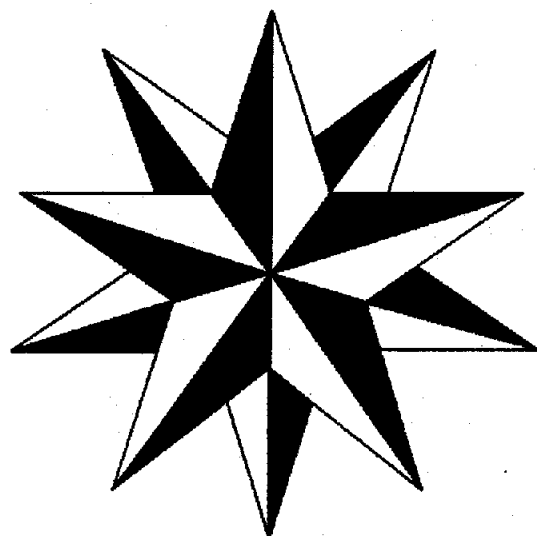


**A DEMONSTRATION OF THE UTILITY OF SAS IN
CREATING EXAMPLES TO AID IN THE TEACHING
OF STATISTICAL CONCEPTS**



**Earnest Davenport, Jr.
University of Minnesota
Department of Educational Psychology**

A Demonstration of the Utility of SAS in CREATING Examples to Aid in the Teaching of Statistical Concepts

Ernest C. Davenport, Jr., University of Minnesota
Department of Educational Psychology, 320 Burton Hall, Minneapolis, MN 55455

The use of SAS for statistical analysis and data management is well established. The current paper makes use of SAS in a different context. Here SAS is used to create examples to aid the teaching of statistical concepts. Due to time and space limitations the teaching of only one concept is illustrated; the Central Limit Theorem (CLT). Several side issues are also addressed as they relate to the example developed below. Note that concepts other than the CLT can also be modelled to allow instructors to develop material for teaching a wide range of statistical concepts such as: analysis of variance, regression analysis, factor analysis, reliability, validity, hypothesis testing, critical values, etc.

Demonstration of the Central Limit Theorem

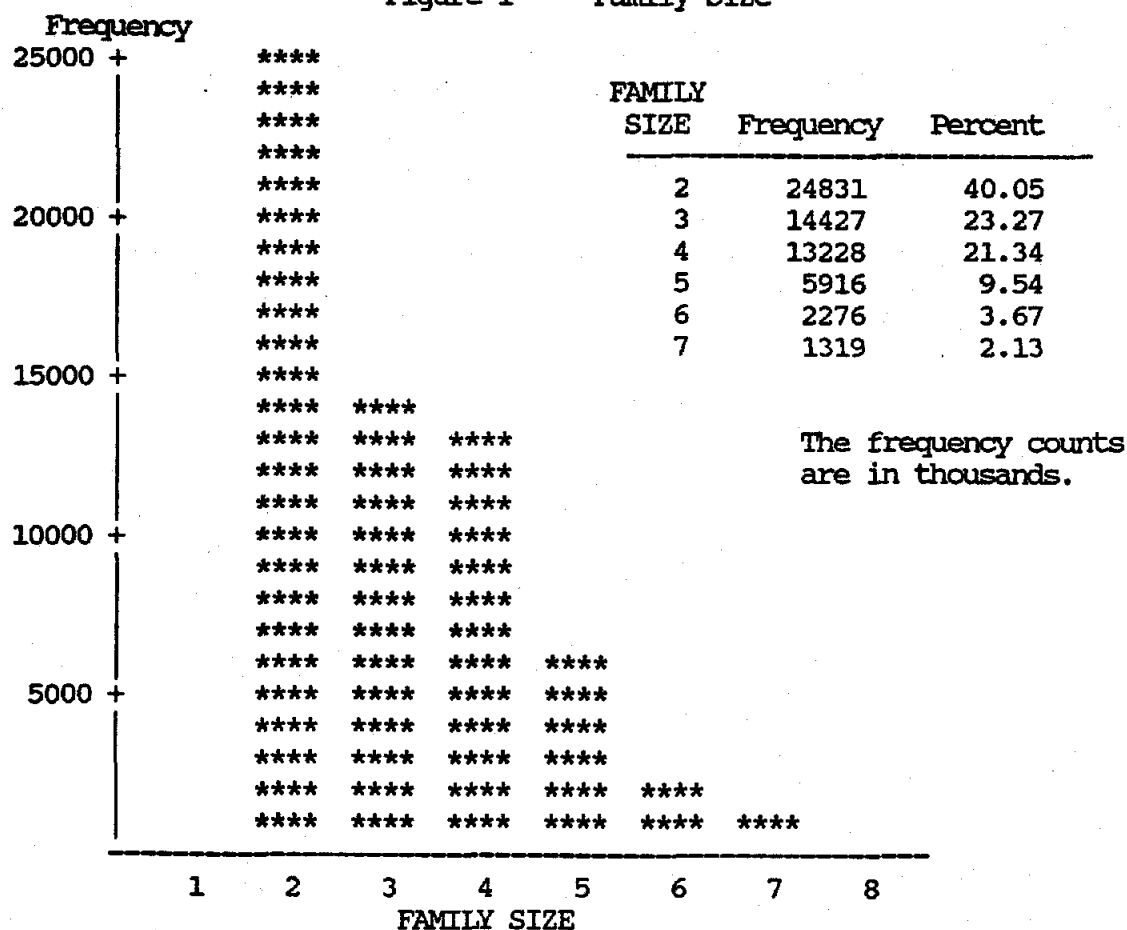
The Central Limit Theorem states, "...the random sampling distribution of means tends toward a normal distribution irrespective of the shape of the population of observations sampled and that the approximation to the normal distribution becomes increasingly close with increase in sample size." (Minium, 1978, p.224). CLT forms the basis of statistical inference involving means and is a part of virtually all statistics courses which feature hypothesis testing. Providing concrete examples that provide insight into the theorem and its implications can be very valuable to students.

