Introduction to Probability Theory and Statistics for Linguistics

Marcus Kracht
Department of Linguistics, UCLA
3125 Campbell Hall
405 Hilgard Avenue
Los Angeles, CA 90095–1543
kracht@humnet.ucla.edu

Contents

1 Preliminaries and Introduction

Some Useful Hints. This text provides a syllabus for the course. It is hopefully needless to state that the manuscript is not claimed to be in its final form, and I am constantly revising the text and it will grow as material gets added. Older material is subject to change without warning. One principle that I have adopted is that everything is explained and proved, unless the proof is too tedious or uses higher mathematics. This means that there will be a lot of stuff that is quite difficult for someone interested in practical applications. These passages are marked by S in the margin, so that you know where it is safe to skip. If you notice any inconsistencies or encounter difficulties in understanding the explanations, please let me know so I can improve the manuscript.

Statistics and probability theory are all about things that are not really certain. In everyday life this is the norm rather than the exception. Probability theory is the attempt to extract knowledge about what event has happened or will happen in presence of this uncertainty. It tries to quantify as best as possible the risks and benefits involved. Apart from the earliest applications of probability in gambling, numerous others exist: in science, where we make experiments and interpret them, in finance, in insurance and in weather reports. These are important areas where probabilities play a pivotal role. The present lectures will also give evidence for the fact that probability theory can be useful for linguistics, too. In everyday life we are frequently reminded of the fact that events that are predicted need not happen, even though we typically do not calculate probabilities. But in science this is absolutely necessary in order to obtain reliable result. Quantitative statements of this sort can sometimes be seen, for example in weather reports, where the experts speak of the "probability of rain" and give percentages rather than saying that rain is likely or unlikely, as one would ordinarily do. Some people believe that statistics requires new mathematics, as quantum mechanics required a new kind of physics. But this is not so. The ordinary mathematics is quite enough, in fact it has often been developed for the purpose of applying it to probability. However, as we shall see, probability is actually a difficult topic. Most of the naive intuitions we have on the subject matter are either (mathematically speaking) trivial or false, so we often have to resort to computations of some sort. Moreover, to apply the theory in a correct fashion, often two things are required: extensive motivation and a lot of calculations. I give an example. To say that an event happens with the probability $\frac{1}{6}$ means that it happens in 1 out of 6 cases. So if we throw a die six times we expect a given number, say 5, to appear once, and only once. This means that in a row of six, every one of the numbers occurs exactly once. But as we all know, this need not happen at all! This does not mean that the probabilities are wrong. In fact probability theory shows us that any six term sequence of numbers between 1 and 6 may occur. Any sequence is equally likely. However, one can calculate that for 5 to occur not at all is less likely than for it to occur once, and to occur a number $n > 1$ of times also is less likely. Thus, it is to be expected that the number of occurrences is 1. Some of the events are therefore more likely than others. But if that is so, throwing the die 60 times will not guarantee either that 5 occurs exactly 10 times. Again it may occur less often or more. How come then that we can at all be sure that the probabilities we have assigned to the outcomes are correct? The answer lies in the so called law of the large numbers. It says that if we repeat the experiment more often than the chance of the frequency of the number 5 deviating from its assigned probability gets smaller and smaller; in the limit it is zero. Thus, the probabilities are assumed *exactly* in the limit. Of course, since we cannot actually perform the experiment an infinite number of times there is no way we shall actually find out whether a given die is unbiased, but at least we know that we can remove doubts to any desirable degree of certainty. This is why statisticians express themselves in such a funny way, saying that something is certain (!) to occur with such and such probability or is likely to be the case with such and such degree of confidence. Finite experiments require this type of caution.

At this point it is actually useful to say something about the difference between *probability theory* and *statistics*. First, both of them are founded on the same model of reality. This means that they do not contradict each other, they just exploit that model for different purposes. The model is this: there is a certain space of events that occur more or less freely. These can be events the happen without us doing anything like “the sun is shining” or “there is a squirrel in the trashcan”. Or they can be brought about by us like “the coin shows tails” after we tossed it into the air. And, finally, it can be the result of a measurement, like “the voice onset time is 64 ms”. The model consists in a set of such events plus a so-called probability. We may picture this as an oracle that answers our question with “yes” or “no” each time we ask it. The questions we ask are predetermined, and the probabilities are the likelihood that is associated with a “yes” answer. This is a number between 0 and 1 which tells us how frequent that event is. *Data* is obtained by making an *experiment*. An experiment is in this scenario a question put to the oracle. An array of experiments yields data. Probability theory tells us how likely a particular data or set of data is.

In real life we do not have the probabilities, we have the data. And so we want tools that allow us to estimate the probabilities given the data that we have. This is what statistics is about. The difference is therefore merely what is known and what is not. In the case of an unbiased die we already have the probabilities; and so we can make predictions about a particular experiment or series thereof. In science it is the data we have and we want to know about the probabilities. If we study a particular construction, say tag questions, we want to know what the probability is that a speaker will use a tag question (as opposed to some other type of construction). Typically, the kind of result we want to find is even more complex. If, for example, we study the voice onset time of a particular sound, then we are interested to find a number or a range thereof. Statistics will help in the latter case, too, and we shall see how.

Thus, statistics is the art of guessing the model and its parameters. It is based on probability theory. Probability theory shows us why the particular formula by means of which we guess the model is good. For example, throw a die 100 times and notice how many times it shows 5. Let that number be 17. Then statistics tells you that you should guess the probability of 5 at 17/100 = .17. Probability tells you that although that might not be right, it is your best bet. What it will in fact prove is that if you assign any other probability to the outcome 5 then your experiment becomes less likely. This argument can be turned around. Probability theory tells you that the most likely probability assignment is .17. All this is wrapped up in the formula that the probability equals the frequency. And this is what you get told in statistics.

Literature. The mathematical background is covered in [5] and in [2]. Both texts are mathematically demanding. As for R, there is a nice textbook by Peter Dalgaard, himself a member of the R team, [1]. This book explains how to use R to do statistical analysis and is as such a somewhat better source than the R online help. In this manuscript I shall give a few hints as to how to use R, but I shall not actually introduce R nor do I intend to give a comprehensive reference. For that the book by Dalgaard is a good source and is recommended as complementary reading. For linguistic interests one may use [3]. There is a lot of more specialised literature which I shall point out in the sequel.

Part I

Basic Probability Theory

2 Counting and Numbers

We begin with a few very basic facts about counting elements in a set. We write $\mathbb{N}$ for the set of natural numbers. This set contains the numbers starting with 0. Thus

$$(1) \qquad \mathbb{N} = \{0, 1, 2, 3, 4, \dots\}$$

The **cardinality of a set** tells us how large that set is. If the set is finite, the cardinality is a natural number. We write $|A|$ for the cardinality of A. If B is also a set (not necessarily different) then $A \cup B$ is the set that contains the members of A and B. Since an element can belong to both but is only counted once we have

$$(2) \qquad |A \cup B| = |A| + |B| - |A \cap B|$$

The set $A \times B$ contains all pairs $\langle x, y\rangle$ such that $x \in A$ and $y \in B$. The set A^B contains all functions from B to A.

$$(3) \qquad |A \times B| = |A| \times |B|$$

$$(4) \qquad |A^B| = |A|^{|B|}$$

We say that A and B have the same cardinality if there is a one–to–one and onto function $f : A \to B$. Equivalently, it suffices to have functions $f : A \to B$ and $g : B \to A$ such that for all $x \in A$, $g(f(x)) = x$ and for all $y \in B$, $f(g(y)) = y$.

Theorem 1 $|\wp(A)| = 2^{|A|}$. *In other words, there are as many subsets of A as there are functions from A into a two–element set.*

Proof. For convenience, let $T = \{0, 1\}$. Clearly, $|T| = 2$. Let $X \subseteq A$. Then let $q(X)$ be the following function. $q(X)(u) = 1$ if $u \in X$ and $q(X)(u) = 0$ otherwise. Then $q(X) : A \to T$. Now let $g : A \to T$ be a function. Put $p(g) := \{u \in A : g(u) = 1\}$. Then $p(g) \subseteq A$. All we have to do is show that p and q are inverses of each other. (1) Let $X \subseteq A$. $p(q(X)) = \{u : q(X)(u) = 1\} = \{u : u \in X\} = X$. (2) Let $f : A \to T$. Then $q(p(f)) = q(\{u : f(u) = 1\})$. This is a function, and $q(p(f))(v) = 1$ iff $q(\{u : f(u) = 1\})(v) = 1$ iff $v \in \{u : f(u) = 1\}$ iff $f(v) = 1$. And $q(p(f))(v) = 0$ iff $f(v) = 0$ follows. Hence $f = q(p(f))$. ⊣

One of the most important kinds of numbers are the **binomial coefficients**. We shall give several equivalent characterisations and derive a formula to compute them.

Definition 2 *The number of k element subsets of an n element set is denoted by* $\binom{n}{k}$ *(pronounce: n choose k).*

We do not need to require $0 \leq k \leq n$ for this to be well-defined. In that case it is easily seen that the number is 0.

Theorem 3 *The following holds.*

① $\binom{n+1}{k+1} = \binom{n}{k} + \binom{n}{k+1}$.

② $\sum_{k=0}^{n} \binom{n}{k} = 2^n$.

Proof. Consider first $n = 1$. Here, if $k = 0$, $\binom{1}{1} = 1$, and $\binom{0}{1} + \binom{0}{0} = 0 + 1 = 1$, as promised. Now let A be an $n + 1$ element set, and let $a \in A$. Then $A - \{a\}$ is an n element set. Now choose a subset X of A of cardinality $k + 1$. (Case 1). $a \in X$. Then $X - \{a\}$ is an n element subset of $A - \{a\}$. Conversely, for every $H \subseteq A - \{a\}$ that has k elements the set $H \cup \{a\}$ has $k+1$ elements. (Case 2). $a \notin X$. Then X is a $k+1$ element subset of $A-\{a\}$. Conversely, every $k+1$ element subset of $A-\{a\}$ is a $k + 1$ element subset of A, and this finishes the proof. The second claim follows from the observation that the numbers are nonzero only when $0 \leq k \leq n$ and from the fact that the number of subsets of an n element set is 2^n. ⊣

Theorem 4 *For all complex numbers x and y and natural numbers n:*

$$(5) \qquad (x + y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$$

Proof. For $n = 1$, the claim is that

$$(6) \qquad x + y = (x + y)^1 = \binom{1}{0} x^0 y^1 + \binom{1}{1} x^1 y^0 = x + y$$

Now suppose the claim has been established for n. Then

$$\begin{aligned}
(x+y)^{n+1} &= (x+y)(x+y)^n \\
&= (x+y)\left(\sum_{k=0}^{n}\binom{n}{k}x^k y^{n-k}\right) \\
&= x\left(\sum_{k=0}^{n}\binom{n}{k}x^k y^{n-k}\right) + y\left(\sum_{k=0}^{n}\binom{n}{k}x^k y^{n-k}\right) \\
&= \sum_{k=0}^{n}\binom{n}{k}x^{k+1}y^{(n+1)-(k+1)} + \sum_{k=0}^{n}\binom{n}{k}x^k y^{(n+1)-k} \tag{7} \\
&= \left(\sum_{k=1}^{n+1}\left(\binom{n}{k} + \binom{n}{k-1}\right)x^k y^{(n+1)-k}\right) + y^{n+1} \\
&= \sum_{k=0}^{n+1}\binom{n+1}{k}x^k y^{(n+1)-k}
\end{aligned}$$

$\dashv$

This is a very important theorem. Notice that we can derive the second part of the previous theorem as follows. Put $x := y := 1$. Then we get

$$2^n = (1+1)^n = \sum_{k=0}^{n}\binom{n}{k} \tag{8}$$

Another formula that we get is this one. Put $x := 1$ and $y := -1$. Then

$$0 = (1-1)^n = \sum_{k=0}^{n}\binom{n}{k}(-1)^k \tag{9}$$

How to use R to do the calculations. Arithmetical expressions are written using +, * and so on. Help is provided in the R manual on how to do this. I shall only sketch a few useful tricks. R has a function `choose` which allows to calculate the binomials. For example, $\binom{23}{16}$ must be entered by `choose (23, 16)` (and you get the value 245157. If you want to assign that to a variable, say `x`, you will have to type

(10)

```
> x <- choose (23, 16)
```

and hit ‘enter’. (Here, > is the prompt; this is not what you type, it is already present on your screen. You type only what comes after the prompt.)

Now let us see how we can check the validity of Theorem 3. We want to calculate the sum of all $\binom{10}{i}$ where runs from 0 to 10. Write as follows:

(11)

```
> x <- 0
> for (i in 0:10) x <- x + choose (10, i)
> x
[1] 1024
```

This means the following. The variable `x` is assigned the value 0. Now, `i` is made to visit all values from 0 to 10 and each time add the result of $\binom{10}{i}$ to the value of `x`. Finally, when all is done we ask R for the value of `x`. The expression `0:10` is known as a **value range**. It denotes the sequence (!) of numbers 0, 1, 2, $\cdots$, 10. The start of the sequence is the left hand number, the end is given by the right hand number. The value range `0:10` is therefore distinct from `10:0`, which denotes the sequence 10, 9, 8, $\cdots$, 0.

There is an interesting way to get the same result. First we are going to create a vector of length 11 that contains the entries $\binom{10}{0}$, $\binom{10}{1}$, $\binom{10}{2}$ and so on. The way to do this is very easy:

(12)

```
> z <- choose (10, 0:10)
```

Next issue

(13)

```
> sum(z)
[1] 1024
```

and you get back the prompt. This last example shows the use of vectors and how to generate them with a formula. Generally, if you put a range in place of the variable it will create a vector containing the values of the formula with each member of the sequence inserted in turn. It you want to see the graphical shape of your vector you may type `plot (z)`. It opens a window and it gives you a graphical display of the values in ascending order.

To save graphics, here is a sample dialog. Save in a file the following data (as you see it) into a file, say falling.txt:

(14)

```
distance height
2 10
2.5 8
3.5 7
5 4
7 2
```

(There is no need to align the numbers). Now issue the following:

(15)

```
> d <- read.table("falling.txt",header=T)
```

The last argument is important since it declares that the first line contains the header (rather than being part of the data). Now you can do the following:

(16)

```
> pdf (file = "graph.pdf")
> plot (d)
> dev.off ()
```

This will cause .pdf to be stored in your temporary workspace. When you quit R you will be asked whether you want to save the workspace. If you enter ‘y’ then you will find a file called graph.pdf that contains your data. It is possible to use other device drivers (for example PostScript). Please read the manual for ”plot” as well as ”pdf” for further option. Also, get help on ”read.table” to find out how you can prepare your data to be read by R.

3 Some Background in Calculus

The most important notion in calculus is that of a *limit*. Consider the following two sequences:

(17) $1, 1/2, 1/3, 1/4, 1/5, \ldots$

(18) $1, -1, 1, -1, 1, -1, \ldots$

(19) $0, 1, 2, 3, 4, \ldots$

What separates the first from the second and third sequence is that the members of the sequence get closer and closer to 0 as time moves on. From time n on the distance to 0 is at most $1/n$ for any of the subsequent members. We call 0 the **limit** of the sequence. Indeed, it is even the case that the members get closer to each other because they all zoom in on the same value. This is to say that from a certain time point on every member is close to every other member of the sequence. In the present case this distance is again at most $1/n$. (This is the Cauchy-property and is equivalent to having a limit.) The second sequence is a little bit similar: here all the members of the sequence are either 1 or -1. Thus, the members of the sequence zoom in on two values, their distance among each other is sometimes 2 sometimes 0. Finally, the last sequence has no such property at all. The members of the sequence are not within a fixed corridor from each other. Sequences may show any mixture of these behaviours, but these three cases may be enough for our purposes.

Let us first look at a function $f : \mathbb{N} \to \mathbb{R}$. These functions are also called sequences and often written down in the manner shown above. This function is said to be **convergent** if for every ε there is a $n(\varepsilon)$ such that for all $n, n' \geq n(\varepsilon)$

(20) $|f(n) - f(n')| < \varepsilon$

If f is convergent there is a real number a such that the following holds: for all ε there is a $m(\varepsilon)$ such that if $n \geq n(\varepsilon)$ then

(21) $|f(n) - a| < \varepsilon$

This number is called the **limit** of the function f, and we write

(22) $a = \lim_{n\to\infty} f(n)$

Now suppose that f is a function from real numbers to real numbers. We shall use the notion of limit to define continuity of a function. There is a definition which runs as follows: f is continuous in x_0 if, given any sequence $(x_n)_{n\in\mathbb{N}}$ with limit x_0 the sequence of values $(f(x_n))_{n\in\mathbb{N}}$ also is convergent with limit $f(x_0)$. Notice that the sequence $(x_n)_{n\in\mathbb{N}}$ need not contain x_0 itself, but must get infinitely close to it. The previous definition is somewhat cumbersome to use. Here is a better one. We say that f is **continuous** at x_0 if for every $\varepsilon > 0$ there is a $\delta(\varepsilon) > 0$ such that if $|x - x_0| < \delta(\varepsilon)$ and $|y - x_0| < \delta(\varepsilon)$ then

$$(23) \qquad |f(x) - f(y)| < \varepsilon$$

The nice aspect of this definition is that f need not be defined at x_0. There is again exactly one value that f can be given at x_0 to make it continuous, and that value is called the **limit** of f in x_0:

$$(24) \qquad \lim_{x\to x_0} f(x)$$

Now if f is continuous in a point then we can know its value at that point if we study the values at points close to it. To be really exact, if we want to make an error of at most ε we should study values at distance at most $\delta(\varepsilon)$. If one wants to suppress explicit mention of the dependency, one write as follows.

$$(25) \qquad f(x_0 + \Delta) \approx f(x_0)$$

And this means that the error becomes small if Δ is *small enough*. Δ is a real number, any number, but preferably small. Now imagine a number dx_0 that is so small that it is not zero but smaller than every real number > 0. In physics this is known as a *virtual number*, in mathematics we call them **infinitesimals**. A number greater than zero but less than every real number $r > 0$ is called infinitesimal. Although you may find this intuitive, at second though you may find this contradictory. Do not worry: mathematician used to think that this is impossible, but it has been shown to be consistent! What rebels in you is only the thought that the line that you see has no place for them. But who knows? Now, write $\approx$ to say that the numbers are different only by an infinitesimal amount.

$$(26) \qquad f(x_0 + dx_0) \approx f(x_0)$$

The two are however different, and the difference is very small, in fact smaller than every real > 0. This number is denoted by $(df)(x_0)$.

$$(27) \qquad (df)(x_0) := f(x_0 + dx_0) - f(x_0)$$

We shall use this definition to define the derivative. For a given x_0 and $\Delta \neq 0$ define

$$g(\Delta) := \frac{f(x_0 + \Delta) - f(x_0)}{\Delta} \tag{28}$$

If f is not everywhere defined, we need to make exceptions for Δ, but we assume that f is defined in at least some open interval around x_0. The function g as given is defined on that interval except for Δ. We say that f has a **derivative** in x_0 if g is continuous in x_0, and we set

$$f'(x_0) := \lim_{\Delta \to 0} g(\Delta) \tag{29}$$

To be precise here: the notion of limit applies to real valued functions and yields a real value. Now, the same can be done at any other point at which f is defined, and so we get a function $f' : \mathbb{R} \to \mathbb{R}$, called the **derivative** of f. Another notation for f' is $\frac{df}{dx}$. The latter will become very useful. There is a useful intuition about derivatives. f has a derivative at x_0 if it can be written as

$$f(x_0 + \Delta) \approx f(x_0) + \Delta f'(x_0) \tag{30}$$

where the error has size Δ. For $\approx$ we may ignore the error if infinitesimally small, write

$$f(x_0 + dx_0) = f(x_0) + f'(x_0)dx_0 \tag{31}$$

Sticking this into (27) we get

$$f'(x_0) = \frac{(df)(x_0)}{dx_0} =: \frac{df}{dx}(x_0) \tag{32}$$

where now we have inserted a function that yields very, very small values, namely df. However, when divided by another function that also yields very, very small values, the quotient may actually become a real number again!

We can do real calculations using that, pretending dx to be a number.

$$\frac{d(x^2)}{dx} = \frac{(x + dx)^2 - x^2}{dx} = \frac{2xdx - (dx)(dx)}{dx} = 2x + dx \tag{33}$$

If you want to know the real number that this gives you, just ignore the addition of dx. It's just $2x$. To be exact, we may write

$$\frac{d(x^2)}{dx} \approx 2x \tag{34}$$

Indeed, the latter is a real valued function, so it is our candidate for the derivative (which after all is a real valued function again). Hence, the derivative of the function $f(x) = x^2$ at x_0 is the function $2x_0$. In this vein one can deduce that the derivative of x^n is nx^{n-1}.

However, we can use this also for very abstract calculations. Suppose that you have a function $f(g(x))$. You take x and apply g and then apply f. Now, let us calculate:

$$f(g(x + dx)) = f(g(x) + (dg)(x)) = f(g(x)) + (df)((dg)(x)) \tag{35}$$

It follows that (suppressing the variable)

$$d(f \circ g) = (df) \cdot (dg) \tag{36}$$

This is often given the following form.

$$\frac{df}{dx} = \frac{df}{dy}\frac{dy}{dx} \tag{37}$$

Here, y is an arbitrary variable. In our case, $y = g(x)$, so y depends in value on x. Understood as a fraction of numbers this is just an instance of ordinary expansion of fractions.

Here is another useful application.

$$\begin{aligned} d(f + g) &= (f + g)(x + dx) - (f + g)(x) \\ &= (f(x + dx) - f(x)) + (g(x + dx) - g(x)) \\ &= (df)(x)dx + (dg)(x)dx \end{aligned} \tag{38}$$

From this we derive

$$\frac{d(f + g)}{dx} = \frac{df}{dx} + \frac{dg}{dx} \tag{39}$$

$$\begin{aligned} d(fg) &= (fg)(x + dx) - (f + g)(x) \\ &= f(x + dx)g(x + dx) - f(x)g(x) \\ &= ((f(x) + (df)(x)dx)(g(x) + (dg)(x)) - f(x)g(x) \\ &= f(x)g(x) + (df)(x)g(x)dx + f(x)dg(x)dx \\ &\qquad + (df)(x)(dg)(x)dxdx - f(x)g(x) \\ &= ((df)(x)g(x) + f(x)(dg)(x))dx + (df)(x)(dg)(x)(dx)^2 \end{aligned} \tag{40}$$

We get that

$$\frac{d(fg)}{dx} = (df)g + f(dg) + (df)(dg)dx \quad = (df)g + f(dg) \tag{41}$$

Recall that the derivative is a real valued function, so we may eventually ignore the infinitesimally small $(df)(dg)dx$. We derive a last consequence. Suppose that $f(g(x)) = x$; in other words, f is the inverse of g. Then, taking derivatives, we get

$$\frac{df}{dg}\frac{dg}{dx} = 1 \tag{42}$$

We are interested in the derivative of f as a function of a variable, say y. y is implicitly defined (via g). In fact, $dy = dg$. Then, multiplying

$$\frac{d(g^{-1}}{dy} = \left(\frac{dg}{dx}\right)^{-1} \tag{43}$$

For example, the derivative of e^x is e^x. The inverse of this function is $\ln y$. Thus

$$\frac{d(\ln y}{dy} = \left(\frac{d(e^x)}{dx}\right)^{-1} = \frac{1}{e^x} = \frac{1}{y} \tag{44}$$

Theorem 5 *The following laws holds for the derivatives:*

1. $\frac{d(x^\alpha)}{dx} = \alpha x^{\alpha-1}$, $\alpha \in \mathbb{R}$.
2. $\frac{d(f+g)}{dx} = \frac{df}{dx} + \frac{dg}{dx}$.
3. $\frac{d(fg)}{dx} = f\frac{dg}{dx} + g\frac{df}{dx}$.
4. $\frac{d(f\circ g)}{dx} = \frac{df}{dg} \cdot \frac{dg}{dx}$.
5. $\frac{d(g^{-1})}{dx} = \left(\frac{dg}{dx}\right)^{-1}$.
6. $\frac{d(e^x)}{dx} = e^x$.
7. $\frac{d(\sin x)}{dx} = \cos x$.

This is as much as we need in the sequel. We shall look at the last two claims in particular. First, notice that

$$d(e^x) = e^{x+dx} - e^x = e^x(e^{dx} - 1) \tag{45}$$

Thus we establish that

$$\frac{d(e^x)}{dx} = e^x \frac{e^{dx} - 1}{dx} \tag{46}$$

It is actually the definition of e that

$$\lim_{\Delta \to 0} \frac{e^{\Delta} - 1}{\Delta} = 1 \tag{47}$$

This settles the claim as follows.

$$\frac{e^{dx} - 1}{dx} \approx \lim_{\Delta \to 0} \frac{e^{\Delta} - 1}{\Delta} = 1 \tag{48}$$

Now, using complex numbers, notice that $e^{ix} = \cos x + i \sin x$ so that $\sin x = (e^{ix} - e^{-ix})/2i$. This gives

$$\begin{aligned}
\frac{d(\sin x)}{dx} &= \frac{d((e^{ix} - e^{-ix})/2)}{dx} \\
&= \frac{1}{2i}\left(ie^{ix} - (-i)e^{i(-x)}\right) \\
&= \frac{1}{2i}\left(i(\cos x + i \sin x) + i(\cos(-x) + i \sin(-x))\right) \\
&= \frac{1}{2i}\left(2i \cos x\right) \\
&= \cos x
\end{aligned} \tag{49}$$

We have used $\cos(-x) = \cos x$ and $\sin(-x) = -\sin x$.

Now we turn to integration. Integration is a technique to calculate the area beneath some graph of a function. Before we approach the problem of integration we shall first look at the notion of area. A **measure** is a function μ that assigns to subsets of a space some real number, called the **measure** of that set. Not every set needs to have a measure. There are three conditions we wish to impose on μ in order to qualify for a measure.

1. $\mu(\emptyset) = 0$.
2. If $A \cap B = \emptyset$ then $\mu(A \cup B) = \mu(A) + \mu(B)$.
3. If A_n, $n \in \mathbb{N}$, are pairwise disjoint, then

$$\mu(\bigcup_{n\in\mathbb{N}} A_n) = \sum_{n\in\mathbb{N}} \mu(A_n)$$

We can for example define the following measure on sets of real numbers. For an interval $[a, b]$ with $a \leq b$ the measure is $b - a$. In particular, if $a = b$ we get $\mu(\{a\}) = \mu([a, a]) = 0$. It follows that every finite set has measure 0, and so does every set that can be enumerated with the natural numbers. But if an interval $[a, b]$ is not a singleton it actually has more members than can be enumerated! Now, it follows that every finite or countable disjoint union of intervals has a measure. These sets are called the **Borel sets of finite measure**. Another characterization is as follows.

Definition 6 *Let $\mathcal{B}$ be the least set of subsets of $\mathbb{R}$ which contains the finite intervals and is closed under complement and countable unions.*

The set $(-\infty, y]$ is a Borel set but its measure is infinite. The rational numbers are also Borel, and of measure 0.

The same definition of Borel set is defined for $\mathbb{R}^n$. We do this for $n = 2$. We start with rectangles of the form $[a_1, b_1] \times [a_2, b_2]$ which we declare to be sets of measure $(b_1 - a_1)(b_2 - a_2)$ and then close under infinite unions and complement. However, for purposes of integration we define the measure as follows. In the following definition, $b_1 \leq 0 \leq b_2$.

$$\mu([a_1, b_1] \times [0, b_2]) := b_2(b_1 - a_1) \tag{50}$$

$$\mu([a_1, b_1] \times [b_1, 0]) := b_1(b_1 - a_1) \tag{51}$$

Thus, the interval $[3, 4] \times [0, 2]$ has measure 2, while $[3, 4] \times [-2, 0]$ has measure -2. To see why we need this odd looking definition, take a look at integration. Suppose you are integrating the function $f(x) = -2$ between 3 and 4. Instead of the value 2 (which would give you the area) it is actually defined to be -2. The area below the line counts negatively for the integration.

Once we have the definition of this measure, we can define as follows:

$$\int_{x=a}^{b} f(x)dx = \mu(\{\langle x, y\rangle : a \leq x \leq b \text{ and: } 0 \leq y \leq f(x) \text{ or } f(x) \leq y \leq 0\}) \tag{52}$$

We abbreviate the set to the right by $[f(x)]_a^b$. This definition has a number of immediate consequences. First,

$$\int_{a}^{b} f(x)dx = \int_{a}^{c} f(x)dx + \int_{c}^{b} f(x)dx \tag{53}$$

For a proof, notice that $\mu([f(x)]_c^c) = f(c) \cdot 0 = 0$, by definition. Now, notice that the measure is additive:

$$\begin{aligned} \mu([f(x)]_a^b) &= \mu([f(x)]_a^c \cup [f(x)]_c^b) \\ &= \mu([f(x)]_a^c \cup ([f(x)]_c^b - [f(x)]_c^c) \\ &= \mu([f(x)]_a^c) + \mu([f(x)]_c^b - [f(x)]_c^c) \\ &= \mu([f(x)]_a^c) + \mu([f(x)]_c^b) \end{aligned} \tag{54}$$

Another important consequence is that integration is the inverse of differentiation:

$$\begin{aligned} d\left(\int_{a}^{x} f(y)dy\right) &= \int_{a}^{x+dx} f(y)dy - \int_{a}^{x} f(y)dy \\ &= \int_{x}^{x+dx} f(y)dy \end{aligned} \tag{55}$$

Assuming that f is continuous, the values of f in the interval are only infinitesimally apart from $f(x)$, so we may assume that they all lie in an interval $[f(x) - adx, f(x) + adx]$, a a real number. Then we have

$$dx(f(x) - adx) \leq \int_{x}^{x+dx} f(y)dy \leq dx(f(x) + adx) \tag{56}$$

Now we have

$$f(x) - adx \leq \frac{d(\int_a^x f(y)dy)}{dx} \leq f(x) + adx \tag{57}$$

The left and right are only infinitesimally apart, so in terms of real numbers they are equal. We conclude

$$\frac{d(\int_a^x f(y)dy)}{dx} = f(x) \tag{58}$$

This greatly simplifies integration, since differentiation is on the whole easier to do, and once a function f is known to be a derivative of some other function g, the integral of f is known to be g (plus a constant) as well. We derive a particular consequence, the rule of **partial integration**. Consider having to calculate the integral $\int_a^b fgdx$. Suppose we know how to integrate f but not g. Then we can proceed as follows. Suppose that $\int_a^x f(x)dx = h(x)$. Then

$$\int_a^b f(x)g(x)dx = h(x)g(x)|_a^b - \int_a^b h(x)g'(x)dx \tag{59}$$

For a proof, just take derivatives on both sides:

$$f(x)g(x) = \frac{d}{dx}\left(\int_a^x f(y)g(y)dy\right) \tag{60}$$

$$\begin{aligned} &\frac{d}{dx}\left(h(y)g(y)|_a^x - \int_a^x h(y)g'(y)dy\right) = \quad \frac{d}{dx}(h(x)g(x)) - h(x)g'(x)) \\ =&f(x)g(x) + h(x)g'(x) - h(x)g'(x) \\ =&f(x)g(x) \end{aligned} \tag{61}$$

4 Probability Spaces

The definition of probability spaces is somewhat involved. Before we can give it in its full generality, let us notice some special cases. Intuitively, certain events happen with some probability. If I throw a coin then it will either show heads or tails. Moreover, we assume that the probability with which heads comes up is $\frac{1}{2}$. This means that we expect that half of the time we get heads and half of the time we get tails. Similarly, throwing a die we have six different outcomes and we expect each of them to be occur as often as the others; throwing two dice, a green and a red one, each outcome where the green die shows some number i and the red die a number j is equally likely. There are 36 such outcomes. Each one therefore has probability $\frac{1}{36}$. To give yet another example, suppose that we throw two dice, but that our outcomes are now the sum of the points we have thrown. These are the numbers 2, 3, . . ., 12, but the probabilities are now

(62)
$$\begin{array}{llllll} p(2) & = \frac{1}{36} & p(3) & = \frac{2}{36} & p(4) & = \frac{3}{36} \\ p(5) & = \frac{4}{36} & p(6) & = \frac{5}{36} & p(7) & = \frac{6}{36} \\ p(8) & = \frac{5}{36} & p(9) & = \frac{4}{36} & p(10) & = \frac{3}{36} \\ p(11) & = \frac{2}{36} & p(12) & = \frac{1}{36} & & \end{array}$$

You may check that these probabilities sum to 1. This time we have actually *not* assumed that all outcomes are equally likely. Why is that so? For an explanation, we point to the fact that the events we look at are actually derived from the previous example. For notice that there are five possibilities to throw a sum of 6. Namely, the sum is six if the first die shows 1 and the second 5, the first shows 2 and the second 4, and so on. Each possibility occurs with the probability $\frac{1}{36}$, and the total is therefore $\frac{5}{36}$. This example will occur later on again.

