# Revision

- 2 main branches of statistics?
- Descriptive statistics?
- Inferential statistics?
- Population vs Sample?
- Data analysis process?
- Levels of measurement?

# PAST YEAR

Determine which of the four levels of measurement (nominal, ordinal, interval and ratio) is most appropriate. [5 marks]

a. Heights of women basketball players in a tournament.

b. Ratings of superior, above average, average, below average or poor for blind dates.

c. Noon temperatures (in degrees Celsius) in a Johor Bahru this week.

d. A movie critic's classification of "drama, comedy, adventure".

e. Consumer Reports magazine ratings of "best buy, recommended, not recommended".

f. Distances travelled by students who commute to faculty.

g. The actual contents (in ml) of cola in Pepsi cans.

h. IQ scores, where the score is considered to be a measure of intelligence.

i. IQ scores, where the score is considered to be the number of points scored on the IQ test.

j. The number of bugs made when a programmer develop a coding for a project.

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

INNOVATIVE ● ENTREPRENEURIAL ● GLOBAL

www.utm.my

# Presenting Data

Prepared by Razana Alwee / Updated by Suhaila Mohamad Yusuf

# Types of Data Set

- Univariate
- Multivariate
  - 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 consists only two variables or two different characteristics of observations.
  - 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.
  - Frequency distribution
  - Bar chart
  - Pie chart

# Frequency Distribution

- When data is collected from a survey or designed experiment, they must be organized into a manageable form.
- Data that is not organized is referred to as **raw data.**
- We can construct a frequency table to organize data.

# 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.

$$\text{relative frequency} = \frac{\text{frequency}}{\text{number of observations in the data set}}$$

# Example

## Data

| 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

**Staff distribution in Faculty of Computing**

## Frequency Table

| **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)

# Bar Charts

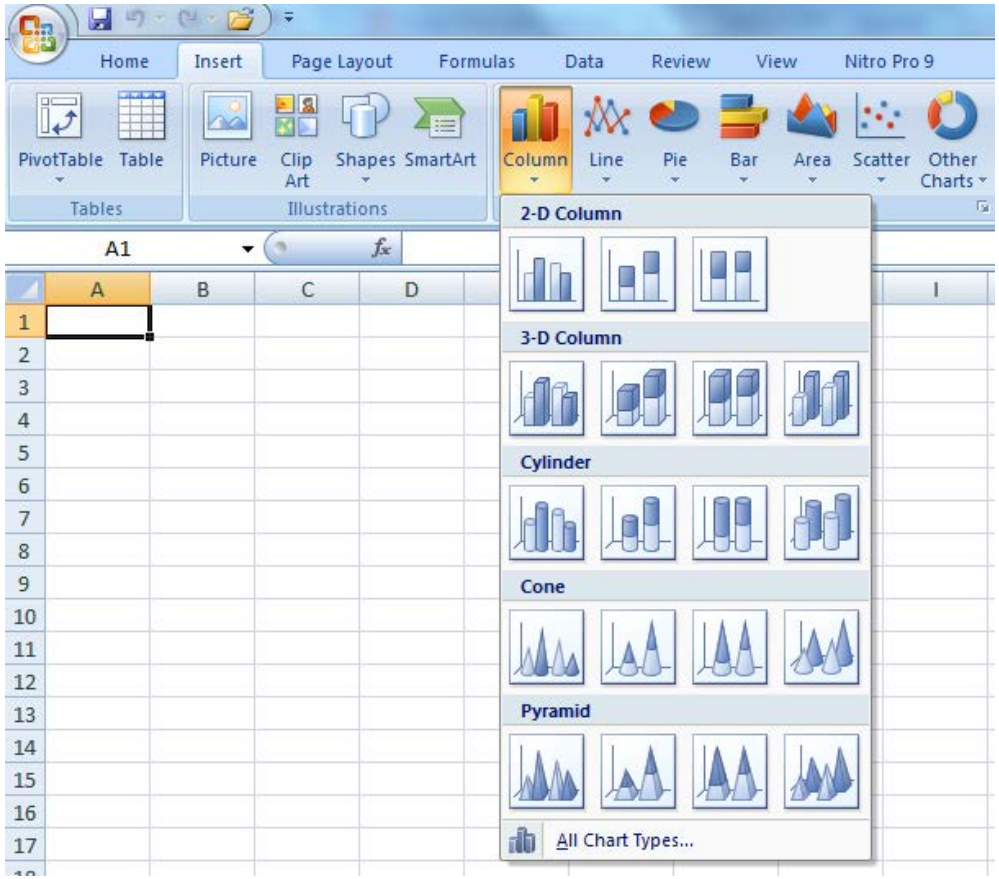
- A bar chart is a graph of the frequency distribution of categorical data.
- Each category in the frequency distribution is represented by a bar or rectangle.

# Bar Charts

- 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.

