MATH189 - Final Report

David Justo
A12382001
MATH 189
djusto@ucsd.edu

Hudson Cooper
A12199791
MATH 189
hncooper@ucsd.edu

Joseph Gilby
A14248571
MATH 189
jgilby@ucsd.edu

Vikram Khare
A12693563
MATH 189
vkhare@ucsd.edu

Jonathan Lim
A13018363
MATH 189
jzlim@ucsd.edu

Young Jin Yun
A12083802
MATH 189
yyun@ucsd.edu

1 Introduction

The Large Hadron Collider (LHC) is the world's largest, most powerful, and most famous particle collider. The purpose of a particle collider is to accelerate subatomic particles like protons and neutrons to high enough velocities that when they collide, they fragment into numerous smaller particles to be detected and studied. Often these finer particles are not observed at a high enough frequencies to study outside of a collider. The operation of the LHC has lead to many important advancements in particle physics, including the first detection of the Higgs Boson in 2012, which was a particle predicted to exist under the Standard Model of Particle Physics (developed in the 1960s). The primary organization behind the LHC is CERN, The European Organization for Nuclear Research, and it is located in a northwest suburb of Geneva. A growing concern for CERN is the performance of the algorithms which sift and process the data generated by the LHC. In hopes of dealing with this issue, they have been looking towards machine to play a larger role in their pipeline. For the task of reverse engineering the paths of particles based on detector data, the Conference and Workshop on Neural Information Processing Systems (NIPS), IEEE World Congress on Computational Intelligence (WCCI), and CERN have partnered with Kaggle to present the TrackML Challenge. As described in the competition proposal, increases in "readout rates" and the total number of subatomic particle collisions, after upgrades to the LHC, will lead to an "explosion in combinatorial complexity." A fixed budget in addition to these problems necessitates new methods in delineate the particle trajectories. It is thus the hope of these organizations that a crowdsourced solution can be found through the Kaggle competition. Although the competition problem statement is enticing, the provided data set is complicated, and there is clear reason to believe that there is structure to be uncovered, which is the primary purpose of this report. Namely, we wish to examine the statistics that describe the behavior and geometry of the paths and collisions that the particles undergo. We believe that it will be useful to use these statistics to inform any future machine learning work that may be carried out, rather than applying clustering and classification based approaches blindly.

2 Data

The TrackML challenge delivers a massive dataset of independent events. Each event contains measurements of the resultant particles from collisions between protons conducted at the Large Hadron Collider at CERN. The overall goal of the competition is to correctly associate all the collisions, hereby referred to as 'hits', to their respective particles in order to uncover coherent paths that the particles take through the detector chamber. The dataset comes with the recorded hits and the ground truth, which contains the association between the particles, their initial characteristics, and the correct path. The test set does not contain the associated path.

Because the dataset is so large and rich in features, it is split up into several csv files for each event. Each event is identified with a unique 9 digit number, and each training event has four csv files containing information about the cells, the hits, the ground truth, and the particles.

The cell file contains information pertaining to the grid array of cells that detect the events.

Thus each cell is the smallest possible granularity of position. The features in the cell files are:

- **hit_id**: The identifier that corresponds to different detector cell hits
- **ch0, ch1**: Coordinates of the cell in the detector module
- **value**: The value of the charge deposited by the particle

The hits file details the actual event hits. The features in the hits files are:

- **hit_id**: The identifier that corresponds to different detector cell hits
- **x,y,z**: measured x,y,z position in millimeter of the hit in global coordinates
- **volume_id**: numerical identifier of the detector group
- **layer_id**: numerical identifier of the detector layer inside the group
- **module_id**: numerical identifier of the detector module inside the layer

The truth file contains information about the ground truth. This file maps hits to particles and the particle state at each hit.

- **hit_id**:The identifier that corresponds to different detector cell hits
- **particle_id**: numerical identifier of the generating particle as defined in the particles file. A value of 0 means that the hit did not originate from a reconstructible particle, but e.g. from detector noise.
- **tx, ty, tz**:true intersection point in global coordinates (in millimeters) between the particle trajectory and the sensitive surface.
- **tpx, tpy, tpz**: true particle momentum (in GeV/c) in the global coordinate system at the intersection point. The corresponding vector is tangent to the particle trajectory at the intersection point.
- weight: per-hit weight used for the scoring metric; total sum of weights within one event equals to one.

The particle file contains information about the particle itself. Each one of the entries in this file refer to particles that exit in the ground truth.

- **particle_id**: numerical identifier of the particle inside the event.
- **vx, vy, vz**: initial position or vertex (in millimeters) in global coordinates.
- **px, py, pz**: initial momentum (in GeV/c) along each global axis. q: particle charge (as multiple of the absolute electron charge).
- **nhits**: number of hits generated by this particle.

There are 8850 such training events split across 5 files for manageability. However, the training events have a size of 74 Gigabytes, which is too much to even load onto our machines. Because of the difficulty of working with the entire dataset at once, the train_sample.zip file was used for data exploration, which consists of just the first 100 training events, and has a much more reasonable size of less than a Gigabyte.

An important notice is that this data is simulated, which explains how details such as ground truths are precisely known. It appears that properties like charge and weight in practice are calculated from estimations. Due to the lack of details, however, it is unknown how algorithms developed on this data set may be affected.

3 Background

This section will discuss physics and functions of the LHC, and how they are leveraged in collecting data from the collisions of subatomic particles. In overview, particles are accelerated to high energy before being made to collide. Observations made by various detectors and sensors combined with principles of physics may be used to expand the data and identify the particles.

3.1 Detectors

The detectors measures the positions of the particles that cross them (hits), and are built from silicon slabs called modules, which are arranged into cylinders and disks. Modules are

either rectangular or trapezoidal, and those in the same group have similar properties such as granularity. Each slab has a module, volume and layer id, where group of modules form a volume, and groups of volumes form a layer. The detectors are augmented by a electromagnetic calorimeter, which measures the energy of electrons and photons they interact with. When the particles pass through the detectors, they deposit a charge, which is recorded and included in the dataset.

3.2 Particle Pathing

Electromagnets are used to generate strong magnetic fields in the particle collider. Particles produced in collisions normally travel in a straight line, but curve due to the magnetic field. Features of this curve including curvature and momentum provide more details about the particle. For example, it is expected the direction of curvature curve corresponds with charge of particle, due to the right-hand rule.

3.3 Radiation

The measures of radiation emitted by particles during these collisions are used to obtain some of the features of the dataset. First, charged particles traveling faster than light in a medium emit Cherenkov radiation at an angle, which may be used to calculate velocity of the particle. Mass can be calculated from the measures of velocity and momentum, and is most important in determining identity of the particle. Next, the particle emits transition radiation when it travels in between two electrical insulators with different resistances. This radiation is related to energy of the particle, which is also useful in identifying the particle.

4 Investigations

As previously stated, the present report pursues a myriad of hypothesis and explanatory endeavors on the TrackML dataset. These do not necessarily conform nor contribute directly to the original purpose of the dataset: to produce fast implementations that predict the paths followed by particles resulting from accelerating protons at the LHC. Nonetheless, the findings resulting from these statistical explorations may be insightful in the process of building statistical learning models that do attempt to tackle the TrackML challenge. Furthermore, these may be also of interest to a wider audience with an interest in high-energy particle physics.

