Investigation of b quark production and meson/baryon hadronisation ratio at the LHC.

Nat Karmios   Carlos Barredo   Lucas George   Nima Sabha
Supervised by Alexander Richards

2017

Abstract

The particle accelerator at CERN is used to record data on the number of baryons and mesons that are produced by collisions between a range of naturally occurring particles; this data has been used to develop Monte-Carlo simulators, such as Pythia, which allows one to generate collision data with one's own parameters, varying the center of mass energy and filtering out irrelevant particles.

This research is centered in the field of quantum chromodynamics (QCD), specifically concerning the ratio between the number of b-mesons and b-baryons produced when a heavy quark hadronises as the centre of mass energy varies, as well as how the pseudorapidity of such particles varies with charge energy. Pythia has been used to carry out these simulations.

1 Introduction

Quantum chromodynamics (QCD) is the modern theory of particle interactions involving the strong nuclear force; it was first proposed by David Politzer, Frank Wilczek and David Gross in the early 1970s. They were awarded with the Nobel Prize in Physics in 2004 as a result of their work on this field. In this project, we will use the theory of QCD to describe the interactions between two heavy quarks in a high-energy particle accelerator during their hadronisation process.

Before describing QCD, we must clarify some concepts which will be dealt with throughout the paper. Firstly, a hadron is a composite particle made from quarks which are held toghether by the strong nuclear force. Within hadrons, we encounter two main sub-categories: baryons and mesons. Baryons are formed by three quarks whereas mesons are formed by a quark and an anti-quark. Examples of baryons include protons and neutrons whereas mesons include pions and kaons.

Another idea we must introduce, is that of a Monte-Carlo data simulator. The Monte-Carlo method is a computational algorithm which utilises random sampling as a means of creating numerical data simulations of a chosen process. We will be using this technique within the software of Pythia in order to generate the data needed for the investigation. This method was chosen since the nature of obtaining the data needed experimentally is way beyond the financial allowances of our project. Using the Monte-Carlo method however, allows us to simulate thousands of events and obtain results almost immediately, which has been way more convenient considering the time limitations of the project.

The last thing to be aware of, is what we mean by the center of mass frame, which we will deal with in 3.1. A center of momentum frame, when referring to a system, is the unique inertial frame of reference in which the sum of the linear momenta of all the particles within the system disappears. The center of momentum frame refers to the relative velocities and momenta of a chosen number of particles, rather than defining a location. The center of mass frame is, consequently, an inertial frame in which the center of mass of a particle remains at the origin. This however, does not necessarily mean that the center of mass is at the origin of the coordinate axes, but it does imply that the center of mass of a given particle is always at rest.

The theory of QCD describes quarks as being held together by gluons which are particles that act as mediators of strong interactions between these, as well as taking part in the strong interactions themselves. Gluons are the force carriers for this theory just like photons are for the electromagnetic force in quantum electrodynamics (QED). The force carried by gluons, does not weaken over distance, as opposed to gravity for example, but instead remains uniform, on the order of several thousand Newtons. This therefore implies that at no point do any quarks separate from one another, meaning that they can only be observed on a hadron level. Quarks are said to have three different 'colours' (red green and blue) and six different 'flavours' (up, down, charm, strange, top and bottom). These terms are merely analogous, referring to both the charge of the quark as well as describing the species of elementary particle the quark is. Quarks can be classified into two further categories: heavy and light. Heavy quarks include charm, bottom and top quarks whereas the up, down and strange quarks are said to be light. Heavy quarks are 'heavy' because their masses are large on the typical QCD scale regarding the gauge coupling parameter, Λ. This denotes the strength of the force exerted by a particle during an interaction, measured in MeV. $\Lambda_{MS} = 217^{+25}_{-23}$ MeV[1].

[1] https://www.phenix.bnl.gov/WWW/publish/xiewei/RBRC_Workshop_Dec/heavyworkshop/Dima.pdf

2 Method

2.1 Program Development

2.1.1 Setting up Pythia

When initially researching Pythia, we found that it was possible to compile it into a Python library[2] - since Python is a much faster language to develop in, and some of us have past experience with it, this was much more ideal for our project compared to using the original C++ version. However, it was evident that this was intended for use on a Linux system, whereas we use Windows systems; for initial tests, we had Pythia compiled on a remote server, however this made the development process painstakingly slow.

