FUNDAMENTALS OF STATISTICS (4ECON013C)

TUTORIAL 1

Learning outcomes:

- calculate the central tendency and variation measurements;
- explain the limitations of these measurements;
- be able to identify the outliers in the dataset;
- provide the five-number summary;
- plot various graphs, including histogram, boxplot, line graph, etc.;
- understand the Chebyshev's rule.

1.1 Central Tendency measurements

Mean, median and **mode** are the three principal measures of central values.

Mean

Sample mean:	*Population mean:*
$\bar{x} = \frac{\sum x_i}{n}$	$\mu = \frac{\sum x_i}{N}$

Example.

Find the mean of the following dataset:

35 40 25 38 24 25 32 50 (1.1)

$$\bar{x} = \frac{35+40+25+38+24+25+32+50}{8} = \frac{269}{8} = 33.625$$

Weighted mean - a type of average that assigns different levels of importance (weights) to different values in a dataset, giving more significance to certain values than others.

$$\bar{x} = \frac{\sum w_i x_i}{w_i}$$

where w – weights, x – observation values.

Example. An investor is building up a particular stock in his portfolio. He purchases the same stock at different prices over time.

Price (x)	# of stocks (w)	x*w
\$50	50	2,500
\$55	100	5,500
\$48	70	3,360
Sums	220	11,360
Weighted mean		11,360/220 = **\$51.64**

Median is the middle (50th percentile) value, and its location is found by $\frac{n+1}{2}$.

If n is odd, there is an explicit middle value.

If n is even, you need to take the average of the two values either side of the midpoint (location).

Example. Find the median for the following dataset:

35 40 25 38 24 25 32 50

Step 1. Sort the data in ascending order: 24, 25, 25, 32, 35, 38, 40, 50.

Step 2. Location of the median is $\frac{8+1}{2} = 4.5$

Step 3. Take the average of 4th and 5th values: Median = (32+35)/2 = 33.5

Mode is the most frequent data point. In the above example, 25 is the mode, because it is occurring twice.

1.2 Measures of spread (variation)

Measures of spread (variation) are also important values when drawing conclusions. We will cover 4 of them: **range, inter-quartile range (IQR), variance** and **standard deviation.**

Range

The range is the difference between the largest and smallest value in the dataset.

$$\text{Range} = \text{Maximum} - \text{Minimum} = x_{(n)} - x_{(1)}$$

here, $x_{(n)}$ indicates the largest value and $x_{(1)}$ is the smallest one.

Example. For the dataset in (1.1), the range is

$$x_{(8)} - x_{(1)} = 50 - 24 = 26$$

Interquartile range (IQR)

IQR is defined as mid 50% range and found by: $IQR = Q_3 - Q_1$

where Q_3 and Q_1 are the third (upper) and first (lower) quartiles respectively.

To find the positions of quartiles, the following formulas are used:

Position of Q_1: (n-1)*0.25 + 1

Position of Q_3: (n-1)*0.75 + 1

Note that there are many different methods of finding the positions of quartiles, but the formula given here is consistent with the built-in functions of MS Excel.

Example. The sorted dataset from (1.1) is given below:

24, 25, 25, 32, 35, 38, 40, 50.

Let's find the position of Q_1: (8-1)*0.25 + 1 = 2.75. For Q_3: (8-1)*0.75 + 1 = 6.25

Now, using interpolation method, we will find these quartiles:

Q_1 = 25 + 0.75*(25-25) = 25 Q_3 = 38 + 0.25*(40-38) = 38.5

Hence, IQR = 38.5 – 25 = 13.5

Outliers are extreme values which are far away from the rest of the dataset. One method of deciding whether a value is an outlier is to find out if this value is outside this range:

$$(Q_1 - 1.5*IQR, Q_3+1.5*IQR)$$

If the value is outside, then it is an outlier.

Five-number summary indicates the following 5 measures:

Minimum	Q_1	Q_2 (Median)	Q_3	Maximum

Variance & Standard deviation

Variance and standard deviation are much better and more useful measures of the spread of the dataset.

Population variance: *Sample variance:*

$$\sigma^2 = \frac{\sum(x_i-\mu)^2}{N} = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{N}}{N} \qquad s^2 = \frac{\sum(x_i-\bar{x})^2}{n-1} = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1}$$

Standard deviation is simply the square root of variance:

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

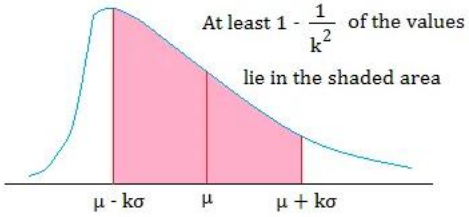
Example. Using the same (1.1) dataset, let's find the sample variance and sample standard deviation.

x	$(x - \bar{x})$	$(x - \bar{x})^2$
35	1.375	1.891
40	6.375	40.641
25	-8.625	74.391
38	4.375	19.141
24	-9.625	92.641
25	-8.625	74.391
32	-1.625	2.641
50	16.375	268.141
Total		S_{xx} = 573.875

$$s^2 = \frac{573.875}{8-1} = 81.98,\ s = \sqrt{81.98} = 9.05$$

1.3 Chebyshev's rule

Chebyshev's rule (or theorem) is a statistical principle that states a minimum proportion of any data set will fall within a certain number of standard deviations from the mean. This theorem applies to a broad range of probability distributions.



Chebyshev's theorem

Minimum proportion of observations that are within k standard deviations of the mean:

$$\mathbf{1 - \frac{1}{k^2}}$$

where k equals the number of standard deviations in which you are interested. K must be greater than 1.

Example. Suppose you randomly asked 45 schoolteachers in Tashkent about their monthly salary, and you obtained a sample mean of $420 and standard deviation of $40. Find the minimum proportion of all schoolteachers who earn between $360 and $480.

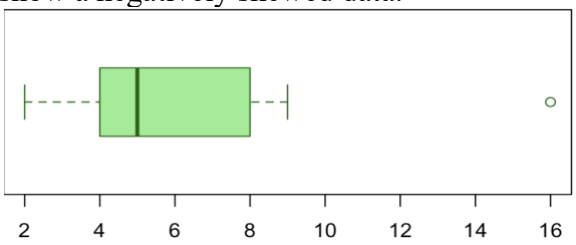
360 = 420 – k*40, and k = 1.5 (you can find through 480 = 420 + k*40 too)

At least $1 - \frac{1}{1.5^2} = 0.556$ or 55.6% percent of schoolteachers earn between $360 and $480.

TASKS

1. Decide whether the statement makes sense (or is clearly true) or does not make sense (or is clearly false). Explain clearly; not all of these statements have definitive answers, so your explanation is more important than your chosen answer.

a. A data set of incomes has modes of \$50,000 and \$80,000.
b. A researcher studying an income distribution obtains the same value of \$75,000 for the mean, median, and mode.
c. Jennifer received an SAT score that was equal to the first quartile and the 35th percentile.
d. The house key lengths of 15 statistics students are measured and rounded to the nearest centimeter, and all 15 values are the same, so the standard deviation is 0 cm.
e. When there are outliers in the dataset, the median is a better measure of central tendency compared to the mean.
f. The following boxplot show a negatively skewed data.



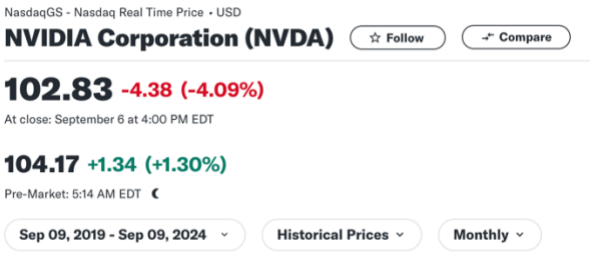
2. The monthly incomes of a sample of middle-management employees at KPMG Uzbekistan are given below:

Monthly salary (in USD)	# of employees
\$2000	5
\$1900	4
\$3000	3
\$2500	6

a. Compute the sample mean, median and mode.
b. If $S_{xx} = \sum(x_i - \bar{x})^2 = 2{,}797{,}778$, then find the sample variance and standard deviation.

3. The following data provides the monthly closing stock price of NVIDIA Corporation (NVDA) rounded to the nearest integer:

Month	Close Price (\$)
Sep 1, 2024	\$103
Aug 1, 2024	119
Jul 1, 2024	117
Jun 1, 2024	124
May 1, 2024	110
Apr 1, 2024	86
Mar 1, 2024	90
Feb 1, 2024	79
Jan 1, 2024	62
Dec 1, 2023	50
Nov 1, 2023	47
Oct 1, 2023	41
Sep 1, 2023	44



a. Find the range.
b. Find the IQR.
c. Provide the five-number summary.

4. Annual incomes (defined as X, in thousand $) of randomly surveyed households are given in the following table:

52	23	66	41	93	27	98	84
22	77	80	45	30	99	42	46
88	26	53	58	78	47	44	66

If $\sum x = 1{,}385$ and $\sum x^2 = 93{,}905$, do the following:
a. Find the mean and median.
b. Compute sample variance and standard deviation.
c. Plot the box plot and comment on the presence of outliers.

5. Suppose two-bedroom apartment rental prices in Tashkent have an average of $350 with a standard deviation of $60. If nothing is known about the shape of the distribution, at least what percentage of rental prices lie between $260 and $440?

6. The mean score on an accounting test is 80, with a standard deviation of 10. Suppose the scores follow a bimodal distribution. Between which two scores must this mean lie to represent at least 96% of the data set? Note that the student scores can range between 0 and 100.

7. Work in groups of 3-4 students to discuss this task.
All 100 first-year students at a small college take three modules in the Core Studies program. Two modules are taught in large lectures, with all 100 students in a single class. The third module is taught in 10 classes of 10 students each. Students and college administrators get into an argument about whether classes are too large. The students claim that the mean size of their Core Studies classes is 70. The administrators claim that the mean class size is only 25. Can both sides be right? Explain.

HOMEWORK

8. For the distribution drawn below comment on the skewness of the plots. Identify the points for mean, median and mode.

9. The manager of a local RV sales lot has collected data on the number of RVs sold per month for the last five years. That data is summarized below:

# of Sales	0	1	2	3	4	5	6
# of Months	2	6	9	13	21	7	2

What is the weighted mean number of sales per month?

10. The data is given regarding the distance in miles between exits on I-75 in Kentucky. Find the modal distance.

11	4	10	4	9	3	8	10	3	14	1	10	3	5
2	2	5	6	1	2	2	3	7	1	3	7	8	10
1	4	7	5	2	2	5	1	1	3	3	1	2	1

What is the modal distance?

11. A bored carpenter counts the actual number of nails in 10 boxes of nails and records his findings as: 230, 235, 302, 287, 312, 323, 265, 319, 342, and 298. What can we say about the shape of the distribution of the number of nails? (skewed to the left)

12. Suppose Adam has assessment components midterm exam, group project and final exam with 30%, 30% and 40% weights respectively. He received 58 from midterm and 72 from the group project. What is the minimum score he needs from final exam to receive at least 70 overall mark?

13. The following data is given.

1	11.5	6	7.2	4	8	9
10	6.8	8.3	2	2	10	1

a. Find the range.
b. Find the IQR.
c. Provide five-number summary.