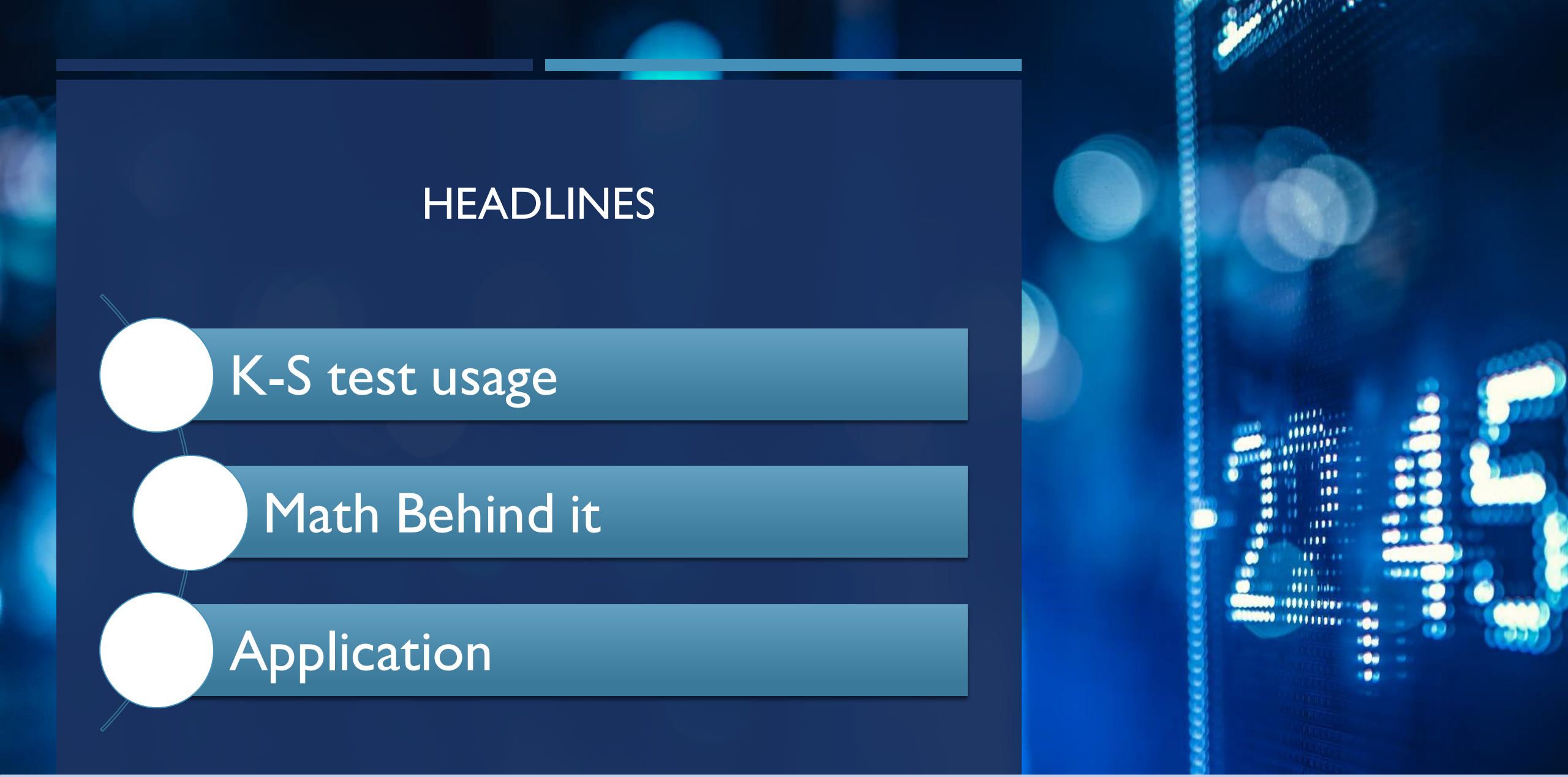
KOLMOGOROV–SMIRNOV TEST

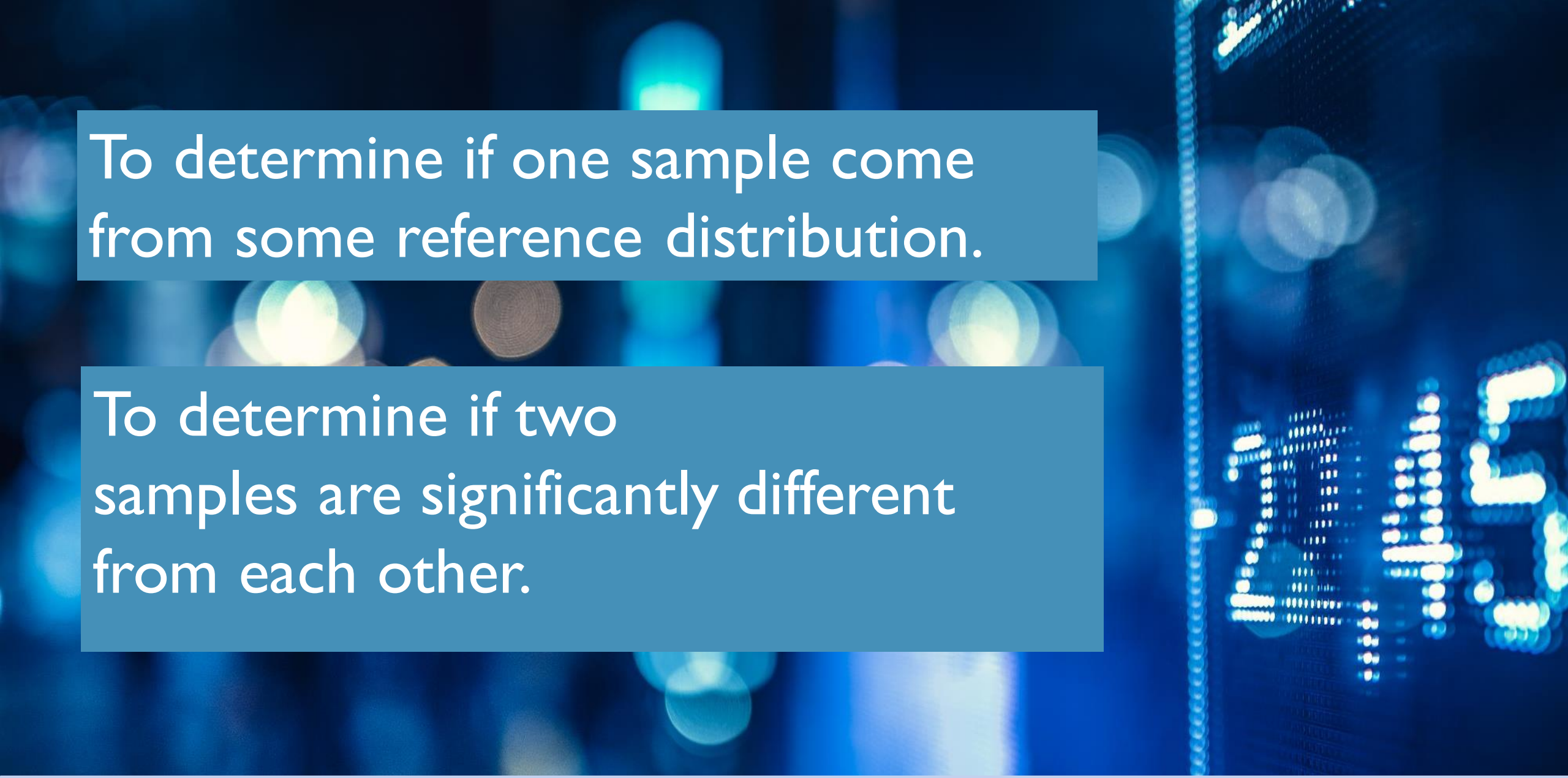
ŞADİ UYSAL

BOGAZICI UNIVERSITY



HEADLINES

- K-S test usage
- Math Behind it
- Application


To determine if one sample come from some reference distribution.
To determine if two samples are significantly different from each other.

