İstatistiksel Analiz Uygulamaları Ara Sınav Ödevi

Ramazan Erduran, İlkay Şafak Baytar, Talha Kurt, Mert Hasan Kılıç, Müstecep Berca AKBAYIR

09 Aralık 2021

### Kullanılan Kütüphaneler ve açıklamaları:

**library(dplyr) (tidyverse)**: Verileri manipüle edebilmek için kullandık.  
**library(readr)**: Verilerimizi içeri aktarmak için kullandık.  
Örneklemimiz:

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(readr)  
dataset\_ex <- read\_delim("~/GitHub/Statistical-Analyze/datasets/dataset\_ex.csv",   
 delim = "\t", escape\_double = FALSE,   
 trim\_ws = TRUE)

## Rows: 2240 Columns: 29

## -- Column specification --------------------------------------------------------  
## Delimiter: "\t"  
## chr (3): Education, Marital\_Status, Dt\_Customer  
## dbl (26): ID, Year\_Birth, Income, Kidhome, Teenhome, Recency, MntWines, MntF...

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

head(dataset\_ex, 3) #kitlemiz

## # A tibble: 3 x 29  
## ID Year\_Birth Education Marital\_Status Income Kidhome Teenhome Dt\_Customer  
## <dbl> <dbl> <chr> <chr> <dbl> <dbl> <dbl> <chr>   
## 1 5524 1957 Graduation Single 58138 0 0 04-09-2012   
## 2 2174 1954 Graduation Single 46344 1 1 08-03-2014   
## 3 4141 1965 Graduation Together 71613 0 0 21-08-2013   
## # ... with 21 more variables: Recency <dbl>, MntWines <dbl>, MntFruits <dbl>,  
## # MntMeatProducts <dbl>, MntFishProducts <dbl>, MntSweetProducts <dbl>,  
## # MntGoldProds <dbl>, NumDealsPurchases <dbl>, NumWebPurchases <dbl>,  
## # NumCatalogPurchases <dbl>, NumStorePurchases <dbl>,  
## # NumWebVisitsMonth <dbl>, AcceptedCmp3 <dbl>, AcceptedCmp4 <dbl>,  
## # AcceptedCmp5 <dbl>, AcceptedCmp1 <dbl>, AcceptedCmp2 <dbl>, Complain <dbl>,  
## # Z\_CostContact <dbl>, Z\_Revenue <dbl>, Response <dbl>

set.seed(300) #kodu her çalıştırdığımızda örneklemimizin değişmemesi için kullandığımız bir fonksiyon  
data\_sample <- dataset\_ex[sample(nrow(dataset\_ex),300),replace=F]  
head(data\_sample, 3)

## # A tibble: 3 x 29  
## ID Year\_Birth Education Marital\_Status Income Kidhome Teenhome Dt\_Customer  
## <dbl> <dbl> <chr> <chr> <dbl> <dbl> <dbl> <chr>   
## 1 738 1981 2n Cycle Single 42395 1 1 04-04-2014   
## 2 10846 1978 Master Married 43783 1 0 29-05-2014   
## 3 9240 1971 PhD Single 60504 0 1 07-02-2014   
## # ... with 21 more variables: Recency <dbl>, MntWines <dbl>, MntFruits <dbl>,  
## # MntMeatProducts <dbl>, MntFishProducts <dbl>, MntSweetProducts <dbl>,  
## # MntGoldProds <dbl>, NumDealsPurchases <dbl>, NumWebPurchases <dbl>,  
## # NumCatalogPurchases <dbl>, NumStorePurchases <dbl>,  
## # NumWebVisitsMonth <dbl>, AcceptedCmp3 <dbl>, AcceptedCmp4 <dbl>,  
## # AcceptedCmp5 <dbl>, AcceptedCmp1 <dbl>, AcceptedCmp2 <dbl>, Complain <dbl>,  
## # Z\_CostContact <dbl>, Z\_Revenue <dbl>, Response <dbl>

### Çocuk var/yok sütunu ekleme:

Herhangi bir kütüphane kullanmadan base paketteki **ìfelse()** fonksiyonu ile 0’dan fazla çocuğu olanları 1 olmayanları 0 olarak kodlayacak şekilde **Kid\_Status** değişkenine atadık:

data\_sample$Kid\_Status <- ifelse(data\_sample$Kidhome>0, 1, 0)

### 9.3 Faktörleri Tablolama ve Olasılık Tabloları Oluşturma:

Bir faktörü tablolamak veya birden fazla faktörden bir olasılık tablosu oluşturmak için kullanılan fonksiyon:

head(table(data\_sample$Income,data\_sample$Education))

