

Unit 2

Probability, Density Curves & Normal Distributions

Variability and Randomness

Statistics involves the study of variability. But how can we work with something that involves so much uncertainty?

To understand this, we consider the idea of random behaviour.

The key lies in the fact that random behaviour is unpredictable in the short-run, but has a **regular** and **predictable** distribution in the long-run.

Probability

Roll a die, toss a coin, or buy a lottery ticket with your lucky numbers. The outcome cannot be predicted ahead of time, but can nevertheless be described by a regular pattern, one which emerges only after many repeated trials.

This fact is the foundation for the study of **probability**.

Probability

We toss a coin and record the proportion of heads that has been observed after each toss. Assuming the coin is fair, the likelihood of observing a Head is the same as that for observing a Tail. There is a 50% chance of either outcome.

Suppose we observe the following sequence of tosses:

H T T H T H

We record the proportions 1.0, 0.5, 0.33, 0.5, 0.4, 0.5

Probability

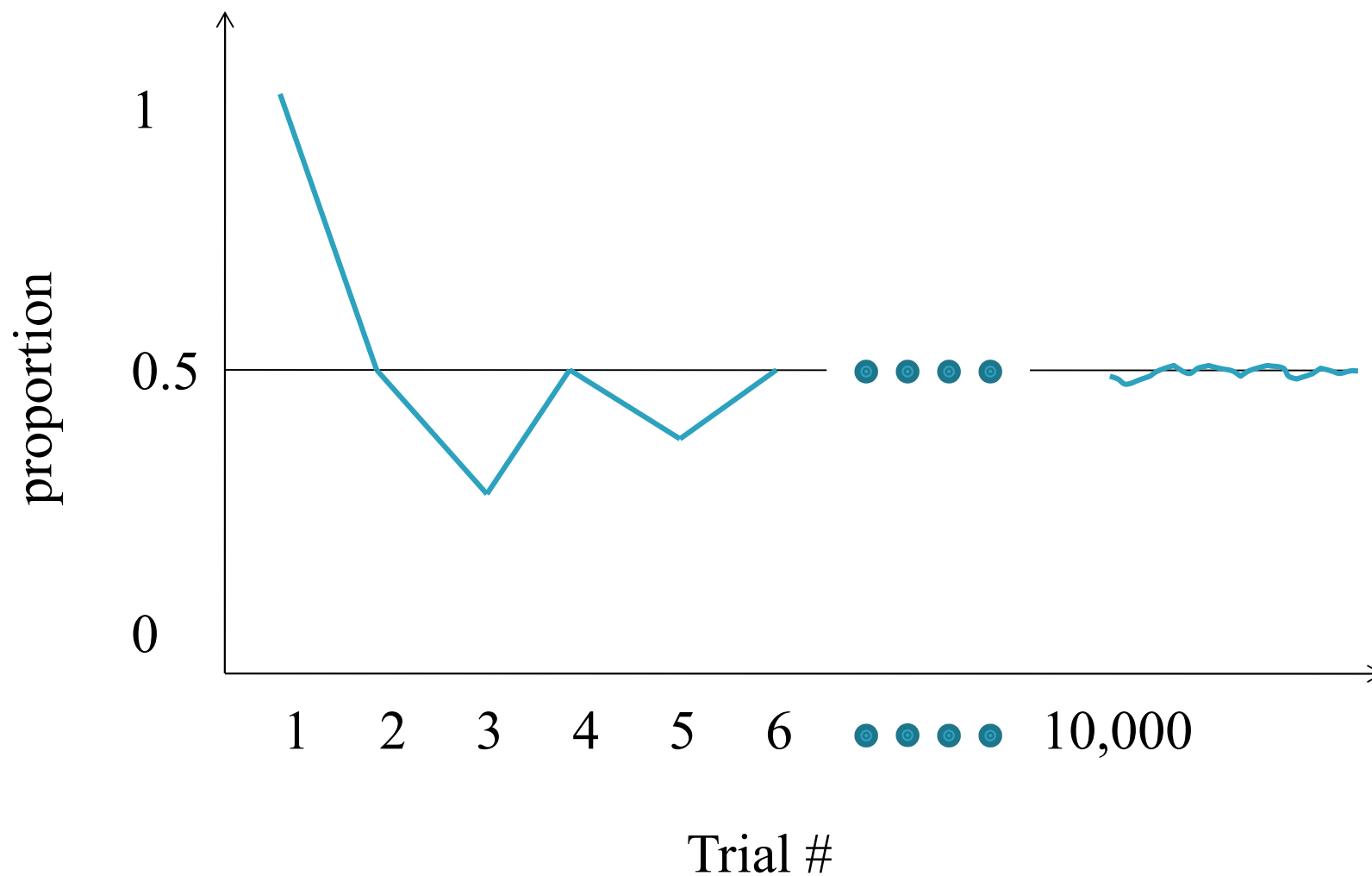
The proportions vary quite a bit early on, but in the long-run, we are bound to see proportions very close to 0.5 consistently.

Eventually the proportion gets close to 0.5 **and stays there**. (Toss a fair coin 10,000 times and you will almost certainly observe between 4,800 and 5,200 Heads.)

We say that 0.5 is the **probability** of observing a Head.

U2-5

Probability



Randomness

We call a phenomenon **random** if individual outcomes are uncertain but there is a regular distribution of outcomes in a large number of repetitions.

A **random experiment** is any process or activity in which there is uncertainty, and which has two or more possible outcomes.

Probability

Note the term “random” is not the same in statistics as it is in everyday language. We often associate “randomness” with a haphazard or even chaotic event.

This may be due to the fact that we do not observe the phenomenon enough times to see a long-run pattern, after which regularity would be sure to emerge.

The **probability** of any outcome of a random phenomenon is the proportion of times the outcome would occur in an infinitely long series of trials.

Proportion vs. Probability

What is the difference between proportions and probability?

A proportion is a known or observed value, while a probability is a theoretical value of a proportion after an infinitely long series of trials.

We speak of proportions in the present tense, whereas probability always relates to future events.

Probability

Probability theory is the branch of mathematics that describes random behaviour. We must deal with mathematical models, because probability itself can never be observed.

We can't toss a coin forever, but our understanding of random behaviour enables us to describe what would happen if we did.

Probability Model

Since we can't perform an experiment infinitely many times, we use a **probability model** to describe random behaviour.

A probability model has two components:

- a list of possible outcomes
- a probability for each outcome

Sample Space

The **sample space S** of a random phenomenon is the set of all possible outcomes.

Sample spaces can be very simple. If we toss a coin one time, our sample space is simply $S = \{H, T\}$.

They can also be very complicated. Consider the set of all possible combinations of six numbers drawn in Lotto 6/49 from a population of 49 numbers. There are almost 14 million of them!

Sample Space

If we toss a coin three times, the sample space is

$$S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$$

If we are interested in the outcomes of the Valour FC soccer team's next two games, the sample space is

$$S = \{WW, WT, WL, TW, TT, TL, LW, LT, LL\}$$

Note that **order matters** – winning the first game and losing the second (WL) is not the same thing as losing the first and winning the second (LW).

Sample Space

If we roll two six-sided dice, the sample space of outcomes is

$$S = \{11, 12, 13, 14, 15, 16, \\ 21, 22, 23, 24, 25, 26, \\ 31, 32, 33, 34, 35, 36, \\ 41, 42, 43, 44, 45, 46, \\ 51, 52, 53, 54, 55, 56, \\ 61, 62, 63, 64, 65, 66\}$$

If we were specifically interested in the sum of the numbers on the two dice, our sample space would be

$$S = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$$

Practice Question

A fair six-sided die has 3 faces that are painted blue, 2 faces that are red and 1 face that is green. We will toss the die twice. The outcome of interest is the colour facing upward on each of the two tosses. What is the sample space for this experiment?

- (A) $S = \{BB, BR, BG, RR, RG, GG\}$
- (B) $S = \{B, R, G\}$
- (C) $S = \{0, 1, 2\}$
- (D) $S = \{B, B, B, R, R, G\}$
- (E) $S = \{BB, BR, BG, RB, RR, RG, GB, GR, GG\}$

Probabilities of Outcomes

Suppose the sample space for an experiment is

$$S = \{O_1, O_2, \dots, O_n\}$$

The probability of outcome O_i is denoted as p_i . Probabilities of the outcomes must satisfy the following two conditions:

- (i) $0 \leq p_i \leq 1$ for all $i = 1, 2, \dots, n$
- and (ii) $p_1 + p_2 + \dots + p_n = 1$

Events

An **event** is any subset of outcomes in the sample space. For example, when flipping a coin three times, the event that we get exactly two Tails can be written as

$$A = \{HTT, THT, TTH\}$$

The event that Valour FC doesn't lose either of their next two games is

$$B = \{WW, WT, TW, TT\}$$

Events

When rolling two dice, the event “At Least One 4” is

$$A = \{14, 24, 34, 41, 42, 43, 44, 45, 46, 54, 64\}$$

The event “Sum is 9” is

$$B = \{36, 45, 54, 63\}$$

Complements

The **complement** A^c of an event A is the event consisting of all outcomes in the sample space which are not contained in A .

For example, when flipping a coin, $H^c = T$.

When flipping a coin three times, the complement of the event $A = \{\text{exactly two tails}\}$ is

$$A^c = \{HHH, HHT, HTH, THH, TTT\}$$

Complements

Since all probabilities add up to one, it follows that

$$P(A^c) = 1 - P(A)$$

For example, $P(H) = 1 - P(T)$.

$$P(\text{Bombers win their next game}) = 1 - P(\text{Bombers lose or tie})$$

$$P(\text{sum of two dice is 5 or higher}) = 1 - P(\text{sum is 4 or less})$$

Practice Question

A recently married couple plans to have two children. The outcome of interest is the gender of each of the two children. Consider the event that exactly one of the couple's children will be a boy. Which of the following is the complement of this event?

- (A) two boys
- (B) two girls
- (C) one girl
- (D) at least one girl
- (E) zero or two girls

Probability Distributions

A **probability distribution** gives the possible values of some variable, as well as the probability of each value.

Some probability distributions are very simple and intuitive. If we roll a fair six-sided die and let X be the number showing on the top face, then the probability distribution of X is as follows:

x	1	2	3	4	5	6
$P(X = x)$	1/6	1/6	1/6	1/6	1/6	1/6

Probability Distributions

We roll a fair six-sided die with three faces painted blue, two faces painted red, and one face painted green. The outcome of interest is the colour you roll. The sample space is $S = \{B, R, G\}$ and the probability distribution is as follows:

Colour	Blue	Red	Green
Probability	$1/2$	$1/3$	$1/6$

Probability of Events

The probability of any event can be found by adding the probabilities of all outcomes contained in that event.

For example, when flipping three coins, the probability of getting two tails is

$$P(2 \text{ Tails}) = P(HTT) + P(THT) + P(TTH)$$

Probability Distributions

Let X be the sum of the numbers rolled on two fair dice.
The probability distribution of X is as follows:

x	2	3	4	5	6	7	8	9	10	11	12
$P(X = x)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

The probability of rolling a sum greater than 8 is

$$\begin{aligned}
 P(X > 8) &= P(X = 9) + P(X = 10) + P(X = 11) + P(X = 12) \\
 &= 4/36 + 3/36 + 2/36 + 1/36 = 10/36 = 0.2778
 \end{aligned}$$


Random Variables

The variable X (the sum of the rolls of two dice) is called a **random variable**.

A random variable is a numerical description of the outcome of a statistical experiment.

Example

The NHL's Atlantic Division consists of eight teams. Prior to the start of the season, the probabilities of each team winning the division are determined, and are shown below, where k is some constant:

Team								
Probability	k	0.05	$4k$	$2k$	0.02	0.04	0.29	$3k$

What is the probability that a Canadian team wins the division?

Example

First, we find the value of the constant k . We know all probabilities add up to one (someone has to win the division), so we have

$$k + 0.05 + 4k + 2k + 0.02 + 0.04 + 0.29 + 3k = 1$$

$$\Rightarrow 10k + 0.40 = 1 \Rightarrow 10k = 0.60 \Rightarrow k = 0.60/10 = 0.06$$

Therefore,

$$P(\text{Canadian team wins}) = P(\text{Montreal}) + P(\text{Ottawa}) + P(\text{Toronto})$$

$$= k + 0.05 + 4k = 0.06 + 0.05 + 4(0.06) = 0.06 + 0.05 + 0.24 = 0.35$$

Example

The NHL's Pacific Division consists of eight teams. Prior to the start of the season, the probabilities of each team winning are determined, and are shown below, with some values missing:

Team								
Probability	0.08	0.27	0.09	0.05	???	0.01	0.04	???

What is the probability that an American team wins the division?

Example

The probability an American team wins is

$$P(\text{Anaheim}) + P(\text{LA}) + P(\text{San Jose}) + P(\text{Seattle}) + P(\text{Vegas})$$

Since we don't know two of these probabilities, we can equivalently find the probability as follows:

$$\begin{aligned} P(\text{US team wins}) &= 1 - P(\text{Canadian team wins}) \\ &= 1 - [P(\text{Calgary}) + P(\text{Edmonton}) + P(\text{Vancouver})] \\ &= 1 - (0.08 + 0.27 + 0.09) = 1 - 0.44 = 0.56 \end{aligned}$$

Practice Question

A gas station recently sent out “scratch and save” coupons through the mail. The coupons offer customers varying discounts off a 30-litre fuel purchase. The discounts, and the probabilities of receiving them, are shown below:

Discount	\$1	\$2	\$3	\$4	\$5
Probability	$7k$	0.32	$3k$	0.08	$2k$

What is the probability of receiving a discount of at least \$3?

- (A) 0.18 (B) 0.29 (C) 0.33 (D) 0.42 (E) 0.60

Practice Question

The number of courses X taken by students at a large university in one semester has the probability distribution shown below, with some values missing:

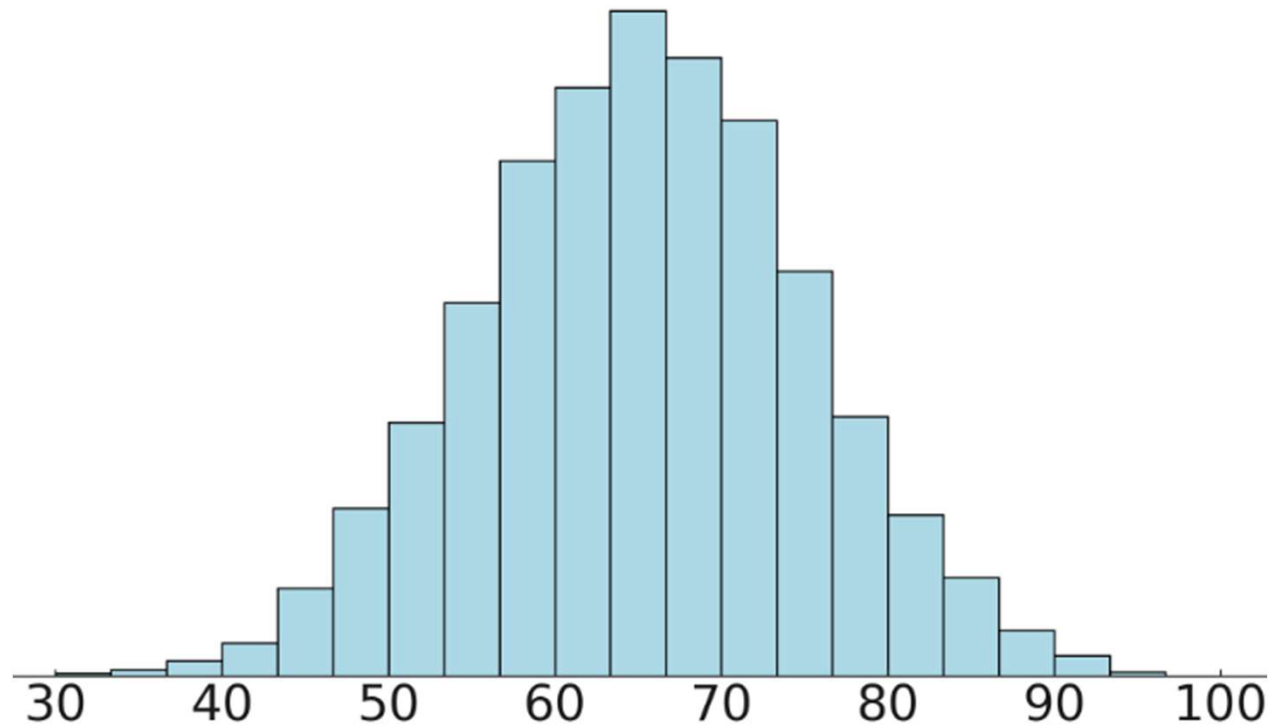
# of Courses	1	2	3	4	5
Probability	???	0.16	???	0.37	0.28

What is the probability that a randomly selected student is taking an odd number of courses?

- (A) 0.32 (B) 0.38 (C) 0.47 (D) 0.53 (E) 0.60

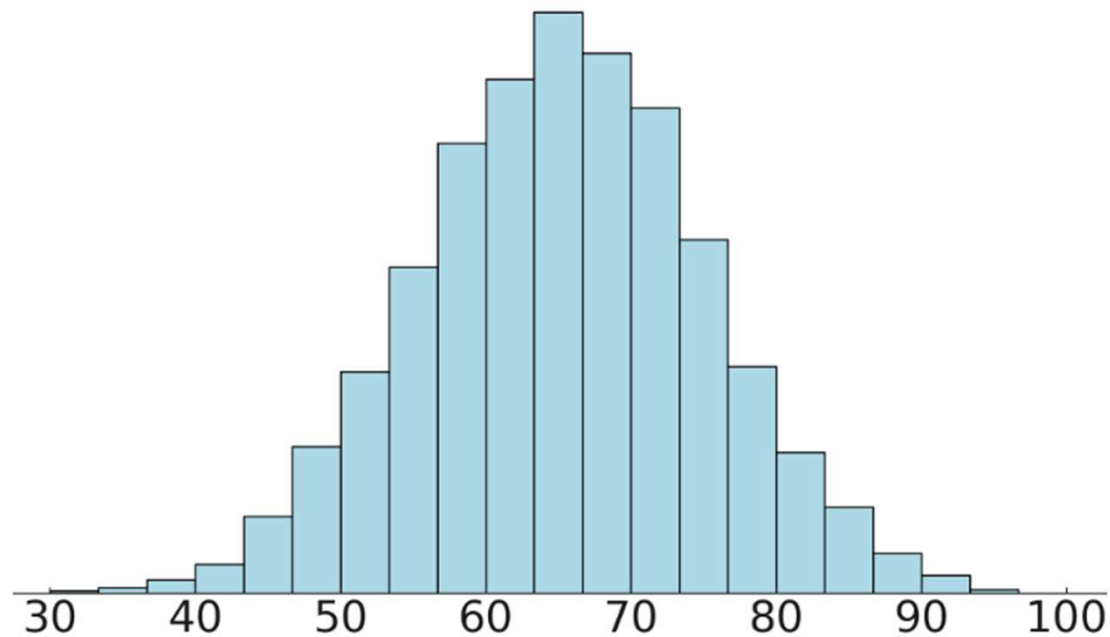
Density Curves

The following is a histogram of test scores for 5,000 high school students taking a provincial math exam:



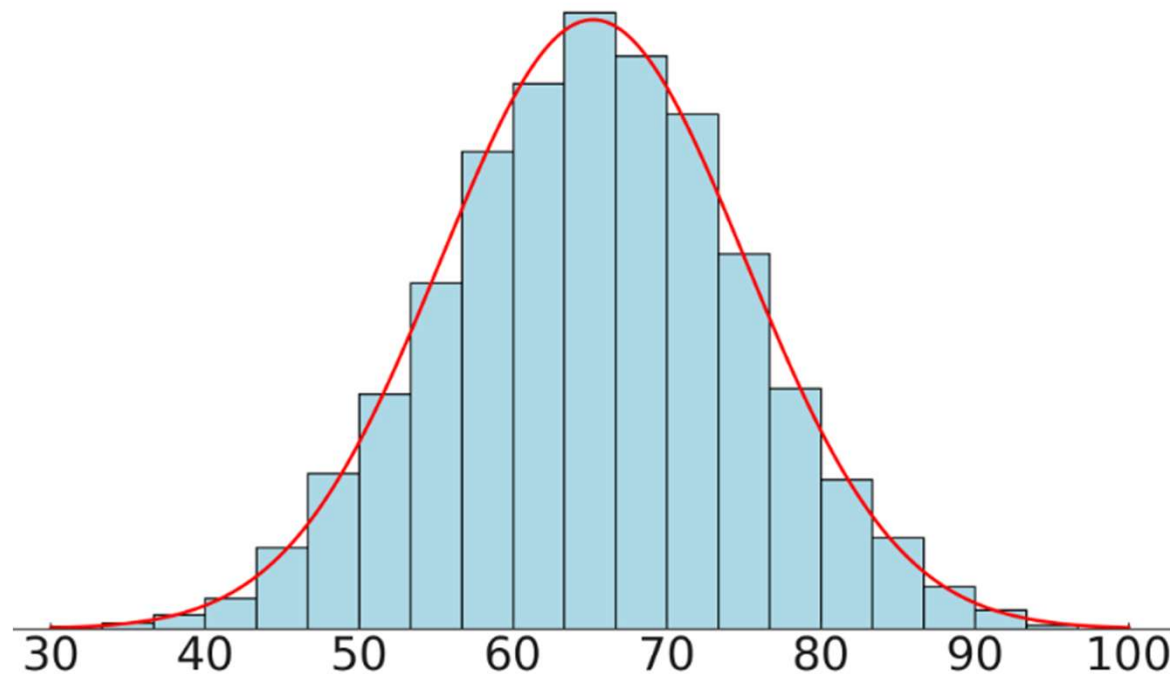
Density Curves

Scores of many students on this exam have a fairly regular distribution. The distribution is symmetric with a fairly smooth pattern descending from the center on each side. There are no gaps or outliers.



