CHAPTER : 1

papergrid

Date: / /

Ch : DESCRIPTIVE STATISTICS

GOAL: Graphical and numerical representation of data

TOPICS : 1) Basic Concepts of Data Acquisition
2) Analysis of Univariate Data
3) Analysis of Bivariate Data

* BASIC CONCEPTS OF DATA ACQUISITION

→ Statistical unit or entity : A single "object" is called a stat. unit.
eg : Robotics student XYZ in course statistics in WS2023.

→ Population : set of all objects under consideration.
eg : entire class of robotics students in statistics in WS2023.

→ Variables : properties of a statistical unit one is interested in.
a.k.a attributes / characteristics

→ Occurence : specific value a variable takes on.

Example :

	Matr. no.	Citizenship	Civil Status	Age (variable)	
statistical unit ←	3313019	UK	Single	21	Population
	⋮			⋮	
	3312008	India	single	18	
				↳ Occurences	

→ Key Attributes : An attribute or minimal combination of attributes that identifies a statistical unit uniquely.

→ Possible Occurences :

→ Scale of Variables : • Categorical / nominal scale : only labels
↳ language
↳ Citizenship
• Ordinal scale : sorting possible
↳ Army rank
↳ Placement in a sporting event.
• Cardinal / metric scale : Calculation possible
↳ Stock & commodity prices
↳ Age, weight, length

NOTE: Categorical scale can be multi
but not the other 2 sc

papergrid
Date: / /

→ Multivalued Variables : sometimes a variable can take 2/more values at the same time.

eg: A student
↳ more than 1 citizenship
↳ speaks more than 1 language

Q Exercise

Nominally scaled	Ordinally Scaled	Cardinally scaled
• Gender ↳ m/w/d	• Job → team member	• Income → 0€, 150K€, 30
• Job → programmer	junior chief	• Age → 2, 15, 32 (y
firefighter	senior chief	• Temperature (in K): 0
professor	• Academic degree	• Grades
• First name → Rob,	→ B Eng, M Eng, PHD	
Tia, Ria	• Knight, Duke, Count	
	↳ Title	

* ANALYSIS OF UNIVARIATE DATA

- raw data – all occurences written down as a list with no specific order.
multiset / bag

- sorted lists – raw data sorted w.r.t specific criterion (eg: in ascending order)
atleast ordinal scale data required.
every occurence is written down only once. plus we write frequencies.

- **Univariate frequency** : $h(a_j)$ of the occurence a_j appears in raw data.
- **relative frequency** : $f(a_j) = h(a_j)/n$ of occurence a_j tells us fraction with which a_j occurs in raw data.

raw data → sorted list

eg : 23, 1, 1, 5, 3, 6, 3, 3, 1 → 1, 1, 1, 2, 3, 3, 3, 3, 5, 6

a_j	1	2	3	4	5	6	Σ
$h(a_j)$	3	1	4	0	1	1	10
$f(a_j)$	0.3	0.1	0.4	0	0.1	0.1	1

$h_j = h(a_j)$ = absolute frequencies

$f_j = f(a_j) = \frac{h(a_j)}{n}$

- **Cumulative frequencies**

n : number of univariate statistical units $(x_1, \ldots, x_n)$ and
k : number of different occurences $(a_1, \ldots, a_k)$

- **Absolute cumulative frequency**

$H_j = h_1 + h_2 + \ldots + h_j$

- **relative cumulative frequency**

$F_j = f_1 + f_2 + \ldots + f_j = H_j/n$

Generally $H_k = n$

$F_k = 1$

- **Cumulative frequency Distribution**

$$H(x) = \begin{cases} 0 & \text{if } x < a_1 \\ H_j & \text{if } a_j \le x < a_{j+1} \\ n & \text{if } x \ge a_k \end{cases}$$

$$F(x) = \frac{H(x)}{n} = \begin{cases} 0 & \text{if } x < a_1 \\ F_j & \text{if } a_j \le x < a_{j+1} \\ 1 & \text{if } x \ge a_k \end{cases}$$

Date: / /

Exercise: total no. of objects produced in one hour:

No. of objects:	3	4	5	6	7	8	9	10
Abs. frequency:	10	15	30	30	25	20	15	5

$F(3)$, $F(5.5)$, $F(10)$ = ?

