}^nC_m\left\{1 - {}^{n-m}C_1\left(1-\frac{1}{n-m}\right)^k + \dots - (-1)^{i\ (n-m)}C_i\left(1-\frac{i}{n-m}\right)^k + \dots\right\}\left(1-\frac{m}{n}\right)^k$$

At this stage it is clear that to handle all these combinatorial expressions we need some easier way to compute them rather than by using factorials. Factorials are huge monsters to compute!

5. Conditional Probability

The notion of conditional probability was historically introduced as a stratagem for simplifying the computation of probabilities by considering the event under consideration together with other information. Abstracting into a rigorous definition isn't difficult in the language and notation of set theory.

5.1 Defining Conditional Probability

Before we give a formal definition, let's consider the following two examples:

5.1.1 Example

The four numbers {13, 33, 19, 111} are scribbled on four identically shaped paper. A blindfolded person select two papers one after the other.

Assuming the first selected paper is replaced, the probability of 19 showing up on the 2nd draw is $\frac{1}{4}$.
Assuming the first selected paper is not replaced , the probability of 19 showing up on the 2^{nd} draw is still $\frac{1}{4}$.
This seems odd because we feel it should be 1/3 in the second case. However a moment's thought will reveal that if we know what the first draw is, then the value of $\frac{1}{4}$ for the probability will change. If 19 showed up on the first draw, then its probability of showing up again is 0, otherwise it is 1/3!

5.1.2 Example

Each individual in a population was contacted and the past family history with reference to diabetes checked. The following table was compiled:

	Females	**Males**	**Total**
Positive History	623	551	1174
Negative History	1211	1088	2299
Total	1834	1639	3473

Common sense will help us appreciate the validity of the following two arguments immediately:
Given that a person, picked randomly from this population, has a diabetes history, the probability that the person is female is: (623/3473)/(1174/3473)= 623/1174.
Given that a person, picked randomly from this population, is male, the probability that his family has no history of diabetes is: (1088/3473) / (1639/3473) = 1088/1639.
In both cases above, in a natural way we have used the idea of conditional probability which as can be seen usually involves two events A and B, one of which is known to have

happened. We are interested then in seeing how the probability of the other event is modified.

5.1.3 Definition
Given events A and B, with $\mathbb{P}[B]>0$ the conditional probability of A given B is given by:

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}$$

Here we insist that $\mathbb{P}[B]>0$ because we want to divide with it in the above equation. When it is zero, we need a different definition which will be encountered much later on in more advanced work in statistics and probability.
We have already mentioned the fact that events of probability 0 are not "impossible" sets. They are set which when compared with their complements are "infinitely" less likely to happen. This "anomaly" has to be understood solely as a mathematical problem. The numbers in the interval [0 , 1] are in some cases not enough to paste over all subsets of some probability space. The smaller ones would have to be assigned probability 0, even if their occurrence is by no means impossible.

5.2 Independence

We consider 2 ideas which already came up earlier:

1. In a number of cases we multiplied probabilities of 2 events when we wanted the probability of their intersection. Why? One could argue that the probabilities of two successive choices from the same population without replacement are completely unaffected by one another.

2. Addition of probabilities moves parallel to union of event with the disjoint proviso. What about multiplication? Intersection would be its natural counterpart in terms of set operations.

These considerations justify the formal definition

5.2.1 Definition
Two events A and B are said to be independent if and only if:

$$\mathbb{P}[A \cap B] = \mathbb{P}[A] \times \mathbb{P}[B]$$

or equivalently, independence is captured by

$$\mathbb{P}[A|B] = \mathbb{P}[A]$$

Note that independence of 2 events A and B in the probabilistic sense says that knowledge of the occurrence of A leaves our knowledge, or lack of it, concerning the other event unchanged.

5.3 Theorem of Total Probability

Many times calculating the probability of an event becomes much easier in the knowledge of another event, and conditional probability helps us formalize this technique. An immediate extension of this idea for simplifying calculation of probabilities would consist of partitioning the whole probability space and then conditioning a given event on each component of the partition. An implementation of this idea is made use of when tree diagrams are employed in computing probabilities.

5.3.1 Theorem (Law of Total Probability)

Given event B and sequence of events $A_1, A_2, \ldots, A_n, \ldots$ which partition Ω in such a way that $\Omega = \bigcup_{i=1}^{\infty} A_i$, with the $A_i's$ being disjoint with $\mathbb{P}[A_i] > 0 \ \ \forall i$, we have that:

$$\mathbb{P}[B] = \sum_{i=1}^{\infty} \mathbb{P}[B|A_i]\mathbb{P}[A_i]$$

Proof

From the given information we can express the set B as follows:

$$B = \bigcup_{i=1}^{\infty} B \cap A_i$$

Hence, it follows that:

$$\begin{aligned} \mathbb{P}[B] &= \sum_{i=1}^{\infty} \mathbb{P}[B \cap A_i] \\ &= \sum_{i=1}^{\infty} \mathbb{P}[B|A_i]\mathbb{P}[A_i] \end{aligned}$$

Here we are using the definition of conditional probability and the concept of disjoint sets. ■

5.4 Switching the Order of Events in Conditional Probability

One idea, suggested immediately after looking at the formula for conditional probability, is whether it is possible to exchange the role of the consequent and conditioning events. The following equation, derived first by the English pastor Thomas Bayes (1701 - 1761), shows how easy it is

$$\mathbb{P}[A|B]\mathbb{P}[B] = \mathbb{P}[A \cap B] = \mathbb{P}[B|A]\mathbb{P}[A]$$

In this section we shall generalize this result.

5.4.1 Bayes' Theorem

Given event B, with $\mathbb{P}[B]$ and a sequence of events $A_1, A_2, \ldots, A_n, \ldots$ which partition $\boldsymbol{\Omega}$ in such a way that

$\boldsymbol{\Omega} = \bigcup_{i=1}^{\infty} A_i$, with the $A_i's$ being disjoint with $\mathbb{P}[A_i] > 0 \quad \forall i$, we have that:

$$\mathbb{P}[A_n|B] = \frac{\mathbb{P}[B|A_n]\mathbb{P}[A_n]}{\sum_{i=1}^{\infty} \mathbb{P}[B|A_i]\mathbb{P}[A_i]}$$

Proof

To prove the result we use the formula of total probability and the definition of conditional probability:

$$\mathbb{P}[A_n|B] = \frac{\mathbb{P}[B \cap A_n]}{\mathbb{P}[B]} = \frac{\mathbb{P}[B|A_n]\mathbb{P}[A_n]}{\sum_{i=1}^{\infty} \mathbb{P}[B|A_i]\mathbb{P}[A_i]}$$

■

This theorem has consequences belying its theoretical simplicity. In probability and statistical theory there is an important line of development linked very intimately to subjective probability. It "relates" probabilistically an event to a sequence of possible "causes". Bayes' result allows you to change the role of events. This idea has been strongly opposed by a school of statisticians and probabilists who say that employing it in practice is tantamount to swapping the order of cause and effect, a grave error which is philosophically untenable. However one has to emphasize that the mathematical derivation of the result was never under criticism. It is the interpretation which is under attack.

In spite of all the controversy, Bayes' Theorem is a much used result in theory, research and practice. It enables researchers to work out probabilities which cannot be worked out directly. Examples of its use abound in medical practice and research and in production engineering. Furthermore, a large corpus of results has been obtained and assembled into what is known as Bayesian statistics. In recent years there have been a number of great and important successes registered by Bayesian methods in statistics.

To understand better this development, one must appreciate that subjectivism in probability treats probabilistic statements as statements about our "ignorance".

5.4.2 Example

How would one compute the probability that a person will die of cancer (event B) given that the person smokes (event A)?

Surely you cannot take all smokers and find how many will die of cancer. But it is easy to find an estimate of the probability that a dead smoker actually died of smoking. Check through the records of N dead smokers and established how many died of cancer (n). The probability of smoking and that of dying are easy to calculate and then we use:

$$\mathbb{P}[B|A] = \frac{\mathbb{P}[A|B]\mathbb{P}[B]}{\mathbb{P}[A]}$$

5.4.3 Example

Gaklamus Co. Ltd. manufacture acetylene cylinders in 3 different plants: in Bulebel, in Xewkija and in San Gwann. Tests to check for defective outlet valves, and records of volume of sales, from the 3 plants gave the following table:

Plant	**Probability of a faulty valve**	**Percentage Sales**
Bulebel	0.01	20%
Xewkija	0.02	50%
San Ġwann	0.03	30%

A customer complains about a faulty valve in a cylinder. What is the probability that the cylinder in question originated from the Xewkija plant?

Let A represent the event "cylinder is from Xewkija",
Let B represent the event "cylinder is faulty".

Then

$$\begin{aligned}
\mathbb{P}[A|B] &= \frac{\mathbb{P}[B|A]\mathbb{P}[A]}{\mathbb{P}[B]} \\
&= \frac{0.02 \times 0.5}{0.01 \times 0.2 + 0.02 \times 0.5 + 0.03 \times 0.3} \\
&= 0.4762
\end{aligned}$$

6. Random Variables

6.1 Introduction

Many statistics books, especially the older ones, fail to give a mathematically clear and precise definition of the term "random variable". They will give many practical examples of it, like amount of rainfall, the height of 18-year olds, the number of rainy days in a week, the cholesterol level of 80-year olds...
These are all random variables in the sense that they vary in value according to the particular day, the adolescent, the week, the senior citizen you measure. Stopping short of saying what sort of mathematical object they are will not do.
We need to clarify the notion of this type of variability and see how we can model it with reference to some probability space which we have defined with patience and accuracy earlier.

6.2 Examples

Consider the following situations:

6.2.1 John and Mary play the following game with dice. After a throw, an odd number makes John give Mary 11 times the number. An even number makes Mary give John 9 times the number. How do we model John's and Mary's gains?

6.2.2 A vehicle uses 3 engines each with probability .01 of breakdown and operating independently of the others. The vehicle operates only with 3 running engines.
Furthermore onboard instrumentation can only tell the number of engines not running. Repairing each engine costs €20. How do we model prospective repair costs?

6.2.3 In example 2.2.4 with the CH_3OH molecule we saw a H atom, C atom or OH ion detach with probabilities 0.4, 0.3 , 0.1 respectively. Absorption of 10 units can knock off the H atom. 15 units can make the C atom detach while 25 units force the OH to leave. How do we model this?

6.2.4 How do we model scoring on a dart-board when an experienced player is playing?

6.2.5 How do we model the gains of a person who gets 2 to the power of the no of throws it takes for the first heads to appear when a dice is thrown in euros?

Modelling the basic situations using properly defined probability spaces is something we have already done earlier. Now we need to introduce a new idea. We have to define a quantity which depends on the particular possibility we have on hand. The examples chosen are clearly not the most useful one could find, but their relative simplicity allows us to better exemplify the underlying basic notions.

6.3 Definition of Random Variables

So we turn to formulating a precise mathematical definition of the notion of random variable. There are two main concepts involved which we introduce first again through the use of examples.

First we see how a random variable is nothing else but a function. In all these cases involving random variables we have a number of possibilities, whose occurrence is governed by some probability law, something which we saw earlier is modeled by a probability space $(\Omega,\mathcal{F},\mathbb{P})$. For each possible outcome we have a specific value. This situation can be captured mathematically by means of a function which maps the points in Ω into the real line $\mathbb{R}$.

In the examples mentioned above:

6.3.1 Example

From 6.2.1, the set of all possible outcomes is given by: $\Omega = \{1,2,3,4,5,6\}$

Let X be the random variable which models John's gains and Let Y be the random variable which models Mary's gains. Thus the random variables X and Y are defined as follows:

$$\begin{aligned} X(2i+1) &= -11 \times (2i+1) \\ X(2i) &= 9 \times (2i) \\ Y(\omega) &= -X(\omega) \end{aligned}$$

6.3.2 Example

From 6.2.2, the set of all possible outcomes is given by:

$$\Omega=\{(0,0,0),(1,0,0),(0,1,0),(0,0,1),(1,1,0),(1,0,1),\ 0,1,1),(1,1,1)\}$$

Let X be the random variable which gives the repair cost.

No Engine Failure{ $X((0,0,0))=0$ Three Engine Failures{ $X((1,1,1))=60$

$$\text{One Engine Failure} \begin{cases} X((1,0,0)) = 20 \\ X((0,1,0)) = 20 \\ X((0,0,1)) = 20 \end{cases}$$

$$\text{Two Engine Failures} \begin{cases} X((1,1,0)) = 40 \\ X((1,0,1)) = 40 \\ X((0,1,1)) = 40 \end{cases}$$

6.3.3 Example

From 6.2.3, the set of all possible outcomes is given by Ω = {0,1,2,3,4,5},
where: 0 → none, 1 → 1st H, 2 → 2nd H, 3 → 3rd H, 4 → C, 5 → OH

No atom detaches ⇒ { X(0) = 0

$$\text{Hydrogen atom} \begin{cases} X(1) = 10 \\ X(2) = 10 \\ X(3) = 10 \end{cases}$$

Carbon Atom detaches ⇒ $X(4) = 15$

OH ion detaches ⇒ $X(5) = 25$

6.3.4 Example

From 6.2.4, the set of all possible outcomes is given by:

$$\Omega = \{(x, y) : x^2 + y^2 = R^2\}$$

where R is the radius of the dartboard. If X is the random variable which represents the score then $X((x,y))$= sector number where the point $(x,\ y)$ lies.

6.3.5 Example

From 6.2.5, the set of all possible outcomes is the set of natural numbers:

$$\Omega = \mathbb{N}$$

Furthermore, the random variable X, which model the gains of this individual is defined as follows:

$$X(n) = 2^n$$

There is a second concept involved in defining random variables which we now consider. It is related to the idea of a σ-field. We have seen that the σ-field captures the information which we can have about probabilistically governed events. Forming some subsets of Ω but not others was at the root of this.

A random variable, being defined on Ω, can itself split it into subsets. It *should not be able to do this more finely than the given σ-field*. We now exemplify it through the examples we have constructed.

6.3.6 Example
From 6.2.1 and 6.3.1 we have that Ω = { 1,2,3,4,5,6},
The σ-field of Ω is in fact the power set.

Recall that:
$X(\{2i+1\})$ = -11$\times$ (2i+1)
$X(\{2i\})$ = 18i.

In this case we can form subsets as we please because all the individual points are distinguishable.

6.3.7 Example
Following 6.2.2 and 6.3.2,

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

In this case the σ-field is defined as follows:

$$\mathcal{F} = \{(0,0,0), \{(1,0,0),(0,1,0),(0,0,1)\}, (1,1,0),(1,0,1),(0,1,1)\}, (1,1,1), \text{all unions, all complements}, \Omega, \phi\}$$

Recall that:
$X((0,0,0)) = 0$,
$X((1,0,0))=X((0,1,0))=X((0,0,1))=20$
$X((1,1,0))= X((1,0,1)) =X((0,1,1)) =40$
and $X((1,1,1))=60$.

We note that in this case X does not split the subsets {(1,0,0),(0,1,0),(0,0,1)} & {(1,1,0), (1,0,1), (0,1,1)} into smaller parts.

6.3.8 Example
From 6.2.3 and 6.3.3 we have that Ω = { 0, 1, 2, 3, 4, 5 }

The corresponding σ-field is given by:

$$\mathcal{F} = \{0, \{1,2,3\}, 4, 5, \text{all unions, all complements}, \Omega, \phi\}$$

Recall that the random variable was defined as follows:

$$X(0)=0,\ X(1)=X(2)=X(3)=10,\ X(4)=15,\ X(5)=25$$

In this case X respects the σ-field on Ω by not splitting the subset {1, 2, 3} more finely than the corresponding σ-field.

The abstract property we have been driving at above can be captured with mathematical precision as follows:

6.3.9 Definition
Given a probability space (Ω, $\mathcal{F}$,$\mathbb{P}$) the mapping $X : \Omega \to \mathbb{R}$ is said to be a random variable if : $\forall\ x \in \mathbb{R}$, $\{\omega : X(\omega) > x\} \in \mathcal{F}$.

