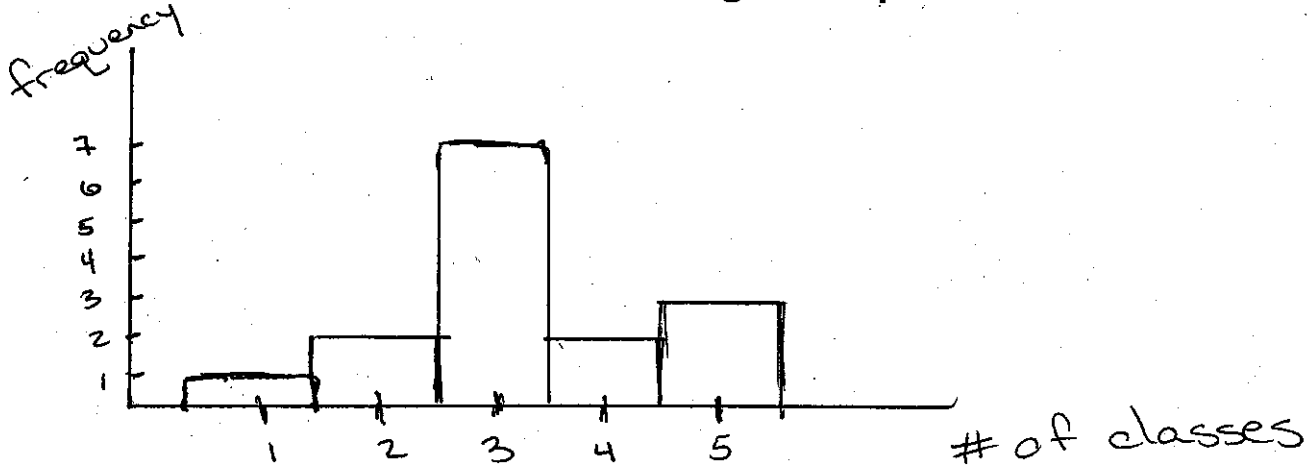
Ch 2: Statistical Graphs → Displaying data

Types of Graphs we will study:

- Dot Plot
- Histogram
- Stem and Leaf Plot
- Boxplot

1. Histogram
   a. By Hand

ex) We are interested in the number of classes a person in this class is taking this quarter.



* A histogram is like a bar chart, but the bars or boxes are contiguous (i.e. they touch each other)

Notes:
- Histograms are really useful for visualizing large data sets
- Horizontal axis = what the data repr. i.e. the variable
- Vert. axis = frequency or relative freq.

Stem and Leaf Plot

- Order data from smallest to largest
- Divide each data into stem and leaf
- Create plot

DATA: Scores on first exam for Precalculus

33	55	68	74	88	94
42	61	69	78	88	94
49	63	69	80	90	94
49	67	72	83	92	96
53	68	73	88	94	100
55					

The leaf is the last digit
The stem is everything before the last digit

Stem	Leaf
3	3
4	2 9 9
5	3 5 5
6	1 3 7 8 8 9 9
7	2 3 4 8
8	0 3 8 8 8
9	0 2 4 4 4 4 6
10	0

*Generally used only for small data sets

An outlier is a piece of data that doesn't fit with the rest

The **Median** is a number that separates the data into two halves. Half of the data is above the median and half of the data is below the median.

Calculating the median:

- Order the data from smallest to largest * Sort first

If there are an odd number of data values: the one in the middle is the median

DATA:

1, 1, 3, (5,) 6, 7, 9

$n = 7$

To find median: $\frac{n+1}{2}$

$\frac{7+1}{2} = 4 \rightarrow 4^{th}$ number

$\boxed{M = 5}$

If there are an even number of data values: the median is the average of the 2 middle #s

DATA:

1, 1, 3, (5, 6,) 7, 9, 9

$n = 8$

To find median: $\frac{n+1}{2}$

$\frac{8+1}{2} = 4.5 \rightarrow$ avg 4^{th} & 5^{th} #s

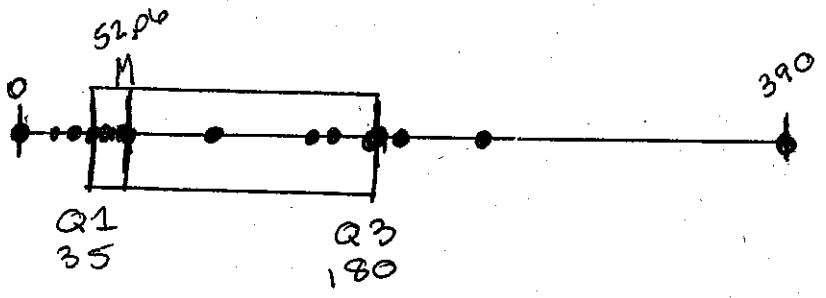
$M = \frac{5+6}{2} = \boxed{5.5}$

Boxplot

We are interested in the amount of money a student in this class spent on books for this quarter.

Five Number Summary

Smallest Value	0
Quartile 1	35
Median	52.06
Quartile 3	180
Largest Value	390



IQR = Interquartile Range = Q3 – Q1

ex) 180 − 35 = 145

The middle 50% of the data is between Q1 and Q3. in other words it's in the "box"

IQR and Outliers

If a data value is more than 1.5* ~~IRQ~~ IQR below Q1 or above Q3, it is considered to be an **outlier**.

data < Q1 − 1.5(IQR)

or data > Q3 + 1.5(IQR)

ex) Q1 − 1.5(IQR) = 35 − (1.5)(145)

= −182.50

No outliers this dir.

Q3 + 1.5(IQR) = 180 + 1.5(145)

= 397.50

No outliers

Boxplot on Calculator

33	55	68	74	88	94
42	61	69	78	88	94
49	63	69	80	90	94
49	67	72	83	92	96
53	68	73	88	94	100
55					

Min = 33
Q1 = 61
Median = 73
Q3 = 90
Max = 100

IQR = 90 - 61
= 29

Outliers?

$Q1 - 1.5(IQR) = 61 - 1.5(29)$
$= 17.5$

No outliers here

$Q3 + 1.5(IQR) = 90 + 1.5(29)$
$= 133.5$

No outliers

Ch. 2: Statistical Measures

A. Measures of the "Center" of Data

1. **Mean** = average

* Add all data values & divide by sample size

Population mean: μ = Greek letter "mu"

Sample mean $\bar{x}$ = "x bar"

The data below shows the number of paperback books bought by shoppers at a bookstore.

# of books	Freq
1	11
2	10
3	16
4	6
5	4
6	2
10	1

Total = 50

Calculate the mean $\bar{x}$:

$$\frac{(1\cdot 11)+(2\cdot 10)+(3\cdot 16)+(4\cdot 6)+(5\cdot 4)+(6\cdot 2)+10}{50} \approx 2.9 = \bar{x}$$

