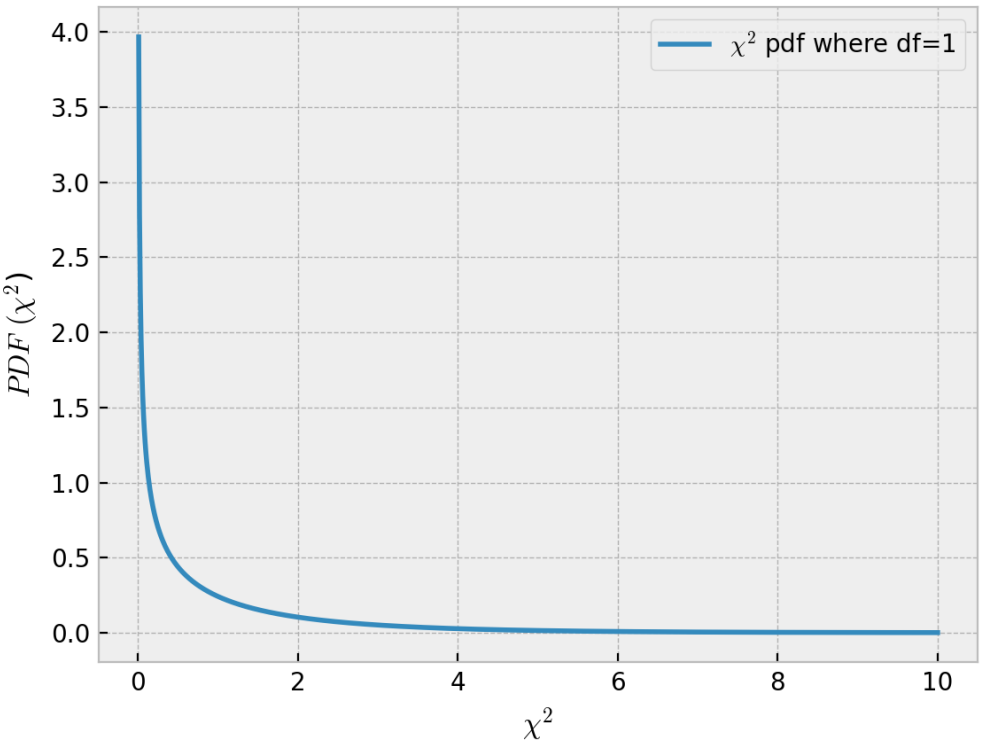
UNIVERSITY OF COPENHAGEN

Problem Set 2
Advanced Methods in Applied Statistics



Daniel Ramyar
NJT478

March 25, 2019

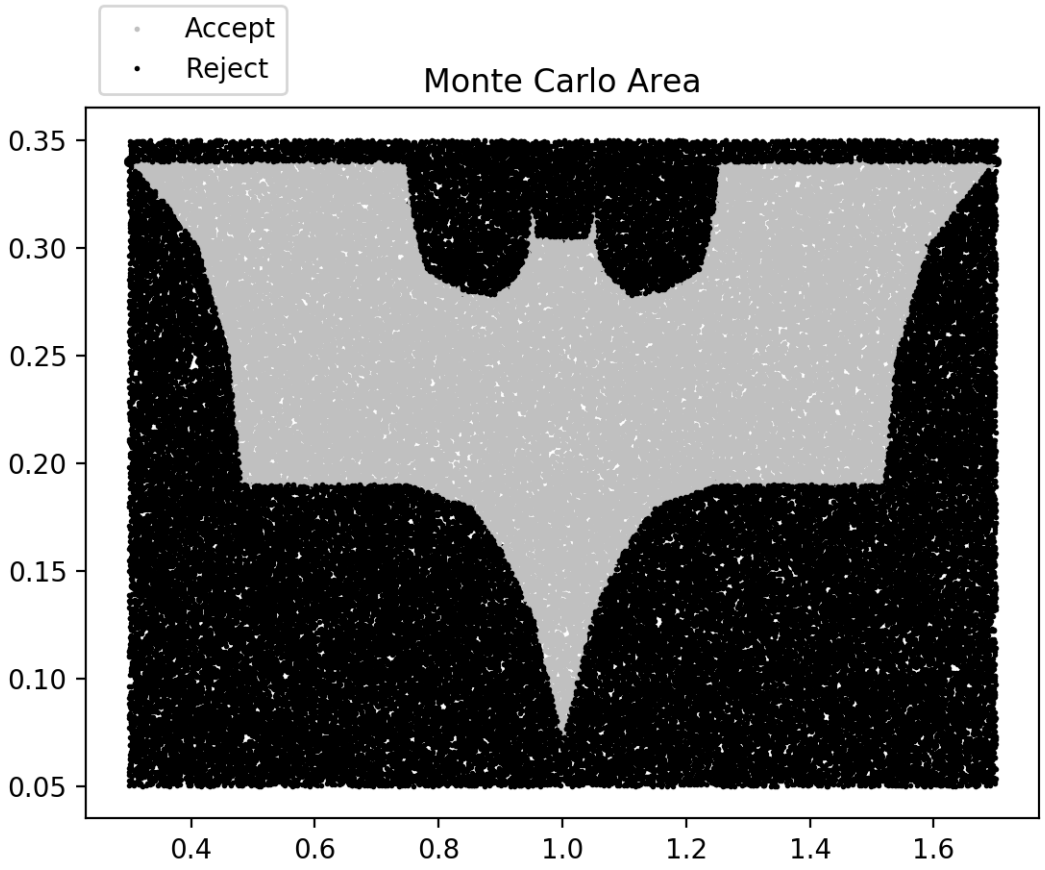


Figure 1: 10^5 generated points where the grey markers landed inside the object and the black outside

Problem 2

We are given a data set `ProblemSet2_Problem2.txt` with data points (x, y). We notice that the data points goes around in a circle so in order to do linear interpolation we separate the data points into an upper and lower part.

Next we generate 10^5 random numbers and check whether they fall inside or outside the enclosed area. The area is then calculated using by using the fact that the probability landing inside the area is $p = \frac{Areaofobject}{Areaenclosingobject}$ and we get p by looking at the ratio of points which landed inside and total generated points and the area enclosing the object is set to 1.4×0.3.

The area of the object is then estimated to be **0.16**. The generated points and object is shown in Figure 1.

Facility	Total % produced	% Defective	% Updated Defective	$\frac{P(D\|A_i)P(A_i)}{P(D)}$
A_1	35	2	2.2	21.4
A_2	15	4	5.2	18.3
A_3	5	10	15.5	15.3
A_4	20	3.5	3.9	21.4
A_5	25	3.1	3.1	23.7

Table 1:

Problem 3

a)

To find probability that the device came from the i'th facility given its defective we can write up bayes theorem

$$P(A_i|D) = \frac{P(D|A_i)P(A_i)}{P(D)} \tag{1}$$

where $P(D|A_i)$ is the probability the device was defective given it came from the i'th facility, $P(A_i)$ is the probability the device came from i'th facility and $P(D)$ is the probability the device is defective which is given by

