İST268 DÖNEM ÖDEVİ

Ahmet Ürkmez - Barış Karataş

2200329023 - 2200329066

2022-05-08

a) Bir adet "ANOVA" verisi bularak ANOVA çözümlemesi yapınız. Gruplar arasında anlamlı farklılık varsa LSD testi yapınız.

Verilerin Vektör Olarak Tanımlanması

New Mexico'daki bir arkeolojik alandaki dört farklı kazı alanı, önemli arkeolojik keşifler için aşağıdaki derinlikleri (cm) verdi:

X1 = Alan I'deki derinlikler X2 = Alan II'deki derinlikler

X3 = Alan III'teki derinlikler X4 = Alan IV'teki derinlikler

```
X1 <- c(93, 120, 65, 105, 115, 82, 99, 87, 100, 90, 78, 95, 93, 88, 110)
X2 <- c(85, 45, 80, 28, 75, 70, 65, 55, 50, 40)
X3 <- c(100, 75, 65, 40, 73, 65, 50, 30, 45, 50, 45, 55)
X4 <- c(96, 58, 95, 90, 65, 80, 85, 95, 82)
```

İki Boyutlu Verinin Oluşturulması ve Yazdırılması

```
Sites <- c(X1, X2, X3, X4)
Groups <- c(rep("Alan I", 15), rep("Alan II", 10),
            rep("Alan III", 12), rep("Alan IV", 9))

Data <- data.frame(Sites, Groups); head(Data)

##   Sites Groups
## 1    93 Alan I
## 2   120 Alan I
## 3    65 Alan I
## 4   105 Alan I
## 5   115 Alan I
## 6    82 Alan I
```

Normallik Varsayımının Test Edilmesi

H_0: Verilerin takip ettiği dağılım ile normal dağılım arasında fark yoktur.

H_1: Verilerin takip ettiği dağılım ile normal dağılım arasında fark vardır.

Her bir alan için elde edilen “p-value” değerleri “0,05” anlamlılık düzeyinden büyük olduğu için H_0 hipotezi reddedilemez, verilerin normal dağılımı takip ettiği varsayılabilir.

```
shapiro.test(X1); shapiro.test(X2); shapiro.test(X3); shapiro.test(X4)

## 
##  Shapiro-Wilk normality test
## 
## data:  X1
## W = 0.98653, p-value = 0.996

## 
##  Shapiro-Wilk normality test
## 
## data:  X2
## W = 0.96729, p-value = 0.8647

## 
##  Shapiro-Wilk normality test
## 
## data:  X3
## W = 0.94757, p-value = 0.6017

## 
##  Shapiro-Wilk normality test
## 
## data:  X4
## W = 0.87404, p-value = 0.1358
```

Varyansların Homojenliğinin Test Edilmesi

H_0: $\sigma^2_{X1} = \sigma^2_{X2} = \sigma^2_{X3} = \sigma^2_{X4}$

H_1: En az bir σ^2_{Xi} değeri diğerlerinden farklıdır. (i={1, 2, 3, 4})

Alan verilerinin normal dağılımı takip ettiğine dair güçlü kanıtlarımız olduğu için homojenlik incelemesinde Bartlett testinden yararlanacağız. Bartlett testi sonucunda elde edilen p değeri “0.05” anlamlılık düzeyinden büyük olduğu için H_0 hipotezi reddedilemez. Varyansların %5 anlamlılık düzeyinde eşit olduğu varsayılabilir.

```
bartlett.test(Sites ~ Groups, data = Data)

## 
##  Bartlett test of homogeneity of variances
## 
## data:  Sites by Groups
## Bartlett's K-squared = 1.7355, df = 3, p-value = 0.6291
```

Gruplar Arası Farkın Test Edilmesi

H_0: $\mu_{X1} = \mu_{X2} = \mu_{X3} = \mu_{X4}$

H_1: En az bir μ_{Xi} diğerlerinden farklıdır. (i = {1, 2, 3, 4})

