ملخص STAT 101

هذا الملخص لا يغني
عن الكتاب او المحاضرات
هو فقط ترتيب وربط
للافكار لتسهيل الفهم

Rayan

CHAPTER 1

Descriptive statistics

1.0 - 1.1 - 1.2 - 1.3 - 1.4 - 1.5

Data :- a collection of information collected by means of experiments observations, or real life events and stored in a proper format (data derived from data)

Statistics :- a branch of science deals with collection, organization, presentation analysis, interpretation of data and take the appropriate divisions. 1.0

Types of statistics 1.1

Descriptive

consists of methods and techniques which are used for presenting and summerizing data in tables or graph forms and provide some numerical measures for it.

* Examples:-
- The study of some human characteristics includes (average height or weight) based on data taken from a group of people is a descriptive study.

*forms of tables:-
- frequency table
- frequency distribution table

*forms of graphs:-
- Pie chart
- Bar chart
- Histogram
- Polygon

* forms of numerical measures:
- averages
- Percintiles
- despersions
- and other measures...

descriptive	inferential
- does not make conclusions - organizes data and describes them	- does make conclusions - Takes the data we organized and draws conclusions

Inferential

some methods and techniques that can be used to make conclusions about the entire population using the observations from the samples taken from the population.

* methods based on probality and distribution theory to make inferences:-
- point estimation
- interval estimation
- hypothesis testing
- statistacal modeling
- clustering
- and many other methods

Population :-

set of all things - have at least one common feature - that will be subjected to a study to obtain infrences for a specific problem.

* the number of elements is called the size (N) (M)
* The elements of a population is called an individualls

Examples:-
- set of all student in a country
- Set of all palm trees in Ksa

Samples :-

a subseat of population which is used to collect information and to make infrences about the population

* the number of elements is called the size and denoted by (n) or (m)

Examples:-
- the set of selecting Qassim terees from Ksa Palm trees.

* since using the whole population to make inferences is hard, we use a subset of it (sample)

Relationship between sample and population

Parameter	statistic
- certain quantity or quality for describing a characteristic in a given population that summerizes the data for the entire population - in a statistical study the mesure is unknown	- certain quantity or quality for describing a characteristic in a given sample that summerizes the data for the entire sample - in a statistical study the mesure is known
one uses a statistic to make inferences about a parameter	

Variable :-

a map or a function X defined on the population or sample and takes value in an arbitrary set M

X : population or sample ⟶ M

* the variable measures a characteristic, feature, or factor in the population

* not all individuals have the same exact all characteristics but they may share some of them

Types of variables

according to type

Qualitative

variables that takes non numerical values or numerical values (can't perform mathmatical operations on the values) which indicate an attribute or property.

Examples:-
- married status
- eye color
- ID number of ksu students

Quanitative

variables that takes numerical values and we can perform mathmatical operations on them

* can be ordered increasing or decreasing.

Examples:-
- Height
- weight
- tempreture
- revenue

according to numbers

Discrete

variables takes finite or infinite countable numbers

Examples:-
- Quan Discrete
  * number of accidents in a city
  * number of laptops sold a day
- Qual Discrete
  * blood groups of people
  * types of cars

continuous

variable takes uncountable number of values

Examples:-
- Quan continuous
  * weight of person
  * distance between cities
- Qual continuous
  * all spectrum colors resulting from white light

Raw data :- data that is not organized well.

(ungrouped data)

1.2

Examples:-

1.2.1 Qualitative

June	July	January	December
march	march	April	September
August	May	May	February
July	February	June	June
April	February	november	August
July	April	June	January
december			

the birth of 25 members of a community

1.2.2 Discrete Quantitative

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

number of children in 40 family in a society

1.2.3 continuous Quantitative

170	180	175	176	172
173	183	171	169	174
180	193	186	189	192
167	175	170	178	191
165	177	183	181	179

the height of 25 adult man in cm

Types of Organization of raw data

1- Frequency Table :- (Qual or discrete Quan)

class	Frequency	Relative frequency	Percent Frequency
all the classes of the variable are mentioned	The number of times the data occures	The Relative Frequency (R.F) = The frequency of a class / The sum of frequencies	Percent frequency = R.F x 100%
sum	$\sum$ frequency	$\sum$ Relative frequency = 1	$\sum$ percent frequency = 100%

Examples :-

- look at example 1.2.4 page 12 for qualitative data frequency ~~distribution~~ table.
- look at example 1.2.6 page 13 for quantitative discrete frequency table.

2- frequency distribution table

(continuous Quan Data)

* unwated to use frequency dist table for data less than 32 ($n \geq 32$)

1- Range : $x_L - x_S$

2- K 5-20 classes , if not provided : $K = [3.322 \log n]$

$K = [x]$ = integer number $\leq x$

3- class boundarie length $C = \frac{R + \text{one measuring unit}}{K}$

4- class limit length = C - 1

5- Lower Limit of class boundarie = -0.5 from Lower Limit of class limit | Higher Limit of class boun = +0.5 from Higher limi of class lim

6- class midpoint = $\frac{\text{upper limit} + \text{lower limit}}{2}$

7- find the frequency

8- find the Relative frequency

9- find ACF : 1- Start with 1st frequency 2- add 1st and 2nd Frequency 3- Add 3rd frequency to the previous ...

10- find DCF : 1- Start with $\sum f_i$ 2- subtract 1st frequency 3- subtract 2nd frequency from previous DCF ...

Example :- look at data page 16

class limit	class boundarie	Midpoint	Frequency	RF	ACF	DCF
1 - 5	0.5 → 5.5	3	7	0.14	7	50
6 - 10	5.5 → 10.5	8	9	0.18	7+9 = 16	50 - 7 = 43
11 - 15	10.5 → 15.5	13	14	0.28	7+9+14 = 30	50-7-9=34
16 - 20	15.5 → 20.5	18	12	0.24	7+9+14+12 = 42	50-7-9-14 = 20
21 - 25	20.5 → 25.5	23	8	0.16	7+9+14+12+8 = 50	50-7-9-14-12 = 8
total	-	-	50	1	-	-

Types of graphs

1-3

1- Pie chart :-

a simple way of representing the portions of each class or category of data on a circular disk so that each categorie is allocated a circular sector representing it.

the measure of angle = $RF_{(i)} \times 360$

Angles drawn are not accurate!

Example:-

data from 1.2.6 in the book

category (AB) the Angle is : $0.05 \times 360 = 18$ degree

" (B) " " : $0.1 \times 360 = 36$ degree

" (A) " " : $0.45 \times 360 = 162$ degree

" (O) " " : $0.40 \times 360 = 144$ degree

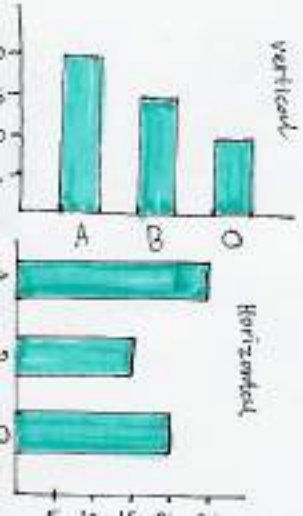
2- Bar chart (class + frequency)

a representation of data of discrete variable with finite values (quan or qual) this is done through vertical or horizontal bars so that each statement has a height equal to the frequency of it.

