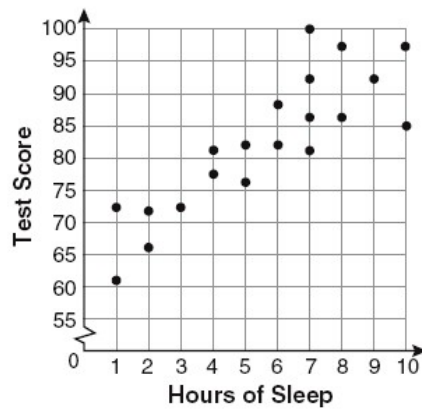
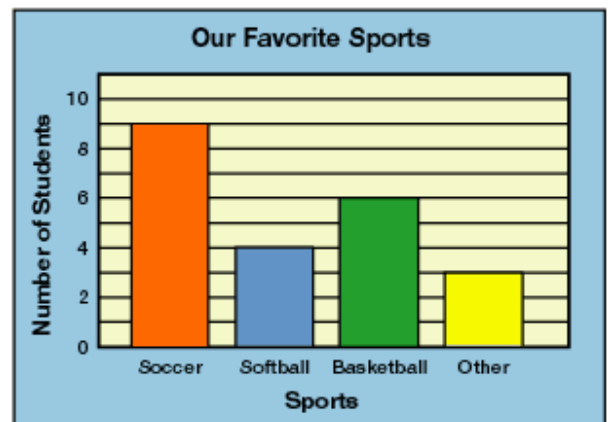
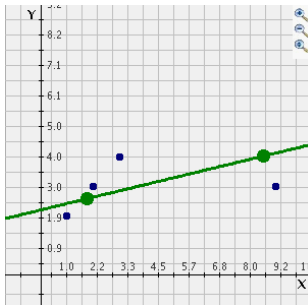


1-Variable Statistics & 2-Variable Statistics



Lesson: Organizing and Graphing Data (One-variable statistics)

Raw Data: the unprocessed information collected for a study

Example from Textbook Page 92:

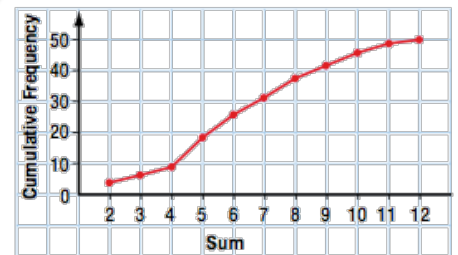
Here are the sums of the two numbers from 50 rolls of a pair of standard dice.

11	4	4	10	8	7	6	6	5	10	7	9	8	8
4	7	9	11	12	10	3	7	6	9	5	8	6	8
2	6	7	5	11	2	5	5	6	6	5	2	10	9
6	5	5	5	3	9	8	2						

Frequencies:

Sum	Tally	Frequency	Cumulative Frequency
2	IIII	4	4
3	II	2	4+2=6
4	III	3	6+3=9
5	IIII	9	9+9=18
6	III	8	18+8=26
7		5	26+5=31
8	I	6	31+6=37
9		5	37+5=42
10	IIII	4	42+4=46
11	III	3	46+3=49
12	I	1	49+1=50

Always the same as the number of rolls!!



Cumulative Frequency Graph will always be **INCREASING** (↑)

Relative Frequency: shows the frequency of a data group as a fraction or percent of the whole data set

Example from Textbook Page 97: Marks on data management test:

$$\text{interval size} = \frac{\text{possible range of data}}{\# \text{ of intervals}}$$

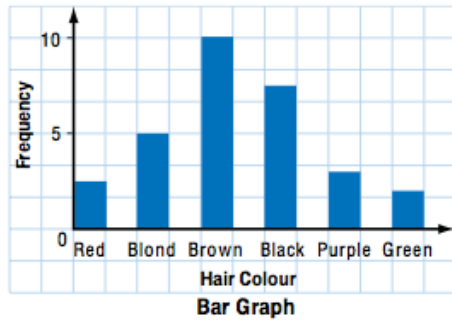
$$\text{e.x. } \frac{96-39}{13} = 4.38 \sim 5 \text{ (always round up)}$$

78	81	55	60	65	86	44	90
77	71	62	39	80	72	70	64
88	73	61	70	75	96	51	73
59	68	65	81	78	67		

Marks(%)	Midpoint	Tally	Frequency	Relative Frequency (round to 3 decimal places)
34.5-39.5	37	I	1	1 ÷ 30 = 0.033
39.5-44.5	42	I	1	0.033
44.5-49.5	47	-	0	0
49.5-54.5	52	I	1	0.033
54.5-59.5	57	II	2	0.067
59.5-64.5	62	IIII	4	0.133
64.5-69.5	67	IIII	4	0.133
69.5-74.5	72	I	6	0.200
74.5-79.5	77	IIII	4	0.133
79.5-84.5	82	III	3	0.100
84.5-89.5	87	II	2	0.067
89.5-94.5	92	I	1	0.033
94.5-99.5	97	I	1	0.033
Total			30	