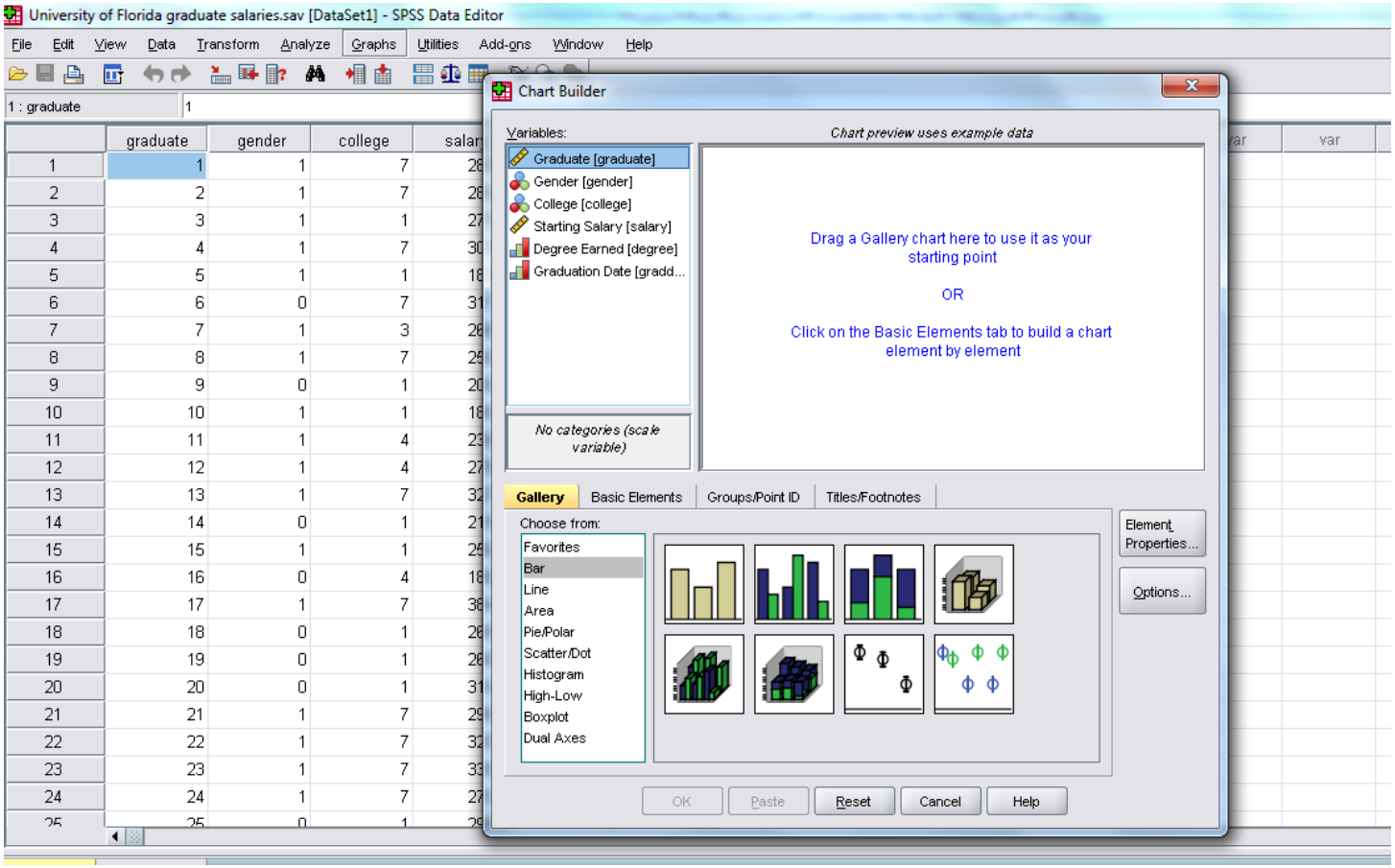
# Statistical Tools

- Excel

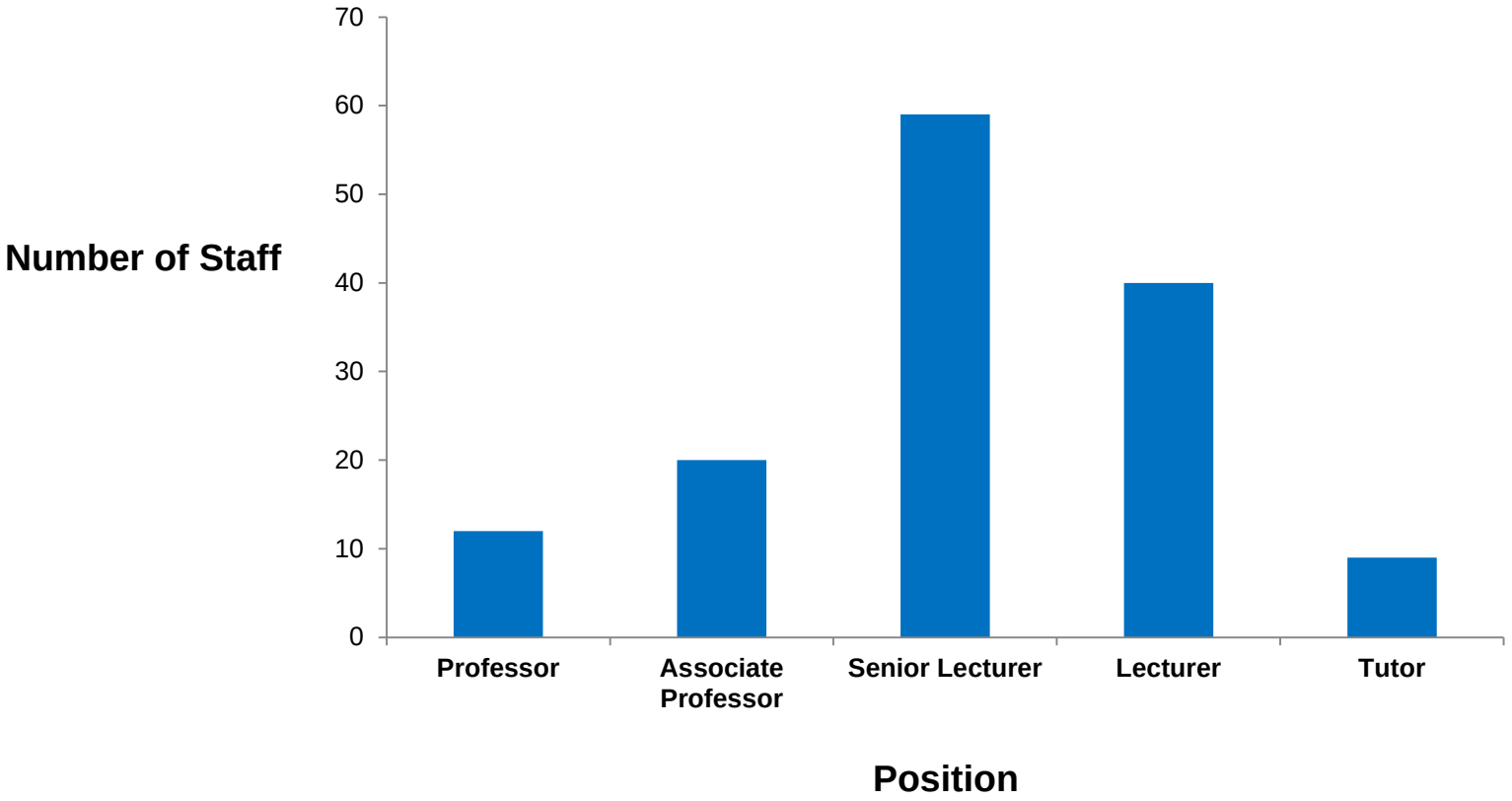


# Statistical Tools

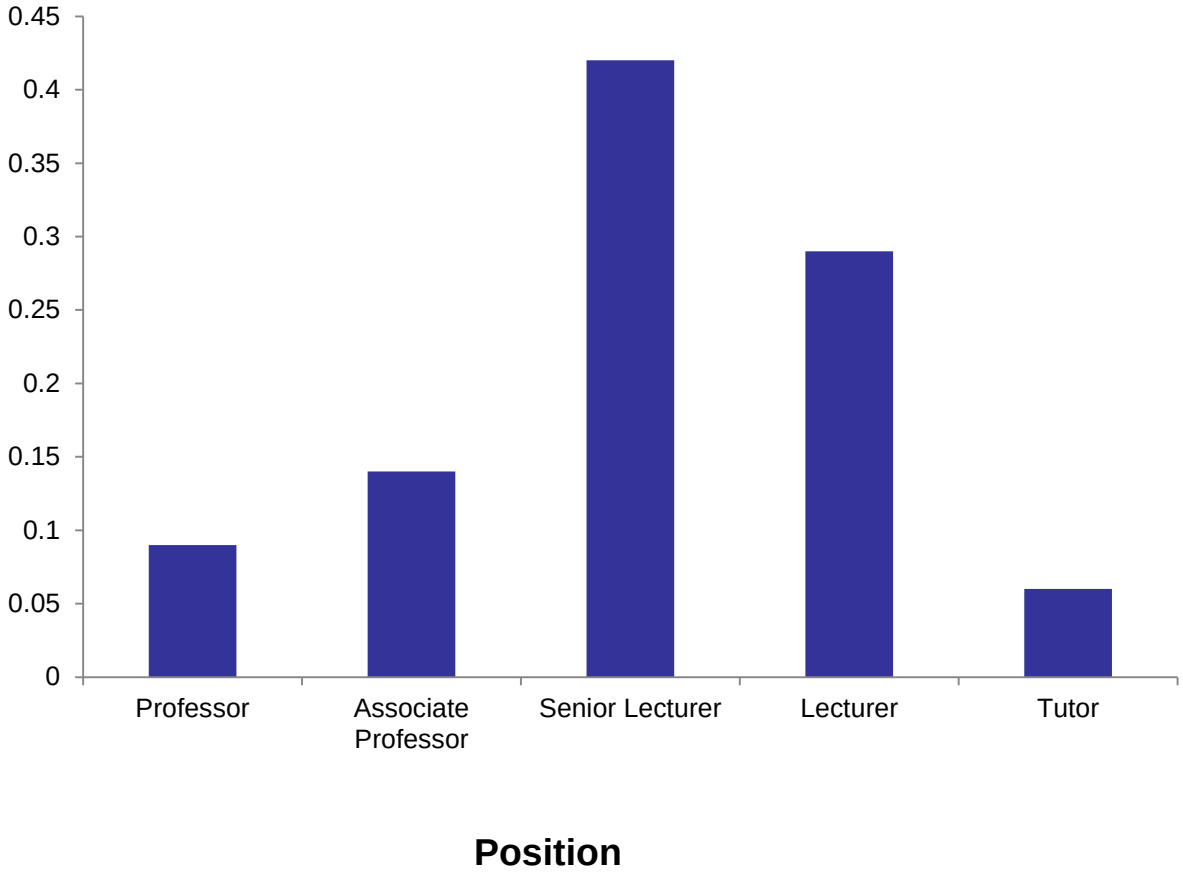
- SPSS



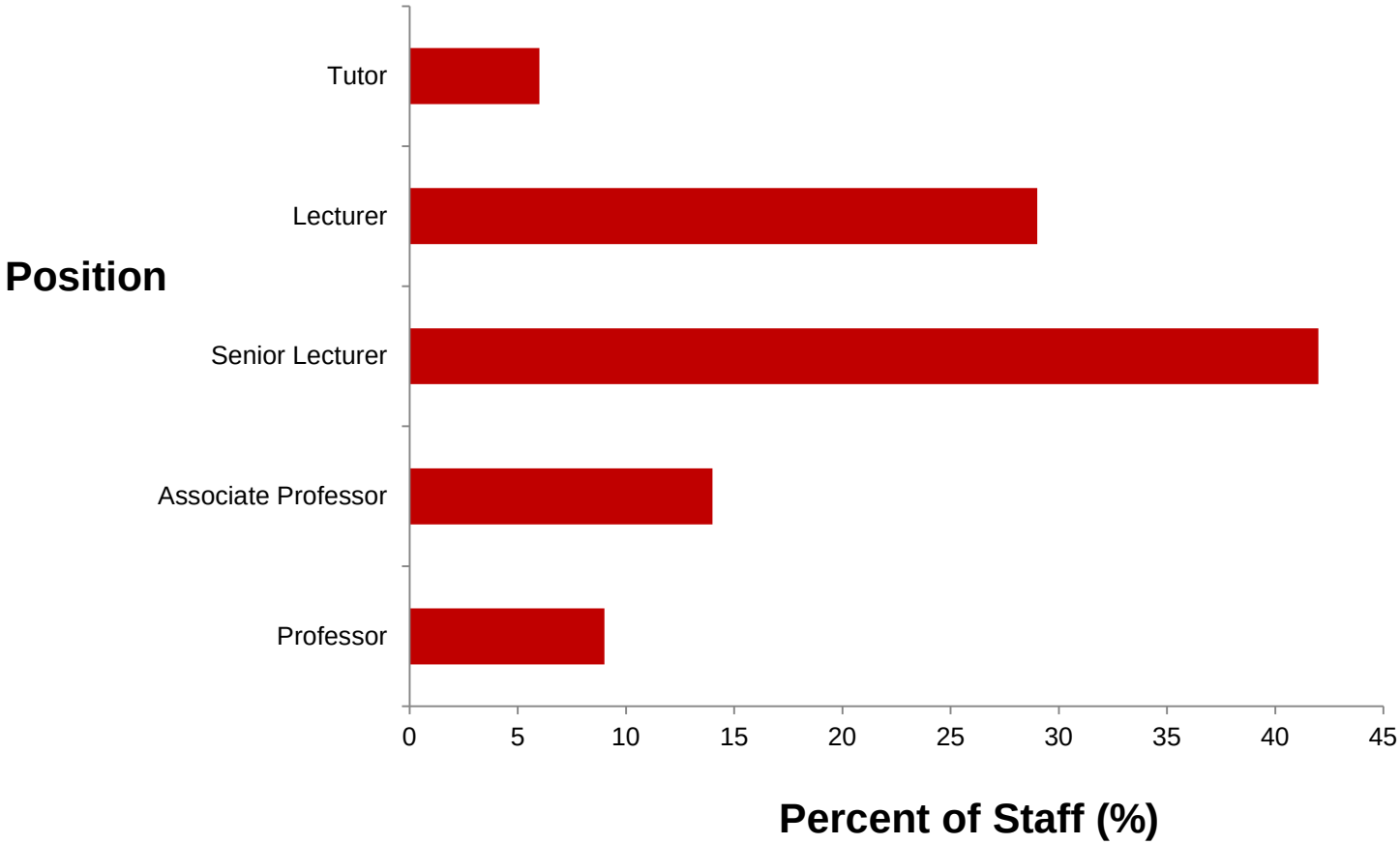
# Example


Number of Staff
70
60
50
40
30
20
10
0
Professor
Associate Professor
Senior Lecturer
Lecturer
Tutor
Position

# Example


Relative frequency
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Professor
Associate Professor
Senior Lecturer
Lecturer
Tutor
Position

# Example


Position
Tutor
Lecturer
Senior Lecturer
Associate Professor
Professor
0
5
10
15
20
25
30
35
40
45
Percent of Staff (%)

# Comparative Bar Charts

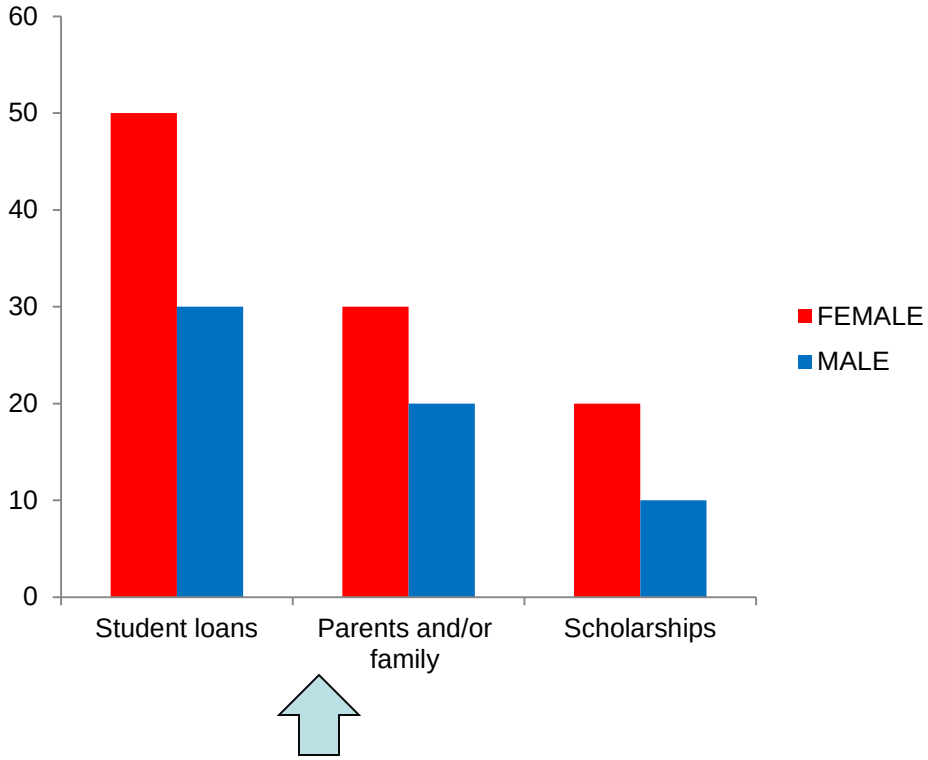
- Bar chart can also be used to give a visual comparison of two or more groups.
- When constructing a comparative bar graph we use the **relative frequency rather than the frequency** to construct the scale on the vertical axis so that we can make **meaningful comparisons** even if the sample sizes are 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** |

