Simon Beylat 01/04/2021

METEORITE DATA RECORDINGS

Mario Veruete

Exam Data Engineering

1 Questions

Before starting to answer the question, we just let you know that the Python program uses those those libraries:

- Panda
- Numpy
- Matplolib
- GeoPanda
- LinearRegression
- Sjoin
- Rtree
- Gaussian_kde

So, if you need, my code is commented so you can just read through it and find the parts related to each question. Before any data processing, I decided to clean the data by removing all the data with missing information, so we have 38115 meteorites

Here the link Github of my python code.

1.

For the first question, we have to make a histogram of the mass distribution of the meteorites. For this we used Panda and we managed to get the graph below.

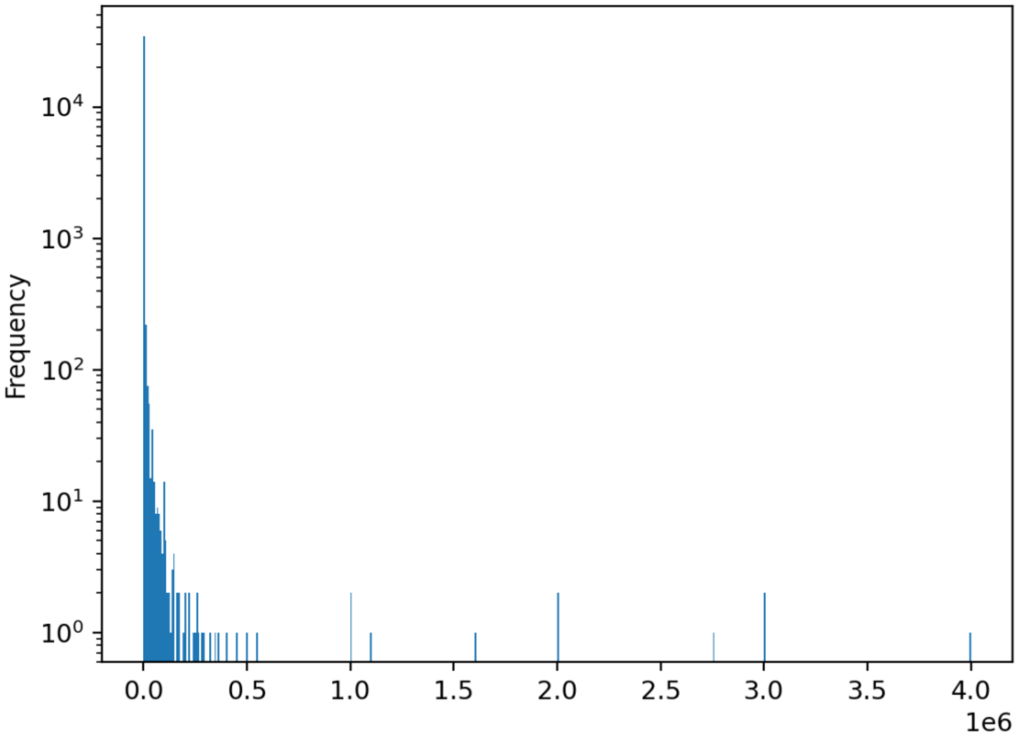


Figure 1. Histogram of the distribution of the mass of meteorites

We see that most of meteorites are between 0 and 50 000 grams. We also see that there are some more big meteorites but there is only one which has been reported. With python we can see that 37684 meteorites have a weight inferior to 50 000 grams.

The second part of the first question is to select only meteorites with a mass of 50 000 grams or less. So, we built another graph that you can see again just below.

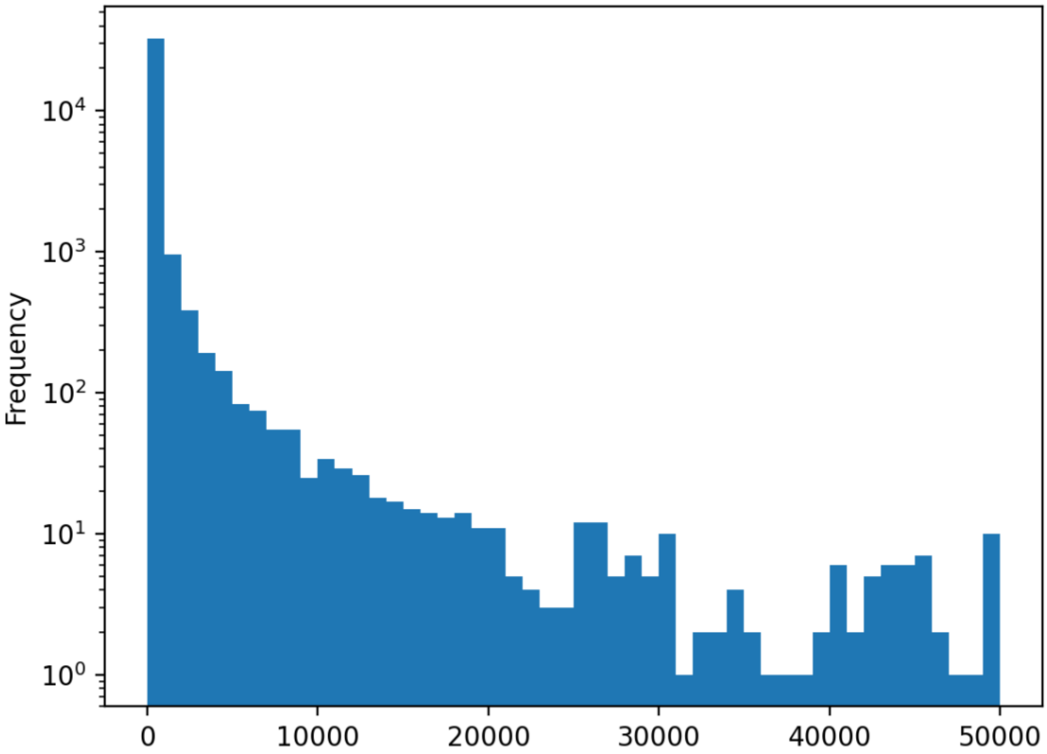


Figure 2. Histogram of the distribution of mass lower than 50 000 grams of meteorites

Here the graph is much more readable. We have a distribution that is descending. We have also put our two graphs in logarithmic axis for a better reading.

To conclude this first question here we have obtained the two graphs with the distribution of the meteorites between the masses, and we see that the graph is decreasing. Moreover, the meteorites of more than 50kg are present only 1,13% of the meteorites recorded. we can conclude that neglecting these meteorites would not be a problem if we wanted to do more processing on the weights in order to avoid the problems of exceptional cases

2.

