CHAPTER 2

Data Description

Prepared & updated by: Razana/Suhaila/Nzah

Types of Data Set

- Univariate
- Multivariate
  - o Bivariate

Univariate Data Set

- Univariate data set consists of observations on a single variable made on individuals in a sample or population.
- **Variable** is any characteristic whose value may change from one individual or object to another.
- A univariate data set is **categorical** (or qualitative) if the individual observations are categorical response.
  - Example: recorded calculator brand.
- A univariate data set is **numerical** (or quantitative) if each observation is a number.
  - Example: recorded number of calculator purchased.

Multivariate Data Set

- Multivariate data set consists of observations on two or more variables made on individuals in a sample or population.
  - Example: recorded height, weight, pulse rate and systolic blood pressure for each individual in a group.
- A bivariate data are special case of multivariate data set because it focus simultaneously on only two variables or two different characteristics.
  - Example: recorded both height and weight for each individual in a group.

Displaying Categorical Data

- An appropriate graphical or tabular of data can be an effective way to summarize and communicate information.
- Example:
  - Frequency distribution or frequency table
  - Bar chart
  - Pie chart
  - Dot-Plot

Frequency Distribution

- A frequency distribution for categorical data is a table that displays the possible categories along with the associated frequencies and/or relative frequencies.
- The **frequency** for a particular category is the number of times the category appears in the data set.
- The **relative frequency** for a particular category is the fraction or proportion of the observations resulting in the category.

$$\text{Relative frequency} = \frac{\text{frequency}}{\text{Total no. of cases}}$$

- Frequency distributions are portrayed as **frequency tables, dot plot, histograms,** or **polygons**.

Example

Data – Staff info at School of Computing

Staff	Position	Blood Type	Weight	Height	Qualification
1	Senior Lecturer	A	60	165	PhD
2	Lecturer	B	55	150	Master
3	Professor	O	65	170	PhD
4	Associate Professor	AB	70	175	PhD
5	Associate Professor	O	61	160	PhD
6	Senior Lecturer	O	58	155	PhD
7	Senior Lecturer	B	48	167	PhD
8	Lecturer	A	68	174	Master
9	Lecturer	A	55	150	Master
10	Associate Professor	AB	62	163	PhD
11	Professor	O	58	165	PhD
...	...	...	...	...	...
140	Tutor	O	45	150	Master

Example

Frequency Table: Staff distribution in School of Computing

Position	Number of Staff (Frequency)
Professor	12
Associate Professor	20
Senior Lecturer	59
Lecturer	40
Tutor	9
Total	140

Example

Position	Frequency	Relative Frequency
Professor	12	12 ÷ 140 = 0.09
Associate Professor	20	0.14
Senior Lecturer	59	0.42
Lecturer	40	0.29
Tutor	9	0.06
Total	140	1.00

Example

- Data

Age	NetUse	Age	NetUse	Age	NetUse
26.0	Yes	45.0	No	55.0	No
48.0	Yes	19.0	No	37.0	No
67.0	Yes	82.0	No	43.0	Yes
44.0	No	83.0	No	29.0	Yes
52.0	No	20.0	Yes	57.0	Yes
52.0	No	89.0	No	36.0	No
51.0	Yes	88.0	Yes	52.0	No
52.0	No	72.0	Yes	56.0	Yes
77.0	No	82.0	Yes	66.0	Yes
40.0	No	34.0	No	46.0	No

Example

- List the variables involved and calculate the frequency

Response	Frequency
No	17
Yes	13

- Other types of response with missing value:
  - DK (Don't know)
  - NAP (Not applicable)
  - NA (No answer/response)

Example

- Calculate the percent, valid percent and cumulative percent

Response	Frequency	Percent	Valid Percent	Cumulative Percent
No	17	56.7	56.7	56.7
Yes	13	43.3	43.3	100.0
Total	30	100.0	100.0	

- Valid percent: excludes data with missing value (ie. DK, NAP, and NA)

Exercise

The data below represent the color of M&Ms in a bag of plain M&Ms. Construct a frequency distribution of the color of plain M&Ms.

Yellow	Orange	Brown	Green	Green
Blue	Brown	Red	Brown	Brown
Orange	Brown	Red	Brown	Red
Green	Brown	Red	Green	Yellow
Yellow	Red	Red	Brown	Orange
Yellow	Orange	Red	Orange	Blue
Brown	Red	Yellow	Brown	Red
Brown	Yellow	Yellow	Blue	Yellow
Yellow	Brown	Yellow	Green	Orange

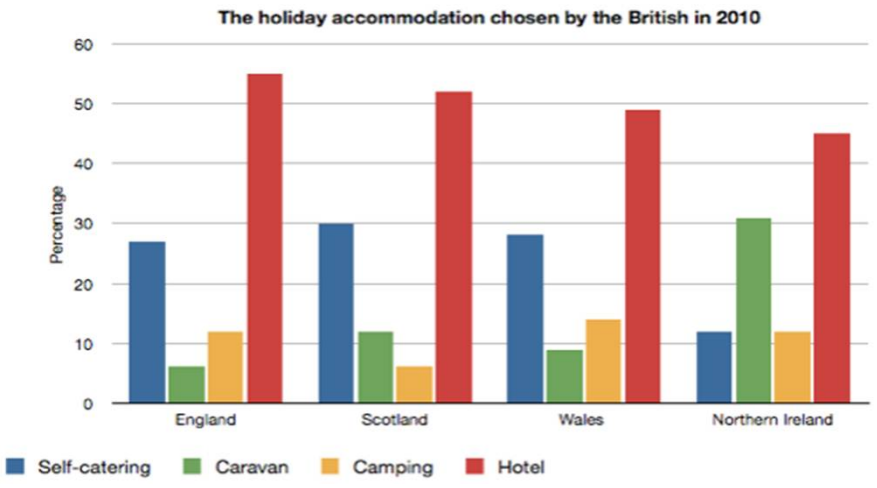
Solution

Color Type	Frequency	Relative Frequency
Yellow	10	0.22
Blue	3	0.07
Orange	6	0.13
Green	5	0.11
Brown	12	0.27
Red	9	0.20
Total	45	1.00

