Methods to Find Estimators

In the previous part of the lectures, we were considering some properties of estimators, and specifying what are good estimators, what are some desirable properties for a good estimator. And our examples were to consider some estimators for parameters, and to see if they are good or bad, if they possess some nice properties or not.

Now, we want to consider some methods to construct estimators with good properties. That is, given a Statistical Model with some unknown parameters, we want to find some good estimators for that parameters.

In this part we will consider two Klassik Methods for estimation: Maximum Likelihood Method (Maximum Likelihood Estimation, MLE) and Method of Moments (Method of Moments Estimation, MME). There are also some other methods to construct estimators, for example, Bayesian Methods, and we will consider them in the appendix.

10.1 Maximum Likelihood Method

Example, MLE: Assume we are tossing a coin 10 times, and we have obtained the following result:

$$\text{THHHTHTHHT} \tag{10.1}$$

Let p be the probability of Heads. What is your best guess for p?
Think before reading the rest.
If this seems hard, try to answer first to the same question in the case of the output

$$\text{HHHHHHHHHH}. \tag{10.2}$$

We will consider the first example (10.1). We model our situation by the $\text{Bernoulli}(p)$ model, where p is the probability of, say, Heads. That is, we assume that our coin is producing results according to the distribution $\text{Bernoulli}(p)$, i.e., the probability of having H with our coin is p.

We want to estimate our p based on the observation given above. Let us think like this - is it possible that our coin is fair, i.e., $p = 0.5$, having the observation (10.1)? The answer is - yes, it is possible. Even, it is possible to have the result from (10.2) with a fair coin. **But the probability of having that result with a fair coin is small compared to when tossing a biased coin**. Let us calculate that, first for the second observation, and then, for the first one. If our coin is fair, then the probability of H is $\frac{1}{2}$, and the probability of having 10 heads when tossing the coin 10 times is $(\frac{1}{2})^{10} = \frac{1}{1024} \approx 0.00097$.

The probability of having the observation (10.1) is again, surprisingly, $(\frac{1}{2})^{10}$ (first we need to have a T, and the probability of that is 0.5, next - we need H, and the probability is 0.5 again, so the probability to have TH is $0.5 \cdot 0.5$ and so on). Now what will be this probability, if we will assume that $p = 6/10 = 3/5$? The probability will be

$$\mathbb{P}(\text{having THHHTHTHHT} \mid p = 0.6) = \mathbb{P}(T|p = 0.6) \cdot \mathbb{P}(H|p = 0.6) \cdot \mathbb{P}(H|p = 0.6) \cdot \ldots \cdot \mathbb{P}(T|p = 0.6) =$$

$$= (\mathbb{P}(T|p = 0.6))^4 \cdot (\mathbb{P}(H|p = 0.6))^6 = 0.4^4 \cdot 0.6^6 \approx 0.00119.$$

We see that this is higher than in the case of the fair coin. So **it is more possible that the observation** (10.1) **is from the coin with** $p = 0.6$ **than from the coin with** $p = 0.5$. It can be shown that, actually, no other value of p will give more chance to this observation. So we guess that the most possible value for our unknown p is $p = 0.6$. But, of course, we can be wrong - although the chances are smaller, but our observation can be from the fair coin. The Maximum Likelihood Method takes as the estimate the value of the parameter making the observed sample to have the highest chance to appear. This is just like searching for your (lost) pair of socks. Well, maybe your socks are between you Calculus and Stat textbooks, but the chances are higher that the socks are in the wardrobe. Silly, example, I agree. I know that you do not have neither Calculus nor Stat textbooks ☺

If we will consider the second observation, (10.2), and consider the case $p = 1$, then the probability of having this observation is

$$\mathbb{P}(\text{having HHHHHHHHHH} \mid p = 1) = \mathbb{P}(H|p = 1) \cdot \mathbb{P}(H|p = 1) \cdot \mathbb{P}(H|p = 1) \cdot ... \cdot \mathbb{P}(H|p = 1) = 1.$$

So $p = 1$ is the value giving the most chances for our observation (10.2) to appear. Although, of course, this can be observed also with a coin with $p = 0.8$. *But the chance of this is smaller*.

One of the methods to construct estimators is the Maximum Likelihood Method (MLE). Here we consider a parametric model $\{\mathbb{P}_\theta : \theta \in \Theta\}$, assuming either that $\mathbb{P}_\theta$ is discrete and given by the PMF $f(x|\theta)$ or that $\mathbb{P}_\theta$ is continuous, given by the PDF $f(x|\theta)$.

Now assume that we are given a random sample

$$X_1, X_2, ..., X_n \overset{IID}{\sim} \mathbb{P} \in \{\mathbb{P}_\theta : \theta \in \Theta\}.$$

Definition 10.1. *The **Likelihood Function** for the Parametric Model $\{\mathbb{P}_\theta : \theta \in \Theta\}$ and the random sample $X_1, ..., X_n$ is the Joint PD(M)F of $X_1, ..., X_n$, considered as a function of the parameter θ, i.e., it is given by*[1]

$$\mathcal{L}(\theta) = \mathcal{L}_n(X_1, ..., X_n|\theta) = f(X_1|\theta) \cdot f(X_2|\theta) \cdot ... \cdot f(X_n|\theta),$$

*and the **Log-Likelihood Function** is the function*

$$\ell(\theta) = \ell(X_1, ..., X_n|\theta) = \ln \mathcal{L}(\theta) = \sum_{k=1}^{n} \ln f(X_k|\theta),$$

in the case when $f(x|\theta) > 0$.

Also we define the Negative Log-Likelihood Function to be

$$\ell(\theta) = -\ln \mathcal{L}(\theta).$$

The advantage of the negative log-likelihood function is because most computer software are designed to find the **minimum** of a function, and our aim will be, see the rest of the section, to find the maximum point of the likelihood function, which is the same as the maximum point of the log-likelihood function, and the same as the minimum point of the negative log-likelihood function[2].

[1] Since X_k-s are independent

[2] Also sometimes one defines the normalized negative log-likelihood function:

$$\ell(\theta) = -\frac{1}{n} \cdot \ln \mathcal{L}(\theta).$$

Another advantage is that when we use calculus tools to check if the point is the maximum point, we need to use negative definiteness of the Hessian matrix, and for the minimum point we check the positive definiteness of the Hessian, and checking positive definiteness is easier and preferable.

One can also consider the function

$$\mathcal{L}_n(\theta) = \mathcal{L}_n(x_1, ..., x_n, \theta) = f(x_1|\theta) \cdot f(x_2|\theta) \cdot ... \cdot f(x_n|\theta)$$

as the Likelihood function, with x_k instead of r.v. X_k. This is OK, and it helps us to find the Maximum Likelihood Estimate, not the Estimator (see the definition next). We will make difference between two things: to find an **Estimator** and to find an **Estimate**.

Again, the Likelihood function is just the Joint PDF, but considered not a function of $X_1, ..., X_n$, but as a function of the unknown parameter θ. I.e., the variable of our Likelihood function is θ.

Definition 10.2. *The **Maximum Likelihood Estimator** (MLE) of the parameter θ is the value of θ that maximizes the likelihood function for the given random sample $X_1, ..., X_n$, the global maximum point (in case it exists) of $\mathcal{L}(X_1, ..., X_n|\theta)$:*

$$\hat{\theta}_n = \hat{\theta}_n^{MLE} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, \mathcal{L}(\theta).$$

Definition 10.3. *The **Maximum Likelihood Estimate** (MLE) of the parameter θ is the value of θ that maximizes the likelihood function for the given realization (observation) $x_1, ..., x_n$ of $X_1, ..., X_n$, the global maximum point (in case it exists) of $\mathcal{L}(x_1, ..., x_n|\theta)$:*

$$\hat{\theta}_n = \hat{\theta}_n^{ML} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, \mathcal{L}(x_1, ..., x_n|\theta).$$

So in case you will calculate ML Estimator, you will obtain a function of your random sample $X_1, ..., X_n$, and in this case you can study the statistical properties of your estimator (consistency, unbiasedness, effectiveness etc). In case you will calculate ML Estimate, you will obtain just a number as your estimate for your unknown parameter, and that will be all (and you cannot give any other properties, that will be just a number).

Clearly, there are two ways to obtain a ML Estimate - one is to obtain first the ML Estimator, then to plug the observed values of our random variables, and the other approach is just to maximize the Likelihood function calculated at that observation, $\mathcal{L}(x_1, ..., x_n|\theta)$. We will do in the first way, since it will allow us to examine the properties of the Estimator first (and, after making sure it is a trustable one with nice properties, we can just plug the observed values and get the estimate).

Since the maximum point of[3] $\mathcal{L}$ is just the maximum point of $\ln \mathcal{L}(\theta) = \ell(\theta)$, that is,

$$\underset{\theta \in \Theta}{\operatorname{argmax}}\, \mathcal{L}(\theta) = \underset{\theta \in \Theta}{\operatorname{argmax}}\, \ell(\theta),$$

usually one tries to maximize the log-likelihood function, since it is given in the sum form, which is much easier to deal with.

REMARK, ML IDEA: So what is the idea of ML Method? - We have an observation (data) coming from a parametric model. ML Method tries to choose that value of the parameter that will make the observed data most likely.

[3] But not the maximum value!

EXAMPLE, ML FOR DISCRETE MODEL: For example, in the discrete case, the Likelihood Function is given by

$$\mathcal{L}(\theta) = \mathcal{L}(x_1, ..., x_n|\theta) = \mathbb{P}(X_1 = x_1, ..., X_n = x_n)$$

(θ is hidden behind the distribution of each X_k) and ML Method tries to find the value of the parameter θ that makes the observation $x_1, ..., x_n$ most likely.

Need to expand and explain this example. No candies = No explanation!

REMARK, IDEA OF THE MLE: Assume we have a dataset generated from a Normal Distribution with the known variance. We want to estimate the value of the Mean μ. The dataset is generated by **R**:

$$0.5798511, -3.1452570, -4.7189382, -6.1291531, 0.4131273, -3.2731419, -2.1459219, -2.5730876,$$

$$-0.8604049, -2.7885335, -4.2254977, -2.0445037, -2.6230195, -3.1152881, -3.3123097, 0.2227296, -2.4562443,$$

$$2.9767443, -3.9351425, -4.4120513, -3.9335637, -3.1549264, 0.5045312, -4.0657712, -2.7856526, -4.3379948,$$

$$-2.7271007, -1.7534743, -0.7400594, -3.8701907, -1.2397111, -2.7222809, -6.3212004, -1.1558864, -2.1475976,$$

$$0.3064523, 1.5176192, -1.8878617, -3.0870704, -5.7957451$$

Fig. 10.1 shows the generated dataset with dots on the OX axis. On the same graph we have two PDFs for Normal Distributions, one is for the $\mathcal{N}(2, 2^2)$, with blue color, and the other one is in red, for $\mathcal{N}(-2.4, 2^2)$. Now, is it possible that our dataset is generated from the $\mathcal{N}(2, 2^2)$? Aha, it is possible. Say, is it possible that the second datapoint, -3.1452570, is generated from $\mathcal{N}(2, 2^2)$? Yes, it is possible, *but the probability of that is small*. Now, what about generating both $-3.1452570, -4.7189382$, the second and third observations, from $\mathcal{N}(2, 2^2)$? Of course, this is possible (every real number can be generated from $\mathcal{N}(2, 2^2)$, in fact), but the probability of that is verrrrry smaall. Now, what about generating the whole dataset from $\mathcal{N}(2, 2^2)$? It is possible, but the probability is veeeeeerrrry smmmmmaaaall. And what about our dataset being generated from the $\mathcal{N}(-2.4, 2^2)$, with red PDF? It is again possible, and the probability of that is muuch higher! In fact, the dataset is actually generated from that distribution, using **R**, you will find the code below.

Now, the idea of the MLE is simple: it is trying to find that value of the parameter, which will make the probability of having the observed dataset the maximal one.

R CODE, ABOVE MLE EXPLANATION EXAMPLE:

```
n <- 40 #no. of observations
x <- rnorm(n, mean = -2.4, sd = 2) #generating normal random numbers
y <- rep(0,n) #taking n 0-s, to form and plot the points (x_i,0) on the plain
plot(x,y, pch = 19, xlim = c(-10,10), ylim = c(0,0.25), xlab = "x", ylab = "y")


t <- seq(from = -10, to = 10, by = 0.01) #running parameter, to plot PDFs
z1 <- dnorm(t, mean = 2, sd = 2)
par(new = TRUE)
plot(t,z1, type = "l", col = "blue", lwd = 2, xlim = c(-10,10), ylim = c(0,0.25), xlab = "x", y
z2 <- dnorm(t, mean = -2.4, sd = 2)
```

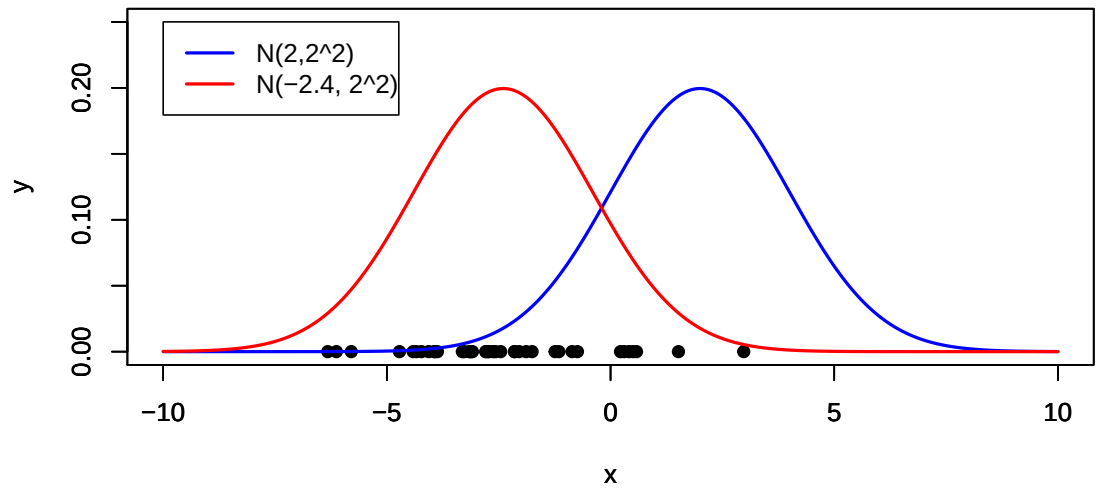

Fig. 10.1: Some generated dataset, and 2 Normal PDFs

```
par(new = TRUE)
plot(t,z2, type = "l", col = "red", lwd = 2, xlim = c(-10,10), ylim = c(0,0.25), xlab = "x", yl
legend(-10,0.25, legend = c("N(2,2^2)", "N(-2.4, 2^2)"), col = c("blue", "red"), lwd = 2)
```

REMARK, STEPS FOR MLE: So, to find the ML Estimator, one needs to do the following steps:

- Identify the parametric Model, Identify/write down the PD(M)F $f(x|\theta)$ of that Model
- Assume the data is generated by a random sample $X_1, ..., X_n$ from that Model;
- Form the Likelihood Function:

$$\mathcal{L}(\theta) = f(X_1|\theta) \cdot f(X_2|\theta) \cdot ... \cdot f(X_n|\theta)$$

- If necessary, form the Log-Likelihood Function,

$$\ell(\theta) = \ln \mathcal{L}(\theta) = \sum_{k=1}^{n} \ln f(X_k|\theta)$$

- Find the critical points of $\mathcal{L}(\theta)$ or $\ell(\theta)$ (choose the easiest one)
- Identify the global maximum point among the critical points (by using some calculus tools)

10.1.1 MLE for Bernoulli Model

We consider the model $\{\text{Bernoulli}(p) : p \in [0,1]\}$, and assume

$$X_1, X_2, ..., X_n \overset{\text{IID}}{\sim} \text{Bernoulli}(p).$$

Our aim is to estimate the unknown parameter p, based on the given random sample $X_1, ..., X_n$. We use here the MLE method to construct an estimator. Since our model is discrete, we calculate the PMF:

$$f(x|p) = \text{PMF}(x,p) = \begin{cases} p, & x = 1 \\ 1-p, & x = 0. \end{cases}$$

This function can be written as[4] $f(x|p) = p^x \cdot (1-p)^{1-x}$ for $x \in \{0,1\}$.

Now, the likelihood function will be

$$\mathcal{L}(p) = \prod_{k=1}^{n} f(X_k|p) = \prod_{k=1}^{n} p^{X_k} \cdot (1-p)^{1-X_k} = p^{X_1+...+X_n} \cdot (1-p)^{n-(X_1+...+X_n)} = p^{n\overline{X}} \cdot (1-p)^{n(1-\overline{X})},$$

where $\overline{X} = \frac{\sum_{k=1}^{n} X_k}{n}$ is the Sample Mean of $X_1, ..., X_n$. Our aim is to find the maximum point of this function with respect to the variable $p \in [0,1]$.

We will consider the case when $p \in (0,1)$, leaving the reader to ponder about the cases when $p = 0$ or $p = 1$. Now, since $f(x|p) > 0$ for any $x \in \{0,1\}$ and $p \in (0,1)$, we can calculate the log-likelihood function:

$$\ell(p) = \ln \mathcal{L}(p) = n \cdot \overline{X} \cdot \ln p + n \cdot (1-\overline{X}) \cdot \ln(1-p), \qquad p \in (0,1).$$

Now,

$$\ell'(p) = \frac{n\overline{X}}{p} - \frac{n(1-\overline{X})}{1-p}.$$

We solve

$$\ell'(p) = 0,$$

and obtain the unique critical point $p = \overline{X}$. To prove that this point is actually the global maximum point of ℓ, we calculate

$$\ell''(p) = -\frac{n\overline{X}}{p^2} - \frac{n(1-\overline{X})}{(1-p)^2}.$$

Clearly, $0 \leqslant \overline{X} \leqslant 1$ (since all X_k-s are Bernoulli r.v.s, and they can take only the values 0 or 1), so

$$\ell''(p) \leqslant 0, \qquad p \in (0,1).$$

So ℓ is concave, and we know that the critical point of a concave function is the global maximum point. So $p = \overline{X}$ is the unique global maximum point of ℓ:

$$\hat{p} = \hat{p}_n^{ML} = \overline{X} = \frac{X_1 + ... + X_n}{n}.$$

This is the MLE estimator for the parameter p in the Benoulli case. Calculus Rulez!

