

Data and Sampling Distributions

مفهوم Sample و Population

- تعتبر العينة (Sample) مجموعة فرعية من مجموعة البيانات الكبيرة (population).
- أما (Population) فهي تمثل مجموعة البيانات الكبيرة والمعرفة.
- 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، لنفترض أنك قمت 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):

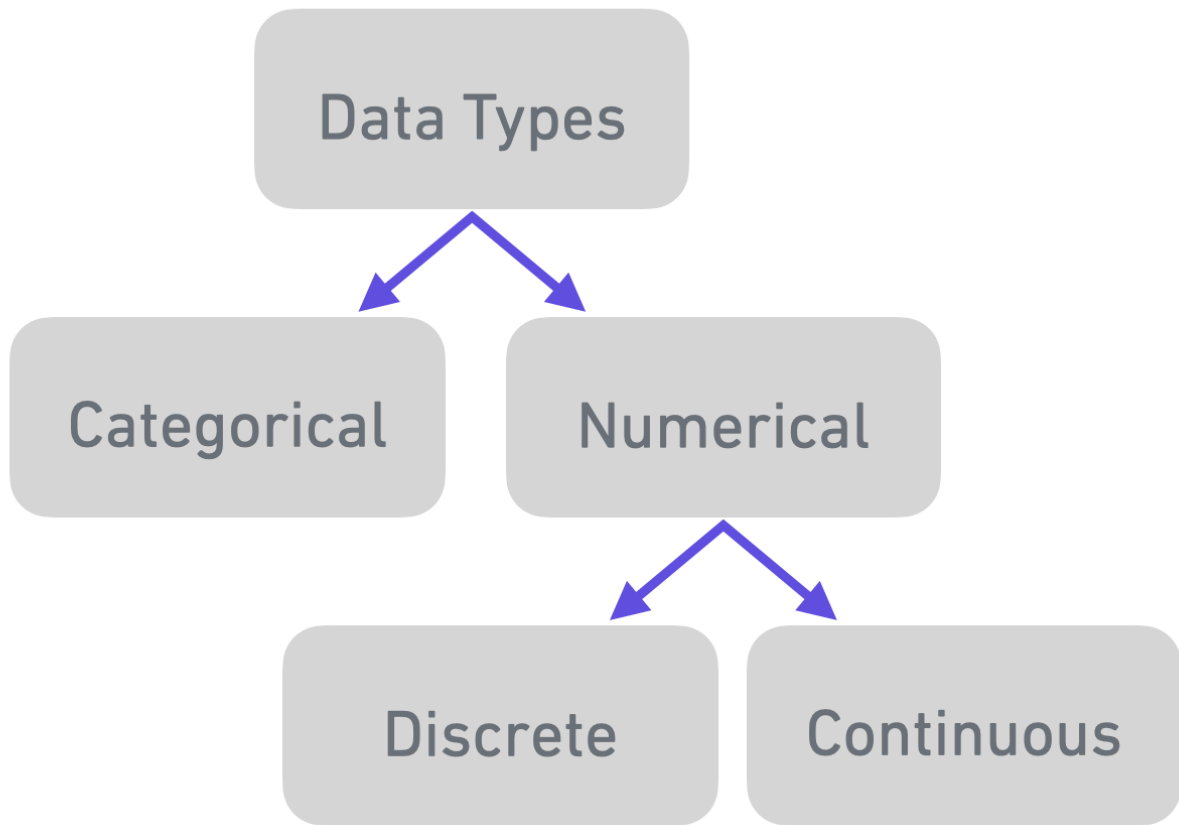
- أن تكون العينة عشوائية Randomness
- عندما يتم اختيار كل عنصر في العينة من population عن طريق الاحتمال (by chance).
- أن تكون العينة مُثَلَّة Representativeness
- عندما يتم اختيار العينة من population بحيث تكون ممثلة لجميع عناصر population.

أنواع المتغيرات

يؤثر نوع المتغير على اختيار الطرق الإحصائية و نوع الرسومات البيانية (Statistical and Visualization Approaches).

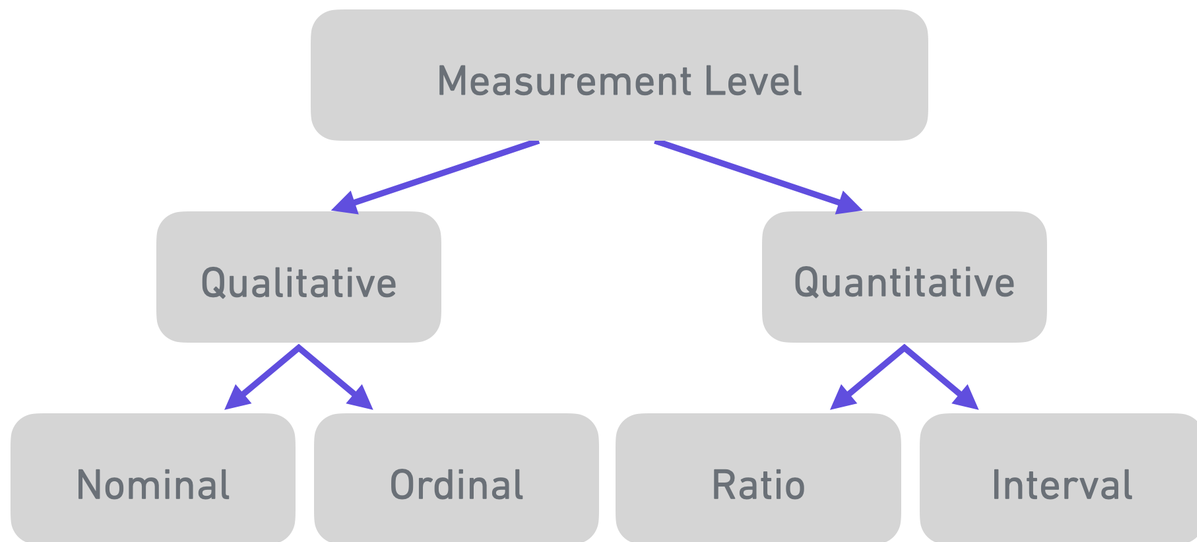
تختلف المتغيرات بناء على:

- 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 لا تحتوي true zero، مثال: درجة الحرارة.