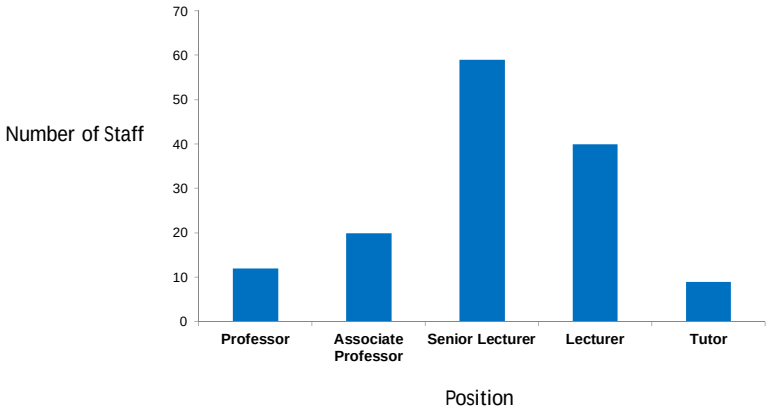
Bar Chart

- A bar chart is a graph of the frequency distribution of categorical data.
- Each category in the frequency distribution is represented by a bar or rectangle.
- How to construct:
  - Draw a horizontal line, and write the category names or labels below the line at regularly spaced intervals.
  - Draw a vertical line, and label the scale using either frequency or relative frequency.
  - Place a rectangular bar above each category label. The height is determined by the category's frequency or relative frequency, and all bars should have the same width.
- What to look for:
  - Frequently and infrequently categories.

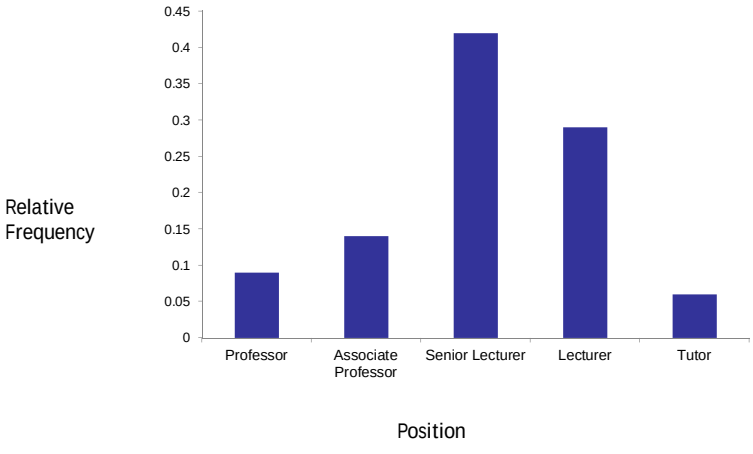
Example



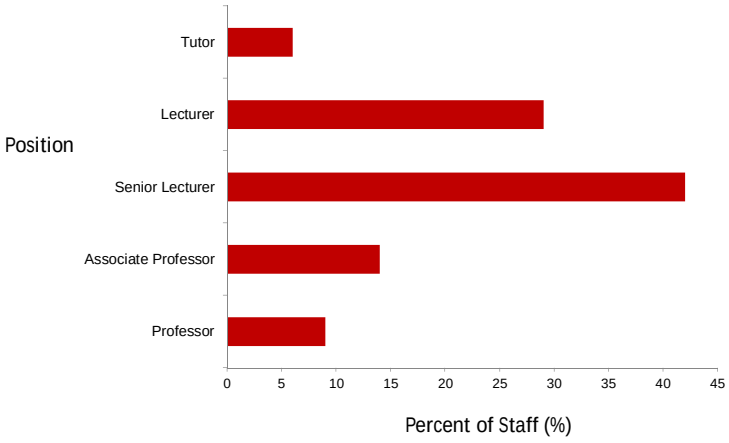
Example



Example

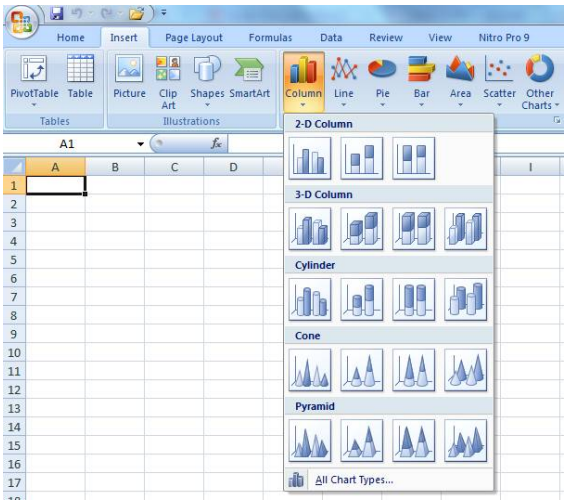


Example



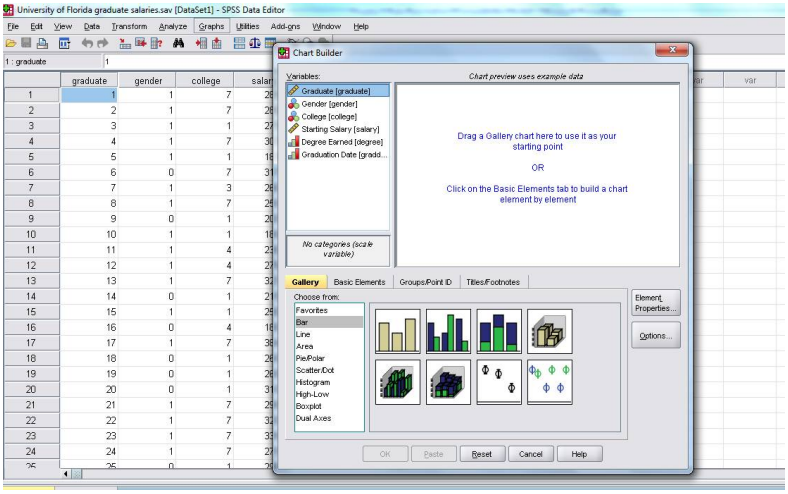
Statistical Tools

- Excel:



