

SHANNON'S ENTROPY USAGE AS STATISTIC IN ASSESSMENT OF DISTRIBUTION

Lorentz Jäntschi

MINISTRY OF EDUCATION AND RESEARCH



TECHNICAL UNIVERSITY
OF CLUJ-NAPOCA

Sorana D. Bolboacă



Technical University of Cluj-Napoca, Romania

Iuliu Hațieganu University of Medicine and Pharmacy, Romania

Introduction

- General null hypothesis H_0 : *data follow a specific distribution*
 - Kolmogorov-Smirnov (KS)
 - Anderson-Darling (AD)
 - Cramér-von-Mises (CM)
 - Kuiper V (KV)
 - Watson U^2 (WU)
 - Shannon's H1 – introduced here as statistic

Computing of statistics

$$AD = -n - \frac{1}{n} \sum_{i=0}^{n-1} (2 \cdot i + 1) \cdot \ln(f_i \cdot (1 - f_i))$$

$$KS = \sqrt{n} \cdot \max_{0 \leq i < n} \left(f_i - \frac{i-1}{n}, \frac{i}{n} - f_i \right)$$

$$KV = \sqrt{n} \cdot \left(\max_{0 \leq i < n} \left(f_i - \frac{i-1}{n} \right) + \max_{0 \leq i < n} \left(\frac{i}{n} - f_i \right) \right)$$

$$CM = \frac{1}{12n} + \sum_{i=0}^{n-1} \left(\frac{2i+1}{2n} - f_i \right)^2$$

$$WU = \frac{1}{12n} + \sum_{i=0}^{n-1} \left(\frac{2i+1}{2n} - f_i \right)^2 - n \left(\frac{1}{2} - \frac{1}{n} \sum_{i=0}^{n-1} f_i \right)^2$$

$$H1 = - \sum_{i=0}^{n-1} f_i \cdot \ln(f_i) + (1 - f_i) \cdot \ln(1 - f_i)$$

where

- n: sample size
- f_i : cumulative distribution function (of the distribution being tested) associated with the i^{th} (from 0 to n-1) observation sorted in ascending order

