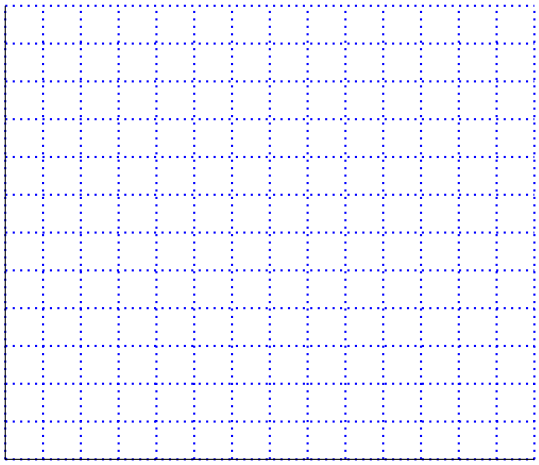
1. Creating Histograms

Definition 1.1.

(1) _____are the entities about which information (data) is collected. Individuals may be people, but they may also be groups, animals, or things.
(2) A _____is a characteristic or trait that can take on different values for different individuals. A particular variable may be either qualitative (e.g., gender) or quantitative (e.g., age).
(3) _____is the process of finding main features of the given data. Begin by looking at each variable and then the relationships between the variables. Graphs and numerical summaries are useful.
(4) _____of a variable gives information (as a table, graph, or formula) about how often the variable takes certain values or intervals of values.
(5) A _____is a graphical representation of a frequency distribution for a single numerical variable. Bars are drawn over each class interval on a number line. The areas of the bars are proportional to the frequencies (or relative frequencies) with which data fall into the class intervals.

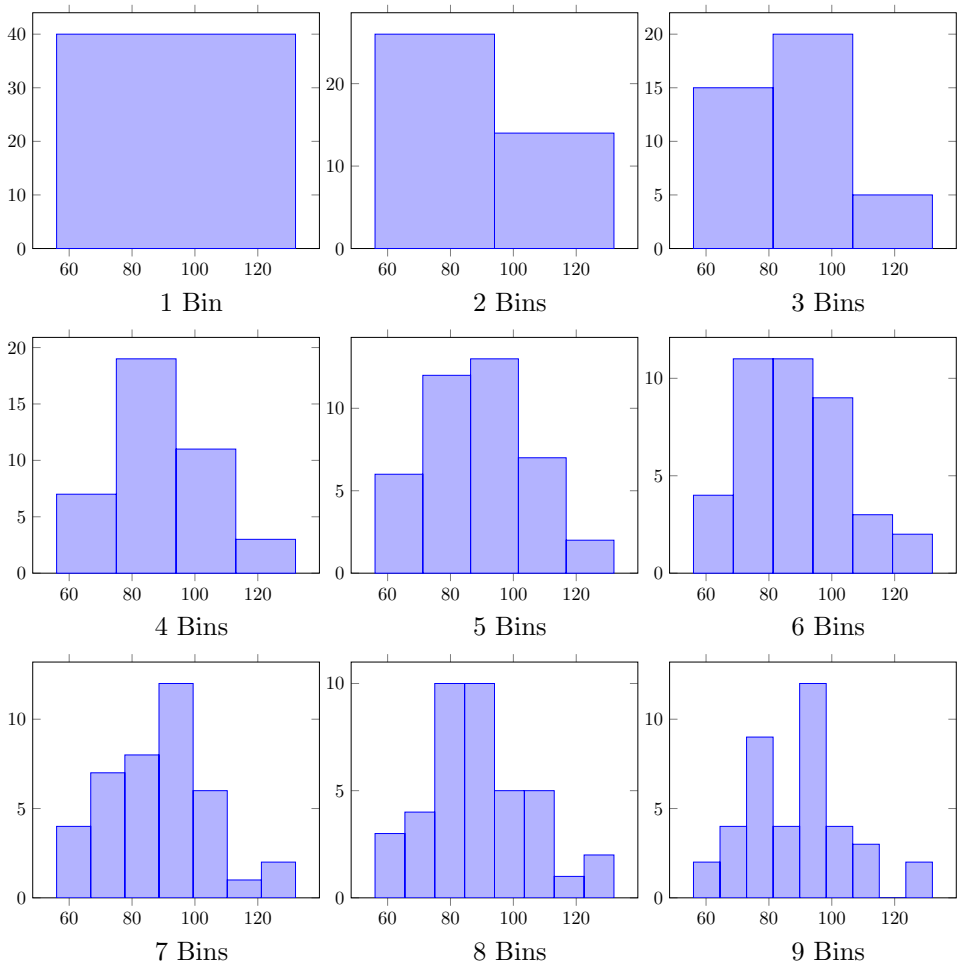
Example 1.2.
Display the date below in a histogram

