Unit 1: Organizing Data: Look for Patterns & Departures from Patterns

Chapter 1: Exploring Data

1.1A: Displaying Distributions with Graphs

Individuals: - commonly known as "subjects" in a statistical study.
- can be people, animals, or other items.

Variables: - the data that are measured in the study.
- must be well-defined and common units must be decided before doing the survey.

1. **Categorical Variables**: - variables with non-numerical values.

   Example: Yes or No, Colour Preference

2. **Quantitative Variables**: - variables with non-numerical values.

   Example: Exam Scores, Heights

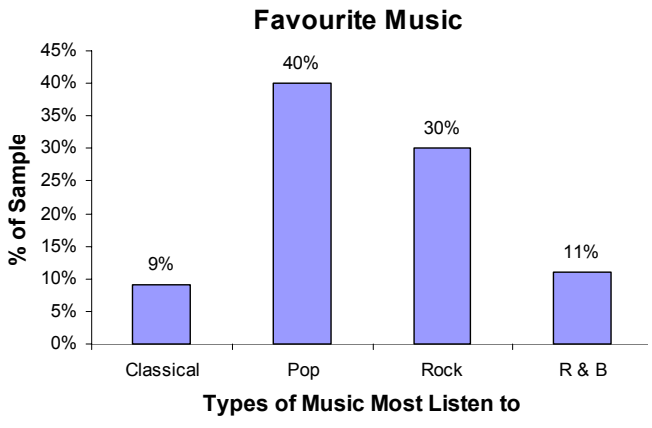
There are two ways to do a proper **Exploratory Data Analysis**:

1. Examination of Variables BEFORE Studying Relationship among the Variables.
2. Examination of Graphs BEFORE Summarizing Specific Aspects of the Results Numerically.

Distribution: - a graph that properly display the patterns and relationships between individuals and variables or amongst variables.

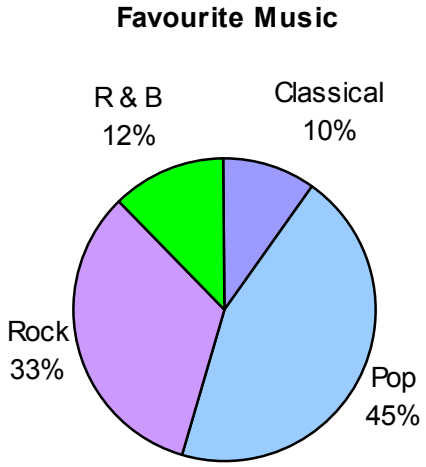
To display Categorical Variables, it is best to use **Bar** or **Pie Chart**:

Bar Chart: - for Non-Continuous categories (bars can be switched around in different order). Thus, gaps between bars must be shown.



Pie Chart:

$$\text{Interior Sector Angle} = \%\ \text{of Population} \times \frac{360^\circ}{100\%}$$



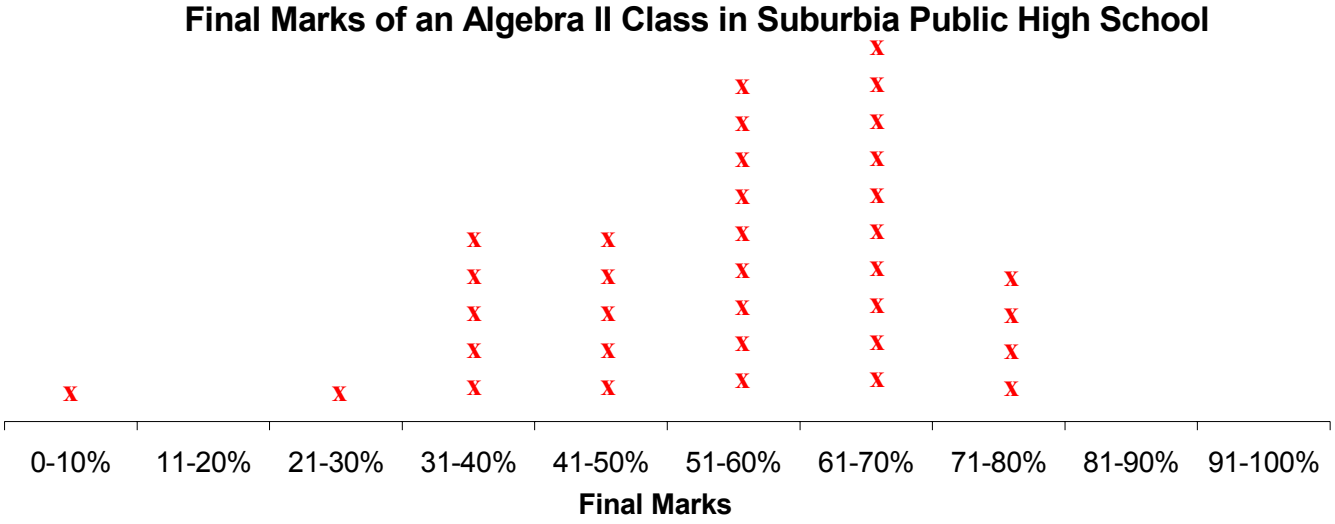
To display Quantitative Variables, it is best to use **Dotplots, Histograms, Stemplots, Back-to-Back Stemplots, or Time Plot**:

In order to illustrate the various types of plots, and in subsequent sections, we will use the following data:

Final Marks in Percentage for an Algebra II Class in Suburbia Public School

Student #	Final Mark	Student #	Final Mark	Student #	Final Mark	Student #	Final Mark
1	50	10	56	19	63	28	56
2	32	11	32	20	72	29	9
3	77	12	50	21	67	30	65
4	59	13	29	22	70	31	95
5	65	14	60	23	50	32	34
6	45	15	45	24	68	33	60
7	60	16	43	25	42	34	45
8	55	17	50	26	65	35	61
9	32	18	34	27	50	36	74

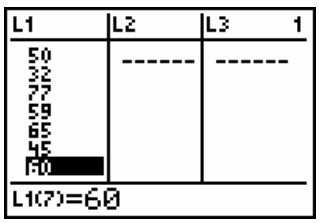
Dotplot: - using dots to quickly and plot a range of data.



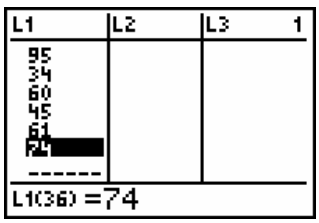
Entering Data using TI-83 Plus Calculator:

STAT **ENTER**

Enter Values



Check if you entered the Correct Number of Data Values



36th Score entered

Range: - the difference between the maximum and the minimum scores.

Sorting Data using TI-83 Plus Calculator:

STAT

Select option 2

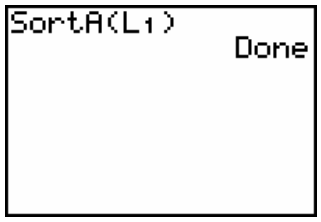
ENTER

2nd

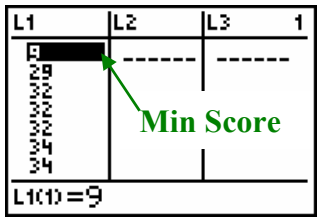
L1

1

ENTER

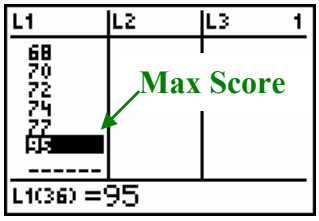


STAT ENTER



Examine the Result

Hold Down



Range = 95% – 9%

Range = 86%

Histogram: - continuous data on the *x*-axis.
- bars cannot be in different order, Thus, there are no gaps between bars.

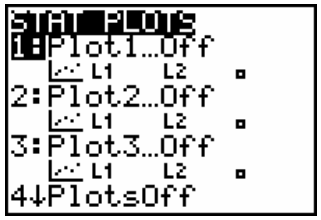
To plot Histogram on TI-83 Plus Calculator:

1. Because we are plotting from a set of data, not from an equation, we have to turn on STAT PLOT.

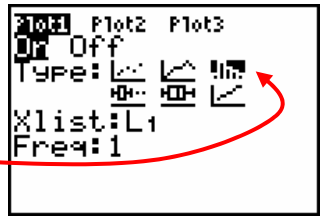
2nd

STAT PLOT

Y =



ENTER

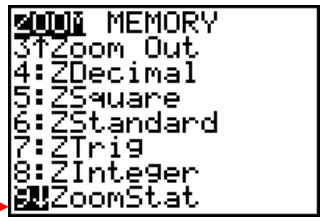


Select Histogram

2. Use ZoomStat to estimate Window Settings

ZOOM

Select Option 9



then fine tune Windows

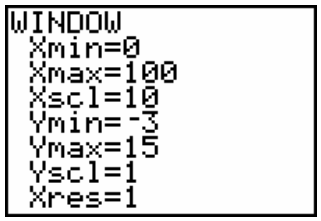
***x*: [x_{min}, x_{max}, x_{scl}]**

***y*: [y_{min}, y_{max}, y_{scl}]**

WINDOW

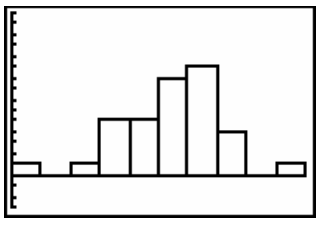
x: [0, 100, 10]

y: [–5, 10, 1]



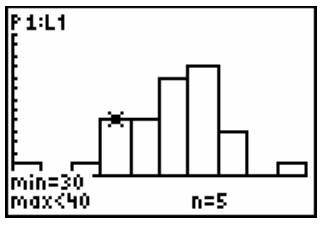
3. Graph Histogram

GRAPH

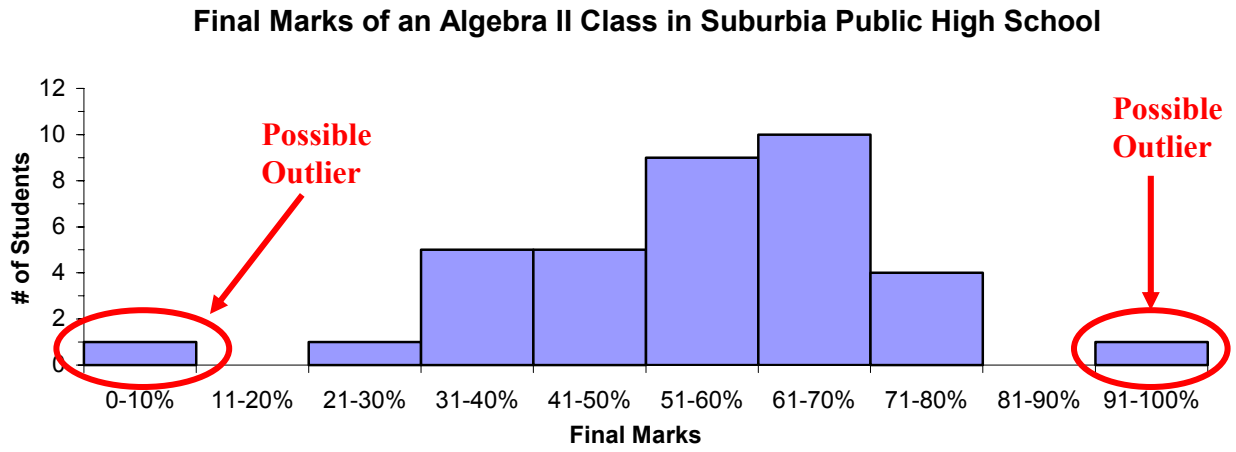


4. Trace to Verify

TRACE



Outlier: - data point outside the general pattern of the graph.
- *Note:* the scale used in both axes can affect identification of outliers.



Describing Distribution

a. **Center**: - the location of various central tendencies (mean, median, and mode)

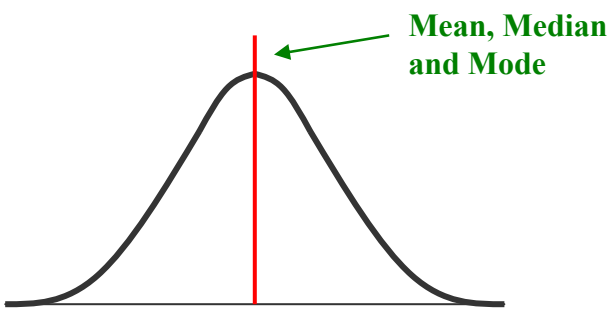
- **i.** **Mean**: - arithmetic average that is easily affected by extreme scores and outliers.
- **ii.** **Median**: - the middle score when the data are lined up from least to greatest.
  - use when there are extreme scores or outliers; not easily affected by them.
- **iii.** **Mode**: - the most frequent score.
  - used often for categorical variable but most often misused.

b. **Spread**: - how the scores spread out across a given range.

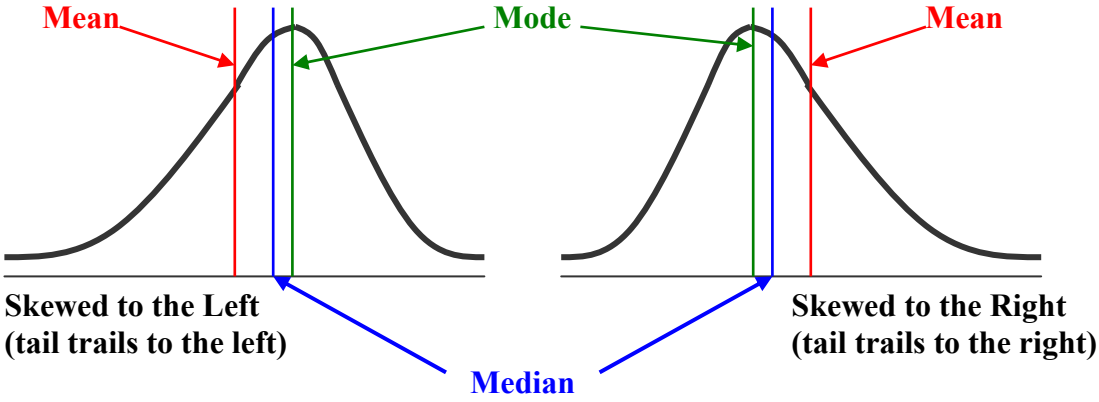
Narrow or Small Spread **Wide or Big Spread**

c. **Shape**: - there are various shape to a distribution:

- **i.** **Symmetrical**: - all central tendencies are the same.

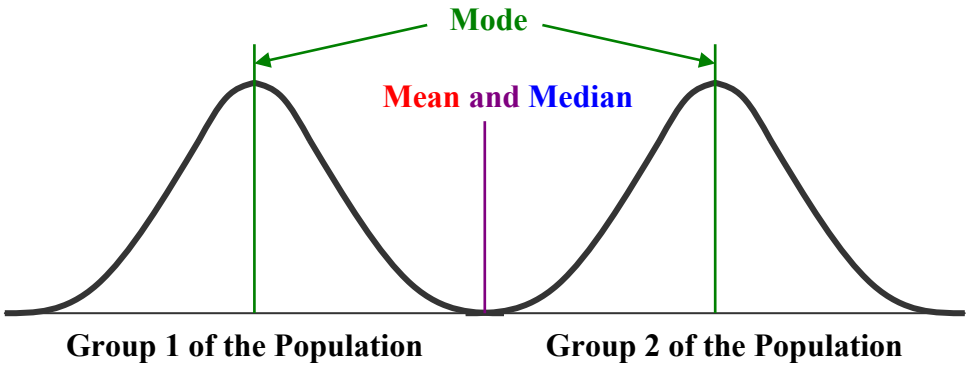


ii. **Skewed**: - some extreme scores pull the tail either to the left or the right.



Mean is most affected by extreme scores. Therefore, it is closer to the tail.
Median is less affected by extreme scores. Therefore it is closer to the mode.

iii. **Bivariant**: - shows two distinct groups of population.



1.1A Assignment: **pg. 9 #1.3; pg. 16–17 #1.5 and 1.7**

1.1B: Displaying Distributions with Graphs (Continued)

Stemplot: - similar to the dotplot; appears sideways and using digits instead of dots.

Final Marks in Percentage for an Algebra II Class in Suburbia Public School

Stem	Leaf
0	9
1	
2	9
3	2 2 2 4 4
4	2 3 5 5 5
5	0 0 0 0 0 5 6 6 9
6	0 0 0 1 3 5 5 5 7 8
7	0 2 4 7
8	
9	5

Rounding: - is allowed when there are two many digits.

Example: 4.782 ⟶ 4.8

2	
3	
4	8
5	

Splitting Term: - is allowed when there are too many numbers in the stem and the range is small.

Example:

3	1 2 3 3
3	7 8 9 9 9 9
4	2 3 4 4
4	5 6 6

Back to Back Stemplot: - when there are two separate groups like male and female.

Example:

Boys (inches)	Height (ft)	Girls (inches)
6 6 5 4 3 2	6	2 3
11 10 9 9 8 7	5	2 3 4 5 6 8 9
8	4	5 7 8 9

Timeplot: - a broken line graph that plots data against time using at least two columns.

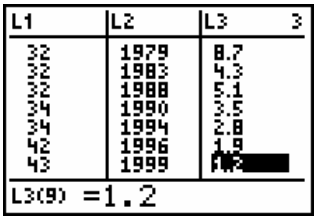
Example: Use the following table to graph a timeplot.

Year	1970	1972	1979	1983	1988	1990	1994	1996	1999
Canadian Inflation Rate	12.3%	9.1%	8.7%	4.3%	5.1%	3.5%	2.8%	1.9%	1.2%

To plot Timeplot on TI-83 Plus Calculator:

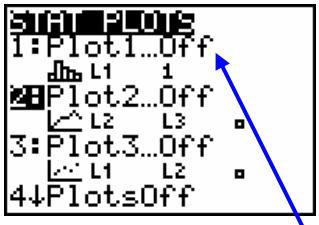
1. Enter the two columns in L_2 and L_3

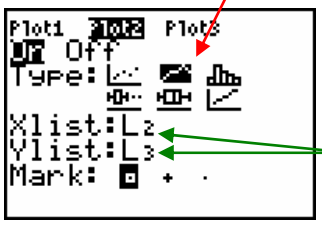
STAT ENTER



2. Turn Off Plot 1, turn On Stat Plot 2 and select Timeplot

2nd STAT PLOT Y = ENTER





Plot 1 Off

Use L_2 and L_3

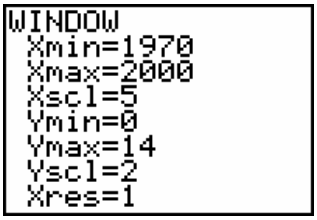
2nd L_2 2 2nd L_3 3

3. Set Windows

x: [1970, 2000, 5]

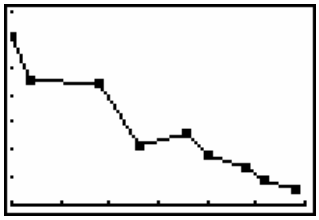
y: [0, 14, 2]

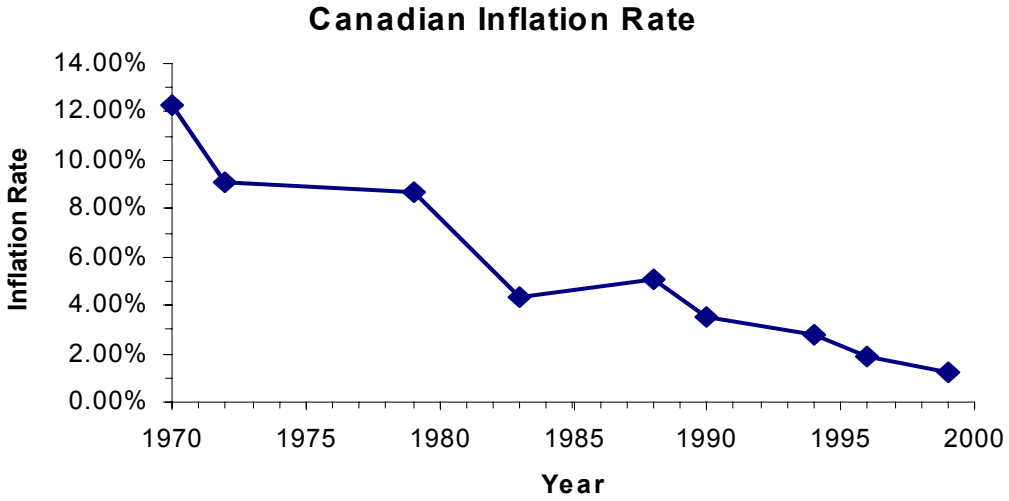
WINDOW



4. Graph Timeplot

GRAPH





1.1B Assignment: pg. 25–28 #1.15, 1.17 and 1.19

1.2A: Describing Distributions with Numbers

Measuring Center

1. **Mean** ($\bar{x}$): - the arithmetic average
   - mean is **Non-Resistant (easily affected) by extreme scores.**
   - **skewed distribution pushes the mean towards the tail of the scores.**

$$\bar{x} = \frac{x_1 + x_2 + x_3 + ... + x_n}{n}$$

or

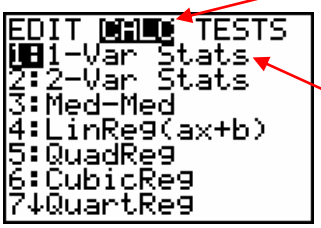
$$\bar{x} = \frac{1}{n}\sum x_i$$

Summation of all x

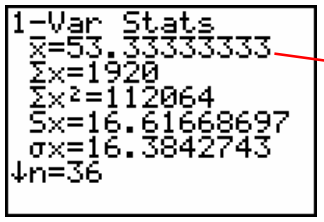
Example: Determine the mean of the *Final Marks in Percentage for an Algebra II Class in Suburbia Public School* from section 1.1A.

To find the Mean using TI-83 Plus Calculator:

STAT **Select CALC on Menu** **ENTER**



Select 1-Var Stats



$\bar{x}$ = 53.33

2. **Median** (*M*): - the middle score when the data are arranged from the least to the greatest.
   - Median is **Resistant (not easily affected) by extreme scores.**
   - **Median is closer to the mode compared to the mean in skewed curves.**

Example: Find the Median of 3, 5, 6, 7, 7, 9

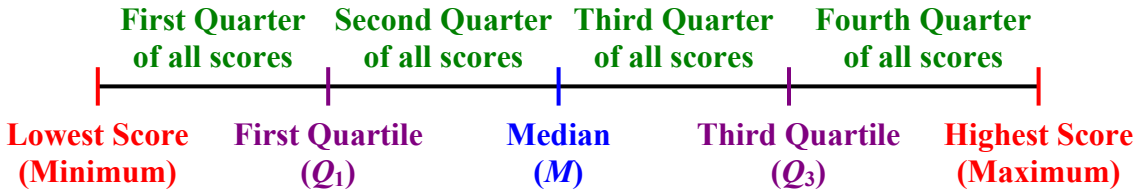
3, 5, 6, 7, 7, 9

$$M = \frac{6+7}{2}$$

***M* = 6.5**

Since there are even number of scores, we have to average the two middle scores.

Quartiles: - the middle half of a set of scores.



