Topic 2: Summary Measure For Populations

▲Recall, parameters are characteristics of data.
▲Recall, parameters summarize data.

We need to be able to compute such measures for individual and "_____" data.

"The most commonly used parameters for interpreting and understanding the meaning of values in populations are measures of central tendency and v_____. They summarize data for logical presentation."

Measures of Location

- There are several measures of "_____ tendency":

(i) The **Mode**: The value that occurs ____ often. Or in a frequency distribution, it is the point or class mark corresponding to the value with the highest frequency.

Example: Age of students in driving school:

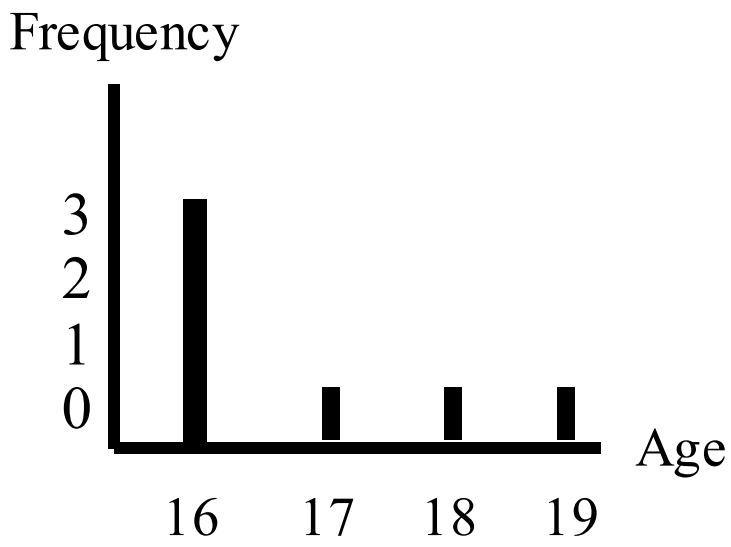
{16, 16, 17, 18, 16, 19.}

Mode=___

<u>Problems with Mode</u> (as a measure of central tendency):

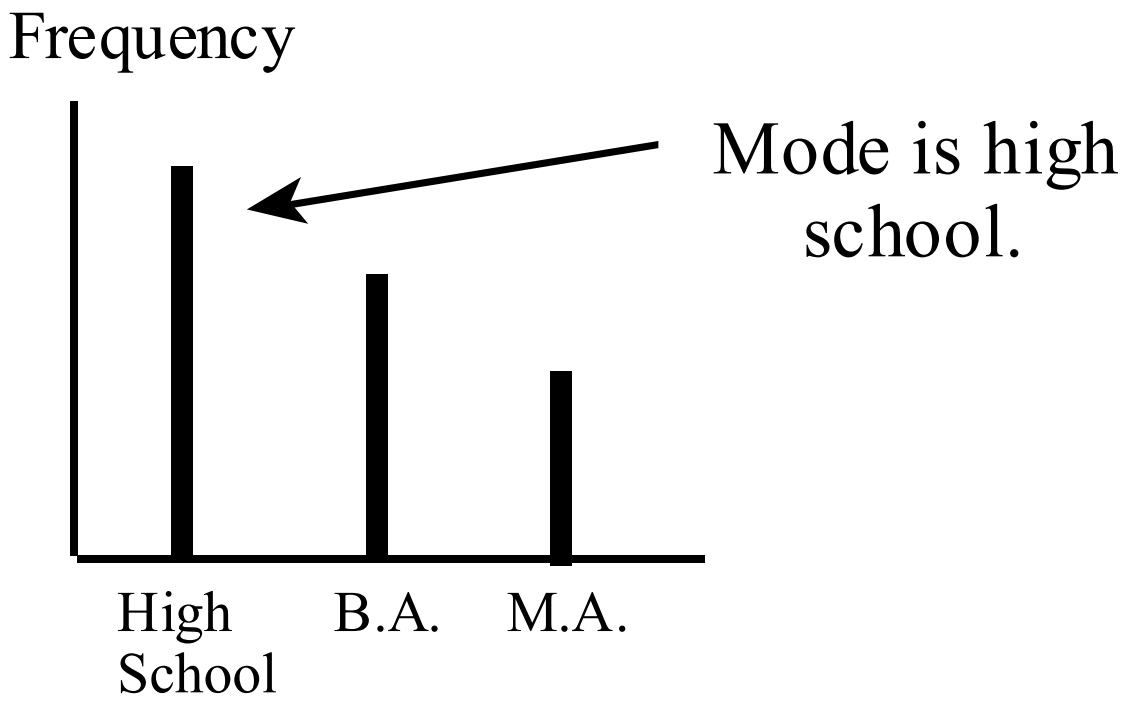
(1) Mode may not be near the_____of the data.

(2) Data may have ____ than one mode.



Mode is not representative of central location of the distribution of ages.

However, for purely **descriptive** purposes, the mode can be useful in representing the ____ frequently occurring value:



Example: Using the mode with a frequency distribution:

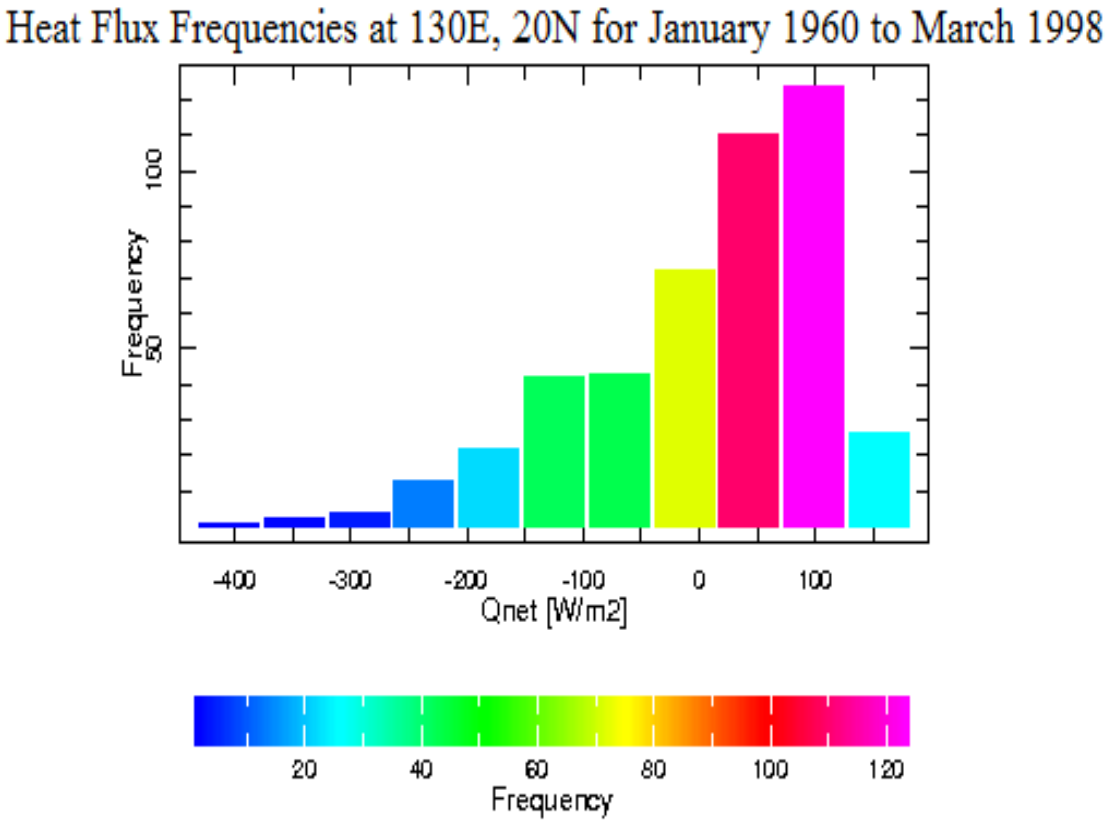
Income	Frequency
80 ≤X < 100	
100 ≤X < 120	
120 ≤X < 140	
140 ≤X < 160	
160 ≤X < 180	
180≤X < 200	
***200*≤X < 220**	
	$\sum f_i = 40$

