acemiadam.com

Örneklemin Standart Sapmasının Hesaplanması - Acemi Adam

Resul Çelik

6–8 minutes

İlk başta herkesin anlayabileceği şekilde yazmaya çalıştım fakat konu çok dağıldı. Bu yüzden olasılık ve istatistik alan birinin anlayabileceği şekilde anlatmaya karar verdim. Bu dersleri alan adam zaten konuyu biliyor diyebilirsiniz ama öyle değil. Bir örneklemin standart sapması hesaplanırken paydayı örneklemin genişliğine(n) değil de (n-1)'e neden böldüğümüzü çok az öğrenci biliyordur tahminimce. Ufuk açması açısından ve istatistiğin temelindeki bir konuyu anlamamız açısından çok faydalı olacağını düşündüğüm bir yazı diyebilirim.

Standart Sapma Nedir ve Nasıl Hesaplanır?

Elinizde bir deneyle alakalı farklı zamanlarla alınmış veriler var. Genelde bir sayıya yakın çıkıyor ama hep aynı sayı çıkmıyor. İşte bu farklılığa standart sapma deniyor. Mesela deneyi 100 defa yaptık. Varyansının formülü de şöyle oluyor;

$$\sigma^2 = \sum_{i=1}^{100} \frac{(x_i - \mu)^2}{100}$$

Varyans demek standart sapmanın karesi demektir. Yani bunun kökünü aldığınızda standart sapmayı bulursunuz. Burada herhangi bir sorun yok fakat bu deney popülasyonunun değil de buradan aldığımız bir örneklemin standart sapmasını hesaplamaya çalışırsak formül biraz farklı oluyor. Mesela örneklem için 10 tane deney sonucu alalım. Örneklemin varyans formülü;

$$s^2 = \sum_{i=1}^{10} \frac{(x_i - \bar{x})^2}{9}$$

10 tane deney alıyoruz ama 9'a bölüyoruz. Bunu hiçbir zaman anlayamadım. Hayat ve dersler bana bunu anlamam ve sorgulamam için vakit bırakmadı. Geçenlerde araştırdım.

Bessel's Correction

Sayı ile ifade ettiğim için karışıklık olur mu bilmiyorum ama şöyle anlatayım. Eğer gerçek veri kümenizden n tane olay içeren bir

örneklem aldıysanız standart sapma hesaplarken (n-1)'e bölüyorsunuz. Bunun (n-1)'e bölünmesine Bessel'in düzeltmesi diyorlar ki adam iyi de yapmış. Grafikle göstermeye çalışacağım.

Mesela bir olay için 100 gözlem yaptık ve elimizde 100 veri var. Gözlem değeri sayı olarak en az 0 ve en çok 4.000.000 çıkıyor. Ortalaması da 999.999 oluyor. Standart sapmanın mantığını anlamak için bir doğruda bu deney sonuçlarını gösterip 10 tane örneklem alalım.

Birinci doğruda 100 tane nokta düşünün. Bunların hepsinden ortalama çıkartılıp(büyük turuncu nokta) kareleri alınıyor. İkinci grafikte ise 10 tane örnek alınıp(küçük kırmızı nokta) ortalaması alınıyor(büyük kırmızı nokta) ve aradaki mesafeler bulunuyor. İşte burada şunu görüyoruz. Popülasyon yani tüm gözlemlerin olduğu doğruda bu mesafeler genişliyor(Okların mesafelerine bakabilirsiniz). Çünkü popülasyonda ortalamaya uzak değerler var yani dağılma fazla. Yukarıdaki oklarda olduğu gibi 100 deney içinde bu mesafelerin karelerini alıp toplasak ve 100'e bölsek standart sapmayı buluyoruz. Fakat bunu örneklem için yaptığımızda mesafeler kısa olduğu için eğer örneğimizde olduğu gibi 10 tane örnek aldıysak ve 10'a bölersek popülasyonun standart sapmasından daha küçük bir standart sapma buluyoruz. Çünkü bizim örneklem verimiz daha derli toplu. Daha derli toplu olunca da daha küçük sapmalar meydana geliyor. İşte bu yüzden örneklemin standart sapmasını popülasyon standart sapması gibi hesaplarsak bu hesapladığımız standart sapmadan beklediğimiz değer popülasyon standart sapmasıyla uyumlu olmuyor. Peki bunun çözümü ne?

