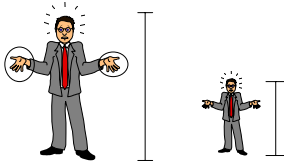
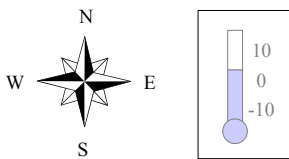


There are several TYPES of variables that reflect characteristics of the data

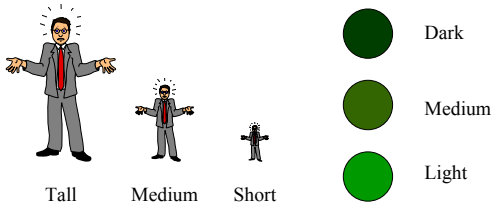
Ratio scale



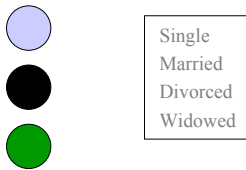
Interval scale



Ordinal scale



Nominal scale



Another useful classification

Continuous

Eg height 150 to 210cm range
Bill - 174.25 cm



Discrete

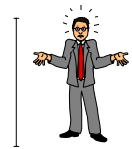
Eg # of hands 0 to 3 range
Bill - 2 hands

2 more important issues with data

Accuracy →

Precision →

Let's say Bill's real height is 174.25 cm.



Accurate Precise	Not Accurate Not Precise	Not Accurate Precise
174.25	172	170.25
174.25	178	170.25
174.25	171	170.25
174.25	174	170.25
174.25	182	170.25
174.25	168	170.25

Frequency Distribution

→ raw frequency
→

→ relative frequency
→

Name	Height (cm)	Hair Colour
Anne	168	Brown
Rishi	178	Black
Bill	183	Brown
Cristin	172	Brown
Rich	175	Black

Variable: Hair Colour

Sample size = 5

Frequency of Black Hair = 2 }
 Frequency of Brown Hair = 3 }

Relative Frequency of Black Hair = $2/5 = 0.4$ }
 Relative Frequency of Brown Hair = $3/5 = 0.6$ }

Variable: Height

Sample size = 5

Frequency of 168 cm = 1
 Frequency of 172 cm = 1
 Frequency of 175 cm = 1
 Frequency of 178 cm = 1
 Frequency of 183 cm = 1

Relative Frequency of 168 cm = $1/5 = 0.2$
 Relative Frequency of 172 cm = $1/5 = 0.2$
 Relative Frequency of 175 cm = $1/5 = 0.2$
 Relative Frequency of 178 cm = $1/5 = 0.2$
 Relative Frequency of 183 cm = $1/5 = 0.2$

Make categories

Eg. Number above and number below **mid-point of range**

Range: Maximum - Minimum

$$183 \text{ cm} - 168 \text{ cm} = 15 \text{ cm}$$

Mid-point: half way between Min and Max

$$\begin{aligned} &= \text{Min} + (\text{Range} / 2) \\ &= 168 \text{ cm} + 7.5 \text{ cm} \\ &= 175.5 \text{ cm} \end{aligned}$$

Frequency of Heights Below 175.5 cm = 3
Frequency of Heights Above 175.5 cm = 2

Relative Frequency of Heights Below 175.5 cm = $3/5 = 0.6$
Relative Frequency of Heights Above 175.5 cm = $2/5 = 0.4$

Could make THREE categories

Divide range by 3: $15 \text{ cm} / 3 = 5 \text{ cm}$

Category 1: 168 cm to 168 cm + 5 cm
168 cm to 173 cm

Category 2: 174 cm to 174 cm + 5 cm
174 cm to 179 cm

Category 3: 180 cm to 180 cm + 5 cm
180 cm to 185 cm

Frequency of Heights in 168 cm to 172 cm = 2
 Frequency of Heights in 173 cm to 178 cm = 2
 Frequency of Heights in 179 cm to 184 cm = 1

