8-1970

An Interesting Relation Between the Binomial and Normal Functions

George Lee Love
Prairie View A&M College

AN INTERESTING RELATION BETWEEN THE BINOMIAL AND NORMAL FUNCTIONS

PRAIRIE VIEW AGRICULTURAL AND MECHANICAL COLLEGE
GRADUATE SCHOOL

WORKSHOP SHEET III & IV
THESIS (OR ESSAY) REPORT
*TURN IN THIS FORM WITH YOUR COMPLETED THESIS OR ESSAY

NAME Love, George
DEGREE Master of Science

[redacted] Odessa, Tex.
(PERMANENT HOME ADDRESS)
DEPARTMENT Mathematics

DATE SUBMITTED August 11, 1970

TITLE OF THESIS OR ESSAY: An Interesting Relation Between the Binomial and Normal Functions

PREVIOUS DEGREES: B.S. (DEGREE) Prairie View (COLLEGE)
May, 1967 (DATE)

(DEGREE) (COLLEGE)

(DATE)

RECORD:
NUMBER OF PAGES: 35
NUMBER OF TABLES OR CHARTS 1
NAME OF TYPIST: George Love

UNDERGRADUATE MAJOR Mathematics
UNDERGRADUATE MINOR Chemistry
GRADUATE MAJOR Mathematics
GRADUATE MINOR Secondary Education

APPROVAL: ______________________
(SIGNATURE OF SUPERVISING PROFESSOR)

BRIEF SUMMARY OF THESIS (OR ESSAY)
(NOT TO EXCEED 100 WORDS)

(THIS SUMMARY IS A PERMANENT BIBLIOGRAPHICAL RECORD. IT SHOULD BE WRITTEN CAREFULLY).

The primary concern in this paper is showing that the binomial function has as its limit, as n gets large, the standard normal function. The proof of the main theorem is done using two approaches. The first being one of a geometry or slope approach and the second approach is that of area.

PRAIRIE VIEW A & M COLLEGE

PRAIRIE VIEW, TEXAS

Final Examination

of

George Lee Love

Tuesday, August 11, 1970

Time: 3:00 P.M. Room: 106 OS

Committee in Charge

Dr. A. D. Stewart	Chairman Professor of Math.
Mrs. Evelyn Thornton	Graduate Advisor Asst. Professor of Mat
Mr. Clyde Christopher	Asst. Prof. of Math.
Dr. H. G. Hendricks	Department of Educatio
Mr. Waymon Webster	Department of Educatio

An Interesting Relation Between the Binomial and Normal Functions

By

George Lee Love

A Thesis

Submitted to the Department of Mathematics in Partial

Fulfillment For the Requirements For the Degree of

Master of Science

in the

Graduate Division

Prairie View Agricultural and Mechanical College

Prairie View, Texas

August, 1970

The Thesis

For the Degree of Master of Science

by

George Lee Love

has been approved for the

Department of Mathematics

by

[redacted]

Head of Department

[redacted]

Advisor

[redacted]

Committee Member

[redacted]

Committee Member

[redacted]

Committee Member

[redacted]

Committee Member

Date

ACKNOWLEDGEMENT

The writer wishes to acknowledge sincere thanks to Mrs. Evelyn Thornton, Mr. Clyde Christopher, Department of Mathematics, and Mr. Waymon Webster, Dr. H. G. Hendricks, Department of Education, instructors at Prairie View A & M College, Prairie View, Texas for their personal interest and guidance for the preparation of this paper. The writer is in deepest gratitude to Dr. A. D. Stewart, who guided the writer through his mathematical study.

These persons mentioned above and others have made this paper possible.

G. L. L.

DEDICATION

This research paper is dedicated to my wife, Laverne, my mother, Mrs. Constance Love, and to my mother-in-law, Mrs. Luciler Mosley, for their encouragement, understanding, and love that has been most uplifting during the course of my graduate studies.

G. L. L.

TABLE OF CONTENTS

Chapters Page

I. Introduction.. 1

II. A Study of the curves under certain conditions and a list of auxiliary theorems.................................... 7

III. Main Theorem; two approaches of showing that the binomial has as it limit the normal function............................ 17

IV. Applications.. 30

V. Conclusion.. 35

Chapter I

INTRODUCTION

Some experiments are composed of repetition of independent trials, denoted by n, each with two possible outcomes. The binomial probability distribution may describe the variation that occurs from one set of trials of such a binomial experiment to another. The formula for the binomial provides an adequate approximation providing n is small. For large n, this distribution function can be quite laborious to work with. However, when n is large, the binomial distribution can be adjusted so that it is closely approximated by the standard normal distribution. Although the formula for the standard normal distribution looks a bit unfriendly, bristling as it does with roots, exponents, and transendental numbers like π and e, these are not important features. For our purposes, the important features are that tables of the normal are widely available, and that the transition from a binomial probability problem to a normal probability problem is easy to make once one knows how.

In this paper, the writer will achieve these results by using two approaches. The first method used will be a geometric approach and the second by use of the De Moivre-Laplace Theorem. This theorem was stated by De Moivre in 1733 for the case $p = \frac{1}{2}$ and proved for arbitrary numbers by Laplace in 1812.

In chapter I, terms are defined and meaning of symbols used in this paper will be given.

Chapter II will include some properities of the two distributions. Also a set of auxiliary theorems and lemmas is stated without proof.

Chapter III will give two approaches of showing the binomial distribution has as its limit, as n gets large, the normal distribution.

Chapter IV will give some applications of the theorems proven in Chapter III.

Finally, in Chapter V, the writer will summarize what we have achieved in this research.

SYMBOLS

Below are the symbols that will be used throughout this paper and their meaning.

1.	Def.	Definition
2.	Thm.	Theorem
3.	H_n	Hypothesis number n
4.	C_n	Conclusion number n
5.	$\ni$	"such that"
6.	$\exists$	"there exist"
7.	$\forall$	"for all"
8.	$\Longrightarrow$	"implies"
9.	$\varepsilon > 0$	epsilon a positive number
10.	$\eta > 0$	nu a positive number
11.	$\lvert a \rvert$	absolute value of "a"
12.	$\sum_n$	Summation of n
13.	$<$	less than
14.	$>$	greater than
15.	$\leq$	less than or equal to
16.	$\geq$	greater than of equal to
17.	σ	Standard deviation
18.	σ^2	variance
19.	μ	mean
20.	lim	Limit
21.	f(t)	Distribution Function
22.	P(n)	Probability of n

23. $\binom{n}{x}$	Probability of n things x at a time
24. L. C. D.	Least Common Denominator
25. exp.	exponent
26. p	probability of success
27. q	probability of failure
28. Fig.	Figure
29. log	logarithm
30. $\cong$	approximately

DEFINITIONS

The following are definitions that will be used throughout the remaining portions of this paper.

Symbol and/or terminology	Meaning
1. Binomial Distribution Function	1. $b(x; n,p) = \binom{n}{x} p^x (1 - p)^{n - x}$ $x = 0, 1, 2, \ldots n$
2. Combination of n things x at a time	2. ${}_nC_x = \binom{n}{x} = \frac{n!}{x!\,(n - x)!}$
3. Probability of an event	3. If an experiment can result in any one of n different, equally likely outcomes, and exactly m of these outcomes correspond to event A, then the probability of event A is $P(A) = \frac{m}{n}$.
4. Random Variable	4. A number determined by the outcome of an experiment.
5. Probability Function	5. Let X be a random variable with possible numbers $x_1, x_2, \ldots x_t$ and associated probabilities $f(x_1), f(x_2) \ldots . f(x_t)$. Then the set f whose elements are the ordered pairs $(x_1, f(x_i))$, $i = 1,2,\ldots .t$ is called the probability function of X.
6. Standard Normal Function	6. Let X be a random variable that assumes all the real numbers. Then X is called a standard normal random variable if the probability assigned to the interval from a to b is the area from a to b between the x-axis an the normal curve defined by $Y(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$.

7. Riemann Sum

7. Let $\Delta(m)$ denote the greatest of the m differences $a_k - a_{k-1}$. Consider one of the intervals (a_{k-1}, a_k) and select in it arbitary point $x_k \not> a_{k-1} = x_k = a_k$. The point x_k will be described as the "designated" point in the kth cell for each m cell (a,b) is divided and form the sum

$$S(m) = \sum_{k=1}^{m} (a_k - a_{k-1})\, f(x_k).$$

8. Slope of a Line

8. The slope m of a line L on the points $P(x_1, y_1)$ and $Q(x_2, y_2)$ is given by

$$m = \frac{y_2 - y_1}{x_2 - x_1}, \quad x_2 \neq x_1 .$$

9. Parameter

9. (a) A constant so long as we consider one operation.

   (b) A variable when one generalizes all such operations.

Chapter II

This chapter will include some properties of the binomial distribution and some properties of the standard normal curve. Also included in this chapter are auxiliary theorems and lemmas stated without proof.

