Inference_project

Ignacio Almodóar & Limingrui Wan

1. Introduction

1.1 Description of the project and motivation.

The aim of this project is to garner knowledge in the vehicles market, specifically about electric vehicles, which is a segment that is growing year by year.

The electric vehicle market has grown a lot in the last years and each year different brands are releasing new models with different specs and capacities. The automobile industry has been around us for almost 130 years, which means that almost everybody has at least a little bit of knowledge about the market in terms of traditional cars (combustion engines), if not probably people around you (relatives or friends) might know enough to give you recommendations that suits your needs.

However, what about electric cars? Most people struggle even telling prices for models that have already been in the market for more than five years, like Model S from Tesla or BMW i3, which are probably the ones that most people have seen several times on the streets, specially if you live in the city center.

As electric cars has always have that label of "expensive cars" or "useless noways", people still do not have the necessity to start considering them as feasible options to buy when looking for a new car. Which ends up with an uninformed society when it comes to this "new age" of cars.

Nevertheless this mentality has been changing, specially in the last years. It is not difficult to see several electric cars in one day, which means that people are starting to care about electric cars and to learn about them. However, it is sometimes difficult to compare between electric cars and traditional ones, because it is not part of what people is used to, and the prices might vary a lot for different models that might look the same just looking at the specs. Also, as new brands have come, many people do not know if their products are worth the price or not.

With this said, we have considered that this study could be useful for both companies and consumers. On one hand, companies that want to start developing electric cars might find helpful this study in order to consider several aspects while designing an electric car in terms of prices and characteristics. On the other hand for those consumers who want to take a peek in the electric cars market, allowing them to compare different prices in terms of power, efficiency or even number of seats and to see where those prices are going while the markets grows.

If you want to take the research on electric vehicles to the next level, the information that is given in this study might be very useful in order to compare between the traditional car market and the electric one. Also, you could make predictions about prices for the next years and which aspects are significant evolving.

1.2 Description of the data set

This data set has been taken from kaggle, which is a crowd-sourced platform to attract data scientists from all around the world to solve data science, machine learning and predictive analytics problems. It has more than one million of members, whereas more than half of the community are active members.

Kaggle enables data scientists and other developers to engage in running machine learning contests, write and share code, and to host data sets.

Even though we have taken the set from Kaggle, all the data that contains this data set is scrapped from an online electric cars gauge web Electric Vehicle Database, where you have different sections for finding and compare different electric vehicles.

This data set originally contains 177 different car models with 11 different aspects for each of them. These aspects are considered the most relevant ones while searching for information of different electric cars. These are also the ones that are directly related to the price and the ones that can be easily compared to one or another.

1.3 Population of our sample size

While searching through the web from where this data set has been scrapped, you can see that they have different sections for electric cars:

- Most recent
- Cheapest EV
- Most Efficient
- Quickest 0-100
- Longest range

In our case we have taken the ones listed for the cheapest electric car vehicles section, which necessary do not have only cheap cars, however it might show the most affordable ones from the ones with similar characteristics.

Notice that this data set is continuously updating, which means that you might get different results depending on when you do your research.

In short, we can consider that our population is the whole electric brand new market, whereas our sample is a summary of those who are considered the cheapest one for their characteristics.

1.4 Description of the variables

As we have mentioned, this data set contains eleven columns which basically tells us the most important aspects for every car.

On each of these different columns we will find different types of variables, some of them might be continuous, some others might be discrete with very few possible values, where ass others are discrete with a higher range of values.

On the other hand we have some of them that does not gives us much information or are difficult to understand if you are not very much into electric cars.

In essence, these are the variables that our data set contains:

- Name: This variable contains each car model name, including the brand and the exact model. This column strictly consists on string values, so we are not going to apply inference directly on it.
- Subtitle: Indicates the type of car and the capacity of the battery. The capacity of the batteries is measured in kWh, which is a unit of energy equal to one kilowatt of power sustained for one hour.
- Acceleration: Measures the time (in seconds) necessary for the model to accelerate from 0-100km/h. This is always a good indicator about the power that the engine provides in relation to the car weight. It is a continuous variable with range between 2.1 and 22.4 seconds.
- TopSpeed: Shows the maximum speed that the vehicle can reach. It is measured in Km/h and it shows how well does the car leverages the potential of its batteries. It is also a continuous variable with values between 123 and 410 km/h.
- Range: Shows the approximate distance a vehicle can travel with a 100% charge, which is always a good indicator about the real capacity of the batteries that the vehicle has in terms of common-daily use. It is measured in km and it is a continuous variable with range 95-970 km.
- Efficiency: This is a relatively difficult variable to understand. It calculates the battery energy consumption used by the vehicle for propulsion and on-board systems, which basically tells us the average consumption of energy per kilometers. This could be easily compared as the average liters per kilometer consumption of fuel for a combustion engine powered car. It is a continuous variable measured in Wh/km. Its range is 104-281 Wh/km

