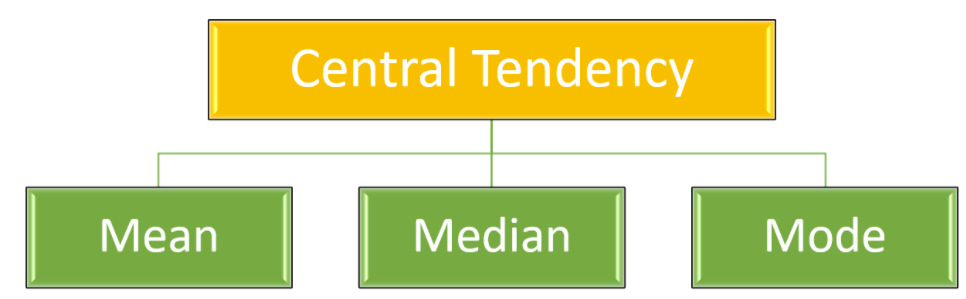
## Central Tendency (Measure of Centre)

**Merkezi Eğilim (Merkez Ölçüsü)**

**Description (Açıklama**)

The central tendency concept is that one single value can best describe the data. Mean, median, and mode are the three important parameters in statistics. Essentially, all three of them refer to a single aspect called the **Central Tendency**. Let's take a closer look at this.

Merkezi eğilim kavramı, tek bir değerin verileri en iyi şekilde tanımlayabilmesidir. Mean(ortalama), medyan ve mode istatistikteki üç önemli parametredir. Esasen, üçü de Merkezi Eğilim adı verilen tek bir yönü ifade eder. Buna daha yakından bakalım.



The mean (the average) is most the famous measure of central tendency that you are probably familiar with, but there are also others, such as the median and the mode. The mean, median and mode are all valid measures of central tendency, but under various conditions, one measure of central tendency might become more appropriate than others.

Ortalama (ortalama), muhtemelen aşina olduğunuz en ünlü merkezi eğilim ölçüsüdür, ancak medyan ve mod gibi başkaları da vardır. Ortalama, medyan ve mod geçerli merkezi eğilim ölçüleridir, ancak çeşitli koşullar altında bir merkezi eğilim ölçüsü diğerlerinden daha uygun olabilir.



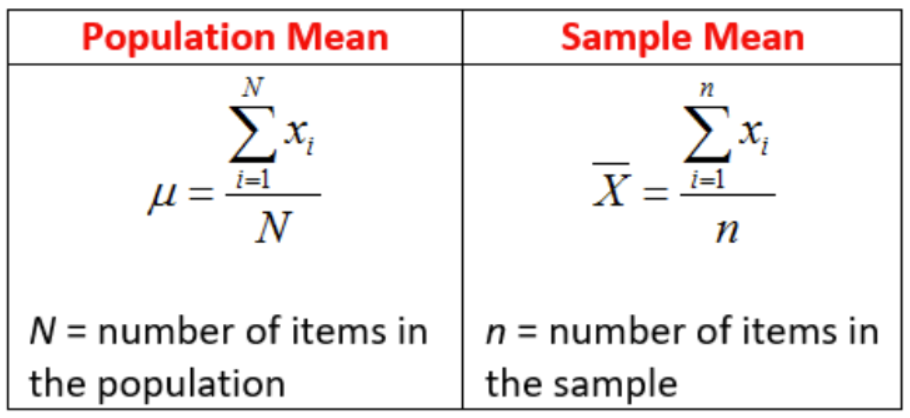
**Central Tendency (Measure of Centre)**

**Mean ( 𝜇 )**

**The mean is equal to the sum of the values in the dataset divided by the number of values**. The number of values in the dataset will be equal to the population or sample size. The table below table gives the formula for the population mean and the sample mean.

Mean(ortalama), veri kümesindeki değerlerin toplamının değer sayısına bölünmesine eşittir. Veri kümesindeki değerlerin sayısı, popülasyon veya örneklem boyutuna eşit olacaktır. Aşağıdaki tablo, popülasyon ortalaması ve örnek ortalaması için formülü verir.





One of the major disadvantages of using mean rather than using median or mode is, the mean is particularly sensitive to the effect of extreme values. Extreme values are also called **outliers**. We will make a technical description of outliers, however, these are the values that are unusual by being relatively small or large in numerical value compared to the rest of the dataset. For example, consider people's salaries at a factory below.

Medyan veya mod kullanmak yerine ortalama kullanmanın en büyük dezavantajlarından biri, ortalamanın özellikle uç değerlerin etkisine duyarlı olmasıdır. Aşırı değerlere aykırı değerler de denir. Aykırı değerlerin teknik bir tanımını yapacağız, ancak bunlar, veri kümesinin geri kalanına kıyasla sayısal olarak nispeten küçük veya büyük olduğu için olağandışı değerlerdir. Örneğin, aşağıdaki bir fabrikadaki insanların maaşlarını düşünün.





