

• Data •

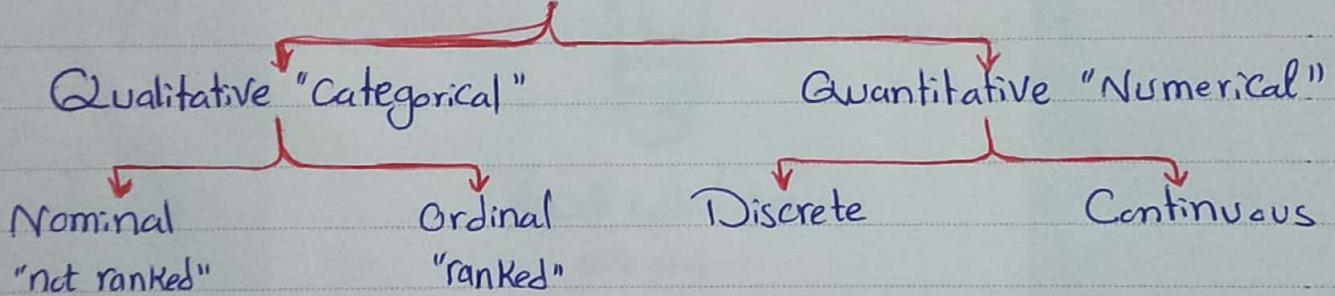
* Data is a Collection of facts or observations about Something.

* Data matters as :

- 1) Helps us understand things as they are. يساعدنا في فهم الأمور كما هي.
- 2) Predict Future behavior ; to guid business. تنبئ بسلوك السوق future market behavior.

ـ ملخص أنواع البيانات

• Data •



ـ النتائج التي نحصل عليها من التحليلات:

• Population :

↳ Contains every member of a group.

• Sample :

↳ a Subset of Population Members.

عن التكلم عن البيانات فإننا نستطيع أن نتأمل بها عن طريق علم الإحصاء (Statistics) وعندما نريد أن نتوقع شيء من ملأ البيانات فإننا نحتاج علم الاحتمالات (probability) وأحياناً نفتح الفرق بعد السطح.

• Statistics.

* Statistics : is the science of manipulating , analyzing and summarizing data.

- الإحصاء هو علم يتطلب وتحليل وفهم البيانات.
- الإحصاء في ذاته ينقسم إلى فرعين رئيسين وهما :

① Descriptive statistics :

↳ Summarizing the data through the given observations ~~through~~ Using a Sample.

الإحصاء الوصفي : وهو الفرع الذي يتم ب幫طه ملخص عن البيانات.

لوصف بعض الخصائص عن هذه البيانات.

للتوصيم فكرتك الرئيسية على أخذ عينة (Sample) من المجتمع (population).

الموحد ومن ثم على تأثيرها.

نقوم بعمل تأثير البيانات من ملأ معهم مقاييس.

* Measurements to summarize data:

① Measurements of Central tendency :

↳ Describes the location of data.

↳ Fails to describe the shape "distribution" of the data.

نقوم بوصف "أو بـ طار ملخص عن مكان البيانات" ولكننا لا نستطيع.

أن نعطي وصف عن شكل توزيع البيانات.

هذه المعايير ثلاثة :

(Mean , Median , Mode)

- Mean : \bar{X} , $E(X)$

↳ Calculated average

$$\bar{X} = \frac{\sum X_i}{n}$$

مقدمة الفم

$$\mu = \frac{\sum X_i}{N}$$

في حالة انتهاج

- Median :

↳ Middle Value القيمة الوسطى في الترتيب من بينها.

- Mode :

↳ Most frequent data "Value". القيمة الأكثر تكراراً.

notes :

- The mean can be influenced by outliers; but not the median.
- The median is much closer to most of values in the sample.

② Measurements of Dispersion :

↳ Describes the shape of the data and how it's spreaded out.

نوعي ونوعي ونوعي ونوعي ونوعي ونوعي .

