Modelling Mondays

Argyris Stringaris

2024-05-06

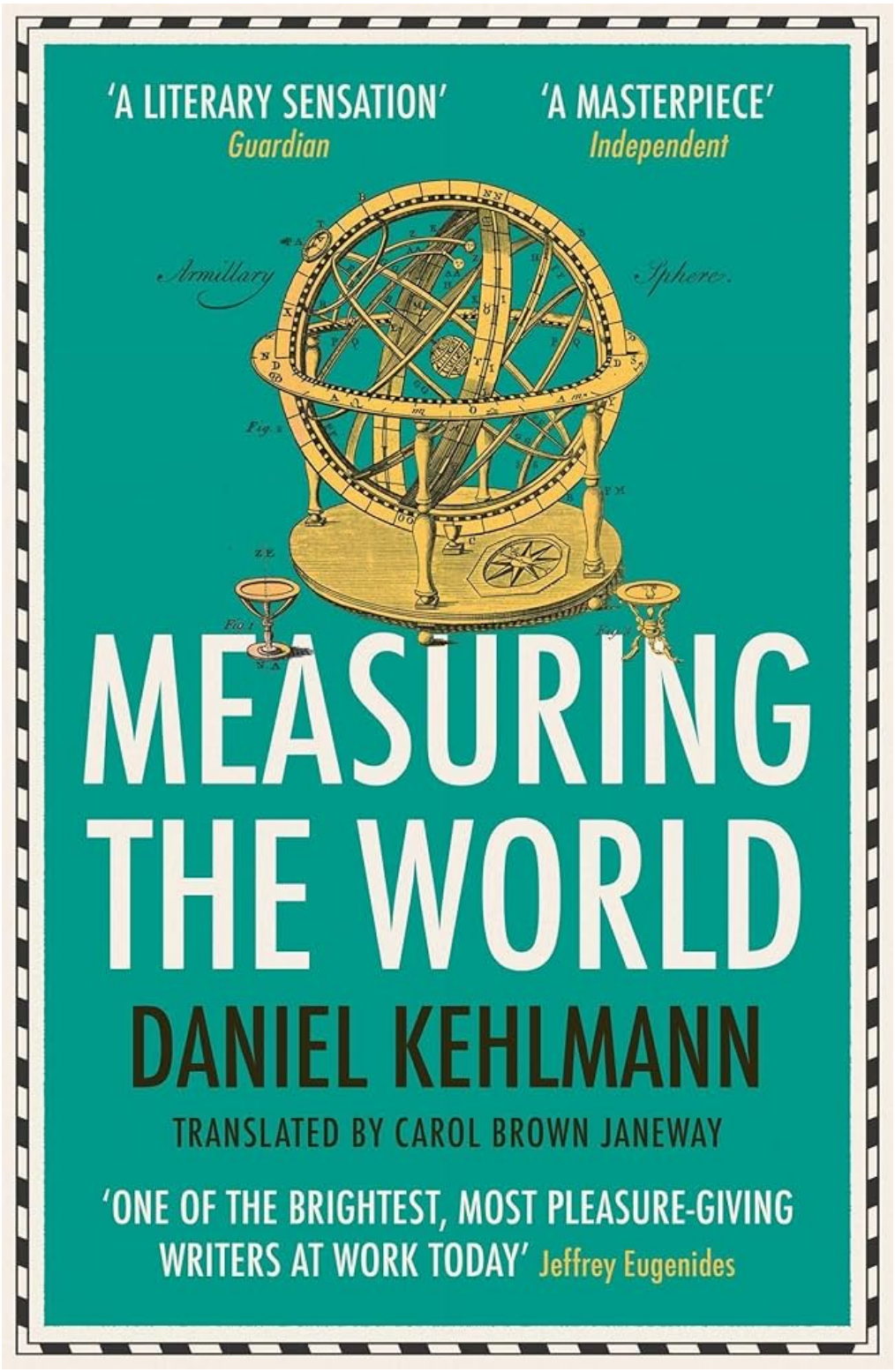
Table of contents

Week 1: The Likelihood Function

Motivation

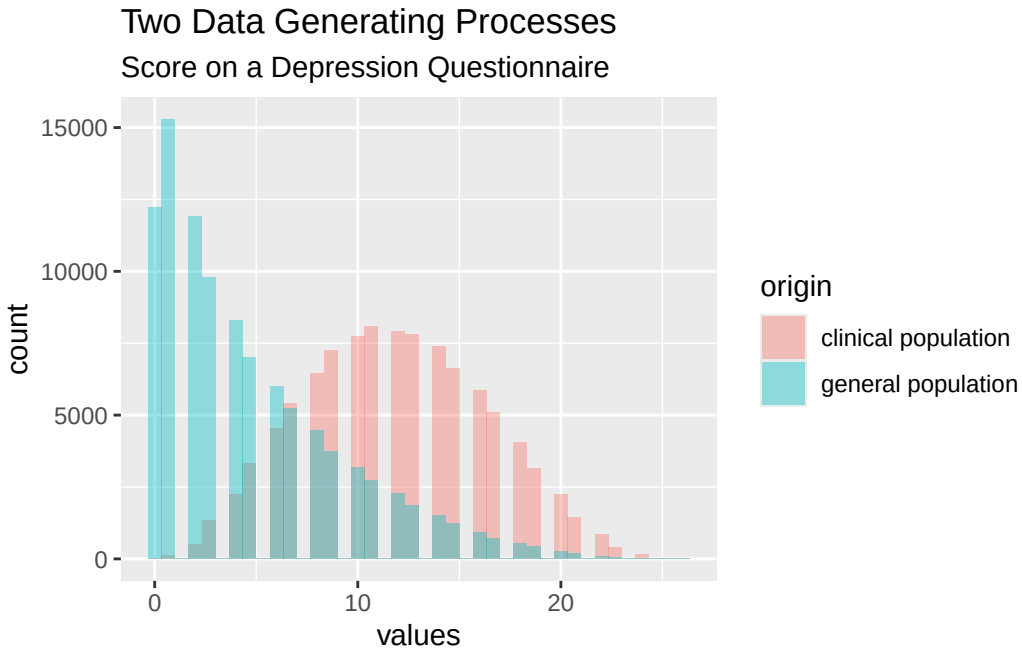
What is Data Generating Process (DGP) and what is a likelihood function?

Typically, we think of a DGP as a mathematical formula that gives rise to a distribution. For example, the IQ curve can be generated through the Gaussian, named after Karl Friedrich Gauß–very much worth reading about also in the novel The Measuring of the World, by Kehlmann (where the parallel lives of Gauß and Humboldt are presented).


'A LITERARY SENSATION'
Guardian
'A MASTERPIECE'
Independent
Armillary Sphere.
MEASURING THE WORLD
DANIEL KEHLMANN
TRANSLATED BY CAROL BROWN JANEWAY
'ONE OF THE BRIGHTEST, MOST PLEASURE-GIVING WRITERS AT WORK TODAY' Jeffrey Eugenides

But in a more abstract way, the question is, what are the mechanisms through which a set of data are generated, be it voting patterns, brain data or league games.

Consider, for example, a sample of the general population filling in a questionnaire about depression. Figure 1a. shows a typical pattern, that of a right-skewed truncated distribution. The "mechanism" that gives rise to the right skew is the fact that there are far more people without many symptoms and hence many people close to the zero mark. It is also truncated because scores can't go below zero and can't go above the max of the sum of the scale. By contrast, Figure 1b, shows the depression scores of a clinical population.



