Temel İstatiksel Kavramlar

Uz. Dr. Mehmet AKMAN
Marmara Üniversitesi Tıp Fakültesi Aile Hekimliği Anabilim Dalı, İstanbul

Özet

Temel istatistik bilgisi, bilimsel araştırma sürecine dahil olan her bilim insanının sahip olması gereken bir bilgidir. Araştırmalarda elde ettiğimiz verileri özetlememiz, yorumlamamız ve elde ettiğimiz sonuçların tesadüfü olmadığını gösterebilmemiz istatistiksel analizlerle mümkündür. Bu derlemede temel istatistiksel kavramların (değişken, merkezi ölçütler ve yayılım ölçütleri) örneklerle açıklanması hedeflenmiştir.

Anahtar kelimeler: Temel istatistik, değişken, merkezi ölçütler, yayılım ölçütleri

Basic statistical concepts

Abstract

Basic statistical knowledge is essential for any scientist, who involved in a scientific research process. Statistical analyses provide us to summarize and interpret our data, and also enable to show our results are not incidental. In this review, basic statistical concepts (variable, measures of central tendency and measures of variability) are explained and an example for each concept is presented.

Key Words

Basic statistics, variable, measures of central tendency, measures of variability

I. İstatistiğe giriş

Çoğu bilimsel araştırmanın amacı veri toplayarak araştırma konusu hakkında bilgi elde etmektir. Herhangi bir araştırmanın verisi, o araştırmanın incelediği bir veya daha fazla değişkene yönelik gözlemlerden oluşur. O halde bir araştırma sırasında incelenen özelliği değişken olarak adlandırabiliriz. Örneğin A hastalığı ile ilgili yapacağımız bir çalışmada; A hastalığına yakalanmış hastalarla ilgili hangi bilgileri toplayacağımızı düşünürken değişkenleri oluşturmuş oluruz: cinsiyet, yaş, eğitim durumu, boy veya başvuru şikayetleri gibi. Burada sayılan hasta özelliklerinin her biri değişkendir. O değişkenlerin aldığı değerler de **veri** olarak adlandırılır.

Örnek

12 kadın, 18 erkek katılımcıya hepatit B ile ilgili bir anket uygulanmıştır.

Bu örnekte "cinsiyet" değişken olarak, araştırmaya katılan her bir erkek veya kadın katılımcının cinsiyet bilgisi ise bir "veri" olarak kabul edilir. Bir başka deyişle, incelenen toplam 30 kişilik grupta cinsiyet değişkeni altında, 12 kadın verisi ve 18 erkek verisi yer almaktadır.

Verilerin bir araya gelerek oluşturdukları veri seti (genellikle veri tabanı olarak adlandırılır) toplumun ilgili kısmını (**popülasyon**) temsil edecek bir **örneklem**den elde edilir. Bir veri tabanını incelerken hedefimiz veriyi mantıklı bir şekilde sıkıştırarak yararlı bilgilerin açığa çıkmasını sağlamaktır. İstatistik verilerin, toplanması, özetlenmesi, analizi ve yorumlanmasına ilişkin bir dizi yöntemleri içinde barındıran bir alandır. Bir başka deyişle araştırma amacımıza ulaşabilmek için istatistiksel teknikleri kullanırız. Veri farklı biçimlerde karşımızı çıkabilir. Uygun istatistiksel analiz yöntemini seçmeden önce her bir değişkenin ne tip değişken olduğunu bilmemiz gerekir. Her bir değişken, dolayısıyla ilgili veriler, başlıca iki tipten biri ile uyumludur: Kategorik veya sayısal.

Kategorik (niteliksel) değişkenler

Bazı özelliklerine göre birbirinden farklı kategorilere ayrılabilen değişkenlerdir. Bu özellikler nitelik olarak belirtilmiştir. Bu verilere "kategorik *veya* niteliksel değişkenler" denir. Her bir olguya ait veri birbirinden bağımsız kategorilerden sadece birine girebiliyorsa kategorik değişkenden bahsedilebilir.