How many hours did data acquisition take at least?

Statistical unit: production hours (1 hour)

Variable: number of objects

n = sum of all abs. frequencies = 150

$a_1 = 3$, $a_2 = 4$, ..., $a_8 = 10$

$$F(n) = \begin{cases} 0 & \text{if } n < a_1 \\ H_j/n & \text{if } a_j \le n < a_{j+1} \\ 1 & \text{if } n \ge a_k \end{cases}$$

$x = a_1$

$$\rightarrow F(3) = \frac{H_1}{150} = \frac{10}{150} = \frac{1}{15} = 0.067$$

$$\rightarrow F(5.5) = \frac{H_3}{150} = \frac{55}{150} = \frac{11}{30} = 0.167$$

$$\rightarrow F(10) = 1$$

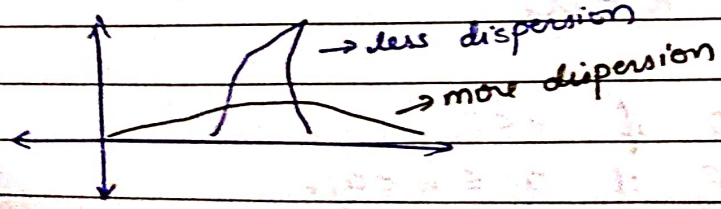
- **Numerical Representation of Univariate Data:**

univariate data can be numerically represented by

→ measures of central tendency

→ measures of dispersion

→ measures of concentration



- **Measures of central tendency**
  ↳ describes where data is primarily located
  ↳ whole data sample summed to a single value
→ imp measures of C.T
  - mode
  - median
  - mean

- **MODE**
→ $\bar{x}_{mod}$ is the most frequent value.
→ Adv: mode can already be used with nominal data
→ Disadv.: only makes sense if value is unique.

eg: data sample: (2, 3, 1, 1, 5, 3, 6, 3, 3, 1)

a_j	1	2	3	4	5	6
h_j	3	1	4	0	1	1

$\bar{x}_{mod} = 3$

- **MEDIAN**
→ $\bar{x}_z$ separates lower half from upper half.
→ median appears as a value in the data list. (unlike mean)
→ if data sample is sorted in asc. order, median $\bar{x}_z$ is defined by

$$\bar{x}_z = \begin{cases} x_{\frac{n+1}{2}} & \text{if } n \text{ odd} \\ x_{\frac{n}{2}} & \text{if } n \text{ even} \end{cases}$$

Alternate: $$\bar{x}_z = \begin{cases} x_{\frac{n+1}{2}} & \text{if } n \text{ odd} \\ x_{\frac{n}{2}+1} & \text{if } n \text{ even} \end{cases}$$

→ Adv.: median can be used with ordinal data
Dis.: not unique for even n.

Date: / /

- **p Quantile**

→ in case of ordinal data, concept of median can be generalized to p-quant

→ at most p·n of the values of in data sample are smaller and at most (1-p)·n of values are larger than value of the p quantile.

→ p = 0.25 first quartile Q1
p = 0.5 second quartile Q2 = median
p = 0.75 third quartile Q3

Exercise : Determine the quartiles Q1, Q2, Q3 of the following data samples :

a) (2, 7, 8, 11, 13, 17)
Q1 Q2 Q3

at most 25% values smaller than Q1
at most 75% values larger than Q1
∴ 2 ≠ Q1
Q1 = 7
similar logic for Q3

b) (1, 5, 66, 234, 440, 489, 500)
Q1 Q2 Q3

c) (1, 3, 5, 66, 111, 234, 440, 489, 500, 777)
Q1 Q2 Q3

- **MEAN**

↳ computation of mean only possible for cardinal data.

→ Arithmetic mean $\quad \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

→ Geometric mean $\quad \bar{x}_G = (x_1 \cdot x_2 \cdot \ldots \cdot x_n)^{\frac{1}{n}}$ if n = even product $\neq$ -ve

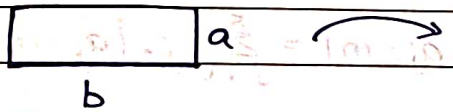
→ Harmonic $\quad \bar{x}_H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$

⇒ **ARITHMETIC MEAN** : $\boxed{\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i}$

→ for given data, arithmetic mean solves minimization problem.