In general, we always want to consider the DGP so as to:

a) understand what gives rise to the data.

b) mathematically describe (at least) how the data arise.

c) estimate parameters (related to b)

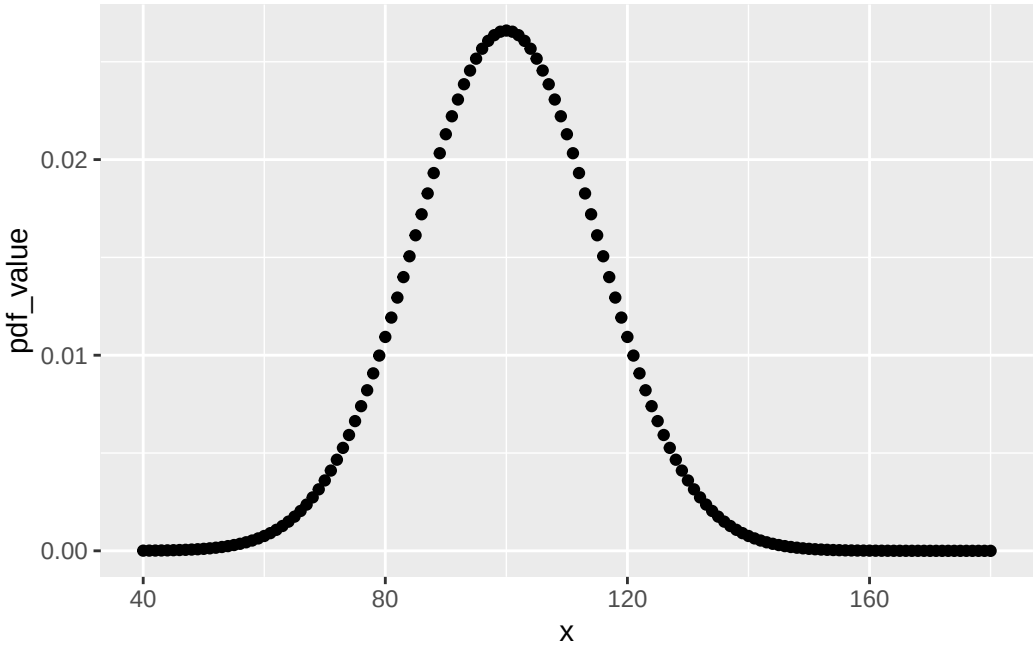
d) simulate the process to study it better.

The omniscient person: knowing the DGS and the correct parameter.

This is someone who knows the function and its probability, is certain about the DGP. Let's say that they know that they are dealing with the normal distribution, which is formalised as:

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

where, x is the point of interest of the probability density function, θ is the mean (location parameter) of the normal distribution, and σ is the standard deviation (spread parameter).



You will all recognise this as the standard IQ curve.

Please note from Equation 1 that here the point is that the situation is phrased as:

$$f(x|\theta, \sigma)$$

i.e. we ask what the probability is of obtaining these data given the parameters θ.

The situation where you are certain about the correct parameter and only need to know the frequency of individual values or set of values is a very convenient one to be in. Often however, in the real world we may have an intuition about what the DGP might be but not know the parameter(s). That is when we ask about the likelihood.

A real person: having data, intuiting the DGS, and not knowing the parameter.

Consider having collected some data, having some intuition about the DGS and needing to find out the parameter amongst a set of parameters.

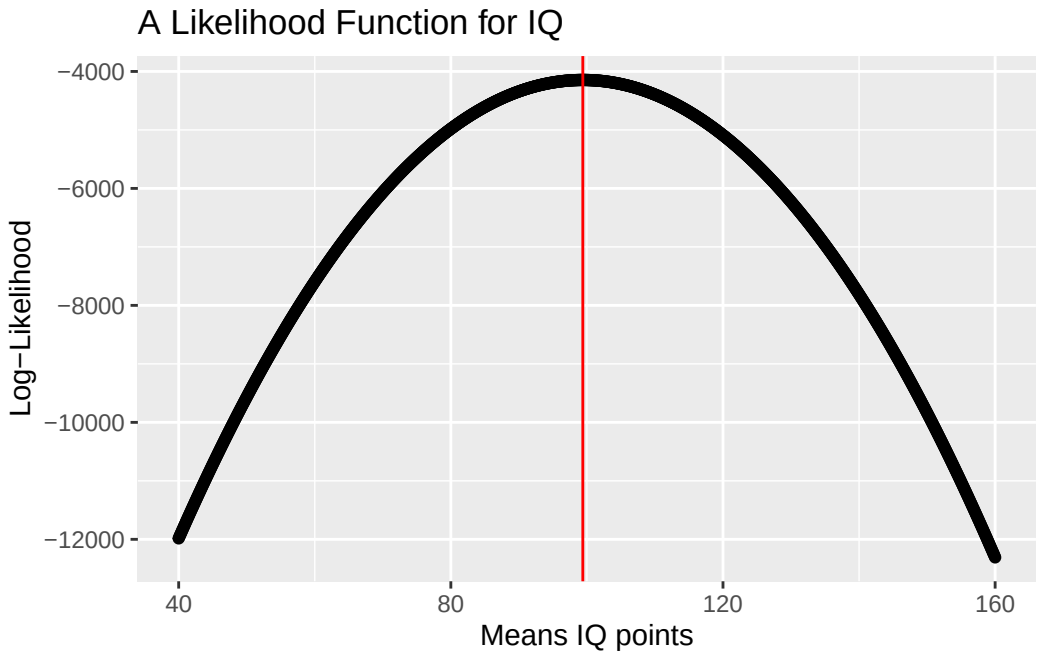
This is a more likely situation which I will illustrate here by trying to recover the mean parameter from synthetic data.

For the moment it is safe to say that what were trying to do here is to invert the process above, i.e. what you do with the probability density function. Instead of asking what data are likely to occur given a parameter (such as the mean and sd) that you *already* know about, here you ask, what is the most likely parameter that has given rise to the data I have.

$$L(\mu, \sigma^2 | x_1, x_2, \dots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) \tag{2}$$

Equation 2 states precisely that: what is the likelihood of this mean and variance, given all these data points? Equation 2 on the right-hand side contains the PDF, as above, but what it says is that it takes the probability at each step and multiplies them altogether, this is what that giant Greek Π stands for, the product.

Notice that when I tried this with fewer data points, I was able to get the likelihood, but when I increased them, I needed the natural log. Try it for yourself.



Likelihood for the linear regression model

Now let's turn to the simple linear regression model. Let's start by asking how to think formally of the data generating mechanism of any linear model. It should be a

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{3}$$

where, ϵ_i follows a normal distribution with mean zero and variance σ^2

$$L(\beta_0, \beta_1 | x_1, y_1, x_2, y_2, ..., x_n, y_n) = \prod_{i=1}^{n} f(y_i | \beta_0 + \beta_1 x_i) \tag{4}$$

where

$$f(y_i | \beta_0 + \beta_1 x_i) \tag{5}$$

is the probability density function (PDF) of the normal distribution with mean ($\mu_i = \beta_0 + \beta_1 x_i$) and constant variance $sigma^2$

To demonstrate this, I will first create synthetic data

In the code chunk below, I explain how the outer product and vectorisation works.

Week 2: The Binomial Distribution

Before we move over to more complex models, let's consider the binomial or Bernoulli distribution. Here are two situations. In the first one, you are a Creator (let's say a game creator, rather than The Creator); in the other you are a detective.

The binomial from a creator's point of view.

Let us assume that you are trying to create a game for which you must create sequences of binary events, let's say decisions between a state H and a state T. Basically, you want either of these two to appear with a probability that is on average $\pi = 0.5$ i.e. a 50% chance of appearing.

I want to take us back to something which whilst obvious, is often forgotten, namely that probabilities are things that we can understand "in the long run".

Here is what I mean. The way to create the game above is to invoke the binomial distribution. This is the following:

$$f(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k} \tag{6}$$

where $\binom{n}{k}$ is the binomial coefficient (will explain this) and then come the probabilities.

Let's explain this. The binomial coefficient, basically says: if I have n objects and want to choose k of them, how many ways can this be done. Think of the following example. I have the letters ABCD; how many ways are there to combine two letters (sequnce doesn't matter, e.g. AB = BA) There are 6 possible ways: AB AC AD BC BD CD to play around with it, look at the code below.

```
[1] "We have 1 way(s) to choose 0 objects out of 4"
[2] "We have 4 way(s) to choose 1 objects out of 4"
[3] "We have 6 way(s) to choose 2 objects out of 4"
[4] "We have 4 way(s) to choose 3 objects out of 4"
[5] "We have 1 way(s) to choose 4 objects out of 4"
```

In Equation 3, this quantity is then multiplied with the product of $p^k \ x \ (1-p)^{n-k}$. This product is a sequence of possible events of success and failure, for a probability p . If you substitute numbers between 1 and, say, 4 (representing possible outcomes in 4 coin tosses) into them, you would get

$p^1 \ x \ (1-p)^{4-1}$

$p^2 \ x \ (1-p)^{4-2}$

$p^3 \ x \ (1-p)^{4-3}$

$p^4 \ x \ (1-p)^{4-4}$

Each of these sequences is then multiplied with the number of ways k objects can be chosen out of n total objects (e.g. the number of 4 times Heads in 10 throws).

