ESTATÍSTICA I

Licenciatura em Gestão do Desporto

2nd 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/

LECTURE 1: FUNDAMENTAL CONCEPTS OF STATISTICS

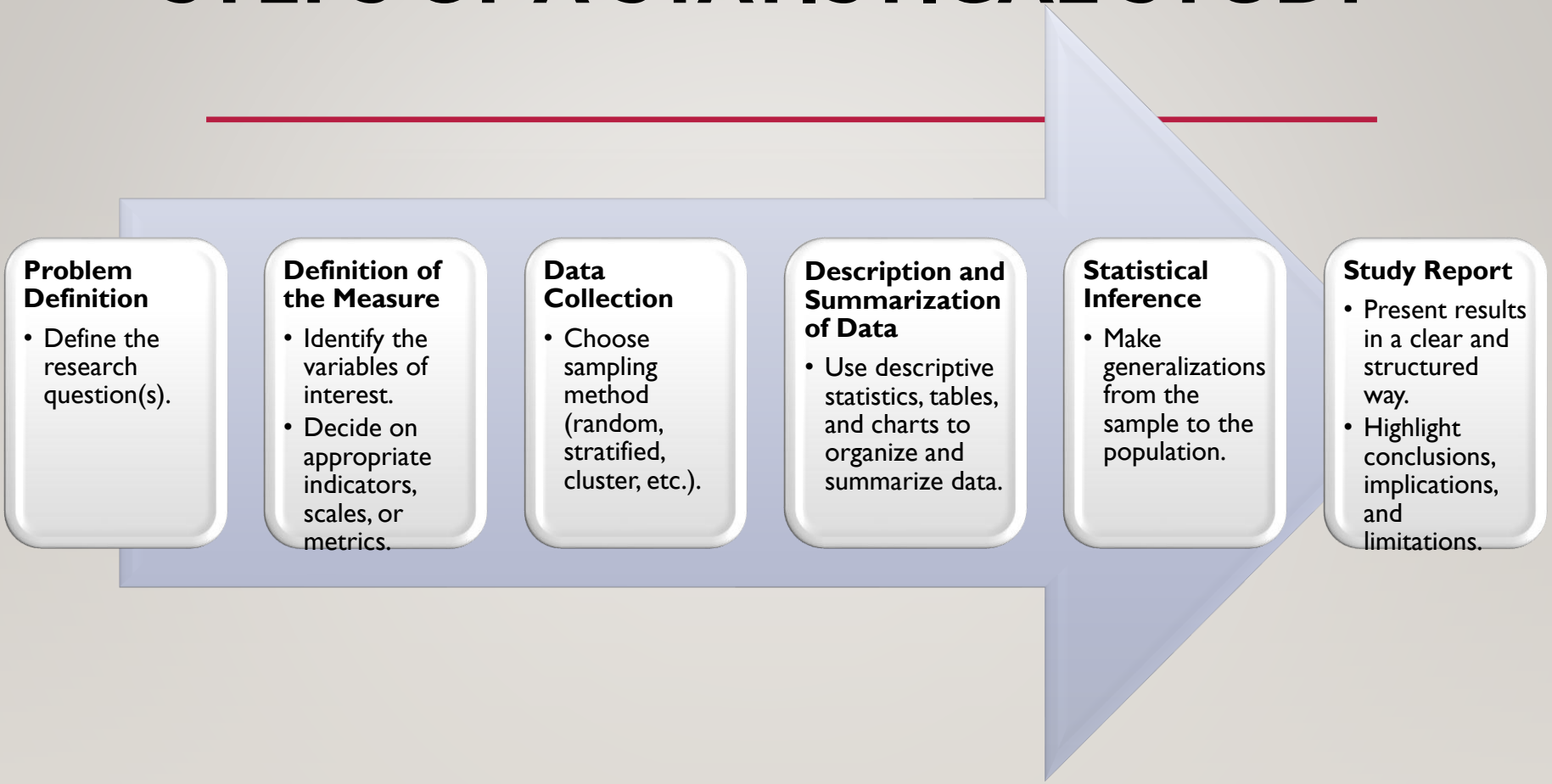
WHAT IS STATISTICS?

- The science of collecting, organizing, analysing, and interpreting data.

- Applications: **Economics, Management**, Healthcare, Social Sciences, Engineering, and more.

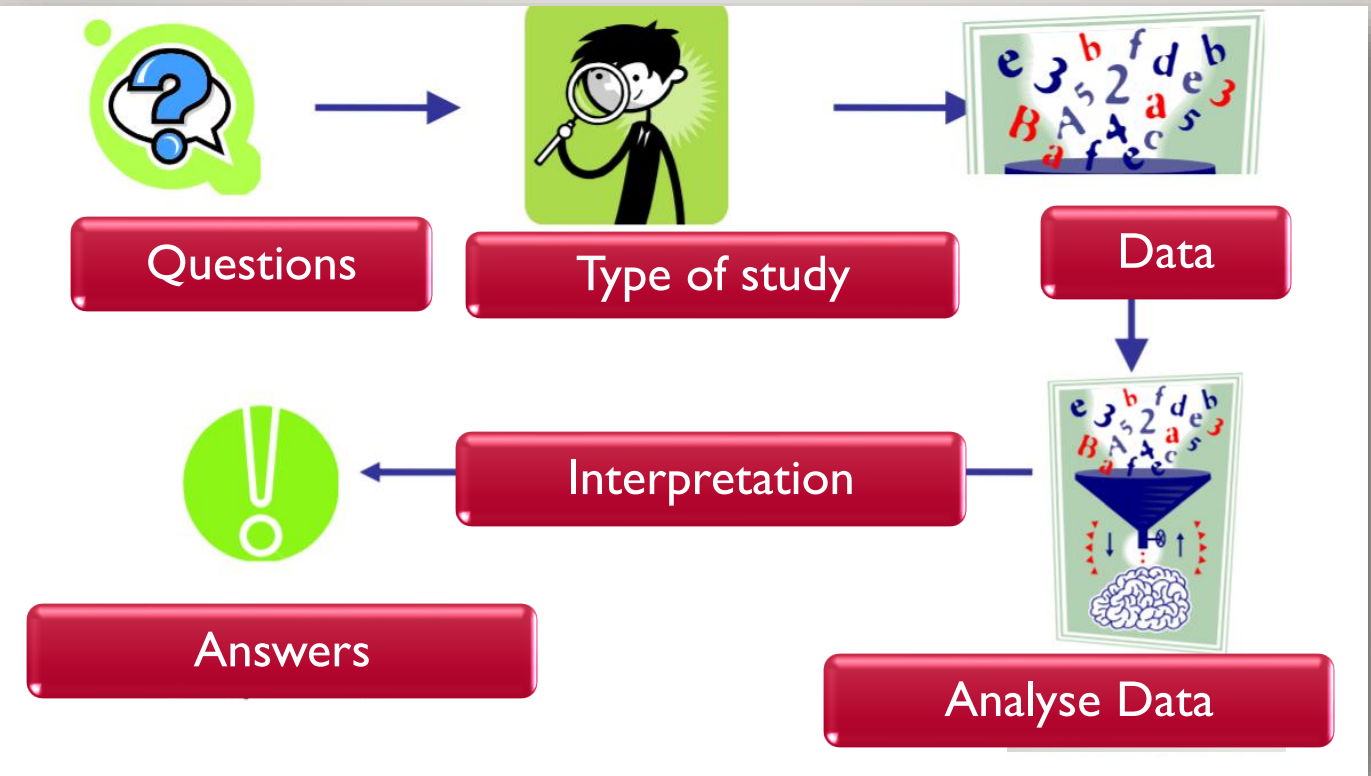
- Purpose in **Management and Economics**: support decision-making under uncertainly, identify patterns, and predict trends.

