Chapter 2: Descriptive Statistics: Tabular and Graphical Presentations

2.1 Summarizing Qualitative Data (labels or names)

* We summarize qualitative data by tabular and graphical methods

Frequency distribution: is a tabular summary of data showing the number (frequency) of items in each several non overlapping classes (categories).

Relative frequency

$$\text{Relative frequency of a class} = \frac{\text{frequency of the class}}{n}$$

where n is the number of observations (sample size)

Percent frequency of a class = relative frequency X 100.

* Note that:
  1. Sum of the frequencies in any frequency distribution = n
  2. Sum of relative frequencies in any relative frequency distribution = 1
  3. Sum of of the percentages in a percent frequency distribution = 100

Example: The following table is a frequency distribution that samaries how 50 persons are distributed across five soft drinks:

Soft drink	Frequency
Fanta	19
Cola	8
7up	5
Pepsi	13
Sprite	5

Frequency distribution

(a) * Find the sample size (number of observations)

(b) * Find the relative frequency of each category (class)

(c) * Find the percent frequency for each category.

Soft drink	frequency	Relative frequency	Percent frequency
Fanta	19	$\frac{19}{50} = \frac{38}{100} = 0.38$	$0.38 \times 100 = 38$
Cola	8	$\frac{8}{50} = \frac{16}{100} = 0.16$	$0.16 \times 100 = 16$
7 up	5	$\frac{5}{50} = \frac{10}{100} = 0.10$	$0.10 \times 100 = 10$
Pepsi	13	$\frac{13}{50} = \frac{26}{100} = 0.26$	$0.26 \times 100 = 26$
Sprite	5	$\frac{5}{50} = \frac{10}{100} = 0.10$	$0.10 \times 100 = 10$
	$n = 50$	total = 1.00	total = 100

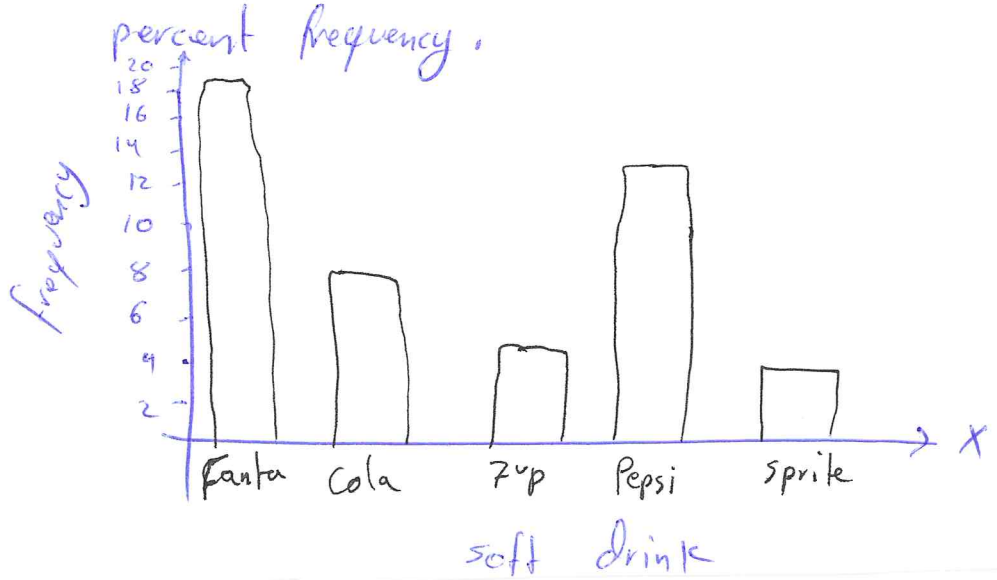
(a) number of observations $(n = 50)$

Bar Graphs (qualitative data)

A bar graph or bar chart is a graphical device for depicting data summarized in a frequency, relative frequency or percent frequency distribution.

Horizontal axis: we specify the labels that are used for the classes (categories).

Vertical axis: we specify a frequency, or relative frequency or percent frequency.



Pie chart

provide another graphical device for presenting relative frequency and percent frequency distributions for qualitative data.