A priori, which one of these outcomes would you expect to be more likely for a fair coin?

```
[1] 0.2500 0.3750 0.2500 0.0625
```


```
[1] 0.2500 0.3750 0.2500 0.0625
```

Exercise 1.1

Modify the above code to create a game where a coin is tossed 100 times.

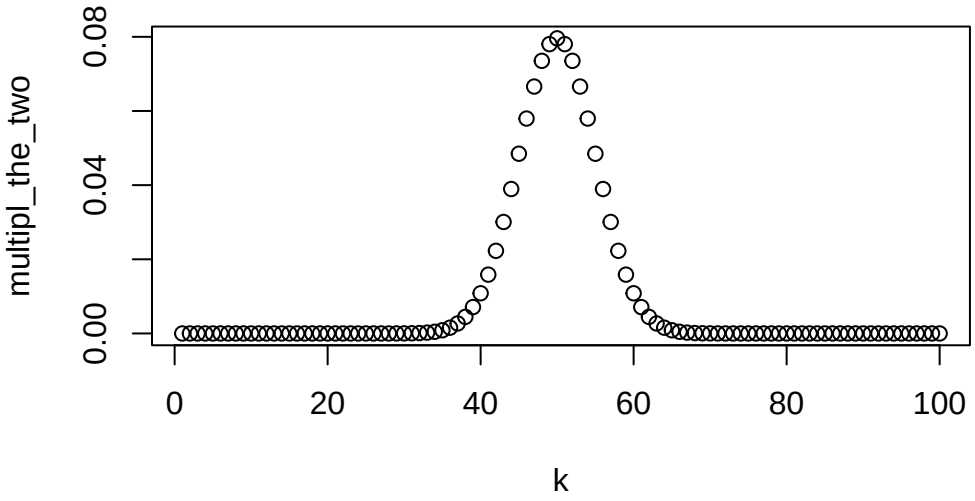
a) Estimate the probabilities for each possible outcome, k and store in vector.

b) Find the outcome, k with the maximum probability

c) Plot all possible outcomes against their probabilities.

d) check against the standard inbuilt R function

```
[1] 50
```



The binomial for a detective

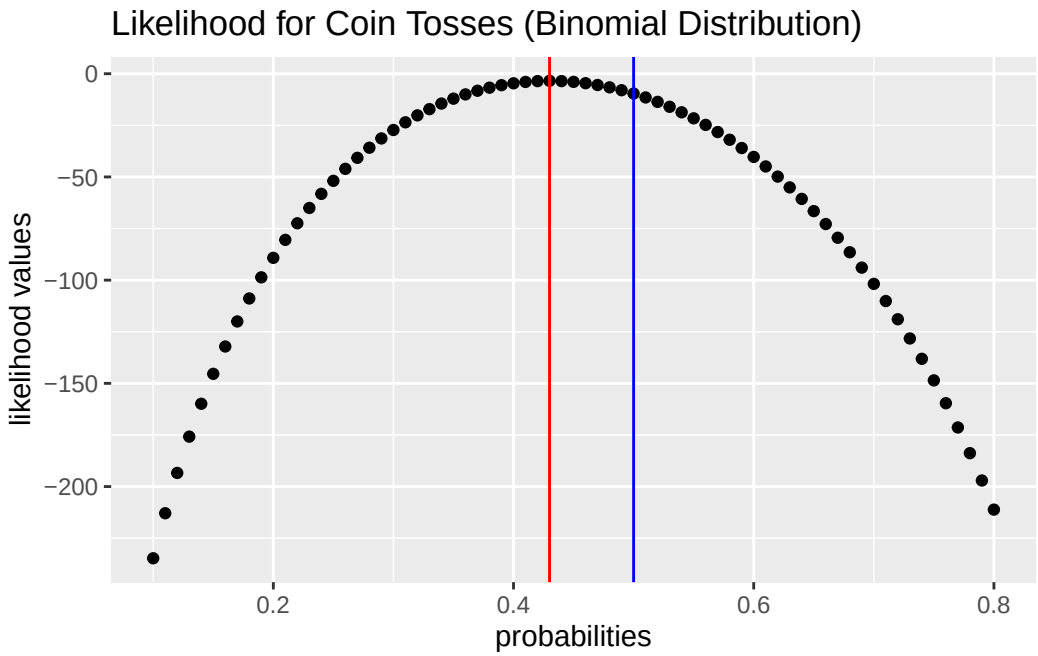
Now, suppose that you are the new detective in town. Your first case is that of "Nick the Shark", against whom there are several allegations of setting up fraudulent games. All you have to go by is a sheet of paper where all the outcomes of coin tosses were recorded by one of your informants. There were 271 outcomes that were Heads. You are now asked to find out whether the coin was fair or not.

You would ask for the help of the statistician, but they are all away at a big conference and you must appear in from of the judge who decides on whether the person can remain in detention or not.

The judge has an exceptional understanding of numbers for a legal person and asks you to prove to her your case that Nick is indeed a swindler, as you say, and not the upright occasionally gambling citizen that the defendant maintains that he is.

You spend the night, writing out all outcomes of the coin tosses, all 630 of them.

```
[1] 0.43
```



This is impressive, you have evidence that the parameter that maximises the likelihood is different to 0.5. But the judge gets back at you and says: all this could easily be due to chance. After all, probability is a matter of doing "experiments in the long run"/

You are stunned at the unprecedented numeracy of a lawyer. She warns you that she will throw out the case and you will not get the arrest warrant issued due to insufficient evidence.

How can you demonstrate that the difference between the blue and the red line is not simply due to chance?

The likelihood ratio test

The question is whether the likelihood at 0.5 (the null) is different to the likelihood at what you found to be the maximum likelihood in the observed data. What if you built the ratio of these two?

Indeed, the likelihood ratio test will allow you to answer the numerate judge's question. Because you have taken logs, the problem simplifies to a subtraction (logging ratios turns to a subtraction).

$$\text{Likelihood Ratio} = -2 * (\text{Log Likelihood}_{0.5} - \text{Log Likelihood}_{ML}) \tag{7}$$

Now, you may wonder about what that *2* is doing there–not to worry about it for the moment, its presence allows you to assume this difference follows a chi-squared distribution. Don't forget the minus sign–you will need this as the values of the chi-squared distribution are all positive.

```
[1] 0.0004451908
```

After this you can go back to the smart judge and convince her that your finding is very unlikely to have occurred by chance. To help you phrase things better to the judge, I have given you the exercise below.

Exercise 1.2

a) How exactly would you phrase your finding? How unlikely is it that Nick's games have occurred by chance?

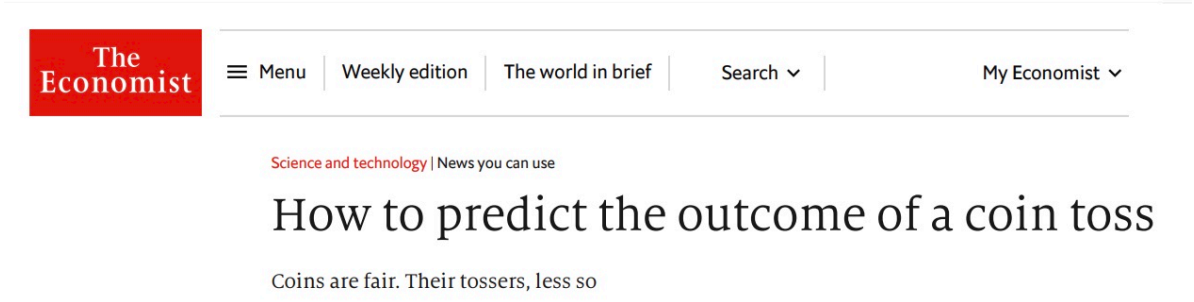
b) How would you construct standard errors and confidence intervals around those estimates? How would phrase the findings about the confidence intervals?

c) Can you plot confidence intervals on the graph with the red and blue line?

Bonus Questions:

d) You do a debrief with your team of detectives and informants. On this occasion, your informant had gathered 630 games. But what if he had sampled less, or more? Can you find out how many games he would have needed to have gathered for you to be able to demonstrate this difference to the judge (e.g. would 30 games be enough)? What do you call such a question in science?

Small diversion: is coin tossing fair?



Check out this article here: https://www.economist.com/science-and-technology/2023/10/15/how-to-predict-the-outcome-of-a-coin-toss

Week 3: Probability theory and The Bayes Theorem

Of Men and Homicides

Before we revisit the above from a Bayesian perspective, we will need a small detour into probability theory.