STEPS OF A STATISTICAL STUDY



Purpose: Transform raw data into meaningful information.

STEPS OF A STATISTICAL STUDY: VISUAL REPRESENTATION



MEANINGS OF "STATISTICS"

Scientific Discipline

Measure:

Numerical summaries that describe characteristics of a dataset.

Examples: mean, variance, and percentages.

Data:

Synonym for numerical information in specific areas.

Examples: heath statistics, industrial statistics, and employment statistics.

POPULATION AND SAMPLE

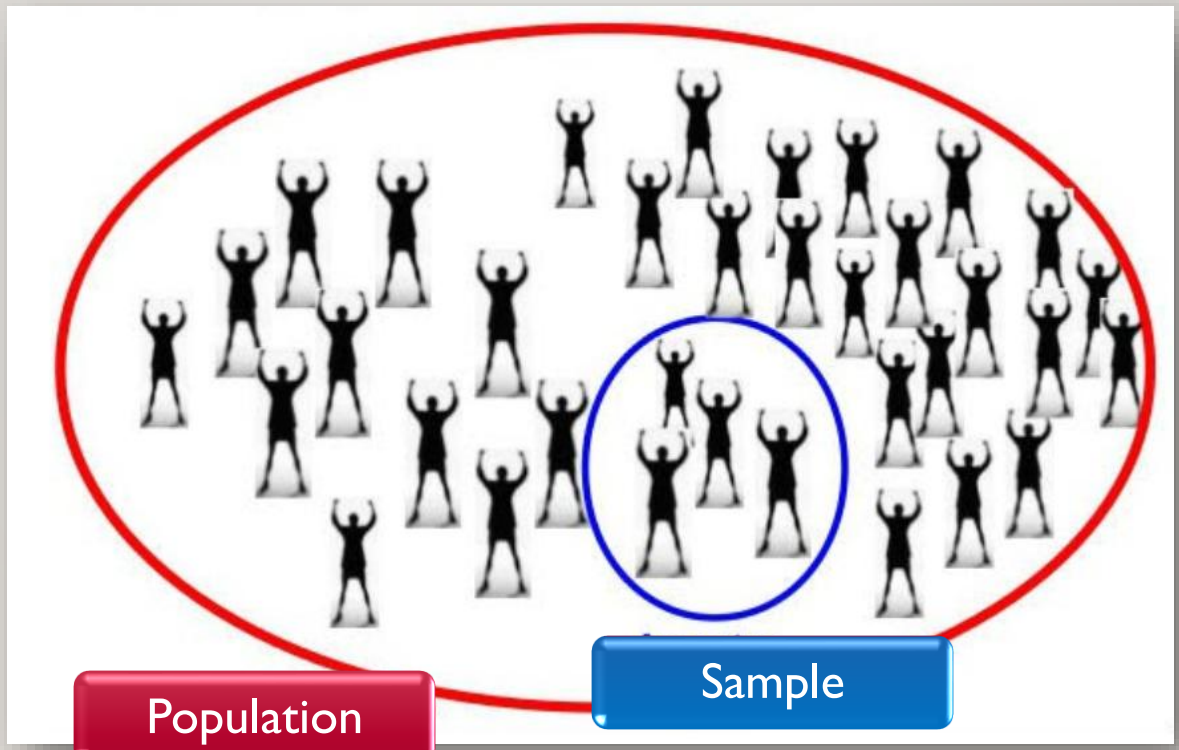
Population/universe: All elements of interest in a study or research.

Types of Population: real vs hypothetical; finite vs infinite.

Sample: A representative subset of a population, used to draw conclusions about the whole.

Example: all customers of a company (population) vs 50 selected customers (sample).

POPULATION AND SAMPLE: VISUAL REPRESENTATION



https://sites.google.com/site/estatisticabasicacc/conteudo/

OBJECTIVES OF STATISTICAL ANALYSIS

Data summarization and reduction

- Organize and condense large datasets into understandable forms
- **Examples**: frequency tables, charts, and summary measures (mean, median, standard deviation)

Inference to other datasets

- Make predictions or generalizations about a population based on sample data
- **Example**: estimating total sales based on a sample of transactions

Identification of relationships between datasets

- Discover correlations, associations, or causal links between variables
- **Example**: analyzing the relationship between marketing spend and customer acquisition

Dimensionality reduction

- Simplify data by reducing the number of variables while retaining essential information
- **Example:** principal component analysis in multivariate data

Classification and discrimination

- Assign data points to categories or groups based on characteristics
- **Example**: categorizing customers into segments (loyal, occasional, new)

Data clustering

- Group similar data points together to identify patterns or natural groupings
- **Example**: grouping products based on sales patterns or customer behavior

TYPES OF STATISTICS

- **Descriptive Statistics**: organizes and summarizes data using tables, graphs, and measures.
- **Inferential Statistics**: draws conclusions about a population from a sample through estimation and hypothesis testing.
- **Example**: average revenue (descriptive) vs sales forecast (inferential).