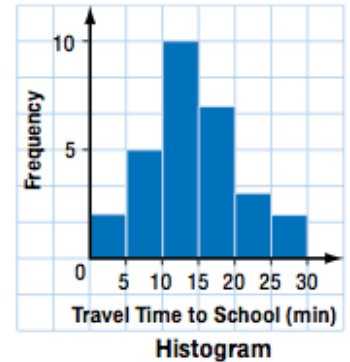
Bar Graph

- a chart or diagram that represents quantities with horizontal or vertical bars whose lengths are proportional to the quantities
- represents all kinds of variables that may not have set order such as hair colour or citizenship



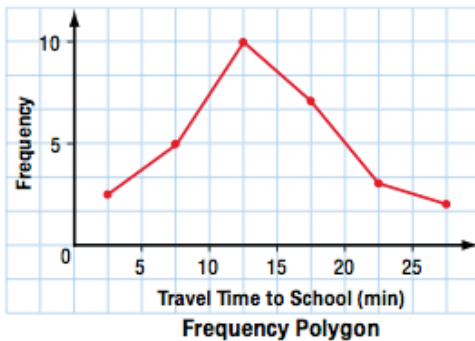
Histogram

- a special form of bar graph which the areas of the bars are proportional to the frequencies of the values of the variable
- used for variables whose values can be arranged in numerical order especially continuous variables such as weight, temperature

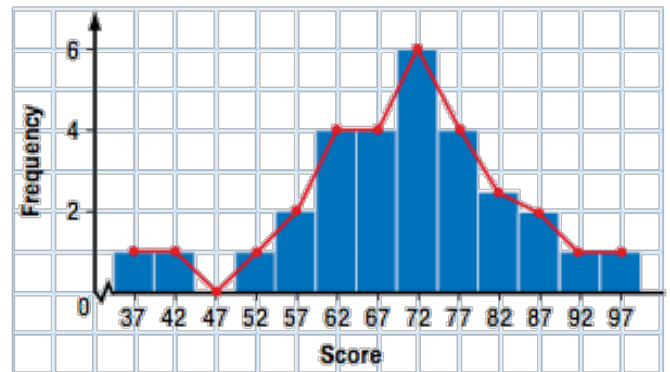


Frequency Polygon




















































- illustrates the same information as a histogram or bar graph by plotting frequencies versus variable values and then joining the points with straight lines




Superimposed frequency polygon and histogram



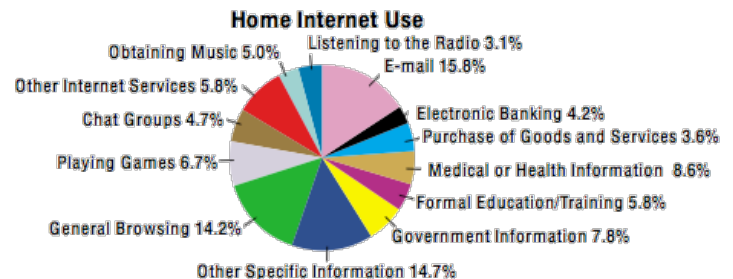
Pictograph

Home Internet Use	
E-mail	      
Electronic Banking	  
Purchase of Goods and Services	 
Medical or Health Information	    
Formal Education/Training	  
Government Information	   
Other Specific Information	      
General Browsing	     
Playing Games	   
Chat Groups	  
Other Internet Services	  
Obtaining Music	 
Listening to the Radio	 

Each  represents 2% of households.

Circle/Pie Graph

Example from Textbook Page 98:

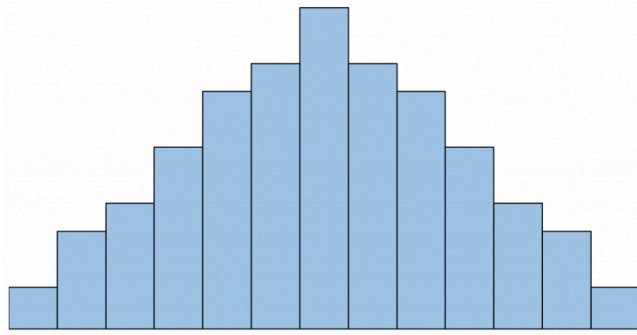


More about Histograms...