books per person

2. **Median** = middle #

ex) $n = 50$

$\frac{50+1}{2} = 25.5$

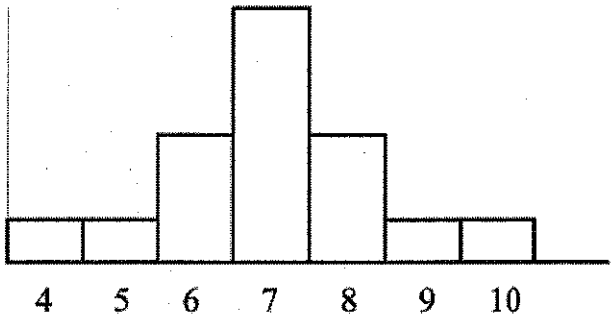
M = avg. of 25^{th} & 26^{th} data

$\boxed{M = 3}$

3. **Mode** = most frequent #

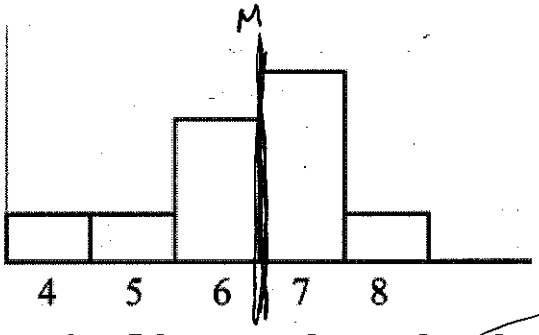
ex) $\boxed{\text{mode} = 3}$

Skewness of the Mean, Median, and Mode



$\bar{x} = 7$
$M = 7$
mode = 7

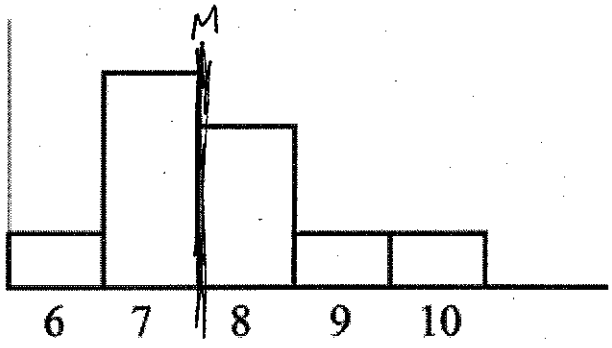
Data is **Symmetrical**
Mean = Median = Mode



$\bar{x} = 6.3$
$M = 6.5$
mode = 7

chopped off on R
long tail on L

Data is **Skewed to the Left**
Mean < Median < Mode



$\bar{x} = 7.7$
$M = 7.5$
mode = 7

long tail to R
chopped off on L

Data is **Skewed to the Right**
Mean > Median > Mode

Statistical Measures (continued)

C. Measures of the Location of Data

Percentiles

The n^{th} percentile is a number that n% of the data is below.

Note: Q1 = 25^{th} percentile
Q3 = 75^{th} percentile
Median = 50^{th} percentile

*We use CRF to find percentiles

Example: 50 students were asked the number of hours of sleep they had gotten the night before.

Amt Sleep (Hours)	Frequency	Relative Frequency	Cumulative Rel. Freq.
4	2	0.04	0.04
5	5	0.10	0.14
→ 6	7	0.14	0.28
→ 7	12	0.24	0.52
8	14	0.28	0.80
9	7	0.14	0.94
10	3	0.06	1.00

← 25%
← 50%
← 75%

Total: 50

What is the 28th percentile? * 28% get 6 hrs or less
We want the # that 28% of the data is less than

*We avg. 6 & the next highest data value i.e $\frac{6+7}{2} = 6.5$

What is the 75th percentile? What is another name for this?

52% are less than 8
80% are less than 9
~~80% is too much~~

So the 75th percentile is 8

What is the median? = 50th percentile

7

What is the first quartile? = 25th percentile

6

- What is the 80th percentile? $\frac{8+9}{2} = 8.5$

To calculate a percentile:

- If you see the percent in the ~~last~~ CRF column, average that data value in the first column and the next higher data value

- If you don't see the percent in the ~~last~~ CRF column, take the data value (in the first column) for the next higher percent

Percentiles

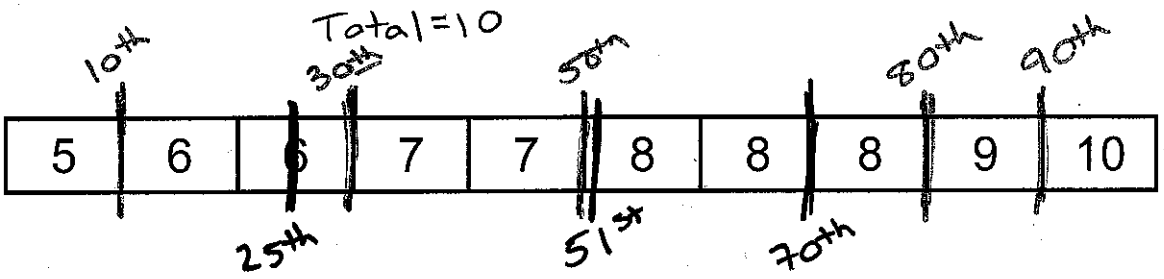
The n^{th} percentile is a number that n% of the data is **below**.