The present report focuses on two major themes: hits and paths. We investigate the number of hits of per particle and accumulated within events in order to describe their behavior. Particle paths are also examined in order to gain an understanding of how the particles interact with the geometry and structure of the detector chamber itself. In total, our investigations pursued a total of **X** hypotheses or guiding questions all of which fall under one of the two major themes described above

4.1 Hits

4.1.1 Are the total hits per event normally distributed?

This section seeks to investigate the distribution of the total number of hits per event. Understanding this property may lead to useful insights when building predictive models for the behavior of particles in the LHC. Precisely speaking, the following question is put forth: "Are the total hits per event normally distributed?"

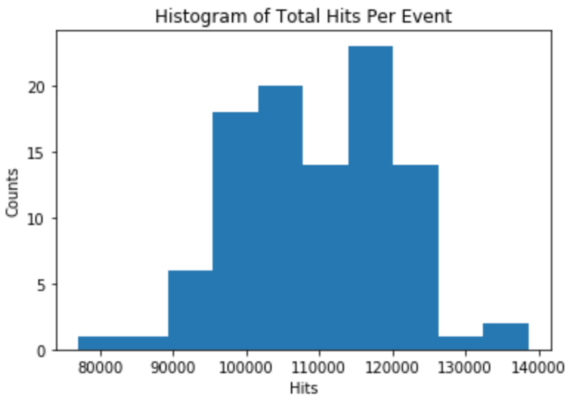

Figure 1: A Histogram of the total number of hits per event

Figure 1 above shows the distribution of values for the total number of hits per event. The distribution, at a glance, does not appear to follow a normal distribution because there is no clear single mode. Nonetheless, this shape may be an artifact of the number of bins chosen for the histogram. After all, if one were to

ignore the lack of a clearly defined mode, we see that the ends of the distribution do behave analogous to what one would expect out of a normal distribution.

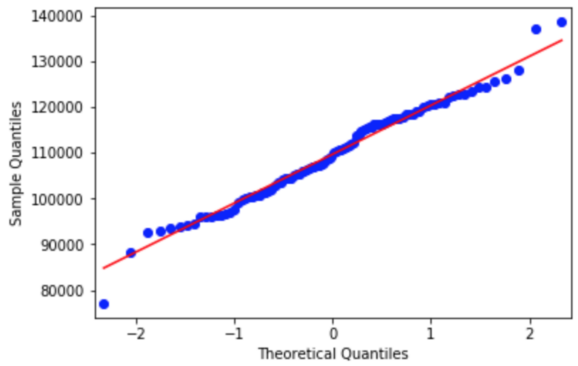

Figure 2: QQ-Plot of normality for the total number of hits per event

The QQ-Plot above suggests a drastically different story from the histogram previously shown. The sample quantiles appear to align, quite closely, with the theoretical quantiles of a normal distribution. Granted, the points at the extremes do appear to deviate from the expectation, but these are so few that they may not be indicative of a departure from normality. Therefore, this motivates the need for a chi-square test to further validate that the number of hits per event indeed follows a normal distribution.

As suggested above, a chi-square test was performed to test for the goodness-of-fit of normality. With an significance value of 0.05, a p-value of around 0.09 was obtained.Note that two parameters were estimated, namely $\hat{\mu}$ and $\hat{sigma}$, and the estimation affects the degrees of freedom in the goodness-of-fit test. Thus, we fail to reject the null-hypothesis which allows us to safely assume normality.

4.1.2 Are the number of hits in each event also normally distributed?

Motivated by the results of the previous section, this section proceeds to explore whether or not the distribution of hits per particle is also normally distributed.

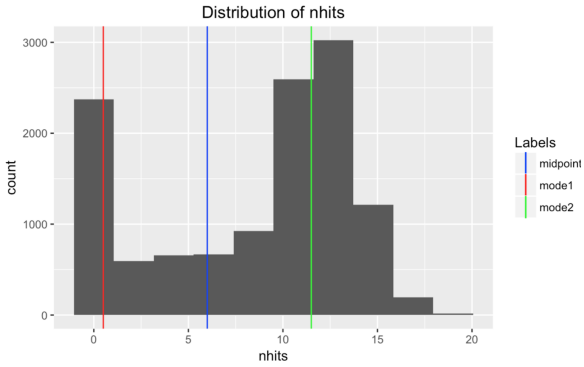

Figure 3: Distribution of hits for a randomly selected event

At a glance, Figure 3 strongly suggests that the distribution is not normal. In fact, the plot suggests bi-modality at the extremes. To account for this bi-modality, the new hypothesis is to determine if there are two distribution overlaid on top of each other. In a naive attempt to split these two distributions, the two modes were used to determine an approximate midpoint that could be used to split the data. This midpoint was used on data from another randomly selected event to obtain a left and right set. Looking at Figure 3, it would seem to suggest that the left side may follow an exponential distribution while the right side may follow a normal distribution.

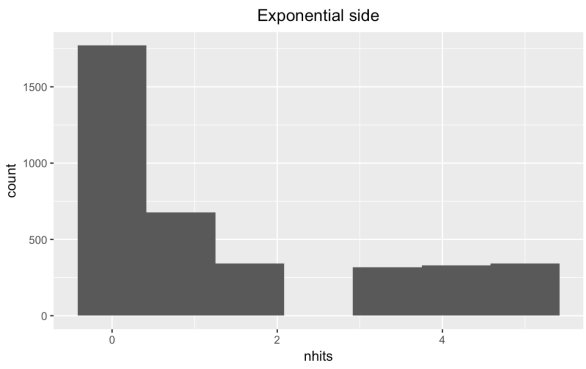

Figure 4: Left side distribution of hits for a randomly selected event

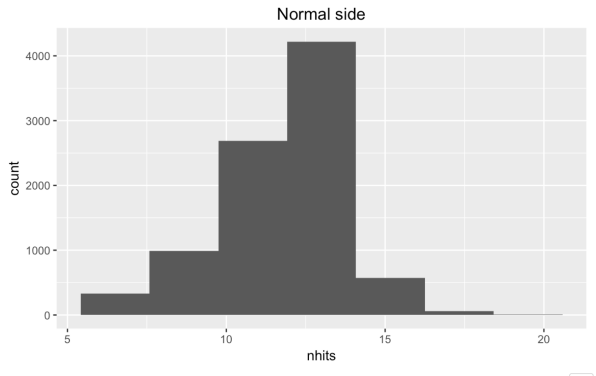

Figure 5: right side distribution of hits for a randomly selected event

This naive approach should truncate the left tail for the normal distribution and over populate the right tail of the exponential side. As expected the tail of the left side distribution in Figure 4 is visually much thicker than the tail of an exponential distribution. On the other hand, the right side distribution in Figure 5 seems to lack symmetry as the right side is much thinner. The results for chi square and KS tests for the two distributions are listed below

Table 1: Chi-square results

Distribution	Result
Exponential side	X-squared = 1071.7, df = 6, p-value < 2.2e-16
Normal side	X-squared = 3017.6, df = 6, p-value < 2.2e-16

Results from Chi-square test

Table 2: KS test results

Distribution	Result
Exponential side	D = 1, p-value < 2.2e-16
Normal side	D = 0.98324, p-value < 2.2e-16

Results from Chi-square test

Both of these goodness of fit tests fail. Since the tails of these distribution are most likely the culprits, kurtosis and skewness tests are done to visually verify this deviation.