2.1.2 Docker containerization

After weighing some options, we decided to use Docker to run our simulations. The reasons for this were twofold: Docker allowed for the safe and contained management of dependencies, both for the operating system and for Python, and this also meant that what was initially a Unix-only program can now run freely on any platform supporting Docker. We have made a base image containing the Pythia library, which is available on Dockerhub[3].

2.1.3 The Pythia Application Programming Interface (API)

Looking into Pythia's manual, we found certain options in the simulator's configuration that are useful to our investigation: "quiet mode" prevented automatic printing to the console, allowing us to just display useful information; we could configure the charge energy, so we could see the effect of collision energy on the output data; there was also certain options - `HardQCD:all` and `HardQCD:hardbbar` - that disabled the production of all particles other than b mesons and baryons, preventing the simulation of irrelivant particles and thus speeding up the generation of data.

[2] http://home.thep.lu.se/ torbjorn/pythia82html/PythonInterface.html
[3] https://hub.docker.com/r/natk/python-pythia/

2.1.4 A complete program

The basic program generates a set of data from Pythia, writing to a file at each step. In order to make this program more accessible to others, we decided to extend our program's functionality. Firstly, we added a step where, once data is generated, the resulting file is uploaded to GitHub Gists, where it is stored in the cloud and can be easily shared via a link. Secondly, we wrapped our program in a Flask[4] server. The program currently uses BASIC authentication with a password, and serves JSON objects at the following endpoints:

- `[GET] /` (root) - Just serves `{"success": true, "message": "Hello there!"}`; this is just a ping message to test that the server is alive.
- `[GET] /status` - Serves a status message, giving information about the state of each data generator instance.
- `[POST] /login` - Generates an auth token that can be used on secure methods. This takes the password that is defined in `cfg.py`.
- `[POST] /test_auth` - Tests authentication; this is used to check if an auth token is still valid, in case the server has restarted, in which case all tokens would be reset.

  (Requires auth token)
- `[POST] /add` - Adds a generator; takes a JSON object with some variables, such as the energy level and number of collisions for histogram data. It responds with some information about the newly-created generator, such as the process ID (PID).

  (Requires auth token)
- `[POST] /delete` - Deletes a generator, canceling it if it is currently running. Responds with a success message and the deleted generator's PID.

  (Requires auth token)

Note: methods that require an authentication token use BASIC authentication, with a blank username and the auth token as the password.

We added Gunicorn[5] to the server's Docker container for some extra security and stability, and then used an Apache server as a public access point. To use this API, we have developed an interface page using ReactJS; we have hosted it through GitHub Pages[6].

[4] Flask is a lightweight HTTP microframework for Python; http://flask.pocoo.org/
[5] Gunicorn *'Green Unicorn'* is a Python WSGI HTTP Server for UNIX; http://gunicorn.org/
[6] https://natkarmios.github.io/ParticleSimulations/ - Our server is hosted at particles.api.karmios.com

3 Analysis

3.1 Variables and Pseudorapidity

When generating this data, there were various parameters to be considered before further ones were introduced. Firstly, we looked at the number of baryons and mesons produced per collision, and the center of mass energy of the collision itself, which ranged between 7.0 to 14.0 TeV. We were then introduced to another variable to take into account; the pseudorapidity of the baryons and mesons produced. We will refer to the pseudorapidity as the greek letter η (eta) for the rest of the paper. To determine η, we will first need to measure the angle that the particle produced from the collision has relative to the beam axis (the z-axis). Once we know this angle, theta, we can then use the formula

$$\eta \equiv -\ln\left[\tan\left(\frac{\theta}{2}\right)\right] \tag{1}$$

to determine the value of η. As we see from said formula, when θ is $\frac{\pi}{2}$ (so that the particle emitted is perpendicular to the beam axis), the value of η will be 0. Inversely, as θ gets closer and closer to zero, η tends towards infinity. In this project, we will deal with particles of pseudorapidity in the range of $|\eta| < 4.8$, which accounts for 95-98% of particles emitted from these types of collisions.

Pseudorapidity is particularly useful in hadron colliders such as the LHC, where the composite nature of the colliding protons means that interactions rarely have their center of mass frame coincident with the detector rest frame, and the complexity of the physics means that η is far quicker and easier to estimate than the normal, perpendicular, rapidity of the particles emitted.[7]

For our plots we initially thought of Excel as an appropriate software to generate the scatter diagrams and histograms required. This turned out not to be the case because a proprietary software such as Excel is not properly equipped to handle such large volumes of data that we are using. Considering Monte-Carlo data generators could simulate up to 10000 collisions, we decided to leave behind the idea of Excel and therefore we decided to make use of R, a statistical analysis programming language to interpret our data. The use of R has given us a greater degree of control over the presentation of the data with less time wasted, due to the simple and highly configurable command of the system as opposed to a graphical user interface(GUI).

