4_hafta_temel_istatistik

Hakan Mehmetcik

2025-10-22

Temel İstatistik Kavramlarına Giriş

Tanımlayıcı/Betimsel İstatistik

İstatistiğin Amacı: İstatistik, verilerden sorularımızı yanıtlamamıza yardımcı olmak için vardır. Yani, verileri analiz ederek, onları anlamaya çalışırız. Fakat ilk olarak verileri anlamamız ve özetlememiz gereklidir.

Verileri özetlemenin bir yolu, **tanımlayıcı/betimsel istatistikler** kullanmaktır. Bu istatistikler, veri setindeki genel durumu anlamamıza yardımcı olur. Ancak, tanımlayıcı istatistiklerin sonuçlarına dayanarak kesin kararlar vermememiz gerekir. Bunun yerine, tanımlayıcı istatistikler bize veri setindeki değişkenler arasındaki ilginç ilişkileri keşfetme fırsatı sunar.

> **Not**
>
> Veri analizi bağlamında, **tanımlayıcı/betimesel istatistik** ve **yorumlayıcı/çıkarımsal istatistik** iki önemli kavramdır. Bu iki istatistik türü, verileri farklı şekillerde kullanmamıza olanak tanır.
> **Tanımlayıcı/betimsel İstatistik** Tanımlayıcı istatistikler, bir veri setinin genel özelliklerini özetlemeye yönelik yöntemlerdir. Bu istatistikler, ortalama, medyan, mod, varyans gibi ölçümleri içerir ve verilerin dağılımı hakkında bilgi verir. Ancak, tanımlayıcı istatistiklerden kesin kararlar vermek mümkün değildir; bunlar yalnızca veri setindeki genel durumu anlamamıza yardımcı olur.
> **Yorumlayıcı/çıkarımsal İstatistik** Yorumlayıcı istatistik ise verileri analiz etmekle ilgilidir; yani verileri özetlemek yerine, örnek verilerden tüm popülasyon hakkında çıkarımlar yapmayı amaçlar. Yani, yorumlayıcı istatistik, örnek verileri kullanarak popülasyon hakkında sonuçlar çıkarmak veya çıkarımlarda bulunmak için kullanılır. Bu bağlamda, yorumlayıcı istatistik, popülasyon hakkında ne kadar güvenilir sonuçlar çıkarabileceğimizi belirlemek için korelasyonlar, olasılık, regresyon gibi çeşitli istatistiksel

yöntemler kullanır.
Özetle:

- **Tanımlayıcı İstatistik**: Veri setinin genel özelliklerini özetler, kesin kararlar vermez.
- **Yorumlayıcı İstatistik**: Verileri analiz eder ve örnek verilerden popülasyon hakkında çıkarımlar yapar.

Kategorik Değişkenlerle Kullanılabilen Tanımlayıcı İstatistikler

1. **Frekans (Frequency)**:
   - Kategorik değişkenlerin her bir seviyesinin kaç kez tekrarlandığını gösterir. Örneğin, bir anket sonucunda "evet" ve "hayır" cevaplarının sayısı.
2. **Yüzde (Percentage)**:
   - Her bir kategorinin toplam içindeki oranını gösterir. Frekansların toplam gözlem sayısına bölünmesiyle elde edilir. Örneğin, "evet" cevabının yüzdesi, "evet" sayısının toplam cevap sayısına bölünmesi ile hesaplanır.

$$p = \frac{\text{kategori içindeki birey sayısı}}{\text{örneklem büyüklüğü}}$$

3. **Mod (Mode)**:
   - En sık rastlanan kategori veya değer. Kategorik veriler için en yaygın olan seviyeyi belirtir. Örneğin, bir sınıftaki en çok tercih edilen renk.
4. **Çapraz Tablo (Contingency Table)**:
   - İki veya daha fazla kategorik değişken arasındaki ilişkiyi gösterir. Her bir kategorinin kesişimindeki frekansları içerir. Örneğin, cinsiyet ve hayatta kalma durumu arasındaki ilişkiyi gösteren bir tablo.

