# 1 Is probability theory well-founded or do you believe in the theory of evolution?

I'll start with the least squares methods through the maximum likelihood estimation; this requires some (at least superficial) knowledge of probability theory.

I was once asked if I believed in evolutionary theory. Take a short break, think about how you would answer. Being puzzled by the question, I have answered that I find it plausible. Scientific theory has little to do with faith. In short, a theory only builds a model of the world, and there is no need to believe in it. Moreover, the Popperian criterion[] requires a scientific theory be able to be falsifiable. A solid theory must possess, first of all, the power of prediction. For example, if you genetically modify crops in such a way that they produce pesticides themselves, it is only logical that pesticide-resistant insects would appear. However, it is much less obvious that this process can be slowed down by growing regular plants side by side with genetically modified plants. Based on evolutionary theory, the corresponding modelling has made this prediction[], and it seems to have been validated[].
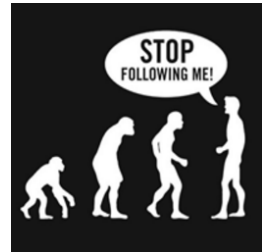
$$P(a) = .4 \qquad P(a \wedge b) = 0 \qquad P(b) = .3 \qquad P(a \vee b) = .8$$

| Bets of the Agent 2 | | Result for the Agent 1 | | | |
|---|---|---|---|---|---|
| Event | Bet amount | $a \wedge b$ | $a \wedge \neg b$ | $\neg a \wedge b$ | $\neg a \wedge \neg b$ |
| $a$ | 4-6 | -6 | -6 | 4 | 4 |
| $b$ | 3-7 | -7 | 3 | -7 | 3 |
| $\neg(a \vee b)$ | 2-8 | 2 | 2 | 2 | -8 |
| | | -11 | -1 | -1 | -1 |

- $0 \leq P(a) \leq 1$

- $P(true) = 1$, $P(false) = 0$

- $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$

In text form, they look like this:

1. All probabilities range from 0 to 1.

2. Certainly true statements have probability of 1 and certainly false probability of 0.

3. The third axiom is the very intuitive disjunction axiom. All cases where the statement $a$ is true, together with those where $b$ is true, cover all those cases where the statement $a \vee b$ is true; however the intersection $a \wedge b$ is counted twice in the sum, therefore it is necessary to subtract $P(a \wedge b)$.

In 1931, de Finetti proved a very strong proposition:

*If a bookmaker is guided by beliefs which break the axioms of the theory of probability, then there exists such a combination of bets of the player which guarantees the loss for the bookmaker (a prize for the player) at each bet.*

Probability axioms can be considered as the limiting set of probabilistic beliefs that some agent can adhere to. Note that if a bookmaker respects Kolmogorov's axioms, it does not imply that he will win (leaving aside the fees), however, if he does not respect the axioms, he is guaranteed to lose. Other arguments have been put forward in favour of the probability theory; but it is the practical success of probability-based reasoning systems that has proved to be very attractive.

# 2 Maximum likelihood through examples

## 2.1 First example: coin toss

Let us consider a simple example of coin flipping, also known as the Bernoulli's scheme. We conduct $n$ experiments, two events can happen in each one ("success" or "failure"): one happens with probability $p$, the

other one with probability $1 - p$. Our goal is to find the probability of getting exactly $k$ successes in these $n$ experiments. This probability is given by the Bernoulli's formula:

$$P(k; n, p) = C_n^k p^k (1 - p)^{n-k}$$

Let us take an ordinary coin ($p = 1/2$), flip it ten times ($n = 10$), and count how many times we get the tails:

$$P(k) = C_{10}^k \frac{1}{2^k} \left(1 - \frac{1}{2}\right)^{n-k} = \frac{C_{10}^k}{2^n}$$

