Probability and Statistics (IT302) Class No. 29

21st October 2020 Wednesday 11:15 AM - 11:45 AM

The Normal Approximation of The Binomial Distribution

1. The binomial distribution is applied to a discrete random variable.
2. Each repetition, called a trial, of a binomial experiment results in one of two possible outcomes, either a success or a failure.
3. The probabilities of the two (possible) outcomes remain the same for each repetition of the experiment.
4. The trials are independent.

The binomial formula, which gives the probability of x successes in n trials, is

$$P(x) = {}_{n}C_{x}\, p^{x} q^{n-x}$$

Source : Prem Mann, *Introductory Statistics, 8/E*

The Normal Approximation of The Binomial Distribution Contd.

Usually, the normal distribution is used as an approximation to the binomial distribution when *np* and *nq* are both greater than 5 -- that is, when $np > 5$ and $nq > 5$

Table 6.5 The Binomial Probability Distribution for n = 12 and p = .50

x	$P(x)$
0	.0002
1	.0029
2	.0161
3	.0537
4	.1208
5	.1934
6	.2256
7	.1934
8	.1208
9	.0537
10	.0161
11	.0029
12	.0002

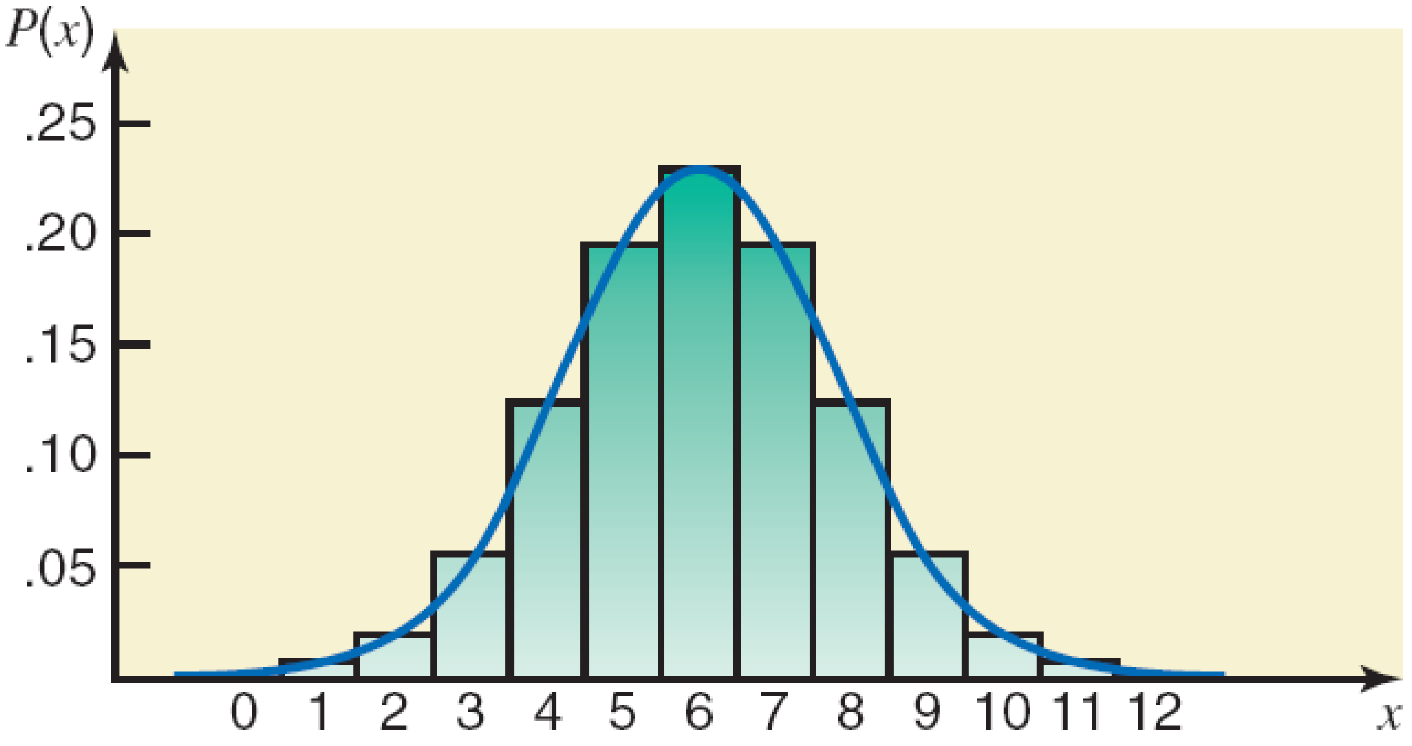


Figure : Histogram for the probability distribution of Table 6.5.

Source : Prem Mann, *Introductory Statistics, 8/E*.

Example 6-20

According to an estimate, 50% of the people in the United States have at least one credit card. If a random sample of 30 persons is selected, what is the probability that 19 of them will have at least one credit card?

Solution: n = 30, p = .50, q = 1 – p = .50
x = 19, n – x = 30 – 19 = 11

From the binomial formula, $P(19) = {}_{30}C_{19}(.5)^{19}(.5)^{11} = .0509$

Let's solve this problem using the normal distribution as an approximation to the binomial distribution. np = 30(.50) = 15 > 5 and nq = 30(.50) = 15 > 5.

We can use the normal distribution as an approximation to solve this binomial problem.

