Mario Cigada

ORIGANOVA

An introduction to Statistics with origami
Version 2.2
With examples using R

English translation by Tim Vaughan

ORIGANOVA

Contents

Chapter 0
An introduction[2] to statistics with origami
ORIGANOVA

In these few pages I would like to tell you something about numbers and statistics, and I would like to do this playing with paper.

Origami is the Japanese word that describes the activity of folding paper. ANOVA stands for ANalysis Of VAriance which is an important, sophisticated tool used in statistical analyses. From these two words comes the unusual title, Origanova, that explains the slightly crazy idea of using origami to explain some important concepts in statistics such as mean, variance and inference.

No particular specialist knowledge is required to use this book, however, it is essential to have to hand a few sheets of A4 paper and a few sheets of paper that are about 10 cm square. Sheets of paper of A4 size are those commonly used in photocopiers in Europe (210 mm x 297 mm; 8.27 inches x 11.69 inches; weighing 80 g /square metre). If you are in the US you can find A4 sheets in specialist shops, or you can cut them from "legal" format paper (216 mm x 356 mm; 8 ½ inches x 14 inches). In fact, for some folds also "letter" format paper (216 mm x 279 mm; 8 ½ inches x 11inches) is fine. The square sheets of paper can be found in toy shops or in stationery shops and are described as origami paper. Otherwise, the coloured blocks of paper that are used for making notes are fine to use, as long as they are exactly square. If you have a choice, pick paper that is a bit thicker than photocopier paper. You will also need a ruler or a set square, a pencil, a rubber, and either scissors or a utility knife.
In this version of Origanova, in addition to a new layout and numerous corrections, I decided to use the US and English notation for numbers; using a dot rather than a comma for decimals.

However, I decided to keep the measurements in millimetres and grams. This version also contains a couple of additional chapters (to be precise one chapter and a pair of appendices) and designs that are a little more beautiful. I hope that you will find it easier to make the folds.

Speaking to some readers of the previous version, I realized that one "flaw" of the book is that it is full of concepts. I do not think I can eliminate this problem, not least because it reflects my way of working. I prefer to have only a few pages to study, even if this forces me to go slowly in my reading and to retrace my steps occasionally to remind myself of something, or to find links between the concepts. I hope that this style is not too annoying.

Another novelty is that I decided to include examples of calculations using a "real" statistics programme, called R. This is a free, but very very powerful and versatile programme, that goes far beyond the needs of this book. Consequently I do not intend to explain exhaustively how to use R; for our purposes we will use R simply as a "super calculator".

When you start the program a window like the one below, called console, will appear on the screen. The screen may be slightly different depending on your operating system.

```
R version 3.0.2 (2013-09-25) -- "Frisbee Sailing"
Copyright (C) 2013 The R Foundation for Statistical
Computing
Platform: i386-w64-mingw32/i386 (32-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

>
```

Next to the cursor there is a "prompt" represented by the character ">". If you want to use R in conjunction with this book you simply need to copy the words written in `Courier` font in the book and put them next to the prompt in R and then press Enter. For example if you want to calculate 1 plus 1 write 1 + 1 next to the prompt > in this way:

```
>1+1
```

Pressing enter you get the following:

```
[1] 2
>
```

That is to say, the result. The result is preceded by [1] because R notifies you that, in this case, the result consists of a single number. After this R goes back to the beginning and shows the " prompt" as if it is saying: "I am ready for another question".
I forgot to mention; you can find R and additional packages on CRAN (Comprehensive R Archive Network), which can be found at the following internet address:

http://cran.r-project.org/

where there is also a lot of information, as well as manuals and examples.
My laziness makes it easy for me to accommodate people who are similar to me. So I have prepared a file with all the functions ready to use. R calls this type of file "script" . You can open the origanova.R script file using the Open Script command which can be found in the R Files menu. Then highlight the function you want to use and copy it with a " copy and paste" command into the console window.

In R for Windows you run the highlighted part of the script simply by pressing Ctrl + r. I do not know what command is required with R for Mac or Linux; there will be an equivalent command but you will have to find it out for yourself. In fact it is easier to do than to explain. Try it yourself; you won't break anything.

In the script , the character `#` defines comments : everything that follows this character until the end of the line will not be executed.

One last point: in the script before a chart, you will find the command:

```
win.graph()
```

that puts the graph in a new window. This enables you to save the previous graphs and to compare them. For the sake of simplicity, this command has been omitted from the text of this book.

If you think that sooner or later you might have to use statistics, the effort involved in learning to use a statistics software will be much rewarded. In this case, it will probably be very handy to know that the character `?` is used to get help with R. So, for example, if you write

```
?mean
```

in the console, R will tell you everything about the function `mean()`. In this book we will use many functions, but we will not stop to analyse all the possible options offered by each of them . Therefore it might be useful to use the help function to find out what else they have to offer.
If none of this interests you, don't worry; you can safely ignore the R commands and just look at the results.

So let's start!

Chapter 1
A paper computer

Statistic is a practical tool , created to manipulate numbers for practical purposes. But we can also use statistics to play, by creating an imaginary situation, rather like a fairy tale.

Once upon a time there was a man who made sweets. Having prepared sweets of many different colours he put the sweets into a machine that put them into bags. The packaging machine was a little old and rather imprecise. Whereas sometimes the bags were completely full, at other times they were half-empty and the children who received those bags complained. In order to understand what was happening with his packaging machine the man took all the bags of sweets that were in stock and weighed them one by one on the scales. The first weighed 2 kg, the second 3 kg, and so on. Here are the weights of all the bags.

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

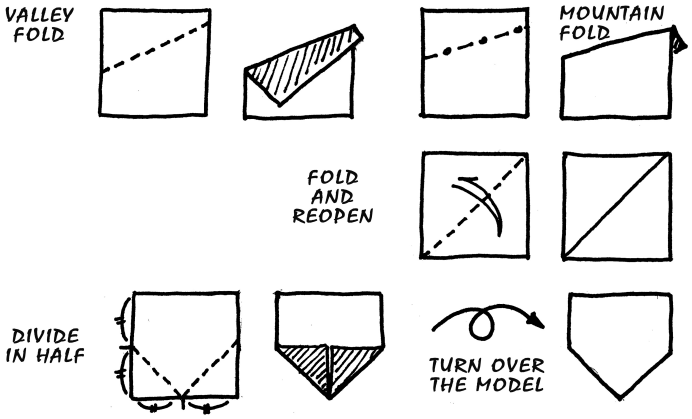
Do you think there too many bags? Well then, let's take just the first 5 bags:

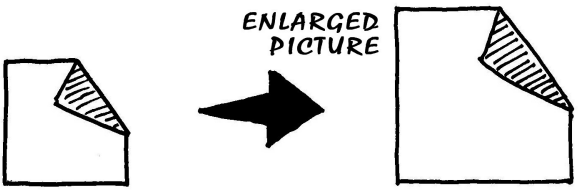
2 3 3 5 2

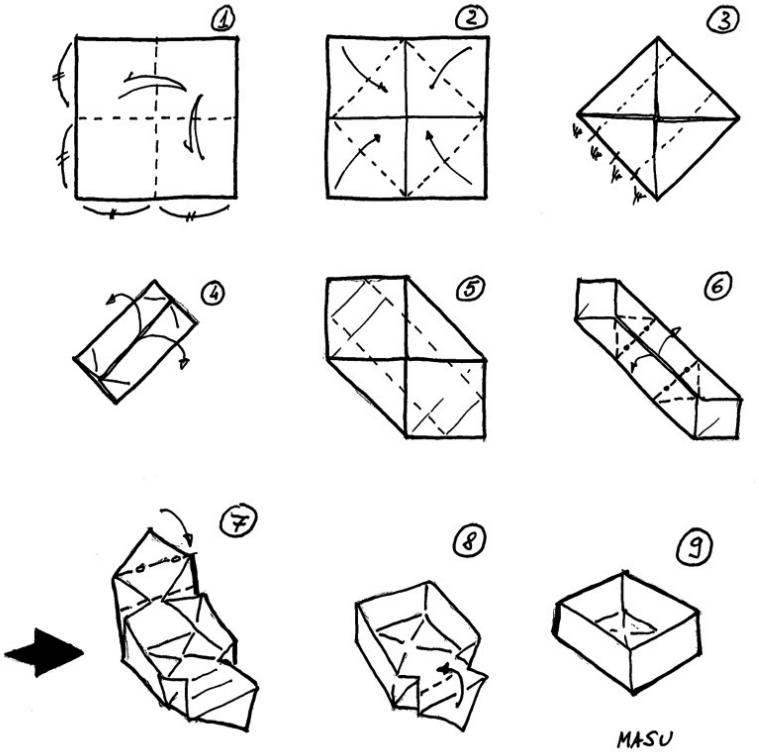
We could pretend that Gervase had a very small stock of sweets. I told you that the manufacturer of the sweets was called Gervase, didn't I? I didn't? Well, I have told you now.

To represent a kilogram I have decided to use a classic fold in traditional origami: the masu. Originally, the masu was a container that was used as a unit of measurement in Japan. So let's begin to make some little masu. Maybe you can get a few friends to help.

On the following pages you will find an explanation of how to make a masu. The drawings will become clearer if you bear in mind that all over the world standard symbols are used to describe origami folds. In addition, let me remind you to be accurate in making the folds and, after folding the paper, to go along the creases with the back of your finger nail.
Here are the most common basic symbols and folds.







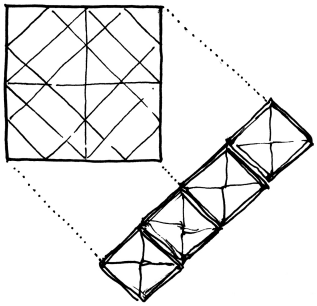
The masu can be used as a container or, turned upside down, can be used like a building block to make things. It is interesting to note that the length of each side of the masu is equal to the length of the sheet of paper we started with, multiplied by the square root of 2, and then divided by 4.

This will be evident if you re-open a masu and look carefully at the creases, especially if you remember that $\sqrt{2}$ is the length of the diagonal of a square of length 1.

So, if you have used a sheet of paper that is 10 cm by 10 cm of the masu will be: $10 \times \sqrt{2} \approx 14.1$ and $14.1 \div 4 \approx 3.5$

i.e. approximately 3.5 cm by 3.5 cm.

Of course, if you want to know the exact value we can get our friend R to help, by typing:



```
> 10*sqrt(2)/4
[1] 3.535534
```

As already mentioned, what is written after the symbol " > " is typed into the "console" R. The result is shown in the line below, and is preceded by [1] because, in this case, the result is composed of a single number.

Returning to our game with paper; how many masu do we have to make? Let's see…

2 to represent the first bag of sweets;
plus 3 for the second bag of sweets;
plus 3 for the third bag of sweets;
plus 5 for ...

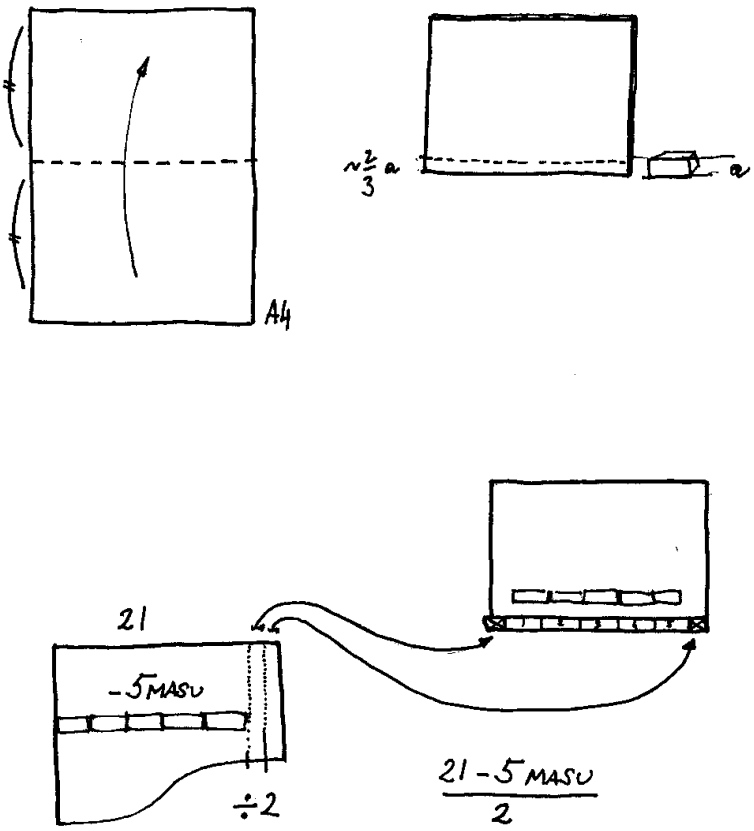
Are you already fed up? Then I'll show you a trick. Let's make only 5 masu and put them in a row so we can measure how long they are. The length is a bit more than 17.5 cm:

3.5 x 5 = 17.5

This is because earlier we had rounded the numbers a little, and because the folds take up some space. But, it does not matter; the important thing is that the total length is less than the short side of an A4 sheet of paper (i.e. less than 21 cm). For this reason the sheets of paper used to make the masu should be about 10 cm square.

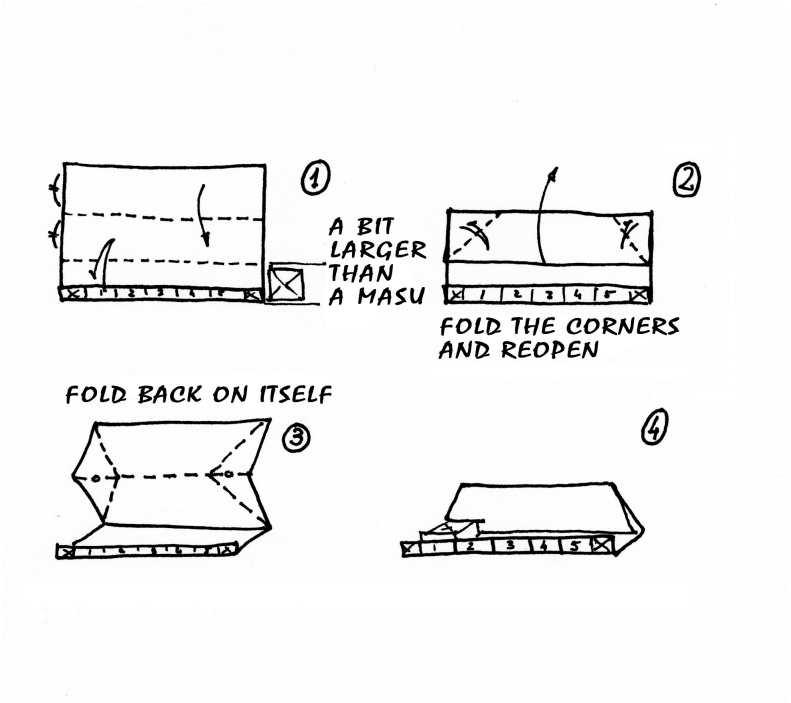
Now we take a sheet of A4 paper (or letter format paper) and fold it in half nd then we make a fold that is a bit less than the height of one of the masu like this:.

Now I need someone to do a calculation. The A4 sheet of paper, as we have said, should have a short side of about 21cm (for letter



paper it is 22 cm) . Subtract from this length the length of the 5 masu. Divide the result by 2 and mark this distance on each edge of the folded paper. Then divide the middle section into 5 equal parts and number them from 1 to 5, as though you are making a ruler.
We can say that a masu placed in position 1 is worth 1 kg, a masu placed in position 2 is worth 2 kg and so on. This is not an unusual thing; in basic arithmetic we make use of positional numbers (based on position). For example, with the number 371, a 3 in the

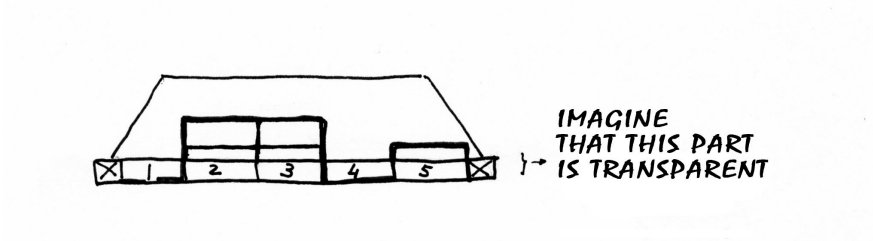
position of the hundreds is worth three hundred, a 7 in the position of the tens is worth seventy and a single unit is worth 1.
Let's go back to what we were building. In order to hold the masu in position and to give the structure more strength it is better to make the following folds.



With the 5 masu we can represent all the stock of Gervase by arranging them in the following way :
• 2 masu for the two bags of 2 kg;
• 2 masu for the two bags of 3 kg; and
• 1 masu for the one bag of 5 kg.

This method of representing the data is called a histogram. Using the masu you have built, you can enjoy yourself representing other sets of numbers. A histogram can be made with masu, or you can

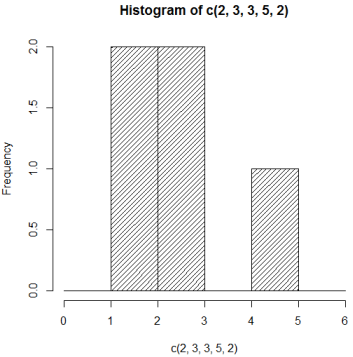
draw it on paper; but usually it is easier to get a computer to draw it for you.



For example, with our friend R you just have to type:

```
>hist(c(2,3,3,5,2),breaks=0:6,density=20)
```

Let me say a quick word about the seemingly mysterious c, which appears in the expression above . It is part of the command `c()` that asks R to link everything that is in brackets (separated by a comma) in order to create a single object; an object that mathematicians call a vector. We will talk about the concept of vectors in Chapter 9.



Please ignore, at the moment, the meaning of the options `breaks` e `density`; if you are curious I remember you that you can ask it to R just typing

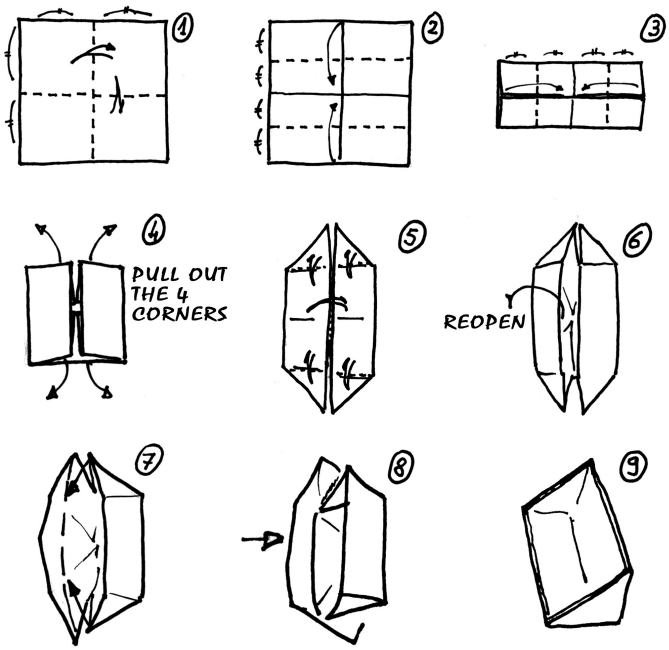
```
?hist
```

What I would like to tell you is that a histogram has a number of very interesting features. First, you will have already noticed that in order to save time in our small example the 5 masu represent 15. Indeed, in a bigger problem we could use a masu to represent a

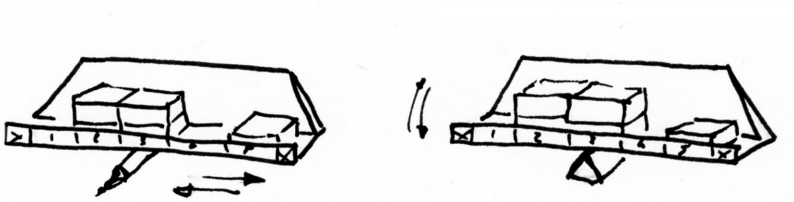
hundred sacks of flour, a thousand camels, and much more. In this way the data are summarized, so you can see at a glance how they are organized.

Now try to find the point of equilibrium of the histogram that represents the sweet stock of Gervase. It can be done in two ways. First, you can put a round pencil under the structure and roll it to the left and to the right until you find the point at which the histogram balances.

Alternatively, we can do this fold (which also serves as the roof of a masu house if you want to play at building things) and use it as a fulcrum. The fold come from a beautiful book called " Origami Omnibus ", written by Kuniko Kasahara (see reference [1])



Have you found the point of equilibrium? Also here it is not easy; we have to be satisfied with an approximate solution, but that's okay. It seems to me that the point of equilibrium is the number 3 on the scale that we wrote on the paper. This is the average of the weights of the bags of sweets.
Maybe some of you already knew what the average was; probably



you had been taught to calculate it by adding together the weights of the bags and dividing by the number of bags, like this:

$(2+3+3+5+2)\div 5=3$

It is no coincidence that it is the same number; because the average is really the centre of gravity of the histogram.
Think back to the diagram on page 13, and let's write it in a more orderly way:

Weight (kg)	*Number of bags*	*Welght x Number*
1	0	0
2	2	4
3	2	6
4	0	0
5	1	5
Total	5	15

So, in more general terms, the arithmetic mean is calculated by dividing the number at the bottom of the third column by the number at the bottom of the second column:

$$15 \div 5 = 3$$

In other words, it multiplies the value of each observation by the number of times it occurs. All of these products are added together and then divided by the number of observations. This is equivalent to:

$$(2+3+3+5+2) \div 5 = 3$$

But be careful not to get confused. Sometimes the number of times that an observation occurs is called its *weight*. In our example, it just so happens that the observations represent weights (physical weights, being the bags of sweets), and are multiplied by the *weights* (mathematical weights, being the number of occurrences).

