STATISTICAL LABORATORY

ISEG

Lisbon School of Economics & Management
Universidade de Lisboa

Applied Mathematics for Economics and Management
1st Year/1st Semester
2025/2026

CONTACT

Professor: Elisabete Fernandes
E-mail: efernandes@iseg.ulisboa.pt

https://doity.com.br/estatistica-aplicada-a-nutricao

https://basiccode.com.br/produto/informatica-basica/

PROGRAM

1. Fundamental Concepts of Statistics

2. Exploratory Data Analysis

3. Organizing and Summarizing Data

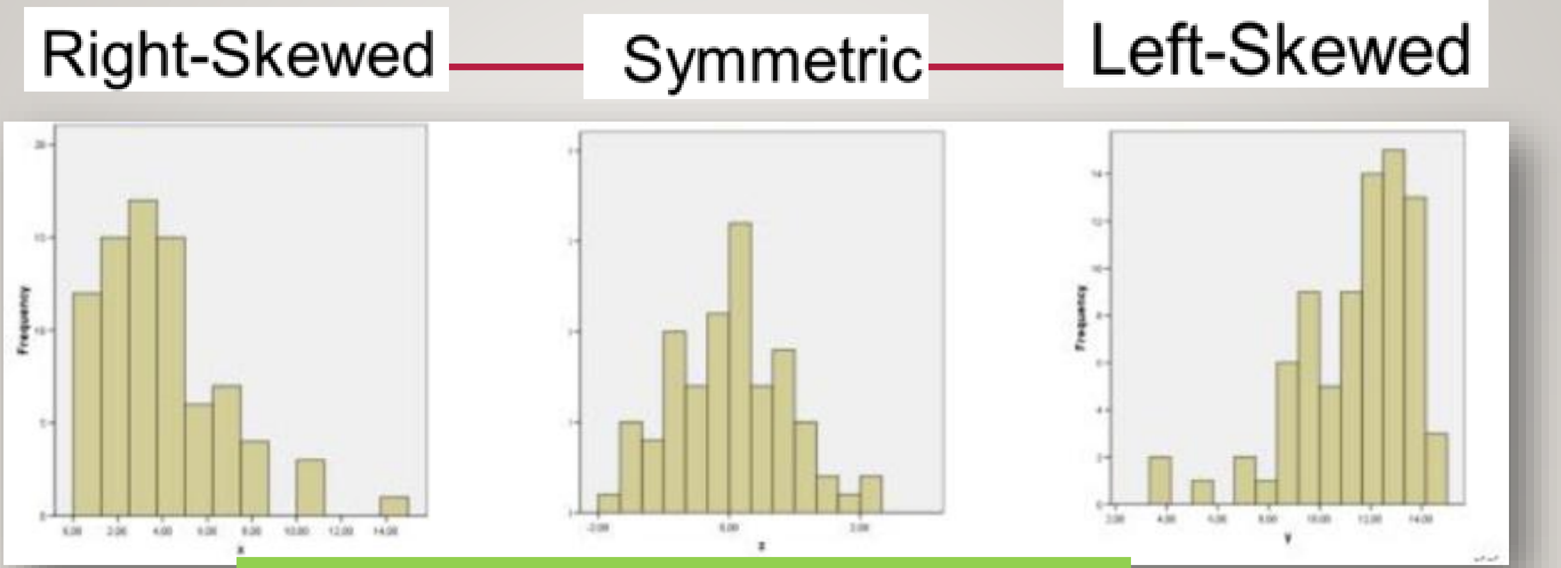
4.Association and Relationships Between Variables

5. Index Numbers

6.Time Series Analysis

LECTURE 3: EXPLORATORY DATA ANAYSIS (CONTINUED)

SKEWNESS IN HISTOGRAMS



Right-Skewed

Symmetric

Left-Skewed

Right-skewed (positively skewed) Distribution:

- Tail extends to the right (higher values).
- Most data concentrated on the left.
- Skewness > 0

Left-skewed (negatively skewed) Distribution:

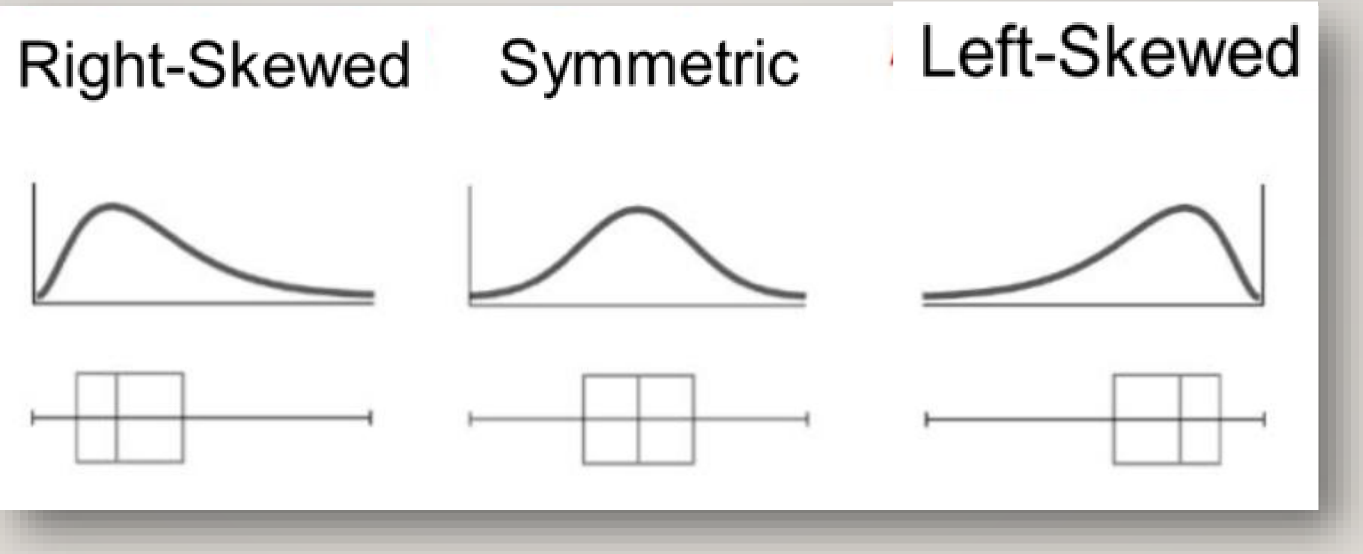
- Tail extends to the left (lower values).
- Most data concentrated on the right.
- Skewness < 0

Symmetric Distribution:

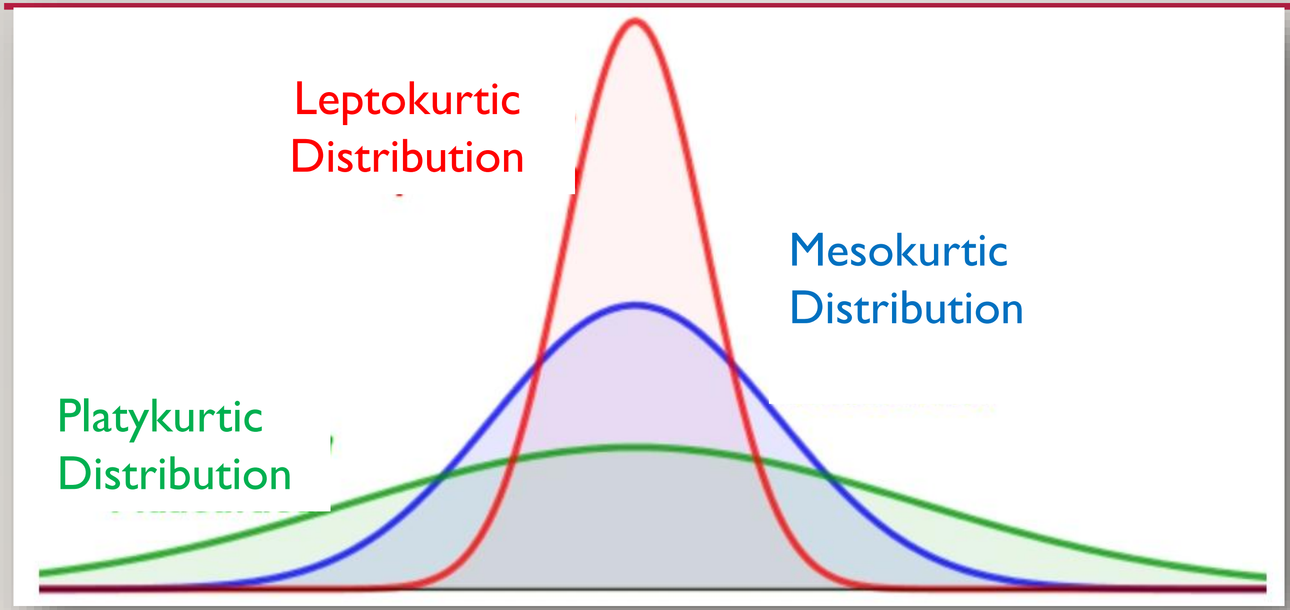
- Tails roughly equal on both sides.
- Data evenly distributed around the center.
- Skewness ≈ 0

SKEWNESS IN BOXPLOTS

The boxplot will be discussed in a subsequent slide.



KURTOSIS VISUALIZATION

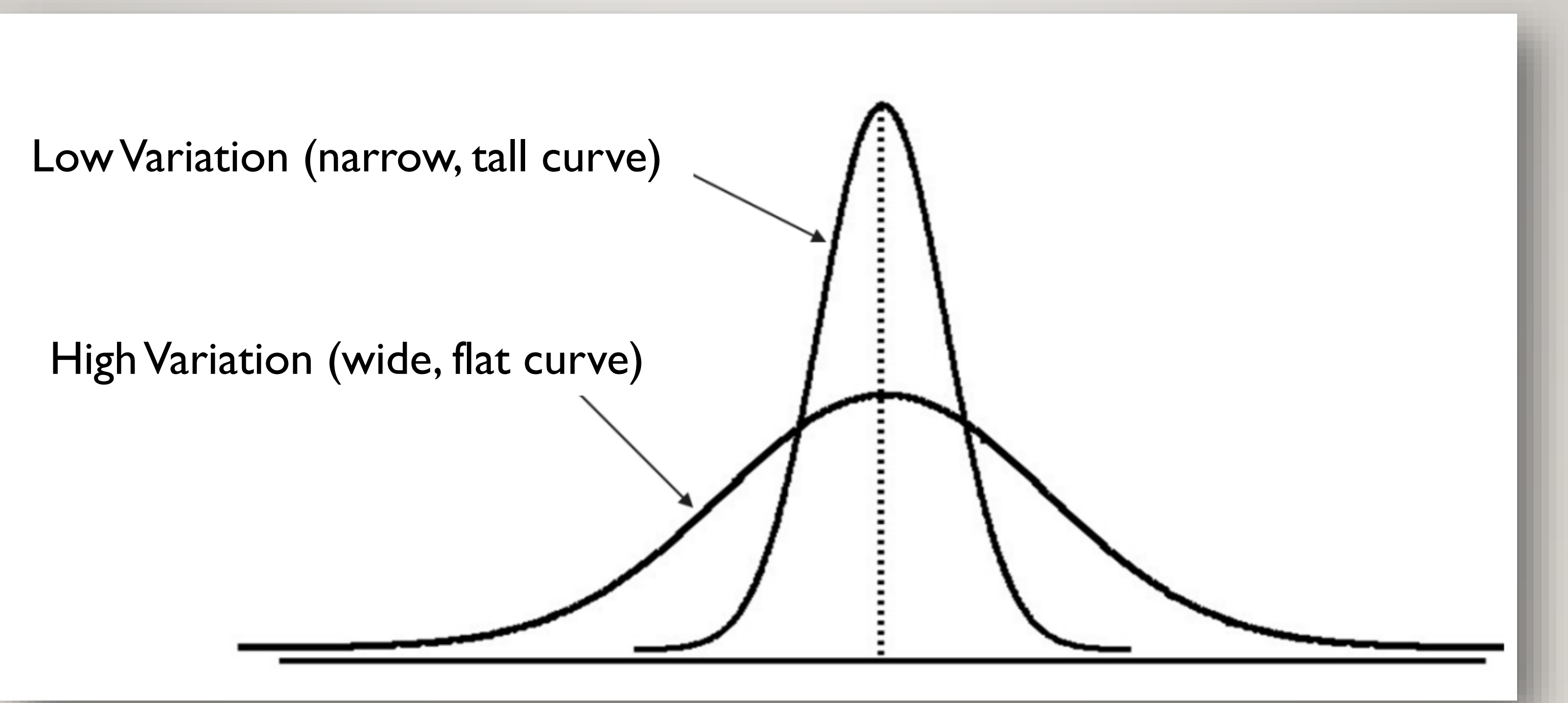


Leptokurtic Distribution (high peak, heavy tails):
•Most data concentrated near the mean.
•More extreme values (outliers) likely.