- Modal Interval is **{180 ≤X < 200}.**
- This is the mode because it is the interval with the most values within it.
- 190 is also considered the "mode" because it is the mid-point of **{180 ≤X < 200}**.

Both are correct!

The Mode:

The mode is defined as the most frequently observed value. For grouped data, the mode is the most commonly observed category, and for ungrouped data, the mode is the value which occurs most frequently.



The histogram is unimodal and is negatively skewed.

"Another measure of central tendency is the Median":

(ii) The **Median**: The _____ value in a set of numbers arranged in order of magnitude.

I.e. the "_____" ranked value in the data.
(Put into ascending order first!)

Example: Sales: N=9 (odd number of values)

{14, 15, 18, 19, **20**, 21, 23, 25, 26}

_____

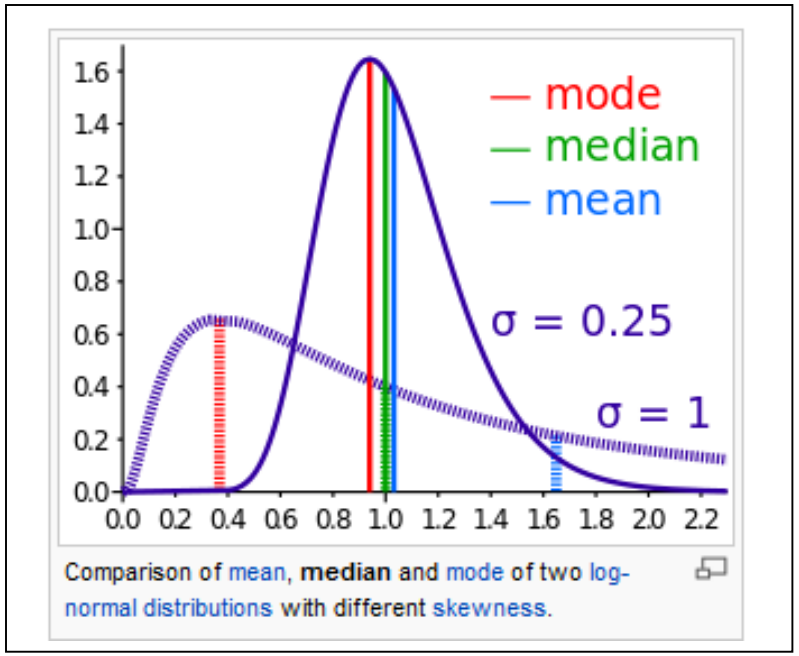
Example: Sales: N=12 (____ number of values)

{16, 16, 17, 18, 19, **20, 21**, 22, 23, 24, 25, 27}

median = 20.5

6th 7th

Note: $\frac{20+21}{2} = 20.5$



Comparison of mean, **median** and mode of two log-normal distributions with different skewness.

Example: Using the median with a frequency distribution:

Income	Frequency	Cumulative Frequency
80 ≤X < 100		**2**
100 ≤X < 120	_	**8**
120 ≤X < 140	_	**16**
140 ≤X < 160	_	**22**
160 ≤X < 180	_	**25**
180 ≤X < 200	**13**	**38**
200 ≤X < 220	**2**	**40=N**
	$\sum f_i = 40$	

N=40
40\2=20 which is an even number.

Find the interval that contains the 20^{th} and 21^{st} ranked observations.

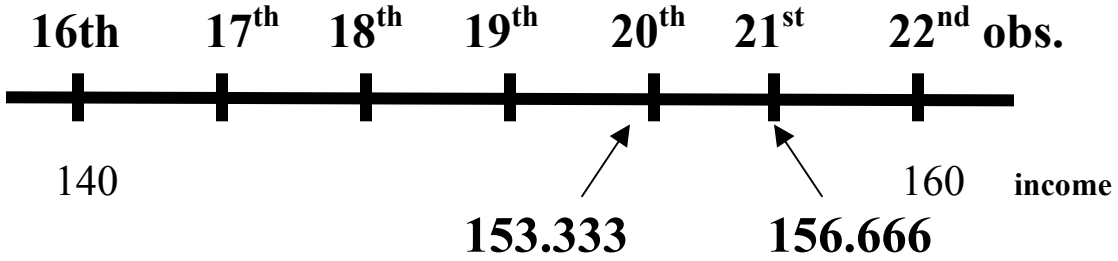
Median is ____-way between the 20th and 21st ranked observations

⇨Median class is (140 to 160).

Note: Could take the middle of the interval as the median: ___.

