Methods of point estimation

Chrysafis Vogiatzis

Lecture 17-18

Learning objectives

After these lectures, we will be able to:

- Find point estimators for unknown parameters.
- Use the method of moments to find point estimators for unknown parameters.
- Use maximum likelihood estimation to find point estimators for unknown parameters.
  - Compare and identify when it is easiest to use likelihood and when log-likelihood.
- Propose new point estimators for unknown parameters based on these three methods.
- Calculate the unknown rate of an exponential distribution, or the unknown success probability of a Bernoulli distribution using the three methods.

Motivation: "I guess it is exponentially distributed. But what is λ?"

Motivation: Estimating the mortality risk

We call **mortality risk** of a hospital the probability of death occurring for any patient admitted to the hospital. The question we need to answer is "what is the mortality risk" of a given hospital? It depends on many factors, such as the type of conditions the patients admitted in this hospital have; the equipment of the hospital; the personnel of the hospital; among many, many others. Our intuition says the following, though: could we not *observe* the hospital for a period of time and then deduce what the risk is based on the obtained data? Is this fair/unfair/correct/misleading? What is the number of deaths in a hospital truly distributed as?

Estimation

During Lectures 15 and 16, we saw what makes a good point estimator $\hat{\Theta}$. We would like to have:

- small **bias** (zero, if possible). $bias = E\left[\hat{\Theta}\right] - \theta.$

- small **variance** (minimum among all estimators). $Var\left[\hat{\Theta}\right]$.
- small **mean square error**. $MSE = bias^2 + Var\left[\hat{\Theta}\right]$.
- We also defined the **relative efficiency** of two estimators $\hat{\Theta}_1, \hat{\Theta}_2$ as $\frac{MSE\left(\hat{\Theta}_1\right)}{MSE\left(\hat{\Theta}_2\right)}$.

So, *given* two or more estimators, you may calculate these items and infer which one to use/which one is better. However, where do these estimators come from? When faced with the problem of recognizing a parameter based on data, what can we do? In this series of lecture, we will work on deriving, using, and comparing three methods of point estimation:

1. **Method of moments** estimators.
2. **Maximum likelihood** estimators.
3. **Bayesian** estimators.

In this set of notes, we only deal with the first two. The third one is addressed in Lecture 19. Before we get to their details, we provide a definition and a motivating example.

Definition 1 (Method of estimation) *Assume we are provided a population X distributed with unknown parameter(s) θ. We want to estimate θ. Given a series of observations (sample) $X_1, X_,2 \ldots, X$, how to come up with a "good" point estimator $\hat{\Theta}$?*

Mortality risk

Let us go back to our original motivating example with calculating/estimating the mortality risk of a hospital based on observations. Say, we have been observing the hospital over the last 2 months, and we have observed 18 deaths in the first 150 patient admissions. What would we estimate the mortality rate as?

Some more examples we may consider?

- How to estimate the rate of an exponential distribution?

  "We know the time between accidents in a factory is exponentially distributed. How do we find out what the rate is?"

- How to estimate the probability (proportion) of a binomial distribution?

"We know the number of students graduating from the College of Engineering is binomially distributed. How do we find out what the probability of success (graduation) is?"

- How to estimate the mean and variance of a binomial distribution?

"We know exam grades in IE 300 are normally distributed. But, what is μ and σ^2?"

Method of moments

Methods

We begin the section with the definition of **sample** (empirical) **moments** and **population moments**. Assume we have a population X distributed with pdf $f(x)$. We have managed to collect a set of samples from the population $X_1, X_2, \ldots, X_n$. Then:

Definition 2 (Population moments) *The k-th population moment of a continuous population X (also referred to as the k-th moment of $f(x)$) is calculated as*

$$E\left[X^k\right] = \int_{-\infty}^{+\infty} x^k f(x) dx.$$

The same logic applies to the k-th population moment of a discrete population X, with a summation rather than an integration:

$$E\left[X^k\right] = \sum_{x \in X} x^k p(x) dx.$$

Definition 3 (Sample (empirical) moments) *The k-th sample moment of X (also referred to as the k-th empirical moment of X) is calculated as*

$$\frac{1}{n} \sum_{i=1}^{n} X_i^k,$$

where $X_1, X_2, \ldots, X_n$ are samples from the population X.