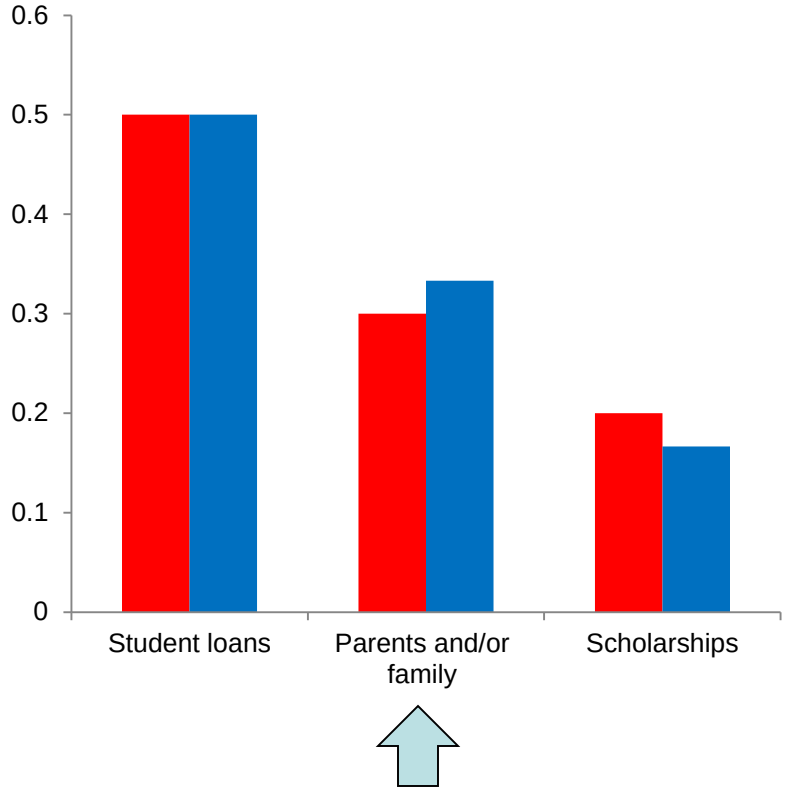
# Example

## Frequency



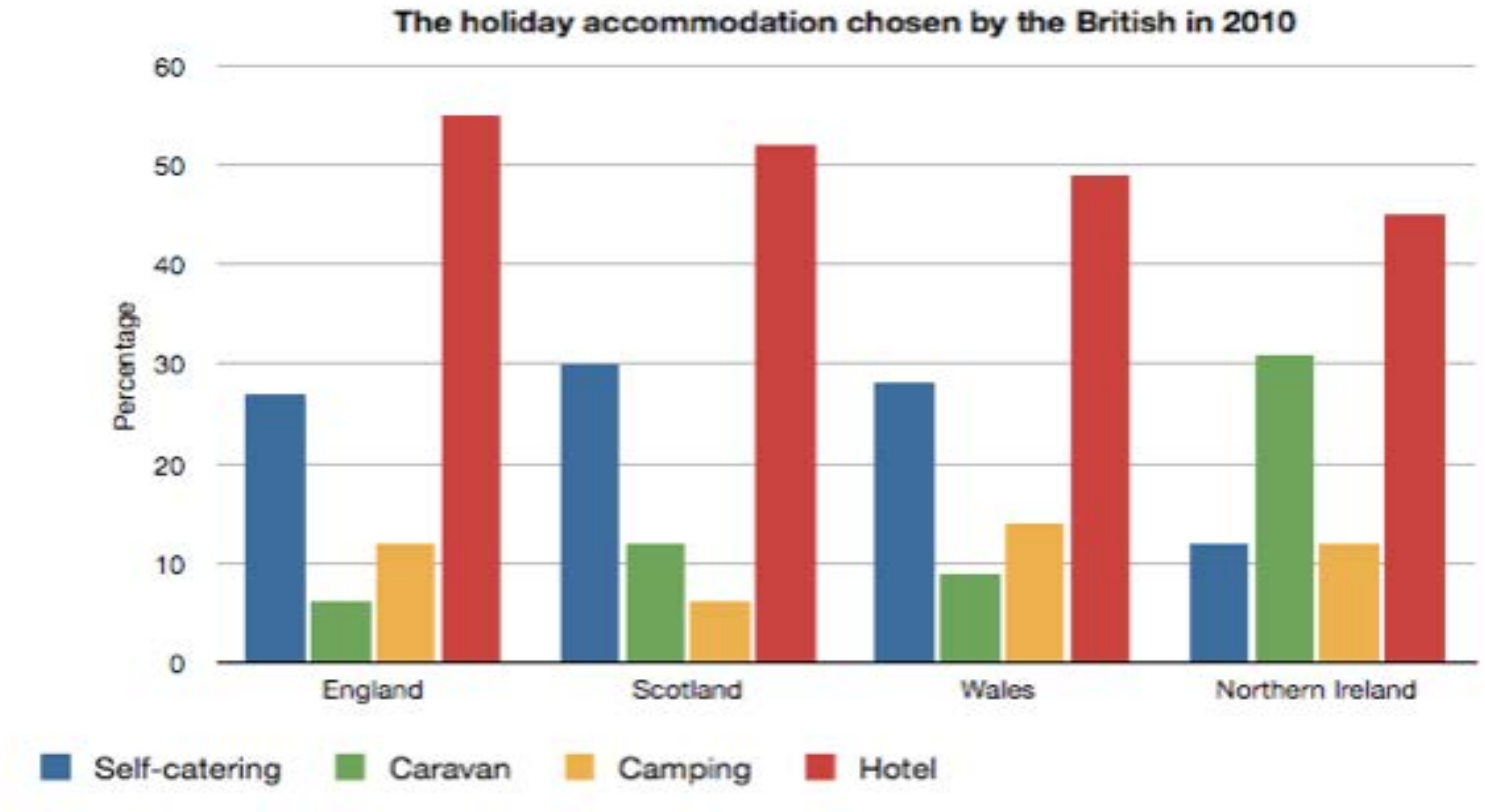

**Inaccurate comparative bar chart**

## Relative frequency




**Accurate comparative bar chart**

# Example


The holiday accommodation chosen by the British in 2010
Percentage
60
50
40
30
20
10
0
England
Scotland
Wales
Northern Ireland
Self-catering
Caravan
Camping
Hotel

# 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.

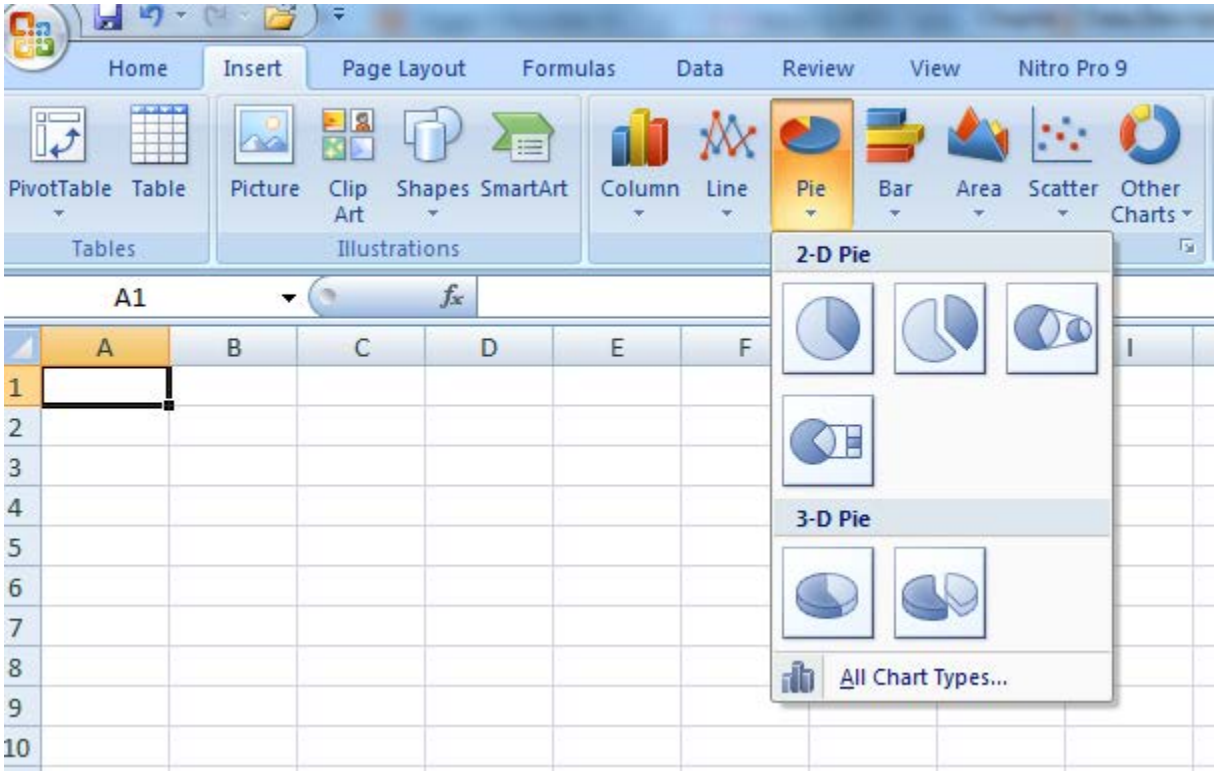
# Pie Charts

- 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.

