Statistics 1

By: Eng. Esraa Madhi
Nourah Almutlaq

Statistics is a branch of mathematics that involves the use of quantified models, summaries, and representations to examine and interpret a collection of data obtained from experiments or real-world observations. The primary advantage of employing statistics is its ability to transform complex data into information that is more comprehensible.

In the realm of Data Science, data processing stands as the cornerstone of any project. Discussing the extraction of insights from data is essentially a conversation about investigating the potentialities inherent in the data. Within Data Science, this exploration and investigation are known as Statistical Analysis.

Different Types of Analytics in Statistics

Analytics types	Definition
1. Descriptive Analytics – **What happened?**	It tells us what happened in the past and helps businesses understand how they are performing by providing context to help stakeholders interpret data.
2. Diagnostic Analytics – **Why did it happen?**	It goes beyond descriptive data to assist you in comprehending why something occurred in the past. (check correlation)
3. Predictive Analytics – **What is likely to happen?**	It forecasts what is likely to happen in the future and provides businesses with data-driven actionable insights. (Statistics is a building block of data science)
4. Prescriptive Analytics – **What should be done ?**	It makes recommendations for actions that will capitalise on the predictions and guide the potential actions toward a solution. Prescriptive analytics is the final and most advanced level of analytics.

Descriptive Analytics - Statistical Concepts for Data Scientists

1. Population and sample

- 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.

لإنشاء العينة (Sample)، هناك شرطين مهمين في إنشاء العينة :

- 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.

---

2. Measures of central tendency - Descriptive Analysis - Univariate Measures

https://youtu.be/0-C_H5J3uJU

https://youtu.be/0-C_H5J3uJU

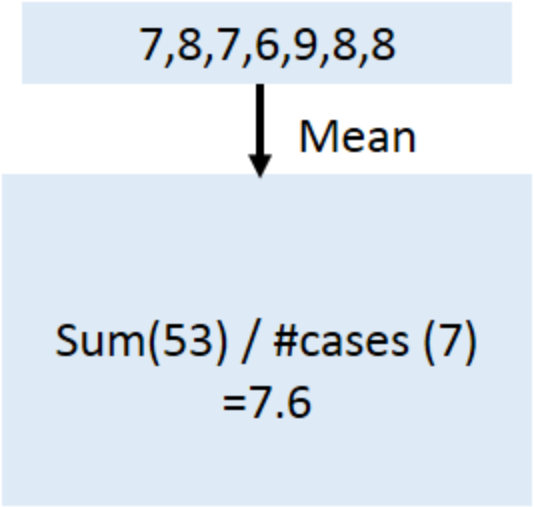
أولا: المتوسط Mean ويسمى Average.

- 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.*

1, 3, 3, **6**, 7, 8, 9

Median = **6**

1, 2, 3, **4**, **5**, 6, 8, 9

Median = (4 + 5) ÷ 2

= **4.5**

- 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

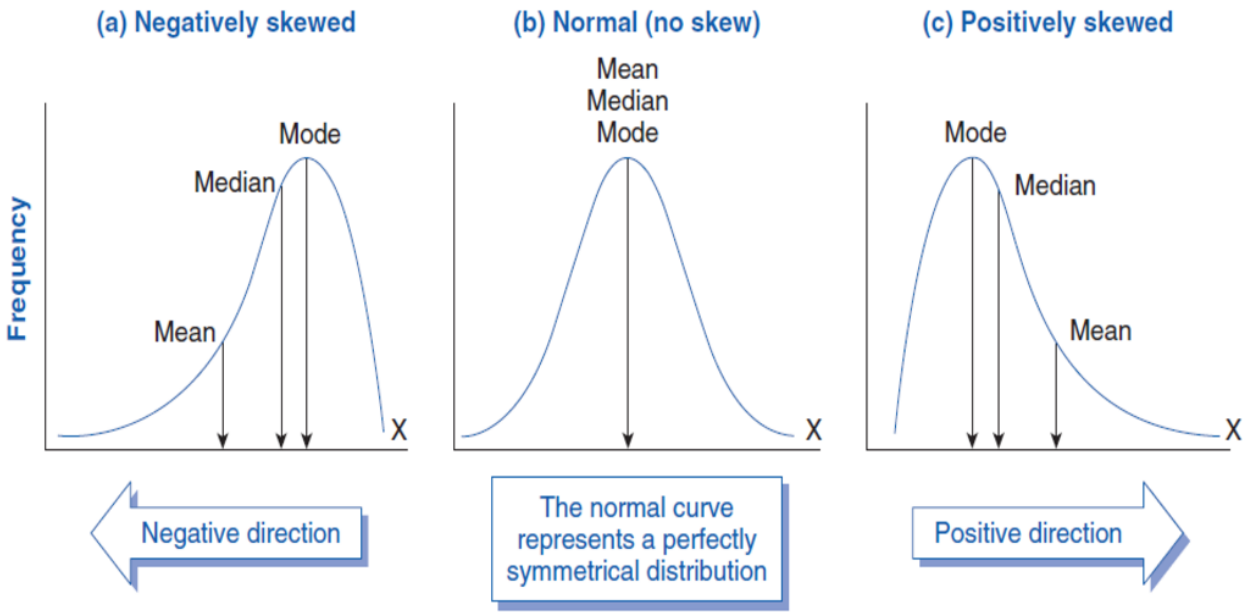
هي القيم التي يتكرر ظهورها من بين مجموعة البيانات ويمكن حساب mode على البيانات سواء numerical أو catagorical.

