Advanced Methods in Applied Statistics
Exam 2023

Philip Kofoed-Djursner
(Tkv976)

31th of March 2023

I, Philip Kofoed-Djursner, expressly vow to uphold my scientific, academic, and moral integrity by working individually on this exam and soliciting no direct external help or assistance.

1 Problem 1

3.0/3.0

1.1 Problem 1a & 1b

For the first column in the data set, the data was found to be compatible with the function

$$f(x : a, b, c) = sin(ax) + ce^{bx} + 1. \tag{1}$$

All the parameters in this problem were found using a log-likelihood (llh) method. This function was ensured to be normalized by doing the definite integral and inserting the limits given in the problem. The same is true for the second function. For the first column, the found best-fit parameters were;

$$a = 3.899 \pm 0.002, \tag{2}$$

$$b = -0.365 \pm 0.013, \tag{3}$$

$$c = 7100 \pm 1900. \tag{4}$$

For the second column of the data set, the compatible function was

$$f(x : a, b) = 1 + ax + bx^2. \tag{5}$$

The found best-fit parameters were;

$$a = 0.47 \pm 0.05, \tag{6}$$

$$b = 2.58 \pm 0.16. \tag{7}$$

For the third column of data, two compatible functions were found. Both the Poisson distribution and the Binomial distribution can fit the data well. This is of course because the distributions become the same in the limit $(1 - p)np \approx \lambda$. The Poisson distribution results were the ones chosen to be presented here, as the Binomial did not seem to have a unique best fit, as any values of n and p would lead to a good fit as long as the limit above holds.

$$f(k : \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}. \tag{8}$$

The found best-fit parameter was;

$$\lambda = 9.18 \pm 0.04. \tag{9}$$

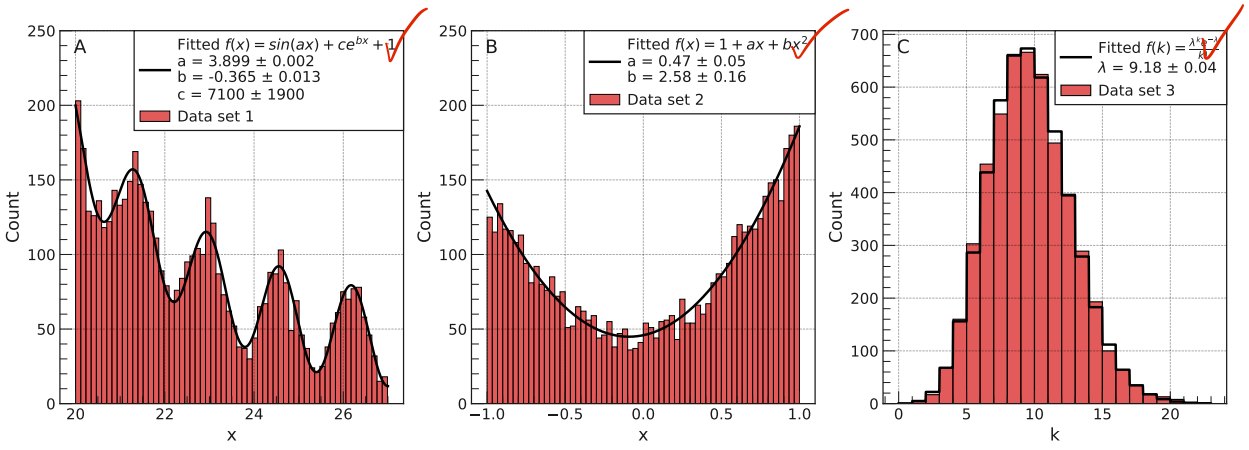

Figure 1: Datasets and fitted functions. **A** is column 1 for the dataset and was fitted to Equation 1. **B** is column 2 for the dataset and was fitted to Equation 5. **C** is column 3 for the dataset and was fitted to Equation 8. The Poisson distribution was plotted as continuous even though it is discrete for better visibility. The PDFs are scaled to match the histograms.

The values and errors quoted were computed by the Migard method implemented through IMinuit.