تقنيات تمثيل البيانات (Visualization Techniques)

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

أولاً: تمثيل المتغيرات من نوع Categorical

نقوم باختيار أحد الطرق الأربعة التالية:

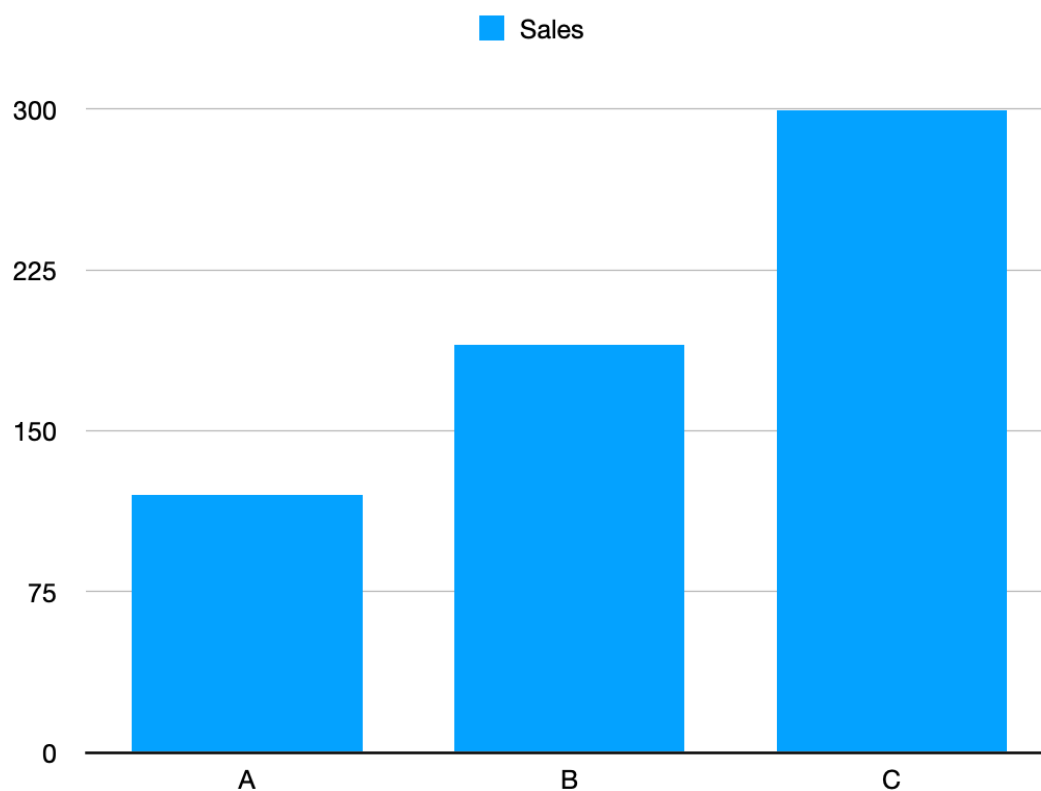
الطريقة الأولى: Frequency Distribution Table

Sales	Company
120	A
190	B
300	C

610

Total

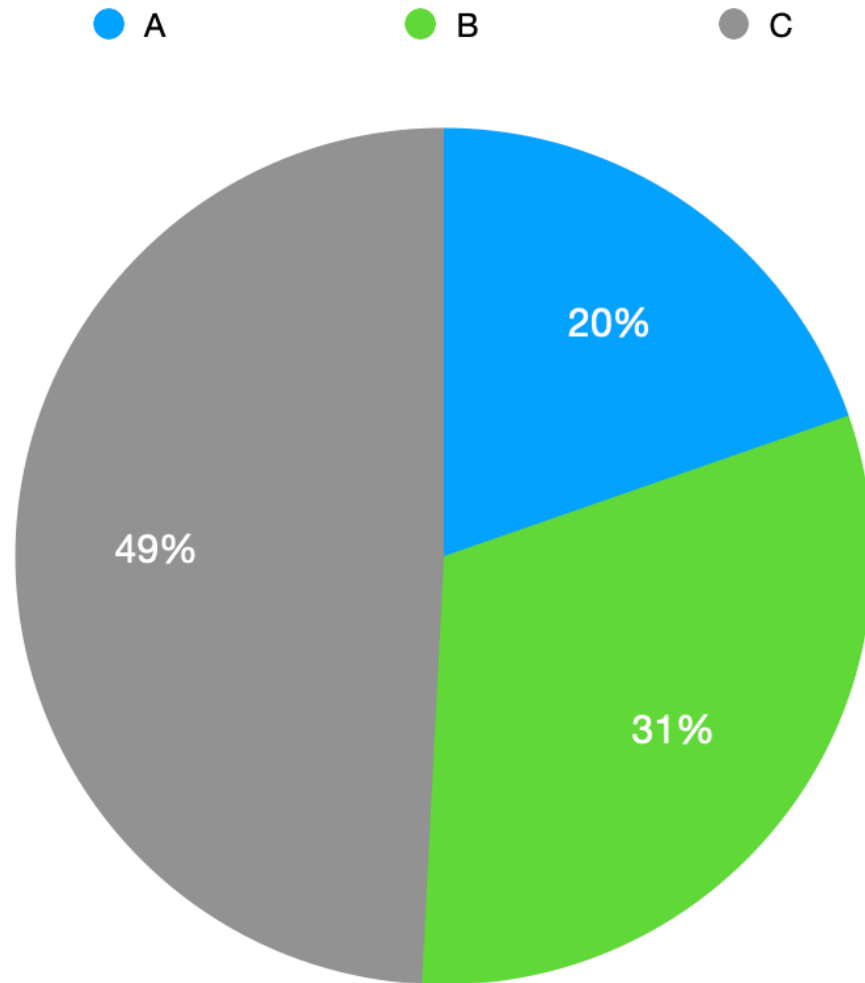
الطريقة الثانية: Column Chart أو Bar Chart
حيث يمثل العمود الواحد التكرار (Frequency) لكل نوع.



