BASIC STATISTICS:

A USER ORIENTED APPROACH

(*Manuscript*)

Spyros MAKRIDAKIS
and
Robert L. WINKLER

CHAPTER 5

CHAPTER 5

WIDELY-USED PROBABILITY DISTRIBUTIONS

5.1 Introduction

Statistics deals with uncertainty, and it is important for you to understand and be able to deal with uncertainty. As we indicated in Chapter 4, probability can be thought of as the language of uncertainty. That is why some knowledge of probability is needed before we discuss how statistics can help us make inferences and decisions in the face of uncertainty. The material presented in Chapter 4 concerning probability and probability distributions should give you an idea of how probability can be used to measure and deal with uncertainty.

As you shall see in later chapters, many statistics problems from different areas of application can be treated with the same statistical procedures. For example, questions concerning the proportion of voters who favor a certain candidate are similar (from a statistical viewpoint) to questions about the proportion of income tax returns containing arithmetic errors or about the proportion of drivers who do not fasten their seat belts. In these and similar applications, a probability distribution known as the binomial distribution can be used to describe the uncertainty with which we are dealing.

In real life there are innumerable events or phenomena. Fortunately, a great number of these events or phenomena can be described by a few theoretical probability distributions. These distributions, such as the binomial distribution, are applicable in a wide variety of situations. Because of their wide use, it is worthwhile to devote some time to such distributions, and in this chapter we have chosen the three distributions which seem to describe or approximate the largest number of real-world phenomena or events.

The three distributions covered in this chapter are the binomial, Poisson, and normal distributions. We have already indicated some situations for which the binomial distribution is appropriate. The Poisson distribution is used in problems involving things like the arrival and servicing of cars at a gas station and the number of telephone calls beings handled at a switchboard. The third distribution to be discussed here, the normal distribution, is the most important of the three in terms of applicability in statistical problems. Many real-world phenomena, such as ages, weights, IQ scores, and exam grades, have distributions that can often be approximated closely by the normal curve.

Familiarity with the distributions discussed in this chapter will be very helpful when we switch from probability to statistics. The normal distribution is particularly important and will be encountered over and over again in the remainder of the book. A few other distributions that are very useful in statistics will be covered in later chapters as the need for them arises.

5.2 The Binomial Distribution

Consider the following situations:

1. A medical researcher is working on the development of a new drug for treatment of high blood pressure. Some concern has been expressed about the possibility that the drug could cause a potentially harmful side effect in some people. The drug is administered to a sample of people with high blood pressure. Of the 80 people in the sample, 12 suffer the side effect and the remaining 68 do not.

2. A firm manufactures cameras that use an electronic device to choose automatically the appropriate shutter speed for a picture. To investigate the reliability of this device, the firm's quality control department takes a sample of 15 cameras from the assembly line and tests each camera to see if the electronic device is working properly. In one of the 15 cameras, the device does not work properly.

3. In the development of standardized tests, the degree of difficulty of questions is of interest. One such question is included in a sample test given to 1,000 high school seniors. The test results show that 714 of the students answered the question correctly and 286 gave wrong answers.

In all of these situations, there are two possible outcomes for each member of the sample. A person either suffers the side effect or does not, the electronic device in a camera is either good or defective, and a student's answer is either correct or incorrect. For convenience, we will label the two possible outcomes S (success) and F (failure), but you should not attach any value judgments to these labels. For example, in the quality control example we could label a defective component S and a good component F.

An experiment with two possible outcomes is called a Bernoulli trial. The probabilities associated with the two outcomes S and F will be labeled $p = P(S)$ and $q = P(F)$. Since there are only two outcomes, $q = 1-p$. For instance, if S and F represent heads and tails in the tossing of a fair coin, then $p = q = 1/2$. In the quality control example, if approximately 5 percent of the electronic devices produced by the camera manufacturer's supplier are defective, then $p = .05$ and $q = .95$.

We are now prepared to define a binomial experiment. The conditions necessary for a binomial experiment are as follows:

1. The experiment consists of n Bernoulli trials (trials having only two possible outcomes).
2. The probability of success, p, is the same for all trials (and q, the probability of failure, is therefore also the same for all trials).
3. The trials are independent.

The first condition requires that each trial has only two possible outcomes and denotes the number of trials by n. The first key element to look for in identifying an experiment as a binomial experiment is the limitation to only two possible outcomes per trials. For example, suppose that a sample of registered voters is asked to state a preference between the two candidates who are running for a particular office. If each voter must choose one candidate or the other, then the first condition for a binomial experiment is satisfied. However, in polls of this nature, other answers, such as "undecided", are often allowed. If there are more than two possible answers, then the experiment is not binomial.

A word of caution is in order here. The number of outcomes in an experiment can depend on how the experimenter decides to define the outcomes. In a survey of voting intentions, it is possible to consider "prefers Candidate A" as one outcome and to combine all other responses ("prefers Candidate B", "undecided") as a second outcome. In an experiment involving the recording of the level of a pollutant in a sample of air, many different levels are possible. The experimenter, however, may only be interested in whether the level exceeds a particular value, such as 20 parts per million. Here 20 parts per million might be the legal limit for this particular pollutant. Thus, the experimenter might define a success to be "level exceeds 20 parts per million" and a failure to be "level does not exceed 20 parts per million". The point is simply that the restriction to only two outcomes may be a logical consequence of the experiment (as in a question with only two permissible answers) or it may be created by the way the experimenter chooses to define the outcomes (as in the pollution example).

The second condition for a binomial experiment states that the trials must be identical in the sense that p is the same for all trials. If a con artist switches from a fair coin to a loaded coin in the middle of a series of 10 tosses, the experiment represented by the 10 tosses is not a binomial experiment. If the same loaded coin is used for all 10 tosses, then the trials are identical. For another example, suppose that 72 percent of the registered voters in California are in favor of a decrease in the state income tax. In a random sample of 1,000 registered voters, the chance of selecting a person who favors a tax decrease is the same from trial to trial : $p = 0.72$.

The third condition for a binomial experiment is independence. This means that the outcome on any trial cannot affect the outcomes on other trials. In tossing a coin, we say that the coin has no memory. In an opinion poll, allowing respondents to hear each other's answers might violate independence. If everyone seems to be favoring Candidate A, a respondent who actually prefers Candidate B may claim to prefer A in order to go along with the crowd. It is important to avoid introducing potential biases like this into an experiment.

The random variable of interest in a binomial experiment is x, the number of successes observed in n trials. The number of successes can be as small as 0 or as large as n. The probability distribution of x is called a binomial distribution, which is given by the formula

$$P(x) = \frac{n!}{x!\,(n-x)!}\, p^x q^{n-x} \quad \text{for } x = 0, 1, 2, \ldots, n,$$

where n is the number of trials, p is the probability of success on any single trial, and $q = 1-p$ is the probability of failure on any single trial. The symbol n! is called "n factorial", and it represents the product of n, n-1, n-2, and so on down to 1. for instance, $4! = 4.3.2.1 = 24$.

The p^x term in the binomial formula represents the probability associated with the x successes. The probability is p for a single success, and independence allows us to multiply all of these ps (there are x of them), to arrive at p^x. In a similar fashion, the q^{n-x} term represents the probability associated with the n-x failures. If there are x successes in n trials, the remaining n-x trials must be failures, and the probability of failure is q on any single trial.

The $\frac{n!}{x!(n-x)!}$ term represents the number of possible ways of getting x successes in n trials. For example, the x successes could occur on the first x trials, the last x trials, or in various other ways. This term is sometimes denoted by $\binom{n}{x}$ or (C_x^n), and it is known as a binomial coefficient. You need not worry about why the binomial coefficient takes on this particular form involving factorials.

To illustrate the use of the binomial formula to calculate probabilities, we will first consider some simple examples with small values of n. Suppose that a fair coin is tossed three times, with success corresponding to heads and failure corresponding to tails. Here we have $n = 3$, $p = 0.5$, and x = the number of times heads appears in three tosses. The probability of observing heads exactly twice is

$$P(2) = \frac{3!}{2!\,(3-2)!}(0.5)^2(0.5)^1$$

$$= \frac{3.2.1}{2.1.1}(0.5)^2(0.5)$$

$$= 0.375.$$

The probability that x = 3 is

$$P(3) = \frac{3!}{3!\,(3-3)!}(0.5)^3(0.5)^0$$

$$= \frac{3.2.1}{3.2.1.1}(0.5)^3$$

$$= 0.125.$$

Note that 0! = 1 and anything raised to the zero power equals 1. If we were interested in the probability of <u>two or more</u> heads in three tosses, we would compute

$$\begin{aligned} P(2 \text{ or } 3) &= P(2) + P(3) \\ &= 0.375 + 0.125 \\ &= 0.50. \end{aligned}$$

The entire binomial distribution for this example is shown in Figure 5.1.

For another example, suppose that an insurance company has issued life insurance policies to five people in a particular high-risk profession. The policies are in force for three years, and the probability that a person in this high-risk category will die within a three-year period is 0.15. What is the probability that the insurance company will have to pay off on no more than one of the five policies? There are two possible outcomes for each person - death within three years, or survival for the entire three-year period. The placement of the five people in the same high-risk category indicates that p is about the same for all five. Also, suppose that none of the five work with each other and accidents with multiple deaths among the five can be ruled out. Thus, independence seems reasonable, and the conditions of a binomial experiment are met. Letting x represent the number of deaths among the five insured people, we have a binomial distribution for x with n = 5, p = 0.15, and q = 0.85. The probabilities for x = 0 and x = 1 are

$$P(0) = \frac{5!}{0!\,5!}\,(0.15)^0\,(0.85)^5$$

$$= \frac{5.4.3.2.1}{1.5.4.3.2.1}\,(0.85)^5$$

$$= 0.4437$$

and

$$P(1) = \frac{5!}{1!\,4!}\,(0.15)^1\,(0.85)^4$$

$$= \frac{5.4.3.2.1}{1.4.3.2.1}\,(0.15)(0.85)^4$$

$$= 0.3915.$$

The probability that the insurance company will have to pay off on no more than one of the five policies is therefore

$$\begin{aligned} P(x \text{ is not greater than } 1) &= P(0 \text{ or } 1) \\ &= 0.4437 + 0.3915 \\ &= 0.8352. \end{aligned}$$

The chance of having to pay off on all five policies is very remote:

$$P(5) = \frac{5!}{5!\,0!}\,(0.15)^5\,(0.85)^0$$

$$= \frac{5.4.3.2.1}{1.4.3.2.1}\,(0.15)^5$$

$$= 0.000076.$$

The entire probability distribution for x is given in Figure 5.2.

Computing binomial probabilities by hand is burdensome unless n is very small. It is easy to use a computer for such calculations. Alternatively, when n is not too large, tables of the binomial distribution can be used to find some probabilities. Table 1 gives binomial probabilities for selected values of n and p, and the table is preceded by instructions and examples. In the quality control example presented at the beginning of this section, suppose that approximately 5 percent of the electronic devices produced by the camera manufacturer's supplier are defective. The probability that the electronic device does not work properly in one camera out of a sample of 15 cameras can be found from Table 1. This probability, corresponding to $n = 15$, $p = .05$, and $x = 1$ in Table 1, is 0.366.

When n is large enough that Table 1 can no longer be used, two options are available for determining binomial probabilities. The first, mentioned earlier, is to use the computer. The second is to use a different distribution to approximate the binomial. Two approximations to the binomial distribution will be discussed later in this chapter.

What can we say about the shape of binomial distributions? The distribution in Figure 5.1 ($n = 3$, $p = 0.5$) is symmetric, but the distribution in Figure 5.2 ($n = 5$, $p = 0.15$) is skewed to the right. Various binomial distributions are shown in Figures 5.3-5.8 to give you an idea of how n and p affect the shape of the distribution when $p = 0.5$, the distribution is symmetric. For a given n, as p moves further from 0.5, the skewness increases - skewness to the right when p is less than 0.5, to the left when p is greater than 0.5. For a given p different from 0.5, the skewness decreases as n increases.

In Figures 5.3-5.8, the mean and standard deviation of x are displayed along with the graph of the probability distribution. The mean, variance, and standard deviation of the binomial distribution can be found from the formulas

$$\mu = np,$$
$$\sigma^2 = npq,$$
and
$$\sigma = \sqrt{npq}.$$

Suppose that the probability is 0.20 that a person who takes a certain drug will suffer a harmful side effect. Let x represent the number of people suffering side effects in a sample of 80 people who take the drug. The mean and standard deviation of x are

$$\mu = 80(0.20) = 16$$
and
$$\sigma = \sqrt{80(0.20)(0.80)} = 3.58.$$

In summary, many real-world situations can be viewed in terms of identical, independent trials with only two outcomes per trial. These situations are called binomial experiments. Given n, the number of trials, and p, the probability of one of the outcomes (success), the probability of observing exactly x successes can be found from the binomial formula:

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

where $q = 1-p$ is the probability of the other outcome (failure) and x must be between 0 and n. The mean and standard deviation of x are given by the formulas

$$\mu = np$$

and

$$\sigma = \sqrt{npq}.$$

To calculate binomial probabilities or the binomial mean and standard deviation, we must know n and p. Later in the book, we will consider situations for which p is not known. For example, in an opinion poll the experimenter does not know the proportion of voters who favor Candidate A but hopes to use the results of the poll to make inferences about this proportion. As you shall see, the binomial distribution is used to help make such inferences.

5.3 The Poisson Distribution

The Poisson distribution is similar to the binomial in that it can be represented as a distribution of x, the number of times a particular event occurs. For the binomial distribution, x is the number of times an event (success) occurs in a series of dichotomous trials. For the Poisson distribution, x is the number of times an event occurs over some continuum, such as time. The following two examples illustrate situations for which the Poisson distribution might be appropriate.

1. A hospital administrator is considering expanding the emergency room staff on Friday and Saturday evenings, when the emergency room tends to be very crowded. For various periods during several Friday and Saturday evenings, the number of arriving patients is counted. Over a total of 20 hours, 341 patients arrive at the emergency room.