- 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.

رابعا: الانحراف (Skewness)

Skewness is a metric for symmetry, or more specifically, the lack of it. If a distribution, or data collection, looks the same to the left and right of the centre point, it is said to be symmetric.

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



الشكل (Right Skewness) أو مايسمى (Positive Skewness).

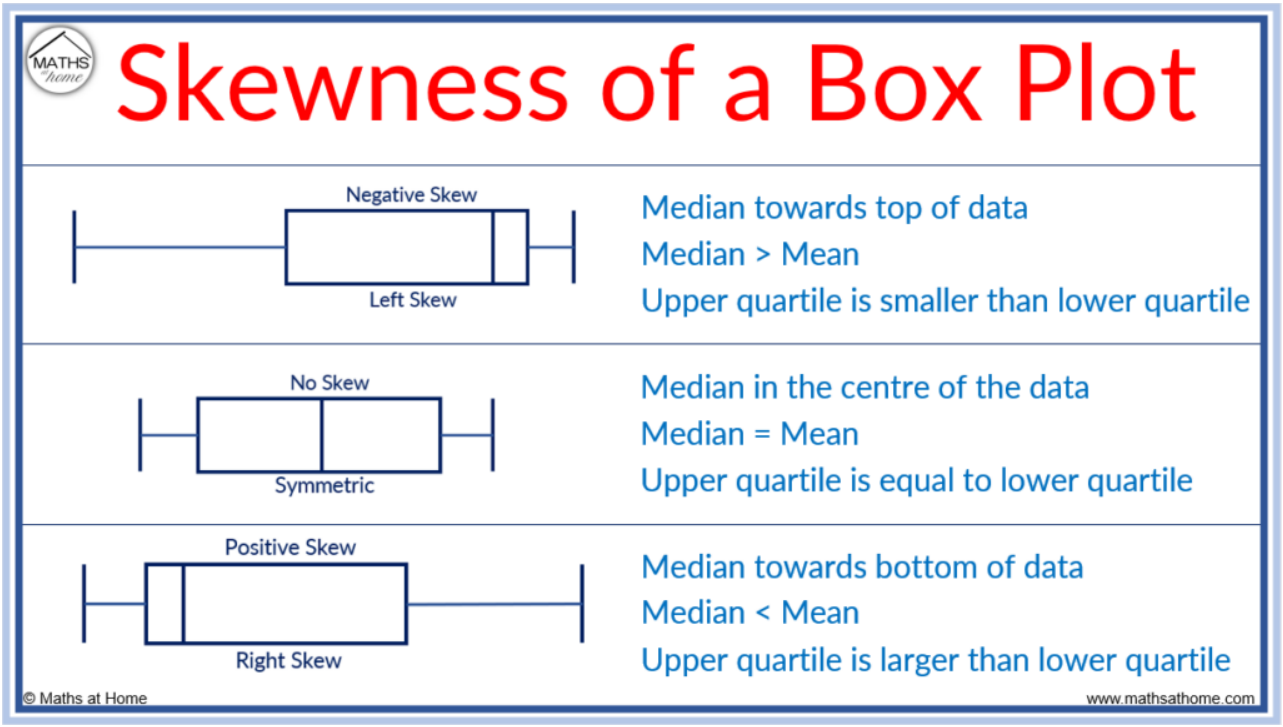
- عندما نقول أن الرسم البياني Right Skewness فهذا يدل على أن Mean > Median.
- يمثل mode أعلى قيمة بالرسم البياني (قمة الرسم البياني).
- من الرسم البياني يمكن ملاحظة تركّز البيانات في الجهة اليسرى.
- مايهمنا من الرسم البياني هو ذيل الرسم البياني ويمثل Outliers حيث نلاحظ تركزها جهة اليمين.