PROPERTIES OF THE BINOMIAL DISTRIBUTION

In this section we study the shapes of graphs of binomial distributions produced under two conditions: (1)for a fixed number of trials n, but different values of p; and (2) for fixed numbers for putting values of p, but different values of n. We study especially how the graphs change shape as n grows large. Such a study helps us understand the family of binomial distributions, and it also helps us understand other sequences of probability distribution functions, because the changes within the binomial family resemble the changes within many other families of distributions.

(1) Fixed n, varying p. As p varies, the shape of the graph of the binomial distribution changes. Figures 1 (a) through (i) illustrate this $n = 5$. For p near zero or near one (fig. 1 (a) and (b)), the probability spikes up at $x = 0$ and $x = n$, respectively. The abscissa corresponding to the largest ordinate is called the mode. For p more centrally located, the binomial ordinate $b(x)$, increases with each successive x until the largest $b(x)$ is achieved (fig. 1 (c), (d), (g) and (h)) and then, except possibly for a tie at $x = 1$ (fig. 1 (e), (f) and (i)), $b(x)$ decreases as x continues to increase. Thus, unless two adjacent ordinates are tied, there is just one largest ordinate, and the ordinates decrease steadily as we move to the right or to the left from the mode.

When $p = \frac{1}{2}$ (fig. 1 (i), the distribution is symmetric about $\frac{n}{2}$; if n is odd two central values of x have equal ordinates (fig. 1 (i); if n is even the ordinate at the middle x is the largest (fig. (2)). If $p \neq \frac{1}{2}$ the distribution is asymmetric.

In figures 1 and 2 the fulcrum ▲ on the horizontal axis shows the mean, μ for each distribution. You can see that the means of the binomial distributions are within one unit of the abscissa with largest probability (the mode). In binomial distribution, the mean and mode are always within one unit of each other. Furthermore, if np is an integer the mode and mean are identical.

(2) Fixed p, increasing n. As n increases, the successive binomial distributions (a) "walk" to the right, (b) flatten, and (c) "spread," We discuss these features in turn.

(a) "Walking." As n increases, the mean μ moves to the right a distance p for each unit increase in n because $\mu = np$. The mode and the other large ordinates are near the mean, so the central mass of the distribution also "walks" to the right as n increases.

Later we shall try to obtain a limiting shape for the binomial distribution as n grows large, and to achieve this end we must prevent the distribution from walking off. We can do this by replacing the random variable X by a new variable X - np. For if X is replaced by X - np, then μ is replaced by μ - np, or zero. Hence this adjustment keeps the successive distributions centered at the origin and prevents walk-off.

(b) Flattening. Consider further the unadjusted random variable X. As the means walk, the distributions flatten (fig. 2 (a) through e). We wish to study the rate of flattening. For large n, the sizes of the central ordinates are inversely proportional to $\sqrt{n}$. Let us illustrate this fact graphically. To do this, let us first recall that y = mx is an equation of a straight line through the origin with slope m. When y is a constant times x, y varies directly as x. If $x = 1/\sqrt{n}$, then we usually say that y varies inversely as n . But we can also say that y varies directly as $1/\sqrt{n}$; and when we plot y against $1/\sqrt{n}$, we get a straight line through the origin. The point is that we have a linear relation if we regard $1/\sqrt{n}$ as an independent variable. In other words, one way to show that y is inversely proportional to n is to show that y is directly proportional to $1/\sqrt{n}$.[1] We use this idea in 3. (Fig.)

Mode proportional to $1/\sqrt{n}$. To return to the main discussion, Figure 3 shows how the middle ordinates of symmetric binomial distributions ($p = \frac{1}{2}$) decrease as n grow. The relation is smooth when n is taken as even. (A similar smooth relation holds for n odd.) The modal value of x is $\frac{n}{2}$. When we choose the horizontal axis as the axis of $1/\sqrt{n}$, we see that, as n grows, p(mode) decreases, following a curve that is almost a straight line through the origin. (A scale of values of n is marked below the axis.) The points on the curve for $p = \frac{1}{2}$ have coordinates ($1/\sqrt{n}$, $b(\frac{1}{2}n)$), n even. Our binomial table can be used to check a point on the curve. For n = 24, $b(12) \approx 0.161$ and and $1/\sqrt{24} \approx 0.204$.

Similarly, the relation between P(mode) and $1/\sqrt{n}$ is approximated by a straight line through the origin for binomial distributions with $p = \frac{1}{5}$ (for smoothness, we have chosen values of n that are multiples of 5, and then the mode is $\frac{n}{5}$). Our binomial table can be used to check a point for n = 25. Then $1/\sqrt{n} = 0.20$ and b(5) 0.196. This graph is adequate to illustrate the approximation; the modal ordinate and its neighbors decrease inversely a $\sqrt{n}$.

[1]Frederick Mosteller et. al., Probability With Statistical Applications, (Reading, Massachusetts, : Addison-Wesley Publishing Company, Inc., 1965), p. 263.

When we study the limiting distribution of X as n grows, we shall need to prevent the binomial distributions with large n from collapsing onto the x-axis. As figure 3 illustrates, as n grows, $1/\sqrt{n}$ tends to zero.

(c) Spreading. We recall that the sum of the ordinates is always 1. Naturally, if the distributions flatten as n increases, and the total probability must remain constant, successive distributions must spread out. They spread at a rate proportional to $\sqrt{n}$. For the standard deviation is a measure of spread, and its value for a binomial distribution, $\sqrt{npq}$, is proportional to $\sqrt{n}$ when p is constant.

The total range of the binomial is from o to n, and increases at a rate proportional to n. But we know from Chebyshev's theorem (given in the list of auxiliary theorems) that there is very little probability near the ends of the distribution compared with the amount within a few standard deviations of the mean. The standard deviation is sensitive to the rate at which the central mass of the binomial distribution spreads (the 75% or the 99% near the mean), and it is this central mass that we want to study.

To summarize; as n grows, (a) successive binomial distribution walk to the right at a rate proportion to n; (b) the modal ordinates flatten at a rate proportional to $1 / \sqrt{n}$; and (c) the distributions spread out, that is their standard deviations increase in proportion to $\sqrt{n}$. (See the accompaning figures on the next three successive pages).

PROPERTIES OF THE NORMAL CURVE

From an inspection of the graph (fig. 4), we note that the normal curve:

(a) is symmetric about the y-axis;

(b) has its highest point a ($0, 1/\sqrt{2\pi}$), where $1/\sqrt{2\pi} \approx 0.40$;

(c) is concave downward between x = -1 and x = +1, and concave upward for x out sides the interval;

