Data and Sampling Distributions

مفهوم Population و Sample

- تعتبر العينة (**Sample**) مجموعة فرعية من مجموعة البيانات الكبيرة (population)، حيث تمثل مجموعة البيانات الكبيرة والمعرفة (population).
  - **مثال:** يمثل عدد موظفين شركة X حجم **Population** بينما يمثل عدد موظفين قسم IT حجم **sample.**
- يرمز لحجم population بالرمز N أما sample فيرمز لها بالرمز n.
- الأرقام التي نحصل عليها من population تسمى parameters.
- الأرقام التي نحصل عليها من sample تسمى statistics.
- في الغالب يصعب ملاحظة و حصر حجم Population بعكس sample حيث يمكن بسهولة حصر العدد وتحتاج وقت وتكلفة أقل.
  - **مثال:** تخيل أنك تريد دراسة آراء سكان مدينة الرياض عن تقنية معينة، سوف تقوم بجمع الآراء خلال حضورك في مؤتمر تقني في مدينة الرياض، بالتالي لن يكون من السهل حصر حجم Population وهو عدد الأشخاص التقنيين من سكان مدينة الرياض، حيث أن هناك الكثير من التقنيين لم يتواجدوا في هذا المؤتمر، أيضا لن تستطيع مقابلة جميع الحضور في هذا المؤتمر، لذا يمكن القول بأن الأشخاص الذين حصلت على إجاباتهم يمثلون sample.

إنشاء العينة (Sample)

هناك شرطين مهمين في إنشاء العينة (Sample):

- أن تكون العينة عشوائية **Randomness**

عندما يتم اختيار كل عنصر في العينة من population عن طريق الاحتمال (by chance).

- أن تكون العينة مُمثّلة **Representativeness**

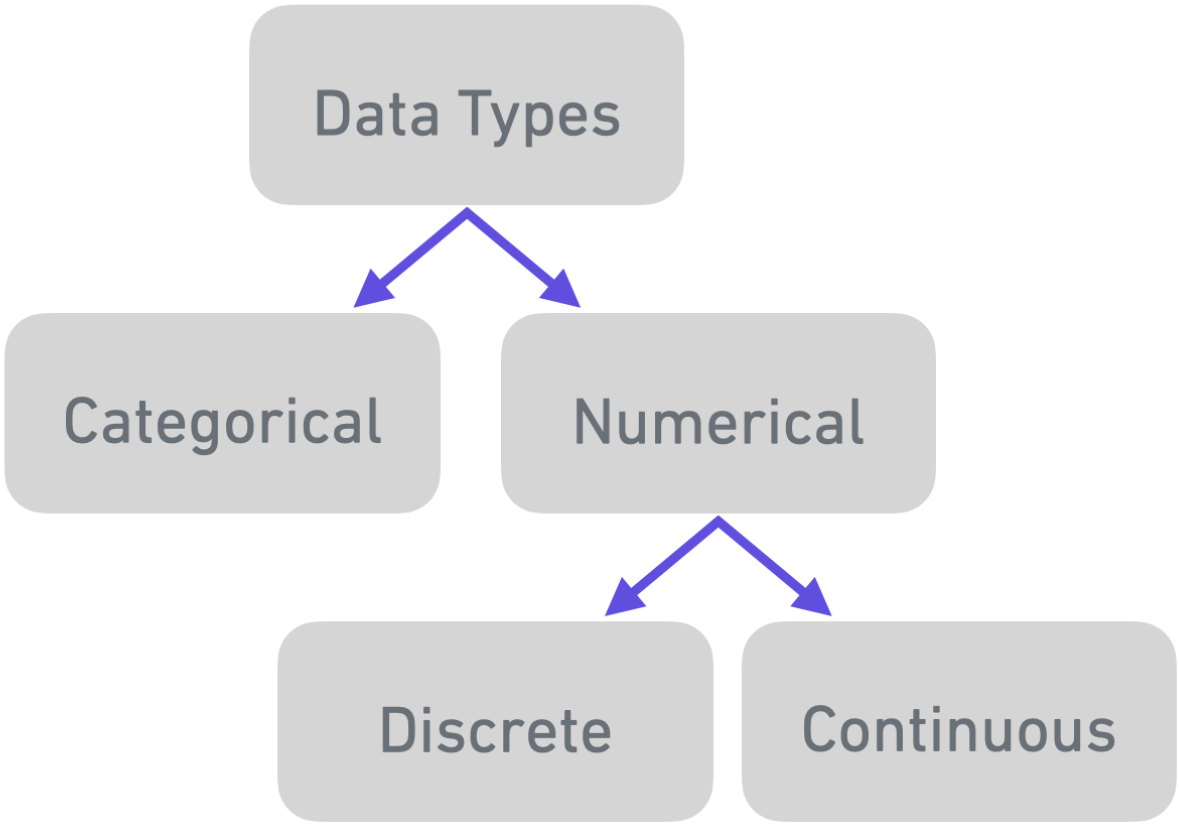
عندما يتم اختيار العينة من population بحيث تكون ممثلة لجميع عناصر population.

أنواع المتغيرات

يؤثر نوع المتغير على اختيار الطرق الإحصائية و نوع الرسومات البيانية (Statestical and Visulization Approaches).

