List 3 Report

Statistics and Linear Models

Dawid Dieu

30 December 2022

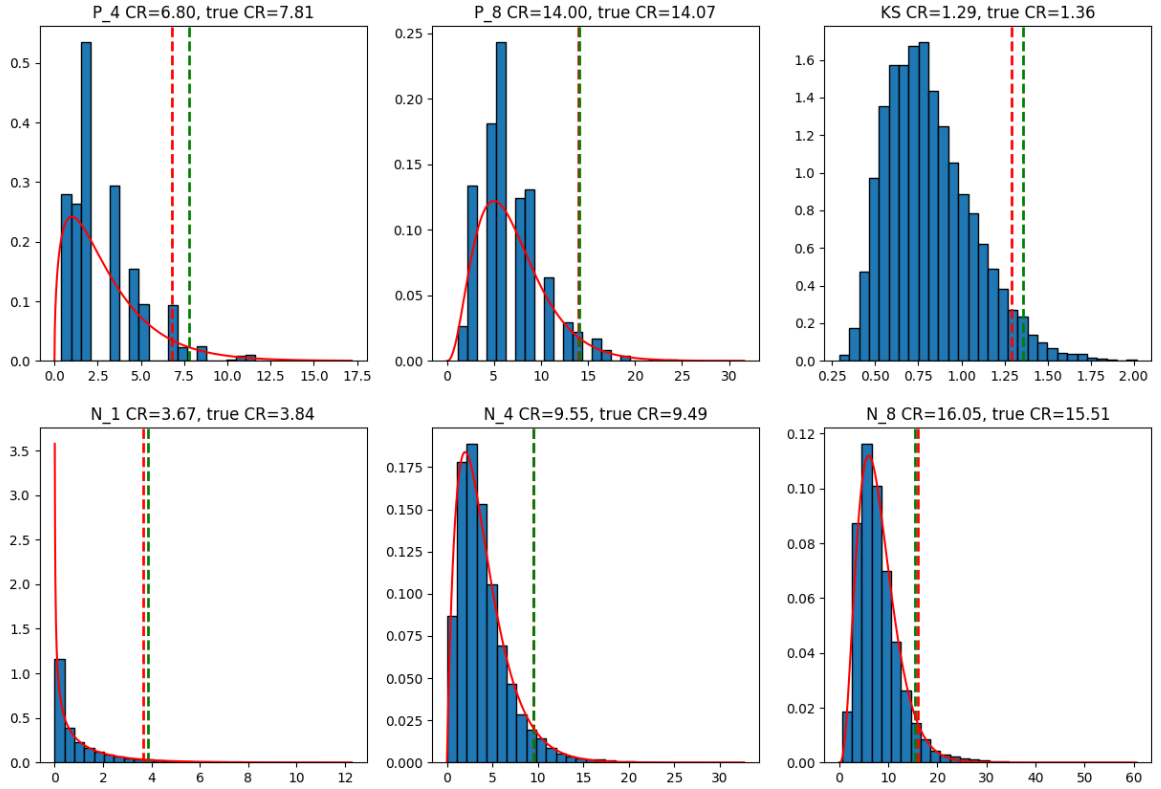
Goal

This report aims to show three different statistical tests: the classical Pearson's chi-square test, the Neyman's smooth test and the Kolmogorov-Smirnov test. We generate data from the uniform and some arbitrary distributions. Then we calculate the statistics mentioned above to investigate the behaviour of the critical values and the power functions.

Task 1

In the first task, we are asked to generate $n = 10$ observations from the uniform $U(0,1)$ distribution. Then, we calculate the values of the statistics $P_4, P_8, N_1, N_4, N_4,$ and KS. We repeat the experiment 10 000 times for $n = 20,30,...,100$ to find the tests' critical values. We compare them with 0.95-quantiles of the respective limiting distributions.

First, we present histograms where $n = 10$. The red line shows the theoretical chi-square distribution, and it is well aligned with the histograms of the Pearson and Neyman statistics. The red dashed line represents the critical value calculated from the histogram. The green one is the true critical value taken from the proper distributions.



