CHAPTER 6. NORMAL APPROXIMATION.

If we toss a coin 1000 times, take heads to be success, and ask how many successes occur, we have a binomial experiment. We can then use the binomial formula to get the probability of exactly 450 successes as

$$\binom{1000}{450}\left(\frac{1}{2}\right)^{1000}.$$

Suppose now that we want the probability of getting between 450 and 550 successes (inclusive). This will be given by the expression:

$$\binom{1000}{450}\left(\frac{1}{2}\right)^{1000} + \binom{1000}{451}\left(\frac{1}{2}\right)^{1000} + \ldots + \binom{1000}{549}\left(\frac{1}{2}\right)^{1000} + \binom{1000}{550}\left(\frac{1}{2}\right)^{1000}.$$

Without a computer or programmable calculator, it would be difficult for us to evaluate this expression directly. There is a fundamental fact of probability theory, however, which enables us to get a numerical answer easily and quickly.

Let us return to Figures 5.2 and 5.3 (in Chapter 5) for the binomial distribution $b(x;n,p)$. Recall that, as n increases, the graph of Figure 5.2 flattens, with the height of its highest bar proportional to $\frac{1}{\sqrt{n}}$, while the graph of Figure 5.3 becomes sharper, with the height of its highest bar proportional to $\sqrt{n}$. This leads us to look for some intermediate kind of graph where, by choosing the horizontal scale properly, we can make the height of the highest bar remain nearly constant as n increases, while the probability of each bar continues to be given by its

total area. For example, if we take a horizontal scale so that individual bars have width $= \frac{1}{\sqrt{n}}$, we can expect the height of the highest bar to remain nearly constant, since the area of the highest bar is nearly proportional to $\frac{1}{\sqrt{n}}$ (and approximately equal to $\frac{1}{\sqrt{2\pi pqn}}$, see Chapter 5), while the width of the graph as a whole will range from 0 to $\sqrt{n}$ along its horizontal axis.

It turns out that we get our most useful result if we take the horizontal scale so that individual bars have width $\frac{a}{\sqrt{n}}$ where $a = \frac{1}{\sqrt{pq}}$ (that is to say, each bar has width $1/\sqrt{npq}$), and if we then move the origin of the horizontal scale to $x = np$, the approximate position of the highest bar. We call this a type C graph. The horizontal scale is then given by $z = \frac{x-np}{\sqrt{npq}}$. We find that, as n increases, the over-all shape of the graph is more and more closely given by a certain fixed bell-shaped curve. The curve is called the standard normal curve or standard Gaussian curve. This fact is called the approximation to the binomial distribution by the normal curve, or, more briefly, the normal approximation. (It is also sometimes called the DeMoivre-Laplace limit theorem. It is a special case of a more general result to be given in Chapter 16, the central limit theorem.) We can use a table of values for areas under the normal curve to solve problems about binomial distributions (just as we can use trigonometric tables to solve problems about triangles).

The standard normal curve is given by the equation $y = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$. The curve is symmetric about the y-axis, and has the shape indicated in Figure 6.1.

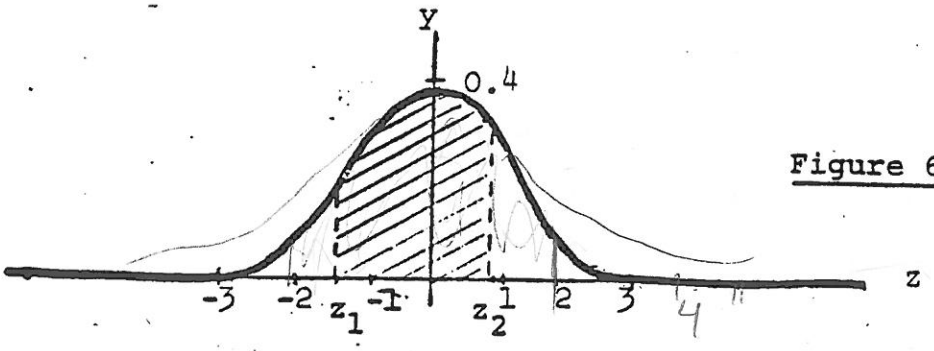


Figure 6.1.

The curve is never zero, but outside the interval $-3 \leq z \leq 3$, $y < 0.005$, and outside the interval $-5 \leq z \leq 5$, $y < 0.000002$. For $z = 1$, $y \approx 0.24$, and for $z = 2$, $y \approx 0.05$. The total area under the curve is exactly 1 (because one can show, by calculus, that $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = 1$.)

If z_1 and z_2 are two values of z, then $\underline{\text{Normal Area}}_{z_1}^{z_2}$ will stand for the numerical area of the region shaded in the figure. (In the notation of calculus, $\underline{\text{Normal Area}}_{z_1}^{z_2} = \int_{z_1}^{z_2} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$.) If one is given values of z_1 and z_2, one can find a value for $\underline{\text{Normal Area}}_{z_1}^{z_2}$ from published tables or on a programmable calculator. Usually, such tables give $A(z) = \underline{\text{Normal Area}}_0^z$ where z is positive. Hence, for example, if we want $\underline{\text{Normal Area}}_{z_1}^{z_2}$ with z_1 negative and z_2 positive, we use the symmetry of the curve and take $\underline{\text{Normal Area}}_0^{-z_1} + \underline{\text{Normal Area}}_0^{z_2} = A(-z_1) + A(z_2)$. Sometimes, instead of $A(z)$, tables give the quantity $\Phi(z)$ which measures the entire area to the left of z under the curve. Then $\underline{\text{Normal Area}}_{z_1}^{z_2} = \Phi(z_2) - \Phi(z_1)$. Of course, for $z \geq 0$,

$\Phi(z) = A(z) + \frac{1}{2}$, and for $z < 0$, $\Phi(z) = \frac{1}{2} - A(-z)$. A table for $A(z)$ is given at the end of this chapter.

We can also use the symbol "∞" and write $\underline{\text{Normal Area}}_{-\infty}^{\infty} = 1$ for the total area under the curve. Then $A(\infty) = 1/2$, and $\Phi(z) = \underline{\text{Normal Area}}_{-\infty}^{z}$.

We use this curve in the following way. Suppose that we have a binomial experiment with n trials and with probability of success p. We wish to get the probability that the number of successes falls between x_1 and x_2 (inclusive). (In the example just above, $n = 1000$, $p = \frac{1}{2}$, $x_1 = 450$, and $x_2 = 550$.) With a bar graph of type A (see Chapter 5), this probability would be the total area of the bars lying between $x_1 - \frac{1}{2}$ and $x_2 + \frac{1}{2}$. Hence, to apply the normal approximation (using a graph of type C), we first find

$$z_1 = \frac{x_1 - \frac{1}{2} - n\,p}{\sqrt{npq}}$$

and

$$z_2 = \frac{x_2 + \frac{1}{2} - n\,p}{\sqrt{npq}}.$$

The value of $\underline{\text{Normal Area}}_{z_1}^{z_2}$ then gives us the desired answer.

Applying this to the example above, we get

$$z_1 = \frac{450 - \frac{1}{2} - 500}{\sqrt{250}} = -3.19$$

$$z_2 = \frac{550 + \frac{1}{2} - 500}{\sqrt{250}} = 3.19$$

From a table for $A(z)$, we get

$$\underline{\text{Normal Area}}\ {}^{3.19}_{-3.19} = 2 \cdot A(3.19) = 2(0.4993) = 0.9986.$$

We thus see that if we toss a coin 1000 times, it is virtually certain that the number of heads will be between 450 and 550.

