

COMPUTER GRAPHICS IN UNDERGRADUATE STATISTICS

by

Edward J. Wegman

Department of Statistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 857

January, 1973

COMPUTER GRAPHICS IN UNDERGRADUATE STATISTICS

by

Edward J. Wegman

University of North Carolina at Chapel Hill

Introduction

Considerable interest has been generated recently in the use of computers for undergraduate statistics curricula; see for example [1], [2], [3] and [4]. The numerical capabilities of the computer in this context are usually well recognized and put to use. However, the potential, in many situations, for use as a graphical display device has been less well recognized. In this paper, we should like to point out some of the potential, discuss briefly two computer-supported graphic display systems and finally report on an experiment recently performed to give some real data on the possible benefits to the student of computer-generated graphic display.

Some Potential Benefits

A successful mathematician or statistician invariably has a well-developed intuition, easily capable of visualizing various mathematical structures including distributions, histograms, converging sequences, random walks, and other such geometrically formulated ideas. The beginning mathematics or statistics student, on the other hand, has no such intuition and must develop it based partly on illustrations in the textbook and those attempted by his instructor at the chalkboard. A cursory examination of most textbooks reveals a relatively small amount of illustrations and,

moreover, of those found, most will lack precision, being the result of an artist's attempt to "clean up" an unsatisfactory drawing submitted by the author. The limited quality and quantity of instructors' chalkboard drawings is undoubtedly widely known and regretted. This situation, then, leaves the student faced with developing his visual intuition in spite of the weaknesses of the illustrations available to him.

The computer has four attributes which allow it to overcome the weaknesses of textbook and chalkboard presentations and, hence, to supplement the introductory statistics course, enabling students to develop their intuition far more rapidly.

1. Precision The computer calculations may be made with sufficient precision so that, for all practical visual purposes, the student has an exact drawing. This is in contrast to the very approximate drawings made at the chalkboard and sometimes found in the textbook.
2. Speed The computer, appropriately programmed, can quickly make a drawing, not only more accurately, but also more speedily than an instructor could at the chalkboard. Thus, an instructor, using a computer can make many accurate illustrations with a computer in the same time as a single illustration at the chalkboard. For example, many of us have seen a normal density with mean = 0 and variance = 1,¹ but how many of us have seen a normal density with mean = 0 and variance = .2 or with mean = 1 and variance = 1? Such an illustration very quickly shows the effect of a shift in parameters for the normal density curve, whereas a simple statement or even mathematical proof of this fact may well leave the elementary statistics student unconvinced.

¹ Actually, there is probably a sizable percentage of Ph.D. statisticians who have *never* seen an accurate drawing of any normal probability density.

3. Adaptability A class of students may display a lack of facility in certain areas. With computer-generated illustrations, it is possible to amplify the presentations in this area of difficulty. If the source of trouble can be pinpointed, additional illustrations may be quickly prepared.

4. Dynamism With a computer, we have the possibility of generating moving illustrations. There are many dynamic conceptions in mathematics and statistics, for example, sequences converging, random walks being carried out, Monte Carlo simulations and numerical methods, to name a few. With a properly programmed computer, one may actually generate these illustrations quickly, possibly even real-time in some circumstances.

It is clear, then, that an appropriately programmed computer integrated into an undergraduate statistics course can assist greatly in developing visual intuition, undoubtedly a great asset in a deeper comprehension of the course material.

Two Graphic Display Systems

The first system we will describe is a software package developed by H.T. Beaujon² for the IBM 360 - IBM 2250 combination. The main feature of the 2250 is a cathode-ray tube (CRT) on which computer generated graphic and alphanumeric information is displayed. The area on the face of the CRT on which images may be displayed is a 12 inch square consisting of discrete points. These points are equally spaced, numbering 2^{10} in both vertical and horizontal directions for a total of 2^{20} points. Other significant

2

Mr. Beaujon is a graduate student in the Department of Computer Science at UNC. His work was supervised by members of both the Computer Science and Statistics departments.

features include an alphanumeric keyboard which can be used to enter messages into the 2250 buffer and onto the CRT display area, a programmed function keyboard and a lightpen detect.

