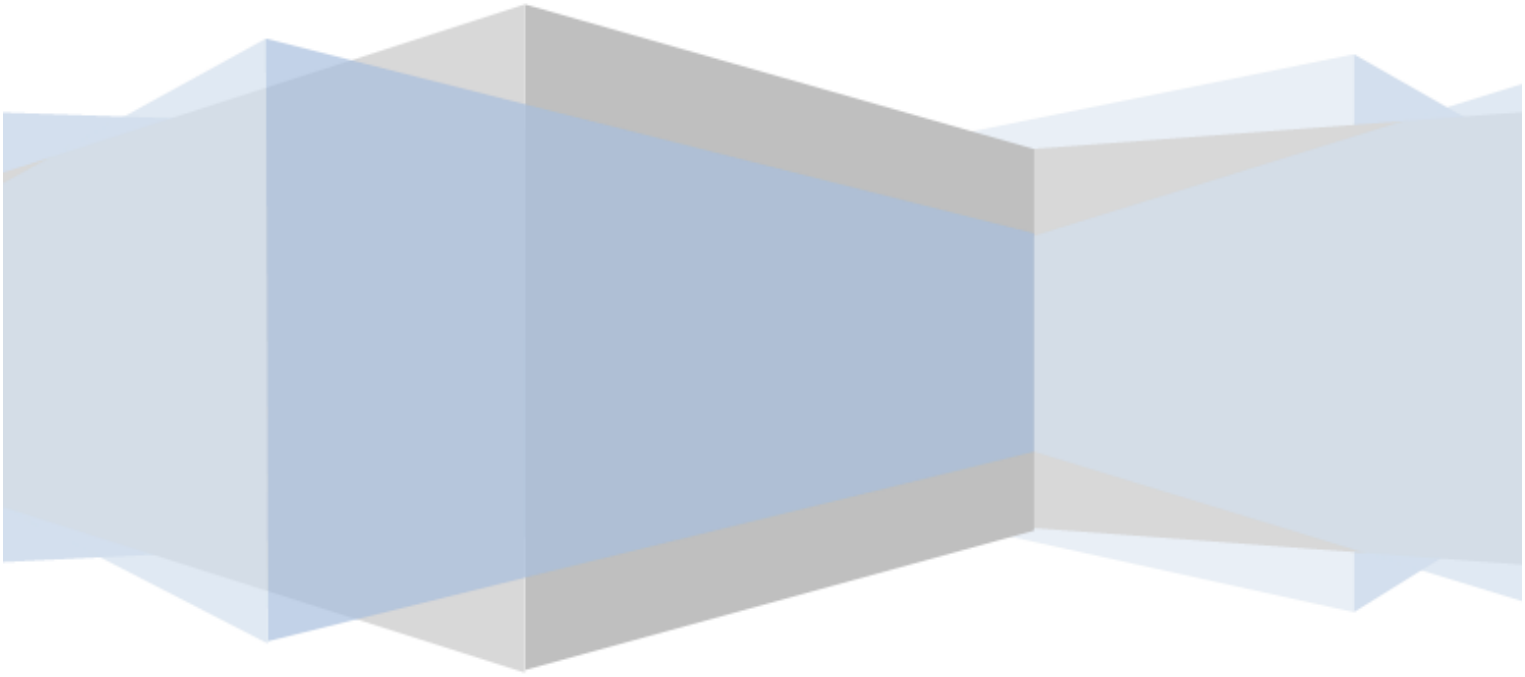
CMS, CERN

Statistical Analysis of Upper Bound using Data with Uncertainties



Tng Jia Hao Barry
Supervisor: Dr. David d'Enterria
Jun-Aug 2014

Contents

Please note that the code for this project may be assessed at the following location:
http://nbviewer.ipython.org/gist/anonymous/477d82a9d4c4ea84f0e2

1 Introduction

1.1 Motivation from Physics

Why might we deal with upper bounds in particle physics?

Recent literature in the area of deep inelastic scattering has proposed and studied an upper bound for the ratio $R(W^2, Q^2) = \sigma_L(W^2, Q^2)/\sigma_T(W^2, Q^2)$. Specifically, this ratio of cross sections for longitudinally and transversely polarized photons has been proposed to obey a numerical upper bound of

$$R(W^2, Q^2) \leq 0.37248 =: c$$

for all values of W^2 and Q^2. A detailed derivation can be found in [1]. Subsequent measurements published from the H1 and ZEUS collaborations have provided readings of R at different Q^2 values, and as discussed in [2], the data for R comes close to the predicted bound, but the bound is still respected within experimental errors. As this bound is a rigorous prediction of the color dipole model, we would like to give a more quantitative assessment of the situation.

We will be using 30 data points from the H1 and ZEUS collaborations as published in [3]. For each data point, we have the following values of primary concern to us:

- Q^2
- R_{obs}
- R_{minus} which indicates how much below is the lower end of the 68% error bar relative to the actual reading
- R_{plus} which indicates how much above is the upper end of the 68% error bar relative to the actual reading

The data is tabulated in Appendix A. Note that R_{obs} is not the actual realized value of R. Rather, it is what we observed after the realized value of R has been perturbed by errors. The 68% confidence interval then meant to include the actual realized value of R with 68% probability.

1.2 Statistical Formulation

The actual problem statement that we will work with

Let us now frame our discussion in a more statistical language. Note that the formulation which we are using does not correspond exactly the physical situation above exactly:

Statistical Formulation: We shall treat our quantity of interest R to be a continuous random variable[1] and let $X_1, \dots, X_n$ be i.i.d. measurements of R. However, what we have as data are $Y_1, \dots, Y_n$ where $Y_i := X_i + \epsilon_i$ with the ϵ_i representing the errors. In other words, the values of R_{obs} tell us the Y_i's although we would rather be in an error-free world and directly working with the X_i's to study the distribution of R. The hypothesis here $\mathcal{H}$ is that the support of R is contained in $[0, c]$.

[1] For our analysis, we will *not* a priori constrain R to be non-negative.

In particular, the data points from H1 or ZEUS represent probing of R at *distinct* Q^2 bins but our statistical framework is for i.i.d. probing of R. However, for the purpose of having some numerical data to test implementation ideas, let us treat the data points as though they were generated in accordance to our statistical formulation. To further simplify our analysis, we shall make the following two assumptions:

Assumption 1: The ϵ_i's are independent of each other and of the X_i's.

