1

Model fitting

Lecture 19. *Thursday 3rd December, 2020. Compiled: Wednesday 30th December, 2020.*

The next step one must make is now **make sense** of data. Therefore, we want to **extract meaningful information** out of it by the mean of a model. However, there are also many **challenges** when collecting data: one should think well about what kind of data is needed, as well as the measurement process and its interpretation. We are going to analyze these arguments using as paradigmatic example COVID-19, but our considerations can actually be applied to *any* spread process.

1.1 Data Collection

When dealing with epidemic processes one usually speaks about **incidence data**, like the one depicted in figure 1.1.

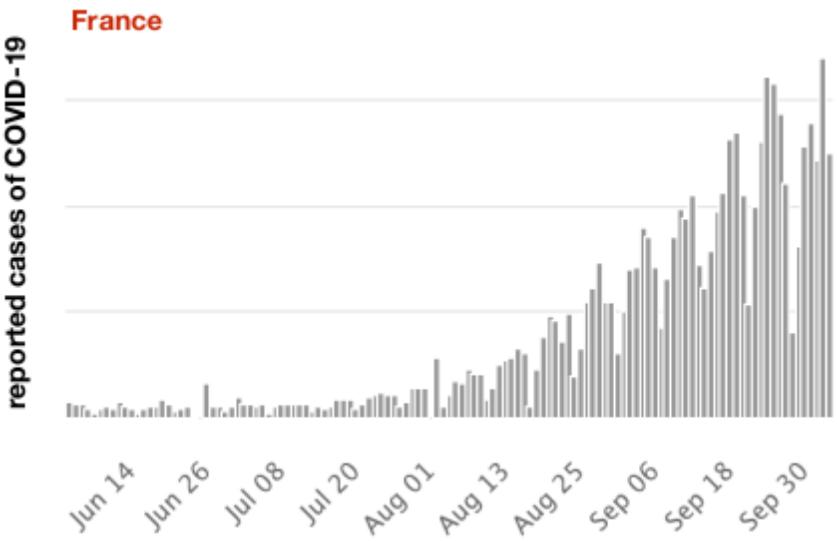


Figure 1.1: COVID-19 curve incidence data in France. For each day, number of reported new cases are shown.

We want to **model** these data, so in other words we want to make **make sense** out of it by interpreting raw numbers. In order to do it, we need first to understand how it was collected and, recalling that data actually gives **partial information**, this must be *completed* by the means of a **model**.

The main **goals** one wants to pursue by using models are:

- **nowcasting**: provide assessment regarding the *present* and *close future* of the epidemics. Therefore, one may need to understand what is and what will be the extension/distribution by groups/regions of the epidemic that is present at that moment, given the partial information returned by data.

- **forecasting**: prediction in *longer term*. For instance, one would like to predict hospitals occupancy, when and how high will be the epidemic peak, how many people will be infected over the next weeks or months, when epidemic will end and what will be its final size.

- **medical and biological understanding**: at the very beginning we have no medical/biological knowledge about the epidemic. Hence, we want to study for instance how it propagates and through what vectors (human-to-human, zoonotic, vector-borne, direct transmission, fomite, aerosol, droplets, etc...), the role of asymptomatic/pre-symptomatic in transmission, susceptibility and rate of symptoms by age group.

- **exploration of counterfactuals and hypothetical scenarios**: we run our model to perform *scenario analysis* and understand what is the best strategy to use in the future. It may be related to vaccination, pharmacological interventions, lockdown, travel restrictions and their impact on the future spreading. Indeed, these are *long-term* projections, despite one may want to explore the case and what would have occurred if a decision had not been made. Hence these arguments are valid also for the past: for instance we want to quantify the impact of lockdown in spreading.

Let us discuss **what** is the *data* we usually work with. When we speak about **incidence in a given area and at time** t, we refer to the **fraction** of population resident in that area that has contracted the disease at time t. Hence, formalizing[1]:

$$\textbf{incidence} = \frac{\text{number of people hit by flu}}{\text{population at risk}} \tag{1.1}$$

Let us take a look closer to the **numerator**. Obviously it is impossible to have a complete information about how many people have the flu at this moment. But, first, one has to face the first problem of **case definition**: set of criteria used in making a decision as to whether an individual has a disease or any other health event of interest. Some possible *criteria* may involve: clinical (e.g. symptoms), laboratory characteristics (e.g. exams, test results). Moreover personal information are taken into account, such as whether this individual travelled to regions at risk/had contact with people at risk can be classified using three levels: confirmed, probable, possible. Cases definition can be either more *sensitive* or *specific*, and it has to be tuned according to the risk assessment. A **sensitive** case definition will detect many cases but may also count as cases individuals who do not have the disease (*possible overestimation*). On the other hand, **specific** case definition is more likely to include only persons who truly have the disease under investigation but also more likely to miss some cases (*possible underestimation*). This is summarized in Fig. 1.2.

