İstatistik

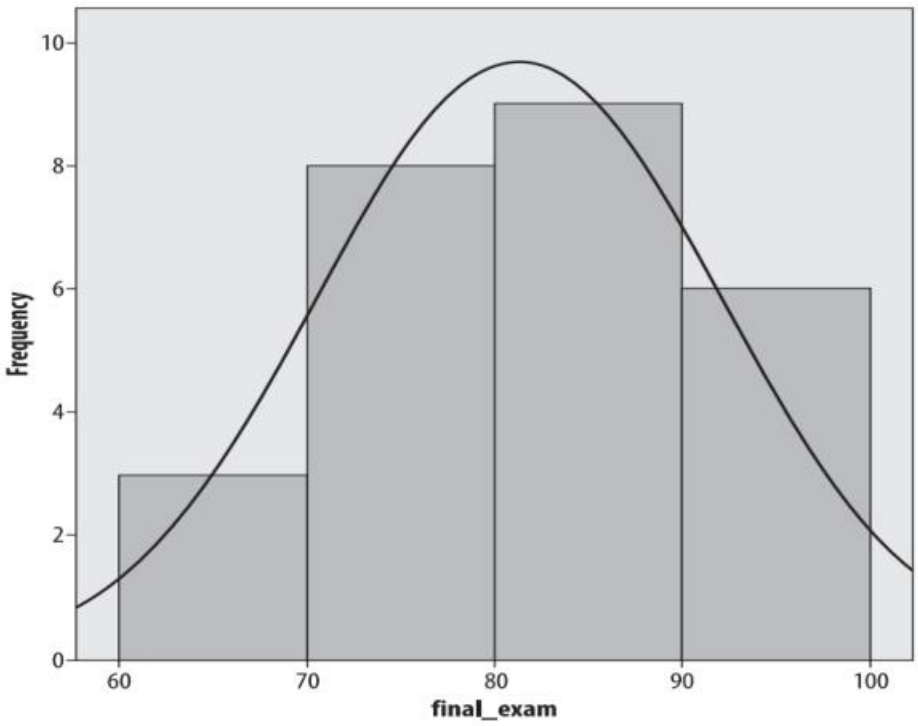
İstatistiksel Model:

İstatistiksel model veriler arasındaki bir ilişki ya da matematiksel bir eşitliktir. Veriye ve ölçümlere dayanarak bir model oluşturulabilir. Örneğin, boyu fazla olan insanların ağırlıklarının da fazla olduğu gözlemleniyorsa, insanların boy ve ağırlıkları arasındaki ilişki hakkında bir model oluşturabiliriz. Burada modelimiz, "insanların boyu ve ağırlıkları arasında pozitif bir ilişki vardır" olabilir. Bu ilişkiyi matematiksel bir ifade şeklinde de belirtebiliriz. Örneğin, a ve b adında 2 katsayı kullanılarak "ağırlık = a x boy + b" şeklinde bir eşitlik oluşturulabilir ve bunun için en ideal a ve b değerlerini bulabiliriz. Burada ise modelimiz direkt bu eşitlik olacaktır.

Bir model oluşturduktan sonra o modeli tahminlerde bulunmak için kullanabiliriz. Örneğin, boy ve ağırlık ilişkisi modelini kullanarak benden daha uzun bir insanın genelde benden daha kilolu olduğunu ileri sürebilirim. Ya da boy ve ağırlık ile ilgili bir matematiksel eşitlik olan modeli kullanarak boyu bilinen bir insanın kilosunu tahmin edebilirim.

Histogram

Histogram, bir veri kümesinin frekans dağılımını gösteren grafiktir. Yani bir veri kümesinde belirli aralıklarda kaçar tane verinin bulunduğunu görmemizi sağlar. Aşağıda örnek bir histogram grafiği gösterilmiştir:



Bu grafikte x ekseninde öğrencilerin final sınavı notları, y ekseninde ise notları belirli aralıklarda olan öğrencilerin sayılar vardır. Örneğin, notu 70 ve 80 arasında olan 8 tane öğrenci vardır. Bu grafiğe bakarak sınıftaki not dağılımını görebiliriz.