The programmed function keyboard consists of thirty-two pushbutton keys, each of which can be made to correspond to a set of instructions in the computer program. The keyboard has a replaceable overlay which labels each key. The lightpen detect is a penlike device which allows the user to enter information into the system by pointing the device at a lighted portion of the CRT and depressing the associated foot switch. Control is then transferred to a computer program.

The statistical display system developed by Beaujon is an interactive system designed as an aid in teaching the elementary properties of statistical distributions. The system may be used either in a classroom context where the teacher reinforces his material with graphical displays or in a laboratory situation where one or more students, appropriately supervised, are in command of the system and go through a number of prepared exercises. (Several such exercises are included in the appendix.) The system has the following graphical capabilities:

- (1) For a number of standard discrete statistical distributions, one can display either the probability function or the probability distribution function.
- (2) For a number of standard continuous statistical distributions, one can display either the probability density function or the probability distribution function.

(3) From a number of standard statistical distributions a random sample can be drawn and the resulting histograms can be displayed.

(4) As many as 7 of these displays may be exhibited simultaneously.

For (1), (2), and (3), the parameters may be freely varied. In addition, (3) has several interesting capabilities. The program is written so that the statistical distribution may be sampled and the standardized statistic,

$$\frac{\sqrt{n} (\bar{X} - \mu)}{\sigma} ,$$

formed. A Monte Carlo experiment may be performed by repeating this sampling procedure and constructing the appropriate histogram. If the sampling procedure is repeated enough, the resulting histogram should approximate a standard normal density illustrating the Central Limit Theorem.³ For (3), the sample size, the number of samples and the class interval width as well as the distribution parameters may be set by the operator.

The mode of operation of the system is determined by the programmed function keyboard, e.g. the density, distribution or sampling mode as well as the parametric family are selected on this keyboard. The parameters are entered via the alphanumeric keyboard, and a display may be erased in whole or in part via the lightpen or via the programmed function keyboard. A standard set of axes is initialized with program initialization but these may be altered by use of the programmed function and alphanumeric keyboards. Operation of the system is straightforward, almost self-explanatory. Our

³Strictly speaking, the CLT says $\sqrt{n}(\bar{X}-\mu)/\sigma$ converges in law to a standard normal random variable. Convergence of densities requires additional assumptions.

experience with freshman and sophomore statistics students suggests that 10 to 20 minutes is sufficient to begin using the system fruitfully. The system has been operational for about 3 years and its use has been routinely incorporated by the author in several elementary statistics courses at the freshman and sophomore level.

This system capitalizes on the precision, speed and adaptability aspects we mentioned in the previous section. In particular, generation of a picture on the screen takes only a matter of moments. One one evening, the author was able to make 108 photographic transparencies in approximately 90 minutes. In closing the discussion of this system, we shall point out several important usages of the system:

1. The exact illustration of various statistical distributions in both density and distribution format.
2. The illustration of the effect of a change of parameter within a family. For example, a family of T density functions is bounded between Cauchy and standard normal densities.
3. The illustration of limit results such as the Central Limit Theorem, the Binomial-Normal approximation and histogram convergence.

For more detailed usages and a complete operational description, we refer to Beaujon [5].

A second rather exciting graphical display system is that being developed by T.L. Isenhour and Steve Woodward.⁴ Their main hardware consists of a

⁴ Isenhour is an associate professor and Steve Woodward is a systems engineer both with the Department of Chemistry at U.N.C. in Chapel Hill.

stand-alone Raytheon 706 computer, which is normally used for data acquisition within the chemistry department. In addition, they have a Heathkit color television rewired as a 3-gun oscilloscope, a Bolex 16mm motion picture camera and a home-built, electromechanical interface which allows the computer to control both the color oscilloscope and the camera.

The software consists of 6 assembly language subroutines which are accessible in a FORTRAN language driving program. The software-hardware combination allows a square grid of 1024 points per side to be displayed on the screen. Thus, as with the 2250, there are 2^{20} accessible points in the square grid. In addition, there are 3 levels of intensity (as well as off) for each color allowing the programmer to choose 63 different colors. (In a film, unfortunately all of these colors are not distinguishable. Nevertheless, a substantial range of colors is available to the programmer.) The software system allows for both points and lines to be drawn. In addition, it allows entire lines of print to be displayed in either horizontal or vertical directions as well as allowing individual characters to be drawn. Finally, the software allows the programmer to control the camera's shutter and advance.