Example:

Data	Frequency	Rel. Freq.	CRF
5	1	0.1	0.1
6	2	0.2	0.3
7	2	0.2	0.5
8	3	0.3	0.8
9	1	0.1	0.9
10	1	0.1	1.0

← 0.25

← 0.51, 0.7



- 50^{th} percentile (median)? CRF 0.5 → 7

  $\frac{7+8}{2} = 7.5$

- 25^{th} percentile (Q1)? = 6

- 70^{th} percentile? = 8

- 51^{st} percentile? = 8

B. Measures of the Spread of Data → How much variation is there in the data?

1. Standard Deviation

* most common measure of variation
* how spread out is the data from the mean?

Population St. Dev. = σ = Greek letter "sigma"

Sample St. Dev. = s_x or just s

How to Calculate on Calculator

Ch 2 Standard Deviation (continued)

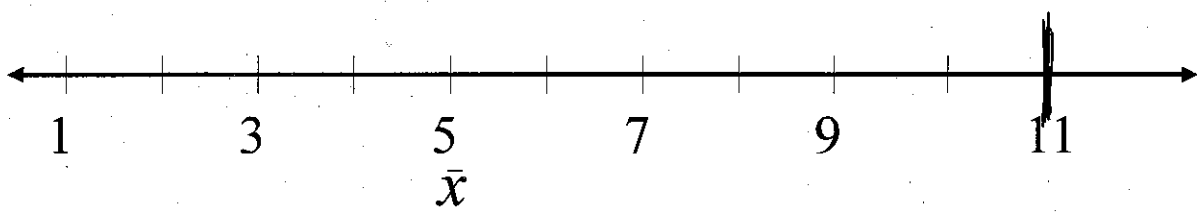
Given a set of data with:

mean $\bar{x} = 5$

st. dev $s = 2$

1. What value is 3 standard deviations above the mean? 11

we also say "to the right of"



2. What value is 2 standard deviations below the mean?

$5 - 2 \cdot 2 = 1$

we also say "to the left of"

3. What value is 1.5 standard deviations above the mean?

$5 + 1.5(2) = 5 + 3 = 8$

Example: Two students, John and Ali, from different high schools, wanted to find out who had the highest G.P.A. when compared to his school. Which student had the highest G.P.A. when compared to his school?

Student	GPA	School Mean GPA	School St. Dev.
John	2.5	2.0	1.0
Ali	77	75	10

*We can also use std. devs to compare data from 2 different data sets

z-score: $z = \frac{x - \bar{x}}{s_x} = \frac{\text{data} - \text{mean}}{\text{std. dev.}}$

ex) John: $\frac{2.5 - 2.0}{1.0} = 0.5$

So John is 0.5 std. devs. above the mean

Ali: $\frac{77 - 75}{10} = 0.2$

So Ali is 0.2 std devs above the mean

So, relatively speaking, John's GPA is better.

2. Variance = (Standard Deviation)^2

The variance is another method of measuring the spread of data. It is widely used in the theory of statistics, but we will use the standard deviation.

How are s_x and σ_x calculated

① For each piece of data x we find

deviation $= x - \bar{x}$

② variance = avg. of the squares of the deviations

$$\frac{\sum_{\text{all data}} (x-\bar{x})^2}{n} = \sigma^2$$

③ Pop. Std. Dev. = square root of variance

$$\sigma_x = \sqrt{\frac{\sum (x-\bar{x})^2}{n}}$$

④ For samples it's better to divide by $n-1$

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

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

To Calculate Standard Deviation by Hand

Data: 1, 3, 5, 7

1. Compute the mean $\bar{x}$: $\frac{1+3+5+7}{4} = 4$

2. Compute the deviations:

Data	Mean $\bar{x}$	Deviations Data – $\bar{x}$	(deviations)^2
1	4	-3	9
3	4	-1	1
5	4	1	1
7	4	3	9
		Sum:	20

sample Variance = $s^2 = \frac{\Sigma(\text{deviations})^2}{n-1} = \frac{20}{3} = 6.6667$

sample Standard Deviation = $s = \sqrt{\text{variance}} = \sqrt{\frac{20}{3}} = 2.5820$