Kategorik değişkenler kendi içinde iki gruba ayrılabilir: **nominal** veya **ordinal** (sıralı) değişkenler. Bir kategorik değişkenin nominal veya ordinal olmasını belirleyen temel özellik kategoriler arasında bir hiyerarşinin varlığı veya yokluğudur.

Nominal değişken, kategorilerin hiyerarşik bir sırada değil sadece kendi isimleriyle sınıflandırıldığı değişkenlerdir. Örneğin: Cinsiyet (kadın-erkek), göz rengi (kahverengi, mavi, yeşil, siyah, ela), mezun olduğu üniversite (İstanbul, Hacettepe, Marmara, Ege, Uludağ vb) gibi.

Bu noktada bilimsel araştırmalarda sık kullanıldığından, sadece iki kategoriye ayrılabilen kategorik değişkenlere "**dikotom** değişkenler" dendiğini belirtmek yerinde olur. Örneğin kadın/erkek, hasta/sağlam, sigara içer/içmez gibi.

Ordinal değişkenlerde bir grup kategorik değişkenin alabileceği değerler arasında hiyerarşik bir sıralama söz konusudur. Kategoriler bir şekilde sıralanmışlardır. Örneğin: Hizmetten memnuniyet durumu: "çok memnun- memnun- kararsız- memnun değil- hiç memnun değil"

Semptomların şiddeti: "ağır-orta-hafif" veya hastalık evresi : "yok, hafif, orta, ileri" gibi.

Sayısal (niceliksel) değişkenler

Sayısal değişkenler belli bir aralıkta, herhangi bir sayısal

değeri alabilir. Bu aralığın bir birini takip eden, eşit miktarlarda artış gösteren bir yapısı olduğu kabul edilir. Örneğin 100 cm'lik bir cetveli ele alalım. Bu cetvelde 0'dan 100'e kadar rakamlar birer cm aralıklarla dizili bulunmaktadır. Bu eşit aralıklı ve bir birini takip eden diziyi verilerimizin yer alabileceği aralık olarak belirlersek, verilerimiz 0 ila 100 cm arasında herhangi bir değer alabileceklerdir.

Sayısal değişkenler **kesikli** ve **sürekli** olmak üzere ikiye ayrılabilir. Kesikli değişkenlerde temel özellik her bir verinin sadece belli sayısal değerler alabilmesidir. Örneğin: belirli bir süre içerisinde aile hekimi başına yapılan muayene sayısı veya son beş yılda kişi başına düşen hastalık epizodu sayısı gibi.

Sürekli değişkenlerde ise o verinin temsil edeceği değer için herhangi bir sınırlama yoktur. Örneğin, yaş, boy veya kilo ölçümleri, hemoglobin değerleri gibi.

Farklı tipteki değişkenlerin bir birinden ayrılması

Çoğunlukla değişkenin sayısal veya kategorik olmasına göre farklı istatistiksel analiz yöntemleri kullanılır. Sayısal ve kategorik değişken arasındaki ayrım genellikle kolay gibi görünse de bazen güçlükler barındırabilir. Örneğin çok sayıda sıralanmış kategorisi olan bir değişkenle karşılaştığımızda (mesela 10 kategorili bir ağrı ölçeği), bu değişkeni kesikli bir sayısal değişkenden ayırmak kolay olmayabilir. Bu ayrımı kesikli ve sürekli sayısal değişkenler arasında yapmak daha da güç olabilir. Ancak bu ayrımların yapılmış olması çoğu istatistiksel analizin sonuçları üzerinde çok az bir etkiye sahip olacaktır. Yaş sıklıkla kesikli bir değişken gibi ele alınır, oysaki gerçek anlamda bir sürekli değişkendir. Burada en önemli nokta sayısal bir değişkeni veri tabanına ilk kez kaydederken kategorik olarak kaydetmemeye özen gösterilmesidir. Çünkü sayısal bir değişkenin sonrasında istenildiği şekilde kategorik veriye dönüştürülmesi mümkünken, bir kez kategorik girilmiş bir verinin sayısal veriye dönüştürülmesi mümkün değildir.