From studying the way the errors are handled in [3], we are justified to make the following assumption:

Assumption 2: The error bars indicate the 68% frequentist confidence intervals of a shifted Gaussian[2]. In other words, we assume $Y_i = X_i + \epsilon_i$ with $\epsilon_i \sim N(\mu_i, \sigma_i^2)$ where the parameters can be computed by

$$\sigma_i = \frac{R_{plus} + R_{minus}}{2} \quad and \quad \mu_i = \frac{R_{minus} - R_{plus}}{2}$$

We will now study two main approaches to our statistical problem. If we happen to know some information about the distribution of R - for example the family of distribution that R belongs to – our problem reduces to one of parameter estimation, for which many tools are available. This will be studied in section 2. Unfortunately, we might not be in such a situation (and this is generally the story), in which case more sophisticated techniques are needed. Section 3 presents the use of Dirichlet Process to perform nonparametric Bayesian analysis as a potential solution. Section 4 documents other miscellaneous techniques considered during the course of this project.

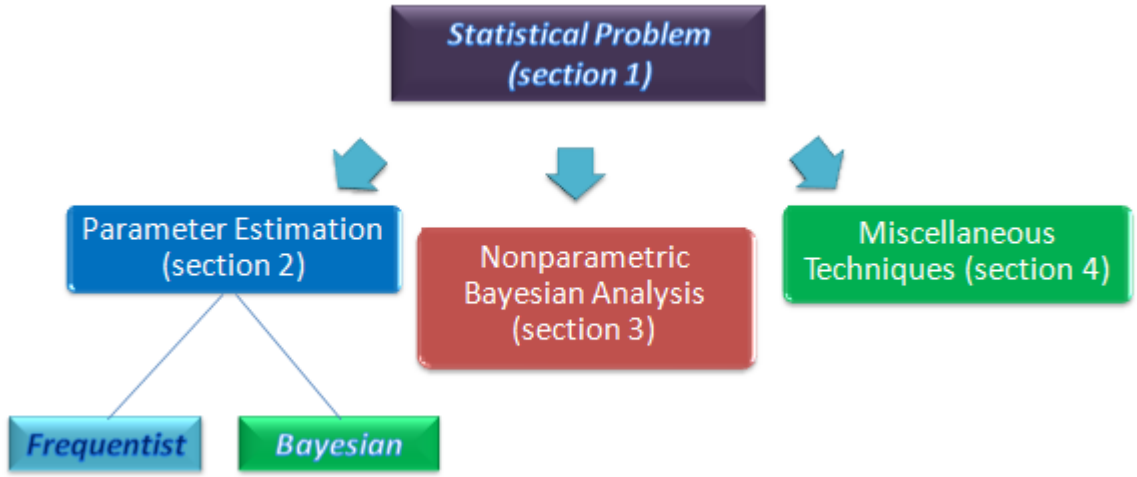


Figure 1: Outline of write-up

[2] This must mean that the error bars are constructed without constraining R to follow the hypothesis $\mathcal{H}$ of having support in $[0, \theta]$ (or anything about the distribution of R). That is, the error bars are constructed as though X_i is a fixed parameter.

2 Approach I – Parameter Estimation

2.1 The Set-up

In this section, we study possible approaches to our problem if we are fortunate enough to know the family of distribution that R belongs to. Although the approaches here apply more generally, for the purpose of concrete illustration, let us make the following specific assumption.

Assumption 3: $R \sim Unif[0, \theta]$ where θ is unknown.

Notice that we are making a switch from hypothesis testing to parameter estimation. While hypothesis testing and parameter estimation are two sides of the same coin, it would be more meaningful in this context to provide estimations for θ than testing $\mathcal{H}$.

There are two possible routes to take – Frequentist or Bayesian.

2.2 The Frequentist Approach

Here, we proceed with a relatively standard approach of calculating the MLE and use the profile-likelihood approach for constructing our confidence intervals. However, as we will see, our first attempt at using the MLE technique will run into slight theoretical problems.

2.2.1 First Application of MLE Technique

Letting f and g_i represent the probability density function (pdf) of $Unif[0, \theta]$ and $N(\mu_i, \sigma_i^2)$ respectively, we can derive the pdf of Y_i as follows:

$$\begin{aligned} h_i(y_i|\theta) &= \int_{-\infty}^{\infty} f(y_i - u) g(u) du \\ &= \int_{y-\theta}^{y} \frac{1}{\theta} g(u) du \\ &= \frac{1}{\theta}\left[\Phi\left(\frac{y_i - \mu_i}{\sigma_i}\right) - \Phi\left(\frac{y_i - \theta - \mu_i}{\sigma_i}\right)\right] \end{aligned}$$

where Φ is cumulative distribution function (cdf) of the standard normal. Thus, the loglikelihood function is

$$l(\theta) = -n \ln \theta + \sum_{i=1}^{n} \ln\left[\Phi\left(\frac{y_i - \mu_i}{\sigma_i}\right) - \Phi\left(\frac{y_i - \theta - \mu_i}{\sigma_i}\right)\right]$$

Let $\hat{\theta}$ denote a maximum likelihood estimator (that is, the value of $\theta > 0$ that maximizes $l(\theta)$ - assuming a global maximum exists). Then for each θ, we have the test statistic

$$T_\theta := 2\left(l(\hat{\theta}) - l(\theta)\right)$$

Wilk's theorem tells us that under suitable hypotheses, T_θ is well-approximated by the χ^2 distribution. However, note that Wilk's theorem as is stated and proven in literature works with i.i.d.

data, which does not fit the description of our data. The correct approach would then be to use Monte Carlo techniques to estimate the distribution of T_θ.

Let us for the moment assume that T_θ is still well-approximated by χ^2, because we will quickly propose a way to remedy the defect of non-i.i.d. data. We may then compute for each θ a p-value which is given by

$$P(T_\theta \geq t_\theta) \; under \; T_\theta \sim \chi^2$$

where t_θ is the observed value of the test statistic. The 95% confidence interval is then the values of θ for which the p-value is ≥ 0.05.

For our set of data, this procedure gives the MLE of $\hat{\theta} = 0.328$ and the 95% confidence interval of $[0.257, 0.416]$. See figure 2.

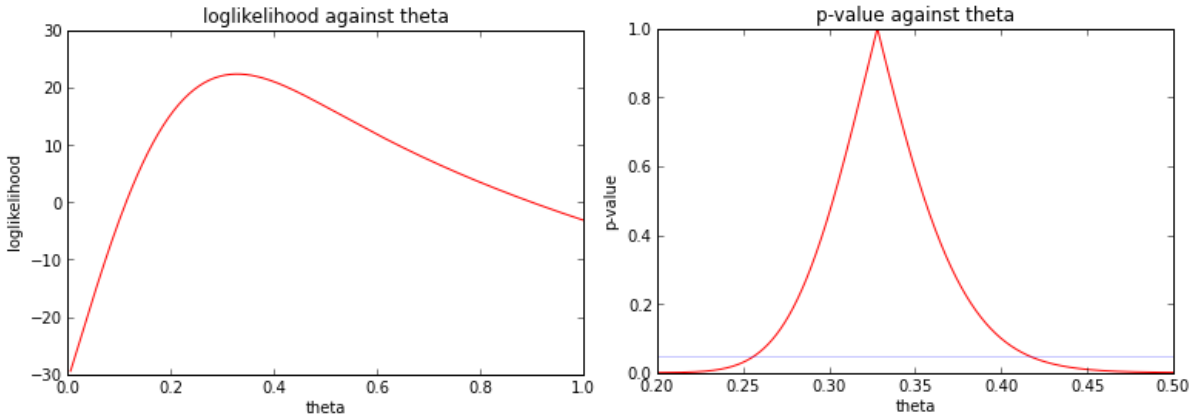


Figure 2: (Left) Plot of loglikelihood function $l(\theta)$, which achieves its maximum at $\hat{\theta} = 0.328$. (Right) Plot of p-values for each θ. The values of θ whose p-values is ≥ 0.05 give us the 95% confidence interval of $[0.257, 0.416]$.

