

CHAPTER 5 – EXPLORING DATA DISTRIBUTIONS

A sample of bags of M&M candies is chosen and the number of blue M&Ms in each bag is counted. The results are shown in the table below:

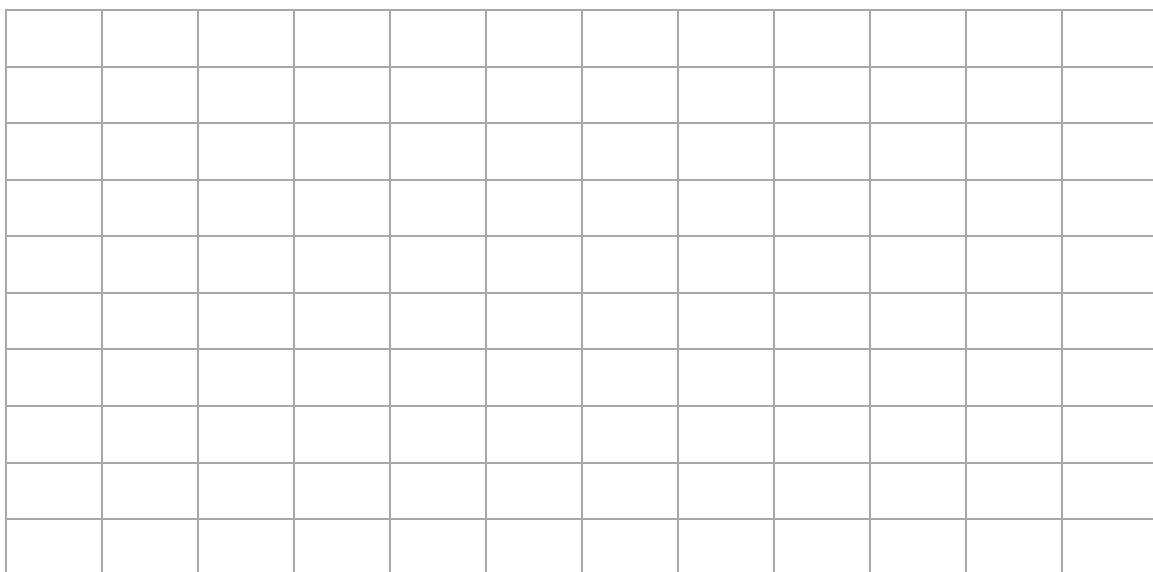
No. of blue M&Ms	7	8	9	10	11
No. of bags	3	4	5	8	2

Individuals are the objects described by a set of data. These individuals may be people, animals or things.

A ***variable*** is any characteristic of an individual. A variable can take on different values for different individuals. Some variables are numeric and others are not.

What are the individuals and what is the variable in the experiment above?

Display this information in a ***histogram***:



Exploratory data analysis is the process of looking at data to describe the main features. Begin by looking at each variable and then the relationships between the variables. Graphs and numerical summaries are useful.

The **distribution** of a variable tells us what values the variable takes and how often it takes these values.

A **histogram** is a graph of the distribution of outcomes for a single numerical variable. The height of each bar is the number of observations in the class of outcomes covered by the base of the bar. All classes should have the same width and each observation must fall into exactly one class.

Steps to creating 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.

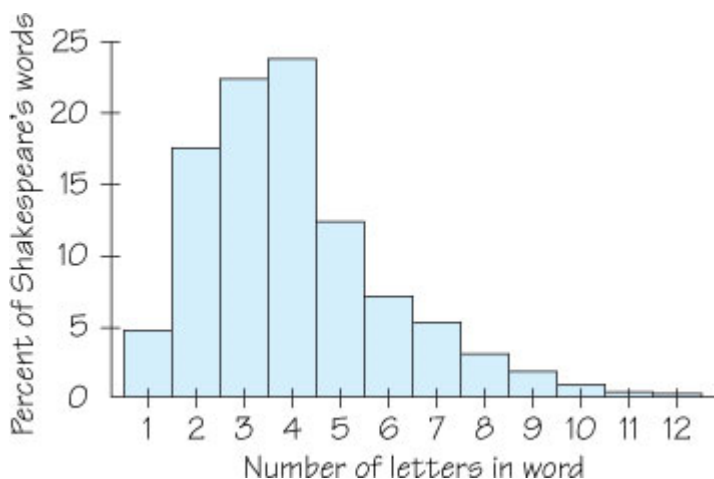
If you have data that ranges from a low value of 14 to a high value of 36, what will the classes be if the class size is 4?