Gördüğümüz gibi final sınavı notları belirli aralıklara ayrılmış. Bu aralıklara **bin** adı verilir. Yukarıdaki histogram grafiğinde her bir bin'in genişliği 10'dur. Bu yöntem, belirli aralıktaki sayıları aynı gruba koyarak grafiği sadeleştirir ve dağılımın net bir şekilde ortaya çıkmasını sağlar.

Peki her grafikteki bin genişliğini değiştirirsek ne olur? Aşağıdaki grafikte bin genişliği 10 yerine 5 olarak alınmıştır:

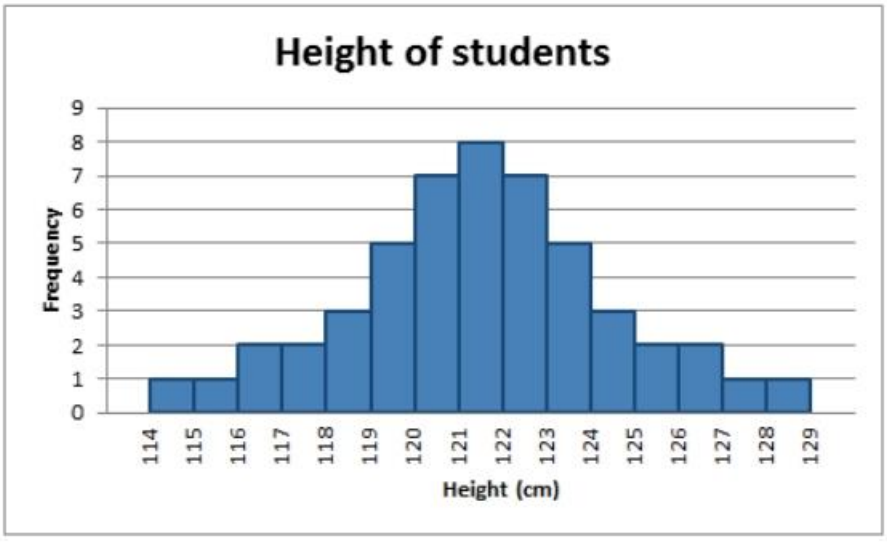
Gördüğümüz gibi dağılımın yapısı biraz değişti. Örneğin, önceki grafikte en çok öğrenci 80 ile 90 notları arasında iken bu grafikte 70 ile 75 arasında! Peki en ideal bin genişliğini belirlemek için bir formül var mı? Bazı formüller bulunsa da kesin bir formül yok. Yani bir verinin histogram grafiğine bakarken farklı bin genişlikleri kullanmamızda yarar var.

İstatistiksel Dağılımlar

İstatistiksel dağılım, bir verideki ölçülerin nasıl dağıldığını gösterir. Önceki konumuzda gördüğümüz **histogram** da bir istatiksel dağılım gösterme yöntemidir ve belirli aralıktaki ölçülerin veri üzerinde sayıca dağılımını gösterir. İstatistiksel dağılımların çok farklı çeşitleri vardır. Biz burada birkaç dağılım türünü göreceğiz:

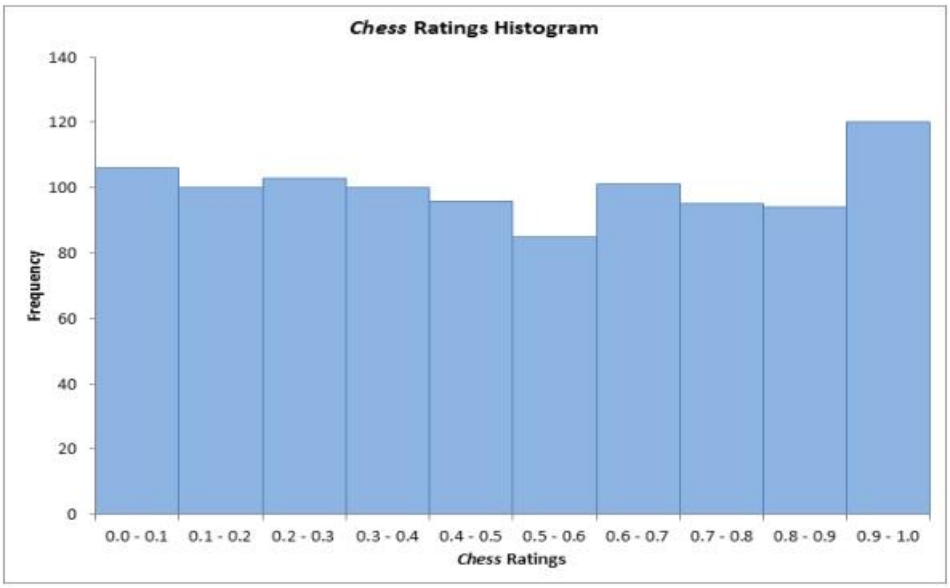
Normal Dağılım

Normal Dağılım, en çok karşılaşılan dağılım türüdür. Merkezinden yanlara simetrik bir şekilde yayılan bir dağılım türüdür. Aşağıdaki şekilde de gösterildiği gibi merkezinde en yüksek değeri alır ve yanlara doğru azalır. Aşağıdaki grafikte öğrencilerin boy uzunluğu verilerine ait bir dağılım gösterilmiştir. Bu tarz dağılımlar genellikle normal dağılım şeklinde görülür.



Tekdüze (Uniform) Dağılım

Her aralığa yaklaşık aynı sayıda verinin düştüğü dağılıma tekdüze (uniform) dağılım denir. Aşağıdaki grafikte satranç oyununda alınan skorların dağılımı gösterilmiştir. Bu grafikte, her aralığa düşen skor sayısı eşit olmasa da birbirlerine yakınlar. Görüleceği üzere bu grafikte normal dağılımdaki gibi bir merkez belirleyemiyoruz çünkü aralıklara düşen sayılar birbirine oldukça çok yakın durumdadır.



Ortalama, Varyans ve Standart Sapma

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 not ortalaması (60 + 80 + 90 +100 + 70) / 5 = 80 olarak bulunur.

Varyans:

Bir veri setindeki tüm verilerin, veri setinin ortalamasına olan uzaklıklarının toplamıdır. σ^2 sembolü, yani standart sapmanın karesi ile gösterilir. Varyans, verilerin birbirinden ne kadar uzaklıkta dağılmış olduklarını ölçer. Hesaplanırken önce ortalama bulunur, sonra tüm verilerin ortalama ile 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 ise 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 ise bu farkların karelerini alacak olursak;

(-20) ^2 = 400, 0 ^2 = 0, 10 ^2 = 100, 20 ^2 = 400, (-10) ^2 = 1000 olarak bulunmuştur.

Fark edecek olursak farkların karesini aldığımız zaman sayılar negatif olmaktan çıkmıştır. Bu da bize uzaklık bilgisini, yani negatif olmayan bilgiyi sağladı. Şimdi ise bulduğumuz farklara ait kare değerlerini toplayalım:

400 + 0 + 100 + 400 + 100 = 1000 Bulunan bu sayıyı da toplam veri sayısına

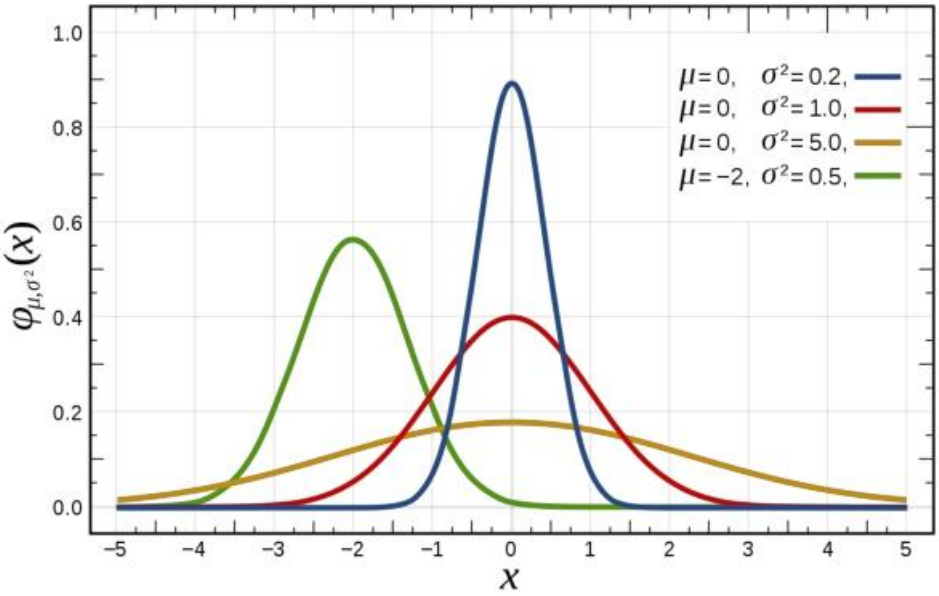
bölelim: 1000 / 5 = 200

Standart Sapma

Standart sapma, varyansın kareköküdür. Peki neden? Varyansı hesaplarken farkların karesini aldık. Farkların kareleri alındıktan 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:

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

Aşağıdaki grafikte farklı ortalama ve standart sapma değerlerine sahip normal dağılımlar gösterilmektedir. Dağılımların ortalama değerleri (μ değeri ile gösterilmekte) gördüğümüz gibi normal dağılımların merkezleri, yani tepe noktalarıdır. Varyanslarına (σ^2 sembolü ile gösterilmekte) bakacak olursak varyansı fazla olanların daha geniş ve daha fazla yayılmış olduğunu, varyansı az olanların ise daha dar ve daha keskin olduğunu görürüz.



Hipotez Testi ve NULL Hipotezi

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, %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, yapılan bir tahminin yani bir hipotezin doğru olup olmadığının test edilmesidir. Bir hipotez testinde birbirine zıt olan iki tane hipotez kurulur. Bu hipotezlerden birisi reddedilirse diğeri doğru kabul edilir.

H0 (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 olmadığını düşünelim. Bu sınıflardaki öğrencilerin not ortalamalarının farklı olup olmadığını test etmek isteyelim. Bu durumda, sıfır hipotezi, "A sınıfının ve B sınıfının not ortalamaları 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 ve bu yüzden sıfır hipotezi reddedilir ve alternatif hipotez kabul edilir. Ancak A sınıf ortalaması 77, B sınıfının ortalaması 78 ise arada pek fark yoktur ve bu fark şansa bağlı olarak kabul edilir. Böyle bir durumda ise sıfır hipotezi reddedilemez ve kabul edilir.

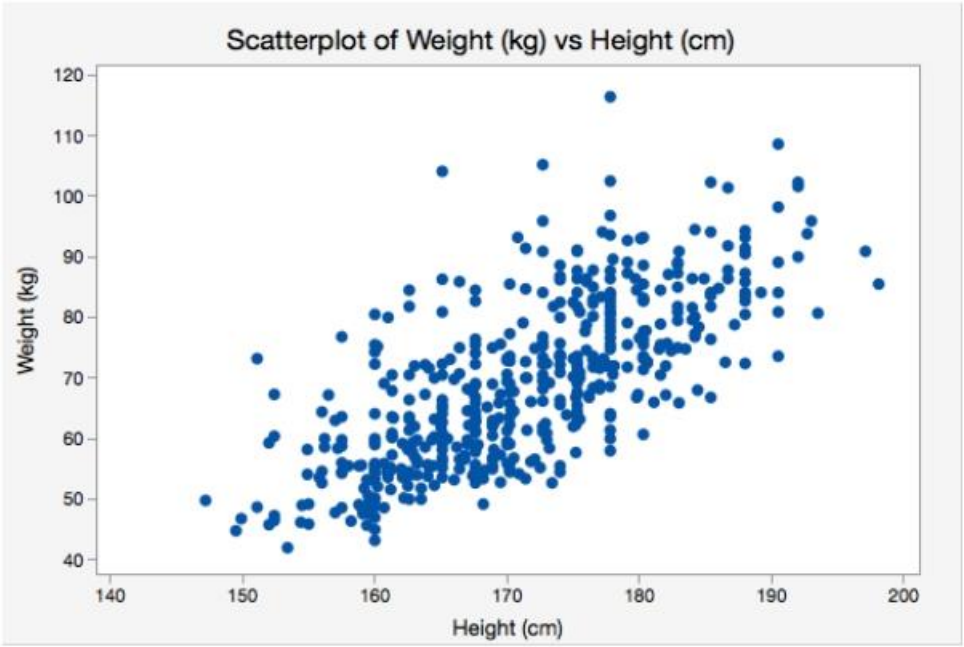
P - Değeri

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 söylenebilir ve sıfır hipotezini reddedebiliriz.