Monte-Carlo building of statistic-probability map

For $0 \leq k \leq 1000 \cdot K$

$f_i \leftarrow \text{Random}_{\text{Uniform}[0,1]}$, for $0 \leq i < n$

$(f_i)_{0 \leq i < n} \leftarrow \text{Sort}_{\text{ASC}}((f_i)_{0 \leq i < n})$

$\text{Observed}_k \leftarrow \text{Formula}((f_i)_{0 \leq i < n})$

The formula of
each statistic
enters here

EndFor

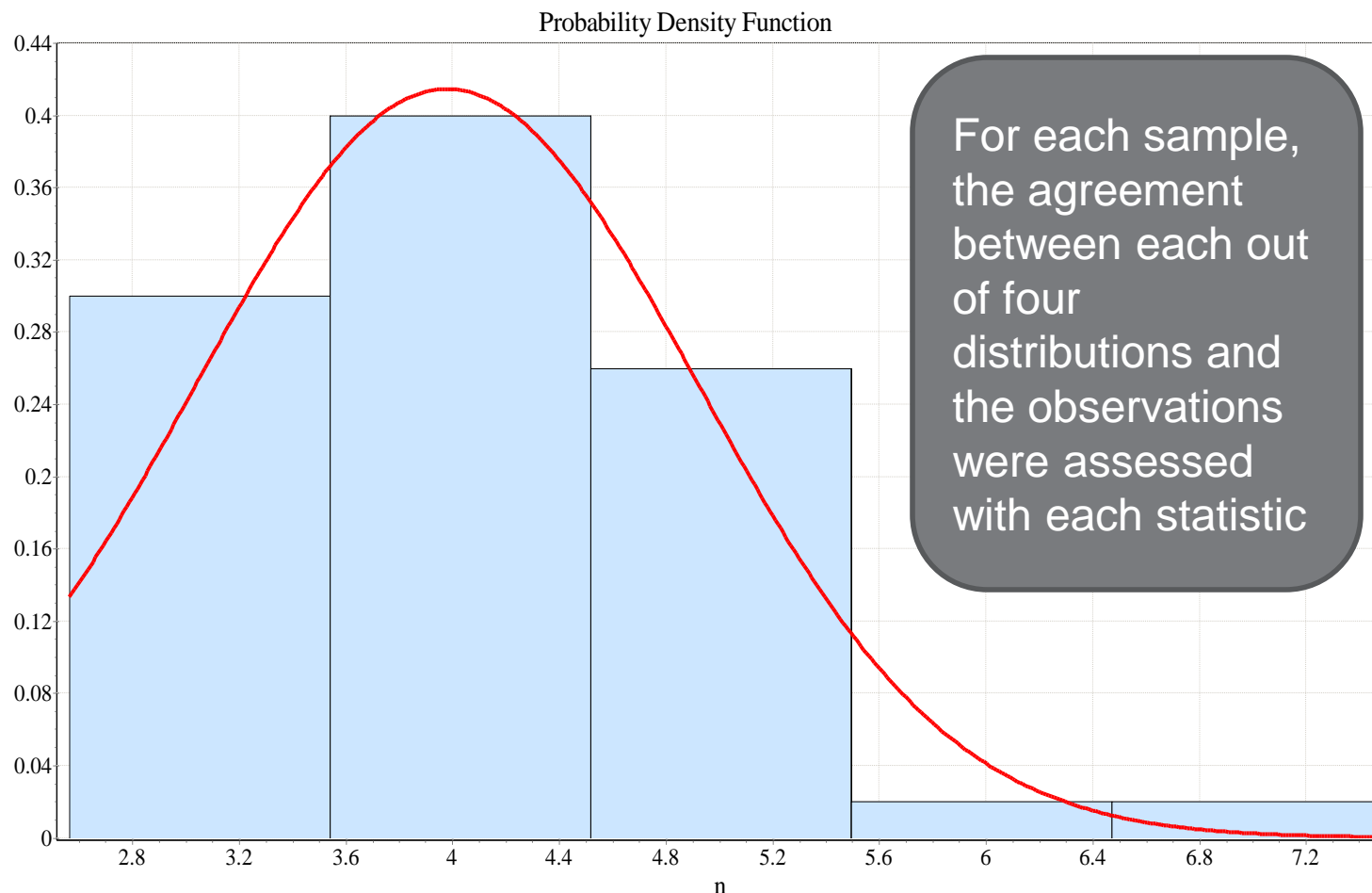
$(\text{Observed}_k)_{0 \leq k < K} \leftarrow \text{Sort}_{\text{ASC}}((\text{Observed}_k)_{0 \leq k < K})$

For $1 \leq j \leq 999$

$\text{Statistic}_{j/1000} \leftarrow \text{Mean}(\text{Observed}_{1000 \cdot K \cdot j - 1}, \text{Observed}_{1000 \cdot K \cdot j})$

EndFor

Material: 50 samples of properties and activities measurements of chemicals



Experimental values taken from literature

Sample of samples sizes are log-normal distributed ranging from $n = 13$ to $n = 1714$

Common distributions were included into analysis. Two of them are two-parametrical (LN and N) and two are three-parametrical

Log-Normal	$\text{LN}(x; \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{\left \frac{\ln x - \mu}{\sigma}\right ^2}{2}\right)$
Normal	$\text{N}(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\left \frac{x - \mu}{\sigma}\right ^2}{2}\right)$
Gauss-Laplace	$\text{GL}(x; \mu, \sigma, q) = \frac{p}{2\sigma} \frac{\Gamma^{1/2}(3/q)}{\Gamma^{3/2}(1/q)} \exp\left(-\frac{\left \frac{x - \mu}{\sigma}\right ^q}{\left(\frac{\Gamma(1/q)}{\Gamma(3/q)}\right)^{q/2}}\right)$
Fisher-Tippett	$\text{FT}(x; \mu, \sigma, q) = \frac{1}{\sigma\left(1 + \kappa \frac{x - \mu}{\sigma}\right)^{1+1/\kappa}} \exp\left(-\left(1 + \kappa \frac{x - \mu}{\sigma}\right)^{-1/\kappa}\right)$

Combining multiple tests

COMBINING INDEPENDENT TESTS OF SIGNIFICANCE

I have made several tests of significance on independent data. The tests I have made concern the difference between mean scores under two different treatments. Retaining the same order of differencing in the numerator of my t -test, I get t values of $-.68$, $+1.53$, $+2.21$, $+1.85$ with 6, 18, 22, and 25 degrees of freedom respectively. I would like to make a combined test of significance to test whether the difference between the means is positive. I do not want to try to form a single value of t from the original raw

If the test based on the product of probabilities from different trials is adopted, allowance can be made for the occasional occurrence of negative values of t by calculating the corresponding probability for a single tail. Negative values of t will then supply probabilities greater than .5, which do not greatly reduce the product and, since they introduce two additional degrees of freedom, will tend to lower its significance.