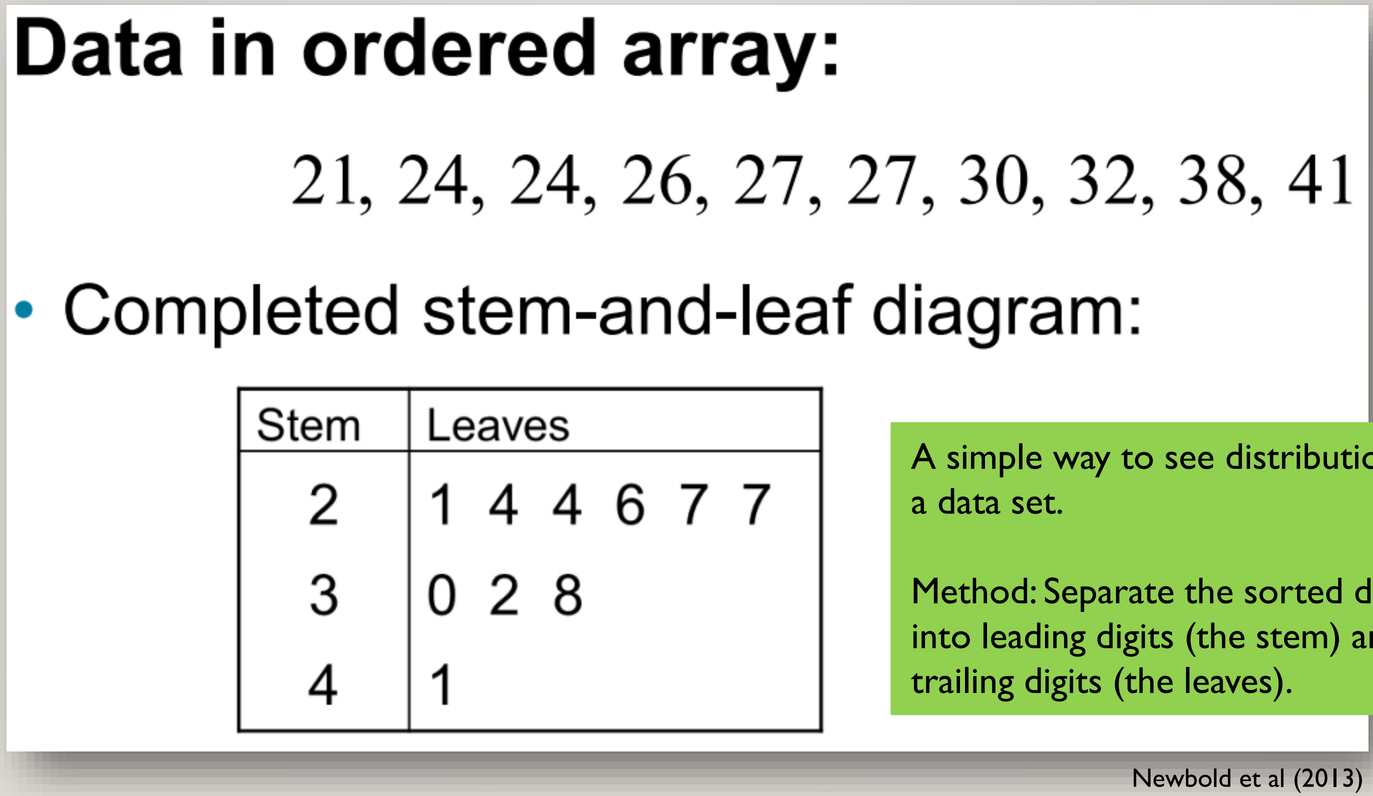
Platykurtic Distribution (low peak, light tails):
•Data more evenly spread.
•Fewer extreme values.

Mesokurtic Distribution (moderate peak, normal tails):
•Moderate concentration of data around the mean.

VARIATION VISUALIZATION


Low Variation (narrow, tall curve)
High Variation (wide, flat curve)

STEM-AND-LEAF DIAGRAM EXAMPLE



Data in ordered array:

21, 24, 24, 26, 27, 27, 30, 32, 38, 41

- Completed stem-and-leaf diagram:

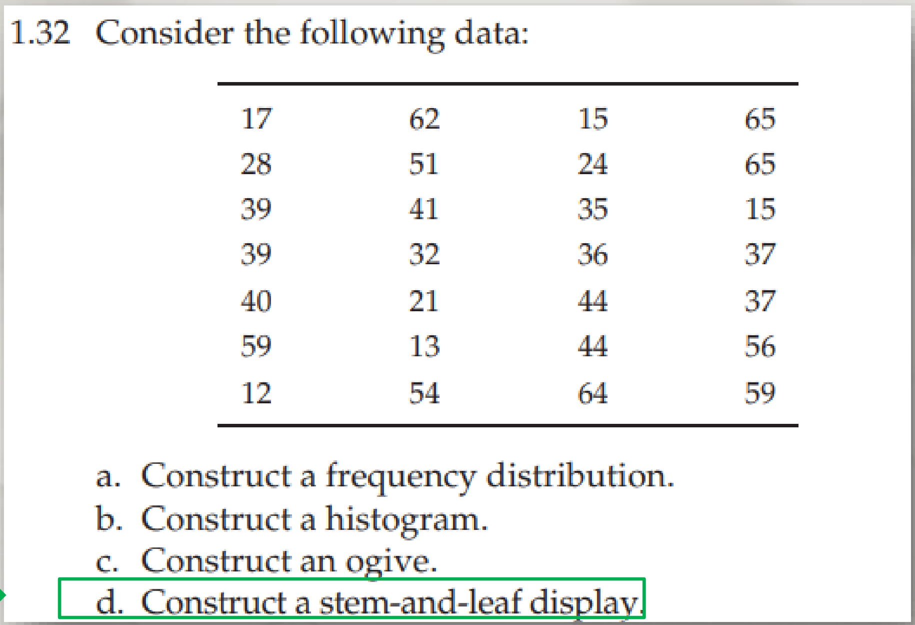
Stem	Leaves
2	1 4 4 6 7 7
3	0 2 8
4	1

A simple way to see distribution details in a data set.

Method: Separate the sorted data series into leading digits (the stem) and the trailing digits (the leaves).

Newbold et al (2013)

EXERCISE 1.32



1.32 Consider the following data:

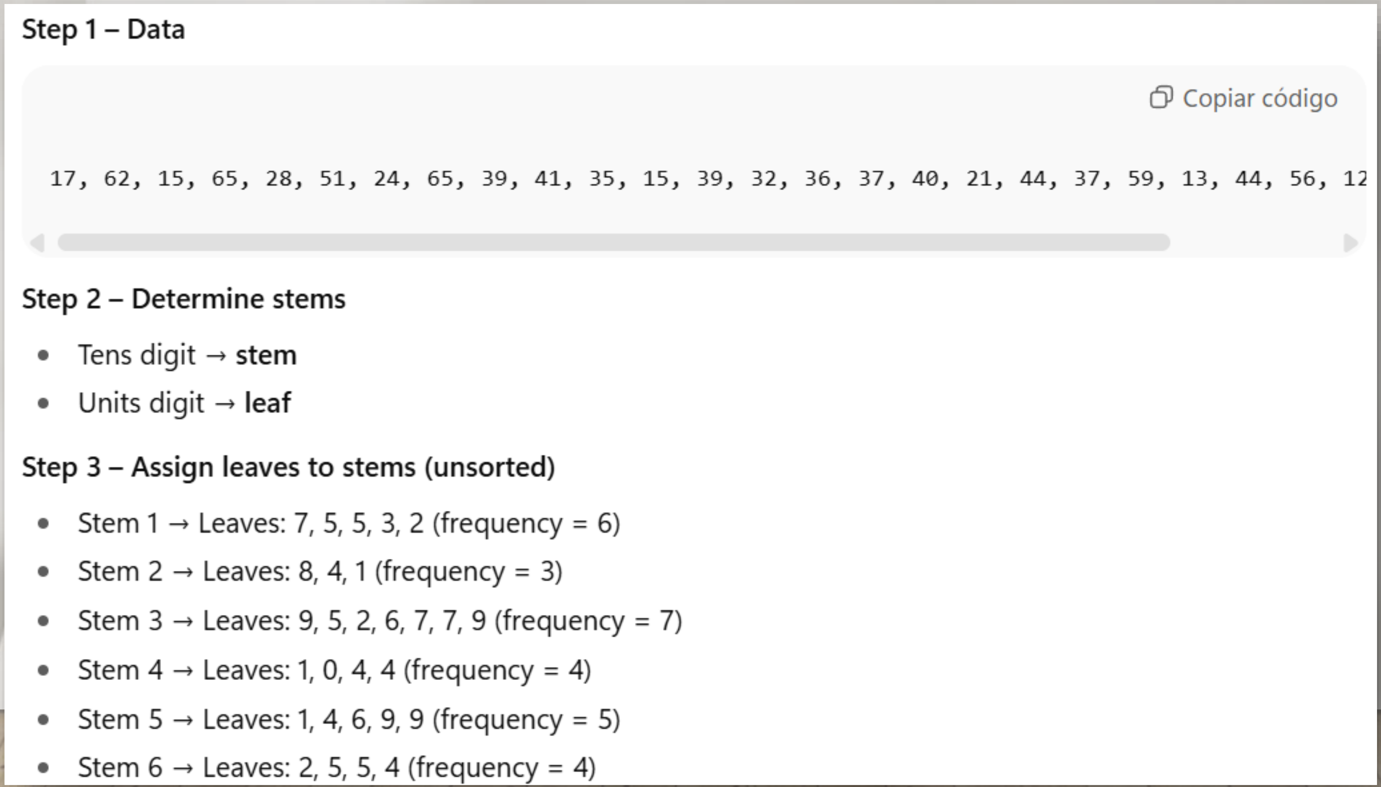
17	62	15	65
28	51	24	65
39	41	35	15
39	32	36	37
40	21	44	37
59	13	44	56
12	54	64	59

a. Construct a frequency distribution.
b. Construct a histogram.
c. Construct an ogive.
d. Construct a stem-and-leaf display.

Newbold et al (2013)

EXERCISE 1.32 D): SOLUTION

✅ Answer:



Step 1 – Data

```
17, 62, 15, 65, 28, 51, 24, 65, 39, 41, 35, 15, 39, 32, 36, 37, 40, 21, 44, 37, 59, 13, 44, 56, 12
```

Step 2 – Determine stems

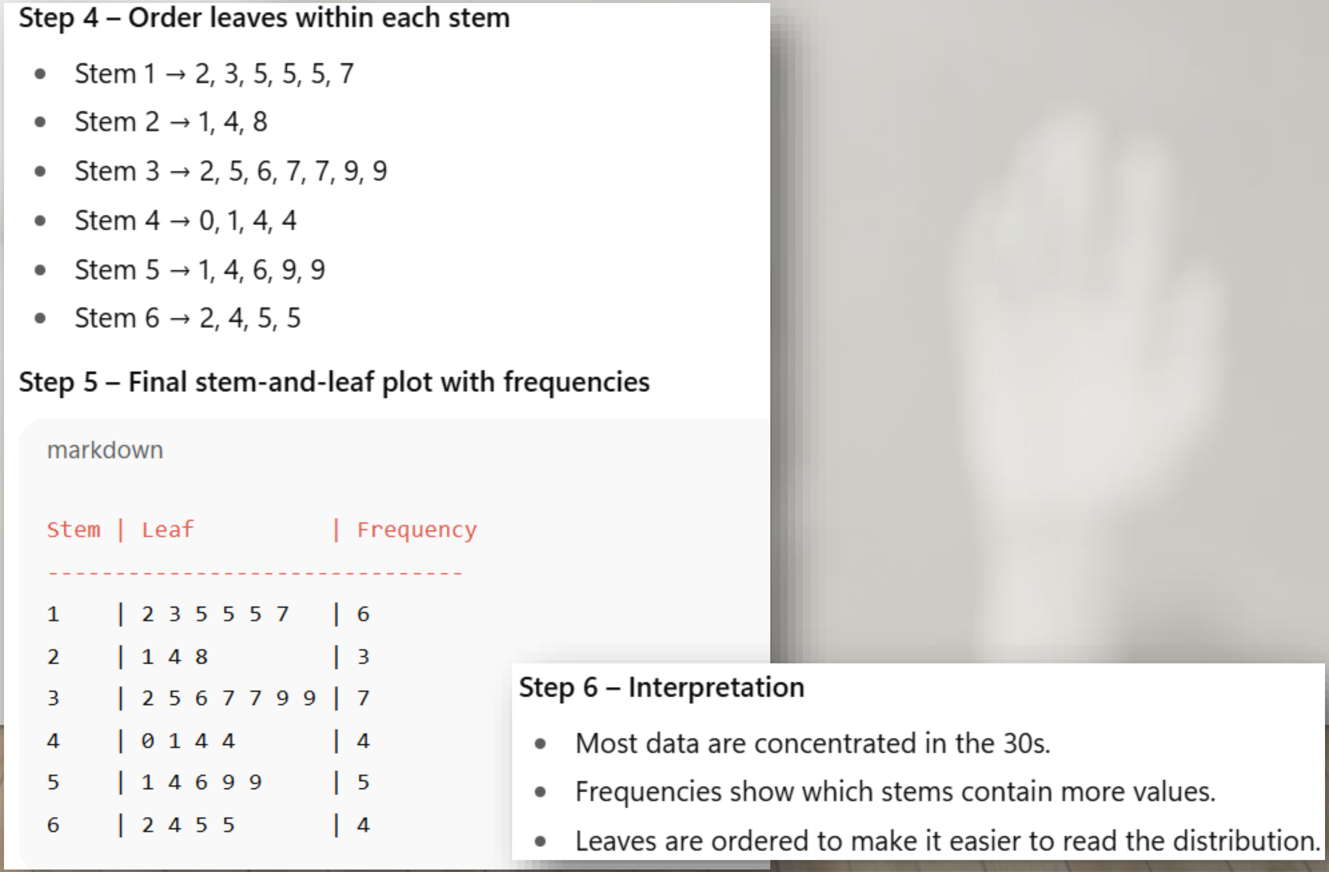
- Tens digit → **stem**
- Units digit → **leaf**