$$P(D) = \sum_{i}^{N} a_i \cdot d_i \tag{2}$$

where a is the percentage produced in the i'th facility and d is the percentage of defective devices produced in the i'th facility.

Looking at Table 1 we have calculated the probability a device coming from i'th facility given it's defective and we see that for A_2 it's 18.3% and we see that if we have a defective device it's most likely from A_5 $P(A_5|D) = 23.7\%$.

b)

Our goal now is to change our defective percentages such that $P(A_i|D) = 20\%$ so that it's equaly likely that our defective decice came from any of the 5 facilities. We can achieve this by rescaling such that

$$\frac{P(D|A_i)P(A_i)}{P(D)} = \frac{\max\{P(D|A_i)P(A_i)\}}{P(D)} \tag{3}$$

$$P(D|A_i) = \frac{\max\{P(D|A_i)P(A_i)\}}{P(A_i)} \tag{4}$$

where $P(D|A_i)$ will be our new defective values which can be found in Table 1.

c)

Repeating the previous question with the new values we get Table 2

Facility	Total produced	Defective rate	Updated defective rate
A_1	0.27	0.02	0.022
A_2	0.1	0.04	0.060
A_3	0.05	0.1	0.119
A_4	0.08	0.035	0.074
A_5	0.25	0.022	0.024
A_6	0.033	0.092	0.180
A_7	0.019	0.12	0.313
A_8	0.085	0.07	0.070
A_9	0.033	0.11	0.180
A_10	0.02	0.02	0.298
A_11	0.015	0.07	0.397
A_12	0.022	0.06	0.270
A_13	0.015	0.099	0.397
A_14	0.008	0.082	0.744

Table 2:

Problem 4

a)

Using the 8th row of the the given data we construct a kernel density estimator with the sklearn package using a epanechnikov kernel with a bandwidth of 0.4. The plot can be seen in figure 2.