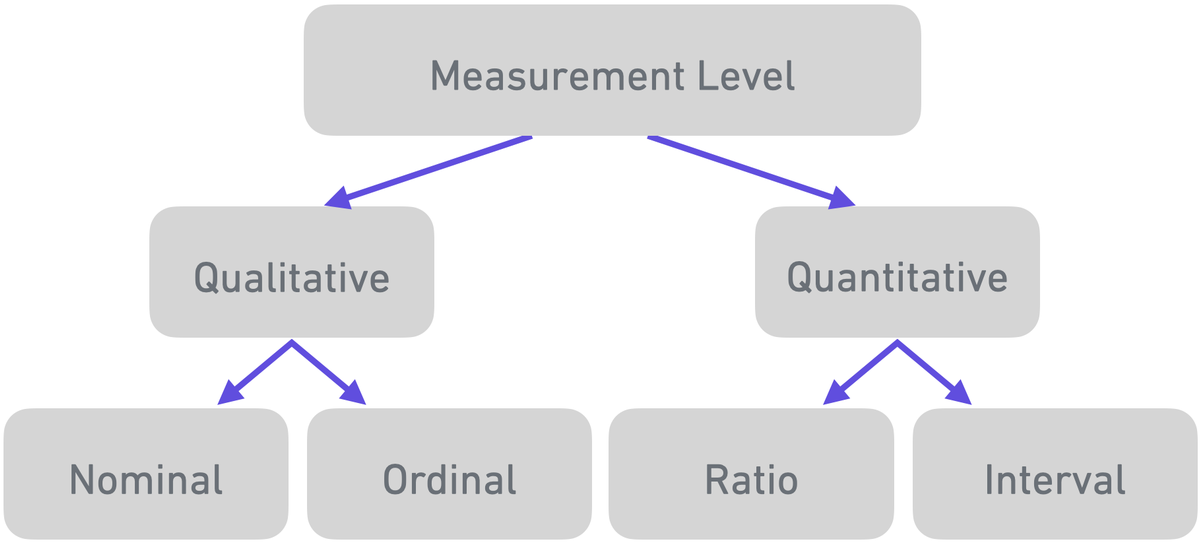
تختلف المتغيرات بناء على:

- **نوع البيانات (Data Types)**



- بيانات Categorical : تُصنف البيانات إلى مجموعات، وهي تمثل الأسئلة التي تكون الإجابة عليها بنعم أو لا، مثال: نتيجة الطالب (نجاح، رسوب).
- بيانات Numerical : تمثل البيانات الرقمية، ولها نوعين:
  - بيانات Discrete : تمثل الأرقام المنتهية finite numbers مثل: عدد الطلاب.
  - بيانات Continuous : تمثل الأرقام الغير المنتهية infinite numbers مثل: الوزن.

- **مستوى القياس (Measurement Level)**



- بيانات Qualitative
  - بيانات Nominal: هي البيانات الغير رقمية والتي ليس لها ترتيب معين، مثل: فصائل الدم، الجنس (ذكر، أنثى).
  - بيانات Ordinal: هي البيانات الغير رقمية والتي لها ترتيب معين، مثل: درجات الطلاب (A+, A, B+, B).
- بيانات Quantitative
  - بيانات Ratio : هي البيانات الرقمية والتي تحتوي true zero، مثال: المسافة والزمن.
  - بيانات Interval : هي البيانات الرقمية والتي لاتحتوي true zero، مثال: درجة الحرارة.

تقنيات تمثيل البيانات (Visualization Techniques)

نقوم بتحديد أفضل طريقة لتمثيل للمتغيرات بناء على Measurement Level و Data Types.

أولا: تمثيل المتغيرات من نوع Catagorical

نقوم باختيار أحد الطرق الأربعة التالية:

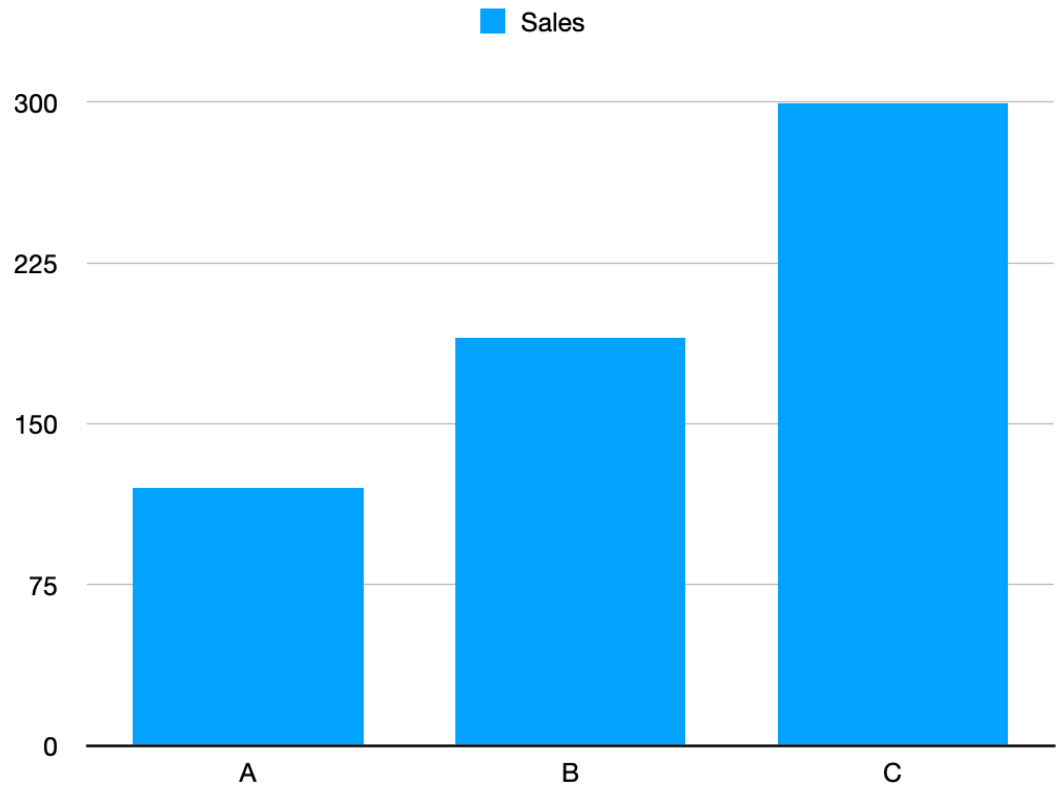
الطريقة الأولى: Frequency Distribution Table

