

The Normal Distribution

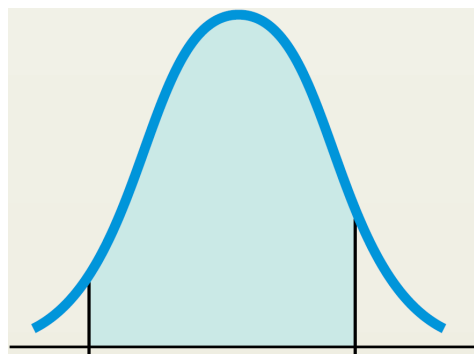
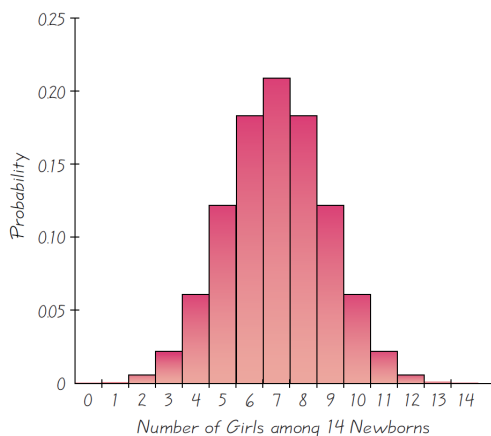
Continuous Distributions

A **continuous random** variable is a variable whose possible values form some interval of numbers. Typically, a continuous variable involves a measurement of something, such as the height of a person, the weight of a newborn baby, or the length of time a car battery lasts.

EXAMPLES:

1. Let x = the amount of milk a cow produces in one day. This is a continuous random variable because it can have any value over a continuous span. During a single day, a cow might yield an amount of milk that can be any value between 0 gallons and 5 gallons. It would be possible to get 4.123456 gallons, because the cow is not restricted to the discrete amounts of 0, 1, 2, 3, 4, or 5 gallons.
2. The measure of voltage for a smoke-detector battery can be any value between 0 volts and 9 volts. It is therefore a continuous random variable.

In the histograms we have seen thus far, the frequencies, percentages, proportions, or probabilities were represented by the heights of the rectangles, or by their areas. In the continuous case, we also represent probabilities by areas — not by areas of rectangles, but by areas under continuous curves.



Continuous curves such as the one shown on the right are the graphs of functions called **probability densities**, or informally, continuous distributions. Probability densities are characterized by the fact that the area under the curve between any two values a and b gives the probability that a random variable having this continuous distribution will take on a value on the interval from a to b .

The total area under the curve must equal 1

EXAMPLE: Verify that $f(x) = \frac{x}{8}$ can serve as the probability density of a random variable defined over the interval from $x = 0$ to $x = 4$. Then find the probabilities that a random variable having the given probability density will take on a value

(a) less than 2;

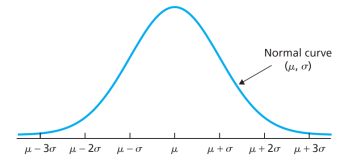
(b) less than or equal to 2.

The Normal Distribution

Normal distributions are extremely important because they occur so often in real applications and they play such an important role in methods of inferential statistics.

If a continuous random variable has a distribution with a graph that is symmetric and bell-shaped, as in the Figure on the right, and it can be described by the function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

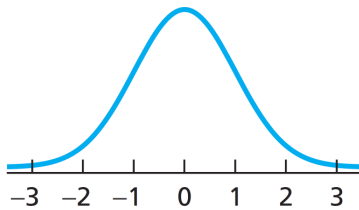


we say that it has a normal distribution.

Standardizing a Normally Distributed Variable

Now the question is: How do we find areas under a normal curve? Conceptually, we need a table of areas for each normal curve. This, of course, is impossible because there are infinitely many different normal curves — one for each choice of μ and σ . The way out of this difficulty is standardizing, which transforms every normal distribution into one particular normal distribution, the **standard normal distribution**.

A normally distributed variable having mean 0 and standard deviation 1 is said to have the **standard normal distribution**. Its associated normal curve is called the **standard normal curve**, which is shown in the Figure below.