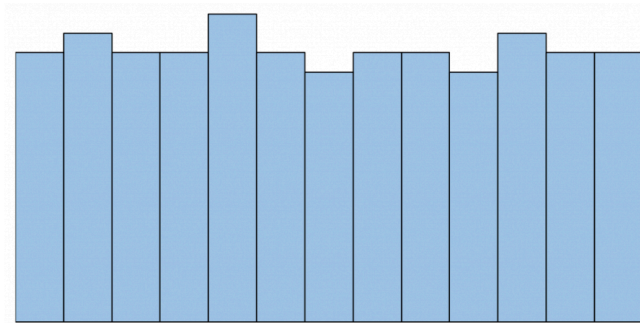
How to describe the shape of histograms:

- ☒ **x-axis** displays the **values** of the dataset
- ☒ **y-axis** displays the **frequencies** of each value

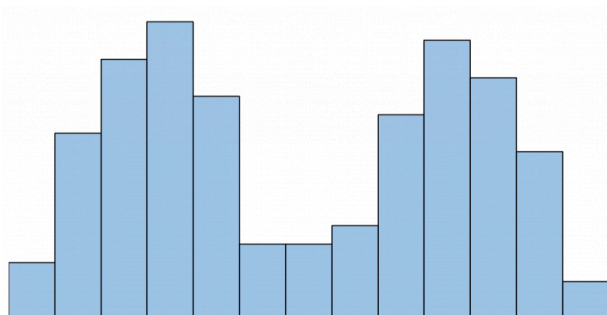
- 1) **Bell-Shaped**: a histogram is bell-shaped if it resembles a “bell” curve and has one single peak in the middle of the distribution. The most common real-life example of this type of distribution is the normal distribution.



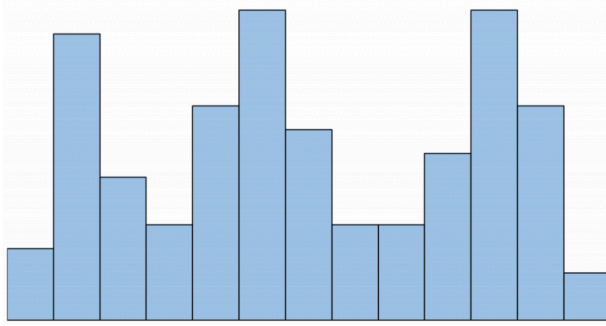
- 2) **Uniform**: A histogram is described as “uniform” if every value in a dataset occurs roughly the same number of times. This type of histogram often looks like a rectangle with no clear peaks.



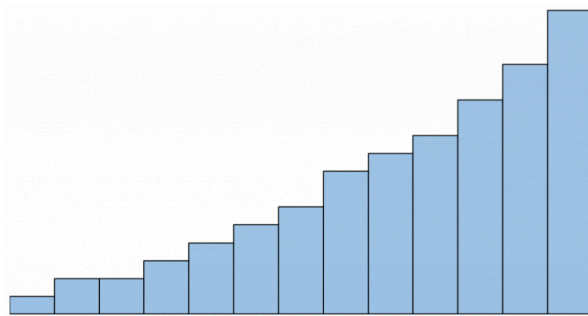
- 3) **Bimodal**: A histogram is described as “bimodal” if it has two distinct peaks.



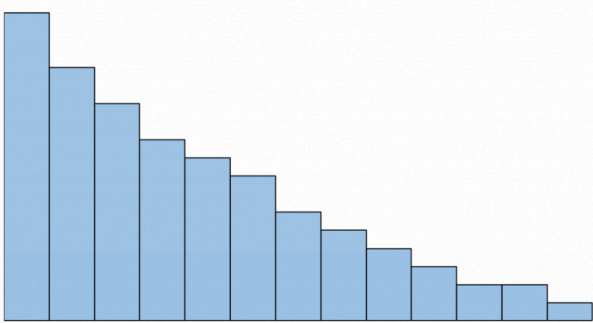
- 4) **Multimodal**: A histogram is described as “multimodal” if it has more than two distinct peaks.



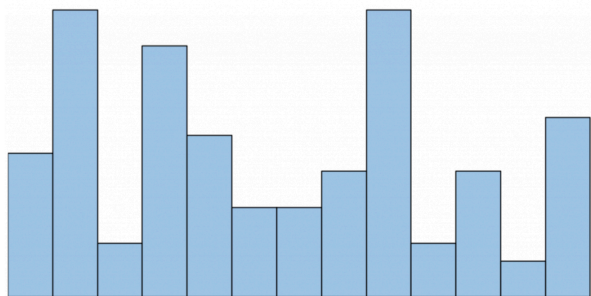
- 5) **Left Skewed**: A histogram is left skewed if it has a “tail” on the left side of the distribution. Sometimes this type of distribution is also called “negatively” skewed.



