# General Introduction

The Purpose of statistics: Statistics has many uses, but perhaps its most important purpose is to help us make decisions about issues that involve uncertainty.

## Definition of Statistics:

**1.** Numerical Facts

1. Average price for one-bedroom apartment at the city of Rocklin is $ 895.
2. 80% of Sierra students graduate in 2 years.

**2. C O D A** **C**ollection, **O**rganization, **D**escription, **A**nalysis, and interpretation of data.

**C**ollection: **D**ata **S**ampling

**O**rganization: **F**requency Table (Bar-chart, Pie-chart),
**H**istogram, **F**requency **P**olygon, **O**give **C**urve

**D**escription: **M**ean, **M**ode, **M**edian,
**R**ange, **V**ariance, **S**tandard **D**eviation **SD**,
**Q**uartiles, **P**ercentiles, **B**ox Plot

**A**nalysis: **C**orrelation and **R**egression, **E**stimation,
**T**est of **H**ypothesis, Analysis of **V**ariance

## Types of Statistics:

**Descriptive**: **C**ollection, **O**rganization, **D**escription

**Inferential**: **A**nalysis and interpretation of data

## What is the statistics all about?

***1.*** It is about how we test if a new drug is effective in treating cancer.
***2.*** It is about opinion polls, pre-election polls, and exit polls.
***3.*** It is about sports, where we rank players and teams primarily through their statistics.
***4.*** It is about the market research and the effectiveness of advertising
***5.*** It is about how agricultural inspectors ensure the safety of the food supply.

## Population versus Sample:

**Population**: Entire elements or subjects under study that share one or more **common characteristic** such as age, gender, major or race. (Keyword all/every), All college students, All Sierra College students, All male Sierra College Students who are taking statistics and majoring in business. Two Elements: **T**ime and **P**lace

**Sample:** A portion of population.

**Census**: The collection of data from every element in a population.

**Parameter vs. statistic:** A numerical measurement describing some characteristic of a
**P**opulation (called parameter) vs. a **S**ample (called statistic)
Hint: Use Greek Alphabet for parameter and lower case English for statistic.
$\mu$ = avg. $\sigma$ (sigma) = st. dev $\chi^2$ = Chi-squared…. $\bar{X}$, s, r

***Extra Practice:*** *Answer questions **A** from **page 1** of practice problem **part 1***

## Types of Data:

**Qualitative** (Names, Labels ...) pass / fail, democrat/republican/independent, yes/no, grades (A,B,C,D,F)

**Quantitative**: **1**. **D**iscrete (Countable): number of accidents in Rocklin each day, number of emergency call to 911 center each day, number of students that will pass Abe stat class

**2**. **C**ontinuous (Measurable): Speed, weight, time, capacity, length, volume, area

***Extra Practice:*** *Answer questions **B** from **page 1** of practice problem **part 1***

## Types of Sampling: R_S_S_C_C

**1**. **R**andom: Every member of population has equal chance to be selected.
**How**? Every member will be assigned a different number, and we select random numbers by a computer or a table and match those with the members' numbers.

**2. S**ystematic: We select some starting point and then select every kth (such as every $20^{th}$) member in the population.
**How?** Every $10^{th}$ customer or client will be selected to be asked questions.

**3. S**tratified: **Subdivide** the population into at least two different subgroups (strata) sharing the same characteristics (such as gender or age bracket), then we draw a sample from each stratum.
**How? a)** divide the police officers in Sacramento into male and female group
**b)** select a random sample of each and collect data regarding the years in service.

**4. C**luster: Divide the population into sections (or clusters), and then randomly select some of those clusters, then choose all the members from those selected clusters.
**How?** To see the customer feedback to a new menu
**a) divide** Sacramento in different zones,
**b) randomly** select some of those zones
**c)** collect data from **all** fast-food branches in those selected zones.

