Kategorik Veri Analizi

Özellikle sosyal bilimlerde yapılan pek çok araştırmada birimlerden toplanan veriler anket yoluyla elde edilmektedir. Anket çalışmalarında soruların genellikle Likert ölçekli olacak şekilde düzenlendiği ve/veya ağırlıklı olarak nitel değişkenlere ilişkin ölçeklerden oluştuğu, sorular arasında kesikli nicel değişkenlerin de sıkça yer aldığı görülmektedir. Kategorik değişken olarak tanımlanan bu tür değişkenlerle yapılan analizlerde "Kategorik Verilerin Analizi" kapsamındaki istatistik yöntemlerin uygulanması gerekmektedir.

Değişken genel olarak nicel (kantitatif) ve nitel (kalitatif) değişkenler olarak sınıflandırılabilir. Nicel değişkenler doğrudan sayılarla ölçülebilen (ücret, fiyat, ağırlık, uzunluk, vb.), nitel değişkenler ise doğrudan sayılarla ifade edilemeyen (cinsiyet, medeni durum, meslek, marka tercihi, vb.) değişkenlerdir.

Nicel değişken belirli bir aralıkta tüm noktaları gerçel değer olarak alabiliyorsa sürekli, sadece belirli noktaları değer olarak alabiliyorsa kesikli değişken olarak tanımlanır. Buna göre, sıcaklık, ağırlık, basınç gibi değişkenler sürekli, ailedeki çocuk sayısı, sınıftaki öğrenci sayısı, bir oteldeki oda sayısı, kesikli nicel değişkenler olarak sınıflandırılır.

Kategorik değişken, sadece sınırlı değerler veya kategorilerle ölçülebilen değişkendir. Tanımdan kesikli nicel ve nitel değişkenlerin kategorik değişken olarak sınıflandırılabileceği anlaşılmaktadır.

Kategorik veriler sosyal ve biyomedikal bilimlerde oldukça yaygın biçimde kullanılmasına rağmen kullanım açısından asla bu alanlarla sınırlı değillerdir.

Kategorik veriler aşağıda sayılan alanlarda da sıklıkla görülmektedir.

* Eğitim alanında bir sınav sorusu için öğrenci yanıtlarının doğru/yanlış şeklinde kategorilerle değerlendirilmesi gibi.
* Pazarlama alanında bir ürüne ilişkin lider markalar arasında tüketici tercihlerinin "A markası, B markası ve C markası" şeklindeki kategorilerle ifade edilmesi gibi.

(*) Mühendislik bilimi ve endüstriyel kalite kontrol gibi kantitatif alanlarda da görülmektedir. Üretilen parçaların belirli standarda uygun olup olmaması şeklinde sınıflandırılması, belirli bir yiyeceğin tadının ne kadar iyi olduğu ya da bir çalışanın belirli bir görevi ne kadar kolay yerine getirdiği gibi belirli karakteristiklerin öznel olarak iyileştirilmesi örnek olarak verilebilir.

Ölçek Türleri

Ölçek, araştırmanın amacına uygun istatistik yöntemin kullanılabilmesi için değişken değerlerinin toplanma biçimidir. Nicel ve nitel değişkenlerin 4 farklı ölçek türü bulunmaktadır.

Nitel değişkenler → Nominal ya da ordinal ölçeklidir.
Nicel değişkenler → Aralık veya oran ölçeğinde ölçülmektedir.

1) Nominal Ölçek: Birimlerin doğrudan sayılarla ölçülemeyen özelliklerini (nitel değişkenleri) kategorilerine ayıran ölçeklerdir. Farklı kategoriler 0, 1, 2 gibi kodlarla temsil edilmektedir. Ancak bu kodların sayısal bir anlamı yoktur. Örneğin; medeni durum; evli, bekar, dul, boşanmış olarak sınıflandırılabilecek 4 kategorili bir değişkendir ve veri kümesi oluşturulurken bu değişken "evli: 0", "bekar: 1", "dul: 2", "boşanmış: 3" olarak kodlanabilmekte ancak bu sayısal kodlar sayısal bir büyüklüğü ifade etmemektedir. Cinsiyet, işletmede çalışılan departman, oy verilen siyasi parti değişkenleri diğer nominal ölçekli değişkenlere örnek olarak verilebilir.

