Probability Models - 1

Remark 1. Here is a short primer on standard mathematical notation used in this chapter.

- A **set** is a collection of objects. As such it is an excessively general concept. In each context, we try to specify what a given set is supposed to be. However, it is often useful to keep the details generic. Sets are often denoted by upper case letters (mostly Roman, but sometimes Greek), and their elements by lower case letters. To write that the object a belongs to the set A, a common notation is $a \in A$. We can combine sets. For example, if A and B are sets, the collection of objects that belong to either one or the other (or, possibly, both), is denoted by $A \bigcup B$ (the *union* of A and B). The collection of points that belong to both is denoted by $A \bigcap B$ (the *intersection* of A and B). Finally, the collection of points belonging to A, but not to B is often denoted by $A \backslash B$.
- A **function** is a rule that associates to each element of a set (called the *domain*), exactly one element from a second set (called the *range*). The two sets may be made up of the same elements (usually numbers), but we keep them distinct for clarity. If A is the domain, and B is the range, a function called f, could be written as $f: A \longrightarrow B$
- A **subset** of a set is, as the name implies, a collection of *some* of the objects belonging to a given set. For technical reason, we include the set itself, as well as the **empty set** (an empty collection) among the subsets of any given set. Again, there are commonly used notations: to state that the collection B is a subset of the collection A, we write $B \subseteq A$, or $B \subset A$ (the first notation implies that we do not rule out that $B = A$, but, in practice, people use the two notation somewhat loosely). There is a standard notation of subsets, when they are characterized by some property. For example, suppose we have a function $X: A \longrightarrow B$. The subset of elements a in A such that $X(a) \in C$, where C is some subset of B, is often written as $\{a | X(a) \in C\}$.

Remark 2. A useful shorthand notation is the following. Often, we have to take the sum of a group of numbers. To streamline the notation, the first step is to give each of these numbers a "name", as in x_1, x_2, ..., x_n where n is the count of how many numbers we have at hand. The sum of these n numbers is often written as $x_1 + x_2 + ... + x_n$. However, this is a somewhat clumsy notation, so a shorthand has been introduced a long time ago: the same sum will be written as

$$\sum_{k=1}^{n} x_k$$

(this is called *summation notation*). Sometimes there is a formula for these numbers, which tells us what number will be in position k. In this case the summation notation is even handier. For example, the sum of the first n whole numbers, that is $1 + 2 + ... + n$, can be written as $\sum_{k=1}^{n} k$ (incidentally, we have a formula for this sum:

$$\sum_{k=1}^{n} k = \frac{n(n+1)}{2}$$

as – legend has it – proved by 8-year-old Gauss). Similarly, the sum of the first 20 even numbers could be written as $\sum_{k=1}^{20} 2 \cdot k = 2 + 4 + ... + 38 + 40$, and the sum of the squares of the first n numbers would be written as $\sum_{k=1}^{n} k^2$.

1 Preliminaries: What is Probability

The word "probability" has many colloquial meanings, and this does not help in clarifying its use in Mathematics and Statistics. While everyone has at least one intuitive understanding of its meaning, for scientific purposes, we should think of it as an array of mathematical tools to be used for quantitative modeling of applications, much like we think of algebra, calculus, differential equations, and so on.

The connection with reality will have to be considered on a case by case basis (just as we do with other mathematical constructions). In particular, we will concentrate, in this course, on the "classical" way of connecting mathematical probability and real life. This is not the only way, and it may not be the most appropriate under every circumstance. It can be argued that in applications to many social and economic problems a "subjective", or "Bayesian" approach would be more adequate. While the choice of "translation": has been (and, to some extent, still is) a hotly debated issue, we will not get into that, content in stating that our approach is pragmatic ("apply the method that seems to work best to a given situation"), and that our examples will all fit the "classical" model reasonably well.

1.1 Classical Interpretation of Probability

Classical probability addresses observations that can be repeated in (essentially) identical conditions, with each observation not influencing any of the other (for example, if you are flipping a coin, you must make sure you don't flip it in a way that will force it to come up with the same side, or the opposite side, and the previous flip - a trick that we all could perform on demand). For example, physical measurements can be performed many times, taking care in ensuring similar conditions every time. Experience suggests that, as you add new observations, even though these numbers keep fluctuating, their **arithmetic mean seems to stabilize**, settling closer and closer to some constant value. This value is interpreted as "the theoretical mean", or the "true value" of the quantity. In particular, if the observations merely count whether an event occurs or not, thus taking values 1 (if it does) and 0 (if it doesn't), this "limiting" value is interpreted as **the probability** of occurrence of the event.

Note 3. The fact that the average of a sequence of observations seems to move closer and closer to a specific number is not incredibly surprising, even if it is definitely not inevitable. Assume, for simplicity, that the observations will all remain between two fixed values, $a < b$. One can easily cook up a sequence whose average jumps around all the time, but that requires some clever choices. For example, suppose that every time, at step n, the average gets really close to $\frac{a+b}{2}$ (we chose the midpoint for simplicity and definiteness), the next n observations all turn up to be equal to a, or all equal to b. However, such a behavior "does not look random", but rather pointing to some goal-oriented production of outcomes.

On the other hand, consider a fairly "regular" sequence. As the simplest case, assume that the numbers are roughly evenly distributed in the interval, so that n observations are, more or less, given by $a + \frac{b-a}{n}$, $a + 2\frac{b-a}{n}, ..., b - \frac{b-a}{n}, b$. The sum of these numbers is $a + \frac{1}{n} \cdot \frac{n(n-1)}{2}(b-a)$, and, dividing by n, we end up with $a + \frac{n-1}{n} \cdot \frac{b-a}{2}$, which is very close to the midpoint $a + \frac{b-a}{2} = \frac{a+b}{2}$, and more and more so, as n gets larger and larger. The intuition behind exercises such as this is that if the sequence does not have special features that prevent the average from settling down, then this settling down is going to happen.

Note that this approach, **does not provide a definition** of probability, and that it is somewhat of the category of approaches that "work when they work". In other words, if you observe such behavior, you accept the interpretation, and if this behavior does not seem to occur, you decide that this is an experiment where the setup is not applicable (often suspecting that mistakes were made - but, again, this is all based on the belief that "if things are done right, this is what should happen", hence, if they don't happen, things were not done right, a good example of *circular argument*). Still, the very vague intuition behind the preceding Note suggests that our idea of "randomness" is at odds with a situation where this "settling down of the mean" does not occur.

The other problem with this approach is that not every observation can be repeated indefinitely many times, under identical circumstances. This has suggested alternate approaches to the interpretation of probabilities in real life situations (e.g., so-called "subjective" probabilities). Without going any further in this debate, we will be content in taking it for granted that, even in such one-shot (or "few-shots") instances, we will be able to make use of probabilities and means, to make useful statements, as discussed on a case by case basis.

