Part I (Section 2) Statistics

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 points, **first** find **Q2** = median **and then** new medians **Q1, Q3** on either side of Q2.

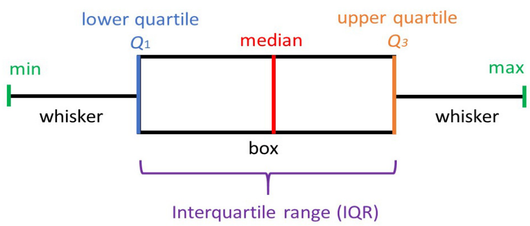
Example 1: **Odd** number of data 2, 5, 11, 16, 8, 9, 3, 7, 5, 4, 13

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

Example 2: **Even** number of data points 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**

[Quartiles calculator](#) [*Boxplot graph maker*](#)

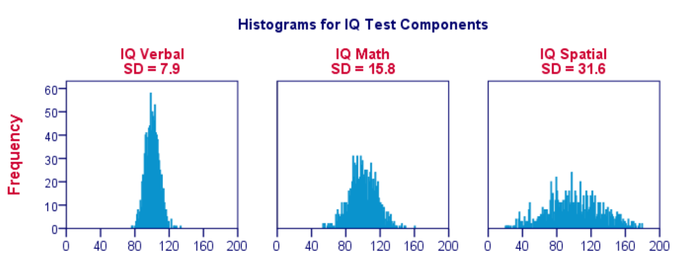


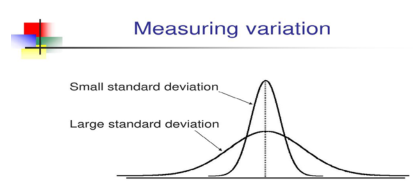
Extra Practice: Answer questions on columns A-G on page 1 from practice problem part 1

C) Measure of Variation (**R**ange, **S**tandard Deviation, **V**ariance)

Range: The difference between the lowest and highest values. In {4, 6, 9, 3, 7} the lowest value is 3, and the highest is 9, so the *range* is 6

Standard Deviation (σ, s) : It measures the **average dispersion** of data **around the mean.**
Observation: By looking at left image you see three IQ scores, the **Verbal score has the least variability** or spread around mean (smallest standard deviation and **spatial scores had the largest variability** or spread around mean (largest standard deviation). The right image shows the larger standard deviation results in wider curve due more dispersion.





Example: Consider the 3 random delivery time (in days) taken by 2 different companies **X**, and Y for shipment from point A to B.

	X	**Y**
Mean	5	5
Median	5	5
Mode	5	none

At first it seems there are not that much of difference between the delivery times of these two companies but let's look at their actual data and their plots on Dot-Plot.

	X	Y
Delivery time	5	5
Delivery time	5	0
Delivery time	5	10

X **Dot Plot** **Y**

```
X:          x
            x
____  _____x________
  0         5        10

Y:  x________x________x
    0        5        10
```

Now, it seems that there is **no dispersion** for company X, but an **average dispersion of 5** for company **Y**, suggesting that company X is more reliable meeting the average delivery time.

How to find standard deviation for a given set of data such as 5, 3 , 9, 10, 3? Using $s = \sqrt{\dfrac{\sum (x-\bar{x})^2}{n-1}}$

Step 1) Create a table with 4 columns with the following headings as, x, $\bar{x}$, $(x-\bar{x})$, $(x-\bar{x})^2$

Step 2) list the data in column 1 from the far left and find the mean for data and put it next column.

$5+3+9+10+3=30$ $\qquad \bar{x}=30/5=6$

Step 3) Subtract the mean from each data and put it in the next column.
Step 4) Square the numbers in the third column and put them into the last column
Step 5) Find the summation of the last column.

x	$\bar{x}=6$	$(x-\bar{x})$	$(x-\bar{x})^2$
5	6	-1	1
3	6	-3	9
9	6	3	9
10	6	4	16
3	6	-3	9
$\sum x=30$			$\sum (x-\bar{x})^2$ =44

Step 6) Divide the summation of last column **by** $n-1$ and take square root

$$s = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}} = \sqrt{\frac{44}{5-1}} = \sqrt{11} = 3.31$$

[**Sample Standard deviation calculator**]

. **Variance** (s^2): Variance is the **square of standard deviation** **Variance** = $s^2 = 3.317^2 = 11$.