Or ⇨Could _____ data is spread evenly through the interval:

20/6=3.333

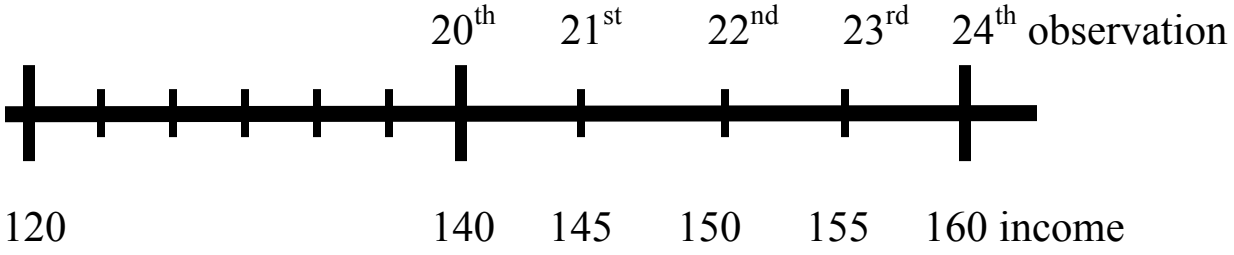


Divide interval into 6 ____ parts. Pick the value between the 20th and 21st observations.

$$\text{Median} = \left(\frac{153.333 + 156.666}{2} \right) = 155$$

Example: 20^{th} observation on the boundary or 20^{th} and 21^{st} observations in different intervals:

Income	Frequency	Cumulative Frequency
$80 \le X < 100$	_	2
$100 \le X < 120$	__	14
$120 \le X < 140$	_	20
$140 \le X < 160$	_	24
$160 \le X < 180$	__	34
$180 \le X < 200$	3	37
$200 \le X < 220$	3	40=N
	$\sum f_i = 40$	



We could _____observation 20 is on the boundary of the interval. We could also assume that the interval {140 to 160} is divided into 4 parts.

The median is calculated as follows:

$$\text{Median} = \left(\frac{140 + 145}{2} \right) = 142.5$$

If we assumed that the 20^{th} and 21^{st} observations are in different _____:

Median 120 to 160 **or** Median =140.

Monthly Sales Summary — July 2011

Tuesday, August 2, 2011

Region / District	Units	Total Volume	Average Price	6 Month Average	Median Price
Residential					
● Single Family					
Greater Victoria					
Victoria	25	$12,991,300	$519,652	$621,921	$480,000
Victoria West	1	$440,000	$440,000	$443,019	$440,000
Oak Bay	19	$16,384,551	$862,345	$878,931	$740,000
Esquimalt	15	$6,595,000	$439,667	$469,087	$415,000
View Royal	4	$2,607,500	$651,875	$554,075	$696,250
Saanich East	55	$33,451,900	$608,216	$654,985	$590,000
Saanich West	20	$10,344,000	$517,200	$567,056	$532,500
Central Saanich	15	$9,281,400	$618,760	$602,821	$534,900
North Saanich	8	$5,489,900	$686,238	$704,703	$677,500
Sidney	11	$5,321,000	$483,727	$488,759	$480,000
Highlands	1	$643,500	$643,500	$573,229	$643,500
Colwood	13	$6,372,999	$490,231	$503,013	$482,000
Langford	30	$14,815,565	$493,852	$508,104	$473,000
Metchosin	5	$3,028,000	$605,600	$651,776	$603,000
Sooke	18	$7,130,588	$396,144	$408,694	$412,500
Waterfront (all districts)	14	$12,706,400	$907,600	$1,071,544	$880,000
Total Greater Victoria	**254**	**$147,603,603**	**$581,117**	**$615,439**	**$535,000**
Other Areas					
Shawnigan Lake / Malahat	1	$338,000	$338,000	$426,518	$338,000
Gulf Islands	11	$5,434,525	$494,048	$466,101	$440,000
Uplsland / Mainland	12	$5,153,755	$429,480	$453,854	$415,500
Waterfront (all districts)	5	$4,115,000	$823,000	$862,788	$640,000
Total Other Areas	**29**	**$15,041,280**	**$518,665**	**$515,932**	**$450,000**
Total Single Family	**283**	**$162,644,883**	**$574,717**	**$606,341**	**$529,900**

The average price for single-family homes sold in Greater Victoria last month was $581,117, down from $629,292 in June. The median price also declined to $535,000 while the six-month average declined to $615,439. There were 13 single family home sales of over $1 million in July including two on the Gulf Islands. The overall average price for condominiums last month was $315,371, down from $320,172 in June. The average for the last six months declined to $327,762. The median price for condominiums in July also declined to $289,000. The average price of all townhomes sold last month declined to $412,178 from $444,768 in June. The median price also declined to $385,000 while the six month average declined to $443,341.

MLS® sales last month included 283 single family homes, 147 condominiums, 47 townhomes and 19 manufactured homes.

(III) **The Mean**: The "balancing" point of data.

Calculate "mean" or average in several ways:

(i) _____ Arithmetic Mean: $\mu = \frac{1}{N}\left[X_1 + X_2 + \cdots + X_N\right] = \frac{1}{N}\sum X_i$

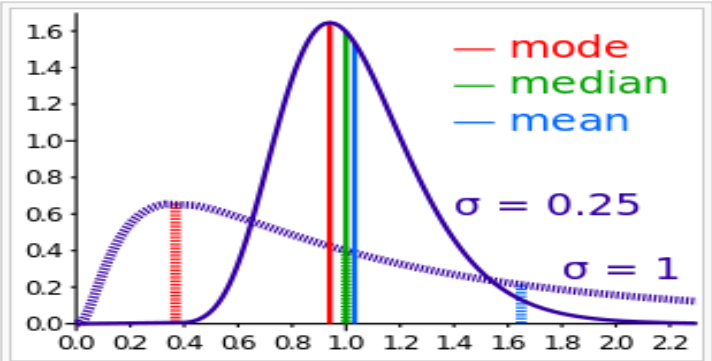
Example: Temperatures: {22, 20, 16, 24, 18, 16, 21, 19, 23} N=9

$\mu = 19.889$

Note for comparison:

Putting data is ascending order: 16, 16, 18, 19, 20, 21, 22, 23, 24

↳ _____=16 (occurs twice)
↳ _____=20 (5th observation)



Comparison of mean, **median** and mode of two log-normal distributions with different skewness.

Property: data values "balance" about __. ***Proof***:

$$\left(X_1 - \mu\right) + \left(X_2 - \mu\right) + \left(X_3 - \mu\right) + \cdots + \left(X_N - \mu\right)$$

$$= \left(X_1 + X_2 + X_3 + \cdots + X_N\right) - \left(\mu + \mu + \mu + \cdots + \mu\right)$$

$$= \sum_{i=1}^{N} X_i - N\mu$$

$$= \left(N\mu - N\mu\right) = 0$$

Since $\dfrac{\sum_{i=1}^{N} X_i}{N} = \mu$, hence: $\sum_{i=1}^{N} X_i = N\mu$.

Problems with the _____:

(A) Affected by _____ in the data.

↳One strange observation misrepresents the data.)

(B) Each data point gets _____ weight when calculating μ–
⇨unrealistic

↳(Does not take into account _____ or importance of certain values.)