Using this system an instructor has the potential to write a relatively simple FORTRAN program, and, in a matter of days, to have returned to him a 16mm color motion picture. The system does not operate real time. However, the time required to make a film is not excessive. The filming time depends on the complexity of the program, but with a complicated program, approximately one hour is needed to film a 3 minute sequence. Unfortunately, this system is still in an experimental stage and so not available for routine

usage. Professor Isenhour and Mr. Woodward are now attempting to speed up the process for real-time usage. Nevertheless, the system as it stands is quite a remarkable development. Not the least remarkable aspect of this system is the price: it was developed for a cost of less than \$20,000 exclusive of the Raytheon computer. The actual hardware, of course, cost much less than that. The potential of such a system is almost mind-boggling. An inexpensive, color graphic display terminal with the potential for moving displays and with a *simple* access language (such as FORTRAN or BASIC) cannot help but excite both instructor and student alike.

A Modest Experiment

A controlled experiment was carried out in the fall of 1972 attempting to evaluate the influence, if any, of graphical aids such as we have been describing. To preface our description of the experiment, we would like to observe that large classes tend to make individual student use of terminals impractical. Congestion quickly builds up on only a few terminals and providing many terminals becomes impractically expensive. To reduce the problem, the author has photographed approximately 130 color transparencies of various displays on the 2250 system. While use of slides reduces somewhat the flexibility allowed by the system, the instructor has a good deal of flexibility left and has the added benefit of economy.

In the fall of 1972, this author was assigned to teach a section of introductory statistics at the freshman to sophomore level. This course is designed to fulfill a general mathematics requirement and is one of several options that students may take. Generally speaking, the students taking this

course are *not* the ones who are successful in mathematical disciplines, i.e. most of the clients are majors in the arts, in English, in languages, in social sciences and in nursing, and frequently there is a sense of hostility towards mathematics.

Normally, this instructor integrates the use of slides or the use of computer terminals periodically throughout the semester. However, for this course the students were informed of the author's intention for this study and all of the class agreed to participate. Showing of the slides was delayed until after the second of three tests was given. With results of these two tests in hand, the students were carefully paired off, one group to see slides, the other not to see them. The composite average score for those not seeing the slides was 66.1% and for those seeing the slides it was 66.3%. The groups were balanced also according to sex, so that each had 11 males and 7 females. Students were also queried about their previous computing experience, but since there was no substantial computing experience in the class, no assignment was made for this variable.

Two days prior to the third examination a 2 1/2 hour work session was scheduled. Twenty eight slides illustrating various distributions and histograms were shown to one group of 18 students.⁵ The duration of the slide show was 19 minutes and accompanied by a simple description of the slides being projected. Subsequent to this the remainder of the class was called in and a 45 minute, fairly intensive review session was held,

⁵ We shall henceforth refer to those seeing the slides as GROUP 2; those not seeing the slides, GROUP 1.

followed by a 30 minute question and answer period. The material reviewed included a discussion of the properties of the distributions just shown to GROUP 2 as well as other material relevant to the upcoming test. In order to minimize clues to GROUP 1, we attempted to integrate our discussion carefully. No special emphasis was placed on the properties of distributions, neither, however, were these properties neglected.

A ten question true-false quiz was included in this third examination and a second ten question true-false quiz was included in the final examination about 3 weeks later. (See the appendix for the quizzes). The raw results of these quizzes are included in Table 1.

The data, of course, was collected in a form amenable to a straightforward one-way analysis of variance or, equivalently, for the two-sample situation we have here, to a T-test. Since we had ANOVA routines readily available, these were used in our analysis. The results are summarized below.

- I. FOR THE OVERALL GROUP. There was a significantly higher score on the first quiz in favor of GROUP 2. $F=10.77$ based on 1 and 34 degrees of freedom was significant at .005 level.
- II. FOR THE FEMALES ONLY. Again we had significance in favor of GROUP 2, this time at the .05 level. $F = 5.27$ based on 1 and 12 degrees of freedom.
- III. FOR THE MALES ONLY. Here we had significance in favor of GROUP 2 at the .025 level. $F = 6.72$ based on 1 and 20 degrees of freedom.
- IV. FOR THE 9 PAIRS OF STUDENTS WITH LOWEST SCORES. Here we had significance in favor of GROUP 2 but only at the .1 level. $F = 3.31$ based on 1 and 16 degrees of freedom.
- V. FOR THE 9 PAIRS OF STUDENTS WITH HIGHEST SCORES. Again significance in favor of GROUP 2 at the .005 level. $F = 10.74$ based on 1 and 16 degrees of freedom.