Example 6-20: Solution Contd

Step 1. Compute μ and σ for the binomial distribution.

$$\mu = np = 30(.50) = 15$$

$$\sigma = \sqrt{npq} = \sqrt{30(.50)(.50)} = 2.73861279$$

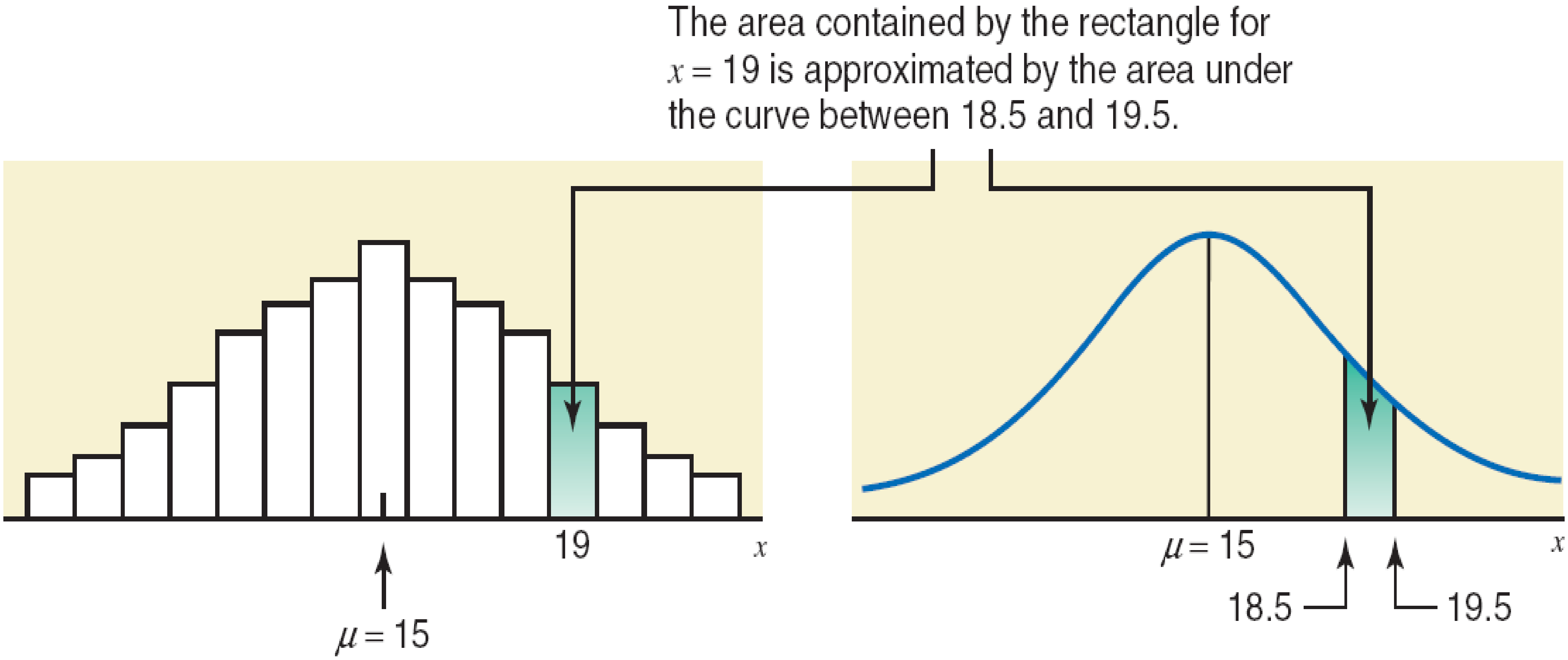
Step 2. Convert the discrete random variable into a continuous random variable (by making the **correction for continuity**).

Continuity Correction Factor Definition

The addition of .5 and/or subtraction of .5 from the value(s) of x when the normal distribution is used as an approximation to the binomial distribution, where x is the number of successes in n trials, is called the continuity correction factor.

Source : Prem Mann, *Introductory Statistics, 8/E*.

Figure 6.51



Source : Prem Mann, *Introductory Statistics, 8/E*

Example 6-20: Solution Contd.

Step 3. Compute the required probability using the normal distribution.

For x = 18.5:

$$z = \frac{18.5 - 15}{2.73861279} = 1.28$$

For x = 19.5:

$$z = \frac{19.5 - 15}{2.73861279} = 1.64$$

$$\begin{aligned} P(18.5 \le x \le 19.5) &= P(1.28 \le z \le 1.64) \\ &= .9495 - .8997 \\ &= .0498 \end{aligned}$$

Source : Prem Mann, *Introductory Statistics, 8/E*.

Example 6-20: Solution Contd.

Thus, based on the normal approximation, the probability that 19 persons in a sample of 30 will have at least one credit card is approximately .0498.

Using the binomial formula, we obtain the exact probability .0509.

The error due to using the normal approximation is .0509 - .0498 = .0011.

