Assignment-7

Shivanshu Shekhar

January 6, 2024

1 Assignments

1. Implement the function to generate normally distributed random numbers in Python. Plot a histogram of the random numbers generated and compare to the theoretical pdf.

2. Use the Chi-squared test and determine if 100 random numbers generated above belong to a gaussian (zero mean, unit variance) distribution. Vary the bin sizes and the number of random numbers and determine the optimal test. Apply the test to random numbers generated from the above function and offset by 0.01. When can the chi-squared test distinguish between the correct random numbers and the offset ones?

3. Repeat Q2 with the KS test

4. Consider now that the function in Q1 was the sum of two random distributions:

$$x = \begin{cases} randn() & 99.9\% \text{ of the time} \\ randn() * 500 & 0.1\% \text{ of the time} \end{cases}$$

5. Consider the following function:

$$f(x, y) = u^2 + v^2 \tag{1}$$

where

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} \cos(\alpha) & \sin(\alpha) \\ -\sin(\alpha) & \cos(\alpha) \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix} \tag{2}$$

where $\alpha = \pi \sin\Big(10(\sqrt{(}x^2 + y^2) - 0.5)\Big)$. We wish to compute:

$$I = \int_{|f|<1} dxdy$$

(a) Write a Python function to implement this function of two variables, x and y.

(b) Plot a contour plot (use `contourf` to get filled contours) of the function with contour values `[0.0,1.0]` and consider the region corresponding to $|f| < 1$.

(c) Find a covering function that contains all the places where $|f| < 1$, and use it to estimate the integral above.

2 Chisquare and KS test

Chi-square Test:

The Chi-square (χ^2) test is a statistical test used to determine if there is a significant association between two categorical variables. It is particularly useful for comparing observed and expected frequencies in a contingency table.

Test Statistic Formula:

The test statistic (χ^2) is calculated using the following formula:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Where:

- χ^2 is the test statistic.
- O_i is the observed frequency for category i.
- E_i is the expected frequency for category i.

Degrees of Freedom:

The degrees of freedom (df) for the Chi-square test depend on the dimensions of the contingency table. For a 2×2 table, $df = 1$. For larger tables, $df = (rows - 1) \times (columns - 1)$.

Null Hypothesis:

The null hypothesis (H_0) for the Chi-square test is that there is no significant association between the variables.

Kolmogorov-Smirnov (KS) Test:

The Kolmogorov-Smirnov test is a non-parametric test used to compare the cumulative distribution function (CDF) of a sample with a reference probability distribution or another sample.

Test Statistic Formula:

For a one-sample KS test, the test statistic (D) is given by:

$$D = \max(|F(x) - S(x)|)$$

Where:

- $F(x)$ is the empirical distribution function of the sample.
- $S(x)$ is the cumulative distribution function of the reference distribution.
- max is taken over all data points.

For a two-sample KS test, the test statistic is calculated similarly, comparing the CDFs of two samples.

Critical Values:

The critical value for the KS test depends on the significance level and the sample size.

Null Hypothesis:

The null hypothesis (H_0) for the KS test is that the sample is drawn from the same distribution as the reference distribution.

3 Solutions

3.1 Chi Squared Test

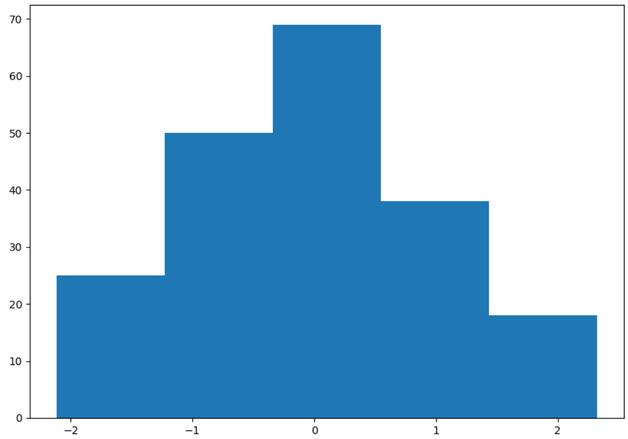

Figure 1: Box Muller Histogram

We generate 200 points sampled from a standard Gaussian using the Box Muller Algorithm. We further use 5 bins to plot them as shown in Figure: 1.

We then conduct a Chi-squared test on these points to see when the Chi-squared statistic gives a high p-value.

We plot the Chi-squared value and p-value in Figures: 2 & 3 respectively.