	Disease is truly present	Disease is truly absent	Total
complies to case definition	a	b	all cases
does not comply to case definition	c	d	all non-cases
	all 'diseased'	all 'non-diseased'	all people in the study sample

Sensitivity = [a / (a+c)]

Specificity = [d / (b+d)]

Figure 1.2: Case definition can be either more sensitive or specific. In the first case we may end up overestimating the number of cases, whereas when a test is more specific this might lead to underestimation.

[1] We will use *flu* as example, since we have much data of it

Let us see **how** data is **collected**: in France there is a network of *General Practitioners* (see Fig. 1.3) that are volunteers and daily send a report to Health Minister concerning all the cases they have visited during a workday. However, for some diseases, for instance measles, every family doctor is obliged to report the case. Let us continue analyzing the *flu* case. The number of cases reported is indeed the number of cases seen by General Practitioners defined under the basis of some clinical criteria. These are *possible cases*: the guarantee can be returned only after a laboratory confirmation which is available only for a small proportion of cases.

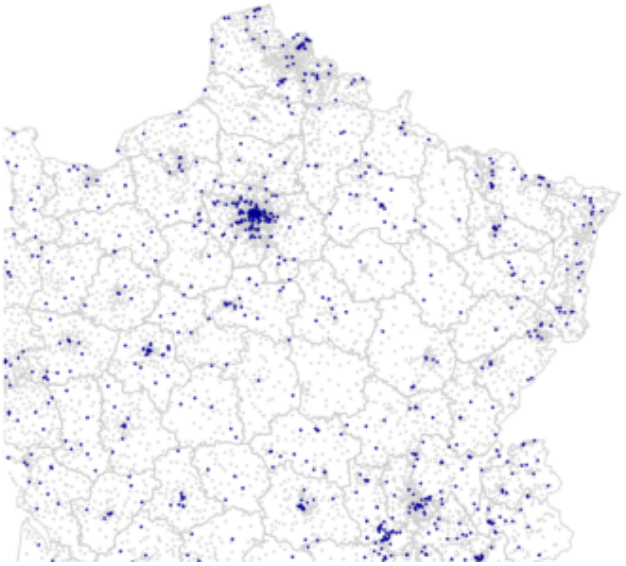


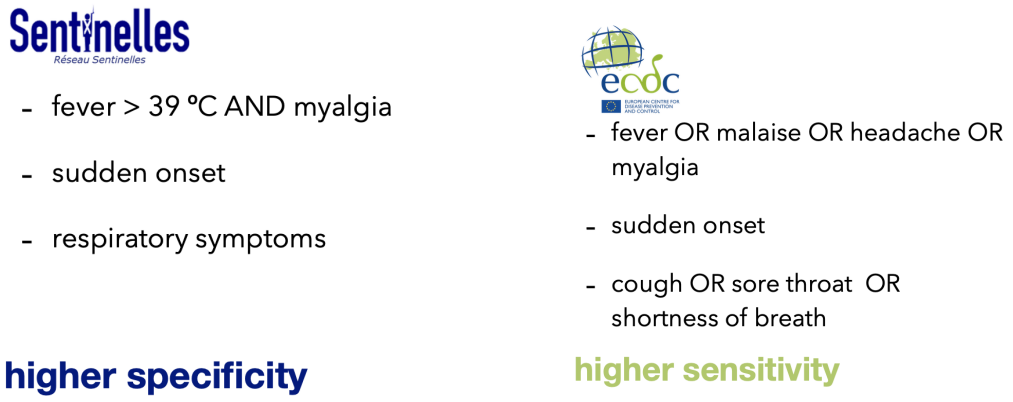
Figure 1.3: The Surveillance Network in france (SN) is based on a fraction of General Practitioner ($\sim 1\%$), who are volunteers.

