A NEW APPROACH TO PROBABILITY AND STATISTICS INSTRUCTION[1]

Zaven A. Karian, Department of Mathematics and Computer Science, Denison University

Symbolic manipulators provide a dynamic educational environment that can increase the productivity of novices as well as expert statisticians. They enable students to ask for a random sample of any size, compute various sample statistics, plot a histogram, see the relationship between the sample statistics and population parameters or the relationship between the sample histogram and the probability histogram—all without interrupting the flow of the reasoning process. Such flexibility develops deeper insights and the opportunity to explore different models significantly enhances intuition. This paper describes a statistics package based on the Maple symbolic computing system and gives several examples of its educational use.

1 Introduction

Statisticians were among the first mathematical scientists to use computers in their teaching and in various statistical analyses. Initially, when computation was confined to numeric calculations, it was natural to use computers to calculate sample statistics. With the advent of methods for the generation of random samples from various distributions, computer-based statistical simulations developed into a powerful technique with applications to industry and education (see, for example, Karian and Dudewicz (1999)). More recently, sophisticated graphic methods have allowed us to observe patterns by visualizing attributes of theoretical distributions and empirical data. A number of authors have already shown the value of symbolic computing to statistical research and pedagogy (see for example, Andrews and Stafford (1993), Baglivo (1995), Boglivo, Pagano and Spino (1993), Karian (1992), Karian and Tanis (1999) and Nash (1995)).

Widespread use of a symbolic manipulator in statistics should, minimally, offer the advantages of extensive statistical tables and a compendium of mathematical and statistical formulas and techniques, presented in nearly uniform notation. In a sense, it should all be there: probability density and distribution functions (in symbolic and numeric form), approximations, integrals, etc. Standard references can only present equations, tables, and figures as static entities: integrals are expressed in a "standard" form that might not be compatible with the statistician's experience or knowledge or with the problem under consideration. For this reason, the use of such references can be laborious and possibly counterproductive.

The manipulation of mathematical symbols, something that is done regularly in probability and statistics, somehow has not yet found its way into standard statistical packages such as MINITAB and SPSS. Moreover, the statistics packages that are bundled with symbolic computing systems such as *Mathematica* and *Maple*, are very

[1] Supported by a grant from the Mellon Foundation

limited in scope and provide only small fraction of the functionality that is necessary for statistical analyses. This paper will focus on a statistics package, developed by the author, and diverse ways in which this package can be used in statistics instruction. Section 2 gives an overview of the package and Sections 3 through 11 describe its various uses.

2 The Package

The most incomprehensible aspect of stats, the statistics package bundled with *Maple*, is its lack of *symbolic* features even though it is embedded in a symbolic computing environment. This package allows users, in a somewhat torturous way, to have access to such entities as probability density functions (p.d.f.s) of "standard" probability distributions but it does not allow users to manipulate (e.g., integrate) these expressions. By contrast, the stat package (distinguished from stats, the *Maple* package), gives users access to all structures in full symbolic form. The universal form of data representation throughout `stat` is the *Maple* list structure.

The package consists of 140 procedures and can be loaded in as a *Maple* package with the commands `> read("stat.all"):` and `> with(stat):` A detailed description of the package, its contents, and its uses is given in Appendix B of Karian and Tanis (1999). The package can be obtained through the author's web page (http//www.denison.edu/karian/).

3 Simple Simulations

Almost all statistical packages (e.g., MINITAB or SPSS) have built-in features that allow users to extract random samples from well-known distributions. What about sampling from arbitrary discrete or continuous distributions? If the statistical procedures were embedded within a symbolic computing environment, as is the case here, we could easily pass an expression representing a probability density function (p.d.f.) as a parameter to a procedure and have the procedure return a random sample from the specified distribution.

The interaction given below illustrates different ways of generating random samples. First, the command, A := Die(6,8); is used to simulate 8 rolls of a 6-sided die followed by Freq which gives the frequencies of 1, 2, …, 6 in the sample. DiscreteS([0, 1/2, 1, 1/2], 8); simulates 8 flips of a coin by sampling from the distribution of a random variable that assumes values 0 and 1, each with probability 1/2.