2) Ordinal (Sıralı) Ölçek: Kategorilerinin kendi aralarında sıralı olduğu değişkenlerdir. Özellikle anket çalışmalarında yer alan ifadelerle ilgili görüşlerin alınmasında sıklıkla kullanılmaktadır. Anket çalışmalarında değişkenlerin sıralı olan kategorileri genellikle üç, beş veya yedi kategoriyle sınırlanmakta ve "Likert" ölçekli değişkenler olarak adlandırılmaktadır.
Herhangi bir ifade ile ilgili düşüncelere başvurulduğunda; tamamen katılmıyorum – katılmıyorum – kararsızım – katılıyorum – tamamen katılıyorum kategorilerinin yer aldığı bir anket sorusu ordinal bir değişkenin kategorileri olarak kabul edilir.
Örneğin; araştırma konusu olan bireyin eğitim düzeyi, ilköğretim, lise, üniversite ve lisansüstü kategorileriyle ölçüldüğünde yine sıralama sözkonusudur.

3) Aralık Ölçeği: Kesin olmayan bir başlangıç noktası olan ve bu başlangıç noktasından itibaren ölçeğin eşit aralıklara bölündüğü, gerçek sıfırı olmayan bir ölçek türüdür. Bu ölçeğe ait en klasik örnek Fahrenheit ve Celcius ölçekleridir. Suyun donma derecesi Celcius ölçeğinde 0°C derece, Fahrenheit ölçeğinde 32°F 'dir. 0 derece sıcaklığın yok olduğunu gösteren bir başlangıç değeri değildir yani burada sıfır yokluk anlamında kullanılmaz. Aralık ölçeğinde değişken değerleri arasındaki uzaklık hesaplanabilir ancak birbirlerinin katı olarak yorumlanamaz. Örneğin; 30°F, 60°F 'ın yarısı olarak yorumlanamaz.

4) Oran Ölçeği: Mutlak yokluğu gösteren bir başlangıç noktası bulunan, bu nedenle de değerler arasındaki farklar kadar oranları da hesaplanabilen ölçeklerdir. Oran ölçeğinin aralık ölçeğinden farkı, gerçek bir sıfır (başlangıç) noktasının ve bölünebilir özelliğinin olmasıdır. Oran ölçekli değişkenlere tüm aritmetik işlemlerin uygulanabilmesi bu ölçeği diğer ölçeklerden daha güçlü kılmaktadır. Uzunluk, ağırlık, ücret, fiyat, kâr gibi değişkenler oran ölçeğiyle ölçülen değişkenlerdir.

Nicel { Sürekli, Kesikli }

Nitel { Ordinal, Nominal }

Kesikli, Ordinal, Nominal } Kategorik

Kesikli nicel ve nitel değişkenler, kategorik değişken sınıfına girmektedir.

Kategorik Verilerin Analizinde Kullanılan Olasılık Dağılımları

Sürekli değişkenlere uygulanan birçok analizde dağılım varsayımı bulunmaktadır. Örneğin; normal dağılım varsayımı, regresyon modellerinde ve ANOVA (Analysis of Variance) sağlanması gereken en önemli varsayımlardan biridir.

Kategorik verilerin analizinde ise verilerin Binom, Multinomial ve Poisson kesikli dağılımlara uygunluğu araştırılmaktadır.

1) Binom Dağılımı:

İki mümkün sonucu (bu mümkün sonuçlar "başarılı" ve "başarısız" olarak tanımlanır.) olan bir deneyde her denemenin "başarılı" olma olasılığının p olduğu durumda, n deneme sonucunda başarı sayısı olan x'in dağılımına Binom dağılımı denir. Dağılımın olasılık fonksiyonu aşağıdaki gibidir:

$$p(x) = \frac{n!}{x!\,(n-x)!}\, p^x (1-p)^{n-x} \quad , \quad x = 0, 1, \ldots, n$$

Burada her deneme birbirinden bağımsızdır yani bir denemenin sonucu diğer bir denemeyi etkilemez. Bu denemeler "Bernoulli Denemeleri" olarak bilinir. Dağılımın beklenen değeri ve varyansı aşağıdaki gibi hesaplanır:

$$E(x) = np$$

$$V(x) = np(1-p)$$

Örnek: Tesadüfen seçilen bir tüketiciye bir markanın bilinirliğinin araştırılması amacıyla marka ile ilgili "Doğru" ve "Yanlış" şıklarının yer aldığı 10 soruluk bir test uygulanmıştır. Daha önce tüketicilere uygulanan bu test sonucunda soruların doğru yanıtlanma oranı 0,70 olarak belirlenmiştir. Tüketicinin tüm soruları doğru yanıtlama olasılığı kaçtır?

