Öğrenci adı: Merve Varlı
Öğrenci numarası: 139719023

Siyasal Bilgiler Fakültesi

İstatistik (IST2083)

R ile COVID-19-NY-SBU Verisetinde İstatistiki Araçların Kullanımı

Giriş

Bu doküman, İstatistik dersi çerçevesinde öğrenmiş olduğumuz fonksiyonların ve araçların belirli bir verisetindeki kullanımlarını görmek için proje ödevi olarak hazırlanmıştır. İlgili veri seti, Stony Brook Üniversitesi'nin COVID-19 çalışmalarında elde ettiği klinik verilere dayanmakta olup 1384 COVID-19 positif hastanın sağlık verilerini anonim şeklinde içermektedir.

Ben bu projemde verisetinin bize sağlamış olduğu 6 farklı sütunu (veri tipini) kullandım. Bunlar: hastanın son durumu, hastanın bulunduğu yaş aralığı, hastanın cinsiyeti, hastanın ilk gelmiş olduğu birim, hastanın hastanede yatış aldığı süre ve hastanın atardamarlarındaki oksijen yüzdesi. Bu veri tiplerinin olası değerleri Tablo 1'de verilmiştir.

Sütun Adı	Veri Tipi
last.status	Kategorik, string: ["discharged", "deceased"]
age.splits	Kategorik, string: ["[18, 59]", "(59, 74]", "(74, 90]"]
gender_concept_name	Kategorik: string: ["FEMALE", "MALE"]
visit_concept_name	Kategorik, string: ["Inpatient Visit", "Emergency Room Visit"]
length_of_stay	Numerik, int
Oxygen saturation in Arterial blood by Pulse oximetry	Numerik, int

Table 1: Kullanılan verisetindeki sutünlar ve tuttuğu değerleri.

Verisetindeki diğer sütunlar temizlenmiş ve yalnızca bu kısmın kullanıldığı modifiye dataset şu bağlantı aracılığı ile GitHub'da saklanmıştır. Orijinal veriseti ise Referans [2]'da görülebilir.

Dokümanın bundan sonraki bölümlerinde ilk olarak istatistiki araçlar kullanılarak veriseti hakkında genel yorumlar yapılacak ve daha sonra COVID-19 postif vakalara yönelik hipotezler sunulup R ile ilgili veriseti üzerinde test edileceklerdir.

Verisetinin R Studio'ya Geçirilmesi

Aşağıdaki komut aracılığı ile GitHub'da sakladığımız CSV dosyamızı okuyoruz ve R'ın anlayabileceği bir formatta *covid_19* değişkenine kaydediyoruz.

```
covid_19 <- read.csv("https://raw.githubusercontent.com/varlimerve/covid19_statistics/
    main/ClinicalDataOfPositivePatients_COVID-19-NY-SBU.csv")
```

Numerik Verilerdeki Dağılım

```
sd(covid_19$Oxygen.saturation.in.Arterial.blood.by.Pulse.oximetry)
sd(covid_19$length_of_stay)
```

Standart sapma komutu *sd()* ile hastaların hastane yatış süreleri verisini ve atardamarlarındaki oksijen yüzdesi verisinin dağılımını incelediğimizde aşağıdaki verileri buluyoruz.

- Atardamar kanlarındaki oksijen yüzdeliği verisinin standart sapması: 5.689748
- Hastane yatış süreleri verisinin standart sapması: 12.09952

Bu sonuçlar bizlere hastane yatış süreleri verilerinin daha dağınık olduğunu söylemekle birlikte, COVID-19 hastalarının kanlarındaki oksijen seviyelerinin çok büyük bir değişkenliğe (varyansa) sahip olmadığını gösteriyor.

