STATISTICAL LABORATORY

Lecture 2: Exploratory Data Analysis

Yiu Lim LUI, Allen

Lisbon School of Economics & Management
Universidade de Lisboa

allenlui@iseg.ulisboa.pt

https://sites.google.com/view/allenlui

Applied Mathematics for Economics and Management
1st Year / 1st Semester – 2026/2027

Course Overview & Program

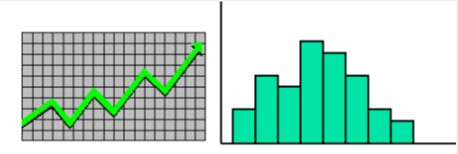
Course Program

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

Descriptive Statistics Overview

Descriptive Statistics Overview

Present data: e.g., Tables and graphs



Summarize data: e.g., Sample mean $\bar{x} = \frac{(x_1+x_2+...+x_n)}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$

Data Representation

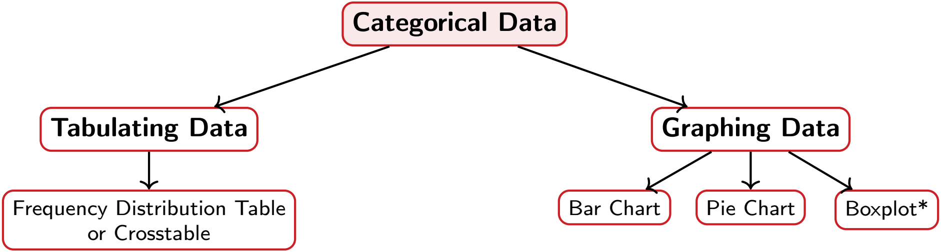
How do we represent data?

- **Tables:** Frequency distributions.
- **Graphs:** Bar chart, pie chart, histogram, boxplot, line chart, etc.
- **Choice depends on:**
    - Type of variable (Categorical vs. Numerical)
    - Purpose of the analysis

Reference

Newbold et al. (2013) *Statistics for Business and Economics*.

Tables and Graphs for Categorical Variables



Tabulating Data:

- **Frequency Table:** Shows distribution of 1 variable (e.g., Gender).
- **Contingency Table/Crosstabulation:** Joint distribution of 2 variables (e.g., Gender and Education level).

***Boxplot Note:**

- Not suitable for nominal variables (lack order).
- Can be used for ordinal variables with 5+ categories or numerical variables.

Frequency Distribution

Definition

A **frequency distribution** is a table used to organize data.

- The **left column** (classes or groups) includes all possible categories/responses.
- The **right column** lists the frequencies (number of observations) for each class.

Relative Frequency

A **relative frequency distribution** is obtained by dividing each frequency by the total number of observations (n) and multiplying by 100%:

$$\text{Relative Frequency} = \frac{\text{Absolute Frequency}}{n}$$

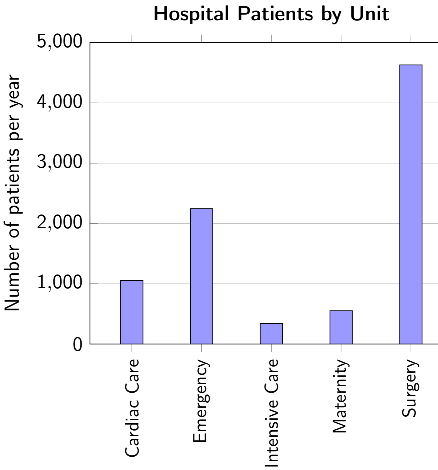
Frequency Distribution Example: Hospital Patients by Unit

Table: Summarize Data by Category

Hospital Unit	Absolute Freq.	Percent (%)
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**	**100.00**

BAR CHART EXAMPLE

Hospital Unit	Number of Patients	% of Total (Rounded)
Cardiac Care	1,052	12%
Emergency	2,245	25%
Intensive Care	340	4%
Maternity	552	6%
Surgery	4,630	53%

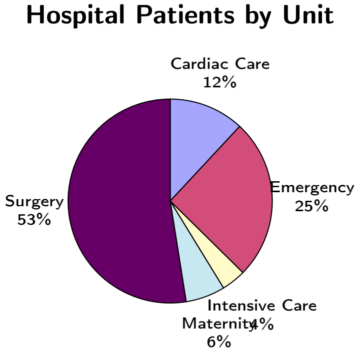


Newbold et al. (2013)