The Kolmogorov–Smirnov statistic quantifies a distance between the **empirical distribution function** of the sample and the **cumulative distribution function(cdf)** of the reference distribution, or between the empirical distribution functions of two samples.

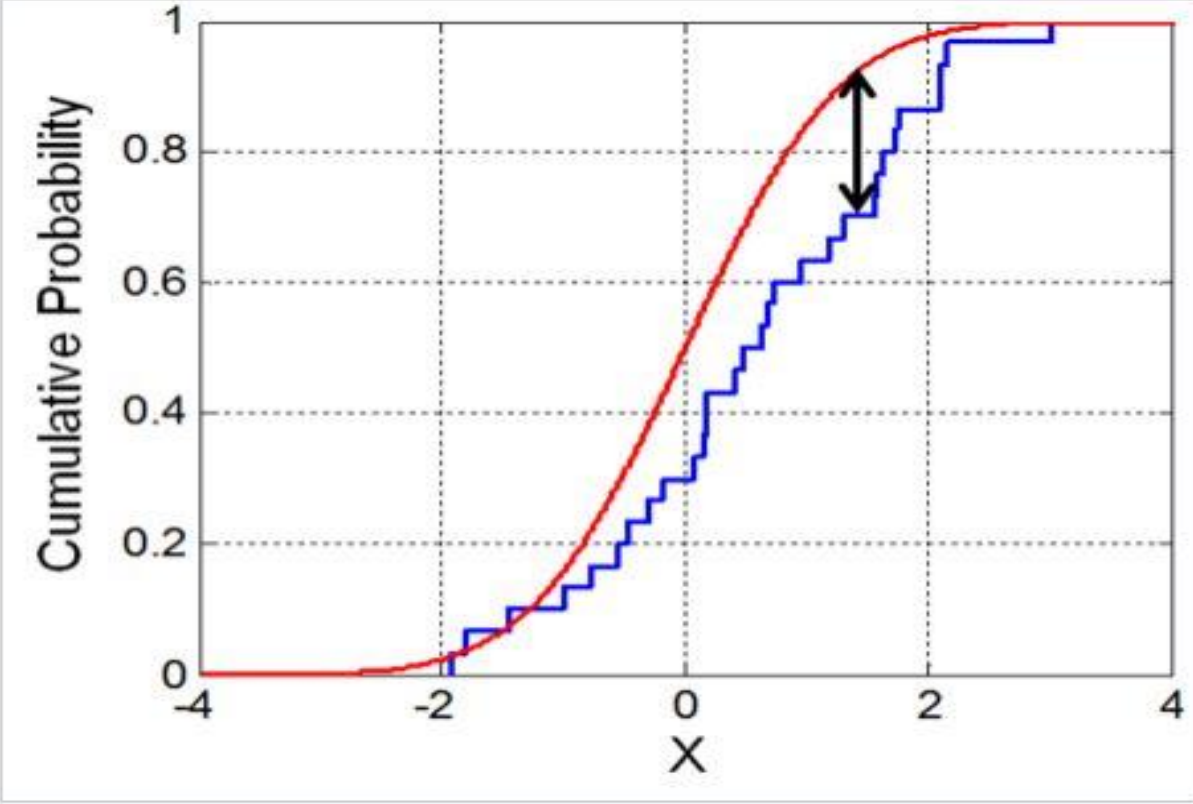


Illustration of the Kolmogorov–Smirnov statistic. Red line is CDF, blue line is an ECDF, and the black arrow is the K–S statistic.

Suppose you are given the following 100 observations.