Returning to our folded paper, perhaps you noticed that we have built a machine that calculates averages! A paper computer that calculates averages and works without batteries! Just put in the masu, make a histogram, find the point of equilibrium, and then use the scale to read off the average.

What is that you said? It only works with numbers ranging from 0 to 5. Well, all computers have their limits. My computer (which cost a lot of money and consumes a lot of energy) is not able to calculate the difference between 10 to the power 308 and 10 to the power 308 minus 1. We can also ask R the same question. To ask if two things are equal the symbol " = " should be repeated twice otherwise R thinks that we want to assign what is to the left of the symbol " = " to the result of calculation to the right of " = "

```
> (10^308)==(10^308-1)
[1] TRUE
```

So, also R is unable to calculate the difference between 10 to the power 308 and 10 to the power 308 minus 1.

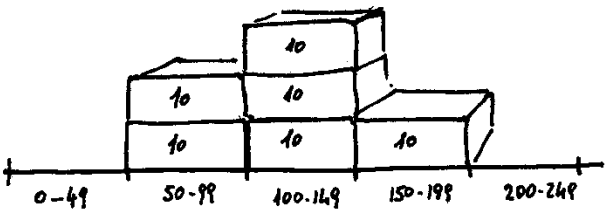
Anyway, to handle larger numbers of masu we just need to build smaller masu. Alternatively, we can use a larger sheet of paper (or use the A4 sheet of paper folded the other way round).

But in fact our machine seems to have another limitation; it only works with whole numbers. This is an interesting observation; it is true that in theory all that is necessary is to build smaller masu. But just think; if Gervase's scales had weighed the bags in grams rather than kilograms we would have had to make masu that were one-thousandth of the size of those that we made, and I can guarantee that making a masu with a sheet of paper that is one-tenth of a millimetre wide is rather difficult.

However, I remind you that it is just coincidence that in our example one masu equalled 1 kg. Nothing prevents us from representing the following set of numbers in a histogram:

138 113 134 195 87 70 75 195 91 116 145 126 174 149 131 83 53 138 173 163 104 129 121 51 144 50 72 76 194 137 112 136 96 146 142 131 135 132 113 132 69 102 76 137 167 83 60 103 118 120 52 69 149 56 52 161 83 158 153 136

in this way:



It is easier to understand how to do this if we start by putting the numbers in ascending order. It is not essential, it just makes it easier to demonstrate the idea.

50 51 52 52 53 56 60 69 69 70 72 75 76 76 83 83 83 87 91 96
102 103 104 112 113 113 116 118 120 121 126 129 131 131 132
132 134 135 136 136 137 137 138 138 142 144 145 146 149 149
153 158 161 163 167 173 174 194 195 195

If we subtract the smallest number from the largest number we get the range.

195 – 50 = 145

Now we have to decide how many groups to divide the ranges into. There are several rules of thumb, such as the table below:

less than 30 observations - the histogram needs just a few groups
less than 100 observations – a maximum of 8 groups
from 101 to 250 observations – a maximum of 10 groups
from 251 to 1000 observations – a maximum of 12 groups

In the example I have decided to use four groups (one of which is empty). Please note that the limits of the groups have to be chosen so as to leave no ambiguity in the assignment of the observations to the groups. I have also decided to use one masu for every 10 observations in the group. Okay, I cheated; the number of observations for each group is exactly divisible by 10 in order to avoid cutting a masu in half. However, remember that I could always decide that a masu is worth 3, 13, or some other number of observations.

Also in the appendix there is a reference to a cube model , which is the height of 2 masu and made with two sheets of paper (it weighs

the same as 2 masu). Combining cubes and masu you can make the system even more flexible.

For example, this is how we could work with the 30 decimal numbers shown below:

```
> x<- c( 4.37348,19.2912,20.7221,12.1345,14.8025,
+ 22.2741, 12.2369,15.5669,18.0976, 17.748,
+  14.5603,13.6388, 8.89321, 12.5463,15.3694,
+  17.5275,13.7957,19.9823,14.256,15.1928,11.2705,
+  20.9492,14.1695,23.5501,14.5677,15.3169,10.7054,
+  14.4311, 16.1116,13.4388)

> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.373  13.490  14.690  15.250  17.690  23.550
```

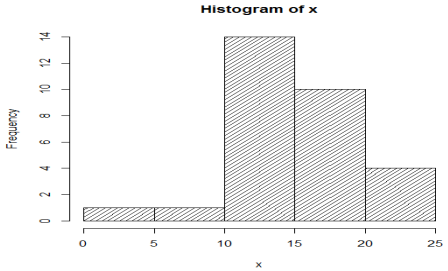
Highest value = 4.37348 Lowest value = 23.5501 Range = 19,1766

They could be represented in a masu histogram using these six groups:

1	From 4.0 to 7.5	1 observation	1 masu
2	From 7.51 to 11.0	2 observations	1 cube
3	From 11.01 to 14.5	10 observations	5 cubes
4	From 14.51 to 18.0	10 observations	5 cubes
5	From 18.01 to 21.5	5 observations	2 cubes + 1 masu
6	From 21.51 to 23.55	2 observations	1 cube

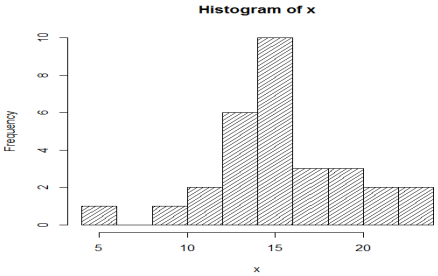
Or we can get R to draw a histogram:

```
hist(x,density=20)
```



Now if we change the class num- ber, the histogram becomes like this:

```
> hist(x,nclass=10,density=20)
```

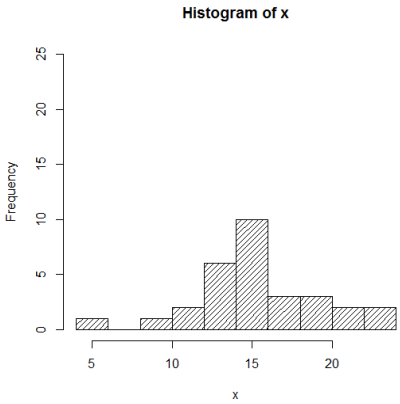


But it's different! Yes and if we force R to plot the Y scale from 0 to 25 , with the option

```
ylim=...
```

it become like like this:

```
hist(x,nclass=10,ylim=c(0,25),density=20)
```



I would ask you to take a few minutes to compare the shapes of the graphs; they look different yet they represent the same set of data.

At this point I hope to make you understand that it is very important when making a histogram to decide carefully about the number of groups and the scales of the axes. This is a very important rule, not only when you make a graph , but also when you look at a chart done by others; pay attention to the scales.

Chapter 2
Measuring dispersion

Going back to our friend, Gervase. He discovered that his packaging machine was producing bags of sweets that weighed an average of 3 kg each. At this point, I hear the familiar comment: "Statistics just tell us lies; if one man eats a whole chicken while another man eats nothing, in terms of statistics they have eaten half a chicken each". This comment, in addition to being old (it is attributed to Trilussa, 1871- 1950), is also wrong. The mean is, in fact, half a chicken each, but statistics is not just about the mean. Returning to the example of Gervase's stock; the story tells us that after a while the bags of sweets were sold, and that it was children who came to buy them. It is true that the bags weighed an average of 3 kg, but go and tell that to the two children who had the misfortune to get bags of sweets weighing only 2 kg. I guarantee that they were pretty disappointed, especially when the child who happened to get the 5 kg bag, a chubby child who later had many problems with his teeth, began making fun of them. In any event Gervase was very upset by the situation.

So we need to invent a way to calculate how the weights of the bags are *dispersed* in relation to the average weight. To do this we could calculate how much each value deviates from the mean. Of course, we can ask R to help us.

```
> (d<-c(3,3,3,3,3)-c(2,3,3,5,2))
[1]  1  0  0 -2  1
```

One of the convenient things about computers is that you can let them do all the repetitive tasks. But here I exaggerated and I did a few things all together in a single line. As I have already mentioned the function `c()` asks R to link everything that is in brackets in order to create a single object called a vector. Then I asked R to do the subtraction with the usual symbol " - " between the two vectors.

3 3 3 3 3
and
2 3 3 5 2

Then I put the result in a variable that I called "d" using the operator " arrow " which I got by using the symbols " < " and " - ". Finally, I kindly asked R to display the result by enclosing the entire expression in brackets.

But, in fact, I could have written it in a simpler way, as follows:

```
> (d<-3-c(2,3,3,5,2))
[1]  1  0  0 -2  1
```

In this case R "understands " that if I ask him to subtract the vector 2 3 3 5 3 2 from single number, it means that I expect him to do 5 subtractions, as in the previous example.

That is:

3 – 2 = 1
3 – 3 = 0
3 – 3 = 0
3 – 5 = - 2
3 – 2 = 1

Now we calculate the average of these results. To do this I have stored the result of the 5 subtractions in the variable d.

```
> mean(d)
[1] 0
```

Well, look at that! It's zero!
It is always zero whatever numbers you choose.

If you think about it for a moment it is obvious. The numbers are a bit larger and a bit smaller than the average, in a way that is exactly

balanced. You will remember that the average is simply the centre of gravity.

Now a mathematical formula that always gives a result of zero is of little use. So, to get something more interesting we could square the differences: a square is never negative (except for the imaginary number *i,* the square of which is -1).

Here it is:

```
> sum(d^2)/4
[1] 1.5
```

You made a mistake! someone will say; you divided by 4 when you should have divided by 5, because there were 5 bags. It was not a mistake, we need to divide by the number of observations minus 1, and to complicate life even more the result of this calculation (n - 1) is given the grand name *degrees of freedom*. To find out why you have to be patient for a few pages; I'll explain it in chapter 9. For the moment just trust me. The number that we calculated is called the *variance*.

The sum of the squared deviations from the mean is also called the *deviance*. So variance = deviation / degrees of freedom (I have set out all the formulas using normal mathematical notation in Appendix I).

So Gervase discovered that his machine packed bags that weighed an average of 3 kg, with a variance of 1.5 kg^2 . Kilograms squared? Yes, having squared the differences we now find ourselves with the kilograms squared. But what is a kilogram squared? Do sweets squared taste better? (Don't get me confused: sweets squared are not necessarily square sweets). However, we simply do not know if sweets squared are better. But to make our life easier we can take

the square root of the variance and thus find a measure of dispersion with the same units as we started with.

```
> sqrt(sum(d^2)/4)
[1] 1.224745
```

To calculate the square root with R using the `sqrt()` function. You will certainly have already noticed that R uses the symbol " `^` " to calculate powers.

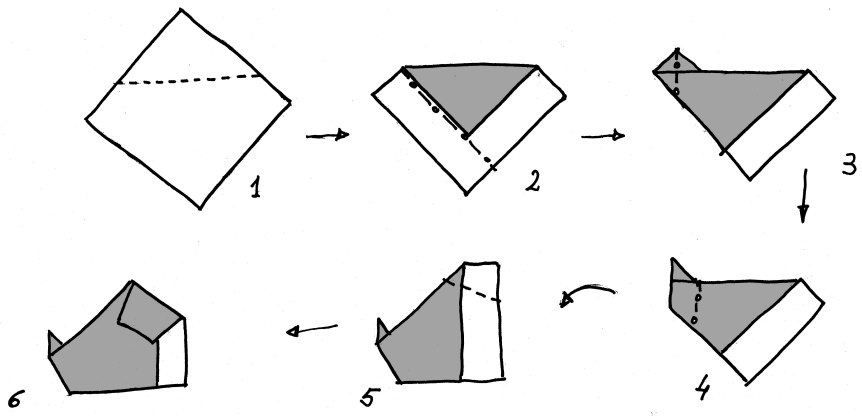
What we have calculated is called *standard deviation*, and, of course, can be calculated directly with R by using the function `sd()`.

```
> sd(c(2,3,3,5,2))
[1] 1.224745
```

Gervase says: "Okay the bags weigh an average of 3 kg, but the dispersion is high; the standard deviation is 1.2 kg, nearly half of the average weight. The packaging machine is completely worn out!

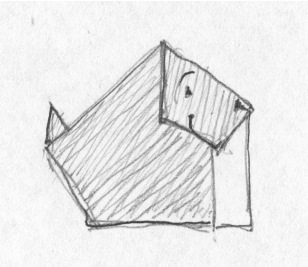
Chapter 3
Measures of position, dispersion and association

The average is a measure of *position* because it tells us where the histogram is positioned, while the standard deviation is a measure of *dispersion*. There are many other measures of position and dispersion; to get to know some of them let's go back to playing with paper. In the bibliography [5] there is a book by Nick Robinson from which I took the fold for this little dog.

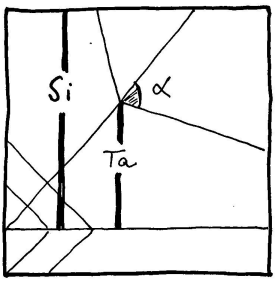


Now please watch carefully; this fold is very elegant in its simplicity. It is better if you use origami paper which is coloured on one side, starting with the coloured side down. If you want you can add some details with a pen, as in the drawing.
This fold has a special characteristic: the first and the last steps do not have any precise reference points, instead the decision about where to make the folds is left to the aesthetic sense of the person who is doing the folding. This is not uncommon in origami; art cannot be bound by rules that are too rigid. On the contrary, maybe the essence of aesthetics lies in finding the right balance between freedom and constraints.

But leave aside the philosophy and return to our little dog. Try to make a few copies, say about fifteen, varying the first and the last fold. It is fun to see how it changes the end result. As it is a simple fold it should not take you too long to create a few.

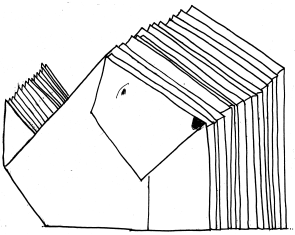


Maybe you have already noticed that there are three variables (four if we take into account the size of the paper). I have shown them in the figure below with the letters Si (size) , Ta (height in italian is “taglia”) and α (alpha). Reopen a puppy and look at the folds: Si depends on where you made the first fold while Ta and α describe where and how you bent the head of the dog in step 5. (According to Nick Robinson if α exceeds 90° the dog becomes a mammoth).



It is fascinating to think that three numbers can fully describe the "biometrics" of our little dog. It is as if we were creating a new breed of dog, and we had the good fortune to be able to model the whole anatomy with just three numbers. This brings to mind many more games and experiments, but I don’t want to digress. Now I would like to focus on one of the three variables: the one in the drawing which is named Ta (the height of the dog) . It is as if we had "caught " a dozen examples of our new breed (Canis Origamicus) and now we want to study them or describe them based on size alone.

If you stand your dogs next to each other, it should not be too difficult to sort them by height, as shown in the diagram.

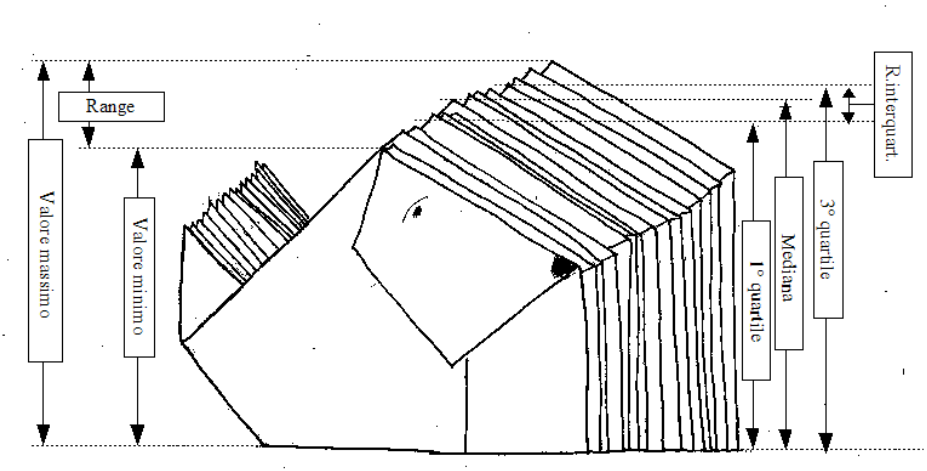


Now it is easy to identify the dog in the middle (number 8), the one for which 7 puppies are smaller and 7 puppies are larger. The size of this little dog is the *median* of our sample. Of course someone will protest at this point: I did 14 dogs and there is not a little dog "in the middle". It is true when I said "about fifteen" I did not mean an exact number of examples. However, it isn't a problem if there are an even number of dogs; just measure the 2 dogs "in the middle" (in the case of 14 examples use the seventh and the eighth), add the values and divide by 2. In other words, the median of our sample in this case is the average of the two middle values.

The median is a nice way to describe the position of our collection of observations without doing lots of calculations. Yes, you guessed it, the median is another index of *position*, just like the mean. Moreover, if we had wanted to calculate the mean height of our sample, we would have had to have measured all the dogs, while the median for a sample, no matter what the size, requires just one or two measurements.
Remember , however, that measuring only the position of a sample exposes us to the risk of a blunder (if one man eats a chicken, while another man eats nothing ...)

Let's go on. The median divides a sample into two groups: one half " small" and the other half "large". But nothing prevents us from taking each of the two groups and repeating the process by dividing each half into two quarters. The size of the dog that separates the two halves into quarters is called the *quartile*. The *first quartile*

separates the quarter of the dogs that are “very small" from the three quarters of the dogs that are big. The *third quartile* separates the quarter of the dogs that are large from the three quarters that are small. And the *second quartile*? It is just another way (not often used) to refer to the median.



If we calculate the difference between the 3rd and 1st quartile we get the *interquartile range*. The *range* , as we had already seen in chapter 1, is the difference between the tallest dog and the shortest dog. Range and interquartile range are two other measures of *dispersion*, like standard deviation.

In some cases , especially when the sample is very large, it is better to divide the sample into 100 parts rather than 4 parts. The values that mark these parts are called *percentiles*. I was thinking that to illustrate this concept it would be necessary to fold 200 or 300 dogs, to arrange them by height and to measure the size of those that ... .

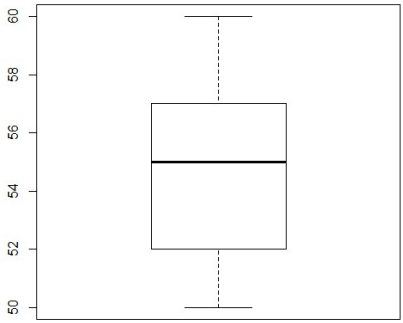
But, perhaps you can just imagine the situation without bothering to make all those dogs. Let's just look at a few examples; the 3rd percentile is the value that separates the smallest 3 per cent of the observations from the largest 97%; the 50th percentile is the median; the 90th percentile indicates the value that is exceeded by only 10 % of the samples; and so on.

There is a very elegant way to represent median and quartiles of a collection of numbers: the box-and-whisker plot. It is easy to do with R. Here are the heights of 15 dogs that I made using sheets of paper that were 95 mm square.

```
Ta<-c(52,52,60,59,50,55,57,57,57,56,52,57,55,54,51)
```

With this simple command we can create the graph.

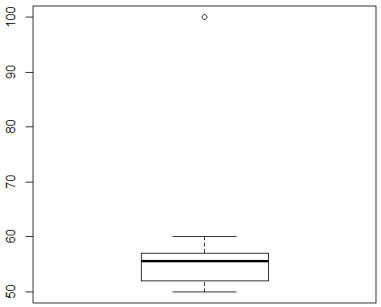
```
> boxplot(Ta)
```



The box represents the first and the third quartile (so the box contains 50% of the observations). The line within the box is the median, while the whiskers extend from the minimum value to the maximum value.

Now, let's imagine there is a another dog, the sixteenth, that is 100 mm high. I know it is impossible to make a dog so high with paper that is only 95 mm high, but do what my little cousin did. She desperately wanted to play with me so she made a dog with a sheet of her own paper... so, add it to the data set and redo the chart.

```
> Ta<-c(Ta,100)

> boxplot(Ta)
```



There is one difference: the " abnormal " value is reported as a single dot, separate from the whisker. But how did R understand that the little dog had been folded my little cousin?

He did not " get it" , for reasons that you will be able to understand from the things that we will look at in the next chapter.
It is very unlikely that such a high value would be found by chance in a group like ours. So the rule is that the whisker cannot extend beyond the median more than one and a half times the interquartile range. Any values beyond this limit are represented as isolated dots.