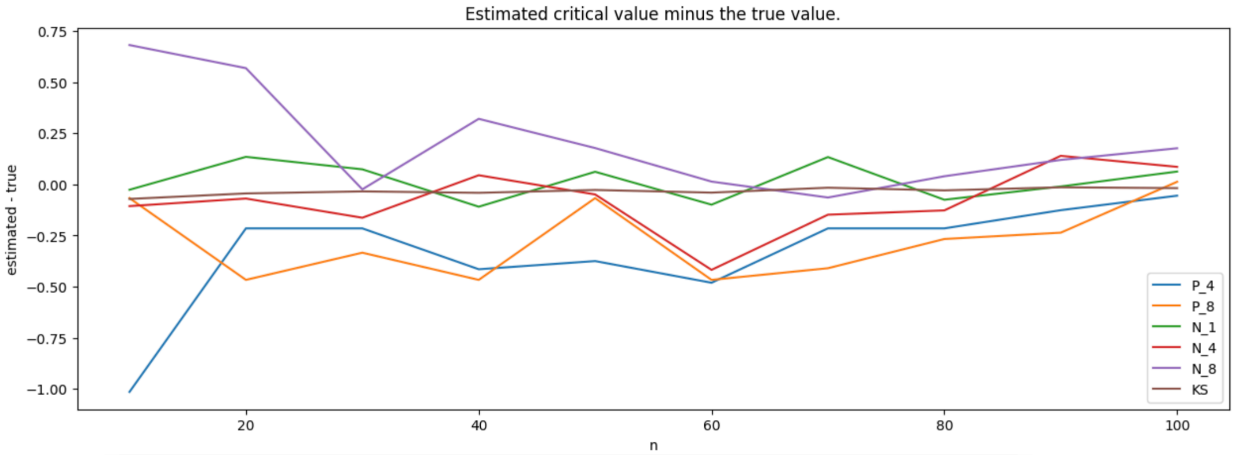
Despite the n being very small, the critical values are close.

Here is the table with the theoretical values.

P_4	P_8	N_1	N_4	N_8	KS
7.814728	14.06714	3.841459	9.487729	15.507313	1.358099

Here is the table with the estimated critical values minus the theoretical values from the table above. We can see that when we increase the sample size, the values get closer to 0, meaning that the estimated value converges to the theoretical value. The KS statistic is the most stable and has the smallest variance.

n	$\hat{P}_4 - P_4$	$\hat{P}_8 - P_8$	$\hat{N}_1 - N_1$	$\hat{N}_4 - N_4$	$\hat{N}_8 - N_8$	$\hat{KS} - KS$
10	-1.015	-0.067	-0.026	-0.106	0.682	-0.071
20	-0.215	-0.467	0.135	-0.069	0.569	-0.044
30	-0.215	-0.334	0.074	-0.163	-0.024	-0.034
40	-0.415	-0.467	-0.109	0.045	0.321	-0.041
50	-0.375	-0.067	0.062	-0.049	0.178	-0.027
60	-0.481	-0.467	-0.099	-0.418	0.014	-0.040
70	-0.215	-0.410	0.134	-0.148	-0.064	-0.016
80	-0.215	-0.267	-0.075	-0.127	0.040	-0.029
90	-0.126	-0.236	-0.010	0.140	0.120	-0.014
100	-0.055	0.013	0.063	0.086	0.177	-0.018
Variance	0.074	0.033	0.008	0.024	0.063	0.00028
Mean	-0.3327	-0.2769	0.0149	-0.0809	0.2013	-0.0334



Task 2

The accept/reject von Neumann algorithm can help us to generate random numbers from an arbitrary distribution. We can use it to generate samples in tasks 3 and 4.

Let $Y \sim f_Y(y)$ and $V \sim f_V(v)$, where f_Y and f_V have common support

$$M = sup_y \frac{f_Y(y)}{f_V(y)} < \infty \implies \frac{f_Y(y)}{M f_V(y)} \leq 1.$$

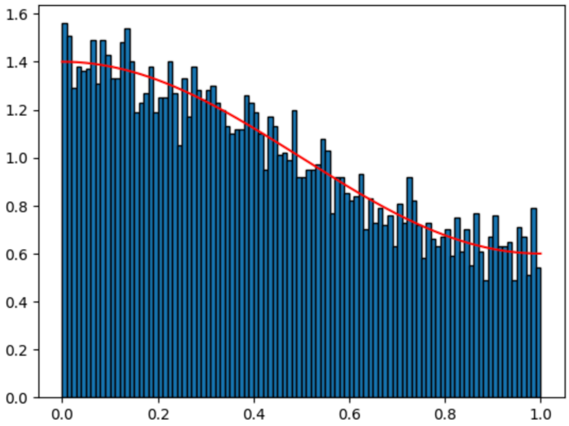
To generate a random variable $Y \sim f_Y$, one has to:

1. Generate two independent variables $u \sim U(0,1)$, $v \sim f_V$.
2. If $u < \frac{f_Y(v)}{M f_V(v)}$, then set $Y = v$. Otherwise, go back to step 1).