PIE CHART EXAMPLE (SUPPLEMENTARY)

Hospital Unit	Number of Patients	% of Total
Cardiac Care	1,052	12%
Emergency	2,245	25%
Intensive Care	340	4 %
Maternity	552	6 %
Surgery	4,630	53%

(Percentages are rounded to the nearest percent)

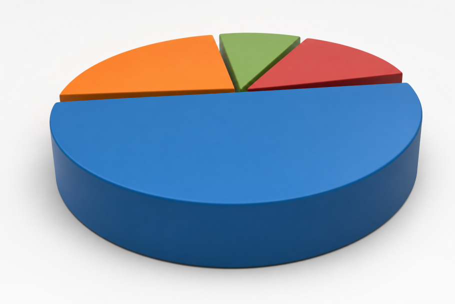


- **Bar charts and Pie charts** are often used for qualitative (categorical) data.
- Height of bar or size of pie slice shows the frequency or percentage for each category.

Newbold et al. (2013)

Problem of 3D chart

- 3D Effects Distort Perspective

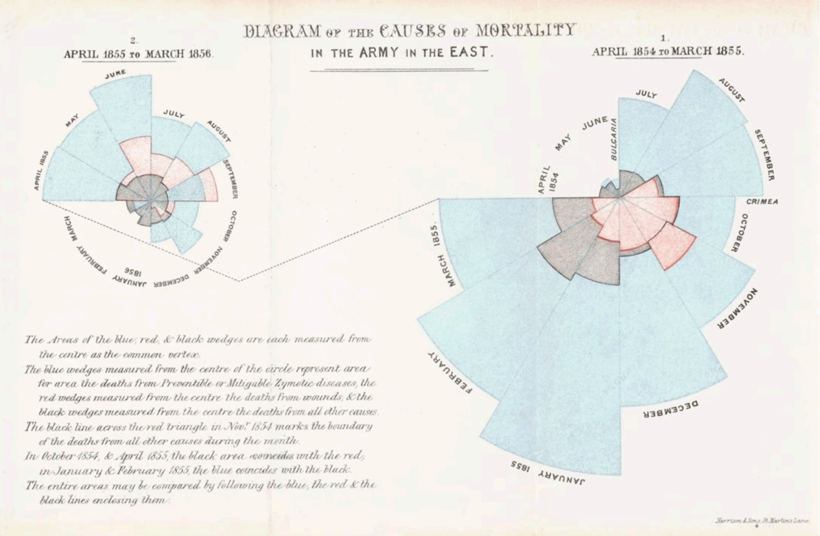


- A company uses a 3D pie chart to show market share. The leading brand appears enormous because its slice is placed in the foreground.
- In reality, the difference between competitors is only 5%.
- The Issue: The human eye is notoriously bad at accurately judging relative areas and 3D volumes compared to simple linear lengths (bars).

History: When Data Saves Lives (Florence Nightingale)

- The Crimean War (1854): More soldiers were dying from preventable hospital infections, poor sanitation, and disease than from actual battle wounds..
- Florence Nightingale didn't just treat patients-she meticulously collected data on mortality causes and invented early graphical diagrams (the Rose chart).
- The Lesson for Us: Raw numbers sitting in hospital ledgers don't change behavior. Visualizing data clearly exposes the truth.

History: When Data Saves Lives (Florence Nightingale)



Video: `youtube.com/watch?v=JZh8tUy_bnM`

Exercise 1: Transportation to Campus

Excercise

A survey was conducted among 30 university students about their preferred mode of transportation to campus:

Sample Data ($n = 30$):
Bus, Car, Walking, Bus, Metro, Car, Car, Bicycle, Walking, Bus, Bus, Car, Metro, Car, Walking, Walking, Bicycle, Bus, Car, Metro, Walking, Car, Bicycle, Car, Bus, Bus, Metro, Walking, Walking, Car.

Tasks

1. Construct a frequency table for *Mode of Transportation* (Absolute and Relative Frequencies).
2. Draw a bar chart to represent the distribution.

