Resumé on Maximum Likelihood Estimation (MLE)

Author: Igor Francetic, MScE Student, email: igor.francetic[at]unil.ch

Contents

1 Introduction

The aim of this brief resume is to highlight main futures of maximum likelihood estimation (from now on MLE). This work is based on M. Verbeek's book "A guide to modern econometrics"[1]. The author is solely responsible for all statements made in his work.

2 Distribution of random variables

The MLE method is based on assumptions about the (joint) distribution of one (or more) random variable(s) (from now on, r.v.). In this chapter are treated the ways to characterize the distribution of a r.v. and some tipps to manipulate them. The distribution of a r.v. can be visualized in a few different ways. Obviously, we need to discriminate between discrete and continuous r.v.'s since they do not behave the same way. As in the differential/integral calculus, the discrete case is simplier.

2.1 The discrete case

For a discrete r.v. (i.e. has a finite number of possible realizations), the pmf links the probability associated with each possible realization to the realizations themselves. For a fair die thrown one time, we have $f(y) = P(Y = y) = \frac{1}{6} \forall y \in \{1, 2, 3, 4, 5, 6\}$, where Y is the r.v. itself and y are the possible realizations (1, 2, 3, 4, 5 and 6 for a six faced die). Figure 1 illustrates pmf for the single die.

For some other discrete variables (e.g. daily logreturns of Novartis stockprice), if we plot frequency over different ranges of return, we would probably observe the well known bell shaped curve associated to the Normal distribution. Frequencies can be translated into relative frequencies (or probabilities in the sample) dividing by the total number of observations (in the example, 50). Figure 2 illustrates such an example. Another intersting example is the case of a one shot two dice throw (i.e. 36 possible combinations). If we look at the sum of the dice faces, in figure 3 we see how the Normal comes in and changes the distribution of figure 1. Note that if we sum the probabilities associated to the possible outcomes we obtain 1, since all possible outcomes are taken into account. In other terms:

$$\sum f(y) = 1$$

[1] Verbeek M. (2012), "A guide to modern econometrics", John Wiley and Sons, UK.

Finally, one last important concept is the cumulative distribtion function (cdf), wich simply represents the total frequencies adding them across categories. Formally, the cdf for a discerete r.v. is expressed as

$$F(y) = P(Y \leq y)$$

and is stepwise. In figure 4 we find the cdf associated to the pmf in figure 3.

2.2 The continous case

For a continous r.v. (i.e. has an infinite number of possible realizations) the pdf embodies, in terms of density, the probability of a given range of outcomes. The concept can be expessed formally as follows:

$$P(a < Y < b) = \int\limits_a^b f(y)dy$$

where $f(y)$ is the pdf.
In fact, by definition, even if the r.v. is defined in an interval (eg. continous values between 0 and 1) in a continous setting there is an infinite number of possible outcomes and thus the probability of one particular outcome is zero. Therefore, we can only look at certain intervals (say from a to b) and integrating the pdf over that interval yields the probability of an outcome falling there.
Note that here the rule for discrete probabilites (section 2.1) switches to an integral rule instead of a sum rule. In fact, if we integrate over the whole range of the r.v. (in the example above, the interval $\{0, 1\}$, in general over $\{-\infty, +\infty\}$) we get 1. In other words, in our example:$\int_0^1 f(y) = 1$. Finally, by definition, $f(y) \geq 0 \, \forall y$.

In the continous case, the cdf is formally expressed as

$$F(y) = P(Y \leq y) = \int_{-\infty}^{y} f(t)dt$$

The cdf is important to understand the pdf because the latter is the derivative of the former, $F(y)' = f(y)$. Moreover, it has some nice properties:

1. left limit (approaching $-\infty$) equals 0 while right limit (approaching $+\infty$) equals 1;
2. is nondecreasing in y;
3. is right-continous.

Figure 5 shows cdf and figure 6 the pdf of a uniform distribution (defined over the continous interval $\{a, b\}$ wich could be our $\{0, 1\}$ case).

