SNS COLLEGE OF ENGINEERING

Kurumbapalayam (Po), Coimbatore – 641 107

An Autonomous Institution

Accredited by NBA – AICTE and Accredited by NAAC – UGC with 'A' Grade
Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

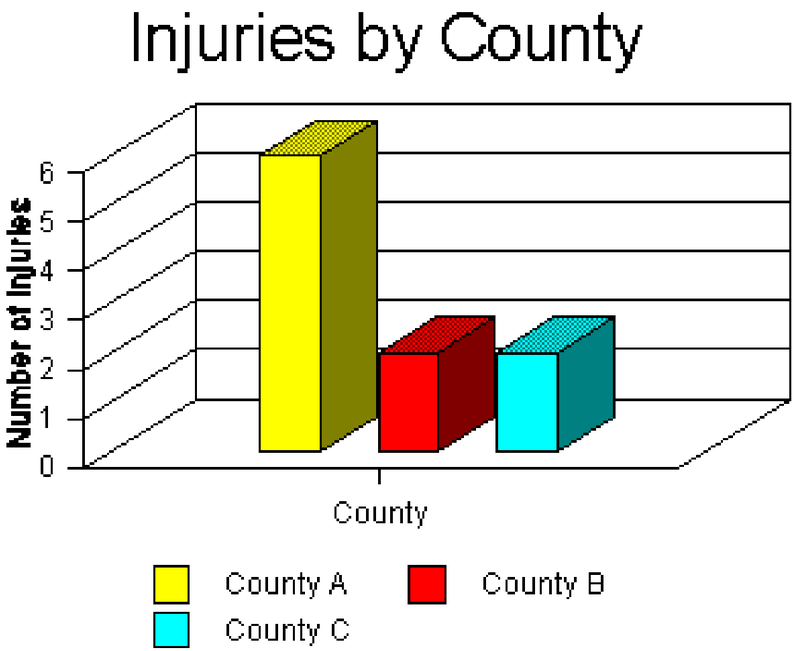
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

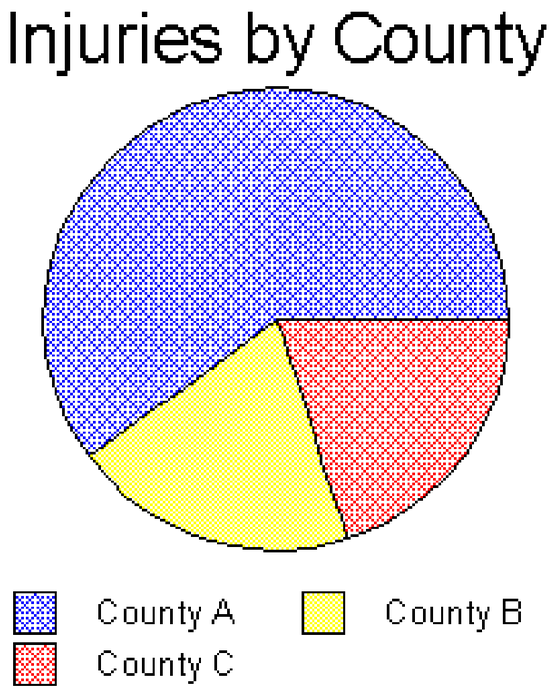
COURSE NAME :19CS407 DATA ANALYTICS WITH R