The two commands pdf:=x/10; and Sample:=DiscreteS(pdf, 1..4, 5); define and then generate a random sample from the discrete distribution with p.d.f.

$$f(x) = x/10, \qquad x = 1,2,3,4.$$

In the final four commands a larger sample (of size 300) is generated from this distribution, a histogram of the sample is computed as the plot structure EH, the probability histogram is obtained as the plot structure PH, and the two histograms are

displayed (Figure 1); the histogram depicted in solid lines is the probability histogram.

```
> A := Die(6,8);
```

$$A := [4,3,4,6,5,3,6,3]$$

```
> Freq(A, 1..6);
```

$$[0,0,3,2,1,2]$$

```
> A := Die(4, 400):
> Freq(A, 1..4);
```

$$[101,107,87,105]$$

```
> Coins := DiscreteS( [0, 1/2, 1, 1/2], 8);
```

$$Coins := [1,1,0,1,1,0,0,1]$$

```
> pdf := x/10;
```

$$pdf := \frac{1}{10}x$$

```
> Sample := DiscreteS(pdf, 1..4, 5);
```

$$Sample := [3,4,1,1,1]$$

```
> Sample := DiscreteS(pdf, 1..4, 300):
> EH := Histogram(Sample, 0.5..4.5, 4): PH := ProbHist(pdf, 1..4):
> display({EH,PH}); # See Figure 1.
```

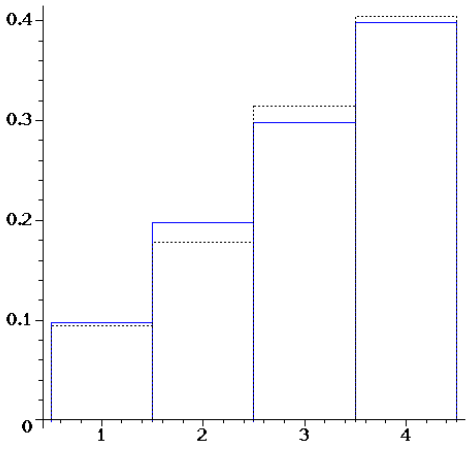

Figure 1: Comparison of sample and probability histograms

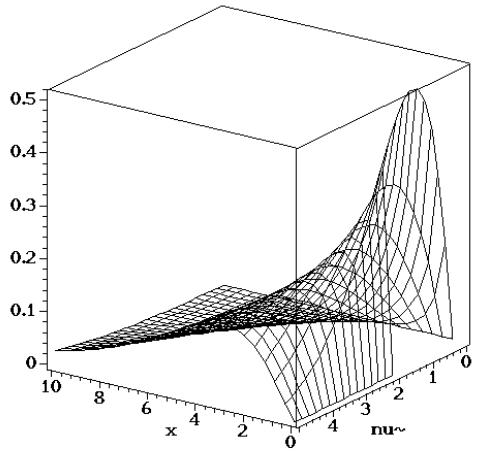

Figure 2: p.d.f.s of the χ^2 distribution

4 Investigating the Properties of a Distribution

Suppose that we are introducing students to the chi-square distribution in a first mathematical statistics course. There are a number of questions that immediately present themselves: What is the shape of the distribution? How is the shape influenced by its parameter ν ? What are the mean and variance of this distribution?

It may well be best for students to be left alone at this stage to answer as many of these questions as they can, with or without the mathematical proofs that may eventually be supplied by the instructor or a suitable text.

Through a set of *Maple* commands, such as the ones given below, a student can consider, in a natural way, all the questions that were posed. First, the chi-square p.d.f. is accessed by ChisquarePDF(nu, x); and then the mean and variance of the distribution are obtained. (Note that on some occasions simplify has to be applied to computations.) The animate and plot3d reveal a surprise — the p.d.f. undergoes a dramatic change in shape about $\nu = 2$ (the result of the plot3d command is given in Figure 2). To determine the mode, if it exists, or to determine the values of ν for which it does not exist, the p.d.f. is differentiated and its critical points determined. Clearly, the 0 critical point is not of interest and it can be seen that the p.d.f. attains a maximum when $x=\nu-2$, a condition that cannot arise for $\nu < 2$ because x must be positive.

```
> assume(nu>0):  f := ChisquarePDF(nu,x);
```

$$\frac{x^{1/2\nu-1}e^{-1/2x}}{\Gamma(1/2\nu)2^{1/2\nu}}$$

```
> mu := int(x*f, x=0..infinity);
```

