Exploring probability and statistics using computer graphics

Tijms, H.C.

1990

document version
Publisher's PDF, also known as Version of record

Faculteit der Economische Wetenschappen en Econometrie

Serie Research Memoranda

Exploring Probability and Statistics Using Computer Graphics

Henk C. Tijms

Research Memorandum 1990 - 45

September 1990

vrije Universiteit *amsterdam*

EXPLORING PROBABILITY AND STATISTICS USING COMPUTER GRAPHICS *)

by
Henk Tijms
Dept. of Econometrics, Vrije University

1. Introduction

In the teaching of probability and statistics microcomputers can do more than take the drudgery out of statistical computing:

- they can be used as well to teach basic concepts and ideas
- they make it possible to perform laboratory experiments
- they enable students to discover basic principles themselves.

In particular, computer graphics are very effective to gain an understanding of basic concepts having a pictorial representation. Many of the basic ideas in probability and statistics seem exceedingly difficult for most students to grasp. It is extremely important to give students a sound intuitive feeling for basic concepts such as randomness and the normal curve before teaching them formal statistical theory. For example, key concepts such as the law of large numbers, random walks, and the central limit theorem can be made to come alive before one's eyes through computer graphics. Direct experience and actual experimentation is the best way the student can obtain a feeling for these basic concepts. It cannot be emphasized enough how important it is that students obtain at an early stage a feeling for probabilistic reasoning. People are not born with a natural feeling for probabilistic thinking, while probability reasoning is essential for solving many real-world problems. The micro is the ideal tool for students to develop a sound probabilistic intuition. Computer graphics - the most powerful feature of the micro - should be exploited. Most people find it much easier to grasp

*) Invited paper presented at the Third International Conference on the Teaching of Statistics, University of Dunedin, New Zealand, August 1990.

concepts in visual terms and through direct experiences than in terms of formulas. This paper discusses how we use graphical software to introduce the beginning student in both a motivating and coherent way to the following very basic concepts:

- *The law of large numbers.* Computer animations of the experiment of coin and dice tossing is the best way to obtain insight in random fluctuations. Using an interactive simulation program it is possible to fight misconceptions that even short runs of the coin-tossing experiment should reflect the theoretical 50: 50 ratio of heads to tails.
- *Random walks.* The graphical display of random walks provides the student with a lot of insight into random fluctuations. The random walk showing the actual number of heads minus the expected number in the coin-tossing experiment is very instructive. Using this random walk, a natural link can be made with the central limit theorem.
- *The central limit theorem.* A visual demonstration of this very important concept in probability and statistics can be given. In an interactive way the student can generate the probability histograms for different sample sizes and see how fast the probability histogram gets close to the normal curve as the sample size gets larger.
- *Statistical graphs.* Plotting the density graphs of various distributions such as the binomial, Poissons and chi-square enables the student to discover limiting relations between these distributions and the normal distribution.

