İSTATİSTİK

Ortalama, Varyans ve Standart Sapma

Ortalama

Ortalama, bir veri setindeki tüm verilerin (sayıların) toplamının veri sayısına bölümüdür. **μ** sembolü ile gösterilir. Hesaplanması aşağıda gösterilmiştir:

$$\mu = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Örneğin, 5 öğrencinin notları 60, 80, 90, 100 ve 70 ise bu veri setinin, yani öğrencilerin notlarının, ortalaması:
 (60 + 80 + 90 + 100 + 70) / 5 = 80 olarak bulunur.

Varyans

Varyans, bir veri setindeki tüm verilerin, veri setinin ortalamasına olan uzaklıklarının ortalamasıdır. σ^2 sembolü, yani standart sapmanın karesi, ile gösterilir. Varyans, verilerin ne kadar birbirinden uzak yanı dağılmış olduklarını ölçer. Hesaplanırken önce ortalama bulunur, sonra tüm verilerin ortalama ile olan farklarının kareleri alınarak toplanır ve çıkan sayı toplam veri sayısına bölünür. Hesaplanması aşağıda gösterilmiştir:

$$\sigma^2 = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_n - \mu)^2}{n}$$

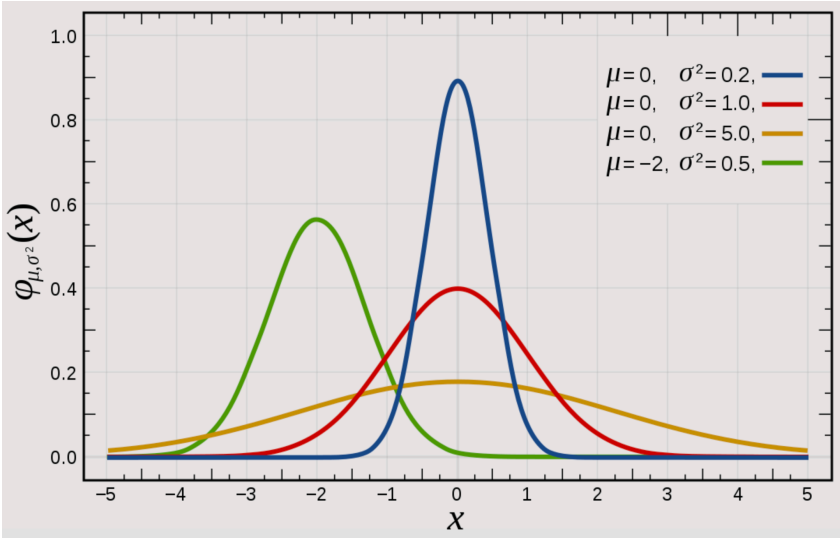
Hadi öğrencilerin notlarının varyansını hesaplayalım! Ortalamayı az önce 80 olarak bulduk. Şimdi tüm sayıların ortalama ile olan farklarını hesaplayalım: *60 - 80 = -20, 80 - 80 = 0, 90 - 80 = 10, 100 - 80 = 20 ve 70 - 80 = -10.* Farkları -20, 0, 10, 20 ve -10 olarak bulduk, şimdi bu farkların karesini alalım: *(-20)^2 = 400, 0^2 = 0, 10^2 = 100, 20^2 = 400 ve (-10)^2 = 100.* Farkların karelerini de 400, 0, 100, 400 ve 100 olarak bulduk. Fark edersek farkların karelerini aldığımız zaman sayılar negatif olmaktan çıktı. Bu bize uzaklık bilgisini, yani negatif olamayan bilgiyi sağladı. Bulduğumuz kareleri toplayalım: *400 + 0 + 100 + 400 + 100 = 1000.* Bu sayıyı da toplam veri sayısına bölelim: *1000 / 5 = 200.* Evet! Bu dağılımın varyansı 200. Eğer öğrencilerin notları 70, 75, 80, 85 ve 90 olsaydı varyans kaç çıkardı? İsterseniz kendiniz hesaplayın, sonra devam edelim. Bu notların varyansını hesapladığımızda sonuç 50 çıkacaktır. Gördüğümüz gibi veriler birbirine daha yakın olduğunda varyans daha az olmakta.