Given the symptoms of flu:

- no symptoms ($\sim 30\%$);
- upper respiratory symptoms, e.g. nasal stuffiness, runny nose, sore throat, sneezing, hoarseness, ear pressure, or earache ($\sim 60\%$);
- lower respiratory symptoms, e.g. cough, breathing difficulty, and chest discomfort ($\sim 2\%$);
- fever ($\sim 35\%$);

The main concept is that it is important to understand according to what criteria data is collected in order to deal better with observables, indeed observables are a *proxy* for the real data. One should note that they may be different even if they regard the same quantity (look Fig. 1.4): for $ECDC$ incidence data we might observe a peak in autumn because of respiratory infections, whereas this might not be present in *Sentinelles* reports, given they do not classify it as case of flu. The $ECDC$ case definition, as one can imagine, has higher sensitivity and this can return overestimation of the number of cases. Conversely, according to *Sentinelles* cases definition we might underestimate their number.

Let us take a look closer to the **denominator**. It refers to the *catchment population*, i.e. all the people living in the catchment area of the General Practitioner reporting the cases, who would usually seek healthcare at the site when they get sick. Therefore, the denominator for the area a that can be computed at a first



(a) "Sentinelles" clinical case definition for flu.

(b) "ECDC" clinical case definition for flu.

Figure 1.4

approximation is[2]:

$$\text{denominator}_a = \text{Population}_a \frac{GP_{SN,a}}{GP_a}$$

where the ratio $\frac{GP_{SN,a}}{GP_a}$ is the proportion of General Practitioners that contribute to the *Surveillance Network* ($\sim 1\%$ according to Fig. 1.3). Moreover, another **problem** that biases our observable is given by the **consultancy rate**. Since many people are asymptomatic or paucisymptomatic, the rate of people going to be examinated by family doctor is highly variable by age: young people, except very little children, are more likely to not go, whereas adults need to go in order to have permission to stay home from work. Raw numbers depend also on family doctors density, on the health-care system (how expensive is going to the GP), and on the period of the year that brings specific diseases. In conclusion, even though data might look simple at a first glance, dealing with it needs to take into account many variables all together and some assumptions are more likely to be made if data is not available. Another important point is that the **confirmed flu cases** are a very small subset among *I*nfluenza *L*ike *I*lness (ILI) people (symptomatic), people that go to General Practitioner, Infected people which either can be detectable or not and that, obviously, need to be recorded as infected.

Another *characteristic* of the **case definition** is that it might be **variable in time**, specially when the range of symptoms is still unknown. It was the case of **COVID-19** at the very beginning. In addition, case definition is **matter of authority**: once cases are reported in hospitals, some papers concerning viral loads, symptoms, evidences are published. Health authority needs to collect and through them in order to define better the case definition, keeping in mind a sort of trade off: if case definition is *high sensitive*, there might be false positives and also cause panic among people. Conversely, if the definition is *too much specific*, we risk to let infectious people go around and spread the infection. This tuning depends on the goals one may want to pursue.

With regards to **COVID-19**, case definition was therefore varying in time being the range of symptoms unknown. Moreover, more problems arose since the disease at the *very beginning* was not wide spread and it was still unclear the region where it was spreading: the denominator related to the catchment population was kind of difficult to estimate at that time. The *reporting rate* was highly variable in time: at the beginning, the tracing system is able to intercept all the case, but the surveillance

[2]Horvitz DG, Thompson DJ. A JASA. 1952;47:663–85

system might saturate even though the case definition remains unchanged. Moreover, due to the change in time of case definition, number of cases can be always retrospectively corrected: real time analysis numbers are indeed biased and this is why sometimes we observe spikes in incidence curves.

1.2 Epidemic Modeling and Bayesian Inference

Typically, the steps are the following ones:

- **Model design/implementation**: decide the model ingredients that synthesise available medical, biological, information etc. We can consider different models, ingredients that describe our **hypotheses**. These can be for instance helpful to simplify our problem and must be *clearly* stated at the beginning.

- **Model calibration**: estimate model parameters from available data, a.k.a. model *fitting*

- **Model validation**: confirming that model output is sufficiently accurate in reproducing the data. It is done as the result of *model calibration*, when we can tell whether our model well represents data also taking into account secondary aspects.

Let us introduce now some concepts about **bayesian inference** and **Maximum Likelihood** that will help us throughout this process. Once we observe data, one wants to introduce an **Observation model** $\mathcal{O}$ according to which we assume to have acquired our data. In our case, we assume it to be a *Binomial* process with probability p: every case, according to a probability p goes to the doctor. In addition we need to define an **Epidemic Model** (e.g. SIR, or alternatively $SEIR$, $SIRS$...) and a **vector of parameters** $\vec{\theta} = (\beta, \mu, ...)$ where some are to be inferred, while others are assumed to kept fixed. This obviously depends on our data and on our **Initial Conditions** I_0, namely the number of infectious. All these quantities help us to define the curve represented in Fig.1.5.

