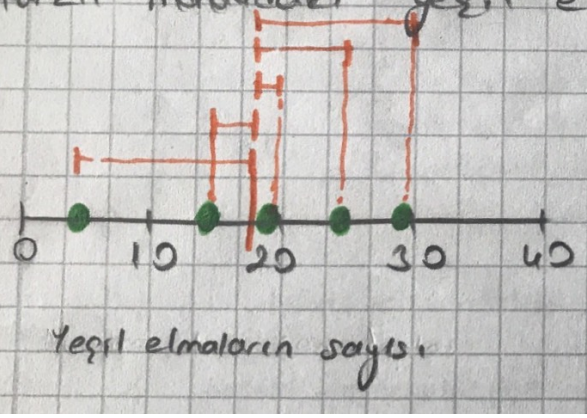
Youtube - Statquest with Josh Starmer

Covarrance, Clearly Explained!!!

Varyans Hatırlatması

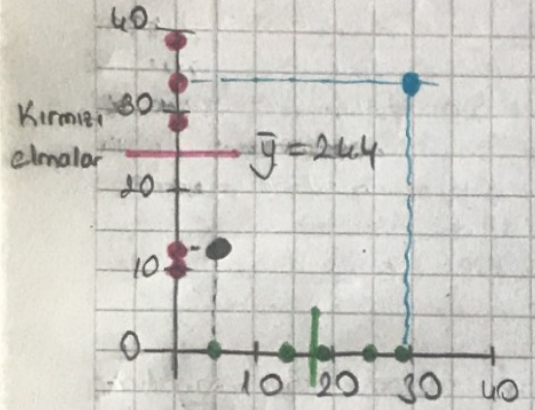
5 farklı marketdaki yeşil elmaları saydığımızı düşünelim.



Estimated mean : $\bar{x}$

Varyans : $\frac{\Sigma (x-\bar{x})^2}{n-1} = 101.8$

Yeşil elmalara ek olarak aynı manavda kırmızı elmaları da saydığımızı düşünelim.

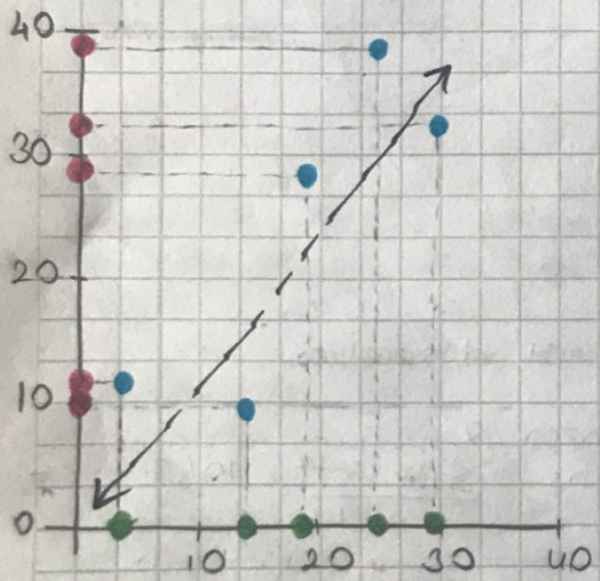


Kırmızı elmalar için varyans: $\frac{\Sigma (y-\bar{y})^2}{n-1} = 160.3$

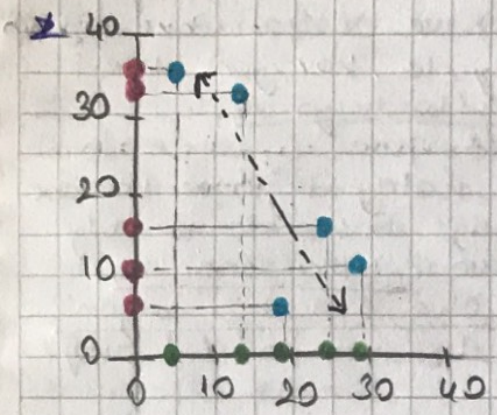
Bu ölçümler aynı manavdan alındıkları için bunları çiftler olarak (pair) düşünebiliriz.

Örneğin; soyah ile ölçtüğüm ölçümler aynı manavdan geliyor. Her ikisi de kendi ortalama değerlerinden daha az. Mavi ile ölçtüğüm de aynı manavdan geliyor çünkü, her ikisi de ortalamadan daha büyük. Bu ölçümler çiftler halinde yapıldığına göre sorumuz şu; "İkili olarak alınan ölçümler bize bireysel ölçümlerin söylemediği bir şey mi söylüyor?" Covariance bu soruyu cevaplamanın bir yoludur.