For this question we need to construct another graph determining the number of meteorites versus time per year. Second, we must look for an approximation of the points by a linear fit of this form (y=ax+b). Finally, we must conclude and justify this approach to predicting the future of meteorite data. First we took all the data from the database and plotted the data with years and numbers on the axes. And this is what we got.

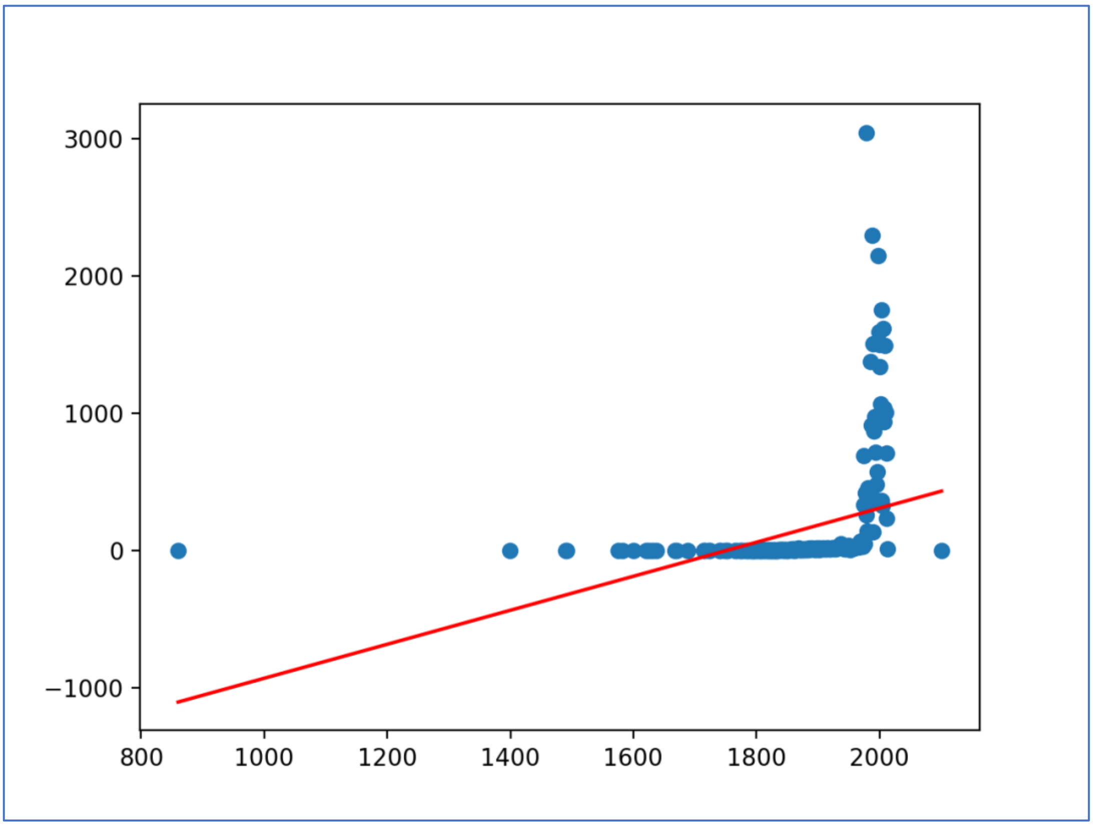


Figure 3. Distribution of the number of meteorites reported by year.

We notice that from the year 1975 there is a strong increase in the number of meteorites. This seems logical because the technological advances have allowed us to locate many more meteorites.

We also notice that the linear approximation is very questionable because the information before 1975 are taken into account but may not represent the reality

We decided to make another graph. This time, we took the data from 1975 to 2050.