##   
## 2n Cycle Basic Graduation Master PhD  
## 1730 0 0 1 0 0  
## 4023 0 0 0 0 1  
## 6560 0 0 0 1 0  
## 7500 0 1 0 0 0  
## 9722 0 1 0 0 0  
## 13533 1 0 0 0 0

### 9.4 Kategorik Değişkenlere Bağımsızlık Testi:

verilen değişkenlerin bağımsız olup olmadığına ilişkin kullanılan ki kare testi fonksiyonudur.  
H0: Veriler bağımsızdır.  
Hs: Veriler bağımsız değildir.

summary(data\_sample)

## ID Year\_Birth Education Marital\_Status   
## Min. : 1 Min. :1943 Length:300 Length:300   
## 1st Qu.: 2578 1st Qu.:1959 Class :character Class :character   
## Median : 5202 Median :1969 Mode :character Mode :character   
## Mean : 5346 Mean :1968   
## 3rd Qu.: 7969 3rd Qu.:1976   
## Max. :11071 Max. :1995   
##   
## Income Kidhome Teenhome Dt\_Customer   
## Min. : 1730 Min. :0.0000 Min. :0.0000 Length:300   
## 1st Qu.:37237 1st Qu.:0.0000 1st Qu.:0.0000 Class :character   
## Median :55478 Median :0.0000 Median :0.0000 Mode :character   
## Mean :53441 Mean :0.4267 Mean :0.4833   
## 3rd Qu.:70897 3rd Qu.:1.0000 3rd Qu.:1.0000   
## Max. :96547 Max. :2.0000 Max. :2.0000   
## NA's :6   
## Recency MntWines MntFruits MntMeatProducts  
## Min. : 0.00 Min. : 0.0 Min. : 0.00 Min. : 1.0   
## 1st Qu.:23.00 1st Qu.: 28.0 1st Qu.: 2.00 1st Qu.: 19.0   
## Median :50.00 Median : 217.5 Median : 10.00 Median : 98.5   
## Mean :49.18 Mean : 352.1 Mean : 25.23 Mean :188.0   
## 3rd Qu.:74.50 3rd Qu.: 612.5 3rd Qu.: 33.25 3rd Qu.:272.5   
## Max. :99.00 Max. :1492.0 Max. :193.00 Max. :951.0   
##   
## MntFishProducts MntSweetProducts MntGoldProds NumDealsPurchases  
## Min. : 0.00 Min. : 0.0 Min. : 0.00 Min. : 0.000   
## 1st Qu.: 3.00 1st Qu.: 2.0 1st Qu.: 10.00 1st Qu.: 1.000   
## Median : 15.50 Median : 10.0 Median : 29.00 Median : 2.000   
## Mean : 40.30 Mean : 28.7 Mean : 44.51 Mean : 2.353   
## 3rd Qu.: 46.75 3rd Qu.: 33.0 3rd Qu.: 56.25 3rd Qu.: 3.000   
## Max. :258.00 Max. :263.0 Max. :362.00 Max. :15.000   
##   
## NumWebPurchases NumCatalogPurchases NumStorePurchases NumWebVisitsMonth  
## Min. : 0.000 Min. : 0.000 Min. : 0.00 Min. : 1.000   
## 1st Qu.: 2.000 1st Qu.: 0.750 1st Qu.: 3.00 1st Qu.: 3.000   
## Median : 4.000 Median : 2.000 Median : 5.00 Median : 5.000   
## Mean : 4.343 Mean : 2.667 Mean : 6.21 Mean : 5.253   
## 3rd Qu.: 6.000 3rd Qu.: 5.000 3rd Qu.: 9.00 3rd Qu.: 7.000   
## Max. :27.000 Max. :11.000 Max. :13.00 Max. :20.000   
##   
## AcceptedCmp3 AcceptedCmp4 AcceptedCmp5 AcceptedCmp1   
## Min. :0.00 Min. :0.0000 Min. :0.00000 Min. :0.00   
## 1st Qu.:0.00 1st Qu.:0.0000 1st Qu.:0.00000 1st Qu.:0.00   
## Median :0.00 Median :0.0000 Median :0.00000 Median :0.00   
## Mean :0.06 Mean :0.1067 Mean :0.08333 Mean :0.11   
## 3rd Qu.:0.00 3rd Qu.:0.0000 3rd Qu.:0.00000 3rd Qu.:0.00   
## Max. :1.00 Max. :1.0000 Max. :1.00000 Max. :1.00   
##   
## AcceptedCmp2 Complain Z\_CostContact Z\_Revenue Response   
## Min. :0.00000 Min. :0.00 Min. :3 Min. :11 Min. :0.00   
## 1st Qu.:0.00000 1st Qu.:0.00 1st Qu.:3 1st Qu.:11 1st Qu.:0.00   
## Median :0.00000 Median :0.00 Median :3 Median :11 Median :0.00   
## Mean :0.02333 Mean :0.01 Mean :3 Mean :11 Mean :0.16   
## 3rd Qu.:0.00000 3rd Qu.:0.00 3rd Qu.:3 3rd Qu.:11 3rd Qu.:0.00   
## Max. :1.00000 Max. :1.00 Max. :3 Max. :11 Max. :1.00   
##   
## Kid\_Status   
## Min. :0.0   
## 1st Qu.:0.0   
## Median :0.0   
## Mean :0.4   
## 3rd Qu.:1.0   
## Max. :1.0   
##

