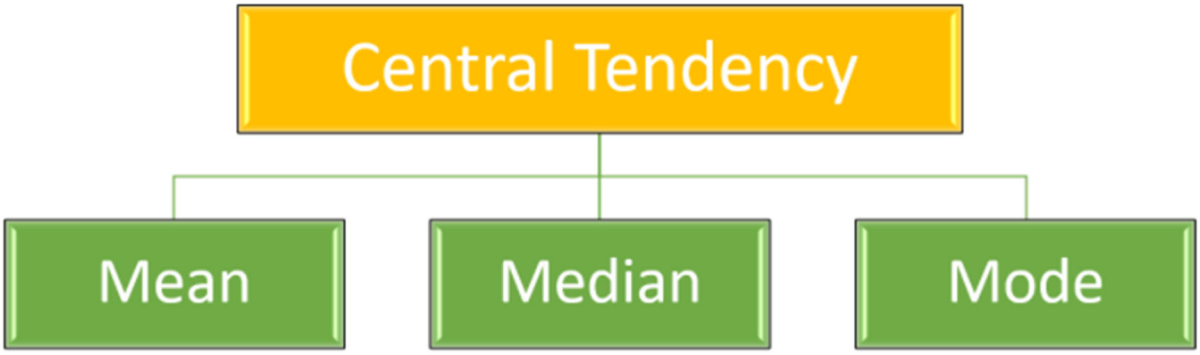
İSTATİSTİK

Fundamentals of Statistics-2

Central Tendency (Measure of Centre) - Merkezi Eğilim (Merkez Ölçüsü)

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



Mean (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.

We understand "mean", "median" and "mode" from the central tendency concept.

Select one:

- (•) True ✔
- ( ) False

Mean (Ortalama)

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

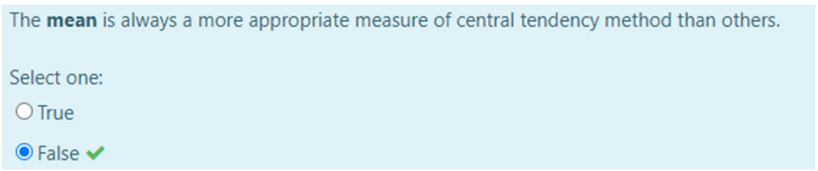
Population Mean	Sample Mean
$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$
N = number of items in the population	n = number of items in the sample

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

Staff	Salary (thousand $)
1	102
2	33
3	26
4	27
5	30
6	25
7	33
8	33
9	24

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.

İpucu:
Çeşitli koşullar altında, bir merkezi eğilim ölçüsü diğerlerinden daha uygun hale gelebilir.



Median

Medyan, küçükten büyüğe sıralanmış bir veri kümesinin orta puanıdır. Aykırı değerler medyanı daha az etkiler. Medyanı hesaplamak için (yukarıdaki tablo ile) aynı verilere sahip olduğumuzu varsayalım. Öncelikle verileri küçükten büyüğe sıralamamız gerekiyor.

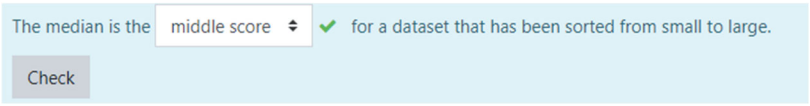
Staff	Salary (thousand $)
1	24
2	25
3	26
4	27
5	30
6	33
7	33
8	33
9	102

Medyan orta skordur, bu durumda 30 bin dolardır. 30 bin dolar orta skor çünkü ondan sonra 4 puan ve ondan önce 4 puan var. Bu, tek sayıda skorlarınız olduğunda işe yarar, ancak çift sayıda örneklem

büyüklüğünüz olduğunda ne olur? Sadece 10 skorun olsa bile mi? Bu durumda, ortadaki iki skoru almamız ve sonucun ortalamasını almamız gerekiyor. Yani, aşağıdaki örneğe bakarsak:

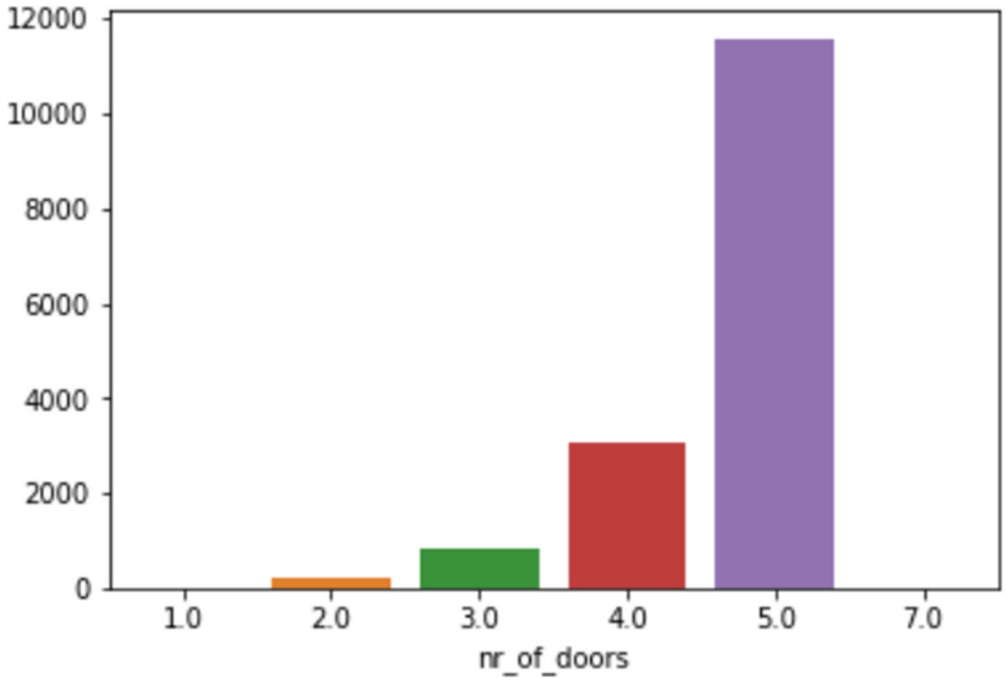
Staff	Salary (thousand $)
1	24
2	25
3	26
4	27
5	30
6	33
7	33
8	33
9	50

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



Mode

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.



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.

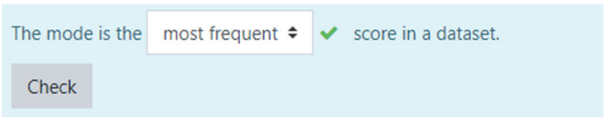
Staff	Salary (thousand $)
1	102
2	33
3	26
4	27
5	30
6	25
7	33
8	33
9	24

Mod, bir veri kümesinde en sık görülen skordur. 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



Python ile Ortalama, Medyan ve Modu hesaplayın

Python ile ortalama, medyan ve mod değerlerini kolayca hesaplayabiliriz. Ortalama ve medyan için numpy kütüphanesini ve mod için baharatlı kütüphaneyi kullanıyoruz. Python ile aldığımız değerleri manuel olarak hesapladığımız değerlerle karşılaştırabilirsiniz.

```
import numpy as np
from scipy import stats
salary = [102, 33, 26, 27, 30, 25, 33, 33, 24]
mean_salary = np.mean(salary)
print("mean:", mean_salary)
median_salary = np.median(salary)
print("median:", median_salary)
mode_salary = stats.mode(salary)
print("mode:", mode_salary)
>>>
mean: 37.0
median: 30.0
mode: ModeResult(mode=array([33]), count=array([3]))
```

Check Yourself Soruları

Which statistical number is the middle number of the data set?

Select one:
- mode
- median ✔ Congrats! You are right.
- mean
- range

Richard took a sample of 100 pieces of data. She added up all of the pieces of data and then divided by 100. What measure of center is she working on?

Select one:
- range
- median
- mean ✔ Congrats! You are right.
- mode

A set of data is given: 3, 7, 10, 10, 16
Jason calculated a measure of center and got 9.2, which measure of center did he just calculate?

Select one:
- mode
- mean ✔ Congrats! You are right.
- range
- median

A set of data is given: 3, 7, 10, 10, 16. What is the mode?

Select one:
- 7
- 16
- 3
- 10 ✔ Congrats! You are right.

A set of data is given: 100, 55, 95, 150, 101, 99, 53, 57, 70. What is the median?