 Relative Frequency of Heights in 168 cm to 172 cm = $2/5 = 0.4$
 Relative Frequency of Heights in 173 cm to 178 cm = $2/5 = 0.4$
 Relative Frequency of Heights in 179 cm to 184 cm = $1/5 = 0.2$

Mother's age and babies birth weight data from Massachussetts

19 2523	25 2977	24 3274	30 3699	34 1135	20 2296
33 2551	25 2977	28 3303	24 3728	25 1330	21 2301
20 2557	29 2977	20 3317	19 3756	25 1474	26 2325
21 2594	19 2977	22 3317	24 3770	27 1588	31 2353
18 2600	27 2992	22 3317	23 3770	23 1588	15 2353
21 2622	31 3005	31 3321	20 3770	24 1701	23 2367
22 2637	33 3033	23 3331	25 3790	24 1729	20 2381
17 2637	21 3042	16 3374	30 3799	21 1790	24 2381
29 2663	19 3062	16 3374	22 3827	32 1818	15 2381
26 2665	23 3062	18 3402	18 3856	19 1885	23 2395
19 2722	21 3062	25 3416	16 3860	25 1893	30 2410
19 2733	18 3076	32 3430	32 3860	16 1899	22 2410
22 2750	18 3076	20 3444	18 3884	25 1928	17 2414
30 2750	32 3080	23 3459	29 3884	20 1928	23 2424
18 2769	19 3090	22 3460	33 3912	21 1928	17 2438
18 2769	24 3090	32 3473	20 3940	24 1936	26 2442
15 2778	22 3090	30 3475	28 3941	21 1970	20 2450
25 2782	22 3100	20 3487	14 3941	20 2055	26 2466
20 2807	23 3104	23 3544	28 3969	25 2055	14 2466
28 2821	22 3132	17 3572	25 3983	19 2082	28 2466
32 2835	30 3147	19 3572	16 3997	19 2084	14 2495
31 2835	19 3175	23 3586	20 3997	26 2084	23 2495
36 2836	19 3175	36 3600	26 4054	24 2100	17 2495
28 2863	21 3203	22 3614	21 4054	17 2125	21 2495
25 2877	30 3203	24 3614	22 4111	20 2126	
28 2877	20 3203	21 3629	25 4153	22 2187	
17 2906	17 3225	19 3629	31 4167	27 2187	
29 2920	17 3225	25 3637	35 4174	20 2211	
26 2920	13 3232	16 3643	19 4238	17 2225	
17 2920	24 3232	29 3651	24 4593	25 2240	
17 2920	28 3234	29 3651	45 4990	20 2240	
24 2948	26 3260	19 3651	28 709	18 2282	
35 2948	20 3274	19 3651	29 1021	18 2296	

Range of the Birth Weight data:
 Minimum: 709 g
 Maximum: 4990 g
 Difference: 4281 g

 Let's say we want to look at the distribution of data
 across 10 categories.

 Each category would span 428.1 g, but for convenience
 we'll round to 430 g.

 Also, instead of starting our first category at 709 g we'll
 use 700g

<u>Category</u>	<u>Range</u>	<u>Freq.</u>	<u>Rel. Freq.</u>
1	700-1130	3	0.015873016
2	1131-1560	3	0.015873016
3	1561-1990	14	0.074074074
4	1991-2420	29	0.153439153
5	2421-2850	34	0.17989418
6	2851-3280	44	0.232804233
7	3281-3710	33	0.174603175
8	3711-4140	23	0.121693122
9	4141-4750	4	0.021164021
10	4751-5000	2	0.010582011