Step 3 – Assign leaves to stems (unsorted)

- Stem 1 → Leaves: 7, 5, 5, 3, 2 (frequency = 6)
- Stem 2 → Leaves: 8, 4, 1 (frequency = 3)
- Stem 3 → Leaves: 9, 5, 2, 6, 7, 7, 9 (frequency = 7)
- Stem 4 → Leaves: 1, 0, 4, 4 (frequency = 4)
- Stem 5 → Leaves: 1, 4, 6, 9, 9 (frequency = 5)
- Stem 6 → Leaves: 2, 5, 5, 4 (frequency = 4)

EXERCISE 1.32 D): SOLUTION

✅ Answer:



Step 4 – Order leaves within each stem

- Stem 1 → 2, 3, 5, 5, 5, 7
- Stem 2 → 1, 4, 8
- Stem 3 → 2, 5, 6, 7, 7, 9, 9
- Stem 4 → 0, 1, 4, 4
- Stem 5 → 1, 4, 6, 9, 9
- Stem 6 → 2, 4, 5, 5

Step 5 – Final stem-and-leaf plot with frequencies

markdown

```
Stem | Leaf          | Frequency
-------------------------------
1    | 2 3 5 5 5 7   | 6
2    | 1 4 8         | 3
3    | 2 5 6 7 7 9 9 | 7
4    | 0 1 4 4       | 4
5    | 1 4 6 9 9     | 5
6    | 2 4 5 5       | 4
```

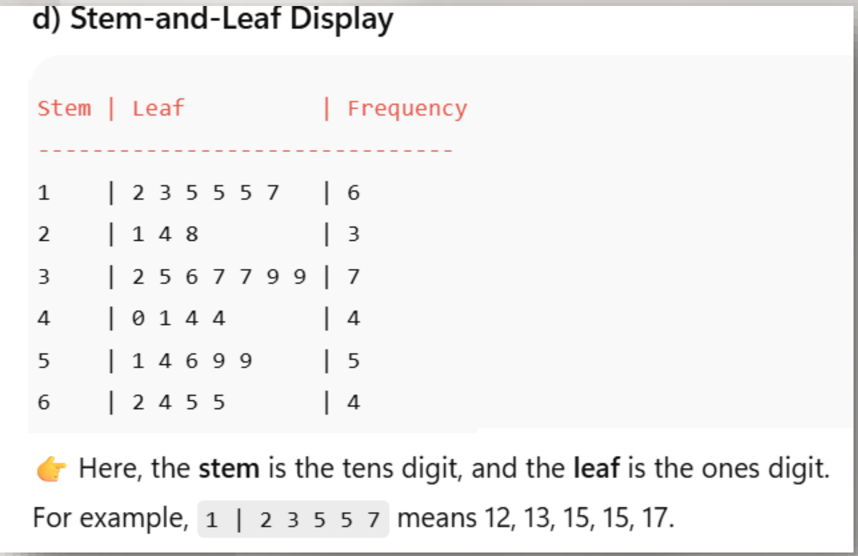
Step 6 – Interpretation

- Most data are concentrated in the 30s.
- Frequencies show which stems contain more values.
- Leaves are ordered to make it easier to read the distribution.

EXERCISE 1.32 D): SOLUTION

✅ Answer:

SPSS Output: Steam and Leaf Plot

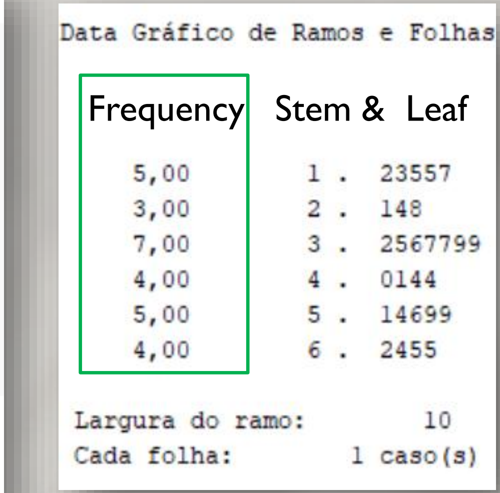


d) Stem-and-Leaf Display

```
Stem | Leaf            | Frequency
-------------------------------
1    | 2 3 5 5 5 7     | 6
2    | 1 4 8           | 3
3    | 2 5 6 7 7 9 9 | 7
4    | 0 1 4 4         | 4
5    | 1 4 6 9 9       | 5
6    | 2 4 5 5         | 4
```

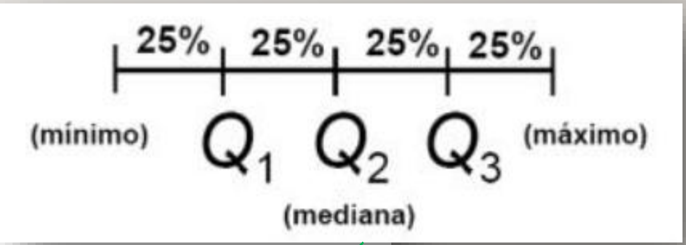
👉 Here, the **stem** is the tens digit, and the **leaf** is the ones digit.

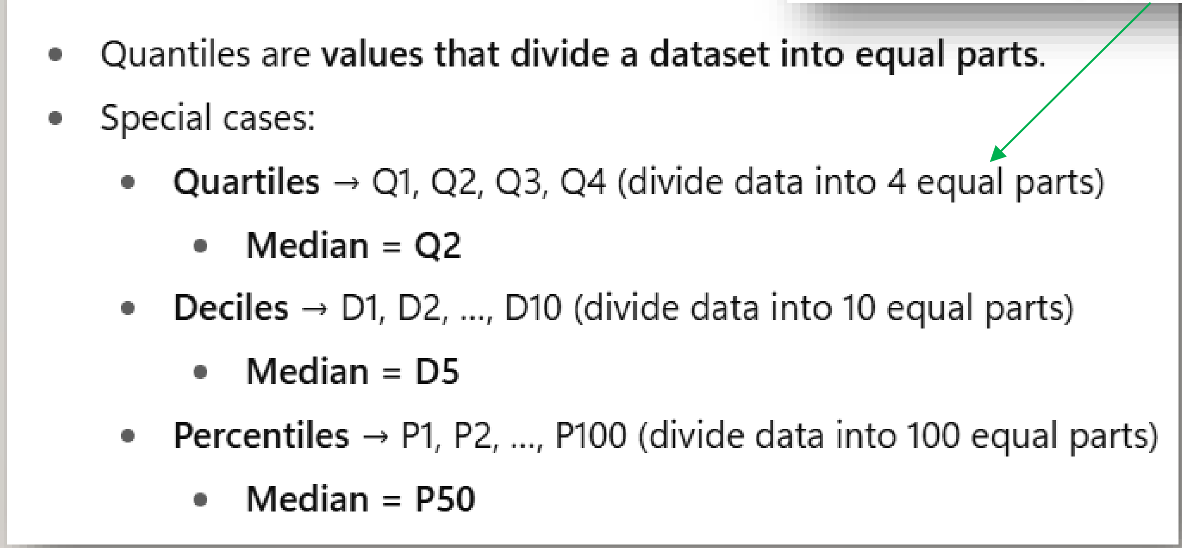
For example, `1 | 2 3 5 5 7` means 12, 13, 15, 15, 17.



```
Data Gráfico de Ramos e Folhas

 Frequency    Stem &  Leaf

     5,00        1 .  23557
     3,00        2 .  148
     7,00        3 .  2567799
     4,00        4 .  0144
     5,00        5 .  14699
     4,00        6 .  2455

 Largura do ramo:         10
 Cada folha:         1 caso(s)
```

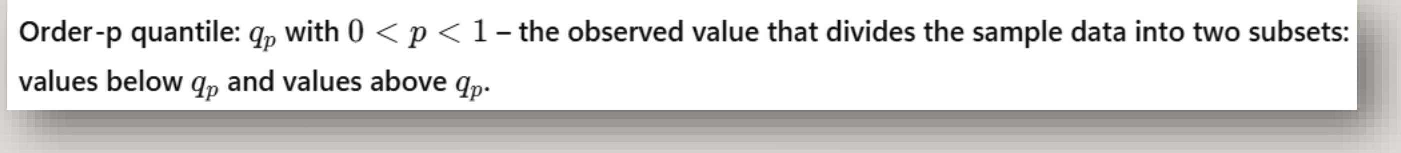
QUANTILES: DEFINITION





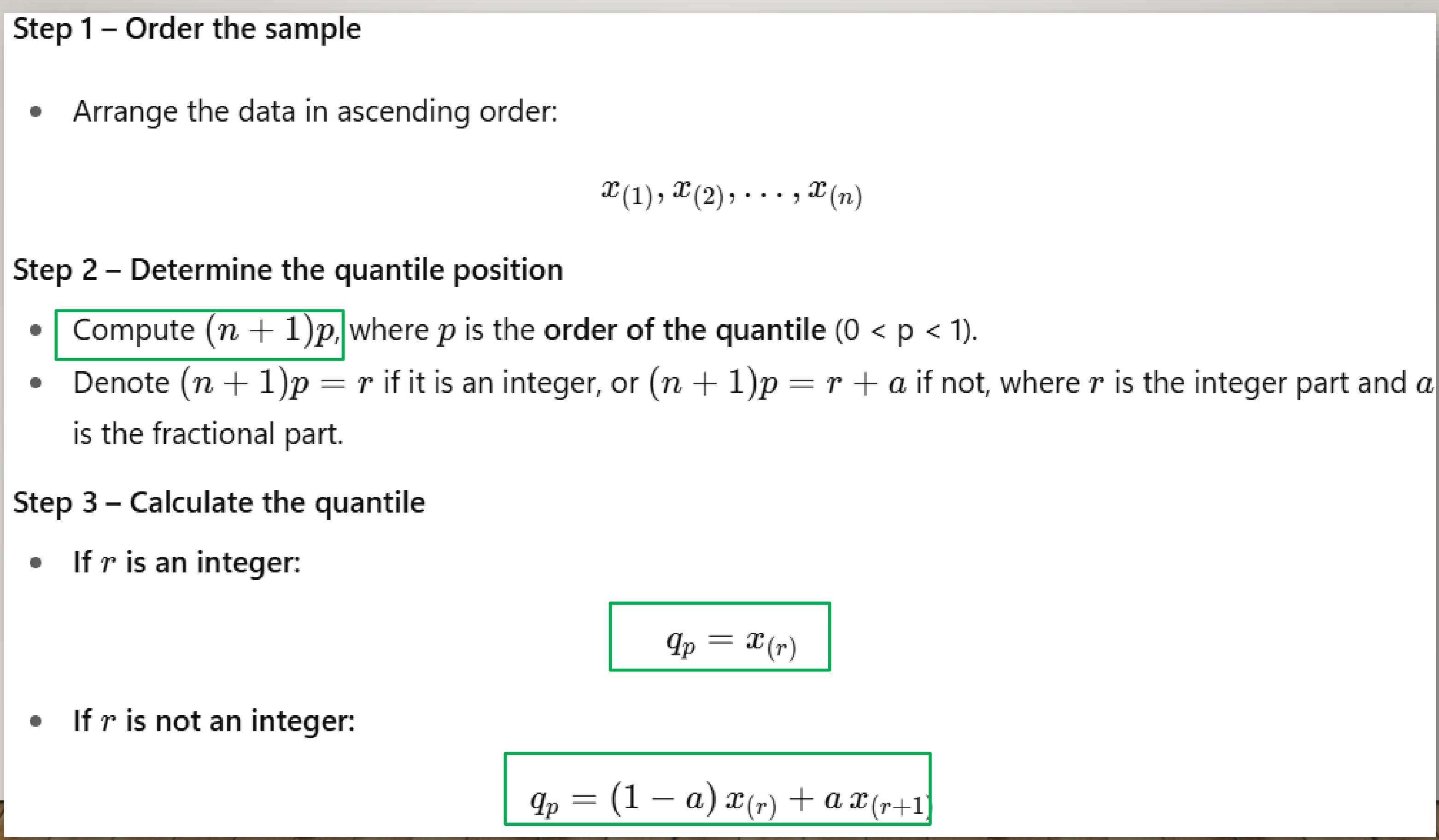
1️⃣ What are Quantiles?

- Quantiles are **values that divide a dataset into equal parts.**
- Special cases:
  - **Quartiles** → Q1, Q2, Q3, Q4 (divide data into 4 equal parts)
    - **Median = Q2**
  - **Deciles** → D1, D2, ..., D10 (divide data into 10 equal parts)
    - **Median = D5**
  - **Percentiles** → P1, P2, ..., P100 (divide data into 100 equal parts)
    - **Median = P50**



Order-p quantile: q_p with $0 < p < 1$ – the observed value that divides the sample data into two subsets: values below q_p and values above q_p.