**5**.**C**onvinence: Use the results are readily available.
**How?** A math instructor asks some of his students if they use student solution manual to do their homework.

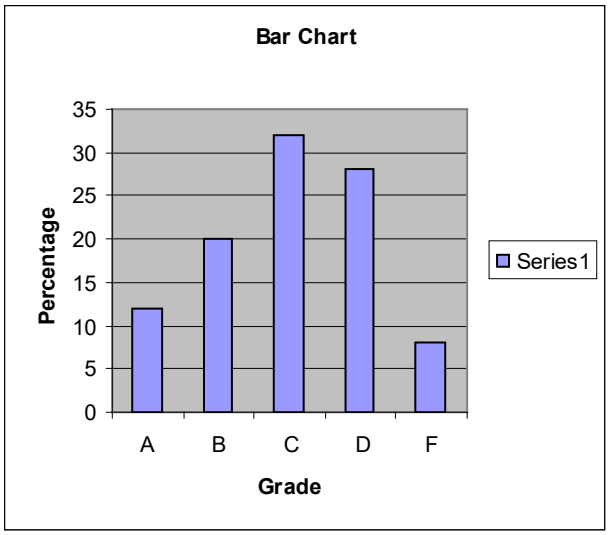
***Extra Practice:*** *Answer questions **C** on **page 1** of practice problem **part 1***

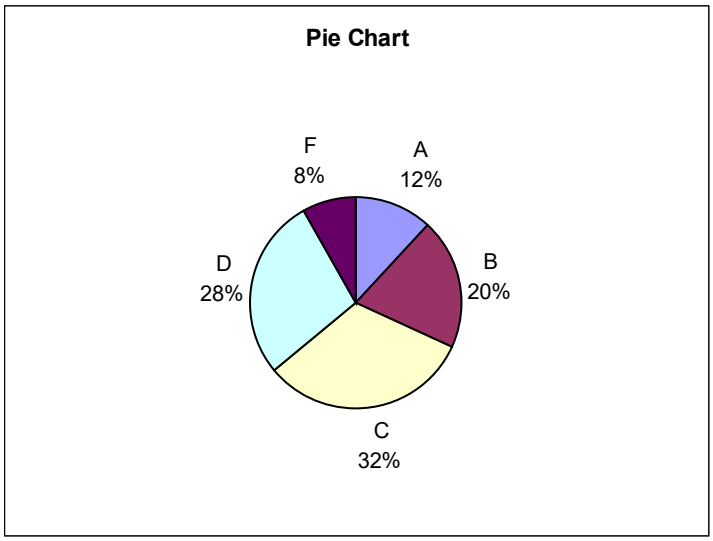
# Qualitative Data

**Example 1.**

| Grade | $f$ = Students | Rel. freq % $\frac{f}{n}\times 100$ | Angles $360^\circ$ ( Rel. freq) |
|---|---|---|---|
| **A** | **6** | $(6/50)\times 100 = \mathbf{12}$ | $.12\times 360 = \mathbf{43.2}^\circ$ |
| **B** | **10** | $(10/50)\times 100 = \mathbf{20}$ | $.20\times 360 = \mathbf{72}^\circ$ |
| **C** | **16** | $(16/50)\times 100 = \mathbf{32}$ | $.32\times 360 = \mathbf{115.2}^\circ$ |
| **D** | **14** | $(14/50)\times 100 = \mathbf{28}$ | $.28\times 360 = \mathbf{100.8}^\circ$ |
| **F** | **4** | $(4/50)\times 100 = \mathbf{8}$ + | $.8\times 360 = \mathbf{28.8}^\circ$ + |
| | $n = \sum f = 50$ | 100? | $360^\circ$ ? |