Let us return to a single die. There are six outcomes, denoted by the numbers from 1 to 6. In addition to them there are what we call *events*. These are certain sets of outcomes. For example, there is an event of throwing an even number. The latter comprises three outcomes: 2, 4 and 6. We expect that its probability is exactly half, since the probability of each of the outcomes is exactly $\frac{1}{6}$. In the language of probability theory, **events** are sets of outcomes. The probability of an event is the sum of the probabilities of the outcomes that it contains. Thus, in the finite case we get a set Ω of outcomes, and a function $p : \Omega \to [0, 1]$ such that

(63)
$$\sum_{\omega\in\Omega} p(\omega) = 1$$

For a subset $A \subseteq \Omega$ we put

$$(64) \qquad P(A) := \sum_{\omega \in A} p(\omega)$$

Notice that we have used a different letter here, namely P. We say that P is the **probability function** and that p is its **density** or **distribution**. The latter is applied only to individual events, while P is applied to sets of outcomes. If ω is an outcome, then $P(\{x\}) = p(x)$. Since there is no risk of confusion, one also writes $P(x)$ in place of $P(\{x\})$. From this definition we can derive a few laws.

$$(65) \qquad P(\varnothing) = 0$$

$$(66) \qquad P(\Omega) = 1$$

$$(67) \qquad P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

The first is clear: if we sum over an empty set we get 0. The second is also clear, it follows from (63). The third needs proof. Perhaps it is easier to start with a different observation. Suppose that $A \cap B = \varnothing$. Then

$$(68) \qquad P(A + B) = \sum_{\omega \in A \cup B} p(\omega) = \sum_{\omega \in A} p(\omega) + \sum_{\omega \in B} p(\omega) = P(A) + P(B)$$

Now if $A \cap B \neq \varnothing$ notice that $A \cup B = A \cup (B - A)$, and the two sets are disjoint. So $P(A \cup B) = P(A) + P(B - A)$. Also, $B = (B - A) \cup (B \cap A)$, with the sets disjoint, and this gives $P(B) = P(B - A) + P(A \cap B)$. Together this yields the formula $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. Finally, we note that if A_i, $1 \leq i \leq n$, are pairwise disjoint sets then

$$(69) \qquad P(\bigcup_{i=1}^{n} A_i) = P(A_1) + P(A_2) + \cdots + P(A_n)$$

This is in fact all that needs to be said if Ω is finite.

However, when Ω is infinite we need to look harder. Suppose, namely, that Ω is the set of natural numbers. Suppose further that each of the numbers is equally probable. Then the probability of each number is actually 0. However, this means that the probability of every subset of the numbers is 0 as well. So, the approach of assigning probabilities to outcomes fails; instead, in probability theory one does not assign probabilities to outcomes but rather to events. For example, if every number has the same probability, the set of even numbers has the probability $\frac{1}{2}$, likewise the set of odd numbers. This means that the probability that a randomly chosen natural number is even is $\frac{1}{2}$, while the probability that it is equal to 377 is 0. In order to make the approach work out correctly we need to restrict the domain of the function that assigns the probabilities. We shall require that the sets that receive probabilities form a boolean algebra. This is a bit more general than typically assumed, where additionally it is required that they are closed under intersection and union over countably many sets (such algebras are called σ–**algebras**).

To give a nontrivial example, let $U(k, n)$ be the set of numbers that leave the remainder k when divided by n. Give probability $\frac{1}{n}$ to these sets. Let $\mathfrak{A}$ be the set of all finite unions and intersections of the sets $U(n, k)$ for all $n \in \mathbb{N}$ and $k < n$. This is a boolean algebra. It is closed under the required operations.

Definition 7 *A* ***probability space*** *is a triple* $\langle \Omega, \mathfrak{A}, P\rangle$*, where* Ω *is a set, the set of* ***outcomes****,* $\mathfrak{A} \subseteq \wp(\Omega)$ *a boolean algebra, the algebra of* **events** *and* $P : \mathfrak{A} \to [0, 1]$ *a function satisfying the following.*

1. $P(\varnothing) = 0$

2. $P(\Omega) = 1,$

3. *If* A_i*,* $i \in I$*, are pairwise disjoint sets and* $|I| \leq \omega$ *then*

$$P(\bigcup_{i\in I} A_i) = \sum_{i\in I} P(A_i)$$

A note on notation. $\mathfrak{A}$ shall always be an algebra of sets over Ω. Thus, we shall not write $0_{\mathfrak{A}}$ but rather $\varnothing$. The operations on $\mathfrak{A}$ are union ($\cup$), intersection ($\cap$) and relative complement ($-$). In particular, notice that $-A = \Omega - A$.

We give some examples.

The Laplace space. Let Ω be a finite set containing n elements. $\mathfrak{A} := \wp(\Omega)$. Finally, put $P(A) := \frac{|A|}{n}$. In this space, every outcome has the same probability, namely $\frac{1}{n}$. The above examples (tossing an unbiased coin, throwing a die) are of this form.

The Bernoulli space. Let $\Omega = \{0, 1\}$, $\mathfrak{A} = \wp(\Omega)$. With $p := p(1)$, we have $q := p(0) = 1 - p$. Here 1 represents the event of "success", while 0 represents failure. For example, betting schemes work as follows. Team A plays against Team B in a match. Team A is expected to win with probability 75%, or 0.75. Hence $p = 0.75$ and $q = 0.25$. The Bernoulli space is the smallest of all probability spaces, but its usefulness in probability theory can hardly be overestimated. One says that this is the probabilities of tossing a "biased coin", where the bias is q/p against you. If the coin is actually unbiased, then $p = q = 1/2$ so that the bias is 1. In the above example the bias is $0.75/0.25 = 3$. Indeed, if the betting office is convinced that the odds are 3:1 that A wins and you are betting ten dollars that B wins instead, it will offer to pay (at most) 40 dollars to you if B wins while cashing your money when it loses. If the betting office is not trying to make money then it will pay exactly 40 dollars. In general, if the odds are $r : 1$, then for every dollar you bet against the event you get $r + 1$ if you win and nothing otherwise. To see that this is fair, notice that on average you win 1 out of $r + 1$ (!) times, you get back the sum you placed and additionally win r dollars for every dollar, while you lose and then lose the dollars you placed on the bet. This scheme will in the long run make no one richer than he was originally (on average). The proof for this will have to be given. But intuitively it is clear that this is the case. Notice that betting offices do aim at making money, so they will make sure that you lose in the long run. To give an example, in French roulette there are 37 numbers (from 0 to 36). 0 plays a special role. If you bet 10 dollars on a number different from 0, say 15, then you get paid 360 dollars if 15 shows, while you get nothing when it doesn't. This is slightly less than the actual odds (which are 36:1, which means that you should get 370 dollars), to make sure that the casino is on average making money from its customers.

Discrete spaces. A space is **discrete** if $\mathfrak{A} = \wp(\Omega)$. So, every conceivable set is an event, and therefore has a probability assigned to it. In particular, if x is an outcome, then $\{x\}$ is an event (these two things are often confused). Therefore, we may put $p(x) := P(\{x\})$. Then we get

$$(70) \qquad P(A) = \sum_{\omega \in A} p(\omega)$$

This however is only defined if Ω is either finite or countably infinite. However, we shall not encounter spaces that are not countably infinite, so this is as general as we need it.

We shall present some abstract constructions for probability spaces. Let $f : X \to Y$ be a function, and $U \subseteq X$ and $V \subseteq Y$. Then put

(71) $$f[U] := \{h(x) : x \in U\}$$

(72) $$f^{-1}[V] := \{x \in U : f(x) \in V\}$$

$f[U]$ is called the **direct image** of U under f, and $f^{-1}[V]$ the **preimage** of V under f. Now, if $\mathfrak{B} \subseteq \wp(V)$ is a boolean algebra, so is $\{f^{-1}[B] : B \in \mathfrak{B}\}$. Here is a proof. (1) $f^{-1}[\varnothing] = \varnothing$, (2) $f^{-1}[V \cup W] = f^{-1}[V] \cup f^{-1}[W]$. For $x \in f^{-1}[V \cup W]$ iff $f(x) \in V \cup W$ iff $f(x) \in V$ or $f(x) \in W$ iff $x \in f^{-1}[V]$ or $x \in f^{-1}[W]$. (3) $f^{-1}[Y - V] = X - f^{-1}[V]$. For $x \in f^{-1}[Y - V]$ iff $f(x) \in Y - V$ iff $f(x) \in Y$ and $f(x) \notin V$ iff $x \in X$ and not $x \in f^{-1}[V]$ iff $x \in X - f^{-1}[V]$. (4) $f^{-1}[V \cap W] = f^{-1}[V] \cap f^{-1}[W]$. Follows from (2) and (3), but can be proved directly as well.

Now suppose we have a probability function $P : \mathfrak{A} \to [0, 1]$. Then we can only assign a probability to a set in $\mathfrak{B}$ if its full preimage is in $\mathfrak{A}$. Thus, we call $f : \Omega \to \Omega'$ **compatible with** $\mathfrak{A}$ if $f^{-1}[B] \in \mathfrak{A}$ for all $B \in \mathfrak{B}$. In this case every set in $\mathfrak{B}$ can be assigned a probability by

(73) $$P'(B) := P(f^{-1}[B])$$

This is a probability function: (1) $P'(\Omega') = P(f^{-1}[\Omega']) = P(\Omega) = 1$, (2) $P'(\varnothing) = P(f^{-1}[\varnothing]) = P(\varnothing) = 0$, (3) If A and B are disjoint, so are $f^{-1}[A]$ and $f^{-1}[B]$, and then $P'(A \cup B) = P(f^{-1}[A \cup B]) = P(f^{-1}[A] \cup f^{-1}[B]) = P(f^{-1}[A]) + P(f^{-1}[B]) = P'(A) + P'(B)$.

Proposition 8 *Let $\langle \Omega, \mathfrak{A}, P \rangle$ be a finite probability space, $\mathfrak{B}$ a boolean algebra over Ω' and $f : \Omega \to \Omega'$ a surjective function compatible with $\mathfrak{A}$. Put $P'(B) := P(f^{-1}[A])$. Then $\langle \Omega', \mathfrak{A}', P' \rangle$ is a probability space.*

We shall put this to use as follows. $\mathfrak{A}$ is finite and has atoms A_1, ..., A_n. Then let $\Omega' = \{1, \dots, n\}$ and define f by $f(x) := i$, where $x \in A_i$. This is well defined: every element is contained in one and only one atom of the algebra. This function is compatible with $\mathfrak{A}$. For let $S \subseteq \Omega'$. Then

(74) $$f^{-1}[S] = \bigcup_{i \in S} A_i \in \mathfrak{A}$$

Here is an example. Suppose that $\Omega = \{1, 2, 3, 4, 5, 6\}$, and that

(75) $$\mathfrak{A} = \{\varnothing, \{1, 2\}, \{3, 4, 5, 6\}, \Omega\}.$$

This is a boolean algebra, and it has two atoms, $\{1,2\}$ and $\{3,4,5,6\}$. Now define $f(1) := f(2) := \alpha$ and $f(3) := f(4) := f(5) := f(6) := \beta$. Then $f(\varnothing) = \varnothing$, $f(\{1,2\}) = \{\alpha\}$, $f(\{3,4,5,6\}) = \{\beta\}$, and $f(\Omega) = \{\alpha,\beta\}$. Finally, assume the following probabilities: $P(\{1,2\}) = \frac{1}{3}$, $P(\{3,4,5,6\}) = \frac{2}{3}$. Then we put $P'(\{\alpha\}) := \frac{1}{3}$ and $P'(\{\beta\}) = \frac{2}{3}$. Notice that the original space was drawn from a simple Laplace experiment: throwing a die, where each outcome has equal probability. However, we considered only four events, with the appropriate probabilities given. The resulting space can be mapped onto a Bernoulli space with $p = \frac{1}{3}$.

There is an immediate corollary of this. Say that $\langle\Omega, \mathfrak{A}, P\rangle$ is **reducible** to $\langle\Omega', \mathfrak{A}', P'\rangle$ if there is a function $f : \Omega \to \Omega'$ such that $\mathfrak{A} = \{f^{-1}[B] : B \in \mathfrak{A}'\}$ and $P'(B) = P(f^{-1}[B])$ for all $B \in \mathfrak{A}'$. Thus the second space has perhaps less outcomes, but it has (up to isomorphism) the same event structure and the same probability assignment.

Proposition 9 *Every finite probability space is reducible to a discrete probability space.*

This means that in the finite setting it does not make much sense to consider anything but discrete probability spaces. But the abstract theory is nevertheless to be preferred for the flexibility that it gives.

Next we look at another frequent situation. Let Ω_1 and Ω_2 be sets of outcomes of experiments E_1 and E_2, respectively. Then $\Omega_1 \times \Omega_2$ is the set of outcomes of the experiment where both E_1 and E_2 are conducted. For example, suppose we are tossing a coin and throw a die. Then the outcomes are pairs $\langle\ell, m\rangle$ where $\ell \in \{\mathrm{H}, \mathrm{T}\}$ and $m \in \{1,2,3,4,5,6\}$. Now, what sort of event do we have to consider? If $A_1 \in \mathfrak{A}_1$ and $A_2 \in \mathfrak{A}_2$ we would like to have the event $A_1 \times A_2$. However, one can show that the set of these event is not a boolean algebra since it is in general not closed under negation and union. To take an easy example, let $\Omega_1 = \Omega_2 = \{0,1\}$ and $\mathfrak{A}_1 = \mathfrak{A}_2 = \wp(\{0,1\})$. The set $\{\langle 0,1\rangle, \langle 1,0\rangle\}$ is the union of the two sets $\{0\} \times \{1\} = \{\langle 0,1\rangle\}$ and $\{1\} \times \{0\} = \{\langle 1,0\rangle\}$. But it is not of the form $A \times B$ for any A, B.

Instead we simply take all finite unions of such sets:

$$\text{(76)} \qquad \mathfrak{A}_1 \otimes \mathfrak{A}_2 := \left\{ \bigcup_{i=1}^{p} A_i \times B_i : \text{ for all } i\text{: } A_i \in \mathfrak{A}_1, B_i \in \mathfrak{A}_2 \right\}$$

The probabilities are assigned as follows.

(77) $$(P_1 \times P_2)(A \times B) := P_1(A) \cdot P_2(B)$$

It is a bit tricky to show that this defines probabilities for all sets of the new algebra. The reason is that in order to extend this to unions of these sets we need to make sure that we can always use disjoint unions. We perform this in case we have a union of two sets. Notice that $(A \times B) \cap (A' \cap B') = (A \cap A') \times (B \cap B')$. Using this and (67) we have

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

This is reminiscent of the fact that the intersection of two rectangles is a rectangles. So, if we take the sum of the probabilities we are counting the probability of the intersection twice. The latter is a rectangle again.

The probabilities of the right hand side are defined; so is therefore the one to the left.

Definition 10 *Let $\mathcal{P}_1 = \langle \Omega_1, \mathfrak{A}_1, P_1 \rangle$ and $\mathcal{P}_2 = \langle \Omega_2, \mathfrak{A}_2, P_2 \rangle$ be probability spaces. Then $\mathcal{P}_1 \otimes \mathcal{P}_2 := \langle \Omega_1 \times \Omega_2, \mathfrak{A}_1 \otimes \mathfrak{A}_2, P_1 \times P_2 \rangle$ is a probability space, the so–called **product space**.*

We give an application. Take a Bernoulli experiment with $p = 0.6$. This defines the space $\mathcal{P}$. Suppose we do this experiment twice. This we can alternatively view as a single experiment, where the probability space is now $\mathcal{P} \otimes \mathcal{P}$. It is not hard to verify that the algebra of events is the powerset $\wp(\{0, 1\} \times \{0, 1\})$. Also, we have

(79) $$p(\langle 0, 0 \rangle) = 0.36, p(\langle 0, 1 \rangle) = p(\langle 1, 0 \rangle) = 0.24, p(\langle 1, 1 \rangle) = 0.16$$

The probabilities sum to 1 as is readily checked.

5 Conditional Probability

Suppose that some person A has three children, and suppose that the probability of a child being a boy is simply $\frac{1}{2}$. The probability of having exactly one boy and therefore two girls is $\frac{3}{8}$. Now suppose you know that A has at least one girl, what is the probability that he has exactly one boy? The probability cannot be the same again. To see this, let us note that the probability of having three sons is zero under this condition. If we didn't know there was at least one girl, the probability would have been $\frac{1}{8}$. So, some probabilities have obviously changed. Let us do the computation then. The probability of having at least one girl is $\frac{7}{8}$. The probability of having exactly one boy is $\frac{3}{8}$. If there is exactly one boy there also is at least one girl, so the probabilities compare to each other as 3:7. Thus we expect that the probability of having exactly one boy on condition of having at least one girl is $\frac{3}{7}$. How exactly did we get there? Let us consider an event A and ask what its probability is on condition that B. There are in total four cases to consider. A may or may not be the case, and B may or may not be the case. However, as we have excluded that B fails to be the case, we have effectively reduced the space of possibilities to those in which B holds. Here the odds are $P(A \cap B) : P((-A) \cap B)$. Thus the probability that A holds on condition that B, denoted by $P(A|B)$ is now

$$(80) \qquad P(A|B) = \frac{P(A \cap B)}{P(A \cap B) + P((-A) \cap B)} = \frac{P(A \cap B)}{P(B)}$$

Definition 11 *The **conditional probability** of A on condition that B is denoted by $P(A|B)$ and is computed as*

$$(81) \qquad P(A|B) = \frac{P(A \cap B)}{P(B)}$$

This is known as **Bayes law of conditional probabilities**. We can derive a series of important conclusions. First, it allows to compute $P(A \cap B)$ by

$$(82) \qquad P(A \cap B) = P(A|B)P(B)$$

Furthermore, as $A = (A \cap B) \cup (A \cap (-B))$, the sets being disjoint, we have

$$(83) \qquad P(A) = P(A|B)P(B) + P(A| - B)P(-B)$$

This means that the probability of an event can be computed on the basis of the conditional probabilities for a family of sets B_i, $i \in I$, provided that the latter are a partition of Ω. (In words: the B_i must be pairwise disjoint and nonempty, and $\bigcup_{i \in I} B_i = \Omega$.)

Furthermore, reversing the roles of A and B in (81), notice that

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A \cap B)}{P(B)} \cdot \frac{P(B)}{P(A)} = P(A|B) \cdot \frac{P(B)}{P(A)} \tag{84}$$

Thus, as long as the individual probabilities are known (or the odds of A against B) then we can compute the probability of B on condition that A if only we know the probability of A on condition that B. This formula is very important. To see its significance, suppose we have a biased coin, with $p = 0.4$, the probability of getting H. Now, toss the coin ten times. Suppose you get the sequence

$$\mathrm{H, T, T, H, H, T, H, H, T, T} \tag{85}$$

Thus, rather than getting H the expected 4 times we get it 5 times. We can calculate the probability of this happening: it is

$$\binom{10}{5} \cdot 0.4^5 \cdot 0.6^5 = 0.201 \tag{86}$$

Suppose however that the coin is not biased. Then the probability is

$$\binom{10}{5} \cdot 0.5^5 \cdot 0.5^5 = 0.236 \tag{87}$$

Thus, the event of getting H 5 times is more likely when we assume that our coin is in fact not biased. Now we want to actually answer a different question: what is the probability of it being biased with $p = 0.4$ as opposed to not being biased if the result is 5 times H? To answer this, let B be the event that the coin is biased with $p = 0.4$. Let F be the event that we get H 5 times. Let N be the event that the coin is not biased. We assume (somewhat unrealistically) that either B or N is the case. So, $P(B) + P(N) = 1$. Put $\alpha := P(B)$. We have

$$P(F|B) = 0.201, P(F|N) = 0.236 \tag{88}$$

We want to have $P(B|F)$. This is

$$P(B|F) = P(F|B) \cdot \frac{P(B)}{P(F)} = 0.201 \cdot \frac{\alpha}{P(F)} \tag{89}$$

So, we need to know the probability $P(F)$. Now, $P(F) = P(F \cap B) + P(F \cap N) = P(F|B)P(B) + P(F|N)P(N) = 0.201\alpha + 0.236(1 - \alpha) = 0.236 - 0.035\alpha$. Thus we get

$$(90) \qquad P(B|F) = 0.201 \cdot \frac{\alpha}{0.236 - 0.035\alpha}$$

If both B and N are equally likely, we have $\alpha = 1/2$ and so

$$(91) \qquad P(B|F) = 0.201 \cdot \frac{1}{2(0.236 - 0.0185)} = 0.201 \cdot \frac{1}{0.438} = 0.4621.$$

Thus, the probability that the coin is biased is 0.4621 and unbiased with the probability 0.5379, on the assumption that it is either biased with $p = 0.4$ or with $p = 0.5$, with equal likelihood for both hypotheses.

The latter kind of reasoning is very frequent. One has several hypotheses H_1, $H_2, \cdots, H_n$, with “a priori” probabilities $P(H_i)$, $i = 1, 2, \cdots, n$, and computes the probabilities of the outcome B of the experiment. These are the probabilities $P(B|H_i)$. Then one conducts the experiment and gets the result B. Now one asks: what is the probability of hypothesis H_i now that B actually happened? Thus one wants to establish $P(H_i|B)$. These are the “a posteriori” probabilities of the H_i. Abstractly, this can be done as follows. We have

$$(92) \qquad P(H_i|B) = P(B|H_i)\frac{P(H_i)}{P(B)}$$

The only thing we need to know is $P(B)$. Here we do the same as before. We have assumed that the hypotheses H_i obtain with probabilities $P(H_i)$, and that these probabilities add up to 1, so that one of the hypotheses obtains. Thus, $\Omega = \bigcup_{i=1}^n H_i$, the sets pairwise disjoint. We therefore have $B = \bigcup_{i=1}^n (B \cap H_i)$. Now we get

$$(93) \qquad P(B) = \sum_{i=1}^{n} P(B|H_i)P(H_i)$$

Entering this into (92) we get

$$(94) \qquad P(H_i|B) = P(B|H_i)\frac{P(H_i)}{\sum_{i=1}^n P(B|H_i)P(H_i)}$$

If A does not depend on B we expect that its conditional probability $P(A|B)$ equals $P(B)$. This means that $P(A \cap B) = P(A|B)P(B) = P(A)P(B)$. This leads to the following definition.

Definition 12 *Let A and B be events of a probability space $\mathcal{P} = \langle \Omega, \mathfrak{A}, P\rangle$. We call A and B* ***independent*** *if $P(A \cap B) = P(A) \cdot P(B)$. Furthermore, let $\mathfrak{B}_1$ and $\mathfrak{B}_2$ be two subalgebras of $\mathfrak{A}$. $\mathfrak{B}_1$ and $\mathfrak{B}_2$ are* ***independent*** *if for all $B_1 \in \mathfrak{B}_1$ and $B_2 \in \mathfrak{B}_2$ $P(B_1 \cap B_2) = P(B_1) \cdot P(B_2)$.*

We present an example that we shall be using later on. Consider the space $\mathcal{P} \otimes \mathcal{Q}$, where $\mathcal{P} = \langle \Omega, \mathfrak{A}, P\rangle$ and $\mathcal{Q} = \langle \Omega', \mathfrak{A}', P'\rangle$. The sets $A \times B$ have been assigned the probabilities $P_2(A \times B) := P(A)P(B)$. This means that

$$\begin{aligned} P_2(A \times \Omega') &= P(A) \cdot P'(\Omega') = P(A) \\ P_2(\Omega \times B) &= P(\Omega) \cdot P'(B) = P(B) \end{aligned} \tag{95}$$

Now

$$\begin{aligned} P_2((A \times \Omega') \cap (\Omega \times B)) &= P_2((A \cap \Omega) \times (\Omega \cap B)) \\ &= P_2(A \times B) = P(A) \cdot P(B) \\ &= P_2(A \times \Omega') \cdot P_2(\Omega \times B) \end{aligned} \tag{96}$$

Proposition 13 *The sets $A \times \Omega'$ and $\Omega \times B$ are independent for all $A \in \mathfrak{A}$ and $B \in \mathfrak{A}'$ in the space $\mathcal{P} \otimes \mathcal{Q}$.*

Moreover, let $\mathfrak{B}_1$ be the algebra of sets of the form $A \times \Omega'$ and $\mathfrak{B}_2$ the subalgebra of sets of the form $\Omega \times B$. First we shall show they actually are subalgebras.

Proposition 14 *Let $\mathfrak{A}$ and $\mathfrak{B}$ be nontrivial boolean algebras. The map $i_1 : A \mapsto A \times 1_B$ is an embedding of $\mathfrak{A}$ into $\mathfrak{A} \otimes \mathfrak{B}$. Similarly, the map $i_2 : B \mapsto 1_A \times B$ is an embedding of $\mathfrak{B}$ into $\mathfrak{A} \otimes \mathfrak{B}$.*

Proof. First, the map i_1 is injective: for let A, C be sets such that $A \times 1_B = C \times 1_B$. Since $1_B \neq \varnothing$ this means that there is a $b \in 1_B$. For every $a \in A$, $\langle a, b\rangle \in A \times 1_B$, hence $\langle a, b\rangle \in C \times 1_B$, so $a \in C$. Similarly, for every $c \in C$, $\langle c, b\rangle \in C \times 1_B$, so $\langle c, b\rangle \in C \times 1_B$, whence $c \in A$. Therefore, $A = C$. $i_1(A \cup C) = (A \cup C) \times 1_B = (A \times 1_B) \cup (C \times 1_B) = i_1(A) \cup i_1(C)$. Also $i_1(-A) = (-A) \times 1_B = A \times (-1_B) \cup (-A) \times (-1_B) \cup -(A \times 1_B) = (-A) \times 1_B = -i_B(A)$. Similarly for the second claim. ⊣

The algebras $i_1(\mathfrak{A})$ and $i_2(\mathfrak{A})$ are independent, as we have just shown. They represent the algebra of events of performing the experiment for the first time

($\mathfrak{B}_1$) and for the second time ($\mathfrak{B}_2$). The reassuring fact is that by grouping the two executions of the experiment into a single experiment we have preserved the probabilities and moreover have shown that the two experiments are independent of each other in the formal sense.

Theorem 15 *Let $\mathcal{P} = \langle \Omega, \mathfrak{A}, P\rangle$ and $\mathcal{Q} = \langle \Omega', \mathfrak{A}', P'\rangle$ be probability spaces. Then the algebras $i_{\Omega'}[\mathfrak{A}] = \{A \times \Omega' : A \in \mathfrak{A}\}$ and $j_{\Omega}[\mathfrak{A}'] = \{\Omega \times B : B \in \mathfrak{A}'\}$ are independent subalgebras of $\mathfrak{A} \otimes \mathfrak{A}'$.*

Postscript

The case discussed above about adjusting the probabilities raises issues that are worth addressing. I choose again the case where what we know is that the coin is unbiased or biased with $p = 0.4$. The probability $P(p = 0.4)$ is denoted by α. Now, let us perform n experiments in a row, ending in k times H. (It is possible to perform the argument—with identical numerical result—using a particular experimental outcome rather than the event "k times K". This is because we are lumping together individual outcomes that always receive the same probability. Thus, in general one should be aware of a potential confusion here.) Let us write $\beta(n, k)$ for the probability that this happens in case the coin is biased, and $\nu(n, k)$ for the probability that this happens if the coin is unbiased. We have

$$\nu(n,k) = \binom{n}{k}\frac{1}{2^n} \tag{97}$$

$$\beta(n,k) = \binom{n}{k}0.4^k 0.6^{n-k} \tag{98}$$

The unconditional a priori probability of 'k times H' is now

$$\begin{aligned} &\nu(n,k)P(p = 0.5) + \beta(n,k)P(p = 0.4) \\ =&\binom{n}{k}(0.5^n(1-\alpha) + 0.4^k 0.6^{n-k}\alpha) \end{aligned} \tag{99}$$

We are interested in the a posteriori probability of that the coin is biased with $p = 0.4$ based on the outcome ”k times H”. It is given by

$$\begin{aligned}
&P(p = 0.4|k \text{ times H})\\
=&P(k \text{ times H}|p = 0.4)\frac{P(p = 0.4)}{P(k \text{ times H})}\\
=&\binom{n}{k}0.4^k 0.6^{n-k}\frac{\alpha}{\binom{n}{k}(0.5^n(1-\alpha) + 0.4^k 0.6^{n-k}\alpha)}\\
=&0.4^k 0.6^{n-k}\frac{\alpha}{0.5^n(1-\alpha) + 0.4^k 0.6^{n-k}\alpha}\\
=&\frac{\alpha}{(0.5/0.4)^k(0.5/0.6)^{n-k}(1-\alpha) + \alpha}
\end{aligned} \tag{100}$$

For a real number ρ, write

$$f_\rho(\alpha) := \frac{\alpha}{\rho(1-\alpha) + \alpha} \tag{101}$$

Our particular case above was $n = 10$ and $k = 5$. Here the a posteriori probability of α is

$$f(\alpha) = 0.201\frac{\alpha}{0.236 - 0.035\alpha} \tag{102}$$

So, in this case $\rho = 0.236/0.201$. This is an update on our prior probabilities. Several questions arise. First, does it matter which prior probabilities we had? The answer is easy: it does! To see this, just insert a different value into the function. For example, put in $\alpha = 0.25$. Then you get

$$\begin{aligned}
f(0.25) =&\frac{0.201}{4(0.236 - 0.035/4)}\\
=&\frac{0.201}{0.944 - 0.035}\\
=&\frac{0.201}{0.909}\\
=&0.2211
\end{aligned} \tag{103}$$

So, if you already had the opinion that the coin was unbiased with probability 0.75, now you will believe that with probability 0.7789. Second question: is there a prior probability that would not be changed by this experiment? Intuitively, we would regard this as the ideal assumption, the one that is absolutely confirmed by the data. So, we ask if there are α such that

(104) $$f(\alpha) = \alpha$$

Or

(105) $$0.201\frac{\alpha}{0.236 - 0.035\alpha} = \alpha$$

The first solution is $\alpha = 0$. Excluding this we can divide by α and get

(106) $$\begin{aligned}\frac{0.201}{0.236 - 0.035\alpha} &= 1\\ 0.201 &= 0.236 - 0.035\alpha\\ 0.035\alpha &= 0.035\\ \alpha &= 1\end{aligned}$$

Thus, there are two (!) a priori probabilities that are not affected by the data: $\alpha = 0$ and $\alpha = 1$. They represent the unshakeable knowledge that the coin is unbiased ($\alpha = 0$) and the unshakeable knowledge that it is biased ($\alpha = 1$). Especially the last one seems suspicious. How can the fact that we get 5 times H not affect this prior probability if it is the less likely outcome? The answer is this: it is not excluded that we get this outcome, and if we get it and simply know for sure that our coin is biased, what should ever make us revise our opinion? No finite evidence will be enough. We might say that it is unwise to maintain such an extreme position, but probability is not a normative science. Here we are just concerned with the revision suggested by the experiment given our prior probabilities.