Standart Sapma

Standart sapma, varyansın karekökü dür. Peki neden? Varyansı hesaplarken farkların karesini aldık. Peki karelerini aldıktan sonra karekökünü almak kulağa hoş gelmiyor mu? Farkların karelerini aldıktan sonra karekök

alınarak sayı tekrar aynı boyuta döndürülür ve bu işlem de bize yine verilerin birbirinden ne kadar uzak olduğunu gösteren standart sapmayı verir. Hesaplanması aşağıda gösterilmiştir:

$$\sigma = \sqrt{\frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_n - \mu)^2}{n}}$$



Yukarıdaki grafikte farklı ortalama ve standart sapmalara sahip normal dağılımlar gösterilmekte. Dağılımların ortalama değerleri (μ sembolü ile gösterilmekte) gördüğümüz gibi normal dağılımların merkezleri, yani tepe noktalarıdır. Varyanslarına (σ^2 sembolü ile gösterilmekte) bakacak olursak varyansı fazla olanların

daha geniş ve yayılmış olduğunu, varyansı az olanların ise daha dar ve keskin olduğunu görürüz.

Hipotez Testi ve Null Hipotez

Hipotez Nedir?

Hipotez, bir araştırmadan önce yapılan tahmin, ileri sürülen iddiadır. Örneğin, bir madeni para atıldığında %50 ihtimalle tura ve %50 ihtimalle yazı gelip gelmediğini araştırmak istiyorsak hipotezimiz "madeni para atıldığında %50 ihtimalle tura ve %50 ihtimalle yazı gelir." olabilir. Ve bu hipotezin tersi, yani "madeni para atıldığında %50 ihtimalle tura ve %50 ihtimalle yazı gelmez." ifadesi de bir hipotezdir.

Hipotez Testi

Hipotez testi ise yapılan tahminin yani hipotezin doğru olup olmadığının test edilmesidir. Bir hipotez testinde iki tane birbirine zıt olan hipotez bulunur. Bu hipotezlerden biri reddedilirse diğeri doğru kabul edilir.

Sıfır Hipotezi ve Alternatif Hipotez

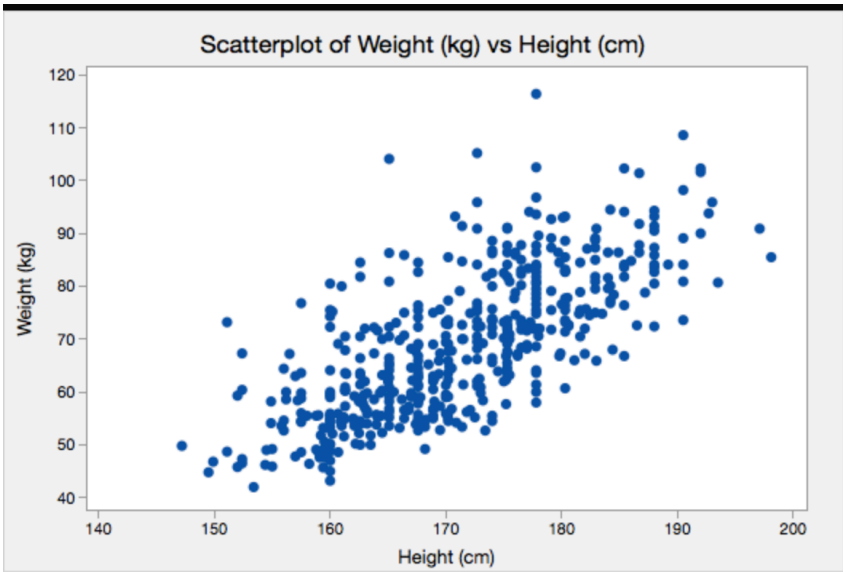
Sıfır hipotezi (null hipotez), test edilen iki grubun arasındaki farkın önemli olmadığını savunur. Örneğin, A sınıfı ve B sınıfı adında iki sınıf olduğunu düşünelim. Bu sınıflardaki öğrencilerin not ortalamalarının farklı olup olmadığını test etmek isteyelim. Bu durumda sıfır hipotezi "A sınıfının ve B sınıfının not ortalamalarının arasında bir fark yoktur." olur. **Alternatif hipotez** ise sıfır hipotezinin tersidir. Yani bu durumda alternatif hipotez, "A sınıfının ve B sınıfının not ortalamalarının arasında fark vardır." olur.
A sınıfının ortalaması 60, B sınıfının ortalaması 80 ise arada fark olduğu açıktır. Bu yüzden sıfır hipotezi reddedilir ve alternatif hipotez kabul edilir. Ancak A sınıfının ortalaması 77, B sınıfının ortalaması 78 ise arada pek fark yoktur ve bu fark şansa bağlı olarak kabul edilir. Bu sıfır hipotezi reddedilemez ve kabul edilir.

