describing the data we have

A framework to describe quantitative data: *goal-write a paragraph*
Describe the <u>Shape</u>, <u>Center and Spread</u>, and <u>Unusual Features</u>   *to summarize a data set*

<u>Shape</u> – How is the data distributed? We need to see a (picture to) determine the shape.

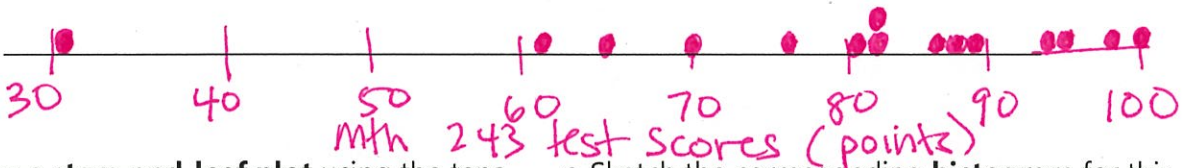**There are four types of graphs for quantitative variables**
- Dot Plot
- Stem-and-leaf Plot
- Histogram
- Boxplot

Example 1: Here is a set of 15 exam scores for a fictional MTH 243 Statistics class at PCC.

31, 62, 65, 70, 76, 81, 82, 82, 87, 88, 89, 94, 95, 98, 100

a. Draw a **dot plot** for this data.   *Include titles and units for all graphs.*


Mth 243 test scores (points)

b. Draw a **stem-and-leaf plot** using the tens digits as the stem and the ones digits as the leaves.

stems   *leaves*

```
3 | 1
4 |
5 |
6 | 2 5
7 | 0 6
8 | 1 2 2 7 8 9
9 | 4 5 8
10 | 0
```

Mth 243 test Scores (points)

Key
3|1 means 31 or 3.1

Split stems p.48
back-to-back p.70.

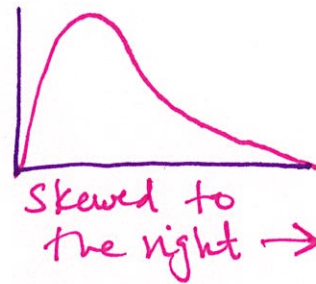c. Sketch the corresponding **histogram** for this data using a bin width of 10. Scale and label your graph appropriately.
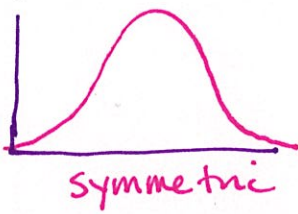

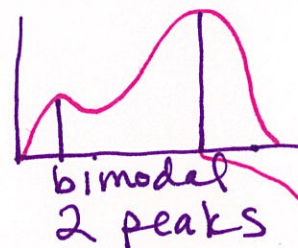*Mth 243 test scores*
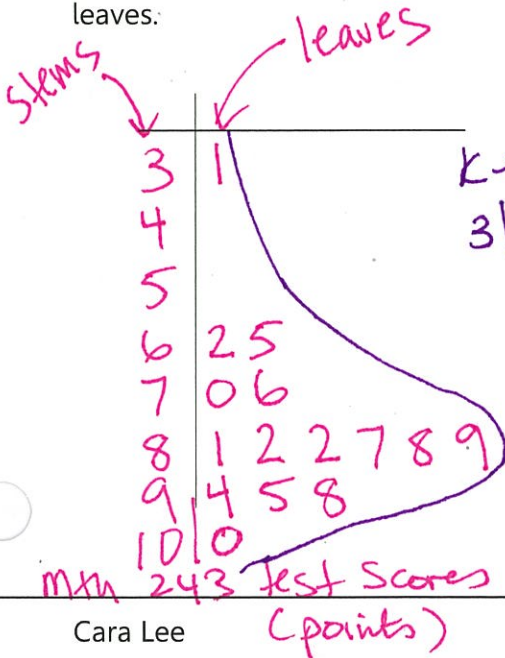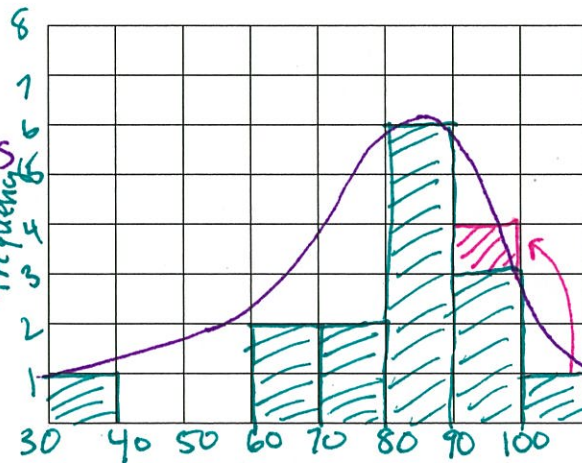
GeoGebra won't go over the max value