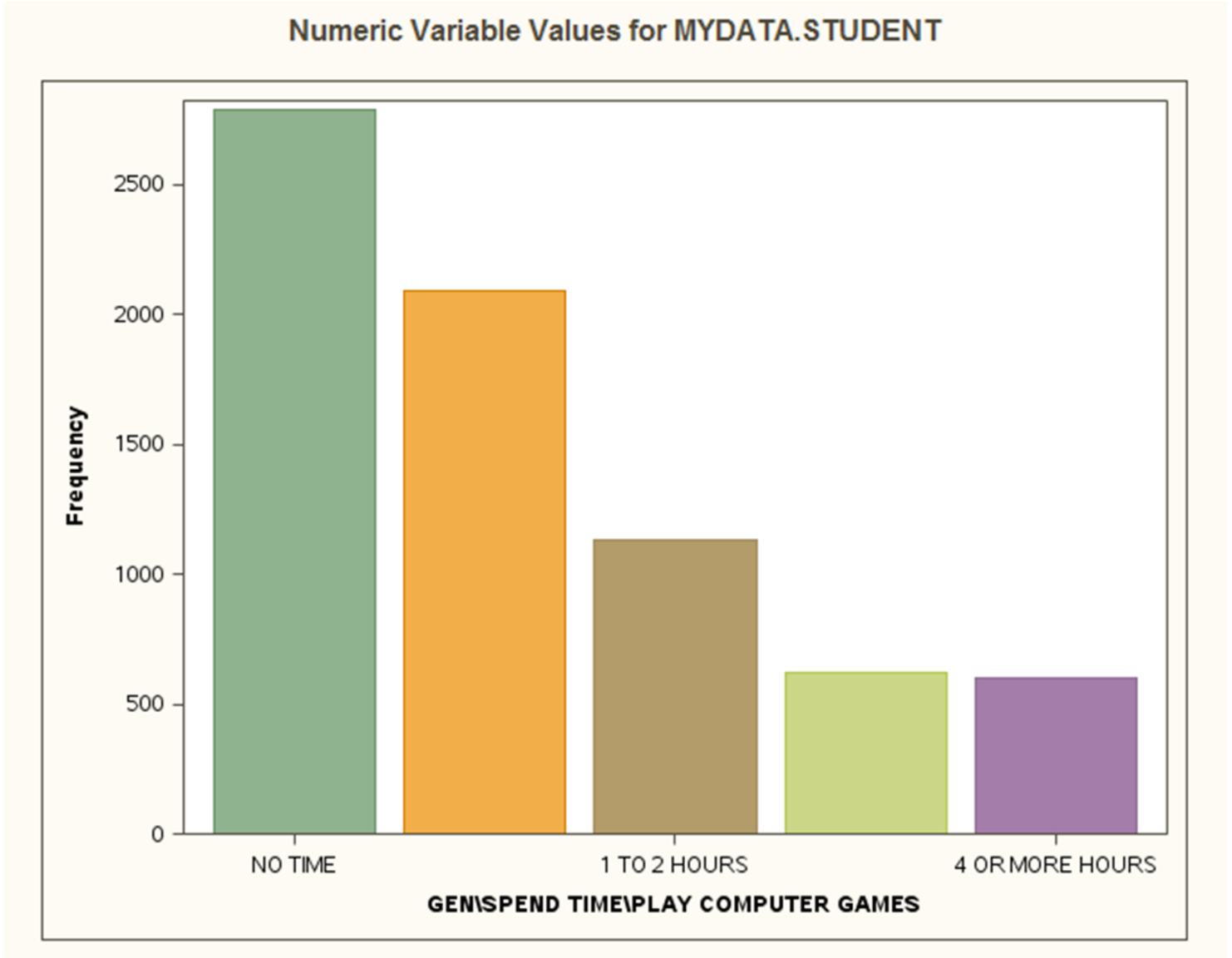
II YEAR /IV SEMESTER

Unit 1- Introduction

Topic : Descriptive Univariate Analysis


Injuries by County
Number of Injuries
0
1
2
3
4
5
6
County
County A
County B
County C


Injuries by County
County A
County B
County C


Numeric Variable Values for MYDATA.STUDENT
Frequency
2500
2000
1500
1000
500
0
NO TIME
1 TO 2 HOURS
4 OR MORE HOURS
GEN\SPEND TIME\PLAY COMPUTER GAMES

Descriptive Univariate Analysis

- ✓ frequency tables
- ✓ statistical measures
- ✓ plots

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.

Table 2.2 Univariate absolute and relative frequencies for "company" attribute.

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

Univariate absolute and relative frequencies for height.