الشكل (Left Skewness) أو مايسمى (Negative Skewness)

- عندما نقول أن الرسم البياني Left 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 اليمنئ.
- من الرسم البياني يمكن ملاحظة أن ذيل الرسم البياني أو Outliers تتركز جهة اليسار.

الشكل (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).



How to Understand and Compare Box Plots – mathsathome.com

ماهو أفضل مقياس من المقاييس السابقة؟

- لايوجد مقياس أفضل من الآخر، والطريقة الأفضل هي استخدام جميع هذه المقاييس في نفس الوقت.
    - If data is Categorical (Nominal or Ordinal) it is impossible to calculate mean or median. So, go for mode.
    - If your data is quantitative then go for mean or median. Basically, if your data is having some influential outliers or data is highly skewed then median is the best measurement for finding central tendency. Otherwise go for Mean.

---

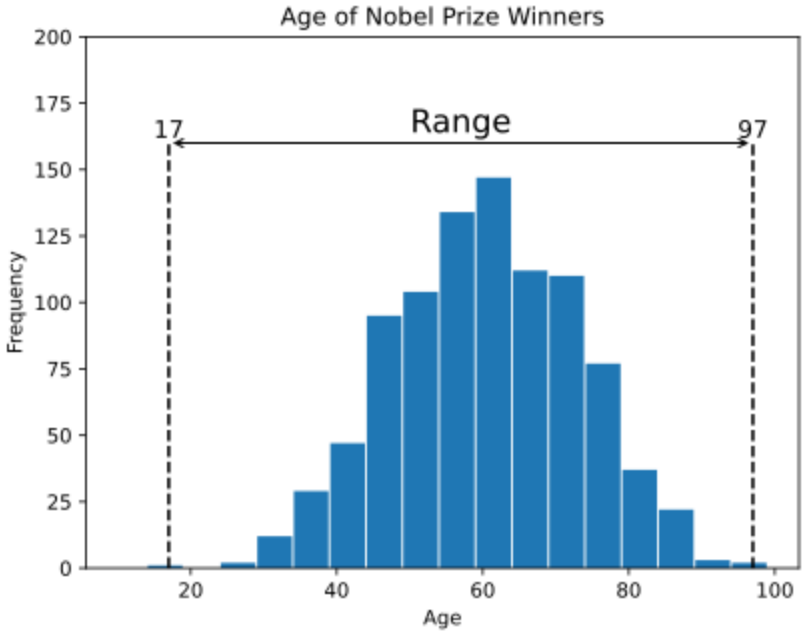
2. Measure of variability (قياس انتشار البيانات)- Descriptive Analysis - Univariate Measures

How Data is Spread Out. Measures of how far away the values in the observations (data points) are from each other.

There are different measures of variation. The most commonly used are:

1. Range
2. Quartiles and Percentiles
3. Interquartile Range
4. Standard Deviation

1. **Range**:
    - The difference between the highest and lowest value in the dataset.
    - Range is the simplest measure of variation.