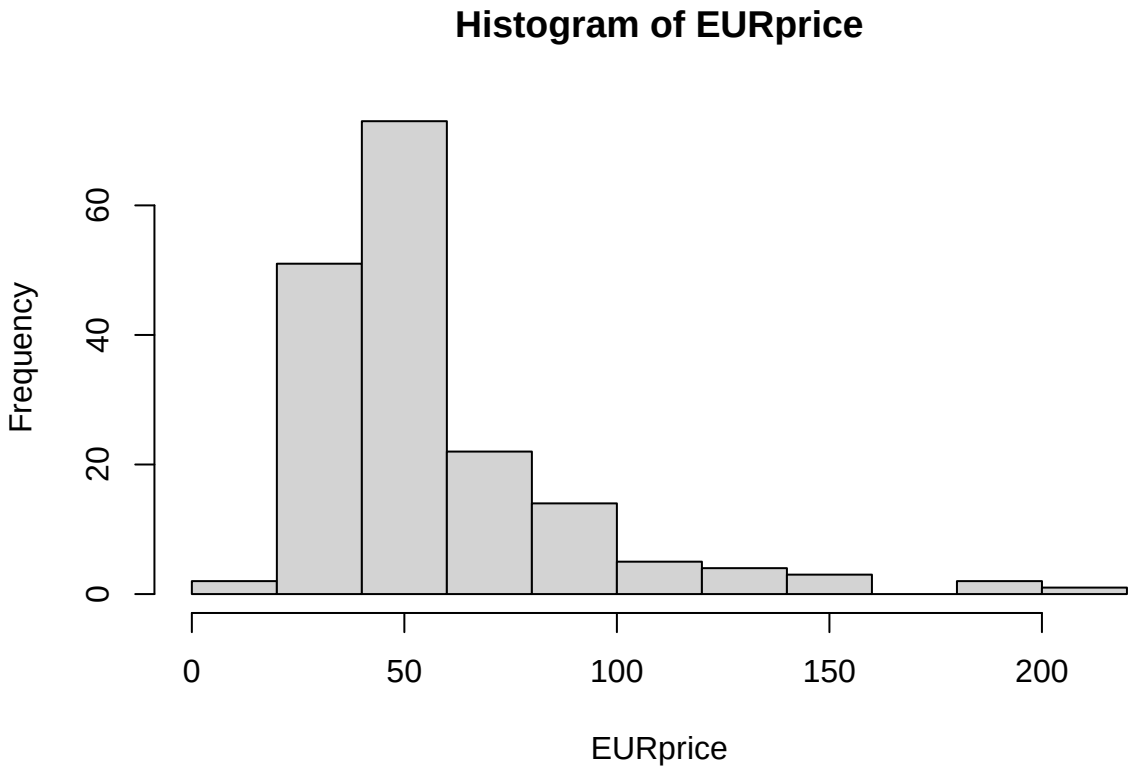
- FastChargeSpeed:This is also a strange variable because it is measured in km/h and it rates the average charging speed over a session from 10% to 80%. It is a continuous variable that fluctuates between 0 and 980 km/h for 0 meaning that it does not have fast charge mode.
- Drive: It defines which pair of wheels are the driving wheels of the vehicle, which is a very important attribute while buying a car. It is a discrete variable with three values (front wheel, Rear wheel and all wheel)
- NumberOfSeats: Indicates the number of legally seats available in the car. It is a discrete variable that oscillates between 2 and 7.
- PriceinGermany: It shows the retail price un Germany for a brand new model. It is a continuous variable and it is measured in Euros. The prices vary between 18460 and 215000 €.
- PriceinUKs: Same concept as before whereas in this case it for UK and it is measured in pounds. Notice that it is not strictly the conversion between euros and pounds, for most of them, the price will be higher in the UK.