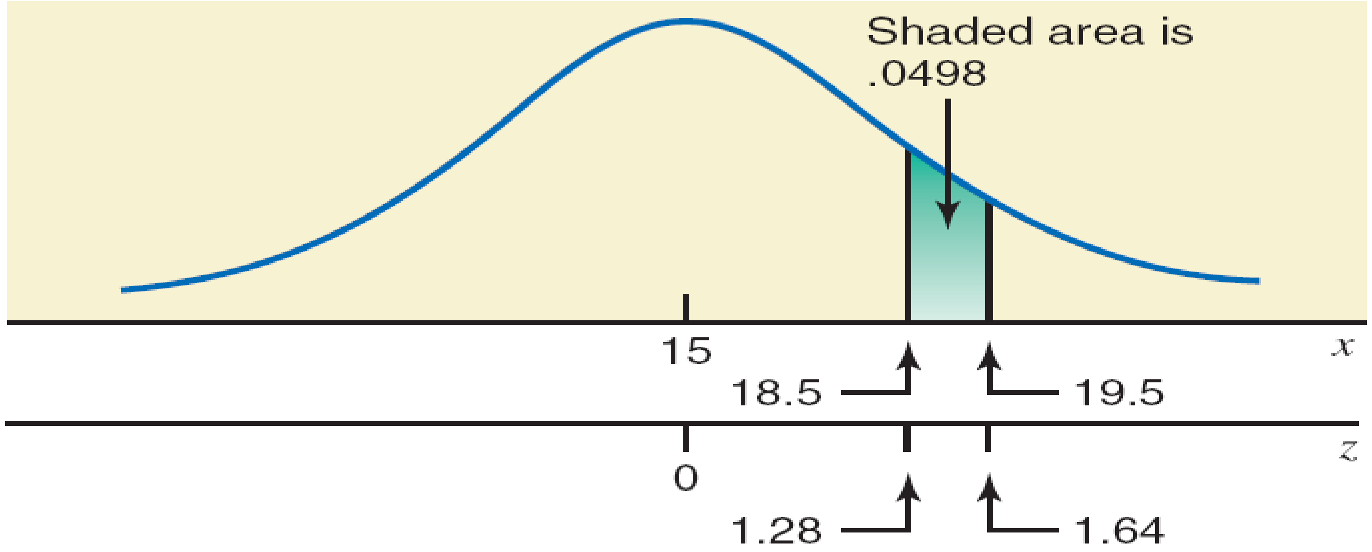


Figure Area between *x* = 18.5 and *x* = 19.5.

Source : Prem Mann, *Introductory Statistics, 8/E*

Example 6-21

According to an Arise Virtual Solutions Job survey, 32% of people working from home said that the biggest advantage of working from home is that there is no commute (*USA TODAY*, October 7, 2011). Suppose that this result is true for the current population of people who work from home. What is the probability that in a random sample of 400 people who work from home, 108 to 122 will say that the biggest advantage of working from home is that there is no commute?

Source : Prem Mann, *Introductory Statistics, 8/E*

Example 6-21: Solution

n = 400, p = .32, q = 1 – .32 = .68

$$\mu = np = 400(.32) = 128$$

$$\sigma = \sqrt{npq} = \sqrt{400(.32)(.68)} = 9.32952303$$

For x = 107.5:

$$z = \frac{107.5 - 128}{9.32952303} = -2.20$$

For x = 122.5

$$z = \frac{122.5 - 128}{9.32952303} = -.59$$

S**ource** : Prem Mann, *Introductory Statistics, 8/E*

Example 6-21: Solution Contd.

$$P(107.5 \le x \le 122.5) = P(-2.20 \le z \le -.59)$$
$$= .2776 - .0139 = .2637$$

Thus, the probability that 108 to 122 people in a sample of 400 who work from home will say that the biggest advantage of working from home is that there is no commute is approximately .2637.

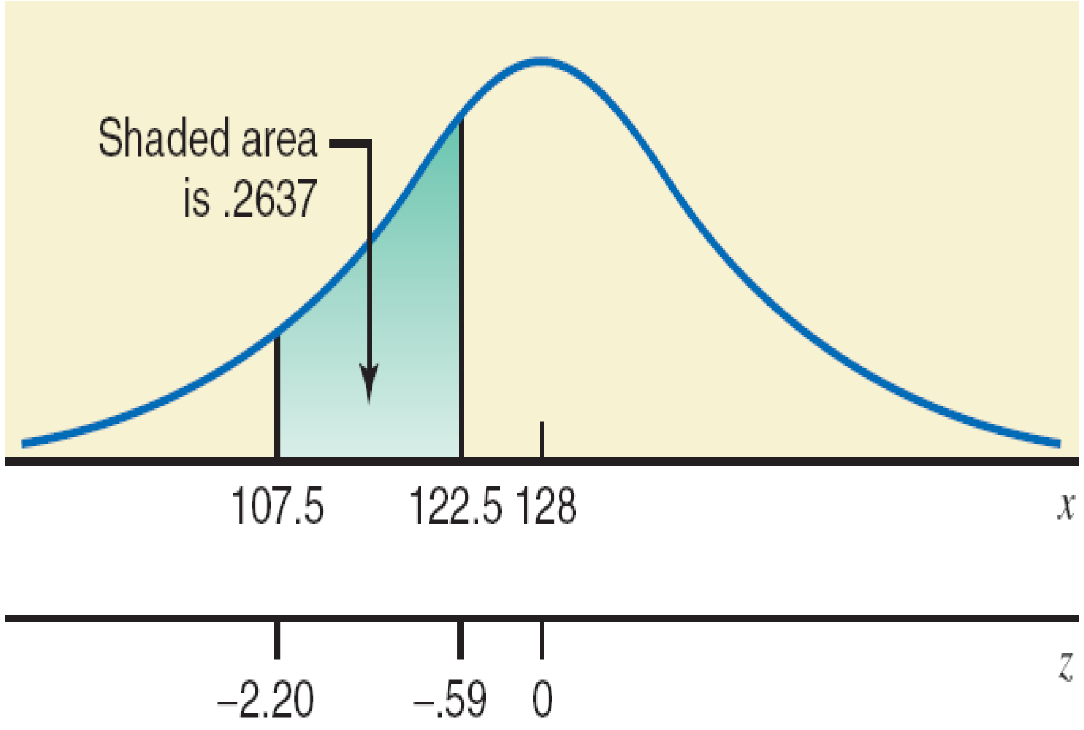


Figure Area between x = 107.5 and x = 122.5

Source : Prem Mann, *Introductory Statistics, 8/E*

Example 6-22

According to a poll, 55% of American adults do not know that GOP stands for Grand Old Party (*Time*, October 17, 2011). Assume that this percentage is true for the current population of American adults. What is the probability that 397 or more American adults in a random sample of 700 do not know that GOP stands for Grand Old Party?