1.2 Axiomatic Approach

The previous discussion only concerns the applicability of probabilistic methods to actual real-life circumstances. By themselves, these methods are now considered simply a mathematical area, like algebra, calculus, differential equations, and so on. The (idealized) job of a mathematical probabilist is to prove theorems, starting from a definite list of axioms, not to discuss what these theorems may imply in a given situation. Even an *applied* mathematician, will, in general, work on proving theorems about a specific mathematical object that was originally presented as a model for an application: once the original mapping between theory and application has been agreed upon, it is no longer an object of debate.

2 Introduction

Consider some of the examples we are going to work on, and think about what would allow us to draw broader conclusions than the mere listing of the observations.

- → **Polling:** we recorded answers from a tiny portion of the American population, and would like to draw conclusions about the population from this small sample. It seems like we need a model that will explain how a small number of individuals might be "representative" (and what this word actually is supposed to mean) of the whole.
- → **Measuring:** we need a theoretical method that could describe how repeated observations of the same quantity can result in different numbers, and what we can conclude about the "true" value (also to be defined) of our quantity from this puzzling array of numbers.
- → **Survival Analysis:** we measured the lifetimes of, say, a few light bulbs, and now we need to argue what the results (which are most likely very spread out) should mean in terms, for example, of a warranty that we are supposed to state (what is the lifetime we should guarantee, and how much free replacements will this choice cost us?)
- → **Correlation:** We have a few pairs of numbers, and believe there is a quantitative connection between them. We need a model to express this connection, and we need a method to make reasonable, and defensible, statements about his connection (e.g., allow us to reliably predict what the second member of the pair will be if the first is equal to a given number).

3 Probability as a Modeling Tool

It has turned out that Probability Theory does indeed provide tools to address the previous issues. They are not miracle tools: they are rather rational guidelines that allow us to draw conclusions, and, at the same time, determine the limitations of these conclusions, as well as tests to verify that their applicability is warranted.

A full fledged course in Probability is a tall undertaking, so we limit ourselves to basic facts and "rules", relying on the fact that they turn out to be fairly intuitive. After providing the language, we will turn to the application, at least as a framework, to each of the previous examples.

3.1 Probability Spaces and Random Variables

3.1.1 The Sample Space

We start by assuming that all possible "outcomes" of an "experiment" (that is observations, like the ones mentioned in the examples) can be thought as an element in a set which we will call the "Sample Space". Note that by "outcome" we mean a very broad and generic thing, not merely the number that we may get. For example, if the experiment consists in tossing a coin, "outcome" could encompass anything that happened as we did the toss, from the trajectory that the coin went through to the fluttering of a butterfly in the Amazon, which, by a popular paradoxical statement, could, in complicated ways, have had an impact on our experiment. We will denote this (presumably gigantic) set by some symbol (a popular symbol in the literature is Ω).

3.1.2 Random Variables

In practice, we don't evaluate but a minimal part of the data that would constitute an "outcome" as described above. We will observe only one or a few numbers (for example, the results of our measurement of the speed of light), or, possibly, some qualitative aspect (for example, the eye color of individuals, if we were trying to study the eye color of people living in the United States). Thus, there will be generally many "outcomes" producing the same observation. We will call this connection a "Random Variable". A moment of thought shows that a Random Variable can be thought of as a *function*, associating a number or a quality to every element of the Sample Space. In other words, a random variable can be thought of as a function

$$X : \Omega \longrightarrow E$$

where E lists the possible observations (numbers for the measurement of the speed of light, a list of possible eye color in the eye color experiment).

The important fact is that we can now consider as our object of study subsets of Ω, of the form $\{\omega | X(\omega) = a\}$, or $\{\omega | X(\omega) \in A\}$, where the symbol $X(\omega) \in A$ means "the observed value of X belongs to the collection of possible observations we called A" – for example, our measurements fell within a specified interval. We will call these sets "Events" (that's what actually "happens", as far as we are concerned).

The simplest examples are random variables that take only a finite number of possible values. For example, a coin toss, where we could count 1 if we guessed the outcome right, and 0 if we didn't, or the toss of a die, where we can denote by X the points that end up on the top side of the die, so that X takes values 1,2,..,6. These are easier to handle, but, bear in mind that they are not always the most useful (one notable exception being the theory of games of gambling).

3.1.3 Events and Probabilities

Now, the point of our model is that we cannot be really sure what we will actually observe, once we perform our experiment. Indeed, even the very same experiment, when repeated, will result in a different result. However, there are results that we may find reasonable, and others that we don't really expect – for example we don't seriously expect the speed of light to be observed as 40 miles per hour. To quantify this observation, we will **assume** that we can associate a number to each event, expressing a quantitative measure of its likelihood. We call this number the "Probability of the Event", and will write stuff like

$$P[X \in A] = p$$

where p is the appropriate number. **By tradition and convention**, these "probabilities" are numbers comprised between 0 and 1 (inclusive), with higher numbers denoting higher likelihood of the event occurring.

There are some reasonable logical rules that need to be followed to make sure statements like "the probability of the event is 0.35" are consistent with each other. These rules go by the name of "axioms of probability" and are very natural once we stop and think about what these statements are supposed to mean.

Now it turns out that, as long as we have decided to concentrate on a specific random variable, we can forget about Ω, which is a very abstract concept anyway, and focus on events of the form we just mentioned. This means that, in practical terms, we are associating probabilities to subsets of E, as in $p(A) = P[X \in A]$. This system of probabilities on E is called the *distribution* of X, and is what we will actually be working with.

We denote the probabilities of all events defined by a Random variable X, as the "Distribution of X". When the possible values for X are just a finite collection, we can determine the distribution by assigning the probability of each outcome, as in $P[X = x_k] = p_k$, if the possible values are denoted by $x_1, x_2, ..., x_n$. In particular, some simple examples can be modeled by assuming that all outcomes have the same probability, that is $p_k = \frac{1}{n}$. This is sometimes called the "classical model", and is commonly used in evaluating the odds in games of pure chance. However, in statistical, "real life" applications, such models rarely play more than an ancillary role, if even that.

3.1.4 Axioms of Probability

We list the axioms here.

1. $0 \leqslant p(A) \leqslant 1$ for any event A
2. $p(\emptyset) = 0$, $p(\Omega) = 1$
3. If A and B are *disjoint* (that is, there is no common element to A and B), $p(A \bigcup B) = p(A) + p(B)$

If events are referring to Random Variables taking only a finite number of values, we don't need anything more. It is however convenient to consider variables that can take an "infinite number" of values, e.g., that can take any real number as value. To handle these consistently, it turns out to be necessary to extend the third axiom to the case where instead of two we have a *sequence* of disjoint events, A_1, A_2, A_3, ... (where the sequence continues indefinitely). We don't really need this extension for the situations we will encounter, but it is useful to bear it in mind, when moving to more advanced topics in Probability and in Statistics.