In Figure 1, the fitted functions and the data are shown. I scaled the PDFs to match the data. To get a measure of the goodness of fit, a χ^2 p-value was calculated from the fitted parameters of the three functions. Degrees of freedom were defined as the number of bins in the plot minus the parameters fitted in the llh fit.

$$p_1 = 0.66, \tag{10}$$

$$p_2 = 0.76, \tag{11}$$

$$p_3 = 0.90. \tag{12}$$

The p-values are denoted by the dataset they correspond to. To get the errors on the histogram, Poisson errors were assumed. All fits are in great agreement with the data.

The first function (Equation 1) was pretty tricky to fit and required quite a lot of hand-tuning of initial parameters to get a good result. Especially "a" was hard to find a good value for. This is simply because the function has a lot of local minima when fitted to the data. These features can be seen from the Raster scan in Figure 2. The Raster scan was for the "a" and "b" variables, as "c" seemed to be the least hard to fit.

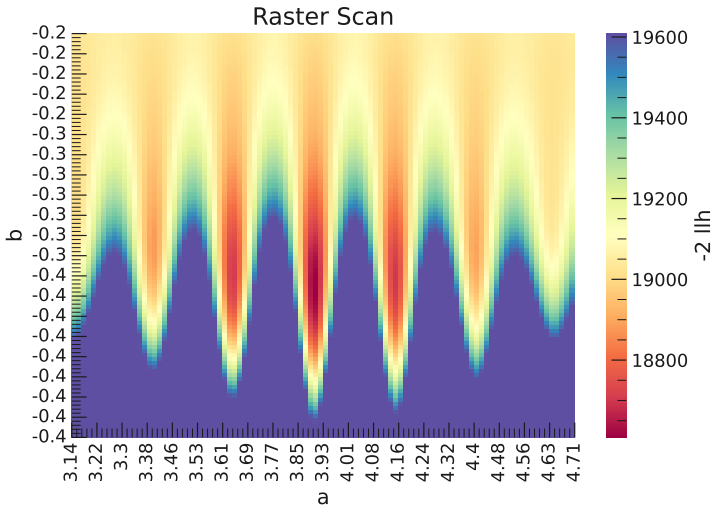

Figure 2: Raster scan of a and b in Equation 1 and c = 6000. The color bar only ranges from the -2 llh minimum to 1000 more. Therefore, structure does exist within the blue region but that is not of interest.

2 Problem 2

2.1 Problem 2a

A good way to visualize data on a globe is to use the Mollweide projection. In Figure 3, the 139 provided data points are shown.

To quantify if the data is isotropic, a perfect isotropic standard must be created to compare the data and pseudo-experiments against. In Figure 4, the 2000 isotropic Monte Carlo points are shown. To compare the distributions auto-correlation is done for both the standard, data, and pseudo-experiments. the Auto-correlation was defined as

$$C(\{\mathbf{n}_i\},\varphi) = \frac{2}{N_{tot}(N_{tot}-1)} \sum_{i=1}^{N_{tot}} \sum_{j=1}^{i-1} \Theta(cos\varphi_{ij} - cos\varphi). \tag{13}$$

Where $cos\varphi_{ij} = n_i \cdot n_j$ is the angle between two points in Cartesian coordinates, Θ is the step function, and φ is the angle wherein clustering is searched for. In my implementation, I modified it slightly to sum over the entire matrix. This computes more sums but the calculation of $cos\varphi_{ij}$ was much faster, so it was a net gain. $cos\varphi$ was computed at 300 points equally spaced between -1 and 1. The standard, data,