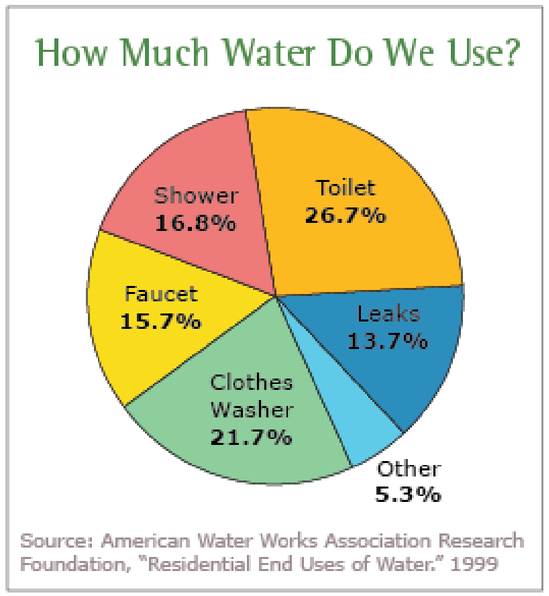
Height	Abs. freq.	Rel. freq.	Abs. cum. freq.	Rel. cum. freq.
158	1	1/14 = 7.14%	1	7.14%
163	1	7.14%	2	14.28%
165	1	7.14%	3	21.42%
168	1	7.14%	4	28.56%
172	2	14.29%	6	42.85%
173	1	7.14%	7	49.99%
175	1	7.14%	8	57.13%
180	2	14.29%	10	71.42%
185	1	7.14%	11	78.56%
190	1	7.14%	12	85.70%
192	1	7.14%	13	92.84%
195	1	7.14%	14	99.98%

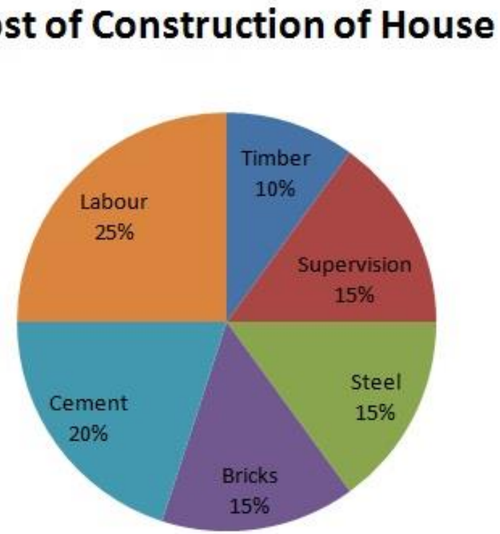
Descriptive Statistics

- Distribution functions from populations can be either probability distribution functions or probability density functions, depending on the data type of the attribute.
- A discrete attribute, such as the integer data type, has a probability mass function, while a continuous attribute, such as the real data type, has a probability density function.
- The reason for this distinction is that in a continuous space the probability of being an exact value is zero

Univariate Data Visualization

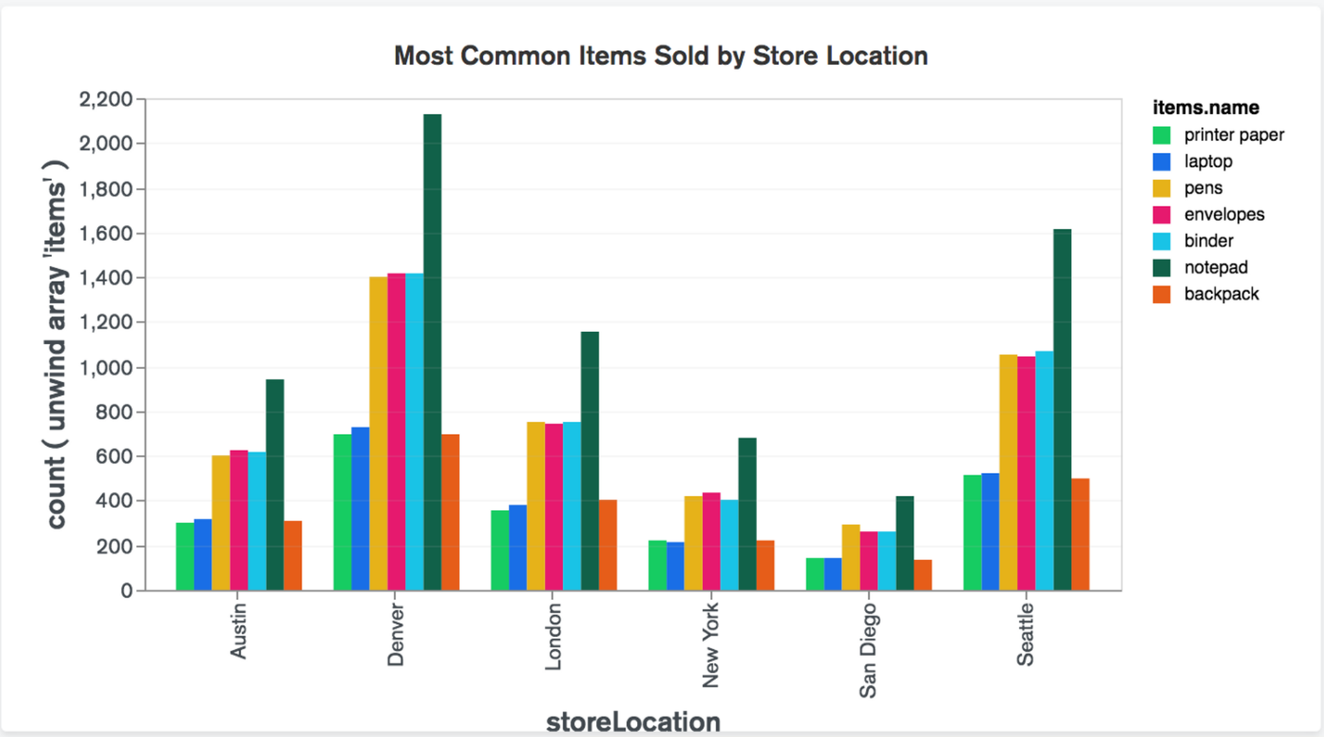
- **Pie chart** These are used typically for nominal scales. It is not advisable to use them with scales where the notion of order exists – in other words for ordinal and quantitative scales – although this is possible