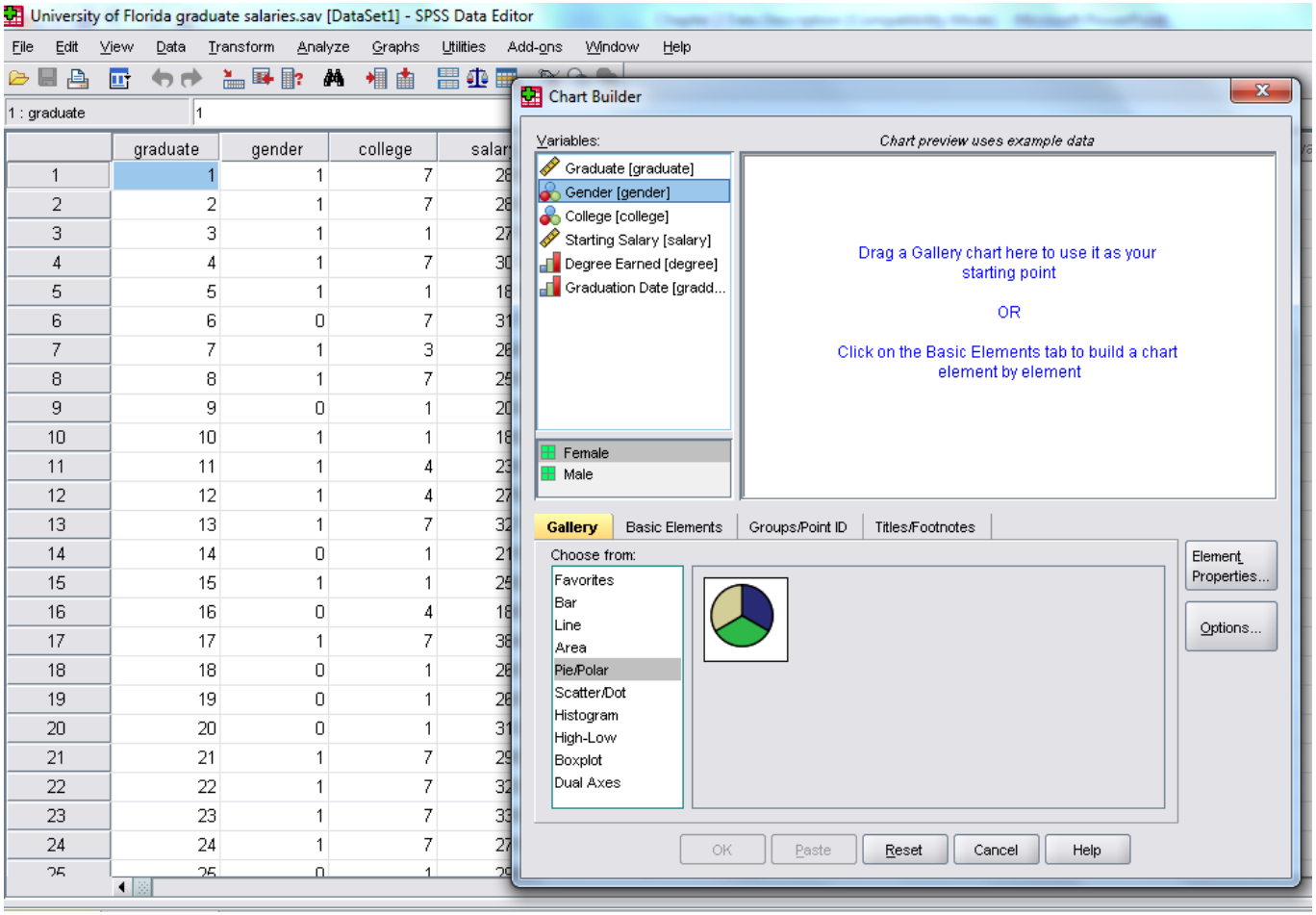
# Statistical Tools

- Excell



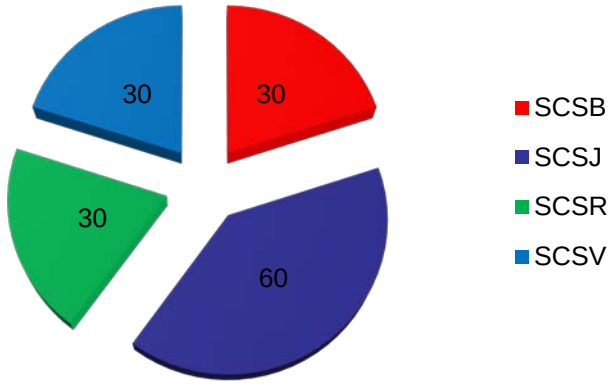
# Statistical Tools

- SPSS

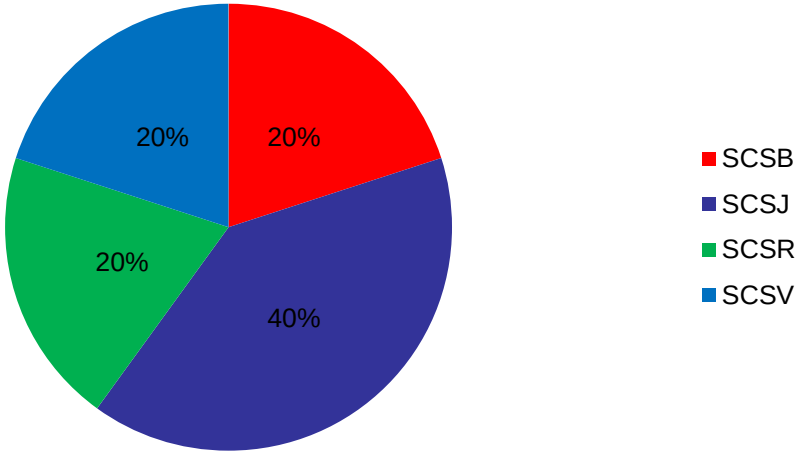


# Example

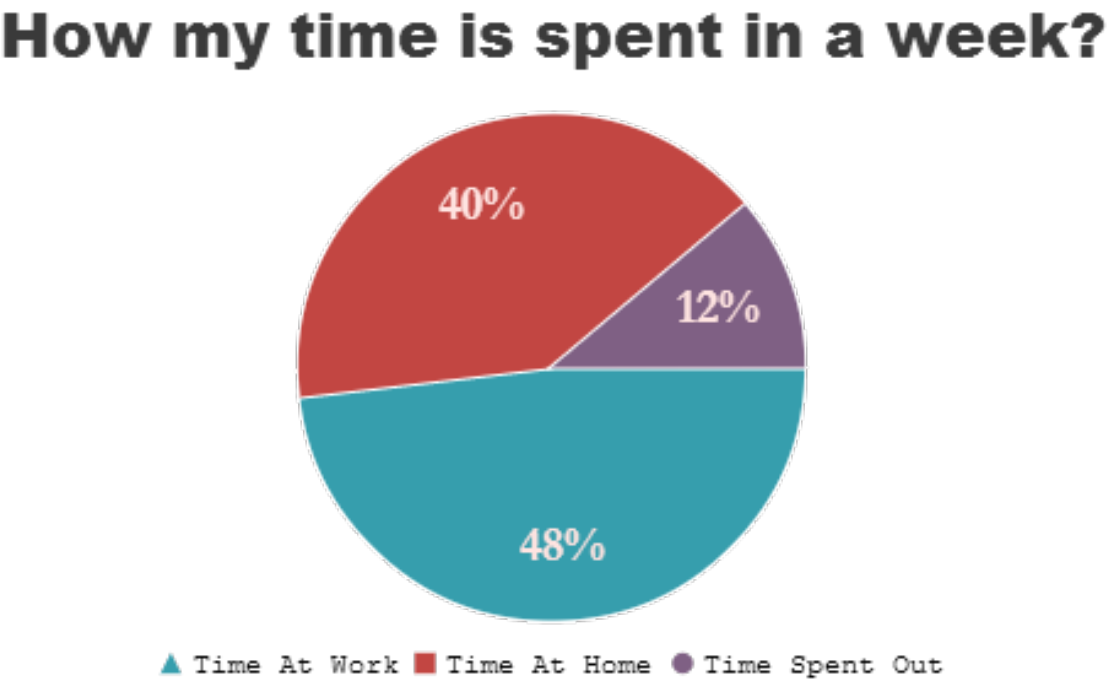
**Number of Students for Year 2013/2014 Intake**



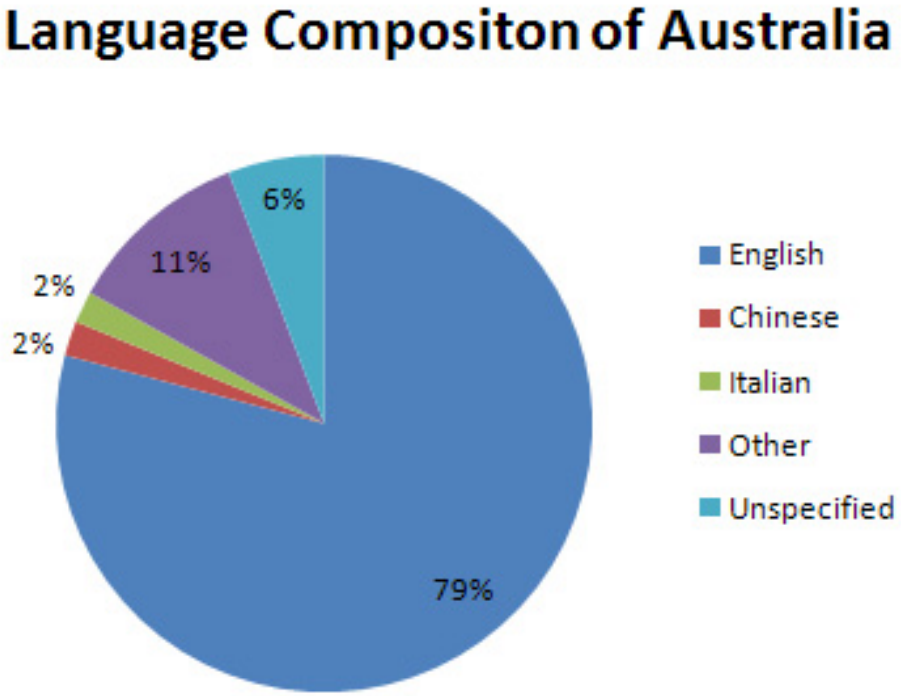

**Percentage of Number of Students for Year 2013/2014 Intake**



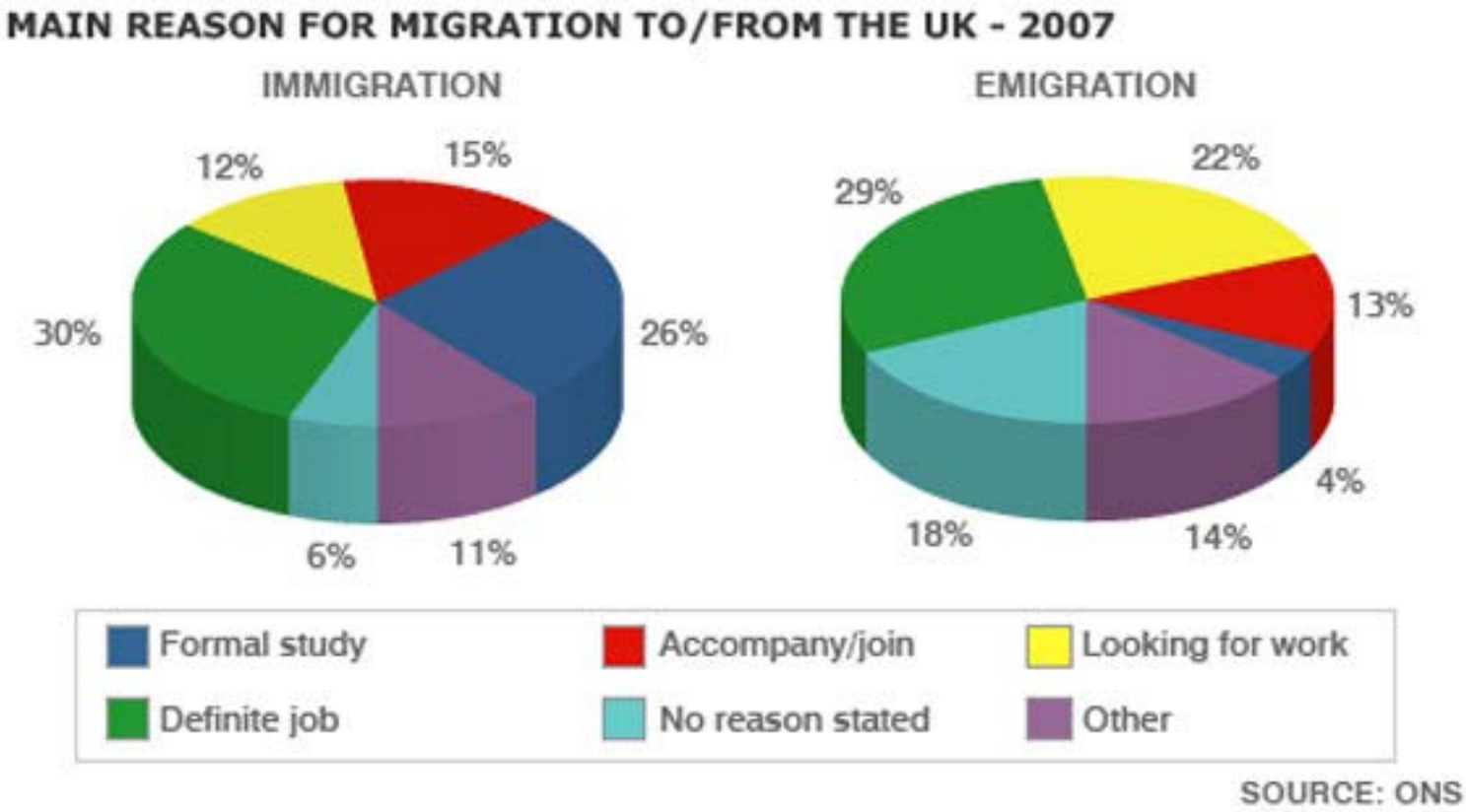
# Example


How my time is spent in a week?
40%
12%
48%
Time At Work
Time At Home
Time Spent Out

# Example


Language Compositon of Australia
6%
11%
2%
2%
79%
English
Chinese
Italian
Other
Unspecified

# Example


MAIN REASON FOR MIGRATION TO/FROM THE UK - 2007
IMMIGRATION
12%
15%
30%
26%
6%
11%
EMIGRATION
22%
29%
13%
4%
18%
14%
Formal study
Accompany/join
Looking for work
Definite job
No reason stated
Other
SOURCE: ONS

# Displaying Numerical Data

## 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** |

# Example

| Marks | Frequency | Relative Frequency | Percent | Cumulative percent |
|---|---|---|---|---|
| 1 | 3 | 0.100 | 10.0 | 10.0 |
| 2 | 3 | 0.100 | 10.0 | 20.0 |
| 3 | 2 | 0.067 | 6.7 | 26.7 |
| 4 | 2 | 0.067 | 6.7 | 33.4 |
| 5 | 3 | 0.100 | 10.0 | 43.4 |
| 6 | 4 | 0.133 | 13.3 | 56.7 |
| 7 | 4 | 0.133 | 13.3 | 70.0 |
| 8 | 4 | 0.133 | 13.3 | 83.3 |
| 9 | 3 | 0.100 | 10.0 | 93.3 |
| 10 | 2 | 0.067 | 6.7 | 100.00 |
| Total | 30 | 1.000 | 100.00 | |