$$p(x=10) = \frac{10!}{10!\,0!} (0{,}70)^{10} \cdot (1-0{,}70)^{0} = 0{,}028 \,//$$

2) Multinomial Dağılım:

Multinomial (çok terimli) dağılım, ikiden fazla (k tane) mümkün sonuçlu deneyler için uygulanan kesikli bir dağılımdır. Binom dağılımının genelleştirilmiş şeklidir.

n_i, i. mümkün sonucun kaç kez meydana geldiğini göstersin ($i=1,2,\dots,k$).

p_i, her denemede i. olayın meydana gelme olasılığıdır ve $\sum_{i=1}^{k} p_i = 1$'dir.

Bu durumda $(n_1, n_2, \dots, n_k)$, $(n, p_1, p_2, \dots, p_k)$ parametreli bir multinomial dağılım gösterir.

$$(n_1, n_2, \dots, n_k) \sim Mult(n, p_1, \dots, p_k)$$

Dağılımın olasılık fonksiyonu, $x_i \geq 0$, $x_1+x_2+\dots+x_k=n$ koşulları altında,

$$P(n_1=x_1, n_2=x_2, \dots, n_k=x_k) = \frac{n!}{x_1! \dots x_k!} p_1^{x_1} \dots p_k^{x_k}$$ fonk ile gösterilir.

Genel olarak her bir mümkün sonuç Binom dağılımı gösterir.

$$n_i \sim Bin(n, p_i)$$

Dolayısıyla n_i'lerin beklenen değerleri ve varyansları Binom dağılımındaki gibi hesaplanır.

$$E(x_i) = np_i$$
$$V(x_i) = np_i(1-p_i)$$

Örnek: 8 indirim marketinin (A, B, C, D, E, F, G, H) pazar payları sırasıyla: 0,12 ; 0,12 ; 0,04 ; 0,12 ; 0,18 ; 0,18 ; 0,06 ; 0,18'dir. Tesadüfen seçilen 50 tüketiciden 5'inin A, 7'sinin B, 4'ünün C, 6'sının D, 8'inin E, 7'sinin F, 3'ünün G ve 10'unun ise H marketinden alışveriş yapma olasılığını hesaplayınız.

$n=50$

$n_1=5$	$p_1=0,12$
$n_2=7$	$p_2=0,12$
$n_3=4$	$p_3=0,04$
$n_4=6$	$p_4=0,12$
$n_5=8$	$p_5=0,18$
$n_6=7$	$p_6=0,18$
$n_7=3$	$p_7=0,06$
$n_8=10$	$p_8=0,18$

$P(n_1=5, n_2=7, n_3=4, n_4=6, n_5=8, n_6=7, n_7=3, n_8=10)$ $=P(x)$ olsun.

$$P(x) = \frac{50!}{5!\,7!\,4!\,6!\,8!\,7!\,3!\,10!} \cdot (0,12)^5 (0,12)^7 (0,04)^4 (0,12)^6 (0,18)^8 (0,18)^7 (0,06)^3 (0,18)^{10}$$

$P(x) = 0,000007$ olur.

3) Poisson Dağılımı: Bir olayın belirli bir zaman periyodunda (t), x kez meydana gelme olasılığının hesaplanmasını sağlayan bir olasılık dağılımıdır. Bu durumda olasılık fonk. aşağıdaki gibidir;

$$p(x) = \frac{e^{-\lambda}\lambda^{x}}{x!} \quad , \quad x=0,1,2,\dots,n \qquad (e=2,718)$$

Burada λ, bir olayın t zaman periyodunda ortalama meydana gelme sayısıdır. Poisson dağılımına uygun bir değişkenin beklenen değeri ve varyansı λ'ya eşittir.

$$E(x) = V(x) = \lambda$$

Uygulamada bazen sayım değerlerinin varyansı ortalamayı aşmaktadır. Örneğin; Poisson olasılık fonksiyonundan hareketle belirli sayıda kaza olması olasılığı hesaplanırken, tüm haftalarda ortalama kaza sayısının sabit ve 2 olduğu varsayılsın. Uygulamada bu varsayım geçerliliğini yitirmektedir. Dolayısıyla varyans, ortalama değerine eşit olmamaktadır. Bu durum Aşırı Yayılım (Overdispersion) olarak tanımlanır. Olasılık hesaplamalarında Aşırı yayılım faktörü Poisson dağılımının yetersiz kalmasına neden olmakla birlikte, Poisson dağılımı kategorik verilerin analizinde en çok temel alınan kesikli olasılık dağılımı olma özelliğini korumaktadır. Dağılımın şekli λ'ya bağlıdır. x, λ'nın üstünde değerler aldıkça olasılık fonksiyonundan elde edilecek olasılık değerinin azaldığı görülecektir.