Now, the properties of this estimator: We know that

- $\hat{p}_n^{MLE}$ is an unbiased estimator for p;
- $\hat{p}_n^{MLE}$ is strongly consistent (so also consistent);

[4] You can check just by calculating the values at $x = 0$ and $x = 1$ or use the relation $\text{Bernoulli}(p) = \text{Binom}(1,p)$.

- $\hat{p}_n^{MLE}$ is efficient;
- The Quadratic Risk $MSE(\hat{p}_n^{MLE}, p) = Var_p(\hat{p}_n^{MLE}) = \frac{Var_p(X_k)}{n} = \frac{p(1-p)}{n}$, and this risk tends to 0 at a rate $\frac{1}{n}$.

■

Remark: Again, if you are interested in getting an **estimator** using the ML Method, then use the MLE method for $\mathcal{L}_n$ with random sample $X_1, ..., X_n$. But if you are interested only in getting just one estimate based on a single observation, without considering how good is your estimate, you can maximize $\mathcal{L}_n$ by using the values $X_k = x_k$. Then you will ge just a number for an estimate of your parameter.

EXAMPLE, BERNOULLI MODEL MLE: Assume we have a box full of red and blue balls. We do not know the number of balls inside, so we do not know the probability of drawing red (blue) ball. Assume we perform a 4 independent trials: we draw a ball at random, fix its color, and return to box. Say, we have the following outcome: RRBR. Now we want to estimate, based on this observation, the probability of drawing red (blue). Intuitively, since the number of red balls drawn is larger, then the probability of a red ball is higher. Let us estimate that probability by ML Method. Here ML Method will give that value of our probability, that makes our observation RRBR most likely.

Clearly, we have a Bernoulli distribution: denote $X = 1$, if we draw red ball, and $X = 0$, if blue. Let p be the unknown probability to draw red ball, then the probability of blue ball is $1 - p$. In this case, $X \sim Bernoulli(p)$, and we need to estimate p.

We have the following independent observations: $x_1 = 1$, $x_2 = 1$, $x_3 = 0$, $x_4 = 1$. First let us construct an estimator. We take

$$X_1, X_2, X_3, X_4 \overset{IID}{\sim} Bernoulli(p).$$

The PMF of each X_k is

$$f(x|p) = p^x(1-p)^{1-x}, \qquad x = 0, 1,$$

so the Likelihood function is:

$$\mathcal{L}(p) = f(X_1|p) \cdot f(X_2|p) \cdot f(X_3|p) \cdot f(X_4|p) =$$
$$= p^{X_1}(1-p)^{1-X_1} \cdot p^{X_2}(1-p)^{1-X_2} \cdot p^{X_3}(1-p)^{1-X_3} \cdot p^{X_4}(1-p)^{1-X_4} =$$
$$= p^{X_1+X_2+X_3+X_4} \cdot (1-p)^{4-X_1-X_2-X_3-X_4} = p^S \cdot (1-p)^{4-S},$$

where $S = X_1 + X_2 + X_3 + X_4$. Btw, S can take only values $0, 1, 2, 3, 4$ (since X_k-s are either 0 or 1).

We need to find the global maximum point of $\mathcal{L}(p)$, and that global maximum point is exactly the global maximum point of the log-Likelihood function

$$\ell(p) = \ln \mathcal{L}(p) = S \cdot \ln p + (4 - S) \cdot \ln(1-p), \qquad p \in (0, 1).$$

Now, to find the maximum point of ℓ, we calculate its derivative:

$$\ell'(p) = \frac{S}{p} - \frac{4-S}{1-p} = \frac{S-4p}{p(1-p)}.$$

We find all critical points by solving $\ell'(p) = 0$, and this gives $p = \frac{S}{4}$. Now, we need to check that this point is the global maximum point, which can be seen from the value of ℓ', which is positive, if $p \in (0, \frac{S}{4})$ and negative in $(\frac{S}{4}, 1)$, so our function ℓ increases up to the point $p = \frac{S}{4}$, then decreases[5].

Hence, the MLE is

$$\hat{p} = \hat{p}^{MLE} = \frac{S}{4} = \frac{X_1 + X_2 + X_3 + X_4}{4}.$$

Using our observation, and plugging $X_k = x_k$, $k = 1, 2, 3, 4$, we can get the estimate for p:

$$\hat{p} = \frac{1+1+0+1}{4} = \frac{3}{4}.$$

Example, Stock Price Rate of the Return Prediction: Of course, I will not predict the RoR (rate of the return) of the Stock price ☺ People are earning billions on this! But will do some simple estimation.

I want to have a Bernoulli model, so I consider the following problem: estimate the probability p that the weekly RoR for the FB stock will be higher than 1%, i.e., higher than 0.01. This problem can be modelled by a Bernoulli distribution, since we have two cases: either the RoR is higher than 0.01 (this is our "success") or it is less than or equal to 0.01 (this is the "failure"). So if X is the weekly RoR for the FB Stock, then $X \sim \text{Bernoulli}(p)$: that is, $X = 1$ with probability p, and this means that the $\text{RoR} > 0.01$, and $X = 0$ with probability $1 - p$, when $\text{RoR} \leqslant 0.01$.

Now, we have a problem, and we need to model it. As very professional statisticians, we first design the experiment. Based on what we will estimate? - Of course, we need to have some observations, observations on the weekly RoR-s, so we need to calculate weekly RoRs for some weeks for which we have data. And, of course, we need to take care how many observations we need to have/calculate. This question will be answered in the next section, when talking about Confidence Intervals. So at this moment we assume that it is enough to have weekly RoRs for 1 year, that is, for 52 weeks. So we will calculate the weekly RoRs for the last 52 weeks. Before calculating these numbers, we need to do the modeling, choose an estimator, check that it is a good estimator. Then, if we will have all these, it will remain to calculate that numbers, plug into the estimator, and get the estimate for p.

Now, we model this by

$$X_1, X_2, ..., X_{52} \overset{IID}{\sim} \text{Bernoulli}(p), \qquad p \in [0, 1]$$

where X_1 will be the result for the first considered week (i.e., X_1 will be 1, if the first weeks RoR will be > 0.01, and X_1 will be 0 otherwise). Similarly, X_{52} will be the result for the 52-nd week. Before observing these numbers, they are r.v.s[6]!

What we will do with these? Of course, our aim is to estimate p, so we need to construct a good estimator! And, of course, we will choose the MLE for p, so our estimator is

$$\hat{p} = \frac{X_1 + X_2 + ... + X_{52}}{52}.$$

We know that this estimator shares all possible good properties we have talked about, so we can trust the the estimate obtained from this estimator. Perfect! It remains to get the observations.

Now, we open the FB historical data page at Yahoo Finance: https://finance.yahoo.com/quote/FB/history?p=FB . We choose here Time Period: 1Y, Frequency: Weekly, and then click on the "Download Data". That will download the file *FB(1).csv* to your download folder (well, FB(1).csv, because I clicked twice ☺). Now we import the data into **R**:

```
treshold <- 0.01 #the treshold, 0.01 in our problem; you can check also with 0, say
```

```
prices <- read.csv(file.choose()) #Choose the Downloaded File "FB(1).csv"
adjclose <- prices$Adj.Close #We choose the Adjusted Close Prices
ror <- (adjclose[-1]- adjclose[-length(adjclose)])/adjclose[-length(adjclose)] #RoRs
hist(ror) # Just to see the picture
x <- (ror > treshold) # will give a vector, with true, if ror > treshold
no <- sum(x) # will calculate the number of cases when ror > treshold, that is, x_1+...+x_{52}
p_hat <- no/length(x) # this is giving \hat{p}, the estimate for p
p_hat
```

The result is (I am running this on Nov 25, 2018) $\hat{p} = 0.3962264$. So our estimate is that with probability almost 0.4, the weekly RoR of FB stock will be > 0.01. In other words, with probability 0.4, the FB stock price will increase in more than 1% in a week.

Not a billion dollar result, but kdzgi for some millions.

10.1.2 MLE for the Gaussian Distribution

Assume our parametric model is $\{\mathcal{N}(\mu, \sigma^2) : (\mu, \sigma^2) \in \mathbb{R} \times [0, +\infty)\}$ with the unknown parameters[7] μ and σ^2. We will denote our parameter $\sigma^2 = \theta$ not to get confused when calculating the derivatives[8], and $(\mu, \theta) \in \Theta = \mathbb{R} \times [0, +\infty)$. Assume we have a random sample

$$X_1, ..., X_n \overset{\text{IID}}{\sim} \mathcal{N}(\mu, \theta),$$

and we want to get a ML Estimator for (μ, θ). We form the Likelihood Function: since $X_k \sim \mathcal{N}(\mu, \theta)$, then the PDF of X_k will be[9]

$$f(x|\mu, \theta) = \frac{1}{\sqrt{2\pi\theta}} \cdot \exp\left\{-\frac{(x-\mu)^2}{2\theta}\right\},$$

so the likelihood function is

$$\mathcal{L}(\mu, \theta) = \mathcal{L}(X_1, ..., X_n|\mu, \theta) = f(X_1|\mu, \theta) \cdot f(X_2|\mu, \theta) \cdot ... \cdot f(X_n|\mu, \theta) =$$

$$= \frac{1}{\sqrt{(2\pi\theta)^n}} \cdot \exp\left\{-\frac{\sum_{k=1}^n (X_k - \mu)^2}{2\theta}\right\} = \left(\frac{1}{2\pi\theta}\right)^{n/2} \cdot \exp\left\{-\frac{\sum_{k=1}^n (X_k - \mu)^2}{2\theta}\right\}.$$

Now, the Log-Likelihood Function is

$$\ell(\mu, \theta) = \ln \mathcal{L}(\mu, \theta) = \ln\left[\left(\frac{1}{2\pi\theta}\right)^{n/2} \cdot \exp\left\{-\frac{\sum_{k=1}^n (X_k - \mu)^2}{2\theta}\right\}\right] =$$

$$= \ln\left(\frac{1}{2\pi\theta}\right)^{n/2} + \ln\exp\left\{-\frac{\sum_{k=1}^n (X_k - \mu)^2}{2\theta}\right\} = -\frac{n}{2} \cdot \log(2\pi\theta) - \frac{1}{2\theta} \cdot \sum_{k=1}^n (X_k - \mu)^2.$$

To find the maximum point of the Log-Likelihood function, we solve the system

$$\frac{\partial \ell}{\partial \mu} = 0 \qquad \text{and} \qquad \frac{\partial \ell}{\partial \theta} = 0.$$

[7] You can consider also the case, when the parameters are μ and σ. Try to do that! Also you can consider the case, when one of the parameters is known, and do the MLE for the unknown one.

[8] We need to calculate the derivative of our Likelihood function w.r.t. σ^2. Say, if you will have σ^2 and calculate that derivative, you will get 1, not 2σ. This is because our variable is σ^2.

[9] Here $\exp(a) = e^a$

Calculating the derivatives and solving the system will yield to a unique critical point (μ, θ) with

$$\mu = \overline{X} \qquad \text{and} \qquad \theta = \frac{\sum_{k=1}^{n}(X_k - \overline{X})^2}{n},$$

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

Now, using some calculus tools (use!!) one can prove that the critical point obtained above is actually the global maximum point of the Log-Likelihood function. Hence, the ML Estimator is

$$\hat{\mu}^{MLE} = \overline{X} \qquad \text{and} \qquad \hat{\sigma^2}^{MLE} = \hat{\theta}^{MLE} = \frac{\sum_{k=1}^{n}(X_k - \overline{X})^2}{n}.$$

Now, about the statistical properties of this estimator:

- $\hat{\mu}^{MLE}$ is unbiased, but $\hat{\sigma^2}^{MLE}$ is biased;
- Both $\hat{\mu}^{MLE}$ and $\hat{\sigma^2}^{MLE}$ are consistent;
- $\hat{\mu}^{MLE}$ is efficient, but $\hat{\sigma^2}^{MLE}$ is not efficient[10].

10.1.3 MLE for the Exponential Model

We consider the model $\{Exp(\lambda) : \lambda \in (0, +\infty)\}$, and assume

$$X_1, X_2, ..., X_n \overset{IID}{\sim} Exp(\lambda).$$

Our aim is to estimate the unknown parameter λ, based on the given random sample $X_1, ..., X_n$. We use here the MLE method to construct an estimator.

Performing the calculations, one will obtain $\hat{\lambda} = \frac{1}{\overline{X}}$. It is not so easy to study the properties of this estimator, but it can be proved that:

- $\hat{\lambda}$ is a biased estimator[11] for λ:
$$\mathbb{E}_\lambda(\hat{\lambda}) = \frac{n}{n-1}\lambda.$$
- $\hat{\lambda}$ is strongly consistent, because of the Strong LLN
- What about the Quadratic Risk?

■

[10] Please look at the previous paragraph for the proofs/ideas

[11] This can be shown by using a formula found in Fikhtengolts, Part 3, page 451 (Russian edition): we need to calculate

$$\mathbb{E}_\lambda(\hat{\lambda}) = \mathbb{E}_\lambda\left(\frac{n}{X_1+X_2+...+X_n}\right) = \int\limits_{[0,+\infty)^n} \frac{n}{x_1+x_2+...+x_n} \cdot \lambda^n \cdot e^{-\lambda x_1} \cdot ... \cdot e^{-\lambda x_n}\, dx_1...dx_n = \int\limits_{[0,+\infty)^n} \frac{n}{x_1+x_2+...+x_n} \cdot \lambda^n \cdot e^{-\lambda(x_1+...+x_n)}$$

In fact, it can be shown that

$$\mathbb{E}\left(\frac{1}{\overline{X}}\right) = n \cdot \int_0^{+\infty} \frac{1}{t} \cdot \frac{\lambda^n t^{n-1} e^{-\lambda t}}{(n-1)!} = \frac{n}{n-1}\lambda.$$

10.1.4 MLE for the Cauchy Model

We consider a (one-)parametric family of Cauchy Distributions $\{Cauchy(\theta) : \theta \in \mathbb{R}\}$ given by PDFs[12]

$$f(x|\theta) = \frac{1}{\pi(1+(x-\theta)^2)}, \qquad \theta \in \mathbb{R}.$$

Assume we have a random sample from that model,

$$X_1, ..., X_n \overset{IID}{\sim} Cauchy(\theta),$$

and we want to estimate the unknown parameter θ by using the MLE.

In this case we will obtain an equation for MLE, which cannot be solved explicitly. So it is not possible to obtain the ML Estimator explicitly. For the ML Estimate, one needs to solve the obtained equation numerically. Also, one can have multiple solutions for that equation!

Do the calculations! ■

10.1.5 MLE for the Uniform Distribution

Consider the parametric family of distributions $\{Unif([0,\theta]) : \theta \in (0,+\infty)\}$, and assume we have a random sample from one of the distributions of that family:

$$X_1, X_2, ..., X_n \overset{IID}{\sim} Unif([0,\theta])$$

for some $\theta \in (0,+\infty)$. Our aim is to estimate θ.

For that, we construct the ML Estimator for θ.

We know that the PDF for a r.v. with uniform distribution in $[0,\theta]$ is

$$f(x|\theta) = \begin{cases} \frac{1}{\theta}, & \text{for } 0 \leqslant x \leqslant \theta \\ 0, & \text{for } x \notin [0,\theta]. \end{cases}$$

So if we will consider $f(x|\theta)$ as a function of $\theta \in (0,+\infty)$, we will get for $x > 0$,

$$f(x|\theta) = \begin{cases} \frac{1}{\theta}, & \text{for } \theta \geqslant x \\ 0, & \text{for } 0 < \theta < x. \end{cases}$$

Now, the Likelihood Function for the ML Estimator will be

$$\mathcal{L}(\theta) = \mathcal{L}(X_1, ..., X_n|\theta) = f(X_1|\theta) \cdot ... \cdot f(X_n|\theta) = \begin{cases} \frac{1}{\theta^n}, & \text{for } \theta \geqslant X_{(n)} \\ 0, & \text{for } 0 < \theta < X_{(n)}, \end{cases}$$

where, as usual, $X_{(n)} = \max\{X_1, ..., X_n\}$ is the n-th order statistics. This can be explained in the following way: $f(X_1|\theta)$ is non-zero for $\theta \geqslant X_1$, $f(X_2|\theta)$ is non-zero for $\theta \geqslant X_2$ etc. . So $\mathcal{L}(\theta)$ will be non-zero, if $\theta \geqslant X_k$ for any $k = 1, ..., n$, which is equivalent to $\theta \geqslant \max\{X_1, X_2, ..., X_n\} = X_{(n)}$.

Now, our function is not differentiable (and is not continuous either) on $\theta \in (0,+\infty)$, so we cannot calculate the maximum point by using derivatives.

[12] See https://en.wikipedia.org/wiki/Cauchy_distribution . Cauchy Distribution is important in Physics. It can be proved that it has no expected value and variance. Also, if X, Y are independent from standard Normal Distribution, then $\frac{X}{Y}$ has a Standard Cauchy Distribution (with $\theta = 0$).

But clearly (see also the graph!! Do that Graph!) $\mathcal{L}(\theta) = 0$ for $\theta \in (0, X_{(n)})$, and $\mathcal{L}(\theta) = \frac{1}{\theta^n}$ decreases on $[X_{(n)}, +\infty)$, so the unique maximum point is

$$\hat{\theta} = \hat{\theta}^{ML} = X_{(n)},$$

and this is our ML Estimator for θ.

REMARK, MLE FOR UNIFORM, INTUITION: The intuition behind the MLE estimate for the parameter θ in the above case is the following: assume we have an observation $0.3, 0.5, 1.2, 3.21, 1.71$ from a Uniform Distribution $\mathrm{Unif}[0, \theta]$. Then, the ML Estimate (not the Estimator!) for θ is $\hat{\theta} = \max\{0.3, 0.5, 1.2, 3.21, 1.71\} = 3.21$. So what is the intuition behind?

First, naturally, θ cannot be less than 3.21. Say, our dataset cannot be generated from $\mathrm{Unif}[0, 3]$, since in that case we will not have 3.21 (the probability of having that number is 0). What about larger values? Say, why not $\mathrm{Unif}[0, 50]$? Of course, our dataset can be generated from $\mathrm{Unif}[0, 50]$. But the thing is that the chance of being generated from $\mathrm{Unif}[0, 50]$ is muuuch smaaaler, than from $\mathrm{Unif}[0, 3.21]$ - when generating from $\mathrm{Unif}[0, 50]$ we have high probability to get a number larger than, say, 4: if $X \sim \mathrm{Unif}[0, 50]$, then

$$\mathbb{P}(X > 4) = \int_4^{+\infty} f_X(x)dx = \int_4^{50} \frac{1}{50}dx = \frac{44}{50} = 0.88,$$

so with probability 88%, the point generated from $\mathrm{Unif}[0, 50]$ will be larger than 4! So having a point $\leqslant 4$ generated from $\mathrm{Unif}[0, 50]$ has a probability 12%. And think about - we generate 5 numbers, and every time we get numbers less than 4 - what is the probability of that, will it be high? Of course, no! (Btw, what is the probability of that?)