Test sonucunda elde edilen p değeri 0'dan önemli ölçüde farklı değildir ve "0.05" anlamlılık düzeyinden küçüktür. Aynı zamanda elde edilen "F = 15.139" değeri, tablo değeri olan "$F_{0.05;\,3,42} \cong F_{0.05;\,3,40} = 2.84$" değerinden büyük olduğundan H_0 hipotezi reddedilir. Gruplar arasında fark olduğunu %5 anlamlılık düzeyinde söyleyebiliriz.

```
oneway.test(Sites ~ Groups, data = Data, var.equal = T)

##
##  One-way analysis of means
##
## data:  Sites and Groups
## F = 15.139, num df = 3, denom df = 42, p-value = 7.991e-07
```

Çoklu Karşılaştırma Testi

Matristeki "0.05" anlam düzeyinden küçük elemanları inceledeğimizde farkı oluşturan alanların "Alan I" ve "Alan IV" olduğunu görebiliriz. Önemli arkeolojik keşifler için elde edilen derinlikler "Alan I" ve "Alan IV "de daha büyüktür.

(Alan I = 94.6667 cm, Alan II = 59.3 cm , Alan III = 57.75 cm ve Alan IV = 82.88889 cm)

```
pairwise.t.test(Sites, Groups, p.adj="bonferroni",data = Data)

##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  Sites and Groups
##
##          Alan I  Alan II Alan III
## Alan II  2.9e-05 -       -
## Alan III 5.1e-06 1.0000  -
## Alan IV  0.5898  0.0203  0.0077
##
## P value adjustment method: bonferroni
```

b) Bir adet "Bağımsız iki örneklem" verisi bularak ve iki grubu karşılaştırınız. İki ortalama arası fark için %95 güven aralığını hesaplayınız.

Verilerin Vektör Olarak Tanımlanması

Aşağıdaki veriler, kamyonet modellerinin bağımsız rastgele örnekleri için perakende fiyatını (bin dolar olarak) temsil etmektedir.

X1 = I. Model kamyon   X2 = II. Model kamyon

```
X1 <- c(17.4, 23.3, 29.2, 19.2, 17.6, 19.2, 23.6, 19.5, 22.2, 24.0, 26.4,
23.7, 29.4, 23.7, 26.7, 24.0, 24.9)
X2 <- c(17.5, 23.7, 20.8, 22.5, 24.3, 26.7, 24.5, 17.8, 29.4, 29.7, 20.1,
21.1, 22.1, 24.2, 27.4, 28.1)
```

İki Boyutlu Verinin Oluşturulması

```
Data <- data.frame(c(X1, X2), c(rep("I. Model", 17), rep("II. Model", 16)))
colnames(Data) <- c("Fiyat", "Model")

attach(Data); head(Data)

##   Fiyat    Model
## 1  17.4 I. Model
## 2  23.3 I. Model
## 3  29.2 I. Model
## 4  19.2 I. Model
## 5  17.6 I. Model
## 6  19.2 I. Model
```

Normallik Varsayımının İncelenmesi

H_0: Verilerin takip ettiği dağılım ile normal dağılım arasında fark yoktur.

H_1: Verilerin takip ettiği dağılım ile normal dağılım arasında fark vardır.

Elde edilen "p-value" değerleri incelendiğinde "0.05" anlamlılık düzeyinden büyük olduğu için H_0 hipotezi reddedilemez verilerin normal dağılımı takip ettiği varsayılabilir.

```
shapiro.test(X1); shapiro.test(X2)

##
##  Shapiro-Wilk normality test
##
## data:  X1
## W = 0.94083, p-value = 0.3281

##
##  Shapiro-Wilk normality test
##
## data:  X2
## W = 0.96167, p-value = 0.6921
```

Varyansların Homojenliğinin İncelenmesi

H_0: $\sigma_{X1}^2 = \sigma_{X2}^2$

H_1: $\sigma_{X1}^2 \neq \sigma_{X2}^2$

I. Yöntem

Elde edilen "p-value" değeri "0.05" anlamlılık düzeyinden büyüktür. Aynı zamanda "F = 0.92157" değeri "$F_{\alpha/2\,=\,0.025;16,15}$= 2.8360467" değerinden küçük olduğu için varyansların eşit olduğu varsayılabilir. Aralığın "1" i içermesi durumu da farkın anlamsız olduğunu belirtir.