A graph is *symmetric* if the right and left sides of the histogram are approximately mirror images of each other.

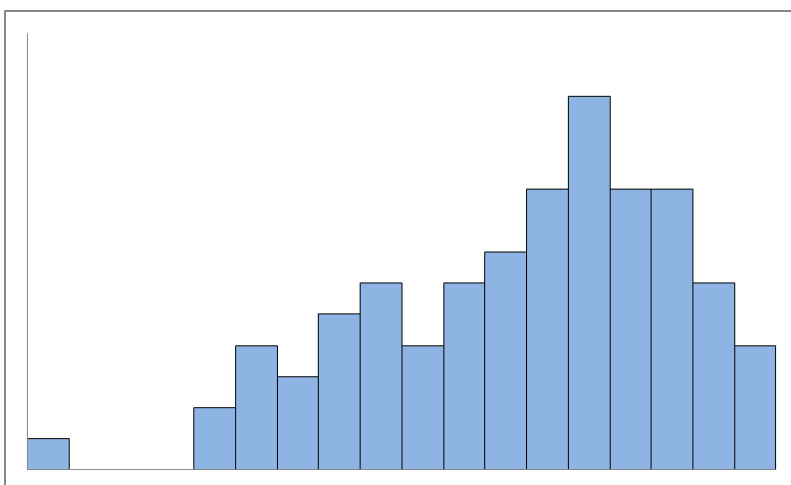
A graph is *skewed to the right* if the longer tail is on the right side. This is also called positively skewed

A graph is *skewed to the left* if the longer tail is on the left side. This is also called negatively skewed.

An *outlier* is an individual data value that falls outside the overall pattern.



The histogram shows the lengths of words used in Shakespeare's plays as a percent of all the words in his plays. What is the overall shape of this distribution?



Comment on the shape of the histogram on the left

A **stemplot** is a display of the distribution of a variable that attaches the final digits of the observations as leaves on stems made up of all but the final digit.

To Make A Stemplot

1. Separate each observation into a **stem** (consisting of all but the rightmost digit) and a **leaf** (the rightmost digit).
2. Write the stems in a vertical column with the smallest at the top. Include all the stem values from smallest to largest, even if some are not used. Draw a vertical line at the right of this column.
3. Write each leaf in the row to the right of its stem in increasing order

Arrange the data into a stemplot 26, 13, 8, 56, 11, 10, 11, 12, 24, 5, 14

The **mean** of the n numbers x_1, x_2, \dots, x_n is $\bar{x} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n}$

This is also the “balance” point for our dot plots and histograms.

The **median** of the n numbers x_1, x_2, \dots, x_n is the number in the middle when the n number are arranged in order of size and there are an odd number of values. When there are an even number of values, the median is the mean of the two middle numbers.

Half of our data is above the median and half is below.

The **mode** of the n numbers x_1, x_2, \dots, x_n is the number that occurs the most often. If no number occurs more often than any other number, there is no mode. If two numbers both occur the most often, then there are two modes.

For the previous data (26, 13, 8, 56, 11, 10, 11, 12, 24, 5, 14), where would the data “balance”? Where is the data cut in half? Which value occurred the most often?

What is the mean, median and mode for the number of blue jelly beans that we graphed earlier? Note these points on the histogram.

No. of blue M&Ms	7	8	9	10	11
No. of bags	3	4	5	8	2

The *range* is a measure of spread of a set of observations. It is obtained by subtracting the smallest observation from the largest.