IMPORTANCE OF SAMPLING STUDIES

Sampling: More efficient, less costly, and less time-consuming than surveying the entire population.

Census: Covers entire population, but is expensive, time-consuming, and often impractical.

Example: Surveying 1,000 households from a city (sampling) vs all households in the entire country (census).

Types of studies: census vs sampling – advantages and disadvantages

DATA COLLECTION

Sources: surveys, interviews, administrative databases, sensors, and company records.

Processes: coding, recording, and validating data.

Data Coding: transforming answers into numerical or categorical codes.
Data Recording: storing information systematically (Excel, SPSS, SQL).
Data Validation: checking consistency, completeness, and accuracy before analysis.

TYPES OF DATA

1. Cross-Sectional Data

Observations for multiple units collected at **one point in time** (one or more variables).

Example: survey of customer satisfaction across 100 stores in January 2025.

2. Time Series Data

Observations collected **over time** for a single unit (one or more variables).

Example: monthly sales revenue of a company from January 2020 to December 2024.

3. Panel Data (Longitudinal Data)

Combines **cross-sectional and time series data.** Observations for multiple units over multiple periods.

Example: annual income of 500 households from 2018 to 2024.

STATISTICAL UNIT, PARAMETER, AND STATISTICS

STATISTICAL UNIT: Each element of the population or sample.

Sample: $(x_1, x_2, \ldots, x_n)$

PARAMETERS: Characteristics of the population.

Example: population mean μ and population standard deviation σ.

STATISTICS: Measures calculated from a sample.

Example: sample mean $\bar{x}$ and sample standard deviation s.

RANDOM EXPERIMENT

A **random experiment** is a process or action whose outcome **cannot be predicted with certainty** in advance, even under identical conditions.

Characteristics:

- Multiple possible outcomes
- Repeatable under same conditions
- Set of all possible outcomes = **Sample Space**

Examples:

- Roll a Die → Outcomes: 1, 2, 3, 4, 5, 6
- Coin Toss (Coin Flip) → Outcomes: Heads or Tails
- Selecting a Random Employee → Outcome: Age of selected person

VARIABLES

Variable: A characteristic or property **observed in a random experiment.**

Random variable: Theoretical concept that assigns values to the outcomes of a random experiment.

Examples: weekly sales and number of defective items.

Empirical variable: Observed in practice, based on collected data.

Examples: age, weight, and number of products sold.

Random Variable (X) vs **Empirical Variable (x)**

Sample: $(x_1, x_2, \ldots, x_n)$

LEVELS OF MEASUREMENT

Nominal

- Categories without order.
- **Examples:** gender and nationality.

Ordinal

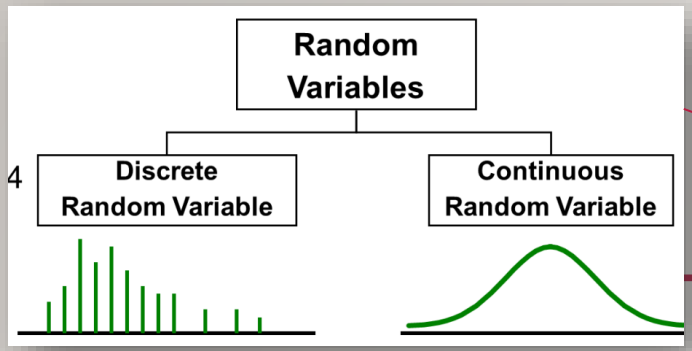
- Ordered categories.
- **Example:** education level.

Interval

- Differences between values are meaningful, but there is **no absolute zero**.
- **Example:** temperature in Celsius °C (0 °C ≠ no temperature).

Ratio

- An **absolute zero exists**, and ratios are meaningful.
- **Examples:** income, age, and weight. (0 kg = no weight).



CLASSIFICATION OF VARIABLES

$$Y = ax + b$$

Number of Values

Discrete Variables:

- Take finite or countable values.
- **Examples:** Number of children and number of defective items.

Continuous Variables:

- Can take infinite values within a range.
- **Examples:** Height and weight.

Explanatory Orientation

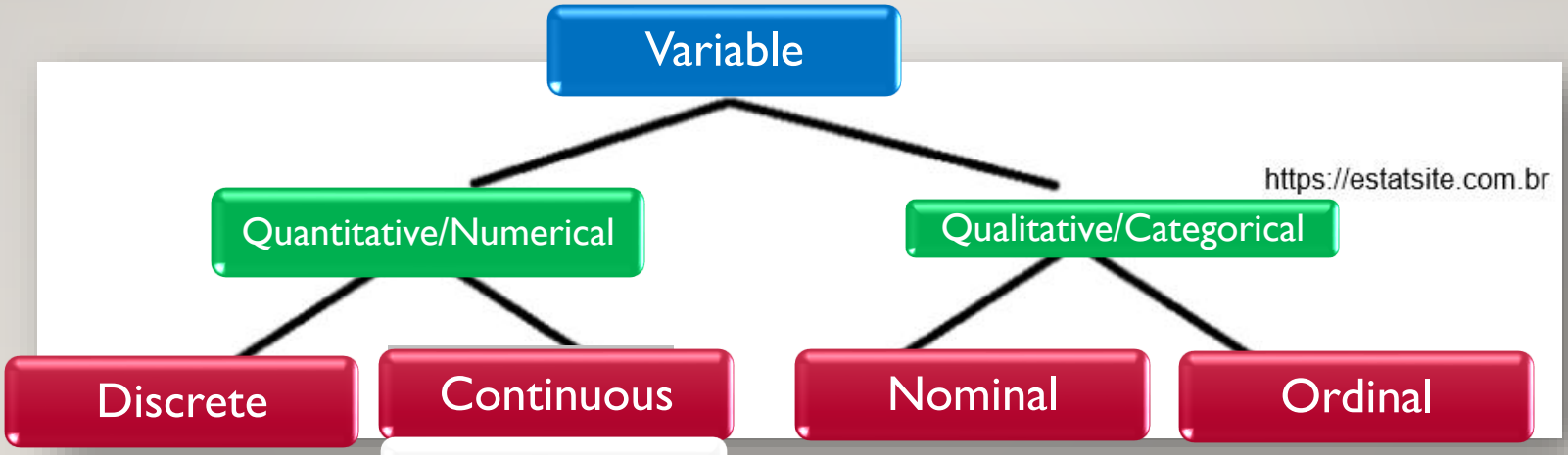
Explanatory Variable (Independent / Predictor Variable)

- A variable used to explain or predict changes in another variable.
- Represents the potential cause, influence, or input.
- Usually placed on the **x-axis** in graphs.
- **Examples:** marketing expenditure, study hours, and product price.