2. A biologist is studying the degree to which a certain organism is present in a reservoir that is used for recreational purposes and as a source of water for a small metropolitan area. Samples of water are taken from the reservoir, and the samples are analyzed. Some samples do not contain the organism, but the organism is observed once, twice, or even three times in other samples. Overall, the organism is found 46 times in 150 cubic centimeters of water.

In the first example, the arrival of patients occurs over time, a single dimension. In the second example, the organism is present at various places in the reservoir, a three-dimensional space. For these situations to be considered as Poisson experiments, two more conditions must be satisfied. The events must occur randomly but at the same average rate, and occurrences must be independent. For example, when patients made appointments with a doctor, their arrivals are somewhat regular, but arrivals at an emergency room would be expected to be more random.

Also, the emergency room experiment considered only Friday and Saturday evenings, for which the average arrival rates would be expected to be about the same. Tuesday morning from, say, 3a.m. to 6a.m., would probably have only a small number of patients as compared with 9p.m. to midnight on Friday night. For the reservoir example, randomness with the same average rate means that the organisms are distributed more or less randomly throughout the reservoir, and we would hope that the biologist would select samples of water from various places in the reservoir instead of just taking 150 cubic centimeters in one fell swoop. Independence implies that the arrivals at the emergency room and the organisms do not tend to occur in clusters. If a major fire occurs, many patients may come to the hospital, and their arrivals would be related (via the fire).

A Poisson experiment, then, can be defined in terms of three conditions:

1. Occurrences of an event are observed over time, space, or some other continuum.
2. The event occurs randomly but at the same average rate.
3. The occurrences are independent.

The random variable of interest in a Poisson experiment is x, the number of times the event of interest occurs. In our examples, x represents the number of arrivals at the emergency room in the 20-hour period and the number of times the organism appears in the 150-cubic-centimeter sample of water. The probability distribution of x is called a Poisson distribution, which is given by the formula

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!} \text{ for } x = 0, 1, 2, \ldots,$$

where λ (Greek lambda) is the average rate at which x would be expected to occur. The symbol e represents a constant, the base of the natural logarithm system, which is approximately 2.718. The symbol x! is a factorial term like the term encountered in a binomial formula.

The Poisson formula may look forbidding to some readers, but it won't pose a serious problem. If you are not familiar with the constant e, don't worry. Poisson probabilities can be found from a table or determined by using a computer or calculator with a built-in e.

Before the use of the Poisson formula is illustrated, a note of caution is in order. The value of λ must be in units comparable to the "size" of the experiment. For instance, if the long-run average arrival rate of patients at the emergency room on Friday and Saturday evenings is 18 per hour and the experiment consists of 20 hours, then λ is 18(20) = 360 per 20 hours. If the organism is present once in 3 cubic centimeters on average, λ is ($^1/_3$)(150) = 50 per 150 cubic centimeters.

As in the case of the binomial distribution, computing Poisson probabilities by hand is tedious in most instances, but it is relatively easy to use a computer for such calculations. Alternatively, tables are available. Table 2 gives Poisson probabilities for selected values λ, and the table is preceded by instructions and examples. In the emergency room example, suppose that the long-run average arrival rate of patients at the emergency room on Friday and Saturday evenings is 18 per hour. What is the probability that exactly 20 patients will come in between 9p.m. and 10p.m. one Friday evening? The length of time is one hour, λ is 18 for a period of one hour, and the probability is

$$P(20) = \frac{18^{20}e^{-18}}{20!}.$$

Looking at Table 2 under $\lambda = 18$ and $x = 20$, we find the value of 0.0798.

In the reservoir example, suppose that the organism is present once in 3 cubic centimeters of water on average. What is the probability of finding less than 3 organisms in a sample of 12 cubic centimeters of water? The appropriate λ is $(1/3)(12) = 4$ per 12 cubic centimeters, and the probability is

$$P(x \text{ is less than } 3) = P(0) + P(1) + P(2).$$

From Table 2 with $\lambda = 4$ and $x = 0$, 1, and 2, we find

$$P(0) = 0.0183,$$
$$P(1) = 0.0733,$$
and
$$P(2) = 0.1465.$$

Therefore,

$$\begin{aligned} P(x \text{ is less than } 3) &= 0.0183 + 0.0733 + 0.1465 \\ &= 0.2381. \end{aligned}$$

Some Poisson distributions are shown in Figures 5.9-5.11 to show you how the shape of the distribution changes as λ varies. The mean, variance, and standard deviation of the Poisson distribution are given by the formulas

$$\mu = \lambda,$$
$$\sigma^2 = \lambda,$$
and
$$\sigma = \sqrt{\lambda}.$$

In the emergency room example with a time period of one hour and $\lambda = 18$ per hour, the expected number of patients arriving is

$$\mu = 18,$$

and the standard deviation of x is

$$\sigma = \sqrt{18} = 4.24.$$

The Poisson distribution is useful not just in its own right, but also as an approximation for the binomial distribution under certain circumstances. The approximation is good when the number of trials n is large and when the probability p is small. How large is large and how small is small? That depends on how accurate you would like the approximation to be. A rough rule of thumb is to use the approximation when $n \geq 20$ and $p \leq 0.05$.

Suppose that records show that approximately 1.5 percent of the drivers assigned to a particular "low risk" category by an insurance company have one or more accidents in a year. If 800 drivers are currently in this category, what is the probability that 10 of them have an accident during the next year? This is a binomial experiment with $n = 800$ and $p = 0.015$, and the desired probability is

$$P(10) = \frac{800!}{10!\ 790!} (0.015)^{10} (0.985)^{790}.$$

This can be calculated with the aid of a computer or a good hand calculator (its value is 0.1052), but it is easier to use a Poisson approximation.

To use the Poisson distribution to approximate the binomial distribution, simply set λ equal to np. (Note that this equates the means of the distributions.) The probability that 10 of the 800 low-risk drivers have an accident during the next year can be approximated by using a Poisson probability for $x = 10$ with $\lambda = np = 800(0.015) = 12$. From Table 2 with $\lambda = 12$,

$$P(10) = 0.1048.$$

You can see that this Poisson approximation is very close to the binomial probability of 0.1052, and it is much easier to use the Poisson tables than to compute the binomial probability.

A few binomial distributions and the corresponding Poisson approximations are shown in Figures 5.12-5.14. The approximation gets better as n increases and p decreases, as suggested earlier. When n is large but p is not small, the normal distribution provides a better approximation to the binomial distribution; this approximation will be discussed in the following section.

5.4 The Normal Distribution

A histogram representing the student ages used in Chapter 2 (see Table 2.5) is shown in Figure 5.15. In Figure 5.16, a bell-shaped curve known as the normal curve is superimposed on the histogram, and you can see that the histogram is very close to the curve. The curve represents the normal distribution, which is widely used in statistics. Not only ages, but many other variables, such as heights, weights, IQ scores, and exam grades, often have distributions that follow the normal curve quite closely. Moreover, an important result known as the Central Limit Theorem, which will be discussed in Chapter 7, provides additional justification for the application of the normal distribution in statistics.

A normal distribution can be specified by giving two values, the mean and the standard deviation of the distribution. The curve is bell-shaped, and the "bell" is centered at the mean. Notice that the curve is symmetric; the half to the right of the mean is a perfect mirror image of the half to the left of the mean. As the standard deviation changes, the "bell" becomes taller and thinner (for smaller standard deviations) or shorter and fatter (for larger standard deviations). Some examples are shown in Figure 5.17.

The formula for the normal curve is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \text{ for } -\infty < x < \infty,$$

where μ is the mean of the distribution and σ is the standard deviation. As noted in the discussion of the Poisson distribution, the symbol e represents a constant, the base of the natural logarithm system, which is approximately 2.718. The symbol π represents another constant, approximately 3.1416, which you may have encountered in geometry in formulas for the area and circumference of a circle.

Fortunately, although the formula for the normal curve is included here for completeness, you will not need to use this formula. Normal probabilities can be found from a table or a computer. Since the normal curve is continuous, probability is represented by area under the curve. The total area under the normal curve is one. Furthermore, for any normal distribution, the area under the curve between μ-σ and μ+σ is approximately 0.68 (Figure 5.18), the area between μ-2σ and μ+2σ is about 0.95 (Figure 5.19), and the area between μ-3σ and μ+3σ is slightly greater than 0.997 (Figure 5.20). Thus, about 68 percent of the probability is within one standard deviation from the mean, 95 percent is within two standard deviations, and 99.7 percent is within three standard deviations. Even though the normal formula holds for any value of x, no matter how large or small, values of x more than three standard deviations from the mean are quite unlikely to occur.

To be able to find probabilities for normal distributions, first you need to concentrate on a special normal distribution, the standard normal distribution. This is a normal distribution with mean zero and standard deviation one. Because a standardized random variable, or a standard score, is usually denoted by z, z is used here to represent a variable with a standard normal distribution. Table can be used to find probabilities for z.

In Table 3, cumulative probabilities such as $P(z > 1.52)$ are given. To find $P(z > 1.52)$, look up $z = 1.52$ and read the probability, $P(z > 1.52) = 0.0643$. This probability is shown in Figure 5.21.

For negative values of z, the symmetry of the normal curve can be used. For example, the area to the left of -1.52 is the same as the area to the right of +1.52, as shown in Figure 5.22. But the area to the left of -1.52 is found by taking $P(z > -1.52) = 0.9357$ and subtracting it from one:

$$P(z < -1.52) = 1 - P(z > -1.52) = 1 - 0.9357 = 0.0643$$

Table 3 gives cumulative probabilities for both positive and negative values of z.

Often the probability that z is between two values is of interest. How can $P(0.43 < z < 1.75)$ be found? The area to the right of 1.75 is 0.0401, from Table 3, and the area to the right of 0.43 is 0.3336. Thus, the area between 0.43 and 1.75 must be $0.3336 - 0.0401 = 0.2935$.

For a final example involving standard normal probabilities, let us find the shaded area in Figure 5:23, which represent $P(-1.5 < z < 0.5)$. From the normal table, the area to the right of 0.5 is 0.3085. The area to the right of -1.5 is 0.9332.

Subtracting the two values gives

$$P(-1.5 < z < 0.5) = 0.9332 - 0.3085 = 0.6247$$

When you are finding a normal probability, it is a good idea to draw a rough sketch to keep track of the area you want. Such a sketch reduces the chance of making simple mistakes such as finding the area to the left of a value when you really want the area to the right. You will be using the normal distribution very often in later chapters, so it is a good idea to become familiar with computing normal probabilities. Even though a computer can do this work for you, you should learn how to read a table of normal probabilities and how to find specific areas under the normal curve.

Probabilities for any normal distribution can be related to standard normal probabilities, since z can be related to x by subtracting μ from x and then dividing by σ:

$$z = \frac{x - \mu}{\sigma}.$$

For example, suppose that scores on a college entrance examination are normally distributed with mean 600 and standard deviation 50. What is the probability of a person scoring less than 640? This is $P(x < 640)$, but $x = 640$ corresponds to a standard score of

$$z = \frac{x - \mu}{\sigma} = \frac{640 - 600}{50} = 0.80.$$

Since an x less than 640 corresponds to a z less than 0.80,

$$P(x < 640) = P(z < 0.80) = 1 - P(x > 0.80) = 1 - 0.2119 = 0.7881$$

from Table 3. This probability is shown in Figure 5.24. Table 3 gives probabilities only for z, but these can be used to find probabilities for x from any normal distribution.

What proportion of students taking the college entrance examination receive scores between 580 and 680? Converting from x to z yields

$$P(580 < x < 680) = P\left(\frac{580 - \mu}{\sigma} < \frac{x - \mu}{\sigma} < \frac{680 - \mu}{\sigma}\right)$$

$$= P\left(\frac{580 - 600}{50} < z < \frac{680 - 600}{50}\right) = P(-0.40 < z < 1.60).$$

Next, from Table 3,

$$P(z > -0.40) = 0.6554,$$
$$P(z > 1.60 = 0.0548,$$

and $P(z > -0.40) - P(z > 1.60) = 0.6554 - 0.0548 = 0.6006.$

Therefore,

$$P(-0.40 < z < 1.60) = 0.6006.$$

Table 3 can also be used to find percentiles of a normal distribution. For instance, what is the 90th percentile of the distribution of scores on the college entrance examination? This is a value of x which has an area of 0.90 to the left and 0.10 to the right. The cumulative probability at x is 0.10. If you look at the cumulative probabilities in Table 3, you will find that the closest values to 0.10 are 0.1003 (for $z = 1.28$) and 0.0985 (for $z = 1.29$). Since 0.1003 is closer to 0.10 than 0.0985, we will take 1.28 as the 90th percentile of z. By manipulating

$$z = \frac{x - \mu}{\sigma}$$

algebraically, we get x as a function of z:

$$x = \mu + z\sigma.$$

Then, if 1.28 is the 90th percentile of z, the 90th percentile of x is

$$\mu + 1.28\sigma = 600 + 1.28(50) = 664.$$

This means that 90 percent of those taking the examination score less than 664. Any score above 664 is in the top 10 percent of the distribution of scores, as shown in Figure 5.25.

In Section 5.2 the Poisson approximation to the binomial distribution was discussed. The normal distribution can also be used to approximate the binomial distribution. The Poisson approximation is good when the number of trials n is large and the binomial probability p is small. The normal approximation is good when n is large and p is <u>not</u> too small or too large. One rough rule of thumb is to use the normal approximation when both $np \geq 5$ <u>and</u> $n(1-p) \geq 5$ are satisfied.

