Chapter 11

Statistics

Section 11.1
Measures of Central Tendencies and Spread

Understanding data is one of the objectives of *statistics*, which is the collection, organization, and interpretation of numerical data. Measuring the central tendencies and spread of data helps in the interpretation of the numbers collected.

MEAN

Popularly known as "average", the mean of a set of data is found by *adding* and then *dividing*. It is computed by using the equation

$$MEAN = \frac{S}{n}$$

Where S = the sum of all numbers
n = the number of observations

Example: Roberto's exam grades for a math class are as follows: 86, 73, 55, 96, 100, 75, 73, 66, 80, and 80. Find the mean.

First, organize data from lowest to highest

55, 66, 73, 73, 75, 80, 80, 86, 96, 100 Number of observations (exams): 10

Add all the numbers: 55 + 66 + 73 + 73 + 75 + 80 + 80 + 86 + 96 + 100 = 784

$$MEAN = \frac{S}{n} = \frac{784}{10} = 78.4$$

Example: In the above example, what grade should Roberto earn on the final exam to increase the mean of his scores to 80?

The sum of his first 10 exams: 784
The sum of 10 exams plus final: 784 + x
Where x is the desired grade for the final
New mean: 80
New number of exams, n: 11

$$80 = \frac{784 + x}{11}$$
$$80(11) = 784 + x$$
$$880 = 784 + x$$
$$880 - 784 = x$$
$$x = 96$$

Roberto must get at least **96** on the final to reach a mean of **80.**

MEDIAN

After organizing the data from lowest to highest, the value in the middle of the data is the *median*. Officially it is: *The middle value in a distribution, above and below which lie an equal number of values.*

Example: Find the median of the following observations: 30, 45, 22, 65, 45, 76, 3, 15, 88, 65, 44

First, organize data from lowest to highest: 3, 15, 22, 30, 44, 45, 45, 65, 65, 76, 88

5 are lower 5 are higher

Number of observations: 11

Because it is an odd number of observations, the median must be the 6th number (5 above and 5 below the median).

Answer: 45

Example: When the number of observations is even, the median is found by averaging the two closest numbers to the center. Find the median of:

86, 73, 55, 96, 100, 75, 73, 66, 80, and 80

First, organize data from lowest to highest:

55, 66, 73, 73, 75, 80, 80, 86, 96, 100 Number of observations: 10

The two closest numbers to the center are 75 and 80. Add both and divide by 2.

$$\frac{75 + 80}{2} = 77.5$$

MODE

The *mode* is: *The value occurring most frequently in a series of observations*.

Example: Find the mode in the following data: 4.5, 6.5, 7.2, 8.1, 6.5, 9.2, 6.6, 7.2, 5.5, 9.1, 7.2, 4.8

First, organize data from lowest to highest:

4.5, 4.8, 5.5, 6.5, 6.5, 6.6, 7.2, 7.2, 7.2, 8.1, 9.1, 9.2 Number of observations: 12

The value occurring most often is 7.2 (three times).

A set of data could have multiple modes (**bimodal**, **trimodal**...) Multiple modes have a negative effect on central tendencies.

CENTRAL TENDENCIES

A set of data is said to have central tendencies if the value of the mean, median, and mode are the same or nearly the same.

Example: Is the data in the example above displaying central tendencies?

4.5, 4.8, 5.5, 6.5, 6.5, 6.6, 7.2, 7.2, 7.2, 8.1, 9.1, 9.2

Mean = $\frac{82.4}{12}$ = 6.9 Median = $\frac{6.6 + 7.2}{2}$ = 6.9 Mode = 7.2

Answer: The mean and median show exact central tendencies, and the mode is off by only +0.3. The data shows central tendencies.

Example: Does the data show central tendencies?

86, 73, 55, 96, 100, 75, 73, 66, 80, and 80

First, organize data from lowest to highest:

55, 66, 73, 73, 75, 80, 80, 86, 96, 100 Number of observations: 10

Mean = $\frac{784}{10}$ = 78.4 Median = $\frac{75 + 80}{2}$ = 77.5 Two Modes: 73, 80

Answer: The mean and median show central tendencies for the data, but the mode does not.