→ in contrast to mode / median, arith. mean does not have to occur in data sample.

⇒ **GEOMETRIC MEAN** :



(area)
$a \times b = s^2$
$\Rightarrow s = \sqrt{ab}$
↳ geometric mean

(can also do this for cubes etc)

→ $\boxed{\bar{x}_G = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}}$

→ only defined for +ve values.

$\boxed{\bar{x} \geq \bar{x}_G}$ ⇒ AM vs. GM

⇒ **HARMONIC MEAN** : $\boxed{x_H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}}$

nominal data → mode
ordinal data → median
cardinal data → mean

$\boxed{\min \leq x_H \leq x_G \leq \bar{x} \leq \max}$

- **Measures of Dispersion**

→ describe how data is dispersed around a central tendency.

→ most important dispersion parameters are:

- the range
- mean deviation
- empirical variance and standard deviation

- **Range**

→ a.k.a width = largest value − smallest value

$a_k - a_1$

- **Mean Deviation**

→ a.k.a mean absolute deviation (mad)

$$m.a.d = \frac{1}{n} \sum_{i=1}^{n} |x_i - m| \qquad m = \text{arithmetic mean } \bar{x} \text{ med./mode}$$

$$m.a.d = \frac{1}{n} \sum_{i=1}^{n} |x_i - m| = \frac{1}{n} \sum_{j=1}^{k} h_j \cdot |a_j - m| = \sum_{j=1}^{k} f_j \cdot |a_j - m|$$

- **Empirical variance (s^2) and standard deviation (s)**

→ $$s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n} \left(\sum_{i=1}^{n} x_i^2 \right) - \bar{x}^2$$

→ $$s = \sqrt{s^2}$$

→ $$s^2 = \frac{1}{n} \sum_{i=1}^{n} (m_i - \bar{x})^2 = \frac{1}{n} \sum_{j=1}^{k} h_j \cdot (a_j - \bar{x})^2 = \sum_{j=1}^{k} f_j \cdot (a_j - \bar{x})^2$$

- Relative dispersion − generally defined as a ratio

→ $\frac{\text{absolute dispersion}}{\text{mean}}$

→ coefficient of variation $V = \frac{s}{\bar{x}}$ → std. dev. / → mean

Exercise

	a_j	1	2	3	4	5	6
$\Sigma = 10$	h_j	1	2	3	2	1	1
$\Sigma = 1$	f_j	0.1	0.2	0.3	0.2	0.1	0.1

a) mode, median, ~~mean~~, GM, HM, AM.
b) range, m.a.d, variance, std dev, coeff of var.

a) mode = 3 median = $\frac{3+3}{2} = 3$ (sorted list : 1, 2, 2, 3, 3, 3, 4, 4, 5, 6)

arithmetic mean = $\frac{\Sigma n_i}{n} = \frac{33}{10} = 3.3$

Geometric mean = $(1.2.2.3.3.3.4.4.5.6)^{1/10} = 2.96$

HM : $\frac{10}{\frac{1}{1}+\frac{1}{2}+\frac{1}{2}+\dots+\frac{1}{6}}$

$= 2.59$

b) Range = $6-1 = 5$ mean abs. dev. = $\frac{\Sigma|n_i - m|}{n} = \frac{|3.3-1|+|3.3-2|+..|33-6|}{10}$

$= \frac{2.3+1.3+1.3+(0.3)3+(0.7)2+..}{10}$

$= \frac{11.6}{10} = 1.16$

$S^2 = 2.01$ $S = 1.42$

$V = \frac{1.42}{3.33} = 0.43$

- **Merging Aggregates**

→ for each group G_ℓ, we know the size of the group n_ℓ, the arithmetic mean $\bar{n}_\ell$ and empirical variance s^2_ℓ.