H_0 reddedilemez, varyanslar %5 anlamlılık düzeyinde homojendir.

```
var.test(X1, X2)

##
##  F test to compare two variances
##
## data:  X1 and X2
## F = 0.92157, num df = 16, denom df = 15, p-value = 0.87
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.3249501 2.5689027
## sample estimates:
## ratio of variances
##          0.9215737
```

II. Yöntem

Farklı bir bakış açısı olarak Levene testi sonucunda elde edilen "p-value" değeri için aynı yorum yapılabilir.

```
library(lawstat)

levene.test(Fiyat, Model, location = "mean")

##
##  Classical Levene's test based on the absolute deviations from the mean
##  ( none not applied because the location is not set to median )
##
## data:  Fiyat
## Test Statistic = 0.087871, p-value = 0.7689
```

"t-test" İstatistiğinin Elde Edilmesi

$H_0: \mu_{X1} = \mu_{X2}$

$H_1: \mu_{X1} \neq \mu_{X2}$

Test sonucunda elde ettiğimiz "p-value" değeri "0.05" anlamlılık düzeyinden büyüktür ve hesap değeri olan "|t| = 0.443699" değeri, tablo değeri olan "$t_{\alpha/2\,=\,0.025;31}$= 2.039513" değerinden küçük olduğu için H_0 hipotezi reddedilemez. I. ve II. model kamyonların arasında fiyat bakımından %5 anlamlılık düzeyinde fark olmadığı söylenebilir.

```
t.test(X1 ,X2, mu = 0, alternative="two.sided", var.equal = T,
conf.level=0.95)

##
##  Two Sample t-test
##
## data:  X1 and X2
## t = -0.43699, df = 31, p-value = 0.6651
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.214866  2.080307
```

```
## sample estimates:
## mean of x mean of y
##  23.17647  23.74375
```

c) Bir adet "Bağımlı iki örneklem" verisi bulunuz ve iki bağımlı örneklem karşılaştırması yapınız.

Verilerin Vektör Olarak Tanımlanması

Aşağıdaki bağımlı veriler, Dobson biriminde ozon kolonunun kalınlığını temsil etmektedir.

(Dobson Birimi: Standart sıcaklık ve basınçta bir mili-santimetre ozon.)

Jan = Ocak ayında aylık ortalama kalınlık

Feb = Şubat ayında aylık ortalama kalınlık

Veriler, 15 yıllık rastgele bir örneklem için yıllara göre eşleştirilmiştir.

```
Jan <- c(360, 324, 377, 336, 383, 361, 369, 349, 301, 354, 344, 329, 337,
387, 378)
Feb <- c(365, 325, 359, 352, 397, 351, 367, 397, 335, 338, 349, 393, 370,
400, 411)
```

İki Boyutlu Verinin Oluşturulması

```
Data <- data.frame(c(rep("Ocak", 15), rep("Subat", 15)), c(Jan, Feb))
colnames(Data) <- c("Aylar", "Ozon")

head(Data)

##   Aylar Ozon
## 1  Ocak  360
## 2  Ocak  324
## 3  Ocak  377
## 4  Ocak  336
## 5  Ocak  383
## 6  Ocak  361
```

Normallik Varsayımını İncelenmesi

H_0: Verilerin takip ettiği dağılım ile normal dağılım arasında fark yoktur.

H_1: Verilerin takip ettiği dağılım ile normal dağılım arasından fark vardır.

Elde edilen "p-value" değeri "0.05" anlamlılık düzeyinden büyük olduğu için H_0 hipotezi reddedilemez, verilerin normal dağılımı takip ettiği varsayılabilir.

```
Diff <- with(Data, Ozon[Aylar == "Ocak"] - Ozon[Aylar == "Subat"])

shapiro.test(Diff)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Diff
## W = 0.95448, p-value = 0.5975
```

"t-test" İstatistiğinin Elde Edilmesi

$H_0: \mu_{Diff} = 0$

Elde edilen "p-value" değeri "0.05" anlamlılık düzeyinden küçüktür ve "|t| = 2.3968" istatistiği tablo değeri olan "t = 2.144787" değerinden büyük olduğu için H_0 hipotezi reddedilir. Aylara göre ozon tabakasının kalınlığının değiştiğini %5 anlamlılık düzeyinde söyleyebiliriz. Aralığın "0" ı içermemesi de farkın anlamlı olduğunu gösterir. Aralık negatif olduğuna göre, şubat ayında ozon kolonunun kalınlığı daha fazladır.

```
t.test(Diff, mu=0, alternative="two.sided", conf.level = .95)

##
##  One Sample t-test
##
## data:  Diff
## t = -2.3968, df = 14, p-value = 0.03106
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -27.791017  -1.542317
## sample estimates:
## mean of x
## -14.66667
```

d) Bir adet "RxC tablosu" bulunuz ve bağımsızlık çözümlemesi yapınız. İlişki anlamlı ise ilişki katsayısını hesaplayınız.