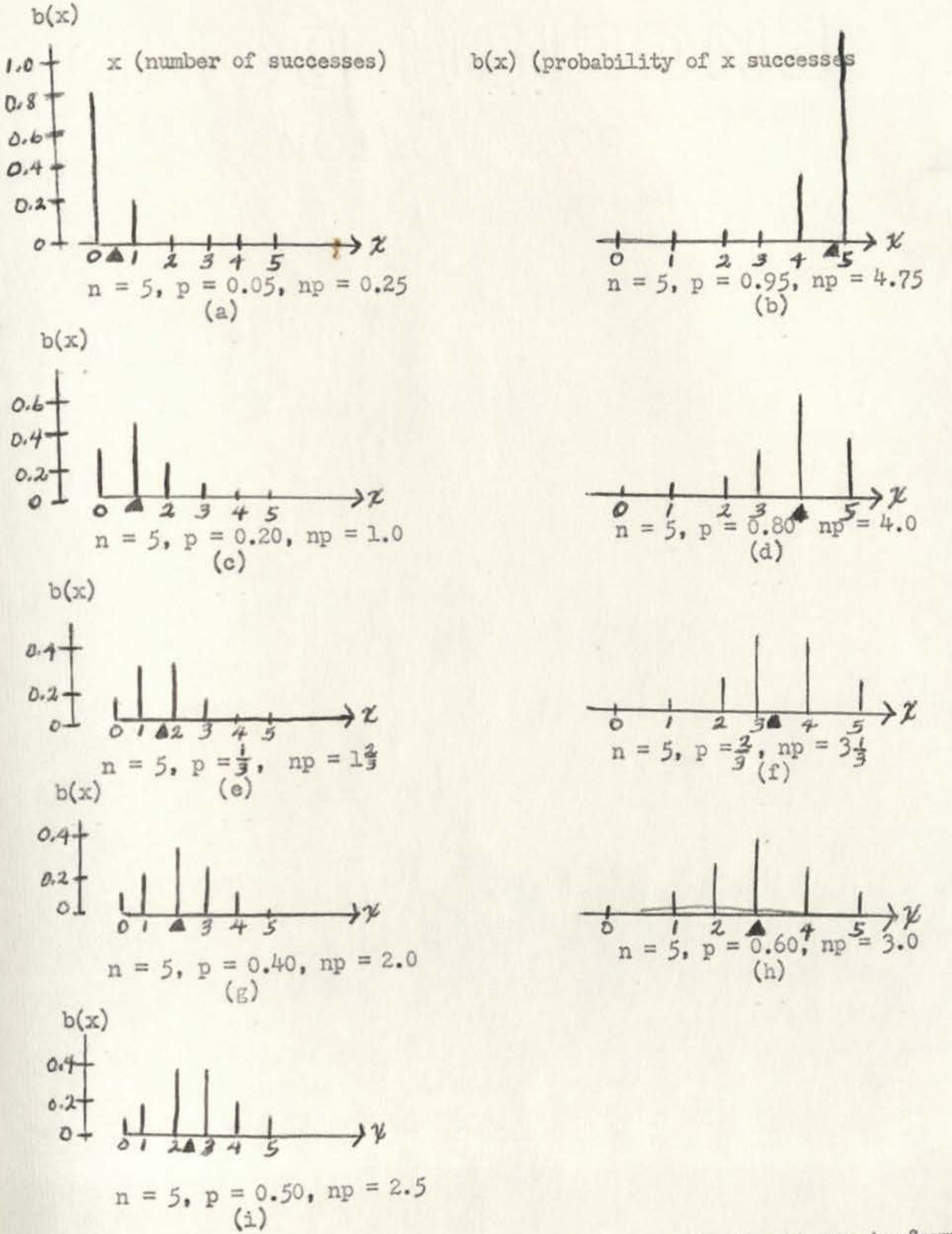


Figure 1. Binomial Distribution for n = 5, displaying the change in form as p varies.

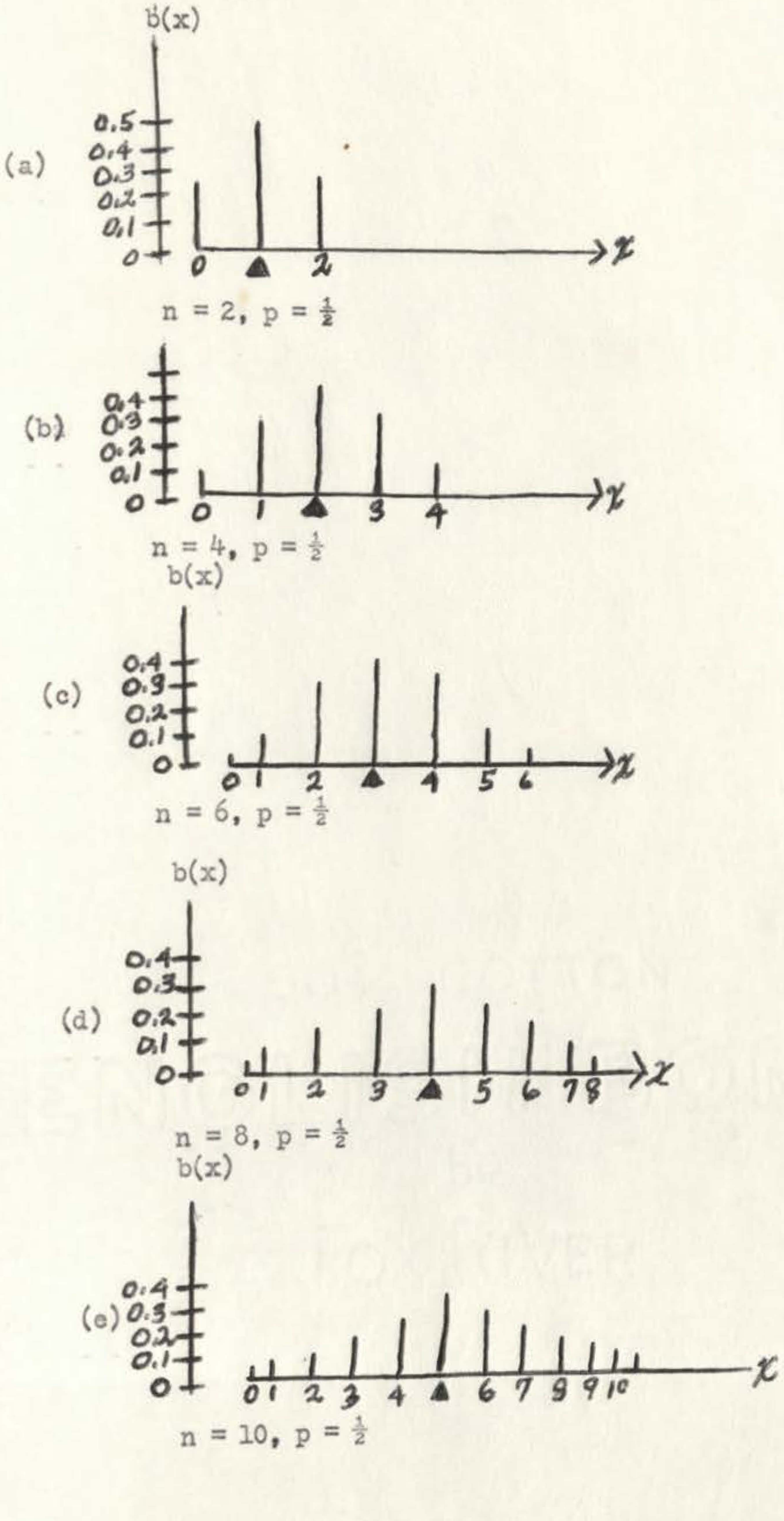


Figure 2. Walking, flattening, and spreading as n increases.

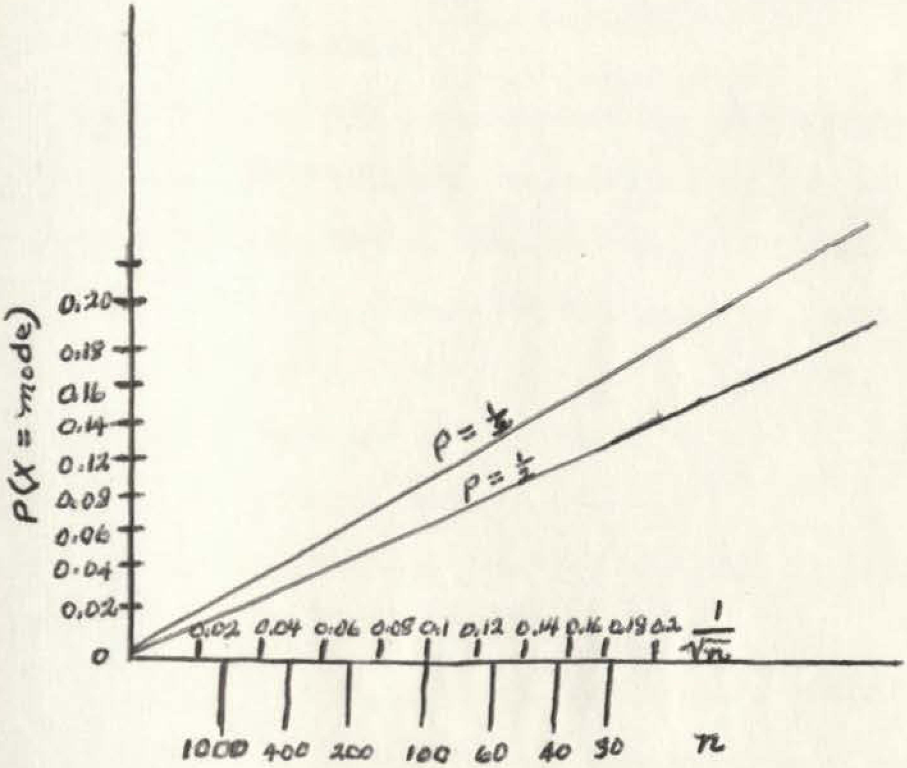


Figure 3. Plot of P(X = mode) against $1/\sqrt{n}$ to show the nearly straight-line relationship for $p = \frac{1}{2}$ and $p = \frac{1}{5}$.

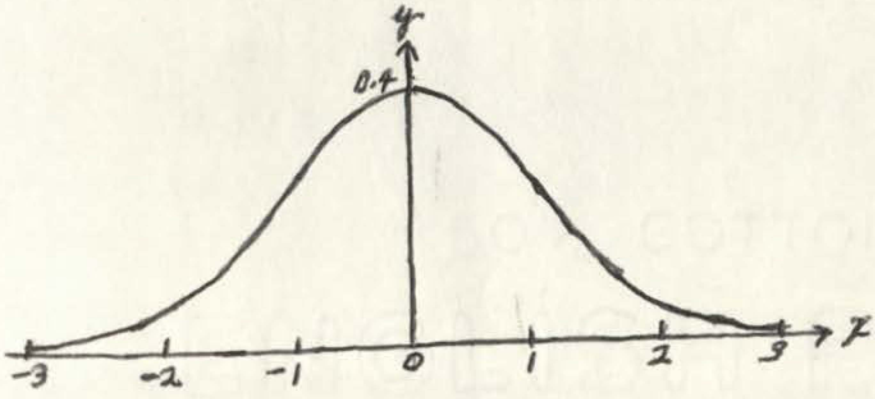


Figure 4. The Normal Curve, whose equation is $Y(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$.