Univariate Data Visualization

- **Bar charts** These are used typically for qualitative scales. When the notion of order exists, the classes should be displayed in the horizontal bar, typically in increasing order of magnitude.
- Many authors argue that bar charts are better for comparing values between different classes than pie charts because it is easier to see that one bar is bigger than another than to see that one pie slice is larger than another



Univariate Data Visualization

- **Line charts** Like area charts, these are used when the horizontal bar uses a quantitative scale with equal lag between observations. In particular, they are used to deal with the notion of time

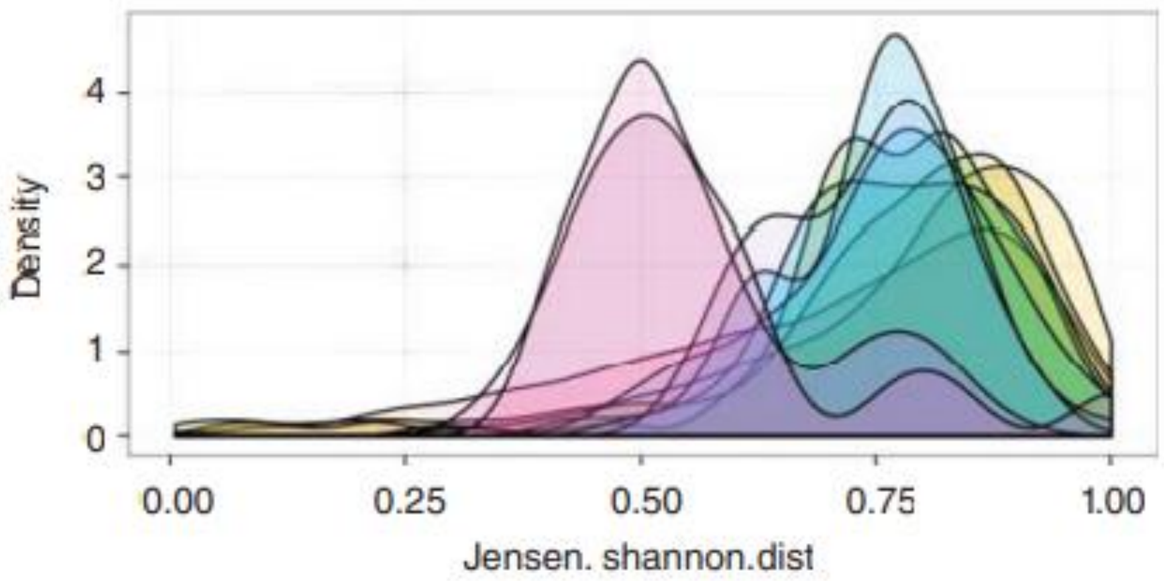


Figure 2.3 An example of an area chart used to compare several probability density functions.