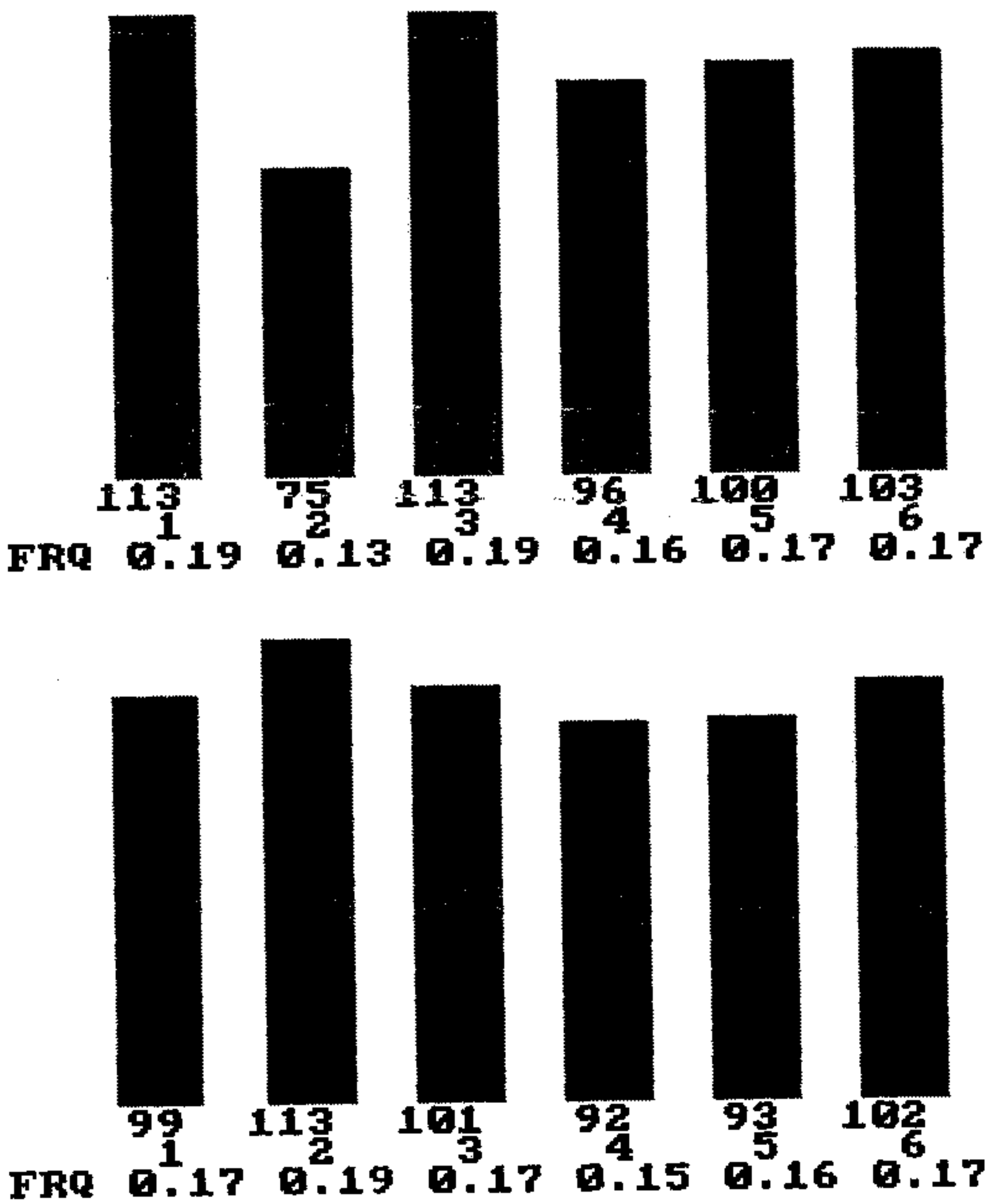
2. Law of Large Numbers

Many people tend to expect that even short runs of the coin-tossing experiment should reflect the theoretical 50:50 ratio of heads to tails. This misconception is known as the gamblers fallacy since some gamblers feel that the probability of tails becomes larger after a run of heads. Mere exposure to the theoretical laws of probability may not be sufficient to overcome such misconceptions. However, an interactive simulation program on the microcomputer is ideally suited to fight misconceptions such as the gamblers fallacy.

Computer animation of the experiment of coin and dice tossing is the best way to obtain a feeling for randomness. Within a few seconds students can simulate and repeat this statistical experiment on the micro. By looking at the graphical representation of the results, they see what randomness means. By doing the experiment of tossing a fair coin and observing the relative frequency at which heads appears, the student will see that this relative frequency may still significantly differ from the value 1/2 after a large number of tosses. As the number of tosses grows, the relative frequency will eventually approach the value 1/2 according to the law of large numbers. To see before one's eyes this experimental demonstration of the law of large numbers is very instructive. Also, it is instructive to see that the relative frequency typically approaches the value of 1/2 in a rather irregular way. To make the above concepts alive, we developed in Kalvelagen and Tijms (1990) a software module that simulates the experiment of rolling a die. The program gives the user the option of using either a fair die or a loaded die. For the case of a loaded die, the user has to specify the probabilities of each of the six possible outcomes of any given throw of the die. By assigning positive probabilities to only two outcomes, the coin-tossing experiment is a special case of the die-tossing experiment. Once the data have been specified, a computer animation of the die-tossing experiment is given in real time. In Figure 1 the final results are displayed of two computer simulations each consisting of 600 throws of a fair die. Such pictures give the student a good feeling for random fluctuations.

The software for the die-tossing experiment is elementary, but it is very instructive. Students like to do the experiment themselves and in this way they learn what randomness means. Much can sometimes achieved by simple means!

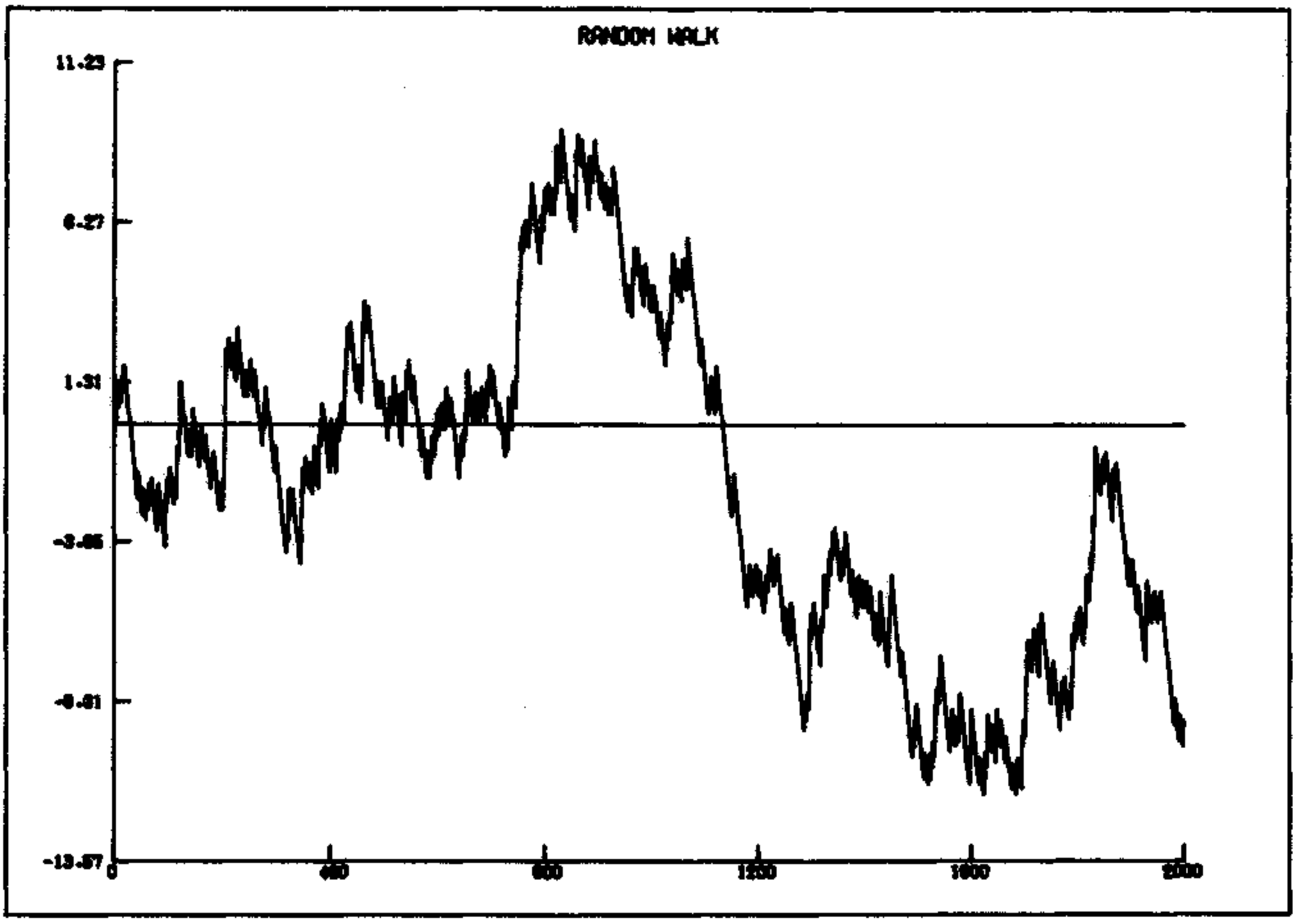
Figure 1 Two simulations of the die-tossing experiment



