KATEGORİK VERİLERİN ANALİZİ
(Uyum İyiliği, Bağımsızlık ve Dağılıma Uygunluk Testleri)

Günümüzde yapılan birçok araştırmada nitel değişkenler kullanılmaktadır. Göz rengi, saç rengi gibi bazı değişkenler yapıları gereği nitel olarak gözlenirler. Bunun yanında nicel olan bazı değişkenler de sınıflandırılarak nitel hale dönüştürülebilir. Örneğin belli bir bitkinin boyu göz önüne alındığında; bu bitkinin boyu cm olarak ölçülebileceği gibi belli bir cm den uzun ya da kısa olmak üzere iki sınıfa ayrılarak da incelenebilir. Böylece aslında nicel olan bir değişken nitel olarak gözlenmiş olur.

Nitel yapıya sahip değişkenler üzerinde yapılan gözlemler genel olarak, belli özelliğe sahip olanların sayısı biçiminde ifade edilir. Örneğin bir sınıftaki renkli gözlü öğrencilerin sayısı ya da boyu 160 cm den kısa olanların sayısı gibi. Elimizde bu şekilde ölçülen değişkenlere ait gözlemler olduğunda bunlara ilişkin sonuç çıkarımı için daha önceki bölümlerde ifade ettiğimiz testleri kullanmak uygun değildir. Bu gibi durumlarda Ki-kare testleri kullanılmaktadır.

1. Uyum İyiliği Testleri

Birçok araştırmada gözlenen frekansların H_0 hipotezinde öne sürülen teorik frekanslara uygun olup olmadığı araştırılmak istenir. Örneğin belli bir hastalığa yakalanan kişilerin kan gruplarının dağılımının normal kitledeki dağılıma uygun olup olmadığı araştırılmak istenebilir.

Bu şekilde gözlenen frekansların beklenen teorik frekanslar uyup uymadığı araştırıldığında aşağıda verilen χ^2 değeri kullanılır,

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_{gi} - f_{bi})^2}{f_{bi}}$$

Burada,

f_{gi} : Gözlenen (deneysel) frekans

f_{bi} : Beklenen (kuramsal) frekans

k : Sınıf sayısı

dır.

Eşitlikten de kolayca anlaşılacağı üzere eğer gözlenen frekanslar beklenen frekanslara yakın ise χ^2 değeri küçük, aksi halde χ^2 değeri büyük çıkacaktır. Buna göre χ^2 değerinin büyük çıkması H_0 hipotezinin reddedilmesine neden olacaktır. Hesaplanan bu χ^2 değeri yaklaşık olarak χ^2 dağılımına sahip olur ve χ^2 tablosu kullanılarak hipotez reddedilir ya da edilemez. Eğer beklenen frekanslar hesaplanırken örneklem verilerinden yararlanarak parametre tahmini yapılmamışsa hesaplanan bu χ^2 değeri yaklaşık olarak serbestlik derecesi $k-1$ olan χ^2 dağılımına sahip olacaktır. Aksi takdirde kaç tane parametre tahmini yapıldı ise o kadar serbestlik derecesinden düşülerek serbestlik derecesi hesaplanacaktır.

Örnek: Dört çocuklu 100 ailede erkek çocukların dağılımı aşağıda verildiği gibidir.

Erkek Çocuk Sayısı	Aile Sayısı
0	7
1	23
2	36
3	20
4	14

Erkek çocukların dağılımın Binom dağılımına uyumunu 0.05 anlam düzeyinde test ediniz.

Çözüm: Hipotez aşağıda verildiği gibi olacaktır.

H_0: Erkek çocukların sayısının dağılımı Binom dağılımına uygundur
H_1: Erkek çocukların sayısının dağılımı Binom dağılımına uygun değildir

İlk olarak Binom dağılımının parametreleri olan n ve p değerleri belirlenmelidir.