About 90% of homicides in Europe are committed by men. How justified is it to say that "men are murderers". Think about this question also by replacing men with immigrant men, foreign men, or foreigners more generally. Think about what may be true in the aggregate (and generates stereotypes) and what is valuable at the individual level. Let's try to tackle this problem in a number of ways.

Let there be a population where,

the probability of homicide in a country be 2.3 in 100,000, i.e. $P(homicide) = 2.3$*10^-5

the probability of being a man in that population be 50%, i.e. $P(man) = 0.5$, and

the probability that if there is a homicide the perpetrator is a man be 90%, i.e. $P(man|homicide) = 0.9$

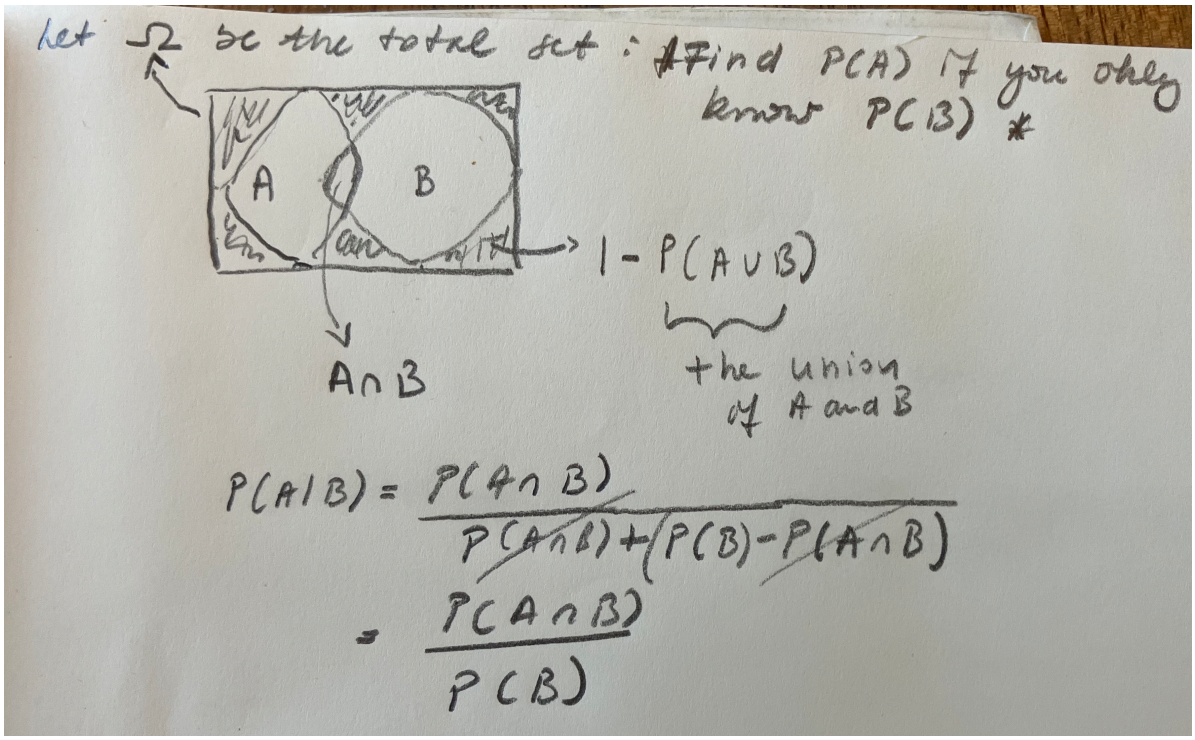
The question is what is the probability that if I see a man on the street, he is a murderer, i.e. $P(homicide|man)$.

Let's arrive at this step by step.

First, let's remind ourselves what the probability is of two events occurring together:

$$P(A \cap B) = P(A) \cdot P(B)$$

This says that the co-occurrence of two events is the product of each event. However, this is only true if the two events are independent of each other, i.e. the occurrence of A has nothing to do with the occurrence of B. Is this the case here with men and homicides. What more general rule can we apply? Let's try to understand this graphically.



When you re-arrange this, you arrive at the very important following formula.

$$P(A \cap B) = P(A \mid B) \cdot P(B)$$

This formula allows you to calculate conditional probabilities if you have the joint ones and a prevalence, and vice versa. But this does not quite help us because we don't know the joint probability of homicides. For this we employ a trick. Re-arranged equation 9, also holds this way:

$$P(A \cap B) = P(B \mid A) \cdot P(A)$$

which then means that,

$$P(A \mid B) \cdot P(B) = P(B \mid A) \cdot P(A) \tag{8}$$

So, now you can estimate any of the two conditional probabilities, if you know the rest.

Let's apply this to our homicide example. Remember we need to estimate: $P(homicide|man)$

Therefore, rearranging and substituting our terms into Equation 10, we get:

$$P(homicide \mid man) = \frac{P(man \mid homicide).P(homicide)}{P(man)} \tag{9}$$

CONGRATULATIONS: you have just entered the world of the Reverend Thomas Bayes! This is his theorem applied to homicides!

As we will see further down, Bayes links back to the likelihood that we have been discussing and allows us to use priors and

Exercise 3. 1

a) Calculate conditional probability. b) Do so for a country like Brazil too, where the probability of homicide is about 10-fold higher. c) Comment on whether calling men, foreigners etc murderers may be considered stereotyping. What does it mean for individual prediction and what does it mean for public health and safety.

Men, Homicides and marginal probabilities

Now let's look at the problem from a different angle. Let's create a table that captures the above in a representative sample, n = 100,000 of the population in Brazil, where the probability of homicide is about 20/10^5, the gender ratio is assumed to be equal and the probability that a homicide is committed by a man is 0.9.

```
        homicides
genders     No   Yes
  Female 50033     2
  Male   49947    18
```

How do you calculate here $P(homicide|male)$. Do it by hand. Do it also after substituting the European homicide probability given above. Do you get the same results?

```
[1] "P(homicide | male) = 0.00036"
```

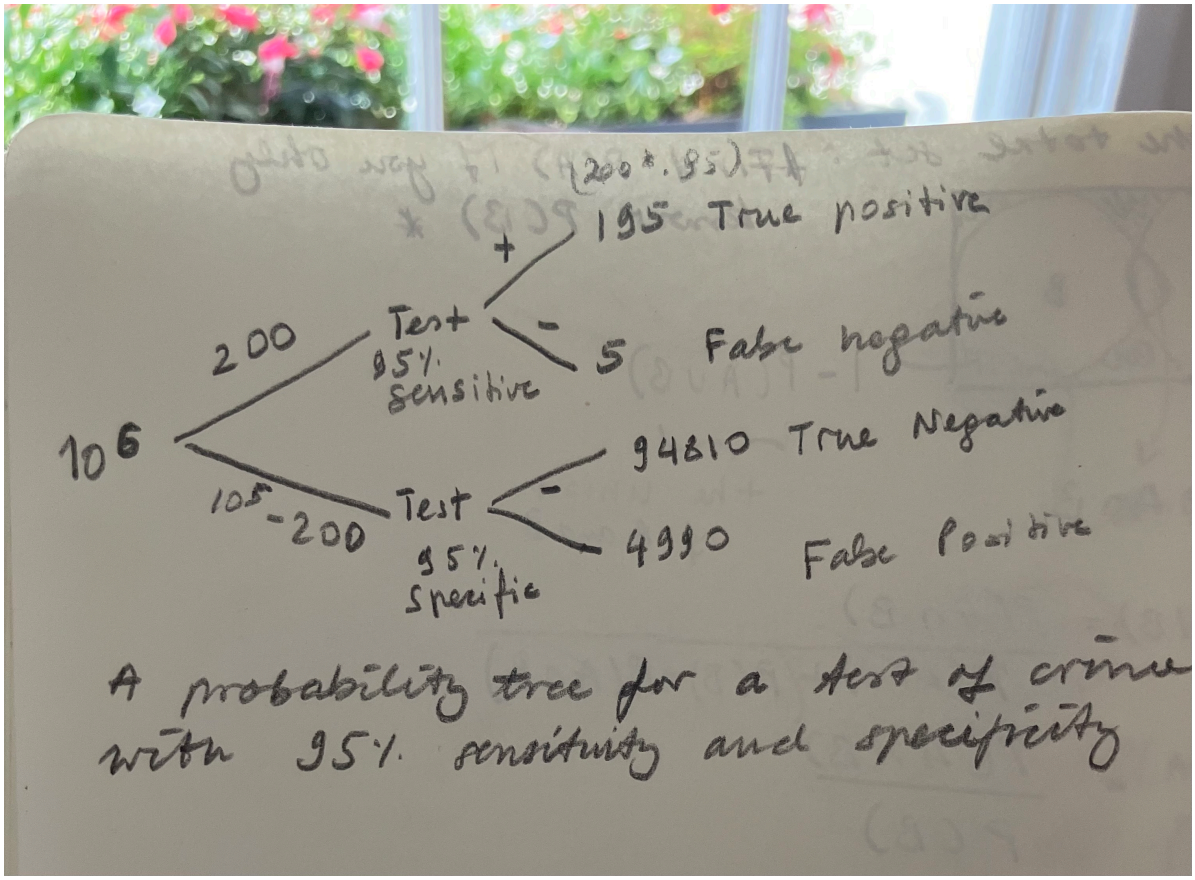
Congratulations, you have just used a **marginal probability**, namely you summed the two outcomes for men, the Nos and Yess, and used them as denominators. This is fundamental in Bayesian statistics, as we shall see in the next few sessions. Here it is very simple. By the way, do it for girls too, what is it? It is an order of magnitude less, as you might expect, but both are very low. Ask yourselves, would gender be a good test to detect suicides?