"Range - Variance - Standard deviation"

- Range : difference between max. and min. values.

$$\text{Range} = \text{max.} - \text{min.}$$

البيانات

Variance: Sum of Squares distance from each point to the mean.

الفكرة هنا في اعطاء رقم يعبر عن مدى ابعاد القيم (Values) عن القيمة

الرسومية لذا هل البيانات متجهة في اتجاه مماثلة او موجعة

إذا كانت قيمة Variance كبيرة، فإنها تعرف أن البيانات موزعة

وإذا كانت قيمة مغيرة فإنها تعرف أن البيانات تقترب من بعضها

من التعريف:-

$$\text{population} \leftarrow \sigma^2 = \frac{\sum (x - \bar{x})^2}{N}$$

نما في حالة الـ Sample ففيه المقدمة

(Bessel's Correction) وهذا يسمى بـ (عدد العناصر - 1)

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

هذا يعنينا قيمة تكون وحدة كيلوغرام مربع (Kg²) لذلك نستخدم الـ (Standard deviation)

Standard deviation: Square-root of the variance.

فإنها هو جذر وحدة القياس في نفسها مثل البيانات.

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

هل هناك فرق بين الـ Variance والـ (Std) :

هذه الفرق لا تؤثر على البيانات بنفسها ووفقاً لها الـ (mean) مثلاً وإنما

هي قيمة تغير عن هذه توزع البيانات ولكن الـ (Std) يفضل هنا

يعبر قيمة ملخص القيم - يعني ترتيبها -

③ Measurement of Variability :

→ Has the advantages that every data point is considered ; not aggregated.

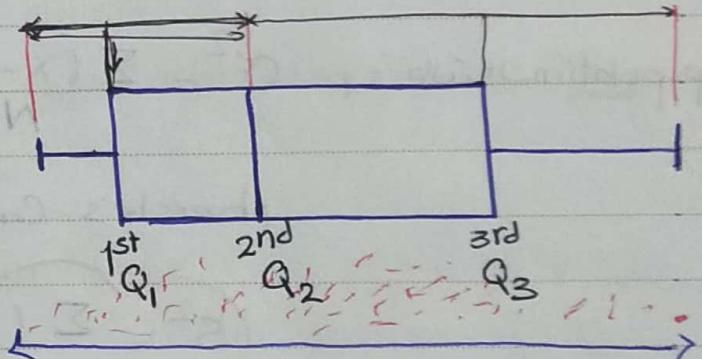
لعن المات (box-plots) لبيان المدى.

• Inter-quartile range (IQR) :

median ال Median ①

Q1 ال Median ال Q3

$$\text{IQR} = Q_3 - Q_1$$



• Fences & Outliers :

$$\text{Fences} = 1.5 \times (\text{IQR})$$

Fences < outlier

"descriptive statistics" مفهوم احصائي

② Inferential Statistics : "Statistical inference"

→ Used to interpret the meaning of descriptive statistics, Helps us to make prediction.

هذا الفرع هو الفرع الاستدلالي والذي يعتمد على "توقعات بناءً على ما أخرجناه من احصائيات descriptive statistics population" خارج نمونة (sample) من كل

نستخدم الاحتمالات (probability) لنتعرف باختصار نسبة تغير العينة عن المجموع "معقول" كافية للاستبعان عن كل المجموع أو لا".

الاحداث التنجلياتية (ها هي مفهوم رئيسيان :

- * Hypothesis Testing.
- * Confidence level.