# Example

- Continuous data

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

# 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.

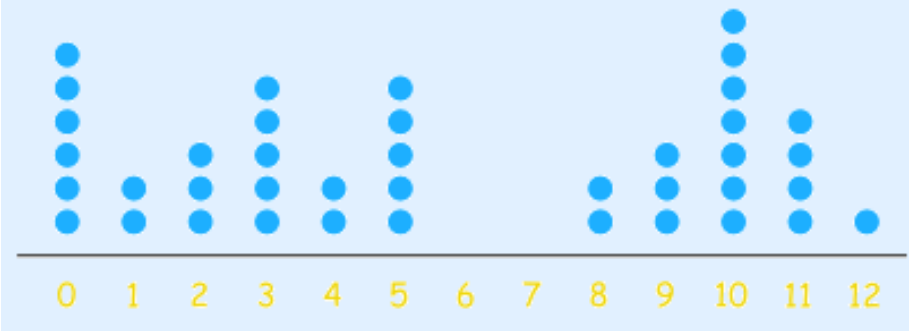
# Example

- 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,



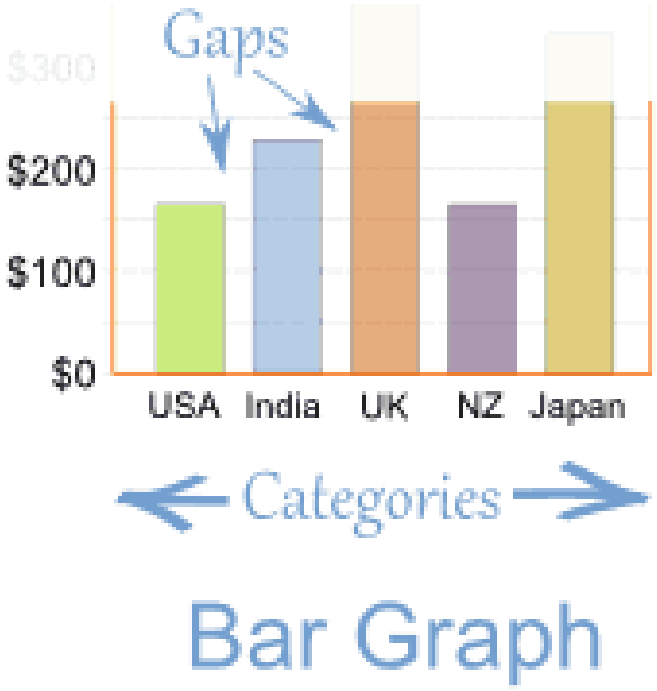
# Histograms

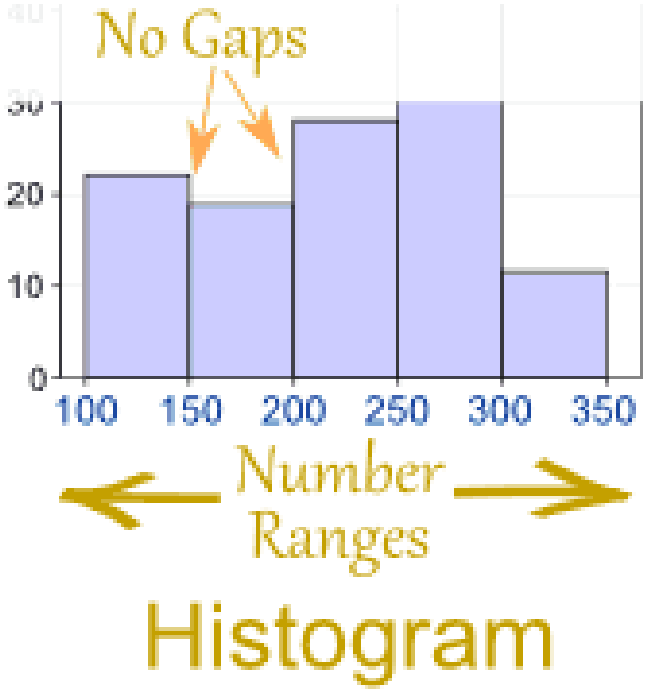
- A histogram is the most commonly used graph to show frequency distributions.
- A histogram consists of a set of rectangles that represent the frequency in each categories.
- It represents graphically the frequencies of the observed values

# 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

# Difference between Histograms and Bar Chart



# Difference between Histograms and Bar Chart

## Histogram or Bar Graph

| Situation | Bar Graph or Histogram? |
|---|---|
| We want to compare total income of five different people. | **Bar graph**. Key question: What is the revenue for each person? |
| We have measured revenues of several people. We want to compare numbers of people that make from 0 to 10,000; from 10,000 to 20,000; from 20,000 to 30,000 and so on. | **Histogram**. Key question: How many people are in each class of revenues? |
| We want to compare heights of ten basketball players on a team. | **Bar graph**. Key question: What is the height of each player? |
| We have measured several players. We want to compare numbers of players that are from 5-5.5 feet high; from 5.5-6; from 6-6.5 and so on. | **Histogram.** Key question: How many players are there in each class of heights? |

# The Construction of a Histogram

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

# The Construction of a Histogram

- 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)

# The Construction of a Histogram

- 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

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

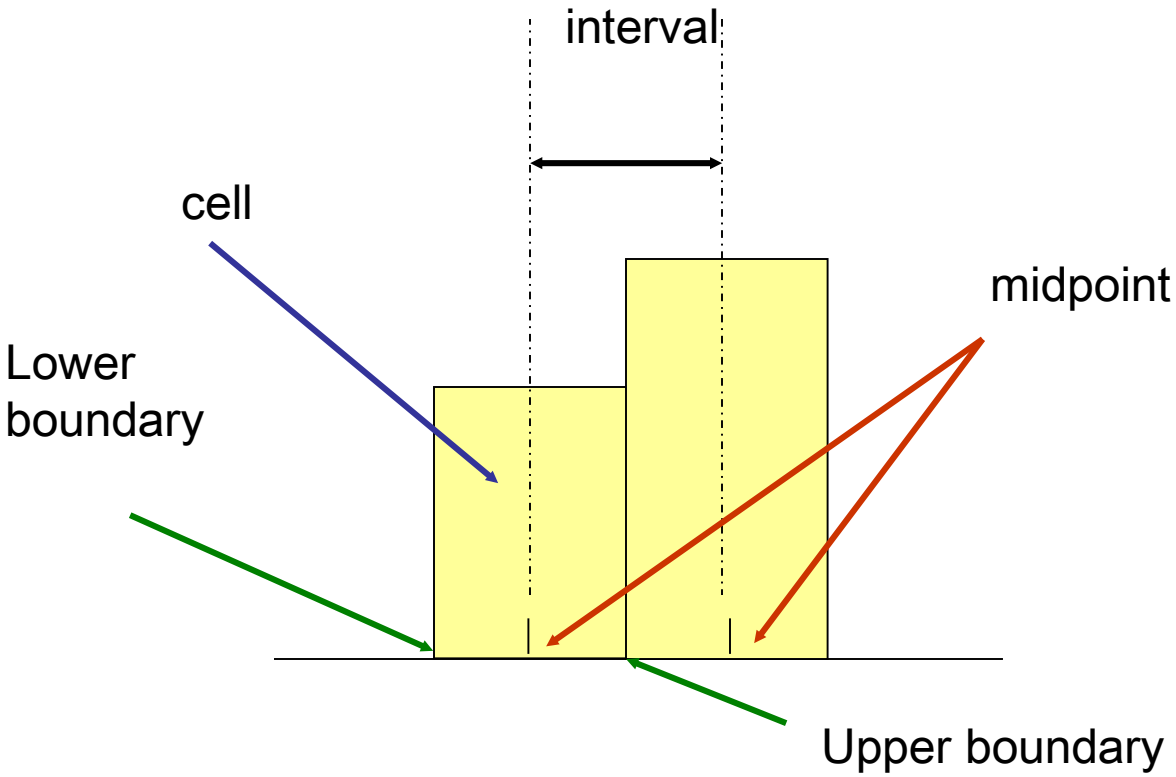
(do not round answer)

$MP_l$ = midpoint for lowest cell

# The Construction of a Histogram

- 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
- Post the cell frequency
- Construct the histogram

# The Construction of a Histogram



# Example – Steel Shaft Weight (kg)

| | | | | | | |
|---|---|---|---|---|---|---|
| 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 (h) | 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 (l) | 2.563 | 2.554 | 2.544 | | |

# Example – 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 |

# Example – range

- Determine the range

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

# Example - cell interval

- Determine the cell interval
  - Sturgis' rule

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

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

  - The closest odd interval for the data is 0.005

# Example - cell interval

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

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

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

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

# Example - cell midpoints

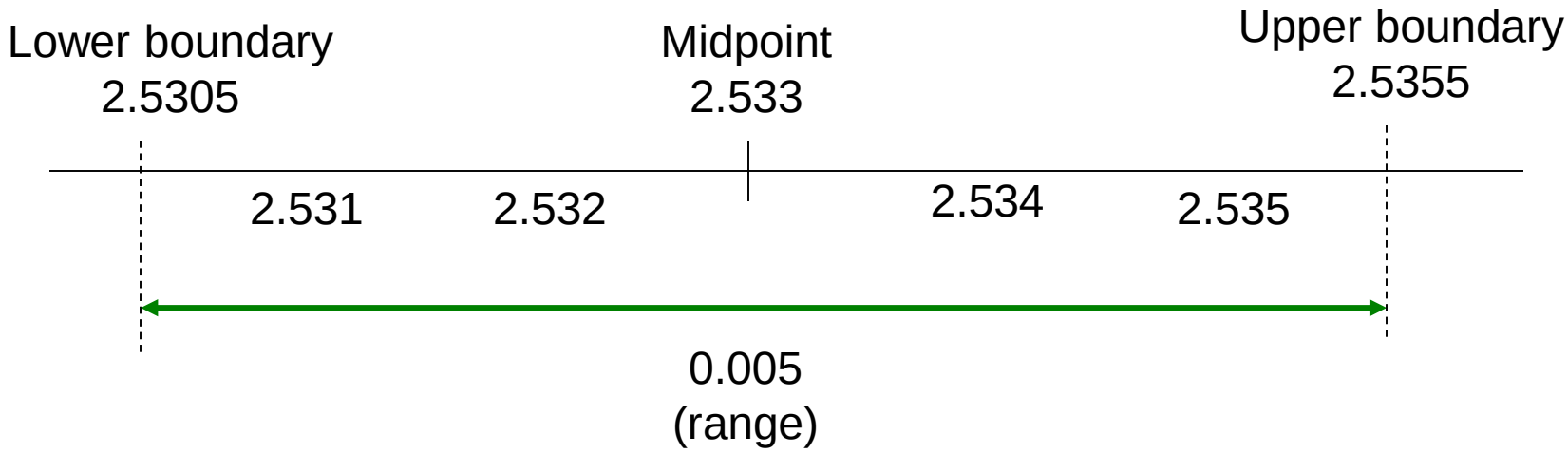
- Determine the cell midpoints
- The simplest technique is to select the lowest data point (2.531) as the midpoint value for the first cell.
- A better technique is to use the formula

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

| Cell midpoint |
|---|
| 2.533 |
| 2.538 |
| 2.543 |
| 2.548 |
| 2.553 |
| 2.558 |
| 2.563 |
| 2.568 |
| 2.573 |