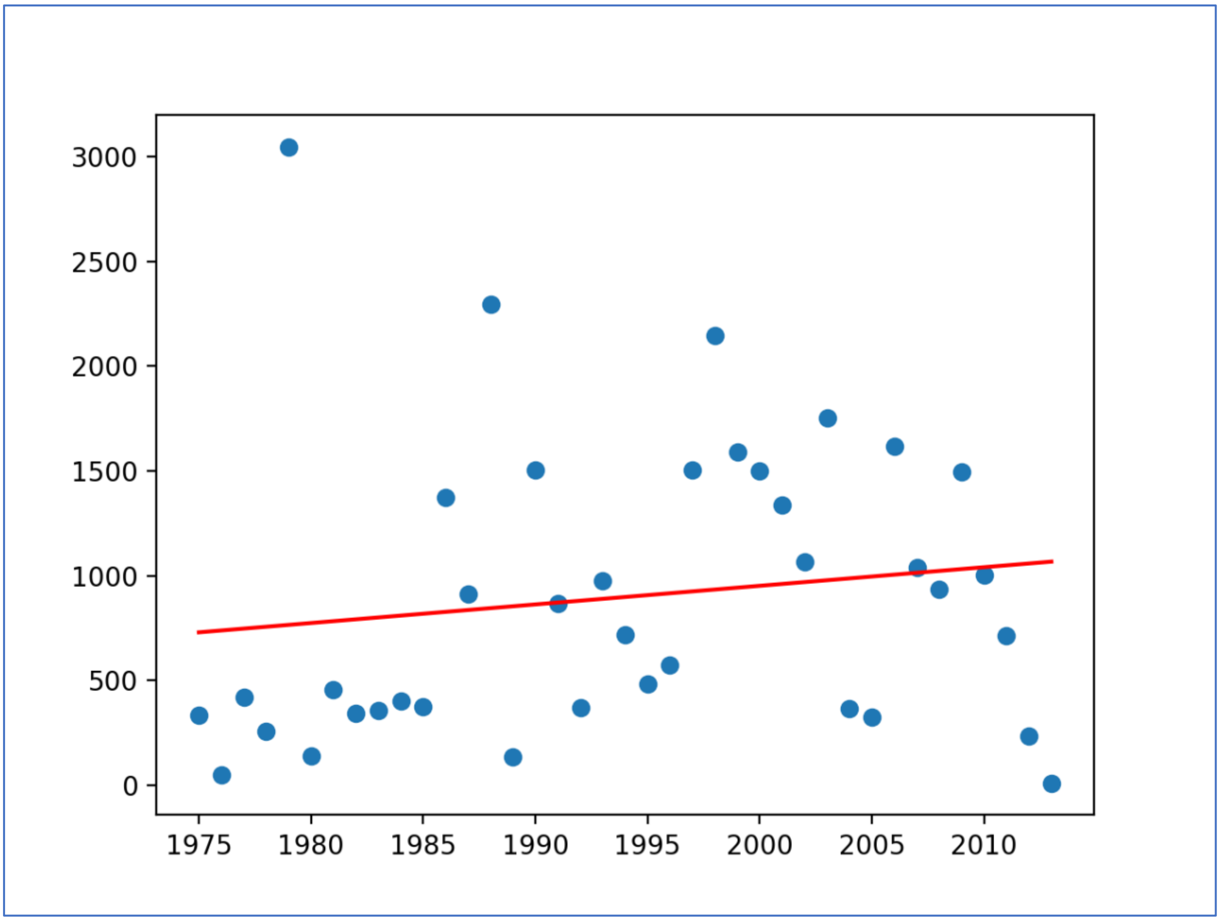


Figure 4. Distribution of the number of meteorites between 1975 and 2020

We see that most of the points are still very different but certainly better represent the reality. So, we have a completely different linear approximation. Thanks to this linear approximation, we can say that the number of meteorites in 2021 is 1147.

The linear approximation is maybe not the best choice because if we take into account all the meteorites, we see clearly that the law y=ax+b does not correspond to the behavior of the number of meteorites per year. this is probably due to a lack of data. However, when we take the data between 1975 and 2050, we notice that even if the linear approximation is closer, we still have a non-negligible lack of precision. Maybe with more years of measurement the linear aproximation will be more reliable. we also notice that in this case the law y=ax+b and it is very close to a constant law. It would seem logical that the number of meteorites seen each year is more or less constant and follows a uniform law

3.

For this question we need to create a graph of the spatial repartition of the landing of meteorites in Oman a middle east country.

Just to have some basic visualization here is a map showing where Oman is on google maps.

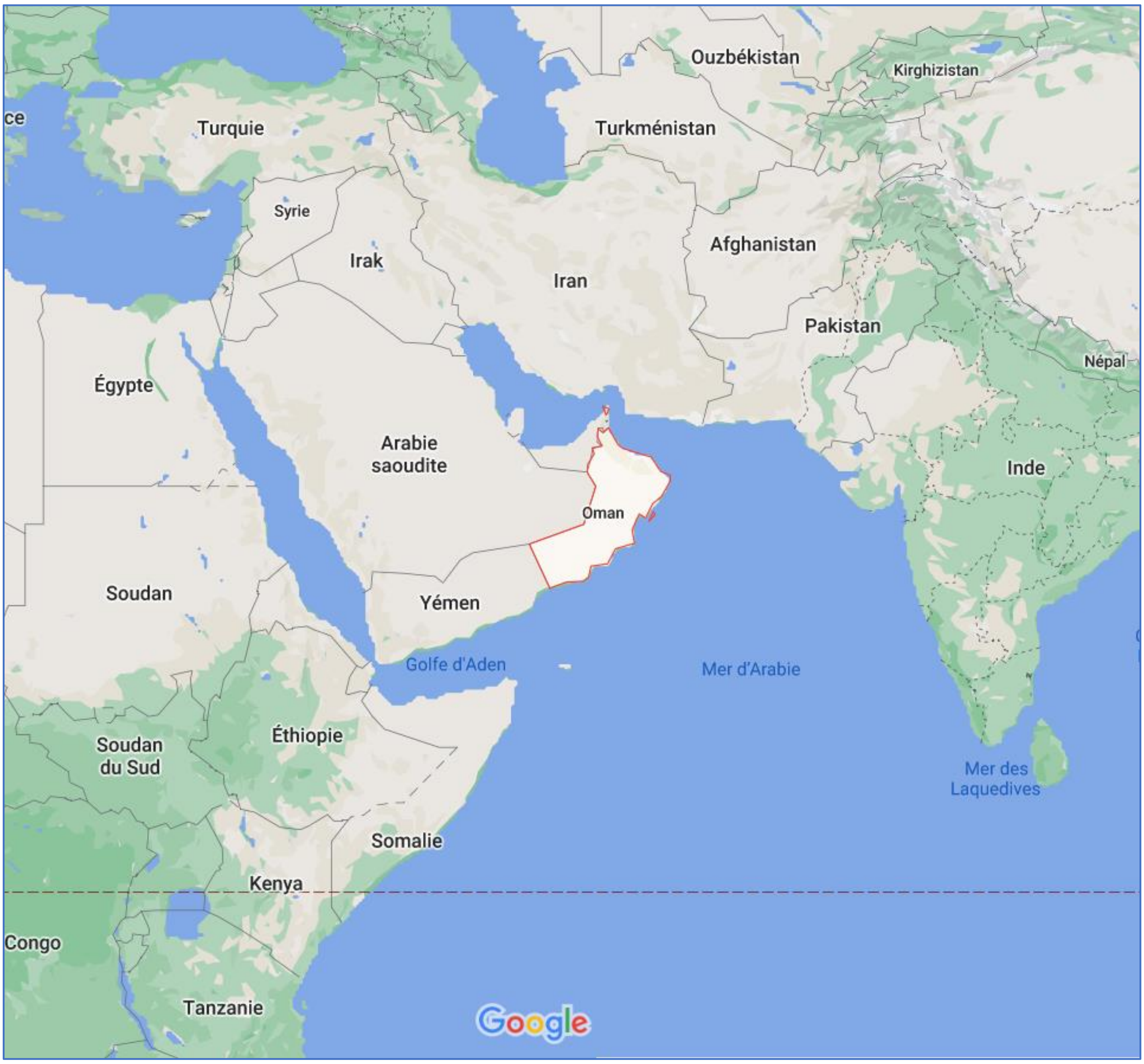


Figure 5. Localization of Oman on Google Maps

We import into our program the shape of the country with a geo panda library, as well as all important geographical points. Then we take the meteorite landing geopoints and choose only those that are inside the virtual country of Oman. And then we plot everything. So we see this graph below the virtual shape of Oman with all the meteorites and their landing points inside.