To construct a pie chart:

1) draw a circle
2) divid a circle into parts

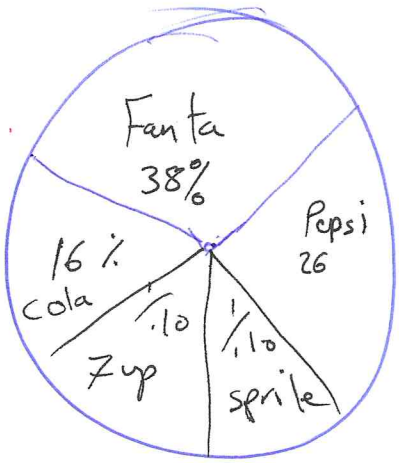
Fanta: $0.38 \times 360^\circ = 136.8^\circ$

Cola: $0.16 \times 360^\circ = 57.6^\circ$

7up: $0.10 \times 360^\circ = 36.0^\circ$

Pepsi: $0.26 \times 360^\circ = 93.6^\circ$

Sprite: $0.10 \times 360^\circ = 36^\circ$



2.2 Summarizing Quantitative data:

* We summarize quantitative data by tabular and graphical methods.

Frequency distribution: is a tabular summary of data showing the number (frequency) of item in each several non overlapping classes (categories).

* This holds for quantitative and qualitative data

* But for quantitative data: there are three steps to define the classes for a frequency distribution:

1. **Number of nonoverlapping classes:** usually (5-20) classes
   - For small data we use 5 classes.
   - For large data we need more classes.
2. **Width of each class** = $\frac{\text{Largest data value} - \text{Smallest data value}}{\text{Number of classes}}$
   - can be approximately accordingly to the preference of the person developing the frequency distribution. $9.2 \approx 10$
3. **class limits:** must be chosen so that each data item belongs to one and only one class.

* The lower class limit: identifies the smallest possible data value assigned to the class

* The upper class limit: = = largest = = = = = =

Class midpoint: is the value halfway between the lower and upper class limits.

Example: Consider the following data

12	14	19	18
15	15	18	17
20	27	22	23
22	21	33	28
14	18	16	13

* Consider the number of class = 5
* largest value = 33
* Smallest value = 12
* Width of each class = $\frac{33-12}{5} = 4.2 \approx 5$

The class (category)	Frequency	Relative frequency	Percent frequency
10 – 14 (lower class limit ← 10, upper class limit → 14)	4	$\frac{4}{20} = \frac{20}{100} = 0.20$	$0.20 \times 100 = 20\%$
15 – 19	8	$\frac{8}{20} = \frac{40}{100} = 0.40$	$0.40 \times 100 = 40\%$
20 – 24	5	$\frac{5}{20} = \frac{25}{100} = 0.25$	$0.25 \times 100 = 25\%$
25 – 29	2	$\frac{2}{20} = \frac{10}{100} = 0.10$	$0.10 \times 100 = 10\%$
30 – 34	1	$\frac{1}{20} = \frac{5}{100} = 0.05$	$0.05 \times 100 = 5\%$
	$n = 20$ total	Total = 1.00	Total = 100%

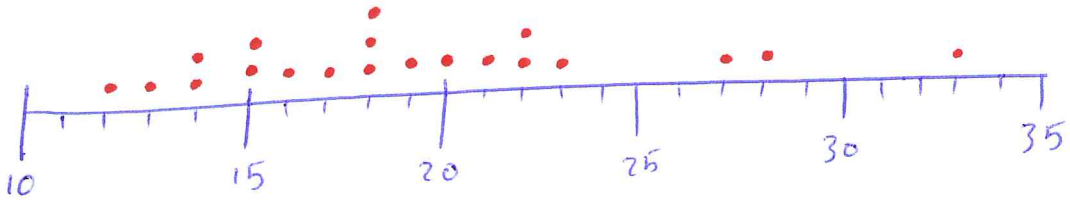
↖ this table is relative frequency and percent frequency distributions for the data above.

