SNS COLLEGE OF TECHNOLOGY

Coimbatore-35.
An Autonomous Institution

COURSE NAME : DATA ANALYTICS

II YEAR/ IV SEMESTER

UNIT – II Getting Insights from Data

Topic: *Univariate Analysis*

Dr.K.Sangeetha
HoD
Department of Computer Science and Engineering

Descriptive Statistics (cont..)

Descriptive Statistics :

- ❖ Descriptive Univariate Analysis
  - Univariate Frequencies
  - Univariate Data Visualization
  - Univariate Statistics
  - Common Univariate Probability Distributions

Descriptive Statistics (cont..)

Descriptive Univariate Analysis :

❖ three types of information can be obtained: frequency tables, statistical measures and plots.

1. Univariate Frequencies :

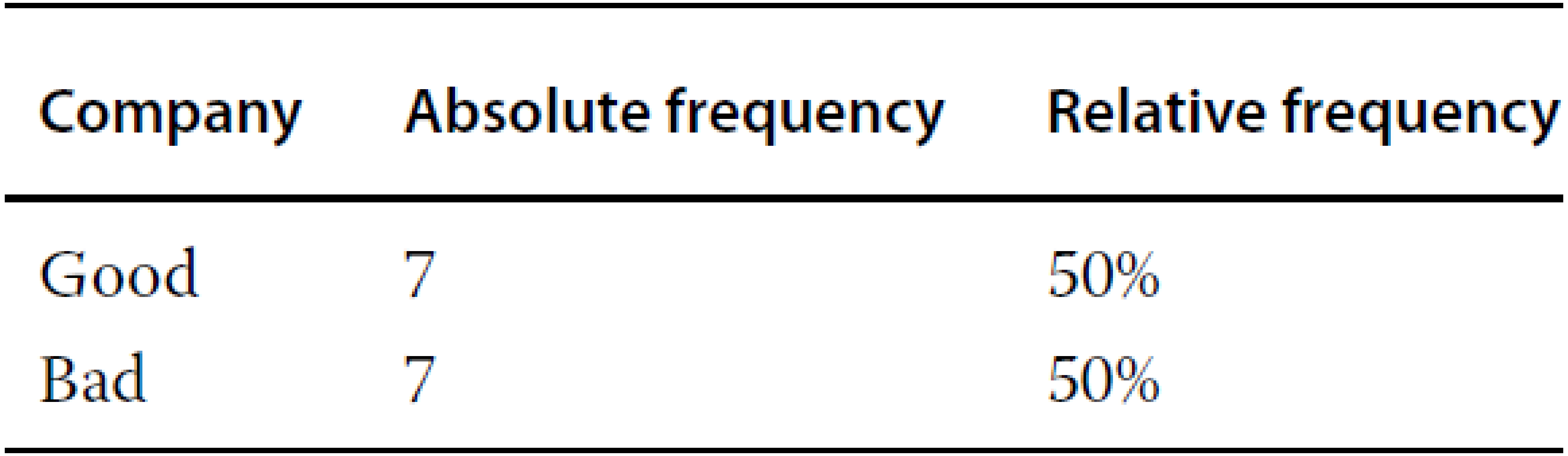
•A frequency is basically a counter.

•The **absolute frequency** counts how many times a value appears.

•The **relative frequency** counts the percentage of times that value appears.

Descriptive Statistics (cont..)

1. **Univariate Frequencies :**



Company	Absolute frequency	Relative frequency
Good	7	50%
Bad	7	50%

Descriptive Statistics (cont..)

2. Univariate Data visualization:

Five different types of charts**:**

❖ Pie chart - These are used typically for nominal scales.

