INTRODUCTORY STATISTICS

STATS101
SINGAPORE MANAGEMENT UNIVERSITY
Tutor: Ravins Cheung (A+)

Module 1: Introduction
1.Tabulating and graphing data distributions
2.Measures of dispersion
3.Measures of central tendency
4.Mean and variance for grouped/sample data
5.Summation Notation
6.Recap

ZOOKAL

Tabulating and graphing data distributions

Objectives

- Helps to organize and present data
- Describes analysis of statistics

ZOOKAL

Tabulating Categorical Data

Representing data that has been categorized in a table form.

30 people were randomly surveyed, and their responses were collated

Marital Status	Absolute Freq	Relative Freq
Single	13	0.325
Married	18	0.450
Others	9	0.225

ZOOKAL

Tabulating Categorical Data

Test scores of 20 students were collated as below.

12	**13**	**17**	**21**	**24**	**24**	**26**	**27**	**27**	**30**
32	**35**	**37**	**38**	**41**	**43**	**44**	**46**	**53**	**58**

Class	Frequency	Relative Frequency	Cumulative Frequency
(10,20]	3	0.15	0.15
(20,30]	7	0.35	0.5
(30,40]	4	0.2	0.7
(40,50]	4	0.2	0.9
(50,60]	2	0.1	1

Note: (10; 20] refers to values between 10 (exclusive) and 20 (inclusive) etc.

ZOOKAL

Graphing Data Distributions

Represents data (raw/tabulated) in graphs for ease of interpretation

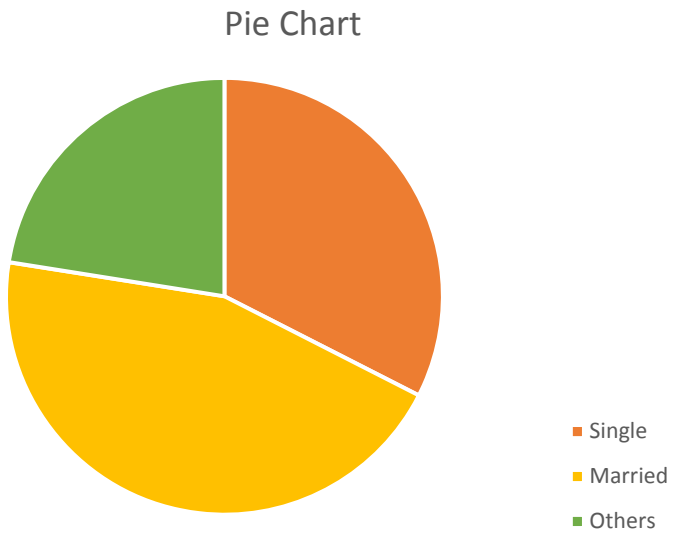
Types of graphs

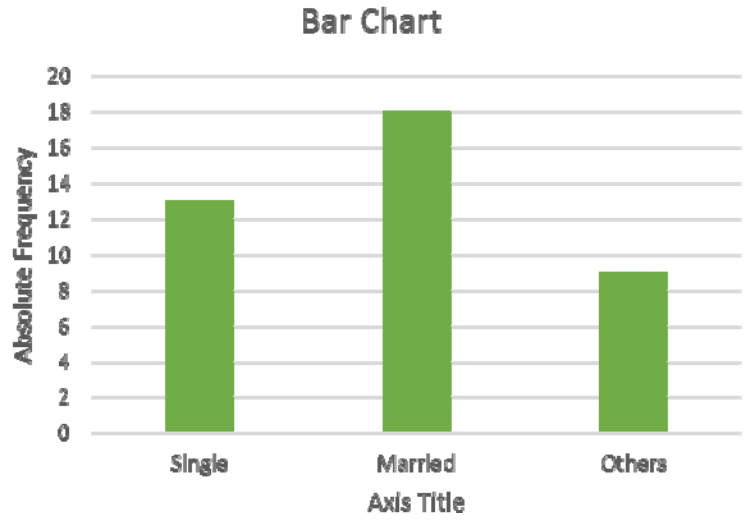
1. Pie Chart
2. Bar Chat
3. Side-by-side chart
4. Absolute frequency histogram
5. Relative frequency histogram
6. Cumulative frequency polygon
7. Stem and leaf diagram
8. Boxplots