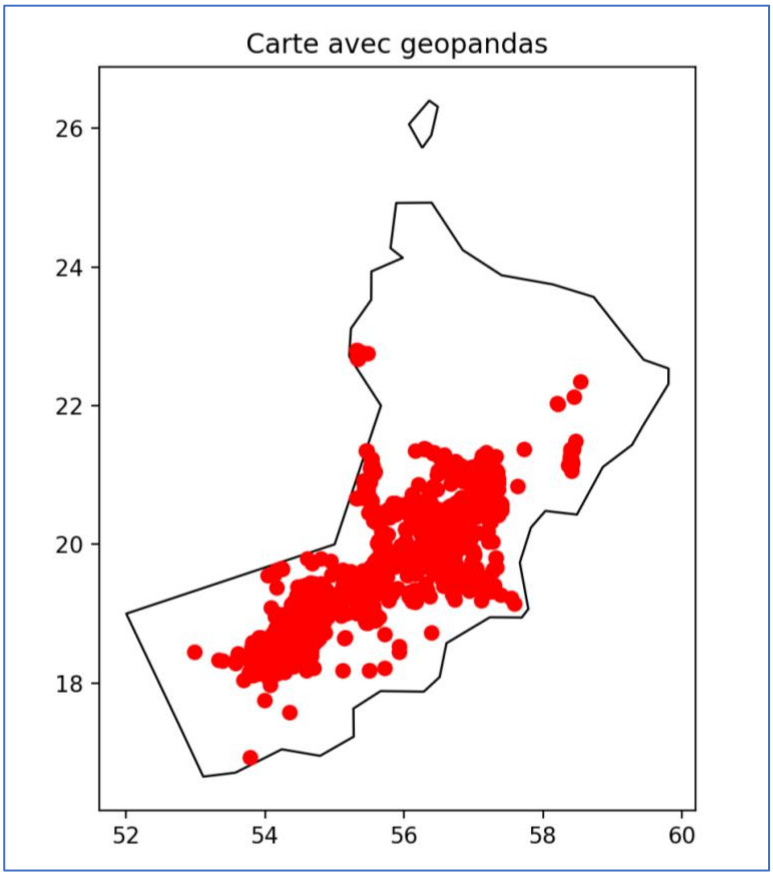


Figure 6. Oman's virtual map with Panda (python) with all the meteorites landing points.

4.

For this one, we have to propose a distribution of the meteorite landing in Oman. We can see on the graph the distribution of density of meteorites

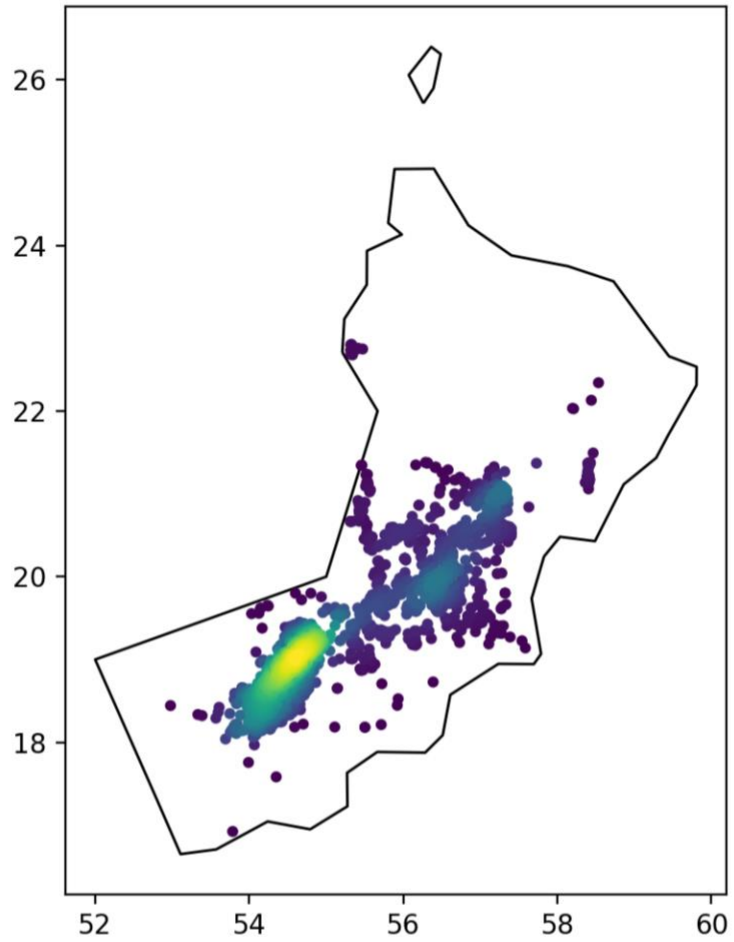


Figure 7. Color wave graph showing the repartition of the localization where landed meteorites in Oman.

To be sure our hypothesis is correct we also tried to build a graph about the density from the real GPS position of each meteorites. We divided this density into one graph on X and the other on Y. So, at the end we got two graphs, one for each axis.

Figure 8. Density repartition on Y of the position of the landing of meteorites in Oman