By definition, the first population moment of X is the population mean, and the first sample moment of X is the sample average. On the other hand, the second population moment of X is **not** the population variance; instead, $E\left[X^2\right]$ is only part of the calculation of the variance:

$$Var\left[X\right] = E\left[X^2\right] - \left(E\left[X\right]\right)^2.$$

Similarly, the second sample moment of X is **not** the sample variance!

Calculating population moments

Assume $f(x) = \frac{1}{2}(1 - \alpha \cdot x)$ where α is some parameter. What are the first three population moments?

- first moment:

$$E[X] = \int_{-1}^{+1} x \cdot f(x)dx = \int_{-1}^{+1} x \cdot \frac{1}{2}(1 - \alpha \cdot x)\, dx = -\frac{\alpha}{3}.$$

- second moment:

$$E\left[X^2\right] = \int_{-1}^{+1} x^2 \cdot f(x)dx = \int_{-1}^{+1} x^2 \cdot \frac{1}{2}(1 - \alpha \cdot x)\, dx = \frac{1}{3}$$

- third moment:

$$E\left[X^3\right] = \int_{-1}^{+1} x^3 \cdot f(x)dx = \int_{-1}^{+1} x^3 \cdot \frac{1}{2}(1 - \alpha \cdot x)\, dx = -\frac{\alpha}{5}$$

Calculating sample (empirical) moments

Assume we have collected $n = 5$ samples from the population distributed with $X_1 = 0.7, X_2 = 0.77, X_3 = 0.65, X_4 = 0.5, X_5 = 0.83$. What are the first three sample moments?

- first moment:

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{5}(0.7 + 0.77 + 0.65 + 0.5 + 0.83) = 0.69.$$

- second moment:

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 = \frac{1}{5}\left(0.7^2 + 0.77^2 + 0.65^2 + 0.5^2 + 0.83^2\right) = 0.48886.$$

- third moment:

$$\frac{1}{n}\sum_{i=1}^{n} X_i^3 = \frac{1}{5}\left(0.7^3 + 0.77^3 + 0.65^3 + 0.5^3 + 0.83^3\right) = 0.354189.$$

The method

The main idea behind the method is the following: **we want to match empirical (sample) moments of a distribution to the population moments**. Before we apply the method, we make a couple of observations.

Observation 1 The k-th moment of $f(x)$, $E\left[X^k\right]$ depends only on the unknown parameters $\theta_1, \theta_2, \ldots, \theta_m$.

Observation 2 The k-th moment of the sample, $\frac{1}{n}\sum_{i=1}^{n} X_i^k$ depends only on the data (the sample itself)!

So, if the 1st population moment is expected to *match* the 1st sample moment, and the 2nd population moment is expected to *match* the 2nd sample moment, and so on, then.. how many moments do we need to be able to solve a system of equations?

Based on the above discussion, we are now ready to formally state the method of moments. Assume we have m unknown parameters $\theta_1, \theta_2, \ldots, \theta_m$. The *method of moment estimators* $\hat{\Theta}_1, \hat{\Theta}_1, \ldots, \hat{\Theta}_m$ can be obtained by:

1. Get the first m [1] moments of $f(x)$ and of the sample.
2. Equate them.
3. Solve a system of equations with m unknowns (parameters θ_i)!

The solution obtained are the **method of moment estimators** $\hat{\Theta}_i$ for each parameter θ_i.

[1] Need to take more than m if some moments are zero or produce equations on the same variables as the previous ones.

Our first method of moments estimator

Recall earlier the population X distributed with $f(x) = \frac{1}{2}(1 - \alpha \cdot x)$ where α is some (unknown) parameter. We have collected a sample of $n = 5$ observations from X and we found the observations to be $X_1 = 0.7, X_2 = 0.77, X_3 = 0.65, X_4 = 0.5, X_5 = 0.83$. What is the method of moments estimator for α?