❖ Bar charts - These are used typically for qualitative scales

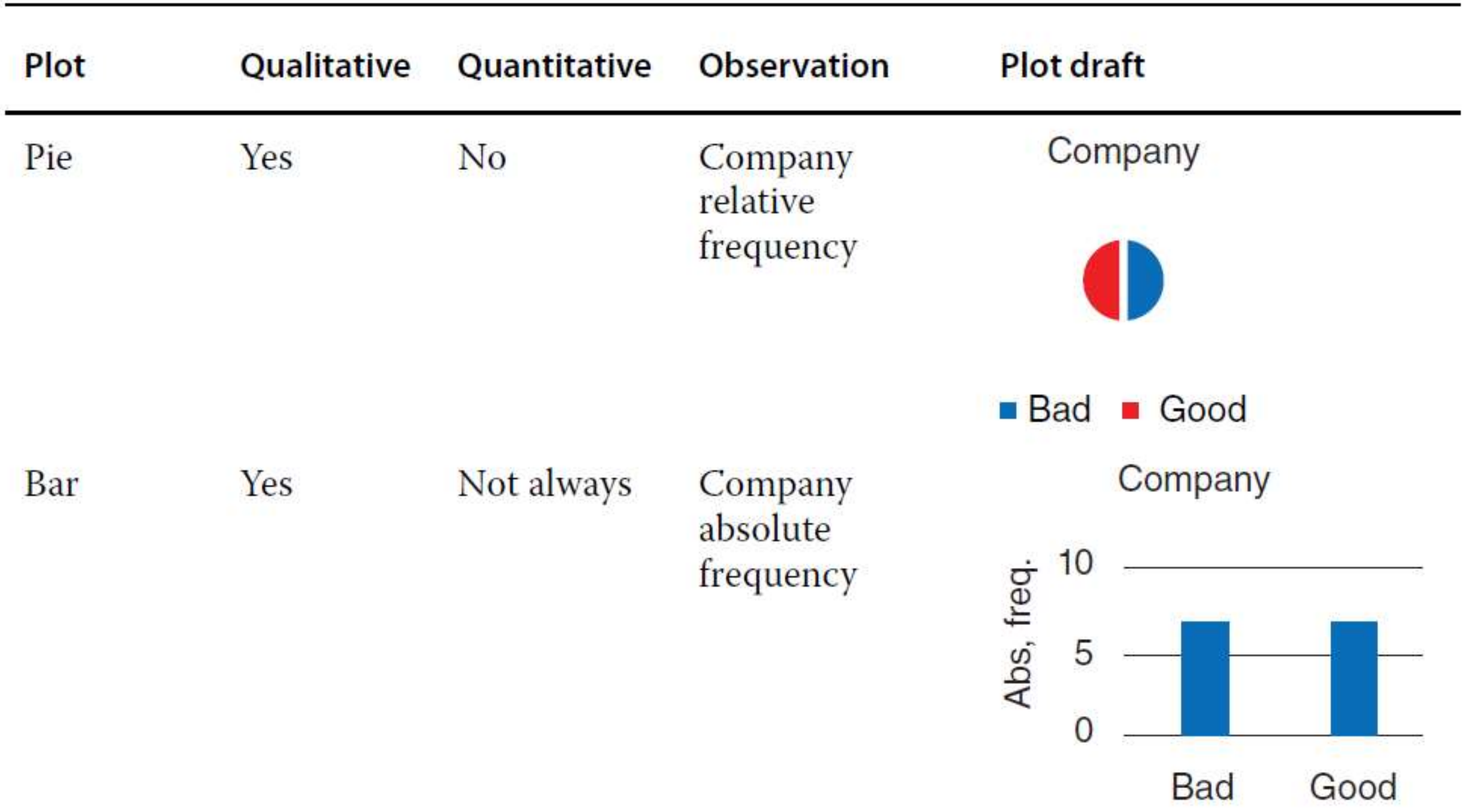
❖ Line charts -

❖ Area charts - used to compare time series and distribution functions.

❖ Histograms - used to represent empirical distributions for attributes with a quantitative scale.

Descriptive Statistics (cont..)

2. Univariate Data visualization:

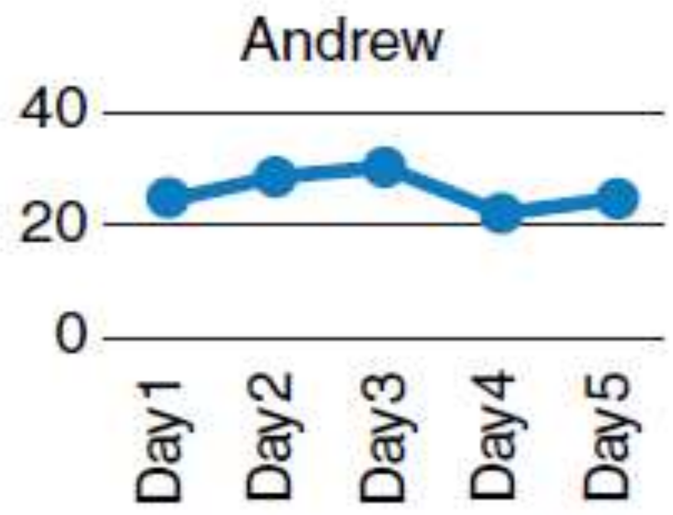
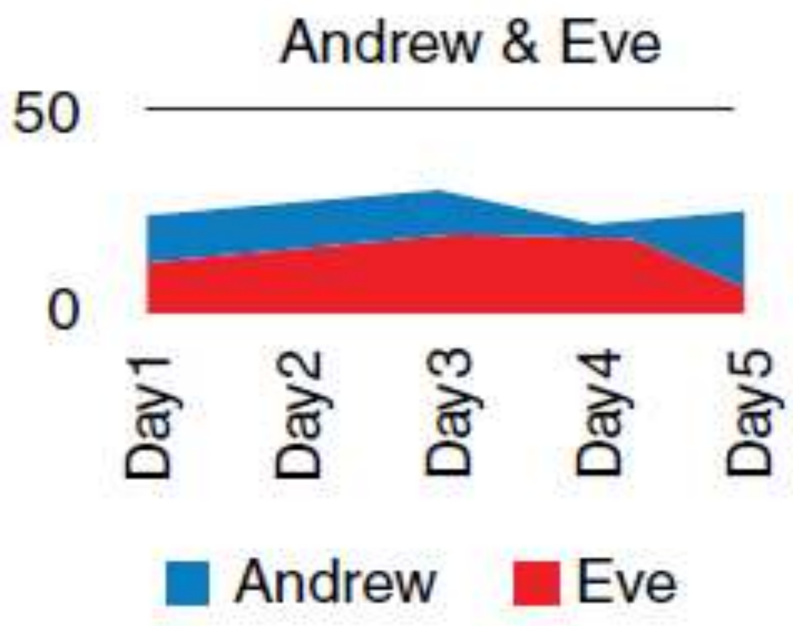