- It's method of making decisions about the parameters of population ; based on Random Sampling

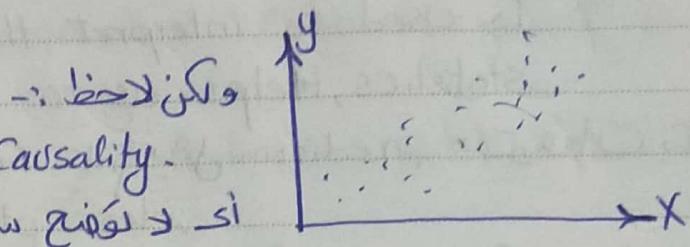
هناك نوعان كاملاً ...

~~~~~  
ما يكتفي به في ادنى حد هو (Uni-variate)

- Uni-variate data : one variable "one type of data".

- Bi-variate data ; measures "Compares" two types of data to find a relationship between them.  
"Correlation"

لعرف العلاقة بين معلومتين من المعايير (Scatter plot)  $(x, y)$ .



Scatter plot can't show Causality.

أى دلائل يوضح سبب العلاقة بين اد  $(x, y)$   
وأماماً يوضح أن هناك علاقة "ولكن لا نعرف سببها".

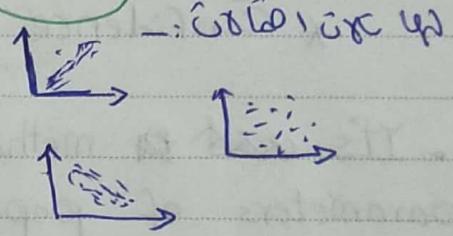
• **Co-variance** الاتجاه يرغم على معاييره.

• **Covariance**:

↳ A way used to mathematically quantify the relationship between two variables.

$$\text{Cov}(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N}$$

$$\text{Cov}(x, y) = \begin{cases} +ve & \text{positive correlation} \\ \approx 0.0 & \text{no} \\ -ve & \text{negative} \end{cases}$$



ولكن هنا الأدلة تكون غير معتبرة بشكل كبير  
لذلك ننتهي اد

• **Pearson Correlation factor**:

$(-1, 1)$  نفس الاتجاه دلائل فعل قوي بين.

$$P_{xy} = \begin{cases} 1 & +ve \text{ Correlation} \\ 0 & \text{no} \\ -1 & -ve \end{cases}$$

$$P_{xy} = \frac{\text{Cov}(x, y)}{\sigma_x \sigma_y}$$

SENA

## Probability .

\* Study of how likely an event is to occur.

\* note :

- Each trial is called an experiment
- ~ outcome - - a simple event
- Sum of every possible event is "sample space".

- قبل دراسة الاحداث سدرس السبل والمتغيرات؟ ~~لأنها~~ لأنها  
لبيان عن تكرر احداث.

## Permutations :

. على "ترتيب اجزاء العينة بترتيب معين" هذا الترتيب مهم اي ان

الحدث باكتئاف هو مفرغ.

$\Rightarrow abc \neq acb \neq bac \neq bca \neq cba \neq cab$

$${}^n P_r = \frac{n!}{(n-r)!}$$

$n \rightarrow$  total no. of elements

$r \rightarrow$  no. of desired elements.

## Combination "Combinatorics" :

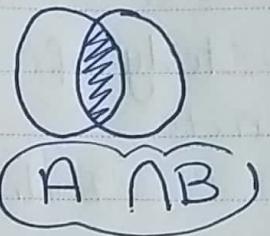
. لـ  $abc$ ,  $acb$  =  $acb$  =  $bac$  هذا الترتيب غير مهم.

$${}^n C_r = \frac{n!}{(n-r)! r!}$$

## \* Probability operations :

• union

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



• Complement:

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

• Multiplication rule :

\* both (A) & (B) occurs.

1) ~~Dependent~~ events :

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

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

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

(Sample space)

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

Law of Total probability :

$$P(A) = \sum_n P(A|B_n) \cdot P(B_n)$$

2) Independent events ;

