Lecture 6
DATA DESCRIPTION

FPT Education

FPT UNIVERSITY

Department of Mathematics

Võ Văn Nam

Contents

1. **Numerical summaries of data**

2. **Stem-and-leaf diagrams**

3. **Box-plots**

4. **Histograms**

1. Numerical summaries of data

- Sample mean
- Sample median
- Sample mode
- Sample variance
- Sample standard deviation
- Sample range

Sample mean

- The **sample mean**, often denoted as $\bar{x}$, is a measure of central tendency that represents the average value of a set of sample data.
- The formula for the sample mean is given by:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Example. Let us consider the weights of the eight observations collected from the prototype engine connectors: 12.6, 12.9, 13.4, 12.3, 13.6, 13.5, 12.6, and 13.1. *Find the sample mean.*

Sample median

- The **sample median** is a measure of central tendency that represents the middle value of a dataset when it is ordered from least to greatest.
- If the data set has an
  - even number of entries: median is the average of the two middle data entries.
  - odd number of entries: median is the middle data entry.

Example. The prices (in dollars) for a sample of round-trip flights from Chicago, Illinois to Cancun, Mexico are listed.

$$872 \quad 432 \quad 397 \quad 427 \quad 388 \quad 782 \quad 397$$

Find the median of the flight prices.

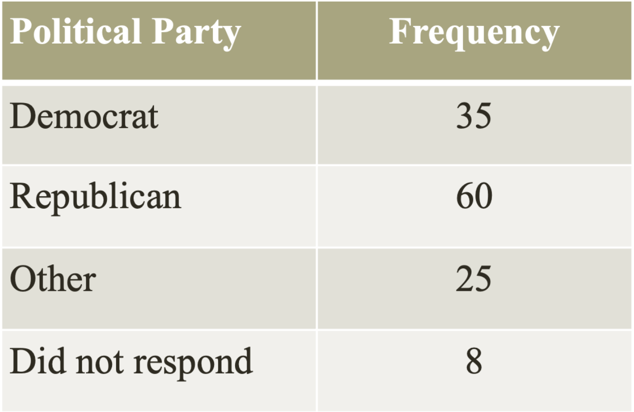
Sample mode

- The **sample mode** is a measure of central tendency that represents the most frequently occurring value in a dataset.
- A dataset can have:
  - No mode: All values occur with the same frequency.
  - Unimodal: One value occurs more frequently than others.
  - Bimodal: Two values occur with the highest frequency.
  - Multimodal: More than two values occur with the highest frequency.

Example

At a political debate a sample of audience members was asked to name the political party to which they belong. Their responses are shown in the table.

What is the mode of the responses?



Political Party	Frequency
Democrat	35
Republican	60
Other	25
Did not respond	8

Sample variance

- **Sample variance** is a measure of how spread out the values in a sample are around the sample mean.
- The formula for the sample variance is given by

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

- The sample standard deviation: s

Example. Let us consider the weights of the eight observations collected from the prototype engine connectors: 12, 13, 9, 12, 10 and 12.
Find the sample standard deviation.

Sample range

- **Sample range** is the difference between the maximum and minimum data entries in the set.
- The data must be quantitative.
- If the n observations in a sample are denoted by $x_1, x_2, ..., x_n$, the sample range is

$$r = \max_{1 \leq i \leq n} x_i - \min_{1 \leq i \leq n} x_i$$

Example. Let us consider the weights of the eight observations collected from the prototype engine connectors: 12, 13, 9, 12, 10 and 12.

Find the sample standard deviation.

2. Stem-and-leaf diagrams

Stem-and-leaf diagram

- A **stem-and-leaf diagram** is a good way to obtain an informative visual display of a data set where each number xi consists of at least two digits.

- To construct a stem-and-leaf diagram, use the following steps:

  1. Divide each number x_i into two parts: a stem, consisting of one or more of the leading digits, and a leaf, consisting of the remaining digit.
  2. List the stem values in a vertical column.
  3. Record the leaf for each observation beside its stem.
  4. Write the units for stems and leaves on the display.

Example

The listing scores of 12 students in a TOEIC test are listed below

$$55 \quad 115 \quad 225 \quad 240 \quad 330 \quad 335 \quad 385 \quad 400 \quad 405 \quad 405 \quad 495 \quad 495$$

The stem-and-leaf diagram

Stem	Leaves
5	5
11	5
22	5
24	0
33	0 5
38	5
40	0 5 5
49	5 5

3. Box-plots

Three quartiles

An ordered set of data is divided into four equal parts, the division points are called quartiles:

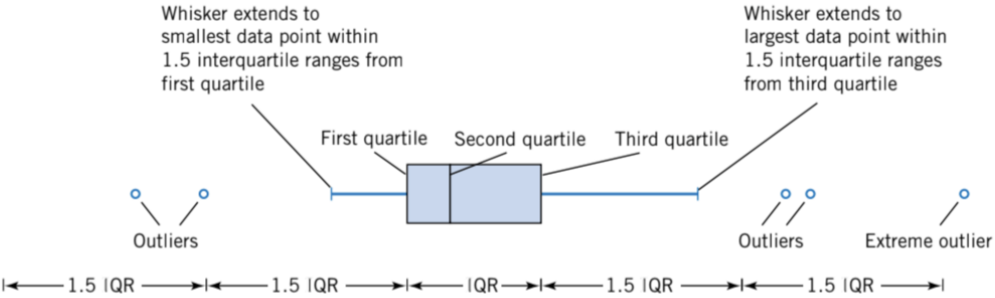
- The first quartile, q_1 or Q_1: is a value that has approximately 25% of the observations below.
- The sample median or second quartile, q_2 or Q_2, has approximately 50% of the observations below its value.
- The third quartile, q_3 or Q_3, has approximately 75% of the observations below its value.
- The interquartile range, $\text{IQR} = Q_3 - Q_1$

Example. Use the given sample data to find the sample quartiles, the sample mode and the IQR.

$$55, \quad 52, \quad 52, \quad 52, \quad 49, \quad 74, \quad 67, \quad 55$$

Box-plots

A **box-plot** is a visual display that describes important features of data: three quartiles, the minimum/maximum values, and unusual observations (outliers).