Plot	Qualitative	Quantitative	Observation	Plot draft
Pie	Yes	No	Company relative frequency	
Bar	Yes	Not always	Company absolute frequency	

Company
■ Bad ■ Good
Company
Abs, freq.
10
5
0
Bad
Good

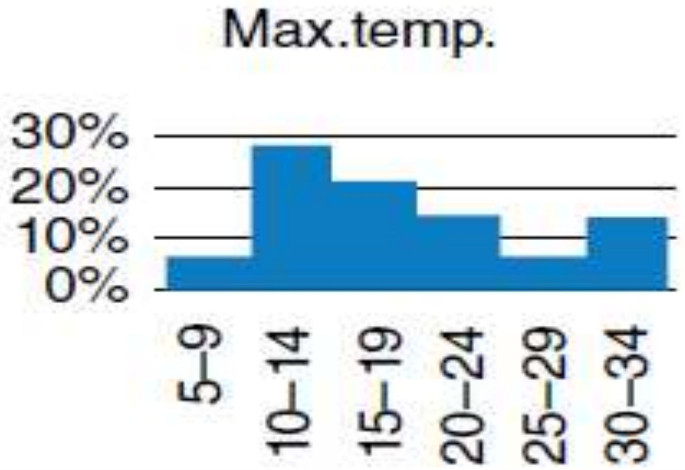
Descriptive Statistics (cont..)

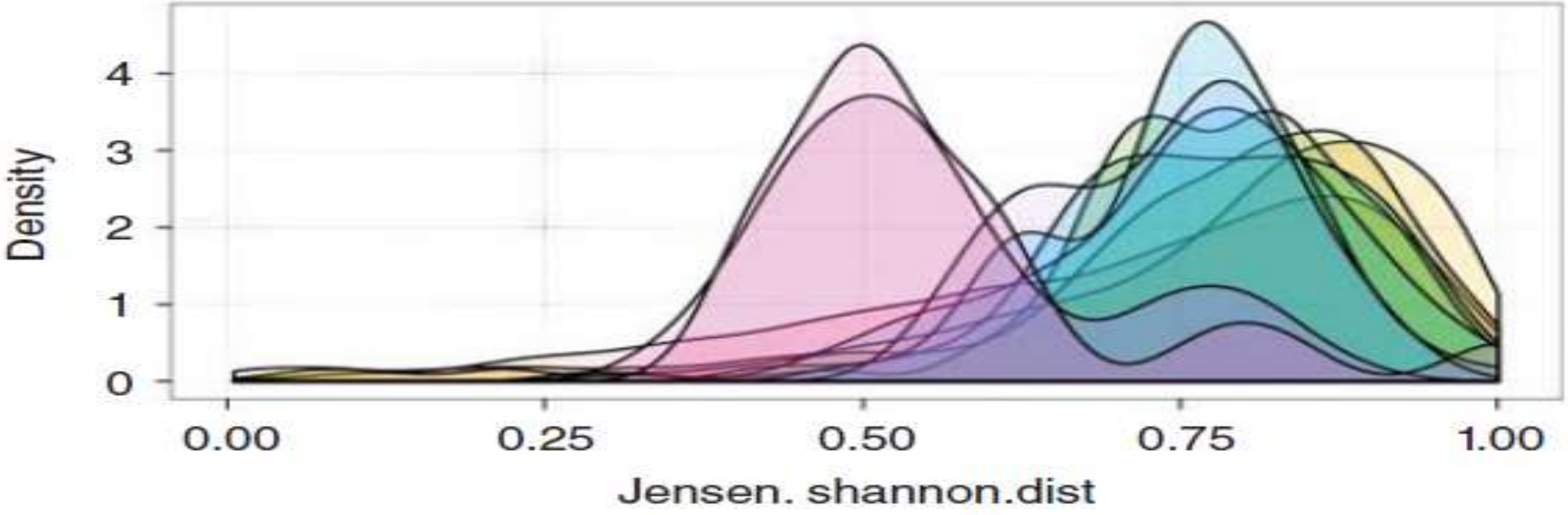
2. Univariate Data visualization:

Line	No	Yes	Andrew's 5-day max. temperatures	
Area	No	Yes	Andrew and Eve 5-day max. temperatures	

Descriptive Statistics (cont..)

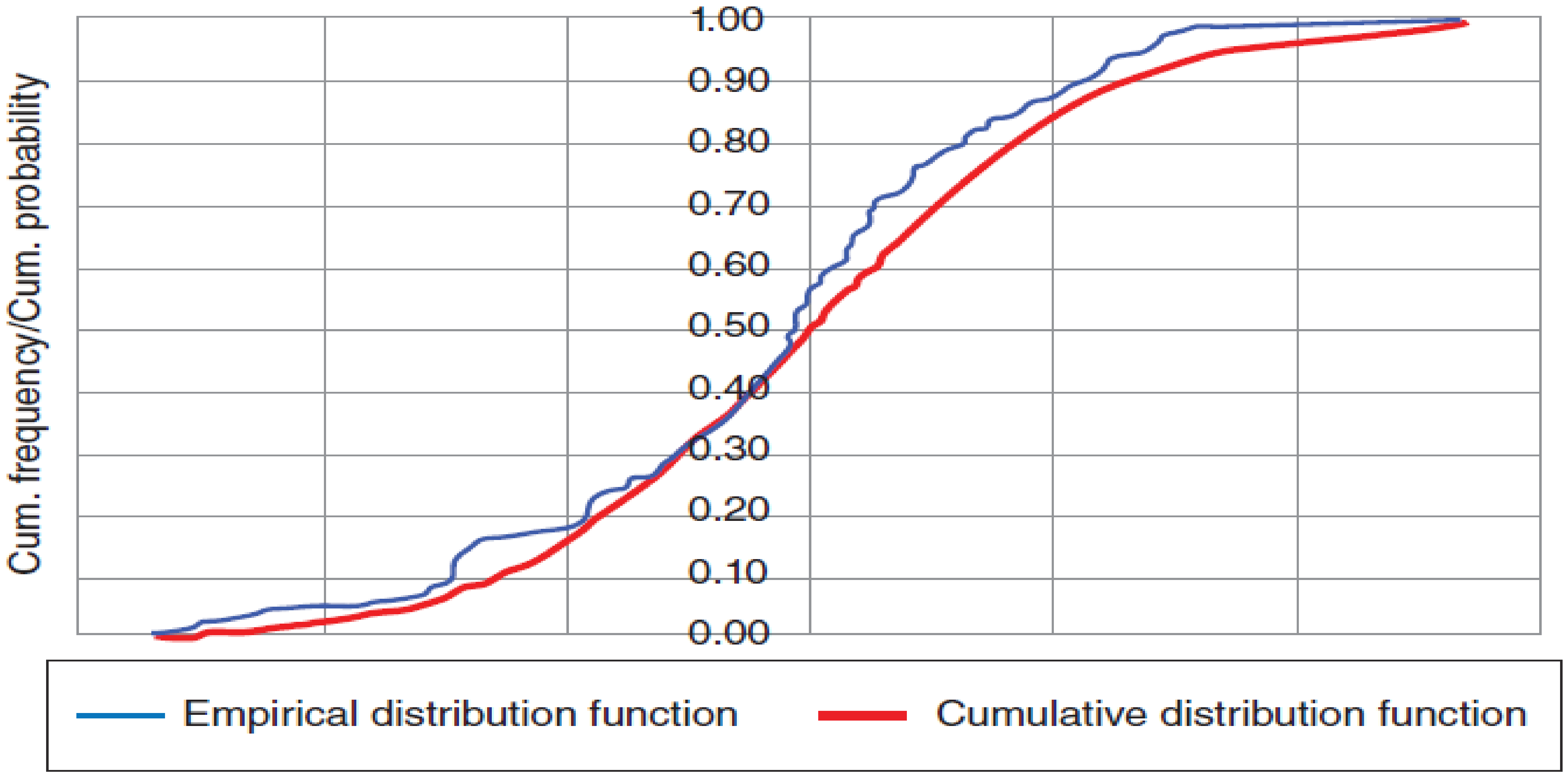
2. Univariate Data visualization:

Histogram	No	Yes	Max. last day temperatures of the 14 contacts	



Descriptive Statistics (cont..)