Örnek 1: Kitap okuma Oranları ve Bilim Haberlerine İlgi

```
# Kitap okuma kategorileri ve sayıları
books_readers <- c("no_books"=395, "print_only"=577, "digital_only"=91, "print_and_digital

books_readers  # Kitap okuyanların sayısını görüntüle
```

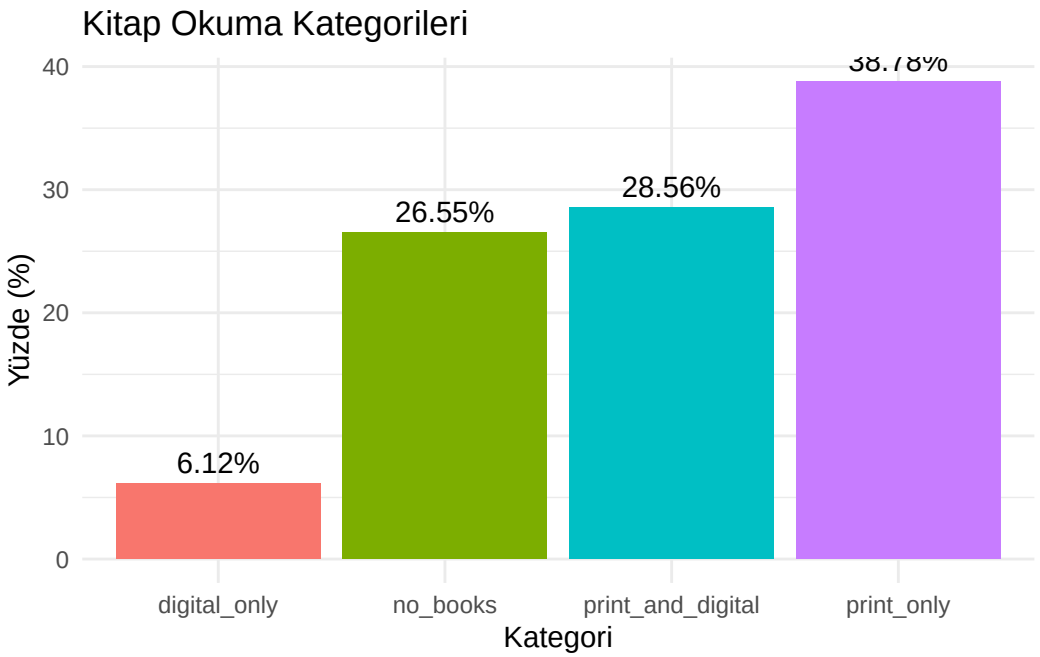
```
       no_books        print_only      digital_only print_and_digital
            395               577                91               425
```

Frekans dışında yüzde olarak da ifade edilebilir:

```r
# Kitap okuma kategorilerinin oranlarını yüzde olarak hesapla ve görüntüle
books_readers_percent <- round(prop.table(books_readers) * 100, 2)
books_readers_percent
```

```
       no_books        print_only      digital_only print_and_digital
          26.55             38.78              6.12             28.56
```

```r
# Veri çerçevesine dönüştür
books_df <- as.data.frame(books_readers) %>%
  rownames_to_column(var = "category") %>%
  rename(count = books_readers) %>%
  mutate(percent = round((count / sum(count)) * 100, 2))

# Bar grafiği oluştur
ggplot(books_df, aes(x = category, y = percent, fill = category)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = paste0(percent, "%")), vjust = -0.5) +
  labs(title = "Kitap Okuma Kategorileri",
       x = "Kategori",
       y = "Yüzde (%)") +
  theme_minimal() +
  theme(legend.position = "none")
```