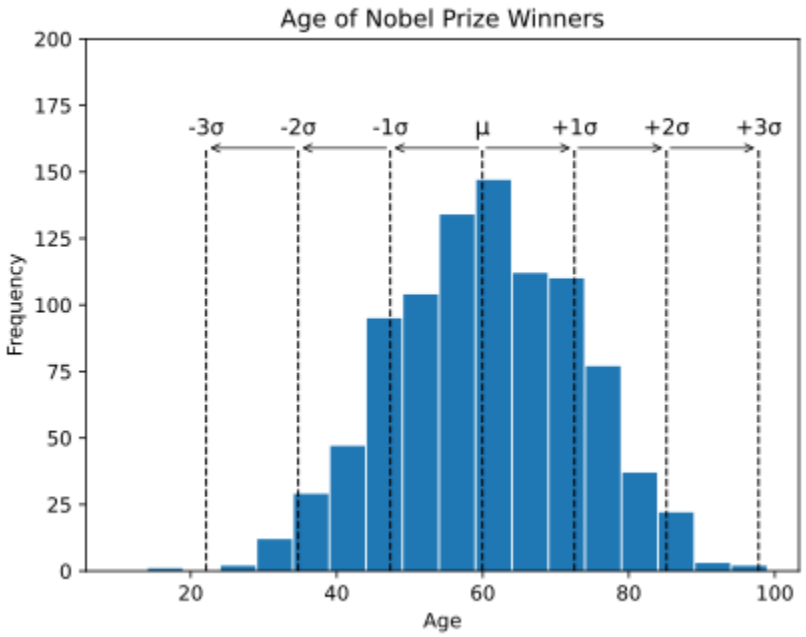


The youngest winner was 17 years and the oldest was 97 years. The range of ages for Nobel Prize winners is then 80 years.

2. **Standard Deviation**
    - Standard deviation is the most used measure of variation.
    - Standard deviation (σ) measures how far a 'typical' observation is from the average of the data (μ).

Standard Deviation Formula

Population	Sample
$\sigma = \sqrt{\frac{\sum(X - \mu)^2}{N}}$	$s = \sqrt{\frac{\sum(X - \bar{x})^2}{n - 1}}$
X - The Value in the data distribution μ - The population Mean N - Total Number of Observations	X - The Value in the data distribution $\bar{x}$ - The Sample Mean n - Total Number of Observations



Each dotted line in the histogram shows a shift of one extra standard deviation.
If the data is **normally distributed:**

- Roughly 68.3% of the data is within 1 standard deviation of the average (from μ-1σ to μ+1σ)
- Roughly 95.5% of the data is within 2 standard deviations of the average (from μ-2σ to μ+2σ)
- Roughly 99.7% of the data is within 3 standard deviations of the average (from μ-3σ to μ+3σ)

Example

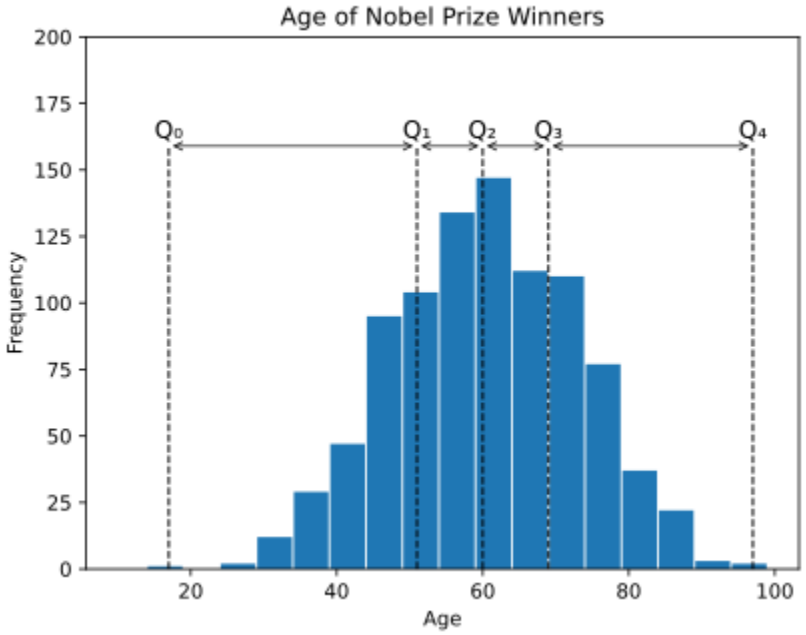
For example, let's calculate the standard deviation of the following numbers: 4, 8, 6, 5, 3.

