Chapter 1: Descriptive Statistics – PART 2

Manuella Lech Cantuaria
Victoria Blanes-Vidal
The Maersk Mc-Kinney Moller Institute
Applied AI and Data Science

LECTURE PLANNING

Lesson	Week	Date	TOPICS	Teacher
1	35	1/Sep	Introduction to the course Descriptive statistics – Part I	MLC
2	36	8/sep	Descriptive statistics – Part II	MLC
3	37	15/Sep	Probability distributions	MLC
4	38	22/Sep	Hypothesis testing (one sample)	VBV
5	39	29/Sep	Hypothesis testing (two samples)	VBV
6	40	6/Oct	ANOVA one-way	VBV
7	41	13/Oct	R class (Introduction to R and descriptive statistics) **Point-giving activity (in class)**	MLC+VBV
-	42	20/Oct	NO CLASS (Autum holidays)	
8	43	27/Oct	R class (hypothesis testing + ANOVA)	MLC
9	44	3/Nov	ANOVA two-way	VBV
-	45	10/Nov	NO CLASS	
10	46	17/Nov	Regression analysis	VBV
11	47	24/Nov	Notions of experimental design and questions **Point-giving activity (in class)**	VBV+MLC
12	48	1/Dec	Multiple regression	MLC

Descriptive statistics

Inferential statistics

Chapter 1 Overview

1.1. Statistics: Descriptive and Inferential

1.2. Variables and Types of Data

1.3. Measures of:

Central Tendency (Location)

Variation (Dispersion)

Position

1.4. Data representation: frequency distributions, histograms and other graphs

1.5. Shapes of frequency distributions: Skewness and kurtosis

OpenIntro book:

- Chapter 1 (Pag 22-24)
- Chapter 2

Let’s already warm up!

Consider you have a dataset composed only by integer values, and there is no repeated value in this dataset.

Which of these statements you can affirm is correct?

a) The median of this dataset is certainly an integer.

b) The mean of this dataset is certainly a positive number.

c) The mode is a good measure of central tendency for this dataset.

d) The standard deviation is certainly a positive number.

1-3 Measures of Central Tendency (location), Variation (dispersion) and Position

- The data distribution can be mainly described by three different characteristics:
  - Measures of location
  - Measures of dispersion
  - Measures of position

Measures of Position

Study to better understand characteristics of software developers working in Odense

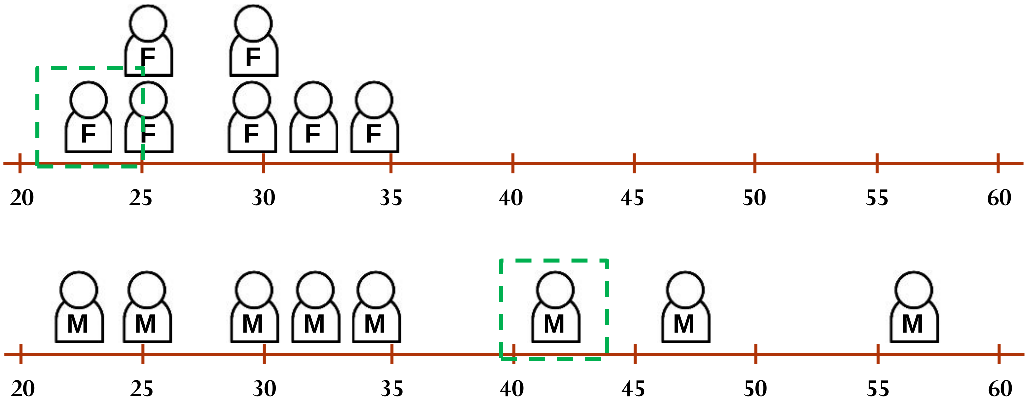
Software developer	Sex	Age	Preferred language
1	M	32	Python
2	M	41	HTML
3	F	23	SQL
4	M	56	Python
5	F	32	Python
6	M	34	HTML
7	M	47	SQL
8	F	25	Python
9	F	29	JavaScript
10	F	29	Python
11	M	30	JaveScript
12	M	23	Python
13	F	34	Python
14	F	25	HTML
15	M	25	SQL

Age of male and female software developers

Male	Female
32	23
41	32
56	25
34	29
47	29
30	34
23	25
25	

Measures of Position

Measures of position indicate the position of a value, relative to other values in a set of observations.