- 6) **Right Skewed**: A histogram is right skewed if it has a “tail” on the right side of the distribution. Sometimes this type of distribution is also called “positively” skewed.



- 7) **Random**: The shape of the distribution can be described as “random” if there’s no clear pattern in the data at all.



Organizing and Graphing Data Practice Assignment

Continuous Data

$$\text{interval size} = \frac{\text{possible range of data}}{\# \text{ of intervals}}$$

Relative Frequency Distribution

1. Given the final marks from last year’s Grade 12 Data Management class.

42	49	50	51	63	66	67	68	69	70
73	74	77	78	80	83	84	85	85	88
89	90	91	93	95	96	97	98	99	100

a) Using the formula of determining the interval size above, construct a frequency table with 7 intervals, includes the column for the relative frequency

Intervals for Marks (%)	MIDPOINT	TALLY	FREQUENCY	RELATIVE FREQUENCY

b) Construct a histogram. (Be sure to label the graph clearly)



c) What proportion of students had marks between 60% and 69%?

This means exclusively

d) Do you find the method of determining the interval from a) effective for analyzing this scenario? Why?

e) If you are going to re-do the frequency table, how would you do it this time?

Intervals for Marks (%)	MIDPOINT	TALLY	FREQUENCY	RELATIVE FREQUENCY

f) Why would you choose to organize the data using the selected interval style from e)?

Lesson: Measures of Central Tendency (One-variable statistics)

When you analyze data that has been collected, the first measure to find is the “average” result for the collection.

There are three measures of central tendency that we can calculate.

- **Mean** – add all the data and divide by the number of data items.
- The symbol used for sample mean is \bar{x} .
- The symbol used for population mean is μ .

The formula used to calculate the sample mean is $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$, where n represents the total number of values and x represents the observed values.

- **Median** – the middle piece of data when the items are ranked from least to greatest
To calculate the median, arrange the data values for least to greatest. Calculate $\frac{n+1}{2}$, where n represents the number of values, to determine the location of the middle value. If the data presents an even number of values, average the two values in the middle position.
- **Mode** – the most frequencyly occurring data value
To find the mode, determine the most frequent value. It is possible for there to exist more than one mode.

When given a set of numbers:		
1. Given a set of numbers: 4, 2, 7, 6, 9		
a. Find the mean.	b. Find the median.	
2. Given a set of numbers: 5, 2, 7, 4, 9, 6		
a. Find the median.		
3. Given a set of number: 1, 1, 2, 3, 4, 5 Find the mode.	4. Given a set of number: 1, 1, 2, 2, 3, 3, 4, 4 Find the mode.	5. Given a set of number: 1, 1, 2, 2, 3, 4, 5 Find the mode.

When given a set of numbers involving frequencies:

6. Given a set of numbers involving frequencies:

x	2	5	4	9
<i>frequencies</i>	3	2	1	4

a. Find the mean.

b. Find the median.

c. Find the mode.

When given a group of distribution:

7. Given a group distribution:

x	0-9	10-19	20-29	30-39
<i>frequencies</i>	3	2	1	4

a. Find the mean.

b. Find the median.

c. Find the mode.

When given a set of continuous data:

8. Given a set of continuous data:
1.2111, 1.1212, 1.1213, 2.1111, 2.1212, 3.1314, 4.1516
Find the mode.

Practice Measure of Central Tendency:

Knowledge Questions:

- 1) Given the following set of data:

4,9,4,5,3,8,10,12

- a) Find the mean of the given set of data. [6.875]
- b) Find the median of the given set of data. [6.5]
- c) Find the mode of the given set of data. [4]

- 2) Determine the interval size if 8 groups are intervals are needed: [6]

12,23,25,27,27,32,34,39,40,41,41,45,47,49,50,51,52,56,59,60,66,70

Application Questions:

- 3) A class consists of 50 students, out of which 30 are girls. The mean of marks scored by girls in a test is 73 out of 100, and that of boys is 71 out of 100. Determine the mean score of the whole class. [72.2]

- 4) The mean of the following distribution is 50. Find the value of a and the frequencies of 30 and 70. [5, 28,24]

x	$frequency$
10	17
30	$5a + 3$
50	32
70	$7a - 11$
90	19

Challenging Question:

- 5) Given three positive integers, a , b , and c . Their average is 20, where $a \leq b \leq c$. If the median is $(a + 11)$, what is the least possible value of c ? [25]

Lesson: Measures of Central Tendency (One-variable statistics)

Weighted Mean: A **weighted** mean is calculated when some of the data items in an average are worth more than others.