Density Curves

We can fit a smooth curve through the tops of the bars in the histogram to get an approximation of the distribution.



Density Curves

We call this curve a mathematical model for the distribution. It is an idealized description of the data and offers an overall picture of the distribution without considering small irregularities.

In general, a smooth curve is easier to work with than a histogram.

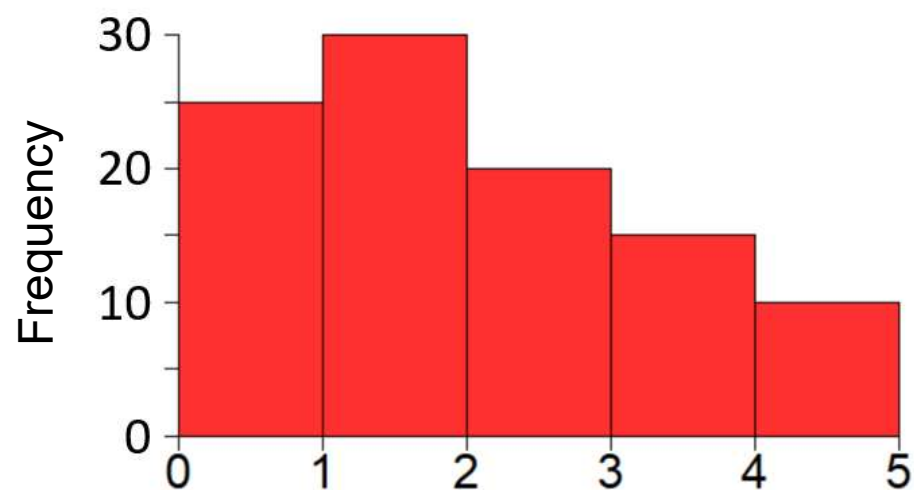
Density Curves

The picture we get from a histogram depends strongly on the number of classes and interval lengths we choose. This is not so when using a smooth curve; there are no intervals.

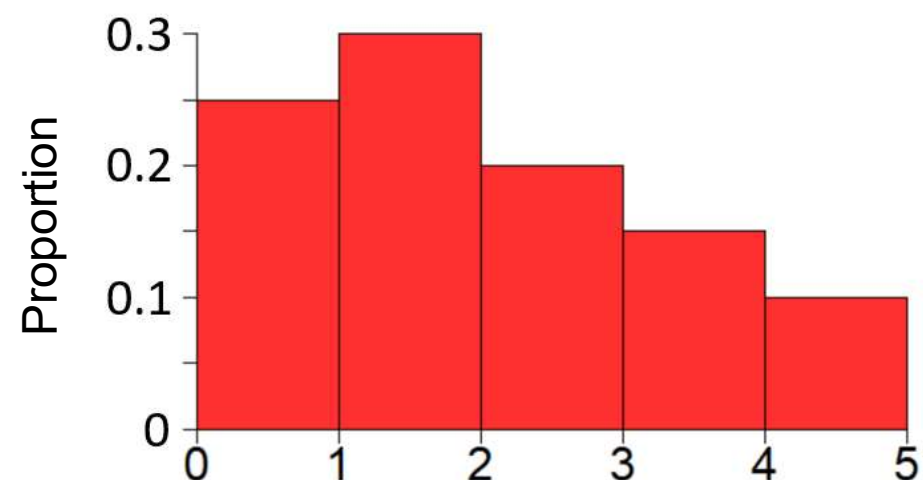
Our eyes respond to the **area** of bars in a histogram. A histogram can be scaled to have a total area equal to one, so that the areas of the bars represent relative frequencies or proportions of observations falling within each class.

Density Curves

total area = 100

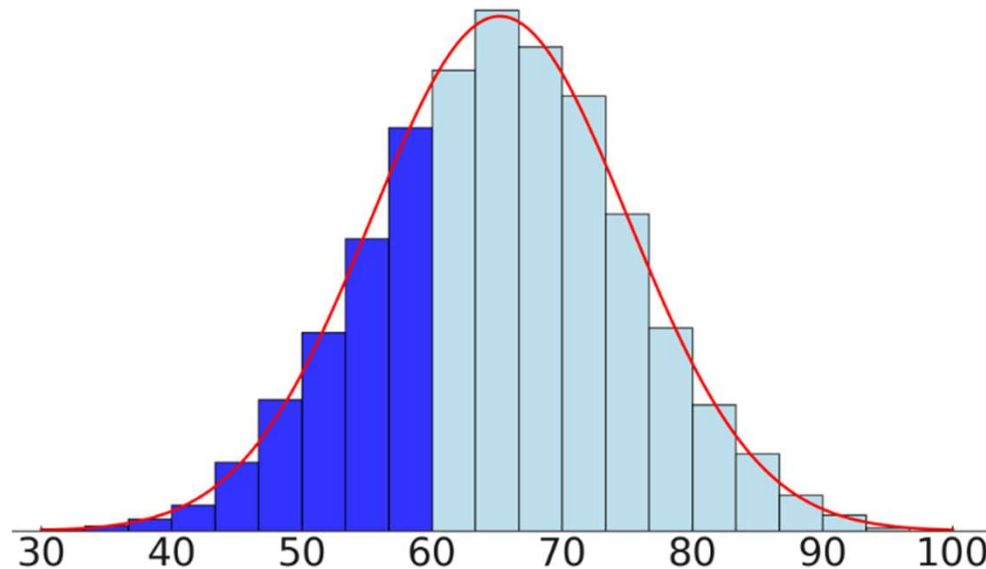


total area = 1



Density Curves

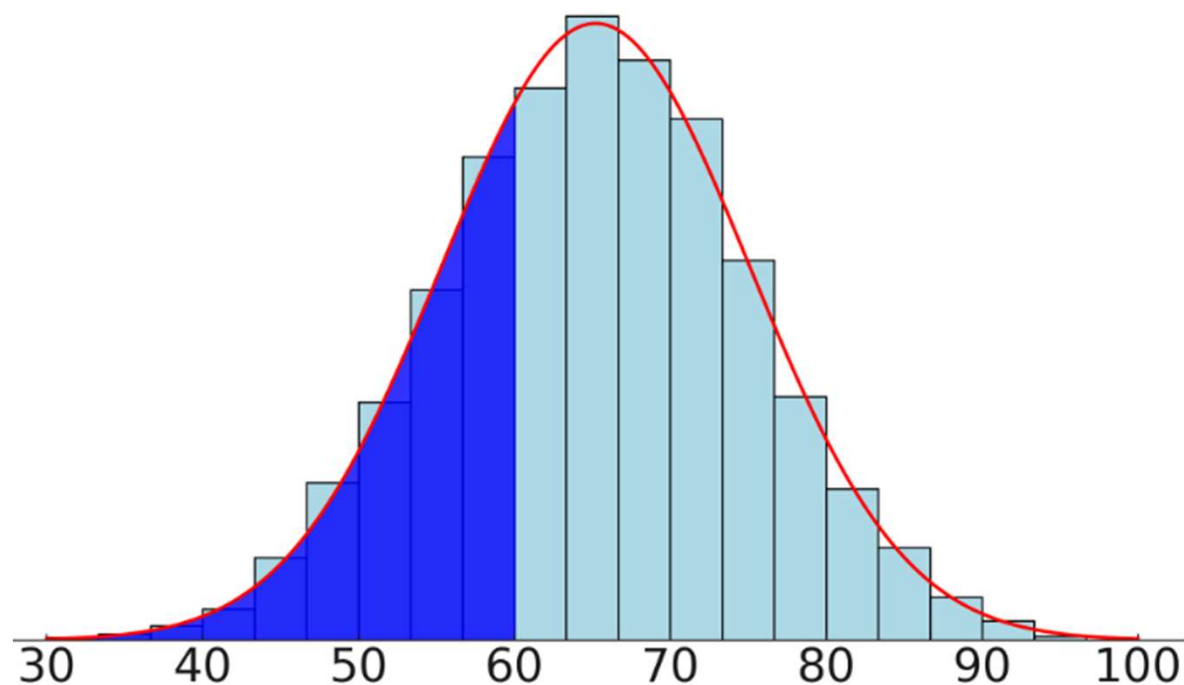
Suppose we are interested in the proportion of students with scores less than 60:



This represents 1,500 students, or a proportion of $1,500/5,000 = 0.3000$.

Density Curves

What if we look at the **area** to the left of 60 underneath the curve?



Density Curves

Just as we did with a histogram, we can scale the area underneath the curve so that it is equal to one. Now the areas underneath the curve represent the proportion of observations falling within the range of interest.

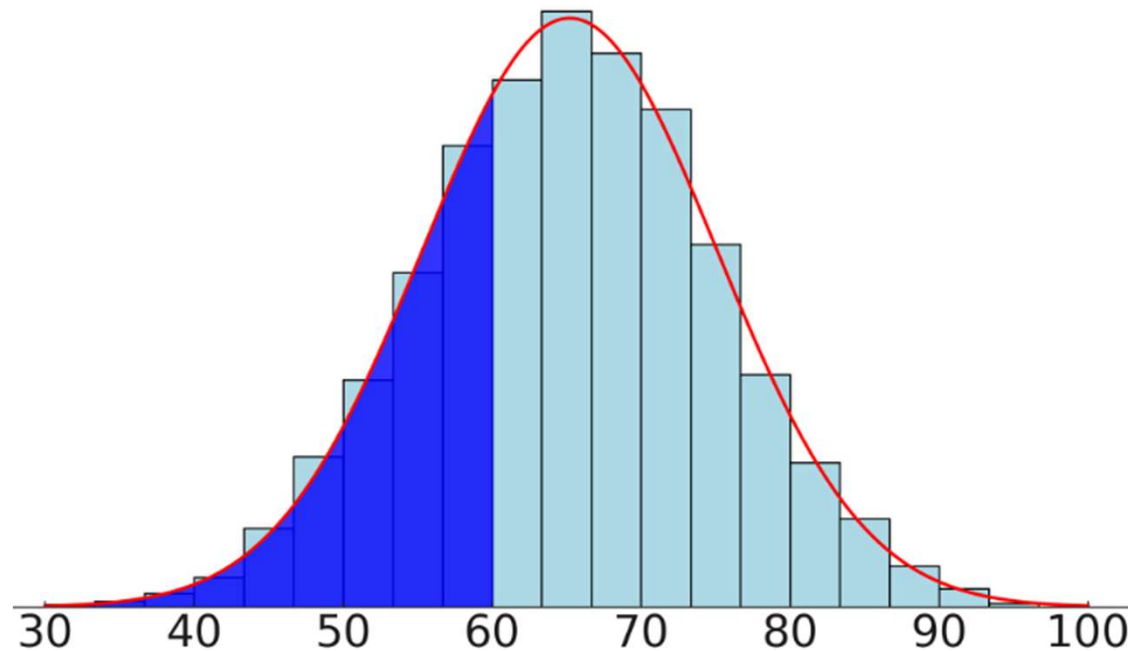
$$\text{proportion of values in interval} = \frac{\text{area of interval}}{\text{total area under curve}}$$

↑
= 1

\Rightarrow proportion of values in interval = area of interval

Density Curves

The area shaded under the curve is equal to 0.3012. (We will learn how to find these areas shortly.) Note that the proportion we estimate using the area under the curve is only 0.0012 away from the true proportion.



Density Curves

The curve describing the overall shape of a distribution is called a **density curve**.

All density curves have the following three properties:

- The curve lies strictly on or above the x -axis.
- The total area underneath the curve and above the x -axis is equal to **one**.
- The curve represents a proper function, i.e., for each value of x there is a unique value of y .

Density Curves

The area under the curve and above the x -axis covering any interval represents the proportion of all values falling within that interval.

The previous example illustrates that the area under a density curve offers a very good approximation to the areas of the corresponding intervals in a histogram.

Density Curves

Now suppose we have a very large population.

For example, suppose we were to construct a histogram representing the heights of **all Canadian men**, the annual incomes of **all workers in Manitoba**, or the GPAs of **all U of M students**.

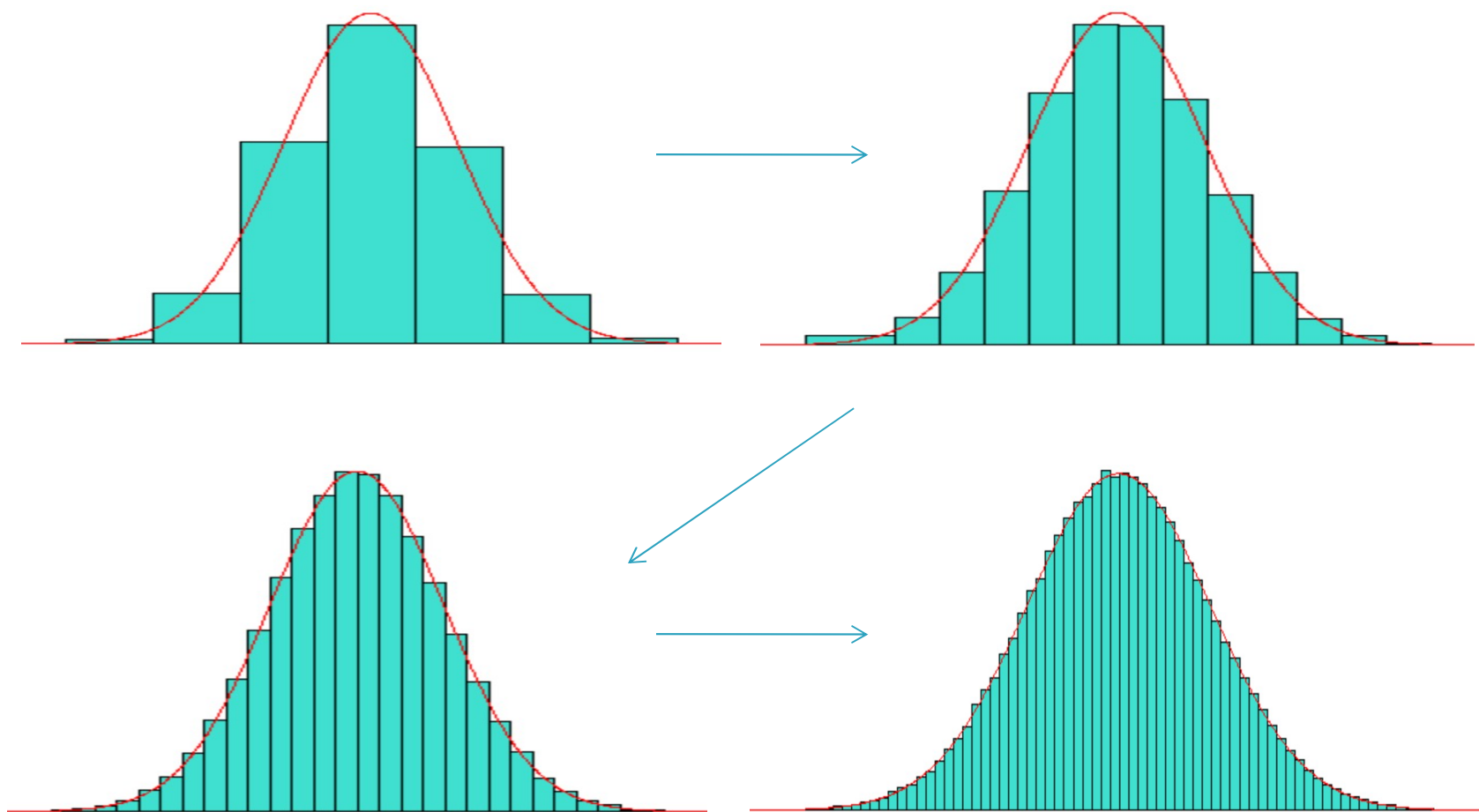
Density Curves

If the population is infinitely (sufficiently) large, we can construct a histogram with as many intervals as we want.

Suppose we construct histograms, repeatedly reducing the length of each interval, so that the number of intervals gets very large.

The representation is then essentially a density curve.

Density Curves



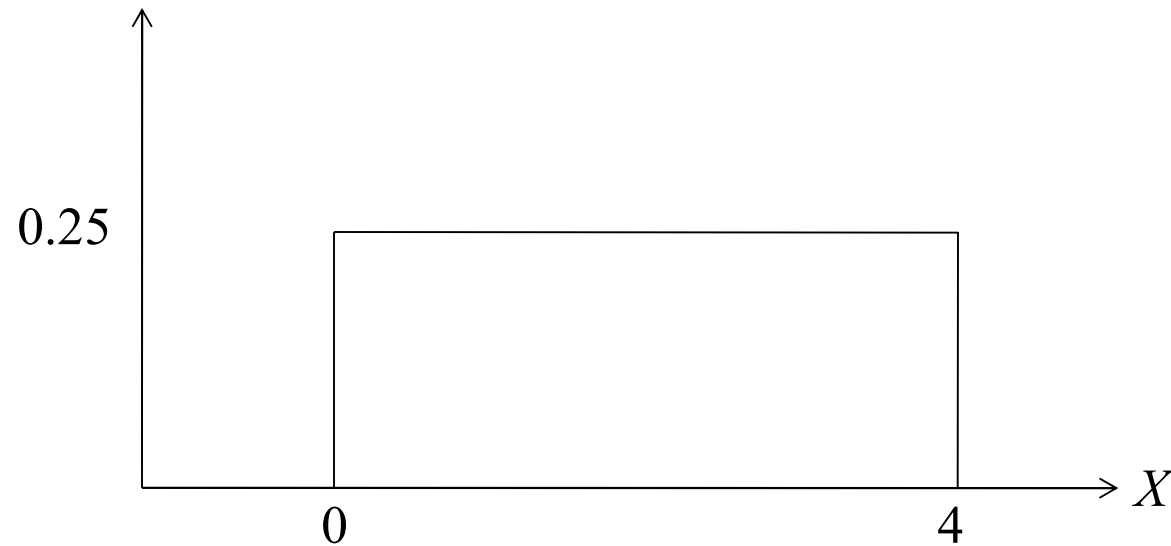
Density Curves

The graphical tools we have used thus far are applicable to samples or finite populations.