The midpoints of the classes

10 – 14 → 12
15 – 19 → 17
20 – 24 → 22
25 – 29 → 27
30 – 34 → 32

* Relative frequency of class = $\frac{\text{Frequency of the class}}{n}$
* Percent frequency of a class = relative frequency of class × 100
* True limits
  - upper true limits: 14.5, 19.5, 24.5, 29.5, 34.5
  - lower true limits: 9.5, 14.5, 19.5, 24.5, 29.5

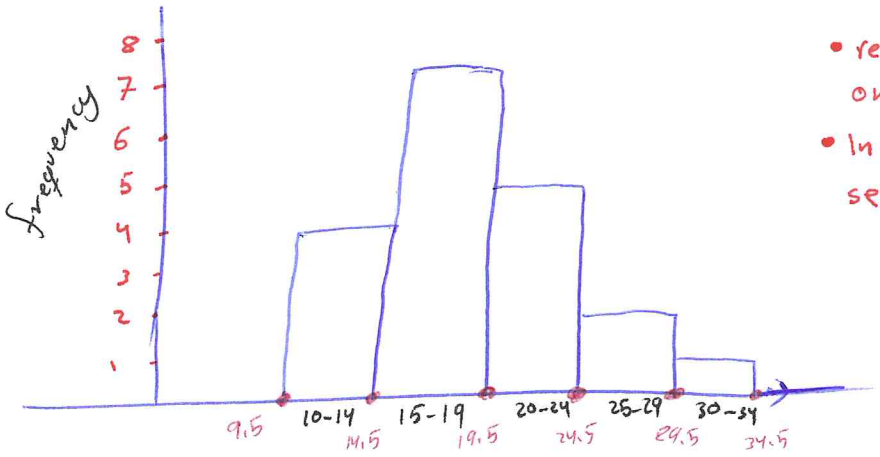
Dot Plot : one of the simplest graphical summaries of data. The horizontal axis shows the range for the data.



Histogram

- to represent quantitative data.
- can be prepared for data summarized in frequency, relative frequency or percent frequency distribution.

→ The variable of intrest on the horizontal axis and the frequency (or relative or percent frequency) on vertical axis.

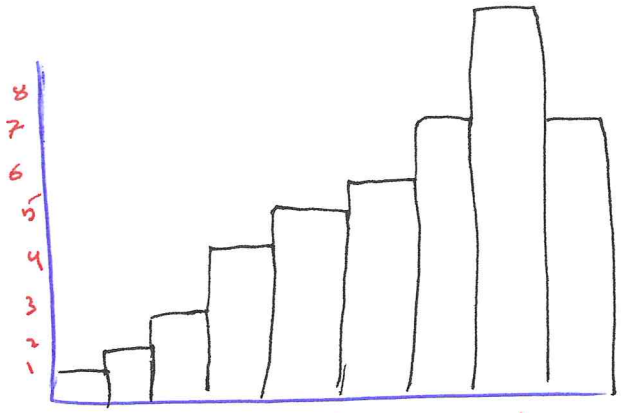


- rectangles are on touch one another in histogram
- In Bar, there is a separation.

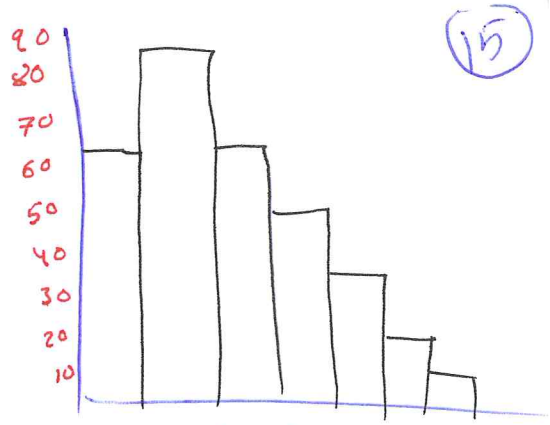
→ Histogram provides information about the shape or form of a distribution.

* A histogram is said to be skewed to the left if its tail extends farther to the left

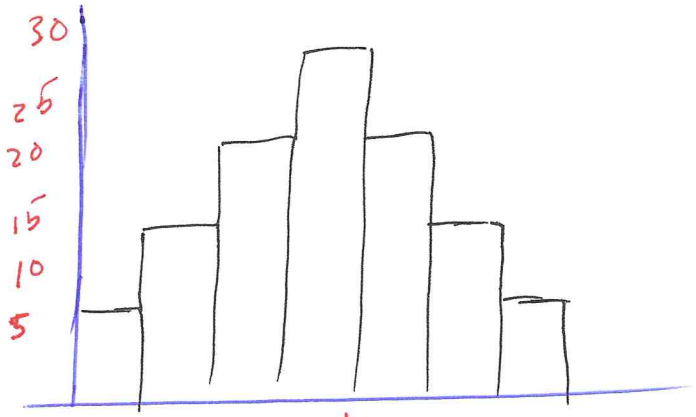
* = = = = = = = = right = = = = = = = right.