Örneğin, A ve B sınıflarının not ortalamalarını karşılaştırdığımız hipotezlerimizi hatırlayalım. Sıfır hipotezi, ortalamaları arasında pek 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 reddedilemeyeceğine bakılır. P değeri, sınır sayıdan küçük ise sıfır hipotezi reddedilir ve gruplar arasında fark vardır denilir. Ancak p – değeri sınır sayıdan büyük ise sıfır hipotezi reddedilemez ve bu gruplar arasında fark yoktur, benzerdir denilir.

Kovaryans

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



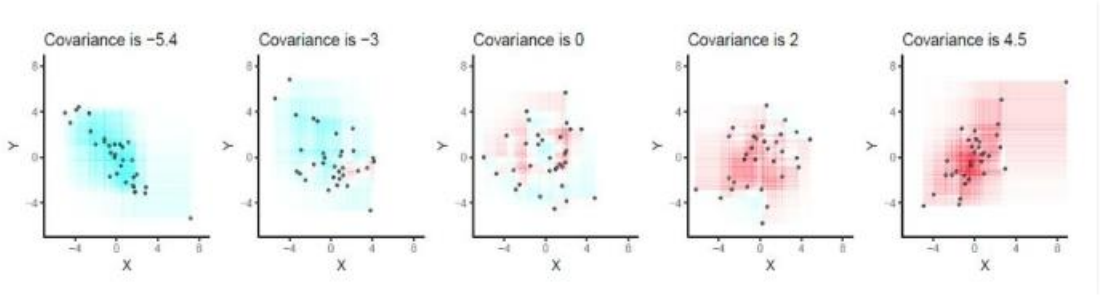
Bu grafikte her nokta bir insana karşılık geliyor ve her insanın yani noktanın bir boy ve ağırlık değerleri var. Grafiğe baktığınızda boy ve ağırlıkların arasında bir ilişki görebiliyor musunuz? Evet, genelde boyu fazla olan insanların ağırlığı da fazla, boyu az olan insanların ağırlığı da az oluyor.

Yani boy ve ağırlık arasında **pozitif** (biri arttıkça diğeri de artan) bir ilişki var. Biri artarken diğeri azalsaydı **negatif** bir ilişki olacaktı. Birinin artması ya da azalması diğerini etkilemiyor olsaydı aralarında bir ilişki olmayacaktı.

Kovaryans konusuna tekrardan geri dönecek olur isek; 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:

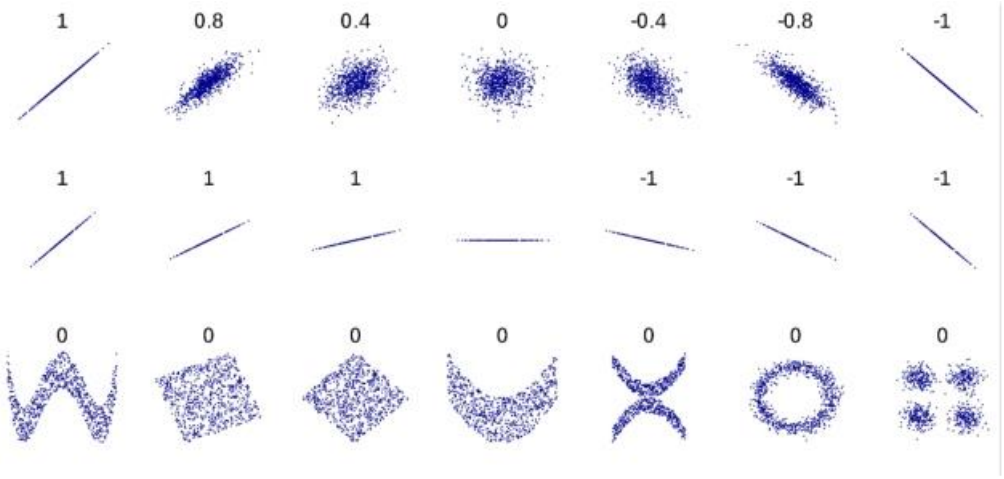
$$cov(X, Y) = \frac{(x_1 - \mu_x).(y_1 - \mu_y) + (x_2 - \mu_x).(y_2 - \mu_y) + \cdots + (x_n - \mu_x).(y_n - \mu_y)}{n}$$