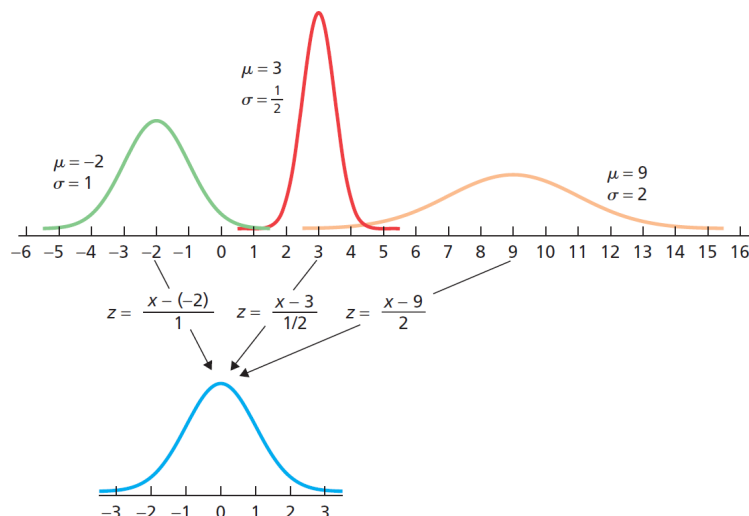
Standardized Normally Distributed Variable

The standardized version of a normally distributed variable x ,

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

has the standard normal distribution.

We can interpret the above statement in several ways. Theoretically, it says that standardizing converts all normal distributions to the standard normal distribution, as depicted in the Figure below.



Basic Properties of the Standard Normal Curve:

Property 1: The total area under the standard normal curve is 1.

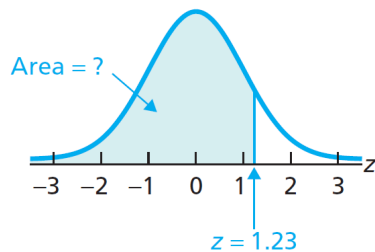
Property 2: The standard normal curve extends indefinitely in both directions, approaching, but never touching, the horizontal axis as it does so.

Property 3: The standard normal curve is symmetric about 0.

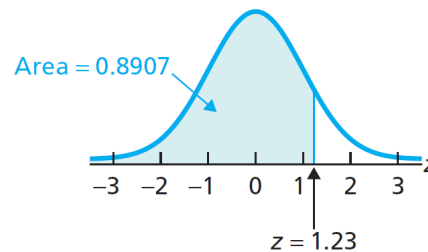
Property 4: Almost all the area under the standard normal curve lies between -3 and 3.

Areas Under the Standard Normal Curve

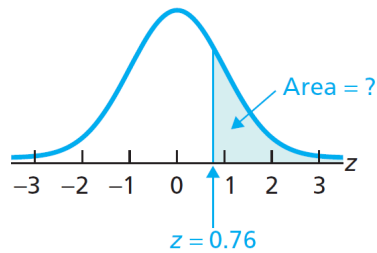
EXAMPLES: Determine the area under the standard normal curve.



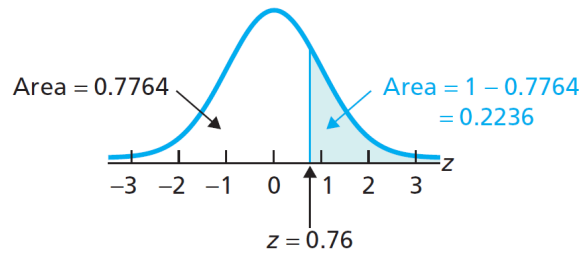
(a)



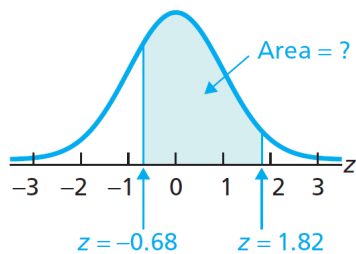
(b)



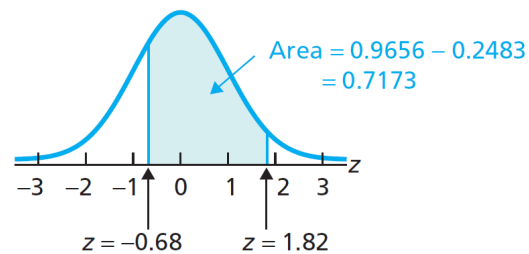
(a)



(b)



(a)



(b)