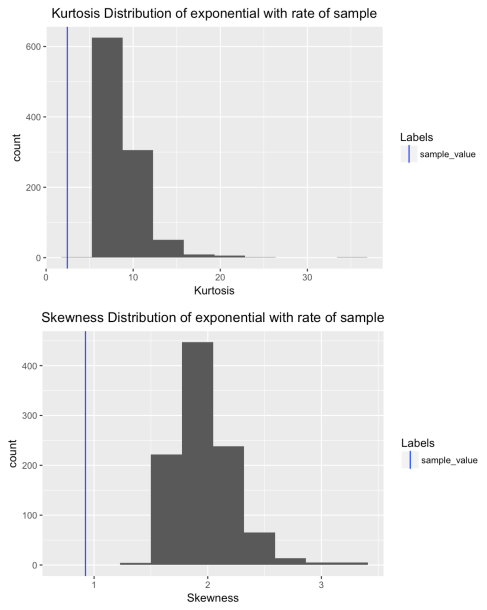

Figure 6: (top)kurtosis and skewness(bottom) of left side data compared with kurtosis and skewness of 1000 exponential samples with a rate equivalent to our sample

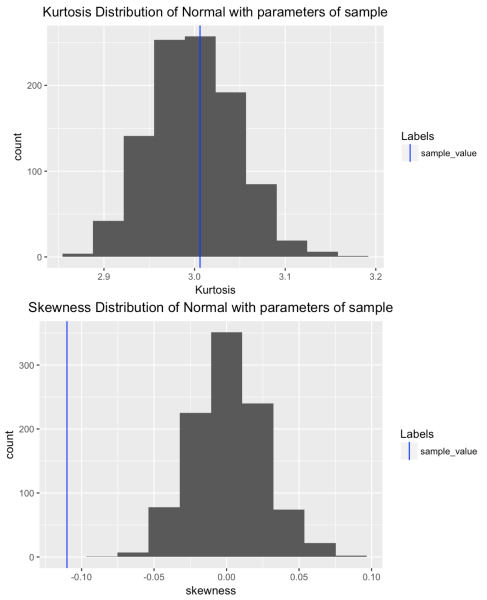

Figure 7: kurtosis(top) and skewness(bottom) of right side data compared with kurtosis and skewness of 1000 normal samples with a parameters equal to our sample

Indeed, the kurtosis and skewness of the exponential side is completely off which can be explained by the additional points that get added to the tail from the truncation process. On the other hand, the normal side is not as affected since although it loses its left tail, it

is expected for a normal distribution to have thin tails. The skewness, however, clearly fails which supports the asymmetry observed.

4.1.3 Does initial momentum predict the number of hits of a particle?

After observing the distribution of hits, the logical step was to find a variable that could predict the amount of hits. The first variable chosen to be observed was momentum, however, since this variable was given as x y z coordinates, the l2 norm was used instead as a metric for the measure of momentum.

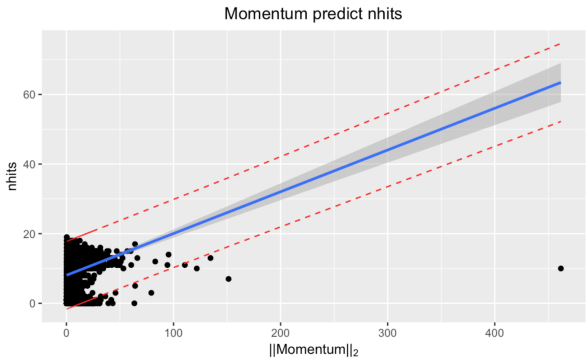

Figure 8: Linear model of momentum predicting hits

Figure 8 shows that the l2 norm of momentum does an extremely poor job for predicting hits. The statistics for this model is listed in the table below

	Estimate	Std.Error	t-value	p-value
(intercept)	8.053854	0.048705	165.36	<2e-16
momentum	0.119943	0.006206	19.33	<2e-16

Table 3: Summary statistics for linear model: momentum predict hits

R-squared	0.02956
Adjusted R-squared	0.02948

Table 4: R-squared and adjusted R-squared values for the linear model

In order to see if perhaps there are any non-linear relations, a zoomed in plot is shown (Figure 9).

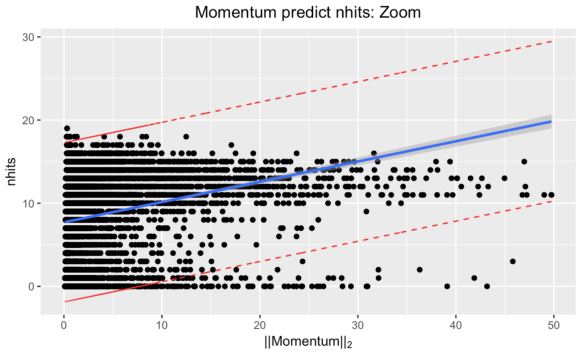

Figure 9: Linear model of momentum predicting hits

Figure 9 shows again no linear relation between the features. Although this model suggests that there may be no linear relation between our metric for momentum and hits, we also consider whether momentum is different between particles that had at least one hit versus those that had no hits at all. Thus, the new hypothesis re-frames the prediction into a classification problem of whether the l2 norm of momentum can predict if a particle falls into the group of hits or no hits. It should be noted that a different set of data was randomly sampled to test for the new hypothesis.

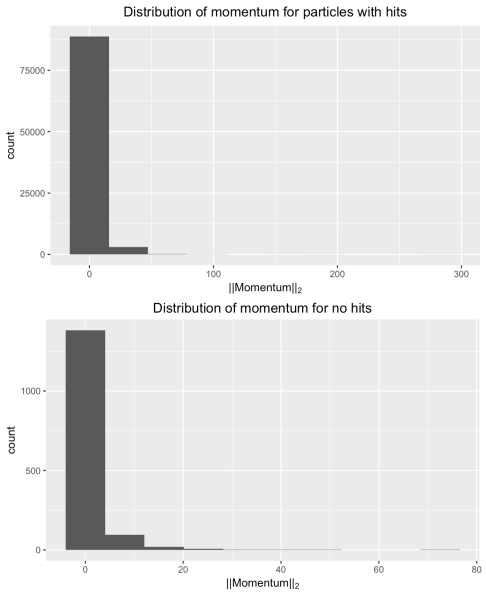

Figure 10: (Top) the distribution of momentum for the particles with no hits. (Bottom) the distribution of momentum for the particles with at least 1 hit.

At a glance, both of the distribution may seem very similar since the majority of particle

momenta in both cases are extremely packed around zero. However, the mean momentum for particles with no hits and hits comes out to be 1.55 and 3.679 respectively. To test that these means are indeed significantly different, we perform a two-tailed Wilcoxon rank sum test. Indeed, W = 30122000, p-value < 2.2e-16 which confirms that the mean momentum for these two groups are in fact different. One caveat to notice, however, is that the sample size for these two groups are drastically different. Specifically, they have 1698 and 103305 data points respectively. As a result, the standard deviation for these two data sets are quite different failing an equal test for variance F = 0.33094, num df = 1513, denom df = 92217, p-value < 2.2e-16. Thus, although the similar distribution requirement is met, the Wilcoxon test requirement of "similar" variances was violated and the result should be taken with caution.

4.1.4 Does charge predict the number of hits of a particle?

This section examines whether the proportion of positively versus negatively charged particles are different across values of the number of hits. Using the same split data set of hits and no hits, the charge for particles are visualized below.