From Section 5.1, the mean and standard deviation of a binomial distribution are

$$\mu = np$$

and $$\sigma = \sqrt{npq}.$$

For a normal approximation, we simply take a normal distribution with the same mean, np, and the same standard deviation, $\sqrt{npq}$. The z score corresponding to a given value of x is therefore

$$z = \frac{x - \mu}{\sigma} = \frac{x - np}{\sqrt{npq}}.$$

Suppose that 40 percent of the voters in a large city favor a proposition to reduce the tax rate in the city. If a poll is taken with a random sample of 150 voters, what is the probability that exactly 65 of the voters in the sample favor the proposition? This is a binomial probability with n = 150, p = 0.40, and x = 65:

$$P(65) = \frac{150!}{65!\,85!}(0.40)^{65}(0.60)^{85}.$$

Our binomial table does not include values for n = 150. Instead of trying to compute this probability from the binomial formula, let's use a normal approximation. With n = 150 and p = 0.40, the mean and standard deviation are

$$\mu = 150(0.40) = 60$$

and $$\sigma = \sqrt{150(0.40)(0.60)} = \sqrt{36} = 6.$$

Probability is represented by area under the curve when we are working with a normal distribution. Since x = 65 is a single point, there is no area corresponding to that point. The problem is that a continuous distribution (the normal) is being used to approximate a discrete distribution (the binomial). To use the approximation, we must convert x = 65 to an interval: $64.5 < x < 65.5$. Now all of the area under the normal curve can be accounted for in a systematic fashion. For example, consider the area under the curve between 64 and 65. The portion between 64.0 and 64.5 is assigned to x = 64, while the portion between 64.5 and 65.0 is assigned to x = 65. Similarly, the area from 65.0 to 65.5 is assigned to x = 65, and the area from 65.5 to 66.0 is assigned to x = 66. The conversion of the single binomial x to an interval from $x - \frac{1}{2}$ to $x + \frac{1}{2}$ is called a continuity correction because it enables us to use the continuous normal distribution to approximate the discrete binomial distribution.

Now, let us calculate the approximation to P(65). The first step is to convert x = 65 to an interval, the second step is to shift from x to a standard z score, and the next step is to find the probability from the normal table:

$$\begin{aligned} P(65) &= P(64.5 < x < 65.5) \\ &= P\left(\frac{64.5-60}{6} < z < \frac{65.5-60}{6}\right) \\ &= P(0.75 < z < 0.92) \\ &= 0.2266 - 0.1788 = 0.0478. \end{aligned}$$

If we want that the probability that <u>less than</u> 65 of the voters in the sample favor the proposition,

$$\begin{aligned} P(\text{less than } 65) &= P(x < 64.5) \\ &= P(z < \frac{64.5 - 60}{6}) \\ &= P(z < 0.75) = 1 - P(P > 0.75) = 1 - 0.2266 = 0.7734. \end{aligned}$$

"Less than 65" goes down to 64.5 when the continuity correction is included. In contrast, "more than 65" would start at 65.5 in the normal approximation:

$$\begin{aligned} P(\text{more than } 65) &= P(x > 65.5) \\ &= P(z > \frac{65.5 - 60}{6}) \\ &= P(z > 0.92) = 0.1788 \end{aligned}$$

To give you some feeling for how the "goodness" of the normal approximation is affected by n and p, some binomial distributions are presented with their normal approximations in Figures 5.26 - 5.29. As you can see, the approximation gets better as n increases and as p is closer to 0.5.

Figure 5.1 : Binomial Distribution for Number of Heads in Coin Example (n=3, p=0.5).

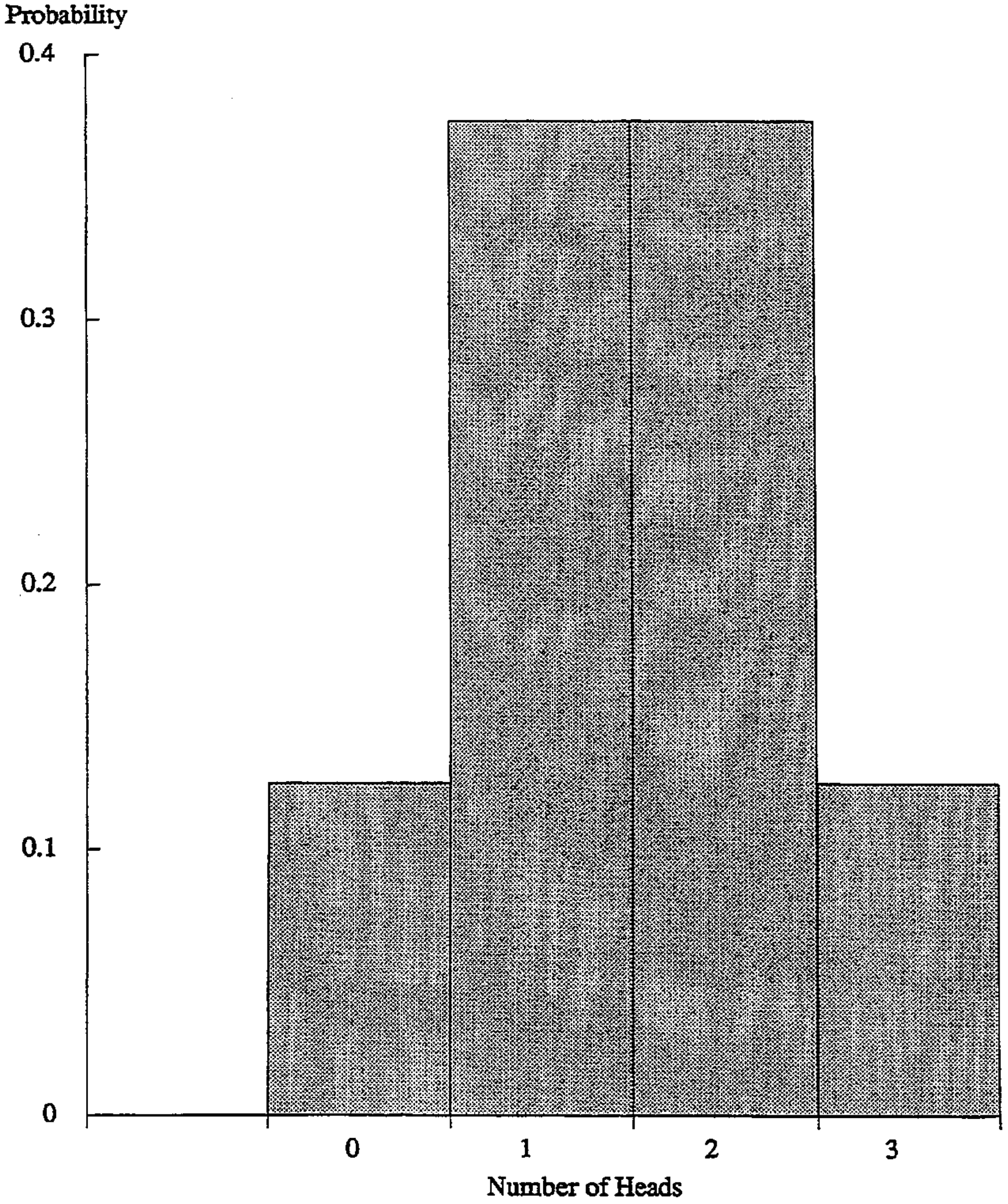


Figure 5.2 : Binomial Distribution for Number of Deaths in Insurance Example (n=5, p=0.15).

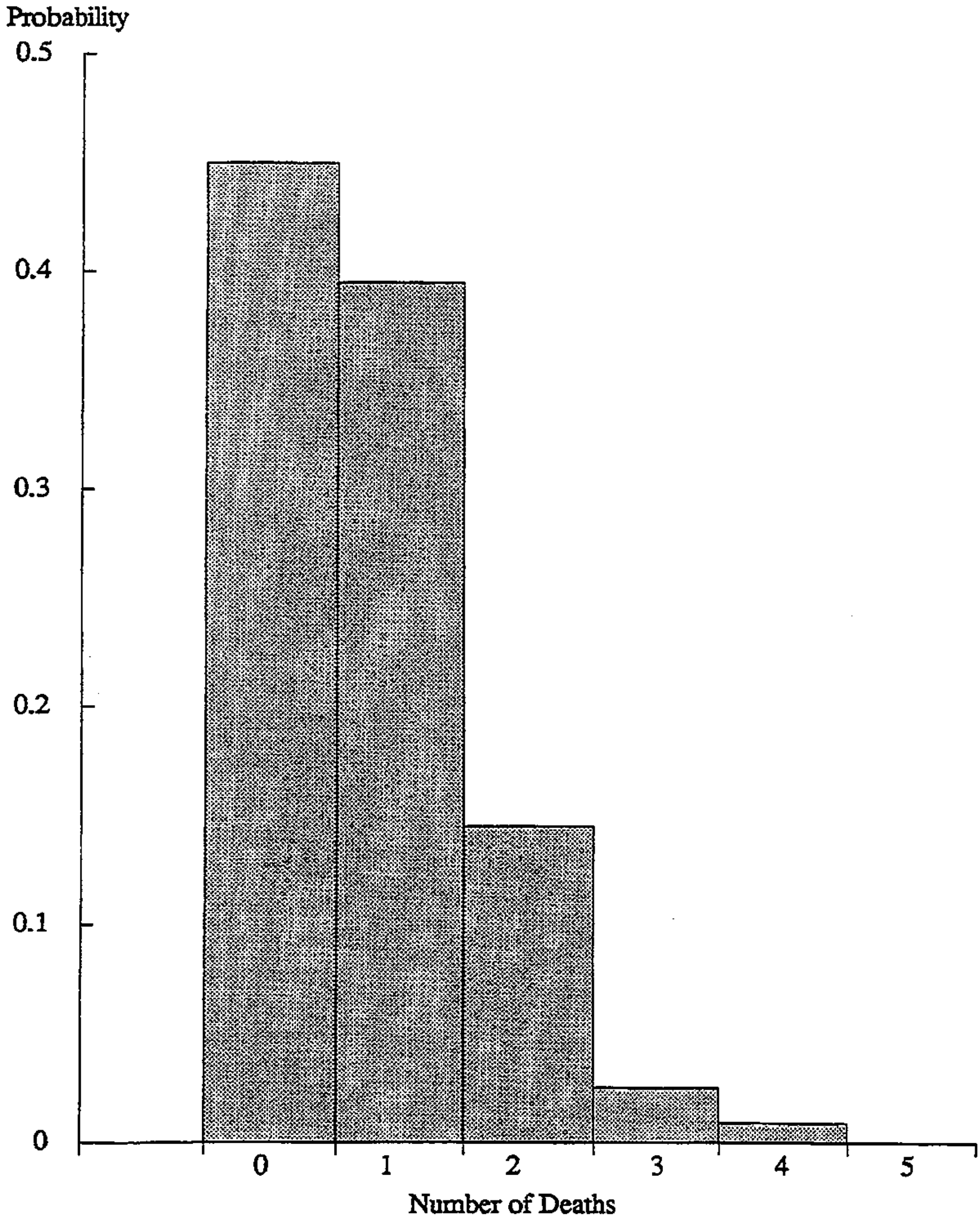


Figure 5.3 : Binomial Distribution with n=5, p=0.5.

Figure 5.4 : Binomial Distribution
with n=5, p=0.3.
Probability
0.4
0.3
0.2
0.1
0
0
1
2
3
4
5
x

Figure 5.5 : Binomial Distribution with n=5, p=0.1.

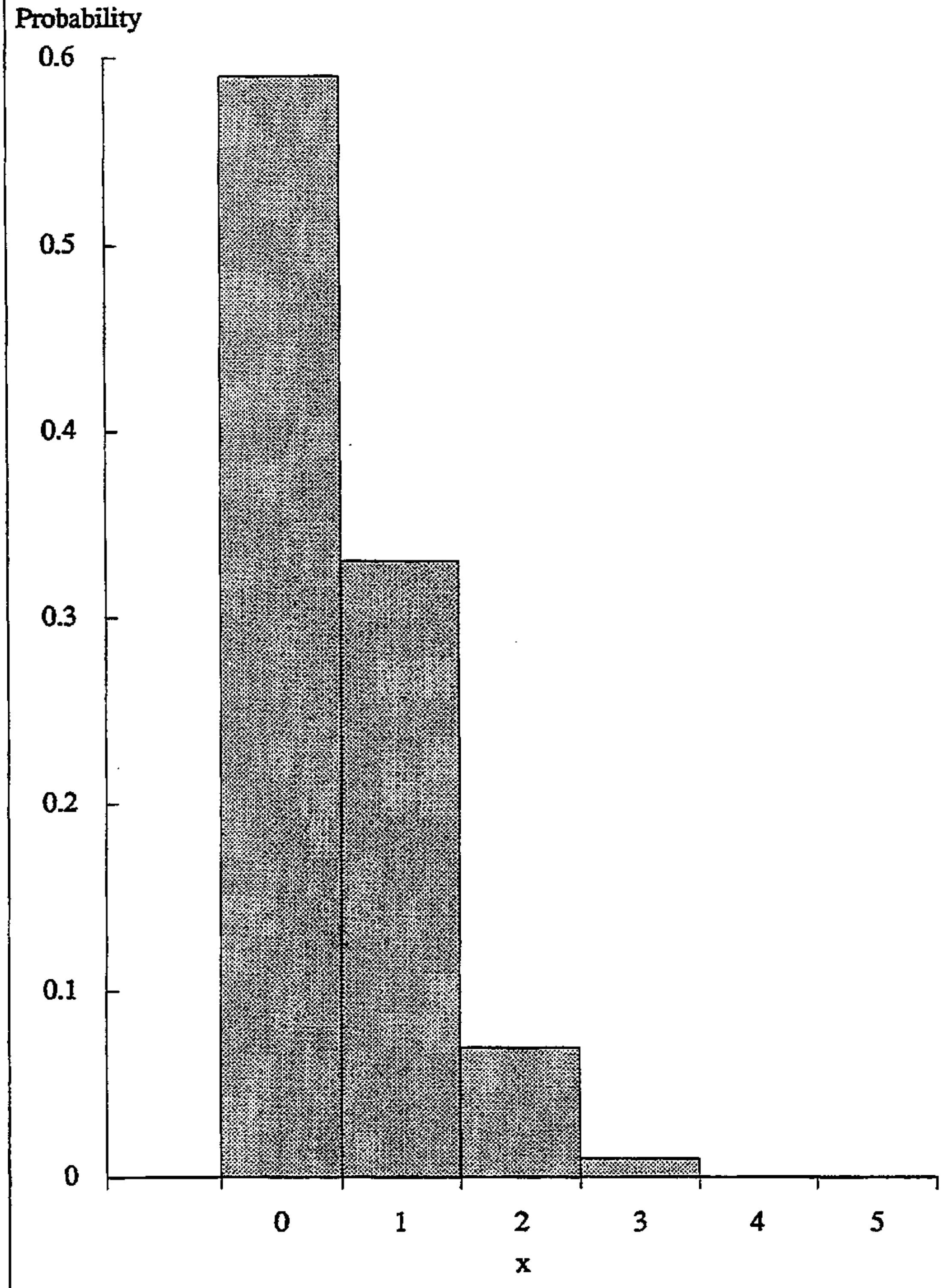


Figure 5.6 : Binomial Distribution with n=15, p=0.5.