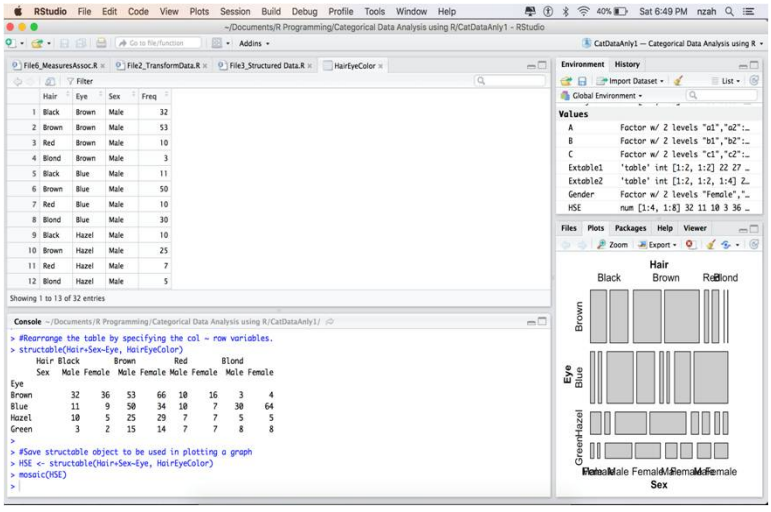
Statistical Tools

- SPSS:



Statistical Tools

- *R* :



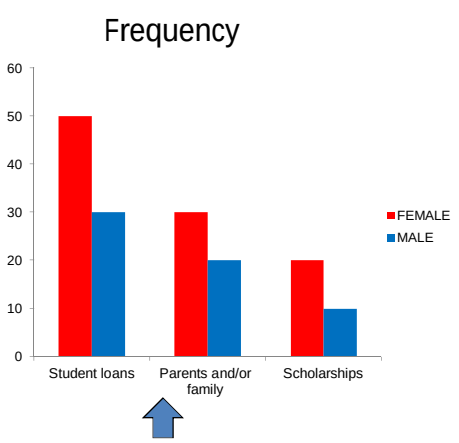
Comparative Bar Charts

- Bar chart can also be used to give a visual comparison of two or more groups.
- When constructing a comparative bar graph we use the relative frequency rather than the frequency to construct the scale on the vertical axis so that we can make meaningful comparisons even if the sample sizes are not the same.

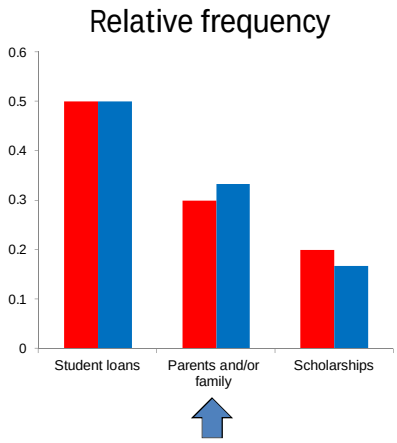
Example

Source of Funding	Frequency		Relative Frequency	
	Female	Male	Female	Male
Student Loans	50	30	0.5	0.50
Parents and/or family	30	20	0.3	0.33
Scholarships	20	10	0.2	0.17
Total	**100**	**60**	**1.00**	**1.00**

Example



Inaccurate comparative bar chart



Accurate comparative bar chart

Pie Charts

- In a pie chart, a circle is used to represent the whole data set, with "slices" of the pie representing the possible categories.
- The size of the slice for a particular category is proportional to the corresponding frequency or relative frequency.
- Pie charts are most effective for summarizing data sets when there are not too many different categories.

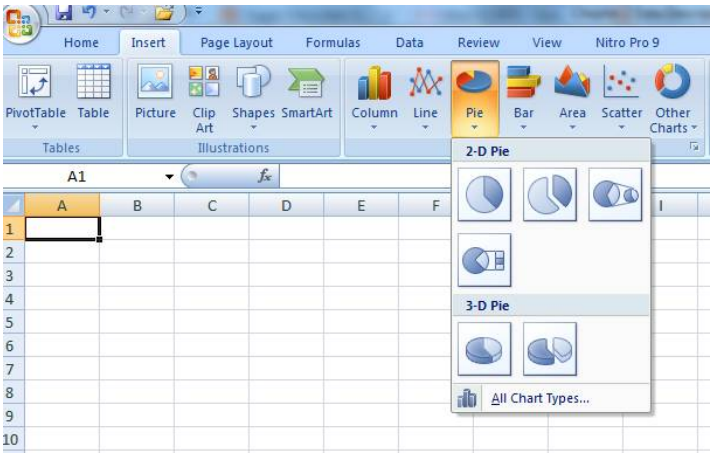
Pie Charts

- How to construct
  - Draw a circle to represent the entire data set
  - For each category, calculate the "slice" size.
    slice size = 360 (category relative frequency)
  - Draw a slice of appropriate size for each category.

- What to look for
  - Categories that form large and small proportions of the data set.

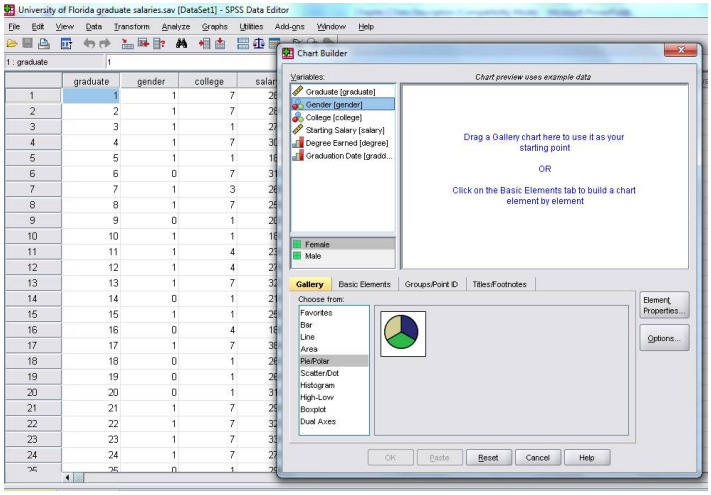
Statistical Tools

- Excel:



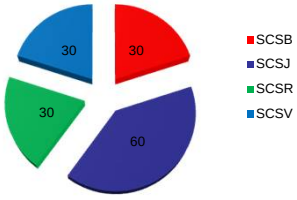
Statistical Tools

- SPSS:

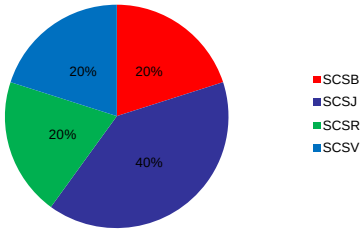


Example

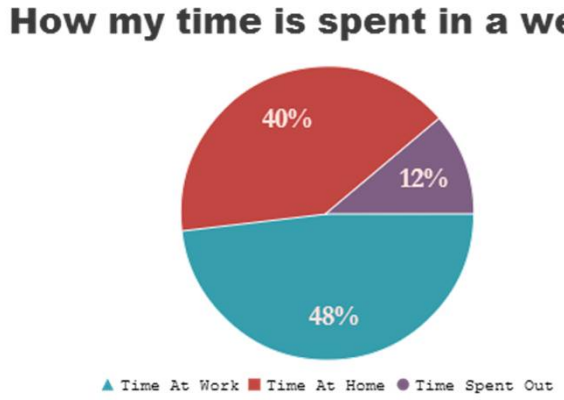
Number of Students for Year 2013/2014 Intake



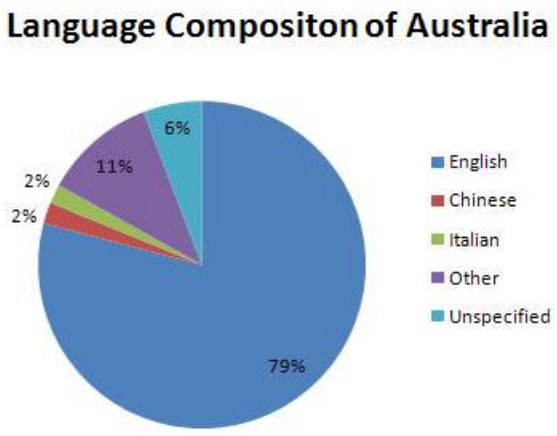
Percentage of Number of Students for Year 2013/2014 Intake



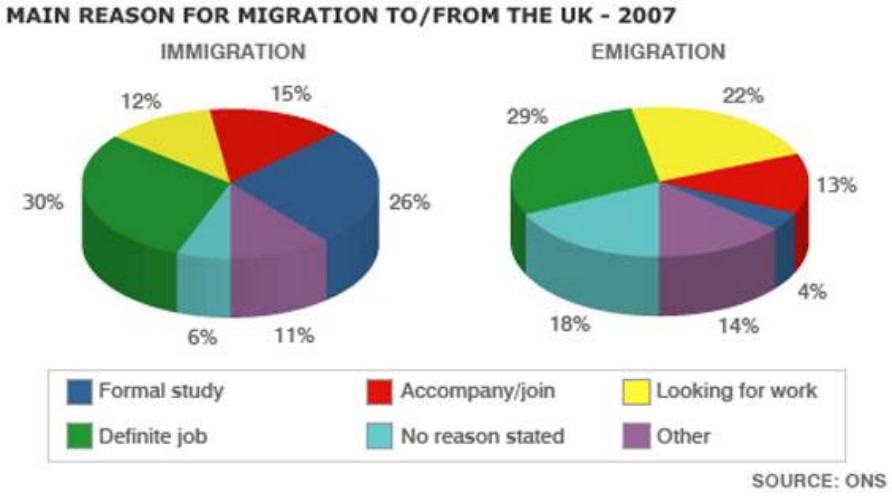
Example



Example



Example



Exercise

Travel Survey

92 people were asked how they got to work:

- 35 used a car
- 42 took public transport
- 8 rode a bicycle
- 7 walked

i) Calculate the relative frequency.

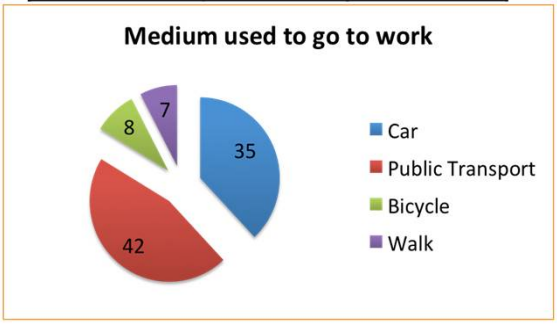
ii) Draw the appropriate chart.

Solution

i) Relative frequency:

Medium used to go to work	Frequency	Relative frequency
Car	35	0.38
Public Transport	42	0.46
Bicycle	8	0.09
Walk	7	0.07
Total	92	1.00

ii) Pie chart:



Displaying Numerical Data

- Ungrouped data comprise a listing of the observed values.
- Grouped data represent a lumping together of the observed values.
- The data can be discrete or continuous.

Example

- Ungrouped data

0	1	3	0	1	0	1	0
1	5	4	1	2	1	2	0
1	0	2	0	0	2	0	
2	1	1	.....				