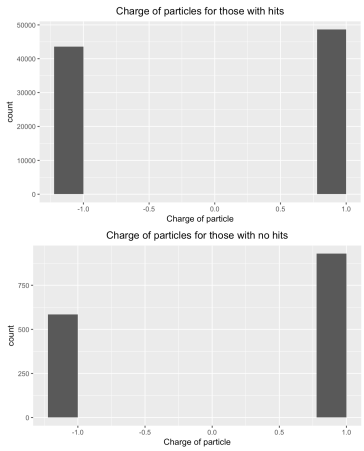

Figure 11: (Top) the distribution of charge for the particles with no hits. (Bottom) the distribution of charge for the particles with at least 1 hit

It is easy to see that the proportion of positively charged particles appears more frequently for the particles with no hits. Indeed, the proportion of positively charged particles were 0.527 and 0.6136 for those with hits and those with no hits respectively. To test the significance, a z-test of proportions was completed. A z-score of 19.59 and a p-value < 2e-16, confirms the result at a significance level of $\alpha = .05$. Again, although the sample sizes are drastically different, they are still extremely large in both cases this has little effect on the z-test statistic.

4.1.5 Is the proportion of particles never detected per event normally distributed?

Under the assumption that each proton collision is similar in nature, and that the detectors work as normal for each event, it makes sense that proportion of particles generated in the simulation which are never detected would be approximately normally distributed for a large quantity of particles per event as a mean of i.i.d. indicator variables.

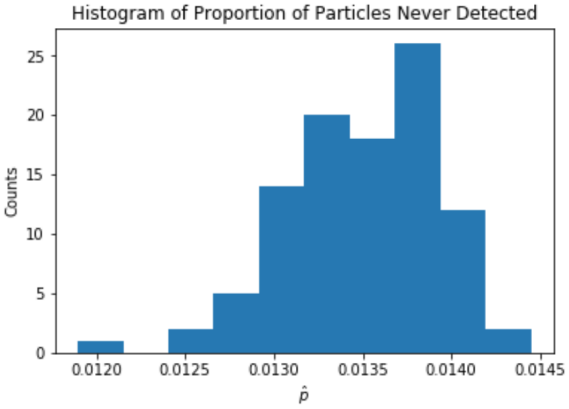

Figure 12: Histogram of Proportion of Particles Never Detected

Figure 12 showcases the proportions generated from 100 events. The histogram looks approximately normal, though the mode is not at the center of the distribution. Thus, more sophisticated techniques will be employed. Referring to the quantile-quantile plot in Figure 13, it can be observed that most quantiles for the sample match with the quantiles behind the estimated normal distribution. This provides further evidence that the proportion is normally distributed. Finally, a Pearson's Chi-Square Goodness of fit test was conducted with bins merged together to ensure that both expected and observed counts exceed 5. The

resulting test statistic was 4.837 along with a p-value of 0.1852. Note that two parameters were estimated, namely $\hat{\mu}$ and $\hat{sigma}$ and this affects the degrees of freedom in the goodness-of-fit test. Following the standard significance level of $\alpha = 0.05$, the test failed to reject the null hypothesis, and so there is quantitative evidence in favor of the distribution being normal. Thus, it is safe to conclude that the proportion of no-hits particles follows a normal distribution with estimated mean and sample standard deviation 0.01350 and 0.0004351, respectively.

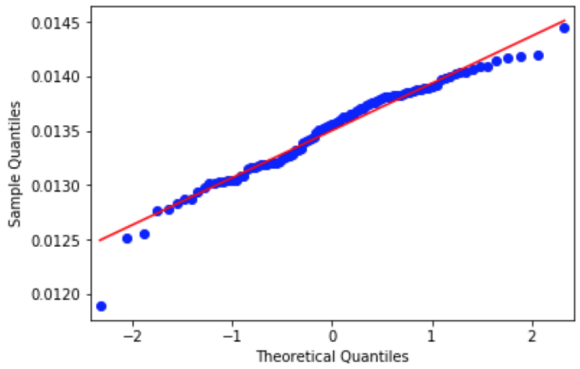

Figure 13: QQ-Plot of the proportion of particles never detected

4.2 Paths

4.2.1 Analysis of the initial momenta for the particles

Our analysis of the paths taken by particles through the chamber begins with an investigation of the initial momenta of the particles in order to see in what directions the particles go immediately following the initial collision.

Since particles collide in non-radial direction along the tube, it may make sense to analyze the momenta in terms of both cylindrical and spherical coordinates with the z direction being in the non-radial direction of the tube. Thankfully, the initial momenta are already provided in this intuitive format.

In a spherical analysis, since, particles collide along the z direction, it could make sense that the radial direction, θ is random and independent, meaning that it is uniformly distributed. A quick glance at figure 14 provides evidence towards this hypothesis.

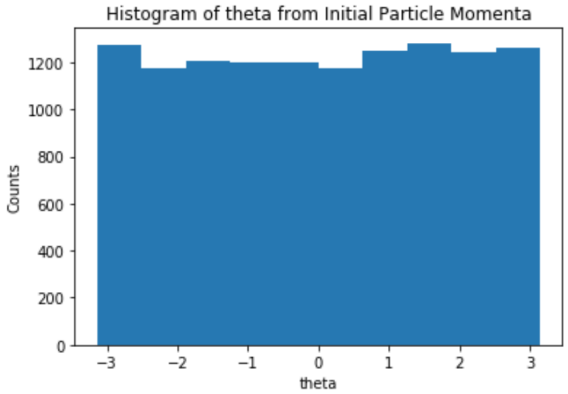

Figure 14: Histogram of theta from Initial Particle Momenta

In order to quantitatively verify the hypothesis, a Chi-Square Goodness of Fit test for uniformity is employed. The resulting test statistic and p-value were 12.1733 and 0.2037, respectively. Also note that no parameters were estimated in this test since it is always true that $\theta \in [-\pi, \pi]$. At a significance level of $\alpha = 0.05$, the test failed to reject the null hypothesis that the distribution is uniform. Hence, it is safe to conclude that θ follows a uniform distribution on $(-\pi, \pi)$.

From an intuitive grasp on how collisions occur, there may be two possibilities on what happens to all particles composing the initially

collided protons in terms of ϕ (polar angle from positive z axis). Either, the particles could scatter radially (peaked at $\frac{\pi}{2}$ radians) from the z axis, or they could bounce nearly directly off of or pass through each other, giving peaks around 0 and π radians. A glance at figure 15 seems to indicate that the plurality of particles follow the latter case. This may be a reflection of the high velocity of the collision where most constituent particles would be more likely to follow the same axis they were following beforehand. Since this histogram includes two possible phenomena that are occurring, it may be interesting to see if there is a method to separate out the data from one another, though the technique is beyond the scope of this paper. From observation, it looks as if the particles that do scatter away from the z-axis follows a uniform distribution around ϕ.

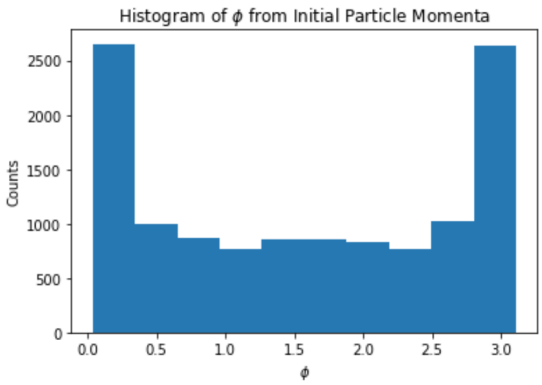

Figure 15: Histogram of Phi from Initial Particle Momenta

In the spherical coordinate scheme, it is also important to asses the magnitude of the variable as well. There is no hypothesis motivating what the distribution would look like, so an exploration of the data will have to happen.

A first glance at figure 16 of the magnitudes is concerning because it is difficult to decipher the distribution. One revelation is how the overwhelming majority of the data is close to zero, yet, there is still a large range for the data. This is because the histogram plotter will still attempt to include the entire range of the data, even if the bins do not show up on the graph.