EXAMPLE: Find z if

(a) the standard-normal-curve area between 0 and z is 0.4484

(b) the standard-normal-curve area to the left of z is 0.9868

Solution:

(a) $z = 1.63$

(b) $0.9868 - 0.5000 = 0.4868 \implies z = 2.22$

EXAMPLE: Find z if the standard-normal-curve area

(a) between 0 and z is 0.4788

(b) to the left of z is 0.8365

(c) between $-z$ and z is 0.8584

(d) to the left of z is 0.3409

EXAMPLE: Find z if the standard-normal-curve area

(a) between 0 and z is 0.4788

(b) to the left of z is 0.8365

(c) between $-z$ and z is 0.8584

(d) to the left of z is 0.3409

Solution:

(a) $z = 2.03$

(b) $0.8365 - 0.5 = 0.3365 \implies z = 0.98$

(c) $0.8584/2 = 0.4292 \implies z = 1.47$

(d) $0.5 - 0.3409 = 0.1591 \implies z = -0.41$

EXAMPLE: If a random variable has the normal distribution with $\mu = 82.0$ and $\sigma = 4.8$, find the probabilities that it will take on a value

(a) less than 89.2

(b) greater than 78.4

(c) between 83.2 and 88.0

(d) between 73.6 and 90.4

EXAMPLE: If a random variable has the normal distribution with $\mu = 82.0$ and $\sigma = 4.8$, find the probabilities that it will take on a value

(a) less than 89.2

(b) greater than 78.4

(c) between 83.2 and 88.0

(d) between 73.6 and 90.4

Solution:

(a) We have

$$z = \frac{89.2 - 82}{4.8} = 1.5$$

therefore the probability is $0.4332 + 0.5 = 0.9332$.

(b) We have

$$z = \frac{78.4 - 82}{4.8} = -0.75$$

therefore the probability is $0.2734 + 0.5 = 0.7734$.

(c) We have

$$z_1 = \frac{83.2 - 82}{4.8} = 0.25 \quad \text{and} \quad z_2 = \frac{88 - 82}{4.8} = 1.25$$

therefore the probability is $0.3944 - 0.0987 = 0.2957$.

(d) We have

$$z_1 = \frac{73.6 - 82}{4.8} = -1.75 \quad \text{and} \quad z_2 = \frac{90.4 - 82}{4.8} = 1.75$$

therefore the probability is $0.4599 + 0.4599 = 0.9198$.

Applications of the Normal Distribution

EXAMPLE: Intelligence quotients (IQs) measured on the Stanford Revision of the Binet-Simon Intelligence Scale are normally distributed with a mean of 100 and a standard deviation of 16. Determine the percentage of people who have IQs between 115 and 140.

Solution: We have

$$z_1 = \frac{115 - 100}{16} = 0.9375 \quad \text{and} \quad z_2 = \frac{140 - 100}{16} = 2.5$$

therefore the probability is $0.4938 - 0.3264 = 0.1674$. It follows that 16.74% of all people have IQs between 115 and 140. Equivalently, the probability is 0.1674 that a randomly selected person will have an IQ between 115 and 140.

EXAMPLE: One of the larger species of tarantulas is the *Grammostola mollicoma*, whose common name is the Brazilian giant tawny red. A tarantula has two body parts. The anterior part of the body is covered above by a shell, or carapace. From a recent article by F. Costa and F. Perez-Miles titled Reproductive Biology of Uruguayan Theraphosids (*The Journal of Arachnology*, Vol. 30, No. 3, pp. 571-587), we find that the carapace length of the adult male *G. mollicoma* is normally distributed with mean 18.14 mm and standard deviation 1.76 mm.

(a) Find the percentage of adult male *G. mollicoma* that have carapace length between 16 mm and 17 mm.

(b) Find the percentage of adult male *G. mollicoma* that have carapace length exceeding 19 mm.

EXAMPLE: One of the larger species of tarantulas is the *Grammostola mollicoma*, whose common name is the Brazilian giant tawny red. A tarantula has two body parts. The anterior part of the body is covered above by a shell, or carapace. From a recent article by F. Costa and F. Perez-Miles titled Reproductive Biology of Uruguayan Theraphosids (*The Journal of Arachnology*, Vol. 30, No. 3, pp. 571-587), we find that the carapace length of the adult male *G. mollicoma* is normally distributed with mean 18.14 mm and standard deviation 1.76 mm.