Hastane Kalış Günlerinin En Çok Tekrar Eden Elemanı

İstatistikte mod olarak bilinen bu araç bize bir dizideki en çok tekrar eden elemanı verir. R'da yapmış olduğum denemelerde gördüğüm üzere *mod()* fonksiyonu istediğimiz istatistiki modu vermemektedir. Böylece, internet üzerinden daha önce yazılmış bir mod alma fonksiyonu kullanarak ilgili veri setinin numerik değerlerinin en çok tekrar edenlerini bulmuş oldum.

```
getmode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}
getmode(covid_19$covid_19$length_of_stay)
```

Görülmektedir ki, en çok karşılaşılan hasta kalış süresi ayakta olarak nitelendirilebilecek "1" gündür.

Atardamar Kanındaki Oksijen Yüzdesinin Ortanca Değeri

İlgili değer 95 olarak bulunarak (ve standart sapması dikkate alındığında) sağlıklı insanlarda olması gereken %90-%100 bandını bize tekrar teyit etmiştir.

```
    median(covid_19$Oxygen.saturation.in.Arterial.blood.by.Pulse.oximetry)
```

Hastanede Kalış Günlerinin İstatistiki Özeti

Aşağıdaki kod bloğunda çağırılmış olan fonksiyonu kullanarak örnek bir veri dizisine yönelik istatistiki bilgiler bir tablo halinde alınabilirdir.

```
    summary(covid_19$length_of_stay)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max
1.000	2.000	6.000	9.656	12.000	96.000

Table 2: Çağırılan *summary()* fonksiyonun çıktısı.

Ayrıca bu özet bilgi ile görülen üçüncü çeyrek ve ilk çeyreğin farkı olan 10.000, çeyrekler arası uzaklık olarak adlandırılır. Yalnızca bunun istenmesi durumunda aşağıdaki kod kullanılabilirdir.

```
    IQR(covid_19$length_of_stay)
```

Hipotezler

1. **Kadınların COVID-19 dolayısı ile hastanede kalış süreleri erkeklerden daha fazladır.**

   Bu hipotezin test edilebilmesi için öncelikle R üzerinde gerekli ayarlamalar yapılmalıdır. Bu ayarlamalar, yalnızca kadınların ve yalnızca erkeklerin bulunduğu alt-tablolar yaratmak ve bu tablolar üzerinde histogram verisi elde etmektir. Önemli bir nokta, bu alt-tabloların yaratılışı esnasında oluşabilecek bilinmeyen değerler (N/A) silinmelidir. Aşağıdaki kod ile yaratılmış test düzeni bu hipotezin testi için kullanılmıştır.

```r
# Get a sub-table for only female patients.
dist_female <- covid_19[(covid_19$gender_concept_name=="FEMALE"), ]
# Draw the histogram of the data.
hist(dist_female$length_of_stay,
     main = "Female Length of Stay Distribution",
     xlab = "Length of Stay (day)",
    )

# Remove NA from dataset to calculate mean for female.
dist_female_wo_na <- na.omit(dist_female)
mean(dist_female_wo_na$length_of_stay)

# Get a sub-table for only male patients.
dist_male <- covid_19[(covid_19$gender_concept_name=="MALE"),]
# Draw the histogram of the data.
hist(dist_male$length_of_stay,
     main = "Male Length of Stay Distribution",
     xlab = "Length of Stay (day)",
     )

