PROCEEDINGS
of the
FOURTH INTERNATIONAL
CONFERENCE
ON TEACHING STATISTICS

VOLUME 1

25-30 July 1994
Marrakech, Morocco

Sponsored by
The International Statistical Institute
and
The National Institute of Statistics and Applied Economics, MOROCCO
Maple and the Computer Provide Synergism for Learning Probability and Statistics

Professor Elliot A. Tanis
Students Brian Goodman, Todd Busman, Joshua Levy
Hope College
Holland, Michigan 49422-9000
U.S.A.

There are many statistical packages that provide powerful methods for analyzing data both numerically and graphically. Maple provides additional opportunities for using the computer. Maple is a computer algebra system that has symbolic computing capabilities. It provides the basic tools for enhancing student learning in a probability and statistics course. In biology, synergism occurs when two (or more) substances or organisms achieve an effect of which each is individually incapable. We believe that Maple has similar potential for student understanding and learning.

A laboratory for a mathematical statistics and probability course has been developed that uses Maple as a basic ingredient. The central objective of the laboratory exercises is to improve students’ understanding of basic concepts. It does this in part by comparing empirical evidence with theoretical concepts. It also takes advantage of the symbolic computing and graphing capabilities of Maple. The following two applications illustrate this.

- The computer can be used as an experimental tool.

Let \( X_1, X_2, \ldots, X_n \) be a random sample of size \( n \) from a normal distribution with mean \( \mu \) and variance \( \sigma^2 \), \( N(\mu, \sigma^2) \). Then \( T = (\bar{X} - \mu)/(S/\sqrt{n}) \) has a \( t \) distribution with \( \nu = n - 1 \) degrees of freedom. Now, what can be said about the distribution of \( T \) if the sample is not from a normal distribution? Students can easily simulate samples from a uniform distribution or an exponential distribution or some other distribution to begin to answer this question. And they can consider the effect of the sample size, \( n \), on their answers.

- Maple can help solve tough and/or messy problems.

Let \( X \) equal the number of flips of a coin that are needed to observe successive heads. The p.d.f. of \( X \) is \( f(x) = f_{x-1}/2^x \), \( x = 2, 3, 4, \ldots \), where \( f_x \) is the \( x \)th Fibonacci number. That is, with \( f_1 = 1 \) and \( f_2 = 1 \), then \( f_x = f_{x-1} + f_{x-2} \), \( x = 3, 4, 5, \ldots \). It is an interesting exercise to show, with pencil and paper, that \( \sum_{x=2}^{\infty} f(x) = 1 \), \( E(X) = 6 \), and \( \text{Var}(X) = 22 \). Maple can do this for you with very simple commands. It is then an instructive exercise to simulate these results, comparing the empirical data with the theoretical model.

Using a few examples such as these, the role of Maple (and computer algebra systems, in general) will be discussed. Both positive enhancements of a probability and statistics course as well as potential pitfalls will be pointed out.
MAPLE AND THE COMPUTER PROVIDE SYNERGISM FOR LEARNING PROBABILITY AND STATISTICS

Elliot A. Tanis
Todd A. Busman
Bryan R. Goodman
Joshua D. Levy

Hope College
Holland, Michigan
49422-9000
U.S.A.
Introduction

Many statistical packages provide powerful methods for analyzing data both numerically and graphically. *Maple* provides additional opportunities for using the computer. *Maple* is a computer algebra system that has symbolic computing capabilities and provides the basic tools for enhancing student learning in a probability and statistics course. In biology, synergism occurs when two (or more) substances or organisms achieve an effect of which each is individually incapable. *Maple* has similar potential for student understanding and learning.

A laboratory for a mathematical statistics and probability course has been developed that uses *Maple* as a basic ingredient. The central objective of the laboratory exercises is to improve students’ understanding of basic concepts.