4 çocuklu ailelerden örnek alındığı için $n = 4$ dür. Fakat, p değeri bilinmediği için tahmin edilmesi gerekir.

$$\hat{p} = \frac{Toplam\ erkek\ çocuk\ sayısı}{Toplam\ çocuk\ sayısı} = \frac{0 \times 7 + 1 \times 23 + 2 \times 36 + 3 \times 20 + 4 \times 14}{4 \times 100}$$
$$= \frac{211}{400} \cong 0.53$$

Hesaplanan bu p değeri ve verilen n değeri kullanılarak beklenen frekansları bulmak için ilk olarak aşağıdaki olasılıklar hesaplanmalıdır.

$$P(X = 0) = \binom{4}{0} 0.53^0 0.47^4 = 0.0488$$
$$P(X = 1) = \binom{4}{1} 0.53^1 0.47^3 = 0.2201$$
$$P(X = 2) = \binom{4}{2} 0.53^2 0.47^2 = 0.3723$$
$$P(X = 3) = \binom{4}{3} 0.53^3 0.47^1 = 0.2799$$
$$P(X = 4) = \binom{4}{4} 0.53^4 0.47^0 = 0.0789$$

Beklenen frekansları bulmak için bulunan bu olasılıklar aile sayısı olan 100 değeri ile çarpılmalıdır.

f_{bi}
$0.0488 \times 100 \cong 5$
$0.2201 \times 100 \cong 22$
$0.3723 \times 100 \cong 37$
$0.2799 \times 100 \cong 28$
$0.0789 \times 100 \cong 8$

→ 100 aileden 5 tanesinin 0 erkek çocuğa sahip olması beklenir

$\sum_{i=1}^{k} f_{gi} = \sum_{i=1}^{k} f_{bi} = n$ olacak şekilde yuvarlatmalar yapılmalıdır.

$$\chi_h^2 = \sum_{i=1}^{5} \frac{(f_{gi} - f_{bi})^2}{f_{bi}} = \frac{(7-5)^2}{5} + \frac{(23-22)^2}{22} + \frac{(36-37)^2}{37} + \frac{(20-28)^2}{28} + \frac{(14-8)^2}{8}$$
$$= 7.65$$

Tablo değerine bakabilmek için serbestlik derecesi belirlenmelidir. Bir parametre örnekten tahmin edildi buna göre,

$$sd = k - 1 - tahmin\ edilen\ parametre\ sayısı$$
$$= 5 - 1 - 1 = 3$$

olup $\chi^2_{3,\alpha=0.05} = 7.81$ olarak bulunur.

$\chi_h^2 < \chi_t^2$ olduğundan H_0 reddedilemez. Yani, erkek çocukların dağılımı Binom dağılımına uygundur şeklinde yorum yapılabilir.

χ^2 testleri, uyum iyiliği testlerinin yanında bağımsızlık ve homojenliklerin test edilmesinde de kullanılır. χ^2 testi yapabilmek için gözlemler çapraz tablolar ile özetlenirler.

a) Tek Gruplu Çapraz Tablo

Burada incelenen gözlemler belirli bir özelliğe göre düzeylere ayrılmıştır. Genellikle uyum iyiliği testlerinde kullanılır.

Sınıflar	Frekanslar
1	f_1
2	f_2
3	f_3
4	f_4

Test istatistiği;

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_{gi} - f_{bi})^2}{f_{bi}}$$

biçiminde hesaplanır. Burada;

χ^2 : Ki-kare değeri
f_{gi} : Gözlenen (deneysel) frekans
f_{bi} : Beklenen (kuramsal) frekans
k : Sınıf sayısı

$$\sum_{i=1}^{k} f_{gi} = \sum_{i=1}^{k} f_{bi} = n$$

dir.

b) 2 × 2 Çapraz Tablo

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(f_{ij} - f'_{ij})^2}{f'_{ij}}$$

ya da

X / Y	1	2	Toplam
1	$f_{11} = \boldsymbol{a}$	$f_{12} = \boldsymbol{b}$	$n_{1.} = \boldsymbol{A}$
2	$f_{21} = \boldsymbol{c}$	$f_{22} = \boldsymbol{d}$	$n_{2.} = \boldsymbol{B}$
Toplam	$n_{.1} = \boldsymbol{C}$	$n_{.2} = \boldsymbol{D}$	n

olmak üzere 2 × 2 çapraz tablo için;

$$\chi^2 = \frac{(ad - bc)^2 n}{ABCD}$$

ile hesaplanır. Görüldüğü gibi, bu formül ile beklenen frekansların hesaplanması gerekmez, sadece gözlenen frekanslar kullanılır.