1. **Find the Mean:**

$$\text{Mean} = \frac{(4+8+6+5+3)}{5} = \frac{26}{5} = 5.2$$

2. **Calculate Each Difference from the Mean, Then Square:**
   - $(4-5.2)^2 = 1.44$
   - $(8-5.2)^2 = 7.84$
   - $(6-5.2)^2 = 0.64$
   - $(5-5.2)^2 = 0.04$
   - $(3-5.2)^2 = 4.84$
3. **Sum of Squares and Calculate Variance:**
   - $\text{Sum of Squares} = 1.44 + 7.84 + 0.64 + 0.04 + 4.84 = 14.8$
   - $\text{Variance} = \frac{14.8}{5} = 2.96$
4. **Calculate Standard Deviation:**
   - $\text{Standard Deviation} = \sqrt{2.96} \approx 1.72$

3. **Quartiles and Percentiles**
   - Quartiles and percentiles are ways of separating equal numbers of values in the data into parts.



Quartiles are values that separate the data into four equal parts.

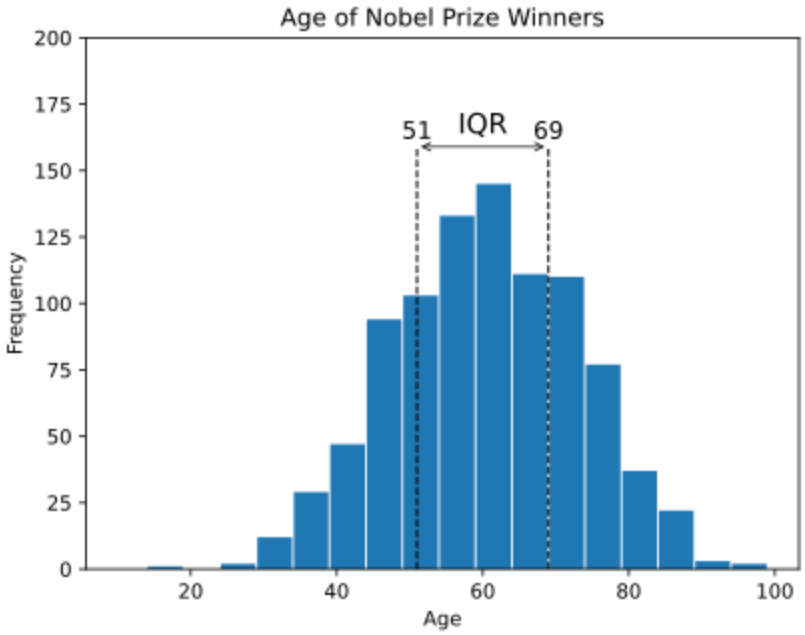
- The quartiles (Q_0,Q_1,Q_2,Q_3,Q_4) are the values that separate each quarter.
- Between Q_0 and Q_1 are the 25% lowest values in the data. Between Q_1 and Q_2 are the next 25%. And so on.
  - Q_0 is the smallest value in the data.
  - Q_1 is the value separating the first quarter from the second quarter of the data.
  - Q_2 is the middle value (median), separating the bottom from the top half.
  - Q_3 is the value separating the third quarter from the fourth quarter
  - Q_4 is the largest value in the data.

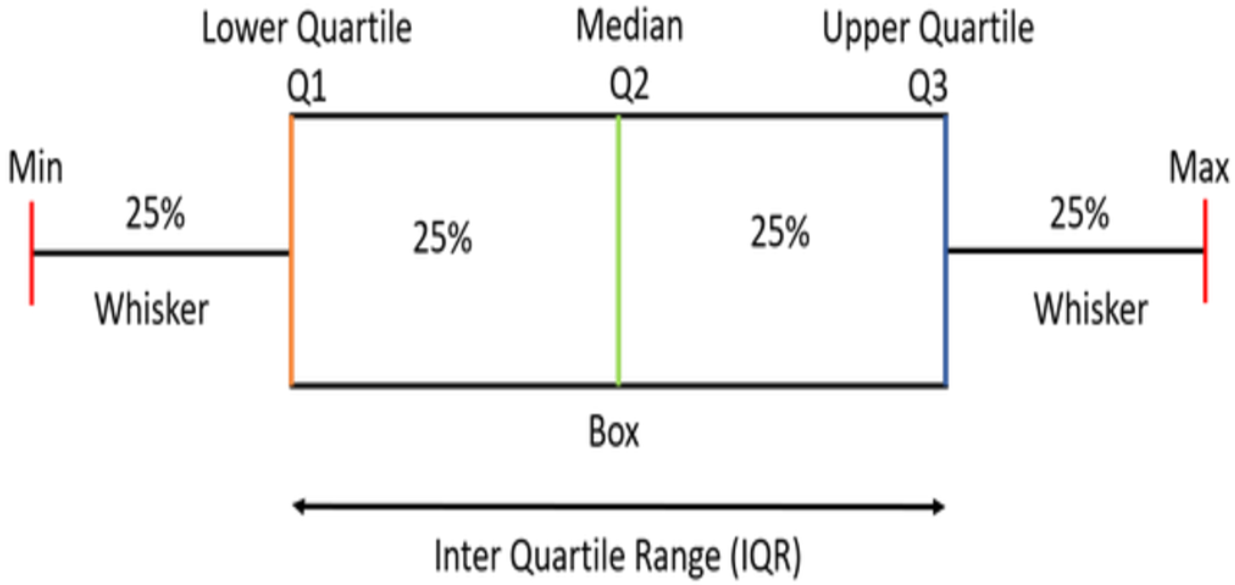
Percentiles are values that separate the data into 100 equal parts.