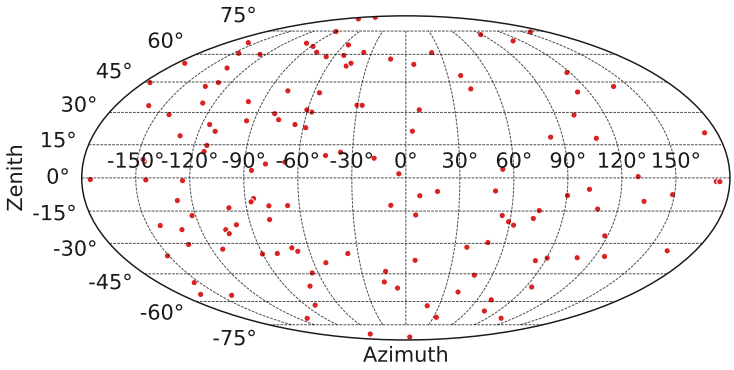

Figure 3: The data points plotted on a Mollweide projection

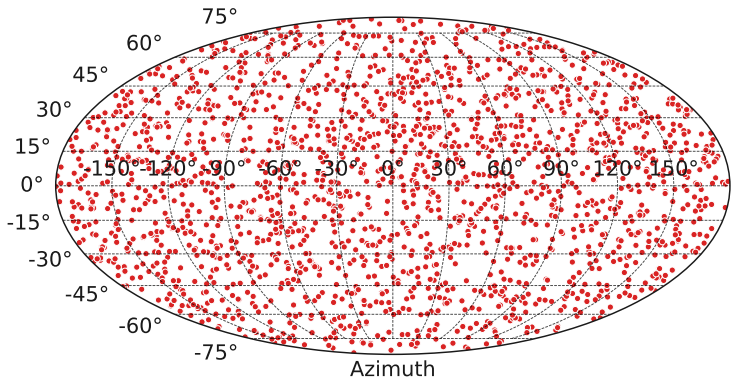

Figure 4: All 2000 isotropic Monte Carlo points used for the background plotted on a Mollweide projection

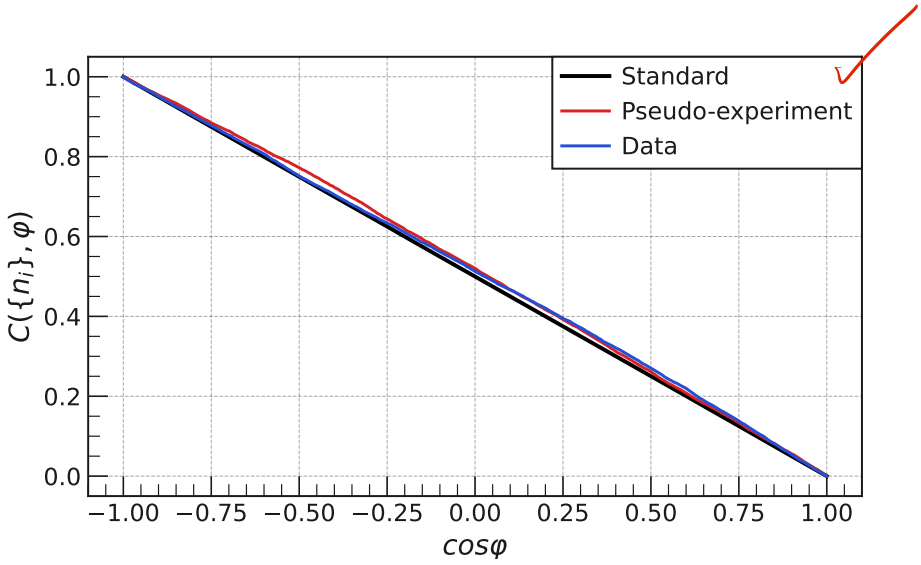

Figure 5: Auto correlation results for the standard (background), the data, and a single isotropic pseudo-experiment.

