WESTMINSTER
International University in Tashkent

Fundamentals of Statistics (FoS)
WEEK 1

Module leader: Olmos Isakov
Email: o.isakov@wiut.uz

Lecture Outline

- Introduction
- Science of Statistics
- Applications of Statistics
- Central tendency measures
- Measurements of variation
- Outliers, five-number summary
- Chebyshev's rule

WEEKLY SESSIONS

OVERALL MARK = 40%* In-class Test + 60% * Final Exam

Pass Mark $\geq$ 40

Condoned Credit → 30 $\leq$ Mark $\leq$ 39 (Eligibility applies)

Textbooks

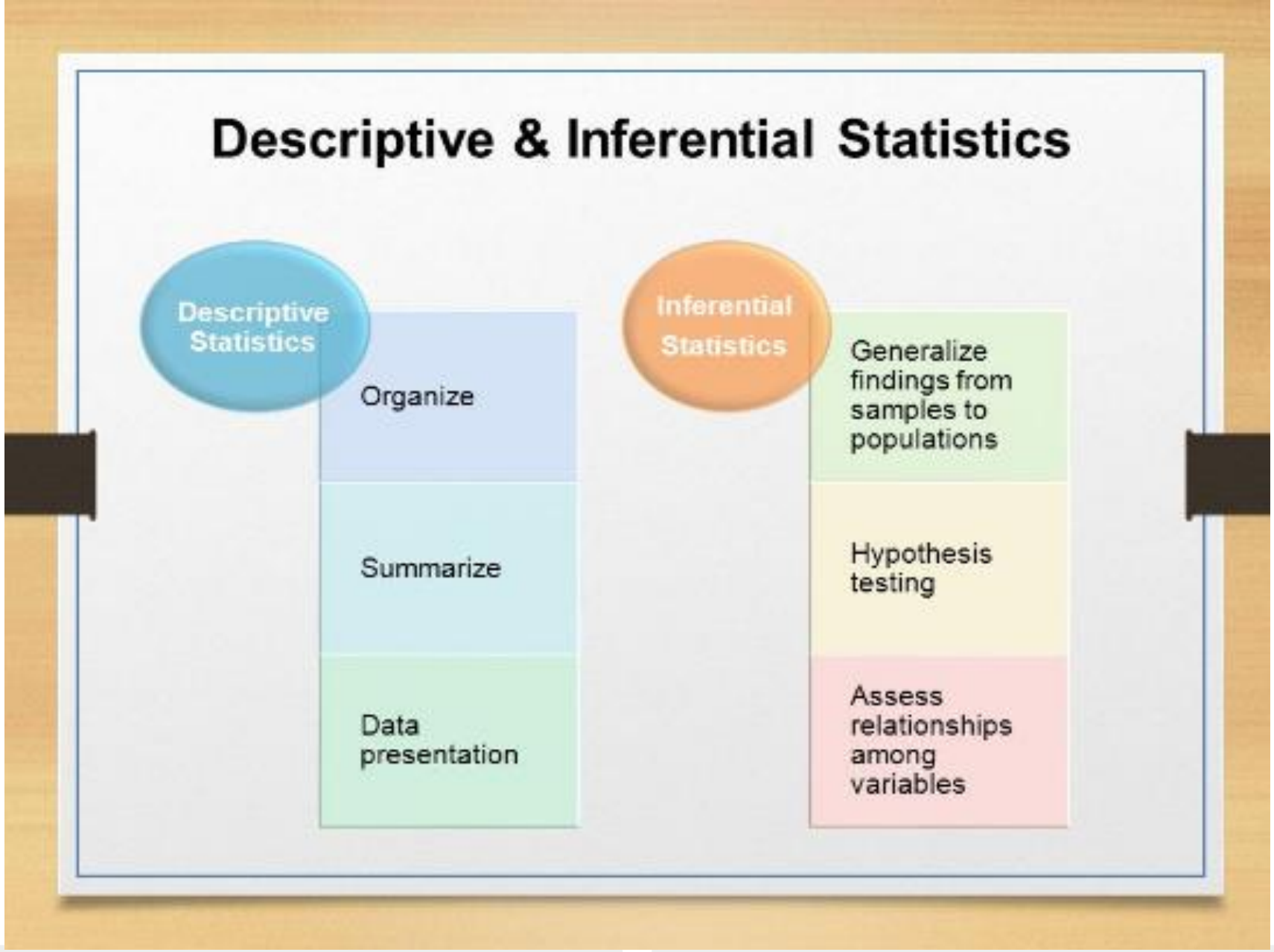
1. Statistics 12th or 13th Edition, Authors: James T. McClave & Terry Sincich, ISBN 978-0-321-75593-3.
2. Basic Statistics for Business & Economics. Authors: Lind, Marchal & Wathen. ISBN 978-1-260-18750-2.
3. An Introduction to Statistical Methods and Data Analysis. Authors: Ott & Longnecker. ISBN 978-0-495-01758-5
4. Statistical Reasoning for Everyday Life. Authors: Bennett, Briggs, Triola, ISBN 978-1-292-04021-9
5. Statistics for Business and Economics. Author: Newbold, Carlson & Thorne. ISBN 978-1292436845

Fundamentals of Statistics

Data → collecting, analysing, presenting } Making inferences, conclusions

Statistics & Data Careers:

- Statistician
- Actuary
- Data Analyst
- Data Scientist
- Quantitative Analyst ("Quant")
- Biostatistician


Descriptive & Inferential Statistics
Descriptive Statistics
Organize
Summarize
Data presentation
Inferential Statistics
Generalize findings from samples to populations
Hypothesis testing
Assess relationships among variables

Descriptive Statistics

Which Group is Smarter?

IQ scores of Group A:

102, 115, 128, 109, 131, 89, 98, 106, 140, 119, 93, 97, 110

IQ scores of Group B:

127, 162, 131, 103, 96, 111, 80, 109, 93, 87, 120, 105, 109

Descriptive Statistics

Average IQ scores:

Group A	Group B
110.54	110.23

They're roughly the same!

With a summary descriptive statistics,
it is much easier to answer our question.

Central Tendency Measurements

Mean: **average of all values (μ – *for population*, $\bar{x}$ – *for sample*)**

Median: **The middle value when a variable's values are ranked in order**

Mode: **The most common (frequent) data point (unimodal vs bimodal)**

Mean (Average)

Most commonly called the “average.”

μ – for population

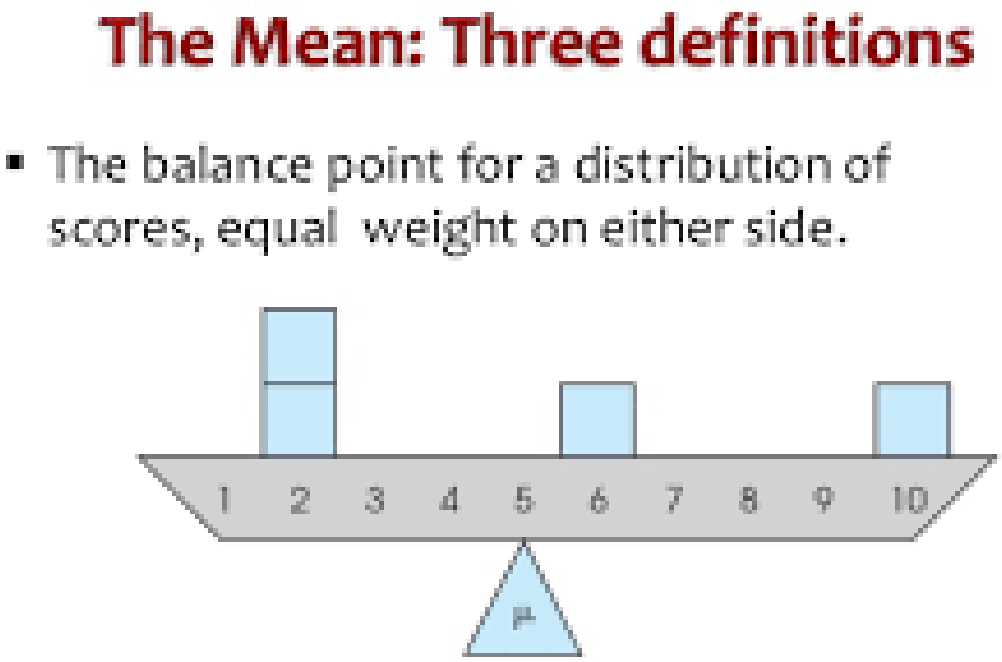
$\bar{x}$, $\bar{y}$ - for sample

Add up the values for each case and divide by the total number of cases.

$$\overline{X} = \frac{X_1+X_2+X_3+ \ldots+X_n}{n} = \frac{\sum X_i}{n}$$

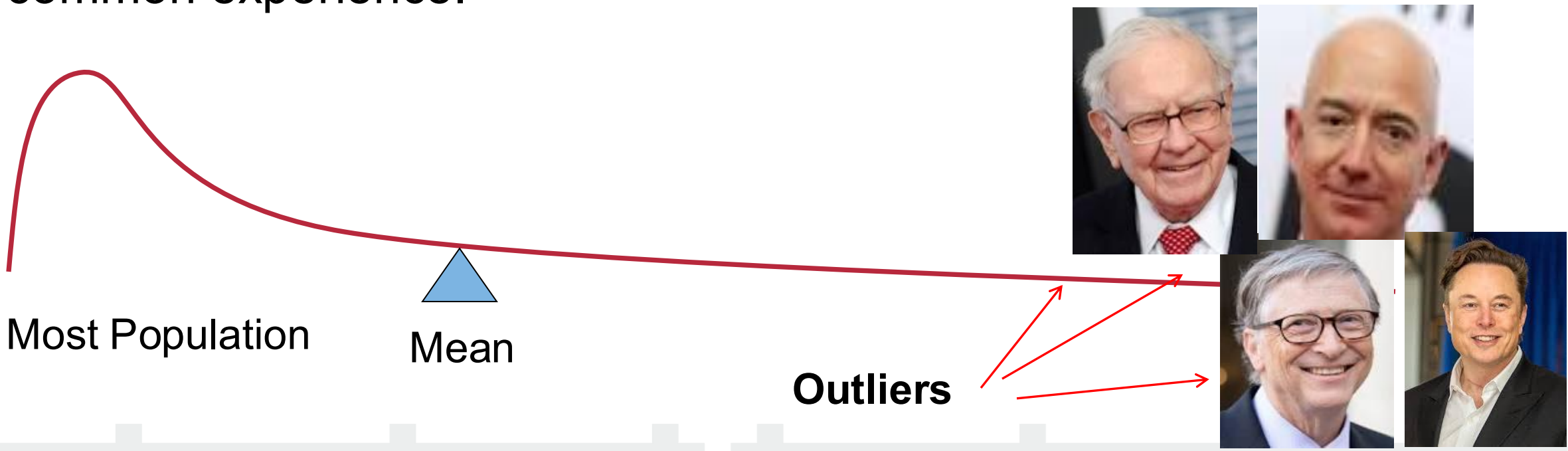
Ex. X: {2, 4, 6}

$$\overline{X} = \frac{X_1+X_2+X_3}{3} = \frac{2+4+6}{3} = 4$$



Limitations of mean

1. Means can be badly affected by outliers (data points with extreme values unlike the rest)
2. Outliers can make the mean a bad measure of central tendency or common experience.



Median: 50th Percentile

The middle value when a variable's values are ranked in order; the point that divides a distribution into two equal halves.

When data are **listed in order,** the median is the point at which 50% of the cases are above and 50% below it.

The 50^{th} percentile.

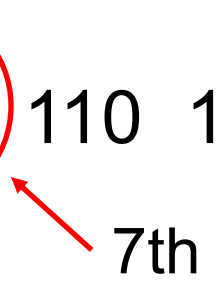
Position of Median = $\frac{n+1}{2}$ (if not integer, then take the average of middle two values).

Example: Median

Class A--IQs of 13 Students

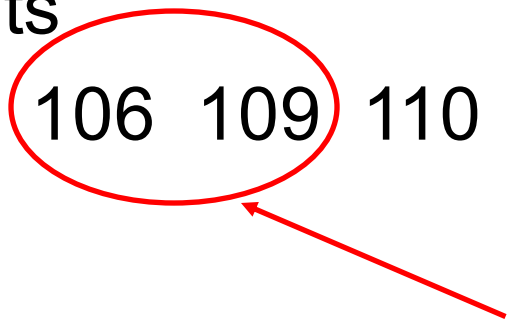
89 93 97 98 102 106 109 110 115 119 128 131 140

Median = 109

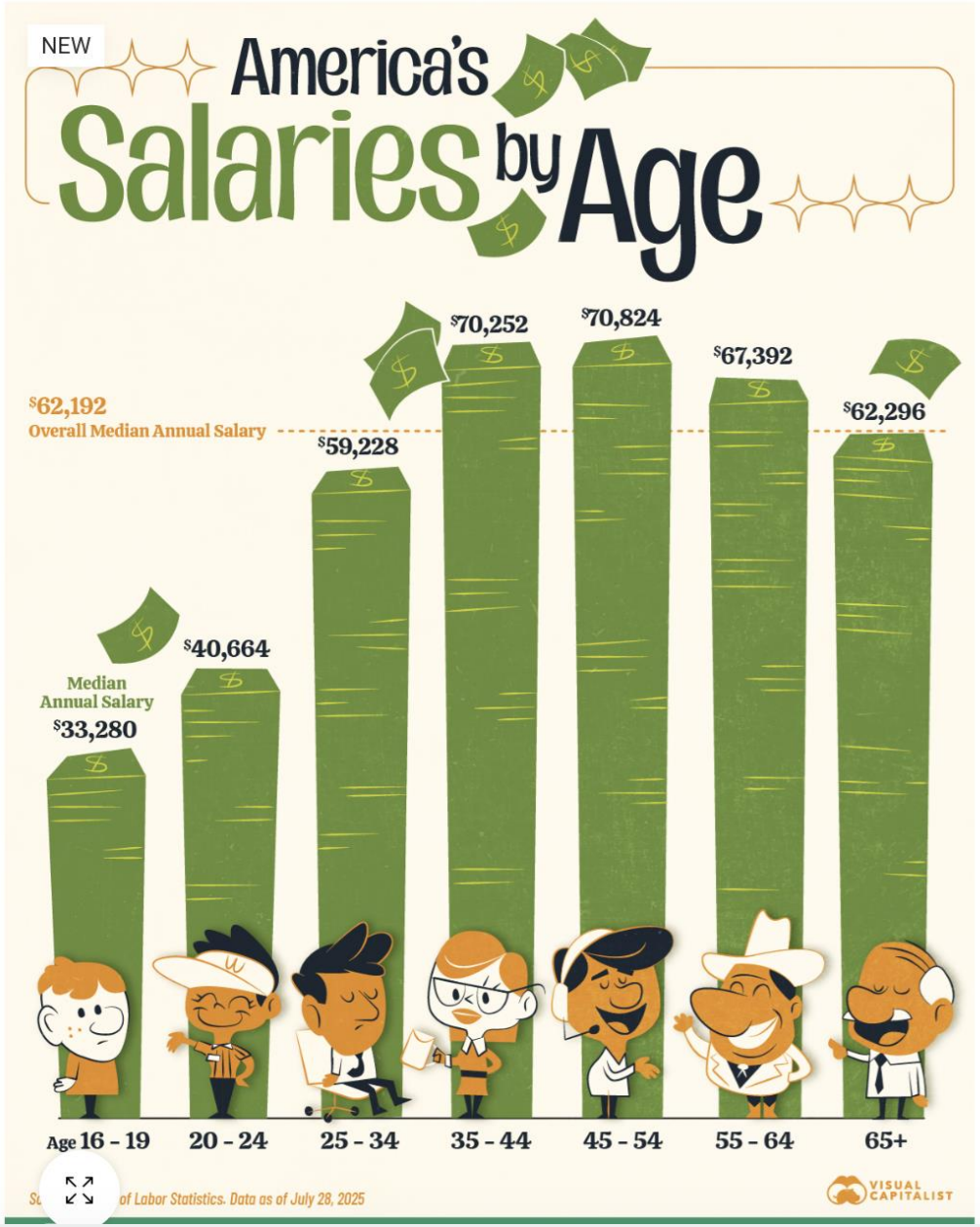


Class A--IQs of 12 Students

89 93 97 98 102 106 109 110 115 119 128 131 ~~140~~



$$\text{Median} = \frac{106+109}{2} = 107.5$$



Charted: Median U.S. Salaries by Age Group

1 Credit

- 1 + ADD TO CART

Compare Products

SKU Number: N/A

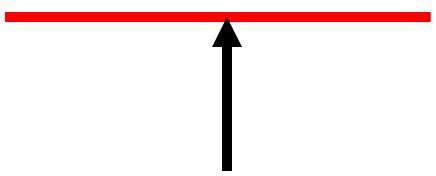
Categories: Money , United States

Mode

- The most common data point is called the mode.
- The combined IQ scores for Classes A & B:

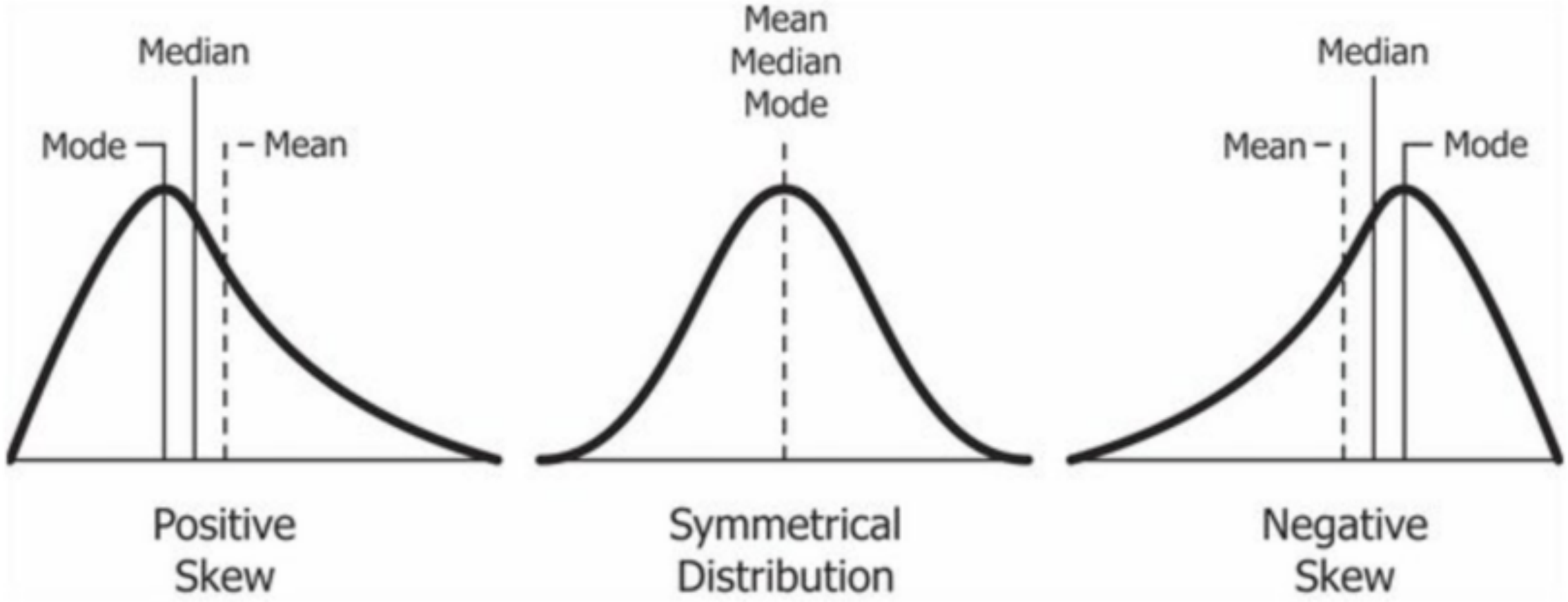
  80 87 89 93 93 96 97 98 102 103 105 106 109 109 109 110 111 115

  119 120 127 128 131 131 140 162



A la mode!!

BTW, It is possible to have more than one mode!



A general relationship of mean and median under differently skewed unimodal distribution

Why Variation Matters

Suppose you are a financial advisor for an elderly couple looking to invest in a stock portfolio. Each portfolio given below has an equal mean and median annual return of 5%. Which portfolio would you recommend to this couple?

Portfolio A (%)	4.8	4.85	4.9	4.95	4.99	5	5.01	5.05	5.1	5.15	5.2
Portfolio B (%)	-20	-10	-5	0	2	5	8	10	15	20	30

Measures of Variation:

1) Range 2) IQR
3) Variance 4) Standard deviation

Range: **The spread between the lowest and highest values of a variable.**

Range = Maximum – Minimum

Portfolio A	4.8	4.85	4.9	4.95	4.99	5	5.01	5.05	5.1	5.15	5.2
Portfolio B	-20	-10	-5	0	2	5	8	10	15	20	30

Portfolio A: Range = 5.2 – 4.8 = 0.4 percentage points
Portfolio B: Range = 30 – (-20) = 50 percentage points

Inter-Quartile Range (IQR)

IQR: **The spread between the 1st and 3rd quartile (*middle 50% range*)**

Interquartile Range (IQR)

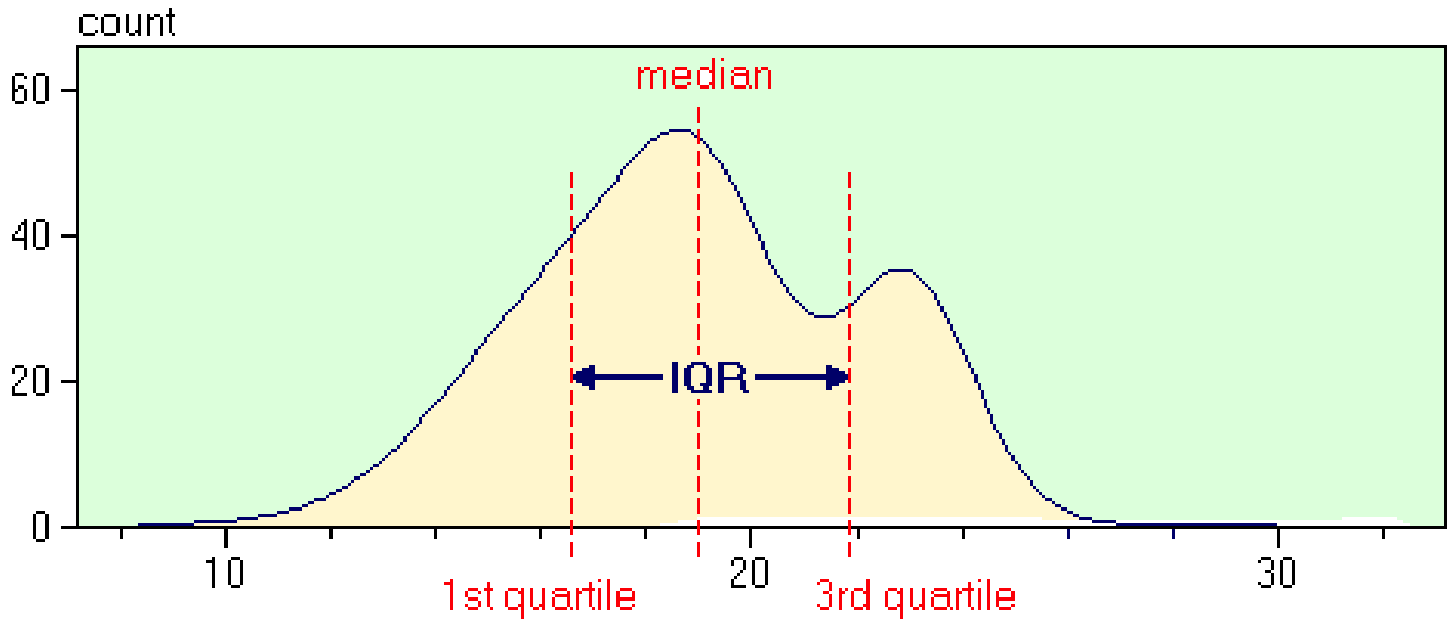
Interquartile range (IQR), also called the midspread, middle 50%, is a measure of statistical dispersion, being equal to the difference between 75th and 25th percentiles, or between upper and lower quartiles.

$$IQR = Q_3 - Q_1$$

Q1 - 25th Percentile
Q2 - 50th Percentile (Median)
Q3 - 75th Percentile

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

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



Example: IQR

Monthly sales data at local auto dealer is given:

Jan	Feb	Mar	Apr	May	June
30	25	34	32	40	50
Jul	Aug	Sep	Oct	Nov	Dec
46	52	35	41	36	48

Solution. Sort the data in ascending order:

25	30	32	34	35	36	40	41	46	48	50	52

Position of $Q_1 = (12-1)*0.25 + 1 = 3.75$ $\quad Q_1 = 32 + 0.75*(34-32) = 33.5$

Position of $Q_3 = (12-1)*0.75 + 1 = 9.25$ $\quad Q_3 = 46 + 0.25*(48-46) = 46.5$

IQR = 46.5 – 33.5 = 13

Outliers

25	30	32	34	35	36	40	41	46	48	50	52

If there is any observed value **outside** the following range:

(Q_1 – 1.5*IQR, Q_3+1.5*IQR), then it is called an outlier.

For our example,

Q_1 – 1.5*IQR = 33.5 – 1.5*13 = 14

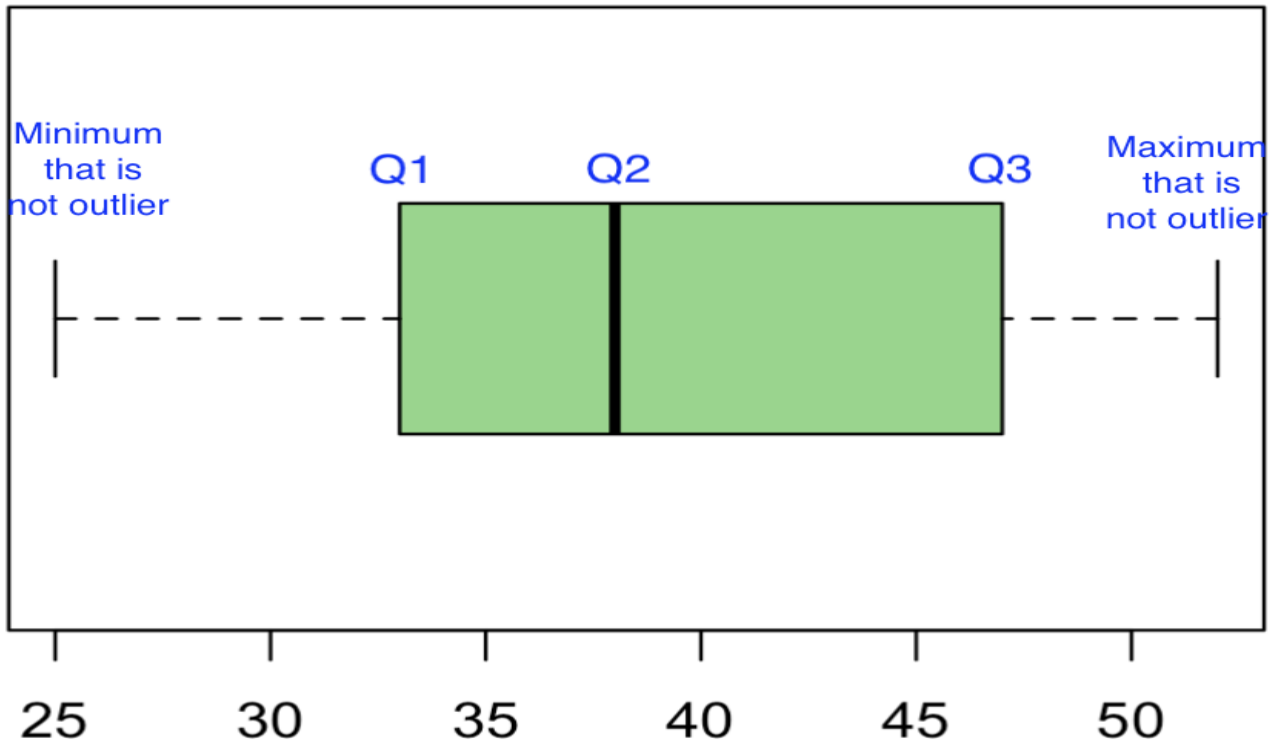
Q_3 + 1.5*IQR = 46.5 + 1.5*13 = 66

So, there is no outlier in this dataset.

Five number summary:

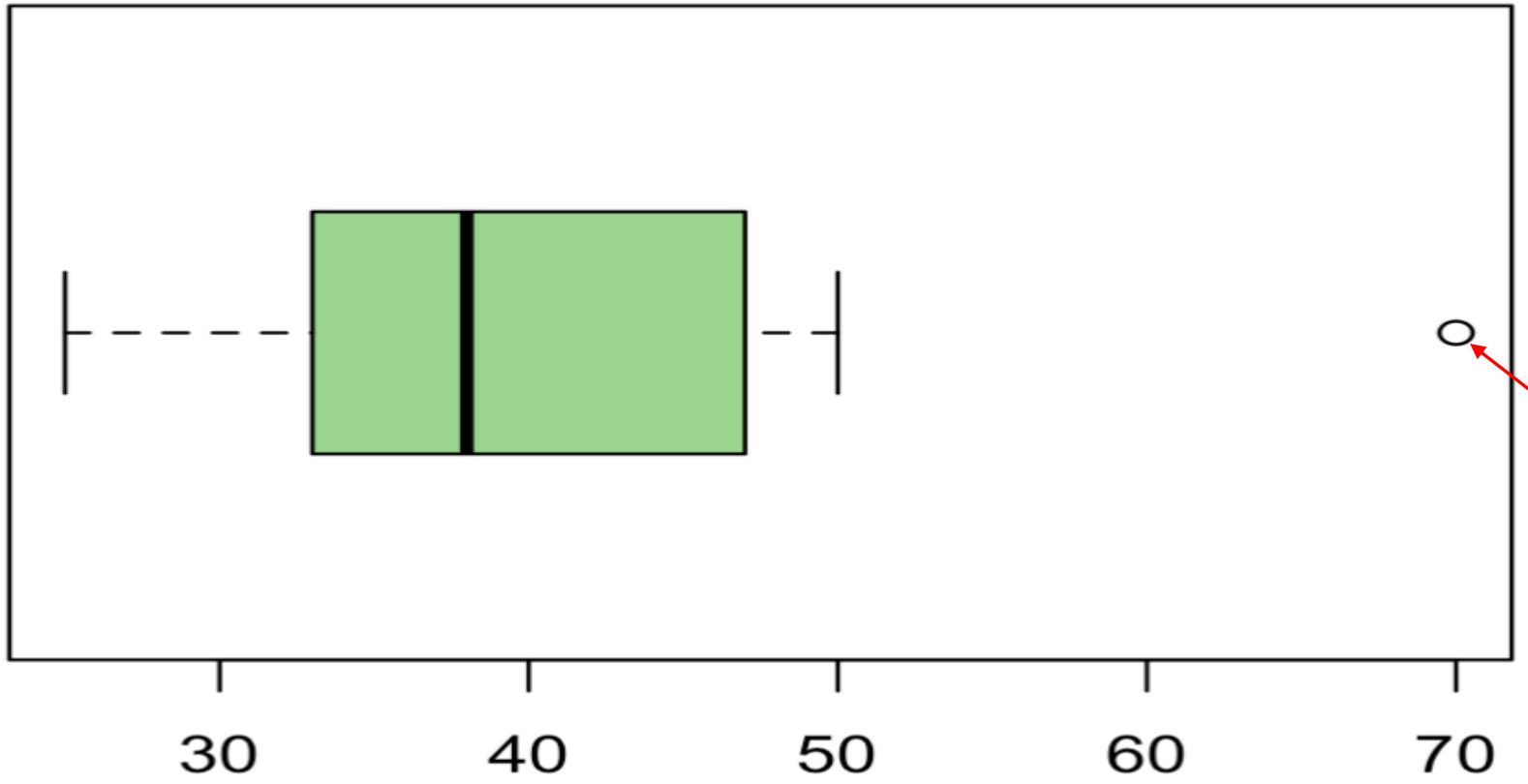
Min Q1 Q2 (Median) Q3 Max

25	30	32	34	35	36	40	41	46	48	50	52



Example with an outlier

25	30	32	34	35	36	40	41	46	48	50	70



Outlier

Variance & Standard deviation

- A measure of the spread of the recorded values on a variable.
- The larger the variance, the further the individual cases are from the mean.

Population Variance

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

σ^2 = population variance

x_i = value of i^{th} element

μ = population mean

N = population size

Sample Variance

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

s^2 = sample variance

x_i = value of i^{th} element

$\bar{x}$ = sample mean

n = sample size

Quick reminder:

the square root of the variance is standard deviation (σ *or* s)

Example

Portfolio A	4.8	4.85	4.9	4.95	4.99	5	5.01	5.05	5.1	5.15	5.2
Portfolio B	-20	-10	-5	0	2	5	8	10	15	20	30

x	$x - \bar{x}$	$(x - \bar{x})^2$
4.8	-0.2	0.040
4.85	-0.15	0.023
4.9	-0.1	0.010
4.95	-0.05	0.002
4.99	-0.01	0.000
5	0	0.000
5.01	0.01	0.000
5.05	0.05	0.002
5.1	0.1	0.010
5.15	0.15	0.023
5.2	0.2	0.040

y	$y - \bar{y}$	$(y - \bar{y})^2$
-20	-25	625
-10	-15	225
-5	-10	100
0	-5	25
2	-3	9
5	0	0
8	3	9
10	5	25
15	10	100
20	15	225
30	25	625

Portfolio A:

$$s_A^2 = \frac{0.15}{10} = 0.015$$

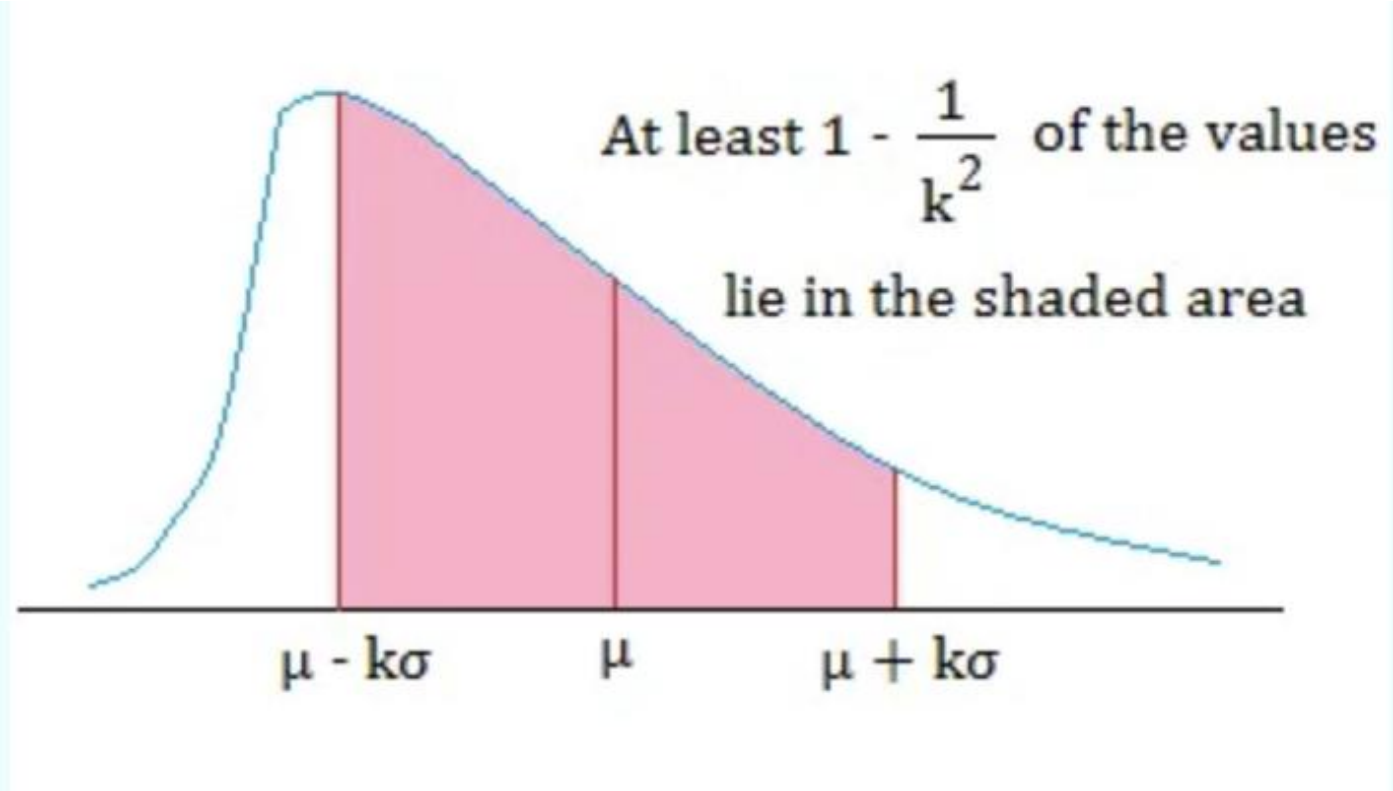
$$s_A = 0.39\%$$

Portfolio B:

$$s_B^2 = \frac{1968}{10} = 196.8$$

$$s_B = 14.03\%$$

The **Chebyshev's rule** applies to any data set regardless of the shape of the frequency distributions



At least **75%** measurement falls within two standard deviations of the mean

At least **89%** measurement falls within three standard deviations of the mean

Example: Chebyshev's rule

Suppose average income in Tashkent is $400 and standard deviation is $75. If no information is given about the shape of the distribution:

- What percent of population in Tashkent earn between $250 and $550?

$$(\$250, \$550) \Rightarrow 400 \pm 150 \Rightarrow 150 = 75*k \Rightarrow k = 2$$

$$(1 - \frac{1}{2^2}) = 0.75 \Rightarrow \text{At least } 75\%$$

- What is the income interval for the 89% of population around the mean?

$$(1 - \frac{1}{k^2}) = 0.89 \Rightarrow k = 3$$

$$400 \pm 3*75 \Rightarrow (\$175, \$625)$$

MS Excel functions

=average()	mean
=median()	median
=mode()	mode
=var() or =var.s()	sample variance
=var.p()	population variance
=stdev() or stdev.s()	sample standard deviation
=stdev.p()	population standard deviation
=max()–min()	range

=QUARTILE.INC(range, 3) - QUARTILE.INC(range, 1) ⟹ IQR

=PERCENTILE.INC(range, 0.75) - PERCENTILE.INC(range, 0.25) ⟹ IQR

THANK YOU