To calculate the *quartiles* Q_1 and Q_3 ,

1. Use the median to split the data set into two halves – an upper half and a lower half.
2. The *first quartile* Q_1 is the median of the lower half.
3. The *third quartile* Q_3 is the median of the upper half.

The *interquartile range* (or *IQR*) is $Q_3 - Q_1$

The *five-number summary* of a distribution consists of the following:

Minimum Q_1 M Q_3 Maximum

A boxplot is a graph of the five-number summary,

An *outlier* is a data value that is less than $Q_1 - 1.5(\text{IQR})$ or is greater than $Q_3 + 1.5(\text{IQR})$.

Find the range, the quartiles, the IQR and outliers (if any) for the data below. Display the data in a boxplot.

26, 13, 8, 56, 11, 10, 11, 12, 24, 5, 14

There are 4 different cases when finding the quartiles:

7 (or 11 or 15 or ...) data points:

1 2 3 4 5 6 7

8 (or 12 or 16 or ...) data points:

1 2 3 4 5 6 7 8

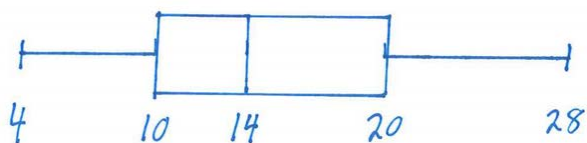
9 (or 13 or 17 or ...) data points:

1 2 3 4 5 6 7 8 9

10 (or 14 or 18 or ...) data points:

1 2 3 4 5 6 7 8 9 10

The boxplot below is for the number of grams of sugar in in a serving of 9 different cereals. If there are two modes, 12 and 20, what is a possible data set?



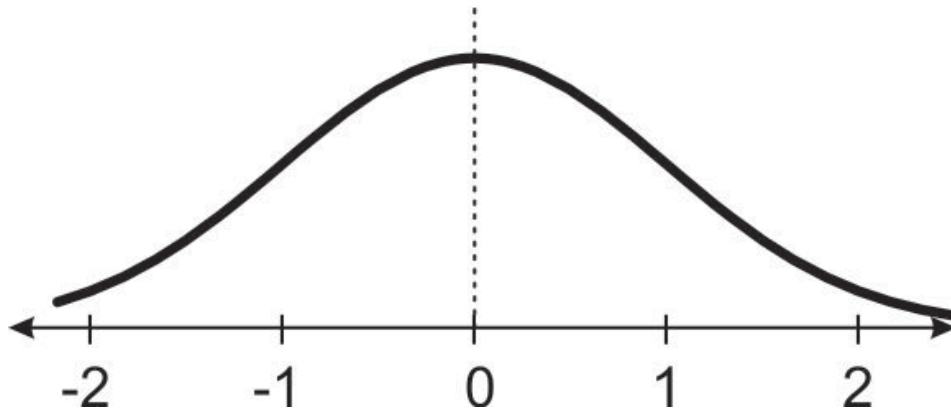
Another way to measure spread is the standard deviation, S . This will determine, on “average”, how far away are the measurements from the mean. To find the standard deviation in a set of n numbers,

- Find the mean of the numbers, \bar{x} .
- Subtract the mean from each of the values, $x - \bar{x}$. Note that the sum of these values must be zero.
- Square the result of the subtractions, $(x - \bar{x})^2$
- Add up the squares, $\sum(x - \bar{x})^2$
- Divide by $n - 1$ and then take the square root, $S = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$

Find the standard deviation in the golf scores for Amy, Bart, and Caleb. Whose scores vary the least?

Amy	70	74	69		
Bart	74	76	84	70	
Caleb	84	83	85	87	86

Many natural and social phenomena produce a continuous distribution with a bell-shaped curve.