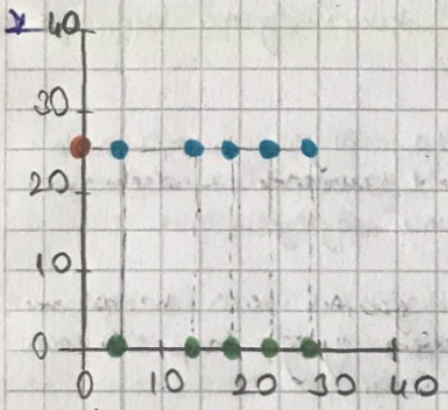
Bu ölçümler aynı manavdan geldiğine göre bunları birleştirip noktalar halinde ifade edebiliriz.



Genel konuşacak olursak X ekseninde düşük değere sahip olan değerler Y ekseninde düşük değere sahip, X ekseninde büyük değere sahip olan değerler de Y ekseninde büyük değere sahip. Ortadaki line ile bu ilişkiyi ifade edebiliriz. Bu çizgi pozitif eğimi ifade ediyor. Yani X artarken Y de birlikte artıyor.



Verimiz bu şekildeyse; düşük X değerleri yüksek Y değerleriyle ilişkili, yüksek X değerleri ise düşük Y değerleriyle ilişkili. O zaman çizilecek eğri negatif bir eğime sahip olacaktır. Negatif bir trend olacaktır aralarında. X'in değeri artarken Y'nin değeri azalacaktır.



Verimiz bu şekilde de olabilir. Bu durumda X'in her değeri için Y aynı değere sahip. Yani, pozitif ya da negatif X ile Y arasında herhangi bir trend yok. Y'nin ölçüm değerinin artı olduğu manada hangi X değeri vardır bilemeyiz bu durumda.

Benzer bir grafiği sabit bir X, değişken Y değerleri için çizdiğimizde de yaşarız. 10 kırmızı elma saydıysak aynı manada kaç yeşil elma vardır bilemeyiz.

Kovaryansın arkasındaki ana fikir, üç tür ilişkiyi sınıflandırabilmektir. Pozitif trendleri, negatif trendleri, trendsizliği.

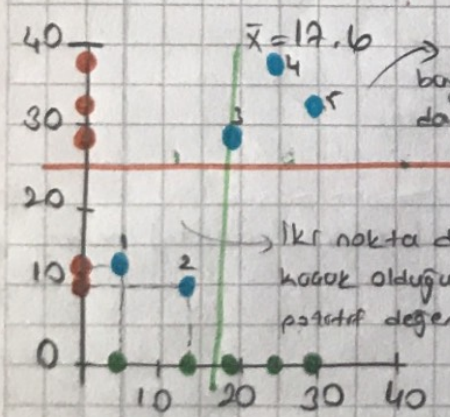
Kovaryans, kendi başına ilginç değildir. Hiçbir zaman kovaryansı hesaplamayacağız. Kovaryans, daha ilginç bir şeyin, korelasyonun bir hesaplama basamağıdır.

Kovaryansın Hesaplanması

$$\frac{\Sigma (x-\bar{x})(y-\bar{y})}{n-1}$$

1. case

Pozitif trend olan bir grafikte hesaplama yapacak olursak;



Üç nokta da ortalamadan büyük olduğu için bu kutucuktan da pozitif değerler elde edeceğiz.

İki nokta da ortalamalardan küçük olduğu için (-).(-) = (+) pozitif değer elde edeceğiz bu kutucuktan.

$$\frac{\underbrace{(3-17.6)\times(12-24.4)}_{181} + \underbrace{(-4.6)\times(-14.4)}_{66.2} + \underbrace{1.4\times4.6}_{6.4} + 55 + 155}{5-1} \rightarrow \text{pay}$$

$$= 116 \rightarrow \text{payda}$$

11b, yani pozitif bir değer bulduğumuza ram X ve Y arasındaki ilişkinin pozitif olduğu konusunda yorumuz. Başka bir deyişle, kovaryans değeri pozitif olduğu için trendi "pozitif" olarak sınıflandırıyoruz.

* Kovaryans değerinin kendisini yorumlamak çok kolay değil ve kontekse bağlı. Kovaryans değeri bize ilişkinin dik olup olmadığı hakkında bir bilgi vermez. Sadece eğrinin pozitif olduğu bilgisini verir. Daha önemlisi; kovaryans değeri bize noktaların line'a yakın olup olmadığı bilgisini de vermez. Pozitif kovaryans değeri bize sadece ilişkinin pozitif olduğu bilgisini verir.

