Statistics: Definitions and Theorems

Probability Theory and Statistics, Fall 2023

translated from (a subset of) the notes of András Tóbiás, which are based on lectures of Marianna Bolla.

Translated by: Padmini Mukkamala

Last updated: December 2, 2023

P(a theorem likely to come on the exam makes it into the translated notes) $= 1$
P(an example makes it | time when I read it $> 10pm$) $= 0.01$
P(an example makes it | time when I read it $< 8pm$) $= 0.5$
P(an example makes it $|8pm <$ time when I read it $< 10pm$) $= \vartheta$, where ϑ is an unknown parameter. Lets take a i.i.d sample X_1, X_2.....

Contents

Lecture 21

Introduction to Statistics

We encountered problems until now, where we knew the distribution (and its specific parameters) of quantities we were measuring or dealing with, for example the outcome of a fair coin toss has Bernoulli distribution with parameter $\frac{1}{2}$, number of Heads in 5 tosses of a fair coin will have $Bin(5, \frac{1}{2})$ distribution. In statistics instead, this parameter is unknown, although we have some idea about the distribution, like Binomial or Geometric. For example, if we find a coin but don't know if it is fair coin or not, then we can think of the outcome Heads as the random variable $1(p)$, where p is the unknown probability of turning up Heads. If we toss this coin 5 times and count the Heads, then it will have $Bin(5, p)$ distribution, where the parameter p is again unknown. Heights of students can be said to have $N(\mu, \sigma^2)$ distribution with two unknown parameters, waiting time at a bank counter can be said to have $Exp(\lambda)$ distribution, where λ is unknown etc. Our goal is to determine or estimate the unknown parameter(s).
The first step to such an estimation is to take samples. For the above coin example where we count number of Heads in 5 tosses, a single sample would be the number of Heads in 5 independent tosses of the coin. We will reapeat this experiment many times, i.e, take many 5 tosses of the coin. We will denote with $X_1, X_2, ..., X_n$ the number of Heads in each collection of 5 tosses, the outcomes of the n samples. We note that each of these samples is independently collected, and because we are repeating the same experiment, it is reasonable to assume that these random variables are also identical.

Let $X_1, X_2, ..., X_n$ be n independent, identically distributed random variables with an unknown distribution. Then the vector $\mathbb{X} = (X_1, X_2, ..., X_n)$ is known as independent indentically distributed sample of n elements. We will call this an **i.i.d sample of n elements** for short.

Once we have a sample, there are two main ways or methods for analysing our data. The first is Estimation, and the second is Hypothesis Testing. In Estimation, we are interested in estimating the unknown parameter (either with an exact value, or with an interval) or some function of it (like the population mean, population standard deviation), while in Hypothesis Testing, we assess the probability of a certain Hypothesis about the parameter (or a function of it, like mean, standard deviation) being true. Examples of Hypothesis are: the coin we have is a fair coin, the average of height of female students at BME is 165cm, the average waiting time at a bank counter is 15 min etc.
The distribution of each component in the i.i.d sample $(X_1, X_2, \ldots, X_n)$ depends on an unknown parameter ϑ. We say that the range of ϑ is θ, that is, $\vartheta \in \theta \subseteq \mathbb{R}^d$. For a given $\vartheta \in \theta$, we denote the cumulative distribution function of X_1 with F_ϑ. Note: in the past $F_X(x)$ denoted the cumulative distribution of the random variable X, but here $F_\vartheta(x)$ is $P(X_1 \leq x)$, that is, it is the cumulative distribution function of X_1 and not of ϑ. Further, depending on whether X_1 is discrete or continuous, $p_\vartheta(x)$ or $f_\vartheta(x)$ denote the probability mass function or the density function.
For example, for a coin with unknown probability of heads, $p_\vartheta(0) = 1 - \vartheta$ and $p_\vartheta(1) = \vartheta$, $\vartheta \in [0, 1]$. For height distributions, since (μ, σ^2) is the vector of unknown parameters, so,

$$f_{(\mu,\sigma^2)}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad \mu \in \mathbb{R}, \sigma^2 > 0.$$

is the density function.
We define Range as before,

$$\text{Range}_\vartheta(X_i) = \{x \in \mathbb{R} |\ f_\vartheta(x) > 0\} \subseteq \mathbb{R}.$$

The density function above is replaced with mass function in case of a discrete distribution.

Let $(X_1, X_2, \ldots, X_n)$ be an i.i.d. sample of n elements. Then, $(x_1, x_2, \ldots, x_n)$, where $\forall i,\ x_i \in Range_\vartheta(X_i)$, is said to be an **observation** or **realization** of the sample.