2.3 Independence and joint distributions

One important result that we may want to remember is about conditional probabilities. If we have two events (say A and B) and want to find the probability of A given that B occoured (we write it $P(A|B)$), the calculation goes as follows: $P(A|B) = \frac{P(A \cap B)}{P(B)}$.

For example, if we throw a die and know that we got an even number (event B), what's the probability that we get a 6 (event A)? $P(A \cap B)$ is equivalent to $P(A)$ because 6 is an even number and it's probability is $\frac{1}{6}$. On the other hand, $P(B) = P(\frac{\{2,4,6\}}{\{1,2,3,4,5,6\}}) = \frac{1}{2}$. Finally, $P(A|B) = \frac{1/6}{1/2} = \frac{1}{3}$. In this case, obiously, there's a relationship between the two results, since if we know that the number is even, the chance of get a 6 grows compared to the simple one shot throw.

Bayes' rule, then, tells us that to switch conditioning from $P(A|B)$ to $P(B|A)$ we simply have multiply by the quotient of the two probabilities, namely $\frac{P(B)}{P(A)}$.
Mooving on to independence, we say that two events are such if and only if $P(A \cap B) = P(A) * P(B)$. In terms of set theory, if A and B are independent their two sets are disjoint and non overlapping. Thus, if we return to the the conditional probability formula, it's easy to prove that $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)*P(B)}{P(B)} = P(A)$.

If we apply the concepts above to the joint distribution of two r.v.'s, we can finally state that **if two r.v.'s are independent, their joint distribution (pdf) is the product of their pdf's.**

Formally, for any two or more independent r.v.'s y_i, $f(y_1, y_2 ... y_N) = \prod\limits_{i=1}^{N} f(y_i)$.

If the two r.v.'s are not independent, applying the first rule in section 2.3, from $f(y|x) = \frac{f(y,x)}{f(x)}$ follows that the joint distribution $f(y, x) = f(y|x) * f(x)$.

3 Likelihood and loglikelihood functions

MLE is all about maximizing what is called a likelihood function or, equivalently, it's log-transformation (log-likelihood)[2]. Thus, we need at least to understand what is a likelihood function and eventually why we may want to take the log-transformed function for our estimation. That's the purpose of chapter 3.

3.1 Introductory example

Loosely speaking, the likelihood function tells us how likely to be thrue something (say the parameter θ) is, given a set of observations (x_i) on the phenomenon.

The following example is taken from Verbeek's book and it's offered in a discrete setting. Suppose we have a box full of red and white balls and are interested in the ratio of red balls (θ). To have a guess about that ratio, we draw a sample of N balls from the box, and each time we pick a ball we keep track of the color and then replace it in the box and shake it (this way, the sample results to be independently drawn, i.e. the probability to get a red ball is not affected by the previous draw). Suppose that in general we keep track of the observation as follows:

- $x_i = 1$ if the ball is red
- $x_i = 0$ if the ball is white

This way, it's easy to state that overall $P(x_i = 1) = \theta$. Moreover, in the sample, $\sum x_i = N_1$ where N_1 is the number of red balls and $N - N_1$ is the number of white balls.

The same sample in terms of white and red balls can show up in some different ways. The probability of a given sample is a function of both, the (unknown) parameter and the sample itself. For example, suppose that:

1. True parameter $\theta = 0.5$
2. $N = 3$
3. $N_1 = 1$

You have the following ways to get the specific sample (B1= ball number 1, S1= sample number 1):

	B1	B2	B3
S1	1	0	0
S2	0	1	0
S3	0	0	1

We don't care about the order of appearance, but from the previous introduction we know that in case of independent events, the joint probability of the events (all events occour) is just the product of the single probabilities. In our case, the joint probability of the sample with 1 red and 2 white balls is given by: $0.5 * 0.5 * 0.5 = 0.5^1 * 0.5^2$ (I divided it in to pieces in order to highlight the probability of different colours, since they're equal).

For any sample size (N) and sum of red balls in the sample (N_1) , with an unknown θ, the formula becomes:

$$P(N_1, N - N_1) = \theta^{N_1} * (1 - \theta)^{N-N_1}$$

[3] and is called the likelihood function θ or $\mathcal{L}(\theta)$.