There is a temptation to use the data as follows. Surely, the outcome suggests that the posterior probability of bias is $f(\alpha)$ rather than α. So, knowing this we may ask: had we started with the prior probability $f(\alpha)$ we would now have reached $f^2(\alpha) = f(f(\alpha))$, and had we started with the latter, we would now have reached $f(f^2(\alpha)) = f^3(\alpha)$ and so on. Repeating this we are probability getting into a limit, and this is the hypothesis that we should eventually adopt. It turns out that our function has the following behaviour. If $\alpha = 1$ then $f(\alpha) = 1$, so we already have an equilibrium. If $\alpha < 1$ then $f(\alpha) < \alpha$, and by iteration we have a descending sequence

(107) $$\alpha > f(\alpha) > f^2(\alpha) > f^3(\alpha) > \dots$$

The limit of this series is 0. So, the equilibrium solution $\alpha = 1$ is unstable, while the solution $\alpha = 0$ is stable, and it represents the firm belief that the coin is unbiased. So we think that this is what we should adopt simpliciter.

Seeing it this way, though, is to make a big mistake. What we are effectively doing is using the same data several times over to revise our probabilities. This is tantamount to assuming that we performed the experiment several times over with identical result. In fact we did the experiment *once* and therefore can use it only once to update our probabilities. To jump to the conclusion that therefore we have to adopt the firm belief (in fact knowledge) that the coin is unbiased is unwarranted and dangerous. In probability, knowledge is time dependent; probabilities are time dependent. Data advises us to change our probabilities in a certain way. Once we did that, the data has become worthless for the same purpose. (It can be used for a different purpose such as revising other people's probabilities.) This cannot be overestimated. For example, suppose you read somewhere that spinach is good for you, presumably because it contains a lot of iron, and that this has been proved by experiment. The more you read the same claim the more you are inclined to believe that it is true. However, underlying this tendency is the assumption that all this is based on different experiments. Suppose, namely, you are told that everybody who writes this bases himself or herself—directly or indirectly—on the same study conducted some decades ago. In that case it is actually so that reading this claim for the second time already should not (and hopefully will not) make you think it becomes more plausible: the experiment has been conducted once—that's it. If however a new study supports this then that is indeed a reason to believe more strongly in the claim.

We can also prove formally that this is what we should expect. Notice the following equation

$$\begin{aligned} f_\rho(f_\rho(\alpha)) &= \frac{\frac{\alpha}{\rho(1-\alpha)+\alpha}}{\rho\left(1 - \frac{\alpha}{\rho(1-\alpha)+\alpha}\right) + \frac{\alpha}{\rho(1-\alpha)+\alpha}} \\ &= \frac{\alpha}{\rho(\rho(1-\alpha)+\alpha-\alpha)+\alpha} \\ &= \frac{\alpha}{\rho^2(1-\alpha)+\alpha} \\ &= f_{\rho^2}(\alpha) \end{aligned} \tag{108}$$

Let us look at the number ρ for the experiment that we win k out of n times. It is

$$\pi_{k,n} = (0.5/0.4)^k(0.5/0.6)^{n-k} \tag{109}$$

It is observed that

$$\pi_{2k,2n} = \pi_{k,n}^2 \tag{110}$$

So we conclude that $f^2(\alpha) = f_{\pi_{10,20}}(\alpha))$, corroborating our claim that adjusting the probabilities twice yields the same probabilities as if we had performed the experiment twice with identical result.

6 Random Variables

Let $\mathcal{P} = \langle \Omega, \mathfrak{A}, P\rangle$ be a probability space and $X : \Omega \to \mathbb{R}$. We call X a **random variable** if for every $a \in \mathbb{R}$ and $I = [a, b]$ we have $X^{-1}(\{a\}) \in \mathfrak{A}$ and $X^{-1}[I] \in \mathfrak{A}$. The reason the condition has been brought in is that we want to have that $X^{-1}[A] \in \mathfrak{A}$ whenever A is a certain set of reals (in general a so called Borel set, but in fact it is enough to require that the preimage of a singleton and of a closed interval is an event). If $\mathcal{P}$ is discrete then any function into the real numbers is a random variable. We give an example. Suppose that a doctor has two kinds of patients, one with insurance A and one with insurance B. If a patient has insurance A the doctor gets 40 dollars per visit, if the patient is from insurance B he gets 55. The function $X : \{A, B\} \to \mathbb{R}$ defined by $f(A) := 40$ and $f(B) := 55$ is a random variable over the space $\langle\{A, B\}, \wp(\{A, B\}), P\rangle$ where $P(A) = p$ and $P(B) = 1 - p$. Suppose that $p = \frac{1}{3}$; how much money does the doctor get on average from every patient? The answer is

$$(111) \qquad \frac{1}{3} \cdot 40 + \frac{2}{3} \cdot 55 = 50$$

This value is known as the **expected value** or **expectation** of X.

Definition 16 *The* ***expectation*** *of a random variable X is defined by*

$$(112) \qquad \mathsf{E}(X) := \sum_{x \in \mathbb{R}} x \cdot P(X = x)$$

where $P(X = x) = P(X^{-1}(\{x\}))$.

We see that for this to be well–defined, $X^{-1}(\{x\})$ must be an event in the probability space. We shall consider a special case where P has a density function p. Then the formula can be rendered as follows.

$$(113) \qquad \mathsf{E}(X) := \sum_{\omega \in \Omega} X(\omega) \cdot p(\omega)$$

There are many applications of this definition. For example, the words of English have a certain probability, and the function X that assigns to each word a length is a random variable for the discrete space over all words of English. Then $\mathsf{E}(X)$ is the expected length of a word. This has to be kept distinct from the notion of an average length, which is

$$(114) \qquad \frac{\sum_{\vec{x}\in L} X(\vec{x})}{|L|}$$

where L is the set of English words. The average is taken over the words regardless of their probabilities. It says this: suppose you take out of L an arbitrary word, what length will it have? The expectation is different. It says: suppose you hit somewhere on a word (say you open a book and point randomly at a word in it). What length will it have? To answer the latter question we need to know what the likelihood is to hit a specific word. Let us continue the example of the doctor. The average payment of the insurances is 47.50 dollars, but the doctor gets more since the higher insured patients are more likely to show up. Another doctor might have a different ratio and his expected payment per patient may differ accordingly.

Suppose we have a random variable X on a space $\mathcal{P}$. From the definition it is clear that X is compatible with the algebra of events and therefore this defines a new probability space on the direct image $X[\Omega]$, as given in Section 4. For example, instead of talking about insurance as outcomes for the doctor we may simply talk about payments directly. The space is now $\{40, 55\}$ and the probabilities are $P'(40) = \frac{1}{3}$ and $P'(55) = \frac{2}{3}$. This is a different way of looking at the same matter, and there is no obvious advantage of either viewpoint over the other. But this construction helps to explain why there are spaces different from the Laplace space. Suppose namely that there is a space Ω of outcomes; if we know nothing else the best assumption we can make is that they are all equally likely to occur. This is the null hypothesis. For example, when we are shown a die we expect that it is fair, so that every number between 1 and 6 is equally likely. Now suppose we throw two dice simultaneously and report only the sum of the values. Then we get a different space, one that has outcomes $2, 3, \cdots, 12$, with quite different probabilities. They arise as follows. We define a random variable $X(\langle x, y\rangle) := x + y$ on the set of outcomes. The probability of a number z is simply $X^{-1}(\{z\})$. The new space is no longer a Laplace space, but it arose from a Laplace space by transforming it via a random variable. The Laplace space in turn arises through the null hypothesis that both dice are unbiased.

Suppose that $A \subseteq \Omega$ is a set. Then let

$$(115) \qquad I(A)(\omega) := \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{else.} \end{cases}$$

The function $I(A)$ is known as the **characteristic function** of A. It is obviously a random variable, so its expectation can be computed. It is

$$\mathsf{E}\, I(A) = \sum_{\omega\in\Omega} p(\omega) I(A)(\omega) = \sum_{\omega\in A} p(w) = P(A) \tag{116}$$

Proposition 17 $\mathsf{E}\, I(A) = P(A)$.

If X and Y are random variables, we can define $X+Y$, αX and $X \cdot Y$ as follows.

$$(X+Y)(\omega) := X(\omega) + Y(\omega) \tag{117}$$

$$(\alpha X)(\omega) := \alpha \cdot X(\omega) \tag{118}$$

$$(X \cdot Y)(\omega) := X(\omega) \cdot Y(\omega) \tag{119}$$

These functions are not necessarily random variables as well.

Proposition 18 *Suppose that X and Y are random variables. If $X+Y$ and αX are random variables, then*

$$\mathsf{E}(X+Y) = \mathsf{E}(X) + \mathsf{E}(Y) \tag{120}$$

$$\mathsf{E}(\alpha X) = \alpha\, \mathsf{E}(X) \tag{121}$$

Proof. Direct verification. We do the case where P has a density.

$$\begin{aligned} \mathsf{E}(X+Y) &= \sum_{\omega\in\Omega} (X+Y)(\omega) p(\omega) \\ &= \sum_{\omega\in\Omega} (X(\omega) + Y(\omega)) p(\omega) \\ &= \sum_{\omega\in\Omega} X(\omega)p(\omega) + Y(\omega)p(\omega) \\ &= \sum_{\omega\in\Omega} X(\omega)p(\omega) + \sum_{\omega\in\Omega} Y(\omega)p(\omega) \\ &= \mathsf{E}(X) + \mathsf{E}(Y) \end{aligned} \tag{122}$$

Also

$$
\begin{aligned}
\mathsf{E}(\alpha X) &= \sum_{\omega\in\Omega}(\alpha X)(\omega)p(\omega)\\
&= \sum_{\omega\in\Omega}\alpha X(\omega)p(\omega)\\
&= \alpha\sum_{\omega\in\Omega}X(\omega)p(\omega)\\
&= \alpha\,\mathsf{E}(X)
\end{aligned} \tag{123}
$$

⊣

These formulae are very important. We give an immediate application. Suppose we have a Bernoulli space $\mathcal{P}$ and a random variable X. Its expectation is $\mathsf{E}(X)$. The expectation per patient is the same no matter how often we do the experiment. In the case of the doctor, the average payment he gets from two patients is 50 per patient, therefore 100 in total. Indeed, let us perform a Bernoulli experiment twice. Then we may also view this as a single space $\mathcal{P} \otimes \mathcal{P}$. Now we have two random variables, $X_1(\langle x, y\rangle) = X(x)$, and $X_2(\langle x, y\rangle) = X(y)$. The first variable returns the value for the first experiment, and the second the value for the second experiment. The variable $\frac{1}{2}(X_1 + X_2)$ takes the mean or average of the two. We have

$$\mathsf{E}\left(\frac{1}{2}(X_1 + X_2)\right) = \mathsf{E}(X) \tag{124}$$

Equivalently, the expectation of $X_1 + X_2$ is $2\,\mathsf{E}(X)$. This is easily generalised to n–fold iterations of the experiment.

The expectation may not be a value that the variable ever attains; an example has been given above. Moreover, one is often interested in knowing how far the actual values of the variable differ from the expected one. This is done as follows.

$$\mathsf{V}(X) = \mathsf{E}(X - \mathsf{E}(X))^2 \tag{125}$$

This is know as the **variance** of X. By way of example, let us calculate the variance of the identity I function on a Bernoulli experiment. First we have to establish its expectation (recall that $p = p(1)$ and $q = p(0) = 1 - p$).

$$\mathsf{E}\,I = p \cdot 1 + q \cdot 0 = p \tag{126}$$

It has two outcomes, 0 and 1, and the random variable assigns 0 to 0 and 1 to 1. So

$$\begin{aligned}\mathsf{V}\,I &= p\cdot(1-\mathsf{E}\,I)^2 + q\cdot(0-\mathsf{E}\,I)^2\\ &= p(1-p)^2 + q(-p)^2\\ &= pq^2+qp^2 \qquad (127)\\ &= pq(q+p)\\ &= pq\end{aligned}$$

This will be useful later. Also, the **standard deviation** of X, $\sigma(X)$, is defined by

(128) $$\sigma(X) := \sqrt{\mathsf{V}(X)}$$

Notice that $X - \mathsf{E}(X)$ is a random variable returning for each ω the difference $X(\omega) - \mathsf{E}(X)$. This difference is squared and summed over all ω, but the sum is again weighted with the probabilities. In the case of the doctor, we have $(X - \mathsf{E}(X))(A) = -10$ and $(X - \mathsf{E}(X))(B) = 5$. Hence

(129) $$\mathsf{V}(X) = \frac{1}{3}\cdot(-10)^2 + \frac{2}{3}\cdot 5^2 = \frac{150}{3} = 50$$

(That this is also the expected value is a coincidence.) Hence $\sigma(X) = \sqrt{50} \approx 7.071$. Thus, the payment that the doctor gets standardly deviates from the expected value 50 by 7.071. The deviation measures the extent to which the actual payments he receives differ from his expected payment. If the deviation is 0 then no payment is different, all payments equal the expected payment; the larger the deviation the larger the difference between individual payments is to be expected.

The formula for the variance can be simplified as follows.

(130) $$\mathsf{V}(X) = \mathsf{E}(X^2) - (\mathsf{E}(X))^2$$

For a proof notice that

$$\begin{aligned}\mathsf{E}(X-\mathsf{E}\,X)^2 &= \mathsf{E}(X-\mathsf{E}\,X)(X-\mathsf{E}\,X)\\ &= \mathsf{E}(X^2 - 2X\cdot\mathsf{E}\,X + (\mathsf{E}\,X)^2)\\ &= \mathsf{E}(X^2) - 2\,\mathsf{E}((\mathsf{E}\,X)\cdot X) + (\mathsf{E}\,X)^2 \qquad (131)\\ &= \mathsf{E}(X^2) - 2(\mathsf{E}\,X)(\mathsf{E}\,X) + (\mathsf{E}\,X)^2\\ &= \mathsf{E}(X^2) - (\mathsf{E}\,X)^2\end{aligned}$$

We shall use the notation $X = x$ for the set of all outcomes ω such that $X(\omega) = x$.

Definition 19 *Let X and Y be random variables on a space $\mathcal{P}$. X and Y are said to be **independent** if for all $x, y \in \mathbb{R}$: $P(X = x \cap Y = y) = P(X = x) \cdot P(Y = y)$.*

Theorem 20 *Let X and Y be independent random variables. Then $\mathsf{E}(X \cdot Y) = \mathsf{E}\, X \cdot \mathsf{E}\, Y$ and $\mathsf{V}(X + Y) = \mathsf{V}\, X + \mathsf{V}\, Y$.*

Proof. For the first, assume that X assumes the values $\{x_i : i \in I\}$ and that Y assumes the values $\{y_j : j \in J\}$.

$$\begin{aligned}
\mathsf{E}(X \cdot Y) &= \mathsf{E}\left(\sum_{i \in I} x_i I(A_i)\right)\left(\sum_{j \in J} y_j I(B_j)\right) \\
&= \mathsf{E}\left(\sum_{i \in I, j \in J} x_i y_j I(A_i \cap B_j)\right) \\
&= \left(\sum_{i \in I, j \in J} x_i y_j\, \mathsf{E}\, I(A_i \cap B_j)\right) \\
&= \left(\sum_{i \in I, j \in J} x_i y_j P(A_i \cap B_j)\right) \qquad (132) \\
&= \left(\sum_{i \in I, j \in J} x_i y_j P(A_i) P(B_j)\right) \\
&= \left(\sum_{i \in I} x_i P(A_i)\right)\left(\sum_{j \in J} y_j P(B_j)\right) \\
&= (\mathsf{E}\, X) \cdot (\mathsf{E}\, Y)
\end{aligned}$$

Notice that the independence assumption entered in form of the equation $P(A_i \cap B_j) = P(A_i) \cdot P(B_j)$. Also, the expectation does distribute over infinite sums just in case the sum is absolute convergent. For the second claim notice first that if X and Y are independent, so are $X - \alpha$ and $Y - \beta$ for any real numbers α and β. In particular, $X - \mathsf{E}\, X$ and $Y - \mathsf{E}\, Y$ are independent, so $\mathsf{E}((X - \mathsf{E}\, X)(Y - \mathsf{E}\, Y)) = \mathsf{E}(X - \mathsf{E}\, X) \cdot \mathsf{E}(Y - \mathsf{E}\, Y) = 0$. From this we deduce the second claim as follows.

$$\begin{aligned}
\mathsf{V}(X + Y) &= \mathsf{E}((X - \mathsf{E}\, X) + (Y - \mathsf{E}\, Y))^2 \\
&= \mathsf{E}(X - \mathsf{E}\, X)^2 - 2\,\mathsf{E}(X - \mathsf{E}\, X)(Y - \mathsf{E}\, Y) + \mathsf{E}(Y - \mathsf{E}\, Y)^2 \qquad (133) \\
&= \mathsf{V}\, X + \mathsf{V}\, Y
\end{aligned}$$

⊣

This is but a special version of a more general result, which is similar in spirit to Theorem 15.

Theorem 21 *Let $\mathcal{P}$ and $\mathcal{Q}$ be probability spaces, and X and Y random variables over $\mathcal{P}$ and $\mathcal{Q}$ respectively. Then define the following random variables X^1 and Y^2:*

(134)
$$\begin{aligned} X^1(\langle\omega_1, \omega_2\rangle) &:= X(\omega_1) \\ Y^2(\langle\omega_1, \omega_2\rangle) &:= Y(\omega_2) \end{aligned}$$

Then X^1 and Y^2 are independent random variables over $\mathcal{P} \otimes \mathcal{Q}$.

The proof is relatively straightforward. $X^1 = \omega_1 = (X = \omega_1) \times \Omega'$ and $Y^2 = \omega_2 = \Omega' \times (Y = \omega_2)$, and so by definition

(135) $$P(X^1 = \omega_1 \cap Y^2 = \omega_2) = P(X^1 = \omega) \cdot P(X^2 = \omega_2)$$

7 Expected Word Length

We shall present some applications of the previous definitions. The first concerns the optimal encoding of texts. The letters of the alphabet are stored in a computer in the so–called ASCII code. Here each letter receives an 8bit sequence. In connection with coding we say that a **letter code** C is a map which assigns to each member a of an alphabet A a member of B^*, the so–called **code word** of a. A string $x_1x_2\cdots x_n$ is then coded by $C(x_1)C(x_2)\cdots C(x_n)$. There obviously are quite diverse kinds of codes, but we shall only consider letter codes. The Morse code is a letter code, but it is different from the ASCII in one respect. While some letters get assigned rather long sequences, some get quite short ones (e gets just one, the dot). One difference between these codes is that the expected length of the coding is different. The Morse code is more efficient in using less symbols on average. This is so since more frequent letters get assigned shorter sequences than less frequent ones. In statistical terms we say that the expected word length is smaller for the Morse code than for the ASCII code. This can be computed once the probabilities of the letters are known. Rather than doing that, we shall describe here a general method that allows to generate an optimal letter coding (in the sense that it minimises the expected word length) on the basis of the probabilities for the letters. We are of looking for a so–called prefix free code; this is a code where no prefix of a code word is a code word. This ensures unique readability. The Morse code is prefix free; this is so since we assign to each letter the Morse sequence plus the following silence. (You have to realize that the Morse code uses in total three letters: $\cdot$, $-$ and silence.) The **compression factor** is the inverse of the expectation of the code length. Let X be the random variable which assigns to each $a \in A$ the length of $C(a)$. Then

$$(136) \qquad \mathsf{E}(X) = \sum_{a\in A} X(a)p(a)$$

Since the length of the original symbol is 1 for each $a \in A$, a symbol is replaced on average by $\mathsf{E}\,X$ many symbols. This is why we call its inverse the compression factor.

As the probabilities of the letters may be different, the codings of letters may result in different compression factors. Obviously, it is best to assign the shortest words to those letters that are most frequent. We shall present an algorithm that produces a letter code into $\{\mathtt{0}, \mathtt{1}\}^*$ that minimises the expected word length (and therefore maximises the compression factor). This is the so–called **Huffman**

code. (As we shall see, there typically are several such codes for any given alphabet. But they result in the same compression factor.) We give an example. Let the alphabet be $\{\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}\}$ with the following frequencies:

(137) $p(\mathtt{a}) = 0.1, p(\mathtt{b}) = 0.2, p(\mathtt{c}) = 0.3, p(\mathtt{d}) = 0.4$

To start we look for a pair of letters whose combined frequency is minimal. For example, a and b together occur in 3 out of 10 times, and no other pair of letters occurs that rarely. We introduce (temporarily) a new letter, x, which replaces both a and b. The new alphabet is $\{\mathtt{x}, \mathtt{c}, \mathtt{d}\}$. The frequencies are

(138) $p(\mathtt{x}) = 0.3, p(\mathtt{c}) = 0.3, p(\mathtt{d}) = 0.4$

We repeat the step: we look for a pair of letters whose frequency is minimal. There is only one possibility, namely x and d. We add a new letter, y, which represents both x and c. This gives the alphabet $\{\mathtt{y}, \mathtt{d}\}$, with frequencies

(139) $p(\mathtt{y}) = 0.6, p(\mathtt{d}) = 0.4$

We repeat the same step again. This time there is no choice, the pair is bound to be y and d. Let z be a new letter, with frequency 1. Now we start to assign code words. We assign the empty word to z. Now z actually represents a pair of letters, y and d. Therefore we expand the word for z as follows: we append 0 to get the word for y and we append 1 to get the word for d. We repeat this for y, which represents x and c. We append 0 for x and 1 for c. As x represents both a and b, we append 0 to get the code for a and 1 to get the code for b. These are now the code words:

(140) $C(\mathtt{a}) = \mathtt{000}, C(\mathtt{b}) = \mathtt{001}, C(\mathtt{c}) = \mathtt{01}, C(\mathtt{d}) = \mathtt{1}$

The expected word length is

(141) $L(C) = 0.1 \times 3 + 0.2 \times 3 + 0.3 \times 2 + 0.4 \times 1 = 1.9$

The compression is therefore $1/1.9 = 0.5263$. Suppose that instead we had used the following **block code**, which is a code that assigns words of equal length to each letter.

(142) $C'(\mathtt{a}) = \mathtt{00}, C'(\mathtt{b}) = \mathtt{01}, C'(\mathtt{c}) = \mathtt{10}, C'(\mathtt{d}) = \mathtt{11}$

Then the expected word length is

(143) $$L(C') = 0.1 \times 2 + 0.2 \times 2 + 0.3 \times 2 + 0.4 \times 2 = 2$$

with compression rate 0.5. Thus, the Huffman code is better than the block code. Notice that we claimed above that there are several optimal codes. Indeed, there are several places where we made a choice. For example, we chose to append `0` to represent `y` and `1` to present `d`. Obviously we could have chosen `1` to represent `y` and `0` to represent `d`. This would have given the code

(144) $$C(\mathtt{a}) = \mathtt{100}, C(\mathtt{b}) = \mathtt{101}, C(\mathtt{c}) = \mathtt{11}, C(\mathtt{d}) = \mathtt{0}$$

Every time we choose a pair of letters we face a similar choice for the code. Also, it may occur that there are two or more pairs of letters that have the same minimal frequency. Again several possibilities arise, but they all end up giving a code with the same expected word length and compression.

We shall briefly comment on the possibility of improving the compression. One may calculate that a somewhat better compression can be reached if the text is cut into chunks of length 2, and if each 2 letter sequence is coded using the Huffman code. Thus the alphabet is $A \times A$, which contains 16 'letters'. A still better coding is reached if we divide the text into blocks of length 3, and so on. The question is whether there is an optimal bound for codes. It is represented by the **entropy**. Suppose that $A = \{\mathtt{a}_i : 1 \leq i \leq n\}$, and that p_i is the frequency of letter $\mathtt{a}_i$.

(145) $$H(p) := -\sum_{i=1}^{n} p_i \log_2 p_i$$

For our original alphabet this is

(146) $$-(0.1 \times \log_2 0.1 + 0.2 \log_2 0.2 + 0.3 \log_2 0.3 + 0.4 \log_2 0.4) = 1.8464$$

The optimal compression that can be reached is therefore 0.5416. These results hold only if the probabilities are independent of their relative position, that is to say, if some letter occurs at position i the conditional probability of another letter to occur at position $i + 1$ (or at any other given position) is its probability. One says that the source has no memory.

Suppose that we have an alphabet of 3 letters, `a`, `b` and blank. The probabilities are as follows.

(147) $$p(\mathtt{a}) = \frac{1}{2}, p(\mathtt{b}) = \frac{1}{3}, p(\Box) = \frac{1}{6}$$

A text is a string over this alphabet. We assume (somewhat unrealistically) that the probability of each letter in the text is independent of the other letters. So, a occurs after a with the same probability as it follows b or blank, namely $\frac{1}{2}$. Later we shall turn to Markov models, where it is possible to implement context restrictions on the frequencies. Words in a text are maximal subsequences which do not contain □. We shall establish first the probability that a certain word occurs in a text. For example, the word a occurs with the probability $\frac{1}{10}$, while b occurs with probability $\frac{1}{15}$. To see this, let x be an occurrence of blank (□) in the text which precedes that letter. One would ordinarily assume that the probability that the next symbol is a is exactly $\frac{1}{2}$. But this is not so. For we have placed the blank such that it precedes a letter, in other words such that the next symbol is *not* a blank. This therefore calls for the following conditional probability:

$$(148)\qquad P(\{\mathtt{a}\}|\{\mathtt{a},\mathtt{b}\}) = \frac{\frac{1}{2}}{\frac{5}{6}} = \frac{3}{5}$$

Now, it is not enough that the next letter is a, we also must have that the letter following it is blank. Therefore, we only have an occurrence of the *word* a if the next letter is □, which adds a factor of $\frac{1}{6}$. We write $p_w(\vec{x})$ for the probability that the word $\vec{x}$ occurs. Hence we get $p_w(\mathtt{a}) = \frac{1}{10}$. Likewise, the probability of the word b is $\frac{2}{5} \cdot \frac{1}{6} = \frac{1}{15}$. Here are now the probabilities of words of length 2:

$$(149)\qquad \begin{array}{llll} p_w(\mathtt{aa}) & = \frac{1}{20} & p_w(\mathtt{ab}) & = \frac{1}{30} \\ p_w(\mathtt{ba}) & = \frac{1}{30} & p_w(\mathtt{bb}) & = \frac{1}{45} \end{array}$$

The general formula is this:

$$(150)\qquad p(x_1x_2\cdots x_n) = \frac{1}{5}\prod_{i=1}^{n} p(x_i)$$

The additional factor $\frac{1}{5}$ derives, as explained above, from two facts: first, that we have chosen a position which is followed by a nonblank (a factor 6/5), and from the fact that the sequence we are considering must be followed by a blank (a factor 1/6).

Now let X be a random variable, assigning each word its length. Thus $(x_1x_2\cdots x_n) = n$. We want to know its expectation. This will give us an estimate of the length of a randomly chosen word. We shall first derive a formula for the probability that a word has length n. For $n = 1$ it is $\frac{1}{10} + \frac{1}{15} = \frac{5}{30} = \frac{1}{6}$. For $n = 2$ it is

$$(151)\qquad \frac{1}{20} + \frac{1}{30} + \frac{1}{30} + \frac{1}{45} = \frac{9+6+6+4}{180} = \frac{5}{36}$$

This suggests $\frac{5^{n-1}}{6^n}$ as a general probability. Indeed, there is an easy way to see this. The probability that the ith letter is either a or b is $\frac{5}{6}$. We have n letters, therefore the probability is $\left(\frac{5}{6}\right)^n$. Finally, the next letter must be the blank, so we have to multiply by $\frac{1}{6}$. It is checked that

$$\begin{aligned}\sum_{n=1}^{\infty}\frac{5^{n-1}}{6^n} &= \frac{1}{6}\sum_{n=0}^{\infty}\left(\frac{5}{6}\right)^n \\ &= \frac{1}{6}\cdot\frac{1}{1-5/6} \\ &= \frac{1}{6}\cdot 6 \\ &= 1\end{aligned} \tag{152}$$

To see this, notice that

Lemma 22 *Let $p \neq 1$ be a real number.*

$$\sum_{n=0}^{m} p^n = \frac{1-p^{m+1}}{1-p} \tag{153}$$

Proof. By induction on m. For $m = 0$, the sum extends over $p^0 = 1$. The term on the right is $\frac{1-p^1}{1-p} = 1$. Now assume the formula holds for m. Then we have

$$\begin{aligned}\sum_{n=0}^{m+1} p^n &= \frac{1-p^{m+1}}{1-p} + p^{m+1} \\ &= \frac{1-p^{m+1}+p^{m+1}(1-p)}{1-p} \\ &= \frac{1-p^{m+1}+p^{m+1}-p^{m+2}}{1-p} \\ &\frac{1-p^{m+2}}{1-p}\end{aligned} \tag{154}$$

⊣

Notice that the formula holds for all reals, except for $p = 1$, for which the sum is simply m. To finish the proof, notice that if $|p| < 1$ then the value of $\frac{1-p^m}{1-p}$ approaches $\frac{1}{1-p}$ for growing m. So the infinite sum actually equals $\frac{1}{1-p}$.

So indeed this is a probability function. We are now ready for the expected length:

$$\mathsf{E}(X) = \sum_{n=0}^{\infty} \frac{n}{6} \cdot \frac{5^n}{6^n} \tag{155}$$

Now, we need to solve an sum over a series of the form np^n.

Proposition 23 *Let $p \neq 1$ be a real number.*

$$\sum_{n=0}^{m} np^n = \left(\frac{m+1}{p-1} - \frac{p}{(p-1)^2}\right) p^{m+1} + \frac{1}{(p-1)^2} \tag{156}$$

Proof. Let $m = 0$. Then the left hand side equals 1. The right hand side equals

$$\begin{aligned}\left(\frac{1}{p-1} - \frac{p}{(p-1)^2}\right) \cdot p + \frac{1}{(p-1)^2} &= \frac{p(p-1) - p + 1}{(p-1)^2} = \\ &= \frac{P^2 - 2p + 1}{(p-1)^2} \\ &= 1\end{aligned} \tag{157}$$

Now we proceed to the inductive step.

$$\begin{aligned}&p^{m+2}\left(\frac{m+2}{p-1} - \frac{p}{(p-1)^2}\right) - \frac{1}{(p-1)^2} \\ &\qquad - p^{m+1}\left(\frac{m+1}{p-1} - \frac{p}{(p-1)^2}\right) + \frac{1}{(p-1)^2} \\ &= p^{m+1}\left(\frac{p(m+2) - (m+1)}{p-1} - \frac{p^2 - p}{(p-1)^2}\right) \\ &= p^{m+1}\left(\frac{(p-1)(m+1) + p}{p-1} - \frac{p}{(p-1)}\right) \\ &= (m+1)p^{m+1}\end{aligned} \tag{158}$$

This is as promised. $\dashv$

As m grows large (156) approaches

$$\frac{1}{(1-p)^2} \tag{159}$$

With $p = \frac{5}{6}$ this becomes $6^2 = 36$. We insert this into (155) and get

$$\text{(160)} \qquad \mathsf{E}(X) = \frac{36}{6} \approx 6$$

A final remark. The expected word length does not depend on the individual frequencies of the letters. All that it depends on is the probability of the blank.

If we convert this to a probability space, we will get $\langle \mathbb{N}, \wp(\mathbb{N}), P \rangle$ with $p(n) = \frac{p^n}{1-p}$ and $P(A) := \sum_{k \in A} p(k)$. As we have seen the probabilities sum to 1. The exponential decline in probability occurs in many other circumstances. There is an observation due to **Zipf** that the frequency of words decreases exponentially with their length. This cannot be deduced from our results above for the fact that the letter probabilities are not independent.