Example

- Ungrouped data

2.559	2.556	2.566	2.546	2.561	2.570	2.546
2.565	2.543	2.538	2.560	2.560	2.545	2.551
2.568	2.546	2.555	2.551	2.554	2.574	2.568
2.572	2.550	2.556	2.551	2.561	2.560	2.564
2.567	2.560	2.551	2.562	2.542	2.549	2.561

Example

- Grouped data

  Frequency distribution

Data	Frequency
0	15
1	20
2	8
3	5
:	:
:	:

Example

- Grouped data

  Frequency distribution

Data	Frequency
2.531 - 2.535	6
2.536 – 2.540	8
2.541 – 2.545	12
2.546 – 2.550	13
:	:
:	:

Frequency Distributions

- A frequency distribution shows the number of observations falling into each of several ranges of values.
- Frequency distributions are portrayed as **frequency tables, histograms, or polygons**.
- Frequency distributions can show either the actual number of observations falling in each range or the percentage of observations. In the latter instance, the distribution is called a **relative frequency distribution.**

Example

- Discrete Data:

5	7	7	1
3	2	8	6
8	2	4	4
9	10	2	6
3	1	6	6
9	9	7	5
7	10	8	1
5	8		

Marks	Tally	Frequency
1	///	3
2	///	3
3	//	2
4	//	2
5	//	2
6	////	5
7	////	4
8	////	5
9	//	2
10	//	2
Total		**30**

Frequency distribution table (details):

Marks	Frequency	Relative Frequency	Cumulative Frequency (CF)	Percent (%) of CF
1	3	0.100	0.1	10.0
2	3	0.100	0.2	20.0
3	2	0.067	0.267	26.7
4	2	0.067	0.334	33.4
5	3	0.100	0.434	43.4
6	4	0.133	0.567	56.7
7	4	0.133	0.7	70.0
8	4	0.133	0.833	83.3
9	3	0.100	0.933	93.3
10	2	0.067	1.00	100.00
Total	30	1.000	100.00	

Example

Frequency distribution table for grouped data (discrete data):

Frequency Distribution Tables → Grouped Data

Arrange the following data into a frequency distribution table:

65 73 64 85 66 77 82 93 86 63 58 63 62 79 61 74

Class	Class centre *CC*	Tally	Frequency *f*
50-59	54.5	I	1
60-69	64.5	𝍸 II	7
70-79	74.5	IIII	4
80-89	84.5	III	3
90-99	94.5	I	1

Count scores

16

Double check

Add up frequency column = 16

Example

- Continuous data:

5.1 14.6 10.3 17.0
12.1 18.3 21.4 15.4
22.4 29.7 15.3 19.5
23.3 28.1 16.9 24.9

Class Interval	Frequency
5 to <10	1
10 to <15	3
15 to <20	6
20 to <25	4
25 to <30	2

Example

- Frequency table for grouped (continuous data):

HEIGHT OF PUPILS (CM)	FREQUENCY
$150 \leq x < 55$	2
$155 \leq x < 160$	6
$160 \leq x < 165$	9
$165 \leq x < 170$	5
$170 \leq x < 175$	1

Dotplots

- **When to use:** Small numerical data sets.
- **How to construct**:
  1. Draw a horizontal line and mark it with an appropriate measurement scale.
  2. Locate each value in the data set along the measurement scale, and represent it by a dot. If there are two or more observations with the same value, stack the dots vertically.
- **What to look for:**
  - ✓ A representative or typical value in the data set.
  - ✓ The extent to which the data values spread out.
  - ✓ The nature of the distribution of values along the number line.
  - ✓ The presence of unusual values in the data set.

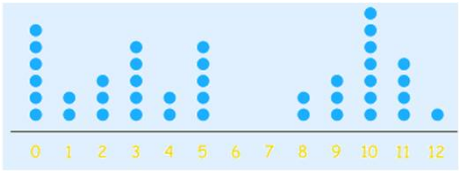
Dotplots

- A graphical data using dots.
- **Example**: Minutes To Eat Breakfast

A survey of "How long does it take you to eat breakfast?" has these results:

Minutes:	0	1	2	3	4	5	6	7	8	9	10	11	12
People:	6	2	3	5	2	5	0	0	2	3	7	4	1

Which means that 6 people take 0 minutes to eat breakfast (they probably had no breakfast!), 2 people say they only spend 1 minute having breakfast, etc. The dot plot shown below,



Histograms

- A histogram is the most commonly used graph to show frequency distributions.
- A histogram consists of a set of rectangles that represent the frequency in each categories.
- It represents graphically the frequencies of the observed values

Histograms

- What to look for:
  - Central or typical value
  - Extent of spread or variation
  - General shape
  - Location and number of peaks
  - Presents of gap and outliers

Histograms

How to construct:

① Collect data and construct a tally sheet:

- The number of cells should be between 5 and 20.
- Use 5 to 9 cells when the number of observation <100.
- Use 8 to 17 (between 100 and 500).
- Use 15 to 20 (>500).

② Determine the range:

- $R = X_h - X_l$
- Where R = range, X_h = highest number, X_l = lowest number

Histograms

③ Determine the cell interval:

- The distance between adjacent cell midpoints.
- An odd interval is recommended, so that the midpoint values will be the same number of decimal places as the data values.
- Sturgis' rule:

$$i = \frac{R}{1 + 3.322 \log n}$$

(n = number of observations)

- Trial and error, $h = R/i$
  (h = number of cells, R = range)

Histograms

④ Determine the cell midpoint:

- The lowest cell midpoint must be located to include the lowest data value in its cell.
- The simplest technique is to select the lowest data point as the midpoint value for the first cell.
- Use formula:

$$MP_l = X_l + \frac{i}{2}$$

(do not round-up/down the answer)

MP_l = midpoint for lowest cell

Histograms

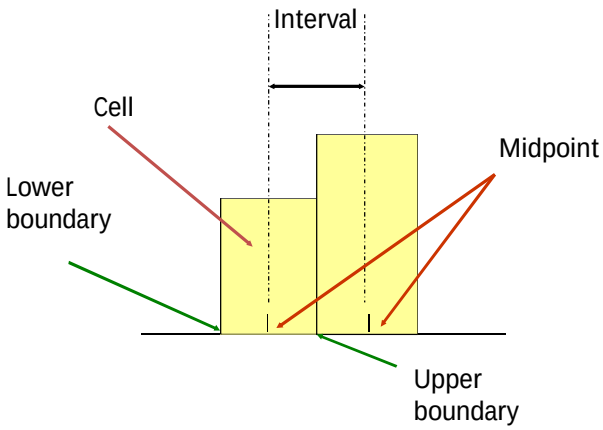
⑤ Determine the cell boundaries:

- Cell boundaries are the extreme or limit values of a cell (upper boundary and lower boundary)
- All the observations that fall between the upper and lower boundaries are classified into that particular cell.
- The boundary values are an extra decimal place or significant figure in accuracy than the observed values

⑥ Post the cell frequency.