If the source of data was **population**, use a σ notation and divide the summation of last column by n

$$\sigma = \sqrt{\frac{\sum (x-\bar{x})^2}{n}} = \sqrt{\frac{44}{5}} = \sqrt{8.8} = 2.96 \qquad \sigma^2 = 2.96^2 = 8.8$$

[**Population Standard deviation calculator**]

Estimating standard deviation. If we do not have any data in hand, use this formula: $s = range / 4 = (\max - \min) / 4$

Example: You found out that the shortest and the longest time to get to work from home was respectively 40 to 55 minutes, use this information to estimate standard deviation. $s = range / 4 = (55-40)/4 = 15/4 = 3.75$

Extra Practice: Answer questions on columns A-G on page 1 from practice problem part 1

Practice: Find **standard deviation and variance** for a given set of data 6, 3 , 7, 4

x	$\overline{x} = 5$	$(x - \overline{x})$	$(x - \overline{x})^2$
6	5		
3	5		
7	5		
4	5		
$\sum x =$			$\sum (x - \overline{x})^2 = 10$

$s = 1.826$ and variance is $s^2 = 1.826^2 = 3.334$

Extra Practice: Answer questions on columns A-G on page 1 from practice problem part 1

What is the **Empirical Rule**?

The empirical rule, also referred to as the three-sigma rule or 68-95-99.7 rule, is a statistical rule which states that for a normal distribution, almost all observed data will fall within three standard deviations (denoted by σ) of the mean or average (denoted by μ).

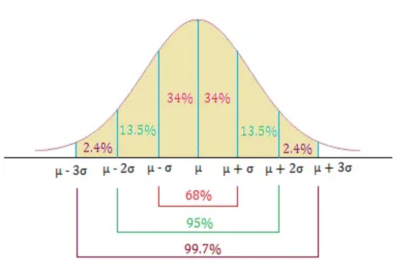
In particular, the empirical rule predicts that 68% of observations falls within the first standard deviation (μ ± σ), 95% within the first two standard deviations (μ ± 2σ), and 99.7% within the first three standard deviations (μ ± 3σ).

Empirical Rules is valid if and only if the **boxplot or histogram is centered**, here are the **three** empirical rules.

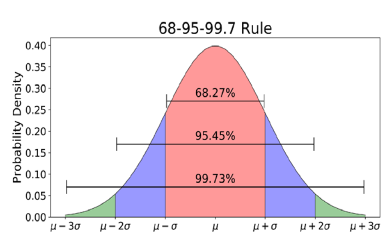
$68\% = \overline{x} \pm \sigma$ or $68\% = \overline{x} \pm s$ **68 %** of data are within 1σ *or* s of the mean ($\overline{x}$)

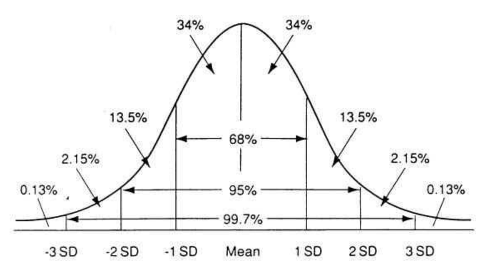
$95\% = \overline{x} \pm 2\sigma$ or $95\% = \overline{x} \pm 2s$ **95 %** of data are within 2σ *or* $2s$ of the mean ($\overline{x}$)

$99.7\% = \overline{x} \pm 3\sigma$ or $99.7\% = \overline{x} \pm 3s$ **99.7 %** of data are within 3σ *or* $3s$ of the mean ($\overline{x}$)



Empirical rule





Example: Find all three empirical rules for Abe Stat class if the average was 72 and the standard deviation was 8, assuming that Boxplot was centered.