Let us investigate somewhat closer the frequencies of words. We have just seen that if the letters are randomly distributed the probability of a word decreases exponentially with its length. On the other hand, there are exponentially many words of length n. Let us define a bijective function f from the natural numbers onto the set of words such that the frequency of $f(n+1)$ is less or equal to the frequence of $f(n)$. In other words, f orders the words according to their frequency, and it is sometimes called the **rank function**; $f(1)$ is the most frequent word, followed by $f(1)$ with equal probability (or less), followed by $f(2)$, and so on. It is not necessarily the case that a more frequent word is also shorter. To continue our example,

$$\text{(161)} \qquad p_w(\mathsf{aaaa}) = \frac{1}{10} \cdot \frac{1}{2^4} = \frac{1}{16} > p_w(\mathsf{bbb}) = \frac{1}{10} \cdot \frac{1}{3^3} = \frac{1}{270}$$

On a larger scale, however, it is true that the less frequent an item is the longer it is. To make the computations simple let us assume that all symbols different from the blank have the same frequency α. Furthermore, let there be d many nonblank letters. This means that the probability that a given word has length n is exactly $\alpha^n \cdot (1 - \alpha)$. And so the probability of any given word is $\frac{\alpha^n(1-\alpha)}{d^n}$, since there are d^n words of length n. Now, if $k < \ell$ then we have $p_w(k) \leq p_w(\ell)$ since the correlation between length and probability is monotone decreasing. (If letters are unequally distributed this would not be strictly so, as we have seen above.) The item number $k_n := \frac{d^{n+1}-1}{d-1}$ is the first to have length n. Solving this for n we get

$$\text{(162)} \qquad n = (\log_d(k_n(d-1)+1)) - 1$$

For a real number r, let $\ulcorner r \urcorner$ denote the largest integer $\leq r$. We derive for the length $\ell(k)$ of the kth item on the following probability rank:

$$\llcorner(\log_d(k(d-1)+1)-1\lrcorner \leq \ell(k) < \llcorner(\log_d(k(d-1)+1)-1\lrcorner + 1 \tag{163}$$

For large k we have

$$(\log_d(k(d-1)+1)-1) \approx \log_d k(d-1) \tag{164}$$

Noticing that $\log_d x = (\log_2 d)(\log_d x)$ and $\log_d k(d-1) = \log_d(d-l) + \log_2 k$ we can say that

$$\ell(k) \approx \beta + \alpha \log_2 k \tag{165}$$

for some α and β. Notice that this an asymptotic formula, and where the original function was taking only positive integer values, this one is a real valued function which moreover is strictly monotone increasing. Nevertheless, we have shown that there is some reason to believe that the length of a word increases logarithmically as a function of its rank in the probability scale. Finally, let us insert this into the formula for probabilities. Before we do this, however, note that the original formula was based on the length being a discrete parameter, with values being positive integers. Now we are actually feeding real numbers. This has to be accounted for by changing some numerical constants. So we are now operating we the assumption that the probability of a word depends only on its length x and equals

$$\gamma \cdot 2^{\theta x} \tag{166}$$

for some γ and θ to be determined. Insert the formula for the length based on its rank:

$$\begin{aligned} p(k) &:= \gamma 2^{\theta(\beta + \alpha \log_2 k)} \\ &= \gamma 2^{\theta\beta} 2^{\alpha \log_2 k} \\ &= \gamma 2^{\alpha\beta} k^{\alpha} \\ &= \xi k^{\alpha} \end{aligned} \tag{167}$$

where $\xi := \gamma 2^{\alpha\beta}$. Actually, the value of ξ can be expressed as follows. By definition of the Riemann zeta–function,

$$\zeta(x) := \sum_{k=1}^{\infty} \frac{1}{k^x} \tag{168}$$

we immediately get

(169) $\xi = \zeta(-\alpha)^{-1}$

This is done to ensure that the probabilities sum to exactly 1. Thus, the higher an item is on the rank, the less probable it is, and the longer it is. The length increases logarithmically the probability decreases exponentially with the rank. This is known as **Zipf's Law**. Again, notice that we have not strictly speaking derived it. Our assumptions were drastic: the probabilities of letters are all the same and they do not depend on the position the letters occur in. Moreover, while this ensures that probabilities can be equal among words of adjacent rank, we have fitted a curve there that decreases strictly from one rank to the next. Thus, on this model no two words have the same probability.

8 The Law of Large Numbers

The probabilities that occur in the definition of a probability space are simply numbers. But these numbers have a concrete meaning. They say that an event A occurs with chance $P(A)$ and does not occur with chance $1 - P(A)$. If $P(A) = 0.7$ we expect that in 7 out of A cases, A is the case. If we perform the experiment, however, it can only do one thing: occur or not occur. The probability becomes certainty. Thus it is absolutely useless to assign probabilities to an experiment that can only be conducted once. However, if we can arbitrarily repeat the experiment, we can actually make sense of the probabilities as follows. If $P(A) = 0.7$ we expect that in 7 out of 10 experiments A obtains. Now, we have seen earlier that it does not mean that when we perform the experiment 10 times that A must hold exactly 7 times. To see this, let us calculate the probabilities in detail. The experiment is a Bernoulli experiment with $p = 0.7$. The chance that A obtains exactly 10 times is, for example, 0.7^{10}. Let α_i be the event that A occurs i times exactly:

$$\begin{array}{lll}
P(\alpha_0) & = 0.3^{10} & = 0.00000590 \\
P(\alpha_1) & = \binom{10}{1} \cdot 0.3^9 \cdot 0.7^1 & = 0.00013778 \\
P(\alpha_2) & = \binom{10}{2} \cdot 0.3^8 \cdot 0.7^2 & = 0.00144670 \\
P(\alpha_3) & = \binom{10}{3} \cdot 0.3^7 \cdot 0.7^3 & = 0.00900169 \\
P(\alpha_4) & = \binom{10}{4} \cdot 0.3^6 \cdot 0.7^4 & = 0.03675691 \\
P(\alpha_5) & = \binom{10}{5} \cdot 0.3^5 \cdot 0.7^5 & = 0.10291935 \\
P(\alpha_6) & = \binom{10}{6} \cdot 0.3^4 \cdot 0.7^6 & = 0.20012095 \\
P(\alpha_7) & = \binom{10}{7} \cdot 0.3^3 \cdot 0.7^7 & = 0.26682793 \\
P(\alpha_8) & = \binom{10}{8} \cdot 0.3^2 \cdot 0.7^8 & = 0.23347444 \\
P(\alpha_9) & = \binom{10}{9} \cdot 0.3^1 \cdot 0.7^9 & = 0.12106082 \\
P(\alpha_{10}) & = 0.7^{10} & = 0.02824752
\end{array} \tag{170}$$

We can see two things: none of the outcomes is impossible, but the outcome α_7 is more likely than the others. The events $\alpha_6 \cup \alpha_7 \cup \alpha_8$ together have probability 0.7, roughly. If we deviate by 1 from the expected outcome, the probability is 0.7; if we deviate by up to 2 from the expected result the probability is even larger: it exceeds 0.9.

Now suppose we repeat the experiment 100 times, what do we get? Rather than do the calculations (which involve quite large numbers) we give the answer right away: the likelihood that the mean deviates from the expected value, 0.7, by at most 10, becomes larger. This means that the mean is less likely to deviate from the expected value as the number of iterations get larger. This is known as the law of large numbers. We shall prove it rigorously. We begin with an easy observation.

Lemma 24 (Chebyshev) *Let X be a positive random variable and $\varepsilon > 0$. Then $P(X \geq \varepsilon) \leq \mathsf{E}(X)/\varepsilon$.*

Proof. Let $I(A)$ be the function such that $I(A)(x) = 1$ iff $x \in A$ and 0 else. Then

(171) $$X \geq X \cdot I(X \geq \varepsilon) \geq \varepsilon I(X \geq \varepsilon)$$

This is seen as follows. Suppose $X(\omega) < \varepsilon$. Then $X(\omega) \geq X(\omega) \cdot I(X \geq \varepsilon)(\omega) \geq \varepsilon I(X \geq \varepsilon)$, because $I(X \geq \varepsilon)(\omega) = 0$. Now assume that $X(\omega) \geq \varepsilon$. Then $I(X \geq \varepsilon)(\omega) = 1$ and so $X\omega) \geq X(\omega) \cdot I(X \geq \varepsilon) \geq \varepsilon I(X \geq \varepsilon)$. Now we obtain

(172) $$\mathsf{E}\,X \geq \mathsf{E}(\varepsilon I(X \geq \varepsilon) = \varepsilon P(X \geq \varepsilon)$$

This holds because in general if $X \geq Y$, that is, if for all ω: $X(\omega) \geq Y(\omega)$, then $\mathsf{E}\,X \geq \mathsf{E}\,Y$. And $\mathsf{E}\,I(A) = P(A)$ for every A. For

(173) $$\mathsf{E}\,I(A) = \sum_{\omega\in\Omega} I(A)(\omega) \cdot p(\omega) = \sum_{\omega\in A} p(\omega) = P(A)$$

⊣

The following are immediate consequences.

Corollary 25 *Let X be a random variable. Then*

❶ $P(|X| \geq \varepsilon) \leq \mathsf{E}(|X|)/\varepsilon$.

❷ $P(|X| \geq \varepsilon) = P(X^2 \geq \varepsilon^2) \leq (\mathsf{E}\,X^2)/\varepsilon^2$.

❸ $P(|X - \mathsf{E}\,X|) \geq \varepsilon) \leq (\mathsf{V}\,X)/\varepsilon^2$.

Proof. The first follows from Lemma 24 by noting that $|X|$ is a random variable taking only positive values. The second follows since X^2 also is a positive random variable. Finally, notice that the variance of X is the expectation of $X - \mathsf{E}\,X$, so the third follows from the second if we substitute $X - \mathsf{E}\,X$ for X in it. ⊣

Let X_n be the following random variable on $\{0,1\}^n$:

$$(174)\qquad X_n(\langle \omega_1, \omega_2, \cdots, x_n\rangle) := \sum_{i=1}^{n} x_i$$

The expectation is pn, as we have noted earlier. We ask what the probability is that it deviates more than εn from this value. This is equivalent to asking whether the mean of X_1 deviates more than ε from p if the experiment is repeated n times. We calculate

$$(175)\qquad P\left(\left|\frac{X_n}{n} - p\right| \geq \varepsilon\right) \leq \mathsf{V}\left(\frac{X_n}{n}\right)/\varepsilon^2$$

Let us therefore calculate the variance of X_n/n. To obtain it, let us recall that if Y and Z are independent, $\mathsf{V}(Y+Z) = \mathsf{V}\,X + \mathsf{V}\,Z$. Define variables $X^k : \{0,1\}^n \to \{0,1\}$ by

$$(176)\qquad X^k(\langle x_1, x_2, \cdots, x_n\rangle) := x_k$$

These variables are independent. To see this, let A be the set of all $\omega = \langle x_1, x_2, \cdots, x_n\rangle$ such that $x_j = a$. The probability of this set is exactly p if $a = 1$ and q otherwise. For the set has the form

$$(177)\qquad \{0,1\}^{j-1} \times \{a\} \times \{0,1\}^{n-j-1} = \{0,1\} \times \cdots \{0,1\} \times \{a\} \times \cdots \{0,1\} \times \cdots \times \{0,1\}$$

and so by Proposition 21, $P_n(A) = P(\{a\})$. Similarly, let B be the set of n tuples such that $x_k = b$. Then its probability is $P(\{b\})$. Finally, the probability of $A \cap B$ is $P(\{a\}) \cdot P(\{B\})$, by the same reasoning.

Now that we have established the independence of the X^j, notice that $X_n = X^1 + X^2 + \cdots + X^n$. For each of the X^1 we know that they have the same expectation and variance as the identity on the Bernoulli experiment, which by (127) has variance pq. Therefore,

$$(178)\qquad \mathsf{V}\,X_n = \sum_{\omega \in \{0,1\}^n} p_n(\omega)(X_n(\omega) - pn)^2 = npq$$

Inserting this back into (175) we get

$$(179)\qquad P\left(\left|\frac{X_n}{n} - p\right| \geq \varepsilon\right) \leq \frac{pq}{n\varepsilon^2}$$

Observe that $pq \leq 1/4$ so that we can make the estimate independent of p and q:

$$P\left(\left|\frac{X_n}{n} - p\right| \geq \varepsilon\right) \leq \frac{1}{4n\varepsilon^2} \tag{180}$$

Now choose ε as small as you like. Furthermore, choose the probability δ of deviation by at most ε as small as you like. We can choose an n independent on p and q such that performing the experiment at least n times will guarantee with probability $1 - \delta$ that the mean of the variable X will deviate from $\mathsf{E}\,X$ by at most ε. Indeed, just choose

$$n \geq \frac{1}{4\varepsilon^2 \cdot \delta} \tag{181}$$

and then

$$P\left(\left|\frac{X_n}{n} - p\right| \geq \varepsilon\right) \leq \frac{1}{4(4\delta\varepsilon^2)^{-1}\varepsilon^2} = \frac{4\delta\varepsilon^2}{4\varepsilon} = \delta \tag{182}$$

In mathematics, the fact that for large n the probability approaches a certain value is expressed as follows.

Definition 26 *Let $f(n)$ be a function from natural numbers to real numbers. We write $\lim_{n\to\infty} f(n) = b$ iff the following holds: for every $\varepsilon > 0$ there is an $n(\varepsilon)$ such that for all $n \geq n(\varepsilon)$ we have*

$$|f(n) - b| < \varepsilon \tag{183}$$

This says in plain words that for any error ε we choose there is a point from which on the values of the sequences are found within the error margin ε away from the value b. Such a statement is often found in statistics. We first name an interval (the **confidence interval**) within which values are claimed to fall and then we issue a probability with which they will actually fall there. The probability is often either a number close to 1, or it is a small number, in which case one actually gives probability that the values will *not* fall into the named interval.

This entails that the values of the sequences get closer to each other. Given this definition, we may now write

Theorem 27 *Let $\mathcal{P}$ be a Bernoulli space, and let X be a random variable. Define as above the random variable $X_n := \sum_{i=1}^{n} X^i$ on the n–fold product of $\mathcal{P}$ with itself. Then*

$$\lim_{n\to\infty} P\left(\left|\frac{X_n}{n} - \mathsf{E}\,X\right| \leq \varepsilon\right) = 0 \tag{184}$$

In plain language this means that with large *n* the probability that in an *n*–fold repetition of the experiment the mean of the results deviates from the expectation by any given margin is as small as we desire.

9 Limit Theorems

There is another way to establish bounds for n, and it involves a different technique of approximating the values of the binomials. Recall that the probability to get k out of n times the result 1 is

$$(185) \qquad P_n(k) = \binom{n}{k} \cdot p^k(1-p)^{n-k}$$

We shall show here that for large n the value of $P_n(k)$ can be approximated by a continuous function. As we shall see, there are several advantages to these theorems. One is that we have to know the values of only one function, namely $e^{-x^2/2}$ in order to compute these probabilities. However, the latter is quite a difficult function which cannot be calculated easily. This is why it used to be tabulated before there were any computers. However, one might think that when computers evaluate $P_n(k)$ they could do this without the help of the exponential function. This is not quite correct. The trouble is that the numbers occurring in the expression are very large and exceed the memory of a computer. For example, in $P_{1000}(712)$ we have to evaluate $\frac{1000!}{712! \cdot 288!}$. The numbers are astronomical! (These large numbers can be avoided through sophisticated methods, but the problem remains essentially the same: doing a lot of multiplications means accumulating errors.) Using the exponential function we can avoid all this. Moreover, the error we are making typically is quite small.

Theorem 28 (Local Limit Theorem) *Let* $0 < p < 1$. *Then*

$$(186) \qquad P_n(k) \approx \frac{1}{\sqrt{2\pi npq}} e^{-(k-np)^2/(2npq)}$$

uniformly for all k such that $|k - np| = O(\sqrt{npq})$.

We are not giving a proof of the theorem here. However, we shall explain the phrase that the formula holds uniformly for all k such that $|k - np| = O(\sqrt{npq})$. The claim is that the continuous function to the right approximates $P_n(k)$ for large n. This in turn means that for any error $\varepsilon > 0$, as small as we like, and any k there is an $n(\varepsilon)$ such that for all $n \geq n(\varepsilon)$ we have

$$(187) \qquad \left| P_n(k) - \frac{1}{\sqrt{2\pi npq}} e^{-(k-np)^2/(2npq)} \right| < \varepsilon$$

Moreover, it is claimed that $n(\varepsilon)$ can be chosen independently of k as long as in the limit $|k-np| \leq C\sqrt{npq}$. (But evidently, $n(\varepsilon)$ does depend on ε because the smaller the error the larger we have to choose the value $n(\varepsilon)$.) In practice this means that the approximation is good for values not too far apart from np; additionally, the closer p is to $1/2$ the better the approximation. It should be borne in mind that we only have an approximation here. Often enough one will use these formulae for finite experiments. The error that is incurred must be kept low. To do this, one has to monitor the difference $k - np$ as well as the bias p!

Now let us get back to the formula (186). Put

$$x := \frac{k - np}{\sqrt{npq}} \tag{188}$$

Then (186) becomes

$$P_n(np + x\sqrt{npq}) \approx \frac{1}{\sqrt{2\pi npq}} e^{-x^2/2} \tag{189}$$

Notice that while k is a discrete parameter (it can only take integer values), so is $np + x\sqrt{npq}$ in the left hand side of the equation. On the right hand side, however, we have a continuous function. This has in important consequence. Suppose we want to compute the sum over certain k within an interval. We shall replace this sum by a corresponding integral of the function on the right. In general, this is done as follows. There is a theorem which says that if f is a continuous function then for any two values x_1 and x_1 there is a ξ such that $x_1 \leq \xi \leq x_1$ and

$$f(\xi) = \frac{1}{x_1 - x_0} \int_{x_0}^{x_1} f(x)dx \tag{190}$$

(This is known as the mean value theorem.) It is particular interest because if the difference $x_1 - x_0$ gets small, also $f(\xi) - f(x_0)$ becomes small, so that we can write

$$f(x_0) \approx \frac{1}{x_1 - x_0} \int_{x_0}^{x_1} f(x)dx \tag{191}$$

We have used $\approx$ here to abbreviate the statement: for any $\varepsilon > 0$ there is a δ such that if $x_1 - x_0 < \delta$ then

$$\left| f(x_0) - \frac{1}{x_1 - x_0} \int_{x_0}^{x_1} f(x)dx \right| < \varepsilon \tag{192}$$

Equivalently, this is phrased as

$$f(x_0) = \lim_{x_1 \to x_0} \frac{1}{x_1 - x_0} \int_{x_0}^{x_1} f(x)dx \tag{193}$$

We now enter (191) into (189). The difference $x_{k+1} - x_k$ is exactly $\Delta = \sqrt{npq}^{-}1$ and so

$$\begin{aligned} P_n(np + x_k\sqrt{npq}) &\approx \frac{1}{\Delta} \cdot \frac{1}{\sqrt{2\pi}} e^{-x_k^2/2} \\ &\approx \frac{\sqrt{npq}}{\sqrt{2\pi npq}} \int_{x_k}^{x_k+\Delta} e^{-y^2/2} dy \\ &\approx \frac{1}{\sqrt{2\pi}} \int_{x_k}^{x_k+\Delta} e^{-y^2/2} dy \end{aligned} \tag{194}$$

As said above, this is justified for the reason that if is n large the function does not change much in value within the interval $[x_k, x_k + \Delta]$, so we may assume that by exchanging it for the constant function we are not getting a big error. (In fact, by the limit theorems we can make the error as small as we need it to be.) This leads to the following.

$$Q_n(a, b] := \sum_{a < x \leq b} P_n(np + x\sqrt{npq}) \tag{195}$$

Theorem 29 (De Moivre–Laplace)

$$\lim_{n \to \infty} \left| Q_n(a, b] - \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2} dx \right| = 0 \tag{196}$$

Thus, whatever the values are for p (and q), all we have to do is calculate a certain integral of the function $\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. In particular, putting

$$\Phi(x) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2} dy \tag{197}$$

we get

$$Q_n(a, b] \approx \Phi(b) - \Phi(a) \tag{198}$$

Unfortunately, the function cannot be given an analytical expression, so one cannot calculate it directly. Instead, must has to either use tables or the computer for the values of the function.

There are results that give estimates on the deviation of $\Phi(x)$ from the actual probabilities. Define first the following function.

$$F_n(x) := P_n(-\infty, x] = P\left(\frac{X_n - np}{\sqrt{npq}} \leq x\right) \tag{199}$$

This is a function from real numbers to real numbers, but it assumes only discrete values. For if $x_k \leq x < x_{k+1}$ then $F_n(x) = F_n(x_k)$, because the variable to the right can only assume the values x_m, $0 \leq m \leq n$. As can be seen, this function can be replaced by a sum:

$$F_n(x) = \sum_{x_k \leq x} P_n(k) \tag{200}$$

Notice that while P_n is a function from numbers to reals, F_n is a function from reals to reals, and uses the rescaled values.

Theorem 30 (Berry–Esseen)

$$\sup_{-\infty \leq x \leq \infty} |F_n(x) - \Phi(x)| \leq \frac{p^2 + q^2}{\sqrt{npq}} \tag{201}$$

In plain words this means that the greatest difference between the sum $F_n(x)$ (which is the actual probability) and the integral of the normal distribution is less than or equal to $\frac{p^2+q^2}{\sqrt{npq}}$. This is different from the limit since we additionally require that the value of $|F_n(x) - \Phi(x)|$ does not exceed the value of $\frac{p^2+q^2}{\sqrt{npq}}$. Thus, the error can be made uniformly small if n is chosen large. However, notice again that for small values for either p or q the estimates are rather poor, which in turn means that large n have to be considered. For example, let $p = 0.1$. Then $\frac{p^2+q^2}{\sqrt{pq}} = 0.811/\sqrt{0.09} = 0.811/0.3 \approx 2.7$, while for $p = q = 0.5$ we have $\frac{p^2+q^2}{\sqrt{pq}} = \frac{2p^2}{p} = 2p = 1$. The best bound is therefore $\frac{1}{\sqrt{n}}$, which means that to get error at most 0.1 you need $n \geq 100$, to get error at most 0.01 you need $n \geq 10000$. It should be stressed that one normally computes $P_n(a, b]$, which is $F_n(b) - F_n(a)$.

Again this number can be approximation by $\Phi(b) - \Phi(a)$. For the error the upper bound is twice the value given above:

$$\begin{aligned} &|(F_n(b) - F_n(a)) - (\Phi(b) - \Phi(a))| \\ \leq &|F_n(a) - \Phi(a)| + |F_n(b) - \Phi(x)| \\ \leq &2\frac{p^2 + q^2}{\sqrt{npq}} \end{aligned} \tag{202}$$

Notice, however, that the function $\Phi(x)$ has been obtained by transforming the values of k using (188). It is customary to express this transformation using expectation and standard deviation. Notice that np is the expectation of X_n; and that $\sqrt{npq}$ is the standard deviation of X_n. Let us drop the index n here. Then we may say the following. Let μ denote the expectation of X and σ its variance (which we may either obtain directly or by doing 'enough' experiments). The distribution of the variable X (for large n) is $\frac{1}{\sqrt{2\pi}}e^{-((x-\mu)/\sigma)^2}$.

Let us return to our experiment of Section 8. We shall use the exponential function to calculate the probabilities. To do this, notice that from (188) we calculate as follows.

$$\begin{aligned} x_0 &= (0 - 10 \cdot 0.7)/(\sqrt{10 \cdot 0.7 \cdot 0.3}) = -4.8305 \\ x_1 &= (1 - 10 \cdot 0.7)/(\sqrt{10 \cdot 0.7 \cdot 0.3}) = -4.1404 \\ x_2 &= (2 - 10 \cdot 0.7)/(\sqrt{10 \cdot 0.7 \cdot 0.3}) = -3.4503 \\ x_3 &= (3 - 10 \cdot 0.7)/(\sqrt{10 \cdot 0.7 \cdot 0.3}) = -2.7603 \\ x_4 &= (4 - 10 \cdot 0.7)/(\sqrt{10 \cdot 0.7 \cdot 0.3}) = -2.0702 \\ x_5 &= (5 - 10 \cdot 0.7)/(\sqrt{10 \cdot 0.7 \cdot 0.3}) = -1.3801 \\ x_6 &= (6 - 10 \cdot 0.7)/(\sqrt{10 \cdot 0.7 \cdot 0.3}) = -0.6901 \\ x_7 &= (7 - 10 \cdot 0.7)/(\sqrt{10 \cdot 0.7 \cdot 0.3}) = 0 \\ x_8 &= (8 - 10 \cdot 0.7)/(\sqrt{10 \cdot 0.7 \cdot 0.3}) = 0.6901 \\ x_9 &= (9 - 10 \cdot 0.7)/(\sqrt{10 \cdot 0.7 \cdot 0.3}) = 1.3801 \\ x_{10} &= (10 - 10 \cdot 0.7)/(\sqrt{10 \cdot 0.7 \cdot 0.3}) = 2.0702 \end{aligned} \tag{203}$$

(204)

k	$F_{10}(k)$	$\Phi(x_k)$
0	0	0.000
1	0.000	0.001
2	0.001	0.002
3	0.010	0.003
4	0.047	0.02
5	0.150	0.08
6	0.350	0.25
7	0.617	0.75
8	0.851	0.92
9	0.972	0.98

From our earlier estimates we expect that the error be less than

$$\frac{0.09 + 0.49}{\sqrt{10 \cdot 0.7 \cdot 0.3}} = \frac{0.58}{\sqrt{2.1}} = 0.40 \tag{205}$$

In effect, the precision is much better.

To view the effect via R, we define the following functions:

(206)

```
Hx <- function (n, p) ((0:n) - n *p)/((n * p *
    * (1 - p) ** 0.5)
Hy <- function (n, p) ((choose (n, 0:n) *
    * (0.7 ** (0:n)) * (0.3 ** (n - (0:n)))
    * ((2 * pi * n * p * (1 - p)) ** 0.5)
```

This defines for given n and p the vectors consisting of the points $\langle (i{-}np)/\sqrt{npq}, \binom{n}{i} p^i q^{n-i} \sqrt{2\pi npq} \rangle$. If you call the first coordinate x, the second coordinate will actually be approximately e^{-x^2}. To view the approximation, plot the function for increasing n (but identical p).

Part II

Elements of Statistics

10 Estimators

Suppose you are tossing a coin but you are not sure whether it is biased. How can you find out what the bias is? Evidently, there is no way to get a definitive answer. We have calculated the probabilities before and noticed that no sequence is impossible unless $p = 0$ or $p = 1$. Still we might want to know with at least great degree of confidence what the bias is. The way to this is as follows. Let θ be the bias, so it has values in $\Omega = [0, 1]$. Let $\omega = \langle \omega_1, \omega_2, \cdots, \omega_n \rangle$, and put $a(\omega) := |\{i : \omega_i = 1\}|$, and $b(\omega) := |\{i : \omega_i = 0\}| = n - a(\omega)$. Given θ we define the probability of an outcome as

(207) $P_\theta(\omega) = \theta^{a(\omega)}(1 - \theta)^{b(\omega)}$

The probability is a function of θ. As a point of notation and terminology: we may construe the situation in two ways. The first is that we have a family $\{P_\theta : \theta \in \Theta\}$ or probability functions on the space Ω^n. The other is that the space is $\Omega^n \times H$, where H is the following space: $H = \langle \Theta, \mathcal{B}(\Theta), \mu \rangle$, where $\mathcal{B}(\Theta)$ is the set of Borel sets over Θ and μ is the measure of the set. To simplify matters, $\mathcal{B}(\Theta)$ contains all finite unions of intervals of the form $(a, b]$ plus all the sets $\{a\}$, $0 \leq a, b \leq b$. Moreover, for an interval $(a, b]$ we put $\mu((a, b]) := b - a$ and $\mu(\{a\}) := 0$. $\mu(X)$ is known as the **Lebesgue–measure** of the set X.

In the latter case the outcomes are of the form $\zeta = \langle \omega, \theta \rangle$. Write $\pi_1(\zeta) = \omega$ (the first projection) and $\pi_2(\zeta) = \theta$. Then a statement of the form '$\omega \in S$' now is short for '$\pi_1(\omega) \in S$'. Additionally, however, we may introduce the notation H_θ for $\pi_2(\zeta) = \theta$. Thus, H_θ is the statement 'the bias is θ'. In order to make the notation perspicuous, we shall continue to use ω for the first part and θ for the second. The following become alternative notations, they are identical by the way things have been set up:

(208) $P_\theta(A) = P(A|H_\theta)$

It is quite frequent in textbooks to make the silent transition between families of probabilities over Ω^n and the space $\Omega^n \times H$ by writing statements such as $P(A|H_\theta)$. The latter are meaningless in the original space because H_θ is not an event of that space. On the other hand, if we perform an experiment we can only get at the value of ω, so one likes to think that one is dealing with the space $\mathcal{P}^n$ and that θ remains implicit in the definition of the probability. Since the numbers are the same, both viewpoints can be used simultaneously.

It is often the case that one does not need $P(\omega|H_\theta)$ but rather $P(\omega)$. How does one obtain this probability? Intuitively, $P(\omega)$ is the 'sum' of all $P(\langle\omega,\theta\rangle)$ where $\theta \in \Theta$. Also, in general,

$$P(\langle\omega,\theta\rangle) = P(\omega|H_\theta)P(H_\theta) \tag{209}$$

If we have only finitely many θ, say $\theta_1, \cdots \theta_n$, then we can simply write

$$P(\omega) = \sum_{i=1}^{n} P(\langle\omega,\theta_i\rangle) = \sum_{i=1}^{n} P(\omega|H_{\theta_i})P(H_{\theta_i}) \tag{210}$$

In the present case the set of values is the unit interval, or more generally, some set Θ of real numbers. In general one assumes that $P(H_\theta)$ does not depend on θ (we have a Laplace space). Then we may write $P(H_\theta) = d\theta$. The sum turns into an integral:

$$P(\omega) = \int_\Theta P(\omega|H_\theta)d\theta \tag{211}$$

Let us say that an **estimator** is a function T_n on Ω^n with values in Θ. (Notice that Θ can be any set of reals, but for the present example we obviously have $\Theta = [0,1]$.) The following is an estimator:

$$T_n(\omega) := a(\omega)/n \tag{212}$$

This is known as the **maximum likelihood estimator**. To see why, notice the following. We claim that $a(n)/n$ is actually the number that maximizes $P(H_\theta|\omega)$. To see this, notice that

$$P(\omega|H_\theta) = P(H_\theta|\omega) \cdot \frac{P(\omega)}{P(H_\theta)} \tag{213}$$

Suppose that we have no prior knowledge about the probabilities $P(H_\theta)$. Thus we assume that all H_θ are equally likely. Then the left hand side is maximal iff $P(H_\theta|\omega)$ is. That is, we have maximized the probability of H_θ under the hypothesis that ω precisely when we have maximized the probability of ω under the condition that H_θ. The latter probability is given by (207). Thus we aim to find the k such that (207) is maximal. There are two ways of doing this. One is to calculate the derivative of the function:

$$\begin{aligned}\frac{d}{d\theta}\left(\theta^{a(\omega)}(1-\theta)^{b(\omega)}\right) &= \left(\frac{d}{d\theta}\theta^{a(\omega)}\right)\cdot(1-\theta)^{b(\omega)} + \theta^{a(\omega)}\left(\frac{d}{d\theta}(1-\theta)^{b(\omega)}\right)\\ &= a(\omega)\theta^{a(\omega)-1}(1-\theta)^{b(\omega)} + b(\omega)(-1)\theta^{a(\omega)}(1-\theta)^{b(\omega)-1}\\ &= \theta^{a(\omega)-1}(1-\theta)^{b(\omega)-1}(a(\omega)(1-\theta) - b(\omega)\theta)\end{aligned} \tag{214}$$

(Recall that $\frac{d}{dx}(fg) = \left(\frac{d}{dx}f\right)g + f\left(\frac{d}{dx}g\right)$.) The maximum is attained at points where the value is 0. Apart from $\theta = 0$ or $\theta = 1$ this is only the case when

$$a(\omega)(1-\theta) - b(\omega)\theta = 0 \tag{215}$$

Equivalently,

$$a(\omega) = (a(\omega) + b(\omega))\theta \tag{216}$$

Since $a(\omega) + b(\omega) = n$ this becomes

$$\theta = \frac{a(\omega)}{n} \tag{217}$$

So, the maximum of $P(\omega|H_\theta)$ is attained when $a(\omega)$ equals $n\theta$. To be exact we would have to look at the cases $\theta = 0$ and $\theta = 1$. They are however completely straightforward: if $\theta = 0$ then the probability of ω is 0 except if $a(\omega) = 0$; if $\theta = 1$ then the probability of ω is 0 except for $a(\omega) = n$. Both validate the law that $a(\omega) = n\theta$. A last point is to be mentioned, though, and that is that $n\theta$ need not be an integer. However, we are interested in obtaining θ from $a(\omega)$ and not conversely, and a real number is fine.