We can say $102K dollars is an outlier. The total salary of employees is $333K and the sample size is nine. The mean salary for these nine staff is $37K (333/9=37). Inspecting the raw data, however, suggests that this mean value may not be the best way to accurately reflect a staff's typical salary, as most staff receive salaries between $24K and $33K thousand dollars. We would like to have a better measure of central tendency in this situation. Therefore, taking the median might be a better measure of central tendency.

102 bin doların uç değer olduğunu söyleyebiliriz. Çalışanların toplam maaşı 333 bin dolar ve örneklem büyüklüğü dokuz. Bu dokuz personelin ortalama maaşı 37 bin dolar (333/9=37). Bununla birlikte, ham verileri incelemek, çoğu personel 24 bin dolar ile 33 bin dolar arasında maaş aldığından, bu ortalama değerin bir personelin tipik maaşını doğru bir şekilde yansıtmanın en iyi yolu olmayabileceğini gösteriyor. Bu durumda daha iyi bir merkezi eğilim ölçüsüne sahip olmak istiyoruz. Bu nedenle, medyanı almak daha iyi bir merkezi eğilim ölçüsü olabilir.

**💡Tips:**

* Under various conditions, one measure of central tendency might become more appropriate than others.
* Under various conditions, one measure of central tendency might become more appropriate than others.

## Central Tendency (Measure of Centre)

### Median

**The median is the middle score for a dataset that has been sorted from small to large.** Outliers less affect the median. Suppose we have the same data to calculate the median.

Medyan, küçükten büyüğe sıralanmış bir veri kümesinin orta puanıdır. Aykırı değerler medyanı daha az etkiler. Medyanı hesaplamak için aynı verilere sahip olduğumuzu varsayalım.

Firstly, we need to sort the data from small to large.



The median is the middle score, in this case, it is $30K. $30K is the middle score because there are 4 scores after it and 4 scores before it. This works well when you have an odd number of scores, but what will happen when you have an even number of sample size? Even if you only had 10 scores? In this case, we simply have to take the middle two scores and average the result. So, if we look at the example below:

Medyan orta puandır, bu durumda 30 bin dolardır. 30 bin dolar orta puan çünkü ondan sonra 4 puan ve ondan önce 4 puan var. Bu, tek sayıda puanınız olduğunda işe yarar, ancak çift sayıda örneklem büyüklüğünüz olduğunda ne olur? Sadece 10 puanın olsa bile mi? Bu durumda, ortadaki iki puanı almamız ve sonucun ortalamasını almamız yeterlidir. Yani, aşağıdaki örneğe bakarsak:





Now, we should take the 5th ($30K) and 6th ($33K) scores in our dataset and average them to get a median of $31.5K.

Şimdi, veri kümemizdeki 5. (30.000 $) ve 6. (33.000 $) puanları almalı ve medyan 31.5K elde etmek için ortalamalarını almalıyız.



**💡Tips:**

* The median is the middle score. If the sample size is 9, the fifth element is the median.

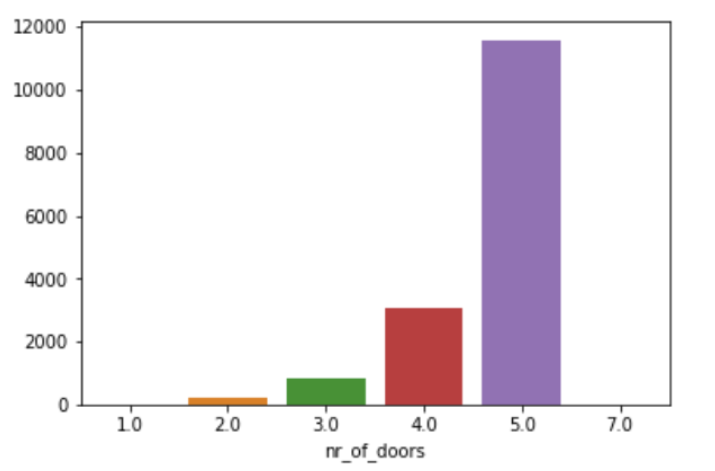
## Central Tendency (Measure of Centre)

### Mode

**The mode is the most frequent score in a dataset**. It represents the highest bar in a histogram or bar chart or. Therefore, sometimes you can consider the mode as being the most popular option. The mode is normally used for categorical data where we want to know which category is the most common. An example of a mode is presented below.

Mod, bir veri kümesinde en sık görülen puandır. Bir histogramdaki veya çubuk grafikteki en yüksek çubuğu temsil eder veya. Bu nedenle, bazen modu en popüler seçenek olarak düşünebilirsiniz. Mod normalde hangi kategorinin en yaygın olduğunu bilmek istediğimiz kategorik veriler için kullanılır. Bir mod örneği aşağıda sunulmuştur





The histogram shows the number of doors among used cars where the sample size is around 15000. We can say the most popular option is the 5-door cars. Therefore, the mode for this dataset is 5.

To find out the mode, suppose we have the same dataset.