3.2 Setting up and working with R

The process of learning how to use R was relatively straightforward once the `.csv` file was obtained from the online interface. Even though none of us had any past experience with this type of software, finding the commands required in online instruction pages allowed us to create the plots we needed.

Generally, we followed this process:

- Set the working directory to the location of the `.csv` file
- Import the `.csv` file with `data = read.csv()`
- Define a variable for the set of data for each variable (e.g. `energy=data$Energy`)
- Make plots of the desired data, for example:

  Use `hist()` function to plot a histogram of desired data set (e.g. `hist(numberOfBaryons)`)

  Use `plot()` to plot a scatter graph for two sets of data (e.g. `plot(energyLevel, numberofBaryons)`)

Plotting the scatter plots proved to be more challenging than obtaining the histograms showing the pseudorapidity distribution. We wanted to obtain information such as the value of the product moment correlation coefficient, as well as the line of regression for these scatter plots. Hence, the installation of packages within the software was necessary.

[7] w.hep.shef.ac.uk/edaw/PHY206/Site/2012_course_files/phy206rlec7.pdf; Rapidity and Pseudoratpidity lecture by E. Daw

To address these needs we used the following commands:

- `aggregate()` - finds the sum of b-baryons and b-mesons for each increase in charge energy. We assigned these values to a final data set using the `cbind()` function.
- `ggpoint()` (from `gglot2`) - allows creation of a scatter plot with two y variables, assigning an individual colour to each variable and labelling the axes appropriately using the `'color'` parameter and `labs()` function respectively.
- `geom_smooth()` - allows plotting of a least squares regression line with single y values.
- `cor()` - finds the product moment correlation coefficient using values in our final data set. For example, `cor(energy, baryons)` with our previously assigned variables finds the PMCC for our baryons plot.
-
```
ggplot(datfinal, aes(x=energy)) +
geom_point(aes(y=baryons), color="red") +
geom_point(aes(y=mesons), color="green") +
labs(x="Centre of Mass Energy / TeV", y="Number of Particles")
```
  This colour-codes the hadrons and finalises the merged scatter plot, with both hadrons against the center of mass energy.

3.3 Plots and Analysis

3.3.1 Histograms

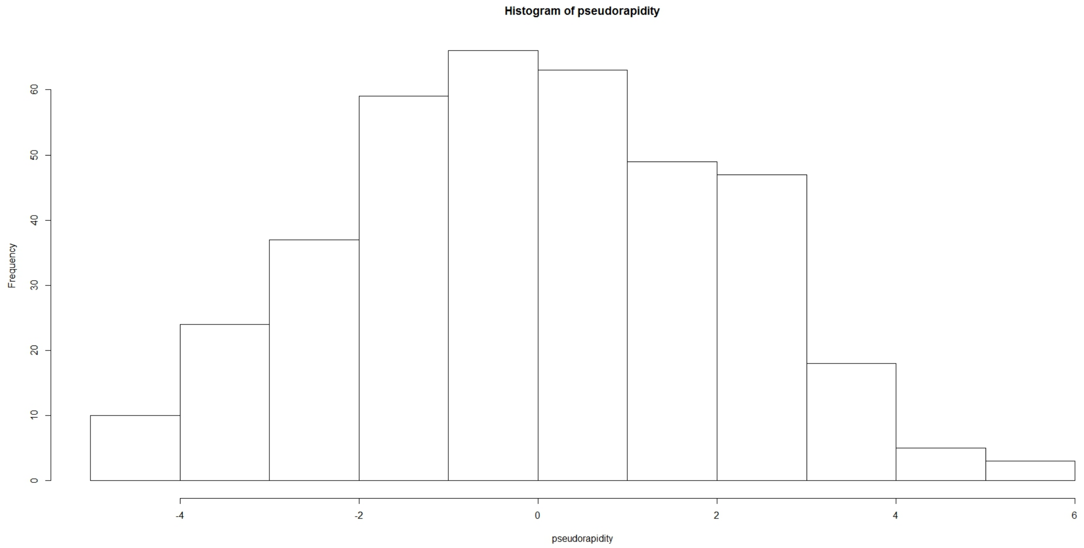

Figure 1: Histogram of the Pseudorapidity distribution of Baryons.