First Quartile (Q_1) = Median between (Min) and (M)
Third Quartile (Q_3) = Median between (M) and (Max)

Example: Determine the median, first and third quartiles of the *Final Marks in Percentage for an Algebra II Class in Suburbia Public School* from section 1.1A.

The Median is found by averaging out the middle two scores (18th and 19th scores).

9, 29, 32, 32, 32, 34, 34, 42, 43, 45, 45, 45, 50, 50, 50, 50, 50, 55, 56, 56, 59, 60, 60, 60, 61, 63, 65, 65, 65, 67, 68, 70, 72, 74, 77, 95

$$M = \frac{55+56}{2}$$ **$M = 55.5$**

The First Quartile is the middle two scores between the minimum and the 18th score (9th and 10th scores).

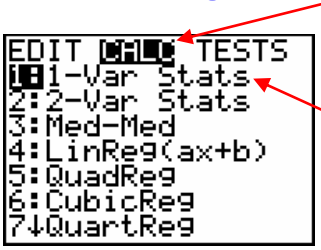
$$Q_1 = \frac{43+45}{2}$$ **$Q_1 = 44$**

The Third Quartile is the middle two scores between the 19th score and maximum (27th and 28th scores).

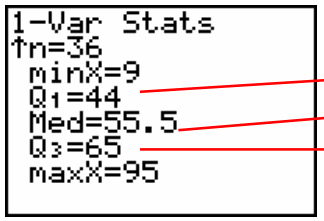
$$Q_3 = \frac{65+65}{2}$$ **$Q_3 = 65$**

To find the Median, Q_1 and Q_3 using TI-83 Plus Calculator:

STAT → **Select CALC on Menu** → **ENTER** → many times



Select 1-Var Stats

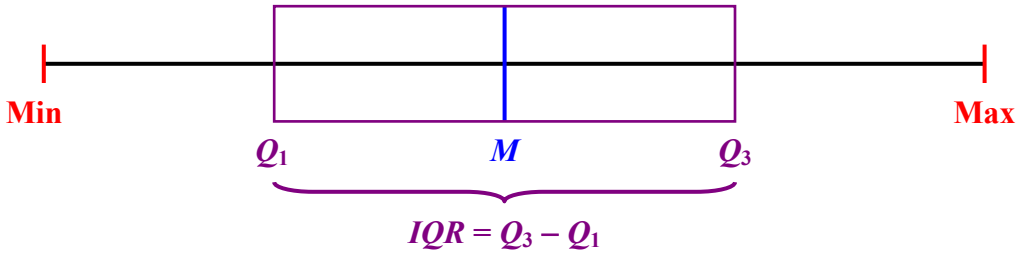


$Q_1 = 44$
$M = 55.5$
$Q_3 = 65$

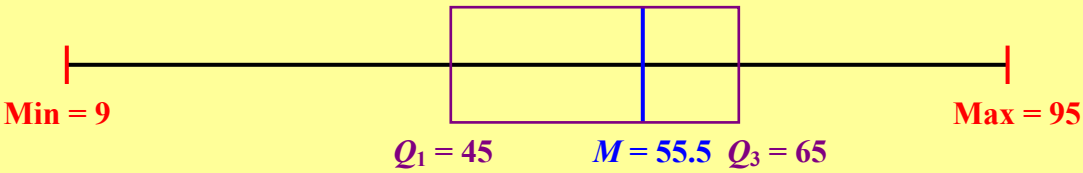
Five Number Summary: - Min, Q_1, M, Q_3, Max

Interquartile Range (*IQR*): - the range between Q_1 and Q_3.

Boxplot: - a plot that graphically show the five number summary.



Example: Create a boxplot and the *IQR* for the *Final Marks in Percentage for an Algebra II Class in Suburbia Public School* from section 1.1A.



$IQR = Q_3 - Q_1$

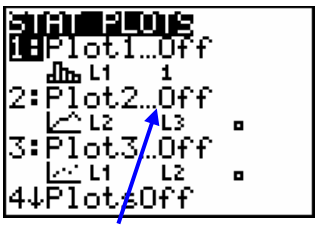
$IQR = 65 - 45$

$IQR = 20$

To plot a Boxplot using TI-83 Plus Calculator:

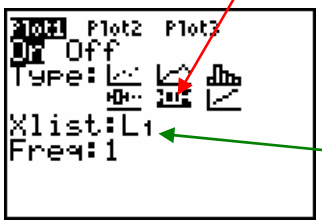
1. Turn Off Plot 2, turn On Stat Plot 1 and select Boxplot

2nd STAT PLOT Y =



Plot 2 Off

ENTER

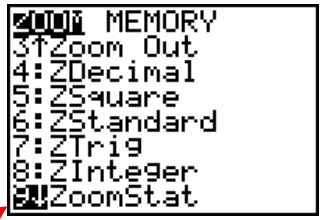


2nd L_1 1

Use L_1

2. Use ZoomStat to estimate Window Settings

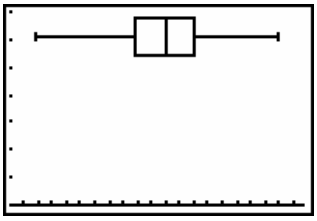
ZOOM



Select Option 9

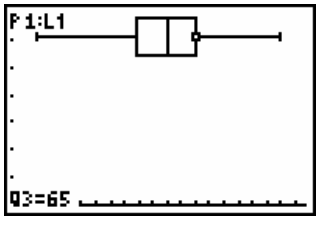
3. Graph Boxplot

GRAPH



4. Trace to Verify

TRACE



or

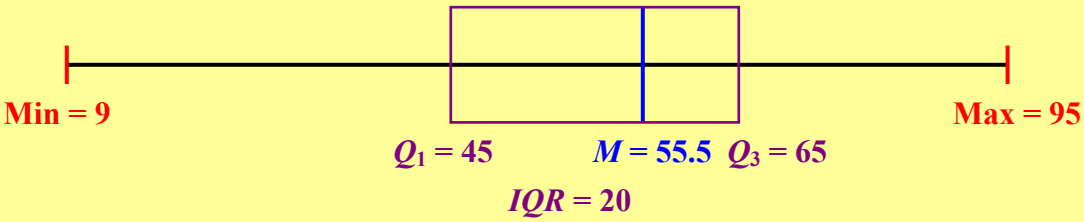
Using *IQR* to test for Outliers

Outlier is defined when a score (x) is:

$$x > (1.5 \times IQR) + Q_3 \quad \text{or} \quad x < Q_1 - (1.5 \times IQR)$$

Basically, it is an outlier when it is located 1.5 times *IQR* above Q_3 or below Q_1.

Example: Determine if the minimum and the maximum is an outlier for the *Final Marks in Percentage for an Algebra II Class in Suburbia Public School* from section 1.1A.



For the minimum at $x = 9$:
$9 < Q_1 - (1.5 \times IQR)$
$9 < 45 - (1.5 \times 20)$

✔ **$9 < 15$**

For the maximum at $x = 95$:
$95 > (1.5 \times IQR) + Q_3$
$95 > (1.5 \times 20) + 65$

✘ **$95 < 95$**

Therefore, the minimum 9 is an outlier and the maximum 95 is not.

Modified Boxplot: - a boxplot that excludes outliers as tested by *IQR*.

To plot a Modified Boxplot using TI-83 Plus Calculator:

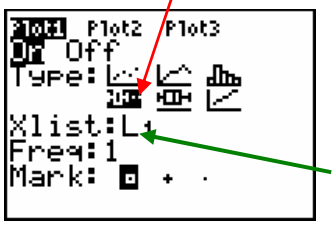
1. Turn On Stat Plot 1 and turn Off Other Plots, and select Modified Boxplot

2nd

STAT PLOT

Y =

ENTER

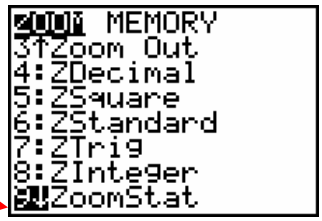


Use L_1

Select Option 9

2. Use ZoomStat to estimate Window Settings

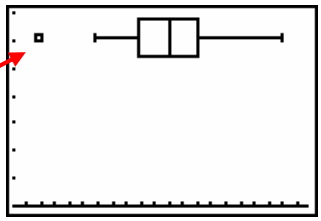
ZOOM



3. Graph Modified Boxplot

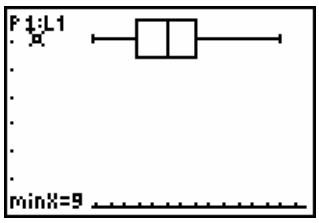
GRAPH

Outlier



4. Trace to Verify

TRACE or



1.2A Assignment:
pg. 34 #1.25
pg. 36–37 #1.27 and 1.29
pg. 42 #1.33

1.2B: Describing Distributions with Numbers (Continued)

Measuring Spread

1. **Variance** (s^2): - average of the **squares of the deviation**, $(x_i - \bar{x})^2$,

differences between each scores and the mean

$$s_x^2 = \frac{(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + (x_3-\bar{x})^2 + \ldots + (x_n-\bar{x})^2}{n-1}$$

or

$$s_x^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2$$

Need to Square the deviation because $\sum (x_i - \bar{x}) = 0$

Degree of Freedom ($n-1$): - only $n-1$ scores can vary freely from the mean.

2. **Standard Deviation** (s): - a measure of how spread out the scores are from the mean.

$$s_x = \sqrt{\frac{1}{n-1}\sum (x_i - \bar{x})^2} \quad \text{or} \quad s_x = \sqrt{s_x^2}$$

Need to take the Square Root to return scores to their original dimensions.

Example: Determine variance and standard deviation for the *Final Marks in Percentage for an Algebra II Class in Suburbia Public School* from section 1.1A.

1. Find the mean, $\bar{x} = 53.3333$

STAT ▶ ENTER

```
1-Var Stats
 x̄=53.33333333
 Σx=1920
 Σx²=112064
 Sx=16.61668697
 σx=16.3842743
↓n=36
```

2. Substitute into the variance formula.

$$s_x^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2$$

$$s_x^2 = \frac{(9-53.333)^2 + (29-53.3333)^2 + (32-53.3333)^2 + \ldots + (95-53.3333)^2}{36-1}$$

Use L_2 = "$L_1 - 53.3333$" for deviation $(x_i - \bar{x})$

Use L_3 = "$L_2{}^2$" for square of deviation $(x_i - \bar{x})^2$

L1	L2	L3
9	-44.33	------
29	-24.33	
32	-21.33	
32	-21.33	
32	-21.33	
34	-19.33	
34	-19.33	

L2 ="L1-53.3333"

L1	L2	L3
9	-44.33	1965.4
29	-24.33	592.11
32	-21.33	455.11
32	-21.33	455.11
32	-21.33	455.11
34	-19.33	373.78
34	-19.33	373.78

L3 ="L2²"

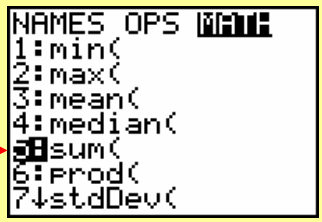
Cursor must be on the column heading before entering formula

" must be used to enter formula — ALPHA — MEM " +

Sum up column L_3 to acquire $\sum(x_i-\bar{x})^2$

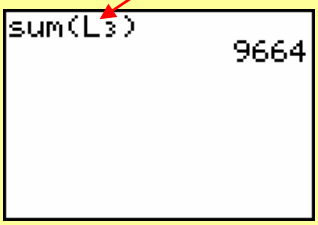
2nd LIST STAT

Select MATH on Menu



Select Option 5

2nd L_3 3



$\sum(x_i-\bar{x})^2 = 9664$

Finally,

$$s_x^2 = \frac{1}{n-1}\sum(x_i-\bar{x})^2$$

$$s_x^2 = \frac{1}{36-1}(9664)$$

$s_x^2 = 276.1143$

For Standard Deviation,

$$s_x = \sqrt{s_x^2}$$

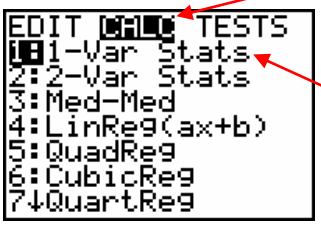
$$s_x = \sqrt{276.1143}$$

$s_x = 16.6167$

To find Standard Deviation using TI-83 Plus Calculator:

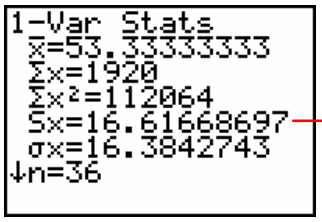
STAT

Select CALC on Menu



Select 1-Var Stats

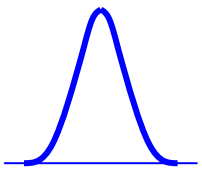
ENTER



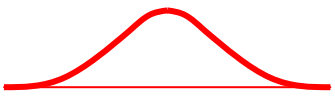
$s_x = 16.6167$

Properties of Standard Deviation

- s_x should only be used when the mean, $\bar{x}$, is decided as the center. This is due to the fact that standard deviation is depended on the mean.
- $s_x = 0$ happens when all scores are the same. There is no deviation.
- s_x is very non-resistant to extreme scores because its calculation is based on the value of the mean.



Small Standard Deviation



Big Standard Deviation

1.2B Assignment: pg. 46–47 #1.35 and 1.37; pg. 50–51 #1.41, 1.43 and 1.45

Chapter 1 Review Assignment
pg. 53–60 #1.49, 1.51, 1.53, 1.55 and 1.59

Chapter 2: The Normal Distribution

2.1A: Density Curves

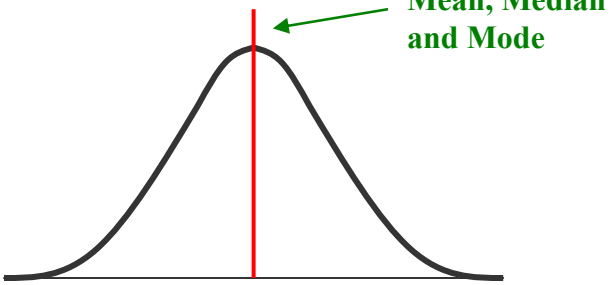
Density Curve: - a smooth curve that replaces a histogram in order to represent a large set of data.
- area under the curve represent proportion of scores or percentage of population.
- entire area of the curve is 1 or 100%

1. **Mean of Density Curve (μ)**: - uses different symbol because it represents a large population.

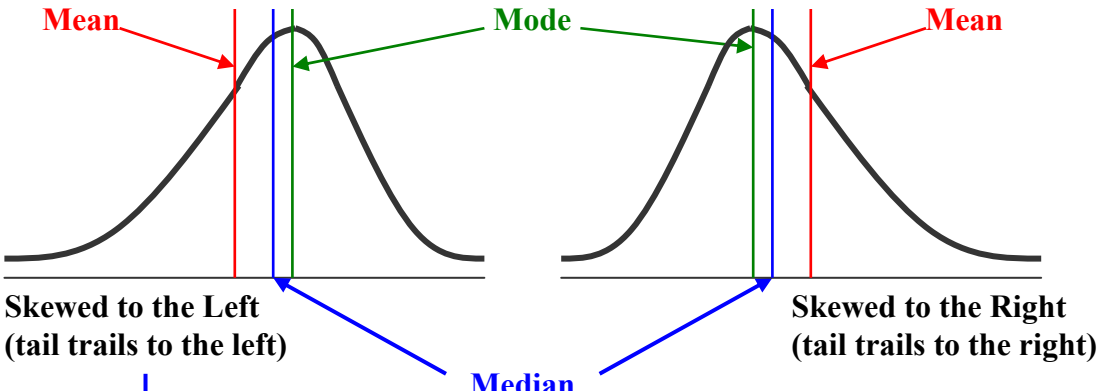
2. **Standard Deviation of Density Curve (σ)**: - again the different symbol denotes a large population.

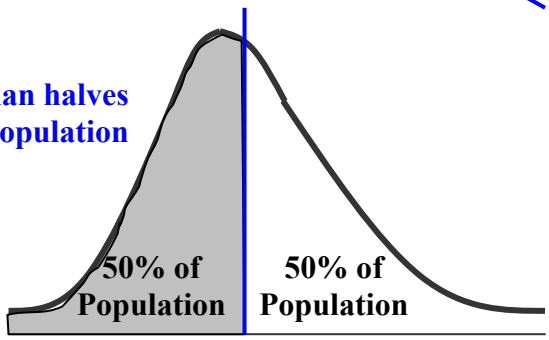
3. **Median of Density Curve (*M*)**: - for symmetrical curves, mean, median and mode is at the same place, in the middle.
- for skewed curves, the median is often closer to the mode (peak) of the curve, whereas the mean is closer to the tail.

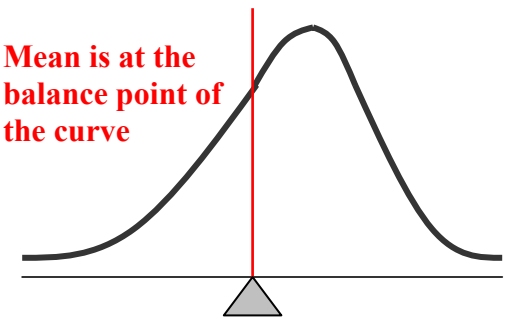
Symmetrical Density Curve



Skewed Density Curve







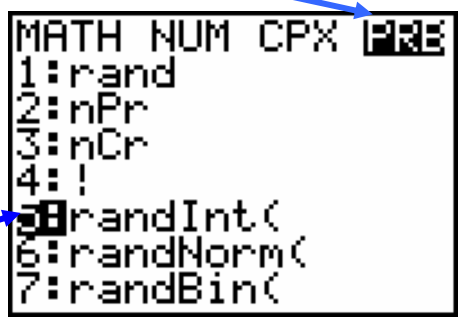
Simulation: - an experimental to generate data in order to compare with theoretical probability.

randInt (Random Integer): generates and displays a random integer within a specified limit and trials.

randInt (Lower Integer Limit, Upper Integer Limit, Number of Trials)

To access randInt:

1. Press **MATH**
2. Use ▶ to access PRB
3. Select Option 5



Example 1: Using the TI-83 Plus, run simulations of rolling a 6-sided dice 120 times, and graph the result.

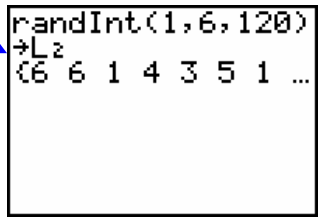
1. randInt(1, 6, 120) and store results in L_2

STO→

2nd

L_2 (2)

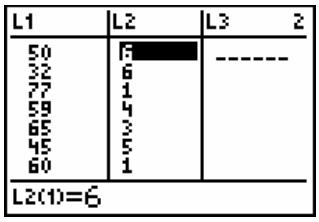
ENTER



2. Check L_2 to see the results

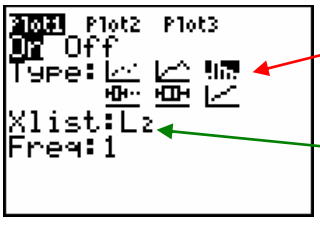
STAT

ENTER



3. Turn ON Stat Plot to Histogram

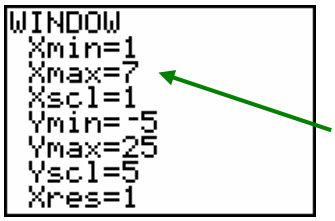
2nd **STAT PLOT** (Y =) **ENTER**



Select Histogram

2nd **L_2** (2)

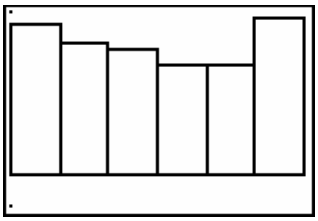
4. Set Windows **WINDOW**



x: [1, 7, 1]
y: [–5, 25, 5]

x_{max} must be one more than the upper integer limit

5. Graph the Histogram **GRAPH**

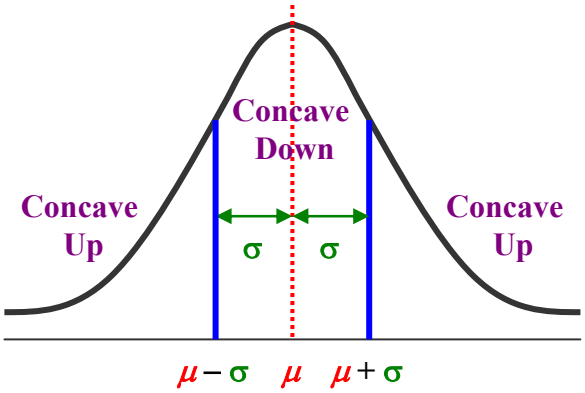


2.1A Assignment: pg. 71–72 #2.1, 2.4 and 2.5

2.1B: Normal Distribution

Normal Distribution (Bell Curve): - a symmetrical density curve that has been normalized for standard use and exhibits the following characteristics.

1. The distribution has a mean (μ) and a standard deviation (σ).
2. The curve is symmetrical about the mean, median and the mode.
3. The standard deviations (σ) are at the inflection points on either side of the curve.
4. Most of the data is within ±3 standard deviation of the mean.
5. The area under the curve represents probability. The total area under the entire curve is 1 or 100%.
6. The probability under the curve follows the 68-95-99.7 Rule.
7. The curve gets really close to the *x*-axis, but never touches it.



Common Usage in Statistics due to:

1. Good Description for Real data.
2. Good Approximation of Chance Outcome (Experimental Simulations match Theoretical Calculation).
3. Its Statistical Inference works well for distributions that are roughly symmetrical.