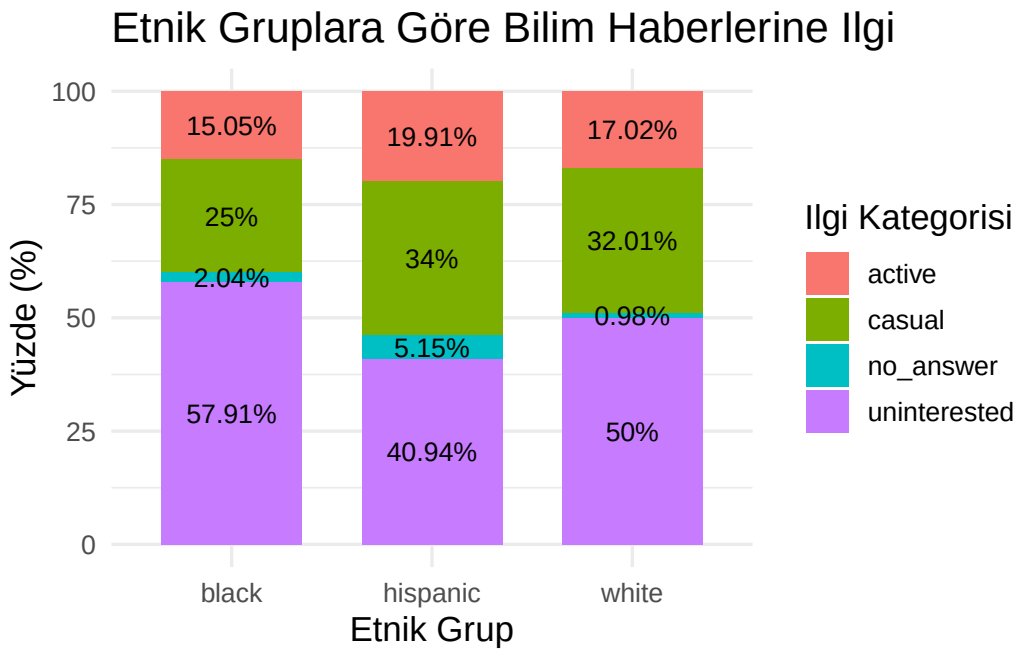


Örnek 2: Etnik Gruplar ve Bilim Haberlerine İlgi:

```r
library(dplyr)
library(tidyr)
library(ggplot2)

# Veriler
white <- c("active"=487, "casual"=916, "uninterested"=1431, "no_answer"=28)
black <- c("active"=59, "casual"=98, "uninterested"=227, "no_answer"=8)
hispanic <- c("active"=89, "casual"=152, "uninterested"=183, "no_answer"=23)

# 1. Veri çerçevesi oluşturma
my_table <- as.data.frame(rbind(white, black, hispanic))

# 2. Oranları hesapla (satır bazında)
prop_table <- round(prop.table(as.matrix(my_table), margin = 1) * 100, 2)
prop_table_df <- as.data.frame(prop_table)

# 3. Oranları tablo olarak görüntüle
prop_table_df
```

	active	casual	uninterested	no_answer
white	17.02	32.01	50.00	0.98
black	15.05	25.00	57.91	2.04
hispanic	19.91	34.00	40.94	5.15

```r
# Uzun formata dönüştür
prop_long <- prop_table_df %>%
  rownames_to_column(var = "group") %>%
  pivot_longer(cols = -group, names_to = "category", values_to = "percent")

# Bar Plot oluştur
ggplot(prop_long, aes(x = group, y = percent, fill = category)) +
  geom_col(width = 0.7) +
  geom_text(aes(label = paste0(percent, "%")),
            position = position_stack(vjust = 0.5),
            size = 3.5) +
  labs(
    title = "Etnik Gruplara Göre Bilim Haberlerine İlgi",
    x = "Etnik Grup",
    y = "Yüzde (%)",
    fill = "İlgi Kategorisi"
  ) +
  theme_minimal(base_size = 13)
```