(d) extends without limit to the left and to the right, and approaches the x-axis very rapidly as we move away from x = 0 in either direction.

(e) the total area under the curve and above the x-axis equals 1.

(f) The arithemetic mean, median, and mode coincide.

(g) The standard deviation cuts the curve at the points of inflection (the points on the steep slope where the curvature reverses).

(h) The first and third quartiles are equally distant from the median.

(i) Within one standard deviation taken plus and minus from the mean, 68.26% of the items fall.

(j) Two standard deviation taken plus and minus from the mean contain 95.45% of the items.

(k) The average deviation is .7979 of the standard deviation.

(l) The semi-interquartile range equals the probable error, which equals .6745 of the standard deviation.[2]

Within the normal probability curve there exist certain fixed ratios, relations, proportions and distributions which may be stated in standard statistical and mathematical terms and universally applied with a high degree of accuracy for all distributions of data approximating the form of the normal curve. For any data which fall approximately in the form of the normal curve and are correctly based on the logical principles underlying the normal curve, the ratios stated above will hold true and may logically be deduced from the samples as applicable to the population from which they came.

---

[2] Elmer B. Mode, Elements of Statistics, (Englewood Cliffs, N. J.,: Prentice-Hall, Inc., 1951), p. 119.

AUXILIARY THEOREMS

The following theorems are stated without proof and will be used throughout the remaining parts of this paper.

Auxiliary Theorem I

H_1: $f(x) = e^{-\frac{1}{2}x^2}$

C: $\lim_{c} \int_{-a}^{a} e^{-\frac{1}{2}x^2}\,dx = \sqrt{2\pi}$

Auxiliary Theorem II

H_1: $q + p = 1$

H_2: $f(x) = \frac{n!}{x!(n-x)!} p^x q^{n-x}$

C_1: $0 \leq f(x) \leq 1$

C_2: $\sum_{x=0}^{n} \frac{n!}{x!(n-x)!} p^x q^{n-x} = 1$

Auxiliary Theorem III

H_1: $Y(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}$ is the probability density function

C; $\lim \int_{-a}^{a} Y(x)\,dx = 1$

Auxiliary Theorem IV
(Taylor's Formula)

H_1: Let $f(x)$ be a function of x defined over the interval (a,b)

H_2: $f(x)$ is differentiable n times on (a,b)

H_3: x_o a fixed point and x a variable point contained in (a,b)

H_4: $f^{(k)}$ is the kth derivative of $f(x)$

C_1: $f(x) = f(x_o) + \frac{x - x_o}{1!} f'(x_o) + \frac{(x - x_o)^2}{2!} f''(x_o) + \dots + \frac{(x - x_o)^k}{k!} f^{(k)}(x_o) + \dots + \frac{(x - x_o)^{n-1}}{(n-1)!} f^{(n-1)}(x_o) + R_n$

where R_n is the remainder term.

Auxiliary Theorem V
(Stirling's Formula)

$$H_1:\ 1 < e^{\frac{\theta}{12}} < e^{\frac{1}{12n}}$$

$$H_2:\ \lim_n e^{\frac{\theta}{12}} = 1$$

$$H_3:\ S(n) = n^n e^{-n}\sqrt{2\pi n}$$

$$C:\ \frac{n!}{S(n)} = e^{\frac{\theta}{12n}}$$

Auxiliary Theorem VI

$$H_1:\ \frac{n!}{S(n)} = e^{\frac{\theta}{12}}$$

$$H_2:\ P_x(K/n) = \frac{n!}{k!(n-k)!}\, p^k (1-p)^{n-k}$$

$$C:\ P_1(K/n) = \frac{1}{\sqrt{np(1-p)}}\ \frac{1}{\sqrt{2\pi}} \left(\frac{np}{k}\right)^{k+\frac{1}{2}} \left(\frac{n(1-p)}{n-k}\right)^{n-k+\frac{1}{2}}$$

Auxiliary Theorem VII
(Duhamel's Lemma)

$$H_1:\ S_n = \sum_{i=1}^{N(n)} T_{ni},\quad \Sigma_n = \sum_{i=1}^{N(n)} T'_{ni},\ n = 1,2,\ldots$$

$$H_2:\ \lim_n \Sigma_n = L, \text{ for every } \epsilon\ \exists\, n(\epsilon) \ni n > n(\epsilon) \Rightarrow \left| \frac{T_{ni}}{T'_{ni}} - 1 \right| < \epsilon$$

$$C:\ \lim_n S_n = L$$

Chapter III

In this chapter, the main results will be presented. It will be shown that the binomial function has the normal function as its limit as n gets large. Two approaches are presented. The first approach will be the geometric or slope method. The second approach involves areas.

PROOF OF APPROACH I

We may expect, that the fitted normal curve will give a fair approximation to the binomial except possibly at the extremities of the range. When the terms of the binomial are arranged symmetrically with respect to the mean, that is when $p = q$, the approximation is considerably better than when either p or q is small compared with the other.

Because of the central role played by the normal law in statistical theory, it will be beneficial to prove that the limiting form of the binomial function is the normal function.

The variable is first changed to t, where $t = (x - np)/\sigma$, $\sigma = (npq)^{\frac{1}{2}}$ so that the step of t is $1/\sigma$. This step becomes small as n gets large. The slope of the straight line joining the tops of two successive t ordinates is then equated to the slope of a continuous approximating curve at the midpoint. To maintain the area under the curve unaltered, the ordinates of the points are multiplied by σ at the same time the abscissae are divided by σ. Figure 5 will indicate only one subdivision of the binomial curve but this is then true for all subdivisions. Thus, we formulate the normal curve in its entirety.

The De Moivre-Laplace Theorem

Let X_1, X_2,....X_n be a sequence of random variables, where X_n is the number of successes in a binomial experiment with n trials, each with the probability of success p, $0 < p < 1$. Let Z_n, $n = 1,2,3,...$ be a corresponding sequence of adjusted random variables where

$$Z_n = \frac{X_n - np}{\sqrt{npq}}$$ and let z be a constant. Then

as n gets large $P(Z_n \geq z)$ approaches the area to the fight of z for the standard normal distribution.

Theorem I

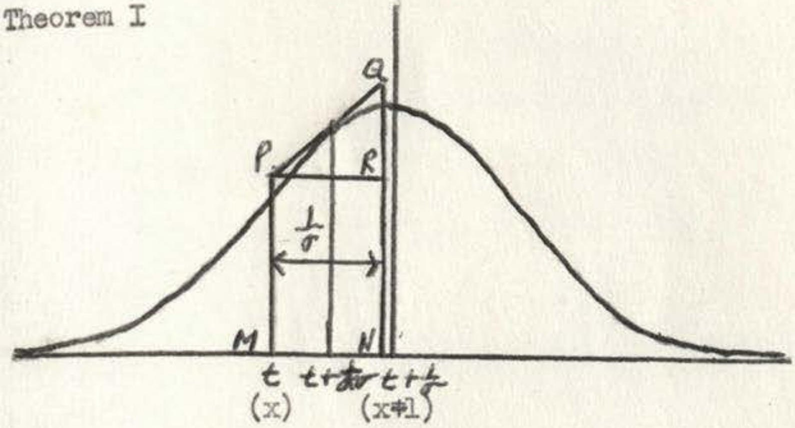


Figure 5

H_1: $f(t) = \dfrac{n!}{x!(n-x)!} p^x q^{n-x}$

H_2: $x = np + \sigma t$

H_3: Slope of PQ $= \dfrac{NQ - MP}{MN} = \dfrac{f(t+\frac{1}{\sigma}) - f(t)}{\frac{1}{\sigma}}$

H_4: $Y(t) = \frac{1}{2}\left[f(t+\frac{1}{\sigma}) + f(t)\right]$

C: $Y(t) = \dfrac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}$

Proof:

(1) Using H_1 and Fig. 5, $f(t+\frac{1}{\sigma}) = \dfrac{\sigma n!}{(x+1)!\,(n-x+1)!} p^{x+1} q^{[n-(x+1)]}$

(2) And since $x = np + \sigma t$, it follows that

(3) $$f(t+\tfrac{1}{\sigma}) = \frac{\sigma\left(\frac{(x-\sigma t)}{p}\right)!}{\left[(np+\ t)+1\right]!\left[\left(\frac{x-\sigma t}{p}\right) - \left[(np+\sigma t)+1\right]\right]} p^{[(np+\sigma t)+1]} q^{\left[\frac{x-\sigma t}{p} - [(np+\sigma t+1]\right]}$$

But by definition,

(4) $$f(t+\tfrac{1}{\sigma}) = \binom{\frac{x-\sigma t}{p}}{np+\sigma t+1} p^{[(np+\sigma t)+1]} q^{\left[\left(\frac{x-\sigma t}{p}\right) - [(np+\sigma t)+1]\right]}$$

Since p = q, adding exponents we obtain,

(5) $$f(t+\tfrac{1}{\sigma}) = \binom{\frac{x-\sigma t}{p}}{np+\sigma t+1} pq^{[(np+\sigma t)+1] + \left[\left(\frac{x-\sigma t}{p}\right) - [(np+\sigma t)+1]\right]}$$

Finding the L.C. D. in the exponents we obtain

(6) $f(t+\frac{1}{\sigma}) = \binom{\frac{x-\sigma t}{p}}{np+\sigma t+1} pq^{\left[\frac{np^2+p\sigma t+p+x-\sigma t-np^2-p\sigma t-p}{p}\right]}$

Simplifying

(7) $f(t+\frac{1}{\sigma}) = \binom{\frac{x-\sigma t}{p}}{np+\sigma t+1} pq^{\frac{x-\sigma t}{p}}$

By H_2,

(8) $f(t+\frac{1}{\sigma}) = \binom{n}{x+1} pq^n$, Also by def.