Histogram örneklem büyüklüğü 15000 civarında olan kullanılmış arabalar arasında kapı sayısını göstermektedir. En popüler seçeneğin 5 kapılı arabalar olduğunu söyleyebiliriz. Bu nedenle, bu veri kümesi için mod 5'tir.

Modu bulmak için aynı veri setine sahip olduğumuzu varsayalım.



**The mode is the most frequent score in a dataset**. So, we can say $33K is the mode in our dataset. Because there are 3 different staff who receives $33K. Let's look at mean, median and mode values for the same dataset (for nine staff):

Mod, bir veri kümesinde en sık görülen puandır. Yani, veri kümemizdeki modun $33K olduğunu söyleyebiliriz. Çünkü 33 bin dolar alan 3 farklı personel var. Aynı veri seti için ortalama, medyan ve mod değerlerine bakalım (dokuz personel için):



mean: $37K

median: $30K

mode : $33K

**Tips:**

* The mode is the most popular option.

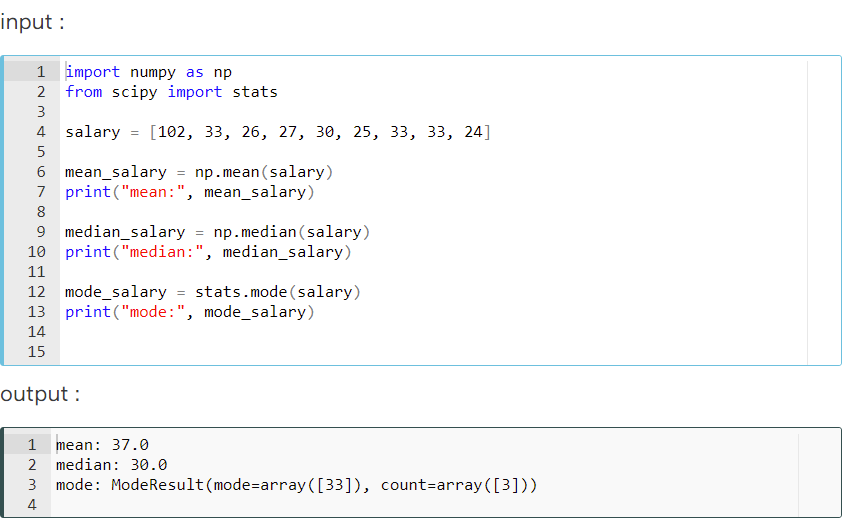
## Central Tendency (Measure of Centre)

### Calculate Mean, Median and Mode with Python

We can easily calculate mean, median and mode values with python. We use the numpy library for the mean and median, and the **scipy** library for the mode. You can compare the values we get by Python with the values we calculated manually.

Python ile ortalama, medyan ve mod değerlerini kolayca hesaplayabiliriz. Ortalama ve medyan için numpy kütüphanesini ve mod için **scipy** kütüphaneyi kullanıyoruz. Python ile elde ettiğimiz değerleri manuel olarak hesapladığımız değerlerle karşılaştırabilirsiniz.





## Dispersion (Measure of Spread)

### Introduction

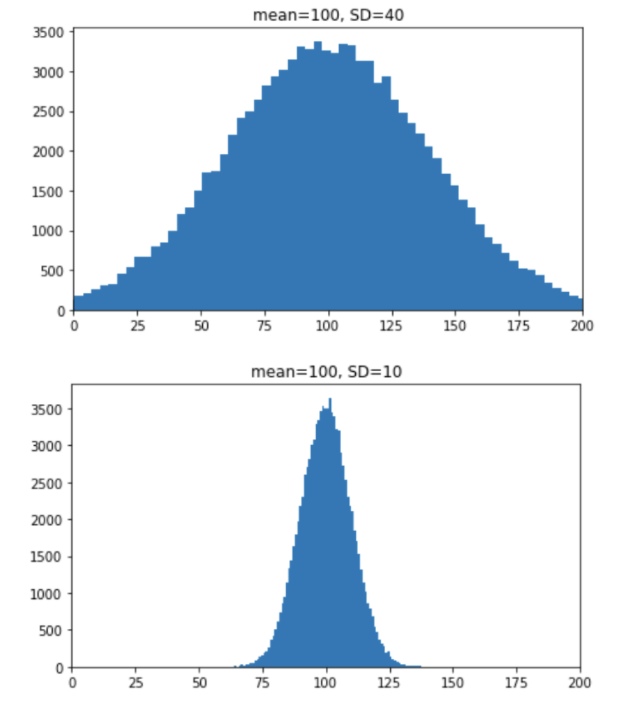
In statistics, the measure of central tendency gives a single value that represents the entire data; however, a single value can not describe the observation exactly. At this point, the dispersion helps us to study the variability of the items. Dispersion is a way to explain how a dataset is distributed. When a dataset has a small value, the values in the dataset are tightly clustered; when it is large the items in the set are widely scattered.

As can be seen in the two histograms below, there may be different distributions with the same mean value (𝜇 =100). The first population is much more dispersed than the second population, however, the mean value for both populations is the same. Therefore, we can say a dispersion explains something more than the central tendency does.