Explained Variable (Dependent / Response Variable)

- The variable whose variation we want to understand or predict.
- Represents the effect, outcome, or result.
- Usually placed on the **y-axis** in graphs.
- **Examples:** sales revenue, exam score, and demand for a product.

CLASSIFICATION OF VARIABLES: VISUAL REPRESENTATION



Example: Number of children

Interval or Ratio

Example: Height and weight

Example: Gender

Example: Satisfaction level

EXERCISE 1.1

1.1 A mortgage company randomly samples accounts of their time-share customers. State whether each of the following variables is categorical or numerical. If categorical, give the level of measurement. If numerical, is it discrete or continuous?

a. The original purchase price of a customer's time-share unit
b. The state (or country) of residence of a time-share owner
c. A time-share owner's satisfaction level with the maintenance of the unit purchased (1: very dissatisfied to 5: very satisfied)
d. The number of times a customer's payment was late

Newbold et al (2013)

EXERCISE 1.1: SOLUTION

✅ Answers:

a. **Numerical, Continuous (Ratio)** → Purchase price, any value, true zero.

b. **Categorical (Nominal)** → State/country, categories with no order.

c. **Categorical (Ordinal)** → Satisfaction scale 1–5, ordered categories.

d. **Numerical, Discrete (Ratio)** → Number of late payments, count values.

EXERCISE 1.2

1.2 Visitors to a supermarket in Singapore were asked to complete a customer service survey. Are the answers to the following survey questions categorical or numerical? If an answer is categorical, give the level of measurement. If an answer is numerical, is it discrete or continuous?

a. Have you visited this store before?
b. How would you rate the level of customer service you received today on a scale from 1 (very poor) to 5 (very good)?
c. How much money did you spend in the store today?

Newbold et al (2013)

EXERCISE 1.2: SOLUTION

✅ Answers:

a. **Categorical (Nominal)** → Yes/No, no natural order.

b. **Categorical (Ordinal)** → Rating scale 1–5, ordered categories.

c. **Numerical, Continuous (Ratio)** → Money spent, decimal values possible, true zero.

EXERCISE 1.5

1.5 A number of questions were posed to a random sample of visitors to a London tourist information center. For each question below, describe the type of data obtained.

a. Are you staying overnight in London?
b. How many times have you visited London previously?
c. Which of the following attractions have you visited?

Tower of London
Buckingham Palace
Big Ben
Covent Garden
Westminster Abbey

d. How likely are you to visit London again in the next 12 months: (1) unlikely, (2) likely, (3) very likely?

Newbold et al (2013)

EXERCISE 1.5: SOLUTION

✅ Answers:

a. Are you staying overnight in London?

- **Categorical (Nominal)** → Yes/No, no inherent order.

b. How many times have you visited London previously?

- **Numerical, Discrete (Ratio)** → Count of visits, zero possible, only integer values.

c. Which of the following attractions have you visited?

- **Categorical (Nominal, Multiple Response)** → Each attraction is a yes/no question; categories with no order.

d. How likely are you to visit London again in the next 12 months?

- **Categorical (Ordinal)** → Likert-type scale (e.g., very unlikely → very likely), ordered categories.

FORMAL REPRESENTATION OF DATA

$(x_1, x_2, x_3, ..., x_n)$ ou $x_i (i = 1, 2, ..., n)$

n observations of one variable

n observations of two variables

$[(x_1, y_1), (x_2, y_2), (x_3, y_3), ..., (x_n, y_n)]$

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

n observations of p variables

Contingency tables

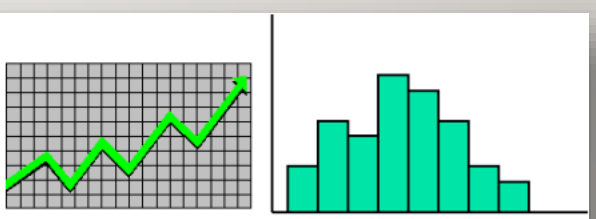
Movies Attended	Gender		Total
	Men	Women	
0	20	40	60
1	40	30	70
2 or more	10	10	20
Total	70	80	150

E.g. A survey of 150 adults classified each as to gender and the number of movies attended last month. Each respondent is classified according to two criteria—the number of movies attended and gender.

LECTURE 1: DESCRIPTIVE DATA ANALYSIS

DESCRIPTIVE STATISTICS

- Present data
  - e.g., Tables and graphs



- Summarize data
  - e.g., Sample mean $= \frac{\sum X_i}{n}$

Newbold et al (2013)