* the bar chart is useless if the values are large
* the width is uniform for all bars
* the width doesn't matter
* the bars must be seperated

Types of bar charts

1- normal



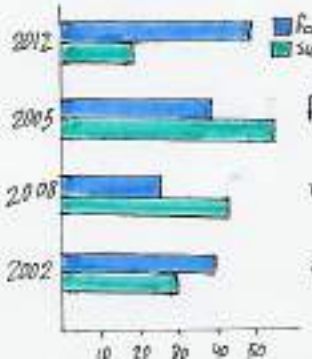
2- Two direction bar chart

a bar chart where we can represent both positive and negative values.

check pg 20 for data

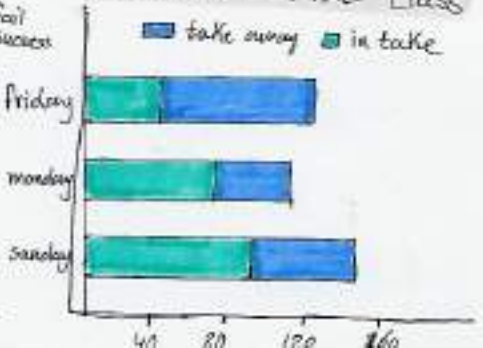
3- Multiple bar chart

a bar chart where we can use it to represent multiple/inter related values by clustering



4- component or stacked bar chart

a bar chart where we can represent each component by a section in the bar and the size is proportional to it's contribution in the class



3- Histogram :-

a graph in which class boundaries are marked on a vertical axis and is constructed by drawing a rectangular column above each actual category so that it's height equals the frequency.

class boundarie + frequency (continuous)

* Similar to a bar chart but the bars must touch and the classes are close to each other

types of histograms

based on shape

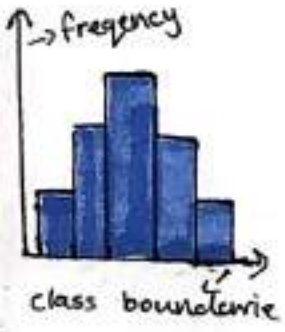
Symmetric

divided into 2 identical sections with respect to a column in the middle

* the highest frequency is in the middle then it decreases from both left and right boundarie
* some have the bell shape
* the tails from left and right are equal

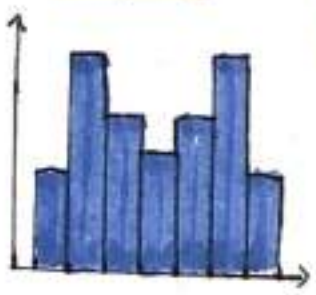
unimodal

one peak



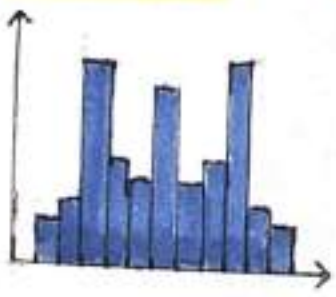
bimodal

two peaks



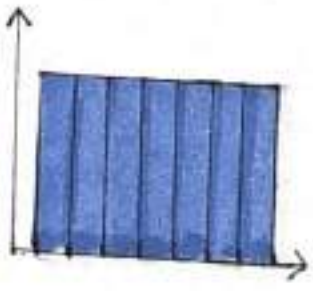
multimodal

more than two peaks



uniform

no peak (same frequency)



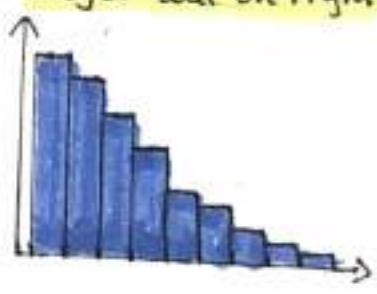
(non-symmetric) skewed

skewedness :- bins on one side have ~~one~~ high frequency which ~~decrease~~ decrease as we move to the other side

* the side with the lower frequency is said to have a longer tail

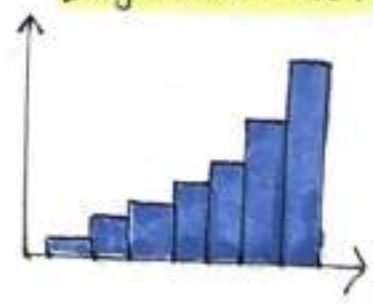
right skewed

Longer tail on right



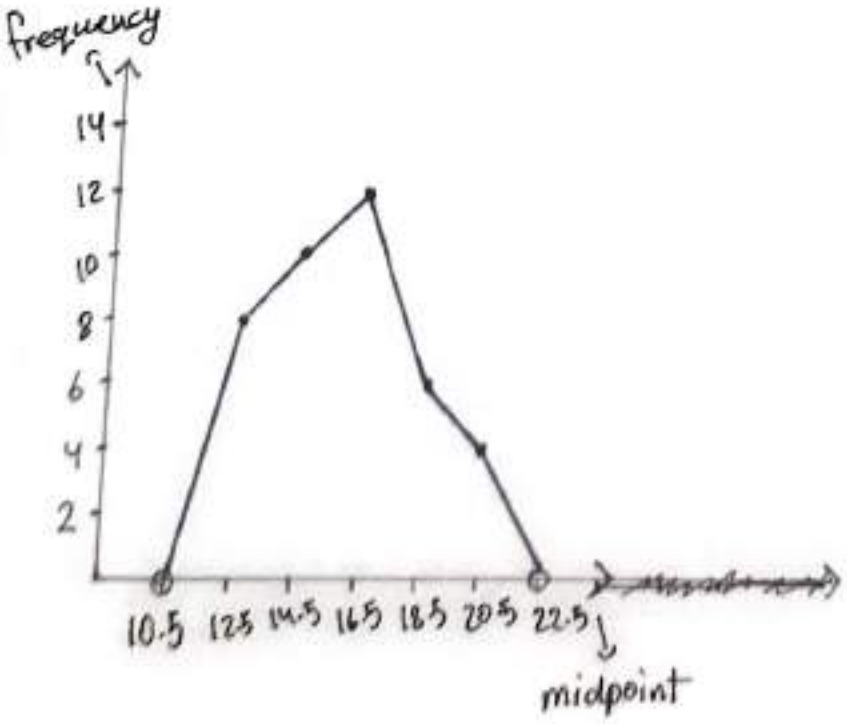
left skewed

Longer tail on left



4- Polygon

(continuous) class midpoint + frequency



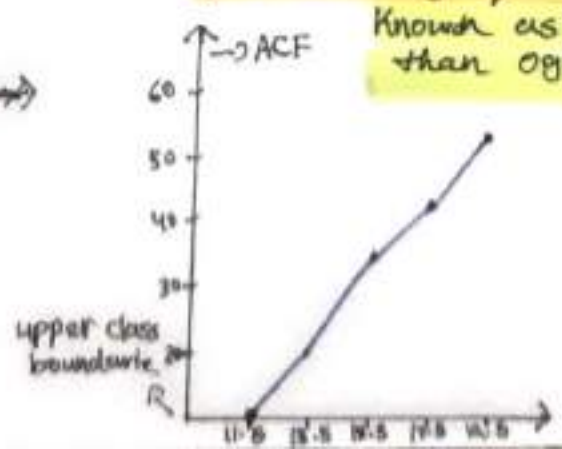
Descending cumulative frequency table (DCFP)