- *Maple* can help solve tough and tedious problems.
- *Maple* provides the tools for solving problems symbolically.
- Visualization provides insights.

Examples will illustrate these points.
Examples

1. In New Zealand a coin has a kiwi on one side and Queen Elizabeth the Second on the other side. Flip such a coin successively. Let \( X \) equal the number of flips that are required to observe a kiwi on successive flips. Determine the p.d.f., mean, variance, and standard deviation of \( X \). Support your theoretical results using simulation.
Let $X$ equal the number of flips of a coin that are required to observe a kiwi, $K$, on consecutive flips. The letter $H$ refers to the head of Queen Elizabeth the Second.
- On the $n$th flip, for any $n$, the number of possible heads equals the number of possible kiwis.

- Let $f_{nH}$ equal the number of heads on the $n$th flip.

- Let $f_{nC}$ equal the number of kiwis on the $n$th flip that complete a two-kiwi sequence.

- Let $f_{nP}$ equal the number of kiwis on the $n$th flip that have the potential for starting a two-kiwi sequence.

- Thus,

$$f_{nH} = f_{nC} + f_{nP}.$$ 

<table>
<thead>
<tr>
<th>$n$</th>
<th>$f_{nC}$</th>
<th>$f_{nP}$</th>
<th>$f_{nH}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>8</td>
<td>13</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>13</td>
<td>21</td>
</tr>
<tr>
<td>8</td>
<td>13</td>
<td>21</td>
<td>34</td>
</tr>
<tr>
<td>9</td>
<td>21</td>
<td>34</td>
<td>55</td>
</tr>
<tr>
<td>10</td>
<td>34</td>
<td>55</td>
<td>89</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Let $F_n$ equal the $n$th Fibonacci number where

$$F_1 = 1, F_2 = 1$$

and

$$F_n = F_{n-1} + F_{n-2}, \quad n = 3, 4, 5, \ldots$$

The p.d.f. of $X$ is

$$f(x) = \frac{F_{x-1}}{2^x}, \quad x = 2, 3, 4, \ldots$$

To find the sum, the mean, and the variance of $X$, recall that for $-1 < r < 1$,

$$\frac{a}{1 - r} = \sum_{x=0}^{\infty} ar^x$$

$$= a + ar + ar^2 + ar^3 + ar^4 + \cdots$$

$$\frac{a}{(1 - r)^2} = \sum_{x=1}^{\infty} axr^{x-1}$$

$$= a + 2ar + 3ar^2 + 4ar^3 + \cdots$$

$$\frac{2a}{(1 - r)^3} = \sum_{x=2}^{\infty} ax(x - 1)r^{x-2}$$

$$= 2a + 3 \cdot 2ar + 4 \cdot 3ar^2 + \cdots$$

Also recall that

$$F_x = \frac{1}{\sqrt{5}} \left[ \left( \frac{1 + \sqrt{5}}{2} \right)^x - \left( \frac{1 - \sqrt{5}}{2} \right)^x \right].$$
Claim: $\sum_{x=2}^{\infty} f(x) = 1$

$$\sum_{x=2}^{\infty} f(x) = \sum_{x=2}^{\infty} \frac{F_{x-1}}{2^x}$$

$$= \sum_{x=2}^{\infty} \frac{1}{\sqrt{5}} \left[ \left( \frac{1 + \sqrt{5}}{2} \right)^{x-1} - \left( \frac{1 - \sqrt{5}}{2} \right)^{x-1} \right] \left( \frac{1}{2^x} \right)$$

$$= \frac{2}{\sqrt{5}(1 + \sqrt{5})} \sum_{x=2}^{\infty} \frac{(1 + \sqrt{5})^x}{4^x} - \frac{2}{\sqrt{5}(1 - \sqrt{5})} \sum_{x=2}^{\infty} \frac{(1 - \sqrt{5})^x}{4^x}$$