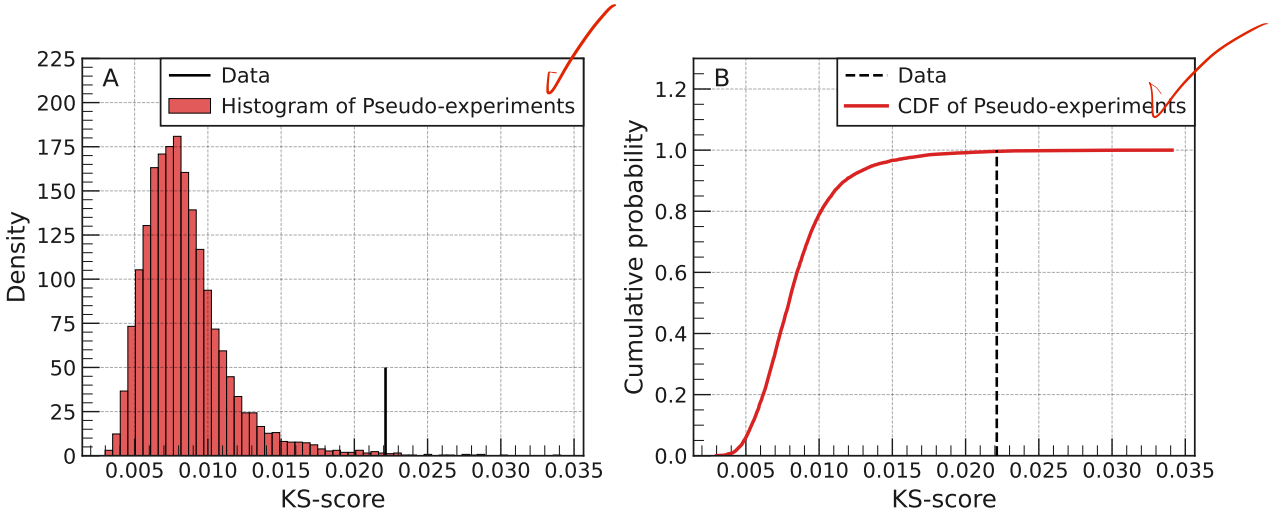

Figure 6: Results of the Kolmogorov-Smirnov test on all 5000 isotropic pseudo-experiments. **A** Histogram of the KS-scores resulting from the pseudo-experiments and the data's score drawn as a line. **B** CDF of the KS-scores and the data's score indicated as a line.

and one pseudo-experiment's auto-correlation results are shown in Figure 5. It can be seen that both the data and the pseudo-experiment's auto-correlation deviates from the standard, so to quantify this deviation a Kolmogorov-Smirnov (KS) test was done. The KS-score can be compared between the data and the pseudo-experiments to see how likely it is they come from the same underlying hypothesis.

In Figure 6A, a histogram of the KS scores for pseudo-experiments is shown and also includes the value for the data. 5000 pseudo-experiments were done each of 139 points. In Figure 6B, the same data is shown but as a cumulative distribution function (CDF). From this, I calculated the p-value that the data was isotropic which was p = 0.002. This is close to a 3σ deviation from the expected value. It is therefore highly unlikely that the data is isotropic.

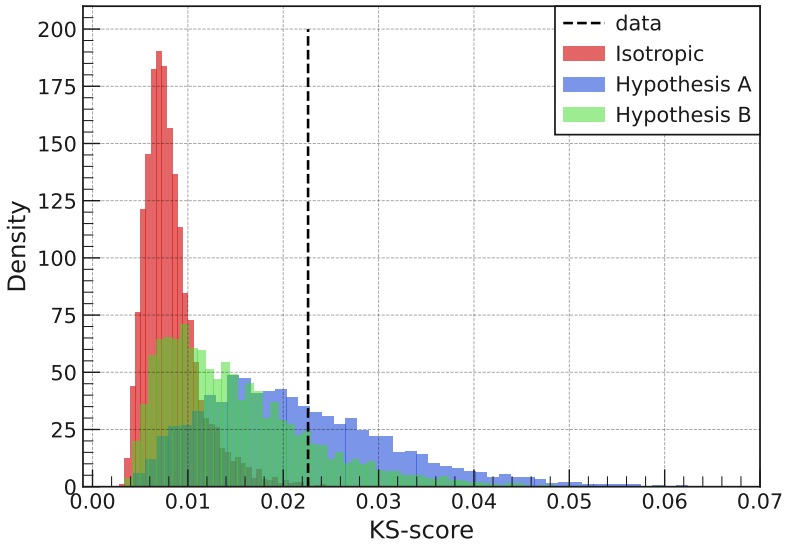

Figure 7: Results of the Kolmogorov-Smirnov test on all 5000 isotropic pseudo-experiments, 5000 pseudo-experiments following hypothesis A, and 5000 pseudo-experiments following hypothesis B against the isotropic background. The data's KS score is shown as a line

2.2 Problem 2b

0.9/1.0

In problem 2b the same ideas are employed, but the pseudo-experiments are made from different distributions. The two alternative hypotheses were tested and found

$$P_{Hyp\ A} = 0.39. \tag{14}$$

$$P_{Hyp\ B} = 0.15. \tag{15}$$

Allowing an arbitrary confidence level of 5%, both hypotheses describe the data. In Figure 7, a histogram of the KS-scores is shown. The p-value is calculated as the probability of getting the same or larger deviation from isotropy.

The only thing you missed was a mention of needed post-trial corrections because you are testing 3 hypotheses.