Exercise 1 Solution: Frequency Table

Table: Frequency Table for Mode of Transportation ($n = 30$)

Mode of Transportation	Absolute Frequency	Relative Frequency
Bus	7	0.233
Car	9	0.300
Walking	7	0.233
Metro / Underground	4	0.133
Bicycle	3	0.100
Total	**30**	**1.000**

$$\text{Relative Frequency} = \frac{\text{Absolute Frequency}}{n}$$

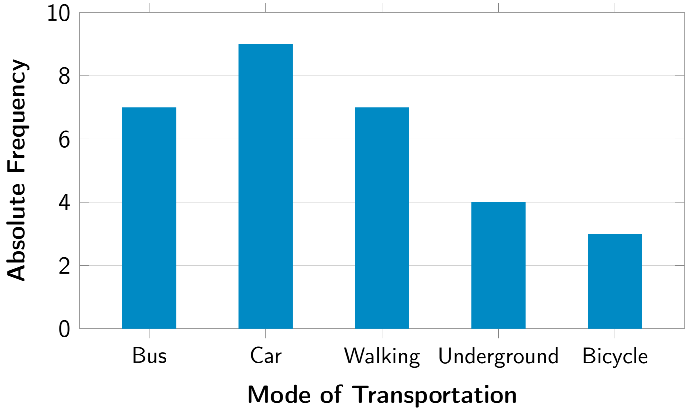
Exercise 1 Solution: Cumulative Frequencies (Supplementary)

Table: Frequency Table with Absolute and Relative Cumulative Frequencies

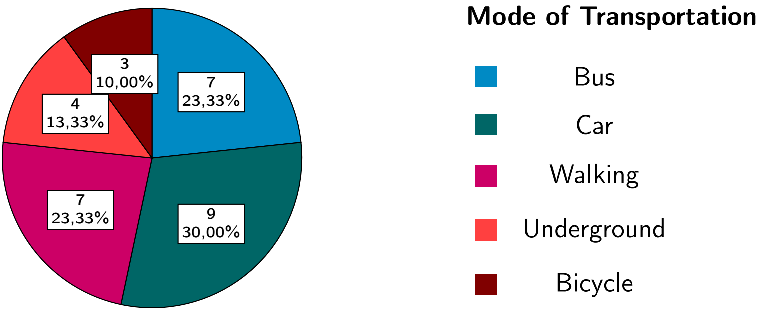
Mode	**Abs. Freq.**	**Rel. Freq.**	**Cum. Abs. Freq.**	**Cum. Rel. Freq.**
Bus	7	0.233	7	0.233
Car	9	0.300	16	0.533
Walking	7	0.233	23	0.766
Metro/Underground	4	0.133	27	0.899
Bicycle	3	0.100	30	$0.999 \approx 1.000$
Total	**30**	**1.000**	–	–

EXERCISE 2: SOLUTION

✓ **Answer: Bar Chart (Absolute Frequency)**



EXERCISE 2: SOLUTION



Graphs to Describe Numerical Variables

Numerical Data

- **Frequency Distributions and Cumulative Distributions**
- **Stem-and-Leaf Display**
  Boxplot
  Histogram
  Ogive
  Bar Chart
  Scatterplot

Continuous Variables

Discrete Variables

- **Univariate (one variable):** Stem-and-Leaf, Boxplot, Histogram, and Ogive (Continuous); Bar Chart (Discrete)
- **Bivariate (two variables):** Scatterplot

Newbold et al. (2013)

Histograms

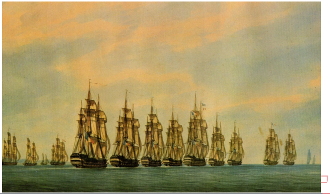
- Histograms provide a view of the **data density**. Higher bars represent where the data are relatively more common.
- Histograms are sometimes confused with bar charts. In a histogram, each bin is for a different range of values, so altogether the histogram illustrates the distribution of values.
- But in a bar chart, each bar is for a different category of observations (e.g., each bar might be for a different population), so altogether the bar chart can be used to compare different categories. Some authors recommend that bar charts always have gaps between the bars to clarify that they are not histograms.