Solution: n = 700, p = .55, q = 1 – .55 = .45

$$\mu = np = 700(.55) = 385$$
$$\sigma = \sqrt{npq} = \sqrt{700(.55)(.45)} = 13.16244658$$

For x = 396.5: $z = \dfrac{396.5 - 385}{13.16244658} = .87$

P(x ≥ 396.5) = P(z ≥ .87) = 1.0 - .8078 = .1922

Thus, the probability that 397 or more American adults in a random sample of 700 will not know that GOP stands for Grand Old Party is approximately .1922.

Source : Prem Mann, *Introductory Statistics, 8/E*

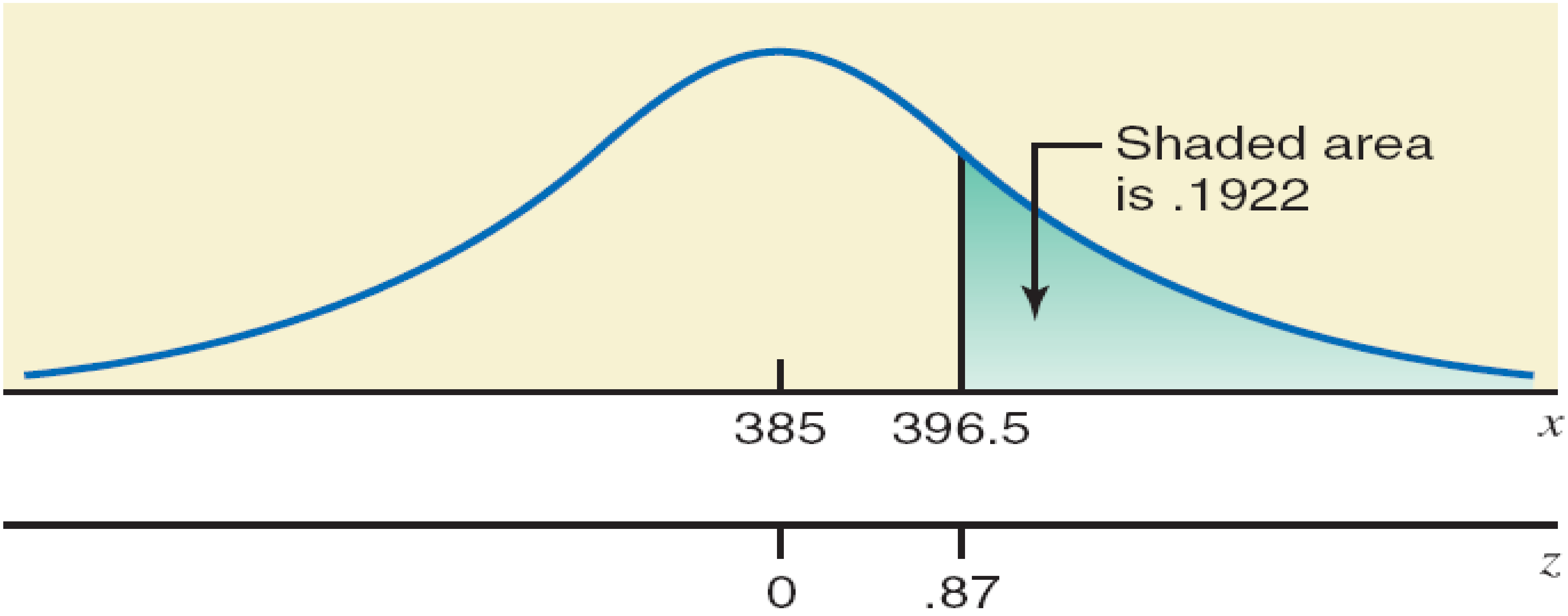


Figure Area to the right of x = 396.5

Source : Prem Mann, *Introductory Statistics, 8/E*

Additional Material

Binomial Random Variable

A binomial random variable represents the number of successes in a series of independent trials. The sample proportion is found by dividing the number of successes by the number of trials. Since the sample proportion is approximately normally distributed whenever $np \geq 10$ and $n(1 - p) \geq 10$, the number of successes is also approximately normally distributed under these conditions. Therefore, the normal curve can also be used to compute approximate probabilities for the binomial distribution.

Recall that if X is a binomial random variable with n trials and success probability p, then:

- The mean of X is $\mu_x = np$
- The variance of X is $\sigma_x^2 = np(1-p)$
- The standard deviation of X is $\sigma_x = \sqrt{np(1-p)}$

Source: https://www.cpp.edu/~jcwindley/classes/sta1200/powerpoints/Section_06_04__ESS_STATS2e.pptx

Normal Approximation

Binomial probabilities can be computed exactly using the techniques described earlier. If the number of trials is large, using these methods by hand is extremely difficult because many terms have to be calculated and added together. If the following conditions are met, binomial probabilities can be approximated using a **normal distribution**.