Previous breakdown ok as long as I have measured weight to the nearest gram,

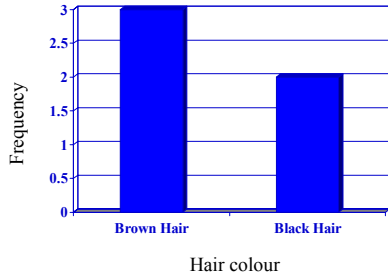
BUT, if I've measure to the nearest 0.1 gram

--> my categories may miss some observations

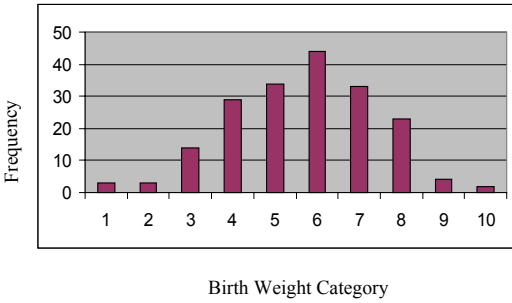
So need to adjust...

<u>Category</u>	<u>Range</u>	<u>Range</u>
1	700-1130	700-1130.9
2	1131-1560	1131-1560.9
3	1561-1990	1561-1990.9
4	1991-2420	1991-2420.9
5	2421-2850	2421-2850.9
6	2851-3280	2851-3280.9
7	3281-3710	3281-3710.9
8	3711-4140	3711-4140.9
9	4141-4750	4141-4750.9
10	4751-5000	4751-5000.9

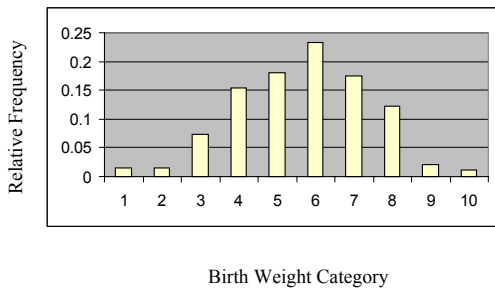
Histogram -



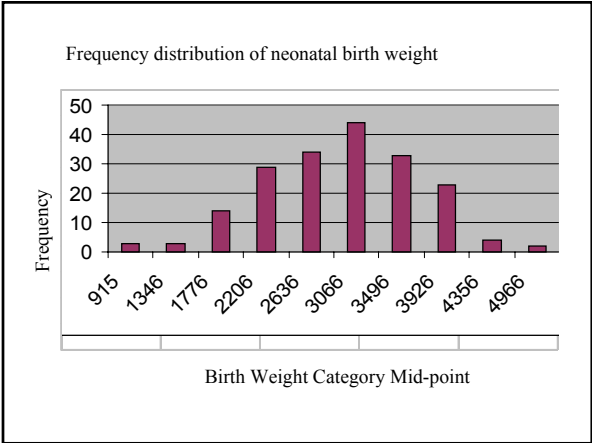
Frequency distribution of neonatal birth weight



Frequency distribution of neonatal birth weight



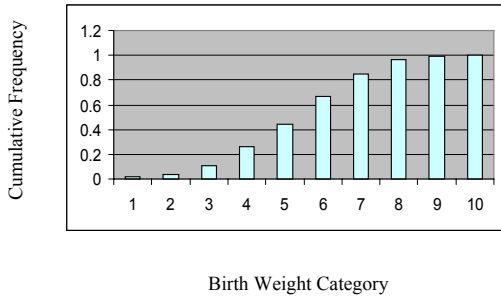
Category	Range	Mid-point
1	700-1130	915
2	1131-1560	1346
3	1561-1990	1776
4	1991-2420	2206
5	2421-2850	2636
6	2851-3280	3066
7	3281-3710	3496
8	3711-4140	3926
9	4141-4750	4356
10	4751-5000	4966