The formula used to calculate a weighted mean is $\overline{x}_w = \frac{\sum w_i x_i}{\sum w_i}$

Example: The overall performance of a student in Data Management was recorded as follows:

CATEGORY	WEIGHT	STUDENT MARK %
Knowledge and Understanding	24%	75%
Application	18%	70%
Thinking	14%	60%
Communication	14%	80%
Culminating Task (ISP)	15%	70%
Final Exam	15%	55%

a) Calculate the student's term mark. (The term mark does not include the culminating task or the final exam.)

b) Calculate the student's final course mark.

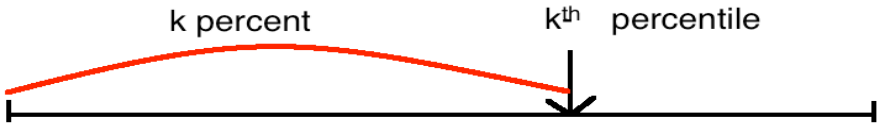
Time to determine your own updated mark in this course from TeachAssist:

Categories	Knowledge (22%)	Application (28%)	Thinking (10%)	Communication (10%)
Unit 1 Assignment				
Unit 2 Day 1 Test				
Unit 2 Day 2 Test				
Total				

Learning Goal: Measures of Spread (Measure of Dispersion)

(One-variable statistics)

PERCENTILE – divide the data into 100 intervals that have equal numbers of values. It is a measure that indicates what percent of the given population scored at or below the measure.



To calculate a particular percentile:

- 1) Multiply the total number of values in the data set by the percentile which will give you the index.
- 2) Order all of the values in the data set in ascending order (least to greatest)
- 3) If the index is a whole number, count the values in the data set from least to greatest until you reach the index, then take the index and the next greatest number and find the average.
- 4) If the index is not a whole number, **round the number up**, then count the values in the data set from least to greatest until you reach the index.
- 5) If the index is at .5, take the mean value of the raw data above and below in the data set from least to greatest.

Textbook page 145:

35	47	57	62	64	67	72	76	83	90
38	50	58	62	65	68	72	78	84	91
41	51	58	62	65	68	73	79	86	92
44	53	59	63	66	69	74	81	86	94
45	53	60	63	67	69	75	82	87	96
45	56	62	64	67	70	75	82	88	98

Percentile Rank: (R) to find the raw score in the data set that represent the percentile

$$R = \frac{p}{100}(n + 1)$$

p = percentile; n = size of the population

Percentile (p): to determine the percentile of a raw score

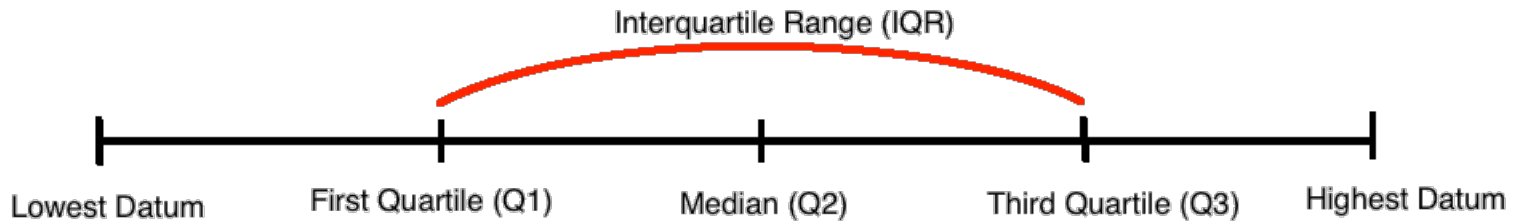
$$p = 100\left(\frac{L + 0.5E}{n}\right)$$

L = number of data less than the data point

E = number of data equal to the data point

- | | | |
|--|---|--|
| a) If a datum scored at 50 th percentile, what was its raw score? | b) What is the 90 th percentile for this data? | c) Does the score of 75 place it at the 75 th percentile? |
|--|---|--|

Quartile and Interquartile Range (IQR):



$$Q2 = \text{Median}$$

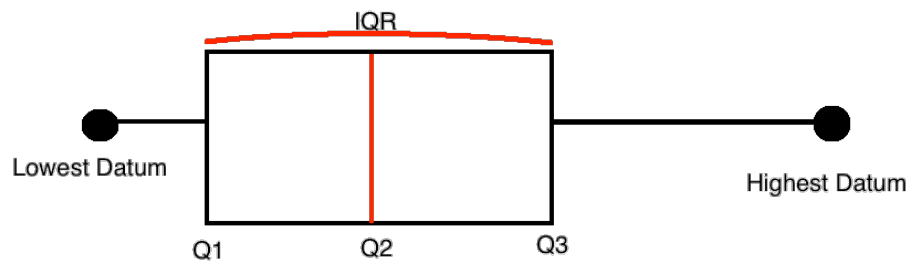
$Q1 = \text{Median of the first half of the data}$