Measures of Position

Measures of position indicate the position of a value, relative to other values in a dataset.

- Percentile
- Decile and Quartile
- Outlier

Measures of Position: Percentiles

- **Percentiles** separate the data set into 100 equal groups.
- A percentile rank for a datum represents the percentage of data values below the datum.

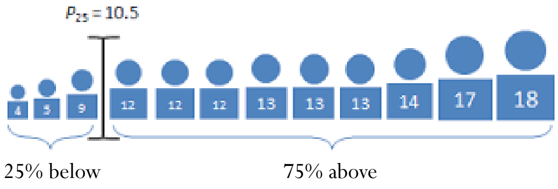
Example: You are the fourth tallest person in a group of 20 (80% of people are shorter than you):

That means you are at the 80th percentile. If your height is 1.85m then "1.85m" is the 80th percentile height in that group.

Measures of Position: Percentiles

$$Percentile = \frac{\text{\# of values below } X}{\text{total \# of values}} \times 100\%$$

Example: What is the percentile of a child whose age is 10 years and half in this group of children?



$$\text{Percentile} = \frac{3}{12}.100 = 25^{th}$$

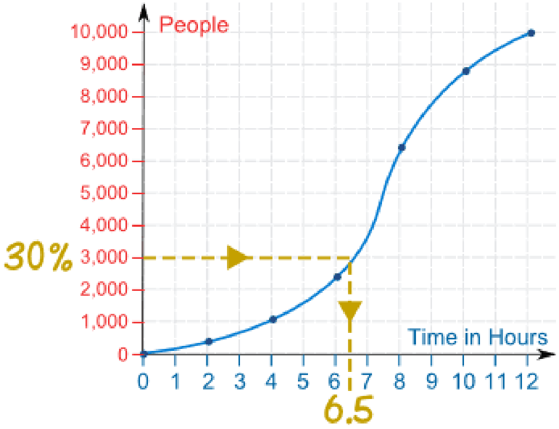
If a child is 10 years and half, he has 3 children out of 12 below him (younger than him), so the percentile rank of that child in this group is 25th percentile (25% of the children are younger than him, and 75% are above his age)

Measures of Position: Example of a Percentile Graph

A total of 10,000 people visited a shopping mall over 12 hours:

Time (hours)	People
0	0
2	350
4	1100
6	2400
8	6500
10	8850
12	10,000

Estimate the 30th percentile (when 30% of the visitors had arrived).



The 30th percentile occurs after about 6.5 hours.

11

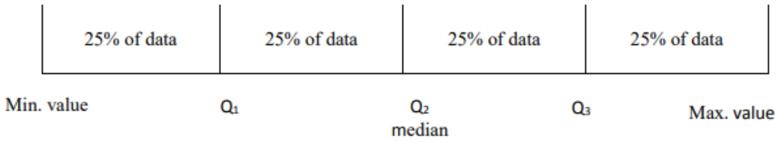
11

Measures of Position: Quartiles and Deciles

- **Deciles** separate the data set into 10 equal groups.

  $D_1=P_{10}$, $D_4=P_{40}$

- **Quartiles** separate the data set into 4 equal groups.



$Q_1=P_{25}$, $Q_2=MD$, $Q_3=P_{75}$

- The **Interquartile Range**, $IQR = Q_3 - Q_1$.

12

12

Quartiles

- There are different methods to calculate quartiles for discrete numbers.
- One of the methods is the following:

STEP 1: Sort the dataset in ascending order

STEP 2: Use the **median** (**Q2**) to divide the ordered data set into two-halves.

- If there is an odd number of data points in the original ordered data set, **do not include** the median (the central value in the ordered list) in either half.
- If there is an even number of data points in the original ordered data set, split this data set exactly in half.

STEP 3: The lower quartile value (**Q1**) is the median of the lower half of the data.

STEP 4: The upper quartile value (**Q3**) is the median of the upper half of the data.

Measures of Position: Outliers

- An **outlier** is an extremely high or low data value when compared with the rest of the data values.
- A data value:
  - less than $Q_1 - 1.5(\text{IQR})$

  Or

  - greater than $Q_3 + 1.5(\text{IQR})$

  can be considered an outlier.

Let’s practice!