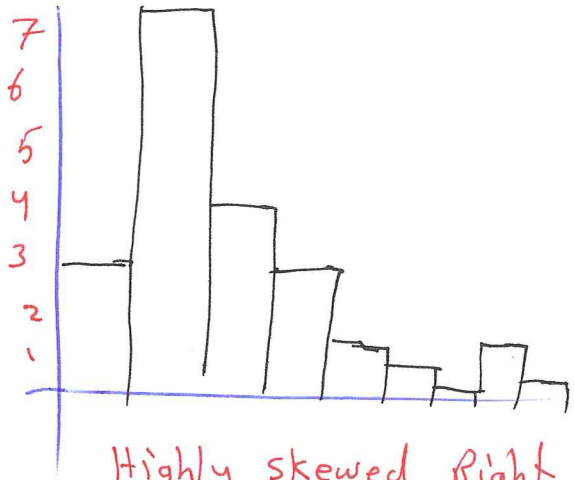
Moderately skewed left



Moderately skewed Right



Symmetric



Highly skewed Right

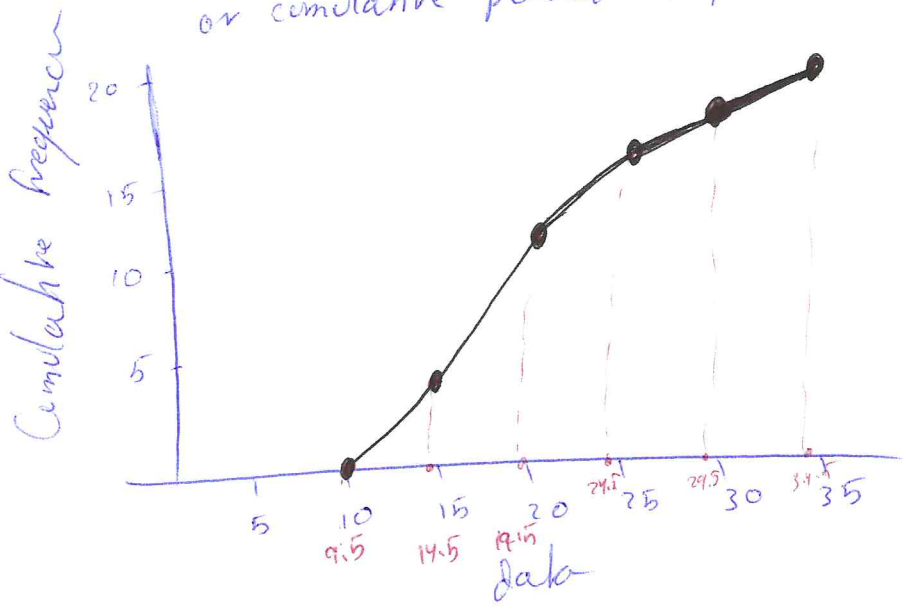
Cumulative distributions: the cumulative frequency distribution shows the number of data with values less than or equal to the upper class limit of each class.

Cumulative classes	Cumulative frequency	Cumulative relative frequency	Cumulative percent freq.
less than or equal 14	4	$4/20 = \frac{20}{100} = 0.20$	$0.2 \times 100 = 20\%$
less than or equal 19	12	$\frac{12}{20} = \frac{60}{100} = 0.60$	$0.60 \times 100 = 60\%$
less than or equal 24	17	$\frac{17}{20} = \frac{85}{100} = 0.85$	$0.85 \times 100 = 85\%$
less than or equal 29	19	$\frac{19}{20} = \frac{0.95}{100} = 0.95$	$0.95 \times 100 = 95\%$
less than or equal 34	$20 = n$	$\frac{20}{20} = \frac{100}{100} = 1.00$	$1.00 \times 100 = 100\%$

* Note that the cumulative frequency distribution shows the proportion of data item.
* Note that the cumulative percent frequency distribution shows the percentage of data item with values less than or equal to the upper limit of each class.

Ogive : is the graph of a cumulative distribution.

- Data values are on horizontal axis.
- Cumulative frequency or cumulative relative frequency or cumulative percent frequencies on the vertical axis.