Density curves describe the distributions of **continuous** variables for “infinite” (i.e., large) populations.

Uniform Distribution

The distribution shown below for some variable X is known as a **uniform distribution**.



Uniform Distribution

The area underneath the curve is equal to

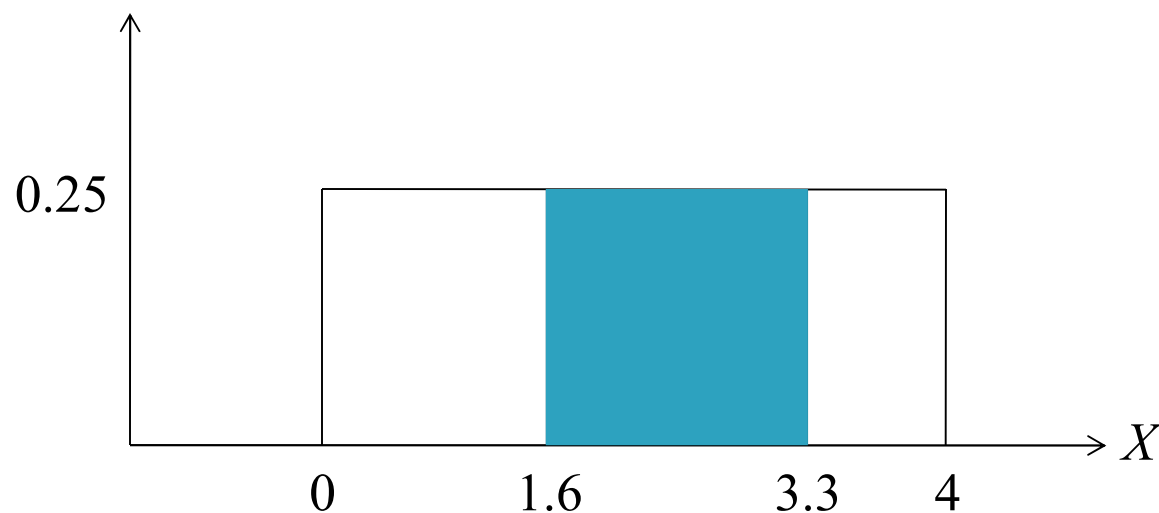
$$\text{Area} = \text{base} * \text{height} = (4 - 0)(0.25) = 1$$

so we know this is a valid density curve.

We can find the proportion of values of X falling in any given interval, because every interval is just another rectangle.

Uniform Distribution

Suppose we are interested in the proportion of values of X falling between 1.6 and 3.3:



This proportion is equal to

$$P(1.6 < X < 3.3) = (3.3 - 1.6)(0.25) = 1.7(0.25) = 0.425$$

R Code

```
> punif(3.3, 0, 4) - punif(1.6, 0, 4)  
[1] 0.425
```

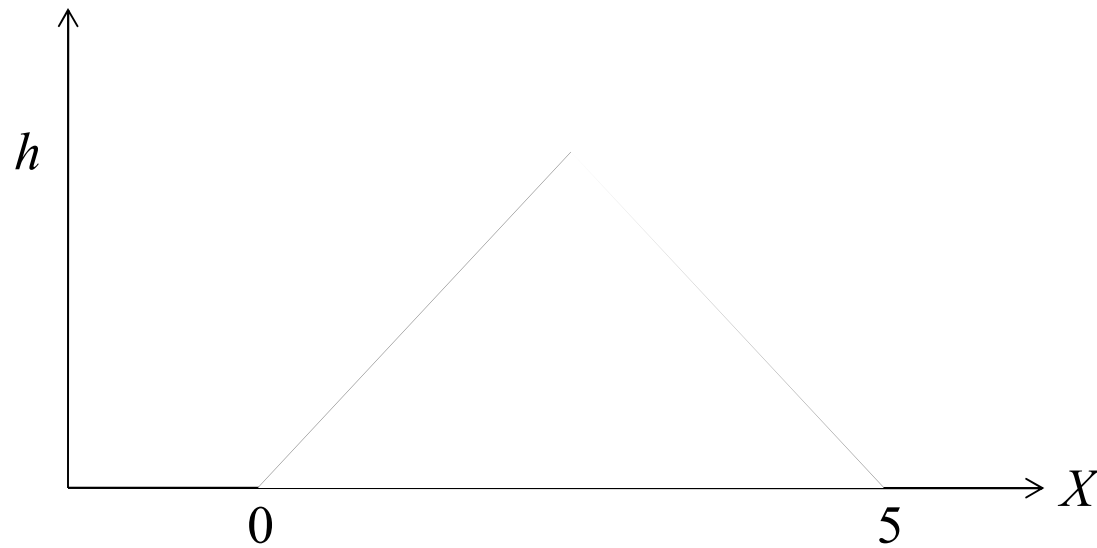
Practice Question

The time X a person spends in the shower follows a uniform distribution on the interval from 5 to 10 minutes. What proportion of the person's showers last between 6.6 and 7.4 minutes?

- (A) 0.04 (B) 0.08 (C) 0.16 (D) 0.32 (E) 0.40

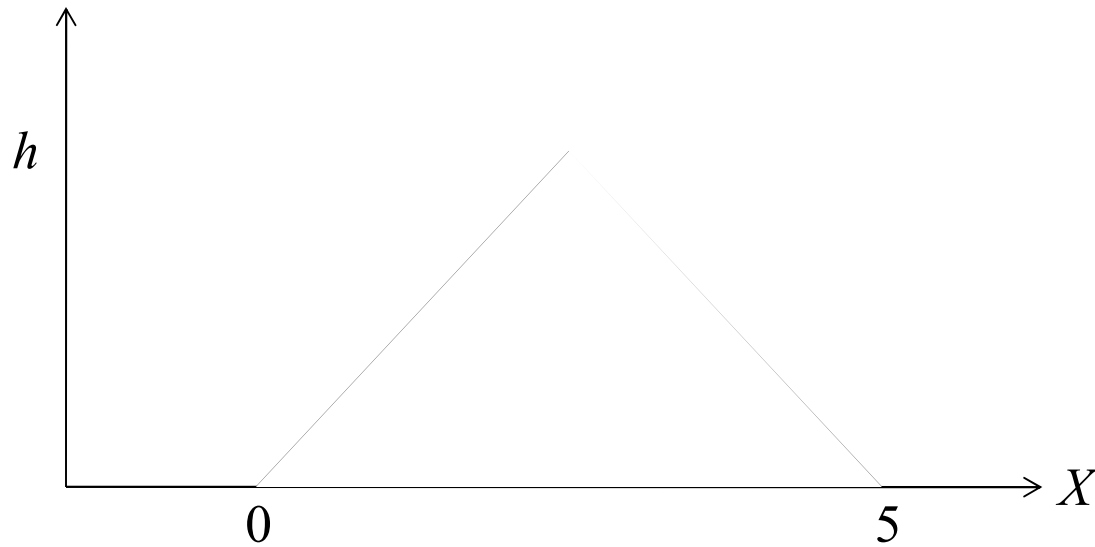
Example

The distribution shown below for some variable X is known as a **triangular distribution**.



Example

What must be the value of the maximum height h in order for this to be a valid density curve?



Example

The area of a triangle is calculated as

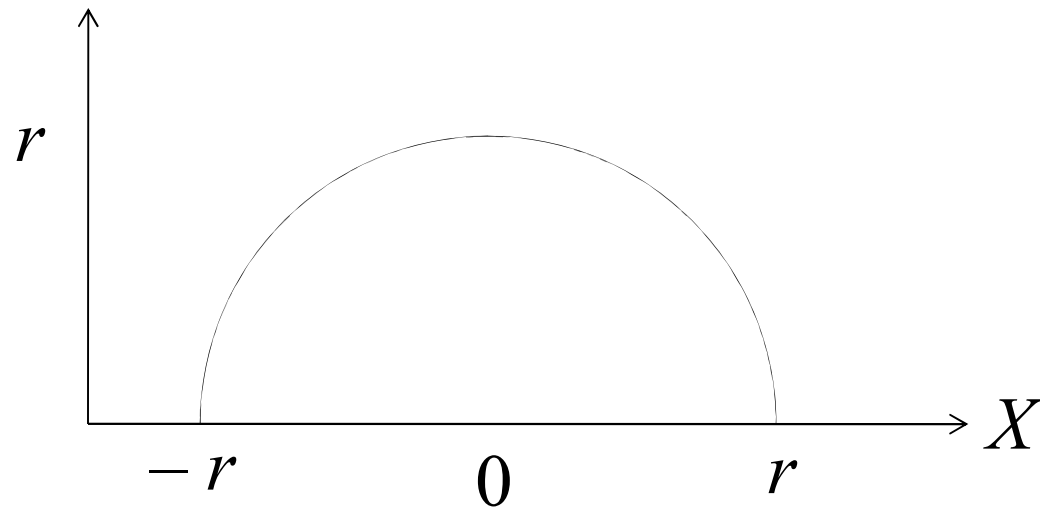
$$\text{Area} = 0.5 * \text{base} * \text{height}$$

and so for this to be a valid density curve,

$$\begin{aligned}\text{Area} &= 0.5(5)(h) = 1 \\ \Rightarrow 2.5h &= 1 \Rightarrow h = 1/2.5 = \mathbf{0.4}\end{aligned}$$

Practice Question

A variable X has a semi-circular density curve as shown below:



Practice Question

What must be the value of the radius r in order for this to be a legitimate density curve? Hint: Recall that the area of a circle is calculated as

$$\text{Area} = \pi r^2$$

- (A) $\sqrt{\frac{\pi}{2}}$ (B) $\sqrt{\frac{1}{\pi}}$ (C) $\frac{1}{2\pi}$ (D) $\sqrt{\frac{2}{\pi}}$ (E) $\sqrt{\pi}$

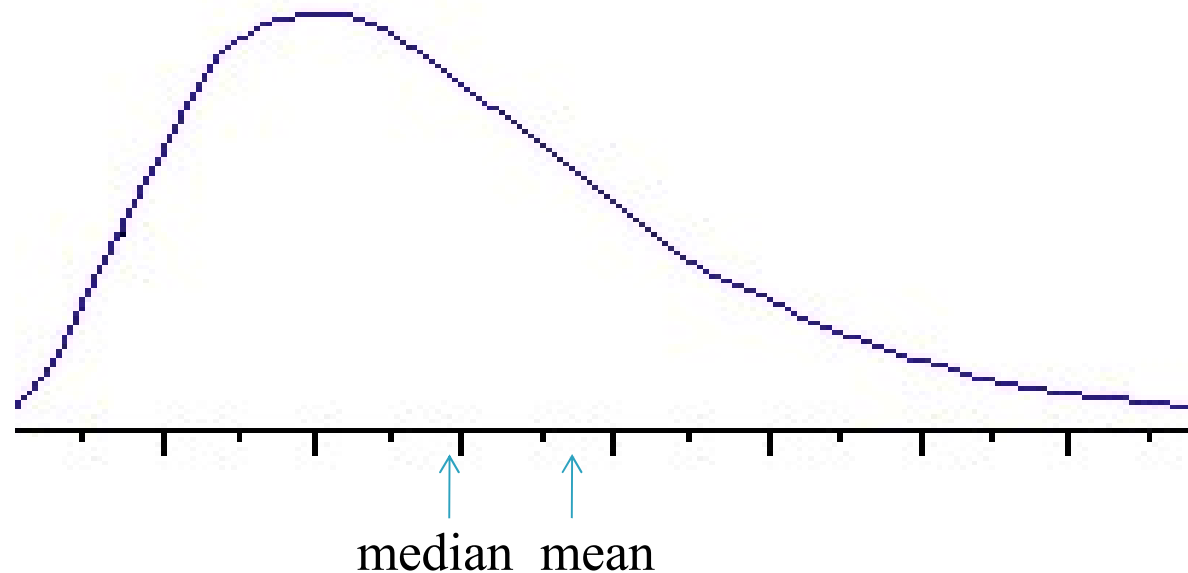
Density Curves – Mean & Median

The **median** of a distribution described by a particular density curve is the point on the x -axis with area 0.5 under the curve to each side.

The **mean** of a continuous distribution is the “balance point” along the x -axis. This is the point at which the distribution would exactly balance if it were made of some solid material.

Density Curves – Mean & Median

Population distributions can also be described by their shape (symmetry, skewness). The mean and median for a symmetric distribution are equal. For a skewed distribution, the mean is closer to the tail:



Density Curves – Variance

The standard deviation of a continuous distribution is mathematically more complicated, but unlike the mean's conceptual definition as the “balance point”, there is no such intuitive description for the standard deviation. It is a measure of spread for the entire population (and the square root of the variance, which does have a “nice” definition).

The variance of a continuous distribution is the average squared deviation of an observation to the mean of the distribution.

Sample Vs. Population

We must distinguish between the **sample mean** \bar{x} and the **population mean**, which we denote as μ . The population mean is the average value of some variable for all units in the entire population.

Similarly, while s is the **sample standard deviation**, we denote the **population standard deviation** as σ (and so σ^2 is the **population variance**).

Parameters & Statistics

We call values such as μ and σ **parameters** of the population. A parameter is a number that describes an entire population.

Because we often deal with very large populations, the values of parameters are usually unknown.

For this reason, we use **statistics** or **estimators** such as \bar{x} and s to estimate the values of parameters. A statistic is a number computed from sample data.

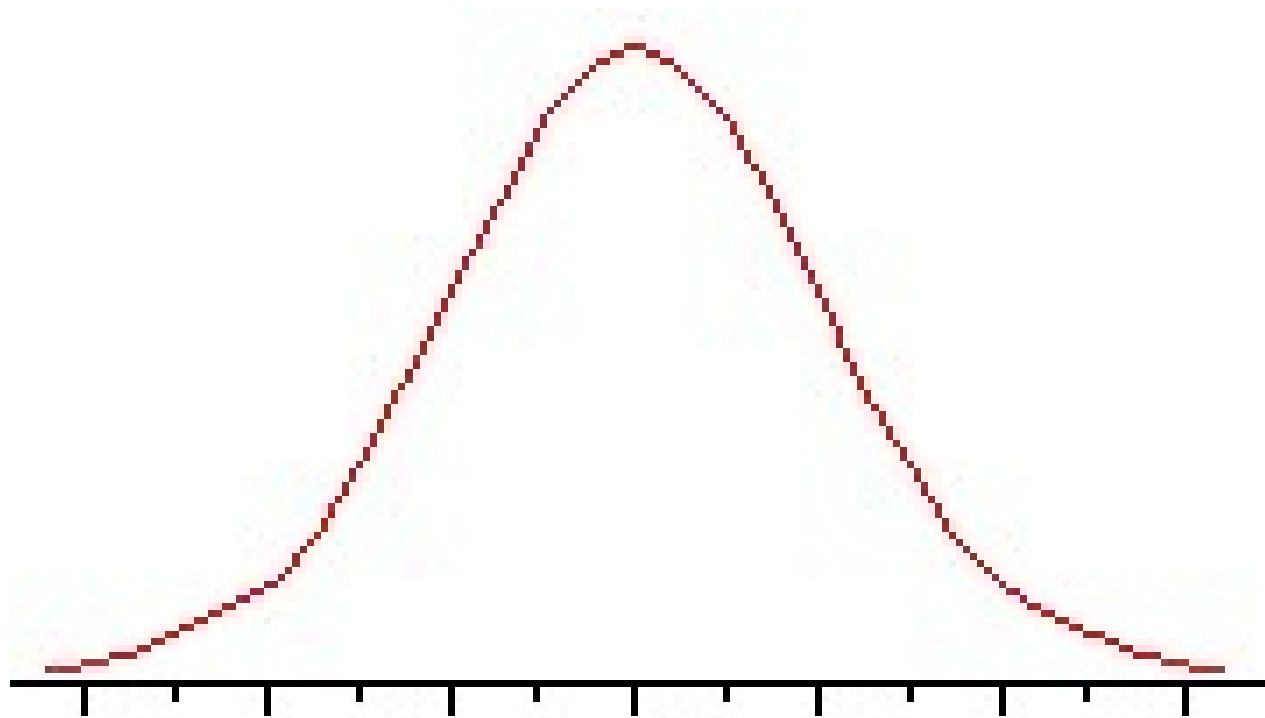
Practice Question

According to the makers of m&m's chocolates, **15%** of all m&m's produced are yellow. You buy a large bag containing **250** m&m's, and you find that 50 of them (**20%**) are yellow. The values in **bold** are, respectively:

- (A) statistic, parameter, parameter
- (B) parameter, parameter, statistic
- (C) statistic, statistic, parameter
- (D) parameter, statistic, parameter
- (E) parameter, statistic, statistic

The Normal Distribution

The density curve we looked at previously is called a **normal distribution curve**.



The Normal Distribution

Normal distributions arise naturally for many different variables:

- heights
- test scores
- diameters
- batting averages
- lengths of pregnancies

The Normal Distribution

The parameters μ (which can take any value, positive or negative) and σ (which must be positive) completely characterize a normal distribution; that is, they tell us everything we need to know about the exact form of the distribution.

There are infinitely many normal distributions – one for every combination of μ and σ .