This is a resume of all the variables that our data set contains. However we might not use some of them in this project as the information that they provide might not be very useful for the conclusions that we want to reach. Due to this, we might filter our data set in order to use only the columns that we find interesting.

2. Model Selection

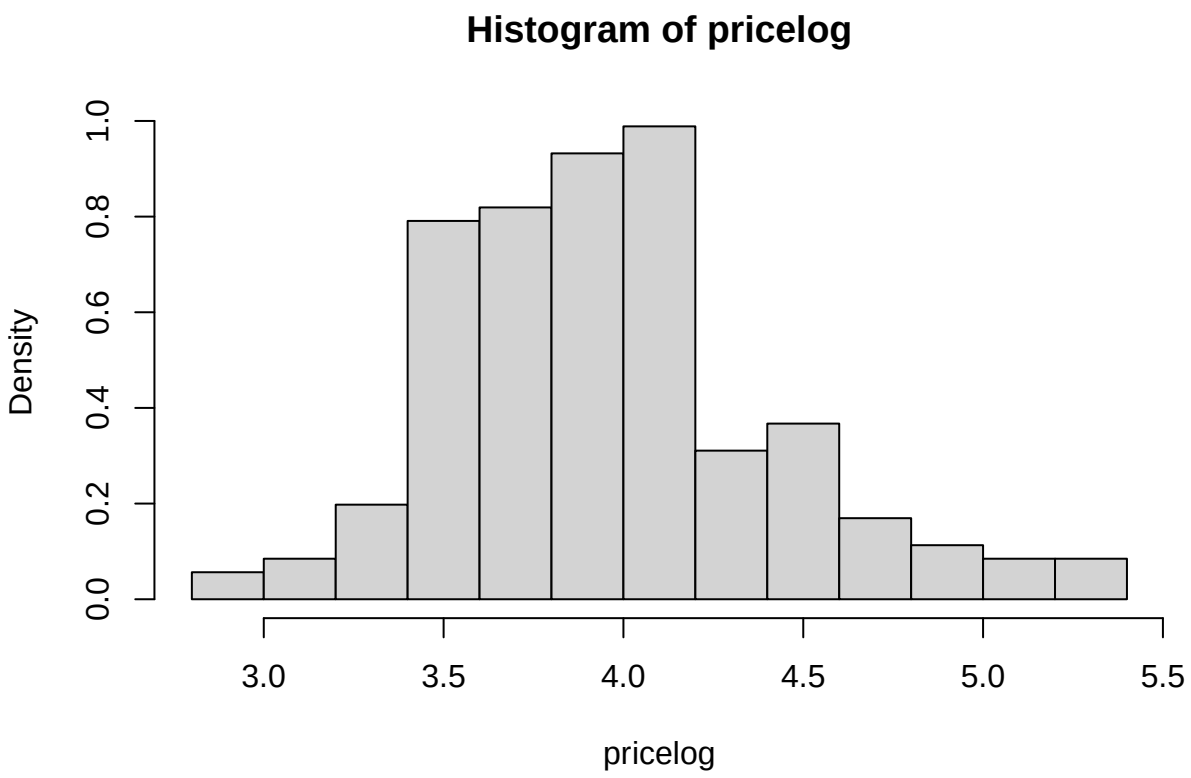
2.1 Pobability distribuction of our continuous random variable "Price in Germany"

In order to get a quick view of how our variable is distributed, we have to plot it. A good way to view it is by plotting a histogram of the variable.



First of all, if we want to model our variable we have to make sure that our variable is normally distributed. By plotting directly our variable that contains the price in EUR for each car in Germany, we can see that it is asymmetric, which means that it violates the assumptions of linear regression.

In order to transform our variable into a normally distributed one, we can apply logarithms to it. This way we make sure that our variable is normally distributed and we can model it with linear regression.



2.2 Estimation of model parameters.

Analyzing these histograms we can assume that our random variable follows a gamma distribution $\Gamma(\alpha, \lambda)$ with parameters:

- Shape k, $\alpha = k$
- Scale θ, $\lambda = \frac{1}{\theta}$

As we want to estimate our model parameters $\hat{\alpha}$ and $\hat{\lambda}$, we are going to use the method of moments in order estimate it.