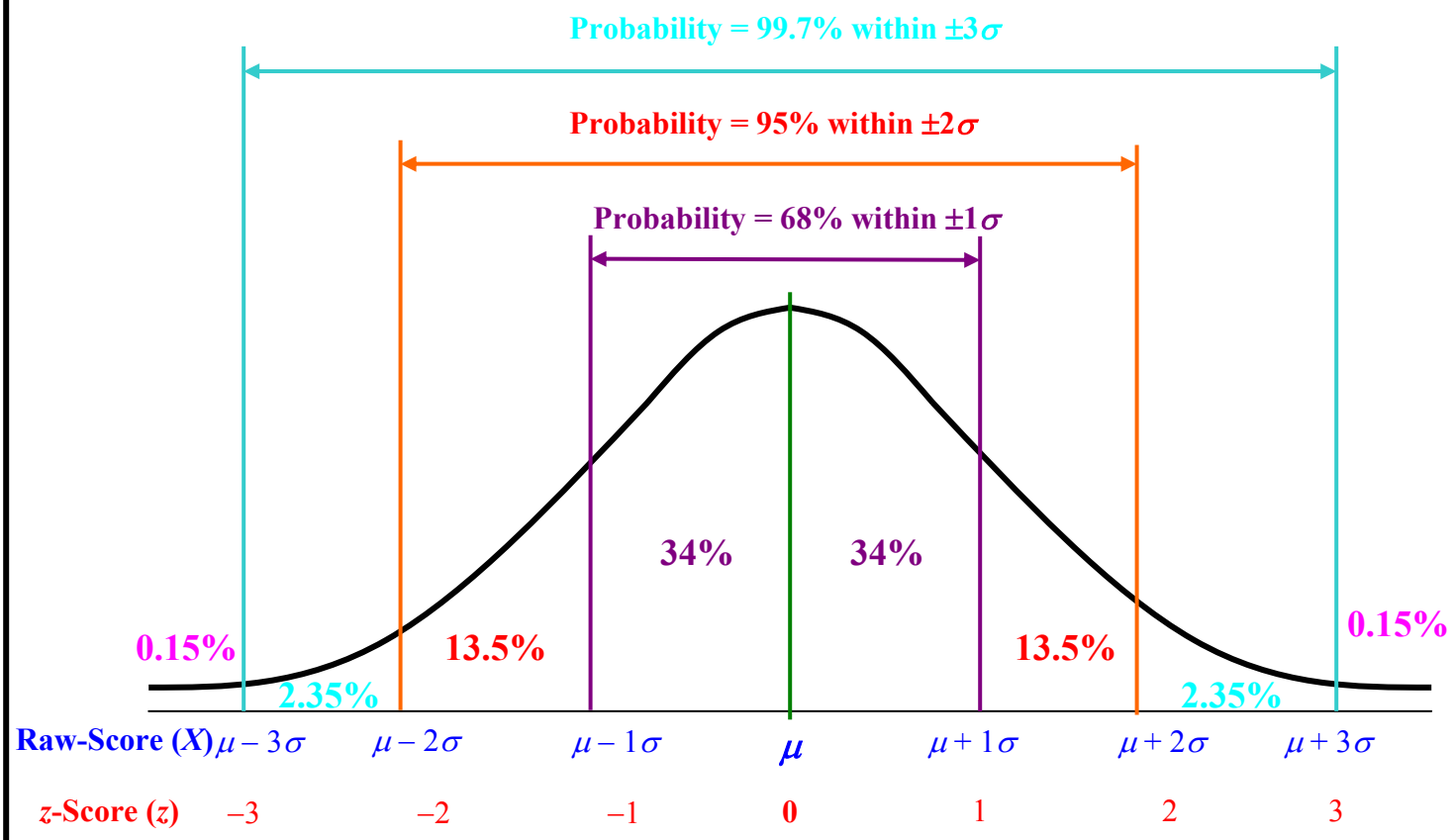
Percentile: - percentage of population **(Area under the Curve)**.

Example: Within the top 15 percentile means 15% of the population scored above you and 85% of the population scored below you.

Raw-Scores (*X*): - the scores as they appear on the original data list.

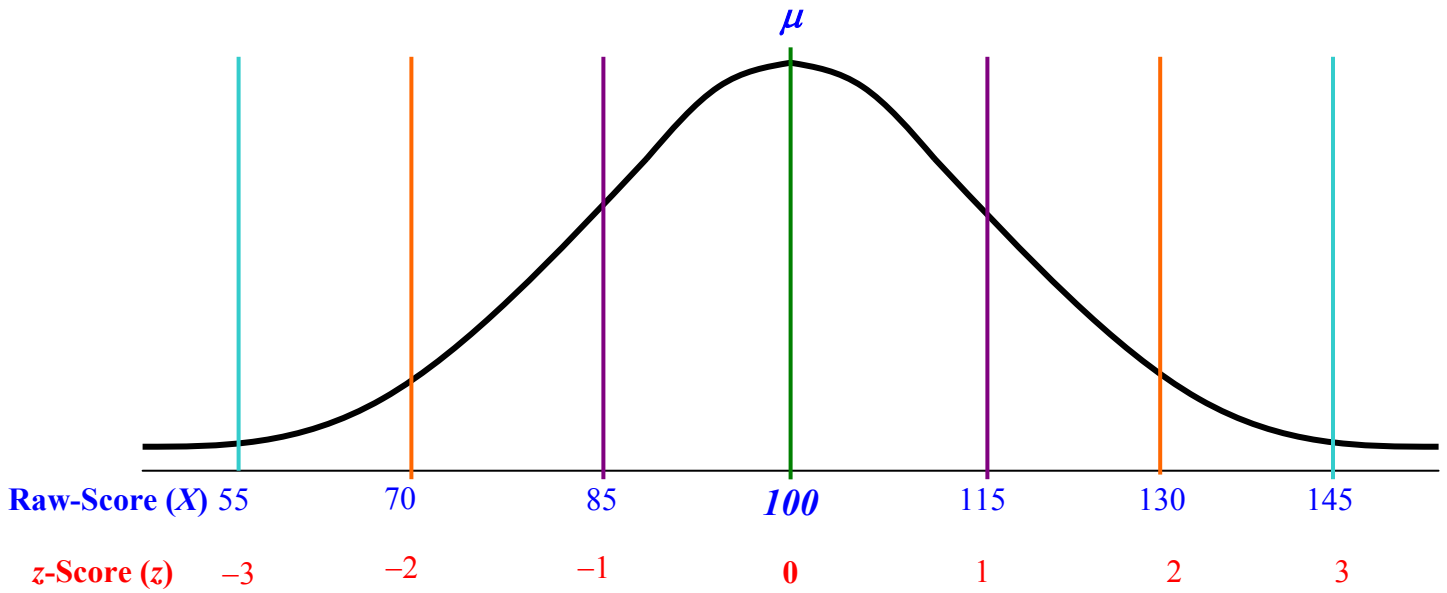
***z*-score (*z*)**: - the number of standard deviation a particular score is away from the mean in a normal distribution.

The 68-95-99.7 Rule of the Area under the Normal Distribution Curve

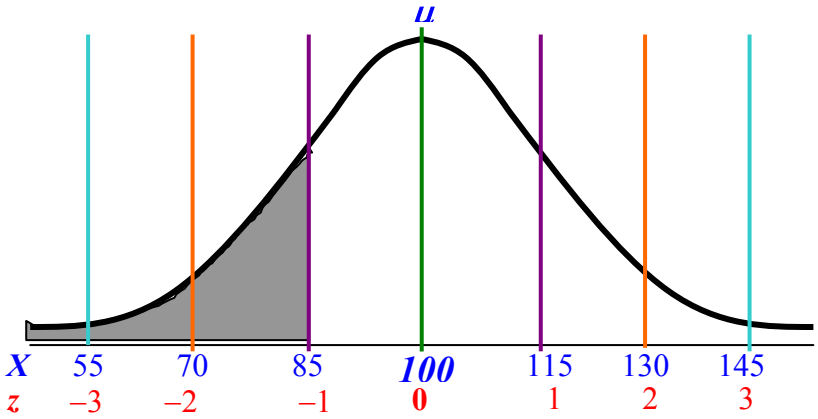


Example 1: The standard IQ test has a mean of 100 and a standard deviation of 15.

a. Draw the normal distribution curve for the standard IQ test.



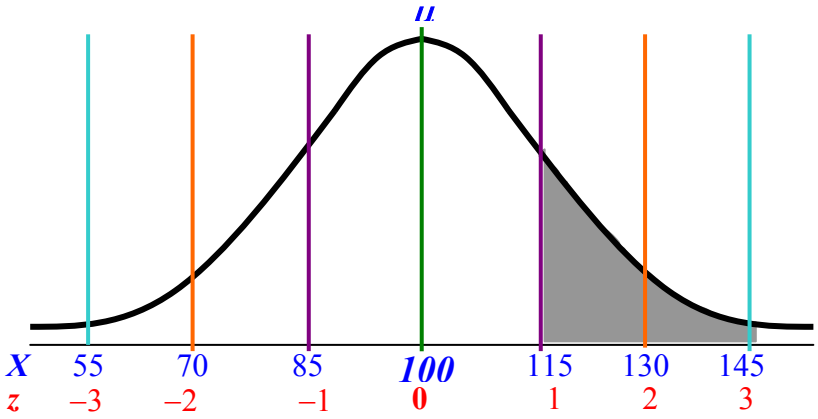
b. What is the probability that a randomly selected person will have an IQ score of 85 and below?



$P(X \le 85) = P(z \le -1)$
$= 13.5\% + 2.35\% + 0.15\%$
$= 16\%$

$P(X \le 85) = 0.16$

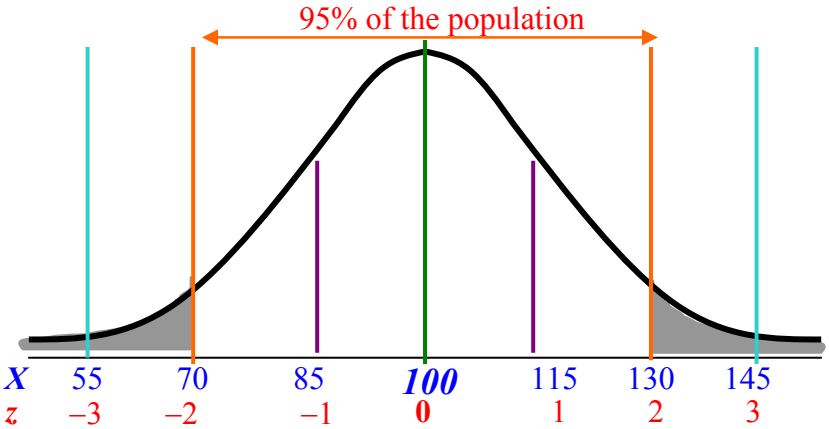
c. What is the probability that a randomly selected person will have an IQ score between 115 to 145?



$P(115 \le X \le 145) = P(1 \le z \le 3)$
$= 13.5\% + 2.35\%$
$= 15.85\%$

$P(115 \le X \le 145) = 0.1585$

d. Find the percentage of the population who has an IQ test score outside of the 2 standard deviations of the mean. Determine the range of the IQ test scores.



$P(z \le -2 \text{ and } z \ge 2) = 100\% - 95\%$

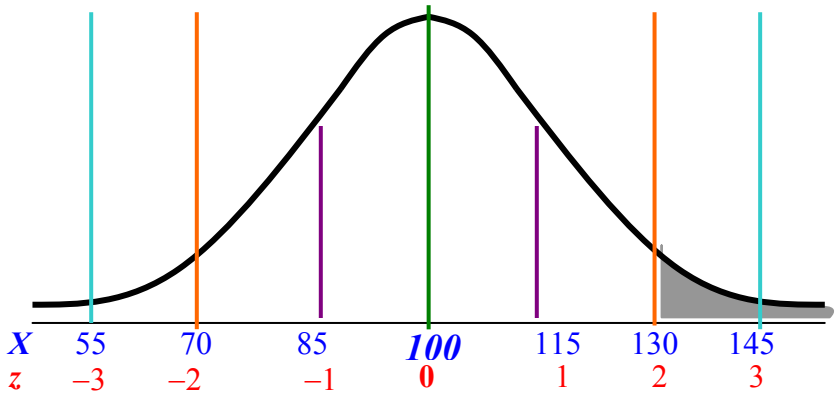
$P(z \le -2 \text{ and } z \ge 2) = 5\%$

From the bell curve, we can see that the ranges are

$P(z \le -2 \text{ and } z \ge 2) = P(X \le 70 \text{ and } X \ge 130)$

The ranges of IQ test scores are
$X \le 70$ and $X \ge 130$

e. In a school of 1500 students, how many students should have an IQ test score above 130?



$P(X > 130) = P(z > 2)$
$= 2.35\% + 0.15\%$
$P(X > 130) = 2.5\%$

Number of students with IQ score > 130 = Total × Probability
$= 1500 \times 0.025$
$= 37.5$

Number of students with IQ score > 130 = 38 students

2.1B Assignment
pg. 73–77 #2.7, 2.8 and 2.9
pg. 80 # 2.14

2.2A: The Standard Normal Distribution

The 68-95-99.7 Rule in the previous section provides an approximate value to the probability of the normal distribution (area under the bell-curve) for 1, 2, and 3 standard deviations from the mean. For *z*-scores other than ±1, 2, and 3, we can use a variety of ways to determine the probability under the normal distribution curve from the raw-score (*X*) and vice versa.

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

where μ **= mean,** σ **= standard deviation,** X **= Raw-Score,** z **=** z**-Score**

Normal Distribution Notation $N(\mu, \sigma)$

Example 1: To the nearest hundredth, find the *z*-score of the followings.

a. $X = 52$, $N(41, 6.4)$

$$z = \frac{X - \mu}{\sigma} \qquad z = \frac{52 - 41}{6.4} = \frac{11}{6.4}$$

$z = 1.72$

b. $X = 75$, $N(82, 9.1)$

$$z = \frac{X - \mu}{\sigma} \qquad z = \frac{75 - 82}{9.1} = \frac{-7}{9.1}$$

$z = -0.77$

Example 2: To the nearest tenth, find the raw-score of the followings.

a. $z = 1.34$, $N(16.2, 3.8)$

$$z = \frac{X - \mu}{\sigma}$$
$$1.34 = \frac{X - 16.2}{3.8}$$
$$(1.34)(3.8) = X - 16.2$$
$$5.092 = X - 16.2$$
$$5.092 + 16.2 = X$$

$X = 21.3$

b. $z = -1.85$, $N(65, 12.7)$

$$z = \frac{X - \mu}{\sigma}$$
$$-1.85 = \frac{X - 65}{12.7}$$
$$(-1.85)(12.7) = X - 65$$
$$-23.495 = X - 65$$
$$-23.495 + 65 = X$$

$X = 41.5$

Example 3: Find the unknown mean or standard deviation to the nearest tenth.

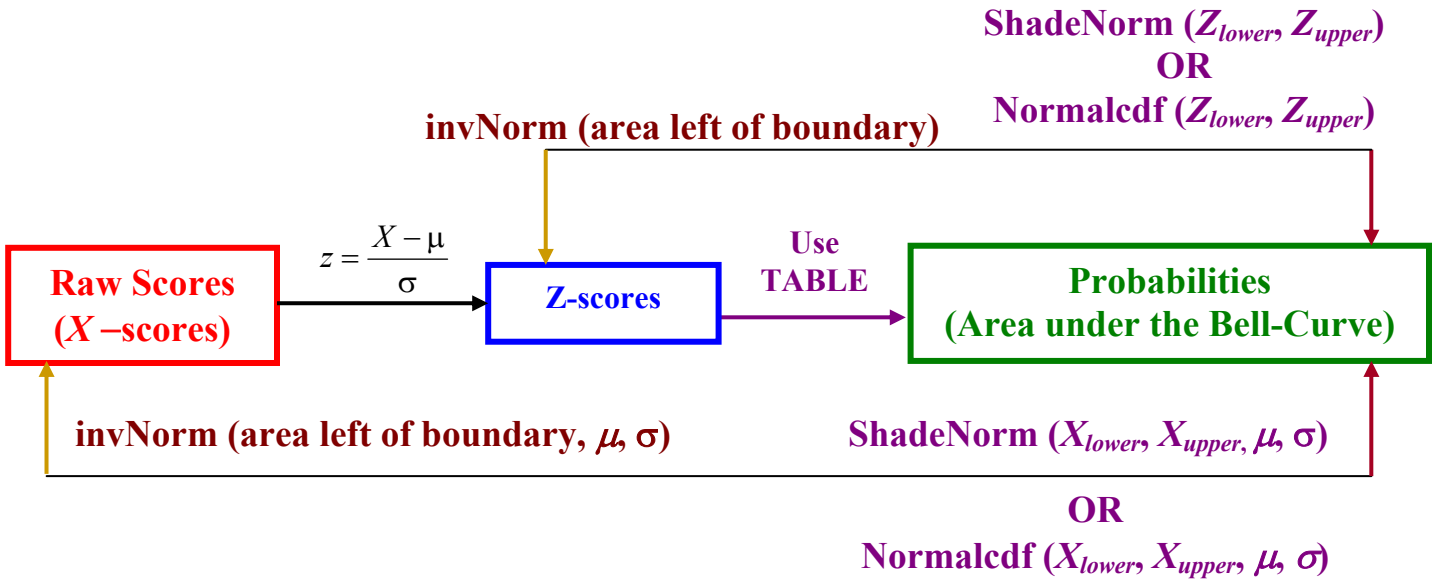
a. $z = -2.33$, $X = 47$, and $N(84, \sigma)$

$$z = \frac{X - \mu}{\sigma}$$
$$-2.33 = \frac{47 - 84}{\sigma}$$
$$\sigma = \frac{47 - 84}{-2.33}$$
$$\sigma = \frac{-37}{-2.33}$$
$$\sigma = 15.9$$

b $z = 1.78$, $X = 38$, and $N(\mu, 8.2)$

$$z = \frac{X - \mu}{\sigma}$$
$$1.78 = \frac{38 - \mu}{8.2}$$
$$(1.78)(8.2) = 38 - \mu$$
$$14.596 = 38 - \mu$$
$$\mu = 38 - 14.596$$
$$\mu = 23.4$$

Normal Distribution Curve Summary



Normalcdf (X_{lower}, X_{upper}, μ, σ) : - use to convert Raw-Score directly to probability with NO graphics.

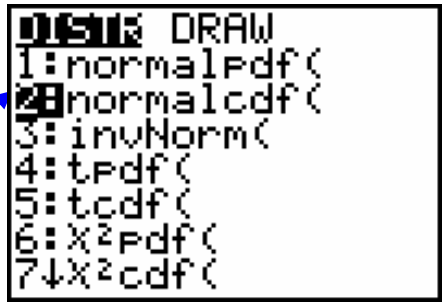
Normalcdf (Z_{lower}, Z_{upper}) : - use to convert *z*-Score to probability with NO graphics

- if X_{lower} or Z_{lower} is at the very left edge of the curve and is not obvious, use -1×10^{99} (–1E99 on calculator).
- if X_{upper} or Z_{upper} is at the very right edge of the curve and is not obvious, use 1×10^{99} (1E99 on calculator).

To access normalcdf:

1. Press **2nd** **DISTR** **VARS**

2. Select Option 2



ShadeNorm (X_{lower}, X_{upper}, μ, σ) : - use to convert Raw-Score directly to probability with graphics.

ShadeNorm (Z_{lower}, Z_{upper}) : - use to convert *z*-Score to probability with graphics

- if X_{lower} or Z_{lower} is at the very left edge of the curve and is not obvious, use -1×10^{99} (–1E99 on calculator).
- if X_{upper} or Z_{upper} is at the very right edge of the curve and is not obvious, use 1×10^{99} (1E99 on calculator).

Before accessing ShadeNorm, we need to select the *WINDOW* setting.

For **ShadeNorm (X_{lower}, X_{upper}, μ, σ)**, select a reasonable setting based on the information provided.

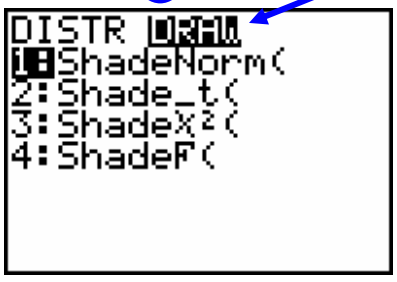
For **ShadeNorm (Z_{lower}, Z_{upper})**, use *x*: [–5, 5, 1] and *y*: [–0.15, 0.5, 0].

To access ShadeNorm:

1. Press **2nd** **DISTR** **VARS**

2. Use ▶ **to access DRAW**

3. Select Option 1

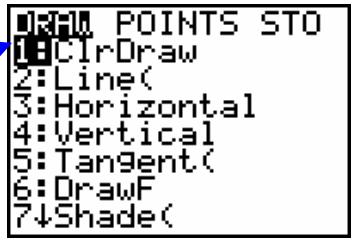


Must Clear the Drawing (ClrDraw) before drawing or graphing again!

To access ClrDraw:

1. Press **2nd** **DRAW** **PRGM**

2. Select Option 1



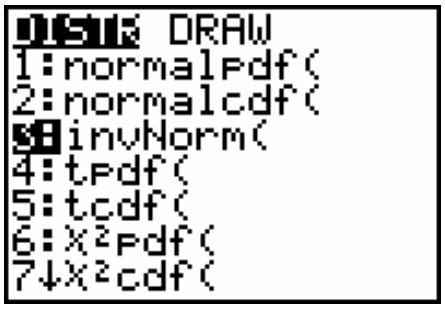
invNorm (area left of boundary, μ, σ) : - use to convert Area under the curve (Probability directly back to Raw-Score with NO graphics.

invNorm (area left of boundary) : - use to convert Area under the curve (Probability) back to *z*-Score with NO graphics.

To access invNorm:

1. Press **2nd** **DISTR** **VARS**

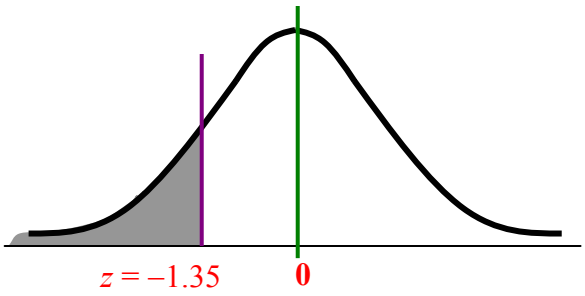
2. Select Option 3

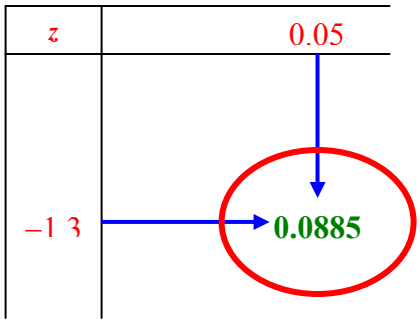


Using the Table of Areas under the Standard Normal Curve (Complete table on the next 2 pages)

1. **Converting *z*-score to Area under the curve (LEFT of the *z*-score boundary)**
   - **a.** Look up the *z*-score from the column and row headings.
   - **b.** Follow that row and column to find the area.

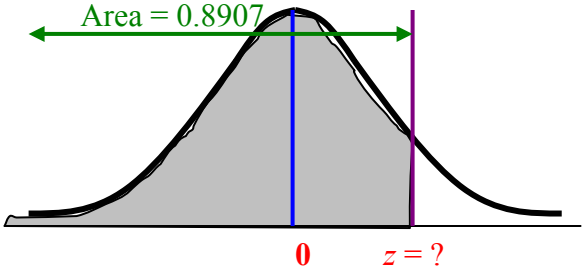
Example: Find $P(z \le -1.35)$.

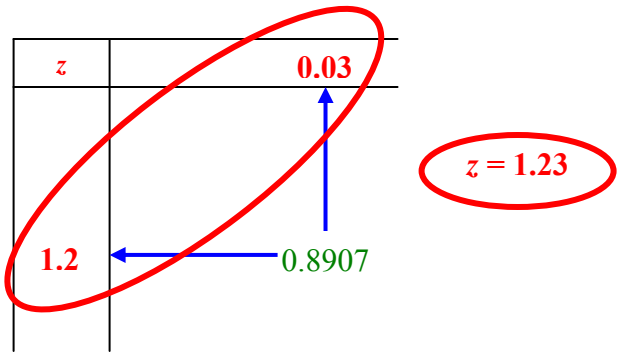




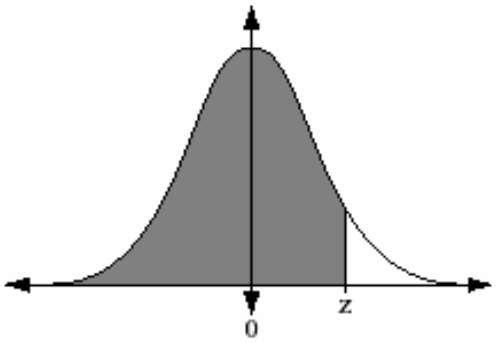
2. **Converting Area under the curve back to *z*-score (LEFT of the *z*-score boundary)**
   - **a.** Look up the Area LEFT of the boundary from INSIDE the table.
   - **b.** Follow that row and column back to the heading and locate the corresponding *z*-score.