Sales	Company
120	**A**
190	**B**
300	**C**

610	**Total**

الطريقة الثانية: Bar Chart أو Column Chart

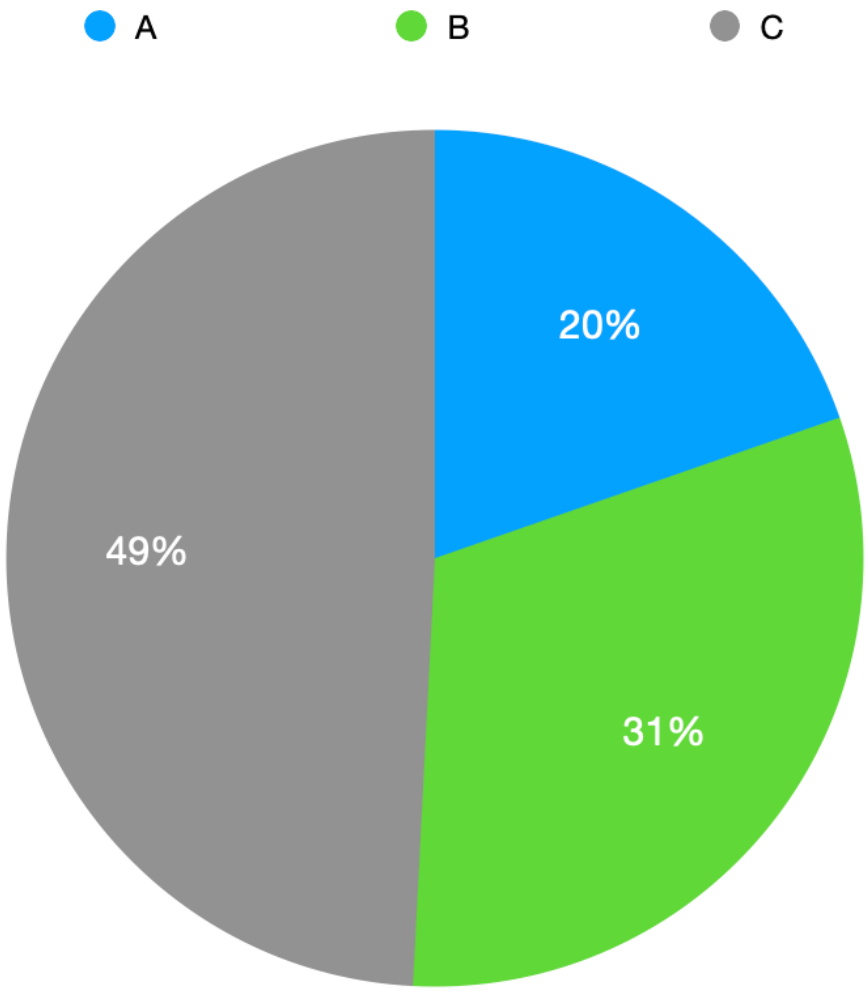
حيث يمثل العمود الواحد التكرار (Frequencey) لكل نوع.



النوع الثالث: Pie Chart

- لرسم هذا النوع من الرسومات نحتاج لحساب نسبة Relative Frequency لكل نوع.
- من أشهر الأمثلة لاستخدامات هذا النوع هو تمثيل Market Share.

Relative Frequency	Sales	Company
120/610 * 100 = 19.6	120	**A**
190/610 * 100 = 31.1	190	**B**
300/610 * 100 = 49.1	300	**C**
610/610 * 100 = 100	610	**Total**