SPREAD

One way to measure data spread is by computing the **range**, and the range is found by subtracting the lowest value from the highest value. The spread illustrates how far the data reaches.

Range = Highest value – Lowest value

Examples: Find the range for the following sets of data:

55, 66, 73, 73, 75, 80, 80, 86, 96, 100 Range = 100 – 55 = 45

3, 15, 22, 30, 44, 45, 45, 65, 65, 76, 88 Range = 88 – 3 = 85

4.5, 4.8, 5.5, 6.5, 6.5, 6.6, 7.2, 7.2, 7.2, 8.1, 9.1, 9.2 Range = 9.2 – 4.5 = 4.7

Practice:

Section 11.2
Introduction to Probability

Probability is a ratio that shows the likelihood of an event taking place. It is the number of successful outcomes compared to the number of possible outcomes that could actually happen.

Because the probability of an event taking place for sure (100%) is 1 (for example, what is the probability of the sun rising tomorrow?), then anything less than 100% probability must be less than 1. In other words, a probability is a decimal, a percent, or a fraction that tells the chance of an event happening in the future.

Example: In a football game a referee tosses a coin to decide which team will kick first. What is the percent probability of your team kicking first?

Two teams, two faces to a coin.

Number of successful outcomes: 1

Number of possible events: 2 Probability = $p = \frac{1}{2} = 0.5 = 50\,\%$

Probability has a very wide use in industry as an applied and research tool. In the field of quality control, for instance, it is used to predict the number of failures.

SIMPLE PROBABILITY

The probability of an event which is always going to occur is 1; the probability of an event which never happens is 0. Furthermore, if the probability of an event is

p, then the probability of the same event not taking place is $(1 - p)$

Example: In a jar Paula has 10 red, 9 white, 3 blue, and 14 yellow jelly beans. What is the fraction probability of Paula picking a white jelly bean from the jar?

36 jelly beans in total (10 + 9 + 3 + 14).

Number of successful outcomes (white jelly beans): 9

Number of possible outcomes: 36

$$p = \frac{number\ \ of\ \ white}{total\ \ number\ \ of\ \ jelly\ \ beans} = \frac{9}{36} = \frac{1}{4}$$

Decimal probability $\frac{1}{4} = 0.25$ Percent probability = $0.25 \times 100 = 25\%$

COMPOUND PROBABILITY

A compound probability is when a probability is taken further. In other words, it is when more than one event is involved.

In the example above the probability could be compounded in two different ways:

1. By including the compound probability of another color.
2. By repeating the same color twice.

Number 1 situation above is a case of either one color OR the other. Because it is an increase, add. Number 2 is a case of one color AND then another. Because it decreases, multiply.

Example: In a jar Raul has 10 red, 9 white, 3 blue, and 14 yellow jelly beans. What is the percent probability of Raul picking a white OR a blue jelly bean from the jar?

36 jelly beans in total (10 + 9 + 3 + 14).

Number of successful outcomes (white and blue jelly beans): 9 + 3 = 12

Number of possible outcomes: 36

$$p = \frac{white + blue}{total} = \frac{9+3}{36} = \frac{12}{36} = \frac{1}{3}$$

Decimal probability $\frac{1}{3} = 0.33$ Percent probability = 0.33 × 100 = 33%

Example: In a jar Brittany has 10 red, 9 white, 3 blue, and 14 yellow jelly beans. What is the percent probability of Brittany picking a white first AND then a blue jelly bean from the jar *without replacing* the white one?

Without replacement

36 jelly beans in total (10 + 9 + 3 + 14).

Number of successful outcomes (white jelly beans): 9 $p = \frac{9}{36} = \frac{1}{4}$

Number of successful outcomes (blue jelly beans): 3

Because one white jelly bean has been eaten, the total number now is: 35 $p = \frac{3}{35}$

Multiply probabilities for white and blue jelly beans: $\frac{1}{4} \times \frac{3}{35} = \frac{3}{140} = 0.0214 \times 100 = 2.14\%$

With replacement

Suppose the white jelly bean was not to Brittany's liking and she puts the white jelly bean back inside the jar. Because the total number of beans would remain at 36, the probability—and answer—for the blue jelly bean will change.