For a second example, take the binomial experiment of tossing a coin 100 times. What is the probability that we get between 45 and 55 heads (inclusive)? Using the normal approximation, we get

$$z_1 = \frac{45 - \frac{1}{2} - 50}{\sqrt{25}} = -1.1,$$

$$z_2 = \frac{55 + \frac{1}{2} - 50}{\sqrt{25}} = 1.1 .$$

Therefore the probability is given by $\underline{\text{Normal Area}}\ {}^{1.1}_{-1.1}$. From tables we find this to be $2 \cdot A(1.1) = 2(0.3643) = 0.729$.

Notice how these two examples agree with the observed weak stability of relative frequencies described in Chapter 1. The second example tells us that in tossing a coin 100 times, there is a 0.73 chance that the observed relative frequency of heads will lie between 0.45 and 0.55; and the first example tells us that in tossing a coin 1000 times, the observed relative frequency is almost certain to lie between 0.45 and 0.55. We further discuss the agreement of normal approximation with weak stability at the end of this chapter.

How accurate is the normal approximation? In both of the above examples, the approximate value turns out to be exact to the number of significant figures shown. As a rule, the approximation will be accurate to two decimal places whenever n is large enough to make both $\frac{np}{q}$ and $\frac{nq}{p} \geq 9$. (As before $q = 1-p$.)

Thus for coin tossing, we get two decimal place accuracy when $n \geq 9$, and for rolling a die to get sixes, we get two decimal place accuracy when $n \geq 45$.

As a further example, consider the following. A die is rolled 50 times. What is the probability of getting at least 5 sixes? We have $n = 50$, $p = 1/6$, and $q = 5/6$. Hence $np = 50/6 = 8.33$. We can use normal approximation with two decimal place accuracy since $\frac{np}{q} = 10 \geq 9$ and $\frac{nq}{p} = 250 \geq 9$. We seek $\underline{\text{Normal Area}}_{z_1}^{z_2}$ where

$$z_1 = \frac{5 - \frac{1}{2} - 8.33}{\sqrt{50\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)}} = -1.45 .$$

We could take $z_2 = \dfrac{50 + \frac{1}{2} - 8.33}{\sqrt{50\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)}} = 16.00$, but since our event puts no upper limit on x, we take $z_2 = \infty$. (The result is the same, since $A(16) = A(\infty) = 0.5$ to many decimal places.) We now have, for our result,

$$\underline{\text{Normal Area}}_{-1.45}^{\infty} = A(1.45) + A(\infty) = 0.43 + 0.5 = 0.93,$$

as the probability of getting at least 5 sixes in 50 rolls of one die.

For still another example, consider the following. A coin is to be tossed n times. How large should n be so that with probability 0.95, the observed relative frequency of heads will lie between 0.49 and 0.51? We first observe from the table of normal area values that $\underline{\text{Normal Area}}_{-z}^{z} = 0.95$ when $z = 1.96$.

Hence, using symmetry, we seek an n such that

$$\frac{x + \frac{1}{2} - np}{\sqrt{npq}} = 1.96$$

where $p = q = 1/2$, and $\frac{x}{n} = 0.51$. Substituting for x, p, and q, we get

$$\frac{0.51n + 0.5 - 0.5n}{\frac{1}{2}\sqrt{n}} = 1.96 \ .$$

Squaring, and then solving the resulting quadratic equation, we get $n = 9504$. (Note that setting $\frac{x - \frac{1}{2} - np}{\sqrt{npq}} = -1.96$ with $\frac{x}{n} = 0.49$ gives the same quadratic equation.)

The reader will find it helpful to <u>memorize</u> normal areas for certain simple values of z_1 and z_2 (just as it is helpful in geometry and calculus to memorize values of trigonometric functions for certain simple angles). In particular,

$$\underline{\text{Normal Area}}\ {}_{-1}^{1} = .683 = \text{approximately } \frac{2}{3} \ ;$$

$$\underline{\text{Normal Area}}\ {}_{-2}^{2} = .954 = \text{approximately } \frac{19}{20} \ ;$$

and

$$\underline{\text{Normal Area}}\ {}_{-3}^{3} = .997 = \text{approximately } \frac{299}{300} \ .$$

We conclude by observing that the normal approximation can be used to find values for the binomial formula itself. For example, if we wish to find a value for

$$b(12;25,0.4) = \binom{25}{12}(0.4)^{12}(0.6)^{13} ,$$

we can proceed as follows. We are looking for the area of one particular bar in the graph of the binomial distribution for $n = 25$ and $p = 0.4$. This will be the area lying between $12 - \frac{1}{2}$ and $12 + \frac{1}{2}$. Thus we want $\underline{\text{Normal Area}}_{z_1}^{z_2}$ where

$$z_1 = \frac{12 - \frac{1}{2} - 10}{\sqrt{25(0.4)(0.6)}} = 0.612 ,$$

and

$$z_2 = \frac{12 + \frac{1}{2} - 10}{\sqrt{25(0.4)(0.6)}} = 1.020 .$$

Using tables we obtain $\underline{\text{Normal Area}}_{.612}^{1.020} = 0.116$. (The exact value, of $b(12;25,0.4)$, to three figures, is 0.114.)

It is common to abbreviate the function $\frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$ as $\varphi(z)$. The area of a single bar for $b(x;n,p)$ (in a type C graph) can also be found as $\varphi(z) \cdot d$, where $d = \frac{1}{\sqrt{npq}}$ is the width of the bar, $\varphi(z)$ is the height of the bar as approximated by the standard normal curve, and $z = \frac{x-np}{\sqrt{npq}}$ is the horizontal value at the center of the bar. Values of φ are easily obtained on an electronic calculator. A table for $\varphi(z)$ is given at the end of this chapter. For the example of $b(12;25,0.4)$ just above, we get

$z = \dfrac{12 - 10}{\sqrt{25(0.4)(0.6)}} = 0.816$, and from tables for φ, we get $\varphi(0.816) = 0.2855$. As $d = \dfrac{1}{\sqrt{25(0.4)(0.8)}} = 0.408$, we have $\varphi(\alpha) \cdot d = (0.2855)(0.408) = 0.116$, the same approximate value as before.

Note on proof. Normal approximation can be proved by taking the binomial formula, putting Stirling's formula in place of the factorials in the binomial coefficient and simplifying. The proof is straightforward, elementary, and similar to the proof given in Chapter 5 that the height of the highest bar in a type A graph is approximately $\dfrac{1}{\sqrt{2\pi pqn}}$. We do not give details here.

Further illustration of normal approximation.