Polygon which connects with a straight line the point (b_i, F_i)

b_i = upper bound of the class

F_i = Ascending frequency of a class

Known as the less than ogive

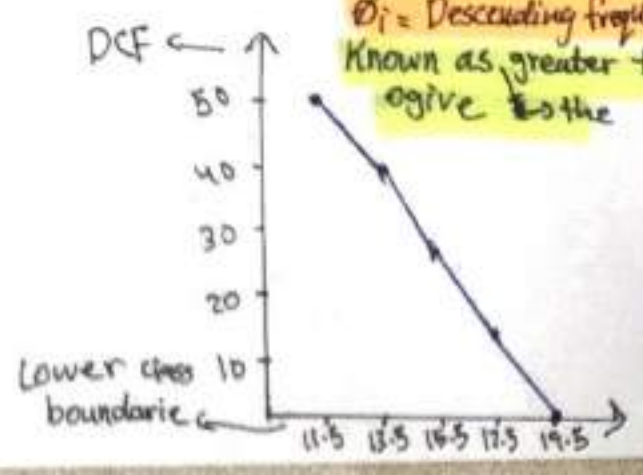


Ascending cumulative frequency table (ACFP)

Polygon which connects with a straight line (b_i, Φ_i) that being a point. b_i = lower bound of a class

Φ_i = Descending frequency

Known as greater than ogive



1.4

Measure of central Tendencies

1- Mean $\bar{X}$ only for qualitative data

Raw data Mean

$$\bar{X} = \frac{\Sigma x_i}{n}$$

Example:-

calculate the mean for this data set

20-18-15-15-14-4

12-11-9-7-6-1 ?

$\bar{X} = \frac{\Sigma x_i}{n}$

$= \frac{20+18+15+15+14+4+12+11+9+7+6+1}{12}$

$= \frac{132}{12} = 11$

frequency table Mean

$$\bar{X} = \frac{\Sigma f_i x_i}{\Sigma f_i} = \frac{\Sigma f_i x_i}{n}$$

Example:-

no. Subject	frequency
0	8
1	18
2	12
3	2
—	40

$\bar{X} = \frac{\Sigma f_i x_i}{\Sigma f_i}$

$= \frac{(0\times8)+(1\times18)+(2\times12)+(3\times2)}{40}$

$= \frac{48}{40} = 1.2$

frequency dist table Mean

$$\bar{X} = \frac{\Sigma f_i x_m}{\Sigma f_i}$$

Example:-

class boundaries	class midpoint	frequency
11.5→13.5	12.5	8
13.5→15.5	14.5	10
15.5→17.5	16.5	12
17.5→19.5	18.5	6
19.5→21.5	20.5	4
—	—	40

$\bar{X} = \frac{\Sigma f_i x_m}{\Sigma f_i}$

$= \frac{(12.5\times8)+(14.5\times10)+(16.5\times12)+(18.5\times6)+(20.5\times4)}{40}$

$= \frac{636}{40} = 15.9$

weighted Mean

$$\bar{X} = \frac{\Sigma w_i x_i}{\Sigma w_i}$$

when weights are equal then ordinary $\bar{X}$ = weight $\bar{X}$

Examples:-

course	grade	point(x)	hour(w)	$\Sigma w_i x_i$
math	B	4	4	16
stat	A	5	3	15
Eng	C	3	3	9
Physics	C	3	4	12
total	—	—	14	52

$\bar{X} = \frac{\Sigma w_i x_i}{\Sigma w_i}$

$= \frac{52}{14}$

$= 3.71$

2- Median $\tilde{x}$ only for Quan data

devids the data in 2 halves after ordering them

Raw data

$\tilde{x} = \frac{x_{n+1}}{2}$ n odd

Example:-

21-22-23-24-26-28-29

$x_1\ x_2\ x_3\ x_4\ x_5\ x_6\ x_7$

$\tilde{x} = x_{\frac{n+1}{2}} = x_{\frac{7+1}{2}} = x_4$

$= 24$

$\tilde{x} = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$ n even

Example:-

21-22-23-24-26-28-29-35

$x_1\ x_2\ x_3\ x_4\ x_5\ x_6\ x_7\ x_8$

$\tilde{x} = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$

$= \frac{x_{\frac{8}{2}} + x_{\frac{8}{2}+1}}{2}$

$= \frac{x_4 + x_5}{2} = \frac{24+26}{2} = 25$

frequency table

$\frac{\Sigma f_i + 1}{2} < x_i$ Σf_i odd

$\tilde{x}$ = the value x_i with the smallest cumulative frequency greater or equal to $\frac{\Sigma f_i + 1}{2}$

Example:- see page 33

Σf_i even

cumulative frequency

if $x_i = \frac{\Sigma f_i}{2}$ then $\tilde{x} = \frac{x_i + x_{i+1}}{2}$

cumulative frequency

if $x_i < \frac{\Sigma f_i}{2}$ and $x_{i-1} < \frac{\Sigma f_i}{2}$ then $\tilde{x} = x_i$

frequency dist table

1- Find median class (smallest cumulative frequency ≥ $\frac{\Sigma f_i}{2}$)

2- $\tilde{x} = L + \frac{\frac{\Sigma f_i}{2} - (\tilde{F} - \tilde{f})}{\tilde{f}} \times C$

L = Lower Limit of median class

$\tilde{F}$ = cumulative frequency of median class

$\tilde{f}$ = frequency of median class

C = class length of median class

Example:- see page 34

3- Mode $\hat{x}$

The highest frequency

Quan and Qual data

* if the highest frequency is constant for all data then the data has no mode.
* multiple values have the highest frequency then the data has more than one mode.

Raw data

Take the value with the highest frequency

Example:-

* A, A, C, A, D, A, B, B, C, D, A, B

The mode is A

* 12, 11, 13, 14, 13, 12, 11, 12, 13, 12, 12, 13, 14, 13

The modes are 12, 13

frequency table

Take the value with the highest frequency

Example:-

Frequency	18	4	16	2
Blood group	A	B	O	AB

The Mode here is A

frequency dist table

1- find the modal class (Frequency is greater than all the others) not extremity

2- $\hat{x} = \hat{L} + \frac{d_1}{d_1 + d_2} \times C$

$\hat{L}$ = Lower class limit of modal

$d_1 = f_{modal} - f_{previous\ class\ directly}$

$d_2 = f_{modal} - f_{next\ class\ directly}$

C = class length

Example:-

class boundaries	50.5-55.5	55.5-60.5	60.5-65.5	65.5-70.5	70.5-75.5
frequency	2	7	8	4	3

modal

$\hat{x} = \hat{L} + \frac{d_1}{d_1 + d_2} \times C = 60.5 + \frac{3-7}{(8-7)+(8-4)} \times 5 = 61.5$