Note 4. The *frequency* interpretation of probability

The axioms connect to the intuitive understanding of what a statement like "the probability of this event happening is 0.4", in terms of *frequency* of this event in a repeated experiment. In fact, frequencies satisfy the axioms of probability. As mentioned before, if we perform an experiment or an observation many times, always in the same circumstances, the *frequency* of an event occurring, that is the fraction of times that event occurs, over the total number of observations, tends to stabilize around its *probability*. For example, one can argue theoretically that the probability of a flipped "fair" coin turning out "Heads" should be 0.5 (or 50%), and if you flip a coin many times, you will often observe that the frequency of heads does come close to 50%.

Remark: "Close" in the last statement can be made quantitative, as we will soon see. In any case, it definitely does **not** imply as much as is sometimes claimed. For example, you may read that somebody tossed a coin 24,000 times and came up with 12,012 heads (it is said to have been done by a major contributor to modern statistics, Karl Pearson). This might have happened, but it is not very likely (unless you are cheating) - one can compute the likelihood that a "fair" coin would do something like this, and it turns out to be a measly 3%.

Note 5. If you are a careful thinker, you will notice the many caveats that are implied in the previous statement. As we shall see, there is a *mathematical statement* that can be interpreted as supporting the assertion above, but the catch is that, in any given real life (as opposed to abstract mathematical) circumstance, you may or may not observe this "regular" behavior. In fact, the assertion above (sometimes called, improperly, the *empirical law of large numbers*), is one of those assertions that hold when they hold, and, when they don't, oh, well... You will sometimes find this assertion presented as a possible *definition* of probability, but it isn't: read it carefully, and you'll find that it is a circular statement – as in "a blue sky is a sky that is blue". We do usually assume that the empirical law will be valid in our experiments. However, we will need to be careful in setting them up, so that they resemble as much as possible the *mathematical* law of large numbers, and, in any case, the proper logical way of thinking is that we have a mathematical model (Probability Theory), and we believe (or hope) that its predictions are significant when applied to real life situation.

Note 6. Another problem with the preceding discussion is that it refers to the ability to repeat an experiment a large number of times. What about one-off situations? E.g., "what is the probability that I will live to age 99?". On the one hand, you can dismiss such a question as meaningless: there is only one "I", and I won't get to repeat my life many times, so I cannot make make sense of the above. On the other hand, people have been trying to apply Probabilities to such statements in a coherent way. This line of thought, called *subjective probability* has useful applications for sure (for example, in what is called *decision theory*), and it leads to a specific development of statistical techniques (the so-called *Bayesian Statistics*), but we will not go there in what is an introductory course. The "frequentist" interpretation of probability, with all its logical caveats, is the more common tool in Statistics (this approach is, indeed, called *Classical Statistics*).

3.1.5 Multiple Random Variables

Often enough, we will be working with more than one Random Variable, that is, we will be observing several values together. In general, the model will consider these Variables as a group, so that events will have the form

$$\{\omega | X_1(\omega) \in A_1, X_2(\omega) \in A_2, ...X_n(\omega) \in A_n\}$$

and the corresponding probabilities are often called the *joint distribution of the variables* X_1, X_2, $...X_n$. Problems involving several variables are more complicated, and it is thus a welcome situation when the joint distribution does not require more than the individual distributions to be determined. This is a *very special case*, when the variables are said to be *independent*, and cannot be taken for granted as the right model, unless we make sure the circumstances warrant it.

3.1.6 "Population" and "Sample"

Statistical terminology can be confusing at times, since it developed in a less than linear way. One example is the terminology in the title of this subsection, which originated specifically in polling applications, but is used in general, even when it is not clear what kind of "population" we may be talking about.

At the most general level, statistical experiments go something like this. We are interested in the feature of some quantity (or quantities), and decide to model it as a random variable X, whose distribution we don't know. To gain information on this distribution, we perform a series of observations of this quantity (or some other closely related one). Each observation results, for example, in a number (if we are considering such a random variable), and we view the collection of these observations as the observations of distinct instances of this random variable – let's say $X_1, X_2, ..., X_n$. Probability theory will help us make a connection between the distribution of X, and the *joint distribution* of these instances, based on how we represent our experiment mathematically. The purpose of statistics is to use the values observed to infer conclusions on the distribution of X.

Now, X is a function defined on the sample space, with values in some other space, say E. The distribution of X associates probabilities to subsets of E, and these probabilities (which are unknown – it is what we are trying to discover) is what is meant by "population". On the other hand, the observed *values* in our experiments, that is the actual values observed for $X_1, X_2, ..., X_n$ are the "sample".

3.2 Independent Random Variables

In general, the joint distribution of our observations could be difficult to calculate from the distribution of X. In a special case, though, the relation is extremely simple. This is also the special case where we obtain the most information about the unknown distribution from the observed sample. The special case we just mentioned is when the joint distribution is such that it is always true that

$$P[X_1 \in A_1, X_2 \in A_2, ...X_n \in A_n] = P[X_1 \in A_1]P[X_2 \in A_2]...P[X_n \in A_n]$$

If that's the case, we say that the Variables are *independent*.

The intuitive idea behind Independence is the following: the observation of any of the variable(s) does not affect the observation(s) of any other(s): whether we observe them one at a time, or together, the probabilities are not affected at all.

We generally assume, for example, that rolling two dice results in two outcomes that are independent, but if, say, the two dice were loaded with magnets, the outcome of one would definitely affect the other, and independence would not be a reasonable assumption.

Independence is not an obvious feature, and should be assumed in a given situation only when it is clear that it is a fair description.

3.3 Initial Probability Models for our Examples

3.3.1 Polling

This is the framework where the terms "population" vs. "sample" originated. Imagine you wanted to learn what Americans think of apple pie. Since there are about 350 to 400 million Americans, it's not realistic to go out and question each one.

Remark 7. Of course, a census is exactly an attempt to do just that. Indeed, the model we discuss here is not relevant for a census, since, by definition, it tries to count every individual. That does not mean that statistics is irrelevant to full census takers, quite the contrary (The US Census Bureau has a powerful statistical team at constant work), just that they do not use the polling model for this specific task (they do a lot of sampling too, of course, besides the full census taken every 10 years). The statistical issues in a census are very difficult to address completely, and are beyond the scope of our introduction (to mention one famous issue, think of the under-counting of particular groups of individuals).

A poll consists in selecting a limited number of individuals, collect their answers, and use the result to "guess" what a full survey could have resulted into. For simplicity, consider the case of a question with two possible answers, "yes", and "no" – with "don't know" or "don't want to answer" as a third possibility. Now, the simplest model for this enterprise is to imagine, instead of a collection of N (350 million plus) individuals, that there are N balls in an urn, part is colored white (and corresponds to "yes"), part is black ("no"), and the rest is green ("don't know"). We take a blindfolded child to the urn, which has been thoroughly shaken so that the balls are all mixed "at random" (in the intuitive sense), and have the child pick n (a reasonable number – for polls it is usually around 500 or 1000) of the balls. The idea is that the *probability* of picking a white ball should be given by the fraction of white balls over the whole bowl. This is intuitive, and, all things being equal, there is no reason to believe otherwise (we are applying the so-called *principle of sufficient reason*: we know of no cause that should make it more or less likely than this to pick a white ball – it is the same logic that has us assign a probability of $\frac{1}{2}$ to "Heads" when tossing a coin).