$$= \vdots$$

$$= 1$$

Claim: $\mu = E(X) = 6$

$$E(X) = \sum_{x=2}^{\infty} x \frac{F_{x-1}}{2^x}$$

$$= \sum_{x=2}^{\infty} x \frac{1}{\sqrt{5}} \left[ \left( \frac{1 + \sqrt{5}}{2} \right)^{x-1} - \left( \frac{1 - \sqrt{5}}{2} \right)^{x-1} \right] \left( \frac{1}{2^x} \right)$$

$$= \frac{1}{2\sqrt{5}} \sum_{x=1}^{\infty} \left[ x \left( \frac{1 + \sqrt{5}}{4} \right)^{x-1} - x \left( \frac{1 - \sqrt{5}}{4} \right)^{x-1} \right]$$

$$= \frac{1}{2\sqrt{5}} \left[ \frac{1}{(1 - \frac{1 + \sqrt{5}}{4})^2} - \frac{1}{(1 - \frac{1 - \sqrt{5}}{4})^2} \right]$$

$$= \vdots$$

$$= 6$$
Claim: \( E[X(X - 1)] = 52 \)

\[
E[X(X - 1)] = \sum_{x=2}^{\infty} x(x - 1) \frac{1}{\sqrt{5}} \left[ \left( \frac{1 + \sqrt{5}}{2} \right)^{x-1} - \left( \frac{1 - \sqrt{5}}{2} \right)^{x-1} \right] \left( \frac{1}{2^x} \right)
\]

\[
= \frac{1}{2\sqrt{5}} \sum_{x=2}^{\infty} x(x - 1) \left[ \left( \frac{1 + \sqrt{5}}{4} \right)^{x-1} - \left( \frac{1 - \sqrt{5}}{4} \right)^{x-1} \right]
\]

\[
= \frac{1}{2\sqrt{5}} \left[ \frac{1 + \sqrt{5}}{4} \sum_{x=2}^{\infty} x(x - 1) \left( \frac{1 + \sqrt{5}}{4} \right)^{x-2} \right]
- \frac{1}{2\sqrt{5}} \left[ \frac{1 - \sqrt{5}}{4} \sum_{x=2}^{\infty} x(x - 1) \left( \frac{1 - \sqrt{5}}{4} \right)^{x-2} \right]
\]

\[
= \frac{1}{2\sqrt{5}} \left[ \frac{2(1+\sqrt{5})}{(1 - \frac{1+\sqrt{5}}{4})^3} - \frac{2(1-\sqrt{5})}{(1 - \frac{1-\sqrt{5}}{4})^3} \right]
\]

\[
= : \quad 52
\]

Claim: \( \sigma^2 = 22 \)

\[
\sigma^2 = E[X(X - 1)] + E(X) - \mu^2
\]

\[
= 52 + 6 - 36
\]

\[
= 22
\]

Thus, \( \sigma = \sqrt{22} = 4.69 \).

1.
2. Let $X_1, X_2, \ldots, X_n$ be a random sample of size $n$ from a normal distribution. Then

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

has a $t$ distribution with $n - 1$ degrees of freedom, where $\bar{X}$ and $S$ are the sample mean and sample standard deviation.

- What can be said about the distribution of $T$ if the sample does not come from a normal distribution?
- If, for example, the distribution from which the sample is taken is uniform or exponential, what can go wrong with the distribution of $T$?
- Can we say anything about the approximate distribution of $T$ in the latter cases and does the sample size affect this answer?

Recall that if $Z$ has a standard normal distribution and $U$ has a chi-square distribution with $r$ degrees of freedom, then

$$T = \frac{Z}{\sqrt{U/r}}$$

has a $t$ distribution with $r$ degrees of freedom, provided that $Z$ and $U$ are independent.
A characteristic of the normal distribution is that the sample mean, $\bar{X}$, and the sample variance, $S^2$, are independent.