The Normal Distribution

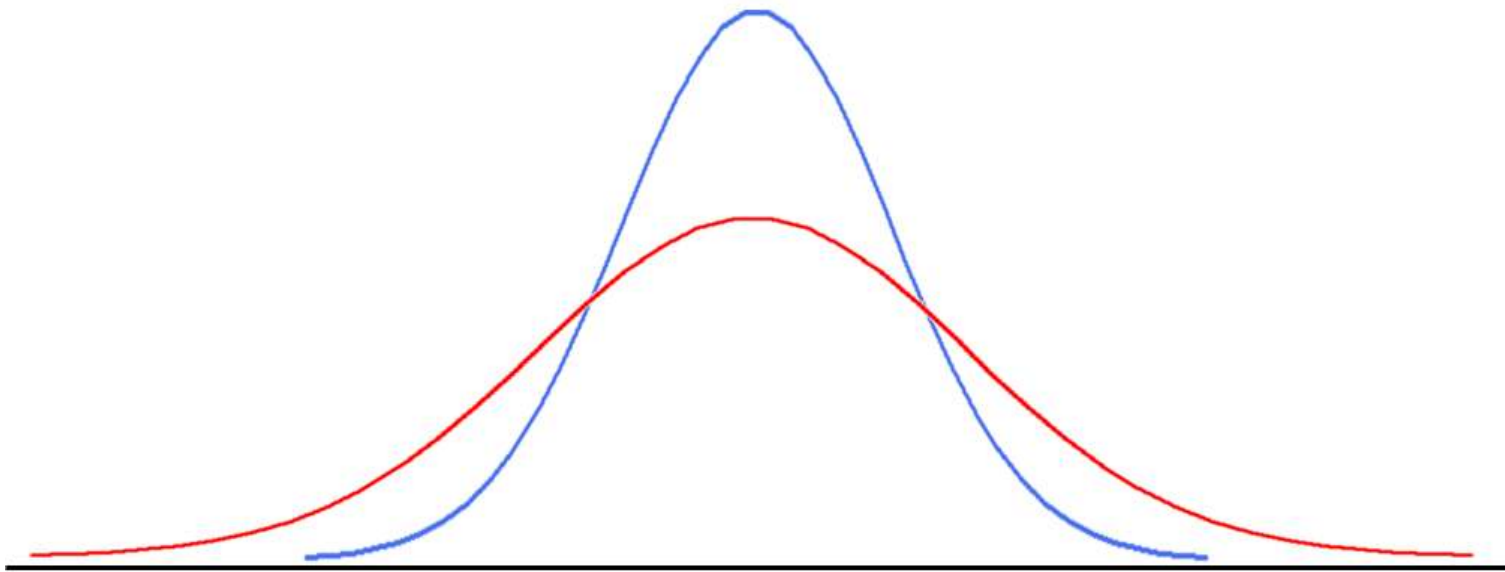
Some characteristics of a normal distribution:

- bell-shaped
- single-peaked
- symmetric about its mean

Note that the curve never actually touches the x -axis, but the height of the curve becomes infinitely small as the value of the variable X gets extremely large or small.

The Normal Distribution

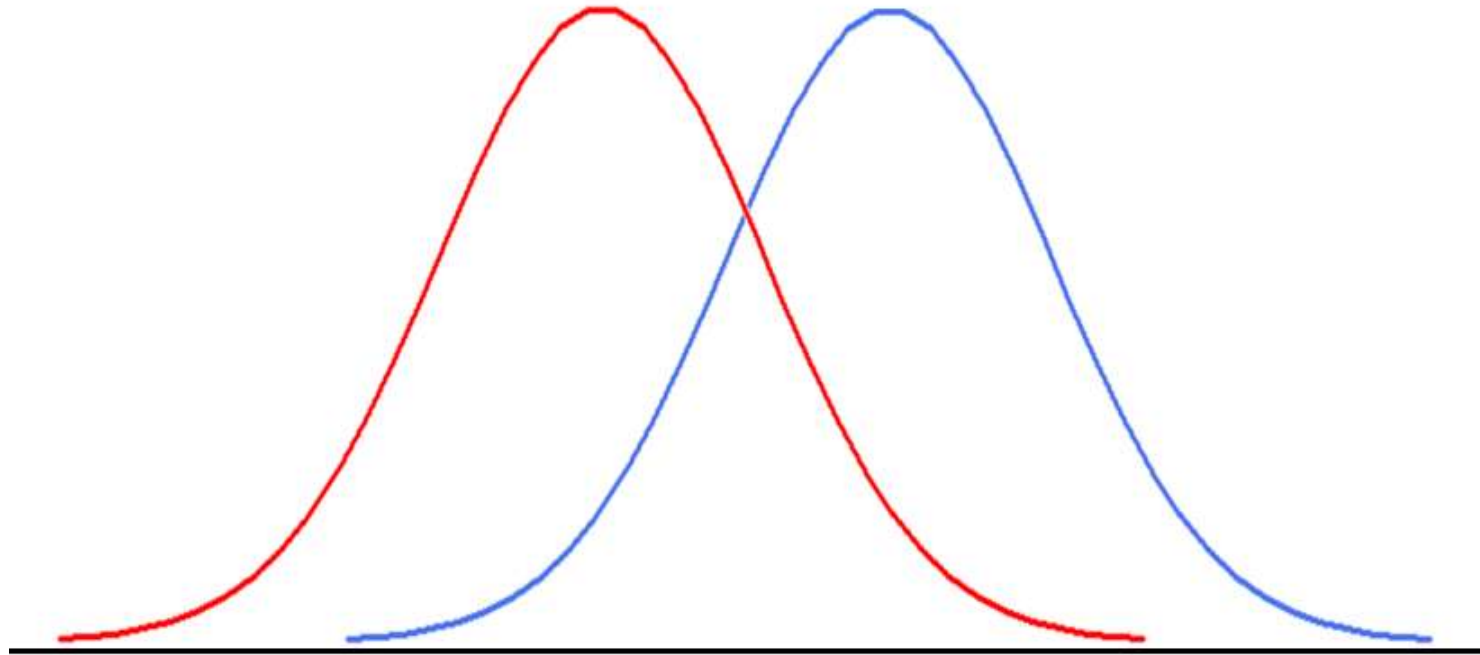
Consider the two normal distributions shown below:



Both distributions are normal with equal means μ , but unequal standard deviations $\sigma_1 > \sigma_2$.

The Normal Distribution

The two normal distributions shown below have equal standard deviations σ , but unequal means $\mu_2 > \mu_1$.



The Normal Distribution

If a variable X has a normal distribution with mean μ and standard deviation σ , we write

$$X \sim N(\mu, \sigma)$$

The symbol " \sim " indicates that the variable X **follows a normal distribution** (indicated by " N ").

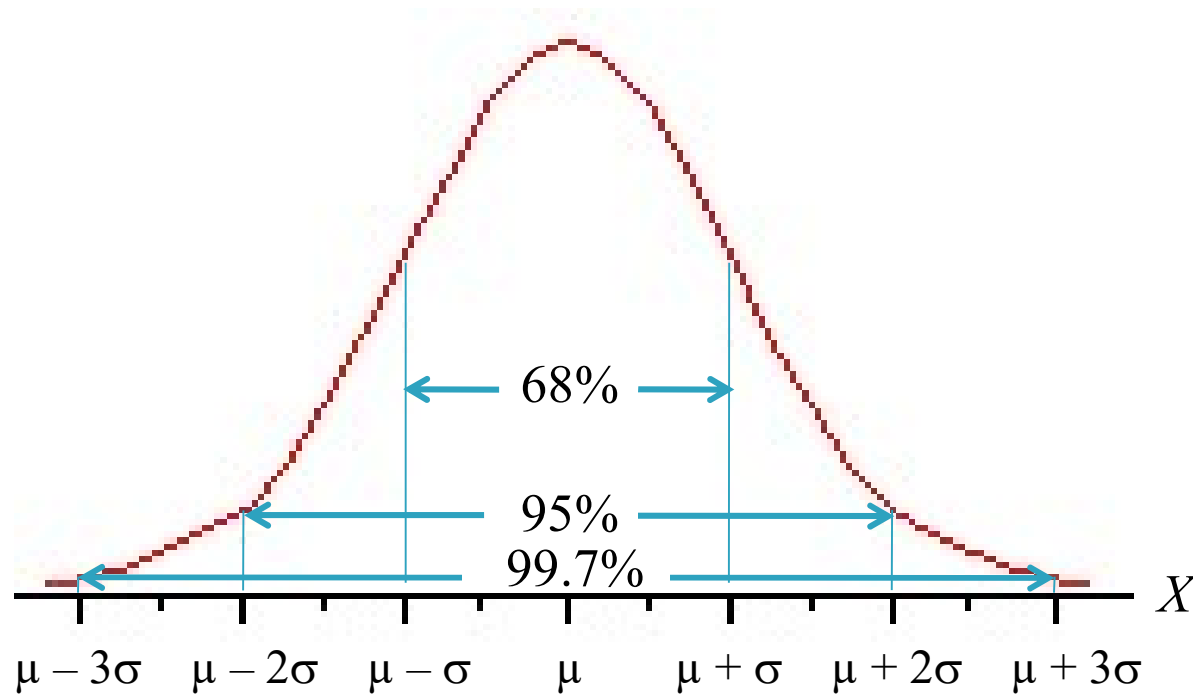
68-95-99.7 Rule

Despite the different means and standard deviations, all normal distributions have the following rule in common: **Approximately**

- 68% of all values fall within one standard deviation of the mean ($\mu \pm \sigma$).
- 95% of all values fall within two standard deviations of the mean ($\mu \pm 2\sigma$).
- 99.7% of all values fall within three standard deviations of the mean ($\mu \pm 3\sigma$).

68-95-99.7 Rule

This is called the **68-95-99.7 rule**.



The Normal Distribution

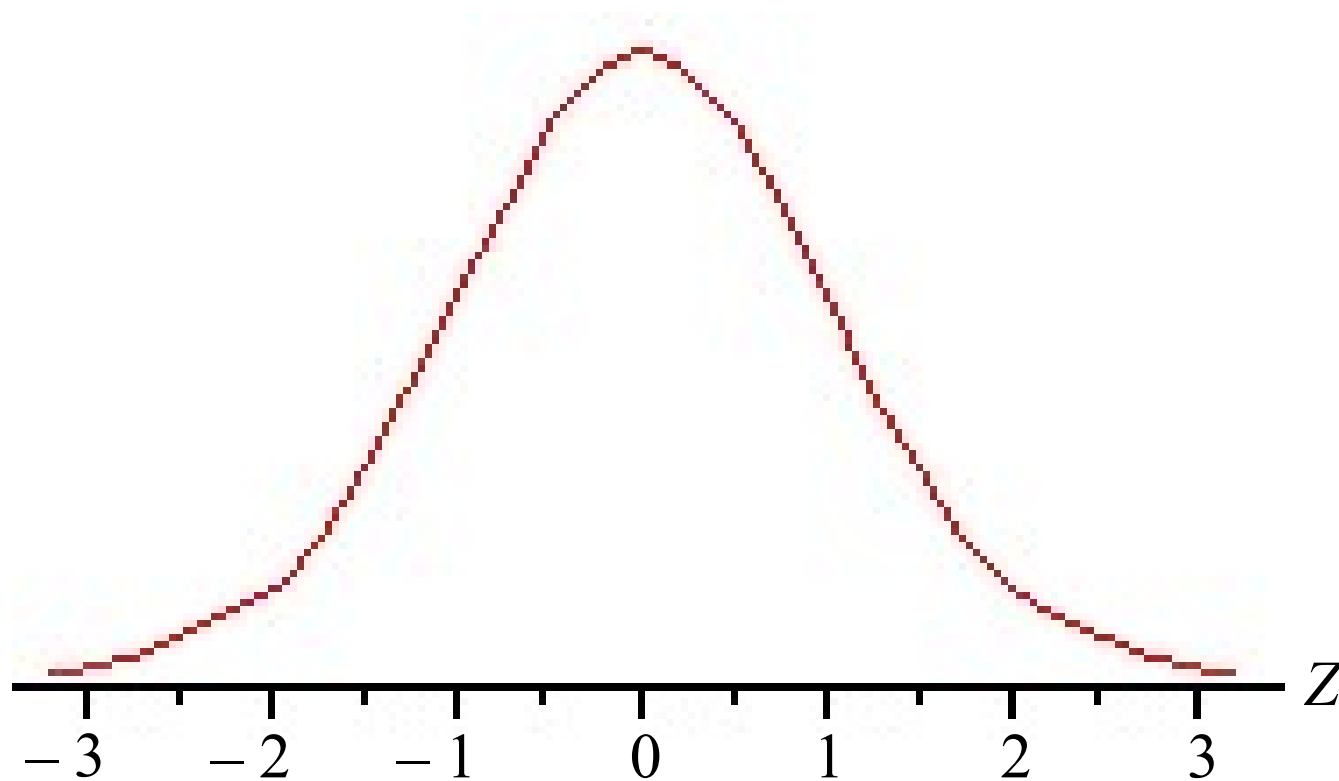
So if we can determine how many standard deviations an observation is away from its mean, we can determine the approximate areas under the curve corresponding to certain intervals.

If we know that $X \sim N(\mu, \sigma)$ and we have an observation x , then how many standard deviations is x away from μ ?

Standard Normal Distribution

There is one very special normal distribution that we will work with directly to find probabilities. This distribution is special because it has a mean of $\mu = 0$ and a standard deviation of $\sigma = 1$. We call this the **standard normal distribution** and denote by **Z** the variable which it represents.

Standard Normal Distribution



Standard Normal Distribution

In fact, we can “transform” any normal variable to have this distribution. This is very helpful, because the standard normal distribution is much easier to work with, and it serves as a frame of reference for all other normal distributions.

Standard Normal Distribution

To transform a normal variable X into a standard normal variable Z , we find the **z-score** (also called the **standardized value of X**). For a given observed value x , the z-score is calculated as

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

Standard Normal Distribution

The numerator tells us how far x is away from μ , and the denominator scales this value to be in standard deviation units.

And so z is in fact the number of standard deviations an observation x is from the mean μ !

Example

Suppose it is known that heights X of adult male Canadians follow a normal distribution with mean 178 cm and standard deviation 6 cm,

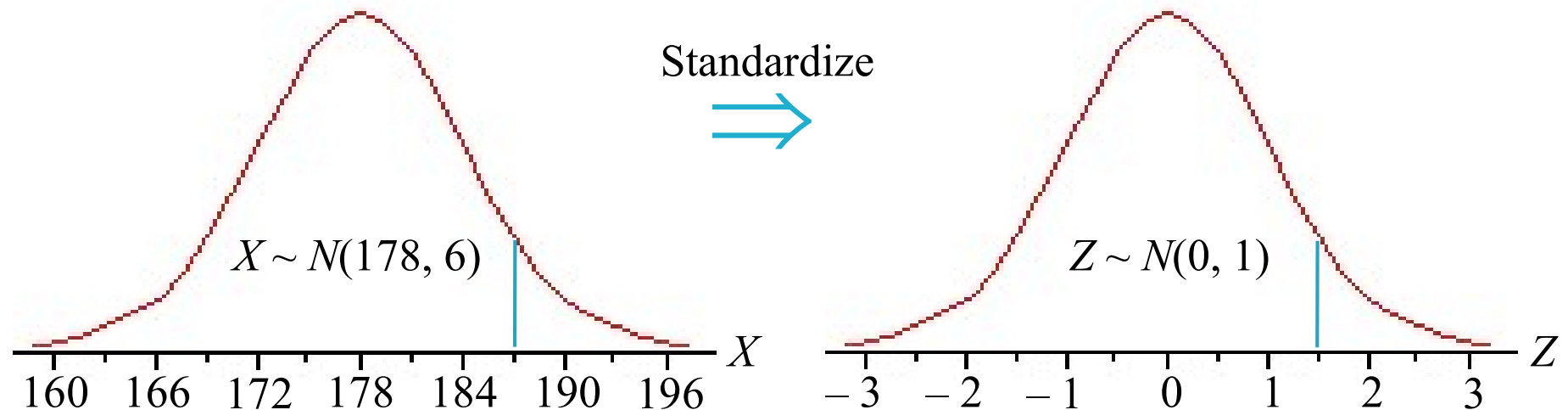
$$X \sim N(178, 6)$$

We measure one man's height to be 187 cm. The z-score for this man's height is

$$z = \frac{x - \mu}{\sigma} = \frac{187 - 178}{6} = 1.5$$

Example

In other words, this man's height is 1.5 standard deviations above the population mean.



Practice Question

Suppose that GPAs at a large university follow a normal distribution with mean 3.10 and standard deviation 0.40. One student has a GPA of 2.30. What is the z-score for this student?

- (A) -0.80 (B) 2.00 (C) -0.50 (D) 0.80 (E) -2.00

Example

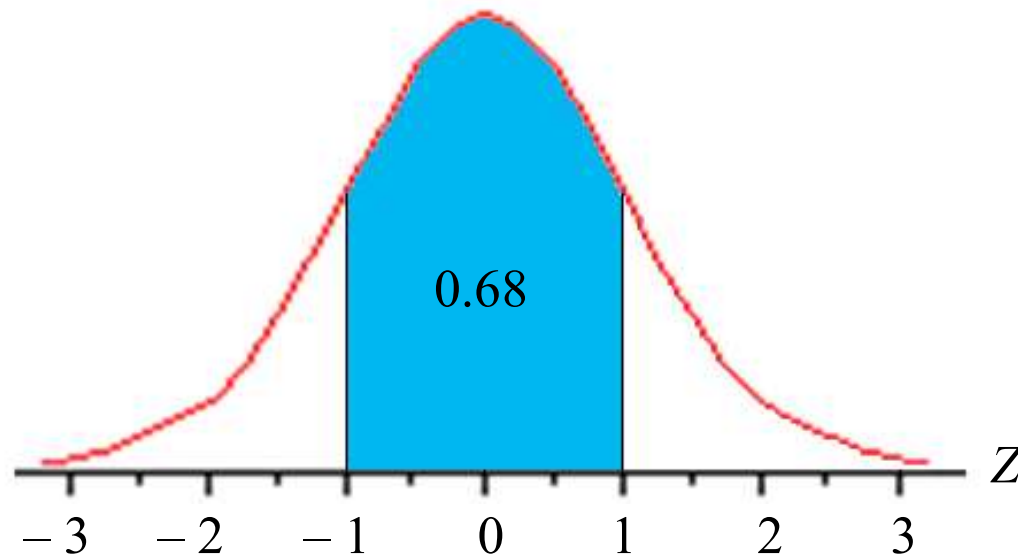
Find the approximate proportions under the standard normal curve:

Note: When finding proportions for any distribution, it is always helpful to sketch the distribution and shade in the area of interest. This will help you visualize the problem to ensure you know what area you are looking for.

Example

(i) $P(-1 < Z < 1)$

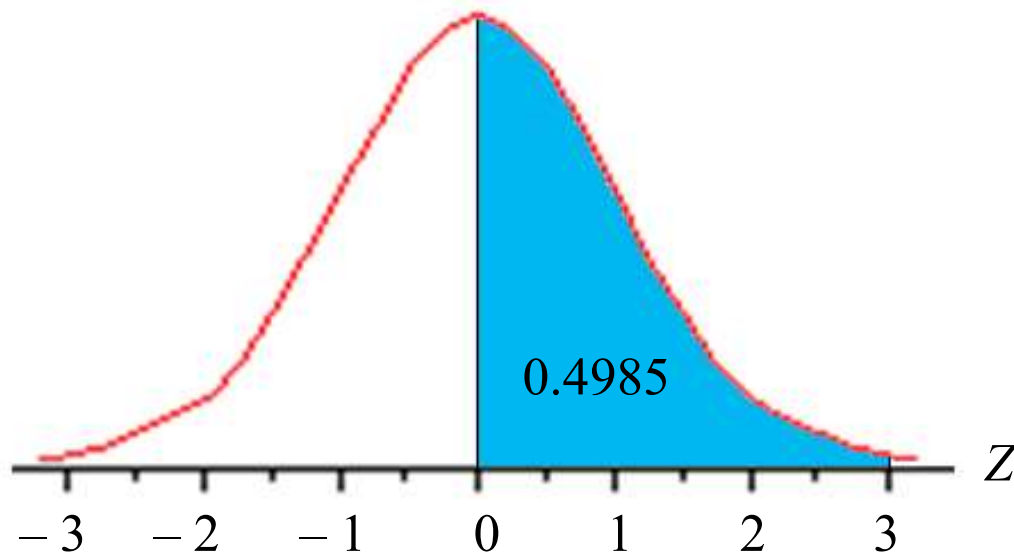
Answer: By definition of our rule, the approximate proportion of observations falling within one standard deviation of the mean ($\mu = 0$) is $P(-1 < Z < 1) \approx \mathbf{0.68}$.