Thus with the illustrative values given in the problem, I obtain the following table:

n	t	P (single tail)	$\log_e 1/P$
6	$-.68$.739	.3025
18	1.53	.0717	2.6353
22	2.21	.01932	3.9466
25	1.85	.03809	3.2698
			<hr/> 10.1542

Doubling the total of the natural logarithms we have chi-square equal to 20.31 for 8 degrees of freedom, corresponding to a probability rather less than 1%. If, as seems here to be the case, consistent deviations in either direction would be deemed relevant, we may say that the probability is less than 2% for the

• Pearson-Fisher Chi-Square

data for various experimental reasons. In *Statistical Methods for Research Workers* R. A. Fisher presents a method of combining tests of significance, but it seems not to apply to the present situation since one of my differences is in the wrong direction. Also I fail to see how Fisher's method takes account of the varying degrees of freedom in the four sets of data. Can you tell me an appropriate procedure for combining these data? If you have any other advice about combining independent tests of significance, I would be glad to get it.

observation of deviations so large and so consistent in direction as those observed.

The differences in the degrees of freedom of the several trials, and also differences in their respective precisions, are, of course, reflected in the values of P to be multiplied together. The inclusion of the first trial lowers the significance of the test, and it may well be that this first trial was on so small a scale that it was likely *a priori* to do so. It would have been open to the experimenter to decide in advance not to conduct any trials with less than 15 degrees of freedom, or, equally legitimately, not to include such trials in his test. Such a decision must, however, be taken in advance and not after inspection of the result of the trials. Generally, therefore, if combination by the product method is intended, it will be worth while to aim at a succession of trials of approximately equal sensitiveness. More usually, however, the test will be employed to give provisional guidance based on all the evidence to hand, and this will often be of very variable precision.

It may be noticed that the same body of data may give a significant combined test in whichever direction the effect is tested. This would, of course, indicate reality of departure from the null hypothesis, but in discrepant directions in different cases, so that the results of the several trials would be heterogeneous.