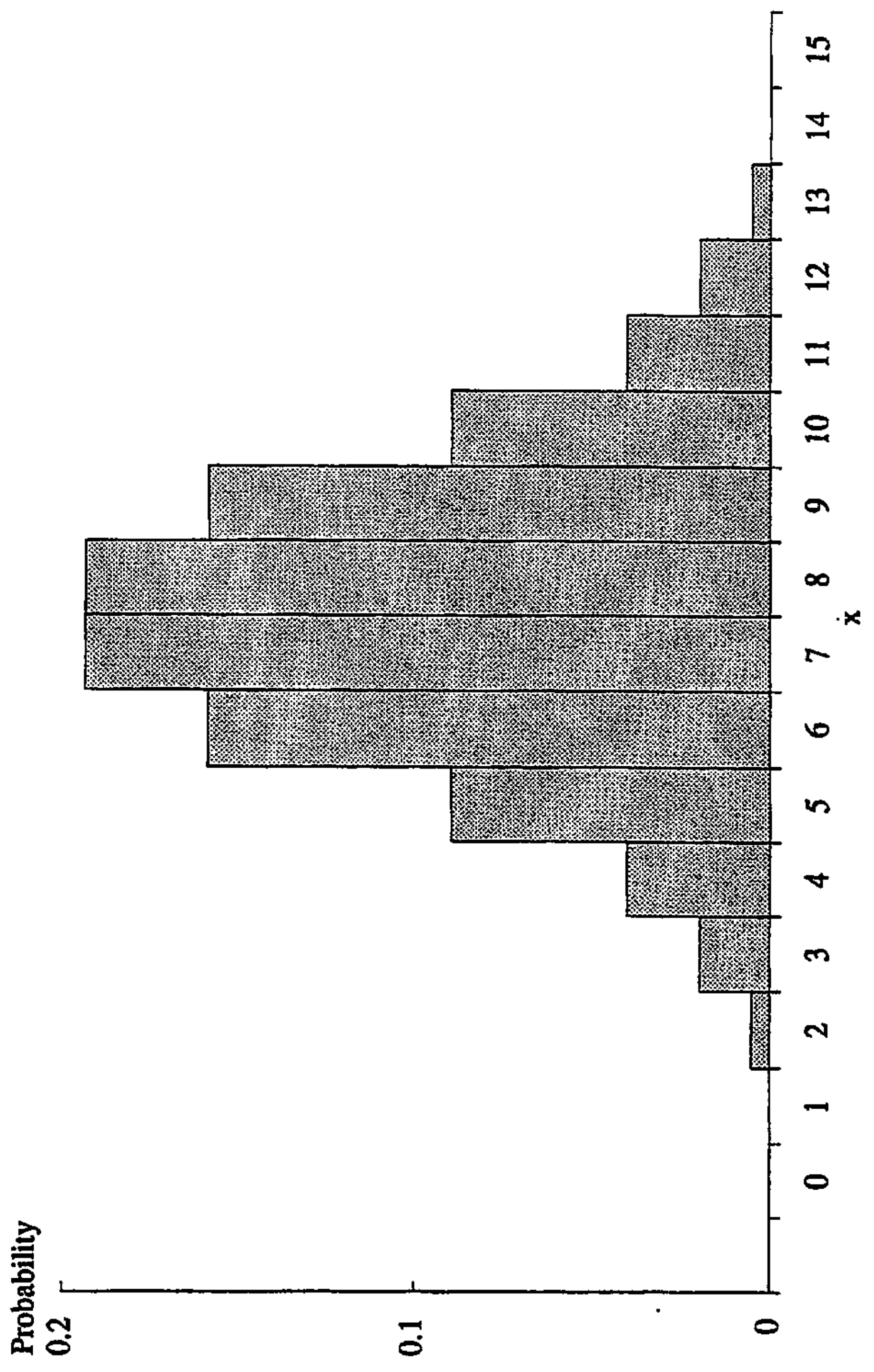
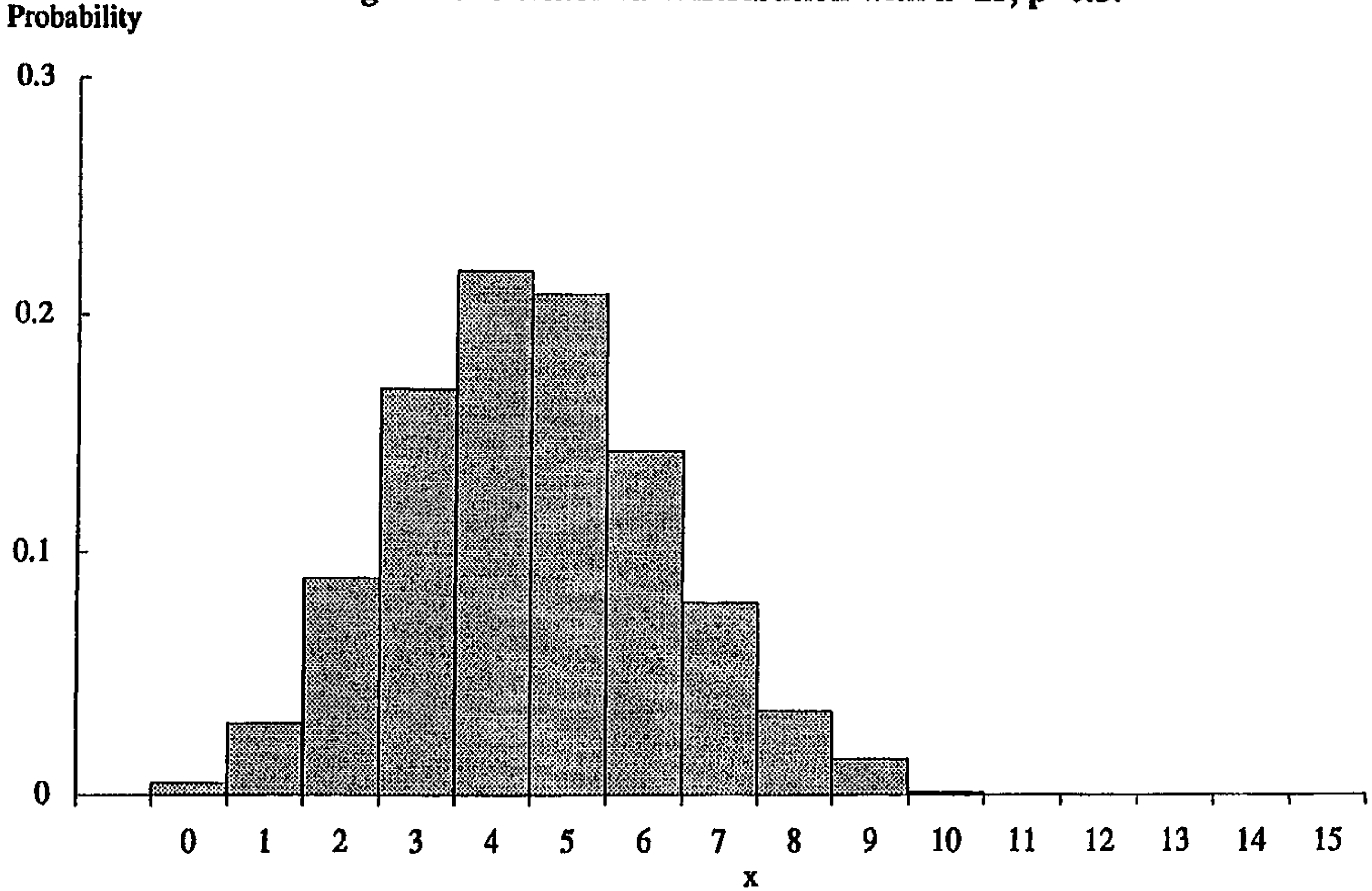


Figure 5.7 : Binomial Distribution with n=15, p=0.3.



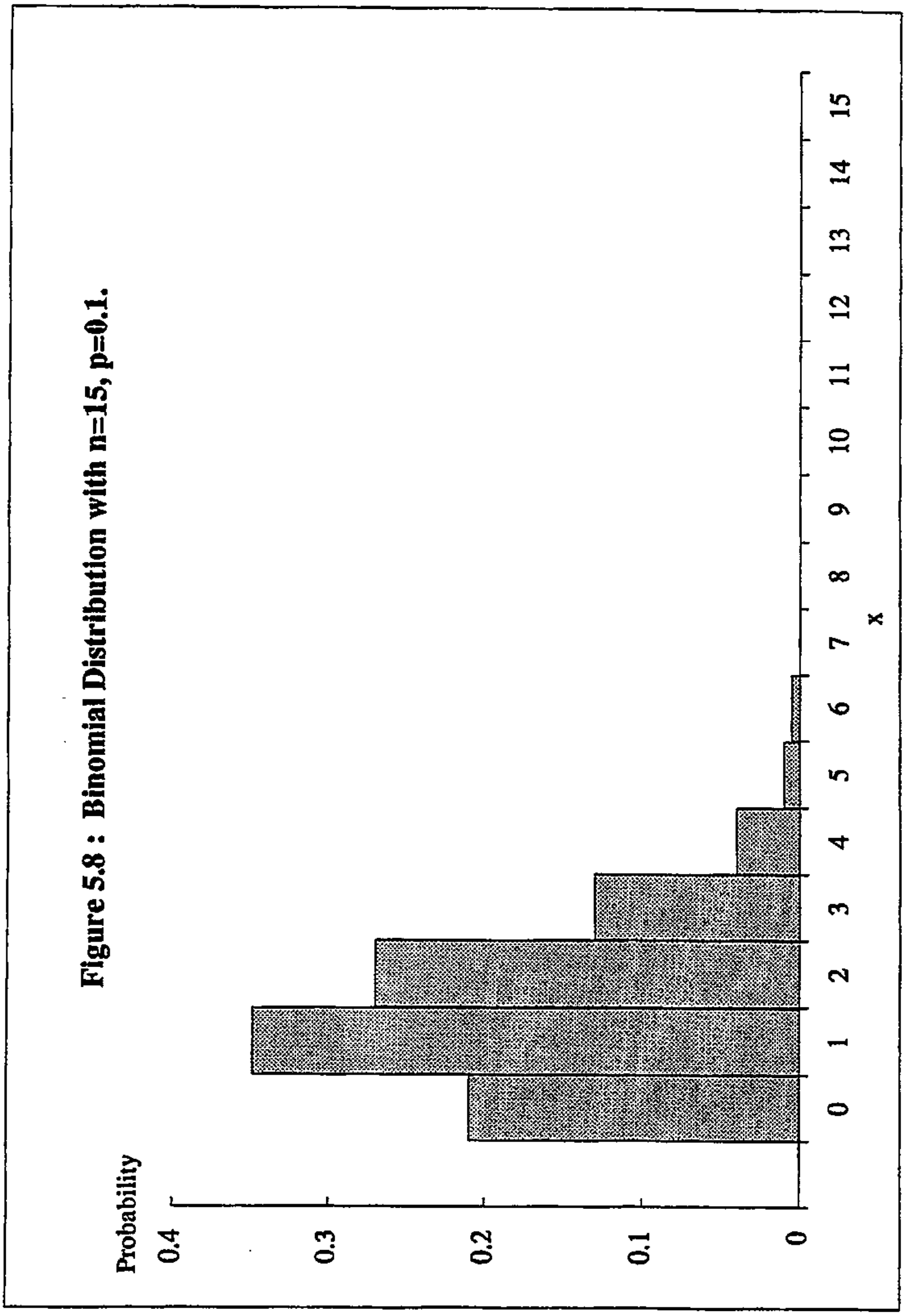


Figure 5.8 : Binomial Distribution with n=15, p=0.1.