Example

(ii) $P(0 < Z < 3)$

Answer: We know that $P(-3 < Z < 3) \approx 0.997$, and so by the symmetry of the normal distribution, we know that $P(0 < Z < 3) \approx 0.997/2 = \mathbf{0.4985}$.



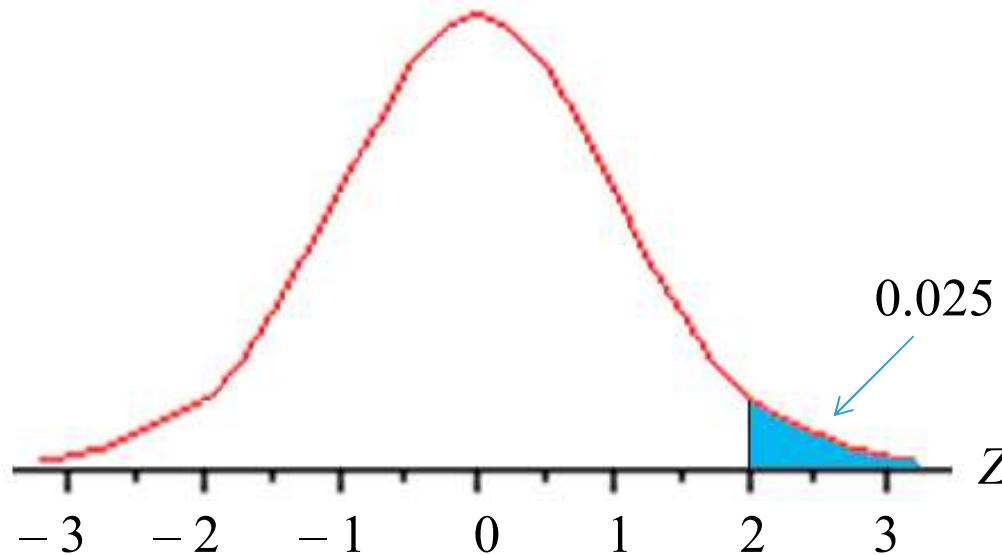
Example

(iii) $P(Z > 2)$

Answer: We know that $P(-2 < Z < 2) \approx 0.95$.

Therefore, $P(Z < -2) + P(Z > 2) \approx 1 - 0.95 = 0.05$.

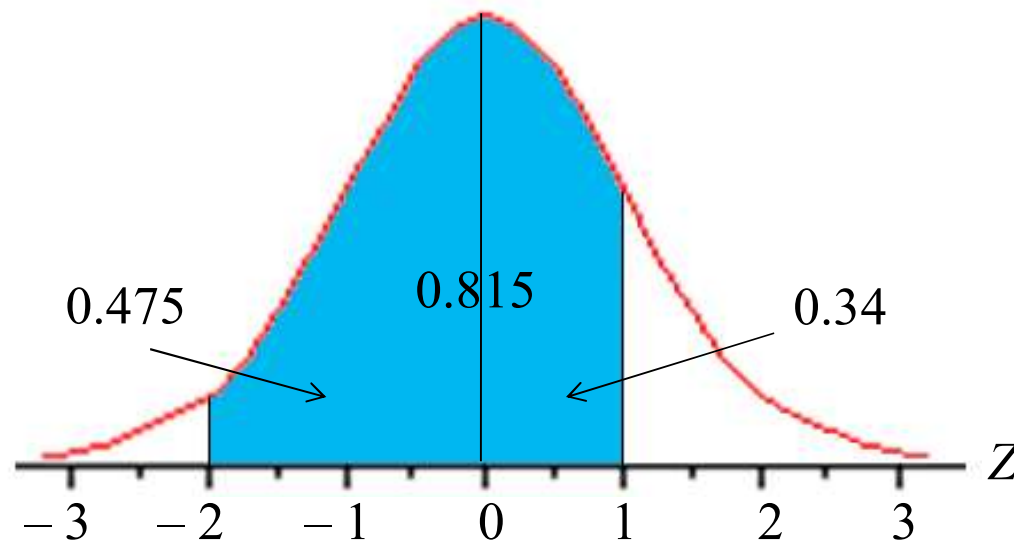
The two areas are equal, so $P(Z > 2) \approx 0.05/2 = \mathbf{0.025}$.



Example

(iv) $P(-2 < Z < 1)$

Answer: We divide this proportion into two areas and we get $P(-2 < Z < 1) = P(-2 < Z < 0) + P(0 < Z < 1) \approx 0.95/2 + 0.68/2 = 0.475 + 0.34 = \mathbf{0.815}$.



Practice Question

What is the approximate proportion $P(Z < -1)$?

- (A) 0.16 (B) 0.17 (C) 0.32 (D) 0.34 (E) 0.84

The Normal Distribution

We can handle such problems for any normal variable $X \sim N(\mu, \sigma)$ using the 68-95-99.7 rule by first standardizing the values of x to find the interval of interest in terms of z .

The standard normal distribution will serve as a frame of reference, which we will use to find all of our proportions for any normal variable (since Z is defined as the number of standard deviations an observation is away from its mean, and this is exactly what the rule applies to).

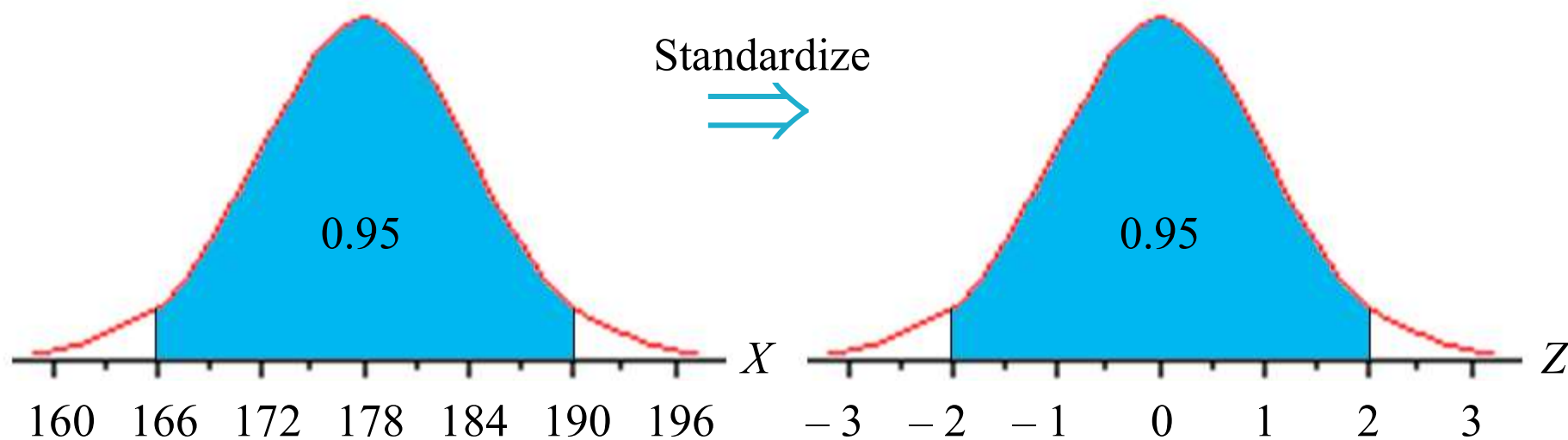
Example

Consider again the distribution of male heights $X \sim N(178, 6)$.

- (i) What proportion of Canadian adult males are between 166 and 190 centimetres tall?

$$\begin{aligned} \text{Answer: } P(166 < X < 190) \\ &= P\left(\frac{166 - 178}{6} < \frac{X - \mu}{\sigma} < \frac{190 - 178}{6}\right) \\ &= P(-2 < Z < 2) \approx \mathbf{0.95}. \end{aligned}$$

Example



Example

Note that we standardized the values of X by subtracting μ from X and then dividing that quantity by σ . We have transformed the variable X into the standard normal variable Z and so the problem is now rephrased in terms of the standard normal distribution.

Note that we did exactly the same thing to all quantities inside the brackets to maintain the inequality.

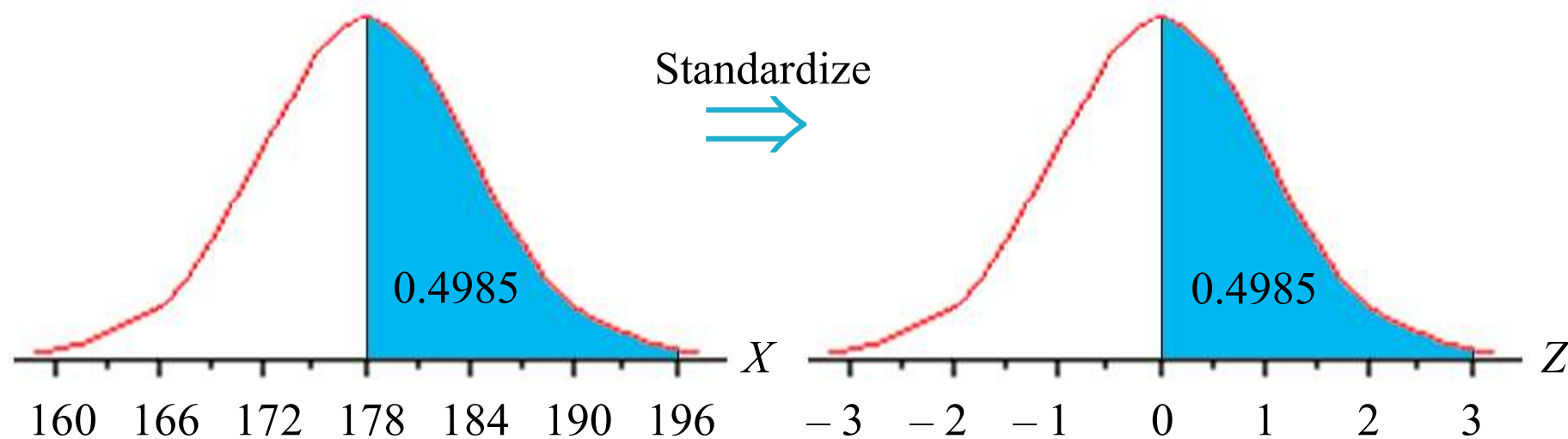
Example

- (ii) What proportion of Canadian adult males are between 178 and 196 centimetres tall?

Answer: $P(178 < X < 196)$

$$\begin{aligned} &= P\left(\frac{178 - 178}{6} < \frac{X - \mu}{\sigma} < \frac{196 - 178}{6}\right) \\ &= P(0 < Z < 3) \approx \mathbf{0.4985}. \end{aligned}$$

Example



Practice Question

Scores on a biology exam follow a normal distribution with mean 70. It is known that approximately 95% of students score between 50 and 90. The standard deviation of scores is approximately equal to:

- (A) 6.7 (B) 10 (C) 13.3 (D) 20 (E) 40

Practice Question

Which of the following normal variables has the density curve with the highest peak?

(A) $X \sim N(1, 3)$

(B) $X \sim N(2, 1)$

(C) $X \sim N(3, 4)$

(D) $X \sim N(4, 2)$

(E) All four curves have the same height.

The Normal Distribution

Note that, for any continuous variable X , the proportion $P(X \leq x)$ is **exactly equal** to the proportion $P(X < x)$!

Why?

The proportion $P(X \leq x)$ is equal to the proportion $P(X < x)$ **plus** the proportion $P(X = x)$. But this latter proportion is **zero**, as it corresponds to the area of a line. So for continuous distributions, it doesn't matter if we are looking for $P(X \leq x)$ or $P(X < x)$.

The Normal Distribution

We are now ready to be able to find **any** proportion corresponding to **any** interval for **any** normal distribution.

Table 1 gives us the proportions under the standard normal curve to the **left** of any value z .

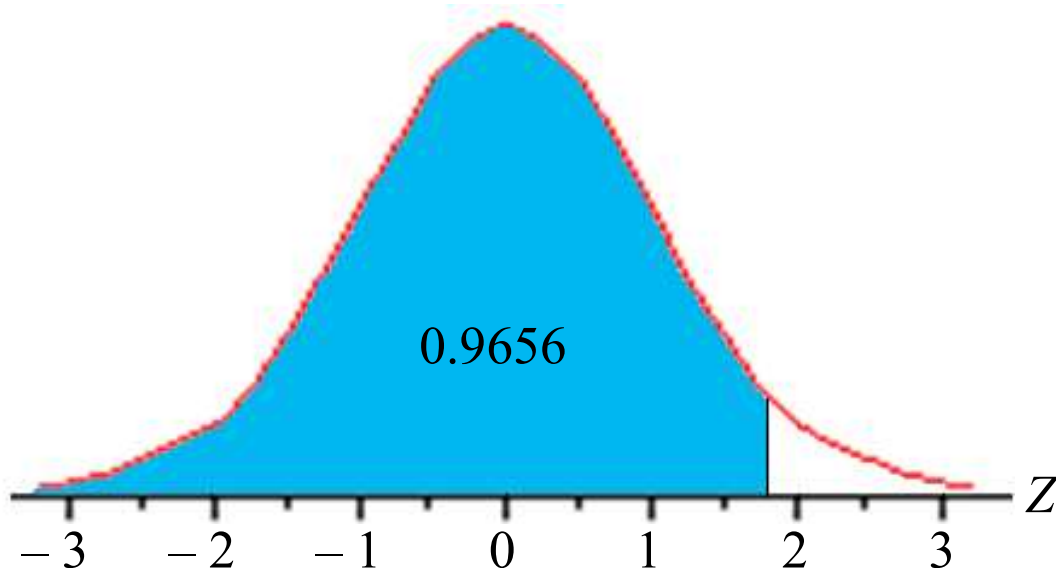
Normal Tables

Table 1 consists of two pages – one for negative values of z and one for positive values.

The numbers in the column under the " z " at the left side of the page represent the units and first decimal place, while the values in the row to the right of " z " represent the second decimal place. The value at the intersection of each row and column is the corresponding proportion to the left of the value z .

Normal Tables

For example, the entry at the intersection of 1.8 and 0.02 is 0.9656. This represents the **exact** proportion under the standard normal curve to the left of the value $z = 1.82$, i.e., $P(Z < 1.82) = 0.9656$.



Normal Tables

The table entries are listed only for values of z between -3.49 and 3.49 . This is because very close to 100% of the area under the standard normal curve falls within this interval. As such, we can take the proportions less than -3.49 or greater than 3.49 to be **zero**. For example,

$$P(Z > 5.17) \approx 0$$

$$P(Z \leq -3.85) \approx 0$$

$$P(Z \geq -6.22) \approx 1$$

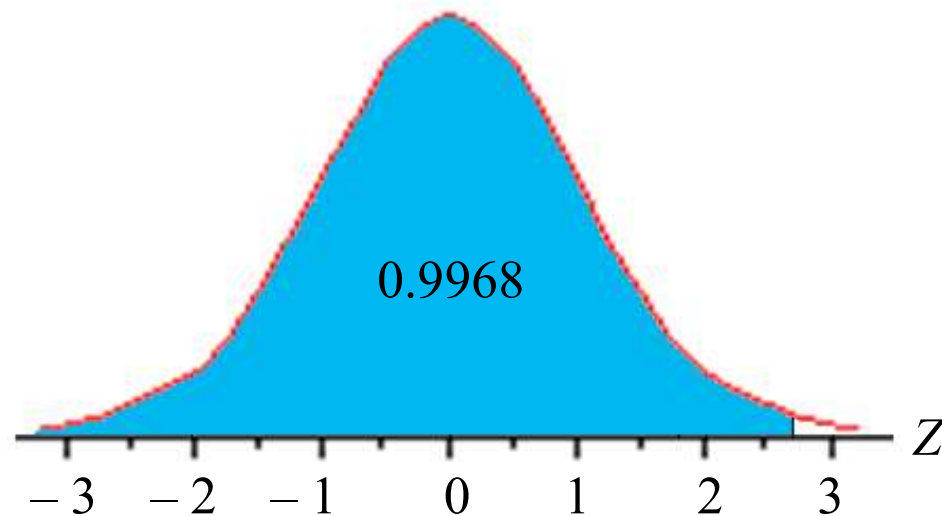
$$P(Z < 4.73) \approx 1$$

Example

Find the proportion under the standard normal curve corresponding to the following intervals:

(i) $P(Z < 2.73)$

Answer: $P(Z < 2.73) = \mathbf{0.9968}$



R Code

```
> pnorm(2.73)  
[1] 0.9968333
```

Example

(ii) $P(Z > 1.16)$

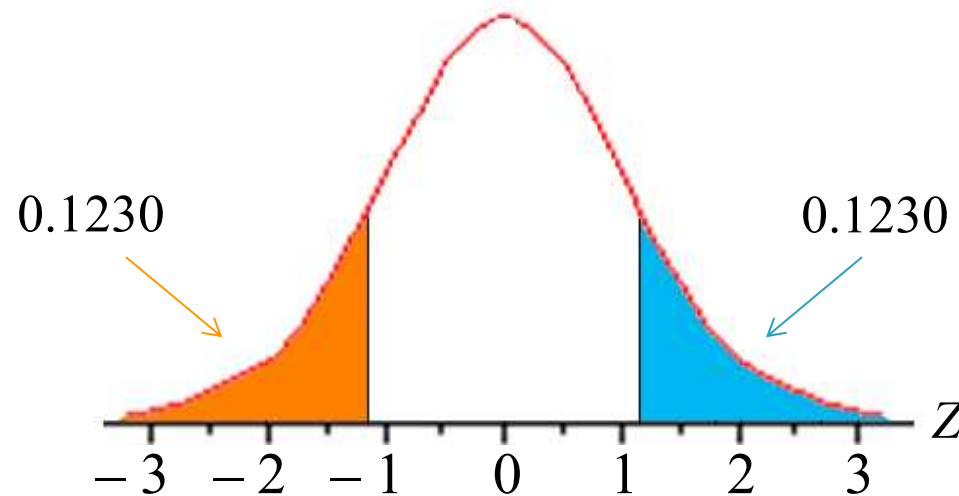
Answer: We cannot just take the proportion for the $z = 1.16$ entry in Table 1, because this gives us the proportion to the left of 1.16 and we want the proportion to the right.

We can solve this problem in one of two ways:

Example

- 1) We can use the symmetry of the normal distribution.

The area to the right of $z = 1.16$ is equivalent to the area to the left of $z = -1.16$.

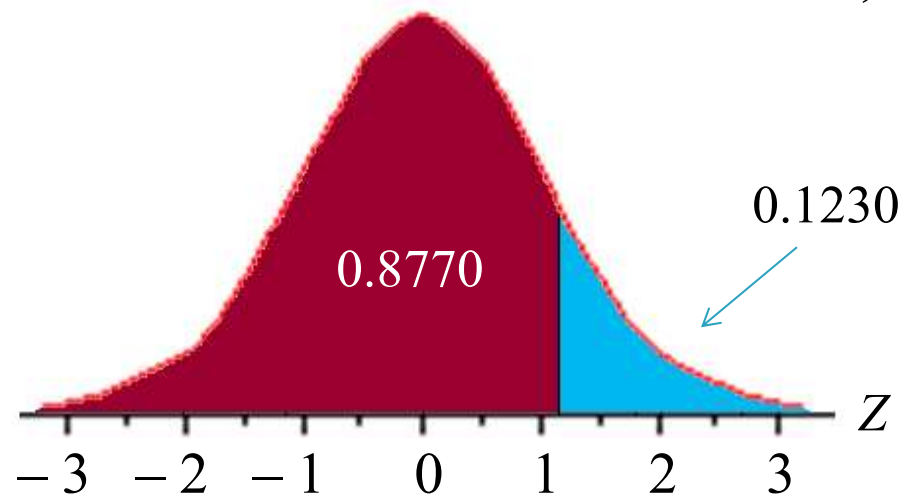


So $P(Z > 1.16) = P(Z < -1.16) = 0.1230$.