→ overall result on merging the groups :

$$n = \sum_{\ell=1}^{m} n_j$$

$$\bar{n} = \sum_{\ell=1}^{m} \frac{n_\ell}{n} \cdot \bar{n}_\ell$$

$$s^2 = \sum_{\ell=1}^{m} \frac{n_\ell}{n} \cdot s^2_\ell + \sum_{\ell=1}^{m} \frac{n_\ell}{n} \cdot (\bar{n}_\ell - \bar{n})^2$$

Exercise Consider a statistical survey where the results of G_1, G_2, G_3 are :

$G_1 : n_1 = 10 \quad \bar{n}_1 = 100 \quad s_1^2 = 400$

$G_2 : n_2 = 5 \quad \bar{n}_2 = 120 \quad s_2^2 = 144$

$G_3 : n_3 = 20 \quad \bar{n}_3 = 60 \quad s_3^2 = 100$

Calculate overall $\bar{n}$ and s^2.

$n = 10 + 5 + 20 = 35$

$$\bar{n} = \frac{10 \times 100}{35} + \frac{5 \times 120}{35} + \frac{20 \times 60}{35} = 80$$

$$s^2 = \frac{10 \times 400}{35} + \frac{5 \times 144}{35} + \frac{20 \times 100}{35} + \frac{10 \times 20^2}{35} + \frac{5 \times 40^2}{35} + \frac{20 \times 20^2}{35}$$

$= 763.43$

$S = 27.63$

* ANALYSIS OF BIVARIATE DATA

→ Bivariate or 2-D data refers to a collection of pairs $((x_1,y_1),\ldots,(x_n,y_n))$ where the values of x and y are recorded simultaneously.

→ Imp concepts : correlation and regression

- **Scatter plot** : data is displayed as a collection of points. it is possible to visually identify correlations b/w the variables.

- **Correlation** :

→ correlation describes a relationship b/w two variables which are of at least ordinal scale.
eg : taller people are usually heavier.

→ correlation is just a statistical observation. NOT a cause-symptom relation

- **Spearman's Rank Correlation Coefficient**

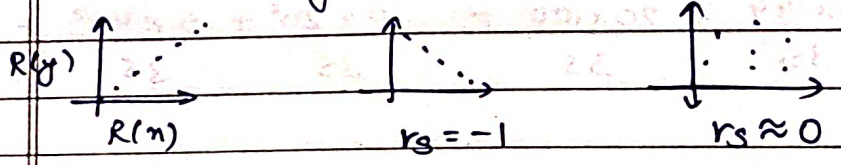
→ Consider bivariate data $(x_1,y_1)\ldots(x_n,y_n)$ where x_i, y_i are occurences of ordinal scale.

→ every x_i recieves a rank $R(x_i)$ and every y_i recieves a rank $R(y_i)$ w.r.t their position.

POSSIBLE CASES

1. The ranks x_i and y_i run in the same direction
   ↳ POSITIVE CORRELATION
2. The ranks x_i and y_i run in opposite direction
   ↳ NEGATIVE CORRELATION
3. There is no relation b/w ranks of x_i and y_i
   ↳ uncorrelated variables

Spearman's rank correlation coeff : r_s measures strength & direction of correlation.



$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^{n} (R(x_i) - R(y_i))^2}{(n-1) \cdot n \cdot (n+1)}$$

$-1 \leq r_s \leq 1$.

$r_s = 1$: ranks run in same direction
$r_s = -1$: ranks run in opp. direction
$r_s = 0$: ranks run without relation

Q Calculate Spearman's rank correlation coeff for the variables "Age" and "Points in test".

Student No.	1	2	3	4	5	6	7	8	9	10	11
Age (years)	38 (4)	47 (2)	44 (3)	51 (1)	35 (5)	29 (6)	22 (7)	14 (9)	12 (10)	19 (8)	9 (11)
Points in test	39 (3)	34 (4)	31 (5)	48 (1)	46 (2)	23 (7)	17 (8)	12 (10)	16 (9)	28 (6)	10 (11)

$$r_s = 1 - 6 \cdot \frac{\Sigma (R(x_i) - R(y_i))^2}{(n-1)(n)(n+1)}$$

$$\Sigma (R(x_i) - (Ry_i))^2 = 1^2 + 2^2 + 2^2 + 0 + 3^2 + 1^2 + 1^2 + 1^2 + 1^2 + 2^2 + 0$$
$$= 1 + 4 + 4 + 9 + 1 + 1 + 1 + 1 + 4 = 26$$

$$\therefore r_s = 1 - \frac{6^1 \times 26^{13}}{(10)(11)(12)_2} = 1 - 0.118 = 0.881$$

most imp

- **Coefficient of Correlation of Bravais Pearson**: takes into account distance of data points and not only ranking
  - → for cardinal data
  - → it is a measure for the linear relationship b/w 2 cardinal variables.
  - → It measures how well the bivariate data can be approximated by a so called linear trend or regression line.