We shall show how X as defined in example 3 is a random variable and we shall give an example Y which is not a random variable.
With reference to example 3 above, we see that:
For $x < 0$, $\{\ \omega : X(\omega) > x\ \} = \ \Omega \in \mathcal{F}$

For $10 > x \geq 0$, $\{\ \omega : X(\omega) > x\ \} = \ \{1,2,3,4,5\} \in \mathcal{F}$

For $15 > x \geq 10$, $\{\ \omega : X(\omega) > x\ \} = \{4,5\} \in \mathcal{F}$

For $25 > x \geq 15$, $\{\ \omega : X(\omega) > x\ \} = \{\ 5\} \in \mathcal{F}$

For $x \geq 25$, $\{\ \omega : X(\omega) > x\ \} = \ \varnothing \in \mathcal{F}$

Now consider the function Y on Ω as follows:
$Y(0)=1,\ Y(1)=1,\ Y(2)=12,\ Y(3)=13\ ,\ Y(4)=15,\ Y(5)=25$

Y is not a random variable. It is able to distinguish between the subsets {2} and {3}. It "splits" Ω in a finer way than $\mathcal{F}$. This means that Y has more information about subsets of Ω than it should have. If Y has equal values on 2 and 3 are the same, then Y satisfies the condition above. By looking at the values which X assumes, on the other hand, we see that we cannot obtain smaller subsets than those contained in $\mathcal{F}$.
Functions which satisfy the condition above are called measurable functions with respect to the σ-field $\mathcal{F}$. Thus a random variable is nothing more than a function defined on a probability space which is measurable.

6.3.10 Note

1. If we replace $>$ with $<$, with $\geq$ or with $\leq$ in the above definition, the resulting definition is equivalent to the original one.

2. A random variable X generates out of some subsets of Ω a σ-field according to the information packed in X's definition.

3. Constant functions satisfy the above condition irrespective of σ-field F.

4. If the σ-field F is the set of all subsets of Ω, the power set of Ω , then all functions on Ω are random variables.

We shall consider a final example.
Our probability space models two independent tosses of a coin:

$$\Omega = \{(H,H), (T,T), (H,T), (T,H)\}$$

There are 2 σ-fields which we define on Ω:

F_1 which is the one corresponding to knowledge up to the first toss:

$F_1 = \{ \varnothing, \Omega, \{(H,H), (H,T)\},\{(T,T),(T,H)\}\}$

and F_2 which is the one corresponding to knowledge up to the 2^{nd} toss, so that F_2 is the power set of Ω.

We define two functions, one of which is a random variable with reference to F_1 and F_2, X, and another which is random variable with reference to F_2 but not to F_1 , Y.

X((H,H)) = 1 , X((H,T)) = 1,
X((T,T)) = 2 , X((T,H)) = 2.

Y((H,H)) = 2 , Y((H,T)) = 1,
Y((T,T)) = 0 , X((T,H)) = 1.

X tells us whether the first toss gave us heads or tails and says nothing about the second toss.
Y counts the number of heads in the two tosses. Notice that Y does not distinguish between a head coming first or second.

By considering more tosses, we see that we can have a larger Ω on which we can define more σ-fields which get larger as sets.

An interesting question would be, if we square a random variable, does it still remain a random variable? Or we could reformulate it more provocatively: By applying some formula to the value of X, can we squeeze more information about the probability space than X has already?

6.4 Probability Distributions

Before the Kolmogorov definition of a random variable was available and accepted widely by the mathematical community from the late 1930's onwards, it was customary not to bother at a precise definition of random variables and concentrate more on the manner in which the values assumed by then were "distributed". How frequently did random variable X assume the value x? This concept was sufficiently fruitful to give a large number of results. In fact it can be said that statistical theory was largely built on it. We shall clarify this notion.
Let us consider the following examples

1. A person wins twice the number showing up when a fair dice is thrown.
   $\Omega = \{1, 2, 3, 4, 5, 6\}$, $X(i) = 2i$. Then the range of X is $\{2, 4, 6, 8, 10, 12\}$
   X makes each value equally probable at $1/6$.

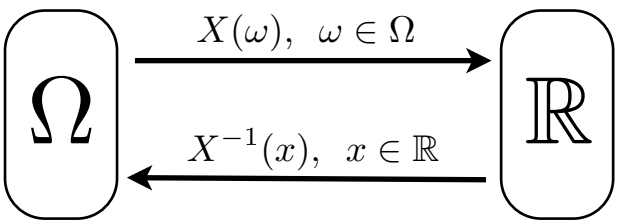
2. Two dice are thrown and we add the 2 values showing up.
   $\Omega = \{(x,y): x,y \in \{1, 2, 3, 4, 5, 6\}\}$ and $X((x,y)) = x+y$.
   The range of X is $\{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$ and the corresponding probabilities are $\{1/36, 2/36, 3/36, 4/36, 5/36, 6/36, 5/36, 4/36, 3/36, 2/36, 1/36\}$

3. A dice is thrown until heads shows up in throw n. A player is rewarded 2^n if n is odd and has to give 3^n if n is even.
   So $\Omega = \mathbb{N}$ and $X(n)$ is defined as follows:
   $$X(n) = \begin{cases} 2^n & n \text{ is even} \\ -3^n & n \text{ is odd} \end{cases}$$

4. The range of X is given by: $\{ 2, 2^3, 2^5, 2^7 , \ldots , 2^{2m+1}, \ldots\} \cup \{ -3^2, -3^4, -3^6, \ldots, -3^{2m},..\}$ with corresponding probabilities: $\{ 2^{-1}, 2^{-3}, 2^{-5}, 2^{-7} , \ldots, 2^{-(2m+1)},..\}$ and
   $\{ 2^{-2}, 2^{-4}, 2^{-6}, \ldots, 2^{-2m},\ldots\}$

5. A dart is thrown as a dartboard of radius R randomly. X gives us the distance of the dart from the centre. $\Omega =$ the disk $\{ (x,y): x^2+y^2 = R^2 \}$. Also $X((x,y)) = \sqrt{(x^2+y^2)}$. The range of X is $[0,R]$. Seeing how the probability is transferred onto this interval is slightly more complicated in this case.

Looking at process above: we see that the probability measure defined on Ω can be lifted to the range of X through X^{-1} in a systematic way. Let us consider random variables whose range consists of a countable set, that is discrete random variables.



Let $X(\Omega)$ be countable $\{ x_1, x_2, \ldots\ldots x_n, \ldots\ldots \}$ where we assume the x_n's are in ascending order.

Then the distribution of X is in effect a probability measure on this countable set given by :

$$\mathbb{P}_X[\{x_n\}] = \mathbb{P}[\omega : X(\omega) = x_n] = \mathbb{P}[X^{-1}(x_n)]$$

for $n = 1$, 2 , ...

An extension of this idea gives us the Probability Distribution Function. Rather than direct probability of a value x_n, here we have the cumulative probability.

6.4.1 Definition
The Probability Distribution Function of a discrete random variable X is a function, which we denote by F_X such that, $F_X : \mathbb{R} \to [0,1]$ and is defined as follows:

$$F_X(x) = \mathbb{P}[\omega : X(\omega) \leq x] = \sum_{n:x_n \leq x} \mathbb{P}[\omega : X(\omega) = x_n]$$

For a continuous random variable, that is to say a random variable whose range is $\mathbb{R}$, $\mathbb{R}^+$ or some interval in $\mathbb{R}$, the situation is more complicated. If we take any point x in the range of X, then $\mathbb{P}$ [$\omega : X(\omega) = x$] $= 0$. So we have to work with intervals.

The distribution of X is also in effect a probability measure in this case and is given by:

$$\mathbb{P}_X[[s,t]] = \mathbb{P}[\omega : s \leq X(\omega) \leq t] = \mathbb{P}[X^{-1}([s,t])]$$

Here we have also the cumulative probability defined by:

6.4.2 Definition
The Probability Distribution Function of a continuous random variable X is a function, which we denote by F_X such that, $F_X : \mathbb{R} \to [0,1]$ and is defined as follows:

$$F_X(x) = \mathbb{P}[\omega : X(\omega) \leq x]$$

How this probability is calculated we see in a later chapter.

Generally, we see that X^{-1} defines a probability measure on subsets of $\mathbb{R}$ or subsets of it as follows:
Assume that an appropriate σ-field of subsets has been defined on $\mathbb{R}$ call it $\mathcal{B}$. Then given $A \in \mathcal{B}$ we define:

$$\mathbb{P}_X[A] = \mathbb{P}[\omega : X(\omega) \in A]$$

It is not very difficult to prove that this set function is actually a probability measure.

The probability measure $\mathbb{P}_X$ is called the distribution of X. In effect it is the image of the probability measure lifted from Ω to $\mathbb{R}$ by X, and calculated through the use of X^{-1}.

7. Discrete Random Variables

7.1 The Uniform Distribution

7.1.1 Definition
A discrete random variable X is said to be uniformly distributed on the real numbers $\{x_1,\ldots,x_n\}$ if its distribution grants equal probability to each point in its range:

$$\mathbb{P}[\omega : X(\omega) = x_i] = \frac{1}{n} \qquad \text{for } 1 \leq i \leq n.$$

To understand this concept better lets us consider the following two examples:

7.1.2 Example
A fair dice is thrown and the random variable X corresponds to the square of the number on top.

X is defined on the familiar probability space Ω={1, 2, 3, 4, 5, 6} with power set and probability of 1/6 for each point. X has the following range:

$$\{1,\ 4,\ 9,\ 16,\ 25,\ 36\} = \{x_1,\ x_2,\ x_3,\ x_4,\ x_5,\ x_6\}$$

on which it is distributed uniformly:

$$\mathbb{P}[\omega : X(\omega) = x_i] = \mathbb{P}[\{i\}] = \frac{1}{6} \ \ \forall i$$

7.1.3 Example
A dart is thrown blindly on a disc of radius R which is divided into twenty equal sectors each radiating out of the centre. The sectors are marked 1 to 20 as the angle they subtend to the centre goes from 0° to 18°, 18° to 36°, .. and so on. X is the random variable corresponding to this score.
In this case:

$$\Omega = \{(r, \theta) : 0 \leq r \leq R,\ \ 0^\circ \leq \theta \leq 360^\circ\}$$

with the probability of subset A = (area of A) /(πR^2).
X is given by $X((\mathrm{r},\boldsymbol{\theta}\,)) = \mathrm{int}(\boldsymbol{\theta}\,/18)$. Furthermore X has the range:

$$\{1,\ 2,\ 3,\ 4,\ 5,\ 6,\ 7,\ 8,\ 9,\ 10,\ 11,\ 12\ ,13\ ,\ 14,\ 15,\ 16,\ 17,\ 18,\ 19,\ 20\}$$

Thus:

$$\mathbb{P}[\omega : X(\omega) = i] = \frac{\mathrm{Area}\{(r, \theta) : 18^\circ \times (i-1) \leq \theta \leq 18^\circ \times i\}}{\pi R^2} = \frac{1}{20}, \ \ \forall i$$

We notice that in both examples, the random variable X split Ω into n equally probable, disjoint subsets which partition Ω and labelled them x_1, x_2, ... , x_n.

7.2 The Binomial Distribution

In probability, sometimes we are faced with situations where an experiment is repeated a finite number of times n. Each experiment can have one of two outcomes: "success" or "failure", where p is the probability of a successful experiment. The individual preforming these experiments would like to be able to compute the probability that k out of n experiments are successful.

The random variable X counts the number of successful experiments and has a binomial distribution if the following conditions are met:

i. The number of observations (sometimes called experiments or trials) n is fixed.
ii. Each observation is independent.
iii. Each observation represents one of two outcomes ("success" or "failure").
iv. The probability of "success" p is the same for each outcome.

If these conditions are met, then X has a binomial distribution with parameters n and p, abbreviated B(n,p).

To illustrate the above concept, consider the following two scenarios:

1. Jane managed to collect 10 fertilized tortoise eggs each with probability 0.3 of hatching. She would like to know the probability that 4 out of 10 eggs will hatch.

2. Michelle planted 5 Daffodil seeds in different pots, each seed has a probability of 0.04 of growing. She would like to calculate the probability that 3 out of 5 seeds will grow.

7.2.1 Definition
A discrete random variable X is said to be binomially distributed with parameters n, p if

$$\mathbb{P}[\omega : X(\omega) = x_i] = {}^nC_i\, p^i(1-p)^{n-i} \qquad \text{for } 1 \le i \le n.$$

We see here that X induces a probability distribution on {0, 1, ..., n} by:

$$\mathbb{P}_X[\{i\}] = {}^nC_i\, p^i(1-p)^{n-i} \qquad \text{for } 1 \le i \le n.$$

Before giving further examples, we shall prove that the above is indeed a probability measure. The formula given above is a probability measure only if $\sum_{i=1}^{n}\mathbb{P}_X[\{i\}] = 1$.

Proof: $$\sum_{i=1}^{n}\mathbb{P}_X[\{i\}] = \sum_{i=1}^{n} {}^nC_i p^i[1-p]^{n-i}$$
$$= \sum_{i=1}^{n}[p+(1-p)]^n$$
$$= 1$$

7.2.2 Example

A coin is tossed three times. We are interested in counting the number of times a head turns up and to calculate its respective probability.

We start by defining the probability space as follows:

$\Omega = \{(x_1, x_2, x_3) : \ x_i = 0 \text{ or } 1\}$

$\mathcal{F}$ is the power set

$\mathbb{P}[0] = \frac{1}{2}$ and $\mathbb{P}[1] = \frac{1}{2}$

Next we define the random variable X which counts the number of heads as follows:

$$X((x_1, x_2, x_3)) = x_1 + x_2 + x_3$$

The range of X is {0, 1, 2, 3} furthermore the probability distribution of X is given by:

$$\mathbb{P}[\omega : X(\omega) = 0] = {}^3C_0 \left(\frac{1}{2}\right)^0 \left(1 - \frac{1}{2}\right)^{3-0} = \frac{1}{8}$$

$$\mathbb{P}[\omega : X(\omega) = 1] = {}^3C_1 \left(\frac{1}{2}\right)^1 \left(1 - \frac{1}{2}\right)^{3-1} = \frac{3}{8}$$

$$\mathbb{P}[\omega : X(\omega) = 2] = {}^3C_2 \left(\frac{1}{2}\right)^2 \left(1 - \frac{1}{2}\right)^{3-2} = \frac{3}{8}$$

$$\mathbb{P}[\omega : X(\omega) = 3] = {}^3C_3 \left(\frac{1}{2}\right)^3 \left(1 - \frac{1}{2}\right)^{3-3} = \frac{1}{8}$$

7.2.3 Example

A cruise liner is power by four diesel engines each with probability 0.01 of breaking down.

As before we start by stating the probability space:

$\Omega = \{(x_1, x_2, x_3, x_4) : x_i = 0 \text{ or } 1\}$, where 1 stands for engine breakdown.

$\mathcal{F}$ is the power set

$\mathbb{P}[0] = 0.99$ and $\mathbb{P}[1] = 0.01$

The random variable X which counts the number of engines that broke down is defined as follows: $X((x_1, x_2, x_3, x_4)) = x_1 + x_2 + x_3 + x_4$, furthermore the probability distribution of X is given by:

$$\mathbb{P}[\omega : X(\omega) = 0] = {}^4C_0 \left(\frac{1}{100}\right)^0 \left(1 - \frac{1}{100}\right)^{4-0} = 0.9606$$

$$\mathbb{P}[\omega : X(\omega) = 1] = {}^4C_1 \left(\frac{1}{100}\right)^1 \left(1 - \frac{1}{100}\right)^{4-1} = 0.0388$$

$$\mathbb{P}[\omega : X(\omega) = 2] = {}^4C_2 \left(\frac{1}{100}\right)^2 \left(1 - \frac{1}{100}\right)^{4-2} = 0.00058806$$

$$\mathbb{P}[\omega : X(\omega) = 3] = {}^4C_3 \left(\frac{1}{100}\right)^3 \left(1 - \frac{1}{100}\right)^{4-3} = 0.00000396$$

$$\mathbb{P}[\omega : X(\omega) = 4] = {}^4C_4 \left(\frac{1}{100}\right)^4 \left(1 - \frac{1}{100}\right)^{4-4} = 1 \times 10^{-08}$$

7.3 The Poisson Distribution

Sometimes there are situations where counting the number of events which occur within a set unit of time, area, volume or length is important. Consider the following examples:

1. the number of radioactive particles disintegrating per period of time
2. the number of cars passing through a junction within a fixed period of time
3. the number of persons joining a queue within a fixed time interval
4. the number of telephone calls received by a bank per hour.