Variations:

(ii) Weighted Arithmetic Mean

In situations in which some values are more _____ than others, a weighted average should be used.

$$\mu_W = \frac{\left[W_1 X_1 + W_2 X_2 + \cdots + W_N X_N\right]}{\left[W_1 + W_2 + \cdots + W_N\right]}$$

$$\mu_W = \frac{\left[\sum_{i=1}^{N} W_i X_i\right]}{\sum_{i=1}^{N} W_i}$$

Example: Grouped Data

Weight

Income	Frequency (f_i)	Mid-Point (X_i)
80 ≤X < 100	_	90
100 ≤X < 120	_	110
120 ≤X < 140	_	130
140 ≤X < 160	_	150
160 ≤X < 180	_	170
180 ≤X < 200	__	190
200 ≤X < 220	2	210
	$\sum f_i = 40$	

If there are K intervals (K=7), then the weighted arithmetic mean is:

$$\mu_W = \frac{\left[\sum_{i=1}^{7} f_i X_i\right]}{\left[\sum_{i=1}^{7} f_i\right]}$$

$$\mu_W = \frac{\left[(90 \times 2) + (110 \times 6) + \cdots + (210 \times 2)\right]}{2 + 6 + 8 + 6 + 3 + 13 + 2}$$

$$\mu_W = 154.6$$

Other "means" are used when the data are all in _____ (or rates of change: i.e. price indices; change in yields on stocks and bonds.)

(iii) **G Mean**: GM gives equal weight to changes of equal importance. It cannot be used if any value is '0' in the data.

$$\mu_G = \left(X_1 \times X_2 \times X_3 \times \cdots \times X_N\right)^{1/N} = \left[\prod_{i=1}^{N} X_i\right]^{1/N}$$

where Π is the product operator.

Example: Two interest rates:

$X_1 = 5.4\%$

$X_2 = 6.7\%$

$\mu_G = (5.4 \times 6.7)^{1/2} = 6.015\%$

$\boxed{\mu_G < \mu_A}$

$Note: \quad \mu = \dfrac{5.4 + 6.7}{2} = 6.05\%$

Besides being used by scientists and biologists, geometric means are also used in many other fields, most notably financial reporting. This is because when evaluating investment returns as annual percent change data over several years (or fluctuating interest rates), it is the geometric mean, not the arithmetic mean, that tells you what the average financial rate of return would have had to have been over the entire investment period to achieve the end result. This term is also so called the Compound Annual Growth Rate or CAGR. Population biologists also use the same calculation to determine average growth rates of populations, and this growth rate is referred to as the Intrinsic Rate of Growth when the calculation is applied to estimates of population increases where there are no density-dependent forces regulating the population.

Financial Return Calculation

For financial investment return calculations, the geometric mean is calculated on the decimal multiplier equivalent values, not percent values (i.e., a 6% increase becomes 1.06; a 3% decline is transformed to 0.97. Just follow the steps outlined in the section below titled Calculating Geometric Means with Negative Values).

The equation is also flipped around when calculating the financial rate of return if you know the starting value, end value, and the time period. This equation is used in these cases when the average rate of return is needed (or population growth rate):

$$\text{Return} = \sqrt[\text{Years}]{(\text{Finalvalue}/\text{origvalue})}$$

Note: If you subtract 1 from the equation above, this is your compound interest rate. To use this equation, if years=5, this is the "fifth root", which is the same as raising to the power of 1/5 or 0.2).

Things to Note:

(i) $\mu_G < \mu$ if all $X_i > 0$ and **not** all the same.

(ii) $$\text{Log } \mu_G = Log\left[\prod_{i=1}^{N} X_i\right]^{1/N}$$

$$= \frac{1}{N} Log\left[\prod_{i=1}^{N} X_i\right] = \frac{1}{N} Log(X_1 \times X_2 \cdots \times X_n)$$

$$= \frac{1}{N}(Log\ X_1 + Log\ X_2 + \cdots + Log\ X_N)$$

$$= \frac{1}{N}\sum_{i=1}^{N} Log(X_i)$$

(Arithmetic mean of logs of data.)

In what sense is the geometric mean a _____ point for the data?

Recall: $(X_1 - \mu) + (X_2 - \mu) + (X_3 - \mu) + \cdots + (X_N - \mu) = 0$

In the case of the geometric mean:

$$\left(\frac{X_1}{\mu_G}\right)\left(\frac{X_2}{\mu_G}\right)\left(\frac{X_3}{\mu_G}\right)\cdots\left(\frac{X_N}{\mu_G}\right)$$

$$= \frac{\left(\prod_{i=1}^{N} X_i\right)}{(\mu_G)^N} = \frac{(\mu_G)^N}{(\mu_G)^N} = 1$$

Since $\mu_G = \left[\prod X_i\right]^{1/N}$ *and* $\left[\mu_G = \left[\prod X_i\right]^{1/N}\right]^N \Rightarrow \mu_G{}^N = \left[\prod X_i\right]$

Everything balances multiplicatively about the 'neutral' ratio of _____ – this is why the geometric mean is appropriate with ratio data.

(iv) **H Mean**: it is the reciprocal of mean of reciprocals.

$$\mu_H = \frac{1}{\left[\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{X_i}\right)\right]}$$

Generally: $\mu_H < \mu_G < \mu_A$

Example: {2, 4, 7, 8}

$\mu = 5.25$

$\mu_G = 4.60$

$\mu_H = 3.93$

In finance

The harmonic mean is the preferable method for averaging multiples, such as the price/earning ratio, in which price is in the numerator. If these ratios are averaged using an arithmetic mean (a common error), high data points are given greater weights than low data points. The harmonic mean, on the other hand, gives equal weight to each data point. [3]

Harmonic Mean

Harmonic mean is another measure of central tendency and also based on mathematic footing like arithmetic mean and geometric mean. Like arithmetic mean and geometric mean, harmonic mean is also useful for quantitative data. Harmonic mean is defined in following terms: *Harmonic mean is quotient of "number of the given values" and "sum of the reciprocals of the given values".*

Harmonic Mean

The ***harmonic mean*** is a better "average" when the numbers are defined in relation to some unit. The common example is averaging speed.

For example, suppose that you have four 10 km segments to your automobile trip. You drive your car:

- 100 km/hr for the first 10 km
- 110 km/hr for the second 10 km
- 90 km/hr for the third 10 km
- 120 km/hr for the fourth 10 km.

What is your average speed? Here is a spreadsheet solution:

Distance	Velocity	Time
km	km/hr	hr
10	100	0.100
10	110	0.091
10	90	0.111
10	120	0.083
40		0.385
		Avg
	103.80 V	

The harmonic mean formula is:

$$HM = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} = \frac{4}{\frac{1}{100}+\frac{1}{110}+\frac{1}{90}+\frac{1}{120}} = 103.8$$

Excel calculates this with the formula **=HARMEAN(100,110,90,120)**. Unfortunately, the formula is not generalized to average velocities if across different distances.