Cumulative Frequency -

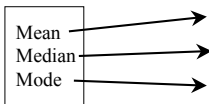
Category	Range	Freq.	Rel. Freq.	Cum. Freq.
1	700-1130	3	0.0158	0.0158
2	1131-1560	3	0.0158	0.0317
3	1561-1990	14	0.07407	0.1058
4	1991-2420	29	0.15343	0.2592
5	2421-2850	34	0.17989	0.4391
6	2851-3280	44	0.23280	0.6719
7	3281-3710	33	0.17460	0.8465
8	3711-4140	23	0.12169	0.9682
9	4141-4750	4	0.02116	0.9894
10	4751-5000	2	0.01058	1.0

Frequency distribution of neonatal birth weight



Measures of Central Tendency

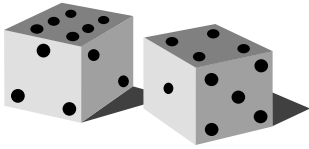
-
-



The Mean:

The mean is calculated by summing the observed values and dividing the sum by the total number of observations.





A die has 6 sides, 1 dot, 2, 3, 4, 5, and 6

$$\mu =$$

$$\bar{X} =$$

$$\mu =$$

$$\bar{X} =$$

	Observation i	Height X_i
Rishi	1	172
Anne	2	185
Bill	3	132
Cristin	4	191
Rich	5	205

19 2523	25 2977	24 3274	30 3699	34 1135	20 2296
33 2551	25 2977	28 3303	24 3728	25 1330	21 2301
20 2557	29 2977	20 3317	19 3756	25 1474	26 2325
21 2594	19 2977	22 3317	24 3770	27 1588	31 2353
18 2600	27 2992	22 3317	23 3770	23 1598	15 2353
21 2622	31 3005	31 3321	20 3770	24 1701	23 2367
22 2637	33 3033	23 3331	25 3790	24 1729	20 2381
17 2637	21 3042	16 3374	30 3799	21 1790	24 2381
29 2663	19 3062	16 3374	22 3827	32 1818	15 2381
26 2665	23 3062	18 3402	18 3856	19 1885	23 2395
19 2722	21 3062	25 3416	16 3860	25 1893	30 2410
19 2733	18 3076	32 3430	32 3860	16 1899	22 2410
22 2750	18 3076	20 3444	18 3884	25 1928	17 2414
30 2750	32 3080	23 3459	29 3884	20 1928	23 2424
18 2769	19 3090	22 3460	33 3912	21 1928	17 2438
18 2769	24 3090	32 3473	20 3940	24 1936	26 2442
15 2778	22 3090	30 3475	28 3941	21 1970	20 2450
25 2782	22 3100	20 3487	14 3941	20 2055	26 2466
20 2807	23 3104	23 3544	28 3969	25 2055	14 2466
28 2821	22 3132	17 3572	25 3983	19 2082	28 2466
32 2836	30 3147	19 3572	16 3997	19 2084	14 2495
31 2836	19 3175	23 3586	20 3997	26 2084	23 2495
36 2836	16 3175	36 3600	26 4054	24 2100	17 2495
28 2863	21 3203	22 3614	21 4054	17 2125	21 2495
25 2877	30 3203	24 3614	22 4111	20 2126	
28 2877	20 3203	21 3629	25 4153	22 2187	
17 2906	17 3225	19 3629	31 4167	27 2187	
29 2920	17 3225	25 3637	35 4174	20 2211	
26 2920	23 3232	16 3643	19 4238	17 2225	
17 2920	24 3232	29 3651	24 4593	25 2240	
17 2920	28 3234	29 3651	45 4990	20 2240	
24 2948	26 3260	19 3651	28 709	18 2262	
35 2948	20 3274	19 3651	29 1021	18 2296	

n = 189

$\sum_{i=1}^{189} X_i = 556540$

$n = 189$

$$\sum_{i=1}^{189} X_i = 556540$$

$$\bar{X} = \frac{\sum X_i}{n} = \frac{556540}{189} = 2944.656$$

Another way to calculate the mean

Suppose you had a frequency distribution for the number of cancerous moles on people who regularly visit Club Med