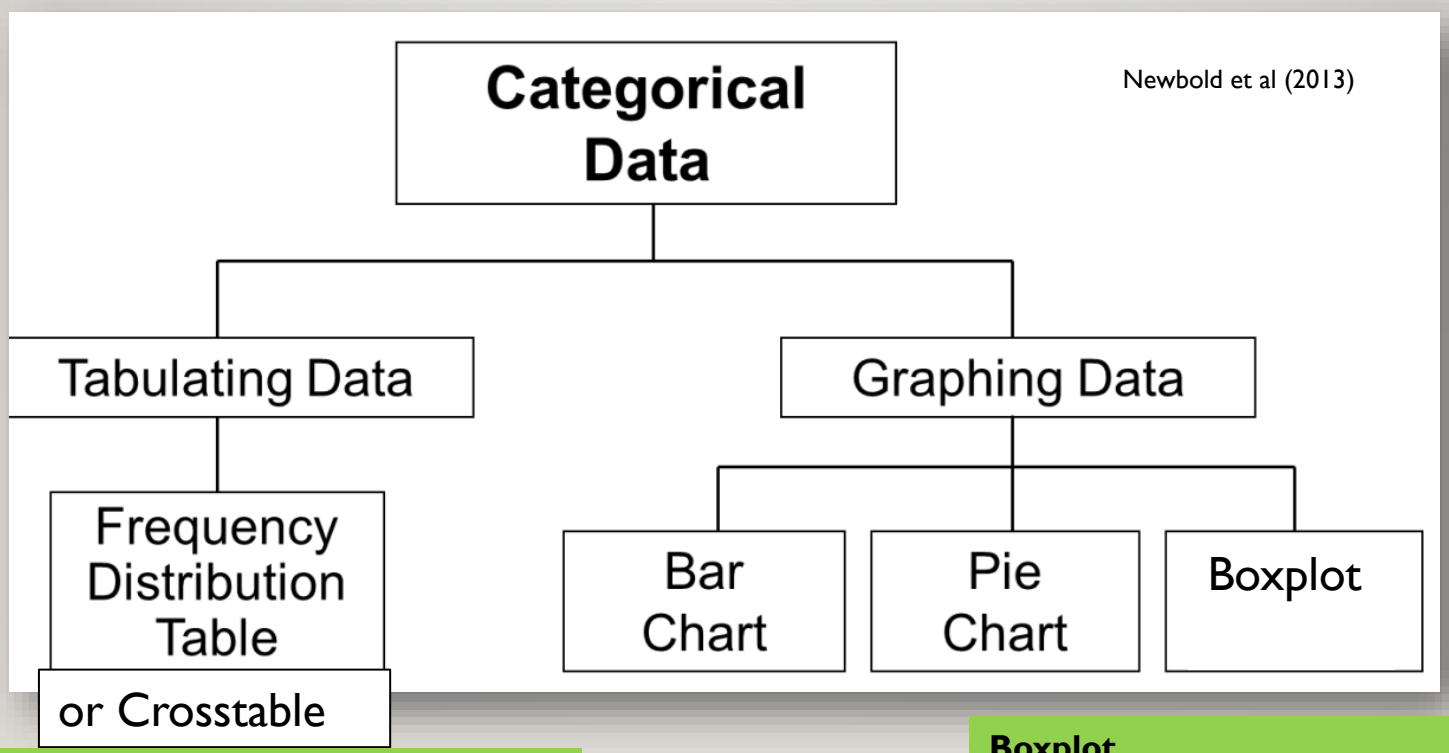
DATA REPRESENTATION

- **Tables**: frequency distributions.

- **Graphs**: bar chart, pie chart, histogram, boxplot, line chart, etc.

- **Choice depends on**: type of variable & analysis purpose.

TABLES AND GRAPHS FOR CATEGORICAL VARIABLES



- **Frequency Table**: shows the distribution of one categorical variable. **Example:** Gender (Male: 40 and Female: 60).
- **Contingency Table / Crosstabulation**: shows the joint distribution of two categorical variables. **Example:** Gender × Education level.

Boxplot

- Not suitable for **nominal variables**, because they **lack order**.
- Can be used for **numerical variables** or **ordinal variables with five or more categories**.

FREQUENCY DISTRIBUTION TABLE EXAMPLE

Summarize data by category

Example: Hospital Patients by Unit

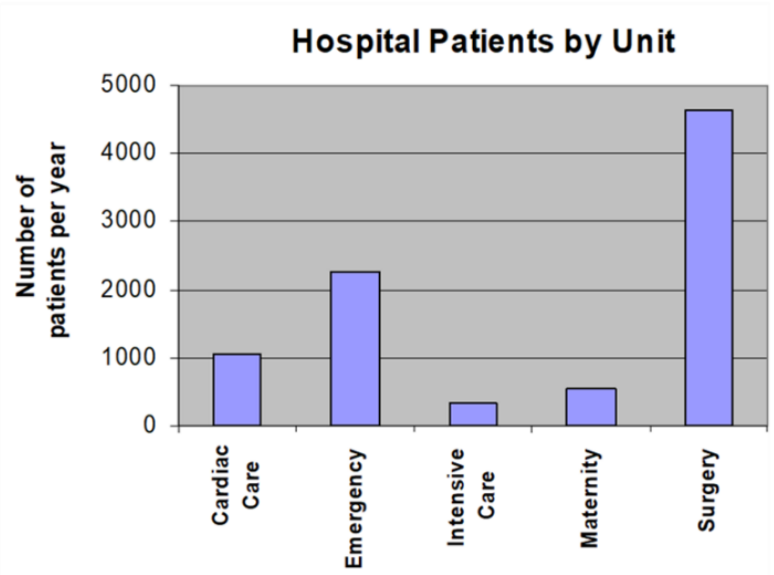
Hospital Unit	Number of Patients	Percent (rounded)
Cardiac Care	1,052	11.93
Emergency	2,245	25.46
Intensive Care	340	3.86
Maternity	552	6.26
Surgery	4,630	52.50
Total:	8,819	100.0

(Variables are categorical)

Newbold et al (2013)

BAR CHART EXAMPLE

Hospital Unit	Number of Patients
Cardiac Care	1,052
Emergency	2,245
Intensive Care	340
Maternity	552
Surgery	4,630

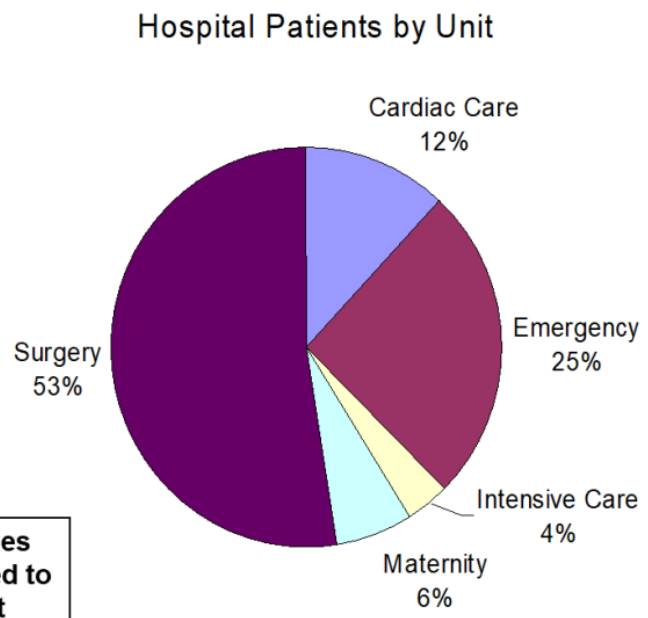


Newbold et al (2013)