The first step is to provide students with a distribution of "non-normal" data. The next step is to provide a mechanism with which the students can take multiple samples of the same size from this distribution. Then, one can create a frequency distribution of these sample estimates. Finally, these values can be evaluated with respect to statistical theories concerning the central limit theorem, expected values, statistical inference, etc. An instructor could have students perform the above tasks, or alternatively, have SAS do most of the work while leaving the students with a meaningful exercise.

The specific example for this paper is based on family size data presented in The 1986 information please almanac (p.749); reproduced here as Figure 1. This data was chosen due to the positive skew of the raw score distribution. The values for family size range from 2 to 7 with 7 depicting seven or more family members. Numbers for family size are represented in thousands. (For this example I use a family size value of 7 to depict "7" family members.) For pedagogical purposes statistics students can be given the following scenario. A government official wishes to confirm the results presented in Figure 1 without having to interview the entire population. She hires 500 research associates (RAs) to canvass the country visiting randomly sampled families. [Here the instructor can address simple random sampling; SRS. Subjects can be anywhere within the country. Costs and benefits for RAs to focus their efforts in one geographical area can be raised (i.e. difference between SRS and cluster sampling). Additionally, other sampling schemes can be discussed (and modelled) relative to the above scenario.] The first 250

RAs are to visit 10 families and the remaining 250 are to visit 15 families to espy the size of the family. (Here the instructor can raise questions about method of data collection; visits versus telephone versus mail, etc.). The resulting data consist of 6250 ($250 \times 10 + 250 \times 15$) observations. In addition to the raw data, each RA reports the mean for their sample.

Figure 1 - Family Size



Program Description and Results

Figure 2 is a SAS program that models the family size data for the above scenario. It demonstrates how to create a sampling distribution and provides results to help explain the CLT. The SAS code and its output is explained in terms of its value to the instructor (process) and/or the students (results).