- **Daily low temperatures** recorded in a town (01/18-01/31, 2005, °F)

Jan. 18 – 11
Jan. 19 – 11
Jan. 20 – 25
Jan. 21 – 29
Jan. 22 – 27
Jan. 23 – 14
Jan. 24 – 11
Jan. 25 – 25
Jan. 26 – 33
Jan. 27 – 22
Jan. 28 – 18
Jan. 29 – 19
Jan. 30 – 30
Jan. 31 – 27

- Is the value Tmin=33°F an outlier?

Chapter 1 Overview

1.1. Statistics: Descriptive and Inferential
1.2. Variables and Types of Data
1.3. Measures of:
Central Tendency (Location)
Variation (Dispersion)
Position
★ **1.4. Data representation: frequency distributions, histograms and other graphs**
1.5. Shapes of frequency distributions: Skewness and kurtosis

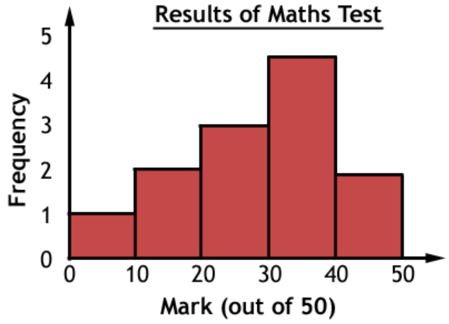
1-4 Data representation: frequency distributions, histograms and other graphs

- When conducting a statistical study, the researcher must gather data for the particular variable under study.
- To describe situations (descriptive statistics) or draw conclusions and make inferences about populations (inferential statistics), the researcher must organize and present the data in some meaningful way.
- We will look more particularly at:
  - ➢ Histograms
  - ➢ Boxplots
  - ➢ Other general graphs

17

Histogram

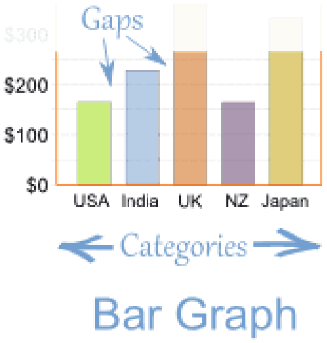
The ***histogram*** is a graph that displays the data by using vertical bars of various heights to represent the frequencies of the classes.

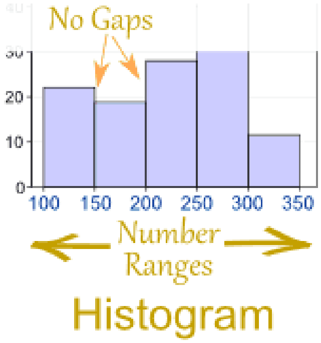


The height of each bar represents the percentage (or counts) of data values in the interval

18

Bar chart and histograms





- Graphical representation of categorical data using rectangular bars where the length of each bar is proportional to the value they represent

- Used to describe the variability of the data
- Divide range of possible measurements into a number of groups
- Count observations in each group

Building a Histogram

- **Daily low temperatures** recorded in a town (01/18-01/31, 2005, °F)

Jan. 18 – 11
Jan. 19 – 11
Jan. 20 – 25
Jan. 21 – 29
Jan. 22 – 27
Jan. 23 – 14
Jan. 24 – 11

Jan. 25 – 25
Jan. 26 – 33
Jan. 27 – 22
Jan. 28 – 18
Jan. 29 – 19
Jan. 30 – 30
Jan. 31 – 27

Building a Histogram

- (1) Develop an ungrouped frequency table

→ Data (minimum measured temperature: T_{min} (F)):