*The general **location** of data is a useful measure, but we need **more**. Extending the idea of the median:*

P_____: *'divide data into 100 equal parts'.*

Rank the data into ascending order.

Calculate _____ in the data list below which a fixed _____ of the data lie.

The K^{th} percentile:

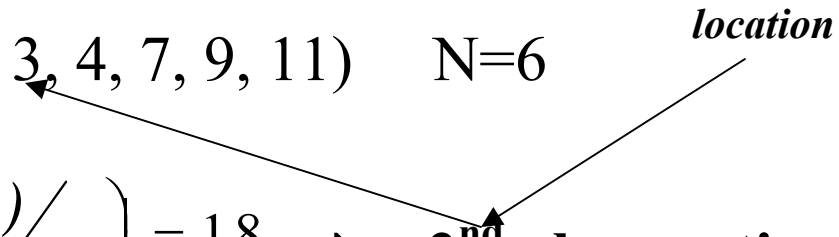
$$P_K = \left((N \times K) / 100 \right)$$

where N = # of observations and K = %.

Then (1) If P_K is integer, add 0.5.
(2) If non-integer, round to next higher integer.

This **locates** the position of the K^{th} percentile.

Example: (1, 3, 4, 7, 9, 11) N=6



$$P_{30} = \left({}^{(6 \times 30)}\!/_{100} \right) = 1.8 \Rightarrow$$ **2^{nd} observation**

The 30^{th} percentile is 3.

(1, 3, 4, 7, 9, 11)

Example: {1, 3, 5, 7, 8, 9, 9, 11 12, 14} N=10

$$P_{17} = \left({(10 \times 17)}/{100} \right) = 1.7$$ ⇨ **2nd observation**

The 17th percentile is '3'.

$$P_{45} = \left({(10 \times 45)}/{100} \right) = 4.5$$ ⇨ 5th **observation**

The 45th percentile is '8'.

❄ ***Kth percentile is the point in the ranked data, below which K% of the data lie.***

Special Cases:

- ♦ 25^{th} Percentile = First Quartile (Q_1)
- ♦ 50^{th} Percentile=Second Quartile (Q_2) = Median
- ♦ 75^{th} Percentile = Third Quartile (Q_3)
- ♦ (Q_3-Q_1) = "Interquartile Range"

Percentiles divide the data into ___ equal parts, each representing one percent of all values.

Eg. The 90^{th} percentile is the value that has 90% of all values below it and 10% above it.

Q_____ divide the data into __equal parts; Only 3 quartile values are necessary to divide the data into 4 parts.

Example : I Range

{3, 5, 7, 2, 1}

Step 1: Order data:

{1, 2, 3, 5, 7} n=5

Step 2: Location

$$Q_3 = P_{75} = \left\lceil (5 \times 75)/100 \right\rceil = 3.75 \quad \Rightarrow 4th$$

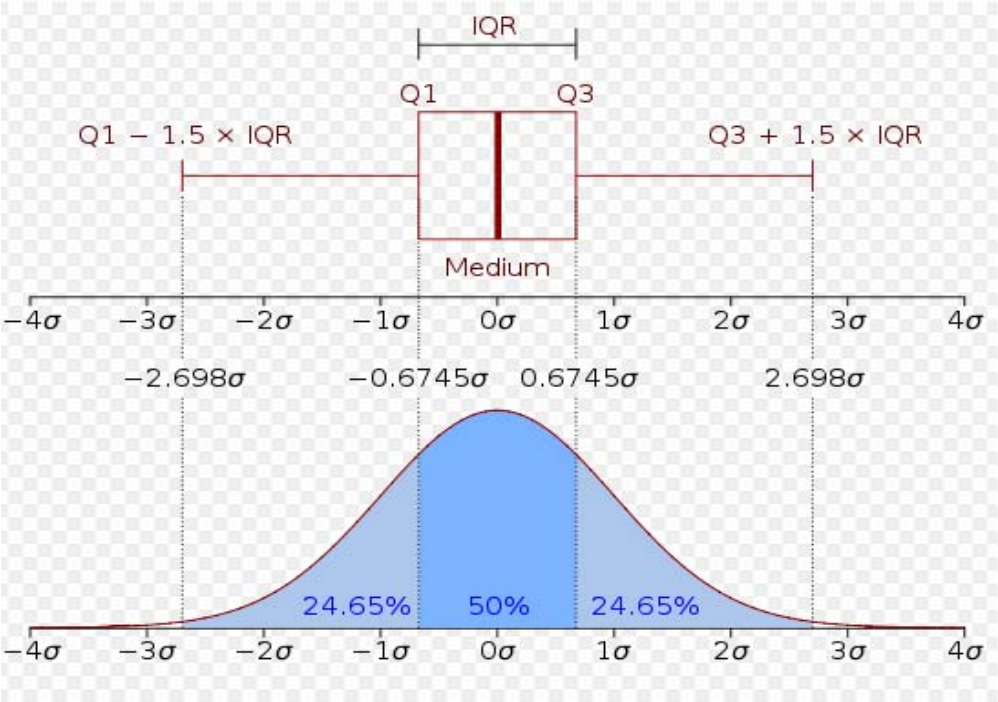
4th observation = 5

$$Q_1 = P_{25} = \left\lceil (5 \times 25)/100 \right\rceil = 1.25 \quad \Rightarrow 2nd$$

2nd observation =2

Step 3: Interquartile calculation:

$$Q_3 - Q_1 = [5 - 2] = 3$$



Unlike (total) range, the interquartile range is a robust statistic, having a breakdown point of 25%, and is thus often preferred to the total range.

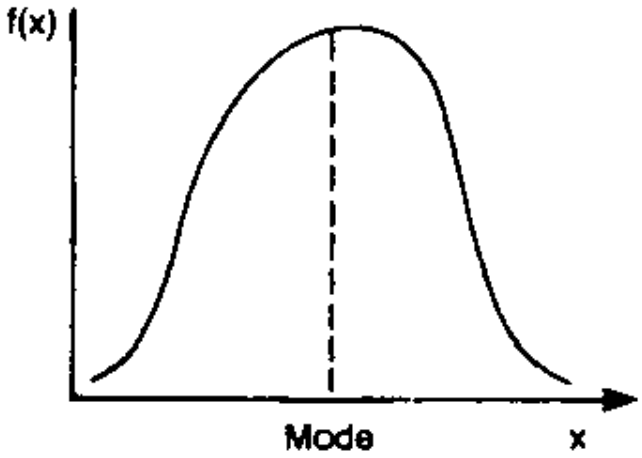
The IQR is used to build box plots, simple graphical representations of a probability distribution.

For a symmetric distribution (so the median equals the midhinge, the average of the first and third quartiles), half the IQR equals the median absolute deviation (MAD).