STEPS TO CALCULATE AN ORDER-P QUANTILE (q_p)



Step 1 – Order the sample

- Arrange the data in ascending order:

$$x_{(1)}, x_{(2)}, \ldots, x_{(n)}$$

Step 2 – Determine the quantile position

- Compute $(n+1)p$, where p is the **order of the quantile** (0 < p < 1).
- Denote $(n+1)p = r$ if it is an integer, or $(n+1)p = r + a$ if not, where r is the integer part and a is the fractional part.

Step 3 – Calculate the quantile

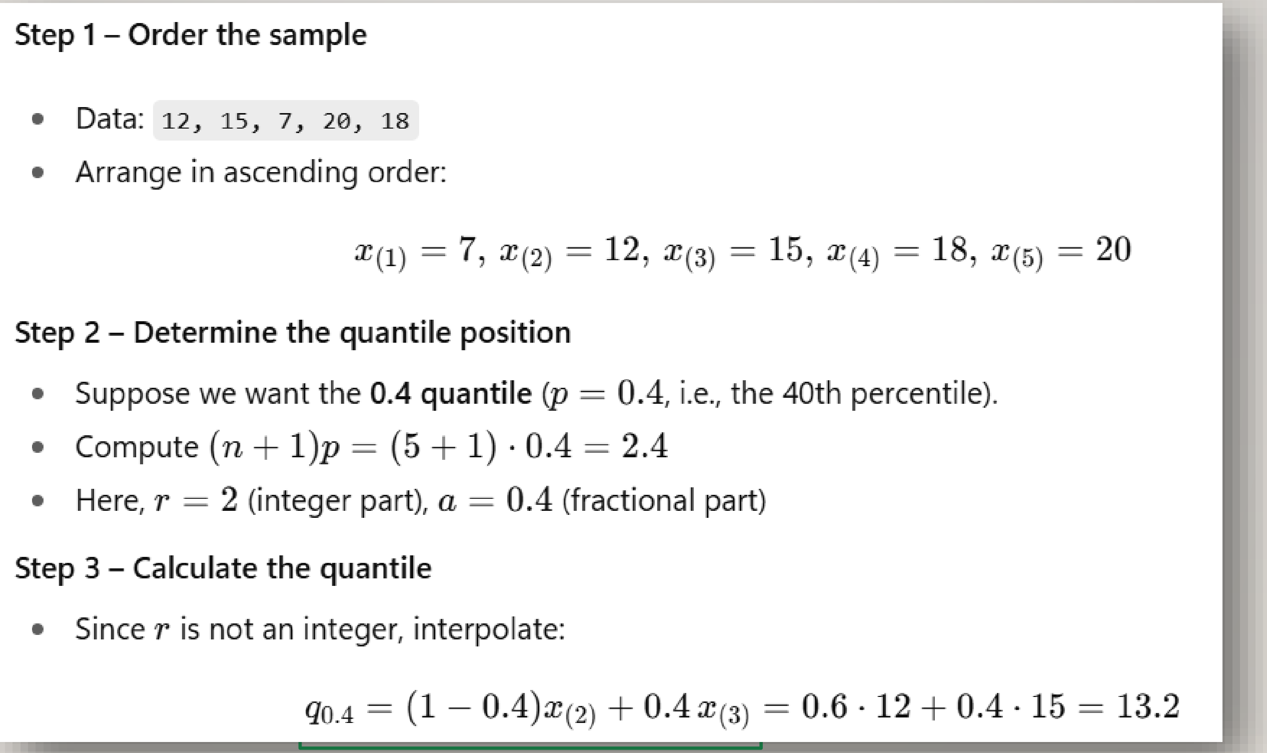
- **If r is an integer:**

$$\boxed{q_p = x_{(r)}}$$

- **If r is not an integer:**

$$\boxed{q_p = (1-a)\, x_{(r)} + a\, x_{(r+1)}}$$

STEPS TO CALCULATE AN ORDER-P QUANTILE (q_p) WITH EXAMPLE



Step 1 – Order the sample

- Data: `12, 15, 7, 20, 18`
- Arrange in ascending order:

$$x_{(1)} = 7,\ x_{(2)} = 12,\ x_{(3)} = 15,\ x_{(4)} = 18,\ x_{(5)} = 20$$

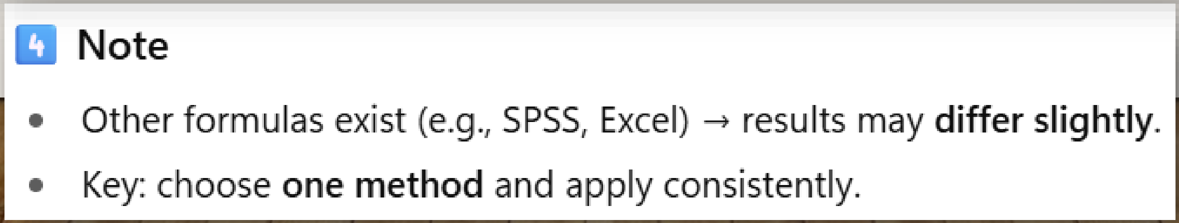
Step 2 – Determine the quantile position

- Suppose we want the **0.4 quantile** ($p = 0.4$, i.e., the 40th percentile).
- Compute $(n+1)p = (5+1) \cdot 0.4 = 2.4$
- Here, $r = 2$ (integer part), $a = 0.4$ (fractional part)

Step 3 – Calculate the quantile

- Since r is not an integer, interpolate:

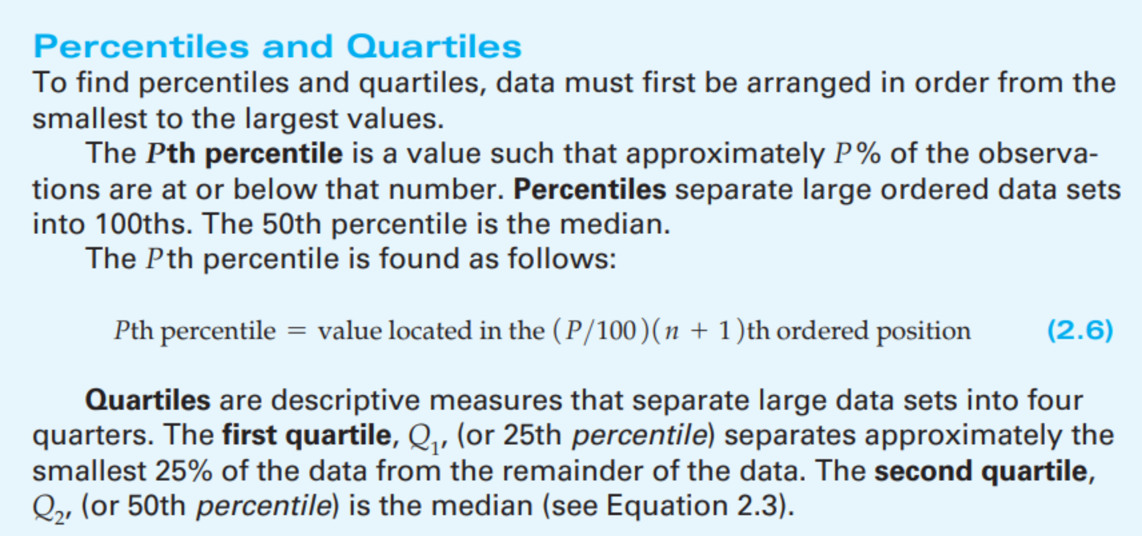
$$q_{0.4} = (1 - 0.4)x_{(2)} + 0.4\,x_{(3)} = 0.6 \cdot 12 + 0.4 \cdot 15 = 13.2$$



4️⃣ Note

- Other formulas exist (e.g., SPSS, Excel) → results may **differ slightly**.
- Key: choose **one method** and apply consistently.

PERCENTILES AND QUARTILES



Percentiles and Quartiles

To find percentiles and quartiles, data must first be arranged in order from the smallest to the largest values.

The **Pth percentile** is a value such that approximately $P\%$ of the observations are at or below that number. **Percentiles** separate large ordered data sets into 100ths. The 50th percentile is the median.

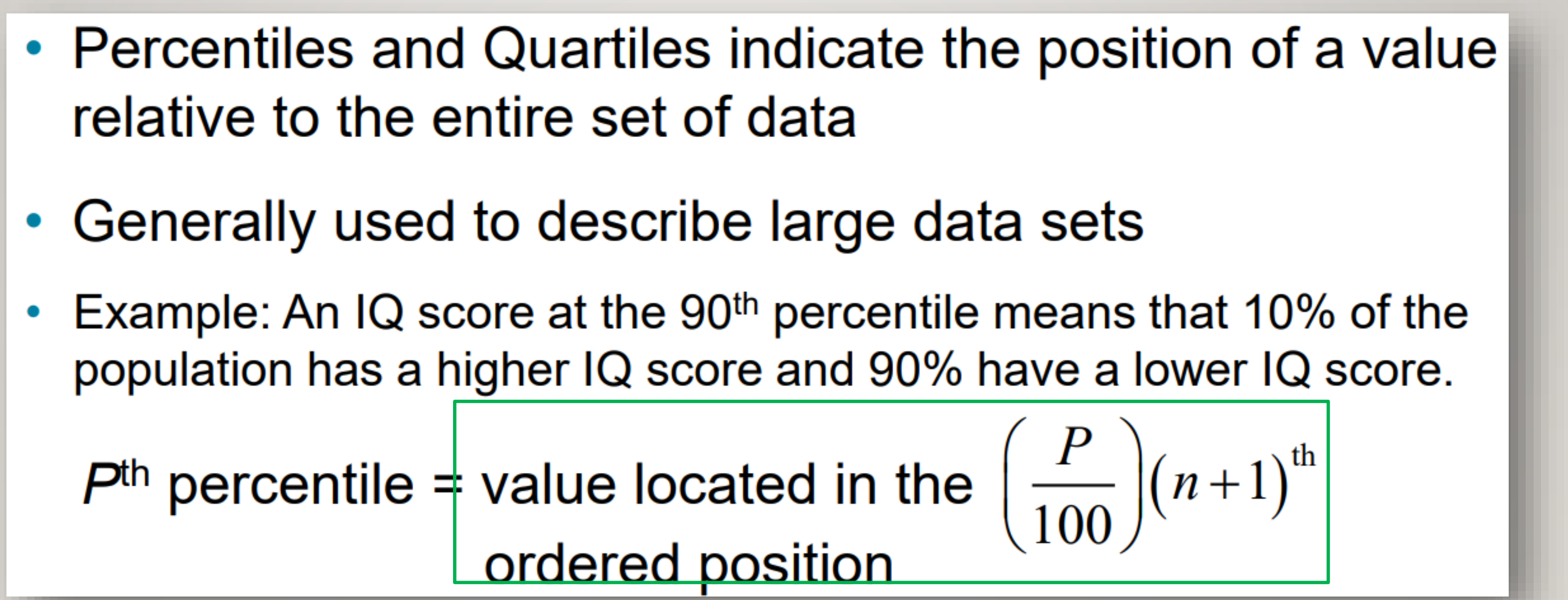
The Pth percentile is found as follows:

$$P\text{th percentile} = \text{value located in the } (P/100)(n + 1)\text{th ordered position} \qquad (2.6)$$

Quartiles are descriptive measures that separate large data sets into four quarters. The **first quartile**, Q_1, (or 25th *percentile*) separates approximately the smallest 25% of the data from the remainder of the data. The **second quartile**, Q_2, (or 50th *percentile*) is the median (see Equation 2.3).