The following figures give graphs for three binomial distributions. The first figure gives the distribution for $n = 4$ and $p = 1/3$, the second for $n = 9$ and $p = 1/3$, and the third for $n = 18$ and $p = 1/3$. In each case, graphs of type A, type B, and type C are given. On each type C graph, the standard normal curve is also drawn. The type A graphs illustrate how the form of the distribution itself changes with increasing n. The type B graphs, where the major portion of the area is confined to a

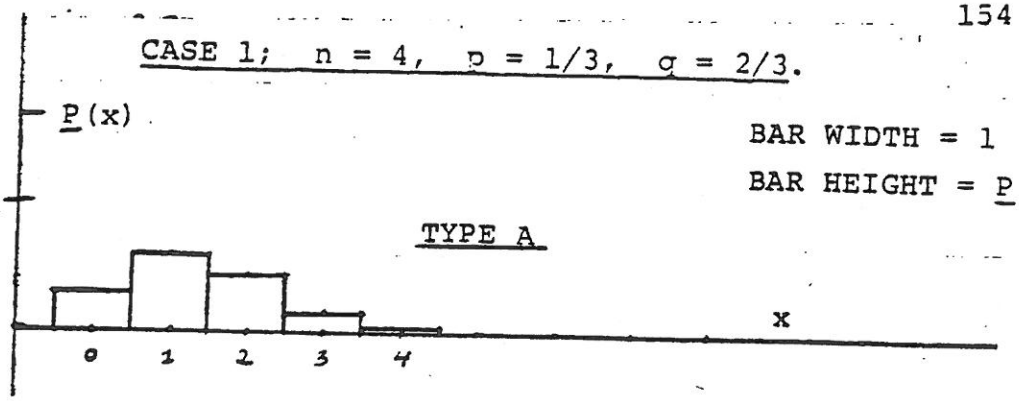
smaller and smaller part of the interval from 0 to 1 around the value 1/3 (= p), show how the calculated distributions agree with the observed stability of relative frequencies described in Chapter 1. The type C graphs, where the silhouette of the graph quickly approaches the standard normal curve, illustrate normal approximation.

In each case, tables are given showing the calculated values from which the graphs are made. Note that with <u>every</u> graph, the probability of an observation falling in a certain interval of x values is given by the total area of the corresponding bars in the graph. Indeed, in each graph, the width of the bars is defined ahead of time (1 for type A, 1/n for type B, and $1/\sqrt{npq}$ for type C), and then the height of the bars is adjusted to make bar area = probability (and hence the total graphical area = 1).

Normal approximation simply says that the areas for the type C graph can be approximated by using areas under the standard normal curve, and that this approximation becomes more and more accurate as n increases. The term 1/2 occurs in the calculation of end-points for normal approximation because we are really approximating the total area of certain bars in the type C graph, and one-half of a bar at each end will be outside the interval of z values that corresponds to the interval of x values for which we seek a probability.

For case 2 (n = 9) and case 3 (n = 18), the bars in the type B graphs have been drawn as single lines because of limited space.

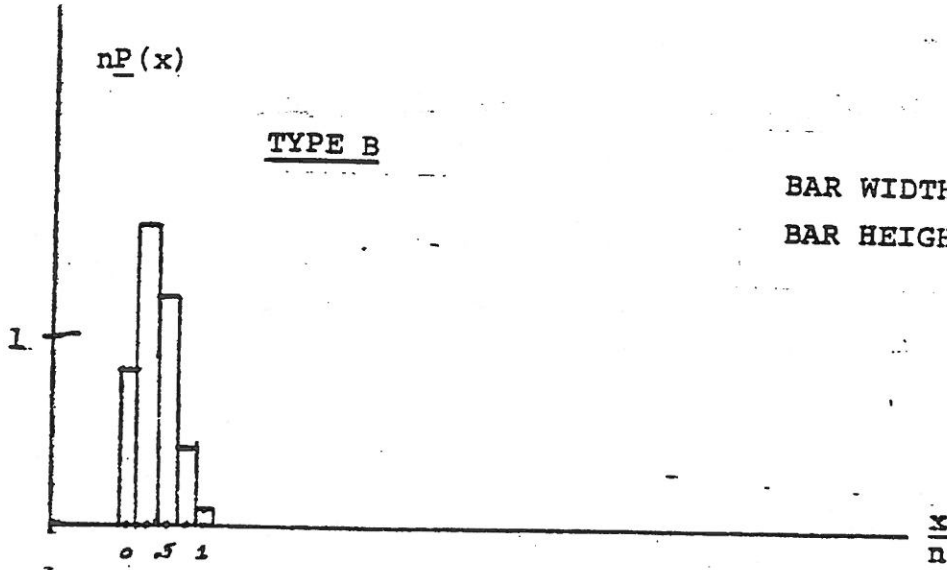
CASE 1; $n = 4$, $p = 1/3$, $q = 2/3$.



BAR WIDTH = 1

BAR HEIGHT = $\underline{P}(x) = \binom{n}{x} p^x q^{n-x}$

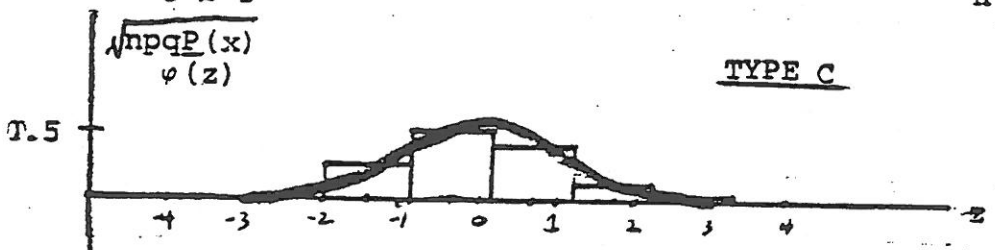
$= \binom{4}{x} \left(\frac{1}{3}\right)^x \left(\frac{2}{3}\right)^{4-x}$



BAR WIDTH = .25

BAR HEIGHT = $n \cdot \underline{P}(x) = 4\underline{P}(x)$

$= \dfrac{1}{\sqrt{4 \cdot \frac{1}{3} \cdot \frac{2}{3}}} = 1.06$



BAR WIDTH = $\dfrac{1}{\sqrt{npq}}$

BAR HEIGHT = $\sqrt{npq} \cdot \underline{P}(x)$

$= 0.94\ \underline{P}(x)$

TABLE

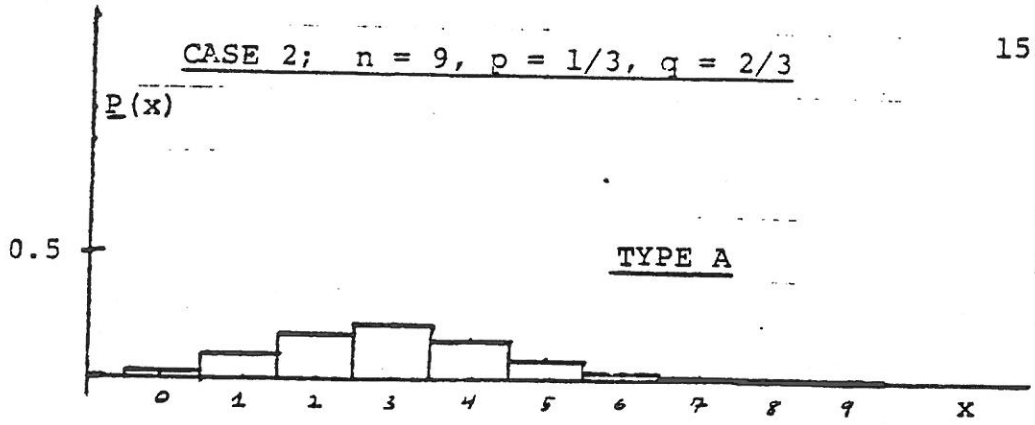
	A		B		C	
x	P(x)	x/n	n P(x)	z	√npq P(x)	φ(z)
0	.20	0	.80	-1.41	.19	.15
1	.40	.25	1.60	-.35	.38	.38
2	.30	.50	1.20	.71	.28	.31
3	.10	.75	.40	1.77	.09	.08
4	.01	1.0	.04	2.83	.01	.01