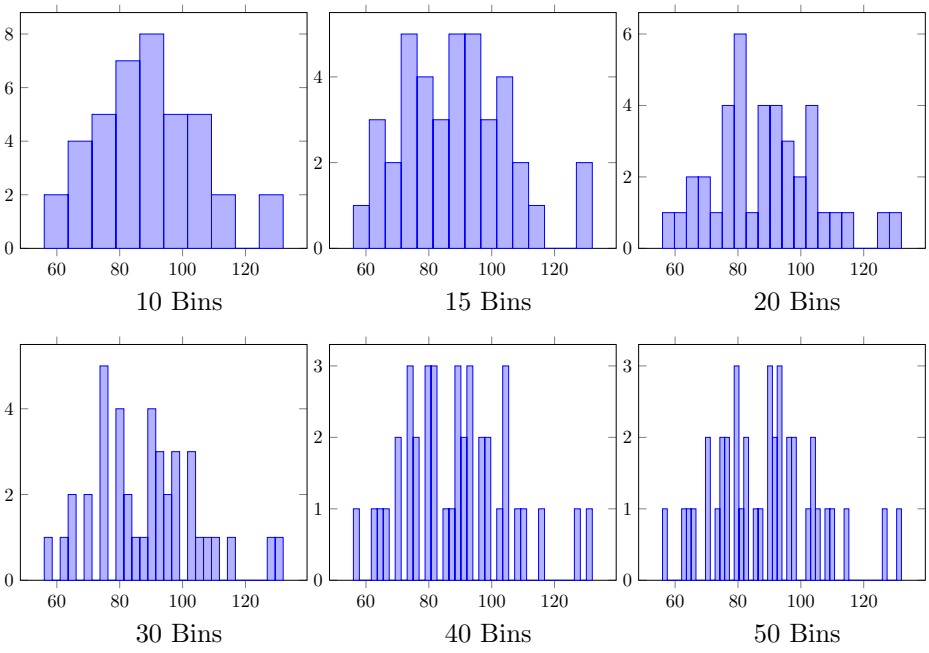
Value	Count
6	8
7	11
8	9
9	5
10	0
11	2



The Histogram is determined by its **Bin Width**.

Example 1.3. 56, 63, 65, 66, 70, 71, 74, 75, 75, 76, 76, 79, 80, 80, 81, 82, 82, 85, 87, 90, 90, 90, 91, 92, 93, 93, 94, 96, 97, 98, 98, 103, 104, 104, 105, 109, 110, 115, 127, 132 are the value in given data. Below are histograms with different bin width.





Algorithm 1.4 (How to draw a Histogram)**.**

(1) Choose the classes. Divide the range of the data into a "reasonable" number of classes of *equal width*.

(2) Count the number of individuals in each class (frequency).

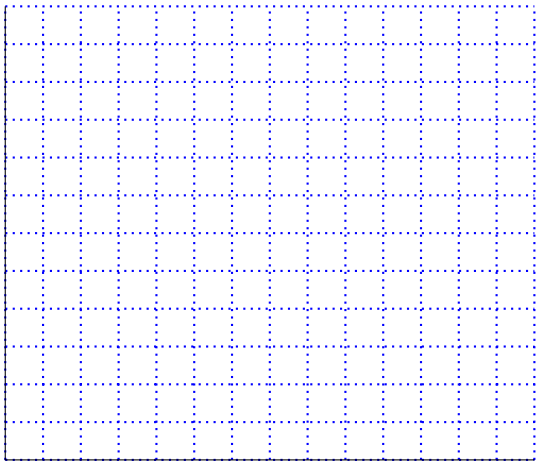
(3) Draw the histogram.

- The vertical axis is the count in each class.
- The horizontal axis represents the classes.

Example 1.5. The following data below is the recorded daily high temperature (in °F) in College Station for March 2006. Display this data in a histogram.