Dağılım (Yayılma Ölçüsü)

İstatistikte, merkezi eğilim ölçüsü, tüm verileri temsil eden tek bir değer verir; ancak tek bir değer gözlemi tam olarak tanımlayamaz. Bu noktada dağılım, maddelerin değişkenliğini incelememize yardımcı olur. Dağılım, bir veri kümesinin nasıl dağıtıldığını açıklamanın bir yoludur. Bir veri kümesi küçük bir değere sahip olduğunda, veri kümesindeki değerler sıkı bir şekilde kümelenir; büyük olduğunda setteki öğeler geniş çapta dağılır.

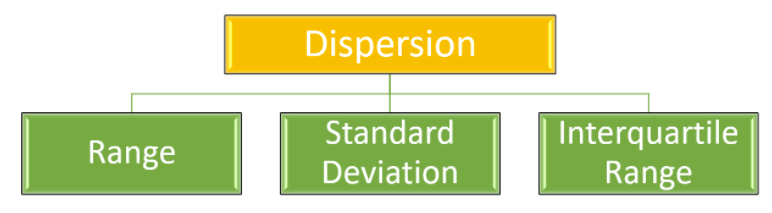
Aşağıdaki iki histogramdan da görüleceği gibi aynı ortalama değere (𝜇 =100) sahip farklı dağılımlar olabilir. Birinci popülasyon, ikinci popülasyona göre çok daha dağınıktır, ancak her iki popülasyon için ortalama değer aynıdır. Bu nedenle, bir dağılımın (dispersion ) merkezi eğilimden (central tendency ) daha fazlasını açıkladığını söyleyebiliriz.



**Range, standard deviation** and **interquartile range** are the three widely used measures of dispersion. Now, we are going to consider these concepts.

Aralık, standart sapma ve çeyrekler arası aralık, yaygın olarak kullanılan üç dağılım ölçüsüdür. Şimdi bu kavramları ele alacağız.





**Tips:**

* A dispersion explains something more than the measure of central tendency does.
* • Bir dağılım, bir şeyi merkezi eğilim ölçüsünden daha fazlasını açıklar



## Dispersion (Measure of Spread)

### Range

The range is the simple measure of dispersion, which is defined as the **difference between the maximum and the minimum values**.

The main advantage of the range is that **it is easy to calculate**.

There are many disadvantages, on the other hand:

**It is highly susceptible to extreme values** and

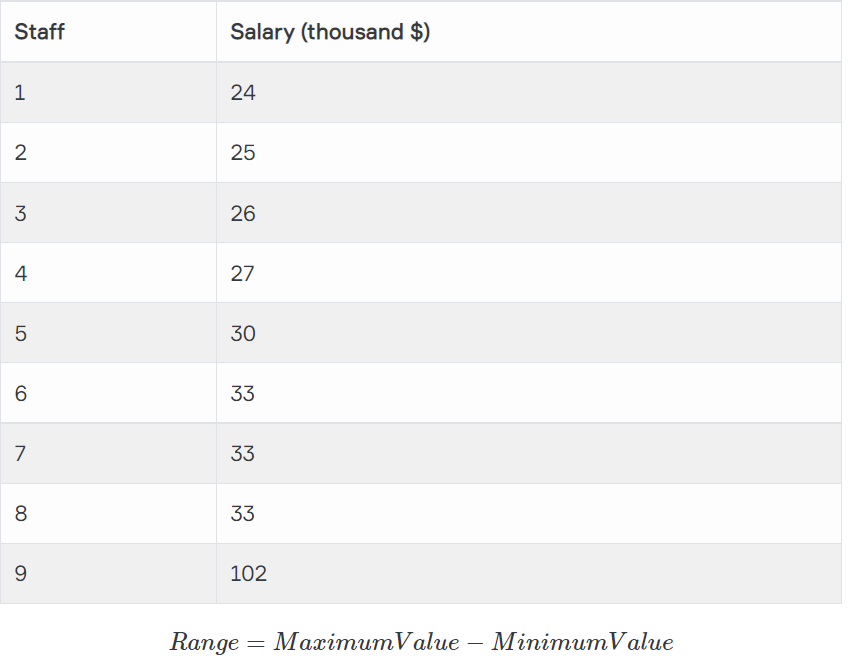
**does not use all the observations in a dataset**.

In this case, if we look at our salary table again:

Aralık, maksimum ve minimum değerler arasındaki fark olarak tanımlanan basit dağılım ölçüsüdür. Aralığın ana avantajı, hesaplanmasının kolay olmasıdır. Öte yandan birçok dezavantajı var. Uç değerlere oldukça duyarlıdır ve bir veri setindeki tüm gözlemleri kullanmaz.

Bu durumda maaş tablomuza tekrar bakarsak





The difference between the maximum value and the minimum value is 102-24=78. We can say the range for this dataset is 78. As you see in the example it gives the difference between the maximum and minimum two values out of nine values and does not use all the observations. And it is highly susceptible to extreme values because $102K affected the range very badly. If we remove that value, the range of the remaining values will be 9 (33-24= 9).