Another route is this. Instead of dealing with the event ω, we think of the probability $P(H_\theta|X_n = a(\omega))$. Again this term is maximal if $P(X_n = a(\omega)|H_\theta)$ is maximal. It is

$$P_\theta(X_n = k) = \binom{n}{k}\theta^{a(\omega)}(1-\theta)^{b(\omega)} \tag{218}$$

The proof would go the same way as before. However, we can also make use of the limit theorems and the function $e^{-x^2/2}$. We have established that with the transformation $x = \frac{k-n\theta}{\sqrt{n\theta(1-\theta)}}$ it suffices for this to find the maximum of

$$f(x) = e^{-x^2/2} \tag{219}$$

This function is symmetric, that is $f(-x) = f(x)$. Moreover, if $0 < x < y$ then $f(x) > f(y)$, so without further calculations we can see that the maximum is attained at $x = 0$. Translating this back we find that $0 = \frac{k-n\theta}{\sqrt{n\theta(1-\theta)}}$, or $k = n\theta$. This means, given θ and n, k must be equal to $n\theta$. Or, as k was given through $a(\omega)$, and we in fact wanted to estimate θ such that H_θ becomes most probable, we must put $\theta := a(\omega)/n$ to achieve this.

It can also be phrased differently. Seen as a function from Ω^n to $[0,1]$ it is actually identical to X_n/n. In general, an estimator is a series of random variables over Ω^n for every n.

We say that T_n is **consistent** if for every $\varepsilon > 0$

(220) $$\lim_{n\to\infty} P_\theta(|T_n - \theta| > \varepsilon) = 0$$

This means the following. For every different outcome we get a possibly different estimate of θ. However, when we take θ as the true bias, the estimator shall not diverge from it in the limit. Of course, we do not know the bias θ, but we want to make sure that with respect to the real bias we get there eventually with any degree of certainty that we want. Since we know that X_n/n approaches θ in the limit we also know that the result is unambiguous: our estimator will converge in the long run and it will converge to the real bias.

Now call T_n **unbiased** if for all $\theta \in \Theta$:

(221) $$\mathsf{E}_\theta T_n = \theta$$

Here, E_θ is the expectation according to the probability P_θ. Finally, we call T_n **efficient** among the class $\mathfrak{U}$ of unbiased estimators if for all $\theta \in \Theta$

(222) $$\mathsf{V}\, T_n = \inf_{U_n \in \mathfrak{U}} \mathsf{V}_\theta\, U_n$$

This simply says that the estimator produces the least possible divergence from the sought value. It is obviously desirable to have an efficient estimator, because it gives us the value of θ with more precision than any other. The following summarizes the properties of the maximum likelihood estimator. It says that it is the best possible choice in the Bernoulli experiment (and an easy one to calculate, too).

Theorem 31 *X_n/n is consistent, unbiased and efficient.*

We have seen already that it is consistent and unbiased. Now,

(223) $$\mathsf{V}\, X_n/n = \frac{\theta(1-\theta)}{n}$$

It has to be shown that the variance of any other unbiased estimator is $\geq \theta(1-\theta)/n$. The proof is not easy, we shall therefore omit it.

Let us return to (213). The probability of H_θ is actually infinitesimally small. This is because there are infinitely many values in any close vicinity of θ, which also are good candidates for the bias (though their probability is slightly less than that of H_θ, a statement that can be made sense of even though all the numbers involved are infinitesimally small!). Thus, we cannot claim with any degree of certainty that the bias is θ. If we give ourselves a probability p of error, then all we can do is say that the bias is inside an interval $[a, b]$ with probability $1 - p$. Evidently, the values of a and b depend on p.

Definition 32 *An interval $[a^*(\omega), b^*(\omega)]$, with a^* and b^* functions from Ω to Θ, is called a* ***confidence interval of reliability*** *$1 - \delta$ or* ***significance level*** *δ if for all $\theta \in \Theta$:*

$$P_\theta(a(\omega) \leq \theta \leq b(\omega)) \geq 1 - \delta \tag{224}$$

So, the confidence level and the significance level are inversely correlated. If the confidence is 0.995 then the significance is 0.005. Very often in the literature one finds that people give the significance level, and this may cause confusion. The lower this number the better, so if the significance level is 0.001 then the probability of the result being in the interval is 99.9 percent, or 0.999!

We want to construct a confidence interval for a given confidence level. For a set $A \subseteq [0, 1]$ let us write H_A for the statement that $\theta \in A$. We are interested in $P(H_A|X_n = a(\omega))$. Using the law of inverted probabilities we can calculate this by

$$P(H_A|X_n = a(\omega)) = P(X_n = a(\omega)|H_A) \cdot \frac{P(X_n = a(\omega))}{P(H_A)} \tag{225}$$

This is less straightforward than it seems, for we now have to determine $P(H_A)$ and $P(X_n = a(\omega)|H_A)$. (Notice that $P(X_n = a(\omega)) = P(X_n = a(\omega)|H_\Theta)$, so the value is determined as well.) To make matters simple, we assume that $A = [a, b]$. Then $P(H_A) = b - a$, as the H_θ are all equiprobable. Now, using (218) we get

$$P(X_n = k|H_A) = \int_a^b \binom{n}{k} \theta^{a(\omega)}(1 - \theta)^{b(n)} d\theta \tag{226}$$

This integral can be solved. However, there is a much simpler solution (though the numbers might be less optimal). By Chebyshev's Inequality, for $T_n := V_n/n$,

$$P_\theta(|\theta - T_n| > \delta) \leq \frac{\mathsf{V}\, T_n}{\delta^2} = \frac{\theta(1 - \theta)}{n\delta^2} \tag{227}$$

Now, put

$$\lambda := \delta\sqrt{\frac{n}{\theta(1-\theta)}} \tag{228}$$

Then

$$P_\theta\left(|\theta - T_n| > \lambda\sqrt{\frac{\theta(1-\theta)}{n}}\right) < \frac{1}{\lambda^2} \tag{229}$$

or, equivalently,

$$P_\theta\left(|\theta - T_n| \leq \lambda\sqrt{\frac{\theta(1-\theta)}{n}}\right) \geq 1 - \frac{1}{\lambda^2} \tag{230}$$

To make this independent of θ, notice that

$$\theta(1-\theta) \leq 1/4 \tag{231}$$

so the equation can be reduced to

$$P_\theta\left(|\theta - T_n| \leq \frac{\lambda}{2\sqrt{n}}\right) \geq 1 - \frac{1}{\lambda^2} \tag{232}$$

Theorem 33 *Let $\mathcal{P}$ be a Bernoulli experiment with unknown bias θ. Based on an n–fold repetition of the experiment with result ω the value $a(\omega)/n$ falls into the interval $[\theta - \frac{\lambda}{2\sqrt{n}}, \theta + \frac{\lambda}{2\sqrt{n}}]$ with probability $1 - 1/\lambda^2$ (or with reliability $1/\lambda^2$).*

Or, put differently, θ is found in the interval $[a(\omega)/n - \frac{\lambda}{2\sqrt{n}}, a(\omega) + \frac{\lambda}{2\sqrt{n}}]$ with probability $1 - 1/\lambda^2$.

Thus, there is a correlation between certainty and precision. If we want to know the value of θ very accurately, we can only do so with low probability. If we want certainty, the accuracy has to be lowered.

11 Tests

Consider the following situation: you have an initial hypothesis H_0 and you consider another hypothesis H_1. Now you run an experiment. What does the outcome of the experiment tell you about the validity of the hypothesis? In particular, will the experiment suggest changing to H_1 or will it suggest to remain with the initial hypothesis? The solution comes in form of a so–called **test**. A test is a method that gives a recommendation whether we should adopt the new hypothesis or whether we should stay with the old one. We shall first analyse the situation of the preceding section. We have repeated an experiment n times and received an element ω. The element ω is also referred to as a **sample**, and Ω^n is the space of **sample points**. We shall agree the following definition.

Definition 34 *A **test** is a function d from the space of sample points to the set $\{H_0, H_1\}$. (d is also called a **decision rule**.) The set $d^{-1}(H_1)$ is called the **critical region** of d. d is **Bayesian** if for all ω, ω' such that $P(H_0|\omega) = P(H_0|\omega')$, $d(\omega) = d(\omega')$.*

It is clear that a test is uniquely determined by its critical region.

Very often one is not interested in the sample points as such but in some other derived value that they determine. For example, when estimating the bias we are really only interested in the value $a(\omega)/n$, because it gives us an approximation of the bias, as we have shown. If we had just taken ω the result would have been no better. The number $a(\omega)/n$ contains all information we need. We shall generalize this now as follows. A **statistic** is a function on Ω^n. A random variable on Ω^n is a statistic, since it is a function from that set into the real numbers. But statistics can go into any set one likes.

Let us give a few examples of statistics. Before we can do so, a few definitions. Suppose ω is a vector of numbers. Then let $\omega_{(i)}$ be the ith element of ω according to the order. For example, if $\omega = \langle 0, 9, 7, 0, 2, 1, 7\rangle$ then $\omega_{(1)} = 0$, $\omega_{(2)} = 0$, $\omega_{(3)} = 1$, $\omega_{(4)} = 2$, $\omega_{(5)} = 7$, $\omega_{(6)} = 7$ and $\omega_{(7)} = 9$. We can write ω in ascending order as follows.

$$(233) \qquad \langle \omega_{(1)}, \omega_{(2)}, \cdots, \omega_{(n)}\rangle$$

where n is the length of ω. In R, the sorting is done using the function `order`.

① The **sample mean**: $\overline{\omega} := \frac{1}{n}\sum_{i=1}^{n}\omega_i$;

② The **sample sum**: $\sum_{i=1}^{n}\omega_i$;

③ The **sample variance**: $s(\omega) := \frac{1}{n-1}\sum_{i=1}^{n}(\omega_i - \overline{\omega})^2$;

④ The **sample deviation**: $d(\omega) := \sqrt{s(\omega)}$;

⑤ The **order statistic**: $o(\omega) := \langle \omega_{(1)}, \omega_{(2)}, \cdots, \omega_{(n)}\rangle$;

⑥ The **sample median**: $m(\omega) := \begin{cases} \omega_{(n+1)/2} & \text{if } n \text{ is odd} \\ 1/2(\omega_{(n/2)} + \omega_{(1+n/2)}) & \text{if } n \text{ is even} \end{cases}$.

⑦ The **sample range**: $\omega_{(n)} - \omega_{(1)}$.

Now, we have a space Θ of values which we want to estimate. The estimator has been construed as a function on the sample space. But we can reduce the space to the space $(X_n/n)[\Omega_n]$. This means that the estimator does not distinguish between different ω as long as they receive the same value under X_n/n. We shall generalize this to the notion of a sufficient statistic. This is a statistic that contains as much information as is needed to get the value for the estimator. Given a statistic T, let $T = t$ denote the set $\{\omega : T(\omega) = t\}$.

Definition 35 *Let $\mathbb{P} = \{P_\theta : \theta \in \Theta\}$ be a family of probabilities on Ω^n and T a statistic. We say that T **sufficient** for $\mathbb{P}$ is for all θ, θ' we have $P_\theta(\omega|T = t) = P_{\theta'}(\omega|T = t)$.*

Trivially, the identity statistics is sufficient. A less trivial example is provided by the statistic $T_n(\omega) := a(\omega)$. To see that it is sufficient note that.

$$\begin{aligned} P_\theta(\omega|T_n(\omega) = t) &= \frac{P_\theta(\omega \cap (T_n = k))}{P_\theta(T_n = k)} \\ &= \frac{P_\theta(\omega)}{P_\theta(T_n = k)} \\ &= \frac{\theta^k(1-\theta)^{n-k}}{\binom{n}{k}\theta^k(1-\theta)^{n-k}} \\ &= \binom{n}{k}^{-1} \end{aligned} \tag{234}$$

This does not depend on θ.

Notice that in general that $P(T = t|\omega) = 1$ if $T(\omega) = t$ and 0 otherwise. Likewise, if $T(\omega) = t$ then $P(\omega \cap (T = t)) = P(\omega)$ and 0 otherwise. The latter has been used in the derivation above.

We shall prove now that a sufficient statistic allows to estimate parameters (or perform tests) with identical precision as the original sample. So, we assume that $P(\omega|(t = T \cap H_\theta)) = P(\omega|(t = T \cap H_{\theta'}))$.

From this we can deduce the following. Put

$$\xi(\omega) := P_1(T = t) \tag{235}$$

Then also $\xi(\omega) = P_\theta(T = t)$, by assumption.

$$\begin{aligned} P(\omega) &= \int_\Theta P(\omega|H_\theta)d\theta \\ &= \int_\Theta P(\omega|(T = t) \cap H_\theta)P((T = t)|H_\theta)d\theta \\ &= \int_\Theta \xi(\omega)P(T = t|H_\theta)d\theta \\ &= \xi(\theta)\int_\Theta P(T = t|H_\theta)d\theta \\ &= \xi(\omega)P(T = t) \end{aligned} \tag{236}$$

Thus, the probability of ω (independently of θ) also is the same fraction of the probability of $T = t$. This allows us to deduce the desired conclusion.

$$\begin{aligned} P(H_\theta|t = T) &= P(t = T|H_\theta) \cdot \frac{P(H_\theta)}{P(t = T)} \\ &= \frac{\xi(\omega)^{-1}P(\omega|H_\theta)P(H_\theta)}{\xi(\omega)^{-1}P(\omega)} \\ &= \frac{P(\omega|H_\theta)P(H_\theta)}{P(\omega)} \\ &= P(H_\theta|\omega) \end{aligned} \tag{237}$$

Thus the probabilities do not depend on ω as long as the statistic is the same. We derive from this discussion the following characterisation of sufficiency.

Theorem 36 *A statistic $T : \Omega^n \to A$ is sufficient iff there are functions $f : A \times \Theta \to \mathbb{R}$ and $g : \Omega^n \to \mathbb{R}$ such that $P(\omega|H_\theta) = f(T(\omega), \theta)g(\omega)$.*

Proof. We have seen above that if T is sufficient, $P(\omega|H_\theta) = P(T = t|H_\theta)\xi(\omega)^{-1}$. Hence, put $g(\omega) := \xi^{-1}$ and $f(T(\omega), \theta) := P(T = t|H_\theta)$, where $t := T(\omega)$. Conversely, suppose the functions f and g can be found as required. Then let ω and t be such that $T(\omega) = t$.

$$\begin{aligned} P(T = t|H_\theta) &= \sum_{T(\omega')=t} P(\omega'|H_\theta) \\ &= \sum_{\omega' \in T^{-1}(t)} f(T(\omega'), \theta)g(\omega') \\ &= f(T(\omega), \theta) \cdot \sum_{\omega' \in T^{-1}(t)} g(\omega') \end{aligned} \tag{238}$$

Notice namely that $f(T(\omega'), \theta) = f(T(\omega'), \theta)$ if $T(\omega') = T(\omega)$. Now,

$$\begin{aligned} P(\omega|T = t \cap H_\theta) &= \frac{P(\omega|H_\theta)}{P(T = t|H_\theta)} \\ &= \frac{f(T(\omega), \theta)g(\omega)}{P(T = t|H_\theta)} \\ &= \frac{t(T(\omega), \theta)g(\omega)}{f(T(\omega), \theta) \sum_{\omega \in T^{-1}(t)} g(\omega)} \\ &= \frac{g(\omega)}{\sum_{\omega' \in T^{-1}(t)} g(\omega')} \end{aligned} \tag{239}$$

This expression is independent of θ. ⊣

Let us note that if T is a sufficient statistic and d a test, we may consider using the test with critical region $T[C] := \{T(\omega) : \omega \in C\}$. Since it is not guaranteed that $C = T^{-1}(T[C])$ this may result in a different test. However, if the test is actually based on probabilities then it cannot differentiate between members of the partition. This is because $P(H_\theta|T = t) = P(H_\theta|\omega)$ implies that if $T(\omega) = T(\omega')$ then also $P(H_\theta|\omega) = P(H_\theta|\omega')$. Now, by definition, if d is Bayesian, $d(\omega) = d(\omega')$. So, tests can be applied to any sufficient statistic. We shall see below that there is also a minimal such statistic, and therefore decisions can be based on just that statistic.

With the tools developed so far we can fully analyse the situation. We need to assume that H_0 is the case with probability p, so that H_1 is true with probability

$q = 1 - p$. Let ω be a single outcome. We have

$$\begin{aligned} P(\omega) &= P(\omega|H_0)P(H_0) + P(\omega|H_1)P(H_1) \\ &= pP(\omega|H_0) + (1-p)P(\omega|H_1) \end{aligned} \tag{240}$$

Therefore we have

$$\begin{aligned} P(H_0|\omega) &= P(\omega|H_0) \cdot \frac{P(H_0)}{P(\omega)} \\ &= P(\omega|H_0) \cdot \frac{p}{pP(\omega|H_0) + (1-p)P(\omega|H_1)} \\ &= \frac{1}{1 + \frac{1-p}{p} \cdot \frac{P(\omega|H_1)}{P(\omega|H_0)}} \end{aligned} \tag{241}$$

and similarly for $P(H_1)$ we have

$$\begin{aligned} P(H_1|\omega) &= P(\omega|H_1) \cdot \frac{P(H_1)}{P(\omega)} \\ &= P(\omega|H_1) \cdot \frac{(1-p)}{pP(\omega|H_0) + (1-p)P(\omega|H_1)} \\ &= \frac{1}{1 + \frac{p}{1-p} \cdot \frac{P(\omega|H_0)}{P(\omega|H_1)}} \end{aligned} \tag{242}$$

Put $r := \frac{P(\omega|H_0)}{P(\omega|H_1)}$ and $c := \frac{p}{1-p}$. Then

$$P(H_0|\omega) := \frac{1}{1 + (cr)^{-1}}, P(H_1|\omega) := \frac{1}{1 + (cr)} \tag{243}$$

Thus, we can effectively determine what probability the hypotheses have given the experiment. But the crucial question is now what we can or have to deduce from that. This is where the notion of a decision method comes in. Call a **threshold test** a test that is based on a single number t, called **threshold** and which works as follows. If $R(\omega) := P(\omega|H_0)/P(\omega|H_1) < t$ then choose hypothesis H_1, otherwise choose H_0. (The case $P(\omega|H_1) = 0$ has to be excluded here; however, in that case H_0 must be adopted at all cost.) Based on the parameter t we can calculate the relevant probabilities.

Here is an example. Suppose the hypotheses are: H_0: the bias is 0.7 and H_1: the bias is 0.5. We perform the experiment 5 times. Let $\omega = \langle 0, 1, 0, 1, 1 \rangle$. Then

$$R(\omega) = \frac{P(\omega|H_0)}{P(\omega|H_1)} = \frac{0.7^3 \cdot 0.3^2}{0.5^5} \tag{244}$$

If we were to base our judgment on the statistic that counts the number of 1s, we get

$$\frac{\binom{5}{3}0.7^3 0.3^2}{\binom{5}{3}0.5^5} \tag{245}$$

which is exactly the same number. There are then only 6 numbers to compute. For given ω, we have

$$R(\omega) = \left(\frac{0.7}{0.5}\right)^{a(\omega)} \left(\frac{0.3}{0.5}\right)^{b(\omega)} = 1.4^{a(\omega)} 0.6^{n-a(\omega)} \tag{246}$$

(247)

Representative	Likelihood Ratio	
$\langle 0, 0, 0, 0, 0\rangle$	0.6^5	0.0778
$\langle 0, 0, 0, 0, 1\rangle$	$1.4 \cdot 0.6^4$	$0, 1814$
$\langle 0, 0, 0, 1, 1\rangle$	$1.4^2 \cdot 0.6^3$	$0, 4324$
$\langle 0, 0, 1, 1, 1\rangle$	$1.4^3 \cdot 0.6^2$	$0, 9878$
$\langle 0, 1, 1, 1, 1\rangle$	$1.4^4 \cdot 0.6$	2.305
$\langle 1, 1, 1, 1, 1\rangle$	1.4^5	5.3782

If we set the threshold to 1 then we adopt the hypothesis that the coin is biased with 0.7 exactly when the number of 1s exceeds 3. If it is does not, we assume that the bias is 0.5. Notice that if we get no 1s then the assumption that it is unbiased is also not likely in general, but more likely than that it is biased with 0.7. However, we have restricted ourselves to considering only the alternative between these two hypotheses.

Of course, ideally we would like to have a decision method that recommends to change to H_1 only if it actually *is* the case, and that it suggests that we remain with H_0 only if it actually *is* the case. It is clear however that we cannot expect this. All we can expect is to have a method that suggests the correct decision with a degree of certainty.

Definition 37 *A test is said to make a* ***type I error*** *when it suggests to adopt H_1 when H_0 is actually the case. A test is said to make a* ***type II error*** *when it suggests to adopt H_0 when H_1 actually is the case.*

It is generally preferred to have a low probability for a type I error to occur. (This embodies a notion of conservativity. One would rather remain with the old hypothesis than change for no good reason.) We have actually calculated above the probabilities for a type I and type II error to occur. They are

$$\text{(248)} \qquad \text{probability of type I error} \qquad \frac{1}{1+\frac{1-p}{p}\cdot\frac{P(B|H_1)}{P(B|H_0)}}$$

$$\text{(249)} \qquad \text{probability of type II error} \qquad \frac{1}{1+\frac{p}{1-p}\cdot\frac{P(B|H_0)}{P(B|H_1)}}$$

Definition 38 *The probability of a type I error is called the* ***significance of the test****, the probability of the nonoccurrence of a type II error the* ***power of the test****. A test T is* ***most significant*** *if for all test T' whose significance exceeds that of T the power of T' is strictly less than that of T; T is* ***most powerful*** *if for all tests T' whose power exceeds that of T the significance of T' is strictly less than the significance of T.*

In statistical experiments the significance is the most common number used. It describes the probability with which the test falsely recommends to change to the new hypothesis. But the power is obviously equally important. Now let us define

$$\text{(250)} \qquad R(\omega) := \frac{P(X=\omega|H_0)}{P(X=\omega|H_1)}$$

This number describes the ratio of the likelihood that ω is the outcome under condition that H_0 divided by the likelihood that ω is the case under condition that H_1. The function R is called the **likelihood ratio statistic**. We shall show that it is sufficient and moreover that it is minimally sufficient.

Definition 39 *A statistic T is* ***minimally sufficient*** *if for all sufficient statistics S there is a function h such that $T(\omega) = h(S(\omega))$.*

Thus, further compression of the data is not possible if we have applied a minimally sufficient statistic.

Theorem 40 *The likelihood ratio statistic is minimally sufficient.*

Proof. First we show that it is sufficient. We use Theorem 36 for that. Define

$$f(R(\omega), \theta) := \begin{cases} R(\omega)^{1/2} & \text{if } \theta = \theta_0 \\ R(\omega)^{-1/2} & \text{if } \theta = \theta_1 \end{cases} \tag{251}$$

Further, put

$$g(\omega) := (P(\omega|\theta_0)P(\omega|\theta_1))^{1/2} \tag{252}$$

Then

$$\begin{aligned} f(R(\omega), \theta_0) &= \frac{P(\omega|\theta_0)}{P(\omega|\theta_1)}(P(\omega|\theta_0)P(\omega|\theta_1))^{1/2} \\ &= P(\omega|\theta_0) \end{aligned} \tag{253}$$

and

$$\begin{aligned} f(R(\omega), \theta_1) &= \frac{P(\omega|\theta_1)}{P(\omega|\theta_0)}(P(\omega|\theta_0)P(\omega|\theta_1))^{1/2} \\ &= P(\omega|\theta_1) \end{aligned} \tag{254}$$

So, the statistic is sufficient. Now take another sufficient statistic T. Take a sample ω and put $t := T(\omega)$. Since T is sufficient,

$$P(\omega|T = t \cap H_0) = P(\omega|T = t \cap H_1) \tag{255}$$

Therefore,

$$\begin{aligned} R(\omega) &= \frac{P(\omega|\theta_0)}{P(\omega|\theta_1)} \\ &= \frac{P(\omega \cap T = t \cap H_0)/P(H_0)}{P(\omega \cap T = t \cap H_1)/P(H_1)} \\ &= \frac{P(\omega|T = t \cap H_0) \cdot P(T = t \cap H_0)/P(H_0)}{P(\omega|T = t \cap H_1) \cdot P(T = t \cap H_1)/P(H)} \\ &= \frac{P(T = t|H_0)}{P(T = t|H_1)} \end{aligned} \tag{256}$$

Now, if $T(\omega) = T(\omega')$ then it follows from this equation that also $R(\omega) = R(\omega')$, and so the map $h : T(\omega) \mapsto R(\omega)$ is well–defined and we have $R(\omega) = h(T(\omega))$. ⊣

A **threshold test** with threshold t is a test that recommends H_0 if $R(\omega) > t$. We notice that the power and significance of the test depend on the value for t.

The higher t the more likely the test is to recommend H_1 thus the higher t the greater the power of the test. The lower t the more likely the test is conservative and therefore the more significant it is. Again we see that there is an antagonism between significance and power. However, we note that no matter what t is chosen, the test is optimally significant within the set of tests of same (or less) power and optimally powerful within the set of tests that have equal (or lesser significance).

Theorem 41 (Neyman & Pearson) *The threshold test is most significant and most powerful.*

Let us continue the previous example. We have calculated the values of $R(\omega)$. Now we wish to know for given threshold t, what the significance and power of the test is.

Here is a useful table of probabilities.

(257)

k	$P(T = k \vert H_0)$	$P(T = 0 \vert H_1)$
0	0.03125	0.0025
1	0.15625	0.02835
2	0.3125	0.1323
3	0.3125	0.3087
4	0.15625	0.36015
5	0.03125	0.16807

From this we draw the following values, assuming $P(H_0) = P(H_1) = 0.5$:

(258)

	H_0		H_1	
t	$R < t$	$R \geq t$	$R < t$	$R \geq t$
0	0	0.5	0	0.5
0.1	0.0150625	0.4849375	0.00125	0.49875
0.4	0.09375	0.40625	0.015425	0.484575
0.9	0.25	0.25	0.0816	0.4184
2	0.40625	0.09375	0.235925	0.264075
5	0.4849375	0.0150625	0.416	0.084
6	0.5	0	0.5	0

The significance and power can be read off the second and the third column; the second is the probability that H_0 obtains but the test advises against it. The third is the probability that H_1 obtains and the test against it (so, take 1 minus the value of the third column to obtain the power). Notice a few extreme cases. If $t = 0$ then the rule never advises to adopt H_1. So, a type I error cannot occur. The significance is optimal. The power on the other hand is 0.5, because that is the probability that H_1 obtains, and we are bound to be wrong then. Now assume that $t = 9$. This value is never reached; we always adopt H_1. So the significance is 0.5, the worst possible value, because if H_0 obtains we get an error. On the other hand, the power is 0.

12 Distributions

We have discussed probability spaces in the first sections. In this section we shall summarize some results on probability distributions and introduce some more, whose relevance have proved essential in probability theory and statistics. We begin with the discrete spaces.

Uniform Distribution This distribution has only one parameter, the size N of the space. The probability of each individual outcome is the same, and it is $1/N$.

Binomial Distribution This distribution has three parameters, p, n and k. The outcomes are the numbers from 1 to n, and

$$(259) \qquad P(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

This distribution arises from the n–fold product of a Bernoulli space by introducing the random variable $X(\langle \omega_1, \omega_2, \cdots, \omega_n \rangle) := \sum_{i=1}^{n} \omega_i$ and then turning the value range into a probability space.

Geometric Distribution The underlying space is the set of natural numbers. This distribution has one parameter, p.

$$(260) \qquad P(k) = \frac{(1-p)^{-k}}{p}$$

Polynomial Distribution The underlying space is the set of natural numbers. There is one parameter, α, and the probabilities are

$$(261) \qquad P(k) = \zeta(\alpha)^{-1} \frac{1}{k^\alpha}$$

where by definition

$$(262) \qquad \zeta(\alpha) := \sum_{k=1}^{\infty} \frac{1}{k^\alpha}$$

Now we turn to continuous distributions. In the continuous case the distribution is defined with the help of its density $f(x)$, and the probability $P((a, b])$ is defined by

$$P((a, b]) = \int_a^b f(x)dx \tag{263}$$

Uniform Distribution The space is a closed interval $[a, b]$. The density is

$$f(x) = \frac{1}{b-a} \tag{264}$$

Normal Distribution The space is $\mathbb{R}$, and the distribution has two parameters, μ and σ, where $\sigma > 0$. The density is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-(x-\mu)^2/2\sigma^2} \tag{265}$$

If $\mu = 0$ and $\sigma = 1$ the distribution is called **standard normal**.

Exponential Distribution The space is $\mathbb{R}^+$, the space of positive real numbers, and there is one parameter, λ, which must be strictly positive (ie $\lambda > 0$).

$$f(x) = \lambda e^{-\lambda x} \tag{266}$$

Gamma Distribution The space is $\mathbb{R}$ and the distribution has two real parameters, α and λ which must both be strictly positive. The density is

$$f(x) = \frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)} \tag{267}$$

This uses the Γ–function, which is defined as follows (for all $t > 0$).

$$\Gamma(t) := \int_0^\infty x^{t-1} e^{-x} dx \tag{268}$$

The Gamma distribution arises as follows. Suppose that X_i, $i = 1, 2, \cdots, n$, are independent random variables which are distributed exponentially with parameter

λ. Then the sum $Y = \sum_{i=1}^{n} X_i$ is a gamma–distributed random variable with parameters $\alpha = n$ and λ. It follows that the exponential distribution is a special case of the gamma–distribution by putting $n = 1$.

We collect some useful facts about the Γ–function. By the method of partial integration,

$$\begin{aligned}\Gamma(t) &= \int_0^\infty x^{t-1}e^{-x}dx \\ &= \frac{x^t}{t}e^{-x}\Big|_0^\infty + \frac{1}{t}\int_o^\infty x^t e^{-x}dx \\ &= \frac{\Gamma(t+1)}{t}\end{aligned} \tag{269}$$

In other words

$$\Gamma(t+1) = (t+1)\Gamma(t) \tag{270}$$

It is useful to note also that $\Gamma(1) = 1$, because $\int_0^\infty e^{-x}dx = -e^{-x}|_0^\infty = -0 + 1 = 1$. From this it follows that $\Gamma(n) = n!$, so this function generalises the factorial. Interesting for purposes of statistics is

$$\begin{aligned}\Gamma(1/2) &= \int_0^\infty x^{-1/2}e^{-x}dx \\ &= \int_0^\infty \frac{\sqrt{2}e^{-y^2/2}}{\sqrt{2}x^{1/2}}dx \\ &= \frac{1}{\sqrt{2}}\int_0^\infty e^{-y^2/2}dy \\ &= \frac{1}{\sqrt{2}}\int_{-\infty}^\infty e^{-y^2/2}dy \\ &= \sqrt{\pi}\end{aligned} \tag{271}$$

To see this, notice that we have put $y := (2x)^{1/2}$, so that $x = y^2/2$. Further, $dy/dx = (2x)^{-1/2}$, or $dx = \sqrt{2x}dy$. This explains the step from the second to the third line. Now, $\frac{1}{\sqrt{2\pi}}\int_{-\infty}^\infty e^{-y^2/2}dy = 1$, and so the remaining equations easily follow. We can derive from these the following result.

