STAT2371

Chapter 2

Descriptive Statistics: Tabular and Graphical Presntations

- Frequency distribution: A table that organizes the data of a variable into classes and frequencies.
- Class: A quantitative or qualitative category of a variable.
- Frequency: The number of data values of a variable in each class.
- Types of frequency distributions:

  ⋆ For qualitative data (variables):

  (a) **Frequency table (F table)**: Classes vs. frequencies.

  (b) **Relative frequency table (RF table)**: Classes vs. relative frequencies.

  (c) **Percent frequency table (PF table)**: Classes vs. percent frequencies.

  ⋆ For quantitative data (variables):

  (a) **Frequency table (F table)**: Classes vs. frequencies.

  (b) **Relative frequency table (RF table)**: Classes vs. relative frequencies.

  (c) **Percent frequency table (PF table)**: Classes vs. percent frequencies.

  (d) **Cumulative frequency table (CF table)**: Upper limits vs. cumulative frequencies.

First: **Summarizing Qualitative Data.**

(a) Frequency tables.

Ex: The following frequency table represents the blood type of a sample of 25 students.

Blood type	**F**
A	5
B	7
O	9
AB	4
Total	25

Note. (1) The table consists of four classes: A (first class), B (second class), O (third class), and AB (fourth class)

(2) The classes are disjoint (non overlapping)

(3) # classes = # quantitative variable values.

(4) Each class has its own frequency which is # students who belong to the class.

(5) Total Frequency (n): is the total number of students of interest which is 25

(6) $\sum F = n =$ number of elements = number of observations = sample size.

(b) Relative frequency tables.

Ex: The RF table for the previous example is below.

Blood type	**RF**
A	0.20
B	0.28
O	0.36
AB	0.16
Total	1

Notes. (1) For each class: $RF = \frac{F}{n}$

(2) $\sum RF = 1$

(3) The relative frequency represents a proportion (probability).

(c) Percent frequency tables.

Ex: The PF table for the previous example is below.

Blood type	**PF**
A	20 %
B	28 %
O	36 %
AB	16 %
Total	100 %

Note. (1) For each class: $PF = \left(\frac{F}{n} \times 100\right)\% = (RF \times 100)\%$

(2) $\sum PF = 100\%$

Remark. We can combine all tables in one table.

Class	**F**	**RF**	**PF**
A	5	0.20	20 %
B	7	0.28	28 %
O	9	0.36	36 %
AB	4	0.16	16 %
Total	25	1	100 %

• Questions about these tables:

Q1: What is the number of students whose blood type is O? 9

Q2: How many students with blood type A **or** B? $5 + 7 = 12$

Q3: How many students with blood type A **and** B? zero

Q4: What is the percentage of students whose blood type is AB? 16%

Q5: What is the probability (proportion) that a student has blood type B? 0.28

How to construct these tables for qualitative variables?

Answer: Determine the values of the variable (classes), then start counting.

∗ **Question:** The following data represent the size (small=S, medium=M, large=L) for 20 T-shirts.

M, M, S, L, L, S, S, S, S, M, L, M, M, M, L, L, L, L, L, S

Construct F, RF, PF tables for these data.

Solution:

Class	**F**	**RF**	**PF**
S	6	0.30	30 %
M	6	0.30	30 %
L	8	0.40	40 %
Total	20	1	100 %

Second: **Summarizing Quantitative Data.**

(i) F, RF, PF tables.

Ex: The following frequency table represents the ages of a sample of 40 BZU employees.

Age	**F**
22 - 30	10
31 - 39	8
40 - 48	7
49 - 57	9
58 - 66	6
Total	40

Note. (1) The table consists of five disjoint classes: 22 - 30 (first class), 31 - 39 (second class), ...

(2) Each class consists of two limits: **lower limit** and **upper limit**.
For example: the first class 22 -30 → lower limit=22, upper limit=30, and so on.