Histograms

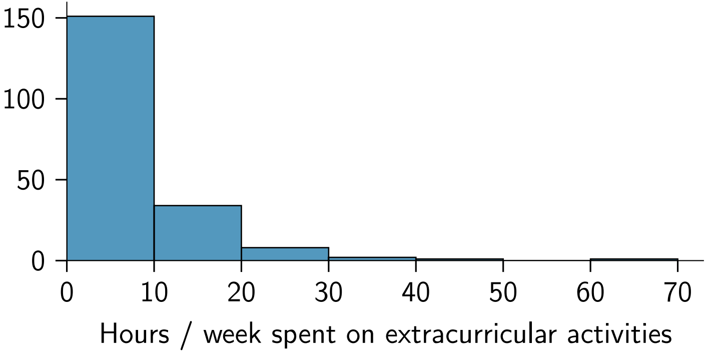
The word **histogram** was coined because the vertical bars look like a row of ship masts. English mathematician and statistician **Karl Pearson** created the term in **1891** (and officially published it in 1895). He combined two ancient Greek words to describe the graph's unique look:

- **Histos ($\iota\sigma\tau$ός):** Meaning a ship's mast, a loom, or anything set upright.
- **Gramma ($\gamma\rho\acute{\alpha}\mu\mu\alpha$):** Meaning a drawing, record, or something written.

Pearson noted that when you line up vertical frequency bars next to each other, the chart looks like a fleet of ships with their upright masts standing side by side.



- Histograms are especially convenient for describing the shape of the data distribution.
- The chosen class width can alter the story the histogram is telling.



RULES FOR BUILDING CLASSES (FREQUENCY TABLE & HISTOGRAM)

1. Number of classes (k):

Sturges' Rule: The number of classes k for a histogram is the **smallest integer** such that:

$$2^k \geq n$$

where n is the sample size.

or

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.

RULES FOR BUILDING CLASSES (FREQUENCY TABLE & HISTOGRAM)

2. Class width (h):

$$h = \frac{\max - \min}{k}$$

- **Always round class width, h, upward. (e.g.** $4.01 \rightarrow 5$, $4.5 \rightarrow 5$**)**
- **Classes must be inclusive and nonoverlapping.**
- Being inclusive here means that the classes, when taken together, must leave no data behind.
- Nonoverlapping means each number belongs to one class only.

RULES FOR BUILDING CLASSES (FREQUENCY TABLE & HISTOGRAM)

3. Where to Start the First Class? (3 options)

- **At the minimum observed value**
  - e.g. data from 12 to 87, $k = 8$, $h \approx 9.4$
  - Classes: $[12, 21.4), [21.4, 30.8), \ldots$
  - Notation:
    - Parenthesis "(" or ")" does not include the boundary value.
    - Square Bracket "[" and "]" include the boundary value.
- **Round down to a "nice" number (e.g. multiple of 5 or 10)**
  - Easier to read $\rightarrow$ start at 10, width $= 10$
  - Classes: $[10, 20), [20, 30), \ldots$
- **General rule**
  - Lower limit $\leq$ minimum
  - Upper limit $\geq$ maximum

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

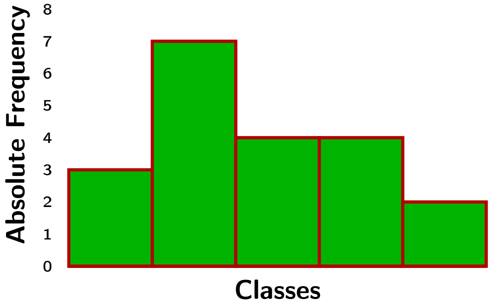
Newbold et al. (2013)

Frequency Distribution Example (Numerical Data)

Data Sample ($n = 20$): 12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

- (Sturges' Rule) $n = 20 \implies k = 5$ classes (since $2^4 = 16 < 20$ and $2^5 = 32 \geq 20$)
- Width $h = \frac{58-12}{5} = 9.2 \implies$ round up to 10

Class Interval	Class Midpoint	Absolute Frequency	Relative Frequency
$[10, 20]$	15	3	0.150
$(20, 30]$	25	7	0.350
$(30, 40]$	35	4	0.200
$(40, 50]$	45	4	0.200
$(50, 60]$	55	2	0.100
Total		**20**	**1.000**



Temperature in Degrees

- If all class widths are equal, histogram heights can simply represent absolute or relative frequencies.
- If class widths differ, using absolute frequency/class width ($\frac{abs.freq.}{class.width.}$) as the bar height ensures that the area of each bar correctly represents the absolute frequency.
- No Gaps between bars.