$68\% = 72 \pm 1(8) = 72 \pm 8$ $64 <$ **68 %** of class got scores between < 80

$95\% = 72 \pm 2(8) = 72 \pm 16$ $56 <$ **95 %** of class got scores between < 88

$99.7\% = 72 \pm 3(8) = 72 \pm 24$ $48 <$ **99.7 %** of class got scores between < 96

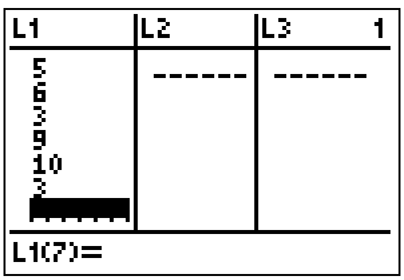
Practice: Find **the mean and standard deviation** for a given set of data such as **5, 3 ,6, 9, 10, 3**.
Use the method discussed in page 2 to get the following answers:
mean $\bar{x} = 6$ **and standard deviation** $s = 2.97$

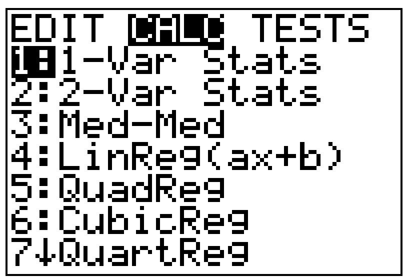
This page is explaining how to get the answers from last practice problems by TI 83/84.

TI-83/84

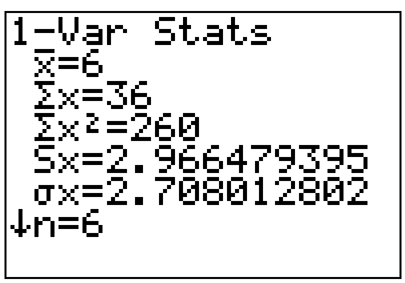
Find the mean, median, Q1, Q3 and standard deviation for 5, 6, 3 , 9, 10, 3, and also draw the Box-Plot.

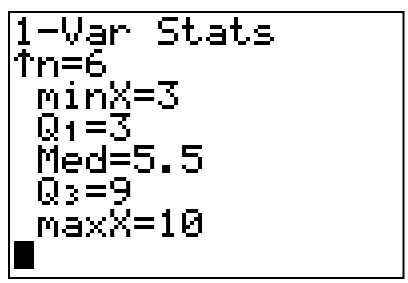
*Inputting data in **L1** (stat $\rightarrow$ Option 1 $\rightarrow$ enter)* *stat $\rightarrow$ calc $\rightarrow$ Option 1 $\rightarrow$ enter*





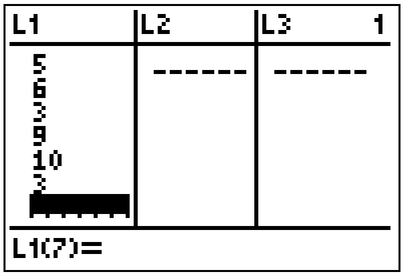
2n d $\rightarrow$ 1 **enter** ***Results*** ***Use down arrow for more Results***