2. **Univariate Data visualization:**
   Empirical and probability distribution functions :



Descriptive Statistics (cont..)

3. Univariate Statistics:

❖ A statistic is a descriptor.

❖It describes numerically a characteristic of the sample or the population.

❖ There are two main groups of univariate statistics:

- Location statistics
- Dispersion statistics.

Descriptive Statistics (cont..)

3. Univariate Statistics:

Location univariate statistics:

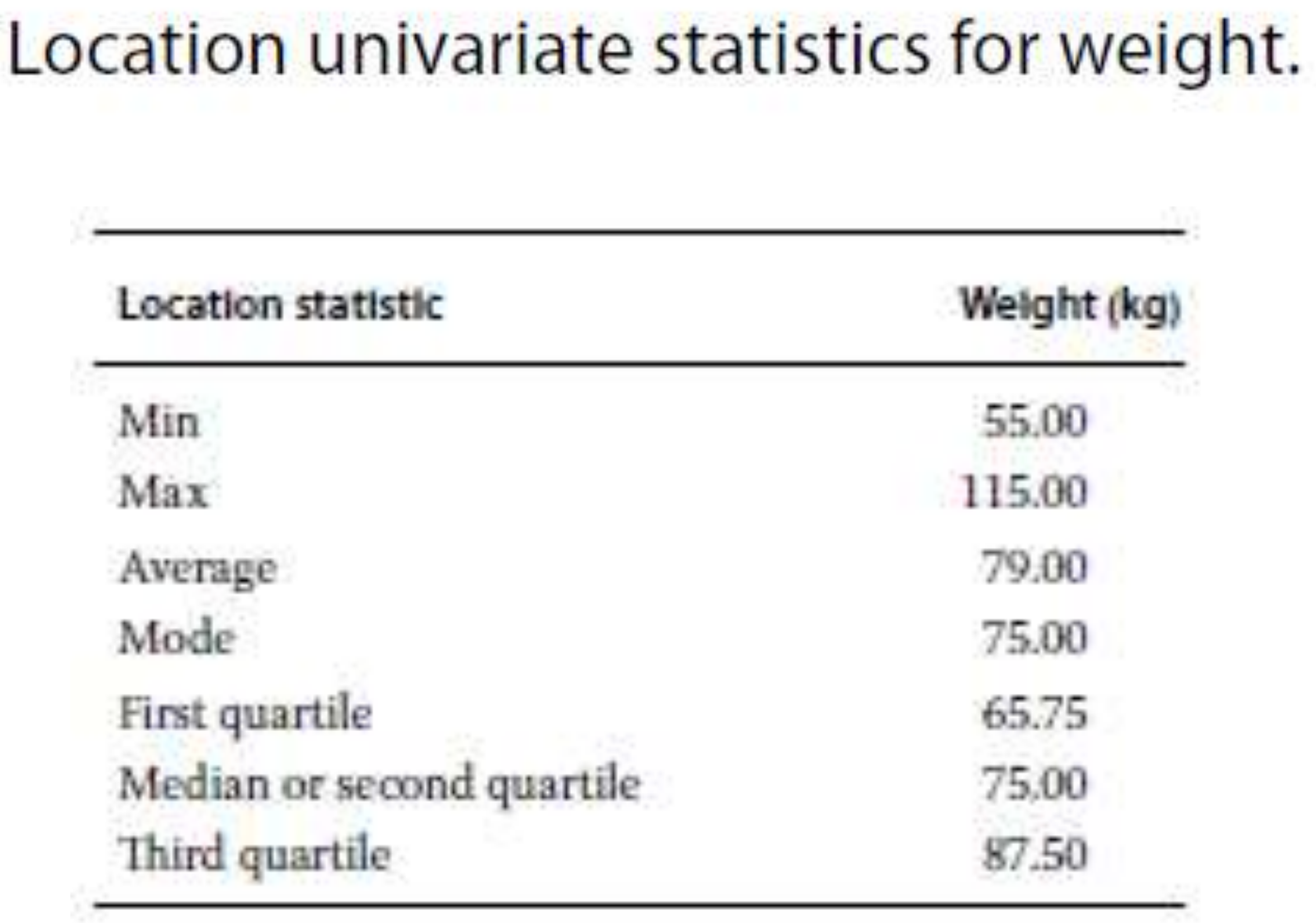
❖identify a value in a certain position.