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



Korelasyonun formülü ise kovaryans ve standart sapma değerlerini bildiğimiz zaman bayağı basitleşmektedir. Korelasyon, iki veri kümesinin kovaryansının, varyanslarının çarpımına bölümüdür. R sembolü ile gösterilir. Formül olarak aşağıda gösterilmiştir:

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

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 ise 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.

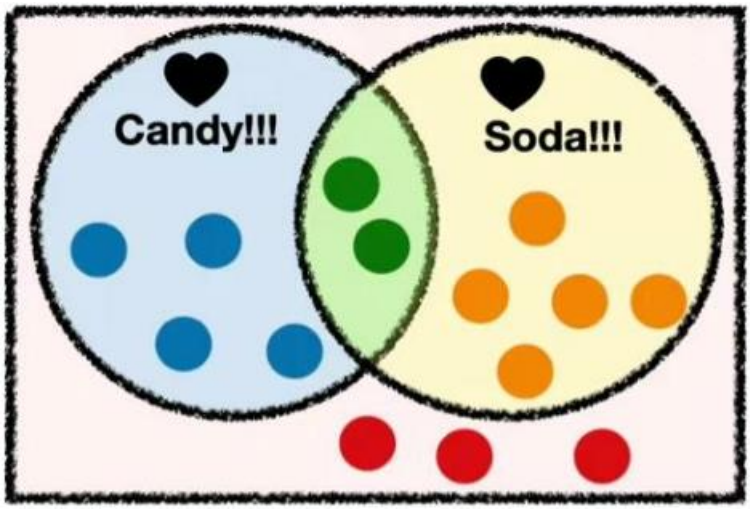
Koşullu Olasılık

Koşullu olasılık, bir koşulun gerçekleştiği bilindikten sonra başka bir koşulun gerçekleşmesi 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. Peki ben gelen sayının bir çift sayı 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 bir 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 ise soda sevenler gösterilmiş. Öncelikle toplam kişi sayısına bakalım: Toplamda 14 kişi var. P(soda) yani bir kişinin soda sevme olasılığı kaçtır? Soda dairesinin içerisinde toplamda 7 kişi var. O zaman, soda seven sayısı / toplam sayı = 7 / 14 yani 1 / 2 kişi. Bir kişinin soda sevme olasılığı %50'ymiş. Bir kişinin hem soda hem şeker sevmesi 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 sevme olasılığı kaçtır? Burada koşulumuz "soda" ve sonucumuz "soda ve şeker". Yani p(soda ve şeker | soda) ya da kısaca p(şeker | soda), p(soda ve şeker) / p(soda) hesabıyla bulunur.

Hesaplarsak p(şeker | soda) = 2/7 olarak bulunur.

Bayes Teoremi

Bayes teoremi, koşullu olasılıklarla türetilmiş bir ifadedir. Kendisi bilinmeyen ancak tersi bilinen bir koşullu bir 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 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 n B) var. Formül 2'deki p(A n B) değerini yalnız bırakalım:

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

Burada bulduğumuz p(A n B) değerine eşit olan değeri formül 1'deki p(A n B)'nin yerine koyacak olursak gösterdiğimiz yukarıdaki Bayes Teoremi formülünü elde ederiz.

KAYNAKÇA

https://academy.patika.dev/tr/courses/istatistik

https://www.youtube.com/@statquest