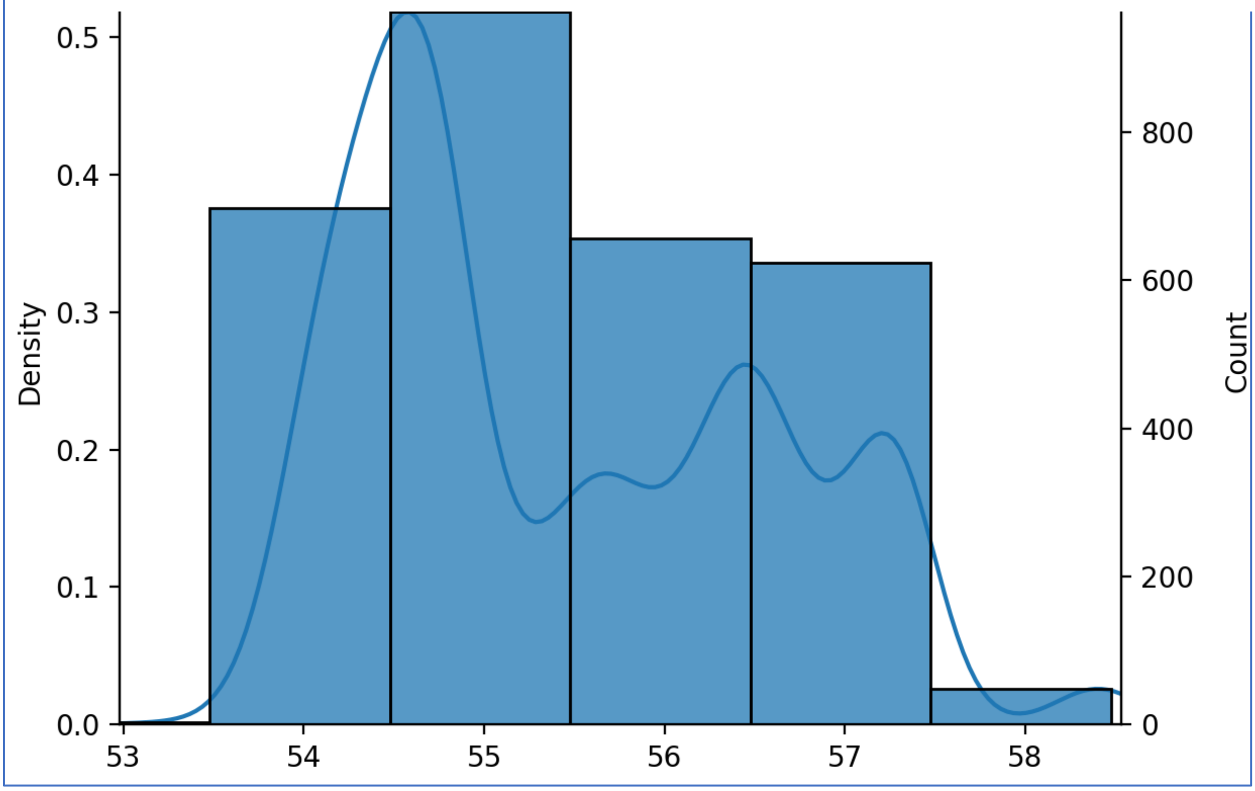


Figure9. Density repartition on X of the position of the landing of meteorites in Oman

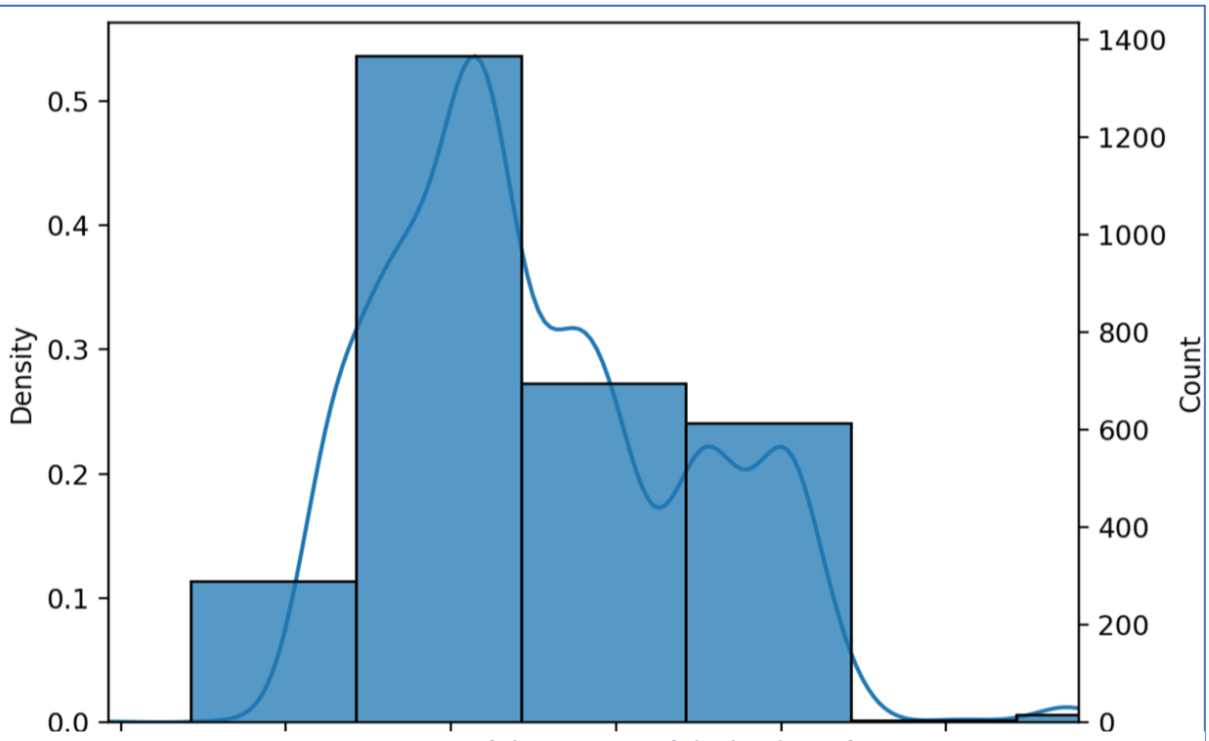


Figure 9. Density repartition on Y of the position of the landing of meteorites in Oman

AAs we can see from the graphs, we can really see that the Gaussian approximation seems to make sense. Ignoring the last points we can approximate it with a normal distribution.