# Example – cell boundaries

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



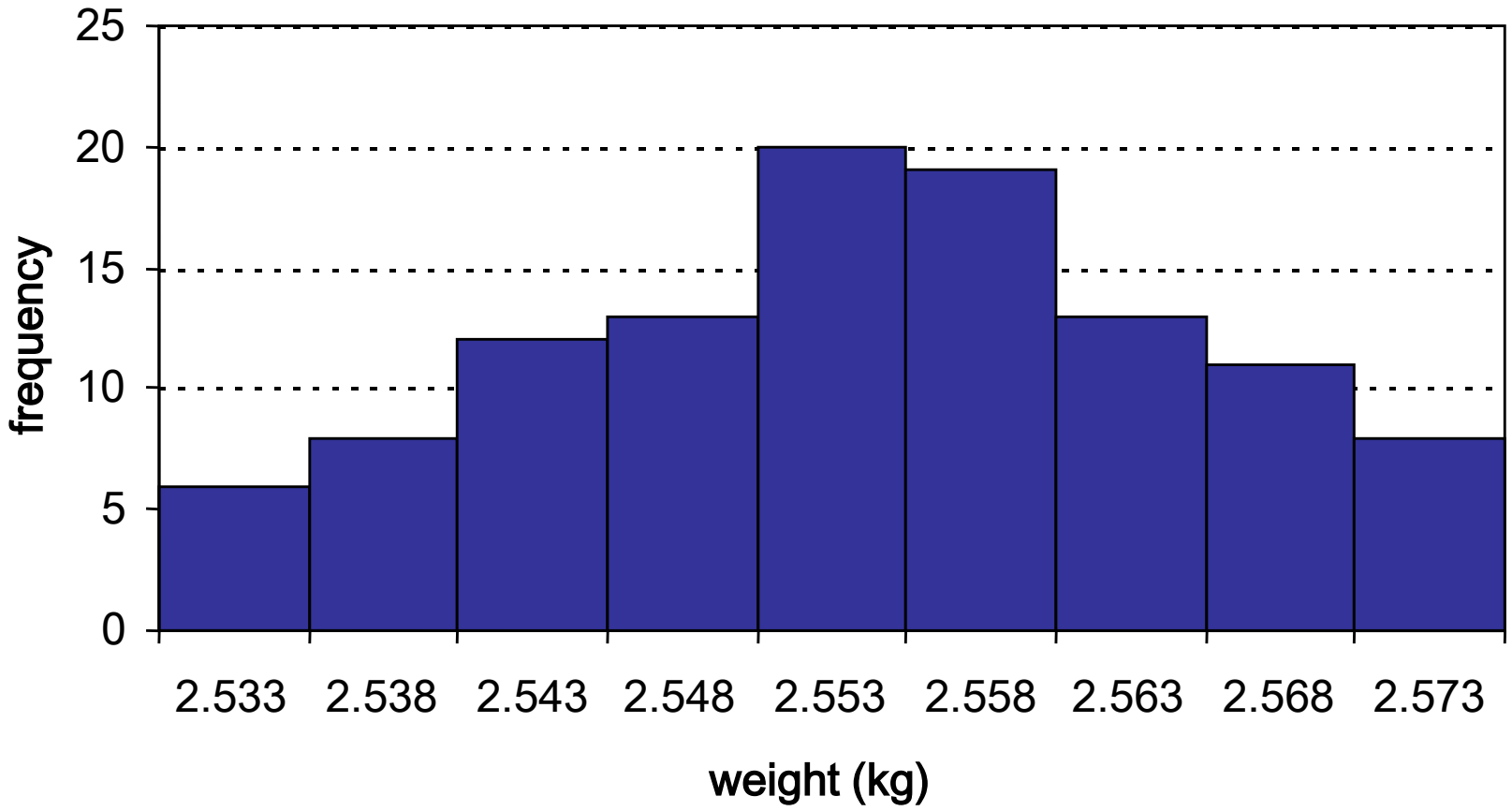
# Example – cell boundaries

| Cell boundaries | Cell midpoint |
|---|---|
| 2.5305 – 2.5355 | 2.533 |
| 2.5355 – 2.5405 | 2.538 |
| 2.5405 – 2.5455 | 2.543 |
| 2.5455 – 2.5505 | 2.548 |
| 2.5505 – 2.5555 | 2.553 |
| 2.5555 – 2.5605 | 2.558 |
| 2.5605 – 2.5655 | 2.563 |
| 2.5655 – 2.5705 | 2.568 |
| 2.5705 – 2.5755 | 2.573 |

# Example – 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 |

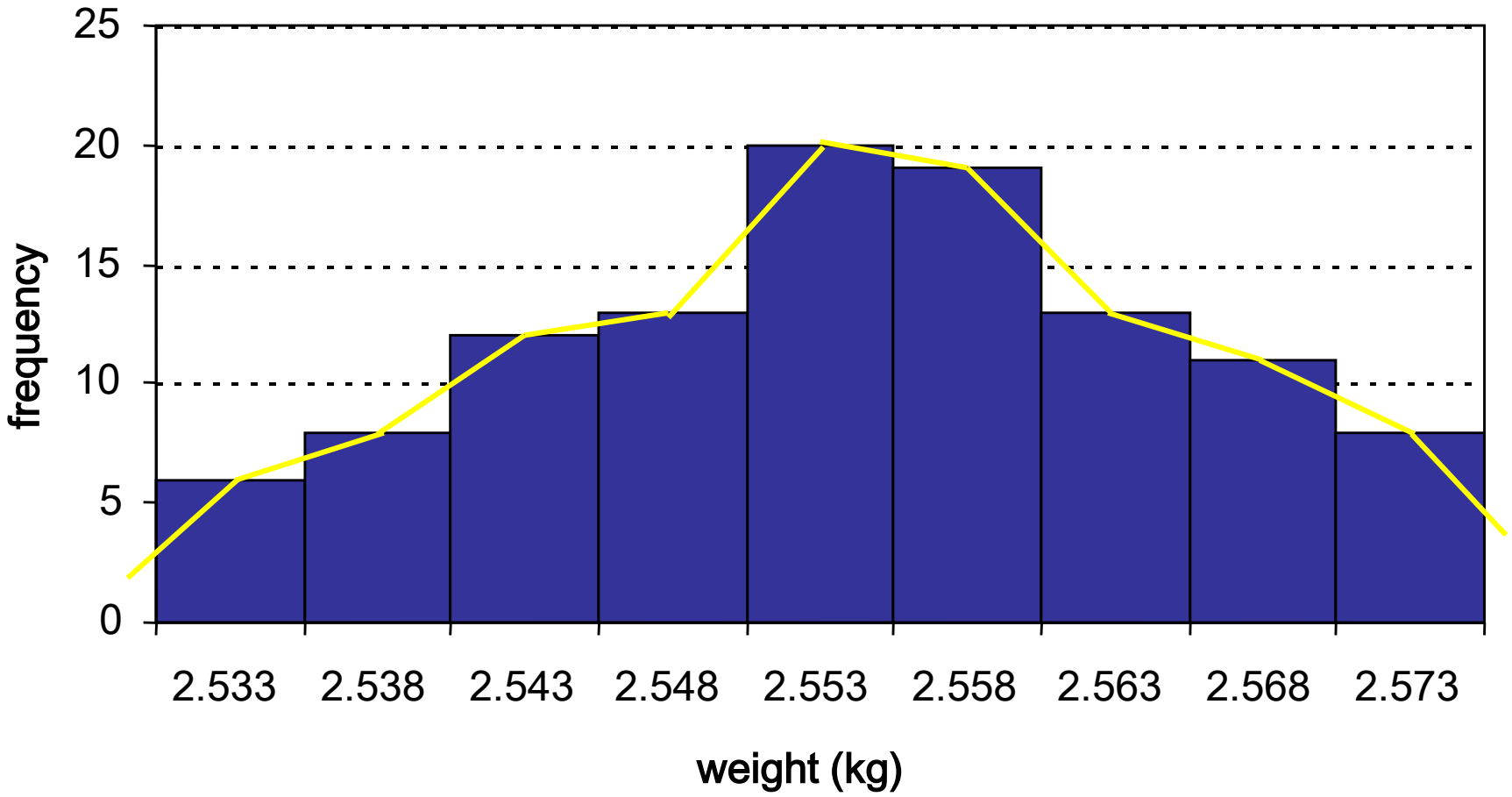
# Example - Histogram



# Frequency Polygon

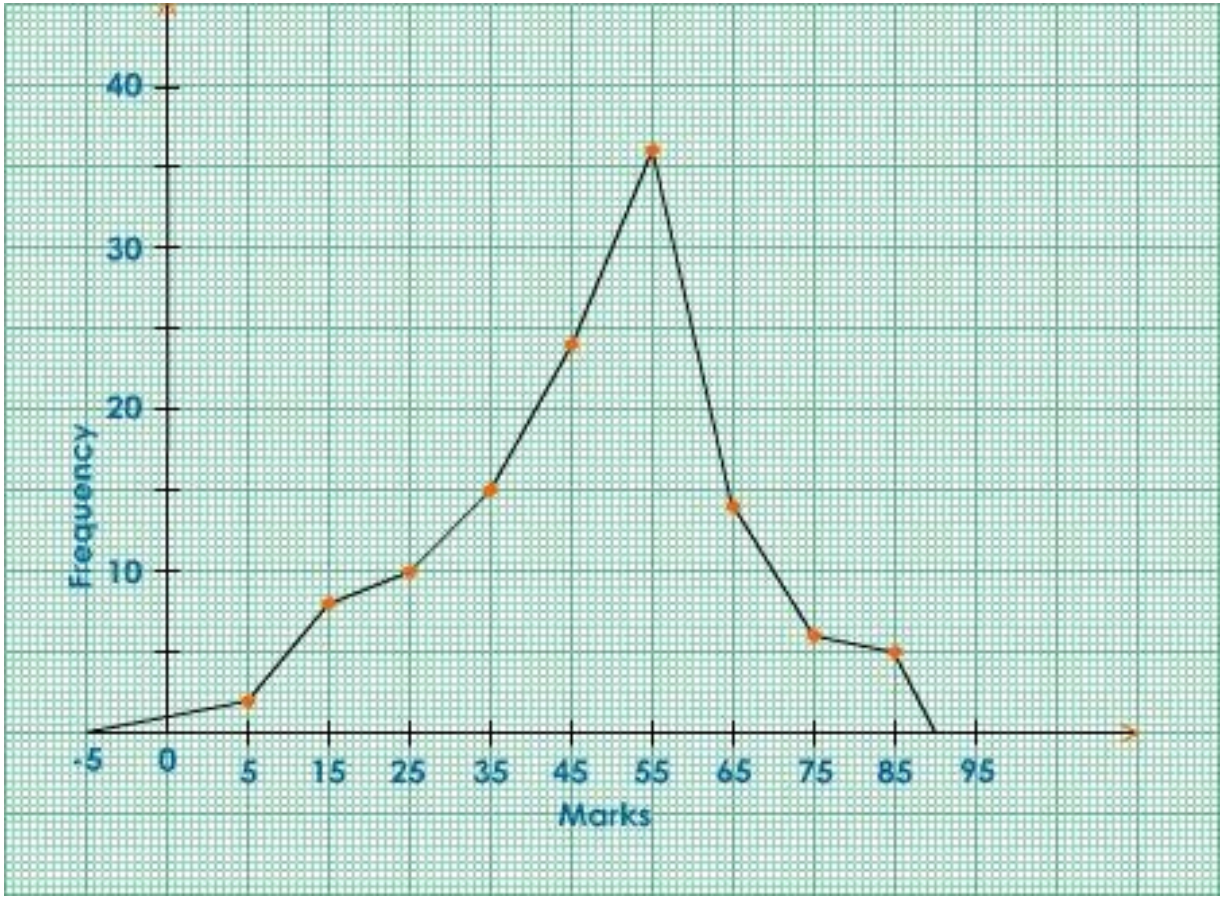
- Relative frequencies of class intervals can also be shown in a **frequency polygon**.
- In a frequency distribution, the mid-value of each class is obtained.
- The frequency is plotted against the corresponding mid-value.
- These points are joined by straight lines.
- These straight lines may be extended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

# Example


25
20
15
10
5
0
frequency
2.533
2.538
2.543
2.548
2.553
2.558
2.563
2.568
2.573
weight (kg)