Relationship between Mean, Median, Mode

* symmetric distribution: $\bar{x} = \hat{x} = \tilde{x}$

* left skewed distribution: $\hat{x} > \tilde{x} > \bar{x}$

* Right skewed distrib: $\bar{x} > \tilde{x} > \hat{x}$

	Mean $\bar{x}$	Median $\tilde{x}$	Mode $\hat{x}$
Advantage	- quick and easy to compute - all values are considered - one value for a set of data	- easy to compute and understand - not affected by outliners or extreme values - can be use if there is missing data not in the middle	- quick and easy to compute - for quan and qual - not affected by extreme values
disadvantage	- not for Qual data - highly affected by outliners since it considers all values - not applicable if data is lost	- doesn't take all values into account - not used in many statistical studies - can't be used for qual data	- There may be two or more modes - sometimes there is no mode at all - may not reflect central tendencies accurately

despersions

1- Percentiles :-

$(P_1, P_2, \dots, P_{99})$ devid the ordered data on 100

P_{50} = devid data to 2 halves 50% P_1 = 1% P_{99} = 99%

$P_{25} = Q_1$ $P_{50} = Q_2$ $P_{75} = Q_3$

*** How to calculate?**

1- $P_r = \frac{r(n+1)}{100} = \frac{K}{\dots}, \frac{S}{\dots}$

2- $P_r = x_K + S(x_{K+1} - x_K)$

* It's useful when the data is large ($n \geq 99$)

Example :-

10	20	36	36	36	38	39	40	40	50
51	53	56	56	60	60	60	60	70	70
70	70	72	80	88	88	92	92	92	95

calculate P_{35} for this data?

$P_{35} = \frac{35(30+1)}{100} = \frac{K}{10} \; \frac{S}{.85}$ | $P_{35} = x_{10} + 0.85(x_{11} - x_{10})$

$= 50 + 0.85(51-50) = 50.85$

2- Deciles :-

$(D_1, D_2, \dots D_9)$ devid the data on 10 $D_5 = Q_2 = P_{50}$

*** How to calculate?**

1- $d_r = \frac{r(n+1)}{10} = \frac{K}{\dots} \; \frac{S}{\dots}$

2- $D_r = x_K + S(x_{K+1} - x_K)$

Example :-

calculate the decile 6th for the above data?

$d_6 = \frac{6(30+1)}{10} = \frac{K}{18} \; \frac{S}{.6}$ | $D_6 = x_{18} + 0.6(x_{19} - x_{18})$

$= 60 + 0.6(70-60) = 66$

3- Quartiles :-

(Q_1, Q_2, Q_3) devid the data on 4

$Q_1 = P_{25}$ $Q_2 = P_{50} = D_5$ $Q_3 = P_{75}$

*** How to calculate?**

1- $q_r = \frac{r(n+1)}{4} = \frac{K}{\dots} \; \frac{S}{\dots}$

2- $Q_r = x_K + S(x_K - x_S)$

* used to determine whether a value x of a given data is an extreme values.

Example :-

calculate the Q_1 for this data

Q_2

Q_3

21 22 23 24 26 28 29

* $q_1 = \frac{1(7+1)}{4} = \frac{K}{2} \; \frac{S}{.0}$ | $Q_1 = x_2 + 0(x_2 - x_2) = x_2 = 22$

* $q_2 = \frac{2(7+1)}{4} = \frac{K}{4} \; \frac{S}{.0}$ | $Q_2 = x_4 + 0(x_5 - x_4) = x_4 = 24$

* $q_3 = \frac{3(7+1)}{4} = \frac{K}{6} \; \frac{S}{.0}$ | $Q_3 = x_6 + 0(x_7 - x_6) = x_6 = 28$

Extreme values

* $LF = Q_1 - 1.5(Q_3 - Q_1)$

* $HF = Q_3 + 1.5(Q_3 - Q_1)$

Five #numbers

smallest value, Q_1, Q_2, Q_3 and largest value

Box plot :- the graphical representation of the 5 numbers

* How to construct a box plot?

1- calculate the five numbers

2- draw an axis

3- draw vertical lines on all 3 Quartiles and join them to make a box

4- calculate the end of the right whisker:

- if data does not have a great value then it ends with highest value (x_n)
- if data has great value then it ends at HF

5 calculate the end of the left whisker:

- if data does not have a small value then it ends with smallest value (x_1)
- if data has a small value then it ends at LF

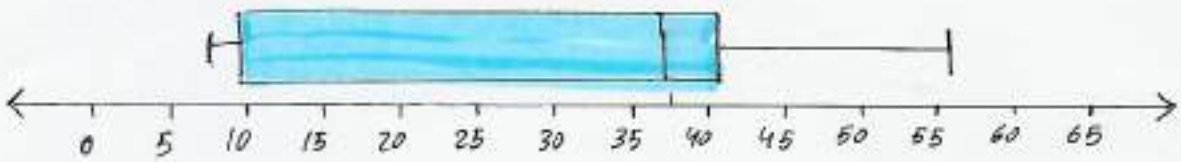
6- extreme values are represented by ~~of~~ (*) or (•)

* for continuous data we can use a box plot to determine if the distribution is symmetric or skewed

left skewed | symmetric | Right skewed

Example:- draw the box plot for these five numbers

7, 9.75, 37.5, 40.75, 56



it does not have extreme values

Measures of variation dispersions

1.5

1- variance S^2

$S^2 \geq 0$ always

Raw data

* if $n=1$ then $S^2 = 0$
* expressed in square ($*^2$)

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

$$S^2 = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1}$$

frequency table

* if $\sum f_i = 1$ then $S^2 = 0$
* expressed in square ($*^2$)

$$S^2 = \frac{\sum f_i (x_i - \bar{x})^2}{\sum f_i - 1}$$

frequency dist

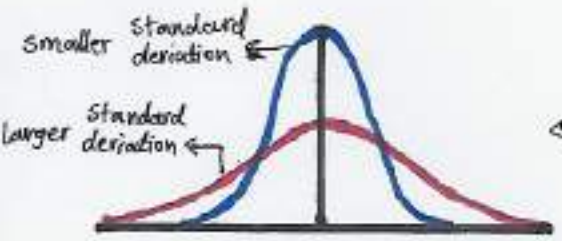
* if $\sum f_i = 1$ then $S^2 = 0$
* expressed in square ($*^2$)

$$S^2 = \frac{\sum f_i (x_m - \bar{x})^2}{\sum f_i - 1}$$

2- Standard deviation S

$S \geq 0$ always

$$S = +\sqrt{S^2}$$



* Standard deviation is the best mesure of despersion
* we use measures of despersion when the mean is used as a central tendency measure
  - small value of standard deviation = variables close to mean
  - large value of standard deviation = variables far from mean
* highly affected by extreme values
* not applicable if data is lost

3- coefficient of variation CV

$$CV = \frac{S}{\bar{x}} \times 100\%$$

$\bar{x} \neq 0$

* a useful measure of variation to compare between sets of data with different units (measures)

4- Range R

Raw data
$R = x_L - x_S$

frequency
$R = x_m - x_1$

grouped data
(distribution table)

$R = x_K - x_1$

Interquartile
$IQR = Q_3 - Q_1$

* it is also called as mid-spread because it gives us the middle 50% of the values

* we can use the 5 numbers to find:-
1- Ranges
2- IQR
3- Median

5- Z-score (Standerd score) Z_{x_i}

* converts data to make it's mean = 0 and S = 1

$$Z_{x_i} = \frac{x_i - \bar{x}}{S}$$

* we use the numerical value to assign a degree of data to tell us it's position compare to the rest of the data