(a) Find the percentage of adult male *G. mollicoma* that have carapace length between 16 mm and 17 mm.

(b) Find the percentage of adult male *G. mollicoma* that have carapace length exceeding 19 mm.

Solution:

(a) We have

$$z_1 = \frac{16 - 18.14}{1.76} \approx -1.22 \quad \text{and} \quad z_2 = \frac{17 - 18.14}{1.76} \approx -0.65$$

therefore the probability is $0.3888 - 0.2422 = 0.1466$. It follows that 14.66% of adult male *G. mollicoma* have carapace length between 16 mm and 17 mm. Equivalently, the probability is 0.1466 that a randomly selected adult male *G. mollicoma* has carapace length between 16 mm and 17 mm.

(b) We have

$$z = \frac{19 - 18.14}{1.76} \approx 0.49$$

therefore the probability is $0.5 - 0.1879 = 0.3121$. It follows that 31.21% of adult male *G. mollicoma* have carapace length exceeding 19 mm. Equivalently, the probability is 0.3121 that a randomly selected adult male *G. mollicoma* has carapace length exceeding 19 mm.

EXAMPLE: As reported in *Runner's World* magazine, the times of the finishers in the New York City 10-km run are normally distributed with mean 61 minutes and standard deviation 9 minutes.

(a) Determine the percentage of finishers with times between 50 and 70 minutes.

(b) Determine the percentage of finishers with times less than 75 minutes.

EXAMPLE: As reported in *Runner's World* magazine, the times of the finishers in the New York City 10-km run are normally distributed with mean 61 minutes and standard deviation 9 minutes.

(a) Determine the percentage of finishers with times between 50 and 70 minutes.

(b) Determine the percentage of finishers with times less than 75 minutes.

Solution:

(a) We have

$$z_1 = \frac{50 - 61}{9} \approx -1.22 \quad \text{and} \quad z_2 = \frac{70 - 61}{9} = 1$$

therefore the probability is $0.3888 + 0.3413 = 0.7301 \implies 73.01\%$.

(b) We have

$$z = \frac{75 - 61}{9} = 1.5556$$

therefore the probability is $0.5 + 0.4406 = 0.9406 \implies 94.06\%$.

EXAMPLE: An article by S. M. Berry titled "Drive for Show and Putt for Dough" (*Chance*, Vol. 12(4), pp. 50-54) discussed driving distances of PGA players. The mean distance for tee shots on the 1999 men's PGA tour is 272.2 yards with a standard deviation of 8.12 yards. Assuming that the 1999 tee-shot distances are normally distributed, find the percentage of such tee shots that went

(a) between 260 and 280 yards.

(b) more than 300 yards.

EXAMPLE: An article by S. M. Berry titled “Drive for Show and Putt for Dough” (*Chance*, Vol. 12(4), pp. 50-54) discussed driving distances of PGA players. The mean distance for tee shots on the 1999 men’s PGA tour is 272.2 yards with a standard deviation of 8.12 yards. Assuming that the 1999 tee-shot distances are normally distributed, find the percentage of such tee shots that went

(a) between 260 and 280 yards.

(b) more than 300 yards.

Solution:

(a) We have

$$z_1 = \frac{260 - 272.2}{8.12} \approx -1.5 \quad \text{and} \quad z_2 = \frac{280 - 272.2}{8.12} \approx 0.96$$

therefore the probability is $0.4332 + 0.3315 = 0.7647 \implies 76.47\%$.

(b) We have

$$z = \frac{300 - 272.2}{8.12} \approx 3.42$$

therefore the probability is $0.5 - 0.4997 = 0.0003 \implies 0.03\%$.

Normal Approximation to the Binomial Distribution

EXAMPLE: Mortality tables enable actuaries to obtain the probability that a person at any particular age will live a specified number of years. Insurance companies and others use such probabilities to determine life-insurance premiums, retirement pensions, and annuity payments. According to tables provided by the National Center for Health Statistics in Vital Statistics of the United States, a person of age 20 years has about an 80% chance of being alive at age 65 years.

Suppose, for instance, that 500 people of age 20 years are selected at random. Find the probability that

- (a) exactly 390 of them will be alive at age 65.
- (b) between 375 and 425 of them, inclusive, will be alive at age 65.