Maksimum değer ile minimum değer arasındaki fark 102-24=78'dir. Bu veri kümesinin aralığı 78'dir diyebiliriz. Örnekte gördüğünüz gibi dokuz değerden maksimum ve minimum iki değer arasındaki farkı verir ve tüm gözlemleri kullanmaz. Ve aşırı değerlere karşı oldukça hassas çünkü 102 bin dolar aralığı çok kötü etkiledi. Bu değeri kaldırırsak kalan değerlerin aralığı 9 olur (33-24=9).



## Dispersion (Measure of Spread)

### Standard Deviation(𝜎)

The most commonly used measure of dispersion is the standard deviation (𝜎). Standard deviation **measures the spread around the mean**. It is also expressed as **the square root of variance**. Therefore we must first describe the variance(𝜎2). Variance is defined as the average of the squared differences from the mean. The formula for the variance and standard deviation is given below.

En yaygın olarak kullanılan dağılım ölçüsü standart sapmadır (𝜎). Standart sapma, ortalama etrafındaki yayılımı ölçer. Varyansın karekökü olarak da ifade edilir. Bu nedenle önce varyansı (𝜎2) tanımlamalıyız. Varyans, ortalamadan farkların karelerinin ortalaması olarak tanımlanır. Varyans ve standart sapma formülü aşağıda verilmiştir.



Variance=

σ2=∑(x−μ)2Nσ2=∑(x−μ)2N

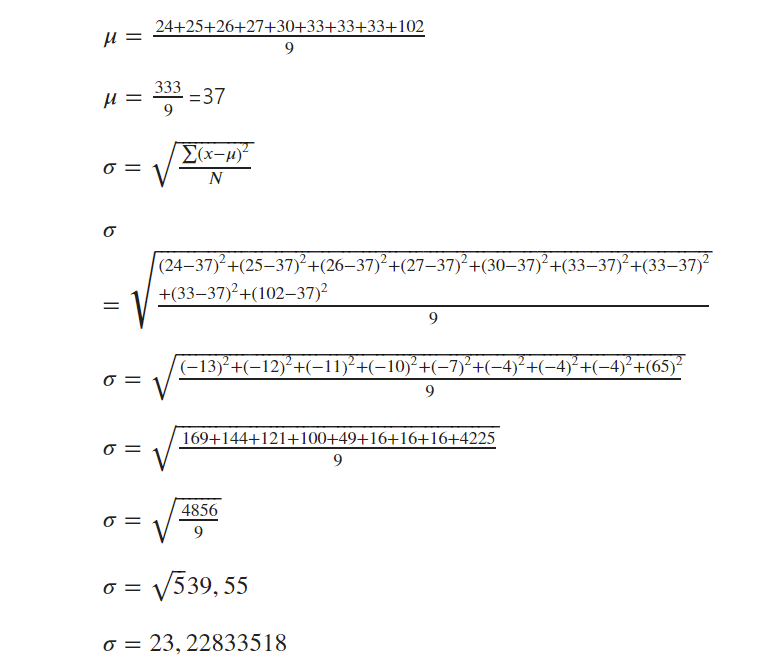
Standard Deviation=

σ=∑(x−μ)2N−−−−−−−−−√σ=∑(x−μ)2N

𝜇 = population mean  
𝑁 = number of items in the population

Let's go back to our salary table again and calculate the standard deviation and the variance for these salaries.





Like the range, also standard deviation is affected by outliers. One value could contribute greatly to the results of the standard deviation. This also means the standard deviation is a good indicator of the existence of outliers.

For example if we remove the outlier from our table, the salary table would be:

Aralık gibi, standart sapma da aykırı değerlerden etkilenir. Bir değer, standart sapmanın sonuçlarına büyük ölçüde katkıda bulunabilir. Bu aynı zamanda standart sapmanın aykırı değerlerin varlığının iyi bir göstergesi olduğu anlamına gelir.

Örneğin aykırı değeri tablomuzdan çıkarırsak maaş tablosu şöyle olur:



* 

And if we recalculate the standard deviation regarding the new salary table:

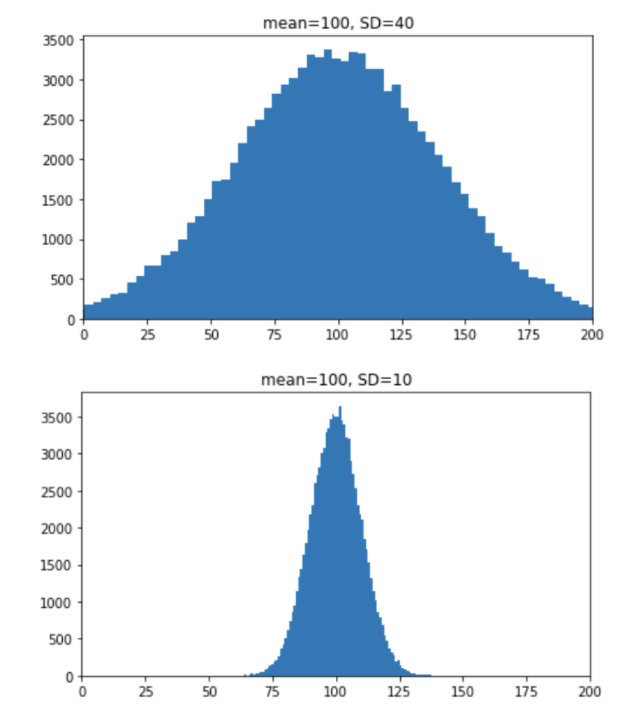
σ=3,58

You can see how only one outlier affects the standard deviation.