Every bell-shaped (NORMAL) curve has the following properties:

- Its peak occurs directly above the mean
- The curve is symmetric about a vertical line through the mean. The curve never touches the x -axis. It extends indefinitely in both directions.
- The area between the curve and the x -axis is always 100% as that is the total probability that something is observed.
- The shape of the curve is completely determined by the mean and the standard deviation.
- The place where the curve changes from concave down [☹] to concave up [☺] is one standard deviation away from the mean.
- The first and third quartiles are located 0.67 standard deviations away from the mean.
- The probability that a data value will fall between two values a and b is given by the area under the curve between a and b .
- The standard normal curve has a mean of 0, standard deviation of 1 and uses the letter Z for the variable.

To find the z -score (how many standard deviations from the mean) for a value X , use the formula

$$Z = \frac{X - \mu}{\sigma}$$

If you are given the z-score, the X value is $X = \mu + \sigma Z$

The weight of an egg is normally distributed with a mean weight of 60 grams with a standard deviation of 5 grams.

(a) Sketch the normal curve for this distribution: _____

(b) Determine the z-score for the following weights:

70 grams

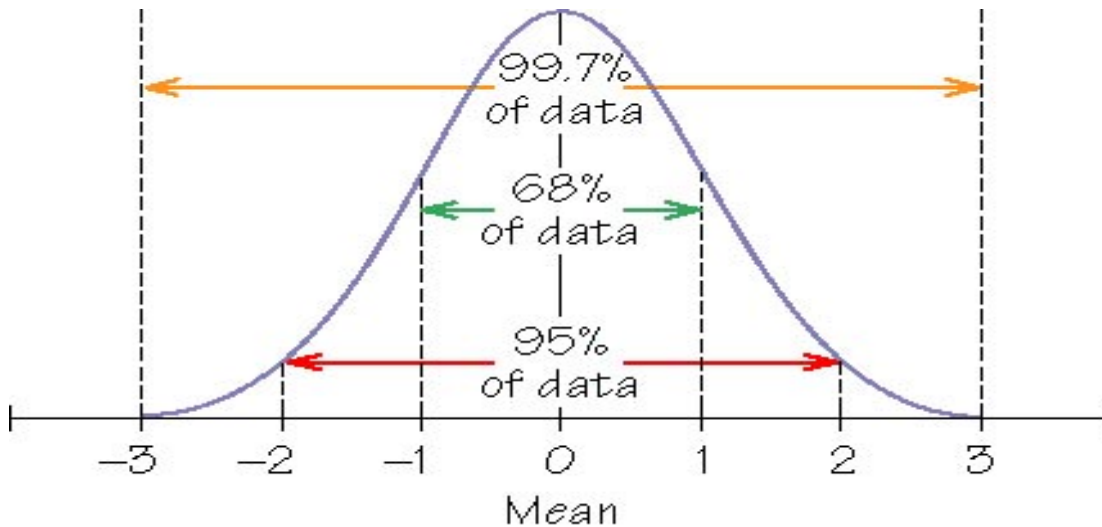
59 grams

63 grams

(c) Where are the first and third quartiles located?

In any normal distribution,

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



The time to run a long race is 48 minutes with a standard deviation of 4 minutes.

- What is the probability that a racer takes longer than 60 minutes?
- What is the probability that a racer takes less than 44 minutes?
- What is the probability that a racer takes between 52 and 56 minutes?
- If there are 5000 runners, how many will finish between 40 and 52 minutes?

SAMPLE EXAM QUESTIONS FROM CHAPTER 5

1. A bag of chocolate chip cookies is opened and the number of chocolate chips in each cookie is counted. The results are

19 22 23 24 24 26 28 29 32 32 35 35 40 42 42 42 45 45 45 46 49 52 55 56 59

(a) Find the following quantities:

Mean = _____

Mode = _____

Median = _____

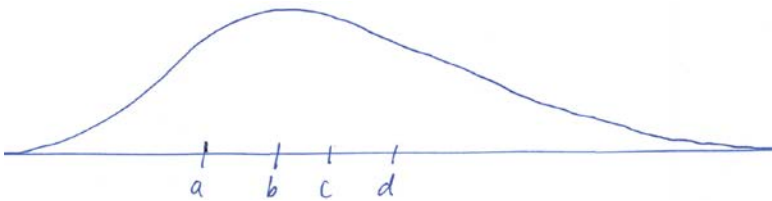
Range = _____

Q_1 = _____

Q_3 = _____

Boxplot

(b) Display the information in a histogram with class size of 6



2a. The distribution above is

- (A) skewed right (B) symmetric (C) skewed left

2b. If c is the median of this distribution, which value is the mean?