We have already found both the first population and the first sample moments. Looking at the earlier solutions, we have $E[X] = -\frac{\alpha}{3}$ and $\frac{1}{n}\sum_{i=1}^{n} X_i = 0.69$. Equating we get:

$$-\frac{\alpha}{3} = 0.69 \implies \hat{\alpha} = -2.07.$$

Observe how we put a "hat" ($\hat{}$) on top of α when we assign a value to it in the end. This is done to signal that this is merely an estimator and is not necessarily its true value.

The general case

In general, letting $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, given any sample of n observations, we may calculate the method of moments estimator for α as:

$$\hat{\alpha} = -3 \cdot \overline{X}.$$

The method of moments estimator for an exponential distribution

Assume we suspect X is a population that is exponentially distributed, but with unknown rate λ. Thankfully, we have collected a sample from that population: $X_1, X_2, \ldots, X_n$. We have one unknown parameter (λ) so we will need one equation.

Let us try the first population moment [2]:

$$E[X] = \frac{1}{\lambda}.$$

[2] Easy to find as it is the expected value of an exponential distribution!

Similarly, we may obtain the first sample moment as:

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}.$$

Recall that it is typical to denote a sample average as $\overline{X}$.

Equating the two (per the method of moments), we get:

$$\hat{\lambda} = 1/\overline{X} = \frac{n}{\sum_{i=1}^{n} X_i}$$

estimator — "recipe" — estimate — #

The method of moments estimator for a normal distribution

Assume we have some normally distributed population with mean μ and variance σ^2. Alas, they are both unknown. However, we have collected n observations (a sample) from the population: $X_1, X_2, \ldots, X_n$. What is the method of moments estimators for μ and σ^2.

We divide this proof in two parts:

1 The population moments.

For the population moments, we need (at least) the first two: $E[X]$ and $E\left[X^2\right]$. The first one is easy, as it is equal to μ. The second one on the other hand is **not the variance**: it is used in the variance calculation! Recall that $\sigma^2 = E\left[X^2\right] - (E[X])^2 \implies E\left[X^2\right] = \sigma^2 + (E[X])^2 = \sigma^2 + \mu^2$. In summary, we have:

$$E[X] = \mu$$
$$E\left[X^2\right] = \sigma^2 + \mu^2.$$

Estimating the rate of earthquakes

We assume that the time between two earthquakes of magnitude greater than or equal to 7 in Japan is exponentially distributed. Here is a list of earthquakes that satisfy these criteria from the last decade and when they have happened:

1	April 16, 2016
2	May 30, 2015
3	October 26, 2013
4	December 7, 2012
5	July 10, 2011
6	April 11, 2011
7	April 7, 2011
8	March 11, 2011
9	March 11, 2011
10	March 9, 2011
11	December 21, 2010
12	February 26, 2010

What is the method of moments estimator for the rate λ?

We first consider the time between the earthquakes. We have 11 such observations (between the first and the second, between the second and the third, etc.). Let us count this in days (for consistency): 322 days, 581 days, 323 days, 516 days, 90 days, 4 days, 27 days, 0 days, 2 days, 78 days, 318 days.
From the method of moments, we want the first population and sample method (as we only have one unknown parameter), so:

$$\begin{cases} E[X] = \frac{1}{\lambda} \\ \frac{1}{n}\sum_{i=1}^{n} X_i = 205.55 \end{cases} \implies \hat{\lambda} = 1 \text{ earthquake per } 205.55 \text{ days.}$$

2 The sample moments. These are easier to calculate as:

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}$$
$$\frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

Here, again, we use $\overline{X}$ to represent the sample average.