**Using Excel to graph the followings**




## Practice 1:

Complete the table and draw the bar chart and the pie chart.(You can use Microsoft Excel to do the graphs)

| Grade | $f$ = Students | Rel. freq % $\frac{f}{n}\times 100$ | Angles $360^\circ$ ( Rel. freq) |
|---|---|---|---|
| **A** | **22** | | |
| **B** | **26** | | |
| **C** | **20** | | |
| **D** | **8** | | |
| **F** | **4** | + | + |
| | $n = \sum f =$ | 100? | $360^\circ$ **?** |

## Descriptive Statistics

**A) Measure of Central Tendency** (**M**ean, **M**edian, **M**ode)

**Mean** ( $\mu$ , $\bar{x}$ )  $x = data$  $\sum$ $= Sum$  N or n = Number of data points

Data: 5, 6, 3 , 9, 7  $\bar{x} = \frac{\sum x}{n} = \frac{5+6+3+9+7}{5} = \frac{30}{5} = 6$

**Median**: The middle data point in a ranked (largest to smallest or smallest to largest) data, **or**
The median cuts the ranked data in half **one half below** it and **one half above** it.

**Example1**: Suppose the median score for the first test was 73, it simply means half the class got below 73 and the other half above it. Examples for odd and even numbered of data.

2, 5, 11, 16, 8, 9, 3, 7, 5  Ranked  2, 3, 5, 5, **7**, 8, 9, 11, 16,  **Median** = **7**

2, 3, 5, 5, **7**, 8, 9, 11, 16, 4  Ranked  2, 3, 4, 5, **5**, **7**, 8, 9, 11, 16, **Median** = $(5+7)/2 = 6$

Hint: If there are extreme values in data set (too large or too low with respect of the rest of data) then median is a better than mean to identify the measure of central tendency.

**Mode**: The data value(s) with the highest occurrence, bimodal, multimodal

2, **8**, 11, 7, **8**, 13  **Mode** = 8

3, **12**, 5, 14, 9, **12**, **7**, 16, **7**  **Bimodal** = 7, 12

**11**, **15**, 7, **2**, 6, 16, **15**, 3, **2**, **11**, 19, 5, 4  **Multimodal** = 2, 11, 15

---

**B) Measure of Positions** (**Q**uartiles, **B**ox-Plot, **P**ercentile, **Z**-score)

Quartiles: Breaking the **ranked data** in 3 quartiles (**Q1, Q2, Q3**)

**Data:** _____25%_____**Q1**_____ 25%_____**Q2**_____25%_____**Q3**_____25%_____

How to find quartiles? 3 steps
**Rank** the data, Find **Q2** = Median, Find the new medians **Q1, Q3** on either side of Q2.

**Example 1**: **Odd** number of data  Data: 2, 5, 11, 16, 8, 9, 3, 7, 5, 4, 13

Ranked Data: 2, 3, ,**4**, 5, 5, **7**, 8, 9, **11**, 13, 16,
Q1 Q2 Q3

**Example 2**: **Even** number of data  Data: 2, 3, 5, 5, 7, 8, 9, 11, 16, 4
Ranked Data 2, 3, **4**, 5, 5, 7, 8, **9**, 11, 16,  **Q2 = Median** = $(5+7)/2 = 6$
**Q1** **Q2** = 6 **Q3**

---

**TI-83/84** *Inputting data in **L1** (stat → Option 1 → enter)*
***then*** *stat → calc → Option 1 → enter → 2n d → 1 → enter*

---

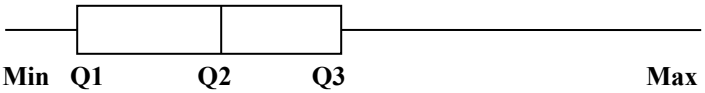
***Extra Practice**: Answer questions on columns **A-G** on **page 3** of practice problem **part 1***

**Box-Plot:** is mainly used for ungrouped data to show how the data are distributed by showing **center**, **spread**, and **skewness**. **Center** is the **Q2**, **Spread** is how wide the box is, **Skewness** explains the distribution of the data by using the longer tail to describe the **Skewness** (for example if the longer tail is on the right, it is called skewed to the right)

To construct a boxplot
**1**. Find the 5-**number summery** of the data that are **Min, Q1, Q2, Q3, Max**
**2**. Plot these points on a **scaled** number line.
**3**. Construct a box by using Q1, Q2, Q3

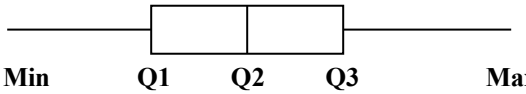
There are many possibilities of where the box in box-plot may be located.

