4_hafta_temel_istatistik

Hakan Mehmetcik

2024-10-22

Temel İstatistik Kavramlarına Giriş

Tanımlayıcı İstatistik

Verileri özetlemenin bir yolu, **tanımlayıcı 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.

İstatistiğin Amacı: İstatistik, verilerden sorularımızı yanıtlamamıza yardımcı olmak için vardır. Yani, verileri analiz ederek, onları anlamaya çalışırız.

> **Note**
>
> Veri analizi bağlamında, **tanımlayıcı istatistik** ve **yorumlayıcı istatistik** iki önemli kavramdır. Bu iki istatistik türü, verileri farklı şekillerde kullanmamıza olanak tanır.
> **Tanımlayıcı 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 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"=4
books_readers  # Kitap okuyanların sayısını görüntüle
```

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

```
# Sadece basılı kitap okuyanların örnek oranını hesaplama
```

1. Frekans:

Kitap Okuma Kategorileri ve Sayıları:

- **Basılı Kitap Okumayan (no_books)**: 395
- **Sadece Basılı Kitap Okuyan (print_only)**: 577
- **Sadece Dijital Kitap Okuyan (digital_only)**: 91
- **Hem Basılı Hem Dijital Kitap Okuyan (print_and_digital)**: 425

Bu frekanslar, her kitap okuma kategorisinde kaç okuyucu bulunduğunu gösterir.

2. Yüzde

Toplam Okuyucu Sayısı:

```
toplam_okuyucu <- sum(books_readers)  # Toplam okuyucu sayısını hesapla
print_only_proportion <- books_readers["print_only"] / toplam_okuyucu  # Sadece basılı kitap
print_only_proportion  # Oranı görüntüle
```

print_only 0.3877688

```
# Kesirli tablolar, iki veya daha fazla kategorik değişken arasındaki tüm olası varyasyonları
# Burada, kitap okuma kategorilerinin oranlarını içeren bir tablo oluşturuluyor
oran_tablosu <- books_readers / toplam_okuyucu
oran_tablosu  # Oluşturulan oran tablosunu görüntüle
```

```
    no_books       print_only     digital_only print_and_digital
  0.26545699       0.38776882       0.06115591       0.28561828
```

3. Mod

- **En Sık Görülen Kategori**: "Sadece Basılı Kitap Okuyan" (577). Bu, en çok okunan kitap türüdür.

4. Çapraz Tablo

Etnik Gruplar ve Bilim Haberlerine İlgi:

```r
# Bilim haberlerini aktif, sıradan veya ilgisiz bir şekilde tüketiyor musunuz?
# Farklı etnik grupların bilim haberlerine olan ilgisini göstermek için kategoriler ve sayıla
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)

# Elde edilen verileri bir veri çerçevesi (data frame) olarak birleştirme
my_table <- as.data.frame(rbind(white, black, hispanic))

# Her bir grubun toplamını hesaplama
my_table$rowsum <- rowSums(my_table)  # Her bir satırın toplamını hesapla
my_table["colsum",] <- colSums(my_table)  # Her bir sütunun toplamını hesapla