86 86 85 83 83 82 82 81 81 80 79 77 77 77 76 76 75 74 74 73 72 72 72 69 69 69 67 65 61 58 51

Classes (Size:___)	Count



2. Interpreting Histogram

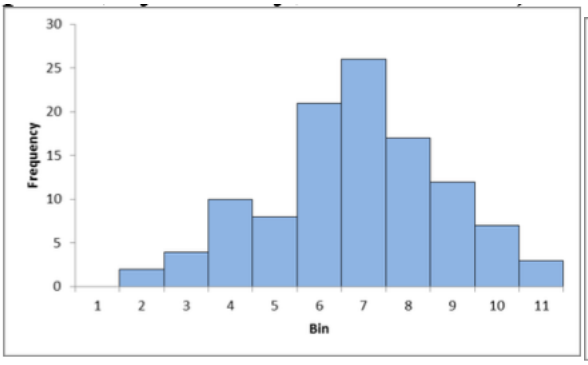
In any graph of data, look for patterns and deviations from the patterns. Ask yourself questions such as:

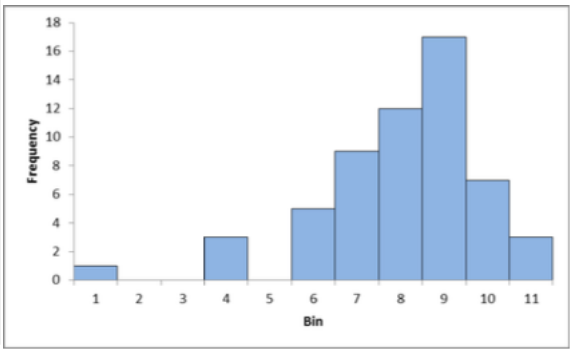
- Does the graph have one or more peaks?
- Is the graph symmetric or skewed?
- Are there outliers?
- Where is the center?
- Is most of the data spread out or close together?

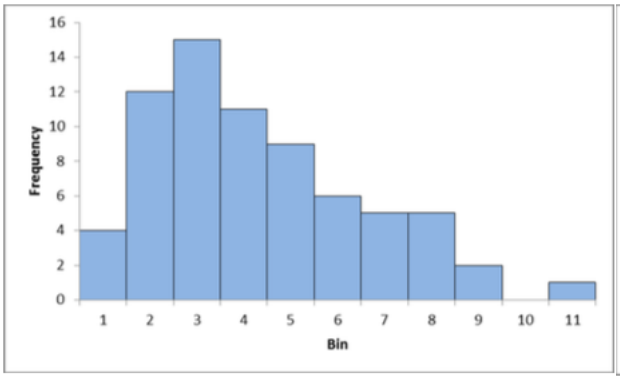
Definition 2.1.

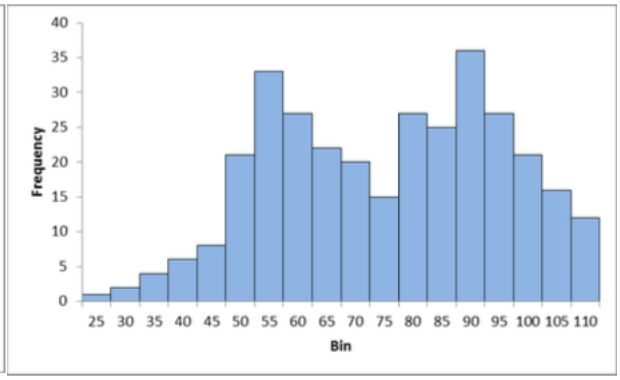
- An _____is an individual value that falls outside the overall pattern such that the value seems striking deviations from the overall pattern.
- A _____distribution is a distribution in which the longer tail of the histogram is on the right side. (Because positive numbers lie on the right side of a number line, such a distribution is also called _____)
- A _____distribution is a distribution in which the longer tail of the histogram is on the left side.(Because negative numbers lie on the left side of a number line, such a distribution is also called "_____.")
- A _____distribution is one in which a vertical line could be superimposed on the histogram and the left and right sides are approximate mirror images of each other.

Example 2.2. Comment on the shapes of the histograms below (including number of peaks, symmetry, and outliers):









3. Creating Stemplot

Definition 3.1.

A _____(or **stem-and-leaf plot**) is a display of the distribution of a variable that attaches the final digits of each observation as a leaf on a stem made up of all but the final digit.