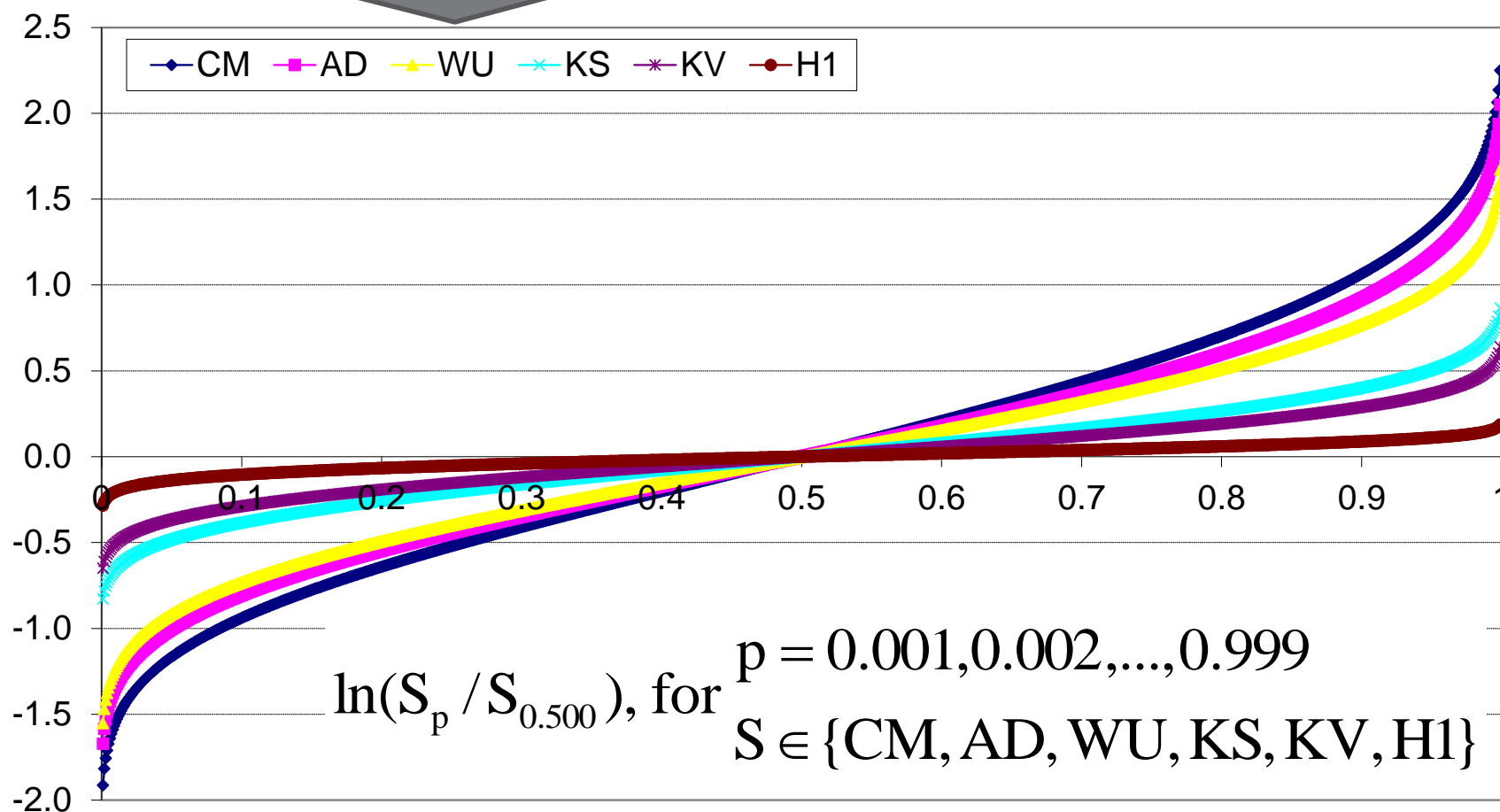
R. A. Fisher

RESULTS

Plots of statistic-probability maps

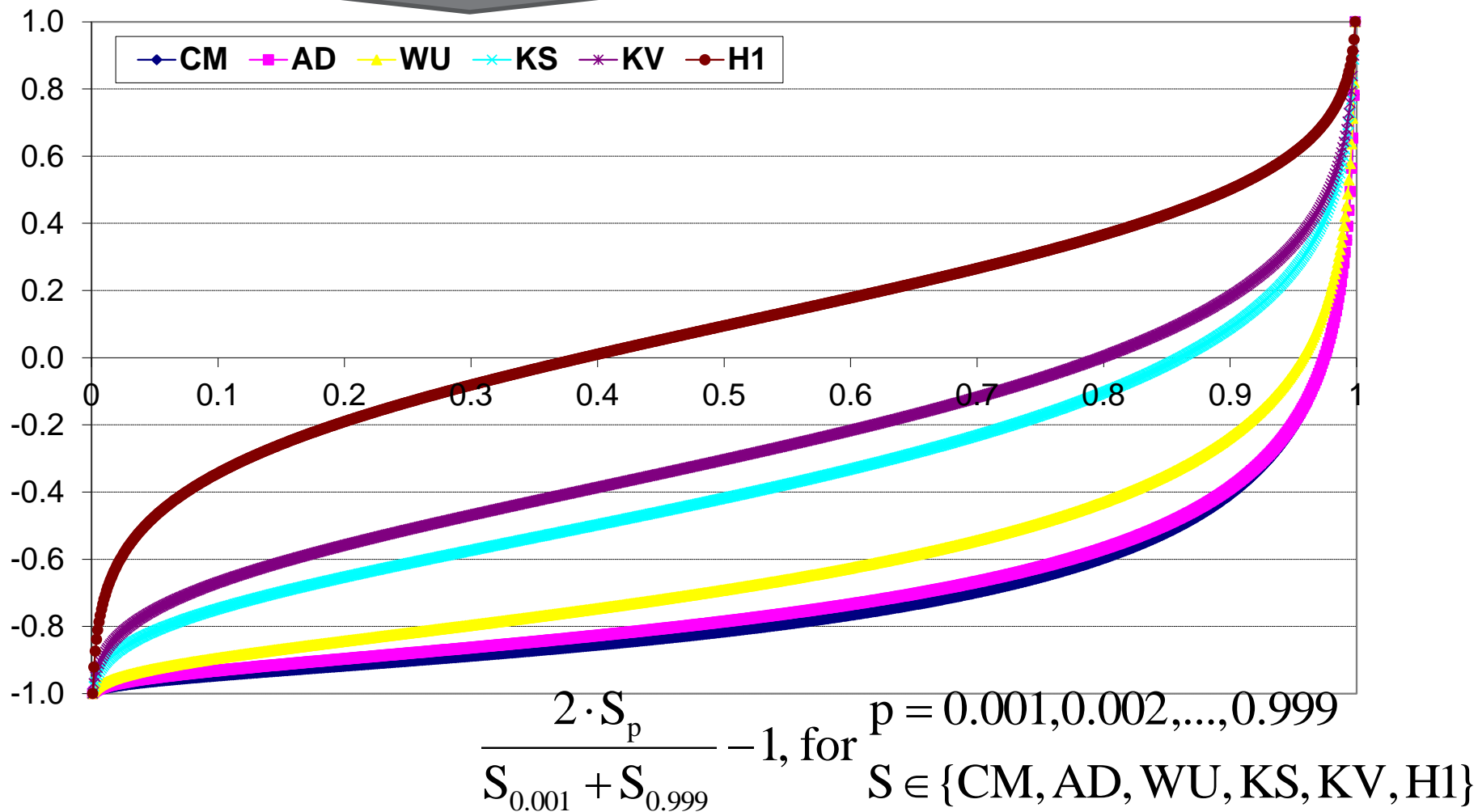
Statistics in log scale for $n = 25$

Statistics represented in logarithmic scale reveals that the highest resolution changing its value is for CM, while H1 tends to vary slowest



Statistics in relative scale for $n = 25$

Statistics represented in relative scale reveals on one hand moving of the inflexion point from near $p = 0$ for CM, AD and WU to the middle ($p=0.5$) for KS, KV and H1 and convergence (when n increases) to symmetrical shape for H1 on the other hand



RESULTS

Scenario 1: combining probabilities from AD, KS, CM, KV and WU

Scenario 2: combining probabilities from AD, KS, CM, KV, WU and H1

No single statistic should be used as 'absolute reference' when agreement between observation and model is assessed; the combined probability from the whole pool of statistics should be used instead

Individual rejections at 5% risk being in error for each statistic

Here is a simple proof why a single statistic should not be used for measuring the agreement between observation and the model

Distribution	AD	KS	CM	KV	WU	H1
Gauss-Laplace	9	12	11	19	17	0
Fisher-Tippett	6	5	4	13	11	3
Lognormal	4	7	4	18	16	3
Normal	8	14	10	21	20	0

50 samples were analyzed

H_0 (data follows a certain distribution) were rejected at 5% risk being in error differently by each statistic