Solution: Let X denote the number of people of the 500 who are alive at age 65. Then X has the binomial distribution with parameters $n = 500$ (the 500 people) and $p = 0.8$ (the probability a person of age 20 will be alive at age 65). In principle, we can determine probabilities for X exactly by using the binomial probability formula,

$$P(X = x) = \binom{500}{x} (0.8)^x (0.2)^{500-x}$$

- (a) We have

$$P(X = 390) = \binom{500}{390} (0.8)^{390} (0.2)^{110}$$

However, obtaining the numerical value of the expression on the right-hand side is not easy, even with a calculator. Such computations often lead to roundoff errors and to numbers so large or so small that they are outside the range of the calculator. Fortunately, we can sidestep the calculations altogether by using normal-curve areas.

- (b) We have

$$\begin{aligned} P(375 \leq X \leq 425) &= P(X = 375) + P(X = 376) + \dots + P(X = 425) \\ &= \binom{500}{375} (0.8)^{375} (0.2)^{125} + \binom{500}{376} (0.8)^{376} (0.2)^{124} + \dots + \binom{500}{425} (0.8)^{390} (0.2)^{75} \end{aligned}$$

Here we have the same computational difficulties as we did in part (a), except that we must evaluate 51 complex expressions instead of 1. Again, the binomial probability formula is too difficult to use, and we will need to use normal-curve areas.

EXAMPLE: A student is taking a true-false exam with 10 questions. Assume that the student guesses at all 10 questions.

- (a) Determine the probability that the student gets either 7 or 8 answers correct.
- (b) Approximate the probability obtained in part (a) by an area under a suitable normal curve.

EXAMPLE: A student is taking a true-false exam with 10 questions. Assume that the student guesses at all 10 questions.

- Determine the probability that the student gets either 7 or 8 answers correct.
- Approximate the probability obtained in part (a) by an area under a suitable normal curve.

Solution: Let X denote the number of correct answers by the student. Then X has the binomial distribution with parameters $n = 10$ (the 10 questions) and $p = 0.5$ (the probability of a correct guess).

- Probabilities for X are given by the binomial probability formula

$$P(X = x) = \binom{10}{x} (0.5)^x (1 - 0.5)^{10-x}$$

Using this formula, we get the probability distribution of X , as shown in the Table below. According to that table, the probability the student gets either 7 or 8 answers correct is

$$P(X = 7 \text{ or } 8) = P(X = 7) + P(X = 8) = 0.1172 + 0.0439 = 0.1611$$

- Referring to Table below, we drew the probability histogram of X . Because the probability histogram is bell shaped, probabilities for X can be approximated by areas under a normal curve. The appropriate normal curve is the one whose parameters are the same as the mean and standard deviation of X , which are

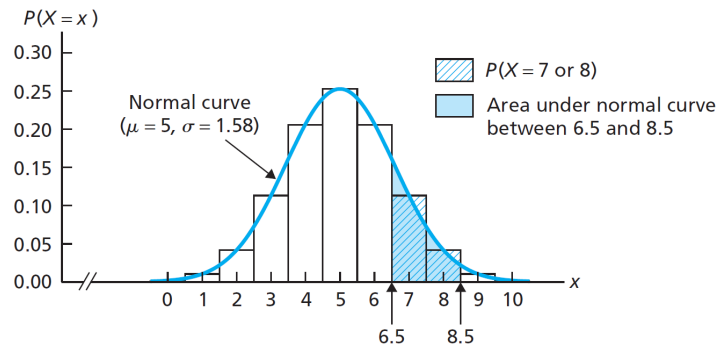
$$\mu = np = 10 \cdot 0.5 = 5$$

and

$$\sigma = \sqrt{np(1-p)} = \sqrt{10 \cdot 0.5 \cdot (1-0.5)} = 1.58$$

Therefore, the required normal curve has parameters $\mu = 5$ and $\sigma = 1.58$.

| Number correct x | Probability $P(X = x)$ |
|-----------------------|---------------------------|
| 0 | 0.0010 |
| 1 | 0.0098 |
| 2 | 0.0439 |
| 3 | 0.1172 |
| 4 | 0.2051 |
| 5 | 0.2461 |
| 6 | 0.2051 |
| 7 | 0.1172 |
| 8 | 0.0439 |
| 9 | 0.0098 |
| 10 | 0.0010 |



The probability $P(X = 7 \text{ or } 8)$ equals the area of the corresponding bars of the histogram, cross-hatched in the Figure above. Note that the cross-hatched area approximately equals the area under the normal curve between 6.5 and 8.5.