النوع الثالث: Pie Chart

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

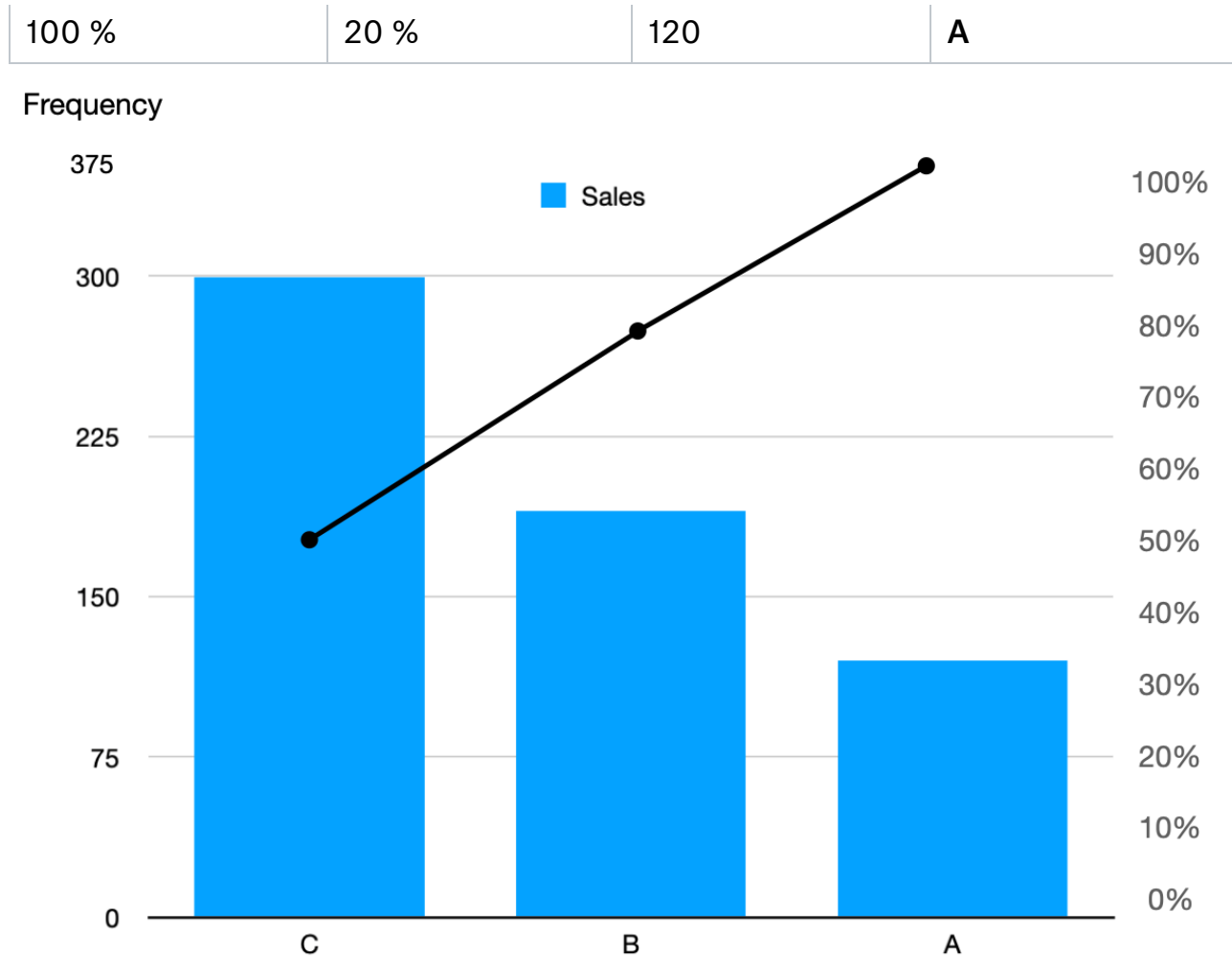
Relative Frequency	Sales	Company
$120/610 * 100 = 19.6$	120	A
$190/610 * 100 = 31.1$	190	B
$300/610 * 100 = 49.1$	300	C
$610/610 * 100 = 100$	610	Total



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

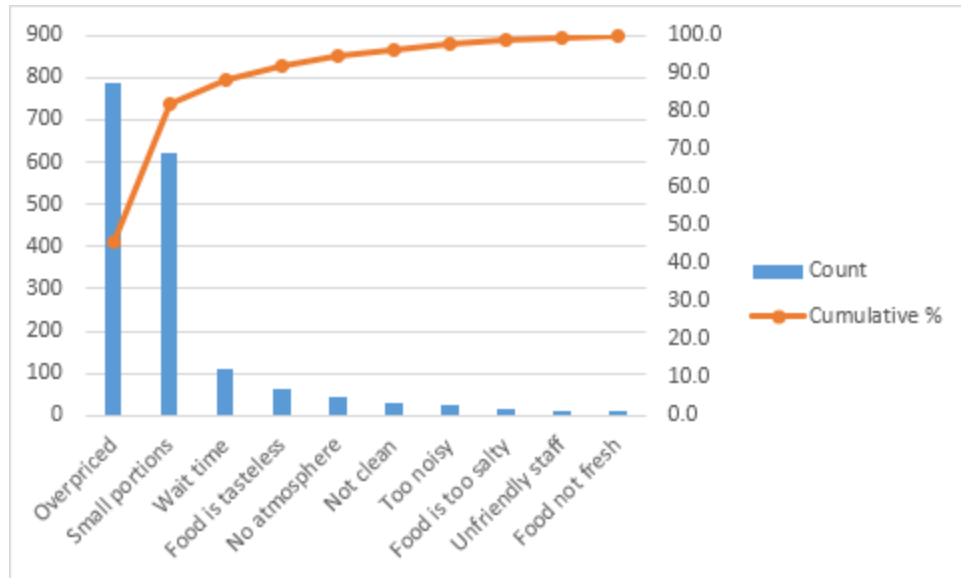


يقوم هذا الرسم على مبدأ **pareto** ويسمى أيضًا **Rule 20-80** وهو ينص على:

- أن 80% من النتائج تأتي من 20% من الإجراءات أو الأسباب .
- يستخدم لقياس الجودة، بحيث يقوم بتقسيم العمل لأجزاء صغيرة والتعرف على الجزء الأكثر تأثيرًا وبالتالي زيادته والتركيز عليه لرفع الجودة.

أمثلة:

- عند إصلاح 20% من مشاكل البرامج فهذا سوف يساعد في حل 80% من مشاكل الزبائن.
- يقوم 20% من الموظفين بإنجاز 80% من العمل.
- يمكن الحصول على 80% من المبيعات عن طريق 20% من الزبائن.



Pareto Chart

في الصورة بالأعلى، عند النظر للجانب الأيمن نلاحظ أن 80% من شكاوي زبائن المطاعم تأتي من 20% من الأسباب (سببين من أصل 10).

ثانياً: 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 كالاتي:

أولاً: 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

ثانياً: 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 $8 = 3 / (1 - 25)$

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

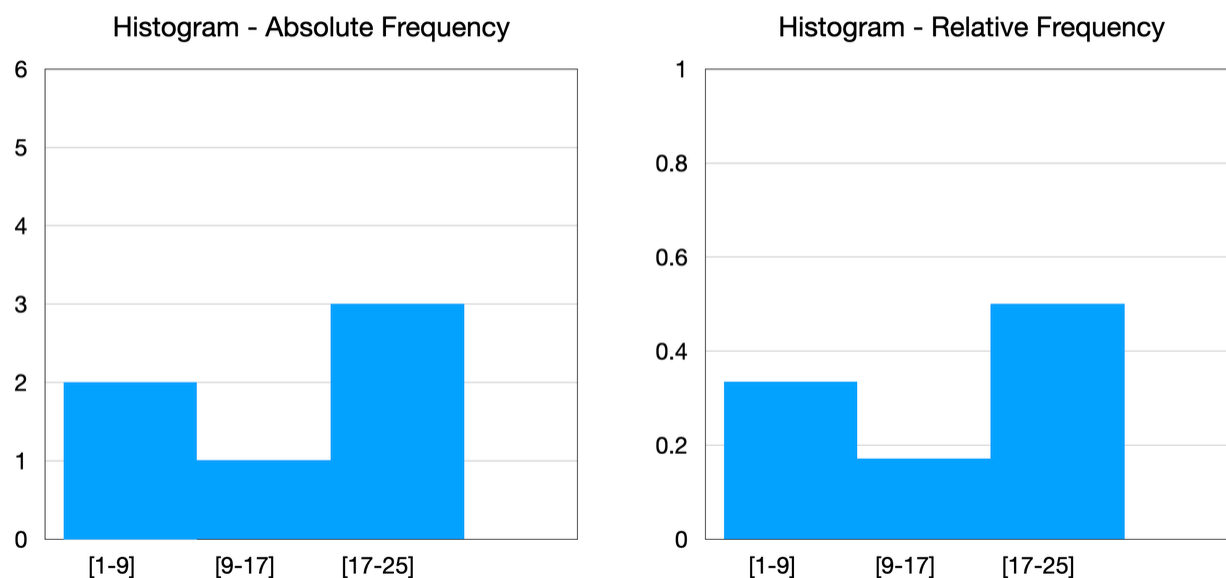
ملاحظة: يتم تقسيم الأرقام بحيث تكون أكبر من قيمة البداية وأصغر من أو تساوي قيمة النهاية.

ثالثاً: تمثيل البيانات

بعد عملنا Summary للبيانات السابقة، الآن 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 الرسميتين متشابهة إلا أن كل واحدة تعطي معلومات مختلفة.

- أي طريقة تعتبر الأفضل؟

○ لا يوجد 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 Horizontal axis تمثل قيم رقمية.
- الأعمدة متصلة لأنها تمثل continuous interval بينما في Bar Chart منفصلة.