Homicides and Probability Trees

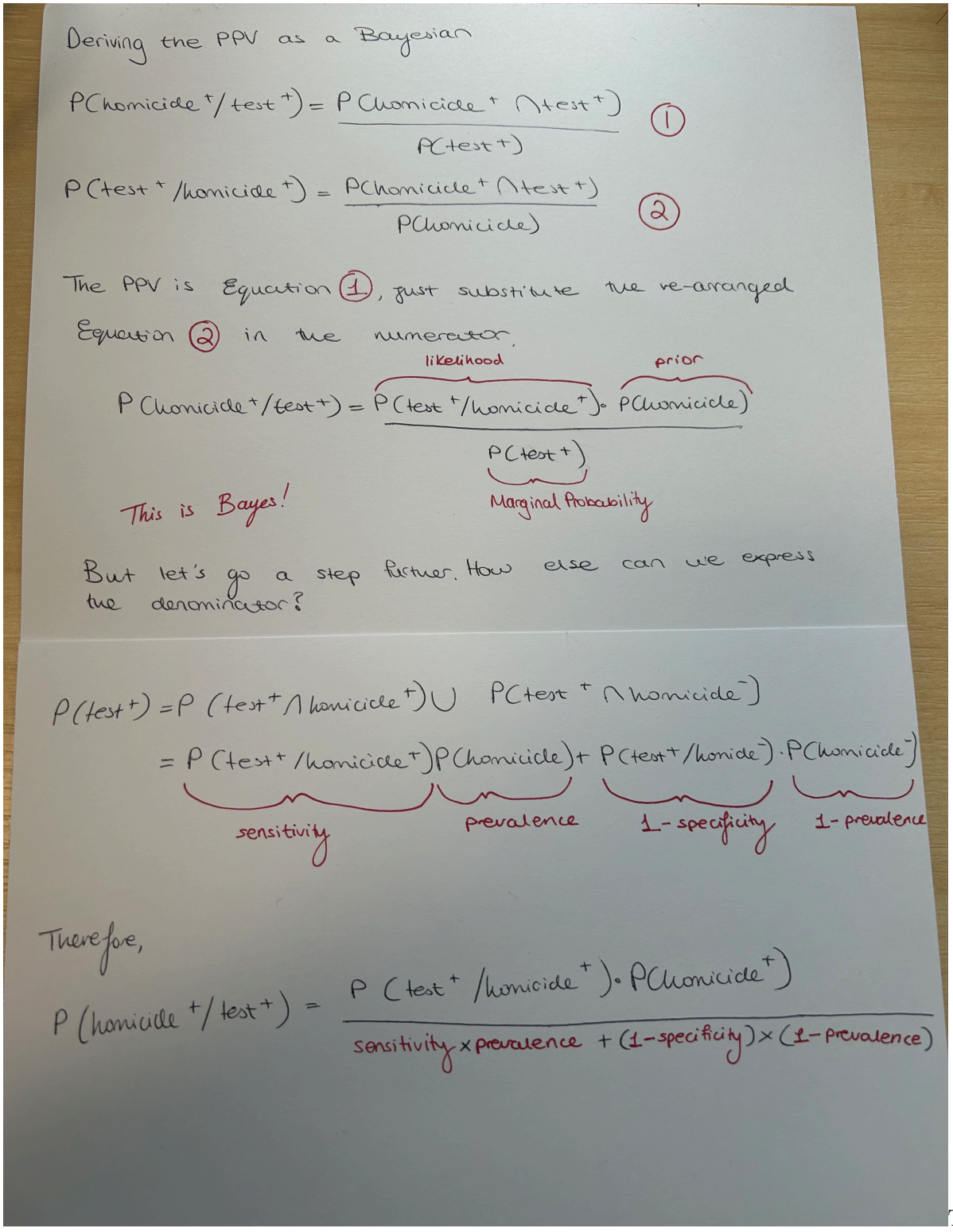
Now let's say that a new company comes and tells you that it has excellent sensitivity and specificity with 95% to detect the scent of a criminal. What conditional probabilities do sensitivity and specificity refer to?

Exactly, Sensitivity is $P(Test+|Criminal+)$ where + denotes having the characteristics (test positivity and being a criminal), i.e. how likely is the test to be positive if you are a criminal. The specificity is $P(Test-|Criminal-)$

Below is a probability tree. Where do you find the sensitivity and where do you find the specificity? And how do you estimate the reverse of the sensitivity? This is the key question, you are not that much interested in how the test performs in criminals, but rather **how the test behaves** in the population you are likely to encounter. This is given by $P(Criminal+|Test+)$, i.e. the probability that you are indeed a criminal if you have a positive test. How do you calculate this and what is your denominator?



This quantity is fundamental to all medical tests and indeed all tests where you want to draw inferences about the goodness of the test in a given population. It is the **Positive Predictive Value and it is a Bayesian quantity.** Using very basic algebra and the rules derived above, I will try to demonstrate this in the picture below.


Deriving the PPV as a Bayesian
P(homicide+/test+) = P(homicide+ ∩ test+) / P(test+) ①
P(test+/homicide+) = P(homicide+ ∩ test+) / P(homicide) ②
The PPV is Equation ①, just substitute the re-arranged Equation ② in the numerator.
likelihood
prior
P(homicide+/test+) = P(test+/homicide+) · P(homicide) / P(test+)
Marginal Probability
This is Bayes!
But let's go a step further. How else can we express the denominator?
P(test+) = P(test+ ∩ homicide+) ∪ P(test+ ∩ homicide-)
= P(test+/homicide+) P(homicide) + P(test+/homicide-) · P(homicide-)
sensitivity
prevalence
1-specificity
1-prevalence
Therefore,
P(homicide+/test+) = P(test+/homicide+) · P(homicide+) / sensitivity x prevalence + (1-specificity) x (1-prevalence)

The

two key equations here are:

$$P(homicide+|test+) = \frac{P(test+|homicide+).P(homicide+)}{P(test+)} \tag{10}$$

Notice the similarity of this equation with that of Equation 11. I have also added in red some nomenclature that we will be encountering very soon, in the next lesson. Now, as I have derived above, there is another way to derive the same quantity using the known properties of the sensitivity and specificity and the prevalence of the population, without needing any other information.

$$P(homicide+|test+) = \frac{P(test+|homicide+).P(homicide+)}{sensitivity.P(homicide+) + (1 - specificity).P(homicide-)} \tag{11}$$

Look at the denominators of both Eq. 12 and 13. These are the marginal likelihoods (also called the evidence). They are cumbersome, but not nearly as complex as what we will be encountering soon, even for solving the same simple binomial problem that we had solved in the last lesson. Indeed, these denominators are often analytically intractable and require approaches such as MCMC algorithms.

We will come to all this.

Meanwhile, I am going to give you some extra code that allows you to play around with sensitivities, specificities, PPV, NPV, but also with the chances of having a disease if someone tells you that you have a negative test (always very important).