PIE CHART EXAMPLE

Hospital Unit	Number of Patients	% of Total
Cardiac Care	1,052	11.93
Emergency	2,245	25.46
Intensive Care	340	3.86
Maternity	552	6.26
Surgery	4,630	52.50

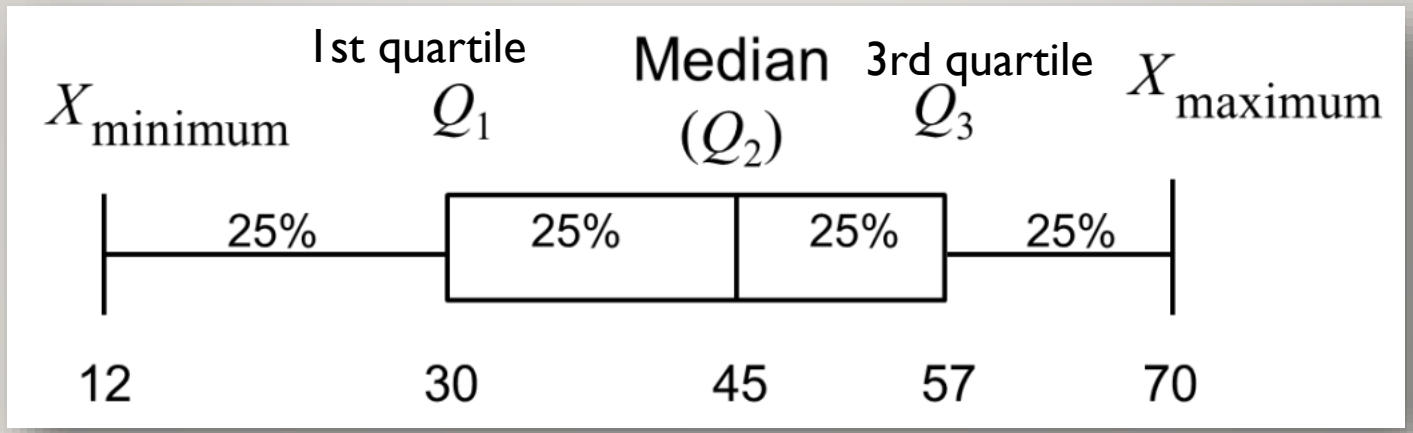
(Percentages are rounded to the nearest percent)



Newbold et al (2013)

- **Bar charts and Pie charts** are often used for qualitative (categorical) data.
- Height of bar or size of pie slice shows the frequency or percentage for each category.

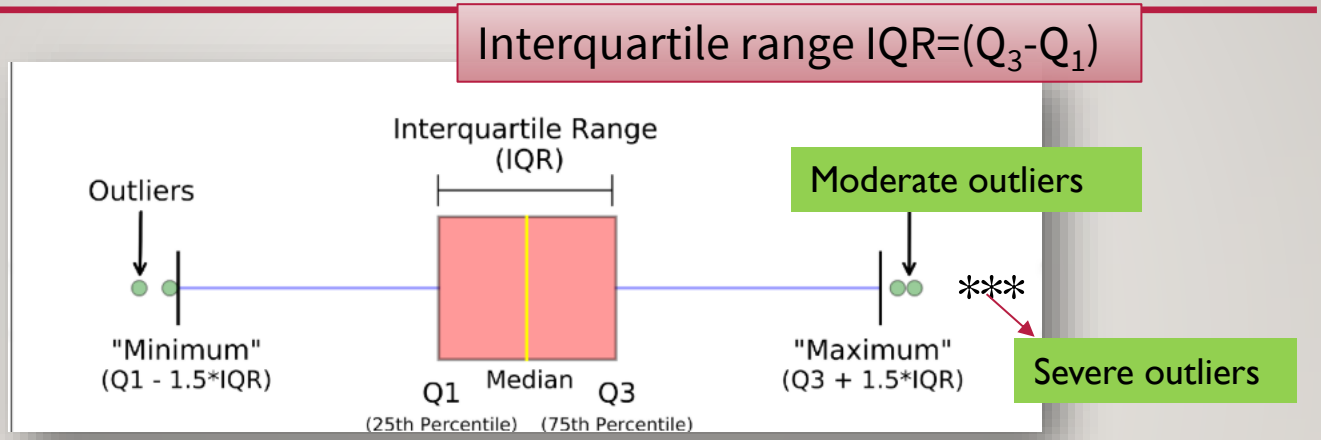
BOX-AND-WHISKER PLOT/ BOXPLOT EXAMPLE



Newbold et al (2013)

The plot can be oriented horizontally or vertically.

BOXPLOT AND OUTLIERS



Moderate outliers (marked with a circle)

$(Q_1\text{-}1{,}5 \times IQR; Q_3\text{+}1{,}5 \times IQR)$ Inner fences

Severe outliers (marked with an asterisk)