We have seen some indices of position and dispersion. There are also indices of *association*. In fact, sometimes it helps to have something that will show us if, when one measure varies maybe there is another one that varies in a similar way Let me explain this with an example. When making nougat, the mixture of sugar and almonds passes between two cylinders, and then the strip that comes out is cut into individual pieces. It is likely to be more convenient to check the weight of the pieces of nougat rather than their length, because the ruler sticks to the nougat so it is tedious taking the measurements (don't lick the ruler please!).

It is reasonable to think that, if the cross section of the nougat is constant, there is a relationship between the length and the weight of the nougat. This relationship may not be completely accurate; for example, it depends on how many almonds happen to be in a

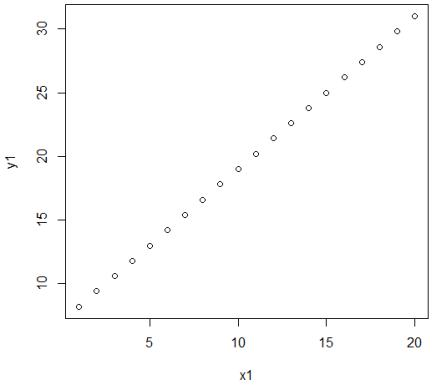
particular piece of nougat. Using statistics we can evaluate the relationship between the weight and the length of a piece of nougat. For example, we can calculate what percentage of the variations in the length can be explained by variations in the weight. This is a measure of association and is generally denoted by the symbol R^2.

Another measure of association is the correlation coefficient, also called "r"; a number that is equal to 0 when the two variables are not associated in any way; equal to 1 when given one measure we can obtain the exact value of the other measure, and that as one measure increases, so the other measure increases; and equal to -1 when one measure increases and the other measure decreases, again with an exact mathematical relationship.

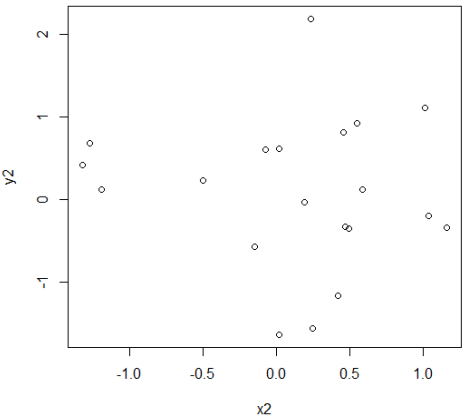
Lovers of formulas can find the formula to calculate r set out in Appendix D, a formula that makes use of another association index, the covariance.

It is not difficult to get help from R to graphically display these things.
With the four commands in the next page we can create a variable x1 containing the integers from 1 to 20. Then we can create another variable y2 by taking x1 and multiplying by 1.2 and adding 7. The graph of the variables is shown, and below is the calculation of the correlation coefficient. This is exactly equal to 1 since it is obtained from x1 and y1 according to an exact calculation: the two variables are perfectly correlated.

```
> x1<-1:20
> y1<-1.2*x1+7

> plot(x1,y1)
> cor(x1,y1)
[1] 1
```



Second example. This time with the function `rnorm()` we can create two vectors with 20 random numbers in each. If we create the graph and calculate the correlation coefficient this time the result is a very low number. You can see from the graph how the points are distributed.




```
> x2<-rnorm(20)
> y2<-rnorm(20)

> plot(x2,y2)
> cor(x2,y2)
[1] -0.08660972
```

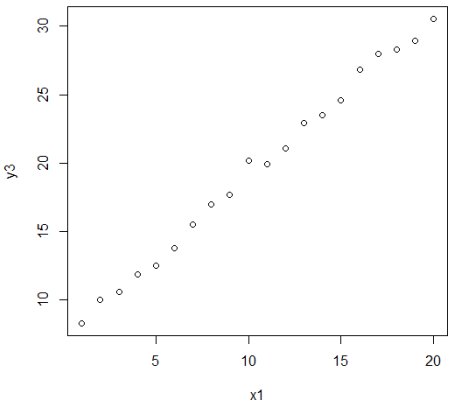
Note that, given the way these commands are written, R will generate different random numbers every time the command is

executed, so your chart and the r value may be different from those shown.

Third example. This time we take x1 and y1again, but to y1 we add a small amount of "randomness" (for now don't worry about how we do it)

```
>y3<-y1+rnorm(20,mean=0,sd=0.5)

> plot(x1,y3)
> cor(x1,y3)
[1] 0.9975803
```

This time the correlation coefficient is no longer 1, but it is still quite a high value and the dots on the graph are fairly well aligned, although not perfectly.

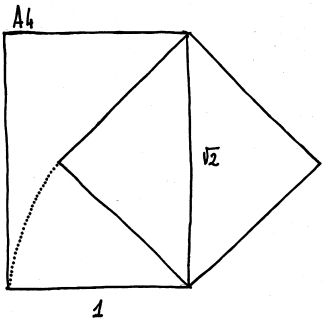


If you want to explore this branch of statistics in more detail, maybe you could try to see if there is any association between the values of Ta, Si and α for the puppies that you folded, but we will leave this for another story.

Chapter 4
Statistical distributions
(infinite masu)

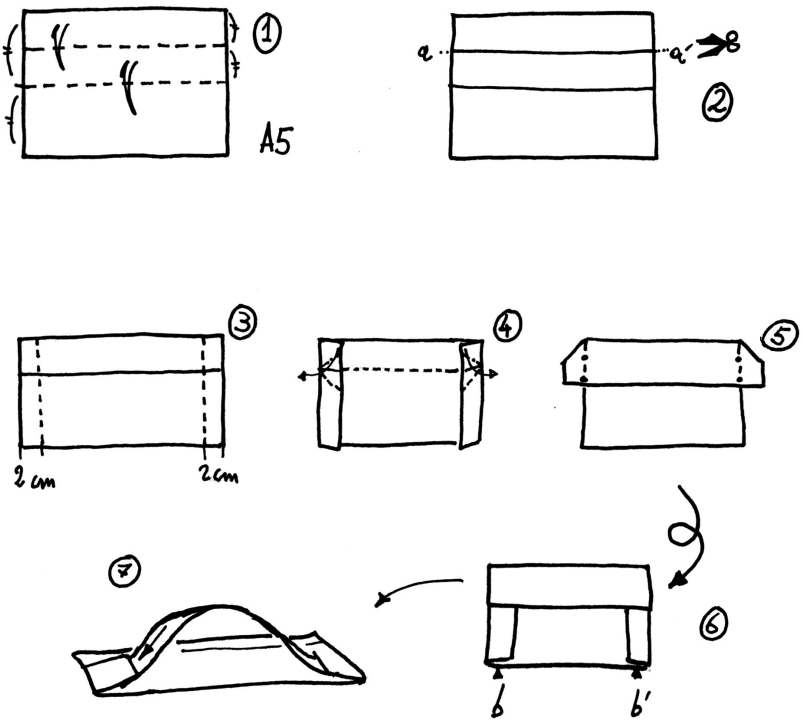
You will certainly remember that, to simplify our lives, we had imagined that Gervase's warehouse contained only 5 bags of sweets. But I like to work with very large quantities, so let's imagine that you have the weight in grams of all bags of sweets that Gervase's machine ever produced, and of all the bags it will produce in the future. In fact, I want to spoil myself, so let's imagine the infinite number of bags already produced and to be produced in the future, all weighed with absolute precision, then we can make a histogram. It is impossible, you say; to make infinite numbers of masu will take an infinite amount of time and an infinite amount of patience. So, before I wear out your patience, I will do a little trick. Let's go back to playing with paper: if we take an A4 sheet of paper and cut it in half we get 2 sheets of A5 size paper.

An interesting feature of the paper we use to make photocopies in Europe (size UNI) is that the long side is as long as the short side multiplied by √2 (I know it is not nice to express it in this way, but it does not matter) . It would be like saying that the short side is equal to the side of the square of which the long side is the diagonal (it is a bit of a tongue twister). However, the interesting thing is that, by dividing a sheet as we did the proportions are exactly the same. So each of the A5 sheets of paper has the same proportions as the original A4 sheet: they are similar rectangles.



In fact, for what we are doing in this chapter it is not important that the sides are in the proportion 1 ÷ √2. So, if you don't have any A4

paper to hand, you can use "letter" format paper cut in half, but the nerd in me could not avoid babbling about √2. Take the A5 sheet of paper (or half-sheet of letter paper) and fold it like this:



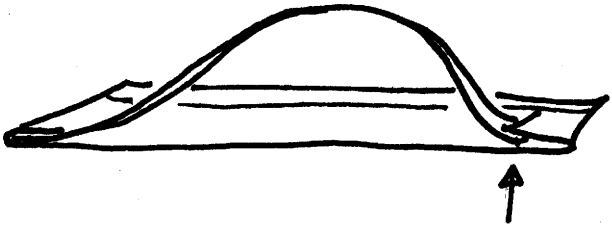
Then cut along the line a-a keeping the small strip of paper to one side and continue the folding. Finally, insert the strip of paper that you put to one side into the two pockets b b_1. Then you have to turn the model so that you can look at it from the side.

Now look at the profile of the strip of paper. It shows a special curve which is very important in statistics and defines exactly the form of the histogram that we wanted to create with an infinite

number of infinitely small masu. Not bad eh ? Just like that: 2 cuts and 3 folds, instead of an endless number of masu made with infinitely small sheets of paper.
A lot of work saved; I can almost feel your gratitude.

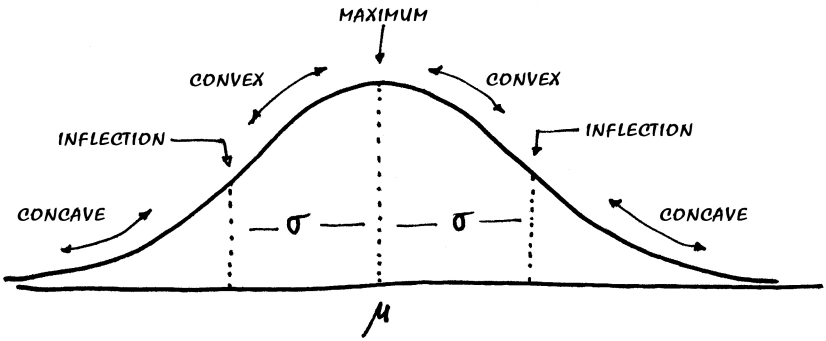
In fact, there are some things to clarify. The curve we have created shows the distribution of an infinite number of masu if Gervase's battered old packaging machine continued to make errors in the *normal* way. What does it mean when we say "make errors in the normal way"? Let's say that "normal" mistakes are those that occur in a totally random way. This is regardless of where the mistakes take place: they could be "mistakes" made by the packaging machine or errors in the measuring instrument, but they must always be random, so that there is nothing that alters the values in a systematic way.

For example, the distribution is *symmetric*, which means that Gervase is honest. In fact, try to move the extreme right of the strip a little bit like this:



We get a different distribution; an asymmetric (skewed) distribution. This would occur if Gervase, occasionally realizing that a bag was too full, took that bag away from the warehouse, but he only took away bags that were too full, and not those that were too empty (he isn't stupid!). Now put the strip of paper back in position in order to return to a symmetrical distribution. Look carefully at the end of the strip; depending on how you have folded the paper the paper model may or may not touch the table surface. Note that there is an important difference between a real normal distribution and the model that we have created: the strip touches the table, but only at an infinite distance. The rest of the strip has an infinite length (like the table) but I'm sure you can imagine that without cutting down an infinite number of trees in order to build an endless strip of paper. By the way, when you have finished playing please remember to throw the paper into the appropriate container for recycling.

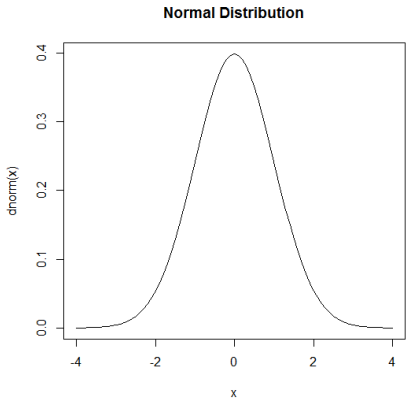
But back to the normal distribution, which is also called a *Gaussian* after the famous mathematician Johann Carl Friedrich Gauss (1777 - 1855). If you look closely you can see that the curve rises first with an upward concavity, then the curvature changes and becomes convex. The curve reaches a maximum, then it falls again with a convex shape and then a concave shape. The highest point of the curve corresponds to the average (try to find the centre of gravity of the Gaussian that you built), while the point where the curve turns from concave to convex (mathematicians call it the *inflection point*) is exactly one standard deviation away from the average.



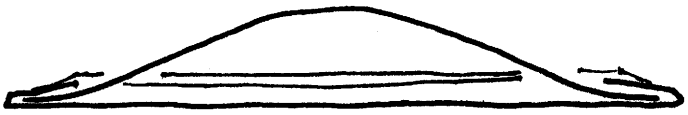
Nice eh! The first time I heard it I enjoyed it a lot; you know it doesn't take much to amuse me.
In the formula of the Gaussian (in the appendix) the symbols μ and σ are used (read mu and sigma), where μ is the mean and σ is the standard deviation. These are called the *parameters* of the Gaussian, because from an average and a standard deviation, one and only one Gaussian can be obtained. When μ=0 and σ=1 the Gaussian is called a standardized Gaussian.

This is how you can ask R to draw a standardized Gaussian:

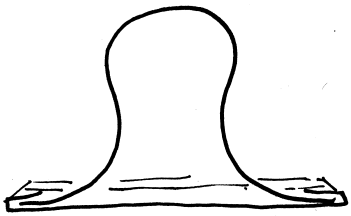
```
x<--40:40
x<-x/10
win.graph()
plot(x,dnorm(x),type="l")
title(main="Normal
Distribution")
```



Now try to build another Gaussian like the one we just made, but before inserting the strip into the two pockets shorten the strip by 2cm. You should get something with this shape:



This is a Gaussian with a larger standard deviation. By comparing the two Gaussians and by moving them you can simulate what happens when the mean and the standard deviation change. When the average changes the Gaussian moves to the right or to the left (the centre of gravity moves) and when the standard deviation changes the Gaussian "widens" or "narrows". In reality it is not the case that it gets wider, remember that the ends stretch to infinity, so Gaussians are all the same width. Let's say that when σ increases the Gaussian becomes a bit overblown.
If you play a little with long and short paper strips of paper you will soon realize that stretching the strip to simulate a reduction in the standard deviation only works up to a certain point. Beyond this point the curve takes a shape like this:



which it is not a Gaussian. In the end we only made a model of a quite complicated mathematical function, and this model, like all models, has certain limitations.

The Gaussian has some very interesting properties. For example, the mean, given that it divides the distribution exactly in two, is equal to the median. Also the range from the mean minus the

standard deviation to the mean plus the standard deviation includes approximately 68% of the total observations. More generally, the following can be calculated:

μ	$\pm 1\ \sigma$	= 68.26 %
μ	$\pm 2\ \sigma$	= 95.45 %
μ	$\pm 3\ \sigma$	= 99.86 %

Then you can calculate the percentiles based on the standard deviation. Let's get R to help us.

```
> pnorm(-3:3)
[1] 0.001349898 0.022750132 0.158655254 0.500000000
[5] 0.841344746 0.977249868 0.998650102
```

The `pnorm()` function does everything we need. It needs a mean which we haven't given it, so we can use the average for the standardized Gaussian (= 0). Then it needs one or more standard deviations, and here there is a very convenient R function " : " (colon) which can be translated as "count from ... to .." in this way:

```
> -3:3
[1] -3 -2 -1  0  1  2  3
```

in our case, from minus 3 to plus 3. In this way R calculates the area under the Gaussian for the mean plus or minus 1 , 2 or 3 standard deviations. Obviously we have to specify whether we are interested in the area to the left or the area to the right of each of the specified points. This is done with the option `lower.tail`. As is evident from the examples below, the default option is to calculate the area to the left. The results of the calculation are 7 numbers, one for each σ. As usual R numbers the results in square brackets and moves to a new line when there is no longer place on a line.

```
> pnorm(-3:3,lower.tail=TRUE)
[1] 0.001349898 0.022750132 0.158655254 0.500000000
[5] 0.841344746 0.977249868 0.998650102

> pnorm(-3:3,lower.tail=FALSE)
[1] 0.998650102 0.977249868 0.841344746 0.500000000
[5] 0.158655254 0.022750132 0.001349898
```

Now let's look at the results. How do we get the area under the curve that lies between $\mu \pm 3\sigma$?

Just subtract from 1 (the area under the standardised Gaussian) the area to the left of $-3\ \sigma$ and the area to the right of $3\ \sigma$ (the areas are equal, because the Gaussian is symmetrical). Here we are:

```
> 1-0.001349898*2
[1] 0.9973002
```

That is, approximately 99.7 % , as reported in the table above.

I'm sure you understand, so there is no need to repeat the calculation for 1σ or 2σ.

Let's go back to Gervase who, given the poor performance of his packaging machine, decided to carry out some drastic exceptional maintenance. So, after convincing Adalgisa, the hen, to choose somewhere other than the scales for laying her eggs; and after having forced shut the box of Christmas decorations that for years had been sitting around, and after sorting out various other small details he picked up some new bags of sweets and here are the weights in grams:

2995 3010 3007 2999 2998 2994 3006 3003 2998 2992
3002 3004 3005 2997 3002 3003 3006 3002 3009 3008
3000 3001 2995 2990 3011

Let's assign these values to a variable in R that we will call "weights " (you can cut and paste the numbers from the page)

```
> pesi<-c(2995, 3010, 3007, 2999, 2998, 2994, 3006,
+   3003, 2998, 2992, 3002, 3004, 3005, 2997, 3002,
+   3003, 3006, 3002, 3009, 3008, 3000, 3001, 2995,
+   2990, 3011)
```

please note that by starting a new line after a comma, R recognized that I was writing an incomplete expression and started the next line with a "+ " , which in this case has nothing to do with addition but simply means "continued from the previous line"

Now it is easy to do some calculations:

```
> mean(pesi)
[1] 3001.48
> sd(pesi)
[1] 5.672448

> mean(pesi)-3*sd(pesi)
[1] 2984.463

> mean(pesi)+3*sd(pesi)
[1] 3018.497
```

Therefore , Gervase concluded that he expected about 99.8 % of his bags of sweets to weigh between 2.984 Kg and 3.018 Kg. Now the children should no longer complain.

Chapter 5
Other distributions

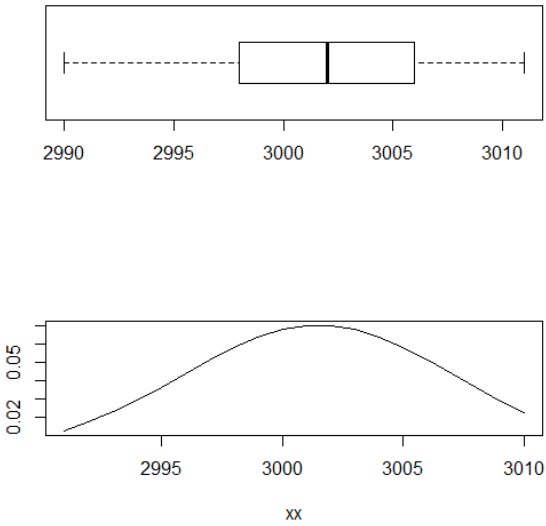
Now it should be clear how convenient it is to use the Gaussian distribution as a model of random events. In fact we do not have an absolute guarantee that the errors in the sweet packaging machine have exactly a Gaussian distribution, but this issue is so important that there are many *tests* (be patient; we will talk about tests in chapter 7) to try to understand if it is too risky to assume that the data has a normal distribution.

Here it is not appropriate for me to do a complete review of the various tests. There is the Shapiro-Wilk test, or the Chi-square test, or the test with the best name of all: the Kolmogorov-Smirnov test.You can get R to help you to understand how to use them or, better still, look at a book (also Wikipedia talks about the tests). There are also several graphical methods to address this problem; in the next chapter, for example, we will see how to superimpose a histogram on a Gaussian.

Another way to try to understand if our data follows a normal distribution is to make a box-and-whisker plot, like the one in chapter 3. If the data follows a normal distribution, the median will be equal to the average, so it should be in the middle of the box. Moreover, given that 50 % of the observations are in the box, the first and third quartiles should not be far from the inflection point of the Gaussian (the mean ± 1standard deviation). The two whiskers should be approximately equal in length and there should be few outliers, that is data beyond the mean ± 1.5 times the interquartile range (which is a number not very different from 2 standard deviations). Below, R is used to create the boxplot of the set of weights from the previous chapter, and a Gaussian with mean and standard deviation equal to those weights.

```
win.graph()
par(mfrow=c(2,1))
boxplot(pesi,horizontal=TRUE)
xx<-2990+1:20
plot(xx,dnorm(xx,
mean=mean(pesi),sd=sd(pesi)),type="l",ylab="")
```

Please note the delightful `mfrow` option in the `par` command, which allows us to create a window with two (or more) graphs together. In this case, the data seem pretty close to the normal distribution and the left whisker is just a little bit shorter than the right one.