Try to understand the basic notion of the Bayesian theorem and apply it to various situations of interest to you, like medical tests, exam results etc. Next time we will pick up again the problem of Nick the Shark and play around with the Bayesian estimation of the finding.

Week 4: Some more on crime and punishment

Contingency tables and expected values (a little detour)

Before we delve into some Bayesian stuff, it may be good to remind ourselves of some very simple principles that would help us decide whether men are more likely to commit crimes according to common standards of significance.

Just to remind you, this was our contingency table:

```
        homicides
genders     No   Yes
  Female 50033     2
  Male   49947    18
```

How would you decide on whether the differences are "significant". You did come up with probabilities for this problem above for males and females. But how would you know that they differ?

We will treat this problem using a standard frequentist approach and then turn to a Bayesian answer later in our meetings. What we need to do is create expected values for each cell in the above example. What would you do?

First, you will need the marginals. There are three types of marginals, the row marginals, the column marginals, and the totals.

The command below gives you the row marginals.

```
marginSums(contingency_table,1)
```

```
genders
Female   Male
 50035  49965
```

This one the column ones

```
marginSums(contingency_table,2)
```

```
homicides
   No   Yes
99980    20
```

and here is the totals

```
marginSums(contingency_table)
```

```
[1] 100000
```

or to create the whole table do

```
addmargins(contingency_table)
```

```
        homicides
genders      No    Yes    Sum
  Female  50033      2  50035
  Male    49947     18  49965
  Sum     99980     20 100000
```

$P(gender == female \cap homicide = no) = P(gender == female \mid homicide == no).P(homicide == no)$

$P(gender == female \cap homicide = yes) = P(gender == female \mid homicide == yes).P(homicide == yes)$

$\sum P(females == yes|homicide)$

Can you think of a way to get expected values now?

There are two principle ways to think about it which are complementary. Either to think of the "rule of three" from primary school maths, or to think in a Bayesian (or rather generally more abstract) way about it. I will start with the simple way.

Let's start with Females who do not kill, how many would we expect? This is the top lefthand corner cell. We **observe** 50167 who have not committed murder and 2 who have. How many would we have expected in each of these two cells? How do we use the term expected? In the sense that ignore the observed column values and say, well there are 50169 (the row total value) overall in 10^5 people (the overall total). How many would there be in the 99980 (the No column, in which the first cell is situated). It follows that we obtain the **expected value** by doing $50169 * 99980/(10^5)$ which is 50159 after rounding.

If you do the same for each one of the cells, you get the following table of **expected** values.

```
exp_table <- chisq.test(contingency_table)$expected
exp_table
```

```
        homicides
genders        No    Yes
  Female 50024.99 10.007
  Male   49955.01  9.993
```

compare the two tables, what do you see?

How can you formalise this into a statistical answer? This will be an exercise for next time (see below). For the moment, and in the interest of abstracting, what exactly did I do here when I used the rule of three?

Let's write out some simple operations in fancy terms.

What is the probability of being female and not being a murderer on the basis of the observed data?

$$P(gender == female \cap homicide == no) = 50167/10^5 = 0.5017 \tag{12}$$

But, let's go even more fancy, let's apply the Bayesian theorem and the relationship between joint and conditional probabilities.

$$P(gender == female \cap homicide == no) = P(gender == female|homicide == no).P(homicide == no) \tag{13}$$

This gives us the same result as you can verify by multiplying the two fractions: 50167/99980 x 99980/10^5.

This is a bit ridiculous and a near tautology, but humour me for a bit. What if I asked you what the following quantity is (which is the equivalent of Equations 6.8 and 6.9 in the Farrell and Lewandowsky book:

$$P(gender == female) = \sum_{\theta} P(\text{female} \mid \text{homiicide}) \cdot P(\text{homicide}) \tag{14}$$

You needn't worry of course, because all you have to do is to calculate eq-16 is to add eq-15 and eq 17 below (which is its counterpart, the cell right next to it, i.e. murdering females.

$$P(gender == female \cap homicide == yes) = P(gender == female|homicide == yes).P(homicide == yes) \tag{15}$$

Which gives you: (2/20*20/10^5) and is the same as doing the following simpler calculation.

$$P(gender == female|homicide == yes) = 2/10^5 = 2x10^{-}5 \tag{16}$$

Now, all you have to do is add the results of

Indeed, when you try to add the results of summing eqs 15 and 17, you get to the marginal probability, which is all that equation 18 is asking you to do, except in fancy formalism:

(50167/99980*99980/10^5) = 0.50169 which is the same as what you would get if you simply divided the marginal for gender with the total i.e. 50169/10^5

PHEW!

Does all this make sense? It is quite simple but can be mind-boggling.

Here is an exercise to consolidate expected values, we will follow up with the Bayesian stuff below.

Exercise 4.1

Once you have solved this exercise, you may get a fundamental insight about model evaluation, at least I did when I grasped this.

a) look at the table above with the observed values, let's call this table O_table and also look at the one with the expected values E_table. Try to conceive of them as locations on some imaginary map. All those numbers are nothing but locations on that multi-dimensional map. Indeed, each table is a matrix and a matrix can be thought of as a location in a space that has the matrices dimensions. The question arises then: how far away is O_table from E_table? What simple mathematical operation would allow you to answer this question?

b) how can you tweak that simple mathematical operation to calculate on the distance.

c) if you want to do the statistical estimation you will need some extra tools. Hint: what is a common way, e.g. used in simple regression estimation to get rid of annoying signs (positive, negative, without taking the absolute though)? another hint: you will need a distribution for deciding.

Back to Bayes and meeting the beta distribution

From the discourse above, we need to remember equation 17, which I am writing here in its general form:

$$P(y) = \sum_{\theta} P(\mathrm{y} \mid \theta) \cdot P(\theta) \tag{17}$$

This is the broad definition of the marginal likelihood and it is going to pop up very often. Equation 20 is simply its instantiation for continuous variables

$$P(y) = \int_{\theta} P(\mathrm{y} \mid \theta) \cdot P(\theta) d\theta \tag{18}$$

Then of course the fundamental Bayesian equation is written as:

$$P(\theta \mid y) = \frac{P(y \mid \theta).P(\theta)}{\sum_\theta P(\mathrm{y} \mid \theta) \cdot P(\theta)} \tag{19}$$

or, for continuous quantities,

$$P(\theta \mid y) = \frac{P(y \mid \theta).P(\theta)}{\int_\theta P(\mathrm{y} \mid \theta) \cdot P(\theta) d\theta} \tag{20}$$

IMPORTANT: $P(\theta \mid y)$ is the posterior distribution, it is what every Bayesian analysis strives for.

I will skip the chapter on analytic methods for obtaining posteriors in the book in favour of a more conceptual understanding of the beta distribution.

You will all know about the debate between Bayesians and Frequentists. It has been raging for years and it usually focuses on the issue of the **priors** and whether Bayesians are unduly **subjective**. Indeed, Bayesians argue that when you try to make a statement about data, you ought to take prior knowledge into account. Perhaps more importantly, they extend the argument to say, well, you should actually update your model as new knowledge accumulates! Only in this way will you be able to be fair to the state of the world.

I won't go into the various arguments, except to say that even frequentists make a lot of decisions that require scientific **judgement**. The most notable one is the likelihood model that we choose. As we have seen, this is key to all modelling of data. Bayesians would say that their approach provides a principled way of assessing the probability of parameters but also of models. How? We shall see below. I have added some materials about voting patterns and confidence intervals that you may want to study.

For the moment, let's revisit, Nick the gambler.

How would you go about evaluating his honesty in a Bayesian way?

You remember that we used the binomial distribution as our likelihood model to estimate the likelihood of what Nick came up with.

For this let's turn to the beta distribution and highlight some interesting features.

I will start at the end, with a re-writing of the equation 6. 24 in the book for obtaining the posterior distribution of a coin toss with n tosses and k heads:

$$P(\theta \mid k, n) = beta(\theta | \alpha + k, \beta + n - k) \tag{21}$$

This basically says that if you toss a coin and want to use a Bayesian approach (and you choose as most Bayesians would) the beta distribution as a prior, all you have do is add something to those priors!