R*C Tablosu

		DISEASES			
	VALUES	***Ulcer***	***Cancer***	***Control***	
B. TYPE	***O***	***983***	***383***	***2892***	***4258***
	A	***679***	***416***	***2625***	***3720***
	B	***134***	***84***	***570***	***788***
		1796	***883***	***6078***	***8766***

İki Boyutlu Verinin Oluşturulması

```
frequencies = matrix(c(983, 383, 2892,
                 679, 416, 2625,
                 134, 84, 570), ncol = 3, byrow = TRUE)

colnames(frequencies) = c('Ulcer', 'Cancer', 'Control')
rownames(frequencies) = c('O', 'A', 'B')

table = as.table(frequencies); table

##   Ulcer Cancer Control
## O   983    383    2892
## A   679    416    2625
## B   134     84     570
```

Ki-Kare Dağılımı Varsayımı İçin Beklenen Sıklıkların Elde Edilmesi

Beklenen sıklık matrisinde "5" ten küçük sıklık yoktur. Ki-kare varsayımında bozulma olmadığı söylenebilir.

```
chisq.test(table)$expected

##      Ulcer    Cancer   Control
## O 872.3897 428.90874 2956.7016
## A 762.1629 374.71595 2583.1211
## B 161.4474  79.37531  547.1773
```

Pearson'un Ki-Kare Testi

H_0: Kan grupları ile hastalıklar arasında anlamlı bir ilişki yoktur.

H_1: Kan grupları ile hastalıklar arasında anlamlı bir ilişki vardır.

Elde edilen "p-value" değeri "0" dan oldukça küçüktür ve "X-squared = 40.543" değeri "$X^2_{0.05:4}$ =9.48773" değerinden büyük olduğu için H_0 hipotezi reddedilir. Kan grupları ile hastalıklar arasında anlamlı bir ilişki olduğu söylenebilir.

```
chisq.test(table)

##
##  Pearson's Chi-squared test
##
## data:  table
## X-squared = 40.543, df = 4, p-value = 3.341e-08
```

İlişki Katsayısının Hesaplanması

H_0: İlişki katsayısı anlamlı değildir.

H_1: İlişki katsayısı anlamlıdır.

Kan grubu ve hastalık sınıflanabilir kategorik değişkenler olduğu için ilişki derecesinin hesaplanmasında “Cramer V” değerini ele alacağız.

Alt sınır ve üst sınır incelendiğinde aralığın “0” ı içermediğini görürüz. Dolayısıyla H_0 hipotezi reddedilir. Kan grubu ve hastalıklar arasında %4’lük anlamlı bir ilişki olduğu söylenebilir.

```
library(DescTools)

CramerV(table, conf.level = .95)

##   Cramer V     lwr.ci     upr.ci
## 0.04808884 0.03107847 0.06131299
```

e) Bir adet “İki Binom oranı” verisi bulunuz ve iki oranı karşılaştırarak iki oran arasındaki fark için %90 güven aralığını hesaplayınız.

	ESKİ BİNA	YENİ BİNA	TOPLAM
1. MAHALLE	30	10	40
2. MAHALLE	35	15	50

(https://personel.omu.edu.tr/docs/ders_dokumanlari/1028_76085_1500.pdf)

Soru: İki mahaldeki eski bina oranlarının farklı olduğu iddia edilmektedir. Bu amaçla I. mahalleden rastgele seçilen 40 binanın 30’u, II. Mahalleden ise 50 binanın 35’i eskidir. %90 güven düzeyine olduğuna göre iddianın doğruluğunu test ediniz. (α=0,10)

H0:P1=P2
H1:P1≠P2

Method

p₁: proportion where Sample 1 = Event
p₂: proportion where Sample 2 = Event
Difference: p₁ - p₂

Descriptive Statistics

Sample	N	Event	Sample p
Sample 1	40	30	0,750000
Sample 2	50	35	0,700000

Estimation for Difference

Difference	90% CI for Difference
0,05	(-0,105066; 0,205066)

CI based on normal approximation

Test

Null hypothesis $H_0: p_1 - p_2 = 0$
Alternative hypothesis $H_1: p_1 - p_2 \neq 0$

Method	Z-Value	P-Value
Normal approximation	0,53	0,599
Fisher's exact		0,643

The test based on the normal approximation uses the pooled estimate of the proportion (0,722222).