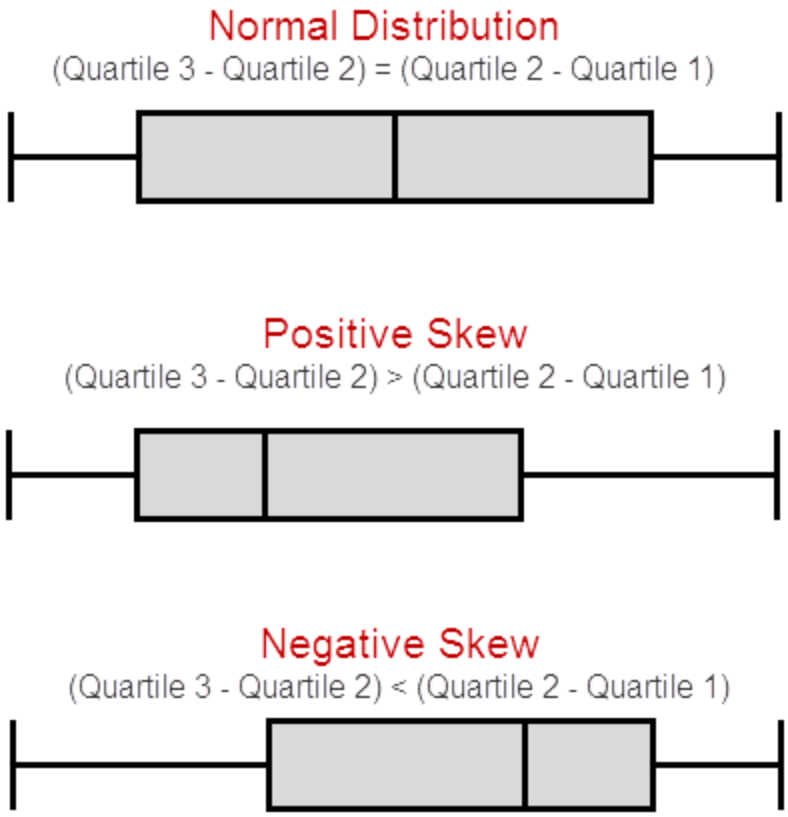
- For example, The 95th percentile separates the lowest 95% of the values from the top 5%
- The 25th percentile ($P_{25\%}$) is the same as the first quartile (Q_1).
- The 50th percentile ($P_{50\%}$) is the same as the second quartile (Q_2) and the median.
- The 75th percentile ($P_{75\%}$) is the same as the third quartile (Q_3)

4. **Interquartile Range**

Interquartile range is the difference between the first and third quartiles (Q_1 and Q_3). The 'middle half' of the data is between the first and third quartile.







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

---

representative for data —> mean , mode, median
The shape of a distribution, or the way in which the data is spread out→ is determined by a combination of its central tendency, variability, and skewness.

Resources:

- https://www.w3schools.com/statistics/
- https://www.analyticsvidhya.com/blog/2021/10/end-to-end-statistics-for-data-science/
- https://www.kdnuggets.com/2020/06/8-basic-statistics-concepts.html
- https://hevodata.com/learn/statistics-for-data-analytics/
- https://makemeanalyst.com/basic-statistics-for-data-analysis/#Basic_Statistics