(3) Each class has a **midpoint**. $\left(\text{midpoint} = \frac{\text{upper limit}+\text{lower limit}}{2}\right)$.
For example: the first class 22 - 30 → midpoint $= \frac{22+30}{2} = 26$, and so on.

(4) The F table has a **class width**.
Class width = the difference between two consecutive lower (or upper) limits.
For our table: Class width $= 31 - 22 = 9$ or $39 - 30 = 9$ or $49 - 40 = 9, \ldots$

(5) The total frequency $= n = 40$

Remark. We can construct RF and PF tables as before.

Age	**RF**	**PF**
22 - 30	0.25	25%
31 - 39	0.20	20%
40 - 48	0.175	17.5%
49 - 57	0.225	22.5%
58 - 66	0.15	15%
Total	1	100%

(ii) CF tables.

Ex: The following is the CF table for the previous example.

Age	**CF**
less than or equal 30 (≤ 30)	10
less than or equal 39 (≤ 39)	18
less than or equal 48 (≤ 48)	25
less than or equal 57 (≤ 57)	34
less than or equal 66 (≤ 66)	40

Notes. (1) The CF table is only for quantitative variables.

(2) The last CF = n

How to construct these tables for quantitative variables?

$*$ **Question:** The following data represents the weights (kg) of 16 BZU students.

52, 60, 79, 60, 71, 45, 64, 74, 80, 59, 91, 50, 60, 76, 84, 75

(a) Construct F, RF, PF tables for these data. (Use four classes)

Solution: (1) Find the **maximum** value and the **minimum** value. Max=91, Min=45

(2) Determine the class width:

Find $\frac{\text{Max}-\text{Min}}{\#\text{classes}} = \frac{91-45}{4} = \frac{46}{4} = 11.5 \rightarrow$ Class width = next integer = 12

(3) Start constructing the classes by finding the lower limits starting from the minimum value using the class width, then find the upper limits.

(4) Start counting to find the frequency for each class.

Class	**F**	**RF**	**PF**
45 - 56	3	0.1875	18.75%
57 - 68	5	0.3125	31.25%
69 - 80	6	0.375	37.5%
81 - 92	2	0.125	12.5%
Total	16	1	100%

(b) Construct a CF table for these data.

Solution:

Weight limit	**CF**
≤ 56	3
≤ 68	8
≤ 80	14
≤ 92	16

(c) How many student weighs **at most** 68 kg? 3+5=8

(d) What is the percentage of students who weigh **at least** 57 kg? $31.25 + 37.5 + 12.5 = 81.25$ %

(e) What is the probability that a student weight is between 57 and 80? $0.3125 + 0.375 = 0.6875$

(f) What is the number of students who weigh more than 80 kg? 2

Graphs & Charts

- We usually use the frequency distributions to construct graphs and charts for the data.
- Types of graphs and charts:

⋆ For qualitative data (variables):

(a) **Bar graph**: Classes (x-axis) vs. F (y-axis).

(b) **Pie chart**: Classes vs. PF.

⋆ For quantitative data (variables):

(a) **Histogram**: ~~Classes~~ Class Boundaries (x-axis) vs. F (y-axis).

First: **Qualitative data.**

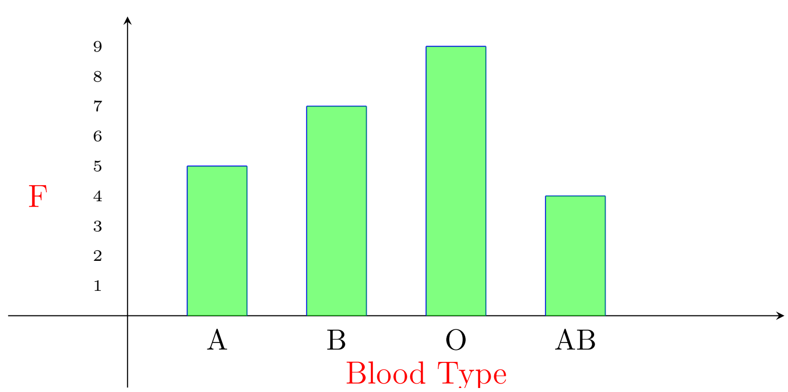
(a) Bar graph. (b) Pie chart.