ZOOKAL

Graphing Data Distributions

Marital Status	Absolute Freq	Relative Freq
Single	13	0.325
Married	18	0.450
Others	9	0.225



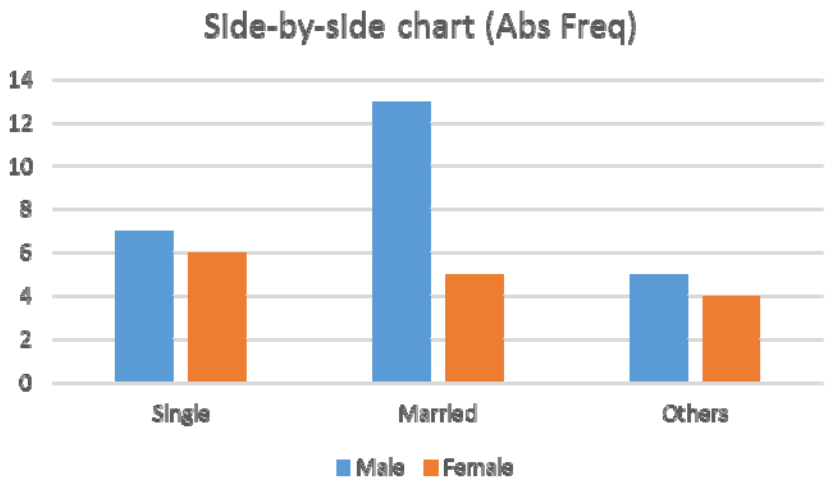


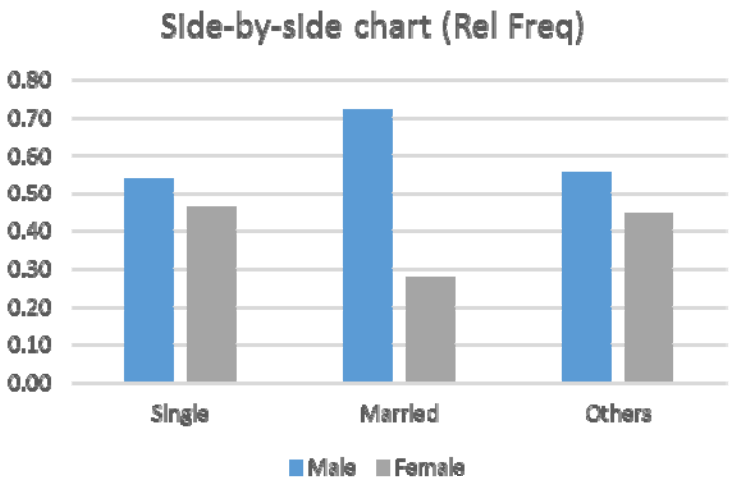
ZOOKAL

Graphing Data Distributions

Marital Status	Male	Female
Single	7	6
Married	13	5
Others	5	4

Marital Status	Male	Female