- (A) a (B) b (C) c (D) d (E) More information is needed

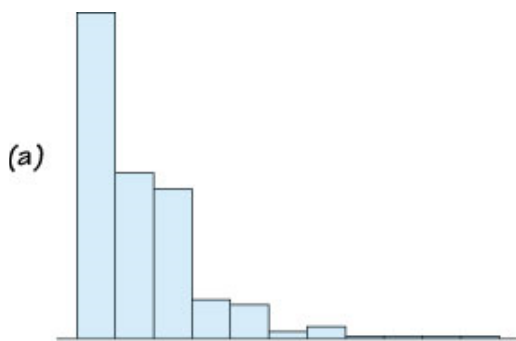
3. Match the distributions to the survey questions

(i) Are you female (0) or male (1)?

(ii) Are you right (0) or left (1) handed?

(iii) What is the length of your arm (to the nearest inch)?

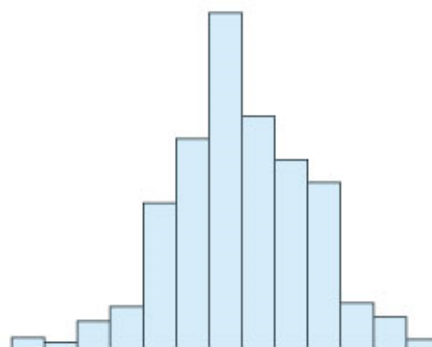
(iv) How many minutes do you study on Saturday evening?



(c)



(d)



4. The data below is the number of winners in a radio contest each month recently. Arrange this in a stemplot.

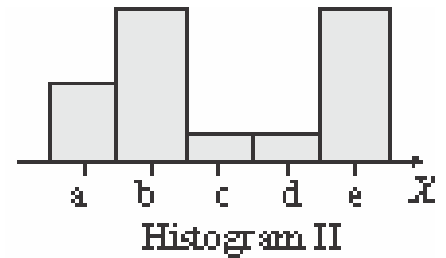
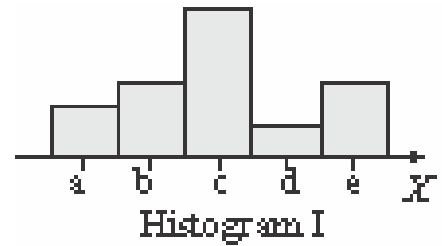
92 105 120 115 98 115 114 116 97 142

5. The number of parking spaces available (checking each hour on the hour) in a parking garage is given below. Find the mean and standard deviation in the number of parking spaces available.

5 4 7 10 4

6. Which histogram has the smaller standard deviation?

- (A) Histogram I
- (B) Histogram II
- (C) They are nearly the same
- (D) There is not enough information to decide



7. A data distribution has a minimum value of 6, $Q_1 = 32$, $M = 41$, $Q_3 = 48$, and the maximum value is 68. Which statements below (if any) are true about this distribution?

- (A) There are no outliers
- (B) There is at least one low outlier
- (C) There is at least one high outlier
- (D) There is not enough information to determine if there are outliers.
- (E) None of these

8. The length of map pencils in a box of used pencils is normally distributed with a mean of 6 inches and a standard deviation of 0.5 inches.

(a) Sketch and label the distribution

(b) What is the probability that a map pencil is shorter than 7 inches.

(c) If the bag has 50 map pencils in it, how many are between 5 and 6 inches long?

9. A normal curve has a mean of 90 and a standard deviation of 10. Where is the 1st quartile located?

- (A) need more information to determine. (B) 100
(C) 86.7 (D) 83.3 (E) 80.0 (F) None of these