Univariate Data Visualization

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

Understanding data distributions give us strong insights about an attribute. We are able to see, for instance, that data are more concentrated in some values or that other values are rare

Univariate Data Visualization

- **Histograms** These are used to represent empirical distributions for attributes with a quantitative scale. Histograms are characterized by grouping values in cells, reducing in this way the sparsity that is common in quantitative scales. You can see in that the histogram is more informative than the bar chart

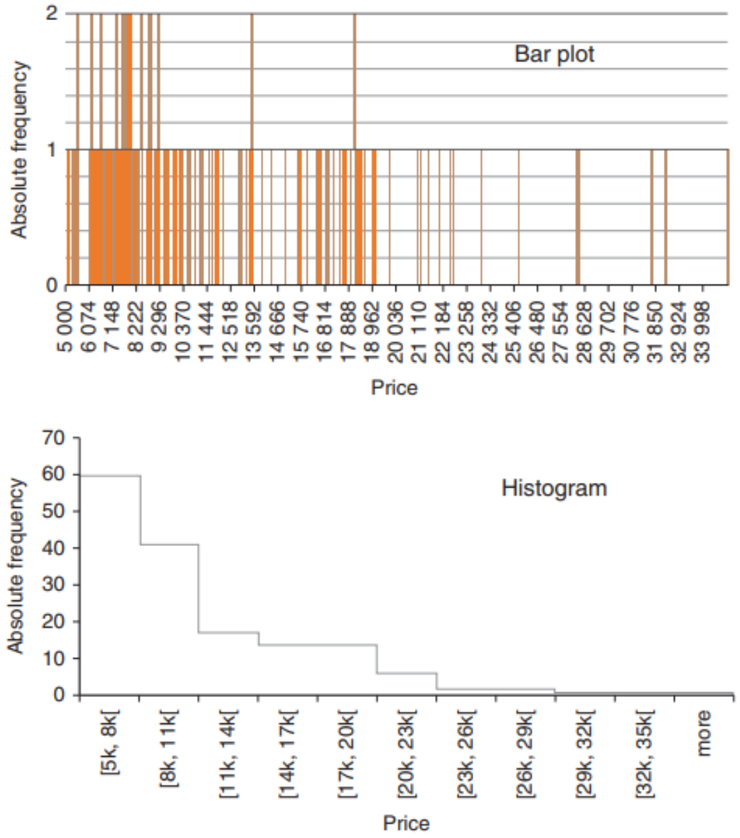


Figure 2.4 Price absolute frequency distributions with (histogram) and without (bar chart) cell definition.

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 and dispersion statistics.

- **Location univariate statistics** Location statistics identify a value in a certain position. Some well known location univariate statistics are the minimum, the maximum or the mean.

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

Dispersion univariate statistics

- A dispersion statistic 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
- median or second quartile: the value that is larger than 50% of all values; the value that splits the sequence into two equal-sized sub-sequences
- Third quartile: the value that is larger than 75% of all values.

Dispersion univariate statistics

Table 2.5 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

Table 2.6 Central tendency statistics according to the type of scale.

	Nominal	Ordinal	Quantitative
Mean	No	Eventually*	Yes
Median	No	Yes	Yes
Mode	Yes	Yes	Yes

*See below.

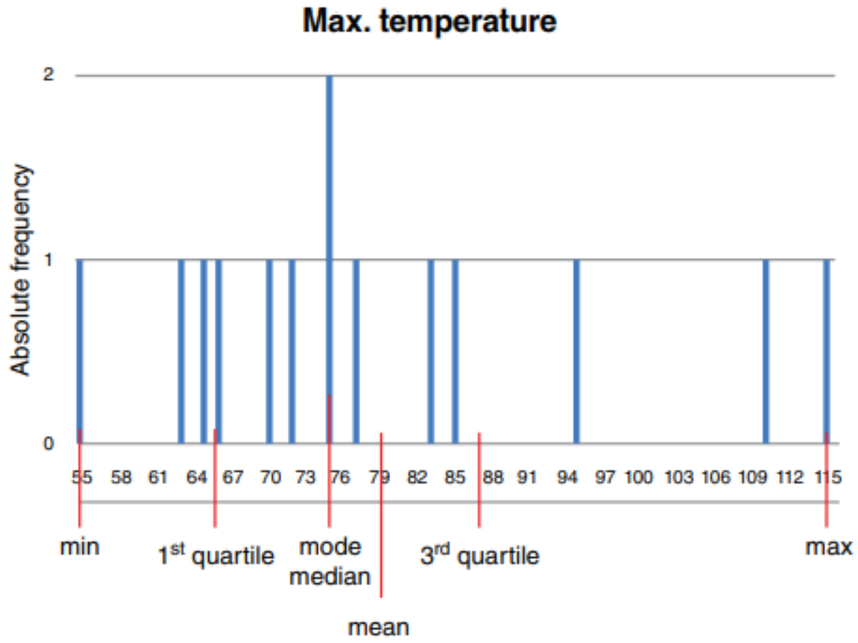


Figure 2.7 Location statistics on the absolute frequency plot for the attribute "weight".