Let $(X_1, X_2, \ldots, X_n)$ be an i.i.d. sample of n elements. We call a function $T \colon \mathbb{R}^n \to \mathbb{R}$, with $T(\mathbb{X}) = T(X_1, X_2, \ldots, X_n)$ a **Statistic**.
When a statistic is used to estimate a population parameter, then it is called an **Estimator**. The value of the estimator evaluated for a specific observation is called an **estimate**.

Lecture 22

Introduction to Statistics (continued.)

Sample Mean, Empirical Sample variance

Sample mean, median etc all are various statistics.

Sample Mean: Let $(X_1, X_2, \ldots, X_n)$ be an i.i.d. sample of n elements. Then the sample mean is a statistic given by,

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

(Note: the book doesn't use the subscript n and only defines it for observations as below).

If $\mathbf{x} = (x_1, \ldots, x_n)$ is an observation of $\mathbf{X} = (X_1, \ldots, X_n)$, then we denote the mean of the observation by $\overline{x_n}$ where $\overline{x_n} = \frac{x_1 + \ldots + x_n}{n}$.
Note: This is the **value of the Statistic** for the given observation.

We further note that the expected value of the sample mean, if it exists, is the expected value of any one component of the sample. That is,

$$E\left[\overline{X_n}\right] = \frac{1}{n}(E(X_1) + \ldots + E(X_n)) = E(X_1).$$

It is comforting to see that the sample mean has the mean of the unknown distribution as its expected value. Then the sample mean does give us a nice way to estimate the population mean (or mean of the unknown distribution).

Sample Variance: Let $(X_1, X_2, \ldots, X_n)$ be an i.i.d. sample of n elements. Then the **sample variance** is a statistic given by,

$$S_n^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X_n})^2$$

Here we call S_n^* the **sample standard deviation**.

Empirical Sample Variance: the variance of the empirical distribution is given by

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X_n})^2$$

Note that the **empirical sample standard deviation** is S_n and this **not** the same as S_n^*.

Why are there two definitions of Variance? In the way that the expected value of sample mean was the expected value of our unknown distribution, we will try to see what is the expected value of S_n^{*2}.

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a i.i.d. sample of n elements such that $E(X_i^2) < \infty$. Then,

$$E(S_n^{*2}) = Var(X_1).$$

Proof:
$E(S_n^{*2}) = \frac{1}{n-1}\sum_{i=1}^{n} E\left(X_i - \overline{X_n}\right)^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(E(X_i^2) - 2E(X_i\overline{X_n}) + E(\overline{X_n}^2)\right)$
$= \frac{1}{n-1}\sum_{i=1}^{n} E(X_i)^2 - \frac{2}{n-1}\sum_{i=1}^{n} E(X_i\overline{X_n}) + \frac{n}{n-1}E(\overline{X_n}^2)$
We will calculate each of the terms above separately. The first term is:

$$\frac{1}{n-1}\sum_{i=1}^{n} E(X_i)^2 = \frac{n}{n-1}E(X_1^2),$$

The second term (without its negative sign):

$$\frac{2}{n-1}\sum_{i=1}^{n} E(X_i\overline{X_n}) = \frac{2}{n-1}\Big(\sum_{i=1}^{n}\frac{1}{n}E(X_i^2) + \frac{1}{n}\sum_{j\neq i}E(X_iX_j)\Big) = \frac{2}{n-1}\Big(E(X_1^2) + (n-1)E(X_1)^2\Big)$$
$$= \frac{2}{n-1}E(X_1^2) + 2E(X_1)^2,$$

While the third term is:

$$\frac{n}{n-1}E(\overline{X_n}^2) = \frac{n}{(n-1)n^2}E\Big[\Big(\sum_{i=1}^{n}X_i\Big)^2\Big] = \frac{1}{n(n-1)}E\Big(\sum_{i=1}^{n}X_i^2 + 2\sum_{1\leq i<j\leq n}X_iX_j\Big)$$
$$= \frac{1}{n(n-1)}\Big(\sum_{i=1}^{n}E(X_1^2) + (n-1)E(X_1)^2\Big)$$
$$= \frac{1}{n-1}E(X_1^2) + E(X_1)^2.$$

In all, then,

$$E[S_n^{*2}] = E(X_1^2)\big(\frac{n}{n-1} - \frac{2}{n-1} + \frac{1}{n-1}\big) + E(X_1)^2(-2+1) = E(X_1^2) - E(X_1)^2 = Var(X_1).$$

Note: the above doesn't work for the empirical sample variance.
And we note that because the expected value of the sample variance is the population variance, it is a good estimator for the population variance.
The **mode** is the statistic denifed as the value with the maximum frequency. If there are several such values, then each of them is considered a mode and the data is multimodal.