Question: Construct a bar graph and a pie chart for the blood type. (previous example)

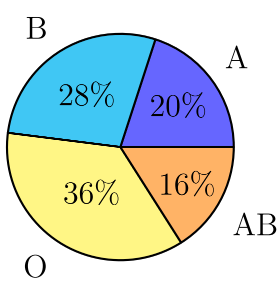
Solution: For bar graph we need F table, and for pie chart we need PF table.

Blood type	F	PF
A	5	20%
B	7	28%
O	9	36%
AB	4	16%
Total	25	100%

Bar graph:



Pie chart:



Second: **Quantitative data.**

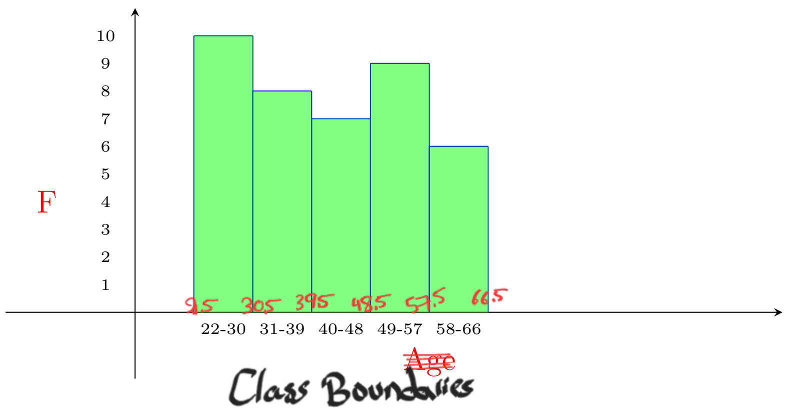
(a) Histogram.

Question: Construct a histogram for the ages. (previous example)

Solution: We need the F table.

Age	F
22 - 30	10
31 - 39	8
40 - 48	7
49 - 57	9
58 - 66	6
Total	40

Histogram:



Crosstabulation

- **Crosstabulation**: a tabular summary of data for two variables.

Ex: The **Type** and **Number of employees** of some firms in Palestine are given below.

# of employees	Governmental	Private	**Total**
10 – 19	4	2	6
20 – 29	8	1	9
30 – 39	9	3	12
40 – 49	5	10	15
50 – 59	6	2	8
Total	32	18	50

- **Answer the following questions.**

(1) What is the number of elements? 50

(2) What is the number of observations? 50

(2) What are the variables? Type of firm and number of employees

(3) What is the type of our study? (census or survey): survey

(4) The # of employees is discrete or continuous? discrete

(5) The midpoint of the class 50 - 59 is: 54.5

(6) The number of governmental schools is : 18

(7) The proportion of schools with number of employees between 20 and 29 is: 0.18

(8) The percentage of private schools with employees between 40 and 49 is: 20 %

(9) The number of schools with employees less than or equal 39 is: 27

(10) Among the private schools, the percentage of schools with employees between 50 and 59 is: 11.11%

Scatter diagram and Trendline

- We will analyze the **simple linear relationships** between two quantitative variables x & y.

Ex: (1) x: # of absences. y: Final grade.

(2) x: Amount of advertising. y: Volume of sales.

(3) x: Age. y: Blood pressure.

⋆ Types of relationships:

(1) **Positive relationship**: When x increases, y increases.

(2) **Negative relationship**: When x increases, y decreases.