Örnek: Trafik kazalarının yoğun olarak yaşandığı bir bölgede ayda ortalama 2 kaza ölümle sonuçlanmaktadır. Belirli bir ayda ölümle sonuçlanan kaza olmaması olasılığını hesaplayınız.

x kesikli rasgele değişkeni, ölümle sonuçlanan kazaların sayısı olsun. Bir ayda hiç kaza olmama durumu $(x=0)$'dır. Ayda ortalama 2 kaza olduğundan $\lambda=2$'dir. Buna göre hiç ölümle sonuçlanan kaza meydana gelmeme olasılığı,

$$p(x=0) = \frac{e^{-2}\cdot 2^{0}}{0!} = 0,135 \text{ 'dir.}$$

Olumsallık (Kontenjans) Tabloları

Olumsallık tabloları genellikle kategoik değişkenler arasındaki ilişkilerin ~~artaya~~ ortaya
çıkarılmasında kullanılan ve doğal sayıların matris formunda düzenlendiği tablo-
lardır. Çapraz tablolar olarak da adlandırılan bu tablolarda yer alan değerler,
değişken kategorilerinin vei kümesinde kaç kez tekrarlandığını gösteren frekanslar
(tekrarlanan değerler) yani sayım verileridir. Tablo tek değişkene ilişkin
frekansları içeriyorsa tek yönlü, iki değişkene ilişkin frekansları içeriyorsa iki yönlü,
üç değişkene ilişkin frekansları içeriyorsa üç yönlü, daha fazla değişkene ilişkin
frekansları içeriyorsa çok yönlü olumsallık tabloları olarak adlandırılmaktadır.
Değişkenlerin kategoi sayıları iki, üç veya çok yönlü tabloların boyutlarının
belirlenmesinde önemlidir.

Genel olarak tablodaki satır değişkeni R(Row), sütun değişkeni C(Colum)
ile gösterilir. Olumsallık tablolarındaki frekanslar (f_{ij}) ile gösterilir.

i : Satır değişkeninin kategorisini
j : Sütun " " belirtir.

1) Tek yönlü Olumsallık Tablo~~sunun~~ Genel Gösterimi

C Değişkeni		
C_1	C_2	C_3
f_{11}	f_{12}	f_{13}

Burada;

f_{11} : C değişkeninin C_1 kategorisindeki birim sayısı
f_{12} : C " C_2 " " "
f_{13} : C " C_3 " " "

olmaktadır.

Örnek (Tek yönlü): Toplam 1200 çalışanı olan bir işletmede çalışanların 900'ü mavi yakalı, 300'ü beyaz yakalıdır. Tek yönlü olumsallık tablosunu düzenleyiniz.

Çalışanların Pozisyonlarına göre Dağılımı

Çalışan Pozisyonu	
Mavi Yakalı	Beyaz Yakalı
900	300

2) a) İki yönlü (2x2) Olumsallık Tablosunun Genel Gösterimi

Değişkenler	C Değişkeni	
R değişkeni	C_1	C_2
R_1	f_{11}	f_{12}
R_2	f_{21}	f_{22}

R değişkeninin kategorileri R_1, R_2 ; C değişkeninin kategorileri C_1 ve C_2 olsun.

f_{11} : (R_1 ve C_1) kategorilerindeki birim sayısı

f_{12} : (R_1 ve C_2) " " "

f_{21} : (R_2 ve C_1) " " "

f_{22} : (R_2 ve C_2) "

2/b) İki yönlü (3x2) Olumsallık Tablosunun Genel Gösterimi

Değişkenler	C Değişkeni	
R Değişkeni	C_1	C_2
R_1	f_{11}	f_{12}
R_2	f_{21}	f_{22}
R_3	f_{31}	f_{32}

Örnek (İki yönlü (2x2) boyutlu): Bir lisansüstü programında yer alan bir dersindeki başarının, öğrencilerin lisans eğitimlerinin sosyal veya sayısal ağırlıklı bir bölümden mezun olmalarına bağlı olup olmadığını araştırmak amacıyla dersi alan 90 öğrencinin başarı durumu ve mezun oldukları bölümler incelenmiştir. 50 öğrencinin sayısal bölüm mezunu olduğu bilinmektedir. 72 başarılı öğrenciden 30'u sözel bölüm mezunudur. Olumsallık tablosunu düzenleyiniz.