Example: $P(z \le ?) = 0.8907$





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



Areas under the Standard Normal Curve

z	0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.00
−3.4	0.0002	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003
−3.3	0.0003	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0005	0.0005	0.0005
−3.2	0.0005	0.0005	0.0005	0.0006	0.0006	0.0006	0.0006	0.0006	0.0007	0.0007
−3.1	0.0007	0.0007	0.0008	0.0008	0.0008	0.0008	0.0009	0.0009	0.0009	0.0010
−3.0	0.0010	0.0010	0.0011	0.0011	0.0011	0.0012	0.0012	0.0013	0.0013	0.0013
−2.9	0.0014	0.0014	0.0015	0.0015	0.0016	0.0016	0.0017	0.0018	0.0018	0.0019
−2.8	0.0019	0.0020	0.0021	0.0021	0.0022	0.0023	0.0023	0.0024	0.0025	0.0026
−2.7	0.0026	0.0027	0.0028	0.0029	0.0030	0.0031	0.0032	0.0033	0.0034	0.0035
−2.6	0.0036	0.0037	0.0038	0.0039	0.0040	0.0041	0.0043	0.0044	0.0045	0.0047
−2.5	0.0048	0.0049	0.0051	0.0052	0.0054	0.0055	0.0057	0.0059	0.0060	0.0062
−2.4	0.0064	0.0066	0.0068	0.0069	0.0071	0.0073	0.0075	0.0078	0.0080	0.0082
−2.3	0.0084	0.0087	0.0089	0.0091	0.0094	0.0096	0.0099	0.0102	0.0104	0.0107
−2.2	0.0110	0.0113	0.0116	0.0119	0.0122	0.0125	0.0129	0.0132	0.0136	0.0139
−2.1	0.0143	0.0146	0.0150	0.0154	0.0158	0.0162	0.0166	0.0170	0.0174	0.0179
−2.0	0.0183	0.0188	0.0192	0.0197	0.0202	0.0207	0.0212	0.0217	0.0222	0.0228
−1.9	0.0233	0.0239	0.0244	0.0250	0.0256	0.0262	0.0268	0.0274	0.0281	0.0287
−1.8	0.0294	0.0301	0.0307	0.0314	0.0322	0.0329	0.0336	0.0344	0.0351	0.0359
−1.7	0.0367	0.0375	0.0384	0.0392	0.0401	0.0409	0.0418	0.0427	0.0436	0.0446
−1.6	0.0455	0.0465	0.0475	0.0485	0.0495	0.0505	0.0516	0.0526	0.0537	0.0548
−1.5	0.0559	0.0571	0.0582	0.0594	0.0606	0.0618	0.0630	0.0643	0.0655	0.0668
−1.4	0.0681	0.0694	0.0708	0.0721	0.0735	0.0749	0.0764	0.0778	0.0793	0.0808
−1.3	0.0823	0.0838	0.0853	0.0869	0.0885	0.0901	0.0918	0.0934	0.0951	0.0968
−1.2	0.0985	0.1003	0.1020	0.1038	0.1056	0.1075	0.1093	0.1112	0.1131	0.1151
−1.1	0.1170	0.1190	0.1210	0.1230	0.1251	0.1271	0.1292	0.1314	0.1335	0.1357
−1.0	0.1379	0.1401	0.1423	0.1446	0.1469	0.1492	0.1515	0.1539	0.1562	0.1587
−0.9	0.1611	0.1635	0.1660	0.1685	0.1711	0.1736	0.1762	0.1788	0.1814	0.1841
−0.8	0.1867	0.1894	0.1922	0.1949	0.1977	0.2005	0.2033	0.2061	0.2090	0.2119
−0.7	0.2148	0.2177	0.2206	0.2236	0.2266	0.2296	0.2327	0.2358	0.2389	0.2420
−0.6	0.2451	0.2483	0.2514	0.2546	0.2578	0.2611	0.2643	0.2676	0.2709	0.2743
−0.5	0.2776	0.2810	0.2843	0.2877	0.2912	0.2946	0.2981	0.3015	0.3050	0.3085
−0.4	0.3121	0.3156	0.3192	0.3228	0.3264	0.3300	0.3336	0.3372	0.3409	0.3446
−0.3	0.3483	0.3520	0.3557	0.3594	0.3632	0.3669	0.3707	0.3745	0.3783	0.3821
−0.2	0.3859	0.3897	0.3936	0.3974	0.4013	0.4052	0.4090	0.4129	0.4168	0.4207
−0.1	0.4247	0.4286	0.4325	0.4364	0.4404	0.4443	0.4483	0.4522	0.4562	0.4602
−0.0	0.4641	0.4681	0.4721	0.4761	0.4801	0.4840	0.4880	0.4920	0.4960	0.5000

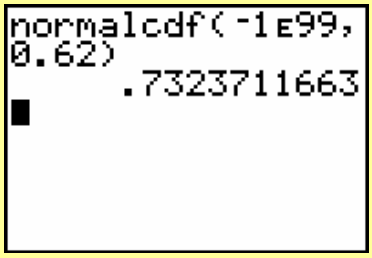
Areas under the Standard Normal Curve

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.5	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Example 4: To the nearest hundredth of a percent, find the probability of the following.

a. $P(z < 0.62)$

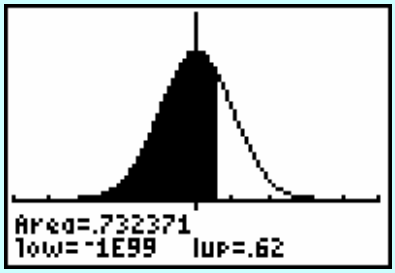
normalcdf $(-1 \times 10^{99}, 0.62)$



$P(z < 0.62) = 73.24\%$

OR

ShadeNorm $(-1 \times 10^{99}, 0.62)$

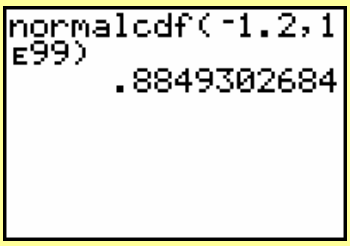


x: [–5, 5, 1]
y: [–0.15, 0.5, 0]

$P(z < 0.62) = 73.24\%$

b. $P(z > -1.2)$

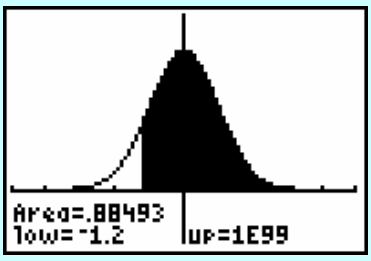
normalcdf $(-1.2, 1 \times 10^{99})$



$P(z > -1.2) = 88.49\%$

OR

ShadeNorm $(-1.2, 1 \times 10^{99})$

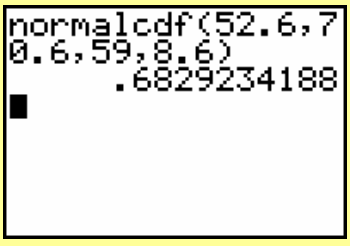


x: [–5, 5, 1]
y: [–0.15, 0.5, 0]

$P(z > -1.2) = 88.49\%$

c. $P(52.6 < X < 70.6)$ given $N(59, 8.6)$

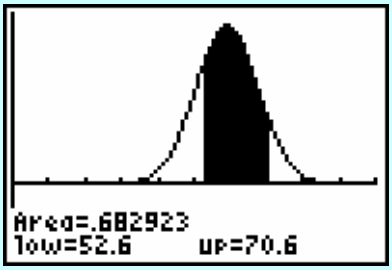
normalcdf (52.6, 70.6, 59, 8.6)



$P(52.6 < X < 70.6) = 68.29\%$

OR

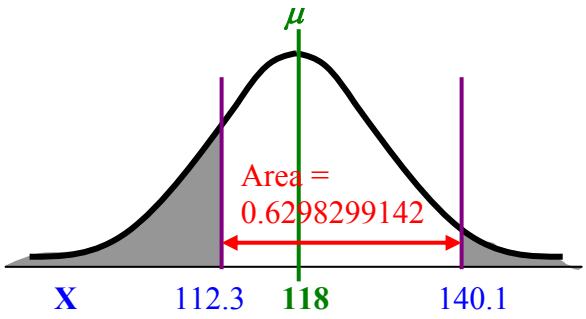
ShadeNorm (52.6, 70.6, 59, 8.6)



x: [0, 100, 10]
y: [–0.02, 0.05, 0]

$P(52.6 < X < 70.6) = 68.29\%$

d. $P(X < 112.3 \text{ and } X > 140.1)$ given $N(118, 12.8)$



normalcdf (112.3, 140.1, 118, 12.8)

```
normalcdf(112.3,
140.1,118,12.8)
      .6298299142
```

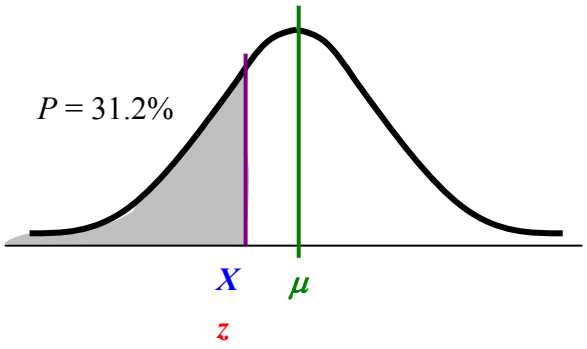
$P(112.3 < X < 140.1) = 0.6298299142$

$P(X < 112.3 \text{ and } X > 140.1) = 1 - 0.6298299142$

$P(X < 112.3 \text{ and } X > 140.1) = 37.0\%$

Example 5: To the nearest hundredth, find the z-score and the raw-score from the following probability.

a. $N(28.9, 3.28)$

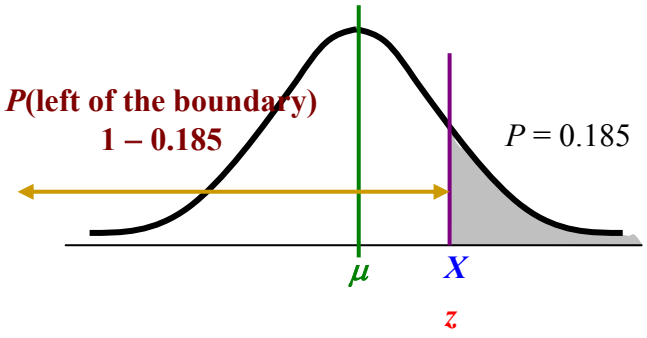


z = invNorm (0.312)
X = invNorm (0.312, 28.9, 3.28)

```
invNorm(0.312)
     -.4901892319
invNorm(0.312,28
.9,3.28)
      27.29217932
```

$z = -0.49$ and $X = 27.29$

b. $N(82.1, 7.42)$



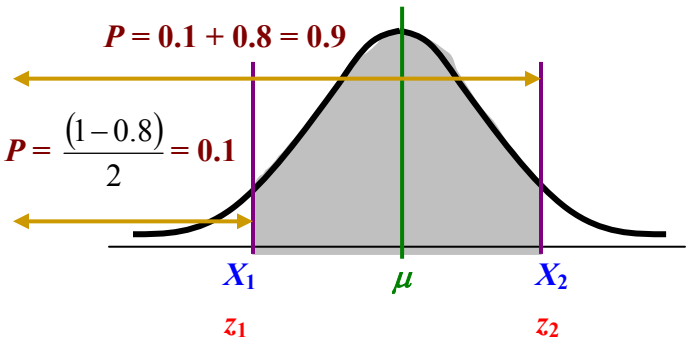
z = invNorm (1 – 0.185)
X = invNorm (1 – 0.185, 82.1, 7.42)

```
invNorm(1-0.185)
      .8964733583
invNorm(1-0.185,
82.1,7.42)
      88.75183232
```

$z = 0.90$ and $X = 88.75$

c. $N(185, 11.3)$

P = 80% symmetrical about the mean



z_1 = invNorm (0.1)
X_1 = invNorm (0.1, 185, 11.3)

```
invNorm(0.1)
     -1.281551567
invNorm(0.1,185,
11.3)
      170.5184673
```

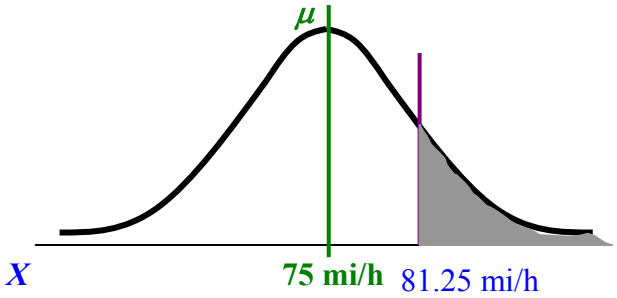
$z_1 = -1.28$ and $X_1 = 170.52$

z_2 = invNorm (0.9) X_2 = invNorm (0.1, 185, 11.3)

$z_2 = 1.28$ and $X_2 = 199.48$

```
invNorm(0.9)
      1.281551567
invNorm(0.9,185,
11.3)
      199.4815327
```

Example 6: There are approximately 12 000 vehicles travelling on a section of Hwy 101 during a non-rush hour everyday. The average speed of these vehicles is 75 miles/hr with a standard deviation of 12 miles/hr. If the posted speed limit on Hwy 101 is 65 miles/hr and the police will pull people over when they are 25% above the speed limit, how many people will the police pull over on any given day?



25% above 65 mi/h = 65 × 125% = 81.25 mi/h

X = 81.25 mi/h, μ = 75 mi/h, σ = 12 mi/h

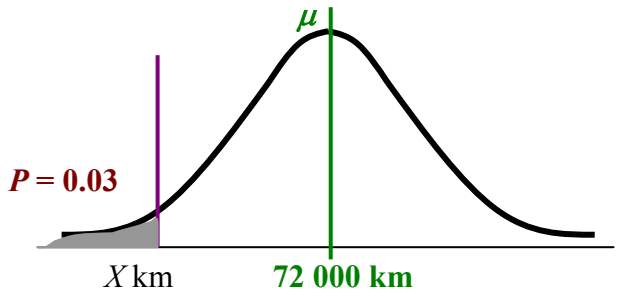
$P(X \geq 80.5$ mi/h) = **normalcdf (81.25, 1 × 10^{99}, 75, 12)**
= 0.3012414225

```
normalcdf(81.25,
1E99,75,12)
      .3012414225
12000*Ans
      3614.89707
```

Number of drivers pulled over = Total × Probability
= 12 000 × 0.3012414225

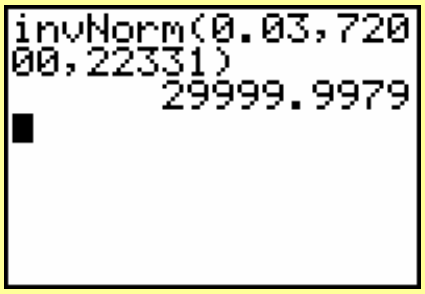
3615 drivers can be pulled over by the police

Example 7: A tire manufacturer finds that the mean life of the tires produced is 72000 km with a standard deviation of 22331 km. To the nearest kilometre, what should the manufacturer's warranty be set at if it can only accept a return rate of 3% of all tires sold?



$P = 0.03$, $\mu = 72000$ km, $\sigma = 22331$ km

X = invNorm(0.03, 720000, 22331)



The warranty should be set at 30000 km

2.2A Assignment

pg. 92 #2.22, 2.24 and 2.25

Worksheet: The Standard Normal Distribution Curve

2.2A Worksheet: The Standard Normal Distribution Curve

1. The mean life expectancy of elephants in a certain protected wilderness area is estimated to be 66 years, with a standard deviation of 4.5 years. In a herd of 255 elephants in this area, how many are expected to live longer than 75 years?

2. The masses of 800 female athletes were measured and found to be normally distributed. The mean mass of these athletes was 55 kg, with a standard deviation of 5 kg.
   a. How many of these athletes had masses between 45 kg and 65 kg?
   b. How many of these athletes had masses greater than 65 kg?
   c. How many of these athletes had masses less than 40 kg?

3. Calculate the percentage of population for the following z-scores.
   a. P ($z < 1.8$) b. P ($z < -2.5$) c. P ($z < 0.7$) d. P ($z < -0.5$)
   e. P ($1.5 < z < 2.3$) f. P ($-1.5 < z < 2.5$) g. P ($-1.5 < z < -0.5$) h. P ($-0.5 < z < 0.5$)

4. A population is normally distributed with a mean of 25.8 and a standard deviation of 1.5. What is the probability that a randomly selected member of population will have the following measures?
   a. greater than 27 b. greater than 28 c. greater than 25 d. greater than 22

5. The mean monthly attendance at a sports arena is 8450, with a standard deviation of 425.
   a. What is the probability that the monthly attendance will be less than 8000?
   b. What is the probability that the monthly attendance will be more than 9000?

6. In a sample of 10 000 Florida oranges, the mean is 985 g and the standard deviation of 52 g.
   a. What percent of the oranges have a mass between 900 g and 1000 g?
   b. How many are expected to have a mass between 900 g and 1000 g?
   c. How many are expected to have a mass less than 1000 g?

7. The mean score on Test A was 65% with a standard deviation of 5%. The mean score on Test B was 63% with a standard deviation of 6%. Stefan achieved 70% on Test A and Renee achieved 68% on Test B. Which student gave the better performance?

8. Major manufacturing companies operate on the principle of preventative maintenance to avoid a complete shutdown of the assembly line if a component fails. The lifetime of one component is normally distributed with a mean of 321 hours and a standard deviation of 23 hour. How frequently should the component be replaced so that the probability of its failing during operation is less than 0.001?

Answers:

1. 6	2a. 760	2b. 20	2c. 1	3a. 0.964 069
3b. 0.006 209	3c. 0.758 036	3d. 0.308 538	3e. 0.056 083	3f. 0.926 983
3g. 0.241 73	3h. 0.382 925	4a. 0.2119	4b. 0.0712	4c. 0.7031
4d. 0.9944	5a. 0.1448	5b. 0.0978	6. 56.24%	7. Stefan
8. 250 hours				

2.2B: Assessing Normality

Normality: - a measure of how well a set of scores fit the standard normal distribution curve.

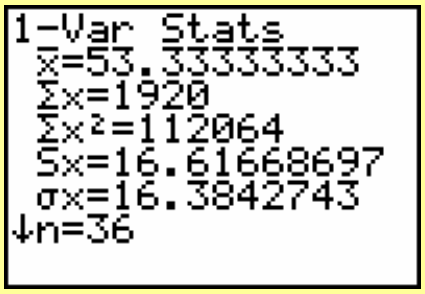
Example 1: Using the data below, assess the normality of the *Final Marks in Percentage for an Algebra II Class in Suburbia Public School*.

Final Marks in Percentage for an Algebra II Class in Suburbia Public School

Student #	Final Mark	Student #	Final Mark	Student #	Final Mark	Student #	Final Mark
1	50	10	56	19	63	28	56
2	32	11	32	20	72	29	9
3	77	12	50	21	67	30	65
4	59	13	29	22	70	31	95
5	65	14	60	23	50	32	34
6	45	15	45	24	68	33	60
7	60	16	43	25	42	34	45
8	55	17	50	26	65	35	61
9	32	18	34	27	50	36	74

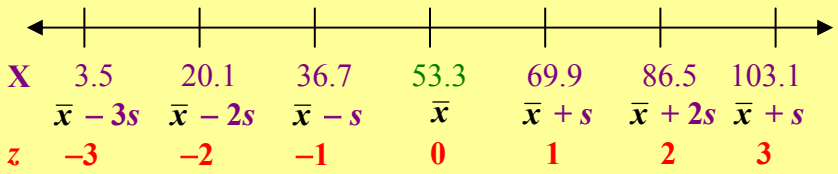
Method 1: Compare Actual Data to Standard Normal Curve's Area Using the 68-95-99.7 Rule

1. Find the mean and standard deviation.

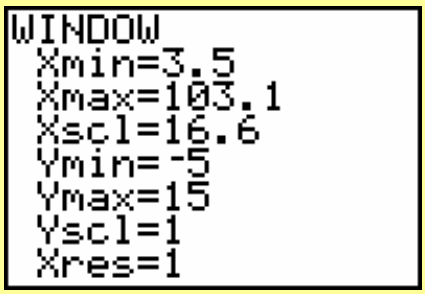


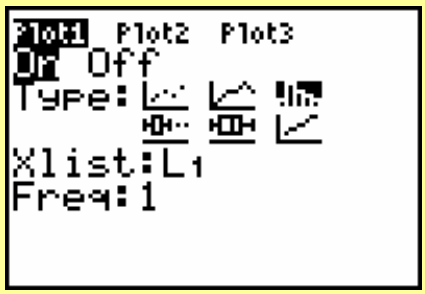
$\bar{x} = 53.3$
$s = 16.6$

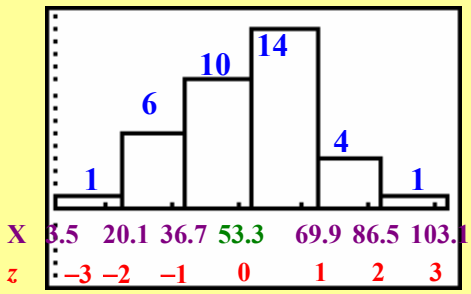
2. Set up the horizontal axis with raw scores corresponding to $-3 \le z \le 3$



3. Set Windows, Graph Histogram, and Trace







x: [3.5, 103.1, 16.6]
y: [−5, 15, 1]

x_{min} = corresponding raw score to $z = -3$
x_{max} = corresponding raw score to $z = 3$
$x_{scl} = s$

4. Calculate Percentage of Population Within each Standard Deviation

Between $-1 \le z \le 1$: $\frac{10+14}{36} \times 100\% = 66.7\%$

Between $-2 \le z \le 2$: $\frac{6+10+14+4}{36} \times 100\% = 94.4\%$

Between $-3 \le z \le 3$: $\frac{1+6+10+14+4+1}{36} \times 100\% = 100\%$

5. Compare to the 68-95-99.7 Rule

The resulting percentages are very close to the 68-95-99.7 Rule (66.7-94.4-100)
Therefore, the scores have a good normality

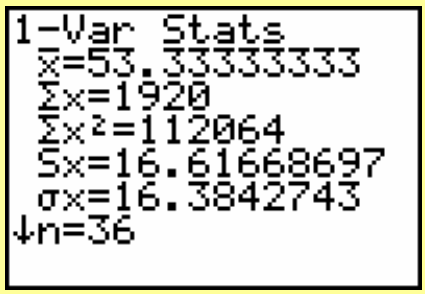
Method 2: Normal Probability Plot

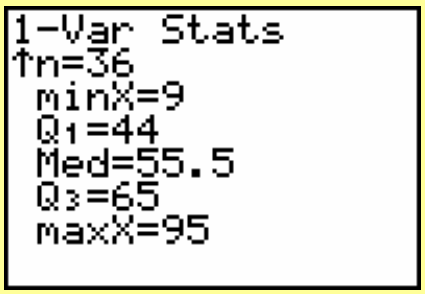
Normal Probability Plot: - compares the individual observed data (x_i) and the z-value of the percentile.

