WESTMINSTER

INTERNATIONAL UNIVERSITY IN TASHKENT

An Accredited Institution of the University of Westminster (UK)

LECTURE 3
MEASURES OF DISPERSION

Saidgozi Saydumarov
Sherzodbek Safarov

Room: ATB 308 QM Module Leaders
Office Hours: ssaydumarov@wiut.uz
by appointment s.safarov@wiut.uz

Lecture outline:

- Range
- Interquartile range
- Variance
- Standard Deviation

Measures of dispersion

- Dispersion measures how "spread out" the data is
- Shows how reliable our conclusions from the measures of location are
- The lower the dispersion the closer the data is bunched around the measure of location
- Measures of dispersion are used by
  - Economists to measure income inequality
  - Quality control engineers to specify tolerances
  - Investors to study price bubbles
  - Gamblers to predict how much they might win or lose
  - Pollsters to estimate margins of error

Untabulated data

Untabulated data – range

Range

A student can take 1 of 2 routes to get to the university

Route A	Route B
15	20
17	15
14	13
16	10
13	17

Both routes have a mean and median time of 15 minutes

Which one would you prefer?

Untabulated data – range

Let’s calculate the range

Range = Maximum – Minimum

Range of Route A = 17 – 13 = 4

Range of Route B = 20 – 10 = 10

	Route A	Route B
Min	13	10
Max	17	20
Range	4	10

Route A has less dispersed or less “spread out” travel time. Route A is preferred over Route B even though they have the same mean and median.

Untabulated data – interquartile range

Interquartile range

Sometimes, the outer values are extreme. In that case, the range between the lower quartile and upper quartile (the interquartile range) is more appropriate than the range between the minimum and maximum values.

Consider **Example 2** from last week's lecture:

The range of the typical route is: 43 – 9 = 34

The range of the alternative route is: 29 – 11 = 18

However, if we exclude the top outlier from both routes, the typical route seems less spread out.

Typical route	Alternative route
9	15
12	13
10	11
11	17
43	29

Untabulated data – interquartile range

Let’s calculate the interquartile range:

Interquartile range: Upper quartile – lower quartile

Typical route: 12 – 10 = 2

Alternative route: 17 – 13 = 4

Using interquartile range, the typical route is less spread out.

Parameter	Typical route	Alternative route
Lower quartile	10	13
Upper quartile	12	17
Interquartile range	2	4

Untabulated data – variance

The range only considers the outer values

The interquartile range discards the outliers but only considers quartile values

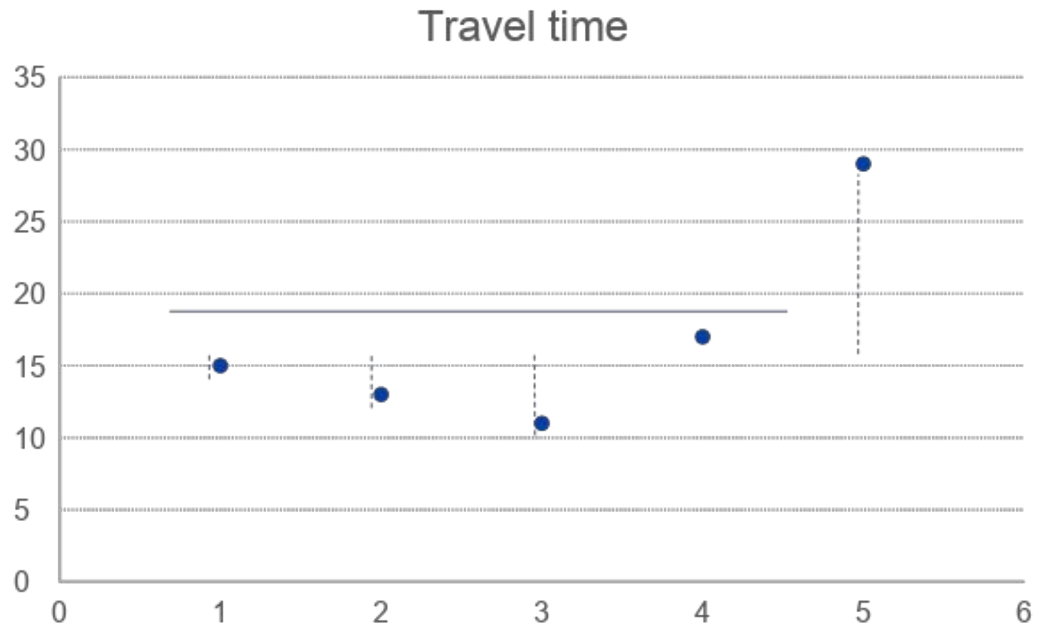
What if we wanted to consider every point when measuring dispersion?

Enter – **Variance**

Variance is the average squared deviations from the mean

Let's plot the travel times of the alternative route on a graph

- The mean is represented by the solid line
- The dashed line is the distance of every observation to the mean



Untabulated data – variance

If we take the average of the distance of each data point from the mean, we get 0 (why is that the case?).

Instead, we take its square to remove the sign.