These situations can be modelled by the so-called Poisson Distribution which is defined as follows.

7.3.1 Definition
A discrete random variable X is said to be Poisson distributed with parameter α if

$$\mathbb{P}[\omega : X(\omega) = i] = \frac{e^{-\alpha}\alpha^i}{i!} \quad \text{for } i = 0, 1, 2, \ldots$$

It is not difficult to check that the above formula does give a probability distribution on the natural numbers.

Proof:

$$\begin{aligned}
\sum_{i=0}^{\infty} \mathbb{P}[\omega : X(\omega) = i] &= \sum_{i=0}^{\infty} \frac{e^{-\alpha}\alpha^i}{i!} \\
&= e^{-\alpha} \sum_{i=0}^{\infty} \frac{\alpha^i}{i!} \\
&= e^{-\alpha} e^{\alpha} \\
&= 1
\end{aligned}$$

■

The Poisson distribution was first introduced by Simeon Denis Poisson (1781-1840). Historically the use of the Poisson distribution was established in approximations to the binomial distributions. Computing binomial probabilities is usually a bit problematical because of the ${}^nC_k$ coefficients. In the Poisson case, for large i , the probabilities die off quickly because $i!$ increases very, very fast in the denominator.

In more recent times this distribution has become extremely important in queuing theory and in the theory of discrete-valued stochastic processes.

In many applications involving discrete random variables whose values can be integers with no clear maximum value, or rather a theoretical infinite number of values, it is natural to see whether the Poisson distribution applies.

7.3.2 Example

Road engineers have been monitoring the flow of traffic during the rush hour in a particular main road by counting the number of vehicles which pass in front of a camera which had been previously installed in a strategic point of this road. Their report claims that the number of vehicles passing in front of this camera within a set time unit follows a Poisson distribution with parameter $\alpha = 9.78$.

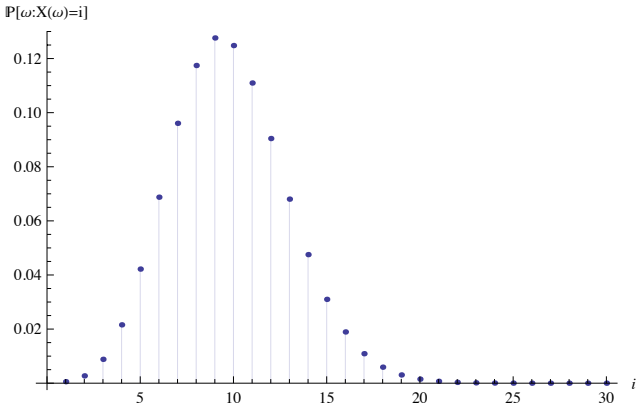
Using this information calculate the probability that 5, 10 or 20 cars pass through the camera within the set time unit.

$$\mathbb{P}[\omega : X(\omega) = 5] = \frac{e^{-9.78}(9.78^5)}{5!} = 0.0421806$$

$$\mathbb{P}[\omega : X(\omega) = 10] = \frac{e^{-9.78}(9.78^{10})}{10!} = 0.124803$$

$$\mathbb{P}[\omega : X(\omega) = 20] = \frac{e^{-9.78}(9.78^{20})}{20!} = 0.00149023$$

The diagram shows the shape of this distribution given that $\alpha = 9.78$.



7.4 The Geometric Distribution

In certain situations one might be interested in counting the number of times an experiment (or an event) is repeated until it is successful. Consider the following examples:

1. The number of times a computer hacker tries to infiltrate a computer network until he is successful.
2. The number of times a thief goes out on procurement trips before being caught by the police. (Assuming the the probability of being caught remains constant.)
3. The number of times an individual sits for the driving exam until he/she passes.

The model here is that of a succession of independent trials which stop with success. Assuming probability p of success and independence between trials, we see that the probability of stopping after one attempt is p, after two attempts is (1-p)p. After 3 attempts, two of which are unsuccessful and the last attempt being successful occurs with probability of (1-p)^2p.

7.4.1 Definition
A discrete random variable X is said to be geometrically distributed with parameter p if

$$\mathbb{P}[\omega : X(\omega) = i] = p(1-p)^{i-1} \qquad \text{for } i \geq 1.$$

We see here that X induces a probability distribution on $\mathbb{N}$ which is guaranteed by the equation:

$$\sum_{i=1}^{\infty} p(1-p)^{i-1} = 1$$

Proof is left as an exercise...

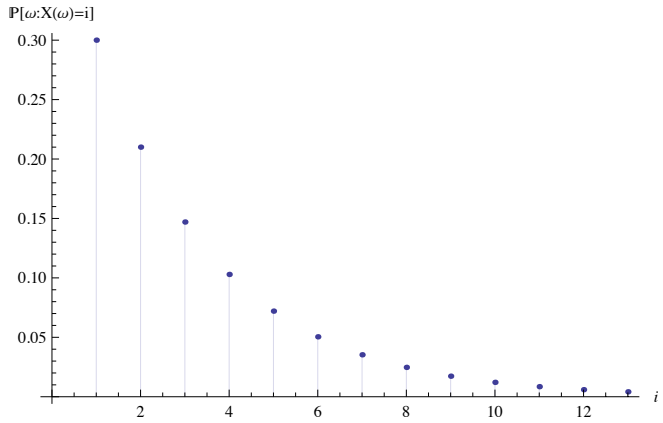
7.4.2 Example
Due to a number of fatal traffic accidents involving young drivers, the Transport Authority decided to make it harder for people to pass the driving exam. It is estimated that the probability of passing the exam is 0.3. Calculate the probability of individual passes the exam at the 3, 4 or 5 attempt.

$$\mathbb{P}[\omega : X(\omega) = 3] = 0.3[1-0.3]^2 = 0.147$$

$$\mathbb{P}[\omega : X(\omega) = 4] = 0.3[1-0.3]^3 = 0.1029$$

$$\mathbb{P}[\omega : X(\omega) = 5] = 0.3[1-0.3]^4 = 0.07203$$

The image illustrates the probability of passing the exam at various attempts.



7.5 The Hypergeometric Distribution

The hypergeometic distribution is applicable when n elements are selected at random from a finite population of size N whose elements can be subdivided in two sub populations one of size k and the other of size N - k . The random variable X counts the number of elements in the sample of size n that belong the subpopulation of size k. Consider the following examples:

1. A random sample of size 100 is chosen from the Maltese population. The number of Gozitans in the sample is counted.
2. Packets with 15 seeds are chosen from 1000 seeds of which 214 are diseased. The number of healthy seeds in the first packet chosen is counted.
3. A production cycle ends when 200 parts are finished, 10 of which are defective. The number of defective parts in the first batch of 20 is counted.

7.5.1 Definition

A random variable X is said to have the hypergeometric distribution with parameters N, n, k if:

$$\mathbb{P}[\omega : X(\omega) = i] = \frac{{}^{k}C_{i}\ {}^{N-k}C_{n-i}}{{}^{N}C_{n}} \qquad \text{for } i = 0, 1, 2, \ldots, k$$

The identity ${}^{N}C_{n} = \sum_{i=0}^{k} {}^{k}C_{i}\ {}^{N-k}C_{n-i}$ assures that the above gives a probability distribution on the set $\{0,1,2,\ldots,k\}$.

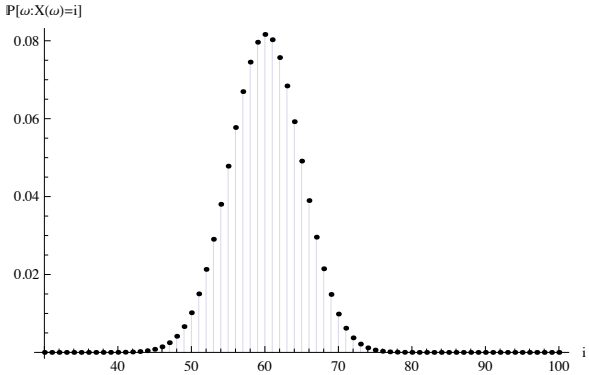
7.5.2 Example

There are 10,000 students currently reading for a degree with the University of Malta. Out of these 10,000 students, 6,000 have a maltese citizenship while the others are foreigners. If a sample of 100 students is chosen at random, calculate the probability that within this sample there are 30, 40 or 60 maltese students.

$$\mathbb{P}[\omega : X(\omega) = 30] = \frac{{}^{6000}C_{30}\ {}^{(10000-6000)}C_{(100-30)}}{{}^{10000}C_{100}} = 7.53516 \times 10^{-10}$$

$$\mathbb{P}[\omega : X(\omega) = 40] = \frac{{}^{6000}C_{40}\ {}^{(10000-6000)}C_{(100-40)}}{{}^{10000}C_{100}} = 0.0000225844$$

$$\mathbb{P}[\omega : X(\omega) = 60] = \frac{{}^{6000}C_{60}\ {}^{(10000-6000)}C_{(100-60)}}{{}^{10000}C_{100}} = 0.0816283$$



7.6 Multinomial Distribution

This distribution is a generalization of the Binomial distribution discussed earlier. Recall that in section 7.2, we discussed experiments which could have only two possible outcomes. However, in this case an experiment can lead to k different outcomes having probabilities $p_1, p_2, \ldots, p_k$.

N independent experiments are conducted and the number $n_1, n_2, \ldots, n_k$ of each outcome noted. Thus one should note that we do not have one reading but rather k readings at a time. This fact transforms the random variable into a random vector. By random vector we shall mean a collection of random variables taken together. Consider the following examples:

1. Samples of 1000 tourist records from last year are selected at random with replacement. Each record was tested to see whether the tourist originated from the US, Britain, Italy, Germany or some other country. The number for each category is noted.
2. A potent drug used as medication is known to have vomiting, diarrhea and dizziness as possible side effects in all combinations (8 in all). Fifty patients are treated with the drug. We're interested in counting the number of patients suffering from each of the side effects combination.
3. 30 persons in favour of introducing abortion in Malta are chosen at random and with replacement. The population is partitioned into 7 social classes and we are interested how many customers come from each of the classes.

For simplicity we shall be using the following notation in this section:

$$^{N}C_{j_1, j_2, \ldots, j_k} = \frac{N!}{j_1!\, j_2! \ldots j_k!}$$

7.6.1 Definition

A random vector of dimension k, $(X_1, X_2, \ldots, X_k)$ is said to have a multinomial distribution with parameters $k, N, p_1, p_2, \ldots, p_k$, where $\sum_{i=1}^{k} p_i = 1$ if for positive integer $j_1, j_2, \ldots, j_k$ satisfying: $\sum_{i=1}^{k} j_i = N$,we have $P[X_i = j_i] = {}^{N}C_{j_1, j_2, \ldots, j_k} p_1^{j_1} \ldots p_k^{j_k}$ for $i = 1,2,\ldots,k$.

7.7 The Negative Binomial Distribution

The negative binomial distribution is a generalization of the geometric distribution. In fact, the negative binomial distribution is the probability distribution of the number of trials (x) which are needed to obtain a number of successes (r) in repeated independent trials. Each trial can only give rise to two mutually exclusive outcomes say, success or failure. Where the probability of success is p and the probability of failure is (1-p). Furthermore these probabilities remain the same from trial to trial.

Consider the following examples:

1. An amateur archer is shooting arrows at a target. We would like to calculate the probability that he hits the target the a second time after the 6th attempt.
2. A person conducting telephone surveys usually dials telephone numbers from a long list and he would like to know the probability of successfully completing the 3rd survey after the 10th telephone call.

If the r^{th} success is to occur on the x^{th} trial, then there must be $r-1$ successes on the first $x-1$ trials, and the probability for this is:

$$^{x-1}C_{r-1}p^{r-1}(1-p)^{x-r},$$

Furthermore the probability of success on the x^{th} trial is p, and the probability that the r^{th} success occurs on the x^{th} trial is :

$$p^{x-1}C_{r-1}p^{r-1}(1-p)^{x-r}$$

7.7.1 Definition

A random variable X has a negative binomial distribution with parameters r and p if

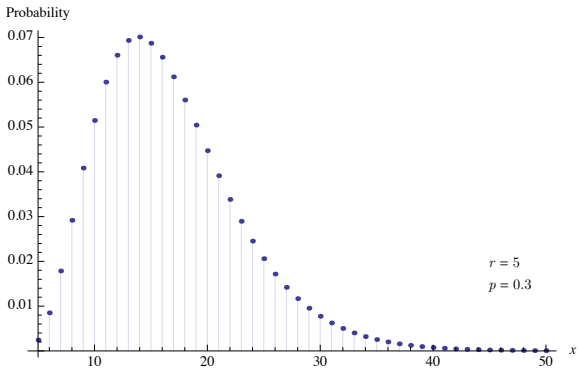
$$\mathbb{P}[\omega : X(\omega) = x] = {}^{x-1}C_{r-1}p^{r}(1-p)^{x-r} \quad \text{for } x = r, r+1, r+2, \ldots$$

Note when $k = 1$ the above becomes the geometric distribution.

7.7.2 Example

John is trowing a tricked dice for which the probability of scoring 6 is 0.3. Find the probability that:

1. he obtains the number 6 for the second time 10 tosses:
$$\mathbb{P}[X = 10] = {}^{10-1}C_{2-1}(0.3)^2(1-0.3)^{10-2}$$
$$= 0.0467$$
2. he obtains the third number 6 in 5 tosses.
$$\mathbb{P}[X = 5] = {}^{5-1}C_{3-1}(0.3)^3(1-0.3)^{5-3}$$



7.8 Expectations

The concept of average is of course a very important one in probability. We shall define it with precision and then proceed to derive expressions for the expectations of distributions introduced earlier.

7.8.1 Definition
Let X be a discrete random variable with range given by $\{x_1, x_2, \ldots, x_n, \ldots\}$ we define the expectation of X as follows:

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} x_i \mathbb{P}[\omega : X(\omega) = x_i]$$

7.8.2 Example:
John goes to the casino and plays the following game:
He throws a fair dice if the numbers 1 or 2 appear he wins 2 euros and 5 euros respectively. If 3 appears he wins nothing,
If 4 appears he looses 4 euros,
If 5 appears he wins 7 euros,
If 6 appears he looses 10 euros.
What is his expected gain?
Let X be the random variable which models his gains, then:

$$X(i) = \begin{cases} 2 & \text{if} \quad i = 1 \\ 5 & \text{if} \quad i = 2 \\ 0 & \text{if} \quad i = 3 \\ -4 & \text{if} \quad i = 4 \\ 7 & \text{if} \quad i = 5 \\ -10 & \text{if} \quad i = 6 \end{cases}$$

$$\begin{aligned} \mathbb{E}[X] &= 2\left(\frac{1}{6}\right) + 5\left(\frac{1}{6}\right) + 0\left(\frac{1}{6}\right) - 4\left(\frac{1}{6}\right) + 7\left(\frac{1}{6}\right) - 10\left(\frac{1}{6}\right) \\ &= 0 \end{aligned}$$

We note the following results which follow easily from the properties of summations:

7.8.2 Properties of Expectations
For any constant α we have that:

$$\mathbb{E}[X + \alpha] = \mathbb{E}[X] + \alpha$$
$$\mathbb{E}[\alpha X] = \alpha \mathbb{E}[X]$$

These two properties of the expectation operator are linear and help us derive with ease a number of important results which would be more difficult to obtain directly.