If the box in box-plot is located to the **far Left**, it suggests that distribution of data are **skewed** to the **Right**



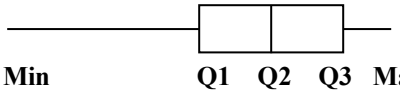

**Skewed** to the **Right**

If the box in box-plot is located to the **Center**, it suggests that distribution of data are **Centered**.




**Centered**

If the box in box-plot is located to the **far Right**, it suggests that distribution of data are **skewed** to the **Left**




**Skewed** to the **Left**

**To see how to do Box-Plot by TI, look at the Youtube links on class website.**

***Extra Practice:*** *Answer questions on columns **A-G** on **page 3** of practice problem **part 1***

---

**TI-83/84** *Inputting data in **L1** (stat → Option 1 → enter)*
***then*** *stat → calc → Option 1 → enter → 2n d → 1 → enter*

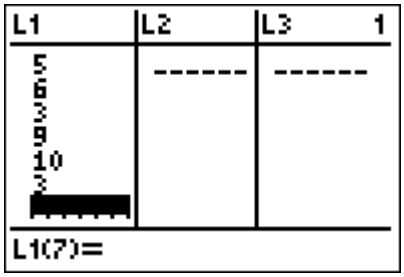
---

***Extra Practice:*** *Answer questions on columns **A-G** on **page 3** of practice problem **part 1***

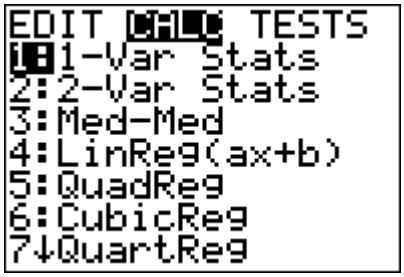
# TI-83/84

Find the mean, median, Q1, Q3 and standard deviation for 5, 6, 3 , 9, 10, 3, and also draw the Box_Plot.