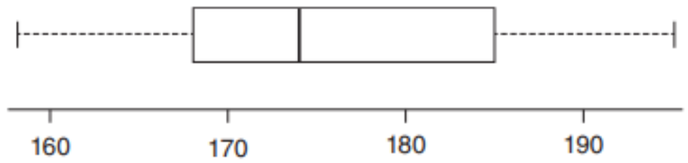


Figure 2.8 Box-plot for the attribute "height".

Dispersion univariate statistics

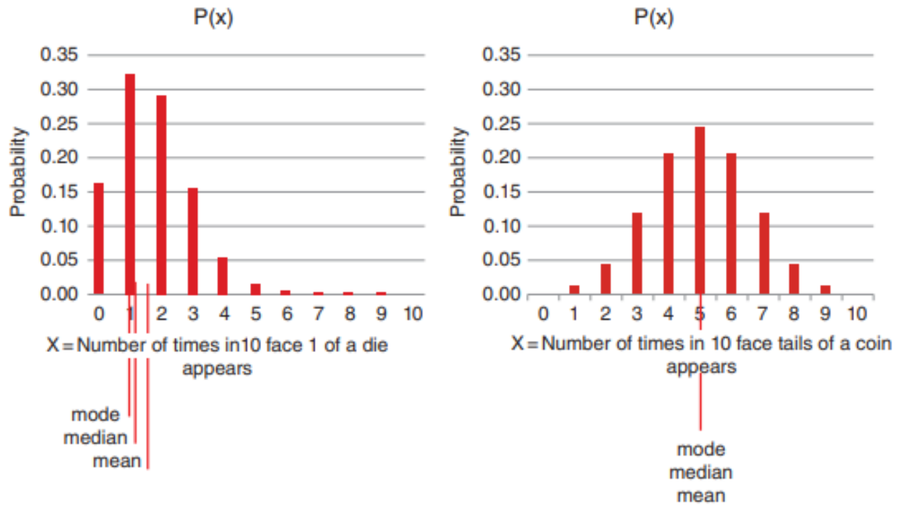


Figure 2.9 Central tendency statistics in asymmetric and symmetric unimodal distributions.

Please circle the number that better fits your experience with the given information

I am satisfied with it
Strongly disagree 1 2 3 4 5 Strongly agree

It is simple to use
Strongly disagree 1 2 3 4 5 Strongly agree

It has good graphics
Strongly disagree 1 2 3 4 5 Strongly agree

It is in accordance to my expectations
Strongly disagree 1 2 3 4 5 Strongly agree

Everything make sense
Strongly disagree 1 2 3 4 5 Strongly agree

Figure 2.10 An example of the Likert scale.

Dispersion univariate statistics

- A dispersion statistic 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. $MAD_x = \frac{\sum_{i=1}^{n} |x_i - \mu_x|}{n},$
- standard deviation: another measure for the typical distance between the observations and their mean $\sigma_x = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu_x)^2}{n}},$

Dispersion univariate statistics

Table 2.7 Dispersion univariate statistics for the "weight" attribute.

Dispersion statistic	Weight (kg)
Amplitude	60.00
Interquartile range	21.75
$\overline{MAD}$	14.31
s	17.38

$$\overline{MAD}_x = \frac{\sum_{i=1}^{n} |x_i - \overline{x}|}{n-1}, \tag{2.3}$$

and

$$s_x = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}}, \tag{2.4}$$

The sample variance is denoted as s^2 and is, as expected, the square of s.