Using simulation, we can simulate 200 (or 400) samples of size $n$, for several values of $n$, and calculate the values of $\bar{x}$, $s$, and

$$ t = \frac{\bar{x} - \mu}{s/\sqrt{n}}, $$

when sampling from

- a $N(0, 1)$ distribution,
- a $U(0, 1)$ distribution,
- an exponential distribution with mean $\theta = 1$.

Graphically we shall:

- Look at scatter plots of $\bar{x}$ vs. $s$;
- Construct the histogram of the data with a $t$ p.d.f. superimposed, and calculate the value of the chi-square goodness of fit test statistic;
- Construct the empirical and theoretical distribution functions and calculate the value of the Kolmogorov-Smirnov goodness of fit statistic.
3. Let $X$ be a random variable of the discrete type with p.d.f. $f(x)$ and space $R$. The factorial moment-generating function is defined by

$$\eta(t) = E(t^X) = \sum_{x \in R} t^x f(x).$$

Recall that

$$\eta(t) = E[e^{X \ln t}] = M(\ln t),$$

where $M(t)$ is the moment-generating function of $X$. Additionally,

$$\eta(1) = 1$$
$$\eta'(1) = E(X)$$
$$\eta''(1) = E[X(X - 1)]$$
$$\vdots$$
$$\eta^{(r)}(1) = E[X(X - 1) \cdots (X - r + 1)]$$

It follows that

$$\mu = E(X) = \eta'(1),$$

$$\sigma^2 = E[X(X - 1)] + E(X) - [E(X)]^2$$
$$= \eta''(1) + \eta'(1) - [\eta'(1)]^2.$$

Furthermore note that the coefficient of $t^x$ in $\eta(t)$ is $f(x) = P(X = x)$. 
Question: Red Rose Tea randomly puts one of 20 different porcelain miniature animals in a box containing 100 tea bags. Let $X$ equal the number of boxes of tea that must be purchased to collect the complete set of 20 animals. Find the p.d.f., mean, and variance of $X$.

Alternative Question: Roll an $n$-sided die. Let $X$ equal the number of rolls needed to observe each face at least once. Find the p.d.f., mean, and variance of $X$.

Let $Y_i$ equal the number of rolls needed to observe the $i$th different (new) face. The distribution of $Y_i$ is geometric with probability of success, with $k = n + 1 - i$,

$$p_i = \frac{n + 1 - i}{n} = \frac{k}{n}, \quad k = n, n - 1, \ldots, 1.$$ 
Furthermore, $Y_1, Y_2, \ldots, Y_n$ are independent and

$$X = Y_1 + Y_2 + \cdots + Y_n.$$ 
The factorial moment-generating function of $Y_i$ is

$$\eta(t) = \frac{(k/n) t}{1 - (1 - (k/n)) t}.$$ 
The factorial moment-generating function of $X$, subscripted by the number of faces on the die, is

$$\eta_n(t) = \prod_{k=1}^{n} \frac{(k/n) t}{1 - (1 - (k/n)) t}.$$
Factorial Moment-Generating Function of a Geometric Random Variable

Let $Y$ have a geometric distribution with probability of success $p$. That is, observe a sequence of Bernoulli trials with $p$ equal to the probability of success on each trial. Then $Y$ is equal to the trial number on which the first success occurs. The p.d.f. of $Y$ is

$$g(y) = (1 - p)^{y-1}p, \ y = 1, 2, 3, \ldots.$$ 

The factorial moment-generating function of $Y$ is

$$E(t^Y) = \sum_{y=1}^{\infty} t^y (1 - p)^{y-1}p$$

$$= \frac{p}{1 - p} \sum_{y=1}^{\infty} [t(1 - p)]^y$$

$$= \frac{p}{1 - p} \left( \frac{t(1 - p)}{1 - t(1 - p)} \right)$$

$$= \frac{pt}{1 - (1 - p)t}, \ |t| < \frac{1}{1 - p}.$$