II. Tanımlayıcı İstatistikler

İstatistik analizin ilk basamağı verilerin tanımlayıcı istatistiklerinin elde edilmesidir. Bir araştırma grubundaki verilerin çeşitli özelliklerinin özetlendiği ortalama, oran, standart sapma vb. rakamlar topluluğuna *tanımlayıcı istatistik* adı verilir. Tanımlayıcı istatistikler, (1) Merkezi ölçütler ve (2) Yayılım ölçütleri olarak iki başlık altında ele alınabilir.

Merkezi ölçütler

Bir veri dizisi ile karşılaştığımızda, o veri ile ilgili bir fikre sahip olmak için akla yatkın bir şekilde verinin özetlenmesine ihtiyaç vardır. Bu amaçla kullanılabilecek ve belli bir özelliğe göre (merkezilik) veri dizisini özetleyen ölçütlerden birisi merkezi ölçütlerdir. Merkezi ölçütler denildiğinde, o veri tabanı içindeki verilerin merkezini belirtmek üzere temsili olarak seçilen sayıdan söz ediliyor demektir. En çok kullanılan merkezi ölçütler ortalama, ortanca (medyan) ve tepe noktası (mod)'dır. Bu ölçütler veri tabanının merkezine yönelik olarak yer tarif eden ölçütlerdir. Bir başka deyişle o veri setini oluşturan değerlerin merkezinin hangi sayıya denk düştüğünü belirtirler.

Aritmetik Ortalama

Bir sayısal değişkene ait olan eldeki tüm verilerin (sayılar) toplanıp, elde edilen toplamın veri sayısına bölünmesi ile elde edilen değere "aritmetik ortalama" (kısaca ortalama) denir. Eğer elimizdeki veriler tüm **popülasyona** aitse, bu verilerden elde edeceğimiz ortalama da **popülasyon ortalaması** olacaktır. Popülasyon ortalaması "μ" simgesi ile gösterilir (mü şeklinde okunur). Buna karşılık elimizdeki veri toplumdan alınmış bir **örneklem**i temsil ediyorsa, bu verilerden elde edeceğimiz ortalamaya örneklem ortalaması denir. Aritmetik ortalaması alınan verilerin her biri birer ölçüm olduğundan, aritmetik ortalama sadece sürekli değişkenlerde kullanılabilir. Matematiksel formül olarak aşağıdaki gibi gösterilir. Burada ortalama "x üzeri çizgi" ile sembolize edilmiştir.

Nasıl hesaplanır?

$$\bar{x} = \frac{x_1 \quad x_2 \quad x_3 + \ldots + x_n}{n}$$

Bu formül şu şekilde kısaltılabilir:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Burada Σ(sigma) 1. Değerden *n*. değere kadar tüm değerlerin toplandığını ifade etmektedir.

Örnek: Aşağıdaki verilen örnekleme ait ortalama değer ne olabilir?

3 9 7 12 0 -4 2 -12 7 0

Çözüm:

Aritmetik Ortalama = [4+9+7+12+0+(-4) +2+(-12) + 7 + 0] /10 = 22/10 = 2.2

Aritmetik Ortalamanın en büyük avantajı değişkene ait tüm verilerin kullanılması ile elde edilen bir merkez ölçüsü olmasıdır. Kolay hesaplanabilir bir merkez ölçütüdür. En önemli dezavantajı ise **aşırı değerlerden** etkilenmesidir. Aşırı değerler, dağılımdaki diğer değerlerden çok farklı olan az sayıdaki değerler olarak tanımlanabilir. Bu değerler, veri dizisinin diğer değerleri ile uyumsuzluk gösterirler. Ölçüm hatalarına, materyalin veya ölçüm aracının bozulmasına, yanlış kaydetmeye bağlı olabilecekleri gibi kişilik farklılıklarından da kaynaklanabilecekleri unutulmamalıdır. Bu değerler istatistiksel olarak fazla bir anlam taşımalar da, bu değerlerin varlığında hesaplanan aritmetik ortalama verinin gerçek merkeziyetini yansıtmaktan uzak olabilir.

ORTANCA (Medyan)

Bir başka merkezi ölçüt ortancadır. Veri tabanı en küçükten en büyük değere doğru sıraya sokulduğunda en ortada yer alan, dağılımın ortasındaki değere **ortanca** adı verilir. Ortanca veri dizisini ortadan ikiye böler. Yani ortancanın üzerideki değerlerle altında değerlerin sayısı birbirine eşittir.