p-Değeri

p-değeri, 0 ile 1 arasında olan ve bir hipotezin güvenilir ve doğru olup olmadığını ölçmemize yardım eden bir sayıdır. Hesapladığımız p-değeri ne kadar küçük çıkarsa iki grup arasında fark olduğunu o kadar güvenli bir şekilde söyleyebilir ve sıfır hipotezini reddedebiliriz. Örneğin, A ve B sınıfının not ortalamalarını karşılaştırdığımız hipotezlerimizi hatırlayalım. Sıfır hipotezi, ortalamaları arasında fark yoktur yani birbirlerine çok yakınlardır diyordu. Alternatif hipotez ise fark vardır diyordu. Bu hipotezlerin hangisinin doğru olduğunu bulmak için öncelikle bir p-değeri sınırı belirlenir. Genellikle bu sınır 0.05 yani %5 olarak belirlenir. Hesaplanan p-değeri bu sayıdan küçük çıkarsa ancak o zaman sıfır hipotezi reddedilebilir. Sonra, ortalamalar hesaplanır. Bundan sonra p-değeri hesaplanır ve sıfır hipotezinin reddedilip edilemeyeceğine bakılır. P-değeri sınır sayıdan küçük ise sıfır hipotezi reddedilir ve bu gruplar arasında fark vardır denilir. Ancak p-değeri sınır sayıdan büyük ise sıfır hipotezi reddedilemez ve bu gruplar arasında fark yoktur, benzerlerdir denilir.

Kovaryans

Kovaryans, iki veri kümesinin birbiriyle olan ilişkisini anlamamıza yarayan bir ölçümdür. Önce iki veri kümesi derken ne demek istediğimize bakalım. Aşağıda bir grup insanın boy ve kiloları grafikte gösterilmiş:

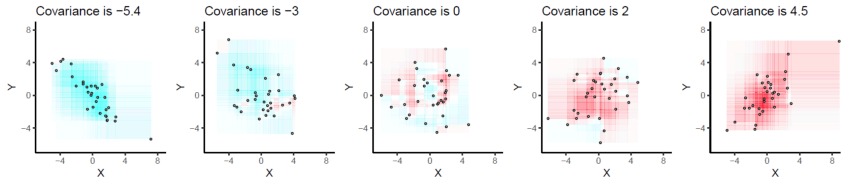


Bu grafikte her nokta bir insana karşılık geliyor ve her insanın yani noktanın boy ve ağırlık değerleri var. Grafiğe baktığınızda boy ve ağırlıkların arasında bir ilişki görebiliyor musunuz? Evet, genelde boyu fazla olan insanların ağırlığı da fazla, boyu az olan insanların ağırlığı da az oluyor. Yani boy ve ağırlık arasında **pozitif**(biri arttıkça diğeri de artan) bir ilişki var. Biri artarken diğeri azalsaydı **negatif** bir ilişli olacaktı.

Birinin artması ya da azalması diğerini etkilemiyor olsaydı da aralarında bir ilişki olmayacaktı. Kovaryansa geri dönelim. Kovaryans hesaplamadan önce iki veri kümesinin de ortalaması hesaplanır. Kovaryans, iki veri kümesindeki her bir verinin ortalamaları ile olan farklarının çarpımının toplanması ve bu sayının toplam veri sayısına bölünmesi ile hesaplanır. Aşağıda formül olarak gösterilmiştir:

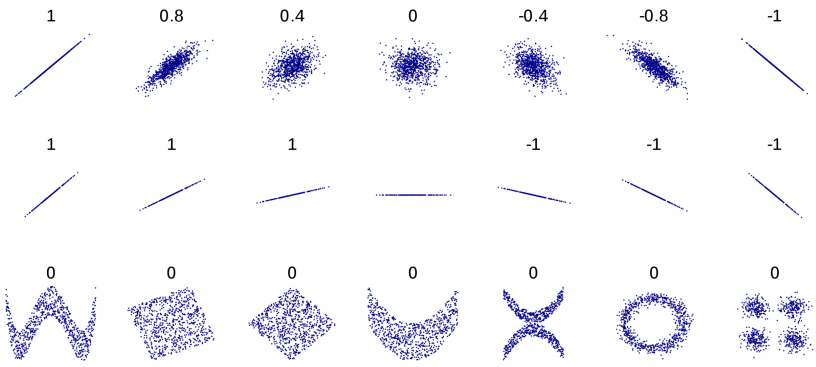
$$cov(X, Y) = \frac{(x_1 - \mu_x).(y_1 - \mu_y) + (x_2 - \mu_x).(y_2 - \mu_y) + \cdots + (x_n - \mu_x).(y_n - \mu_y)}{n}$$