Ordered Sample, Empirical Distribution Function

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a i.i.d. sample of n elements. We denote by $(X_1^*, X_2^*, \ldots, X_n^*)$ the ordering of $X_1, \ldots, X_n$ in increasing order. So,

$$X_1^* \leq X_2^* \leq \ldots \leq X_n^*.$$

Here $(X_1^*, X_2^*, \ldots, X_n^*)$ is called an **ordered sample**.

We note in passing, but will not prove it, that $X_1^*, \ldots, X_n^*$ are not independent.

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a i.i.d. sample of n elements. We define **median** of this as $m_{n,\mathbf{X}} = X_{k+1}^*$ if $n = 2k+1$, and as $m_{n,\mathbf{X}} = \frac{X_k^* + X_{k+1}^*}{2}$ if $n = 2k$.

We note that in the i.i.d sample of n-elements, we do not know the distribution of the X_i's. But we try to answer the question that is it possible to estimate, as a limit as $n \to \infty$, the distribution using the sample? We will show that the empirical distribution function, defined as below, will serve this purpose.

Let $n \in \mathbb{N}$ and let $X_1, \ldots, X_n$ be a i.i.d. sample of n elements. The following function:

$$x \mapsto F_n^*(x) := \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{X_i \leq x\}} = \frac{1}{n}\big|\{i \in \{1, \ldots, n\}|\ X_i \leq x\}\big| \tag{1}$$

is defined as the **empirical distribution function** of this i.i.d. sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ of n elements.

It is important to note that F_n^* is a *random* function: its value depends on the value of the random variables $X_1, \ldots, X_n$.

We can give another definition for the Empirical distribution function using our definition of the ordered sample for the given sample. $F_n^*(x) = \begin{cases} 0, & \text{if } x < X_1^*, \\ \frac{k}{n}, & \text{if } X_k^* \le x < X_{k+1}^*, \quad k = 1, \dots, n-1, \\ 1, & \text{if } x \ge X_n^*. \end{cases}$ From this we can also see that for any occurance or realization of the sample given by $\mathbf{x} = (x_1, \dots, x_n)$, we will get a distribution function (limits are 0,1 at $-\infty$ and ∞ and is monotone increasing and is right continuous).
The next theorem shows us that the Empirical distribution function is a good estimator of the distribution function of X_i's.

Let $n \in \mathbb{N}$ and let $X_1, \dots, X_n$ be i.i.d. random variables with distribution function F. Then for every $x \in \mathbb{R}$, we have,

1. $E(F_n^*(x)) = F(x)$,
2. $Var(F_n^*(x)) = \frac{F(x)(1-F(x))}{n}$, and
3. $P(\lim_{n\to\infty} F_n^*(x) = F(x)) = 1$.

Let $x \in \mathbb{R}$. According to the definition we notice that $n \cdot F_n^*(x)$ is the sum of independent identical indicator random variables $\mathbb{1}_{\{X_i \le x\}}$, with the parameter (and by definition expected value) $P(X_i \le x) = F(x)$. So $nF_n^* \sim Bin(n, F(x))$. Using this,

$$E\big(F_n^*(x)\big) = \frac{1}{n} E\big(nF_n^*(x)\big) = \frac{1}{n} nF(x) = F(x)$$

and

$$Var\big(F_n^*(x)\big) = \frac{1}{n^2} Var\big(nF_n^*(x)\big) = \frac{1}{n^2} nF(x)(1-F(x)) = \frac{F(x)(1-F(x))}{n},$$

which were the first two statements we were required to prove. For (3), we use the strong law of large numbers applied to the indicator random variables $\mathbb{1}_{\{X_i \le x\}}$. Their expected value is

$$E\big(\mathbb{1}_{\{X_1 \le x\}}\big) = P(X_1 \le x) = F(x)$$

(and variance is $nF(x)(1-F(x))$), so, using Strong Law of Large Numbers:

$$P\Big(\lim_{n\to\infty} F_n^*(x) = F(x)\Big) = P\Big(\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \le x\}} = E\big(\mathbb{1}_{\{X_1 \le x\}}\big)\Big) = 1.$$

This ends the proof of (3).