$Q3 = \text{Median of the second half of the data}$

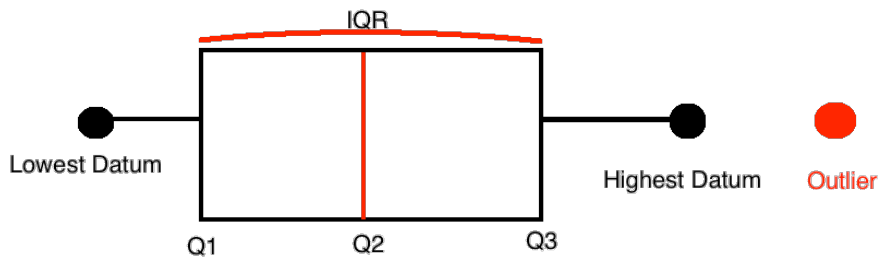
$$IQR = Q3 - Q1$$

$$\text{SemiIQR} = \text{SIQR} = \frac{IQR}{2}$$

Box-And-Whisker Plot:



Modified Box-And-Whisker Plot:



To determine the value(s) of an outlier:

$$1.5 \times IQR$$

If a datum is less than $Q1 - (1.5 \times IQR)$, it is considered as an outlier.

If a datum is greater than $Q3 + (1.5 \times IQR)$, it is considered as an outlier.

Example #1: A survey of movie goers at a screening of “Rocky Horror Picture Show” asked how many times they have seen the movie. The results for 20 respondents were:

3 4 2 8 10 5 1 15 5 16 6 3 4 9 12 3 30 2 10 7

- a) Find the median, Q1, and Q3.
- b) Calculate IQR and SIQR.
- c) Are there any outliers? Explain.
- d) Make a box-and-whisker plot of the results.
- e) Make a modified box-and-whisker plot of the results accounting for any outliers. How do the quartiles compare?

Using the previous example, find the score of:

- a) 20th percentile
- b) 65th percentile
- c) 95th percentile

Learning Goal: Measures of Spread (Measure of Dispersion)

One-Variable Statistics

Deviation is the difference between an individual value in a set of data and the mean for the data.

x = the individual value in a set

μ = the population mean

\bar{x} = the sample mean

For the population, deviation is $x - \mu$

For a sample, deviation is $x - \bar{x}$

The larger the size of the deviations, the greater the spread in the data.

Standard Deviation – measure of variability (How much variation exists from the average)

- low standard deviation indicates the data points are very close to the mean
- High standard deviation indicates the data points are very far from the mean
- To determine if the standard deviation is high or low, compare it to the range of the dataset: if the standard deviation is close to the range, it's considered high; if it's significantly smaller, it's considered low.

Variance - how spread out numbers are

Measure	Sample	Population
Standard Deviation	$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$ <p>n is the number of data in sample</p>	$\sigma = \sqrt{\frac{\sum(x - \mu)^2}{N}}$ <p>N is the number of data in population</p>
Variance	$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$	$\sigma^2 = \frac{\sum(x - \mu)^2}{N}$

Z-Scores - the number of standard deviation that a datum is from the mean

Sample	Population
$z = \frac{x - \bar{x}}{s}$ <p>s=standard deviation of the sample</p>	$z = \frac{x - \mu}{\sigma}$ <p>σ = standard deviation of the population</p>

- Variables below mean have NEGATIVE Z-SCORES
- Variables above mean have POSITIVE Z-SCORES
- Variables equal to mean have ZERO Z-SCORE

Ex. #1 -the number of hours five students spent studying statistics last week: 8 4 9 11 3

a) Calculate Mean

b) Calculate Variance

c) Calculate Standard Deviation

Ex. #2 – The number of summer jobs a sample of six students applied for: 17 15 23 7 9 13
Calculate the Standard Deviation

Ex. #3 – Comparing Consistency of Two Types of Golf Clubs Consistency is the hallmark of a good golfer. Golf equipment manufacturers are constantly seeking ways to improve their products. Suppose that a recent golf innovation is designed to improve the consistency of its users. As a test a golfer was asked to hit 150 shots using a 7-iron, 75 of which were hit with his current club and 75 with the new innovative 7-iron. The distances were measured and recorded. Which 7-iron is more consistent?

Current	
Mean	150.55
Standard Deviation	5.79
Sample Variance	33.55

Innovative	
Mean	150.15
Standard Deviation	3.09
Sample Variance	9.56

Ex. #4 – If the mean is 68.1 and the standard deviation is 15.2, determine the z-scores for Audio Maximizer Ultra 3000 (score 67) and SchmederVox (score 75).