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

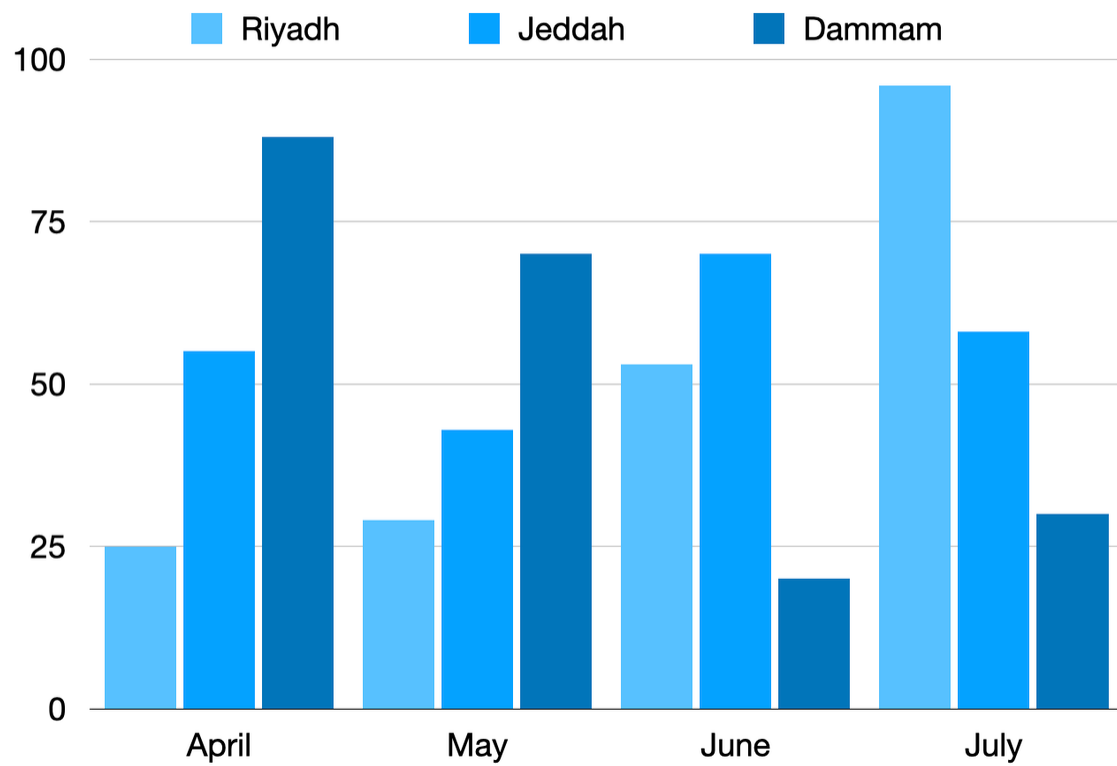
- تمثيل متغيرين (العلاقة بين متغيرين)

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.



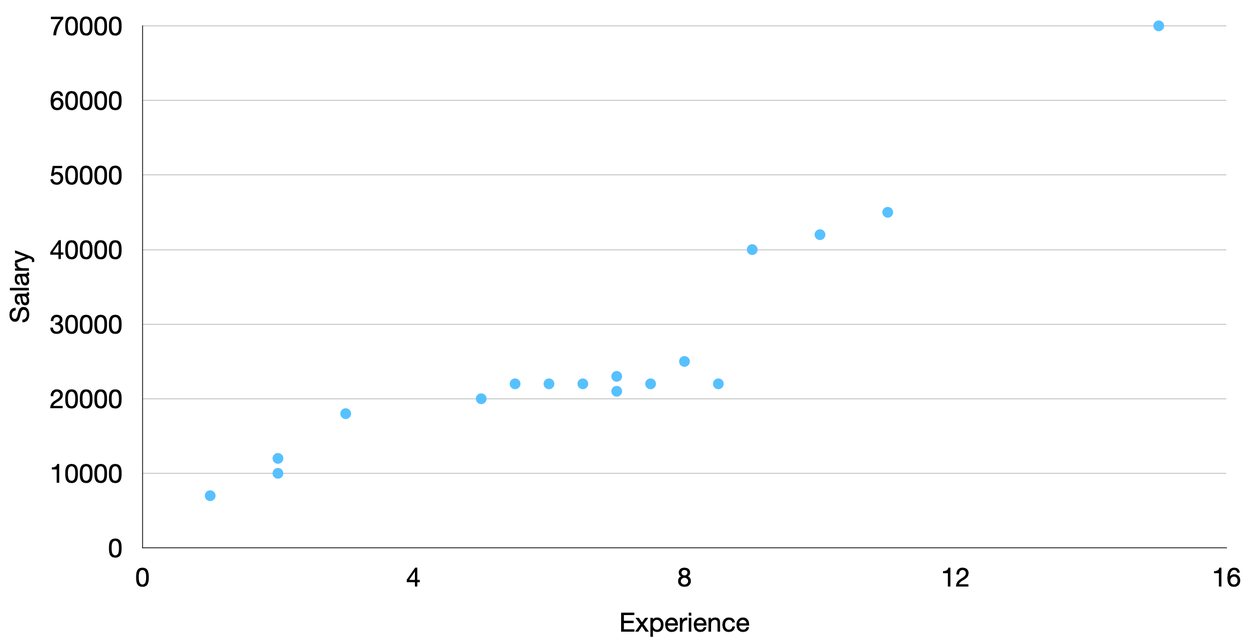
● الرسم البياني السابق يسهل علينا:

- 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

Scatter plot: الطريقة الثانية



- 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

Univariate Measures مقاييس

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

Relationship between variables مقاييس

- قياس التباين Covariance
- قياس معامل الارتباط Linear Correlation Coefficient

Univariate Measures مقاييس

أولاً: قياس Central Tendency

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

أولاً: المتوسط Mean ويسمى Average.

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

Population Mean	Sample Mean
$\mu = \frac{\sum_{i=1}^N x_i}{N}$	$\bar{X} = \frac{\sum_{i=1}^n x_i}{n}$
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 \bar{x} إذا كان في Sample والرمز μ 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 جدة.

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.

ثانياً: الوسيط Median

يمثل الرقم الذي يتوسط مجموعة البيانات أو الأرقام 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 ما يميز الوسيط وهو عدم تأثره بوجود Outliers.

ثالثاً: 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: لا يمكن القول بأن كل الأرقام الخمسة تمثل المنوال، لكن يمكن اختيار رقمين أو ثلاثة أرقام كحد أقصى.