Nasıl hesaplanır?

• Eğer katılımcı sayısı tek ise (n = tek sayı) ortancayı hesaplamak kolaydır. Bu durumda ortanca (n + 1) / 2. değerdir. Bir başka deyişle tam ortada yer alan, her iki tarafındaki denek sayısı eşit olan değerdir.

Örnek: Aşağıdaki veri setinde hangi değer ortancadır?

3 7 8 11 15 20 27 28 29

n= 9 (9+1)/2 = 5
Ortanca =5. sıradaki değer = 15

• Eğer katılımcı sayısı çift ise (n = çift sayı), aslında tam ortanca değere denk düşen bir veri yok demektir. Ancak pratikte bu durumda ortanca en ortadaki iki değerin aritmetik ortalaması alınarak hesaplanır. O halde ortancayı bulmak için (n/2) işleminden elde edilen sıradaki değer ile [(n+2)/2] işleminden elde edilen sıradaki değer toplanıp ikiye bölmemiz yeterli olacaktır.

Örnek: Aşağıdaki veri setinde hangi değer ortancadır?

2 6 10 13 18 22 24 27 31 34

n= 10 10/2 = 5. sıradaki değer→18
(8+2) /2 = 6. sıradaki değer→22
Ortanca = (18 + 22) / 2 = 20

Tekrar hatırlatalım: Ortanca değerin yer aldığı nokta, değerlerin %50'sinin ortanca değerinden küçük, %50'sinin büyük olduğu noktadır. Ortancanın sağladığı en büyük avantaj aşırı değerlerden etkilenmemesidir. Dezavantajı ise tüm veri grubunu kullanmayan bir merkezi ölçüt olmasıdır. Veri özetlenirken hem ortalamayı hem de ortancayı birlikte kullanmak mümkündür ve hatta bu yolla veri hakkında okuyucuya iki istatistiği karşılaştırma olanağı sunulur ve daha fazla bilgi verilmiş olur.

Tepe değeri (Mod)

Bir dağılımda en çok görülen, bir başka deyişle en çok tekrarlanan değere tepe değeri adı verilir.

En sık görülen olay/durum adlandırmak istendiğinde kullanılır. Tepe değerini bulmak için veri dizisi aynı değerler bir arada olacak şekilde gruplandırılır ve her bir gruptaki değerler sayılır. En çok değere sahip grubun (en çok tekrarlanan değer) temsil ettiği değer tepe noktasıdır.

Örneğin:
0 0 2 3 3 4 4 5 5 6 6 6 6 6 6 6 8 8 8 8 8 9 10 10 10

Yukarıdaki veri setinde tepe noktası en çok tekrarlanan değer olan 6'dır.

Eğer bir değişkene ait tüm veriler eşit sıklıkta tekrarlanıyorsa, tepe değeri yoktur. Değişkenin içerdiği veriler arasında diğerlerinden fazla tekrarlanan tek bir değer varsa o veri seti "unimodal " olarak adlandırılır. Eğer en çok tekrarlanan iki değer varsa, veri seti "bimodal" olarak tanımlanacaktır. Tepe değeri, ortalama veya ortancaya göre daha az sıklıkla kullanılan bir merkezi ölçüttür.

Yayılım ölçütleri

Farklı grupların ortalama ve ortanca gibi merkezi dağılım ölçütleri aynı olduğu halde, gruplar birbirlerinden çok farklı olabilirler. Bu nedenle merkezi eğilim ölçütleri yanında, yayılım ölçütleri de çok önemlidir. Bu durumda elimizde sayısal bir veri dizisi olduğunda, bu verinin hem merkeziliği hem de yayılımı hakkında bilgi verirsek, veri dizisini makul bir şekilde özetlemiş oluruz. Yaygınlık ölçütlerinden ilki **dağılım aralığıdır**. En büyük değerle en küçük değerin farkı olarak gösterilebileceği gibi bu iki değer birlikte verilerek de gösterilebilir. Yine dağılım aralığı üzerinden ifade edilebilecek bir diğer yaygınlık ölçütü **persentil**dir. Persentiller, 1 ila 100 aralığında ifade edilirler ve 1. Persentil, tüm verinin %1. Değerine karşılık gelir. Bu bağlamda %25.değer yani 25. Persentil özel bir isim alır 1. **çeyreklik**. O halde 50 persentil 2. Çeyreklik, 75. persentil ise 3. Çeyreklik olarak adlandırılacaktır. **Çeyreklikler arası aralık** dendiğinde ise kastedilen 25 ve 75. Persentil aralığıdır. Şimdi diğer iki önemli yaygınlık ölçütü olan **varyans** ve **standart sapmayı** ayrıntılı olarak gözden geçirelim.