If we had n_w white balls, n_b black, and n_g green ones, we would then assume that the probability of picking these colors are, respectively, $\frac{n_w}{N}, \frac{n_b}{N}, \frac{n_g}{N}$. Since we are picking very few balls out of a huge amount, when we go for the second, third,..., even thousandth ball, the proportions just mentioned will not change appreciably (we are assuming that n is very small not only with respect to N, but also to n_w, n_b, and n_g). Moreover, we can assume that everything has been so well mixed, that the fact of extracting, say, a white ball, does not affect the probability of extracting any given color on the next round. Note that the condition on n being relatively small is crucial here, because if we had, say just 6 balls, 2 white, 3 black, and 2 green, after extracting a withe ball, things would be quite different.

From these considerations, one can compute the probability of extracting any given number of balls of each color. The formula is not too complicated, but it becomes completely unworkable (in the sense of actually computing the probabilities) for the numbers we are looking at here (in fact, it is hardly workable for much smaller values). Fortunately, we will soon meet a solution to this serious problem: as long as all the above holds, and n itself is not too small, there is a direct formula that is approximate, but a very good approximation.

On the other hand, in real life, we do not have a bowl and colored balls when we set up a poll. The challenge to set up a practical polling method that approximates the ball-in-a-bowl ideal is considerable. We will discuss it, not too deeply, of course, after we have familiarized with some of the theoretical (ideal) tool of statistics, at the end of the course.

3.3.2 Measurements

Here we measure a quantity – a physical quantity, for example. This could be the speed of light, the voltage produced by a specific battery, or the concentration of Hydrogen ions in a chemical solution. In all cases, when the measurements are performed carefully and precisely, it turns out that, invariably, repeated measurements yield different results. The likely reason is that, no matter how carefully we work, there are so many factors we cannot control that produce these discrepancies. The differences are usually not very large, and, if we are happy with a rough result, could be ignored, but if precision is required, we need to come up with some kind of reasoned result out of this collection of numbers.

Here things are trickier. We have something to measure, so we assume there is a random variable X describing the possible outcomes of our experiment. These are, in general, real numbers, and the fact that there's so many is due to the fact that as we repeat our experiment, all sorts of additional factors will come in, that we cannot control, and sway our observations one way or another. If we are working carefully, these disturbances will be small, and they will not bias the outcome one way or another. To make this clear, for high precision measurements, even the thermal agitation of the molecules in our instrument will affect the result, but it won't push in any preferred direction.

To come up with a model for this case is clearly hard, but, again, we have nice results that tell us that the combined effect of many small (and essentially independent) disturbances will produce a typical distribution. This is a combination of a mathematical result, and physical insight, but it has proved to work very well.

3.3.3 Survival Analysis

An important application of statistics is in the areas of population dynamics, medical studies, reliability theory, and similar fields, where we have "individuals" (people, animals or other living creatures, machines, and so on) that have a finite "life span" (for living beings, that is clear, for a machine it would be the time to failure). This is a sampling problem, as in the first case (we can only observe the life span of a sample of the population of individuals), but the result is a positive number (the lifetime), and not a simple finite set of options (as in "yes", "no", "don't know"). Also, the "population" in question might be very open ended (for example, when we are studying the time to failure of a piece of equipment that will be manufactured in large and, as yet unknown, quantities).

The appropriate models here are quite different from the previous ones. Depending on the object of our investigation, we can expect the likelihood of living longer than a given time span to decrease in different ways (for example, there may be a high rate of "infant mortality", and a subsequent leveling off of the likelihood of failure, once the first wave of failures has occurred; or there may be a constant "aging" effect"; or a mixture of the two, or yet another pattern). There is a whole literature on this problem, which is, for instance, crucial in industrial production ("time to failure"), ecology, insurance, and more.

3.3.4 Comparing Different Groups

An important special case of the previous examples is when, rather than trying to establish a quantity for a population (e.g., the distribution of life span of a certain ethnic group), we are trying to *compare* the value of a quantity between two or more groups. A typical example is medical testing, where a group of patients may receive a certain treatment, while another group of patients (the *control group* receives a "placebo" – that is a treatment known to be ineffective). The point of such a study is to see whether the treatment is effective, in that it produces a better result than doing nothing. We are in the presence of two samples taken from two "populations" (not so well defined: one is the "population" of those that will undergo treatment, and the other is the "population" of those who get a placebo – we only have the samples, but are assuming that they will tell us about the universe of those who *would* take the treatment and those who *would* not), and we are interested in evaluating the differences.

We will see that there are tools to evaluate the probability of the parallel two experiments yielding significant differences or not. One important point about these tools is that their effectiveness depends on strict assumptions on the distributions of the random variables we are considering. While these assumptions can be justified in many cases, they are very speculative in others, and this incertitude has to be taken into account.

3.3.5 Correlation

A more complex problem arises when we are effectively considering two (or more) different quantities that, we believe, affect each other. For example we might investigate whether family income affects academic performance. Assuming we have a way of quantifying numerically the quantities in play, we may look, for example, at whether increasing one quantity appears to imply an increase (or decrease) of the other, and vice-versa. Note that there is no attempt to show any *causation* – we are not trying to prove that a higher value for quantity A *causes* a higher (or lower) value for quantity B. Clearly, this is a more delicate issue, and is a very rich source of bogus results, whenever the very strict circumstances in which deductions can be made reliably are more or less ignored.

The model for this situation is to assume that we have two random variables that are, in general, *not* independent, and we would like to study the *conditional distribution* of one, conditioned on the other. With additional (strong) assumptions on the distribution of the variables, the problem turns into a fairly simple calculation, known as *Least Mean Squares* estimation. When these strong assumptions are unjustified, there are more complicated tools that can be employed, but a more common approach is to set up a model that is essentially ad hoc, with little theoretical support, but that is analyzed by -the same, simple, Least Mean Squares approach. Again, a critical perspective is very useful in assessing the implications of studies like this.