Single	0.54	0.46
Married	0.72	0.28
Others	0.56	0.44

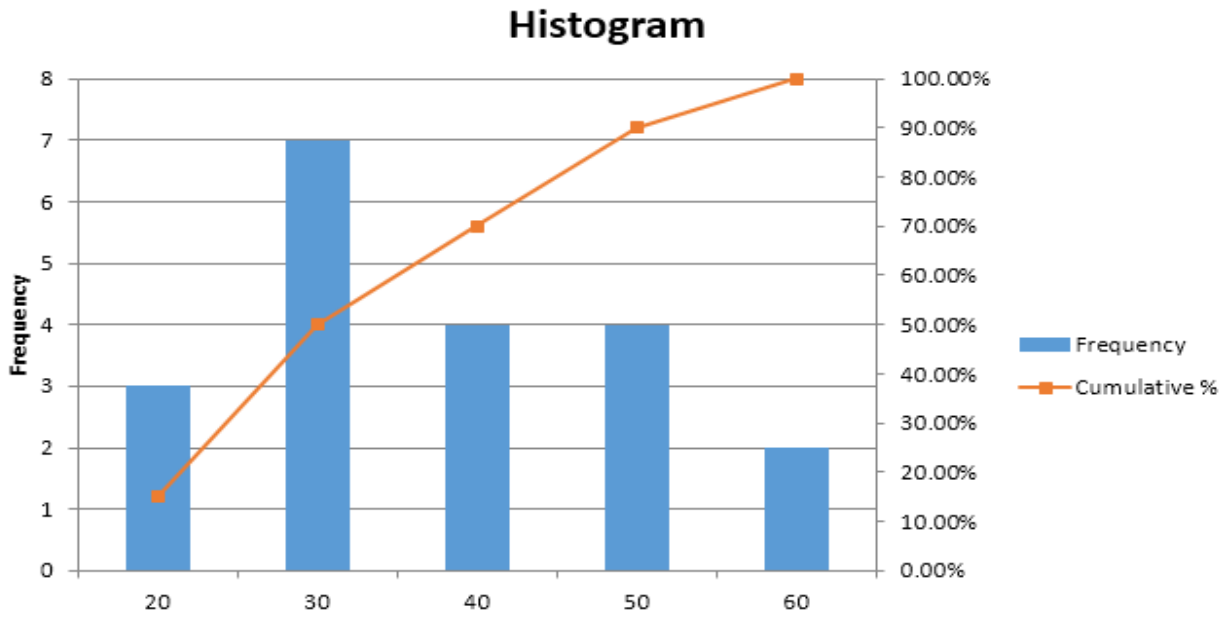




ZOOKAL

Graphing Data Distributions

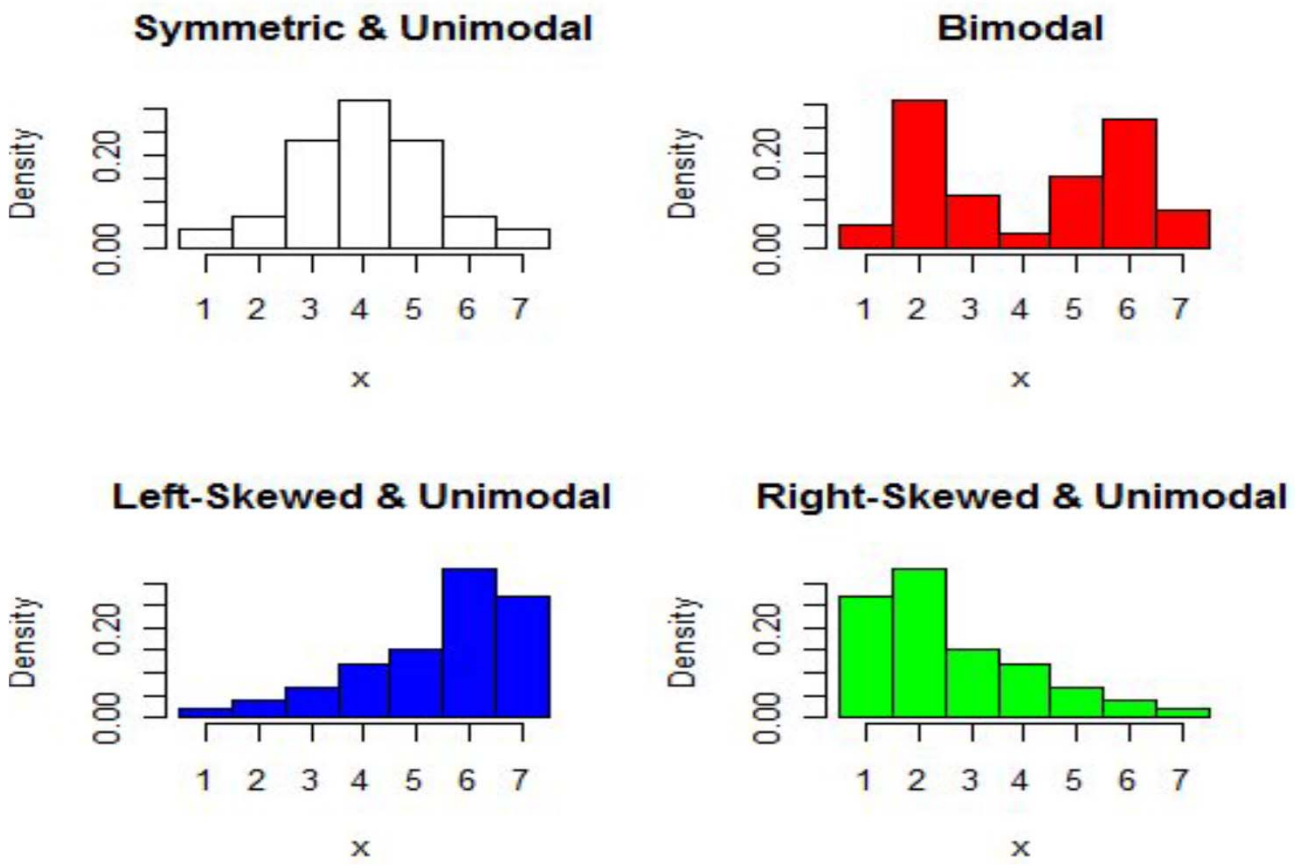
Class	Frequency	Relative Frequency	Cumulative Frequency
(10,20]	3	0.15	0.15
(20,30]	7	0.35	0.5
(30,40]	4	0.2	0.7
(40,50]	4	0.2	0.9
(50,60]	2	0.1	1