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

The variance is (4+16+36+0+144) / 5 = 200 / 5 = 40

Alternative route	Travel time (x)	(x-mean)	$(x\text{-mean})^2$
Day 1	15	-2	4
Day 2	13	-4	16
Day 3	11	-6	36
Day 4	17	0	0
Day 5	29	12	144
Total	*85*	*0*	*200*
Average	*17*	*0?*	*40*

The variance of the alternative route is 40 minutes^2. The unit of variance is squared of the underlying unit. This makes it harder to explain or understand.

Untabulated data – standard deviation

To make it more comparable, we need to take its square root.
Standard deviation is the square root of the variance.

$$standard\ deviation = \sqrt{variance}$$

$$standard\ deviation = \sqrt{40} = 6.3\ minutes$$

The “average distance” from the mean of 17 minutes of the alternative route is 6.3 minutes.

Note: A less computationally intensive way to calculate standard deviation (and variance) is as follows:

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

Tabulated ungrouped data

Tabulated ungrouped data – range

Let’s consider tabulated ungrouped data structures now

To find the **range**, we find the minimum and the maximum and take the difference. Let’s look at **Example 4** from last week’s lecture as a demonstration.

Number of TV sets	3	4	5	6	7	8
Number of days	4	6	7	6	5	2

- Minimum: 3
- Maximum: 8
- **Range:** 8 – 3 = **5**

Tabulated data – interquartile range

Now let's consider **interquartile range**

To compute interquartile range:

Number of TV sets	3	4	5	6	7	8
Number of days (freq.)	4	6	7	6	5	2
Cumulative Frequency	*4*	*10*	*17*	*23*	*28*	*30*

Recall from previous week that

- Lower quartile: 4
- Upper quartile: 6

Interquartile range: 6 – 4 = **2**

Tabulated data – variance

- Finding the variance for tabulated data is similar to that of untabulated data. We just have to account for the frequency information provided.

The mean of this example was 5.3

$$variance = \frac{\sum[f * (x - \bar{x})^2]}{\sum f}$$

$$= \frac{63.9}{30} = 2.13\ tv\ sets^2$$

Standard deviation is its square root

$$std\ dev = \sqrt{2.13} = 1.46\ tv\ sets$$

No. of TV sets (observation)	No. of days (frequency)	$(x\text{-mean})^2$	$f(x\text{-mean})^2$
3	4	$(3\text{-}5.3)^2 = 5.1$	4*5.1 = 20.6
4	6	$(4\text{-}5.3)^2 = 1.6$	6*1.6 = 9.6
5	7	$(5\text{-}5.3)^2 = 0.1$	7*0.1 = 0.5
6	6	$(6\text{-}5.3)^2 = 0.5$	6*0.5 = 3.2
7	5	$(7\text{-}5.3)^2 = 3.0$	5*3.0 = 15.0
8	2	$(8\text{-}5.3)^2 = 7.5$	2*7.5 = 14.9
Total	*30*		*63.9*

Tabulated grouped data

Tabulated grouped data - range

Let's consider tabulated grouped data structures

The **range** is still the difference between the minimum and the maximum. However, we do not consider the midpoints.

We take the lower boundary of the first group for minimum and the upper boundary of the last group for maximum

Minimum = \$0

Maximum = \$50

Range = 50 – 0 = **50**

Expenditure on food	Number of respondents
$\$0 \le x < \$5$	2
$\$5 \le x < \$10$	6
$\$10 \le x < \$15$	8
$\$15 \le x < \$20$	14
$\$20 \le x < \$30$	12
$\$30 \le x < \$40$	6
$\$40 \le x < \$50$	2

Tabulated data – variance

Now lets consider **variance**. Recall from previous week that mean was 19.9

Let's use the computationally less intensive formula:

$$var = \frac{\sum f * mid^2}{\sum f} - mean^2$$

$$= \frac{24787.5}{50} - 19.9^2 = 99.7$$

Variance is 99.7 dollars^2

$$std\ dev = \sqrt{var}$$

$$= \sqrt{99.7} = \$9.99$$

Expenditure on food	Number of respondents	Midpoint	Mid^2	F*Mid^2
\$0 ≤ x < \$5	2	2.5	6.25	12.5
\$5 ≤ x < \$10	6	7.5	56.25	337.5
\$10 ≤ x < \$15	8	12.5	156.25	1250
\$15 ≤ x < \$20	14	17.5	306.25	4287.5
\$20 ≤ x < \$30	12	25	625	7500
\$30 ≤ x < \$40	6	35	1225	7350
\$40 ≤ x < \$50	2	45	2025	4050
Total	*50*			*24787.5*

The standard deviation of the data is approximately \$10.

Essential readings:

- Jon Curwin…, “Quantitative methods…”, Ch 6
- Glyn Burton…, “Quantitative methods…”, Ch 2.4
- Richard Thomas, “Quantitative methods…”, Ch 1.8-1.11
- Mik Wisniewski…, “Foundation Quantitative…”, Ch 7
- Clare Morris, “Quantitative Approaches…”, Ch 6
- Louise Swift “Quantitative methods…”, Ch DD2.