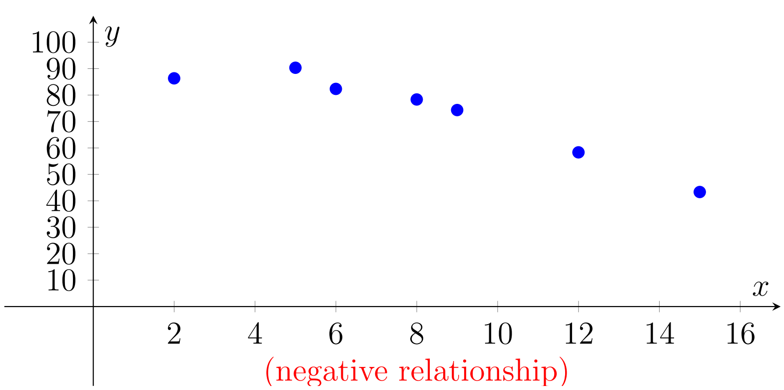
(3) **No relationship**: x and y are not related.

- Scatter diagram: A graph used to determine the type of relationship between the variables x & y by plotting the ordered pairs (x_i, y_i)

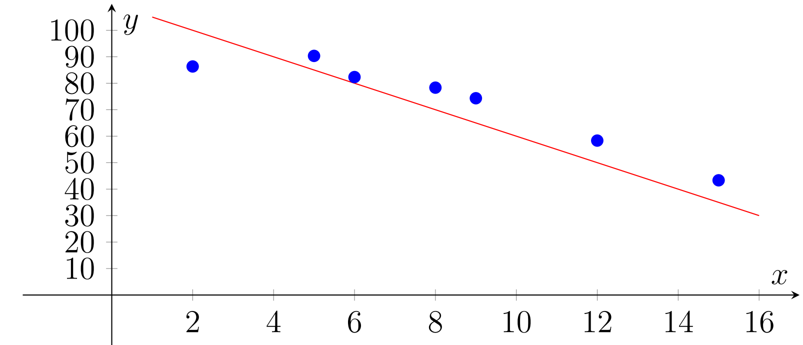
- Trendline: a line that provides an approximation of the relationship between x and y

Ex: The following represents the # **of absences** x and **final grades** y of a sample of of 7 students accompanied with the scatter diagram.

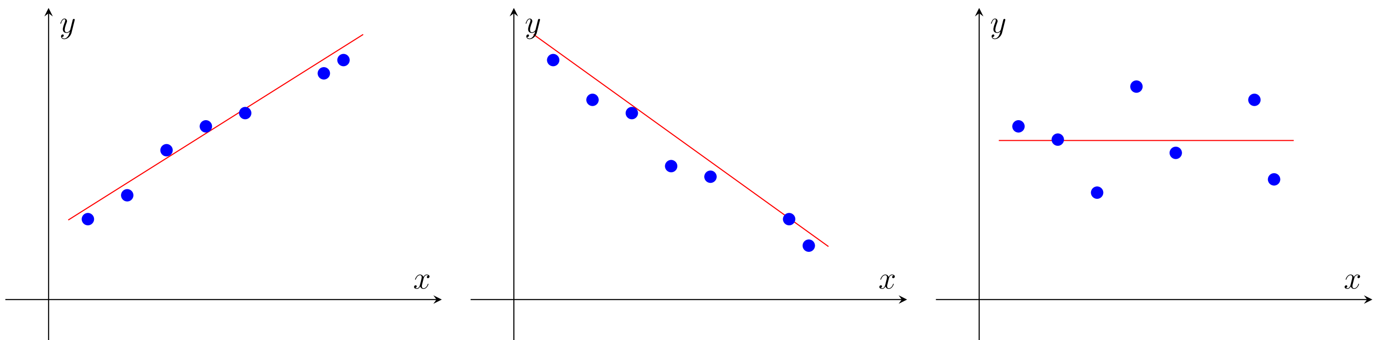
x	y
6	82
2	86
15	43
9	74
12	58
5	90
8	78



(negative relationship)



(negative relationship)



(positive relationship) (negative relationship) (no relationship)