2.2.2 The Usefulness of i.i.d. data

Most statistical techniques and theory rely on having i.i.d. data – for example, the application of Wilk's theorem in section 2.2.1. Unfortunately,our data $Y_1, \ldots, Y_n$ is not i.i.d. due to the presence of non-identically distributed errors. However, we would find it useful if we are able to construct i.i.d. data out of what we are given – and we can do this if we are willing to intentionally add further errors as follows.

Constructing i.i.d. data: We have on hand $Y_1, \ldots, Y_n$ with $Y_i = X_i + \epsilon_i$. We only require assumptions 1 and 2 for this construction. Define $\sigma := \max_{1 \leq i \leq n} \sigma_i$ and for each fixed i, independently draw the random variables $\delta_i \sim N(-\mu_i, \sigma^2 - \sigma_i^2)$. Finally, we set

$$W_i := Y_i + \delta_i$$

Then $W_1, \ldots, W_n$ are i.i.d. from the distribution that is the convolution of the distribution of R with $N(0, \sigma^2)$. ■

Note that the same data $Y_1, \dots, Y_n$ can produce various $W_1, \dots, W_n$, and so depending on which $W_1, \dots, W_n$ we end up with, the result of our statistical tests may differ. In other words, $(W_1, \dots, W_n)|(Y_1, \dots, Y_n)$ is not a constant vector. Hence, in performing statistical tests with $W_1, \dots, W_n$, it is advisable to repeat the test with many productions of $W_1, \dots, W_n$. If all the results cluster around the same conclusion, we can be sure that the conclusion is applicable to our original data $Y_1, \dots, Y_n$. On the other hand, if the results point in a spread of direction, then it implies that we have added too much artificial error that it dilutes the information present in our original data. Which situation occurs depends on the size of $\sigma^2 - \sigma_i^2$ for all i, which reflects the amount of error we artificially add for the sake of i.i.d. data.

2.2.3 MLE Technique Redux

Using the technique just described, we constructed 40 sets of i.i.d. data $W_1, \dots, W_n$ from our single set of given original data $Y_1, \dots, Y_n$ and perform the analysis of section 2.2.1 on each of them. The results are shown in figure 3.

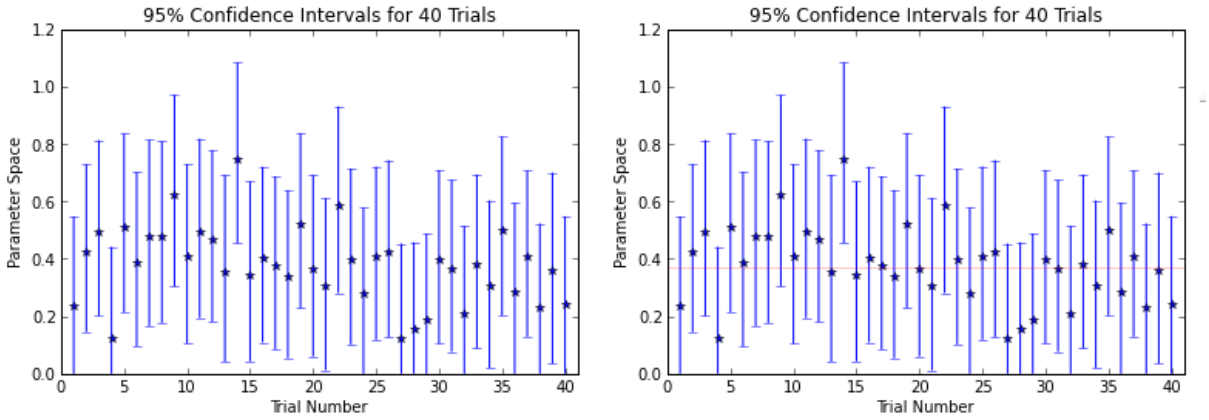


Figure 3: (Left) Plot of the 95% confidence intervals for 40 trials ran on the same set of given data. (Right) The hypothesized bound of 0.37248 drawn in red.

The amount of fluctuations between different trials is due to the effect of the added noise which scales with $\sigma^2 - \sigma_i^2$. If the fluctuations are large, we would be wary about using the confidence intervals from a single trial. In fact, in such a situation, it may be a better idea to return to the method of section 2.2.1 paired with the use of Monte Carlo to replace Wilk's theorem.

We also note that the confidence intervals are generally wider than that of section 2.2.1. This is expected because the total amount of noise in the system has (artifically) increased. The widths of confidence intervals scale with σ^2. Therefore, if σ^2 is small, we would be in good shape both in terms of fluctuations and widths of confidence intervas.

2.3 The Bayesian Approach

The Bayesian approach is simply a direct application of the standard Bayesian paradigm. Let our data be $Y_1 = y_1, \dots, Y_n = y_n$ and let us recall that

$$f(y_1, \ldots, y_n|\theta) = \frac{1}{\theta^n} \prod_{i=1}^{n} \left[\Phi\left(\frac{y_i - \mu_i}{\sigma_i}\right) - \Phi\left(\frac{y_i - \theta - \mu_i}{\sigma_i}\right) \right]$$

and that Bayes' rule gives us

$$f(\theta|y_1, \ldots, y_n) = \frac{f(y_1, \ldots, y_n|\theta) f(\theta)}{\int_0^\infty f(y_1, \ldots, y_n|\theta') f(\theta') d\theta'}$$

Under lack of any prior information, we may, for example, choose $f(\theta)$ to be a flat (improper) prior $f(\theta) \propto 1$, or the Jeffrey's prior[3] corresponding to the uniform distribution $f(\theta) \propto 1/\theta$, although we are not exactly dealing with the uniform distribution here. We would like to stress that other priors are possible.

It would be not worthwhile (or perhaps even impossible) to obtain a closed-form solution for the posterior distribution. Instead, we would construct the posterior distribution numerically using a computer. Figure 4 displays the results of Bayesian analysis on our data using the two priors suggested above. We compute $P(\theta > 0.37248 \mid data)$ as an indication of how likely our hypothesis $\mathcal{H}$ is false now, *assuming the particular prior distribution*. The respective values are 0.175 and 0.145.