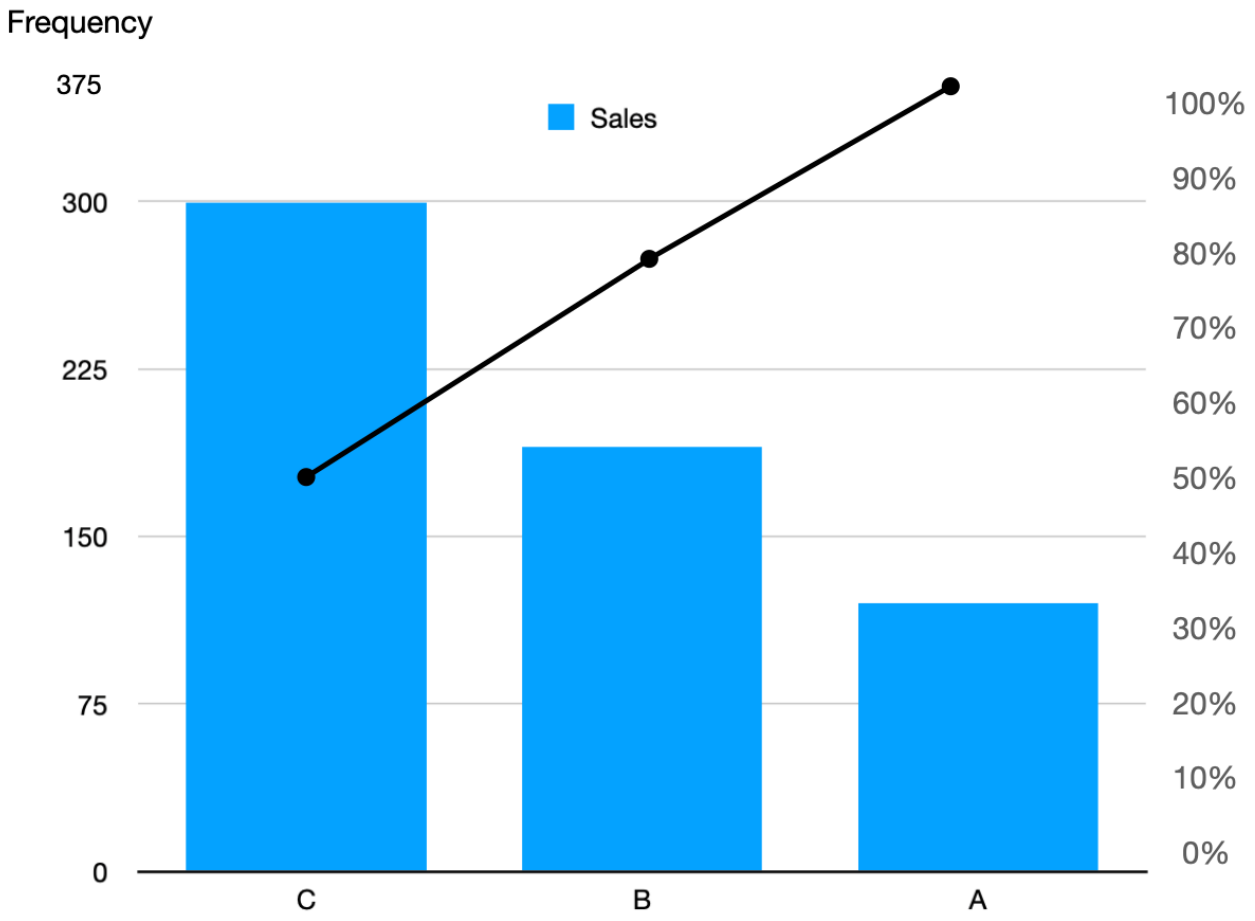


النوع الرابع: Pareto Chart

- هو عبارة عن نوع خاص من Bar Chart تم ترتيبه بشكل تنازلي.
- يحتوي curve يمثل Cumulative Frequency.
- يحتوي Cumulative Frequency وهي تمثل مجموع Relative Frequency.
- يتميز Pareto Chart بأنه يجمع بين مميزات Bar Chart و Pie Chart.

Cumulative Frequency	Relative Frequency	Sales	Company
49 %	49 %	300	**C**
80 %	31 %	190	**B**

100 %	20 %	120	**A**



ثانيا: تمثيل المتغيرات من نوع Numerical

- **تمثيل متغير واحد**

الطريقة الأولى: Frequency Distribution Table

لنفرض أن لدينا الجدول التالي ويمثل عدد الإصابات بمرض COVID-19.

Number of Cases
1
5
11
18
20
25

الطريقة الثانية: Histogram Chart

هل يمكن تمثيل البيانات السابقة برسم بياني؟

نلاحظ أننا لو أردنا تمثيل هذه البيانات برسم بياني فلن نحصل على معلومات مفيدة، لذا نحتاج لعمل عدد من الخطوات كالآتي:

أولا: حساب تكرار الحالات، عندها سوف نحصل على الجدول التالي:

Frequency	Number of Cases
1	1
1	5
1	11
1	18
1	20
1	25
6	**Total**

ثانيا: تلخيص البيانات

يعتبر الجدول السابق غير مناسب لأي تحليل، لذا الحل المناسب عند التعامل مع بيانات رقمية هو تقسيم هذه الأعداد في عدة مجموعات (intervals) ليسهل عمل Summary لهذه الأعداد وبالتالي تمثيلها.

- **لكن، كيف يتم اختيار هذه intervals ؟**

بشكل عام يتم تقسيمها من 5 إلى 20 مجموعة، لكن هذا قد يختلف من حالة لأخرى بحسب كمية البيانات.

يتم تقسيم البيانات عن طريق المعادلة:

```
(أعلى قيمة - أقل قيمة) / عدد المجموعات
```

في المثال السابق لو كنا نريد تقسيم البيانات لثلاثة مجموعات فسنحصل على (25 - 1)/3 = 8

> ***ملاحظة:*** *عند ظهور رقم يحتوي فواصل يمكننا عمل تقريب للعدد.*

Relative Frequency	Frequency	End	Start
2/6 = 0.33	2	9	1
0.17	1	17	9
0.50	3	25	17

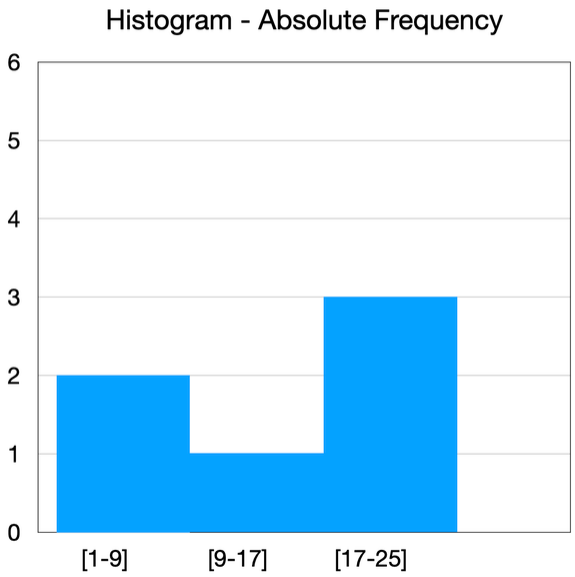
> ***ملاحظة:*** *يتم تقسيم الأرقام بحيث تكون* ***أكبر من*** *قيمة البداية و* ***أصغر من أو تساوي*** *قيمة النهاية.*