$z = \dfrac{x - np}{\sqrt{npq}} = 1.06\left(x - \frac{4}{3}\right)$

$\varphi(z) = \dfrac{1}{\sqrt{2\pi}} e^{-z^2/2}$

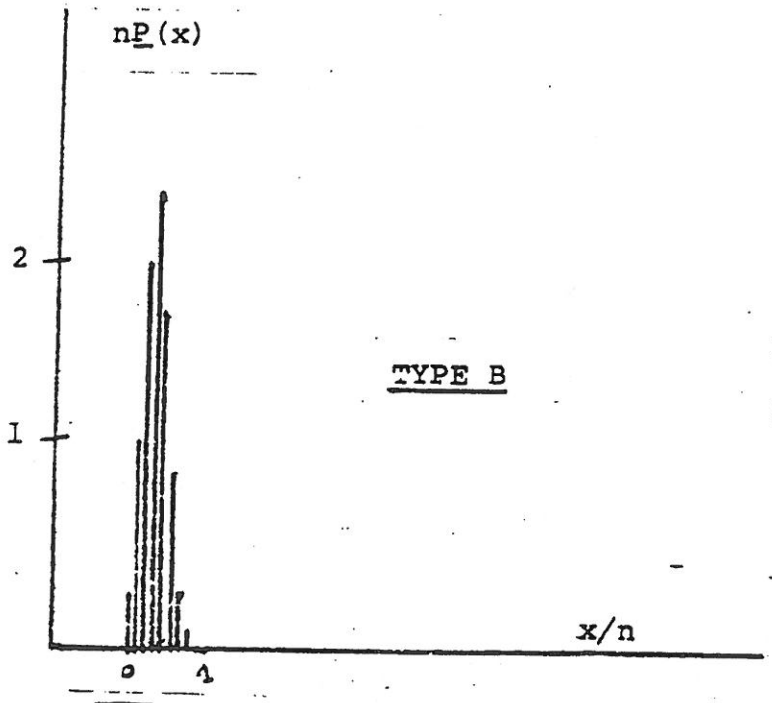
Figure 6.2

CASE 2; n = 9, p = 1/3, q = 2/3



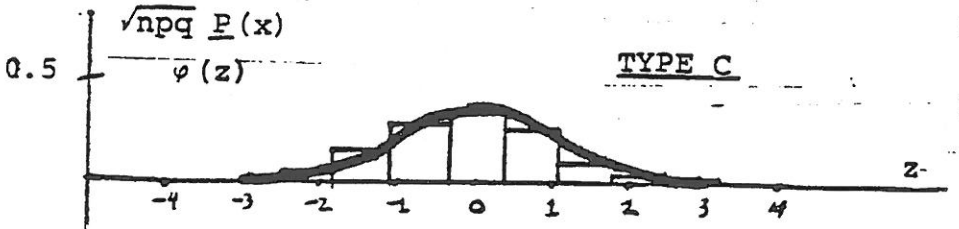
BAR WIDTH = 1

BAR HEIGHT = $\underline{P}(x) = \binom{9}{x}\left(\frac{1}{3}\right)^x\left(\frac{2}{3}\right)^{9-x}$



BAR WIDTH = $\frac{1}{n} = .11$

BAR HEIGHT = $n \cdot \underline{P}(x) = 9\underline{P}(x)$



BAR WIDTH = $\frac{1}{\sqrt{npq}} = 0.71$

BAR HEIGHT = $\sqrt{npq} \cdot \underline{P}(x)$

$= 1.41\ \underline{P}(x)$

$z = \frac{x-np}{\sqrt{npq}} = 0.71(x-3)$

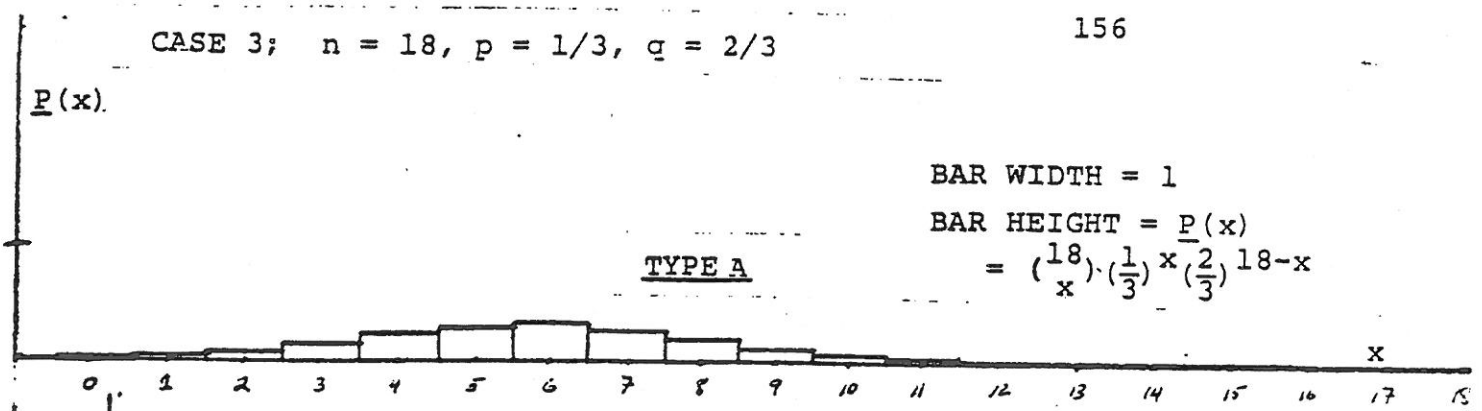
$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$

TABLE

	A	B		C		
x	P(x)	x/n	n·P(x)	z	√npq·P(x)	φ(z)
0	.03	0	.27	-2.13	.04	.04
1	.12	.11	1.08	-1.42	.17	.15
2	.23	.22	2.07	-.71	.32	.31
3	.27	.33	2.43	0	.38	.40
4	.20	.44	1.80	.71	.28	.31
5	.10	.55	.90	1.42	.14	.15
6	.03	.66	.27	2.13	.04	.04
7	.01	.77	.09	2.84	.01	.01
8	.00	.88	.00	--	.00	.00
9	.00	.99	.00	--	.00	.00

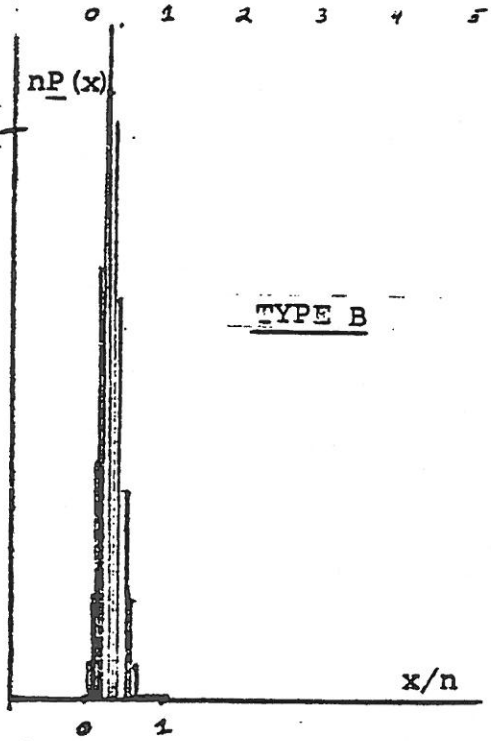
Figure 6.3

CASE 3; $n = 18$, $p = 1/3$, $q = 2/3$



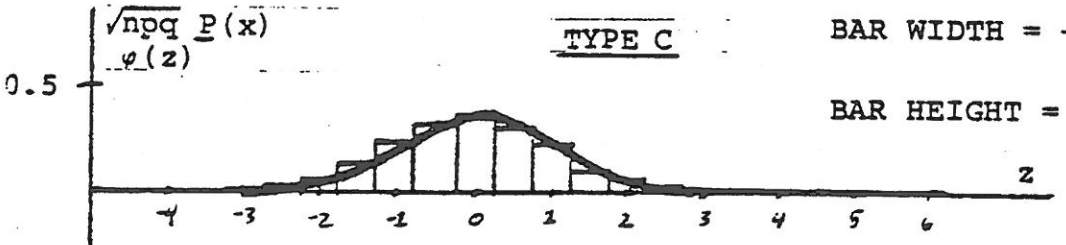
BAR WIDTH = 1

BAR HEIGHT = $\underline{P}(x) = \binom{18}{x}\left(\frac{1}{3}\right)^{x}\left(\frac{2}{3}\right)^{18-x}$