Common Univariate Probability Distributions

- The uniform distribution is a very simple distribution.
- The frequency of occurrence of the values is uniformly distributed in a given interval of values. An attribute x that follows a uniform distribution with parameters a and b, respectively the minimum and maximum values of the interval, is denoted as:

$$x \sim \mathcal{U}(a, b)$$

Common Univariate Probability Distributions

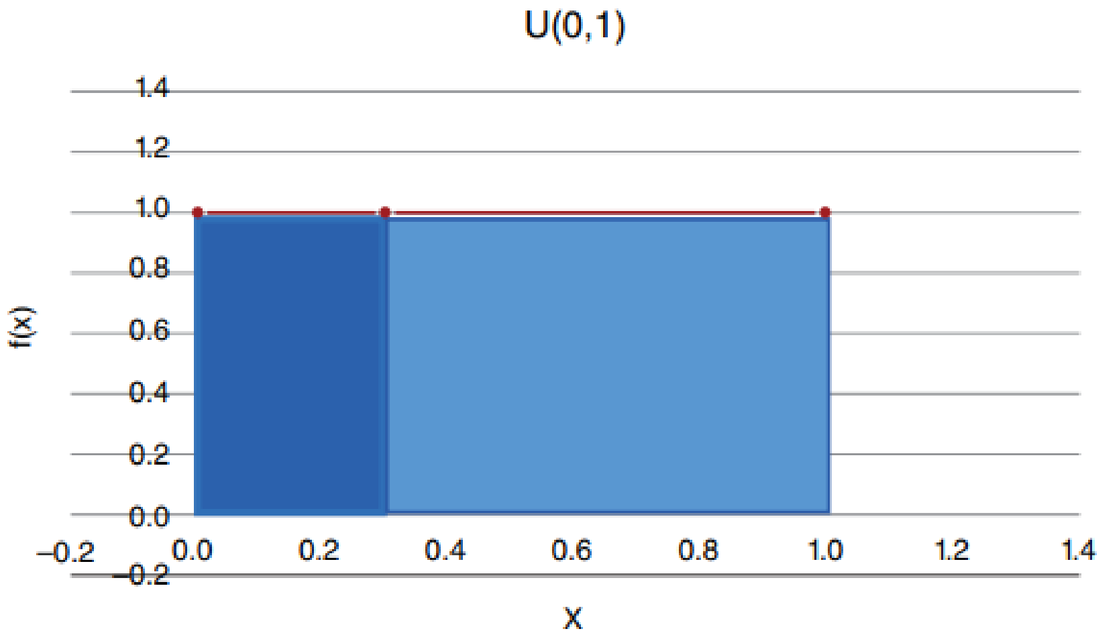


Figure 2.12 The probability density function, $f(x)$ of $x \sim \mathcal{U}(0, 1)$.

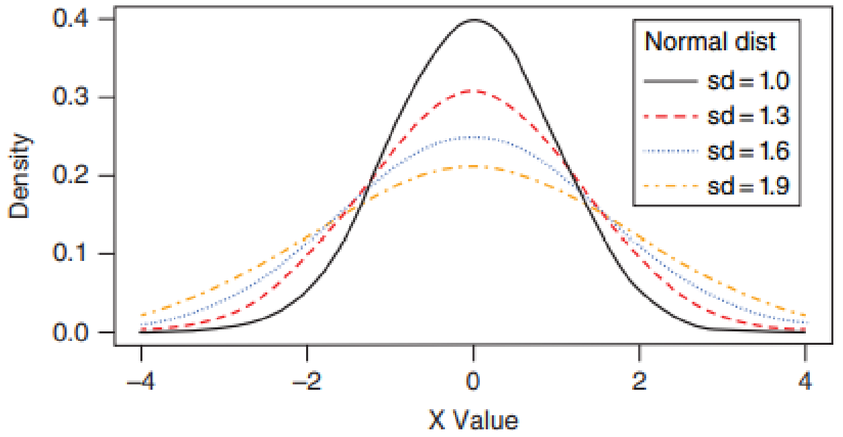


Figure 2.13 The probability density function for different standard deviations, $\mathcal{N}(0, \sigma = sd)$.

Common Univariate Probability Distributions

- ✓ **The normal distribution**, also known as Gaussian distribution, is the most common distribution, at least for continuous attributes.
- ✓ This happens due to an important theorem in statistics, known as the central limit theorem, which is the basis of many of the methods used in inductive learning.
- ✓ Physical quantities that are expected to be the sum of many independent factors (say, people's heights or the perimeter of 30-year-old oak trees) typically have approximately normal distributions.
- ✓ The normal distribution is a symmetric and continuous distribution

Assessment 1

To Analyze dataset using Univariate analysis

References

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

Thank You