The Kolmogorov Goodness-of-Fit Test (Kolmogorov-Smirnov one-sample test)

Introduction

- A test for goodness of fit usually involves examining a random sample from some unknown distribution in order to test the null hypothesis that the unknown distribution function is in fact a known, specified function.

- We usually use Kolmogorov-Smirnov test to check the normality assumption in Analysis of Variance.

- A random sample $X_1, X_2, \ldots, X_n$ is drawn from some population and is compared with $F^*(x)$ in some way to see if it is reasonable to say that $F^*(x)$ is the true distribution function of the random sample.

- One logical way of comparing the random sample with $F^*(x)$ is by means of the **empirical distribution function** $S(x)$

Definition Let $X_1, X_2, \ldots, X_n$ be a random sample. The **empirical distribution function** $S(x)$ is a function of x, which equals the fraction of $X_i s$ that are less than or equal to x for each x, $-\infty < x < \infty$, i.e

$$S(x) = \frac{1}{n} \sum_{i=1}^{n} I_{\{x_i \leq x\}}$$

- The empirical distribution function $S(x)$ is useful as an estimator of $F(x)$, the unknown distribution function of the $X_i s$.
- We can compare the empirical distribution function $S(x)$ with hypothesized distribution function $F^*(x)$ to see if there is good agreement.
- One of the simplest measures is the largest distance between the two functions $S(x)$ and $F^*(x)$,measured in a vertical direction.This is the statistic suggested by Kolmogorov (1933).

Kolmogorov-Smirnov test (K-S test)

- The data consist of a random sample $X_1, X_2, \ldots, X_n$ of size n associated with some unknown distribution function,denoted by $F(x)$
- The sample is a random sample
- Let $S(x)$ be the empirical distribution function based on the random sample $X_1, X_2, \ldots, X_n$. Let $F^*(x)$ be a completely specified hypothesized distribution function
- Let the test statistic T be the greatest (denoted by "sup" for supremum) vertical distance between $S(x)$ and $F^*(x)$.In symbols we say

$$T = \sup_x \mid F^*(x) - S(x) \mid$$

For testing

$H_0 : F(x) = F^*(x)$ **for all x from** $-\infty$ **to** ∞
$H_1 : F(x) \neq F^*(x)$ **for at least one value of x**

If T exceeds the 1-α quantile as given by Table then we reject H_0 at the level of significance α. The approximate p-value can be found by interpolation in Table.

Example

A random sample of size 10 is obtained: $X_1 = 0.621, X_2 = 0.503, X_3 = 0.203, X_4 = 0.477, X_5 = 0.710, X_6 = 0.581, X_7 = 0.329, X_8 = 0.480, X_9 = 0.554, X_{10} = 0.382$.The null hypothesis is that the distribution function is the uniform distribution function whose graph in Figure 1.The mathematical expression for the hypothesized distribution function is

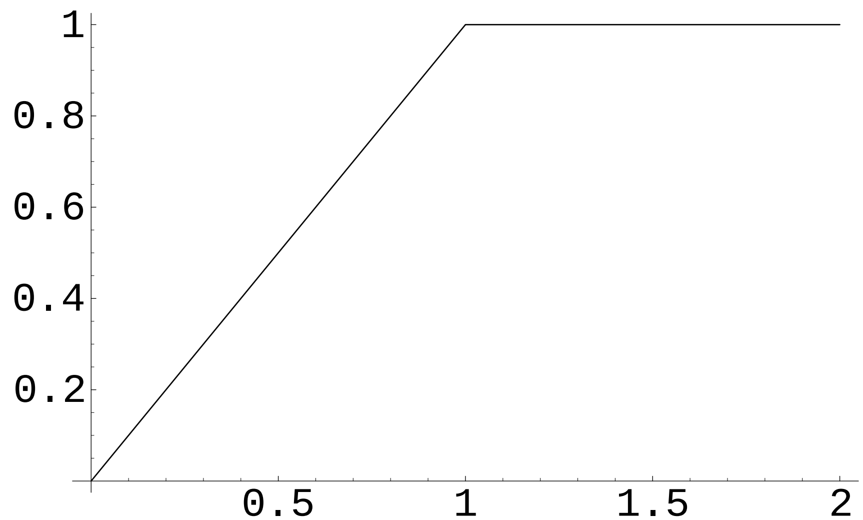
$$F^*(x) = \begin{cases} 0, & \textbf{if } x < 0 \\ x, & \textbf{if } 0 \le x < 1 \\ 1, & \textbf{if } 1 \le x \end{cases}$$