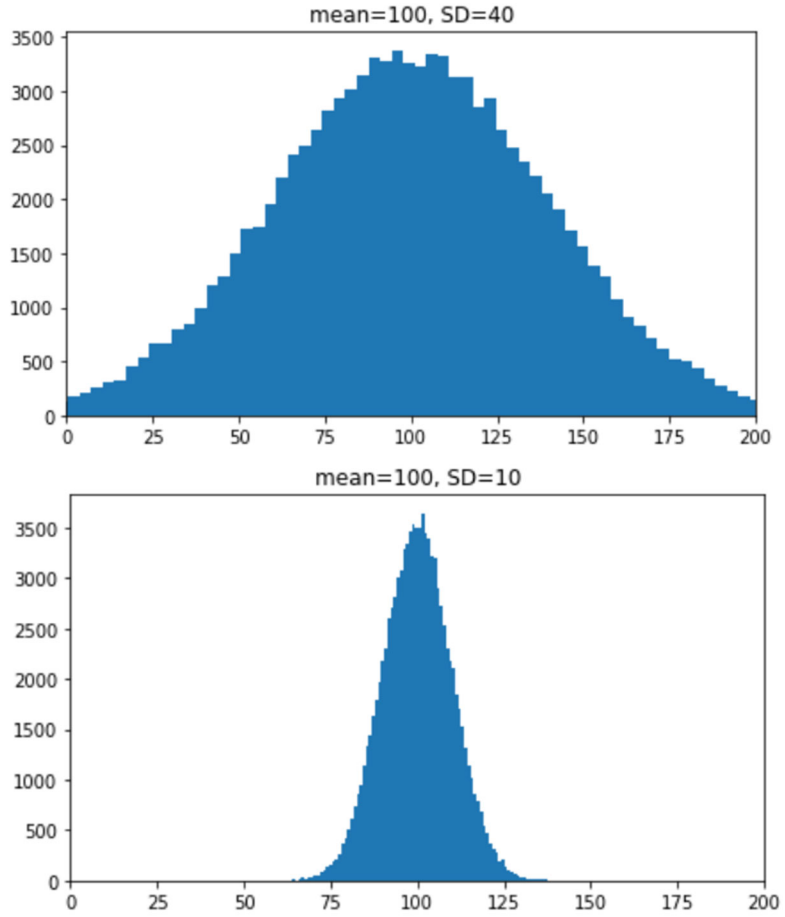
Select one:
- 95 ✔ Congrats! You are right.
- 57
- 100
- 150

Dispersion (Measure of Spread) - Dağılım (Yayılma Ölçüsü)

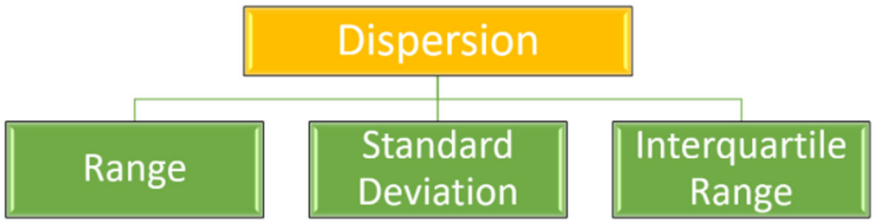
Introduction

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 merkezi eğilimden daha fazlasını açıkladığını söyleyebiliriz.



Aralık (Range), standart sapma (standard deviation) ve çeyrekler arası aralık (interquartile range), yaygın olarak kullanılan üç dağılım ölçüsüdür. Şimdi bu kavramları ele alacağız.



Tips:
A dispersion explains something more than the measure of central tendency does.

Central tendency explains something more than a dispersion does.

Select one:

- ○ True
- ◉ False ✔

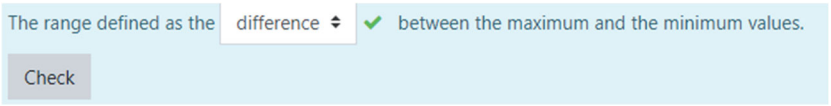
Range (Aralık)

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:

Staff	Salary (thousand $)
1	102
2	33
3	26
4	27
5	30
6	25
7	33
8	33
9	24

$$\text{Range} = \text{MaximumValue} - \text{MinimumValue}$$

Maksimum değer ile minimum değer arasındaki fark 102-24=78'dir. Bu veri seti için aralık 78 diyebiliriz. Örnekte gördüğünüz gibi dokuz değerin maksimum ve minimum iki değeri 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).