*
مؤشر لاعرف ان قيمة x_i عن المتوسط

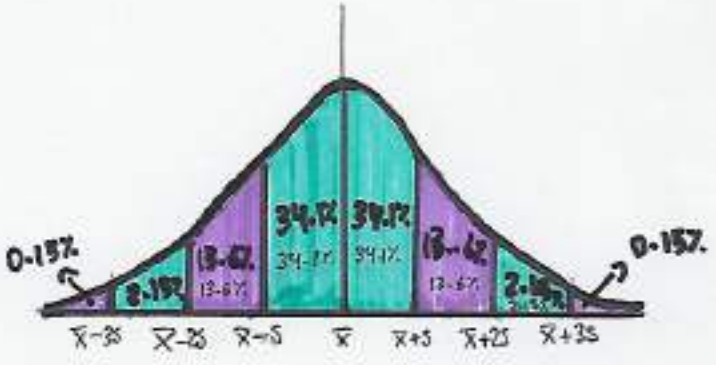
$Z_{x_i} > 0$ يعني أن x_i اكبر من المتوسط

$Z_{x_i} < 0$ " " " اصغر "

The empirical Rule

* if a data set has an aproximate bell shape RF histogram

- 68.2% of data lies in $\bar{x} \pm S$
- 95.4% of data lies in $\bar{x} \pm 2S$
- 99.7% of data lies in $\bar{x} \pm 3S$



chapter 2

Probability

2.1 - 2.2 - 2.3 - 2.4

The fundemental principle of counting

2.1

has | 2 Rules

Multiplicative Rule

If we have A ways of doing something and have B ways of doing another then there are A×B ways of doing both

* "and usually requires us to multiply"
* independent events

Addition Rule

if we have A ways of doing something and have B ways of doing another and we can't do both at the same time then there are A+B to choose 1 action

* "or usually requires us to add"
* mutually exclusive events

for shortcuts we use factorial notation

which is

$n! = n \times (n-1) \times (n-2) \times (n-3) \times \cdots \times 2 \times 1$

1- $0! = 1$ 2- $1! = 1$

3- $n \geq 2 \rightarrow n! \approx n^n e^{-n} \sqrt{2\pi n}$

Some special cases:-

Permutation	combination
- Any arrangment of r distinct objects, from a set of $n \in \mathbb{N}$ (natural numbers) different objects.	- Any unordered group of r distinct objects from a set of n different objects.
$nP_r = \frac{n!}{(n-r)!}$; $0 \leq r \leq n$ | $nP_n = n!$	$nC_r = \frac{n!}{r!(n-r)!} = \frac{nPr}{r!} = \binom{n}{r}$; $0 \leq r \leq n$
care about the order	don't care about the order
عادة يكتب بالسؤال in order فستخدم تباديل	عادة يكتب بالسؤال unordered فستخدم توافيق

cardinal number of a set $|\Omega|$

Let Ω be a given set, then the number of all elements in Ω is $|\Omega|$

* Ω is infinite and it's countable (elements $\in \mathbb{N}$) then $|\Omega| = \infty$.
* Ω is infinite and it's not countable (elements $\notin \mathbb{N}$) then $|\Omega| = \mathfrak{P}$ (Ω has continuous capacity)

Types of Experiments

2.2

Regular (systematic)

we know it's outcomes in advance

Random (stochastically)

we don't know it's outcomes in advance but we can determine the set of results

* each possible outcome is called an ~~elementary~~ elementary event
* 2 or more outcomes are said to have the same chance of appearing

Probability science :-

a branch of mathmatics that deals with theoretical mathematical models of random experiments.

theoritical mathmatical model (probability space) $[\Omega, \mathcal{A}, P]$

Ω

The set of all possible results of a random experiments (space of elementry events)

- **Discrete**: finite or countable Infinite events
- **continuous**: uncountable number of outcomes

Types of events

- **Simple**: if it contains one elementry event (outcome)
- **compound**: if it contains two or more elementry events

$\mathcal{A}$

(Algebra of events) a collection of events of interest with these condition :-

1- * $\Omega \in \mathcal{A}$ or $\mathcal{A} \subseteq \Omega$

2- * any 2 elements A and B $\in \mathcal{A}$ then $A \cup B \in \mathcal{A}$.

3- * any element $A \in \mathcal{A}$ then $\bar{A} \in \mathcal{A}$

* if this condition is true :- $A_1, A_2, A_3, A_n \ldots \in \mathcal{A}$ and $A_1 \cup A_2 \cup A_3 \ldots$ then $\mathcal{A}$ is σ-algebra
* elements of $\mathcal{A}$ (algebra or σ-algebra) are called events

* $\mathcal{A} = 2^{\Omega}$

P

(probability function) a measure of the likelihood a specific event will occur

- **RF**: if the experiment A repeating N times and the event has n trials

  $P(A) = \frac{n(A)}{N}$

  * Relative frequency is not a probability but they're approximite probability
  * But when repeated again and again they become actual probability
- **classical**: if all the events have the same chance in appearing (fair)

  $P(A) = \frac{|A|}{|\Omega|}$ ← compound

  $P(A) = \frac{1}{|\Omega|}$ ← simple
- **thearm**: if the experiment is not fair and Ω finite

  $P(A) = \sum P(\omega_i)$

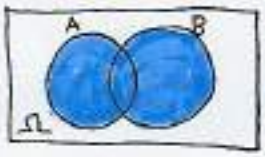
Some operations on events

2.2 - 2.3 - 2.4

1- UNION of two events:- $A \cup B$

an event containing all elementry events that belong to A or B or to both.

$A \cup B = \{w : w \in A \text{ or } w \in B\}$



Example:- $A=\{1,3\}$

$A \cup B = \{1,2,3,4\}$ $B=\{2,4\}$

Additive Law in probability:-

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

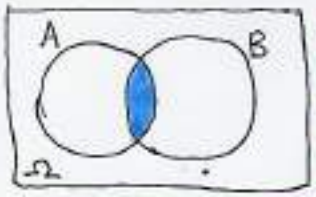
$P(A \cup B) = P(A) + P(B)$ ⇒ if mutually exclusive

$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$

2- Intersection of 2 events:- $A \cap B$

an event containing all elementry events that belong to A and B or both occuring together

$A \cap B = \{w : w \in A \text{ and } w \in B\}$



Example:-

$A=\{TT, HT\}$ $B=\{TT, TH\}$

$A \cap B = \{TT\}$

Multiplicative law in probability:-

$P(A \cap B) = P(B) \times P(A|B)$

$P(A \cap B) = P(A) \times P(B|A)$

$P(A \cap B) = P(A) \times P(B)$ => if independent

3- Complement of an event :- $\bar{A}$

an event that occurs when A doesn't

$\bar{A} = \{w : w \in \Omega, w \neq A\}$

$\Omega = \{1,2,3,4,5,6\}$ $A=\{1,2\}$

$\bar{A} = \{3,4,5,6\}$

$P(\bar{A}) = 1 - P(A)$

De morgan's Law :-

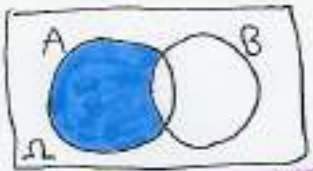
$P(\bar{A} \cup \bar{B}) = \overline{P(A \cap B)}$