The standard deviation is also useful when comparing the spread of two different datasets that have the same mean. The dataset with the smaller standard deviation has a narrower spread of measurements around the mean and therefore usually has relatively less high or low values. In the following example, the standard deviation for the first population is 40, however, the standard deviation for the second one is 10. You see the second population has a narrower spread of measurements around the mean.

Yalnızca bir aykırı değerin standart sapmayı nasıl etkilediğini görebilirsiniz.

Standart sapma, aynı ortalamaya sahip iki farklı veri kümesinin dağılımını karşılaştırırken de yararlıdır. Daha küçük standart sapmaya sahip veri kümesi, ortalama etrafında daha dar bir ölçüm dağılımına sahiptir ve bu nedenle genellikle nispeten daha az yüksek veya düşük değerlere sahiptir. Aşağıdaki örnekte, birinci popülasyon için standart sapma 40'tır, ancak ikincisi için standart sapma 10'dur. İkinci popülasyonun ortalama etrafında daha dar bir ölçüm dağılımına sahip olduğunu görüyorsunuz.



**Tips:**

* The data with the smaller standard deviation has a narrower spread of measurements around the mean.

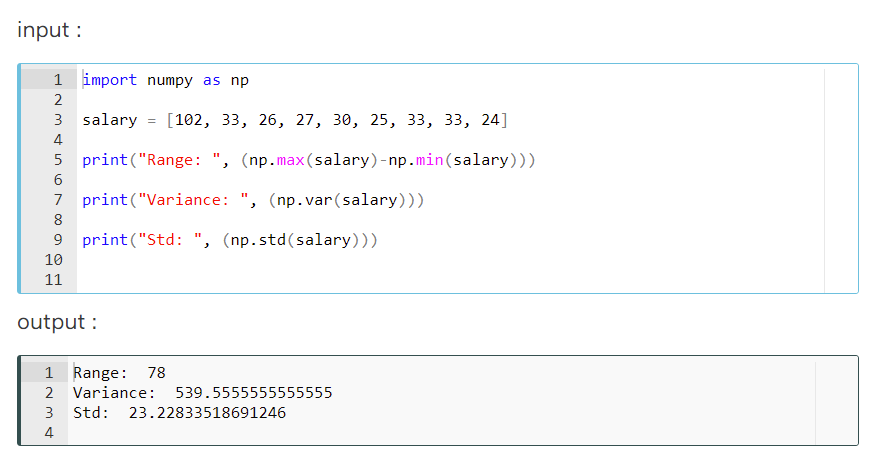
• Daha küçük standart sapmaya sahip veriler, ortalama etrafında daha dar bir ölçüm dağılımına sahiptir.



## Dispersion (Measure of Spread)

### Calculate Range, Variance and Standard Deviation with Python

We can easily calculate range, variance and standard deviation values with numpy. You can compare the values we get by numpy with the values we calculated manually.



## Dispersion (Measure of Spread)

### Inter\* Quartile Range (IQR)

Quartiles are the values that divide a group of numbers into quarters. Q1 or the 25th percentile is the first quartile and defined as the middle number between the smallest number and the median of the dataset. Q2 is the second quartile which is the median of the whole dataset. Q3 or 75th percentile is the third quartile which is the middle value between the median and the highest value of the dataset

Çeyrekler, bir sayı grubunu dörde bölen değerlerdir. Q1 veya 25. yüzdelik dilim ilk çeyrektir ve en küçük sayı ile veri kümesinin medyanı arasındaki orta sayı olarak tanımlanır. Q2, tüm veri setinin medyanı olan ikinci çeyrektir. Q3 veya 75. yüzdelik dilim, veri kümesinin medyanı ile en yüksek değeri arasındaki orta değer olan üçüncü çeyrektir.



For example, a dataset consists of those numbers: 0,4,5,7,8,9,10,12,13,14,15,16,20.

The median (Q2) is the value in the middle of the list. In this case, 10 is the median number.

The first quartile (Q1) is the middle number in between the smallest number (0) and the median (10) which is 7. In other words, the middle number between 0 and 10 is 7.

The third quartile (Q3) is the middle value between the median (10) and the highest value (20) in this case that will be 14. In other words, the middle number between 10 and 20 will be 14.

Örneğin, bir veri kümesi şu sayılardan oluşur: 0,4,5,7,8,9,10,12,13,14,15,16,20.

