Week 5 – Continuous Probabilities

For this portfolio entry I decided to use the topic that was taught in the fifth week of this module; Continuous Probabilities. For this assignment I will use lecture notes and the code base from the lecture where we measured the height of *Pokemon*. As length is a great way of exploring continuous probabilities, I will continue use a similar dataset to the one we used in the lecture and explore data on the height of NBA Basketballers. In this portfolio entry I will attempt to fit different Probability density functions on the data in the hope of finding the best fit.

The first actions I took during this task were the common rudimentary tasks that are essential to every data analysis exercise. This is cleaning and processing the data, after loading the data to ensure the data was clean, I erased the empty fields in the height column.



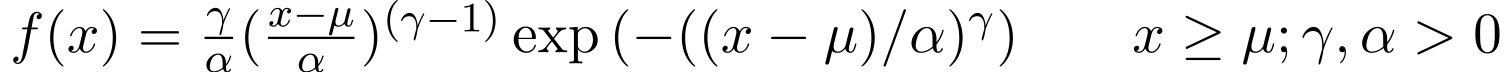
Once the data was ready the first exercise was to visualise the data, I decided to use a histogram over a bar chart for this due to abundance of data.

A graph with blue and white lines

Description automatically generated

From this visualisation we can get a rough idea of the shape of our data, we can see the majority of NBA players are around 2.04-08m mark. We can also see there is a single peak meaning our data is unimodal.

During the lab section of the class there were a number of different distributions that we did not look at. One of the notable distributions was weibull\_min. ‘Weibull plot is a graphical technique to determining if the dataset comes from a population that is logically fit by a 2-parameter Weibull distribution.’



The Weibull plot is used to commonly find the best estimate for the shape parameter of a 2-parameter Weibull distribution, this falls into line with our search to fit a distribution.

A graph with a red line

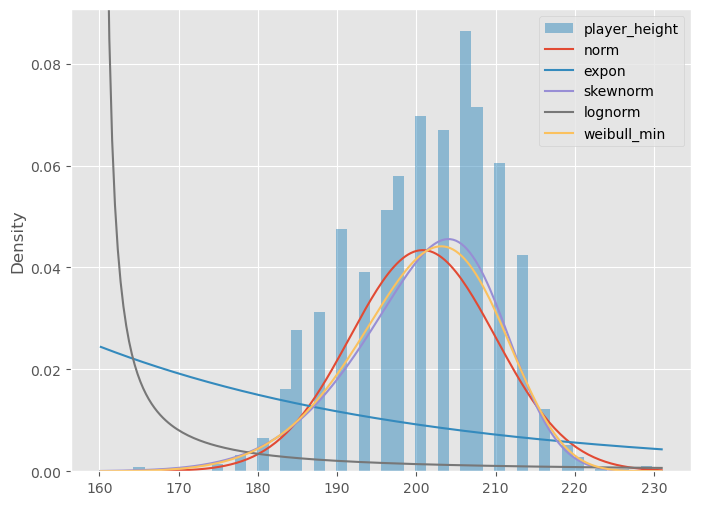
Description automatically generated

We can see from our analysis that the skewnorm approach provides us with the following results.

A graph of a graph with a red line

Description automatically generated

As you can see the results for the skewnorm example far outweigh the other examples and we can see a much better density distribution.



Skewnorm can be efined as ‘maximum likelihood estimation of the shape parameter of a univariate skew-normal distribution’

skewness tells you the amount and direction of skew(departure from horizontal symmetry) If we look at our distributions shape we can digest that its unimodal as it only has one peak. We can also see that its skewed to the right meaning that its positiviely skewed.

To further interpret the level of skewness of our data we can rely on Bulmer’s explanation where he suggests that

If skewness is less than −1 or greater than +1, the distribution can be called highly skewed.

If skewness is between −1 and −½ or between +½ and +1, the distribution can be called moderately skewed.

If skewness is between −½ and +½, the distribution can be called approximately symmetric.

From this analysis it can be concluded that with the breath of data the best approach to fit a Probability density function is to use the Skewnorm approach as it distributes the data more evenly than any of the other approaches.

References

<https://discovery.cs.illinois.edu/learn/Exploratory-Data-Analysis/Histograms/>

<https://www.spotfire.com/glossary/what-is-a-histogram-chart#:~:text=A%20histogram%20is%20a%20graph,Karl%20Pearson%2C%20an%20English%20mathematician>.

<https://brownmath.com/swt/sources.htm#so_Bulmer1979>

<https://books.google.co.uk/books/about/Principles_of_Statistics.html?id=dh24EaSrmBkC&redir_esc=y>

<https://www.geeksforgeeks.org/weibull-plot/>

A graph with a red line

Description automatically generated