Equating the two, we get the following system of equations:

$$\mu = \overline{X} \implies \hat{\mu} = \overline{X}.$$

$$\mu^2 + \sigma^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \implies \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} X_i^2 - n\mu^2}{n} = \frac{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}{n}.$$

Estimating the grade distribution of an exam

Assume we want to estimate the grade distribution of an exam *before* we grade all of the exams! If we expect the distribution to be normally distributed, we could grade the first five exams (at random) and get their grades $X_1 = 80, X_2 = 97, X_3 = 50, X_4 = 67, X_5 = 84$. Then, we calculate $\frac{1}{n}\sum_{i=1}^{n} X_i = 75.6$ and $\frac{1}{n}\sum_{i=1}^{n} X_i^2 = 5970.8$. Finally, from the method of moments, we have:

$$\begin{cases} E[X] = \frac{1}{n}\sum_{i=1}^{n} X_i = 75.6 \\ E\left[X^2\right] = \frac{1}{n}\sum_{i=1}^{n} X_i^2 = 5970.8 \end{cases} \implies \begin{cases} \hat{\mu} = 75.6. \\ \hat{\sigma}^2 = 255.44. \end{cases}$$

The method of moments estimator for a Bernoulli distribution

Finally, let X be a Bernoulli random variable with probability of success p, that is unknown. How to estimate it using the method of moments? Well, we resort to the following setup. Let us run n experiments of that Bernoulli random variable and let's mark each of them as X_i with a 1 (when successful) or a 0 (when failed). Then:

$$E[X] = p$$
$$\frac{1}{n}\sum_{i=1}^{n} X_i$$

Equating the two, we get that

$$p = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Fair or unfair?

Assume you have an unfair coin, but you have no idea how unfair it is – that is, p is not known. Say you toss the coin $n = 10$ times and get 7 Heads, 3 Tails. What is the estimator you get for the probability of getting Heads from the method of moments?

Let Heads be equal to 1 and Tails equal to 0. Then: $\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{7}{10} = 0.7$. From the method of moments $\hat{p} = 0.7$.

A few extra examples

This is an example from the slides. In the slides, we mention that the distribution is normal; but this is not necessary!

A delivery problem

We believe the times it takes to deliver a package are identically distributed with the same unknown mean μ and variance σ^2. We have collected information on 10 packages and the time to delivery (in hours) are: 49.1, 47.9, 48.6, 50.4, 49.5, 49.8, 48.2, 50.3, 45.2, 46.2. What are good mean and variance estimators for the normal distribution using the method of moments?

We have two unknown parameters (mean and variance), so we will need at least two population and sample moments. Let us take the first two:

- Population 1st moment:

$$E\left[X^1\right] = E[X] = \mu$$

- Sample 1st moment:

$$\frac{1}{10}\sum_{i=1}^{10} X_i^1 = 48.52$$

- Population 2nd moment:

$$E\left[X^2\right] = Var[X] + (E[X])^2 = \\ = \sigma^2 + \mu^2$$

- Sample 2nd moment:

$$\frac{1}{10}\sum_{i=1}^{10} X_i^2 = 2356.844$$

Equating the two and solving the system of equations, we get $\hat{\mu} = 48.52$ and $\hat{\sigma}^2 = 2.6536$.

A discrete distribution

Assume we have a discrete random variable X defined over $0, 1, 2, 3, 4$ and distributed with probabilities $p(0) = \frac{\theta_1}{3}, p(1) = \frac{\theta_1}{6}, p(2) = \frac{\theta_1}{6}, p(3) = \frac{\theta_2}{2}, p(4) = \frac{\theta_2}{2}$.

Now, assume we have collected a sample of $n = 10$ observations: $0, 1, 1, 3, 4, 2, 2, 3, 4, 1$. Based on this, what is the method of moments estimators for θ_1 and for θ_2?