3.4 Indexes for the Distribution of Random Variables

If we think of the distribution of Random Variable as a sort of "mass" allocated among the events associated with it, we can consider the "Center of Mass" of this distribution. Traditionally, this Center of Mass is called the *Expectation* (or *Expected Value*) of the Random Variable (even though, it's got nothing to do with anything we might "expect"), also called its *mean* (please, do not mistake this for the *sample mean*, that is the average of observed values in sampling: expectations are theoretical values, while sample means are the result of empirical observations). The expected value of a random variable X is usually denoted by $E[X]$

For the simple case of a Random variable with finitely many values, say x_1, x_2, ..., x_n, this would be *the weighted mean of the values*, with probabilities being the weights (note that -the probabilities add to 1)

$$E[X] = x_1 \cdot P[X = x_1] + x_2 \cdot P[X = x_2] + ... + x_n P[X = x_n] = \sum_{k=1}^{n} x_k \cdot P[X = x_k]$$

More general cases require calculus (possibly, *advanced* calculus) to be properly defined, so we'll be content saying that we have a way to to find the mean in such situations, and will be told what it is. The following facts about expectations are easy to prove in the finite case, and can be proved in the most general situation as well:

- $E[aX + bY] = aE[X] + bE[Y]$

for any two random variables X, and Y, and any two real numbers a, b

- If $X \leqslant Y$, then $E[X] \leqslant E[Y]$

It also turns out to be useful to consider the expectation of powers of a Random Variable, $E[X^m]$, for positive whole numbers m (called "Absolute Moments"). Sometimes, it is the expectation of powers of $X - E[X]$ (called "Centered Moments") that are preferred. One reason of this interest is that, *for some specific distributions*, we can determine the specific distribution if we know its moments (and sometimes, only a few moments are necessary). It is important to realize that this statement is true only for specific types of Random Variables. If we don't have enough information on its features, moments may be of limited or minimal interest. The first centered moment is $E\Big[(X - EX)^2\Big] = \text{Var}[X]$, and is called the *variance* of the distribution (again, we are referring her to the *theoretical variance* of a random variable – what we called "population variance" and "sample variance" refer to empirical observations). Just as in the discussion at the end of the file on notations (check the link from the introduction to the HTML or the corresponding PDF file), which concerned the "population variance" and the "sample mean", one can prove, for the *theoretical* variance and the expectation, following essentially the same steps, that

$$E\Big[(X - EX)^2\Big] = E\big[X^2\big] - (E[X])^2$$

This observation has an unexpected consequence: since the left hand side is, essentially, a sum of squares, it cannot be negative (and, indeed, it is zero only if X can take one value only). Hence, the same holds for the right hand side, that is, no matter what distribution X has,

$$E\big[X^2\big] - (E[X])^2 \geqslant 0, \text{that is } E\big[X^2\big] \geqslant (E[X])^2, \text{or } \sqrt{E\big[X^2\big]} \geqslant |E[X]|$$

This can work as a check as to whether you did a calculation wrong: if you get a "negative variance", you must have made a mistake somewhere. This applies, in particular, to the case of "empirical distribution" calculations, which we will discuss in more detail in the next few sections.

The (positive) square root of the variance is called the *standard deviation*.

Remark. The variance (and, hence, the standard deviation) is a measure of "dispersion" of a distribution (just as its namesake, with respect to a sample). In particular, being a sum of numbers that are never negative, it is equal to zero only if all the numbers are equal to zero – that is if the "distribution" is trivial, since X can take one value only. However, the value of variance as a measure of dispersion depends on the details of the distribution. In particular, statements that you may come across, like "values more than three standard deviations away from the mean are extremely unlikely", assume a very particular type of distribution (so called "normal distributions", very common in applications, but by no means exclusive), and are unwarranted if you are working with a different case.

While, as noted, the sum of any number of random variables, has an expectation that is the sum of the expectations of the individual variables. That is due to the *linear* structure of the expectation. The variance, involving the *square* of the random variable, is very different, and you should not ever assume that variances simply add, except in special circumstances, as discussed in our next module.

Remark 8. While we will not consider these cases (at least not in detail), we should note that variance, mean, and, in fact, all moments, may not always be defined. This is the case when the set of values our variable takes is infinite, and the probabilities associated with large values are not small enough. The useful fact to know is that, if $E[X^m]$ is well defined for some integer m, then also all moments $E[X^n]$, with $n < m$ are well defined. However it can well happen that only a finite set of moments is well defined – or even none at all.

3.5 Examples of Distributions of Random Variables

1. "Bernoulli Distribution". Suppose X may only take two values, say a, and b. Then its distribution is determined simply by knowing the probability of each value, and we will have $P[X = a] = p$, $P[X = b] = 1 - p$. An application of the definition shows that

$$EX = ap + b(1 - p)$$

so that knowledge of EX allows an easy reconstruction of p, and hence of the distribution. In most cases, we work with $a = 1$, $b = 0$, which is the same as counting the occurrences of case a. In this case, $E[X] = p$. Also, in this case, a similar calculation shows that $\mathrm{Var}[X] = p(1 - p)$

2. "Binomial Distribution". Suppose we repeat a Bernoulli experiment as described above, over and over again, in the same circumstances, and making sure that **each repetition does not influence any other** (that is, we require independence), and count how many times we end up with outcome a. To find the distribution of this count we need some tricks from "combinatorial analysis", and the result is a formula, for the probability of k "successes" out of n tries:

$$P[X = k] = \frac{n \cdot (n-1) \cdot (n-2) \cdots (n-k+1)}{1 \cdot 2 \cdot 3 \cdots \cdot k} p^k (1-p)^{n-k}$$

The coefficient in front of $p^k(1-p)^{n-k}$ is called a *Binomial Coefficient*, appears in many other formulas (you may have met the *Pascal Triangle*, which is a simple method to calculate Binomial Coefficients), and is usually denoted by $\binom{n}{k}$ ("n *choose* k"). Being the sum of n independent Bernoulli variables, a binomial variable has $E[X] = np$. It also turns out, because of the assumed independence of the Bernoulli variables, that the variance of a binomial variable is equal to $\mathrm{Var}[X] = np(1 - p)$

3. "Uniform Distribution" between two values $a < b$ (the "standard" case is $a = 0, b = 1$). If X has this distribution, the probability of it taking values *outside* the interval $[a, b]$ is zero. The probability of taking values in an interval $[c, d]$, with $a \leqslant c \leqslant d \leqslant b$ is proportional to $d - c$, and the proportionality constant makes sure that $P[a \leqslant X \leqslant b] = 1$, that is, $P[c \leqslant X \leqslant d] = \frac{d-c}{b-a}$. Note that if $c = d$, the result is $P[X = c] = 0$. In particular, if $a = 0 \leqslant c \leqslant d \leqslant b = 1$, $P[c \leqslant X \leqslant d\Big] = d - c$

   Remark 9. Most programming languages, and most spreadsheets have a built-in function (called `rand`, or `rnd`, or similar names) that will produce a number (a different one every time you call the function), in a way that is practically the same as picking a number "at random" from a uniform distribution on [0, 1]. There are ways to "transform" this output so that the outcome will behave like the observation of random variable with any prescribed distribution, which is why this usually the only such function included in a package. One of the advantages of *Gnumeric*, is that, out of the box, it has pre-programmed random number generators for a very large of number of different distributions, without the need for a user to code extra instructions. More popular spreadsheets, like LibreOffice/OpenOffice Calc or Excel do not have similar functions out of the box.

