Statistical evidence methodology for model acceptance based on records

Mahdi Doostparast

(with Dr. Mahdi Emadi)

Department of Mathematics and Statistics

McMaster University

- Record data

- Statistical evidence

- Statistical evidence methodology for model acceptance based on records

Record data

Let $\{X_i, i \geq 1\}$ be a sequence of independent and identically distributed continuous random variables. An observation X_j will be called an upper record value if its

$$X_j > X_i \text{ for every } i < j$$

By defination, X_1 is an upper record value.

Then the (upper) record time $\{T_n, n \geq 1\}$ sequence is defined in the following manner:

$$T_1 = 1 \quad T_n = \min\{j > T_{n-1} : X_j > X_{T_{n-1}}\} \quad n \geq 2$$

The record value sequence defined by

$$R_n = X_{T_n}, \quad n = 1, 2, 3, \cdots$$

Example.

$$10,\ 11.5,\ 10.6,\ 16.3,\ 14.5,\ 17.25$$

$$R_1 = 10,\ R_2 = 11.5,\ R_3 = 16.3, R_4 = 17.25$$

$$T_1 = 1,\ T_2 = 2,\ T_3 = 4,\ T_4 = 6$$

Motivated by the reported frequency of record weather conditions, Chandler (1952) began studying the distributions of record data for independent and identically distributed sequence of random variables.

1. Ahsanullah, M. (1995). *Record Statistics*, Huntington, NY: Nova Science Publishers. Inc.

2. Arnold, B. C., Balakrishnan, N. and Nagaraja, H. N. (1998) *Records*, John Wiley, New York.

The joint distribution of the first m record value is given by

$$f(\underline{r}) = f(r_m) \prod_{i=1}^{m-1} h(r_i), \quad r_1 < r_2 < ... < r_m, \qquad (1)$$

where $\underline{r} = (r_1, r_2, ..., r_m)$ and $h(r_i) = \dfrac{f(r_i)}{1 - F(r_i)}$.

Statistical evidence

In Neyman-Pearson statistical theory, a test of two hypothesis H_1 and H_2 is represented as a procedure for choosing between them. But in applications, when an optimal test chooses H_2, it is often taken to mean that data are evidence favoring H_2 over H_1. This interpretation can be quite wrong.

When is it right to say that the observations are evidence in favor of one hypothesis *vis-a-vis* another?

The answer to this fundamental question has been known for at least century. However, neither the question nor its simple answer is to be found in most modern statistics text books. The reason is that

The decision-making paradigms since the work of Neyman and Pearson, have been formulated not in terms of interpreting data as evidence, but in terms of choosing between alternative course of action.

This lead to the current state of affairs in which the dominant (Neyman-Pearson) *theory* view common statistical procedures as decision-making tools, while

Much of statistical *practice* consists of using the same procedures for a different purpose, namely, interpreting data as evidence.

We need a mesure of support of H_1 against H_2.

Emadi and Arghami (2003) have studied some measures of support for statistical hypotheses

Let $\eta(> 0)$ be any measure of support of H_1 against H_2. Large (Small) values of η are interpreted as evidence given by data in favor of H_1 (H_2).

Misleading Evidence

The probabilities of observing strong misleading evidence under H_1 and H_2 are

$$M_1 := P(\eta < \frac{1}{k} | H_1 \text{ is true}), \tag{2}$$

and

$$M_2 := P(\eta > k | H_2 \text{ is true}), \tag{3}$$

respectively.

Weak Evidence

The probabilities of weak evidence under H_1 and H_2 are

$$W_1 := P(\frac{1}{k} < \eta < k | H_1 \text{ is true}), \tag{4}$$

and

$$W_2 := P(\frac{1}{k} < \eta < k | H_2 \text{ is true}), \tag{5}$$

respectively

Exponential Model

A random variable X is said to have an Exponential distribution, denoted by $X \sim Exp(\sigma)$, if its cdf is

$$F(x;\sigma) = 1 - e^{-\frac{x}{\sigma}}, \quad x \geq 0, \quad \sigma > 0, \tag{6}$$

and hence the pdf is given by

$$f(x;\sigma) = \frac{1}{\sigma} e^{-\frac{x}{\sigma}}, \quad x \geq 0, \quad \sigma > 0. \tag{7}$$

The Exponential distribution is applied in a wide variety of statistical procedures, especially in life testing problems. Data for survival and reliability analysis, as well as for biomedical and life testing studies have been modeled extensively by Exponential model.