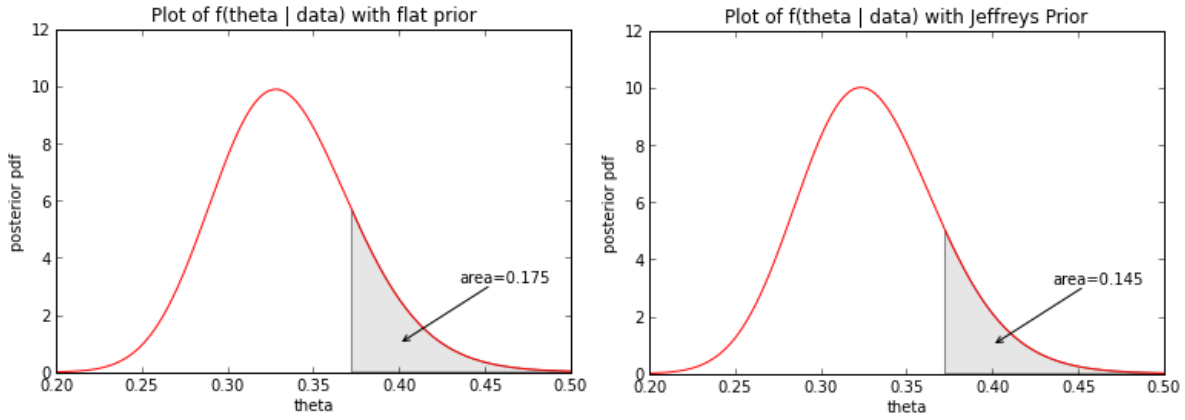


Figure 4: Plots of the posterior distribution assuming a flat prior (left) and the Jeffrey's prior for uniform distribution (right). The maxima are at the values of 0.328 (same as MLE) and 0.323. In addition, $P(\theta > 0.37248 \mid data)$ are respectively 0.175 and 0.145.

Remark: Under a flat prior, the posterior distribution is essentially a normalized form of the likelihood function. We appear to be doing something similar to the frequentist MLE approach, but the interpretation is very different. Moreover, this approach allows for a variety of priors to incorporate various beliefs about θ. ■

[3] The Jeffrey's prior is a common choice for uninformative priors due to the various "nice" properties that it possess.

2.4 Testing the Uniform Assumption

Sometimes, we might not know that the distribution of R belongs to the uniform family, but we suspect that it would be a good guess. Here, we propose a test to check that this assumption is at least not a completely horrendous one. The basic idea is to use the 1-sided Kolmogorov-Smirnov test for goodness-of-fit. However, two deviations from the standard application is required for our situation.

Firstly, we have an estimated parameter θ. That is, we are not testing goodness-of-fit to one specific distribution, but rather to a family of distribution indexed by the parameter. Modifications to address this complication have been published in literature – for example, see Lilliefor's test for testing goodness-of-fit to the family of normal distribution. The procedure that we will be describing follows closely to the approach adopted by Lilliefor. Specifically, we will describe a test statistic that is a slight modification of the standard KS test statistic, use Monte Carlo to study its distribution, and then use the observed test statistic value to compute a p-value.

Secondly, i.i.d. data is required for the KS test (and even the Lilliefor's modification) but $Y_1, \dots, Y_n$ are not i.i.d.. We will use the same technique as described in section 3.2 to artificially create i.i.d. data $W_1, \dots, W_n$ and then perform the test using this latter set of data. As discussed in section 3.2, it would be advisable to generate many sets of $W_1, \dots, W_n$ and perform the test on each set.

Remark: Unfortunately, our test will be weaker than desired. Ideally, we take the distribution of R and perform a goodness-of-fit test with the uniform distribution. However, because of the errors, we have to take the distribution of R convolved with normal, and we perform a goodness-of-fit with the uniform convolved with the normal. The normal component clearly fits perfectly with the normal component. Therefore, we are artificially boosting the goodness-of-fit. ■

The standard KS test would use the test statistic

$$D_{KS} \coloneqq \sup_x \left| F_{emp}(x) - F_{theo}(x) \right|$$

where F_{emp} is the empirical CDF and F_{theo} is the CDF of the specific distribution being testing for. Here, we do not have a specific distribution, so we define our test statistic slightly differently as

$$D \coloneqq \sup_x \left| F_{emp}(x) - F(x) \right|$$

where F is the CDF of $Unif\left[0, \hat{\theta}\right]$ convolved with $N(0, \sigma^2)$. Basically we use the "most likely" distribution in place of a specific theoretical distribution.

Next, in standard KS test, the null hypothesis is that the data is drawn from a specific distribution and given that null hypothesis, the distribution of D_{KS} has been well-studied so p-values can be readily computed. In our case, we would need to use Monte Carlo techniques to investigate the distribution of D. Now, our null hypothesis is a family of distributions rather than a specific one. We thus have to first fix some $\theta = \theta_0$ and then with this specific null hypothesis distribution, use Monte Carlo to produce the distribution of $D|\theta_0$. If we are fortunate, the distribution of $D|\theta_0$ is the same for all feasible choices of θ_0 (this is the case for the normal distribution in Lilliefor's modification).

Otherwise, we would compute the distribution of $D|\theta_0$ for many choices of θ_0 and hope that the p-values for the various choices of θ_0 all point to the same conclusion.

Here is the result of applying the above theory to our data – solely for illustration purposes since we have little reason to suspect that the uniform family would be a good guess. We use the Monte Carlo technique with 1000 samples to estimate the distributions of $D|\theta_0$ for $\theta_0 = 0.3$ and $\theta_0 = 0.4$. Having done this, we construct our i.i.d. data set $W_1, \dots, W_n$ and compute the p-value corresponding with the choice $\theta_0 = 0.3$ and a p-value corresponding with the choice $\theta_0 = 0.4$. This is then reflected in the diagram below with a cross on each of the horizontal line. We then repeat with another set of $W_1, \dots, W_n$ to obtain another pair of crosses, and so on, for a total of 40 trials. The result is shown in figure 5.

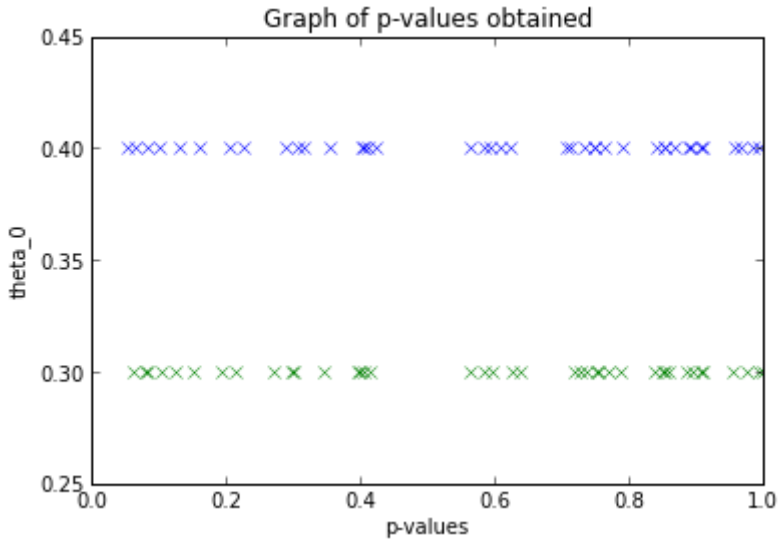


Figure 5: For each of our 40 generated sets of $W_1, \dots, W_n$, we compute the p-values where the distributions of $D|(\theta_0 = 0.3)$ and $D|(\theta_0 = 0.4)$ have been estimated by Monte Carlo techniques.