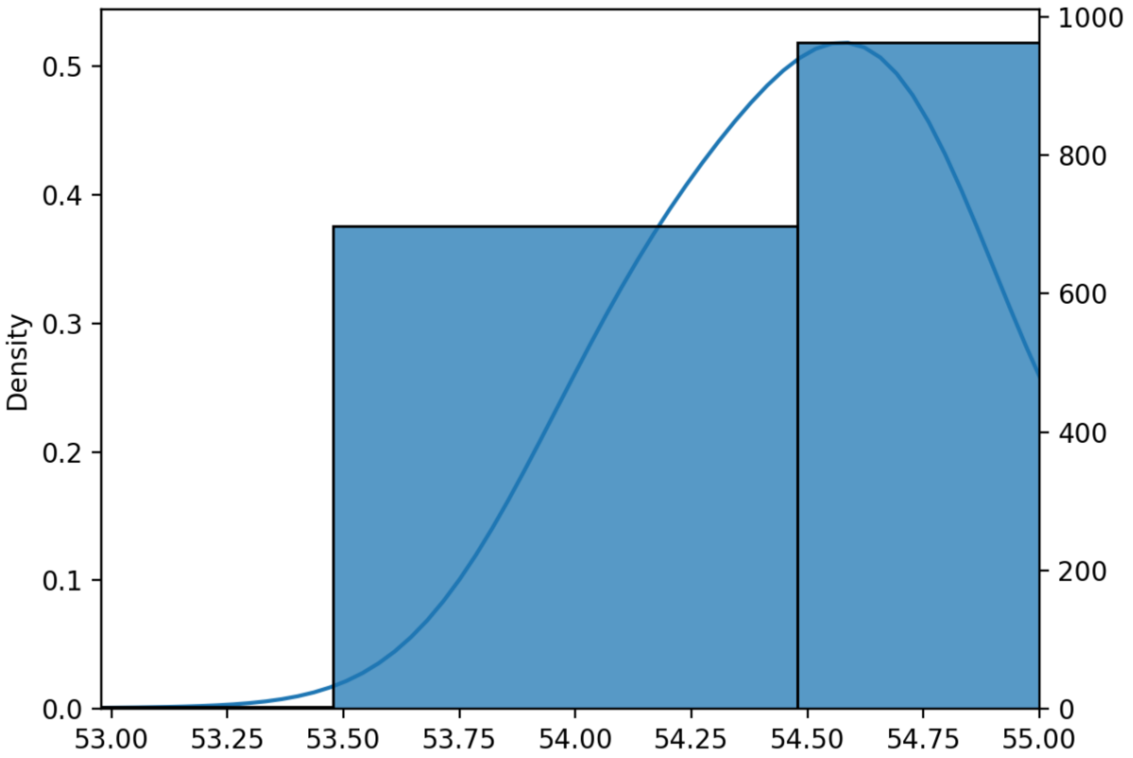


Figure 11. Density repartition on X of the position of the landing of meteorites in Oman after rejection

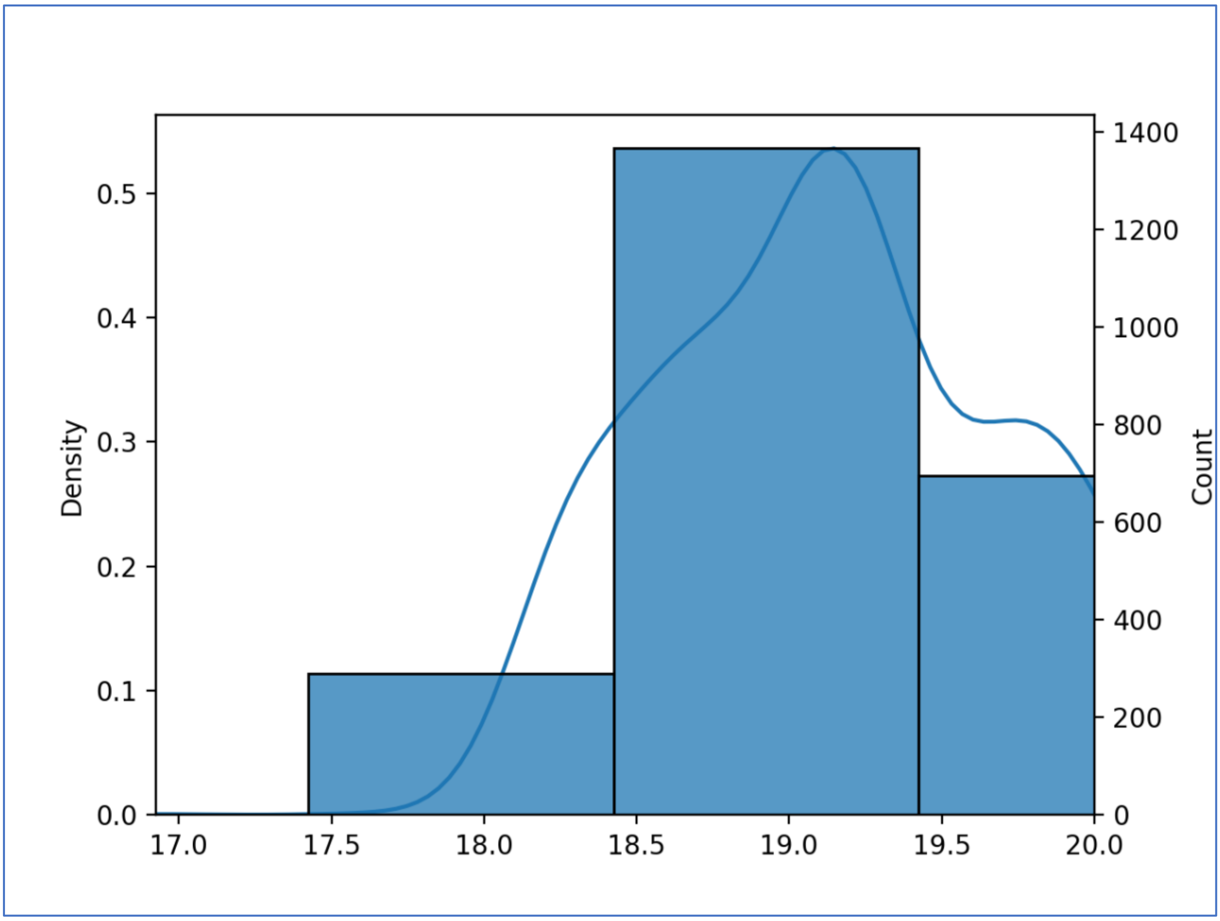
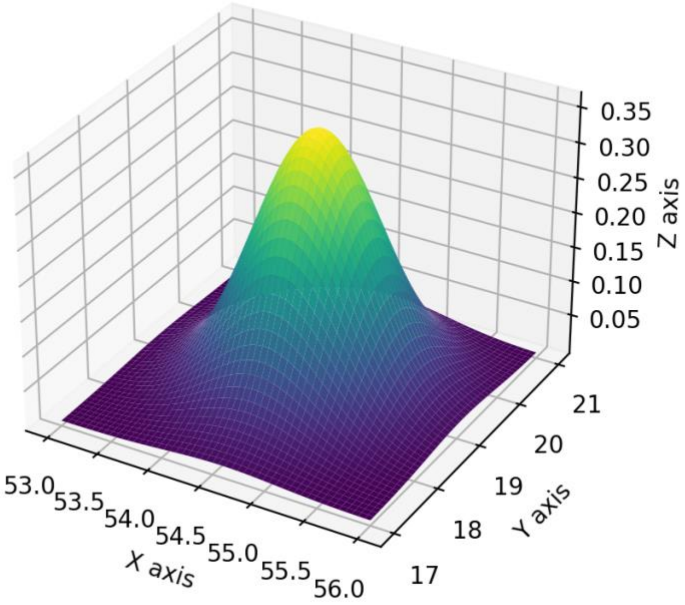


Figure 12. Density repartition on Y of the position of the landing of meteorites in Oman after rejection

With those new graphs we definitively see that the gaussian approach can be a good solution. Now we could use them to determine new landings as asked in the next question. we can also make a 2D graph to see the distribution



5.

Now we choose a gaussian distribution we need for the last question to calculate the probability of having a meteorite landing on a specific circle. So, first we need to determine the parameters of our Gaussian. We want to get those to determine a normal law which then can be transformed to a reduced normal centered law. Which this one can gives us information about probability.