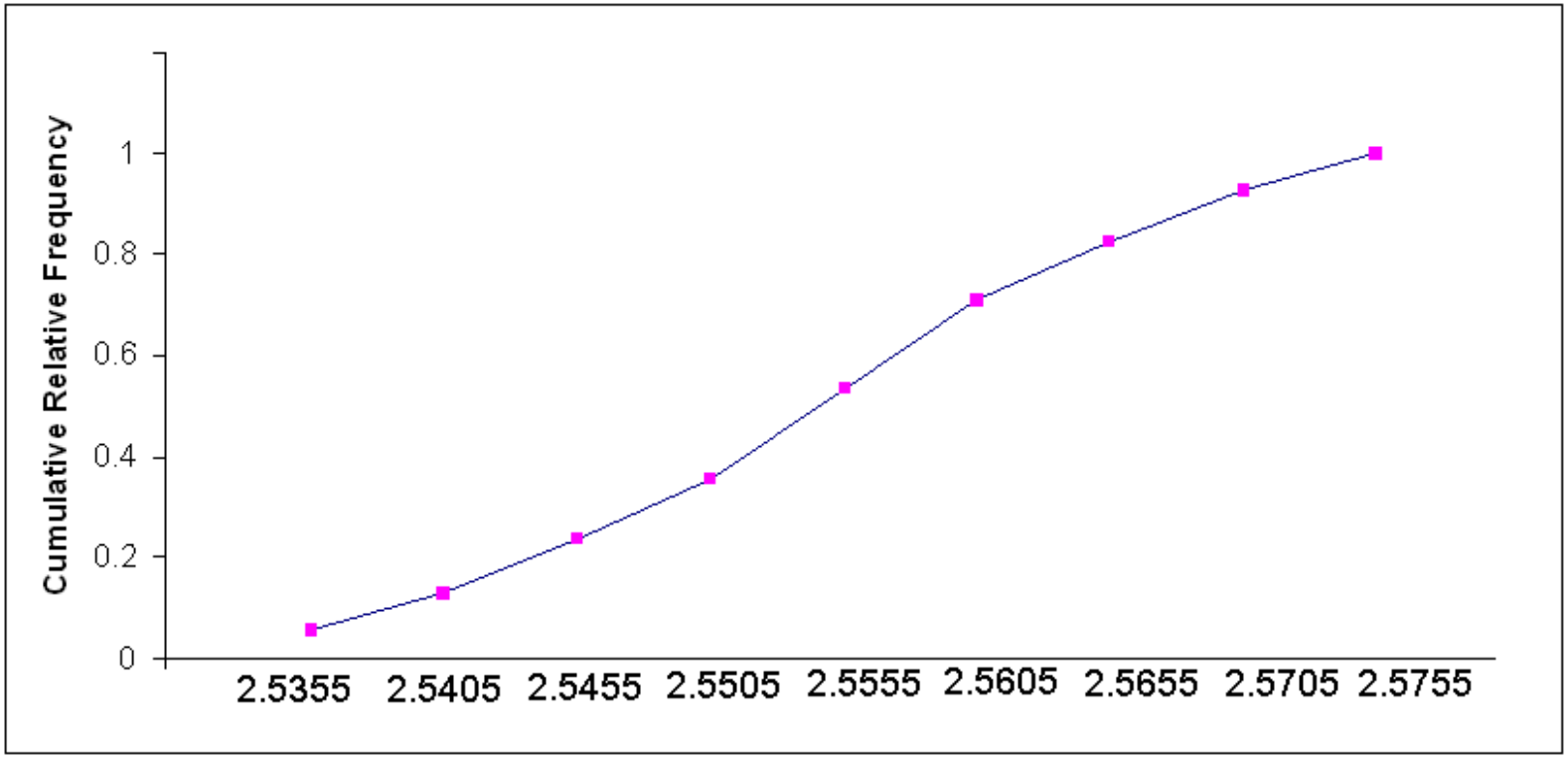
# Example

| 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 |



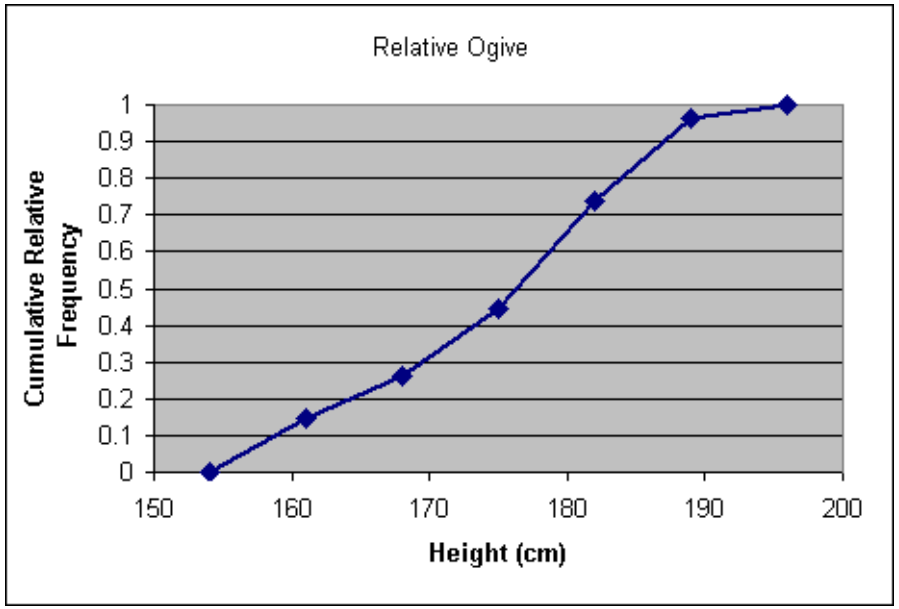
# Ogive

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



# Example

| 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)

# Stem-and-Leaf

- 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.

# Stem-and-Leaf

- What to look for

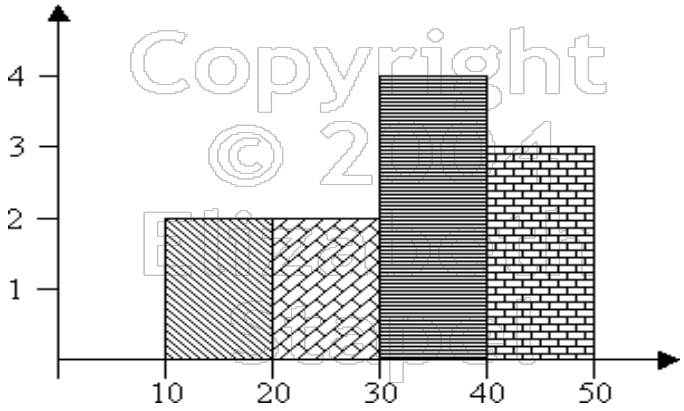
  The display conveys information about

  - 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

# Histogram vs Stem-and-leaf

- Given the following data: 12, 13, 21, 27, 33, 34, 35, 37, 40, 40, 41

| Frequency class | Frequency |
|---|---|
| 10 - 19 | 2 |
| 20 - 29 | 2 |
| 30 - 39 | 4 |
| 40 - 49 | 3 |




| stem | leaf |
|---|---|
| 1 | 2 3 |
| 2 | 1 7 |
| 3 | 3 4 5 7 |
| 4 | 0 0 1 |

**Stem – left-hand column that contains the tens digit**

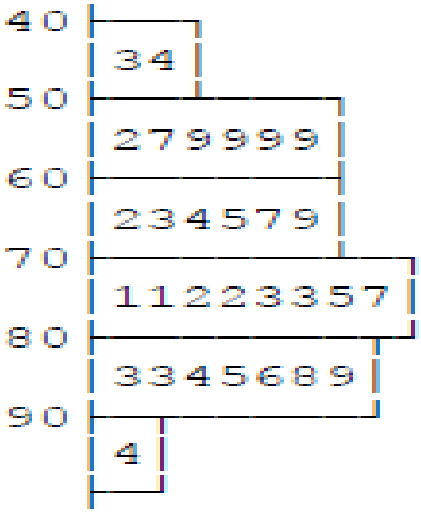
**Leaves – right hand column listing the one's digit for each of the tens digit (stem values)**

# Histogram vs Stem-and-leaf

- In stem-and-leaf:
  - Original values can still be determined
  - Horizontal leaves in the stem-and-leaf plot correspond to the vertical bars in the histogram
  - The leaves have lengths equal to the numbers in the frequency table.

# Stem-and-leaf

- Stem-and-leaf can be converted to histogram:
  - Just place bars over each list in the leaves. Example (with different data set):



# Stem-and-leaf

- Some issues:
  - What about when smallest and largest values in data set is big (e.g. 418 to 1272)? What should be the suitable values for stem?
    - Choise for stem values could be 100s, so that the stem values would be:

      4 5 6 7 8 9 10 11 12
  - It is nice to add legend to indicate stem values are multiple of 10, 100, 0.01, etc. Example of legend stem-and-leaf with legend:

```
4 | 34
5 | 279999
6 | 234579
7 | 11223357
8 | 3345689
9 | 4
```

NOTE: 6 | 3 means 63 ← **legend**

# Example

- Data set

| | | |
|---|---|---|
| 109 | 96 | 94 |
| 95 | 93 | 93 |
| 89 | 84 | 80 |
| 78 | 73 | 70 |
| 55 | 41 | 40 |

| | |
|---|---|
| **10** | **9** |
| **9** | **33456** |
| **8** | **049** |
| **7** | **038** |
| **6** | |
| **5** | **5** |
| **4** | **01** |

Stem: Tens
Leaf: Ones

# Example

- Data set

| | |
|---|---|
| **0.553** | **0.570** |
| **0.576** | **0.601** |
| **0.606** | **0.606** |
| **0.609** | **0.611** |
| **0.615** | **0.628** |
| **0.654** | **0.662** |
| **0.670** | **0.672** |
| **0.690** | **0.693** |
| **0.749** | **0.844** |
| **0.933** | |

**SPSS Output**

Stem-and-Leaf Plot

| Frequency | Stem & | Leaf |
|---|---|---|
| 3.00 | 5 . | 577 |
| 7.00 | 6 . | 0000112 |
| 6.00 | 6 . | 567799 |
| 1.00 | 7 . | 4 |
| 2.00 | Extremes | (>=.84) |

Stem width: .100
Each leaf: 1 case(s)

# Example

- Data set

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 3977 | 3423 | 3586 | 4010 | 3785 | 3447 | 5761 | 6176 |
| 2773 | 3239 | 3239 | 3323 | 5653 | 5384 | 4991 | 3686 |
| 3208 | 5011 | 6230 | | | | | |

| | |
|---|---|
| 2 | 773 |
| 3 | 977,423,586,785,447,239,239,323,686,208 |
| 4 | 010,991 |
| 5 | 761,653,384,011 |
| 6 | 176,230 |

Stem: Thousands
Leaf: Ones

# Example

- Data set

| | |
|---|---|
| 3977 | 3423 |
| 3586 | 4010 |
| 3785 | 3447 |
| 5761 | 6176 |
| 2773 | 3239 |
| 3239 | 3323 |
| 5653 | 5384 |
| 4991 | 3686 |
| 3208 | 5011 |
| 6230 | |

**SPSS Output**

Stem-and-Leaf Plot

| Frequency | Stem & Leaf |
|---|---|
| 1.00 | 2 . 7 |
| 6.00 | 3 . 222344 |
| 4.00 | 3 . 5679 |
| 1.00 | 4 . 0 |
| 1.00 | 4 . 9 |
| 2.00 | 5 . 03 |
| 2.00 | 5 . 67 |
| 2.00 | 6 . 12 |

Stem width: 1000.00

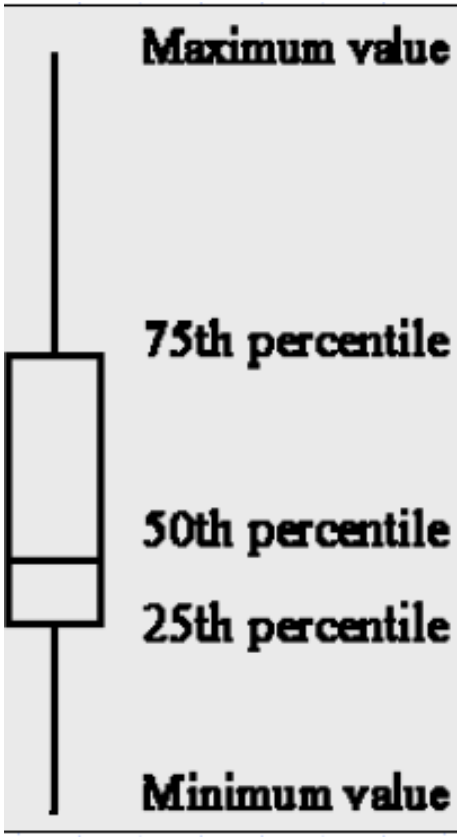
Each leaf: 1 case(s)

# Box Plots

- It would be nice to have a method of summarizing 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.
- A boxplot is one such technique.
- It is compact, yet it provides information about center, spread, and symmetry or skewness of the data.

# Box Plots

- 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.