Lesson: Linear Correlation

Two-Variable Statistics

Scatter Plot

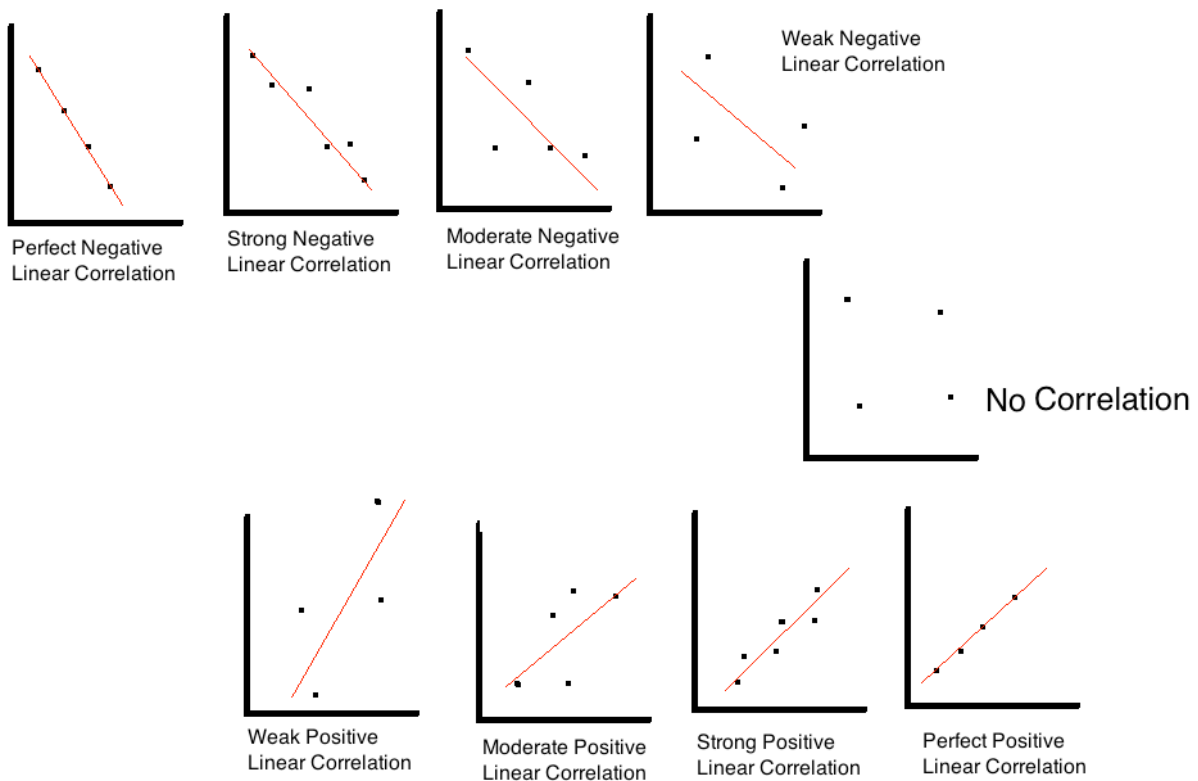
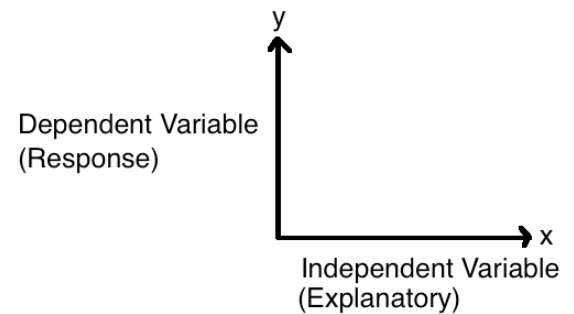
- graph that describes relationship between two variables

Line of Best Fit

- passes as close as possible to all the points on a scatter plot and represents the relationship between two variables

Linear Correlation

- changes in one variable that are proportional to changes in another
- weak, moderate, or strong (strength of the relationship)
- positive or negative (direction of the relationship)

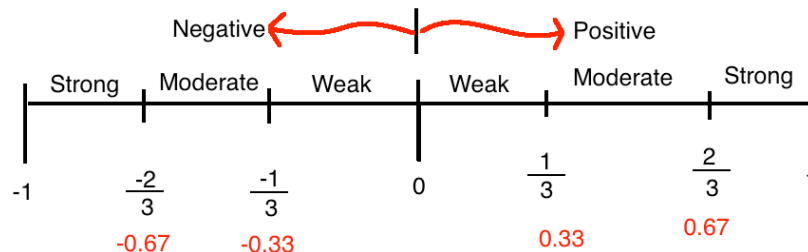


Coefficient of Correlation

- the strength between the two variables

$$r = \frac{n \sum xy - (\sum x)(\sum y)}{\sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2]}}$$