Combined rejections at 5% risk being in error for each scenario

Distribution	Scenario 1	Scenario 2
Gauss-Laplace	19	19
Fisher-Tippett	13	13
Lognormal	20	18
Normal	21	21

50 samples were analyzed

H_0 (data follows a certain distribution) were rejected at 5% risk being in error differently by each scenario of combining statistics

DISCUSSION

By taking the 5% risk being in error as the threshold of rejecting H_0 (data follows a certain distribution)

- The scenario (1) not including H_1 have the tendency to reject the H_0 more often than any single statistic
- The scenario (2) including H_1 have the tendency to reject the H_0 closely to the highest rejection rate of any single statistic (was obtained the same rejection rate as KV)

CONCLUSIONS

- Shannon's statistic seems to have the tendency to fail to reject H_0 more often than all another investigated statistics
- However, its use in a battery of statistics in testing the H_0 , it changes the outcome not significantly (2 out of 73 less rejections of H_0) and making it more closely to the maximum rejection rate of a single statistic



2nd International Electronic Conference on Entropy and Its Applications

15-30 November 2015
chaired by Dr. Dawn E. Holmes

sponsored by



THANK YOU !

For any query, please do not hesitate to contact us:
lorentz.jantschi@gmail.com & sbolboaca@umfcluj.ro



Cluj-Napoca (Romania) by night