Integrating both kernels from -2° C to 4° C should result result in 1 since they both are normalized and doing so indeed results in 1. Integrating both kernels from -2° C to 0° C results in the area 0.2 for the 1997 data and 0.22 for the 2017 data. Integration was done using scipys integration tools.

b)

Sampling 1000 numbers from the 1997 kde over the range -1° C to 2° C (the samples can be seen in figure 2) we calculate the likelihood ratio

$$\frac{\mathcal{L}(H_0|x)}{\mathcal{L}(H_1|x)} = 1.96 \cdot 10^{116} \tag{5}$$

since our 1000 samples came from the 1997 kde which was our null hypothesis we would expect $\mathcal{L}(H_0|x)$ to be larger than $\mathcal{L}(H_1|x)$ which in turn would make our ratio large which we indeed find.

The samples can be found in the file `ramyar_KDE_1000_samples.txt`

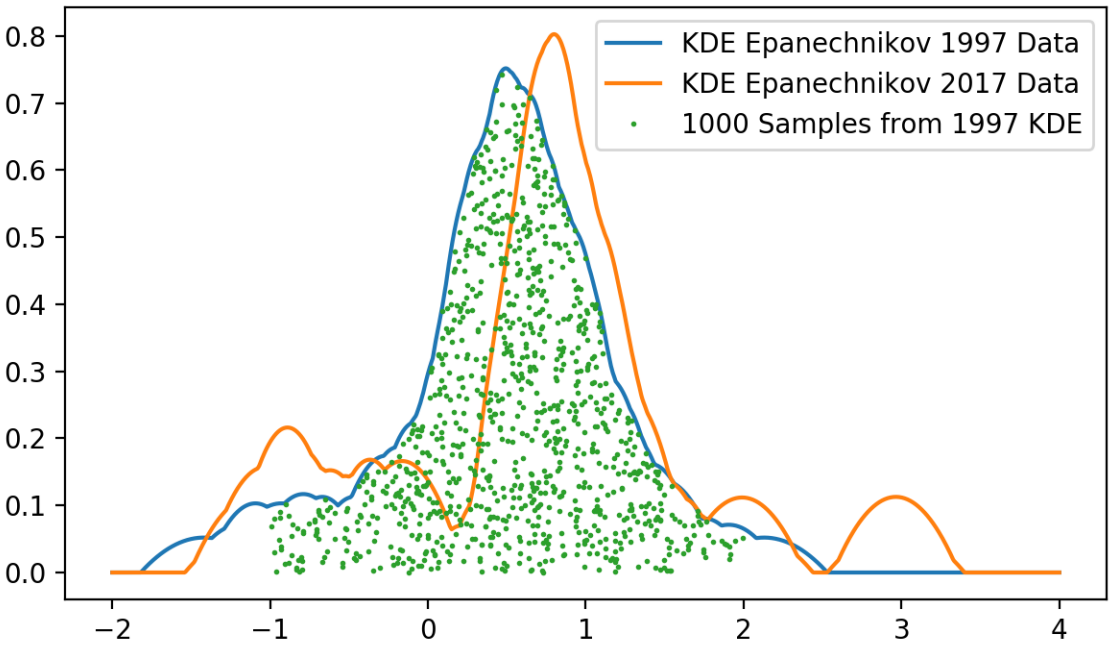


Figure 2: Kernel density estimators for 1997 and 2017 data as well as 1000 sample points.

Problem 5

a)

First we solve our expected pdf using some solver on the internet we get

$$f(t;b,\sigma_t) = \int_0^\infty \frac{\mathrm{e}^{-\frac{(t-t')^2}{2{\sigma_t}^2}}\mathrm{e}^{-\frac{t'}{b}}}{\sqrt{2\pi}\sigma_t b} = \frac{\mathrm{e}^{\frac{{\sigma_t}^2}{2b^2}-\frac{t}{b}}\left(\mathrm{erf}\left(\frac{\sqrt{2}bt-\sqrt{2}{\sigma_t}^2}{2b\sigma_t}\right)+1\right)}{2b} \tag{6}$$

where erf is the error function.