ماهو أفضل مقياس من المقاييس السابقة؟

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

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

- لقياس عدم التماثل لابد من قياس الانحراف (Skewness).
- يعد الانحراف (Skewness) مؤشر على كثافة و تركّز 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

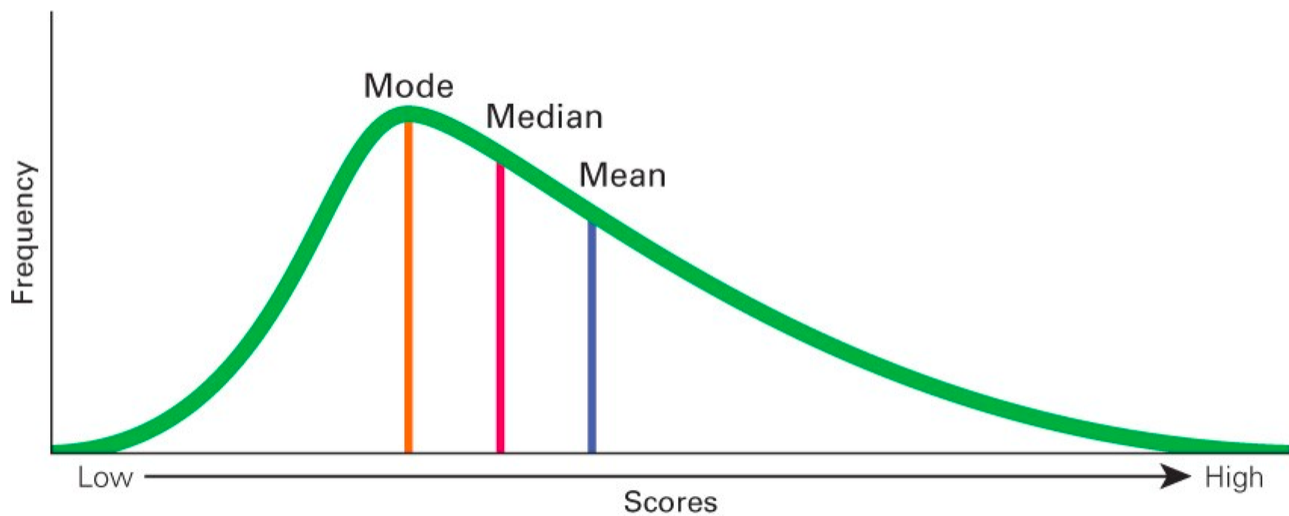
X_i = random variable

\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 ما يسمى (Postive Skewness).

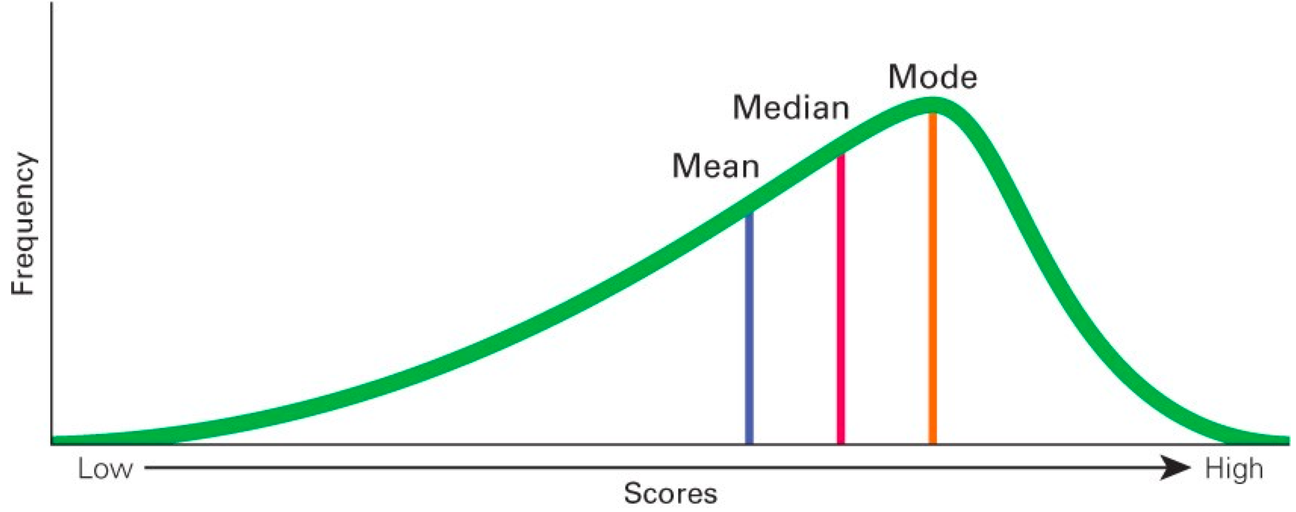


(a) Right-skewed distribution

- عندما نقول أن الرسم البياني Right Skewness فهذا يدل على أن $\text{Mean} > \text{Median}$.

- يمثل mode أعلى قيمة بالرسم البياني (قمة الرسم البياني).
- من الرسم البياني يمكن ملاحظة تركّز البيانات في الجهة اليسرى.
- ما يهمننا من الرسم البياني هو ذيل الرسم البياني ويمثل Outliers حيث نلاحظ تركّزها جهة اليمين.