3. Random Walks

Let's consider the coin-tossing experiment using a fair coin. Many people erroneously believe that a run of heads should be followed by a run of tails so that heads and tails even out. In placing their bets many gamblers use some system that take into account any imbalance between the past number of heads and tails. It is illusionary to think that such a system can be of any help. The coin has no memory. Suppose a fair coin has been tossed 100 times and has landed heads 60 times. Then in the next 100 tosses the absolute difference between the number of heads and tails may still increase, while the relative difference decreases. For example, this occurs when heads appears 51 times in the next 100 tosses. The only thing one can be sure of is that the relative frequencies of heads and tails will eventually be equal, but there is not something as a law of averages for the absolute difference between heads and tails. In fact, the absolute difference between the number of heads and tails tends to increase as the number of tosses gets larger. This surprising fact can convincingly be shown by simulation and a graphical display of the results on the screen. For a simulation experiment consisting of 2,000 tosses of a fair coin, Figure 2 displays the graph of the random walk giving the actual number of heads minus the expected number. By trying themselves different simulation runs of various lengths on the micro, the students will learn by experience that realisations such as in Figure 2 are no exception, but are typical for the coin-tossing experiment. The longer the run, the bigger and bigger the waves of the random walk and the relatively more rare the crossings of the x-axis. This finding is rather counterintuitive, but can be mathematically explained from the central limit theorem. However, before turning to this basic theorem, it is important that the student first learns by direct experience that a series of independent probability trials tends to behave better and better in the average sense but wilder and wilder in the absolute sense as the number of trials increases. This is a basic lesson in probability!

Figure 2 **A random walk for the coin-tossing experiment**



Once the students have learned by experimentation on the micro that the random walk of the actual number of heads minus the expected number tends to exhibit bigger and bigger fluctuations as the number of tosses grows, they may be ready to learn more about the mathematical explanation of this behavior. At this point a natural link can be made with the central limit theorem. Taking for the moment the central limit theorem for granted, how do we explain mathematically the behavior exhibited in Figure 2? Defining the random variable $X_i=1$ if the i^{th} toss gives heads and $X_i=0$ otherwise, it follows that the actual number of heads minus the expected number can be represented as $X_1+\ldots+X_n-n\mu$, where $\mu=1/2$ denotes the mean of the X_i's. Since the normal distribution has about 68% of its mass within one standard deviation of the mean, a simple consequence of the central limit theorem is that

$$P\{|X_1 + \ldots + X_n - n\mu| > \sigma\sqrt{n}\} \approx 0.32 \qquad \text{for n large enough,}$$

where $\sigma=1/2$ denotes the standard deviation of the X_i's. In words, the probability that after n tosses the random walk $X_1 + \ldots + X_n-n\mu$ will take on a value larger than $\sigma\sqrt{n}$ is about 32% when n is large. This gives the mathematical explanation that the absolute difference between the number of heads and tails tends to increase as the number of tosses grows. This result is not contradictory to the fact that the difference between the relative frequencies of heads and tails tends to zero, since the absolute difference between the number of heads and tails roughly tends to increase proportionally to the square root of the number of tosses. This important but subtle point is best understood by students through experimental studies using computer graphics.