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} \qquad -1 \leq r \leq 1$$

$r = 1$: data exactly located on a line with +ve slope
$r = -1$: data exactly located on a line with −ve slope
$r = 0$: data shows no linear relationship at all.

Q Determine the coeff. of correlation for the variables "Inflation" and "Jobless rate":

Year	2001	2002	2003	2004	2005
Inflation (%)	2	3	3	2	5
Jobless rate (%)	4	7	2	3	4
$x_i - \bar{x}$	-1	0	0	-1	2
$y_i - \bar{y}$	0	3	-2	-1	0

$$\bar{x} = \frac{2+3+3+2+5}{5} = 3$$

$$\bar{y} = \frac{4+7+2+3+4}{5} = 4$$

$$r = \frac{(-1)(0) + (0)(3) + (0)(-2) + (-1)(-1) + (2)(0)}{\sqrt{1+0+0+1+4} \cdot \sqrt{0+9+4+1+0}}$$

$$= \frac{1}{\sqrt{6} \cdot \sqrt{14}} = \frac{1}{9.165} \approx 0.109 \rightarrow \text{no linear correlation}$$

- **Empirical covariance** (cov)

→ is an auxiliary measure for the joint variability of two cardinally scaled variables.

(for bivariate data) →

$$C.O.V(x,y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

→ coeff. of covariance $r = \frac{cov(x,y)}{S_x \cdot S_y}$ ↳ standard deviations

relationship b/w coeff. of Spearman & Bravais Pearson

$$r_S = \frac{\sum_{i=1}^{n} (R(x_i) - \overline{R(x)}) \cdot (R(y_i) - \overline{R(y)})}{\sqrt{\sum_{i=1}^{n} (R(x_i) - \overline{R(x)})^2} \cdot \sqrt{\sum_{i=1}^{n} (R(y_i) - \overline{R(y)})^2}}$$

if ranks are unique $\overline{R(x)} = \overline{R(y)} = \frac{n+1}{2}$

y: x_i	$R(x_i)$	
1000	1	$\overline{R(x)} = 3$
800	2	
100	5	
50	4	
10	5	

- **Regression Analysis:**

→ in contrast to correlation analysis, regression analysis assigns two different roles for x and y.

→ x: cause y: symptom $y = f(x)$ → functional relationship; x → independent; y → dependent symptom

- **Computing a linear trend line:**

→ searching for linear function of the type:

$$y = f(x) = ax + b$$

with minimal distances to the data $(x_1, y_1), \ldots (x_n, y_n)$

a, b: coefficients of regression

→ to find a, b: OPTIMIZATION PROBLEM:

$$Q(a,b) = \sum_{i=1}^{n} \underbrace{(y_i - f(x_i))}_{\text{residual}}^2 = \sum_{i=1}^{n} (y_i - ax_i - b)^2 \rightarrow \text{minimize}$$

$$= \sum_{i=1}^{n} (y_i - ax_i - b)^2 \rightarrow \text{minimize}$$

$$\frac{\partial Q}{\partial a} = \sum_{i=1}^{n} -2x_i (y - ax_i - b) = 0$$

$$\frac{\partial Q}{\partial b} = \sum_{i=1}^{n} -2(y_i - ax_i - b) = 0$$

Computing, we get (slope) $$a = \frac{\sum_{i=1}^{n} x_i \cdot y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right) \cdot \left(\sum_{i=1}^{n} y_i\right)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)}$$

(y intercept) $$b = \frac{1}{n}\left(\sum_{i=1}^{n} y_i - a \sum_{i=1}^{n} x_i\right)$$

→ Alternative expressions

$$a = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{cov(x,y)}{s_x^2} = \frac{r s_y}{s_x}$$

$$b = \bar{y} - a\bar{x}$$

Q A distribution center runs ten shrink wrap machines on different velocity x (m/min) the variable y indicates the number of stops caused by cracks. Compute a b