Standard Deviation(σ) – (Standart Sapma (σ))

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 ($\sigma 2$) tanımlamalıyız. Varyans, ortalamadan farkların karelerinin ortalaması olarak tanımlanır. Varyans ve standart sapma formülü aşağıda verilmiştir.

Variance= $\sigma^2 = \frac{\sum(x-\mu)^2}{N}$ Standard Deviation= $\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}}$

μ = popülasyon ortalaması,

N = popülasyondaki öğe sayısı

Tekrar maaş tablomuza dönelim ve bu maaşlar için standart sapma ve varyansı hesaplayalım.

Staff	Salary (thousand $)
1	102
2	33
3	26
4	27
5	30
6	25
7	33
8	33
9	24

$$\mu = \frac{24+25+26+27+30+33+33+33+102}{9}$$

$$\mu = \frac{333}{9} = 37$$

$$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}}$$

$$\sigma = \sqrt{\frac{(24-37)^2+(25-37)^2+(26-37)^2+(27-37)^2+(30-37)^2+(33-37)^2+(33-37)^2+(33-37)^2+(102-37)^2}{9}}$$

$$\sigma = \sqrt{\frac{(-13)^2+(-12)^2+(-11)^2+(-10)^2+(-7)^2+(-4)^2+(-4)^2+(-4)^2+(65)^2}{9}}$$

$$\sigma = \sqrt{\frac{169+144+121+100+49+16+16+16+4225}{9}}$$

$$\sigma = \sqrt{\frac{4856}{9}}$$

$$\sigma = \sqrt{539,55}$$

$$\sigma = 23,22833518$$

Aralık (range) 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.

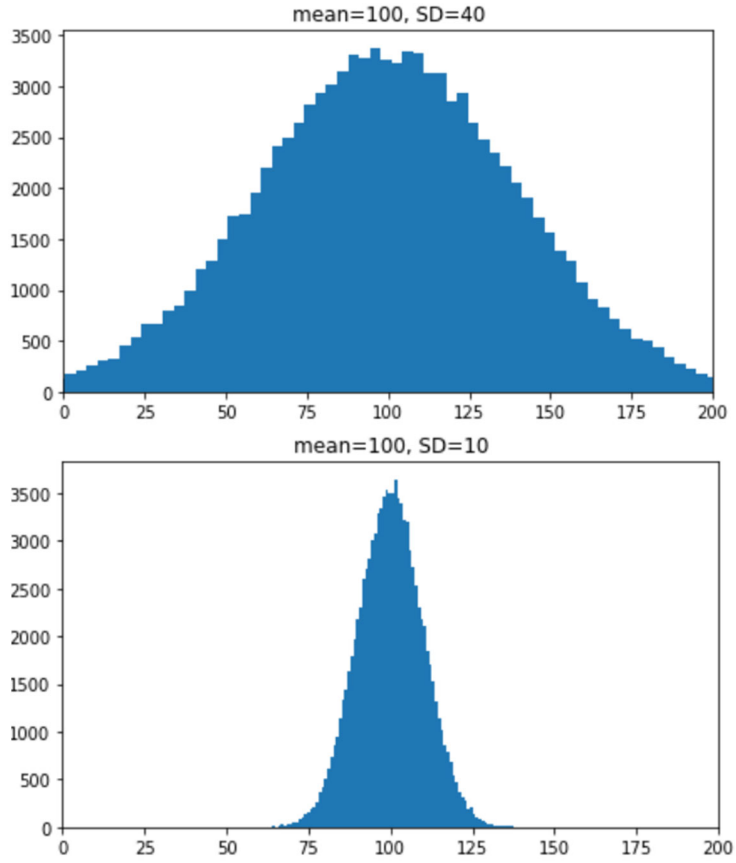
Örneğin aykırı değeri tablomuzdan çıkarırsak maaş tablosu şöyle olur:

Staff	Salary (thousand $)
1	24
2	25
3	26
4	27
5	30
6	33
7	33
8	33

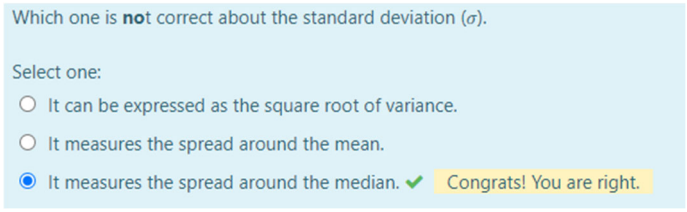
Yeni maaş tablosuna ilişkin standart sapmayı yeniden hesaplarsak: σ=3,58

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, ilk 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.