Varyans

Bir veri dizisinin yayılımını ölçmenin bir yolu da her bir değerin ortalamadan ne kadar uzakta olduğunu (saptığını) bulmaktır. Bu uzaklık arttıkça veri dizisinin yayılımı da artacaktır. Ancak uzaklıkların ortalamasını bir ölçüt olarak kullanabilmek imkansızdır, çünkü pozitif ve negatif uzaklıklar bir birini götürecek ve sonuç sıfır olacaktır. Bu sorunu aşmak için her bir uzaklığın (sapmanın) karesi alındıktan sonra ortaya çıkan değerlerin aritmetik ortalaması alınır. Sonuçta elde edilen değere **varyans** denir. Varyans s^2 olarak sembolize edilir. O halde varyans formülü aşağıdaki gibi olacaktır:

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

Bu formülde aritmetik ortalamadan faklı olarak, paydada n değil n-1'in yer alır. Bunun nedeni n-1 ile örneklem üzerinden popülasyon varyansının daha iyi kestirilebilmesidir.

Standart sapma

Bir veri tabanında yer alan değerlerin çeşitliliği o veri setinin yaygınlığını belirler. Çeşitliliğin (veya yaygınlığın) en sık kullanılan ve en yararlı ölçütlerinden birisi "**standart sapma**"dır. Standart sapma varyansın kare kökü alınarak bulunur ve s ile sembolize edilir.

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

Standart Sapmayı hesaplarken öncelikle veri tabanının ortalamasını bulunur. Her bir değerden ortalama çıkarılır. Elde edilen her bir farkın karesini alınır ve bu değerler toplanır. Kareler toplamı veri tabanında yer alan değerlerin toplam sayısından bir eksiğine bölünür (n-1). Elde edilen rakamın karekökü alınır.

Örnek:
Aşağıdaki veri tabanı için standart sapmayı bulalım.

2 3 7 8

Ortalama = (2+3+7+8)/4 = 5

Her bir değerden ortalamayı çıkarırsak
2-5 = -3
3-5 = -2
7-5 = 2
8-5 = 3

Sonuçların karesini alıp toplarsak
9+4+4+9=26

Bu örnekte n = 4 olduğuna göre n-1 = 3 olacaktır. O halde 26/3 = 8.66 bulunur

Son olarak 8.66'nin kare kökünün alırsak 2.94 buluruz.

S = 2.94

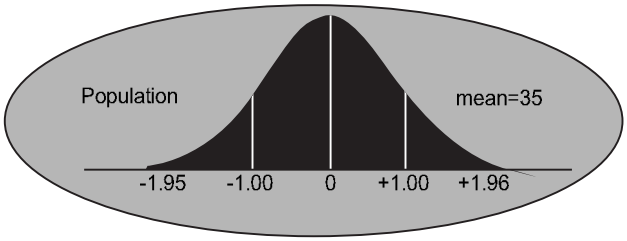
Ortalama ve standart sapma uç değerler içermeyen **simetrik** verilerin özetlenmesinde çok uygun ölçütlerdir. Araştırma verilerinin sunumunda veya herhangi bir bilimsel makalede ortalama ile birlikte standart sapma değeri sıklıkla birlikte ifade edilir. Bir standart sapma, ortalamadan standart (veya tipik) bir miktar sapma (veya uzaklık) anlamına gelir. En kaba anlamıyla "ortalamadan ortalama bir uzaklığı" ifade eder. Standart sapma aynı zamanda verilerin çoğunun hangi aralıkta yer alacağını, ortalamaya göre, göreceli olarak söyler.