Medyan (Q2), listenin ortasındaki değerdir. Bu durumda, 10 medyan sayıdır.

İlk çeyrek (Q1), en küçük sayı (0) ile ortanca (10) arasındaki ortadaki sayıdır ki bu da 7'dir. Yani 0 ile 10 arasındaki orta sayı 7'dir.

Üçüncü çeyrek (Q3), medyan (10) ile en yüksek değer (20) arasındaki orta değerdir ve bu durumda 14 olacaktır. Yani 10 ile 20 arasındaki orta sayı 14 olacaktır.

**Inter Quartile Range(IQR)** is the **difference between Q3 and Q1**.

Çeyrekler Arası Aralık (IQR), Q3 ve Q1 arasındaki farktır.

In this case:

IQR=Q3-Q1

IQR=14-7=7

As you remember we have already mentioned the extreme values and named them as outliers. In statistics, an outlier is a data point that differs significantly from other observations. IQR helps us to make a technical description of outliers. **A definition of the outlier is, any data point more than 1.5 interquartile ranges (IQRs) below the first quartile or above the third quartile.**

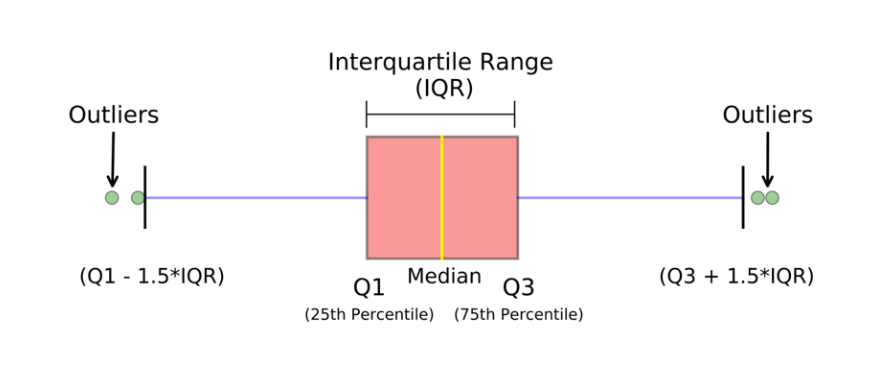
Hatırladığınız gibi, uç değerlerden daha önce bahsetmiştik ve bunları aykırı değerler olarak adlandırmıştık. İstatistikte aykırı değer, diğer gözlemlerden önemli ölçüde farklı olan bir veri noktasıdır. IQR, aykırı değerlerin teknik bir tanımını yapmamıza yardımcı olur. Aykırı değerin bir tanımı, birinci çeyreğin altında veya üçüncü çeyreğin üzerinde 1.5 çeyrekler arası aralıktan (IQR) fazla herhangi bir veri noktasıdır.



In this case, we can say:

**Outliers: (Q1 - 1.5 \* IQR) or (Q3 + 1.5 \* IQR)**

The following picture shows the relationship between IQR and outliers.



**Tips:**

* Outlier is, any data point more than 1.5 IQR below the Q1 or above the Q3.
* Aykırı değer, Q1'in altında veya Q3'ün üzerinde 1.5 IQR'den fazla herhangi bir veri noktasıdır.

## Dispersion (Measure of Spread)

### Practice IQR

Imagine we have the following number list.

number\_list = [1, 5, 10, 15, 40]

Now we will try to find which numbers on our list are the outlier.

We have the following summary:

minimum number = 1

maximum number =40

median=10

Q1 = 5

Q3 = 15

IQR = Q3-Q1

IQR= 15-5 = 10

Therefore, (1.5 \* IQR) = 15

To determine if there are any outliers, we must consider the numbers that are 1.5\*IQR beyond the quartiles.

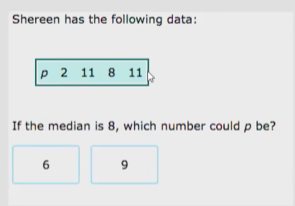
Q1 – (1.5 \* IQR) = 5-15 = -10

Q3 + (1.5 \* IQR) = 15+15 = 30

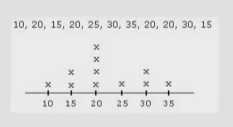
The last number in our list is 40. And it is outside of the interval from (–10) to (30), therefore 40 is an outlier. The rest of the numbers in the list are not outliers.

**KONUNUN ÖZETİ (KAHOOT ÜZERİNDEN)**

Mean outlierlara karşı sensitivedir.

Soru: eğer median 8 ise p hangi rakam olmalıdır? 

Cevap: Sort edersek –> 2, 8, 11, 11… Eğer 9 olursa 8 iyice sola kayacak ve ortada 9 kalacak. 6 olmalı ki 8 ortadaki rakam olsun.

Soru : Mode burada kaçtır? 

Cevap : Mode en fazla tekrar eden eleman sayısı idi. Bu durumda mode, 20’dir.,

Soru : Ne zaman mean ile median aynı olur?

Cevap : When the distribution is symmetric. Hatta bu durumda mode da eşit olur.