Example

- 2) We can use the fact that the entire area under the curve is one.

From Table 1, we see that $P(Z < 1.16) = 0.8770$.



So $P(Z > 1.16) = 1 - P(Z < 1.16)$
 $= 1 - 0.8770 = 0.1230.$

R Code

```
> pnorm(1.16, lower.tail = FALSE)  
[1] 0.1230244
```

```
> 1 - pnorm(1.16)  
[1] 0.1230244
```

Practice Question

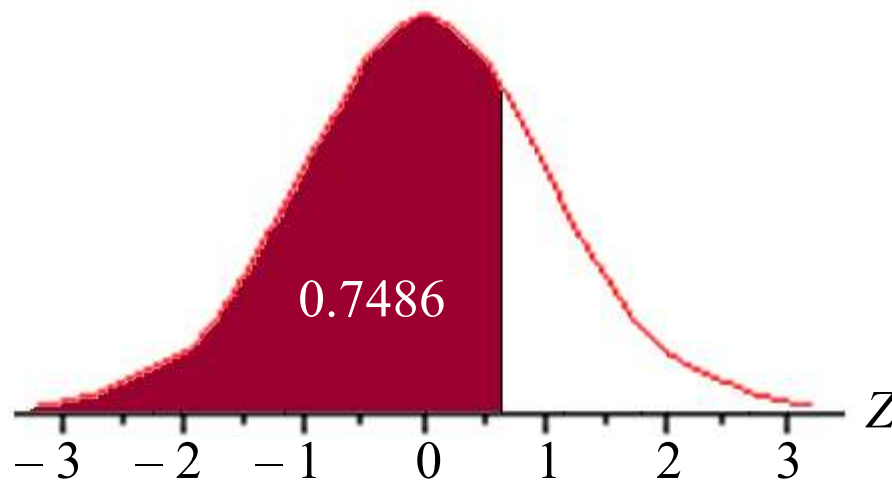
What is the proportion $P(Z > -0.48)$?

- (A) 0.4322
- (B) 0.3156
- (C) 0.5940
- (D) 0.2748
- (E) 0.6844

Example

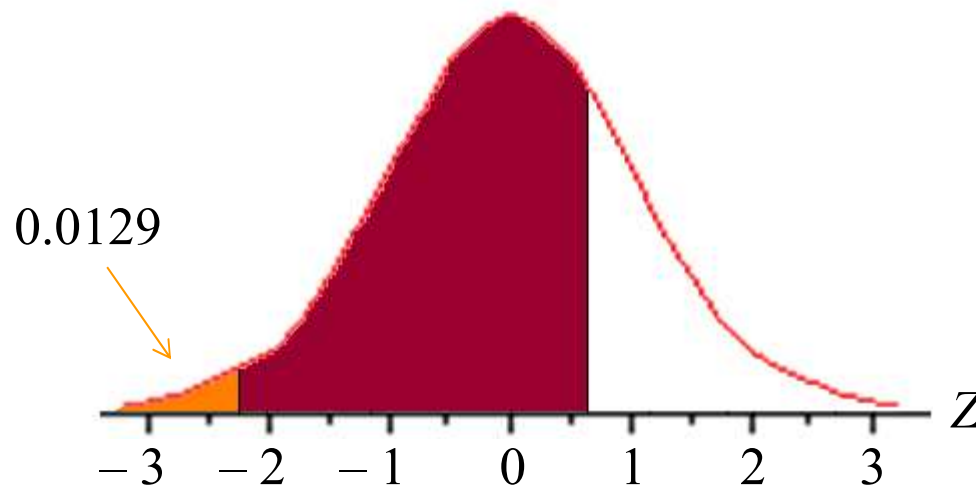
(iii) $P(-2.23 < Z < 0.67)$

Answer: Sketching the areas for this type of question is particularly helpful. First, we see that **$P(Z < 0.67) = 0.7486$** .



Example

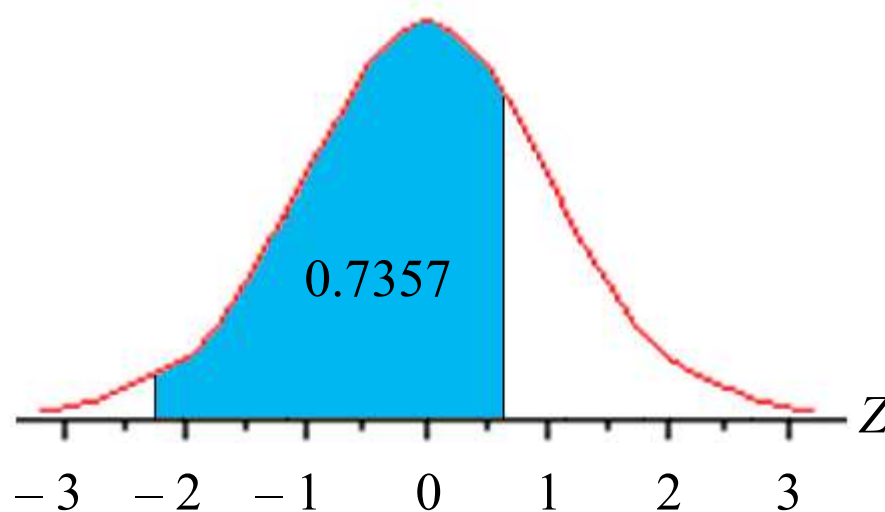
But this area is too large, as it also includes the area below -2.23 , which is where we want to cut it off. So we need to subtract $P(Z < -2.23) = 0.0129$.



Example

So the proportion of interest is equal to

$$\begin{aligned} P(-2.23 < Z < 0.67) &= P(Z < 0.67) - P(Z < -2.23) \\ &= 0.7486 - 0.0129 = \mathbf{0.7357}. \end{aligned}$$



R Code

```
> pnorm(0.67) - pnorm(-2.23)  
[1] 0.7356974
```


Practice Question

What is the proportion $P(0.40 \leq Z \leq 2.77)$?

- (A) 0.3418
- (B) 0.2987
- (C) 0.4210
- (D) 0.5612
- (E) 0.8263

The Normal Distribution

We have now seen several examples finding proportions under the standard normal curve.

In summary, for any values b and c ($b \leq c$):

- $P(Z < b) = \text{Table entry corresponding to } z = b$
- $P(Z > b) = 1 - \text{Table entry for } b$
- $P(b < Z < c) = \text{Table entry for } c - \text{Table entry for } b$
- $P(Z = b) = 0$

Our goal is always to rephrase the problem so that it only involves areas to the **left** so we can use Table 1.

The Normal Distribution

Now that we know how to find any area, we can find the **actual** proportions we approximated with the 68-95-99.7 rule:

$$\begin{aligned} P(-1 < Z < 1) &= P(Z < 1) - P(Z < -1) \\ &= 0.8413 - 0.1587 = \mathbf{0.6826} \end{aligned}$$

Similarly (check as an exercise):

$$\begin{aligned} P(-2 < Z < 2) &= \mathbf{0.9544} \quad \text{and} \\ P(-3 < Z < 3) &= \mathbf{0.9974} \end{aligned}$$

“Backwards” Normal - Percentiles

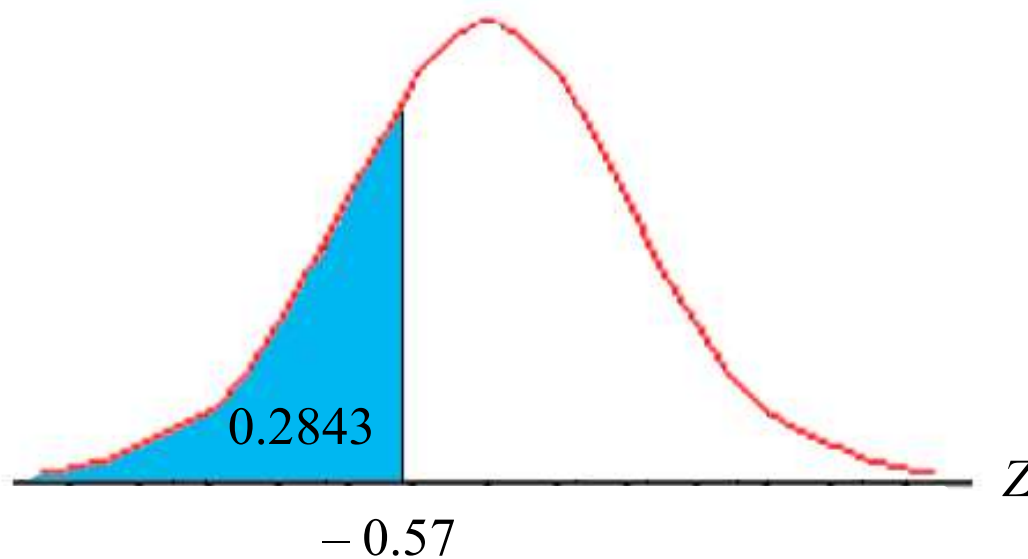
Now suppose we want to go “backwards”, i.e., find the value z such that the proportion $P(Z < z)$ is equal to some specified value.

- (i) Find the value z such that $P(Z < z) = 0.2843$.

Answer: Rather than looking in Table 1 at a specific value of z , we search the body of the table for the proportion 0.2843 and determine to which value of z this proportion corresponds.

Example

We see from Table 1 that $P(Z < -0.57) = 0.2843$, and so $z = -0.57$.



R Code

```
> qnorm(0.2843)  
[1] -0.5701146
```

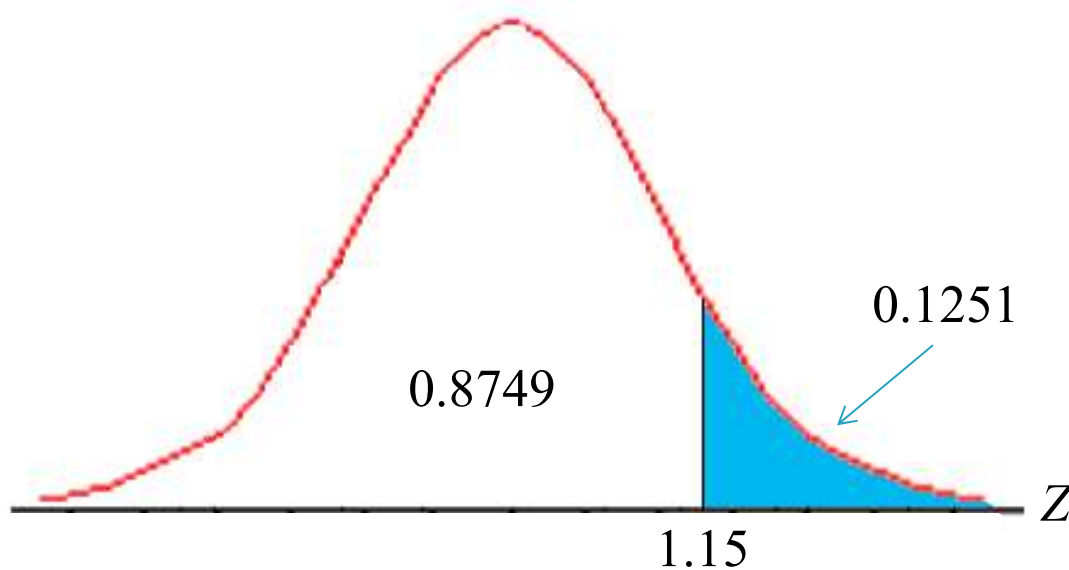
Example

(ii) Find the value z such that $P(Z > z) = 0.1251$.

Answer: We can't just search for the value 0.1251 in Table 1, because this will give us the value of z with area 0.1251 to its **left**. We want the value with 0.1251 to its **right**, and so we search for the value of z with $1 - 0.1251 = 0.8749$ to its left, i.e., $P(Z < z) = 0.8749$.

Example

We see from Table 1 that $P(Z < 1.15) = 0.8749$, and so $z = 1.15$.



R Code

```
> qnorm(0.1251, lower.tail = FALSE)  
[1] 1.149864
```

```
> qnorm(0.8749)  
[1] 1.149864
```

Example

- (iii) Find the interquartile range of the standard normal distribution.

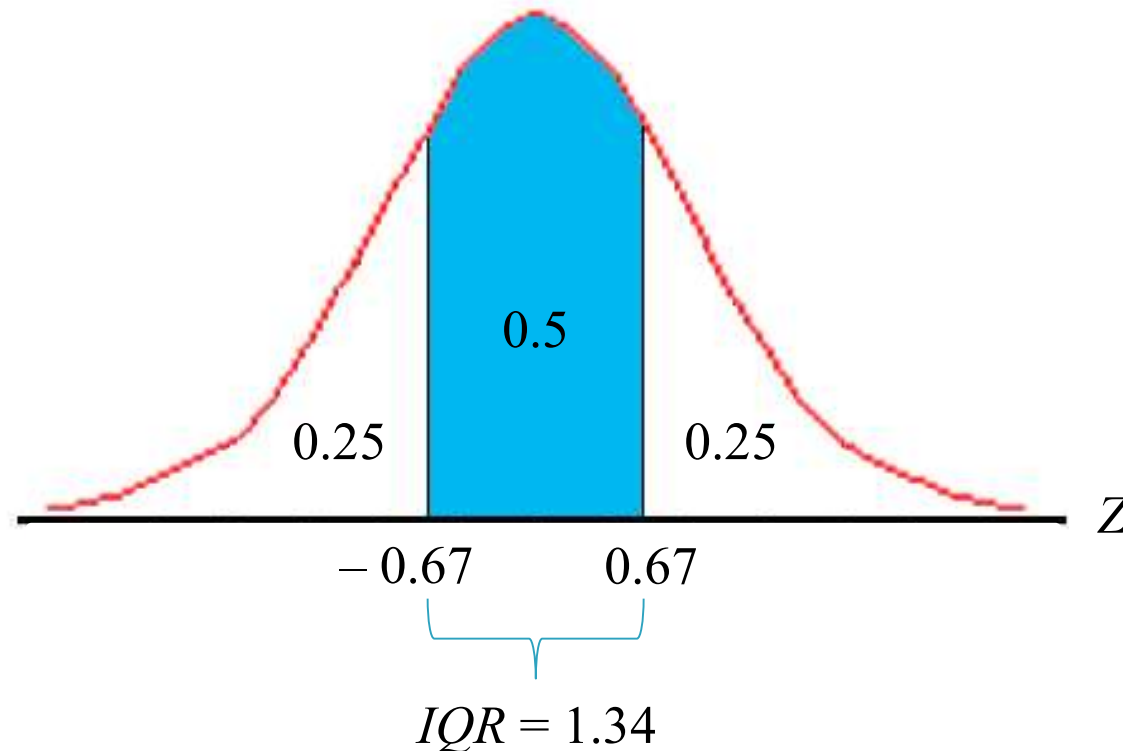
Answer: The first quartile is the value of z such that $P(Z < z) = 0.2500$. We see from Table 1 that this value of z is closest to $z = -0.67$, and so $Q1 = -0.67$. Similarly, by symmetry, the value of z with area 0.7500 to its left is $Q3 = 0.67$.

(If the exact proportion is not on the table, take the closest value of z . If the proportion is exactly halfway between two values on the table, take the average of the two z values to either side.)

Example

The interquartile range is therefore

$$IQR = Q3 - Q1 = 0.67 - (-0.67) = \mathbf{1.34}.$$



R Code

```
> qnorm(0.75) - qnorm(0.25)  
[1] 1.34898
```

Practice Question

What is the value z such that $P(-z < Z < z) = 0.8664$?

- (A) $z = 1.83$
- (B) $z = 0.78$
- (C) $z = 1.50$
- (D) $z = 1.11$
- (E) $z = 2.04$

Normal Distribution – Exact Proportions

We can find proportions for any normal distribution by first standardizing our normal variable X .

For example, suppose it is known that pulse rates of adult females follow a normal distribution with mean 74 beats per minute and standard deviation 12 beats per minute.

Example

- (i) What proportion of adult females have pulse rates above 57 beats per minute?

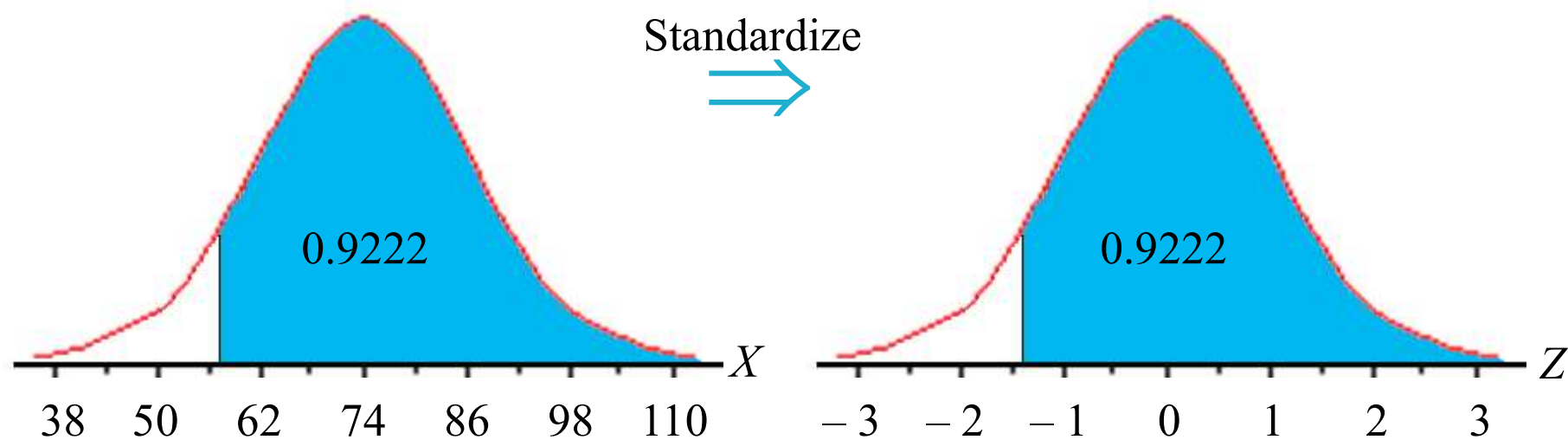
$$\begin{aligned}\text{Answer: } P(X > 57) &= P\left(Z > \frac{57 - 74}{12}\right) \\ &= P(Z > -1.42) = 1 - P(Z < -1.42) \\ &= 1 - 0.0778 = 0.9222.\end{aligned}$$

R Code

```
> pnorm(57, 74, 12, lower.tail = FALSE)  
[1] 0.9217098
```

```
> 1 - pnorm(57, 74, 12)  
[1] 0.9217098
```


Example



Example

Note that we can think of these proportions we're calculating as probabilities:

The **proportion** of adult females with pulse rates above 57 beats per minute is 0.9222.

Therefore, if we randomly select an adult female, the **probability** her pulse rate **will be** above 57 beats per minute is 0.9222.

Example

- (ii) In a sample of 500 adult females, about how many would we expect to have a pulse rate over 57 beats per minute?

Answer: We would expect about 92.22% , or $0.9222(500) \approx 461$ of them to have a pulse rate over 57 beats per minute.