Let X be a binomial random variable with n trials and success probability p. If $np \geq 10$ and $n(1-p) \geq 10$, then X is approximately normal with mean $\mu_x = np$ and standard deviation $\sigma_x = \sqrt{np(1-p)}$.

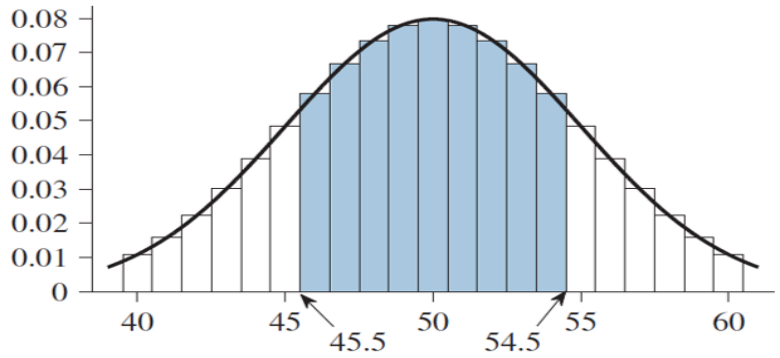
The binomial distribution is **discrete**, whereas the normal distribution is **continuous**. The **continuity correction** is an adjustment, made when approximating a discrete distribution with a continuous one, which can improve the accuracy of the approximation.

Source: https://www.cpp.edu/~jcwindley/classes/sta1200/powerpoints/Section_06_04__ESS_STATS2e.pptx

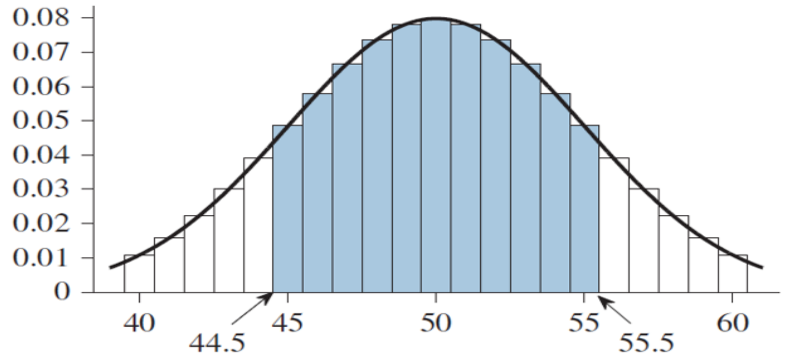
Continuity Correction

Suppose a fair coin is tossed 100 times, let X represent the number of heads that result. Then X has a binomial distribution with $n = 100$ trials and success probability $p = 0.5$. If we wanted to compute the probability X is between 45 and 55 [i.e. $P(45 < X < 55)$], the probability will differ depending on whether the endpoints 45 and 55 are included.

To compute $P(45 < X < 55)$, the areas of the rectangles corresponding to 45 and to 55 should be excluded. To approximate this probability with the normal curve, compute the area under the curve between **45.5** and **54.5**.



To compute $P(45 \leq X \leq 55)$, the areas of the rectangles corresponding to 45 and to 55 should be included. To approximate this probability with the normal curve, compute the area under the curve between **44.5** and **55.5**.



Source: https://www.cpp.edu/~jcwindley/classes/sta1200/powerpoints/Section_06_04__ESS_STATS2e.pptx

Example 1: Normal Approximation

Let X be the number of heads that appear when a fair coin is tossed 100 times. Use the normal curve to find $P(45 \leq X \leq 55)$.

Solution:

The number of trials is $n = 100$. Since the coin is fair, the success probability is $p = 0.5$. Therefore, $np = (100)(0.5) = 50 \geq 10$ and $n(1 - p) = (100)(1 - 0.5) = 50 \geq 10$. We may use the normal approximation.

We compute the mean and standard deviation of X: $\mu_x = np = (100)(0.5) = 50$

$$\sigma_x = \sqrt{np(1-p)} = \sqrt{(100)(0.5)(1-0.5)} = 5$$

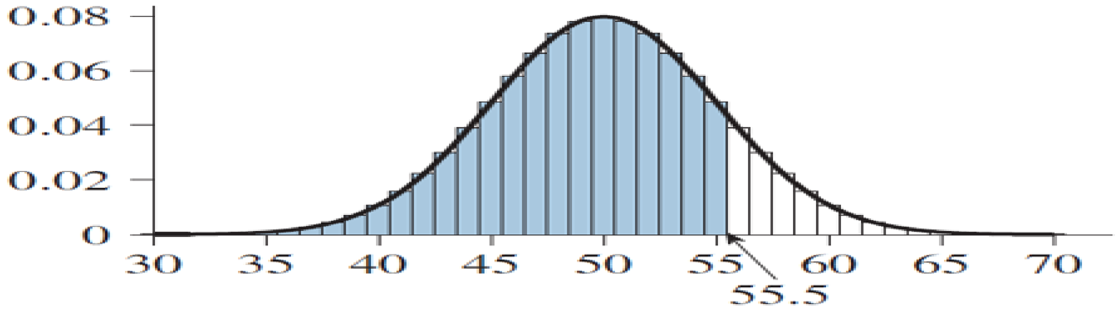
Because the probability is $P(45 \leq X \leq 55)$, we want to *include* both 45 and 55. Therefore, we set the left endpoint to 44.5 and the right endpoint to 55.5.We use tables or technology to find that the probability is 0.7287.