$$\mu := \nu$$

```
> Var := simplify(int(x^2*f, x=0..infinity)-mu^2);
```

$$Var := 2\nu$$

```
> animate(f, x=0..25, nu=1..15): plo3d(f, x=0..10, nu=0..5);
> df := simplify(diff(f, x));
```

$$\frac{2^{-1-1/2\nu}x^{1/2\nu-2}e^{-1/2x}(\nu-2-x)}{\Gamma(1/2\nu)}$$

```
> solve(df=0,x);
```

$$0, \nu-2$$

5 Confidence Intervals

Students' understanding of confidence intervals will improve if they can look at many intervals in a fixed situation. For example, for a given significance level, α students can consider 100(1- α)% confidence intervals of the mean, μ, based on random samples of size n from $N(\mu,\sigma^2)$, the normal distribution with mean μ and variance σ^2. Different random samples will lead to different 100(1- α)% confidence

intervals, some of which will contain μ. It is worthwhile to show students that the proportion of intervals that contain μ will be approximately 1- α. Additionally, it is possible to graphically illustrate how knowledge of σ^2 affects the confidence intervals. In the following interaction random samples of size 5 are generated from $N(40,12)$ and with the iteration provided by the *Maple* seq command, ListOfSamples becomes a list of 50 such samples. Only the outputs associated with the ConfIntPlot commands are shown; these are plots that display the confidence intervals of the mean with the true mean, 40, marked as a vertical line making it easy to determine how many of the 50 intervals contain 40 (in the illustration below, when α is set to 0.2, μ is in 38 of the 50 80% confidence intervals). In the first graph (Figure 3, left) no knowledge of σ^2 is assumed whereas in the second graph (Figure 3, right) σ^2 is set to 12.

```
> ListOfSamples := [seq(NormalS(40,12,5), i=1..50)]:
> CIVarUnknown := ConfIntMean(ListOfSamples, 80):
> ConfIntPlot(CIVarUnknown, 40); # See Figure 4 - left
> CIVarKnown := ConfIntMean(ListOfSamples, 80, 12):
> ConfIntPlot(CIVarKnown, 40); # See Figure 4 - right
```

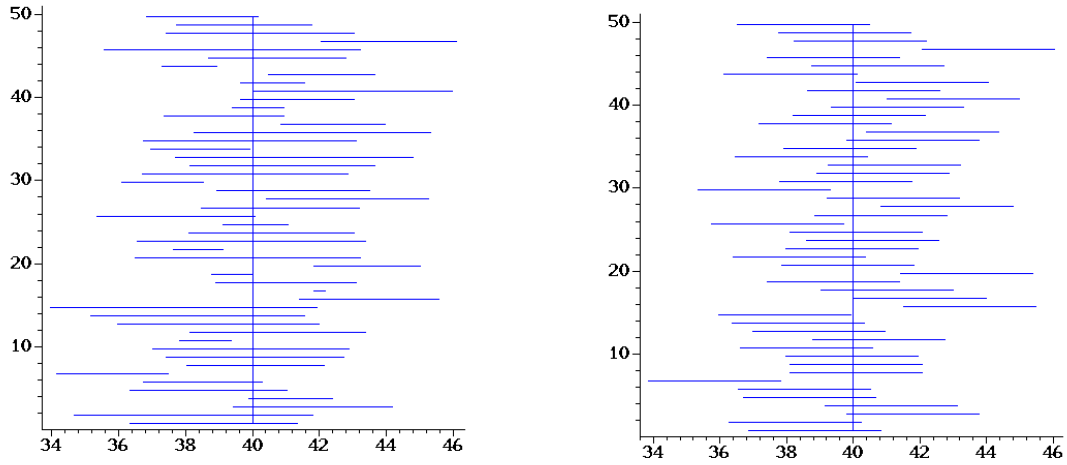

Figure 3: Confidence intervals with σ^2 unknown (left) and σ^2 known (right)

6 Understanding an Important Theorem

To impress upon students the broad applicability of the Central Limit Theorem (CLT) we can empirically show that if $\overline{X}$ is the mean of a random sample of size *n* from *any* distribution with mean μ and variance σ^2, then for large *n* the distribution of $\overline{X}$ can be approximated by $N(\mu,\sigma^2/n)$. We can be more convincing in our attempt to illustrate this if we choose an arbitrary distribution, preferably one whose shape is as distinct from the "bell shape" as possible. We could, for example, choose the distribution with density function

$$f(x) = (3/2)x^2, -1 \le x \le 1$$

which is U-shaped, extract random samples of small size (say n=4), compute the means of these samples, and compare their distribution to the original one.