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



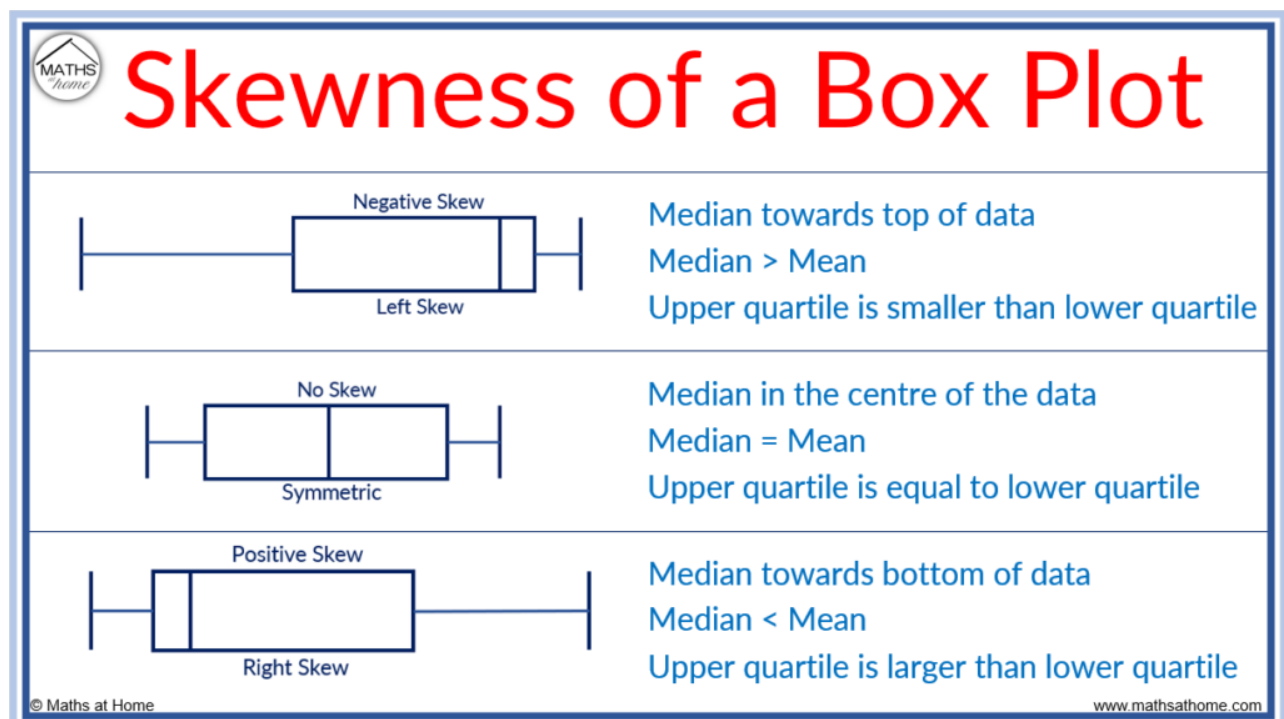
(b) Left-skewed distribution

- 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 $\text{Mean} = \text{Median} = \text{Mode}$.
- يعتبر توزيع البيانات (Symmetrical Distribution).



Normal Distribution

$$(\text{Quartile 3} - \text{Quartile 2}) = (\text{Quartile 2} - \text{Quartile 1})$$



Positive Skew

$$(\text{Quartile 3} - \text{Quartile 2}) > (\text{Quartile 2} - \text{Quartile 1})$$



Negative Skew

$$(\text{Quartile 3} - \text{Quartile 2}) < (\text{Quartile 2} - \text{Quartile 1})$$



Understanding and interpreting box plots | by Dayem Siddiqui | Dayem Siddiqui | Medium

How Data is Spread Out ثالثاً: قياس انتشار البيانات

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

أولاً: 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 لا تظهر قيم سالبة لأن التشتت يعبر عن المسافة وهي لا تكون سالبة.
- تضخيم أثر الفوارق (large difference).

ثانياً: حساب الانحراف المعياري Standard Deviation

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

ثالثاً: حساب Coficient of Variation.

يمثل نتيجة Standard Deviation بالنسبة إلى Mean لذلك يسمى Relative Standard 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 Standard deviation الطريقة الشائعة لحساب variability في مجموعة بيانات واحدة ولكن نحتاج لحساب coefficient of variation عند المقارنة بين مجموعتين أو أكثر من البيانات لأن المقارنة من خلال Standard deviation غير مجدية.

Standard deviation and coefficient of variation

Pizza price example

	NY Dollars		Pesos
\$	1.00	MXN	18.81
\$	2.00	MXN	37.62
\$	3.00	MXN	56.43
\$	3.00	MXN	56.43
\$	5.00	MXN	94.05
\$	6.00	MXN	112.86
\$	7.00	MXN	131.67
\$	8.00	MXN	150.48
\$	9.00	MXN	169.29
\$	11.00	MXN	206.91

	Dollars		Pesos
Mean	\$ 5.50	MXN	103.46
Sample variance	\$ ² 10.72	MXN ²	3793.69
Sample standard deviation	\$ 3.27	MXN	61.59
Sample coefficient of variation	0.60		0.60

- does not have a unit of measurement
- universal across datasets
- perfect for comparisons

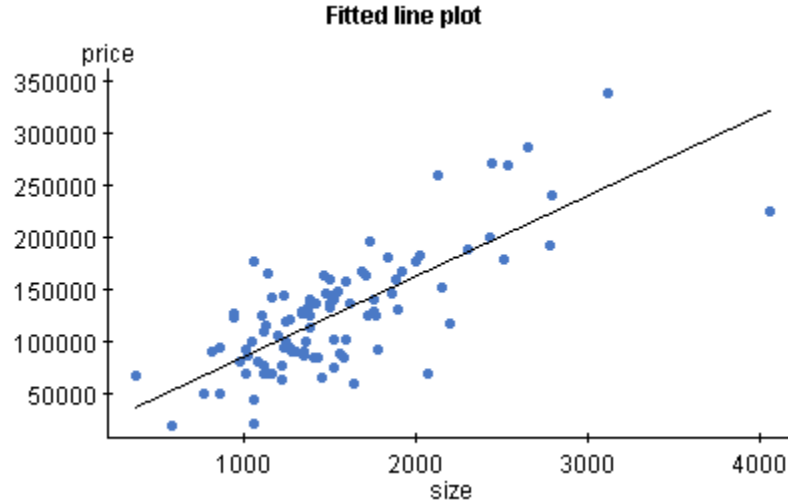
في المثال السابق، إذا قمنا بالمقارنة بين الأسعار في المجموعتين نلاحظ أن :

- نتيجة variance لا تمثل الأسعار الواقعية بسبب وجود التربيع.
- نتيجة Standard Deviation أكثر واقعية عند مقارنة الأسعار، لكن نلاحظ اختلاف Standard Deviation لنفس المطعم.
- لذلك الحل هو استخدام coefficient of variation حيث نلاحظ تماثل الأرقام وهي النتيجة الأقرب للواقع لأن مجموعتي البيانات لها نفس variability حيث انها. تشير لنفس المطعم ولكن باستخدام عملات مختلفة.
- نلاحظ أن coefficient of variation ليس له وحدة قياس.