Finally, if we apply a log transformation the the likelihood function we get the loglikelihood, wich in our case is[4]:

$$log\mathcal{L}(\theta) = N_1 log(\theta) + (N - N_1) * log(1 - \theta)$$

Figure 7 displays the loglikelihood function of θ for the example above.

Conceptually we can relate probabilities and likelihood as follows: $\mathcal{L}(\theta|x) = \mathcal{P}(x|\theta)$. That is, the likelihood of θ given the sample x is nothing but the probability of the sample realization for exactly that value of θ. This is particularly clear in the example above, where we constructed the likelihood of the sample realization given the parameter.

[2] Note that "log" has meaning of natural logarithm, wich is a monotonic transformation and preserves the characteristics of the function, thus maximizing the likelihood is the same as maximizing the log-likelihood.

[3] Note that the independence of the draws plays a crucial role for the statement.

[4] We applied some simple properties of the log:

1. log(a*b)=log(a)+log(b)
2. log(a^b)=b*log(a)

3.1.1 ML estimation for example 3.1

To obtain the MLE estimator for θ we need to maximize the function with respect to θ wich corresponds to set the first partial derivative to zero (FOC's) as follows:

$$\arg\max_{\theta} log\mathcal{L}(\theta|x) \rightarrow N_1 * \frac{1}{\theta} - (N - N_1) * \frac{1}{1-\theta} = 0$$

$$N_1 - N_1\theta = N\theta - N_1\theta \rightarrow \hat{\theta}_{MLE} = \frac{N_1}{N}$$

Then, to be sure that the point found is a maximum, we need to check the sign of the second partial derivative of the expression above (SOC's) wich needs to be stricktly negative.

The MLE estimation in our example unsurprisingly turns out to indicate that, given a random sample, the best guess about the ratio of red balls in the box is the ratio of red balls in the sample. The same key concept applies basically to all situations, within the necessary assumptions framework.

Finally, note that to maximize the log-likelihood we set the first partial derivative of the function with respect to the parameter of interest to zero. For some unknown parameter λ given the sample x, we define $\frac{\partial log\mathcal{L}(\lambda|x)}{\partial\lambda}$ (the gradient of the log-likelihood with respect to the parameter λ) as the score vector $s(\lambda|x)$ wich is also the sum of single contributions to the score $\sum s(\lambda|x_i) = \sum s_i(\lambda)$. To get the ML estimate we set the score to zero.

3.2 General approach and continous variables

Section 3.1 was about a simple discrete case. Here we try to extend the concept to any kind of variable, including the continous case. A key concept here is the shape of the distribution of our unknown parameter, conditional on the sample. In 3.1 we were able to write down the true distribution of independent draws with replacement and derive $\mathcal{L}(\theta|x)$, starting conceptually from the equivalent expression $\mathcal{P}(x|\theta)$[5]. In general is not always possible to derive such an easy expression (e.g. if we have a continous variable, where the probability of a single value is zero), and thus we need to make some assumptions, wich may seem restrictive but finally allow to get a full set of informations.

Typically, we can assume that our unknown parameter λ, conditionally on the sample x, is normally distributed with some mean and variance: $\lambda|x \sim \mathcal{N}(\mu, \sigma^2)$. Since we know the pdf of a Normal distribution, we can write down the likelihood function at least for an independent and identically distributed sample (iid). A widely used example is the standard OLS linear model where we assume a Normal distribution for the error term $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

With normal and independent error terms and usual assumptions we know that:

- $E(y_i|x_i) = \beta x_i$
- $V(y_i|x_i) = \sigma^2$

Thus, we're ready to write down the likelihood. For a single observation we have[6]

$$f(y_i|x_i, \beta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{(-\frac{1(y_i - \beta x_i)^2}{2\sigma^2})}$$