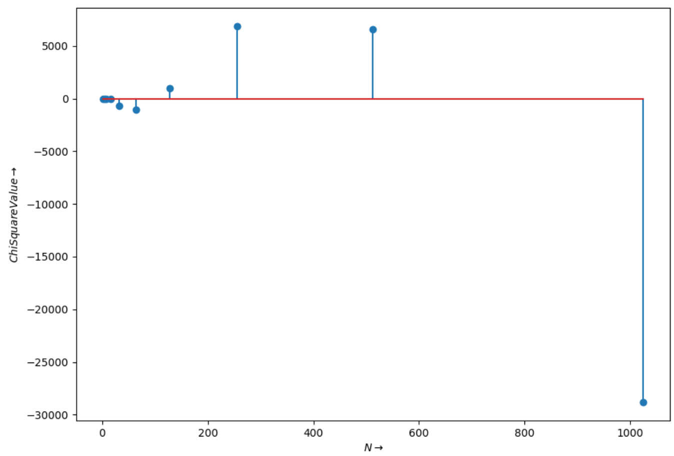

Figure 2: Chi-squared Values

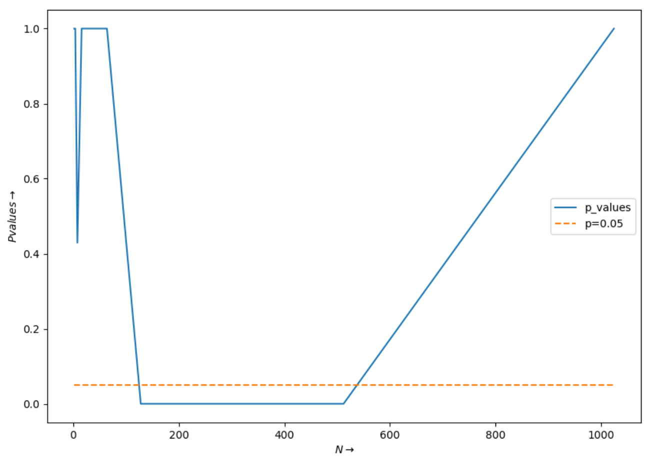

Figure 3: p-values for Chi-squared test

We see that the Chi-squared test requires about 600 points to give the correct statistics.

If we shift the point by a costant we see that the Chi-squared test gives the correct answer after 600 points as shown in Figure: 4

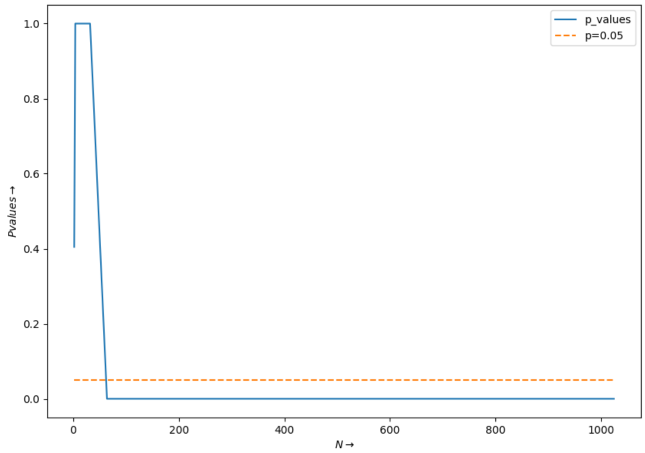

Figure 4: p-values for Chi-squared test

3.2 KS Test

We repeat the same things for KS statistics, we get the corresponding Figures: 5, 6, 7:

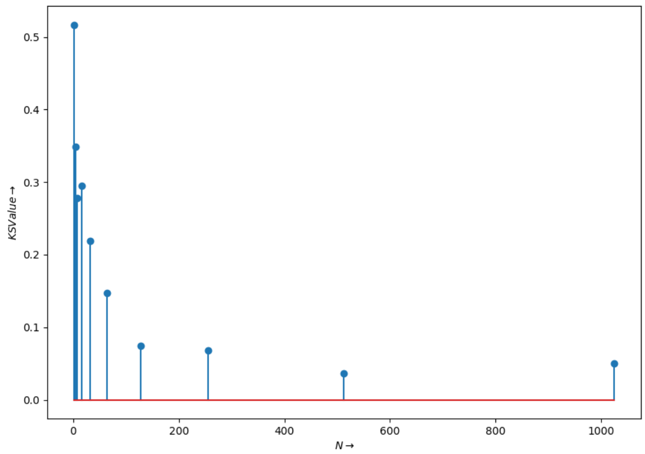

Figure 5: KS test values

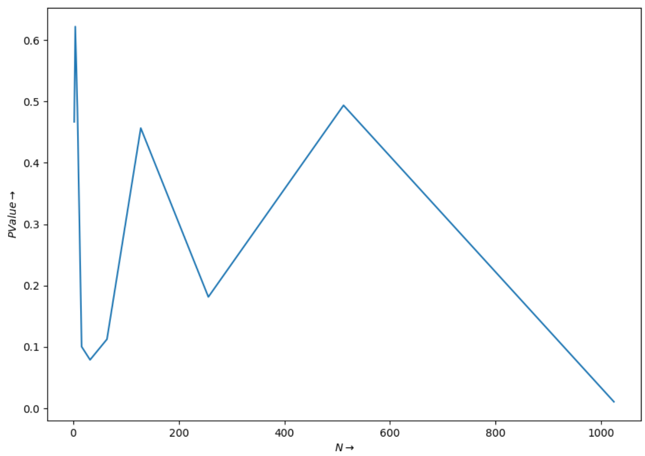

Figure 6: p-values for KS test

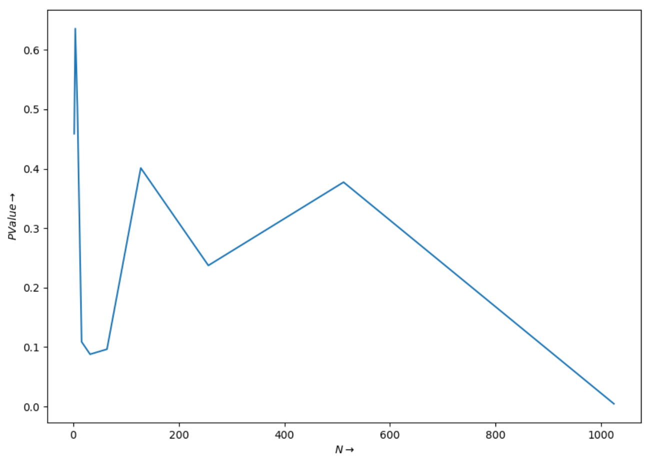

Figure 7: p-values for KS test

We see that KS test is a lot more sensitive than Chi-Squared test.

3.3 Area under the function

We see from Figures: 8, 9, 10 that the function can be bounded a square between -2 to 2. We then uniformly sample points between -2 to 2 and then calculate the probability of the sample point lying inside the function contour. and we multiply this probability with the area of the square to get our final area.

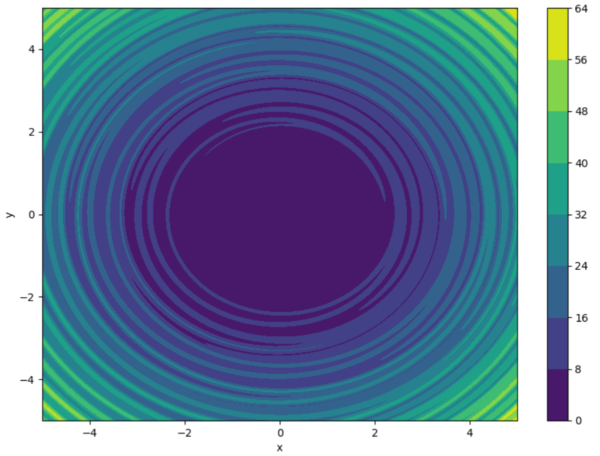

Figure 8: Contour of $f(x, y)$

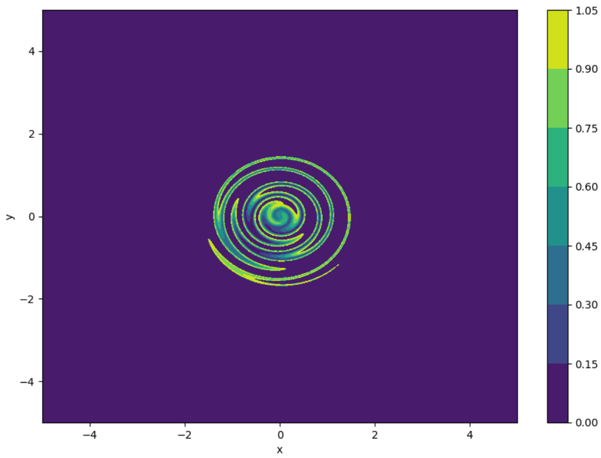

Figure 9: Region of Interest

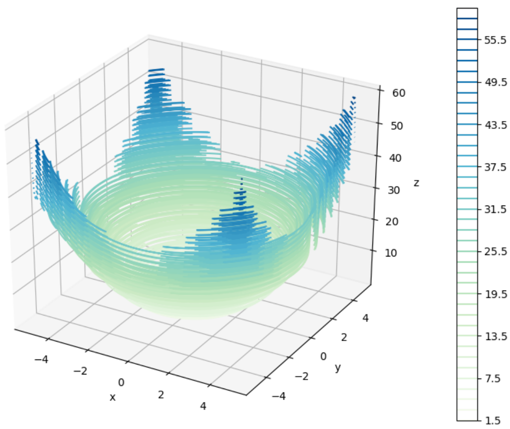

Figure 10: 3D plot of $f(x, y)$