- in a normal distribution curve, z and X form a linear relationship $z = \frac{X-\mu}{\sigma}$.

Therefore, a linear plot shows that the scores are close to normality.

1. Compare Mean ($\bar{x}$) and Median (M) using 1-Var Stats





$\bar{x} = 53.3$ **Mean and Median**
$M = 55.5$ **are fairly close**

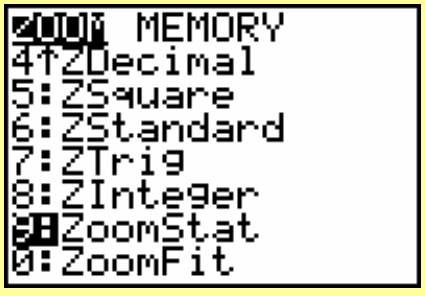
2. Plot Normal Probability Plot

2nd **STAT PLOT** **ENTER** **ZOOM** **ENTER**

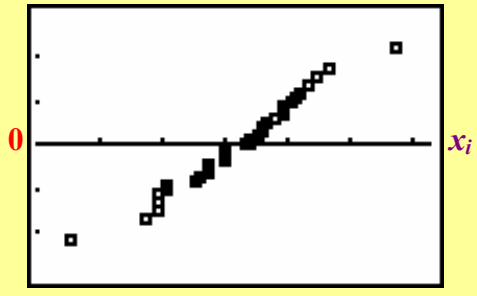
Y =

Use ZoomStat

The plot shows a linear relationship. Therefore, the scores have a good normality.



z-scores



2.2B Assignment
pg. 97–98 #2.27 to 2.30

Chapter 2 Review Assignment
pg. 102–103 #2.39, 2.41, 2.43, 2.45

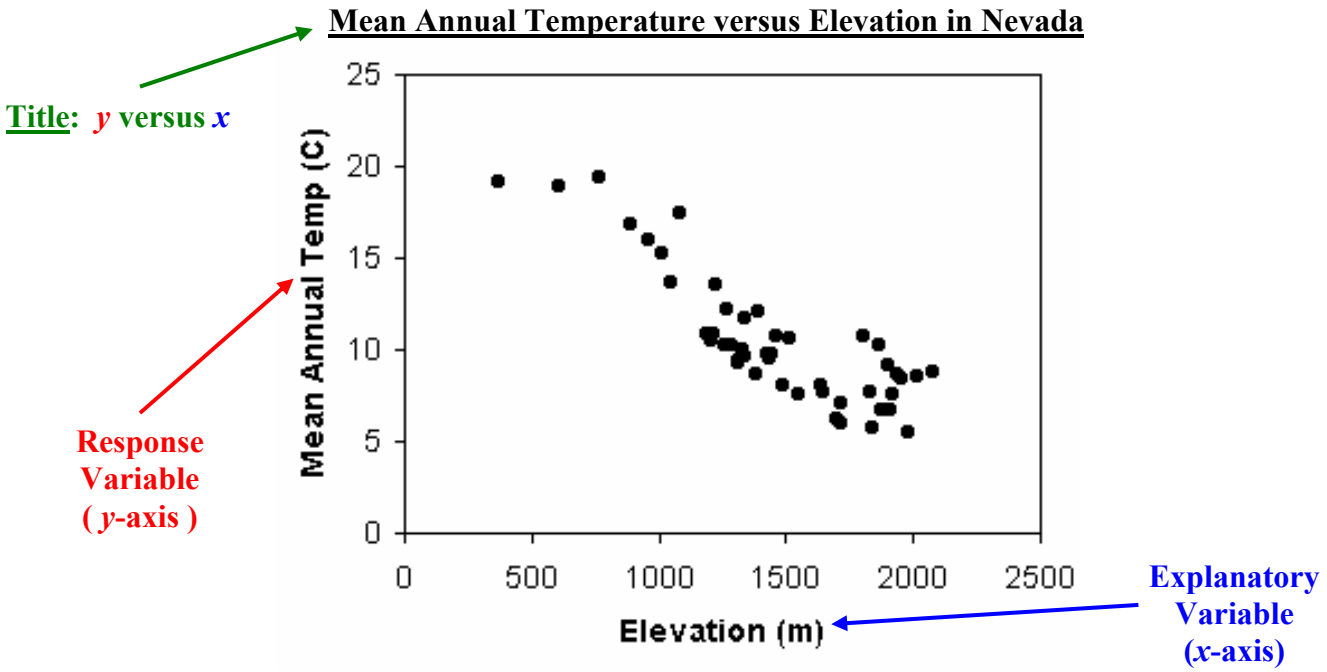
Chapter 3: Examining Relationships

3.1: Scatter Plots

Response Variable: - the dependent variable that measures the outcome of an experiment or a study.
- located on the *y*-axis of a graph.

Explanatory Variable: - the independent variable where the experimenter changes to observe the change in the response variable.
- it is used to explain the observed outcome.
- located on the *x*-axis of a graph.

Scatter Plot: - a plot using dots to show the relationship between an explanatory variable and a response variable.



Interpretation of Scatter Plots

1. **Direction**: - whether the graph is positive or negative associated.
   a. **Positive Association**: - generally as *x* increases, *y* increases. (*x* ↑ *y* ↑)
   b. **Negative Association**: - generally as *x* increases, *y* decreases. (*x* ↑ *y* ↓)

2. **Form**: - identified the locations and the reasons for **clusters** along with the **shape**.
   a. **Cluster**: - a group of data points.
   b. **Shape**: - whether the dots assume a linear or a curve line.

3. **Strength**: - how strong is the relationship between the two variables plotted (as indicated by how close the points in the scatterplot lie to form a line).

4. **Deviation**: - identify any outliers that fall beyond the overall pattern.

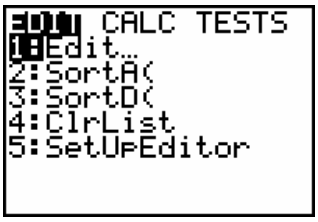
Example 1: Graph the scatterplot between the number of absences and final scores. Describe the scatterplot in terms of its direction, form, strength and deviation.

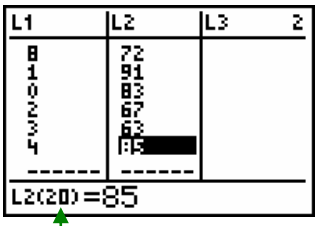
Comparison between Number of Absences and Final Score for an Algebra I Class in Suburbia Public High School

Student ID Number	Number of Absences	Final Marks (%)	Student ID Number	Number of Absences	Final Marks (%)
1	3	75	11	2	52
2	6	67	12	3	65
3	8	51	13	3	88
4	1	88	14	4	67
5	2	80	15	8	72
6	4	78	16	1	91
7	10	42	17	0	83
8	7	55	18	2	67
9	3	70	19	3	63
10	5	65	20	4	85

Entering Data using TI-83 Plus Calculator:

STAT **ENTER** **Enter Values**





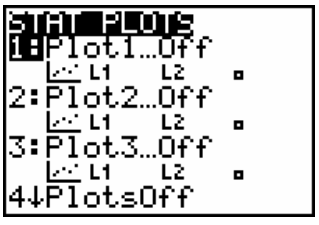
Check if you entered the Correct Number of Data Values

20th Score entered

To plot Scatter Plot on TI-83 Plus Calculator:

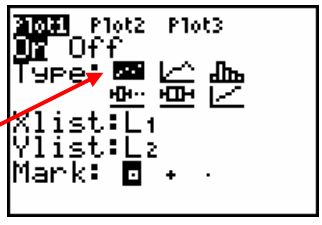
1. Because we are plotting from a set of data, not from an equation, we have to turn on STAT PLOT.

2nd **STAT PLOT** **Y =**



ENTER

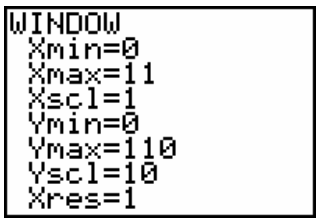
Select Scatter Plot



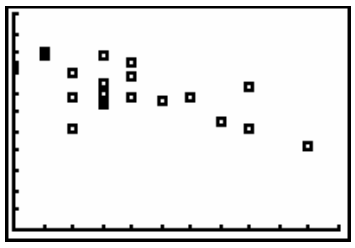
2. Set Window by identifying the Minimum and Maximum Values of Both Variables

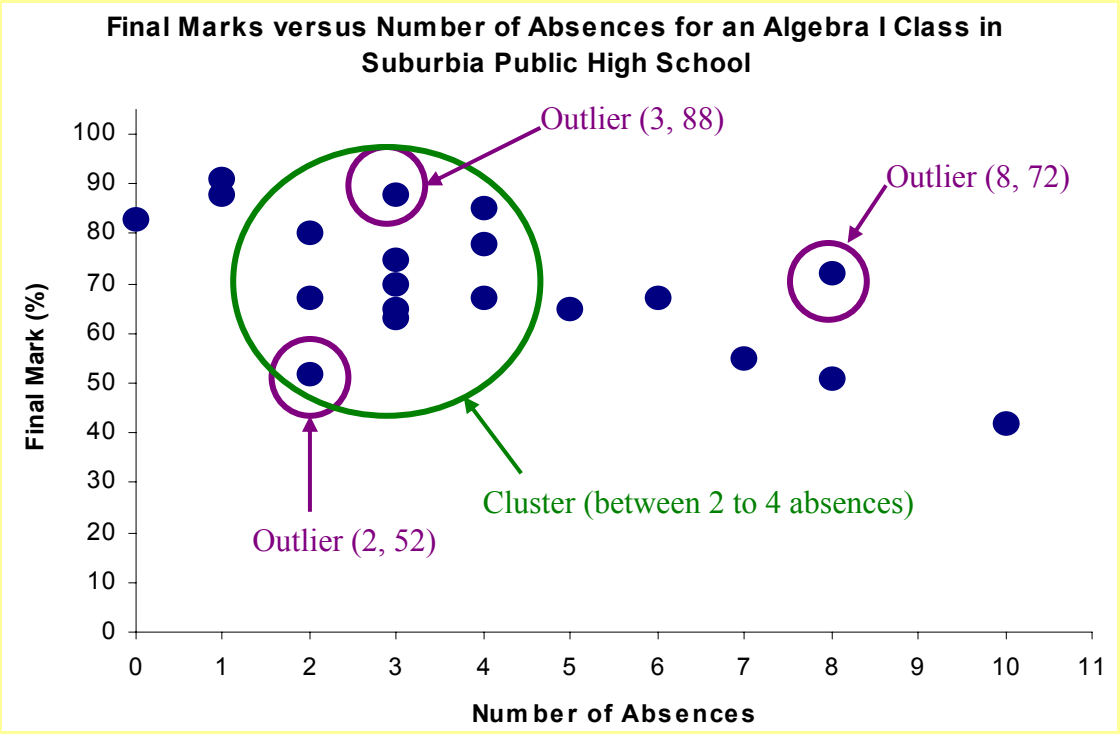
WINDOW

x: [0, 11, 1]
y: [0, 105, 10]



3. Graph Scatter Plot **GRAPH**





Direction: As the number of absences increases, the final mark decreases (Negative Association)

Form: There is a cluster between 2 to 4 absences. (Most students have around 2 to 4 absences). The general shape is linear.

Deviation: There are 3 outliers (2, 52), (3, 88) and (8, 72). They lie outside the general trend of the plot.

Strength: Due to the large amount of outliers, 3 out of 20 data points, there is only a **weak relation** between the two variables.

3.1 Assignment: **pg. 113 #3.5; pg. 115–116 #3.7, 3.8; pg. 120–121 #3.9**

3.2: Correlation

Correlation (***r***): - a measure of the strength and direction of the two variables in a linear relationship.

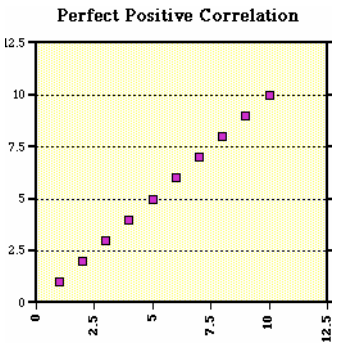
Correlation (*r*)

$$r = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

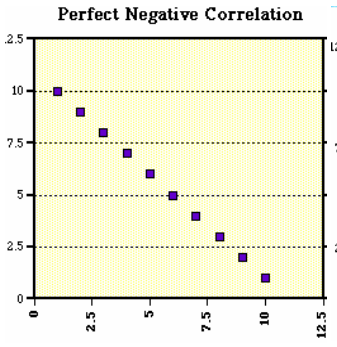
- the formula gives an average of the product of the standardized value (z) of both variables.
- when z_x and z_y have a positive association, they are either both positive and negative. Their product will always be positive.
- when z_x and z_y have a negative association, one of them is positive while the other one is negative. Their products will always be negative.

Properties of Correlation

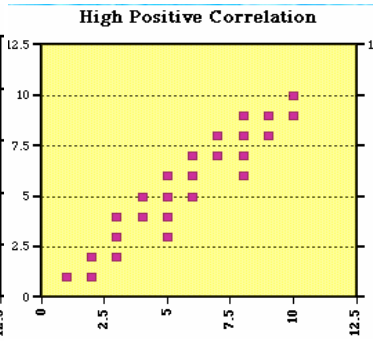
1. Positive r means positive association. Negative r means negative association. (The sign of r indicates the direction of the scatter plot.)
2. All correlation has values between $-1 \le r \le 1$. When r is close to zero, there is very little or no correlation between the variables. The closer r is equal to -1 or 1, the stronger the relations are between the two variables.



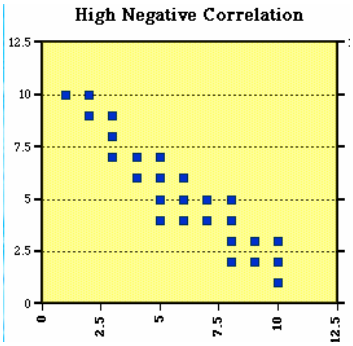
$r = 1$



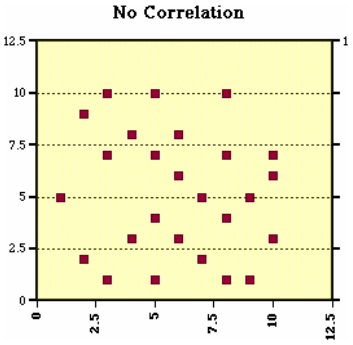
$r = -1$



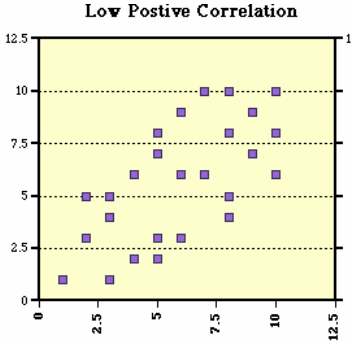
$r \approx 0.9$



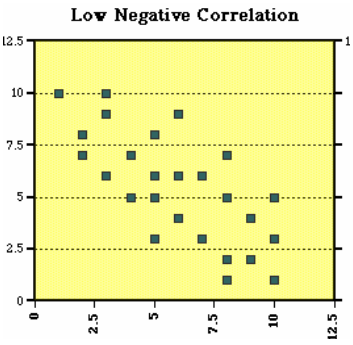
$r \approx -0.9$



$r = 0$



$r \approx 0.6$



$r \approx -0.6$

3. r is unitless, and is unaffected by change in units due to its measure of standardized value.

4. Correlation only measures the strength of a linear relationship. It cannot be used to measured curve relationship.

5. Due to the fact r uses standardized values, which include means and standard deviations, it is easily affected by the presence of outliers.

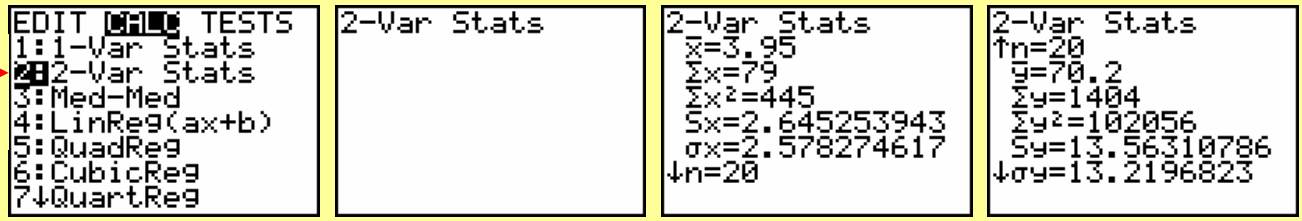
Example 1: Calculate the correlation of the scatter plot of the *Number of Absences and Final Scores for an Algebra I Class in Suburbia Public High School* from Section 3.1. Comment on the direction and the strength and direction using the correlation.

1. Find the mean and standard deviation of both variables (in L_1 and L_2):

STAT → ENTER

Scroll Down

Select Option 2



2. Substitute into the correlation formula.

$$r = \frac{1}{n-1}\sum\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$$

The data points are (3, 75), (6, 67), …, (4, 85)

$\bar{x} = 3.95$
$s_x = 2.6453$
$\bar{y} = 70.2$
$s_y = 13.5631$

$$r = \frac{1}{20-1}\left[\left(\frac{3-3.95}{2.6453}\right)\left(\frac{75-70.2}{13.5631}\right)+\left(\frac{6-3.95}{2.6453}\right)\left(\frac{67-70.2}{13.5631}\right)+\ldots+\left(\frac{4-3.95}{2.6453}\right)\left(\frac{85-70.2}{13.5631}\right)\right]$$

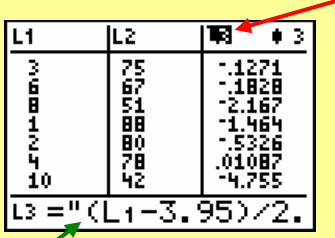
Use L_3 = "(L_1 – 3.95) / 2.6453 * (L_2 – 70.2) / 13.5631" for $\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$

Sum up column L_3 to acquire $\sum\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$

Cursor must be on the column heading before entering formula

2nd → LIST / STAT → → 2nd → L_3 → 3

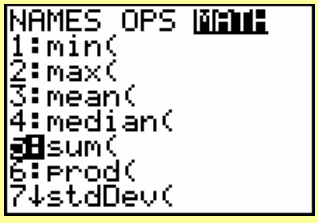
Select MATH on Menu

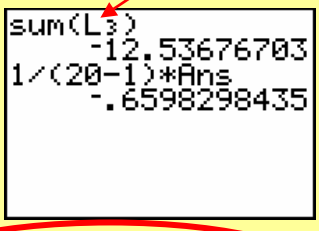


" must be used to enter formula

ALPHA → MEM / " → +

Select Option 5





$r = -0.6598$
(Low Negative Correlation)

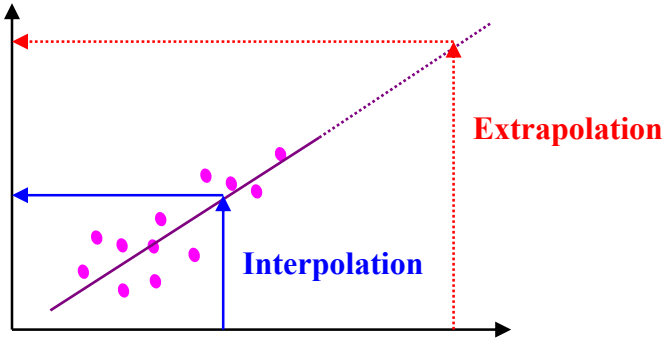
3.2 Assignment: pg. 130 #3.18; pg. 135–136 #3.25, 3.29

3.3A: Least Square Regression

Regression Line: - a straight line that explains how *y* changes as *x* changes.
- often use to predict values of *y* for a given value of *x* or vice versa.

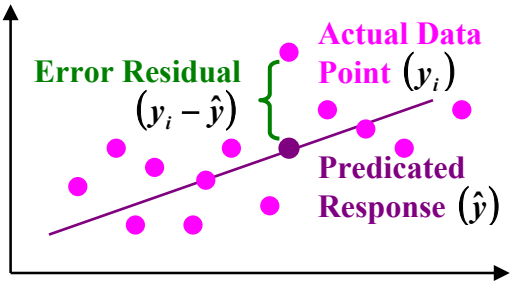
Extrapolation: - when values are predictable outside the experimental data points.

Interpolation: - when values are predicted within the range of experimental data points.



Least Square Regression: - a method to find a line that explains in specific setting, the relationship between two variables.

Least Square Regression Line (LSRL): - a regression line that minimizes the sum of squares of the vertical direction of the data points from the line.



LSRL is achieved when $\sum (y_i - \hat{y})^2$ = Minimum

Equation of the Least Square Regression Line

	Textbook Notation	AP Exam Notation
	$\hat{y} = a + bx$	$\hat{y} = b_0 + b_1 x$
	$\hat{y}$ = predicted response of *y* value for any *x* value	
Slope	$b = r\frac{s_y}{s_x}$	$b_1 = r\frac{s_y}{s_x}$
	when $r = \pm 1$, then $b = \pm\frac{s_y}{s_x} = \pm\frac{\Delta y}{\Delta x}$ (true slope)	
y-intercept	$a = \bar{y} - b\bar{x}$	$b_0 = \bar{y} - b_1\bar{x}$

Example 1: Using the *Number of Absences and Final Scores for an Algebra I Class in Suburbia Public High School* data from Section 3.1,

a. find the least square regression line using the formulas.
b. find the least square regression line using the TI-83 Plus Graphing Calculator.
c. predict the final scores when there are 7 absences.
d. determine the number of absences of a student's final mark is 75%.

a. Find the least square regression line using the formulas.

From Example 1 in 3.2, we have determined the followings,

$r = -0.6598$
$\bar{x} = 3.95$
$s_x = 2.6453$
$\bar{y} = 70.2$
$s_y = 13.5631$

Substitute into formulas

Slope

$$b_1 = r\frac{s_y}{s_x}$$

$$b_1 = (-0.6598)\frac{(13.5631)}{(2.6453)}$$

$$b_1 = -3.38296$$

y-intercept

$$b_0 = \bar{y} - b_1\bar{x}$$

$$b_0 = (70.2) - (-3.3830)(3.95)$$

$$b_0 = 83.56285$$

Finally, substitute into equation of the least square regression line, $\hat{y} = b_0 + b_1 x$

or using the context of the question

$$\hat{y} = 83.56285 - 3.38296x$$
$$FM = 83.56295 - 3.38296A$$

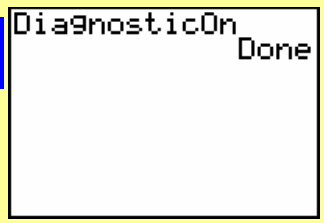
b. Find the least square regression line using the TI-83 Plus Graphing Calculator.