Örneğin ampirik bir kural olarak vakaların yaklaşık %95'i ortalamanın iki standart sapma aralığı içerisinde yer alacağını söyleyebiliriz. Standart sapma büyüdükçe dağılımın yaygınlaşacağı unutulmamalıdır. Standart sapma uzaklık belirttiğinden ve uzaklık asla negatif bir değerle ifade edilemeyeceğinden, negatif bire değer almaz. Standart sapmanın alabileceği en düşük rakam sıfırdır. Standart sapmanın sıfır olmasının anlamı veri tabanındaki tüm değerlerin tam olarak bir birine eşit olmasıdır. Yani hiç sapma yoktur. Standart sapma, ortalama gibi, aşırı değerlerden etkilenir. Standart sapmanın birimi orijinal verinin birimi ile aynıdır.

Normal (simetrik) dağılım (çan eğrisi)

Normal dağılımda verilerin büyük bir kısmı ortalamanın etrafında büyük bir yığın halinde toplanmıştır ve ortalamanın her iki tarafında ortalamadan uzaklaştıkça veri sayısı azalmaktadır. Bu özelliğinden dolayı verilerin dağılımı bir çan eğrisini andırır. Normal dağılım gösteren değişken değerleri için aşağıdaki kurallar geçerlidir:

- Değerlerin 1/3'ü ortalama ± 1 SD içindedir.
- Değerlerin %95'i ortalama ± 2 SD içindedir.
- Değerlerin hemen hemen tümü ortalama ± 3 SD içindedir.



Şekil 1: Normal Dağılım

Normal dağılım, daha güçlü olan **parametrik testleri** kullanabilmenize, ortalama ve standart sapma değerlerini verebilmenize imkan sağladığı için önemlidir. Yaşamdan alınmış pek çok değişkene (örneğin standardize test skorları, boy, kilo, kan basıncı ölçümleri, kolesterol, kan şeker vb) ait verilerin dağılımı normal dağılımla (çan eğrisi) benzerlik gösterir.

Normal dağılımın özelliklerini şu şekilde sıralayabiliriz: eğri sürekli, çan şeklinde, ortalama baz alındığında simetriktir. Ortalama, ortanca, tepe noktası dağılımın ortasında yer alır ve bir birlerine eşittirler (**ortalama = ortanca = tepe noktası**). Eğri unimodaldir (tek mod-tepe noktası) ve asla x-eksenine teğet geçmez. Normal eğrinin altındaki alan 1'e eşittir.

Böylece temel istatistik bilgisi içerisinde önemli yeri olan değişken kavramını ve merkezi ölçütler ile yayılım ölçütlerini gözden geçirdik. Araştırma verilerinin özetlenmesinde hangi ölçütlerin kullanılabileceğinin ve bu ölçütlerin özelliklerinin bilinmesi, uygun ölçütün seçilmesinde son derece önemlidir. Doğru seçilmiş ölçütlerle özetlenen veriler, emek yoğun bir süreç olan araştırma sürecinin bilim dünyasına hak ettiği biçimde aktarılmasına yardımcı olacak esas unsurlardan bir tanesidir.

Kaynaklar

1. Özcebe H. Tanımlayıcı istatistikler. "Halk Sağlığı Temel Bilgiler" içinde. (ed) Bertan M, Güler Ç. Ankara, Güneş, Yayınevi. 1995; 100-115.
2. Sümbüloğlu K, Sümbüloğlu V. Frekans dağılımları ve tanımlayıcı ölçüler. "Biyoistatistik" içinde. (ed) Sümbüloğlu K, Sümbüloğlu V. 8.baskı. Ankara, Hatiboğlu. 1998; 7-29.
3. Petrie A, Sabin C. Handling Data. In "Medical Statistics at a Glance". (eds) Petrie A, Sabin C. Oxford, Blackwell Science. 2000; 8-26.
4. Jaisingh L. Descriptive statistics. In "Statistics for the Utterly Confused". (ed/editörler) Jaisingh L. New York, McGraw Hill. 2000:1-119.