*Inputting data in **L1** (stat → Option 1 → enter)*

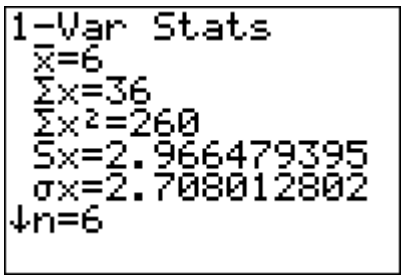



*stat → calc → Option 1 → enter*

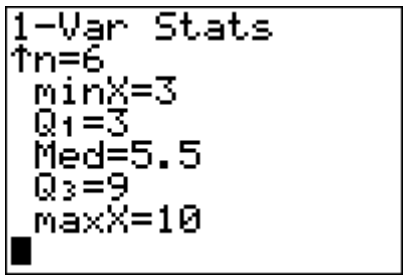



*2n d → 1*


*Results*

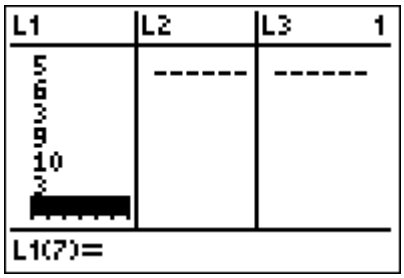



*Use down arrow for more Results*

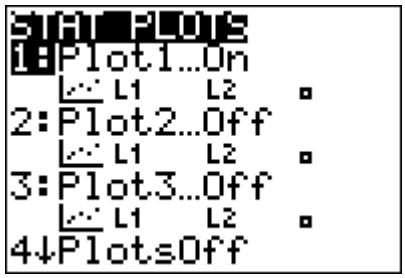



**Doing the Box Plot by TI**

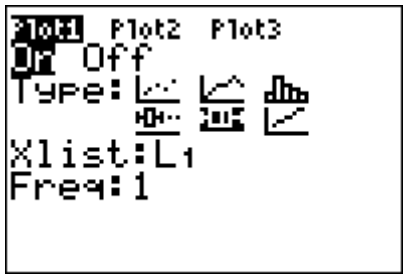
***Inputting data in L1***




***2nd STAT Plots***

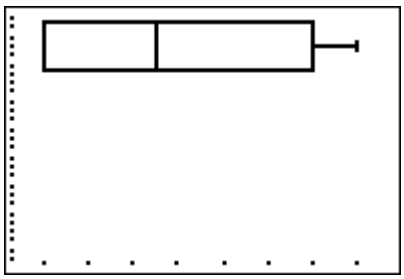


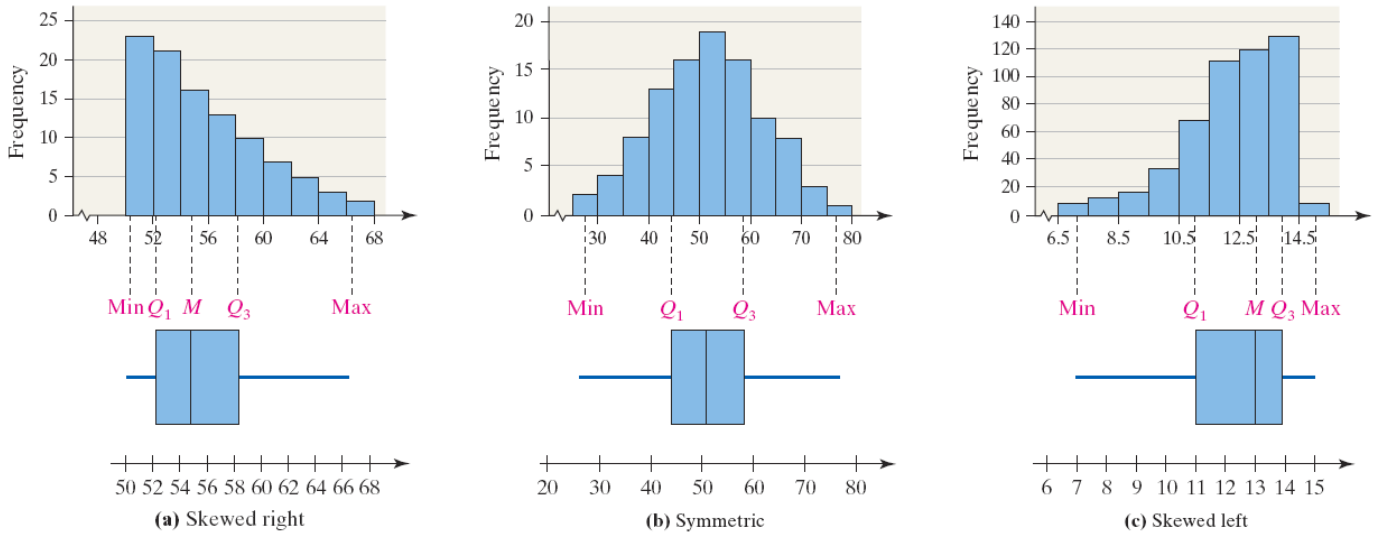

***Choose the fifth option***




***Press ZOOM 9***

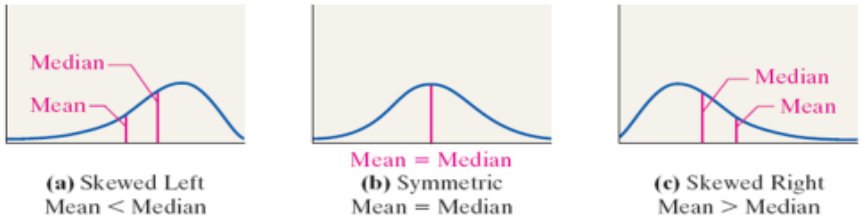

***Result***

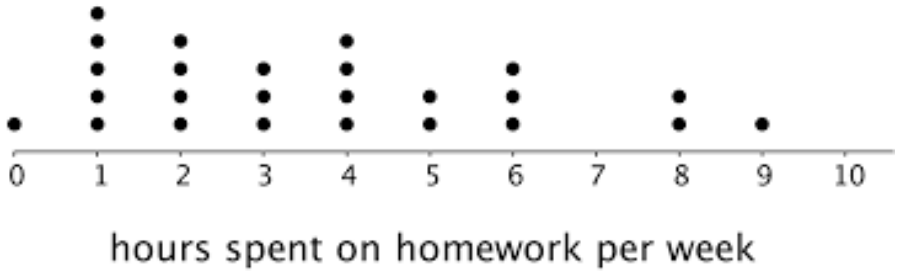





**Relation Between the Mean, Median, and Distribution Shape**

| Distribution Shape | Mean versus Median |