We would have hoped for the crosses of each color to cluster very closely around a certain p-value. It would then imply that our choice of $W_1, \dots, W_n$ is not as important as the original data $Y_1, \dots, Y_n$ in deciding what the ultimate p-value would be. We can then somewhat unambiguously assign a p-value to the original data $Y_1, \dots, Y_n$. With this p-value (one for each θ_0), we can go about doing our analysis in the usual way (e.g. following Lilliefor). Unfortuantely, this is not the situation here, and we may attribute this to the relatively wide error bars of data point 1, which results in a large amount of artificial noise that has to be added to all the other data points for the sake of i.i.d. data. This large amount of artificial noise, of course, dilutes the information present in our original data. We would be hesitant to draw conclusions on whether the uniform assumption should be rejected or not.

3 Approach II – Nonparametric Bayesian Analysis

In a more general context, we might not know the family of distribution which R belongs to. In this case, we propose a Bayesian approach to study the probability distribution of R. Crudely speaking, we begin with a certain prior distribution of R, proceed to incorporate the data, and obtain a posterior distribution of R. One could use the posterior as an estimate of the actual distribution of R and examine $P(R > c \mid data)$ or one could compare the prior and posterior (i.e. $P(R > c)$ with $P(R > c \mid data)$) to see if the data lends support for or against the upper bound.

3.1 Theory

Let $PM(X)$ be the set of all possible probability measures on X. Note that $PM(X)$ is an infinite-dimensional space. In reality, the actual distribution of X is one single element of $PM(X)$. However, when we are unsure about the actual distribution of X, the Bayesian approach would be to assign some prior probabilities to each element of $PM(X)$ as a belief that it could be the actual distribution of X. The **Dirichlet Process** is a way to assign probabilities to each probability distribution in $PM(X)$.

3.1.1 The Stick-Breaking Construction

Before explaining how the DP assign probabilities to probability distributions, let us examing the following stick-breaking process. First, fix $\alpha > 0$ and a probability distribution h. The output of this stick-breaking construction will be a probability distribution g. The process is as follows

1. Grab a stick of unit length.
2. Let $B_1, B_2, \ldots \sim i.i.d.\, Beta(1, \alpha)$. Let $P_1, P_2, \ldots \sim h$
3. Break off proportion B_1 of the stick and place it vertically at the x-position P_1.
4. Break off proportion B_2 of the remaining stick and place it vertically at the x-position P_2.
5. And so on.

Since $B_1, B_2, \ldots, P_1, P_2, \ldots$ are random, the output g is not deterministic – it could have been either of the two graphs below (figure 6), for example.

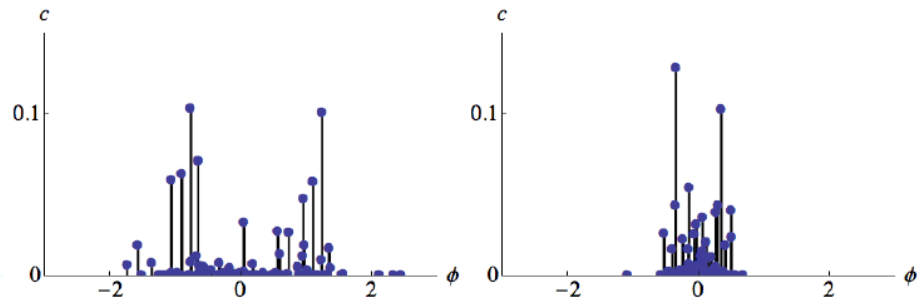


Figure 6: Just two of the infinitely many possible output g from our stick-breaking construction. Source of figure: [5]

For a fixed choice of a, h, there is a certain probability[4] that g turns out to be the first graph, and another probability that g turns out to be the second graph, another probability that g turns out to be some other graph, and so on. We write this probability distribution on probability distributions assigned by this stick-breaking process by

$$g \sim DP(\alpha, h)$$

It can be shown (see [6]) that in expectation, g will be the same as h.

3.1.2 Performing Inference I

Let us consider the following set-up (where $\alpha > 0$, distribution h are known inputs):

$$g \sim DP(\alpha, h)$$
$$X_1, X_2, \ldots \mid g \sim g \ \ (i.i.d.)$$

The first line is simply saying that we have assigned probabilities to probability distributions according to the stick-breaking process. The second line is saying that given the output g, which is a specific probability distribution, all the $X_1, X_2, \ldots$ are drawn i.i.d. from that probability distribution.

This set-up is handy for the following inference situation. Suppose $X_1, X_2, \ldots \sim R$ where R is of unknown distribution and suppose for the moment that we could observe $X_1, \ldots, X_n$. Although R is unknown, we could assign prior probabilities to possible probability distributions through the choice of some α, h and write the set-up above. Now using $X_1, \ldots, X_n$, we would like to obtain posterior probabilities to possible probability distributions. The following useful result is given in literature (see for example, [6]):

Fact 3.1: The posterior probabilities over possible probability distributions is also described by DP:

$$g \mid X_1, \ldots, X_n \sim DP\left(\alpha + n, \frac{1}{\alpha + n}\left[\alpha h + \sum_{i=1}^{n} \delta_{X_i}\right]\right)$$

where δ_a represents the dirac delta function located at a (that is, the infinite spike located at a of strength 1). ■

The fact that the posterior is also described by DP makes the choice of DP appealing (basically, the DP is its own conjugate prior). This explains why we chose to assign our prior probabilities using a DP. After all, realize that even if we look at all possible choices of α and h, the DP does not capture all possible ways of assigning prior probabilities. For example, any continuous probability distribution will always get a prior probability of zero, because the output g always have only countably many values of x for which $g(x) \neq 0$.

[4] This probability is actually zero, just like $P(X = x) = 0$ for any continuous random variable X. So I should say something like "probability that it looks very similar to the first graph", and basically describe the "probability" using some sort of density function.

Now, the DP is an artificial construct to help us assign prior probabilities to probability distributions. The posterior that we are truly interested in is $X_{n+1} \mid X_1, \dots, X_n$ (equivalently, $R \mid X_1, \dots, X_n$). The following useful result is also given in literature.