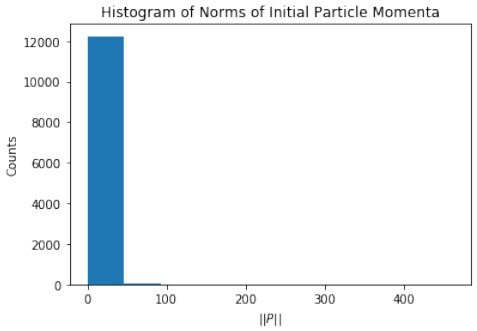

Figure 16: Histogram of initial particle momenta norms.

To assess more of what is going on with the vast majority of the data, it might be necessary to exclude some data points. To do this, the inter-quartile range is studied as in figure 17. It is important to note that the óutliersfiltered out in this analysis are not to be considered outliers for statistical purposes. This was only done for exploratory reasons. Regardless, a striking phenomena occurs in the inter-quartile range. Namely, the bins follow a concave up, decreasing shape. This suggests that for all the data points, it might be a good idea to make a data transformation.

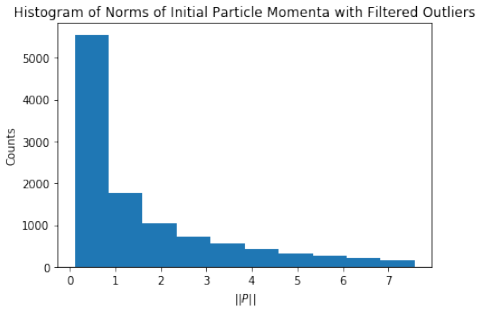

Figure 17: Histogram of initial particle momenta norms filtered for the inter-quartile range.

Figure 18 provides a histogram of the initial magnitudes for the momenta of the particles after a log data transformation was enacted. It could be suggested that the log data follows a normal distribution, though there is a clear right skew on the histogram which discounts the idea. Nevertheless the standard techniques for assessing normality can be employed.

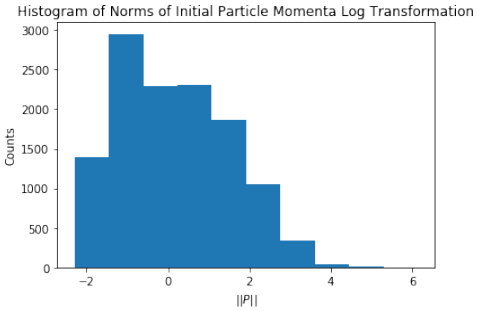

Figure 18: Histogram of initial particle momenta norms with a natural-log transformation on the data points.

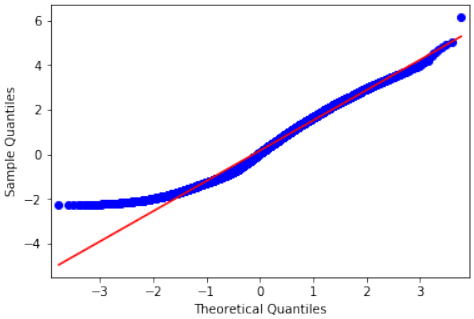

Figure 19: Q-Q Plot of initial particle momenta norms with a natural-log transformation

To begin the analysis, a quick glance at the QQ-Plot in figure 19 showcases how there is a right skew on the data despite the apparent normality that it follows. Finally, a Chi-Square Goodness of Fit Test for normality was conducted with a test statistic of 673.0985 and a P-value of 3.9278×10^{-142}. Also note that two parameters were estimated, $\hat{\mu}$ and $\hat{\sigma}$ so the chi square test has two addition degrees of freedom removed. Given an $\alpha = 0.05$ level of significance in the test, the null hypothesis is rejected in favor of the alternative hypothesis. Therefore, we cannot conclude that the data follows a normal distibution.

4.2.2 What is the typical geometry of a path?

The paths of 100 random particles are plotted below in global X,Y,Z coordinates. The Z axis corresponds to the central axis of the detection chamber, and the X and Y dimensions are perpendicular radial axes.

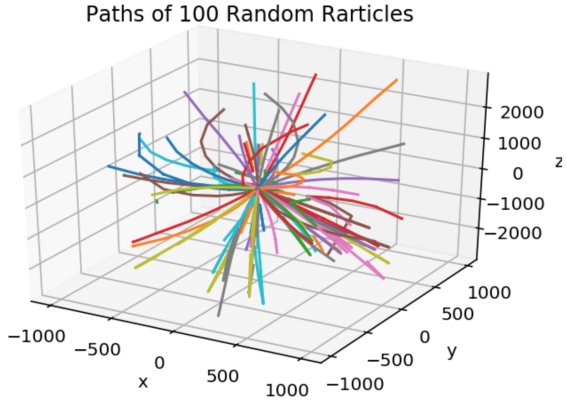

Figure 20: Paths of 100 Random Particles

The plot shows particles spreading out in all directions from the origin. Most paths appear to travel in smooth paths, without bouncing drastically. Some paths appear relatively straight, while others form helical arcs, both clockwise and counter-clockwise when viewed from the positive Z dimension. Because a strong magnetic field is oriented along the Z direction in the collision chamber, these helical paths are products of the force of the magnetic field on the particles, and the chirality of the helices would b e expected to depend on the charge of the particles. Below are 3D plots of the particles, separated by positive and negative charges.

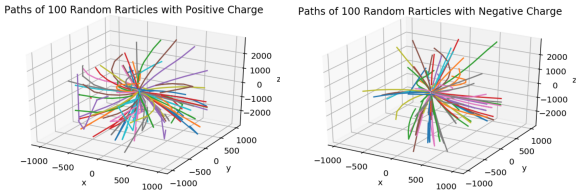

Figure 21: Paths of 100 Random Particles separated by charge

There is a clear visual difference in the chirality of the helical paths. Particles with positive charge move in a clockwise direction when view from the positive Z direction, while

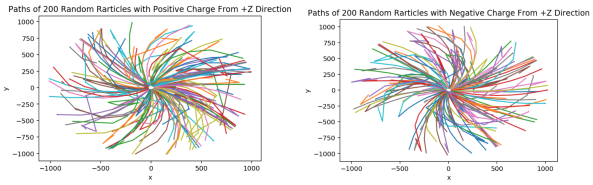

Figure 22: Paths of 200 Random Particles divided by charge in the Z-direction

particles with negative charge move counterclockwise.

4.2.3 What is the distribution of the net displacement of a path?

The total displacement of each particle in an event was examined. Similar to the momentum case, positions were give in x y z coordinates, therefore the l2 norm of the resulting directional vector was used.

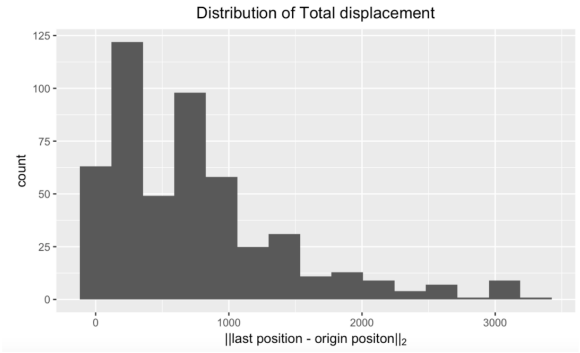

Figure 23: Distribution of total displacement

Figure 23 portrays the distribution of total displacement. It is not clear what distribution this may represent. Since no clear distribution can be seen from the overview, the next step of the procedure was to split total displacement by charge and observe if any differences can be detected.