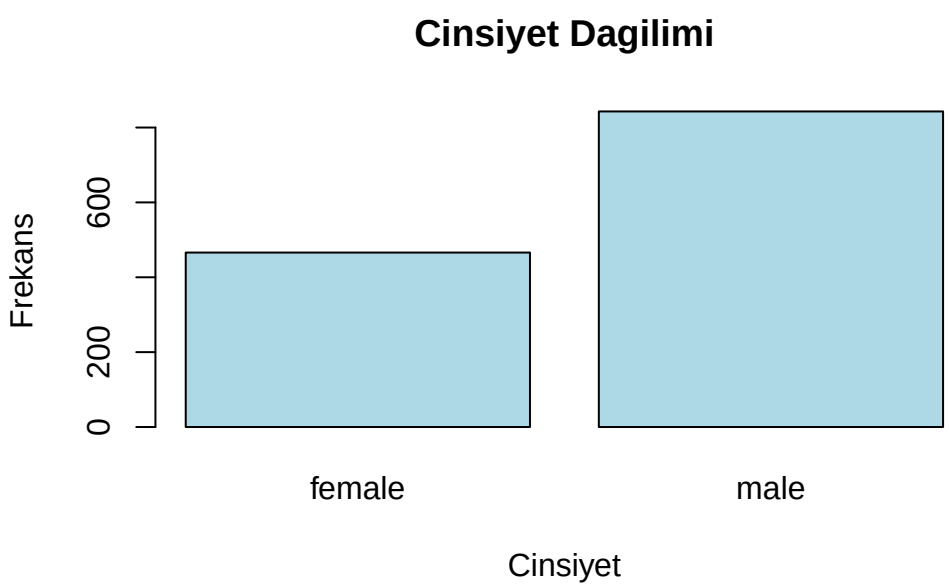


Örnek 3: Titanik'te Hayatta Kalma Oranları

```
# Titanic verisini okuma
titanic <- read.csv("https://raw.githubusercontent.com/bio304-class/bio304-course-notes/ma

# Gerekli kütüphaneleri yükleyin
library(tidyverse)

# Cinsiyet dağılımı tablosu
table(titanic$sex)
```

```
female   male
   466    843
```

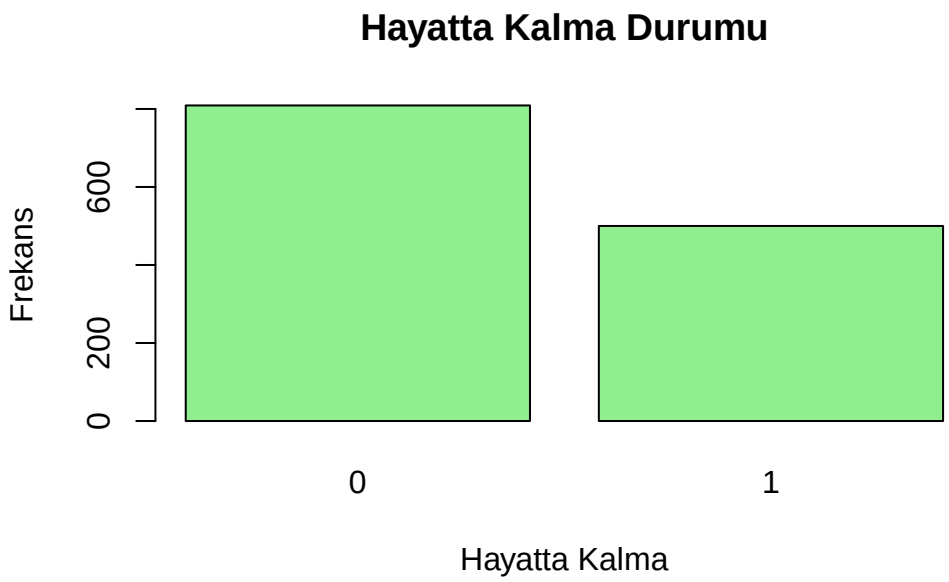
```
# Cinsiyet dağılımını gösteren çubuk grafiği
barplot(table(titanic$sex), main = "Cinsiyet Dağılımı", xlab = "Cinsiyet", ylab = "Frekans
```




```
# Hayatta kalma durumu tablosu
table(titanic$survived)
```

```
  0   1
809 500
```

```
# Hayatta kalma durumunu gösteren çubuk grafiği
barplot(table(titanic$survived), main = "Hayatta Kalma Durumu", xlab = "Hayatta Kalma", yl
```