Replacing white jelly bean $p = \frac{3}{36} = \frac{1}{12}$

Multiply new probabilities for white and blue jelly beans: $\frac{1}{4} \times \frac{1}{12} = \frac{1}{48} = 0.0208 \times 100 = 2.08\%$

Example: A group of 4 boys and 3 girls rush to the school cafeteria to have lunch. What is the percent probability that the first and last in line would be girls?

Number of successful outcomes (first in line a girl): 3 $p = \frac{3}{7}$

Number of possible events: 7

Because one girl is already at the head of the line, subtract one girl from the group.

Number of successful outcomes (last in line a girl): 2 $p = \frac{2}{6} = \frac{1}{3}$

Number of possible events: 6

Multiply probabilities for first *and* last a girl: $\frac{3}{7} \times \frac{1}{3} = \frac{3}{21} = 0.1428 \times 100 = 14.28\%$

Practice:

CONTENTS OF THIS PAGE HAVE BEEN REMOVED TO PROTECT COPYRIGHT

Section 11.3
The Counting Principle, Permutations, and Combinations

THE COUNTING PRINCIPLE

When more than one event is taking place, the counting principle is used to determine the possible outcomes. The counting principle states that all the possible choices should be multiplied.

Example: A student has 7 skirts and 6 blouses in her closet that she may combine to produce one outfit. How many ways can she combine all the skirts and all the blouses?

The short answer would be: $7 \times 6 = 42$

In the long answer consider that the first skirt will match with 6 blouses, and the second skirt will do the same, and the third the same, and so on. Seven skirts matching 6 times each is 42.

Example: A printer has 16 colors and needs to select two of the colors for the cover of a book, one for the text and one for the background. In how many different ways can the cover be printed?

When the printer selects one color for background, there are 15 colors left for the text. If this is done for all 16 colors:

Each of the 16 colors, with 15 other selections: $16 \times 15 = 240$ possibilities

PERMUTATIONS

A permutation is when a set is rearranged—in other words, an event in which one thing is substituted for another.

Example: How many ways can the letters in the name CARLOS be arranged?

Trying it manually by moving letters around would take a long time; however, the following equation can be used to solve it

$$P = n! \qquad \text{Where } n! \text{ is n-factorial}$$

A factorial is:

The product of all the positive integers from a given number to 1

Because CARLOS has 6 letters: $P = 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 720$ ways

COMBINATIONS

If order in selection is important, permutations is the way to solve a problem, but if order is NOT important, then the equation for combinations can be used:

$$C = \frac{n(n-1)(n-2)(n-3)\ldots}{k!}$$

Where ***n*** is the total number in the group taken "***k***" at a time. This way ***n*** is multiplied by the reduced "***k***" value in the numerator, then divided by "***k***" factorial. See example below.

By order it is meant that selecting, for example, three students for a task (k), the order could be Moe, Pat, and Jerry or Jerry, Moe, and Pat.

Example: In a group of 12 students, the teacher wants to make groups of three for a project. How many different groups of three could she make?

The question here is how many combinations are there in 12 items taken 3 at a time.

Reduced by "***k***" (3) is represented as 12, 11, 10

$$C = \frac{12(12-1)(12-2)}{3!} = \frac{12 \times 11 \times 10}{3 \times 2 \times 1} = \frac{1320}{6} = 220$$

Example: In a group of 25 students, how many different groups of 6 could be made for a project?

$$C = \frac{25 \times 24 \times 23 \times 22 \times 21 \times 20}{6!} = \frac{127512000}{6 \times 5 \times 4 \times 3 \times 2 \times 1} = 177{,}100$$

Practice:

CONTENTS OF THIS PAGE HAVE BEEN REMOVED TO PROTECT COPYRIGHT

Section 11.4
Tools for Exploring Data

GRAPHS

Someone once said “a picture is worth a thousand words.” If the statement is true, then it is probably the reason why in statistics so many graphs are used.