Source: https://www.cpp.edu/~jcwindley/classes/sta1200/powerpoints/Section_06_04__ESS_STATS2e.pptx

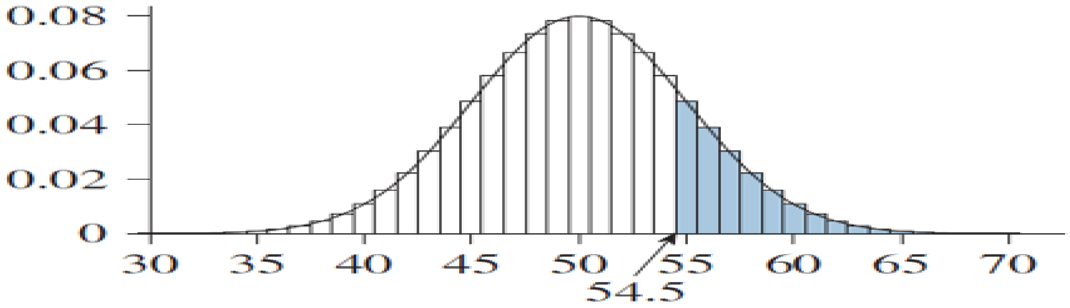
Example: Continuity Correction

A fair coin is tossed 100 times. Let X be the number of heads that appear.

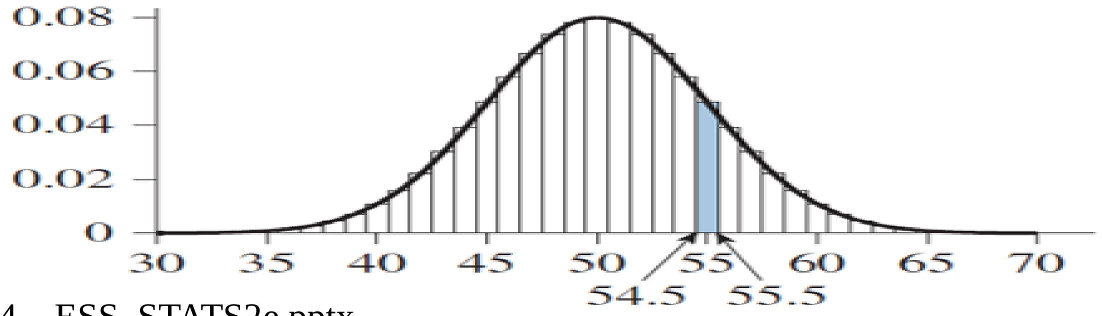
To approximate $P(X \leq 55)$, we would find the area to the left of 55.5.



To approximate $P(X \geq 55)$, we would find the area to the right of 54.5.

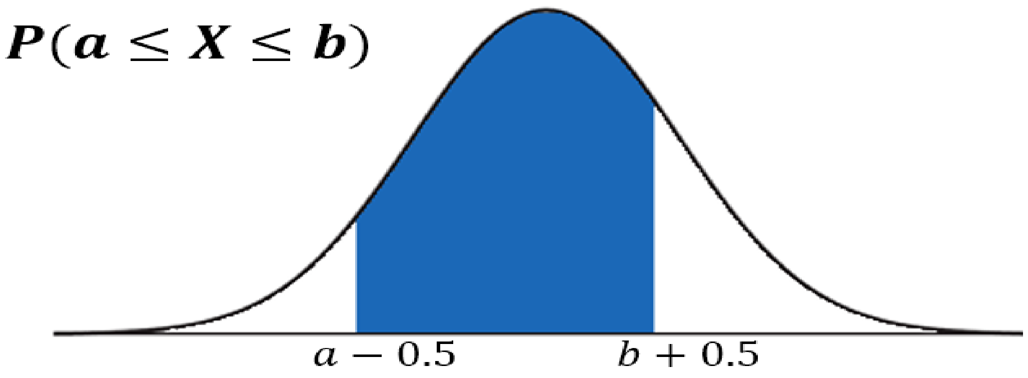


To approximate $P(X = 55)$, we would find the area between 54.5 and 55.5.

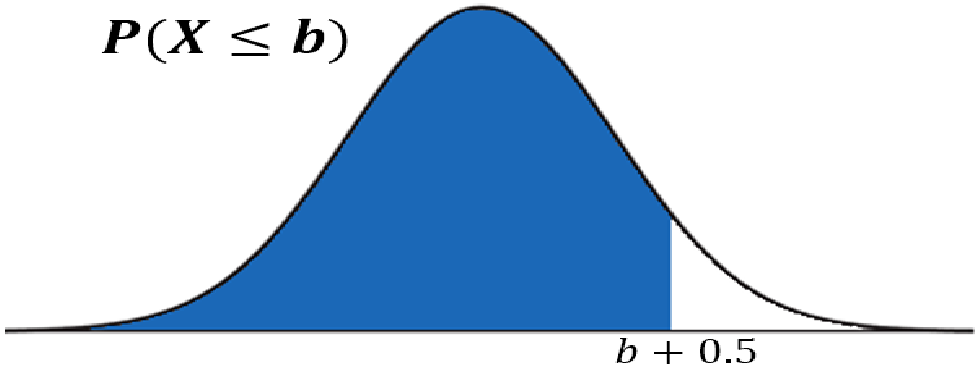


Source: https://www.cpp.edu/~jcwindley/classes/sta1200/powerpoints/Section_06_04__ESS_STATS2e.pptx