Formally , the hypotheses are given by

$H_0 : F(x) = F^*(x)$ **for all x from** $-\infty$ **to** ∞
$H_1 : F(x) \neq F^*(x)$ **for at least one value of x**

where $F(x)$ is the unknown distribution function common to the X_is and $F^*(x)$ is given by above equation.

Figure 1



The Kolmogorov test for goodness of fit is used.The critical region of size $\alpha = 0.05$ corresponds to values of T greater than the 0.95 quantile 0.409,obtained from Table for n=10.

The value of T is obtained by graphing the empirical distribution function $S(x)$ on the top of the hypothesized distribution function $F^*(x)$,as shown in Figure 2.The largest vertical distance separating the two graphs in Figure 2 is 0.290,which occurs at $x = 0.710$ because $S(0.710) = 1.000$

and $F^*(0.710) = 0.710$.In other words,

$$\begin{aligned} T &= \sup_x \mid F^*(x) - S(x) \mid \\ &= \mid F^*(0.710) - S(0.710) \mid \\ &= 0.290 \end{aligned}$$

Since T=0.290 is less than 0.409,the null hypothesis is accepted.The p-value is seen,from Table,to be larger than 0.20.

Figure 2

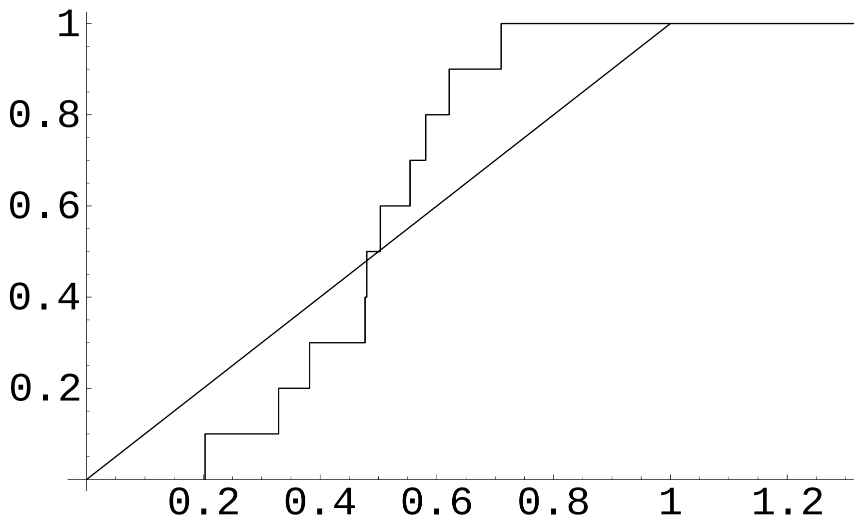
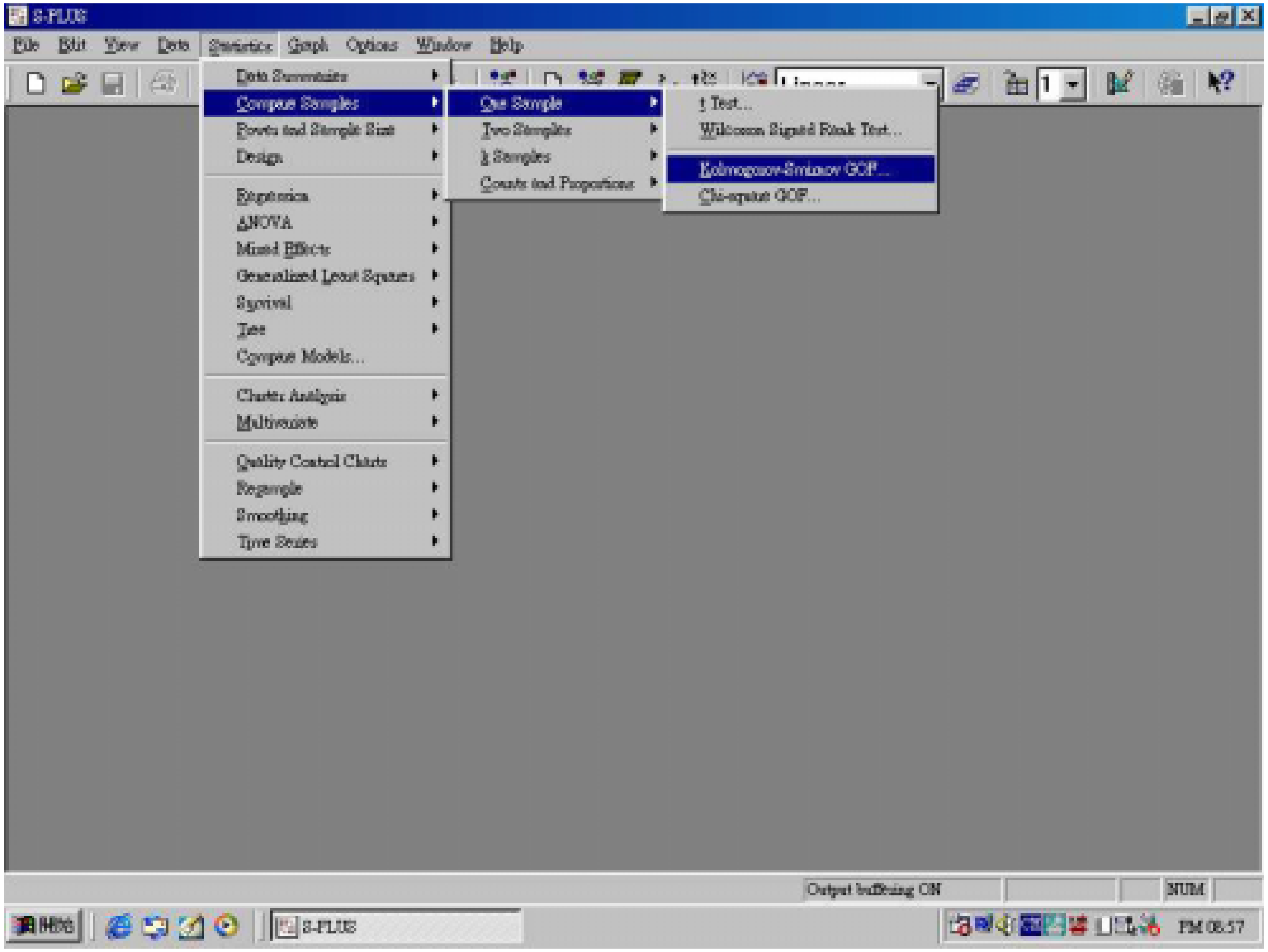
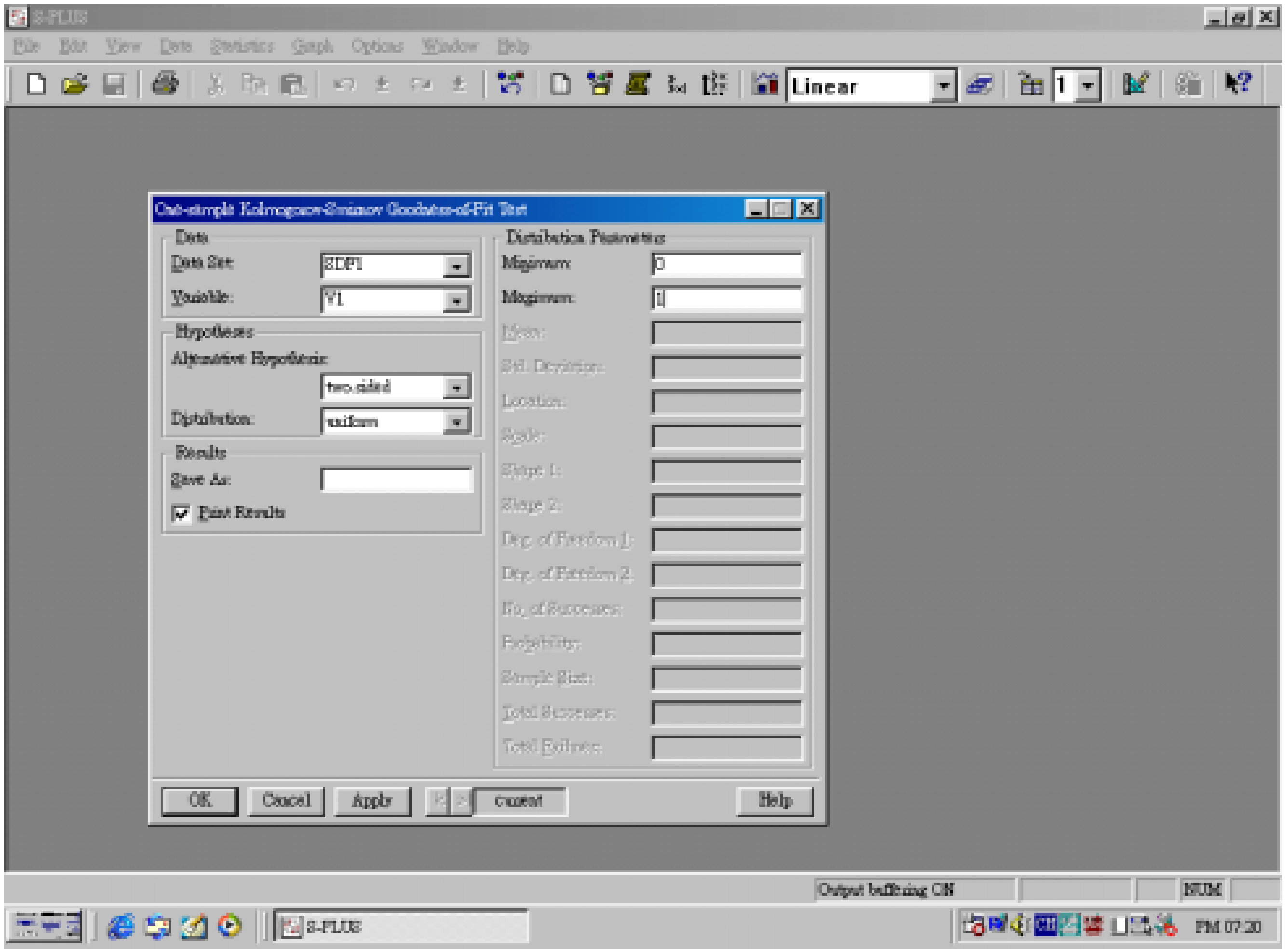