Algorithm 3.2 (How to make stem-plot)**.**

(1) Separate each observation into a stem, which consists of all but the final (rightmost) digit, and a leaf, which is the final digit. Stems may have as many digits as needed, but each leaf contains only a single digit. If this produces too many stems, consider rounding the data to the tens place and using the tens digit as the leaf (or rounding to the hundreds place and using the hundreds digit as the leaf, or something similar).
(2) Write the stems in a vertical column, with the smallest at the top, and draw a vertical line at the right of this column. Include all stems, even if they are not used.
(3) Write each leaf in the row to the right of its stem. Arrange the leaves from smallest to largest.

Example 3.3.

Display the following data in a stemplot:
2 5 7 9 11 15 16 18 18 23 23 25 25 28 29 29 34 35 37 39 40 43 44 45 45 57 70

Example 3.4.

Round the following data to the nearest 10, drop the ending zero and display the result in a stemplot.
118 122 160 161 203 210 216 247 250 266 301 302 304 313 316 321 328 333 334 335 349 393 403 411 605 1111

4. Describing Center: Mean and Median

Definition 4.1.

Given n observations, $x_1, x_2, \cdots, x_n$, _____, $\overline{x}$ of a set of the observations is one of the representation of a center of distribution, calculated as

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

_____, M, of a set of the observations is one of the representation of a center of distribution, calculated as below procedure;

(1) Arrange all observations (including any repeated values) in increasing order (from smallest to largest).
(2) If the number of observations n is odd, the median M is the center observation in the ordered list. If the number of observations is even, the median M is the mean of the two center observations in the ordered list.

_____of a set of observations is the observation that occurs most frequently. You can have no mode, or, if there is a tie, you can have multiple modes.

All of these measurements will have **the same unit** as the data values.

Example 4.2.

The following are scores on an honors exam from a class of 17 students:

$$32, 71, 72, 77, 77, 83, 84, 85, 87, 89, 90, 92, 95, 96, 98, 99, 100$$

What is the "average" score on the honors exam?

Are there any obvious outliers? If so, remove the outlier and recalculate the mean, median, and mode. What does this tell us?

Each of these three measurements is a measure of the "center" of the data. Discuss the helpfulness of each of these measures in understanding the performance of the class on this exam.

5. Describing Spread: The Quartiles

Definition 5.1.

The _____is a measure of variability of a set of observations. It is obtained by subtracting the smallest observation from the largest observation.

$$\text{range} = \text{maximum - minimum}$$

The _____is a set of three observations, Q_1, Q_2, and Q_3 cutting data into four groups.

Max:=the largest observation

Q_3:=is the median of the data above the median$M$.

$Q_2 := M$

Q_3:=is the median of the data below the median$M$.

Min:=the smallest observation

Lastly, the **interquartile range** (or **IQR**) is $Q_3 - Q_1$.

Example 5.2.

The following are scores on an honors exam from a class of 17 students:

$$32, 71, 72, 77, 77, 83, 84, 85, 87, 89, 90, 92, 95, 96, 98, 99, 100$$

Find the following for the given data:

Min	Max	Range	Mean	Median	Mode	Q1	Q2	Q3	IQR

Example 5.3.

The following data is the recorded daily high temperature (in °F) in College Station for March 2006.

$$86, 86, 85, 83, 83, 82, 82, 81, 81, 80, 79, 77, 77, 77, 76, 76, 75, 74, 74, 73, 72, 72, 72, 69, 69, 69, 67, 65, 61, 58, 51$$

Find the following for the given data:

Min	Max	Range	Mean	Median	Mode	Q1	Q2	Q3	IQR

Reminder: The numbers given in these examples were already placed in numerical order. If your data is not given to you in numerical order, you must *first* **put the data in numerical order** before calculating the median and the quartiles.

Example 5.4. If the mean, median and mode of the wealth of students in this class are given, how will these be changed that (if at all) knowing that Bill Gates is one of the secret students contained in the data? What effect will this have on the histogram? Discuss the usefulness of each measure of center.

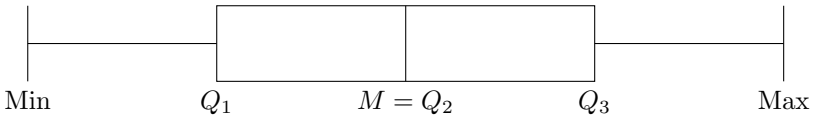
6. The Five-Number Summary and Boxplots

Definition 6.1.