We note that the definition of expectation allows us to define immediately the expectation of any function of a random variable. We consider only the case of a random variable raised to some power m, X^m :

$$\mathbb{E}[X^m] = \sum_{i=1}^{\infty} x_i^m \mathbb{P}[\omega : X(\omega) = x_i]$$

We note that in general $\mathbb{E}[X^m] \neq (\mathbb{E}[X])^m$. It is true is very special cases like when X is uniformly distributed. We next derive the expectations of the distributions we introduced earlier.

7.8.3 Uniform Distribution

For a random variable X having a uniform distribution, its expectation its computed as follows:

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{i=1}^{n} x_i \mathbb{P}[\omega : X(\omega) = x_i] \\
&= \sum_{i=1}^{n} x_i \frac{1}{n} \\
&= \frac{1}{n} \sum_{i=1}^{n} x_i
\end{aligned}$$

7.8.4 Binomial Distribution

For a binomially distributed random variable X with parameters n, p the expectation is:

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{i=1}^{n} i \, \mathbb{P}[\omega : X(\omega) = i] \\
&= \sum_{i=1}^{n} i \, {}^{n}C_i \, p^i (1-p)^{n-i} \\
&= \sum_{i=1}^{n} \frac{i(n!)}{i!(n-i)!} p^i (1-p)^{n-i} \\
&= np \sum_{i=1}^{n} \frac{(n-1)!}{(i-1)!(n-i)!} p^{i-1} (1-p)^{([n-1]-[i-1])} \\
&= np \sum_{i=1}^{n} {}^{n-1}C_{i-1} \, p^{i-1} (1-p)^{([n-1]-[i-1])} \\
&= np \sum_{i=0}^{n} {}^{n-1}C_i \, p^i (1-p)^{n-1-i} \\
&= np(p+1-p)^{n-1} \\
&= np
\end{aligned}$$

7.8.5 Poisson Distribution

Let X be a Poisson distributed random variable with parameter α , the the expectation is:

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{i=1}^{\infty} i\,\mathbb{P}[\omega : X(\omega) = i] \\
&= \sum_{i=1}^{\infty} \frac{i\,e^{-\alpha}\alpha^i}{i!} \\
&= \alpha\,e^{-\alpha} \sum_{i=1}^{\infty} \frac{\alpha^{i-1}}{(i-1)!} \\
&= \alpha\,e^{-\alpha} \sum_{i=0}^{\infty} \frac{\alpha^{i}}{i!} \\
&= \alpha\,e^{-\alpha}\,e^{\alpha} \\
&= \alpha
\end{aligned}$$

7.8.6 Geometric Distribution

Let X be geometrically distributed with parameter p. Then:

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{i=1}^{\infty} i\mathbb{P}[\omega : X(\omega) = i] \\
&= \sum_{i=1}^{\infty} ip(1-p)^{i-1} \\
&= p\sum_{i=1}^{\infty} i(1-p)^{i-1} \\
&= p\left(\frac{\partial}{\partial x}\sum_{i=1}^{\infty} x^i\right)_{x=1-p} \\
&= p\left(\frac{\partial}{\partial x}\left[\frac{x}{1-x}\right]\right)_{x=1-p} \\
&= p\left(\frac{x}{(1-x)^2} + \frac{1}{1-x}\right)_{x=1-p}
\end{aligned}$$

As soon as we substitute for x and simplify we obtain:

$$\mathbb{E}[X] = \frac{1}{p}$$

7.8.7 Hypergeometric Distribution

Let X have a hypergeometric distributed with parameters N, n, k , then the expectation is:

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{i=1}^{k} i\,\mathbb{P}[\omega : X(\omega) = i] \\
&= \sum_{i=1}^{k} i \frac{{}^{k}C_i\ {}^{N-k}C_{n-i}}{{}^{N}C_n} \\
&= \sum_{i=1}^{k} \frac{k!}{(i-1)!(k-i)!} \frac{{}^{N-k}C_{n-i}}{{}^{N}C_n} \\
&= \frac{k}{{}^{N}C_n} \sum_{i=1}^{k} \frac{(k-1)!}{(i-1)!(k-i)!}\ {}^{N-k}C_{n-i} \\
&= \frac{k}{{}^{N}C_n} \sum_{i=1}^{k} {}^{k-1}C_{i-1}\ {}^{N-k}C_{n-i} \\
&= \frac{k}{{}^{N}C_n} \sum_{i=0}^{k-1} {}^{k-1}C_i\ {}^{N-k}C_{n-(i+1)} \\
&= \frac{k\ {}^{N-1}C_{n-1}}{{}^{N}C_n} \\
&= \frac{nk}{N}
\end{aligned}$$

7.8.8 Negative Binomial Distribution

Let X have a negative binomial distribution with parameters r and p, then the expectation is computed as follows:

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{x=r}^{\infty} x\ {}^{x-1}C_{r-1}\, p^r (1-p)^{x-r} \\
&= \sum_{x=r}^{\infty} \frac{x(x-1)!}{(r-1)!(x-r)!} p^r (1-p)^{x-r} \\
&= rp^r \sum_{x=r}^{\infty} \frac{x!}{r!(x-r)!} (1-p)^{x-r} \\
&= rp^r \sum_{x=r}^{\infty} {}^{x}C_r (1-p)^{x-r} \\
&= rp^r \sum_{x=0}^{\infty} {}^{x+r}C_r (1-p)^{x} \\
&= rp^r \left(\frac{1}{(1-[1-p])^{r+1}} \right) \\
&= \frac{r}{p}
\end{aligned}$$

Note:

Here we are using the standard result:

$$\sum_{k=0}^{\infty} {}^{s+k-1}C_{s-1}\ \varphi^k = \frac{1}{(1-\varphi)^s}$$

7.9 Variances

Random variables model variability according to different possibilities. It follows that measuring the degree of variability of a random variable should be an important theme in the study of random variables. In defining a method for measuring the variability of random variables we consider the following points:

1. The variability of some quantity has to be measured relative to some fixed point (the most natural point for a random variable is its mean).
2. The variability of a random variable itself varies from possibility to another, so it makes sense to average out this variability if we want some collective measure.
3. Variability about the mean is on average zero: by definition the random variable varies as much above the mean as below.

These 3 points suggest a natural candidate for measuring the variability of a random variable: the mean deviation squared from the mean, which we call variance.

7.9.1 Definition

The variance of a random variable, when it exists is defined by:

$$\mathbb{V}\text{ar}[X] = \sum_{i=1}^{\infty} (x_i - \mu)^2 \, \mathbb{P}[\omega : X(\omega) = x_i],$$

where $\mu = \mathbb{E}[X]$.

Furthermore the above can be re-written as follows:

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

Two important properties which can be easily proven from the above definition are:
For any constant real number α , we have:

$$\mathbb{V}\text{ar}[X + \alpha] = \mathbb{V}\text{ar}[X] \qquad\qquad \mathbb{V}\text{ar}[\alpha X] = \alpha^2 \, \mathbb{V}\text{ar}[X]$$

7.9.2 Uniform Distribution of first n integers

Let X be uniformly distributed on (1, 2, ..., n).
Using the concepts introduced in the previous section we can show that:

$$\mathbb{E}[X] = \frac{(n+1)}{2} \qquad\qquad \mathbb{E}[X^2] = \frac{1^2 + 2^2 + \cdots + n^2}{n} = \frac{n(n+1)(2n+1)}{6n}$$

Hence:

$$\begin{aligned}\mathbb{V}\text{ar}[X] &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \\ &= \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4} \\ &= \frac{(n^2-1)}{12}\end{aligned}$$

7.9.3 Binomial Distribution

To find the variance of a binomially distributed random variable we use the following procedure:

$$\begin{aligned}\mathbb{E}[X(X-1)] &= \sum_{k=0}^{n} k(k-1)\,{}^nC_k\, p^k(1-p)^{n-k} \\ &= p^2 \sum_{k=0}^{n} k(k-1)\,{}^nC_k\, p^{k-2}(1-p)^{n-k} \\ &= p^2 \frac{\partial^2}{\partial x^2}\left(\sum_{k=0}^{n} {}^nC_k\, x^k(1-p)^{n-k}\right)_{x=p} \\ &= p^2 \frac{\partial^2}{\partial x^2}\left([x+1-p]^n\right)_{x=p} \\ &= p^2\left(n(n-1)(x+1-p)^{n-2}\right)_{x=p} \\ &= n(n-1)p^2\end{aligned}$$

Furthermore:

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

Thus:

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

7.9.4 Poisson Distribution

Let X be a Poisson distribute random variable with parameter α , then:

$$\begin{aligned}
\mathbb{E}[X(X-1)] &= \sum_{k=1}^{\infty} k(k-1)e^{-\alpha}\frac{\alpha^k}{k!} \\
&= \alpha^2 e^{-\alpha}\sum_{k=0}^{\infty}\frac{k(k-1)\alpha^{k-2}}{k!} \\
&= \alpha^2 e^{-\alpha}\frac{\partial^2}{\partial\alpha^2}\sum_{k=0}^{\infty}\frac{\alpha^k}{k!} \\
&= \alpha^2 e^{-\alpha}e^{\alpha} \\
&= \alpha^2
\end{aligned}$$

Hence, $\mathbb{E}[X^2] = \mathbb{E}[X(X-1)] + \mathbb{E}[X]$
$= \alpha^2 + \alpha$

Therefore: $\mathbb{V}\text{ar}[\text{X}] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$
$= \alpha$

7.9.5 Geometric Distribution

Given that X is a geometrically distributed random variable, then:

$$\begin{aligned}
\mathbb{E}[X(X-1)] &= \sum_{k=1}^{\infty} k(k-1)p(1-p)^{k-1} \\
&= p(1-p)\sum_{k=1}^{\infty} k(k-1)(1-p)^{k-2} \\
&= p(1-p)\frac{\partial^2}{\partial p^2}\sum_{k=1}^{\infty}(1-p)^k \\
&= p(1-p)\frac{\partial^2}{\partial p^2}\left(\frac{1-p}{p}\right) \\
&= \frac{2(1-p)}{p^2}
\end{aligned}$$

So, $\mathbb{V}\text{ar}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$
$= \frac{2(1-p)}{p^2} + \frac{1}{p} - \frac{1}{p^2}$
$= \frac{(1-p)}{p^2}$

7.9.6 Negative Binomial Distribution

Given that X is a random variable which has a negative binomial distribution with parameters r and p, then:

$$\begin{aligned}
\mathbb{E}[X(X+1)] &= \sum_{x=r}^{\infty} x(x+1)\,{}^{x-1}C_{r-1}\,p^r(1-p)^{x-r} \\
&= r(r+1)\sum_{x=r}^{\infty}\frac{(x+1)!}{(r+1)!(x-r)!}\,p^r(1-p)^{x-r} \\
&= r(r+1)p^r\sum_{x=r}^{\infty}{}^{(x+1)}C_{r+1}\,(1-p)^{x-r} \\
&= r(r+1)p^r\sum_{x=0}^{\infty}{}^{(x+r+1)}C_{r+1}\,(1-p)^{x} \\
&= r(r+1)p^r\left(\frac{1}{(1-[1-p])^{r+2}}\right) \\
&= \frac{r(r+1)}{p^2}
\end{aligned}$$

Hence:

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

Finally we can deduce that:

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

7.10 Computation of Probabilities, Expectations and Variances

We consider a set of problems related to calculating probabilities with reference to discrete distributions.

7.10.1 Example

The number of days in a working week of 5 days when values of shares of Glormu Galea Enterprises manage to end up higher at the end of the trading day than at the start have been noted for 60 successive weeks:

4	4	4	3	4	3	2	3	2	4
2	4	3	2	2	4	5	3	2	3
5	2	2	0	2	2	2	3	4	3
4	3	3	3	3	4	3	2	3	3
2	3	1	3	4	2	3	3	3	3
5	3	2	3	3	3	3	4	2	3

1. Fit an appropriate distribution.
2. Calculate the probability that on two successive weeks there are no days when an increase in value occurs.
3. Find the probability that on two successive weeks there has been an increase for at least four days.

Solution

1. From the given sample we can compute the mean and variance:

Sample size: 60

Mean: 2.95

Variance: 0.9464

Furthermore it can be noticed that the numbers in the above data set can only be whole numbers between 0 and 5. It therefore sounds plausible to fit a binomial distribution.
From the previous sections we know that $\mathbb{E}[X] = np$, while $\mathbb{V}\text{ar}[X] = np(1-p)$

Obviously we know that $n = 5$, so

$$\mathbb{E}[X] = np$$
$$2.95 = 5p$$
$$p = 0.59$$

Thus fitted distribution is binomial with parameters: $n = 5, p = 0.59$

2. $\mathbb{P}[\omega : X(\omega) = 0] \times \mathbb{P}[\omega : X(\omega) = 0] = (1-p)^5 \times (1-p)^5 = 0.0001342$

3. $\mathbb{P}[\omega : X(\omega) \geq 4] \times \mathbb{P}[\omega : X(\omega) \geq 4] = \left[p^5 + {}^5C_4 p^4 (1-p) \right]^2$

7.10.2 Example

During a study of traffic flows along a by-pass road connecting the two heavily populated towns of Birkirara and Msida, the number of cars passing during one minute on weekdays at lunchtime were measured.
For 120 times the number of cars were counted by specially positioned observers and their readings are reproduced below:

5	2	0	1	3	7	2	4	1	4
0	1	4	3	5	3	9	5	2	5
3	9	3	3	7	3	7	2	7	5
6	3	4	7	5	3	2	3	3	7
4	3	5	4	3	2	4	2	2	5
3	2	4	2	3	0	4	3	3	3
5	1	6	5	4	7	1	5	6	2
3	6	1	4	4	2	3	5	4	5
5	2	3	3	3	6	2	4	3	1
4	4	5	2	3	6	4	4	4	4
1	3	7	4	7	0	4	3	7	2
4	10	2	4	2	7	5	4	1	1

1. Fit a suitable distribution to the above data.
2. Find the probability that four cars pass in one minute.
3. Find the probability that at more than 5 cars pass in one minute.
4. Find the probability that 240 cars pass in one hour.

Solution

1. From the given data we observe:

Sample size: 120

Mean: 3.7333

Variance: 3.9541

Seeing that the variance and the mean are quite close the Poisson assumption is not a bad one. So in this case we may estimate the parameter $\alpha \cong 3.73$

2. $\mathbb{P}[\omega : X(\omega) = 4] = \dfrac{e^{-\alpha}\alpha^4}{4!} = 0.1935$.

This can be evaluated in Matlab by using the command: poisspdf(4, 3.73).

3. $\mathbb{P}[\omega : X(\omega) > 5] = 1 - \mathbb{P}[\omega : X(\omega) \le 5]$

$$= 1 - \sum_{i=0}^{5} \mathbb{P}[\omega : X(\omega) = i]$$
$$= 1 - 0.8258$$
$$= 0.1742$$

This can be evaluated in Matlab by using the command: 1-poisscdf(5,3.73).

4. We start by noting that the time interval is now 60 minutes. This means that the parameter α has to be suitable altered.

   If $\alpha \cong 3.73$ for one minute, than $\alpha \cong 223.8$ for 60 minutes.

   Let Y be the random variable which counts the number of cars which pass in 60 minutes, then:

   $\mathbb{P}[\omega : Y(\omega) = 120] = \frac{e^{-\alpha}\alpha^{120}}{120!} = 9.1772\text{e-}15$

8. Continuous Random Variables

8.1 Introduction

Discrete random variables we saw are easy to model because one had a countable set of values, which we usually identify with some subset of $\mathbb{N}$. Each point i having its probability p_i, we saw, allowed us to treat the probability as though it was lifted from the underlying probability space up to the range of our random variable. However there are many random variables whose range cannot be discrete.