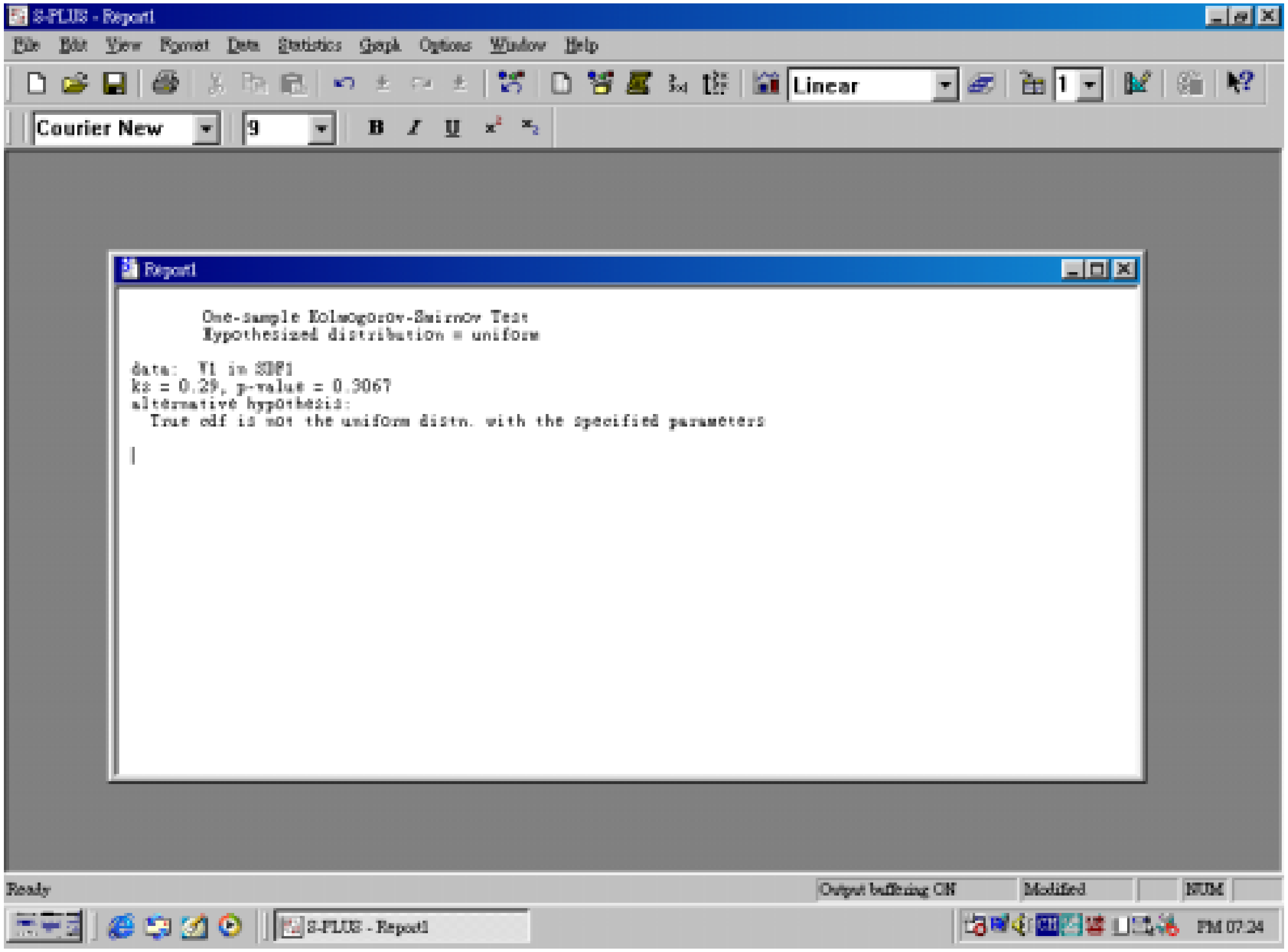
Table (Quantiles of the Kolmogorov Test Statistic)