# cancerous moles (X)	Frequency (f)
0	8
1	4
2	8
3	10
4	2
5	1

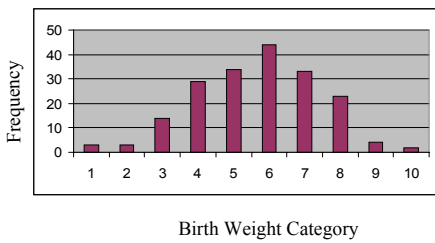
# cancerous moles (x)	Frequency (f)	f * x
0	8	0
1	4	4
2	8	16
3	10	30
4	2	8
5	1	5
n = 33		$\Sigma f * x = 63$

$n = \Sigma f's$
 $\Sigma X's = \Sigma f*x$

$\bar{X} =$

The Mode: the most frequently occurring value in a set of

Frequency distribution of neonatal birth weight



Category	Range	Freq.	Rel. Freq.
1	700-1130	3	0.015873016
2	1131-1560	3	0.015873016
3	1561-1990	14	0.074074074
4	1991-2420	29	0.153439153
5	2421-2850	34	0.17989418
6	2851-3280	44	0.232804233
7	3281-3710	33	0.174603175
8	3711-4140	23	0.121693122
9	4141-4750	4	0.021164021
10	4751-5000	2	0.010582011

Mid-point is 3065.5 --> report the MODE as 3065.5

The Median: the middle measurement of a set of data
--> data must be ordered

Observation (X)	Heights (cm)	Ordered Heights (cm)
1	178	123
2	143	143
3	123	168
4	189	173
5	187	178
6	205	187
7	168	189
8	173	198
9	198	205

Median is 178 cm

Observation (X)	Heights (cm)	Ordered Heights (cm)
1	178	123
2	143	143
3	123	162
4	189	168
5	187	173
6	205	178
7	168	187
8	173	189
9	198	198
10	162	205

Middle observation is 5.5 --> median is midway between observation 5 and observation 6

Median is $(173+178)/2 = 175.5$

General formula for Median:
If n is an **odd** number:

General formula for Median:
If n is an **even** number:

# cancerous moles (X)	Frequency (f)	Cumulative Frequency
0	8	8
1	4	12
2	8	20
3	10	30
4	2	32
5	1	33

$M = X_{(\frac{n+1}{2})} = X_{17} = 2$

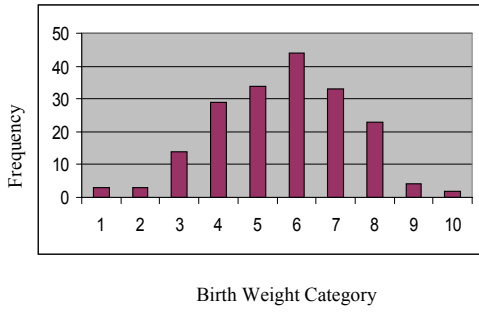
0	2	3
0	2	3
0	2	3
0	2	3
0	2	3
0	2	3
0	2	4
0	2	4
1	3	5
1	3	
1	3	
1	3	

Category	Range	Freq.	Cum. Freq.
1	700-1130	3	3
2	1131-1560	3	6
3	1561-1990	14	20
4	1991-2420	29	49
5	2421-2850	34	83
6	2851-3280	44	127
7	3281-3710	33	160
8	3711-4140	23	183
9	4141-4750	4	187
10	4751-5000	2	189

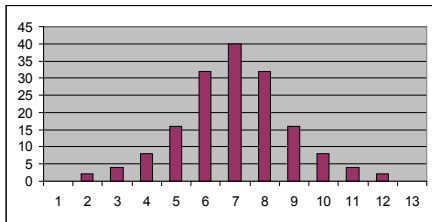
$M = X_{(n+1)/2} = X_{190/2} = X_{95}$

Median =
(lower limit of class) + ((0.5*n - cum.freq.) / #obs in interval)(interval size)
= 2851 + ((0.5*189- 83)/44) * (430)
= 2851 + (94.5-83)/44 * 430
= 2963.4