c) R xC Tipinde Çapraz Tablo

R : Satır sayısı C : Sütun sayısı olmak üzere,

X / Y	1	2	⋯	C	**Toplam**
1	f_{11}	f_{12}	⋯	f_{1C}	$\boldsymbol{n_{1.}}$
2	f_{21}	f_{22}	⋯	f_{2C}	$\boldsymbol{n_{2.}}$
⋮	⋮				⋮
R	f_{R1}	f_{R2}	⋯	f_{RC}	$\boldsymbol{n_{R.}}$
Toplam	$\boldsymbol{n_{.1}}$	$\boldsymbol{n_{.2}}$	⋯	$\boldsymbol{n_{.C}}$	$\boldsymbol{n}$

$$\sum_{i=1}^{R} n_{.i} = \sum_{j=1}^{C} n_{.j} = n$$

$$\sum_{i=1}^{R}\sum_{j=1}^{C} f_{ij} = \sum_{i=1}^{R}\sum_{j=1}^{C} f_{ij}{}' = n$$

$$\chi^2 = \sum_{i=1}^{R}\sum_{j=1}^{C} \frac{\left(f_{ij} - f'_{ij}\right)^2}{f'_{ij}}$$

f_{ij} :gözlenen frekans

f'_{ij} :beklenen frekans

χ^2 çözümlemesi yapılırken dikkat edilmesi gereken bir takım kurallar vardır;

1) Yapılacak tahminlerin güvenilir olması için beklenen frekansların en az 5 olması istenir. Eğer 5'ten az beklenen (kuramsal) frekans değeri varsa, bu frekansın yer aldığı satır ya da sütun tablodaki uygun bir satır ya da sütun ile birleştirilir.

2) Uygulamada genellikle gözlem sayısı $n < 50$ ve serbestlik derecesi 1 olduğunda Yates düzeltmesi yapılması gerektiği söylenir. Bu durumda χ^2 değişkeninin düzeltilmiş ve düzeltilmemiş değerlerinin karşılaştırılmasında yarar vardır; eğer her iki χ^2 değeri birbirinden çok farklı ise denek sayısı yetersizdir, örneklem genişliği artırılmalıdır.

3) i.inci satırdaki gözlenen frekans ile i.inci satırdaki beklenen frekans toplamları birbirine eşit olmalıdır. Aynı şekilde j.inci sütundaki gözlenen frekans ile j.inci satırdaki beklenen frekans toplamları birbirine eşit olmalıdır.

4) 2×2 düzeninde çapraz tabloda 5'den az sayıda gözlenen frekans varsa, önemlilik testi Fisher'in özel eşitliği uygulanarak yapılır.

Sayımla Belirtilen Kitlelerde Bağımsızlık Kontrolü

İki ya da daha çok grubun bağımsız olup olmadığı kontrol edilir. Veriler çapraz tablo üzerinde gösterilebilir.

1. Hipotez kurulur.

$p_{ij} = (p_{i.})(p_{.j})$

p_{ij} :birimin i inci satır ve j inci sütuna düşme olasılığı

$p_{i.}$:birimin i inci satıra düşmesi olasılığı

$p_{.j}$:birimin j inci sütuna düşmesi olasılığı

H_0: Değişkenler bağımsızdır.(Gruplar arası fark yoktur.)
H_1: Değişkenler bağımlıdır.(Gruplar arası fark vardır.)

2. Test istatistiği belirlenir.

$R \times C$ tablosu düzenlenir ve

f'_{ij} kuramsal sıklığı (i inci satır, j inci sütuna düşmesi gereken frekans) şöyle bulunur:

$$f'_{ij} = (\hat{p}_{i.})(\hat{p}_{.j})n$$

Burada,

$p_{i.} = \frac{n_{i.}}{n}$ ve $p_{.j} = \frac{n_{.j}}{n}$

dir. Buradan,

$$f'_{ij} = \frac{(n_{i.})(n_{.j})}{n}, i = 1, \dots, R(satır) \quad , \quad j = 1, \dots, C(sütun)$$

$$\chi^2 = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{\left(f_{ij} - f'_{ij}\right)^2}{f'_{ij}}$$