That seems super simple, but requires a lot of maths to arrive at. We will discuss all this at the next lesson. You may want to look here until then: https://www.statlect.com/probability-distributions/beta-distribution

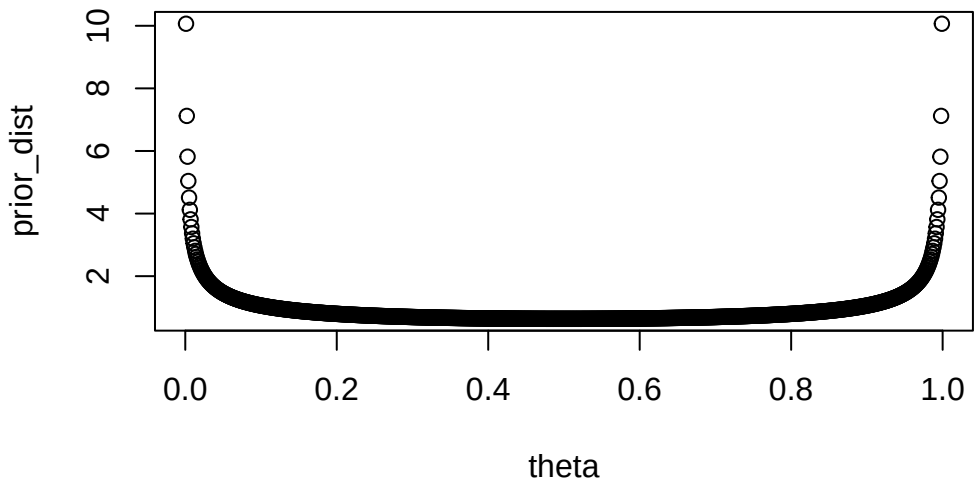
But let's start with some basics.

Let's get an intuition for the beta. I am writing out its probability density function:

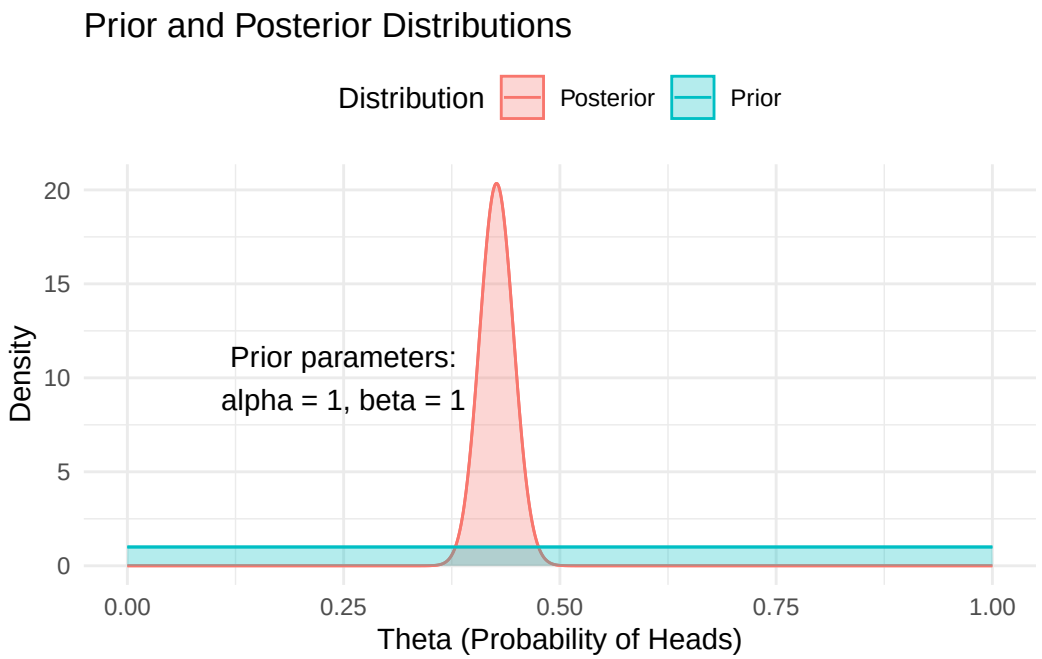
$$f(x;\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \tag{22}$$

We can glean that it has bits that look like the binomial in the numerator (and actually also in the denominator). We will delve into this next time.

For the time being let's play around with the beta distribution for different values of its two parameters. I am using the code below.



Let now return to Nick… Play around with the priors by tuning the α and β to various values and see what happens.




```
[1] 0.4264264
```

Now what do you do with this information? Can you get confidence intervals? Yes!

What can you with these data, can you use them for prediction of future tosses.

Exercise 4.2

a) What is your understanding of what Ipsos, the polsters are saying here. Check out page 3, does it make sense?

https://www.ipsos.com/sites/default/files/2017-03/IpsosPA_CredibilityIntervals.pdf

b) Can you see the relationship with this paper?

https://www.tandfonline.com/doi/pdf/10.1080/01621459.2018.1448823?casa_token=phCtUIGpXcsAAAAA:TnbmBljQ5CaMfCreH_qxMLeIvdEKpJD_tTDPEE0cYK3a_-q0JBHYb3CUqkKHzf2V-gBYW64r5nEfyQ

Week 5: Self Study

Week 6: Inference in Bayes through sampling.

This week we will try to familiarise ourselves with the notion and practicalities of sampling from the posterior. We are going to make small steps into this as it can become quite complex and is probably best revisited as the practical need arises (and it will indeed come up a lot in what we do).

Why use sampling?

Notice the denominator of Equation 22, also called the evidence or the total space of the . Such an integral can be fairly simple, so that mere mortals like ourselves might stand a chance to solve analytically. But it can become quite complex, some times so complex that even evolved machines need help and special tricks to solve. But that does not matter and the reason is that we can make some reasonable assumptions that will help us arrive at knowledge about our parameters of inference in any given model.

This week, we will discuss one form of sampling, namely sampling from the posterior. We will assume that we have arrived at a posterior and will sample from it. It is to demonstrate the principle of sampling. We will only allude to what will become our main engine to support our Bayesian inference, namely Monte Carlo methods and chiefly the Metropolis-Hastings algorithms. We will however try to use today's example to fit our first models in Stan a probabilistic programming languages and brms, a package in R that talks to Stan.

Some simple sampling in the case of Nick

The simple sampling from the posterior comes from chapter 3 of the McElreath book, Rethinking Statistics–a great resource by the way.

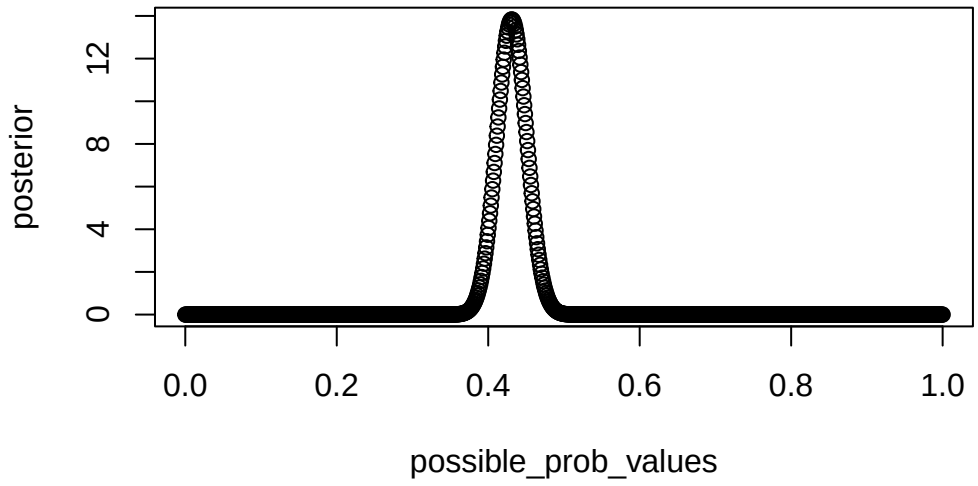
Let's consider the example of Nick the gambler again. We remember 271 heads out of 630 tosses was what our detective learnt about Nick's gambling games. In chpaters 2 and 3 we looked a lot at the likelihood, which as we recall is represented by the binomial probability density function. Please do refer to the formalisms above; here I will only remind us that we can use the dbinom function in base R for it. Let's then build a Bayesian model as we did above, in Week 4.