# Remove NA from dataset to calculate mean for male.
dist_male_ma_na <- na.omit(dist_male)
mean(dist_male_ma_na$length_of_stay)
```

   Bulgular aşağıdaki gibidir.

   - Kadınların ortalama kalış süresi 7.831919 gün.
   - Erkeklerin ortalama kalış süresi 10.65303 gün.

   İlgili histogram görüntüleri Figür 1'de ve Figür 2'de gösterilmiştir.

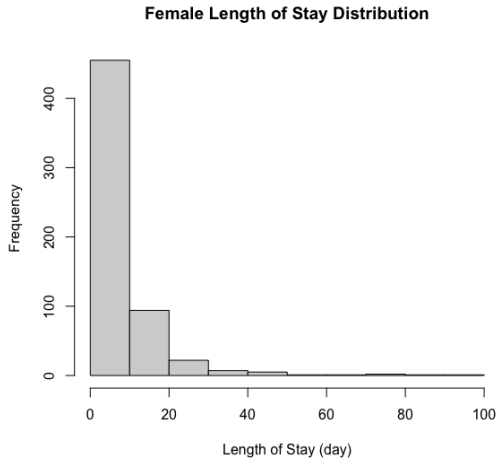

Figure 1: COVID-19 pozitif kadın hastaların hastande kalış süreleri.

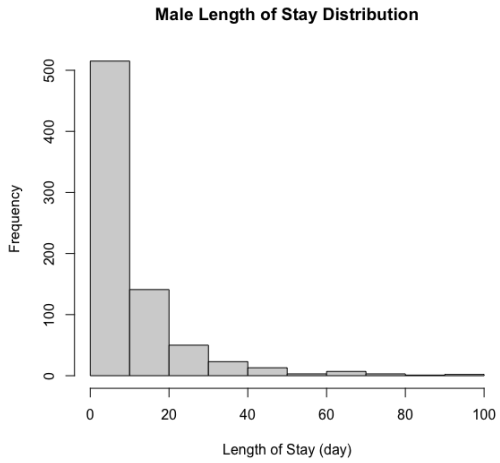

Figure 2: COVID-19 pozitif erkek hastaların hastande kalış süreleri.

Görülmektedir ki, **hipotez yanlıştır** ve erkek hastaların kalış süresi kadınlarınkinden daha fazladır.

2. **COVID-19 pozitif hastalarda ölüm ihtimalinin fazla olduğu kişilerin kanlarındaki oksijen seviye daha düşüktür.**

   Bu hipotezin test edilebilmesi için öncelikle R üzerinde gerekli ayarlamalar yapılmalıdır. Bu ayarlamalar, yalnızca taburcu olmuşların ve yalnızca ölmüşlerin bulunduğu alt-tablolar yaratmak ve bu tablolar üzerinde histogram verisi elde etmektir. Aşağıdaki kod ile yaratılmış test düzeni bu hipotezin testi için kullanılmıştır.

```
# Create the sub-tables.
table_last_status_dead <- covid_19[(covid_19$last.status== "deceased"),]
table_last_status_live <- covid_19[(covid_19$last.status== "discharged"),]

# Get the summary statistical informations.
summary(table_last_status_dead$Oxygen.saturation.in.Arterial.blood.by.Pulse.
    oximetry)
summary(table_last_status_live$Oxygen.saturation.in.Arterial.blood.by.Pulse.
    oximetry)
```

Bulgular Tablo 3 ve Tablo 4 üzerinde verilmiştir.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max
64	93	95	94.46	98	100

Table 3: Taburcu Edilmiş Hastaların Atardamarlarındaki Oksijen Yoğunluğu

Min.	1st Qu.	Median	Mean	3rd Qu.	Max
55	85	92.000	89.3	95.5	100

Table 4: Ölmüş Hastaların Atardamarlarındaki Oksijen Yoğunluğu

Görülmektedir ki, **hipotez doğrudur**. Daha sonradan öldüğünü bildiğimiz bireylerin kanlarındaki oksijen seviyeleri %89.3'tür. Buna karşılık taburcu edilmiş bireylerin ise %94.46 olduğu görülmüştür. Böylece atardamarlardaki oksijen seviyesinde yaşanılan %5.14'lük düşüşün kişilerin ölümüne yol

açtığı çıkarımı yapılabilir. Diğer taraftan daha sonra öldüğünüz bildiğimiz bireylerin kanındaki minimum oksijen değeri %55 iken taburcu edilen bireylerde bu değer %64'tür. Böylece atardamarlardaki oksijen seviyesindin minimum değerinde yaşanılan %9'luk düşüş de çıkarsamamıza destek olmaktadır.

3. **COVID-19 pozitif hastalarda yaş aralığı arttıkça ölüm oranı da artmaktadır.**

   Bu hipotezin test edilebilmesi için öncelikle R üzerinde bir bar-plot oluşturulmuş ve gözlemsel olarak oranlarına bakılmıştır. Daha sonra ise bu bar-plotta gözlemsel olarak çıkarılan sonuç matematiksel olarak desteklenmiştir.