Using $f(t;b,\sigma_t)$ we calculate the maximized ln-likelihoods for the 100 pseudo-experiments for each of the hypotheses and then calculate $-2\ln\lambda$ the results are then plottet as a histogram with bin width 1.

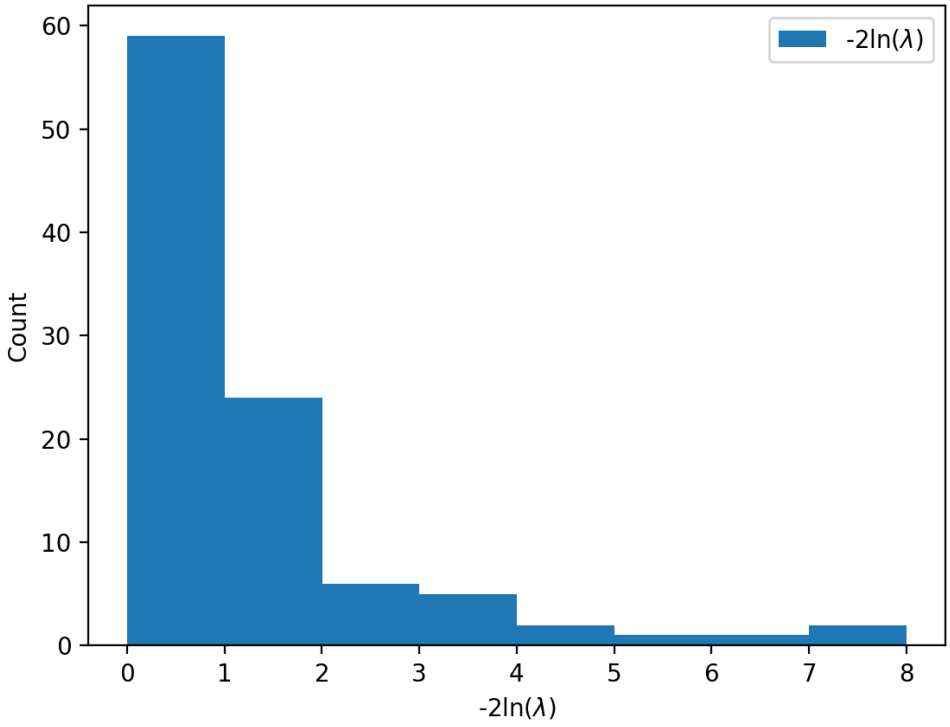


Figure 3: A histogram of calculated $-2\ln\lambda$ values

b)

To figure out whether or not the $-2\ln\lambda$ values are χ^2 distributed we will perform a χ^2 test. Since we only have two variables in λ we only have 1 degree of freedom (df) and therefore we will compare it to the χ^2 distribution with 1 df (this is our null hypothesis).

We get the expected count in each bin from the formula

$$\left(\int_{x_1}^{x_2} \frac{e^{-x/2}}{\sqrt{2\pi x}} dx\right) * 100 \tag{7}$$

where for the first bin we would integrate from 0 to 1 and the expected count would be 68 and so on.

Setting a significance level of $\alpha = 0.05$ which corresponds to a threshold value of 14.07 (since we have 8 bins we have 7 df in our χ^2 test where the threshold value can we found in a table).

We then do the χ^2 test and get a χ^2 value of 14.03 which is lower than our threshold value and we can therefore not reject our null hypothesis and the distribution is therefore a χ^2 distribution with 1 df.

Counting the pseudo-experiments with a value $-2ln(\lambda) > 2.706$ we get 11 values.

To figure out if 11 values > 2.706 is consistent with what we expect to see after 100 throws we will use a binomial distribution with $n = 100$ and $p \approx 0.1$ which was calculated by taking the integral of the χ^2 distribution for 1 df from 2.706 to ∞.