Now, let's see. At first glance we have two estimators.. But we know better than that. We probably remember that $\sum_{x=0}^{4} p(x) = 1$, which implies that:

$$\sum_{x=0}^{4} p(x) = 1 \implies \frac{\theta_1}{3} + \frac{\theta_1}{6} + \frac{\theta_1}{6} + \frac{\theta_2}{2} + \frac{\theta_2}{2} = 1 \implies \frac{2\theta_1}{3} + \theta_2 = 1.$$

Based on that, if we knew, say θ_1 we could obtain θ_2 right away. Let us get the first moments and equate them:

$$E[X] = 0 \cdot \frac{\theta_1}{3} + 1 \cdot \frac{\theta_1}{6} + 2 \cdot \frac{\theta_1}{6} + 3 \cdot \frac{\theta_2}{2} + 4 \cdot \frac{\theta_2}{2} = \frac{\theta_1 + 7\theta_2}{2}.$$

$$\frac{1}{n} \sum_{i=1}^{10} X_i = \frac{1}{10} (0 + 1 + 1 + 3 + 4 + 2 + 2 + 3 + 4 + 1) = 2.1.$$

Finally, we have a system of equations at our hands!

$$\begin{cases} \frac{2\theta_1}{3} + \theta_2 = 1 \\ \frac{\theta_1 + 7\theta_2}{2} = 2.1 \end{cases} \implies \begin{cases} \hat{\theta}_1 = \frac{42}{55} \\ \hat{\theta}_2 = \frac{27}{55} \end{cases}$$

Maximum likelihood estimation

Basics

Recall that we already have a population X distributed with pdf $f(x)$. Also recall that the pdf has one or more unknown parameters θ. We may then write that the pdf is actually a function of x and θ as $f(x, \theta)$. That is, we need inputs for both the value x and the parameter(s) θ before evaluating $f(x)$. Finally, we have already collected a sample of n observations from the population, let them be $X_1, X_2, \ldots, X_n$.

This brings us to the definition of the **likelihood function**.

Definition 4 (Likelihood function) *The **likelihood function** of a sample of n observations $X_1, X_2, \ldots, X_n$ is defined as*

$$L(\theta) = f(X_1, \theta) \cdot f(X_2, \theta) \cdot \ldots \cdot f(X_n, \theta) = \prod_{i=1}^{n} f(X_i, \theta).$$

Observe how the likelihood function is only a function of θ as $X_i, i = 1, \ldots, n$ are known quantities.

The method

The main idea is pretty simple: for the sample to have been obtained the way it has, then the observations must have been **likely**. Hence, they must be values that maximize the likelihood function! This is summarized in the following statement:

The **maximum likelihood estimators** $\hat{\Theta}$ are
the values that *maximize* the likelihood function.

The maximum likelihood estimators are also referred to as MLE. To find this maximizer, we take the first derivative of the likelihood function and equate it to 0:

$$\frac{\partial L}{\partial \theta} = 0$$

and solve for θ to obtain the estimator.

Our first MLE estimator

Go back again to the population X distributed with $f(x) = \frac{1}{2}(1 - \alpha \cdot x)$ where α is the unknown parameter we would like to estimate. We have a sample from X as $X_1 = 0.7, X_2 = 0.77, X_3 = 0.65, X_4 = 0.5, X_5 = 0.83$. What is the MLE estimator for α?

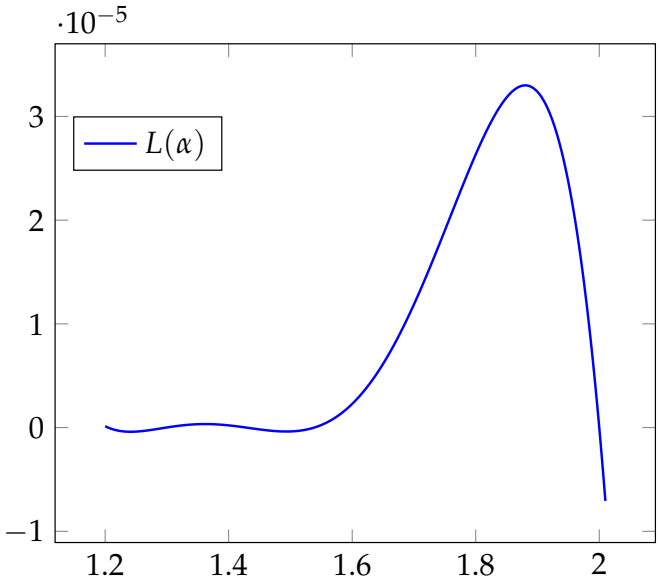
First to build the likelihood function:

$$\begin{aligned} L(\alpha) &= f(X_1) \cdot f(X_2) \cdot f(X_3) \cdot f(X_4) \cdot f(X_5) = \\ &= \frac{1}{32}(1 - 0.7\alpha)(1 - 0.77\alpha)(1 - 0.65\alpha)(1 - 0.5\alpha)(1 - 0.83\alpha) = \\ &= \frac{1}{32} - 0.107813\alpha + 0.147784\alpha^2 - 0.100557\alpha^3 + 0.0339432\alpha^4 - 0.0045436\alpha^5 \end{aligned}$$

Then, we get the first derivative and set it equal to o to find the maximizer. We get:

$$\frac{\partial L}{\partial \alpha} = 0 \implies \alpha = 1.88.$$

This solution could also be found visually! Here is a plot of the likelihood function and the point where it is maximized is easier to find.



Finally, observe how we got a different estimator here compared to the method of moments!

Extension to log-likelihood

Since the likelihood involves a product of n pdf values, it comes as no surprise that our end result may be a little difficult to control and use. This is why we may also introduce the **log-likelihood**:

Definition 5 (Log-likelihood function) *The **log-likelihood function** of a sample of n observations $X_1, X_2, \ldots, X_n$ is defined as*

$$\ln L(\theta) = \ln f(X_1, \theta) + \ln f(X_2, \theta) + \ldots + \ln f(X_n, \theta) = \sum_{i=1}^{n} \ln f(X_i, \theta).$$

Observe how also the log-likelihood function is only a function of θ as $X_i, i = 1, \ldots, n$ are known quantities. Contrary to the simple likelihood function, the log-likelihood is a summation which makes it easier to differentiate.

Our first MLE estimator using log-likelihood

We have:

- pdf $f(x) = \frac{1}{2}(1 - \alpha \cdot x)$;
- sample $X_1 = 0.7, X_2 = 0.77, X_3 = 0.65, X_4 = 0.5, X_5 = 0.83$.

We build the log-likelihood function as:

$$\begin{aligned}\ln L(\alpha) &= \ln f(X_1) + \ln f(X_2) + \ln f(X_3) + \ln f(X_4) + \ln f(X_5) = \\ &= \ln \frac{1}{2}(1 - 0.7\alpha) + \ln \frac{1}{2}(1 - 0.77\alpha) + \ln \frac{1}{2}(1 - 0.65\alpha) + \\ &+ \ln \frac{1}{2}(1 - 0.5\alpha) + \ln \frac{1}{2}(1 - 0.83\alpha).\end{aligned}$$

Here, we note that $\left(\frac{1}{2}(1 - X_i\alpha)\right)' = \frac{X_i}{\alpha X_i - 1}$. Hence, in our case we have:

$$\frac{\partial \ln L}{\partial \alpha} = 0 \implies$$
$$\frac{0.7}{1 - 0.7\alpha} + \frac{0.77}{1 - 0.77\alpha} + \frac{0.65}{1 - 0.65\alpha} + \frac{0.5}{1 - 0.5\alpha} + \frac{0.83}{1 - 0.83\alpha} = 0 \implies$$
$$\implies \alpha = 1.88.$$

The result obtained using the likelihood or the log-likelihood function will be the same.

The MLE estimator for an exponential distribution

Assume we have obtained a sample of n observations $X_1, X_2, \ldots, X_n$ with average $\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$. We also assume that the population

is exponentially distributed with rate λ. What is the MLE estimator for λ?