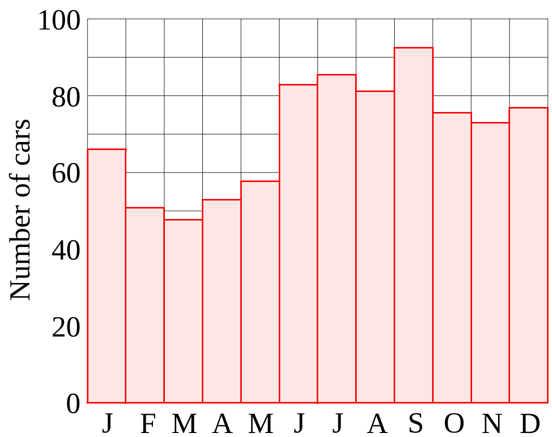
A graph is a way to summarize data into an organized demonstration that could be read quickly.

Bar Graphs and Histograms

As the name indicates, a bar graph uses bars. Bar graphs represent data based on frequency (how often.) In the case of the bar graph, bars show a specific category. The graph below shows the category “number of cars” sold by a dealership during 2006. To build a bar graph, first decide the size of the vertical scale. Because from the table the highest number of cars sold per month is close to 100, that is the limit set for the graph. The division of the scale is set at 10 for neatness. The horizontal scale was set per month because that is the way sales commissions are computed.

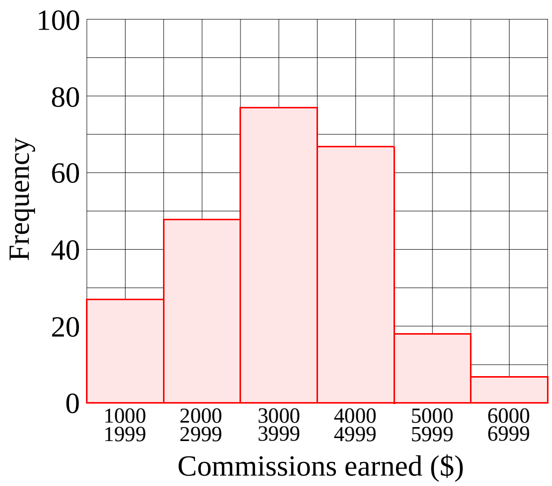
Month	number of cars
Jan.	67
Feb.	51
Mar.	48
Apr.	53
May	58
June	83
July	85
Aug.	80
Sept.	93
Oct.	76
Nov.	73
Dec.	77

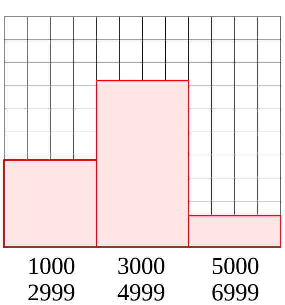
SALES BY MONTH



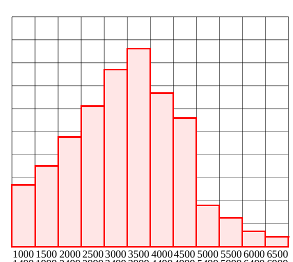
Unlike in the bar graph, in a **histogram**, each bar represents a **range** of data (see graphs below). The intervals selected are the actual commissions earned by salespersons. Notice that selecting the size of the interval is important. If the intervals are too few or too many, the graph loses meaning.

MONTHLY EARNINGS PER SALESPERSON





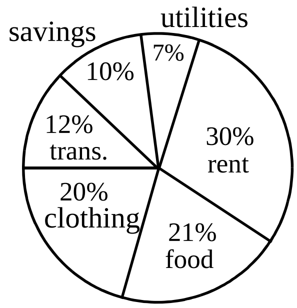
Too few intervals

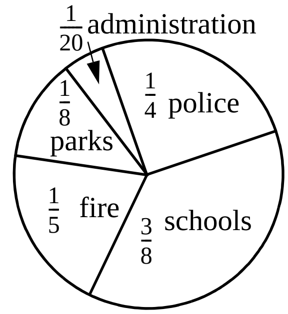


Too many intervals

Circle Graphs

Circle graphs are also called “pie charts.” Although circle graphs have limitations and cannot be expanded like a bar graph, they are an excellent way to present fractions, portions, or percentage as a “piece of a pie.”