Describing Shape

- Symmetric, skewed left or skewed right

symmetric   Symmetric Uniform   Skewed to the left   Skewed to the right →

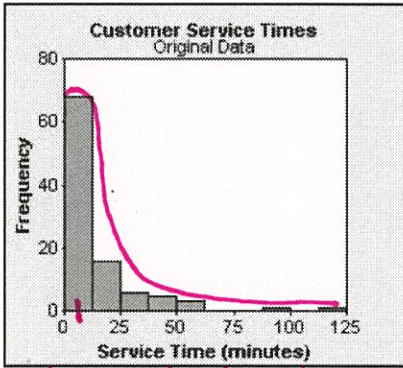- Modes: no mode, unimodal, bimodal, multimodal

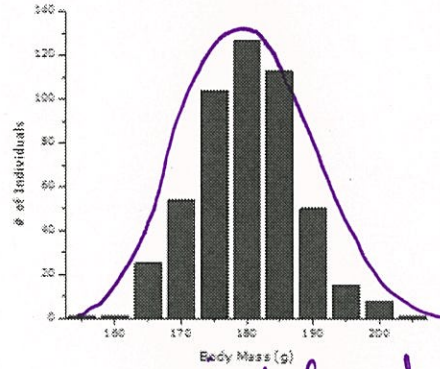no mode no peak   unimodal one peak   bimodal 2 peaks   multimodal
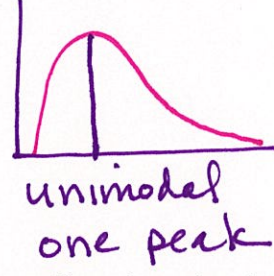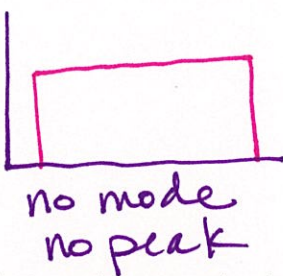
Example 2: In the examples below we'll explore the shape of various histograms. Is it symmetric? What is the shape of the histogram? Where are the modes (if it has any)?

The mode is the center of the peak.

Customer Service Times
Original Data

Skewed to the right and unimodal with the mode at 6.25 minutes.

Unimodal and approximately symmetric. The mode is 180 grams.
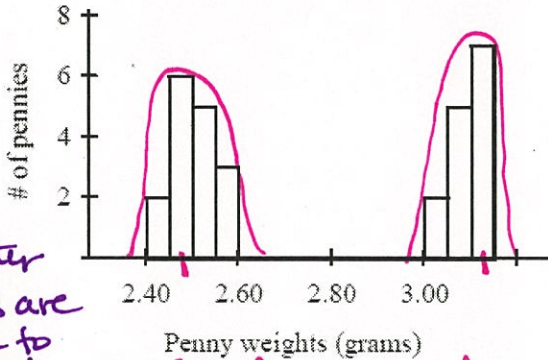
The lighter pennies are skewed to the right. The heavier pennies are skewed to the left.

Bimodal, with modes at 2.475 g and 3.175 g. Usually means that 2 populations have been grouped together - separate into 2 groups.

Phillip Morris Monthly Stock Return, July 1990 - May 1997

Unimodal and slightly skewed to the left. The mode is 2.5%.

<u>Center</u> and <u>Spread</u> – These should always be reported together.