The method of moments consists in leveling population moments and sample moments in order to get an equation or system of equations from where to get our wanted parameters.

For a gamma distribution, we get the following equations:

- $E[x] = \frac{1}{n}\sum_{i=0}^{n} x = \overline{X}$ for the fist population moment
- $E[x^2] = \frac{1}{n}\sum_{i=0}^{n} x_2$ for the second population moment

As we know that $E[x] = \frac{\alpha}{\lambda}$ for a Gamma distribution, from the first function we get that:

$\alpha = \overline{X}\lambda$

Replacing α on the second equation and simplifying it, we obtain:

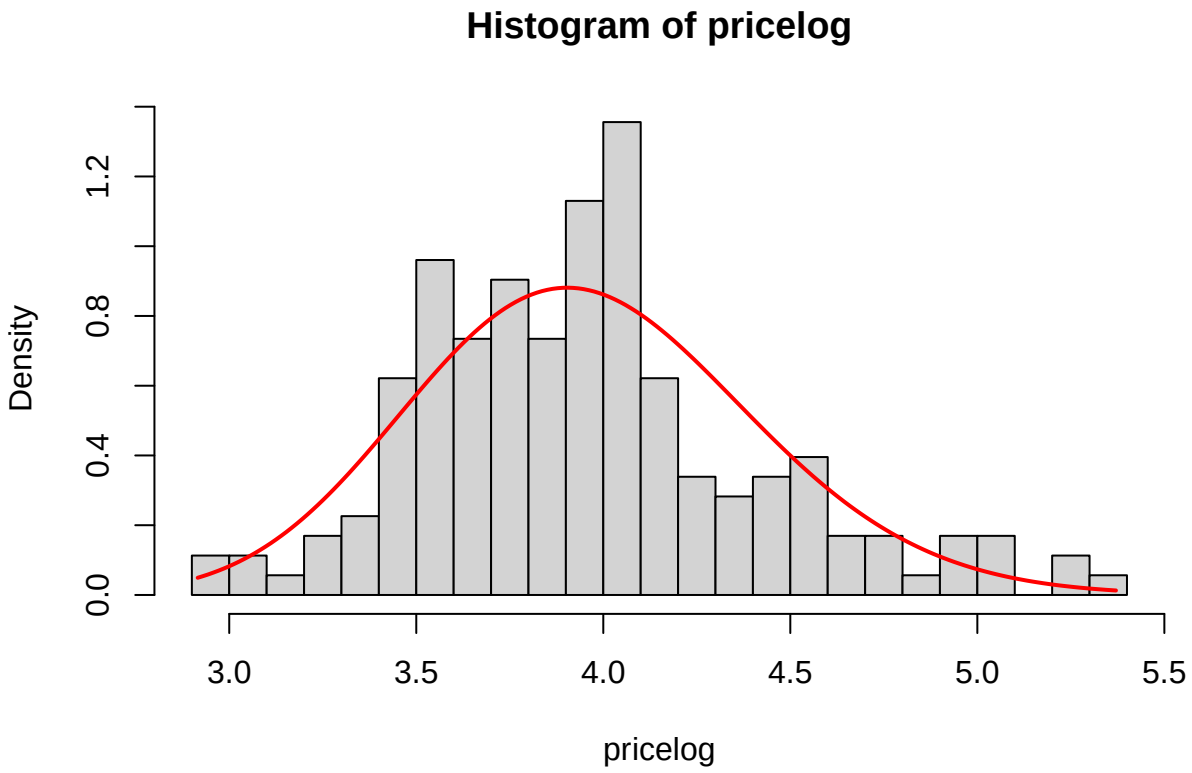
$\hat{\lambda} = \frac{\overline{X}}{\frac{1}{n}\sum_{i=0}^{n} x_i^2 - \overline{X}^2}$

Whereas, using $\alpha = \overline{X}\lambda$, we get:

$\hat{\alpha} = \overline{X}\hat{\lambda} = \frac{\overline{X}^2}{\frac{1}{n}\sum_{i=0}^{n} x_i^2 - \overline{X}^2}$

Using these results we can calculate our estimators using R. Running the code above we can say that our variable "Price in Germany" follows a gamma distribution $\Gamma(\hat{\alpha}, \hat{\lambda})$ with parameters $\hat{\alpha} = 569.5397$ and $\hat{\lambda} = 52.419$

```
## starting httpd help server ... done
```