Theorem 42 *The gamma function assumes the following values.*

① $\Gamma(n) = n!$

② $\Gamma(n/2) = \frac{(2n)!\sqrt{\pi}}{2^{2n}n!}$

Chi–squared distribution This is a special case of the Γ–distribution. The space is $\mathbb{R}^+$. There is a single parameter, n, which assumes values in $\mathbb{N}$. This number is also called the **degrees of freedom**. The density is Γ–distributed with $\lambda = 1/2$ and $\alpha = n/2$.

$$\chi^2_n(x) = \frac{x^{n/2-1}}{\sqrt{2^n e^x}\Gamma(n/2)} \tag{272}$$

This distribution arises as follows. Let X_i, $i = 1, 2, \cdots, n$ be independent random variables which are all standard normally distributed. Then $Y = \sum_{i=1}^{n} X_i^2$ is a random variable, whose distribution is χ^2 with n degrees of freedom.

We shall explain why this is so. First, take the simplest example, $Y = X^2$. We want to know the distribution of Y. Notice that Y can have only positive values; and that if $Y = a$ then X may be either $+\sqrt{a}$ or $-\sqrt{a}$. So, we get the following probability.

$$P(Y \leq a) = 2P(X < \sqrt{a}) \tag{273}$$

Call the probability distribution of Y F. Then we deduce

$$\int_0^a F(y)dy = \frac{2}{\sqrt{2\pi}} \int_0^{\sqrt{a}} e^{-x^2/2} dx \tag{274}$$

Now, $e^{-x^2/2} = e^{-y/2}$, but what about dx? Here we use some magic (which nevertheless is rigorous mathematics!).

$$dy = \frac{dy}{dx}dx = \frac{dx^2}{dx}dx = 2xdx \tag{275}$$

From this we get $dx = dy/2x = dy/2\sqrt{y}$, and so we have

$$\int_a^b F(y)dy = \frac{2}{\sqrt{2\pi}} \int_{x=\sqrt{a}}^{x=\sqrt{b}} e^{-y/2} \frac{dx}{2\sqrt{y}} = \frac{1}{\sqrt{2\pi}} \int_{y=a}^{y=b} \frac{e^{-y/2}}{\sqrt{y}} dy \tag{276}$$

From this we deduce now that

$$F(y) = \frac{e^{-y/2}}{\sqrt{2\pi}\sqrt{y}} \tag{277}$$

This is precisely the value given above. To see this, we only need to recall that $e^{-y/2} = 1/\sqrt{e^y}$ and that $\Gamma(1/n) = \sqrt{\pi}$.

Now take the next case, $Y = X_1^2 + X_2^2$. Here, matter get somewhat more involved. We want to get the probability that $Y = a$. This means that $|X_1| \leq \sqrt{a}$ and that $|X_2| \leq \sqrt{a - X_1^2}$. For both X_1 and X_2 we can either choose a positive value or a negative one.

$$\int_0^a F_2(y)dy = \frac{4}{2\pi}\int_0^{\sqrt{a}}\int_0^{\sqrt{a-x_1^2}} e^{-x_1^2/2}e^{-x_2^2/2}dx_1dx_2 \tag{278}$$

This is a heavy weight to lift. However, notice first that $e^{-x_1^2/2}e^{-x_2^2/2} = e^{-(x_1^2+x_2^2)/2} = e^{-y/2}$. So, we are integrating $e^{-y/2}$ for all values x_1 and x_2 such that $x_1^2 + x_2^2 = y$. If we fix y then the values $\langle x_1, x_2\rangle$ are on a quarter circle, starting with $\langle 0, \sqrt{y}\rangle$ and ending with $\langle \sqrt{y}, 0\rangle$. The length of the line along which we are integrating is exactly $2\pi\sqrt{y}/4$, because the radius of the circle is $\sqrt{y}$. So, we switch to the new coordinates y and φ, where φ is the angle from the x–axis to the vector pointing from the origin to $\langle x_1, x_2\rangle$ (these are called **polar coordinates**). Now,

$$x_1 = \sqrt{y}\cos\varphi, \qquad x_2 = \sqrt{y}\sin\varphi \tag{279}$$

From this we deduce that

$$\frac{dx_1}{dy} = \frac{1}{2\sqrt{y}}\cos\varphi, \qquad \frac{dx_2}{dy} = \sqrt{y}\cos\varphi \tag{280}$$

So we have

$$\begin{aligned}
\int_0^a F_2(y)dy &= \frac{2}{\pi}\int_{y=0}^{y=a}\int_{\varphi=0}^{\varphi=\pi/2} e^{-y/2}dx_1dx_2 \\
&= \frac{2}{\pi}\int_{y=0}^{y=a}\int_{\varphi=0}^{\varphi=\pi/2} e^{-y/2}\sqrt{y}d\sqrt{y}\cos\varphi\frac{2}{\sqrt{y}}\cos\varphi\varphi dy \\
&= \frac{2}{\pi}\int_{y=0}^{y=a}\int_{\varphi=0}^{\varphi=\pi/2} e^{-y/2}\cos^2\varphi d\varphi dy \\
&= \frac{2}{\pi}\int_0^\infty e^{-y/2}dy\int_0^{\pi/2}\cos^2\varphi d\varphi
\end{aligned} \tag{281}$$

Now we have to solve $\int_0^{\pi/2} \cos^2 \varphi d\varphi$. Partial integration yields.

$$\begin{aligned}\int_0^{\pi/2} \cos^2 \varphi d\varphi &= \cos\varphi \sin\varphi|_0^{\pi/2} - \int_0^{\pi/2} (-\sin\varphi)\sin\varphi d\varphi \\ &= \int_0^{\pi/2} \sin^2 \varphi d\varphi \\ &= \int_0^{\pi/2} (1 - \cos^2 \varphi) d\varphi \\ &= \frac{\pi}{2} - \int_0^{\pi/2} \cos^2 \varphi d\varphi \end{aligned} \tag{282}$$

And so it follows that

$$\int_0^{\pi/2} \cos \varphi d\varphi = \frac{\pi}{4} \tag{283}$$

We insert this into (281) and continue:

$$\begin{aligned}\int_0^a F_2(y)dy = \frac{2\alpha}{\pi}\frac{\pi}{4}\int_0^\infty e^{-y/2}dy \quad &= \frac{1}{2}(-2)e^{-y/2}\Big|_0^a \\ = e^0 - e^{-a/2} \\ = 1 - e^{-a/2}\end{aligned} \tag{284}$$

The formula offered above is (with $n = 2$)

$$\chi_2^2(y) = \frac{y^0}{2\sqrt{e^y}\Gamma(2/2)} = \frac{e^{-y/2}}{2} \tag{285}$$

For higher dimensions the proof is a bit more involved but quite similar. We change to polar coordinates and integrate along the points of equal distance to the center. Instead of dx_i, $i = 1, 2, \cdots, n$ we integrate over y, φ_j, $j = 2, 3, \cdots, n$.

F–distribution There are two parameters, n_1 and n_2. If X is χ^2 with n_1 degrees of freedom and Y is χ^2 with n_2 degrees of freedom, then the variable $Z := (X/n_1)/(Y/n_2) = (n_2X)/(n_1Y)$ has a distribution called **F–distribution** with degrees of freedom (n_1, n_2). Its density function is

$$f(x) = \frac{(\Gamma((n_1 + n_2)/2)}{\Gamma(n_1/2)\Gamma(n_2/2)}\left(\frac{n_1}{n_2}\right)^{n_1/2} x^{(n_1/2)-1}\left(1 + \frac{n_1 x}{n_2}\right)^{-(n_1+n_2)/2} \tag{286}$$

The mean of this distribution is $\frac{n_2}{n_2-2}$ (independent of n_1!) and the variance is

$$\frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)^2(n_2-4)} \tag{287}$$

Student (or *t*–)distribution Like the χ^2–distribution, this one has a natural number as its sole parameter. Suppose that X^2 has an F–distribution with degrees of freedom $(1, n)$ and that X is distributed symmetrically around zero (we need this condition because the square root may be positive or negative). Then X is said to have **t–distribution** with $(1, n)$ degrees of freedom.

$$t_n(x) := \frac{\Gamma(\frac{1}{2}(n+1))}{\sqrt{n\pi}\Gamma(n/2)}\left(1+\frac{x^2}{n}\right)^{-(n+1)/2} \tag{288}$$

The mean of this distribution is 0 (as it is symmetric), and the variance is $\frac{n}{n-2}$.

13 Parameter Estimation

In this section we shall generalize the problem and illustrate the general pattern of reasoning in statistics. The problem is as follows. Suppose you have a certain space of outcomes and events, and what you need is the probabilities. This task is generally speaking quite impossible to solve. A similar problem is this. You have a random variable X on a space and you wish to infer its distribution. That is to say, your space may be considered the space of real numbers (or a suitable part thereof), and you want to establish the probability density. That is to say, you want to establish the function $f(x)$ such that the probability that X is found in the interval $[a, b]$ is

(289) $$\int_a^b f(x)dx$$

The way to do this is to run an experiment. The experiment gives you data, and on the basis of this data one establishes the distribution. There are plenty of examples. For example, we measure the voice onset time of the sounds of the English sound [b]. Most likely, this is not a fixed number, but it is distributed around a certain number t, and it is expected to decrease with increasing distance to t. We may measure grammaticality judgments of given sentences in a scale and find a similar pattern. We may estimate the probability distribution of words according to their rank by taking a text and counting their frequency. And so on.

It is to be stressed that in principle there is no way to know which probability density is the best. We always have to start with a simple hypothesis and take matters from there. For example, suppose we measure the voice onset times of [b] with many speakers, and many samples for each speaker. Then we get the data, but which density function are we to fit? Typically, one makes a guess and says that the voice onset time is normally distributed around a certain value, so the probability density is

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

We shall discuss this example, as it is the most widespread. What we need to establish is just two numbers: μ and σ. This radically simplifies the problem. We shall see that there are ways to estimate both numbers in a very simple way. But we shall also see that the proofs that this is so are quite difficult (and we shall certainly not expose them in full here).

First let us look at μ. If the normal distribution is just the limit of the binomial distribution then we know what μ is: it is the mean. Now, suppose that we are perform the experiment n times with result $\vec{s} = \langle s_1, s_2, \cdots, s_n \rangle$. This is our **sample**. Now we define the number

$$m(\vec{s}) := \frac{1}{n}\sum_{i=1}^{n} s_i \tag{291}$$

This is called the **sample mean**. The sample mean, as we stressed before, does not have to be the same as the number μ. However, what we claim that the sample mean is an unbiased and efficient estimator of the mean μ. This means in words the following:

① under the assumption that the mean is μ, the sample mean converges towards μ as n grows large.

② among all possible estimators, the sample mean has the least variance.

We shall prove the first claim only. We introduce n independent random variables $X_1, X_2, \ldots, X_n$, representing the values for the experiments $1, 2, \ldots, n$. It says that if the mean is μ, and we draw a sample of size n, we should expect the sample mean to be μ. The sample mean is $\overline{X} = 1/n \sum_{i=1}^{n} X_i$.

$$\mathsf{E}(\overline{X}) = \frac{1}{n}\sum_{i=1}^{n} \mathsf{E}(X_i) = \frac{1}{n} \cdot n \cdot \mu = \mu \tag{292}$$

This is because each random variable is distributed according to the same distribution, and the mean is μ. If the variance of the distribution is known, then we can actually compute the variance of $\overline{X}$ as well:

$$\begin{aligned} \mathsf{V}(\overline{X}) &= \sum_{i=1}^{n} \mathsf{V}\left(\frac{X_i}{n}\right) \\ &= \frac{n}{n^2}\mathsf{V}(X) \\ &= \frac{\sigma}{n^2} \end{aligned} \tag{293}$$

It is another matter to show that the sample variance also approaches that value, so that the estimator is shown to be efficient. We shall skip that part. Instead we shall offer a way to quantify the certainty with which the sample allows to estimate μ. We base this for the moment on the assumption that we know the value of σ.

The sample mean $\overline{X}$ is also normally distributed. The same holds for $Y := \sqrt{n}(\overline{X} - \mu)/\sigma$. This is called the **adjusted sample mean**. What it does is the following. We adjust the distributions to standard normal distributions by subtracting the mean and dividing by the standard deviation. If we do that, the new sample mean is Y. Moreover, the distribution of $\overline{Y}$ is standardized: the mean is 0 and the variance is 1. Now, suppose we want to announce our result with certainty p. That is to say, we want to announce numbers a and b such that μ falls within the interval $[a, b]$ with probability p. Based on $\overline{Y}$, we must ask for numbers a^* and b^* such that $P(Y \in [a^*, b^*]) \geq p$. Once we have a^* and b^* we get a and b as

$$(294) \qquad a = \frac{a^*\sigma}{\sqrt{n}} + \overline{X}, \qquad b = \frac{b^*\sigma}{\sqrt{n}} + \overline{X}$$

Since the value is expected to be 0 (for Y), the interval is of the form $[-a^*, a^*]$, so that

$$(295) \qquad a = \overline{X} - \frac{a^*\sigma}{\sqrt{n}}, \qquad b = \overline{X} + \frac{b^*\sigma}{\sqrt{n}}$$

Now, to find a^*, we need to use our tables. We want to solve

$$(296) \qquad \int_{-a^*}^{a^*} e^{-x^2/2} dx = \Phi(a^*) - \Phi(-a^*)$$

There is a simpler solution. Observe that the function $e^{-x^2/2}$ is symmetric around the origin. Hence

$$(297) \qquad \begin{aligned} \int_{-a^*}^{a^*} e^{-x^2/2} dx &= 2\int_0^{a^*} e^{-x^2/2} dx \\ &= 2(\Phi(a^*) - \Phi(0)) \\ &= 2\Phi(a^*) - 1 \end{aligned}$$

(Notice that $\Phi(0) = 1/2$.) Recall that we have specified p in advance. Now we have

$$(298) \qquad p = 2\Phi(a^*) - 1$$

From this we get

$$\Phi(a^*) = \frac{p+1}{2} \tag{299}$$

So, the procedure is therefore this. Given p, we establish a^* by lookup (or with the help of the computer), using the formula (299). This we enter into formula (295) and establish the actual confidence interval for μ.

Similar considerations reveal that if μ is known, then the sample variance is an efficient unbiased estimator of the variance σ. However, this is in practice not a frequently occurring situation. Very often we need to estimate both μ and σ. In this case, however, something interesting happens. The sample variance based on μ is this:

$$\frac{1}{n}\sum_{i=1}^{n}(s_i - \mu)^2 \tag{300}$$

But now that we also have to estimate μ, things become more complex. First, it again turns out that μ is approximated by the sample mean. Consider the random variable for the deviation:

$$s_X := \sqrt{\sum_{i=1}^{n}\left(\frac{(X_i - \overline{X})}{n}\right)^2} \tag{301}$$

Let us calculate the expected value of the variance:

$$\begin{align}
\mathsf{E}\, s_X^2 &= \frac{1}{n}\,\mathsf{E}\left(\sum_{i=1}^{n}(X_i - \overline{X})^2\right) \notag\\
&= \frac{1}{n}\left(\mathsf{E}\sum_{i=1}^{n}((X_i - \mu) - (\overline{X} - \mu))^2\right) \notag\\
&= \frac{1}{n}\left[\left(\sum_{i=1}^{n}(X_i - \mu)^2\right) + n\,\mathsf{E}((\overline{X} - \mu)^2) - 2\,\mathsf{E}\left((\overline{X} - \mu)\sum_{i=1}^{n}(X_i - \mu)\right)\right] \notag\\
&= \frac{1}{n}\left[n\sigma^2 + n\frac{\sigma^2}{n} - 2\,\mathsf{E}\left((\overline{X} - \mu)\left(\left(\sum_{i=1}^{n} X_i\right) - n\mu\right)\right)\right] \tag{302}\\
&= \frac{1}{n}\left[n\sigma^2 + \sigma^2 - 2n\,\mathsf{E}((\overline{X} - \mu)(n\overline{X} = n\mu))\right] \notag\\
&= \frac{1}{n}\left[n\sigma^2 + \sigma^2 - 2n\,\mathsf{E}((\overline{X} - \mu)^2)\right] \notag\\
&= \frac{1}{n}\left[n\sigma^2 + \sigma^2 - 2n\cdot\frac{\sigma^2}{n}\right] \notag\\
&= \frac{(n-1)\sigma^2}{n} \notag
\end{align}$$

It turns out that the sample variance is *not* an unbiased estimator of the variance! If it were, the result would have been σ^2. This suggests the following

Definition 43 *The* ***unbiased sample variance*** *for a sample of size n is*

$$\hat{s}_X := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{303}$$

Of course, calling this estimator unbiased calls for a proof. However, notice that $\hat{s}_X^2 = \frac{n}{n-1}s_X^2 = \frac{n}{n-1}\frac{n-1}{n}\sigma^2 = \sigma^2$, so this is easily established. We skip the efficiency part. We notice only that the calculations show that $\hat{s}_X$ is only asymptotically efficient; this means that we cannot be guaranteed for any n that its variance of the sample variance is the least possible value, but for growing n it approaches the least value with any given error.

It is perhaps good to reflect on the problem of the discrepancy between variance and sample variance. As seen, the factor is $\frac{n}{n-1}$. Where does it come from?

Statisticians like to express this in terms of degrees of freedom. Suppose you draw a sample of size n. Independently of the variance of the distribution, the mean can be estimated by the sample mean. Now the variance, based on the mean, is calculated straightforwardly. However, now that we have used the data once to extract the mean, we have eliminated one degree of freedom from it. We have fixed the mean to be the sample mean, and we have used the data to do this. One says that we have removed one degree of freedom. To see this in more clear terms, notice that we my also present our data as follows:

$$(304) \qquad \overline{s}_* := \langle \overline{s}, s_1, s_2, \cdots, s_n \rangle$$

This looks like $n + 1$ data points, but the truth is that the sum of the last n terms is actually $n\overline{s}$. Thus, one of the terms completely depends on the others, and can therefore be dropped. We decide to drop s_1.

$$(305) \qquad \overline{s}_* := \langle \overline{s}, s_2, \cdots, s_n \rangle$$

Finally, when we calculated the variance, we have to add the squares of all terms with the mean subtracted. But the contribution of the first term is guaranteed to be 0. So, our new data contains only $n - 1$ terms rather than n.

If this argumentation sounds fishy, here is another one. In probabilities terms, the situation can also be described as follows. We must rescale our expectations of the variance by the knowledge that the sample mean is what it is. In other words, once the sample mean is calculated, the expectation of the variance is tilted just because the sample isn't drawn freely any more: it has to have to have that mean. We have to calculate the expectation of variance on condition that $\mu = \overline{X}$. This ought to slightly diminish the variance, and it does so by the factor $\frac{n}{n-1}$.

Finally, let us return to our estimation problem. We have to address the question of how we can extract a confidence interval for given p. This is not easy since we must now estimate both μ and σ simultaneously. The solution is to consider the value

$$(306) \qquad T := \sqrt{\frac{n(\overline{X} - \mu)^2}{\hat{s}_X^2}}$$

This number is known to be distributed according to the t–distribution with $n - 1$ degrees of freedom. This is because it is the square root of the quotient of

$n(\overline{X}-\mu)^2/\sigma^2$ (adjusted sample mean) and $(n-1)\hat{s}_X^2/\sigma^2$ (adjusted sample variance). These numbers are independent χ^2–random variables with degree of freedom 1 and n respectively. So, T^2 follows the F–distribution, and T follows the t–distribution. Hence we can do the following. Suppose we wish to have a confidence interval for p. First, $t_{n-1}(x) = t_{n-1}(-x)$ so that the $100p$th confidence interval for the mean is

$$\left[\overline{X} - t_{n-1}((1+p)/2)\sqrt{\frac{\hat{s}_X^2}{n}}, \overline{X} + t_{n-1}((1+p)/2)\sqrt{\frac{\hat{s}_X^2}{n}}\right] \tag{307}$$

For the variance σ^2 we get the following interval

$$\left[\frac{(n-1)\hat{s}_X^2}{\chi^2_{n-1}((1+p)/2)}, \frac{(n-1)\hat{s}_X^2}{\chi^2_{n-1}((1-p)/2)}\right] \tag{308}$$

14 Correlation and Covariance

We begin by a general question. Let us have two random variables X and Y, one of which, say X we consider as observable, while the other one, Y, is hidden. How much information can we obtain about Y from observing X? We have already seen plenty of such situations. For example, we may consider space to consist of pairs $\langle\vec{x}, y\rangle$, where $\vec{x}$ is the result of an n-fold iteration of a Bernoulli experiment with $p = y$. The variable $\vec{x}$ is observable, while y is hidden. We have ways to establish the probability of $\langle\vec{x}, y\rangle$ and they allow us to gain knowledge about y from $\vec{x}$. A special case of this is when we have two random variables X and Y over the same space. There is an important number, called the **covariance**, which measures the extent to which X reveals something about Y. It is defined by

$$\text{(309)} \qquad \mathsf{cov}(X, Y) := \mathsf{E}(X - \mathsf{E}\,X)(Y - \mathsf{E}\,Y)$$

We notice that $\mathsf{cov}(X, Y)$ is symmetric and linear in both arguments:

$$\text{(310)} \qquad \begin{aligned} \mathsf{cov}(X, Y) &= \mathsf{cov}(Y, X) \\ \mathsf{cov}(X, Y_1 + Y_1) &= \mathsf{cov}(X, Y_1) + \mathsf{cov}(X, Y_2) \\ \mathsf{cov}(X, aY) &= a\,\mathsf{cov}(X, Y) \\ \mathsf{cov}(X_1 + X_2, Y) &= \mathsf{cov}(X_1, Y) + \mathsf{cov}(X_2, Y) \\ \mathsf{cov}(aX, Y) &= a\,\mathsf{cov}(X, Y) \end{aligned}$$

If $\mathsf{cov}(X, Y) = 0$ then X are said to be **uncorrelated**. Notice that random variables may be uncorrelated but nevertheless not be independent. An example is $X := \sin\alpha$ and $Y = \cos\alpha$ where $\Omega = \{0, \pi/2, \pi\}$ with equal probability. We find $X(0) = X(\pi) = 0$, $X(\pi/2) = 1$. $Y(0) = 1$, $Y(\pi/2) = 0$ and $Y(\pi) = -1$. Now, $\mathsf{E}\,X = 1/3(0+1+0) = 1/3$ and $\mathsf{E}\,Y = 1/3(1+0-1) = 0$. Finally, $Z := (X-\mathsf{E}\,X)(Y-\mathsf{E}\,Y)$ takes the following values:

$$\text{(311)} \qquad \begin{aligned} Z(0) &= (0 - 1/3)(1 - 0) = -1/3 \\ Z(\pi/2) &= (1 - 1/3)(0 - 0) = 0 \\ Z(1) &= (0 - 1/3)(-1 - 0) = 1/3 \end{aligned}$$

The expectation of Z is

$$\text{(312)} \qquad \mathsf{E}\,Z = 1/3(-1/3 + 0 + 1/3) = 0$$

So, X and Y are uncorrelated. But they are not independent. For example,

$$P(X = 1 \cap Y = 0) = 1/3 \neq P(X = 1)P(Y = 0) = 1/9. \tag{313}$$

The following connects the variance and the covariance:

$$\mathsf{V}(X + Y) = \mathsf{V}\,X + \mathsf{V}\,Y + 2\,\mathsf{cov}(X, Y) \tag{314}$$

For a proof note that

$$\begin{aligned}\mathsf{V}\,X + \mathsf{V}\,Y + 2\,\mathsf{cov}(X, Y) &= \mathsf{E}\,X^2 - (\mathsf{E}\,X)^2 + \mathsf{E}\,Y^2 - (\mathsf{E}\,Y)^2 + \\ &\quad + 2\,\mathsf{E}(X - \mathsf{E}\,X)(Y - \mathsf{E}\,Y) \\ &= \mathsf{E}\,X^2 + 2\,\mathsf{E}(XY) + \mathsf{E}\,Y^2 - (\mathsf{E}\,X)^2 - (\mathsf{E}\,Y)^2 - \\ &\quad - 4\,\mathsf{E}\,Y\,\mathsf{E}\,X + 2\,\mathsf{E}(\mathsf{E}\,X)(\mathsf{E}\,Y) \\ &= \mathsf{E}(X^2 + 2XY + Y^2) - \\ &\quad - ((\mathsf{E}\,X)^2 + 2(\mathsf{E}\,X)(\mathsf{E}\,Y) + (\mathsf{E}\,Y)^2) \\ &= \mathsf{E}(X + Y)^2 - (\mathsf{E}(X + Y))^2 \\ &= \mathsf{V}(X + Y)\end{aligned} \tag{315}$$

To see this, observe that $\mathsf{E}(X\,\mathsf{E}\,Y) = \mathsf{E}\,Y\,\mathsf{E}\,Y = \mathsf{E}\,X\,\mathsf{E}\,Y$.

Additionally, the **correlation coefficient** $\rho(X, Y)$ is defined as

$$\rho(X, Y) := \frac{\mathsf{E}(X - \mathsf{E}\,X)(Y - \mathsf{E}\,Y)}{\sqrt{\mathsf{V}\,X \cdot \mathsf{V}\,Y}} \tag{316}$$

We find that $\rho(X, X) = \mathsf{cov}(X, X)/\,\mathsf{V}\,X$. However, $\mathsf{cov}(X, X) = \mathsf{E}(X - \mathsf{E}\,X)^2 = \mathsf{V}\,X$, so that $\rho(X, X) = 1$. Furthermore, it is easily seen that $\rho(X, -X) = -1$. These are the most extreme cases.

Proposition 44 $-1 \leq \rho(X, Y) \leq 1$.

Proof. I shall give the argument in the discrete case. In that case, we have to show that

$$|\,\mathsf{E}(X - \mathsf{E}\,X)(Y - \mathsf{E}\,Y)| \leq \sqrt{\mathsf{V}\,X \cdot \mathsf{V}\,Y} \tag{317}$$

Now,

$$\text{(318)}\qquad \mathsf{E}(X - \mathsf{E}\,X) = \sum_{\omega}(X(\omega) - \mathsf{E}\,X)p(\omega)$$

$$\text{(319)}\qquad \mathsf{V}(X) = \sum_{\omega}(X(\omega) - \mathsf{E}\,X)^2 p(\omega)$$

Let $\mathbb{R}^{\Omega}$ be the real vector space over Ω. Introduce the vectors $\vec{X}_*$ and $\vec{Y}_*$ by $\vec{X}_*(\omega) := (X(\omega) - \mathsf{E}\,X)/\sqrt{p(\omega)}$ and $\vec{Y}_*(\omega) := (Y(\omega) - Y(\omega))/\sqrt{p(\omega)}$. Then (317) becomes

$$\text{(320)}\qquad \left|\sum_{\omega}\vec{X}_*(\omega)\vec{Y}_*(\omega)\right| \leq \sqrt{\sum_{\omega}\vec{X}_*(\omega)^2 \cdot \sum_{\omega}\vec{Y}_*(\omega)^2}$$

(320) expresses the following:

$$\text{(321)}\qquad |\vec{X}_* \cdot \vec{Y}_*| \leq \sqrt{(\vec{X}_* \cdot \vec{X}_*)(\vec{Y}_* \cdot \vec{Y}_*)} = |\vec{X}_*||\vec{Y}_*|$$

In other words, this is the well known vector identity: the scalar product $\vec{x} \cdot \vec{y}$ is the length of $\vec{x}$ times the length of $\vec{y}$ times the cosine of the angle between them. So the correlation coefficient is the cosine between the random variables viewed as vectors. ⊣

We see from the proof also that $\rho(X, Y)$ is 1 or -1 just in case Y is a linear multiple of X. Thus the correlation coefficient effectively measures to what degree X and Y are linearly dependent. If Y is not linear multiple of X, but some other function, say $Y = X^2$, then the correlation is some number other than 1 or -1. (We have seen above that the correlation coefficient can also be 0.) Also, from a geometrical viewpoint it becomes clear that if the correlation coefficient is 0 the vectors need not be independent. All this says is that $\vec{X}_*$ is orthogonal to $\vec{Y}_*$.

A pair of variables is called **Gaussian** if its distribution is as follows:

$$\text{(322)}\qquad P(X = x \cap Y = y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-m_1)^2}{\sigma_1^2} - \right.\right.$$
$$\left.\left. - 2\rho\frac{(x-m_1)(y-m_2)}{\sigma_1\sigma_2} + \frac{(y-m_2)^2}{\sigma_2^2}\right]\right\}$$

This definition can be extended to vectors of random variables of arbitrary length, but this makes the picture no clearer. It turns out that $\rho = \rho(X, Y)$. If $\rho = 0$ then the function above reduces to

$$P(X = x \cap Y = y) = \frac{1}{2\pi\sigma_1\sigma_2}\exp\left\{-\frac{1}{2}\left[\frac{(x-m_1)^2}{\sigma_1^2} + \frac{(y-m_2)^2}{\sigma_2^2}\right]\right\} \tag{323}$$

We also find that

$$P(X = x) = \frac{1}{\sqrt{2\pi}\sigma_1}e^{-(x-m_1)^2/2\sigma_1^2} \tag{324}$$

$$P(Y = y) = \frac{1}{\sqrt{2\pi}\sigma_1}e^{-(y-m_1)^2/2\sigma_1^2} \tag{325}$$

And so we obtain that

$$P(X = x \cap Y = y) = P(X = x) \cdot P(Y = y) \tag{326}$$

Theorem 45 *Let X and Y be Gaussian random variables. If X and Y are uncorrelated they are independent.*

15 Linear Regression I: The Simple Case

The following sections draw heavily on [6], which is a rather well-written book on regression. It can obviously cover much more than can be done here, so if further clarification is needed, the reader is advised to consult this book.

Consider an experiment where data has been collected, for example, where one measures fuel consumption. At the same time, one has same information available, such as taxation, state where one lives in, age, income, and so on. What one finds is that the null hypothesis, that the consumption is simply independent of all these variables, seems quite unlikely. One is convinced that any of the variables might be a factor contributing the the effect that we are measuring. The question is: what is the precise effect of any of the factors involved?

We consider first a simplified example: we assume that all factors enter linearly. So, we are assuming a law of the following form. The variables that we are given are called X_1, X_2, and so on. They all come from different populations. The variable that we wish to explain is called Y. The variables whose values we consider given are called **predictors**, the variable we want to explain in terms of the predictors is called indexresponse**response**. Ideally we would like to have the correspondence

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \tag{327}$$

In practice, such an exact correspondence is never found. Instead, we allow for an additional error ε, so that the equation now becomes

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon \tag{328}$$

The condition on ε is that it is normally distributed and has mean 0. That is, the density of the distribution of ε is $\frac{1}{\sqrt{2\pi\sigma}} e^{-x^2/s\sigma}$. This last condition may appear arbitrary, but it is a way to ensure that we do not use ε as a garbage can for effects that we are just unable to account for.

When we have made our measurements, it is our first task to find the numbers β_i. This is done as follows. Look at Table 1 for definitions of terms. In that table, n is the number of repetitions of the experiments. The running index for the sums is i, and it runs from 1 to n. We assume to have only two variables, X and Y, from which we get the sample points $\vec{x} = \langle x_1, x_2, \cdots, x_n \rangle$ and $\vec{y} = \langle y_1, y_2, \cdots, y_n \rangle$. These give rise to the statistics shown in the table. They are easily generalised to

Table 1: Definitions of Sample Statistics

Symbol	Statistic	Description
$\overline{x}$	$\sum x_i/n$	Sample average
$\mathrm{S}XX$	$\sum(x_i - \vec{x})^2$	Sample sum of squares
SD_x^2	$\mathrm{S}XX/(n-1)$	Sample variance
SD_x	$\sqrt{\sum \mathrm{S}XX/(n-1)}$	Sample standard deviation
$\mathrm{S}XY$	$\sum(x_i - \overline{x})(y_i - \overline{y})$	Sum of cross-products
s_{xy}	$\mathrm{S}XY/(n-1)$	Sample covariance
r_{xy}	$s_{xy}/\mathrm{SD}_x\mathrm{SD}_y$	Sample correlation

the case where we have more than one variables. We discuss first the case of a single explanatory variable. In this case, we shall have to establish three numbers: β_0, β_1 and σ. It is customary to put a hat on a number that is computed from a sample. Thus, while we assume that there is a number β_0 that we have to establish, based on a sample we give an estimate of the number and call it $\hat{\beta}_0$. For example:

$$(329)\qquad \begin{aligned} \hat{\beta}_1 &:= \frac{\mathrm{S}XY}{\mathrm{S}XX} = r_{xy}\left(\frac{\mathrm{S}XY}{\mathrm{S}XX}\right)^{1/2} \\ \hat{\beta}_0 &:= \overline{y} - \hat{\beta}_1\overline{x} \end{aligned}$$

These numbers estimate our dependency of Y on X_1. We estimate the error of fit as follows:

$$(330)\qquad \mathrm{RSS}(\gamma_0, \gamma_1) := \sum_{i=1}^{n}(y_i - (\gamma_0 + \gamma_1 x_i))^2$$

This is called the **residual sum of squares**. Notice that even with the true numbers β_0 and β_1 in place of γ_0 and γ_1 we get some residual. Now that we have estimated these numbers we compute the residual sum of squares and estimate thus the variance of the error. However, notice that we have taken away two degrees of freedom, as wa have established two numbers already. Thus the following is an unbiased estimator for σ^2:

$$(331)\qquad \hat{\sigma}^2 := \frac{RSS(\hat{\beta}_0, \hat{\beta}_1)}{n-2}$$

It is immediately clear that these are just estimates and therefore we must provide confidence intervals for them. It can be shown that the estimators are unbiased.

Moreover, $\hat{\beta}_0$ and $\hat{\beta}_1$ can be shown to be normally distributed, since they are linear functions of the y_i, which depend on the one hand on the x_i, on the other hand are normally distributed for given x_i, by assumption on the error. Be aware, though, that we have not claimed anything for the real parameters, β_0 and β_1. To give confidence intervals, however, we can only judge the matter on the basis of the data, and then the true values are assumed fixed. Thus, the y_i do depend linearly on the x_i with a given normal error. Thus, to establish the confidence intervals, we only need to estimate the variance:

$$\begin{aligned} \mathsf{V}(\hat{\beta}_1) &= \sigma^2 \frac{1}{\mathrm{S}XX} \\ \mathsf{V}(\hat{\beta}_0) &= \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\mathrm{S}XX} \right) \end{aligned} \tag{332}$$

These are the true values. To get estimates, we replace the true σ by its estimate:

$$\begin{aligned} \widehat{\mathsf{V}}(\hat{\beta}_1) &= \hat{\sigma}^2 \frac{1}{\mathrm{S}XX} \\ \widehat{\mathsf{V}}(\hat{\beta}_0) &= \hat{\sigma}^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\mathrm{S}XX} \right) \end{aligned} \tag{333}$$

Notice that rather than the true variance, we now have the estimated variance, so must also put a hat on the variance. Knowing that the parameters are standardly distributed, we can now give confidence intervals in the standard way, based on a distribution with mean $\hat{\beta}_i$ and variance $\hat{\mathsf{V}}(\hat{\beta}_i)$.

What we would like to know now is how good this actually is in terms of explaining the data. There are several test that describe the significance of the data. We shall discuss them now, including a new one, the R value.

The first test is to establish whether or not adding the variable X_1 as an explanatory variable actually was necessitated by the data. The null hypothesis is that X_1 is not an explanatory variable; formally, this is the hypothesis that $\beta_1 = 0$. Thus the test we are applying is to decided between the following alternatives, with (NH) referring to the null hypothesis and (AH) to the alternative hypothesis:

$$\begin{aligned} &(\mathrm{NH})\ \mathsf{E}(Y|X = x) = \beta_0 \\ &(\mathrm{AH})\text{for some } \beta_1 \neq 0\text{: } \mathsf{E}(Y|X = x) = \beta_0 + \beta_1 x \end{aligned} \tag{334}$$

To assess this, we use the F statistic:

$$F := \frac{(SYY - \mathrm{RSS})/1}{\hat{\sigma}^2} \tag{335}$$

This number is distributed according to the F-distribution with $(1, n-2)$ degrees of freedom. It is thus possible to compute the F value but also the t value (the square root of F) as well as the p value. The latter is, as observed earlier, the probability of obtaining a data that has a t value which is as extreme as the one we get from our data.

Finally, there is an additional number which is of great interest. It is R, where

$$R^2 := 1 - \frac{\mathrm{RSS}}{SYY} \tag{336}$$

The number R^2 is called the **coefficient of determination**. It measures the strength of prediction of X_1 for the value Y. Observe that

$$R^2 = \frac{(SXY)^2}{SXX \cdot SYY} = r_{xy}^2 \tag{337}$$

Thus, R is nothing but the correlation of X and Y. Since the numbers are calculated with correction by the mean, the correlation would be 1 or -1 if $\sigma = 0$. This is hardly ever the case, though.

A final check of the goodness of the approximation is actually a look at the residuals. By definition, these are

$$e_i := y_i - (\beta_0 - \beta_1 x_i) \tag{338}$$

Again, as we are now dealing with approximations, the only thing we can actually compute are

$$\hat{e}_i := y_i - (\hat{\beta}_0 - \hat{\beta}_1 x_i) \tag{339}$$

We have required that these values are normally distributed and that the mean is 0. Both assumptions must be checked. The mean is easily computed. However, whether or not the residuals are normally distributed is not easy to assess. Various ways of doing that can be used, the easiest of which is a plot of the residuals over fitted values (they should look random), or a Q-Q-plot.

We provide an example. If you load the package ”alr3” you will find a data set called `forbes.txt`, which describes data found that correlates the boiling temperature of water as a function of the pressure. It is a simple text file, containing three columns, labeled as ”Temp”, ”Pressure” and ”Lpres”. The data is loaded into R and printed by

```
        For <- read.table("/usr/lib/R/library/
            alr3/data/forbes.txt", header=TRUE)
(340)
        plot(For$Temp, For$Pressure, xlab="Pressure",
            ylab="Temperature")
```

The result is shown in Figure 1. As a next step we compute the regression line. One way to do this is to use `lm` as follows.

```
        attach(For)
(341)   forl <- lm(Temp ˜ Pressure)
        summary(forl)
```

This summary will give us a host of additional information, but for the moment we are happy just with the values of the constants. The **intercept**, β_0, is estimated to be 155.296 and the factor β_1 to be 1.902. Let us do another plot, this time inserting the regression line. For example, with Pressure equal to 22.4 we get Temperature: $155.296 + 1.902 \cdot 22.4 = 197.9008$ (against 197.9) and with pressure 30.06 we get $155.3 + 30.06 \cdot 1.9 = 212.4701$ against the measured 212.2. The fit appears to be good.

```
        x <- c(22.4, 20.06)
        y <- 155.296 + 1.902 * x
        pdf (file = "forbes-wl.pdf")
(342)   plot(Temp, Pressure, xlab="Pressure",
            ylab="Temperature")
        lines(xy.coords(x,y), col="red")
        dev.off ()
```

This produced the graphics in Figure 2. To complete the analysis, we now plot the error, shown in Figure 3. A close look at this plot reveals that there is one point which is actually ”odd”. It should probably be removed from the set because it

Figure 1: The Temperature over Pressure

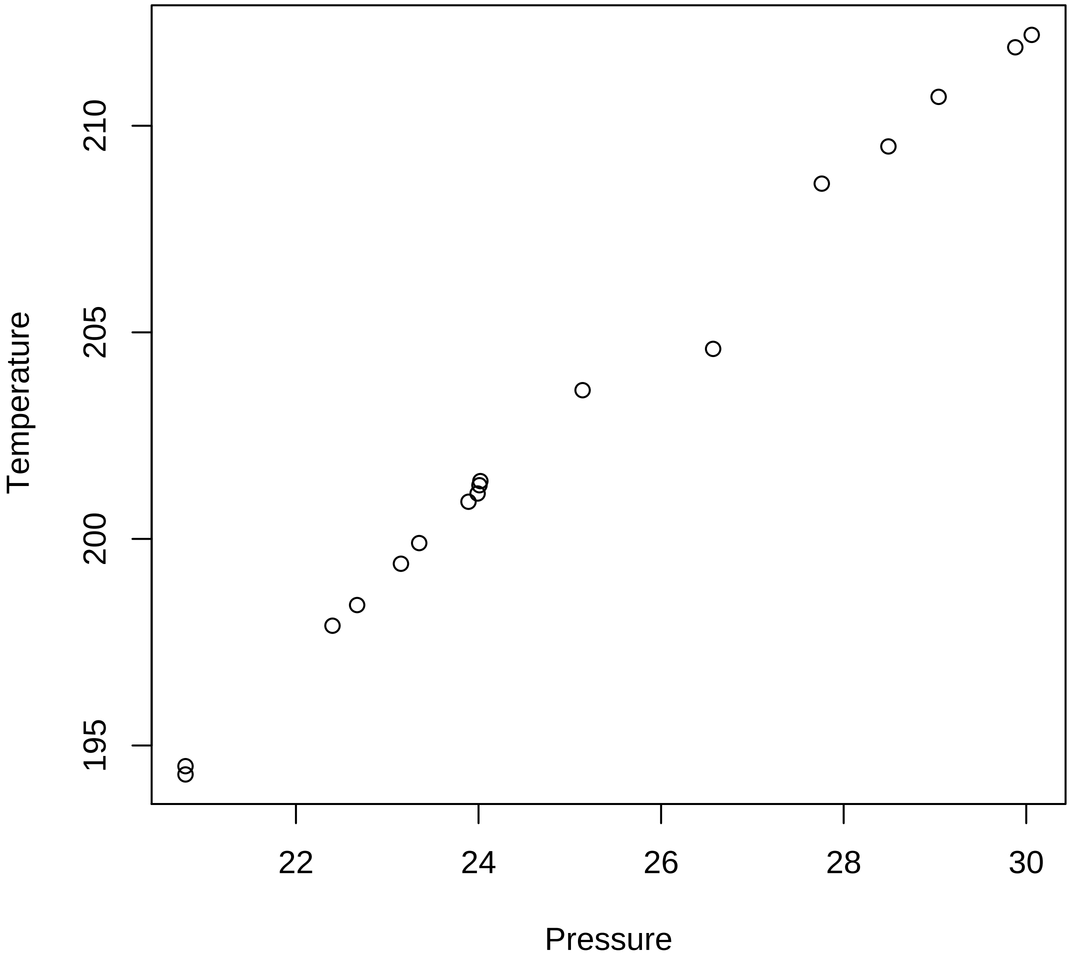


Figure 2: The Temperature over Pressure with Regression Line

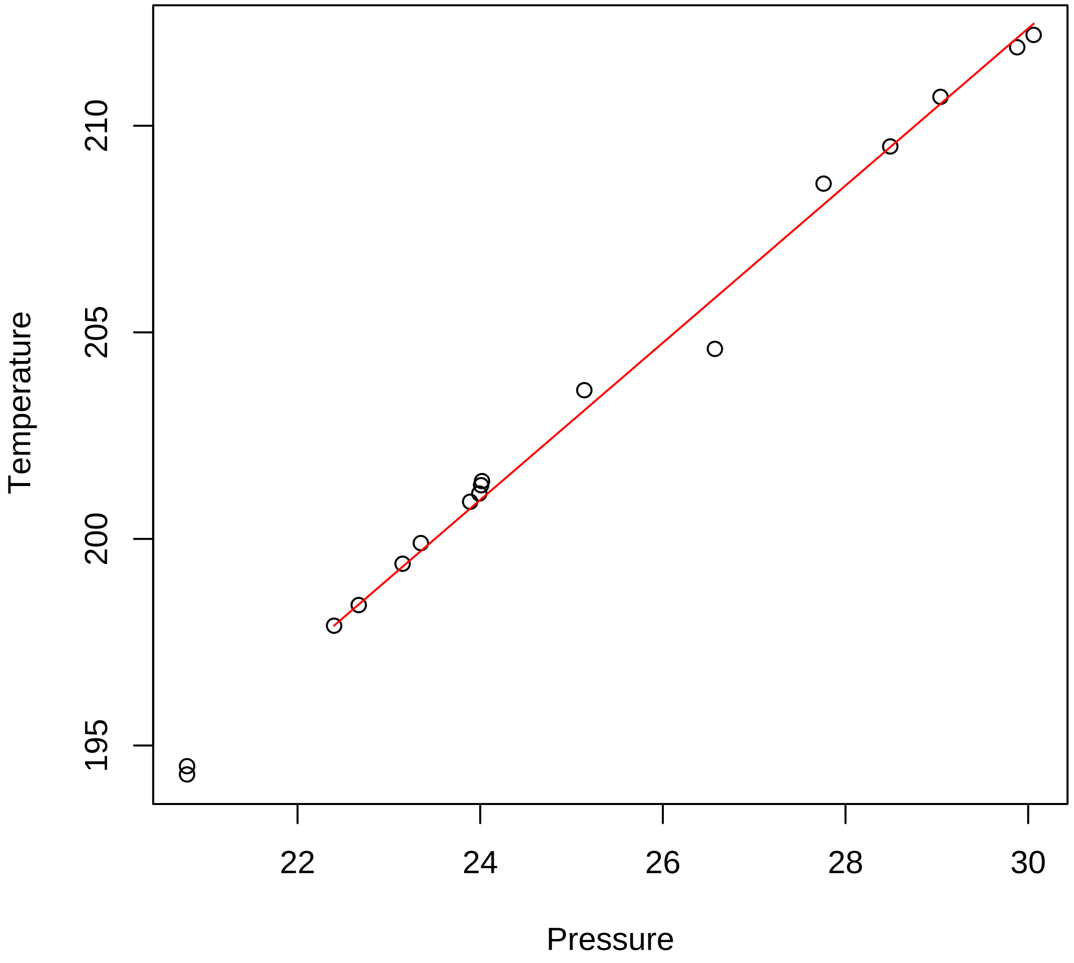
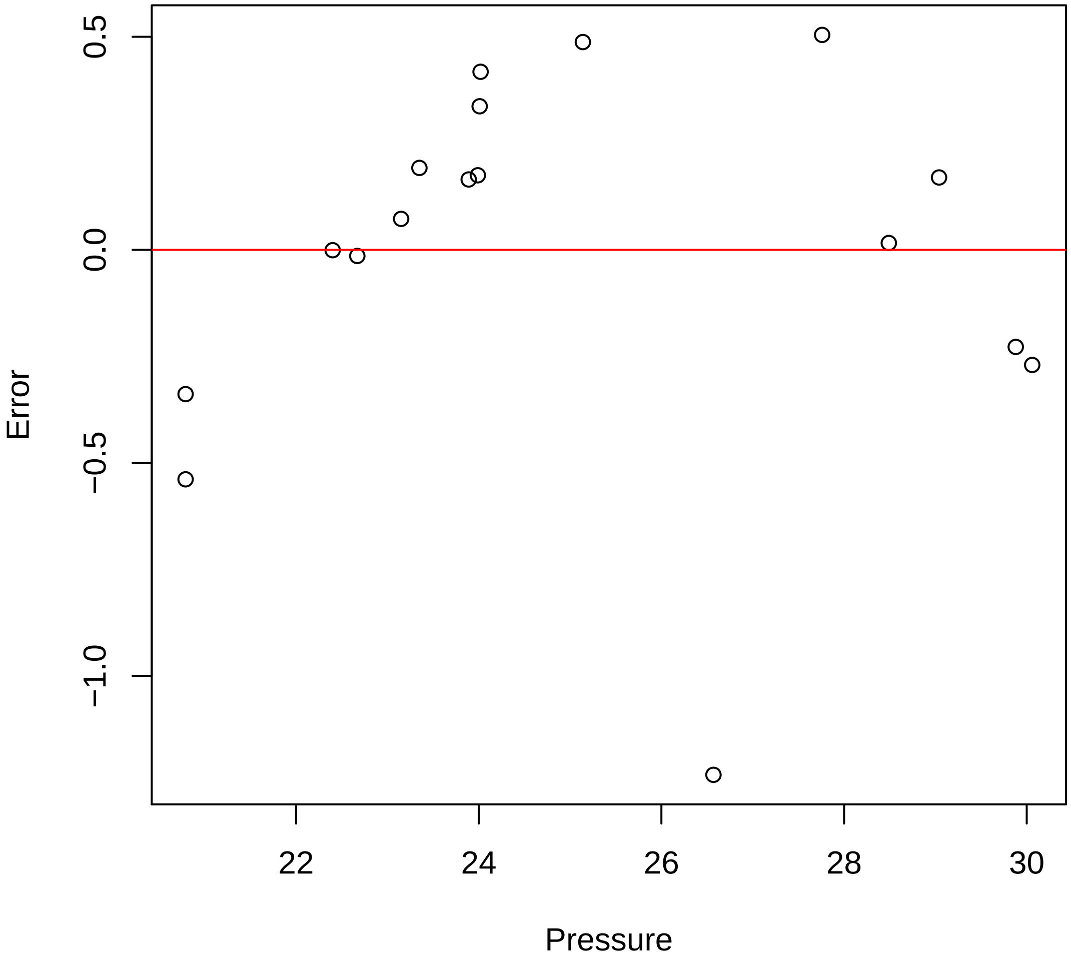


Figure 3: The Temperature over Pressure with Regression Line



seems to be erroneous. Apart from this point, however, the error is not normally distributed. We see that it starts below the line, increases and the decreases again. Thus, the misfit, though at first slight, is actually systematic. We shall return to this problem.

16 Linear Regression II

In this chapter we shall look at the general equation

(343) $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$$

As before, the condition on ε is that it is normally distributed and has mean 0. That is, the density of the distribution of ε is $\frac{1}{\sqrt{2\pi\sigma}} e^{-x^2/s\sigma}$. This allows us to give the equation also another form, namely

(344) $$\mathsf{E}(Y|\vec{X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

This means that the conditional expectation of Y based on the predictors is described by the law above. Taking expectations makes the random term disappear.

In fact, even though it is called linear regression, it is not necessary that X_i equals an observable. Rather, it is allowed to *transform* the explanatory variables in any way. Here is a very simple example. The area A of a rectangle is given by the formula

(345) $$A = \ell \cdot h$$

where ℓ is the length and h the height. Taking logarithms on both sides we get

(346) $$\log A = \log \ell + \log h$$

This is linear law, but it involves the logarithms of length and height. In general, any equation that involves a product of numbers—and in physics there are very many such laws—can be turned into a linear law by taking logarithms.

This leads to the distinction between *predictors* and *terms*. The **predictors** are the variables that enter the equation on the right hand side. The **terms** are the X_i. Terms might among other be the following:

① The **intercept**. Rewrite (347) as

(347) $$Y = \beta_0 X_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$$

where X_0 is a term that always equals 1.

② The **predictors**.

③ **Transformations** of predictors. If X is a predictor, a transformation would be any function thereof, for example $\log X$, X^p for some number p, e^X, $\sin X$, and so on.

④ **Polynomials** of predictors. If X is a predictor, then we have a term of the form $a_0 + a_1X + a_2X^2 + \cdots$.

⑤ **Interactions** of predictors. As seen above, we may have terms that involve several predictors. A pure multiplicative law, however, can be rendered linear by taking the log on both sides.

⑥ **Dummy variables and factors**. If the result depends on a factor rather than a number we can artificially create room for the effect of a factor by introducing a variable that assumes only two values, normally 0 and 1, depending on the presence or absence of the factor. The factors enters into the equation in the form of that variable.

Additionally, we can also transform the response, as we have done above.

This in fact raises two separate questions: the first, which are terms that enter the equation; the second, which are the coefficients, their mean and variance; and third, which predictors should be used. The answer to the last question is deferred to the next section. Here we shall first briefly address the question about coefficients and then talk about transforms.

The determination of the coefficients is basically done through linear algebra. Suppose for simplicity that the terms are the predictors. Next assume thet we have $n + 1$ data points, $\vec{x}_i = \langle x_{i1}, x_{i2}, \cdots, x_{in}\rangle$, as well as the y_i, then we end up with a system of equations of the following form:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in} + \varepsilon \tag{348}$$

This requires linear algebra to solve for the β_i. Additionally, it is possible to estimate not only the β_i but also give estimates of the variance. This allows for the estimation of error. The details of this go beyond the scope of these lectures, however. Luckily, this can be left to the software, in our case R. It is however important to understand what sort of computations R performs.

Significance: the test involves the following hypotheses: the null hypothesis that there is only an intercept, and the alternative that there is a law of the form

(347).

$$\text{(349)}\qquad \begin{aligned}&\text{(NH)}\ \mathsf{E}(Y|\vec{X}) = \beta_0\\ &\text{(AH)}\ \mathsf{E}(Y|\vec{X}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n\end{aligned}$$

R also gives significance values for each individual variable. The is the column `Pr(>|t|)`. Here the alternative is the following.

$$\text{(350)}\qquad \begin{aligned}&\text{(NH)}\ \mathsf{E}(Y|\vec{X}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n \text{ and } \beta_i = 0\\ &\text{(AH)}\ \mathsf{E}(Y|\vec{X}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n\end{aligned}$$

This can lead to the following problem: if the response can be predicted equally well from two predictors, then none of them is significant, since any one of them can be dropped if the other is kept. (A trivial example would be measuring the temperature both in Celsius and in Fahrenheit. This gives two numbers in each case, but each one is as informative as the other.) The tests would predict both predictors to be insignificant, since any of them can be dropped from the ensemble. But the test only reveals what happens if one of them is dropped. After it has been dropped, the other predictors can rise dramatically in significance.

Instead, one can also look at the cumulative sinificance. Here the alternative is the following:

$$\text{(351)}\qquad \begin{aligned}&\text{(NH)}\ \mathsf{E}(Y|\vec{X}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_{i-1} X_{i-1}\\ &\text{(AH)}\ \mathsf{E}(Y|\vec{X}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_i X_i\end{aligned}$$

These numbers are returned in the if `anova` is called in R. The significance is given in the column `Pr(>F)`. This measures whether adding the term to the terms up to number $i - 1$ makes any difference. Obviously, the order in which the terms are presented makes a difference. If X_i is as predictive as X_j, $j > i$, then the cumulative significance of X_j is zero, because all the information has been given already. If one were to interchange them, X_i would be judged insignificant instead.

16.1 General F test

The general theory behind the significance for predictors is as follows. We assume to have two sets of predictors X_1, through X_{p-q} and X_{p-q+1} through X_p. We ask whether after hacing added the first set of predictors the second set is still needed. We formulate a hypothesis:

$$\begin{aligned}&(\mathrm{NH}) Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-q} X_{p-q} + \varepsilon \\ &(\mathrm{AH}) Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon\end{aligned} \tag{352}$$

Doing this, we compute two regression lines and get $\mathrm{RSS}_{\mathrm{NH}}$ with degrees of freedom $\mathrm{df}_{\mathrm{NH}}$ for the null hypothesis, and $\mathrm{RSS}_{\mathrm{AH}}$ with degrees of freedom $\mathrm{df}_{\mathrm{AH}}$ for the alternative hypothesis. Clearly, $\mathrm{RSS}_{\mathrm{NH}} - \mathrm{RSS}_{\mathrm{AH}} \geq 0$. If they are actually equal, the answer is clear: the additional variables are not needed. So, we may actually assume that the difference is not zero. We now compute the value

$$F = \frac{(\mathrm{RSS}_{\mathrm{NH}} - \mathrm{RSS}_{\mathrm{AH}})/(\mathrm{df}_{\mathrm{NH}} - \mathrm{df}_{\mathrm{AH}})}{\mathrm{RSS}_{\mathrm{AH}}/\mathrm{df}_{AH}} \tag{353}$$

This number is $F(\mathrm{df}_{\mathrm{NH}} - \mathrm{df}_{\mathrm{AH}}, \mathrm{df}_{\mathrm{AH}})$ distributed. Hence we can apply an F-test to determine whether or not adding the predictors is needed, and we can do so for an arbitrary group of predictors.

16.2 Lack of Fit

Lack of fit can be diagnosed through the residual. Let us look at a particular case, when the conditional variance of Y is not constant. Then, if the X_i assume certain fixed values, we should under (347) assume that Y has fixed variance σ^2. Thus,

$$\mathsf{V}(Y|\vec{X} = \vec{x}) = \sigma^2 \tag{354}$$

Assume that $\mathsf{V}(Y|\vec{X})$ is however a function of $\mathsf{E}(Y|\vec{X} = \vec{x})$, that is, assume that

$$\mathsf{V}(Y|\vec{X} = \vec{x}) = \sigma^2 g(\mathsf{E}(Y|\vec{X} = \vec{x})) \tag{355}$$

In that case the misfit can often be corrected by transforming the response. There are a number of heuristics that one can apply.

① If $g = ax$, use $\sqrt{Y}$ as a predictor in place of Y.

② If $g = ax^2$ use $\log Y$ in place of Y.

③ If $g = ax^4$ use $1/Y$ in place of Y. This is appropriate if the response is often close to 0, but occasional large values occur.

④ If Y is a proportion between 0 and 1, you may use $\sin^{-1}(\sqrt{Y})$ in place of Y.

Notice that this transformation makes the mean function nonlinear!

17 Linear Regression III: Choosing the Estimators

We have previously talked about information theory in connection with word length. Now we shall engage in a more general discussion of information. The notion of information occurs in connection with two problem areas. The first is the question of how much information a sample contains and the second is the cross-entropy of probability distributions, which is needed to estimate the goodness of fit of approximations.

The Kullbach-Leibler Divergence. Let p and q be two probability functions on a space Ω. Then

$$\text{KL}(p,q) := \mathsf{E}\,\frac{p}{q} = \sum_{\omega\in\Omega} p(\omega)\log_2\left(\frac{p(\omega)}{q(\omega)}\right) \tag{356}$$

In the continuous case:

$$\text{KL}(f,g) := \int_{-\infty}^{\infty} f(x)\log_2\left(\frac{f(x)}{g(x)}\right)dx \tag{357}$$

The **entropy** of a distribution is

$$H(p) := \sum_{omega} -\ln p(\omega)\ln p(\omega) \tag{358}$$

The **cross-entropy** $H(p,q)$ is defined by

$$H(p,q) := \sum_{omega} -\ln p(\omega)\ln q(\omega) \tag{359}$$

Notice that $H(p) = H(p,p)$. The equation (356) now becomes

$$\text{KL}(p,q) = H(p,q) - H(p) \tag{360}$$

The KL really is a distance function. It has the following properties.

Proposition 46

$$\begin{aligned} &KL(p,q) \geq 0 \\ &KL(p,q) = 0 \Leftrightarrow p = q \end{aligned} \tag{361}$$

Proof. We do only the discrete case. Fist, observe that $\log x \leq x - 1$, and that equality only holds when $x = 1$. Now,

$$\begin{aligned} \mathrm{KL}(p, q) &:= \sum_{\omega} p(\omega) \log_2 \left(\frac{p(\omega)}{q(\omega)} \right) \\ &\leq \sum_{\omega} p(\omega) \left(\frac{p(\omega)}{q(\omega)} - 1 \right) \end{aligned} \tag{362}$$

⊣

It is not symmetric, though, and therefore the original definition of [4] actually works with the symmetrified version: $\mathrm{KL}(p, q) + \mathrm{KL}(q, p)$. The latter has the properties above and is additionally also symmetric. Neither of them is a metric, since they do not satisfy the triangle inequality.

Consider now the following question. We have a dependent variable Y and some explanatory variables X_i, $1 \leq m$, and we wish to find the best possible linear model that predicts Y from the X_i. Goodness of fit is measured in terms of the residual sum of squares. If f is the estimator, and we have taken n measurements, each leading to samples $\vec{\omega}_i$, $i \leq n$, for the variables X_i, and to values η_i, $i \leq n$, for the variable Y, then

$$\mathrm{RSS}(f) := \sum_{i=1}^{n} (\eta_i - f(\vec{\omega}))^2 \tag{363}$$

This function measures the distance of the approximating function f from the actual, observed data. We want to find the best possible function f in the sense that the $\mathrm{RSS}(f)$ is minimal. Unless the explanatory variables are really useless, it is always better to add the variable, because it may contribute to improve the residual sum of squares. On the other hand, if we have m measurements and m variables, we can always fit a polynomial of degree $m - 1$ through the data points, and so the residual sum of squares will be 0. Although we are dealing here with linear models, the question that the example raises is a valid one: simply minimising RSS will result in the inclusion of all variables however small their contribution to the function f is. Therefore it is felt that even variables that make no contribution at all in the real distribution will be included because statistically it is likely that they appear to be predictive. To remove this decificiency, it has been proposed to measure not only the RSS but to introduce a punishment for having too many explanatory variables. This leads to the **Akaike Information**

Criterion:

$$\text{(364)} \qquad \text{AIC} = \frac{2k}{n} + \ln\left(\frac{\text{RSS}}{n}\right)$$

In addition to the mean RSS what enters in it is the number *n* of measurements and the number *k* of explanatory variables. Since we want to minimise the AIC, having more variables will make matters worse unless the RSS shrinks accordingly.

The Schwarz-Information Criterion for Gaussian models:

$$\text{(365)} \qquad \text{SIC} = \frac{k}{n}\ln(n) + \ln\left(\frac{\text{RSS}}{n}\right)$$

18 Probabilistic Regular Languages and Hidden Markov Models

Hidden Markov models have gained popularity because of their greater flexibility than ordinary Markov chains. We shall approach Hidden Markov Models through a seemingly different object, namely *probabilistic regular languages*.

Definition 47 *A* ***probabilistic regular grammar*** *is a quintuple $\langle S, N, A, R, P\rangle$, where N and A are disjoint sets, the set of* ***terminal*** *and* ***nonterminal symbols****, respectively, $S \in N$ the* ***start symbol****, $R \subset (A \times N) \cup N \cup \{\varepsilon\}$ a finite set, the set of* ***rules****, and $P : R \to [0, 1]$ a* ***probability assignment*** *such that for all $X \in N$*

(366) $$\sum_{X \to \vec{\alpha}} P(X \to \vec{\alpha}) = 1$$

Although the unary rules $A \to B$ can be dispensed with, we find it useful to have them. A **derivation** is a sequence $\vec{\rho} = \langle \rho_i : 1 \leq i \leq n\rangle$ of rules such that (1) $\rho_1 = A \to \vec{\alpha}$ for some $\vec{\alpha}$, and $A \in N$, (2) for $i > 1$, if $\rho_{i-1} = A \to aB$ for some $A, B \in N$ and $a \in A$ then $\rho_i = B \to bC$ for some $C \in N$, $b \in A$ or $\rho_i = B \to \varepsilon$. The string derived by $\vec{\rho}$ is defined by induction as follows.

(367) $$\sigma(\langle\rangle) := A$$

(368) $$\sigma(\langle \rho_0, \cdots, \rho_i\rangle) := \vec{\alpha}bC, \qquad \text{where}\sigma(\langle \rho_0, \cdots, \rho_{i-1}\rangle) = \vec{\alpha}B$$

The probability assigned to a derivation is

(369) $$P(\vec{\rho}) := \prod_{i=1}^{n} P(\rho_i)$$

For a string $\vec{\alpha}$, put

(370) $$P(\vec{\alpha}) := \sum_{\sigma(\vec{\rho}) = \vec{\alpha}} P(\vec{\rho})$$

This defines a probability distribution on A^*. Here is an example.

(371)

S → aA	1/4
S → bB	3/4
A → bB	1/3
A → b	2/3
B → aA	1/2
B → a	1/2

This grammar generates the language $(\mathtt{ab})^{+}\mathtt{a}? \cup (\mathtt{ba})^{+}\mathtt{b}?$. The string ababa has the following derivation:

(372) $\langle \mathtt{S} \rightarrow \mathtt{aA}, \mathtt{A} \rightarrow \mathtt{bB}, \mathtt{B} \rightarrow \mathtt{aA}, \mathtt{A} \rightarrow \mathtt{bB}, \mathtt{B} \rightarrow \mathtt{a}\rangle$

The associated probability is $1/4 \cdot 1/3 \cdot 1/2 \cdot 1/3 \cdot 1/2 = 1/96$.

Definition 48 *A* ***probabilistic language over*** *A is a discrete probability space over A^*.*

We ask first: under which conditions does the probability distribution define a probabilistic language? Call a nonterminal A **reachable** if the string $\vec{x}A$ is derivable. Call a nonterminal A groundable if there is a derivation $A \vdash_G \vec{y}$, for a terminal string $\vec{y}$.

Theorem 49 *Let G be a reduced probabilistic regular grammar with $N = \{A_i : 1 \leq i \leq m\}$. Then G defines a probabilistic language over A iff all reachable symbols are also groundable.*

Proof. Let H_n be the set of derivations of length n. Then $1 = \sum_{\langle\rho\rangle\in H_1} P(\rho)$. Let C_1 be the set of complete derivations of length 1. Let γ_1 be the probability of those derivations. Then let H_2 be the set of derivations of length 2. We have $\iota_1 = \sum_{\vec{\rho}\in H_2 - C_2} P(\vec{\rho})$. This can be generalized. Let C_n be the set of derivations of length n and let their probabilities sum to γ_n. Then we have

(373) $$1 = \gamma_1 + \gamma_2 + \cdots + \gamma_n + \sum_{r\vec{ho}\in H_n - C_n} P(\vec{\rho})$$

The claim is established once we have shown that

(374) $$1 = \sum_{i=1}^{\infty} \gamma_i$$

Put

(375) $$\pi_n = \sum_{i=1}^{n} \gamma_i$$

Suppose that every reachable symbol is groundable. We shall show that for some m and $c < 1$, $(1 - \pi_{n-m}) \leq c(1 - \pi_n)$. This will show that $\lim_{n\to\infty} \pi_n = 1$. To this end, let $\vec{\rho}$ be a derivation of length n ending in $\vec{x}A$. By assumption A is groundable, so that $A \vdash_G \vec{y}$, for some $\vec{y}$ of length $\ell_{\vec{\rho}}$ and with some probability $p_{\vec{\rho}} > 0$. Put $m := \max\{\ell_{\vec{\rho}} : \vec{\rho} \in H_n - C_n\}$ and $d := \max\{p_{\vec{\rho}} : \vec{\rho} \in H_n - C_n\}$. Then every incomplete derivation of H_n is completed with probability $c := 1-d$. Hence $(1-\pi_{n+m}) \leq c(1-\pi_n)$. This shows the claim. Now suppose that there is a reachable symbol A that is not groundable. Then a derivation $\vec{\rho}$ of the string $\vec{x}A$ cannot be grounded. This means that the sum of the probabilities of all derivations is at most $1 - P(\vec{\rho})$. ⊣

Definition 50 *A probabilistic language is called* ***regular*** *if it is generated by a probabilistic regular grammar.*

We shall explore the relation with Markov processes. The states of the Markov process shall correspond to the letters of the alphabet, so that every sequence that the process traverses is actually a word. First, notice that while derivation of a regular grammar are supposed to terminate, a Markov process goes on forever. This can be remedied as follows. We add a state s to the Markov process and stipulate that the probability of leaving s is zero (so the transition probability of s to s is 1). Such a state is called a **sink**. We assume to have just one sink, and it corresponds to the **blank**. (Notice that the process is not irreducible.) The Markov process is translated into rules as follows. Each state corresponds to a letter, and the state corresponding to the letter a is denoted by a. If $m_{ba} = p$, and $a \neq s$, we add a rule $\rho := N_a \to aN_b$. If $a = s$ then we add the rule $\rho := N_s \to \varepsilon$. Finally, we add a start symbol S. Let P^0 be the initial probability distribution of the letters. Then we add rules $S \to N_a$ with probability $P^1(a)$. We shall establish a one-to-one correspondence with certain random walks and derivations in the grammar.

Let $\vec{x} = \langle x_1, x_2, \cdots, x_n\rangle$ be a random walk. Since the x_i are letters, we may consider it a string over A. Its probability in the Markov process is

$$(376)\qquad P^1(x_1)\sum_{i=1}^{n-1} m_{x_{i+1}x_i}$$

Since trailing s does not change the probabilities, we may just eliminate them. It is not hard to see that $\vec{x}$ has a unique derivation in the grammar, which is

$$(377)\qquad \langle S \to N_{x_1}, N_{x_1} \to x_1N_{x_2}, N_{x_2} \to x_2N_{x_3}, \cdots, N_{x_{n-1}} \to x_{n-1}N_{x_n},$$
$$N_{x_n} \to x_nN_s, N_s \to \varepsilon\rangle$$

The associated probability is the same. It follows that the Markov process assigns the same probability to the string. Now we wish to do the converse: given a regular grammar, obtain a Markov process that generates the strings with identical probability. This turns out be lead to a generalisation. For notice that in general a nonterminal symbol does not develop into a fixed terminal symbol.

Definition 51 *A **Hidden Markov Model (HMM)** is a pair $\langle S, O, M, E\rangle$, such that S is a set, the set of **states**, O the set of **observables**, $M = (m_{b,a})_{a,b\in S}$ a stochastic matrix, and E a function from S to O.*

We shall as usual assume that $S0\{1, 2, \cdots, n\}$. The addition over the Markov process is a map that translates states into observables, though only with a certain probability. Random walks are walks in the set of states. However, one thinks of the random walk as something that is hidden. Instead, one can only observe sequences $\langle o_1, o_2, \cdots, o_z u$, which are obtained by applying the map E. It is not possible to retrieve the random walk from the image under E, since E may conflate several states.

We translate a regular probabilistic grammar into a Hidden Markov model in two steps. First, we apply some reform to our grammar to establish a slightly differennt grammar H. $N_H := (N \times A) \cup N$, $A_H := A$. The start symbol is S. The rules are all rules of the form $A \to \langle A, a\rangle$, and $\langle A, a\rangle \to aB$, where $A \to aB$ is a rule; all rules of the form $A \to B$ if $A \to B$ is a rule of the grammar, and $B \to \varepsilon$ if $B \to \varepsilon$ is a rule of the original grammar. The probabilities are assigned as follows.

(378) $$P_H(A \to \langle A, a\rangle) := \sum_{a\in A, B\in N} P(A \to aB)$$

(379) $$P_H(\langle A, a\rangle \to aB) := P(A \to aB)/P(A \to \langle A, a\rangle)$$

(380) $$P_H(A \to B) := P(A \to B)$$

(381) $$P_H(A \to \varepsilon) := P(A \to \varepsilon)$$

It is easy to see that (complete) derivations in the new grammar are in one-to-one correspondence with derivations in the old grammar. We translate H into a Hidden Markov model. Put $S := N_H$, $O := A \cup \{|\}$, where $| \notin A$. $E(A) := |$ and $E(\langle A, a\rangle) := a$. Finally, the transition probabilities are $m_{\langle A,a\rangle,A} := P_H(A \to \langle A, a\rangle)$, $m_{B,\langle A,a\rangle} := P_H(\langle A, a\rangle \to B)$. As before, we add a sink, and for the rule $A \to \varepsilon$ we put $m_{s,A} := P_H(a \to \varepsilon)$. The Markov process generates random walks of the form $\langle S, \langle S, x_1\rangle, X_1, \langle X_1, x_2\rangle, X_2, \cdots\rangle$, which by applying E are mapped to sequences of the form $x_1|x_2|x_3|\cdots$. They correspond to strings $x_1x_2x_3$ of the known sort. It is easily established that the strings are generated by the Markov process with the appropriate probability. (First it is shown of the derivations, and since derivations are in one-to-one correspondence, this then derives the same result for strings.)

Theorem 52 *A probabilistic language is regular iff it can be generated by a Hidden Markov Model with some initial probability.*

In the literature, it is common to assume a broader definition of a Hidden Markov Model. The most common definition is the following. Instead of a function $E : S \to O$ one simply assumes a relation probability distribution $I : W \to [0,1]$ such that $\sum_{o \in O} I(\langle s, o\rangle) = 1$. This means the following. Given a state s, the output symbol is no longer unique; instead, with some probability $I(\langle s, o\rangle)$ the symbol o will appear. An even more general definition is to make the output symbol contingent on the transition. Thus, there is a probability assignment J from pairs of states and an output state such that $\sum_{o \in O} J(\langle s, s', o\rangle) = 1$ for all $s, s' \in S$. We shall show below that these definitions are in fact not more general. Anything that these models can achieve can be done with a Hidden Markov Model in the sense defined above.

We shall show how to convert a Markov Model of the first kind into an HMM. The reader will surely be able to caryr out a similar reduction of the second one. Let $\mathcal{M} = \langle S, O, M, I\rangle$ be a Markov model. Put $S' := S \times O$, $E(\langle s, o\rangle) := o$. Further,

$$(382) \qquad m'_{\langle s',o\rangle,\langle s,o\rangle} := m_{s',s} \cdot I(\langle s', o'\rangle)$$

We claim that $\mathcal{H} = \langle S', O, M', E\rangle$ is a HMM and that it assigns identical probabilities to all sequences of observables. To that end, we shall do the following. Let $U_s := \{\langle s, o\rangle : o \in O\}$. We claim that

$$(383) \qquad P^k_{\mathcal{H}}(U_s) = P^k_{\mathcal{M}}(s)$$

$$(384) \qquad P^k_{\mathcal{H}}(\langle s, o\rangle) = I(\langle s, o\rangle) P^k_{\mathcal{H}}(U_s)$$

This follows by induction. The first equation is seen as follows.

$$P^{k+1}_{\mathcal{H}}(U_s) = \sum_{o,o'\in O, s'\in S} P^k_{\mathcal{H}}(\langle s, o\rangle) m_{\langle s',o'\rangle,\langle s',o'\rangle} \tag{385}$$

$$= \sum_{o,o'\in O, s'\in S} P^k_{\mathcal{H}}(\langle s', o'\rangle) m_{s,s'} I(\langle s, o\rangle) \tag{386}$$

$$= \sum_{o\in O, s'\in S} P^k_{\mathcal{H}}(U_{s'}) m_{s,s'} I(\langle s, o\rangle) \tag{387}$$

$$= \sum_{s'\in S} P^k_{\mathcal{H}}(U_{s'}) m_{s,s'} \tag{388}$$

$$= \sum_{s'\in S} P^k_{\mathcal{M}}(s') m_{s,s'} \tag{389}$$

$$= P^{k+1}_{\mathcal{M}}(s) \tag{390}$$

The second is shown analogously. Now, the probability that o is emitted in state s is $I(\langle s, o\rangle)$ in $\mathcal{M}$. In $\mathcal{H}$, it is the probability of $P^k_{\mathfrak{H}}(\langle s, o\rangle | U_s) = I(\langle s, o\rangle)$, and the claim is proved.

Now we shall state and prove a much more general result, from which the previous can easily be derived as a corollary. The idea to this theorem is as follows. Suppose we ask about the distribution of a letter at position $k + 1$. In an ordinary Markov model this probability only depends on the probabilities of the letters (including a) at k. We generalize this to allow for any kind of dependency of presvious histories, with the only condition that the histories are classified using regular languages. That is to say, the probability of a with history $\vec{c}$ (a word of length k) at $k + 1$ depends on whether or not $\vec{c}$ is a member of some regular language. This allows to state dependencies that go arbitrarily deep, for example vowel harmony. In what is to follow we write $P^{k+1}(a|\vec{c})$ to mean that the immediate history of a is $\vec{c}$; differently put, if $\vec{c}$ has length k then

$$P^{k+1}(a|\vec{c}) = P(\vec{c}a) \cdot P(\vec{c}) \tag{391}$$

where $P(\vec{c})$ is the probability that the process will generate $\vec{c}$ when initialized.

Theorem 53 *Let $\mathbb{S}$ be a finite partition of A^* into regular languages. Let H be a function from $\mathbb{S} \times A$ to $[0, 1]$ such that $\sum_{a\in A} H(L, a) = 1$. Define probability distributions P^k over A as follows. If $\vec{c}$ is a word of length k and $\vec{c} \in L \in \mathbb{S}$ then $P^k(a|\vec{c}) := H(L, a)$. Then the sequence of probability distributions $\langle P^i : i \in \mathbb{N}\rangle$ can be generated by a HMM.*

Proof. We notice that this definition fixes also the initial distribution; as there is a unique language $L' \in \mathbb{S}$ containing ε, the initial distribution is defined as $P^1(a) := H(L', a)$.

Let $\mathbb{S} = \{L_i : 1 \leq i \leq p\}$. Assume that $\mathfrak{A}_i = \langle Q_i, A, i_i, F_i, \delta_i \rangle$ is a deterministic finite state automaton recognizing L_i. Define $S := Q_1 \times Q_2 \times \cdots \times Q_p \times A$. Put $O := A$ and $E(\langle q, q', \cdots, q^{(p)}, a \rangle) := a$. Finally, the transition matrix is as follows. Let $o := \langle q, q', \cdots, q^{(p)}, a \rangle$ and $o' := \langle r, r', \cdots, r^{(p)}, b \rangle$. Then if for all i: $r^{(i)} = \delta_i(q^{(i)}, a)$, then $m_{o'o} := H(L_j, b)$, where j is the unique j such that $q^{(j)} \in F_j$. (That it is unique follows from the fact that the languages are disjoint.) If $r^{(i)} \neq \delta_i(q^{(i)})$ for some i, then $m_{o'o} := 0$. We show that this defines a stochastic matrix. Let $o'' = \langle s, s', \cdots, s^{(p)}, c \rangle$. If $m_{o''o} \neq 0$ then in fact $s^{(i)} = r^{(i)}$ for all $i \leq p$. Thus

$$(392) \qquad \sum_{o' \in O} m_{o'o} = \sum_{a \in A} H(L, a) = 1$$

Thus we have a HMM. Next we show that $P^{k+1}(a|\vec{c}) = H(L, a)$, where $\vec{c}$ has length k. To this end notice that if $\vec{c}$ is the history of a, we can recover the state as follows: let $q^{(i)}$ be the unique state such that there is a run from i_i to $q^{(i)}$ with the word $\vec{c}$; and let $r^{(i)}$ be the unique state such that there is a run from i_i to $r^{(i)}$ with the word $\vec{c}a$. Then the transition is from $\langle q, q', \cdots, q^{(p)}, c_k \rangle$ to $\langle r, r', \cdots, r^{(p)}, a \rangle$, with probability $H(L_j, a)$ where $L_j \ni \vec{c}$, by construction. This is as it should be. ⊣

Part III

Probabilistic Linguistics

19 Probabilistic Regular Languages and Hidden Markov Models

Hidden Markov models have gained popularity because of their greater flexibility than ordinary Markov chains. We shall approach Hidden Markov Models through a seemingly different object, namely *probabilistic regular languages*.

Definition 54 *A* ***probabilistic regular grammar*** *is a quintuple $\langle S, N, A, R, P\rangle$, where N and A are disjoint sets, the set of* ***terminal*** *and* ***nonterminal symbols****, respectively, $S \in N$ the* ***start symbol****, $R \subset (A \times N) \cup N \cup \{\varepsilon\}$ a finite set, the set of* ***rules****, and $P : R \to [0, 1]$ a* ***probability assignment*** *such that for all $X \in N$*

$$(393) \qquad \sum_{X\to\vec{\alpha}} P(X \to \vec{\alpha}) = 1$$

Although the unary rules $A \to B$ can be dispensed with, we find it useful to have them. A **derivation** is a sequence $\vec{\rho} = \langle \rho_i : 1 \leq i \leq n\rangle$ of rules such that (1) $\rho_1 = A \to \vec{\alpha}$ for some $\vec{\alpha}$, and $A \in N$, (2) for $i > 1$, if $\rho_{i-1} = A \to aB$ for some $A, B \in N$ and $a \in A$ then $\rho_i = B \to bC$ for some $C \in N$, $b \in A$ or $\rho_i = B \to \varepsilon$. The string derived by $\vec{\rho}$ is defined by induction as follows.

$$(394) \qquad \sigma(\langle\rangle) := A$$

$$(395) \qquad \sigma(\langle \rho_0, \cdots, \rho_i\rangle) := \vec{\alpha}bC, \qquad \text{where}\sigma(\langle \rho_0, \cdots, \rho_{i-1}\rangle) = \vec{\alpha}B$$

The probability assigned to a derivation is

$$(396) \qquad P(\vec{\rho}) := \prod_{i=1}^{n} P(\rho_i)$$

For a string $\vec{\alpha}$, put

$$(397) \qquad P(\vec{\alpha}) := \sum_{\sigma(\vec{\rho})=\vec{\alpha}} P(\vec{\rho})$$

This defines a probability distribution on A^*. Here is an example.

$$(398) \qquad \begin{array}{ll} \texttt{S} \to \texttt{aA} & 1/4 \\ \texttt{S} \to \texttt{bB} & 3/4 \\ \texttt{A} \to \texttt{bB} & 1/3 \\ \texttt{A} \to \texttt{b} & 2/3 \\ \texttt{B} \to \texttt{aA} & 1/2 \\ \texttt{B} \to \texttt{a} & 1/2 \end{array}$$

This grammar generates the language $(\mathtt{ab})^{+}\mathtt{a}? \cup (\mathtt{ba})^{+}\mathtt{b}?$. The string ababa has the following derivation:

(399) $\langle \mathtt{S} \to \mathtt{aA}, \mathtt{A} \to \mathtt{bB}, \mathtt{B} \to \mathtt{aA}, \mathtt{A} \to \mathtt{bB}, \mathtt{B} \to \mathtt{a} \rangle$

The associated probability is $1/4 \cdot 1/3 \cdot 1/2 \cdot 1/3 \cdot 1/2 = 1/96$.

Definition 55 *A* ***probabilistic language over*** *A is a discrete probability space over A^*.*

We ask first: under which conditions does the probability distribution define a probabilistic language? Call a nonterminal A **reachable** if the string $\vec{x}A$ is derivable. Call a nonterminal A groundable if there is a derivation $A \vdash_G \vec{y}$, for a terminal string $\vec{y}$.

Theorem 56 *Let G be a reduced probabilistic regular grammar with $N = \{A_i : 1 \leq i \leq m\}$. Then G defines a probabilistic language over A iff all reachable symbols are also groundable.*

Proof. Let H_n be the set of derivations of length n. Then $1 = \sum_{\langle\rho\rangle \in H_1} P(\rho)$. Let C_1 be the set of complete derivations of length 1. Let γ_1 be the probability of those derivations. Then let H_2 be the set of derivations of length 2. We have $\iota_1 = \sum_{\vec{\rho} \in H_2 - C_2} P(\vec{\rho})$. This can be generalized. Let C_n be the set of derivations of length n and let their probabilities sum to γ_n. Then we have

(400) $$1 = \gamma_1 + \gamma_2 + \cdots + \gamma_n + \sum_{r\vec{h}o \in H_n - C_n} P(\vec{\rho})$$

The claim is established once we have shown that

(401) $$1 = \sum_{i=1}^{\infty} \gamma_i$$

Put

(402) $$\pi_n = \sum_{i=1}^{n} \gamma_i$$

Suppose that every reachable symbol is groundable. We shall show that for some m and $c < 1$, $(1 - \pi_{n-m}) \leq c(1 - \pi_n)$. This will show that $\lim_{n\to\infty} \pi_n = 1$. To this end, let $\vec{\rho}$ be a derivation of length n ending in $\vec{x}A$. By assumption A is groundable, so that $A \vdash_G \vec{y}$, for some $\vec{y}$ of length $\ell_{\vec{\rho}}$ and with some probability $p_{\vec{\rho}} > 0$. Put $m := \max\{\ell_{\vec{\rho}} : \vec{\rho} \in H_n - C_n\}$ and $d := \max\{p_{\vec{\rho}} : \vec{\rho} \in H_n - C_n\}$. Then every incomplete derivation of H_n is completed with probability $c := 1 - d$. Hence $(1-\pi_{n+m}) \leq c(1-\pi_n)$. This shows the claim. Now suppose that there is a reachable symbol A that is not groundable. Then a derivation $\vec{\rho}$ of the string $\vec{x}A$ cannot be grounded. This means that the sum of the probabilities of all derivations is at most $1 - P(\vec{\rho})$. ⊣

Definition 57 *A probabilistic language is called* **regular** *if it is generated by a probabilistic regular grammar.*

We shall explore the relation with Markov processes. The states of the Markov process shall correspond to the letters of the alphabet, so that every sequence that the process traverses is actually a word. First, notice that while derivation of a regular grammar are supposed to terminate, a Markov process goes on forever. This can be remedied as follows. We add a state s to the Markov process and stipulate that the probability of leaving s is zero (so the transition probability of s to s is 1). Such a state is called a **sink**. We assume to have just one sink, and it corresponds to the **blank**. (Notice that the process is not irreducible.) The Markov process is translated into rules as follows. Each state corresponds to a letter, and the state corresponding to the letter a is denoted by a. If $m_{ba} = p$, and $a \neq s$, we add a rule $\rho := N_a \to aN_b$. If $a = s$ then we add the rule $\rho := N_s \to \varepsilon$. Finally, we add a start symbol S. Let P^0 be the initial probability distribution of the letters. Then we add rules $S \to N_a$ with probability $P^1(a)$. We shall establish a one-to-one correspondence with certain random walks and derivations in the grammar.

Let $\vec{x} = \langle x_1, x_2, \cdots, x_n\rangle$ be a random walk. Since the x_i are letters, we may consider it a string over A. Its probability in the Markov process is

$$(403) \qquad P^1(x_1)\sum_{i=1}^{n-1} m_{x_{i+1}x_i}$$

Since trailing s does not change the probabilities, we may just eliminate them. It is not hard to see that $\vec{x}$ has a unique derivation in the grammar, which is

$$(404) \qquad \langle S \to N_{x_1}, N_{x_1} \to x_1 N_{x_2}, N_{x_2} \to x_2 N_{x_3}, \cdots, N_{x_{n-1}} \to x_{n-1}N_{x_n}, \\ N_{x_n} \to x_n N_s, N_s \to \varepsilon\rangle$$

The associated probability is the same. It follows that the Markov process assigns the same probability to the string. Now we wish to do the converse: given a regular grammar, obtain a Markov process that generates the strings with identical probability. This turns out be lead to a generalisation. For notice that in general a nonterminal symbol does not develop into a fixed terminal symbol.

Definition 58 *A **Hidden Markov Model (HMM)** is a pair $\langle S, O, M, E\rangle$, such that S is a set, the set of **states**, O the set of **observables**, $M = (m_{b,a})_{a,b\in S}$ a stochastic matrix, and E a function from S to O.*

We shall as usual assume that $S 0\{1, 2, \cdots, n\}$. The addition over the Markov process is a map that translates states into observables, though only with a certain probability. Random walks are walks in the set of states. However, one thinks of the random walk as something that is hidden. Instead, one can only observe sequences $\langle o_1, o_2, \cdots, o_z u$, which are obtained by applying the map E. It is not possible to retrieve the random walk from the image under E, since E may conflate several states.

We translate a regular probabilistic grammar into a Hidden Markov model in two steps. First, we apply some reform to our grammar to establish a slightly differennt grammar H. $N_H := (N \times A) \cup N$, $A_H := A$. The start symbol is S. The rules are all rules of the form $A \to \langle A, a\rangle$, and $\langle A, a\rangle \to aB$, where $A \to aB$ is a rule; all rules of the form $A \to B$ if $A \to B$ is a rule of the grammar, and $B \to \varepsilon$ if $B \to \varepsilon$ is a rule of the original grammar. The probabilities are assigned as follows.

(405) $$P_H(A \to \langle A, a\rangle) := \sum_{a\in A, B\in N} P(A \to aB)$$

(406) $$P_H(\langle A, a\rangle \to aB) := P(A \to aB)/P(A \to \langle A, a\rangle)$$

(407) $$P_H(A \to B) := P(A \to B)$$

(408) $$P_H(A \to \varepsilon) := P(A \to \varepsilon)$$

It is easy to see that (complete) derivations in the new grammar are in one-to-one correspondence with derivations in the old grammar. We translate H into a Hidden Markov model. Put $S := N_H$, $O := A \cup \{|\}$, where $| \notin A$. $E(A) := |$ and $E(\langle A, a\rangle) := a$. Finally, the transition probabilities are $m_{\langle A,a\rangle,A} := P_H(A \to \langle A, a\rangle)$, $m_{B,\langle A,a\rangle} := P_H(\langle A, a\rangle \to B)$. As before, we add a sink, and for the rule $A \to \varepsilon$ we put $m_{s,A} := P_H(a \to \varepsilon)$. The Markov process generates random walks of the form $\langle S, \langle S, x_1\rangle, X_1, \langle X_1, x_2\rangle, X_2, \cdots\rangle$, which by applying E are mapped to sequences of the form $x_1|x_2|x_3|\cdots$. They correspond to strings $x_1x_2x_3$ of the known sort. It is easily established that the strings are generated by the Markov process with the appropriate probability. (First it is shown of the derivations, and since derivations are in one-to-one correspondence, this then derives the same result for strings.)

Theorem 59 *A probabilistic language is regular iff it can be generated by a Hidden Markov Model with some initial probability.*

In the literature, it is common to assume a broader definition of a Hidden Markov Model. The most common definition is the following. Instead of a function $E : S \to O$ one simply assumes a relation probability distribution $I : W \to [0,1]$ such that $\sum_{o\in O} I(\langle s,o\rangle) = 1$. This means the following. Given a state s, the output symbol is no longer unique; instead, with some probability $I(\langle s,o\rangle)$ the symbol o will appear. An even more general definition is to make the output symbol contingent on the transition. Thus, there is a probability assignment J from pairs of states and an output state such that $\sum_{o\in O} J(\langle s,s',o\rangle) = 1$ for all $s,s' \in S$. We shall show below that these definitions are in fact not more general. Anything that these models can achieve can be done with a Hidden Markov Model in the sense defined above.

We shall show how to convert a Markov Model of the first kind into an HMM. The reader will surely be able to caryr out a similar reduction of the second one. Let $\mathcal{M} = \langle S,O,M,I\rangle$ be a Markov model. Put $S' := S \times O$, $E(\langle s,o\rangle) := o$. Further,

$$m'_{\langle s',o\rangle,\langle s,o\rangle} := m_{s',s} \cdot I(\langle s',o'\rangle) \quad (409)$$

We claim that $\mathcal{H} = \langle S',O,M',E\rangle$ is a HMM and that it assigns identical probabilities to all sequences of observables. To that end, we shall do the following. Let $U_s := \{\langle s,o\rangle : o \in O\}$. We claim that

$$P^k_{\mathcal{H}}(U_s) = P^k_{\mathcal{M}}(s) \quad (410)$$

$$P^k_{\mathcal{H}}(\langle s,o\rangle) = I(\langle s,o\rangle)P^k_{\mathcal{H}}(U_s) \quad (411)$$

This follows by induction. The first equation is seen as follows.

$$(412)\qquad P^{k+1}_{\mathcal{H}}(U_s) = \sum_{o,o'\in O,s'\in S} P^k_{\mathcal{H}}(\langle s,o\rangle)m_{\langle s',o'\rangle,\langle s',o'\rangle}$$

$$(413)\qquad = \sum_{o,o'\in O,s'\in S} P^k_{\mathcal{H}}(\langle s',o'\rangle)m_{s,s'}I(\langle s,o\rangle)$$

$$(414)\qquad = \sum_{o\in O,s'\in S} P^k_{\mathcal{H}}(U_{s'})m_{s,s'}I(\langle s,o\rangle)$$

$$(415)\qquad = \sum_{s'\in S} P^k_{\mathcal{H}}(U_{s'})m_{s,s'}$$

$$(416)\qquad = \sum_{s'\in S} P^k_{\mathcal{M}}(s')m_{s,s'}$$

$$(417)\qquad = P^{k+1}_{\mathcal{M}}(s)$$

The second is shown analogously. Now, the probability that o is emitted in state s is $I(\langle s,o\rangle)$ in $\mathcal{M}$. In $\mathcal{H}$, it is the probability of $P^k_{\mathfrak{H}}(\langle s,o\rangle|U_s) = I(\langle s,o\rangle)$, and the claim is proved.

Now we shall state and prove a much more general result, from which the previous can easily be derived as a corollary. The idea to this theorem is as follows. Suppose we ask about the distribution of a letter at position $k+1$. In an ordinary Markov model this probability only depends on the probabilities of the letters (including a) at k. We generalize this to allow for any kind of dependency of presvious histories, with the only condition that the histories are classified using regular languages. That is to say, the probability of a with history $\vec{c}$ (a word of length k) at $k+1$ depends on whether or not $\vec{c}$ is a member of some regular language. This allows to state dependencies that go arbitrarily deep, for example vowel harmony. In what is to follow we write $P^{k+1}(a|\vec{c})$ to mean that the immediate history of a is $\vec{c}$; differently put, if $\vec{c}$ has length k then

$$(418)\qquad P^{k+1}(a|\vec{c}) = P(\vec{c}a)\cdot P(\vec{c})$$

where $P(\vec{c})$ is the probability that the process will generate $\vec{c}$ when initialized.

Theorem 60 *Let $\mathbb{S}$ be a finite partition of A^* into regular languages. Let H be a function from $\mathbb{S}\times A$ to $[0,1]$ such that $\sum_{a\in A} H(L,a) = 1$. Define probability distributions P^k over A as follows. If $\vec{c}$ is a word of length k and $\vec{c}\in L\in\mathbb{S}$ then $P^k(a|\vec{c}) := H(L,a)$. Then the sequence of probability distributions $\langle P^i : i\in\mathbb{N}\rangle$ can be generated by a HMM.*

Proof. We notice that this definition fixes also the initial distribution; as there is a unique language $L' \in \mathbb{S}$ containing ε, the initial distribution is defined as $P^1(a) := H(L', a)$.

Let $\mathbb{S} = \{L_i : 1 \leq i \leq p\}$. Assume that $\mathfrak{A}_i = \langle Q_i, A, i_i, F_i, \delta_i \rangle$ is a deterministic finite state automaton recognizing L_i. Define $S := Q_1 \times Q_2 \times \cdots \times Q_p \times A$. Put $O := A$ and $E(\langle q, q', \cdots, q^{(p)}, a \rangle) := a$. Finally, the transition matrix is as follows. Let $o := \langle q, q', \cdots, q^{(p)}, a \rangle$ and $o' := \langle r, r', \cdots, r^{(p)}, b \rangle$. Then if for all i: $r^{(i)} = \delta_i(q^{(i)}, a)$, then $m_{o'o} := H(L_j, b)$, where j is the unique j such that $q^{(j)} \in F_j$. (That it is unique follows from the fact that the languages are disjoint.) If $r^{(i)} \neq \delta_i(q^{(i)})$ for some i, then $m_{o'o} := 0$. We show that this defines a stochastic matrix. Let $o'' = \langle s, s', \cdots, s^{(p)}, c \rangle$. If $m_{o''o} \neq 0$ then in fact $s^{(i)} = r^{(i)}$ for all $i \leq p$. Thus

$$(419) \qquad \sum_{o' \in O} m_{o'o} = \sum_{a \in A} H(L, a) = 1$$

Thus we have a HMM. Next we show that $P^{k+1}(a|\vec{c}) = H(L, a)$, where $\vec{c}$ has length k. To this end notice that if $\vec{c}$ is the history of a, we can recover the state as follows: let $q^{(i)}$ be the unique state such that there is a run from i_i to $q^{(i)}$ with the word $\vec{c}$; and let $r^{(i)}$ be the unique state such that there is a run from i_i to $r^{(i)}$ with the word $\vec{c}a$. Then the transition is from $\langle q, q', \cdots, q^{(p)}, c_k \rangle$ to $\langle r, r', \cdots, r^{(p)}, a \rangle$, with probability $H(L_j, a)$ where $L_j \ni \vec{c}$, by construction. This is as it should be. $\dashv$

Bibliography

[1] Peter Dalgaard. *Introductory Statistics with R*. Statistics and Computing. Springer, Berlin, Heidelberg, 2002.

[2] James L. Johnson. *Probability and Statistics for Computer Science*. Wiley, Hoboken, New Jersey, 2003.

[3] Keith Johnson. *Quantitative Methods in Linguistics*. Blackwell, Oxford, 2005. to appear.

[4] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79 – 86, 1951.

[5] A. N. Shiryayev. *Probability Theory*. Springer, Berlin, Heidelberg, 1984.

[6] Sanford Weisberg. *Applied Linear Regression*. Wiley Series in Probability and Statistics. Wiley and Sons, Hoboken, 2005.