```r
# Install the package to use bar-plots.
install.packages("ggplot2")
library(ggplot2)

# Create a barplot for deceased and discharged in range of age.
barplot(table(covid_19$last.status))
table(covid_19$age.splits)
barplot(table(covid_19$age.splits))
table_1 <- table(covid_19$last.status, covid_19$age.splits)
addmargins(table_1)
prop.table(table_1, margin = 2)
barplot(table_1, legend.text = TRUE)
```

İlgili bar grafiği Figur 3'de görülmektedir. Bu figürden görülebileceği üzere siyah alanların beyaz alanlara (yani, ölümlerin taburculara) oranının en yüksek olduğu durum (74, 90] yaş aralığına aittir. Yaş aralığının artması, ölüm oranlarını da arttırmıştır.

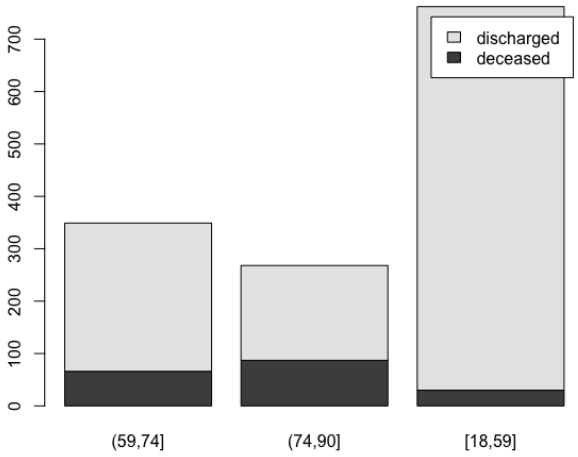

Figure 3: COVID-19 pozitif vakalarda ölüm ve taburcu oranlarının bar grafiğine yansıması.

İlgili gözlemimizi matematiksel olarak test edelim.

```r
table_2 <- table(covid_19$last.status, covid_19$age.splits)
table_2[1,]/table_2[2,]
```

Görülmektedir ki, **hipotez doğrudur** ve hastaların yaşı arttıkça ölümlerin taburcuya olan oranı yükselmektedir. Yaşın artışı ölüm ihtimalini de arttırmaktadır denilebilir.

18...59	60...74	75...90
0.04198361	0.23321555	0.48066298

Table 5: Ölmüş Hastaların Taburcu Olmuş Hastalara Bölümünun Yaşlara Göre Dağılımı

4. **COVID-19 pozitif vakalarda hastanelerin acil bölümüne başvurmada erkeklerin oranı kadınlara göre daha fazladır.**

   Bu hipotezin test edilebilmesi için öncelikle R üzerinde bir bar-plot oluşturulmuş ve gözlemsel olarak oranlarına bakılmıştır. Daha sonra ise bu bar-plotta gözlemsel olarak çıkarılan sonuç matematiksel olarak desteklenmiştir.