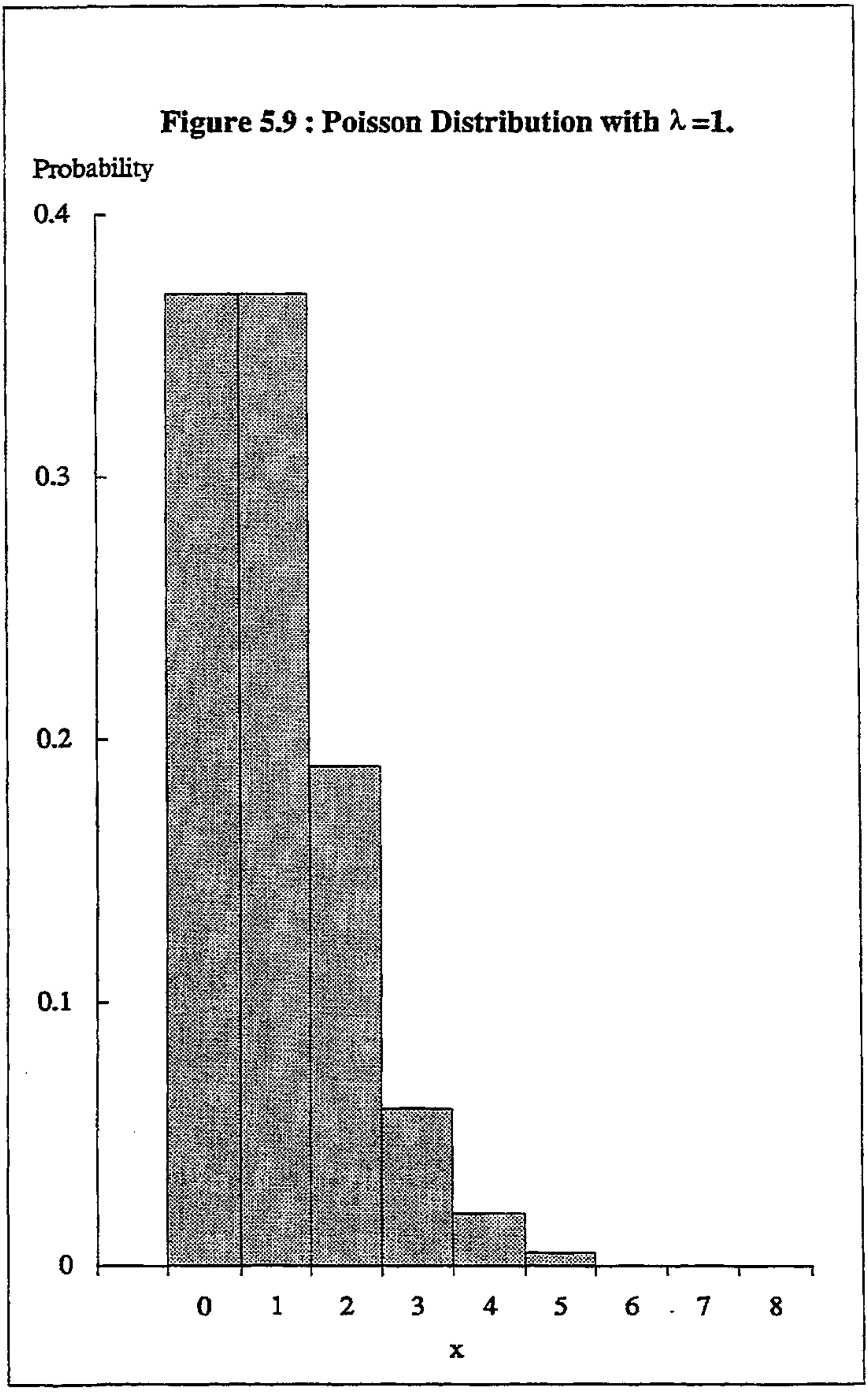
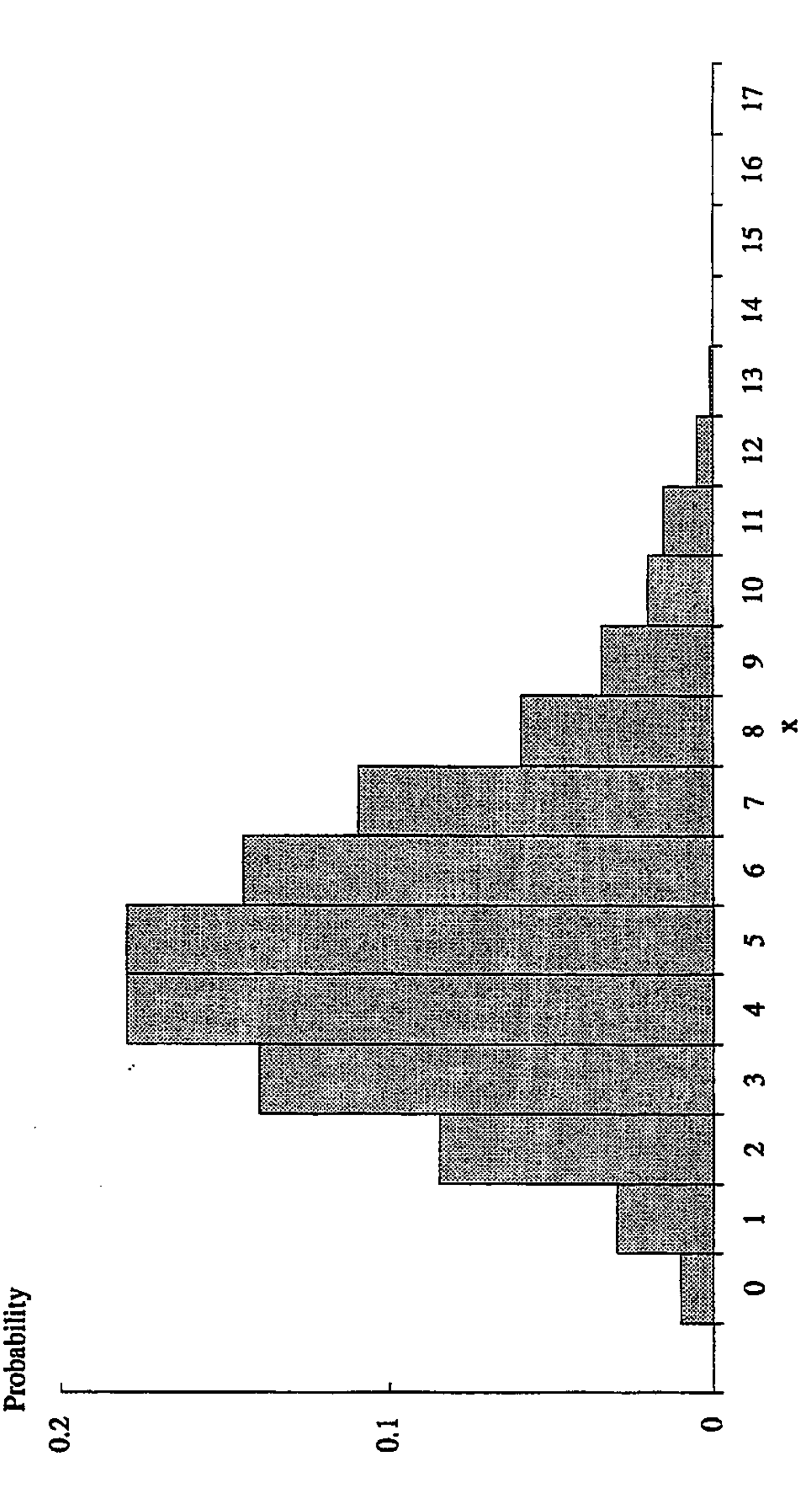


Figure 5.9 : Poisson Distribution with λ =1.
Probability
0.4
0.3
0.2
0.1
0
0
1
2
3
4
5
6
7
8
x

Figure 5.10 : Poisson Distribution with λ =5.



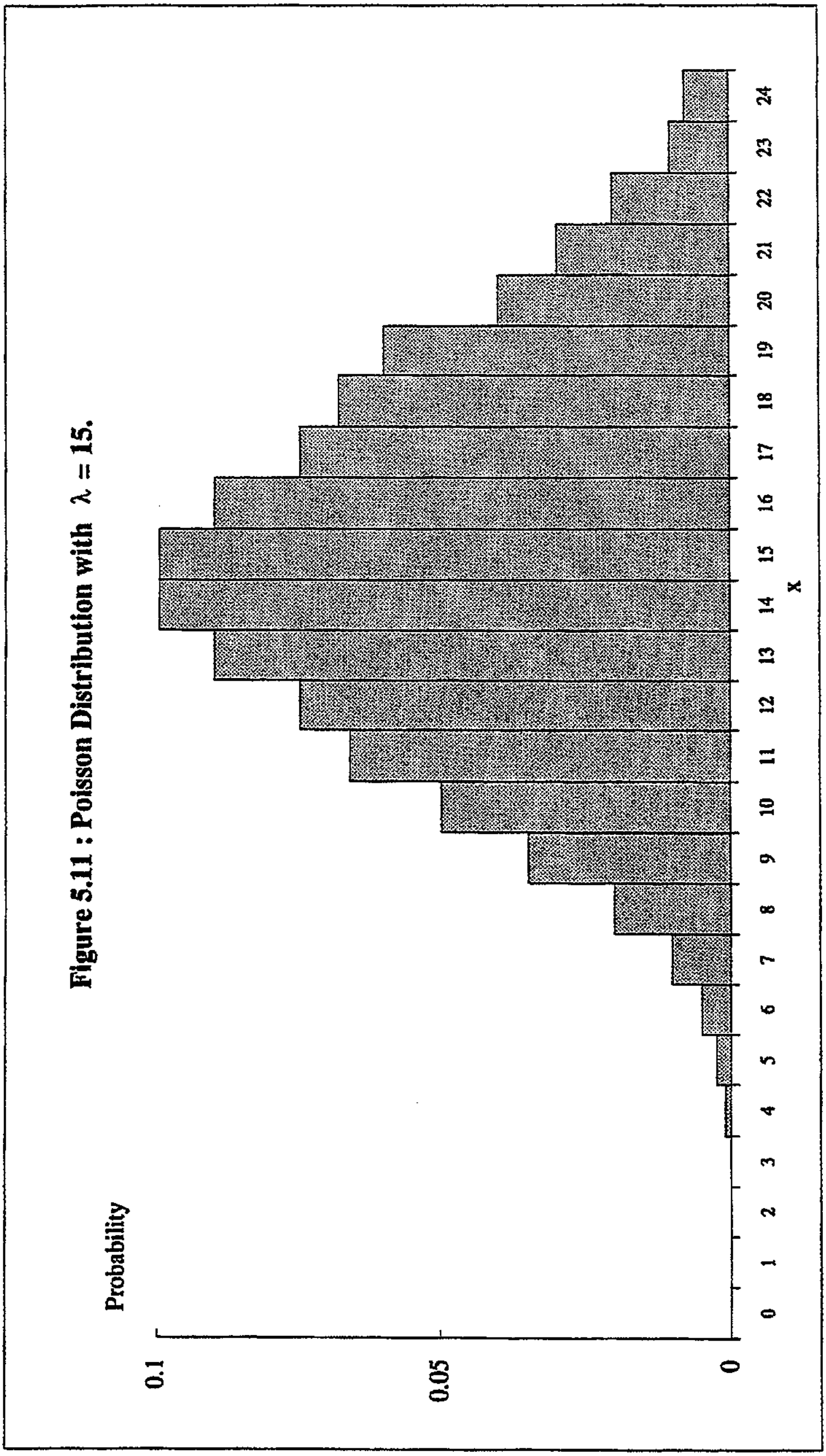
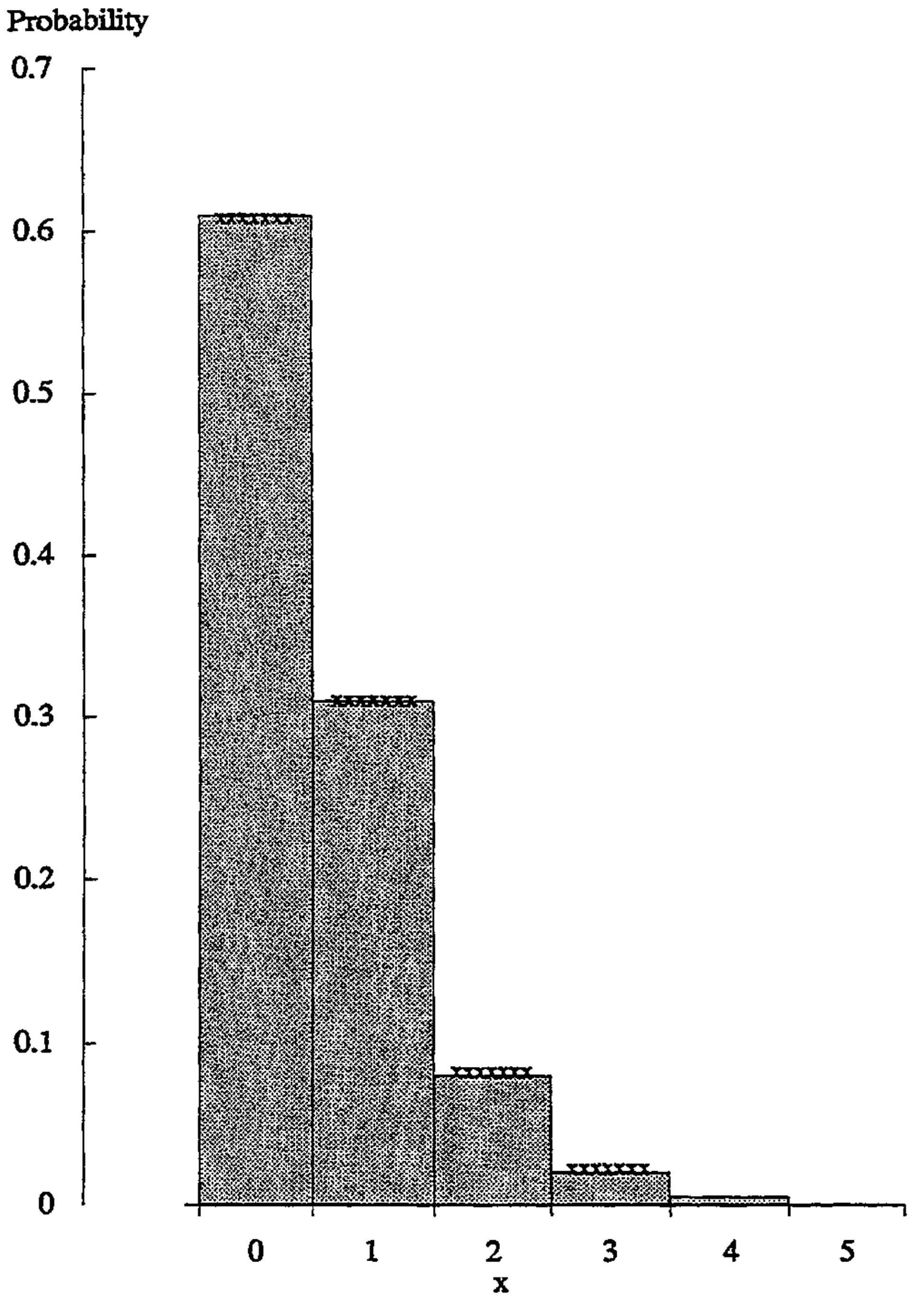


Figure 5.11 : Poisson Distribution with $\lambda = 15$.