Figure 24 displays the previous graph separated by charge. Although the distributions of net displacement for both positive and negative charge look fairly similar, the means were given to be 756.4 and 722.9 respectively. Since these are not normal, a wilcoxon rank sum test was performed to detect any shift. With a statistic W = 30523 and p-value = 0.6656, we conclude that the means were not significantly different at an α of .05. We verify the validity of this test by observing that the standard deviations were 655 and 678 respectively

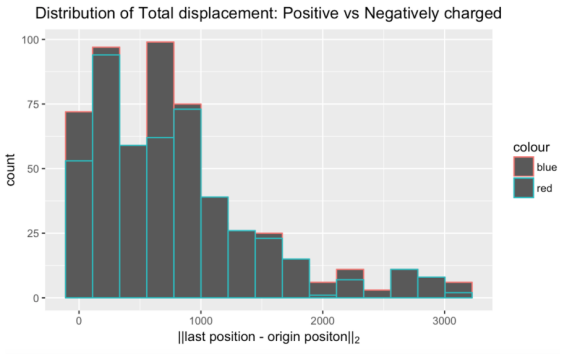

Figure 24: Distribution of Total displacement: Positive vs Negatively charged

and that an F-test of equal variance produces F_232,267 = .9337, p-value = 0.5919, which confirms that the standard deviations can be assumed equal. Figure 25 is a normalized version of figure 24, portrayed to adjust the slight difference of sample size and to confirm the blatant similarities.

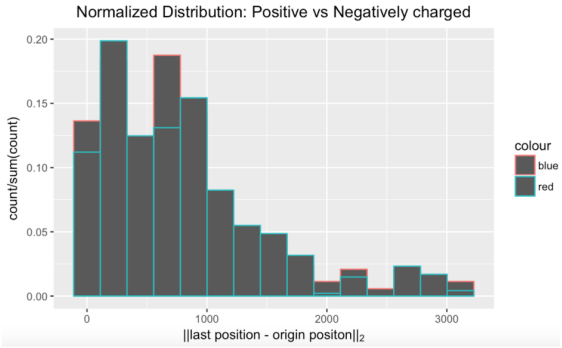

Figure 25: Normalized Distribution: Positively vs Negatively charged particles

4.2.4 Do particles bounce off of detectors, or do they go straight through?

This section considers the behaviors of particle paths as they travel between detectors corresponding to consecutive hits. Let $\vec{h_i}$ be the vector corresponding to the location of the ith hit of some particle. Then $v_{i,\vec{i+1}}$ $= \vec{h_{i+1}} - \vec{h_i}$ is the vector corresponding to the straight line path between hit i and hit $i+1$. Consider $v_{\vec{1},2}$ and $v_{\vec{2},3}$ for a random sample of 500 particles associated with at least 3 hits. By comparing these vectors with $\vec{n_2}$, the normal vector to the detector corresponding to the second hit, one may

examine the changes in the paths of particles, approximated with piecewise linear, as they collide with the second detector.

Let θ_{in} be the angle that $\vec{v_{1,2}}$ makes with $\vec{n_2}$:

$$\theta_{\text{in}} = arccos\left(\frac{\vec{v_{1,2}} \cdot \vec{n_2}}{|\vec{v_{1,2}}|\ |\vec{n_2}|}\right)$$

Let θ_{out} be the angle that $\vec{v_{2,3}}$ makes with $\vec{n_2}$:

$$\theta_{\text{in}} = arccos\left(\frac{\vec{v_{2,3}} \cdot \vec{n_2}}{|\vec{v_{2,3}}|\ |\vec{n_2}|}\right)$$

θ_{in} corresponds to the *angle of incidence* while θ_{out} corresponds to the *angle of reflection.*

Furthermore, let θ_{between} be the angle that $\vec{v_{1,2}}$ makes with $\vec{v_{2,3}}$:

$$\theta_{\text{between}} = arccos\left(\frac{\vec{v_{1,2}} \cdot \vec{v_{2,3}}}{|\vec{v_{1,2}}|\ |\vec{v_{2,3}}|}\right)$$

θ_{between} should be close to 0 for particles that transmit straight through the detector, while θ_{between} should be larger for particles that reflect off of the detector. The figure below is a scatter plot for θ_{in} vs θ_{out}, with points colored by the value of θ_{between}.

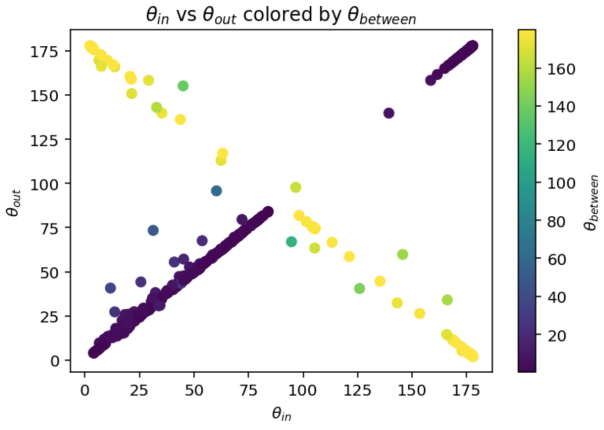

Figure 26: θ_{in} vs θ_{out} colored by $\theta_{between}$

This plot appears to consist of two distinct linear portions, one descending from $(0, 180)$ to $(180, 0)$, and the other ascending from $(0, 0)$ to $(180, 180)$. By hypothesis, transmitted particles should have $\theta_{\text{between}} \approx 0$ while reflected particles should have a higher value for θ_{between}. The color codes on the plot appears to indicate that the two linear portions of the plot can be well separated by a threshold at $\theta_{\text{between}} = 90$, so we conduct further analysis under the assumption that transmitted particles correspond to $\theta_{\text{between}} \leq 90$ and reflected particles correspond to $\theta_{\text{between}} > 90$. The figures below show θ_{in} vs θ_{out} for the transmitted particles (left) and the reflected particles (right).

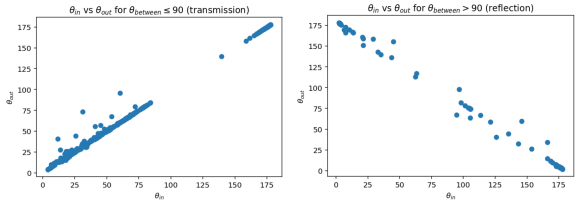

Figure 27: θ_{in} vs θ_{out} for the transmitted particles (left) and the reflected particles (right).

By the law of reflection, particles which reflect off of the detector should have θ_{out} approximately equal to 180 - θ_{in} if they do not scatter, and particles which are transmitted straight through the detector should have θ_{out} approximately equal to θ_{in} if they do not refract. We conduct a Least Absolute Deviation (LAD) regression analysis to test the null hypothesis that the slope for transmitted particles equals 1 and the slope for reflected particles equals -1 with intercepts of 0 and 180, respectively. P-values are reported below, as are residual plots. LAD was chosen over Ordinary Least Squares (OLS) due to the apparent sparsity of deviations from the dominant lines. These outliers may be due to the non-linear paths of some of the particles, and identifying these outliers may be beneficial in future work for distinguishing between particles that take straight paths and those that take spiraling paths.

Group	Coef	StdErr	t	P > |t|	[.025	.975]
Transmitted	1.0002	0	.717	.474	1.000	1.001
Reflected	-1.000004	0	-0.010	.992	-1.001	-.999

Table 5: Results of t test

With a significance level of $\alpha = .05$, we cannot reject the null hypothesis given by the law

of reflection, providing evidence that there are no significant scattering or refraction effects.