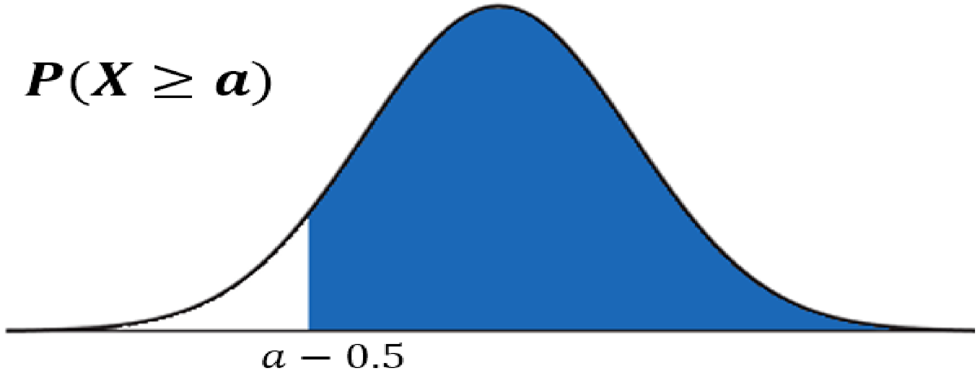
Areas to Use When the Continuity Correction is Applied



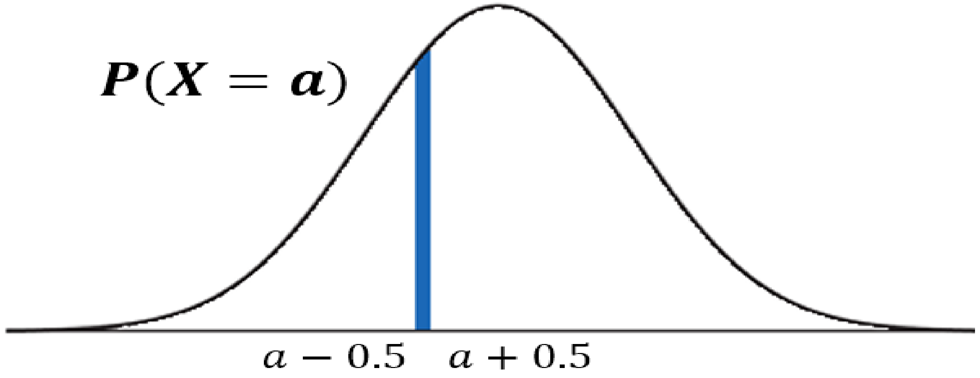
Find the area between $a - 0.5$ and $b + 0.5$



Find the area to the left of $b + 0.5$



Find the area to the right of $a - 0.5$



Find the area between $a - 0.5$ and $a + 0.5$

Source: https://www.cpp.edu/~jcwindley/classes/sta1200/powerpoints/Section_06_04__ESS_STATS2e.pptx

Example 2: Normal Approximation

The Statistical Abstract of the United States reported that 66% of students who graduated from high school enrolled in college. One hundred high school graduates are sampled. Let X be the number who enrolled in college. Find $P(X \leq 75)$.

Solution:

The number of trials is $n = 100$ and the success probability is $p = 0.66$. Therefore, $np = (100)(0.66) = 66 \geq 10$ and $n(1-p) = (100)(1-0.66) = 34 \geq 10$. We may use the normal approximation.

We compute the mean and standard deviation of X:

$$\mu_x = np = (100)(0.66) = 66$$

$$\sigma_x = \sqrt{np(1-p)} = \sqrt{(100)(0.66)(1-0.66)} = 4.73709$$

Because the probability is $P(X \leq 75)$, we compute the area to the left of 75.5. Using tables or technology, we find that the probability is 0.9775.

Source: https://www.cpp.edu/~jcwindley/classes/sta1200/powerpoints/Section_06_04__ESS_STATS2e.pptx

Continuity Corrections Procedures

1. When using the normal distribution as an approximation to the binomial distribution, always use the continuity correction.

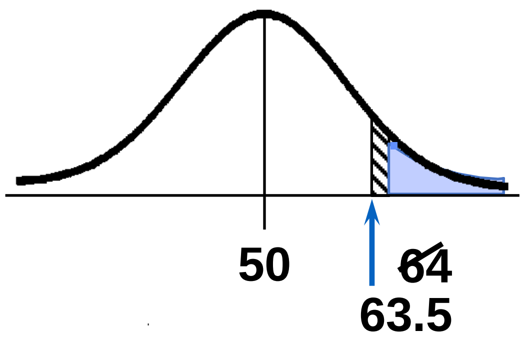
2. In using the continuity correction, first identify the discrete whole number *x* that is relevant to the binomial probability problem.

3. Draw a normal distribution centered about μ, then draw a vertical strip area centered over *x* . Mark the left side of the strip with the number *x* - 0.5, and mark the right side with *x* + 0.5. For *x* = 64, draw a strip from 63.5 to 64.5. Consider the area of the strip to represent the probability of discrete number *x*.