summary(table(data\_sample$MntWines,data\_sample$MntFruits))

## Number of cases in table: 300   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 21265, df = 17928, p-value = 1.801e-62  
## Chi-squared approximation may be incorrect

p-value < alpha=0.05 için %95 güven düzeyiyle söylenebilir ki veriler bağımsız değildir.

### 9.5 Çeyreklikleri Hesaplama:

**`quantile()** fonksiyonunu değişkenimizin çeyrekliklerini bulmak için kullanıyoruz:

quantile(data\_sample$Income, na.rm=T)

## 0% 25% 50% 75% 100%   
## 1730.00 37237.25 55478.00 70897.00 96547.00

### 9.9 Ortalama için Güven Aralığı Hesaplama:

**n<30** olabilmesi için 30 örneklem çekip ortalamaya göre güven aralığı bulduk ki t testini uygulayabilelim.  
Bir diğer etmen olarak ise **“nuemric”** olmayan verileri çıkardık ki hesaplama süresince bize sıkıntı çıkarmasın.  
güven aralığı testi yaptık. Bunun sonucunda gelirin ortalamsına ilişkin güven aralığı %95 güven düzeyinde:  
[1502.769;3058.393] aralığındadır.

set.seed(30)  
data\_sample\_30 <- data\_sample[sample(nrow(data\_sample),30),]  
data\_sample\_30\_dbl <- select(data\_sample\_30, -c(Education, Marital\_Status, Dt\_Customer))  
data\_sample\_30\_dbl

## # A tibble: 30 x 27  
## ID Year\_Birth Income Kidhome Teenhome Recency MntWines MntFruits  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1212 1973 52845 1 0 7 384 25  
## 2 6182 1984 26646 1 0 26 11 4  
## 3 5585 1972 21359 1 0 1 12 2  
## 4 7698 1976 51650 0 1 81 152 3  
## 5 8629 1956 46984 1 1 71 19 1  
## 6 1945 1962 71434 0 1 4 711 36  
## 7 1772 1975 79174 0 0 2 1074 37  
## 8 7244 1951 NA 2 1 96 48 5  
## 9 1991 1967 44931 0 1 0 78 0  
## 10 1506 1975 42160 1 1 26 48 9  
## # ... with 20 more rows, and 19 more variables: MntMeatProducts <dbl>,  
## # MntFishProducts <dbl>, MntSweetProducts <dbl>, MntGoldProds <dbl>,  
## # NumDealsPurchases <dbl>, NumWebPurchases <dbl>, NumCatalogPurchases <dbl>,  
## # NumStorePurchases <dbl>, NumWebVisitsMonth <dbl>, AcceptedCmp3 <dbl>,  
## # AcceptedCmp4 <dbl>, AcceptedCmp5 <dbl>, AcceptedCmp1 <dbl>,  
## # AcceptedCmp2 <dbl>, Complain <dbl>, Z\_CostContact <dbl>, Z\_Revenue <dbl>,  
## # Response <dbl>, Kid\_Status <dbl>

#T testini yapabilmek için 36. satırdaki 3 değişkeni yokettik.  
t.test(data\_sample\_30\_dbl$Income)

##   
## One Sample t-test  
##   
## data: data\_sample\_30\_dbl$Income  
## t = 14.127, df = 27, p-value = 5.443e-14  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
## 47426.53 63543.90  
## sample estimates:  
## mean of x   
## 55485.21