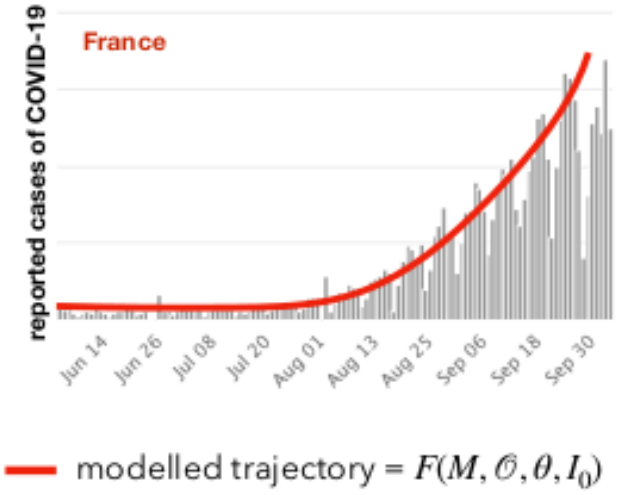


Figure 1.5: Curve fitting the incidence curve presented before (see Fig. 1.1).

We therefore use a **Maximum Likelihood approach** that returns from a probabilistic formulation: indeed the relation between model and data is probabilistic, and our aim is to identify the trajectory and thus the parameters, $\vec{\theta}$, that are **more**

probable given the data we have. One should note that, in the *bayesian framework* probability is used as a **measure of uncertainty**.

When we are dealing with single random variable A we talk about *univariate probability*: the probability that A takes value a is defined as $p(A = a) = p(a)$ and the normalization condition holds $\sum_a p(a) = 1$. For *continuous variables* it can be rewritten as $\int p(a)\,\mathrm{d}a = 1$. When random variables are more than one, e.g. A, B, the joint probability that A takes value a and B takes value b is written as $p(A = a, B = b) = p(a, b)$. The marginal probability $p(a) = \sum_b p(a, b)$ and is the probability that A takes value a regardless b: we indeed summed over all the possible $p(b)$. For *continuous variables* it can be rewritten as $p(a) = \int p(a, b,)db$

Some of the **basic properties** we will deal are the following:

- **Conditional probability** of a from random variable A, given that the outcome of a random variable B was b is $p(A = a|B = b) = p(a|b)$.
- **Bayes Theorem** allows us to rewrite the conditional probability as follows $p(a|b) = \frac{p(a,b)}{p(b)}$.
- **Chain rule**: $p(a, b, c) = p(a|b, c)p(b|c)p(c)$.

We are going now to introduce some variables and a short overview over statistical inference. As said, the latter helps us in drawing conclusions from numerical data about quantities that are not observed: for instance we see that a disease is more frequent in adults and we want to infer its prevalence in children population. Some of these unobserved quantities can be $\tilde{\mathbf{y}}$: potentially observable quantities such as future observations of a process (e.g. predictions), and $\tilde{\theta}$ that are quantities not directly observable such as *parameters* that govern hypothetical process. **Bayesian statistical conclusions** about a parameter θ or unobserved data $\tilde{y}$ are made in terms of *probability statements*. These are expressed as **conditional** probabilities on the observed values of y: $p(\theta|y)$.

In other words, we want to obtain a distribution for θ conditioned to y: $p(\theta|y)$. In order to pursue our goal:

- we need a model (M) that provides us the joint probability distribution of θ and y: $p(\theta, y)$;
- given the model M, thanks to Bayes' Theorem we can write $p(\theta, y) = p(y|\theta)p(\theta)$, with $p(\theta)$ that is the **prior distribution** and $p(y|\theta)$ that is the sampling distribution;
- we use the Bayes rule to *condition* on the known value of the data y, namely:

$$p(\theta, y) = p(\theta|y)p(y) = p(y|\theta)p(\theta)$$

The **unnormalized posterior density**, namely the expression that helps us inferring the parameters we need, is:

$$p(\theta|y) \propto p(\theta)p(y|\theta) \tag{1.2}$$

It is unnormalized since we do not care about the normalization term: it is constant.