There are are in fact many practical examples:

1. daily rainfall in mm,
2. weight or height of persons,
3. daily relative humidity,
4. lifetimes of industrial components.

The problem with such variables is that the range of X has "too many" points! So many points that each point has to have 0 probability, but so many points that when taken together they can recover some strictly positive probability. This problem is not new. We saw it with geometric probability. And we can be guided by the same idea. Using area of points to serve as probability.

Let's consider the following scenario:
100,000 individuals were randomly selected from a population and their height (in cm) was measured. Both histograms below were plotted using the same sample:

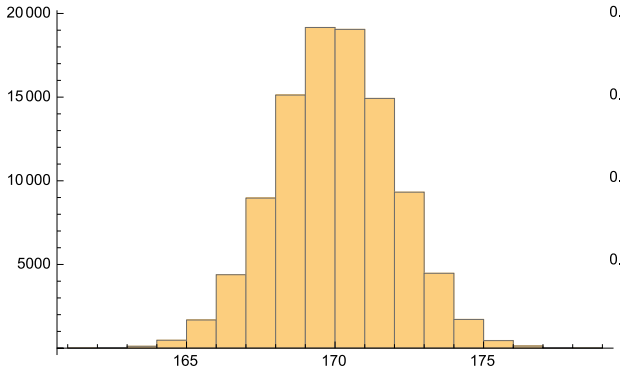


Fig. 1

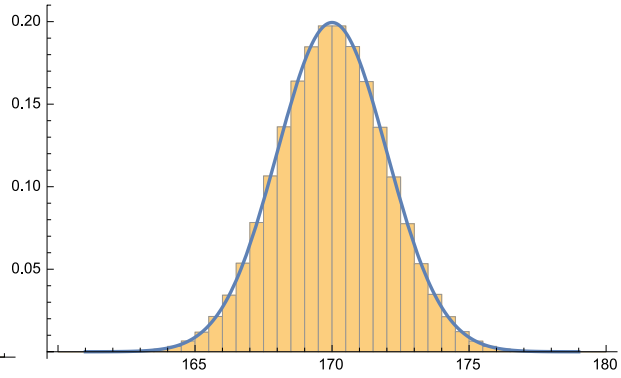
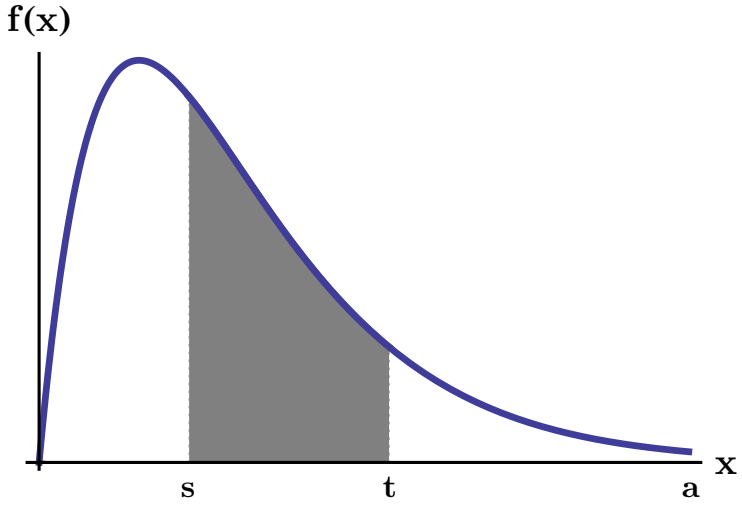


Fig. 2

Let's concentrate on fig.1. By definition, the sum of the height of the bars is equal to 100,000. To obtain the histogram in fig.2, we first reduced the bin size (i.e. we reduced the width of each bar) and divided the height of each bar by 100,000. Clearly this means that the sum of the new height of the bars is now equal to 1. Furthermore we notice that as the bin size goes to zero, the shape of the histogram is approaching the shape of the curve whose area is again equal to 1. This curve is called the Probability Density Function and will be discussed in the following section.

8.2 Probability Density Functions and Distribution Functions

Let us suppose that we have the interval [0, a] and we want to spread a probability measure on it. One simple way is to define a function over the interval, under which the area equals 1.



Then the probability of the interval [s, t] would be the area on top of this interval. Any function defined on some interval [a, b] such that:

1. $f(x) \geq 0$ for all x satisfying $a \leq x \leq b$
2. $\int_a^b f(x)dx = 1$

provides us with a way for defining probabilities over any interval [a , b].

We could extend this process to all of $\mathbb{R}$ or just its positive part $\mathbb{R}_+$. In these cases we would have $\int_{-\infty}^{\infty} f(x)dx = 1$ or $\int_0^{\infty} f(x)\,dx = 1$ respectively.

Such functions are called **density functions**. Their integrals, suitably defined, give what are called **probability distribution functions** F, which are defined by:

$$F(z) = \int_a^z f(x)dx$$

If f is defined on $\mathbb{R}$, the limit a in the integral is replaced by $-\infty$. In the case of $\mathbb{R}_+$ it is replaced by 0.

Note that $F(z)$ is an increasing function in z which starts from 0 and increases to a supremum of 1. It is immediate from the fundamental theorem of calculus that the derivative of F is f :

$$\frac{dF(z)}{dz} = f(z)$$

8.2.1 Definition

A random variable X is said to be continuous with density function *f*, if:

1. the range of X is some interval $[a,\ b]$ where a and b could be infinite.
2. for all z we have that $P[\omega : X(\omega) < z] = \int_a^z f(x)\,dx$.

To use a more general notation we shall assume that density functions are defined on $\mathbb{R}$ and we set the density function to be zero outside its proper domain. Thus if $f(x)$ is defined on $[a,\ b]$, then we extend $f(x)$ to be 0 for $x < a$ and for $x > b$. We can then use

$$\int_{-\infty}^{\infty} f(x)\,dx = 1\,.$$

8.2.2 Theorem

If X is a random variable with probability density function $f(x)$ then:

1. $\mathbb{P}[\omega : X(\omega) = x] = 0$,
2. $\mathbb{P}[\omega : X(\omega) \geq x] = \mathbb{P}[\omega : X(\omega) > x] = \int_x^{\infty} f(x)dx$,
3. $1 - \mathbb{P}[\omega : X(\omega) < x] = 1 - \mathbb{P}[\omega : X(\omega) \leq x]$,
4. $\mathbb{P}[\omega : s \leq X(\omega) \leq t] = \int_s^t f(x)dx$.

Proof

The proofs of these results follow immediately from the properties of the integral. In fact more than through a formal proof, we can see that the equations above are true by definition.

The area of a function over a point is 0. Also we have results of the type:

$$\int_s^t f(x)\,dx = \int_{-\infty}^t f(x)\,dx - \int_{-\infty}^s f(x)\,dx\,.$$

We can repeat the arguments we had about discrete random variables inducing probability measures on $\mathbb{N}$ or subsets of it. For continuous random variables we need to work with $\mathbb{R}$ or subsets of it.

So let X be a random variable with density function$f(x)$. Then X induces a probability measure $\mathbb{P}_X$ on $\mathbb{R}$ as follows:

Take any interval (s,t) , then $\mathbb{P}_X[(s,t)] = \mathbb{P}[\omega : s < X(\omega) < t]$, that is we take any interval , take its inverse image under X. This gives us a set in Ω. This set we know is in the σ-field. So we can find its probability. It thus transfers its probability to the interval (s,t) .

This process explains how one can ignore the underlying probability space and work with distribution on the real line $\mathbb{R}$.

Denoting the probability distribution function of random variable X by F_X, we notice also that $\mathbb{P}_X[(s,t)] = \mathbb{P}[\omega : s < X(\omega) < t] = F_X(t) - F_X(s)$

8.3 Uniform Distribution

Consider the following examples:

1. A string of length l snaps randomly at any point from 0 to l with uniform probability. We are interested in the distance of the point where the strings breaks from 0.
2. A signal starts transmitting and stays on for a minimum of α seconds and a maximum of β seconds. It can stop at any intermediate time with uniform probability.

8.3.1 Definition

A random variable X is said to be **uniformly distributed** on the interval $[\alpha,\beta]$ if for all $\alpha \le s < t \le \beta$ we have:

$$\mathbb{P}[\omega : s < X(\omega) < t] = \int_s^t \frac{1}{\beta - \alpha} dx$$

We note the probability distribution function $F(z)$:

$$F(z) = \int_\alpha^z \frac{dx}{\beta - \alpha} = \frac{z - \alpha}{\beta - \alpha}$$

That the function $1/(\beta-\alpha)$ can serve as a density function follows immediately from elementary properties of its integral. We note that there exists no continuous uniform distribution on an unbounded subset of $\mathbb{R}$.

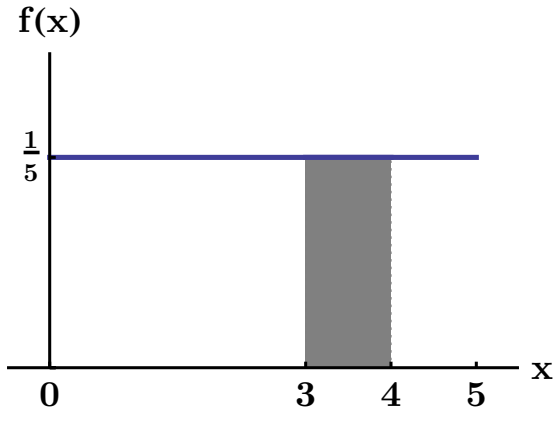
8.3.2 Example

An inelastic string of length 5cm is stretched until it breaks in two parts. The point where the string breaks is assumed to be uniformly distributed. Let X be the random variable which measures the length of the longer part of the string. Evaluate $\mathbb{P}[\omega : 3 \le X(\omega) \le 4]$.

Solution

The required probability is computed as follows:

$$\mathbb{P}[\omega : 3 \le X(\omega) \le 4] = \int_3^4 \frac{x}{5-0} dx = 0.7 .$$



8.4 Exponential Distribution

Consider the following examples:

1. The duration of telephone conversations is a random variable which takes short values with greater frequency than long ones.
2. The lifetimes of many household and industrial appliances, like batteries, electronic components have distributions which when scaled suitably, tend to favour the smaller values to the higher ones.
3. The waiting time between the arrivals of customers in a queue, or between the arrivals of cars at a junction or between the arrivals of planes at an airport.

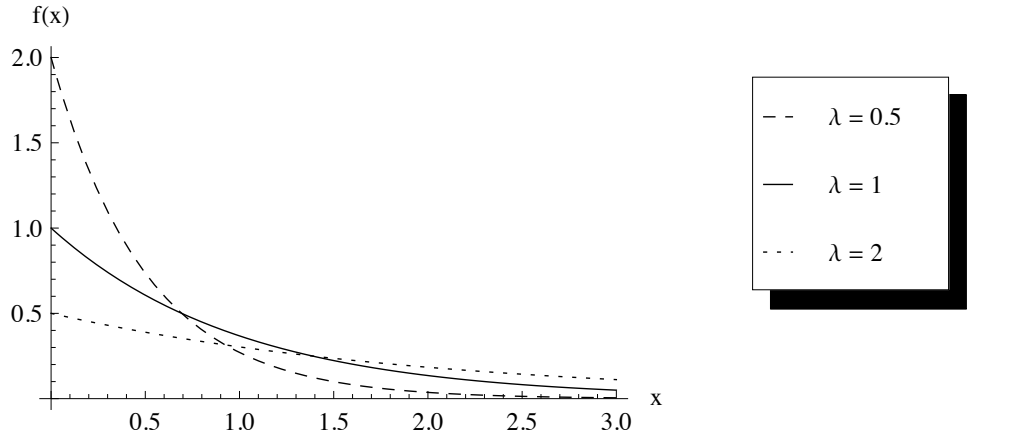
8.4.1 Definition

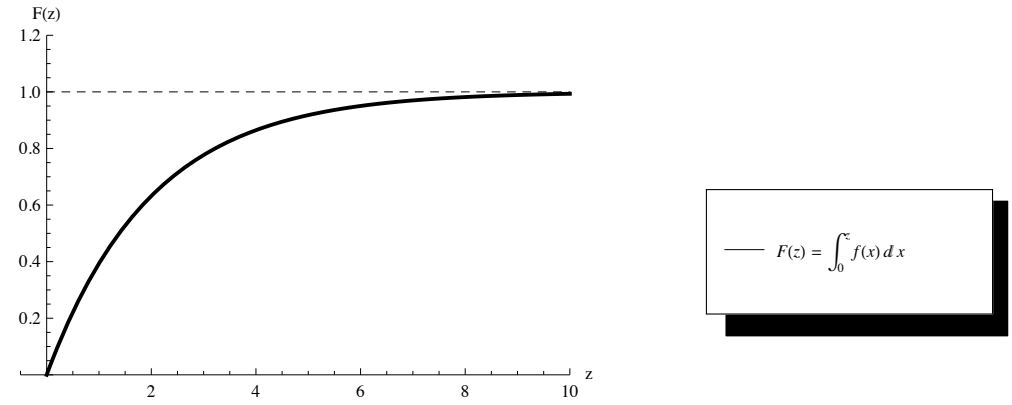
A random variable X is said to be exponentially distributed with parameter λ if for all $0 < s < t$, it satisfies the equation:

$$\mathbb{P}[\omega : 0 < X(\omega) < t] = \int_0^t \frac{1}{\lambda} e^{\frac{-x}{\lambda}}\, dx\,.$$

We note the probability distribution function $F(z)$ is defined:

$$F(z) = \int_0^z \frac{1}{\lambda} e^{\frac{-x}{\lambda}}\, dx = 1 - e^{\frac{-z}{\lambda}}\,.$$





We note that $F(0) = 0$ and $\lim_{z\to\infty} F(z) = 1$.

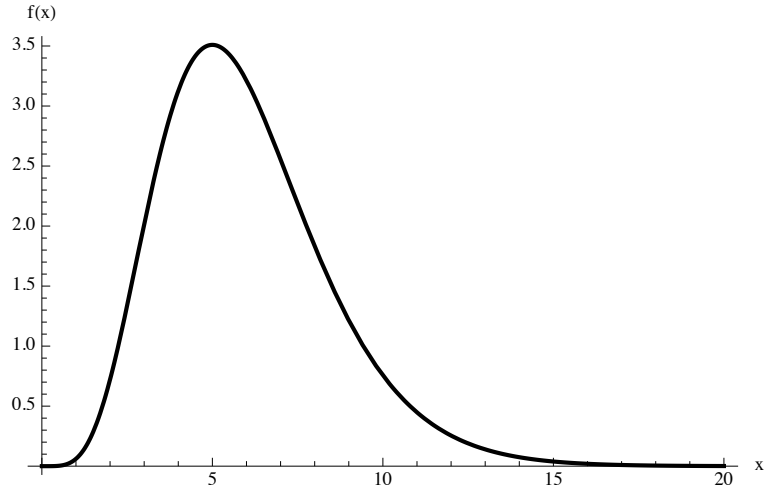
Some distributions in fact have a shifted exponential distribution:

$$\mathbb{P}[\omega : s < X(\omega) < t] = \int_s^t \frac{1}{\lambda} e^{\frac{-(x-\delta)}{\lambda}}\, dx \quad \text{for } \delta < s < t$$

with $F(z) = 0$ for $z \le \delta$ and $F(z) = 1 - \exp\left(\frac{z-\delta}{\lambda}\right)$.

8.5 Gamma Distribution

The exponential distribution has a distinctive shape always sloping down which makes the very lowest values the most probable. It is not difficult to see that there are distributions which do decline sharply in frequency as we go for the high values. But this still does not make the lowest values the most probable. Relative frequency starts from 0 for the lowest values and increases to some intermediate value and then after peaking, die away exponentially. The shape we are considering is shown in the figure:



8.5.1 Examples:

1. The overall time it takes three successive tasks on an car as it is being built on an assembly conveyor belt.
2. The age of mothers when they give birth to their first child.
3. The duration of a telephone conversation.