Task 3

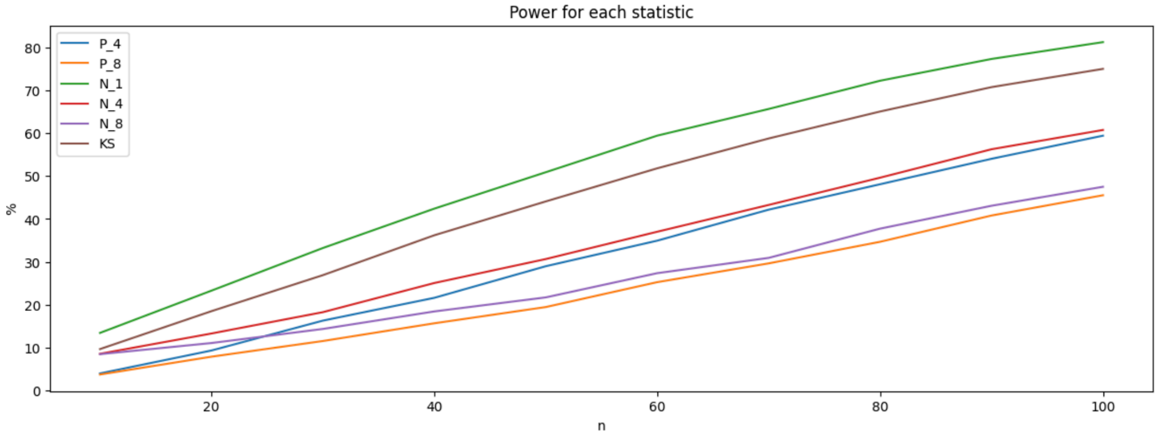
In the third task, we are asked to generate $n = 10$ observations from the density $C_1(u,0.4) = 1 + 0.4cos(\pi u), u \in (0,1)$. Then, we calculate the values of the statistics$P_4, P_8, N_1, N_4, N_4,$ and KS. We repeat the experiment 10 000 times for $n = 20,30,...,100$ to find the values of the power functions of the tests under consideration. We use $\alpha = 0.05$.

We can use the von Neumann algorithm described in the previous task to generate numbers from such distribution. On the right side, we have a histogram of such generated data to confirm that the algorithm works as we intend.



Now let's look at power functions. The power grows when we increase the sample size. With more samples, the distribution seems less and less uniform. We can see that statistics P_1 and KS yields the highest power, 81% and 75%, respectively.

n	P_4	P_8	N_1	N_4	N_8	KS
10	4.02	3.77	13.46	8.59	8.51	9.69
20	9.36	7.92	23.30	13.32	11.10	18.52
30	16.32	11.57	33.23	18.29	14.38	26.91
40	21.63	15.70	42.40	25.08	18.46	36.20
50	28.99	19.48	50.88	30.66	21.73	44.08
60	34.94	25.28	59.40	37.02	27.37	51.79
70	42.16	29.65	65.61	43.26	30.93	58.74
80	48.08	34.68	72.19	49.61	37.74	65.02
90	54.01	40.80	77.26	56.23	43.07	70.70
100	59.39	45.52	81.18	60.72	47.49	74.95