$P(\bar{A} \cap \bar{B}) = \overline{P(A \cup B)}$

4- Difference Between 2 events :- $A \setminus B$

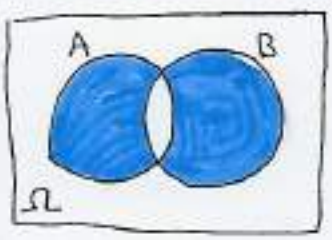
the event of the occurance of A but not B

$A \setminus B = A \cap \bar{B} = \{w : w \in A \text{ and } w \notin B\}$



$P(A \setminus B) = P(A) - P(A \cap B)$

5- exactly one of event δ $A \triangle B$



the event of A and B occur but not together $A \triangle B = \{x : x \in A \cap \bar{B} \text{ or } x \in B \cap \bar{A}\}$

$= \{x : x \in A \cup B \text{ and } x \notin A \cap B\}$

$A \triangle B = (A \setminus B) \cup (B \setminus A)$

6- impossible event $\emptyset$

The event of an occurance being impossible

Example:-

$A \cap \bar{A} = \emptyset$

certain event Ω

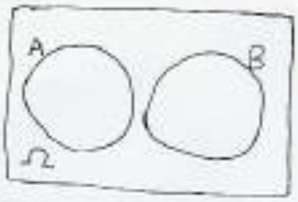
An event that is sure to happen

Example :- * $A \cup \bar{A} = \Omega$

* flipping a coin once and getting a head and tail

Mutually Exclusive event

they can't occur at the same time



that means $A \cap B = \emptyset$

Example:-

getting on a plane or not getting on a plane

conditional probability $P(A|B)$ 2.4

the probability of A occuring given that B also occurs

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

independent ← $P(A|B) = P(A)$

independent events

The occurance of B doesn't effect the occurance of A

* باختصار، حدثين مالهم دخل ببعض مثل اللي اداوم أو تتخرج بدري. ما يأثرون على بعض

* Mutually exclusive events are not the same as independent

* A and $\bar{B}$ → independent

* $\bar{A}$ and B → independent

* $\bar{A}$ and $\bar{B}$ → independent

conditions for partitioning Ω:

1- $Z_i \neq \emptyset$ → no impossible events

2- $Z_i \cap Z_j = \emptyset$ (mutully exclusive)

3- $\cup Z_i = \Omega$ → union gives Ω

Bayes' Theorem

an equation that relates two conditional probabilities of the form $P(B|A)$ and $P(A|B)$

$$P(Z_i|B) = \frac{P(Z_i)P(B|Z_i)}{\sum P(Z_k)P(B|Z_k)}$$

Total probability theory

$$P(B) = \sum P(Z_k)P(B|Z_k)$$

توضيح المعادلة فوق

$P(B) = P(B \cap Z_1) + P(B \cap Z_2) + P(B \cap Z_3) + \cdots + P(B \cap Z_K)$

$= P(Z_1)P(B|Z_1) + P(Z_2)P(B|Z_2) + P(Z_3)P(B|Z_3) + \cdots + P(B \cap Z_K)P(Z_K)$

CHAPTER 3

Random variables and Probability distribution

3.1 3.2 3.3

Random variable

3.1 3.2 3.3

* denoted by X, Z, Y and the values by x, z, y
* $\{W \in \Omega ; X(w) \leq x\} \in \mathcal{F}$ $\forall x \in \mathbb{R}$
  $(X \leq x)$ ↑ اختبار

Discrete

finite or infinite countable

Proving X is a discrete random variable has 2 conditions:-

1- $P(X=x) \geq 0$

2- $\Sigma P(X=x) = 1$

Probability mass function (P.M.F) $f_x(x)$

(تقدر تقول احتمال ان x_i حدوث)

Properties and conditions:-

1- $P(X=x) \geq 0$

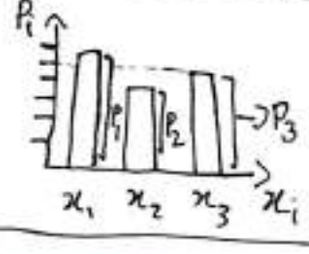
2- $\Sigma P(X=x) = 1$

can be represented in a ① table or ② graph

1-

x Random variable	x_1	x_2	..	x_n	Σ
$P(X=x) = f_x(x)$	P_1	P_2	..	P_n	1

2-



mean:-

$\mu = E(X) = \Sigma x_i P(X=x)$

variance:-

$\sigma^2 = E(X^2) - \mu^2 = \Sigma x_i^2 P(X=x) - \mu^2$

standard deviation:-

$\sigma = +\sqrt{\sigma^2}$

Distribution function $F_x(x)$

$F_x(x) = \Sigma P(X=x)$

* $0 \leq F_x(x) \leq 1$
* $\lim_{x \to -\infty} F_x(x) = 0$ (قيمة قريبة من 0), $\lim_{x \to +\infty} F_x(x) = 1$ (قيمة قريبة من 1)

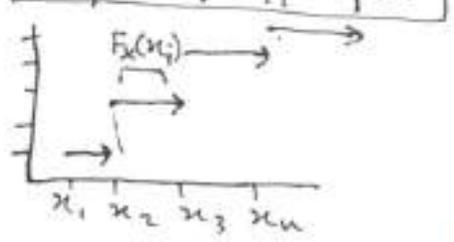
$$F_x(x) = \begin{cases} 0 & x < x_1 \\ P_1 & x_1 \leq x < x_2 \\ P_1+P_2 & x_2 \leq x < x_n \\ P_1+P_2+P_n = 1 & x \geq x_n \end{cases}$$

can be represented in a ① table or ② graph

1-

For	$x < x_1$	$x_1 \leq x \leq x_2$	....
$F_x(x)$	0	P_1	...

2-



continuous

uncountable set of numbers

Probability density function (P.d.F)

Properties and conditions:-

1- $P(a < x < b) \geq 0$

2- $\int_{-\infty}^{+\infty} f_x(x)\,dx = 1$

* $P(a \leq x \leq b) = \int_a^b f_x(x)\,dx$
* $f_x(x) = \frac{d}{dx} F_x(x)$

distribution function

$F_x(x) = \int_{-\infty}^{x} f_x(u)\,du$

Mean

$\mu = E(X) = \int_{-\infty}^{+\infty} u f_x(u)\,du$

variance

$\sigma^2 = \int_{-\infty}^{+\infty} x^2 f(x)\,dx - \mu^2$

some special distribution

Exponential

times between events occurring

* $\lambda > 0$

P.d.F

$$f_x(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases}$$

distribution function

$$F_x(x) = \begin{cases} 1-e^{-\lambda x} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases}$$

mean: $\mu = \frac{1}{\lambda}$

variance: $\sigma^2 = \frac{1}{\lambda^2}$

standard deviation: $\sigma = \frac{1}{\lambda}$

Normal

natural phenomena in Real life

P.d.F

$f_x(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left[\frac{x-\mu}{\sigma}\right]^2}$

distribution function

$F_x(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}\left[\frac{x-\mu}{\sigma}\right]^2}$