The residuals are shown in the image below.

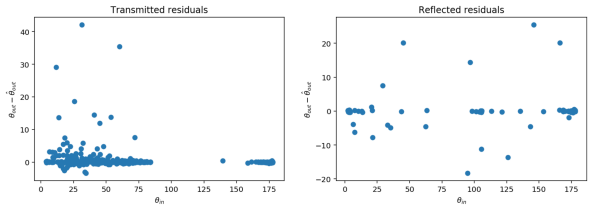

Figure 28: Residuals

4.2.5 Is the proportion of transmitted particles significantly different than .5?

The test statistic and p-value for the two-tailed test of whether the proportion of transmitted particles is significantly different than .5 are reported below:

Proportion transmitted	z-score	p-value
.796	4.729	4.188e-49

Table 6: Results of z test

With a significance level of $\alpha = .05$, we strongly reject the null hypothesis that the proportion of transmitted particles is equal to the proportion of reflected particles, suggesting that a large proportion of particles pass straight through detectors.

4.2.6 Do momentum, angle of incidence, and/or charge predict whether a particle reflects or transmits?

We test whether initial momentum of a particle, θ_{in}, and its charge are significant predictors of whether the particle will transmit or reflect. A logistic regression model was built to predict the binary value corresponding to whether a particle is transmitted (coded with 1) or reflected (coded with 0). The model used the following predictors: initial momentum, the difference in degrees of θ_{in} from the plane of the detector, and charge coded as a dummy binary variable-with positive charge coded as 1 and negative charge coded as 0. Results are provided below.

With a significance level of $\alpha = .05$, only $|90 - \theta_{\text{in}}|$ is a significant predictor of the transmission rate of a particle. The coefficient corresponding to $|90 - \theta_{\text{in}}|$ is -.0266, predicting that for every degree of increase of degree from the tangent plane of the detector, an expected difference of the log odds of transmission is -.0266, corresponding to a decrease in odds of transmission by 2.6%.

Group	Coef	StdErr	z	P > |z|	[.025	.975]
Intercept	2.9180	0.413	7.064	0.000	2.108	3.728
Momentum	-0.0111	0.024	-0.468	0.640	-0.058	0.035
$\|90 - \theta_{\text{in}}\|$	-.0266	0.006	-4.676	0.000	-0.038	-0.015
Charge	.1215	0.257	0.472	.637	-0.383	.0626

Table 7: Results of z test

5 Theory

5.1 Chi-Square Goodness-Of-Fit Test

The Chi-Square Goodness-Of-Fit test is a hypothesis test for a discrete distribution. A distribution table is calculated for the expected data. For m the number of categories or values for the response and N_j is the number of observations that appear in category j, $j = 1, ..., m$. These counts are then compared to what would be expected as

$$\mu_j = np_j,$$
$$p_j = \mathbb{P}(\text{an observation is in category} j)$$

and that $\sum p_j = 1$ so $\sum_j \mu_j = n$. The test statistic χ^2 is a measure of the discrepancy between the sample and expected counts. It is calculated as:

$$\sum_{j=1}^{m} \frac{(j\text{th sample count} - j\text{th Expected count})^2}{j\text{th Expected count}}$$
$$= \sum_{j=1}^{m} \frac{(N_j - \mu_j)^2}{\mu_j}$$

5.2 p-value

The *observed significance level*, or *p-value*, is the chance that a calculated test statistic is as large as, or larger, than observed. In context, the test statistic is compared against the χ^2 distribution. A correct probability model has an approximate chi-squared distribution with $m - k - 1$ degrees of freedom, where m is the number of categories and k is the number of parameters estimated to obtain the expected counts.

χ^2_{m-k-1} is a continuous distribution on the positive real line with a long right tail. Increasing degrees of freedom causes the distribution to become more symmetric.

If the p-value is small, then there is a low probability that the observed data actually follows the distribution. When this is the case, a residual plot can help determine where the observed data deviates from the expected distribution. For each category the standardize residuals is

$$\frac{\text{sample count} - \text{Expected count}}{\sqrt{\text{Expected count}}} = \frac{N_j - \mu_j}{\sqrt{\mu_j}}$$

The denominated standardizes the residuals in order to have equal variances.

5.3 Skewness

In the context of the distribution of a dataset, skewness refers to the degree and direction of deviation away from a standard normal distribution. The graph of a distribution with little to no skew is symmetrical about the mean and mode. A negative distribution, or one that is skewed to the left, is one such that its mean is less than its mode. A distribution that is skewed left consequentially has a tail to the right. A positive distribution, or one that is skewed to the right, has a mean that is greater than its mode, and has a tail to the left.

5.3.1 Sample Skewness

For the purposes of estimating population skewness, we used the biased estimate. Given a sample $x_1, ..., x_n$, the skewness estimate is defined as

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^3.$$

That is, the sample mean of the cube of the standardized sample.

5.4 Kurtosis

Similar to skewness, kurtosis is a descriptor of the shape of a distribution. More specifically, kurtosis describes the sharpness and height of the central peak. A high kurtosis value indicates a tall, pointy peak, with more probability mass in the tails, while a low kurtosis value indicates a less extreme distribution and more evenly distributed probability value. Thus, outliers have a large effect on kurtosis, and a large outlier disproportionately increases kurtosis. A normal distribution has a kurtosis of 3 and is said to be mesokurtic. A high kurtosis value of more than 3 is referred to as a leptokurtic distribution. A low kurtosis value of less than 3 is a platykurtic distribution.

5.4.1 Sample Kurtosis

For the purposes of estimating population kurtosis, we used the biased estimate. Given a sample $x_1, ..., x_n$, the kurtosis estimate is defined as

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^4.$$

That is, the sample mean of the fourth power of the standardized sample.

5.5 Histograms

A histogram is a useful visual tool to display the distribution of data by the discretization of the range of the variable and by portraying the counts in each bin. The peaks of a histogram allows us to quickly identify the modes of the distribution. In general, data that approximates a normal distribution are known as unimodal, whereas data that are multi-modal may represent a uniform distribution or some other unknown distribution. Symmetry can be defined as the ability to have a mirror image when the distribution is split in the middle. On the other hand, when a distribution has a tail or an uneven center of mass, it is defined to be skewed, with skewed left meaning a left tail and vice versa. Histograms are also extremely convenient for detecting outliers. Specifically, in R or python, the x axis for histograms are automatically adjusted which allows for immediate outlier detection if the axis is shown to be far larger than necessary for the histogram. Disconnected bins also offer outlier detection, however, disconnected bins that are relatively close to the main distribution should not be in general removed since these are valuable points of data. Histograms allow the ability to quickly find the probability of being in specific thresholds for the range of values of our variable. In essence, histograms are a multinomial approximation of the pdf, which makes the count of observations within specific bins the probability of that range of values. The

Bin width for a histogram is essential for obtaining a good approximation of the distribution. Too small a bin width will put too much emphasis on random fluctuations in the data, whereas too large a bin width will not capture any information of the distribution.

5.6 Quantiles

Quantiles are a method of dividing up a statistical distribution into equally sized proportions. The most familiar quantile is the median, which splits the distribution into two equal probability masses of 50% each. Another set of familiar quantiles is the quartiles that divide up the distribution into four equally sized partitions of 25% each.

5.7 Quantile-Quantile Plot

Quantile-quantile plots are a graphical method used to compare two data distributions. Each point corresponds to the quantile of the first distribution against the same quantile of the second one. Two samples are the same if the plot matches a line of slope 1 and intercept 0, and departures from a linear form indicate differences in the distributions.