Glivenko-Cantelli Theorem: Let $n \in \mathbb{N}$ and let $X_1, \dots, X_n$ be a i.i.d. random variables with distribution function F. Then the function $x \mapsto F_n^*(x)$ converges with probability 1 to the distribution function $x \mapsto F(x)$, or,

$$P\Big(\lim_{n\to\infty} \sup_{x\in\mathbb{R}} \big|F_n^*(x) - F(x)\big| = 0\Big) = 1.$$

We will not prove this theorem but note this is stronger result than the one we obtained by using the Strong Law of Large numbers. Here also not only does the F^* distribution converge almost surely to F we also have uniform convergence.

Estimator

Let $\mathbf{X} = (X_1, \dots, X_n)$ be a i.i.d. sample of n elements, where their distribution depends on an unknown parameter $\vartheta \in \theta \subseteq \mathbb{R}^d$, and let $\psi \colon \mathbb{R}^d \to \mathbb{R}^m$ be a function, where $m \ge 1$. Further, let $T(\mathbf{X}) = T(X_1, \dots, X_n)$ be a statistic for the sample. The statistic T if used to estimate ψ is then called an **Estimator**.

We often use statistics (either the standard ones like mean, median, mode or something else) to estimate the unknown parameter, or a function (ψ) that depends on the unknown parameter ϑ.

We will look now at some properties that are useful to have in an estimator. We say that the statistic $T(\mathbf{X})$,

1. is an **unbiased** estimator of the parameter $\psi(\vartheta)$ if for every $\vartheta \in \theta$,
$$E_\vartheta(T(X_1, \ldots, X_n)) = \psi(\vartheta),$$
where E_ϑ denotes the expected value using P_ϑ as the distribution,
2. is an **asymptotically unbiased** estimator of the parameter $\psi(\vartheta)$ if for every $\vartheta \in \theta$,
$$\lim_{n\to\infty} E_\vartheta(T(X_1, \ldots, X_n)) = \psi(\vartheta),$$
3. is a **consistent** estimator of the parameter $\psi(\vartheta)$, if for every $\vartheta \in \theta$,
$$P_\vartheta\big(\lim_{n\to\infty} T(X_1, \ldots, X_n) = \psi(\vartheta)\big) = 1,$$
4. is an **efficient** estimator of the parameter $\psi(\vartheta)$ if, $T(\mathbf{X})$ is an unbiased estimate of the parameter $\psi(\vartheta)$, and for any unbiased estimator $T'(\mathbf{X}) = T'(X_1, \ldots, X_n)$ of the parameter $\psi(\vartheta)$,
$$Var_\vartheta(T(\mathbf{X})) \leq Var_\vartheta(T'(\mathbf{X})),$$
where Var_ϑ denotes the Variance computed with P_ϑ.

We notice that an unbiased estimator is also an asymptotically unbiased estimator.

Lecture 23

Maximum Likelihood Estimate, Method of Moments

Maximum Likelihood Estimate

An important method used to estimate the unknown parameter is the **Maximum Likelihood Estimate**. We take an i.i.d. sample of n elements $\boldsymbol{X} = (X_1, \ldots, X_n)$, and further consider an observation $\boldsymbol{x} = (x_1, \ldots, x_n)$. The distribution of X_i's depends on some unknown parameter ϑ. This distribution's mass function is given by p_ϑ in the discrete case, or the density function by f_ϑ in the continuous case. The key idea behind the maximum likelihood estimate is to find the parameter ϑ^* which will maximize the probability of the observation $\boldsymbol{x}$. In the discrete case, we can directly work with the probability of the observation, but in the continuous case, we try to find the ϑ^* which maximizes the value of the density function. We will formalize these ideas below.

It is important to note however that once we have $\vartheta_* = \vartheta_*(x_1, \ldots, x_n)$, the result of the estimate is the statistic:
$$T(X_1, \ldots, X_n) = \vartheta_*(X_1, \ldots, X_n)$$
So, it is important to note here that the result of the estimate is a **function of the i.i.d. random sample** $\boldsymbol{X}$ and not of the observation $\boldsymbol{x}$.
The function that we will optimize (using the mass function in the discrete case, and the density function in the continuous case) is called the Likelihood function.

Let $X_1, \ldots, X_n$ be an i.i.d sample of n elements, whose distribution is dependent on an unknown parameter $\vartheta \in \theta \subseteq \mathbb{R}^d$.