1. (p=0.599) > 0.10 → p değeri anlamlılık değerinden (α) büyüktür.
2. |Z|=|0.53|=0.53 test istatistiği $Z_{\alpha/2}$=$Z_{0.05}$=1.645 tablo değerinden küçüktür.
3. (-0.105, 0.205) güven aralığı 0'ı içeriyor.

Z testi sonucuna göre H_0 hipotezi reddedilemez. Yani, iki mahalledeki eski bina oranları arasında %90 güven düzeyinde anlamlı bir farklılık bulunmamıştır.

f) Bir adet sıklık çizelgesi üzerinde Poisson ya da Binom dağılımına uygun olup olmadığını uyum iyiliği testi ile inceleyiniz.

Dört çocuklu 100 ailede erkek çocukların dağılımı aşağıda verildiği gibidir. Erkek çocukların dağılımın Binom dağılımına uyumunu 0.10 anlam düzeyinde test edelim.

Erkek çocuk sayısı	Aile sayısı
0	7
2	23
3	36
4	20
5	14

(https://acikders.ankara.edu.tr/pluginfile.php/92910/mod_resource/content/0/ders8.pdf)

***H0*: Verilerin dağılımı binom dağılımına uygundur.**
***HS*: Verilerin dağılımı binom dağılımına uygun değildir.**

Binom Olasılıklarının Hesaplanması:

↓	C1	C2	C3
	Erkek çocuk sayısı	Aile Sayısı	p
1	0	7	0,048797
2	1	23	0,220105
3	2	36	0,372305
4	3	20	0,279889
5	4	14	0,078905

Ki-Kare Uyum İyiliği Testi

Chi-Square Goodness-of-Fit Test for Observed Counts in Veriable: Aile sayısı

Observed and Expected Counts

Category	Observed	Test Proportion	Expected	Contribution to Chi-Square
0	7	0,048797	4,8797	0,92132
1	23	0,220105	22,0105	0,04449
2	36	0,372305	37,2305	0,04067
3	20	0,279889	27,9889	2,28027
4	14	0,078905	7,8905	4,73054

1 (20,00%) of the expected counts are less than 5.

Chi-Square Test

N	DF	Chi-Sq	P-Value
100	4	8,01728	0,091

Sıklık dağılımında 5’ten küçük beklenen sıklık sayısı yukarıda belirtildiği gibi %20 olduğu için toplam sınıf sayısının %20’sini aşmamış ve düzey birleştirmesine gerek kalmamıştır.

p değeri bilinmediği ve örneklemden tahmin edildiği için t ‘‘1’’ değerini almakta ve serbestlik derecesi sd=k-t-1=5-1-1=3 olarak elde edilmektedir.

$\boldsymbol{\chi 2} = 8,01728 > \boldsymbol{\chi 2}_{3;0,1} = 6,251$ veya $p=0,091 < \alpha=0,1$ olduğu için H0 reddedilir. %10 anlamlılık düzeyinde verilerin dağılımı Binom dağılımına uyum sağlamamaktadır.

Kaynakça

Birinci, ikinci ve üçüncü soru için,

https://college.cengage.com/mathematics/brase/understandable_statistics/7e/students/datasets/tvis/frames/frame.html

Dördüncü soru için,

https://www.openepi.com/RbyC/RbyC.htm

Beşinci soru için,

https://personel.omu.edu.tr/docs/ders_dokumanlari/1028_76085_1500.pdf

Altıncı soru için,

https://acikders.ankara.edu.tr/pluginfile.php/92910/mod_resource/content/0/ders8.pdf

(5. ve 6. Soru için nasıl yapıldığı en azından olmadığı için değerlendirebiliriz bu seferlik demiştiniz. Yorumlama ve hipotez kurma kısmında da “ders pdf lerimizden” yararlandık. Kendi gördüğümüz notasyona ve içeriğe bağlı kaldık.) (O pdflerin çözümleri kafa karıştırıcıydı, sadece veriyi çektik.)