Now, about the properties of this estimator:

- The Estimator $\hat{\theta}^{ML} = X_{(n)}$ is biased. To show this, first we calculate the PDF for $X_{(n)}$, and for that we calculate the CDF of $X_{(n)}$:

$$F_{X_{(n)}}(x) = \mathbb{P}(X_{(n)} \leqslant x) = \mathbb{P}(\max\{X_1, ..., X_n\} \leqslant x) = \mathbb{P}(X_1 \leqslant x, X_2 \leqslant x, ..., X_n \leqslant x)$$

$$\overset{X_k \text{ are indep}}{=\!=} \mathbb{P}(X_1 \leqslant x) \cdot \mathbb{P}(X_2 \leqslant x) \cdot ... \cdot \mathbb{P}(X_n \leqslant x)$$

$$\overset{X_k \text{ are identically distributed}}{=\!=\!=} [\mathbb{P}(X_1 \leqslant x)]^n \overset{X_k \sim \mathcal{U}([0,\theta])}{=\!=} \begin{cases} 0, & \text{if } x < 0; \\ \left(\frac{x}{\theta}\right)^n, & \text{if } x \in [0, \theta]; \\ 1, & \text{if } x > 1. \end{cases}$$

So the PDF of $X_{(n)}$ is

$$f_{X_{(n)}}(x) = \begin{cases} \frac{n}{\theta^n} \cdot x^{n-1}, & \text{if } x \in (0, \theta); \\ 0, & \text{otherwise.} \end{cases}$$

Now,

$$\mathbb{E}_\theta(\hat{\theta}_n^{ML}) = \mathbb{E}_\theta(X_{(n)}) = \int_{-\infty}^{+\infty} x \cdot f_{X_{(n)}}(x)dx = \int_0^\theta x \cdot \frac{n}{\theta^n} \cdot x^{n-1} = \frac{n}{n+1} \cdot \theta.$$

The bias of this estimator is

$$\mathrm{bias}(\hat{\theta}_n^{ML}, \theta) = \frac{n}{n+1}\theta - \theta = -\frac{\theta}{n}.$$

Since the bias is less that zero, then our estimator is underestimating the actual value of the parameter.

Intuitively, we have that the values of X_k are from $[0,\theta]$ a.s., so the max of them, $X_{(n)}$ will take values in $[0,\theta]$, so the average value of $X_{(n)}$ over a lot of observations cannot be equal to θ, it will be less than θ, it will underestimate θ.

- Now, let us talk about the consistency. We need to check if $\hat{\theta}_n^{ML} \xrightarrow{\mathbb{P}} \theta$ as $n \to \infty$: for any $\varepsilon > 0$,

$$\mathbb{P}(|\hat{\theta}_n^{ML} - \theta| \leqslant \varepsilon) = \mathbb{P}(\theta - \varepsilon \leqslant \hat{\theta}_n^{ML} \leqslant \theta + \varepsilon) = \mathbb{P}(\theta - \varepsilon \leqslant X_{(n)} \leqslant \theta + \varepsilon) =$$

$$= F_{X_{(n)}}(\theta + \varepsilon) - F_{X_{(n)}}(\theta - \varepsilon) = 1 - \left(\frac{\theta - \varepsilon}{\theta}\right)^n \to 1, \qquad n \to +\infty,$$

so

$$\mathbb{P}(|\hat{\theta}_n^{ML} - \theta| > \varepsilon) = 1 - \mathbb{P}(|\hat{\theta}_n^{ML} - \theta| \leqslant \varepsilon) \to 0.$$

This means that $\hat{\theta}_n^{ML} \xrightarrow{\mathbb{P}} \theta$, and our estimator is consistent;

- In fact, our estimator $X_{(n)}$ is strongly consistent. To prove this, we use the following fact: $X_{(n)}$ is an increasing sequence and $X_{(n)} \leqslant \theta$ a.s. (since each $X_k \leqslant \theta$ a.s., because of $X_k \sim \mathcal{U}(0,\theta)$). So by monotone convergence theorem,

$$X_{(n)} \xrightarrow{a.s.} X$$

for some r.v. X, and $X \leqslant \theta$ a.s. . Now we want to prove that $X = \theta$ a.s.. To this end, for each $\varepsilon > 0$, we calculate

$$\mathbb{P}(X_{(n)} > \theta - \varepsilon) = 1 - \mathbb{P}(X_{(n)} \leqslant \theta - \varepsilon) = 1 - \prod_{k=1}^{n} \mathbb{P}(X_k \leqslant \theta - \varepsilon) = 1 - \left(\frac{\theta - \varepsilon}{\theta}\right)^n \to 1,$$

as $n \to +\infty$. Now, $X \geqslant X_{(n)}$ for each n, since $X_{(n)} \uparrow X$, so

$$1 \geqslant \mathbb{P}(X > \theta - \varepsilon) \geqslant \mathbb{P}(X_{(n)} > \theta - \varepsilon) \to 1,$$

implying that $\mathbb{P}(X > \theta - \varepsilon) = 1$. This holds for any $\varepsilon > 0$, and this will give[13] that $\mathbb{P}(X \geqslant \theta) = 1$. And with $X \leqslant \theta$ a.s. this will give

$$\mathbb{P}(X = \theta) = 1.$$

Ufff....

- Having the distribution of $X_{(n)}$, we can calculate also the quadratic risk:

Do the rest, use the bias-variance decomposition! ■

I should have warned you that no need to read about the properties of this estimator. But I forgot, sorry ☺ Although I am pretty sure you have skipped the above part ☺

10.1.6 MLE for Uniform Distribution, another parametric model

We consider the parametric model $\{\mathtt{Unif}(\theta, \theta+2) : \theta \in \mathbb{R}\}$.

In this case we will obtain

$$\mathcal{L}(\theta) = \begin{cases} \frac{1}{2^n}, & X_{(n)} - 2 \leqslant \theta \leqslant X_{(1)} \\ 0, & \text{otherwise} \end{cases}$$

[13] Prove this!

So every point of $[X_{(n)}-2, X_{(1)}]$ is a max point[14] of $\mathcal{L}$, so in this case we have an infinite number of estimators. For example, we can take

$$\hat{\theta} = \alpha \cdot (X_{(n)} - 2) + (1-\alpha) \cdot X_{(1)},$$

where $\alpha \in [0,1]$ is arbitrary. ■

Exercise: Find the MLE for the model $\{\mathtt{Unif}([\theta, 5\theta] : \theta \in (0, +\infty))\}$.

10.1.7 Properties of MLE

Why people are using the MLE? Because it gives very nice estimators, and the idea behind MLE is simple and powerful. Now, let us state some nice properties of MLE Estimators. We have studied the properties of the estimators obtained above for our Klassik models, and here we want to give the general MLE properties.

Theorem 10.1. *Under some regularity conditions*[15] *on the Parametric family PD(M)F* $f(x|\theta)$*, the MLE* $\hat{\theta}^{MLE}$ *of the parameter* θ *possesses the following properties:*

1. $\hat{\theta}^{MLE}$ *is asymptotically unbiased;*
2. $\hat{\theta}^{MLE}$ *is consistent;*
3. $\hat{\theta}^{MLE}$ *is asymptotically efficient*[16].

So, basically, if you have some large amount of observations, the MLE estimator will work the best, you will (almost) never beat the MLE results by other Estimators! That's why MLE is very popular!

Another important property for the MLE is its Invariance:

Sometimes, when dealing with parametric models, one is interested not (or not only) in the estimation of the parameter itself, but of some function of the parameter. Say, we have considered above the MLE for the Gaussian model $\mathcal{N}(\mu, \sigma^2)$, with two parameters - μ and σ^2. But one can also be interested in estimating the Standard Deviation σ, say. The next proposition says that one can calculate the MLE estimator for σ, having the MLE estimator for σ^2.

Proposition 10.1. *Assume we work with the parametric family of distributions with parameter* $\theta \in \Theta$ *and* $\hat{\theta}_n$ *is the MLE for* θ*. If* g *is any function of the parameter, then the MLE for the parameter* $g(\theta)$ *will be* $g(\hat{\theta})$*, that is,*

$$\textit{MLE for } g(\theta) = g(\textit{MLE for } \theta),$$

or, in other words,

$$\widehat{g(\theta)}^{MLE} = g(\hat{\theta}^{MLE}).$$

10.1.8 MLE Example, Multinomial Case

[14] Prove that this interval is not empty. Recall that X_k takes values from $[\theta, \theta+2]$ a.s.
[15] Give a citation here!
[16] Give the definition here

EXAMPLE, MLE FOR MULTINOMIAL CASE: We consider the following problem: assume we have 3 political parties in Armenia, A, B and C. We want to estimate the percentage p_A of persons who supports A, the percentage p_B of persons who supports B, p_C - for supporters for C, and persons p_D who do not support any of these three parties (we assume, to avoid complications, that each person is either supporting exactly one of A, B, C or neither one). To that end we randomly pick 100 persons (with replacement, to have independent trials), and ask about their preferences. Here are the results:

Parties	A	B	C	NONE
Number of supporters	23	35	17	25

based on this result, we want to estimate p_A, p_B, p_C and p_D.

We model this problem in the following way. Let X be the choice of a random person - we can assume $X = 1$, if the person is preferring A, $X = 2$, if B is his/her choice, $X = 3$ for C, and $X = 0$, if he/she is none of the parties are preferable[17]. The probability that he/she is preferring the party A is p_A, the probability of preference for B is p_B, and correspondingly, p_C and p_D for C and NONE. Clearly, $p_A + p_B + p_C + p_D = 1$.

So our political orientation distribution (for a person) is

Values of X	1	2	3	0
$\mathbb{P}(X = x)$	p_A	p_B	p_C	$1 - p_A - p_B - p_C$

and we have 3 parameters: p_A, p_B and p_C. We want to construct the ML estimates for these parameters.

To that end, we construct the Likelihood function, the joint probability of our observation: the probability that we will have 23 supporters of party A is p_A^{23}, the probability to have 35 supporters of B is p_B^{35}, the probability to have 17 supporters of party C is p_C^{17}, and, finally, the probability of having 25 non-party-supporters is $(1 - p_A - p_B - p_C)^{25}$. And, in totality, the probability to have the complete data we have observed is

$$\mathcal{L}(p_A, p_B, p_C) = p_A^{23} \cdot p_B^{35} \cdot p_C^{17} \cdot (1 - p_A - p_B - p_C)^{25}.$$

We need to find the global maximum point for this function, and that will be our MLE estimates for the parameters. We pass to log-Likelihood to simplify calculations:

$$\ell(p_A, p_B, p_C) = \ln \mathcal{L}(p_A, p_B, p_C) = 23 \cdot \ln p_A + 35 \cdot \ln p_B + 17 \cdot \ln p_c + 25 \cdot \ln(1 - p_A - p_B - p_C).$$

To find the global max point of ℓ, we solve the system:

$$\begin{cases} \dfrac{\partial \ell}{\partial p_A} = 0 \\ \dfrac{\partial \ell}{\partial p_B} = 0 \\ \dfrac{\partial \ell}{\partial p_C} = 0 \end{cases} \Leftrightarrow \begin{cases} \dfrac{\partial \ell}{\partial p_A} = \dfrac{23}{p_A} - \dfrac{25}{1 - p_A - p_B - p_C} = 0 \\ \dfrac{\partial \ell}{\partial p_B} = \dfrac{35}{p_B} - \dfrac{25}{1 - p_A - p_B - p_C} = 0 \\ \dfrac{\partial \ell}{\partial p_C} = \dfrac{17}{p_C} - \dfrac{25}{1 - p_A - p_B - p_C} = 0 \end{cases}$$

From this system we can deduce that

$$\frac{23}{p_A} = \frac{35}{p_B} = \frac{17}{p_C} \overset{\text{denote}}{=\!=} \frac{1}{\lambda}.$$

Hence,

$$p_A = 23\lambda, \qquad p_B = 35\lambda, \qquad p_C = 17\lambda.$$

From the first equation of the system above, plugging these values, we obtain

$$\frac{1}{\lambda} - \frac{25}{1-75\lambda} = 0,$$

so $\lambda = \frac{1}{100}$. This means that

$$p_A = \frac{23}{100}, \qquad p_B = \frac{35}{100}, \qquad p_C = \frac{17}{100}, \qquad \text{and} \qquad p_D = \frac{25}{100}.$$

This is the ML Estimates for our parameters! K Vashim Uslugam, Dami i Gospoda!

Of course, we need to prove that this point is a global max point of ℓ, but "we" means "me" + "you". I have calculated my part, the rest is yours ☺

10.2 Method of Moments

Now we describe another method to construct estimates and estimators. Although MLE is very nice, and, as stated above, one cannot (almost cannot) beat MLE, but in many cases MLE calculation is very hard to obtain analytically, and even in many cases it is not possible (and one relies on numerical methods to calculate the approximate MLE). The point is that optimization problems (say, for the multivariable case) are hard problems to solve, even with computers. So we want to have some alternative method that will be easier to use (although, the result will work much worse than MLE).

We assume again that we have a random sample

$$X_1, ..., X_n \overset{IID}{\sim} \mathbb{P}_\theta$$

from the family of distributions $\{\mathbb{P}_\theta : \theta \in \Theta\}$, and our aim is to estimate θ.

Before describing the MoM Method, let us give the definitions of the Theoretical Moment of the Distribution and the Empirical Moment of that Distribution.

Definition 10.4. *Assume $X \sim \mathbb{P}_\theta$. The k-the order **Theoretical Moment** of X is the expected value $\mathbb{E}(X^k)$.*

Since X comes from the parametric family with the parameter θ, usually, the Theoretical Moments depend on θ. Also, let me recall the formulas to calculate that expectations:

- If $\mathbb{P}_\theta$ is continuous with the PDF $f(x|\theta)$, then the k-the order Moment is equal to
$$\mathbb{E}(X^k) = \int_{-\infty}^{+\infty} x^k \cdot f(x|\theta)\, dx.$$
- If $\mathbb{P}_\theta$ is discrete with possible values x_i and PMF $f(x_i|\theta) = \mathbb{P}_\theta(X = x_i)$, then the k-the order Moment is equal to
$$\mathbb{E}(X^k) = \sum_{x_i} (x_i)^k \cdot f(x_i|\theta).$$

Now, about the Empirical Moments:

Definition 10.5. *Assume we have a random Sample $X_1, ..., X_n \overset{IID}{\sim} \mathbb{P}_\theta$. The k-the order **Empirical Moment** of $\mathbb{P}_\theta$, using the observations X_k, is the mean*

$$\frac{(X_1)^k + (X_2)^k + ... + (X_n)^k}{n}.$$

The empirical moment is independent of the parameter, it is just calculated using the observations only.

Now, the idea of the Method of Moments is the following: to estimate the parameters of our model, solve equations of the form

$$k\text{-th order Theoretical Moment} = k\text{-th order Empirical Moment}$$

for some values of k-s. The obtained values of parameters are the Method of Moments Estimators (MME) for that parameters.

Usually, the number of equations is equal to the number of parameters in our model. Say, if we have a 1-parametric family, then we need to use just one equation of the above type to find that parameter's value. One begins by taking $k = 1$, and tries to solve

$$1\text{-st order Theoretical Moment} = 1\text{-st order Empirical Moment},$$

or, using the Moments values,

$$\mathbb{E}(X) = \frac{X_1 + X_2 + ... + X_n}{n} = \overline{X}.$$

The left-hand side usually depends on the unknown parameter θ. If we can solve this equation for θ and find its value expressed by $\overline{X}$, then we are done, and we are happy. This is our MoM Estimator, MME. Otherwise, if this equation will not give us the value of our parameter, we try to solve

$$2\text{-nd order Theoretical Moment} = 2\text{-nd order Empirical Moment},$$

i.e.,

$$\mathbb{E}(X^2) = \frac{(X_1)^2 + (X_2)^2 + ... + (X_n)^2}{n}.$$

If this will give the value for θ, then that value is the MME for θ, and we are done, and we are happy. In the other case, you continue to try with the 3-rd, 4-th,... moments until getting happy.

In the multi-parameter case, say, if we work with 2-parametric model (say, $\mathcal{N}(\mu, \sigma^2)$), we need to solve 2 equations to get the values for two unknowns. We start by

$$\begin{cases} 1\text{-st order Theoretical Moment} = 1\text{-st order Empirical Moment} \\ 2\text{-nd order Theoretical Moment} = 2\text{-nd order Empirical Moment} \end{cases}$$

that is,

$$\begin{cases} \mathbb{E}(X) = \frac{X_1+X_2+...+X_n}{n} = \overline{X} \\ \mathbb{E}(X^2) = \frac{(X_1)^2+(X_2)^2+...+(X_n)^2}{n} \end{cases}$$

If this is giving the values of our parameters, we are happy. Otherwise, we try with

$$\begin{cases} 2\text{-nd order Theoretical Moment} = 2\text{-nd order Empirical Moment} \\ 3\text{-rd order Theoretical Moment} = 3\text{-rd order Empirical Moment} \end{cases}$$

that is,

$$\begin{cases} \mathbb{E}(X^2) = \frac{(X_1)^2+(X_2)^2+...+(X_n)^2}{n} \\ \mathbb{E}(X^3) = \frac{(X_1)^3+(X_2)^3+...+(X_n)^3}{n} \end{cases}$$

hoping to get happy this time. If this is not working, you already know how to get happy - continue this way!

BTW, there is an anecdote on this:

```
У профессора спрашивают: – Скажите, а как вы
определяете: какую оценку поставить студенту на
экзамене? – Ко мне заходит студент, я задаю ему
вопрос, он на него не отвечает, мне становится все
ясно, я ставлю ему "два" и он уходит. – Ну а если
он отвечает на этот вопрос? – Я задаю ему второй
вопрос, он на него не отвечает, мне становится все
ясно, я ставлю ему "два" и он уходит. – А если он
отвечает и на этот вопрос? – Я задаю ему еще один
вопрос, он на него не отвечает, мне становится все
ясно, я ставлю ему "два" и он уходит. – И до каких
пор это продолжается? – А пока мне не станет все
ясно.
```

10.2.1 MoM for $\{Unif([0,\theta]) : \theta > 0\}$

Our model is: we are given

$$X_1, ..., X_n \overset{IID}{\sim} Unif[0,\theta],$$

and we want to estimate θ by MoM.