1. if the distribution of X_1 for any $\vartheta \in \theta$ is discrete, and p_ϑ denotes the mass function with the parameter ϑ, then, for any observation $(x_1, \ldots, x_n)$, we let,

$$L_\vartheta(x_1, \ldots, x_n) = \prod_{i=1}^{n} p_\vartheta(x_i).$$

2. if the distribution of X_1 for any $\vartheta \in \theta$ is continuous, and $x \mapsto f_\vartheta(x)$ denotes the density function with the parameter ϑ, then for any observation $(x_1, \ldots, x_n)$, we let,

$$L_\vartheta(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_\vartheta(x_i).$$

The function $\mathbb{R}^n \to \mathbb{R}$, $(x_1, \ldots, x_n) \mapsto L_\vartheta(x_1, \ldots, x_n)$ is said to be the **likelihood-function** (pertaining to the parameter ϑ).

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a i.i.d sample of n elements with distribution dependent on an unknown parameter $\vartheta \in \theta \subset \mathbb{R}^d$. Further assume that (for every $\vartheta \in \theta$) the L_ϑ likelihood function exists. Let $\mathbf{x} = (x_1, \ldots, x_n)$ be an observation of the sample $\mathbf{X} = (X_1, \ldots, X_n)$.

The **maximum likelihood estimate** of ϑ **in terms of the observation** $(x_1, \ldots, x_n)$ is the parameter $\vartheta_*(x_1, \ldots, x_n) \in \theta$, for which the likelihood function $L_\vartheta(x_1, \ldots, x_n)$ attains a maximum. Formally,

$$\vartheta_*(x_1, \ldots, x_n) = \arg\max_{\vartheta \in \theta} L_\vartheta(x_1, \ldots, x_n). \tag{2}$$

In other words, $\vartheta_*(x_1, \ldots, x_n)$ is the parameter from θ for which given any $\vartheta' \in \theta$,

$$L_{\vartheta_*(x_1,\ldots,x_n)}(x_1, \ldots, x_n) \geq L_{\vartheta'}(x_1, \ldots, x_n). \tag{3}$$

Although it is possible that the likelihood function has several maximums or none at all, but in practice, this rarely happens.
As stated earlier, our maximum likelihood estimate is a statistic, not its value for some observation $(x_1, \ldots, x_n)$, but a function defined on the random sample $\mathbf{X} = (X_1, \ldots, X_n)$. The following definition serves this purpose:

Using the maximum likelihood estimate **in terms of the observation** as given the previous definition, we now define the function,

$$T \colon \mathbb{R}^n \to \mathbb{R}, \quad (x_1, \ldots, x_n) \mapsto \vartheta_*(x_1, \ldots, x_n)$$

This statistic

$$T(X_1, \ldots, X_n) = \vartheta_*(X_1, \ldots, X_n)$$

is called the **maximum likelihood estimate** of ϑ.

We notice that since the likelihood function is a product of functions, taking its logarithm gives a nice sum which is in practice easier to deal with.

If $(x_1, \ldots, x_n) \mapsto L_\vartheta(x_1, \ldots, x_n)$ is a likelihood dunction, then the function,

$$\mathbb{R}^n \to \mathbb{R}, \quad (x_1, \ldots, x_n) \mapsto l_\vartheta(x_1, \ldots, x_n) = \ln L_\vartheta(x_1, \ldots, x_n)$$

is called the **log-likelihood-function**.

If $(x_1, \ldots, x_n) \mapsto L_\vartheta(x_1, \ldots, x_n)$ is a likelihood-function and $(x_1, \ldots, x_n) \mapsto l_\vartheta(x_1, \ldots, x_n) = \ln L_\vartheta(x_1, \ldots, x_n)$ is a corresponding log-likelihood-function, then for a $\vartheta \in \theta$, $L_\vartheta(x_1, \ldots, x_n)$ is a maximum only if $l_\vartheta(x_1, \ldots, x_n)$ is a maximum.

Proof: The statement follows directly from the fact that logarithm is a strictly increasing function.
Steps for finding the Maximum-Likelihood-Estimate when the log-likelihood function is differentiable:

1. For a fixed observation $x_1, \ldots, x_n$ and a $\vartheta \in \theta$ parameter, determine the likelihood-function $L_\vartheta(x_1, \ldots, x_n)$.
2. Take its logarithm to get the $l_\vartheta(x_1, \ldots, x_n)$ function value.
3. Differentiate the function $\vartheta \mapsto l_\vartheta(x_1, \ldots, x_n)$ with respect to ϑ.
4. Find places where this derivative is 0.
5. Check all the places where it is zero to see if it is a local maximum (this can be done by checking if the derivative changes from a positive value to a negative value in a small interval around this point, OR by checking if the second derivative is negative). If there are many local maximums then we find the biggest of these as the global maximum. [1]
6. If we do have a global maximum $\vartheta_*(x_1, \ldots, x_n)$, then, $\vartheta_*(X_1, \ldots, X_n)$ is the required maximum likelihood estimate.