One should note that data affects the posterior **only** through $p(y|\theta)$. If we keep fixed y and let θ vary, this is the **Likelihood function** $\mathcal{L}(\theta) = p(y|\theta)$

Example 1: Hemophilia

Let us consider now an **example**. *Hemophilia* is a hereditary disease associated to a gene of the chromosome X. This is recessive inheritance: a man who inherits the gene is affected, a woman who inherits the gene on only one X is not affected. Let us recall, for the sake of completeness, that a man has chromosomes XY, while woman XX.
We want to deal with the following **problem**: given that a woman has an affected brother and a father not affected, she can be a carrier of the gene on either one X. We want to estimate whether she is a carrier, and we define $\theta = 1$ as the situation where she actually is, while $\theta = 0$ describes a situation where she is not. Since we do not have *any* other information, **a priori** one should not introduce any bias, hence $p(\theta = 1) = p(\theta = 0) = 0.5$.
Our empirical **data** consists on the fact that she has got two sons, and neither of the two is affected: $y_1 = 0$ and $y_2 = 0$. The **Likelihood** is therefore:

$$p(y_1 = 0, y_2 = 0|\theta = 1) = 0.5^2$$
$$p(y_1 = 0, y_2 = 0|\theta = 0) = 1$$

Multiplying these terms, we can obtain the **posterior**, namely:

$$p(\theta = 1|y_1, y_2) = \frac{p(y_1, y_2|\theta = 1)p(\theta = 1)}{p(y_1, y_2|\theta = 1)p(\theta = 1) + p(y_1, y_2|\theta = 0)p(\theta = 0)}$$
$$= \frac{0.25 \cdot 0.5}{0.25 \cdot 0.5 + 0.5} = 0.2$$

So the probability that she is a carrier, given our observation is quite low.
However, most of the times, it might happen that **new data is available**: for example the same woman has a third son, which is not affected $y_3 = 0$. Obviously, one does not want to lose all the information obtained so far, hence we **update the prior**. The **prior** becomes:

$$p(\theta = 1) = 0.2, \qquad p(\theta = 0) = 0.8$$

which is the posterior of before. The **Likelihood** follows the same argument as before $P(y_3 = 0|\theta = 1) = 0.5$. Out of these expression we can compute the **posterior**, which is:

$$p(\theta = 1|y_3) = \frac{p(y_3|\theta = 1)p(\theta = 1)}{p(y_3|\theta = 1)p(\theta = 1) + p(y_3|\theta = 0)p(\theta = 0)} = \frac{0.5 \cdot 0.2}{0.5 \cdot 0.2 + 0.8} = 0.111$$

Indeed, the probability for the woman to be a carrier is even lower. We want to stress once again that, updating the posterior, we have not lost any information that was previously obtained.

Example 2: Bernoulli trial

Let us consider **another example**, with *Bernoulli trials*. Recalling that for a **Binomial distribution** we have n independent trials with two possible complementary outcomes (either failure or success) and we observe y successes. The probability of a success is θ, consequently for a failure is $1 - \theta$ and this is the parameter one may want to estimate. For instance, given a number n of observations with y successes, we want to infer whether a coin is fair $\theta = 1/2$.
In the case, we do not have any information so we can use a **uniform prior**

for the paramterθ: $\mathcal{U}[0,1]$. The **Likelihood** is a *Binomial distribution* with parameters:

$$P(y|\theta) = \texttt{Bin}(y|n,\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$$

Note as we do not write the dependence on n on the left side because is part of the experimental design and considered fixed. All probabilities will be conditional on n. The **posterior** becomes:

$$p(\theta|y) \propto \theta^y(1-\theta)^{n-y}$$

that is nothing more that a *Beta* distribution Beta$(y+1, n-y+1)$. Note that $\binom{n}{y}$ does not depend on θ therefore can be disregarded.

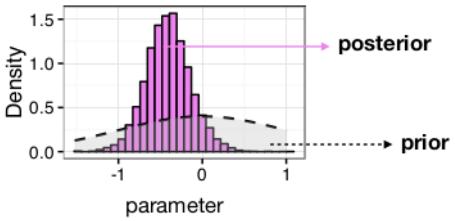


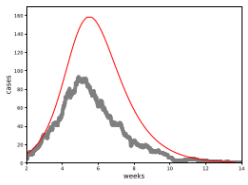
Figure 1.6: When we work with Bayesian inference we usually want to infer the most likely value for a parameter thanks to the *posterior*, given the observation (*data*) and our previous knowledge (*prior*).