Çözüm

Bir popülasyonu tahmin etmek için kullanılan örneklem verisine estimator(tahmin edici) deniyor. Ve bunun beklenen değeri(Expected Value) popülasyonun verisine eşit olmak zorunda. Mesela yukarıdaki örnekte örneklemim ortalamasını buldunuz(büyük kırmızı nokta).Eğer siz bu 100 deneyden 10'ar 10'ar örneklemler alsanız ve her birinin ortalamalarını bulsanız. Bu ortalamaların ortalamasının popülasyonun ortalamasına eşit olması beklenir. (Turuncu büyük nokta) Formül olarak;

$$E[\bar{x}] = \mu$$

Bu formülden yola çıkarak örneklemin varyansının(standart sapmanın karesi) beklenen değerinin de popülasyonun varyansına eşit olması lazımdır. Yani;

$$E[s^2] = \sigma^2$$

Şimdi örneklemin varyasyonunu n'e bölerek hesaplayalım. Eğer bu hesabın beklenen değeri popülasyon varyasyonuna eşit çıkmazsa bir şeyleri değiştirmek gerekecek.

$$s^2 = \sum \frac{(x_i - \bar{x})^2}{n} \qquad E[s^2] = \sigma^2 \qquad \sigma^2 = \sum \frac{(x_i - \mu)^2}{n}$$

$$\mathrm{E}[\sum \frac{(x_i - \bar{x})^2}{n}] = \sum \frac{(x_i - \mu)^2}{n}$$

Beklenen değer hesaplarken sabit sayıları parantez dışına alabiliriz.

$$\frac{1}{n}\mathrm{E}[\sum (x_i - \bar{x})^2] = \sum \frac{(x_i - \mu)^2}{n}$$

$$\frac{1}{n}\mathrm{E}[\sum ({x_i}^2 - 2x_i\bar{x} + \bar{x}^2)] = \sum \frac{(x_i - \mu)^2}{n}$$

Toplama sembolü ve beklenen değer parantezi parçalara ayrılabilir.

$$\frac{1}{n}(\mathrm{E}\left[\sum {x_i}^2\right] - 2E\left[\sum x_i\bar{x}\right] + E\left[\sum \bar{x}^2\right]) = \sum \frac{(x_i - \mu)^2}{n}$$

① ② ③

Buraya kadar kolay anladım ama sonrasında biraz karıştırdığım için numaralandırdığım ifadeleri detaylı anlatacağım. 1 numaralı ifadeden başlayalım. Burada örneklemdeki verilerin karelerinin toplamının beklenen değeri soruluyor. Her verinin karesinin beklenen değeri eşit olduğu için bu ifade n tane beklenen değer toplamına eşit olur.

$$\mathrm{E}\left[\sum {x_i}^2\right] = n\mathrm{E}[{x_i}^2]$$

2 numaralı ifadeyi biraz daha açıklıyorum.

$$E\left[\sum x_i\bar{x}\right] = \mathrm{E}[\bar{x}\sum x_i] \quad (2.1)$$

$$\bar{x}n = \sum x_i \quad (2.2)$$

$$E\left[\bar{x}\sum x_i\right] = \mathrm{E}[n\bar{x}^2] = n\mathrm{E}[\bar{x}^2] \quad (2.3)$$

2.1 numaralı eşitlikten şunu anlamalıyız; Örneklemin ortalaması örneklem içinde hep aynı olacağı için toplam sembolünde sabit gibi davranır ve dışarı çıkarabiliriz. Fakat bunun beklenen değeri örneklemden örnekleme değişeceği için beklenen değer parantezinden çıkamaz.