The median is the corresponding measure of central tendency.

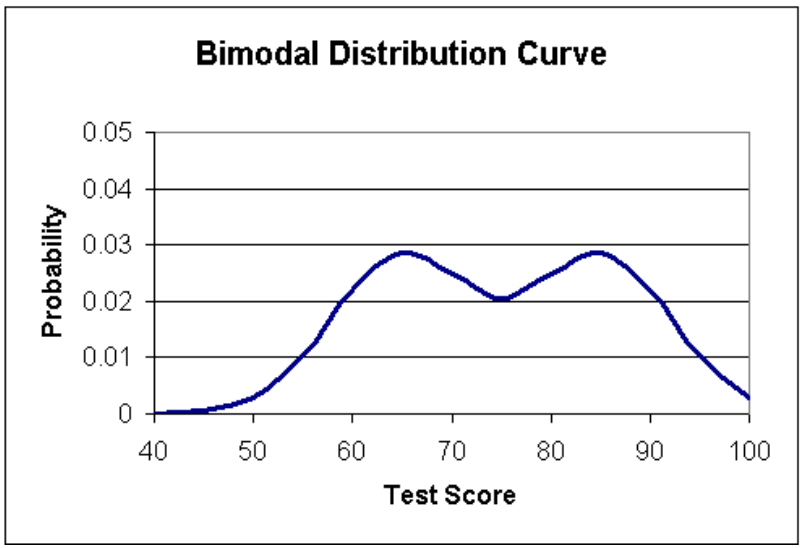
The Shape of A Distribution

● Often having a method for describing the _____ of a frequency distribution is often more helpful than just being able to describe the _____ location of a set of data.



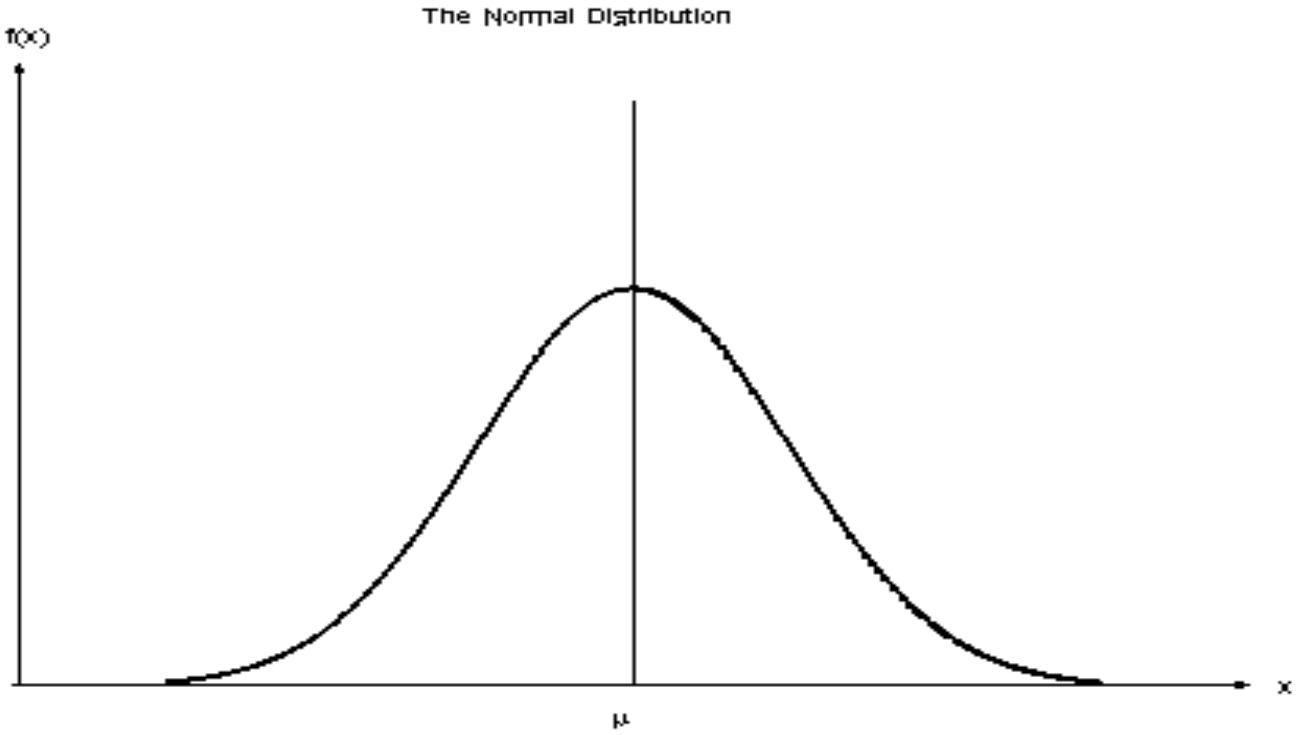
Most "real world" distributions are called _____ distributions, implying one mode or peak.

A distribution with 2 peaks is called a bi distribution.



A frequency distribution is said to be **s**_____ if its shape is the same on both _____ of the _____: Median = Arithmetic Mean.
If a distribution is uni-modal and symmetric then:

Mean=Mode=Median



An asymmetric frequency distribution is said to be _____:

1) to the ____ (down) if: mean < median< mode.
2) to the _____ (up) if: mean >median> mode.