R:satır sayısı

C:sütun sayısı

f_{ij}:gözlenen frekans

Yates Düzeltmeli Formülü

$$\chi_H^2 = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{(\left|f_{ij} - f'_{ij}\right| - 0{,}5)^2}{f'_{ij}}$$

2×2 düzeninde çapraz tablolar için;

$\chi_H^2 = \frac{(ad-bc)^2 n}{ABCD}$, Yates düzeltmeli $\chi_H^2 = \frac{(|ad-bc|-n/2)^2 n}{ABCD}$

3. Yorum ve karar

$\chi_H^2 > \chi^2_{T(\alpha, s'=(R-1)(C-1))}$ ise H_0 red edilir.

Bir değişkenin ikiden çok düzeyi olduğunda ya da ikiden çok grup karşılaştırıldığında, gruplar arası fark önemli ise, gruplardan hangisinin diğerlerinden önemli derecede farklı olduğunu araştırmak gerekir. Bu nedenle en büyük χ^2 değerine sahip olan grup işlemden çıkarılır. Geriye kalan gruplar için, hipotez kabul edilinceye kadar işlemlere devam edilir.

Örnek: Yurtta kalıp kalmamayla başarı arasında bir ilişki var mıdır? $\alpha = 0.01$ anlam düzeyinde test ediniz.

	Başarısız	Başarılı	Çok başarılı	Toplam
Yurtta kalan	45	35	20	100
Yurtta kalmayan	15	85	10	110
Toplam	60	120	30	200

$f'_{11} = \frac{100*60}{210} = 28.6$ $f'_{12} = \frac{100*120}{210} = 57.1$ $f'_{13} = \frac{100*30}{210} = 14.3$

$f'_{21} = \frac{110*60}{210} = 31.4$ $f'_{22} = \frac{110*120}{210} = 62.9$ $f'_{23} = \frac{110*30}{210} = 5.7$

1. H_0:Yurtta kalıp kalmamayla başarı arasında ilişki yoktur.
   H_1: Yurtta kalıp kalmamayla başarı arasında ilişki vardır.

2. $\chi^2_H = \sum_{i=1}^{R}\sum_{j=1}^{C} \frac{\left(f_{ij}-f'_{ij}\right)^2}{f'_{ij}} = \frac{(45-28.6)^2}{28.6} + \frac{(35-57.1)^2}{57.1} + \cdots + \frac{(10-15.7)^2}{15.7} = 38.62$

3. $\chi^2_{T(\alpha,sd=(R-1)(C-1)=0.01,(2-1)*(3-1)=2)} = 9.21$

   $\chi^2_H > \chi^2_T$ olduğundan H_0 red edilir.

Yorum: Yurtta kalıp kalmamayla başarı arasında bir ilişkinin olduğu %99 güven düzeyinde söylenebilir.

Örnek. Rasgele seçilen 100 kişiye yapılan bir ankete göre yeni bir ürünün beğenilip beğenilmemesinin cinsiyete göre dağılımı aşağıdaki tabloda verilmiştir. Cinsiyetin ürünün beğenilip beğenilmemesiyle ilgisi var mıdır, $\alpha = 0.05$ anlam düzeyinde test ediniz.

	Beğendi	Beğenmedi	Toplam
Kadın	50=a	12=b	62=A
Erkek	20=c	18=d	38=B
Toplam	70=C	30=D	100=n

2×2' lik bir tablo $\longrightarrow$ f_{ij}' lerden herhangi biri 5' ten küçük değil.

H_0:Cinsiyet ile ürünü beğenip beğenmeme arasında ilgi yoktur.
H_1: Cinsiyet ile ürünü beğenip beğenmeme arasında ilgi vardır.

$$\chi_H^2 = \frac{(ad - bc)^2 n}{ABCD}$$

$$\chi_H^2 = \frac{(50 \times 18 - 12 \times 20)^2 \times 100}{62 \times 38 \times 70 \times 30} = 8.804$$

$$\chi^2_{T(0.05,(2-1)\times(2-1)=1)} = 3.84$$

$\chi_H^2 > \chi_T^2$ olduğundan H_0 red edilir. Yani, Cinsiyetin ürünü beğenip beğenmeme üzerinde önemli bir etkisi olduğu söylenebilir.