Machine	1	2	3	4	5	6	7	8	9	10	$\bar{x} = 26$
x_i	21	22	23	24	25	26	27	28	30	34	$\bar{y} = 50$
y_i	30	30	30	40	50	50	60	60	70	80	
$x_i - \bar{x}$	-5	-4	-3	-2	-1	0	1	2	4	8	
$y_i - \bar{y}$	-20	-20	-20	-10	0	0	10	20	20	30	
$(x_i-\bar{x})(y_i-\bar{y})$	100	80	60	20	0	0	10	20	80	240	
$(x_i-\bar{x})^2$	25	16	9	4	1	0	1	4	16	64	

$$a = \frac{610}{140} = 4.36$$

$$b = \bar{y} - a\bar{x} = 50 - (4.36 \times 26) = -63.36$$

- **Nonlinear Regression**

→ linear regression is not always suitable for relation b/w x and y.

→ In case of monotonically increasing / decreasing regression functions, suitable substitution can reduce the problem to linear regression.

eg:

x_i	1	2	3	4
y_i	8	18	30	51

→ To compute a, b

$\tilde{x}$	1	4	9	16

→ Regression function is of the type $y = f(x) = ax^2 + b$

$\Rightarrow y = f(\tilde{x}) = a\tilde{x} + b$ where $\tilde{x} = x^2$

$$a = \frac{\sum (\tilde{x}_i - \bar{\tilde{x}}) \cdot (y_i - \bar{y})}{\sum (\tilde{x}_i - \bar{\tilde{x}})^2} = 2.82 \qquad b = \bar{y} - a\bar{\tilde{x}} = 5.6$$

$\therefore y = f(x) = 2.82x^2 + 5.6$

- **Quality of Regression Analysis** : we compare empirical values y_i against the corresponding values $f(x_i)$ of the regression function

→ estimates $\hat{y}_i = f(x_i)$

residual $\hat{u}_i = y_i - \hat{y}_i$

→ coefficient of determination $R^2 = 1 - \frac{\sum_{i=1}^{n} \hat{u}_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$

↓ measure of how much reality (empirical data) meets a (statistical model)

$0 \leq R^2 \leq 1$

if $R^2 = 1$: Perfect regression function, all residuals are 0.

$R^2 < 0.1$: Obtained regression function is nonsense

→ Alternate Formula : to find R^2

In case of a linear regression function $y = f(x) = ax + b$

$$R^2 = \frac{s_{\hat{y}}^2}{s_y^2} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = a^2 \frac{s_x^2}{s_y^2} = r^2$$

↳ coeff. of correlation of Bravais Pearson

- **Time Series Analysis**

→ time series is a time ordered collection of sequentially observed data pairs.

$(t_i, x_i)_{i=1}^{n} = ((t_1, x_1), \ldots, (t_n, x_n))$

where t represents points in time $t_1 < t_2 < \ldots < t_n$

x : cardinally scaled

→ time series show temporal patterns → seasonal cycles
→ long term trends
→ random fluctuations

→ time dependent variable $\boxed{x = T + C + S + R}$

T: trend : long term tendency

C: cycle : medium term periodic component

S: season : short term periodic component

R: random component : irregular fluctuations

Date: / /

→ Trend : T - long term tendency of a time series, often in terms of a straight line (regression analysis)

→ Component Cycle : C - medium term oscillations. cycle C is a periodic component with cycle length or period larger than the length of S. Standardized C shows oscillations without trend T.

→ Component Season : S - short term oscillations with typical pattern and well known wavelengths. typical season period is one year.

→ ~~Time~~ Random Component : R - irregular & unpredictable effects. $R = x - (T+C+S)$. Standardized R fluctuates around 0 line.

⇒ COURSE OF ACTION : Time Series

1) Determine T by regression analysis
2) Determine smooth component $G = T+C$ by filtering time series based on moving averages.
3) $C = G - T$
4) Find S by averaging over difference $x - G = S + R$
5) $R = x - T - C - S$

Q Determine the moving averages x_j^* of the following time series according to a time window of length 5:

t_i	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
x_i	3.00	4.07	5.13	6.20	7.27	3.33	4.40	5.47	6.53	7.60	3.67	4.73	5.80	6.87	7.93
x_i^*	X	X	5.13	5.2	5.27	5.33	5.4	5.47	5.53	5.6	5.67	5.73	5.8	5.87	X