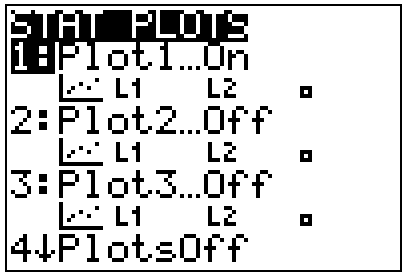


Doing the Boxplot by TI

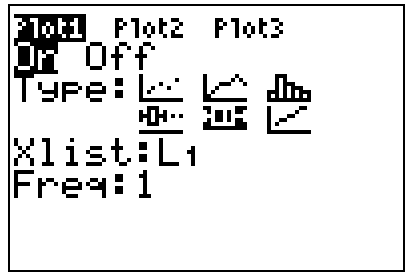
Inputting data in L1



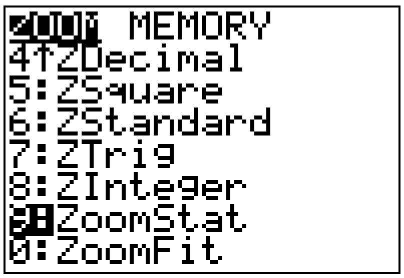
2nd STAT Plots



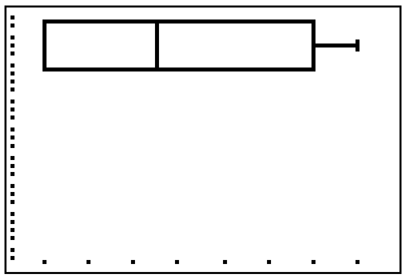
Choose the fifth option



Press ZOOM 9



Result

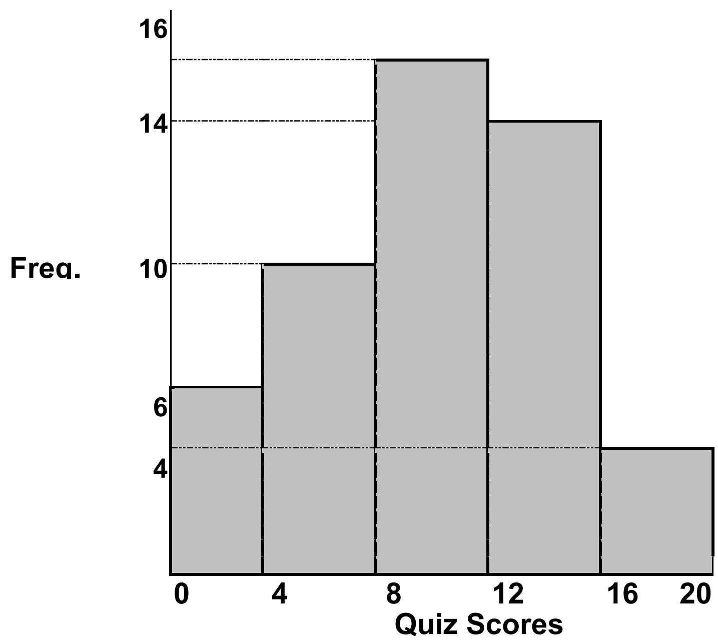


Grouped Data (in Frequency table form)

The table below shows the quiz scores of 50 students that are given in 5 groups.

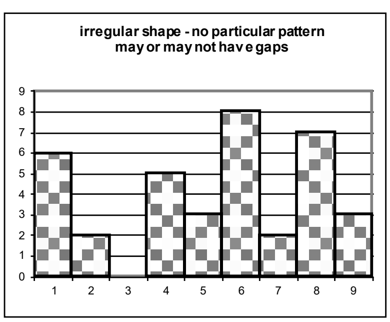
Quiz Score	Freq (f) = Students			
0 – 4	6			
4 - 8	10			
8 – 12	16			
12 – 16	14			
16 – 20	4			

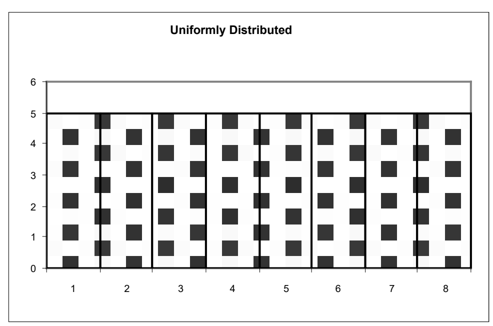
Use the quiz scores on x-axis, frequency on the Y-axis to draw blocks for a shape that is called **Histogram**

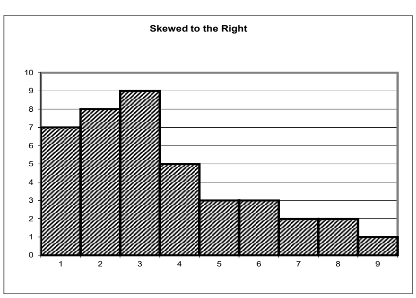


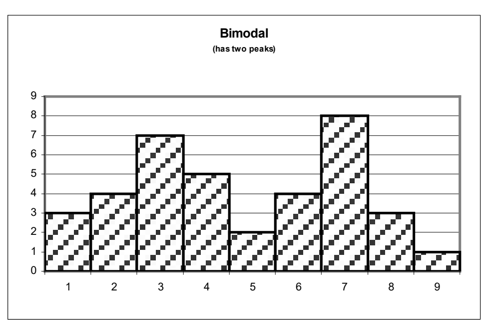
Histogram looks close to a Centered or bell-shaped distribution.