4. "Exponential Distribution" describes random variables that can take any non negative value, and are such that for any $x \geqslant 0$, $P[X > x] = e^{-\lambda x}$. Here $\lambda > 0$ is a parameter that determines the details of the distribution. Of course, $P[X \leqslant x] = 1 - e^{-\lambda x}$. It turns out that, for such a random variable, $EX = \frac{1}{\lambda}$, so that, again, the expected value determines the distribution. Incidentally, we also have $\mathrm{Var}[X] = \frac{1}{\lambda^2}$. This distribution is important in the study of reliability and survival, as it is the borderline case between "aging" and "rejuvenating".

5. "Normal (or Gaussian) Distribution". This is a very important distribution, that appears constantly in applications. The reason is that, thanks to a theorem, sometimes referred to as "The (Second) Fundamental Theorem of Statistics", but more properly as the *Central Limit Theorem*, it does indeed apply to many different cases. This leads sometimes to use it always, as if there was no other plausible model. While it is an extremely common model, it has its scope – a vast one, to be sure, but not universal. We will discuss this more precisely with examples. The explicit form of this distribution is nowhere as simple as the other examples we mentioned here. In fact, we have to use tables (or pre-programmed routines in our computers) to evaluate probabilities associated with it. It is determined by its expected value (often denoted by μ), and its variance (often denoted by σ^2). A common notation, which we will use, is, in the case $\mu = 0, \sigma = 1$, which is called the *Standard Normal Distribution*,

$$P[X \leqslant x] = \Phi(x)$$

3.5.1 A Note on "Continuous" Distributions

In the previous examples we met two "discrete" distributions: the Bernoulli and the Binomial distributions apply to variables that can only take a finite number of values. For these, it is customary to provide the *Probability Mass Function*, that is the probability that any given value may be taken (that's exactly what we did). The remaining three are examples of so-called "continuous" distributions: the variables can take any real number as value (with some limitations, possibly, as for the Uniform distribution, taking as values any real number between a and b, and the Exponential Distribution where only positive values are allowed).

For these, it is customary to provide the *Cumulative Distribution Function*, which is the function $F_X(x) = P[X \leqslant x]$ (as you can see, Φ is the Cumulative Distribution Function for a Standard Normal, or Gaussian, Random Variable). Alternatively, especially in the case of positive Random Variables, as are commonly used in Survival Analysis, it is sometimes convenient to refer to the *Survival Function* (also called (*Reliability Function*), defined as

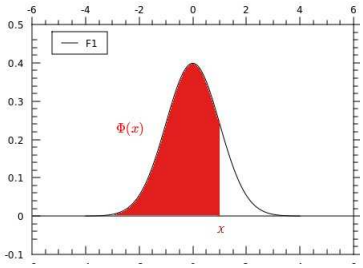
$$R_X(x) = P[X > x] = 1 - P[X \leqslant x] = 1 - F_X(x)$$

You will notice that $e^{-\lambda x}$ is the Survival Function for an Exponential Random Variable, with parameter λ.

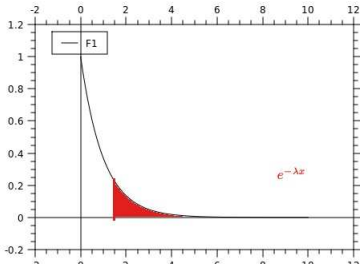
Whichever of the two is used, it is easy to compute probabilities such as

$$P[a < X \leqslant b] = F_X(b) - F_X(a) = R_X(a) - R_X(b)$$

The way continuous distributions are defined depends on their complexity. In all common cases, the probabilities above are defined as given by the area between the graph of a function (called the *density*) and the horizontal axis. For example, the function Φ mentioned above gives the area filled in red in the following graph:



The curve is the graph of the function $\frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}$ (this is the well-known "Bell Curve"). Similarly, the survival function of an exponential, for example one with $\lambda = 1$, is the area filled in red in the following graph:



In general, the curve is the graph of the function $\lambda e^{-\lambda x}$, limited to the set of $x \geqslant 0$.

While one has to consider, in general, more complicated continuous distributions, we will limit ourselves to those that are most common in statistics, and they all share one feature: their Cumulative Distribution Functions can always be found as the area below a continuous function, as in the examples above. This continuous function is called the *density* of the distribution. Hence, all continuous distributions we will consider will have densities.

Remark. Since probabilities are never negative, a density must never be negative either. Also, the area below its complete graph must add up to 1, since that would be the probability of "anything happening".

3.5.2 A Note on "Probability Zero"

This is a somewhat delicate point. For "discrete" Random Variable, the concept is easy: a value can be taken by it, or it can't - in the latter case, its probability is zero. Things are subtler for continuous variables. The problem lies in the fact that we cannot pretend to observe such a variable with complete precision. For example, let's suppose we are measuring the time to failure of a piece of equipment. "Time" is the quintessential "continuous" quantity, since we think of it as capable to be subdivided indefinitely. Nonetheless, our ability to measure it is limited by the precision of our instruments, and this is finite. So, we may measure time "up to a 10th of a second", "up to a millionth of a second", and even more, but never "up to infinite precision". This is the same conundrum encountered when we face "real numbers". A real number is defined by an infinite sequence of decimal digits, and we cannot express such a sequence explicitly (an infinite sequence is too long for beings with a finite life span). What we can do is approximate such a number as much as we wish, but always up to a point.

Now, let's look at what happens if we wish to pinpoint the probability that our piece of equipment will fail exactly at time t, assuming the distribution of its time to failure, X, is exponential. We can calculate, for example,

$$P[t-\varepsilon < X \leqslant t+\varepsilon] = e^{-\lambda(t-\varepsilon)} - e^{-\lambda(t+\varepsilon)}$$

If we try to take smaller and smaller value for ε, we see that this probability becomes smaller and smaller, ever closer to zero. The conclusion is that "the probability of X being *exactly* equal to t is zero". This seems paradoxical: even if it is unlikely that the time to failure will turn out to be *exactly* t, it *will turn out to be some specific number*, yet the probability of it taking this number "is zero". Fact is, we will never know or observe this number: we will observe a small interval, like the one above, whose size depends on our measuring instrument, and the probability of X taking a value in this interval may be small, but is not zero. Thus, for continuous Random Variable, "probability zero" does not correspond to an event conceptually impossible: it corresponds to an event which cannot be directly observed, but which can be observed approximately, with a probability that is smaller and smaller, the better our approximation is.

3.6 The Empirical Distribution

You may have noticed a striking connection between the "mean" and the "variance" mentioned above, and the indexes by the same name we met in the Descriptive Statistics chapter. This is no coincidence, of course. In fact, the discussion there can be rephrased in the language of probabilities, if we look at a sample as follows.