We will see two examples of the maximum likelihood estimate, one in the discrete case, and the other in the continuous case.
First we will see it for Poisson distribution. We will follow the above steps for a random i.i.d. sample of n elements $X_1, \ldots, X_n$, all with Poisson distribution with an unknown parameter $\vartheta > 0$. Let $(x_1, \ldots, x_n)$ be an observation of $(X_1, \ldots, X_n)$ (note that then it is a vector of non-negative integers).

1. Since
$$p_\vartheta(x_i) = P_\vartheta(X_i = x_i) = \frac{\vartheta^{x_i}}{x_i!} e^{-\vartheta}, \qquad i = 1, \ldots, k,$$
so
$$L_\vartheta(x_1, \ldots, x_n) = \prod_{i=1}^{n} p_\vartheta(x_i) = \prod_{i=1}^{n} \frac{\vartheta^{x_i}}{x_i!} e^{-\vartheta} = \frac{\vartheta^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} e^{-n\vartheta}.$$
2. The log-likelihood function is then,
$$l_\vartheta(x_1, \ldots, x_n) = \ln\left(\frac{\vartheta^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} e^{-n\vartheta}\right) = \ln\vartheta \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \ln(x_i!) - n\theta.$$
3. We will take the derivative of the log-likelihood function with respect to ϑ:
$$\frac{\partial}{\partial\vartheta} l_\vartheta(x_1, \ldots, x_n) = \frac{1}{\vartheta} \sum_{i=1}^{n} x_i - n.$$
4. We note here that it was indeed useful to use log-likelihood function, that the computations are easier and cleaner. The zeros of the derivative:
$$\frac{\partial}{\partial\vartheta} l_\vartheta(x_1, \ldots, x_n) = 0 \qquad \Leftrightarrow \qquad \frac{1}{\vartheta} \sum_{i=1}^{n} x_i - n = 0 \qquad \Leftrightarrow \qquad \vartheta = \frac{1}{n} \sum_{i=1}^{n} x_i = \overline{x_n}.$$
Or, the only zero of the derivative is the value of the sample mean for the given observation.
5. We note that the second derivative of the log-likelihood function is negative for any $\vartheta \in \theta$:
$$\frac{\partial^2}{\partial\vartheta^2} l_\vartheta(x_1, \ldots, x_n) = -\frac{1}{\vartheta^2} \sum_{i=1}^{n} x_i < 0.$$
So the zero is a local maximum. Also, since the range of the parameter ϑ is an open interval $\theta = (0, \infty)$, so the obtained maximum is a global maximum.

[1] If the parameter space θ is closed, then its boundary must also be inspected to find maximums.

6. So the log-likelihood function obtains its maximum at

$$\vartheta_*(x_1,\dots,x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x_n}$$

And so, the maximum likelihood estimate is:

$$\vartheta_*(X_1,\dots,X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X_n}.$$

Now we will look at a continuous example with Exponential distribution. Let $X_1,\dots,X_n$ be a i.i.d. sample of n elements with exponential distribution with an unknown parameter $\vartheta > 0$. Let $(x_1,\dots,x_n)$ be an observation of $(X_1,\dots,X_n)$ (note that then it is a vector of positive real numbers).

1. The density function of the elements of the sample is

$$f_\vartheta(x_i) = \vartheta e^{-\vartheta x_i}, \qquad i = 1,\dots,n,$$

so the likelihood-function is:

$$L_\vartheta(x_1,\dots,x_n) = \prod_{i=1}^{n} f_\vartheta(x_i) = \prod_{i=1}^{n} \vartheta e^{-\vartheta x_i} = \vartheta^n e^{-\vartheta\sum_{i=1}^n x_i}.$$

2. The log-likelihood-function is therefore,

$$l_\vartheta(x_1,\dots,x_n) = \ln\left(\vartheta^n e^{-\vartheta\sum_{i=1}^n x_i}\right) = n\ln\vartheta - \vartheta\sum_{i=1}^{n} x_i.$$

3. Taking the derivative of the log-likelihood function with respect to ϑ:

$$\frac{\partial}{\partial\vartheta}\ln l_\vartheta(x_1,\dots,x_n) = \frac{n}{\theta} - \sum_{i=1}^{n} x_i.$$