However, having a reference distribution is so helpful that statisticians have looked for other distributions that are suitable for describing different events. For example, we said that the Gaussian describes a continuous measurement to which a random error is applied. On the other hand, the *binomial* distribution can be used to describe events with only 2 possible alternatives. Like the time Gervase decided to produce sweets with a hole; how likely is the sweet to have a hole in it and how likely is it to be produced without a hole.

In general, the *Poisson* distribution is said to be suitable to describe rare events. For example, it was used by Colonel von Bortkiewicz (1868- 1931) in the late 1800s to describe the number of deaths annually from horse kicks in each unit of the Prussian army.

The *uniform* distribution describes events that all have the same probability, while the *Weibull* distribution is often used to describe the incidence of failures.

Many stories have been written about these and other distributions, but they are not for this book.

Chapter 6
Sample mean and population mean
(How much liquorice juice did you put in?)

One day Gervase set off on a journey. He had to participate in an internship about sweets organized by his friend Chalkbeard . He had left the laboratory in the loving care of his best worker: Tony. At that time a batch of wonderful liquorice blackberries, made with Gervase's secret recipe was being processed. Except that, in the bustle of departure, Gervase had forgotten to leave precise instructions as to how to proceed with the work. In particular, Tony could not establish how much liquorice juice Gervase had already added to the pot.

In fact Gervase had already put exactly 500 millilitres of liquorice juice into the pot that contained 50 litres of syrup. So 500 millilitres divided by 50 litres (that is to say 50000 ml), gives exactly 0.01 i.e. a concentration of 1%. But Tony did not know this and did not want to disturb Gervase by sending him a carrier pigeon (mobile phones were not in use at that time).

Therefore Tony decided to take a small sample from the pot and to analyse it to determine the exact concentration of liquorice juice. The analysis of the sample yielded these five results:

0.01 0.015 0.02 0.008 0.022

with an average of 0.015.

But shouldn't it be 0.01? We already know that the mean was exactly 0.01, but Tony did not know that and it is possible that due to an imperfect mix of the ingredients, or due to some inaccuracy in the measuring instruments, the average of a small sample of measures was not exactly equal to 0.01. Life is full of infinite absurdities, which, strangely enough, do not even need to appear

plausible, since they are true (in this way we have also quoted Pirandello [6]). So Tony thinks that 750 ml of liquorice juice was put in the pot, while we know that was not the case. 'Statistics lie' I can almost hear a little voice say; but once again it is not true. We simply have to be very careful not to confuse the average calculated on the basis of a sample with the true mean of the whole pot. The average derived from a sample, calculated in whatever way (it makes no difference if it is calculated by hand, with a computer or with our paper computer) is an average based on a sample. For this reason it is called the *sample mean* and is generally denoted by a small line above the variable name. For example, the mean of x is $\bar{x}$.

On the other hand the mean of the whole pot is usually called the *population mean* or the *real mean* and you can never know exactly what its value is. It is denoted by the Greek letter μ (mu) and is one of the parameters of the Gaussian. Already, it is logical to think that, if we were to analyse the whole pot, with an infinite number of samples, we would not always get the same value, but the measures we get would be normally distributed, like a Gaussian, with a mean and a standard deviation. This is because the movement of the molecules of liquorice juice in a pot is inherently variable and can only be described with statistical methods. To use the jargon it is said that it is a *stochastic* phenomenon.
I can assure you that, when I started studying statistics, by myself, it took me a long time to understand why the mean was represented in one part of the book by the symbol $\bar{x}$, while elsewhere the symbol μ was used. Now it should be obvious that $\bar{x}$ is simply the result of a calculation, while μ is something that we do not know but that we would like to estimate. In this way we can use it as a parameter of a Gaussian and as a model of the whole universe of data that we are analyzing.
Typically, we are in the same position as Tony; we cannot know the population mean (we cannot analyze the whole pot), we can only

calculate the average of a sample. But, you say, it would be helpful if there were some sort of relationship between the two: in italian they have the same name![*] In fact, one of the purposes of statistics is precisely to help us to *estimate* the true mean. To use statistical jargon, the process of estimating parameters is called *inference* and this particular branch of statistics is called inferential statistics.

How do we do it? It is simple. First of all we calculate the standard deviation of the samples (we will call this *s*) and *s* is divided by the square root of the sample size. Tony had 5 samples taken from the pot, so the standard deviation calculated with 5 samples can be calculated using R.

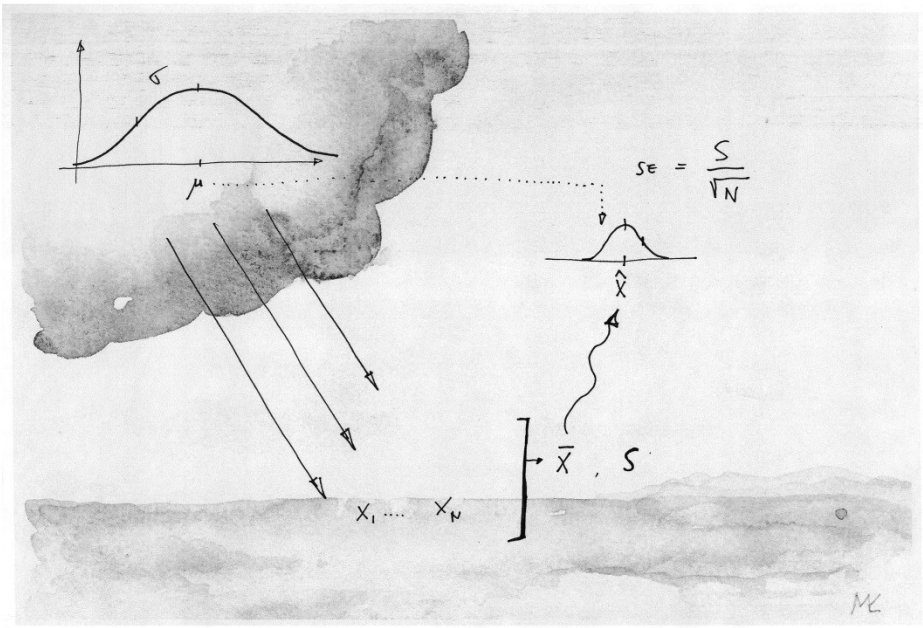
```
samples<-c(0.01, 0.015, 0.02, 0.008, 0.022)

> mean(samples)
[1] 0.015

> sd(samples)
[1] 0.006082763

> sd(samples)/sqrt(5)
[1] 0.002720294
```

The answer is approximately 0.0027, and this value is called the *standard error.*
So Tony did not know the true mean, but statistics tells us that the probability of finding it is a Gaussian distribution (also this!) with a mean equal to the sample mean and a standard deviation equal to the standard error. Then (according to the table in Chapter 4) there is a 95% probability that the true mean is between 0.015 plus or minus twice the standard error. Therefore it is between 0.0204 and 0.0096 and, in fact, the true mean (which we know) is between these two values. In conclusion, Tony knows with 95% probability

* In English the term average is referred to a sample while the term mean is related to the true mean, but often they are interchangeable.

that Gervase put between 480 ml and 1020 ml of liquorice juice into the pot. You might say that this is a little vague. This is not the fault of statistics; we can either reduce the standard deviation or increase the size of the sample. It is obvious that if the sample becomes larger, the estimate improves, so as n tends to infinity the standard error tends to zero and the sample mean becomes equal to the true mean. Similarly, it is obvious that if we mix the contents of the pot well and use accurate methods of analysis *s* becomes smaller and the estimate improves. However, often researchers need statistics because the phenomena that they are studying is inherently uncertain and there is no way of reducing *s*.

This way of working is common in statistics and is called the *interval estimate.* When we estimate something that is impossible to calculate exactly, we calculate a range that contains, with a known probability (typically 95%) the true value. This range is called the *confidence interval* and, from the above, you've probably figured out that it can be calculated for all kinds of estimates.



I will try to explain it another way, and with a little drawing. There is a universe that we do not know (if we knew it there would be no need to use inferential statistics) which I have put in a cloud. From this we extract a sample of n observations. With this sample we carry out calculations to find, for example, the mean and standard deviation. The square bracket in the drawing is our computer, R. We assume (*inference* is the wavy line), that our sample mean is the best possible estimate of the true mean μ (to indicate the estimated average a "little hat" is put on the variable, like this $\hat{x}$). Finally, by calculating the standard error we estimate what mistake we could make by deciding that the figures that we calculated describe the universe we started with (dotted line).

Please note that this general reasoning applies to anything we want to estimate with statistics; the important thing is to choose the right distribution to use and to decide how to sample it.

At this point there could be a long discussion about inference. The fact is that this reasoning is valid only if certain things are valid: i.e. there are some *assumptions*. In this particular case, the only assumption is that the errors are distributed according to a Gaussian. But be careful; each inference depends on specific assumptions that will be relevant to each particular case, otherwise our conclusions may be completely wrong. Furthermore, the results of the inference depend not only on the assumptions, but on how we selected the sample. Precise rules exist on how to carry out sampling, which I cannot tell you about here; just remember that sampling should not be done " haphazardly".

But there is another interesting thing. There is a theorem called the central limit theorem that shows that whatever the distribution of our starting universe (well σ^2 should not be infinite), if we extract n samples many times and every time we calculate a sample mean, all these averages will tend to be distributed according to a

Gaussian, and we have already seen how easy it is to use the Gaussian model.

Now, please go and get the histogram that we made with the masu on cha.1 and also the Gaussian on cha.4 that I told you was like a Gaussian histogram made with an infinite number of infinitely small masu. Now we can be more precise: in fact we can check that our paper Gaussian has a σ equal to about 1 masu, while the s calculated on cha.2 was 1.2 because we had used A4 paper, pieces of paper about 10 cm long, and had folded the paper into 2 cm lengths...these figures having been chosen to make the numbers work.
Now we can try to superimpose the Gaussian on the histogram, knowing that our estimate $\hat{x} = 3$ is affected by a standard error of $1.2 \div \sqrt{(5)} = 0.54$ Therefore, the true mean could be between 1.92 kg and 4.08 kg (with about 95 % confidence).

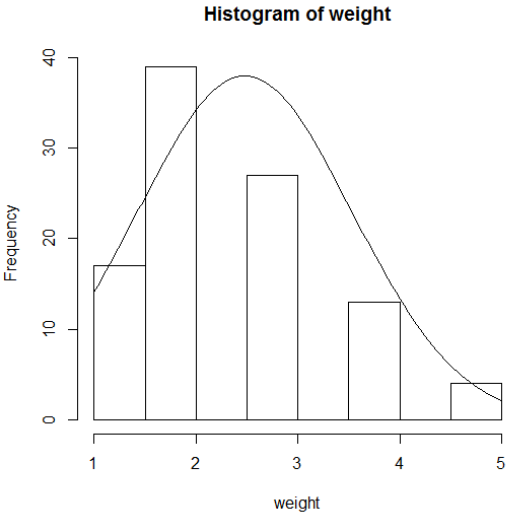
Do you remember the large warehouse with all of Gervase's bags of sweets mentioned on cha.1? Now, with the help of R all of those numbers no longer look so frightening.

```
 > weight<-c(2, 3, 3, 5, 2, 3, 3, 2, 2, 3, 2, 3, 1, 2,
+  3, 3, 4, 3, 4, 2, 4, 3, 1, 5, 1, 3, 1, 2, 2, 2, 4,
+  3, 2, 2, 4, 3, 5, 3, 2, 1, 4, 3, 2, 3, 2, 3, 1, 4,
+  5, 1, 1, 3, 3, 1, 2, 2, 1, 4, 3, 2, 2, 2, 2, 2, 3,
+  4, 2, 2, 2, 1, 2, 2, 3, 2, 2, 3, 4, 1, 2, 3, 3, 4,
+  2, 2, 2, 1, 3, 3, 1, 4, 1, 2, 1, 2, 1, 2, 2, 4, 2,
+  2)

> mean(weight)
[1] 2.48
> sd(weight)
[1] 1.049098
> sd(weight)/sqrt(length(weight))
[1] 0.1049098
```

So the mean is equal to 2.48, the standard deviation is equal to 1.05, and the standard error is equal to approximately 0.10.

So the true mean should be between 2.28 and 2.68 kg (with approximately 95 % probability) .
And finally we can draw the histogram of a Gaussian and with μ=2.48 σ=1.05

```
> win.graph()
> hist(weight)
> plot(function(x) length(weight)*
   dnorm(x,mean=mean(weight),sd=sd(weight)),
   from=1,to=5,add=TRUE)
```



I admit that the last R instruction is a little bit cryptic. The fact is that the function `plot()` has a lot of options and possibilities. If you are curious and want to know more try writing:

```
>?methods(plot)
```

and in particular:

```
>?plot.function
```

Chapter 7
The verification of a test
(the soft toffee)

You should know that one of the real specialities of Gervase is toffee that, in addition to being delicious, is really soft and juicy. However, things do not always go well. Once, before Berta had learned to spin, Gervase was producing soft toffee that stuck to the teeth, until one day his young assistant (Tony, again) stumbled upon an ingredient that, when combined with the toffee mixture produced a toffee that was even softer and melted more in the mouth. I am sure that the more curious among you will want to know what it was; unfortunately it was so long ago that it has been forgotten. However, Tony prepared a few of these new sweets and gave them to Gervase to try. Gervase liked them but said to Tony that he had always produced his sweets with his traditional recipe and before changing the recipe he wanted to make sure that the new sweets were really softer. How can we be sure that these new candies are softer because of the new ingredient in Tony's mixture?

"You used a different mould, you regulated the heating of the mixture in a particular way, then there is the cooling temperature, and then of course, each sweet has its own melting characteristics" Gervase said.
"Let's do this: let's prepare two lots of sweets, one with the old recipe and one with the new recipe, trying to heat them the same amount, using the same mould and cooling them in the same way. Then we can measure the melting characteristics and make a comparison."
"Okay" Tony answered.
Do you want to know how to measure the melting characteristics of toffee? It is easy; you do the dragon spit test. You put the toffee in a glass full of dragon spit and measure how long it takes to melt completely.

Here are the melting times of 10 sweets made with Tony's recipe. We'll call them A, for Tony, whose real name was Antony.
While we're at it, we might as well put them in R and calculate the mean.

```
> A<- c(72, 82, 65, 83, 50, 61, 83, 68, 52, 75)
> mean(A)
[1] 69.1
```

Then we collect the melting times of 10 sweets made with the traditional recipe, which we put in group G, for Gervase.

```
> G<- c(89, 71, 76, 81, 75, 79, 60, 62, 70, 61)
> mean(G)
[1] 72.4
```

But how do you tell which ones are the juiciest? The average melting time in dragon spit of group A is less than the average for group G, but there are a couple of values in G below the average of A. Are these exceptions? So, should we do another test? Are you going to get all the dragon spit that is needed? Be careful, you can't just throw away the used spit wherever you like: it must be disposed of carefully because it is a pollutant.

In reality, things are much simpler. All that is needed is a test. In statistics we talk about hypothesis testing because, in effect, we make a hypothesis and then verify that hypothesis. Or should I say, we try to falsify that hypothesis. In fact, the hypothesis is always a hypothesis of no difference. In our case the hypothesis is that G is equal to A and is called the *null hypothesis*, known as H0 to its friends. So H0: A = G
But if A equals G the difference between the averages should be zero. However, we must remember that the two means are only estimates so we have to take this into account by calculating the standard error of the difference between the means.

Then we calculate the difference between the means divided by the standard error of this difference. I know very well that we have not yet learned how to calculate the standard error of a difference between means, but I am very keen on formulas. The formula, like all the others, is in the appendix.

What is very interesting is that the number that comes out also follows a well-known distribution, that is, it follows a well-known mathematical function. The distribution was described for the first time by William S. Gosset (1876- 1937) in 1908 while working for Alec Guinness & Co. the beer company. I told you that statistics is very useful - it is also connected with making beer!

Dr. Gosset published his findings under the alias "Student", so the distribution is called Student's t. This distribution allows us to calculate how likely it is that a certain value of t (from Student) occurred by chance. Since t is a difference between the means (divided by an error), it is like saying that it is a coincidence that there is a difference, which is to say there is no difference. That is a bit like saying that there is a probability that H0 is true.
Then, if this probability is sufficiently low we can conclude that H0 is probably false, and so A is different from G.

I know that you have probably lost the thread of the argument, Let's recap with an outline.

Basically someone has already taken the trouble to:

1. invent a formula that measures the difference between two samples
2. demonstrate that the result follows a distribution
3. calculate the values of this distribution
4. arrange them in a table (or include them in a program such as R), ranked by the probability that H0 is true; i.e., it is true that there are differences between the means.

All that remains for us to do is to:

a) define a H0 (something equals something else) which really we want to falsify
b) calculate the statistical test (t , in our example)
c) look up in the table, or get R to calculate, the probability that H0 is true for the value (t)
d) look at whether the probability is high or low
e) if the probability is low to reject H0 (the two are different)
f) if the probability is high then we say that we cannot reject H0 (so probably the two things are not different).

Like all sciences statistics has its own jargon, and like all languages statistical jargon has a reason to exist. In fact, saying that we reject H0 (point "e" above) is a bit like saying that H0 is false, but it is more correct to say that it is probably false. On the other hand, saying that if the probability is high we "cannot reject H0" (point "f" above) seems a bit Byzantine, but in fact it is not, because I have not been entirely clear with you. The probability that you find in the table is not the probability that H0 is true, but the probability of making a mistake by saying that it is false, and this is not the same thing.
For example, we have obtained a high P (and thus are not able to reject H0) because we used a sample that was too small. Is there a way to check this? It involves calculating the *power* of the test that we used, and R has many tools for carrying out such calculations, for example a beautiful package is called `pwr`; but its use is beyond the scope of this little book.

But let's go back to our toffee. This is the value of the Student's t calculated by R for groups G and A

```
> t.test(A,G)

        Welch Two Sample t-test

data:  A and G
t = -0.6753, df = 16.988, p-value = 0.5086
alternative hypothesis: true difference in means is
not equal to 0
95 percent confidence interval:
 -13.610851   7.010851
sample estimates:
mean of x mean of y
     69.1      72.4
```

So t is equal to about minus 0.67. This corresponds to a probability (P) of 0.5086 , that is to say about 50%. So there is a 50 % chance of making a mistake by saying that G is not equal to A (rejecting H0). Therefore we should not reject the null hypothesis: Tony's new ingredient does not change the melting time of the sweets significantly.

What we have done so far applies to many statistical tests; all that changes is the statistical test and the distribution it refers to. Let's look at a few examples.

In our case we are comparing the melting times of 2 samples of extra-soft toffee. The measurements of the first sample are independent of the measurements of the second sample so the Student t test is okay.

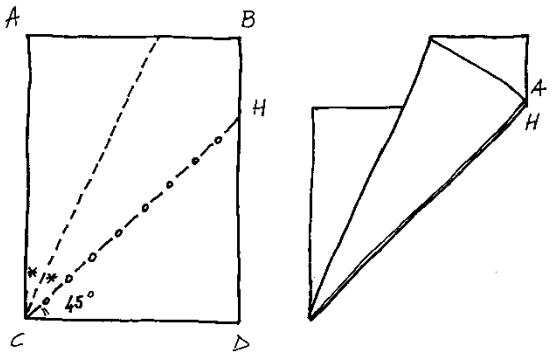
But once Gervase found himself in trouble with the elves who, tempted by strings of liquorice ran the risk of getting high blood pressure. Indeed, some argue that eating too much liquorice causes blood pressure to increase so Gervase called a medical friend who measured the blood pressure of the elves before and after the shift making strings of liquorice. Everyone knows that when they work

with strings of liquorice the elves taste a little here and there; you can't take it away from them.

In this case we cannot overlook the fact that the blood pressures refer to the same elves, that is, the x-th measure before working with the liquorice corresponds to the x-th measure after working with the liquorice because both relate to elf x . In this case you cannot use the Student t test in the form we saw above, but you have to use another formula that is called the *paired Student t*. The distribution it refers to remains the same, but the way you calculate t and the degrees of freedom change.

I can assure you that it is not Byzantine. Let's look at an example. Let's create two variables with R:

```
M<- c(5,23,18,9,12,25,19,14)
N<- c(7,24,19,11,15,25,20,15)
```



If we want to compare them with the Student t we must first know where the data came from in order to decide whether to use the Student t for paired data or the student t for unpaired data. We have already seen that for an A4 sheet of paper the sides are in the proportions $1 \div \sqrt{2}$. This can be verified by doing the fold above and measuring the distance between A and H (the reason for this should be obvious if you consider that the valley fold divides the angle

ACH exactly in half, while the mountain fold is the diagonal of a square).

Imagine that you want to compare the accuracy of two manufacturers of paper, then M and N are the measurements, in microns, of the distance between A and H for 10 sheets of paper taken from 10 different reams of paper for two suppliers, Manuel and Nando . Using the unpaired t (the R t-test for unpaired data is the "default " option, so it is not necessary to specify `paired = "FALSE"`. You can check this using `?t.test`) .

```
> t.test(M,N)

        Welch Two Sample t-test

data:  M and N
t = -0.4194, df = 13.854, p-value = 0.6813
alternative hypothesis: true difference in means is
not equal to 0
95 percent confidence interval:
 -8.413401  5.663401
sample estimates:
mean of x mean of y
   15.625    17.000
```

t= - 0.419 P= 0.68 df=16
so we cannot reject H0.