Figure 1, left shows what a probability density graph looks like.
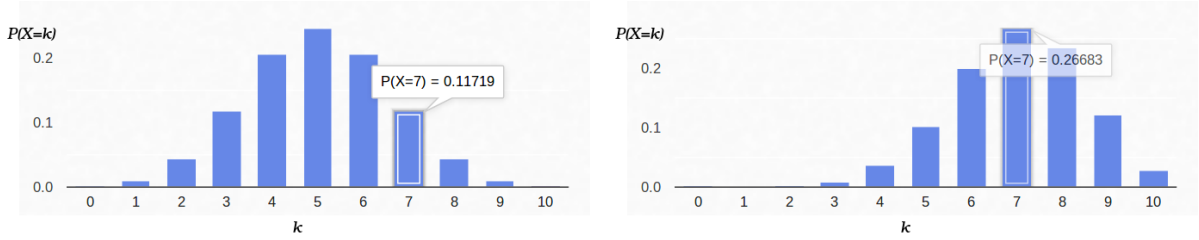


Рис. 1: **Left:** probability density graph for the Bernoulli's scheme with $p = 1/2$. **Right:** probability density graph for the Bernoulli's scheme with $p = 7/10$.

Thus, if we have fixed the probability of "success" ($1/2$) and also fixed the number of experiments (10), then the possible number of "successes" can be any integer between 0 and 10, but these outcomes are not equiprobable. It is clear that five "successes" are much more likely to happen than none. For example, the probability encountering seven tails is about 12%.

Now let us look at the same problem from a different angle. Suppose we have a real coin, but we do not know its distribution of a priori probability of "success"/"failure". However, we can toss it ten times and count the number of "successes". For example, we have counted seven tails. Would it help us to evaluate $p$?

We can try to fix $n = 10$ and $k = 7$ in the Bernoulli's formula, leaving $p$ as a free parameter:

$$\mathcal{L}(p) = C_{10}^7 p^7 (1 - p)^3$$

Then the Bernoulli's formula can be interpreted as the plausibility of the parameter being evaluated (in this case $p$). I have even changed the function notation, now it is denoted as $\mathcal{L}$ (likelihood). That is being said, the likelihood is the probability to generate the observation data (7 tails out of 10 experiments) for the given value of the parameter(s). For example, the likelihood of a balanced coin ($p = 1/2$) with seven tails out of ten tosses is approximately 12%. Figure 2 plots the likelihood function for the observation data with 7 tails out of 10 experiments.
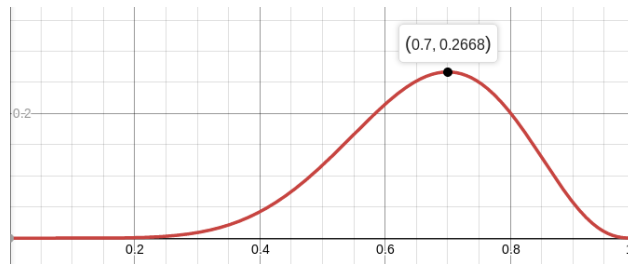


Рис. 2: The plot of the likelihood function $\mathcal{L}(p)$ for the observation data with 7 tails out of 10 experiments.

So, we are looking for the parameter value that maximizes the likelihood of producing the observations we have. In our particular case, we have a function of one variable, and we are looking for its maximum. In order to make things easier, I will not search for the maximum of $\mathcal{L}$, but for the maximum of $\log \mathcal{L}$. The logarithm is a strictly monotonous function, so maximizing both is equivalent. The logarithm has a nice property of breaking down products into sums that are much more convenient to differentiate. So, we are looking for the maximum of this function:

$$\log \mathcal{L}(p) = \log C_{10}^7 + 7 \log p + 3 \log(1 - p)$$

That's why we equate it's derivative to zero:

$$\frac{d \log \mathcal{L}}{dp} = 0$$

The derivative of $\log x = \frac{1}{x}$, therefore:

$$\frac{d \log \mathcal{L}}{dp} = \frac{7}{p} - \frac{3}{1-p} = 0$$

That is, the maximum likelihood (about 27%) is reached at the point $p = 7/10$. Just in case, let us check the second derivative:

$$\frac{d^2 \log \mathcal{L}}{dp^2} = -\frac{7}{p^2} - \frac{3}{(1-p)^2}$$

In the point $p = 7/10$ it is negative, therefore this point is indeed a maximum of the function $\mathcal{L}$:

$$\frac{d^2 \log \mathcal{L}}{dp^2}(0.7) \approx -48 < 0$$

Figure 1 shows the probability density graph for the Bernoulli's scheme with $p = 7/10$.