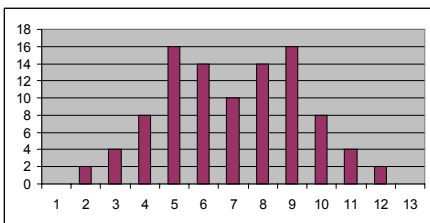
Frequency distribution of neonatal birth weight



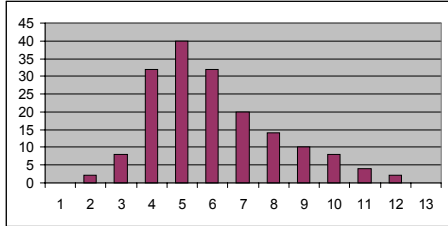
Symmetrical, unimodal distribution



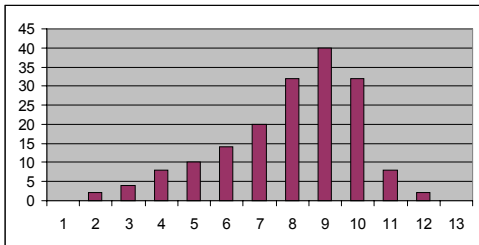
Symmetrical, bimodal distribution



Asymmetric distribution

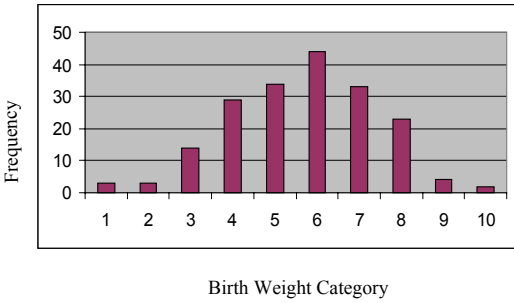


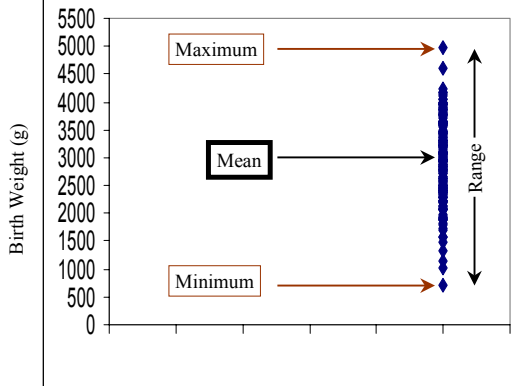
Asymmetric distribution

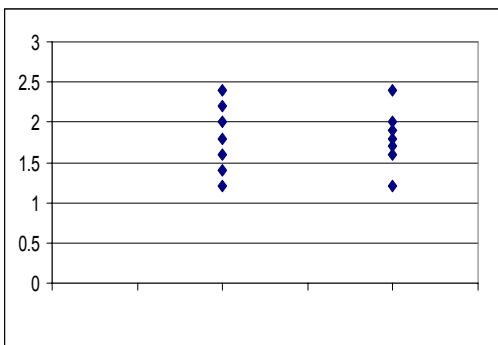


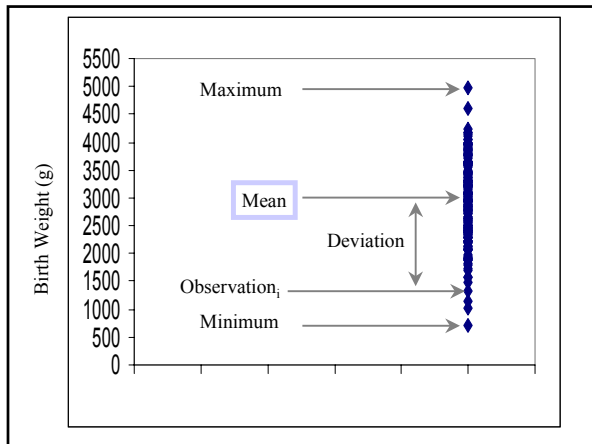
Measures of Dispersion and Variability

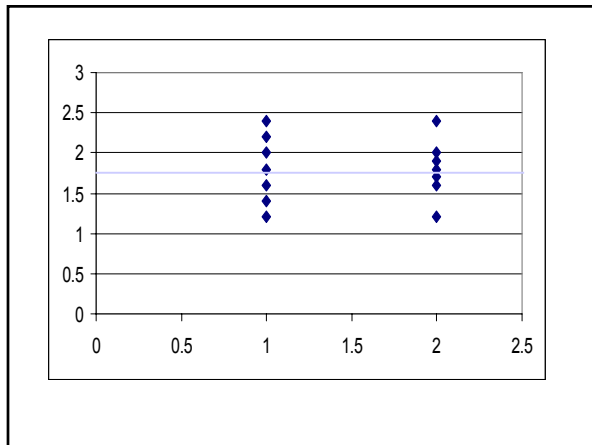
Frequency distribution of neonatal birth weight











Average Deviation from the Mean

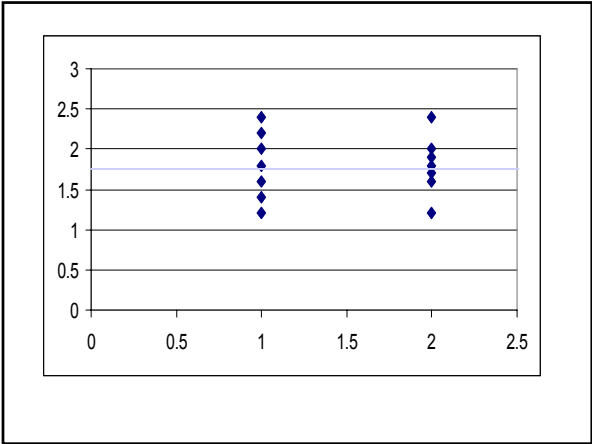
--> on average, how much do the individual observations differ from the mean?

$$\frac{\sum_{i=1}^n (X_i - \bar{X})}{n}$$