### 9.11 Oran Testi:

Oran testini yapabilmek için önce Hipotezlerimizi kurduk:  
H0: Q = 50.000  
Hs: Q > 50.000 {geliri 50.000 den fazla olanları bulmak için kurduğumuz seçenek hipotezi}

prop.test(nrow(data\_sample[data\_sample$Income>50000,]),300,p=0.5,alternative="greater")

##   
## 1-sample proportions test with continuity correction  
##   
## data: nrow(data\_sample[data\_sample$Income > 50000, ]) out of 300, null probability 0.5  
## X-squared = 6.75, df = 1, p-value = 0.004687  
## alternative hypothesis: true p is greater than 0.5  
## 95 percent confidence interval:  
## 0.5275895 1.0000000  
## sample estimates:  
## p   
## 0.5766667

H0 reddedildi gelirin 50000 fazla olma olasılığı 0.5 ten fazladır.

### 9.16 Parametrik Olmayan İki Örneklem Konum Testleri:

Parametrik olamyan iki örneklem bulmak için önce normallik testi yaptık.  
Balık ve et satışlarını baz alarak yaptığımız bu testte iki değişkenin de normal dağılmadığı kanısına vardık.  
Test etmek üzere hipotezlerimizi kurduk: **H0: balık satışı ile et satışı arasında fark yoktur.**  
**Hs: balık satışı ile et satışı arasında fark vardır.**

sample\_meat <- as.data.frame(list(data\_sample$MntMeatProducts))  
colnames(sample\_meat) <- c("meat")  
sample\_fish <- as.data.frame(list(data\_sample$MntFishProducts))  
colnames(sample\_fish) <- c("fish")  
shapiro.test(sample\_meat$meat) #p<alpha=0.05 için normal dağılmıyor

##   
## Shapiro-Wilk normality test  
##   
## data: sample\_meat$meat  
## W = 0.78752, p-value < 2.2e-16

shapiro.test(sample\_fish$fish) #p<alpha=0.05 için normal dağılmıyor

##   
## Shapiro-Wilk normality test  
##   
## data: sample\_fish$fish  
## W = 0.71122, p-value < 2.2e-16

wilcox.test(sample\_meat$meat,sample\_fish$fish, alternative = "two.sided") #p<0.05 için H0 reddedilir.

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: sample\_meat$meat and sample\_fish$fish  
## W = 68491, p-value < 2.2e-16  
## alternative hypothesis: true location shift is not equal to 0

Test sonucuna baktığımızda ise;  
“%95 güven düzeyiyle söylenebilir ki et ve balık satışları arasında fark vardır.” yorumu yapılabilir.

### 9.17 Korelasyon Testi:

Aralarındaki ilişkiyi test etmek üzere **şarap (data\_sample$MntWines)** ve **gelir (data\_sample$Income)** değişkenlerinden örneklemler çektik.  
Daha sonra korelasyon testine tabii tuttuk.

sample\_wine <- as.data.frame(list(data\_sample$MntWines))  
colnames(sample\_wine) <- c("wine")  
  
sample\_income <- as.data.frame(list(data\_sample$Income))  
colnames(sample\_income) <- c("income")  
  
cor.test(sample\_income$income,sample\_wine$wine)

##   
## Pearson's product-moment correlation  
##   
## data: sample\_income$income and sample\_wine$wine  
## t = 19.987, df = 292, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.7071724 0.8045197  
## sample estimates:  
## cor   
## 0.7600781

Sonuç olarak **“Gelir düzeyi ile şarap alımı arasında %76 lık doğrusal bir ilişki vardır.”**

### 9.18 Gruplar Arası Oran Testi:

Testimize başlamadan önce hiptoezlerimizi kurduk:  
H0: Aylık web sitesi ziyaretinin 3’ten fazla olma olasılığı ile aylık dükkan ziyaretinin 5’ten fazla olma olasılığı eşittir.  
Hs: Pweb < Pstore

prop.test(nrow(data\_sample[data\_sample$NumWebPurchases>3,]),300,p=0.5)

##   
## 1-sample proportions test with continuity correction  
##   
## data: nrow(data\_sample[data\_sample$NumWebPurchases > 3, ]) out of 300, null probability 0.5  
## X-squared = 2.0833, df = 1, p-value = 0.1489  
## alternative hypothesis: true p is not equal to 0.5  
## 95 percent confidence interval:  
## 0.4851125 0.6004263  
## sample estimates:  
## p   
## 0.5433333

prop.test(nrow(data\_sample[data\_sample$NumStorePurchases>5,]),300,p=0.5)