❖location univariate statistics are the minimum, the maximum or the mean

- **minimum:** the lowest value
- **maximum:** the largest value
- **mean:** the average value,
- **node:** the most frequent value;
- **first quartile**: the value that is larger than 25% of all values
- **median or second quartile**: the value that is larger than 50% of all values;
- **third quartile**: the value that is larger than 75% of all values.

Descriptive Statistics (cont..)

2. Univariate Statistics:

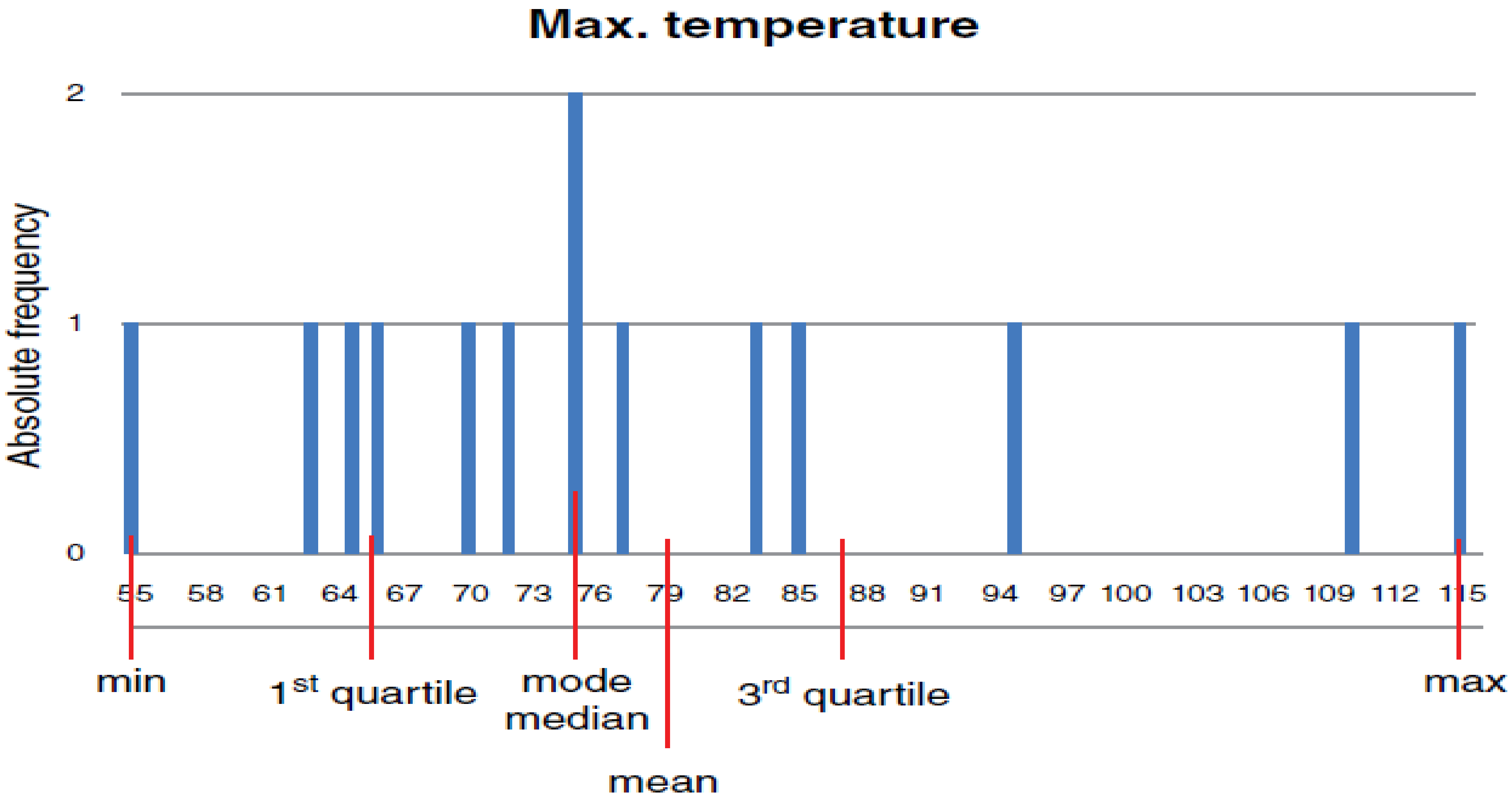


Location univariate statistics for weight.

Location statistic	Weight (kg)
Min	55.00
Max	115.00
Average	79.00
Mode	75.00
First quartile	65.75
Median or second quartile	75.00
Third quartile	87.50

Descriptive Statistics (cont..)

2. Univariate Statistics:



Location statistics on the absolute frequency plot for the attribute “weight”.

Descriptive Statistics (cont..)