-0.16	-0.68	-0.32	-0.85	0.89	-2.28	0.63	0.41	0.15	0.74
1.30	-0.13	0.80	-0.75	0.28	-1.00	0.14	-1.38	-0.04	-0.25
-0.17	1.29	0.47	-1.23	0.21	-0.04	0.07	-0.08	0.32	-0.17
0.13	-1.94	0.78	0.19	-0.12	-0.19	0.76	-1.48	-0.01	0.20
-1.97	-0.37	3.08	-0.40	0.80	0.01	1.32	-0.47	2.29	-0.26
-1.52	-0.06	-1.02	1.06	0.60	1.15	1.92	-0.06	-0.19	0.67
0.29	0.58	0.02	2.18	-0.04	-0.13	-0.79	-1.28	-1.41	-0.23
0.65	-0.26	-0.17	-1.53	-1.69	-1.60	0.09	-1.11	0.30	0.71
-0.88	-0.03	0.56	-3.68	2.40	0.62	0.52	-1.25	0.85	-0.09
-0.23	-1.16	0.22	-1.68	0.50	-0.35	-0.35	-0.33	-0.24	0.25

Do they come from N(0,1)?

We want to compare the empirical distribution function of the data, F_{obs}, with the cumulative distribution function associated with the null hypothesis, F_{exp} (expected CDF).

Null Hypothesis:The samples come from Normal Distribution

The Kolmogorov-Smirnov statistic is

$$D_n = \max_x |F_{\text{exp}}(x) - F_{\text{obs}}(x)|.$$

In practice, order the data:

-3.68	-2.28	-1.97	-1.94	-1.69	-1.68	-1.60	-1.53	-1.52	-1.48
-1.41	-1.38	-1.28	-1.25	-1.23	-1.16	-1.11	-1.02	-1.00	-0.88
-0.85	-0.79	-0.75	-0.68	-0.47	-0.40	-0.37	-0.35	-0.35	-0.33
-0.32	-0.26	-0.26	-0.25	-0.24	-0.23	-0.23	-0.19	-0.19	-0.17
-0.17	-0.17	-0.16	-0.13	-0.13	-0.12	-0.09	-0.08	-0.06	-0.06
-0.04	-0.04	-0.04	-0.03	-0.01	0.01	0.02	0.07	0.09	0.13
0.14	0.15	0.19	0.20	0.21	0.22	0.25	0.28	0.29	0.30
0.32	0.41	0.47	0.50	0.52	0.56	0.58	0.60	0.62	0.63
0.65	0.67	0.71	0.74	0.76	0.78	0.80	0.80	0.85	0.89
1.06	1.15	1.29	1.30	1.32	1.92	2.18	2.29	2.40	3.08

Given observations $x_1, \ldots, x_n$ the empirical distribution function $F_{\text{obs}}(x)$ gives the proportion of the data that lies below x,

$$F_{\text{obs}}(x) = \frac{\#\text{observations below } x}{\#\text{observations}}.$$

Then compute the empirical distribution function:

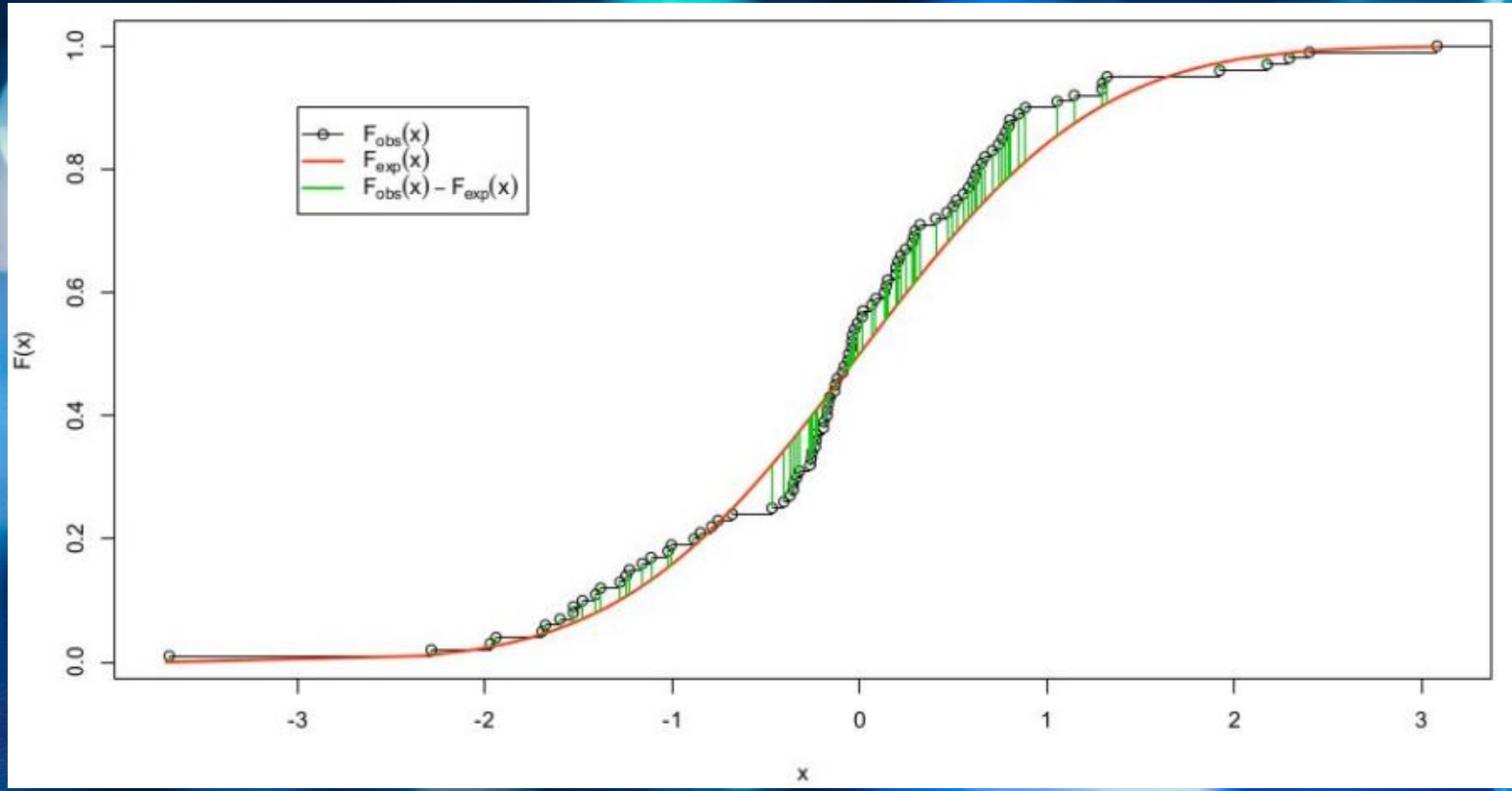
$$F_{\text{obs}}(-3.68) = \frac{1}{100}, \quad F_{\text{obs}}(-2.28) = \frac{2}{100}, \ldots, \quad F_{\text{obs}}(3.08) = 1$$

0.000	0.011	0.024	0.026	0.045	0.047	0.055	0.064	0.064	0.070
0.080	0.084	0.101	0.107	0.110	0.123	0.133	0.154	0.158	0.189
0.198	0.215	0.226	0.249	0.321	0.343	0.356	0.362	0.363	0.369
0.375	0.396	0.399	0.400	0.407	0.409	0.410	0.423	0.425	0.432
0.432	0.434	0.437	0.447	0.449	0.453	0.464	0.468	0.476	0.477
0.484	0.484	0.485	0.490	0.496	0.505	0.508	0.526	0.535	0.553
0.557	0.560	0.577	0.577	0.582	0.588	0.597	0.610	0.614	0.617
0.627	0.658	0.680	0.692	0.698	0.711	0.720	0.727	0.732	0.735
0.743	0.748	0.761	0.771	0.777	0.783	0.788	0.789	0.803	0.812
0.854	0.874	0.902	0.903	0.907	0.973	0.985	0.989	0.992	0.999

For each observation x_i compute $F_{\text{exp}}(x_i) = P(Z \leq x_i)$.