On the other hand, if we imagine that we have 10 suppliers of paper and that in a new contract the proposed price for the paper is based on the precision of their cutting, we can try to test whether the economic incentive has any effect by comparing the precision of the cutting for each supplier before (M) and after (N) the changes to the contract. In this case you have to use the Student paired t test:

```
> t.test(M,N,paired=TRUE)

        Paired t-test

data:  M and N
t = -4.2451, df = 7, p-value = 0.003816
alternative hypothesis: true difference in means is
not equal to 0
95 percent confidence interval:
 -2.1409 -0.6091
sample estimates:
mean of the differences
                 -1.375
```

t= - 4.24 P= 0.003 df=8
so we can reject H0.

So you can see. A different source of data leads (in this case) to opposite conclusions. This is very important. A computer cannot know where you got your data from and how you made the measurements. This is definitely something that you have to decide.

Sometimes it is necessary to analyze data that are not measurements, but numbers that relate to qualitative variables. Do you remember the sweets with the holes? Sweets either have holes or not; you cannot measure half a hole or two point seven holes. Again, in this case you cannot use the Student t but you have to use another statistical test, such as the *chi-squared test* (to its friends it is known as χ^2 and for R it is `chisq.test()`).

An interesting application of the chi square test is to see if it is reasonable to assume that a sample of observations come from a normally distributed population.

If the data are not normally distributed, then you cannot use many statistical tests. You have to resort to a new family of tests: *nonparametric tests*.

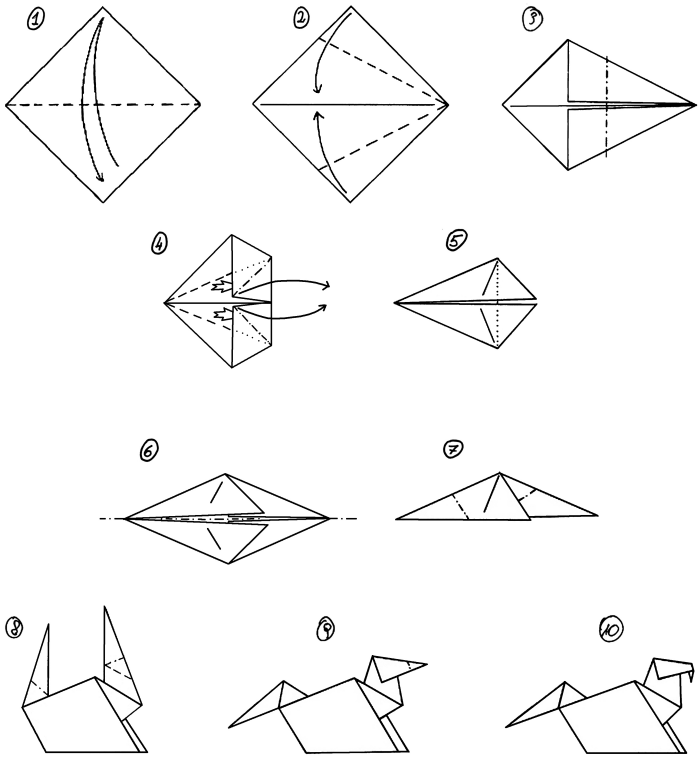
The important thing is that all these tests operate in the same way. You define a H0, you try to falsify it, you calculate a statistical test and you look it up in a table or on a computer. Once you know the mechanism it is the same every time. Just be careful to choose the right test.

Often, talking about probability I referred to low probability and high probability. But how high and how low? Usually we use 5% (in some cases 1%), namely 0.05 (or 0.01). If P is less than these values we feel we are permitted to reject H0.
But beware, there is still a 5 % chance of making a clumsy mistake, that is, to consider two samples to be different when they are not. This is called a *type I error*, also called an *alpha error* (α). Obviously there is also another kind of error which is when H0 is false but we do not reject it. This is called a *type II error* or *beta error* (β).

Reality	My decision	
	Reject H0	Do ot reject H0
H0 is true	α error	OK
H0 is false	OK	β error

Chapter 8
Paper Palaeontology

Let's play another game. Let's play at being scientists! Palaeontologists in search of dinosaurs.



The fold above is nice in terms of its simplicity. It reminds me of one of the first reptiles that adapted itself to amphibian life. I think its name was *Dentonus Cartaceus* , probably due to the fang with which it looked for food in the mud of the sea floor. It had a

beautiful strong tail which helped it to move and two front flippers, almost capable of functioning as rudimentary legs when it came out of the water. It was a rather lazy animal and used its fang to attach itself to ammonites and to get carried through the water rather than swimming. Hence the saying that is famous among the ammonites, "to scrounge like a dentone". You have never heard this expression? Well in fact it hasn't been used for a while, the ammonites having been extinct since the Cretaceous period ...

So if we make some Dentonus, in addition to inventing its characteristics and habitat (you realized that I invented it didn't you?), we can also study its shape. One of the jobs of a scientist is to measure things. If we begin measuring things someone will say "how boring this is". It may be that it is not very exciting, but without measurement science doesn't go anywhere. Indeed, if you look into the history of scientific discoveries you will find that, almost always, they are related to the collection of measurements. For example, Galileo spent a long time measuring the speed of falling objects, throwing iron balls from the tower of Pisa, while that scoundrel Kepler was able to formulate his famous laws because he had the data patiently collected by Tycho Brahe.
The fact is that even when taking measurements we can make mistakes. Statistics can come in handy to find out if our measurement errors are acceptable or not.
How does it work? You may have noticed that, like Nick's dog in chapter 3, also for Dentonus, some folds are left to the aesthetic sense of the person folding the paper. So if you like we can make some little Dentonus with different behaviours: one that is looking around; one that is delving in the mud; one that is looking up; one with its tail down, a little depressed about the ammonites; and one that is all excited, with a straight tail.
Then let's play at being palaeontologists. We have found these beautiful dentonus fossils and have to measure them. But if we are to do things properly, we must make sure that the various teams of palaeontologists around the world use the same method and the

same expertise in measuring the fossils. Dentonus are not like mushrooms and are not all in the same place, so, either we make sure that everyone uses the same method to measure them, or we must carry all of them to the same laboratory and measure them with the same system. It is pretty obvious that, besides our little game, the problem is actually a very serious one and involves all *multicentre trials*, to use the scientific jargon. For example, if we want to study a relatively rare disease, we have two options: either to spend a lot of time collecting a sufficient number of cases, or we reach an agreement with different hospitals and to put together the observations from each hospital. In the latter case it is absolutely essential that the measurements are done in the same way, otherwise measurement errors can totally invalidate the project. In fact, from one point of view, this chapter should have been the first chapter of the book; if Gervase had not been sure that the weights obtained from the scales were correct, or if Tony's way of measuring the melting time of the fudge had been different to Gervase's method then everything we have said and done so far would not stand up.

A good way to evaluate reproducibility is to repeat each measurement at least twice under different conditions that are known to be important, for example with different tools or with different operators, and then to try to measure any errors that were introduced by the repetition.

Be careful! To conduct a test such as one of those explained in chapter 7 and to apply it 'upside down', i.e. to look for a P greater than 0.05 and to conclude that there are no differences, therefore the measurements are reproducible, is totally wrong. In fact, to say that "there is a high probability (greater than 5%) of being wrong in saying that there is no difference between the two sets of measures " is not the same as saying that "the two groups are equal." Absolutely not! When P is not significant (> 0.05) it is like saying " we do not know ". For example, one possible cause of this might

be the size of the sample, as mentioned in chapter 7. The subject is too lengthy to cover in detail here, but intuitively it is easy to understand that:

- the variability of a phenomenon
- the significance of a test about a difference
- the sample size
- type α and type β errors

are all things that affect the situation.

If you think about it:
• the more a phenomenon is variable the greater the amount of data that will be required to find a difference.
• the bigger the difference the fewer the number of observations that will be needed.
But then, someone will say, if you dramatically increase the number of observations P will become more significant. Exactly! A difference, no matter how small, will become important if the number of measurements is high enough. For this reason statistics based only on P can be misleading: I could demonstrate that my medication to reduce blood pressure is more effective than a drug currently on the market, simply by taking millions and millions of measurements. It is too bad that the difference, though statistically significant, is only 0.1 mmHg (millimetres of mercury) which is clinically irrelevant. One of my teachers once said: "Statistics based on P is statistics for the rich" . This is because taking many measures costs time, effort and money as you might have realised when you had to measure the dogs in chapter 3.

But now let's go back to our problem of reproducibility.
Another idea is to use the measures of association (remember chapter 3) such as the correlation coefficient. Hopefully, a simple example will help you to understand that this is not a good idea. Imagine that 10 dentonus were measured from nose to tail by two researchers Bill and Bull:

```
Bill<-c(10,12,15,18,13,9,11,10,16,14)

Bull<-c(8.0,10.2,13.5,16.8,11.3,6.9,9.1,8.0,14.6,12.4)
```

Now if we ask R to put the values side by side, it is easy to see that at least one of the two has not done a very good job.

```
> cbind(Bill,Bull)
      Bill Bull
 [1,]   10  8.0
 [2,]   12 10.2
 [3,]   15 13.5
 [4,]   18 16.8
 [5,]   13 11.3
 [6,]    9  6.9
 [7,]   11  9.1
 [8,]   10  8.0
 [9,]   16 14.6
[10,]   14 12.4
```

Look at the numbers. It is not acceptable that the same dentonus is 10 cm long according to Bill but only 8 cm long according to Bull, and so on. Yet, if we ask R to calculate the correlation coefficient between Bill and Bull:

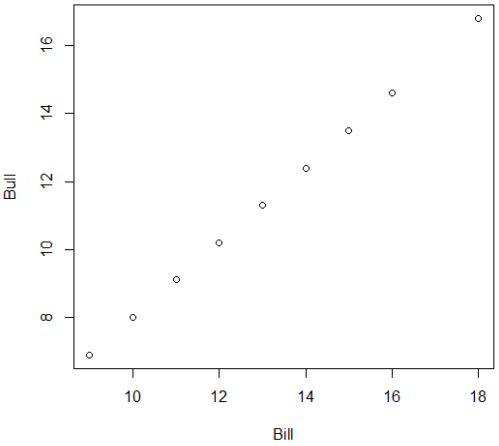
```
> cor(Bill,Bull)
[1] 1
```

The answer is 1, namely a perfect correlation!! Why is that?
The answer is simple. I created Bull's numbers artificially by increasing Bill's numbers by 10% and then subtracting three centimetres:

```
> Bill*1.1-3
 [1]  8.0 10.2 13.5 16.8 11.3  6.9  9.1  8.0 14.6 12.4
```

So it makes sense that r is equal to 1 and that the points on the graph are perfectly aligned. But remember chapter 1: Be careful of the scales! This just goes to show that the concept of association

between two measurements is very different from what we are looking for here. We are not interested in whether the measurements are associated, but whether there is agreement about the measurements, whether the measurements are reproducible or whether they match. In fact the terms reproducibility, reliability and agreement are not exact synonyms, but they indicate concepts that are quite similar to each other, but very different from the concept of association.



So how do you assess reproducibility?
There are so many ways to do it. For example, there is a graphical method proposed by Bland and Altman [11] which is quite nice, or there is the Intraclass Correlation Coefficient (ICC) [12]. A review of all the methods is beyond the scope of this text, I think it is sufficient to examine just one: The Concordance Correlation Coefficient, known as CCC to its friends.

One of the easiest ways to calculate the CCC is to use an R package called `epiR`. In fact we have not yet talked about R packages, they were only mentioned in passing in chapter 7. The fact is that this is something that is easier to use than to explain. Basically there are many additional packages that make R even more powerful and even more versatile. To use a package you need to install it by downloading it from a server, using the "Install package(s)" function in the "Packages"menu. At this point the package is on

your PC, but to use it you have to load it into the memory, for example, with the following command:

```
require(epiR)
```

At this point you can execute the function:

```
epi.ccc(Bill,Bull)
```

The result is something a bit unusual which is called a list and contains (among other things) the following:

`rho.c`	the value of CCC and, in particular
`$rho.c$est`	the estimate of CCC
`$rho.c$lower`	the lower margin of the confidence interval
`$rho.c$upper`	the upper margin of the confidence interval
`C.b`	The correction factor C_b

The fact is that the list is unusual and contains many other objects, which in turn can contain other objects. The $ character is used by R to indicate which object or sub-object we are interested in. Let's look at an example. Let's assign the result of the above analysis to the variable bla, like this:

```
> bla<-epi.ccc(Bill,Bull)
```

Now we can ask the R value of CCC

```
> bla$rho.c
        est     lower     upper
1 0.8490153 0.6655893 0.9357144
```

the result is a list of three elements. If we want only the lower limit of the confidence interval we can do this:

```
> bla$rho.c$lower
[1] 0.6655893
```

The CCC is calculated simply by multiplying the value of r by a correction factor C_b which is also called correction for bias . For those of you who love DIY, below is the R function for the calculation of the CCC, which I wrote based on formulas from Marubini (2005) [13].

```
cf.lin<-function(a,b){
# Lin CCC (Concordance Correlation Coefficient)
n<-length(a)
ma<-mean(a)
mb<-mean(b)
sa<-sd(a)
sb<-sd(b)
# correlation coefficient
r<-cor(a,b)
# bias correction
c<-(2*sa*sb)/(sa^2+sb^2+(n/(n-1))*(ma-mb)^2)
cat("r=",r,"\n")
cat("CF=",c,"\n")
cat("CCC=",(c*r),"\n")
# CCC
}
```

You will have noticed that R can " learn " to do new calculations by defining what is called a function, with the function `function ()`. Don't be confused by the repetition; it is very simple. After executing the lines above in the console, everything that is in brackets will be run by R whenever we call up the function `cf.lin()`.It is only necessary to do this:

```
> cf.lin(Bill,Bull)
r= 1
CF= 0.8490153
CCC= 0.8490153
```

OK, so r is equal to 1 as we have already seen, but the CCC is about 0.85, a value that is fairly high, but different to 1.

Chapter 9
ANOVA
(more extra soft toffee)

One day Bortolo, the other assistant of Gervase, perhaps because he was a little jealous, insisted that the ideal ingredient to improve the melting of toffee was sarsaparilla. So, no sooner said than done, Gervase prepared another 10 sweets with Bortolo's recipe. Here are the values:
B = 79 52 80 68 61 68 74 71 76 73
Now we have a problem: what are we going to compare B with?
• with G?
• with A?
• is it the same, given that we have already "proved" that there is no difference between the two of them?
But in fact we have not demonstrated that there is no difference. We have established that we shouldn't reject the null hypothesis that there is no difference between them. Fortunately, there is a beautiful statistical instrument that seems to have been made to get us out of this situation. It can help us to answer questions such as these (even when the problems are more serious or more complex). It is the *analysis of variance*, known as ANOVA to its friends.

Before we continue our discussion of ANOVA, I must introduce a new concept to you: the concept of *vectors*.
It is simple; you take some numbers put them next to each other and that is a vector. So B is a vector, as are G and A. In general, vectors are written in bold so let's write **B** is a vector, as are **G** and **A**. In this way it is written correctly and the fussy side of me is happy.
Maybe some of you have already heard of vectors when studying forces in physics. In that case it may have been explained to you that vectors are like arrows with a length, a direction and an angle. There is no contradiction between the two definitions: if you draw an arrow on a system of axes, putting the end of the arrow at the

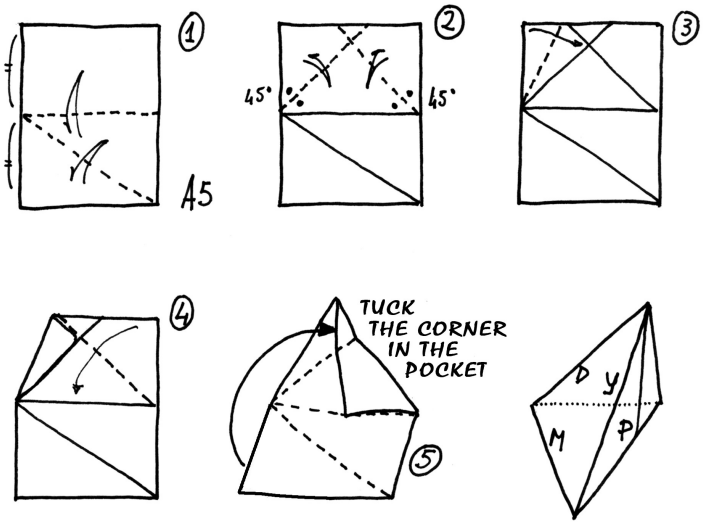
point 0 0, then the tip of the vector arrow will be at a certain point xy; for example, x = 13, y = 78. These two numbers together form the vector
13 78
If the arrow was in three-dimensional space its vector would have 3 elements, and would be made up of three numbers (x, y and z). So a vector with five elements like an arrow in 5 dimensional space. What is that you said? Five dimensional space doesn't exist. Well, nothing prevents us from imagining space with 5, 7 or 256 dimensions. Imagination costs nothing!

Let's take the 3 vectors **G , A** and **B** and "put them together" to make a longer vector which we will call **Y**.
Y = 89 71 76 81 75 79 60 62 70 61 72 82 65 83 50 61 83 68 52 75 79 52 80 68 61 68 74 71 76 73

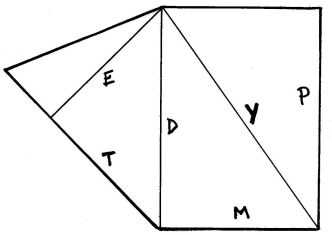
Now all that is needed is to construct 30 dimensional space with paper and to put our vector in it. I have to confess that I am unable to do origami in 30 dimensions. But it is not a problem; have you never seen a map? What does a map have to do with this? Well, a map is an example of representing something that is three dimensional in 2 dimensions. Here, all we need to do is to build a three-dimensional representation (3d) of something with 30 dimensions (**Y**).

Let's start with an A5 sheet of paper and do the fold in the next page.
Let's look for a moment at the fold; it is a triangular pyramid and all of its faces are right angled triangles.
The edges of the pyramid have the proportions $1, \sqrt{(2)}, \sqrt{(3)}$.

Six pyramids like this form a cube, or rather, with three pyramids like this plus 3 other pyramids with the mirror form we can make a cube.



For convenience, it is best to mark the edges of the pyramid with letters, so reopen the model to stage 5, then to turn it so that the mountain folds are at the top. Mark the folds which represent the edges of the pyramid, following the pattern shown here:

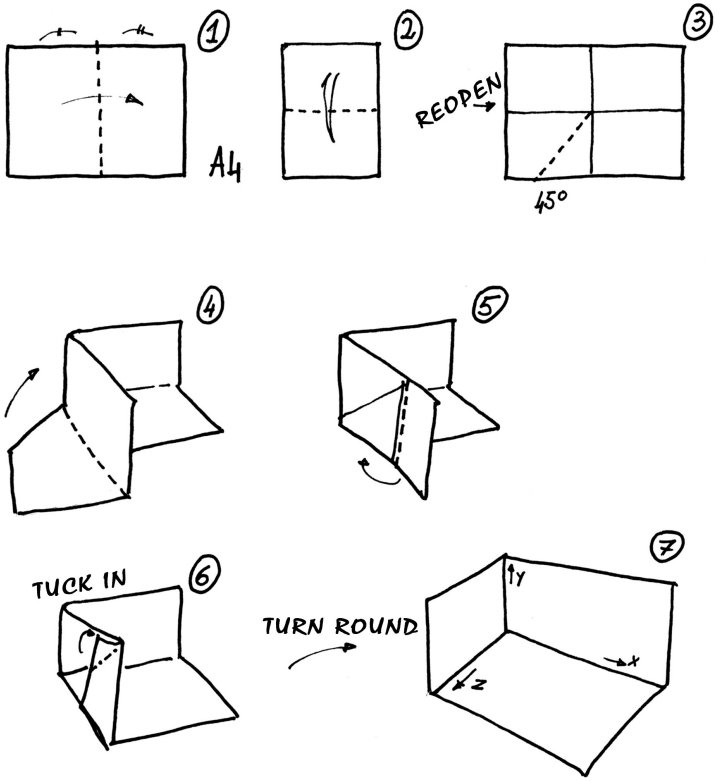


Now reconstruct the pyramid and for a moment look only at the triangle formed by the edges Y, D and M. Let's imagine that the edge that we marked with the letter " Y " represents our vector **Y** . To be more precise, we

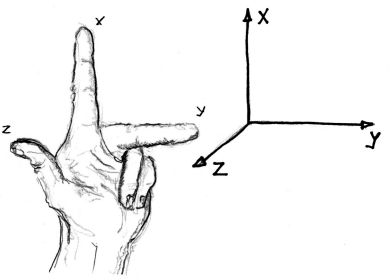
should say a 3d projection of vector **Y**, but no doubt you understood anyway.

Now imagine **Y** in space and have a go at moving it in cartesian space. Put the corner of edges Y and M at the origin, while the other end of Y (where it joins D) can be moved where you want.

You can construct a model of cartesian space with a sheet of A4 paper and can even write on the names of the three axes x, y and z.