Figure 1 shows that the distribution of pseudorapidities of baryons produced in a proton-proton (pp) collision at 12.0 TeV. We immediately notice that the data distribution resembles that of a slightly skewed normal distribution. A possible reason for the skew of the distribution lies in the fact that the highest frequency density is centered around values of η very close to 0, in other words, the baryons originated from the collision are emitted at angles almost perpendicular to the beam axis. Furthermore, we see a higher than expected value for the frequency density and hence the number of baryons (frequency) at a value of $2 < \eta < 3$, which puts to question the validity of the Monte-Carlo data simulator. Of course this is assuming that the data obtained in the real world followed a similar skewed normal distribution, which of course doesn't have to be the case. We haven't been able to find published data to compare these plots to, so we can't confirm the reliability of these.

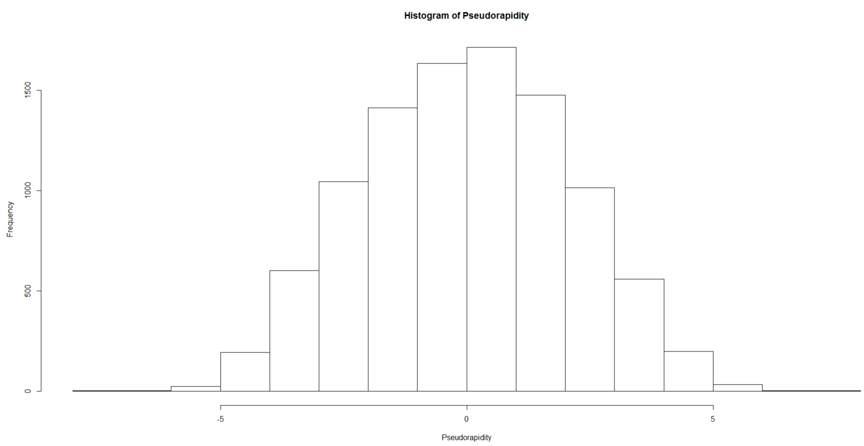

Figure 2: Histogram of the Pseudorapidity Distribution distribution of Mesons.

The histogram in Figure 2 shows the distribution of pseudorapidities of mesons produced in a collision at $12.0 TeV$. From the plot, we can clearly see that the distribution of mesons follows a normal distribution pattern, which is expected since, as described in 3.1, 95% to 98% of hadrons lie within a value of $|\eta| < 4.8$. We see that very few mesons exceed a value of $|\eta| > 5$ meaning that, very rarely is a meson emitted at an angle of 1.013 degrees or more relative to the beam axis. It is true that as theta tends to 0, the value of η tends towards infinity however this increase is extremely slow as described above.

From these plots, it is concluded that the pseudorapidity distributions of produced charged hadrons- baryons and mesons in this case- from inelastic proton (pp) collisions fairly resemble that of Gaussian distribution curve. It is true, that the histogram in Figure 1 shows a slightly skewed normal distribution however for analytical purposes at this stage, we shall refer to both of them as Gaussian. The implications of these distributions are beyond the breadth of this paper, however we have confirmed that these distributions exist in the production of charged hadrons during proton-proton collisions.
We provide a Gaussian distribution plot in the appendix for further reference in terms of comparing our experimental data with this fit so that the slight skew can be noticed more clearly.

3.3.2 Scatter Diagrams

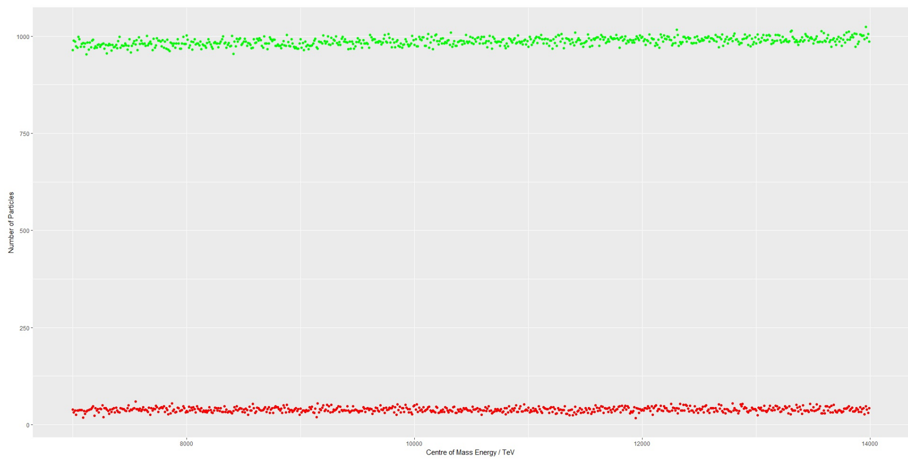

Figure 3: Scatter plot showing number of mesons and baryons produced at different center of mass energies.