4. Central Limit Theorem

The central limit theorem is the most important result in probability and statistics. Every beginning student should know about this theorem. Using the random walk discussed in the previous section the student's interest could be aroused in this theorem, but how to explain it? The central limit theorem is extremely difficult to prove. Moreover, the proof will not substantially help the student to understand the working of the theorem. It will not give the student a clear insight into how large n should be before the sum $X_1+\ldots+X_n$ of n independent and identically distributed random variables $X_1,\ldots,X_n$ is approximately normally distributed. An intuitive feeling for the central limit theorem is best obtained through experimental studies based on computer graphics.

A possible way to demonstrate graphically the central limit theorem is to use computer simulation. In an interactive way many samples may be drawn from a given probability distribution and a graphical display of the probability histogram of the sum may be given for different sizes. It will then be seen that the probability histogram gets close to the bell-shaped form of the normal density as the sample size increases. The drawback of the simulation approach is that for any fixed sample size n many observations of the sum are needed before the simulated probability density of $X+...+X_n$ is sufficient close to the true density. How many observations are needed is often not clear. Hence the simulation approach has the drawback that the law of large numbers interferes with the central limit theorem. To avoid this complication which obscures the working of the central limit theorem, we advocate an analytical approach in combination with computer graphics. Restricting to discrete probability distributions allows to compute analytically the true probability density of the sum $X_1 + ... + X_n$ for any fixed value of n. How to do this can be easily explained to the beginning student, cf. Kalvelagen and Tijms (1990). An illustrative case is obtained by using the good old die and taking

X_k = the number of points shown at the k^{th} rolling of a die.

The sum $X_1+...+X_n$ then represents the total number of points obtained when the die is rolled n times.

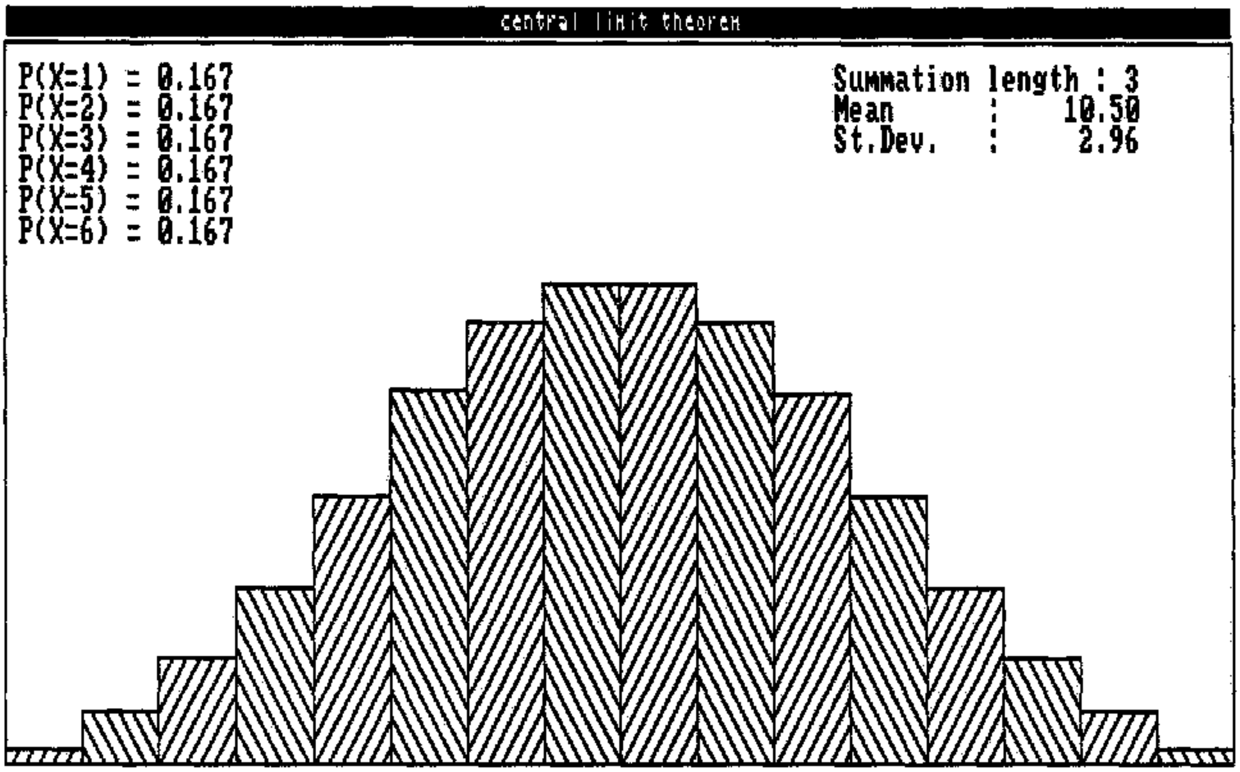
The software module we developed in Kalvelagen and Tijms (1990) gives the student the option of using a fair die or a loaded die. For a loaded die, the student has to specify for $j=1,...,6$ the probability $p(j)$ of getting j points at any given throw of the die. For example, by choosing $p(1)=p$ and $p(2)=1-p$ for some $0<p<1$, the student has the option to verify experimentally that the binomial probability density tends to the normal curve as the number of trials gets large.