-0.1

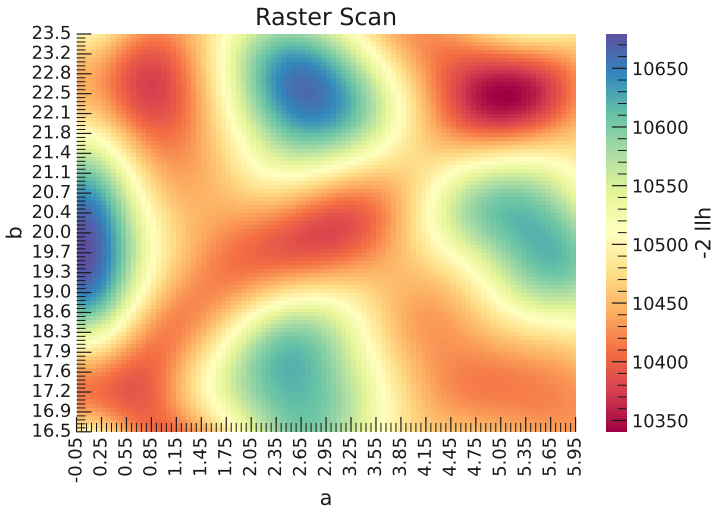

Figure 8: Raster scan perform after the first fit. Ranges are from $a = 2.95 \pm 3$ and $b = 20.00 \pm 3.5$.

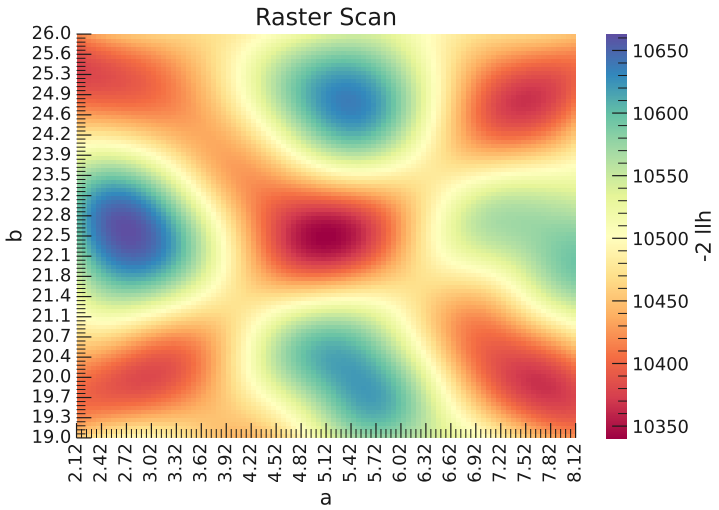

Figure 9: Raster scan perform after the second fit. Ranges are from $a = 5.12 \pm 3$ and $b = 22.45 \pm 3.5$. This Raster scan is centered on the global minimum

3 Problem 3

First, the function was fitted to the data and the found best-fit parameters were

$$a = 2.95 \pm 0.12, \tag{16}$$

$$b = 20.00 \pm 0.10. \tag{17}$$

In Figure 8, is shown the first raster centered on the above-found parameters. It can be seen that the parameters do lie in a negative log-likelihood minimum, but it is not the global minimum. At around $a = 5$ and $b = 23$ a better minimum can be seen. So I reran the fitting with starting parameters closer to the better minimum. The new best-fit parameters were

$$a = 5.12 \pm 0.09, \tag{18}$$

$$b = 22.45 \pm 0.08. \tag{19}$$

In the specified range around these values, a new raster scan was done, which is shown in Figure 9. Within the scan range, this is the best minimum. I also did a full scan of the parameter space and found this to be the global minimum. Still, the parameter land space does contain a lot of local minima.

To see how the fitted function looked on the data I also plotted that in Figure 10. The χ^2 p-value was $p = 0.82$ with assumed Poisson errors. The fit seems to be in great agreement with the data.

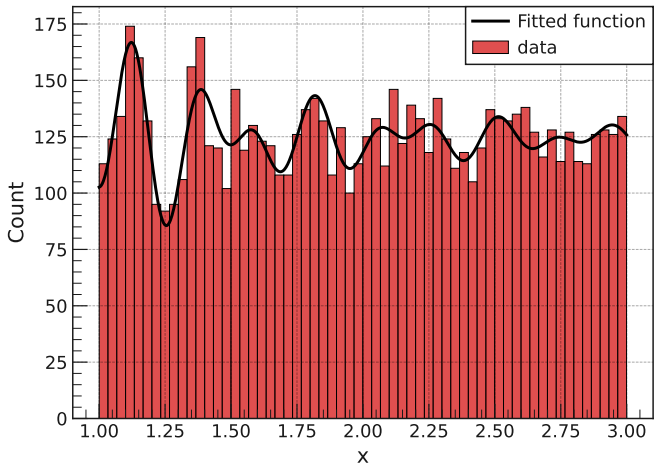

Figure 10: Data set plotted as a histogram with the function with best-fit parameters plotted on top.