ZOOKAL

Graphing Data Distributions

Identifying Distribution Shapes



ZOOKAL

Graphing Data Distributions

Quartiles

$$
\begin{aligned}
Q2 &= \text{second quartile of the } x_i\text{'s} \\
&= \begin{cases} \frac{1}{2}\left(x_{(k)} + x_{(k+1)}\right), & \text{if } n = 2k, \\ x_{(k+1)}, & \text{if } n = 2k+1. \end{cases}
\end{aligned}
$$

Note that $Q2$ is also referred to as the median of the x_i's.

The first quartile, denoted $Q1$, *may be defined* as the median of x_i values less than or equal to $Q2$.

The third quartile, denoted $Q3$, *may be defined* as the median of x_i values greater than or equal to $Q2$.

ZOOKAL

Graphing Data Distributions

Survey of travelling time to school was conducted on 25 students

12	19	3	13	32
33	25	8	15	30
19	19	4	5	24
21	32	17	10	2
35	11	18	10	9

0	2 3 4 5 8 9
1	0 0 1 2 3 5 7 8 9 9 9
2	1 4 5
3	0 2 2 3 5

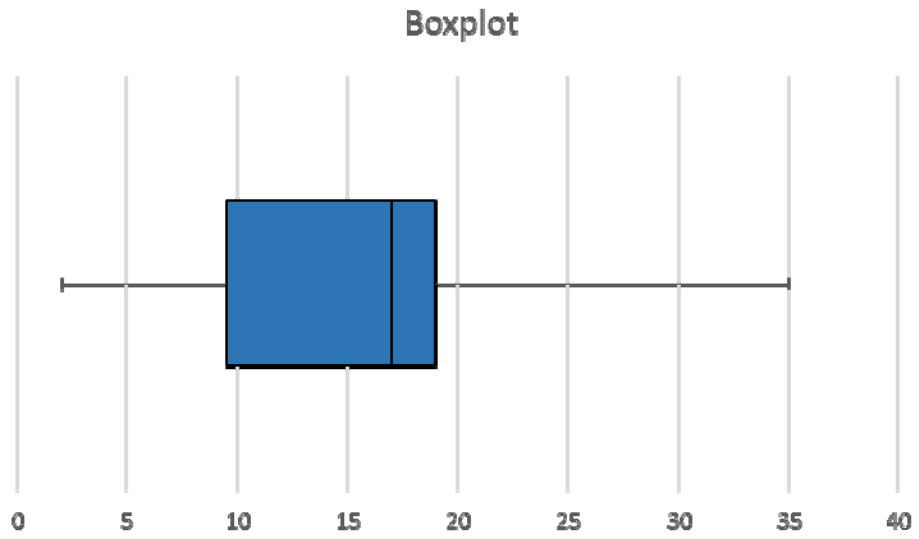
Q1=9.5 , Q2=17, Q3=19, Interquartile Range (IQR)=9.5