Fact 3.2: The posterior predictive distribution is given by

$$X_{n+1}|X_1, \dots, X_n \sim \frac{1}{\alpha + n}\left[\alpha h + \sum_{i=1}^{n} \delta_{X_i}\right]$$

where δ_a represents the dirac delta function located at a. ■

3.1.3 Dirichlet Mixture Model

Unfortunately, we do not observe $X_1, \dots, X_n$. Instead, we observed $Y_1, \dots, Y_n$ which we can describe with the following set-up (where $\alpha > 0$, distribution H, μ_i, σ_i are known inputs):

$$\begin{aligned} &g \sim DP(\alpha, h) \\ &X_1, X_2, \dots \mid g \sim g \ \ (i.i.d.) \\ &Y_i \mid X_i \sim N\left(X_i + \mu_i, \sigma_i^2\right) \ (independent\ of\ X_j, Y_j\ for\ j \neq i) \end{aligned}$$

We call this set-up the Dirichlet Mixture Model because when we are **not** given X_i, then Y_i can be seen as being drawn from a mixture of normals, with the mixing eventually described via the Dirichlet process.

3.1.4 Performing Inference II

In our statistical formulation (of section 1.2), the posterior predictive distribution that we are interested in is $X_{n+1}|Y_1, \dots, Y_n$. We use the DP set-up above to assign prior probabilities to probability distributions, apply Bayesian analysis on the data to obtain posterior probabilities to probability distributions, and then integrate over all possible probability distributions to obtain the posterior predictive distribution for R. This is what should happen in theory, but we will use a numerical approximation method instead along with fact 3.2.

Let $\vec{x} = (x_1, \dots, x_n)$ and $\vec{y} = (y_1, \dots, y_n)$. For a fixed x_{n+1} and $\vec{y}$, we like to find

$$\begin{aligned} &f(x_{n+1} \mid \vec{y}) \\ &= \int f(x_{n+1} \mid \vec{x}, \vec{y}) f(\vec{x} \mid \vec{y})\, d\vec{x} \\ &= \int f(x_{n+1} \mid \vec{x}) f(\vec{x} \mid \vec{y})\, d\vec{x} \\ &= \mathbb{E}_{\vec{x}}[f(x_{n+1} \mid \vec{x}) \mid \vec{y}] \end{aligned}$$

We can apply a Monte Carlo approximation and compute

$$\frac{1}{S}\sum_{s=1}^{S} f\left(x_{n+1} \mid x_{s,1}, \dots, x_{s,n}\right)$$

where $\left(x_{s,1}, \dots, x_{s,n}\right) \sim X_1, \dots, X_n | Y_1, \dots, Y_n$.

We will use Gibbs Sampling to approximately obtain S independent samples of

$$(x_{s,1}, \dots, x_{s,n}) \sim X_1, \dots, X_n | Y_1, \dots, Y_n$$

where Gibbs Sampling propose that we may obtain one sample as follows:

1. Start with any initial $(x_{0,1}, \dots, x_{0,n})$.
2. For $s = 1,2, \dots$ and for $1 \leq i \leq n$, we sample $x_{s,i}$ from the conditional distribution
$$f(x_i \mid x_{s,1}, \dots, x_{s,i-1}, x_{s-1,i+1}, \dots, x_{s-1,n}, y_1, \dots, y_n)$$

So basically, we update the current vector to obtain the next vector one term at a time. The idea is that we have a Markov chain with states $(x_{s,1}, \dots, x_{s,n})$ where the stationary distribution over the states is precisely the desired distribution. Thus, if we let this Markov Chain run long enough, the state that we obtain is akin to sampling from the desired distribution.

In practice, to obtain S independent samples instead of one, we would run the Markov Chain for a very long time and:

- Ignore the first k_1 vectors to give the process some time to move from an initial choice to actual sampling from the desired distribution (the "burn-in" period)
- Consider only every k_2 vectors to ensure independence between samples ("thinning")

The choices of k_1 and k_2 involve more or less some "black magic".

To complete our description of the method, we should figure out how to compute
$$f(x_i \mid x_{s,1}, \dots, x_{s,i-1}, x_{s-1,i+1}, \dots, x_{s-1,n}, y_1, \dots, y_n)$$

This is done by observing that, by Bayes' Rule,

$$\begin{aligned} & f(x_i \mid x_{-i}, y_1, \dots, y_n) \\ & = f(x_i \mid x_{-i}, y_i) \\ & \propto f(y_i \mid x_i, x_{-i}) f(x_i \mid x_{-i}) \\ & = f(y_i \mid x_i) f(x_i \mid x_{-i}) \\ & \propto f(y_i | x_i) \left[\alpha h(x_i) + \sum_{j \neq i} \delta_{x_j}(x_i) \right] \end{aligned}$$

where we used fact 3.2 in the last line. The normalizing constant is

$$\begin{aligned} C &:= \int f(y_i | x_i) \left[\alpha h(x_i) + \sum_{j \neq i} \delta_{x_j}(x_i) \right] dx_i \\ &= \sum_{j \neq i} f(y_i \mid x_j) + \alpha \, \mathbb{E}_{x_i}[f(y_i | x_i)] \end{aligned}$$

So to sample x_i from $f(x_i \mid x_{-i}, y_i)$, we do the following: with mutually exclusive probabilities $f(y_i | x_j)/C$ (where $j \neq i$), we will set x_i to equal x_j. Otherwise, we sample x_i from the density

proportional to $f(y_i|x_i)h(x_i)$. In our implementation, sampling from $\propto f(y_i|x_i)h(x_i)$ is done using a standard Metropolis-Hastings algorithm.

3.2 Implementation Results

To implement our theory, we have to pick a prior α and h. For h, a reasonable choice would be to use the old Kernel Density Estimate constructed using a set of older data (see [1]) and using the procedure described in section 4.1. The data points used for the prior are compiled in Appendix C. For α, we picked an arbitrary choice of $\alpha = 1$.

Our implementation is largely approximation based (or perhaps more accurately, numerically based). In our implementation, we made the following choices. The theoretically ideal situation, of course, would be to set all of these parameters to ∞.

Implementation Parameter	Choice
Number of Monte Carlo Samples S	500
"burn-in" period for Gibbs sampling k_1	500
"thinning" parameter for Gibbs sampling k_2	300
Metropolis-Hastings "discard" parameter	100

Figure 7 shows the prior distribution vis-à-vis the posterior. Figure 8 shows the posterior distribution with the x-axis zoomed in. The probability of $R > c$ has shrunk significantly after incorporating the data. In fact, $P(R > c \mid data)$ is merely 0.00875 – the new data supports the upper bound.