Aslında formüle de baktığımızda kovaryansın, iki veri kümesinin ortalamalarından olan sapmalarının çarpımını hesapladığını görürüz. Bu da bize aralarındaki ilişki ile ilgili bilgi verir. Aşağıda farklı veri kümelerinin kovaryansları verilmiştir. Kovaryans negatif ise ilişkinin de negatif (sol kısım), pozitif ise ilişkinin de pozitif (sağ kısım), sıfıra yakın ise bir ilişkinin olmadığını (orta kısım) görebiliyoruz.



Korelasyon

Korelasyon da kovaryans gibi iki veri kümesinin birbiriyle olan ilişkisini gösteren bir ölçümdür. Ancak korelasyon, kovaryans gibi sadece ilişkinin pozitif mi negatif mi olduğunu göstermez, ilişkinin ne kadar güçlü olduğunu da gösterir. Korelasyon her zaman -1 ve 1 sayıları arasında olduğu için ilişkinin ne kadar güçlü olduğunu kolayca anlayabiliriz ve farklı ilişkileri karşılaştırabiliriz. Korelasyon 1'e veya -1'e ne kadar yakınsa o kadar güçlüdür, 0'a yaklaştığında ise zayıflar. Korelasyon pozitifse ilişki pozitif, negatifse ilişki negatiftir. Aşağıda bazı korelasyon örnekleri var:



Korelasyonun formülü ise kovaryans ve standart sapmayı bildiğimiz zaman bayağı basitleşiyor. Korelasyon, iki veri kümesinin kovaryansının, varyanslarının çarpımına bölümüdür. r sembolü ile gösterilir. Formül olarak aşağıda gösterilmiştir:

$$r = \frac{cov(X, Y)}{\sigma_x . \sigma_y}$$

Koşullu Olasılıklar

Olasılık

Olasılık, bir olayın sonucunda ortaya çıkabilecek sonuçların ihtimallerini gösteren bir ölçüdür. "p" sembolü ile gösterilir. Olasılık, 0 ile 1 arasında olur. Bir sonucun olasılığı 0 ise görülmesi imkansız, 1 ise görülmesi kesindir. Örneğin, bir madeni para havaya

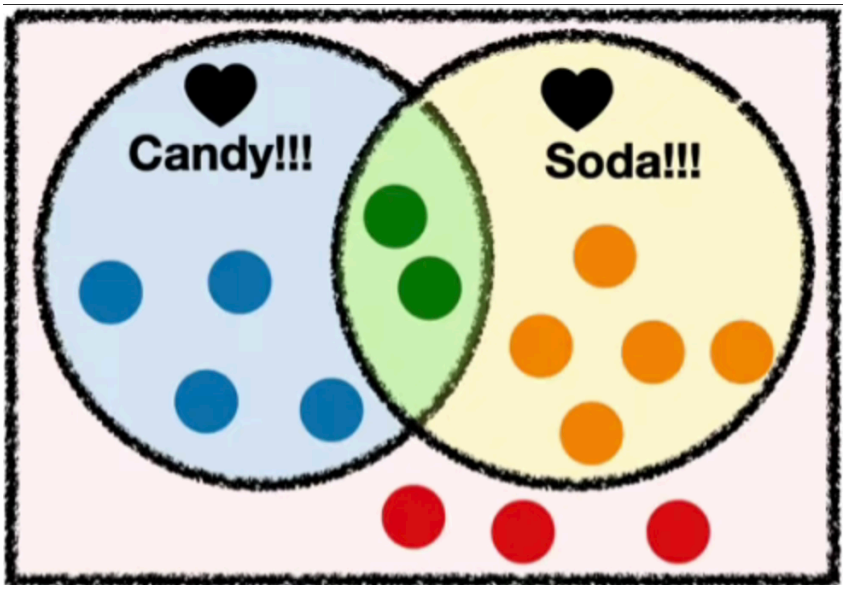
atıldığında 2 sonuç ortaya çıkabilir, yazı ya da tura. Eğer para normal bir para ise yazı çıkma olasılığı 0.5 yani %50, tura çıkma olasılığı da 0.5 yani %50'dir. Hem yazı hem de tura gelme olasılığı 0 yani %0'dır çünkü sonuç ya yazı ya da tura olabilir. Olasılık, **p(sonuç)** olarak gösterilir. Örneğin, para atıldığında yazı gelme olasılığı "p(yazı)" şeklinde gösterilebilir