Soru : ……….is a way to explain how a dataset is distributed.

Cevap : dispersion

**Central tendency** de bir tane metrik değer bir datanın dağılımının gösterirken (ya meandir ya mediandır ya mode dur) ama dağımımızın şekli saçılımı farklı olduğu halde mean ve median aynı olduğu durumlar oluyordu. O zaman bu iki datasetini birbirinden nasıl birbirinden ayırt edebileceğim? O zaman dipersion dediğimiz özellikler imdadımıza yetişiyor.

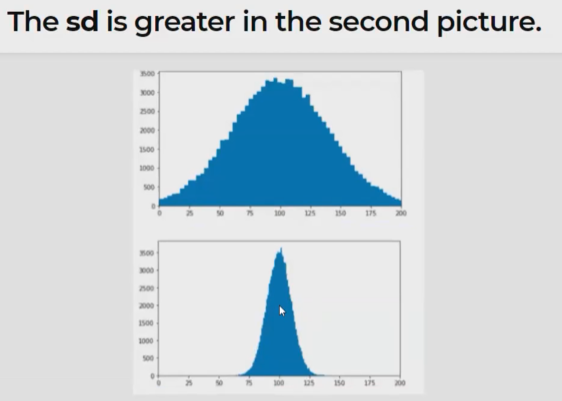
Bu özellikler: Range, Standard Deviation ve IQR.

**Range** dediğimiz şey minimum ve maksimum değerler sözkonusu oluyor. Elimdeki datasetimin minimum ve maksimum aldığı değerlere bakıyorum. Demek ki bu aralıkta değişen değerler var elimde diyebiliyorum.

**Standard Deviation**’da bu dağılımın mean’ini hesap ediyordum. Data setiminin her bir elemanı mean’den ne kadar uzaklaşmış ona bakıyorum. Ama ne kadar uzaklaştığına sadece farktan bakarsak olmaz. Elemanların karelerini alıyor odan devam ediyorduk. Onları topluyor ve eleman sayısı-1’e bölüyorduk sample tarafında. Ama popülasyonda normal n’e (eleman sayısına) bölüyorduk.

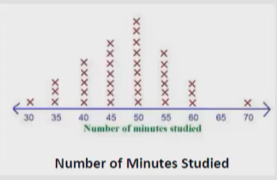
**IQR**: Interquartile Range bize biraz daha farklı birşeyler söylüyor. Datasetimi quartile lara bölüp (4’te birlere bölüp) her bir 4 te bir içerisine hangi değerlerin düştüğünü görüp, burdan dışarda outlierlarımın kalıp kalmadığını görüp bana data setim hakkında önemli şeyler söylüyor.

Histogram ile IQR arasındaki fark neydi? His. 2 boyutluydu frekansı veriyordu ve alt tarafta interval’ları veriyordu ama IQR ise tamamen bir summarizationdur. Elimizdeki data setinin özetlenmiş rakamlarını ifade eder.

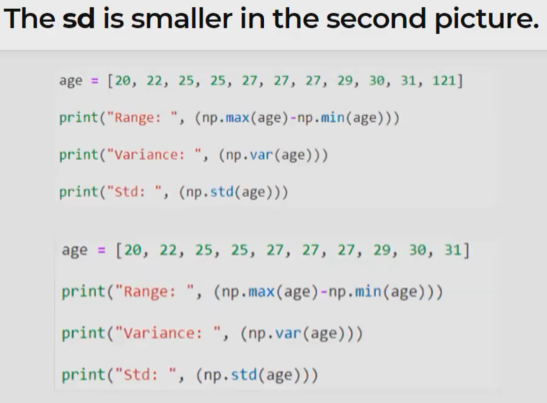
Soru : 

Cevap : False.. Alttaki resimde standart deviation küçüktür. Yani dağılım azdır.

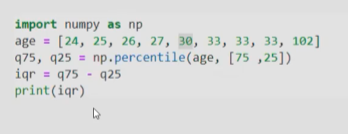
Standart deviation’ın yol arkadaşı meandir. Her elemanın mean’den uzaklaşma durumunu verir. Alttaki resimde elemanların mean’ etrafında toplandığını görüyoruz.

Soru : What is the range of this data? 

Cevap : Max ile min arasdındaki farka bakacağız. 70-30 = 40

Soru: True/False ? 

Cevap : True. İlk resimde sağda büyük bir rakam var ve o rakam dağılımı arttırırıyor. Dolayısıyla sd de artıyor.

Soru: what is the output? 

Cevap : median 30. Sol tarafın medianı yani Q1🡪 26, sağ tarafın medianı yani Q3🡪 33.

IQR : Q3-Q1 = 7

Diyelim ortada 30’dan 2 tane olsaydı. 🡪 24, 25, 26, 27, 30, 30, 33, 33, 33, 102

O zaman soldaki Q1 i hesaplarken soldaki 30’u baz alacaktım. Sağdaki Q3 ü hesaplarken de sağdaki 30’u baz alacaktım.