ZOOKAL

Graphing Data Distributions

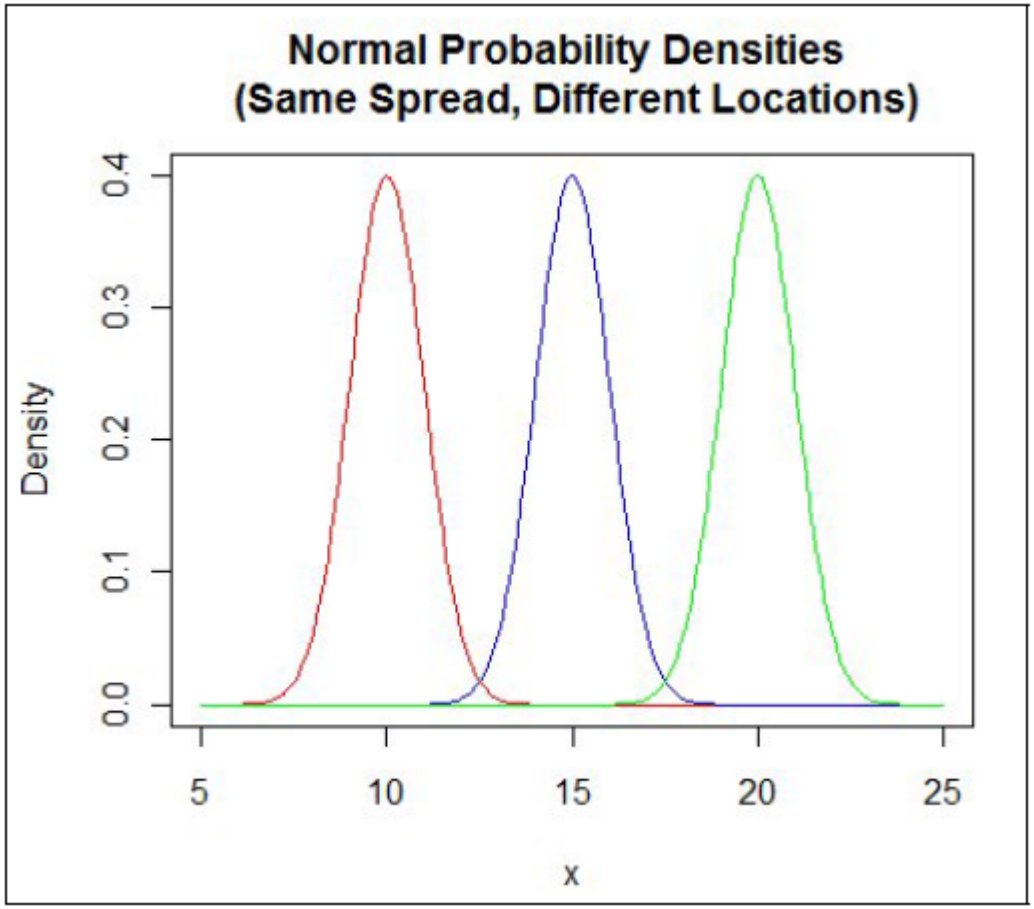
Boxplot

Q1=9.5 , Q2=17, Q3=19, Interquartile Range (IQR)=9.5, min=2, max=35, lower limit = Q1-1.5 x IQR, upper limit = Q3 + 1.5 x IQR