Newbold et al (2013)

PERCENTILES AND QUARTILES

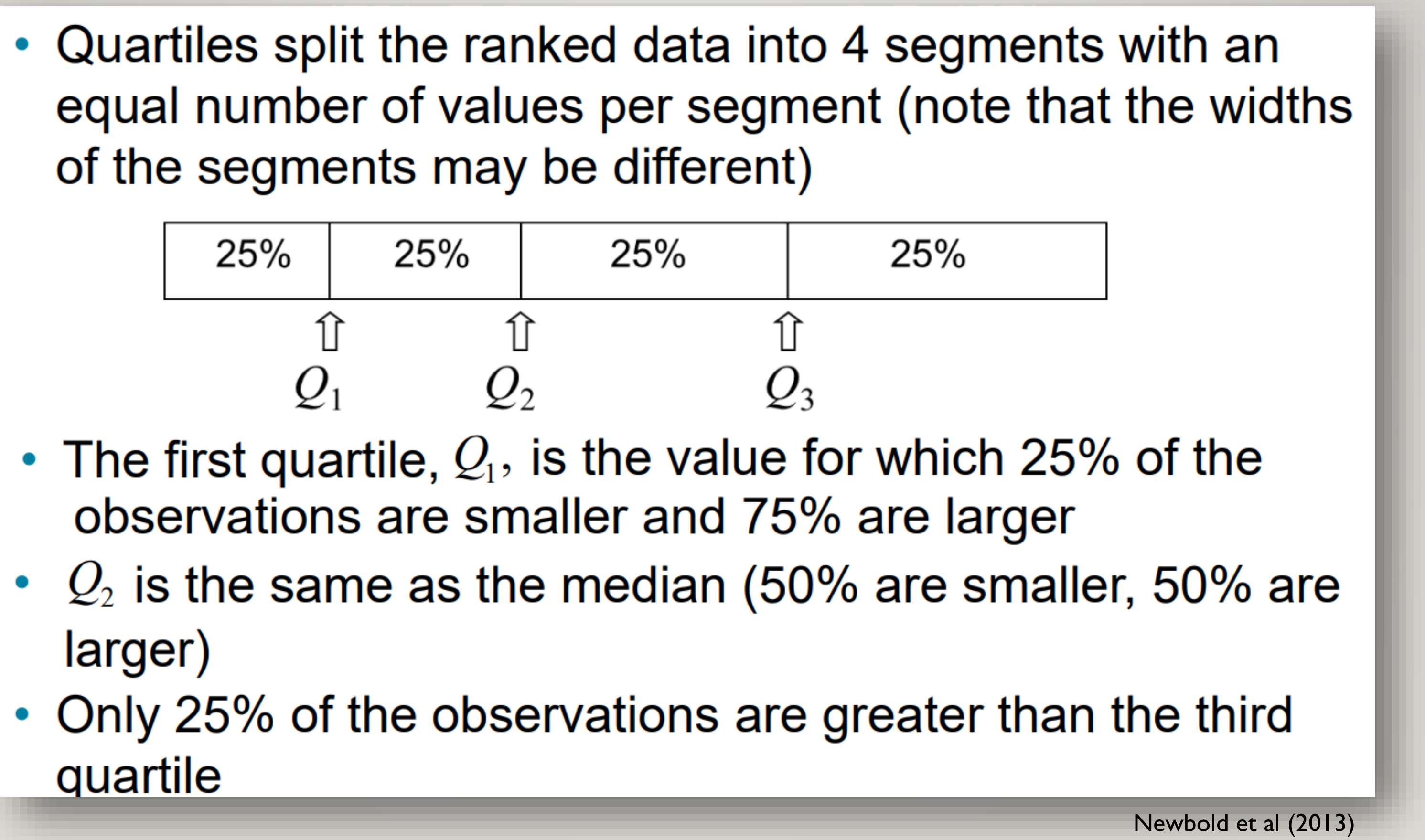


- Percentiles and Quartiles indicate the position of a value relative to the entire set of data
- Generally used to describe large data sets
- Example: An IQ score at the 90th percentile means that 10% of the population has a higher IQ score and 90% have a lower IQ score.

P^{th} percentile = value located in the $\left(\frac{P}{100}\right)(n+1)^{th}$ ordered position

Newbold et al (2013)

QUARTILES



- Quartiles split the ranked data into 4 segments with an equal number of values per segment (note that the widths of the segments may be different)

25%	25%	25%	25%

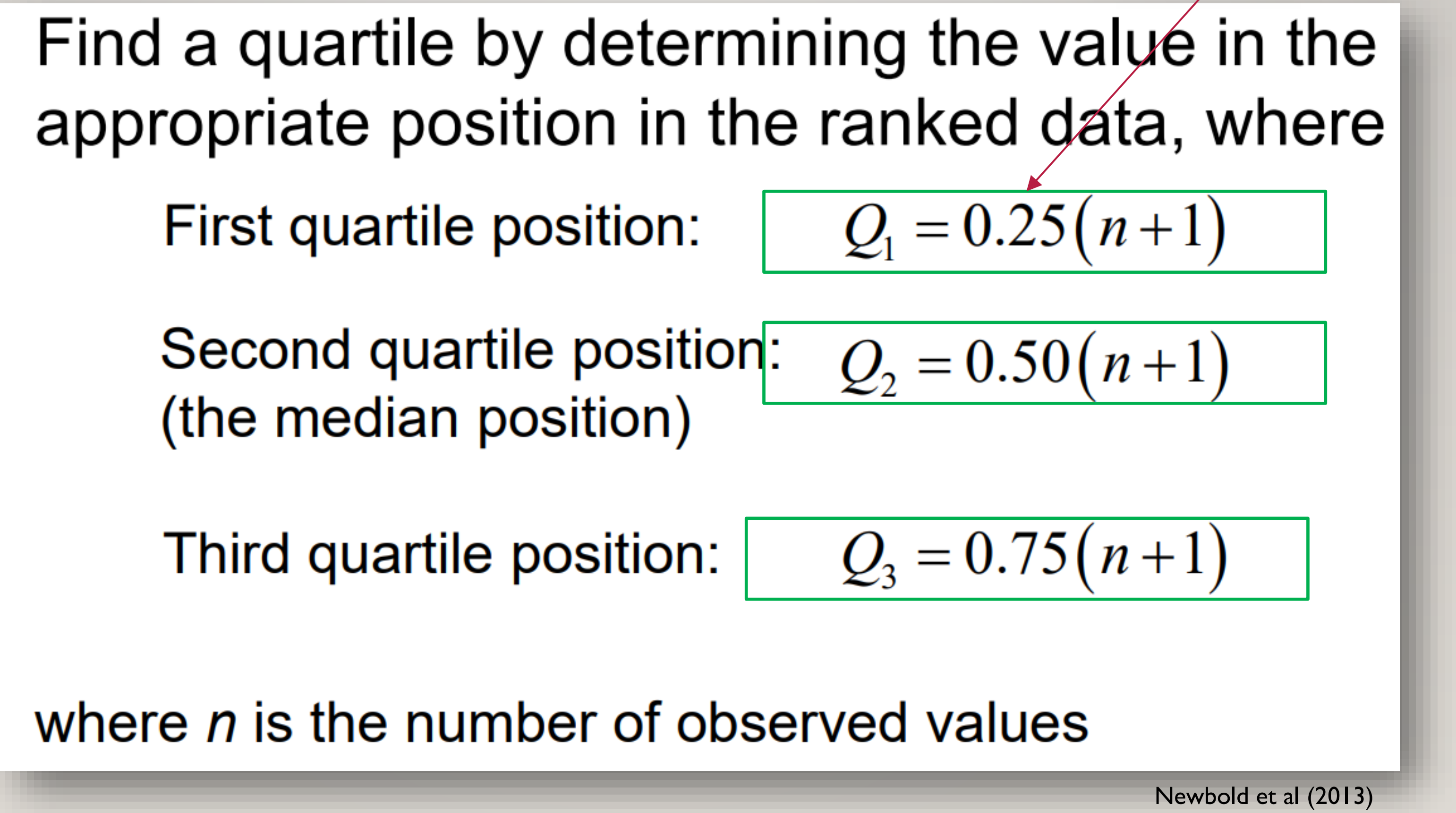
⇧ Q_1 ⇧ Q_2 ⇧ Q_3

- The first quartile, Q_1, is the value for which 25% of the observations are smaller and 75% are larger
- Q_2 is the same as the median (50% are smaller, 50% are larger)
- Only 25% of the observations are greater than the third quartile

Newbold et al (2013)

QUARTILE FORMULAS

$(n+1)p$



Find a quartile by determining the value in the appropriate position in the ranked data, where

First quartile position: $Q_1 = 0.25(n+1)$

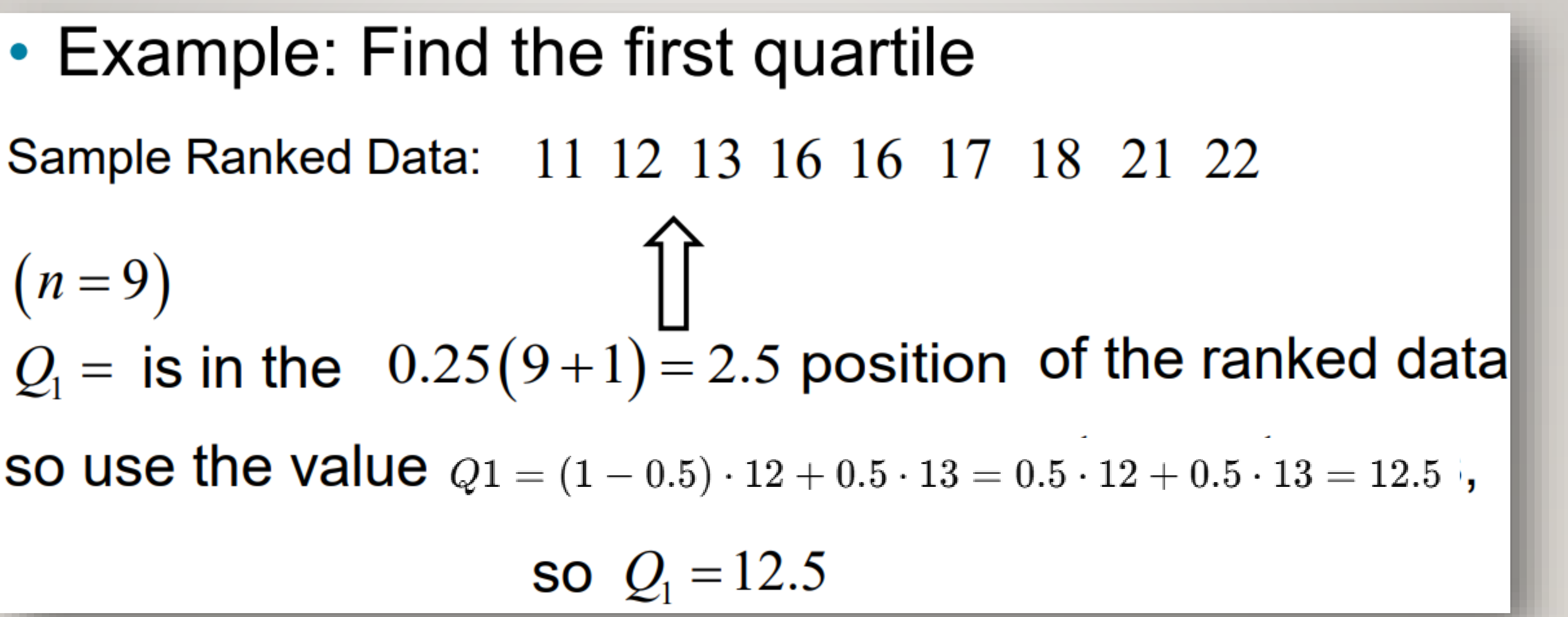
Second quartile position: $Q_2 = 0.50(n+1)$
(the median position)

Third quartile position: $Q_3 = 0.75(n+1)$

where *n* is the number of observed values

Newbold et al (2013)

QUARTILE: EXAMPLE



- Example: Find the first quartile

Sample Ranked Data: 11 12 13 16 16 17 18 21 22

⇧

$(n=9)$

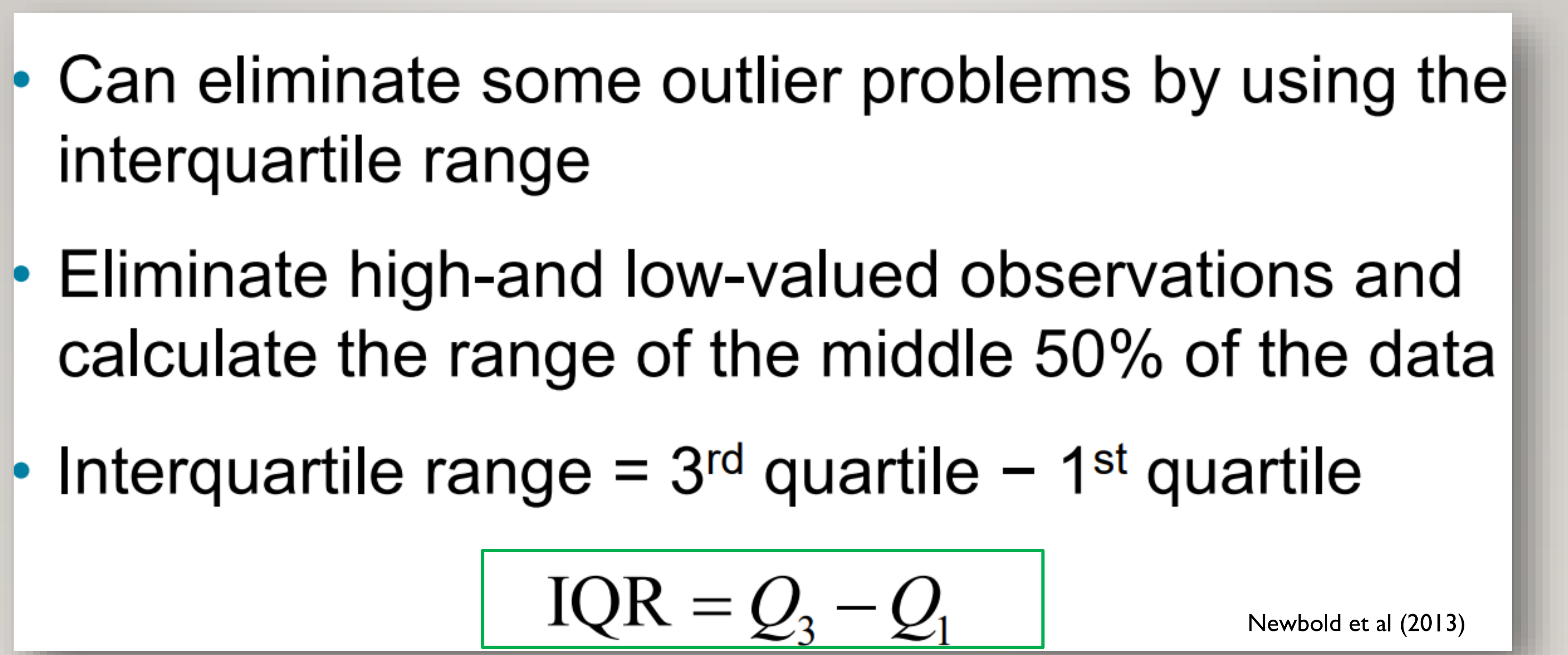
$Q_1 =$ is in the $0.25(9+1)=2.5$ position of the ranked data

so use the value $Q1 = (1-0.5)\cdot 12 + 0.5\cdot 13 = 0.5\cdot 12 + 0.5\cdot 13 = 12.5$,

so $Q_1 = 12.5$

Newbold et al (2013)