Our Goal

Statistical evidence for model acceptance based on records

Suppose, we can observe the sequence of record values $R_1 = r_1, R_2 = r_2, \cdots, R_m = r_m$ from Exponential distribution.

Two hypothesis

$$\begin{cases} H_1 & : \quad \sigma = \sigma_1 \\ H_2 & : \quad \sigma = \sigma_2 \end{cases} \tag{8}$$

are under consideration where $0 < \sigma_1 < \sigma_2$.

Let λ be the likelihood ratio for the competing hypotheses H_1 and H_2, i.e.

$$\lambda = \frac{L_1}{L_2}, \tag{9}$$

where L_i is likelihood function under H_i.

We will use λ as a measure of support H_1 against H_2.

The likelihood function is given by

$$L(\sigma; \underline{r}) = \left(\frac{1}{\sigma}\right)^m e^{-r_m/\sigma}, \quad \sigma > 0.$$

The likelihood ratio for the competing hypothesis H_1 and H_2 is given by

$$\lambda = \left(\frac{\sigma_2}{\sigma_1}\right)^m e^{r_m(\frac{1}{\sigma_2} - \frac{1}{\sigma_1})}. \tag{10}$$

Misleading evidences are given by

$$M_1 = 1 - F_{\chi^2_{(2m)}}\left(2\frac{\ln(k) + m\ln(\sigma_2/\sigma_1)}{1-\frac{\sigma_1}{\sigma_2}}\right), \quad (11)$$

$$M_2 = F_{\chi^2_{(2m)}}\left(2\frac{-\ln(k) + m\ln(\sigma_2/\sigma_1)}{\frac{\sigma_2}{\sigma_1}-1}\right). \quad (12)$$

where $F_{\chi^2_v}(.)$ is cdf of a chisquare distribution with v degree of freedom.

It can be shown that by

$$\frac{2R_m}{\sigma} \sim \chi^2_{(2m)}.$$

We have

i. $\lim_{\sigma_2 \to +\infty} M_1 = \lim_{\sigma_2 \to +\infty} M_2 = 0$

ii. $\lim_{\sigma_2 \to \sigma_1^+} M_1 = \lim_{\sigma_2 \to \sigma_1^+} M_2 = 0$

iii. The point of global maximum of M_1 and M_2 can be obtained as a solution of the following non-linear equations

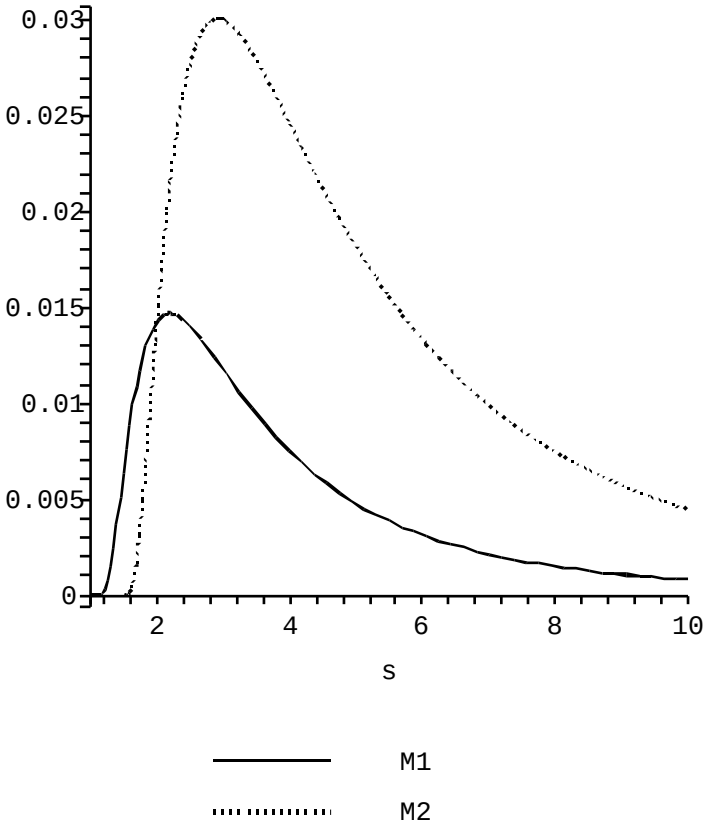
$$m - m \ln \sigma_1 + \ln k = m\frac{\sigma_2}{\sigma_1} - m \ln \sigma_2, \qquad (13)$$