How can we modify the exponential density function to cater for the shape required? It seems sensible to pull down the high end of the exponential distribution at the origin by multiplying it by some power of x:

$$f(x) = Ax^{\beta} e^{-\frac{x}{\lambda}} \text{ for } x \ge 0 \,,$$

f is clearly positive but we need to choose A to ensure that the area under f is 1, and also we need to have the flexibility to make the "scaling" in x for the power part be allowed to be different from that in the exponential part. Firstly we note that from the theory of the gamma function in real function theory we have:

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx$$

which is equal to $(\alpha - 1)!$ for integer α .

This equation could be used for defining the gamma function for values of α which are not integers. This integral equation does the job perfectly because it's used to define a real valued function of α .

This together with the point made earlier give us the basis for defining the gamma distribution.

8.5.2 Definition

Random variable X is said to have the gamma distribution with parameters $\alpha > 0$, and $\lambda > 0$ if

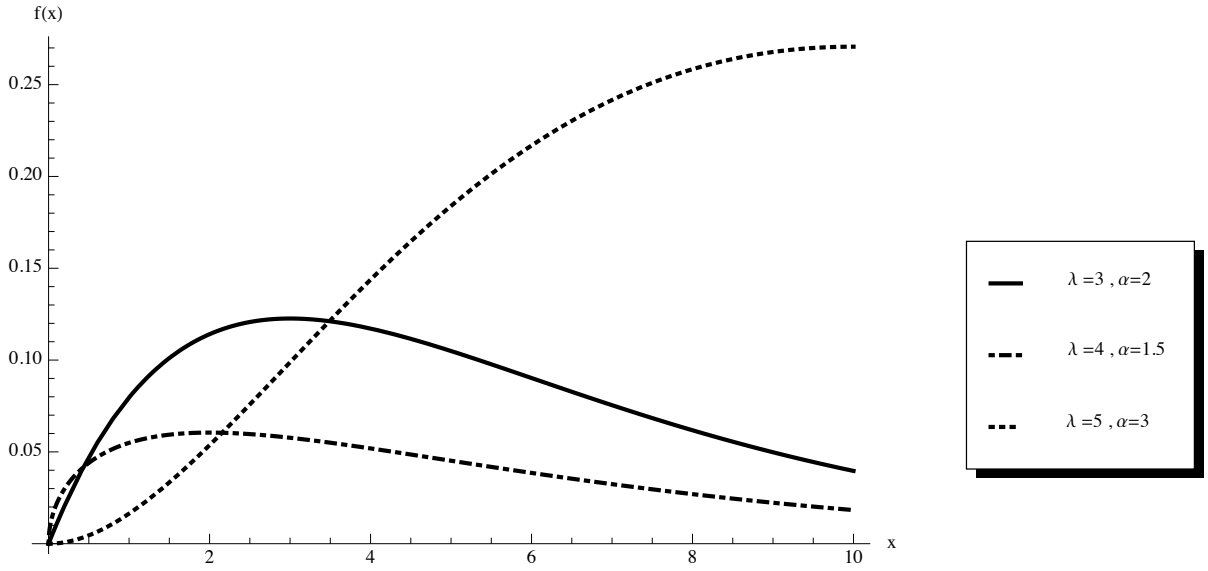
$$\mathbb{P}[\omega : s < X(\omega) < t] = \frac{1}{\Gamma(\alpha)} \int_s^t \frac{1}{\lambda} \left(\frac{x}{\lambda} \right)^{\alpha-1} e^{-\frac{x}{\lambda}} dx$$

for $0 \le s \le t < \infty$

The corresponding probability distribution function is given by:

$$F(z) = \mathbb{P}[\omega : 0 \le X(\omega) < z] = \frac{1}{\Gamma(\alpha)} \int_0^z \frac{1}{\lambda} \left(\frac{x}{\lambda} \right)^{\alpha-1} e^{-\frac{x}{\lambda}} dx$$

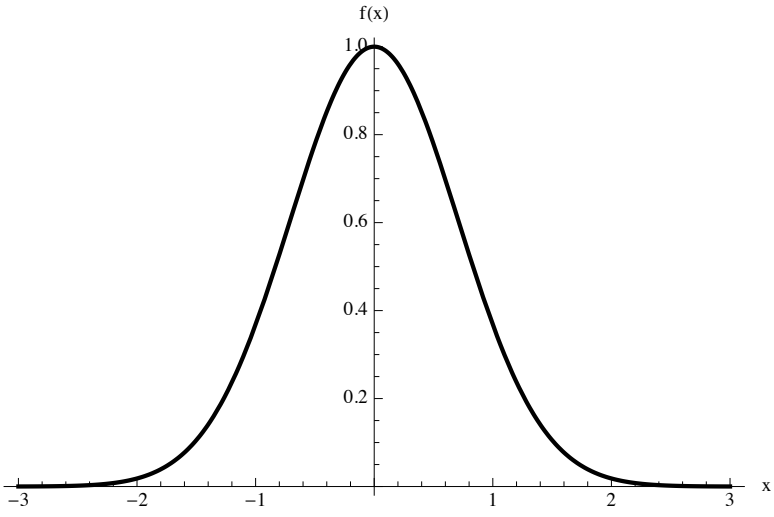
Below we can observe the density function $f(x) = \frac{1}{\Gamma(\alpha)} \frac{1}{\lambda} \left(\frac{x}{\lambda} \right)^{\alpha-1} e^{-\frac{x}{\lambda}}$ for different values of α and λ :



Calculating the probabilities for given parameters and limits s and t has to be done numerically in general. There are a few cases when the integral can be worked out as a closed formula using integration by parts. But for fractional values of α this cannot be done. There are many software packages available which offer numerical integration routines. The integrand is smooth and numerical integration works out well with it.

8.6 Normal Distribution

The gamma distribution displays a lack of symmetry which in a number of applications would not seem natural. Gauss introduced a symmetrical distribution when he was developing a theory of errors. This distribution would be centered on zero. A function which displays the right shape is $f(x) = \exp(-x^2)$:



This function has the famous "bell" shape. The function is everywhere positive. However we have first to make sure that it caps an area of 1.
For reasons to become clearer later it is more appropriate to consider $\exp(-x^2/2)$. Furthermore from standard theory we know that:

$$\int_{-\infty}^{\infty} e^{\frac{-x^2}{2}}\, dx = \sqrt{2\pi}$$

Thus dividing the said integral by $\sqrt{2\pi}$ gives us an area equal to 1. For wider applicability we would need to have the ability to shift the graph by a parameter μ . This will be the value around which the distribution will be centered. Furthermore we shall require to be able to control the width of the graph. To do this we need to do the transformation:

$$x \to \frac{x-\mu}{\sigma}$$

To preserve the area we need to divide by σ. This gives us the normal distribution with shift parameter μ (mean) and dispersion parameter (standard deviation) σ.

8.6.1 Definition

X is said to be normally distributed with parameters μ and $\sigma > 0$, if

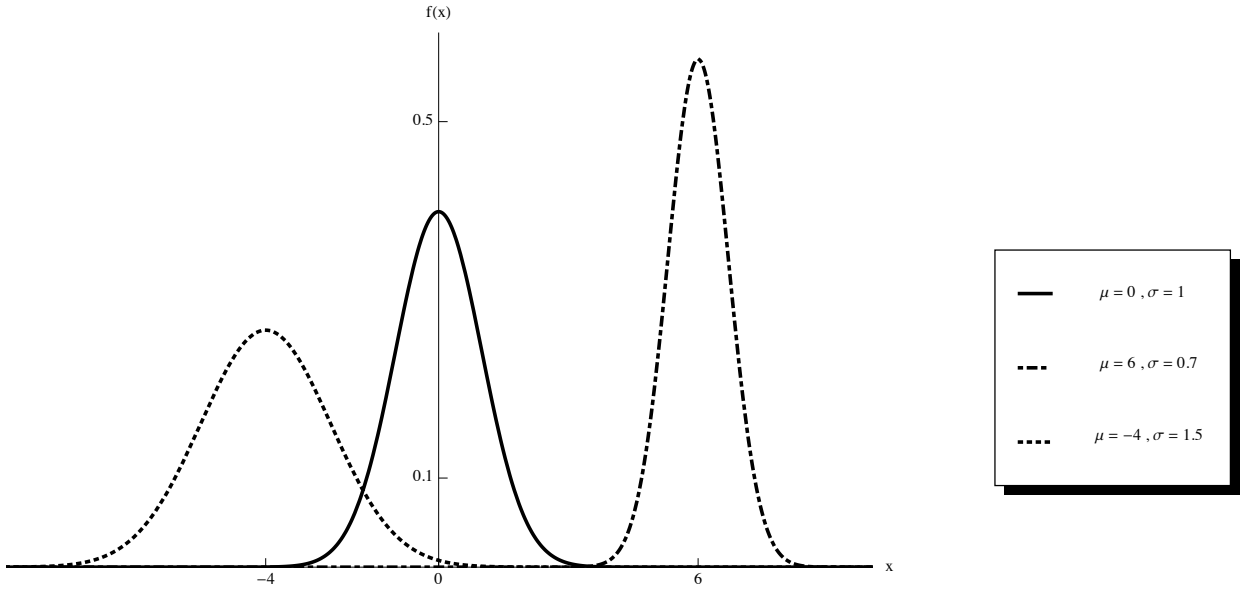
$$\mathbb{P}[\omega : s < X(\omega) < t] = \frac{1}{\sigma\sqrt{2\pi}} \int_s^t e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx ,$$

for $-\infty < s \le t < \infty$. The corresponding probability distribution function is

$$F(z) = \mathbb{P}[\omega : 0 \le X(\omega) < z] = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx .$$

We note that the normal density function is not a perfect integral. It has to be evaluated numerically. Traditionally the probabilities under the normal curve were computed using tables. However nowadays it is easier to use numerical software packages like Excel, Matlab or Mathemetica to evaluate the above integrals.

Below we can observe the density function $f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ for different values of μ and σ :



In the above diagram we notice that that curve with the solid black line has parameters $\mu=0$ and $\sigma=1$. A normal distribution with these parameters is called the *standard normal distribution.*

In fact it is only necessary to work out half of it because symmetry considerations can help get probabilities on both sides in terms of probability on one side:

$$\frac{1}{\sqrt{2\pi}}\int_{-s}^{t} e^{-\frac{x^2}{2}}\,dx=\frac{1}{\sqrt{2\pi}}\int_{0}^{s} e^{-\frac{x^2}{2}}\,dx+\frac{1}{\sqrt{2\pi}}\int_{0}^{t} e^{-\frac{x^2}{2}}\,dx\,.$$

Furthermore if we have parameters μ and σ we can use the result:

$$\frac{1}{\sigma\sqrt{2\pi}}\int_{s}^{t} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx=\frac{1}{\sigma\sqrt{2\pi}}\int_{(s-\mu)/\sigma}^{(t-\mu)/\sigma} e^{-\frac{x^2}{2}}\,dx\,,$$

to see how we can always translate normal probabilities into probabilities for the standard normal.

In all the more popular statistical packages it is possible to obtain directly normal probabilities so that the need for tables giving probabilities under the standard normal distribution has been rendered obsolete.

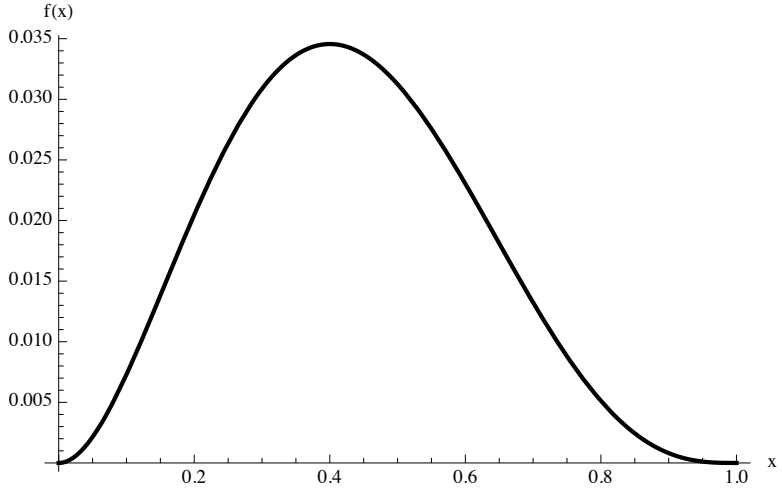
8.7 Beta Distribution

The gamma distribution has the right shape for a substantial number of data. However theoretically it should not be used for data whose range is limited to some bounded interval.

8.7.1 Examples:

1. Relative humidities at noon for different days.
2. Strictly speaking a woman's age at childbirth is biologically limited to the age range 13-50.
3. Working efficiencies of engines of the same design.

The idea is thus to limit the distribution to an interval range. We fix this interval to [0,1] and the shape we require is given by functions of the type:



The integral of functions like $f(x) = x^{\alpha-1}(1-x)^{\beta-1}$ give what are called beta functions:

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,dx\,.$$

It can be shown that $B(\alpha,\beta) = \dfrac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$.

8.7.2 Definition

A random variable X is said to be beta-distributed with parameters α and β if for $0 \le s \le t \le 1$:

$$\mathbb{P}[\omega : s < X(\omega) < t] = \frac{1}{B(\alpha,\beta)} \int_s^t x^{\alpha-1}(1-x)^{\beta-1}\,dx$$

The corresponding probability distribution function is given by:

$$F(z) = \mathbb{P}[\omega : 0 \le X(\omega) < z] = \frac{1}{B(\alpha,\beta)} \int_0^z x^{\alpha-1}(1-x)^{\beta-1}\,dx$$

for $0 \le z \le 1$.

For the integrals of the functions as given above, no explicit formulas involving common functions are known. So to obtain probabilities under these curves one has to use numerical integration. Many statistical packages include routines which compute the area under the beta distribution.

8.8 Using software Packages

In the previous section we were confronted with integrals which cannot be evaluated analytically. It is therefore necessary to use a software package to help us evaluate the integral numerically. In this section we shall use Excel, Matlab, Mathematica and R to evaluate such integrals.

8.8.1 Gamma Distribution

Let us assume that we would like to compute the following:

(i) $\mathbb{P}[\omega : 0 < X(\omega) < 5] = \int_0^5 \frac{1}{\lambda}\left(\frac{x}{\lambda}\right)^{\alpha-1} e^{-\frac{x}{\lambda}} dx$,

(ii) $\mathbb{P}[\omega : 7 < X(\omega) < 10] = \int_7^{10} \frac{1}{\lambda}\left(\frac{x}{\lambda}\right)^{\alpha-1} e^{-\frac{x}{\lambda}} dx$,

in both cases we shall be assuming that $\alpha = 1.2$ and $\lambda = 2.3$.

Excel

(i) =GAMMADIST(5, 1.2, 2.3, TRUE)
(ii) =GAMMADIST(10,1.2,2.3,TRUE)-GAMMADIST(7,1.2,2.3,TRUE)

MATLAB

(i) gamcdf(5, 1.2, 2.3)
(ii) gamcdf(10, 1.2, 2.3)-gamcdf(7, 1.2, 2.3)

Mathematica

(i) CDF[GammaDistribution[1.2, 2.3], 5]
(ii) CDF[GammaDistribution[1.2, 2.3], 10] - CDF[GammaDistribution[1.2, 2.3], 7]

R

(i) pgamma(5,1.2,1/2.3)
(i) pgamma(10,1.2,1/2.3)-pgamma(7,1.2,1/2.3)

8.8.2 Normal Distribution

Let us assume that we would like to compute the following (where $\mu = 2$, $\sigma = 0.4$):

(i) $\mathbb{P}[\omega : 0 < X(\omega) < 7.4] = \frac{1}{\sigma\sqrt{2\pi}} \int_0^{7.4} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$

(ii) $\mathbb{P}[\omega : -2.59 < X(\omega) < 3.8] = \frac{1}{\sigma\sqrt{2\pi}} \int_{-2.59}^{3.8} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$

Excel

(i) =NORMDIST(7.4,2,0.4,TRUE)-NORMDIST(0,2,0.4,TRUE),
(ii) =NORMDIST(3.8,2,0.4,TRUE)-NORMDIST(-2.59,2,0.4,TRUE).