* in a graph it has a bell shaped
* one peak in the center = mean = mode = median

standardizing Normal distribution :-

to help compute areas under P.d.F curve

$$Z = \frac{X-\mu}{\sigma}$$

some Special distributions

Binomial

2 possible outcomes success (p) and failure (q) repeated n times

* $1 > p > 0$ * $1-p=q$
* trials are independent

P.M.F

$P(X=x) = \binom{n}{k} p^k q^{n-k}$

mean: $\mu = E(X) = np$

variance: $\sigma^2 = npq$

Distribution function

$F_x(x) = \Sigma P(X=x)$

* $\Sigma P(X=x) = 1$

Poission

events occuring in an amount of specific time

* $\lambda > 0$

P.M.F

$P(X=x) = \frac{x^x e^{-\lambda}}{K!}$

Distribution function

$F_x(x) = \Sigma P(X=x)$

mean: $\mu = E(X) = \lambda$

variance: $\sigma^2 = \lambda$

CHAPTER 4

Introduction to statistical inference

4.1 4.2 4.3 4.4 4.5 4.6

in INFrential statistics we have something called an Estimator:-

a statistic (function of the (random sample) whose value depends on the particular sample is drawn on the population

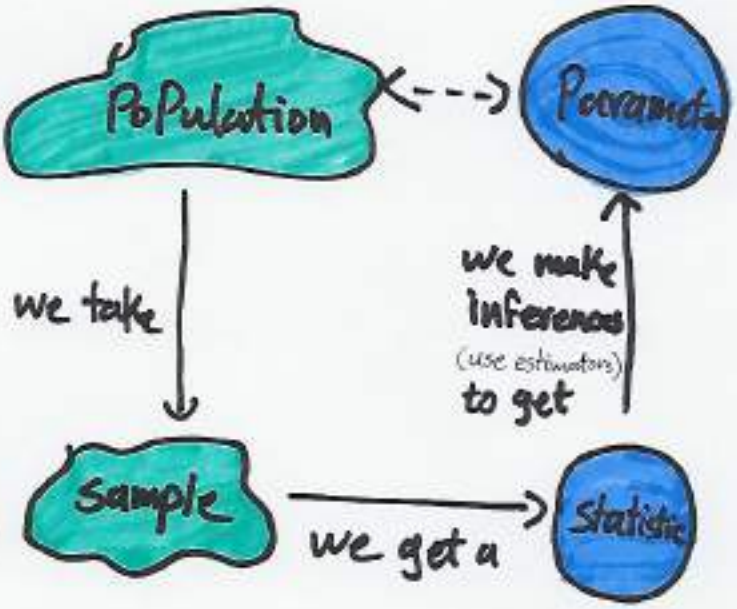
in a sample space Ω we take a simple random sample of Ω denoted by $\mathcal{X}$ (Ω = population, $\mathcal{X}$ = Sample)

the mean of $\mathcal{X}$ is $\overline{X} = \frac{X_1 + X_2 + \dots X_n}{n}$ which is an estimator

the variance of $\mathcal{X}$ is $S^2 = \frac{\sum (x_i - \overline{X})^2}{n-1}$ which is an estimator

* we use the estimators to make predictions about the value of a population. explained in this graph

Since the population is HUGE we can't get an accurate value of parameter so we use samples



((شرح الفكرة ال estimators وكيف نستخدمها لايجاد ال Parameter))

((الفرق بين $\overline{X}$ و S^2 في الجزء الأخير من الصفحة مهم جدًا لا تنسى))

Standard Normal Probabilities Z_α

4.1 4.2 4.3

الفيديو هذا يشرح الفكرة أفضل !

The value on the real axis for which is the area between the left part of the curve of $p(x)$ and the straight $x = Z_\alpha$ equal to α

confidence Level	Confidence coefficient	Reliability coefficient	
$(1-\alpha)\%$	α	$Z_{1-(\frac{\alpha}{2})}$	$Z_{1-(\frac{\alpha}{2})}$
0.90	0.10	$Z_{0.95}$	1.645
0.95	0.05	$Z_{0.975}$	1.96
0.98	0.02	$Z_{0.99}$	2.33
0.99	0.01	$Z_{0.995}$	2.575

Normal distribution عند $Z = \frac{X-M}{\sigma}$ وتطلع لنا قيمة. بهذي الحالة يعطيك القيمة الاخيرة وبيك تروح القيمة الاولية

اسحبوا على كل ذا الكلام. إحنا تشابه ٣ قلنا بأن عشان نخليه قياسي نستعمل ذا القانون

The central limit theorem

When the sample size is large ($n \geq 30$)
then the sample distribution follows a normal distribution

Sample distribution of mean

mean: $\mu_{\bar{x}} = \mu$

standard error: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

Standarizing the distribution:

$$Z = \frac{\bar{X} - \mu_{\bar{x}}}{\frac{\sigma}{\sqrt{n}}} = \frac{\bar{X} - \mu_{\bar{x}}}{\sigma_{\bar{x}}}$$

Sample distribution of proportion

mean: $\mu_p = p$, standard error: $\sigma = \sqrt{\frac{\hat{P}(1-\hat{P})}{n}}$

Standerizing the distribution:

$$Z = \frac{\bar{X} - \mu_{\bar{x}}}{\sigma_p}$$

Types of estimations

Point estimation:- an estimate of the Population Parameter by a single number

Interval estimation:- Predicted to have the value of Population Parameter with a certain Probability

Estimation of the Population mean

interval estimation

*confidence interval:- a range of values that have the true value of population Parameter with specific Probability (confidence level) (level of confidence)

case	Population	Sample size	Standered deviation S.D	confidence interval
1	normal	any	Known	$\bar{x} \pm Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$
2	any	large $n \geq 30$	Known	$\bar{x} \pm Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$
3	any	large $n \geq 30$	unKnown	$\bar{x} \pm Z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$

cases and their confidence interval

But replaced with Sample S.D

* margin error:- the maximum error of using the value of statistics instead of Parameter

$$\delta_\mu = Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \text{ or } \delta_\mu = Z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$$

to find sample size:

$$n = \left[\frac{\sigma Z_{1-\frac{\alpha}{2}}}{\delta_\mu}\right]^2$$

Point estimation

Parameter = statistic

$\mu = \bar{X}$

Estimation of the Population Proportion

* confidence interval for Population Proportion:-

$$P = \hat{P} \pm Z_{1-\frac{\alpha}{n}} \left(\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}\right)$$

$P = \frac{N(A)}{N}$, $\hat{P} = \frac{n(A)}{n}$

* Margin Error:-

$$\delta_p = \hat{P} \pm Z_{1-\frac{\alpha}{n}} \left(\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}\right)$$

* sample size:-

1- $\hat{P}$ is given:-

$$n = \left[\frac{Z_{1-\frac{\alpha}{n}}}{\delta_p}\right]^2 \hat{P}(1-\hat{P})$$

2- $\hat{P}$ not given

$$n = 0.25 \left[\frac{Z_{1-\frac{\alpha}{n}}}{\delta_p}\right]^2$$

3- L is given

L = maximum length

$$n = \left[\frac{Z_{1-\frac{\alpha}{n}}}{L}\right]^2$$

Statistical Hypothesis

4.4 4.5 4.6

an argument about a specific statistical question, it can be true or wrong

شرح بسيط للاف فكر

Null Hypothesis H_0

a statment under an investigation or testing

* we usually use =
* represent a statement of "no effect" or "no difference"

باختبار null نفترض شركة ادعت ان حليبها وزنه 1Kg هذا هو null وان اخذنا عينة وطلع 0.99Kg هذا فبنوي اختبار alternate لاعرفين عن طريق test statistic شي اسمه ونشوف اي فرضية صحيحة