The _____of a distribution consists of the following:

Min	Q1	Q2=Median	Q3	Max

A _____is a graph of the five-number summary.



Example 6.2.

Display the exam scores and high temperatures (from previous examples) in a boxplot.

Scores:

$$32, 71, 72, 77, 77, 83, 84, 85, 87, 89, 90, 92, 95, 96, 98, 99, 100$$

Temps:

$$86, 86, 85, 83, 83, 82, 82, 81, 81, 80, 79, 77, 77, 77, 76, 76, 75, 74, 74, 73, 72, 72, 72, 69, 69, 69, 67, 65, 61, 58, 51$$

Definition 6.3.

One definition of an **outlier** is a data value x such that

$$Q_1 - 1.5 > x \text{ or } Q_3 + 1.5 < x$$

Are their outliers in the exam grades?

Example 6.4.

The CDC reports that the number of new AIDS cases per year from 1990 to 2004 in Iowa are:

$$75, 118, 156, 104, 110, 104, 97, 76, 60, 78, 79, 80, 75, 76, 69.$$

Show this data in a boxplot, with labels. Are there outliers?

Example 6.5.

The CDC reports that the number of new Lyme disease cases per year from 1990 to 2005 in Iowa are:

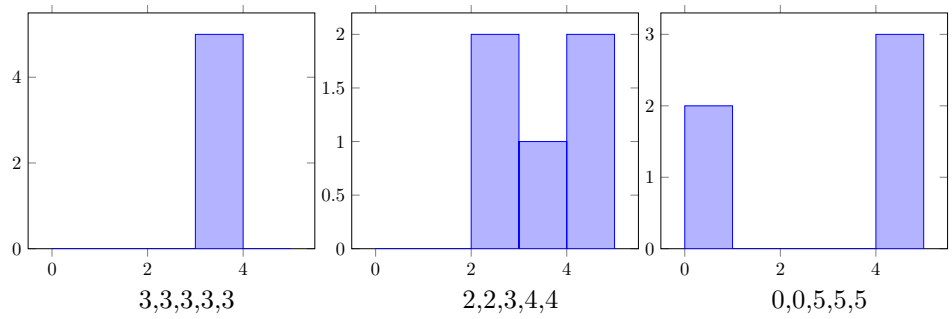
$$16, 22, 33, 8, 17, 16, 19, 8, 27, 24, 34, 54, 65, 72, 49, 88$$

Show this data in a boxplot, with labels. Are there any outliers?

Example 6.6.

Six numbers have min=5, max=19, mode=17 and med=16. Find a possible data set for these results.

7. Describing Spread: The Standard Deviation



For all three of these, the mean (average) is 3. However their distributions are different.

Definition 7.1.

The **standard deviation** s gives us a way of saying on average how far away the actual data values are from the mean. In other words, it is a measure of spread. It is defined as below;

$$s := \sqrt{\frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1}}$$

Another measure of spread is called **variance** Var, which is $Var := s^2$.

Example 7.2.

For the graphs above, which one has highest standard deviation? The lowest?

Example 7.3.

The table to the right gives the average monthly temperature (in °F) for four different months for San Diego. Find the mean, standard deviation, and variance for the temperature.

	Jan	Apr	Jul	Oct
San Diego	65	68	76	75
Chicago	29	59	84	64

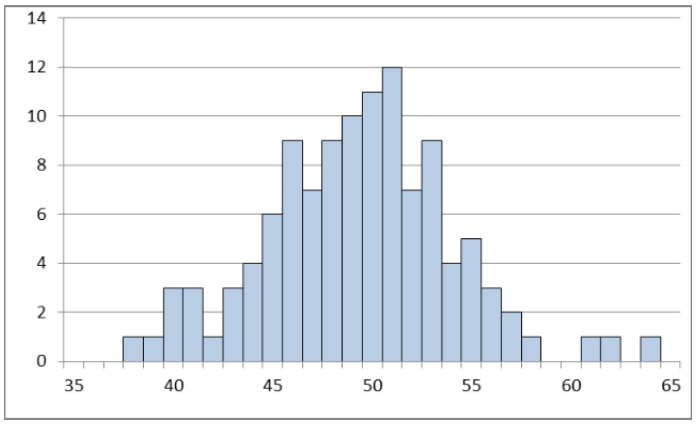
x	$\overline{x}$	$x - \overline{x}$	$(x - \overline{x})^2$
65			
68			
76			
75			
sum	X	X	
divide by $n-1$	X	X	
take $\sqrt{\ }$	X	X	

x	$\overline{x}$	$x - \overline{x}$	$(x - \overline{x})^2$
29			
59			
84			
64			
sum	X	X	
divide by $n-1$	X	X	
take $\sqrt{\ }$	X	X	

Which city's temperature varies more?