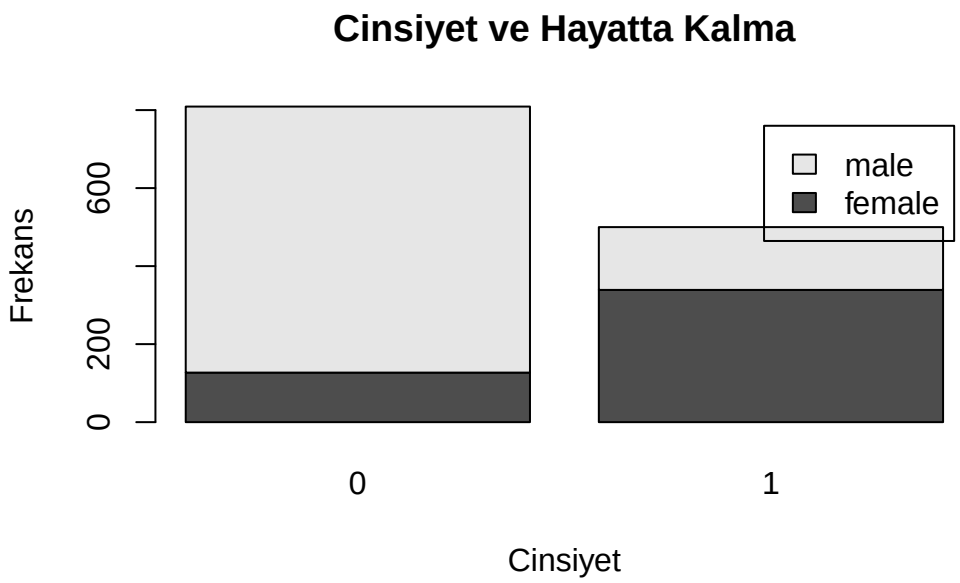

```r
# Cinsiyet ve hayatta kalma durumu arasındaki tablo
table_1 <- table(titanic$sex, titanic$survived)

# Toplamları ekleyerek tabloyu gösterme
addmargins(table_1)
```

```
          0    1  Sum
female  127  339  466
male    682  161  843
Sum     809  500 1309
```

```r
# Oran tablosunu gösterme
prop.table(table_1, margin = 2)
```

```
                0         1
female 0.1569839 0.6780000
male   0.8430161 0.3220000
```

```r
# Cinsiyet ve hayatta kalma durumu için çubuk grafiği
barplot(table_1, legend.text = TRUE, main = "Cinsiyet ve Hayatta Kalma", xlab = "Cinsiyet"
```