VI. FOR THE OVERALL GROUP ON THE SECOND QUIZ. A slightly better average for GROUP 2 but no significance. $F = .51$ based on 1 and 30 degrees of freedom.

TABLE 1

GROUP 1 - NO SLIDES					GROUP 2 - SLIDES				
Student No.	Score 1st Quiz*	Score ⁶ 2nd Quiz*	Sex	Avg. Score 1st 2 Exams**	Student No.	Score 1st Quiz*	Score ⁶ 2nd Quiz*	Sex	Avg. Score 1st 2 Exams**
1	7	8	M	66	19	7	6	M	66.5
2	6	7	M	76.5	20	8	4	M	76
3	6	7	M	55.5	21	7	7	M	58
4	5	7	F	53	22	7	4	F	47
5	7	7	M	60	23	8	8	M	59
6	7	5	M	49	24	8	10	M	61.5
7	4	4	M	66	25	7	7	M	71
8	6		F	81.5	26	9		F	81.5
9	6	8	F	71	27	10	8	F	73
10	7	6	M	67.5	28	5	6	F	67.5
11	5	8	M	74	29	7	9	M	72.5
12	6	8	F	42	30	7	8	M	28.5
13	8	7	M	80	31	8	7	M	80.5
14	7	9	M	78	32	7	7	M	80.5
15	6	7	F	54	33	6	10	M	51
16	7	6	F	75	34	7		F	79
17	8	6	F	57.5	35	9		F	61.5
18	6	7	M	83.5	36	9	8	F	79.5

* - out of 10 .

** - out of 100

⁶ There was a slight attrition of students on the 2nd quiz, the reason being that taking the 2nd quiz was made optional.

To state our conclusions and comments in summary:

1. From the proceeding results I and VI, it seems clear that even a very modest amount of supplementary graphical material of the type made available by the graphic display packages provides a significantly increased score for a short term. The knowledge obtained from a "one-shot" presentation decays over a longer period, which no doubt we could reasonably expect. It is fair to assume that a continued program integrating the use of slides or computer graphics would show a significant difference between the groups over a longer term.
2. From results II and III, we may conclude that males appear to respond somewhat better than females, although both male and female students respond well.
3. From results IV and V, it appears as though better students respond more favorably than poorer students. This difference is fairly marked and strikes the author as an interesting and perhaps unexpected result.
4. The experiment just related is open to criticism on two counts. First is that if GROUP 1 had received 20 additional minutes of instruction, perhaps their score would have been significantly improved. The second is that performance on a 10 question true-false quiz may well be biased by the nature of its construction in favor of those having seen the slides. The answer to both of these objections is to reiterate the experimental conditions. Those students in GROUP 2 saw 28 slides in 19 minutes. We feel this was a negligible amount of time compared to the some 42 hours all students had been exposed to up to this time, including an intensive review. Solutions to all the questions

asked in both quizzes had been discussed with *all* students in the review session. No mention of these solutions was made during the narrative of the slides. Hence it is the feeling of this author that the difference in average scores can be attributed *substantially* to the viewing of exact computer generated illustrations and that these other two effects contributed only a *negligible* amount to this difference. This, of course, is a subjective judgment and other studies eliminating these and other undesirable characteristics are to be encouraged.

5. One final anecdote: subsequent to the first quiz, the author was approached by a student in GROUP 2. The student's comment was that he did not see any benefit to having seen the slides; this in spite of the fact that his score was 8 - much better than the average score of GROUP 1. This suggests that the development of this visual intuition we have earlier mentioned is a subtle phenomenon whose existence is not necessarily clear to the students.

Acknowledgement

The author is indeed grateful to the members of the Statistics 11 class, Fall, 1972. Their fine cooperation has provided some interesting and useful data and their participation is deservedly recognized and commended.