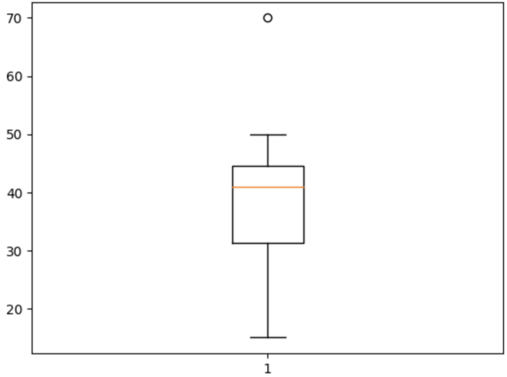
Whisker extends to smallest data point within 1.5 interquartile ranges from first quartile
Whisker extends to largest data point within 1.5 interquartile ranges from third quartile
First quartile
Second quartile
Third quartile
Outliers
Outliers
Extreme outlier
1.5 IQR
1.5 IQR
IQR
1.5 IQR
1.5 IQR

Example

Given a data of ages of 14 random adults from a village:

$$15, 20, 31, 31, 32, 40, 41, 41, 42, 43, 45, 45, 50, 70$$

Draw a box plot for this data.



4. Histograms

Frequency distribution

Construction of **frequency distribution**: divide the range of the data into intervals (called class intervals, cells, or bins). The bins should be of equal width.

Example. The final exam grades of a group of 10 students are given by:

$$2.4, 4.4, 4.6, 5.0, 5.0, 5.8, 6.0, 7.4, 8.2, 9.0$$

- Divide grade ranges into 5 bins: 0 - 2, 2 - 4, 4 - 6, 6 - 8, 8 - 10.
- Count the number of data values in each bin.

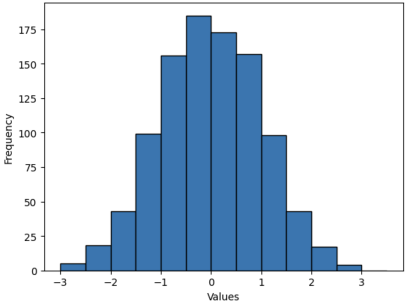
Bin	Frequency
0 - 2	0
2- 4	1
4 - 6	6
6-8	1
8 -10	2

Histograms

The histogram is a visual display of the frequency distribution.

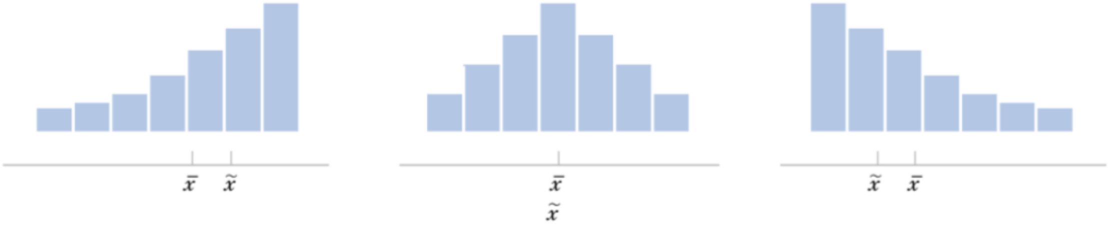
- Label the bin (class interval) boundaries on a horizontal scale.
- Mark and label the vertical scale with the frequencies or the relative frequencies.

Above each bin, draw a rectangle where height is equal to the Frequency (or relative frequency) corresponding to that bin.



Remarks

1. Histograms are very useful to explore the distribution of data.



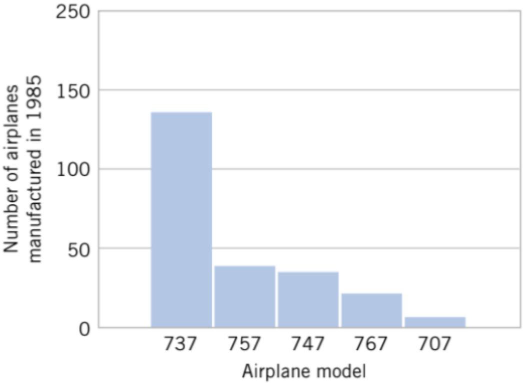
Negative or left skew
(a)

Symmetric
(b)

Positive or right skew
(c)

Remarks (cont')

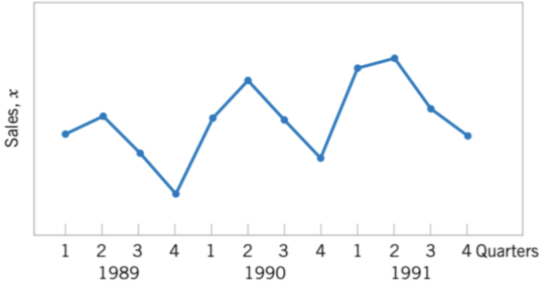
2. Pareto chart: (frequencies are ordered decreasingly)



Times sequence plots

A time series or time sequence is a data set in which the observations are recorded in the order in which they occur.

A **time series plot** is a graph in which the vertical axis denotes the observed value of the variable (say x) and the horizontal axis denotes the time (which could be minutes, days, years, etc.)



Thank you!

namvv14@fe.edu.vn