For N independent observations, we apply the product rule (section 2.3) in order to obtain the joint distribution:

$$f(y_1, ..., y_N|x_i, \beta, \sigma^2) = \prod \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{(-\frac{(y_i - \beta x_i)^2}{2\sigma^2})} \right] = (\frac{1}{\sqrt{2\pi\sigma^2}})^N \prod_{i=1}^{N} e^{(-\frac{(y_i - \beta x_i)^2}{2\sigma^2})} = \frac{1^N}{(2\pi\sigma^2)^{\frac{N}{2}}} \prod_{i=1}^{N} e^{(-\frac{(y_i - \beta x_i)^2}{2\sigma^2})}$$

The last expression is the likelihood function we were looking for: $\mathcal{L}(\beta, \sigma^2|x)$. The corresponding log-likelihood thus is:

$$log\mathcal{L}(\beta, \sigma^2) = ln(\frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}}) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \beta x_i)^2 = ln(1) - \frac{N}{2} ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \beta x_i)^2$$

Simplifying we get

$$log\mathcal{L}(\beta, \sigma^2) = -\frac{N}{2} ln(2\pi) - \frac{N}{2} ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \beta x_i)^2$$

[5] We thought about all the possible ways to get exactly the observed sample and the related probabilities.

[6] Wich is the pdf for a Normal with mean βx_i and variance σ^2.

3.2.1 ML estimation for example 3.2

If we maximize the expression with respect to β unsurprisingly we obtain the well known OLS estimator[7] as follows:

$$\arg\max_{\beta} log\mathcal{L}(\beta,\sigma^2|x) \rightarrow -\frac{2}{2\sigma^2}\sum_{i=1}^{N}(y_i-\beta x_i)*-x_i = \frac{-x_i}{\sigma^2}\sum_{i=1}^{N}(y_i-\beta x_i)=0$$

$$\rightarrow -\sum(x_iy_i)+\beta\sum(x_i^2)=0\rightarrow \frac{\sum(x_iy_i)}{\sum(x_i^2)}=\hat{\beta}_{MLE}=\hat{\beta}_{OLS}$$

Further examples can be found in many books and with different settings.

4 Properties of MLE

The statements below are valid:

- in general, for some unknown parameter λ and given it's MLE estimate $\hat{\lambda}_{MLE}$;
- under weak regularity conditions;
- provided a correct specification of the likelihood function.

4.1 Consistency

MLE is consistent, or $plim(\hat{\lambda}_{MLE})=\lambda$. The proof is not provided here[8].

4.2 Asymptotical efficiency and normality

4.2.1 Fisher information matrix and Cramér-Rao Lower bound

To discuss efficiency of MLE we need to introduce the Cramér-Rao Lower bound, and therefore we define the Fisher information matrix.

In section 3.1.1 we've already introduced the score $s(\lambda)$. It can be shown that $E\left[s(\lambda)\right]=0$, thus $E\left[(s(\lambda)^2\right]=V\left[s(\lambda)\right]$[9]. The variance of the score is referred to as the Fisher Information Matrix, or $\mathcal{I}(\lambda)=E\left[(\frac{\partial log\mathcal{L}(\lambda)}{\partial\lambda})^2\right]$. If $log\mathcal{L}(\lambda)$ is twice differentiable with respect to λ, we can write analogously $\mathcal{I}(\lambda)=-E\left[\frac{\partial^2 log\mathcal{L}(\lambda)}{\partial^2\lambda}\right]$, wich states that the information is the negative of the expectation of the Hessian matrix of the log-likelihood with respect to λ. The Hessian describes the curvature of the log-likelihood at different points, therefore conceptually if the curvature is high around the maximum the variance of the ML estimator will be low.

Now, H. Cramér[10] and C. R. Rao[11] derived the so-called Cramér-Rao lower bound (CRLB), wich states that among the class of unbiased estimators, any estimator $\hat{\lambda}$ with $E\left[\hat{\lambda}\right]=\lambda$ has a variance at least as large as the inverse of $\mathcal{I}(\lambda)$, $CRLB=\mathcal{I}(\lambda)^{-1}$. Formally we can state that

$$\forall\,\hat{\lambda}\,s.d.\;E\left[\hat{\lambda}\right]=\lambda,\;V(\hat{\lambda})\geq\mathcal{I}(\lambda)^{-1}$$

4.2.2 MLE is BUE

The ML estimator is efficient because, under the assumption mentioned above, it attains the CRLB and thus has the lower variance among all the unbiased estimators (best unbiased estimator, BUE). The proof is not provided here. It's interesting to note that from the example in section 3.2.1 follows the proof about the OLS estimator being BLUE (like BUE, with "L" for linear).