Dersteki Başarının Bölümlere Göre Dağılımı

Değişkenler	Dersteki Başarı Durumu	
Bölüm	Başarılı	Başarısız
Sayısal	42	8
Sözel	30	10

3) Üç yönlü Olumsallık Tablosunun Genel Gösterimi

3 kategorik değişkenden oluşan üç yönlü bir olumsallık tablosunda değişkenlerden birincisi (X) iki, ikincisi (Y) iki, üçüncüsü (Z) iki kategorili ise tablonun boyutu (2x2x2) olacaktır. Olumsallık tablosundaki frekanslar (f_{ijk}) ile gösterilir.

i : X değişkeninin kategorisini
j : Y " "
k : Z " "

gösterir.

		Y değişkeni	
Z değ.	X değ.	Y_1	Y_2
Z_1	X_1	f_{111}	f_{121}
	X_2	f_{211}	f_{221}
Z_2	X_1	f_{112}	f_{122}
	X_2	f_{212}	f_{222}
Z_3	X_1	f_{113}	f_{123}
	X_2	f_{213}	f_{223}

2x2x3

Örneğin; f_{113}; X değişkeninin birinci, Y değişkeninin birinci ve Z değişkeninin üçüncü kategorisindeki birimlerin kaç kez tekrarlandığını gösterir.

Örnek (Üç yönlü (2x2x2) boyutlu): Bir fakültenin kadrosunda 90 öğretim elemanı bulunmaktadır. E-öğrenmeye geçiş öncesinde eğitimin etkinliği konusunda görüşler alınmıştır. Bu görüşlerin öğretim elemanının kadrosu ve cinsiyetine bağlı olup olmadığı araştırılacaktır. Öğretim elemanlarının kadrolarına ve cinsiyetlerine göre dağılımı aşağıdaki gibidir:

Öğretim elemanlarının 40'ı öğretim üyesi, 50'si araştırma görevlisidir.

Öğretim üyelerinin 22'si, araştırma görevlilerinin 20'si erkektir.

Kadın öğretim üyelerinin 15, erkek öğretim üyelerinin 21'i uygulanacak e-öğrenim programının etkin olacağını düşünmektedir.

Erkek araştırma görevlilerinin 18'i, kadın araştırma görevlilerinin 25'i e-öğrenim programının etkin olacağını düşünmektedir.

Olumsallık tablosunu düzenleyiniz.

Örnek (Üç yönlü (2x2x2) boyutlu): Bir fakültenin kadrosunda 90 öğretim elemanı bulunmaktadır. E-öğrenmeye geçiş öncesinde eğitimin etkinliği konusunda görüşler alınmıştır. Bu görüşlerin öğretim elemanının kadrosu ve cinsiyetine bağlı olup olmadığı araştırılacaktır. Öğretim elemanlarının kadroları ve cinsiyetlerine göre dağılımı aşağıdaki gibidir: (4)

Öğretim elemanlarının 40'ı öğretim üyesi, 50'si araştırma görevlisidir.

Öğretim üyelerinin 22'si, araştırma görevlilerinin 20'si erkektir.

Kadın öğretim üyelerinin 15, erkek öğretim üyelerinin 21'i uygulanacak e-öğrenim programının etkin olacağını düşünmektedir.

Erkek araştırma görevlilerinin 18'i, kadın araştırma görevlilerinin 25'i e-öğrenim programının etkin olacağını düşünmektedir.

Olumsallık tablosunu düzenleyiniz.

E-öğrenim Etkinliği ile İlgili Görüşlerin Cinsiyet ve Kadrolara Göre Dağılımı

E-öğr. etkinliği

Cinsiyet (Z)	Öğretim Elemanı (X)	Etkin	Etkin değil
Kadın	Öğretim Üyesi	15	3
	Araş. Gör.	25	5
Erkek	Öğretim Üyesi	21	1
	Araş. Gör.	18	2

Buna göre olumsallık tablolarından, satır-sütun toplamlarına ve/veya oranlara, olasılıklara dayanan birçok özet bilgi elde edilebilir. Örneğin yukarıdaki tabloda, kadın araştırma görevlilerinin tüm araştırma görevlileri içindeki oranı (0,6), tüm öğretim elemanları içindeki oranı ise (0,33)'dür.