Or you can use the thumb and two fingers of your hand, as shown.
Be careful, vectors can also have negative values, so if you are using a paper model of cartesian space you have to imagine that the pyramid can also penetrate the planes xy, yz and xz formed by the paper. In our case, the values of **Y** are all positive, being melting times.



We said that the end of **Y** can be in any position in space. Let me explain this a bit better. Imagine that Gervase is giving Tony experimental values one at a time, the vector being in 30 dimensional space. In the world of fantasy 30d space exists, but don't ask me if it is made of paper, plywood or marzipan because I don't know. Whatever it is made of, it is clear that until Gervase has given Tony all of the 30 values, he doesn't know where to put the arrow – the vector. Every value specifies where to place it with regards to a certain axis (a certain dimension) and only when all of **Y** is known can the vector be positioned accurately. It can be said that Y has the freedom to be anywhere in space, so in a space made up of n dimensions it has n *degrees of freedom*.
Now imagine that the edge **M** represents the average of **Y**. But the average is a single number so how is it represented in 30d space? Simply like this:
70,6 70,6 70,6 70,6 70,6 70,6 70,6 70,6…….thirty times.
30 times the same number. And where do you put a vector, with any number of dimensions, all made up of the same number? It has to be on a straight line which passes through, and which is

equidistant from all axes. If there are 2 axes we are on a flat plane (2d) and **M** is on the line that bisects the angle between the x-axis and the y-axis. If we are in 3d, **M** is on the diagonal of a cube that has a vertex at the origin, and so on. In 30d **M** is on the diagonal of a hypercube in 30d. So, whatever the number of dimensions, **M** can only move along a line through the origin, in one dimension. In fact, as soon as Gervase told Tony one of the values of **M**, Tony knew where to put the vector because he knew that all the other 29 values were the same. For this reason M has always only one degree of freedom. I'm sure that at this point you are dying to know what D is. Wait a moment; first I have to tell you a couple of things about vectors.
Vectors have some peculiarities that affect the way we do the mathematical calculations.

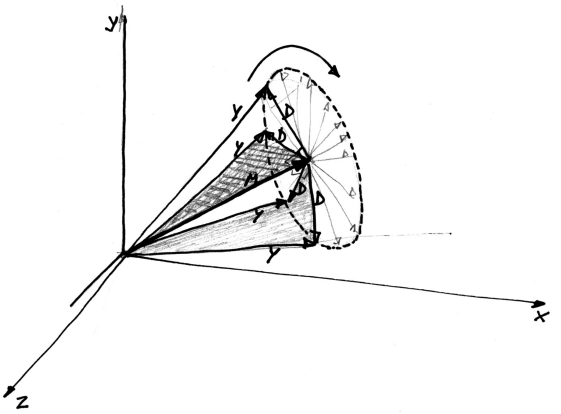
The value of a vector is obtained by summing the squares of all of its elements (if you think about a vector in a 2d plane and the Pythagorean theorem this will be obvious)

The addition and subtraction of two vectors is done by adding and subtracting the corresponding elements of two vectors (we did this in chapter 2 without actually knowing that we were doing calculations with vectors; we were smarter than we thought we were). Otherwise, if the vectors are orthogonal, the sum of the vectors can be obtained simply by drawing the vector that joins the two ends of the vectors (the tips of the two arrows) *.

Therefore,
Y = **M** + **D** then **D** = **Y** – **M**. That is, **D** is the difference between the individual observations and the average.

* Maybe some of you will remember that to add two vector arrows the so-called parallelogram rule can be used. It takes just a moment of thought to realize that what is set out here is the same thing if we remember that a rectangle is a special parallelogram where the two diagonals are the same length.

The value of **D** (the sum of the deviations from the mean squared) is our old friend, the deviance.



Now **M** is on the bisector of the angle between the coordinate axes, so let's keep it there as shown in the diagram; How can **D** move? It can only rotate around **M** like a wheel around an axis. It can only move in a plane which is perpendicular to **M,** that is, in 2d space, and with one degree of freedom less than we started with. In general terms it has n-1 degrees of freedom. I promised you that I would get there and here we are. This is the reason why when you have to calculate the variance you have to divide by n – 1. The reason is simple: to calculate the mean (which we need to calculate the deviations from the mean) we have already taken a degree of freedom and now we have only n-1 to calculate the deviance (and thus the variance and the standard deviation). In fact, if Gervase had said to Tony the values of D, as soon as he reached the penultimate value, Tony, a skilled mathematician, would have stopped him and said:

" Do you want to bet that I can guess the last value?"

" It is easy. I know that the sum of all of the values of D (if they are not squared) is zero (remember this from chapter 2), so I just need to add up all the values that you have said and see how far it is from zero"

In general, if I know the mean of a sample and I know n-1 values I can derive mathematically the n-th value, so this n-th value is not

free to assume any value that it wants: the mean has "eaten" its degree of freedom. This happens because, not knowing the true mean, we are forced to use an estimate, the sample mean to estimate the variance.

But let's go back to our pyramid ; Now let's consider the triangle **Y**, **P** and **E**.
P stands for Prediction: it is the vector with the most probable values of the 3 vectors **G** , **A** and **B**. The mean is the best estimate (as we said), so it is also the most probable value, so **P** is made up of the means of **G** , **A** and **B.** Here it is:

72,4 72,4 72,4 …[10 times] 69,1 69,1 69,1 …[10 times] 70,2 70,2 70,2…[10 times]

Y minus **P** gives us a vector with the deviation from the prediction. It shows us how measurements vary as a result of what happens *within* the 3 groups. It is sometimes called the error **E** because it indicates the error in our estimates.

The triangle below that, bounded by **P** , **M** and **T,** tells us that subtracting the **M**ean from the **P**rediction gives us **T**, that is the vector with the contributions of each new recipe to the melting characteristics of the sweets. Usually it is called the effect of the **T**reatment, or the variation *between* the groups.

So our pyramid shows us how we can break down the total deviance (and thus the total variance) into a deviation due to the effect of the treatment **T** and a deviation due to error, the random component **E**.
All we need to do is to look at the triangle **DTE.**

The clever idea of Sir Ronald A. Fisher (1890- 1962) was to calculate the distribution of the statistical test obtained by dividing the value of **T** by the value of **E**. In fact, if the contribution of the

new recipe is as large as the random component it is logical to assume that the new recipe does not add anything to the melting characteristics of the sweets, while if **T** is much bigger than **E** we can expect to have found something interesting. In fact before calculating the ratio between the value of **T** and the value of **E** both of these numbers need to be divided by the appropriate degrees of freedom, because it is obvious that the number of observations in each group and the number of groups is important, but it doesn't change the underlying logic.
This test takes the name analysis of variance or ANalysis Of VAriance. Almost always the distribution is called F for Fisher or the Snedecor-Fisher distribution because Snedecor proposed some improvements to Fisher's original method.

Together with R let's look at how things work in practice. First of all we put together the vectors that we need. The measurements are put all together in a single vector called **Y**.

```
> Y<- c( 89, 71, 76, 81, 75, 79, 60, 62, 70, 61,
+  72, 82, 65, 83, 50, 61, 83, 68, 52, 75, 79, 52,
+  80, 68, 61, 68, 74, 71, 76, 73)
```

Then you have to prepare a vector that " explains " to R who each measurement belongs to. We will call it **A**, the author of each measurement.

```
A<-rep(c("0G","1A","2B"),each=10)
```

As you probably guessed the function `rep()` helps us whenever we need to produce vectors that are simply repeating something. The use is pretty obvious, however, let me just remind you that for more information all you need to do is to write `?Rep`.
Unfortunately, R has a bad habit (nobody is perfect) in some cases (when handling factors) of rearranging the names in alphabetical order. For this reason I chose to put a number 0, 1, 2 before the letter for each author's measurements.

In any case, if we put **A** and **Y** side by side it is clear what is in **A.**

```
> cbind(Y,A)
      Y    A
 [1,] "89" "0G"
 [2,] "71" "0G"
 [3,] "76" "0G"
 [4,] "81" "0G"
 [5,] "75" "0G"
 [6,] "79" "0G"
 [7,] "60" "0G"
 [8,] "62" "0G"
 [9,] "70" "0G"
[10,] "61" "0G"
[11,] "72" "1A"
[12,] "82" "1A"
[13,] "65" "1A"
[14,] "83" "1A"
[15,] "50" "1A"
[16,] "61" "1A"
[17,] "83" "1A"
[18,] "68" "1A"
[19,] "52" "1A"
[20,] "75" "1A"
[21,] "79" "2B"
[22,] "52" "2B"
[23,] "80" "2B"
[24,] "68" "2B"
[25,] "61" "2B"
[26,] "68" "2B"
[27,] "74" "2B"
[28,] "71" "2B"
[29,] "76" "2B"
[30,] "73" "2B"
```

To carry out an ANOVA R needs the classificatory variables (like A in our case) to be transformed into something called a *factor*. We don't have the space here to explain in detail what a factor is, so just accept that `A<-as.factor(A)` does what we need.

Now, to do the analysis of variance we use a function called `aov()`. The function has an unusual syntax that makes use of the special

character ~ . Often this character is not present on the keyboard but can be obtained (with a Windows PC) by pressing the Alt key and typing the number 126 on the numeric keypad. Alternatively, refer to the computer manual. The result of the function aov ( ) is slightly strange. It is a *list*, which is something that we have already come across that can be saved in a variable enabling us to do many interesting things which I can't go into now. What interests us is what in the jargon is called the *ANOVA table* which you get simply by applying the function `summary()` to the result of the function `aov()`. This is written below.

```
> summary(aov(Y~A))
            Df Sum Sq Mean Sq F value Pr(>F)
A            2   56.5   28.23   0.272  0.764
Residuals   27 2804.9  103.89
```

So, we have a table with two rows: the first is the effect of **A**, which in the origami example we had generically called **T.** The second row is the residual which is also called the error **E**.

`Df`	are the degrees of freedom
`Sum Sq`	is the sum of the squares (the deviance)
`Mean Sq`	is the variance (Sum Sq / Df)
`F value`	is the Fisher F (ratio of the 2 variances)
`Pr(>F)`	the probability of being wrong by saying that **A**, **G** and **B** are different.

In our example the figure is about 76%. So there is not a significant difference between the 3 recipes. In fact, R has already calculated everything, taking away all the excitement, but we can try to follow the calculations with our pyramid. Given **Y** and **A** as set out above, it is easy to generate **M**.

```
M<-rep(mean(Y),30)
```

To get P we use the function `tapply()` which applies a function (in this case the average) to all subgroups that can be found in **Y**

according to the "rule" **A**. It is very powerful and very useful , I recommend studying it quietly. Guess how? Just write `?tapply` .

```
P<-tapply(Y,A,mean)
```

and it does all the calculations we need. Then

```
P<-rep(P,each=10)
```

which repeats each value 10 times (our vectors all live in 30 dimensional space, remember!). Then, as we have already seen, we can write some simple calculations, one for each face of the pyramid:

```
D<-Y-M
T<-P-M
E<-Y-P
```

but also

```
E<-D-T
```

for the picky, we can ask R to check that the two ways of calculating **E** are equivalent. It is a bit like running up and then down a pyramid.

```
(P-M)==(D-T)
```

But then

```
> sum(T^2)
[1] 56.46667
```

and

```
> sum(E^2)
[1] 2804.9
```

or

```
> sum(D^2)-sum(T^2)
[1] 2804.9
```

Please compare them with the values in the ANOVA table.

At this point we have to calculate the degrees of freedom. You have probably already noticed that as you add up the vectors, so you add up the degrees of freedom:

vector	degrees of freedom	in our case
Y	n (number of observations)	30
M	1	1
D	n-1	29
P	k (number of treatments)	3
T	k-1	2
E	(n-1)-(k-1)	27

Then the value of F is obtained from the following:

```
> (56.46667/2)/(2804.9/27)
[1] 0.2717744
```

Again, please check the results in the ANOVA table .
As already mentioned, the calculation of the probability of finding a given value of F by chance (an issue which applies to all the probability distributions) is rather complicated (it requires the solution of an integral) , so we will ask R to help us using the function `pf()`

```
> pf(0.2718,2,27,lower.tail=FALSE)
[1] 0.7640668
```

The function takes as inputs the value of F and the two values of the degrees of freedom. I hope you remember the meaning of the option `lower.tail=FALSE` from chapter 4. In this case it is equivalent to

```
> 1-pf(0.2718,2,27)
[1] 0.7640668
```

And the extra-soft toffee?
Oh yes, I forgot. Gervase discovered it by chance.

Chapter 10
Something on Regression

You may remember that in the chapter 3 I pointed out the possibility of looking for an association between the measurements of Nick Robinson's little dog.
Below is a drawing of one of the dogs " unfolded".
Let's consider the measurements Si and Tg.
I hope you kept the dogs that we used to calculate the median and quartiles. If so, reopen them and measure Si and Tg for each sheet, thus creating two vectors **Si** and **Tg**.

Now we can ask R to calculate the correlation coefficient (r), to see if the two measurements are associated with each other.
We have already said that the correlation coefficient tends to 1 when the variables are positively correlated (if one increases, the other also increases), tends to -1 when the correlation is strong but negative (if one increases the other decreases), and tends to 0 when there is no correlation.

I've already taken the measurements of some dogs, that were made using sheets of paper that were 95mm square. They are in the table.
This time I will use a different method to pass the data to R. So far we have built vectors in the workspace of R with the command `<-`.
This is not a very convenient way to do it, especially when there is a lot of data. So I used a spreadsheet to do the data entry which I saved in the tab-delimited "text" format, calling it dogs.txt.

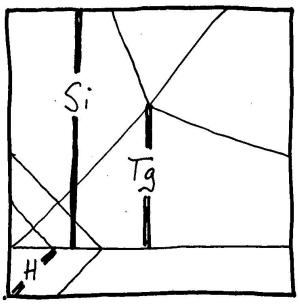
At this point you have to make sure that the file is in the same folder in which you started R, or to determine that the one where the file is located is R's "working directory". How do we do that?
To define the "working directory" (wd) there is the function `setwd()` or you can use the command Change dir ... in the File menu.

For beginners , I think it is easier to use the function:

```
> getwd()
```

with which we find out what the current working directory is. Then we can just copy the file into that directory (folder).

N	Si	Tg	H
1	73	52	31
2	71	52	26
3	79	60	26
4	78	59	22
5	87	50	12
6	70	55	36
7	76	57	26
8	79	57	21
9	76	57	25
10	79	56	23
11	73	52	32
12	81	57	13
13	80	55	20
14	71	54	34
15	70	51	34



Having resolved the problem of the wd we run the command:

```
> dogs<-read.delim("dogs.txt")
```

This creates a special object in the workspace called a *dataframe*, (in our case it is named dogs) that contains all of our data, rather like a large table. At this point we can directly use the four variables in the table by referring to them with their names:

```
dogs$N
dogs$Si
dog$Tg
dogs$H
```

Si and **Tg** are the two columns that interest us, **N** is simply the number of the dog, and we will talk about **H** shortly.

The command

```
> attach(dogs)
```

allows us to avoid writing the prefix `dogs$` every time.

Now, R uses the function `cor()` to calculate r. Forgive the foolishness: we are using R to calculate r (it is an old joke – about 2000 years old [14]). Anyway, here it is:

```
> cor(Si,Tg)
[1] 0.2395227
```

The value of the correlation coefficient is about 0:24, so the two measurements are associated, but fairly weakly. This is logical; there is no reason why those who do the first fold in a certain way should then do the fifth fold in a way that depends on the first fold. They are two "free" steps as we have already said, when speaking about the fold in the chapter 3.*

But now let's try to measure the length H of our little paper dog, which is a measurement to do with the length of the tail of each

* In fact this is not quite true . I realized that there is some sort of relationship between Si and Tg. In particular, if all the dogs are folded by the same person, sometimes the aesthetic sense of the person leads him to choose a size and head shape that is not completely independent of the height of the dog. If you look at the plot of Tg and Si at page 92 it is clear that there is a single point with a very high Si and a very low Tg, whereas all of the other points tend to correlate (weakly) in a positive way .

dog. As you can see, I put them into the spreadsheet and they are already in our data frame, so it is easy to calculate the correlation.

```
> cor(Si,H)
[1] -0.9185351
```

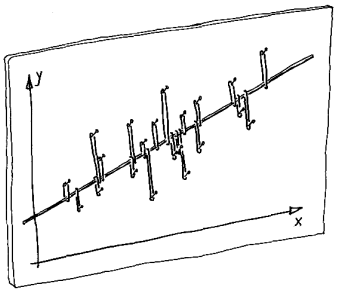
Here, however, there seems to be an association: -0.9 is fairly close to 1. Perhaps some of you have also understood why... But, wait a minute; tell me later.

The fact is that, if the measures are associated, we can ask R to look for a kind of "rule of association". In general, we are talking about a *model* and the easiest model to find is the linear model: the well-known function of a straight line.

$$y = a + bx$$

Well, I have written "well-known" to imitate some of the serious mathematics textbooks. Maybe some of you are familiar with the branch of mathematics which combines geometry and arithmetic in a wonderfully elegant way. Maybe you could have some fun going through Appendix C (don't forget to have some pieces of paper handy) .

Anyway:
a is the point where the line crosses the y-axis
b is the slope of the line.



Given a set of x and y values which are correlated, it is not difficult to estimate a and b. Get a wooden board , draw a Cartesian plane on it and put a nail where each pair of x and y values intersect. Then hang a rubber band on each nail. Then get a nice straight stick and thread it through all the

bands, let go... and you're done! The stick represents the function of the straight line [10] .

What is that you are saying? It is a bit complicated. Well then we'll have to ask R for some help, using the function `lm()`. This function, like `aov()`,produces a result that becomes more interesting if we add in the function `summary()`.

Here it is:

```
> summary(lm(Si~H))

Call:
lm(formula = Si ~ H)

Residuals:
    Min      1Q  Median      3Q     Max
-4.8295 -0.3732  0.1705  0.6705  3.1705

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 91.88336    1.94201  47.314 6.13e-16 ***
H           -0.61746    0.07371  -8.377 1.35e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
1

Residual standard error: 1.999 on 13 degrees of freedom
Multiple R-squared:  0.8437,    Adjusted R-squared:  0.8317
F-statistic: 70.18 on 1 and 13 DF,  p-value: 1.346e-06
```

What a lot of stuff! Don't worry; the column " Estimate" contains the values that we are interested in.

$$Si \approx 91.9 - 0.6H$$

This is called the calculation (*estimate*) of the *regression* of y on x .

Let's stop to think for a moment. I seem to remember that someone told me that it is logical that Si and H are related. In fact, looking at

the piece of paper with the folds, we can see that Si is equal to L (the side of the square we started with), less H divided by the square root of 2. Well done!
But then, given that the side of the square and the square root of 2 do not change we should be able to derive the parameters a and b. In fact, all we need is a little algebra. We said that:

$$Si = L - H \div \sqrt{2}$$

Therefore $a = L$ and $b = -1 \div \sqrt{(2)}$

Let's check

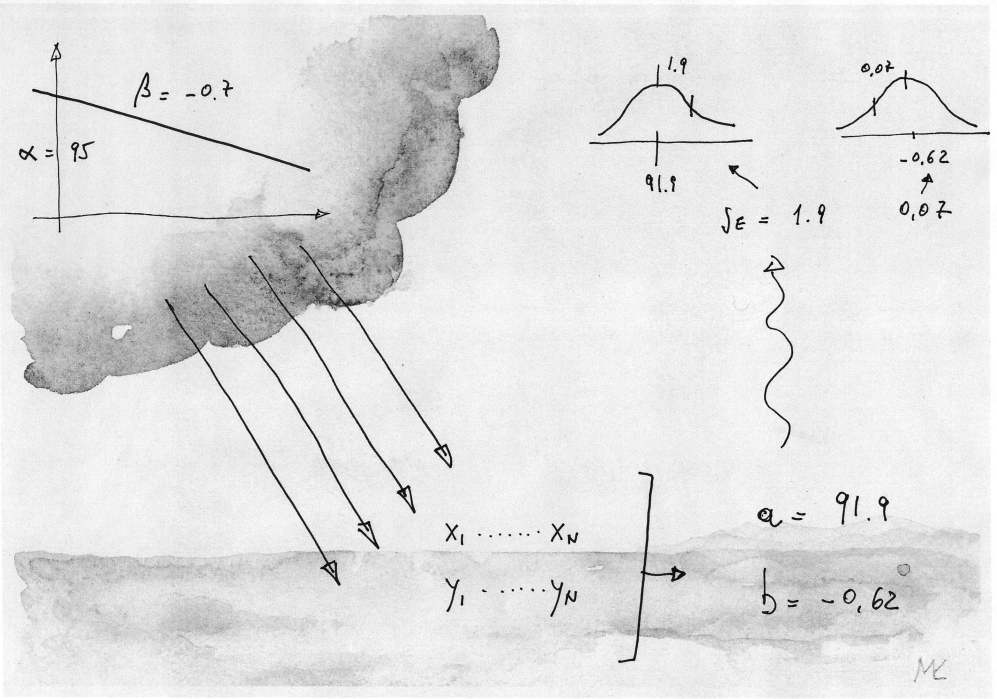
In my case *a*=L=95 whereas we got 91.9

b should be

```
> -1/sqrt(2)
[1] -0.7071068
```

whereas we got – 0.6
This is due to the mistakes I made when measuring with a ruler, the imprecision in making the folds and to rounding errors. If you were better than me with the folds and the measurements maybe you did better, but if you got exactly L - 0.71 I know that you cheated.