BAR WIDTH = $\frac{1}{n} = 0.055$

BAR HEIGHT = $n \cdot \underline{P}(x) = 18\underline{P}(x)$



BAR WIDTH = $\frac{1}{\sqrt{npq}} = 0.50$

BAR HEIGHT = $\sqrt{npq} \cdot \underline{P}(x) = 2\underline{P}(x)$

$$z = \frac{x-np}{\sqrt{npq}} = 0.5(x-6)$$

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

TABLE

	A		B		C	
x	$\underline{P}(x)$	x/n	$n \cdot \underline{P}(x)$	z	$\sqrt{npq} \cdot \underline{P}(x)$	$\varphi(z)$
0	.00	0	0	-3.00	.00	.00
1	.01	.055	.18	-2.50	.02	.02
2	.03	.11	.54	-2.00	.06	.05
3	.07	.165	1.26	-1.50	.14	.13
4	.13	.22	2.34	-1.00	.26	.24
5	.18	.275	3.24	-.05	.36	.35
6	.20	.33	3.60	0	.40	.40
7	.17	.385	3.06	.50	.34	.35
8	.12	.44	2.16	1.00	.24	.24
9	.06	.495	1.08	1.50	.12	.13
10	.03	.55	.54	2.00	.06	.05
11	.01	.605	.18	2.50	.02	.02
12	.00	.66	.00	3.00	.00	

Figure 6.4

Note on stability of relative frequencies. Normal approximation gives us a theoretical verification of the approximate formula $2/\sqrt{n}$ stated in Chapter 1 for the weak stability of relative frequencies. If an event has probability p, then, in n independent trials, by normal approximation, the observed number of occurrences must have approximate probability 0.95 of lying between x_1 and x_2 (inclusive), if x_1 and x_2 are chosen so that, as nearly as possible,

$$\frac{x_1 - \frac{1}{2} - np}{\sqrt{npq}} = -1.96, \quad \text{and} \quad \frac{x_2 + \frac{1}{2} - np}{\sqrt{npq}} = 1.96 .$$

Dividing numerators and denominators by n to get relative frequencies, we have

$$\frac{\frac{x_1}{n} - p - \frac{1}{2n}}{\sqrt{\frac{pq}{n}}} = -1.96, \quad \text{and} \quad \frac{\frac{x_2}{n} - p + \frac{1}{2n}}{\sqrt{\frac{pq}{n}}} = 1.96 .$$

Therefore, $\frac{x_1}{n} - p = -1.96\sqrt{\frac{pq}{n}} + \frac{1}{2n}$

and $\frac{x_2}{n} - p = 1.96\sqrt{\frac{pq}{n}} - \frac{1}{2n}$.

Subtracting to find the length of the interval of relative frequencies between $\frac{x_1}{n}$ and $\frac{x_2}{n}$, we get

$$\frac{x_2}{n} - \frac{x_1}{n} = \frac{2(1.96)}{\sqrt{n}}\sqrt{pq} - \frac{1}{n} .$$

Now pq is maximum when $p = \frac{1}{2}$, and in that case $\sqrt{pq} = \frac{1}{2}$.

Hence $\frac{x_2}{n} - \frac{x_1}{n} = \frac{2(1.96)}{\sqrt{n}}\sqrt{pq} - \frac{1}{n} \leq \frac{1.96}{\sqrt{n}} - \frac{1}{n} < \frac{1.96}{\sqrt{n}} < \frac{2}{\sqrt{n}}$.

Thus, if we assume a model which assigns probability p to a given event, then the probability is at least 0.95 that an observed relative frequency of the event (in n trials) will lie in the interval centered at p whose total length is $\frac{2}{\sqrt{n}}$.

In fact, the above analysis gives us a better formula than $2/\sqrt{n}$ for the weak stability of relative frequencies in the case where we know (or can assume) a specific limiting value λ for the relative frequency. For we have, from above, probability 0.95 that

$$\frac{x_2}{n} - \frac{x_1}{n} \leq \frac{2(1.96)\sqrt{\lambda(1-\lambda)}}{\sqrt{n}} < 4\sqrt{\frac{\lambda(1-\lambda)}{n}}.$$

When $\lambda = 1/2$, this new improved formula, $4\sqrt{\frac{\lambda(1-\lambda)}{n}}$, gives the same interval as before, but when $\lambda \neq 1/2$, it gives a smaller interval centered at λ. For example, if we consider bets on a single number at roulette (as in Chapter 1), if we take $\lambda = 1/38$, and if we ask for a value of n such

that, for that n, the observed relative frequency of winning will nearly always lie within the distance $\frac{1}{36} - \frac{1}{38} = 0.0015$ of λ (this is the interval which gives an average net loss to a bettor making repeated bets of constant size--see Chapter 1), then we have

$$4\sqrt{\frac{\frac{1}{38}\cdot\frac{37}{38}}{n}} = 2(0.0015) = 0.003,$$

or $$n \approx 46{,}000.$$

If we ask the same question using the previous formula $2/\sqrt{n}$, we only get $\frac{2}{\sqrt{n}} = 0.003$, or $n \approx 450{,}000$.

Note. The above deduction of the weak stability law from normal approximation assumes that the conditions for normal approximation hold. In the case of the improved formula, this means that both $\frac{n\lambda}{1-\lambda}$ and $\frac{n(1-\lambda)}{\lambda}$ must be ≥ 9, or that $n \geq \frac{9(1-\lambda)}{\lambda} > \frac{9}{\lambda}$ and $n \geq \frac{9\lambda}{1-\lambda} > \frac{9}{1-\lambda}$. In the case of the cruder formula $2/\sqrt{n}$, it can be shown that it is enough to take $n \geq 9$.

Example. Consider the games of roulette and craps as described in Chapter 1 (see also Exercises 1-4 and 1-5). What is the probability that after 10,000 plays, a bettor is ahead at roulette, and what is the probability that after 10,000 plays, a bettor is ahead at craps? In the game of roulette,

the bettor will be ahead if x, the number of successes, is such that $xw - (1{,}000-x)\ell = x35 - (10{,}000-x)1 > 0$, which is the same as $x \geq 278$. Using normal approximation, we have

$$P(x \geq 278) \approx \underline{\text{Normal Area}}_{z_1}^{\infty},$$

where
$$z_1 = \frac{277.5 - 263.2}{100\sqrt{(\frac{1}{38})(\frac{37}{38})}} \approx 0.89.$$

Therefore $P(x \geq 278) \approx \frac{1}{2} - A(0.89) = \frac{1}{2} - 0.31 = 0.19$. Thus the bettor has almost one chance in five of being ahead after 10,000 plays of roulette.