1. Turn Diagnostic On

2nd **CATALOG** **0**

ENTER

Again

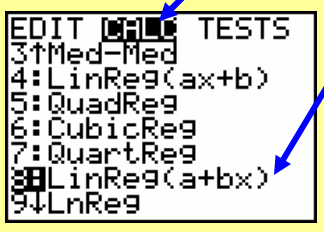


Select DiagnosticOn **ENTER**

Note: After **DiagnosticOn** is selected; it will remain ON even when the calculator is turned Off. However, resetting the calculator will turn the Diagnostic Off (factory setting).

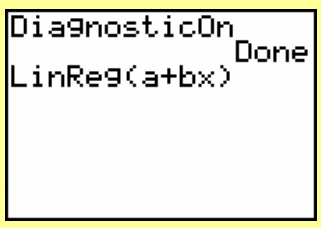
2. Obtain LinReg ($a + bx$), Copy Equation to Scatter Plot. **Y=** **Screen, and Graph Least Square Line on Scatter Plot.**

STAT **Select CALC, use**

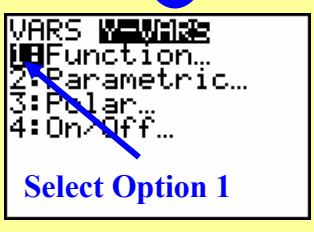


Choose Option 8 Linear Regression ($a + bx$)

ENTER

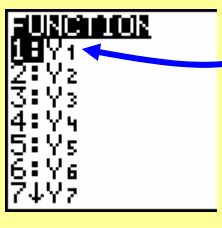


To access Y_1, press **VARS**



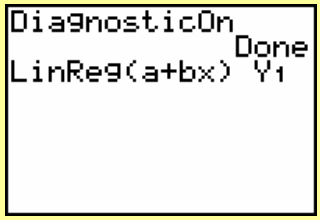
Select Option 1

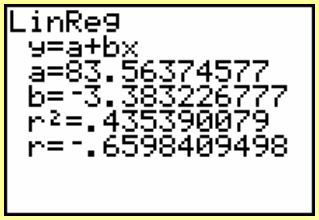
Select Option 1 for Y_1 **ENTER**

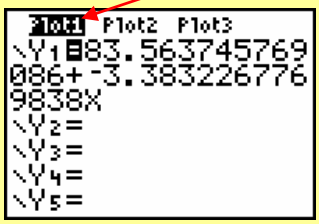


LinReg ($a + bx$) Y_1 calculates the equation and copied into the **Y=** **Screen.**

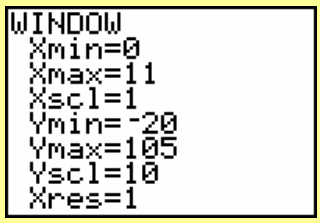
ENTER **Y=** **Be Sure Plot1 is On**



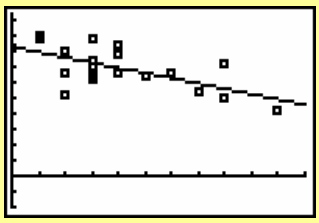




WINDOW x: [0, 11, 1]
y: [–20, 105, 10]



GRAPH



Least Square Line Equation

$y = 83.56375 - 3.38323x$
$r = -0.6598$
(It is the same r as calculated in 3.2)

c. Predict the final scores when there are 7 absences.

Using Least Squares Equation from Formulas,

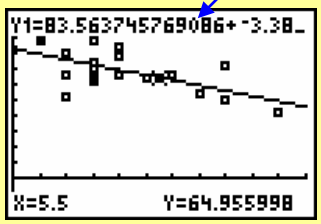
$FM = 83.56295 - 3.38296A$
$FM = 83.56295 - 3.38296(7)$

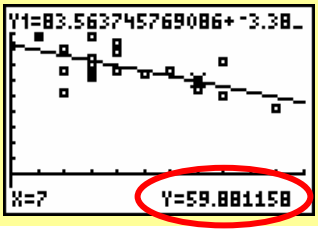
***FM* = 59.88%**

Using Least Squares Equation from Calculator,

TRACE ▼

Be sure the Y_1 Equation is up **Type 7**

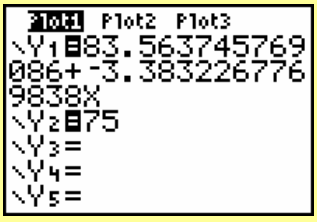


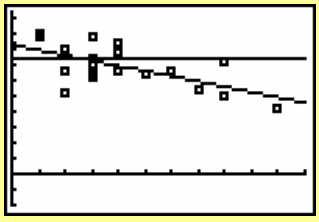


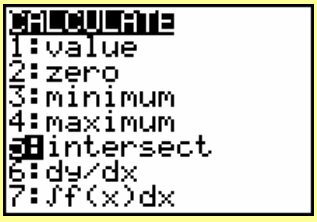
d. Determine the number of absences of a student's final mark is 75%.

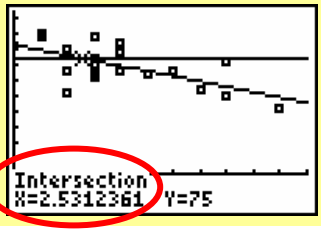
Using Least Squares Equation from Calculator,

Y= **Enter 75 in Y_2** **GRAPH** **2nd** **CALC** **TRACE** **ENTER** **ENTER** **ENTER**









3.3A Assignment: pg. 142–143 #3.31, 3.33 and 3.35

3.3B: Coefficient of Determination

Sum of Squares about the Mean (*SSM*): - total sample variability (variance) of the response variable.

$$SSM = \sum (y - \bar{y})^2$$

Sum of Squares for Error (*SSE*): - total deviation between each response variable to the predicted values off the Least Square Regression Line.
- the larger the *SSE*, the more scattered are the scores compared to the ones predicted by the Least Square Regression Line.

$$SSE = \sum (y - \hat{y})^2$$

Coefficient of Determination (r^2): - the difference between the *SSM* and *SSE* compared to the original *SSM*.
- when express in percentage, it is also referred to as **variance (a measure on how many *y* values correspond to the changes in *x* values)**.

Coefficient of Determination

$$r^2 = \frac{SSM - SSE}{SSM} \quad \text{or} \quad r^2 = (r)^2$$

- when *SSM* and *SSE* are similar, $r^2 \approx 0$. Therefore, $r \approx 0$ when $SSM \approx SSE$
- when $SSE = 0$, then $r^2 = 1$. $r = \pm 1$ and we have a true equation of the least square regression line.
- the square accounts for the amount of variation by the linear relationship on either variables. (*y* versus *x*, and *x* versus *y*)

Example 1: In a study between number of hours of violence watched on TV in a week and the number of violence acts occur on the playground for a group of children found that the correlation is 0.85. The mean and standard deviation of the number of hours of violence watched on TV are 16.2 hours and 5.28 hours respectively. The number of violence acts observed has a mean of 4.8 with a standard deviation of 1.42.

a. Identify the explanatory and response variables.
b. Determine the slope and *y*-intercept of the regression line.
c. If a child were to watch 20 hours of violence on TV in a week, how many violence act can we expect he will undertake on the playground?
d. Comment on the strength of the relationship using *r* and r^2.

a. Identify the explanatory and response variables.

Since the number of hours of violence watched on TV seems to affect behavior on the playground,

Explanatory Variable (*x*) = Number of Hours watched on TV in a week.
Response Variable (*y*) = Number of Violence Acts Observed on Playground.

b. Determine the slope and y-intercept of the regression line.

From the questions, we have the followings,

$r = 0.85$
$\bar{x} = 16.2$
$s_x = 5.28$
$\bar{y} = 4.8$
$s_y = 1.42$

Substitute into formulas

Slope

$$b_1 = r\frac{s_y}{s_x}$$

$$b_1 = (0.85)\frac{(1.42)}{(5.28)}$$

$$\boldsymbol{b_1 = 0.22860}$$

y-intercept

$$b_0 = \bar{y} - b_1\bar{x}$$

$$b_0 = (4.8) - (0.22860)(16.2)$$

$$\boldsymbol{b_0 = 1.09668}$$

Finally, substitute into equation of the least square regression line, $\hat{y} = b_0 + b_1 x$

or using the context of the question

$$\boldsymbol{\hat{y} = 1.09668 + 0.22860x}$$
$$\boldsymbol{VA = 1.09668 + 0.22860TV}$$

c. If a child were to watch 20 hours of violence on TV in a week, how many violence act can we expect he will undertake on the playground?

Using Least Squares Equation from Formulas,

$VA = 1.09668 + 0.22860TV$
$VA = 1.09668 + 0.22860(20)$
$VA = 5.66868$

$\boldsymbol{VA = 6}$ **Violence Acts**

d. Comment on the strength of the relationship using r and r^2.

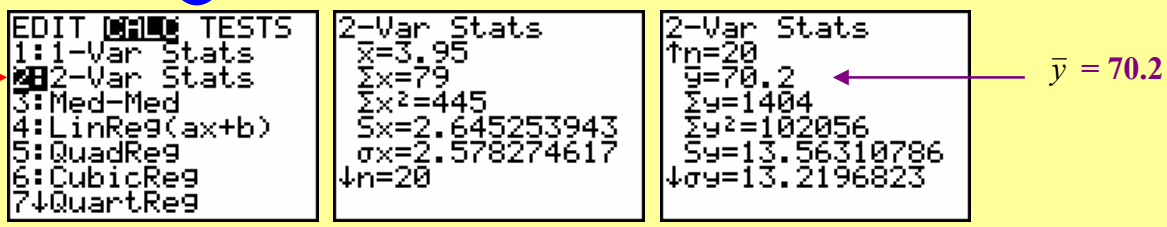
$r = 0.85$ which means $r^2 = 0.7225$. Even though the correlation seems strong, a closer examination of the coefficient of determination indicates a value somewhat less than 1. This can only mean that there are other variables that can affect the amount of violence acts a child can perform other than the total time spent watching violence TV.

Example 2: Using the formula *SSM* and *SSE*, calculate r^2 for the *Number of Absences and Final Scores for an Algebra I Class in Suburbia Public High School* data from Section 3.1.

1. Calculating *SSM*

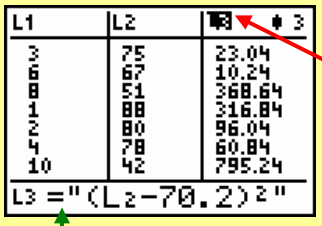
STAT **ENTER**

Select Option 2



$\bar{y} = 70.2$

Use L_3 = "$(L_2 - 70.2)^2$" for $(y-\bar{y})^2$



Cursor must be on the column heading before entering formula

" must be used to enter formula

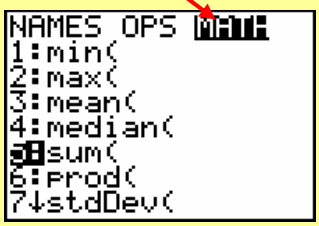
ALPHA **MEM "** **+**

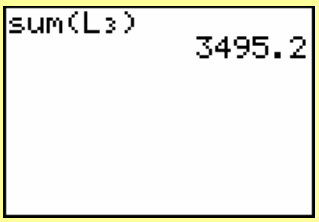
Sum up column L_3 to acquire $\sum(y-\bar{y})^2$

2nd **LIST** **STAT** **2nd** **L_3** **3**

Select MATH on Menu

Select Option 5

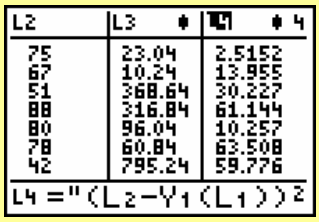




***SSM* = 3495.2**

2. Calculating *SSE*

Use L_4 = "$(L_2 - Y_1(L_1))^2$" for $(y-\hat{y})^2$

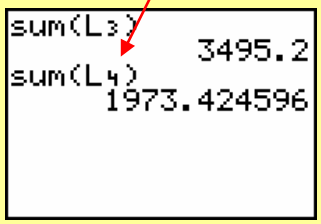


$Y_1(L_1)$ means each *y* value for each of the individual *x* values in L_1

***Note:* LinReg (*a* + *bx*) Y_1 must be ran prior to entering formula for L_4.**

Sum up column L_4 to acquire $\sum(y-\hat{y})^2$

2nd **L_4** **4**

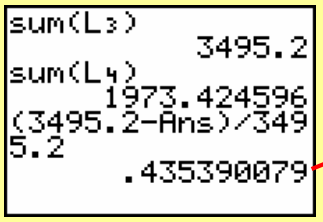


***SSE* = 1973.424596**

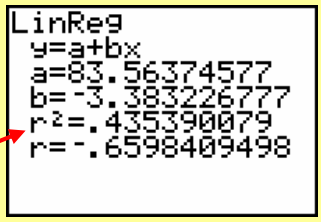
3. Calculating r^2

$$r^2 = \frac{SSM - SSE}{SSM}$$

$$r^2 = \frac{3495.2 - 1973.424596}{3495.2}$$



Verify using LinReg with DiagnosticOn



r^2 = 0.43539

Since r^2 is even closer to zero than *r*, we can definitely say that there is very little correlation between *x* and *y*.

3.3B Assignment: **pg. 150–151 #3.36, 3.37 and 3.38**

3.3C: Residuals

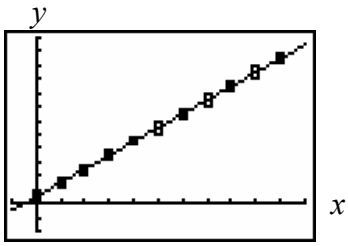
Residuals: - the difference between an observed response variable data to the predicted value from the Least Square Regression Line.

$$\text{Residual} = y - \hat{y}$$

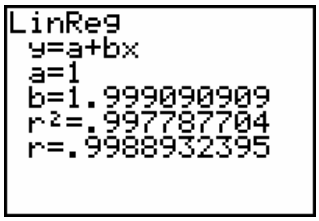
Residual Plot: - a plot of residual on the *y*-axis versus the explanatory variable on the *x*-axis.
- the sum of residuals is always zero.

Roundoff Errors: - calculator or computer software shows sum of residual close to zero but not exactly zero.

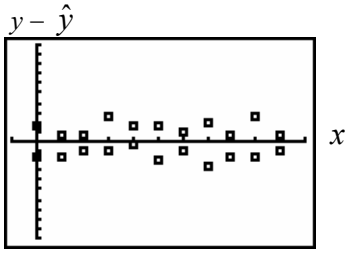
1. Good Residual Plot: - all scores huddle around the mean



Scatter Plot and Least Square Regression Line

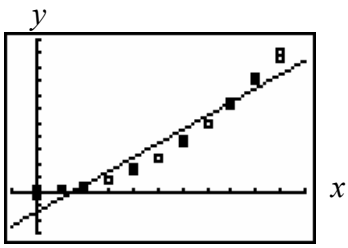


A good fit as shown with the correlation

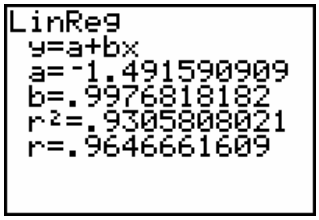


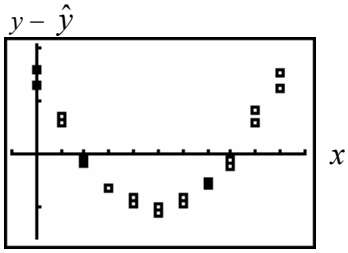
Residual Plot is also good with small residual about 0

2. Curve Residual Plot: - a linear model is not appropriate

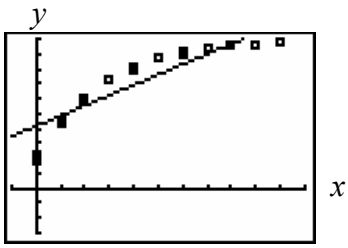


Scatter Plot and Least Square Regression Line. We can clearly see that a curve would be a better fit.

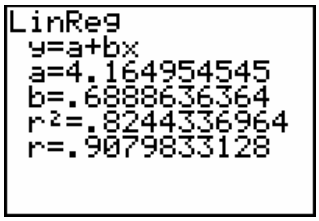


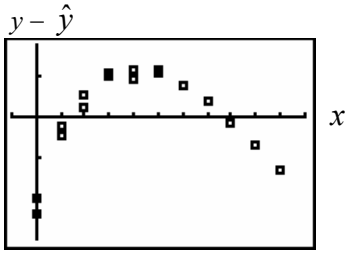


Although correlation is high, a curve residual plot shows that the linear model is not appropriate.



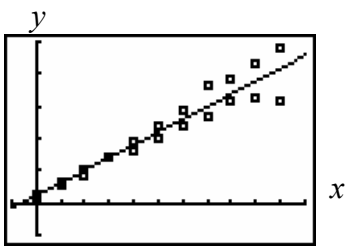
Scatter Plot and Least Square Regression Line. Again, we can clearly see that a curve would be a better fit.



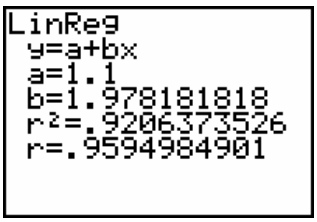


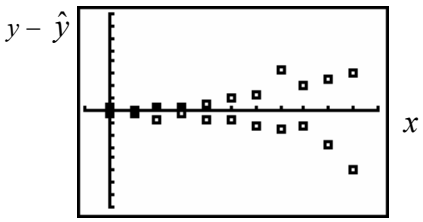
Although correlation is somewhat strong, a curve residual plot shows that the linear model is not appropriate.

3. Locally Well Suited Residual Plot: - the residual has a larger spread at higher or lower *x*-values. Regression model may not be appropriate for those regions.



Scatter Plot and Least Square Regression Line. We can clearly see that the predicted values are not accurate at high *x*-values.

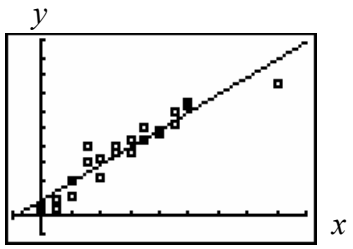




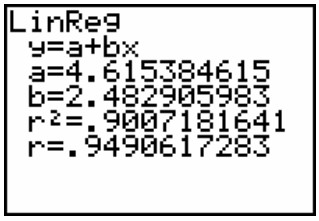
Although correlation is somewhat strong, the residual plot shows that the linear model is only appropriate at small *x*-values. There are too many residual away from 0 at higher *x*-values.

Outliers of Residual Plot: - a residual value that is too far from the zero line compared to other residual value.

Influential Observation: - a lone residual at high *x* value that can practically change the position of the regression line. (It is difficult to tell whether that lone data point is an outlier in the extreme *x* direction.)

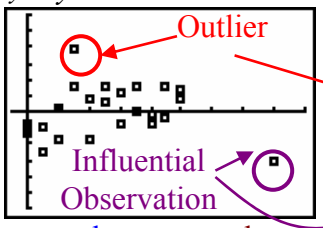


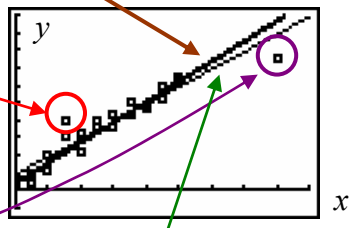
Scatter Plot and Least Square Regression Line. We can see that there is an outlier and an influential observation.



Although correlation is somewhat strong, the residual plot shows clear where the outlier and influential observation are.

y – *ŷ* LSRL without the outlier and influential observation





Original LSRL with outlier and influential observation

Example 1: Graph the residual plot for the *Number of Absences and Final Scores for an Algebra I Class in Suburbia Public High School* data from Section 3.1.

a. Explain if the regression line is appropriate to predict the final marks using number of absences.
b. Verify the sum of the residual is 0 and indicate any roundoff error.
c. Identify and comment on any outlier and influential observations.

Graphing Residual Plot

Use L_3 = "($L_2 - Y_1(L_1)$)" for $(y - \hat{y})$

L1	L2	L3
3	75	1.5859
6	67	3.7356
8	51	-5.498
1	88	7.8195
2	80	3.2027
4	78	7.9692
10	42	-7.731

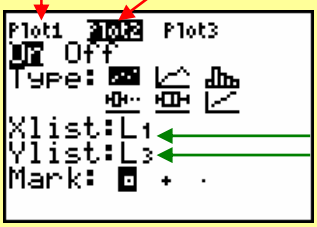
L3 ="L2-Y1(L1)"

$Y_1(L_1)$ means each *y* value for each of the individual *x* values in L_1

***Note:* LinReg ($a + bx$) Y_1 must be ran prior to entering formula for L_3.**

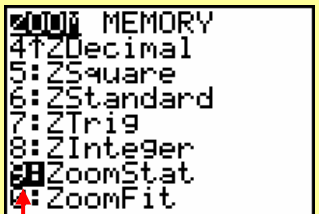
2nd **STAT PLOT** **Y =**

Plot 1 Off **Plot 2 On**



L_1 as *x*
L_3 as *y*

ZOOM

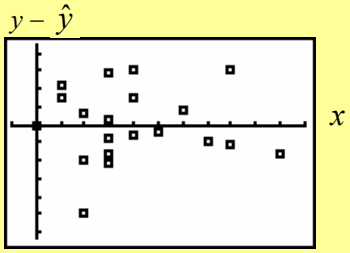


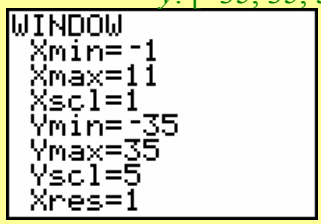
Select Option 9 ZoomStat

ENTER

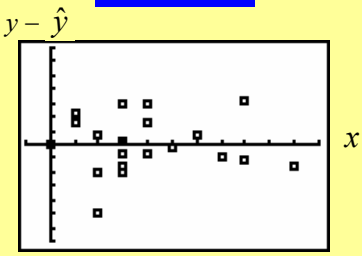
Customize Windows

WINDOW x: [−1, 11, 1]
y: [−35, 35, 5]





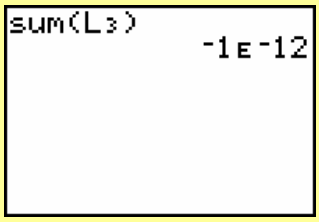
GRAPH



a. Explain if the regression line is appropriate to predict the final marks using number of absences.

The Regression is not appropriate for this data set. There are a few outliers as observed most noticeably at $x = 2$. This is compounded by a possible influential observation at $x = 10$.

b. Verify the sum of the residual is 0 and indicate any roundoff error.