$\rightarrow$  data set doesn't change.

$$P(A \cap B) = P(A) \cdot P(B)$$

$$P(A|B) = P(A)$$

$$P(B|A) = P(B)$$

\* Bayes Theorem:

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

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

$P(A)$   $\Rightarrow$  Total probability.

$$P(A) = P(A|B) \cdot P(B) + P(A|C) \cdot P(C) + \dots$$

$$P(A|B)$$

ما العبرة أن يكون  $B$  بعلبة  
إذا كان  $A$  مصنوع

\* Sampling :

$\hookrightarrow$  The selection of a subset of individuals from the population to estimate the characteristics of the population.

عليك في ذلك من المهم أن تستطيع تحديد خصائص هذا المربع.  
هناك نوع كثيرة للSampling.

- Random Sampling
- Stratified
- systematic
- cluster

## \* Random Variable :

→ variable whose value is determined by a random experiment.

- مثل الاجزء الخطي فإن المتغير ( $X$ ) يتم تحديده من خلال معادلة ( $f(x)$ )  
لذا هنا المتغير يحدده في هذه 形式 بشكل عشوائي كل مرة لذه ينبع من اجراء  
تجربة (mapping) من اكسل (variable) إلى قيم (function).

- عند اجراء كل جزء يجدها نستطيع حساب متوسط القيم التي سترى  
ما هي متوسط هذه القيم من ( $1, 2, \dots, 4$ ) فانا اريد ان احسب المتوسط المتوقع  
من هذه التجربة وصياغة بقرينة الwartungswert Expected value

## \* Expected Value :

→ It's the mean of a random variable

$$E(X) = \mu$$

- كما عرفنا بأن المتوسط ( $\bar{X} = \frac{\sum X_i}{N}$ ) ← وهذا يتم تطبيقه في حالة انى  
اجريت الاختبار وفرجت بالنتائج وبالنهاية اريد ان احسب المتوسط المتوقع الناتج.

- اما في حالة ( $E(X) = \mu$ ) فان هذا هو ال (empirical) او المتوقع من خلال  
الحسابات.

- ولكن حسب او متوقع عادة ( $E(X)$ ) فانا نستلزم ما يعني بالـ  
probability distribution ← وهو عبارة عن جدول ذو زوج

## \* Probability Distribution :

القائمة على الأحداث الممكنة  
Table or Formula that lists the probabilities for each outcome of a random variable ( $X$ ).

أى :  $x$  يمثل أحد أحداث بياناتي يتم فيها درج كل من العناصر طبقاً لـ  $P(x)$ .

مثال : - في حالة تossing على 3 مرات ، ما أحتمالية أن يكون نتائج (HHH) ؟

| $X$    | 0             | 1             | 2             | 3             |
|--------|---------------|---------------|---------------|---------------|
| $P(X)$ | $\frac{1}{8}$ | $\frac{3}{8}$ | $\frac{3}{8}$ | $\frac{1}{8}$ |

probability dist.

$X = 3$   $\rightarrow$  أحتمالية الحصول على 3 نتائج (HHH)

$P(X) = \frac{1}{8}$   $\rightarrow$  تساوى  $(\frac{1}{8})$

لذا  $\{$   $\begin{matrix} HHH & \rightarrow 3 \text{ heads} \\ HTT & \rightarrow \frac{1}{8} \\ TTH & \rightarrow \frac{1}{8} \end{matrix} \}$   
وليس كل الحالات  $\{$   $\begin{matrix} HHH & \rightarrow 3 \text{ heads} \\ HTT & \rightarrow \frac{1}{8} \\ TTH & \rightarrow \frac{1}{8} \\ TTT & \rightarrow \frac{1}{8} \end{matrix} \}$