In the game of craps, the bettor will be ahead if $xw - (10{,}000-x)\ell = x1 - (10{,}000-x)1 > 0$, which is the same as $x > 5000$. Using normal approximation, we have

$$P(x > 5000) \approx \underline{\text{Normal Area}}_{z_2}^{\infty},$$

where
$$z_2 = \frac{500.5 - 4929.3}{100\sqrt{(0.493)(0.507)}} = 1.42.$$

(Here we have used the result of Exercise 4-10f, that $P(\underline{\text{pass}})$ at craps $= 0.49293$.) Therefore $P(x > 5000) \approx \frac{1}{2} - A(1.42) = \frac{1}{2} - 0.42 = 0.08$. Thus the bettor has less than one chance in ten of being ahead after 10,000 plays of craps.

We see that the gambler at roulette has a better chance of being ahead at roulette after 10,000 plays than at craps,

even though the disadvantage, as measured by average loss per play, is greater for roulette than for craps. The explanation of this difference is that the bettor at roulette, while having a greater chance of being ahead, also has a greater chance of losing a substantially larger amount of money than at craps.

This same improved formula, $4\sqrt{\frac{\lambda(1-\lambda)}{n}}$, can also be used for the strong stability of relative frequencies in the range of normally observed values of n. Empirical evidence shows that the formula may thus be taken as an improved formula for describing the empirical facts of both weak and strong stability in situations where we know λ, the limiting value of the relative frequency. For example, if we wish to get an estimate of a value of n such that our roulette bettor's average winnings per play will become and <u>remain</u> negative beyond n, we can set $4\sqrt{\frac{(\frac{1}{38})(\frac{37}{38})}{n}} = 0.0015$ and get $n \approx 185{,}000$. As before with the formula $2/\sqrt{n}$, theoretical considerations show that the added factor $\sqrt{\frac{\log \log n}{2}}$ is necessary in the case of strong stability. This correction factor only makes a difference for extremely large values of n, lying beyond those normally observed.

<u>Conceptual and philosophical comment</u>. We have seen in this chapter that normal approximation gives a theoretical version of the empirically observed weak stability of relative frequencies. This theoretical version says the following (for the case $p = 1/2$). In a binomial experiment of n trials

($n \geq 9$), the probability is 0.95 that the observed relative frequency of success will fall in an interval of length $\frac{1.96}{\sqrt{n}}$ symmetrical about the value 1/2. Our earlier empirical statement of weak stability, as made in Chapter 1, in effect asserted: if we make many repeated over-all trials of a given binomial experiment, then nearly all values of the relative frequency observed in each over-all trial will fall in an interval of length $2/\sqrt{n}$ --where n is the number of individual Bernoulli trials within the binomial experiment. We see that the theoretical counterpart to "nearly all" is "with probability 0.95". Thus the theoretical statement makes an assertion about n trials of an experiment with success probability $p = 1/2$ in terms of a single trial of another experiment with success probability 0.95. The philosophically minded reader will sense here the possible danger of an infinite regress; that is to say, of a situation in which we appear to be theoretically explaining a concept (the concept of probability) in terms of the same concept itself.

The correct resolution of this apparent difficulty is straightforward. Our theory only makes statements about models, that is to say, about probability spaces. Thus all of its statements must be statements about probabilities of certain events in those spaces. A separate question, going beyond the theory, is the question of how we connect the concepts of the theory with observations in the physical world. We need such a connection in order: (i) to show that our theory implies some

empirical fact or law, or (ii) to decide whether a particular theoretical model is satisfactory or not satisfactory in describing and predicting observations in the physical world. The connection between theory and reality is made in a direct and simple way. If the totality of our observed information can be viewed as the result of a single trial of some large experiment, and if our theory ascribes a probability very close to 1 to some event (for a trial of that large experiment), then we say that our theory predicts that event. If we observe that that event does not in fact occur, then we conclude that our theory is unsatisfactory.

The question remains: how close is very close? This depends upon our purposes in using the theory and upon the practical circumstances in which we use it. As we shall see, a large part of the subject of mathematical statistics is concerned with this question. In probability theory and the physical sciences, as a rule of thumb, differences smaller than 0.01 or 0.001 are often considered to be very close for the purpose of casting doubt on a theoretical assumption.

For example, if we have a theoretical model which gives $p = 1/2$ to a certain binomial experiment, if we wish to decide whether the model is reasonable or not, and if the totality of our observed information tells us that in 1000 trials, exactly 425 successes have been observed, we see that an event does not occur (the event that the observed

relative frequency falls within the interval [0.426,0.574]) whose probability, by normal approximation, is greater than 0.999999. We therefore conclude that our theory is incorrect. We shall return to questions and problems of this kind when we take up the study of mathematical statistics in Chapter 9.

Remark on calculations. Normal approximation, as described in this chapter, always uses the correction for bar width (the constant 1/2 in the numerator of the expressions for z_1 and z_2 when $\underline{\text{Normal Area}}_{z_1}^{z_2}$ is being found). It is sometimes convenient, for quick calculation when n is sufficiently large, to omit this correction (see Exercises 6-4 and 6-5 below). The reader should make a general practice of using the correction in careful work, however, since the difference $x-np$ is often small (even when n is large) and the corresponding value of $A(z)$ can be significantly affected by omission of the correction term.

In Chapter 7, we shall study another quite different form of approximation to the binomial distribution. At that time, we shall further discuss the matter of when to use normal approximation and when not. For the moment, if two decimal place accuracy is sufficient, the reader should use normal approximation when (i) it is convenient to do so and (ii) the condition holds that both $\frac{np}{q}$ and $\frac{nq}{p}$ are ≥ 9.