# Sonuçları görüntüleme
my_table  # Hesaplanan tabloyu görüntüle
```

	active	casual	uninterested	no_answer	rowsum
white	487	916	1431	28	2862
black	59	98	227	8	392
hispanic	89	152	183	23	447
colsum	635	1166	1841	59	3701

Yorumlar:

1. **Sadece Basılı Kitap Okuyanların Oranı**:
   - **Oran**: 0.3877688 (yaklaşık %38.78)
   - Bu, toplam okuyucu sayısının yaklaşık %38.78'inin yalnızca basılı kitap okuduğunu göstermektedir. Bu oran, okuyucuların çoğunluğunun basılı kitapları tercih ettiğini ortaya koymaktadır.
2. **Diğer Kategoriler**:
   - **Hiç Kitap Okumayanlar (no_books)**: Yaklaşık %26.55 kişi, hiç kitap okumadığını belirtmiştir.
   - **Yalnızca Basılı Kitap Okuyanlar (print_only)**: %38.78, yalnızca basılı kitap okuduğunu ifade etmiştir.
   - **Yalnızca Dijital Kitap Okuyanlar (digital_only)**: %6.12, yalnızca dijital kitap okuduğunu belirtmiştir.

- **Hem Basılı Hem de Dijital Kitap Okuyanlar (print_and_digital)**: %28.56, hem basılı hem de dijital kitap okuduğunu söylemiştir.

Bu oranlar, kitap okuma alışkanlıklarının dağılımını daha net bir şekilde gösterir. Yalnızca basılı kitap okuma oranı en yüksekken, yalnızca dijital kitap okuma oranı oldukça düşüktür.

3. **Bilim Haberlerine Tüketim**:
   - Her üç grup arasında "ilgili" ve "sıradan" olarak sınıflandırılan birey sayısının yüksek olması dikkat çekicidir. Ancak "ilgisiz" olanların sayısı en yüksektir; bu durum, bilim haberlerine olan ilgisizliği göstermektedir.
4. **Etnik Gruba Göre Dağılım**:
   - Beyaz grubun, diğer gruplara göre bilim haberlerine daha fazla ilgi gösterdiği gözlemlenmektedir. Siyah ve Hispanik grupların "aktif" oranları daha düşüktür.
5. **Cevap Vermeyenler**:
   - Cevap vermeyenlerin sayısı oldukça düşük kalmış, bu da verilerin güvenilirliğini artırmaktadır.

Sonuç

Bu analiz, hem kitap okuma kategorileri hem de etnik grupların bilim haberlerine olan ilgisini anlamak için çeşitli tanımlayıcı istatistikler kullanmaktadır.

- **Frekanslar**: Her kategorideki okuyucu sayısını göstermektedir.
- **Yüzdeler**: Belirli grupların toplam içindeki oranlarını belirtmektedir.
- **Mod**: En sık rastlanan kategoriyi ifade etmektedir.
- **Çapraz Tablo**: İki veya daha fazla değişken arasındaki ilişkileri ortaya koymaktadır.

Örnek 2: Titanik'te Hayatta Kalma Oranları

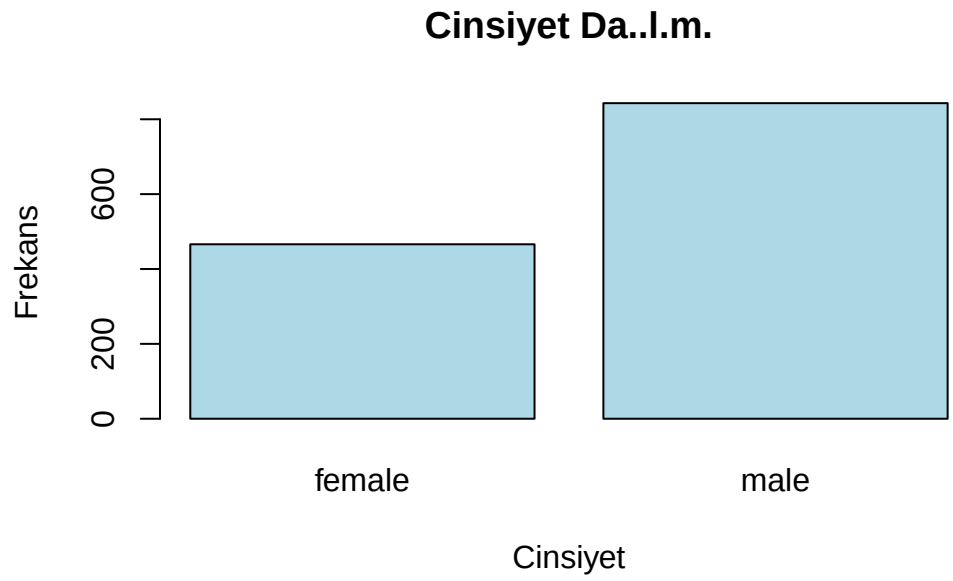
```
# Titanic verisini okuma
titanic <- read.csv("https://raw.githubusercontent.com/bio304-class/bio304-course-notes/maste