```r
# Barplot for visitng room for both female and male.
barplot(table(covid_19$gender_concept_name))
barplot(table(covid_19$visit_concept_name))
table_2 <- table(covid_19$gender_concept_name, covid_19$visit_concept_name)
addmargins(table_2)
prop.table(table_2, margin = 2)
barplot(table_2, legend.text = TRUE)
```

   İlgili bar grafiği Figur 4'de görülmektedir. Bu figürden görülebileceği üzere erkek ve kadın sayılarının hastaneye geliş tipleri ile (acil veya ayakta/bekleyen) bir ilgisi olmadığı açıktır. Bundan dolayı, COVID-19 vakalarda hastanelerin acil bölümüne başvurulma oranında kadın ve erkekler neredeyse eşit olduğu görülmüş olup, **hipotez yanlışlanmıştır**.

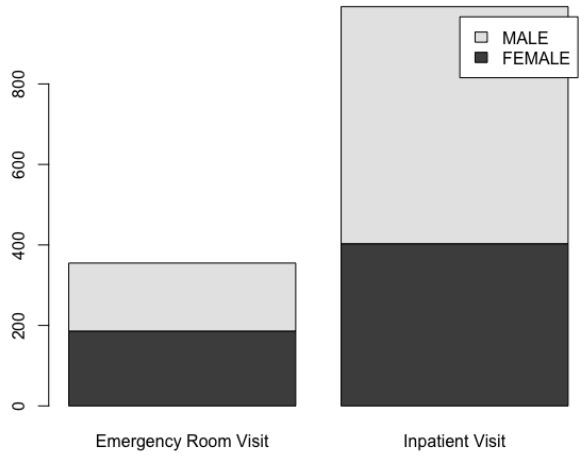

Figure 4: COVID-19 pozitif vakalarda acil veya ayakta girişlerin cinsiyete göre dağılımı.

5. **COVID-19 pozitif hastaların atardamarlarındaki oksijen miktarının azalışı hastanede yattıkları sürenin artışına yol açar.**

   İki veri arasındaki bir bağlantını ortaya çıkabilmesi için bu değerler bir scatter grafiği ile çizilir ve mantıksal bir fonksiyona indirgenmeye çalışılır. Doğru orantı, $y = ax + b$ fonksiyonu ile, ters orantı ise aynı fonksiyonun a değişkeninin negatif oluşu ile (ki bu, eğimin ters yönde olduğu anlamına gelir.) belirtilir. Ancak bilinmelidir ki, iki değer arasındaki ilişki birinci dereceden bir fonksiyon ile

belirtilmek zorunda değildir. Örneğin, ikinci dereceden veya üçüncü dereceden bir fonksiyon da bir ilişkiyi açıklayabilir.

Ben bu hipotezi test etmek için iki veri tipini bir scatter plot'a koydum ve bir fonksiyon bağlantısı aramaya çalıştım. Sizlerin de Figure 5'den göreceğiniz üzere bu iki değer arasında bir bağlantı bulunmamaktadır. Öncelikle, ilgili grafiğimizi R yardımı ile yaratalım.

```
plot(x=covid_19$length_of_stay,
     y=covid_19$Oxygen.saturation.in.Arterial.blood.by.Pulse.oximetry,
 xlab = "Length of Stay",
     ylab = "Oxygen Saturation")
```

Bu kodun bize yaratacağı çıktı olan scatter grafiği Figür 5'da görülmektedir.

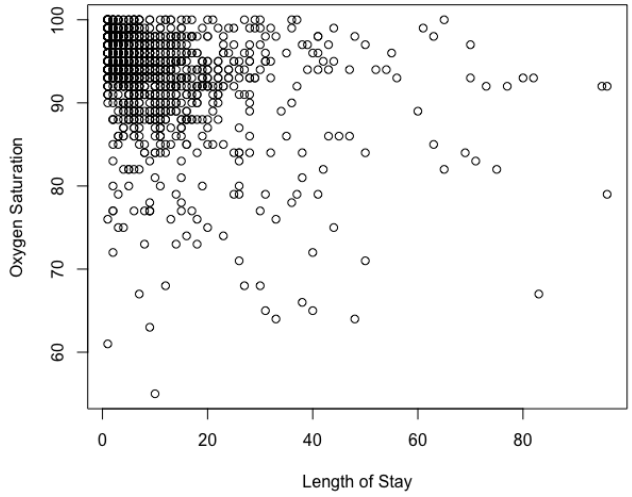

Figure 5: Oksijen yoğunluğun ve hastanade kalış süresi arasındaki grafik.

Böylece görmekteyiz ki, **hipotez yanlıştır**. COVID-19 pozitif hastaların oksijen seviyeleri azaldıkça hastanede yatış sürelerinin artıp azalmadığı, çünkü doğrudan bir ilişkinin bulunmadı görülmüştür.

Referanslar

1. Hakan Mehmetçik, R ile İstatistik Ders Notları, 2022

2. Saltz, J., Saltz, M., Prasanna, P., Moffitt, R., Hajagos, J., Bremer, E., Balsamo, J., & Kurc, T. (2021). Stony Brook University COVID-19 Positive Cases [Data set]. The Cancer Imaging Archive. https://doi.org/10.7937/TCIA.BBAG-2923