هناك نوعان، تبيان للـ probability dist.

العامل منها :-

- Discrete probability distribution. مع الأحداث المتماثلة
- Continuous مع الأحداث المتصلة

### ① Discrete probability distribution :

هذا تكون الأحداث متماثلة "R, T" يمكن لها  $\rightarrow$  القائمة ( $H, T$ )

هذا توزيعات كثيرة تسمى  $\rightarrow$  discrete distribution و  $\rightarrow$  سهل و عملي

لذلك  $\rightarrow$  random variable

• uniform dist.

• Binomial  $\rightarrow$

• Poisson  $\rightarrow$

• Geometric  $\rightarrow$

• hyper geometric dist.

كل واحد فيه له تطبيق  
واستخدام معن

الآن  $\rightarrow$  الآن

probability mass function

$P(X) \Rightarrow PMF$

- Expected Value & Variance of discrete random Variables:

1) Expected Value: it's the theoretical mean :

$$E(X) = \mu$$

$$\mu = E(X) = \frac{\sum x_i \cdot n_i}{N} \rightarrow P_i$$

$n_i \Rightarrow$  عدد نتائج (أي)  $\Sigma$   
 $N \Rightarrow$  عدد النتائج  $\Sigma$

$$\therefore \mu = E(X) = \sum x_i \cdot P_i$$

notes:

$$\mu = E(\text{mean}) \rightarrow E(m^2) = \sum m^2 \cdot P_i$$

$$E(X-Z) = E(X) - E(Z)$$

علاقة

2) Variance :

$$\sigma^2 = E(X^2) - \mu^2$$

$$E(X^2) = \sum x_i^2 \cdot P_i(x)$$

$$E(\text{مربع}) = \sum (\text{مربع}) \cdot P(\text{مربع})$$

· أنواع الخوارزميات .

## ① Bernoulli distribution :

· عبارة عن (discrete prob. dist of a random variable) (X)

~~Success~~ = 1 &

له قيمتان فقط "نجاح وفشل"

Success :  $X=1$

Fail :  $X=0$

حالات

· عن رمي عملة، فإذا باعول "سرواها" بـ "باقطل بنفسه" لو ظهر (head) فـ "أنا

فictor (success) وإذا ظهر tail فـ "أنا fail".

· عن القاتل يجيء زر فـ "أنا أعتبر" إذا حـ 2 الرقم 5 "فـ أنا يعتبر" success و إذا ظهر غير ذلك fail

SENA

$$P(\text{success}) = P, \quad x = 1$$

$$P(\text{fail}) = 1 - P, \quad x = 0$$

$$(PMF) \quad P(X) = P^x (1-P)^{1-x}$$

$$\text{So; } P(X=1) = P^1 = P \quad \left. \begin{array}{l} \text{هذا هو قانون الاحتمالات المتعاقبة} \\ \text{وهو يعادل بولينولي Bernoulli} \\ \text{وهو يعادل طبقات Poisson} \end{array} \right\}$$

## ② Binomial distribution

لنفس السياق ولكن هنا أضفتنا بعد (n) من المعاولات .  
 مثلاً : في العادة عادة ( .. ) ما المعاولة عن تكون head في (L<sub>n</sub> )  
 هنا نستخرج مبدأ الـ Combination أو التوافق ، وهذا بسبب التكرار .

$$P(x) = {}^n C_r \text{ bernall; } = \binom{n}{r} \text{ bernall;}$$

$$(PMF) \quad P(x) = \frac{n!}{x!(n-x)!} P^x (1-P)^{n-x}$$

$$\mu = nP \quad \sigma^2 = np(1-p)$$

Ex:

A balanced six-sided die is rolled (3) times; what's the probability a (5) comes exactly twice?

$$\text{Success} = 5 \quad \therefore p(\text{success}) = \frac{1}{6} \quad 1 - P = \frac{5}{6}$$