Koşullu Olasılık

Koşullu olasılık, bir koşulun gerçekleştiği bilindikten sonra başka bir koşulun gerçekleşme olasılığıdır. Örneğin, 6 yüzlü bir zar atıldığında her bir sayının gelme olasılığı 1/6'dır. Çünkü gelebilecek sayılar 6 tanedir (1, 2, 3, 4, 5, 6) ve her bir sayının gelme olasılığı eşittir. Örneğin, p(1) = 1/6 ya da p(2) = 1/6. Peki ben gelen sayının çift olduğunu biliyorsam ne olur? Bu sefer gelebilecek sayılar 3 tanedir, yani 2, 4 ve 6. Yani gelen sayının çift olduğu biliniyorsa 2 gelme olasılığı nedir? Gelebilecek 3 sayı var, her sayının gelme olasılığı eşit ve 2 de bu sayılardan biri. O zaman bu olasılık 1/3 olur. Koşullu olasılık **p(sonuç|koşul)** olarak gösterilir. Örneğin, gelen sayının çift olduğu bilindiğinde 2 gelme olasılığı yani p(2|çift) = 1/3'tür. Koşullu olasılık formülü aşağıda gösterilmiştir. Yani B koşulu varken A'nın olma olasılığı, A ve B'nin birlikte olma olasılığının B'nin olma olasılığına bölümüdür.

$$p(A|B) = \frac{p(A \cap B)}{p(B)}$$

Aşağıda bir venn şeması var. Bu şemadan hareketle bazı olasılıkları hesaplayalım:



Şemada soldaki dairede şeker (candy) sevenler, sağdaki dairede soda sevenler gösterilmiş. Öncelikle toplam kişi sayısına bakalım: Toplamda 14 kişi var. p(soda) yani bir kişinin soda sevme olasılığı kaçtır? Soda dairesinin içinde toplamda 7 kişi var. O zaman, soda seven sayısı / toplam sayı = 7/14 yani 1/2. Bir kişinin soda sevme

olasılığı %50'ymiş. Bir kişinin hem soda hem şeker sevme olasılığı kaçtır? Soda ve şeker dairesinin kesişiminde 2 kişi var. O zaman, p(soda ve şeker) = 2/14 = 1/7 eder.
Koşullu olasılıklara da bakalım. Örneğin, soda seven birinin şeker de sevme olasılığı kaçtır? Burada koşulumuz "soda" ve sonucumuz da "soda ve şeker". Yani p(soda ve şeker|soda) kısaca p(şeker|soda), p(soda ve şeker) / p(soda) hesabıyla bulunur. Hesaplarsak p(şeker|soda) = 2/7 olarak bulunur.

Bayes Teoremi

Bayes Teoremi

Bayes Teoremi, koşullu olasılıklarla türetilmiş bir ifadedir. Kendisi bilinmeyen ancak tersi bilinen bir koşullu olasılıktan kendisine ulaşmamızı sağlar. Formül aşağıda gösterilmiştir:

$$p(A|B) = \frac{p(B|A).p(A)}{p(B)}$$

Formülün nasıl elde edildiğine bakalım. Aşağıdaki formülü yani koşullu olasılık formülünü biliyoruz. Buna formül 1 diyelim:

$$p(A|B) = \frac{p(A \cap B)}{p(B)}$$

Bu koşullu olasılığın tersini de alalım. Yani A koşulu varken B'nin olma olasılığı. Buna da formül 2 diyelim:

$$p(B|A) = \frac{p(A \cap B)}{p(A)}$$

İki formülde de p(A∩B) var. Formül 2'de p(A∩B) değerini yalnız bırakalım:

$$p(A \cap B) = p(B|A).p(A)$$

Burada bulduğumuz $p(A \cap B)$ değerine eşit olan değeri formül 1'de $p(A \cap B)$'nin yerine koyarsak en üstte gösterdiğimiz Bayes Teoremi formülünü elde ederiz.