11, 11, 11, 14, 18, 19, 22, 25, 25, 27, 27, 29, 30, 33

→

11	3
14	1
18	1
19	1
22	1
25	2
27	2
29	1
30	1
33	1

Building a Histogram

- 2. Construct a grouped frequency table

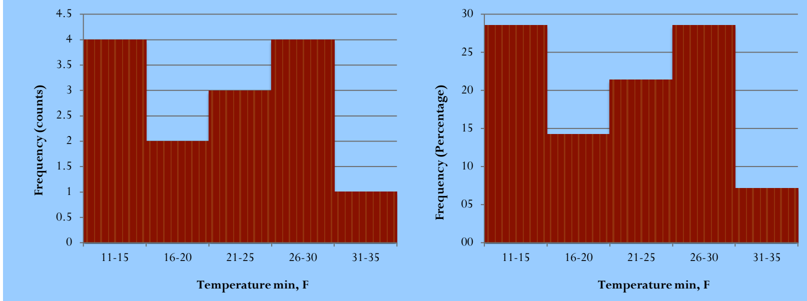
→ Select a set of classes

→

11-15	4
16-20	2
21-25	3
26-30	4
31-35	1

Building a Histogram

- 3. Plot the frequencies of each class



23

Box-plot

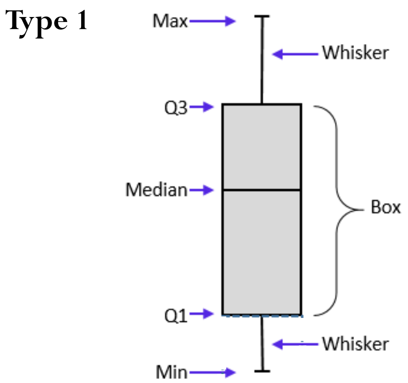
- The **Five-Number Summary** is composed of the following numbers: Low, Q_1, median, Q_3, High
- The Five-Number Summary can be graphically represented using a **Boxplot**.

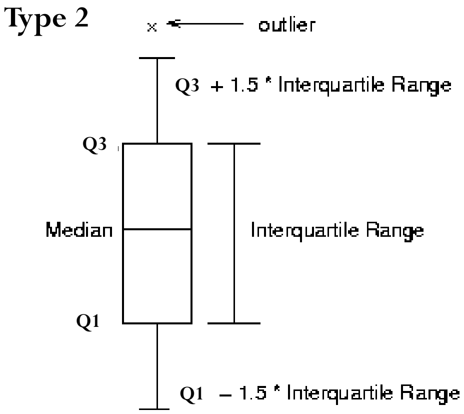
24

24

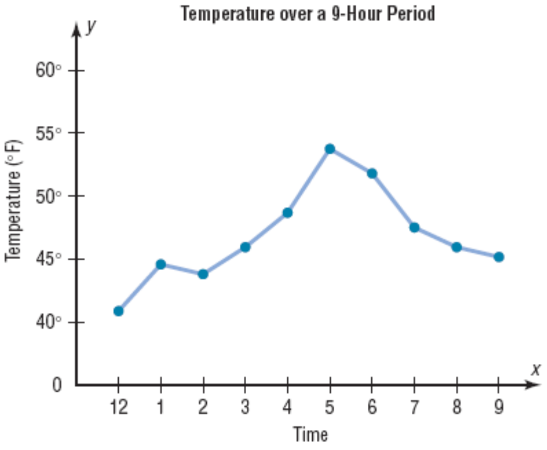
Types of box-plots

“Type 1” and “Type 2”

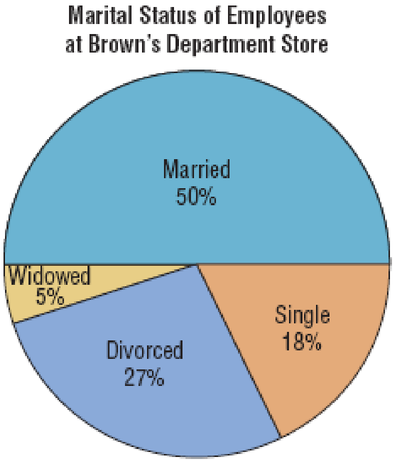



Type 2
outlier
Q3 + 1.5 * Interquartile Range
Q3
Median
Interquartile Range
Q1
Q1 – 1.5 * Interquartile Range

Other Types of Graphs
Time Series Graphs



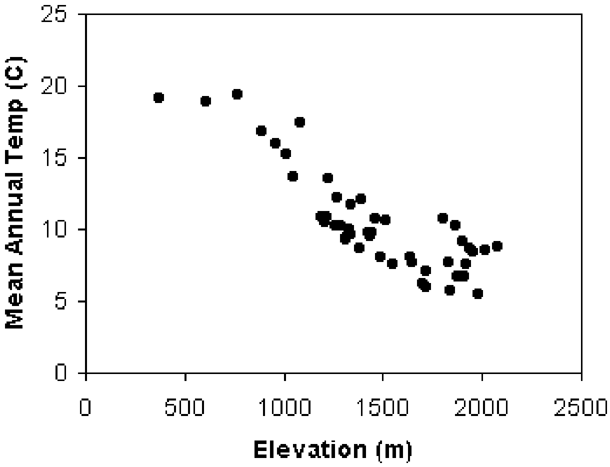
Other Types of Graphs
Pie Graphs



27

27

Scatter Plot

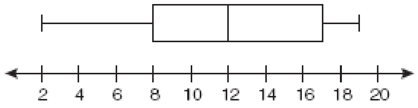


28

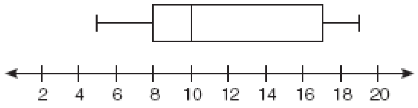
Discuss together with your peers

The data set 5, 14, 9, 19, 6, 9, 19, 7, 8, 17, 9, 18, 10, 17, 12 represents the number of hours spent on the Internet in a week by students in a mathematics class. Which box-plot represents the data?