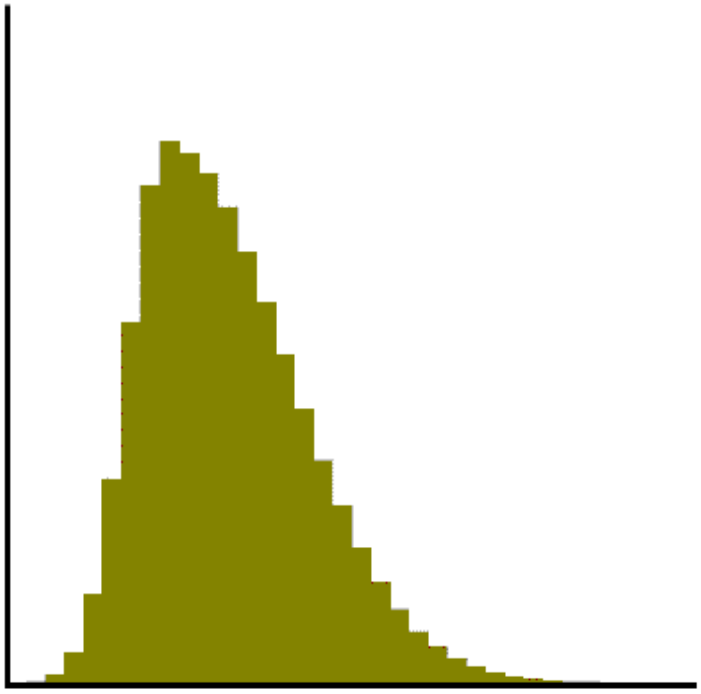


Figure 6.1: Positively Skewed Distribution

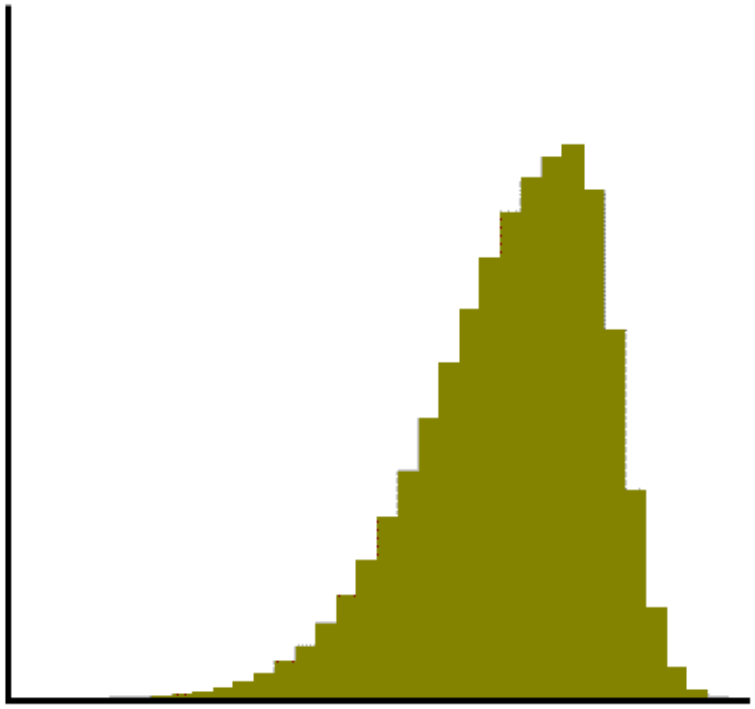


Figure 6.2: Negatively Skewed Distribution

Positively (_____) skewed

Negatively (____) skewed

Measures of Dispersion:

●Measures of central location and general 'shape' are not enough to properly describe a distribution of data because **variability** or **s_____** is ignored.

The Range: the range is the _____ difference between the highest and lowest values in the data set.

$R=(X_{max} - X_{min})$

Advantages:

●Independent of measure of _____ location
●Easy to calculate

Disadvantages:

⬇Ignores all data except 2 items
⬇These values may be "_____." I.e. not representative.

Example: Price of Phones: {10, 37.5, 25, 35, 7, 15, 45, 27}
R=(45-7)=$38

Rank Data: {7 10 15 25 27 35 37.5 45}

(45-7)=__ = RANGE

The Mid-Range(s): Measure of the _____ of the innermost concentrated portion of the data.

- Obtained by _____ a specified proportion of the extreme values at both ____ of the ordered values in the data set.

(Order data and exclude a certain proportion at either end of the data set.)

Use these to avoid _____ problems.

Example: Interest Rates

First, order the data:

{6.3, 6.4, 6.5, 7.0, 7.2, 7.5, 8.5, 9.1, 9.2, 11.4} % N=10

80% Mid-range ⇨ discard top and bottom __% of data (20%)

*Calculate the 10^{th} and 90^{th} p_____.

$P_{90} = (90 \times 10)/100 = 9$ add $0.5 \Rightarrow$ 9.5 position:

$(9.2 + 11.4)/2 = 10.3$

$P_{10} = (10 \times 10)/100 = 1$ add $0.5 \Rightarrow 1.5^{th}$ position:

$(6.3 + 6.4)/2 = 6.35$

R_{80}=(10.3-6.35) =3.95%.

Note: 50% ___-_____= Inter-quartile Range.

A selection of mid-ranges provides a "picture" of the dispersion in data. But, a single summary measure would be more helpful:

The Variance: A single summary measure of data d_____.

- Takes account of ___ N data values
- Adjusts the data to take account of ***t***_____ level. I.e. use the mean, μ, in calculations.

Construction: Calculate deviation from the mean for each observation:

$$\left(X_1 - \mu\right), \left(X_2 - \mu\right), \left(X_3 - \mu\right), \cdots, \left(X_N - \mu\right)$$

We could add these "dispersions" together:

$$\left(X_1 - \mu\right) + \left(X_2 - \mu\right) + \left(X_3 - \mu\right) + \cdots + \left(X_N - \mu\right) = 0$$

But this sum is always ____, and some of the deviations will be positive and some negative:

For example: {1, 2, 3}
↳mean=2

$$(1-2) + (2-2) + (3-2) = (-1) + (0) + (1) = 0$$

So, we need to deal with the positive and negative signs:

Two Options:

Take the "***s***_____" or take "***a***_____" values of the individual deviations.

In each case, we need to scale the sum by dividing by N to get the correct order.

This leads to: **Two Measures of the Variance**:

(A) **Mean _____ Deviation**: square each deviation and divide by N:

$$MSD = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - \mu\right)^2$$

In the case of a population, we also call this the Population Variance: (**σ^2 ---** sigma squared)

*Note the _____: if data are in $, MSD and the variance are in $\$^2$.

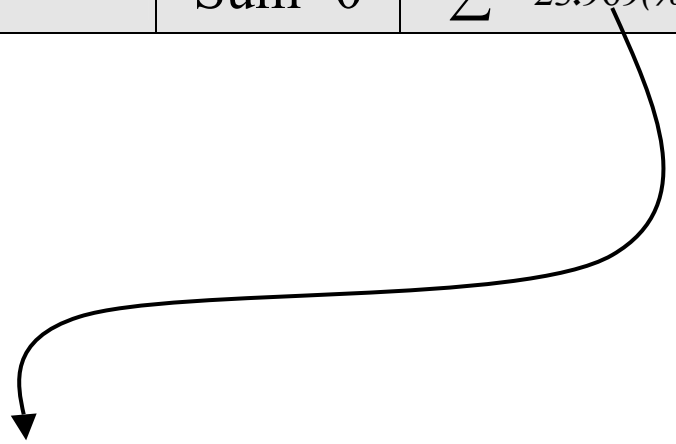
So, we often report the standard deviation, which is the _____ root of the variance: Standard deviation = $\sqrt{\text{Variance}} = \sigma$.