2.2 numaralı ifade çok kolay anlaşılabilir. Örneklemdeki verilerin toplamı örneklem ortalamasıyla örneklemdeki veri sayısının çarpımına eşittir.

2.3 numaralı ifade ise önceki iki ifadeyi birleştirip sabit olan n'i beklenen değer parantezinden dışarı alıyoruz.

Sonrasında 1,2 ve 3 numaralı denklemlerin yerine basitleştirilmiş hallerini yerleştiriyoruz. Ve ortaya aşağıdaki tablo çıkıyor.

$$\frac{1}{n}(n\mathrm{E}[{x_i}^2] - 2nE[\bar{x}^2] + nE[\bar{x}^2]) = \sum \frac{(x_i - \mu)^2}{n}$$

$$\frac{1}{n}(n\mathrm{E}[{x_i}^2] - nE[\bar{x}^2]) = \sum \frac{(x_i - \mu)^2}{n}$$

$$\frac{n}{n}(\mathrm{E}[{x_i}^2] - E[\bar{x}^2]) = \sum \frac{(x_i - \mu)^2}{n}$$

Şimdi burada hatırlamamız gereken iki formül var. Aşağıya yazıyorum bunları ve bunlar varyansın tanımından geliyor. Yukarıdaki son denklemimizdeki beklenen değerler yerine aşağıdaki ifadeleri yerleştiriyoruz.

$$\mathrm{E}[{x_i}^2] - E[x_i]^2 = Var(x_i)$$

$$\mathrm{E}[\bar{x}^2] - E[\bar{x}]^2 = Var(\bar{x})$$

Ve şöyle bir ifade karşımıza çıkıyor;

$$\frac{n}{n}(Var(x_i) + E[x_i]^2 - Var(\bar{x}) - E[\bar{x}]^2) = \sum \frac{(x_i - \mu)^2}{n}$$

Burada karşımıza çıkan ifadeleri birazda olsa tanıyoruz. Sadeleştirmek için şu formüllerle yer değiştiriyorum.

$$Var(x_i) = \sigma^2 \qquad E[x_i]^2 = \mu^2$$

$$Var(\bar{x}) = \frac{\sigma^2}{n} \qquad E[\bar{x}]^2 = \mu^2$$

Örneklemin varyansı popülasyonun varyansının n'e bölünmesiyle ortaya çıkıyor.

$$\frac{n}{n}(\sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2) = \sigma^2$$

Burada gördüğünüz sadeleştirmeleri yaptığınızda karşınıza sonuç

çıkıyor.

$$\frac{n-1}{n}\sigma^2 = \sigma^2$$

Gördüğünüz gibi eşitliği sağlamadı ve bizim varsayarak koyduğumuz kırmızı n eşitliğin solunda kaldı. Ve pay (n-1) olarak gözüktü. Eğer kırmızı n'i ilk yazdığımız denklemi n ile çarpıp (n-1)'e bölersek doğru ifadeyi yakalayacağız demektir. Bu da aslında standart sapma hesaplarken örneklemdeki veri sayısına değil de bunun 1 eksiğine bölmek demektir. Formül olarak vermek gerekirse;

$$s^2 = \sum \frac{(x_i - \bar{x})^2}{n-1}$$

Elimden geldiğince açıklamaya çalıştım. Benim de anlayamadığım bazı noktalar var fakat yine de bazı şeyleri açıklığa kavuşturmak açısından faydalı olabileceğini düşünüyorum. Derslerde genelde bu konu n'e değil (n-1)'e bölüyoruz diyerek geçiliyor. Vize ve final geçmek yerine temellerini anlamak insanı daha çok tatmin ediyor.