⑦ Construct the histogram.

Histograms



Example: Data of Steel Shaft Weight (kg)

Draw the histogram for the following data.

2.559	2.556	2.566	2.546	2.561	2.570	2.546
2.565	2.543	2.538	2.560	2.560	2.545	2.551
2.568	2.546	2.555	2.551	2.554	2.574	2.568
2.572	2.550	2.556	2.551	2.561	2.560	2.564
2.567	2.560	2.551	2.562	2.542	2.549	2.561
2.556	2.550	2.561	2.558	2.556	2.559	2.557
2.532	2.575	2.551	2.550	2.559	2.565	2.552
2.560	2.534	2.547	2.569	2.559	2.549	2.544
2.550	2.552	2.536	2.570	2.564	2.553	2.558
2.538	2.564	2.552	2.543	2.562	2.571	2.553
2.539	2.569	2.552	2.536	2.537	2.532	2.552
2.575 (highest)	2.545	2.551	2.547	2.537	2.547	2.533
2.538	2.571	2.545	2.545	2.556	2.543	2.551
2.569	2.559	2.534	2.561	2.567	2.572	2.558
2.542	2.574	2.570	2.542	2.552	2.551	2.553
2.546	2.531 (lowest)	2.563	2.554	2.544		

① Calculate the frequency – used tally sheet.

Weight	Tabulation	Frequency	Weight	Tabulation	Frequency	Weight	Tabulation	Frequency
2.531	|	1	2.546	||||	4	2.561	𝍸 ||	5
2.532	||	2	2.547	|||	3	2.562	||	2
2.533	|	1	2.548		0	2.563	|	1
2.534	||	2	2.549	||	2	2.564	|||	3
2.535		0	2.550	||||	4	2.565	||	2
2.536	||	2	2.551	𝍸 |||	8	2.566	|	1
2.537	||	2	2.552	𝍸 ||	6	2.567	||	2
2.538	|||	3	2.553	|||	3	2.568	||	2
2.539	|	1	2.554	||	2	2.569	|||	3
2.540		0	2.555	|	1	2.570	|||	3
2.541		0	2.556	𝍸 ||	5	2.571	||	2
2.542	|||	3	2.557	|	1	2.572	||	2
2.543	|||	3	2.558	|||	3	2.573		0
2.544	||	2	2.559	𝍸 ||	5	2.574	||	2
2.545	||||	4	2.560	𝍸 ||	5	2.575	||	2

② Calculate the range.

$$R = X_h - X_l$$
$$= 2.575 - 2.531$$
$$= 0.044$$

③ Calculate the cell interval – use Sturgis' rule.

$$i = \frac{R}{1 + 3.322 \log n}$$
$$= \frac{0.044}{1 + 3.322 \log 110} = \frac{0.044}{1 + 3.322(2.041)} = 0.0057$$

- Trial and error
- Based on the guidelines, 0.005 will give the best presentation of the data.

$$i = 0.003; \quad h = \frac{R}{i} = \frac{0.044}{0.003} = 15$$

$$i = 0.005; \quad h = \frac{R}{i} = \frac{0.044}{0.005} = 9$$

$$i = 0.007; \quad h = \frac{R}{i} = \frac{0.044}{0.007} = 6$$

④ Calculate the cell midpoint.

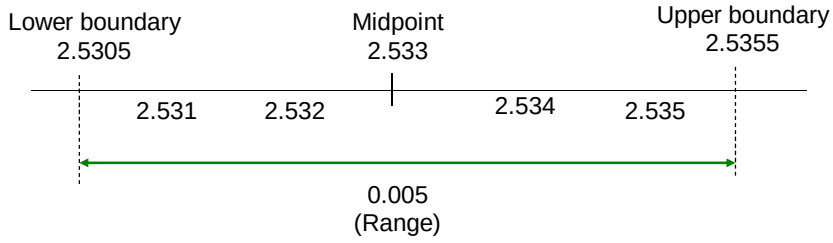
- The simplest technique is to select the lowest data point (2.531) as the midpoint value for the first cell.
- A better technique is to use the formula:

$$MP_l = X_l + \frac{i}{2} = 2.531 + \frac{0.005}{2} = 2.533$$

Cell midpoint
2.533
2.538
2.543
2.548
2.553
2.558
2.563
2.568
2.573

⑤ Determine the cell boundaries.

- The boundary values are an extra decimal place or significant figure in accuracy than the observed values.
- Example:



⑥ Post the cell frequency.

Cell Boundaries	Cell Midpoint	Frequency
2.5305 – 2.5355	2.533	6
2.5355 – 2.5405	2.538	8
2.5405 – 2.5455	2.543	12
2.5455 – 2.5505	2.548	13
2.5505 – 2.5555	2.553	20
2.5555 – 2.5605	2.558	19
2.5605 – 2.5655	2.563	13
2.5655 – 2.5705	2.568	11
2.5705 – 2.5755	2.573	8

⑦ Draw the histogram.

Frequency Polygon

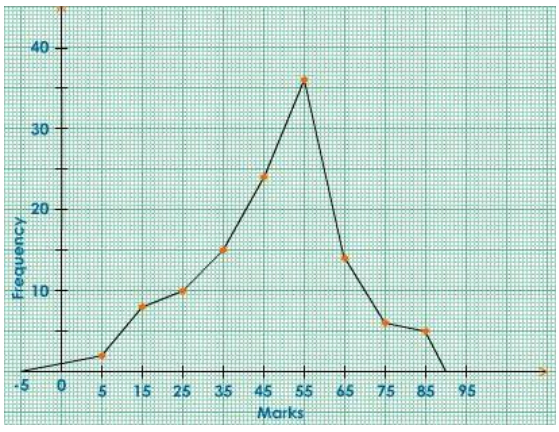
- Relative frequencies of class intervals can also be shown in a **frequency polygon**.
- In a frequency distribution, the mid-value of each class is obtained.
- The frequency is plotted against the corresponding mid-value.
- These points are joined by straight lines.
- These straight lines may be extended in both directions to meet the X - axis to form a polygon.