Using the graphs in the previous section we can determine the two parameters we need. For μ we got (54,58;0,64) respectively for X and Y. And for σ we got (0,3;1,66). With those parameters we can make again another graph this time following perfectly our gaussian law.

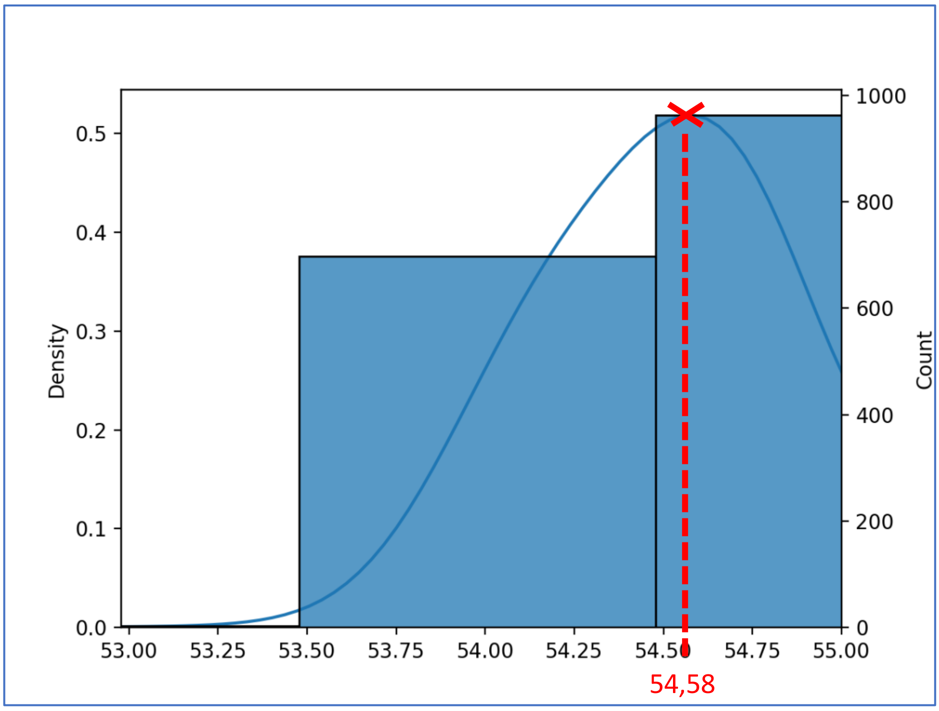


Figure 14. Visualization of the choice of μ for X

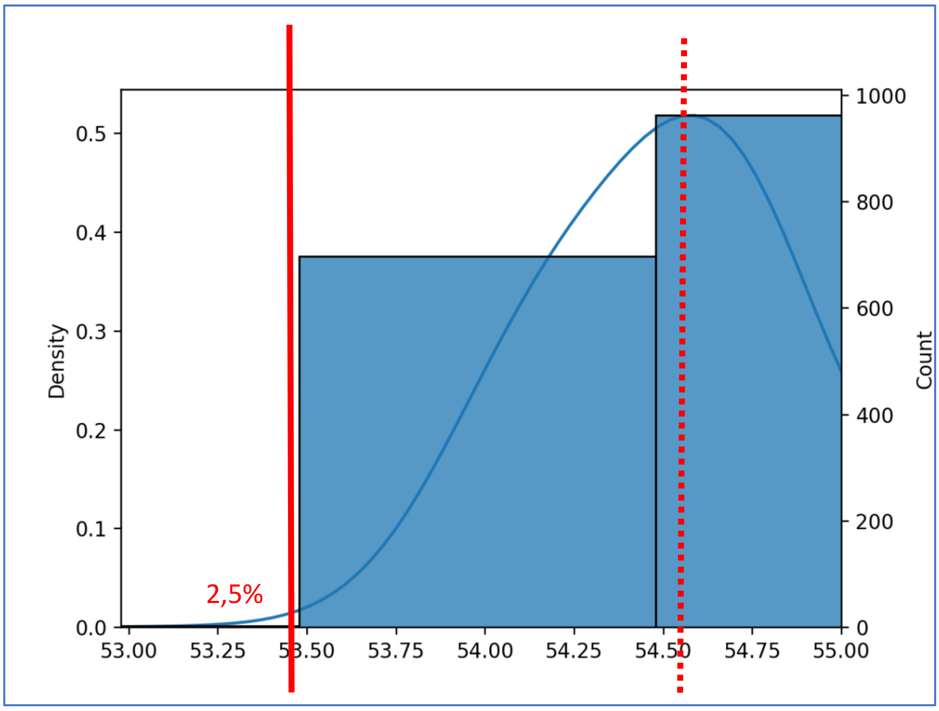


Figure 13. Visualization of the choice to calculate σ for X

Now we use a website to draw the last graphs with the perfect's normal laws.