2.3 The stem-and-leaf Display : Exploratory data analysis:

- The stem-and-leaf display can be used to show the rank order and shape of a data set.
- The stem-and-leaf display is an example of the exploratory data analysis (Methods that use simple arithmetic and easy to draw graphs to summarize data quickly).
- We put the leading digists in a vertical line (to the left)
- We put the last digists of each data value to the right of the vertical line.

Example: Construct a stem-and-leaf display for the following data

108	83	97	101	110
118	95	100	126	117
121	106	116	91	123
82	101	119	122	102

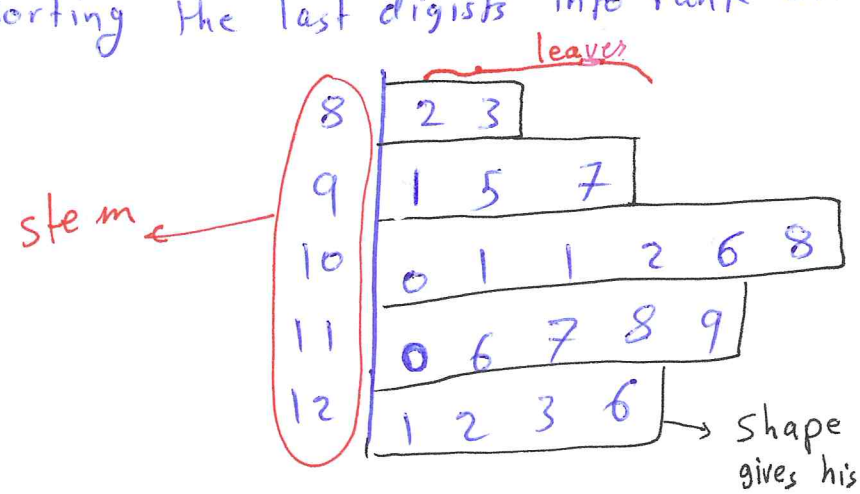
leading digists

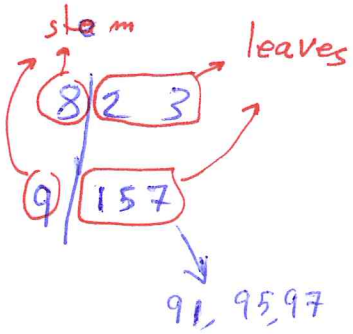
8	2 3
9	5 7 1
10	8 6 1 0 1 2
11	8 6 9 0 7
12	1 6 2 3

121 → leading digists: 12, last digit: 1

83

sorting the last digists into rank order





Shape: rotating this counterclockwise gives histogram with classes:
80-89, 90-99, 100-109, 110-119, 120-129

* The stem-and-leaf provides the same information as histogram. It has two advantages

[1] The stem-and-leaf display is easier to construct by hand.

[2] Within a class interval, the stem-and-leaf display provides more information than histogram because the stem-and-leaf shows the actual data.

---

stretched stem-and-leaf

* Whenever a stem value is stated twice, it means that the first value corresponds to leaf values of 0-4 and the second value corresponds to leaf values of 5-9

8	2 3
8	
9	1
9	5 7
10	0 1 1 2
10	6 8
11	0
11	6 7 8 9
12	1 2 3
12	6

This is similar to a frequency distribution with intervals

80-84
85-89
90-94
95-99
⋮

Example: Construct a stem-and-leaf display for the following data. Use a leaf unit of 10.

1565 1852 1644 1766 1888 1912 2044 1812
1790 1679 2008 1852 1967 1954 1733

leaf unit = 10

15	6
16	4 7
17	3 6 9
18	1 5 5 8
19	1 5 6
20	0 4

156 × 10 = 1560 this is approximation for 1565 "not exact data"

- single digit used to defined each leaf
- The first 3 digits used in this steam-and-leaf display.
- leaf units can be 100, 10, 1, 0.1, ...
- if the leaf unit is not shown, then it is assumed to be 1.

2.4 Crosstabulations and Scatter Diagrams

* Crosstabulations and scatter diagrams are used to summarize data in a way that reveals the relationship between two variables.
* Crosstabulation is a tabular summary of data for 2 variables.

Example: Consider the following quality rating and meal price for 300 restaurants:

Each restaurant provides a quality rating and a meal price.