Example

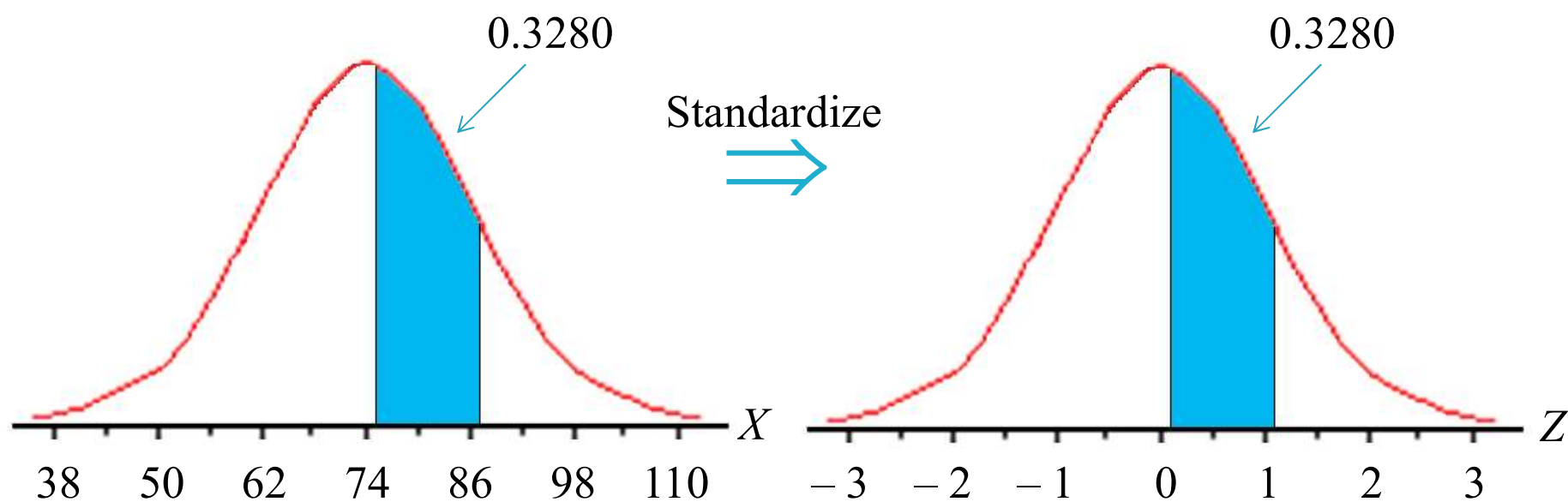
(iii) What proportion of adult females have pulse rates between 75 and 87 beats per minute?

$$\text{Answer: } P(75 < X < 87) = P\left(\frac{75 - 74}{12} < Z < \frac{87 - 74}{12}\right)$$

$$= P(0.08 < Z < 1.08) = P(Z < 1.08) - P(Z < 0.08)$$

$$= 0.8599 - 0.5319 = 0.3280$$

Example



R Code

```
> pnorm(87, 74, 12) - pnorm(75, 74, 12)  
[1] 0.327463
```

Practice Question

Weights of bulldogs follow a normal distribution with mean 40 pounds and standard deviation 8 pounds.

What proportion of bulldogs weigh less than 50 pounds?

(A) 0.8508

(B) 0.8218

(C) 0.9582

(D) 0.8944

(E) 0.9147

Practice Question

Weights of bulldogs follow a normal distribution with mean 40 pounds and standard deviation 8 pounds.

What proportion of bulldogs weigh more than 36 pounds?

(A) 0.6915

(B) 0.5948

(C) 0.3085

(D) 0.7642

(E) 0.4090

Practice Question

Weights of bulldogs follow a normal distribution with mean 40 pounds and standard deviation 8 pounds.

What proportion of bulldogs weigh between 32 and 54 pounds?

- | | | |
|------------|------------|------------|
| (A) 0.7043 | (B) 0.9068 | (C) 0.6077 |
| (D) 0.5029 | (E) 0.8012 | |

“Backwards” Normal – Finding x

(iv) The lowest 10% of pulse rates are below what value?

Answer: Rather than being given a value of x and asked to find a proportion, we are now given the proportion and asked to find the value of x . We want to find the value of x such that $P(X < x) = 0.10$, i.e., the tenth percentile of the distribution of X . We will do this in two steps.

Example

Step 1: Find the value z such that $P(Z < z) = 0.10$.

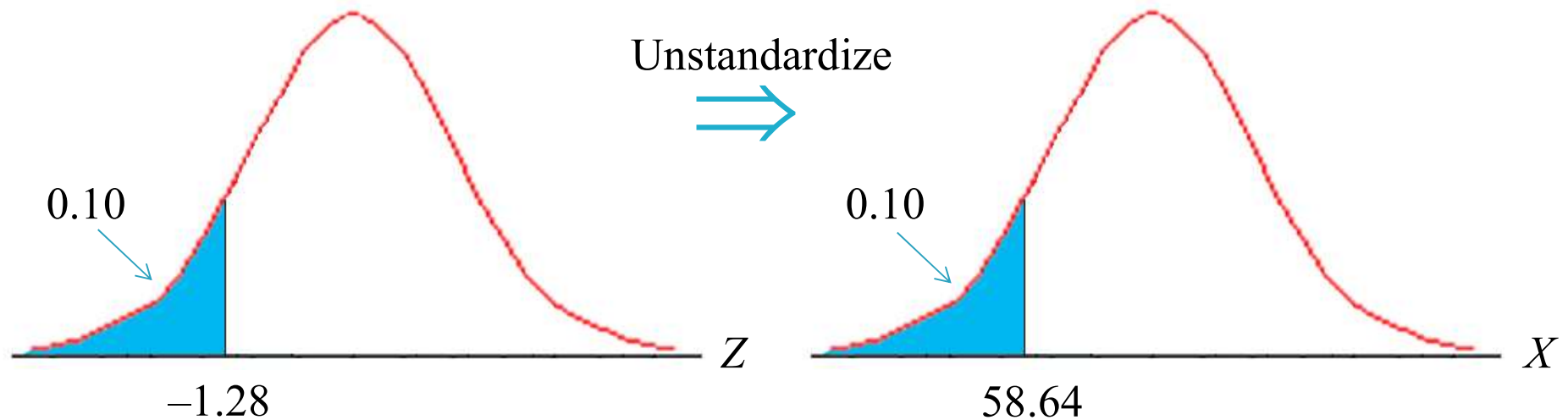
We see from Table 1 that this corresponds to the value $z = -1.28$.

Step 2: Transform z into x :

$$\begin{aligned} z &= \frac{x - \mu}{\sigma} \Rightarrow x = \mu + z\sigma \\ &= 74 + (-1.28)(12) = 58.64 \end{aligned}$$

Example

We have thus determined that $P(X < 58.64) = 0.10$,
i.e., 10% of pulse rates are below 58.64.



R Code

```
> qnorm(0.10, 74, 12)  
[1] 58.62138
```

Example

GPA's of students at a large college follow a normal distribution with mean 3.32 and standard deviation 0.50. In order to graduate with *summa cum laude* honours, a student's GPA must be in the top 3%. What is the minimum GPA for a student to graduate with these honours?

Example

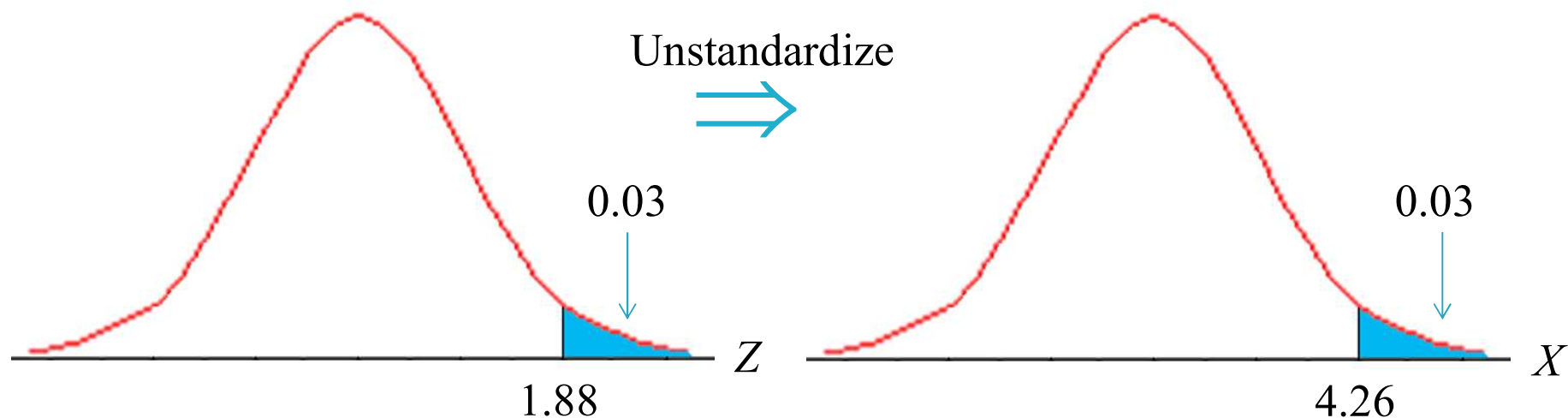
Step 1: Find the value z such that $P(Z > z) = 0.03$
 $\Rightarrow P(Z < z) = 0.97$. We see from Table 1 that this corresponds to the value $z = 1.88$.

Step 2: Transform z into x :

$$\begin{aligned} z &= \frac{x - \mu}{\sigma} \Rightarrow x = \mu + z\sigma \\ &= 3.32 + 1.88(0.50) = 4.26 \end{aligned}$$

Example

We have thus determined that $P(X > 4.26) = 0.03$,
i.e., only 3% of GPAs are above 4.26.



R Code

```
> qnorm(0.03, 3.32, 0.50, lower.tail = FALSE)  
[1] 4.260397
```

```
> qnorm(0.97, 3.32, 0.50)  
[1] 4.260397
```

Practice Question

Scores on an exam follow a normal distribution with mean 73 and standard deviation 12. The professor decides to assign a grade of A+ to the students with the top 9% of scores. What is the minimum score required to receive a grade of A+?

- (A) 88.16 (B) 89.08 (C) 90.20 (D) 91.12 (E) 92.24

Comparing Normal Variables

Percentage grades in a large Economics class follow a normal distribution with mean 65 and standard deviation 10. Percentage grades in a large Psychology class follow a normal distribution with mean 72 and standard deviation 8.

One student got 80% in Economics and 82% in Psychology. In which class did the student do better relative to other students?

Example

To answer this question, we simply calculate the student's z-score for both classes:

$$z_E = \frac{x_E - \mu_E}{\sigma_E} = \frac{80 - 65}{10} = 1.50$$

$$z_P = \frac{x_P - \mu_P}{\sigma_P} = \frac{82 - 72}{8} = 1.25$$

Example

Although the student got a higher percentage grade in Psychology, her score in Economics was actually better relative to the rest of the class.

$$P(Z < 1.50) = 0.9332$$

$$P(Z < 1.25) = 0.8944$$

She did better than 93.32% of students in Economics and 89.44% of students in Psychology.

Practice Question

Heights of NBA players follow a normal distribution with mean 198 cm and standard deviation 7 cm. Heights of NHL players follow a normal distribution with mean 185 cm and standard deviation 5 cm.

Blake Griffin (an NBA player) is 205 cm tall. Mark Scheifele (an NHL player) is 190 cm tall. Which player is taller, relative to other players in their respective sports?

- (A) Griffin
- (B) Scheifele
- (C) Both players are equally tall relative to other players in their sports.

The Normal Distribution

We have seen how to find proportions corresponding to some area under the normal curve, as well as how to determine a value of x corresponding to a given proportion.

There are two other quantities in the z -score formula that we may be asked to find – the mean and standard deviation.

Normal Distribution – Finding σ

Weights of adult leopard seals are known to follow a normal distribution with mean 367 kg. It is known that 33% of all adult seals of this type weigh more than 400 kg. What is the standard deviation of the distribution of weights?

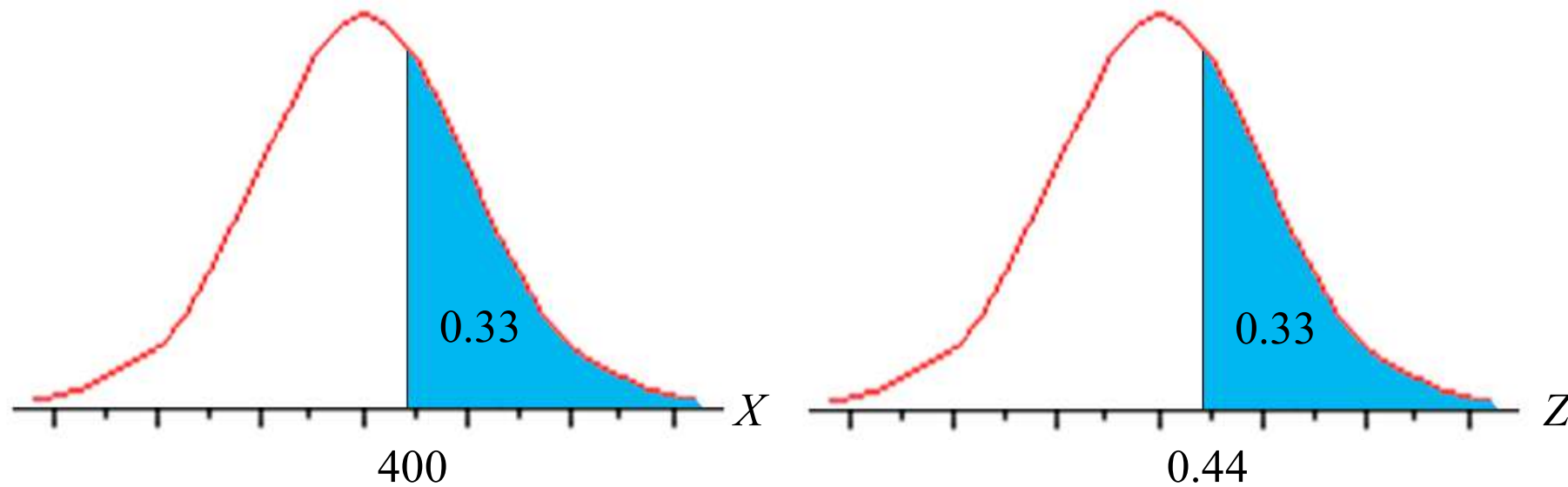
Example

Answer: We know $\mu = 367$ and $P(X > 400) = 0.33$, so $x = 400$. We find the value z such that $P(Z > z) = 0.33 \Rightarrow P(Z < z) = 1 - 0.33 = 0.67 \Rightarrow z = 0.44$.

We use the z -score formula to solve for the value of the standard deviation:

$$z = \frac{x - \mu}{\sigma} \Rightarrow \sigma = \frac{x - \mu}{z} = \frac{400 - 367}{0.44} = 75$$

Example



Normal Distribution – Finding μ

Labels on shampoo bottles claim the bottles contain 425 ml of shampoo. In fact, the bottles are filled by a machine, and fill volumes are known to be normally distributed with standard deviation 3 ml.

The machine has a dial to set the mean fill volume. To what value should the dial be set so that only 5% of bottles contain less than the amount on the label?

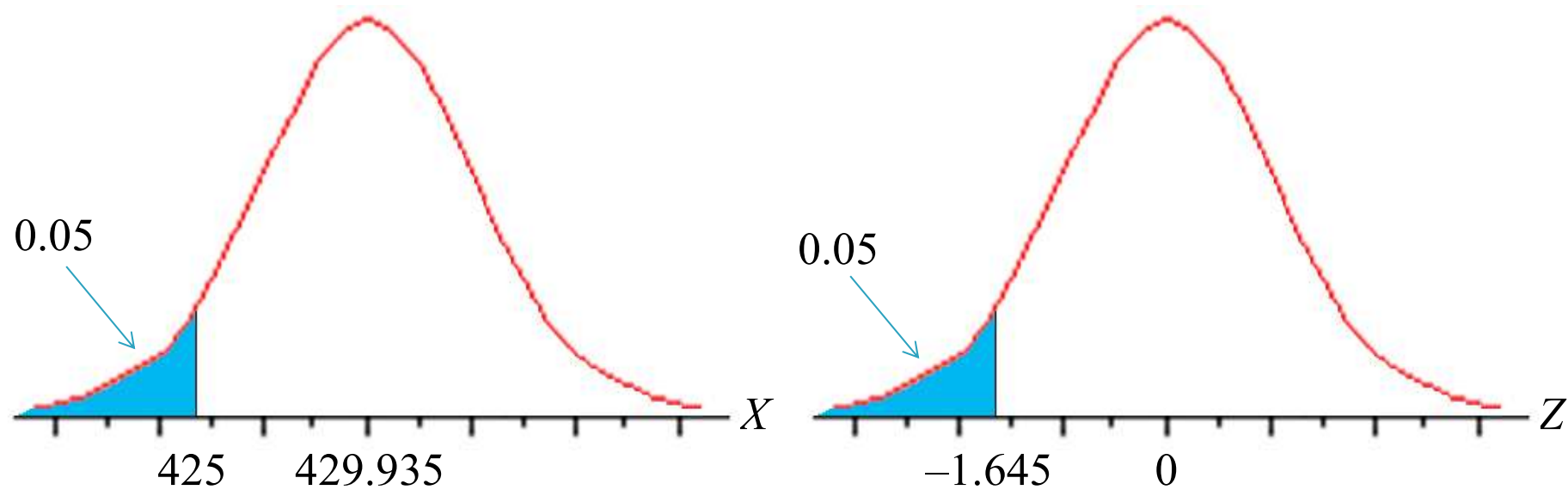
Example

Answer: We know $\sigma = 3$ and $P(X < 425) = 0.05$, so $x = 425$. We find the value z such that $P(Z < z) = 0.05 \Rightarrow z = -1.645$.

We use the z -score formula to solve for the value of the mean:

$$z = \frac{x - \mu}{\sigma} \Rightarrow \mu = x - z\sigma = 425 - (-1.645)(3) = 429.935$$

Example



Practice Question

A coffee machine is regulated so it dispenses an average of μ ounces per cup. If fill volumes are normally distributed with a standard deviation of 0.2 ounces, then what value should μ be set at so that 6-ounce cups will overflow only 2% of the time?

- (A) 5.78 ounces
- (B) 5.59 ounces
- (C) 5.86 ounces
- (D) 5.02 ounces
- (E) 6.41 ounces

The Normal Distribution

We have now seen examples of all four types of normal distribution problem. We could be asked to find any of the four quantities in the z-score formula:

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

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

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

$$\mu = x - z\sigma$$

Practice Question

Monthly electric bills for households in a large city follow a normal distribution with mean \$132 and standard deviation \$27.

- a) What proportion of households have a monthly electric bill less than \$99?

Practice Question

Monthly electric bills for households in a large city follow a normal distribution with mean \$132 and standard deviation \$27.

- b) What proportion of households have a monthly electric bill greater than \$188?

Practice Question

Monthly electric bills for households in a large city follow a normal distribution with mean \$132 and standard deviation \$27.

- c) What proportion of households have a monthly electric bill between \$80 and \$140?

Practice Question

Monthly electric bills for households in a large city follow a normal distribution with mean \$132 and standard deviation \$27.

- d) According to the 68-95-99.7 rule, approximately 99.7% of households have monthly electric bills between what two values?