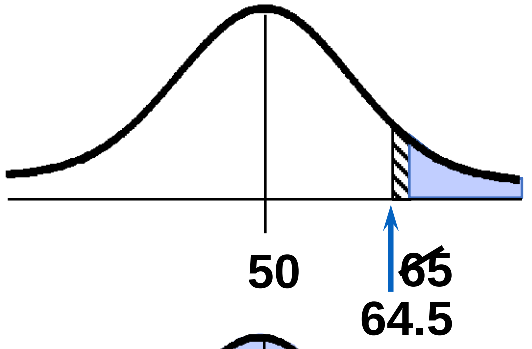
4. <u>Now determine whether the value of x itself should be included in the probability you want. Next, determine whether you want the probability of at least x, at most x, more than x, fewer than x, or exactly x. Shade the area to the right of left of the strip, as appropriate; also shade the interior of the strip itself if and only if x itself is to be included, The total shaded region corresponds to probability being sought.</u>

Source: http://www.stat.rutgers.edu/~mxie/stat211/ftp/lect5-4.ppt

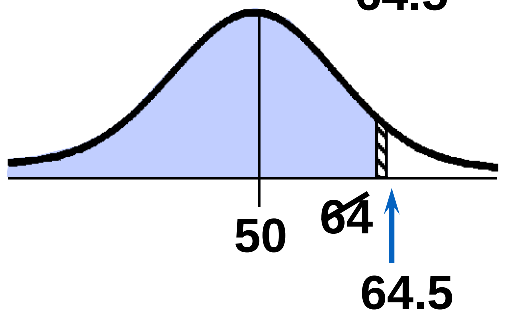
X = **at least** 64
= 64, 65, 66, . . .



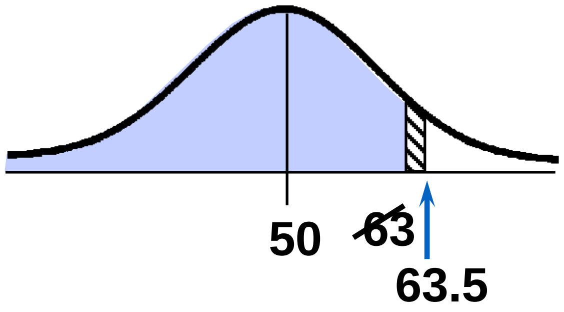
X = **more than** 64
= 65, 66, 67, . . .



X = **at most** 64
= 0, 1, . . . 62, 63, 64

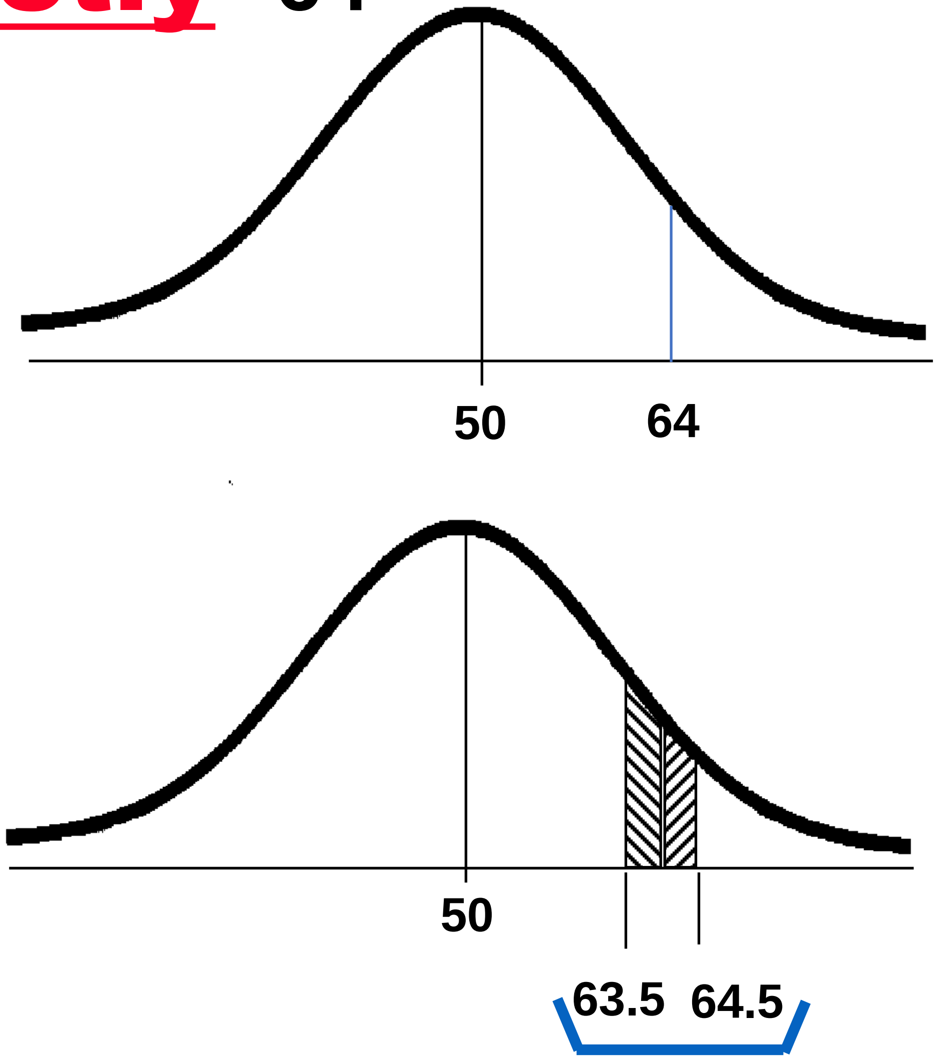


X = **fewer than** 64
= 0, 1, . . . 62, 63



Source: http://www.stat.rutgers.edu/~mxie/stat211/ftp/lect5-4.ppt

x = <u>**exactly**</u> 64



Interval represents discrete number 64

Source: http://www.stat.rutgers.edu/~mxie/stat211/ftp/lect5-4.ppt