Python ile Aralık (Range), Varyans ve Standart Sapmayı Hesaplama

Numpy ile aralık, varyans ve standart sapma değerlerini kolayca hesaplayabiliriz. Numpy ile aldığımız değerleri manuel olarak hesapladığımız değerlerle karşılaştırabilirsiniz.

```
import numpy as np
salary = [102, 33, 26, 27, 30, 25, 33, 33, 24]
print("Range: ", (np.max(salary)-np.min(salary)))
print("Variance: ", (np.var(salary)))
print("Std: ", (np.std(salary)))
>>>
Range: 78
Variance:  539.5555555555555
Std:  23.22833518691246
```

Inter Quartile Range (IQR) - Çeyrek Aralığı (IQR)

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.

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 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.

Kategorik veriler yalnızca kadın / erkek cinsiyetler, otomatik / yarı otomatik veya manuel vites kutuları gibi bir dizi olası kategoriyi temsil eden belirli bir değer kümesini alabilir.

Inter Quartile Range(IQR) is the **difference between Q3 and Q1**. In this case:

IQR=Q3-Q1

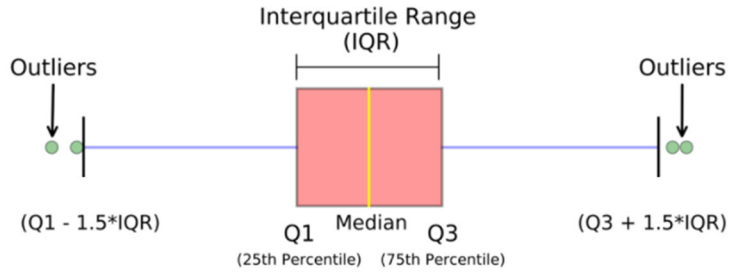
IQR=14-7=7

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)

Aşağıdaki resim, IQR ve aykırı değerler arasındaki ilişkiyi göstermektedir.



Tips:
Outlier is, any data point more than 1.5 IQR below the Q1 or above the Q3.

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.

Select one:
- True ✔
- False

Practice IQR

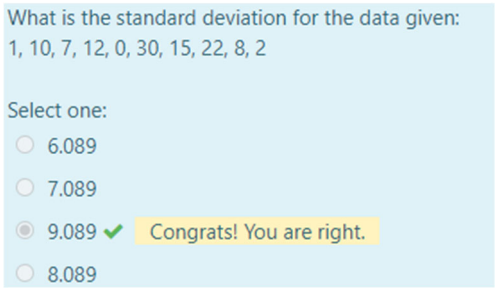
Aşağıdaki numara listesine sahip olduğumuzu hayal edin.
number_list = [1, 5, 10, 15, 40]
Şimdi listemizdeki hangi sayıların aykırı olduğunu bulmaya çalışacağız.
Aşağıdaki özet bilgilere sahibiz:
minimum number = 1
maximum number =40
median=10
Q1 = 5
Q3 = 15
IQR = Q3-Q1
IQR= 15-5 = 10
Bu nedenle, (1.5 * IQR) = 15
Aykırı değerlerin olup olmadığını belirlemek için çeyreklerin ötesinde 1,5*IQR olan sayıları dikkate almalıyız.
Q1 – (1.5 * IQR) = 5-15 = -10
Q3 + (1.5 * IQR) = 15+15 = 30
Listemizdeki son sayı 40'tır. Ve (–10) ile (30) arasındaki aralığın dışındadır, bu nedenle 40 bir aykırı değerdir. Listedeki sayıların geri kalanı aykırı değildir.