First, we build the log-likelihood function as:

$$\begin{aligned}\ln L(\lambda) &= \ln \lambda e^{-\lambda X_1} + \ln \lambda e^{-\lambda X_2} + \ldots + \ln \lambda e^{-\lambda X_n} = \\ &= \ln \lambda - \lambda X_1 + \ln \lambda - \lambda X_2 + \ldots + \ln \lambda - \lambda X_n = \\ &= n \ln \lambda - \lambda \left(X_1 + X_2 + \ldots + X_n\right)\end{aligned}$$

Again, find the maximizer:

$$\frac{\partial \ln L(\lambda)}{\partial \lambda} = 0 \implies \left(n \ln \lambda - \lambda \left(X_1 + X_2 + \ldots + X_n\right)\right)' = 0 \implies$$

$$\frac{n}{\lambda} - \left(X_1 + X_2 + \ldots + X_n\right) = 0 \implies n - \lambda \sum_{i=1}^{n} X_i = 0 \implies \lambda = \frac{n}{\sum\limits_{i=1}^{n} X_i} = \frac{1}{\overline{X}}$$

Observe how we have reached the same result as when using the method of moments. Recall that this is **not necessarily** always the case.

Say we had not wanted to use the log-likelihood and instead used the simple likelihood function $L(\lambda)$:

$$\begin{aligned}L(\lambda) &= \lambda e^{-\lambda X_1} \cdot \lambda e^{-\lambda X_2} \cdot \ldots \cdot \lambda e^{-\lambda X_n} = \lambda^n \cdot e^{-\lambda (X_1 + X_2 + \ldots + X_n)} = \\ &= \lambda^n \cdot e^{-\lambda \sum\limits_{i=1}^{n} X_i}\end{aligned}$$

Take the derivative:

$$\begin{aligned}&\frac{\partial L(\lambda)}{\partial \lambda} = 0 \implies \left(\lambda^n \cdot e^{-\lambda \sum\limits_{i=1}^{n} X_i}\right)' = 0 \implies \\ &\implies n \cdot \lambda^{n-1} \cdot e^{-\lambda \sum\limits_{i=1}^{n} X_i} - \lambda^n \cdot \sum_{i=1}^{n} X_i \cdot e^{-\lambda \sum\limits_{i=1}^{n} X_i} = 0.\end{aligned}$$

Observe how we can simplify quite a bit: we may divide by λ^{n-1} (because we know that $\lambda > 0$). This gives us:

$$n \cdot e^{-\lambda \sum\limits_{i=1}^{n} X_i} - \lambda \cdot \sum_{i=1}^{n} X_i \cdot e^{-\lambda \sum\limits_{i=1}^{n} X_i} = 0$$

We may also divide by $e^{-\lambda \cdot \sum\limits_{i=1}^{n} X_i}$ because it is also definitely positive. This leads to the much more manageable:

$$n - \lambda \cdot \sum_{i=1}^{n} X_i = 0 \implies \lambda = \frac{n}{\sum\limits_{i=1}^{n} X_i} = \frac{1}{\overline{X}}$$

Note how getting the same result with the log-likelihood was significantly easier due to the nature of this probability density function.

The MLE estimator for a Bernoulli distribution

Once again, consider that we have a population producing random variables distributed as Bernoulli with probability of success p. We have obtained a sample of $n = 10$ observations with 7 successes (let them be $X_i = 1$) and 3 failures ($X_i = 0$). What is the MLE estimator for the unknown p?

Based on the MLE method we first need to calculate the likelihood function. Recall that for a Bernoulli random variable its probability mass function (as it is a discrete random variable) is $P(0) = 1 - p$ and $P(1) = p$. Without loss of generality, assume we arrange the observations with the successes first (the first, say, 7 observations) and the failures next (the remaining $10 - 7 = 3$ observations).