Once the underlying probability distribution {p(j)} of the die and the number n of throws have been specified, the computer program calculates the probability density of the sum $X_1+...+X_n$ and displays the graph of this density on the screen. A glance at the screen is sufficient to see whether this density has the bell-shaped form of the normal curve or not. Since the probabilities p(j) of the die can be varied, students can discover themselves that how large n should be before the probability density of the sum $X_1+...+X_n$ is close to the normal curve depends very much on the degree of asymmetry in the underlying probability distribution of the die. It is quite instructive to find out that the more the underlying distribution is asymmetric, the larger n should be. As an illustration, Figure 3 gives the probability densities of the sum $X_1+...+X_n$ for n=3 and n=5 tosses of a fair die. It is removable how fast the probability density of the sum resembles the normal curve when the underlying distribution is symmetric. However, the situation is quite different for an asymmetric distribution. Figure 4 displays the probability densities of the sum $X_1+...+X_n$ for n=5 and n=20 tosses of a loaded die having the asymmetric distribution p(1)=0.2, p(2)=0.1, p(3)=p(4)=0, p(5)=0.3 and p(6)=0.4. It is really fun to learn about the central limit theorem in this experimental way.

5. Statistical Graphs

The normal curve in probability theory is a law of nature. Many probability distributions are closely related to the normal distribution. Students can learn this fact not only from the central limit theorem, but as well by plotting the density graphs of various distributions.