The first circle graph to the right shows a home budget. Notice that all the percents add up to 100%. To build a circle graph you must recall that a circle has 360º; therefore, the piece of the pie devoted to, for example, 7%, has to be proportional to the other pieces of the pie where 100% = 360º:

$$\frac{7}{100} = \frac{angle}{360} \qquad angle = \frac{7 \times 360}{100} = 25.2^{o} \qquad 25.2^{o} \text{ is the angle for } 7\%$$

The second circle graph above shows how a city spends taxes. Similarly, forming a circle graph with fractions equates a fix amount, in this case one (1), with 360º. For example, the angle of the piece of the “pie” devoted to parks will be:

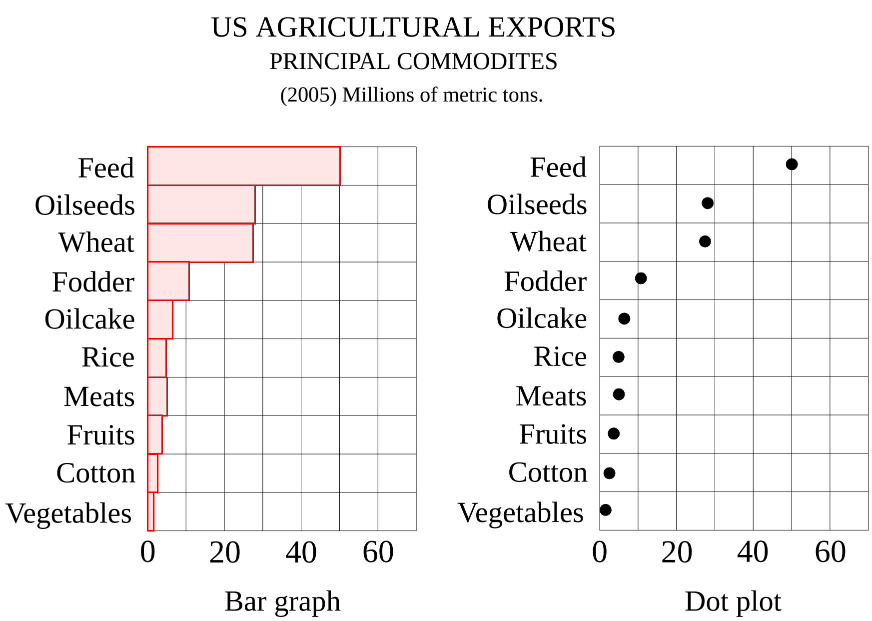
$$\frac{1/8}{1} = \frac{angle}{360} \qquad angle = \frac{\frac{1}{8} \times 360}{1} = \frac{\frac{1 \times 360}{8}}{1} = \frac{\frac{360}{8}}{1} = 45^{o}$$

PLOTS

Plots are built from points that define a location. The location of points, in turn, may or may not develop into recognizable patterns that define the direction a set of data is taking.

Dot plots may be used as an alternative to bar charts, and many researchers prefer them because they are cleaner and easier to read. For example, the dot plot below shows the same data as the bar graph, yet it is simpler to produce and easy to read.

commodity	millions of metric tons
Feed	50
Oilseeds	28
Wheat	27
Fodder	11
Oil cake	6
Rice	5
Meats	5
Fruits	4
Cotton	3
Vegetables	2



Box plots, also called "box and whiskers" plots, besides establishing location, contain data on the median, upper quartile (UQ), lower quartile (LQ), interquartile range (IQR), upper limit (UL), lower limit (LL), and outliers (far-off points).

To build a box plot:

1. Organize the data from lowest to highest.
2. Find the median (number where 50% of all data is above and 50% below, see page 180.)
3. Find the lower quartile (number where 75% of all data is above and 25% below)
4. Find the upper quartile (number where 25% of all data is above and 75% below)

Example: Build a box plot for the data: 86, 73, 55, 96, 100, 75, 73, 66, 80, 40, 120, 86, 78, 92, 67, 82. Find the interquartile range, upper and lower limits, and outliers.

Organize: 40, 55, 66, 67, **LQ** 73, 73, 75, 78, **median** 80, 82, 86, 86, **UQ** 92, 96, 100, 120

Median: 79 Lower quartile: $\frac{67+73}{2} = 70$ Upper quartile: $\frac{86+92}{2} = 89$

Because there are 16 numbers (even amount), the median must be after the 8th number (half the numbers), but before the 9th number, in the middle of 78 and 80, or 79.