Consider a specific sample, that is a collection of numbers $x_1, x_2, x_3, ..., x_n$. Define now a random variable that can take only these values, each with probability $\frac{1}{n}$. Then the distribution of this random variable defines our observations, and is called the *empirical distribution* of the sample (note that, generally, a new sample will produce a different empirical distribution - hence this is called a *random distribution*), and the mean, variance, and so on of this variable are precisely the corresponding quantities we defined for our sample. As we will discuss more in detail, our goal will be to use information from the empirical distribution (e.g., its mean and variance) to estimate properties of the distribution of the random variable we are studying.

Remark. In this line, the "proper" index is the so-called *population* variance, since that is the variance of the empirical distribution, as we defined it. The "correction" (you will recall that the *sample* variance is defined as $\frac{n}{n-1}$ times the *population* variance), is due to a totally different consideration of samples.

When the simple models we will work with happen to be inadequate, one can try to do statistics in terms of the complete empirical distribution. This area, *nonparametric statistics*, is harder technically, and lies outside our scope, but is a very interesting topic. An example that is coded in the Statistics menu of Gnumeric is one of the many procedures developed to test whether a given sample can be reasonably assumed to come from a normal distribution.

4 Finite Probability

The basic ideas of probability and how it works are best understood in simple cases, when random variables can only take a finite (and small) number of values. Unfortunately, this simple environment is of little consequence for practical statistics, since the prevailing tools all refer to continuous distributions (for example, thanks to the Central Limit Theorem, a pivotal result that we will look at in the next part).

Still, you should explore the concepts presented in Chapter 5 from the Online Stat book (see the course map), since it is a way to get a better intuitive feeling of how probability works. Please, take the time to go through that file, and its links, as they cover several points that we may not be work very much on, but that have serious implications.

5 Distributions We Will Use

In most cases, we will have to decide which type of distribution should be associated with our "population" random variable X. Statistical tools will then be used to choose specific constants that are left undecided. With minimal exceptions, the most useful distributions require elaborate computations to be applied, so much so that they are pre-programmed in all spreadsheets and are tabulated in any statistics textbook.

While the available distributions are many, the ones we will focus on are the following. Note that, thanks to theorems like the Central Limit Theorem we just mentioned, direct use of discrete distributions in statistical applications is fairly rare: the conditions for their replacement with an appropriate continuous one are commonly in place, which is a good thing, since it makes the job of analyzing an experiment much easier.

Remark 10. There is another important limit theorem which applies to binomial experiments with n large, but p so small that the the product np has a low value (say, no more than 10, or so). The limiting description is the so-called *Poisson distribution*. It is an interesting case for a number of reasons, but we will not cover it, even if you are encouraged to look it up in some of the external references we quoted.

5.1 Normal (Gaussian) Distribution

This is the most common model, thanks to the Central Limit Theorem. It is so common that it is sometimes "forced" on a problem, even when it would not be prudent to do so. A side discussion of an example of this kind of attitude is in an optional file on *truncated normal distributions*.

The normal distribution is actually a family of distributions, each one characterized by its expectation (let's call it μ), and its variance (let's call it σ^2). The density of a normal distribution with these parameters (the distribution is often denoted symbolically as $N\big(\mu, \sigma^2\big)$) is

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

The cumulative distribution function, cannot be expressed in terms of "elementary" functions (the functions we learned about in algebra, and precalculus), but is pre-programmed in all spreadsheets.

The special case of $N(0,1)$, that is $\mu = 0, \sigma = 1$, is called the *standard* normal distribution. Its cumulative distribution function, denoted often by Φ, is specifically pre-programmed in any spreadsheet, as well as tabulated in textbooks. The reason is the following fact:

Proposition 11. *Suppose X has distribution $N\big(\mu, \sigma^2\big)$. Thus, $E[X] = \mu, \mathrm{Var}[X] = \sigma^2$. Then the random variable*

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

($\sigma = \sqrt{\sigma^2}$) has distribution $N(0,1)$: it is normal, with $E[Z] = 0, \mathrm{Var}[Z] = 1$, and density

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Thus, if you need to know the value of $P[a < X < b]$, and only have access to tables, you can use them to get your answer, because

$$P\Big[a < X < b] = P\Big[\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma}\Big] = P\Big[\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\Big]$$

A notable fact of normal distributions is that the sum of independent normal variables is again normal (this "invariance" is unusual among distributions). Using observations we will make in the next unit, if the variables X_k are distributed according to $N\big(\mu,\, \sigma^2\big)$, then $\sum_{k=1}^n X_k$ is distributed according to $N\big(n\mu,\, n\sigma^2\big)$, and $\frac{1}{n}\sum_{k=1}^n X_k$ is distributed according to $N\Big(\mu, \frac{\sigma^2}{n}\Big)$.

This is the most important distribution in this course (and, possibly, in most common statistical applications), so go to the accompanying link on this topic from the Online Stat Book

5.2 Chi-Square Distribution

It is sometimes useful to consider the distribution of *the sum of squares of independent, identically distributed normal random variables*. Again, a simple transformation allows to reduce any such sum to the sum of squares of *standard normal random variables*.

Traditionally, if the variables $X_1, X_2, ..., X_n$ are distributed as $N(\mu, \sigma^2)$, the distribution of $\sum_{k=1}^{n} X_k^2$ is denoted by χ^2_{n,μ,σ^2}. However, if we start by "normalizing" the variables as before, that is we work with variables $Z_k = \frac{X_k - \mu}{\sigma}$, the resulting sum $\sum_{k=1}^{n} Z_k^2$ has a distribution, denoted by χ^2_n, which is tabulated in any textbook. As usual, your spreadsheet has all of this already pre-programmed.

Note 12. For historical reasons, the number n that identifies each "chi-square" distribution is called its number of *degrees of freedom*.

Note that, given its structure as related to the sum of squares of independent identically distributed standard normal random variables, the sum of a χ^2_n, and a χ^2_m random variable has distribution χ^2_{n+m}.

5.3 Student *t* Distribution

A brilliant statistician, William Gosset, working for the Guinness brewery in Ireland, devised an explicit formula for the distribution of a random variable that appears naturally in statistical problems involving normal variables. The distribution was actually found earlier, but it was Gosset's (independent) work that brought it to the forefront, thanks to the recognition by the leading English statistician of the time, Fisher. He could not use his name, as the company had strict rules against publishing, for fear of publicizing trade secrets, so he published his results under the pseudonym of "Student". Too bad his real name is not nearly as famous.

As we will briefly discuss in the *Estimation* module, we can make good use of the distribution of the quotient of a standard normal random variable by the square root of the ratio of a chi-square random variable with n degrees of freedom, and the number of degrees of freedom: if X is a standard normal random variable, and Y_n has χ^2_n distribution, we are referring to the quotient

$$\frac{X}{\sqrt{\frac{Y_n}{n}}}$$

Since the distribution depends on n, we talk of a t distribution *with n degrees of freedom*. Every statistics textbook carries a table of values for the cumulative distribution function of this family of distributions, and, of course, it is pre-programmed in your spreadsheet as well.