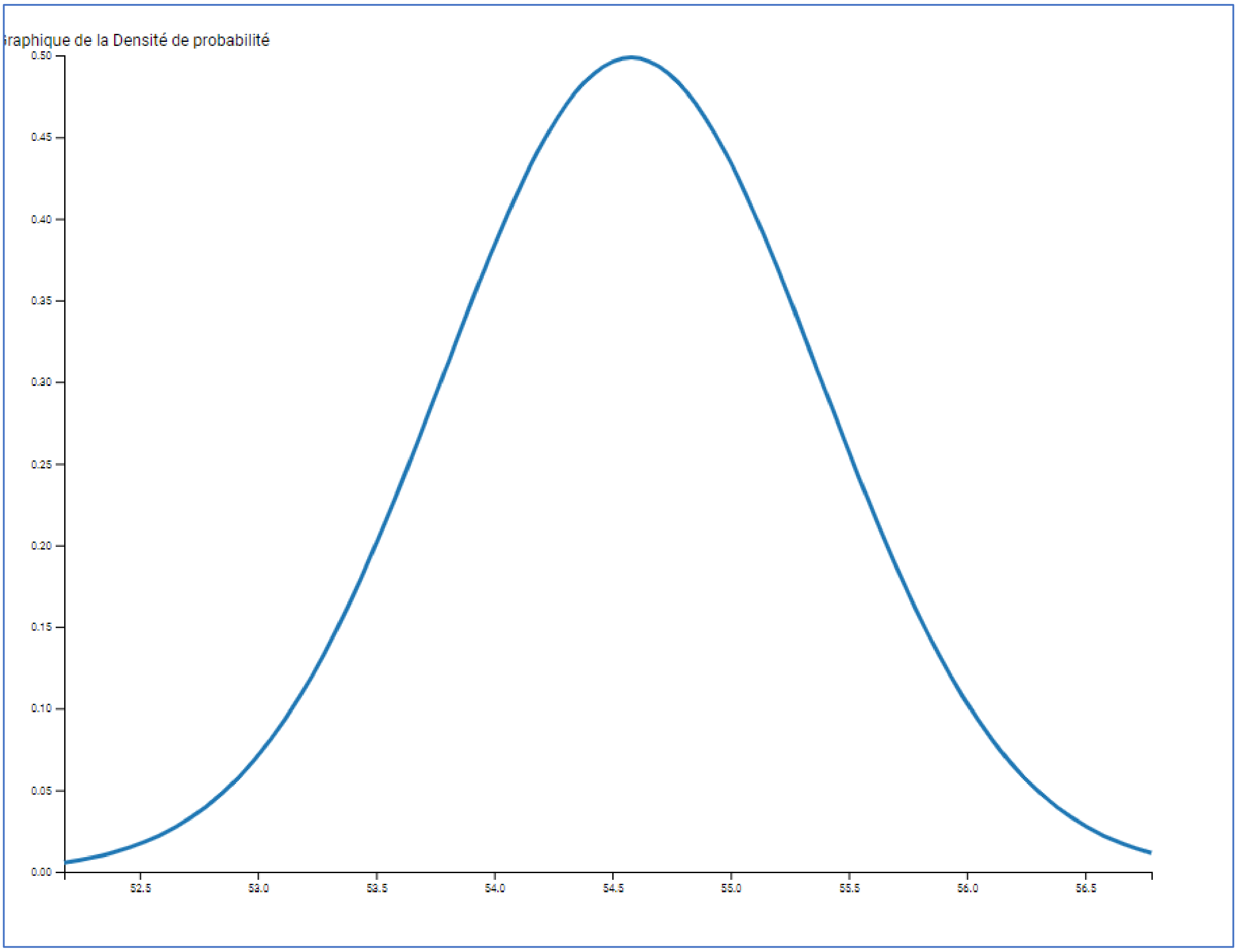


Figure 15. Perfect representation of our approximation for X

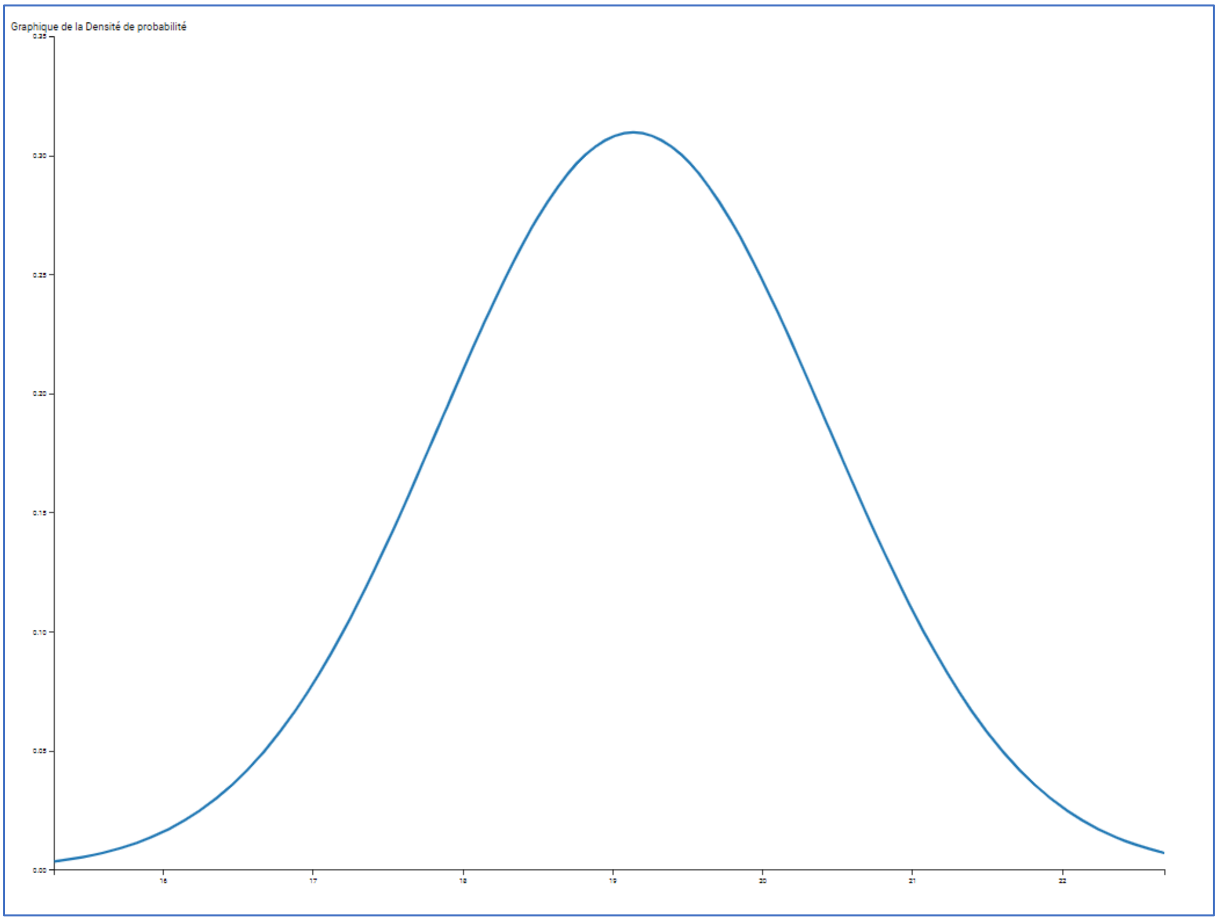


Figure 16. Perfect representation of our approximation for Y

Now we got this me need to determine what is the probability of having a meteorites landing inside the virtual circle determined in the subject.

The problem is that we need to get the probability of having at the same time X and Y respected so we need to group all the previous graphs.

For this we know that X and Y are independent so we can calculate the probability of X and then Y. First we need to reduce and center the normal distribution. We do the same for Y and we obtain respectively P(X)=0,658

and P(Y)=0.5543. Now for the square so P(X&Y) as there are independent it is just P(X)*P(Y). So: 0.3647.

Let's try to find out how to calculate this probability of having a circle instead of a square.

For a circle the probability is less than 36,47%.