Restaurant	Quality Rating	Meal price ($)
1	Good	18
2	Very Good	22
3	Good	28
4	Excellent	38
5	Very Good	33
⋮	⋮	⋮
300	Good	13

a) Construct a crosstabulation for the data
b) Develop a relative and percent frequency distribution for quality rating
c) = meal price
d) Construct a row percenteges for each quality rating category
d) What is the relation ship between the quality rating and meal price
(ie) = meal price
= Column

- Quality rating: is qualitative variable with categories; good, very good, excellent.
- Meal price: is quantitative variable that ranges from 10$ to 49$

Quality rating	Meal Price $10-19	$20-29	$30-39	$40-49	Total
Good	42	40	2	0	84
Very Good	34	64	46	6	150
Excellent	2	14	28	22	66
Total	78	118	76	28	300

Column ⇓
→ row total
row ⇒
→ crosstabulation
column total

* For example restaurant 5 provides a quality rating very good with meal price $33. This restaurant belongs to the cell in row 2 and column 3.

- The greatest number of restaurants in the sample is 64 have a very good rating and meal price in $20 – 29 range.
- Only 2 restaurants with excellent rating and meal price in $10 – 19 range.

(b)

Quality rating	Relative frequency	Percent frequency
Good	$\frac{84}{300} = 0.28$	$0.28 \times 100 = 28$
Very Good	$\frac{150}{300} = 0.50$	$0.50 \times 100 = 50$
Excellent	$\frac{66}{300} = 0.22$	$0.22 \times 100 = 22$
	Total = 1.00	Total = 100

28% of the restaurant were rating good.
50% " " " = very good.
22% " " " " = excellent.

(c)

Meal Price	Relative frequency	Percent frequency
$10 – 19	$\frac{78}{300} = 0.26$	$0.26 \times 100 = 26$
$20 – 29	$\frac{118}{300} = 0.39$	$0.39 \times 100 = 39$
$30 – 39	$\frac{76}{300} = 0.25$	$0.25 \times 100 = 25$
$40 – 49	$\frac{28}{300} = 0.09$	$0.09 \times 100 = 9$
	Total = 1.00	Total = 100

26% of the meal price are in the lowest class $10 – 19
39% " " " " " " " = next higher class and so on

(e) Higher meal prices are associated with the higher quality rest.
(f) lower meal prices " " " " lower " "

→ to know the relationship between the two variables within cross tabulation, we convert the entries into row percentages or column percentages.

(d)

Quality Rating \ Meal price	$10 – 19	$20 – 29	$30 – 39	$40 – 49	Total
Good	$\frac{42}{84} \times 100 = 50$	$\frac{40}{84} \times 100 = 47.6$	$\frac{2}{84} \times 100 = 2.4$	$\frac{0}{84} \times 100 = 0.0$	100
Very Good	$\frac{34}{150} \times 100 = 22.7$	$\frac{64}{150} \times 100 = 42.7$	$\frac{46}{150} \times 100 = 30.6$	$\frac{6}{150} \times 100 = 4.0$	100
Excellent	$\frac{2}{66} \times 100 = 3.0$	$\frac{14}{66} \times 100 = 21.2$	$\frac{28}{66} \times 100 = 42.4$	$\frac{22}{66} \times 100 = 33.4$	100

↑ Row percentages for each quality rating category

* For the lowest quality restaurant (good), we see the greatest percentages are for the less expensive restaurants (50% have $10–19 meal prices and 47.6% have $20–29 meal prices)...
* For the greatest quality restaurants (excellent), we see the greatest percentages are for the more expensive restaurants (42.4% have $30–39 meal prices and 33.4% have $40–49 meal prices).

(e)

Quality Rating \ Meal price	$10 – 19	$20 – 29	$30 – 39	$40 – 49
Good	$\frac{92}{78} \times 100 = 54$	$\frac{40}{118} \times 100 = 34$	$\frac{2}{76} \times 100 = 3$	$\frac{0}{28} \times 100 = 0$
V. Good	$\frac{34}{78} \times 100 = 44$	$\frac{64}{118} \times 100 = 54$	$\frac{46}{76} \times 100 = 61$	$\frac{6}{28} \times 100 = 21$
Excellent	$\frac{2}{78} \times 100 = 2$	$\frac{14}{118} \times 100 = 12$	$\frac{28}{76} \times 100 = 36$	$\frac{22}{28} \times 100 = 79$
	100	100	100	100

↑ Column percentages for each ↓ category (meal price)

Simpson's Paradox

Conclusions drawn from two or more separate cross tabulations that can be reversed when data are aggregated into a single cross tabulation.