i	X_i	$X_i - \bar{X}$	$ X_i - \bar{X} $	$(X_i - \bar{X})^2$
1	1.2			
2	1.4			
3	1.6			
4	1.8			
5	2.0			
6	2.2			
7	2.4			
$\Sigma X = 12.6$				
$n = 7$				
$\bar{X} = \frac{12.6}{7} = 1.8$				

i	X_i	$X_i - \bar{X}$	$ X_i - \bar{X} $	$(X_i - \bar{X})^2$
1	1.2	$1.2 - 1.8 = -0.6$		
2	1.4	-0.4		
3	1.6	-0.2		
4	1.8	0.0		
5	2.0	0.2		
6	2.2	0.4		
7	2.4	0.6		
$\Sigma X = 12.6$				
$n = 7$				
$\bar{X} = \frac{12.6}{7} = 1.8$				

i	X_i	$X_i - \bar{X}$	$ X_i - \bar{X} $	$(X_i - \bar{X})^2$
1	1.2	$1.2 - 1.8 = -0.6$		
2	1.4	-0.4		
3	1.6	-0.2		
4	1.8	0.0		
5	2.0	0.2		
6	2.2	0.4		
7	2.4	0.6		
$\Sigma X = 12.6$				
$n = 7$				
$\bar{X} = \frac{12.6}{7} = 1.8$				



Average Absolute Deviation from the Mean

--> on average, how much do the individual observations differ from the mean?