NORMAL TABLE

z	$\phi(z)$	A(z)	z	$\phi(z)$	A(z)	z	$\phi(z)$	A(z)	z	$\phi(z)$	A(z)	z	$\phi(z)$	A(z)	z	$\phi(z)$	A(z)	z	$\phi(z)$	A(z)
.00	.39804	.00000	.45	.36053	.17364	.00	.26609	.31594	1.35	.16038	.41149	1.80	.07895	.46407	2.25	.03174	.48778			
.01	.39892	.00399	.46	.35880	.17724	.91	.26369	.31859	1.36	.15822	.41309	1.81	.07754	.46485	2.26	.03103	.48809			
.02	.39886	.00798	.47	.35723	.18082	.92	.26129	.32121	1.37	.15608	.41466	1.82	.07614	.46562	2.27	.03034	.48840			
.03	.39870	.01197	.48	.35553	.18439	.93	.25888	.32381	1.38	.15395	.41621	1.83	.07477	.46638	2.28	.02965	.48870			
.04	.39862	.01595	.49	.35381	.18793	.94	.25647	.32639	1.39	.15183	.41774	1.84	.07341	.46712	2.29	.02898	.48899			
.05	.39814	.01994	.50	.35207	.19146	.95	.25406	.32894	1.40	.14973	.41924	1.85	.07206	.46784	2.30	.02833	.48928			
.06	.39822	.02392	.51	.35029	.19497	.96	.25164	.33147	1.41	.14764	.42073	1.86	.07074	.46856	2.31	.02768	.48956			
.07	.39797	.02790	.52	.34849	.19847	.97	.24923	.33398	1.42	.14556	.42220	1.87	.06943	.46926	2.32	.02705	.48983			
.08	.39767	.03188	.53	.34667	.20194	.98	.24681	.33646	1.43	.14350	.42364	1.88	.06814	.46995	2.33	.02643	.49010			
.09	.39733	.03586	.54	.34482	.20540	.99	.24439	.33891	1.44	.14146	.42507	1.89	.06687	.47062	2.34	.02582	.49036			
.10	.39695	.03983	.55	.34294	.20884	1.00	.24197	.34134	1.45	.13943	.42647	1.90	.06562	.47128	2.35	.02522	.49061			
.11	.39654	.04380	.56	.34105	.21226	1.01	.23955	.34375	1.46	.13742	.42786	1.91	.06439	.47193	2.36	.02463	.49086			
.12	.39608	.04776	.57	.33912	.21566	1.02	.23713	.34614	1.47	.13542	.42922	1.92	.06316	.47257	2.37	.02406	.49111			
.13	.39559	.05172	.58	.33718	.21904	1.03	.23471	.34850	1.48	.13344	.43056	1.93	.06195	.47320	2.38	.02349	.49134			
.14	.39505	.05567	.59	.33521	.22240	1.04	.23230	.35083	1.49	.13147	.43189	1.94	.06077	.47381	2.39	.02294	.49158			
.15	.39448	.05962	.60	.33322	.22575	1.05	.22988	.35314	1.50	.12952	.43319	1.95	.05959	.47441	2.40	.02239	.49180			
.16	.39387	.06356	.61	.33121	.22907	1.06	.22747	.35543	1.51	.12758	.43448	1.96	.05844	.47500	2.41	.02186	.49202			
.17	.39322	.06749	.62	.32918	.23237	1.07	.22506	.35769	1.52	.12566	.43574	1.97	.05730	.47558	2.42	.02131	.49224			
.18	.39253	.07142	.63	.32713	.23565	1.08	.22265	.35993	1.53	.12376	.43699	1.98	.05618	.47615	2.43	.02083	.49245			
.19	.39181	.07535	.64	.32506	.23891	1.09	.22025	.36214	1.54	.12188	.43822	1.99	.05508	.47670	2.44	.02033	.49266			
.20	.39104	.07926	.65	.32297	.24215	1.10	.21785	.36433	1.55	.12001	.43943	2.00	.05399	.47725	2.45	.01984	.49286			
.21	.39024	.08317	.66	.32086	.24537	1.11	.21546	.36650	1.56	.11816	.44062	2.01	.05292	.47778	2.46	.01936	.49305			
.22	.38940	.08706	.67	.31874	.24857	1.12	.21307	.36864	1.57	.11632	.44179	2.02	.05186	.47831	2.47	.01889	.49324			
.23	.38853	.09095	.68	.31659	.25175	1.13	.21069	.37076	1.58	.11450	.44295	2.03	.05082	.47882	2.48	.01842	.49343			
.24	.38762	.09483	.69	.31443	.25490	1.14	.20831	.37286	1.59	.11270	.44408	2.04	.04980	.47932	2.49	.01797	.49361			
.25	.38667	.09871	.70	.31225	.25804	1.15	.20594	.37493	1.60	.11092	.44520	2.05	.04879	.47982	2.50	.01753	.49379			
.26	.38568	.10257	.71	.31006	.26115	1.16	.20357	.37698	1.61	.10915	.44630	2.06	.04780	.48030	2.51	.01709	.49396			
.27	.38466	.10642	.72	.30785	.26424	1.17	.20121	.37900	1.62	.10741	.44738	2.07	.04682	.48077	2.52	.01667	.49413			
.28	.38361	.11026	.73	.30563	.26730	1.18	.19886	.38100	1.63	.10567	.44845	2.08	.04586	.48124	2.53	.01625	.49430			
.29	.38251	.11409	.74	.30339	.27035	1.19	.19652	.38298	1.64	.10396	.44950	2.09	.04491	.48169	2.54	.01585	.49446			
.30	.38139	.11791	.75	.30114	.27337	1.20	.19419	.38493	1.65	.10226	.45053	2.10	.04398	.48214	2.55	.01545	.49461			
.31	.38023	.12172	.76	.29887	.27637	1.21	.19186	.38686	1.66	.10059	.45154	2.11	.04307	.48257	2.56	.01506	.49477			
.32	.37903	.12552	.77	.29659	.27935	1.22	.18954	.38877	1.67	.09893	.45254	2.12	.04217	.48300	2.57	.01468	.49492			
.33	.37780	.12930	.78	.29431	.28230	1.23	.18724	.39065	1.68	.09728	.45352	2.13	.04128	.48341	2.58	.01431	.49506			
.34	.37654	.13307	.79	.29200	.28524	1.24	.18494	.39251	1.69	.09566	.45449	2.14	.04041	.48382	2.59	.01394	.49520			
.35	.37524	.13683	.80	.28969	.28814	1.25	.18265	.39435	1.70	.09405	.45543	2.15	.03955	.48422	2.60	.01358	.49534			
.36	.37391	.14058	.81	.28737	.29103	1.26	.18037	.39617	1.71	.09246	.45637	2.16	.03871	.48461	2.61	.01323	.49547			
.37	.37255	.14431	.82	.28504	.29389	1.27	.17810	.39796	1.72	.09089	.45728	2.17	.03788	.48500	2.62	.01289	.49560			
.38	.37115	.14803	.83	.28269	.29673	1.28	.17585	.39973	1.73	.08933	.45818	2.18	.03705	.48537	2.63	.01256	.49573			
.39	.36973	.15173	.84	.28034	.29955	1.29	.17360	.40147	1.74	.08780	.45907	2.19	.03626	.48574	2.64	.01223	.49585			
.40	.36827	.15542	.85	.27798	.30234	1.30	.17137	.40320	1.75	.08628	.45994	2.20	.03547	.48610	2.65	.01191	.49598			
.41	.36678	.15910	.86	.27562	.30511	1.31	.16915	.40490	1.76	.08478	.46080	2.21	.03470	.48645	2.66	.01160	.49609			
.42	.36526	.16276	.87	.27324	.30785	1.32	.16694	.40658	1.77	.08329	.46164	2.22	.03394	.48679	2.67	.01130	.49621			
.43	.36371	.16640	.88	.27086	.31057	1.33	.16474	.40824	1.78	.08183	.46246	2.23	.03319	.48713	2.68	.01100	.49632			
.44	.36213	.17003	.89	.26848	.31327	1.34	.16256	.40988	1.79	.08038	.46327	2.24	.03216	.48745	2.69	.01071	.49643			