|---|---|
| Skewed left | Mean substantially smaller than median |
| Symmetric | Mean roughly equal to median |
| Skewed right | Mean substantially larger than median |




**The dot-plot above shows how many students spending how many hours in homework.**

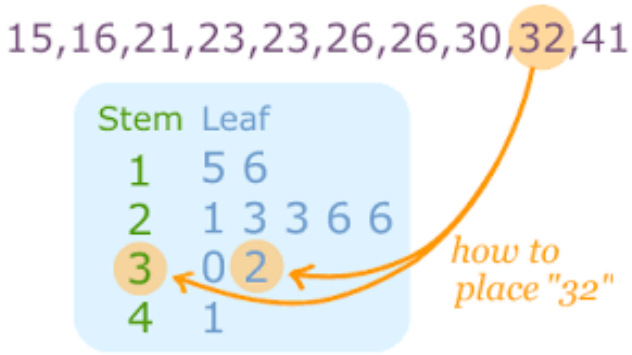
## Construction of a Stem-and-Leaf Plot

**Step 1:** The stem of a data value will consist of the digits to the left of the right-most digit. The leaf of a data value will be the rightmost digit.

**Step 2:** Write the stems in a vertical column in increasing order. Draw a vertical line to the right of the stems.

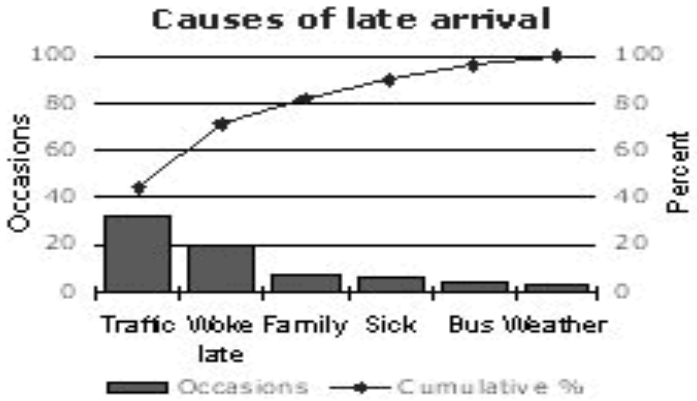
**Step 3:** Write each leaf corresponding to the stems to the right of the vertical line.