The plot above shows the difference between the number of baryons and mesons produced across different center of mass energies ranging from 7.0 to 14.0 TeV. The baryons are shown in red whereas mesons are shown in green. Due to an average difference in the y axis of 950 between the number of b-mesons produced and the number of b-baryons produced, the scale on the graph above is massive compared to the variation of our data for the two hadrons. Therefore, it is difficult to immediately notice any correlation between the number of each hadron produced and the increasing center of mass energy. However, we can say that neither of these datasets are random as the variation in the difference between them is insignificant in comparison to the average size of the gap in the number of particles produced. Conclusively, we know that there must exist a relationship between them.

Figure 4: Scatter diagram of number of b mesons produced across a range of 7.0 to 14.0 TeV

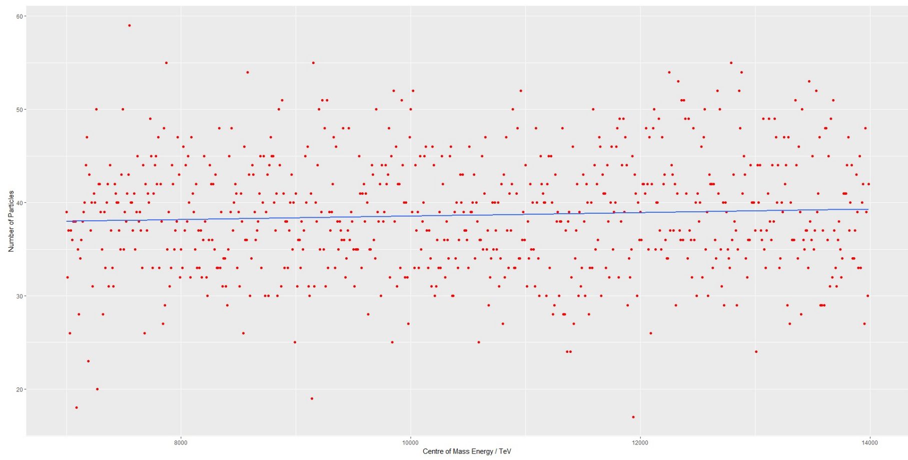

Figure 5: Scatter diagram of number of b baryons produced across a range of 7.0 to 14.0 TeV

These two scatter plots show each group of data individually, number of mesons and baryons respectively against the center of mass energy. The scales of the axes have been modified in such a way so that more accurate interpretations of the data can be made. Using these plots as opposed to the merged one, we were able to plot two lines of regression that reveal the correlation between the number of the type of hadron being produced, depending on the center of mass energy of the collision the hadron is produced at. From the regression lines of each plot, we have then calculated the value for the product moment correlation coefficient, defined as $\frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}}$ for each hadron and hence, we have been able to determine a correlation between these and the center of mass energy.

The plot in Figure 4 has a value of $r = 0.4745392$ which means that there is a fairly strong positive correlation between the center of mass energy and the number of mesons produced and hence these variables are correlated at 47.45%. This justifies the fact that more mesons are produced for higher center of mass energy values. In contrast, the plot in Figure 5 gives a value of $r = 0.0599828$, which shows that there is a negligible correlation between the center of mass energy and the number of baryons produced. Assuming that this number would continue to be constant, we predict that the increasing trend of the number of b-mesons produced would plateau at a certain energy level, for the mass available for the production of the b-mesons is limited as long as the number of b-baryons produced is constant.

4 Conclusion

Ideas regarding the relatively constant number of baryons for different center of mass energies may arise, however we are not able to give a conclusive statement for this relationship. It is true we can say this is the case for values between 7.0 and 14.0 TeV, however for values outside this range, further investigation would be required in order to fully confirm the trend outlined in 3.3.1 .