4.2.3 MLE is asymptotically normally distributed

Like the OLS estimator, MLE is asymptotically normally distributed like $\sqrt{N}(\hat{\lambda}-\lambda)\rightarrow\mathcal{N}(0,V)$. From 4.2.2 follows that $V(\hat{\lambda})=\mathcal{I}(\lambda)^{-1}$. The proof is not provided here.

[7] Here we look at the OLS estimate for β. In fact, when looking at $\hat{\sigma^2}$, OLS and MLE are different at least is small samples (OLS has a degree of freedom correction while MLE doesn't).

[8] A proof can be found here: http://ocw.mit.edu/courses/mathematics/18-443-statistics-for-applications-fall-2006/lecture-notes/lecture3.pdf. The material is part of MIT OCW Course 18-443 and the content is subject to a Creative Commons licence.

[9] For any r.v. X with $E\left[X\right]=0$ it's easy to prove that $V\left[X\right]=E\left[(X-E\left[X\right])^2\right]=E\left[X^2-2XE\left[X\right]+E\left[X\right]^2\right]=E\left[X^2\right]$.

[10] Cramér, H. (1946), Mathematical Methods of Statistics. Princeton, NJ: Princeton Univ. Press.

[11] Rao, C. R. (1945), "Information and the accuracy attainable in the estimation of statistical parameters". Bulletin of the Calcutta Mathematical Society 37: 81–89.

4.3 Inference

With MLE we need to make much stronger assumptions if compared to other methods (e.g. OLS). This effort pays in terms of inference, meaning that MLE allows us to make inference on any feauture of the distribution.

4.4 Nonrobustness to misspecifications

While in section 4.3 we expose the positive aspect of simple MLE, we need to specify that in general it's efficiency comes at the price of nonrobustness. In other words, if the assumption about the likelihood is not correct the ML estimator turns out to be inconsistent.

5 Conclusion

This short paper obviously is not meant to treat the argument in a complete way. For further informations many books are available in addition to the main reference for this paper, "A guide to modern econometrics". One important part missing is the testing procedure. Moreover, some extensions and applications could be useful to enhance comprehension. Corrections and comments are totally welcome (igor.francetic@unil.ch).

6 Figures

Here I put all the figures with corresponding numbers.

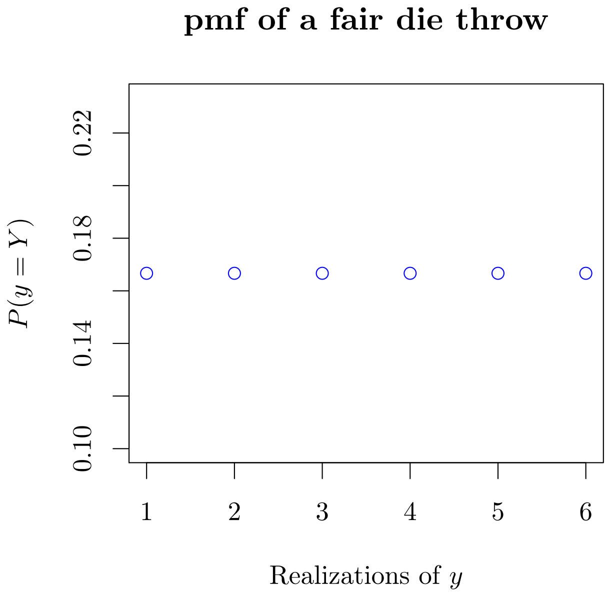

Fig. 1: pmf of a fair die with six faces. Source: author.