## 2.2   Second example: analog-to-digital converter (ADC)

Let us imagine that we have a constant physical quantity that we want to measure; for example, it can be a length to measure with a ruler or a voltage with a voltmeter. In the real world, any measurement gives *an approximation* of this value, but not the value itself. The methods I am describing here were developed by Gauss at the end of the 18th century, when he measured the orbits of celestial bodies [1]. [?]

For example, if we measure the battery voltage $N$ times, we get $N$ different measurements. Which of them should we take? All of them! So, let us say that we have $N$ measurements $U_j$:

$$\{U_j\}_{j=1}^{N}$$

Let us suppose that each measurement $U_j$ is equal to the real value plus the Gaussian noise. The noise is characterized by two parameters — the center of the Gaussian bell and its "width". In this case, the probability density can be expressed as follows:

$$p(U_j) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(U_j - U)^2}{2\sigma^2}\right)$$

That is, having $N$ measurements $U_j$, our goal is to find the parameters $U$ and $\sigma$ that maximize the likelihood. The likelihood (I have already applied the logarithm) can be written as follows:

$$\log \mathcal{L}(U, \sigma) = \log \left(\prod_{j=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(U_j - U)^2}{2\sigma^2}\right)\right) =$$

$$= \sum_{j=1}^{N} \log \left(\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(U_j - U)^2}{2\sigma^2}\right)\right) =$$

$$= \sum_{j=1}^{N} \left(\log \left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{(U_j - U)^2}{2\sigma^2}\right) =$$

$$= -N \left(\log \sqrt{2\pi} + \log \sigma\right) - \frac{1}{2\sigma^2} \sum_{j=1}^{N} (U_j - U)^2$$

---

[1]Note that Legendre has published an equivalent method in 1805, whereas Gauss' first publication is dated by 1809. Gauss has always claimed that he had been using the method since 1795, and this is a very famous priority dispute [?] in the history of statistics. There are, however, numerous evidence to support the thesis that Gauss possessed the method before Legendre, but he was late in his communication.

And then everything is strictly as it used to be, we equate the partial derivatives to zero:

$$\frac{\partial \log \mathcal{L}}{\partial U} = \frac{1}{\sigma^2} \sum_{j=1}^{N} (U_j - U) = 0$$

The most plausible estimation of the unknown value $U$ is the simple average of all measurements:

$$U = \frac{\sum_{j=1}^{N} U_j}{N}$$

And the most plausible estimation of $\sigma$ turns out to be the standard deviation:

$$\frac{\partial \log \mathcal{L}}{\partial \sigma} = -\frac{N}{\sigma} + \frac{1}{\sigma^3} \sum_{j=1}^{N} (U_j - U)^2 = 0$$

$$\sigma = \sqrt{\frac{\sum_{j=1}^{N} (U_j - U)^2}{N}}$$

Such a convoluted way to obtain a simple average of all measurements. . . In my humble opinion, the result is worth the effort. By the way, averaging multiple measurements of a constant value in order to increase the accuracy of measurements is quite a standard practice. For example, ADC averaging. Note that the hypothesis of Gaussian noise is not necessary in this case, it is enough to have an unbiased noise.