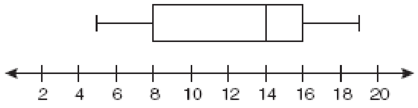
(1)



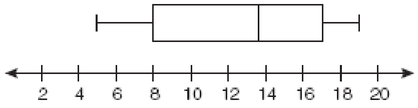
(2)



(3)



(4)



Chapter 1 Overview

1.1. Statistics: Descriptive and Inferential

1.2. Variables and Types of Data

1.3. Data organization and histograms

1.4. Measures of:

Central Tendency (Location)

Variation (Dispersion)

Position

1.5. Data representation: frequency distributions and graphs

1.6. Shapes of frequency distributions: Skewness and kurtosis

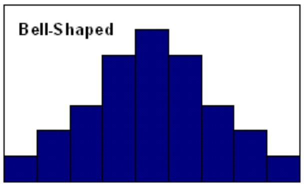
1.5 Shapes of Distributions

- While measures of dispersion are useful for helping us describe the width of the distribution, they tell us nothing about the **shape of the distribution**
- There are two important parameters that can be used to describe the shape of a distribution:
  - **Skewness**
  - **Kurtosis**

31

Skewness

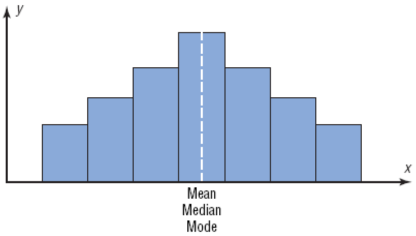
- Skewness of a distribution is a measure of symmetry, or more precisely, the lack of symmetry.
- A distribution, or data set, is symmetric if it looks the same to the left and right of the center point.



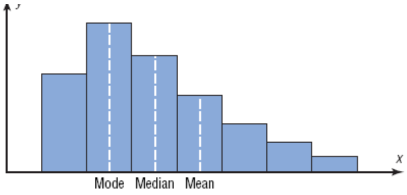
32

Skewness

- **No-skewness**

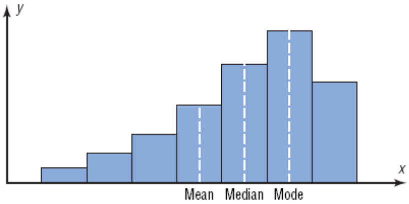


- **Positive skewness**



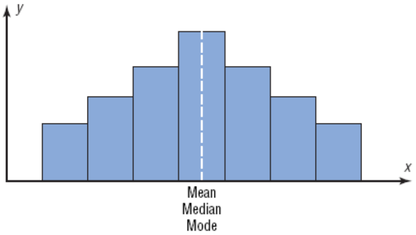
(a) Positively skewed or right-skewed

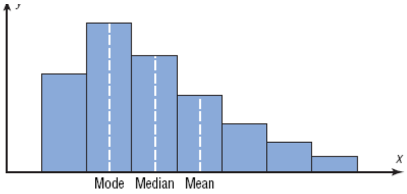
- **Negative skewness**



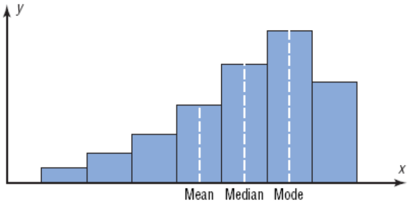
Skewness

- **No-skewness**
  - Same observations below and above the mean
  - Mean and median coincide
- **Positive skewness**
  - There are more observations below the mean than above it
  - When the mean is greater than the median
- **Negative skewness**
  - There are a small number of low observations and a large number of high ones
  - When the median is greater than the mean