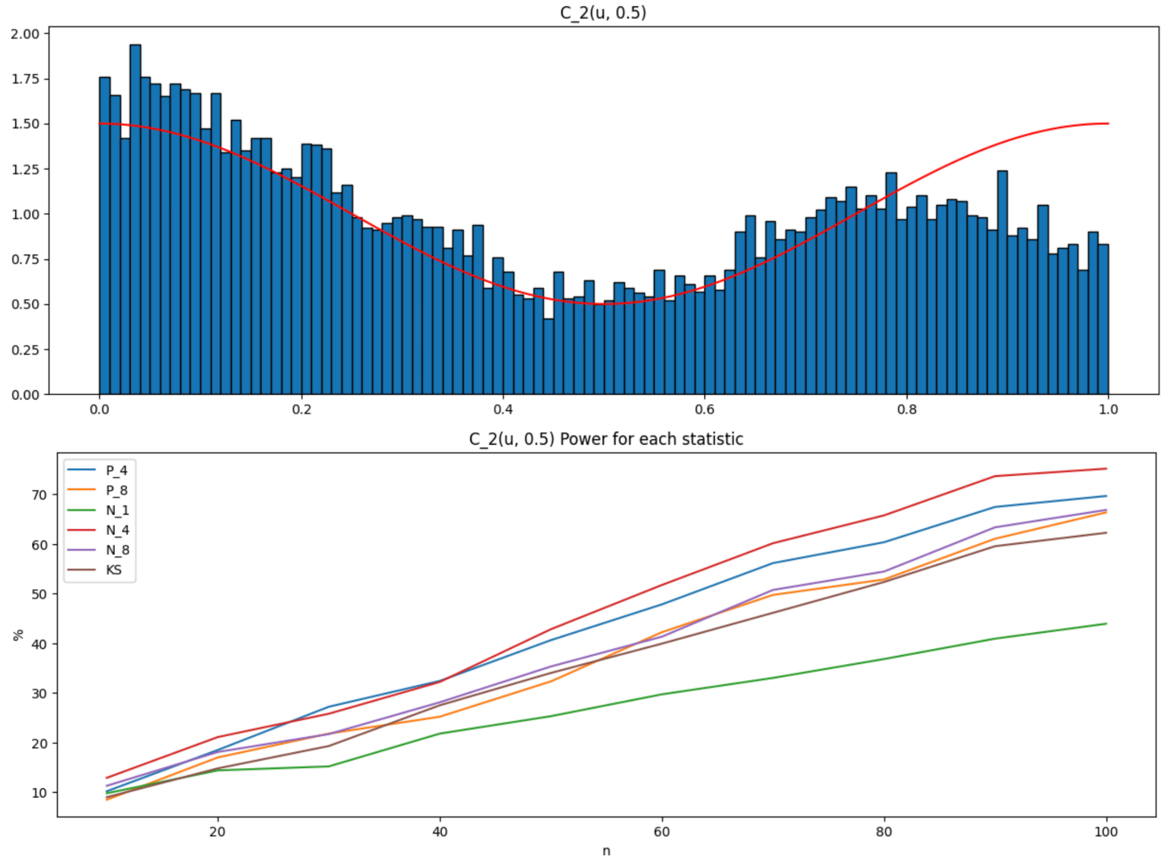


Task 4

In the fourth task, we are asked to generate $n = 10,20,...,100$ observations from the density $C_j(u,\rho) = 1 + \rho cos(j\pi u), u \in (0,1)$. Then we repeat the numerical experiments from the previous task and show charts with power functions.

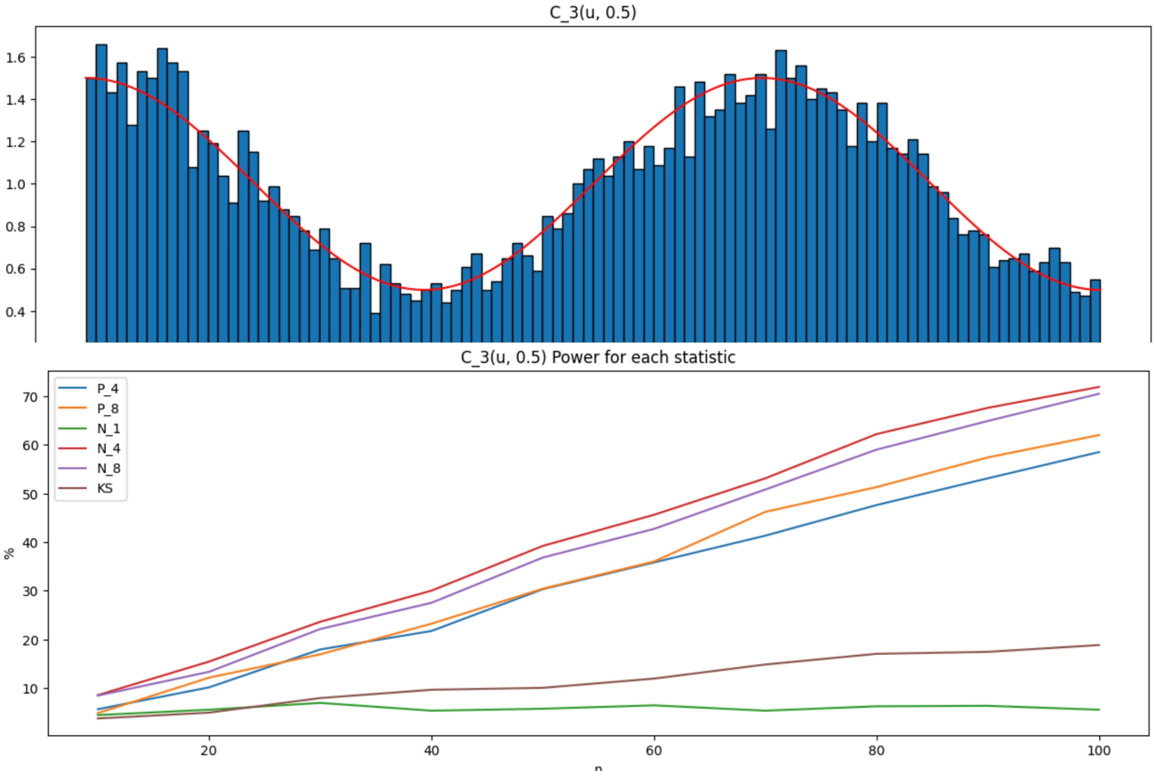
a) $j = 2, \rho = 0.5$

Here the N_4 statistic yields the highest power of 75%.