8. Normal Distribution

Example 8.1.
One hundred fair coins were flipped and the number of heads was counted. This experiment was repeated 114 times and the results are shown in the histogram below.

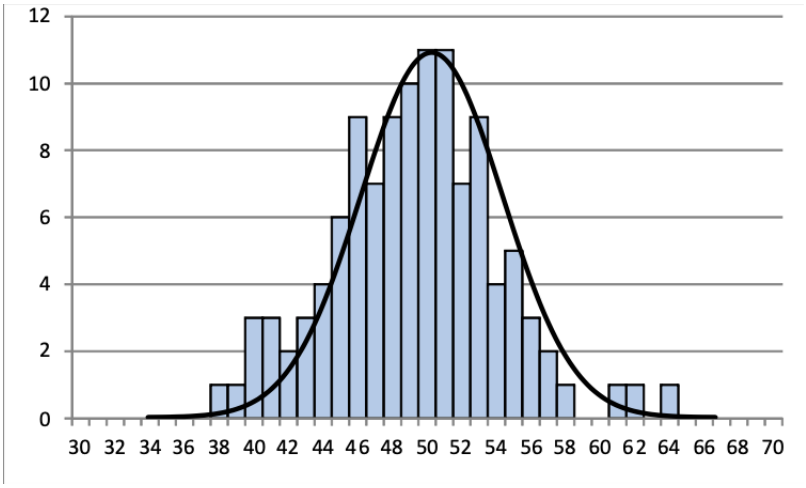


How many times were 45 or fewer heads observed?

What proportion of the time were 45 or fewer heads observed?

What proportion of the time were more than 50 heads observed?

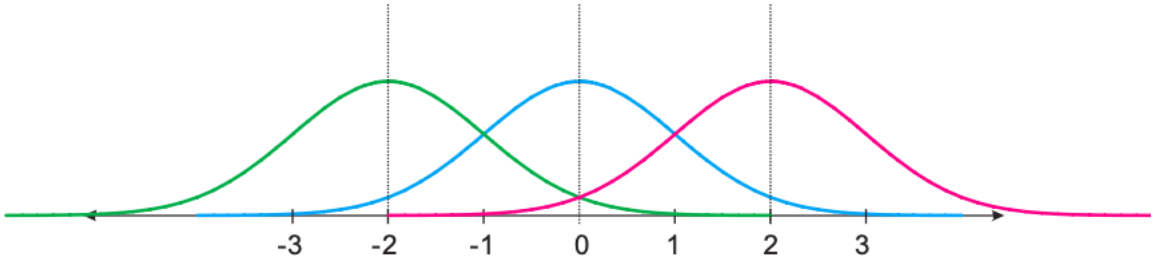
Can we approximate this with a bell-shaped curve? The data has a mean of 50.28 heads and a standard deviation of 5.1 heads. This generates a bell curve with the shape as shown. The reason this is useful is because we will be able to use these bell-shaped curves to find proportions.



Definition 8.2.
Bell-shaped curves like above are called **normal curves** or **normal distributions.**

Some general information about bell curves (also known as normal curves or the normal distribution).

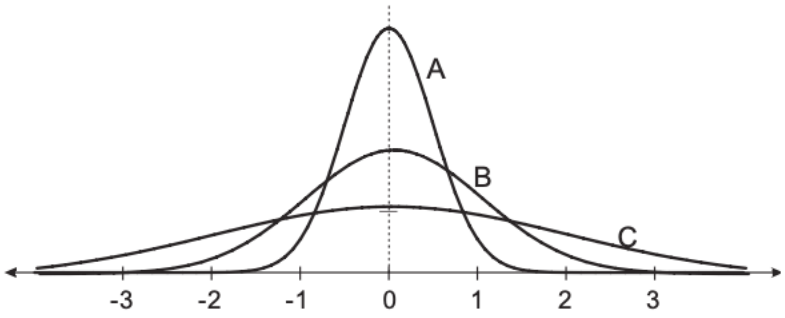
(1) The location of the peak of the curve is where the mean is located. Typically use the symbol μ (mu) for the mean of the distribution.



