MAPLE AND THE COMPUTER PROVIDE SYNERGISM FOR LEARNING PROBABILITY AND STATISTICS

Elliot A. Tanis
Hope College
Holland, Michigan
49422-9000
e-mail: tanis@math.hope.edu
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 tools for solving problems symbolically.
- Visualization - sometimes incorporating simulation - provides insights.

Examples will illustrate these points.
Examples

1. There are 20 classes: 16 with 25 students in each, three with 100 students in each, and one with 300 students.

- What is the average class size?
- If you take a random sample of 15 students, what would you expect the average of their class sizes to be?
2. An example about the intersection of intervals.

- Select a random permutation of the first $2 \cdot M$ positive integers.
- Construct $M$ intervals using, in the random permutation, the first two integers, the next two integers, ..., the last two integers. Order each pair so that the left endpoint is less than the right endpoint.
- Find the probability that at least one of these $M$ random intervals intersects each of the other intervals.
- Let $X$ equal the number of random intervals that intersect each of the $M$ intervals in a random permutation of $2 \cdot M$ integers. Define the p.d.f. of $X$ and find its mean, variance, and standard deviation.
- How does the value of $M$ affect your answers?
Example with $M = 5$: (Underlined intervals intersect all others.)

<table>
<thead>
<tr>
<th>X = 0</th>
<th>X = 1</th>
<th>X = 2</th>
<th>X = 3</th>
<th>X = 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1,2)</td>
<td>(4,7)</td>
<td>(3,5)</td>
<td>(6,9)</td>
<td>(8,10)</td>
</tr>
<tr>
<td>(8,9)</td>
<td>(4,6)</td>
<td>(1,10)</td>
<td>(2,5)</td>
<td>(3,7)</td>
</tr>
<tr>
<td>(6,7)</td>
<td>(5,8)</td>
<td>(3,10)</td>
<td>(2,9)</td>
<td>(1,4)</td>
</tr>
<tr>
<td>(2,10)</td>
<td>(1,8)</td>
<td>(4,5)</td>
<td>(6,7)</td>
<td>(3,9)</td>
</tr>
<tr>
<td>(3,7)</td>
<td>(1,9)</td>
<td>(4,8)</td>
<td>(2,10)</td>
<td>(5,6)</td>
</tr>
</tbody>
</table>
In Random Intervals in the December, 1990, issue of The American Mathematical Monthly, Joyce Justicz, Edward R. Scheinerman, and Peter M. Winkler claim that, for $k < M$, 

$$P(X \geq k) = \frac{2^k}{\binom{2k + 1}{k}}$$

It is possible to find, e.g., with $M = 4$, that 

$$P(X = 0) = \frac{13,440}{40,320} = \frac{1}{3}$$

$$P(X = 1) = \frac{10,752}{40,320} = \frac{4}{15}$$

$$P(X = 2) = \frac{6,912}{40,320} = \frac{6}{35}$$

$$P(X = 3) = \frac{0}{40,320} = 0$$

$$P(X = 4) = \frac{9,216}{40,320} = \frac{8}{35}$$
Means, variances, and standard deviations of $X$ for selected values of $M$ are:

<table>
<thead>
<tr>
<th>$M$</th>
<th>$\mu$</th>
<th>$\sigma^2$</th>
<th>$\sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1.3333</td>
<td>0.8889</td>
<td>0.9428</td>
</tr>
<tr>
<td>3</td>
<td>1.4667</td>
<td>1.7156</td>
<td>1.3098</td>
</tr>
<tr>
<td>4</td>
<td>1.5238</td>
<td>2.7875</td>
<td>1.5125</td>
</tr>
<tr>
<td>5</td>
<td>1.5492</td>
<td>2.6412</td>
<td>1.6252</td>
</tr>
<tr>
<td>10</td>
<td>1.5703</td>
<td>3.0823</td>
<td>1.7557</td>
</tr>
<tr>
<td>15</td>
<td>1.5708</td>
<td>3.1026</td>
<td>1.7614</td>
</tr>
<tr>
<td>20</td>
<td>1.5708</td>
<td>3.1034</td>
<td>1.7616</td>
</tr>
</tbody>
</table>

The mean of $X$, as $M \to \infty$, is

$$E(X) = 0 \cdot P(X = 0) + 1 \cdot [P(X \geq 1) - P(X \geq 2)] +$$

$$2 \cdot [P(X \geq 2) - P(X \geq 3)] + \cdots$$

$$= P(X \geq 1) + P(X \geq 2) + P(X \geq 3) + \cdots$$

$$= \sum_{k=1}^{\infty} \frac{2^k}{(2k + 1) \binom{2k}{k}}$$

$$= \frac{\pi}{2} = 1.570796$$
3. Let $X_1, X_2, \ldots, X_n$ be a random sample of size $n$ from a distribution with mean $\mu$ and variance $\sigma^2$. When $n$ is sufficiently large,

$$W_n = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

is approximately $N(0, 1)$.

- Determine when is $n$ sufficiently large if the p.d.f. of $X$ is

$$f(x) = \frac{3}{2} x^2, \quad -1 < x < 1.$$  

- Determine when $n$ is sufficiently large if the p.d.f. of $X$ is

$$f(x) = \frac{1}{\pi(1 + x^2)}, \quad -\infty < x < \infty.$$
4. 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 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,
- an exponential distribution with mean $\theta = 30$.

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.
5. 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,

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

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

$$p_i = \frac{n + 1 - i}{n}, \quad i = 1, 2, \ldots, n.$$ 

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, with $k = n + 1 - i$,

$$\eta(t) = \frac{(k/n) t}{1 - (1 - (k/n)) t}.$$ 

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

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

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

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

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