##   
## 1-sample proportions test with continuity correction  
##   
## data: nrow(data\_sample[data\_sample$NumStorePurchases > 5, ]) out of 300, null probability 0.5  
## X-squared = 0.56333, df = 1, p-value = 0.4529  
## alternative hypothesis: true p is not equal to 0.5  
## 95 percent confidence interval:  
## 0.4191650 0.5347756  
## sample estimates:  
## p   
## 0.4766667

### 9.19 İki Grubun Ortalamalarına ilişkin Karşılaştırma Testi:

### 9.20 Dağılımlara ilişkin Kolmogrov Smirnov Testi:

Test etmek üzere hipotezlerimizi kurduk:  
H0: Dağılımlar arasında fark yoktur.  
Hs: Dağılımlar arasında fark vardır.  
Hipotezlerimizi kurduktan sonra Normal dağılıp dağılmadığına **shapiro.test()** fonksiyonu ile test ettik.  
Normal dağılmayan verilerimize devamında **ks.test()** fonksiyonu ile Kolmogrov testi uyguladık.

shapiro.test(data\_sample$MntFruits) #normal dağılmıyor

##   
## Shapiro-Wilk normality test  
##   
## data: data\_sample$MntFruits  
## W = 0.69365, p-value < 2.2e-16

shapiro.test(data\_sample$MntFishProducts) #normal dağılmıyor

##   
## Shapiro-Wilk normality test  
##   
## data: data\_sample$MntFishProducts  
## W = 0.71122, p-value < 2.2e-16

ks.test(data\_sample$MntFishProducts,data\_sample$MntFruits)

## Warning in ks.test(data\_sample$MntFishProducts, data\_sample$MntFruits): p-value  
## will be approximate in the presence of ties

##   
## Two-sample Kolmogorov-Smirnov test  
##   
## data: data\_sample$MntFishProducts and data\_sample$MntFruits  
## D = 0.11667, p-value = 0.0337  
## alternative hypothesis: two-sided

#p<0.05 için H0 reddedilir. dağılımlar farklıdır.

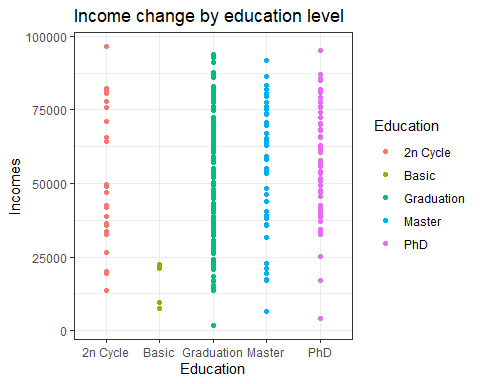
Sonuç oalrak ise:  
“%95 güven düzeyiyle dağılımlar arasında fark olduğu söylenebilir.” yorumu yapılabilir.

### 10.4 ggplot’a Tema Ekleme:

Veri görselleştirebilmek için **library(ggplot2)** kütüphanesini kullandık.  
Değişkenleri daha temiz işleyebilmek için **attach()** fonksiyonunu kullandık.  
Dağılım grafiğini çizdiğimiz verilerin isimlerini ise **labs()** fonksiyonu ile değiştirdik.  
Son olarak ise tema ekleme hususunda **theme\_dark()** temasını kullandık.

library(ggplot2)  
attach(data\_sample)  
ggplot(data\_sample, aes(Education, Income, color=Education)) +  
 geom\_point() +  
 theme\_bw() +   
 labs(title = "Income change by education level ",  
 x = "Education",  
 y = "Incomes")

## Warning: Removed 6 rows containing missing values (geom\_point).



detach(data\_sample)

##### Grafik yorumu:

### 10.11 :

ggplot(data\_sample, aes(data\_sample$Education,data\_sample$Income)) +   
 geom\_bar(stat="identity") +  
 geom\_errorbar(aes(ymin=min(data\_sample$Income,na.rm=T), ymax=max(data\_sample$Income,na.rm=T)), width=.2)

## Warning: Use of `data\_sample$Education` is discouraged. Use `Education` instead.

## Warning: Use of `data\_sample$Income` is discouraged. Use `Income` instead.  
  
## Warning: Use of `data\_sample$Income` is discouraged. Use `Income` instead.  
  
## Warning: Use of `data\_sample$Income` is discouraged. Use `Income` instead.

## Warning: Use of `data\_sample$Education` is discouraged. Use `Education` instead.

## Warning: Use of `data\_sample$Income` is discouraged. Use `Income` instead.

## Warning: Removed 6 rows containing missing values (position\_stack).