The normal curves above all have a standard deviation of 1. Which has the highest mean?

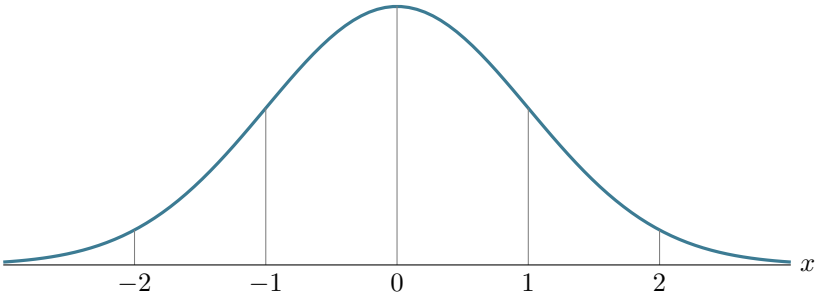
(2) The curve is symmetric about the mean (mirror image around the mean).

(3) The spread is determined by the standard deviation. Typically use the symbol σ (sigma) for the standard deviation of the distribution.



The normal curves above all have a mean of 0. Which has the highest standard deviation? The lowest?

The specific normal curve that has a mean of 0 and a standard deviation of 1 is called the standard normal curve. It is graphed below.



When you are one standard deviation above or below the mean, notice that this is where the curvature changes.

Example 8.3.
A certain type of washing machines has a useful life with a mean of 12 years and a standard deviation of 4 years.

- Draw a normal curve with this mean. Also mark the locations that are one two, and three standard deviations above and below the mean.

- What value is 1.3 standard deviations above the mean?

- What value is 0.8 standard deviations below the mean?

Definition 8.4.
The **standard score** or Z**-score** of a measurement X is how many standard deviations (σ) the measurement is away from the mean (μ).

To calculate the standard score, Z, one can have the formula

$$Z = \frac{X - \mu}{\sigma} \iff X = \sigma Z + \mu.$$

Example 8.5.

(1) In the above example, if a washing machine lasts 17 years, what is its z-score?

(2) How many years would a washing machine last if its Z-score was -1.8 years?

Example 8.6.
A normal distribution has a mean of 50 years and a standard deviation of 8 years.

(1) Find the Z-scores for the following values of X

X	Z
$X = 42$	
$X = 38$	
$X = 60$	

(2) If the Z-score of a measurement is -2.1, what is the value of X?

Theorem 8.7.

In a normal distribution with the mean μ and standard deviation σ,

$$Q_1 = \mu - 0.67\sigma \qquad Q_2 = \mu \qquad Q_3 = \mu + 0.67\sigma.$$

Example 8.8.

Where are the first and third quartiles located on a normal curve with a mean of 50 years and a standard deviation of 8 years?

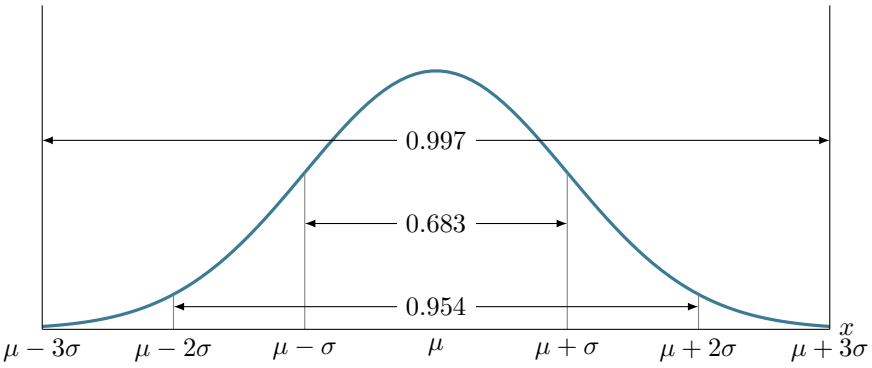
Example 8.9.

For the washing machine example, where the mean was 12 years and the standard deviation was 4 years, what values bracket the middle 50 percent of the data?