Different possible shapes of Histogram









Mean and **Standard Deviation**.

Steps in finding the mean and standard deviation from a frequency table.

1) You need a table with 5 columns the first two columns from the left always will be given.
2) You need to create have three columns as m, $f \times m$ **and** $f \times m^2$
3) The third column is called midpoint = m and to find it to find the average of the two numbers from first column
4) The fourth column $f \times m$ is the product of frequency of each row and its midpoint.
5) The fifth column $f \times m^2$ is the product is the product of fourth column one more time by m
6) You need to get the summation of the 2nd, 4th, and 5th columns.
7) You need these summations to plug into the mean and standard deviation formulas given below, look at the arrows as how where each summation goes.

Mean: $\bar{X} = \dfrac{\sum f \times m}{n} =$ Sample Standard deviation: $s = \sqrt{\dfrac{n\sum f \times m^2 - \left(\sum f \times m\right)^2}{n(n-1)}} =$

Column 1	**Column 2**	**Column 3**	**Column 4**	**Column 5**
Quiz Scores	**Frequency (f)= Students**	m = **midpoint**	$f \times m$	$f \times m^2$
0 – 4 4 - 8 8 – 12 12 – 16 16 – 20	6 10 16 14 4	$(0+4)/2=2$ $(4+8)/2=6$ $(8+12)/2=10$ $(12+16)/2=14$ $(16+20)/2=18$	$6\times2=12$ $10\times6=60$ $16\times10=160$ $14\times14=196$ $4\times18=72$	$6\times2^2=24$ $10\times6^2=360$ $16\times10^2=1600$ $14\times14^2=2744$ $4\times18^2=1296$
	$\sum f = n = 50$		$\sum f \times m = 500$	$\sum f \times m^2 = 6024$

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

Sample Standard deviation: $s = \sqrt{\dfrac{n\sum f \times m^2 - \left(\sum f \times m\right)^2}{n(n-1)}} = \sqrt{=\dfrac{50(6024)-(500)^2}{50(50-1)}} = \sqrt{\dfrac{51200}{2450}} = 4.57$

Variance: $s^2 = 4.57^2 = 20.9$ **Calculator find mean and standard deviation for frequency table**

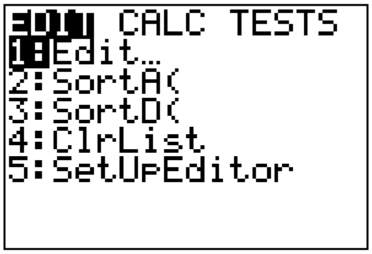
Apply 95% empirical rule:

$95\% = \bar{x} \pm 2\,s = 10 \pm 2(4.57) = 10 \pm 9.14$ $0.86 <$ **95. %** of class got scores <19.14

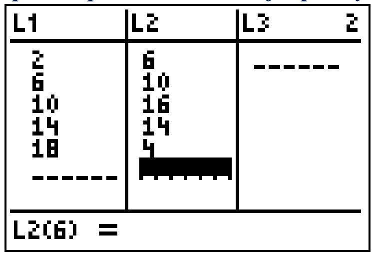
Extra Practice: Answer questions A-d on grouped data from practice problem part 1