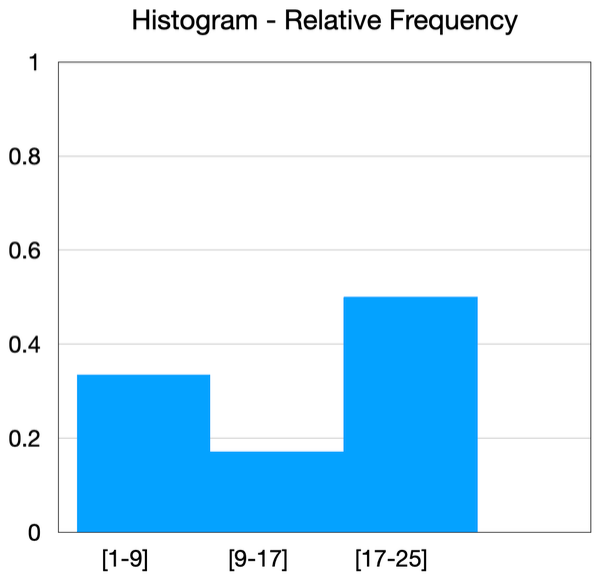
ثالثا: تمثيل البيانات

بعد عملنا Summary للبيانات السابقة، الان يمكننا رسم هذه البيانات في رسوم بيانية.

- رسم البيانات عن طريق Histogram Chart

يعتبر من أشهر الرسوم البيانية المستخدمة لتمثيل البيانات الرقمية.





في الشكل السابق نلاحظ أننا مثلنا Histogram بطرق مختلفة، أحدها باستخدام Absolute Frequency، والآخر باستخدام Relative Frequency، وعلى الرغم من أن الرسمتين متشابهة إلا أن كل واحدة تعطي معلومات مختلفة.

- **أي طريقة تعتبر الأفضل؟**
  - لايوجد طريقة أفضل من الأخرى، ولكن يعتمد اختيارنا للرسم البياني بحسب المعلومات التي نريد عرضها للمستخدم.

نلاحظ أن Histogram Chart يشابه Bar Chart لكن:

- كلا Vertical axis و Horisantel axis تمثل قيم رقمية.
- الأعمدة متصلة لأنها تمثل continouse interval بينما في Bar Chart منفصلة.

ملاحظة: يمكن تمثيل Histogram في مجموعات غير متساوية unequal intervals.

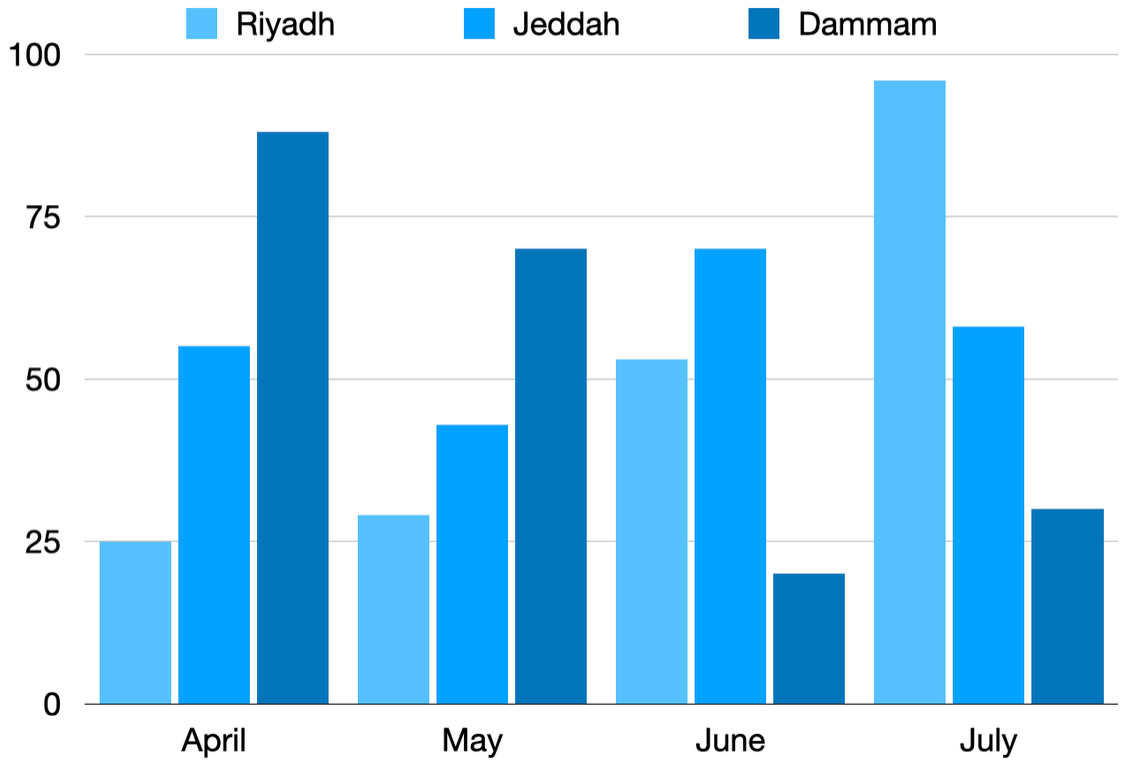
- **تمثيل متغيرين (العلاقة بين متغيرين)**