4 Problem 4

4.1 Problem 4a

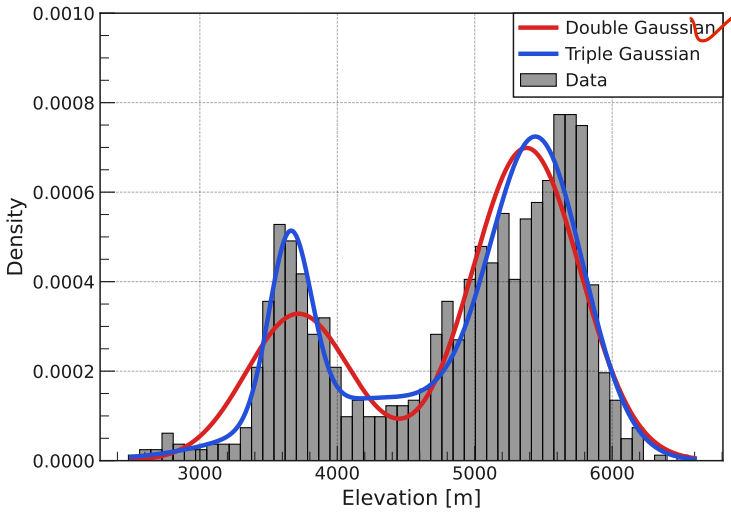

Figure 11: Histogram of data with the double Gaussian and triple Gaussian plotted on top.

In Figure 11 is shown the data as a histogram, the double Gaussian and the triple Gaussian. The two functions are fitted to the data using unbinned log-likelihood fits. To ensure the PDFs were normalized they were defined as

$$f_{Double\ Gaussian}(x) = (1 - P_2)g(x : \mu_1, \sigma_1) + P_2 g(x : \mu_2, \sigma_2), \tag{20}$$

$$f_{Triple\ Gaussian}(x) = (1 - P_2 - P_3)g(x : \mu_1, \sigma_1) + P_2 g(x : \mu_2, \sigma_2) + P_3 g(x : \mu_3, \sigma_3). \tag{21}$$

Where $g(x)$ is a normalized 1-D Gaussian PDF. In Tables 1 & 2 the fitted parameters and their associated errors are shown. Both fits are quite bad and do not describe the features of the data. The calculated χ^2 p-values are

$$P_{Double\ Gaussian} = 6.5 \cdot 10^{-19}, \tag{22}$$

$$P_{Triple\ Gaussian} = 0.00028. \tag{23}$$

Table 1: Best fit parameters and error for the double Gaussian.

	P_2	μ_1	σ_1	μ_2	σ_2
Values	0.689	3720	380	5376	393
Errors (±)	0.016	30	20	18	14

Table 2: Best fit parameters and error for the triple Gaussian.

	P_2	P_3	μ_1	σ_1	μ_2	σ_2	μ_3	σ_3
Value	0.58	0.26	3658	155	5460	338	4290	750
Error (±)	0.04	0.04	18	15	20	19	130	60

4.2 Problem 4b

Wilk's theorem will be used to compare the two fits, so therefore, I assume that it holds.

$$D = -2 \cdot (llh(Double\,Gaussian) - llh(Triple\,Gaussian)) = 67.4. \tag{24}$$

Wilk's theorem states that the -2 times natural log of the likelihood ratios (log-likelihood difference) will follow a χ^2 distribution with degrees of freedom equal to the difference in the number of parameters. In this case, the difference in the number of parameters is 3.

$$\chi^2_{k=3}(x > 67.4) = 1.6 \cdot 10^{-14}. \tag{25}$$

So the probability that the likelihood ratio observed is by chance is $1.6 \cdot 10^{-14}$, so the triple Gaussian is a better model for the data.