The Figure makes clear why we consider the area under the normal curve between 6.5 and 8.5 instead of between 7 and 8. This adjustment is called the *correction for continuity*. It is required because we are approximating the distribution of a discrete variable by that of a continuous variable.

In any case, the Figure shows that $P(X = 7 \text{ or } 8)$ roughly equals the area under the normal curve with parameters $\mu = 5$ and $\sigma = 1.58$ that lies between 6.5 and 8.5. To compute this area,

we convert to z -scores and then find the corresponding area under the standard normal curve. We have

$$z_1 = \frac{6.5 - 5}{1.58} \approx 0.95 \quad \text{and} \quad z_2 = \frac{8.5 - 5}{1.58} \approx 2.22$$

therefore the probability is $0.4868 - 0.3289 = 0.1579$. This is close to $P(X = 7 \text{ or } 8)$, which, as we found in part (a), is 0.1611.

We can now write a general step-by-step method for approximating binomial probabilities by areas under a normal curve.

To Approximate Binomial Probabilities by Normal-Curve Areas

Step 1 Find n , the number of trials, and p , the success probability.

Step 2 Continue only if both np and $n(1 - p)$ are 5 or greater.

Step 3 Find μ and σ , using the formulas $\mu = np$ and $\sigma = \sqrt{np(1 - p)}$.

Step 4 Make the correction for continuity, and find the required area under the normal curve with parameters μ and σ .

EXAMPLE: The probability is 0.80 that a person of age 20 years will be alive at age 65 years. Suppose that 500 people of age 20 are selected at random. Determine the probability that

(a) exactly 390 of them will be alive at age 65.

(b) between 375 and 425 of them, inclusive, will be alive at age 65.

EXAMPLE: The probability is 0.80 that a person of age 20 years will be alive at age 65 years. Suppose that 500 people of age 20 are selected at random. Determine the probability that

- (a) exactly 390 of them will be alive at age 65.
 (b) between 375 and 425 of them, inclusive, will be alive at age 65.

Solution: We have $n = 500$ and $p = 0.8$, therefore

$$np = 500 \cdot 0.8 = 400 \quad \text{and} \quad n(1 - p) = 500 \cdot 0.2 = 100$$

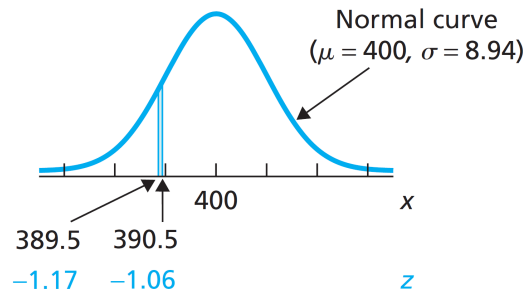
Both np and $n(1 - p)$ are greater than 5, so we can continue. We get

$$\mu = np = 400 \quad \text{and} \quad \sigma = \sqrt{np(1 - p)} = \sqrt{500 \cdot 0.8 \cdot 0.2} = 8.94$$

To make the correction for continuity, we subtract 0.5 from 390 and add 0.5 to 390. Thus we need to find the area under the normal curve with parameters $\mu = 400$ and $\sigma = 8.94$ that lies between 389.5 and 390.5. We have

$$z_1 = \frac{389.5 - 400}{8.94} \approx -1.17 \quad \text{and} \quad z_2 = \frac{390.5 - 400}{8.94} \approx -1.06$$

therefore the probability is about $0.3790 - 0.3554 = 0.0236$ that exactly 390 of the 500 people selected will be alive at age 65.



(b) To make the correction for continuity, we subtract 0.5 from 375 and add 0.5 to 425. Thus we need to determine the area under the normal curve with parameters $\mu = 400$ and $\sigma = 8.94$ that lies between 374.5 and 425.5. As in part (a), we convert to z -scores

$$z_1 = \frac{374.5 - 400}{8.94} \approx -2.85 \quad \text{and} \quad z_2 = \frac{425.5 - 400}{8.94} \approx 2.85$$

and then find the corresponding area under the standard normal curve. This area is $0.4978 \cdot 2 = 0.9956$. So, $P(375 \leq X \leq 425) = 0.9956$, approximately.