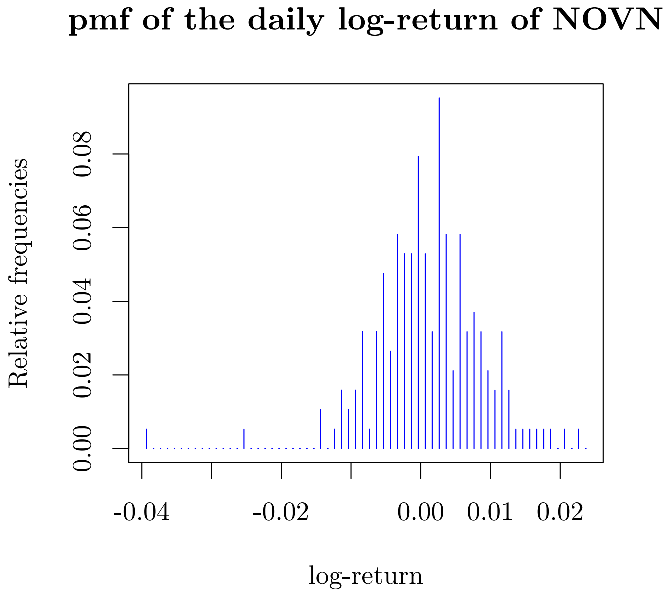

Fig. 2: observed pmf of daily stockprice logreturns of Novartis (Swiss Exchange). Sample represents the whole year 2012. Average is about 0.06%. St. deviation is about 0.0076. Source: yahoo.com/finance, author's elaborations.

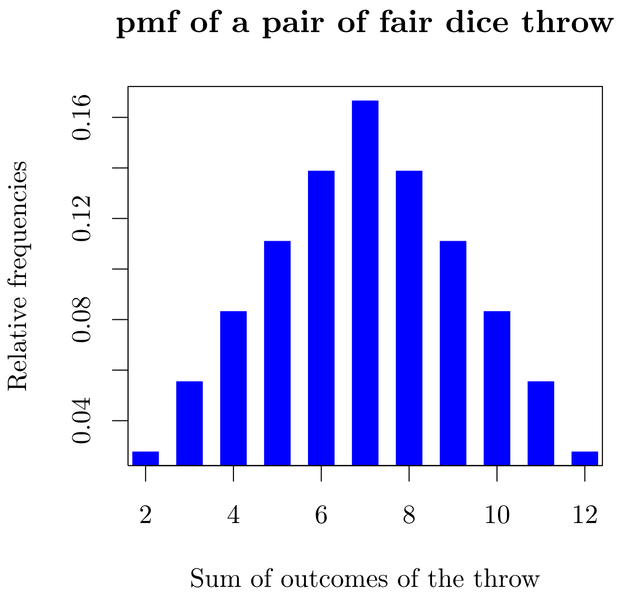

Fig. 3: pmf of a pair of fair dice throw. Source: author.

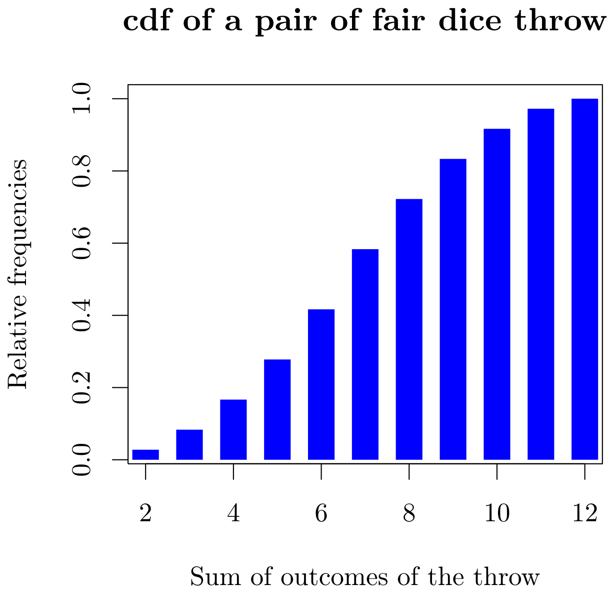

Fig. 4: cdf of a pair of fair dice throw. Source: author.