❖**Dispersion univariate statistics** - measures how distant different values are. The most common dispersion statistics are:

- **amplitude:** the difference between the maximum and the minimum values

- **interquartile range**: is the difference between the values of the third and first Quartiles

- **mean absolute deviation**: a measure for the mean absolute distance between the observations and the mean. Its mathematical formula for the population is:

$$MAD_x = \frac{\sum_{i=1}^{n} |x_i - \mu_x|}{n},$$

Descriptive Statistics (cont..)

❖**Dispersion univariate statistics** - measures how distant different values are. The most common dispersion statistics are:

- **standard deviation:** another measure for the typical distance between the observations and their mean

•The square of the sample deviation is termed the variance and is denoted as σ^2.

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu_x)^2}{n}},$$

Descriptive Statistics (cont..)

4. Common Univariate Probability Distributions

- Each attribute has its own probability distribution.

- Many common attributes follow functions for which the distribution is already known.
- Types :

1. Uniform distribution
2. Normal distribution

Descriptive Statistics (cont..)

4. Common Univariate Probability Distributions

1. Uniform distribution :

- very simple distribution.
- frequency of occurrence of the values is uniformly distributed in a given interval of values.
- minimum and maximum values of the interval, is denoted as:
- continuous distributions the probabilities are calculated per interval.

$$x \sim U(a,b)$$

Descriptive Statistics (cont..)

4. Common Univariate Probability Distributions

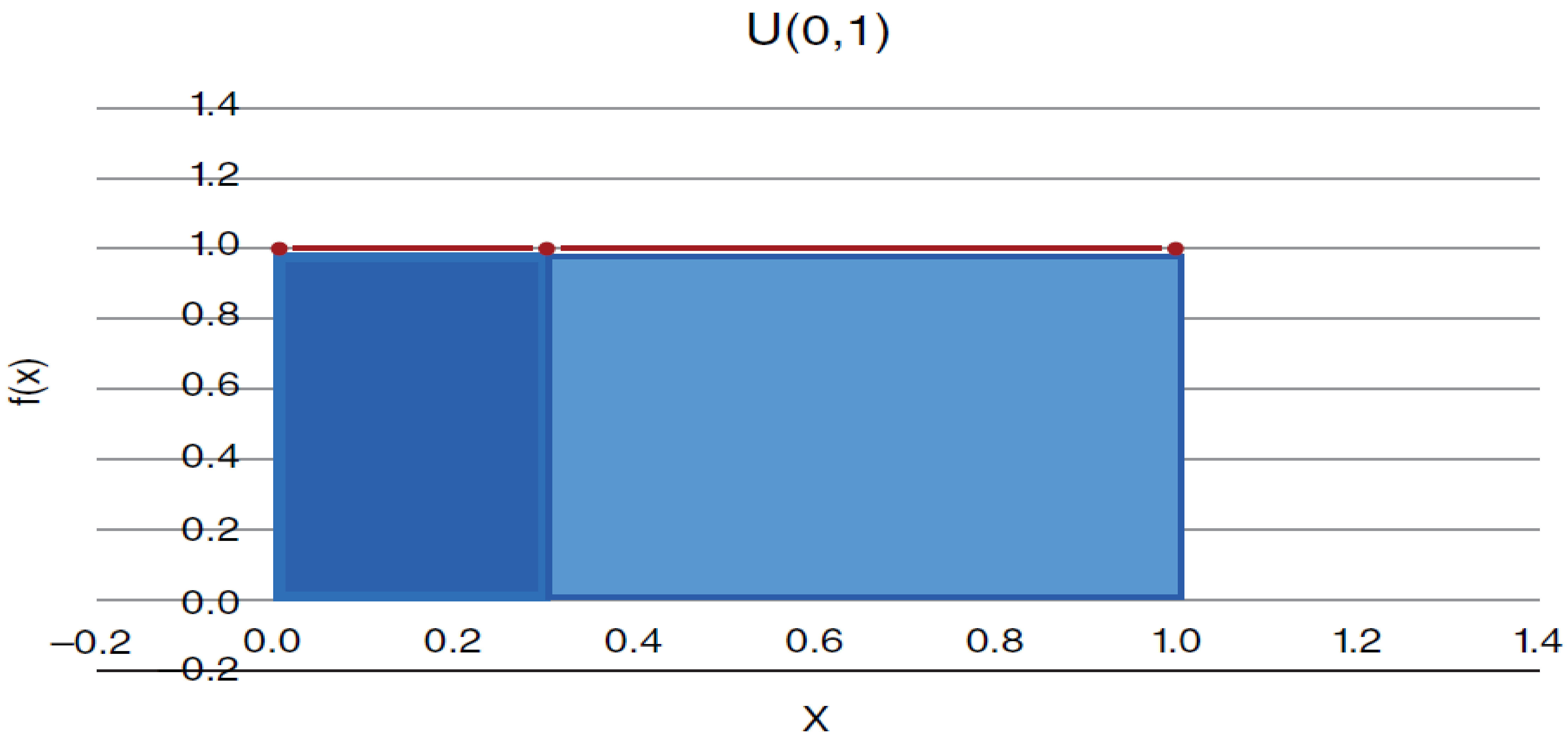
1. Uniform distribution :