<u>Center</u> is the position or location of the data (The <u>average or typical value</u>).
<u>Spread</u> is how much variation is in the data or how spread out the data is.
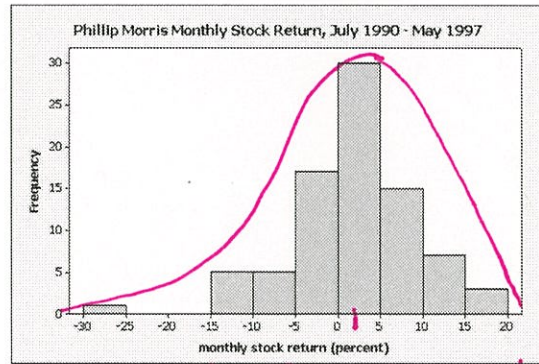
There are two different sets of measures for center and spread:

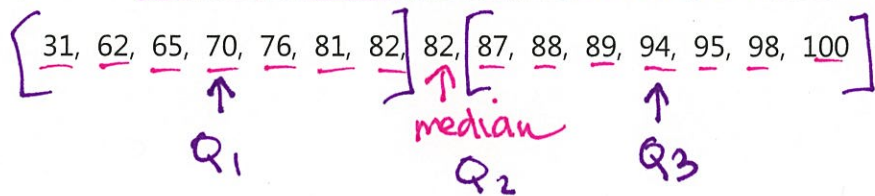| For skewed distributions:	For symmetrical distributions:
Center: Median Spread: IQR (Interquartile Range)	Center: Mean, $\bar{x}$ Spread: Standard Deviation, s → *next packet*

this packet

Median and IQR

Find the median and quartiles of the data. This divides the data into four equal sections.

even # of numbers

$81, 8\!\not{\mathrel{,}}82 \,|\, 87, 88, 89$

$\dfrac{82+87}{2}$

$= 84.5$

Example 1 continued: Using the previous data set (repeated here), find the median and quartiles by hand. The data is already in order. If there is an odd number of values, do not include the median when finding the quartiles. <u>If there is an even number of values, average the middle two.</u>

$$\Big[\, 31,\ 62,\ 65,\ 70,\ 76,\ 81,\ 82 \,\big|\, 82, \big[\, 87,\ 88,\ 89,\ 94,\ 95,\ 98,\ 100 \,\Big]$$

$\uparrow$ Q_1 *median* Q_2 $\uparrow$ Q_3

The median (Q_2) is the middle value or __50__ th percentile. __50__ % of the data are below that value.

The first quartile (Q_1) is the __25__ th percentile. __25__ % of the data are below that value.

The third quartile (Q_3) is the __75__ th percentile. __75__ % of the data are below that value.

5-Number Summary:

Min	31 **points**
Q_1	70 **points**
Median (Q_2)	82 **points**
Q_3	94 **points**
Max	100 **points**

31, 70, 82, 94, 100 **points**

Interquartile Range or IQR (Spread):
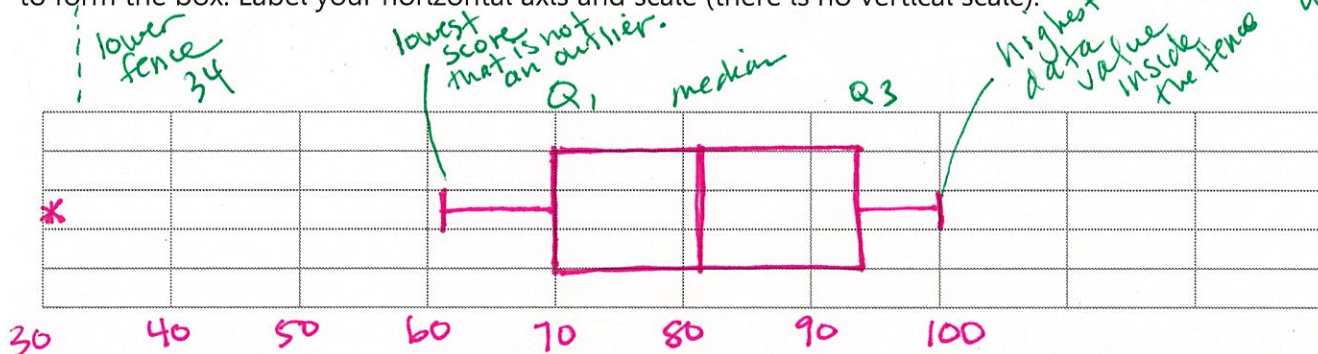The <u>width of the middle 50%</u> of the data

$IQR = Q_3 - Q_1$

$94 - 70 = 24$ **points**

✱ The IQR is a better measure of spread than the range (max - min) because it is less sensitive to outliers. $100 - 31 = 69$ **points**

Cara Lee

How to draw a Boxplot. Some books call this a modified boxplot because outliers are shown.