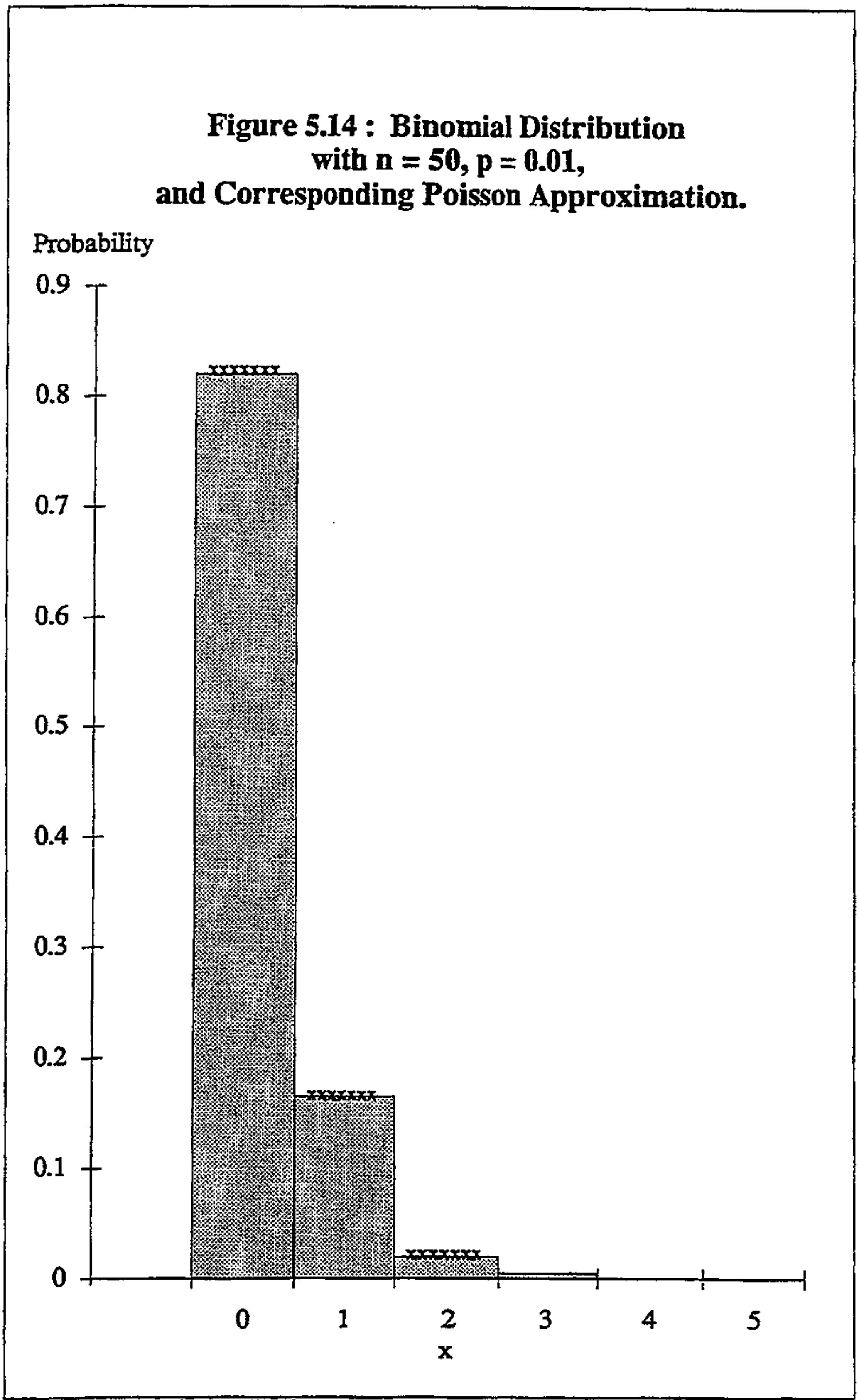
Figure 5.12 : Binomial Distribution with n=20, p=0.05, and Corresponding Poisson Approximation.

— = Binomial Distribution
xxx = Poisson Approximation

Figure 5.13 : Binomial Distribution with n=20, p=0.01, and Corresponding Poisson Approximation.



___ = Binomial Distribution
xxx = Poisson Approximation (Almost a complete overlap)



___ = Binomial Distribution
xxx = Poisson Approximation (Almost a complete overlap)

Figure 5.15 : Histogram of Student Ages (See Table 2.1 for Data).

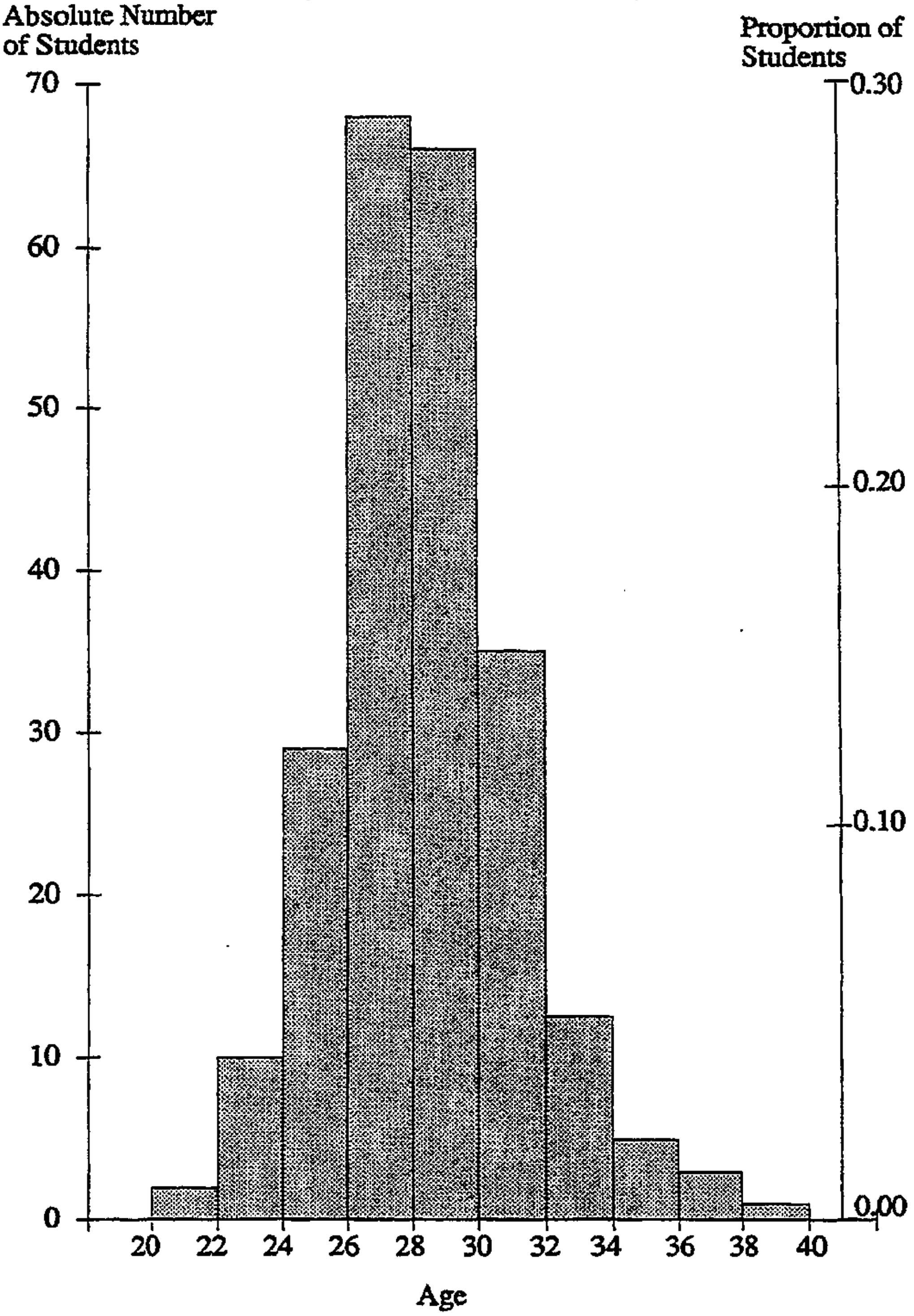


Figure 5.16 : Histogram of Student Ages with Normal Curve.

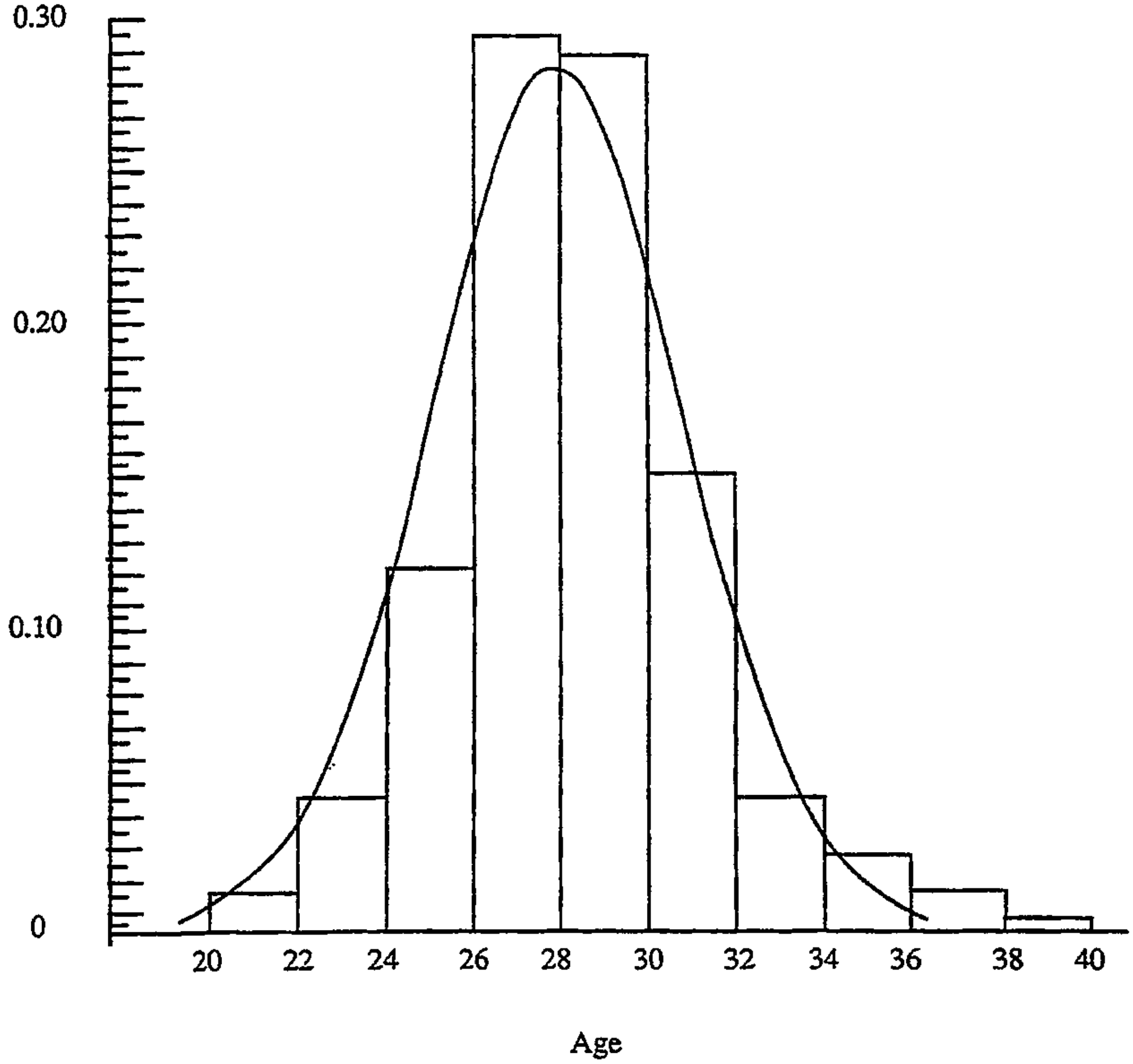


Figure 5.17 : Some Normal Curves.

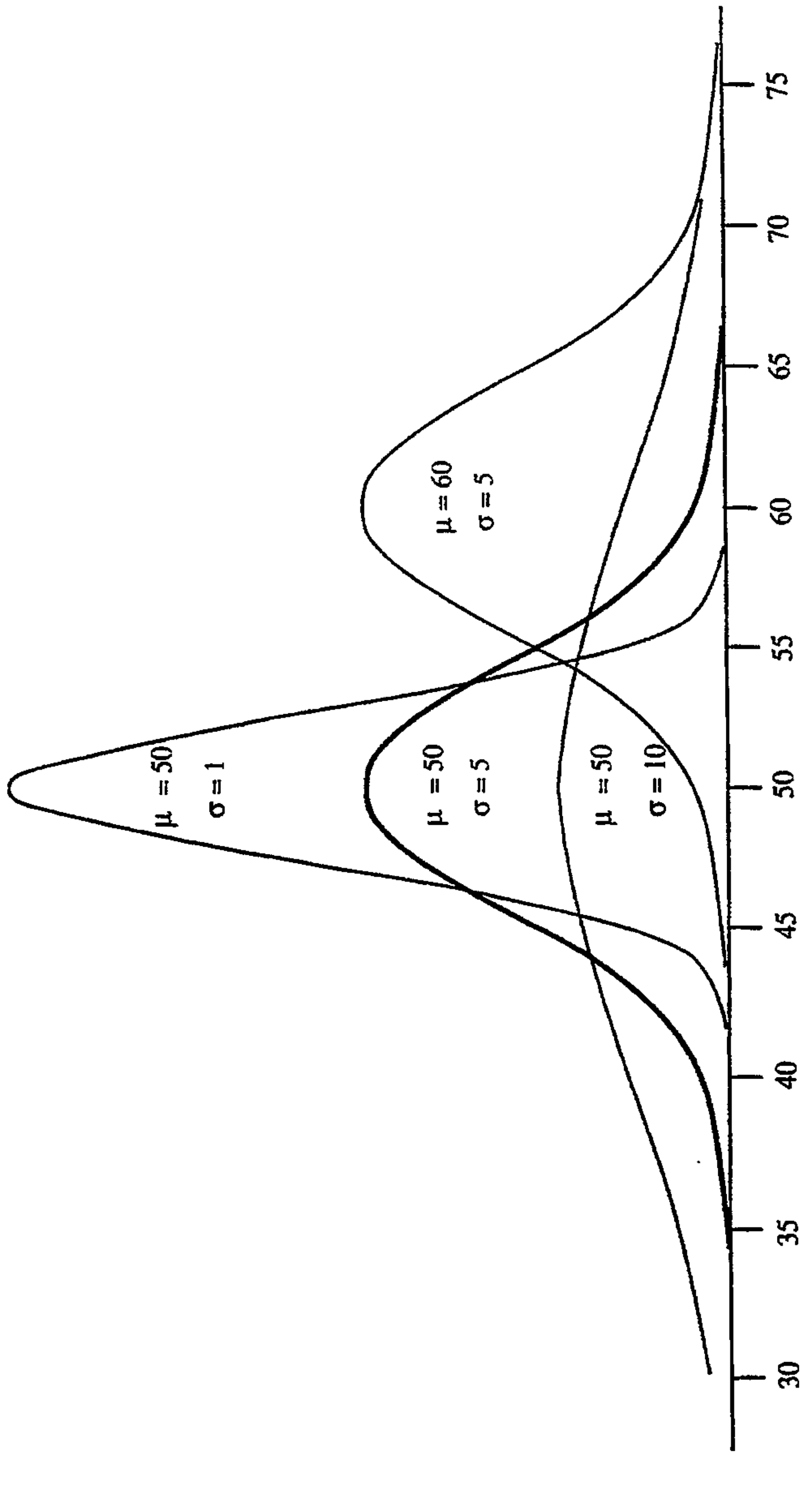


Figure 5.18 : The Area within One Standard Deviation of the Mean in a Normal Curve.