In this case the expected distribution function is standard normal so use the normal table.

We have calculated the maximum absolute distance between the expected and observed distribution functions, in green in the plot below.



Compute the absolute differences between the entries in the two tables.

The Kolmogorov-Smirnov statistic $D_n = 0.092$ is the maximum shown here in blue.

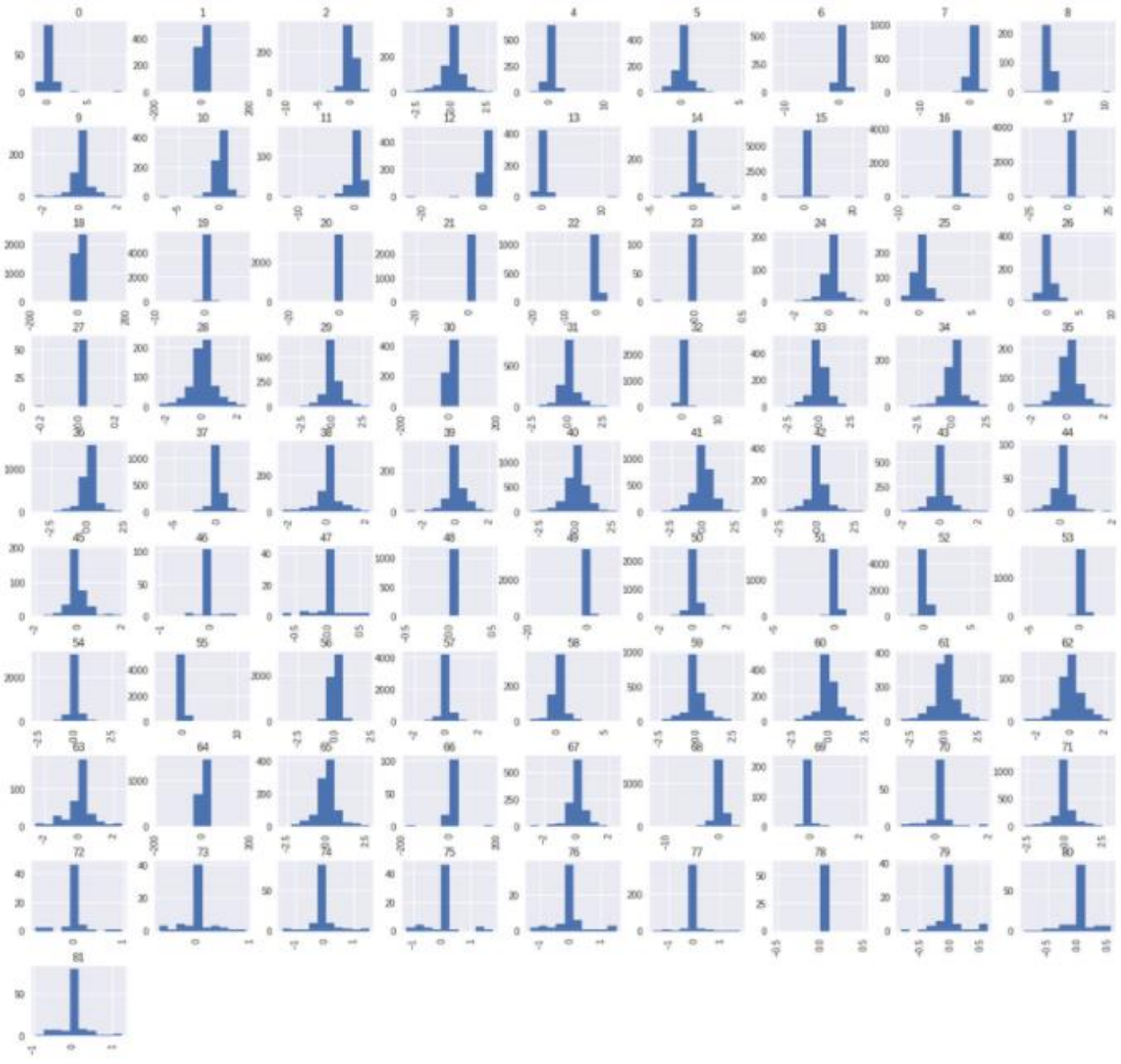
0.010	0.009	0.006	0.014	0.004	0.014	0.015	0.017	0.026	0.031
0.031	0.036	0.030	0.034	0.041	0.037	0.037	0.026	0.031	0.011
0.012	0.005	0.003	0.008	0.069	0.085	0.086	0.083	0.073	0.071
0.064	0.077	0.067	0.061	0.055	0.049	0.039	0.045	0.035	0.033
0.023	0.013	0.006	0.008	0.002	0.008	0.006	0.012	0.014	0.024
0.026	0.036	0.046	0.052	0.054	0.056	0.062	0.052	0.054	0.048
0.054	0.060	0.055	0.061	0.067	0.073	0.071	0.070	0.076	0.082
0.084	0.061	0.049	0.049	0.052	0.048	0.051	0.054	0.058	0.064
0.068	0.071	0.069	0.070	0.074	0.078	0.082	0.092	0.088	0.087
0.055	0.045	0.029	0.037	0.043	0.013	0.015	0.009	0.002	0.001