5.8 Linear Regression

Linear regression is a linear model between two continuous variables. The case of one explanatory variable is called simple linear regression. In its canonical form, a linear regression model is formulated as follows:

$$y = \beta_1 + \beta_2 X$$

In the equation above, y denotes the response variable, the variable being modeled or predicted. The variable X denotes the independent variable, which is used as a *predictor* of y. The term β_2 is, formally, the slope of the linear model. In terms of its intuition, β_2 is to be understood as the relationship between X and y. Lastly, β_1 is the bias term, which enables the model to displace away of the origin.

5.9 Logistic Regression

In the field of Machine Learning, Logistic regression is a linear binary classification model for categorical data. Its output is bounded between 0 and 1 and thus it has a direct interpretation as a probability value. Formally speaking, it is defined as follows.

$$y = \frac{1}{1 + e^{-(\beta_1 + \beta_2 X)}}$$

5.10 Hypothesis Tests

A Hypothesis Test allows the investigator to suppose an initial belief of the world and compare this belief with data observations. If the data falls within reasonable differences of chance, the null hypothesis is unable to be rejected. If the data is absurdly different from what the initial belief was assumed to be, the null hypothesis is rejected and the alternate hypothesis is assumed to be true. The example below is displayed to elucidate this principal.

Example 1. In Hennepin County, a simple random sample of 119 households found an average radon level of 4.6 pCi/l with a standard deviation of 3.4pCi/l. In neighboring Ramsey County, a simple random sample of 42 households has an average radon level of 4.5 pCi/l with a standard deviation as 4.9pCi/l. It is claimed that the households in these two counties have the same average radon level and that the difference observed in the sample averages is due to chance variation in the sampling procedure.

A probability model is first assumed. Let $X_1, \ldots, X_{119}$ and $Y_1, \ldots, Y_{42}$ denote the randon levels for the Hennepin County and Ramsey County respectively. Let μ_h, μ_R and σ_H, σ_R denote their respective means and standard deviations. The *null hypothesis* claims that the respective means of these two counties are the same.

$$H_0 : \mu_H = \mu_R$$

Whereas, the *alternative hypothesis* claims that the respective means will be different.

$$H_1 : \mu_H \neq \mu_R$$

Now by the Central Limit Theorem,

$$\bar{x} \sim N(\mu_H, \frac{\sigma_H^2}{119}),\ \bar{y} \sim N(\mu_R, \frac{\sigma_R^2}{42})$$

$$\implies \bar{x} - \bar{y} \sim (\mu_H - \mu_R, \frac{\sigma_H^2}{119} + \frac{\sigma_R^2}{42})$$

Thus, with the assumption that the null hypothesis is true,

$$\bar{x} - \bar{y} \sim (0, \frac{\sigma_H^2}{119} + \frac{\sigma_R^2}{42})$$

A fair test statistic is then the standardized version $Z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_H^2}{119} + \frac{\sigma_R^2}{42}}} \sim N(0,1)$ known as the Z-test.
Entering the observed values into this Z-statistic gives an observe Z of

$$Z_{\text{observed}} = \frac{4.6 - 4.5}{\sqrt{\frac{3.4^2}{119} + \frac{4.9^2}{42}}} = .12227$$

The p-value offers a measure for how likely this observed data was to occurs given that the null hypothesis was in fact true. In particular, since the alternative hypothesis $\mu_h \neq \mu_R$ leaves possibilities in both end, we look at the probability that the Z statistic could have fallen as far or further away from the mean (0) as the observed Z on both ends of the distribution of the Z statistic. One then obtains

$$P(|Z| > Z_{\text{observed}}) = P(|Z| > 0.122) = 0.9$$

If this value was lower than 5% one could conclude that the data observed is in disagreement with the null hypothesis, hence giving support towards rejecting the null hypothesis. The cutoff of 5% is arbitrary, however, is consistently kept as the standard for hypothesis tests in the scientific field. Note that the p-value is not the chance that the null hypothesis is true: the hypothesis is either true or not. Given the null hypothesis was true, the p-value gives the chance that the observed or worse values occurred.
Now although we reject the null hypothesis when p-values are less than 5%, by definition this conclusion may not be correct. In fact, it will be incorrect 5% of the time since the observed value could have just been incredible unlucky and one of the 5%'s. This is know as the Type I error. This error is controlled and is exactly the $5\% = \alpha$ or the significance level of the test. On the other hand, there is also a Type II error where one might fail to reject the null hypothesis even though the alternative hypothesis may be true. For this error, a particular alternative must be specified. In practice, however, one generally looks at the power of a test which is 1 - Type II error, rather than the actual Type II error. For the previous example one might look at the alternative that $\mu_H - \mu_R = 0.5$. Then

$$\begin{aligned}
&\mathbb{P}(\frac{|\bar{x} - \bar{y}|}{.81} > 1.96) \\
&= \mathbb{P}(|\bar{x} - \bar{y}| > 1.96 * 0.81) \\
&= \mathbb{P}(\bar{x} - \bar{y} > 1.58) + \mathbb{P}(\bar{x} - \bar{y} < -1.58) \\
&= \mathbb{P}(\frac{\bar{x} - \bar{y} - 0.5}{0.81} > 1.34) \\
&\quad + \mathbb{P}(\frac{\bar{x} - \bar{y} - 0.5}{0.81} < -2.58) \\
&= 0.09
\end{aligned}$$

The power is then the probability of rejecting the null hypothesis when in fact the alternative hypothesis is true. In a sense, it is the probability of the test detecting the alternative hypothesis correctly. In this case, the test was not very powerful and a larger sample would be necessary. A friendly table of the mentioned errors is below:

	fail to reject H_0	reject H_0
H_0 true	✓	Type I error
H_A true	Type II error	✓

Table 8: Type I and Type II errors with respective Truth for rows and Decisions as columns

5.11 Residual Plots

In regression analysis, a residual is the vertical distance from a data point to the regression line. A residual plot is a graphical way of representing the residuals of a regression model trained on the dataset in which it was trained. These plots can be used to evaluate, somewhat subjectively, how well a model fits the data. A non-normal distribution of the residuals may indicate, for example, that our modeling assumptions (such as a linear relationship between the variables) are incorrect. Furthermore, unusually large residuals may suggest the presence of outliers that, perhaps, negatively influence our model's predictive capabilities.

5.11.1 Kolmogorov-Smirnov(KS) Test

Given observations $X_1, \ldots, X_n$ and a predetermined distribution P. The Kolmogorov-

Smirnov test is used to test

H_0 : the samples come from P

vs

H_1 : the samples do not come from P

Let F_{exp} be the expected CDF of P, and F_{obs} be the empirical distribution function define below,

$$F_{\text{obs}}(x) = \frac{\#\text{ observations below } x}{\#\text{ observations}}$$

The Kolmogorov-Smirnov statistic is then

$$D_n = \max_x |F_{\text{exp}}(x) - F_{\text{obs}}(x)|$$

The 95% level critical value for this statistic is given by

$$D_{\text{crit},0.05} = \frac{1.36}{\sqrt{n}}$$

This test is used as a non parametric equivalent of the chi-square goodness of fit test.

References

[1] David Rousseau et al. "The TrackML challenge". In: *NIPS 2018 - 32nd Annual Conference on Neural Information Processing Systems*. Montreal, Canada, Dec. 2018, pp. 1–23. URL: `https://hal.inria.fr/hal-01745714`.