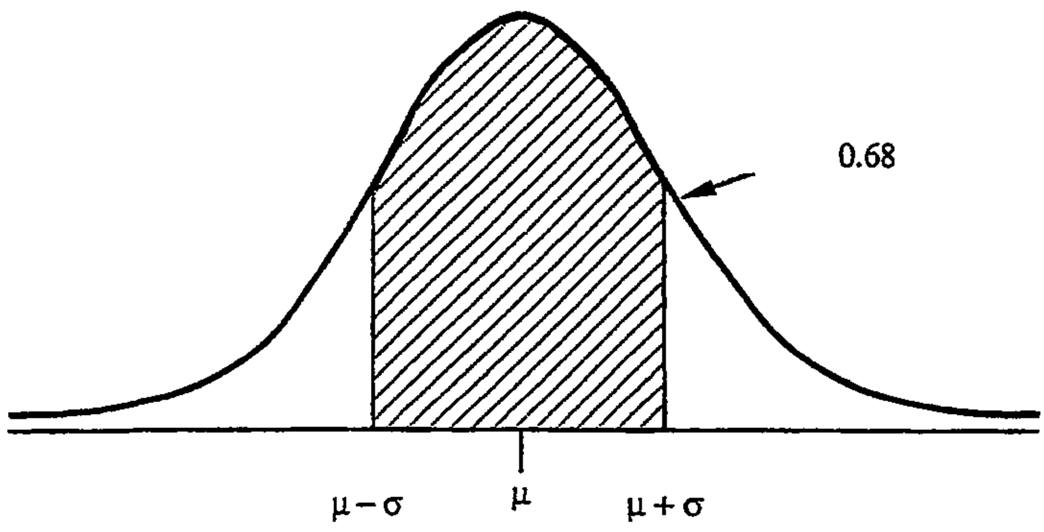


Figure 5.19 : The Area within Two Standard Deviations of the Mean in a Normal Curve.

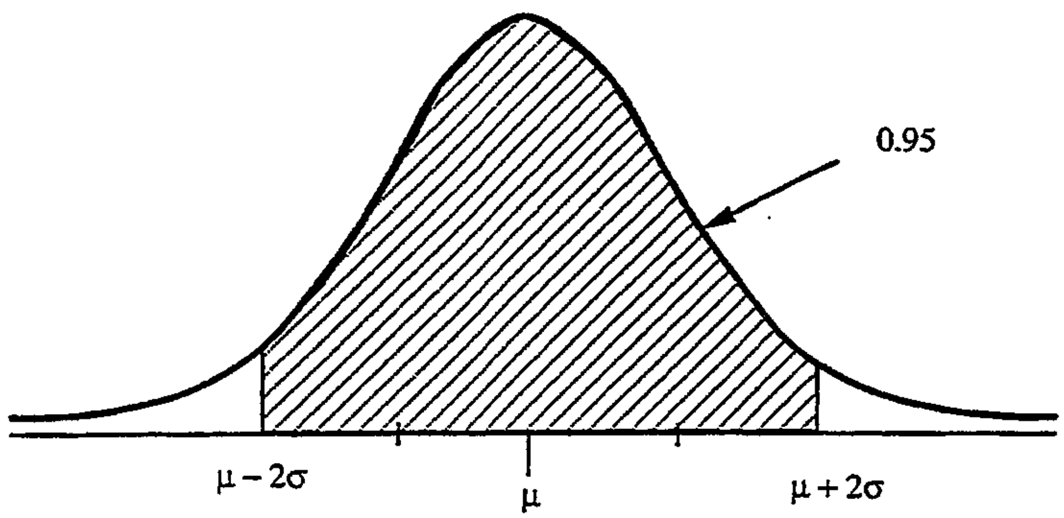


Figure 5.20 : The Area within Three Standard Deviations of the Mean in a Normal Curve.

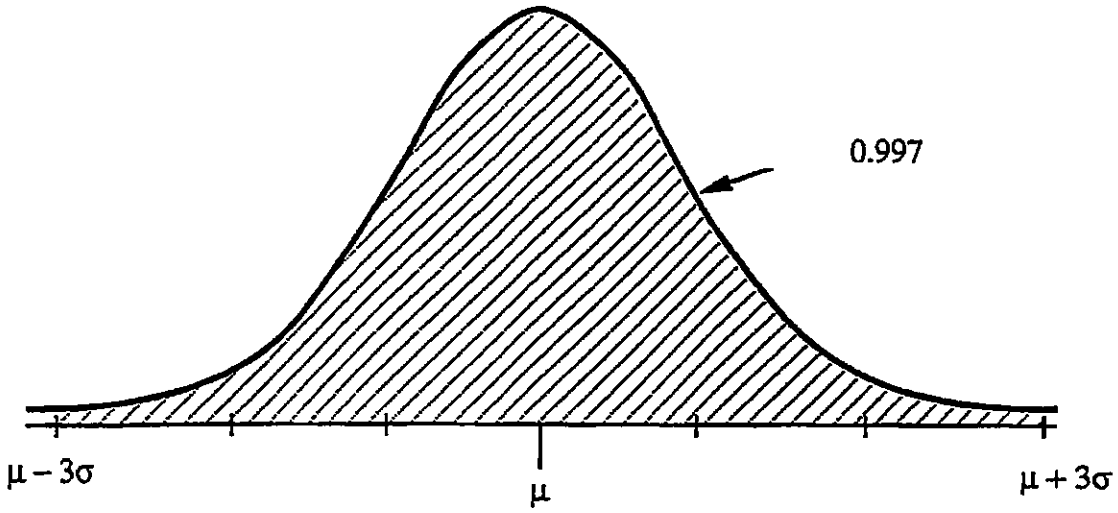


Figure 5.21 : The Area to the Right of $z = 1.52$ in a Standard Normal Distribution.

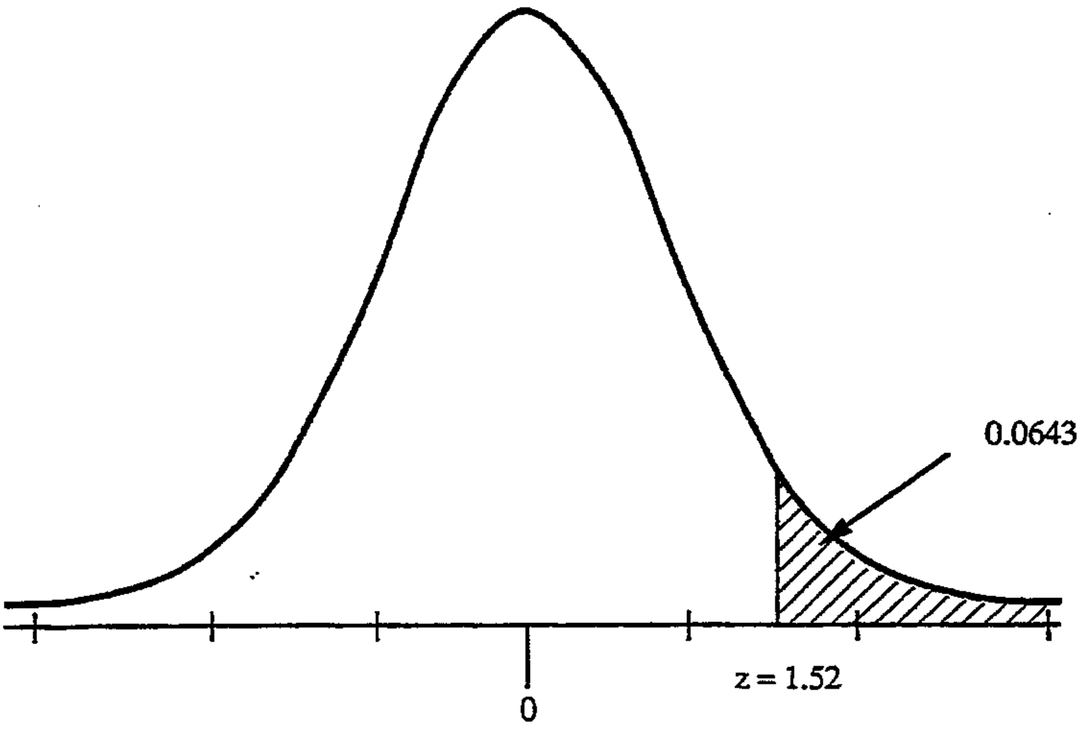


Figure 5.22 : Equal Areas in the Tails of a Standard Normal Distribution.

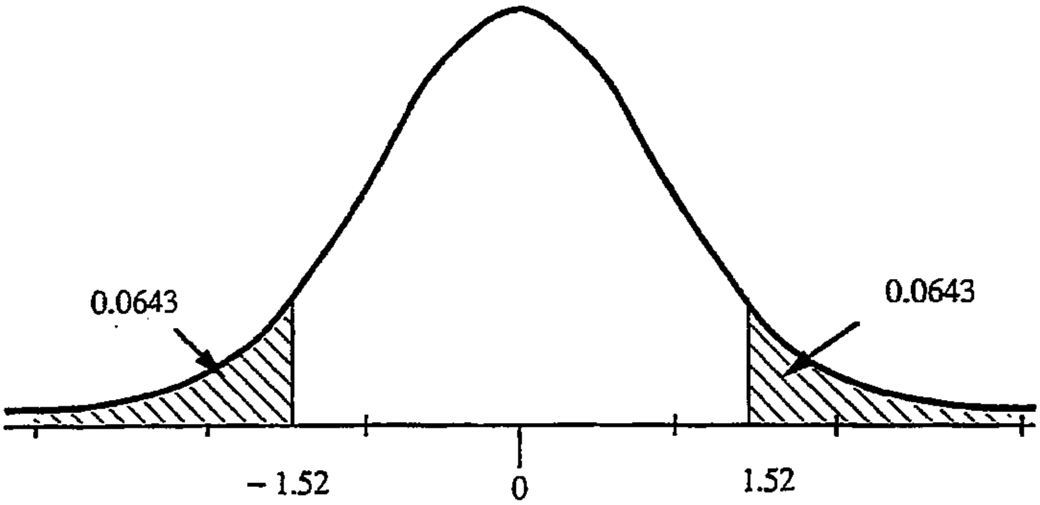


Figure 5.23 : P (- 1.5 < z < 0.5) .

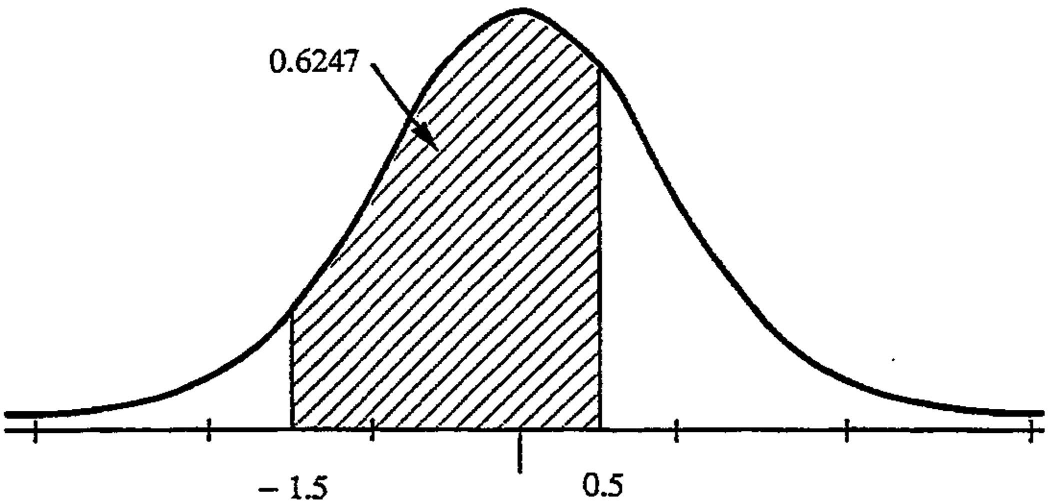


Figure 5.24 : The Probability of a Score less than 640 when $\mu = 600$ and $\sigma = 50$.