Alternate Hypothesis H_1

a statement we will adopt in the situation in which the evidence is so strong that we will reject the null

* help us know if the null hypothesis is true or not
* we usually use $<, \neq, >$

Types of error:-

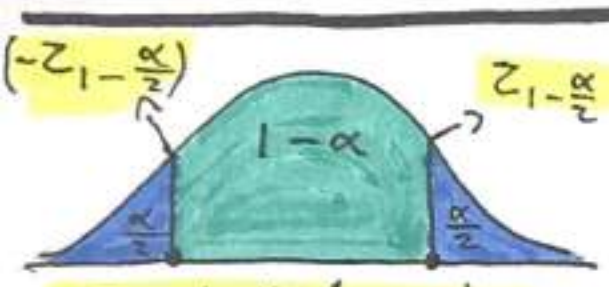
Decision \ actual situation	H_0 true	H_0 is false
reject H_0	correct decision	Type II error: $P(\text{Type II error}) = \beta$
don't reject H_0	Type I error: $P(\text{Type I error}) = \alpha$	correct desicion

α = significane error $1-\beta$ = the power of test

The critical regions and values :-

* a region that produced by the values that corresponds to the rejection of the null hypothesis at some chosen level of significance
* the values that determine the critical region(s) is called critical value

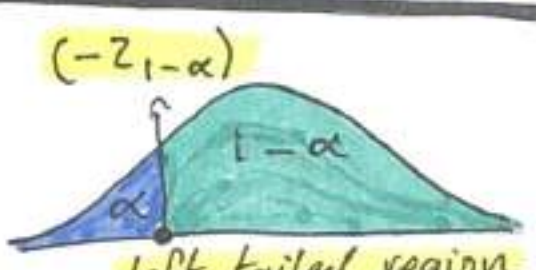
$-(Z_{1-\alpha}), (Z_{1-\frac{\alpha}{2}}), (-Z_{1-\alpha}), (-Z_{1-\frac{\alpha}{2}})$



Two tailed region

$\mu \neq \mu_0$

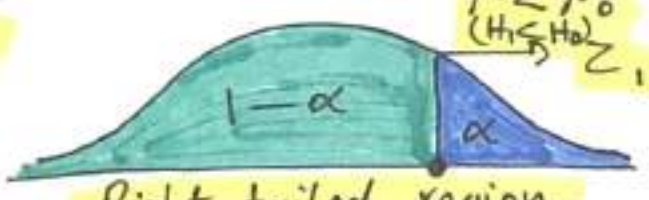
$(H_1 \neq H_0)$



Left tailed region

$\mu < \mu_0$

$(H_1 < H_0)$



Right tailed region

$\mu > \mu_0$

$(H_1 > H_0)$

Accepted region: H_0 تقبل فيه فرضية

rejected region: H_0 نرفض فيه فرضية

* نختار اي critical value نستعمل عن طريق العلاقة بين H_1 و H_0 وايضا عن طريق التوزيع الطبيعي القياسي (الرسمة)

Test statistic

is a statistic whose calculated value from the sample is used to make a decision on Hypothesis test

for mean :-

$$z_0 = \begin{cases} \frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}} & \sigma \text{ Know} \\ \frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}} & \sigma \text{ unKnow and } n \geq 30 \end{cases}$$

for propotion :-

$$z_0 = \frac{\hat{P} - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}}$$

Hypothesis test

Population mean

How to :-

1- determine H_0, H_1
2- determine test statistic
3- determine critical region
4- make a decision wether H_0 is true or not

Pvalue :-

Assuming H_0 is true, the Probability that the test statistic will take values as extreme as or more extreme than the observed test statistic

check example: 4.5.1 & 4.6.1

* the smaller p value the stronger the evidence against H_0

Population Proportion

1- determine H_0, H_1
2- determine test statistic
3- determine critical region
4- make a decision

CHAPTER 5

Correlation and Regression

5.1 [redacted] 5.2

Pearson's correlation Coefficient r :- 5.1

a measure of linear correlation between X & Y

$$r = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\Sigma(x_i-\bar{x})^2}\cdot\sqrt{\Sigma(y_i-\bar{y})^2}}$$ or $$r = \frac{n\cdot\Sigma x_i y_i - \Sigma x_i \cdot \Sigma y_i}{\sqrt{n\cdot\Sigma y_i^2-(\Sigma y_i)^2}\cdot\sqrt{n\cdot\Sigma x_i^2-(\Sigma x_i)^2}}$$

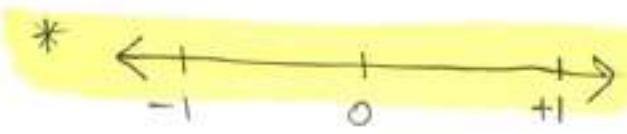
*

Relationship between X & Y	Range of r
no linear $S_y=0$ $S_x=0$	$r = 0$
very weak	$0 < \pm r \leq 0.30$
weak	$0.3 < \pm r \leq 0.5$
moderatly strong linear	$0.5 < \pm r \leq 0.7$
Strong	$0.7 < \pm r \leq 0.86$
Very Strong	$0.86 < \pm r \leq 1$
complete	$\pm r = 1$

* the values ranges between −1 and +1

* +r means $x \propto y$ and it's called Positive monotone

* −r means $x \propto \frac{1}{y}$ and it's called negative monotone

* 

كلما كانت r قريبة من ±1 كلما كان العلاقة واد' r=0 فلا توجد علاقة

* Scatter plot :-

a graph of data give in the form of (x_i, y_i)

* check page 205 For Examples

Equation for linear regression: $\hat{y} = \hat{a} + \hat{b}x$ 5.2

* constant $\hat{b}$ = the slope * constant $\hat{a}$ = Y-intercept

* $\hat{b} = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\Sigma(x_i-\bar{x})^2}$ * $\hat{a} = \bar{y} - \hat{b}\bar{x}$

/ $\bar{y} = \frac{\Sigma y_i}{n_y}$ / $\bar{x} = \frac{\Sigma x_i}{n_x}$

coefficient of determination r^2:

$$r^2 = \frac{SSR}{SS_{tot}} = 1 - \frac{SSE}{SS_{tot}} = \%$$

* $SS_{tot} = \Sigma(y_i-\bar{y})^2 = SSR + SSE$

total sum of squared deviation

* $SSR = \Sigma(\bar{y}-\hat{y}_i)^2$

sum of squared regression error (Explained variation)

* $SSE = \Sigma(y_i-\hat{y}_i)^2$

* the values range $0 \leq r^2 \leq 1$

* $r^2 = 0$ no explanatory value

* $r^2 = 1$ explains 100%

تمثيل معادلة الانحدار الخطي

نحتاج نقطتين فقط

فنختار واسم من دولا

1- $(X_1, \hat{Y}_1)$ نقطة , $(X_2, \hat{Y}_2)$ نقطة

ناخذ قيمة x ونعوضها بالمعادلة

2- $(\bar{X}, \bar{Y})$ نقطة , $(X_1, \hat{Y}_1)$ نقطة

نمثل الـ mean لـ X و Y

وناخذ قيمة x ونعوضها ← Y

* scatter plot نجرين نمثل النقطتين على وتوصل بينهم بخط