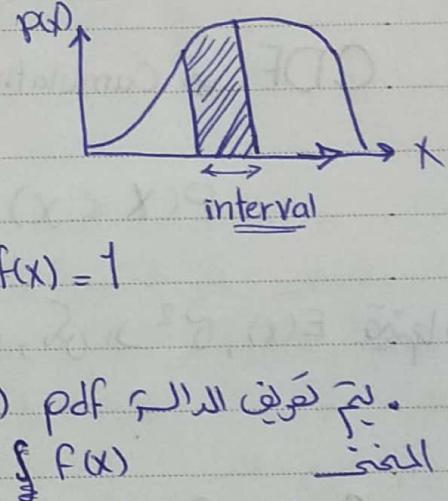
$$n=3 \quad x=2 \quad \therefore P(x=2) = \frac{3!}{2!(3-2)!} \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^{3-2} = 0.0694$$

## ② Continuous random distribution:

مُعَدِّل الدَّاتِ *Continuous data* مثلاً "الاطفال ولادون والاعمار، درجات الحرارة وغيرها".

→ Continuous random variables can take an infinite number of possible values corresponding to every value in an interval.

الدالر رهنا تشي بار (probability density function) ده تقوم بحساب احتمال وقوع قيمة  $x$  متسقة.



- \* For any Continuous distribution:

- $$\therefore f(x) \geq 0 \text{ for all } (x)$$

- The area under the entire curve

equals one  $\sum(p_i) = 1$   $\int_{-\infty}^{\infty} f(x) dx = 1$

• يتم تعريف الدالة  $f(x)$  pdf كـ  $\int_{-\infty}^x f(x) dx$  يدعى المكثفه.

$\int f(x) dx$  المحتوى

للاحظ أن الافتراضات  $(P)$  تمثل الفترات  $\text{intervals}$  متساوية ولهم نفس الحجم والمسافة

لأن  $P(X) = 0.5$  يأوه هنف ونمايؤيد  $2 \leq X \leq 4$ .

• هناك نوعان مختلفان لـ *ستاتيك* في تطبيقات مختلفة، ولكن فقط ستدرس:

- . Uniform distribution.
  - . Normal ↗
  - exponential ↗

$$P_d f = P(X \geq x) = \int_a^b f(x) dx$$

## • Expected Value and Variance:

\* over an interval  $(a, b)$

$$\mu = E(X) = \int_a^b x \cdot f(x) d(x)$$

$$\sigma^2 = E(X^2) - \mu^2$$

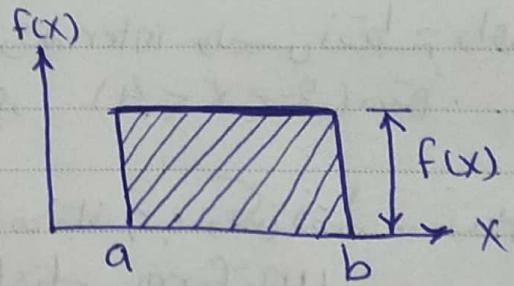
CDF  $\Rightarrow$  Cumulative density function:

$$P(X \leq x) = \int_0^x f(x) dx$$

لذا ان اد  $f(x)$  ستحتطف با متلاف المجموع، لكن اد  $E(x)$  قيمها واحدة ومتلطف با متلاف اد  $f(x)$ .

## ① Uniform distribution :

في هذا المكروبيج تكون كل الفئي (ها) لقص في هيئة المدوث ذو لها نفس الامثل (P).



$$\therefore \text{area} = (b-a) \cdot f(x) = 1$$

$S_{0j}$

$$F(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

$$X = U(a, b)$$

- Mean = median
- Variance

$$E(X) = \frac{a+b}{2}$$

$$\sigma^2 = \frac{1}{12} (b-a)^2$$

\* Suppose  $(X)$  is a random variable that has a uniform dist. with  $a = 200$ ,  $b = 250$ ,

$$\text{Find } P(X > 230) \quad \therefore f(x) = \begin{cases} \frac{1}{50} \\ 0.0 \end{cases}$$

$$\therefore P(X > 230) = \int_{230}^{250} \frac{1}{50} dx = \left( \frac{1}{50} \right) (20) = 0.4$$

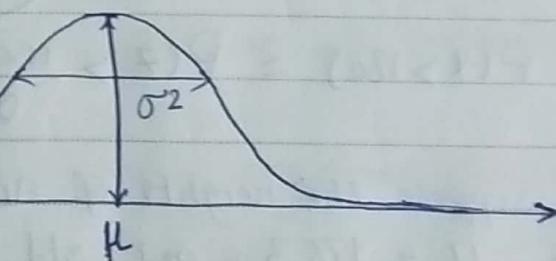
## ② Normal distribution:

وهي توزيع اساسي ينبع بشكل كبير عن التوزيع الطبيعي ل معظم الظواهر لدينا. ولو كان توزيعنا غير ذلك فإنه يتبع عن طريق استدلال Central Limit theorem (CLT).

• (Gaussian distribution) ليس بار.

• يأخذ شكل bell-shape.

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$



مما يدل على  
empirical

دائلاً (μ) يكون في المقدمة وهو الذي نعرفه ببيانات.

ـ يتم التعبير عنها كـ