Figure 3 Probability histograms for a fair die



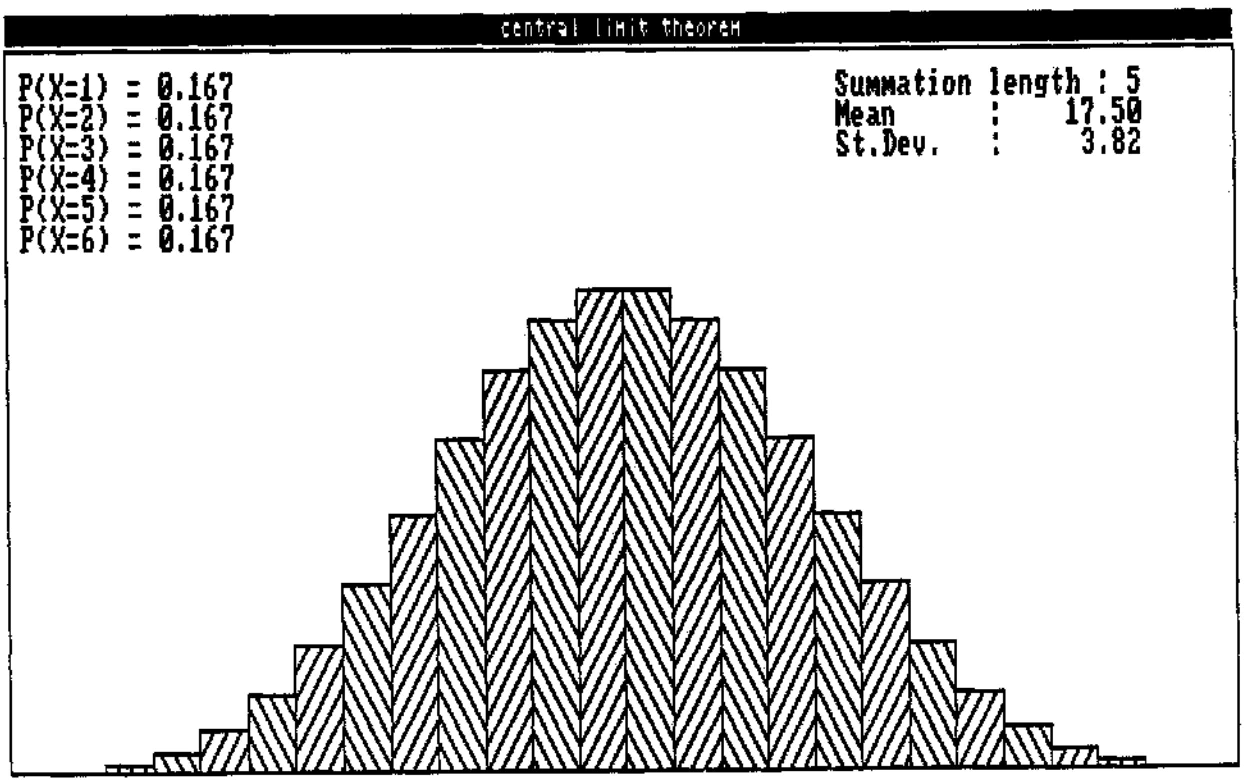
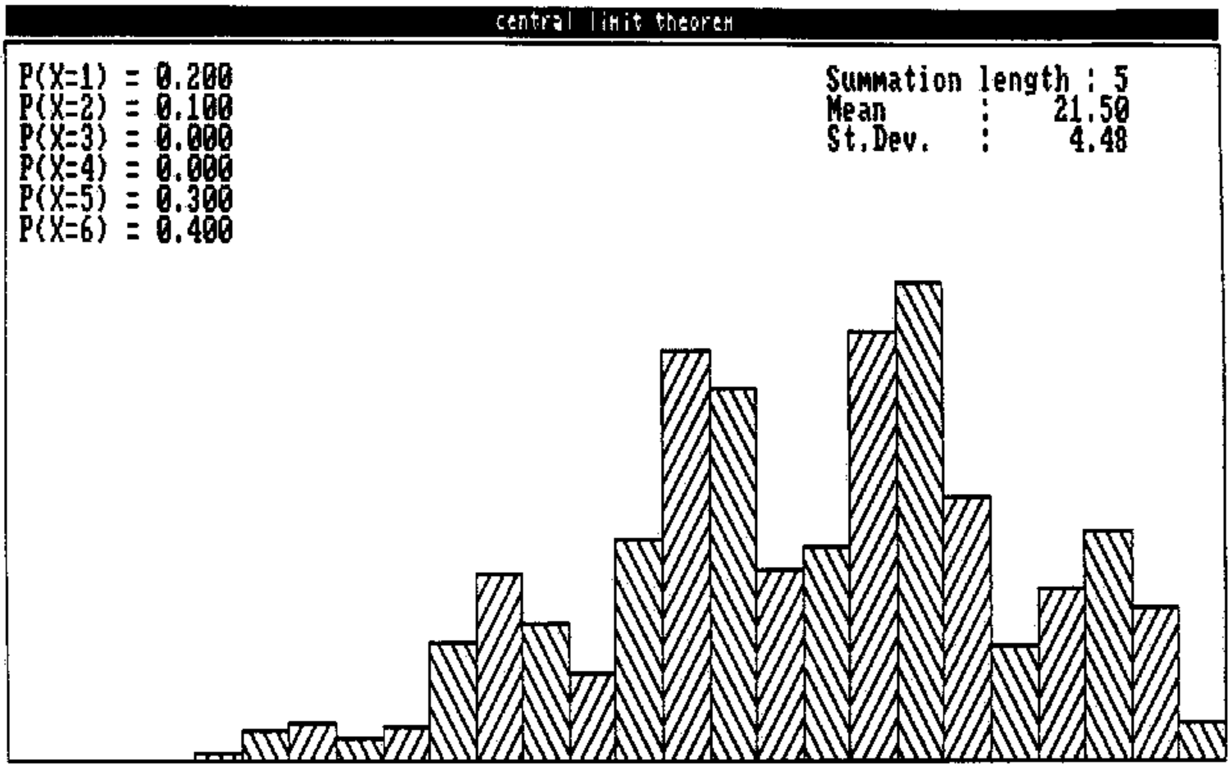
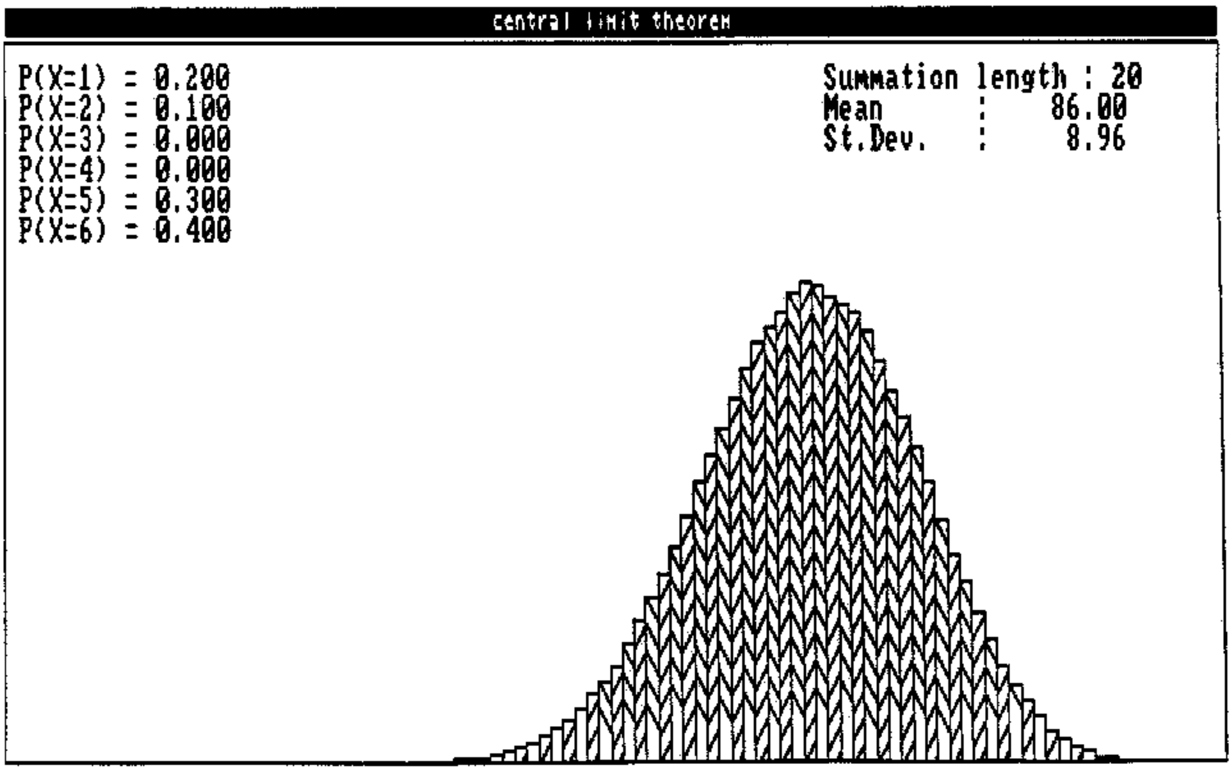