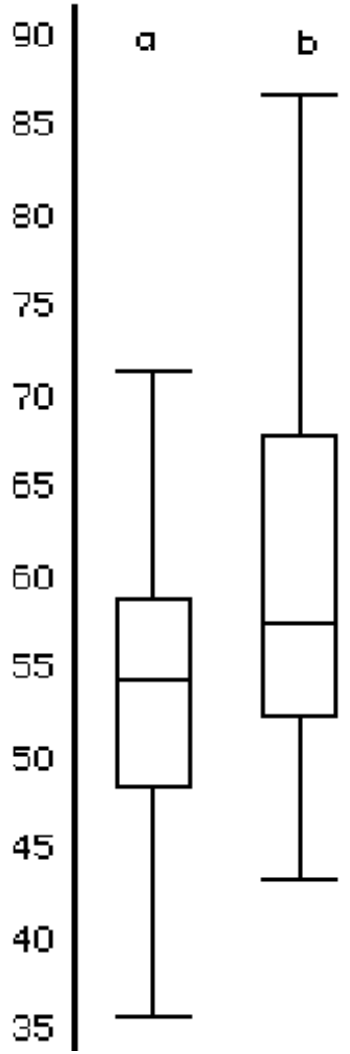


# Example

It is often useful to compare data from two or more groups by viewing box plots from the groups side by side.

The data from *b* are higher, more spread out, and have a positive skew.

That the skew is positive can be determined by the fact that the upper whisker is longer than the lower whisker.



# Box Plots

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

# 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

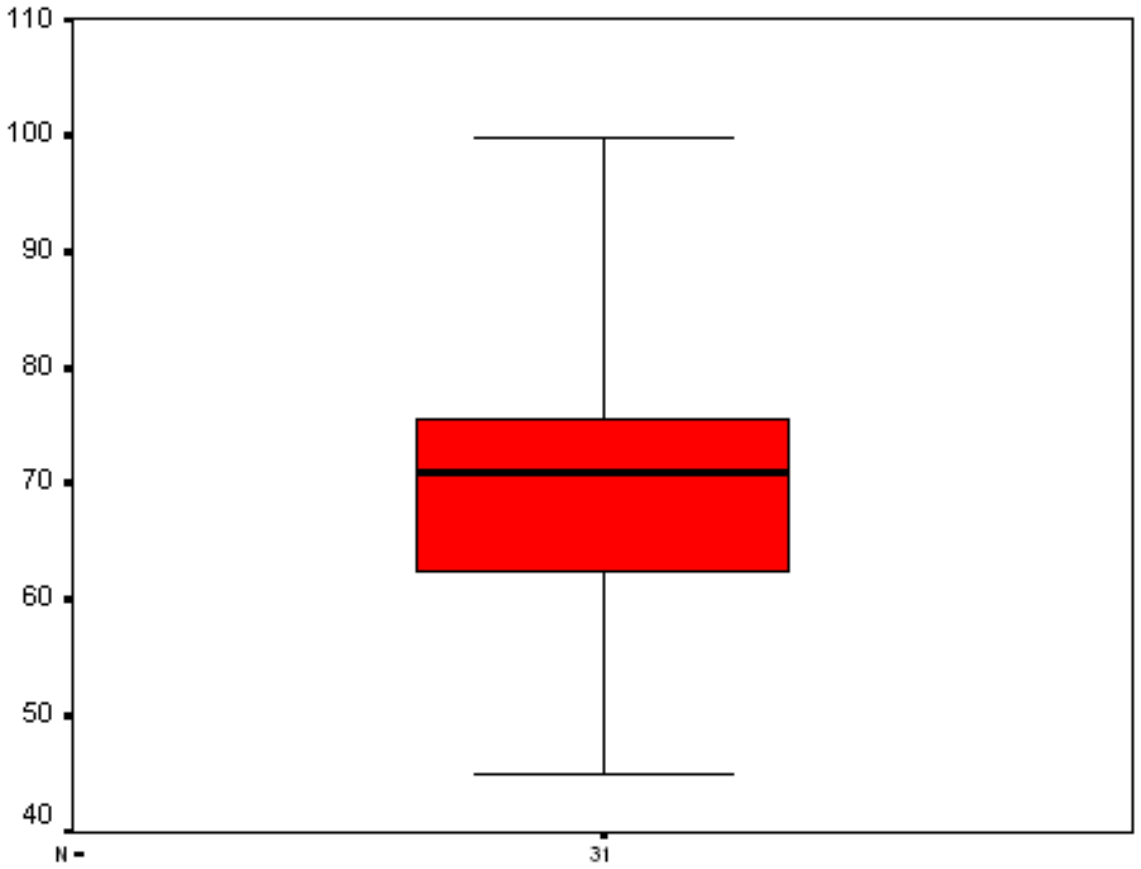
- Data set:

45 48 50 54 57 60 60 62 63 63 64 65 67 68 69
71 (median) 71 72 72 74 74 75 75 76 80 83 84
88 100 100 100

  - Smallest observation = 45,
  - Lower quartile = 62
  - Median = 71
  - Upper quartile = 76
  - Largest observation = 100

# Example

Smallest observation = 45,
Lower quartile = 62
Median = 71
Upper quartile = 76
Largest observation = 100



# 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 outliers by solid circles "●" and extreme outliers by open circles "○" or sometimes asterisk "*", and the whiskers extend on each end to the most extreme observations that are not outliers.

# Box Plots

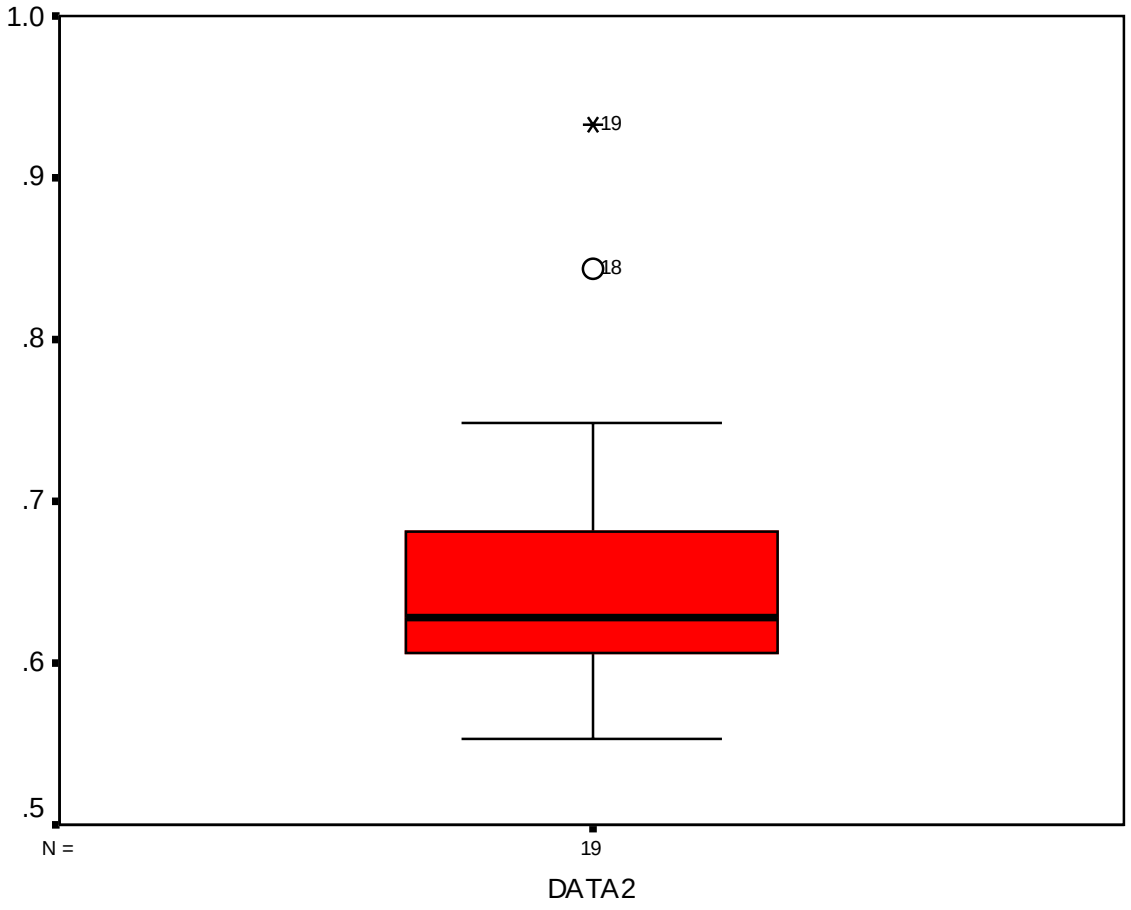
- Construction of a Modified Boxplot
  - Draw a vertical (or horizontal) measurement scale
  - Construct a rectangular box with a lower (or left) edge at the lower quartile and a upper (or right) edge at the upper quartile.

    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 in the data set
  - Draw a * to mark the location if any extreme outliers in the data set

# Example

## Data Set

| | |
|---|---|
| 0.553 | 0.570 |
| 0.576 | 0.601 |
| 0.606 | 0.606 |
| 0.609 | 0.611 |
| 0.615 | 0.628 |
| 0.654 | 0.662 |
| 0.670 | 0.672 |
| 0.690 | 0.693 |
| 0.749 | 0.844 |
| 0.933 | |



# 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** |

# Exercise

- Construct a histogram for the following data:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 57 | 61 | 57 | 57 | 58 | 57 | 61 | 54 |
| 68 | 51 | 49 | 64 | 50 | 48 | 65 | 52 |
| 56 | 46 | 54 | 49 | 51 | 47 | 55 | 55 |
| 54 | 42 | 51 | 56 | 55 | 51 | 54 | 51 |
| 60 | 62 | 43 | 55 | 56 | 61 | 52 | 69 |
| 64 | 46 | | | | | | |

# Exercise

- Construct two histograms corresponding to the frequency table of samples of students cars' age and faculty/staff cars' age obtained from a college. Compare both of them.

| Age | Students | Faculty/Staff |
|---|---|---|
| 0 to < 3 | 23 | 30 |
| 3 to < 6 | 33 | 47 |
| 6 to < 9 | 63 | 36 |
| 9 to < 12 | 68 | 30 |
| 12 to < 15 | 19 | 8 |
| 15 to < 18 | 10 | 0 |
| 18 to < 21 | 1 | 0 |
| 21 to < 24 | 0 | 1 |

# Exercise

- Construct a stem-and-leaf display for the following data set: (*Hint: First round the data to the nearest whole number*)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 63.2 | 67.7 | 59.1 | 64.4 | 88.2 | 79.5 | 81.4 | 59.5 |
| 73.7 | 92.0 | 93.2 | 87.6 | 31.8 | 73.0 | 67.4 | 70.2 |
| 71.0 | 85.9 | 99.0 | 63.8 | 79.5 | 96.9 | 71.9 | 61.6 |
| 54.6 | 63.6 | 85.6 | 67.3 | 83.8 | 70.0 | 67.7 | 73.1 |
| 75.6 | 65.2 | 71.7 | 83.8 | 75.5 | 46.5 | 63.9 | 80.0 |

# Exercise

- The following list shoes the ages of most of the employees at the Vita Needle Company. Construct a boxplot.

| | | | | |
|---|---|---|---|---|
| 76 | 45 | 72 | 77 | 63 |
| 65 | 87 | 73 | 84 | 86 |
| 79 | 86 | 75 | 87 | 74 |
| 39 | 75 | 41 | 82 | 34 |
| 88 | 85 | 79 | 73 | 53 |

# PAST YEAR

Fifteen (15) people were asked to state the number of hours they exercise in a seven days period. The results of the survey are listed below.

**8, 2, 4, 7.5, 10, 11, 5, 6, 8, 12, 11, 9, 6.5, 10.5, 13**

i) Complete the frequency table below (Table 1). (5 marks)

Table 1

| Hours of Exercise | Tally | Frequency |
|---|---|---|
| 0 – 2 | | |
| 3 – 5 | | |
| 6 – 8 | | |
| 9 – 11<br>12 – 14 | | |

ii) Plot a histogram to display the data. (3 marks)

iii) Make some conclusions based on the information displayed in the histogram.

(2 marks)

- ***Project 1 Briefing***
- - grouping
- ***DATA***