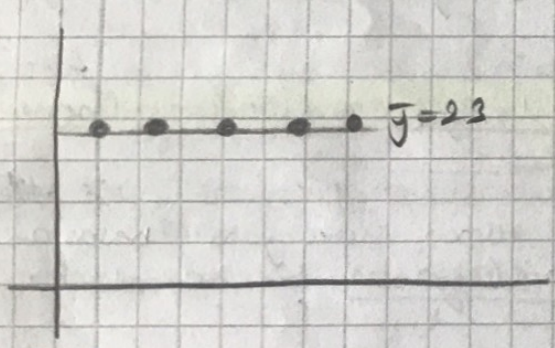
* Kovaryansı yorumlamak zor olsa da bizim için daha ilginç bir değer korelasyonun bir hesaplama basamağı.

2. case

* X ile Y arasında negatif ilişki olan bir grafikte aynı hesaplamaları yaptığımızda sonuç -105.15 çıktı. Buradan X ve Y arasındaki ilişkinin ve çizilecek olan eğrinin eğrinin negatif olduğunu söyleyebiliriz.

3. case

* Farklı X değerleri için aynı Y değerlerini veren grafik için hesaplama yapsaydık $\bar{y}$, kesişim noktalarının üzerinde olacağından $(y-\bar{y})$ hep 0 çıkacaktı, dolayısıyla sonuç da 0 çıkacaktı.

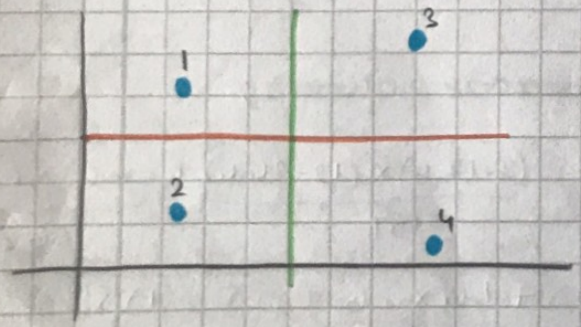


$$\frac{\Sigma (x-\bar{x})\cdot(y-\bar{y})}{n-1} = \frac{(-14,6 \times 0) + 0+0+0+0}{4}$$

$$= 0$$

4. case

* Verilerimiz öyle bir yerleşebilirdi ki pozitif ve negatif değerler birbirini götürüp 0 elde edebilirdik.



$$\frac{\Sigma (x-\bar{x})\cdot(y-\bar{y})}{n-1} = \frac{-100+100+150-150}{4-1}$$

$$= 0$$

Son iki örnekte X ve Y arasında (yeşil ve kırmızı elmalar arasında) herhangi bir ilişki olmadığını anladık.

Neden Kovaryansı yorumlamak zordur?

✳ Yeşil elmaların yine yeşil elmalarla olan kovaryansını hesaplayalım. Ortalama yine 17.6 olacaktır.

$\frac{\Sigma (x-\bar{x}).(y-\bar{y})}{n-1}$ formülünde $y=x$, $\bar{y}=\bar{x}$ olacaktır.

Dolayısıyla formül;

$\frac{\Sigma(x-\bar{x}).(x-\bar{x})}{n-1} = \frac{\Sigma(x-\bar{x})^2}{n-1}$ estimating variyans formülüne dönecektir.

Bu işlemi yaptığımızda 102 sonucunu elde ettik. Bu kovaryans değeri pozitif olduğu için x'in kendisiyle olan ilişkisinin pozitif olduğunu söylüyoruz.

✳ Yeşil elma ölçümlerini 2 katına çıkardığımızda ve bu hesaplamaları tekrar yaptığımızda 408 bulacağız kovaryans değerini. Aradaki ilişki değişmemesine rağmen (kendisiyle olan kovaryansı hesaplıyoruz) kovaryans değeri 4 kat arttı.

Yani; kovaryans değeri verinin scale'ine, büyüklüğüne karşı duyarlı, bu da bizim onu yorumlamamızı zorlaştırıyor.

Aynı şekilde ilişkinin noktalara yakınlığı ve uzaklığı ile kovaryans değeri arasında bir ilişki bulunmuyor.

Scale'e duyarlı olmayan bir şey hesaplamak istersek Korelasyon imdadımıza yetişiyor.

✳ Kovaryans değeri korelasyonun bir hesaplama adımı olduğu gibi Principal Component Analysis (PCA)'nın da bir hesaplama adımı.