9. The 68-95 -99.7 Rule

Theorem 9.1. *In any normal distribution,*

- *About 68 percent of the data is within 1 standard deviation of the mean.*
- *About 95 percent of the data is within 2 standard deviation of the mean.*
- *About 99.7 percent of the data is within 3 standard deviation of the mean.*



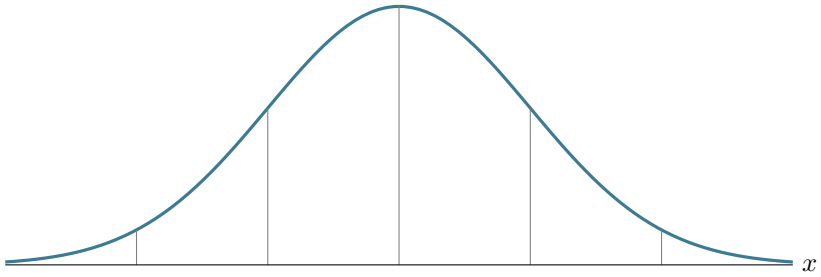
These percentages can also be seen as probabilities. If 68 percent of the data lies within 1 standard deviation of the mean, we can also say that the probability an observation lies within 1 standard deviation of the mean is 68 percent.

Example 9.2.

For the washing machine example, where the mean was 12 years and the standard deviation was 4 years, what range of values make up the middle 95 percent of washing machines?

Example 9.3.

The length of tape on a roll of a certain type of masking tape is normally distributed with a mean of 25 meters and a standard deviation of 50 centimeters.



(1) What is the range of lengths of most (99.7 percent) of the rolls?

(2) What lengths bracket the middle 68 percent of rolls of tape?

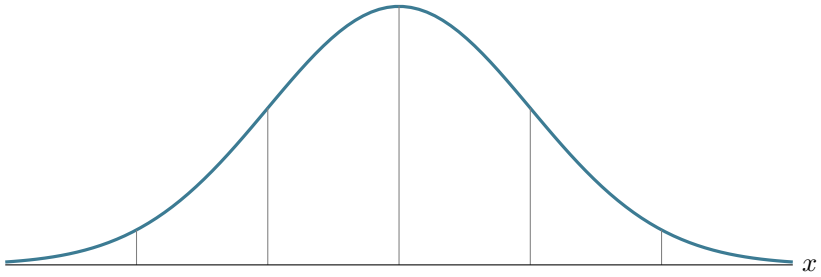
(3) What percent of the rolls are longer than 26 meters?

(4) What percent of the rolls are between 25 and 26 meters?

(5) What is the probability as a percent that a roll of tape is less than 24.5 meters?

Example 9.4.

The amount of time students spend studying on a Sunday is normally distributed with a mean of 145 minutes and a standard deviation of 23 minutes.



(1) What is the range of study times that make up the middle 95 percent of students?

(2) What percent of students study more than 214 minutes?

(3) What percent of students study less than 168 minutes?

(4) What percent of students study between 99 and 122 minutes?

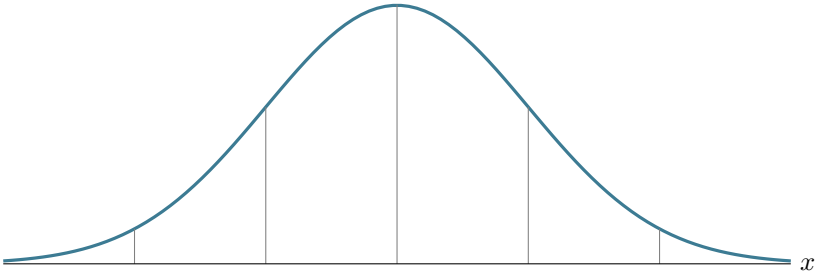
(5) What percent of students study between 122 and 191 minutes?

(6) What study times make up the middle 50 percent of students?

Example 9.5. A class of 460 students will be graded based on the normal curve with

- A grade of "A" assigned to students who are more than two standard deviations above the mean.
- A grade of "B" assigned to students who are between 1 and 2 standard deviations above the mean.
- A grade of "C" assigned to the students within one standard deviation of the mean.
- A grade of "D" assigned to students between 1 and 2 standard deviations below the mean.
- A grade of "F" assigned to students more than two standard deviations below the mean.

If the mean grade in the class is 74, with a standard deviation of 11, what are the grade cut-offs?



How many students receive each grade?