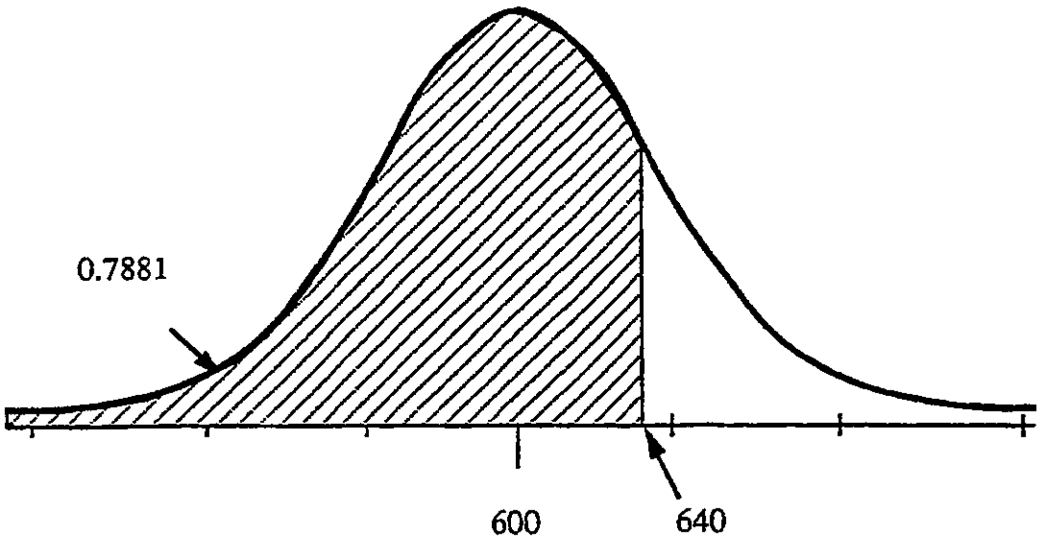


Figure 5.25 : The 90th Percentile of the Normal Distribution with $\mu = 600$ and $\sigma = 50$.

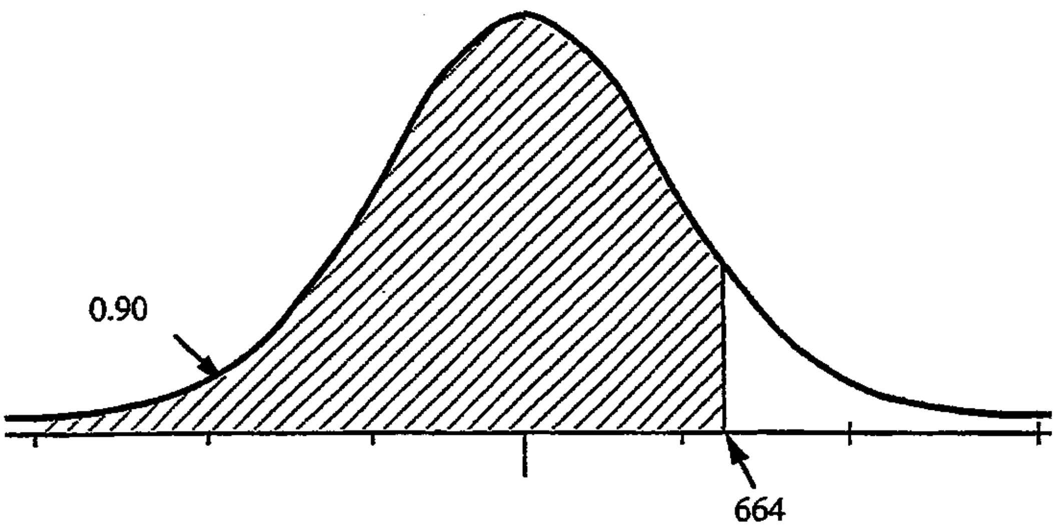


Figure 5.26 : Binomial Distribution with n = 20, p = 0.5, and Corresponding Normal Approximation.

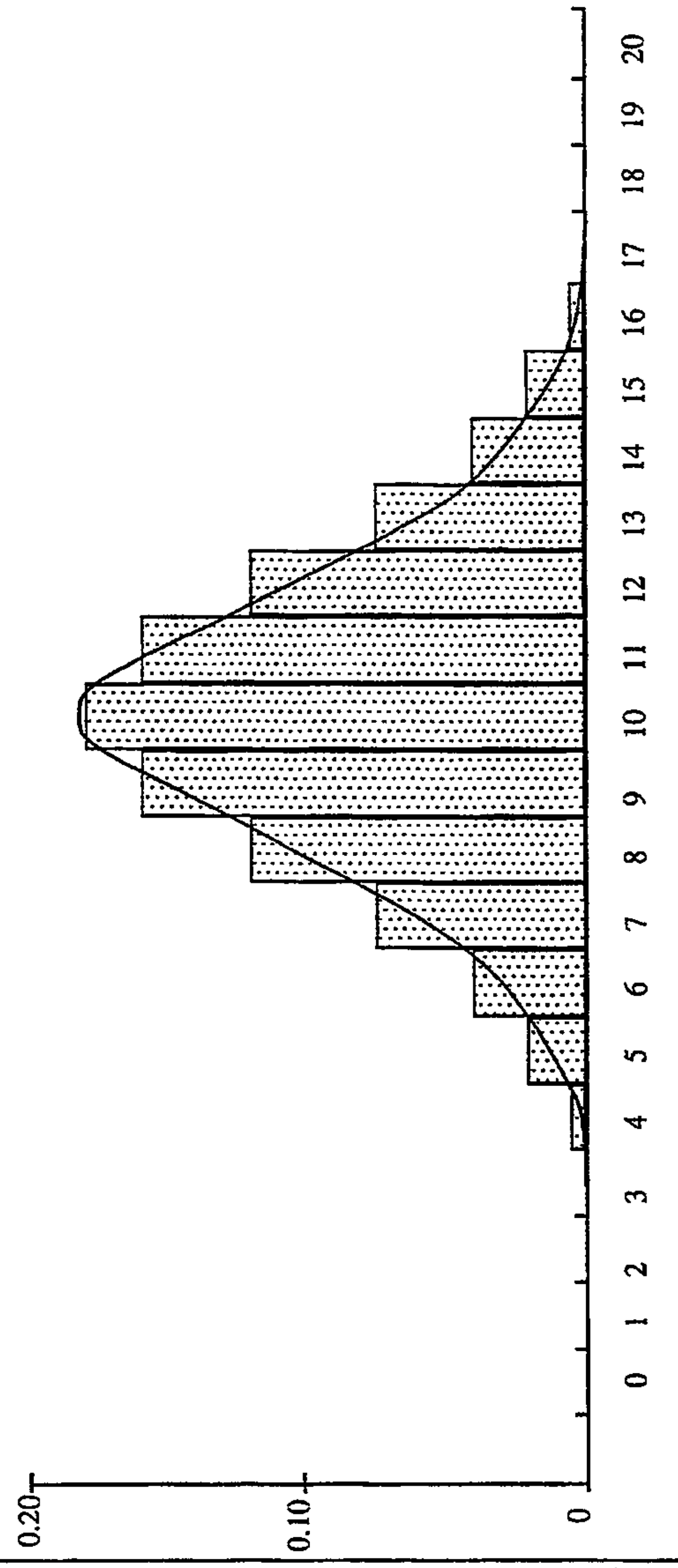


Figure 5.27 : Binomial Distribution with n = 50, p = 0.5, and Corresponding Normal Approximation.

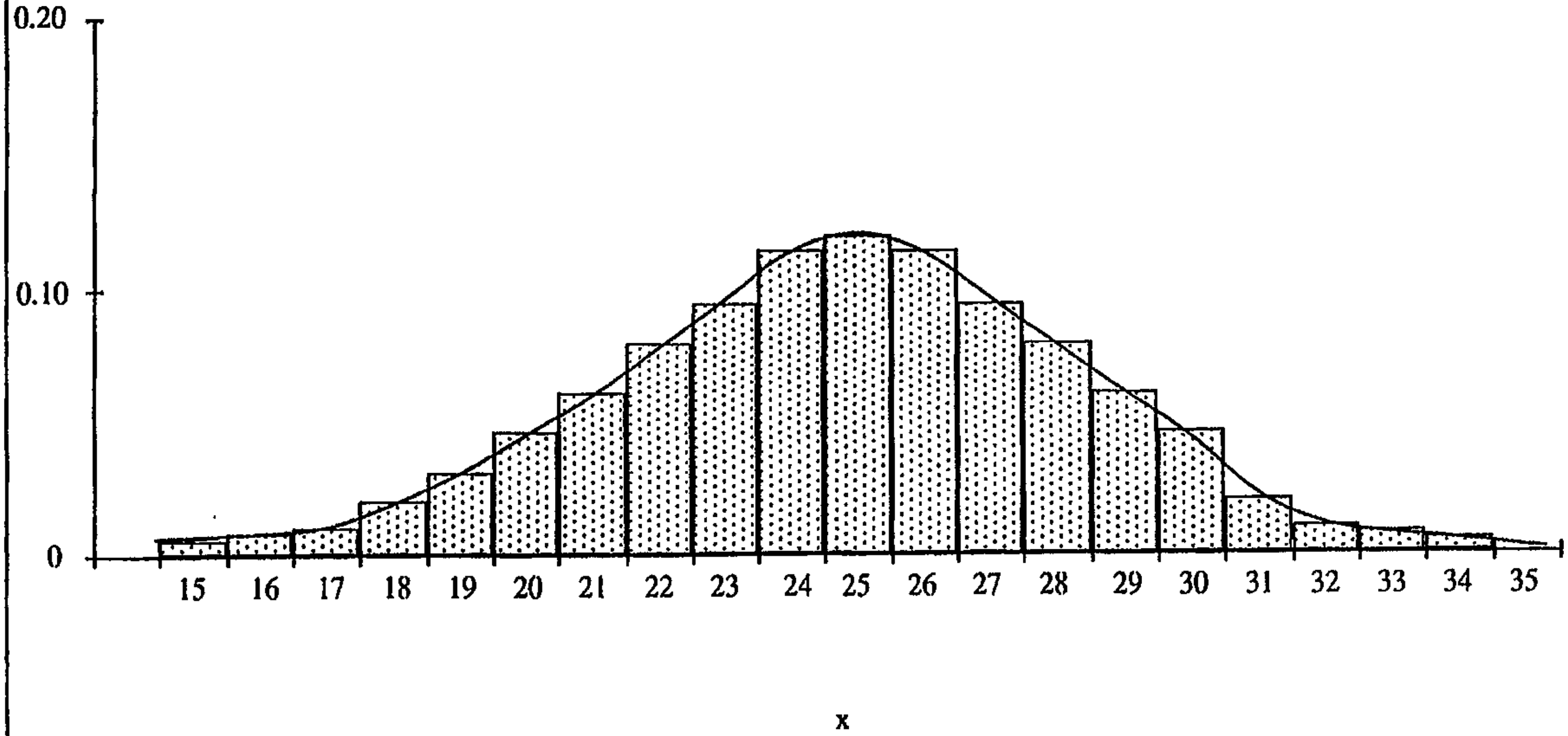


Figure 5.28 : Binomial Distribution with n = 20, p = 0.3, and Corresponding Normal Approximation.

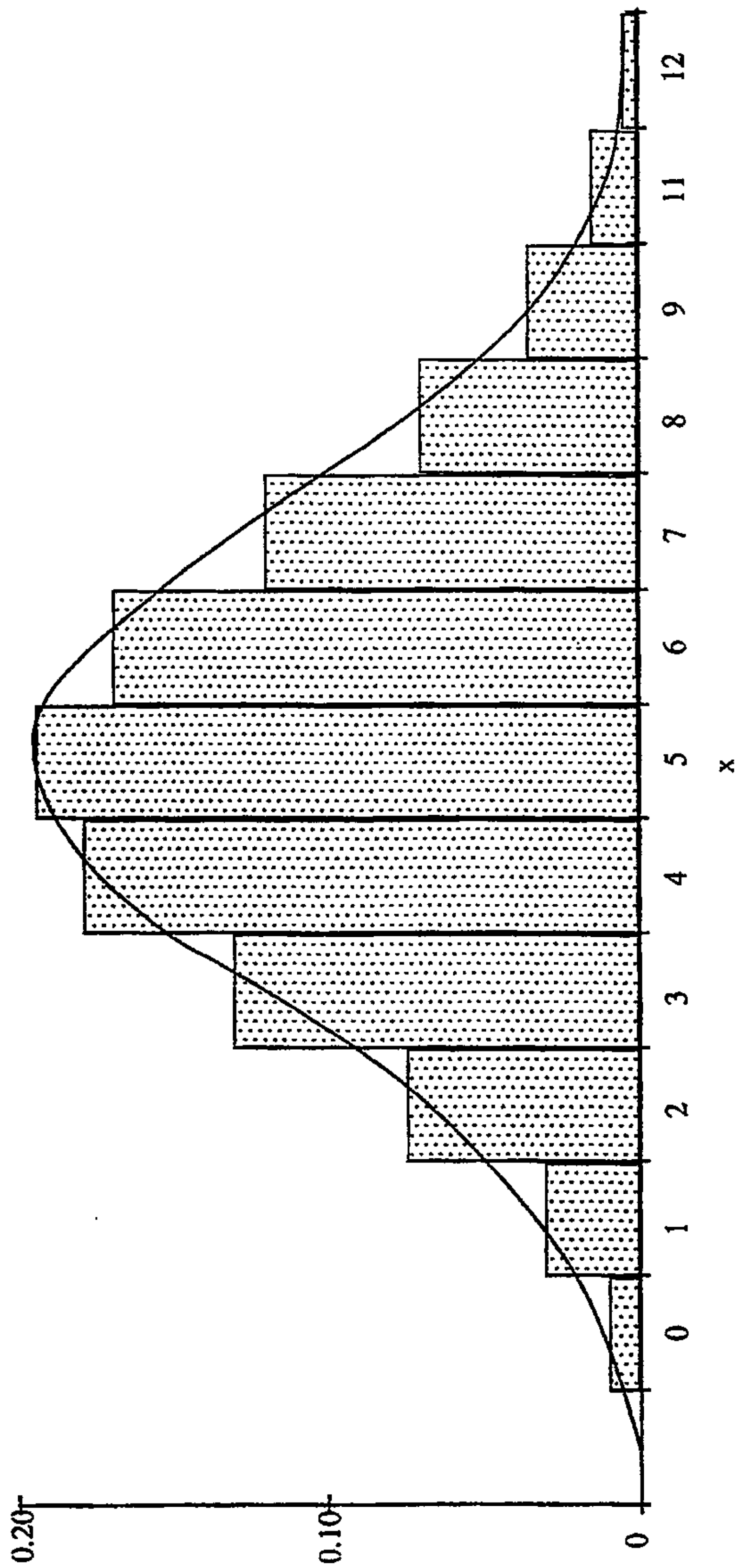


Figure 5.29 : Binomial Distribution with n = 50, p = 0.3, and Corresponding Normal Approximation.