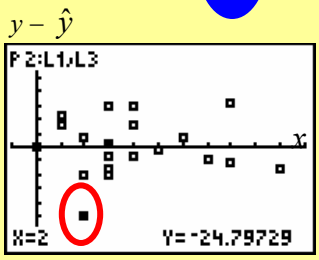


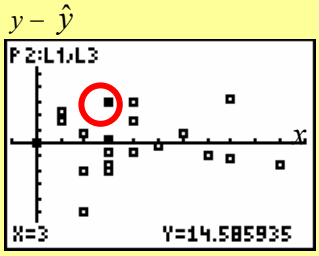
The roundoff error is indicated as -1×10^{-12}

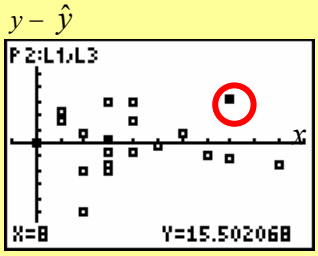
c. Identify and comment on any outlier and influential observations.

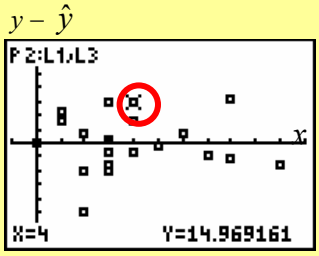
Trace along the Residual Plot and find outliers $(y - \hat{y}) \geq 11$ and $(y - \hat{y}) \leq -11$

TRACE

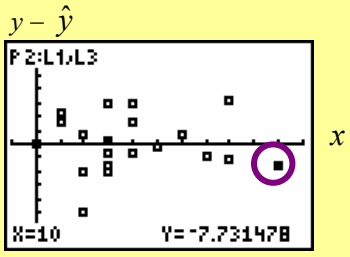








The furthest x is at $x = 10$, which might be an influential observation.



There are 4 outliers corresponding to (2, 52), (3, 88), (8, 72) and (4, 85)

(10, 42) might be an influential observation

3.3C Assignment:
pg. 156–162 #3.39, 3.41, 3.43 and 3.45

Chapter 3 Review Assignment
pg. 168–173 #3.52, 3.55 and 3.57

Chapter 4: More on Two-Variable Data

4.1A: Modeling Non-linear Data: Exponential Function

Exponential Function: - a function where y_n is increased by **multiplying** y_{n-1} with a **common ratio (b)**.

Exponential Function

$$y = ab^x$$

a = initial y value at $x = 0$ (y-intercept)
$[a \neq 0$, y-intercept $\neq 0]$

b = common ratio = $\dfrac{y_n}{y_{n-1}}$

When $b > 1$, $y \uparrow$ as $x \uparrow$ (Exponential Growth)
When $0 < b < 1$, $y \downarrow$ as $x \uparrow$ (Exponential Decay)

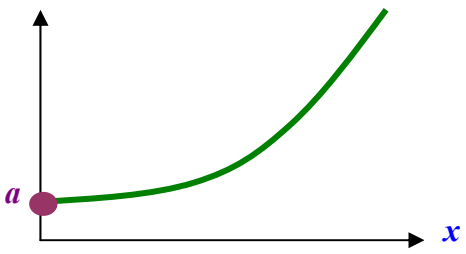
Exponential and Logarithmic Functions

$$y = b^x \leftrightarrow \log_b y = x$$

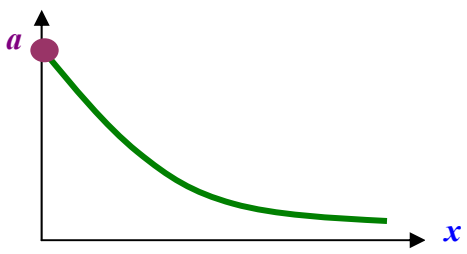
Logarithm Laws

$$\log (XY) = \log X + \log Y$$
$$\log \left(\frac{X}{Y}\right) = \log X - \log Y$$
$$\log X^n = n \log X$$

Exponential Growth ($b > 1$)
($b - 1$) = "Rate of Growth"

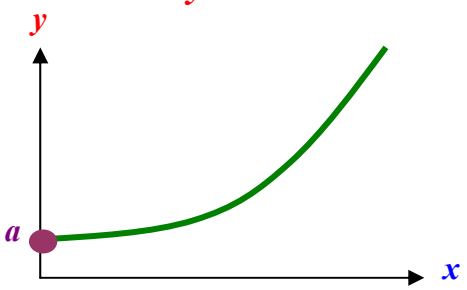


Exponential Decay ($0 < b < 1$)
($1 - b$) = "Rate of Decay"

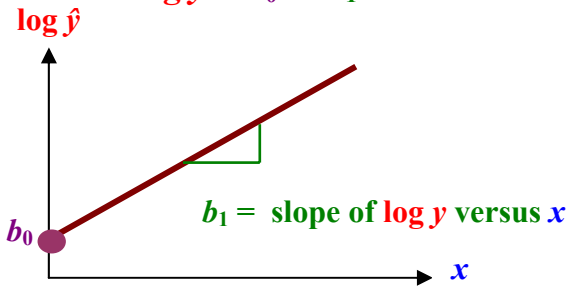


Logarithmic Transformation of Exponential Function: - when the graph of (y versus x) is transformed to ($\log y$ versus x).

$y = ab^x$



$\log \hat{y} = b_0 + b_1 x$



Inverse Transformation of Exponential Function: - when ($\log \hat{y}$ versus x) is transformed back to ($\hat{y}$ versus x) in the powers of 10.

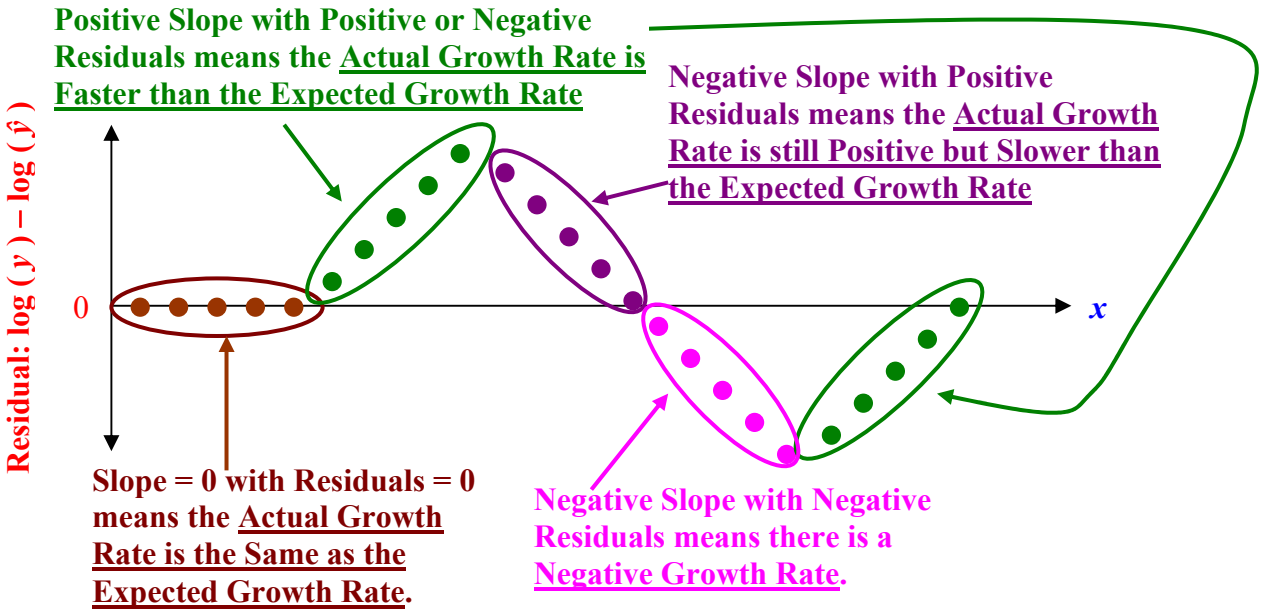
Inverse Transformation

$$\log \hat{y} = b_0 + b_1 x$$
$$10^{\log \hat{y}} = 10^{(b_0 + b_1 x)}$$
$$\hat{y} = \left(10^{b_0}\right)\left(10^{b_1 x}\right)$$

Compare to $y = ab^x$,

y-intercept: $a = \left(10^{b_0}\right)$
Average Common Ratio: $b = \left(10^{b_1}\right)$

Residual of Logarithmic Transformation: - compares the log y (log y actual) with log $\hat{y}$ (log y predicted).



Example 1: The US Department of Labour keeps many statistical information on the country. One of these information is employment of anyone above the age of 16. A sample size of 50,000 households is survey in California every year. The number of households where everyone over the age of 16 is employed over 15 hours a week is recorded.

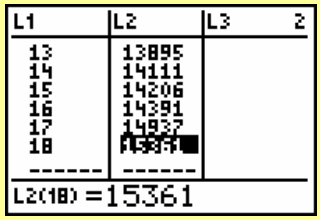
Number of households in California where everyone over the age of 16 is employed

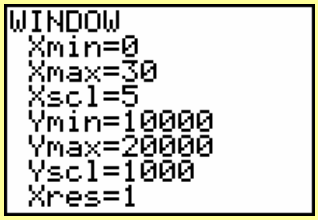
Year	Number of Years since 1980	Fully Employed Household	Year	Number of Years since 1980	Fully Employed Household
1981	1	10938	1990	10	14313
1982	2	10967	1991	11	13992
1983	3	11095	1992	12	13954
1984	4	11631	1993	13	13895
1985	5	12048	1994	14	14111
1986	6	12442	1995	15	14206
1987	7	12955	1996	16	14391
1988	8	13292	1997	17	14937
1989	9	13870	1998	18	15361

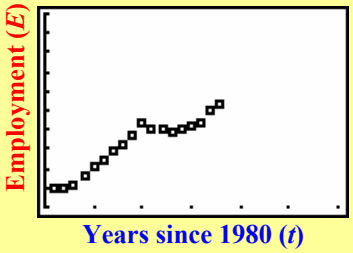
a. Make a scatterplot of the table above. Use x = number of years after 1980.
b. Test for the common ratio. Make a table to record your result.
c. Construct a graph using the exponential logarithmic transformation.
d. Find and graph the least square regression equation of the exponential logarithmic transformation.
e. Using the inverse transformation, provide the exponential equation of the data above. Draw this equation with the original scatter plot. Determine the average rate of growth for employment and state the meaning of the predicted y-intercept.
f. Determine the predicted employment in 2005.
g. Graph and interpret the residual plot of the logarithmic transformation.

a. Make a scatterplot of the table above. Use x = number of years after 1980.

1. Enter Data in L_1 and L_2. **2. Set Windows** **3. Turn StatPlot On and Graph**





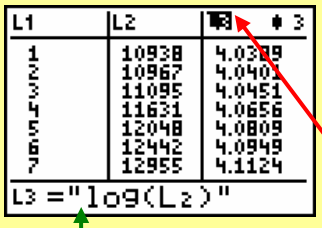


b. Test for the common ratio. Make a table to record your result.

Year	Fully Employed Household	Common Ratio y_n / y_{n-1}	Year	Fully Employed Household	Common Ratio y_n / y_{n-1}
1981	10938	-----	1990	14313	**1.031939**
1982	10967	**1.002651**	1991	13992	**0.977573**
1983	11095	**1.011671**	1992	13954	**0.997284**
1984	11631	**1.048310**	1993	13895	**0.995772**
1985	12048	**1.035852**	1994	14111	**1.015545**
1986	12442	**1.032703**	1995	14206	**1.006732**
1987	12955	**1.041231**	1996	14391	**1.013023**
1988	13292	**1.026013**	1997	14937	**1.037940**
1989	13870	**1.043485**	1998	15361	**1.028386**

c. Construct a graph using logarithmic transformation.

1. Use L_3 = "log (L_2)"

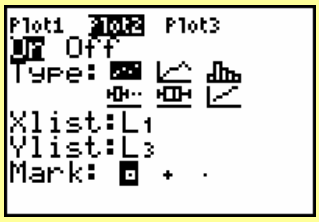


" must be used to enter formula

Cursor must be on the column heading before entering formula

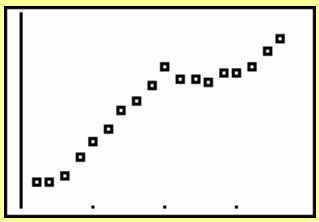
ALPHA **MEM "** **+**

2. Turn Off Plot 1, Turn On Plot 2



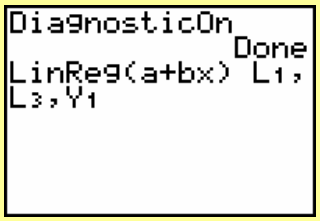
3. Choose ZoomStat to Graph

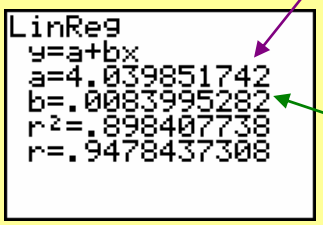
log (E)



d. Find the least square regression equation of the logarithmic transformed graph.

1. Turn DiagnosticOn,
Run LinReg ($a + bx$) on L_1 and L_3
Paste Equation to Y_1



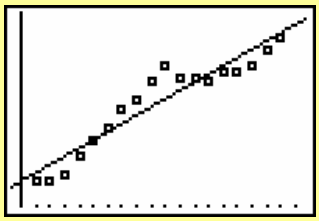


b_0

b_1

2. Graph

log (E)



Years since 1980 (t)

e. Using the inverse transformation, provide the exponential equation of the data above. Draw this equation with the original scatter plot. Determine the average rate of growth for employment and predicted y-intercept of the original scatter plot.

$\log \hat{E} = b_0 + b_1 t$

$10^{\log \hat{E}} = 10^{(b_0 + b_1 t)}$

$\hat{E} = (10^{b_0})(10^{b_1 t})$

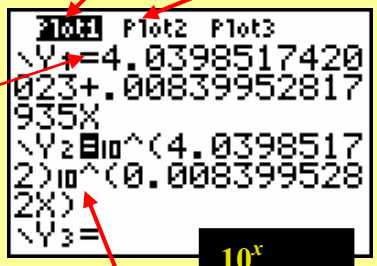
$b_0 = 4.03985172$

$b_1 = 0.0083995282$

$\hat{E} = (10^{4.03985172})(10^{0.0083995282t})$

1. Enter equation into Y_2. Turn Off Y_1
Turn Plot 1 On, and Plot 2 Off.

ENTER

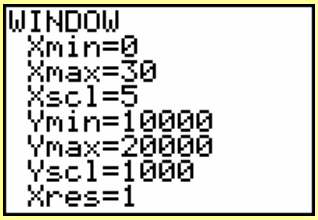


10^x

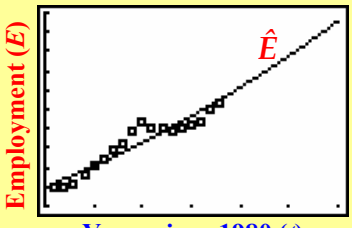
LOG

2nd

2. Reset Windows



3. Graph



Years since 1980 (t)

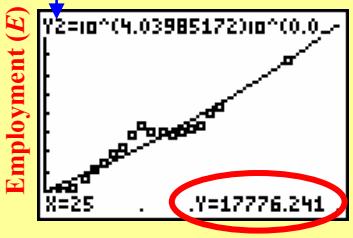
Compare to $y = ab^x$,
y-intercept: $a = 10^{4.03985172} = 10936$ employed in 1980 ($t = 0$)
Average Common Ratio: $b = 10^{0.0083995282} = 1.01952887$
Average Rate of Growth = $(1.01952887 - 1) = 1.953\%$

f. Determine the predicted employment in 2005.

Be Sure Y_2 is up

TRACE

Type 25
(In 2005, $t = 25$)

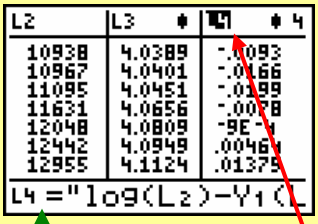


Years since 1980 (t)

In 2005, the predicted employment is 17776

g. Graph and interpret the residual plot of the logarithmic transformation.

1. Use L_4 = "log (L_2) – Y_1 (L_1)"

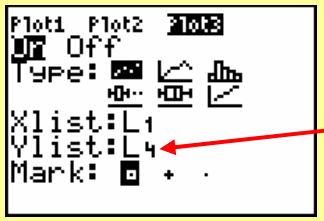


" must be used to enter formula

ALPHA **MEM "** **+**

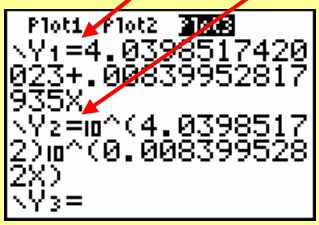
Cursor must be on the column heading before entering formula

2. Turn Off Plot 1 and Plot 2, Turn On Plot 3

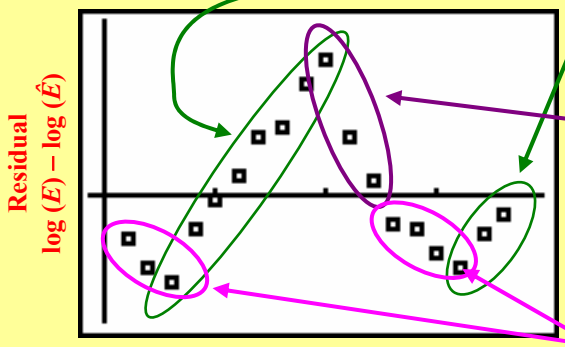


2nd **L_4** **4**

3. Turn Off Y_1 and Y_2



4. Graph using ZoomStat



From 1983 to 1990 and from 1996 to 1998, there is a Fast and Positive Growth Rate in California Employment. (Economic Boom 1986–1990 and Recovery 1983–1985; 1996 –1998)

From 1990 to 1992 there is a Slow Down but still Positive Growth Rate in California Employment. (Economy is Slowing Down)

From 1981 to 1983 and from 1992 to 1996, there is a Negative Growth Rate in California Employment. (Recession)

4.1A Assignment:
pg. 189–190 #4.2

4.1B: Modeling Non-linear Data: Power Function

Power Function: - a function where x is the base raise to a **constant exponent b**.
- it includes most polynomial $\{b > 1, b \in R\}$ and radical functions $\{0 > b > 1, b \in R\}$.
- **all power function begins at (0, 0)**. Therefore, data where the starting point should be at the origin should use power regression instead of exponential regression.

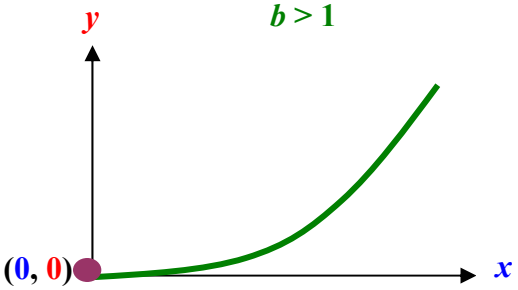
Power Function

$$y = ax^b$$

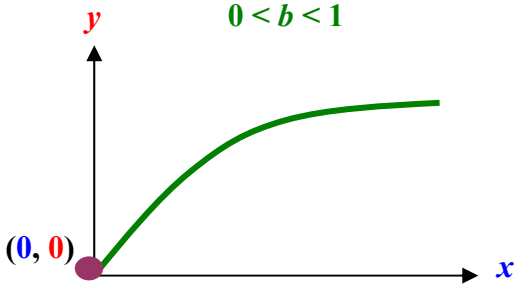
a = Coefficient of the Power Function
($a > 0$ for Logarithmic Transformation)

b = Constant Exponent ($b > 0$)
(If $-1 < b < 0$, then y-intercept $\neq 0$.)
(If $b < -1$, then y-intercept will be undefined.)

Polynomial Function $b > 1$



Radical Function $0 < b < 1$

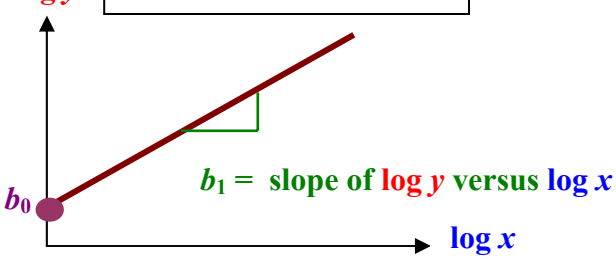


Logarithmic Transformation of Power Function: - when the graph of (y versus x) is transformed to (log y versus log x).

$y = ax^b$
$\log y = \log (ax^b)$
$\log y = \log a + \log x^b$

$$\log y = \log a + b \log x$$

$$\log \hat{y} = b_0 + b_1 \log x$$



Inverse Transformation of Power Function: - when (log $\hat{y}$ versus log x) is transformed back to ($\hat{y}$ versus x) using the powers of 10.

Compare the two equations above:
$\log y = (\log a) + b \log x$
$\log \hat{y} = b_0 + b_1 \log x$

$b_0 = \log a$

$a = 10^{b_0}$ $\quad b = b_1$

Inverse Transformation

$$y = ax^b$$

$$y = \left(10^{b_0}\right)\left(x^{b_1}\right)$$

Example 1: Explain why the data from Example 1of 4.1 A on *Number of households in California where everyone over the age of 16 is employed* is not appropriate using the power function model.

The fact that employment in 1980 ($t = 0$) would not be 0 means that y-intercept of the nonlinear data would not be (0, 0). Since power regression is appropriate only for curves that should have (0, 0) for y-intercept, it is most definitely unsuitable for the employment data set.

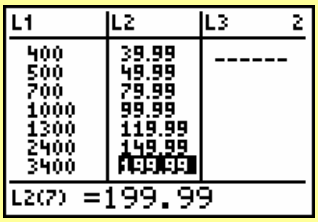
Example 2: A cellular phone company offers the following rates for their National Plan.

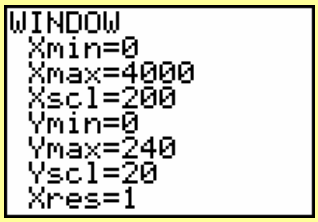
Number of Minutes per Month	Monthly Cost	Number of Minutes per Month	Monthly Cost
400	$ 39.99	1300	$ 119.99
500	$ 49.99	2400	$ 149.99
700	$ 79.99	3400	$ 199.99
1000	$ 99.99		

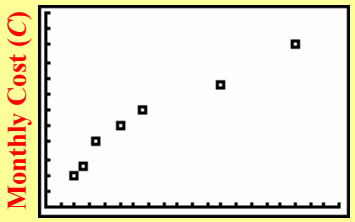
a. Make a scatterplot of the table above.
b. Construct a graph using the power logarithmic transformation.
c. Find and graph the least square regression equation of the power logarithmic transformation.
d. Using the inverse transformation, provide the appropriate equation of the data above. Draw this equation with the original scatter plot. Interpret the meaning of b in the context of the data.
e. The cellular phone company would like to add a plan of 1800 minutes per month, what should be the appropriate cost.
f. Graph and interpret the residual plot of the power logarithmic transformation. Compare it with the power regression and the original scatter plot. Identify any outliers of both graphs and provide a possible explanation to their existence.
g. Verify that the power model is an appropriate model by finding the coefficient of determinations for both the exponential and power regressions of the original data.

a. Make a scatterplot of the table above.