Figure 4	**Probability histogram for a loaded die**



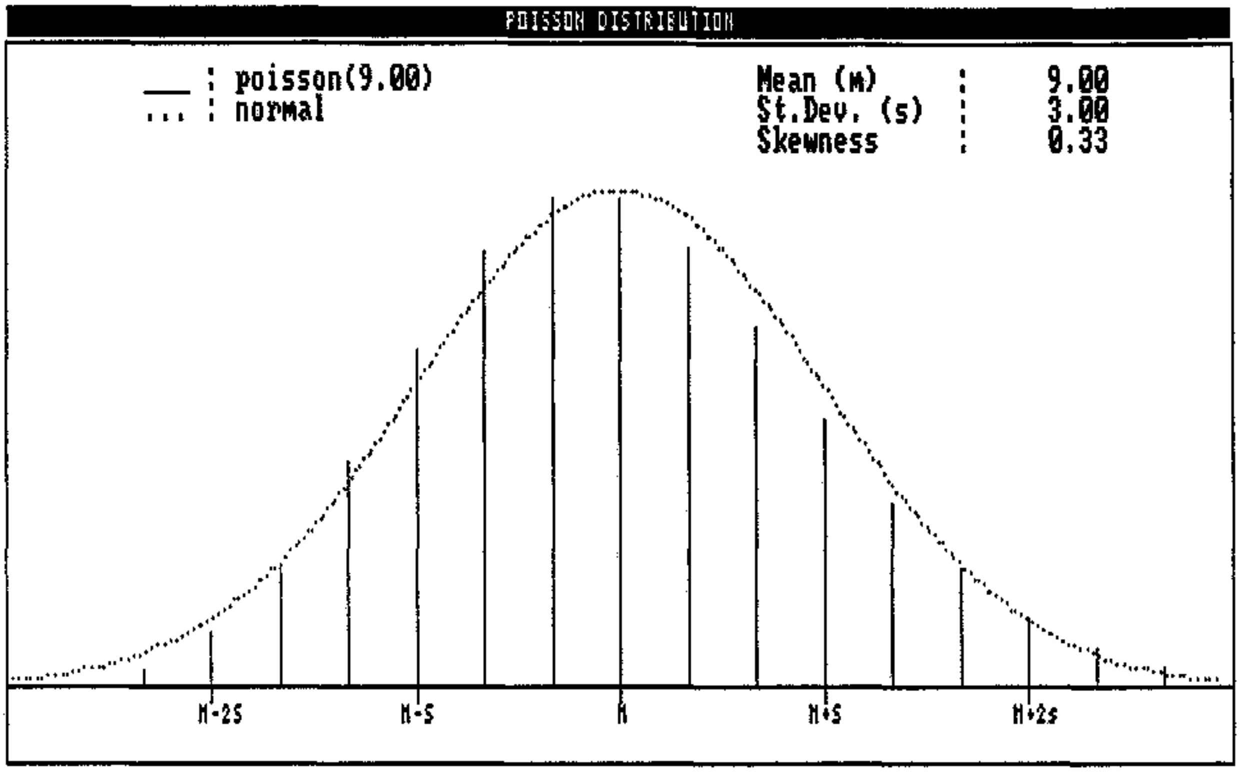


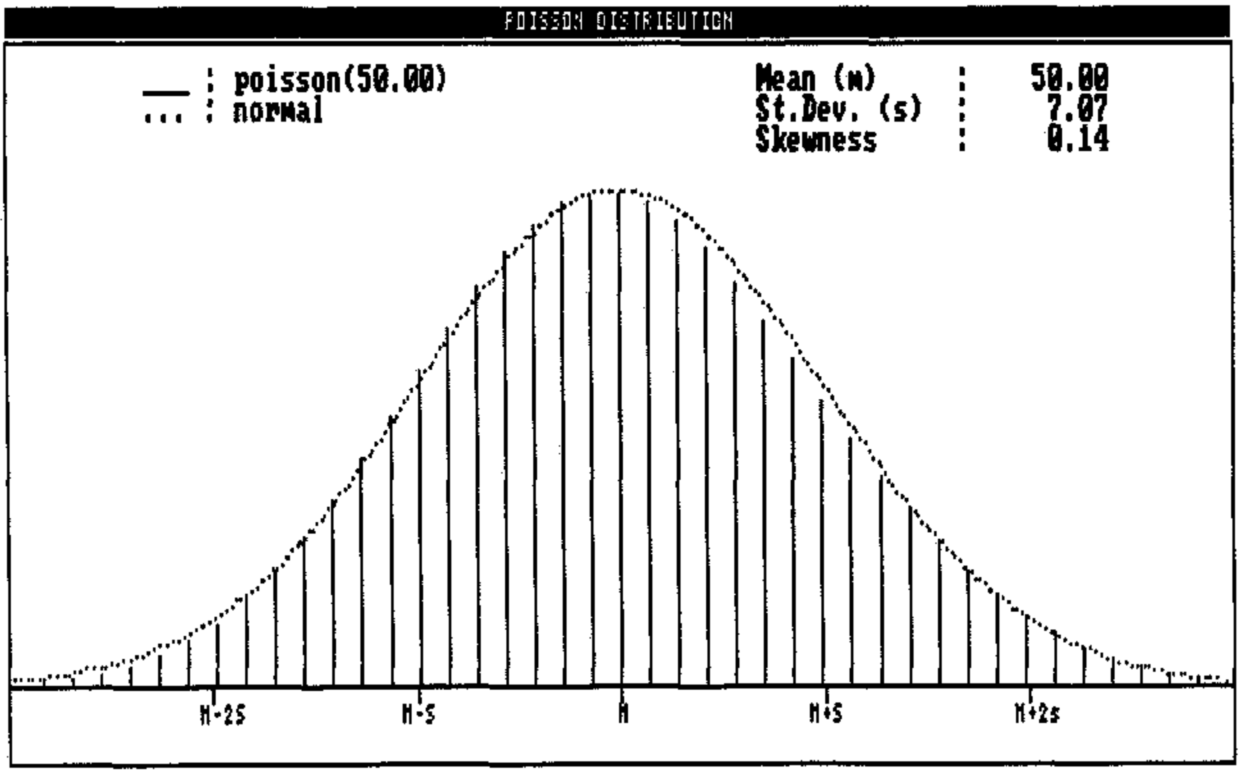
The binomial distribution is the most important family of discrete probability distributions. A binomial random variable with parameters n and p can be interpreted as the total number of successes in n independent trials with probability p of success on each trial. This interpretation enables to explain from the central limit theorem the fact that the binomial density graph approaches the normal curve as n increases. Alternatively, students can discover this limiting relation by plotting the binomial density graphs for various values of n and together with the corresponding graphs of the normal density. At the same time they can experimentally verify that the normal approximation to the binomial density is quite good provided that the binomial mean is at least three (say) deviations away from both 0 and n, the extreme values of the binomial random variable. Assuming that students know that more than 99% of the area under the normal curve is within three standard deviations of the mean, this condition will be intuitively obvious to them; otherwise, there is enough distance between the mean and the extreme values to be able to adopt the shape of the normal curve. The approach of plotting the probability density graphs is particularly useful to discover a limiting relation between the Poisson and the normal distribution. A Poisson random variable X has a single parameter: $P\{X=k\} = e^{-\lambda} \lambda^{k}/k!$, $k=0,1,\ldots$. The mean μ and the standard deviation of the Poisson distribution are given by $\mu=\lambda$ and $\sigma=\sqrt{\lambda}$. Students may argue that approximate normality for a Poisson distribution requires that the mean $\mu=\lambda$ is at least three deviations $\sigma=\sqrt{\lambda}$ away from 0. The condition $\lambda=3\sqrt{\lambda}$ gives $\lambda\geq 9$. Using graphical plots, students can directly verify the validity of this condition. Figure 5 displays the plots for both $\lambda=9$ and $\lambda=50$. A lot of insights is obtained by experimenting with graphical software that displays the density graphs of various distributions together with the corresponding density graphs of the normal distribution. It is very rewarding to learn in this way about the key role of the normal distribution in probability and statistics.