References

- [1] G.P. Weeg (ed.), *Proceedings of a Conference on Computers in Undergraduate Curricula*, University of Iowa, 1970.
- [2] T.E. Kurtz (ed.), *Proceedings of the 1971 Conference on Computers in Undergraduate Curricula*, Dartmouth College, Hanover, N.H., 1971.
- [3] J.W. Hamblen (ed.), *Proceedings of the 1972 Conference on Computers in Undergraduate Curricula*, Southern Regional Education Board, Atlanta, Georgia, 1972.
- [4] E.J. Wegman and B.H. Gere, "Some thoughts on computers and introductory statistics," *Int. J. Math. Educ. Sci. Technol.*, 3, pp. 211-221 (1972).
- [5] H.J. Beaujon, *An Interactive Graphical Display System for Illustrating Elementary Properties of Statistical Distributions*, University of North Carolina at Chapel Hill, 1970.

Appendices

Exercise 1

<u>Action</u>	<u>Comment</u>
1. RESET	pfk
2. GRID	pfk
3. Data: 0,1	enter distribution parameters.
4. NORMAL	pfk
5. Data: 1,1,	enter distribution parameters.
6. NORMAL	pfk; note that the change in the mean of the normal distribution corresponds to a <u>shift</u> .
7. Data: 1,2,	enter distribution parameters.
8. NORMAL	pfk; a change in the variance corresponds to a change in the <u>spread</u> of the curve.
9. REML	pfk; remove the last distribution plotted.
10. Y.2	pfk; double the Y grid parameters.
11. GRID	pfk; new grid.
12. REDRAW	pfk; redraw everything on the new scales.
13. Data: 0,.2,	enter distribution parameters.
14. NORMAL	pfk; a variance less than 1 makes the curve skinnier (and taller!) than $N(0,1)$.
15. Data:-2,.2,	enter distribution parameters
16. NORMAL	pfk.

* * *

Exercise 3

Action	Comment
1. RESET	pfk
2. xlo = 0 xhi = 20 xinc = 20 ylo = 0 yhi = .4 yinc = .4	grid parameters
3. GRID	pfk
4. Data: 4,.5,	enter distribution parameters.
5. BINOM	pfk; binomial distribution.
6. Data: 5,.5,	enter distribution parameters.
7. BINOM	pfk
8. REM NL	pfk; remove the next to last distribution.
9. Data: 6,.5,	enter distribution parameters
10. BINOM	pfk
11. REM NL	pfk; remove the next to last distribution.
12. Data: 12,.5,	enter distribution parameters.
13. BINOM	pfk
14. REM NL	pfk; remove the next to last distribution.
15. Data: 16,.5,	enter distribution parameters.
16. BINOM	pfk
17. REM NL	pfk; remove the next to last distribution.

The discrete points now look pretty much like a normal distribution. Which normal distribution? Try finding it by trial and error. Check your results by calculation.

Introduce new grid parameters and see what the correspondence is for the respective cumulative distribution functions.

Quiz 1

True or False.

- a. Two normal distributions with the same variance but different means have the same shape but different locations.
- b. A normal distribution function has no jumps.
- c. The binomial probability function is symmetric when $p = .5$, but it need not always be.
- d. The Poisson is never symmetric.
- e. The binomial probability function with $n = 16$ and $p = .5$ is bell-shaped substantiating the binomial-normal approximation.
- f. As n increases in a binomial family, the mean and variance of the family both increase.
- g. We say a discrete family is bimodal if the probability function has two spikes of equal height which are taller than any others. The Poisson family is bimodal.
- h. A histogram based on a sample size of 10 is rotten.
- i. As m increases in a Poisson family, the mean increases.
- j. As m increases in a Poisson family, the maximum height of the probability function decreases.

Quiz 2

True or False

- a. The theoretical mean of a normal distribution determines both shape and location.
- b. A normal random variable is continuous because its distribution function has only jumps and elsewhere is flat.
- c. A binomial probability function is symmetric when $p = .5$ and $n = 13$.
- d. A Poisson probability function is symmetric when m is bigger than 25.
- e. For n large and p close to $.5$, the binomial probability function is approximately bell-shaped.
- f. As p increases in the binomial family, the mean increases.
- g. The Poisson family is not bimodal always.
- h. A histogram based on a sample of size 100 is always better than one based on a sample of size 90.
- i. As m increases in a Poisson family, the variance increases.
- j. As n increases in a binomial family, the maximum height of the probability function decreases.