Practice Question

Monthly electric bills for households in a large city follow a normal distribution with mean \$132 and standard deviation \$27.

- e) Only 2.5% of households have a monthly electric bill greater than what?

Practice Question

Monthly water bills for households in the city follow a normal distribution with standard deviation \$18.

- f) Only 20% of households have a monthly water bill less than \$56.88 per month. What is the mean monthly water bill for households in the city?

Practice Question

Monthly electric bills for households in a large city follow a normal distribution with mean \$132 and standard deviation \$27.

Monthly water bills for households in the city follow a normal distribution with mean \$72 and standard deviation \$18.

- g) One household has an electric bill of \$170 and a water bill of \$106. Which bill is higher, relative to other households in the city?

Normal Quantile Plots

Question: Throughout this unit, we have assumed that the variable of interest follows a normal distribution. How can we know this if we don't have information about the whole population?

Answer: We can't! This is an assumption, and our calculations will only be accurate if the assumption is true.

Normal Quantile Plots

Although we can't be sure a variable follows a normal distribution, we can construct a graph that helps us verify this assumption.

The graph we use is called a **normal quantile plot**.

The data values are plotted on the y -axis, against the theoretical expected values for a standard normal variable, which are plotted on the x -axis.

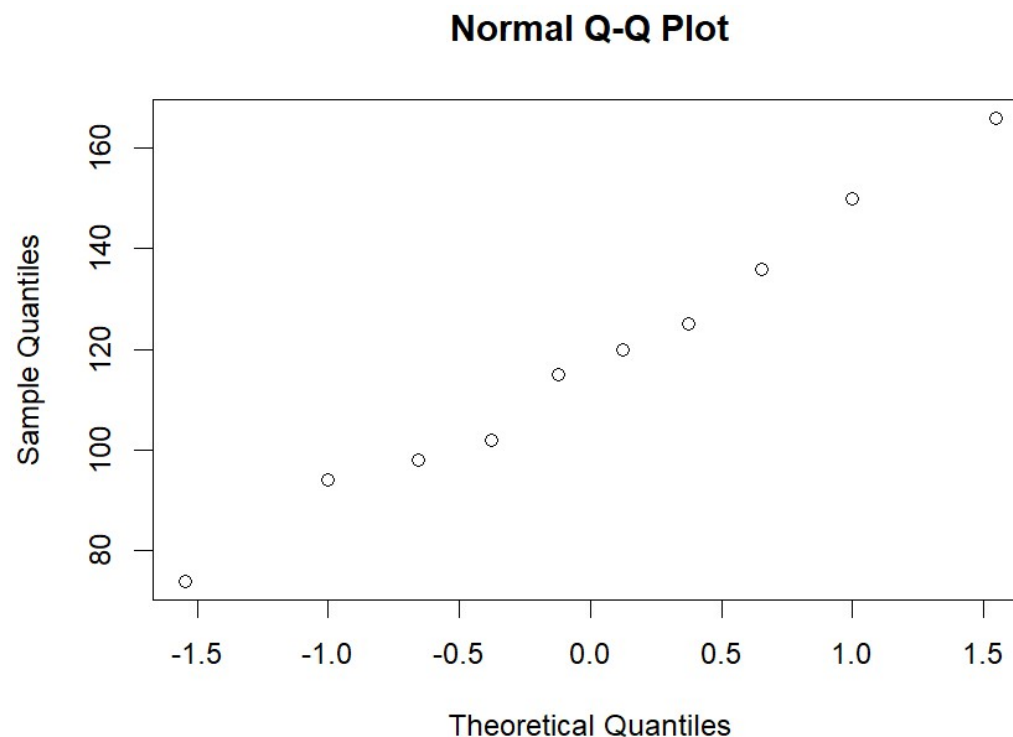
Example

The sentences (in months) of a random sample of ten criminals convicted of a certain crime are shown below:

136	102	84	150	115
98	125	176	120	72

We would like to know if it is a reasonable assumption that sentence times follow a normal distribution.

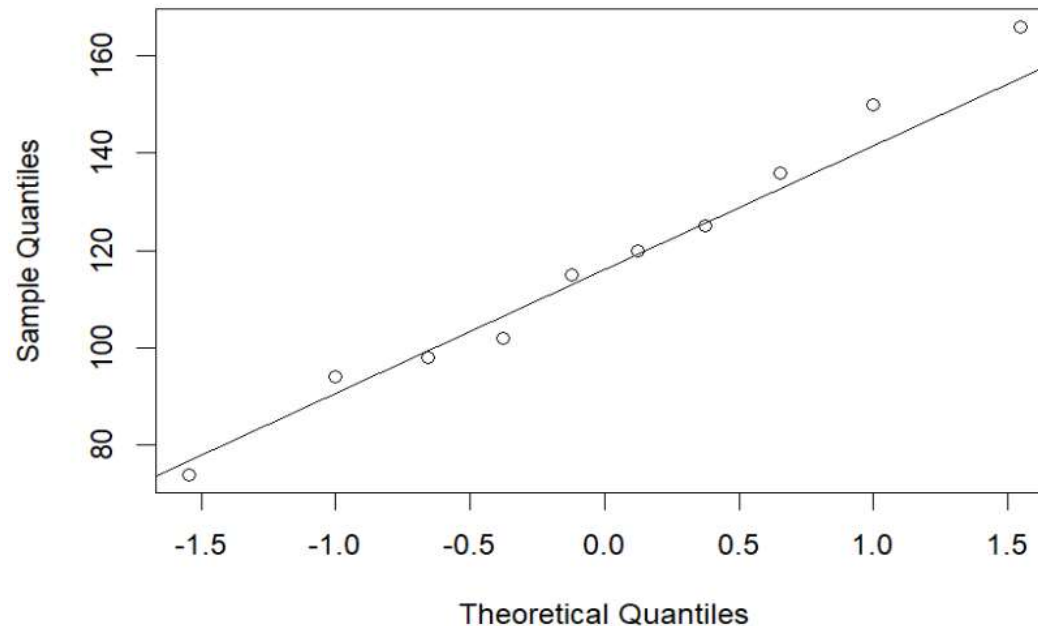
Normal Quantile Plots



```
Sentence <- c(136,102,94,150,115,98,125,  
              166,120,74)  
qqnorm(Sentence)
```

Normal Quantile Plots

If the data come from a normal distribution, we expect the points to fall close to a diagonal line. We see this is the case here. (We added the line to the graph with the command `qqline(Sentence)`.)



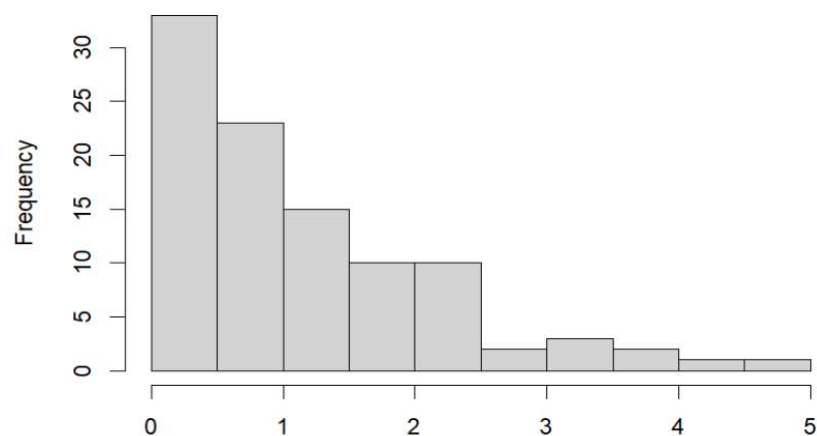
Normal Quantile Plots

The plot therefore validates the assumption of a normal distribution.

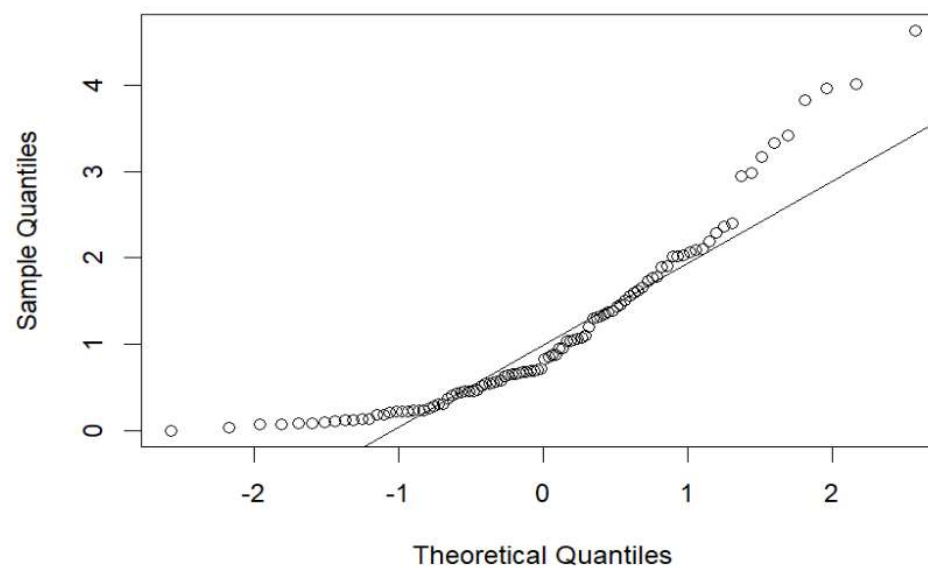
The following normal quantile plots show what we would expect to see on the graph if the distribution was skewed to the right, skewed to the left, or if there was an outlier in the data set.

Normal Quantile Plots

Histogram of X

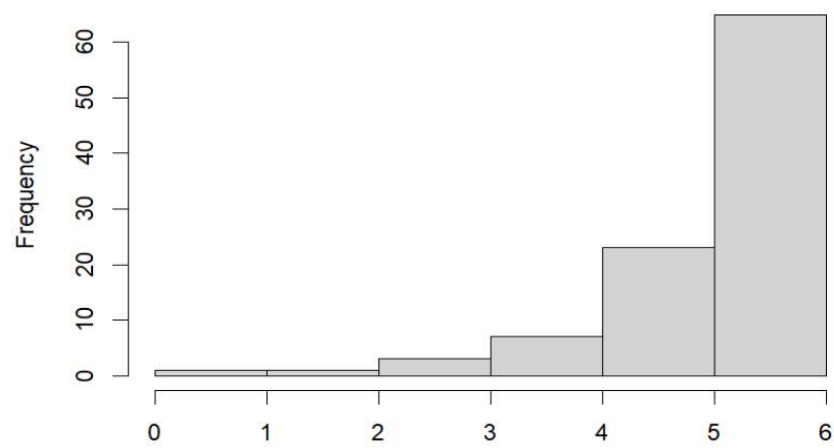


Normal Q-Q Plot

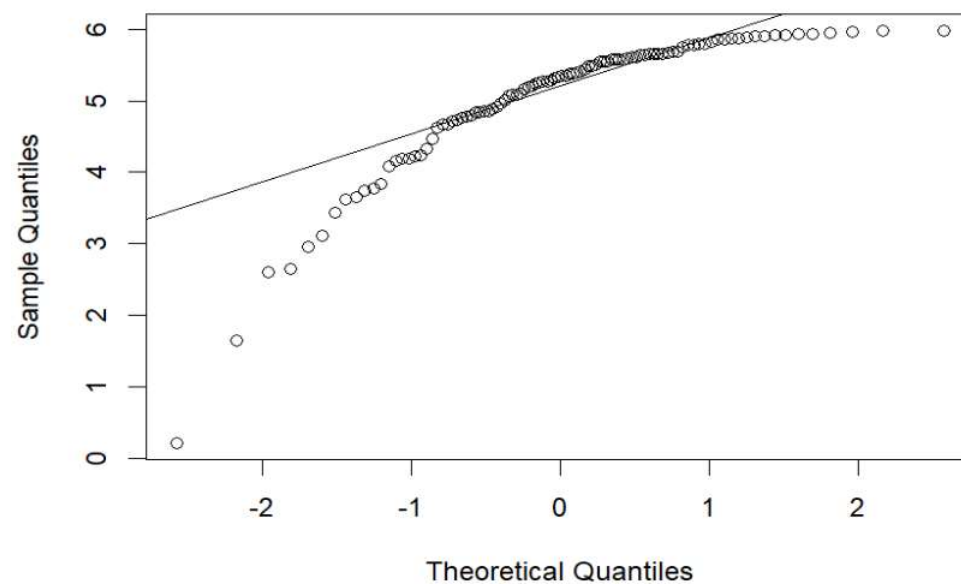


Normal Quantile Plots

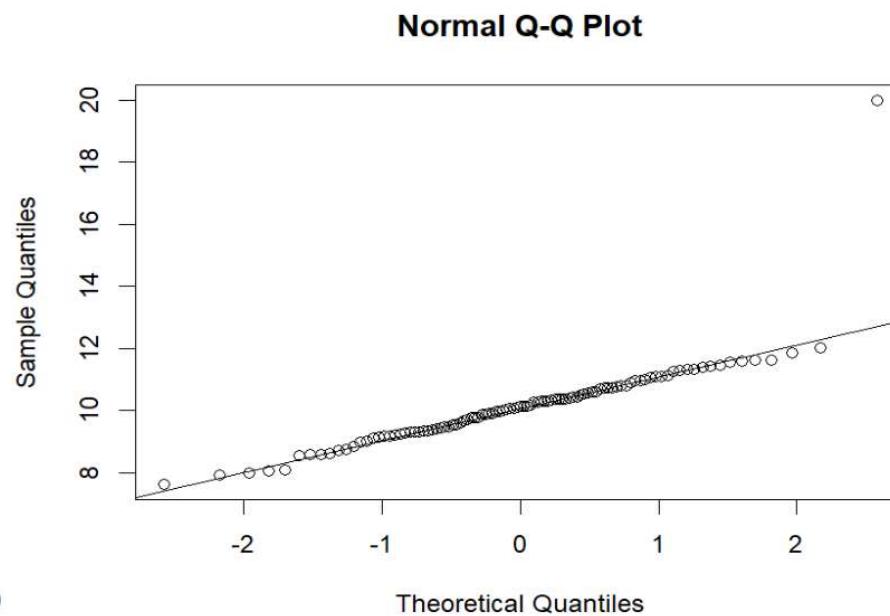
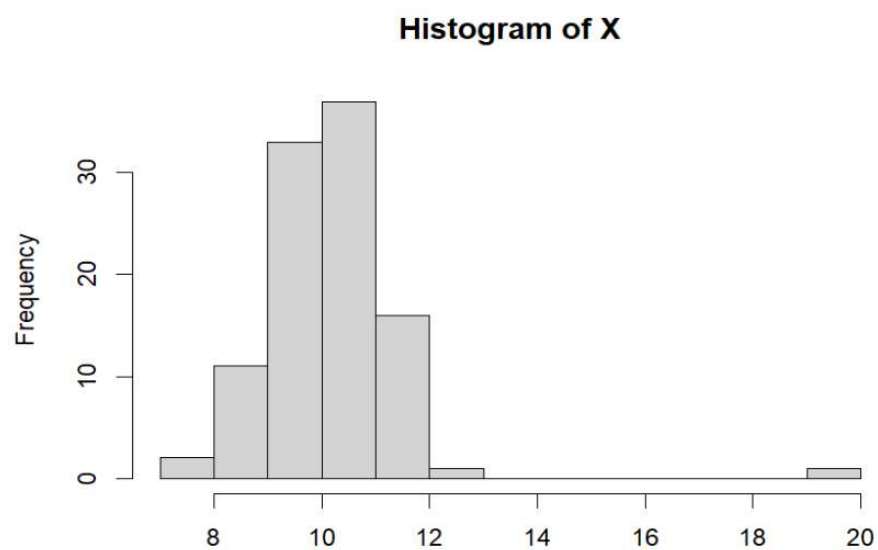
Histogram of X



Normal Q-Q Plot



Normal Quantile Plots



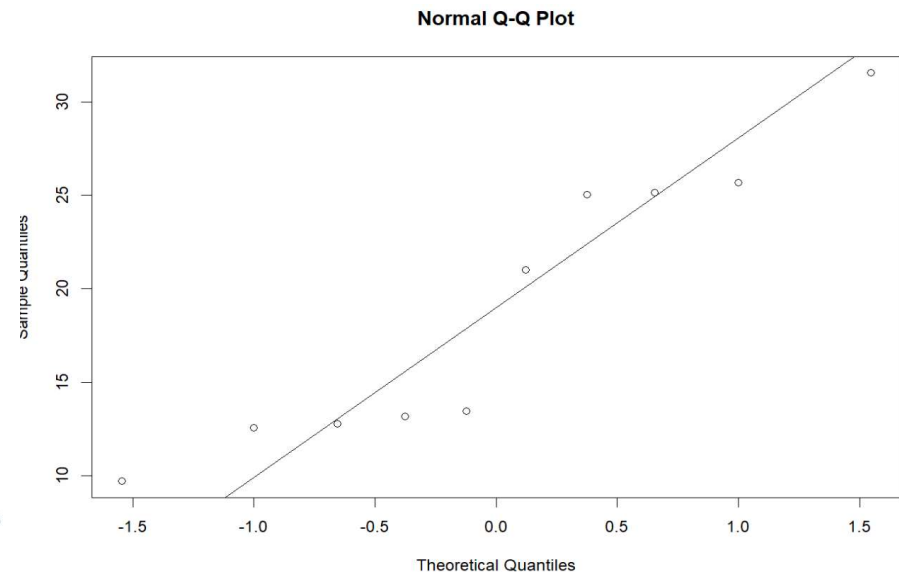
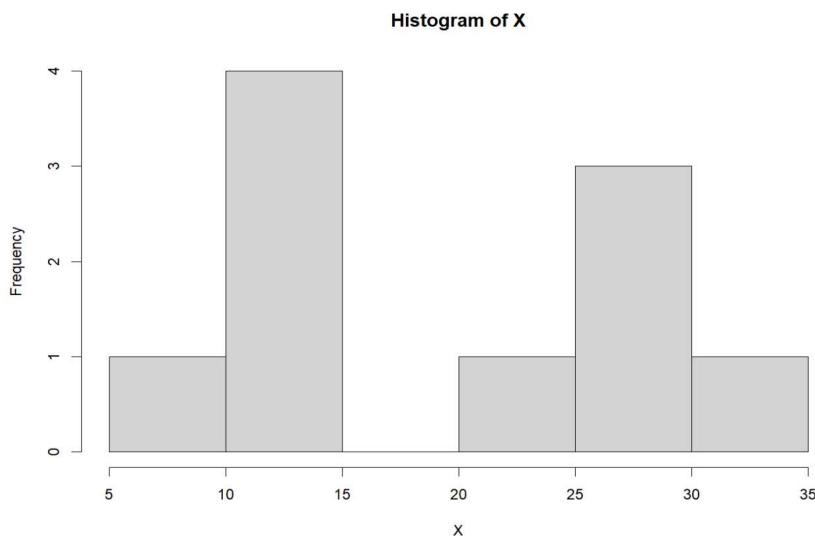
Normal Quantile Plots

Note that if the sample size is small, we can't expect the histogram to look like a bell curve. Similarly, we can't expect the normal quantile plot to show all data values very close to a diagonal line.

In the case of very small samples, we must rely on the assumption of normality.

Normal Quantile Plots

The following sample of 10 observations was randomly generated from a normal distribution:



```
X<-rnorm(10,20,5)  
hist(X)
```

```
qqnorm(X)  
qqline(X)
```