MATLAB

(i) normcdf(7.4, 2, 0.4) - normcdf(0, 2, 0.4),
(ii)normcdf(3.8, 2, 0.4) - normcdf(-2.59, 2, 0.4).

Mathematica

(i) CDF[NormalDistribution[2, 0.4], 7.4] - CDF[NormalDistribution[2, 0.4], 0],
(ii) CDF[NormalDistribution[2, 0.4], 3.8] - CDF[NormalDistribution[2, 0.4], -2.59].

R

(i) pnorm(7.4,2,0.4)-pnorm(0,2,0.4),
(i) pnorm(3.8,2,0.4)-pnorm(-2.59,2,0.4).

Similar (though suitably modified) commands are used for the beta distribution.

8.9 Expectations of Continuous Random Variables

Continuous random variable have ranges which are uncountable. Therefore we cannot average their values by using summations. However, just as in the case when we calculating probabilities, we can try to approximate the average by finite sums.

Suppose we divide the range of a continuous random variable X, which we assume to be an interval R and having probability density function $f(x)$, into n strips of width δ, the divisions being demarcated by the set of ordered $n+1$ points $\{ x_0, x_1, \ldots , x_n\}$. Then when the value of X lies inside the interval $[x_i, x_{i+1}]$ we can say that approximately the value of X is x_i and this with probability $f(x_i)\ \delta x$. Thus the average of X can be approximated by:

$$\sum_{i=1}^{n} x_i f(x_i)\delta x$$

The x_i gives us the value of the random variable, the rest weights that value by its probability.

Now by increasing the number of divisions, and hence forcing the width of the strips δx to tend to 0, we get the usual Riemann integral:

$$\int_R xf(x)\,dx$$

8.9.1 Definition

Given a random variable X with probability density function f and range R, we define its expectation (mean, average) E[X] , μ, whenever it exists by:

$$E[X] = \mu = \int_R xf(x)\,dx$$

One observation we should make is that not all random variables have average. The more common ones do. But not each every one does.

Let us state a few obvious results about the expectation:

8.9.2 Theorem

Let X be a continuous random variable with density function f, mean μ and range R. Then we have:

1. $\mathbb{E}[X + a] = \mathbb{E}[X] + a$
2. $\mathbb{E}[X - \mu] = 0$
3. $\mathbb{E}[aX] = a\mathbb{E}[X]$

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

The proofs of these elementary results follows easily from how we took our approximating sequence.

A generalization of the above definition given us the definition of the m^{th} moment of a random variable:

$$\mathbb{E}[X^m] = \int_R x^m f(x)\,dx$$

8.9.3 Expectation of Uniform, Exponential and Gamma Distributions

We apply the definition of expectation to the distribution we described earlier:

Let X be **uniformly distributed** on $[\alpha,\beta]$. Then:

$$\begin{aligned}\mathbb{E}[X] &= \int_\alpha^\beta \frac{x}{\beta-\alpha}\,dx \\ &= \frac{\beta^2-\alpha^2}{2(\beta-\alpha)} \\ &= \frac{1}{2}(\beta+\alpha)\end{aligned}$$

Let X be **exponentially distributed** with parameter λ. Then

$$\begin{aligned}\mathbb{E}[X] &= \int_0^\infty \frac{x}{\lambda} e^{-\frac{x}{\lambda}}\,dx \\ &= 0 - \int_0^\infty e^{-\frac{x}{\lambda}}\,dx \\ &= \lambda\end{aligned}$$

where we use integration by parts.

Let X be **gamma distributed** with parameters α, λ.
Then:

$$\begin{aligned}\mathbb{E}[X] &= \frac{1}{\Gamma(\alpha)} \int_0^\infty \frac{x}{\lambda}\left(\frac{x}{\lambda}\right)^{\alpha-1} e^{-\frac{x}{\lambda}}\,dx \\ &= \frac{\lambda}{\Gamma(\alpha)} \int_0^\infty \frac{1}{\lambda}\left(\frac{x}{\lambda}\right)^{\alpha} e^{-\frac{x}{\lambda}}\,dx \\ &= \frac{\lambda\Gamma(\alpha+1)}{\Gamma(\alpha)} \\ &= \alpha\lambda\end{aligned}$$

where we are using the result that $\Gamma[\alpha+1] = \alpha\Gamma[\alpha]$

8.9.4 Expectation of Normal and Beta Distributions

Let X be **normally distributed** with parameters μ and σ^2 , then

where we are using the change of variable $u = (\ x$ - $\mu)$ and using the fact that the integral of an odd function over all of $\mathbb{R}$ is 0. Thus $\mathbb{E}[X] = \mu$.

$$\begin{aligned}
\mathbb{E}[X-\mu] &= \int_{-\infty}^{\infty} \frac{x-\mu}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx \\
&= \int_{-\infty}^{\infty} \frac{u}{\sqrt{2\pi}\sigma} e^{\frac{-u^2}{2\sigma^2}} du \\
&= 0
\end{aligned}$$

Let X be **beta distributed** with parameters α and β , then

$$\begin{aligned}
\mathbb{E}[X] &= \int_0^1 \frac{x x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} dx \\
&= \frac{1}{B(\alpha,\beta)} \int_0^1 x^{\alpha}(1-x)^{\beta-1} dx \\
&= \frac{B(\alpha+1,\beta)}{B(\alpha,\beta)} \\
&= \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \\
&= \frac{\alpha}{\alpha+\beta}
\end{aligned}$$

We are here using the result that:

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

8.10 Variances of Continuous Random Variables

We are now in a position to define the variance for continuous distributions. We remember that the idea of variance is that of averaging deviation squared from the mean. That means we take the value x of our random variable, subtract the mean μ i.e. $(x-\mu)$, square it $(x-\mu)^2$, and average it over all possible values x i.e multiply by f(x)δx and integrate.

8.10.1 Definition

Given a random variable X with probability density function f and range $\mathbb{R}$, we define its variance σ^2, whenever it exists by:

$$\sigma^2 = \mathbb{V}\text{ar}[X] = \mathbb{E}[(X-\mu)^2] = \int_{\mathbb{R}} (X-\mu)^2 f(x)dx$$

Its square root, σ, is called the standard deviation.

The reason why we square the deviations is the same as for discrete random variables: if we average deviations from the mean we get 0. To recuperate the original units we square root.

We interpret the standard deviation as a measure of the variability of a random variable. Of two random variables X and Y, the one with the biggest variance is the one which varies most on average.

8.10.2 Theorem

Let X be a continuous random variable with density function f, mean μ and standard deviation σ and range $\mathbb{R}$, then we have:

$\mathbb{V}ar[X+\beta] = \mathbb{V}ar[X]$ and $\mathbb{V}ar[\alpha X] = \alpha^2 \mathbb{V}ar[X]$

for all $\alpha, \beta \in \mathbb{R}$

In particular $\mathbb{V}ar\left[\dfrac{X-\mu}{\sigma}\right] = 1$ and $\mathbb{E}\left[\dfrac{X-\mu}{\sigma}\right] = 0$.

Finally we have: $\mathbb{V}ar[X] = \mathbb{E}\left[X^2\right] - \left(\mathbb{E}[X]\right)^2$

Proof

Using the properties of expectations we have that: $\mathbb{E}[\alpha X + \beta] = \alpha\mathbb{E}[X] + \beta = \alpha\mu + \beta$. Then:

$$\begin{aligned}
\mathbb{V}ar[\alpha X+\beta] &= \int [(\alpha x+\beta) - (\alpha\mu+\beta)]^2 f(x)dx \\
&= \int [\alpha x - \alpha\mu]^2 f(x)dx \\
&= \alpha^2 \int (x-\mu)^2 f(x)dx \\
&= \alpha^2 \mathbb{V}ar[X]
\end{aligned}$$

Next,

$$\begin{aligned}
\mathbb{V}ar[X] &= \int (x-\mu)^2 f(x)dx \\
&= \int (x^2 - 2\mu x - \mu^2) f(x)dx \\
&= \int x^2 f(x)dx - 2\mu \int x f(x)dx + \mu^2 \int f(x)dx \\
&= \mathbb{E}[X^2] - 2\mu^2 + \mu^2 \\
&= \mathbb{E}[X^2] - \mu^2 \\
&= \mathbb{E}[X^2] - (\mathbb{E}[X])^2
\end{aligned}$$

8.10.3 Variance of the Uniform, Gamma and Exponential Distributions

Let X be **uniformly distributed** on $[\alpha,\beta]$. Then we have:

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

Furthermore from section 8.9.2 we have that:

$$\mathbb{E}[X] = \frac{(\beta-\alpha)^2}{4}$$

Thus using $\mathbb{V}ar[X] = \mathbb{E}[X^2] - \big(\mathbb{E}[X]\big)^2$ we obtain:

$$\mathbb{V}ar[X] = \frac{\alpha^2 - 2\alpha\beta + \beta^2}{12}$$

Let X be **gamma distributed** with parameters α,λ.
From section 8.9.2 we know that formula for $\mathbb{E}[X]$. Furthermore:

$$\begin{aligned}
\mathbb{E}[X^2] &= \int_0^\infty \frac{x^2}{\Gamma[\alpha]\lambda} \left(\frac{x}{\lambda}\right)^{\alpha-1} e^{\frac{-x}{\lambda}} dx \\
&= \frac{\lambda^2}{\Gamma[\alpha]} \int_0^\infty \frac{1}{\lambda} \left(\frac{x}{\lambda}\right)^{\alpha+1} e^{\frac{-x}{\lambda}} dx \\
&= \frac{\lambda^2 \Gamma[\alpha+2]}{\Gamma[\alpha]} \\
&= \alpha(\alpha+1)\lambda^2
\end{aligned}$$

Hence,

$$\mathbb{V}ar[X] = \alpha(\alpha+1)\lambda^2 - (\alpha\lambda)^2 = \alpha(\lambda)^2$$

Let X have an **exponential distribution** with parameter λ.

From section 8.9.2 we know that $\mathbb{E}[X]=\lambda$. Furthermore,

$$\mathbb{E}[X^2]=\int_0^\infty \frac{x^2}{\lambda}e^{-\frac{x}{\lambda}}dx=2\lambda^2$$

Thus, $\mathbb{V}ar[X]=\lambda^2$.

8.10.4 Variance of the Normal and Beta Distribution

Let X be **normally distributed** with parameters μ,σ^2. Then its variance can be written directly as:

$$\begin{aligned}\mathbb{V}\text{ar}[X]&=\mathbb{E}[(x-\mu)^2]\\&=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}(x-\mu)^2\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)dx\end{aligned}$$

Next, we change the variable of integration by letting $y=x-\mu$. This gives us:

$$\begin{aligned}\mathbb{V}\text{ar}[X]&=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}y^2\exp\left(-\frac{y^2}{2\sigma^2}\right)dx\\&=\frac{-\sigma^2}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}y\left[-\frac{y}{\sigma^2}\exp\left(-\frac{y^2}{2\sigma^2}\right)\right]dx\\&=\frac{-\sigma}{\sqrt{2\pi}}\int_{-\infty}^{\infty}y\left[-\frac{y}{\sigma^2}\exp\left(-\frac{y^2}{2\sigma^2}\right)\right]dx\end{aligned}$$

Using integration *by parts*, we let $u=y$ and $v'=-\frac{y}{\sigma^2}\exp\left(-\frac{y^2}{2\sigma^2}\right)$. These in turn yield:

$u'=1$ and $v=\exp\left(-\frac{y^2}{2\sigma^2}\right)$.

$$\begin{aligned}\text{Thus, }\mathbb{V}\text{ar}[X]&=\frac{-\sigma}{\sqrt{2\pi}}\left[y\exp\left(-\frac{y^2}{2\sigma^2}\right)\right]_{-\infty}^{\infty}+\frac{\sigma}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp\left(-\frac{y^2}{2\sigma^2}\right)dy\\&=0+\frac{\sigma}{\sqrt{2\pi}}\left(\sigma\sqrt{2\pi}\right)\\&=\sigma^2\end{aligned}$$

Note

In the above the following standard result was used:

$$\int_{-\infty}^{\infty}\exp\left(-\frac{y^2}{2\sigma^2}\right)dy=\sigma\sqrt{2\pi}$$

Let X be **beta distributed** with parameters α,β . Then its variance can be can be found as follows:

$$\begin{aligned}
\mathbb{E}[X^2] &= \int_0^1 \frac{x^2 x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}dx \\
&= \int_0^1 \frac{x^{\alpha+1}(1-x)^{\beta-1}}{B(\alpha,\beta)}dx \\
&= \frac{B(\alpha+2,\beta)}{B(\alpha,\beta)} \\
&= \frac{\Gamma(\alpha+2)\Gamma(\beta)}{\Gamma(\alpha+\beta+2)}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \\
&= \frac{(\alpha+1)\alpha}{(\alpha+\beta+1)(\alpha+\beta)}
\end{aligned}$$

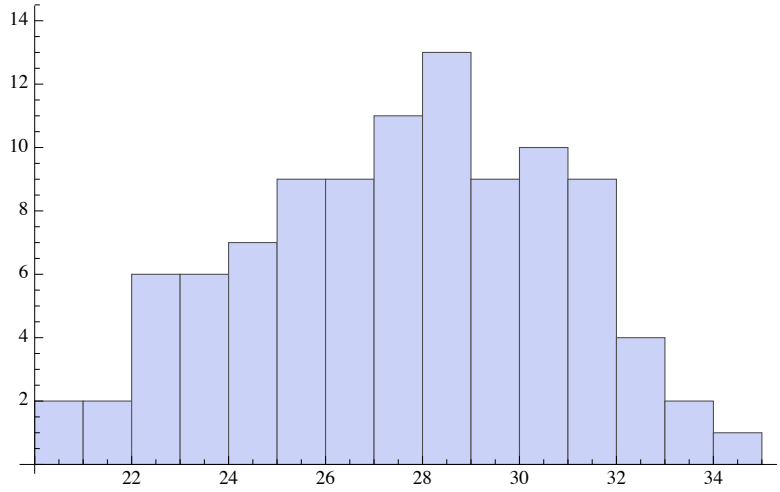
Hence, $\mathbb{V}ar[X] = \dfrac{(\alpha+1)\alpha}{(\alpha+\beta+1)(\alpha+\beta)} - \dfrac{\alpha^2}{(\alpha+\beta)^2} = \dfrac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$

8.10.4 Example

The male of a certain species of insects is eaten up by the female after its first and only opportunity for sexual intercourse. This happens when the male is aged between 15 months up to 35 months. The age when male insects are eaten up is believed to be beta distributed. Use the following set of compiled 100 ages to fit a suitably shifted and magnified beta distribution:

33.4	27.8	27.2	28.4	28.8	30.6	27.4	30.0	31.4	21.1
33.2	32.0	28.8	28.6	29.2	28.4	28.0	31.0	31.8	20.5
22.5	32.1	31.8	30.2	29.8	28.8	28.4	32.0	23.4	34.2
22.4	24.2	31.6	30.8	29.5	28.0	29.2	25.4	24.8	23.9
23.4	24.1	23.4	25.6	26.2	27.2	30.0	25.2	31.1	21.1
24.6	24.4	22.6	26.0	26.6	28.2	30.2	30.6	32.6	27.5
23.6	25.6	25.2	26.8	27.8	29.6	26.6	30.4	22.8	29.8
25.2	25.6	26.2	27.8	27.7	29.8	26.4	31.6	23.2	31.8
26.4	26.6	25.2	27.4	28.4	30.0	29.2	24.8	22.1	20.2
25.8	28.0	27.8	28.8	30.1	27.0	29.4	24.2	22.3	31.1

We start analyzing this data set by simply drawing histogram which is shown below:



Next we can compute some basic results data set which gives us:

$$\text{mean} = 27.495$$

$$\text{variance} = 10.3193$$

To fit the beta distribution we shall have to shrink our interval from [15,35] to [0,1] to do so we use the following transformation:

$$X \to \frac{X-15}{(35-15)}$$

This transformation will give us $\mu \simeq 0.62475$ and $\sigma^2 \simeq 0.0257982$.