The Arbitrariness of Bin Widths

Rules of Thumb vs. Reality

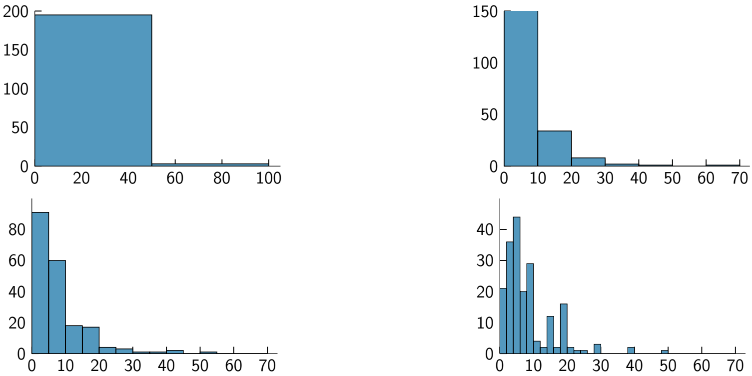
- Rules like Sturges' rule ($k \approx \sqrt{n}$ or $2^k \geq n$) are **merely rules of thumb**, not immutable laws of physics.
- Choosing $k = 5$ versus $k = 20$ can dramatically alter the story your data tells.

The Danger of Poor Bin Selection

- An overly coarse or overly fine bin width can completely hide a frequency distribution with two distinct peaks.
- *Example:* Smoothing out two distinct peaks (e.g., entry-level vs. senior staff salaries) into a single misleading bell curve.

Histograms

Which one(s) of these histograms are useful? Which reveal too much about the data? Which hide too much?



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.

Newbold et al. (2013)

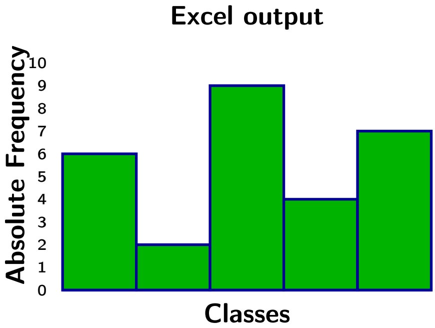
EXERCISE 1.32 A): SOLUTION

Answer:

Class	Midpoint	Absolute Frequency	Relative Frequency
$[10, 21]$	15.5	6	$6/28 \approx 0.214$
$(21, 32]$	27	2	$2/28 \approx 0.071$
$(32, 43]$	38	9	$9/28 \approx 0.321$
$(43, 54]$	49	4	$4/28 \approx 0.143$
$(54, 65]$	60	7	$7/28 \approx 0.250$

Sample size: $n = 28$
Number of classes: $k = 5$ (Sturges's Rule)
Width of each interval $= (\text{Max} - \text{Min})/k = (65 - 12)/5 = 10.6 \sim 11$

EXERCISE 1.32 B): SOLUTION



Exercise 1.36

Problem Statement

The following table shows the ages of competitors in a charity tennis event in Rome:

Age	Percent (%)
18–24	18.26
25–34	16.25
35–44	25.88
45–54	19.26
55+	20.35

Tasks

a. Construct a relative cumulative frequency distribution.
b. What percent of competitors were under the age of 35?
c. What percent of competitors were 45 or older?

Exercise 1.36 Solutions

a) Relative Cumulative Frequency Distribution

Age	Percent (%)	Cumulative Percent (%)
18–24	18.26	18.26
25–34	16.25	$18.26 + 16.25 = \mathbf{34.51}$
35–44	25.88	$34.51 + 25.88 = \mathbf{60.39}$
45–54	19.26	$60.39 + 19.26 = \mathbf{79.65}$
55+	20.35	$79.65 + 20.35 = \mathbf{100.00}$

b) Competitors Under 35

$$18.26\% + 16.25\% = \mathbf{34.51\%}$$

c) Competitors 45 or Older

$$19.26\% + 20.35\% = \mathbf{39.61\%}$$

THANKS!

Questions?