Case : In case of an ~~even~~ even time window eg : $6 = 2k$

$\Rightarrow k = 3$

0	1	2	3	4	5	6	7
3	4.07	5.13	6.20	7.27	3.33	4.40	5.47
X	X	X	4.95				

$$\frac{1}{6}\left(\frac{3}{2} + \frac{4.4}{2} + 4.07 + 5.13 + 6.20 + 7.27 + 3.33\right)$$

• Seasonal Adjustment by Moving averages

→ $x^* \approx G = T + C$

$\therefore x - x^* \approx S + R$

→ Summary :

1) Filter x in terms of x^*
2) Remove smooth component $x^* \approx G = T + C$ from time series to get $S + R$
3) Average $x - x^*$ to obtain seasonal averages
4) Standardize seasonal averages so that they oscillate around the zero line to obtain seasonal pattern S.
5) determine seasonally adjusted time series $x - S$ by removing S from x.

$$x = T + C + S + R$$

$G = T + C$

Q **Exercise :** In a tide station, the water level is measured on four consecutive days at 0:00 am, 8:00am and 4:00 pm respectively. The results of the measurements in cm are as follows –

Day 1	Day 2	Day 3	Day 4
603 | 723 | 480	606 | 720 | 420	600 | 660 | 420	537 | 660 | 600

Apply appropriate moving averages to ~~obt~~ obtain the seasonal pattern and the seasonally adjusted time series.

	0:00	8:00	4:00	0:00	8:00	4:00	0:00	8:00	4:00	0:00	8:00	4:00
t_i	1	2	3	4	5	6	7	8	9	10	11	12
x_i	603	723	480	606	720	420	600	660	420	537	660	600
$G = x_i^*$	X	602	603	602	582	580	560	560	539	538	599	X
$x_i - x_i^*$	X	(121)	−123	4	(138)	−160	40	(100)	−119	−2	(61)	X = S+R
$x_i - S$	584	613	609	587	610	549	581	550	549	518	550	729

($x_i - S$: seasonally adjusted time series)

0:00 am	14 ($\rightarrow \frac{4+40-2}{3}$)	+ 5	19
8:00 am	105 ($\rightarrow \frac{121+138^*}{100+61}{4}$)	+ 5	110
4:00 pm	−134	+ 5	−129
	−15		

$= \begin{pmatrix} S_1 \\ S_2 \\ S_3 \end{pmatrix} = S$ ↓ seasonal pattern

Date: / /

- **Time Series Forecasting by Extrapolation.**

1) Acquire time series data $((t_i, x_i))_{i=1}^{n}$
2) Check preconditions
   - one full cycle is contained in time series dat
   - atleast a couple of seasons
   - seasonal wavelength is well known
3) Determine T, C, S
4) Forecast the time series by forwarding T and repeating periodic components C and S.

$$x = T + C + S$$

eg (Pg 118, STSE_01)

- **Alternative methods of filtering**

- **Filtering by smoothing**

→ a.k.a exponential weighted moving average

→ smoothing factor α where $0 \leq \alpha \leq 1$

larger values of α reduce extent of smoothing

→ $x_1^* = x_1$ $\quad x_{j+1}^* = \alpha \cdot x_j + (1-\alpha) \cdot x_j^*$

→ smoothed once – one level smoothing
smoothed again – 2 level smoothing
smoothed n times – n level smoothing

for forecast: $x_1^* = x_1$ $\quad x_{j+1}^* = \alpha x_j + (1-\alpha) x_j^*$

Q Quaterly production volumes of a garment factory are available for three consecutive years 2021, 2022 and 2023.

2021	2022	2023
10 | 12 | 8 | 14	14 | 16 | 12 | 18	18 | 20 | 16 | 22

Apply exponential smoothing to the time series with $\alpha = 0.2$ in order to give a forecast on the production volume of the first quarter in 2024.

t_i	1	2	3	4	5	6	7	8	9	10	11	12	13
x_i	10	12	8	14	14	16	12	18	18	20	16	22	?
x_i^*	10	10	10.4	9.92	10.73	11.38	12.3	12.24	13.39	14.31	15.44	15.55	16.84

Using $x_1^* = x_1$

and $x_{j+1}^* = \alpha x_j + (1-\alpha) x_j^*$ } forecast