$$x \sim \mathcal{U}(a = 0, b = 1).$$

$$P(x < x_0) = \begin{cases} 0, \text{ if } x_0 < a; \\ \frac{x_0 - a}{b - a}, \text{ if } a \leq x_0 \leq b; \\ 1, \text{ if } x_0 > b. \end{cases}$$

Descriptive Statistics (cont..)

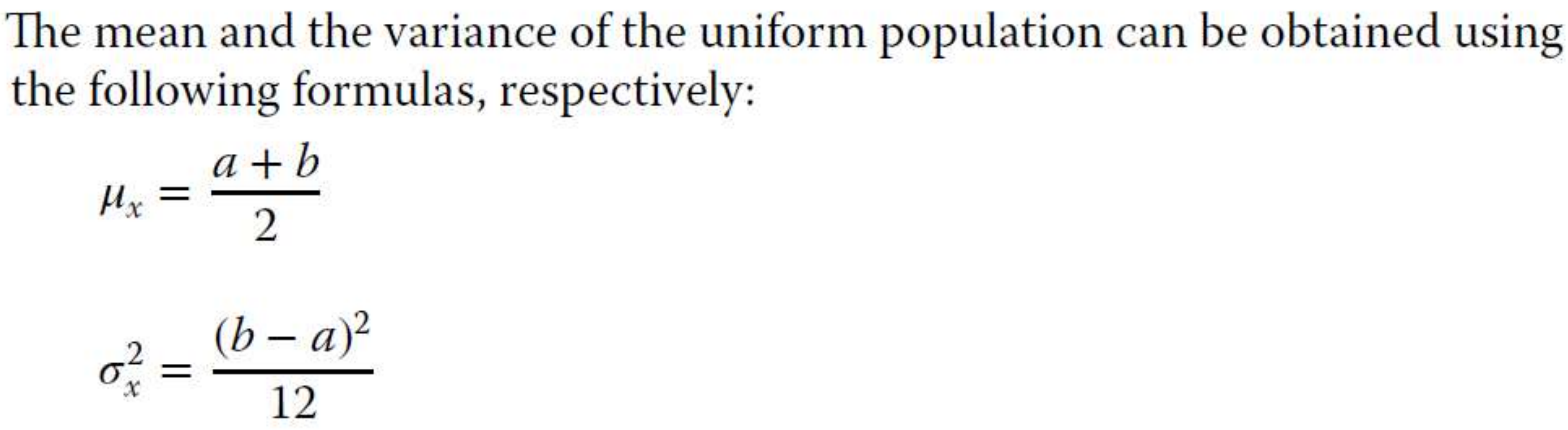
4. Common Univariate Probability Distributions



The probability density function, $f(x)$ of $x \sim \mathcal{U}(0, 1)$.

Descriptive Statistics (cont..)

4. Common Univariate Probability Distributions



The mean and the variance of the uniform population can be obtained using the following formulas, respectively:

$$\mu_x = \frac{a+b}{2}$$

$$\sigma_x^2 = \frac{(b-a)^2}{12}$$

Descriptive Statistics (cont..)

4. Common Univariate Probability Distributions :

The normal distribution:

- ❖ also known as Gaussian distribution
- ❖ normal distribution - symmetric and continuous distribution
- ❖ It has two parameters:

1. the mean
2. the standard deviation.

- ❖ mean - localizes the highest point of the bell-shaped distribution,
- ❖ the standard deviation - defines how thin or wide the bell shape of the distribution is

References

TEXT BOOKS

1. Joao Moreira, Andre Carvalho, Tomás Horvath – “A General Introduction to Data Analytics” – Wiley -2018

REFERENCES

1 Dean J, —Big Data, Data Mining and Machine learning, Wiley publications, 2014.

2 Provost F and Fawcett T, —Data Science for Business, O‘Reilly Media Inc, 2013.

3 Janert PK, —Data Analysis with Open Source Tools, O‘Reilly Media Inc, 2011. .

4 Weiss SM, Indurkhya N and Zhang T, —Fundamentals of Predictive Text Mining, Springer-Verlag London Limited, 2010.

5. Runkler T A, - Data Analytics: Models and Algorithms for Intelligent data analysis,Springer, 2012