n	p=0.80	p=0.90	p=0.95	p=0.98	p=0.99
1	0.900	0.950	0.975	0.990	0.995
2	0.684	0.776	0.842	0.900	0.929
3	0.565	0.636	0.708	0.785	0.829
4	0.493	0.565	0.624	0.689	0.734
5	0.447	0.509	0.563	0.627	0.669
6	0.410	0.468	0.519	0.577	0.617
7	0.381	0.436	0.483	0.538	0.576
8	0.358	0.410	0.454	0.507	0.542
9	0.339	0.387	0.430	0.480	0.513
10	0.323	0.369	**0.409**	0.457	0.489

Operation of S-PLUS




S-PLUS
Linear
Data
Data Set:
SDF1
Variable:
V1
Hypotheses
Alternative Hypothesis:
two.sided
Distribution:
uniform
Results
Save As:
Print Results
Distribution Parameters
Minimum:
0
Maximum:
1
OK
Cancel
Apply
current
Help
Output buffering ON
NUM
S-PLUS



From the result of computer software, we have the same conclusion as above, that is, the unknown distribution function is in fact the uniform distribution function.

Reference

W. J. Conover(1999),"*Practical Nonparametric Statistical*"
,3rd edition, pp.428-433 (6.1), John Wiley & Sons, Inc. New York.