5 Problem 5

5.1 Problem 5a

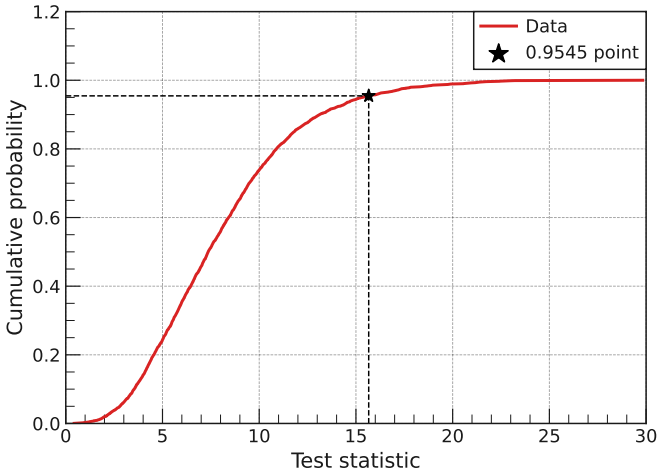

Figure 12: CDF of the test statistic and the point which defines the 0.9545 points. This is the same test statistic value which defines 0.0455 on the survival function.

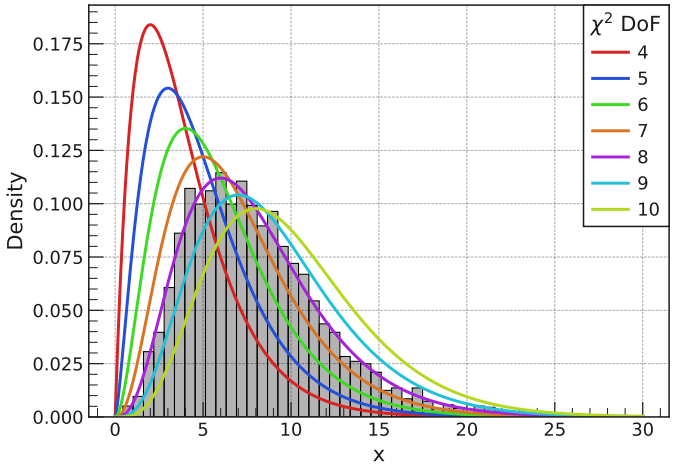

Figure 13: Histogram of the test statistic and χ^2 distributions with DoF ranging from 4 to 10.

In Figure 12, the CDF of the data is shown. From this, I calculated the critical value of the onesided 4.55% threshold. This comes out to be 15.7. To see if this statistical test is distributed as a χ^2 with k=5 I calculated the survival function from 15.7.

$$\chi^2_{k=5}(x > 15.7) = 0.008. \tag{26}$$

This is far from 0.0455, which suggests that they do not match the same distribution. In Figure 13, χ^2 distributions with a different number of degrees of freedom are shown. Here a χ^2 distribution with k=8 seems to fit the test extremely well. The distribution with k=5 does not fit the data. Therefore, I conclude that the test statistic does follow a χ^2 distribution, but with k=8 and not k=5.

5.2 Problem 5b

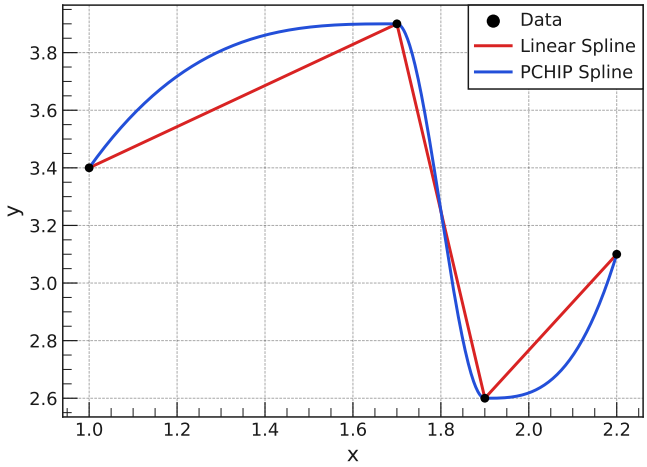

Figure 14: The data, the linear spline and the PCHIP spline plotted

In Figure 14, the data, a linear spline, and a PCHIP spline are plotted. From these splines, the y-value at x=2 can be interpolated.

$$Spline_{linear}(2) = 2.77, \tag{27}$$

$$Spline_{PCHIP}(2) = 2.62. \tag{28}$$

Which spline interpolates the data the best is impossible to know without more context and data.