The availability of *symbolic* manipulations allows us, as shown below, to define the p.d.f., verify that it is indeed a p.d.f., and obtain its mean, variance, distribution function, and a graph of the p.d.f.

```
> f := x -> (3/2)*x^2;
```

$$f := x \to \frac{3}{2}x^2$$

```
> int(f(x), x=-1..1);
```

$$1$$

```
> mu := int(x*f(x), x=-1..1);
```

$$\mu := 0$$

```
> var := int(x^2*f(x), x=-1..1)- mu^2;
```

$$Var := \frac{3}{5}$$

```
> F := int(f(t), t = -1..x);
```

$$F := \frac{1}{2}x^3 + \frac{1}{2}$$

```
> P := plot(f(x), x=-1..1);
```

Next, we want to extract samples of size 4 from this *Continuous* distribution, defined on the interval [-1,1]. This is done by the "inner" command, ContinuousS(f, -1..1, 4), on the following *Maple* line; the outer portion of the line forces 300-fold repetition, making Samples a list of 300 random samples, each of size 4. Since the output associated with this operation is rather extensive, we suppress it by using ":" for the termination symbol. The subsequent command computes Means, the means of the 300 random samples. (Note that ContinuousS and Mean are part of the statistics supplement and not a part of *Maple*.)

```
> Samples := [seq(ContinuousS(f(x), -1..1, 4), I=1..300)]:
> Means := [(Mean(Samples[i]), I=1..300)]:
```

The next three lines compute and store three graphs: The histogram of sample means, the graph of the p.d.f., and the graph of the normal distribution that is prescribed by the CLT. These graphs are then displayed on one set of axes with the display command.

```
> H := Histogram(Means, -1..1, 7): n := NormalPDF(mu, var/4, x):
> N := plot(n, x=-1..1):
> display({P, H, N}); # See Figure 4
```

Figure 4 gives two plots: one obtained through the *Maple* interaction above with n=4 (on the left) and a similar graph obtained with n=8 (on the right).

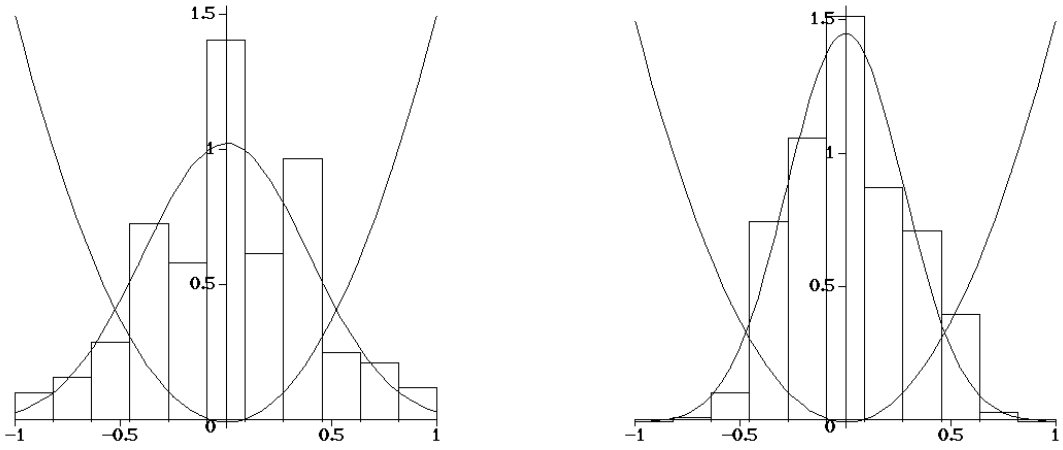

Figure 4: Illustration of the CLT with $n = 4$ (left) and $n = 8$ (right)