TI-83/84

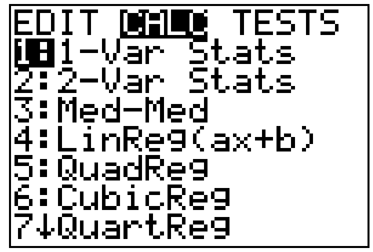
Select stat option 1



Input midpoints in L1 and frequency in L2



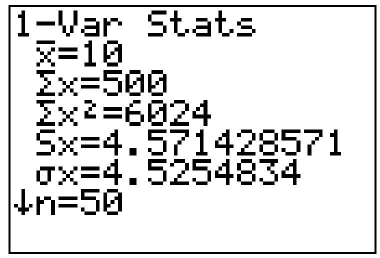
stat → calc → Option 1



2nd 1 , 2nd 2

Press enter

Results



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

Quiz Scores	**Frequency (f)= Students**	m	$f \times m$	$f \times m^2$
0 – 10 10 - 20 20 – 30 30 – 40	8 12 14 6	5 25 	40 180 	200 7350
	$\sum f = n = 40$		$\sum f \times m =$ **780**	$\sum f \times m^2 =$ **19000**

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

Standard deviation: $s = \sqrt{\dfrac{n\sum f \times m^2 - \left(\sum f \times m\right)^2}{n(n-1)}} = \sqrt{\underline{\qquad\qquad\qquad}} = \sqrt{\underline{\qquad\qquad}} = 9.86$

Variance: $s^2 = 9.8^2 = 97.18$

Apply 68% empirical rule: **Answer**: < 68% of quiz scores are between <

Apply 95% empirical rule: **Answer**: < 95% of quiz scores are between <

Practice 2:
Please complete the table and find the mean, standard deviation, and the variance for different men shoe sizes.
Hint: because we do not have group ranges therefore, we use the sizes in first column as midpoint in the third column.

Men shoe sizes	**Freq (f)=**	m	$f \times m$	$f \times m^2$
6	4	6	24	$24 \times 6 = 144$
6.5	6	6.5	39	$6.5 \times 39 = 253.5$
7	8	7	56	$56 \times 7 = 392$
7.5	10	7.5		
8	12			
8.5	16			
9	14			
9.5	12			
10	8			
10.5	6			
11	4	11	44	$44 \times 11 = 484$
	$\sum f = n = 100$		$\sum f \times m =$	$\sum f \times m^2 =$

Mean: $\overline{X} = \dfrac{\sum f \times m}{n} = \underline{\qquad\qquad} =$

Standard deviation: $s = \sqrt{\dfrac{n\sum f \times m^2 - \left(\sum f \times m\right)^2}{n(n-1)}} = \sqrt{\dfrac{\qquad\qquad}{\qquad\qquad}} = \sqrt{\dfrac{\qquad}{\qquad}} =$

Apply 68% empirical rule: **Answer**: < 68% of shoe sizes are between <

Apply 95% empirical rule: **Answer**: < 95% of shoe sizes are between <

Practice 3: Use both formula and the Ti to find the mean, standard deviation, and the variance

Test Scores	**Freq (f)=**	m	$f \times m$	$f \times m^2$
0 – 20	2	10	20	200
20 – 40	8	30	$8 \times 30 = 240$	$8 \times 30^2 = 7200$
40 – 60	14			
60 – 80	32			
80 – 100	24			
	$\sum f = n =$		$\sum f \times m =$	$\sum f \times m^2 =$

Mean: $\overline{X} = \dfrac{\sum f \times m}{n} = \underline{\qquad\qquad} = 67$

Standard deviation: $s = \sqrt{\dfrac{n \sum f \times m^2 - \left(\sum f \times m\right)^2}{n(n-1)}} = \sqrt{\underline{\qquad\qquad\qquad}} = \sqrt{\underline{\qquad\qquad}} = 20.89$

Variance: $s^2 =$

Apply 68% empirical rule: **Answer**: 46.11 < 68% of test scores are between < 87.89

Final Observation: **Mean, Median & Mode**

The **3Ms** of statistics that are commonly used are: **`mean`**, **`median`** and **`mode`**. All these 3 values make a lot of sense and gives the gist of the data scale and throws some **basic light on how the data is distributed** in terms of values and their frequency. Depending on the values or gaps between the 3Ms - the data is said to be Symmetrical or Asymmetrical in nature. Mean, Median, Mode and skewedness on different type of data distribution.