الطريقة الأولى: Cross Table ويسمى Contingency Table

لنفرض أن لدينا البيانات بالجدول التالي (عدد الحوادث المرورية الشهرية بالنسبة للمدن)

City	April	May	June	Total
Riyadh	25	29	53	107
Jeddah	55	43	70	168
Dammam	88	70	20	178
Total	168	142	143	453

- يوضح الجدول السابق العلاقة بين متغيرين من نوع Catagorical (المدن والأشهر)
- يوجد لدينا Sub-Total لكل صف (المجموع بالنسبة للمدن) وعمود (المجموع بالنسبة للشهور).
- أفضل طريقة لتمثيل هذا النوع هو الرسومات Side-by-side bar chart.

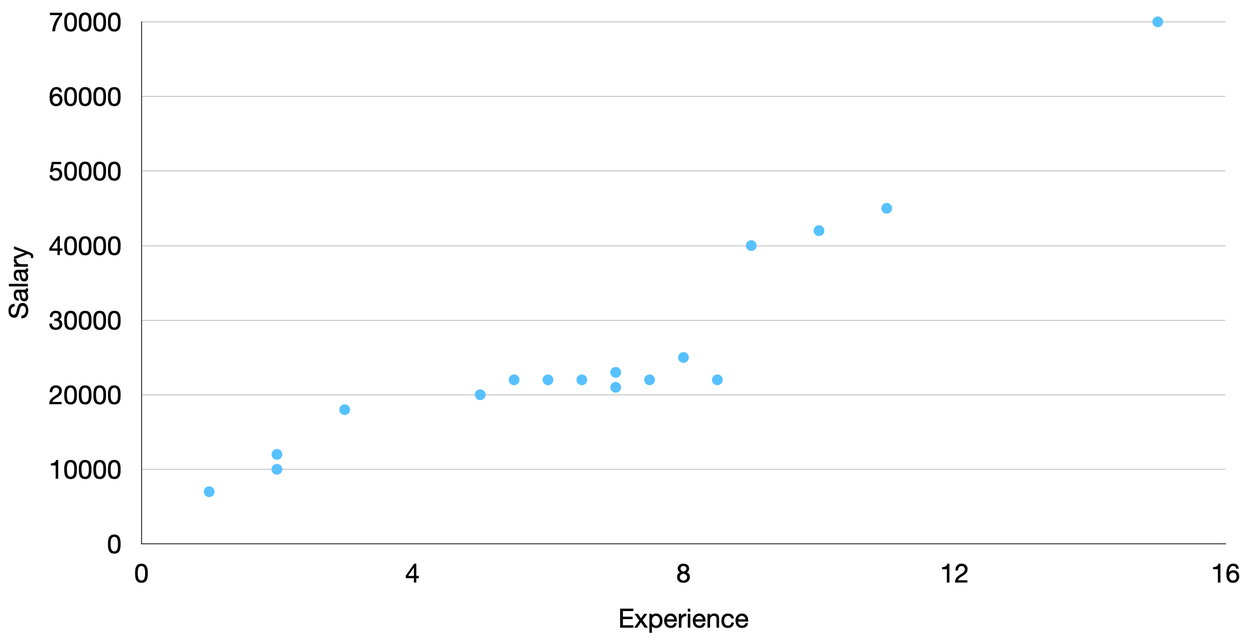


- الرسم البياني السابق يسهل علينا:
  - مقارنة عدد الحوادث بالنسبة للمدينة الواحدة في مختلف الشهور.
  - مقارنة عدد الحوادث في شهر معين بالنسبة لكل المدن.

لنفرض أن لدينا بيانات الموظفين كما في الجدول التالي، ونريد رسم بياني يوضح العلاقة بين الراتب وسنوات الخبرة.

Employee No.	Experience	Salary
1	1	7000
2	2	10000
3	2	12000
4	3	18000
5	5	20000
6	6	22000
7	6.5	22000
8	5.5	22000
9	7.5	22000
10	8.5	22000
11	7	21000
12	7	23000
13	8	25000
14	9	40000
15	10	42000
16	11	45000
17	15	70000

الطريقة الثانية: Scatter plot



- يتم استخدام هذا النوع للتعبير عن العلاقة بين متغيرين من نوع Numerical.
- من خلال الرسم يمكننا معرفة range أو النطاق الخاص بكل متغير.
- كل نقطة تمثل بيانات خاصة تخص (observation) معينة، في الرسم أعلاه تمثل الموظف.
- الهدف من هذا النوع معرفة توزيع البيانات أو تواجد trends (إذا زادت خبرة الموظف زاد الراتب).
- نلاحظ كثافة البيانات في المنتصف وهي تمثل متوسط البيانات (الخبرة (5 - 7)، الراتب (20,000 - 23,000)).
- يمكن من خلال الرسم استخراج القيم الشاذة outliers والتي تخالف بقية البيانات (مثال: موظف بخبرة أعلى من المتوسط وراتب أقل من المتوسط)

المقاييس

- قياس Central Tendency
- قياس عدم التماثل Asymmetry
- قياس انتشار البيانات How Data is Spread Out

أولا: قياس Central Tendency

يمكن قياس Central Tendency عن طريق ثلاثة مقاييس (Mean, Median, Mode).