Example

Histogram & Polygon

Example

Marks	Frequency
0 - 10	2
10 - 20	8
20 - 30	10
30 - 40	15
40 - 50	24
50 - 60	36
60 - 70	14
70 - 80	6
80 - 90	5

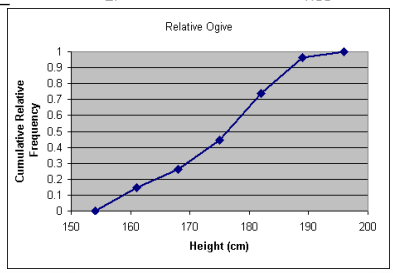


Polygon

Ogive

- A plot of the cumulative frequency against the upper class boundary with the points joined by line segments.

Class num	Class	Frequency	Relative Frequency	(Ogive) Cumulative Freq.	(Relative Ogive) Cumulative Relative Freq.
1	154 and under 161	4	0.15	4	0.15
2	161 and under 168	3	0.11	7	0.26
3	168 and under 175	5	0.19	12	0.44
4	175 and under 182	8	0.30	20	0.74
5	182 and under 189	6	0.22	26	0.96
6	189 and under 196	1	0.04	27	1.00
		27	1.00		



Stem-and-Leaf

- A stem-and-leaf display is an effective and compact way to summarize univariate numerical data.
- Each number in the data set is broken into two pieces, a stem and a leaf.
- The **stem** is the first part of the number and consists of the beginning digit(s).
- The **leaf** is the last part of the number and consists of the final digit(s).

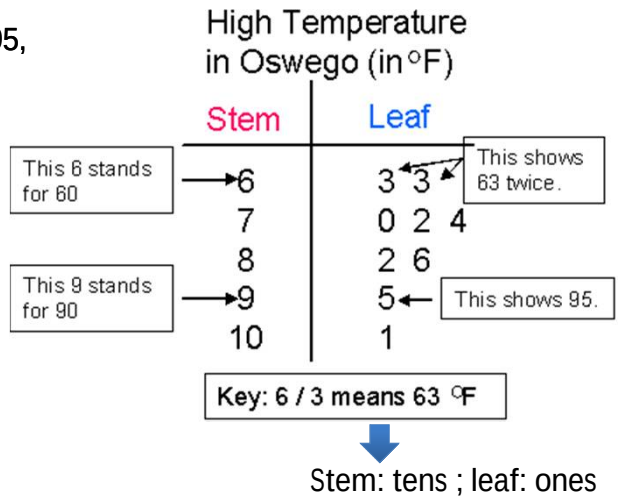
- When to use:
  - Numerical data sets with a small to moderate number of observations (does not work well for very large data sets)
- How to construct:
  - Select one or more leading digits for the stem values. The trailing digits (or sometimes just the first one of the trailing digits) become the leaves.
  - List possible stem values in a vertical column.
  - Record the leaf for every observation beside the corresponding stem value.
  - Indicate the units for stems and leaves someplace in the display.

- What to look for:

  The display conveys information about
  - a representative or typical value in the data set
  - the extent of spread about a typical value
  - the presence of any gaps in the data
  - the extent of symmetry in the distribution of values
  - the number and location of peaks

Example

Dataset:
63, 70, 74, 63, 82, 72, 95, 101, 86



Exercise

a) Make a stem-and-leaf plot for the following list of numbers:

2.1, 2.0, 3.2, 3.7, 1.4, 1.5, 2.6

b) Make a stem-and-leaf plot for the following list of numbers:

23, 19, 28, 45, 36, 47, 48, 50, 51, 13

Solution

a) **Answer**

Stem	Leaf
1	4,5
2	0,1,6
3	2,7

key: 1|4 = 1.4

b) **Answer**

Stem	Leaf
1	3,9
2	3,8
3	6
4	5,7,8
5	0,1

key: 1|3 = 13

Box Plots

- Box plots – a method that can summarize data that gives more detail than just a measure of center and spread and yet less detail than a stem-and-leaf display or histogram.
- It is compact, yet it provides information about center, spread, and symmetry or skewness of the data.

- Box plots graphically display five key statistics of a data set:
  - Minimum
  - First quartile
  - Median
  - Third quartile
  - Maximum
- Very useful in identifying the shape of a distribution and outliers in data set.

Quartiles

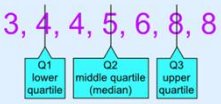
Quartiles are the values that divide a list of numbers into quarters.

- **First** put the list of numbers in order.
- **Then** cut the list into four equal parts.
- The Quartiles are at the "cuts".

Example: 5, 8, 4, 4, 6, 3, 8

Put them in order: 3, 4, 4, 5, 6, 8, 8

Cut the list into quarters:



And the result is:

- Quartile 1 (Q1) = **4**
- Quartile 2 (Q2), which is also the Median, = **5**
- Quartile 3 (Q3) = **8**

How To Calculate Quartile

- Sort the data set so measurements are in order from lowest to highest,

  Y[1], Y[2], …. , Y[N]

- Calculate,

$$i = \frac{P}{100}(N)$$

{Note: Q1: P = 25; Q2: P = 50; Q3: P = 75}

- If *i* is not an integer, round up to the next highest integer *k* and use Y[*k*] as the quartile estimate.
- If *i* is an integer, use (Y[i] +Y[*i*+1]) ÷ 2 as the quartile estimate.

Example

Given set of data:

12, 4, 6, 11, 9,15, 20, 18, 25, 30.

Calculate 1st , 2nd and 3rd quartile.