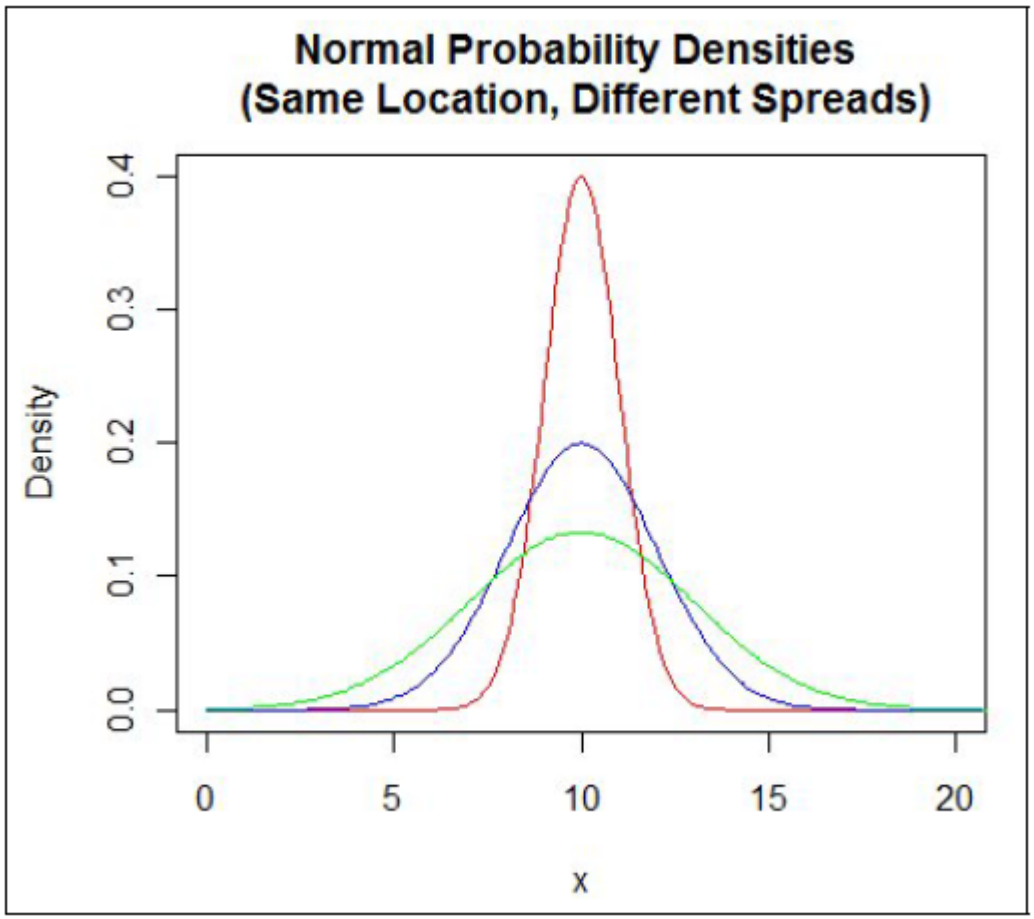
ZOOKAL

Measures of dispersion



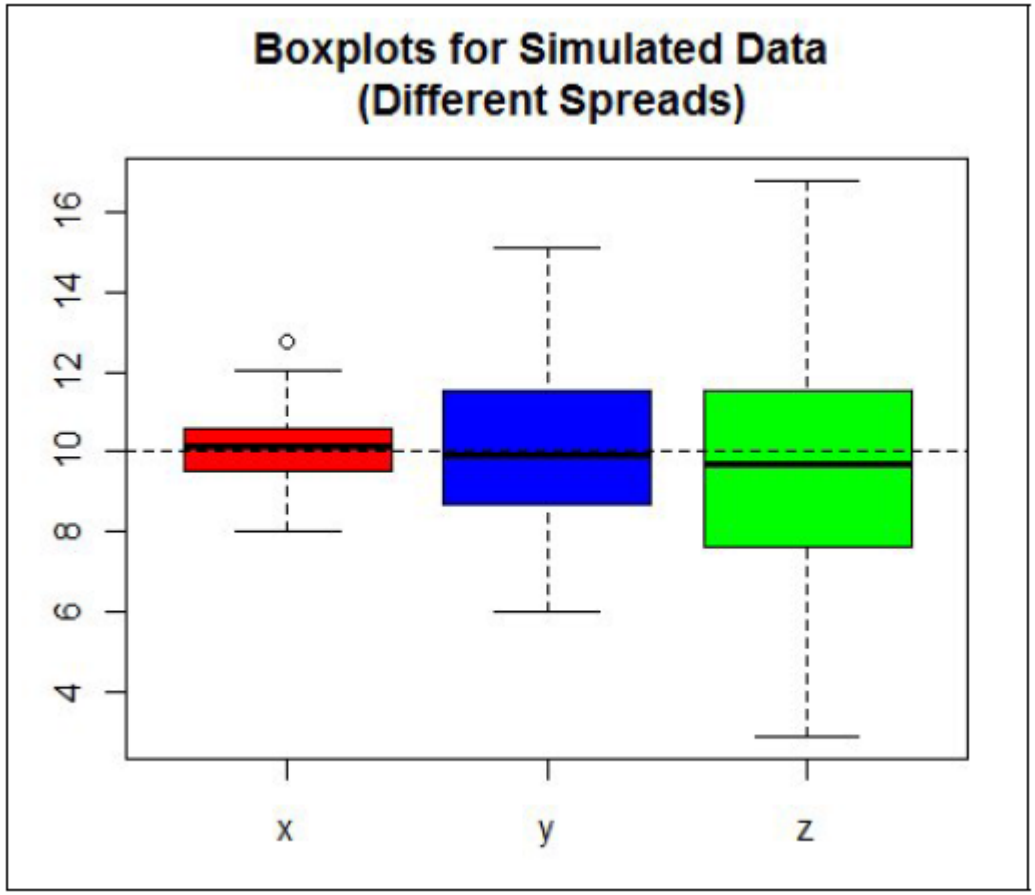
ZOOKAL

Measures of dispersion



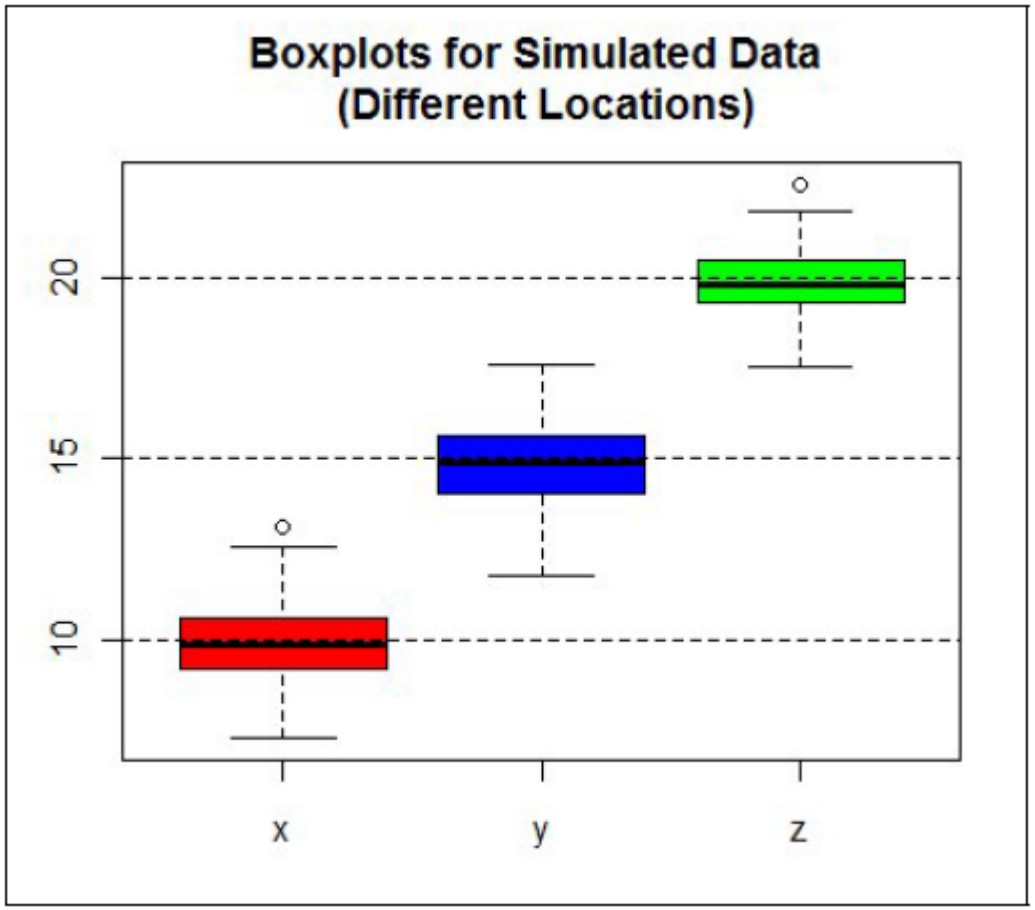
ZOOKAL

Measures of dispersion



ZOOKAL

Measures of dispersion



ZOOKAL

Measures of Central Tendency

- *Mean*

$$\text{mean} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}, \text{ say.}$$

- *Median*

$$\text{median} = \begin{cases} \frac{1}{2}\left(x_{(k)} + x_{(k+1)}\right), & \text{if } n = 2k, \\ x_{(k+1)}, & \text{if } n = 2k+1. \end{cases}$$

