The measures of Center and The measures of the spread

* Mean : هو متوسط القيم
* Median : هو الوسيط لمجموعة من القيم
* Mode : هو اكثر قيمة تكررت فى مجموعة من القيم

مثال : لو عندك مجموعة من الاطفال مكونه من 11 طفل ومحتاج تحسب ال Mean, Median, Mode لاعمارهم :

* هتجمع جميع الاعمار وتقسم على عددهم تحصل على المتوسط
* عشان تحصل على الوسيط هترتب القيم من الاصغر الى الاكبر وتبحث عن القيمة الى فى الوسط وزى ما فى المثال هى القيمة رقم 6 وهيبقى هى دى الوسيط
* بالنسبة لل Mode ببحث عن اكثر عمر متكرر ويبقى هو دا mode

بعض المعلومات عن القيم الى فاتت :

* ليه بنستخدم قيم زى Median or Mode ومش بنكتفى بال Mean لان احيانا فى قيم بيبقى فيها قيم شاذه outliers بتخلى المتوسط اما كبير جدا عن المفروض او ضغير جدا المفروض .

1. when **there is big outlier**, the mean becomes greater than the median, when there is small outlier, the mean becomes smaller than the median.
2. if we **remove small outlier** from the shape , the mean will increase and the median, and vice versa if we add small outlier.
3. if we **remove big outlier** in the distribution, the mean will decrease and also the median and vice versa if we add it.
4. if we **increase big outlier** or small outlier in the distribution, the mean will increase but the median will still the same.
5. **The more spread out a data distribution is, the greater its standard deviation**
6. standard deviation cannot be negative.
7. A standard deviation close to 0 indicates that the data points tend to be close to the mean
8. when there is outlier, we use the median to know the average better than the mean, and also we use the interquartile range to know the spread better than standard deviation
9. standard deviation will be zero when all the points is same
10. **shifting the values or adding values to the dataset** affects on the measures of center like the mean and median but it doesn't affect on the spread measures ( IQR & standard deviation
11. Scaling or multiplying affects on the measures of center and the spread.
12. the mean will be at the right of the median when the distribution is skewed to the right.
13. the mean will be t at the left of the median when the distribution is skewed to the left.
14. standard normal distribution is shape with the mean equal 0 and standard deviation is equal

المعلومات دى تقدر تخليك تحكم او تفهم البيانات بشكل احسن وكمان لو حصل اى تعديلات على البيانات تقدر تتوقع الى ممكن يحصل للقيم الى حصلت عليها.

* من القيم الى محتاج تعرفها برده عن مجموعة من القيم هى الاقل قيمة والاعلى قيمة والفرق ما بينهم يسمى Range .
* وايضا variance وهو مجموع الانجراف عن المتوسط
* و ايضا ٍStandard Deviation وهو متوسط الانحراف عن المتوسط ودا بيكون عندك صورة عن شكل البيانات الى عندك
* ومن اهمية الانحراف المعيارى استخدامه فى معرفة اين تقع القيم وهى الاتى  **empirical rule**:
  1. 68% من القيم تقع بين ( Mean +,- Standard Deviation )
  2. 95% من القيم تقع بين ( Mean +,- 2 \*Standard Deviation )
  3. 99% من القيم تقع بين (Mean +,- 3 \*Standard Deviation )
* ايضا من القيم المهمة لمعرفة خصاشص البيانات هى Q1 , Q2 وايضا interquartile Range الفرق بين Q1,Q3
* فى قاعده مهمة فى خصائص البيانات هى

**Central limit theorem:** says that more measurements, more likely your data will be distributed as you expect.

Vann Diagram : وهى فكرة البيانات المشتركة زى مثلا عندنا شكولاتة والشكولاتة متقسمة شكولاته بالفذدق - وشكولاتة بالمكسرت وفى شكولاته بالفذدق والمكسرات وفى شكولاته ولا فيها فذدق ولا مكسرات

1. **Dot plot** is plot to show the frequency of number.
2. **histogram** displays numerical data by grouping data into "bins" of equal width. Each bin is plotted as a bar whose height corresponds to how many data points are in that bin. Bins are also sometimes called "intervals", "classes", or "buckets"
3. **Stem and leaf plots is to** display numerical data by splitting each data point into a "leaf" (usually the last digit) and a "stem" (the leading digit or digits).
4. **Distribution skewed right** ( like the number of deaths related to the ages, you will notice the most number of people who is old so the curve will be skewed right.
5. **Distribution skewed left :**
6. يوجد بعض المصطلحات المستخدمة فى وصف البيانات زى :
   1. **cluster** is the group of variables together and separate of other groups
   2. **Gaps** is the period between two variables don't have any result.
   3. **peak** is the maximum number achieved in cluster and it is different that maximum because there may be two peaks in the distribution.
7. **line graph** is used with time changing, and also taking about the manipulation happened in X axis and Y axis and the hidden factors may change the results.
8. **scatterplots** to see relationships between variables. Scatterplots are really good for helping us see if two variables have positive or negative association.
   1. ***Form****:* Is the association linear or nonlinear? *Direction:* Is the association positive or negative?
   2. ***Strength****:* Does the association appear to be strong, moderately strong, or weak?
   3. ***Outliers****:* Do there appear to be any data points that are unusually far away from the general pattern?
   4. **The correlation coefficient r :** measures the direction and strength of a linear relationship.
      1. It always has a value between −1, 1.
      2. Strong positive linear relationships have values of *r* closer to 1.
      3. Strong negative linear relationships have values of *r* closer to −1
      4. Weaker relationships have values of *r*  closer to 0.
9. **percentile** is the percent of the all data below the value we want like:
   1. 20% of the drinks have 20 grams of sugar or less.
   2. 50% of the drinks have 30 grams of sugar or less.
   3. 100% of the drinks have 50 grams of sugar or less.
10. **Z - Score: measures** exactly how many standard deviations above or below the mean a data point is.
    1. positive z-score says the data point is above average.
    2. negative z-score says the data point is below average.
    3. z-score close to 00 says the data point is close to average.
    4. If z-score is 0 means that the data point is the mean.
    5. data point can be considered unusual if its z-score is above 3 or below −3.

* **Statistical question** is any question needed to collect data.
* **Types of Baises :**
  1. **Uncovering Samples :** وهي فكرة عدم تغطية كل العينات التى تشمل كل المجموعه مما يتسبب فى خطأ فى النتائح
  2. **Overestimating sample :** وهى اختيار عينة منحازة وتقدر بشكل خطأ)overestimating sample like singers make a vote on his website about his sings(
  3. **volunteering sample**: like ( to ask his relatives to vote for him )

**Types of the studies**

1. **O*bservational study:*** we measure or survey members of a sample without trying to affect them.
2. ***Controlled experiment:*** we assign people or things to groups and apply some treatment to one of the groups, while the other group does not receive the treatment.
3. **Sample study:** specify parameter we need like standard deviation or mean or median.