**Step 4:** Within each stem, rearrange the leaves in ascending order, title the plot, and provide a legend to indicate what the values represent.




| Class A Leaves | Stems | Class B Leaves |
|---|---|---|
| 8 0 | 6 | 0 0 |
| 5 0 | 7 | 0 1 3 3 5 6 7 |
| 6 4 | 8 | 4 5 6 |
| 6 4 4 2 1 0 | 9 | 1 2 |
| 0 0 | 10 | |

**The above stem and leaves plot is used to show the difference in scores between the two classes.**



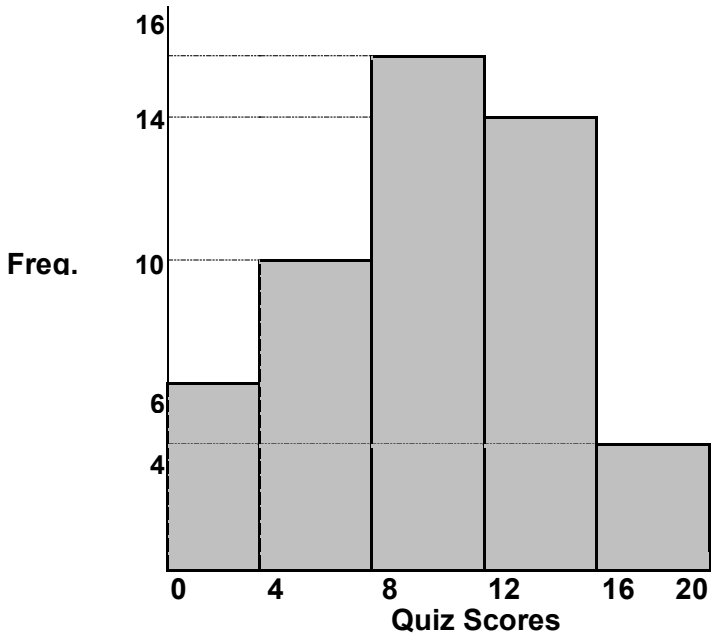

**Pareto Chart: It shows which factors are the major and minor causes for arriving late.**

## Grouped Data (Freq. Table)

The table below shows the quiz scores of 50 students that are given in group.

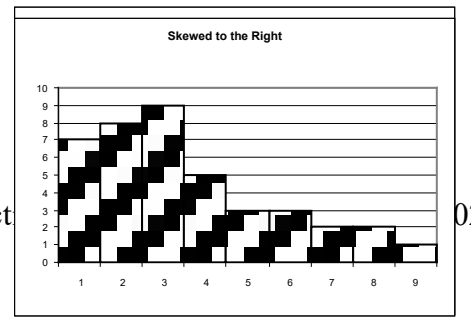
| Quiz Score | Freq ( $f$ ) = Students | | | |
|---|---|---|---|---|
| 0 – 4 | 6 | | | |
| 4 - 8 | 10 | | | |
| 8 – 12 | 16 | | | |
| 12 – 16 | 14 | | | |
| 16 – 20 | 4 | | | |

Use the quiz scores on x-axis, frequency on the Y-axis to draw blocks for a shape that is called **Histogram**

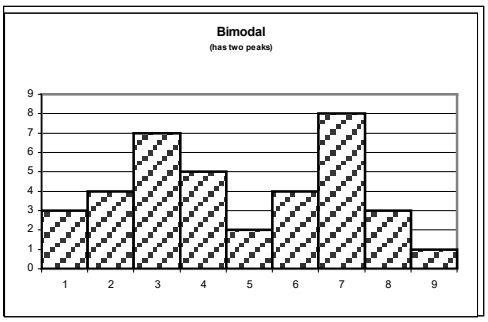



Histogram looks close to a Centered or bell-shaped distribution.