- *Mode*

$$\text{mode} = \text{data value with highest frequency.}$$

ZOOKAL

Measures of Central Tendency

0	2 3 4 5 8 9
1	0 0 1 2 3 5 7 8 9 9 9
2	1 4 5
3	0 2 2 3 5

Median=17, Mean = 17.04, Mode = 19

ZOOKAL

Measures of Central Tendency

Advantages & Disadvantages

Feature	Mean	Median	Mode
Always Exists?	Y	Y	N
Always Unique?	Y	N	N
Not Affected by Outliers?	N	Y	Y
Further Analysis Potential?	Y	N	N

Note

- Use a robust (i.e., resistant) measure of central tendency when outlying values (assuming these are valid) are present.
- The trimmed mean is an example of a robust measure of location

ZOOKAL

Measures of Dispersion

- *Range*

$$\text{range} = x_{(n)} - x_{(1)}$$

- *Interquartile Range*

$$\text{IQR} = \text{Third Quartile} - \text{First Quartile}$$

- *Variance*

$$\text{variance} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

- *Standard Deviation*

$$\text{standard deviation} = \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right)}$$

A relative measure of dispersion is

$$\text{coefficient of variation} = \frac{\text{standard deviation}}{\text{mean}}.$$

ZOOKAL

Measures of Dispersion

0	2 3 4 5 8 9
1	0 0 1 2 3 5 7 8 9 9 9
2	1 4 5
3	0 2 2 3 5

Range=33, Interquartile Range (IQR)=9.5

$$\text{Variance} = \frac{(2^2+2^2+\cdots+35^2) - 25*17.4^2}{25-1} \approx 87.04$$

$$\text{Standard Deviation} \approx \sqrt{87.04} \approx 9.33$$

$$\text{Coefficient of variation} = \frac{9.33}{17.4} \approx 0.54$$

ZOOKAL

Measures of Dispersion

Advantages & Disadvantages

Feature	R	V	SD	IQR	CV
Always Exists?	Y	Y	Y	Y	Y
Always Unique?	Y	N	N	N	N
Not Affected by Outliers?	N	N	N	Y	N
Absolute Measure?	Y	Y	Y	Y	N
Same Units?	Y	N	Y	Y	N

ZOOKAL

Mean & Variance for Grouped Data

Group data refers to data in a frequency distribution

- Let

$$\begin{aligned} m_i &= \text{mid-point of } i\text{-th class}, \\ n_i &= \text{frequency of } i\text{-th class}, \\ k &= \text{number of classes}, \\ n &= \text{total frequency}. \end{aligned}$$

- The grouped data mean is

$$\overline{x}_g = \sum_{i=1}^{k} m_i \times \frac{n_i}{n} = \frac{1}{n} \sum_{i=1}^{k} m_i n_i.$$

- The grouped data variance is

$$s_g^2 = \sum_{i=1}^{k} m_i^2 \times \frac{n_i}{n} - \overline{x}_g^2 = \frac{1}{n} \sum_{i=1}^{k} m_i^2 n_i - \overline{x}_g^2.$$

ZOOKAL

Mean & Variance for Grouped Data

Class	ni	mi	ni*mi	ni* mi^2
(10,20]	3	15	45	675
(20,30]	7	25	175	4375
(30,40]	4	35	140	4900
(40,50]	4	45	180	8100
(50,60]	2	55	110	6050
Sum	20		650	24100

$$\bar{x}_g = \frac{650}{20} = 32.5$$

$$S_g{}^2 = \frac{24100}{20} - 32.5^2 = 148.75$$

ZOOKAL

Summation Notation

Summation Notation

Given numerical values $x_1, \ldots, x_n$, we have:

- $\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n$
- $\sum_{i=1}^{n} (ax_i + b) = (ax_1 + b) + \cdots + (ax_n + b) = a\sum_{i=1}^{n} x_i + nb$

Example

If x_i's are given by $1.75, 2.25, 2.25, 2.25, 1.75, 2.00, 1.50$, we have

$$\sum_{i=1}^{7} x_i = 13.75 \text{ and } \sum_{i=1}^{7} x_i^2 = 1.75^2 + \cdots + 1.50^2 = 27.5625.$$

ZOOKAL

Recap

Module 1: Introduction
Tabulating and graphing data distributions
Measures of dispersion
Measures of central tendency
Mean and variance for grouped/sample data
Summation Notation
Recap

Next Module:
Module 2: Discrete Distribution Problems

ZOOKAL