Coefficient Correlation



Lesson: Linear Regression

Two-Variable Statistics

Linear Regression

- a technique to find $y = ax + b$ for the line of best fit
- “Least Square Method”

$$a = \frac{n \sum xy - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

$$b = \bar{y} - a\bar{x}$$

Covariance

- measure of correlation which reveals whether the relationship between the variables is positive or negative

$$S_{xy} = \frac{1}{n-1} \left(\sum xy - \frac{\sum x \sum y}{n} \right)$$

Residual

- the difference between the values of y at the data point and at the point that lies on the line of best fit and has the same x -coordinate as the data point

$$\text{Residual} = \text{observed} - \text{predict}$$

positive = data above the line of best fit

negative = data below the line of best fit

Coefficient of Determination for [non-linear regression](#)

- r^2 (the square of the correlation coefficient)
- it measures the proportion of the variation in y that is explained by the variation in x
- it tells what percentage of points that are on the line/curve of best fit
- $0 \leq r^2 \leq 1$, $r^2 = 1$ the curve is a perfect fit
- this applies to any type of regression

Linear Correlation and Regression Practice

Example #1 Given the two variables below in the table,

a) determine the linear regression and coefficient of correlation.

x	y	x^2	y^2	xy
2	13			
6	20			
7	27			
$\sum x$	$\sum y$	$\sum x^2$	$\sum y^2$	$\sum xy$

$$r = \frac{n \sum xy - (\sum x)(\sum y)}{\sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2]}}$$

$$a = \frac{n \sum xy - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

$$b = \bar{y} - a\bar{x}$$

b) What do the statistics you have calculated tell you about the relationship between the two variables?

Example #2:

Are the marks one receives in a course related to the amount of time spent studying the subject? To analyze this possibility, a student took a random sample of 10 students who had enrolled in an accounting class last semester. She asked each to report his or her mark in the course and the total number of hours spent studying accounting. These data are listed here.

Hours spent	40	42	37	48	25	44	41	48	35	28	
Marks	77	63	79	86	51	78	83	90	65	47	

a) Calculate the covariance.

$$S_{xy} = \frac{1}{n-1} \left(\sum xy - \frac{\sum x \sum y}{n} \right)$$

) Calculate the coefficient of correlation.

$$r = \frac{n \sum xy - (\sum x)(\sum y)}{\sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2]}}$$

c) What do the statistics you have calculated tell you about the relationship between marks and study times?

Example #3: The table below shows a sample of employees in the company of their ages, in years, and annual income, in thousands of dollars.

Annual Income (\$)	Age (years)			
33	33			
31	25			
18	19			
52	44			
56	50			
60	54			
44	38			
35	29			

- a) Using Least-Squares method, determine the equation of the linear regression.
- b) Predict the annual income for an employee who is 21 and an employee retiring at age 65.
- i) For a 21-year-old employee,
- ii) For a 65-year-old employee,
- c) Refer to b), determine which one is interpolation and which one is extrapolation.
- d) Comment on the accuracy of both estimates.

Lesson: Cause-and-Effect Relationship

Two-Variable Statistics

5 different types and degrees of causal relationship between variables:

1) **Cause-and-Effect Relationship**

A change in x produces a change in y

2) **Common-Cause Factors**

External variable(s) causes two variables to change in the same way

3) **Reverse Cause-and-Effect Relationship**

The dependent and independent variables are reversed in the process of establishing causality

4) **Accidental Relationship**

A correlation exists without any causal relationship between variables that happens by random chance

5) **Presumed Relationship**

A correlation does not seem to be accidental even though no cause-and-effect relationship or common cause factor is apparent.

Try This... Classify the relationships in the following situations.

- a) The rate of a chemical reaction increases with temperature
- b) Leadership ability has a positive correlation with academic achievement
- c) The price of butter and motorcycles have a strong positive correlation over many years
- d) Sales of cellular telephones had a strong negative correlation with ozone levels in the atmosphere over the last decade
- e) Traffic congestion has a strong correlation with the number of urban expressways

Extraneous Variables

- Variables that affect or obscure the relationship between an independent and a dependent variable

Experimental Group

- The group for which the independent variable is changed in an experiment or statistical study

Control Group

- The group for which the independent variable is held constant in an experiment or statistical study

Example1:

A medical researcher wants to test a new drug believed to help smokers overcome the addictive effects of nicotine.

- One hundred people who want to quit smoking volunteer for the study.
- The researcher carefully divides the volunteers into two group, each with an equal number of moderate and heavy smokers.
- One group is given the nicotine patches with the new drug, while the second group uses ordinary nicotine patches.
- Twenty-eight people in the first group quit smoking completely, as do eighteen people in the second group.

Identify the experimental and control groups.