Here we have just one parameter, θ, so we need just one equation. Let's try

$$\text{1-st order Theoretical Moment} = \text{1-st order Empirical Moment},$$

that is,

$$\mathbb{E}(X) = \frac{X_1 + X_2 + ... + X_n}{n} = \overline{X}.$$

We know that for $X \sim Unif[0,\theta]$,

$$\mathbb{E}(X) = \frac{\theta}{2}.$$

So we will get

$$\frac{\theta}{2} = \frac{X_1 + ... + X_n}{n},$$

implying

$$\theta = 2\frac{X_1 + ... + X_n}{n} = 2\overline{X}.$$

This is our estimator by MoM:

$$\hat{\theta}^{MME} = 2\frac{X_1 + ... + X_n}{n} = 2\overline{X}.$$

Exercise: Study the properties of this estimator!

10.2.2 MoM for the Bernoulli Model

We consider the Parametric Model $\{Bernoulli(p) : p \in (0,1)\}$. We are given a random sample from one of the distributions of that family,

$$X_1, X_2, ..., X_n \overset{IID}{\sim} Bernoulli(p),$$

and our aim is to estimate p. Here we use the MoM Method.

Again our model has just one parameter, so we will solve one equation. We start by, as usual, from

$$\text{1-st order Theoretical Moment} = \text{1-st order Empirical Moment},$$

that is,

$$\mathbb{E}(X) = \frac{X_1 + X_2 + ... + X_n}{n} = \overline{X}.$$

In this case, for $X \sim \text{Bernoulli}(p)$,

$$\mathbb{E}(X) = p.$$

So we will have

$$\hat{p}^{MME} = \frac{X_1 + ... + X_n}{n} = \overline{X}.$$

The result coincides with the ML Estimator!

10.2.3 MoM for the Gaussian Model

We consider here 2D parametric model $\{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 \in (0, +\infty)\}$.

Assume our random sample

$$X_1, X_2, ..., X_n \overset{IID}{\sim} \mathcal{N}(\mu, \sigma^2),$$

comes from a Normal Distribution with unknown μ and σ^2, and we need to estimate that parameters basing on our sample. We use here the MoM. To that end, we need to solve the 2 equations of the form "Theor MoMent = Empirical MoMent". We use

$$\begin{cases} \text{1-st order Theoretical Moment} = \text{1-st order Empirical Moment} \\ \text{2-nd order Theoretical Moment} = \text{2-nd order Empirical Moment} \end{cases}$$

that is,

$$\begin{cases} \mathbb{E}(X) = \frac{X_1 + X_2 + ... + X_n}{n} = \overline{X} \\ \mathbb{E}(X^2) = \frac{(X_1)^2 + (X_2)^2 + ... + (X_n)^2}{n} \end{cases}$$

In this case, if $X \sim \mathcal{N}(\mu, \sigma^2)$, then

$$\mathbb{E}(X) = \mu, \qquad \text{and} \qquad \mathbb{E}(X^2) = Var(X) + (\mathbb{E}(X))^2 = \sigma^2 + \mu^2.$$

So we need to solve the system

$$\mu = \frac{X_1 + X_2 + ... + X_n}{n} = \overline{X}; \qquad \sigma^2 + \mu^2 = \frac{X_1^2 + X_2^2 + ... + X_n^2}{n},$$

in order to construct the MoM estimators for μ and σ^2. The solution is:

$$\hat{\mu}^{MME} = \frac{X_1 + X_2 + ... + X_n}{n} = \overline{X},$$

and

$$\widehat{\sigma^2}^{MME} = \frac{X_1^2 + X_2^2 + ... + X_n^2}{n} - \left(\overline{X}\right)^2 = \frac{\sum_{k=1}^{n} (X_k - \overline{X})^2}{n}.$$

The same, as from MLE!!

10.2.4 MoM for $\{\text{Unif}([a, b]) : (a, b) \in R^2, a \leqslant b\}$

Write the model and solution.

10.3 Additions

Question 1: Calculate the MLE/MoME for the parameter p in the $\text{Geometric}(p)$ model.

Question 2: Assume we have

$$X_1, X_2, ..., X_n \overset{\text{IID}}{\sim} \mathrm{Unif}[a, b].$$

Estimate, using MLE/MoME, the parameters a and b.

Question 3: We are uniformly throwing points into the rectangle $[0, a] \times [0, b]$, and we have some observation, i.e., we have a random sample

$$(X_1, Y_1), ..., (X_n, Y_n) \overset{\text{IID}}{\sim} \mathrm{Unif}([0, a] \times [0, b]).$$

Find the MLE for the parameter (a, b).

Question 4: Assume the parametric family of distributions is given through its PDF

$$f(x|\theta) = \begin{cases} \dfrac{x}{\theta}, & x \in [0, \theta] \\ \dfrac{2-x}{2-\theta}, & x \in [\theta, 2] \end{cases} \qquad \theta \in \Theta = (0, 2).$$

Find the MLE/MoME for θ.

Parametric inference and interval estimation

In our last chapter we have constructed and considered point estimators for unknown parameters. If we obtain an estimate for our parameter, then, unfortunately, that is not giving an idea how good is our estimate. Say, I have obtained an estimate $\hat{p} = 0.35$ for the probability p in the Bernoulli model. Well, I do not know the exact value of the parameter p, so I cannot assess how good is my estimate, how close is my estimate to the real value. Of course, if I am using an estimator for p with some nice properties, then I can rely on the estimate I have obtained, but I want to have some quantitative measure how goods is my estimate.

For example, if I will have that the interval $(0.33, 0.36)$ contains my unknown parameter value p, then this is good. And even I can state that $p = 0.345 \pm r$ with $0 \leqslant r \leqslant 0.015$. This is giving the idea about the true value of the parameter. Unfortunately, in Statistics, Probability and real life, usually we are unable to give a fixed interval of small length which will contain the unknown parameter **for sure**. And instead we are giving Probabilistic intervals - random intervals that will cover (contain) the unknown parameter with a high probability.

Another point is the following: assume we are estimating an unknown parameter θ based on a random sample from a distribution from some parametric family of continuous distributions. Since our distributions are continuous, then the estimator $\hat{\theta}$ will be (well, in most of the cases) a continuous r.v., so

$$\mathbb{P}(\hat{\theta} = \theta) = 0,$$

and this means that we will (almost surely) never get the exact value of the parameter θ.

Well, this is a little bit upsetting, but.... ☺ There is way out of this situation. To consider Confidence Intervals.

Before going into the subject, let us give the following definition:

Definition 11.1. *Assume X and Y are two r.v. on the same probability space and $X \leqslant Y$ a.s., that is, $\mathbb{P}(X > Y) = 0$. We will call intervals of the form $[X, Y]$ or (X, Y) or $(X, Y]$ or $[X, Y)$ or $(X, +\infty)$,... , **random intervals**.*

In other words, random interval is an interval for which at least one endpoint is a r.v..

Example, Random Intervals: Assume $X \sim \mathtt{Unif}[0, 3]$. Then the interval

$$[X, X + 3)$$

is a random interval. Also, $(X, 4)$ will be a random interval. And we will have another random interval by taking $[-Y, Y]$, where $Y \sim \mathtt{Bernoulli}(0.4)$.

11.1 Confidence Intervals

Assume again we have a parametric model $\{\mathbb{P}_\theta : \theta \in \Theta\}$, and assume that θ is 1D, i.e., $\Theta \subset \mathbb{R}$.

Assume we have a random sample from one of the distributions $\mathbb{P}_\theta$ from that family,

$$X_1, X_2, ..., X_n \overset{IID}{\sim} \mathbb{P}_\theta,$$

or, in the applications, a realization from that random sample, $x_1, ..., x_n$, and again our aim is to estimate the unknown parameter θ.

In this case we will not consider a single guess for the value of θ, rather we will give an interval containing (covering) θ with some confidence level, with some probability. Since we will construct our interval based on the random sample, that interval will be a **random interval**, i.e., its endpoints will be r.v.'s. Next we give the definition:

Definition 11.2. *Assume* $0 < \alpha < 1$, *and let* $L = L(x_1, ..., x_n, \alpha)$, $U = U(x_1, ..., x_n, \alpha)$ *be two functions with* $L(x_1, ..., x_n, \alpha) \leqslant U(x_1, ..., x_n, \alpha)$ *for all* $(x_1, ..., x_n, \alpha)$. *The random interval*

$$(L, U) = \Big(L(X_1, ..., X_n, \alpha)\ ,\ \ U(X_1, ..., X_n, \alpha)\Big)$$

is called a ***confidence interval*** *(or confidence interval estimator) for* θ *of level* $1 - \alpha$, *if for any* $\theta \in \Theta$,

$$\mathbb{P}(L < \theta < U) \geqslant 1 - \alpha.$$

In the case we have a realization of $X_1, ..., X_n$, *say,* $x_1, ..., x_n$, *then the interval*

$$\Big(L(x_1, ..., x_n, \alpha),\ U(x_1, ..., x_n, \alpha)\Big)$$

will be an ***interval estimate*** *for* θ *for the confidence level* $(1 - \alpha)$.

Equivalently, the random interval (L, U) will be an $(1 - \alpha)$ confidence level interval for θ, if

$$\mathbb{P}((L, U) \ni \theta) \geqslant 1 - \alpha \qquad \text{or, equivalently,} \qquad \mathbb{P}((L, U) \not\ni \theta) \leqslant \alpha.$$

Above, L and U are called lower and upper confidence limits, respectively, for θ.

Remark, CI 1: In the case of interval estimator, θ is not random, but the interval (L, U) is random, so we read

$$\mathbb{P}(L < \theta < U)$$

as not the probability that θ is in (L, U), but the probability that (L, U) will include, will contain, will cover the unknown parameter θ. This is not the probability that the values of θ will lie in the fixed interval (there are no different values for θ, θ has just one value, but, unfortunately, we do not know that value), rather it is the probability that the intervals (L, U) will contain the fixed (but unknown) parameter θ.

The complete way of writing of the above is:

$$\mathbb{P}(\omega \in \Omega : L(\omega) < \theta < U(\omega)) \geqslant 1 - \alpha.$$

Remark, CI 2: In fact, if some random interval (L, U) will be a confidence interval for θ of level $1 - \alpha$, then any random interval containing (L, U), say, $(L - 1, U + 4)$ will be a confidence interval for θ of level $1 - \alpha$, since

$$\mathbb{P}((L - 1, U + 4) \ni \theta) \geqslant \mathbb{P}((L, U) \ni \theta) \geqslant 1 - \alpha.$$

We need to find, in some sense, as narrow interval as possible. This is because we want to estimate θ as exact as possible.

REMARK, CI 3: As we have seen above, for a given confidence level $1-\alpha$, we can find infinitely many random intervals with the property

$$\mathbb{P}(L < \theta < U) \geqslant 1-\alpha.$$

Usually (but not always!) one takes an interval of the symmetric form $(L,U) = (R-K, R+K)$ around some r.v. R, where $K \geqslant 0$ can be both deterministic or a r.v.. In that case one uses also the notation

$$(R \pm K) = (R-K, R+K) \qquad \text{and} \qquad [R \pm K] = [R-K, R+K],$$

or even without parentheses, $R \pm K$, if it is clear (or not so important) whether we are talking about open or closed interval.

REMARK, CI 4: In fact, we are using $\geqslant 1-\alpha$ in the definition, since it is possible that no L, U exists with (say, for discrete case)

$$\mathbb{P}(L < \theta < U) = 1-\alpha.$$

In other cases it can be the case that such (L,U) exists, but it will be very difficult (and even impossible) to find such intervals. That's why we are using $\geqslant 1-\alpha$ usually. In the case when exactly $\mathbb{P}(L < \theta < U) = 1-\alpha$, we call the random interval (L,U) to be the exact $1-\alpha$ level confidence interval for θ.

REMARK, CI 5: Sometimes one defines the closed confidence interval $[L,U]$ for θ of level α in the same fashion:

$$\mathbb{P}(L \leqslant \theta \leqslant U) \geqslant 1-\alpha.$$

Of course, in the continuous case, you will have

$$\mathbb{P}(L \leqslant \theta \leqslant U) = \mathbb{P}(L < \theta < U),$$

so both (L,U) and $[L,U]$ will be confidence intervals for θ.

REMARK, ON THE INTERPRETATION OF CI: If (L,U) is a confidence interval for θ of level $1-\alpha$, then we say that we are $1-\alpha$ or $100\cdot(1-\alpha)\%$ confident that the interval (L,U) contains the true value of θ. This doesn't mean that if we will do different observations, for $100\cdot(1-\alpha)\%$ of cases of θ we will have $\theta \in (L,U)$ - we do not have different values of θ, θ is just a fixed parameter (although unknown). Rather, we think like this: if we will do different observations, and construct the intervals (L,U) for each observation, in the $100\cdot(1-\alpha)\%$ cases that interval will contain the true parameter θ.

REMARK, ON THE INTERPRETATION OF CI: Another remark closely related to the previous one. One constructs CI *before observing the sample*. After observing the sample, we will have some values of L and U, so we will have some fixed, non-random interval, say, $(45,56)$. And it is not correct to state that the probability that $\theta \in (45,56)$ is $1-\alpha$, i.e., it is not correct that $\mathbb{P}(45 < \theta < 56) = 1-\alpha$. This is because θ is a constant. Either it will be in that interval, or will not be!

REMARK, INTERPRETATION OF THE OBSERVED CI: Assume you are reading a statistical analysis report and see the following: the interval $(0.1, 0.3)$ is a 95% CI for the parameter. How to interpret this? We cannot interpret as in 95 percent of cases, our parameter is inside the interval.

The interpretation can be the following: the method of construction of CI's that led to the result $(0.1, 0.3)$ was producing intervals, and that intervals are containing our parameter in 95% of cases (i.e., approx. 95 of 100 intervals produced will contain the unknown parameter). And we have one of the produced intervals, $(0.1, 0.3)$. So there is a high chance that that interval is among that 95%, i.e., it is containing the true parameter.

Fig. 11.1 show the CI illustration. Using the formula for CI (we will learn some formulas later) we generate 100 random intervals. And also we draw the line $y = \theta$ in red, where θ is the real value of the parameter (in real-life examples, we do not have this value). You can notice that most of the intervals contain the real value of the parameter (they intersect the red line), and some intervals are not. In this figure we plotted 95% CI (the code will be provided later), so we need to have that more than approximately 95% of intervals contain the unknown parameter.

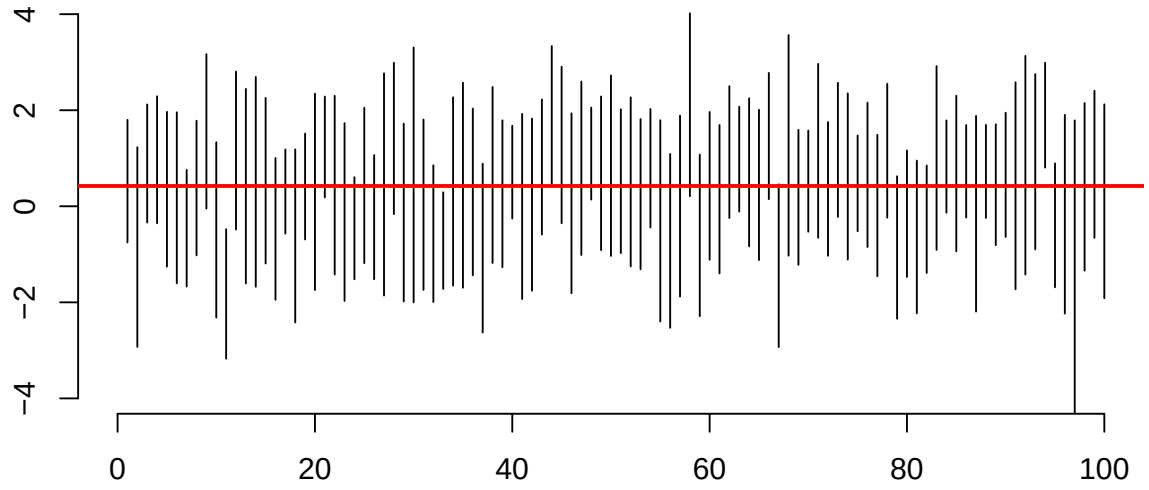

Fig. 11.1: Confidence Intervals Illustration: OX axis shows the number of the generated interval, the red line is the line passing through the real value of the parameter

EXAMPLE, INVERSE PROBLEM, CALCULATION OF THE CONFIDENCE LEVEL: Here we will solve an inverse problem for CI-s. Usually, people are interested in constructing a CI for a given confidence level. Here we will find the confidence level, given the interval.

Assume we have a random sample

$$X_1, X_2, ..., X_{25} \overset{IID}{\sim} \mathcal{N}(\mu, 4)$$

and we want to construct a random interval containing the unknown parameter μ. We know that

$$\overline{X} = \frac{X_1 + ... + X_{25}}{25}$$

is an estimate for μ, so the values of $\overline{X}$ are close to μ with some (supposedly high) probability. So we can guess that the interval $(\overline{X}-1, \overline{X}+1)$ contains the unknown parameter μ with some (supposedly high) probability. So in this case we consider

$$L(X_1, ..., X_{25}) = \overline{X} - 1, \qquad U(X_1, ..., X_{25}) = \overline{X} + 1,$$

and the interval

$$(L, U) = (\overline{X} - 1, \overline{X} + 1)$$

is our interval estimator for μ. Let's find the confidence level. We calculate:

$$\mathbb{P}(L < \mu < U) = \mathbb{P}(\overline{X} - 1 < \mu < \overline{X} + 1) = \mathbb{P}(-1 < \overline{X} - \mu < 1).$$

Now, since $X_k \sim \mathcal{N}(\mu, 4)$ are IID, then

$$X_1 + ... + X_{25} \sim \mathcal{N}(25\mu, 4 \cdot 25),$$

and

$$\overline{X} = \frac{1}{25} \cdot (X_1 + ... + X_{25}) \sim \mathcal{N}(\mu, \frac{4}{25}),$$

hence,

$$\overline{X} - \mu \sim \mathcal{N}(0, \frac{4}{25}),$$

so

$$\frac{\overline{X} - \mu}{\frac{2}{5}} = \frac{\overline{X} - \mu}{\sqrt{\frac{4}{25}}} \sim \mathcal{N}(0, 1).$$

Then,

$$\mathbb{P}(L < \mu < U) = \mathbb{P}(\overline{X} - 1 < \mu < \overline{X} + 1) = \mathbb{P}(-1 < \overline{X} - \mu < 1) = \mathbb{P}(-\frac{5}{2} < \frac{\overline{X} - \mu}{\frac{2}{5}} < \frac{5}{2}) =$$

$$= \Phi(\frac{5}{2}) - \Phi(-\frac{5}{2}) = 2 \cdot \Phi(\frac{5}{2}) - 1 \approx 0.988.$$

So the confidence level is 98.8%. That is, at the 98.8% confidence level (or, in 98.8% cases), the interval $(\overline{X} - 1, \overline{X} + 1)$ will contain the unknown parameter μ.

11.2 Methods to construct Confidence Intervals

In this section we will consider some known methods to construct CIs. In particular, we will construct CIs using Probabilistic inequalities and the Pivots method.

11.2.1 Confidence Intervals based on the Chebyshev Inequality

Let us recall the Chebyshev Inequality:

Proposition 11.1 (Chebyshev Inequality). *For any r.v.* X *with finite* $\mathbb{E}(X)$ *and* $Var(X)$*, and for any* $\varepsilon > 0$*,*

$$\mathbb{P}(|X - \mathbb{E}(X)| \geqslant \varepsilon) \leqslant \frac{Var(X)}{\varepsilon^2}.$$

11.2.2 Confidence Interval for the Bernoulli Model by Chebyshev Inequality

We consider the parametric family of Bernoulli distributions $\{Bernoulli(p) : p \in [0,1]\}$, and assume

$$X_1, ..., X_n \overset{IID}{\sim} Bernoulli(p)$$

for some unknown parameter p, and our aim is to estimate p. Here we want to construct a Confidence Interval for p of level $1-\alpha$, for given $\alpha \in (0,1)$.

Let $\overline{X}_n = \dfrac{X_1 + ... + X_n}{n}$. We know that $\overline{X}_n$ is close to p, so we want to construct our CI around $\overline{X}_n$. We know that

$$\mathbb{E}(\overline{X}_n) = \mathbb{E}(X_1) = p \qquad \text{and} \qquad Var(\overline{X}_n) = \frac{p(1-p)}{n}.$$

Then, by Chebyshev's inequality, for any $\varepsilon > 0$,

$$\mathbb{P}(|\overline{X}_n - \mathbb{E}(\overline{X}_n)| \geqslant \varepsilon) = \mathbb{P}(|\overline{X}_n - p| \geqslant \varepsilon) \leqslant \frac{1}{\varepsilon^2} \cdot Var(\overline{X}_n) = \frac{1}{\varepsilon^2} \cdot \frac{p(1-p)}{n}.$$

It is very simple to prove that[1] $p(1-p) \leqslant \frac{1}{4}$ for any p, so we will have

$$\mathbb{P}(|\overline{X}_n - p| \geqslant \varepsilon) = \frac{1}{\varepsilon^2} \cdot \frac{p(1-p)}{n} \leqslant \frac{1}{\varepsilon^2} \cdot \frac{1}{4n}$$

Here ε is any positive number, so we can choose it as we wish. We choose $\varepsilon > 0$ in such a way to have

$$\frac{1}{\varepsilon^2} \cdot \frac{1}{4n} = \alpha,$$

i.e.,

$$\varepsilon = \frac{1}{2\sqrt{n\alpha}}.$$

In that case,

$$\mathbb{P}(|\overline{X}_n - p| \geqslant \frac{1}{2\sqrt{n\alpha}}) \leqslant \alpha.$$

Now, by passing to the complements[2],

$$\mathbb{P}(|\overline{X}_n - p| < \frac{1}{2\sqrt{n\alpha}}) \geqslant 1 - \alpha,$$

and hence,

$$\mathbb{P}(|\overline{X}_n - p| < \frac{1}{2\sqrt{n\alpha}}) = \mathbb{P}(|\overline{X}_n - p| < \frac{1}{2\sqrt{n\alpha}}) \geqslant 1 - \alpha,$$

which is the same as

$$\mathbb{P}\left(\overline{X}_n - \frac{1}{2\sqrt{n\alpha}} < p < \overline{X}_n + \frac{1}{2\sqrt{n\alpha}}\right) \geqslant 1 - \alpha.$$

[1] The graph of the function $y = p(1-p)$ is an inverted parabola, with the maximum point at $p = 1/2$.
[2] I.e., here we use $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$, where A is some event.

This means that the interval

$$\left(\overline{X}_n - \frac{1}{2\sqrt{n\alpha}}, \overline{X}_n + \frac{1}{2\sqrt{n\alpha}}\right)$$

is a $1-\alpha$ level confidence interval for p. We can write this also in a more compact way:

$$\overline{X}_n \pm \frac{1}{2\sqrt{n\alpha}}.$$

Here, the quantity $\frac{1}{2\sqrt{n\alpha}}$ is called the **Error Margin**.

REMARK, CI FOR BERNOULLI: If we will calculate the length of the above CI, we'll obtain

$$\frac{1}{\sqrt{n\alpha}},$$

and this tends to 0, as $n \to +\infty$, with the rate of $O(\frac{1}{\sqrt{n}})$.

This, in particular, means that if we will increase the Sample Size, then we will get more accurate estimate for p.

REMARK, CI FOR BERNOULLI MODEL: Usually, in many models, the CI length will increase, if we will make α smaller. That is, if we want to have more confidence that the interval contains the unknown value of the parameter, we need to have a larger interval. And, particularly, in our case, the interval length

$$\frac{1}{\sqrt{n\alpha}}$$

will increase, if α will decrease to 0. And, if we want to be 100% sure, we can take the limit $\alpha \to +0$, and obtain that the interval length $\frac{1}{\sqrt{n\alpha}} \to +\infty$, so the method will give the answer: if you want to be 100% sure that the interval contains the parameter value, then the interval is $(-\infty, +\infty)$.

Later, using other methods, we will consider approximate CI for the same Bernoulli model p. That method will provide narrower Confidence Interval, although the result will be true for large n-s.

EXAMPLE, CI FOR BERNOULLI:

Assume we want to estimate the proportion of persons in AUA who prefer tea to coffee. We ask 50 persons, and it turns out that 8 of them prefer tea to coffee. So, based on our observation, we want to estimate the proportion p of persons in AUA who prefer tea to coffee. In this case we want to obtain a CI for p of some confidence level, say, 95% CI.

Here we deal with the Bernoulli model, of course. For a person, we will write $X = 1$, if he/she prefers tea-to-coffee, and $X = 0$ otherwise. So we will have $X \sim \text{Bernoulli}(p)$. Now, we consider a random sample

$$X_1, X_2, ..., X_{50} \overset{\text{IID}}{\sim} \text{Bernoulli}(p),$$

where X_k is the k-th persons preference (X_k is a r.v., this is before the polling, before asking persons).

We want to be 95% confident, so our confidence level is $1-\alpha=0.95$, so $\alpha=0.05$. The CI obtained above for p, using the Chebyshev Inequality, is

$$\left(\overline{X}_n-\frac{1}{2\sqrt{n\alpha}},\overline{X}_n+\frac{1}{2\sqrt{n\alpha}}\right).$$

So after plugging the observed values we will obtain the observed CI

$$\left(\overline{x}_n-\frac{1}{2\sqrt{n\alpha}},\overline{x}_n+\frac{1}{2\sqrt{n\alpha}}\right).$$

Here we need to take $n=50$, $\alpha=0.05$ and $\overline{x}_n=\overline{x}_{50}=\frac{8}{50}=\frac{4}{25}$, since 8 out of 50 responses were 1-s (eight of x_k-s are equal to 1, and others are 0-s). Then the obtained 95% CI will be

$$\left(\overline{x}_n-\frac{1}{2\sqrt{n\alpha}},\overline{x}_n+\frac{1}{2\sqrt{n\alpha}}\right)=\left(\frac{4}{25}-\frac{1}{2\sqrt{50\cdot 0.05}},\frac{4}{25}+\frac{1}{2\sqrt{50\cdot 0.05}}\right)\approx(-0.1562,0.4762).$$

We can write the obtained CI as

$$\frac{4}{25}\pm\frac{1}{2\sqrt{50\cdot 0.05}}=\frac{4}{25}\pm\frac{1}{\sqrt{10}}.$$

In fact, since p cannot be negative, we will obtain the following CI: $[0,0.4762)$. Interpretation is: with 95% confidence, the proportion of persons in AUA preferring tea-to-coffee is less than 0.4762. More that a half are coffee drinkers. Fuuu!

As we have state above, we will construct CI for the Bernoulli Model in other way, which will give us smaller interval, so we will be able to estimate p more accurately.

On of the important usages for CIs is the Sample Size determination: using CIs, we can find the Sample Size to achieve the desired accuracy within some confidence level.

Example, CI, calculation of the Sample Size: Consider the Bernoulli Model above, and now let us solve the following problem: How many observations we need to have in order to be sure that

$$\mathbb{P}(\overline{X}_n-0.1<p<\overline{X}_n+0.1)\geqslant 0.95,$$

or, which is the same, to be sure with 95% confidence that

$$p\in(\overline{X}_n-0.1,\overline{X}_n+0.1).$$

The other, more non-mathematical way of stating the problem is - how many observations we need to have to say with 95% confidence that p is equal to $\overline{X}_n$ within the maximum error ± 0.1?

The confidence level 0.95 corresponds to $\alpha=0.5$. Now, if we will have

$$(\overline{X}_n-0.1,\overline{X}_n+0.1)\supset\left(\overline{X}_n-\frac{1}{2\sqrt{n\alpha}},\overline{X}_n+\frac{1}{2\sqrt{n\alpha}}\right),$$

then we will deduce[3], using the above CI construction,

$$\mathbb{P}(\overline{X}_n-0.1<p<\overline{X}_n+0.1)\geqslant\mathbb{P}(\overline{X}_n-\frac{1}{2\sqrt{n\alpha}}<p<\overline{X}_n+\frac{1}{2\sqrt{n\alpha}})\geqslant 1-\alpha=0.95.$$

So it is enough to have

$$(\overline{X}_n-0.1,\overline{X}_n+0.1)\supset\left(\overline{X}_n-\frac{1}{2\sqrt{n\alpha}},\overline{X}_n+\frac{1}{2\sqrt{n\alpha}}\right),$$

that is, enough to have

$$\frac{1}{2\sqrt{n\alpha}} < 0.1,$$

which is the same as

$$n > \frac{25}{\alpha} = \frac{25}{0.05} = 500. \qquad \blacksquare$$

11.2.3 Confidence Interval for the Exponential Model by the Chebyshev Inequality

Consider the following Exponential model: we consider a family of distributions $\{\mathcal{E}(\theta) : \theta > 0\}$, where $\mathcal{E}(\theta)$ is given by the PDF

$$f(x;\theta) = \begin{cases} \frac{1}{\theta} \cdot e^{-\frac{x}{\theta}}, & \text{if } x \geqslant 0 \\ 0, & \text{otherwise} \end{cases}$$

This differs from our previous definition of the Exponential Distribution by having $\lambda = \frac{1}{\theta}$, and this simplifies calculations[4]

It is easy to calculate that for $X \sim \mathcal{E}(\theta)$,

$$\mathbb{E}(X) = \theta \qquad \text{and} \qquad Var(X) = \theta^2.$$

Assume we have a random sample from one of the distributions $\mathcal{E}(\theta)$,

$$X_1, X_2, ..., X_n \overset{iid}{\sim} \mathcal{E}(\theta).$$

We want to construct a Confidence Interval of level $1-\alpha$ for the parameter θ. We know that the empiric mean

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

is an estimator for $\mathbb{E}(X_k) = \theta$, so θ is in some interval around $\overline{X}_n$. Let us find an interval of that type with some probability estimate. He again we use the Chebyshev inequality.

By that inequality, for any $\varepsilon > 0$,

$$\mathbb{P}(|\overline{X}_n - \mathbb{E}(\overline{X}_n)| \geqslant \varepsilon) \leqslant \frac{Var(\overline{X}_n)}{\varepsilon^2}.$$

Now, clearly, $\mathbb{E}(\overline{X}_n) = \theta$ and $Var(\overline{X}_n) = \frac{\theta^2}{n}$, so we will have

$$\mathbb{P}(|\overline{X}_n - \theta| \geqslant \varepsilon) \leqslant \frac{\theta^2}{n\varepsilon^2}.$$

[4]Well, we could consider the previous model, but we need then to estimate the parameter $\frac{1}{\lambda}$ instead of estimating λ. Otherwise, if we will try to estimate λ, we will try to use $\frac{1}{\overline{X}_n}$ as an estimator, and it is not an easy task to calculate the expectation and variance of this estimator.

Of course, we can use our initial the model $\{Exp(\lambda) : \lambda > 0\}$ and estimate first $\frac{1}{\lambda}$ to get an CI of the form (a, b) with

$$\mathbb{P}(a < \frac{1}{\lambda} < b) \geqslant 1 - \alpha.$$

Then, if our $a > 0$, we can get

$$\mathbb{P}(\frac{1}{b} < \lambda < \frac{1}{a}) \geqslant 1 - \alpha,$$

meaning that $\left(\frac{1}{b}, \frac{1}{a}\right)$ is a $1-\alpha$ level CI for λ. The length of this interval is $\frac{b-a}{ab}$.

If $\alpha > 0$, we choose $\varepsilon > 0$ in such a way that

$$\frac{\theta^2}{n\varepsilon^2} = \alpha,$$

i.e.,

$$\varepsilon = \frac{\theta}{\sqrt{n\alpha}}.$$

Plugging this value in the Chabyshev inequality will yield:

$$\mathbb{P}(|\overline{X}_n - \theta| \geqslant \frac{\theta}{\sqrt{n\alpha}}) \leqslant \alpha,$$

hence, passing to the complements,

$$\mathbb{P}(|\overline{X}_n - \theta| < \frac{\theta}{\sqrt{n\alpha}}) \geqslant 1 - \alpha.$$

This can be written

$$\mathbb{P}(-\frac{\theta}{\sqrt{n\alpha}} < \overline{X}_n - \theta < \frac{\theta}{\sqrt{n\alpha}}) \geqslant 1 - \alpha,$$

and solving in terms of θ will give, for the case $1 - \frac{1}{\sqrt{n\alpha}} > 0$,

$$\mathbb{P}(\frac{\overline{X}_n}{1 + \frac{1}{\sqrt{n\alpha}}} < \theta < \frac{\overline{X}_n}{1 - \frac{1}{\sqrt{n\alpha}}}) \geqslant 1 - \alpha.$$

This means that the interval

$$\left(\frac{\overline{X}_n}{1 + \frac{1}{\sqrt{n\alpha}}}, \frac{\overline{X}_n}{1 - \frac{1}{\sqrt{n\alpha}}}\right)$$

is a $1 - \alpha$ level confidence interval for θ.

One can also calculate the length of this interval and see that it goes to 0 in the rate of $O(\frac{1}{\sqrt{n}})$. ■

11.2.4 Confidence Interval for the mean of Normal Distribution, σ is known: Chebyshev Inequality Model

We consider the parametric model $\{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R}\}$, assuming that σ is known and fixed.

Assume we have a random sample

$$X_1, X_2, ..., X_n \overset{\text{IID}}{\sim} \mathcal{N}(\mu, \sigma^2)$$

for some unknown $\mu \in \mathbb{R}$, and we want to give an interval estimator for μ.

We know that

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

is a consistent (and unbiased) estimator for μ. So we want to find an interval around $\overline{X}_n$ containing the unknown parameter μ with the desired confidence level $1 - \alpha$.

Here we will use the Chebyshev inequality, and later we will give another method to obtain a CI for this model.

By Chebyshev inequality, for any positive ε,

$$\mathbb{P}(|\overline{X}_n - \mathbb{E}(\overline{X}_n)| \geqslant \varepsilon) \leqslant \frac{\mathrm{Var}(\overline{X}_n)}{\varepsilon^2}.$$

Now, since X_k are IID r.v. from $\mathcal{N}(\mu, \sigma^2)$, then

$$\mathbb{E}(\overline{X}_n) = \mu \qquad \text{and} \qquad Var(\overline{X}_n) = \frac{\sigma^2}{n},$$

so we will have

$$\mathbb{P}(|\overline{X}_n - \mu| \geqslant \varepsilon) \leqslant \frac{\sigma^2}{n\varepsilon^2}.$$

Now we take $\alpha \in (0, 1)$ and solve

$$\frac{\sigma^2}{n\varepsilon^2} = \alpha$$

to find ε:

$$\varepsilon = \frac{\sigma}{\sqrt{n\alpha}}.$$

Plugging this value into the inequality above, we will obtain

$$\mathbb{P}(|\overline{X}_n - \mu| \geqslant \frac{\sigma}{\sqrt{n\alpha}}) \leqslant \alpha,$$

and passing to the complements, we will get

$$\mathbb{P}(|\overline{X}_n - \mu| < \frac{\sigma}{\sqrt{n\alpha}}) \geqslant 1 - \alpha.$$

This means that

$$\mathbb{P}(-\frac{\sigma}{\sqrt{n\alpha}} < \overline{X}_n - \mu < \frac{\sigma}{\sqrt{n\alpha}}) \geqslant 1 - \alpha,$$

i.e.,

$$\mathbb{P}(\overline{X}_n - \frac{\sigma}{\sqrt{n\alpha}} < \mu < \overline{X}_n + \frac{\sigma}{\sqrt{n\alpha}}) \geqslant 1 - \alpha,$$

so the interval

$$\left(\overline{X}_n - \frac{\sigma}{\sqrt{n\alpha}}, \overline{X}_n + \frac{\sigma}{\sqrt{n\alpha}}\right)$$

is a $1 - \alpha$-level CI for μ.

REMARK, CI FOR THE NORMAL MODEL MEAN: Later we will consider another approach to construct a CI for μ in this model. In fact, the other approach is more common than this one. This is because, on one hand, it gives smaller Confidence interval for μ, and, on the other hand, that approach can be generalized for the case when σ^2 is unknown (and this case is usually happening in the real-life problems).

Btw, this approach can partially help us also in the case when σ^2 will be unknown. I mean, in the case when we do not know the exact value of σ^2, but we know some bound on it, say, $\sigma^2 \leqslant 10$, then we can use the above approach to construct a CI: we just need to use

$$\mathbb{P}(|\overline{X}_n - \mu| \geqslant \varepsilon) \leqslant \frac{Var(\overline{X}_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \leqslant \frac{10}{n\varepsilon^2}.$$

The rest is the same.

Another remark is that here again, the length of the CI tends to zero as $n \to +\infty$, and with the rate $O(\frac{1}{\sqrt{n}})$.

11.3 Confidence Intervals based on Pivotal Method

The general idea of the Pivotal Method is the following. Assume we have a random sample

$$X_1, X_2, ..., X_n \overset{IID}{\sim} \mathbb{P}_\theta,$$

and we want to construct an $(1-\alpha)$ Confidence Level CI for the unknown parameter θ.

Suppose we can construct a function of $X_1, X_2, ..., X_n$ and our unknown parameter θ,

$$g(X_1, X_2, ..., X_n, \theta)$$

such that **the distribution (or at least, the asymptotic distribution) of** $g(X_1, ..., X_n, \theta)$ **is independent of** θ. In this case we call $g(X_1, ..., X_n, \theta)$ to be a **Pivot** for our model[5].

Usually, we can have infinitely many Pivots for the same Model. But, again usually, one take some standard Pivots for the standard cases (we will consider that standard cases soon).

EXAMPLE, PIVOT IDEA: Well, you can ask: if $g(X_1, ..., X_n, \theta)$ depends on θ, how can its distribution be independent of θ? Nice question, of course, but you have seen and used this kind of things several times up to this point. Say, if

$$X \sim \mathcal{N}(\mu, 10),$$

where μ is our unknown parameter, then

$$g(X, \mu) = X - \mu$$

depends on μ, but

$$g(X, \mu) = X - \mu \sim \mathcal{N}(0, 10),$$

so the distribution of g is independent of μ! (This is not a factorial sign, rather an excitement one, like "Eureka" ☺).

For the same model, another pivot can be, say,

$$g_1(X, \mu) = \frac{X - \mu}{\sqrt{10}} \sim \mathcal{N}(0, 1),$$

or, say,

$$g_2(X, \mu) = 0.3 * (X - \mu) + 12,$$

but I will ask you to find the distribution of the latter.

Also, the following function:

$$h(X, \mu) = \frac{X}{\mu}$$

is not a Pivot, since its distribution is

$$h(X, \mu) = \frac{X}{\mu} \sim \mathcal{N}\left(1, \frac{10}{\mu^2}\right),$$

which depends on our unknown parameter μ.

[5] In fact, usually wise people add here the requirement that g needs to be strictly monotonic in θ, or invertible in terms of θ. This is to ensure that we can solve the inequality

$$a < g(X_1, ..., X_n, \theta) < b$$

for a given a and b.

Now, assume we have a pivot for our model, $g(X_1, X_2, ..., X_n, \theta)$. Then we proceed to find two numbers a and b such that[6]

$$\mathbb{P}(a < g(X_1, X_2, ..., X_n, \theta) < b) = 1 - \alpha.$$

Important thing is that a and b will be independent (they will depend only on the distribution of $g(X_1, X_2, ..., X_n, \theta)$, will be some quantiles of the distribution). Having these a and b, we are almost done. Next we solve

$$a < g(X_1, X_2, ..., X_n, \theta) < b$$

in terms of θ (if possible) to have something like

$$a < g(X_1, X_2, ..., X_n, \theta) < b \qquad \Leftrightarrow \qquad L < \theta < U.$$

(L and U will depend, in general, on $a, b, X_1, ..., X_n$ and α). Then,

$$1 - \alpha = \mathbb{P}(a < g(X_1, X_2, ..., X_n, \theta) < b) = \mathbb{P}(L < \theta < U),$$

so (L, U) is our CI for θ of level $1 - \alpha$.

Now, let us consider concrete realizations of this idea.

11.3.1 Confidence Interval for the mean of Normal Distribution, σ is known

Again we consider the parametric model $\{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R}\}$, assuming that σ is known and fixed.

Assume we have a random sample

$$X_1, X_2, ..., X_n \overset{IID}{\sim} \mathcal{N}(\mu, \sigma^2)$$

for some unknown $\mu \in \mathbb{R}$, and we want to give an interval estimator for μ, but using this time this new, pivoting, method.

We know that

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

is a consistent (and unbiased) estimator for μ, so $\overline{X}_n$ is close to μ and concentrated around μ for large n. So we want to find an interval around $\overline{X}_n$ containing the unknown parameter μ with the desired confidence level $1 - \alpha$. To that end, we do the following trick: we want to obtain from $\overline{X}_n$ and μ a new r.v. Z, the distribution of which is independent of[7] μ: Since $X_k \sim \mathcal{N}(\mu, \sigma^2)$, then $\overline{X}_n \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$, and $\overline{X}_n - \mu \sim \mathcal{N}(0, \frac{\sigma^2}{n})$, implying

$$Z := \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

So Z is standard normal r.v., and its distribution does not depend on μ. Uraaa! We found a Pivooot!

Now, we choose $a, b \in \mathbb{R}$ such that

$$\mathbb{P}(a < Z < b) = 1 - \alpha.$$

In fact, there are infinitely many such a, b-s: we just need to have that the area under the graph of the PDF of standard normal distribution is $1 - \alpha$. Since that PDF is symmetric about 0, then the

[6] Or, if this is unsolvable, satisfying the following inequality:

$$\mathbb{P}(a < g(X_1, X_2, ..., X_n, \theta) < b) \geqslant 1 - \alpha$$

[7] And also for which the set $a < Z < b$ can be easily transformed to the form $L(\overline{X}_n) < \mu < U(\overline{X}_n)$.

interval $[a, b]$ with the above property and *with minimal length* will be[8] the symmetric interval $[-q, q]$, where the area under the PDF graph over $(-\infty, -q]$ and $[q, +\infty)$ is $\frac{\alpha}{2}$:

Give the Graph here!

If we will recall the definition of the theoretical quantiles, then the point $-q$ will be equal to the quantile of level $\frac{\alpha}{2}$, and q will be equal to the quantile of the level $1-\frac{\alpha}{2}$. Usually, the p-level quantile of the standard normal are denoted by z_p:

$$\Phi(z_p) = p,$$

or, which is the same,

$$z_p = \Phi^{-1}(p).$$

In this case, $-q = z_{\frac{\alpha}{2}}$ and $q = z_{1-\frac{\alpha}{2}}$. So we will have

$$\mathbb{P}(z_{\frac{\alpha}{2}} < Z < z_{1-\frac{\alpha}{2}}) = 1-\alpha.$$

We are close to the solution: using the definition of Z, we'll obtain

$$\mathbb{P}(z_{\frac{\alpha}{2}} < \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} < z_{1-\frac{\alpha}{2}}) = 1-\alpha,$$

so

$$\mathbb{P}(z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} < \overline{X}_n - \mu < z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}) = 1-\alpha,$$

implying

$$\mathbb{P}(\overline{X}_n - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{X}_n - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}) = 1-\alpha,$$

and since $z_{\alpha/2} = -z_{1-\alpha/2}$, the last equality can be written in the form

$$\mathbb{P}(\overline{X}_n - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{X}_n + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}) = 1-\alpha.$$

This means that the interval

$$\left(\overline{X}_n - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \overline{X}_n + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right)$$

is an exact $1-\alpha$ level CI for[9] μ.

Here again, the length of the CI tends to 0 with the rate $O(\frac{1}{\sqrt{n}})$. ■

EXAMPLE, CI FOR THE MEAN OF THE NORMAL DISTRIBUTION: Example here

[8] Check this! And we want to have a Confidence Interval as small as possible for the given level of confidence.

[9] Try to compare this CI with the one obtained by using the Chebyshev Inequality: you need to compare

$$2z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \quad \text{with} \quad 2\frac{\sigma}{\sqrt{n\alpha}},$$

so you need to estimate

$$z_{1-\alpha/2} \quad \text{with} \quad \frac{1}{\sqrt{\alpha}}.$$

REMARK, CHOOSING THE SAMPLE SIZE: Here I want to give two important considerations concerning the CIs. These considerations work also for other Model/Parameter CIs too.

The first consideration is that the length of the CI above gets smaller as we increase the number of observations - that is, if we observe more information, we can estimate the unknown parameter more accurately. In fact, the length of the CI for the parameter μ above is

$$2 \cdot z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}},$$

which tends to 0 as $n \to +\infty$. And, if we want to narrow the interval, say, 2 times, we need to increase our sample size 4 times. Well known fact among statisticians!

The other consideration is the choice of the Sample size: if we want to have a CI for μ with the Maximum Margin Error ε, then we need to take

$$z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leqslant \varepsilon,$$

that is,

$$n \geqslant \left(\frac{\sigma \cdot z_{1-\frac{\alpha}{2}}}{\varepsilon} \right)^2.$$

Well, let me give another consideration too, although I promised to give only two. The fact is that when we decrease α, that is, if we increase the confidence level $1-\alpha$, then the length of the interval,

$$2 \cdot z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}},$$

is increasing, since the quantiles $z_{1-\frac{\alpha}{2}}$ are increasing (think why?) - so if you require more confidence by having the same amount of data (n is unchanged), then you will have a larger interval. And, in fact, if $\alpha \to 0+$, then $z_{1-\frac{\alpha}{2}} \to +\infty$, so if you want to be 100% sure that the interval contains the true value of the parameter, you will get the interval $(-\infty, +\infty)$!

REMARK, CI: We can think like this: we know that $X_1 \sim \mathcal{N}(\mu, \sigma^2)$, so $X_1 - \mu \sim \mathcal{N}(0, \sigma^2)$. So basically, $X_1 - \mu$ can serve as a pivot. Of course, it can. Better to transform it to $\frac{X_1-\mu}{\sigma} \sim \mathcal{N}(0,1)$. So

$$Z = \frac{X_1 - \mu}{\sigma}$$

is a pivot with $Z \sim \mathcal{N}(0,1)$. OK, let's see what this will give. Just like above, we can write

$$\mathbb{P}(-z_{1-\alpha/2} < Z < z_{1-\alpha/2}) = 1-\alpha,$$

so

$$\mathbb{P}\left(-z_{1-\alpha/2} < \frac{X_1 - \mu}{\sigma} < z_{1-\alpha/2} \right) = 1-\alpha,$$

implying

$$\mathbb{P}\left(X_1 - z_{1-\alpha/2} \cdot \sigma < \mu < X_1 - z_{1-\alpha/2} \cdot \sigma \right) = 1-\alpha,$$

hence,

$$X_1 \pm z_{1-\alpha/2} \cdot \sigma$$

is an exact $1-\alpha$ confidence level CI for μ.

Everything is correct, the only thing is that the length of this interval is very large, it is equal to $2 \cdot z_{1-\alpha/2} \cdot \sigma$, and this cannot be made smaller, say, by increasing the number of observations. So the use of the pivot above is justified by the fact that the distribution of $\overline{X}_n$ is getting more concentrated around μ as n is increasing, so our interval is getting narrower as we increase n.

And also, we are not using the whole information we have, we are just using X_1. Not a wise idea.

R CODE, COMPARISON OF CI LENGTHS, CHEBYSHEV INEQUALITY AND Z-SCORE METHOD:

```
#Comparing Margins of Errors for Normal mu, when sigma is known,
#Chebyshev inequality and Z-test cases
a <- seq(0.01, 0.2, by = 0.01)
#Plotting y = 1/a
plot(a, 1/a, type = "l", col = "blue", lwd = 2, ylim = c(0,16))
par(new = TRUE)
#Plotting y = z_{1-a/2}
plot(a, qnorm(1-a/2), type = "l", col = "red", lwd = 2, ylim = c(0,16))
```

11.3.2 Confidence Interval for the mean of Normal Distribution, σ is unknown

Here again we consider the parametric model $\{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R}\}$, but in this case we assume that σ is not known.

Assume we have a random sample

$$X_1, X_2, ..., X_n \overset{IID}{\sim} \mathcal{N}(\mu, \sigma^2)$$

for some unknown $\mu \in \mathbb{R}$, and we want to give an interval estimator for μ.

Again, we know that

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

is a consistent (and unbiased) estimator for μ. So we want to find an interval around $\overline{X}_n$ containing the unknown parameter μ with the desired confidence level $1 - \alpha$.

In this case, since σ^2 is unknown also, we need to use some estimate for that, and we will use

$$S^2 = \frac{\sum_{k=1}^{n} (X_k - \overline{X})^2}{n-1},$$

the unbiased estimator for σ^2.

In the previous case we have used the r.v.

$$Z = \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0, 1)$$

to obtain the CI. Here we use the r.v. close to Z, taking S instead of σ:

$$t = \frac{\overline{X}_n - \mu}{\frac{S}{\sqrt{n}}}.$$

Now, it is well known that r.v. t has the Student's t-distribution with $n-1$ degrees of freedom: see for definition, CDF, PDF formulas and some properties at the appendix of this chapter and at https://en.wikipedia.org/wiki/Student's_t-distribution. That is,

$$t \sim t(n-1).$$

Now, in the analogy of the previous case, we choose $a, b \in \mathbb{R}$ such that

$$\mathbb{P}(a < t < b) = 1-\alpha.$$

Again, there are infinitely many such a, b-s. Since the PDF of t-distribution is again symmetric about 0, then the interval $[a, b]$ with the above property and *with minimal length* will be the symmetric interval $[-q, q]$, where the area under the PDF graph over $(-\infty, -q]$ and $[q, +\infty)$ is $\frac{\alpha}{2}$:

Give the Graph here!

In this case $-q = t_{\alpha/2}(n-1)$ and $q = t_{1-\alpha/2}(n-1)$, where $t_{\alpha/2}(n-1)$ and $t_{1-\alpha/2}(n-1)$ are the $\alpha/2$ and $1-\alpha/2$ level quantiles for the t distribution with $(n-1)$ degrees of freedom, $t(n-1)$. Clearly, $t_{\alpha/2}(n-1) = -t_{1-\alpha/2}(n-1)$. So we will have

$$\mathbb{P}(-t_{1-\alpha/2}(n-1) < t < t_{1-\alpha/2}(n-1)) = 1-\alpha,$$

hence,

$$\mathbb{P}(-t_{1-\alpha/2}(n-1) < \frac{\overline{X}_n - \mu}{\frac{S}{\sqrt{n}}} < t_{1-\alpha/2}(n-1)) = 1-\alpha.$$

Again, using some calculations, we will obtain:

$$\mathbb{P}\left(\overline{X}_n - t_{1-\alpha/2}(n-1)\cdot\frac{S}{\sqrt{n}} < \mu < \overline{X}_n + t_{1-\alpha/2}(n-1)\cdot\frac{S}{\sqrt{n}}\right) = 1-\alpha,$$

and the interval

$$\left(\overline{X}_n - t_{1-\alpha/2}(n-1)\cdot\frac{S}{\sqrt{n}}, \overline{X}_n + t_{1-\alpha/2}(n-1)\cdot\frac{S}{\sqrt{n}}\right)$$

is the exact CI for μ, if the variance is unknown. ■

Remark, CI for the mean of $\mathcal{N}(\mu, \sigma^2)$: Let us summarize what we have obtained for this model. The problem is: given a r.s.

$$X_1, X_2, ..., X_n \overset{IID}{\sim} \mathcal{N}(\mu, \sigma^2),$$

and $\alpha \in (0,1)$, we want to construct an $1-\alpha$-level CI for the unknown parameter μ.

We have considered 2 cases: when σ^2 was known and unknown. Here we give the summary:

	σ^2 is known	σ^2 is unknown
Pivot:	$Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$	$t = \frac{\overline{X}-\mu}{S/\sqrt{n}}$
Distrib. of Pivot:	$Z \sim \mathcal{N}(0,1)$	$t \sim t(n-1)$
Region with prob. $1-\alpha$	$-z_{1-\alpha/2} < Z < z_{1-\alpha/2}$	$-t_{1-\alpha/2}(n-1) < t < t_{1-\alpha/2}(n-1)$
CI for μ:	$\overline{X} \pm z_{1-\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}$	$\overline{X} \pm t_{1-\alpha/2}(n-1)\cdot\frac{S}{\sqrt{n}}$

Here S is the Sample Standard Deviation given by

$$S^2 = \frac{\sum_{k=1}^n (X_k - \overline{X})^2}{n-1}.$$

R CODE, CI FOR NORMAL MEAN, VARIANCE IS UNKNOWN:

```
#CI for the Normal mu (sigma is unknown)
real.mu <-0.423
conf.level <- 0.95
a = 1 - conf.level
sample.size <- 10
no.of.intervals <- 100

plot.new()
plot.window(xlim = c(0,no.of.intervals), ylim = c(-4,4))
axis(1)
axis(2)
for(i in 1:no.of.intervals){
  x <- rnorm(sample.size, mean = real.mu, sd = 2.32)
  s <- sd(x)
  lo <- mean(x) - qt(1-a/2, df = sample.size - 1)*s/sqrt(sample.size)
  up <- mean(x) + qt(1-a/2, df = sample.size - 1)*s/sqrt(sample.size)
  segments(c(i), c(lo), c(i), c(up))
}
abline(h = real.mu, lwd = 2, col = "red")
```

11.3.3 Confidence Interval for the Variance of Normal Distribution, mean is known

Now let us consider the parametric model $\{\mathcal{N}(\mu, \sigma^2) : \sigma^2 \in (0, +\infty)\}$, assuming that the value of μ is known.

Assume also that we have a random sample from one of the distributions of that family:

$$X_1, ..., X_n \overset{IID}{\sim} \mathcal{N}(\mu, \sigma^2),$$

for some unknown σ^2. Our aim is to construct a confidence interval for σ^2.

Since we know that $X_k \sim \mathcal{N}(\mu, \sigma^2)$, then $\dfrac{X_k - \mu}{\sigma} \sim \mathcal{N}(0, 1)$. Now we consider the r.v.

$$Y = \sum_{k=1}^{n} \left(\frac{X_k - \mu}{\sigma}\right)^2,$$

which is a sum of squares of n independent standard normal r.v.s. It is known that the distribution of Y is $\chi^2(n)$, i.e., $Y \sim \chi^2(n)$, so the distribution of Y is independent of the parameter σ^2, and Y can serve as a pivot. Now, we take $\alpha \in (0, 1)$, and take some interval (a, b) such that for $Y \sim \chi^2(n)$

$$\mathbb{P}(a < Y < b) = 1 - \alpha.$$

Unfortunately, the PDF of $\chi^2(n)$ is not symmetric, so it is not so easy to find the minimal-length interval satisfying this property. We just take the quantiles $\chi^2_{\alpha/2}(n)$ and $\chi^2_{1-\alpha/2}(n)$ of $\chi^2(n)$ of levels $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ (so that the tails areas under the PDF are $\frac{\alpha}{2}$, i.e., we choose quantiles in such a way that they will separate symmetric area portions from both sides):

$$\mathbb{P}(\chi^2_{\alpha/2}(n) < Y < \chi^2_{1-\alpha/2}(n)) = 1 - \alpha.$$

Now we plug the value of Y and solve inequalities to express σ^2. We'll obtain

$$\mathbb{P}\left(\frac{\sum_{k=1}^{n}(X_k-\mu)^2}{\chi^2_{1-\alpha/2}(n)} < \sigma^2 < \frac{\sum_{k=1}^{n}(X_k-\mu)^2}{\chi^2_{\alpha/2}(n)}\right) = 1-\alpha,$$

so we have found an exact CI of level $1-\alpha$ for σ^2:

$$\left(\frac{\sum_{k=1}^{n}(X_k-\mu)^2}{\chi^2_{1-\alpha/2}(n)}, \frac{\sum_{k=1}^{n}(X_k-\mu)^2}{\chi^2_{\alpha/2}(n)}\right).$$

Try to find the length of this interval and show that it tends to zero! ■

REMARK, CI FOR THE STANDARD DEVIATION: Above we have constructed an $(1-\alpha)$-level CI for the Variance σ^2 of the Normal Distribution model. And we know that

$$\mathbb{P}\left(\frac{\sum_{k=1}^{n}(X_k-\mu)^2}{\chi^2_{1-\alpha/2}(n)} < \sigma^2 < \frac{\sum_{k=1}^{n}(X_k-\mu)^2}{\chi^2_{\alpha/2}(n)}\right) = 1-\alpha.$$

Now, obviously (as $\sigma \geqslant 0$), we will have

$$\mathbb{P}\left(\sqrt{\frac{\sum_{k=1}^{n}(X_k-\mu)^2}{\chi^2_{1-\alpha/2}(n)}} < \sigma < \sqrt{\frac{\sum_{k=1}^{n}(X_k-\mu)^2}{\chi^2_{\alpha/2}(n)}}\right) = \mathbb{P}\left(\frac{\sum_{k=1}^{n}(X_k-\mu)^2}{\chi^2_{1-\alpha/2}(n)} < \sigma^2 < \frac{\sum_{k=1}^{n}(X_k-\mu)^2}{\chi^2_{\alpha/2}(n)}\right) = 1-\alpha,$$

and hence, the random interval

$$\left(\sqrt{\frac{\sum_{k=1}^{n}(X_k-\mu)^2}{\chi^2_{1-\alpha/2}(n)}}, \sqrt{\frac{\sum_{k=1}^{n}(X_k-\mu)^2}{\chi^2_{\alpha/2}(n)}}\right)$$

is an $(1-\alpha)$-level CI for σ in the Normal Model.

PS: BTW, $\sqrt{\chi^2_{\alpha/2}(n)} \neq \chi_{\alpha/2}(n)$, there is no such thing as $\chi_{\alpha/2}(n)$. $\chi^2(n)$ is the name of the distribution, just like the Republic Square ◡̈

11.3.4 Confidence Interval for the Variance of Normal Distribution, mean is unknown

Let us consider again, like in the previous section, the parametric model $\{\mathcal{N}(\mu,\sigma^2) : \sigma^2 \in (0,+\infty)\}$, assuming that the value of μ in this case is unknown.

Assume also that we have a random sample from one of the distributions of that family:

$$X_1,...,X_n \overset{IID}{\sim} \mathcal{N}(\mu,\sigma^2),$$

for some unknown σ^2. Our aim is to construct a confidence interval for σ^2.

Unfortunately, we cannot use the same pivot as in the previous case, since now μ is unknown, and we cannot use the value of μ. In this case we will consider the r.v.

$$Y = \sum_{k=1}^{n}\left(\frac{X_k-\overline{X}_n}{\sigma}\right)^2,$$

which uses $\overline{X}_n$ instead of μ from the previous section. It can be proved (this is not obvious, but not too hard), that

$$Y \sim \chi^2(n-1).$$

Since the distribution of Y is independent of the parameter σ^2, then Y can serve as a pivot. Now, using the same technique as above, we'll obtain an exact $1-\alpha$ level CI for σ^2:

$$\left(\frac{\sum_{k=1}^{n}(X_k-\overline{X}_n)^2}{\chi^2_{1-\alpha/2}(n-1)}, \frac{\sum_{k=1}^{n}(X_k-\overline{X}_n)^2}{\chi^2_{\alpha/2}(n-1)}\right).$$

BTW, this can be written in the following, way, if we will use the Unbiased Sample Variance S^2:

$$\left(\frac{(n-1)\cdot S^2}{\chi^2_{1-\alpha/2}(n-1)}, \frac{(n-1)\cdot S^2}{\chi^2_{\alpha/2}(n-1)}\right).$$

EXAMPLE, CI FOR THE VARIANCE OF NORMAL DISTRIBUTION: If you will navigate to https://www.boschtools.com/us/en/boschtools-ocs/laser-measuring-glm-35-125633-p/, or the Amazon page for the same tool, https://www.amazon.com/Bosch-Distance-120-Feet-GLM-35/dp/B00VI7WBWE, you will find the details about the Bosch Compact Laser Distance Measure, GLM 35. On that pages, under the technical data (at the Amazon site, you can download the user's manual), you will find that the measuring accuracy is $\pm\frac{1}{16}$ inches. Have you ever wondered how people can assess the accuracy of devices like this? Let's look how statistics and CIs can help.

Assume we have a measuring device (say, the distance measure laser or an electronic scale), and we want to estimate the precision (accuracy) of that device. The idea is the following: we are doing the same measurement several times - usually, we will get different numbers (close to each other). Based on that numbers we want to estimate the standard deviation of all measurements that can be obtained by our device, and this will be our device's precision measure.

We assume that the errors of our device are normally distributed with the mean 0, that is, if we will measure something with the exact value μ, then the actual observations will be distributed $\mathcal{N}(\mu,\sigma^2)$ for some σ^2, i.e., our measurements results will be $\mu+\varepsilon$ for some $\varepsilon\sim\mathcal{N}(0,\sigma^2)$. Think why we can assume this.

Now, as we agreed above, we will do several measurements with our device. Say, I am calculating the distance between two fixed stolbs (lamppost) on our street, and doing the calculation 10 times, resulting in the following numbers (distances), in mm:

$$253.2,\ 235.9,\ 244.1,\ 246.6,\ 241.3,\ 249.0,\ 255.4,\ 244.4,\ 239.3,\ 250.0$$

I want to construct a CI for the standard deviation σ, based on these observations.

The above story was to estimate the variance, σ^2. Having the CI for the variance, we will construct easily the CI for the Standard Deviation.

Now, since the mean, the **exact** distance between that two lampposts is not available, we will use the case when the mean is unknown. We model our problem by: we have

$$X_1, X_2, ..., X_{10}\sim\mathcal{N}(\mu,\sigma^2),$$

we do now know the value of μ, and we want to construct the CI for σ^2. Here we have 10 observations, so $n=10$.

We choose the confidence level, say, 95%, so $1-\alpha=0.95$, i.e., $\alpha=0.05$. The CI for the Normal model σ^2, for the level $1-\alpha$ is, as we have found above,

$$\left(\frac{\sum_{k=1}^{n}(X_k-\overline{X}_n)^2}{\chi^2_{1-\alpha/2}(n-1)}, \frac{\sum_{k=1}^{n}(X_k-\overline{X}_n)^2}{\chi^2_{\alpha/2}(n-1)}\right)$$

First we calculate the quantiles of the distribution $\chi^2(9)$ (degrees of freedom, $df = n-1 = 9$, since we do not have the exact value of μ) of level $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$, using **R**: we run

```
qchisq(0.05/2, df = 9)
qchisq(1-0.05/2, df = 9)
```

and get $\chi^2_{0.025}(9) \approx 2.7$ and $\chi^2_{0.975}(9) \approx 19.0$. Now, the observed sample mean is

$$\overline{x} = \frac{253.2+235.9+244.1+246.6+241.3+249.0+255.4+244.4+239.3+250.0}{10} = 245.92$$

and (note that I am using small letters, since already, instead of r.v.s, I am plugging the realized values):

$$\sum_{k=1}^{10}(x_k-\overline{x})^2 = 340.656.$$

Finally, our estimated CI for the σ^2 will be:

$$\left(\frac{340.656}{19.0}, \frac{340.656}{2.7}\right) = (17.91, 126.15).$$

So the estimated CI for σ will be:

$$\text{CI for } \sigma = (\sqrt{17.91}, \sqrt{126.15}) = (4.23, 11.23).$$

The result can be interpreted as: with 95% confidence, the standard deviation of our measuring device is between 4.23mm and 11.23mm. Alas, my device is not as good as Bosch's.

And the complete code is here:

```
loquant <- qchisq(0.05/2, df = 9) #the lower quantile
upquant <- qchisq(1-0.05/2, df = 9) #the upper quantile

x <- c(253.2,235.9, 244.1,246.6, 241.3, 249.0,255.4, 244.4, 239.3, 250.0) #our data
xbar <- mean(x) #mean of the data, we will not use it
s <- sum((x - mean(x))^2)
lobo <- s/upquant #lower bound for the interval for sigma^2
upbo <- s/loquant #upper bound for the interval for sigma^2
lb <- sqrt(lobo) #lower bound for the interval for sigma
ub <- sqrt(upbo) #upper bound for the interval for sigma
c(lb, ub)
```

Remark, CI for the variance of $\mathcal{N}(\mu, \sigma^2)$: Again, as above, let us summarize what we have obtained for this model. The problem is: given a r.s.

$$X_1, X_2, ..., X_n \overset{IID}{\sim} \mathcal{N}(\mu, \sigma^2),$$

and $\alpha \in (0,1)$, we want to construct an $1-\alpha$-level CI for the unknown parameter σ^2.

We have considered 2 cases: when μ was known and unknown. Here we give the summary:

	μ is known	μ is unknown
Pivot:	$\chi^2 = \sum_{k=1}^{n} \left(\frac{X_k-\mu}{\sigma}\right)^2$	$\chi^2 = \sum_{k=1}^{n} \left(\frac{X_k-\overline{X}}{\sigma}\right)^2$
Distrib. of Pivot:	$\chi^2 \sim \chi^2(n)$	$\chi^2 \sim \chi^2(n-1)$
Region with prob. $1-\alpha$	$\chi^2_{\alpha/2}(n) < \chi^2 < \chi^2_{1-\alpha/2}(n)$	$\chi^2_{\alpha/2}(n-1) < \chi^2 < \chi^2_{1-\alpha/2}(n-1)$
CI for μ:	$\left(\frac{\sum_{k=1}^{n}(X_k-\mu)^2}{\chi^2_{1-\alpha/2}(n)}, \frac{\sum_{k=1}^{n}(X_k-\mu)^2}{\chi^2_{\alpha/2}(n)}\right)$	$\left(\frac{\sum_{k=1}^{n}(X_k-\overline{X}_n)^2}{\chi^2_{1-\alpha/2}(n-1)}, \frac{\sum_{k=1}^{n}(X_k-\overline{X}_n)^2}{\chi^2_{\alpha/2}(n-1)}\right)$

11.4 Confidence Interval for the Exponential Model

Assume we have a random sample

$$X_1, ..., X_n \overset{IID}{\sim} Exp(\lambda),$$

and we want to construct an $(1-\alpha)$-level CI for λ.

The idea of construction is to use the fact that a good estimator for $\frac{1}{\lambda}$ - the MLE, is $\overline{X}$.

Now, using $\overline{X}$, we want to construct a pivot. It is easy to show that $\lambda \cdot X_k \sim Exp(1)$, and using the Proposition 8.6, we can state that

$$\lambda \cdot X_1 + \lambda \cdot X_2 + ... + \lambda \cdot X_n \sim Gamma(n, 1),$$

hence,

$$n \cdot \lambda \cdot \overline{X} \sim Gamma(n, 1).$$

This means that the distribution of $n \cdot \lambda \cdot \overline{X}$ is independent on the parameter λ, so ti can serve as a pivot. Now, we solve

$$\mathbb{P}\left(a < n \cdot \lambda \cdot \overline{X} < b\right) = 1 - \alpha$$

to find a and b. Say, we can choose the quantiles cutting equal-area parts from the left- and right-hand tails: $a = \gamma_{\frac{\alpha}{2}}(n, 1)$ and $\gamma_{1-\frac{\alpha}{2}}(n, 1)$, where $\gamma_t(n, 1)$ is the quantile of order t for the $Gamma(n, 1)$ distribution. Then we'll get

$$\mathbb{P}\left(\frac{\gamma_{\frac{\alpha}{2}}(n, 1)}{n \cdot \overline{X}} < \lambda < \frac{\gamma_{1-\frac{\alpha}{2}}(n, 1)}{n \cdot \overline{X}}\right) = 1 - \alpha,$$

so

$$\left(\frac{\gamma_{\frac{\alpha}{2}}(n, 1)}{n \cdot \overline{X}}, \frac{\gamma_{1-\frac{\alpha}{2}}(n, 1)}{n \cdot \overline{X}}\right)$$

will be an $(1-\alpha)$-level CI for λ.

11.5 Asymptotic Confidence Intervals

In many cases the problem of constructing Confidence Intervals for a parameter is not an easy job. One of the approaches then is to construct Asymptotic (or Approximate) Confidence Intervals, which will give that approximate CI's for large n. Let us describe the idea.

Consider again a parametric model $\{\mathbb{P}_\theta : \theta \in \Theta\}$, and assume that θ is 1D, $\Theta \subset \mathbb{R}$.

Assume we have, in contrast to previous cases, an infinite random sample from one of the distributions $\mathbb{P}_\theta$ from that family,

$$X_1, X_2, ..., X_n, ... \overset{IID}{\sim} \mathbb{P}_\theta,$$

or a realization of that random sample, $x_1, ..., x_n, ...$, and again our aim is to estimate somehow the unknown parameter θ.

In this case we want to construct an interval asymptotically containing (covering) θ with some confidence level, with some probability. We already know the definition of the asymptotic CI, and let me give that here again:

Definition 11.3. *Assume $0 < \alpha < 1$, and let, for any n, $L_n = L_n(x_1, ..., x_n, \alpha)$, $U_n = U_n(x_1, ..., x_n, \alpha)$ be two functions with $L_n(x_1, ..., x_n, \alpha) \leqslant U_n(x_1, ..., x_n, \alpha)$ for all $(x_1, ..., x_n, \alpha)$. The sequence of random intervals*

$$(L_n, U_n) = \big(L_n(X_1, ..., X_n, \alpha);\ U_n(X_1, ..., X_n, \alpha)\big)$$

*is called an **asymptotic confidence interval sequence** (or asymptotic confidence interval estimator sequence) for θ of (asymptotic) level $1 - \alpha$, if for any $\theta \in \Theta$,*

$$\liminf_{n \to +\infty} \mathbb{P}(L_n < \theta < U_n) \geqslant 1 - \alpha.$$

One of the methods to construct asymptotic CI's is to use approximate pivots. The idea is explained in the following example.

11.5.1 Asymptotic CI for the Mean of General Distribution

Here we want to construct an Asymptotic CI for the unknown mean of general distribution based on the Slutsky theorem.

Assume we have a family of distributions with unknown mean μ and unknown variance σ^2, i.e., we consider the parametric family $\{\mathbb{P}_\mu : \mu \in \mathbb{R}\}$ with the property that if $X \sim \mathbb{P}_\mu$, then

$$\mathbb{E}(X) = \mu, \qquad \text{and} \qquad Var(X) = \sigma^2.$$

Assume also that we have an infinite random sample

$$X_1, X_2, ..., X_n, ... \overset{IID}{\sim} \mathbb{P}_\mu$$

for some unknown μ, and we want to construct an Asymptotic CI for μ.

By the CLT, we know that

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{D} Z \qquad \text{for some } Z \sim \mathcal{N}(0, 1).$$

Unfortunately, we cannot use this relation, since σ is unknown.

Now, we can use the following estimator for σ^2: we take

$$S_n^2 = \frac{\sum_{k=1}^n (X_k - \overline{X}_n)^2}{n - 1}.$$

So we are considering the following (asymptotic) pivot:

$$t = \frac{\overline{X}_n - \mu}{S_n/\sqrt{n}}$$

To show that t is a nice asymptotic pivot, we need to show that the asymptotic distribution of t is independent on the parameter μ. We write

$$t = \frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} = \frac{\sigma}{S_n} \cdot \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$$

It can be proved that

$$\frac{S_n}{\sigma} \xrightarrow{\mathbb{P}} 1, \qquad \text{and} \qquad \frac{\sigma}{S_n} \xrightarrow{\mathbb{P}} 1,$$

(see the Appendix 11.8 for this chapter) so, by using the Slutsky's theorem we can obtain

$$t = \frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} = \frac{\sigma}{S_n} \cdot \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{D} \mathcal{N}(0,1).$$

To finish, we can use an argument like in the construction of CI for μ for unknown σ^2 case: omitting some small details that can be found above, we can get (here $Z \sim \mathcal{N}(0,1)$)

$$\mathbb{P}\left(\overline{X}_n - z_{1-\alpha/2} \cdot \frac{S_n}{\sqrt{n}} < \mu < \overline{X}_n + z_{1-\alpha/2} \cdot \frac{S_n}{\sqrt{n}}\right) = \mathbb{P}\left(-z_{1-\alpha/2} < \frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} < z_{1-\alpha/2}\right) \to$$

$$\mathbb{P}(-z_{1-\alpha/2} < Z < z_{1-\alpha/2}) = 1 - \alpha,$$

so we have obtained an *exact asymptotic* CI for μ:

$$\left(\overline{X}_n - z_{1-\alpha/2} \cdot \frac{S_n}{\sqrt{n}}, \overline{X}_n + z_{1-\alpha/2} \cdot \frac{S_n}{\sqrt{n}}\right).$$

Cool!