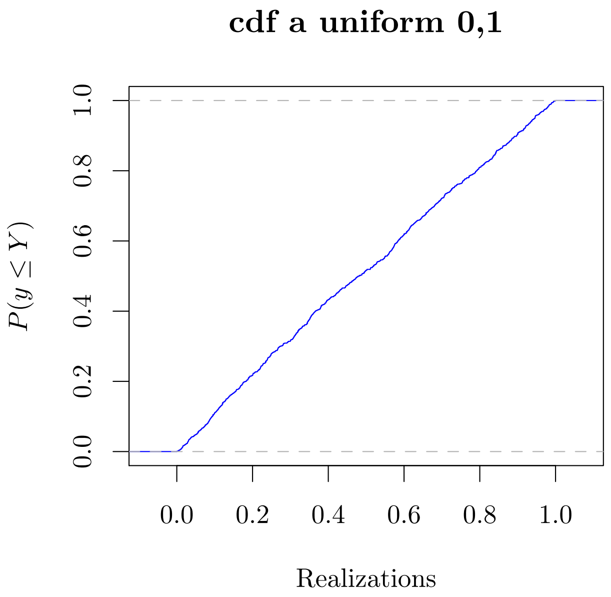

Fig. 5: empirical cdf of 1000 random draws from a uniform $\{0, 1\}$ distribution. Source: author.

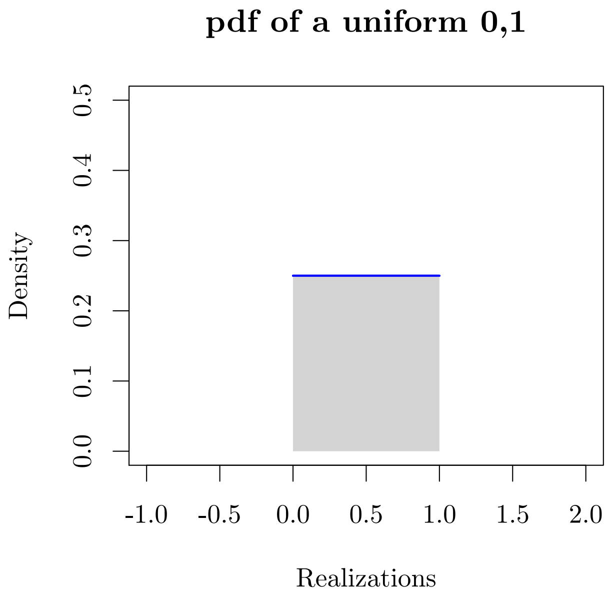

Fig. 6: pdf of a uniform $\{0, 1\}$ distribution. Source: author.

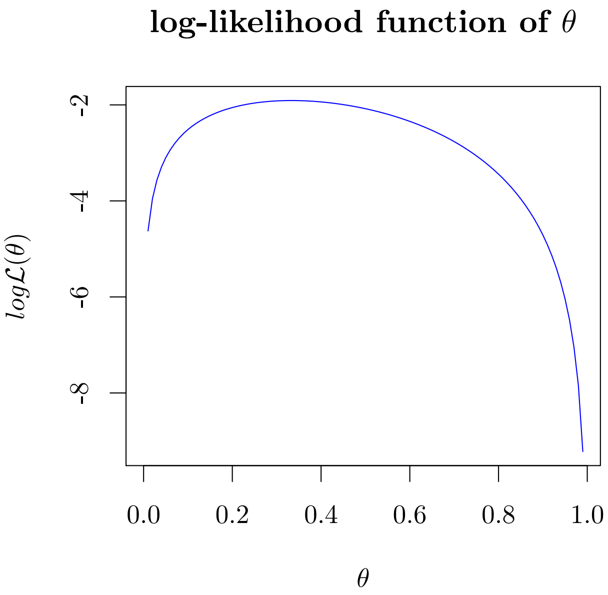

Fig. 7: loglikelihood function of θ from example in section 3.1. Source: author.