From the previous sections we have shown that

$$\mu = \frac{\alpha}{\alpha+\beta} \text{ and } \sigma^2 = \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$$

Thus to estimate the values of α and β we solve simultaneously the following set of equations:

$$\frac{\alpha}{\alpha+\beta} = 0.62475 \text{ and } \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2} = 0.0257982$$

which gives us: $\alpha \simeq 5.05$ and $\beta \simeq 3.03$.

Furthermore our distribution is given by:

$$\begin{aligned}
\mathbb{P}[s < \text{Age} < t] &= \frac{1}{105}\int_{(s-15)/20}^{(t-15)/20} x^4(1-x)^2 dx \\
&= \frac{1}{105}\left[\frac{x^5}{5} - \frac{x^6}{3} + \frac{x^7}{7}\right]_{(s-15)/20}^{(t-15)/20} \\
&= \frac{1}{105\times 20^5}\left[\frac{(t-15)^5 - (t-15)^5}{5} - \frac{(t-15)^6 - (s-25)^6}{3\times 20} + \frac{(t-15)^7 - (s-25)^7}{7\times 20^2}\right]
\end{aligned}$$

9. Convergence in a Probability Setting

9.1 Introduction

Historically the quest for classical limit theorems in probability was initiated by the need to compute probabilities for the binomial distribution involving large numbers. Historically the binomial distribution served as an important referent. It provided a powerful formula for computing probabilities. But also it provided the first limit theorem – the law of large numbers which was to help legitimize the notion of applying probability computations in practical situation. As more Limit theorems were discovered and used, their role became extremely important, and in many ways crucial, in the development of the theory of stochastic processes and sampling theory.

It turns out that in probability there are many situations where limits need to be taken, largely when finding probabilities of infinite sets or when considering random variables which are being approximated by other random variables which are in some sense simpler than the ones at hand. In fact there is more than one type of limits which one meets with in probability and the differences between these type is theoretically deep.

9.2 Chebychev's Theorem

This result comes in handy when we want to use upper bounds for the probability of sets on which a random variable X becomes large.

9.2.1 Theorem

Let X be a nonnegative random variable defined on a probability space $(\Omega,\mathcal{F},\mathbb{P})$. Then given any positive number $k>0$, we have:

$$\mathbb{P}[\omega : X(\omega) > k] \leq \frac{E[X]}{k}$$

Proof

We assume that X is a continuous random variable with density function f :

$$\begin{aligned}
\mathbb{E}[X] &= \int_0^k xf(x)dx + \int_k^\infty xf(x)dx \\
&\geq \int_k^\infty xf(x)dx \\
&\geq \int_k^\infty kf(x)dx \\
&= k\int_k^\infty f(x)dx \\
&= kP[\omega : X(\omega) \geq k]
\end{aligned}$$

An entirely similar argument holds for a discrete random variable where summation is used instead of integration.

The above inequality is called Markov's inequality sometimes. It's a generalization of Chebychev's original form of inequality in effect. There are many equivalent forms which list as corollary.

9.2.2 Corollary

Given any rv X with finite second moment, we have:

1. $\mathbb{P}\left[\left|X\right| \geq k\right] \leq \mathbb{E}[X^2] / k^2$,

2. $\mathbb{P}\left[\left|X - \mu\right| \geq k\right] \leq \sigma^2 / k^2$,

Proof

If we apply the theorem above to |X| which is clearly non-negative we get result 1 immediately.

If we now apply the theorem to the random variable |X - μ|, we again have a non-negative random variable and furthermore $\mathbb{E}\left[\left|X - \mu\right|^2\right] = \sigma^2$.

9.3 Law of Large Numbers

This result is due to James Bernouilli and had a profound influence from the time when it was first published in *Ars Conjectandi* (1713). It managed to arouse much interest because of its "philosophical" implications. There were many attempt to strengthen the result in other direction. In fact it is now known as the weak law of large numbers. We shall first state in roughly its original form.

9.3.1 Theorem

Let A be an event which occurs with probability p. Let n successive independent experiments involving A be conducted, at the end of each of which we register the occurrence or otherwise of A. Let S_n be the total number of occurences. Given any $\epsilon > 0$, as $n \to \infty$ we have: $\mathbb{P}\left[\left|S_n / n - p\right| \geq \epsilon\right] \to 0.$

In other words the average number of successes approaches the true probability p. In practice this result somehow says that asymptotically we can always approximate to within any required limits, provided we do so within the confines allowed by probability.

We shall state and prove a generalization :

9.3.2 Theorem

Let X_n be a sequence of independent, identically distributed random variables with finite mean μ and finite variance σ^2. Then for all fixed $\varepsilon > 0$, we have

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

Proof

Clearly $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \mu$.

Also the independence assumption of X_i's implies that

$$\mathbb{V}ar\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{\sigma^2}{n}.$$

Take any $\delta > 0$. Then for any $n > \sigma^2 / \delta\epsilon^2$, applying the corollary to Chebychev's theorem to $\frac{1}{n}\sum_{i=1}^{n} X_i$ we have that:

$$\mathbb{P}\left[\left| \frac{1}{n}\sum_{i=1}^{n} X_i - \mu \right| > \varepsilon \right] \leq \frac{\sigma^2}{n\varepsilon^2} < \delta$$

This shows clearly that as n tends to ∞, the probability tends to 0.

9.4 Computing Binomial Probabilities

The Law of Large Numbers tells us that repeating an experiment a sufficient number of times gets us as close to the true probability of success via our proportion of successes as we want with high probability.

This result somehow reassures us that repetition is not useless. And it also creates an interest in being able to compute probabilities involved in situations where independent trials are effected in succession. This is familiar territory involving the binomial distribution and we now come to it with a need to compute expressions of the form ${}^{n}C_{k}p^{k}(1\text{-}p)^{n-k}$ with accuracy.

A moment's thought will reveal immediately that it is factorials which can cause problems in such expression:

$${}^{n}C_{k} = \frac{n!}{(n-k)!k!}$$

It must be appreciated that the origin of the classical limits theorems of probability has to be located in the need for good approximations to standard distributions, especially the ubiquitous binomial distribution. The following theorem shows the original interest in the Poisson distribution.

9.5 The Poisson Approximation to Binomial Probabilities

9.5.1 Theorem

Let X_n be a sequence of binomially distributed random variables with parameter n, λ/n. Then the distributions of X_n converge to the Poisson distribution with parameter $\lambda > 0$.

Proof:

We start by expanding the binomial coefficients:

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

Now as $n \to \infty$ we have:

$$ n(n-1)...(n-k+1)\left(\frac{\lambda}{n}\right)^k = \left(1-0\right)\left(1-\frac{1}{n}\right)...1-\left(\frac{(k-1)}{n}\right)\lambda^{\,k} \to \lambda^{\,k} $$

while

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

for which $\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{-k} = 1$ and

$$ \lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{\frac{-n}{\lambda}} = $$

$$ = \lim_{n\to\infty} 1+\frac{n}{\lambda}\frac{\lambda}{n}+\frac{n}{2!\lambda}\left(\frac{n}{\lambda}-1\right)\left(\frac{\lambda}{n}\right)^2+...\frac{n}{k!\lambda}\left(\frac{n}{\lambda}-1\right)..\left(\frac{\lambda}{n}-k+1\right)\left(\frac{\lambda}{n}\right)^k $$

$$ = \lim_{n\to\infty} 1+1+\frac{1}{2!}\left(1-\frac{\lambda}{n}\right)+...\frac{1}{k!}\left(1-\frac{\lambda}{n}\right)..\left(1-\frac{(k-1)\lambda}{n}\right)+....= e $$

Putting everything together we obtain:

$$ \lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{n-k} = \lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{-k}\left(\left(1-\frac{\lambda}{n}\right)^{\frac{-n}{\lambda}}\right)^{-\lambda} = e^{-\lambda} $$

and $\lim_{n\to\infty} {}^nC_k\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k} = \frac{\lambda^k e^{-\lambda}}{k!}$

We next show how this theorem can be used to obtain approximations of binomial probabilities by using the Poisson distribution.

Suppose we need to compute the first 16 values for the binomial distribution with n = 200 and p = 0.01. Comparing the parameters we had in the theorem, we see that we need to take λ = np = 200 × 0.01 =2.
Then we use the approximation:

$$^{200}C_k 0.01^k 0.99^{200-k} = \frac{2^k e^{-2}}{k!}$$

The table below compares the two probabilities:

k	B(k,200,0.01)	Poisson Approximation
0	0.133979674857962	0.135335283236613
1	0.270666009814064	0.270670566473225
2	0.272033009863630	0.270670566473225
3	0.181355339909087	0.180447044315484
4	0.090219701924470	0.090223522157742
5	0.035723356721608	0.036089408863097
6	0.011727364580326	0.012029802954366
7	0.003282985178331	0.003437086558390
8	0.000800020378053	0.000859271639598
9	0.000172394963621	0.000190949253244
10	0.000033260038436	0.000038189850649
11	0.000005802945182	0.000006943609209
12	0.000000923195824	0.000001157268201
13	0.000000134856888	0.000000178041262
14	0.000000018194977	0.000000025434466
15	0.000000002278967	0.000000003391262
16	0.000000000266167	0.000000000423908
17	0.000000000029100	0.000000000049872
18	0.000000000002988	0.000000000005541
19	0.000000000000289	0.000000000000583
20	0.000000000000026	0.000000000000058

It shows how good the approximation really is.

9.6 De Moivre-Laplace Theorem

Another classical limit theorem, whose complicated proof developed over a number of years, concerns the approximation of binomial probabilities using the normal density function. This might seem unnatural to us because we make such a fundamental difference between discrete and continuous random variables. However, the use of this theorem was initially mostly computational but not exclusively so. Eventually it was instrumental in establishing a central role for the normal distribution in the theory of statistics and probability.

We shall state the theorem but not give a proof of the result which is beyond the scope of this course.

9.6.1 Definition

Given a real number x we denote by $\lfloor x \rfloor$ the greatest integer which is smaller or equal to x. By $\lceil x \rceil$ we denote the smallest integer which is greater or equal to x.

9.6.2 Theorem

For integers $k \leq n$, let

$$P_n(a,b) = \sum_{k=\lfloor np+a\sqrt{np(1-p)} \rfloor}^{\lfloor np+b\sqrt{np(1-p)} \rfloor} {}^nC_k p^k (1-p)^{n-k}$$

Then,

$$\lim_{n\to\infty} \sup_{-\infty\leq a<b\leq\infty} \left| P_n(a,b] - \frac{1}{\sqrt{2\pi}} \int_a^b e^{\frac{-x^2}{2}} dx \right| = 0$$

We give a brief description of what the result says:
Choose a probability p,
Take any positive integer n.

Choose a $<$ b and shift the interval $\sqrt{np(1-p)}(a,b]$ by np .

From within this interval pick out all integers k to form the ${}^nC_k\, p^k\, (1\text{-}p)^{n\text{-}k}$ ‘s and add all these probabilities.

Work out integral $\frac{1}{2\pi}\int_a^b \exp(-x^2/2)\,dx$.

Vary a and b over the real line and get the biggest possible difference between the integral and the sum of probabilities.

As n tends to infinity this biggest possible difference tends to 0.
The last two comments show that the convergence we have is uniform over all compact subsets (closed and bounded subsets) of $\mathbb{R}$.

We shall give a brief sketch of how the proof works:

Using Stirling's formula, expansions involving natural logs and Taylor's theorem one obtains the asymptotic formula

$$P_n(k) = {}^nC_k p^k (1-p)^{n-k} \sim \frac{1}{\sqrt{2\pi np(1-p)}} e^{-(k-np)^2/(2np(1-p))} .$$

Interval (a,b] is taken into consideration picking integers therein expressible as

$$np + x\sqrt{np(1-p)} .$$

For integer k, t_k is defined by $k = np + t_k\sqrt{np(1-p)}$, increment

$$\Delta t_k = t_{k+1} - t_k = \frac{1}{\sqrt{np(1-p)}} \quad ,$$

defined so that

$$P_n(np + t_k\sqrt{np(1-p)}) \sim \frac{\Delta t_k}{\sqrt{2\pi}} e^{-t_k{}^2/2} .$$

This allows us to bring Riemann integration into play giving the integral involving the standard normal distribution on the right hand side and some residual sums

The residual sums are shown to decline steadily to 0 as n→∞

9.6.3 Example

As an application of this theorem we give the following:
Find the probability that the number 30 is selected as the first number drawn in Super 6 Lottery draws for less than 11 times, but not more than twice in 253 successive weeks.

We assume that the lottery selects numbers randomly from 1 to 40. Therefore we need:

$$\sum_{k=3}^{10} {}^{253}C_k \left(\frac{1}{40}\right)^k \left(\frac{39}{40}\right)^{253-k}$$

Using the notation in the theorem: $n = 253,\ p = 1/40,\ a = 3$ and $b = 10$. This summation can be approximated by the area under the standard normal distribution curve and bounded by the limits:

$$\frac{3 - 253 \times \frac{1}{40}}{\sqrt{253 \times \frac{1}{40} \times \frac{39}{40}}} = \text{-1.33893} \quad \text{and} \quad \frac{10 - 253 \times \frac{1}{40}}{\sqrt{253 \times \frac{1}{40} \times \frac{39}{40}}} = 1.479874$$

So we compute:

$$\int_{\text{-1.33893}}^{1.479874} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} dx = 0.84025$$

One final note about approximating binomial probabilities has to do with p. It is small (close to 0) and large values (p close to 1) which give the worst approximations.

9.7 Various Types of Convergence

The Law of Large Numbers and other limit theorems which we have seen gave rise to a substantial amount of mathematical literature striving to tighten and generalize these results.

In doing this, the notion of convergence has had to be studied extensively and clarified in depth with the help of other branches of mathematics, notably measure theory and topology.
Convergence in the probabilistic sense can happen in various distinctly different ways.

We can have four main types of converges which we proceed to list:

9.7.1 Convergence In probability

A sequence of random variables X_1, ... X_n, is said to converge in probability to the random variable X if for all $\varepsilon > 0$:

$$\lim_{n\to\infty} P[\omega : \left|X_n(\omega) - X(\omega)\right| > \varepsilon] = 0$$

In other words, the sets on which X_n stays far from X even as n increases, their probability decreases and goes off to 0.

9.7.2 Convergence *P*-Almost Surely

A sequence of random variables X_1, ... X_n, is said to converges P-Almost Surely to the random variable X if

$$\mathbb{P}[\omega : \lim_{n\to\infty} \left|X_n(\omega) - X(\omega)\right| = 0] = 1\,,$$

equivalently:

$$\mathbb{P}[\omega : \lim_{n\to\infty} \left|X_n(\omega) - X(\omega)\right| \text{ does not exist}] = 0\,.$$

In other words the set of point on which the sequence X_n fails to converge has 0 probability.

9.7.3 Convergence In Distribution

A sequence of random variables X_1, ... X_n, is said to converges in distribution to the random variable X if the distribution function of the X_n's tends to the distribution function of X.
This definition is slightly problematic because density functions are not uniquely defined ... but we shall leave it at this.

9.7.4 Convergence In p'th Mean

A sequence of random variables X_1, ... X_n, is said to converge in p'th mean to rv X if for

$$\lim_{n\to\infty} \mathbb{E}[\left|X_n - X\right|^p] = 0$$

In other word the averaged absolute difference between X_n and X raised to the p'th power goes to 0.