and

$$m + \ln k + m \ln \sigma_1 = m\frac{\sigma_1}{\sigma_2} + m \ln \sigma_2, \qquad (14)$$

respectively.

iv. For $\sigma_2 < \sigma_1 e^{\sqrt[m]{k}}$, we have $M_2 = 0$.


0.03
0.025
0.02
0.015
0.01
0.005
0
2
4
6
8
10
s
M1
M2

It may be noticed that

when σ_2 tends to infinity, the distance between populations will increase as much as possible. Hence the probability of misleading tend to zero.

Also,

when σ_2 tends to σ_1, the distance between two populations will decrease as much as possible. So the M_1 and M_2 will be mixed with W_1 and W_2 and they tend to zero. As we will see later, in this case, for determination of true hypothesis we will therefore need more record values or more data (thus generating more record values).

Weak evidences are given by

$$\begin{aligned} W_1 &= F_{\chi^2_{(2m)}}\left(2\frac{\ln(k)+m\ln(\sigma_2/\sigma_1)}{1-\frac{\sigma_1}{\sigma_2}}\right) \\ &\quad -F_{\chi^2_{(2m)}}\left(2\frac{-\ln(k)+m\ln(\sigma_2/\sigma_1)}{1-\frac{\sigma_1}{\sigma_2}}\right). \end{aligned}$$

$$\begin{aligned} W_2 &= F_{\chi^2_{(2m)}}\left(2\frac{\ln(k)+m\ln(\sigma_2/\sigma_1)}{\frac{\sigma_2}{\sigma_1}-1}\right) \\ &\quad -F_{\chi^2_{(2m)}}\left(2\frac{-\ln(k)+m\ln(\sigma_2/\sigma_1)}{\frac{\sigma_2}{\sigma_1}-1}\right). \end{aligned}$$

we have

i. $\lim_{\sigma_2 \to +\infty} W_1 = \lim_{\sigma_2 \to +\infty} W_2 = 0.$

ii. $\lim_{\sigma_2 \to \sigma_1^+} W_1 = \lim_{\sigma_2 \to \sigma_1^+} W_2 = 1.$