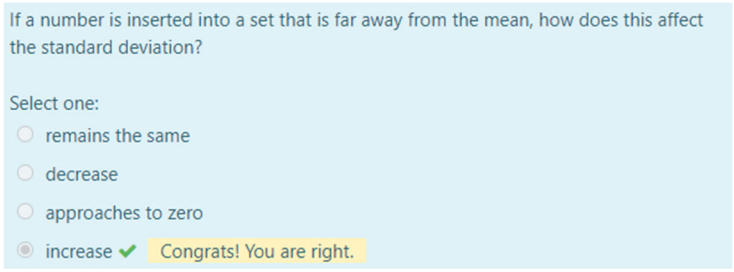
Tamamlayıcı ders videosu https://youtu.be/mk8tOD0t8M0

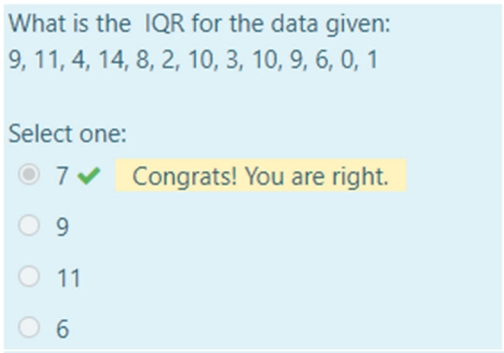
Check yourself

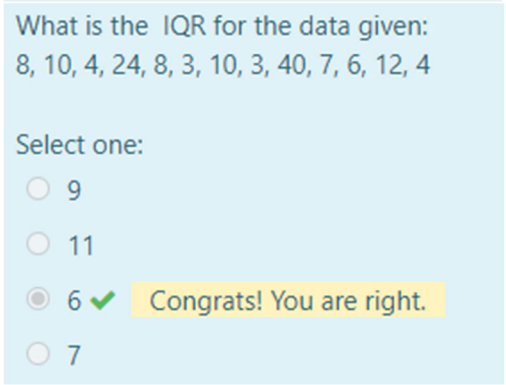
What is the range for the data given:
1, 10, 7, 12, 0, 30, 15, 22, 8, 2

Select one:
- 0
- 22
- 1
- 30 ✔ Congrats! You are right.









Scatter Plot & Box Plot

Scatter Plot

Korelasyon tartışmasına başlamadan önce, iki değişken x ve y arasındaki ilişkiyi göstermenin bir yolunu incelememiz gerekiyor. En yaygın ve en kolay yol bir dağılım grafiğidir.

Bir dağılım grafiği, değişkenler arasındaki ilişkinin yönünü gösterir. Aşağıdakilerden biri olduğunda net bir yön olur:

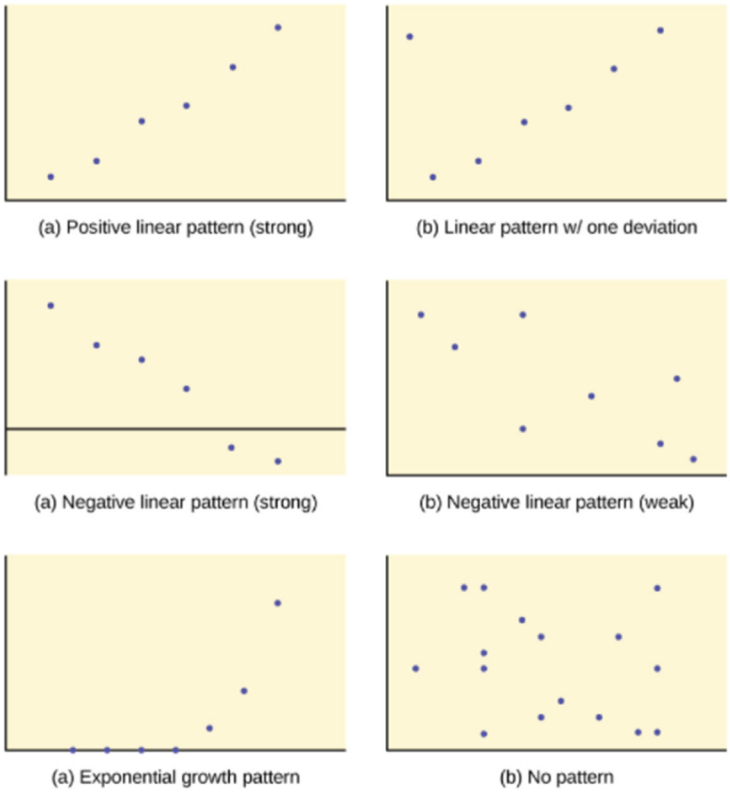
Bir değişkenin yüksek değerlerinin diğer değişkenin yüksek değerleriyle ortaya çıkması veya bir değişkenin düşük değerlerinin diğer değişkenin düşük değerleriyle ortaya çıkması.