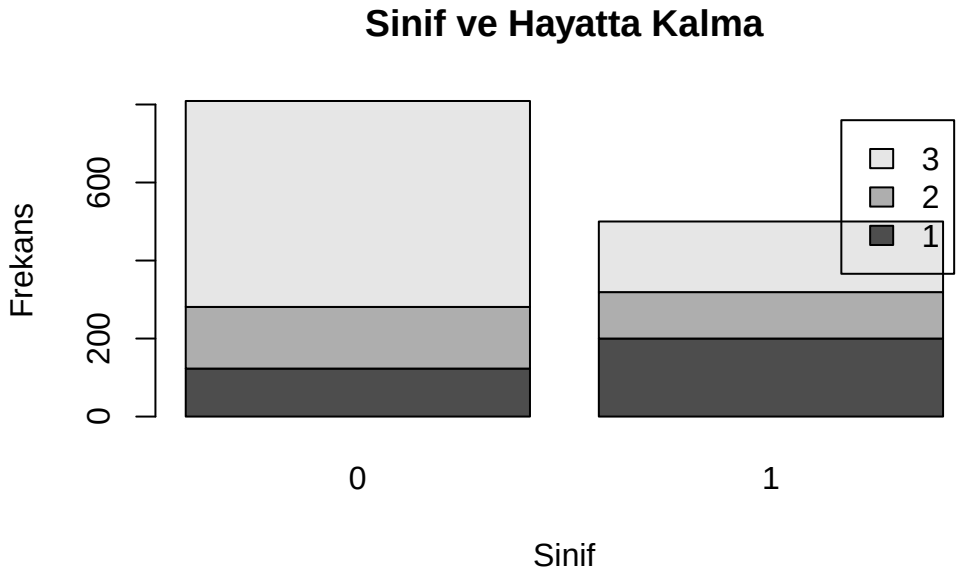

```r
# Sınıf ve hayatta kalma durumu arasındaki tablo
table_2 <- table(titanic$pclass, titanic$survived)

# Toplamları ekleyerek tabloyu gösterme
addmargins(table_2)
```

```
      0    1  Sum
1   123  200  323
2   158  119  277
3   528  181  709
Sum 809  500 1309
```

```r
# Oran tablosunu gösterme
prop.table(table_2)
```

```
           0          1
1 0.09396486 0.15278839
2 0.12070283 0.09090909
3 0.40336134 0.13827349
```

```r
# Sınıf ve hayatta kalma durumu için çubuk grafiği
barplot(table_2, legend.text = TRUE, main = "Sınıf ve Hayatta Kalma", xlab = "Sınıf", ylab
```




```r
# Yaş ve hayatta kalma durumu arasındaki tablo
table_3 <- table(titanic$age, titanic$survived)

# Toplamları ekleyerek tabloyu gösterme
addmargins(table_3)
```

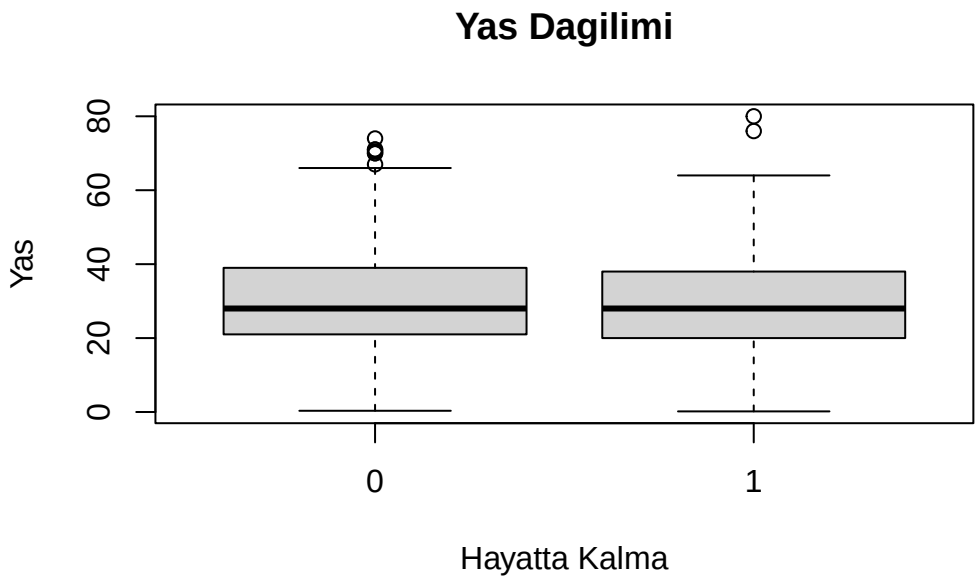
```
          0    1  Sum
0.1667    0    1    1
0.3333    1    0    1
0.4167    0    1    1
0.6667    0    1    1
0.75      1    2    3
```

0.8333	0	3	3
0.9167	0	2	2
1	3	7	10
2	8	4	12
3	2	5	7
4	3	7	10
5	1	4	5
6	3	3	6
7	2	2	4
8	2	4	6
9	6	4	10
10	4	0	4
11	3	1	4
11.5	1	0	1
12	0	3	3
13	2	3	5
14	4	4	8
14.5	2	0	2
15	1	5	6
16	11	8	19
17	13	7	20
18	25	14	39
18.5	3	0	3
19	18	11	29
20	15	8	23
20.5	1	0	1
21	30	11	41
22	23	20	43
22.5	1	0	1
23	16	10	26
23.5	1	0	1
24	25	22	47
24.5	1	0	1
25	23	11	34
26	19	11	30
26.5	1	0	1
27	17	13	30
28	24	8	32
28.5	3	0	3
29	17	13	30
30	25	15	40
30.5	2	0	2
31	11	12	23