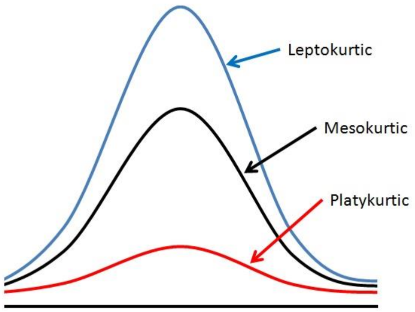
(a) Positively skewed or right-skewed

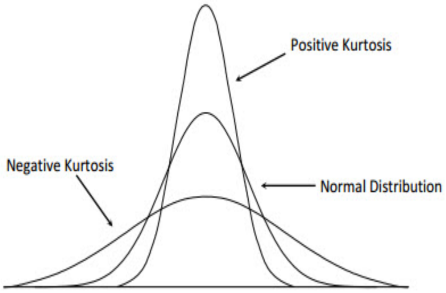


Kurtosis

- **Kurtosis** measures how **peaked** the histogram is
- **Kurtosis** characterizes the relative **peakedness** or **flatness** of a distribution compared to the **normal distribution**

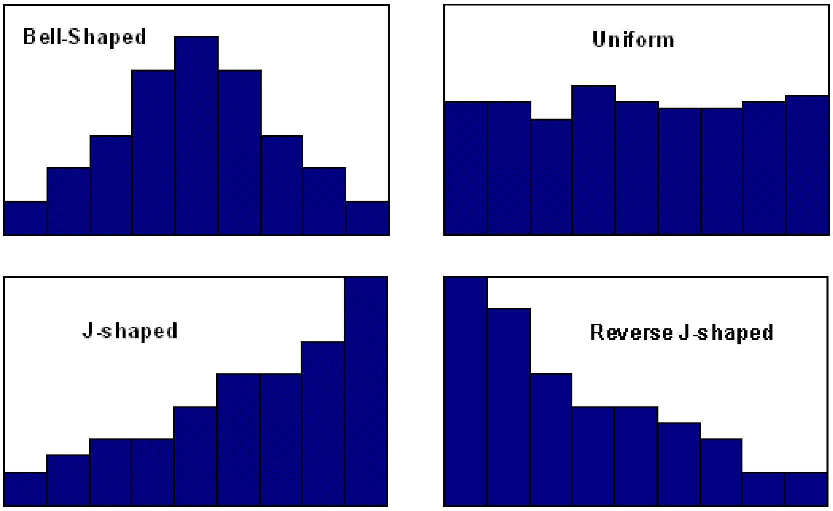
- **Leptokurtic**– positive kurtosis indicates a relatively peaked distribution
- **Platykurtic**– negative kurtosis indicates a relatively flat distribution
- **Mesokurtic** (in between)





35

Some examples of distributions' shapes

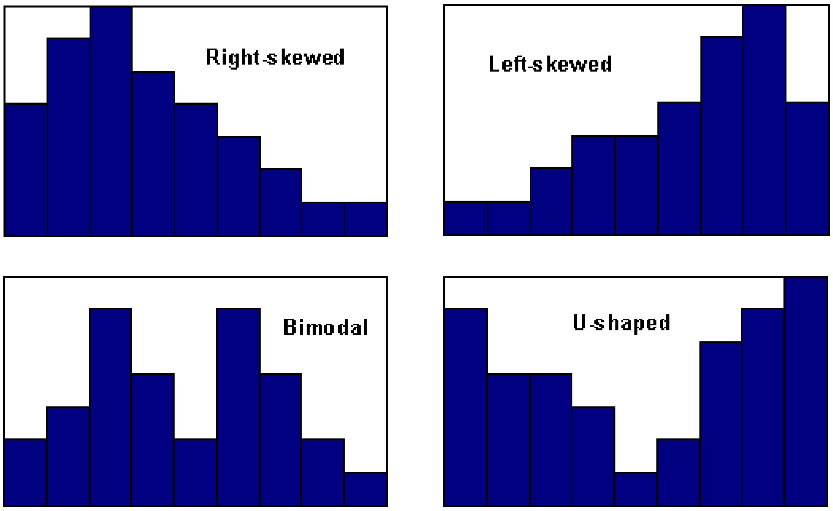


36

Bluman, Chapter 2

36

Some examples of distributions' shapes



37

Bluman, Chapter 2

37

Questions?

Now it is time for you to practice what you learned at the exercises' class!

Rooms: U90, U171, U172, U176

38

38