Solution (i): 1st quartile (Q1)

• Arrange in order:

4 6 9 11 12 15 18 20 25 30

• *N* = 10, *P* = 25 ; *i* = 25 x 10 ÷ 100 = 2.5, k = 3

Y[3] = 9, .: Q1 = 9

Solution (ii): 2nd quartile (Q2)

- Arrange in order:

4 6 9 11 12 15 18 20 25 30

- $N = 10$, $P = 50$; $i = 50 \times 10 \div 100 = 5$, k = 5

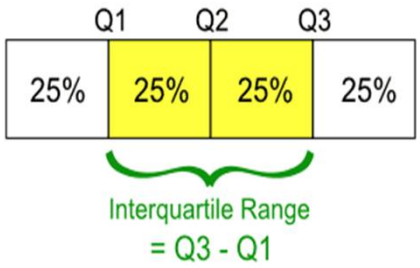
(Y[5] + Y[6]) ÷ 2 = (12 + 15) ÷ 2 = 13.5, .: Q2 = 13.5

Solution (iii): 3rd quartile (Q3)

- $N = 10$, $P = 75$; $i = 75 \times 10 \div 100 = 7.5$, k = 8

Y[8] = 20, .: Q3 = 20

The "Interquartile Range" is from Q1 to Q3:



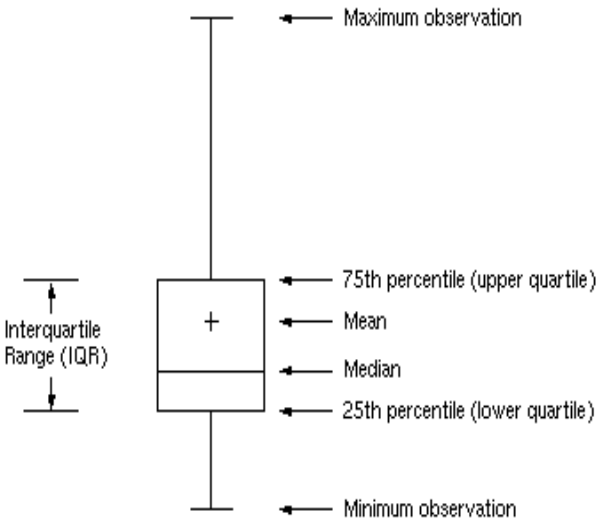
- 2 types of boxplots:
  - Skeletal boxplot
  - Modified boxplot

Skeletal Box Plots

- Construction of a Skeletal Boxplot
  - Draw a vertical (or horizontal) measurement scale
  - Construct a rectangular box with a lower (or left) edge at the lower quartile and a upper (or right) edge at the upper quartile.
  - The box width is then equal to the interquartile range (iqr)

    iqr = upper quartile – lower quartile
  - Draw a horizontal (or vertical) line segment inside the box at the location of the median.
  - Extend vertical (or horizontal) line segments, call whiskers, from each of the box to the smallest and largest observations in the data set.

Example – Skeletal Box plot



Modified Box Plots

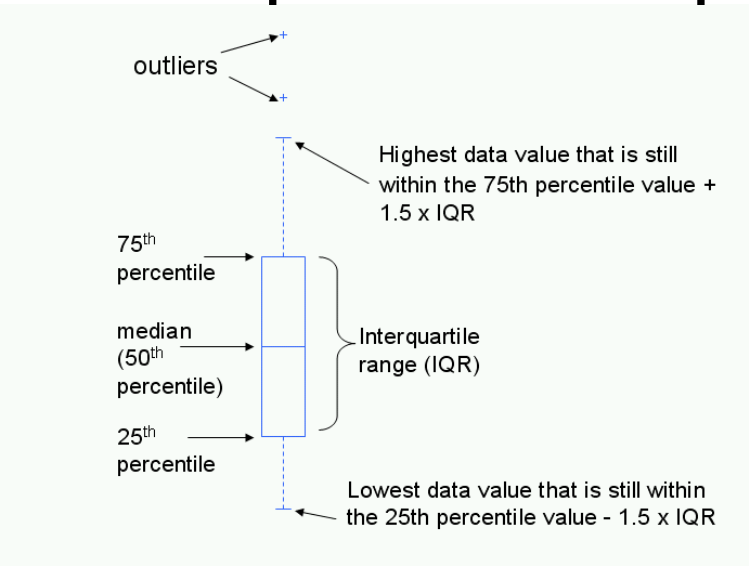
- An observation is an **outlier** if it is more than 1.5(iqr) away from the nearest quartile (the nearest end of the box)
- An outlier is **extreme** if it is more than 3(iqr) from the nearest end of the box, and it is **mild** otherwise.
- A modified boxplot represent mild and extreme outliers, and the whiskers extend on each end to the most extreme observations that are not outliers.

Outliers

- For a modified box plot, the whiskers are the lines that extend from the left and right of the box to the **adjacent values**. The adjacent values are defined as the lowest and highest observations that are still inside the region defined by the following limits:
  - Lower Limit: Q1 – 1.5 × IQR
  - Upper Limit: Q3 + 1.5 × IQR
- In general, values that fall outside of the adjacent value region are deemed outliers.

- Construction of a Modified Boxplot:
  - Draw a vertical (or horizontal) measurement scale
  - Construct a rectangular box with a lower (or left) edge at the lower quartile and a upper (or right) edge at the upper quartile.
  - The box width is then equal to iqr.
  - Draw a horizontal (or vertical) line segment inside the box at the location of the median.
  - **Determine if there are any mild or extreme outliers in the data set.**
  - Draw whiskers that extend from each end of the box to the most extreme observation that is not an outlier.
  - Draw a + to mark the location if any mild outliers or extreme outliers in the data set.

Example – Modified Box plot



Exercise

a) Draw a box plot for the following set of data.
Remember to order the data first, if necessary.

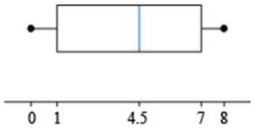
1, 0, 3, 2, 1, 1, 7, 8, 6, 6, 7, 7

b) Draw a box plot for the following set of data.
Remember to order the data first, if necessary.

4.7, 3.8, 3.9, 3.9, 4.6, 4.5, 5

Solution

a) **Answer**



b) **Answer**