This code below generates the posterior distribution.

```r
n_heads <- 271
n_tosses <- 630

possible_prob_values <- seq(from = 0 , to = 1, length.out = 10^3) # these are the values  tha

likelihood <- dbinom(n_heads, n_tosses, prob = possible_prob_values)  # our likelihood

prior <- rep(1: length(possible_prob_values)) # this prior is uniform, corresponds to a beta

Bayes_numerator <- likelihood*prior # this is the numerator of Eqs 21 and 22.

posterior <- Bayes_numerator# /sum(Bayes_numerator) # the outcome of Eqs 21 and 22.
```

For me the simplest way of understanding what is going on is the following: to what value of the parameters (i.e. probabilities) does the maximum value of the posterior correspond to? I plot this here.

```r
plot(possible_prob_values, posterior)
```



You can also ask in the way we have before

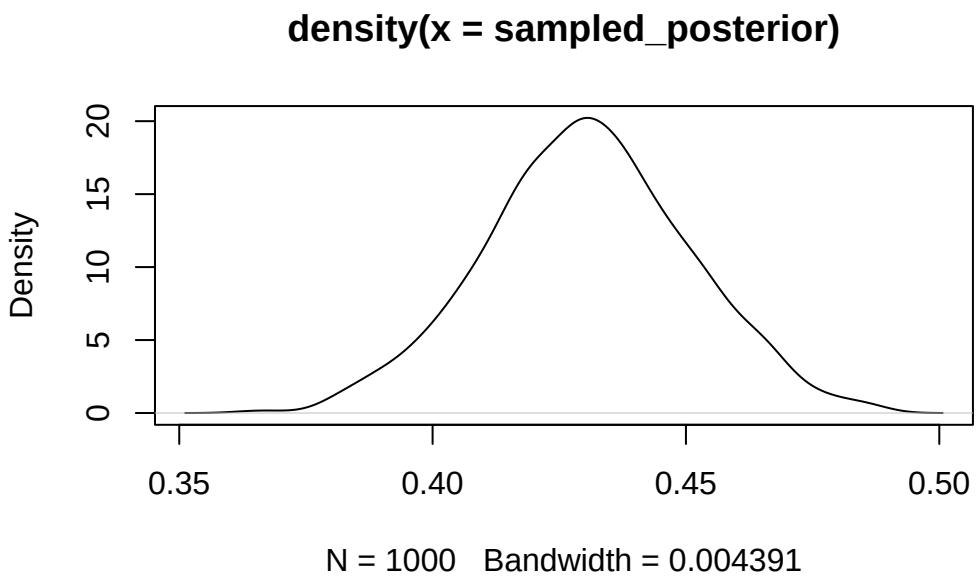
```r
possible_prob_values[which.max(posterior)]
```

```
[1] 0.4314314
```

This number is very close to what we got when we did the maximum likelihood (which should be no surprise, since we used flat priors.

To illustrate that this can also be achieved through a draw 1000 values out of the posterior.

```r
sampled_posterior <- sample(possible_prob_values, # the parameter values to sample from
       posterior, # these are the probabilities that we input
       size = 10^3, # the size of it,
       replace = T # got to do this
       )

plot(density(sampled_posterior))
```



Now this may not sound trivial, but you just did your first sampling from the posterior…

Exercise 6.1

a) Find the values of mean, median, mode and max of the sampled posterior. What do you observe?
b) Instead of sampling from the posterior, sample from the Bayes numerator.
c) Use the sample from (b) to re-run (a), what do you observe?
d) Do (b) and (c) using the likelihood instead of the Bayes numerator.
e) Find the 95% intervals–what would you call those intervals?

Here is a little pointer for 5.1.e

Rethinking: Why 95%? The most common interval mass in the natural and social sciences is the 95% interval. This interval leaves 5% of the probability outside, corresponding to a 5% chance of the parameter not lying within the interval (although see below). This customary interval also reflects the customary threshold for *statistical significance*, which is 5% or $p < 0.05$. It is not easy to defend the choice of 95% (5%), outside of pleas to convention. Ronald Fisher is sometimes blamed for this choice, but his widely cited 1925 invocation of it was not enthusiastic:

> "The [number of standard deviations] for which $P = .05$, or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not."[53]

Most people don't think of convenience as a serious criterion. Later in his career, Fisher actively advised against always using the same threshold for significance.[54]

So what are you supposed to do then? There is no consensus, but thinking is always a good idea. If you are trying to say that an interval doesn't include some value, then you might use the widest interval that excludes the value. Often, all confidence intervals do is communicate the shape of a distribution. In that case, a series of nested intervals may be more useful than any one interval. For example, why not present 67%, 89%, and 97% intervals, along with the median? Why these values? No reason. They are prime numbers, which makes them easy to remember. But all that matters is they be spaced enough to illustrate the shape of the posterior. And these values avoid 95%, since conventional 95% intervals encourage many readers to conduct unconscious hypothesis tests.

Rethinking: What do confidence intervals mean? It is common to hear that a 95% confidence interval means that there is a probability 0.95 that the true parameter value lies within the interval. In strict non-Bayesian statistical inference, such a statement is never correct, because strict non-Bayesian inference forbids using probability to measure uncertainty about parameters. Instead, one should say that if we repeated the study and analysis a very large number of times, then 95% of the computed intervals would contain the true parameter value. If the distinction is not entirely clear to you, then you are in good company. Most scientists find the definition of a confidence interval to be bewildering, and many of them slip unconsciously into a Bayesian interpretation.

But whether you use a Bayesian interpretation or not, a 95% interval does not contain the true value 95% of the time. The history of science teaches us that confidence intervals exhibit chronic overconfidence.[55] The word *true* should set off alarms that something is wrong with a statement like "contains the true value." The 95% is a *small world* number (see the introduction to Chapter 2), only true in the model's logical world. So it will never apply exactly to the real or *large world*. It is what the golem believes, but you are free to believe something else. Regardless, the width of the interval, and the values it covers, can provide valuable advice.

Simplifying posterior sampling

The general point though of Exercise 5.1 is to illustrate something that the Farrell & Lewandowski book mentions in chapter 7, namely equation 7.1. It states that all we need to know to do the sampling is the Bayes numerator, i.e. the likelihood multiplied by the prior.

This is because the evidence, i.e. the denominator (i.e. the marginal likelihood) is a constant in relation to this quantity. Therefore, equation 22:

$$P(\theta \mid y) = \frac{P(y \mid \theta).P(\theta)}{\int_{\theta} P(\mathrm{y} \mid \theta) \cdot P(\theta) d\theta}$$

can be reduced to:

$$P(\theta \mid y) \propto P(y \mid \theta).P(\theta) \tag{23}$$

Where $\propto$ means proportional to, the posterior is proportional to the Bayes numerator, i.e. the likelihood times the prior.

Therefore, we can use algorithms like the Metropolis-Hastings, which we will talk about in more detail in the next few weeks, to arrive at inferences.

For this week though, let's estimate the Nick problem using hardcore Bayesian models…

First off, with brms, the very useful package: *Bayesian Regression Models Using Stan*, an R package that is structured like lme, but allows you to fit a wide range of linear and non-linear models. It uses an Hamilton Monte Carlo estimator, based on similar principles to MHMC that we will be seeing later.

Here is some brms code to run the Nick model.

```
library(brms)

brm(data = list(nheads = 271),
family = binomial(link = "identity"),
nheads | trials(630) ~ 0 + Intercept,
# we are using a flat prior--like the uniform above.
prior(beta(1, 1), class = b, lb = 0, ub = 1),
iter = 2000, warmup = 500, # we will discuss this in more detail above.
seed = 3)
```

Exercise 6.2

a. go through the code above line by line. What is the model, see "nheads | trials(630) ~ 0 + Intercept"

b. run the code above and look at how long it takes, why, what is happening there.

c. assign the model above to an object called "my_first_brms_model". Use summary to view the output. What do you notice about the estimate?

Now let's push the boundaries and try to do all this in Stan.

```
library(rstan)

# Data preparation
data_list <- list(nheads = 271, N = 630)

# Stan model code
stan_code <- "
data {
  int<lower=0> nheads;
  int<lower=0> N;
}
parameters {
  real<lower=0, upper=1> p;
}
model {
  p ~ beta(1, 1);
  nheads ~ binomial(N, p);
}
"

# Compile the model
model <- stan_model(model_code = stan_code)

# Fit the model
fit <- sampling(model,
                data = data_list,
                iter = 2000,
                warmup = 500,
                seed = 3)

# Print the summary of the fit
print(fit)
```

Exercise 6.3

a. Do what you did for Exercise 6.2, but this time for the Stan model.

b. Does the Bayes denominator appear anywhere here?

b. Whose grave is depicted in the photo below. Hint: it is in East London (Photo courtesy of Dr LA)