### Different possible shapes of Histogram




Sect 020



**Mean** and **Standard Deviation**.

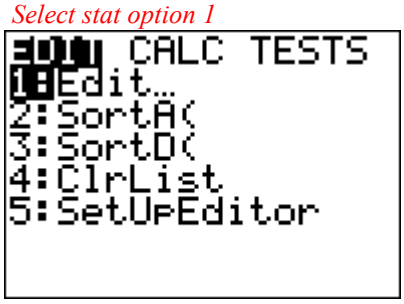
First step is to create a new column called **midpoint** (average of scores in each group). For example, for 0 – 4, the midpoint will be 2, for 4 – 8, the midpoint will be 6. Next step is to open two new columns $f \times m$ and $f \times m^2$ do the necessary calculations, find the summation for each and then use them in the given formulas.

| **X-axis** | | **midpoint** | **Mean** | |
|---|---|---|---|---|
| **Quiz Scores** | **Freq( $f$ )= Students** | m | $f \times m$ | |
| 0 – 4<br>4 - 8<br>8 – 12<br>12 – 16<br>16 – 20 | 6<br>10<br>16<br>14<br>4 | 2<br>6<br>10<br>14<br>18 | 12<br>60<br>160<br>196<br>72 | |
| | $\sum f = n = 50$ | | $\sum f \times m = 500$ | |

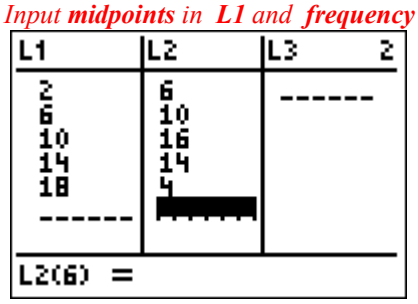
Mean: $\bar{X} = \dfrac{\sum f \times m}{n} = \dfrac{500}{50} = 10$

## TI-83/84

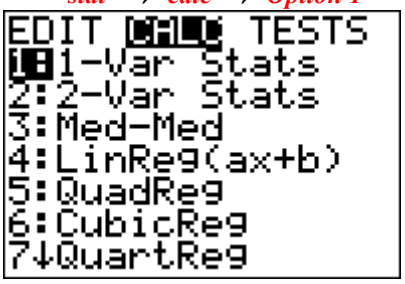
*Select stat option 1*




*Input **midpoints** in **L1** and **frequency** in **L2***



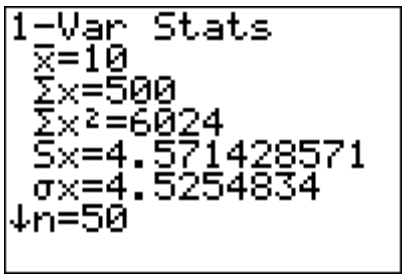

***stat*** → ***calc*** → ***Option 1***




$2^{nd}$ 1 , $2^{nd}$ 2


Press enter


Results



**Practice 1: Use both formula and the Ti to find the mean, standard deviation, and the variance.**

| **Quiz Scores** | **Freq($f$)** | m | $f \times m$ | |
|---|---|---|---|---|
| 0 – 10<br>10 - 20<br>20 – 30<br>30 – 40 | 8<br>12<br>14<br>6 | 25 | 180 | |
| | $\sum f = n =$ | | $\sum f \times m =$ **780** | |

Mean: $\bar{X} = \dfrac{\sum f \times m}{n} =$ _____ = 19.5

**Practice 2: Use both formula and the Ti to find the mean, standard deviation and the variance**

| **Test Scores** | **Freq ($f$)=** | m | $f \times m$ | |
|---|---|---|---|---|
| 0 – 20<br>20 – 40<br>40 – 60<br>60 – 80<br>80 – 100 | 2<br>8<br>14<br>32<br>24 | | | |
| | $\sum f = n =$ | | $\sum f \times m =$ | |

Mean: $\bar{X} = \dfrac{\sum f \times m}{n} =$ _____ = 67

***Extra Practice**: Answer questions **A, B, C, D** on **pages 5, 6** of practice problem **part 1***