Measures of Dispersion

- *Defenition*
- *Range*
- *Interquartile Range*
- *Variance and Standard Deviation*

1

Definition

- *Measures of dispersion* are descriptive statistics that describe how similar a set of scores are to each other
  - The more similar the scores are to each other, the lower the measure of dispersion will be
  - The less similar the scores are to each other, the higher the measure of dispersion will be
  - In general, the more spread out a distribution is, the larger the measure of dispersion will be

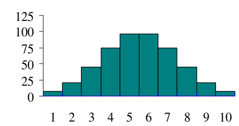
2

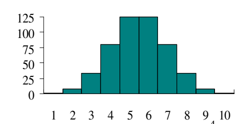
Definition

- Measure of dispersion estimate the spread or variability of a distribution around the centre
- **Dispersion** is a key concept in statistical thinking.
- The basic question being asked is how much do the scores deviate around the Mean?
- The more "bunched up" around the mean the better your ability to make accurate predictions.

Measures of Dispersion

- Which of the distributions of scores has the larger dispersion?





- The upper distribution has more dispersion because the scores are more spread out
  - That is, they are less similar to each other

4

Measures of Dispersion

- If all values are the same, then they all equal the mean. There is no variability.
- Variability exists when some values are different from (above or below) the mean.
- We will discuss the following measures of spread: range, quartiles, variance, and standard deviation

The Range

- The *range* is defined as the difference between the largest score in the set of data and the smallest score in the set of data, $X_L - X_S$
- What is the range of the following data: 4 8 1 6 6 2 9 3 6 9
- The largest score (X_L) is 9; the smallest score (X_S) is 1; the range is $X_L - X_S = 9 - 1 = 8$

6

When To Use the Range

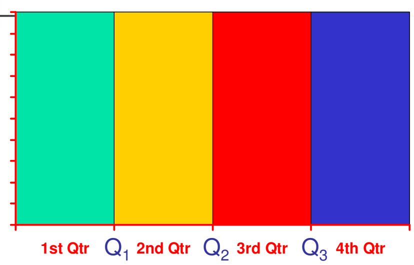
- The range is used when
  - you have ordinal data or
  - you are presenting your results to people with little or no knowledge of statistics
- The range is rarely used in scientific work as it is fairly insensitive
  - It depends on only two scores in the set of data, X_L and X_S
  - Two very different sets of data can have the same range:
    1 1 1 1 9 vs 1 3 5 7 9

7

Quartiles

- Three numbers which divide the ordered data into four equal sized groups.
- Q_1 has 25% of the data below it.
- Q_2 has 50% of the data below it. (Median)
- Q_3 has 75% of the data below it.

Quartiles
Uniform Distribution



9

Obtaining the Quartiles

- Order the data.
- For $\mathbf{Q_2}$, just find the median.
- For $\mathbf{Q_1}$, look at the lower half of the data values, those to the left of the median location; find the *median* of this lower half.

10

Obtaining the Quartiles

- For $\mathbf{Q_3}$, look at the upper half of the data values, those to the right of the median location; find the *median* of this upper half.

Obtaining the Quartiles

- Position of i-th Quartile: position of point
- Qi =(i(n+1))/4
- Eg. Data in Ordered Array:
- 11 12 13 16 16 17 18 21 22
- Position of Q1 **= (1(9+1))/4**
- **=10/4=2.5**
- Q1=12.5

- Position of Q2 **= (2(9+1))/4**
- **=20/4=5**
- Q2=16
- Position of Q3 **= (3(9+1))/4**
- **=30/4=7.5**
- Q3=19.5

The interquartile range

is the distance between the lower quartile and the upper quartile.

2 3 | 5 7 10 11 | 15 16

The interquartile range

- gives the spread of the middle 50% of the data – often the bit you are most interested in

2 3 | 5 7 10 11 | 15 16

- It is not affected by any extreme values.

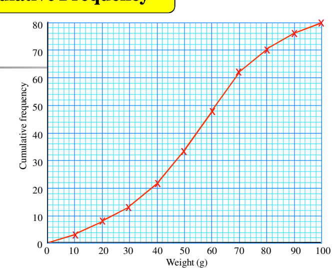
- Difference Between Third & First Quartiles: Interquartile Range =Q3-Q1
- =19.5-12.5=7
- Not Affected by Extreme Values

A pet shop owner weighs his mice every week to check their health.
The weights of the 80 mice are shown below:

weight (g)	Frequency (f)	Cumulative Frequency
$0 < w \le 10$	3	3
$10 < w \le 20$	5	8
$20 < w \le 30$	5	13
$30 < w \le 40$	9	22
$40 < w \le 50$	11	33
$50 < w \le 60$	15	48
$60 < w \le 70$	14	62
$70 < w \le 80$	8	70
$80 < w \le 90$	6	76
$90 < w \le 100$	4	80

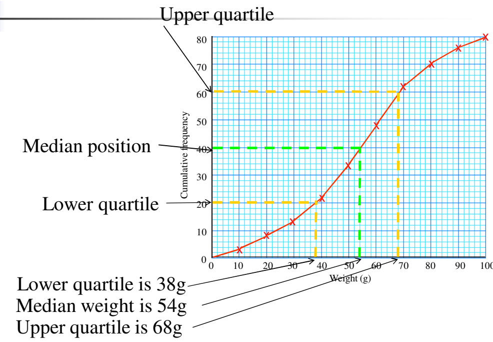
Cumulative Frequency

Weight (g)	Frequency (f)	Cumulative Frequency
$0 < w \le 10$	3	3
$10 < w \le 20$	5	8
$20 < w \le 30$	5	13
$30 < w \le 40$	9	22
$40 < w \le 50$	11	33
$50 < w \le 60$	15	48
$60 < w \le 70$	14	62
$70 < w \le 80$	8	70
$80 < w \le 90$	6	76
$90 < w \le 100$	4	80



- From this graph we can now find estimates of the median, and upper and lower quartiles
- The lower quartile is the 20th piece of data ¼ of the total pieces of data
- The upper quartile is the 60th piece of data ¾ of the total pieces of data

Cumulative Frequency



Variance and Standard Deviation

- variability exists when some values are different from (above or below) the mean.
- Each data value has an associated *deviation from the mean*:

$$x_i - \overline{x}$$

21

Deviations

- what is a *typical* deviation from the mean? *(standard deviation)*
- small values of this typical deviation indicate small variability in the data
- large values of this typical deviation indicate large variability in the data

22

Variance

- Find the mean
- Find the deviation of each value from the mean
- Square the deviations
- Sum the squared deviations
- Divide the sum by *n-1*

(gives typical squared deviation from mean)

23

Variance Formula of Sample

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

24

Standard Deviation

- Standard deviation = √variance
- Variance = standard deviation^2

Standard Deviation Formula
typical deviation from the mean

$$s = \sqrt{\frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

[standard deviation = square root of the variance]

Variance and Standard Deviation
Example from Text

Metabolic rates of 7 men (cal./24hr.) :

1792 1666 1362 1614 1460 1867 1439

$$\bar{x} = \frac{1792+1666+1362+1614+1460+1867+1439}{7}$$

$$= \frac{11{,}200}{7}$$

$$= 1600$$

Variance and Standard Deviation
Example from Text

Observations x_i	Deviations $x_i - \bar{x}$	Squared deviations $(x_i - \bar{x})^2$
1792	1792–1600 = 192	$(192)^2$ = 36,864
1666	1666 –1600 = 66	$(66)^2$ = 4,356
1362	1362 –1600 = -238	$(-238)^2$ = 56,644
1614	1614 –1600 = 14	$(14)^2$ = 196
1460	1460 –1600 = -140	$(-140)^2$ = 19,600
1867	1867 –1600 = 267	$(267)^2$ = 71,289
1439	1439 –1600 = -161	$(-161)^2$ = 25,921
	sum = 0	sum = 214,870

Variance and Standard Deviation
Example from Text

$$s^2 = \frac{214{,}870}{7-1} = 35{,}811.67$$

$$s = \sqrt{35{,}811.67} = 189.24 \text{ calories}$$

Variance Formula of a Population

- When calculating variance, it is often easier to use a computational formula which is algebraically equivalent to the definitional formula:

$$\sigma^2 = \frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N} = \frac{\sum (X-\mu)^2}{N}$$

- σ^2 is the population variance, X is a score, μ is the population mean, and N is the number of scores

Variance Formula of a Population

X	X^2	X-μ	$(X-\mu)^2$
9	81	2	4
8	64	1	1
6	36	-1	1
5	25	-2	4
8	64	1	1
6	36	-1	1
Σ = 42	Σ = 306	Σ = 0	Σ = 12

31

Computational Formula Example

$$\sigma^2 = \frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N} = \frac{306 - \frac{42^2}{6}}{6} = \frac{306 - 294}{6} = \frac{12}{6} = 2$$

$$\sigma^2 = \frac{\sum (X-\mu)^2}{N} = \frac{12}{6} = 2$$

32

Measuring Dispersion of Grouped Data

Daily Renta

Stated Class Limits	Frequency (f)
50 - 59	2.00
60 - 69	3.00
70 - 79	5.00
80 - 89	3.00
90 - 99	2.00
Totals	n = 15.00

Range for grouped data

Range

A. Range = H - L

B. Use the real class limits for H and L

$H - L = 99.5 - 49.5 = 50$

Interquartile Range– Grouped Data

$$Median = L_m + \left(\frac{\frac{n}{2} - F}{f_m} \right) i$$

Where:

n = the **total frequency**

F = the **cumulative frequency** ***before*** class median

Fm= the **frequency** of the class median

f_m

i = the class width

L_m = the **lower boundary** of the class median

Quartiles

Using the same method of calculation as in the Median, we can get Q_1 and Q_3 equation as follows:

$$Q_1 = L_{Q_1} + \left(\frac{\frac{n}{4} - F}{f_{Q_1}} \right) i \qquad Q_3 = L_{Q_3} + \left(\frac{\frac{3n}{4} - F}{f_{Q_3}} \right) i$$

Example: Based on the grouped data below, find the Interquartile Range

Time to travel to work	Frequency
1 – 10	8
11 – 20	14
21 – 30	12
31 – 40	9
41 – 50	7

Solution:
1[st] Step: *Construct the cumulative frequency distribution*

Time to travel to work	*Frequency*	*Cumulative Frequency*
1 – 10	*8*	*8*
11 – 20	*14*	*22*
21 – 30	*12*	*34*
31 – 40	*9*	*43*
41 – 50	*7*	*50*

$$\text{Class } Q_1 = \frac{n}{4} = \frac{50}{4} = 12.5$$

Class Q_1 is the 2[nd] class Therefore,

$$Q_1 = L_{Q_1} + \left(\frac{\frac{n}{4} - F}{f_{Q_1}} \right) i$$

$$= 10.5 + \left(\frac{12.5 - 8}{14} \right) 10$$

$$= 13.7143$$

Qi =(i(n+1))/4

$$\text{Class } Q_3 = \frac{3n}{4} = \frac{3(50)}{4} = 37.5$$

$$Q_3 = L_{Q_3} + \left(\frac{\frac{n}{4} - F}{f_{Q_3}} \right) i$$

$$= 30.5 + \left(\frac{37.5 - 34}{9} \right) 10$$

$$= 34.3889$$

Interquartile Range

$IQR = Q_3 - Q_1$

$IQR = Q_3 - Q_1$

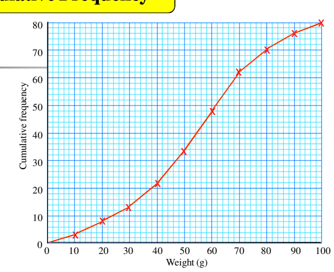
calculate the IQ
$IQR = Q_3 - Q_1 = 34.3889 - 13.7143 = 20.6746$

A pet shop owner weighs his mice every week to check their health.
The weights of the 80 mice are shown below:

weight (g)	Frequency (f)	Cumulative Frequency
$0 < w \le 10$	3	3
$10 < w \le 20$	5	8
$20 < w \le 30$	5	13
$30 < w \le 40$	9	22
$40 < w \le 50$	11	33
$50 < w \le 60$	15	48
$60 < w \le 70$	14	62
$70 < w \le 80$	8	70
$80 < w \le 90$	6	76
$90 < w \le 100$	4	80

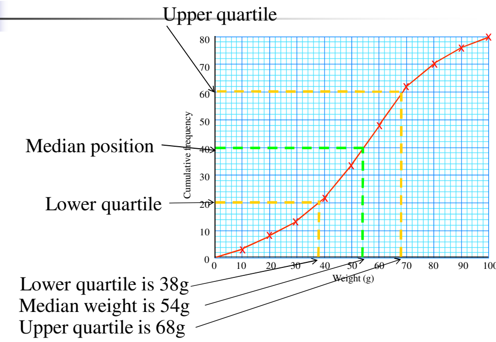
Cumulative Frequency

Weight (g)	Frequency (f)	Cumulative Frequency
$0 < w \le 10$	3	3
$10 < w \le 20$	5	8 13
$20 < w \le 30$	5	22
$30 < w \le 40$	9	33
$40 < w \le 50$	11	48
$50 < w \le 60$	15	62
$60 < w \le 70$	14	70
$70 < w \le 80$	8	76
$80 < w \le 90$	6	80
$90 < w \le 100$	4	



- From this graph we can now find estimates of the median, and upper and lower quartiles
- The lower quartile is the 20th piece of data ¼ of the total pieces of data
- The upper quartile is the 60th piece of data ¾ of the total pieces of data

Cumulative Frequency



Standard deviation for Grouped Data

$$S = \sqrt{\frac{\sum fx^2 - \frac{\left(\sum fx\right)^2}{n}}{n-1}}$$

Variance and SD

Stated Class Limits	Frequency (f)	x
50 - 59	2.00	54.50
60 - 69	3.00	64.50
70 - 79	5.00	74.50
80 - 89	3.00	84.50
90 - 99	2.00	94.50
Totals	n = 15.00	

Standard deviation for Grouped Data

Stated Class Limits	Frequency (f)	x	fx
50 - 59	2.00	54.50	109.00
60 - 69	3.00	64.50	193.50
70 - 79	5.00	74.50	372.50
80 - 89	3.00	84.50	253.50
90 - 99	2.00	94.50	189.00
Totals	n = 15.00		1,117.50

Standard deviation for Grouped Data

Stated Class Limits	Frequency (f)	x	fx	x^2
50 - 59	2.00	54.50	109.00	2,970.25
60 - 69	3.00	64.50	193.50	4,160.25
70 - 79	5.00	74.50	372.50	5,550.25
80 - 89	3.00	84.50	253.50	7,140.25
90 - 99	2.00	94.50	189.00	8,930.25
Totals	n = 15.00		1,117.50	

Standard deviation for Grouped Data

Stated Class Limits	Frequency (f)	x	fx	x^2	fx^2
50 - 59	2.00	54.50	109.00	2,970.25	5,940.50
60 - 69	3.00	64.50	193.50	4,160.25	12,480.75
70 - 79	5.00	74.50	372.50	5,550.25	27,751.25
80 - 89	3.00	84.50	253.50	7,140.25	21,420.75
90 - 99	2.00	94.50	189.00	8,930.25	17,860.50
Totals	n = 15.00		1,117.50		85,453.75

Standard deviation for Grouped Data

$$S = \sqrt{\frac{\sum fx^2 - \frac{\left(\sum fx\right)^2}{n}}{n-1}}$$

Standard deviation for Grouped Data

$$= \sqrt{\frac{85{,}453.75 - \frac{(1{,}117.5)^2}{15}}{15-1}} = \sqrt{\frac{85{,}453.75 - 83{,}253.75}{14}}$$

$$= \sqrt{\frac{2{,}200}{14}} = \sqrt{157.143} = 12.5$$

Alternative formula

$$\sigma_g^2 = \frac{\sum_{i=1}^{K}\left[f_i\left(x_i - \bar{x}\right)^2\right]}{N},$$

Yield per Hectare (in quintals)	Number of Fields
31 – 35	2
36 – 40	3
41 – 45	8
46 – 50	12
51 – 55	16
56 – 60	5
61 – 65	2
66 – 70	2

Yield per Hectare (in quintal)	No. of Fields	Class Marks
31 – 35	2	33
36 – 40	3	38
41 – 45	8	43
46 – 50	12	48
51 – 55	16	53
56 – 60	5	58
61 – 65	2	63
66 – 70	2	68
Total	50	

Yield per Hectare (in quintal)	No. of Fields	Class Marks	$(x_i - \bar{x})$
31 – 35	2	33	– 17
36 – 40	3	38	– 12
41 – 45	8	43	– 7
46 – 50	12	48	– 2
51 – 55	16	53	+3
56 – 60	5	58	+8
61 – 65	2	63	+13
66 – 70	2	68	+18
Total	50		

Yield per Hectare (in quintal)	No. of Fields	Class Marks	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
31 – 35	2	33	– 17	289
36 – 40	3	38	– 12	144
41 – 45	8	43	– 7	49
46 – 50	12	48	– 2	4
51 – 55	16	53	+3	9
56 – 60	5	58	+8	64
61 – 65	2	63	+13	169
66 – 70	2	68	+18	324
Total	50			

Yield per Hectare (in quintal)	No. of Fields	Class Marks	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$f_i(x_i - \bar{x})^2$
31 – 35	2	33	– 17	289	578
36 – 40	3	38	– 12	144	432
41 – 45	8	43	– 7	49	392
46 – 50	12	48	– 2	4	48
51 – 55	16	53	+3	9	144
56 – 60	5	58	+8	64	320
61 – 65	2	63	+13	169	338
66 – 70	2	68	+18	324	648
Total	50				2900

$$\sigma_g^2 = \frac{\sum_{i=1}^{n}\left[f_i(x_i - \bar{x})^2\right]}{N} = \frac{2900}{50} = 58 \text{ and } \sigma_g = +\sqrt{58} = 7.61 \text{ (approx)}$$

Standard deviation for Grouped Data

Variance

$$S^2 = (S)^2 = (12.5)^2 = 156.25 \approx 156.3$$

exercise

In a study on effectiveness of a medicine over a group of patients, the following results were obtained :

Percentage of relief	0 – 20	20 – 40	40 – 60	60 – 80	80 – 100
No. of patients	10	10	25	15	40

Find the variance and standard deviation.

In a study on ages of mothers at the first child birth in a village, the following data were available :

Age (in years) at first child birth	18 – 20	20 – 22	22 – 24	24 – 26	26 – 28	28 – 30	30 – 32
No. of mothers	130	110	80	74	50	40	16

Find the variance and the standard deviation.