أولا: المتوسط Mean ويسمى Average.

- يمثل حاصل جمع العناصر مقسومة على عدد العناصر، كما في المعادلة التالية:

Population Mean	Sample Mean
$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$
N = number of items in the population	n = number of items in the sample

Population Mean And Sample Mean (video lessons, examples, solutions)

نعبر عن Mean بالرمز x إذا كان في Sample والرمز m إذا كان في population.

عيوب هذا النوع من المقاييس:

- يؤثر عليه Outlier، لذا هو غير كاف للوصول للاستنتاجات.

- مثال: عدد إصابات COVID-19 في مدينتي الرياض و جده.

Jeddah	Riyadh	Day
98	100	1
152	150	2
140	145	3
125	130	4
125	500	5

عند حساب المتوسط لمدينة الرياض = 205، أما بالنسبة لمتوسط مدينة جدة = 128، على الرغم من تشابه الأعداد ولكن وجود القيم الشاذة Outliers في مدينة الرياض كان له تأثير في اختلاف المتوسط بين المدينتين.

ثانيا: الوسيط Median

يمثل الرقم الذي يتوسط مجموعة البيانات أوالأرقام المرتبة بشكل تصاعدي، ويمكن الوصول لهذا الرقم عن طريق المعادلة التالية: (n+1/2) حيث n يمثل عدد العناصر.

> ***ملاحظة:*** *إذا كان العدد زوجي نأخذ العددين ونقسمهم على 2.*

- في الجدول السابق، يمكن حساب الوسيط لمدينة الرياض = 145، أما الوسيط لمدينة جدة = 125.
- نلاحظ تقارب القيم بعكس قيم المتوسط، وهذا مايُميز الوسيط وهو عدم تأثره بوجود Outliers.

ثالثا: المنوال Mode

هي القيم التي يتكرر ظهورها من بين مجموعة البيانات ويمكن حساب mode على البيانات سواء numerical أو catagorical.

- في الجدول السابق، يمكن حساب المنوال لمدينة جدة = 125، أما بالنسبة لمدينة الرياض فلا يوجد لها منوال.
- في المثال: لايمكن القول بأن كل الأرقام الخمسة تمثل المنوال، لكن يمكن اختيار رقمين أو ثلاثة أرقام كحد أقصى.

ماهو أفضل مقياس من المقاييس السابقة؟

- لايوجد مقياس أفضل من الآخر، والطريقة الأفضل هي استخدام جميع هذه المقاييس في نفس الوقت.

ثانيًا: قياس عدم التماثل (Measure of Asymmetry) أو الانحراف (Skewness)

- لقياس عدم التماثل لابد من قياس الانحراف (Skewness).
- يعد الانحراف (Skewness) مؤشر على كثافة و تركُز البيانات في أحد الاتجاهات.

- يمكن قياس الانحراف (Skewness) عن طريق المعادلة التالية:

$$\tilde{\mu}_3 = \frac{\sum_i^N (X_i - \bar{X})^3}{(N-1) * \sigma^3}$$

$\tilde{\mu}_3$ = skewness

N = number of variables in the distribution

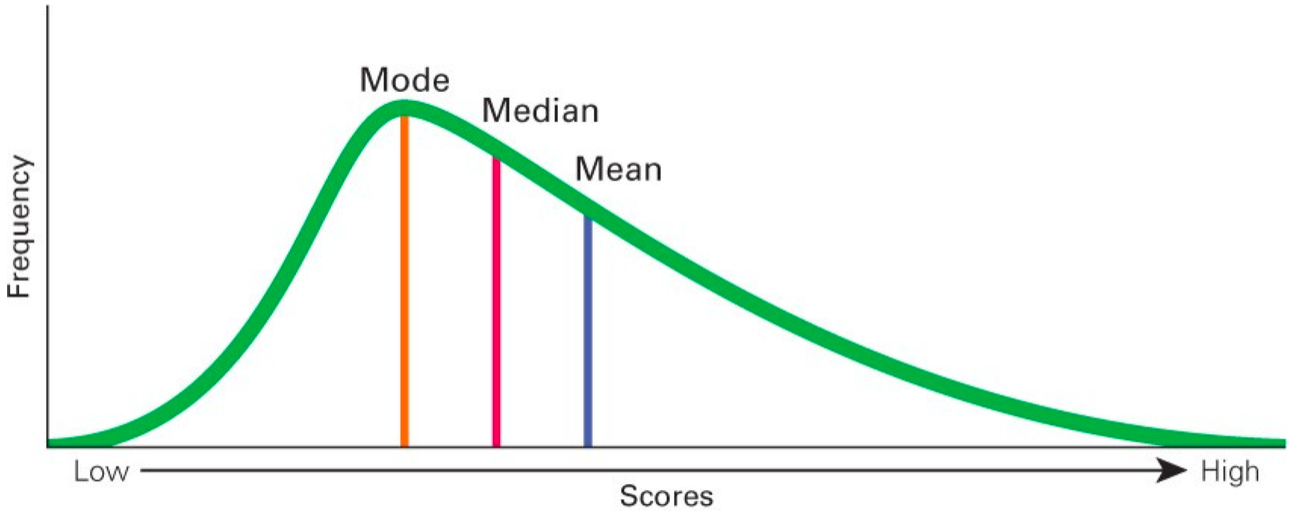
X_i = random variable

$\bar{X}$ = mean of the distribution

σ = standard deviation

descriptive statistics - How was the skewness formula made/derived? - Cross Validated