We are now ready to build the likelihood function:

$$L(p) = \left(\prod_{i=1}^{7} p\right) \cdot \left(\prod_{i=8}^{10} (1-p)\right) = p^7 \cdot (1-p)^{10-7} = p^7 \cdot (1-p)^3 .$$

The derivative of the likelihood function can be found as:

$$\frac{\partial L}{\partial p} = \left(p^7 \cdot (1-p)^3\right)' = 7p^6(1-p)^3 - 3(1-p)^2 p^7.$$

Now, equate this to 0 to get the maximizer:

$$\begin{aligned} 7p^6(1-p)^3 - 3(1-p)^2 p^7 = 0 \implies \\ \implies 7(1-p) - 3p = 0 \implies \hat{p} = 0.7. \end{aligned}$$

In the above, we make the assumption that $p \in (0, 1)$: that is, it cannot be 0 or 1. If we allowed this to be the case, then solving would give three solutions $p = 0, p = 1, p = 0.7$. However, the first two solutions are *minima* rather than *maxima*, and we would still pick $\hat{p} = 0.7$ as our estimator.

A few extra examples

This is an example from the slides. In the slides, we mention that the distribution is normal; but this is not necessary!

A delivery problem

We believe the times it takes to deliver a package are identically distributed with the same unknown mean μ and variance σ^2. We have collected information on 10 packages and the time to delivery (in hours) are: 49.1, 47.9, 48.6, 50.4, 49.5, 49.8, 48.2, 50.3, 45.2, 46.2. What are good mean and variance estimators for the normal distribution using the method of moments?

We have two unknown parameters (mean and variance), so we will need at least two population and sample moments. Let us take the first two:

- Population 1st moment:

$$E\left[X^1\right] = E[X] = \mu$$

- Sample 1st moment:

$$\frac{1}{10}\sum_{i=1}^{10} X_i^1 = 48.52$$

- Population 2nd moment:

$$E\left[X^2\right] = Var[X] + (E[X])^2 = \\ = \sigma^2 + \mu^2$$

- Sample 2nd moment:

$$\frac{1}{10}\sum_{i=1}^{10} X_i^2 = 2356.844$$

Equating the two and solving the system of equations, we get $\hat{\mu} = 48.52$ and $\hat{\sigma}^2 = 2.6536$.

A discrete distribution

Assume we have a discrete random variable X defined over $0, 1, 2, 3, 4$ and distributed with probabilities $p(0) = \frac{\theta_1}{3}, p(1) = \frac{\theta_1}{6}, p(2) = \frac{\theta_1}{6}, p(3) = \frac{\theta_2}{2}, p(4) = \frac{\theta_2}{2}$.

Now, assume we have collected a sample of $n = 10$ observations: $0, 1, 1, 3, 4, 2, 2, 3, 4, 1$. Based on this, what is the method of moments estimators for θ_1 and for θ_2?

Now, let's see. At first glance we have two estimators.. But we know better than that. We probably remember that $\sum_{x=0}^{4} p(x) = 1$, which implies that:

$$\sum_{x=0}^{4} p(x) = 1 \implies \frac{\theta_1}{3} + \frac{\theta_1}{6} + \frac{\theta_1}{6} + \frac{\theta_2}{2} + \frac{\theta_2}{2} = 1 \implies \frac{2\theta_1}{3} + \theta_2 = 1.$$

Based on that, if we knew, say θ_1 we could obtain θ_2 right away. Let us get the first moments and equate them:

$$E[X] = 0 \cdot \frac{\theta_1}{3} + 1 \cdot \frac{\theta_1}{6} + 2 \cdot \frac{\theta_1}{6} + 3 \cdot \frac{\theta_2}{2} + 4 \cdot \frac{\theta_2}{2} = \frac{\theta_1 + 7\theta_2}{2}.$$

$$\frac{1}{n} \sum_{i=1}^{10} X_i = \frac{1}{10} (0 + 1 + 1 + 3 + 4 + 2 + 2 + 3 + 4 + 1) = 2.1.$$

Finally, we have a system of equations at our hands!

$$\begin{cases} \frac{2\theta_1}{3} + \theta_2 = 1 \\ \frac{\theta_1 + 7\theta_2}{2} = 2.1 \end{cases} \implies \begin{cases} \hat{\theta}_1 = \frac{42}{55} \\ \hat{\theta}_2 = \frac{27}{55} \end{cases}$$