The lower quartile must be in the middle of the first 8 numbers (after the 4th, but before the 5th). Add the 4th and 5th number, divide by 2, and the LQ = 70.

The upper quartile must be in the middle of the last 8 numbers (after the 12th, but before the 13th). Add the 12th and 13th number, divide by 2, and the UQ = 89.

The interquartile range is the distance between UQ and LQ: IQR = UQ – LQ = 19

Outliers are those numbers that fall beyond 1.5 the distance of the IQR, or $IQR \times 1.5 = 19 \times 1.5 = 28.5$

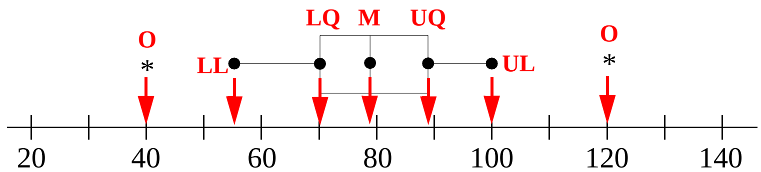
Add 28.5 to the UQ 28.5 + 89 = **117.5** and Subtract 28.5 from the LQ 70 – 28.5 = **41.5**

Any numbers below **41.5** or above **117.5** are not worth considering because they are **outliers**.

Therefore, both 40 and 120 are outliers and the UL falls to 100 and the LL moves up to 55.

Building the box:

On a scale that includes all numbers observed, draw a box with the LQ and UQ as the sides of the box, and draw a line dividing the box at the median (M). Add lines (whiskers) to the LL and UL. Mark the outliers (O) with an asterisk or a cross.



Stem-and-Leaf plots are frequency plots that are used to organize data and determine central tendency. The idea is to split every number into two sections: the left part is the **stem**, the other the **leaf**. For example, the number 123 would be split 12|3, and the number 48 4|8. A column of the stem is matched by a string representing the leaf.

Example: Make a stem-and-leaf plot of the following airplane take-off waiting times, in minutes.

5, 6, 3, 1, 16, 22, 34, 21, 1, 7, 55, 12, 33, 8,
4, 55, 34, 17, 25, 44, 48, 25, 22, 7, 14, 31, 40

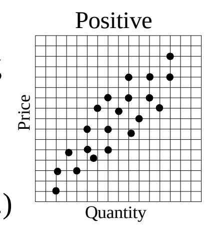
STEM	LEAF
0	1, 1, 3, 4, 5, 6, 7, 7, 8
1	2, 4, 6, 7
2	1, 2, 2, 5, 5
3	1, 3, 4, 4
4	0, 4, 8
5	5, 5

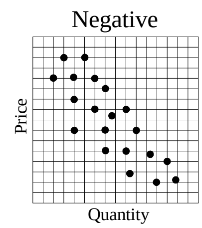
The stem will be the ten digits and the leaf the one digits. The central tendencies may be checked by looking at the distribution. In this particular case the data is “skewed” (leaning) towards the lower values, while the median is 21 minutes and there are 6 modes: Poor central tendencies.

Notice that a stem-and-leaf plot looks like a horizontal histogram: They both have intervals and frequency.

A benefit of the stem-and-leaf plot is that it displays all of the individual values within that interval, while the histogram does not.

Scatter plots are used to summarize how a set of data behaves when compared to another set. Called correlation, it is a way of checking patterns of plotted points using two variables. The more the points tend to bunch up to form a distinctive group that shows direction (like a flock of birds), to the verge that they look like going somewhere, the stronger the correlation (the variables follow each other.) There are three ways of reading scatter plots:

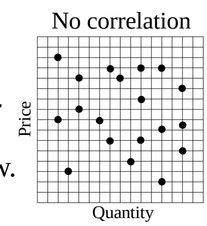




1. If they aim towards the northeast (lower left to upper right), the correlation is positive.

2. If they aim towards the southeast (upper left to lower right), the correlation is negative.

3. If the points are random without showing any particular aim, the correlation is very low.



Practice: