Descriptive Statistics

Elias Zintzaras, M.Sc., Ph.D.

Professor in Biomathematics-Biometry
Department of Biomathematics
School of Medicine
University of Thessaly

Institute for Clinical Research and Health Policy Studies
Tufts University School of Medicine
Boston, MA, USA

Theodoros Mprotsis, MSc, PhD
Teacher & Research Fellow
(http://biomath.med.uth.gr)
University of Thessaly
Email: tmprotsis@uth.gr

Basic Statistical Terms

Basic Statistical Terms

- In medical research and clinical practice, we collect data from a **sample** of individuals to draw conclusions about the broader **population** to which the sample belongs

- **Example:** If we want to investigate the relationship between a pregnant woman's weight gain during pregnancy and the weight of the newborn, we need to study a sample of pregnant women. It is not possible to study all pregnant women

Basic Statistical Terms

The following are some of the most basic terms used in statistical methodology:

- **Population**: a **population** is the entire group that you want to draw conclusions about
- **Sample**: a sample is a subset of individuals, items, or observations selected from a larger group or population
- **Variable** (denoted as X or x): a variable is any characteristic, number, or quantity that can be measured or counted, such as body weight gain
- **Observation**: an observation is **a value of something of interest you're measuring or counting during a study or experiment**: a person's height

Types of variables

A variable is **qualitative** when it takes discrete values and **quantitative** when it takes values on a continuous scale

- **Qualitative variables** include gender (e.g. male, female), hemoglobin stabilization in kidney patients (e.g. stabilized, not stabilized), survival (e.g. survives, dies), treatment outcome (e.g. improved, not improved), and amount of medicine taken (e.g. small, medium, large)
- **Quantitative** include height, weight, blood pressure, age, and so on

Null and alternate hypothesis

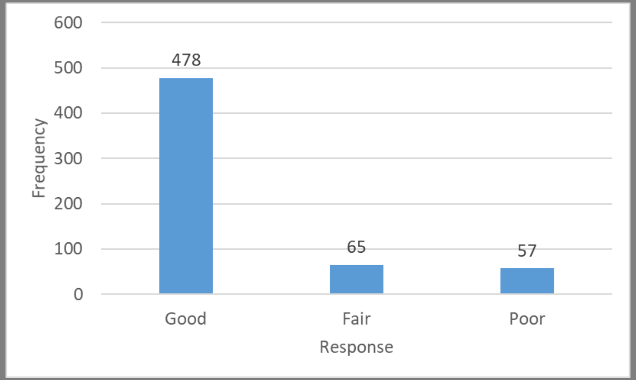
- **Null hypothesis (H_0):**
  The null hypothesis is a statement that there is **no effect, no difference, or no relationship** between variables. The null hypothesis is typically the default or baseline hypothesis that researchers seek to test against

  Example: In a clinical trial comparing a new drug to a placebo, the null hypothesis might state: "There is no difference in effectiveness between the new drug and the placebo"

- **Alternative hypothesis (H_1):**
  The alternative hypothesis is a statement that indicates the **presence of an effect, difference, or relationship** between variables. The alternative hypothesis is considered if the null hypothesis is rejected based on the data.

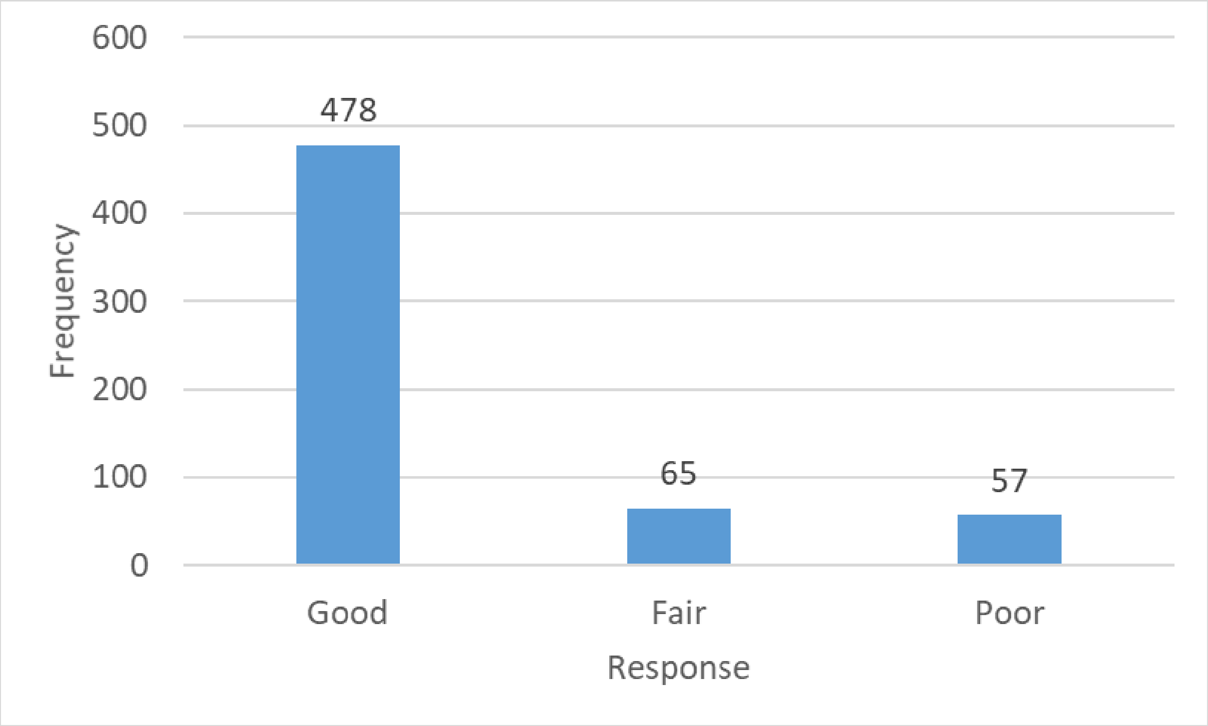
  Example: Using the same clinical trial, the alternative hypothesis might state: "The new drug is more effective than the placebo. Alternatively, there is a difference in the effect between the two drugs"

Graphical Methods of Data Description



Data visualization

It is important to always start the analysis with a graph to visualize the data

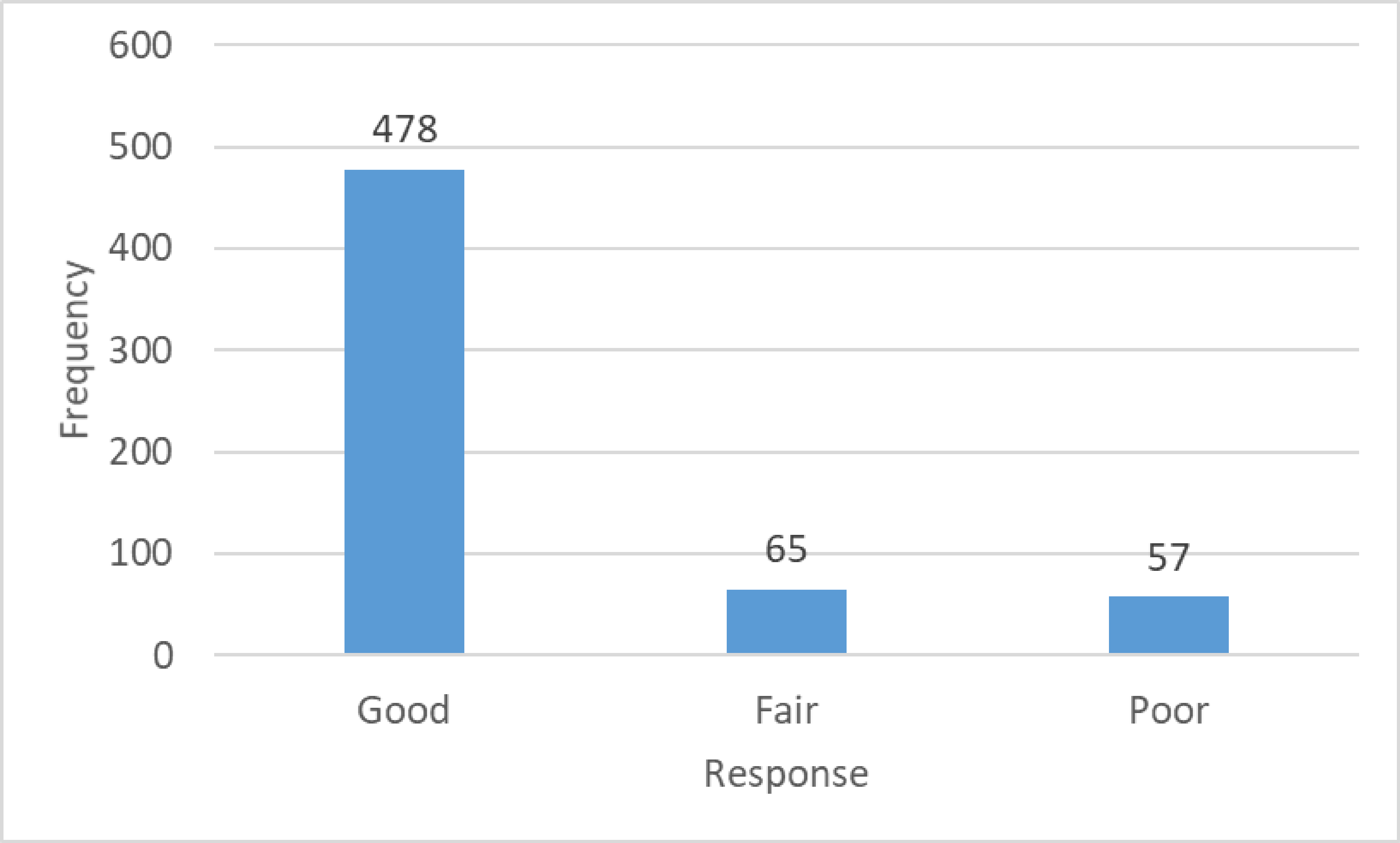


Bar charts

To show the frequency of a **qualitative variable**, we use a **bar chart**.

Example: 600 patients participated in a clinical study. Their responses to the treatment were categorized as **good**, **fair**, or **poor**. The qualitative variable investigated in this study is the **response to treatment**.

Response	N
Good	478
Fair	65
Poor	57
Total	**600**



The height of each bar is proportional to the corresponding frequency

Relative and Percent Frequencies

- The previous frequency table provides us with some information. For example, the value **Good** has a frequency of 478 (i.e., 478 patients had a good response to treatment)
- However, this frequency (i.e., the number 478) has limited meaning on its own if the total number of patients who participated in the treatment is not reported
- To find the **relative frequency** of a value, we **divide** the frequency of that value by the **total number** of observations
- We can then express this result as a percentage (%)

Relative and Percent Frequencies

$$\text{Relative frequency} = \frac{\text{Frequency of value}}{n}$$

$$\text{Percent frequency} = \frac{\text{Frequency of value}}{n} \cdot 100$$

Relative and Percent Frequencies

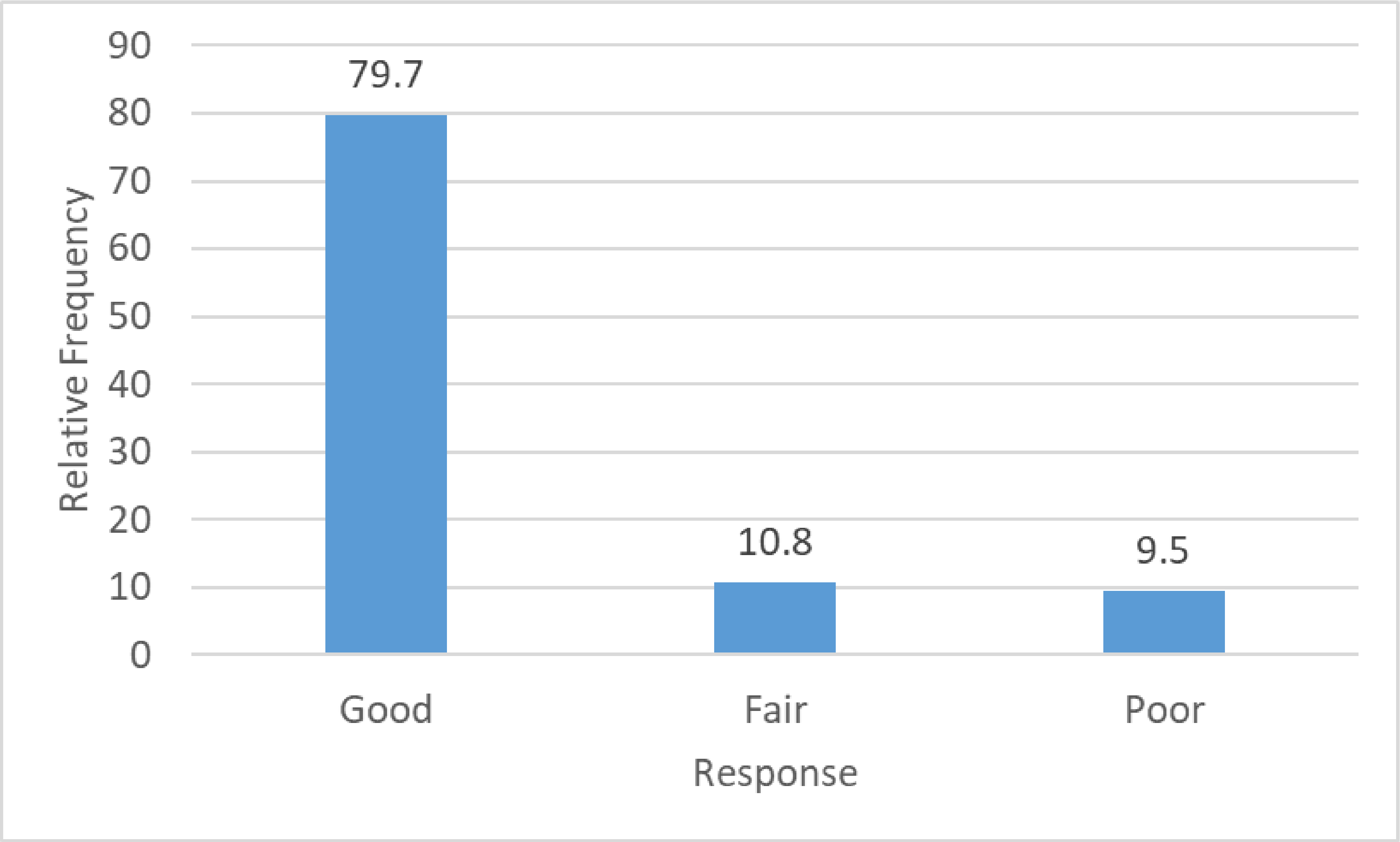
$$\text{Relative frequency “Good”} = \frac{478}{600}$$

$$\text{Relative frequency “Good”} = 0.797$$

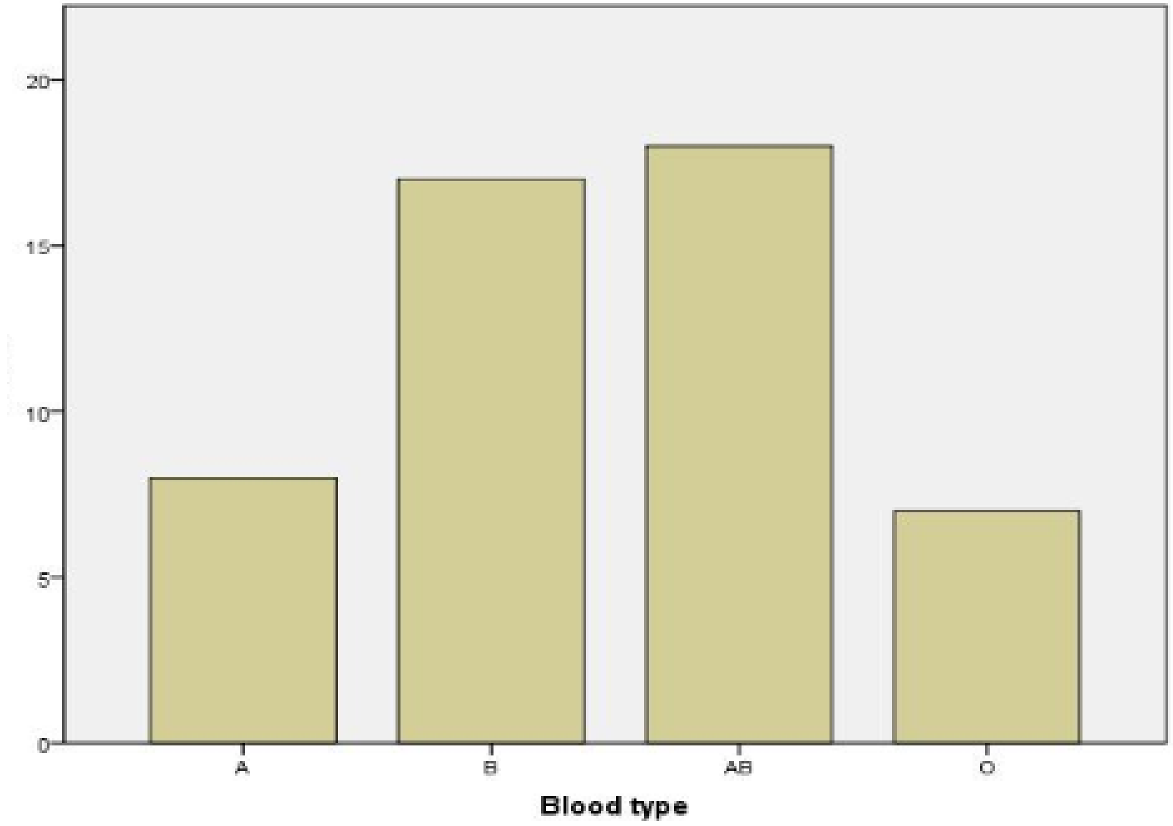
$$\text{Percent frequency “Good”} = 0.797 \cdot 100$$

$$\text{Percent frequency “Good”} = 79.7\%$$

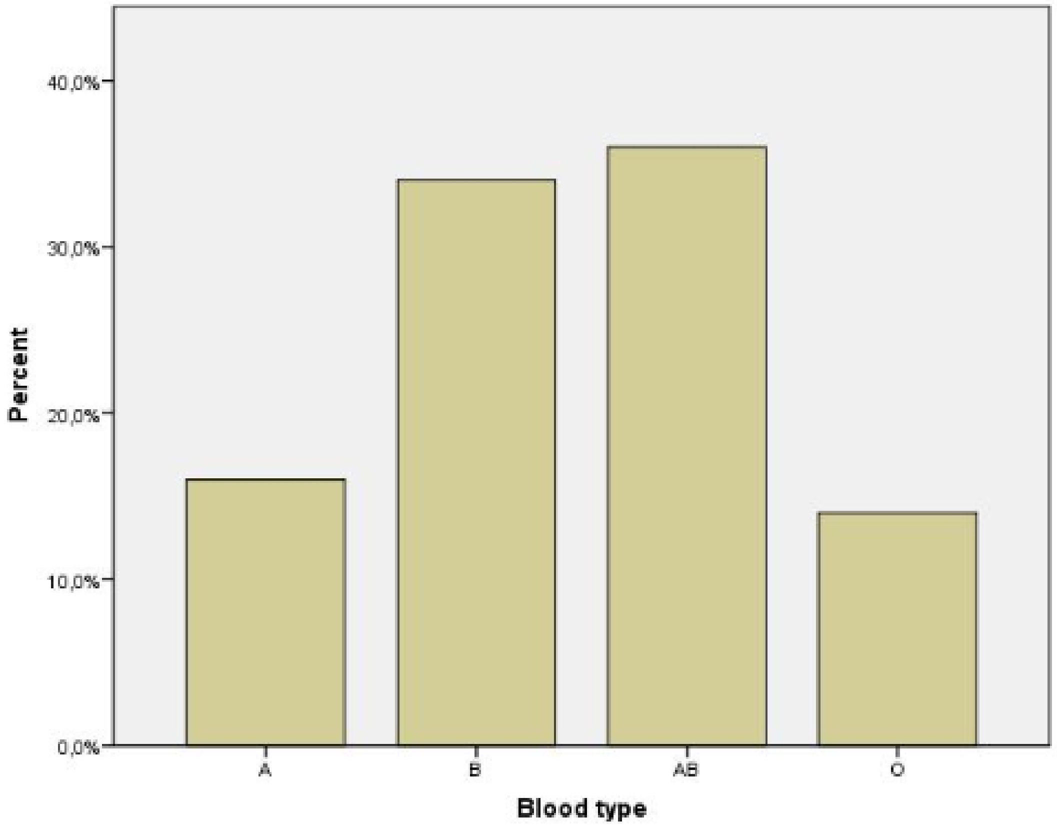
Response	N	Percent Frequencies (%)
Good	478	79.70%
Fair	65	10.80%
Poor	57	9.50%
Total	**600**	**100%**



The height of each bar is proportional to the corresponding relative frequency



20
15
10
5
0
A
B
AB
O
Blood type

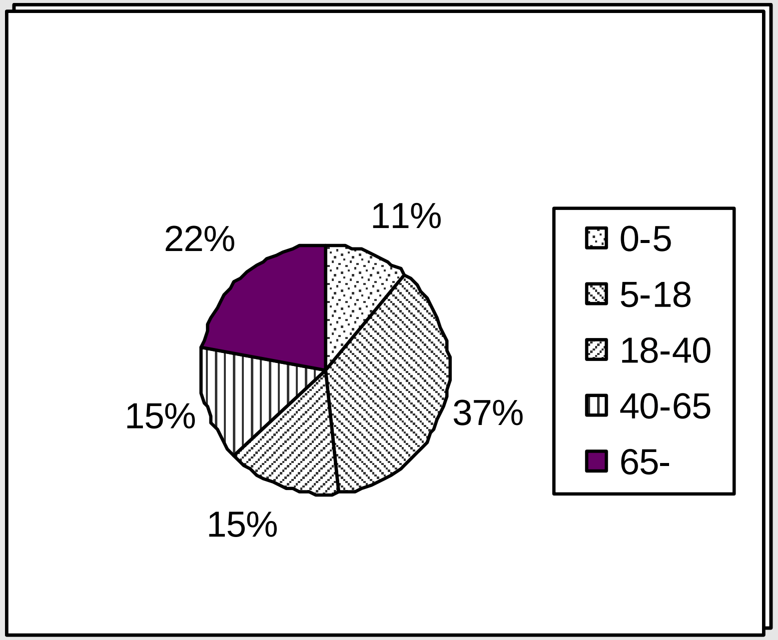


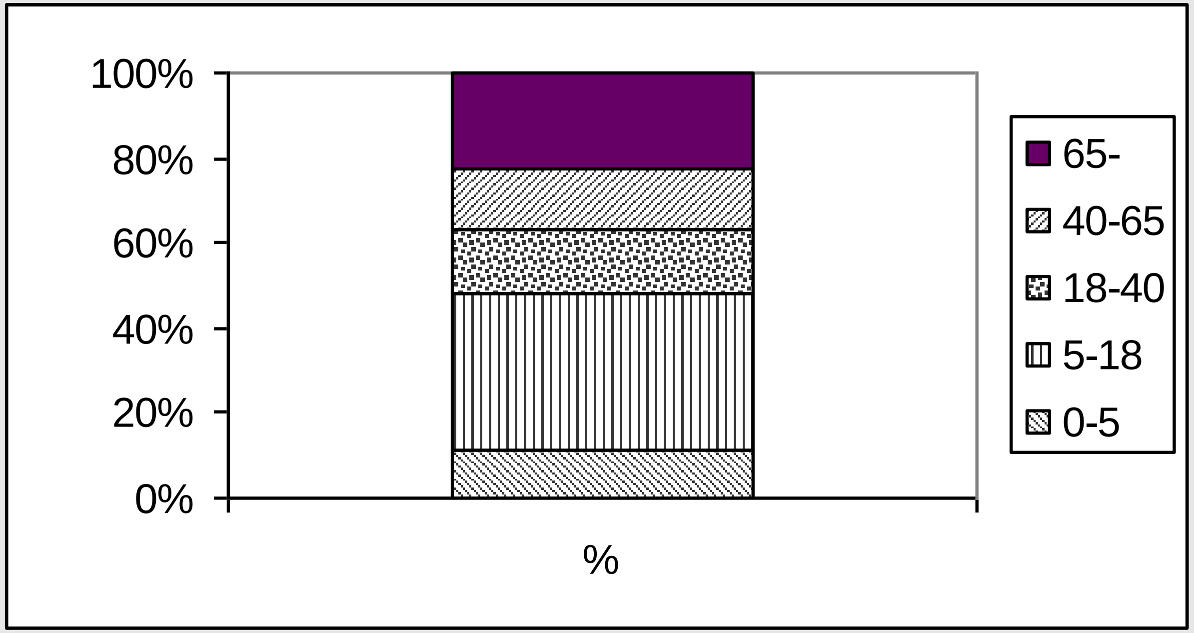
Percent
40,0%
30,0%
20,0%
10,0%
0,0%
A
B
AB
O
Blood type

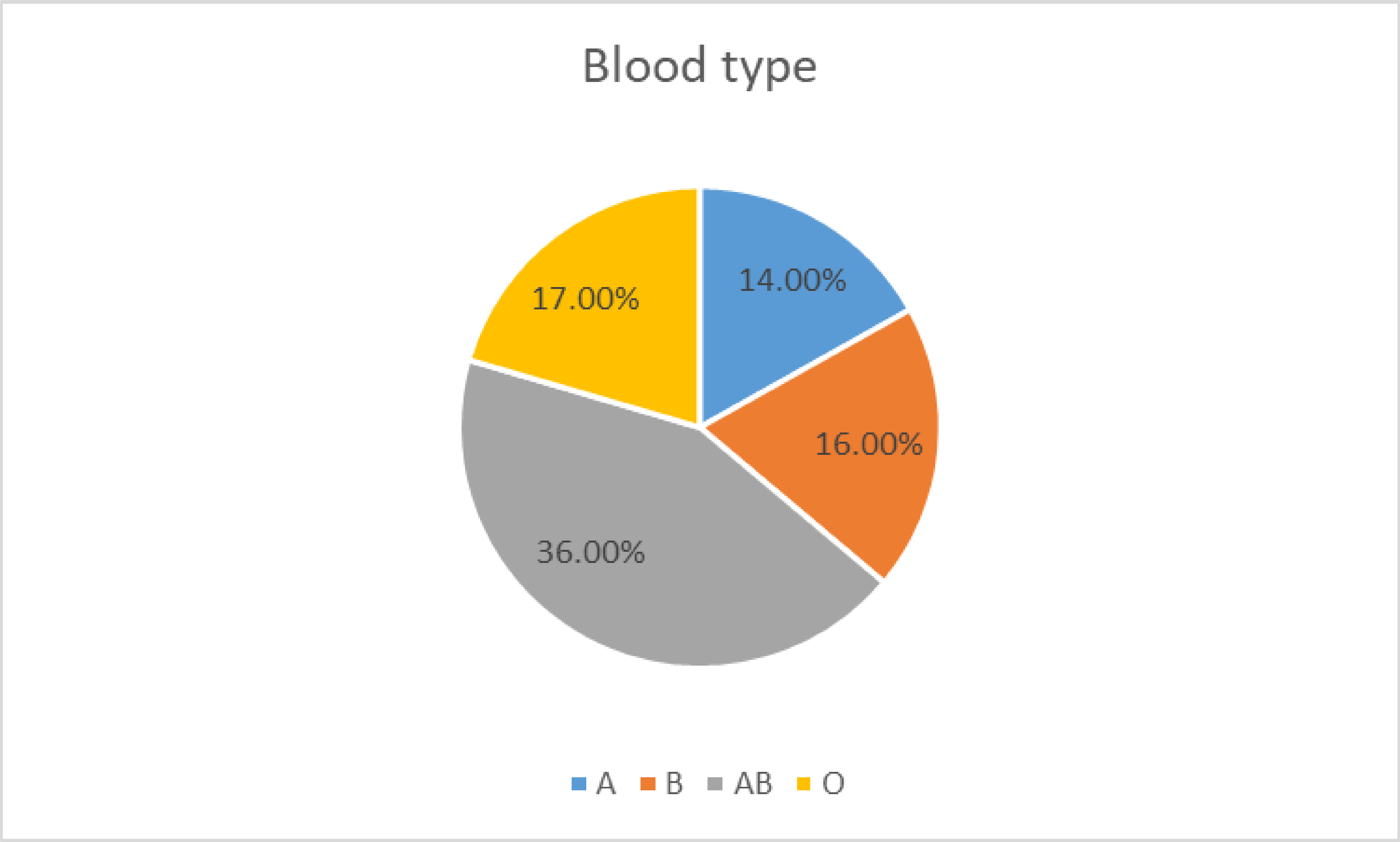
Pie chart

A **pie chart** is a circle divided into slices, where each slice represents the values of a variable for different categories.

Example: The pie chart below shows the age distribution of the population in a city







Blood type
14.00%
16.00%
36.00%
17.00%
A
B
AB
O

Histograms

- When the variable we are studying is **quantitative**, we construct a **frequency distribution** represented by a histogram
- If there are many values, we group them into **5-8 bins (groups)**
- The **horizontal axis** (x-axis) represents the **variable of interest**, such as hemoglobin, age, etc.
- On the **vertical axis** (y-axis), we plot the **simple frequencies**, **relative frequencies**, or **percentage frequencies**

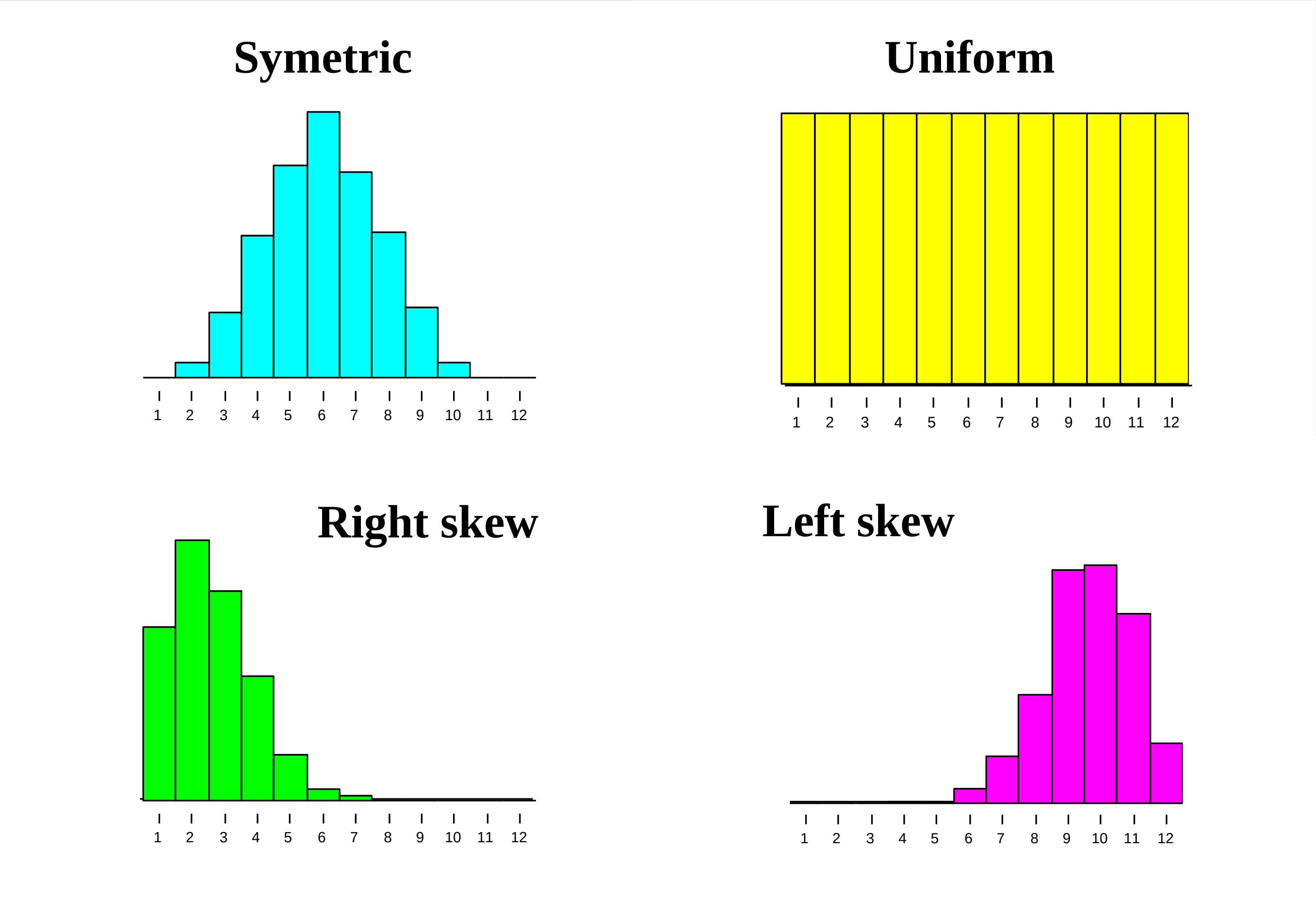
Histograms

- It is the most **useful** graphical representation of **quantitative data**
- A histogram shows the **shape** of the data distribution
- Each bar (rectangle) represents a group (bin) of data values
- The **height** is determined by the frequency, relative frequency, or percent frequency of the observations in that bin
- Unlike bar charts, there are **no gaps** between the bars in a histogram

Skew

A histogram indicates the **skewness** or **symmetry** of the data distribution

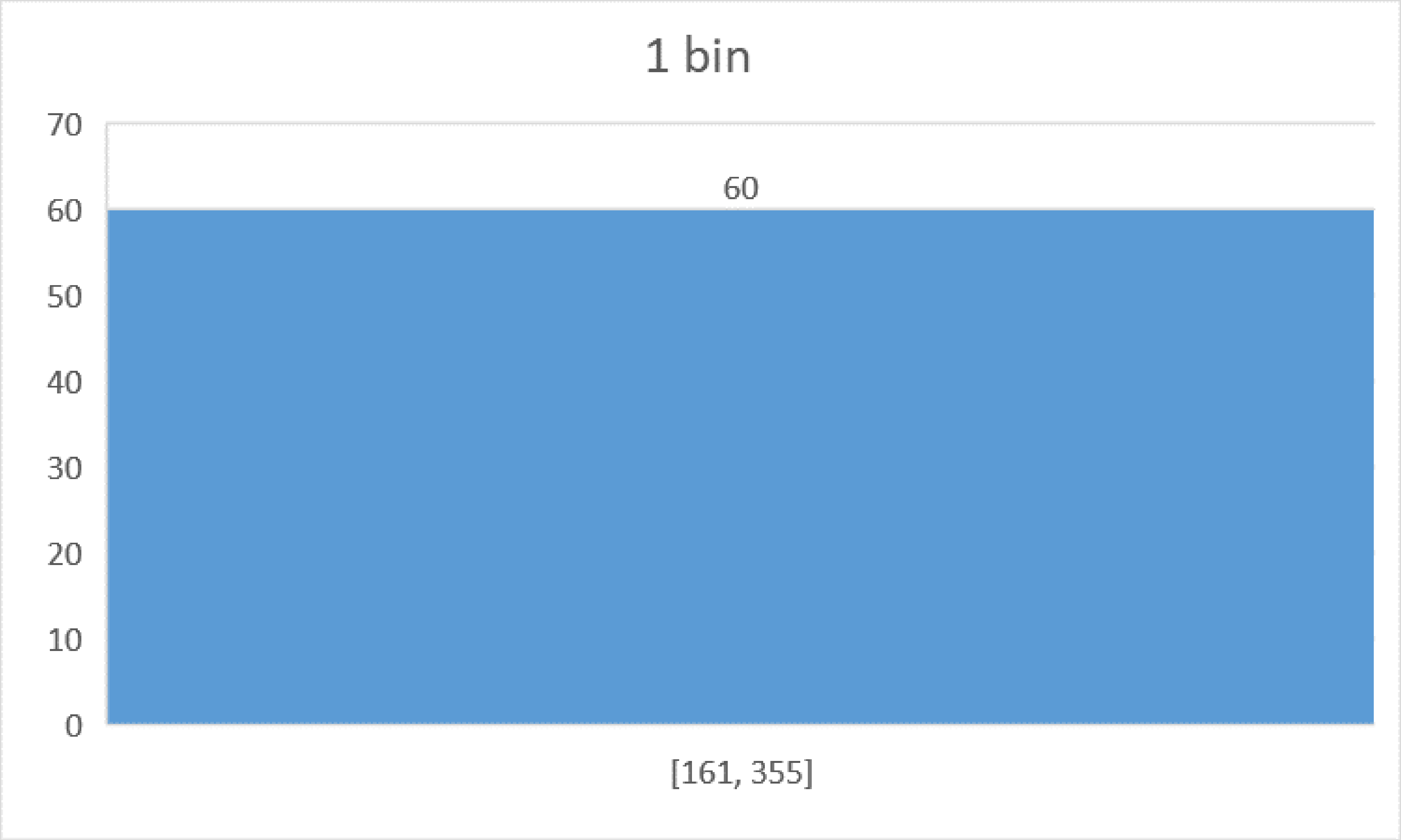
Example: The distribution of systolic blood pressure values among elderly people is positively skewed (right skewed), while the distribution of hemoglobin (Hgb) levels among 20 women is symmetric



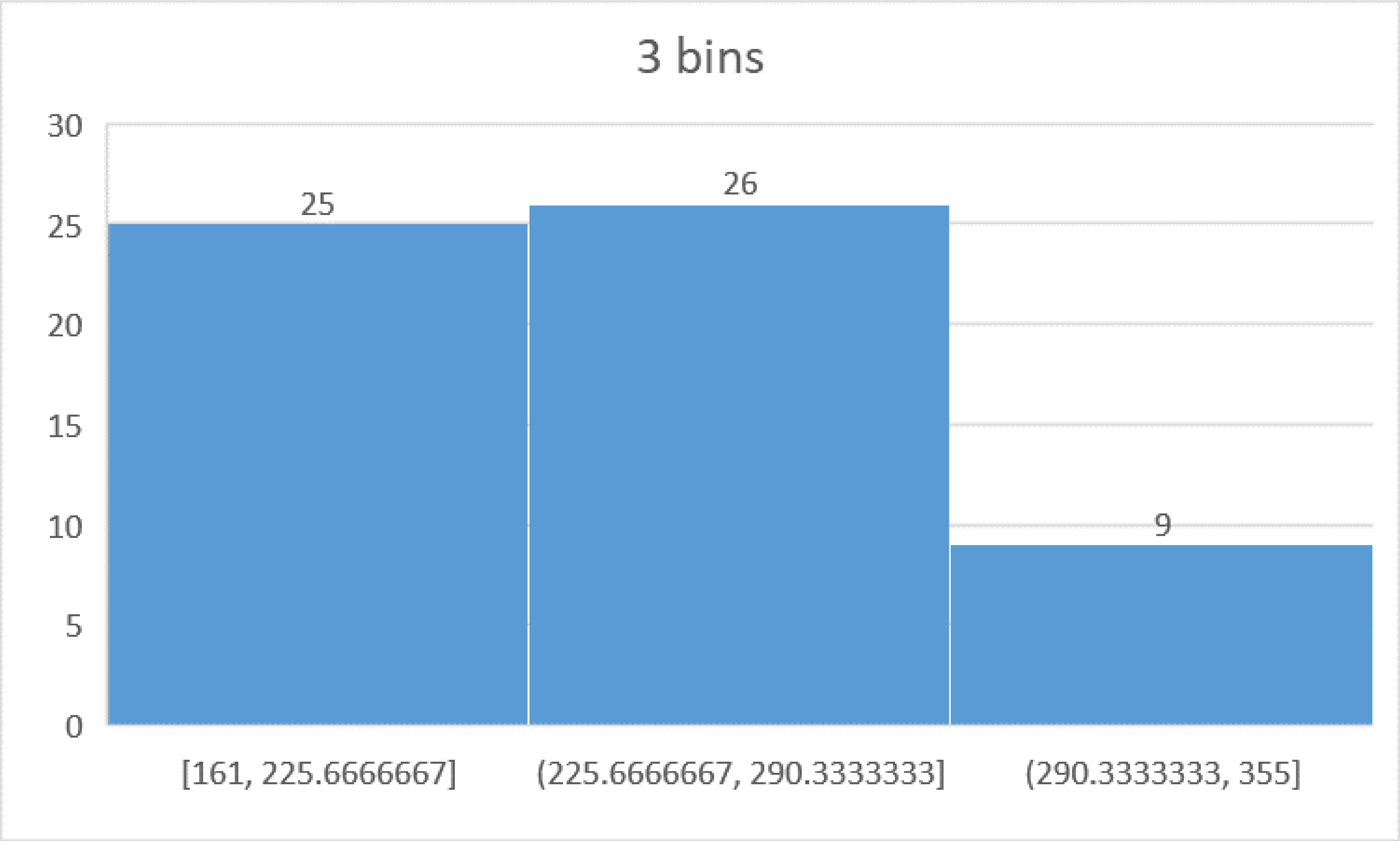
How to create a histogram

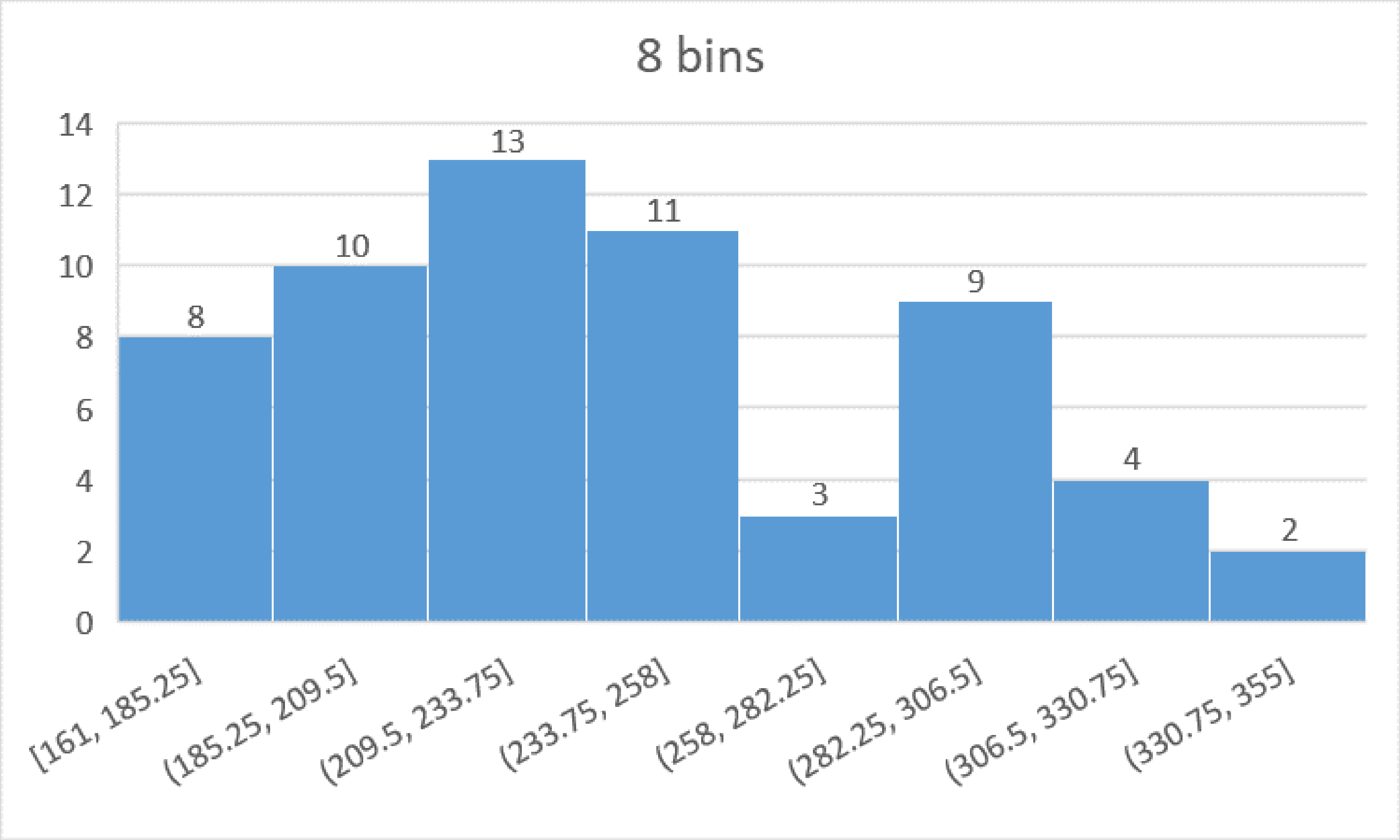
Example: The cholesterol levels of 60 subjects in a clinical trial were measured

212	249	227	218	310	281	330	226
233	223	161	195	233	249	284	284
174	170	256	169	299	210	301	199
258	258	195	227	244	355	234	195
196	354	282	282	286	286	176	195
163	297	211	228	309	309	225	223
195	248	284	173	256	169	209	209
200	258	284	239				



1 bin
70
60
50
40
30
20
10
0
60
[161, 355]


3 bins
30
25
20
15
10
5
0
25
26
9
[161, 225.6666667]
(225.6666667, 290.3333333]
(290.3333333, 355]



8 bins
14
12
10
8
6
4
2
0
8
10
13
11
3
9
4
2
[161, 185.25]
(185.25, 209.5]
(209.5, 233.75]
(233.75, 258]
(258, 282.25]
(282.25, 306.5]
(306.5, 330.75]
(330.75, 355]

Grouping values into eight class intervals (classes or bins)

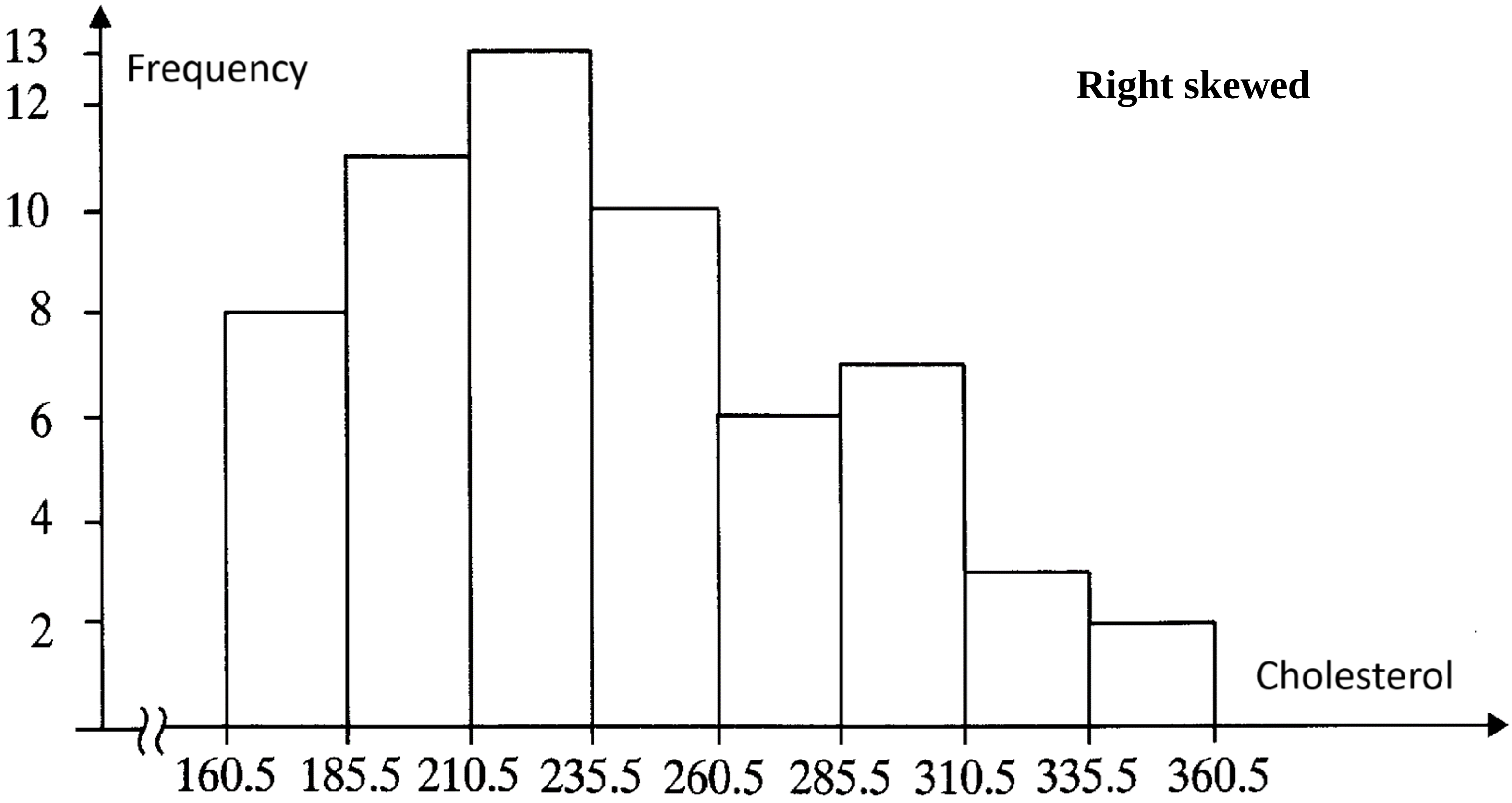
- First, we select the number of class intervals, $\boldsymbol{k = 8}$
- Next, we calculate the width of the class intervals

$$\boldsymbol{c = \frac{R}{k} = \frac{(355-161)}{8} = \frac{194}{8} = 24.25 \cong 25.}$$

- Finally, we determine the class intervals

Minimum x_{min} = 161
Maximum x_{max}= 355

Class intervals	Central value	Class frequency	Relative frequency (%)
[160.5—185.5)	173	8	13.33
[185.5—210.5)	198	11	18.33
[210.5—235.5)	223	13	21.67
[235.5—260.5)	248	10	16.67
[260.5—285.5)	273	6	10.00
[285.5—310.5)	298	7	11.67
[310.5—335.5)	323	3	5.00
[335.5—360.5)	348	2	3.33
Total		**60**	**100**

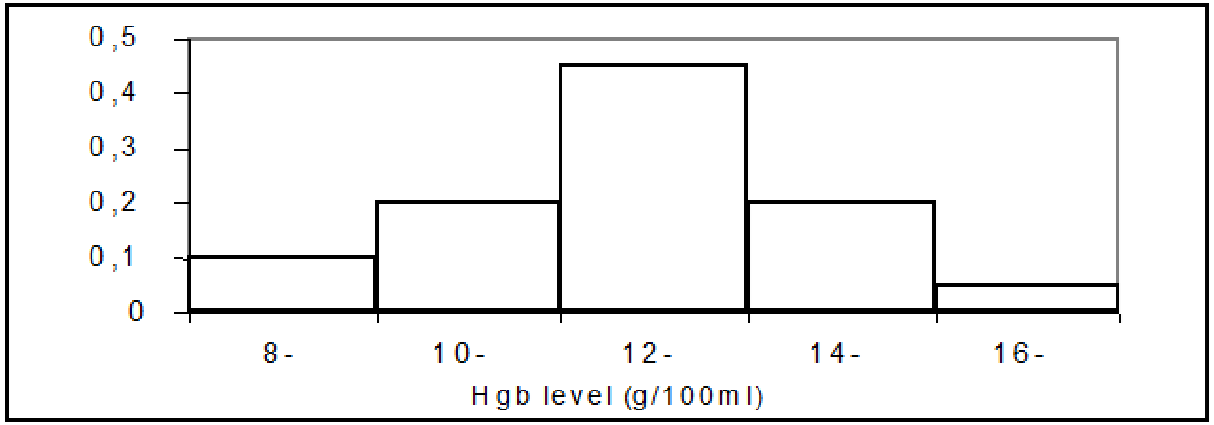

13
12
10
8
6
4
2
Frequency
Right skewed
Cholesterol
160.5
185.5
210.5
235.5
260.5
285.5
310.5
335.5
360.5

Another histogram example

Example: The hemoglobin levels (g/100 ml) of 20 women were measured and are as follows:

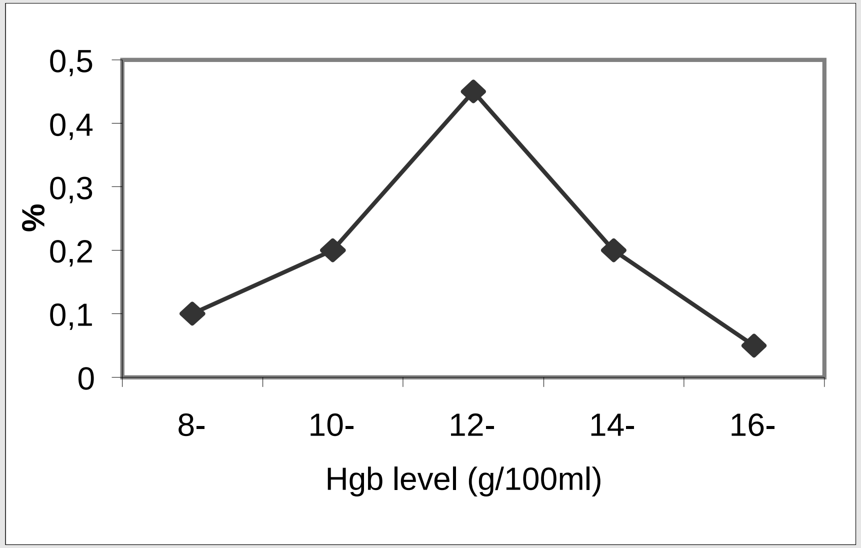
Hgb levels	
8,8	12,9
9,3	12,9
10,5	12,9
10,6	13,3
11,1	13,4
11,4	14,5
12	14,6
12	14,6
12,1	15,1
12,1	16,1

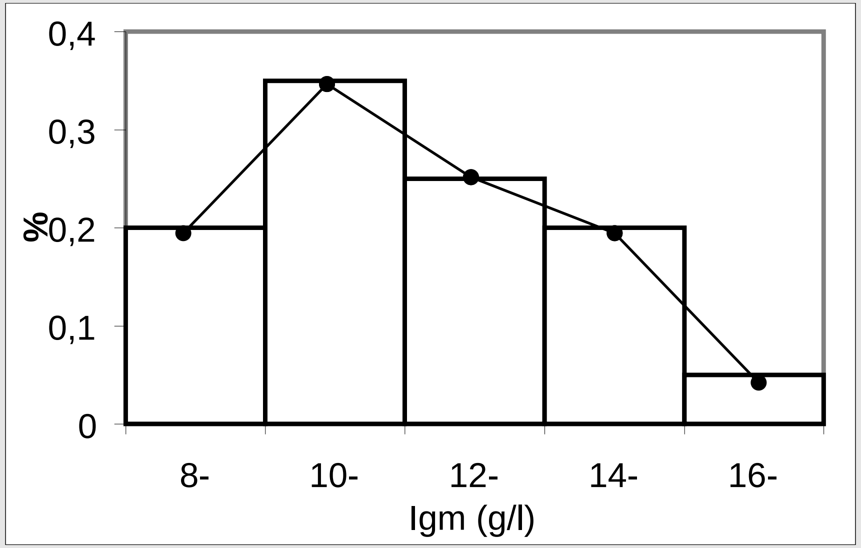
Hgb	Frequency	Proportion
8 -	2	0,1
10 -	4	0,2
12 -	9	0,45
14 -	4	0,2
16 -	1	0,05
Total	20	



Frequency curve

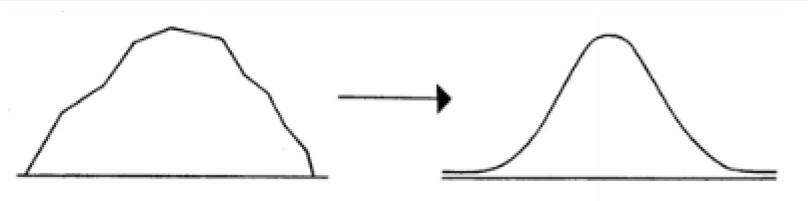
The **histogram**, for practical purposes, can be represented by a curve constructed by joining the **midpoints** of the tops of the rectangles (bars) in the histogram, forming what is called a **frequency curve**





Normal distribution

By increasing the sample size and constructing the histogram with smaller and smaller class widths, the corresponding polygon approaches a smooth curve



- The **normal curve** is **bell-shaped**, **symmetrical**, and its tails approach the horizontal axis smoothly. The mean and median are the same
- The area with the highest density is in the middle of the distribution. In other words, when the values of a variable are **normally distributed**, there are many values around the mean, while there are relatively few values far from the mean

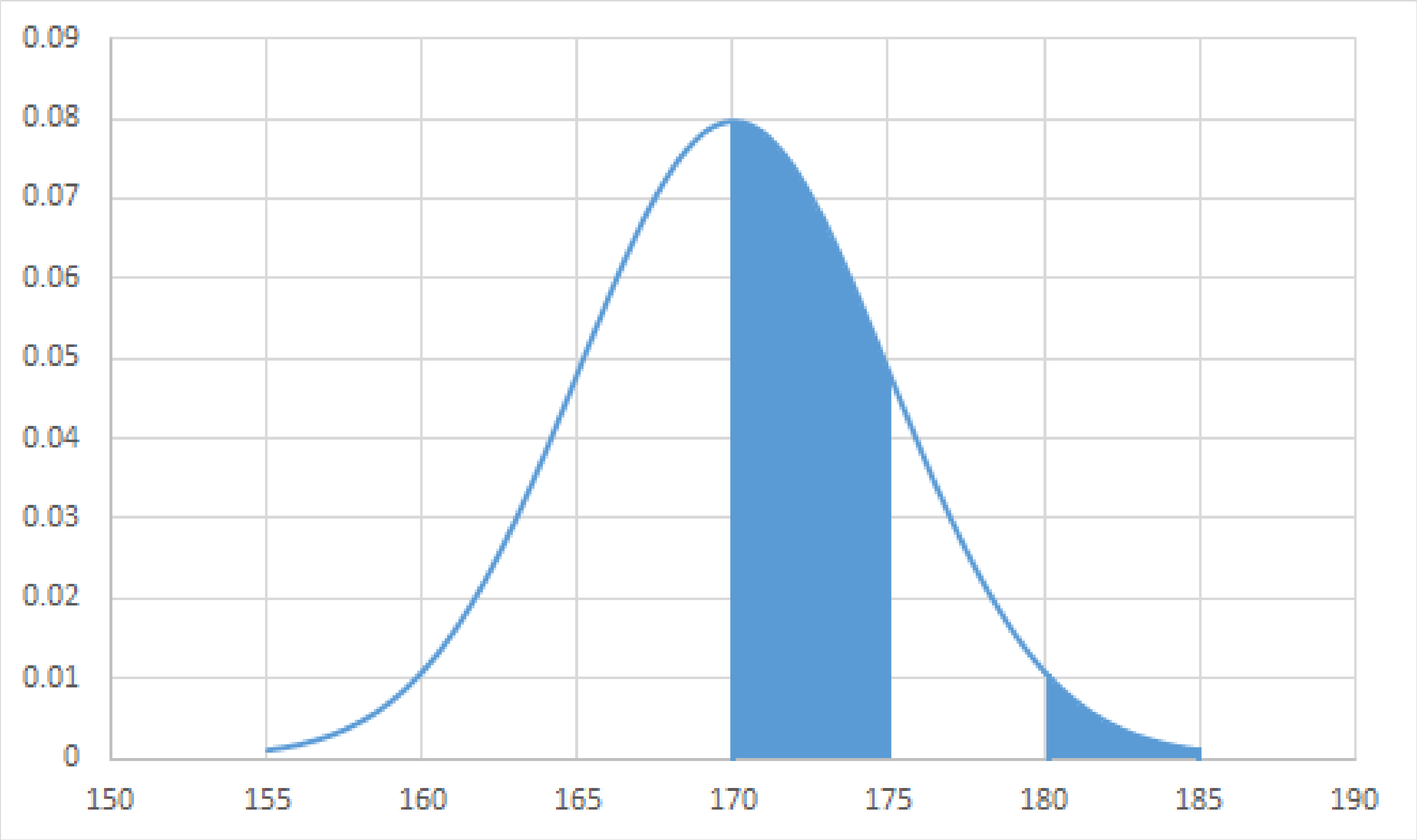
Why do normal distributions matter?

- All kinds of variables in **natural and social sciences** are **normally or approximately normally distributed**. Some examples of these variables are height, birth weight, and work satisfaction.
- Because normally distributed variables are so common, many **statistical tests** are designed for **normally distributed populations**
- Understanding the properties of normal distributions means you can use **inferential statistics** to compare different groups and make estimates about populations using samples

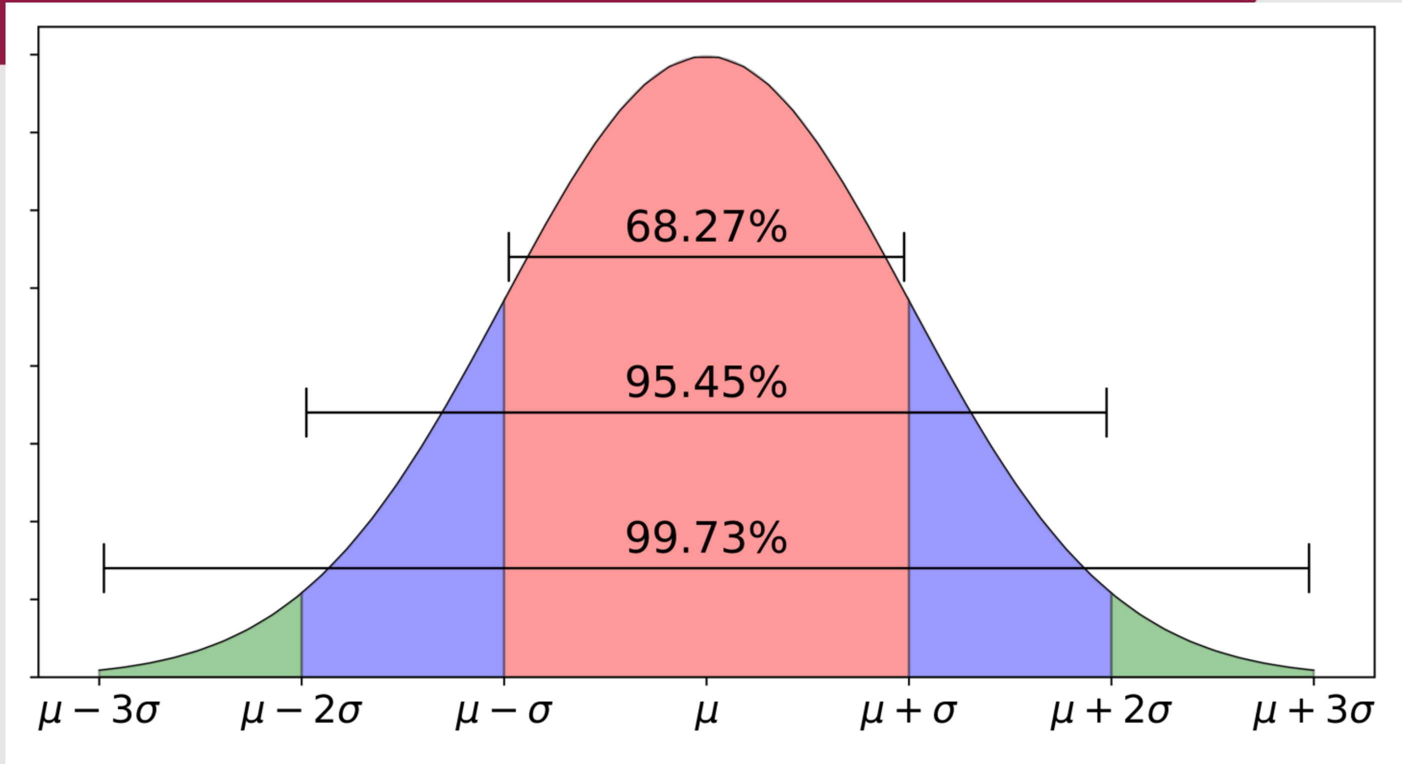
Normal distribution

Example:

- **Height of Greek people**, aged 18 to 25 years
- **Normally distributed**
- **Average height**: 170 cm
- **Standard deviation**: 5 cm
- Given this distribution, there are more people with heights between 170 cm and 175 cm than between 180 cm and 185 cm
- Additionally, very few people are taller than 185 cm or shorter than 155 cm



Normal distribution



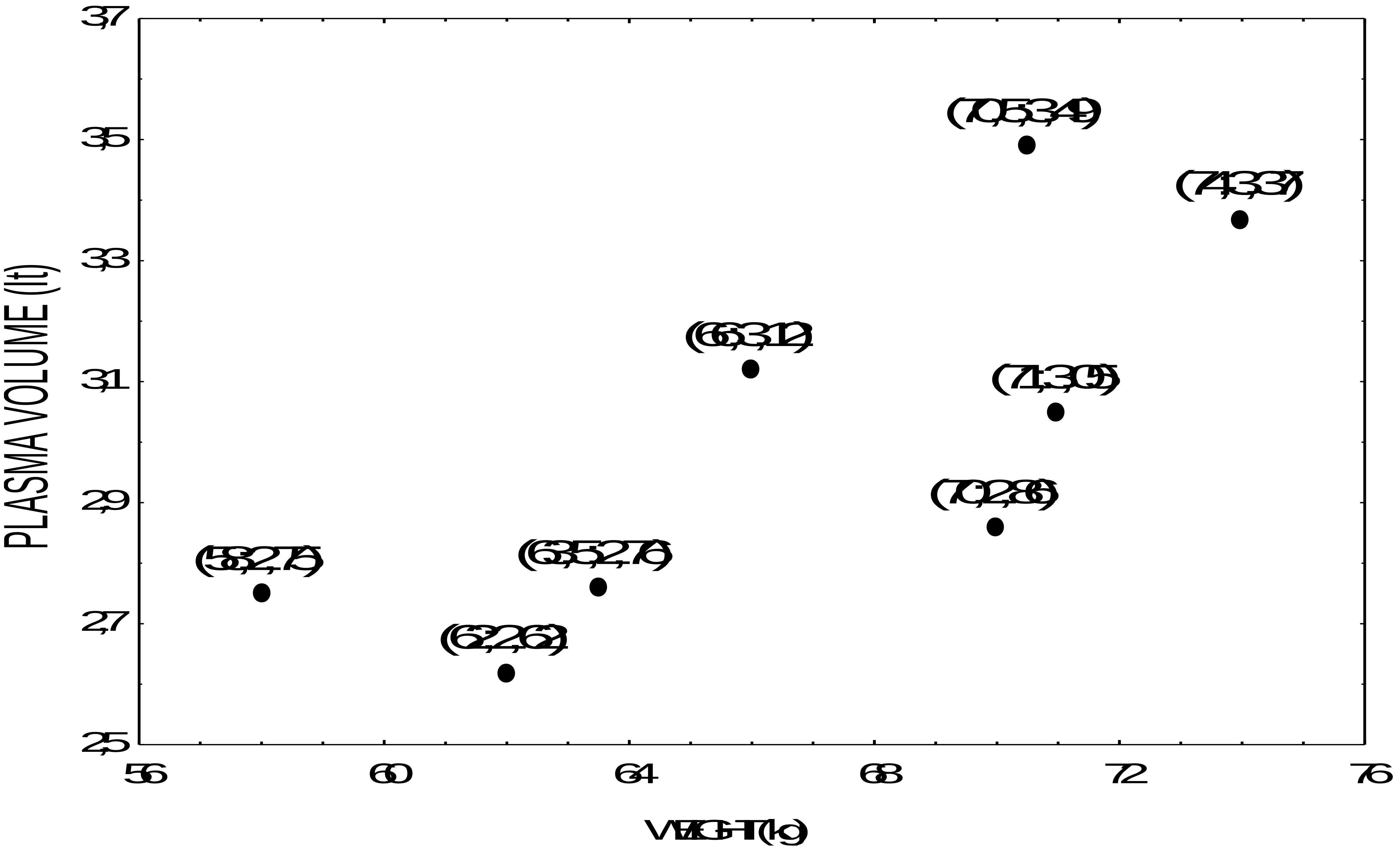
- Around 68% of values are within 1σ standard deviation from the mean μ ($\mu - \sigma, \mu + \sigma$).
- Around 95% of values are within 2σ standard deviations from the mean μ ($\mu - 2\sigma, \mu + 2\sigma$).
- Around 99.7% of values are within 3σ standard deviations from the mean μ ($\mu - 3\sigma, \mu - 3\sigma$).

Scatter plot

When there are observations from **two quantitative variables** and we are interested in the **relationship** between them, the data is presented using a scatter plot

Example: The body weight and plasma volume of 8 healthy men are:

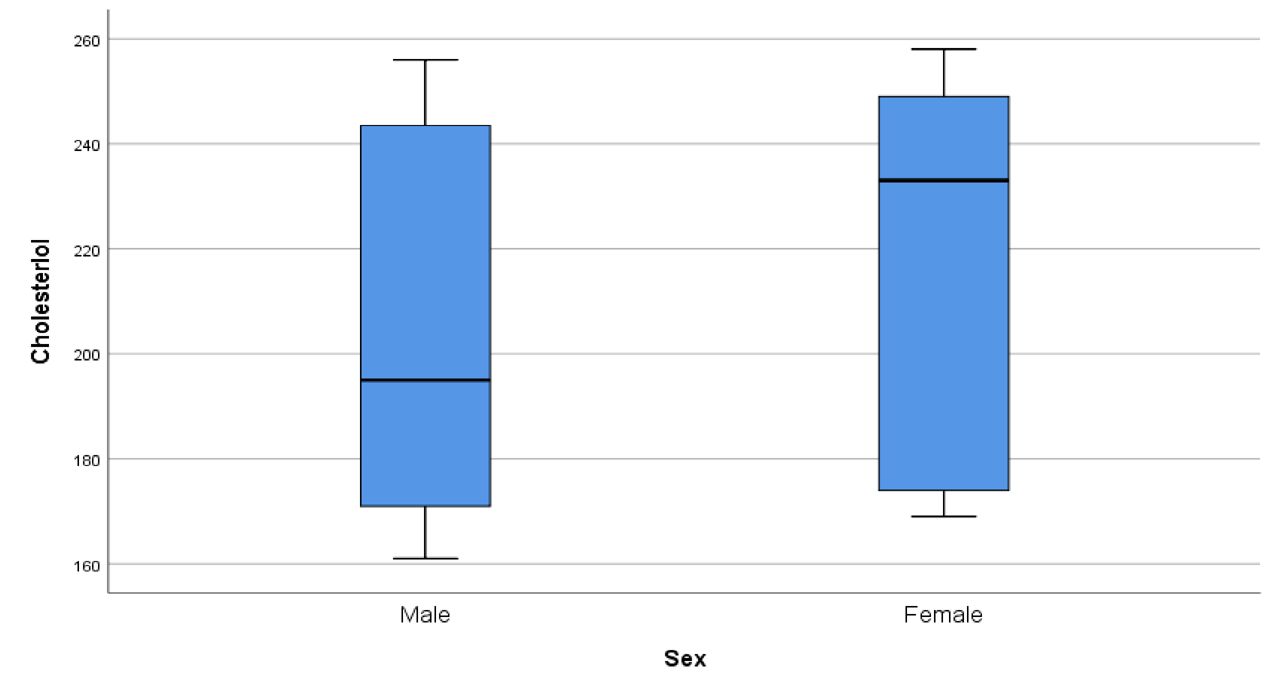
ID	Weight in Kg (x)	Plasma volume in lt (y)
1	58	2.75
2	70	2.86
3	74	3.37
4	63.5	2.76
5	62	2.62
6	70.5	3.49
7	71	3.05
8	66	3.12



3,7
3,5
3,3
3,1
2,9
2,7
2,5
PLASMA VOLUME (lt)
56
60
64
68
72
76

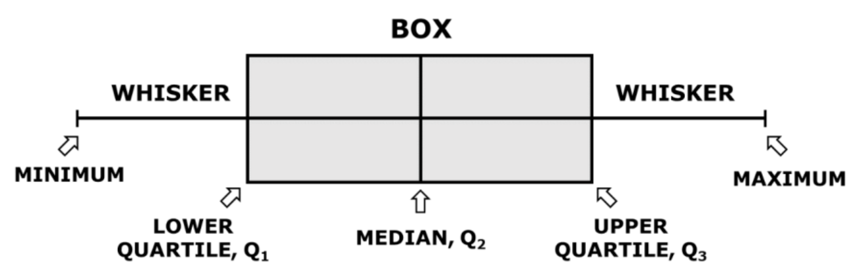
Box plot

- A **box plot** is an easy way to graphically display the **shape** of the data
- **Easy to interpret**
- It shows if the data is **skewed**
- It helps to find **outliers**
- Box plots are useful for **comparing different groups**

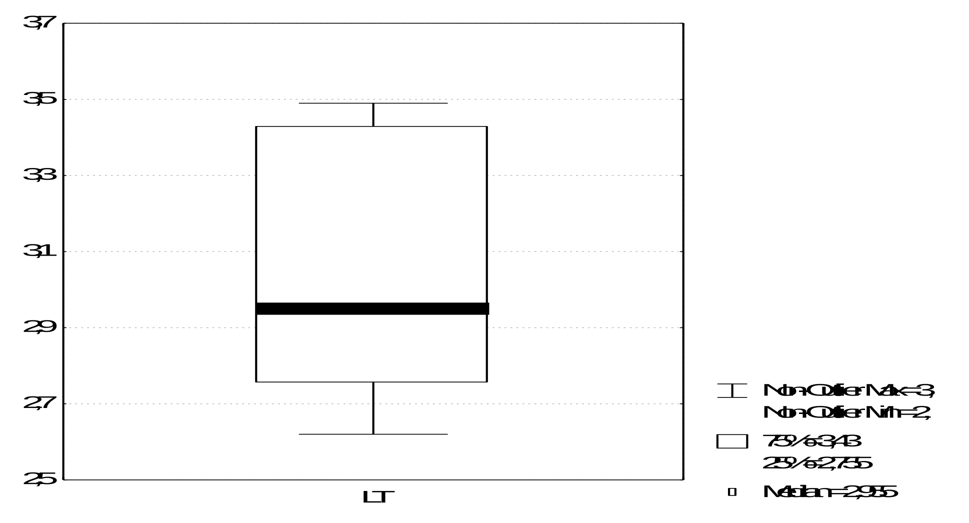


Box plot

- A box plot displays data with a **rectangular box** and **whiskers** extending from it
- The top of the box represents the **75th percentile** (third quartile), and the bottom represents the **25th percentile** (first quartile)
- The **median** is shown by a horizontal line within the box
- The whiskers extend to the maximum and minimum values

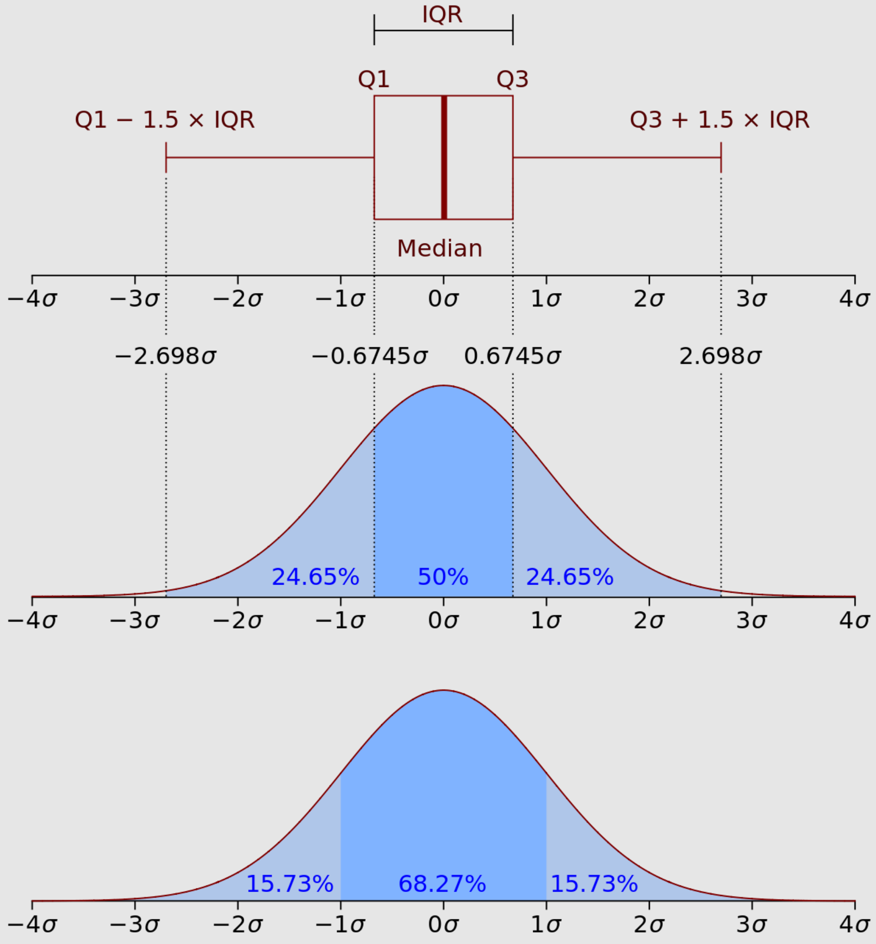


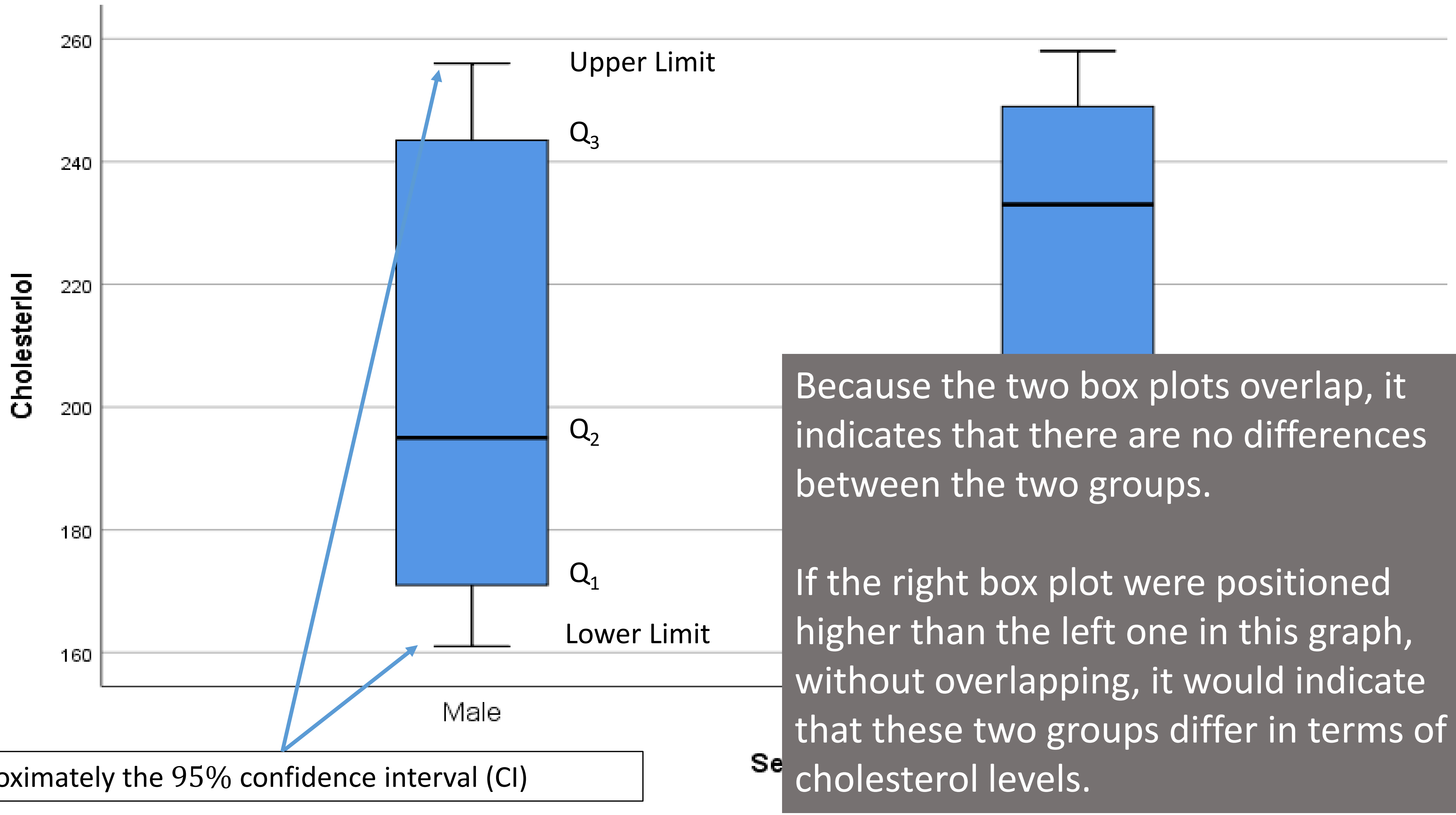
Example: Using data from plasma volumes, the following box plot is produced. Note that the value 7.32 is considered an outlier and is therefore excluded.



Box plot versus histogram

Box plot (with Interquartile Range - IQR) and Probability Density Function (PDF) for a normal distribution N(0, σ²)





Because the two box plots overlap, it indicates that there are no differences between the two groups.

If the right box plot were positioned higher than the left one in this graph, without overlapping, it would indicate that these two groups differ in terms of cholesterol levels.

Quantitative Methods of Data Description

Numerical descriptive measures

- Measures of central tendency

  Median

  Mean

  Quartiles

  Percentiles

- Measures of variation

  Range

  Variance

  Standard deviation

Mean

The simplest way to describe a set of observations from a **continuous variable** is the **mean**, which is the **sum** of all observations **divided** by the number of observations

Example*:* The plasma volumes of 8 healthy men are:

2.75 lt	2.86 lt	3.37 lt	2.76 lt	2.62 lt	3.49 lt	3.05 lt	3.12 lt

x_1=2.75, x_2=2.86, x_3=3.37, x_4=2.76, x_5=2.62, x_6=3.49, x_7=3.05, x_8=3.12

Mean

The sum of the values is:

$\sum x$ = $x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8$ = 2.75 + 2.86 + 3.37 + 2.76 + 2.62 + 3.49 + 3.05 + 3.12 = 24.02

The number of observations is n = 8

Therefore, the mean is calculated as follows:

$$\bar{x} = \frac{\sum x}{n} = \frac{(2.75 + 2.86 + 3.37 + 2.76 + 2.62 + 3.49 + 3.05 + 3.12)}{8} = \frac{24.02}{8} = 3.0025$$

Median – Percentiles - Quartiles

- When there are **one or more extremely small or large observations**, the mean is not the best way to describe the data
- In such cases, the observations are best described by the **median** or **50th percentile**
- To find the **median**, the observations are **sorted** in numerical order (smallest to largest)
- If the number of observations is **odd**, the median is the **middle observation**
- If the number of observations is **even**, the **median** is the **average** of **the two middle values**

Median (odd number of observations)

Example*:* The maximal inspiratory pressure, in cm$H_2$0, (PImax) of 9 cystic fibrosis patients is:

1	2	3	4	5	6	7	8	9
80	85	110	95	95	100	45	95	130

The numbers are ordered from smallest to largest:

1	2	3	4	5	6	7	8	9
45	80	85	95	95	95	100	110	130

Then, the median is the middle value which is the 5th value $\left(\text{since } \frac{9}{2} = 4.5 \cong 5\right)$. Therefore, the median is 95

1	**2**	**3**	**4**	**5**	**6**	7	8	9
45	80	85	95	95	95	100	110	130

Median (even number of observations)

Example*:* The plasma volumes of 8 healthy men are :

1	2	3	4	5	6	7	8
2.75	2.86	3.37	2.76	2.62	3.49	7.32	3.05

The numbers are ordered from smallest to largest:

1	2	3	4	5	6	7	8
2.62	2.75	2.76	2.86	3.05	3.37	3.49	7.32

Then, the median is the average of the 4th and 5th value, which is

median $= \ (2.86 + 3.05)/2 = 2.96$

Percentiles - Quartiles

25^{o} (Q_1) 50^{o} (Q_2) 75^{o} (Q_3)

1	2	3	4	5	6	7	8
2.62	2.75	2.76	2.86	3.05	3.37	3.49	7.32

$$L = \frac{p}{100}(n+1)$$

Where:

L is the position in the sorted data
p is the percentile we are looking for
n is the number of observations

Percentiles - Quartiles

$$L_{25} = \frac{25}{100}(8+1) = 2.25 \qquad L_{50} = \frac{50}{100}(8+1) = 4.5 \qquad L_{75} = \frac{75}{100}(8+1) = 6.75$$

25º (Q_1) 50º (Q_2) 75º (Q_3)

1	2	3	4	5	6	7	8
2.62	2.75	2.76	2.86	3.05	3.37	3.49	7.32

$$L_{50} = 4.5$$

We see that the median is halfway (0.5) between the 4th and 5th observations, whose values are 2.86 and 3.05, respectively, so:

$$Q_2 = 2.86 + 0.5(3.05 - 2.86) = 2.955 = 2.96$$

Percentiles - Quartiles

	25^o (Q_1)			50^o (Q_2)		75^o (Q_3)	
1	2	3	4	5	6	7	8
2.62	2.75	2.76	2.86	3.05	3.37	3.49	7.32

L_{25}=2.25

We see that first percentile (Q_1) is a quarter (0.25) of the way between the 2nd and 3rd observation, whose values are 2.75 and 2.75, respectively, so:

$$Q_1 = 2.75 + 0.25(2.76 - 2.75) = 2.7525 = 2.75$$

Percentiles - Quartiles

	25^o (Q_1)		50^o (Q_2)		75^o (Q_3)		
1	2	3	4	5	6	7	8
2.62	2.75	2.76	2.86	3.05	3.37	3.49	7.32

L_{75}=6.75

We see that the third percentile (Q_3) is three quarters (0.75) of the way between the 6^{th} and 7^{th} observation, whose values are 3.37 and 3.49, respectively, so:

$$Q_3 = 3.37 + 0.75(3.49 - 3.37) = 3.46$$

Percentiles - Quartiles

	25° (Q_1)		50° (Q_2)		75° (Q_3)		
1	2	3	4	5	6	7	8
2.62	2.75	2.76	2.86	3.05	3.37	3.49	7.32

Q_1 = 2.75, Q_2 = 2.96, Q_3 = 3.46

Therefore:

Approximately 25% of 8 healthy adult men have a plasma volume of 2.75 or less.
Approximately 50% of 8 healthy adult men have a plasma volume of 2.96 or less.
Approximately 75% of 8 healthy adult men have a plasma volume of 3.46 or less.

We use the term 'approximate' because these values are not directly within the data.