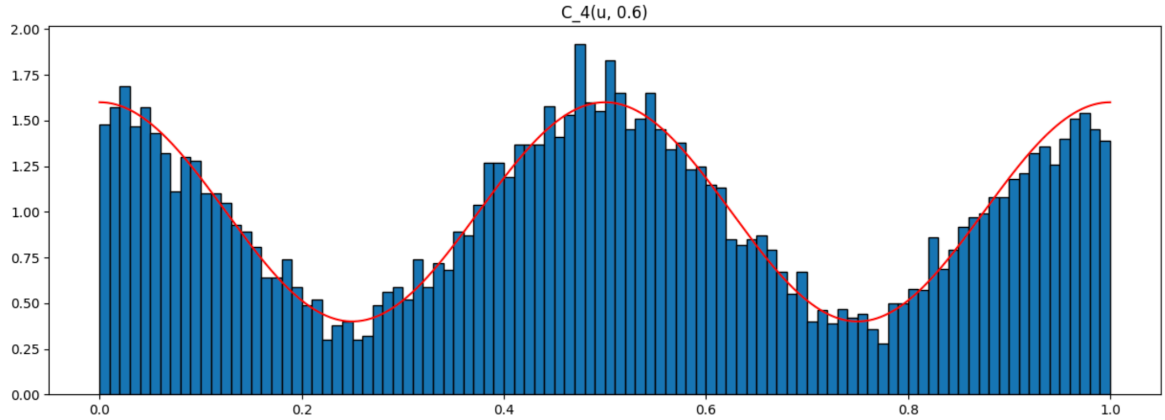
b) $j = 3, \rho = 0.5$

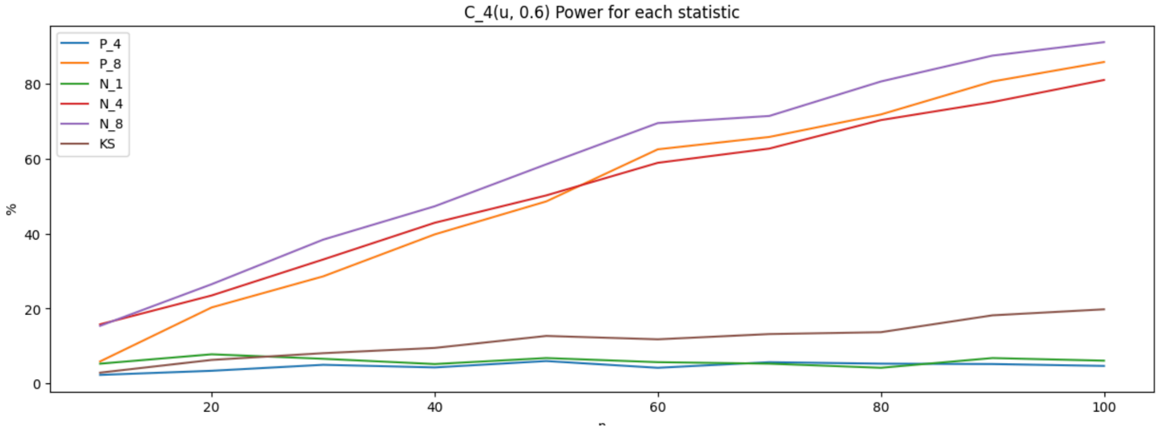
Here also, the N_4 statistic yields the highest power of 72%.



c) $j = 4, \rho = 0.6$

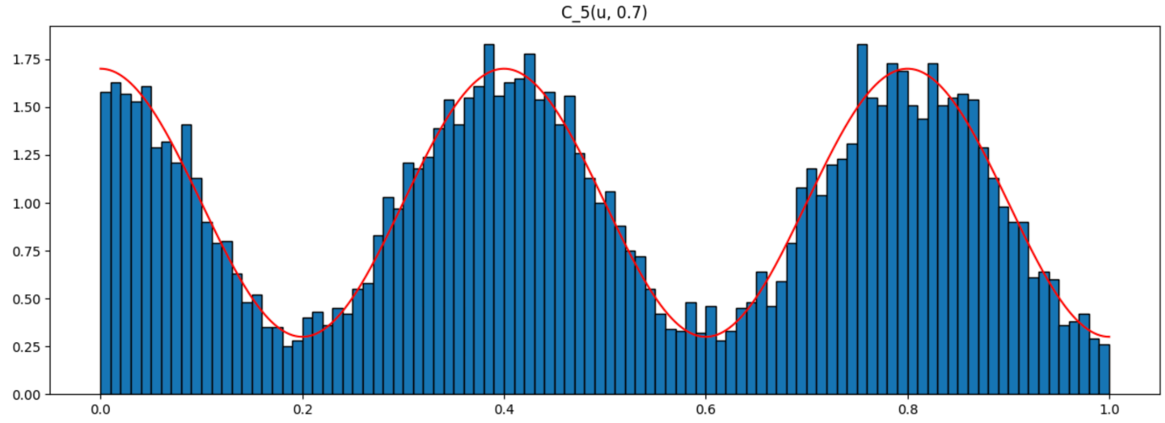
Here the N_8 statistic yields the highest power of 91%. It's noticeable that KS, N_1, and P_4 performed poorly.