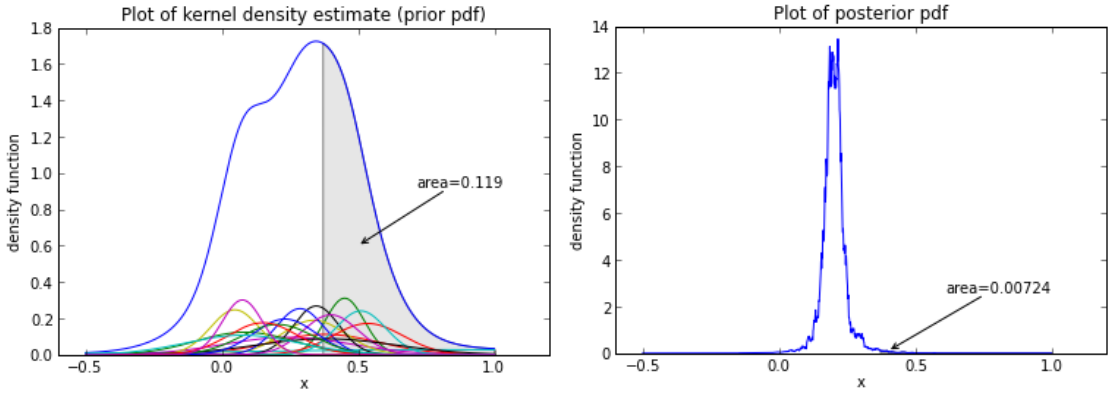


Figure 7: Comparing the prior distribution with the posterior distribution.

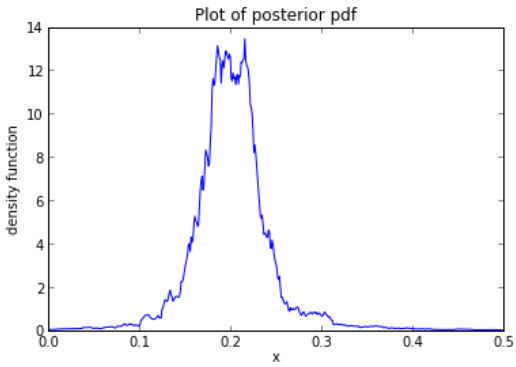


Figure 8: Posterior distribution with x-axis zoomed in.

4 Miscellaneous Techniques

In this section, we document other miscellaneous techniques for our statistical problem.

4.1 Kernel Density Estimation

The Kernel Density Estimation is a way of estimating unknown probability density function from the data. Given $X_1, \dots, X_n$ drawn i.i.d. (and measured precisely) from the unknown distribution with pdf f, the kernel density estimator is an estimator of f given by

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x - x_i}{h}\right)$$

where K is some choice of kernel (whose purpose is sort of to smooth out the data point into a distribution) and $h > 0$ is some choice of smoothing parameter, also called bandwidth. Popular choices of K include for example the normal, triangular or Epanechnikov distribution. The choice of K as well as h has been subjected to extensive study. While there are no "correct" choices, there are many recommended approaches to make these choices.

For our situation, there is a rather canonical choice. The errors, which follow $N(\mu_i, \sigma_i^2)$, already provide us the smoothing function performed by a kernel. In other words, our kernel density estimator would be

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sigma_i}\phi\left(\frac{x - y_i + \mu_i}{\sigma_i}\right)$$

where ϕ is the standard normal pdf. Note that this is a slight modification of the original KDE approach because we allow the choice of h to vary between data points to take into account the varying uncertainties. The constructed $\hat{f}$ is shown in figure 9.

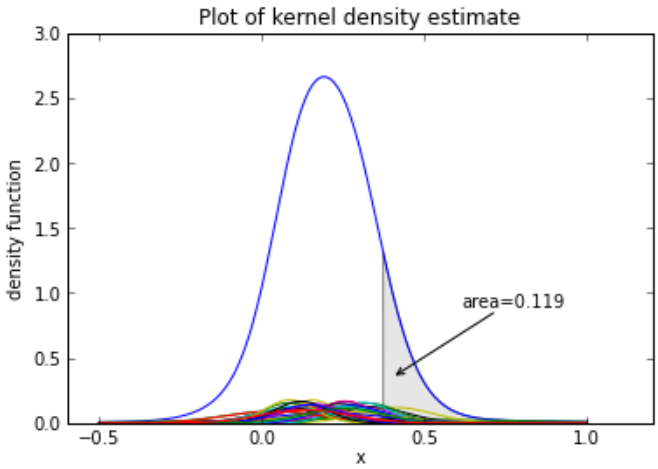


Figure 9: Constructed kernel density estimator. The rainbow of tiny humps represents the 30 data points smoothed out by the uncertainties of the errors.

In the original Kernel Density Estimation theory, it can be shown that under certain regularity conditions that are not too restrictive, $\hat{f} \to f$ uniformly with probability 1. For example, see [4]. We might expect similar asymptotic behavior to hold in our situation, though we will not provide a formal justification.

Note that $\hat{f}$, even if it well-approximates f for large n, could never be the actual distribution f itself. For one thing, R is constrained to be non-negative, but $\hat{f}$ will always have positive area in the region $x < 0$ by construction. Moreover, $\hat{f}$ will also have positive area in the region $x > x_0$ for any x_0, so if we had treated $\hat{f}$ as our actual distribution and test c as an upper bound, we will definitely reject our hypothesis (because a true upper bound means $P(R > c) = 0$). The point is we have to use $\hat{f}$ merely as an approximation. If we believe $\hat{f}$ to be a uniformly excellent approximation of f, we could compute $\int_c^\infty \hat{f}(x)dx$ and observe if it is close to zero. There is a slightly better approach that adjusts for the possibility that $\hat{f}$ is not yet a good approximation for our sample size. Since we know $\int_{-\infty}^0 f(x)dx = 0$ (notice: there is no hat in the f), we can look at $\int_{-\infty}^0 \hat{f}(x)dx$ as a rough indication of how well $\hat{f}$ is approximating f. We should then examine $\int_c^\infty \hat{f}(x)dx$ relative to $\int_{-\infty}^0 \hat{f}(x)dx$. For our data,

$$\int_c^\infty \hat{f}(x)dx = 0.119 \; and \; \int_{-\infty}^0 \hat{f}(x)dx = 0.090$$

and so the hypothesis that $\int_c^\infty \hat{f}(x)dx \to 0$ does not seem too far-fetched.

Of course, the above approach is more of a heuristic treatment than a theoretically-rigorous analysis from which firm conclusions can be drawn.

4.2 Fisher's Method

Here is one attempt to tackle the original motivating problem from physics. The hypothesis is that $R \leq c$ always, and we may see the measurement of R for each bin of Q^2 as an "independent test" of this hypothesis. This sounds like a possible application of Fisher's method. Loosely speaking, for each data point, we would compute the probability that the bound is respected and then combine the information using the Fisher's method to obtain an overall p-value.