7 Relationships Between Random Variables

With some encouragement, students can discover a variety of relationships between specified random variables. For example, the distribution of $Y = Z^2$, where Z is the standard normal random variable, can be considered empirically before it is established that Y has a chi-square distribution with one degree of freedom, $\chi^2(1)$. In the following *Maple* interaction Z1 represents a random sample of size 500 from the standard normal distribution and Y consists of the squares of these observations. Next, the histogram YH of Y and the graph of the $\chi^2(1)$ p.d.f. are obtained as *Maple* plot structure and displayed together in Figure 5 (left).

```
> Z1 := NormalS(0,1,500):
> Y := [seq(Z1[i]^2, i=1..500)]:
> YH := Histogram(Y, 0..12, 16):
> CH1 := plot(ChisquarePDF(1,x), x=0..12):
> display({YH, CH1}); #See Figure 5 - left
```

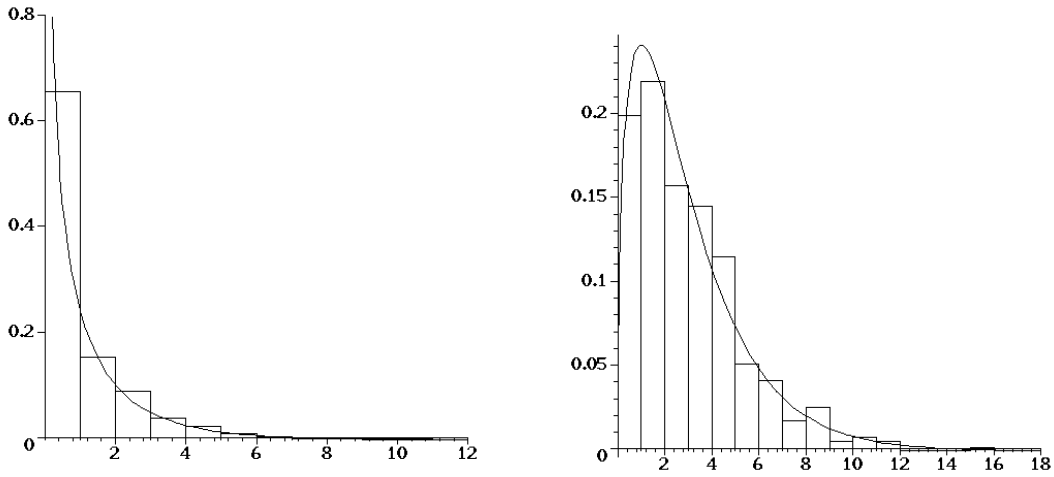

Figure 5: compared with (left) and compared with (right)

To see the more general pattern, two additional random samples, Z2 and Z3, each of size 500 are obtained and Y is set to the sum of squares of three standard normals. This time the histogram of Y is compared to the $\chi^2(3)$ p.d.f. The resulting plot is shown as Figure 5(right).

```
> Z2 := NormalS(0,1,500): Z3 := NormalS(0,1,500):
> Y := [seq(Z1[i]^2 + Z2[i]^2 + Z3[i]^2, i=1..500)]:
> YH := Histogram(Y, 0..18, 24):
> CH3 := plot(ChisquarePDF(3,x), x=0..12):
> display({YH, CH3});
```

REFERENCES

Andrews, D. F. and Stafford, J. E. (1993). "Tools for the Symbolic Computation of Asymptotic Expansions," *J. R. Statist. Soc. B*, V.5, No.3, pp.613--627.

Baglivo, J. (1995). "Computer Algebra Systems: Maple and Mathematica," *The American Statistician*, V.49, No.1, pp.86--92.

Baglivo, J., Pagano, M. and Spino, C.(1993). "*Symbolic Computation of Permutation Distributions*," Proceedings of the Statistical Computing Section, ASA, pp. 218--223.

Karian, Z. A. (Ed.), (1992). *Symbolic Computation in Undergraduate Mathematics Education*, The Mathematical Association of America, Washington, D.C., U.S.A.

Karian, Z. A. and Dudewicz, E.J. (1999). *Modern Statistical, Systems and GPSS Simulation* Second Edition, W. H. Freeman.

Karian, Z. A. and Tanis, E. A. (1999). *Probability and Statistics Explorations with Maple* Second Edition, Prentice Hall, Inc., Englewood Cliffs, NJ 07632, U.S.A.

Nash, J. C. (1995). "Computer Algebra Systems: DERIVE," *The American Statistician*, V.49, No.1, pp.93--99.