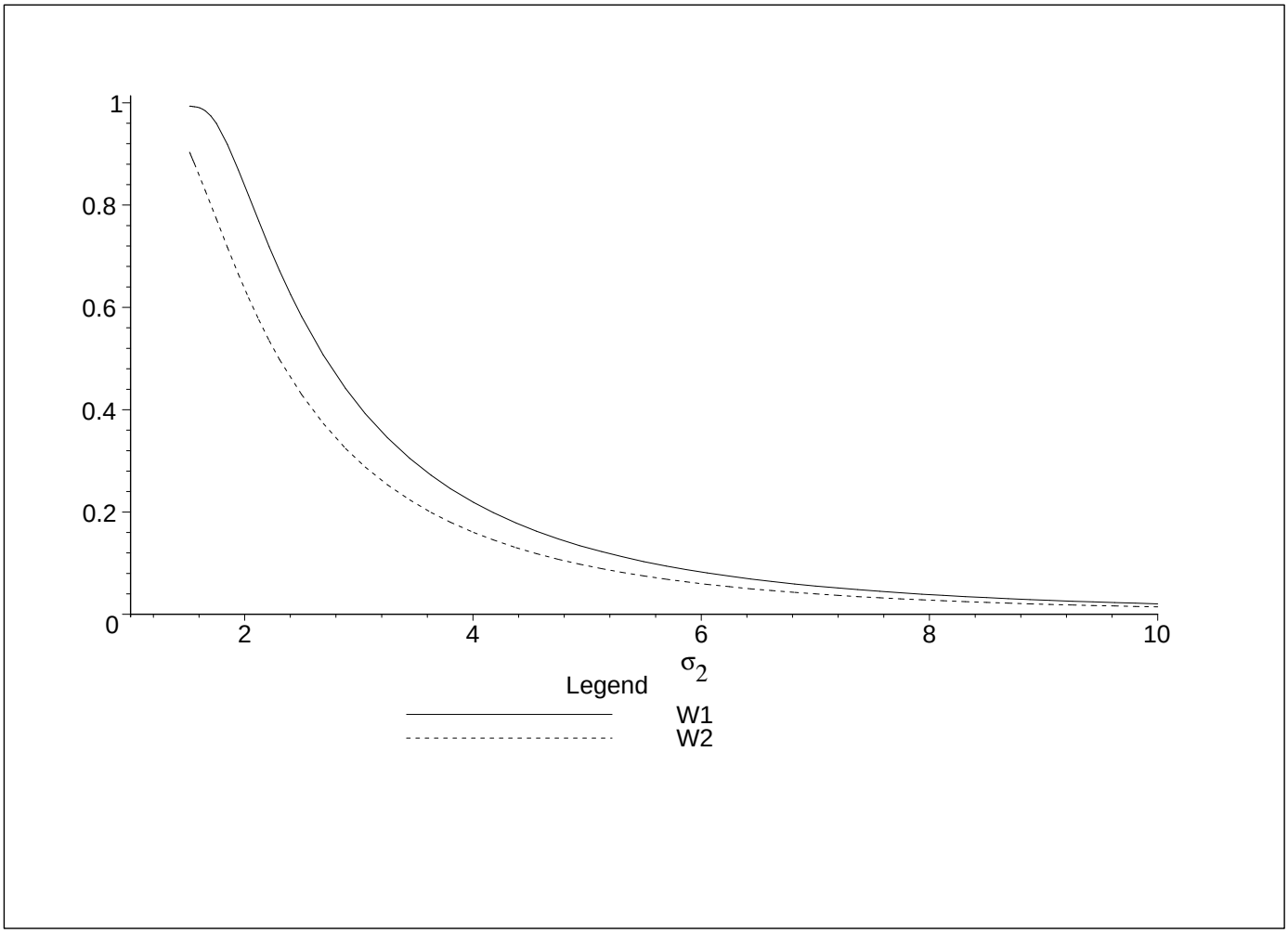
iii. The point of global maximum of W_1 and W_2 can be obtained as a solution of the following non-linear equations

$$\left(\frac{m(\frac{\sigma_2}{\sigma_1}-1)-(\ln k+m\ln\frac{\sigma_2}{\sigma_1})}{m(\frac{\sigma_2}{\sigma_1}-1)-(-\ln k+m\ln\frac{\sigma_2}{\sigma_1})}\right)=\left(\frac{-\ln k+m\ln\frac{\sigma_2}{\sigma_1}}{\ln k+m\ln\frac{\sigma_2}{\sigma_1}}\right)^{\frac{\sigma_2}{\sigma_1}(m-1)/(\frac{\sigma_2}{\sigma_1}-1)}, \tag{15}$$

and

$$\left(\frac{m(1-\frac{\sigma_1}{\sigma_2})-(\ln k+m\ln\frac{\sigma_2}{\sigma_1})}{m(1-\frac{\sigma_1}{\sigma_2})-(-\ln k+m\ln\frac{\sigma_2}{\sigma_1})}\right)=\left(\frac{-\ln k+m\ln\frac{\sigma_2}{\sigma_1}}{\ln k+m\ln\frac{\sigma_2}{\sigma_1}}\right)^{(m-1)/(\frac{\sigma_2}{\sigma_1}-1)}, \tag{16}$$

respectively.


1
0.8
0.6
0.4
0.2
0
2
4
6
8
10
σ2
Legend
W1
W2

When σ_2 tends to infinity, the distance between populations will increase as much as possible. So, even with few data we can make the decision about true hypothesis.

Also,

when σ_2 tends to σ_1, the distance between two populations will decrease as much as possible. Hence we have a few record values to determine true hypothesis, so we need more data.

Thank you.