z	$\phi(z)$	A(z)	z	$\phi(z)$	A(z)	z	$\phi(z)$	A(z)
2.70	.01042	.49653	3.15	.00279	.49918	3.60	.00061	.49984
2.71	.01014	.49664	3.16	.00271	.49921	3.61	.00059	.49985
2.72	.00987	.49674	3.17	.00262	.49924	3.62	.00057	.49985
2.73	.00961	.49683	3.18	.00254	.49926	3.63	.00055	.49986
2.74	.00935	.49693	3.19	.00246	.49929	3.64	.00053	.49986
2.75	.00909	.49702	3.20	.00238	.49931	3.65	.00051	.49987
2.76	.00885	.49711	3.21	.00231	.49934	3.66	.00049	.49987
2.77	.00861	.49720	3.22	.00224	.49936	3.67	.00047	.49988
2.78	.00837	.49728	3.23	.00216	.49938	3.68	.00046	.49988
2.79	.00814	.49736	3.24	.00210	.49940	3.69	.00044	.49989
2.80	.00792	.49744	3.25	.00203	.49942	3.70	.00042	.49989
2.81	.00770	.49752	3.26	.00196	.49944	3.71	.00041	.49990
2.82	.00748	.49760	3.27	.00190	.49946	3.72	.00039	.49990
2.83	.00727	.49767	3.28	.00184	.49948	3.73	.00038	.49990
2.84	.00707	.49774	3.29	.00178	.49950	3.74	.00037	.49991
2.85	.00687	.49781	3.30	.00172	.49952	3.75	.00035	.49991
2.86	.00668	.49788	3.31	.00167	.49953	3.76	.00034	.49992
2.87	.00649	.49795	3.32	.00161	.49955	3.77	.00033	.49992
2.88	.00631	.49801	3.33	.00156	.49957	3.78	.00031	.49992
2.89	.00613	.49807	3.34	.00151	.49958	3.79	.00030	.49992
2.90	.00595	.49813	3.35	.00146	.49960	3.80	.00029	.49993
2.91	.00578	.49819	3.36	.00141	.49961	3.81	.00028	.49993
2.92	.00562	.49825	3.37	.00136	.49962	3.82	.00027	.49993
2.93	.00545	.49831	3.38	.00132	.49964	3.83	.00026	.49994
2.94	.00530	.49836	3.39	.00127	.49965	3.84	.00025	.49994
2.95	.00514	.49841	3.40	.00123	.49966	3.85	.00024	.49994
2.96	.00499	.49846	3.41	.00119	.49968	3.86	.00023	.49994
2.97	.00485	.49851	3.42	.00115	.49969	3.87	.00022	.49995
2.98	.00471	.49856	3.43	.00111	.49970	3.88	.00021	.49995
2.99	.00457	.49861	3.44	.00107	.49971	3.89	.00021	.49995
3.00	.00443	.49865	3.45	.00104	.49972	3.90	.00020	.49995
3.01	.00430	.49869	3.46	.00100	.49973	3.91	.00019	.49995
3.02	.00417	.49874	3.47	.00097	.49974	3.92	.00018	.49996
3.03	.00405	.49878	3.48	.00094	.49975	3.93	.00018	.49996
3.04	.00393	.49882	3.49	.00090	.49976	3.94	.00017	.49996
3.05	.00381	.49886	3.50	.00087	.49977	3.95	.00016	.49996
3.06	.00370	.49889	3.51	.00084	.49978	3.96	.00016	.49996
3.07	.00358	.49893	3.52	.00081	.49978	3.97	.00015	.49996
3.08	.00348	.49897	3.53	.00079	.49979	3.98	.00014	.49997
3.09	.00337	.49900	3.54	.00076	.49980	3.99	.00014	.49997
3.10	.00327	.49903	3.55	.00073	.49981			
3.11	.00317	.49906	3.56	.00071	.49981			
3.12	.00307	.49910	3.57	.00068	.49982			
3.13	.00298	.49913	3.58	.00066	.49983			
3.14	.00288	.49916	3.59	.00063	.49983			

EXERCISES FOR CHAPTER 6.

6-1. In a binomial experiment, a fair coin is tossed 100 times. Let x = the number of times that heads appears. Use normal approximation to estimate the following probabilities:

(a) $P(49 \leq x \leq 51)$;

(b) $P(55 \leq x < 65)$;

(c) $P(60 \leq x)$;

(d) $P(x = 50)$.

6-2. A fair coin is tossed 10 times and comes up heads x times. Use normal approximation to estimate the following probabilities:

(a) $P(4 \leq x \leq 6)$;

(b) $P(4 \leq x \leq 10)$;

(c) $P(x = 6)$.

6-3. A thumbtack has probability 0.6 of landing on its side when it is tossed. It is tossed 10,000 times and lands on its side x times. Use normal approximation to estimate the following probabilities:

(a) $P(5{,}950 \leq x \leq 6{,}050)$;

(b) $P(x \leq 5{,}900)$;

(c) $P(x = 6000)$.

6-4. A fair coin is tossed 5 times and comes up heads x times. For each of the following probabilities, find: (i) the exact value; (ii) the value given by

normal approximation; and (iii) the value given by a form of normal approximation in which the correction for bar width (the constant 1/2 in the numerator) is omitted.

(a) $P(2 \leq x \leq 3)$.

(b) $P(2 \leq x \leq 5)$.

6-5. A fair coin is tossed n times and comes up heads x times.

(a) For $n = 100$, find: (i) the value for $P(45 \leq x \leq 55)$ given by normal approximation, and (ii) the value given by a form of normal approximation in which the correction for bar width (the constant 1/2 in the numerator) is omitted.

(b) For $n = 1000$, do the same as (a) for $P(484 \leq x \leq 516)$.

(c) For $n = 10,000$, do the same as (a) for $P(4,950 \leq x \leq 5,050)$.

Note. In the following exercises normal approximation should be used whenever appropriate.

6-6. On the average, a certain basketball player succeeds in scoring on 80 percent of his free throws. What is the probability that he score on 85 or more of his next 100 throws? What is the probability that he score on exactly 80 of them? What added assumption about the player are you making?

6-7. A fair coin is tossed 1000 times.

(a) What is the most probable number of heads to appear?

(b) Find a value for the probability that exactly that number will in fact appear.

6-8. In a group of 300 families with 3 children each, we observe 495 boys. Assume that the probability that any child is a boy is 1/2 and that being a boy is independent from birth to birth. What is the probability of observing 495 or fewer boys? What is the probability of observing 420 or fewer boys?

6-9. A thumbtack is tossed and either falls on its side (S) or on its back (B). Assume $P(B) = 1/3$. Estimate the number of tosses necessary so that with probability 0.95, the observed relative frequency of B lies within 0.01 of $P(B)$.

6-10. Assume that the probability of a male birth is 0.514. What is the probability that there will be fewer boys than girls in 1000 births? How many births must be observed to make the probability of observing fewer boys than girls less than 0.05?

6-11. Neglecting the correction for bar width (the constant 1/2 in the numerator), use normal approximation to get a value for the number of times that a fair coin should be tossed in order to ensure, with probability 0.95, that the observed relative frequency of heads is within 0.001 of the value 0.5.

6-12. Assume that the probability of a car breaking down between Boston and Worcester on the Massachusetts Turnpike is 0.001 and that this probability is the same for all cars. What is the probability that there are more than 5 and fewer than 10 breakdowns among the next 20,000 cars that make the trip?

6-13. A biased coin is tossed 100 times. Let x = the number of times that heads appears.

(a) Assume that $p = 0.55$, where p is the probability of heads on a single toss. Find the smallest integer a such that $P(|x-55| \leq a) \geq 0.95$.

(b) Assume that p is unknown and that 55 heads are observed. For what values of p can we say that the event $|x-100p| \leq |55-100p|$ has probability no greater than 0.95?

6-14. Consider the game described in Exercise 1-6. (The bettor bets even money that a <u>six</u> will not occur in six rolls of one die.) What is the probability that a bettor, making bets of constant size, is ahead after 100 plays? After 1000 plays? After 10,000 plays?

6-15. A bettor plays a game in which $w = 0.894$, $\ell = 1$, and the true probability of winning is 1/2. His average loss per play, in the long run, will be $1/2(1) - \frac{1}{2}(0.894) = 0.053$. (This is the same as the average loss per play for bets (with $\ell = 1$) on a

single number at roulette.) What is the probability that a bettor is behind after 10,000 plays?

6-16. Show that $\int_{-\infty}^{\infty} \phi(z)dz = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = 1.$

(Hint. Consider $[\int_{-\infty}^{\infty} \phi(z)dz]^2$. Write this as $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \phi(x)\phi(y)dx\ dy$, express in polar coordinates, and evaluate.)