Remark, CI for the General Distribution Mean: Of course, the above CI will contain the unknown parameter μ with probability $1-\alpha$ for the case when n is large. And, when n is large, we know that the $t(n-1)$ distribution is very close to the Standard Normal distribution, so the quantiles of $t(n-1)$ and $\mathcal{N}(0,1)$ will be very close in the case when n is large. That's why usually people use the following asymptotic CI for the mean μ in this general case:

$$\left(\overline{X}_n - t_{1-\alpha/2}(n-1) \cdot \frac{S_n}{\sqrt{n}}, \overline{X}_n + t_{1-\alpha/2}(n-1) \cdot \frac{S_n}{\sqrt{n}}\right).$$

The idea is to use the same intervals in the case when the variance σ^2 is unknown (this is the widely used case in practice) for the Normal Distribution mean and General Distribution Mean (compare this CI with the case of the Normal Distribution's mean's CI obtained above). The only difference is that in the General Distribution case, the CI is asymptotic, and needs to be used for large n (usually large in this context means $n > 30$).

This is because of the laziness of the human kind: lazy people do not want to remember two different CIs, one is enough, why to fill the memory with some extra (non-important) things? Better is to remember the football player names, party dates, and which boys is with which girl.

11.5.2 R Code for Confidence Intervals

R CODE, CI CONSTRUCTION:

```
#CI for the Bernoulli(p)
realprob <- 0.62
conf.level <- 0.95
a = 1 - conf.level
sample.size <- 100
no.of.intervals <- 100

plot.new()
plot.window(xlim = c(0,no.of.intervals), ylim = c(-1,2))
axis(1)
axis(2)
for(i in 1:no.of.intervals){
  x <- rbinom(sample.size, size = 1, prob = realprob)
  lo <- mean(x) - 1/sqrt(sample.size*a)
  up <- mean(x) + 1/sqrt(sample.size*a)
  segments(c(i), c(lo), c(i), c(up))
}
abline(h = realprob, lwd = 2, col = "red")



#CI for the Normal mu (sigma is unknown)
real.mu <-0.423
conf.level <- 0.95
a = 1 - conf.level
sample.size <- 10
no.of.intervals <- 100

plot.new()
plot.window(xlim = c(0,no.of.intervals), ylim = c(-4,4))
axis(1)
axis(2)
for(i in 1:no.of.intervals){
  x <- rnorm(sample.size, mean = real.mu, sd = 2.32)
  s <- sd(x)
  lo <- mean(x) - qt(1-a/2, df = sample.size - 1)*s/sqrt(sample.size)
  up <- mean(x) + qt(1-a/2, df = sample.size - 1)*s/sqrt(sample.size)
  segments(c(i), c(lo), c(i), c(up))
}
abline(h = realprob, lwd = 2, col = "red")
```

```
#Bernoulli(p)
#Assume the real parameter value is p=0.2
#Run several times!
p <- 0.2 #parameter real value, we will try to recover this
n <-20 #no. of observations
```

```
a <- 0.05 #confidence level
x <- rbinom(n,1,p) #binom(1,p)=bernoulli(p)
CILeft <-mean(x)-1/(2*sqrt(n* a))
CIRight <-mean(x)+1/(2*sqrt(n* a))
CI <- c(CILeft,CIRight)
plot.new()
plot(c(min(x)-1/(2*sqrt(n* a)), max(x)+1/(2*sqrt(n* a))), c(-0.1,0.5))
lines(c(CILeft,CIRight),c(0,0), lwd = 3)
points(p,0, col = "red", pch = 16, cex = 1)
```

!! Also add t.test, where one can obtain the confidence interval

11.6 Appendix

11.6.1 Chi-Squared Distribution

Assume

$$X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(0,1).$$

Then the distribution of the sum

$$X = X_1^2 + X_2^2 + ... + X_n^2$$

is called the Chi-Square distribution with n degrees of freedom, and is denoted by $\chi^2(n)$. The PDF of $\chi^2(n)$ r.v. is given by

$$f(x) = \begin{cases} \frac{1}{2^{n/2}\Gamma(n/2)} \cdot x^{n/2-1} \cdot e^{-x/2}, & x \geqslant 0 \\ 0, & \text{otherwise,} \end{cases}$$

see the graphs of PDF of $\chi^2(1)$ and $\chi^2(4)$ variables in Fig. 11.2, 11.3.

Clearly,

$$\chi^2(n) \equiv \text{Gamma}\left(\frac{n}{2}, 2\right).$$

It is easy to calculate that if $X \sim \chi^2(n)$, then

$$\mathbb{E}(X) = n, \qquad \text{and} \qquad Var(X) = 2n.$$

For example,

$$\mathbb{E}(X) = \mathbb{E}(X_1^2 + X_2^2 + ... + X_n^2) = \sum_{k=1}^{n} \mathbb{E}(X_k^2) \overset{X_k \text{ are IID}}{=\!=} n \cdot \mathbb{E}(X_1^2) = n,$$

since

$$\mathbb{E}(X_1^2) = Var(X_1) + \left[\mathbb{E}(X_1)\right]^2 \overset{X_1 \sim \mathcal{N}(0,1)}{=\!=} 1 + 0^2 = 1.$$

R Code: R code to draw the PDF of $\chi^2(k)$ for $k = 1, 3$

```
#ChiSquare PDF
curve(dchisq(x, df=1), from = 0, to = 7)
curve(dchisq(x, df=3), from = 0, to = 7)
```

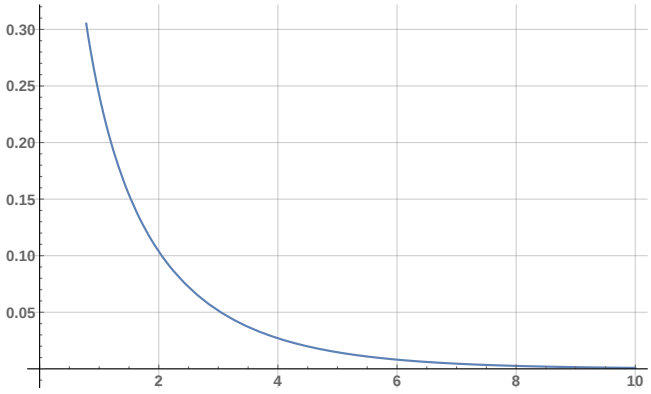

Fig. 11.2: PDF for r.v. with $\chi^2(1)$, 1 degree of freedom

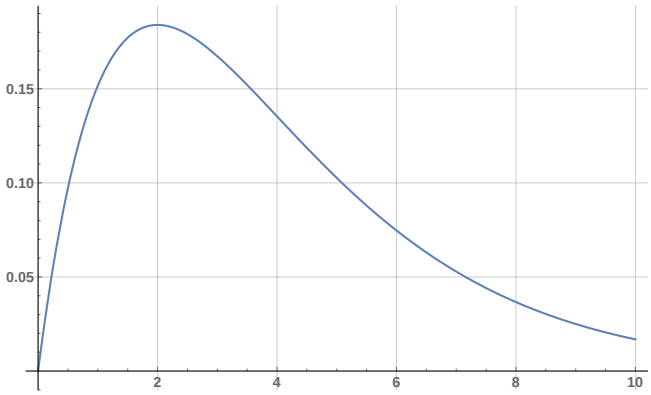

Fig. 11.3: PDF for r.v. with $\chi^2(4)$, 4 degree of freedom

11.6.2 Student's t-distribution

Assume

$$X_0, X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(0,1).$$

Then the distribution of the ratio

$$X = \frac{X_0}{\sqrt{\frac{X_1^2+X_2^2+...+X_n^2}{n}}}$$

is called the Student's t-distribution with n degrees of freedom and is denoted by $t(n)$. In fact, we could define t-distribution as the following: assume $X_0 \sim \mathcal{N}(0,1)$ and $Y \sim \chi^2(n)$. Then the r.v.

$$X = \frac{X_0}{\sqrt{Y/n}}$$

will have a Student's t-distribution with freedom degree n.

The PDF of r.v. $X \sim t(n)$ is

$$f(x) = \frac{1}{\sqrt{n\pi}} \cdot \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \cdot (1+\frac{x^2}{n})^{-\frac{n+1}{2}}, \qquad x \in \mathbb{R}$$

It is known that the PDF of $X \sim t(n)$ is symmetric about 0,

$$\mathbb{E}(X) = 0 \qquad \text{and} \qquad Var(X) = \frac{n}{n-2}, \qquad n > 2.$$

For $n = 1, 2$, the Variance is infinite.

R Code: R Code for PDF

```
#t Distrib PDF
curve(dt(x, df=1), from = -7, to = 7)
curve(dt(x, df=3), from = -7, to = 7)
```

11.6.3 Supplementary

Theorem 11.1. *Assume $X_1, X_2, ..., X_n$ are IID r.v.,*

$$\overline{X} = \frac{X_1 + X_2 + ... + X_n}{n} \qquad \text{and} \qquad S^2 = \frac{\sum_{k=1}^{n}(X_k - \overline{X})^2}{n-1}.$$

Then the followings hold:

1. *If $X_k \sim \mathcal{N}(\mu, \sigma^2)$ for some μ, σ, then $\overline{X}$ and S^2 are independent;*
2. *If the variance of X_k exists and $\overline{X}$ and S^2 are independent, then X_k are Normally Distributed with some parameters.*

Proof. See [1] □

11.7 (18+) Prediction Intervals

Up to this point we were considering CI for the unknown parameter of a distribution, based on the observations from that distribution. Here, in this section, we will construct *Prediction Intervals* (PI). The idea is the following: assume we have some observation from an unknown distribution (more precisely, in our settings, from the parametric family of distributions[10]). Now, can we give some reasonable interval where **a new observation from that distribution** will be with high probability?

More precisely, we assume that we have a random sample

$$X_1, X_2, ..., X_n \overset{IID}{\sim} \mathbb{P}_\theta,$$

where θ is our unknown parameter. And assume $X_{n+1} \sim \mathbb{P}_\theta$ is our next observation (r.v.), which is independent of $X_1, ..., X_n$. For a given prediction level $\alpha \in (0, 1)$, we want to find a random interval (L, U), constructed using our random sample, such that

$$\mathbb{P}(L < X_{n+1} < U) \geqslant 1 - \alpha.$$

We will not talk about the general case, but will consider only the Normal Distribution case, with unknown mean and known/unknown variance. So our problem is the following:

Problem: We are given a random sample

$$X_1, X_2, ..., X_n \overset{IID}{\sim} \mathcal{N}(\mu, \sigma^2),$$

where μ is unknown, and σ^2 is known (case 1) or unknown (case 2), and we want to find a $(1-\alpha)$-level Prediction Interval for our new observation $X_{n+1} \sim \mathcal{N}(\mu, \sigma^2)$, having that X_{n+1} is independent of $X_1, ..., X_n$.

[10] I.e., we know the family of distributions, from which our observation comes, up to some parameter, and we do not know the exact value of the parameter.

Case 1: σ^2 is known. We will use again the same pivoting technique: we will use the sample mean $\overline{X}_n = \frac{X_1+\ldots+X_n}{n}$ to construct the prediction interval for X_{n+1}, i.e., we suppose

$$X_{n+1} \approx \overline{X}_n \pm something.$$

To that end, we consider the r.v. $Y = \overline{X}_n - X_{n+1}$. Y is Normally distributed, as a sum of independent Normally distributed r.v.-s. The expected value of Y is

$$\mathbb{E}(Y) = \mathbb{E}(\overline{X}_n - X_{n+1}) = \mu - \mu = 0,$$

and the variance is

$$Var(Y) = Var(\overline{X}_n - X_{n+1}) = Var(\overline{X}_n) + Var(X_{n+1}) = \frac{\sigma^2}{n} + \sigma^2 = \sigma^2 \cdot \left(1 + \frac{1}{n}\right).$$

Hence, the r.v.

$$Z = \frac{Y}{\sqrt{Var(Y)}} = \frac{\overline{X}_n - X_{n+1}}{\sigma \cdot \sqrt{1+\frac{1}{n}}} \sim \mathcal{N}(0,1),$$

so Z will be our pivot to construct a PI for X_{n+1}. Now, as in CI story, we know that for $Z \sim \mathcal{N}(0,1)$,

$$\mathbb{P}(-z_{1-\alpha/2} < Z < z_{1-\alpha/2}) = 1-\alpha,$$

where $z_{1-\alpha/2}$ is the $1-\frac{\alpha}{2}$ - level quantile of the Standard Normal Distribution. Hence,

$$\mathbb{P}\left(-z_{1-\alpha/2} < \frac{\overline{X}_n - X_{n+1}}{\sigma \cdot \sqrt{1+\frac{1}{n}}} < z_{1-\alpha/2}\right) = 1-\alpha,$$

and, solving for X_{n+1} will yield

$$\mathbb{P}\left(\overline{X}_n - z_{1-\alpha/2} \cdot \sigma \cdot \sqrt{1+\frac{1}{n}} < X_{n+1} < \overline{X}_n + z_{1-\alpha/2} \cdot \sigma \cdot \sqrt{1+\frac{1}{n}}\right) = 1-\alpha,$$

and our PI for X_{n+1} is

$$\overline{X}_n \pm z_{1-\alpha/2} \cdot \sigma \cdot \sqrt{1+\frac{1}{n}}.$$

And the interpretation is the following: if we know the value of σ, then with probability $1-\alpha$, the value of X_{n+1} will be in the above PI.

Remark, PI for the Normal Case: Please note that in the case when $n \to +\infty$, the prediction interval shrinks to

$$\mu \pm z_{1-\alpha/2} \cdot \sigma,$$

and the interpretation is that if we have infinitely many X_k-s, then we can find exactly the mean μ (because of our good friend LLN!), and if we have some $X \sim \mathcal{N}(\mu, \sigma^2)$, then with probability $1-\alpha$,

$$X \in \left(\mu - z_{1-\alpha/2} \cdot \sigma, \mu + z_{1-\alpha/2} \cdot \sigma\right),$$

which, of course, can be seen from $X \sim \mathcal{N}(\mu, \sigma^2)$ by elementary calculations.

Case 2: σ^2 is unknown. If we assume that σ is unknown, then we cannot use the above PI. But we can do the following modifications from the above calculations: we consider the following Statistics (pivot):

$$t = \frac{\overline{X}_n - X_{n+1}}{S_n \cdot \sqrt{1+\frac{1}{n}}}$$

where S_n is the Sample Standard Deviation for $X_1, ..., X_n$ (with the denominator $n-1$). So in the above statistics Z we substitute the unknown σ by its estimate (constructed using $X_1, ..., X_n$), S_n. The heartbreaking and/or breathtaking fact is the the distribution of t is known:

$$t = \frac{\overline{X}_n - X_{n+1}}{S_n \cdot \sqrt{1+\frac{1}{n}}} \sim t(n-1).$$

Hence, we can apply all the above technique and obtain the following $(1-\alpha)$-level PI for X_{n+1}:

$$\overline{X}_n \pm t_{1-\alpha/2} \cdot S_n \cdot \sqrt{1+\frac{1}{n}}.$$

The interpretation is the following: if we do not have the value of σ, then with probability $1-\alpha$, the value of X_{n+1} will be in the above PI.

REMARK, EXAMPLES FOR PI: See https://en.wikipedia.org/wiki/Reference_ranges_for_blood_tests and https://en.wikipedia.org/wiki/Reference_range

11.8 Appendix

First we want to show that this is a consistent estimator for σ^2, and also that s_n is consistent estimator for σ. To this end, we define our ordinary Sample Variance as

$$S_n^2 = \frac{\sum_{k=1}^n (X_k - \mu)^2}{n-1}.$$

If we will use the weak LLN for r.v.'s $Y_k = (X_k - \mu)^2$, we'll obtain that[11]

$$\frac{Y_1 + ... + Y_n}{n} \xrightarrow{\mathbb{P}} \mathbb{E}(Y_1) = \sigma^2.$$

Now,

$$S_n^2 = \frac{Y_1 + ... + Y_n}{n-1} = \frac{n}{n-1} \cdot \frac{Y_1 + ... + Y_n}{n} \xrightarrow{\mathbb{P}} \sigma^2.$$

Next, let us calculate

$$S_n^2 - \hat{\sigma}_n^2 = \frac{1}{n-1} \sum_{k=1}^n \left[(X_k - \mu)^2 - (X_k - \overline{X}_n)^2\right] =$$

$$= \frac{1}{n-1} \sum_{k=1}^n \left[(X_k - \mu - X_k + \overline{X}_n)(X_k - \mu + X_k - \overline{X}_n)\right] =$$

[11] We can use the weak LLN, since $\mathbb{E}(|Y_k|) = \mathbb{E}(Y_k) = Var(X_k) = \sigma^2 < +\infty$.

$$= \frac{1}{n-1}\sum_{k=1}^{n}\left[(\overline{X}_n-\mu)(2X_k-\mu-\overline{X}_n)\right] = \frac{(\overline{X}_n-\mu)}{n-1}\sum_{k=1}^{n}\left[2X_k-\mu-\overline{X}_n\right] =$$

$$= \frac{(\overline{X}_n-\mu)}{n-1}\left[2n\overline{X}_n-n\mu-n\overline{X}_n\right] = \frac{(\overline{X}_n-\mu)}{n-1}\left[n\overline{X}_n-n\mu\right] = \frac{n}{n-1}(\overline{X}_n-\mu)^2.$$

Now, because of weak LLN, $\overline{X}_n-\mu \xrightarrow{\mathbb{P}} 0$, implying that $(\overline{X}_n-\mu)^2 \xrightarrow{\mathbb{P}} 0$, so $S_n^2-\hat{\sigma}_n^2 \xrightarrow{\mathbb{P}} 0$, and since $S_n^2 \xrightarrow{\mathbb{P}} \sigma^2$, then also $\hat{\sigma}_n^2 \xrightarrow{\mathbb{P}} \sigma^2$.

Now, since the function $g(x)=\sqrt{x}$ is continuous in his domain, we will have also

$$\sqrt{\hat{\sigma}_n^2} \xrightarrow{\mathbb{P}} \sqrt{\sigma^2},$$

i.e.

$$\hat{\sigma}_n \xrightarrow{\mathbb{P}} \sigma.$$

Ufff, finally!

Now, continuity of $g(x)=\frac{1}{x}$ in $(0,+\infty)$, and the fact that $\hat{\sigma}_n \in (0,+\infty)$, we invoke $\frac{\sigma}{\hat{\sigma}_n} \xrightarrow{\mathbb{P}} 1$.