3. One-sample Inference

3.1 Estimators for our population mean

unbiasedness

In this section we can use sample mean and sample median value to estimate our population mean.

```
## [1] "the sample mean is  3.9559243422887"
```

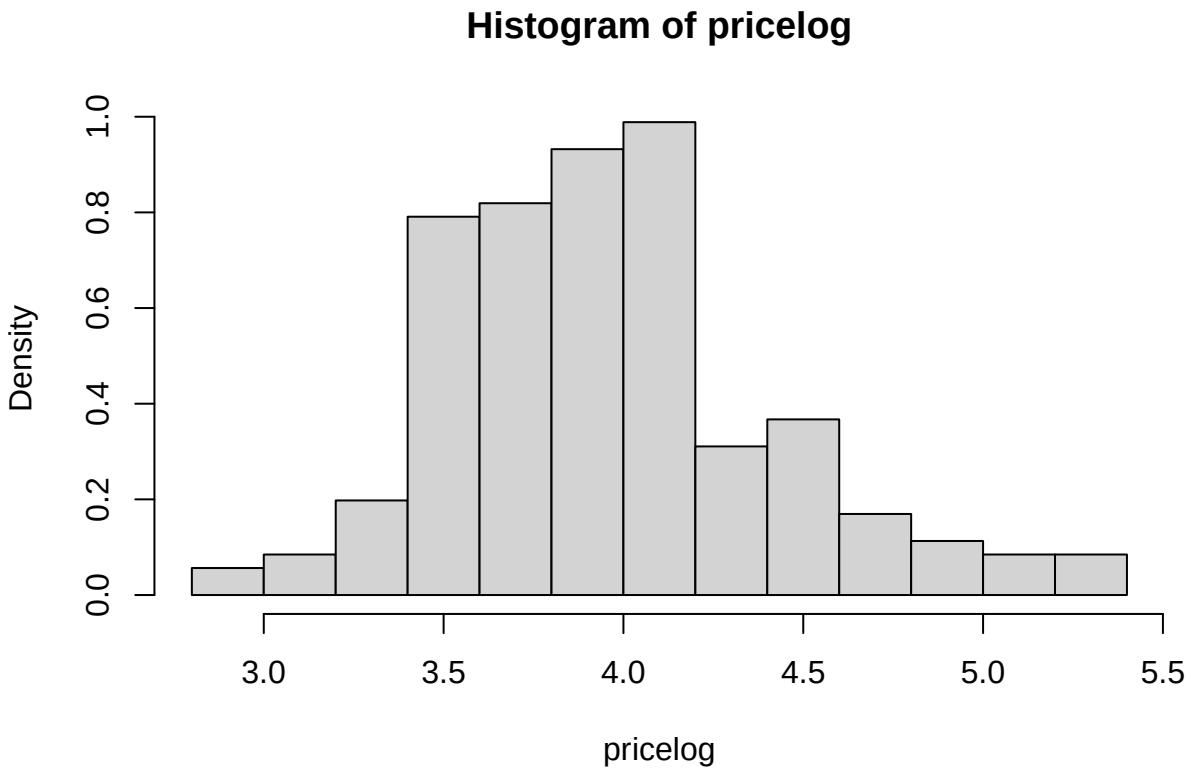
```
## [1] "the sample median is  3.92028874984518"
```

we can quickly give the distribution of the sample mean:

$$\hat{\theta} \sim N(\mu, \sigma^2/n)$$

since we assumed a gamma distribution, we have $\mu = \frac{\alpha}{\beta}$ and $\sigma^2 = \frac{\alpha}{\beta^2}$

obviously the sample mean is an unbiased estimator, as for the sample median value we know that gamma distribution is not symmetric. To proof this we can check again our plot.



More specifically, we can get the pdf of the median value via the formula for order statistics, which is

$$\rho_k(x) = \frac{n!}{(k-1)!(n-k)!}(F(x))^{k-1}(1-F(x))^{n-k}\rho(x)$$

where n is the total number, k is the order number, $\rho(x)$ is the pdf of the general distribution, $F(x)$ is the cdf.