$$\frac{\sum_{i=1}^n |X_i - \bar{X}|}{n}$$

i	X_i	$X_i - \bar{X}$	$ X_i - \bar{X} $	$(X_i - \bar{X})^2$
1	1.2	1.2-1.8 = -0.6	1.2-1.8 = 0.6	
2	1.4	-0.4	0.4	
3	1.6	-0.2	0.2	
4	1.8	0.0	0.0	
5	2.0	0.2	0.2	
6	2.2	0.4	0.4	
7	2.4	0.6	0.6	
$\Sigma X = 12.6$		0.0	$\sum_{i=1}^7 X_i - \bar{X} = 2.4$	
$n = 7$			$= 2.4 / 7 = 0.34$	
$\bar{X} = \frac{12.6}{7} = 1.8$				

Sum of Squared Deviations

$$SS = \sum_{i=1}^n (X_i - \bar{X})^2$$

“Sum of Squares”

i	X_i	$X_i - \bar{X}$	$ X_i - \bar{X} $	$(X_i - \bar{X})^2$
1	1.2	-0.6	0.6	$(-0.6)^2 = 0.36$
2	1.4	-0.4	0.4	0.16
3	1.6	-0.2	0.2	0.04
4	1.8	0.0	0.0	0
5	2.0	0.2	0.2	0.04
6	2.2	0.4	0.4	0.16
7	2.4	0.6	0.6	0.36
$\Sigma X = 12.6$		0.0	0.34	$\sum_{i=1}^n (X_i - \bar{X})^2 = 1.12$
$n = 7$				
$\bar{X} = \frac{12.6}{7} = 1.8$				

Variance

--> mean sum of squares

$$\sigma^2 = \frac{\sum_{i=1}^n (X_i - \mu)^2}{N} \quad \text{Population}$$

$$s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} \quad \text{Sample}$$

i	X_i	$X_i - \bar{X}$	$ X_i - \bar{X} $	$(X_i - \bar{X})^2$
1	1.2	-0.6	0.6	$(-0.6)^2 = 0.36$
2	1.4	-0.4	0.4	0.16
3	1.6	-0.2	0.2	0.04
4	1.8	0.0	0.0	0
5	2.0	0.2	0.2	0.04
6	2.2	0.4	0.4	0.16
7	2.4	0.6	0.6	0.36
$\Sigma X = 12.6$		0.0	0.34	1.12
$n = 7$				
$\bar{X} = \frac{12.6}{7} = 1.8$				
		$s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} = \frac{1.12}{6} = 0.1867$		

Another way to calculate variance

$$s^2 = \frac{\sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n}}{n-1}$$

i	X_i
1	1.2
2	1.4
3	1.6
4	1.8
5	2.0
6	2.2
7	2.4

$$s^2 = \frac{\sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n}}{n-1}$$

i	X_i	X_i^2
1	1.2	1.44
2	1.4	1.96
3	1.6	2.56
4	1.8	3.24
5	2.0	4.00
6	2.2	4.84
7	2.4	5.76
Σ	12.6	23.80

$n=7$

$$s^2 = \frac{\sum_{i=1}^n X_i^2 - \frac{(\sum_{i=1}^n X_i)^2}{n}}{n-1}$$

i	X_i	X_i^2
1	1.2	1.44
2	1.4	1.96
3	1.6	2.56
4	1.8	3.24
5	2.0	4.00
6	2.2	4.84
7	2.4	5.76
Σ	12.6	23.80

$n=7$

$$s^2 = \frac{23.80 - \frac{12.6^2}{7}}{6} = \frac{23.80 - 22.68}{6} = \frac{1.12}{6} = 0.1867$$

Standard Deviation

$$\sigma = \sqrt{\sigma^2}$$

Population

$$s = \sqrt{s^2}$$

Sample

Coefficient of Variation

$$V = \frac{s}{\bar{X}}$$

--> allows comparison of variability among samples measured in different units or scales.