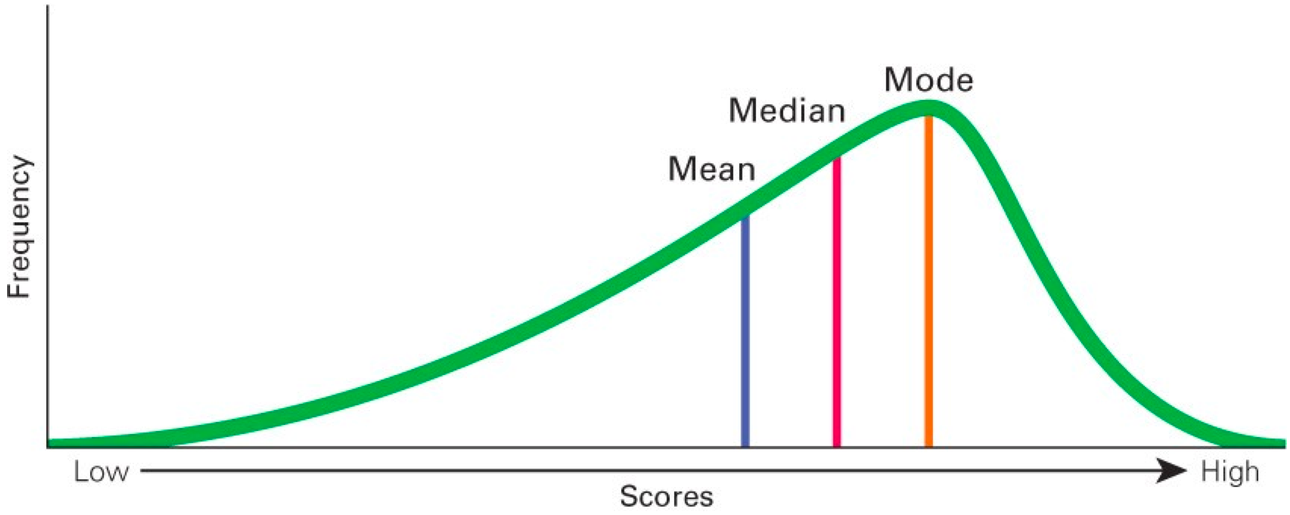
يمثل الشكل التالي(Right Skewness) أو مايسمى (Postive Skewness).



(a) Right-skewed distribution

- عندما نقول أن الرسم البياني Right Skewness فهذا يدل على أن Mean > Median.
- يمثل mode أعلى قيمة بالرسم البياني (قمة الرسم البياني).
- من الرسم البياني يمكن ملاحظة تركّز البيانات في الجهة اليسرى.
- مايهمنا من الرسم البياني هو ذيل الرسم البياني ويمثل Outliers حيث نلاحظ تركزها جهة اليمين.

يمثل الشكل التالي (Left Skewness) أو مايسمى (Negative Skewness)



(b) Left-skewed distribution

- عندما نقول أن الرسم البياني Left Skewness فهذا يدل على أن Mean < Median.
- من الرسم البياني يمكن ملاحظة أن ذيل الرسم البياني أو Outliers تتركز جهة اليسار.

يمثل الشكل التالي (Zero Skewness) أو مايسمى (No Skewness)

- عندما نقول أن الرسم البياني Zero Skewness فهذا يدل على أن `Mean = Median = Mode`.
- يعتبر توزيع البيانات (Symetrical Distribution).

ثالثًا: قياس انتشار البيانات How Data is Spread Out

- يطلق على هذه المقاييس(Measure of variability).
- يمكن حسابها عن طريق:
  - حساب Variance
  - حساب Standerd Deviation
  - حساب Cofficent of Variation.

أولا: حساب التباين Variance

- يقيس تشتت البيانات حول المتوسط (Mean).
- الحصول على نتيجة تباين قليلة يعني قرب البيانات من Mean والعكس صحيح.
- حساب المقاييس (Mean, Median, Mode) يختلف بين population و sample، وعلى هذا الأساس سوف نلاحظ اختلاف حساب Variance لكلا من population و sample كما في المعادلة التالية:

	Population	Sample
Variance	$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$	$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}$
Standard deviation	$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$	$S = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}}$

Variance and Standard Deviation-Definition, Formula, Relation and Example

نلاحظ وجود التربيع في معادلات Variance وذلك لعدة أسباب:

- حتى لاتظهر قيم سالبة لأن التشتت يعبر عن المسافة وهي لاتكون سالبة.
- تضخيم أثر الفوارق (large difference).

ثانيا: حساب الانحراف المعياري Standerd Deviation

يعتبر حساب التباين يعطي نتائج غير مناسبة وذلك بسبب تربيع القيم لذا فإن الحل البديل هو استخدام الانحراف المعياري.

ثالثا: حساب Cofficent of Variation.

يمثل نتيجة Standerd Deviation بالنسبة إلى Mean لذلك يسمى Relative Standerd Deviation ويمكن تمثيله بالمعادلة التالية:

$$\text{Coefficient of Variation (CV)} = \frac{\text{Standard Deviation (s)}}{\text{Sample Mean } (\bar{x})}$$

Coefficient of Variation | Overview, Formula & Examples | Study.com

يعتبر حساب Standerd deviation الطريقة الشائعة لحساب variability في مجموعة بيانات واحدة ولكن نحتاج لحساب cofficent of variation عند المقارنة بين مجموعتين أو أكثر من البيانات لأن المقارنة من خلال Standerd deviation غير مجدية.