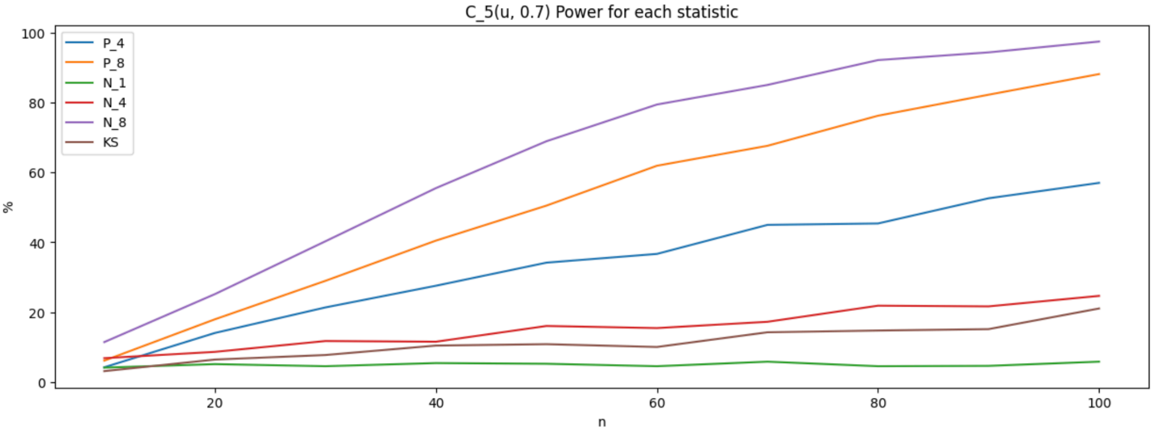




d) $j = 5, \rho = 0.7$

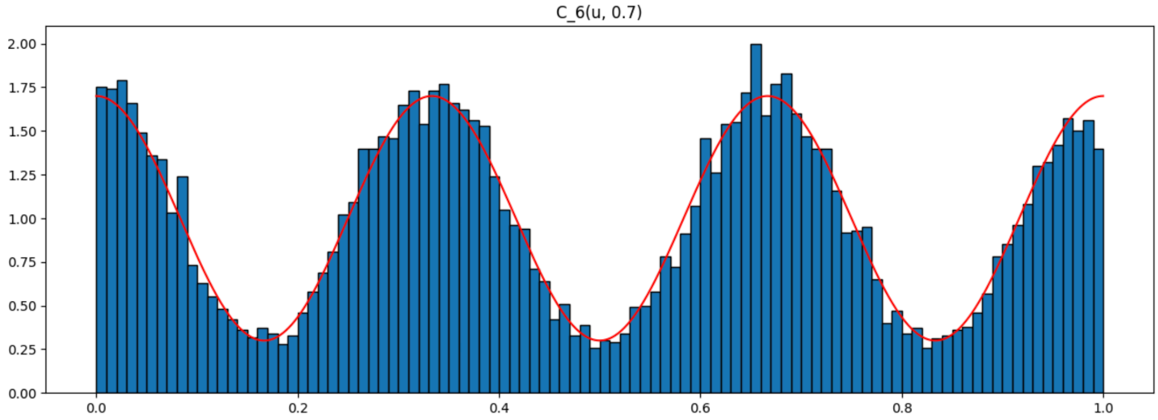
Here also, the N_8 statistic yields the highest power of 97%. The worst performing statistics are N_4, KS, and N_1.

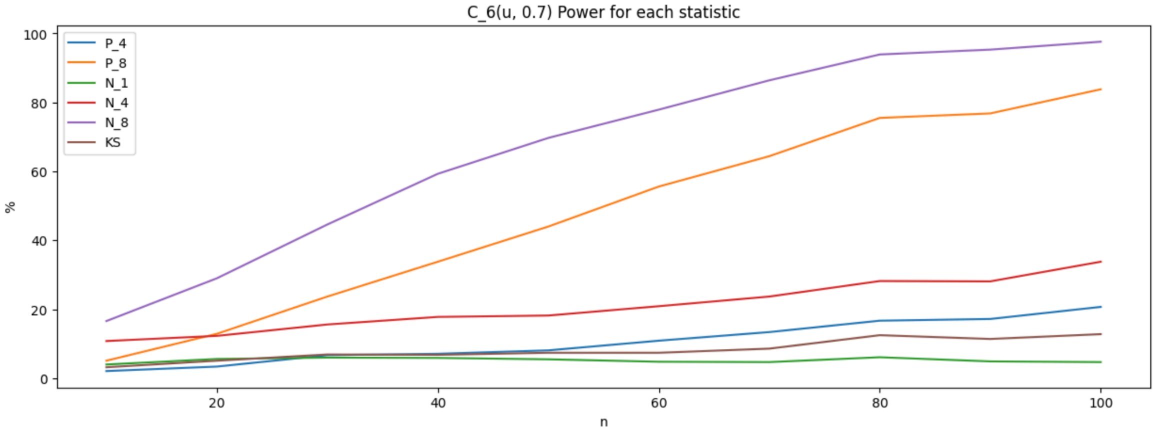




e) $j = 6, \rho = 0.7$

Here also, the N_8 statistic yields the highest power of 97%. P_8 is the second best with 84% power. The worst performing statistics are N_4, P_4, KS, and N_1. One could see that the more peaks in the distribution, the higher power was for the best statistics (like N_8).


C_6(u, 0.7)
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
0.0
0.2
0.4
0.6
0.8
1.0


C_6(u, 0.7) Power for each statistic
P_4
P_8
N_1
N_4
N_8
KS
n