5.4 Exponential Distribution, and Distribution of a Sum of Exponentials

We have mentioned the exponential distribution before. A random variable has an exponential distribution, with parameter $\lambda > 0$, if it has (all equivalent definitions)

- density function $\lambda e^{-\lambda x}$
- cumulative distribution function $1 - e^{-\lambda x}$
- survival/reliability function $e^{-\lambda t}$

It turns out that, if X has exponential distribution with parameter λ (we often write $X \sim \text{EXP}(\lambda)$),

$$E[X] = \frac{1}{\lambda},\ \text{Var}[X] = \frac{1}{\lambda^2}$$

Note that, consequently,

$$\frac{E[X]}{\sqrt{\mathrm{Var}[X]}} = 1$$

It is remarkable that it also turns out that the distribution χ^2_2 turns out to be identical to $\mathrm{Exp}\left(\frac{1}{2}\right)$. Hence, the sum of n independent exponential variables with parameter $\frac{1}{2}$ is distributed like χ^2_{2n}. Since it is easy to show that if $X \sim \mathrm{EXP}(\lambda)$, kX has distribution defined by

$$P[kX > x] = P\left[X > \frac{x}{k}\right] = e^{-\frac{\lambda}{k}x}$$

that is, $\mathrm{EXP}\left[\frac{\lambda}{k}\right]$, it is easy to use the easily available tables and programs for the χ^2 distributions when studying the sum of independent identically distributed exponential random variables: if we are dealing with n independent variables X_k, each distributed like $\mathrm{EXP}(\lambda)$, the sum $\sum_{k=1}^{n} 2\lambda X_k$ is distributed according to χ^2_{2n}.

Remark 13. Both the family of χ^2, and t distributions are, in some way, connected to sum of n independent random variables. Hence, it is not too surprising that, for large values of n, these distributions tend to look more and more like the normal distribution (for the χ^2 case, we have to first subtract the mean, and divide by the standard deviation). If you go and check the tables of these distributions, you will notice that they list values up to some value of n (maybe 100 or 1000). For higher values, the difference with a normal distribution is negligible, and the table stops.

In practice, you would almost always use a spreadsheet (or a dedicated statistical package) to handle problems requiring use of any of these distributions. We won't go into details, as they are somewhat dependent on the specific tool used. An attentive reading of the manual or Help should be enough to figure out their use. Still, it is sometimes faster to adopt old-fashioned tools like tables for quick and dirty estimates.

6 Using Tables To Evaluate Probabilities

Tables, such as the ones we have in our course for the normal and related distributions, are useful even in the computer age, since they provide a bird's eye view of the probabilities associated to various events. Here are a few indications for the two tables we will use most: the table of the Standard Normal Distribution, and the table for the Student distributions (they are organized very differently). Tables for the χ^2 distributions are similar to those for the Student distributions.

6.1 Standard Normal Distribution

There are some differences between tables published in different places, but the basic idea is the same. Mainly, *make sure you look at the small graphic at the top, and its caption, to make sure you know exactly what the numbers in the table are referring to.* In some cases, tables list the probabilities that a standard normal random variable will take values between 0 and values x, $P[0 < X \leqslant x]$. In other cases, the table will give the cumulative distribution function, $\Phi(x) = P[X \leqslant x]$, or the survival function, $P[X > x]$. The table linked from the Map.html file has both of the last two choices as entries.

The peculiar way in which these probabilities are listed sometimes confuses people the first time they see these tables. Rather than giving an abstract description, let's find a few probabilities right away:

> Suppose you had a table of the first type (tabulating values for $P[0 < X \leqslant x]$), and wanted to know the probabilities
>
> 1. $P[0 < X < 1.82]$
> 2. $P[-1.53 < X < 2.15]$
> 3. $P[X < 2.5]$

4. $P[X > 1.17]$

We proceed as follows

1. We look up 1.8 in the leftmost (bold) column. Then we move on the corresponding row until we find ourselves under 0.02: the number we find is the probability we are looking for: 0.4656. In other words, it is something like an "addition table": each probability corresponds to the value you obtain by *adding* its row label with its column label.
2. Here we have to break the problem in two parts: $-1.53 < X < 0$, and $P[0 < X < 2.15]$. For the first, we note that it is equal to $P[0 < X < 1.53] = 0.4370$ (using our trick from #1), while the second is equal to 0.4394. Hence,
$$P[-1.53 < X < 2.15] = 0.4370 + 0.4394 = 0.8764$$
3. This is similar, except we have no finite left endpoint. That's fine: the probability corresponding to half the curve is 0.5, and we will add that to $P[0 < X < 2.5] = 0.4938$ (we looked at the intersection of the 2.5 with the 0.0 column). All in all, 0.9938.
4. This is the area to the *right* of 1.17. But that's exactly the area under the whole right half-curve minus the area from 0 to 1.17. The latter is 0.3790, hence, our answer is $0.5 - 0.3790 = 0.1210$

You can work out how to use of a different table layout for similar problems.

6.2 Student *t* Distribution

As mentioned, there are many t distributions: we need to know the number of "degrees of freedom". To simplify the task, tables for this distribution typically only give selected probabilities for each choice of n. Also, it does make little sense to list values for large values of n, because the difference from those for the standard normal distribution becomes smaller and smaller.

Again, you have to check carefully what exactly the probabilities are referring to: the tables may refer to the *cumulative distribution function* (that is to probabilities like $P[X \leqslant x]$, or to the *survival function* (probabilities like $P[X > x]$). Yet another possibility (this is often -the case in the pro-programmed functions in spreadsheets, by the way) is to have a "two-tailed" probability listed: for a (positive) number x, print the value of $P[-x \leqslant X \leqslant x]$ – the heading, as well as the numbers themselves, should make it clear. Once that is established, we work as follows.

We are focusing on the case when the table lists the cumulative distribution of Student distributions. The columns would be labeled $t_{.\mathrm{xx}}$, were the subscript is number between 0 and 1. Each row corresponds to a value of n, the number of *degrees of freedom*. The entry under $t_{.\mathrm{xx}}$, and n is a number, such that a variable distributed as a Student with n degrees of freedom, has probability .xx to be *less* than that number, if the table is built around the cumulative distribution function. For example:

- Column $t_{.75}$, $n = 5$: 0.727. If T_5 is a random variable with t_5 distribution, $P[T_5 < 0.727] = 0.75$
- Column $t_{.95}$, $n = 30$: 1.697. With a similar notation, $P[T_{30} < 1.697] = 0.95$

If the table is built around the survival function, we would have entries like

- Column $t_{0.25}, n = 5$: 0.727. If T_5 is a random variable with t_5 distribution, $P[T_5 > 0.727] = 0.25$
- Column $t_{.05}, n = 30$: 1.697. With a similar notation, $P[T_{30} > 1.697] = 0.05$

Notice how as n increases, the change in the numbers slows down, and the table becomes sparse (you can easily guess a reasonable interpolation for the missing degrees of freedom).