Plugging the values into wolfram we get a mean of ≈ 10 values and standard deviation of ≈ 3 values so 11 values are consistent with what we expect.

c)

We know our $-2\ln\lambda$ value comes from a χ^2 distribution with 1 df if $\mathcal{L}(\omega)$ is true as the number of datapoints goes to infinity (wilk's theorem). So if we want to reject our null hypothesis we can check whether or not our $-2\ln\lambda$ value is larger than some rejection threshold which is in this case 3σ.

We choose our threshold value such that the area under the χ^2 distribution for 1 df after that value is $1-3\sigma$

$$\alpha = \int_k^\infty \frac{e^{-x/2}}{\sqrt{2\pi x}} dx = \left[erf\left(\sqrt{\frac{x}{2}}\right)\right]_k^\infty = 1 - \left[erf\left(\sqrt{\frac{k}{2}}\right)\right] = 1 - 3\sigma = 0.0027 \tag{8}$$

where k is our threshold value which is calculated to be $k = 9$.

Now calculating the $-2\ln\lambda$ value for the 20000 datapoints we get 0.68 which is less than our threshold value so we cant reject the null hypothesis.

Just to check i plotted the data in a histogram and (6) with the values found for H_0 and H_1 by the maximum likelihood estimation in figure 4 to see if there was a significant difference between the two by eye. They look pretty much the same so i'm confident that the null hypothesis cannot be rejected.

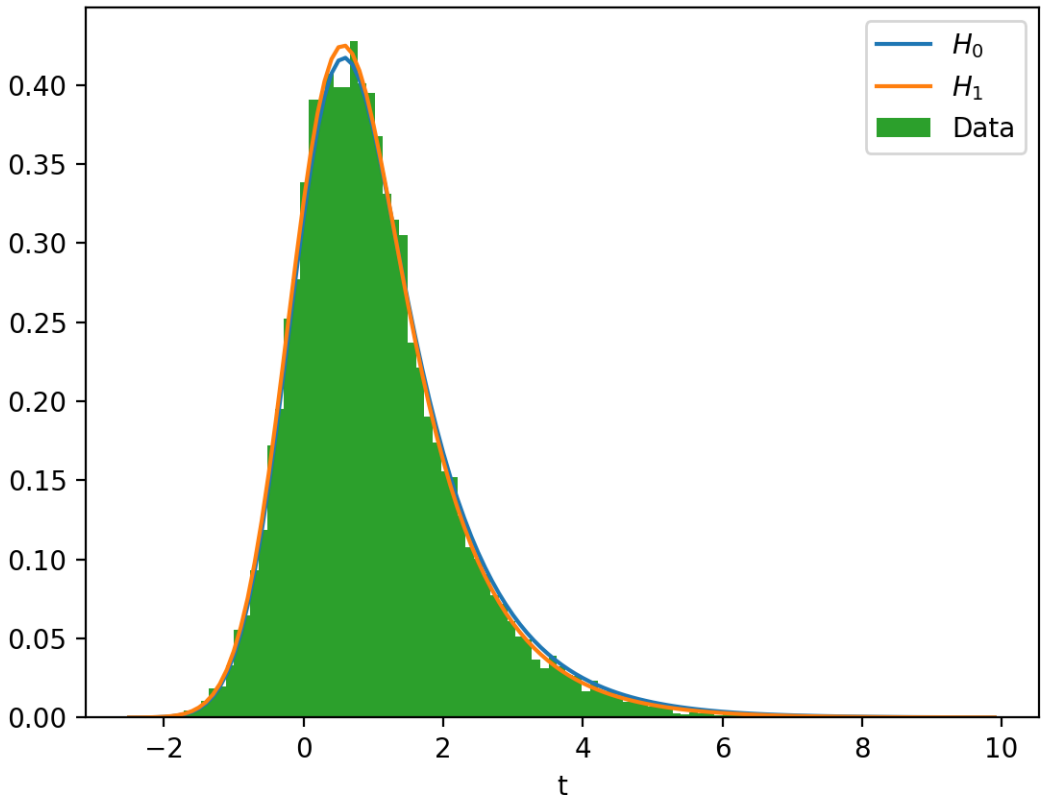


Figure 4: Data from ProblemSet2_Prob5_NucData.txt file plottet as a histogram with 100 bins and the two hypotheses where $b = b_0$ is our null hypothesis and $b \neq b_0$ is H_1