## 2.3 Third exampe, still 1D

Let us re-consider the previous example with a small modification. Let us say that we want to measure the resistance of a resistor. We have a bench top power supply with current regulation. That is, we control the current flowing through the resistance and we can measure the voltage required for this current. So, our "ohmmeter" evaluates the resistance through $N$ meausrements $U_j$ for each reference current $I_j$:

$$\{I_j, U_j\}_{j=1}^{N}$$

If we draw these points on a chart (Figure 3), the Ohm's law tells us that we are looking for the slope of the blue line that approximates the measurements.
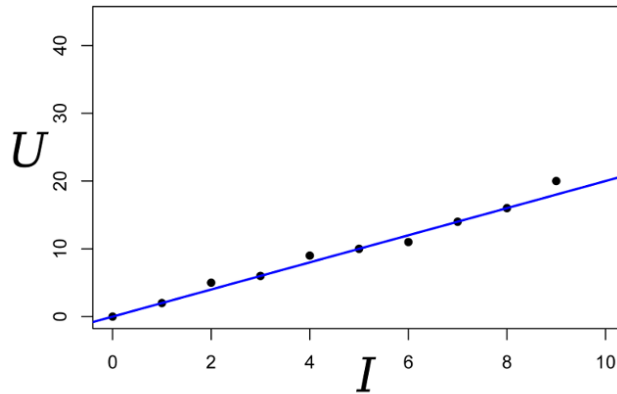


Рис. 3: Having $N$ meausrements $U_j$ for each reference current $I_j$, we are looking for the slope of the blue line that approximates the measurements through the Ohm's law.

Let us write the expression of the (logarithm of) likelihood of the parameters:

$$\log \mathcal{L}(R, \sigma) = \log \left( \prod_{j=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp \left( -\frac{(U_j - RI_j)^2}{2\sigma^2} \right) \right) =$$

$$= -N \left( \log \sqrt{2\pi} + \log \sigma \right) - \frac{1}{2\sigma^2} \sum_{j=1}^{N} (U_j - RI_j)^2$$

As usual, we equate the partial derivatives to zero:

$$\frac{\partial \log \mathcal{L}}{\partial R} = -\frac{1}{2\sigma^2} \sum_{j=1}^{N} -2I_j (U_j - RI_j) =$$

$$= \frac{1}{\sigma^2} \left( \sum_{j=1}^{N} I_j U_j - R \sum_{j=1}^{N} I_j^2 \right) = 0$$

Then the most plausible resistance $R$ can be found with the following formula:

$$R = \frac{\displaystyle\sum_{j=1}^{N} I_j U_j}{\displaystyle\sum_{j=1}^{N} I_j^2}$$

This result is somewhat less obvious than the simple average of all measurements in the previous example. Note that if we take one hundred measurements with $\approx 1A$ reference current and one measurement with $\approx 1kA$ reference current, then the first hundred measurements would barely affect the result. Let's remember this fact, we will need it later.

## 2.4 Fourth example: back to the least squares

$$\{x_j, y_j\}_{j=1}^{N}$$

$$\log \mathcal{L}(a, b, \sigma) = \log \left( \prod_{j=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp \left( -\frac{(y_j - ax_j - b)^2}{2\sigma^2} \right) \right) =$$

$$= -N \left( \log \sqrt{2\pi} + \log \sigma \right) - \frac{1}{2\sigma^2} \sum_{j=1}^{N} (y_j - ax_j - b)^2$$

$$S(a, b) = \sum_{j=1}^{N} (y_j - ax_j - b)^2$$

$$\frac{\partial S}{\partial a} = \sum_{j=1}^{N} 2x_j (ax_j + b - y_j) = 0$$

$$\frac{\partial S}{\partial b} = \sum_{j=1}^{N} 2(ax_j + b - y_j) = 0$$

$$\begin{cases} a \sum_{j=1}^{N} x_j^2 + b \sum_{j=1}^{N} x_j = \sum_{j=1}^{N} x_j y_j \\ a \sum_{j=1}^{N} x_j + bN = \sum_{j=1}^{N} y_j \end{cases}$$

$$a = \frac{N \sum_{j=1}^{N} x_j y_j - \sum_{j=1}^{N} x_j \sum_{j=1}^{N} y_j}{N \sum_{j=1}^{N} x_j^2 - \left( \sum_{j=1}^{N} x_j \right)^2}$$

$$b = \frac{1}{N} \left( \sum_{j=1}^{N} y_j - a \sum_{j=1}^{N} x_j \right)$$