Lines 1-3 is a comment about the purpose of the program. Options are set in line 5 to control the size of the subsequent bar diagrams. A library is named so that resulting data can be saved in a permanent SAS data set. When using SAS random number generators to create data, a different set of data is created on each run. Thus, if additional analyses are needed on a given set of data, it must be recalled from a permanent data set. Line 8 defines two sample size conditions. (This is included so the students can compare results

Figure 2 - The Program

```

1  * THIS PROGRAM DEMONSTRATES FEATURES OF THE CENTRAL LIMIT THEOREM AND
2  SAMPLING DISTRIBUTIONS USING DATA MODELLED FOR THE SEVERELY POSITIVELY
3  SKEWED FAMILY SIZE DATA;
4
5  OPTIONS PS=40  LS=64;
6  LIBNAME STORE '\SASWK\';
7  DATA STORE.SUGI (DROP = N K T);
8      DO SSIZE = 1 TO 2;
9          DO REP = 1 TO 250;
10             N = 10 + 5*(SSIZE = 2);
11             DO K = 1 TO N;
12                 T = UNIFORM(0);
13                 FSIZE = 2 * (T < 0.4005) +
14                     3 * (0.4005 <= T < 0.6332) +
15                     4 * (0.6332 <= T < 0.8466) +
16                     5 * (0.8466 <= T < 0.9420) +
17                     6 * (0.9420 <= T < 0.9787) +
18                     7 * (T >= 0.9787);
19                 OUTPUT;
20             END;
21         END;
22     END;
23 CARDS;
24
25 PROC FREQ DATA = STORE.SUGI; TABLE FSIZE;
26 PROC MEANS DATA = STORE.SUGI; VAR FSIZE;
27 PROC CHART DATA = STORE.SUGI; VBAR FSIZE / MIDPOINTS = 1 TO 8 BY 1;
28
29 PROC MEANS DATA = STORE.SUGI NOPRINT; BY SSIZE REP; VAR FSIZE;
30 OUTPUT OUT = RESULTS MEAN = MSIZE;
31
32 PROC MEANS DATA=RESULTS; BY SSIZE; VAR MSIZE;
33 PROC CHART DATA=RESULTS; BY SSIZE; VBAR MSIZE/MIDPOINTS=2 TO 4.6 BY 0.2;
34
35 PROC SORT DATA=RESULTS; BY SSIZE MSIZE;
36
37 DATA _NULL_ ; SET RESULTS; FILE '\SASWK\CLT'; RETAIN L IC 0;
38     IC = IC + 1;
39     IF L NE SSIZE THEN DO;
40         PUT ' ';
41         PUT ' NEW SAMPLE SIZE CONDITION';
42         PUT ' ';
43         IC = 1; L = SSIZE;
44     END;
45     PUT MSIZE 7.3 @ ;
46     IF IC >= 11 THEN DO;
47         PUT ' ';
48         IC = 0;
49     END;
50
51 RUN;

```

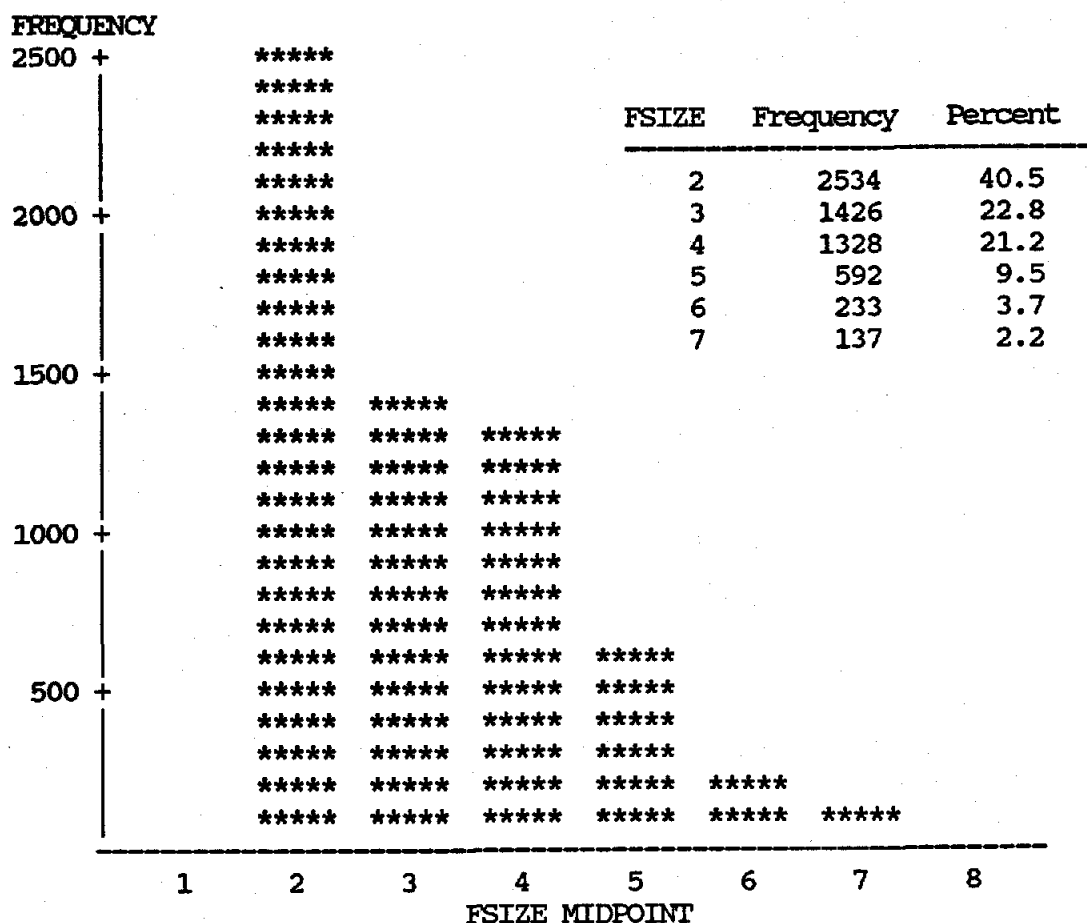
for sampling distributions given samples of differing sizes.) Line 9 sets the number of replications at 250 for each sample size condition. (This models the number of RAs in the scenario.) Line 10 defines samples of size 10 for the first condition and samples of size 15 for the second. "(SSIZE = 2)" is a logical statement. A value of 1 is returned if the statement is true (i.e. SSIZE = 2); or else the value 0 is returned. Multiplying this statement by 5 increases N to 15 for the second SSIZE condition. Line 11 starts the loop to extract the samples. T in line 12 is a temporary continuous random variable with values from 0 to 1 where all points are equally likely. Lines 13 to 18 define the family size variable, FSIZE, to model Figure 1. These are logical statements similar to the one in line 10. Given that all points are equally likely between 0 and 1, a value between 0 and 0.4005 is expected 40.05% of the time. Multiplying this logical statement by 2 should result in the value 2 40.05% of the time. The remaining logical statements in this sequence are deciphered similarly. The OUTPUT statement, line 19, is necessary to force the output to be saved at the given place in the data step. Usually SAS executes each statement in a data step once for each line of input data. At the end of the step the observation is saved in the defined data set. However, since no data are input, SAS would output only the last observation in memory when it completed executing all of the program statements (giving only 1 observation). The END statements in lines 20-22 close loops for each sample (10 or 15 observations), at each replication (RAs), for the two sample sizes, respectively. The CARDS statement, line 23, explicitly ends this DATA step. Note that $250 \times 10 + 250 \times 15 = 6250$ observations will be created. There will be 250 samples of size 10 and another 250 samples of size 15.

The first set of outcomes to be presented to one's students results from lines 25 and 27 of the program. These results are given in Figure 3. The bar diagram and percentages in Figure 3 are similar to the values given in Figure 1 for the population. One can also compare the average value of family size for the population with the corresponding value from the SAS run (see Line 26). For the population this value is $(2 \times 24831 + 3 \times 14427 + 4 \times 13228 + 5 \times 5916 + 6 \times 2276 + 7 \times 1319) / 61997 = 3.19893$. The value from the SAS run (see Figure 3) is 3.196. The instructor can also have students derive standard deviations for both sets of scores. For the population the uncorrected sum of squares is $24831 \times 2 \times 2 + 14427 \times 3 \times 3 + 13228 \times 4 \times 4 + 5916 \times 5 \times 5 + 2276 \times 6 \times 6 + 1319 \times 7 \times 7 = 735282$, the corrected sum of squares is $735282 - 198324 \times 198324 / 61997 = 100857.61$, the variance is 1.62684, and the standard deviation is 1.2754766. The value from the SAS run can similarly be computed by hand (or obtained as a result of the program statement at line 26); 1.28253. (Note that SAS can also be used to compute the population parameters, the hand calculations are included here to illustrate these calculations as possible exercises for students.)

Lines 29 and 30 of the program are the heart of the CLT. It is here that the sample means are computed. "NOPRINT" suspends the printing of results, since it is unnecessary to see the values at this point. "OUTPUT" is used to store the values for subsequent use. "BY" is used so that the mean for each sample (size 10 or 15) will be computed individually. These means are listed in Figure 4. When presenting Figure 4 to students, the instructor should remind them of how the values were obtained (using the scenario given above).

The task of lines 35 to 49 is to produce the sample means in the form shown in Figure 4 (an ordered list, 11 to a line). Line 35 sorts the means within sample size condition. "_NULL_" is used in line 37, because creating a

Figure 3 - Family Size Results from the SAS Run



SAS data set is unnecessary to output the values in the desired form. The ordered sample means are sent to an ascii file entitled CLT in the SASWK subdirectory. IC is a counter that tells when there are 11 means on a line. (Eleven was chosen because with a 7.3 format, this yields a line length of 77 characters.) L detects change in the sample size conditions. The "@" in line 45 holds the line and allows for multiple outputs to the same line (until there are 11 sample means on the line; after which a new line is thrown).

Next, the instructor can give bar diagrams for the means presented in Figure 4. Figure 5 presents these distributions. (The plots are a result of line 33 of the program.) One can note that these distributions are much closer to normality than those for the original data. The instructor can also note the fact that the sample means for the N=15 condition are more compressed than sample means for the N=10 condition. A final point is that both distributions are centered close to the population mean; 3.19893.

Line 32 of the program calculates the mean and standard deviation for the 250 sample means at each sample size condition. When presenting these results the instructor can discuss implications of the sample mean being an unbiased estimator of the population mean. It can be noted (from Figure 4) that some

Figure 4 - The 500 Sample Means

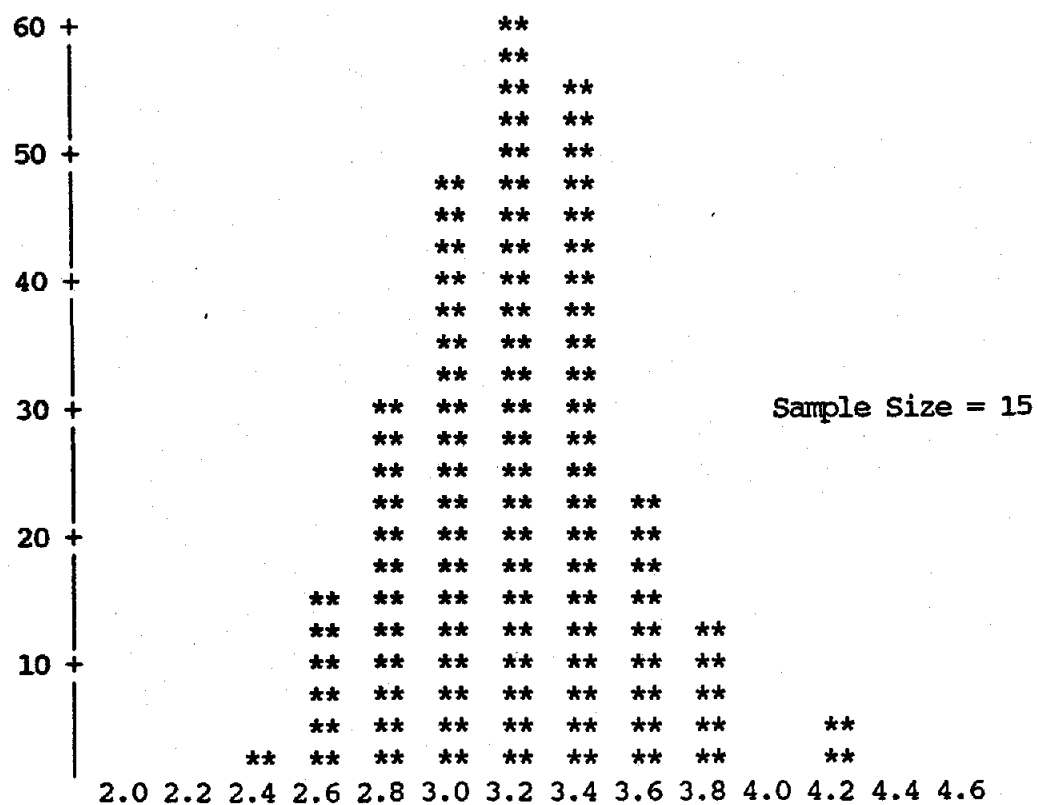
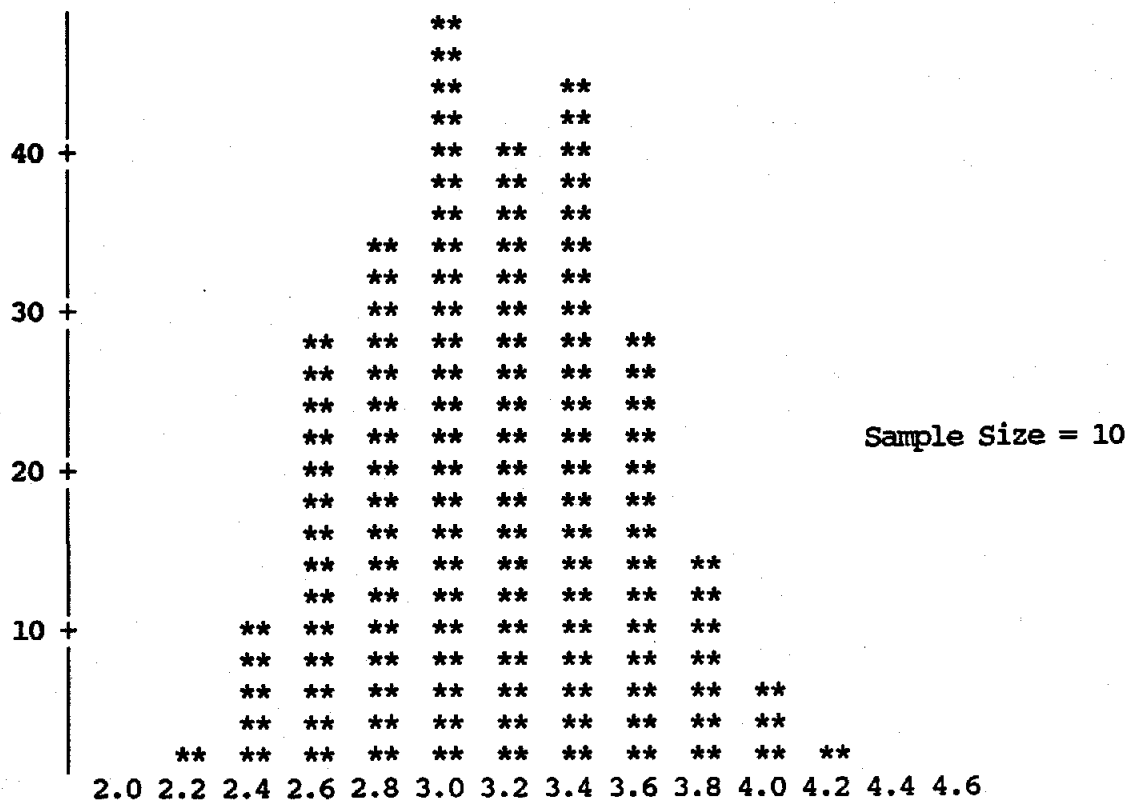
Sampling 10 subjects at a time

[illegible]

Sampling 15 subjects at a time

2.200	2.467	2.467	2.533	2.533	2.533	2.600	2.600	2.600	2.667	2.667
2.667	2.667	2.667	2.667	2.667	2.667	2.733	2.733	2.733	2.733	2.733
2.733	2.733	2.733	2.733	2.733	2.800	2.800	2.800	2.800	2.800	2.800
2.800	2.800	2.800	2.867	2.867	2.867	2.867	2.867	2.867	2.867	2.867
2.867	2.867	2.867	2.933	2.933	2.933	2.933	2.933	2.933	2.933	2.933
2.933	2.933	3.000	3.000	3.000	3.000	3.000	3.000	3.000	3.000	3.000
3.000	3.000	3.000	3.000	3.000	3.000	3.067	3.067	3.067	3.067	3.067
3.067	3.067	3.067	3.067	3.067	3.067	3.067	3.067	3.067	3.067	3.067
3.067	3.067	3.067	3.067	3.067	3.067	3.067	3.133	3.133	3.133	3.133
3.133	3.133	3.133	3.133	3.133	3.133	3.133	3.133	3.133	3.133	3.133
3.133	3.133	3.133	3.200	3.200	3.200	3.200	3.200	3.200	3.200	3.200
3.200	3.200	3.200	3.200	3.200	3.200	3.200	3.200	3.200	3.200	3.200
3.200	3.267	3.267	3.267	3.267	3.267	3.267	3.267	3.267	3.267	3.267
3.267	3.267	3.267	3.267	3.267	3.267	3.267	3.267	3.267	3.267	3.267
3.333	3.333	3.333	3.333	3.333	3.333	3.333	3.333	3.333	3.333	3.333
3.333	3.333	3.333	3.333	3.333	3.333	3.333	3.400	3.400	3.400	3.400
3.400	3.400	3.400	3.400	3.400	3.400	3.400	3.400	3.400	3.400	3.400
3.400	3.400	3.467	3.467	3.467	3.467	3.467	3.467	3.467	3.467	3.467
3.467	3.467	3.467	3.467	3.467	3.467	3.467	3.467	3.467	3.467	3.467
3.467	3.533	3.533	3.533	3.533	3.533	3.533	3.533	3.533	3.533	3.533
3.600	3.600	3.600	3.600	3.600	3.600	3.600	3.600	3.600	3.667	3.667
3.667	3.667	3.733	3.733	3.733	3.733	3.733	3.733	3.800	3.800	3.800
3.867	3.867	3.867	3.867	4.133	4.133	4.200	4.267			

Figure 5 - Bar Diagrams for the Distribution of Sample Means



sample means underestimate the population mean, some overestimate it, and some are essentially equivalent to 3.19893. But, on average they are close; 3.1852 for the N=10 condition and 3.2032 for the N=15 condition. (These results are from the analysis listed on line 32 of the program.) The instructor can also note that the standard error for a sample mean equals the standard deviation of the raw scores divided by the square root of the sample size. For the N=10 case the expected standard deviation of the sample means (standard error of the mean) is $1.2754766 / 3.1622777 = 0.40334$; the comparable value for N=15 is $1.2754766 / 3.8729833 = 0.3293266$. The actual standard deviations for the sample means (from the analysis listed on line 32 of the program) are 0.39665 and 0.33447 respectively for the N=10 and N=15 conditions.

An instructor can also use this data to illustrate the meaning of a "Type 1 error". From theory, 5% of the sample means for the N=10 case should be outside of the bands defined by $3.19893 \pm 1.96 * 0.40334$; 2.40838 and 3.9895. For the N=10 case 11 values (out of the possible 250) are outside of this range: 2.30, 2.30, 2.40, 2.40, 4.00, 4.00, 4.10, 4.10, 4.10, 4.10, and 4.20. Thus, 4.4% of these sample means lie in the 5% rejection region (even though they came from the population of interest). Corresponding values for the N=15 case shows the 95% limits to be $3.19893 \pm 1.96 * 0.3293266$; 2.55345 and 3.8444. For the N=15 condition there are 14 values outside of this range: 2.200, 2.467, 2.467, 2.533, 2.533, 2.533, 3.867, 3.867, 3.867, 3.867, 4.133, 4.133, 4.200, and 4.267. Thus 5.6% of these sample means were unexpected and would lie in the 5% rejection region.

Summary

SAS is a useful tool to create data to illustrate statistical concepts. In addition to providing the data to model techniques and/or "real-world" situations, SAS can also be used to analyze the resultant data. Output from SAS can be easily incorporated into handouts (or short papers) covering the topic of interest. All information provided in the figures in this paper were created by SAS with the exception of the program, Figure 2, and incorporated into the text via cut and paste tactics using a text editor. The use of "real" examples to augment lectures on theoretical aspects of a statistical technique and/or concept should aid students' comprehension by giving them something tangible and meaningful that they can relate to, rather than trying to understand these entities entirely in the abstract.

References

- Houghton Mifflin Company. (1986). Information please almanac: Atlas and yearbook 1986. 39th edition. Boston: Houghton Mifflin Company.
- Minium, E. E. (1978). Statistical reasoning in psychology and education. 2nd edition. New York: John Wiley and Sons.
- SAS Institute Inc. (1985) SAS user's guide: Basics, version 5 edition. Cary, NC: SAS Institute Inc.