(9) $f(t+\frac{1}{\sigma}) = p(x+1, n)$

(10) Using H_1 and Fig.5 again then $f(t) = \frac{\sigma n!}{x!(n-x)!} p^x q^{n-x}$

Then

(11) $f(t) = \frac{\sigma\left(\frac{x-\sigma t}{p}\right)!}{(np+\sigma t)!\left[\left(\frac{x-\sigma t}{p}\right)-np+\sigma t\right]!} p^{np+\sigma t} q^{\left[\left(\frac{x-\sigma t}{p}\right)-np+\sigma t\right]}$

But by definition

(12) $f(t) = \binom{\frac{x-\sigma t}{p}}{np+\sigma t} p^{np+\sigma t} q^{\left[\left(\frac{x-\sigma t}{p}\right)-np+\sigma t\right]}$

Adding exponents

(13) $f(t) = \binom{\frac{x-\sigma t}{p}}{np+\sigma t} pq^{np+\sigma t+\left[\left(\frac{x-\sigma t}{p}\right)-np+\sigma t\right]}$

Finding the L. C. D. in the exponents

(14) $f(t) = \binom{\frac{x-\sigma t}{p}}{np+\sigma t} pq^{\left[\frac{np^2+p\sigma t-x-\sigma t-np^2-p\sigma t}{p}\right]}$

Simplifying

(15) $f(t) = \binom{\frac{x-\sigma t}{p}}{np+\sigma t} pq^{\frac{x-\sigma t}{p}}$

By H_2

(16) $f(t) = \binom{n}{x} pq^n$, Also by def.

(17) $f(t) = p(x,n)$

Using steps 9 and 17 then

(18) $\frac{f(t+\frac{1}{\sigma})}{f(t)} = \frac{p(x+1,n)}{p(x,n)}$

Since this is true then

$$(19)\quad \frac{p(x+1,n)}{p(x,n)} = \frac{\dfrac{\sigma\, n!}{(x+1)!\,[n-(x+1)]!}\, p^{x+1} q^{n-(x+1)}}{\dfrac{n!}{x!\,(n-x)!}\, p^{x}\, q^{n-x}}$$

This implies that

$$(20)\quad \frac{p(x+1,n)}{p(x,n)} = \left[\frac{\sigma\, n!}{(x+1)!\,[n-(x+1)]!}\right]\cdot\left[\frac{x!\,(n-x)!}{\sigma\, n!}\right]\cdot\left[\frac{p^{x+1}\, q^{n-(x+1)}}{p^{x}\, q^{n-x}}\right]$$

$$(21)\quad \frac{p(x+1,n)}{p(x,n)} = \left[\frac{x!\,(n-x)!}{(x+1)!\,[n-(x+1)]!}\right]\cdot\left[\frac{p^{x+1}\, q^{n-(x+1)}}{p^{x}\, q^{n-x}}\right]$$

If we expand this and subtract exponents

$$(22)\quad \frac{p(x+1,n)}{p(x,n)} = \frac{n-x}{x+1}\, p^{x+1-x}\, q^{(n-x+1-n-x)}$$

$$(23)\quad \frac{p(x+1,n)}{p(x,n)} = \frac{n-x}{x+1}\, p\cdot q^{-1}$$

$$(24)\quad \frac{p(x+1,n)}{p(x,n)} = \frac{n-x}{x+1}\cdot\frac{p}{q}$$

Looking at step 18 then

$$(25)\quad \frac{f(t+\frac{1}{\sigma})}{f(t)} = \frac{p(x+1,n)}{p(x,n)} = \frac{n-x}{x+1}\cdot\frac{p}{q}$$

Hence:

$$(26)\quad \frac{f(t+\frac{1}{\sigma})-f(t)}{f(t+\frac{1}{\sigma})+f(t)} = \frac{(n-x)p-(x+1)q}{(n-x)p+(x+1)q}$$

$$(27)\quad \frac{f(t+\frac{1}{\sigma})-f(t)}{f(t+\frac{1}{\sigma})+f(t)} = \frac{np-xp-xq-q}{np-xp+xq+q}$$

$$(28)\quad \frac{f(t+\frac{1}{\sigma})-f(t)}{f(t+\frac{1}{\sigma})+f(t)} = \frac{np-q-x(p+q)}{np+q+(q-p)x}$$ but $p+q=1$, so

$$(29)\quad \frac{f(t+\frac{1}{\sigma})-f(t)}{f(t+\frac{1}{\sigma})+f(t)} = \frac{np-q-x}{np+q+(q-p)x}$$

Since $x = np + \sigma t$, it follows that

$$(30)\quad \frac{f(t+\frac{1}{\sigma})-f(t)}{f(t+\frac{1}{\sigma})+f(t)} = \frac{np-q-np-\sigma t}{np-q+(q-p)(np+\sigma t)}$$

(31) $\frac{f(t+\frac{1}{\sigma}) - f(t)}{f(t+\frac{1}{\sigma}) + f(t)} = \frac{-q - \sigma t}{np + q + qnp + q\sigma t - np - \sigma pt}$

but qnp is σ^2 and np = σ^2, factoring out σt and -1 in numerator

(32) $\frac{f(t+\frac{1}{\sigma}) - f(t)}{f(t+\frac{1}{\sigma} + f(t)} = - \frac{q + \sigma t}{q + 2\sigma^2 + (q-p)\sigma t}$

The ordinate of the approximating curve midway between t and t + $\frac{1}{\sigma}$ nearly the right side of H_4, so that if Y(t) is the ordinate of this approximating curve, we suppose that

(33) $\frac{1}{Y(t)} D_t\, Y(t) = \lim_{\sigma} \frac{\text{slope of PQ}}{\text{mid-ordinate}}$

(34) $\frac{1}{Y(t)} D_t\, Y(t) = \lim_{\sigma} \frac{\sigma\{f(t+\frac{1}{\sigma}) - f(t)\}}{\frac{1}{2}\{f(t+\frac{1}{\sigma}) + f(t)\}}$

(35) $\frac{1}{Y(t)} D_t\, Y(t) = \lim_{\sigma} \frac{-2\sigma(q+\sigma t)}{q + 2\sigma^2 + (q-p)\sigma t}$

Substituting t - $\frac{1}{2\sigma}$ for t

(36) $\frac{1}{Y(t)} D_t\, Y(t) = \lim_{\sigma} \frac{-2\sigma(q+\sigma t - \frac{1}{2})}{q + 2\sigma^2 + (q-p)\sigma t - \frac{q-p}{2}}$

(37) $\frac{1}{Y(t)} D_t\, Y(t) = \frac{-2\sigma(-\frac{1}{2} + q + \sigma t)}{\frac{q}{2} + \frac{p}{2} + 2\sigma^2 + (q-p)\sigma t}$

(38) $\frac{1}{Y(t)} D_t\, Y(t) = \frac{-2\sigma(q(\frac{\sigma}{\sigma}) + \sigma t - \frac{1}{2}(\frac{\sigma}{\sigma})}{\frac{q+p}{2} + 2\sigma^2 + (q-p)\sigma t}$

(39) $\frac{1}{Y(t)} D_t\, Y(t) = \frac{-2\sigma^2(\frac{q}{\sigma} + t - \frac{1}{2\sigma})}{\frac{1}{2} + 2\sigma^2 + (q-p)\sigma t}$

(40) $\frac{1}{Y(t)} D_t\, Y(t) = \frac{-2\sigma^2 \frac{2q-1}{2\sigma} + t)}{(\frac{1}{2}(\frac{2\sigma^2}{2\sigma^2}) + 2\sigma^2 + (\frac{2\sigma}{2\sigma})(q-p)\sigma t}$

(41) $\frac{1}{Y(t)} D_t\, Y(t) = \frac{-2\sigma^2 \frac{2q-1}{2\sigma} + t}{2\sigma^2 (\frac{1}{4\sigma^2} + 1 + \frac{q-p}{2\sigma})t}$

(42) $$\frac{1}{Y(t)} D_t Y(t) = \frac{-\left(\frac{2q - (p + q)}{2\sigma} + t\right)}{\frac{1}{4\sigma^2} + 1 + \frac{q - p}{2\sigma} t}$$

(43) $$\frac{1}{Y(t)} D_t Y(t) = \lim \frac{-\frac{q - p}{2\sigma} - t}{1 + \frac{1}{4\sigma^2} + \frac{q - p}{2\sigma} t}$$

(44) $$\frac{1}{Y(t)} D_t Y(t) = - t$$

(45) $$\log_e Y(t) = \int (-t)\, dt + c$$

(46) $$\log_e Y(t) = -\tfrac{1}{2}t^2 + c$$

(47) $$Y(t) = e^{-\frac{1}{2}t^2 + c}$$ def. of log.