$(Q_1\text{-}3 \times IQR; Q_3\text{+}3 \times IQR)$ Outer fences

CROSSTABLE EXAMPLE

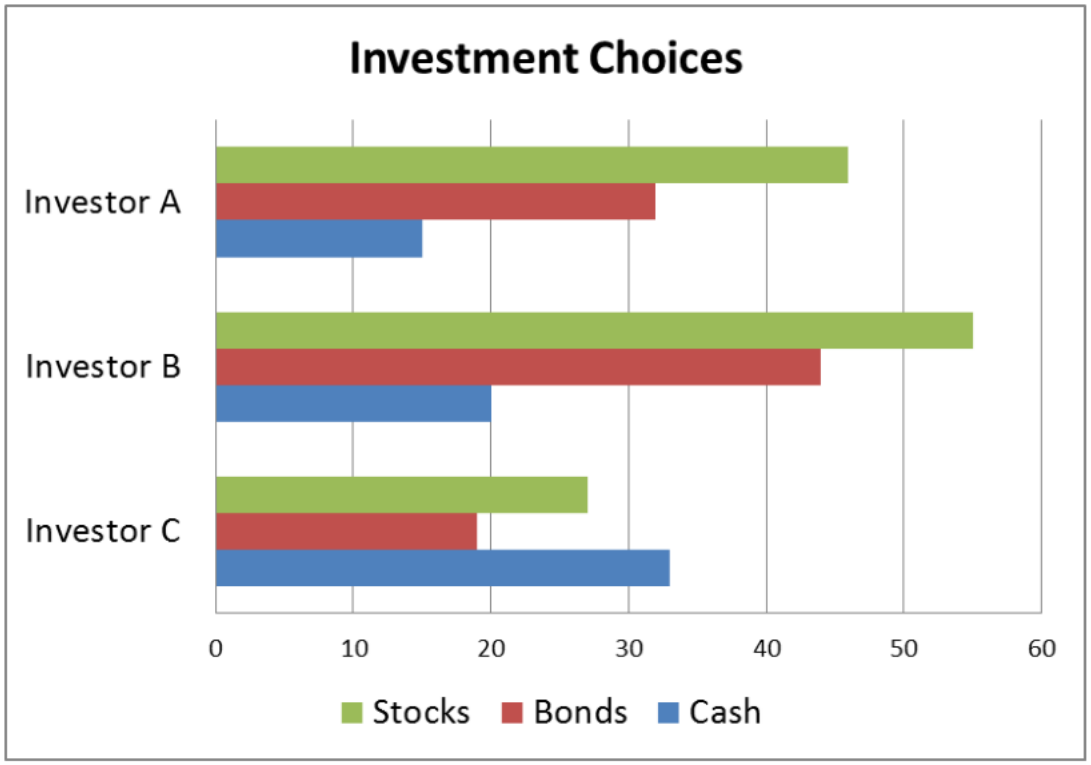
3×3 Cross Table for Investment Choices by Investor (values in $1000's)

Investment Category	Investor A	Investor B	Investor C	Total
Stocks	46	55	27	**128**
Bonds	32	44	19	**95**
Cash	15	20	33	**68**
Total	**93**	**119**	**79**	**291**

Newbold et al (2013)

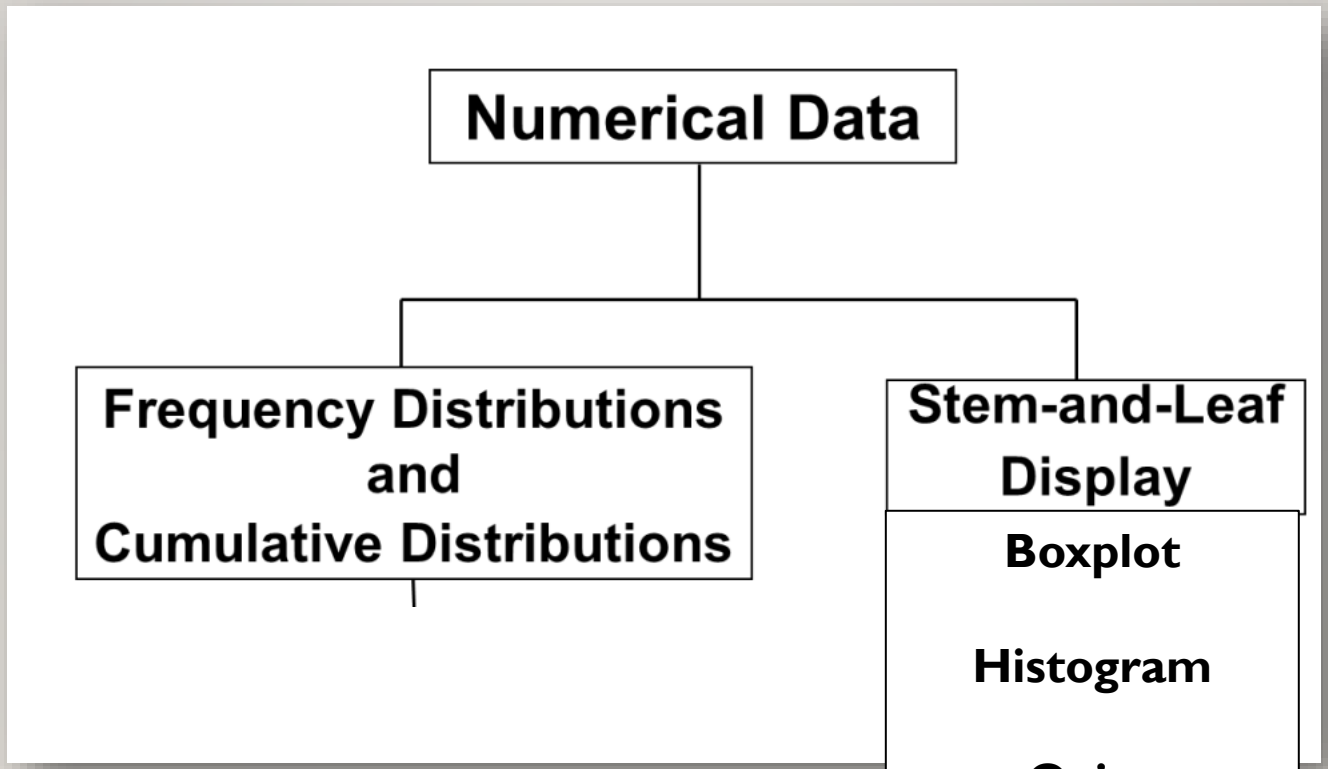
GRAPHING MULTIVARIATE CATEGORICAL DATA

Side by side horizontal bar chart



Newbold et al (2013)

GRAPHS TO DESCRIBE NUMERICAL VARIABLES



Newbold et al (2013)

- **Univariate (one variable):** Stem-and-Leaf, Boxplot, Histogram, and Ogive
- **Bivariate (two variables):** Scatterplot

FREQUENCY DISTRIBUTION EXAMPLE

Data in ordered array:

12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

Interval	Frequency	Relative Frequency	Percentage
10 but less than 20	3	.15	15
20 but less than 30	6	.30	30
30 but less than 40	5	.25	25
40 but less than 50	4	.20	20
50 but less than 60	2	.10	10
Total	20	1.00	100

Newbold et al (2013)

"Square root rule for the number of classes: The number of classes (k) can be estimated as the square root of the number of observations (n):

$$k \approx \sqrt{n}$$

where n is the total number of data points."

n = 20 (sample size)
k = $\sqrt{20}$ =
4.47 ~ 5 (number of classes)

CLASS INTERVALS

- Each class grouping has the same width
- Determine the width of each interval by