Example: Consider the following two cross tabulations

Cross tabulation for School 1

Gender	10th class	5th class	
M	29 (91%)	100 (85%)	129
F	3 (9%)	18 (15%)	21
	32 (100%)	118 (100%)	150

Cross tabulation for school 2

Gender	10th class	5th class	
M	90 (90%)	20 (80%)	110
F	10 (10%)	5 (20%)	15
	100 (100%)	25 (100%)	125

Simpson's Paradox:

Gender	School 1	School 2	
M	129 (86%)	110 (88%)	239
F	21 (14%)	15 (12%)	36
	150 (100%)	125 (100%)	275

In Simpson's paradox, we need to be careful, when drawing conclusions using aggregating data.

Hidden variable is 10th class and 5th class.

Scatter Diagram and Trendline

- A scatter diagram is a graphical presentation of the relationship between two quantitative variables.
- Trendline: is a line that provides an approximation of the relationship.

Example: The following 10 observation are for two quantitative variables x: number of commercials
y: sales ($100s)

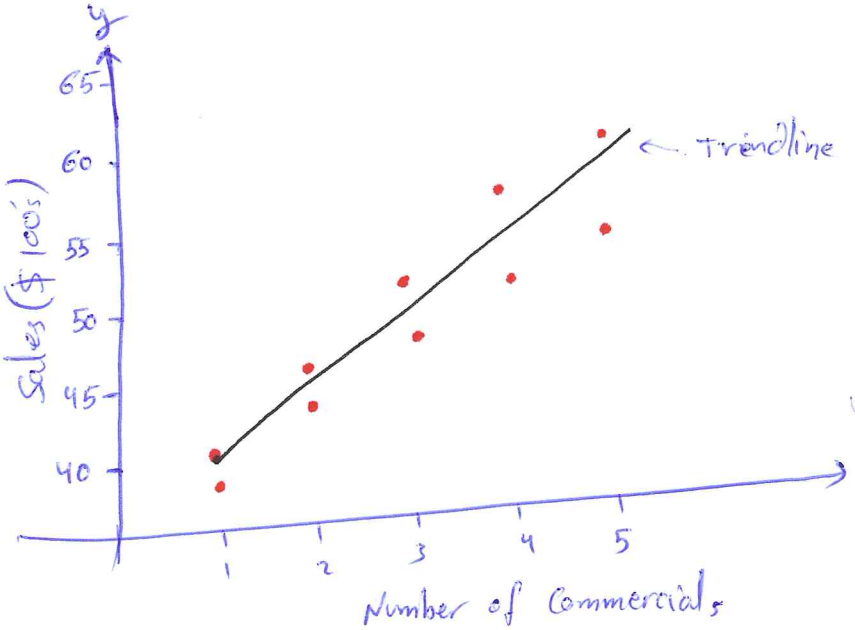
Number of Commercials (x)	Sales ($100s) y
2	50
5	57
1	41
3	54
4	54
1	38
5	63
3	48
4	59
2	46

a) Develop a scatter diagram for the relationship between x and y.

b) What is the relationship, if any, between x and y?

(c) Is the relation perfect?

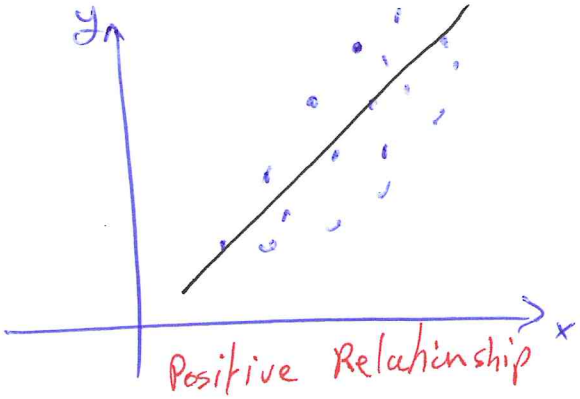
(a)