# Gerekli kütüphaneleri yükleyin
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

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

female male 466 843

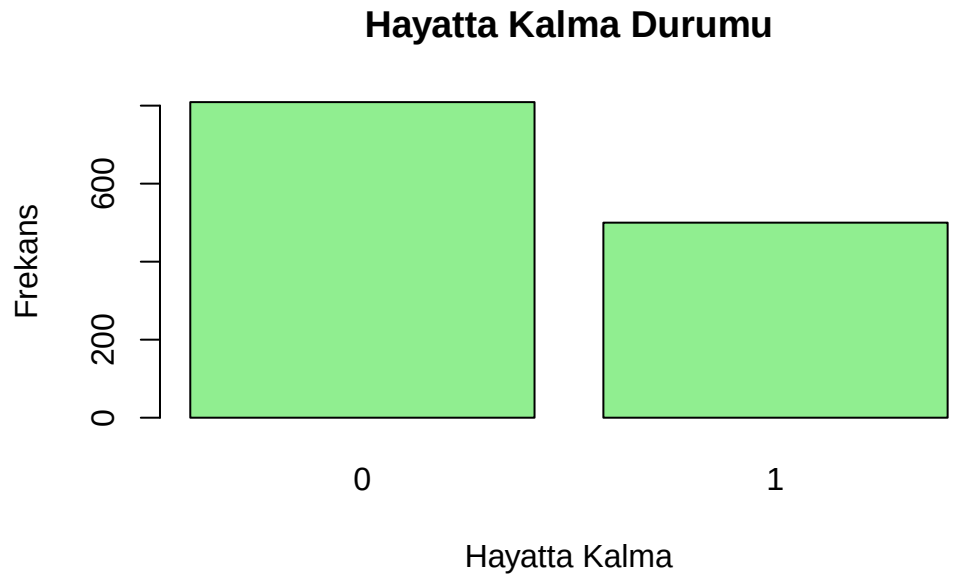
```r
# Cinsiyet dağılımını gösteren çubuk grafiği
barplot(table(titanic$sex), main = "Cinsiyet Dağılımı", xlab = "Cinsiyet", ylab = "Frekans",
```




```r
# 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", ylab
```




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

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

```
        0    1  Sum
```

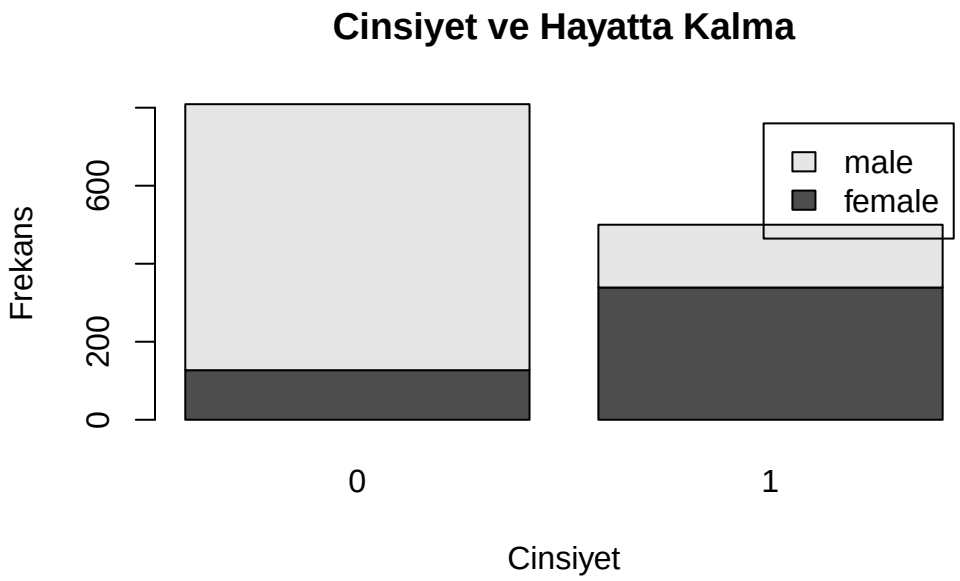
female 127 339 466 male 682 161 843 Sum 809 500 1309

```
# 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", y
```




```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
```