1. Draw the Box. Determine the scale and draw vertical lines at the Median, Q_1 and Q_3. Connect these to form the box. Label your horizontal axis and scale (there is no vertical scale).

lower fence 34 *lowest score is not an outlier.* Q_1 *median* Q_3 *highest data value inside the fence* *upper fence 130*

30 40 50 60 70 80 90 100

Test scores of Math 243 students in points.

2. Calculate the Fences using 1.5*IQR.

We use 1.5 times the Interquartile range on each side of the box to determine the fences. Any data values outside of the fences are considered outliers. The whiskers are drawn to the <u>nearest data values inside each fence.</u>

Upper Fence: $Q_3 + 1.5*IQR$ $\qquad 94 + 1.5(24) = 130$ points

Lower Fence: $Q_1 - 1.5*IQR$ $\qquad 70 - 1.5(24) = 34$ points

Draw in the fences lightly or below the graph. They are not technically part of the graph.

3. Draw the whiskers to the nearest data value inside each fence and make a short vertical bar. Label each value outside the fences (outliers) with a dot.

Activity 1. Use **GeoGebra** to calculate the 1-variable statistics. *GeoGebra.org downloads classic 5*

    View > Spreadsheet > Type in your data (or copy and paste)

    Select your data > Click on the histogram for **One Variable Analysis** > Click **Analyze**

Use **GeoGebra** to draw a histogram and boxplot.

    Use the dropdown menu to select **Histogram.** Change the **options** using the icon in the upper right corner

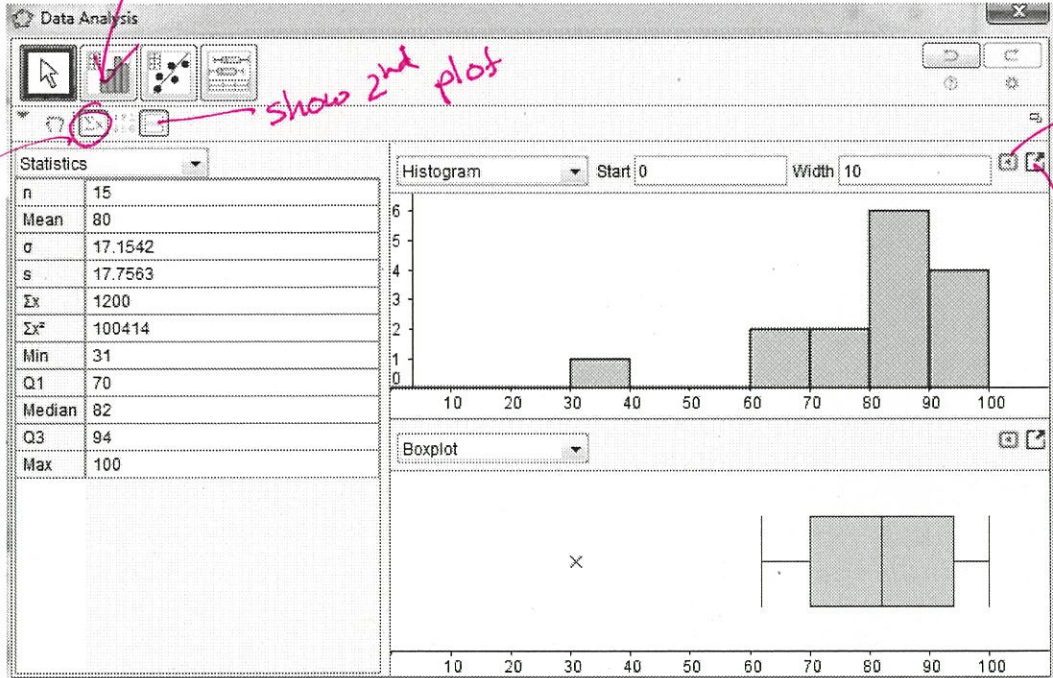   Select **Show 2nd Plot** and select **Boxplot.** Use the same scale for both plots so they match.