INTERQUARTILE RANGE



- Can eliminate some outlier problems by using the interquartile range
- Eliminate high-and low-valued observations and calculate the range of the middle 50% of the data
- Interquartile range = 3rd quartile – 1st quartile

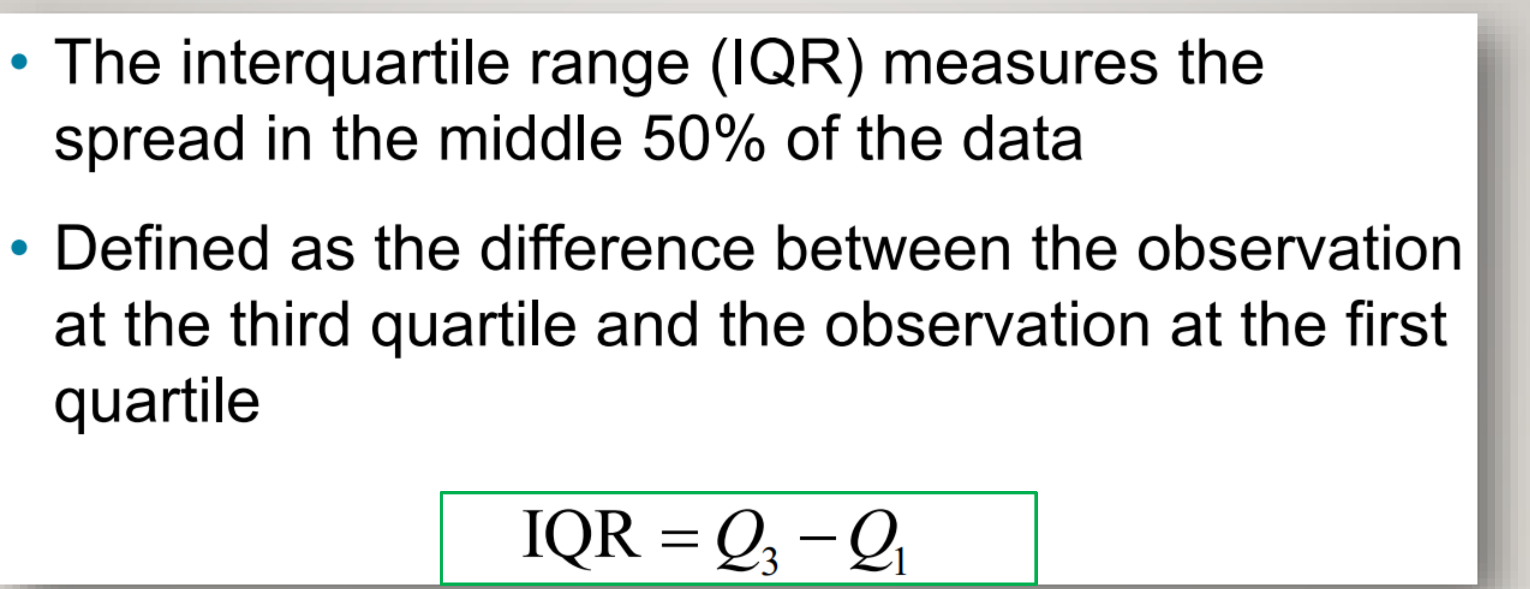
$$\boxed{\mathrm{IQR} = Q_3 - Q_1}$$

Newbold et al (2013)

"The interquartile range (IQR) is **technically defined** as the difference between the third and first quartiles, that is, $IQR = Q3 - Q1$. This is the standard definition found in most statistics textbooks and software (Excel, R, Python).

However, some sources or people may informally refer to the 'interquartile interval' as the interval [Q1, Q3], describing the range that contains the middle 50% of the data. This is not the standard technical definition of the IQR, but rather a way of describing the central spread of the data."

INTERQUARTILE RANGE

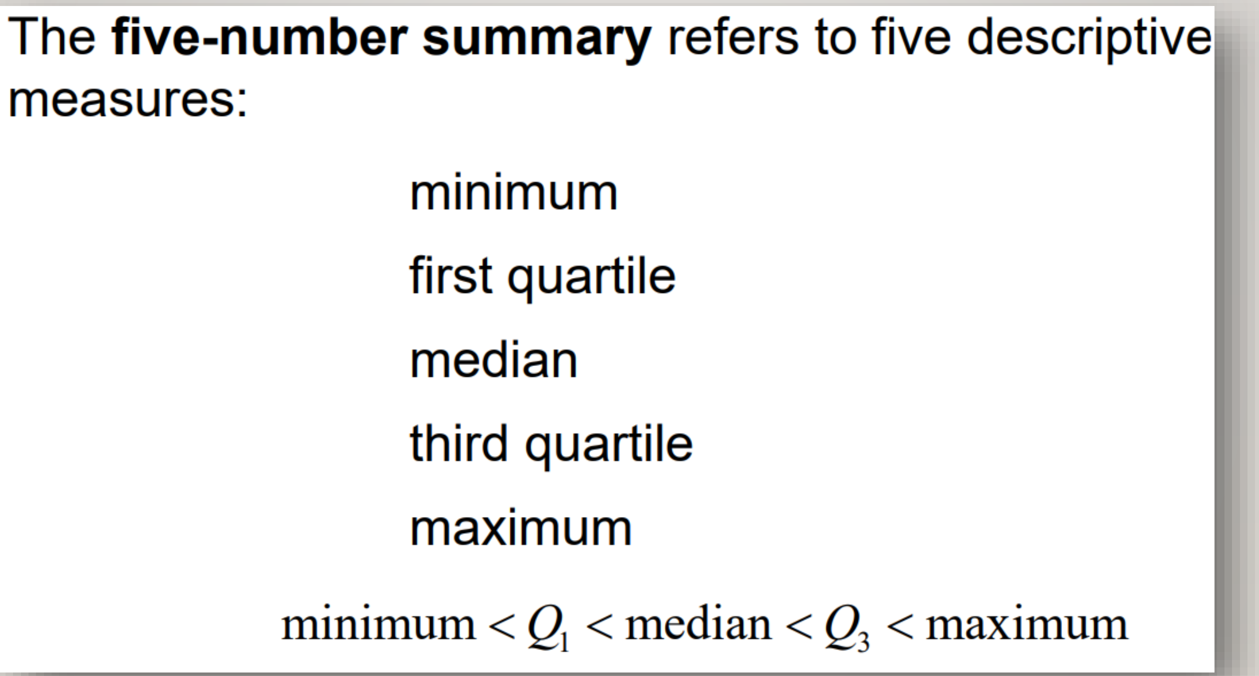


- The interquartile range (IQR) measures the spread in the middle 50% of the data
- Defined as the difference between the observation at the third quartile and the observation at the first quartile

$$\boxed{\text{IQR} = Q_3 - Q_1}$$

Newbold et al (2013)

FIVE-NUMBER SUMMARY



The **five-number summary** refers to five descriptive measures:

minimum

first quartile

median

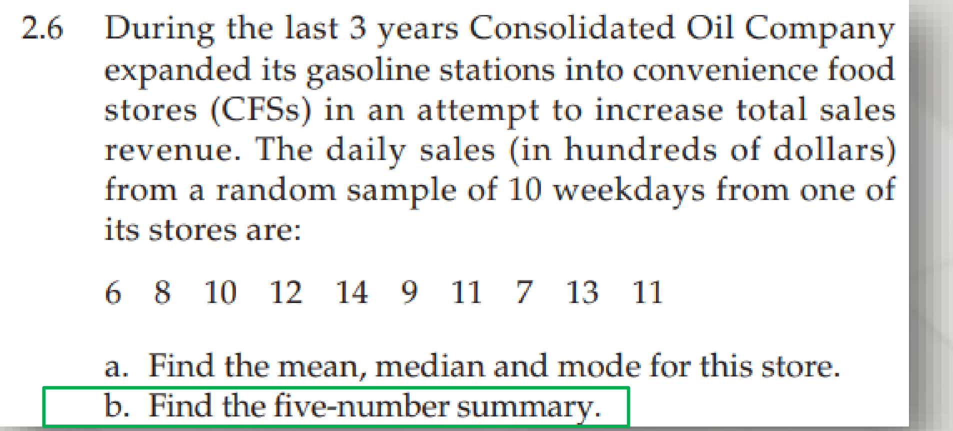
third quartile

maximum

$$\text{minimum} < Q_1 < \text{median} < Q_3 < \text{maximum}$$

Newbold et al (2013)

EXERCISE 2.6



2.6 During the last 3 years Consolidated Oil Company expanded its gasoline stations into convenience food stores (CFSs) in an attempt to increase total sales revenue. The daily sales (in hundreds of dollars) from a random sample of 10 weekdays from one of its stores are:

6 8 10 12 14 9 11 7 13 11

a. Find the mean, median and mode for this store.

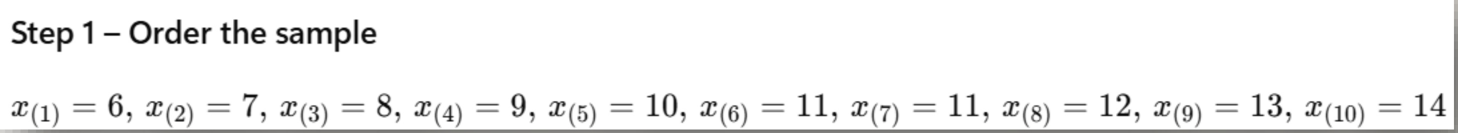
b. Find the five-number summary.

Newbold et al (2013)

EXERCISE 2.6 B): SOLUTION

✅ Answer:

,



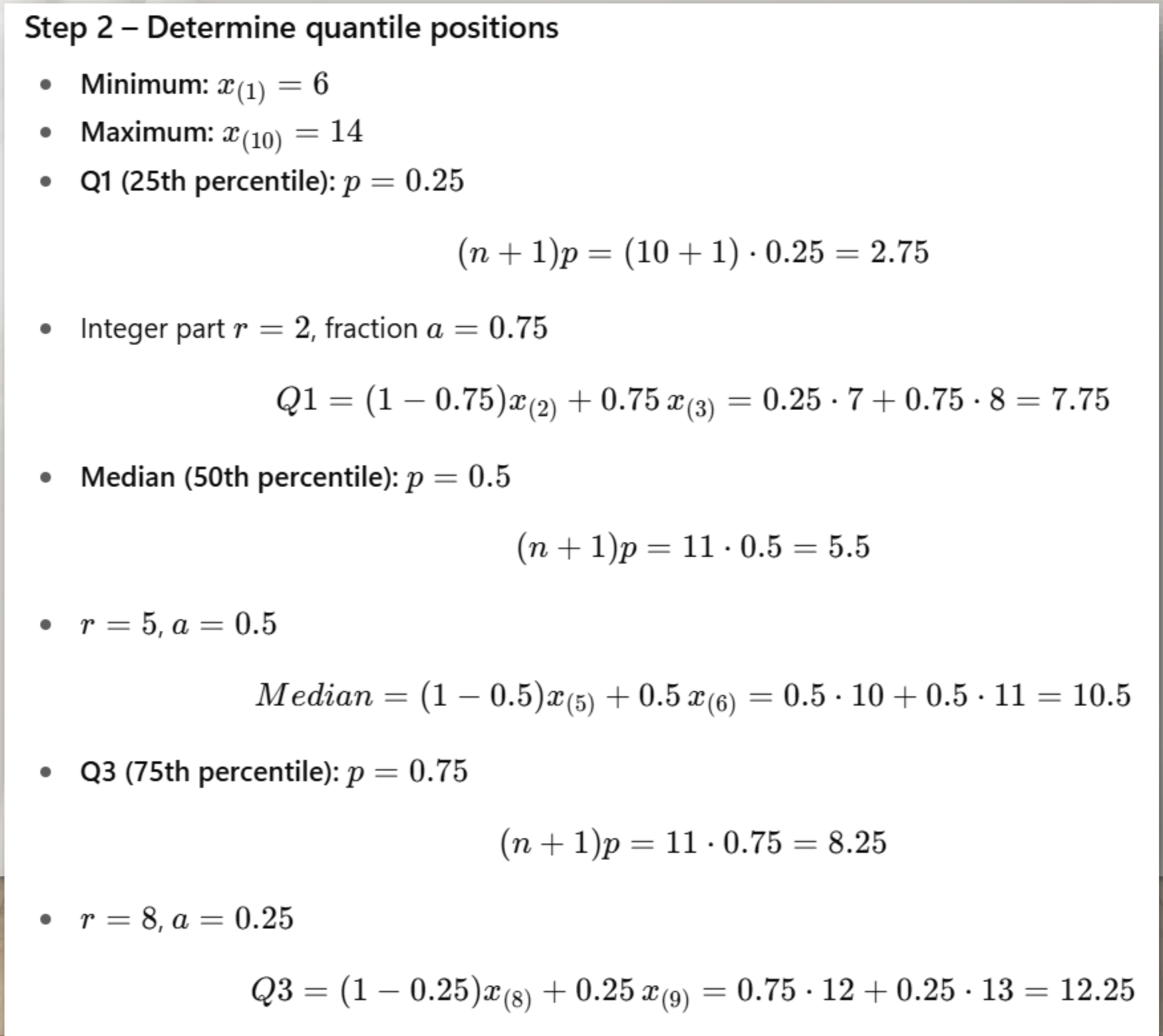
Step 1 – Order the sample

$x_{(1)} = 6,\ x_{(2)} = 7,\ x_{(3)} = 8,\ x_{(4)} = 9,\ x_{(5)} = 10,\ x_{(6)} = 11,\ x_{(7)} = 11,\ x_{(8)} = 12,\ x_{(9)} = 13,\ x_{(10)} = 14$