32	13	11	24
32.5	3	1	4
33	12	9	21
34	10	6	16
34.5	2	0	2
35	10	13	23
36	17	14	31
36.5	1	1	2
37	7	2	9
38	8	6	14
38.5	1	0	1
39	12	8	20
40	12	6	18
40.5	3	0	3
41	9	2	11
42	12	6	18
43	6	3	9
44	7	3	10
45	7	14	21
45.5	2	0	2
46	6	0	6
47	11	3	14
48	4	10	14
49	4	5	9
50	9	6	15
51	5	3	8
52	3	3	6
53	0	4	4
54	5	5	10
55	4	4	8
55.5	1	0	1
56	2	2	4
57	5	0	5
58	2	4	6
59	2	1	3
60	3	4	7
60.5	1	0	1
61	5	0	5
62	3	2	5
63	2	2	4
64	3	2	5
65	3	0	3
66	1	0	1

```
67         1    0    1
70         2    0    2
70.5       1    0    1
71         2    0    2
74         1    0    1
76         0    1    1
80         0    1    1
Sum      619  427 1046
```

```r
# Yaş dağılımını gösteren kutu grafiği
boxplot(titanic$age ~ titanic$survived, main = "Yaş Dağılımı", xlab = "Hayatta Kalma", yla
```



Yorumlar:

- **Grup 1**: 123 kişi ölmüş, 200 kişi hayatta kalmış. Toplamda 323 kişi.
- **Grup 2**: 158 kişi ölmüş, 119 kişi hayatta kalmış. Toplamda 277 kişi.
- **Grup 3**: 528 kişi ölmüş, 181 kişi hayatta kalmış. Toplamda 709 kişi.
- **Toplamlar**:
  - Tüm gruplarda toplam 1309 kişi gözlemlenmiştir.

  - Ölenlerin toplamı 809, hayatta kalanların toplamı ise 500'dür.

- **Hayatta Kalma Oranı**:
  - En yüksek hayatta kalma sayısına sahip grup 1'dir (200 kişi hayatta), en yüksek ölüm sayısına sahip grup ise 3'tür (528 kişi ölmüş).

- **Hayatta Kalma Oranları**: Kadınların hayatta kalma oranı (67.8%) erkeklerin hayatta kalma oranından (32.2%) oldukça yüksektir. Bu, kadınların Titanic faciasında erkeklere göre daha yüksek bir hayatta kalma oranına sahip olduğunu göstermektedir.
- **Ölüm Oranları**: Erkekler için ölüm oranı (84.3%) oldukça yüksekken, kadınlar için bu oran çok daha düşüktür (15.7%). Bu durum, kadınların daha iyi korunmuş olabileceğini veya bazı sosyal faktörlerin etkisiyle hayatta kalma şanslarının artmış olabileceğini düşündürmektedir.

Sayısal Değişkenlerle Kullanılabilen Tanımlayıcı İstatistikler

Tanımlayıcı istatistiklerde veriyi tanımlamak için sıkça kullanılan istatistiksel ölçümler şunlardır:

1. **Merkezi Eğilim Ölçüleri**: Ortalama, Medyan, Mod
2. **Merkezi Dağılım Ölçüleri**: Aralık, standart sapma, varyans
3. **Eğrilik ve Basıklık**: Dağılım grafiklerinin normal dağılımdan farklılaşması
4. **Korelasyon**: İki değişken arasındaki ilişkinin yönü ve gücü.