مقاييس Relationship between variables

أولاً: قياس التباين Covariance

مثال : العلاقة بين أسعار المنازل وحجمها.



House Size Vs. Selling Price

نستطيع القول بأن المتغيرين correlated ويمكن حساب ذلك عن طريق قياس التباين Covariance.

Sample formula

$$S_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x}) * (y_i - \bar{y})}{n-1}$$

Population formula

$$\sigma_{xy} = \frac{\sum_{i=1}^N (x_i - \mu_x) * (y_i - \mu_y)}{N}$$

على عكس variance يمكن القول بأن نتيجة قياس التباين Covariance يمكن أن تكون:

- موجبة وهذا يعني بأن المتغيرين يتحركان بنفس الاتجاه.
مثال: العلاقة بين أسعار المنازل وحجمها، حيث نلاحظ بأن اتجاه العلاقة بالشكل التالي: (حجم المنزل يُحدد سعره).
- سالبة وهذا يعني بأن المتغيرين يتحركان باتجاهين متعاكسين.
مثال: مبيعات ice cream تزداد في فصل الصيف بينما مبيعات umbrellas تزداد في فصل الشتاء وهذا يعني زيادة المبيعات في النوع الأول تعني قلتها في النوع الثاني.
- صفر وهذا يعني بأن كل واحد متغير مستقل عن الآخر (independent).
مثال: العلاقة بين أسعار القهوة في السعودية وأسعار المنازل في الإمارات.

أحيانا يكون نتيجة قياس Covariance أرقام مختلفة في scale بشكل كبير مثل: 0.00039 أو 93,900,259 الحل: هو استخدام المقياس التالي:

ثانيًا: قياس معامل الارتباط Linear Correlation Coefficient

هو مقياس التباين Covariance ولكن تم تعديله بحيث يُوفر لنا النتائج بطريقة يسهل فهمها و تفسيرها.

$$\frac{Cov(x, y)}{Stdev(x) * Stdev(y)}$$



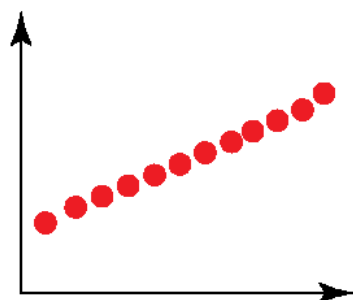
$$\frac{S_{xy}}{S_x S_y}$$



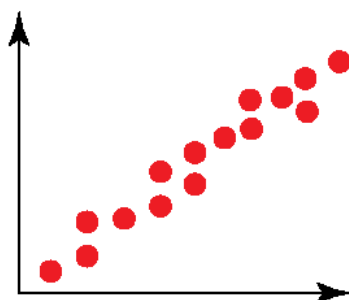
$$\frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Correlation Coefficient قيمة	Correlation Coefficient نوع
1	Perfect Positive Correlation
(1, 0.7)	Strong Positive Correlation
(0.6, 0.4)	Moderate Positive Correlation
(0.3, 0.1)	Weak Positive Correlation
0	No Correlation
-1	Perfect Negative Correlation
(-1, 0)	Imperfect Negative Correlation

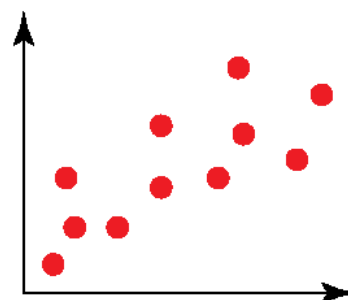
يمكننا اكتشاف **Correlation** من خلال النظر في الرسومات البيانية كالتالي:



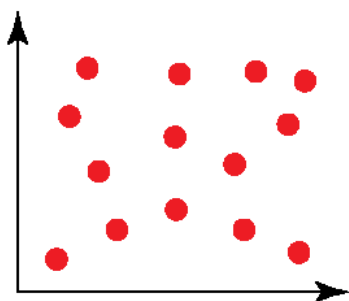
Perfect
Positive
Correlation



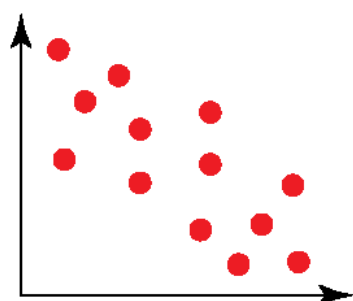
Strong
Positive
Correlation



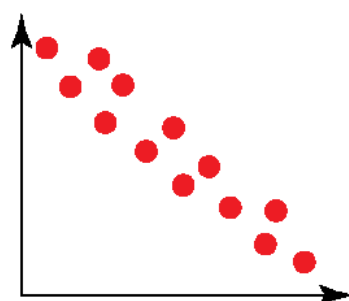
Weak
Positive
Correlation



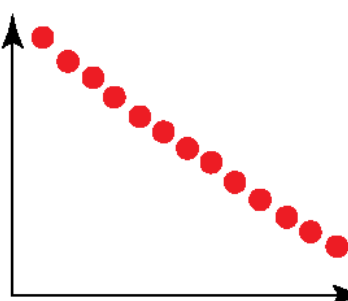
No
Correlation



Weak
Negative
Correlation



Strong
Negative
Correlation



Perfect
Negative
Correlation

Resources:

- Statistics for Data Science and Business Analysis, Packt Publishing, 2021.
 - <https://learning.oreilly.com/videos/statistics-for-data/9781789803259/>
- <https://hbrarabic.com/%D8%A7%D9%84%D9%85%D9%81%D8%A7%D9%87%D9%8A%D9%85-%D8%A7%D9%84%D8%A5%D8%AF%D8%A7%D8%B1%D9%8A%D8%A9/%D9%85%D8%A8%D8%AF%D8%A3-%D8%A8%D8%A7%D8%B1%D9%8A%D8%AA%D9%88/>