EXERCISE 2.6 B): SOLUTION

✅ Answer:

,



Step 2 – Determine quantile positions

- **Minimum:** $x_{(1)} = 6$
- **Maximum:** $x_{(10)} = 14$
- **Q1 (25th percentile):** $p = 0.25$

$$(n+1)p = (10+1) \cdot 0.25 = 2.75$$

- Integer part $r = 2$, fraction $a = 0.75$

$$Q1 = (1-0.75)x_{(2)} + 0.75\,x_{(3)} = 0.25 \cdot 7 + 0.75 \cdot 8 = 7.75$$

- **Median (50th percentile):** $p = 0.5$

$$(n+1)p = 11 \cdot 0.5 = 5.5$$

- $r = 5$, $a = 0.5$

$$Median = (1-0.5)x_{(5)} + 0.5\,x_{(6)} = 0.5 \cdot 10 + 0.5 \cdot 11 = 10.5$$

- **Q3 (75th percentile):** $p = 0.75$

$$(n+1)p = 11 \cdot 0.75 = 8.25$$

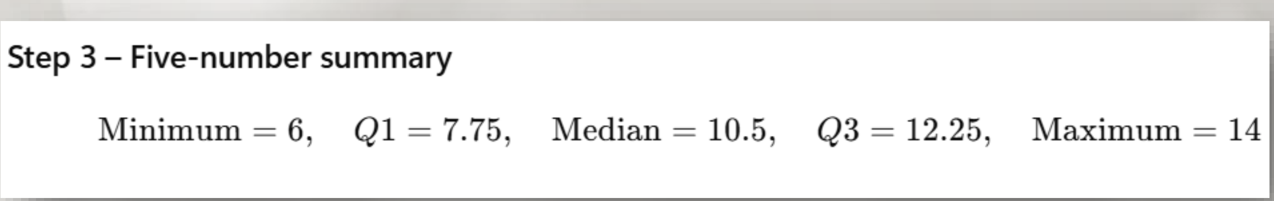
- $r = 8$, $a = 0.25$

$$Q3 = (1-0.25)x_{(8)} + 0.25\,x_{(9)} = 0.75 \cdot 12 + 0.25 \cdot 13 = 12.25$$

EXERCISE 2.6 B): SOLUTION

✅ Answer:

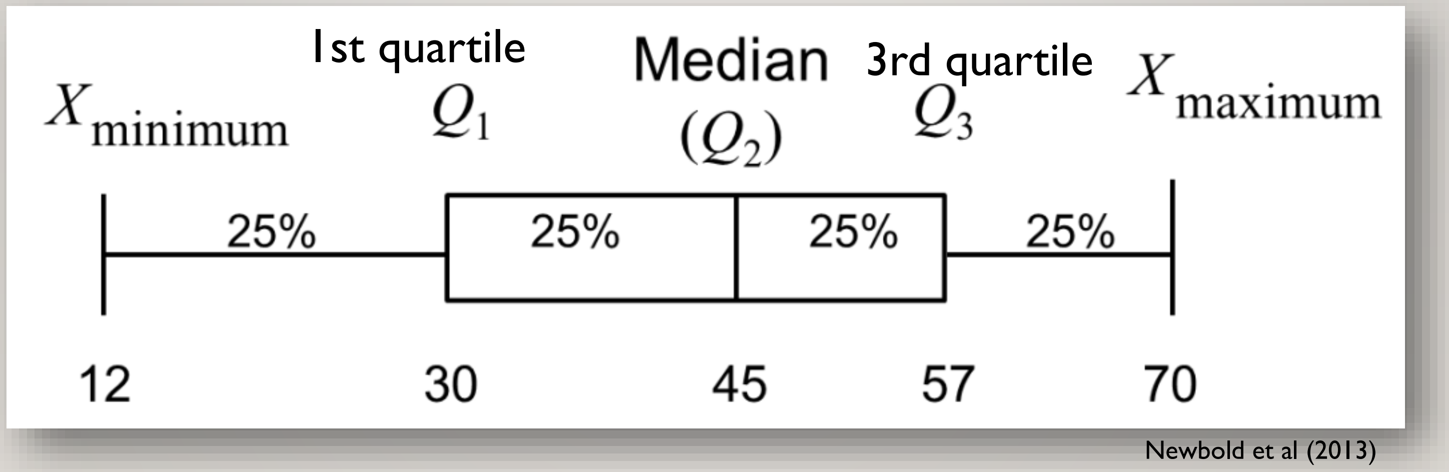
,



Step 3 – Five-number summary

$$\text{Minimum} = 6, \quad Q1 = 7.75, \quad \text{Median} = 10.5, \quad Q3 = 12.25, \quad \text{Maximum} = 14$$

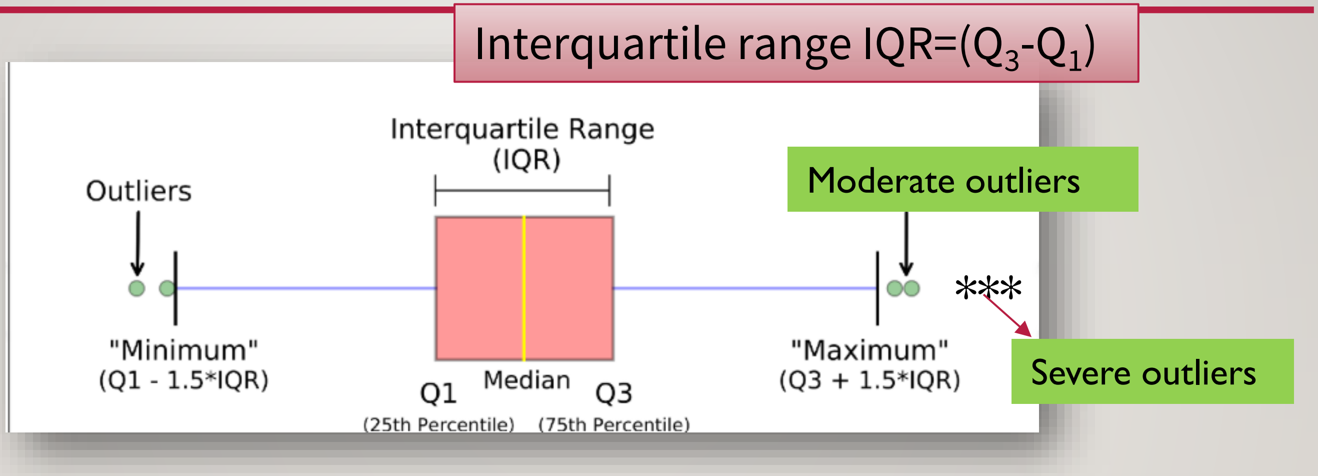
BOX-AND-WHISKER PLOT/ BOXPLOT EXAMPLE



Newbold et al (2013)

The plot can be oriented horizontally or vertically.

BOXPLOT AND OUTLIERS



Interquartile range IQR=(Q_3-Q_1)

Moderate outliers

Severe outliers

Moderate outliers (marked with a circle)

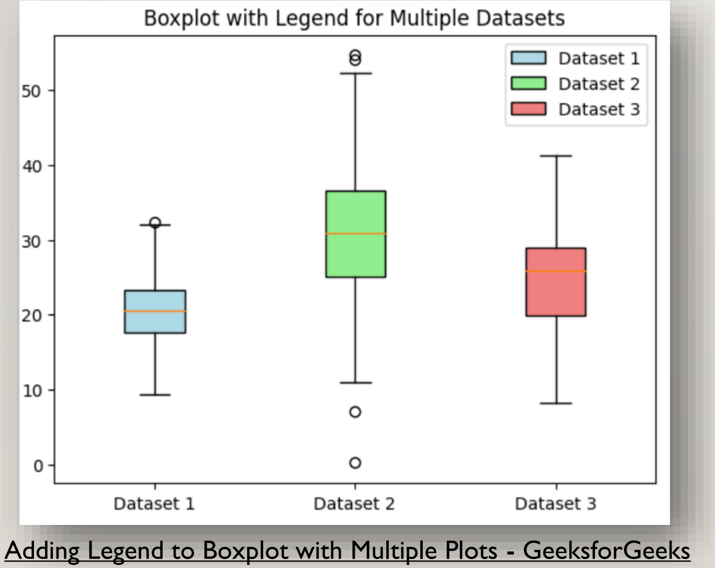
(Q_1-1,5 × IQR; Q_3+1,5 × IQR) Inner fences

Severe outliers (marked with an asterisk)

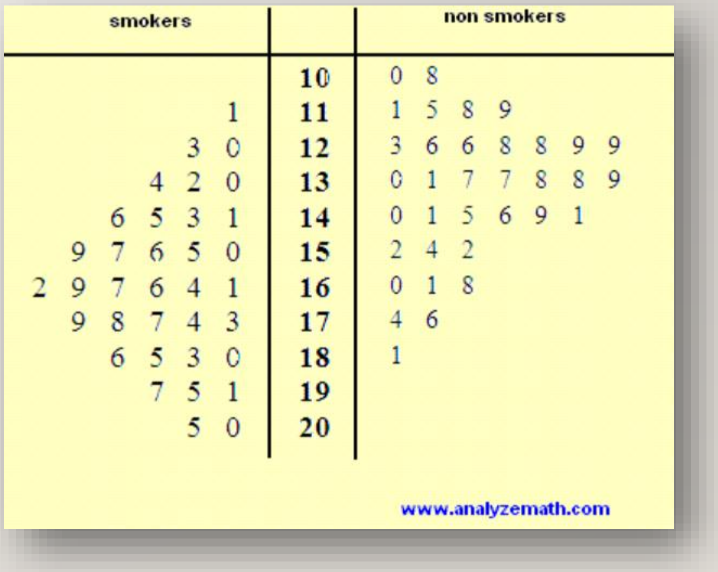
(Q_1-3 × IQR; Q_3+3 × IQR) Outer fences

COMPARING DATA SETS

- Use measures of **central tendency, dispersion, and shape**.
- Visualize with **histograms, boxplots, and stem-and-leaf plots**.
- Identify **patterns, similarities, and differences** between data sets.



Adding Legend to Boxplot with Multiple Plots - GeeksforGeeks



smokers		non smokers
	10	0 8
1	**11**	1 5 8 9
3 0	**12**	3 6 6 8 8 9 9
4 2 0	**13**	0 1 7 7 8 8 9
6 5 3 1	**14**	0 1 5 6 9 1
9 7 6 5 0	**15**	2 4 2
2 9 7 6 4 1	**16**	0 1 8
9 8 7 4 3	**17**	4 6
6 5 3 0	**18**	1
7 5 1	**19**	
5 0	**20**	

www.analyzemath.com

QUANTILE-QUANTILE (Q-Q) PLOTS

Definition

- A Q-Q plot is a graphical tool used to compare the distribution of two datasets or to check if a dataset follows a theoretical distribution.

Purpose

- If the data and the reference distribution have the same shape, the points form a straight line.
- Deviations from the line indicate differences between distributions.

Common Applications

- **Normality Check:** Compare data quantiles with a normal distribution.
- **Distribution Comparison:** Compare two datasets.
- **Outlier Detection:** Identify values that deviate from the expected pattern.