(48) $$Y(t) = e^{-\frac{1}{2} t^2} \cdot e^{c}$$

(49) $$Y(t) = Ae^{-\frac{1}{2}t^2}$$

Now since t ranges from -a to a as x gets large, and since $Y(t)\, dt = 1$

(50) $$Y(t) = \int_{-a}^{a} Ae^{-\frac{1}{2}t}$$

(51) $$Y(t) = A\int_{-a}^{a} e^{-\frac{1}{2}t^2}$$ and because $\int e^{-\frac{1}{2}t^2} = \sqrt{2\pi}$ then

(52) $Y(t) = A\sqrt{2\pi} = 1$, then $A = \frac{1}{\sqrt{2\pi}}$ so,

Therefore

(53) $$Y(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}$$

Q. E. D.

PROOF OF APPROACH II

The sum of successive terms of the binomial equals the area of the corresponding rectangles in its histogram. We may obtain an approximation to this sum by finding the area under the fitted normal curve which these rectangles occupy. A Riemann sum will be built to so illustrate. Graphically, $x = 0,1,2,\ldots.n$ are the mid-points of the bases of these rectangles. Therefore, if we are summing the terms of the binomial in which x ranges from $x = t_1$ to $x = t_2$, inclusive, the corresponding area under the curve will be from $x = t_1 - \frac{1}{2}$ to $x = t_2 + \frac{1}{2}$.

The problem naturally arises of substitution for the exact expression of P an approximate formula which will be easy to understand and use in practice and which, for large n, will give a sufficiently close approximation to P. De Moivre was the first successfully to attack this difficult problem. After him, in essentially the same way, but using more powerful analytical tools, Laplace succeeded in establishing a simple approximate formula which is given in all books on probability.

When we use an approximate formula instead of an exact one, there is always this question to consider: How large is the committed error? If, as is usually done, this question is left unanswered, the derivation of Laplace's formula becomes an easy matter. However, to estimate the error comparatively long and detailed investigation is required. Except of the length, this investigation is not very difficult.

Theorem I'

H_1: $k = np + x_k(n)\sqrt{np(1-p)}$, $A(n) \leq k \leq B(n)$

$$H_2:\ P_2(k|n) = \frac{\frac{1}{\sqrt{np(1-p)}}}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x_k^2(n)}$$

$$H_3:\ P_1(k|n) = \frac{1}{\sqrt{np(1-p)}}\,\frac{1}{\sqrt{2\pi}}\left(\frac{np}{k}\right)^{k+\frac{1}{2}}\left(\frac{n(1-p)}{n-k}\right)^{n-k+\frac{1}{2}},\quad 1 \leq k \leq n$$

$$C_1:\ \frac{P_1(k|n)}{P_2(k|n)} = \left|\frac{\frac{1}{\sqrt{np(1-p)}}\,\frac{1}{\sqrt{2\pi}}\left(\frac{np}{k}\right)^{k+\frac{1}{2}}\left(\frac{n(1-p)}{n-k}\right)^{n-k+\frac{1}{2}}}{\frac{\frac{1}{\sqrt{np(1-p)}}}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x_k^2(n)}} - 1\right| < \mathcal{E}$$

for n sufficiently large

$$C_2:\ P_2(t_1, t_2, n) = \sum_{k=A(n)}^{B(n)} P_2(k|n) = \sum_{k=A(n)}^{B(n)} \frac{\frac{1}{\sqrt{np(1-p)}}}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x_k^2(n)}$$

$$C_3:\ P_2(t_1, t_2, n) = \int_{t_1}^{t_2} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x^2}\, dx$$

Proof:

(1)

(2) Let k assume all numbers from $A(n)$ through $B(n)$ then $x_k(n)$ increases by equal steps of

(3) $$x_k(n) - x_{k-1}(n) = \frac{1}{\sqrt{np(1-p)}} = \Delta_n$$

The length of the steps tends to zero as n gets large. The first of the $x_k(n)$ is $x_{A(n)}(n)$, and since $A(n)$ denotes the smallest number larger than $np + t_1\sqrt{np(1-p)}$, then

(4) $$t_1 < x_{A(n)}(n) = \frac{A(n) - np}{\sqrt{np(1-p)}} \leq t_1 + \Delta_n$$

Similarly, the last of the numbers $x_k(n)$ is $x_{B(n)}(n)$. Since $B(n)$ is defined as the greatest number which is less than $np + t_2\sqrt{np(1-p)}$ it follows that

$$(5)\qquad t_2 - \Delta_n \leq x_{B(n)}(n) = \frac{B(n) - np}{\sqrt{np(1-p)}} < t_2$$

Using H_2 and substitution of H_1 into the expression H_3 and evaluate the logarithm of the reciprocal of $P_1(k\,n)$ that is

$$\frac{1}{P_1(k\,n)} = \frac{1}{\frac{1}{\sqrt{np(1-p)}}\,\frac{1}{\sqrt{2\pi}}\left(\frac{np}{k}\right)^{k+\frac{1}{2}}\left(\frac{n(1-p)}{n-k}\right)^{n-k+\frac{1}{2}}}$$

We have

$$(6)\qquad -\log P_1(k|n) = \log\sqrt{2\pi} - \log\frac{1}{\sqrt{np(1-p)}} + \left(np + x_k(n)\sqrt{np(1-p)} + \tfrac{1}{2}\right)\log\left(1 + x_k(n)\sqrt{\frac{1-p}{np}}\right) + \left(n(1-p) - x_k(n)\sqrt{np(1-p)} + \tfrac{1}{2}\right)\log\left(1 - x_k(n)\sqrt{\frac{p}{n(1-p)}}\right)$$

Now using Taylor's formula, applying it to the evaluation of the two logarithms in the right hand side, we write

$$(7)\qquad W_k = \log P_1(k|n) - \log\sqrt{2\pi} + \log\frac{1}{\sqrt{np(1-p)}}$$

$$(8)\qquad W_k = \left[np + x_k(n)\sqrt{np(1-p)} + \tfrac{1}{2}\right]\left[x_k(n)\sqrt{\frac{1-p}{np}} - \frac{x_k^2(n)}{2}\,\frac{1-p}{np} + \frac{x_k^2(n)}{3}\left(\frac{1-p}{np}\right)^{\frac{3}{2}}\frac{1}{\left(1 + \vartheta_1 x_k(n)\sqrt{\frac{1-np}{np}}\right)^3}\right]$$

$$+\left[n(1-p) - x_k(n)\sqrt{np(1-p)} + \tfrac{1}{2}\right]\left[x_k(n)\sqrt{\frac{p}{n(1-p}} - \frac{x_k^2(n)}{2}\,\frac{p}{n(1-p} - \frac{x_k^3(n)}{3}\left(\frac{p}{n(1-p)}\right)^{\frac{3}{2}}\frac{1}{\left(1 - \vartheta_2 x_k(n)\sqrt{\frac{p}{n(1-p)}}\right)^3}\right]$$

where ϑ_1 and ϑ_2 are two numbers between zero and unity. Only the first three terms are considered.

(9) $np\, x_k(n)\sqrt{\frac{1-p}{np}} - n(1-p)x_k(n)\sqrt{\frac{p}{np(1-p)}} = 0$

In each of the two products there will be exactly two terms free from n. They

(10) $\left[-\tfrac{1}{2}x_k^2(n)(1-p) + x_k^2(n)(1-p)\right] + \left[-\tfrac{1}{2}x_k^2(n)p + x_k^2(n)p\right] = \tfrac{1}{2}x_k^2(n).$

The other terms in the two products denominators include n raised to the power one-half or higher. The total of these terms can be written as

(11) $\frac{u_k(n)}{\sqrt{N}}$. (Step 8)

Observing the numerator $u_k(n)$ it is necessary to show the existence of positive number $M(t_1, t_2, p)$ which depends on t_1, t_2, and p but not on n and k for large n,

(12) $| u_k(n) | < M(t_1, t_2, p)$

$\forall$ k between the limits $A(n) \leq k \leq B(n)$.

(13) Let Υ stand for the greater of $|t_1|$ and $|t_2|$ then $|x_k(n)| < \Upsilon$, every term in $u_k(n)$ is bounded.