In more rigorous terms, we are conducting the following test, which unfortunately is much weaker than what we wish to verify. The null hypothesis is $H_0: R = c$ (note, this is not our grand null hypothesis $\mathcal{H}$), tested against the alternate hypothesis $H_1: R > c$. Under the null hypothesis, $Y_i|H_0 \sim N\left(c + \mu_i, \sigma_i^2\right)$ independently. By setting $p_i \coloneqq P(Y_i \geq y_i|H_0)$, we have – as we expect of p-values – that $p_i \sim Unif[0,1]$ independently under the null hypothesis, and so our test statistic

$$p \coloneqq -2\sum_{i=1}^{n} \ln p_i \sim \chi_{2n}^2$$

from which we could compute our overall p-value.

Remark: How does this coincide our initial intuition of computing the probability that the bound is respected? Under assumption 2 (and the issue raised in footnote 1), $X_i|Y_i \sim N\left(Y_i - \mu_i, \sigma_i^2\right)$ and so $P(X_i \leq c\,|Y_i = y_i) = P(Y_i \geq y_i|H_0)$ as one can verify. However, we prefer the rigorous formulation above because the analysis in this remark, strictly speaking, does not make sense. In the context of assumption 2, the frequentist error bars were constructed without being constrained by $\mathcal{H}$ and we treated X_i as a parameter to be estimated. We then made the Bayesian statement $X_i|Y_i \sim N(Y_i - \mu_i, \sigma_i^2)$. Worse still, we then bring $\mathcal{H}$ back as a null hypothesis when we ask for $P(X_i \leq c\,|Y_i = y_i)$. ■

For our actual data, the p-values have been tabulated in Appendix B and with an observed test statistic of $p = 8.82$, our p-value is 100% (it is actually 99.99999…%). This should not come as a surprise because our test is an extremely weak one. R is hypothesized to be distributed in $[0, c]$ and so many data points will be significantly below c and contribute strongly to protect the null hypothesis. However, note that even if we (illegally) select just the points lying above c (there is actually only one such point), the p-value is still as high as 36.0%. Our quick test thus shows that the data presents really little evidence against the bound.

5 References and Acknowledgements

[1] Ewerz, Carlo *et al.* **Bounds on Ratios of DIS Structure Functions from the Color Dipole Picture.** Phys.Lett. B648 (2007) 279-283 hep-ph/0611076 ECT-06-18, HD-THEP-06-29

[2] Ewerz, Carlo *et al.* **The New F_L Measurement from HERA and the Dipole Model.** Phys.Lett. B720 (2013) 181-187 arXiv:1201.6296 [hep-ph] ZU-TH-02-12, MZ-TH-12-06

[3] ZEUS Collaboration (Abramowicz, H. *et al.*) **Deep inelastic cross-section measurements at large y with the ZEUS detector at HERA.** arXiv:1404.6376 [hep-ex] DESY-14-053

[4] Wied, Dominik; Weißbach, Rafael. **Consistency of the kernel density estimator - a survey**. Statistical Papers (2010), ISSN 0932-5026, Vol. 53, Iss. 1, pp. 1-21, http://dx.doi.org/10.1007/s00362-010-0338-1

[5] Orbanz, P. **Lecture Notes on Bayesian Nonparametrics.** (2014). http://stat.columbia.edu/~porbanz/talks/npb-tutorial.html

[6] Orhan, E. **Dirichlet Process**. (2012). http://www.cns.nyu.edu/~eorhan/notes/dpmm.pdf

I would also like to thank the following people:

- Dr David d'Enterria, supervisor, for his helpful guidance
- Avi Feller, for helpful discussions early on in the project

Appendix A – Data Points

The 30 data points from H1 and ZEUS collaborations.

	Data Point	Q^2	R_{obs}	R_{minus}	R_{plus}
H1	1	1.5	0.16072	0.2967440	0.5584825
	2	2.0	0.11137	0.1124525	0.1671691
	3	2.5	0.12813	0.0849772	0.1277879
	4	3.5	0.28359	0.0854678	0.1176146
	5	5.0	0.39733	0.0973186	0.1281531
	6	6.5	0.32394	0.0953105	0.1224058
	7	8.5	0.26271	0.0920121	0.1215727
	8	12.0	0.31603	0.0754044	0.0990038
	9	15.0	0.24344	0.0669440	0.0914325
	10	20.0	0.29723	0.0722171	0.0955846
	11	25.0	0.24095	0.0680745	0.0944573
	12	35.0	0.14376	0.0645117	0.0873687
	13	45.0	0.22429	0.0752226	0.1045976
	14	60.0	0.25154	0.0816530	0.1159686
	15	90.0	0.06879	0.0695198	0.1025077
	16	120.0	0.15329	0.0896861	0.1158746
	17	150.0	0.24382	0.0954648	0.1159492
	18	200.0	0.12918	0.1064739	0.1196202
	19	250.0	0.23573	0.1399603	0.1620137
	20	346.0	0.10660	0.1239695	0.1066737
	21	636.0	0.24786	0.2144543	0.1220300
ZEUS	22	9.0	0.01	0.13	0.17
	23	12.0	0.01	0.12	0.17
	24	17.0	0.09	0.10	0.11
	25	24.0	0.14	0.07	0.11
	26	32.0	0.08	0.07	0.08
	27	45.0	0.10	0.06	0.10
	28	60.0	0.14	0.09	0.10
	29	80.0	0.25	0.12	0.12
	30	110.0	0.10	0.10	0.16

Appendix B – p-values for Fisher Method

The p-values obtained for section 4.2.

Data Point	p-value
1	0.575019433
2	0.952729406
3	0.981945068
4	0.763348072
5	0.360477668
6	0.626064945
7	0.813127406
8	0.695681721
9	0.929881983
10	0.775665345
11	0.927329964
12	0.997890795
13	0.931207464
14	0.853212727
15	0.999579529
16	0.977529473
17	0.868695438
18	0.981872503
19	0.797486349
20	0.991356402
21	0.845041088
22	0.988790705
23	0.990029291
24	0.995887201
25	0.990884404
26	0.999936722
27	0.999200256
28	0.991679075
29	0.846293804
30	0.968925468

Appendix C – Older Data Points

The 18 data points found in [1] and used for constructing the prior in section 3.2.

Data Point	Q^2	R_{obs}	R_{error}
1	0.40	0.451	0.0713
2	0.59	0.379	0.198
3	0.925	0.814	0.802
4	0.935	0.244	0.232
5	1.40	0.338	0.117
6	1.49	0.347	0.0827
7	1.49	0.287	0.0874
8	1.50	0.0851	0.179
9	1.46	0.536	0.129
10	1.56	0.510	0.0920
11	1.77	0.400	0.101
12	2.90	0.0483	0.0897
13	2.92	0.384	0.253
14	4.3	0.232	0.113
15	6.35	0.205	0.133
16	6.37	0.152	0.126
17	12.55	0.0759	0.207
18	15.85	0.0759	0.0736