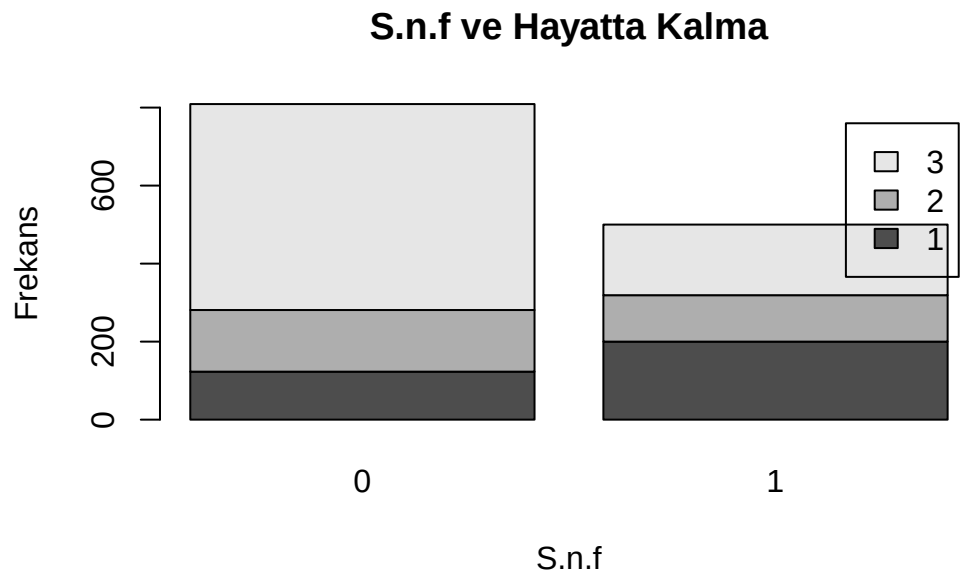
1 123 200 323 2 158 119 277 3 528 181 709 Sum 809 500 1309

```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

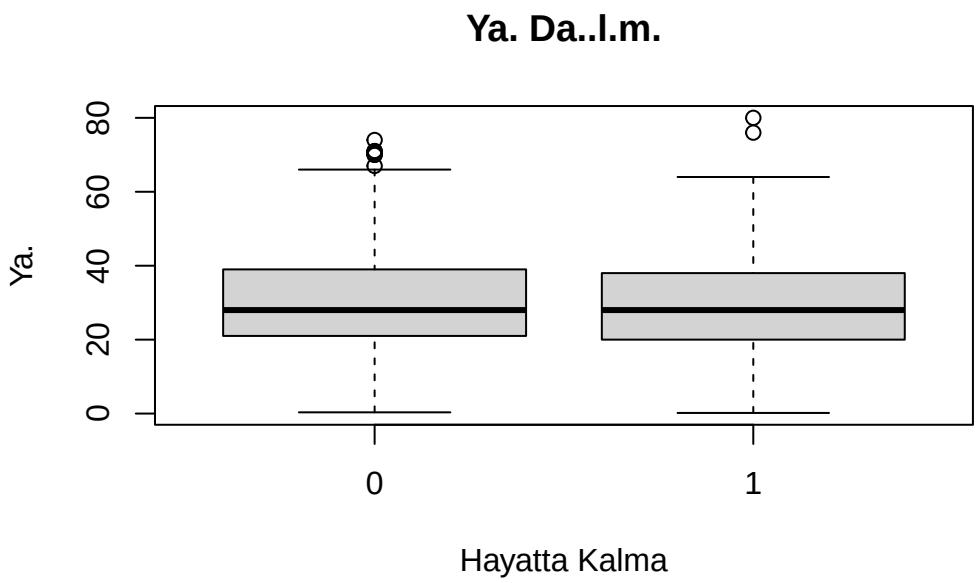
```r
# Sınıf ve hayatta kalma durumu için çubuk grafiği
barplot(table_2, legend.text = TRUE, main = "Sınıf ve Hayatta Kalma", xlab = "Sınıf", ylab =
```




```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", ylab =
```



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

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**:

```
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

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

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

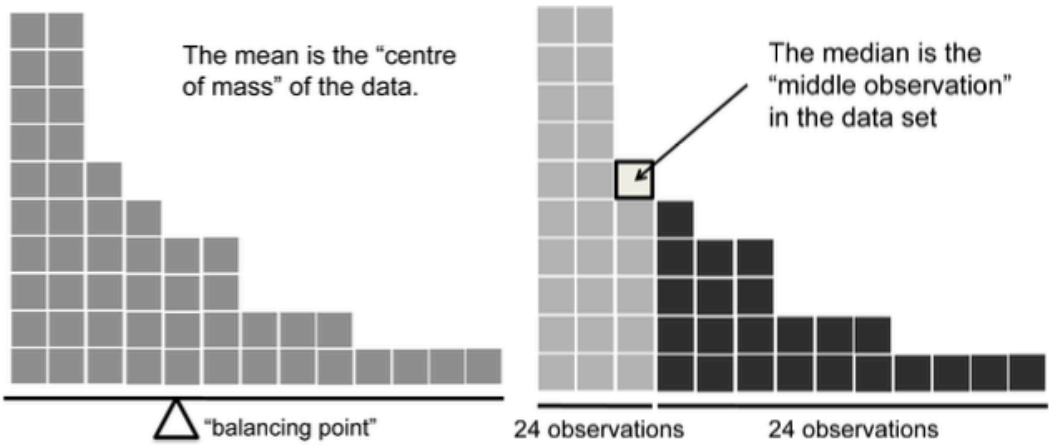
     - **Eğer** n çift ise:

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

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

[1] 49

> **Note**
>
> 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.



3. **Mod (Mode)**:

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

```
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
```

4. **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

5. **Standart Sapma (Standard Deviation)**:

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

$$\text{Standart Sapma} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

- **R Formülü**

```
sd(veri)  # veri, standart sapmasını almak istediğiniz sayısal vektördür.
```

[1] 28.45398

6. **Varyans (Variance)**:

- **Tanım**: Verilerin ortalamadan ne kadar saptığını gösteren bir ölçüdür. Standart sapmanın karesidir. Veriler arasındaki dağılımın ne kadar farklı olduğunu anlamamıza yardımcı olur.
- **Matematiksel Gösterim**:

$$\text{Varyans} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

- **R Formülü**

```
var(veri)  # veri, varyansını almak istediğiniz sayısal vektördür.
```

[1] 809.6289

Population	Sample
$\sigma = \sqrt{\frac{\sum(X - \mu)^2}{N}}$	$s = \sqrt{\frac{\sum(X - \bar{x})^2}{n - 1}}$
X - The Value in the data distribution μ - The population Mean N - Total Number of Observations	X - The Value in the data distribution $\bar{x}$ - The Sample Mean n - Total Number of Observations

x_1	$x_1 - \bar{x}$	$(x_1 - \bar{x})^2$
2	-1.5	2.25
6	2.5	6.25
5	1.5	2.25
3	-0.5	0.25
2	-1.5	2.25
3	-0.5	0.25
= 13.5		

7. **Çeyrekler Aralığı (Interquartile Range, IQR)**:

- **Tanım**: Verilerin ortasındaki yarısını kapsayan yayılımı gösterir. 1. çeyrek (Q1) ile 3. çeyrek (Q3) arasındaki farktır.
- **Matematiksel Gösterim**:
$$\text{IQR} = Q3 - Q1$$
- **R Formülü**

```r
IQR(veri)  # veri, çeyrekler aralığını almak istediğiniz sayısal vektördür.
```

[1] 44.5

8. **Küçük ve Büyük Çeyrekler (Quantiles)**:

- Verilerin belirli bir yüzdesine karşılık gelen değerlerdir. Örneğin, %25'lik çeyrek (Q1) ve %75'lik çeyrek (Q3).
- **R Formulü:**

```r
quantile(veri, probs = c(0.25, 0.75))  # %25 ve %75'lik çeyrekler
```