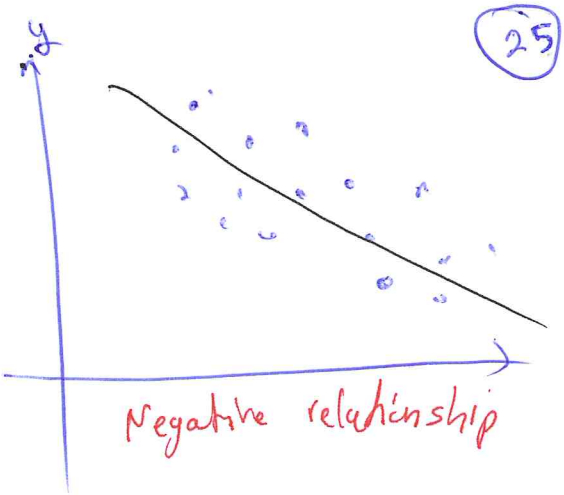
(b) The scatter diagram indicates a positive relationship between x and y.

- Higher sales associated with higher number of commercials.

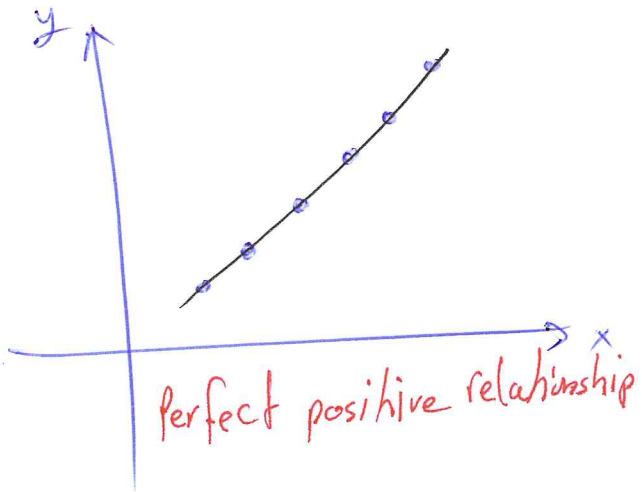
(c) The relation is not perfect. Because not all the points are not on the trendline



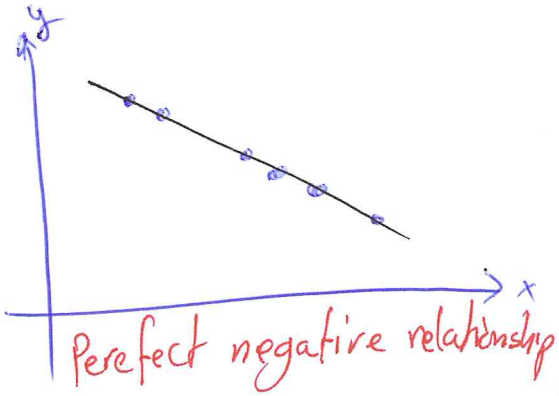
Positive Relationship



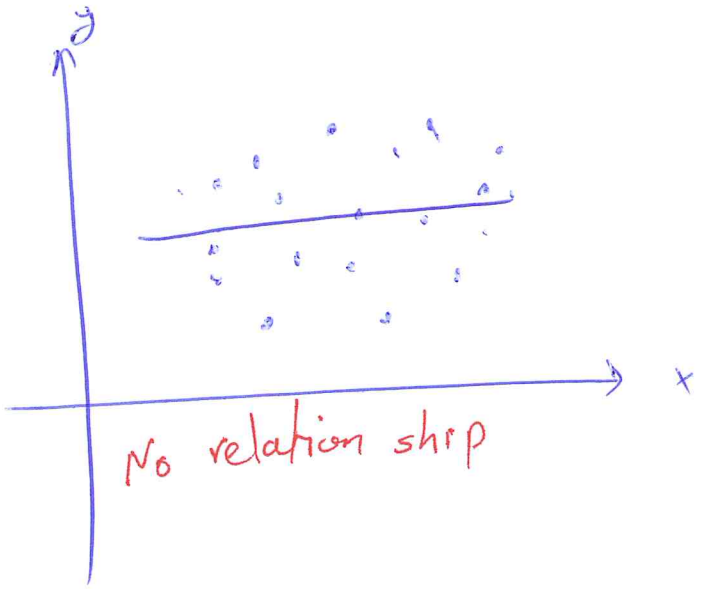
Negative relationship



Perfect positive relationship



Perefect negative relationship



No relation ship