Be careful: this is a very special case, created specifically for educational purposes. In fact, here we are able to calculate the real model that links Si with H. But this is because the dogs were "created" by us, following geometrical rules. In biometrics this hardly ever happens (unless you have God in your team). Typically, all we can do, is to use regression to try to find the best *estimate* of the model, namely the line, without ever knowing exactly the real nature of the things.



Let me explain this a little better. Remember the discussion of the cloud (chapter 6). Here is the same thing, except that in the cloud there are a and b, the parameters of a straight line that, in a case that is more unique than rare, we know exactly:
$a =$ 95 mm, whereas $b = -1 \div \sqrt{(2)} = -0.71$

And indeed, defining them as a and b showed a lack of precision on my part. I should have used the corresponding Greek letters (α and β).

We "sampled" this universe by building the dogs and taking measurements, and got the two vectors **X** and **Y** (which in this case

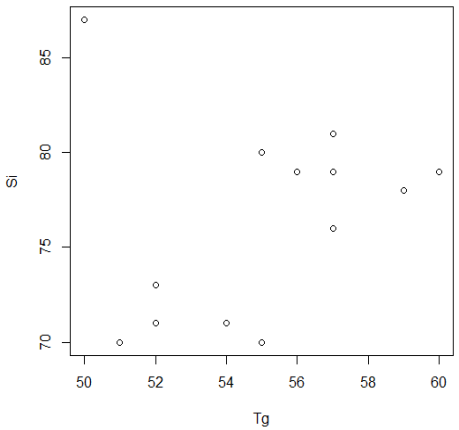
we called **H** and **Si**). We put this data into a machine (our friend R) which provided *estimates* for α and β , which we called *a* and *b*. In fact, *a* and *b* are in the column called Estimate. But in the adjacent column is our old friend: the standard error. Now given what we know about the Gaussian (Chapter 4) it is very rare that the two "real" parameters (α and β) and a and b are separated by more than 2 times the standard error. Indeed

```
> 91.88336+2*1.94201
[1] 95.76738

> -0.61746 - 0.07371*2
[1] -0.76488
```

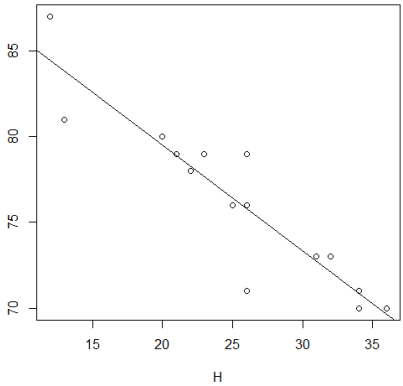
It works!

All this is even more fun if we represent it graphically which is simple with R.

```
> plot(Tg,Si)
```



```
> plot(H,Si)
> abline(lm(Si~H))
```



Let me draw your attention to the second graph where there is an interesting thing. R drew the line that has our parameters a and b. As you can see, to do this you just need the function `abline()` and to give the result of the function `lm()` as an input.
The logical thing would be to save the result of `lm()` somewhere and use the saved object as an argument for `abline()`. But , out of laziness, I prefer to get it to recalculate `lm()` each time. Well it is R that does all the hard work!

Forgive me but, at this point, I have to reiterate another apparently tedious thing. We started the chapter by talking about correlation and we got on to talking about regression. Indeed often these things go hand in hand so to speak, but , please, do not confuse them with each other. They are two different instruments, with different results. Correlation measures the association between two variables, whereas regression seeks to estimate a model that links them, irrespective of their degree of association.

Now, the method that we used is called the *least squares method* because it calculates the parameters of the particular line in which the sum of the squares of the distances to all of our points is the minimum possible. If you think about it a little it is the same thing that the rubber bands did on p. 88, by trying to be as short as possible. Only one thing must be emphasised: the elastic bands should all be vertical, because what we want to minimize is the distance on the y axis.

A bit of jargon. Note that y (in our case Si) is called the *dependent variable*, because we really hope that it depends on H which doubles as the tail of the dog, which we call the independent variable.
There is always a reason for Jargon. In this case we are working with pieces of paper and it is not very important, but using our imagination and thinking of Canis Origamicus, it is not very logical to think that the height of a dog depends on the length of its tail. Perhaps it would be better to say that it is the tail that depends on the height of the dog? I don't know, I don't really know anything about dogs (and those made of paper...), but I would just like to draw your attention to the fact that, in defining the model, it is worth taking a moment to think about what to put "on x" and what to put "on y". Things cannot be interchangeable and the result will not; this is an other important difference between regression and correlation (try it and see).

Perhaps some of you are dissatisfied because I haven't explained to you how R gets a and b.
This may be explained in different ways. Let's return to our friends the vectors.
Whereas the sum of two vectors is not unlike the sum of two numbers, the product of, and the division of, two vectors requires mathematical operations that are a bit more complex, but which give very interesting results.
Now, before continuing, I have to introduce a new character: *the matrix* . Whereas a vector is a set of numbers placed in a row, a matrix is a series of vectors "lined up", that is a table of numbers arranged in rows and columns (maybe in pages, hyper cubes etc. A matrix can also have more than 2 dimensions).

Let's build a matrix **X** in this way:

```
> X
          H
 [1,] 1 31
 [2,] 1 26
 [3,] 1 26
 [4,] 1 22
 [5,] 1 12
 [6,] 1 36
 [7,] 1 26
 [8,] 1 21
 [9,] 1 25
[10,] 1 23
[11,] 1 32
[12,] 1 13
[13,] 1 20
[14,] 1 34
[15,] 1 34
```

with many ones in the first column and the values of **H** in the second column. Then if then we "divide" the vector **Si** by the matrix **X** , we get a and b. I told you that division with matrices is rather strange...
In fact it is a bit like solving a group of equations, one for each row of the matrix.

With R it works like this:

```
> X<-cbind(1,H)

> qr.solve(X,Si)
                      H
91.8833605 -0.6174551
```

Not all software is capable of dividing two matrices (or a vector and a matrix), so it is possible that in some books you will find a slightly different procedure. First calculate the inverse of the product of the matrix **X** and its transpose; then multiply this matrix

(the product of the transpose of **X**) and **Y**. I know, it's a bit like a tongue twister, and I haven't even explained what the transpose of a matrix is. But I can translate it into the language of R, in that for the product of two matrices you use the function `% * %` . For the transposition of a matrix (which is a bit like rotating the matrix on its main diagonal) the function `t()` can be used. The function `solve()` with a single input can be used to calculate the inverse of a matrix.

```
> (solve(t(X) %*% X)) %*% (t(X) %*% Si)
         [,1]
  91.8833605
H -0.6174551
```

To calculate an inverse matrix it is first necessary to calculate a thing called the *determinant* of the matrix. If you studied matrix algebra you will certainly have learnt to calculate a determinant and, just as certainly you will have wondered what the purpose was of all those calculations. Well here is a practical application of matrix algebra. If you didn't study matrix algebra it doesn't matter; as I told you generally all of the calculations can be left to the computer (R). Or there are formulas which do not use the matrices, which can be found in any book of statistics (see for example [8] in the bibliography).

Thank you for following all of this number crunching. Of course I do not claim to have explained exhaustively everything about regression, but perhaps to have explained what is useful, but above all I hope I have left you with the desire to study this really fascinating topic in greater depth.

Chapter 11. And to finish...
A true story

Some time ago the manager of a big Italian company told me that a number of years earlier he had wanted to optimize a certain stage of production using Stepwise Regression. With Stepwise Regression it is possible to choose from among many independent variables those that are most important to predict the behaviour of a dependent variable.

Now you should know that "a number of years" before "some time ago", although not exactly the time of the dinosaurs, was still a time when calculating instruments were not as easy to use as they are now. Computers were expensive objects weighing a few tons that had to stay in air-conditioned areas, and that had to be programmed with bundles of cardboard punch cards. In order to use a computer you had to be authorized to the required “machine time”, and above all it was not easy to find software suitable to do calculations. So my partner decided it would be more practical to do the calculations "by hand"; that is to say, using calculating machines (these were already in use, as were washing machines the internal combustion engine and bicycles). He organized two teams to work in parallel on the problem in hand.

After two weeks of working on calculations both teams reached a result, but the results were different!
Discouraged the manager decided that the plant was fine with the method they had used up to that point, and that Stepwise Regression could stay where it was in the statistics book.

Today it is quite easy to find a program to calculate Stepwise Regression. It can be run on any PC and the result is available in less than a second. However, with this huge capacity for calculation we sometimes run the risk of not having time to understand what the computer is doing. Here, in these few pages I did not intend to

convince you that folding paper was the most practical way to solve the problems of statistics, but I wanted to help you to become acquainted with some concepts of statistics. At the same time, I hoped to get you to feel at least a little of the fun I got from folding paper and making drawings.

If you managed to get to the end I thank you for your patience and for the attention that you gave me. I hope that some ideas will be evident for you,
I hope that some ideas will be clear to you, I recap them here:

- when measuring something not only is data position important, but also data dispersion
- quality of measurement is important (repeatability?)
- pay attention to the scales of graphs
- describing data is different from making inference
- when making inference remember the assumptions
- association is not a cause-effect relationship
- models are not reality
- it is important to make a distinction between independent events and events with conditional probability (see appendix B)

I take this opportunity to thank Laura Antolini, Guido Pacchetti, Piergiorgio Duca, Giorgio, Alfredo e Chiara Cigada, Carlo Alberto Spinicci; Mauro Sette and Remo Cacciafesta who gave me encouragement and valuable advice; Valeria Lovato who helped me with some drawings; and I am especially grateful to Anna Maria Viganò for graphics. But above all I have to thank my wife Flavia and my daughters Irene and Anna who have endured months with the house full of pieces of paper and were forced to listen countless times to ideas, phrases, thoughts and... were patient. Obviously I am responsible for any errors or inaccuracies and I apologize in advance for these.

Appendix A
For origamists

In this book I have limited myself to very simple paper models, on the assumption that the reader knows nothing about origami . It is possible that more experienced origamists are somewhat disappointed by some of the models which are a bit basic. So I thought of referring to some folds that are a little bit more complex, but that can be used in the same way as those presented in the book.

The late Thoki Yenn in the booklet mentioned in the bibliography [3], presented a fold to make the lopsided tetrahedron used in chapter 9 with a sheet of paper with the format 2×1 .

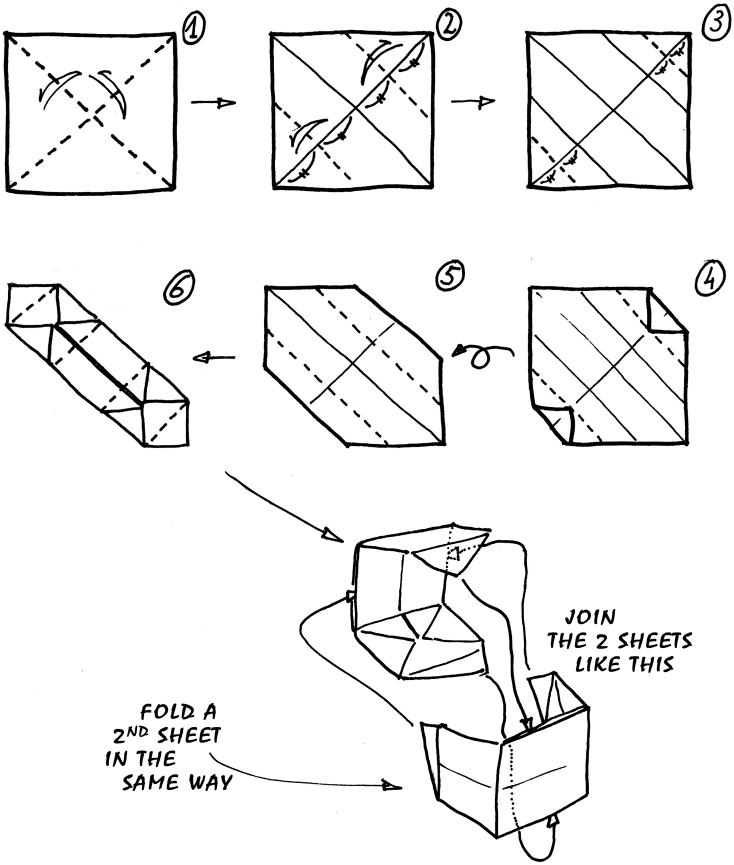
In this way, noted Thoki, a sixth of a cube comes from half a square!
His tetrahedron is more beautiful than mine because it has all of the faces (including the one at the bottom) .
Thoki was an extraordinary character. After his death the British Origami Society decided to host on its website the pages of the web site of Thoki, which risked being dismantled. The site also contains a scan of the quoted text [3].

http://www.britishorigami.info/academic/thok/index.html

In Kasahara's book, that has already been mentioned [1], there is an explanation of different models for a cube with each side equal to $\sqrt{2} \div 4$ (like the masu), that are capable of being made with 2 sheets of paper. In this way you can make histograms that combine cubes and masu. Below there is a diagram.

If you like the origami (and you can understand italian) but don't know Centro Diffusione Origami, contact them immediately; They

have beautiful paper and lots of books, including foreign books, at a good price .
www.origami-cdo.it
If you don'understand italian language, don't worry, I'm shure that in your country there is a similar society.



Appendix B
Probability

In fact all of statistics derives from the theory of probability. Probably for this reason, most manuals on statistics start with a large section on this branch of mathematics.
The fact is that the theory of probability is far from intuitive, so in this booklet I thought you would be happy to avoid the problem up to this point. But, not wishing to adopt the tactics of an ostrich, I would now like to try to introduce something I wrote some time ago, to see if origami can help us understand something about probability.

Probability as an area

Billy Ball is a strange child. He spends his afternoons in the courtyard playing with a ball. He throws it up in the air and lets it fall to the ground, then picks it up and throws it into the air again. He is capable of spending hours and hours in the courtyard with his ball, he enjoys it so much.
One day, as were watching Billy play, Professor Mumble Numble pointed out a curious thing to me.
Look, he said, Billy does not have the strength to throw the ball over the surrounding walls, so every time the ball is thrown, sooner or later it falls in the courtyard which, by the way, is perfectly square. Now if, unbeknownst to Billy, I define an area within the square, I could bet with you how often the ball falls inside that area and how many times it doesn't. Of course, in the long run, the number of times that I win will just be a function of the relationship between the chosen area and the total area of the yard (or between the chosen area and its complement). Accordingly, any punter, including you, would only agree to bet based on the size of chosen area. For example, if the area was one-third of the size of the courtyard, you would only bet on the basis of 2 against 1. This is very similar to the subjectivist definition of probability. However,

please note that I could describe this series of events as a ratio of frequencies; how many times the ball falls into the chosen area divided by the total number of throws. This is analogous to the frequentist definition of probability. At this point, continues Mumble Numble, I ask myself if the Billy-courtyard-ball phenomena follows the axioms of Kolmogorov, in that it is possible to describe the probability as the ratio between areas. Perhaps this is a trivial thing, but nonetheless fun.

This brief chat with Professor Mumble Numble brought my thoughts back to my old love: origami.
The classic piece of origami paper is square (like Billy's courtyard), so I thought we could forget the courtyard, ball and Billy (maybe he could be taught another game!) and describe some aspects of the theory of probability by folding paper.

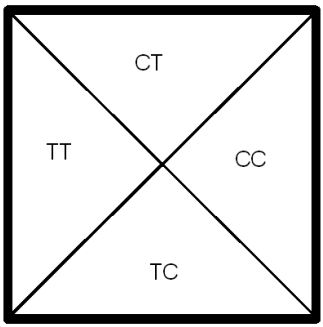
So what are the classical axioms of Kolmogorov?

In terms of probability it means a measure in *event space* with the following characteristics:
a) it is a positive number less than or equal to 1
b) the probability of a certain event is equal to 1
c) the probability that one or other of two events happens is equal to the sum of the two probabilities, if the two events are mutually exclusive.

In the world of origami for the probability of an event we mean the area of part of a square of origami paper.
a) the area is measured as the ratio (or percentage) of the total area of the square
b) the probability of a certain event corresponds to the whole square
c) the probabilities are added by adding the areas, provided that the areas do not overlap.

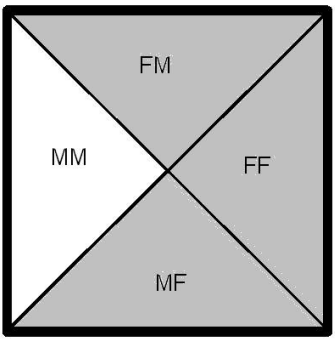
Let us look at some **examples.**
The toss of a coin is represented by a sheet with a fold in the middle. If the coin is not biased the 2 areas corresponding to "heads" (T in italian “Testa”) and "tails" (C in italian “Croce”) both have a value of 0.5. According to the standard notation:
P(T)=P(C)=0.5

If we flip the coin a second time we just have to make a second fold. The event space represents four possible and equally likely events with P=0.25. Please note that the order of the letters within the areas reflects the order of the tosses.



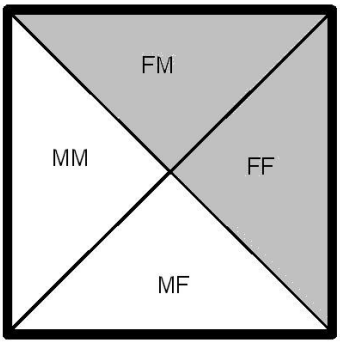
Up to this point it is pretty dull, but the last fold can help us unravel a classic problem of probability due to its apparently paradoxical nature: the problem of the two brothers.

I meet a lady who tells me that she has two children, one of which we know is female. What is the probability that the other is male?



If you answered 50% I am sorry, but you are wrong. Having 2 children is represented in the same way as two tosses of a coin. So take the sheet we used before and where there is a T put M for male and where it says C put F for female. Now we know that one of the children is a girl so the event space that interests us is the shaded area. There are 3 events of equal probability. It is easy to

see that in 2 portions out of 3 one of the children is male, so the result of the problem is 2/3 ≈ 0.66.



This is why the outline of the problem can be misleading. When we hear that the lady has a daughter we "infer" (incorrectly) that this is the first-born, in which case the event space becomes the one drawn here (the order of the letters reflects the order of birth of the children) and the probability that the second child is male, given that the first is a daughter, is 0.5 as you said before. With the standard notation P(M|F) = 0.5. It is counter intuitive simply due to a failure to understand the problem!

Another game: A probability value that is well known to all those who deal with statistics is P = 0.05, often for α errors (remember chapter 7). To see it lets start by folding a sheet of paper in half and then in half again, as in Figure 1.

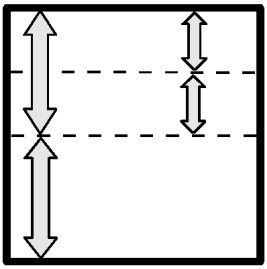


fig.1

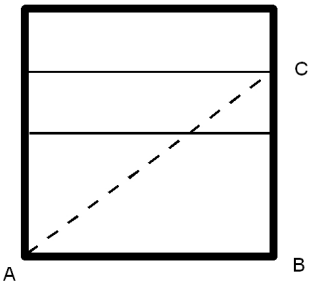


fig.2

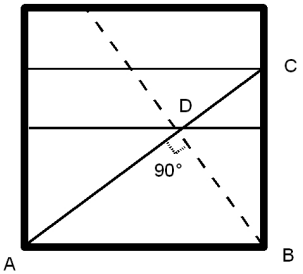


fig.3

Then we make a fold as shown in figure 2.
Now, the triangle ABC is a right angle triangle and, if you think for a moment, you will notice that the 2 sides are in the proportions 3 to 4, so the hypotenuse AC is equal to 5 according to the Pythagorean theorem.

Now we have to make a fold perpendicular to the fold AC (Figure 3) that meets vertex B. It is not difficult to make a 90° angle, just make sure , when you make the fold , that point C falls precisely along the fold AD. We get another triangle ADB. ADB and ABC are similar triangles (Do you know how to prove it? If not you can find out at the end of Appendix D), but ADB is smaller and the hypotenuse coincides with the side of the original square. So ¼ of the side AD is equal to 1/5 of the side of the original square (AB).

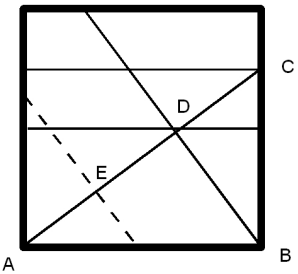


fig.4

Then we fold AD in half (Figure 4).

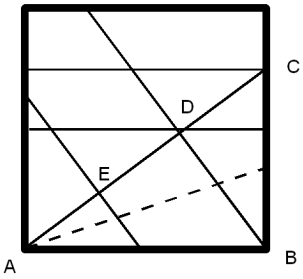


fig.5

then fold along the bisector of the angle CAB so that the side AB lines up with the line AC and point E is on the side of the square, as shown in figure 5.

To do this we have to fold the paper under at point E (in origami this is called a "mountain" fold and is indicated by a row of dots and lines, to distinguish it from a "valley" fold which is indicated by a series of dashes).

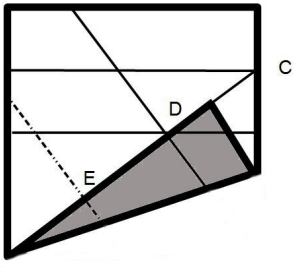


fig.6

The result should be as shown in figure 7.

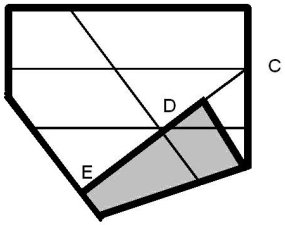


fig.7

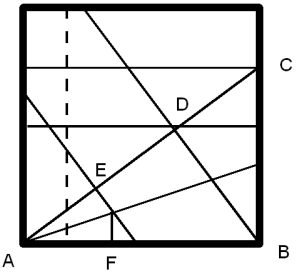


fig.8

Now unfold the sheet of paper. We positioned AE (= ½ AD) on AB creating AF=AE. Then ½ of AF = 1/5 of AB. Now let's do one last fold, taking point A to point F. The result is the shaded area which is our P = 0.05.

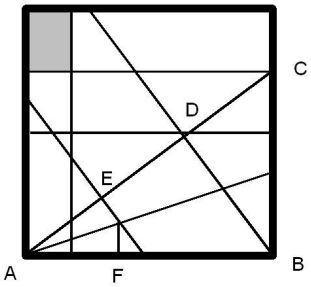


fig.9

$$\frac{1}{5}\times\frac{1}{4}=\frac{1}{20}=0.05$$

You say that you did not expect it to be so big? Neither did I, the first time, but I can assure you that it is correct. So remember that accepting an α error of 5% is not such a small risk.

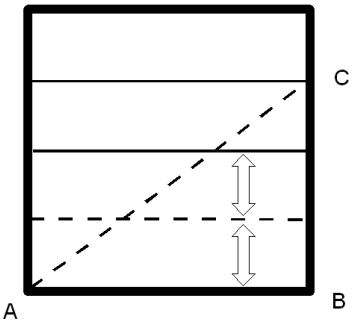


fig.2b

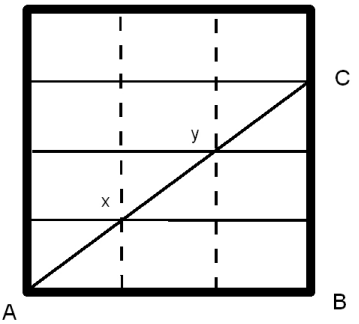


fig.3b

In the book "Origami Omnibus" referred to in note 2, Kasahara tells us that if we start from figure 2 of the previous diagram and add a fold as in figure 2b we get a very easy way to divide the side of a square into three. In fact, the two folds which pass through points X and Y as shown in fig 3b divide the square exactly into three.

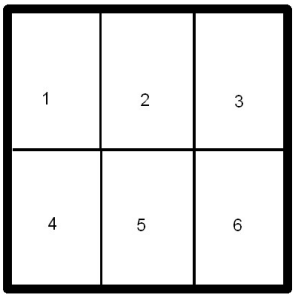


fig.4b

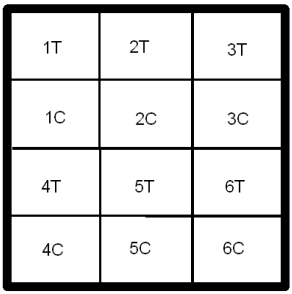


fig.5b

At this point, ignoring the construction folds and taking into consideration only the folds in figure 4b we have a square divided into 6 equal parts. For the reasons mentioned earlier this represents the event space of throwing a dice. If in addition to throwing the dice we toss a coin the event space becomes that shown in Figure

5b. This figure allows us to summarize the effects of different operations on probability.

The sum

This is equivalent to adding areas, and is analogous to the logical operator "or" and the union of sets. For example: what is the probability of having a tail (=C remember that in italian is “Croce”) or a 5?

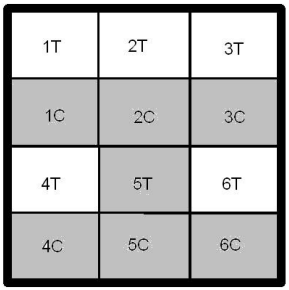


fig.5b

$$P(C \cup 5) = P(C \vee 5) = P(C) + P(5) - P(C \wedge 5)$$

Please note that if you do not take off P(C and 5), that is the probability of simultaneously having a tail and a 5, the area marked 5C in figure 5b will be counted twice.

So $P(C \cup 5) = \frac{1}{2} + \frac{1}{6} - \frac{1}{12} = \frac{7}{12} \approx 0.583$

The product

This is equivalent to adding folds (dividing areas). It is analogous to the logical operator "and" and the intersection of sets. For example: what is the likelihood of having a tail and simultaneously a 5?

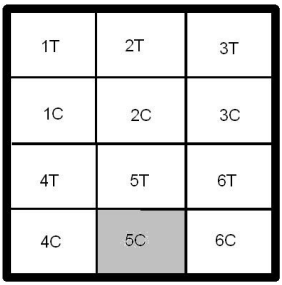


fig.5b

$$P(C5)=P(C\cap 5)=P(C\wedge 5)=P(C)\times P(5)$$

$$P(C\cap 5)=\frac{1}{2}\times\frac{1}{6}=\frac{1}{12}\approx 0.083$$

In simple terms you fold in half (½) the area of P(5).

Be careful. This applies only if the two events (coin and dice) are independent. In fact I could have written:
P (C5) = P (C) x P(5|C) which says that the probability of having a 5 and a tail simultaneously is equal to the probability of having a tail multiplied by the probability of having a 5 given a tail. But as the result of a coin toss does not depend in any way on the outcome of the dice then P (5) = P (5|C) and what we have written above is correct.

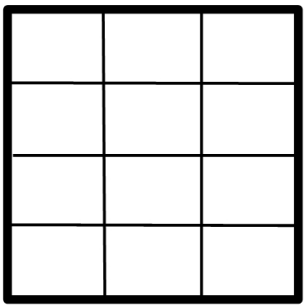
But this is not always the case..

So let's deal with a problem of conditional probability. In this case a useful tool is the so-called Bayes theorem, which can be written as:

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

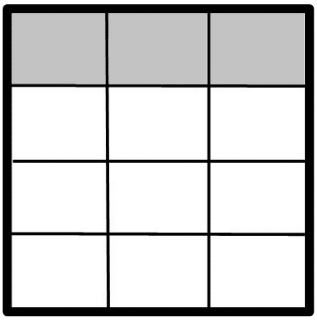
And is read: the probability of A given B is equal to the probability of B given A multiplied by the probability of A and divided by the probability of B.
I know this sounds far-fetched, but I want to see if our origami can help us to understand the meaning of this really important theorem.

Let's prepare a sheet folded into 12, as in Figure 5b, but leaving all of the boxes blank.

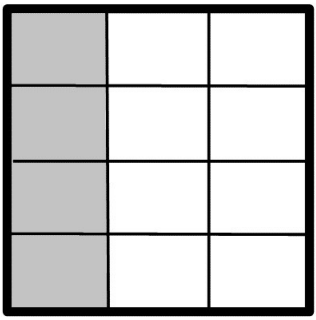


Now let's try to calculate P(A|B). Assuming that P(B|A) is equal to 1/4 and P(A) = 1/3 it is easy. I deliberately chose numbers that were "easy to calculate".

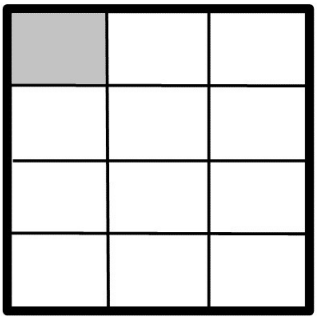
$$\frac{1}{4}\times\frac{1}{3}=\frac{1}{12}$$



P(B|A)

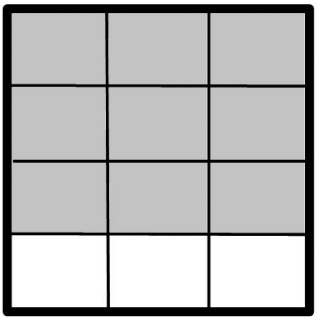


P(A)



P(B|A) x P(A)

This is the term above the fraction line. Now we have to divide by the probability of B. I hope you remember that to divide fractions it is necessary to multiply one by the inverse of the other. Let's make two hypotheses.



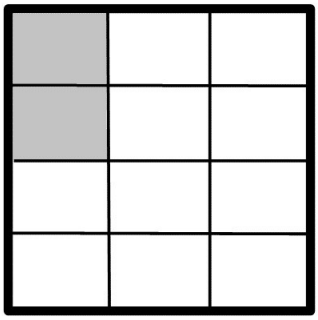
P(A|B)

If P(B) =1/9 we have:

$$P(A|B)=\frac{1}{12}\times\frac{9}{1}=\frac{3}{4}$$

But be careful. If event B was more probable, for example, P(B)=1/2,

we have: $P(A|B)=\frac{1}{12}\times\frac{2}{1}=\frac{1}{6}$



P(A|B)

This tells us that the probability of A given B rises with an increase in the probability of A and the probability of B given A, but drops with an increase in the probability of B.
I can imagine that some of you do not find this astonishing, so let me illustrate it with an anecdote.

During my youth a tragic social problem was heroin addiction, which I imagine you have heard of.
Then some well-meaning people, having noticed that the majority of heroin users were previously consumers of cannabis derivatives (hashish or marijuana), concluded that the "cause" of heroin addiction was linked to the consumption of cannabis, thereby demonstrating a delightful ignorance of Bayes' theorem.

The probability of becoming addicted to heroin after consuming cannabis = P (A|B).
It is certainly linked to the probability of finding a consumer of cannabis among heroin users = P (B|A)
But we have to multiply this probability by the incidence of heroin use = P (A)
And above all we have to divide the result by the incidence of cannabis use P (B)

So, if our sample is from an area where the use of cannabis is rare, then the observation that the consumption of cannabis predisposes someone to heroin use is significant, but if in our sample many people consume cannabis, then our conclusion (finding that many cannabis users among heroin users indicates that cannabis use "predisposes" someone to heroin use) is wrong.

Appendix C
Bits of analysis

Renee Descartes (1596 1650), also known as Cartesius, had the idea of looking for a link between arithmetic and geometry, in a simple and ingenious way.
If we draw two axes perpendicular to one another on a plane, and along each axis we put real numbers we get a fun way to represent the possible links between two variables. In this diagram, which traditionally goes by the name of a *Cartesian plane*, the horizontal axis is generally defined as the axis of the x variable (or *x-axis*) and the vertical axis is that of the y variable (or *y-axis*). The intersection between the two axes has the value 0.0 and is referred to as the *origin*. So the mathematical relationships between x and y come to be drawn rather elegantly on a plane.

It is very easy to represent linear relationships such as

$$y = a + b\,x$$

In this function a and b are the *parameters* of a straight line and, in particular:
a indicates the point at which the line crosses the y-axis
b indicates the slope of the line.

I think it is easier to give a few examples. In chapter 3 we produced a couple of graphs of variables that were related to each other. In the first example y was obtained by increasing x by 20% and adding 7. This corresponds to the function

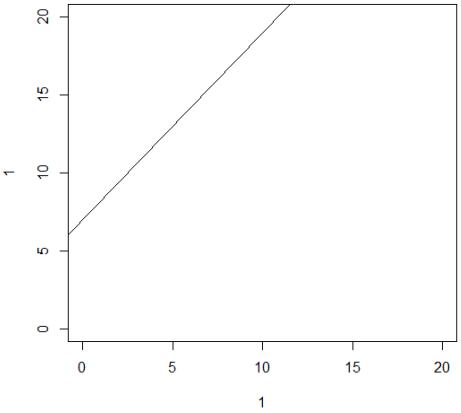
$$y = 1.2\mathrm{x} + 7$$

R can easily help us to represent it in the Cartesian plane with the command below. In fact the first line only serves to draw a "blank"

Cartesian plane ranging from 0 to 20 for both the x-axis and the y-axis, while the function is plotted by `abline()`.

```
plot(x=1,y=1,xlim=c(0,20),ylim=c(0,20),type="n")
abline(7,1.2)
```

The third example is even more interesting. Here we have added uncertainty to the function y, distributed according to a Gaussian with mean = 0 and standard deviation = 0.5.



So if we execute these two commands

```
abline(8,1.2,lty=2)
abline(6,1.2,lty=2)
```

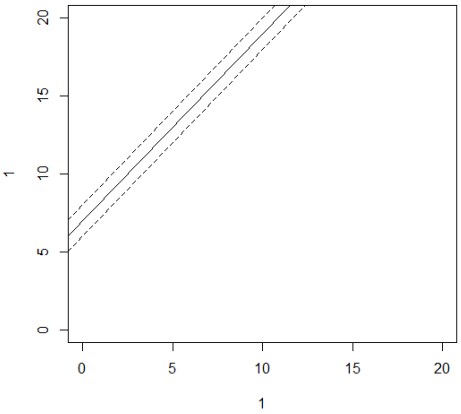
R adds to the chart 2 straight lines, with the same slope but with intercepts at:

$$7+0.5\times2=8$$

and

$$7-0.5\times2=6$$

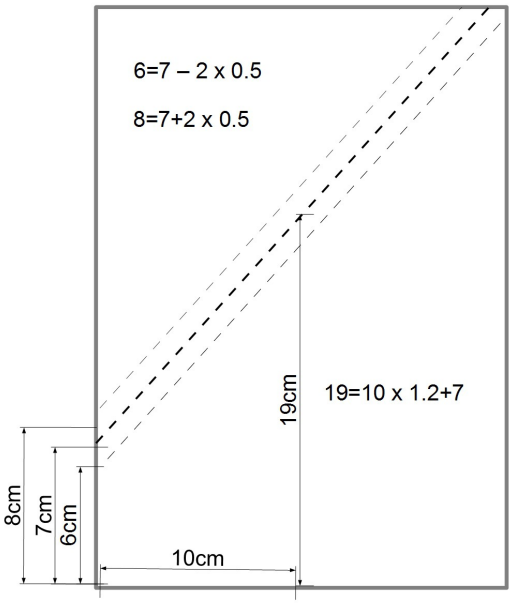
.

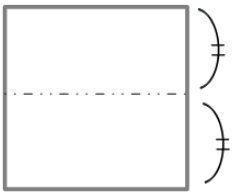


This defines the area in which we expect to find about 95% of the observations (mean ±2SD).

I hope you have to hand a few squares of paper, no matter what size. The length of a side will be denoted by the letter l for length.

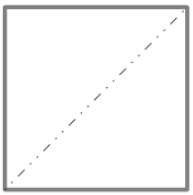
We determine that the angle at the bottom left of our sheet is the origin. Here and on the next page are some examples of how to fold a few simple linear functions and their parameters.
Below is an example of how to fold an A4 sheet of paper to represent the functions described at the end of chapter 3. That is to say the line $y=1.2x+7$ to which we added a Gaussian error with sd = 0.5 using the function RNorm ( ).
As you can see, we have to represent the results of the calculations: a lot of numbers, ruler, measurements ... all for three miserable folds.
Besides not everything can be solved only with the origami. It is better to use R and `abline()`.





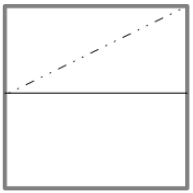
$$y=\frac{1}{2}l$$

a=	½ l
b=	0



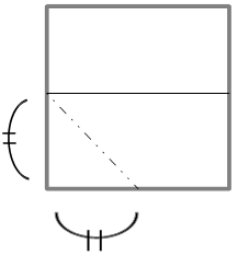
$$y=x$$

a=	0
b=	1



$$y=\left(\frac{1}{2}l\right)\left(\frac{1}{2}x\right)$$

a=	½ l
b=	½



$$y=\left(\frac{1}{2}l\right)(-x)$$

a=	½ l
b=	-1

Appendix D
Formulas

Mean: The sum from i = 1 to n (= first to last) of the n elements of x, divided by n

$$\overline{x}=\left(\sum_{i=1}^{n} x_i\right)\div n$$

Range: The maximum value minus the minimum value

$$R=x_{Max}-x_{Min}$$

Deviance: The sum from i = 1 to n of the squared differences between the n values of x and the mean of x.

$$D=\sum_{i=1}^{n}\left(\overline{x}-x_i\right)^2$$

Variance: Deviance divided by the degrees of freedom

$$S^2=D\div(n-1)$$

or

$$S^2=\left(\sum_{i=1}^{n}\left(\overline{x}-x_i\right)^2\right)\div(n-1)$$

Standard deviation: The square root of the variance

$$S=\sqrt{S^2}$$

or

$$S=\sqrt{\left(\sum_{i=1}^{n}\left(\overline{x}-x_i\right)^2\right)\div(n-1)}$$

Standard error: The standard deviation divided by the square root of n

$$ES = S \div \sqrt{n}$$

Covariance

$$Cov(x, y) = \sum_{i=1}^{N} \frac{(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

Correlation Coefficient

$$r = \frac{Cov(x, y)}{S(x)S(y)}$$

The covariance of x and y divided by the product of their variance

Gaussian

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

Student's t for unpaired data: The difference between the two means divided by the standard error of the difference, that is, the square root of the sum of the 2 variances divided by n. NB it is possible that $n_1 \neq n_2$

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$df = n_1 + n_2 - 2$$

Student's t for paired data: The average of the differences between the two samples divided by the standard error of those differences. NB $n_1 = n_2 = n$

$$t = \frac{\sum (x_1 - x_2) \div n}{\sqrt{s^2 \div n}}$$

$$df = n - 1$$

The answer

It is not difficult. ABC and ADB are both right angled triangles, so they have an angle that is equal. But also angle DAB is equal, therefore the third angle must also be equal, because in all triangles the sum of the angles is equal to 180 °. From this it is also the case that DCB is similar to the first 2 triangles.

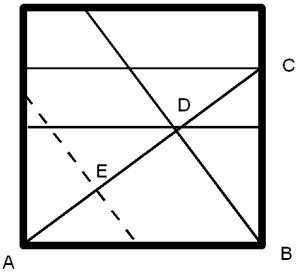


fig.4

Bibliography (with comments?)

[1] Origami Omnibus by Kuniko Kasahara is published by Japan Publications in Tokyo. It is a beautiful book written in English containing a lot of food for thought as well as many fun models.

[2] For those who want to start origami I suggest a book also by Kuniko Kasahara, Origami Facile, ed. il Castello, 1978 Milano.

[3] 13 Thoki Yenn Orikata is a short, but very stimulating book, published by the British Origami Society (www.britishorigami.org.uk/) in April 1985. It was reprinted in 1987 in A4 format. It can be bought from BOS.

[4] A nice book in relation to origami and geometry is that by Tomoko Fuse: Origami Modulare, il Castello, 1988 Milano.

[5] Super quick origami animals by Nick Robinson, Sterling Publisher Co. Inc (New York 2002) brings together many folds of animals that are delightful in their simplicity.

[6] Luigi Pirandello, Six characters in search of an author in "Maschere Nude"; various editors, for example. Garzanti.

[7] In the book by Box, Hunter and Hunter called Statistics for Experimenters ed John Wiley & Sons 1978 New York; you can find a more formal treatment of the geometrical model of the ANOVA.

[8] There are many books available for those who want to study statistics seriously. To choose one that deals more seriously with the subjects I have played with in this book I would pick the book by Lamberto.Soliani that you can find free on the internet at http://www.dsa.unipr.it/soliani/soliani.html. It is very nice and comprehensive.

[9] The objective of this book is to make you more familiar with some statistical concepts as well as to learn to interpret the results of statistical analysis. On this subject there is a nice introductory book written by G.Gigerenzer: Quando i numeri ingannano. Imparare a vivere con l'incertezza. Raffaello Cortina Editore, 2002 Milano.

[10] The idea of using a wooden stick, nails and rubber bands to create a linear regression was not mine, but comes from an article published several years ago in the journal "Le Scienze" N. 204 August 1985 on page 112 in the section (Ri)creazioni al calcolatore by A.K.Dewdney. The article is called "Congegni analogici che risolvono problemi di varia natura e sollevano un sacco di domande".

[11] Bland J.M. Altman D.G. (1986). Statistical Methods for Assessing Agreement between two Methods of Clinical Measurement *Lancet* i: 307-310

[12] Shrout P.E. Fleiss J.L. (1979). Intraclass Correlations: Uses in Assessing Rater Reliability
Psychological Bulletin 86, 2; 420- 428

[13] Marubini E. Pizzamiglio S Verderio P (2005). Agreement between observers: Its measure on a quantitative scale
The International Journal of Biological Markers 20;1, 73-78

[14] In Petronius Arbiter's Satyricon (First century AD) and in particular in Trimalchio's dinner, it was considered funny to give the slave who had to cut the meat , the name "Cut" . In this way, the host could make the elegant (?) joke: "cut, Cut!"