Shortly, the **prior distribution** summarises my a priori knowledge about parameters. It might be defined based on the literature, for instance if we are analysing an outbreak of flu and our goal is to estimate R_0, we may want to look at previous R_0 estimates. If we have no prior knowledge on the problem, however, the best idea is to use a vague, or flat, *noninformative* prior. Instead, the **posterior distribution** is a compromise between data and prior information. Such compromise is increasingly controlled by data as the sample size increases. Posterior *variance* on average is smaller than prior variance: if it occurs, then this denotes either a conflict or an inconsistency between sampling model (i.e. data we obtain) and the prior distribution. The **main information** one wants to obtain from the posterior are:

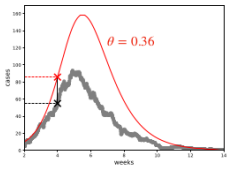
- **Mode** of the posterior, i.e. the most likely parameter given the data.
- **Uncertainty** associated to our estimate, i.e. the $C.I.$ (credibility interval) and usually it is given by the 2.5% and 97.5% quantiles. It is really a relevant quantity: the range according to which the mode spans can lead to really **different** and **opposite outcomes**.

Let us discuss now a more practical problem. We want now to understand how to **fit an incidence curve**, such as the one in Fig. 1.7a where all symptomatic cases that go to the doctor are detected, to estimate R_0. In other words, we have an incidence curve and want to fit a SIR model. The steps to follow are:

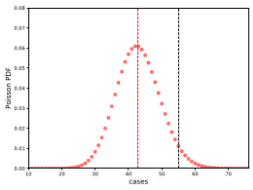
- **Data**. We know that the infection causes *symptoms* for the 50% of cases. We assume that all symptomatic cases go to the doctor and are detected.
- **Observation model**: weekly cases are independently detected with probability $d = 0.5$. Observations y_t are independent. This process is binomial: each case has the 50% to be detected. However, for large numbers, it can be approximated as a Poisson.



(a) Raw data: incidence curve.



(b) Prediction of *total* number of cases for $\theta = 0.36$, i.e. $I(t)$.



(c) Prediction of *detected* number of cases for $\theta = 0.36$ at fourth week, i.e. $y_{t=4}$.

Figure 1.7

- **Model**: SIR, from the literature and peer review journals we know that the average infection duration is $\mu^{-1} = 5.5$ days. We can assume this parameters to be fixed.
- **Parameter** we want to fit is $\theta = \beta$
- **Prior** uniform in $[0, 1]$, since we do not have any information on β.
- **Likelihood** $\mathcal{L}(\theta) = p(y_1, , ...y_t, ..., y_{t_M}|\theta) = \prod_t p(y_t|\theta)$, cases at t-th week are denoted by y_t

Then, for each value of β (our θ to be inferred) we run a simulation of the trajectory of the SIR, fixing μ and I_0 based on available knowledge. Practically, we can approximate observed data to be distributed as a Poisson(λ), where $\lambda = I(t)d$. This is the number of cases we see at t-th week times the detection probability. The total number of cases is distributed as the red line in fig. 1.7b. This returns us the model projection related to the observation.

For instance, at week $t = 4$, we observe $y_{t=4} = 55$ cases. If we assume data is distributed as a Poisson, the sampling distribution is $y_{t=4}|\theta \sim$ Poisson(λ), then $\lambda = I(t)d = 86 \cdot 0.5 = 43$ we expect 43 detected cases. Keeping the parameter θ fixed, we have to compute the likelihood as $\mathcal{L}(\theta = 0.36) =$ Poisson$(55|\lambda)$. This procedure has to be done for every value of the parameter θ. Moreover, one should take into account that since dealing with products is uncomfortable because of really small numbers and for computational simplicity, we take the logarithm[3] of the likelihood, therefore considering the sum: $\log \mathcal{L}(\theta) = \sum_t \log\ p(y_t|\theta)$.

Let us summarize the **basic idea** behind likelihood computation: we want to evaluate the probability of the data given the model and the parameters. In order to **estimate** θ we keep the model M and x_0 fixed and vary θ to compute the probability $p(y|\theta)$. The Likelihood function is $\mathcal{L}(\theta) = p(y|\theta)$, and generally it can span a wide range of orders of magnitude, which can lead to numerical problems. In practice it is better to work with the log-likelihood: $\log \mathcal{L}(\theta) = \log\ p(y_1, ..., y_n|\theta) = \sum_i \log\ p(y_i|\theta)$. The best estimate for θ is actually the one that maximizes the posterior, i.e. the **mode**.

[3]It can be done since $\mathcal{L}(\theta)$ is monotone, without any losing of generality.