1. Merkezi Eğilim Ölçüleri:

1.1 Ortalama (Mean):

- **Tanım**: Verilerin aritmetik ortalaması, tüm değerlerin toplamının gözlem sayısına bölünmesi ile hesaplanır.
- **Matematiksel Gösterim**:

$$\text{Ortalama} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- **R Formülü**:

```r
veri <- c(34, 67, 23, 45, 89, 12, 56, 78, 99, 5, 62, 48, 39, 75, 80, 22, 90, 11, 36,

mean(veri)  # veri, ortalamasını almak istediğiniz sayısal vektördür.
```

```
[1] 51.05
```

1.2 Medyan (Median):

- **Tanım**: Verilerin sıralandıktan sonra ortadaki değeri. Özellikle aşırı değerlerin etkisini azaltır.
- **Matematiksel Gösterim**:
  - **Eğer** n tek ise:

$$\text{Medyan} = x_{(\frac{n+1}{2})}$$

  - **Eğer** n çift ise:

$$\text{Medyan} = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}$$

```r
median(veri)  # veri, medyanını almak istediğiniz sayısal vektördür.
```

```
[1] 49
```

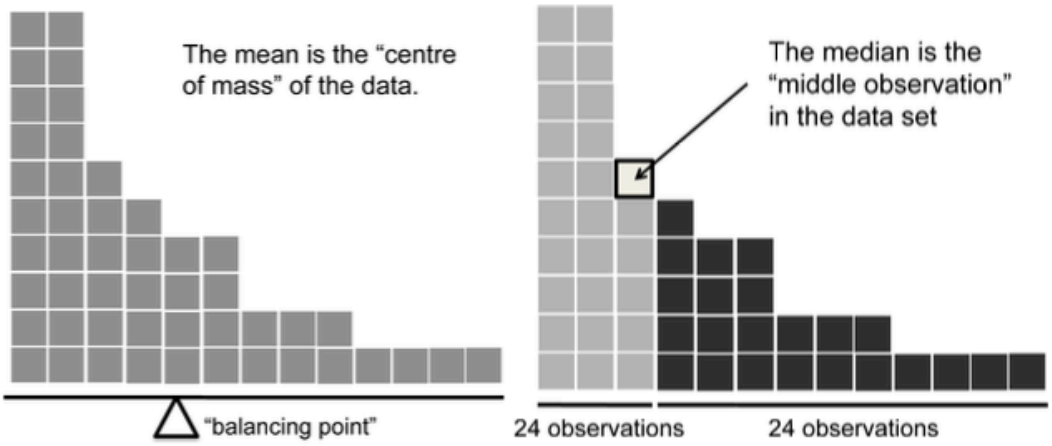
> **Not**
>
> Ortalama, veri setinin "ağırlık merkezi"dir: verilerin histogramını katı bir cisim olarak hayal ederseniz, onu dengeleyebileceğiniz nokta (bir tahterevalli gibi) ortalamadır. Buna karşılık, medyan ortadaki gözlemdir. Gözlemlerin yarısı daha küçüktür ve yarısı daha büyüktür.

1.3 Mod (Mode):

- **Tanım**: En sık rastlanan sayısal değer. Sayısal veriler için de kullanılabilir, ancak genellikle kategorik verilerle daha yaygındır.
- **R Formülü:**

```r
mode_function <- function(x) {
  uniq_x <- unique(x)
  uniq_x[which.max(tabulate(match(x, uniq_x)))]
}
mode_function(veri)  # veri, modunu almak istediğiniz sayısal vektördür.
```

```
[1] 34
```



2. Merkezi Dağılım Ölçüleri

2.1 Aralık (Range):

- **Matematiksel Gösterim**:

$$\text{Aralık} = \max(x) - \min(x)$$

- **R Formülü:**

```
range(veri)  # veri, aralığını almak istediğiniz sayısal vektördür.
```

```
[1]  5 99
```

```
max(veri) - min(veri)  # Aralığın hesaplanması
```

```
[1] 94
```

2.2 Standart Sapma (Standard Deviation):

- **Tanım:** Verilerin ortalamadan ne kadar yayıldığını gösterir. Verilerin ne kadar değişken olduğunu ölçer.