Since step 12 holds for large n and $\forall$ k $\ni$ $A(n) \leq k \leq B(n)$ we can write step 7 as

(14) $W_k = -\log P(k\ n) - \log\sqrt{2\pi} + \log\sqrt{\frac{1}{np(1-p)}}$

$= \tfrac{1}{2}x_k^2(n) + \frac{u_k(n)}{\sqrt{n}}$. This gives

(15) $P_1(k|n) = \frac{\frac{1}{\sqrt{np(1-p)}}}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x_k^2(n)}\, e^{\frac{-u_k(n)}{\sqrt{n}}}$

(16) $P_1(K|n) = P_2(k|n)\, e^{\frac{-u_k(n)}{\sqrt{n}}}$.

From this we conclude that the terms $P_2(k|n)$ satisfy the Lemma of Duhamel. That is $\forall \varepsilon > 0\ \exists$ a $n_2(\varepsilon) \ni$ the inequality $n > n_2(\varepsilon) \Longrightarrow$

(17) $\left| \frac{P_1(k\ n)}{P_2(k\ n)} - 1 \right| < \varepsilon$

$\forall$ k $\ni$ $A(n) \leq k \leq B(n)$. This follows from steps 12 and 16 because both $\Longrightarrow$ that for large n

(18) $$e^{-\frac{M(t_1,t_2,p)}{\sqrt{n}}} < \frac{P_1(k/n)}{P_2(k/n)} = e^{-\frac{u_k(n)}{\sqrt{n}}} \; e^{\frac{M(t_1,t_2,p)}{\sqrt{n}}}$$

In order to satisfy set 17 for all k, n must be large enough so that the two extreme sides in step 18 differ from unity by less than . Now, using the lemma of Duhamel we conclude that if the sum

(19) $$P_2(t_1, t_2, n) = \sum_{k=A(n)}^{B(n)} P_2(k/n) = \sum_{k=A(n)}^{B(n)} \frac{\frac{1}{\sqrt{np(1-p)}}}{\sqrt{2\pi}} \, e^{-\frac{1}{2}x_k^2(n)}$$

tends to the limit L as n is increased. Then similarly, the same number L is the limit of $P_1(t_1, t2, n)$. Hence it is sufficient to show that

(20) $$P_2(t_1, t_2, n) = \sum_{k=A(n)}^{B(n)} \frac{\frac{1}{\sqrt{np(1-p)}}}{\sqrt{2\pi}} \, e^{-\frac{1}{2}x_k^2(n)}$$

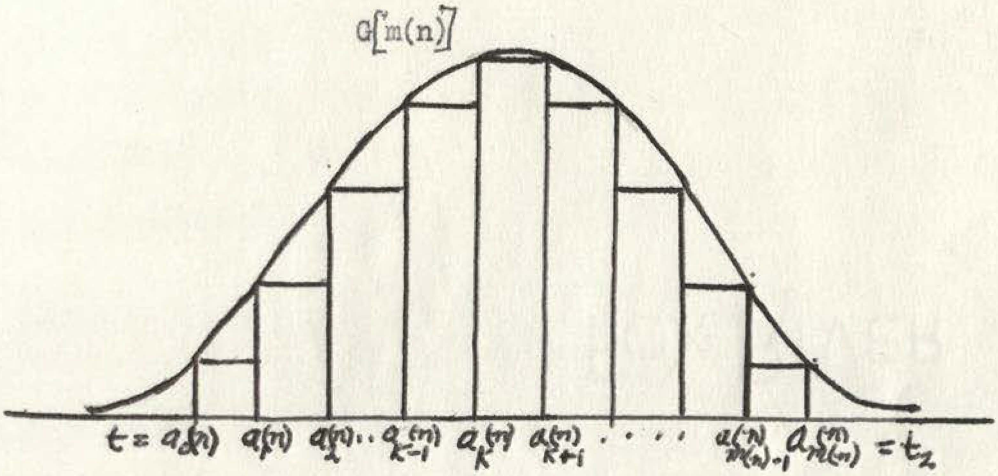


Figure 6

(21) Consider any n and the corresponding sum $P_2(t_1, t_2, n)$ which differ from a Riemann sum by a quantity which tends to zero. Let m(n) denote the number of terms in the sum. Then

$$m(n) = B(n) - A(n) + 1$$

(22) Denote by $G[m(n)]$ the grid composed of $m(n)$ cells extending from t_1 to t_2 and defined as follows. The first point of the grid is $a_o(n) = t_1$. The next point is $a_1(n) = a_o(n) + \Delta_n$, where

$$\Delta_n = \frac{1}{\sqrt{np(1-p)}} .$$

(23) Let the $(k+1)$st point of the grid be $a_k(n) = a_{k-1}(n) + \Delta_n \ \forall\ k = 1,2,,,,m(n)-1$. The last point of the grid will be $a_{m(n)}(n) = t_2$.

(24) By definition, the first $m(n) - 1$ cells of the grid have the same length Δ_n. The length of the last cell say $\delta(n)$, may be equal to, smaller than of greater than Δ_n. In fact

(25) $$\delta(n) = t_2 - a_{m(n)-1}(n) = t_2 - t_1 - [m(n) - 1]\Delta_n$$

$$\delta(n) = t_2 - t_1 - [B(n) - A(n)]\Delta_n$$

(26) In order to obtain the limits for $[B(n) - A(n)]\Delta_n$ we use the meaning of Δn and steps 4 and 5 then subtracting equation 5 from equation 4 we obtain

(27) $$t_2 - t_1 - 2\Delta_n \leq (B(n) - A(n)\Delta_n < t_2 - t_1$$

then from step 25 we get

(28) $$0 < \delta(n) \leq 2\Delta_n$$

(29) Hence the length of the last cell in $G[m(n)]$ also tends to zero.

(30) Since $x_{A(n)}(n)$ lies within the first cell of the grid, This follows from step 4. Thus, we select $x_{A(n)}(n)$ as the designated point of the first cell. Also $x_k(n)$ are all equidistant and the interval between them is equal to Δ_n which is the length of each of the $m(n) - 1$ cells of the grid.

(31) It follows that the point $x_k(n)$ are distributed so that one is in each of the $m(n)$ cells of the grid and namely that

$$a_{k-1}(n) < x_{A(n)+k-1}(n) \leq a_k(n) \quad k = 1,2,\ldots m(n)$$

(32) Choose the point $x_{A(n)+k-1}(n)$ as the designated point of the cell $(a_{k-1}(n), a_k(n)$ and build a Riemann sum call it $P_2'(t_1, t_2, n)$ then

(33) $$\lim_n P_2'(t_1, t_2, n) = \lim \sum_{k=A(n)}^{B(n)-1} \frac{\frac{1}{\sqrt{np(1-p)}}}{\sqrt{2\pi}} e^{-\frac{1}{2}x_k^2(n)} + \delta(n) \frac{e^{-\frac{1}{2}B^2(n)}}{\sqrt{2\pi}}$$

Since the maximum length of the cell of the grid underlying the Riemann sum $P_2'(t_1, t_2, n)$ tends to zero, it follows that

(34)
$$\lim_n P_2'(t_1, t_2, n) = \int_{t_1}^{t_2} \frac{e^{-\frac{1}{2}x^2}}{\sqrt{2\pi}}\, dx$$

Now, comparing $P_2'(t_1, t_2, n)$ with $P_2(t_1, t_2, n)$ we find that

(35)
$$\left| P_2(t_1, t_2, n) - P_2'(t_1, t_2, n) \right| =$$

$$\left| \sum_{k=A(n)}^{B(n)} \frac{\frac{1}{\sqrt{np(1-p)}}}{\sqrt{2\pi}} e^{-\frac{1}{2}x_k^2(n)} - \sum_{k=A(n)}^{B(n)-1} \frac{e^{-\frac{1}{2}x^2}}{\sqrt{2\pi}} + \delta_n \frac{e^{-B(n)^2}}{\sqrt{2\pi}} \right|$$

$$= \left| \frac{1}{\sqrt{np(1-p)}} - \delta_n \right| \frac{e^{-\frac{1}{2}\beta^2}}{\sqrt{2\pi}} \leq \frac{1}{\sqrt{2\pi np(1-p)}}$$

(36) Thus, as n gets large, the difference between $P_2(t_1, t_2, n)$ and $P_2'(t_1,t_2,n)$ tends to zero. Therefore if $P_2'(t_1,t_2,n)$ tends to the integral in step 34 then $P_2(t_1,t_2, n)$ must tend to the same integral also.

(37) Therefore:

$$P_2(t_1, t_2, n) = \int_{t_1}^{t_2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}\, dx$$

Q. E. D.

Chapter IV

APPLICATIONS OF THE NORMAL APPROXIMATION TO BINOMIAL

Problems related to the binomial distribution are fairly easy to solve provided the number of trials, n, is not large. If n is large, the computations involved become exceedingly lengthy; consequently, a good simple approximation to the distribution should prove to be beneficial and useful. Such an approximation exists in the form of the proper normal distribution. For the purpose of investigating this approximation, consider some numerical examples.