To proof if this median value we can compute the expectation of the estimator, so we shall calculate such indefinite integral:

$$\int x\rho_k(x)dx$$

3.2 Estimation of the errors of the two estimators

Since sample mean is an unbiased estimator, so we can calculate the error by CV, which is

$$cv = \frac{sd(\hat{\theta})}{exp(\hat{\theta})} = \frac{\sqrt{\frac{\hat{\alpha}}{n\hat{\beta}^2}}}{\frac{\hat{\alpha}}{\hat{\beta}}} = \sqrt{\frac{1}{n\hat{\alpha}}} = 0.017$$

as for the median value, we can compute the RRMSE $RRMSE = \sqrt{E(\hat{\theta} - \theta)^2}$

3.3 Confidence interval of 95% for our population mean

we are not having the true distribution of the price in population, so we don't know the true variance, thus we don't know the variance of mean value. So we decide to use the method of the pivotal quantity for μ

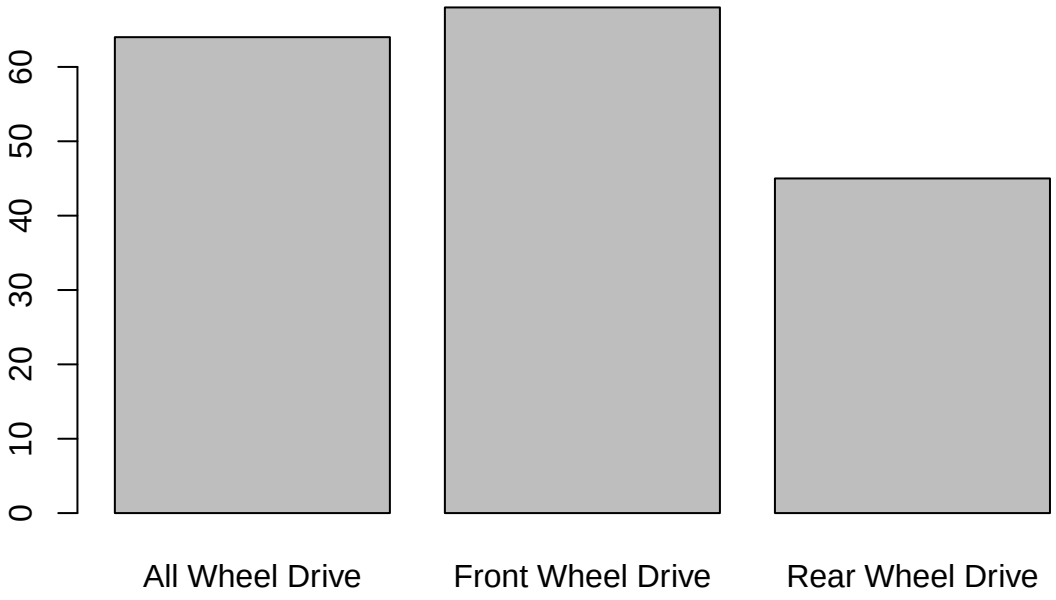
$$CI_{1-\alpha}(\mu) = \overline{x}_n \pm t_{n-1,\alpha/2}\frac{S'}{\sqrt{n}}$$

```
## [1] 3.054793 4.857055
```

3.4 Estimation of the proportion of units in the population that belongs to the category selected

We divide the sample into three groups according to the variable "drive", which implies how the specific car is driven. We consider a Binomial distribution$Bern(p)$, the estimator actually estimate the mean value of the probability in Binomial test, so the distribution of estimated proportion is$N(p, \frac{1(1-p)}{n})$

```
##
##   All Wheel Drive Front Wheel Drive  Rear Wheel Drive
##                64                68                45
```



the estimated proportion is:

```
## [1] "the estimated proportion of All wheel Drive car is  0.361581920903955"
## [1] "the estimated proportion of Front Wheel Drive car is  0.384180790960452"
## [1] "the estimated proportion of Rear Wheel Drive  car is  0.254237288135593"
```

3.5 Variance of our estimator of proportion

the variance of the estimator is:

```
## [1] 0.001304183
## [1] 0.001336644
## [1] 0.00107119
```

3.6 Population proportion with 95% of confidence interval

we can give the proportion by such formula, so we can get such results

$$\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

```
## [1] 0.2908008 0.4323630
## [1] 0.3125242 0.4558373
## [1] 0.1900895 0.3183850
```