Measures of variation

- But we also need a measure of data variation
- The **mean value alone** does not allow us to **differentiate** between samples

Range

- The range is the difference between the largest and smallest observation
- However, it does not show how the remaining observations are distributed between these two

Example*:* The plasma volumes of 8 healthy men are:

2.75, 2.86, 3.37, 2.76, 2.62, 3.49, 3.05, 3.12 lt

Range = max-min = 3.49 - 2.62 = 0.87

The range generally gives us a good indicator of variability when you have a distribution without extreme values

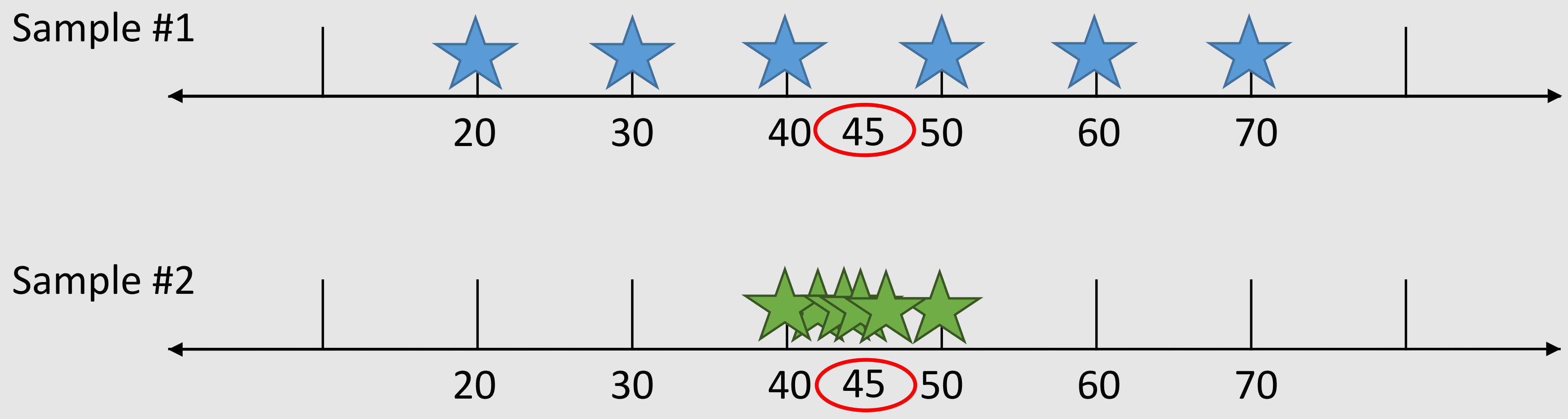
Variance (σ^2 or s^2) and Standard Deviation

	Sample #1	Sample #2
1	20	40
2	30	43
3	40	44
4	50	46
5	60	47
6	70	50
	Mean #1	Mean #2
	$\bar{x} = 45$	$\bar{x} = 45$

Samples with the same mean

Different spread of data

Variance (σ^2 or s^2) and Standard Deviation



Question: What do these two plots tell us about the variance of the data?

Answer: While both have the same mean, Sample #1 shows greater variability

Variance (σ^2 or s^2) and Standard Deviation

- **How far is each point from the mean? (DISTANCE)**
  This is the question that variance and standard deviation help us answer
- The **standard deviation** is simply the **square root** of the variance, making it easy to calculate
- If some points are close to the mean, the variance and standard deviation will be smaller than for points that are further away from the mean
- The mean, variance, and standard deviation are very important when comparing data sets (t-test, anova) or when comparing a data set with a theoretical value
- However, since variance is the square of the differences from the mean, it can be less intuitive. Therefore, we often use the standard deviation to express variance
- Both measures reflect variability in a distribution, but their units differ: **Standard deviation** is expressed in the same units as the **original values** (e.g., minutes or meters). **Variance** is expressed in **much larger units** (e.g., meters squared)
- Symbols
  - Mean $\bar{x}$ (x-bar)
  - Variance σ^2 ή s^2
  - Standard Deviation σ ή s

Note: We are using the sample variance, not the population variance

Variance (σ^2 or s^2) and Standard Deviation

Variance

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

Standard Deviation

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

Plasma Volume: Variance and Standard Deviation (σ^2 και σ)

	Plasma volume x	Mean $\bar{x}$	Plasma volume – Mean $x - \bar{x}$	$(x - \bar{x})^2$
1	2.75	3.0025	-0.2525	0.063756
2	2.86	3.0025	-0.1425	0.020306
3	3.37	3.0025	0.3675	0.135056
4	2.76	3.0025	-0.2425	0.058806
5	2.62	3.0025	-0.3825	0.146306
6	3.49	3.0025	0.4875	0.237656
7	3.05	3.0025	0.0475	0.002256
8	3.12	3.0025	0.1175	0.013806
			$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$	$s^2 = (0.063756 + 0.020306 + 0.135056 + 0.058806 + 0.146306 + 0.237656 + 0.002256 + 0.013806) / 7 = 0.09685$
				$s = \sqrt{0.09685} = 0.311207$

Sample #1: Variance and Standard Deviation (σ^2 και σ)

	x	$\bar{x}$	$x-\bar{x}$	$(x-\bar{x})^2$
1	20	45	-25	625
2	30	45	-15	225
3	40	45	-5	25
4	50	45	5	25
5	60	45	15	225
6	70	45	25	625

$$\sigma^2 = \frac{\sum(x-\bar{x})^2}{n-1}$$

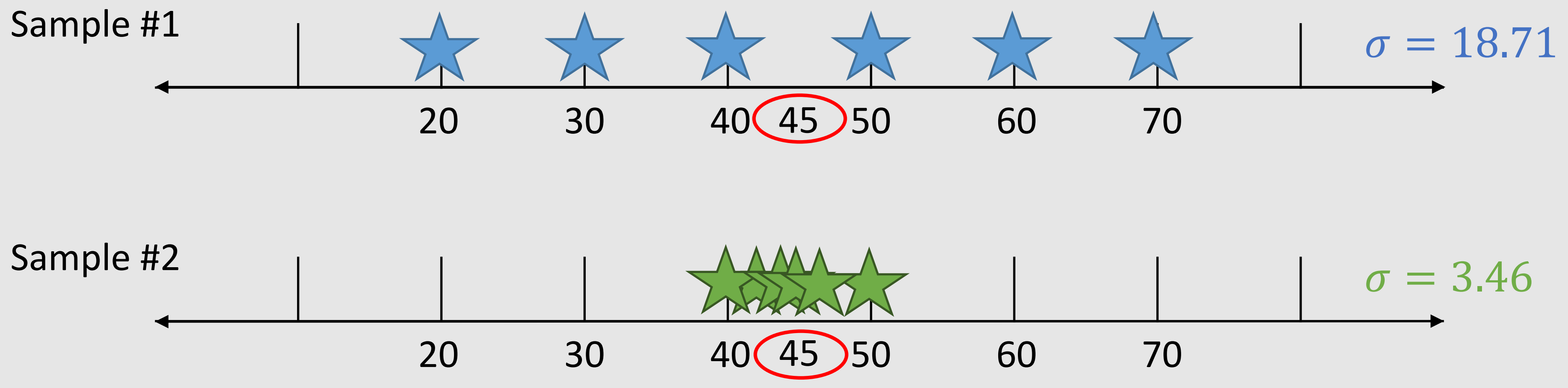
$$\sigma^2 = (625 + 225 + 25 + 25 + 225 + 625)\ /\ 5 = 350$$

$$\sigma = \sqrt{350} = 18.71$$

Sample #2: Variance and Standard Deviation (σ^2 και σ)

	x	$\bar{x}$	$x-\bar{x}$	$(x-\bar{x})^2$
1	40	45	-5	25
2	43	45	-2	4
3	44	45	-1	1
4	46	45	1	1
5	47	45	2	4
6	50	45	5	25
	$\sigma^2 = \dfrac{\sum(x-\bar{x})^2}{n-1}$			$\sigma^2 = (25+4+1+1+4+25) \,/\, 5 = 12$
				$\sigma = \sqrt{12} = 3.46$

Variance and Standard Deviation (σ^2 και σ)



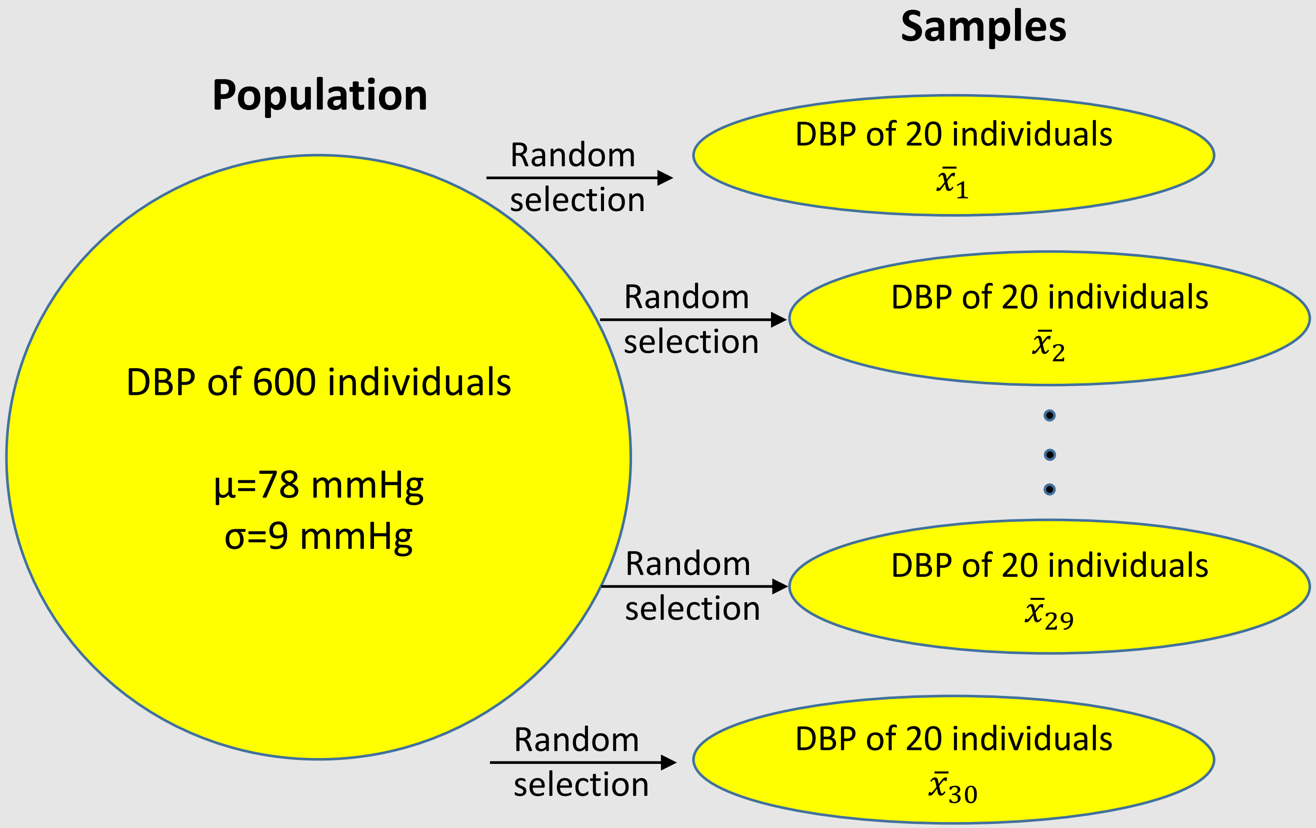
In general, a smaller standard deviation is usually better because it means the data points are closer to the mean, showing more consistency and reliability

Standard Error

- We draw conclusions about a population by collecting a **representative sample**
- Therefore, the mean ($\bar{x}$) and standard deviation (s) of a **sample** are used to **estimate** the mean (μ) and standard deviation (σ) of the **population** from which the sample is drawn
- The mean value of a sample is **unlikely to be exactly the same** as that of the population
- A different sample would likely give a **different** mean, and this difference is due to **sampling variability**

Standard Error

- If we collect **several independent samples** of the same size and calculate the **mean** and **standard deviation of each**, then the mean of the sample means will **approximate** the **population mean**
- The **standard deviation** of the **sample means** is equal to $\frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}}$, where σ is the population standard deviation and n is the sample size



$$Average\ of\ sample\ means\ \mu = \frac{\bar{x}_1 + \bar{x}_2 + \cdots + \bar{x}_{29} + \bar{x}_{30}}{30}$$

$$Standard\ deviation\ of\ sample\ means\ SD = \sqrt{\frac{\sum(\bar{x}_i - \mu)^2}{n}} = \frac{\sigma}{\sqrt{\mathrm{n}}}$$

Standard Error

$$se = \frac{\sigma}{\sqrt{n}}$$

- The **quantity** $\boldsymbol{\frac{\sigma}{\sqrt{n}}}$ is called the **standard error** of the sample mean and measures how well the population mean is approximated by the sample mean
- The standard error (SE) is a function of the variance and the sample size
- A large sample with a small variance produces a small standard error
- Because we rarely know the population standard deviation σ, we use the sample standard deviation s instead
- Therefore, the standard error of the mean is estimated by the quantity $se = \frac{s}{\sqrt{n}}$

Standard Error

Example*:* The plasma volumes, in liters, of 8 healthy men are:

1	2	3	4	5	6	7	8
2.75	2.86	3.37	2.76	2.62	3.49	7.32	3.05

Mean, $\bar{\boldsymbol{x}}$	3.025
Standard Deviation, s	0.311
Standard Error of the mean, $se(\bar{x})$	$se(\bar{x}) = \frac{s}{\sqrt{n}} = \frac{0.311}{\sqrt{8}} = 0.111$

If the sample size approaches the population size, then the standard error (se) tends to zero