1. Enter Data in L_1 and L_2. **2. Set Windows** **3. Turn StatPlot On and Graph**

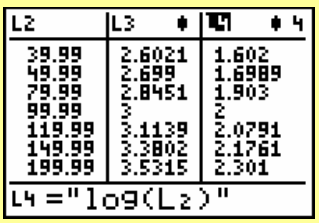


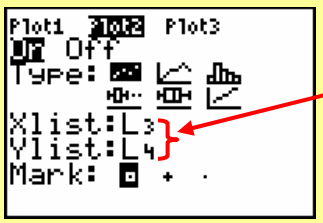




b. Construct a graph using the power logarithmic transformation.

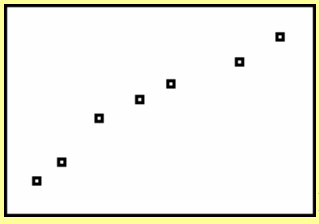
1. Use L_3 = "log (L_1)" and L_4 = "log (L_2)" **2. Turn Off Plot 1, Turn On Plot 2** **3. Choose ZoomStat to Graph**





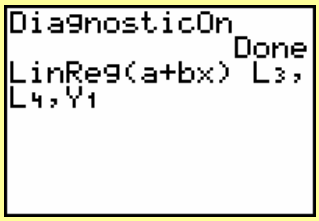
Use L_3 and L_4

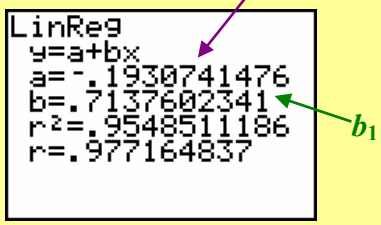
log (C)



c. Find and graph the least square regression equation of the power logarithmic transformation.

1. Turn DiagnosticOn, Run LinReg ($a + bx$) on L_3 and L_4 Paste Equation to Y_1



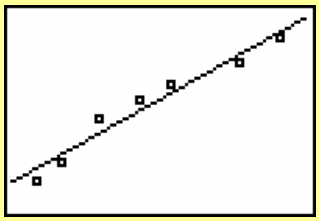


b_0

b_1

2. Graph with LSRL

$\log \hat{C} = b_0 + b_1 \log t$



d. Using the inverse transformation, provide the appropriate equation of the data above. Draw this equation with the original scatter plot.

$C = at^b$

$\log C = \log (at^b)$

$\log C = \log a + \log t^b$

$\log C = \log a + b \log t$

Compare the two equations:

$\log C = \log a + b \log t$

$\log \hat{C} = b_0 + b_1 \log t$

$b_0 = \log a$

$-0.1930741476 = \log a$

$a = 10^{(-0.1930741476)}$

$b = b_1$

$b = 0.7137602341$

Inverse Transformation

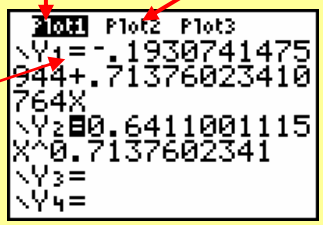
$\hat{C} = at^b$

$\hat{C} = (10^{b_0})(t^{b_1})$

$\hat{C} = (10^{(-0.1930741476)})\ (t^{0.7137602341})$

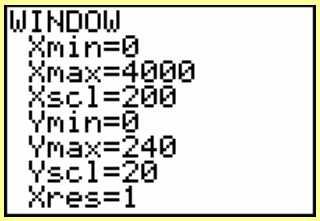
$\hat{C} = 0.6411001115\ (t^{0.7137602341})$

1. Enter equation into Y_2. Turn Off Y_1 Turn Plot 1 On, and Plot 2 Off.

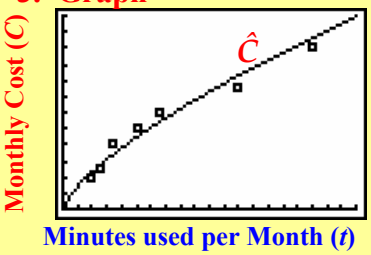


ENTER

2. Reset Windows



3. Graph



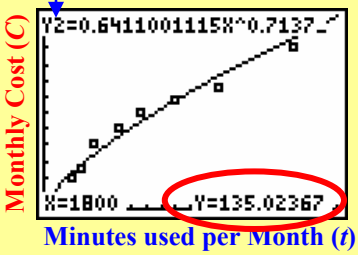
$0 < b < 1$ which means that even though C increases as t increases, but the Cost per Minute decreases as t increases.

e. The cellular phone company would like to add a plan of 1800 minutes per month, what should be the appropriate cost.

Be Sure Y_2 is up

TRACE

Type 1800 for t = 1800 mins



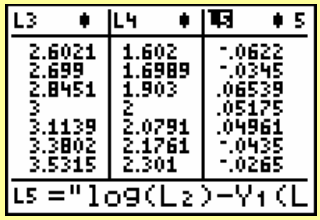
For 1800 minutes per month, the cellular phone company should charge $135.02

f. Graph and interpret the residual plot of the power logarithmic transformation. Identify any outliers and provide a possible explanation to their existence.

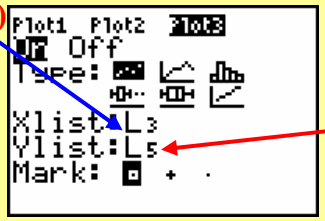
1. Use L_5 = "log (L_2) – Y_1 (L_3)"

Residual = log (C) – log ($\hat{C}$)

(log ($\hat{C}$) is dependent on log x = L_3)

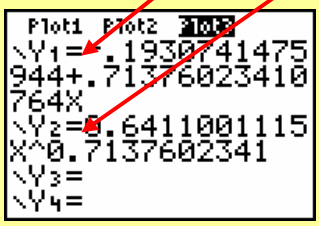


2. Turn Off Plot 1 and Plot 2, Turn On Plot 3

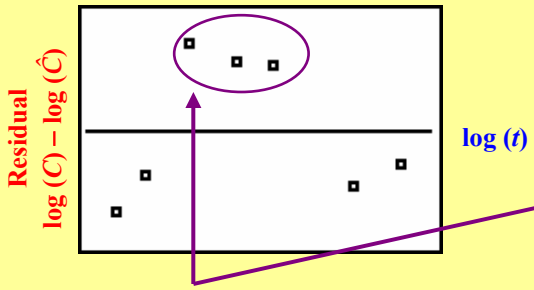


2nd **L_5** **5**

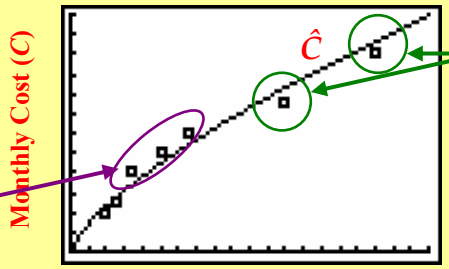
3. Turn Off Y_1 and Y_2



4. Graph of Logarithmic Residual using ZoomStat



5. Graph of Original Scatter Plot with Power Regression

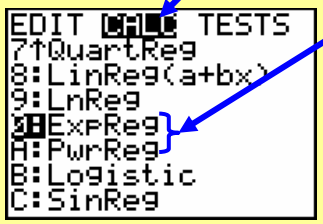


These outliers (2400 min, \$149.99) and (3400 min, \$199.99) are special sales plans for power cell phone users.

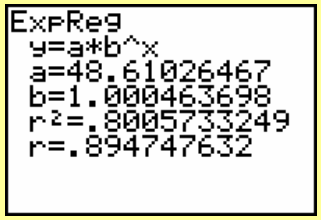
These outliers indicate the plans (700 min, \$79.99), (1000 min, \$99.99) and (1300 min, \$199.99) are above the average cost trend. They might be needed to make up any profits lost on other special plans.

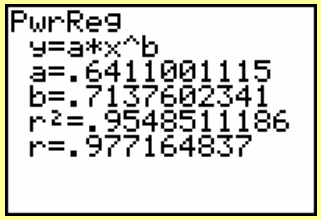
g. Verify that the power model is an appropriate model by finding the coefficient of determinations for both the exponential and power regressions of the original data.

STAT **Select CALC, use**



Choose Option 9 or 0
ExpReg or PwrReg





Exponential Regression r^2 = 0.8005733249 = 80.057%
Power Regression r^2 = 0.9548511186 = 95.485%

(Power Regression has better Variance than Exponential Regression.)

4.1B Assignment:
pg. 197–205 #4.5, 4.7, 4.11 and 4.17

4.2: Interpreting Correlation and Regression

When using correlation and regressions to interpret any scatter plots, one must be careful about their limitations.

1. ***r* is used to interpret Linear Relationships only.**
   - *r* can easily be thrown off with any outliers and influential observations.
   - a residual plot should always be used after a linear regression is made. This is to ensure that a non-linear model is not present even when $r \approx 1$.

2. **Extrapolation should be used Close to the Domain of the Data Set.**
   - linear regression model might be appropriate for a certain domain, prediction at higher *x*-value might not be appropriate.

   Example 1: Height versus Age of Children CANNOT apply to Adults.

3. **Correlation of Average Data Cannot Reflect Individual Changes**.
   - when average data is used, any outliers or individual residuals is muffled. The correlation of the average plot is usually high. But when the regression model is used to identify individual data, the predicted result might not be accurate.

   Example 2: Average Gas Consumption versus Average Temperature of each month will have a higher correlation than if the *x*-variable were to be the Average Temperature Everyday of the year.

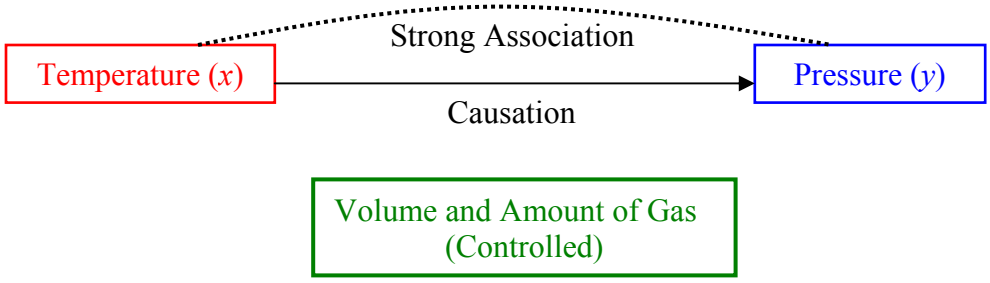
4. **Lurking Variables may Exist even at High Correlation.**

   Lurking Variable: - a variable that can affect the outcome of the responding variable (y) but is either ignored or not controlled when x is manipulated.

   Association is NOT Causation:

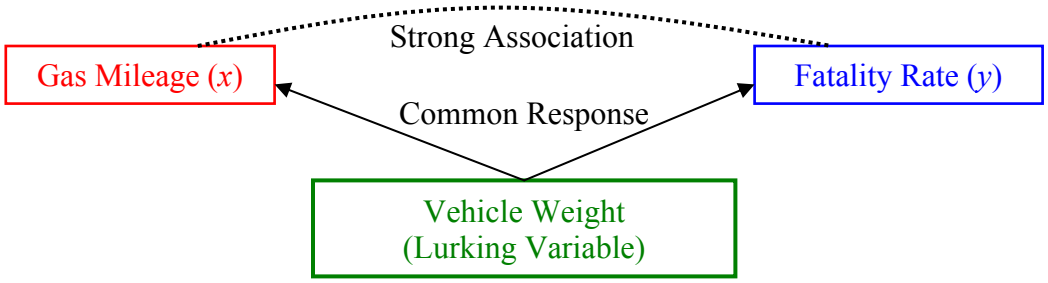
   a. **Causation**: - "a change in x causes a change in y" can only be concluded **when all lurking variables are identified and controlled in an experiment**. The mechanism of causation must be understood as well.

   Example 3: - an increased in temperature causes pressure to increase in a gas canister when the volume and the amount of gas remain constant. This is due to the fact that gaseous particles move faster at higher temperature, causing more collisions or force applied to the inner wall of the canister.



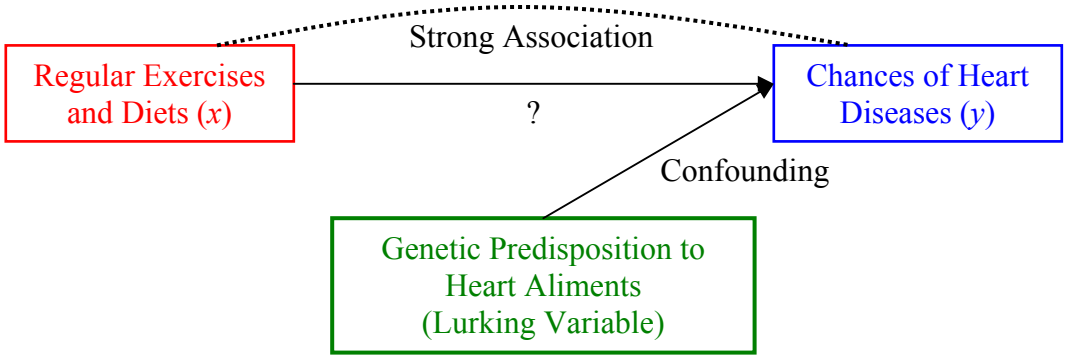
b. Common Response: - y can be predicted from x using the regression, but changing x might not change y definitively. This is due to the presence of **lurking variable that is affecting BOTH the x and y-variables**.

Example 4: - a research by the police department shows that fatality rate in traffic accidents has a strong positive correlation with gas mileage of vehicles. However, a common response lurking variable of vehicle weight can affect both the exploratory (gas mileage) variable and the responding (fatality rate). As vehicle weight increases, gas mileage and fatality rate decreases.



c. Confounding - the effect of x on y is mixed up with the effect of a lurking variable. This time **the lurking variable is affecting on the y-variable ONLY**. As such, it makes the relationship (if any) become unclear.

Example 5: - a study shows that regular exercises and proper diet will lower the chance of heart diseases. However, people who are genetically predisposed to heart aliments are going to develop health problem even if they exercise and maintain a proper diet. In this case, the lurking variable is the level of genetic predisposition to heart disease and it is affecting the y-variable (chance of heart disease) only.



4.2 Assignment:
pg. 214–215 #4.25 and 4.27

4.3: Relations in Categorical Data

Two-Way Table: - a table that describes two categorical variables.

Row Variable: - a variable situated in a row that describes one categorical variable.

Column Variable: - a variable situated in a column that describes the other categorical variable.

Marginal Distribution: - the distribution of the row variable or the column variable that are found at the bottom and right margin of the table.

Conditional Distribution: - the distribution that satisfies a specific condition (basically a specific marginal distribution).

Note: When graphing a complete two-way table, it is most appropriate to use a double bar chart to show the two categorical variables. The relative frequencies of each categorical variable must add up to 100%. However, when describing relationships among categorical variables (like an element out of a particular column or row total), we have to calculate appropriate percentage from the counts given. The total percentages of these percentages from the counts given may not add up to 100%.

Example 1: Competing brands, *Coke*, *Pepsi*, and *Safeway* sells soda in 355 mL cans, 750 mL cans, and 2L bottle. The sales table for the three companies for the month of May at a local Safeway supermarket is as follows.

Soda Sales at a local Safeway for the month of May

	Container Sizes			
Soda Brands	**355 mL**	**750 mL**	**2 L**	**Total**
Coke	850	500	400	1750
Pepsi	750	600	425	1775
Safeway	900	450	525	1875
Total	2500	1550	1350	5400

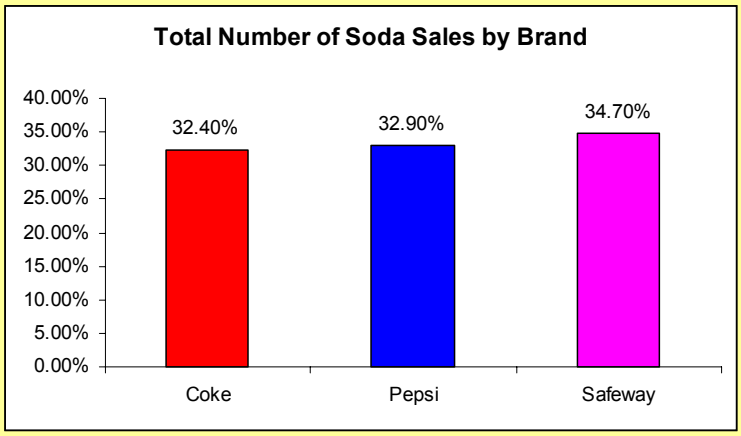
a. Identify the two categorical variable and find the percentage marginal distribution of each.
b. Create two bar charts using the percentages of marginal distribution for each categorical variable.
c. Determine the percentages of 355 mL, 750 mL and 2 L containers that were Safeway brands. Create a bar chart for the data found.

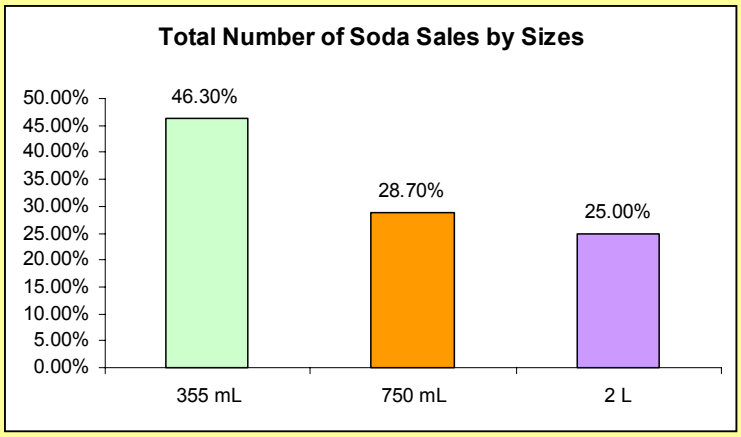
a. Identify the two categorical variables and find the percentage marginal distribution of each.

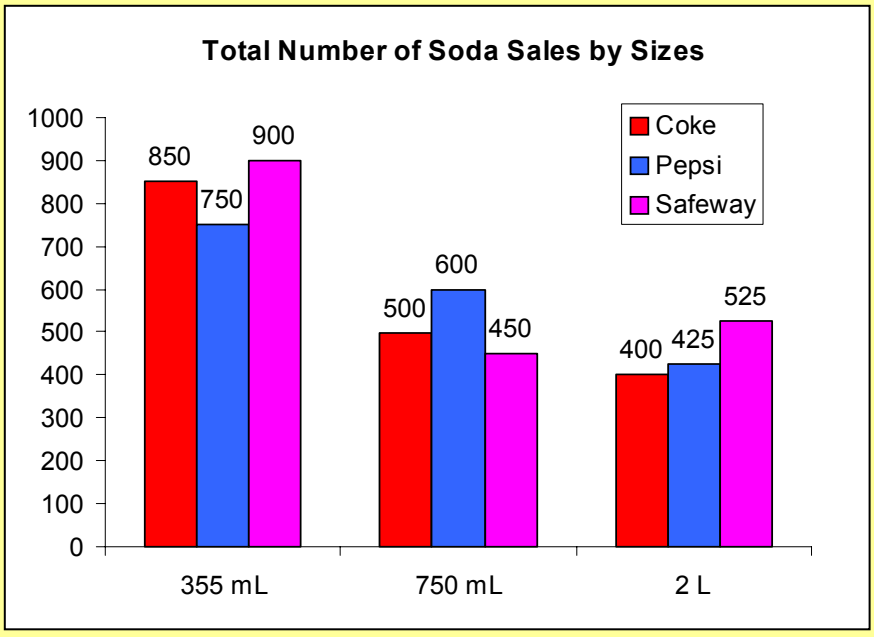
The two categorical variables are the **soda brand** and the **container sizes**.

Soda Brands	Container Sizes: 355 mL	750 mL	2 L	Total
Coke	850	500	400	$\frac{1750}{5400} = 32.4\%$
Pepsi	750	600	425	$\frac{1775}{5400} = 32.9\%$
Safeway	900	450	525	$\frac{1875}{5400} = 34.7\%$
Total	$\frac{2500}{5400} = 46.3\%$	$\frac{1550}{5400} = 28.7\%$	$\frac{1350}{5400} = 25.0\%$	5400

b. Create two bar charts using the percentages of marginal distribution for each categorical variable.







c. Determine the percentages of 355 mL, 750 mL and 2 L containers that were Safeway brands. Create a bar chart for the data found.

The percentage of 355 mL that were Safeway Brand

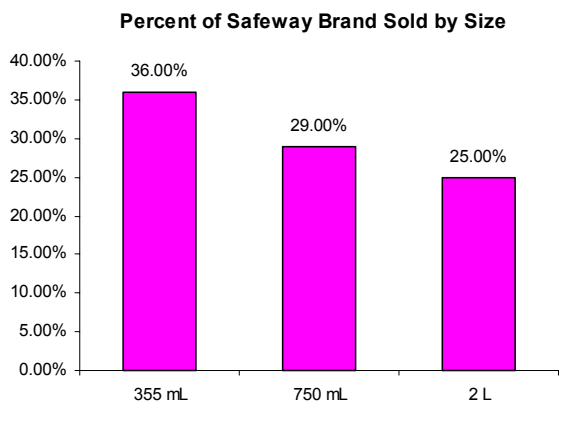
$$\frac{900}{2500} = 36.0\%$$

The percentage of 750 mL that were Safeway Brand

$$\frac{450}{1550} = 29.0\%$$

The percentage of 2 L that were Safeway Brand

$$\frac{525}{1350} = 38.9\%$$



Simpson's Paradox: - the comparison of the data is reversed when separate groups of similar data are combined to form a single aggregated (collective) group.

Example 2: A certain college was found to admit more male students than female students in its law and business faculties.

Business and Law Faculties Admittance Rate

	Admit	Deny
Male	980 (70%)	420 (30%)
Female	560 (56%)	440 (44%)

The college were charged with denying more female applicants than male applicants. The college administration produces the data to defend the charge

Business Faculty Admittance Rate

	Admit	Deny
Male	960 (80%)	240 (20%)
Female	360 (90%)	40 (10%)

Law Faculties Admittance Rate

	Admit	Deny
Male	20 (10%)	180 (90%)
Female	200 (33%)	400 (67%)

Explain why this is a Simpson's Paradox.

An examination of the data shown that when the admittance rate is combined for both the business and law faculties, it clearly reflect a bias against female applicants. However, a closer evaluation when the two faculties' admittance rates are separated from the aggregated data indicates they both admit more female applicants than male. This is a Simpson's Paradox due to the fact the combined data of both faculties reflect a different result versus the data from two separate groups.

4.3 Assignment:
pg. 217 #4.31; pg. 221–222 #4.35 and 4.39; pg. 225 #4.41

Chapter 4 Review:
pg. 233 #4.57