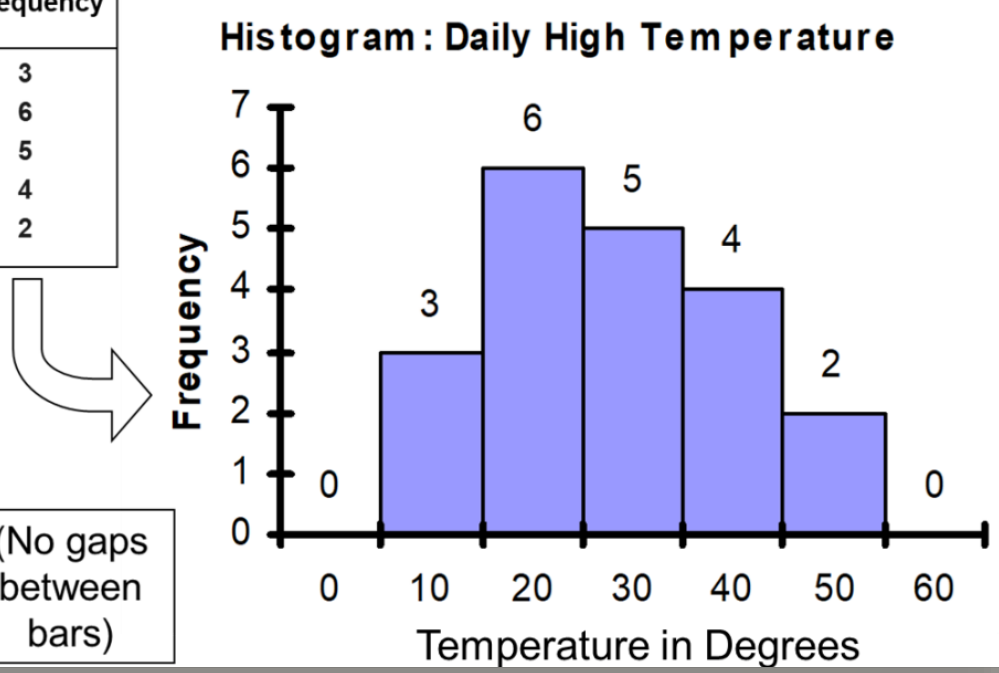
$$w = \text{interval width} = \frac{\text{largest number} - \text{smallest number}}{\text{number of desired intervals}}$$

- Use at least 5 but no more than 15-20 intervals
- Intervals never overlap
- Round up the interval width to get desirable interval endpoints

Newbold et al (2013)

HISTOGRAM EXAMPLE

Interval	Frequency
10 but less than 20	3
20 but less than 30	6
30 but less than 40	5
40 but less than 50	4
50 but less than 60	2



(No gaps between bars)

Newbold et al (2013)

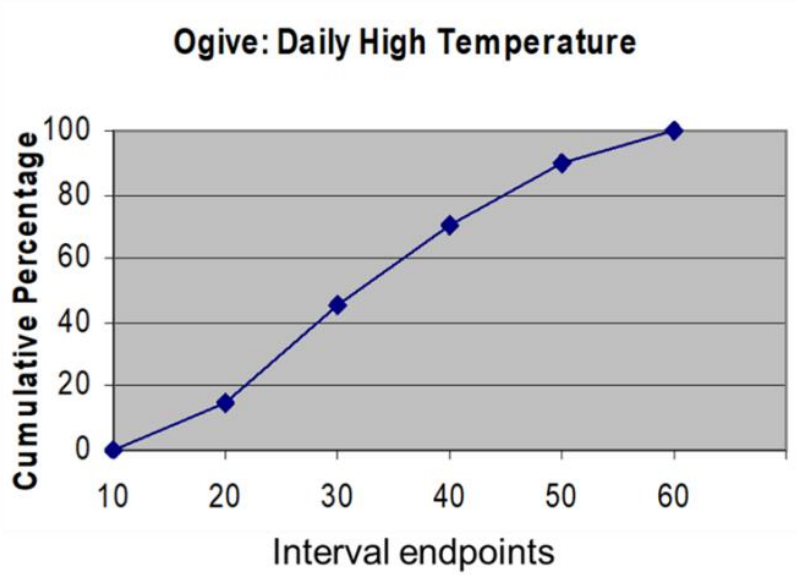
n = 20 (sample size)

k = $\sqrt{20} = 4.47 \sim 5$ (number of classes)

Width of each interval = (58 -12) / 5 = 9,2 ~ 10

THE OGIVE GRAPHING CUMULATIVE FREQUENCIES

Interval	Upper interval endpoint	Cumulative Percentage
Less than 10	10	0
10 but less than 20	20	15
20 but less than 30	30	45
30 but less than 40	40	70
40 but less than 50	50	90
50 but less than 60	60	100



Newbold et al (2013)

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

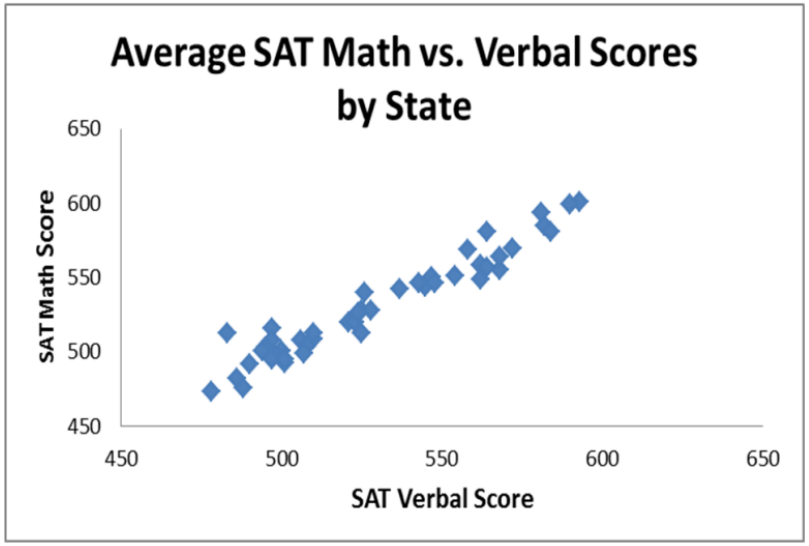
A simple way to see distribution details in a data set.

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

Newbold et al (2013)

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.

THANKS!

Questions?