GeoGebra Results. Select **Copy to Clipboard** to insert a graph into a document.

(handwritten annotations:) 1-variable statistics · shows stats · show 2nd plots · set classes manually · copy to clipboard · set x-min, x-max, x-scale

Data Analysis

Statistics	
n	15
Mean	80
σ	17.1542
s	17.7563
Σx	1200
Σx^2	100414
Min	31
Q1	70
Median	82
Q3	94
Max	100

Histogram — Start 0 — Width 10

Boxplot

(handwritten: average)

How do the mean, median and mode relate to the shape of the distribution? (2.6)

Symmetric Distributions

(handwritten: use the mean for symmetric data)

↑ mean median mode
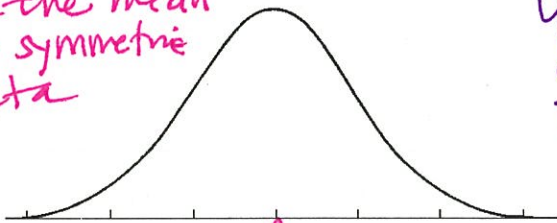
Skewed Distributions

Test for the shape

if the mean and median are close, then the distribution is approximately symmetric.

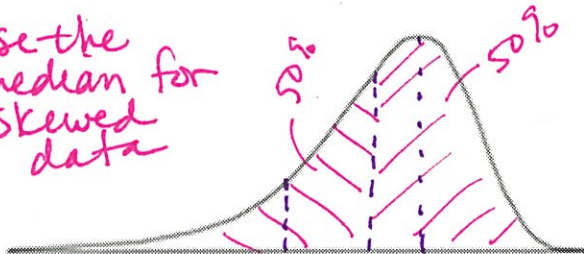if the mean is below the median, the data are skewed to the left.

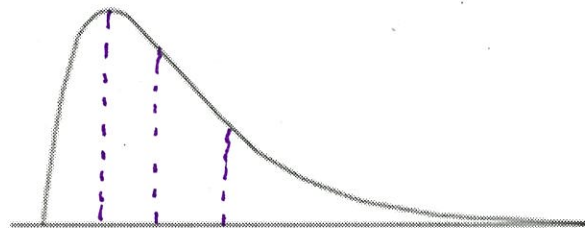if the mean is higher than the median, the data are skewed to the right.

Skewed left   Skewed right

(handwritten: use the median for skewed data)
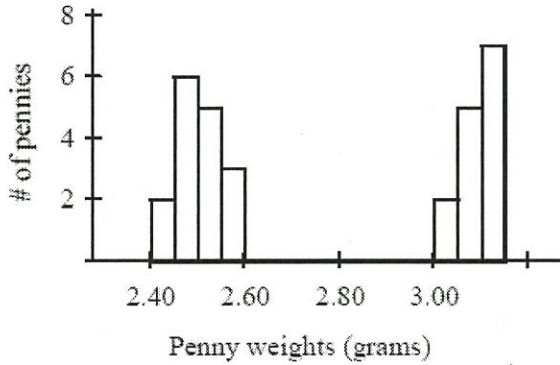
50%   50%

mean → median ↑ mode ↑

mode ↗ median ↗ mean ↑

Unusual Features:

Mention anything unusual about the data or state that there aren't any unusual features.
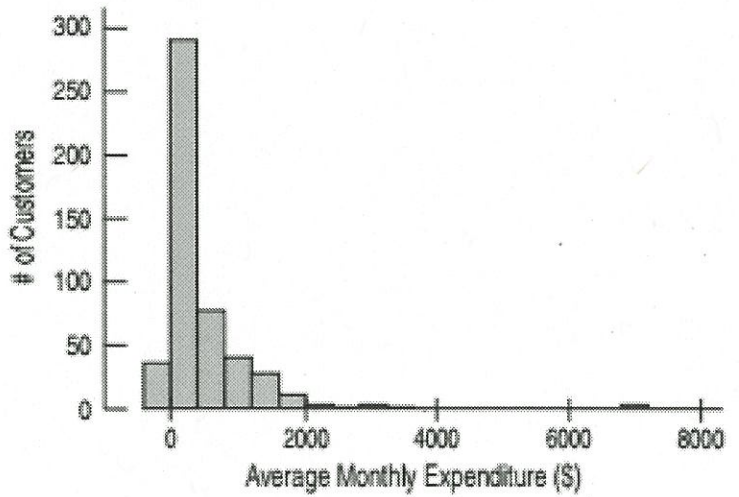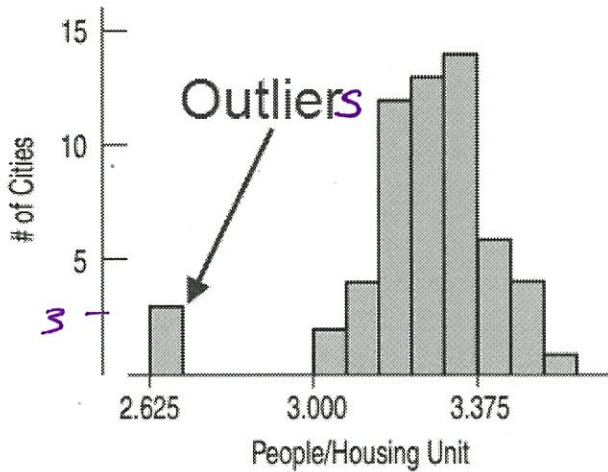- Multiple modes (look for subpopulations)
- Gaps and Outliers

Multiple Modes:

Bimodal – split into 2 populations

Gaps and Outliers:

Should you toss out outliers? *No, look into those cases – identify which cases are outliers and why. If no unusual features, state that.*

Include context

Example 1 Continued: Continuing with our test score data set, write a paragraph describing the distribution. Be sure to talk about the shape, center and spread, and any unusual features (or say that there are none).

(the mode is about 85 points)

The distribution of test scores is unimodal *(shape)* and skewed to the left. The median test *(center + spread)* score is 82 points and the IQR is 24 points. There was one score that was an outlier *(unusual features)* at 31 points.

Practice: The US Bureau of Transportation Statistics reports data on airline flights. Let's look at data giving the <u>percentage of flights cancelled each month between 1994 and 2013.</u> Use a few sentences to describe how often flights are cancelled. Be sure to talk about the shape, center and spread, and any unusual features (or say that there are none).

2.39 - 1.16 = 1.23%

Count	238
Max	20.24
Q3	2.39
Median	1.68
Q1	1.16
Min	0.38
IQR	1.23

(# of Months vs Percent of Flights Cancelled)
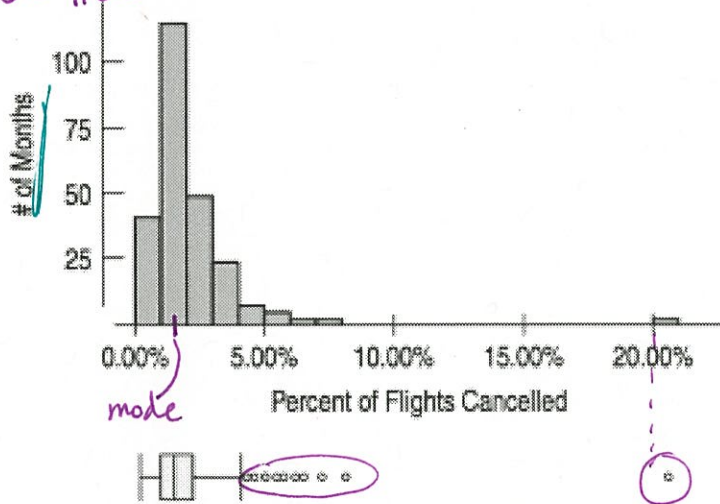
mode

The distribution of the percentage of flights canceled each month between 1994 and 2013 is unimodal and skewed to the right. The mode is about 1.5% of flights canceled in a month. The median is 1.68% of flights canceled, with an IQR of 1.23%. There was one month with over 20% of flights canceled and another grouping of *(months)* outliers with about 4-9% of the flights canceled.