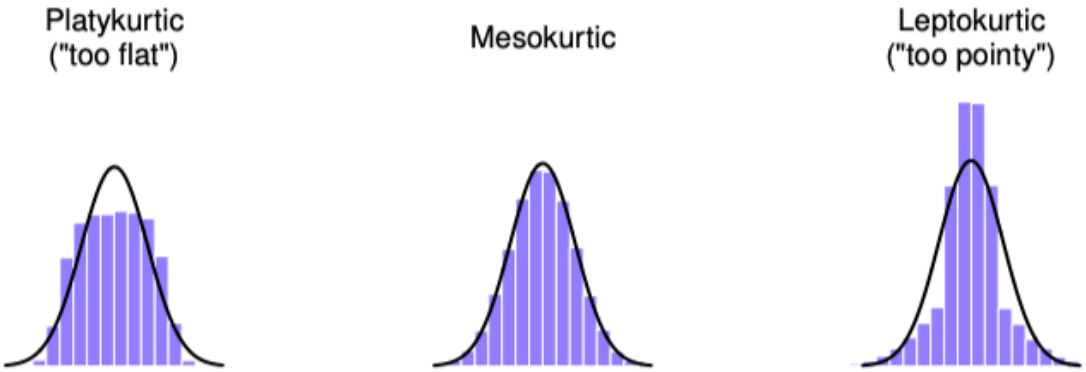
25% 75% 31.25 75.75

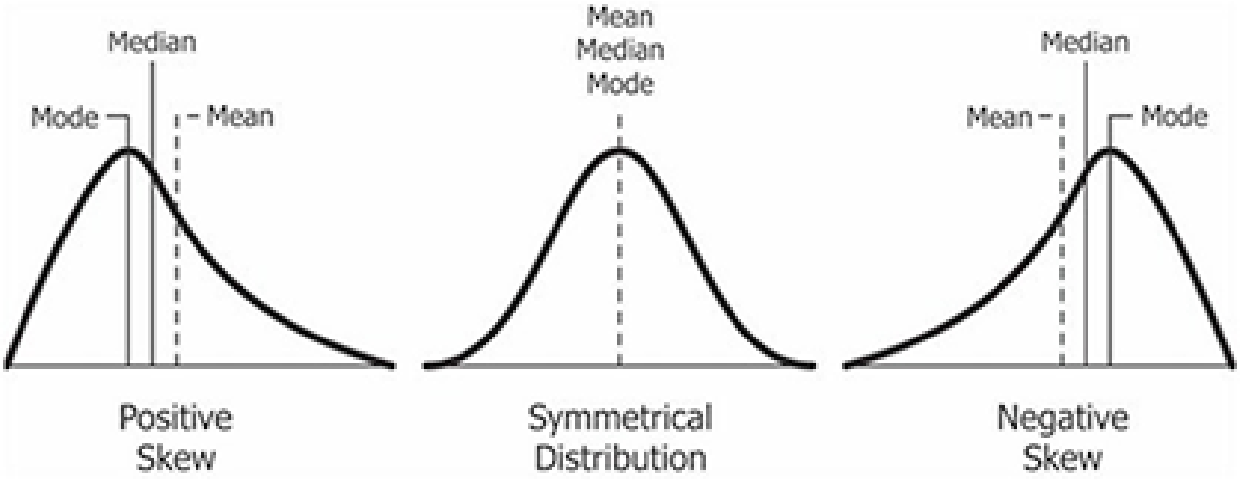
9. **Skewness (Eğrilik)**:

- Verinin dağılımının simetrik olup olmadığını gösterir. Pozitif veya negatif eğrilik değerleri, veri dağılımının sağa veya sola kaydığını belirtir.

10. **Kurtosis (Basıklık)**:

- Verinin dağılımının ne kadar "keskin" veya "düz" olduğunu ölçer. Yüksek kurtosis, veri dağılımının keskin zirvelere sahip olduğunu gösterirken, düşük kurtosis daha düz bir dağılımı işaret eder.


Platykurtic
("too flat")
Mesokurtic
Leptokurtic
("too pointy")


Median
Mode
Mean
Positive
Skew
Mean
Median
Mode
Symmetrical
Distribution
Median
Mean
Mode
Negative
Skew