Bir değişkenin yüksek değerleri ile diğer değişkenin düşük değerlerinin ortaya çıkması.

Scatter Plot
A scatter plot of two variables shows the values of one variable on the Y axis and the values of the other variable on the X axis.

Dağılım grafiğine bakarak ve noktaların bir çizgiye, bir güç işlevine, üstel bir işleve veya başka bir işlev türüne ne kadar yakın olduğunu görerek ilişkinin gücünü belirleyebilirsiniz. Doğrusal bir ilişki için bir istisna vardır. Tüm noktaların "mükemmel uyum" sağlayan yatay bir çizgi üzerine düştüğü bir dağılım grafiğini düşünün. Yatay çizgi aslında hiçbir ilişki göstermez.

Bir dağılım grafiğine baktığınızda, genel deseni ve desenden sapmaları fark etmek istersiniz. Aşağıdaki dağılım grafiği örnekleri bu kavramları göstermektedir.



Scatter plots are well suited for revealing the relationship between two variables.

Select one:
- (●) True ✔
- ( ) False

Box Plot

Kutu çizimleri (**box-and-whisker plots** veya **box-whisker plots** olarak da adlandırılır), veri konsantrasyonunun iyi bir grafik görüntüsünü verir. Ayrıca uç değerlerin çoğu veriden ne kadar uzakta olduğunu da gösterirler. Beş değerden bir kutu grafiği oluşturulur: minimum değer, ilk çeyrek, medyan, üçüncü çeyrek ve maksimum değer. Bu değerleri, diğer veri değerlerinin onlara ne kadar yakın olduğunu karşılaştırmak için kullanırız.

Box Plot
One of the more effective graphical summaries of a data set, the box plot generally shows mean, median, 25th and 75th percentiles, and outliers.

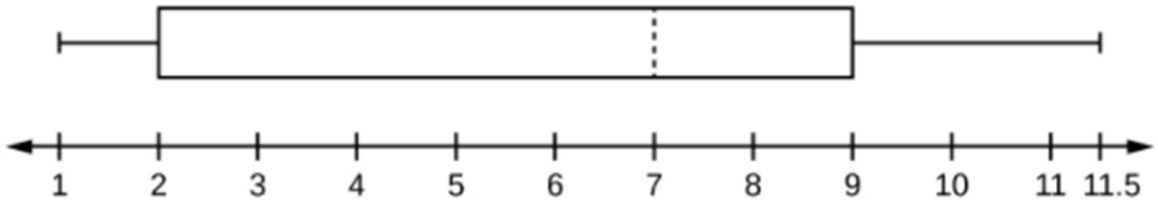
Bir kutu grafiği oluşturmak için yatay veya dikey bir sayı doğrusu ve dikdörtgen bir kutu kullanın. En küçük ve en büyük veri değerleri, eksenin uç noktalarını etiketler. İlk çeyrek kutunun bir ucunu ve üçüncü çeyrek kutunun diğer ucunu gösterir. Verilerin yaklaşık yüzde 50'si kutunun içine düşüyor. "Bıyıklar", kutunun uçlarından en küçük ve en büyük veri değerlerine kadar uzanır. Medyan veya ikinci çeyrek, birinci ve üçüncü çeyrekler arasında olabilir veya biri, diğeri veya her ikisi olabilir. Kutu grafiği, verilerin iyi ve hızlı bir resmini verir.

İpuçları:
Aykırı değerleri işaretleyen noktalara sahip box-and-whisker plots'lar ile karşılaşabilirsiniz. Bu durumlarda, whiskers'lar minimum ve maksimum değerlere uzanmaz.

Aşağıdaki veri kümesini göz önünde bulundurun.

1	1	2	2	4	6.8	7	8	8.3	9	10	10	11.5

İlk çeyrek iki, medyan yedi ve üçüncü çeyrek dokuz. En küçük değer bir, en büyük değer 11.5'tir. Aşağıdaki görüntü, constructed box plot göstermektedir.



İki bıyık (whisker), ilk çeyrekten en küçük değere ve üçüncü çeyrekten en büyük değere kadar uzanır. Medyan kesikli bir çizgi ile gösterilir.

The '*middle*' line drawn inside the box shows the position of the median value ✔ .

The *ends of the 'box'* give the positions of the first and third quartiles ✔ .

The *ends of the 'whiskers'* give the minimum and maximum values ✔ in the data.