Silvestre (2007)

QUANTILE-QUANTILE (Q-Q) PLOTS: CONSTRUCTION & INTERPRETATION

Construction Steps

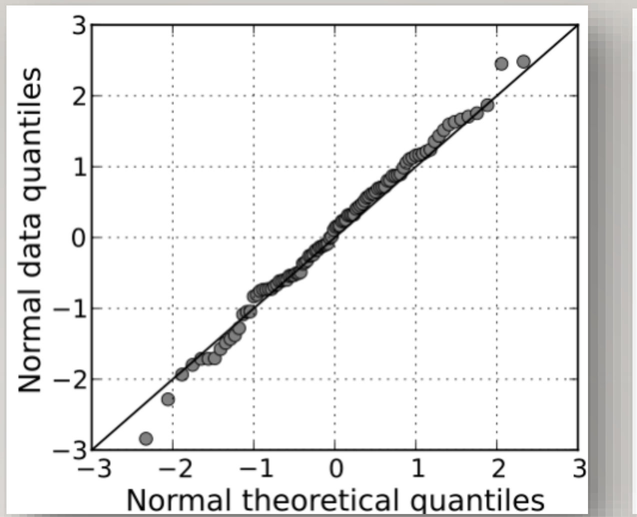
1. Sort the observed data in ascending order.
2. Calculate the corresponding theoretical quantiles.
3. Plot observed quantiles (y-axis) against theoretical quantiles (x-axis).

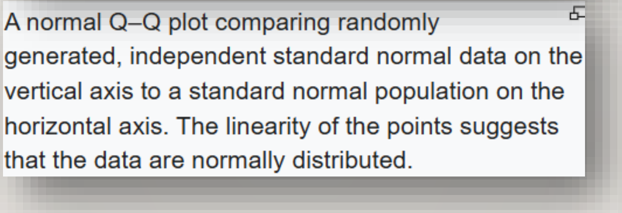
Interpretation

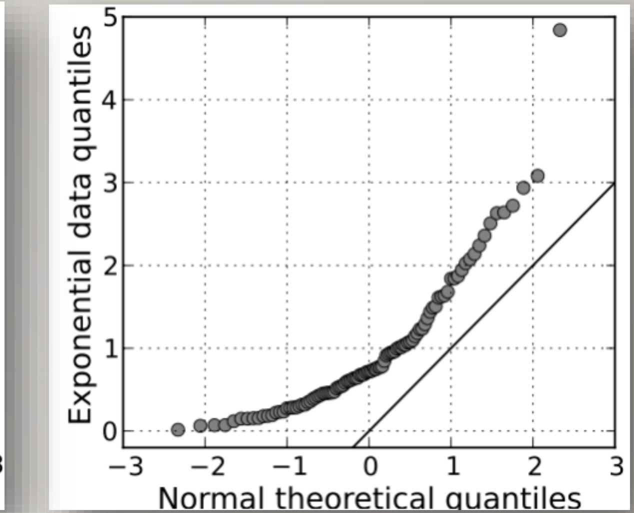
- Points close to a straight line → data follows the reference distribution.
- Points deviating from the line → data differs from the reference distribution.

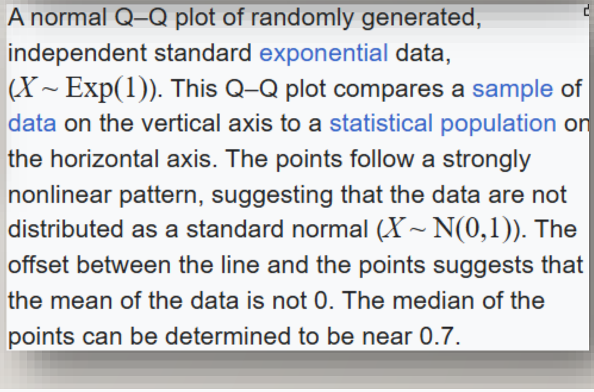
Silvestre (2007)

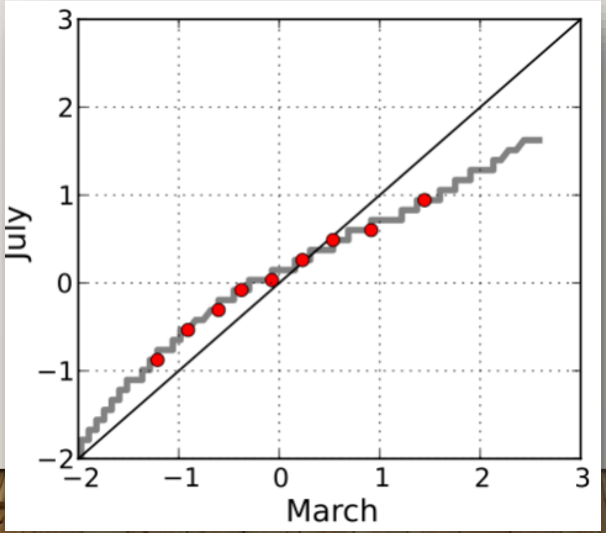
QUANTILE-QUANTILE (Q-Q) PLOTS EXAMPLES

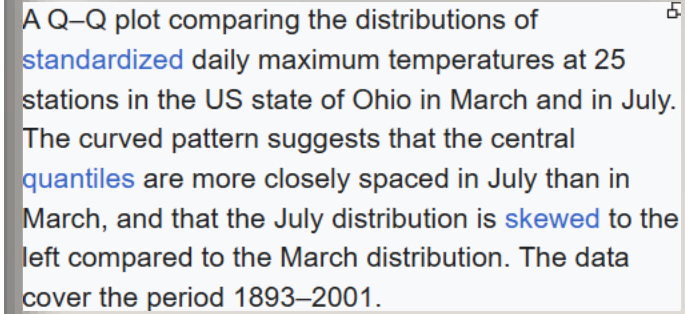



A normal Q–Q plot comparing randomly generated, independent standard normal data on the vertical axis to a standard normal population on the horizontal axis. The linearity of the points suggests that the data are normally distributed.



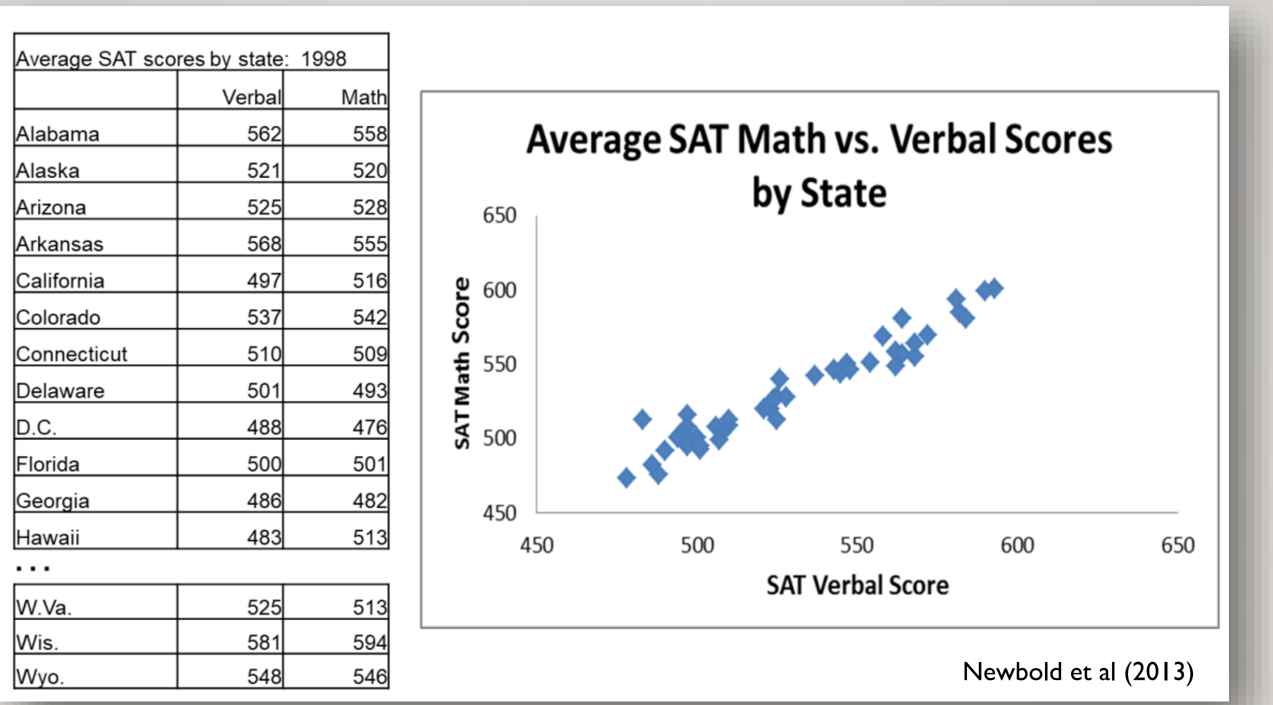

A normal Q–Q plot of randomly generated, independent standard exponential data, $(X \sim \operatorname{Exp}(1))$. This Q–Q plot compares a sample of data on the vertical axis to a statistical population on the horizontal axis. The points follow a strongly nonlinear pattern, suggesting that the data are not distributed as a standard normal $(X \sim \mathrm{N}(0,1))$. The offset between the line and the points suggests that the mean of the data is not 0. The median of the points can be determined to be near 0.7.




A Q–Q plot comparing the distributions of standardized daily maximum temperatures at 25 stations in the US state of Ohio in March and in July. The curved pattern suggests that the central quantiles are more closely spaced in July than in March, and that the July distribution is skewed to the left compared to the March distribution. The data cover the period 1893–2001.

Q–Q plot - Wikipedia

SCATTER DIAGRAM /SCATTERPLOT EXAMPLE



Average SAT scores by state: 1998

	Verbal	Math
Alabama	562	558
Alaska	521	520
Arizona	525	528
Arkansas	568	555
California	497	516
Colorado	537	542
Connecticut	510	509
Delaware	501	493
D.C.	488	476
Florida	500	501
Georgia	486	482
Hawaii	483	513
. . .		
W.Va.	525	513
Wis.	581	594
Wyo.	548	546

Newbold et al (2013)

Scatter Diagrams are used for paired observations taken from two numerical variables.

One variable is measured on the vertical axis and the other variable is measured on the horizontal axis.

TIME PLOT

Definition

A Time Plot is a graphical representation that displays the **temporal sequence or ordering of events**. In statistics, it is often used to **visualize the occurrence or timing of observations**, helping to detect trends, patterns, or irregularities over time.

Purpose

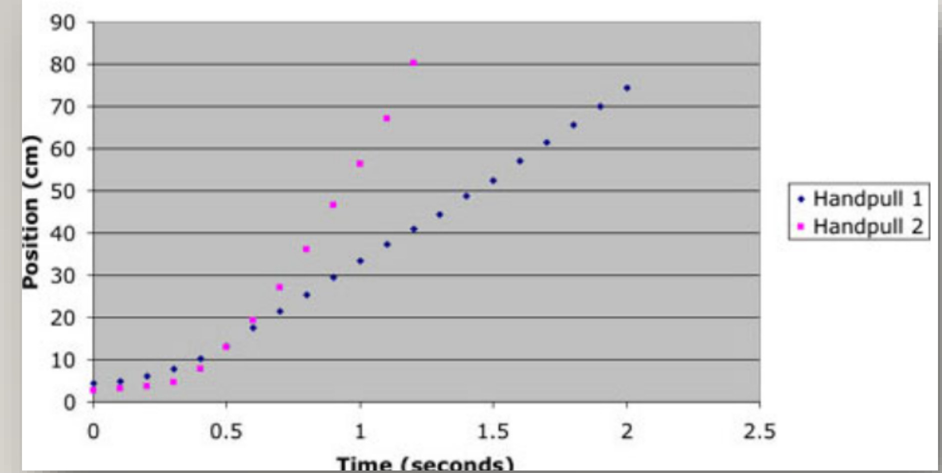
- To **show data or events in chronological order**.
- To **identify patterns, trends, or anomalies** in the timing of events.
- To **support descriptive analysis** of time-related data.

Key Features

- Time or sequence is represented on one axis (usually the x-axis).
- Observations or events are plotted on the y-axis.
- Useful in **quality control, survival analysis, and event studies**.

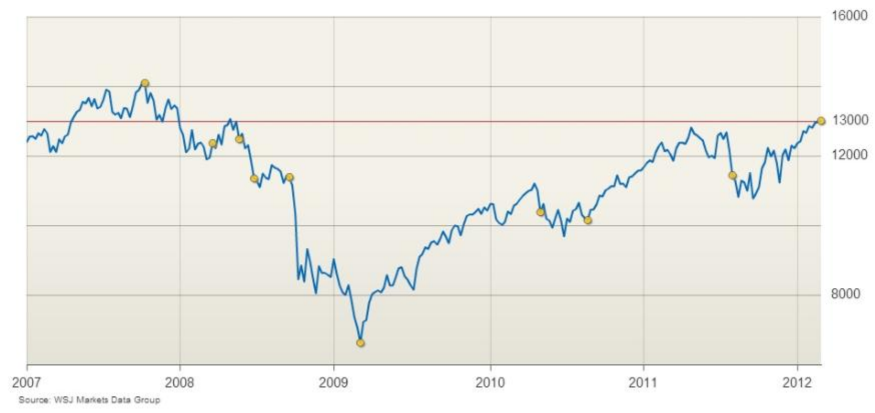
Silvestre (2007)

TIME PLOT EXAMPLES



The following graph shows a physics-related time plot with the position vs. time for two spark tapes pulled through a spark timer at different constant speeds.

Timeplot / Time Series: Definition, Examples & Analysis - Statistics How To



Source: WSJ Markets Data Group

Image: Wall Street Journal

While a time plot can resemble a scatter plot, with a series of dots, you will often see these plots with the dots connected, especially in financial publications like The Wall Street Journal.

THANKS!

Questions?