(B) **Mean Absolute Deviation**: Sum the _____ deviation and divide by N:

$$MAD = \frac{1}{N}\sum_{i=1}^{N}\left|X_i - \mu\right|$$

Example: N=10; Interest Rates; μ=7.91%

i	X_i	$(X_i - \mu)$	$(X_i - \mu)^2$	$\lvert(Xi-\mu)\rvert$
1		-1.61	2.5921	1.61
2		-1.51	2.2801	1.51
3		-1.41	1.9881	1.41
4		-0.91	0.8281	0.91
5		-0.71	0.5041	0.71
6		-0.41	0.1681	0.41
7		0.59	0.3481	0.59
8		1.19	1.4161	1.19
9		1.29	1.6641	1.29
10		3.49	12.1801	3.49
		Sum=0	$\sum$ = 23.969$(\%)^2$	$\sum$ = 13.12$(\%)$



Mean square deviation

$$\text{MSD} = \sigma^2 = \frac{23.969}{10} = 2.3969(\%)^2$$

$$\sigma = 1.5548\%$$

$$\text{MAD} = \frac{13.12}{10} = 1.312\%$$

same units

Mean absolute deviation

Warning: Rounding error can be an issue.

Alternative Formula for the Variance (MSD):

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - \mu\right)^2 \Leftarrow \textit{Expand}$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\left(X_i^2 + \mu^2 - 2\mu X_i\right) \Leftarrow \text{Take summation operator through}$$

$$\sigma^2 = \frac{1}{N}\left[\sum_{i=1}^{N} X_i^2 + \sum_{i=1}^{N}\mu^2 - 2\mu\sum_{i=1}^{N} X_i\right] \Leftarrow \text{Since } \frac{\sum \mathrm{X_i}}{\mathrm{N}} = \mu$$

$$\text{then: } \mathrm{N}\mu = \sum \mathrm{X_i}$$

$$\sigma^2 = \frac{1}{N}\left[\sum_{i=1}^{N} X_i^2 + N\mu^2 - 2\mu(N\mu)\right]$$

$$\sigma^2 = \frac{1}{N}\left[\sum_{i=1}^{N} X_i^2 + N\mu^2 - 2N\mu^2\right]$$

$$\sigma^2 = \frac{1}{N}\left[\sum_{i=1}^{N} X_i^2 - N\mu^2\right]$$

$$\sigma^2 = \frac{1}{N}\left[\sum_{i=1}^{N} X_i^2\right] - \mu^2$$

Note: $\sum_{i=1}^{N} \mu^2 = \mu^2 + \mu^2 + \mu^2 + \cdots + \mu^2 = N\mu^2$

f a random variable X has the expected value (mean) $\mu = \mathrm{E}[X]$, then the variance of X is given by:

$$\operatorname{Var}(X) = \mathrm{E}\left[(X - \mu)^2\right] .$$

That is, the variance is the expected value of the squared difference between the variable's realization and the variable's mean. This definition encompasses random variables that are discrete, continuous, or neither (or mixed). It can be expanded as follows:

$$\begin{aligned}
\operatorname{Var}(X) &= \mathrm{E}\left[(X - \mu)^2\right] \\
&= \mathrm{E}\left[X^2 - 2\mu X + \mu^2\right] \\
&= \mathrm{E}\left[X^2\right] - 2\mu\,\mathrm{E}[X] + \mu^2 \\
&= \mathrm{E}\left[X^2\right] - 2\mu^2 + \mu^2 \\
&= \mathrm{E}\left[X^2\right] - \mu^2 \\
&= \mathrm{E}\left[X^2\right] - (\mathrm{E}[X])^2.
\end{aligned}$$

What About Grouped Data?

For the Variance:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{K}\left[\left(X_i - \mu\right)^2 f_i\right]$$

where:

K =# of groups

f_i = frequency $\Rightarrow \sum f_i = N$

$$MAD = \frac{1}{N}\sum_{i=1}^{K}\left[\left|X_i - \mu\right| f_i\right]$$

NOTE: X_i = midpoint

Example: House Prices ($'000): calculate the variance and SD.

Class (i)	Mid-Point (X_i)	Frequency (f_i)	$(Xi-\mu)^2$	$(Xi-\mu)^2 f_i$
100<X≤200	150		20,736	62,208
200<X≤300	250		1,936	52,272
300<X≤400	350		3,136	47,040
400<X≤500	450		24,336	121,680
		N=50		283,200

$$\mu = \frac{1}{N}\sum X_i f_i$$

$$\mu = \frac{1}{50}\left[(150 \times 3) + (250 \times 27) + (350 \times 15) + (450 \times 5)\right]$$

$$\mu = 294$$

$$\sigma^2 = \frac{1}{50}(283{,}200) = 5664\ (\$'000)^2$$

$$\sigma = 75.26 \quad (\$'000)$$

Issue: Accuracy With Grouped Data

- We used **mid** _____ of groups as representative values for the calculations of mean and variance.
  - The effect on the mean is _____.
- The effect on the _____ is **not**!

Must use **Sheppard's Correction** (for the _____):

$$\sigma_C^2 = \sigma^2 - \left(h^2/12\right)$$

where h = class width

$$\sigma^2 = 5664 \quad (\$'000)^2$$

Example: $\sigma_C^2 = 4830.7 \quad (\$'000)^2$

- Using intervals instead of actual data values, the variance is _____.

- As the interval width gets smaller and smaller, eventually use individual values.

Comparing Dispersions of Two Populations

How do you compare standard deviations if:
(i) S____(average value) differs and/or
(ii) U____ of measurement differ across populations? I.e. £,¢,$.

- **Must construct a u_____ measure which is scale free:**

Coefficient of Variation (CV)

$$C.V. = \left[\frac{\sigma}{\mu} * 100 \right] \%$$

Example:

Population 1 ($)	Population 2 (Kg.)
	104
	107
	102

$\mu = 2.5$

$\sigma = 1.118(\$)$

$C.V. = 44.72\%$

$\mu = 104.33$

$\sigma = 2.055(kg.)$

$C.V. = 1.97\%$

Measuring Skewness

Recall:

(i) Skewed n_____ if mean < median.
(ii) Skewed p_____ if mean > median.

Both measures have the same units.

Pearson's Skewness Measure:

$$\text{Skew} = \left[\frac{\text{mean} - \text{median}}{\sigma} \right] \Rightarrow \text{"+"} \quad \text{"0"} \quad \text{"–"}$$

Unitless measure for comparison.

Example: {25, 22, 31, 35, 30, 27, 28, 45, 50, 100}

$$\mu = 39.3$$

$$median = 30.5$$

$$\sigma = 21.882$$

$$skew = \left[\frac{39.3 - 30.5}{21.882}\right] = 0.402$$

Unitless –O.K. for comparisons.

Pearson's skewness coefficients

Karl Pearson suggested simpler calculations as a measure of skewness:[3] the Pearson mode or first skewness coefficient,[4] defined by

- (mean − mode) / standard deviation,

as well as Pearson's median or second skewness coefficient,[5] defined by

- 3 (mean − median) / standard deviation.