References

[1] E. Kalvelagen and H.C. Tijms (1990), Exploring Operations Research and Statistics in the Micro Lab, Prentice-Hall, Englewood Cliffs, New Jersey.

Figure 5 Poisson densities and the normal curve





HT909AT1

1990-1	B. Vogelvang	Testing For Co-Integration with Spot Prices of Some Related Agricultural Commodities
1990-2	J.C.J.M. van den Bergh P. Nijkamp	Ecologically Sustainable Economic Development Concepts and Model Implications
1990-3	J.C.J.M. van den Bergh P. Nijkamp	Ecologically Sustainable Economic Development in a Regional System: A Case Study in Agri cultural Development Planning in the Netherlands
1990-4	C.Gorter P.Nijkamp P.Rietveld	Employers' Recruitment Behaviour and Re-Employment Probabilities of Unemployed
1990-5	K.Burger	Off-farm income and the farm-household the case of Kenyan smallholders
1990-6	H. Visser	Crowding out and the Government Budget
1990-7	P. Rietveld	Ordinal Data in Multicriteria Decision Making, a Stochastic Dominance Approach to Siting Nuclear Power Plants
1990-8	G. van der Laan P.H.M. Ruys D.J.J. Talman	Signaling devices for the supply of semi-public goods
1990-9	F.A.G. den Butter	Labour Productivity Slowdown and Technical Progress: An empirical analysis for The Netherlands
1990-10	R.W. van Zijp	Neo-Austrian Business Cycle Theory
1990-11	J.C. van Ours	Matching Unemployment and Vacancies: The Efficiency of the Dutch Labour Market
1990-12	B. Vogelvang	Hypotheses Testing Concerning Relationships between Spot Prices of Various Types of Coffee
1990-13	A.F. de Vos I.J. Steyn	Stochastic Nonlinearity: A Firm Basis for the Flexible Functional Form
1990-14	Y.H. van Emmerik D. de Jong W.W.A. Zuurmond D.N. Dukkers-van Emden	Opereren in overleg: geprotocolleerde samenwerking 1e-2e-lijn bij dagchirurgie
1990-15	T.J.J.B. Wolters	Mediation and Collective Bargaining: A Diagnostic Approach
1990-16	E.M.A. Scholten J. Koelewijn	Financieringsproblematiek van startende ondernemingen: een mogelijke verklaring op basis van empirisch onderzoek.
1990-17	E. Hüner H.P. Smit	Saturation and Model Specification of Passen ger car Ownership
1990-18	F.A.G. den Butter	Sociale zekerheid, de wig en economische groei