From the plots in Figures 3, 4 and 5, we have calculated the ratio for the number of b-mesons to the number of b-baryons produced to be directly proportional to the center of mass energy of the collision. There exists an individual ratio for each collision which is the reason why we have calculated the average number of baryons and mesons produced respectively. The mean number of b-baryons produced is 381.57 and the mean number of b-mesons produced is 9869.14. From this, we derive the average ratio of b-mesons to b-baryons produced to be 25.86 : 1.

The implications of this numerical value obtained are yet to be discussed, however these go beyond the objective of this paper. All the calculations and simulations utilized to arrive to this ratio, assume that our current understanding of QCD is correct or accurate enough so that this value holds under certain modifications to the theory.

5 Source Reliability and Acknowledgements

Prior to the composition of this paper and due to the complexity of the topics being dealt with, extensive background research was required in order to make the paper as scientifically accurate as possible. Having said this, we haven't cited any sources directly, however we would like to acknowledge the following documents for clarifying concepts needed for the completion of this paper:

For our background research in the field of QCD, we were suggested by our supervisor to read a range of papers on the subject. These at first seemed rather intimidating due to the obscure nature of the topic. We found however, that some websites had a more basic but perfectly valid explanation of the theory of QCD. This served as an introductory stage to the project and made us all aware of what we would be facing when writing this paper. Our preferred website was, quantum diaries [8] due to the visual aid provided through Feynman diagrams, giving us a greater insight on the nature of the collisions investigated.

D. Kharzeev's slides on heavy quarks and hot quark matter for the Brookheaven National Laboratory in New York[9] concisely described what a heavy quark was, as well as giving us some basic notions about the theory of QCD. Although these slides were produced in 2005, we assumed their reliability since, after further research, the concepts discussed in these matched our research results.

On the other hand, sources such as Cristina Terrevoli's paper on heavy flavour measurements in the ALICE detector at the LHC[10], although reliable, didn't directly help us with any of the aspects discussed throughout this paper and hence, this source was discarded.

One of the most challenging things about this project was perhaps the familiarization with completely new concepts and variables. As outlined in 3.1, understanding concepts such as pseudo-rapidity required external reading and research. E. Daw's 7th lecture for the University of Sheffield on rapidity and pseudorapidity[11] gave us an understandable and clear idea of this concept, which made it easier to write this paper in a more in-depth manner.

[8] http://www.quantumdiaries.org/tag/qcd/

[9] www.phenix.bnl.gov/WWW/publish/xiewei/RBRC_Workshop_Dec/heavyworkshop/Dima.pdf

[10] arxiv.org/pdf/1404.0054.pdf

[11] www.hep.shef.ac.uk/edaw/PHY206/Site/2012_course_files/phy206rlec7.pdf

6 Appendix

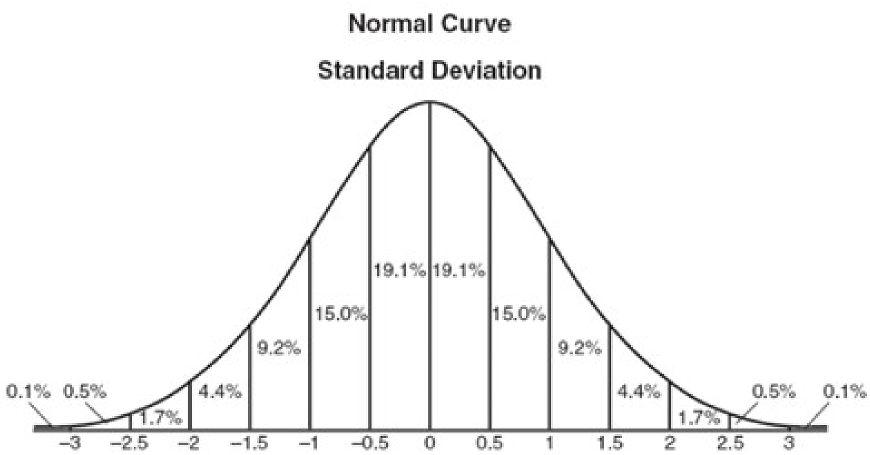

Figure 6: This plot shows a normal distribution curve showing the percentage of data lying within a certain number of standard deviations from the mean.
Source: www.mathplanet.com/Oldsite/media/27934/normal_distribution_500x263.jpg

As we see from the plot, 99.8% of data lies within three standard deviations from the mean, which is concordant with our data since the majority of our data does so too. This supports our analysis in terms of the pseudorapidity values following this distribution as described in 3.3.1.