Let $n = 12$ and $p = \frac{1}{3}$ and construct the graph of the corresponding binomial distribution. By use of the formula

$$P(x) = \frac{n!}{x!(n-x)!} p^x q^{n-x}$$

the P(x) were computed, correct to three decimals, as

$$P(0) = .008 \qquad P(6) = .111$$
$$P(1) = .046 \qquad P(7) = .048$$
$$P(2) = .127 \qquad P(8) = .015$$
$$P(3) = .212 \qquad P(9) = .003$$
$$P(4) = .238 \qquad P(10) = .000$$
$$P(5) = .191 \qquad P(11) = .000$$
$$P(12) = .000$$

Although the graph used earlier for a binomial distribution was a line graph because of the discrete character of the variable x, this distribution will be graphed as a histogram in order to compare it more readily with normal distribution histograms. The graph of the histogram. The graph of the histogram for this distribution is shown in Figure 7. The height of any rectangle is

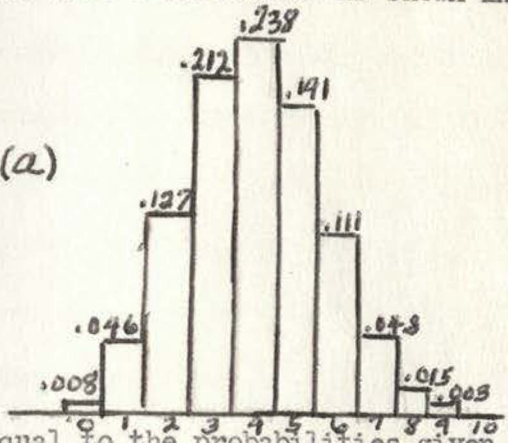
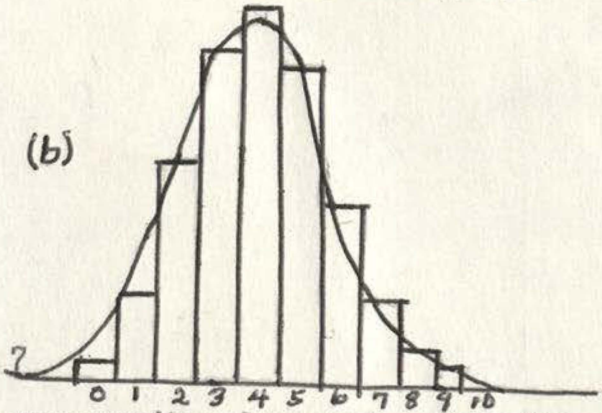


Figure 7



equal to the probabilities given for the corresponding class mark. Since the base length of any rectangle is 1, the area of any rectangle is equal to its height, and therefore these probabilities are also given by the areas of the corresponding rectangles. It therefore appears that this histogram could be fitted fairly well by the normal curve.

Since the normal distribution curve is completely determined by its mean and standard deviation, the natural normal curve to be used here is the one with the same mean and standard deviation as the binomial distribution. Then from the formulas $\mu = np$ and $\sigma = \sqrt{npq}$, it follows tha $= 12\cdot\frac{1}{3} = 4$ and $\sigma = \sqrt{12\cdot\frac{1}{3}\cdot\frac{2}{3}} = 1.63$. A normal curve with this mean and standard deviation was superimposed on Figure 7 (a) to give Figure 6 (a). It appears that the fit is fairly good in spite of the fact that $n = 12$ is small n and advanced theory promises a good fit only for large n.

As a test of the accuracy of the normal curve approximation here and as an illustration of how to use normal curve methods for approximating binomial probabilities, consider these problems related to Figure 7.

A) If the probability that a marksman will hit a target is $\frac{1}{3}$ and if he takes 12 shots, what is the probability that he will score at least 6 hits?

The exact answer, correct to three decimals, is obtained by adding the values in the listings of $P(x)$ for those given from $x = 6$ to $x = 12$, which is found to be .177. Geometrically, this answer is the area of that part of the histogram in Figure 7 lying to the right of $x = 5.5$. Therefore, the approximation of this probability by normal curve methods is merely necessary to find the area under that part of the fitted normal curve lying to the right of 5.5. Since the fitted normal curve has $\mu = 4$ and $\sigma = 1.63$, it follows that

$$\frac{5.5 - 4}{1.63} = 0.92.$$

Now, from the table, the area to the right of = 0.92 is .179; therefore this is the desired approximation to the probability of getting at least 6 hits. Since the exact answer was just computed to be .177, the normal curve approximation here is certainly good.

(B) To test the accuracy of normal curve methods over a shorter interval, calculate the probability that the marksmann will score precisely six hits in twelve shots. The answer, correct to three decimals, is .111. Since this is equal to the area of the rectangle whose base runs from 5.5 to 6.5, to approximate this answer it is necessary to find the area under the fitted normal curve between $x = 5.5$ and $x = 6.5$. Thus by calculating the z value and using the table, one obtains

$$z_2 = \frac{6.5 - 4}{1.63} = 1.53, \qquad A_2 = .4370$$

$$z_1 = \frac{5.5 - 4}{1.63} = 0.92 \qquad A_1 = .3212.$$

Subtracting these two areas gives .116, which compared to the exact probability of .111, is also good. From these two examples it appears that normal curve methods give good approximation even for some situations, such as the one considered here, in which n is not very large.

As an illustration of the use of the normal curve approximation to the binomial distribution when n is large, consider the following problem. To do so, the proper normal curve is now the one with mean and standard deviation given by the formulas

$$\mu = p \qquad \text{and } \sigma = \sqrt{\frac{pq}{n}}.$$

Problem:

Suppose a politician claims that a survey in his district showed that 60 per cent of his constituents agreed with his vote on an important piece of legislation. It it is assumed temporarily that this percentage is correct and if an impartial sample of 400 voters is taken in his district, what is is the probability that the sample will yield less than 50 per cent in agreement?[3]

It is assumed that taking a sample of 400 voters is like playing a game of chance 400 times for which the probability of success in a single game is .6, this problem can be treated as a binomial distribution problem with p = .6 and n = 400. For such a large n the normal curve approximation will be excellent. Using the above formula

$$z = \frac{.5 - .6}{\sqrt{\frac{(.6)(.4)}{400}}} = -4.08$$

Now the sample proportion, x/n, will be less than .5 provided that x is less than -4.08. The probability that $z < -4.08$ is by symmetry equal to the probability that $z > 4.08$, which is considered too small to be worth listing in a table. Thus, if it should happen that less than 50 per cent of the sample

[3] Paul G. Hoel, Elementary Statistics(New York,: John Weley & Sons, Inc., 1965) p. 108.

favored the politician, his claim of 60 per cent backing would certainly be discredited.

Objections may be raised, and rightfully so, the getting of a sample of 400 voters is not equivalent to playing a game of chance 400 times. There is a question concerning the independence of trials and the constancy of the probability that must be answered before one can be thoroughly happy with the binomial distribution model for this problem. Considerations such as these gave rise to the empirical rule for a good approximation.

Now just suppose that we attempted to use the formula for the binomial function to get the solution to this problem. We would have

$$P(X = 1) = \binom{400}{1}\left(\frac{3}{5}\right)\left(\frac{2}{5}\right)^{399}$$

which seems highly improbable because who could conceive of a person raising a number to the 399th power. Hence, it certainly understandable why for large n it is more practical to use the standard normal function.

Chapter V

CONCLUSION

The De Moivre-Laplace Theorem is one form of a quite general set of "central limit theorems." These theorems treat the limiting distributions of sums of random variables, and these limiting distributions are ordinarily normal. The value of the theorems is that they enable us to compute approximate probabilities for sums using the normal distribution without ever knowing the exact distribution of the sum. Exact distributions are often hard to get so we are grateful for such approximations.

In this paper the writer has illustrated the approach to normality of a sequence of adjusted binomial distributions. Since the number of successes, X, in a binomial experiment is an example of a random variable which is itself the sum of several independent random variables, we have also illustrated the more general idea that sums of independent random variables, suitably adjusted, tend to be normally distributed, under quite general conditions.

BIBLIOGRAPHY

Hoel, Paul G., Elementary Statistics, John Wiley & Sons, Inc., New York, 1966.

Kenney, J. F., Mathematics of Statistics, D. Van Nostrand Company, Inc., New York, 1951.

Lindgren, B. W., Introduction to Probability and Statistics, The Macmillan Company, 1969.

Meyer, Paul L., Introductory Probability and Statistical Applications, Addison-Wesley Publishing Company, Inc., Reading, Mass., 1965.

Mosteller, Frederick, Probability with Statistical Applications, Addison-Wesley Publishing Company, Inc., Reading, Mass., 1965.

Neyman, J., First Course in Probability and Statistics, Henry Holt and Company, New York, 1950.

Parzen, Emanuel, Modern Probability Theory and Its Applications, John Wiley & Sons, Inc., New York, 1960.