$$X \sim N(\mu, \sigma^2)$$

هناك نوع خاص من الـ normal dist. يسمى من الـ standard deviation (Standard normal distribution) ويكون فيه  $\mu = 0$  و  $\sigma^2 = 1$ .

\* Standard normal distribution:

لهم احياناً بال (Z-distribution)

$$\left( \begin{array}{l} \mu = 0 \\ \sigma^2 = 1 \end{array} \right)$$

$$Z \sim N(0, 1)$$

كل رقم له تردد متساوى لستة (Table)، كذا في جدول الـ Z-table (Table) من  $X \sim N(\mu, \sigma^2)$  حيث قيمه متساوية في الـ standard dist.

\* Standardizing normal distribution:

لهم احياناً  $P(X)$  ترداد متساوية في فقرة متساوية.

$$P(X > 180)$$

لهم احياناً  $Z$ -dist يقوم بتحويلها إلى الـ Z-dist.

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

$$= P(X > 180) = P\left(Z > \frac{180 - \mu}{\sigma}\right) \text{ لهم احياناً } \mu, \sigma^2 \text{ فمودع.}$$

Ex: Suppose the heights of American males are normally dist. with  $\mu = 176.3$  cm and std.dev  $\sigma = 7.1$  cm.

What's the probability that a randomly selected adult is taller than 180 cm?

$$X \geq 180 \quad \mu = 176.3 \quad \sigma = 7.1$$

$$\therefore P(X > 180) = P\left(Z > \frac{180 - 176.3}{7.1}\right) = P(Z > 0.521) = \sqrt{\text{من قبائل الخروج}}$$

المعنى

\* what is the probability that the male is between 170 and 180 ?

$$= P(170 < X < 180) = P\left(\frac{170-173.1}{7.1} < Z < \frac{180-176.3}{7.1}\right)$$

$$= P(-0.887 < Z < 0.521) = \checkmark$$

\* what is the 90th percentile of the height of a male ?

$$P(Z) = 90\% = 0.9$$

$$\therefore \text{From tables: } Z = \checkmark \quad z = 1.28$$

&

$$\therefore Z = \frac{X - \mu}{\sigma} \Rightarrow X = Z\sigma + \mu = 1.28 \times 7.1 + 176.3 = \checkmark$$

\* Sampling distribution ?

→ we take a sample , and we use the sample mean ( $\bar{X}$ ) to estimate the population mean ( $\mu$ )

أيضاً في التوزيع العيني لها للتوزيع العيني.

أيضاً للتوزيع العيني.

\* We would like to give some measure of uncertainty associated with this value.

\* the value of the mean would vary from sample to sample.

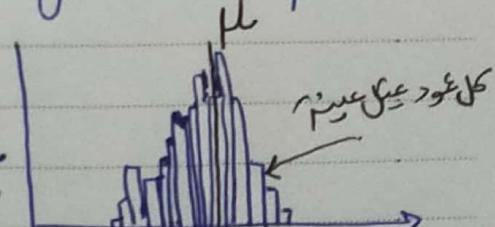
الآن أكثر من العينات وحساب المتوسطان

لذلك خاتماً نلاحظ أن هذه المتوسطات

population mean يشكل توزيعاً مركباً

(CLT)

وهي تسمى بـ



## • Central Limit theorem :

↳ in a Sampling distribution ; the samples means will be normally distributed about the population mean even if the population itself isn't normally distributed.

\* why this is important ?

↳ This give us a normal distribution even if we are Sampling from not normal distribution. So; we can use a well developed statistical inference procedures that are based on a normal distribution.

ANOVA (Hypothetical testing) التجربة المفروضة