- At the 95% level the critical value is approximately given by

$$D_{\text{crit},0.05} = \frac{1.36}{\sqrt{n}}.$$

Null Hypothesis:The samples come from Normal Distribution

- Here we have a sample size of $n = 100$ so $D_{\text{crit}} = 0.136$.
- Since $0.092 < 0.136$ do not reject the null hypothesis.



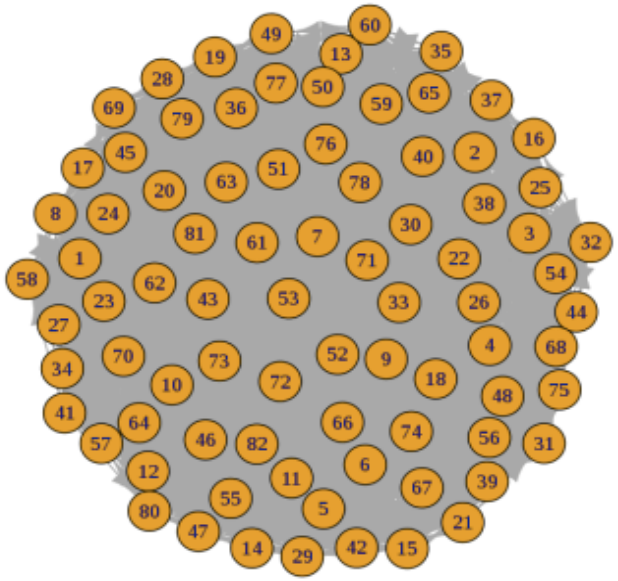
Nodes are the sample number and the vertices length being the KS Test Value.

The goal here is to graph all the possible nodes and vertices.

Two nodes with low KS P-Value would be close and two with High P-Value would be far. This hopefully would create distinguishable clusters.

KOLMOGOROV–SMIRNOV TEST
BOGAZICI UNIVERSITY
ŞADI UYSAL

APPLICATION



We then ended up with a network graph where everybody was connected with everybody else which is not particularly useful.

The next step is to only keep the significant link (lower than a certain threshold)

As we can see on this picture, we got a very exciting result.

THANK YOU

Şadi Uysal
Contact info:
sadiuysalsadi@gmail.com