4. The places where this is zero:

$$\frac{\partial}{\partial\vartheta}\ln l_\vartheta(x_1,\dots,x_n) = 0 \qquad \Leftrightarrow \qquad \frac{n}{\theta} - \sum_{i=1}^{n} x_i = 0 \qquad \Leftrightarrow \qquad \theta = \frac{n}{\sum_{i=1}^n x_i} = \frac{1}{\overline{x_n}}.$$

5. Since

$$\frac{\partial^2}{\partial\vartheta^2}\ln l_\vartheta(x_1,\dots,x_n) = -\frac{n}{\vartheta^2}$$

is negative for every $\vartheta > 0$, the zero we found is a local maximum. Also, since the range of the parameter is an open interval $\theta = (0,\infty)$, so this is also a global maximum.

6. So the maximum-likelihood estimate for the given observation $(x_1,\dots,x_n)$ is

$$\vartheta_*(x_1,\dots,x_n) = \frac{n}{\sum_{i=1}^n x_i} = \frac{1}{\overline{x_n}}.$$

So the maximum likelihood estimate is:

$$\vartheta_*(X_1,\dots,X_n) = \frac{n}{\sum_{i=1}^n X_i} = \frac{1}{\overline{X_n}}.$$

Method of Moments

We will define moments and empirical moments for our next method of estimation: the method of moments.

For any positive integer k and a random variable X, if $E[|X|^k] < \infty$, then the k^{th} **moment** is defined as $E[X^k] \in \mathbb{R}$.

Let k be a positive integer, and let $X_1, \ldots, X_n$ be i.i.d. random variables. Then, for the i.i.d. sample $(X_1, \ldots, X_n)$ of n elements, the k^{th} **empirical moment** is defined as the statistic:

$$\widehat{m}_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k$$

The method of moments is particularly useful when there are more than one parameter. For example if we have normal distribution and μ and σ^2 are both unknown. Then the parameter vector is $\boldsymbol{\vartheta} = (\mu, \sigma^2) \in \mathbb{R}^2$, where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. As another example, consider the uniform distribution on the interval (a, b) where both end points are unknown. Then the parameter vector then is $\boldsymbol{\vartheta} = (a, b) \in \mathbb{R}^2$ for some real numbers $-\infty < a < b < \infty$.
For a given $\mathbf{X} = (X_1, \ldots, X_n)$ i.i.d sample of n elements, $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_k)$, with the range $\theta \subseteq \mathbb{R}^k$. Also assume that for $X_1, \ldots, X_n$, elements of the sample, the first k moments exist. These moments depend on the unknown parameter $\vartheta_1, \ldots, \vartheta_k$:

$$\begin{aligned} m_1 &= g_1(\vartheta_1, \ldots, \vartheta_k), \\ m_2 &= g_2(\vartheta_1, \ldots, \vartheta_k), \\ &\vdots \\ m_k &= g_k(\vartheta_1, \ldots, \vartheta_k) \end{aligned}$$

for some functions $g_1, g_2, \ldots, g_k \colon \mathbb{R}^k \to \mathbb{R}$.

$$(\vartheta_1, \ldots, \vartheta_k) \mapsto \boldsymbol{g}(\vartheta_1, \ldots, \vartheta_k) := (g_1(\vartheta_1, \ldots, \vartheta_k), \ldots, g_k(\vartheta_1, \ldots, \vartheta_k))$$

is then a function $\boldsymbol{g} \colon \mathbb{R}^k \to \mathbb{R}^k$, or a vector space. If this vector space has an inverse, then we represent it with $\boldsymbol{h} = (h_1, \ldots, h_k)$. Using this inverse, and with the knowledge of moments, we can determine $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_k)$ parameter:

$$\begin{aligned} \vartheta_1 &= h_1(m_1, \ldots, m_k), \\ \vartheta_2 &= h_2(m_1, \ldots, m_k), \\ &\vdots \\ \vartheta_k &= h_k(m_1, \ldots, m_k) \end{aligned}$$

In the above situation, the estimate for the parameter vector $\underline{\vartheta} = (\vartheta_1, \ldots, \vartheta_n)$ using the **method of moments** is denoted by $\widehat{\boldsymbol{\vartheta}} = (\widehat{\vartheta}_1, \ldots, \widehat{\vartheta}_k)$, where,

$$\widehat{\vartheta}_i = h_i(\widehat{m}_1, \ldots, \widehat{m}_k), \qquad i = 1, \ldots, k,$$

and $\widehat{m}_i$ denotes the i^{th} empirical moment.

For a differenciable $\boldsymbol{g}$, the inverse $\boldsymbol{h} = (h_1, \ldots, h_k)$ exists only if the determinant of the Jacobian of $\boldsymbol{g}$ is never zero on the set θ.
We illustrate this with an example:
Let $X_1, \ldots, X_n$ be i.i.d random variables with $N(\mu, \sigma^2)$ distribution, where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. Then the first two moments for $k = 1, 2$ for a $\boldsymbol{\vartheta} = (\mu, \sigma^2)$ can be obtained as:

$$\begin{aligned} m_1 &= E_\vartheta(X_1) = \mu, \\ m_2 &= E_\vartheta(X_1^2) = Var_\vartheta(X_1) + (E_\vartheta(X_1))^2 = \sigma^2 + \mu^2. \end{aligned}$$

In this case, $\boldsymbol{g}\colon (\mu, \sigma^2) \mapsto (\mu, \sigma^2 + \mu^2)$ is our map which is invertible. [2]:

$$\mu = m_1,$$
$$\sigma^2 = m_2 - m_1^2,$$

or in other words, the inverse map is $\boldsymbol{h}\colon (m_1, m_2) \mapsto (m_1, m_1 - m_2^2)$. And so the estimate with the method of moments is:

$$\widehat{\mu} = \widehat{m}_1 = \overline{X_n},$$
$$\widehat{\sigma^2} = \widehat{m}_2 - \widehat{m}_1^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X_n}^2.$$

We can also see that $\widehat{m}_1 - \widehat{m}_2^2$ can be obtained in terms of the empirical sample variance.

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X_n})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \frac{2}{n}\sum_{i=1}^{n} X_i\overline{X_n} + \frac{1}{n}\sum_{i=1}^{n}\overline{X}^2 = \sum_{i=1}^{n} X_i^2 - 2\overline{X_n}^2 + \overline{X_n}^2 = m_2 - m_1^2. \quad (4)$$

And this is true for any distribution.
Let us look at another example:
Let $X_1, \ldots, X_n$ be i.i.d. random variables with uniform distribution in the interval (a, b), with the unknown parameter vector, $\boldsymbol{\vartheta} = (a, b)$, where $-\infty < a < b < \infty$. Using the properties of uniform distribution, the first two moments are:

$$m_1 = E_\vartheta(X_1) = \frac{a+b}{2},$$
$$m_2 = E_\vartheta(X_1^2) = Var_\vartheta(X_1) + E_\vartheta(X_1)^2 = \frac{(b-a)^2}{12} + \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2 + 3(a+b)^2}{12}.$$

So then we obtain,

$$\frac{(b-a)^2}{12} = m_2 - m_1^2 \qquad \Rightarrow \qquad b - a = 2\sqrt{3(m_2 - m_1^2)}$$

and

$$a + b = 2m_1.$$

(We also note that the determinant of the Jacobian matrix is never zero in $-\infty < a < b < \infty$.) With this we obtain a set of linear equations for a and b, which in the interval $-\infty < a < b < \infty$ always has a solution. And the solution is:

$$a = m_1 - \sqrt{3(m_2 - m_1^2)},$$
$$b = m_1 + \sqrt{3(m_2 - m_1^2)}.$$

Using the equation $\widehat{m}_2 - \widehat{m}_1^2 = S_n^2$, we get the following estimate using the method of moments:

$$\widehat{a} = \widehat{m}_1 - \sqrt{3(\widehat{m}_2 - \widehat{m}_1^2)} = \overline{X_n} - \sqrt{3S_n^2},$$
$$\widehat{b} = \widehat{m}_1 + \sqrt{3(\widehat{m}_2 - \widehat{m}_1^2)} = \overline{X_n} + \sqrt{3S_n^2}.$$

[2] Here, when computing the Jacobian matrix, one must consider σ^2 as the variable (and not σ), so the Jacobian matrix is:

$$\begin{bmatrix} \frac{\partial g_1(\mu,\sigma^2)}{\partial \mu} & \frac{\partial g_1(\mu,\sigma^2)}{\partial (\sigma^2)} \\ \frac{\partial g_2(\mu,\sigma^2)}{\partial \mu} & \frac{\partial g_2(\mu,\sigma^2)}{\partial (\sigma^2)} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 2\mu & \sigma^2 \end{bmatrix}.$$

The determinant of this σ^2, which in the range θ is never zero.