Complete Summary of Test Scores for a Math 243 Class

<u>Summary Statistics</u>
Mean: 80 points
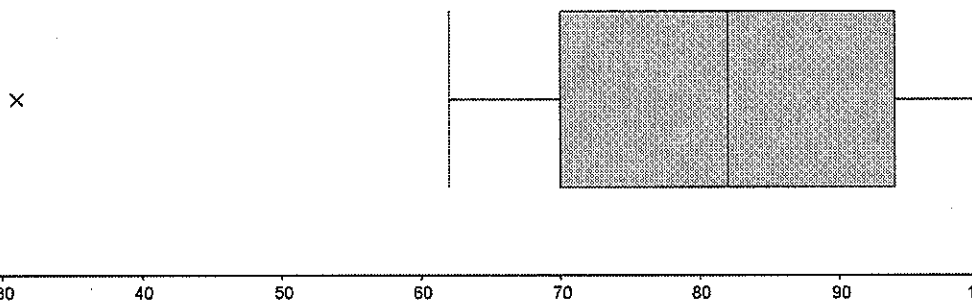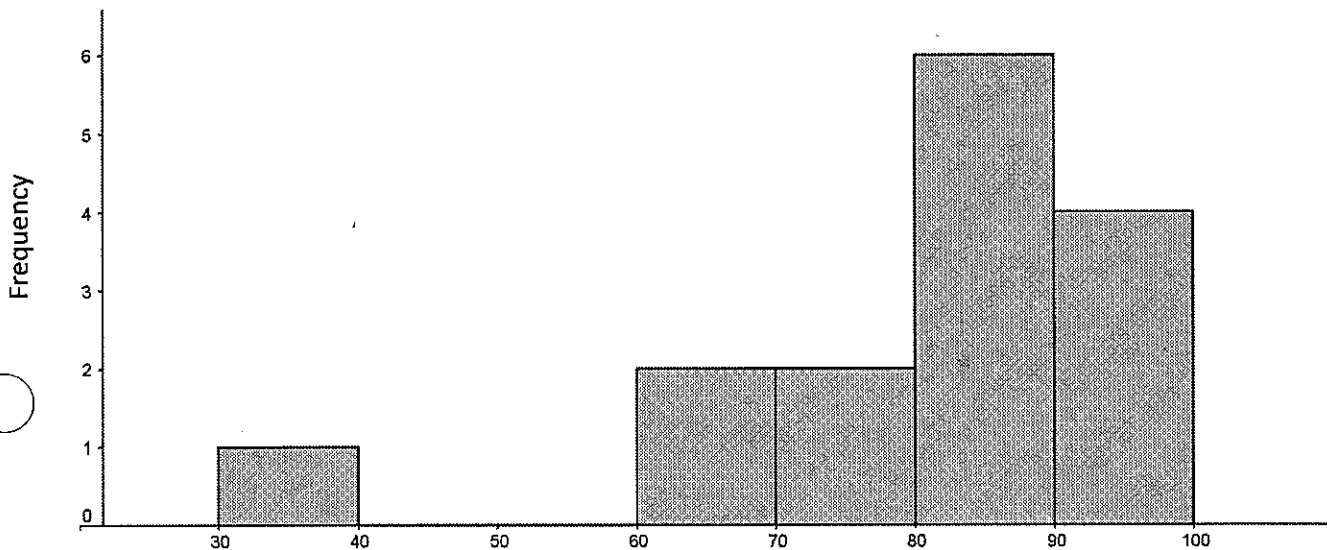Standard Deviation: 17.76 points
Min: 31 points
Q1: 70 points
Median: 82 points
Q3: 94 points
Max: 100 points

Test Scores for Math 243 Students (Points)

The distribution of test scores is unimodal and skewed to the left. The mode is between 80 and 90 points. The median is a score of 82 points and the width of the middle 50% is 24 points (interquartile range). There is one outlier score of 31 points to look into. The student may have gotten behind or had an extremely bad day.