Tugas Pertemuan 6

Husni Mubarok Ramadhan

2023-03-03

2 Statistik deskripsi dengan R

Sebelum memulai dengan konsep dasar analisis data, seseorang harus menyadari berbagai jenis data dan cara untuk mengatur data dalam file komputer.

2.1 Beberapa istilah dasar

Populasi – agregat subjek (makhluk, benda, kasus, dan sebagainya). Untuk studi tertentu, *populasi target* harus ditentukan: pada subjek mana kita akan menggeneralisasi atau menggunakan hasilnya?

Sampel – kumpulan subjek *dalam penelitian*. Secara umum, sampel harus representatif untuk populasi target.

Observasi – unit studi atau *subjek* atau individu. Seringkali manusia, terkadang juga hewan, tumbuhan atau apa pun.

Variabel – kualitas atau kuantitas, diukur atau dicatat untuk setiap subjek dalam sampel (usia, jenis kelamin, tinggi badan, berat badan, tingkat merokok, dll.).

Dataset – seperangkat nilai dari semua variabel yang menarik bagi semua individu dalam penelitian ini. Hasil numerik yang diperoleh dari dataset akan digunakan untuk menarik kesimpulan tentang populasi target.

2.2 Organisasi data

Himpunan data sebagian besar diatur (dan disimpan sebagai file komputer) dalam bentuk *matriks data*.

Matriks data yang mewakili jenis kelamin (1-laki-laki; 0-perempuan), usia, tidak. anak-anak, berat (kg), dan tinggi (cm) 7 orang:

No	Jenis Kelamin	Umur	Nomor anak	Berat	Tinggi
1	0	57	1	65	158
2	1	70	3	100	175
3	0	45	0	71	162
4	0	38	2	58	164
5	0	25	1	81	170
6	1	50	4	68	172
7	1	61	0	85	179

Setiap baris matriks semacam itu mewakili satu pengamatan. Semua baris memiliki panjang yang sama: data yang sama telah direkam untuk semua individu. Setiap kolom mewakili satu variabel. Misalnya, Berat adalah nama variabel, yang mewakili berat badan (dalam kg) seseorang.

2.3 Jenis data Data numerik -Data diskrit – variabel hanya dapat mengambil nilai bilangan bulat (0, 1, 2 dll.)

Contoh: jumlah anak, jumlah teman

-Data kontinu – setiap nilai bernomor nyata (seringkali dalam rentang tertentu) adalah contoh yang mungkin: berat badan, usia

Data kualitatif (non-numerik, kategoris) -Data nominal: kategori tidak berurutan contoh: golongan darah, warna mata.

-Data ordinal atau terurut:ord ered kategori contoh: tingkat merokok, sikap (baik-sedang-buruk).

Pengkodean numerik dari data nominal atau pesanan tidak membuat data numerik!

2.4 Meringkas/menyajikan data Data kontinu/diskrit Statistik lokasi ringkasan: rata-rata, median. Rata-rata sampel adalah rata-rata aritmatika data. Ini dapat dihitung, dengan menjumlahkan semua nilai data dan membagi jumlah dengan ukuran sampel total.

Contoh: Data : 8 6 8 9 9

Mean: $(8 + 6 + 8 + 9 + 9)/5 = 40/5 = 8$ Secara matematis: untuk variabel X, rata-rata sering dinotasikan sebagai ¯x dan dihitung sebagai:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

di mana x1,x2,. . . ,xn menunjukkan pengamatan variabel dan n adalah jumlah pengamatan dalam sampel.

R :

```
x <- c(8,6,8,9,9)
mean(x)
```

```
## [1] 8
```

Jika ada nilai yang hilang:

```
x <- c(1,3,5,2,9,NA,7,10)
mean(x)
```

```
## [1] NA
```

Terkadang menarik untuk mengurutkan nilai variabel dalam urutan naik atau turun. Nomor urutan pengamatan dalam baris seperti itu disebut sebagai peringkat. Median adalah titik tengah dari data yang dipesan – baik pengamatan tengah (jika jumlah pengamatan ganjil) atau rata-rata dari dua pengamatan tengah (jika jumlah pengamatan genap).

Contoh: Tinggi badan 11 orang (dalam cm): 155, 160, 171, 182, 162, 153, 190, 167, 168, 165, 191. Data yang dipesan: 153 155 160 162 165 167 168 170 171 182 191 Median: 167

```
x<-c(155, 160, 171, 182, 162, 153, 190, 167, 168, 165, 191)
median(x)
```

```
## [1] 167
```

Keuntungannya, tetapi kadang-kadang juga kerugian dari median adalah, bahwa itu tidak terpengaruh oleh nilai-nilai ekstrem dalam data. Tidak masalah, seberapa kecil atau seberapa besar nilai yang lebih besar atau lebih kecil dari median.

Baik rata-rata maupun median tidak memberikan informasi yang cukup tentang data: seseorang juga harus tahu tentang variabilitas.

Simpangan baku (SD, s) adalah kuantitas yang mencerminkan variabilitas sampel. Seseorang dapat menafsirkan SD sebagai perkiraan jarak rata-rata dari rata-rata.

Lebih tepatnya, SD didefinisikan sebagai akar kuadrat dari varians (s 2) (jumlah perbedaan kuadrat dari rata-rata dibagi dengan ukuran sampel dikurangi 1 (latter disebut sebagai varians sampel, s2).

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

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Demikian pula dengan rata-rata, simpangan baku sensitif terhadap ekstrem dalam data.

```
x<- c(1,4,5,7,8,11)
mean(x)
```

```
## [1] 6
```

```
median(x)
```

```
## [1] 6
```

```
var(x) # varians
```

```
## [1] 12
```

```
sd(x) # Standar deviasi
```

```
## [1] 3.464102
```

```
x<-c(1,4,5,7,8,110) # ubah pengamatan terakhir dari 11 menjadi 110
mean(x)
```

```
## [1] 22.5
```

```
median(x)
```

```
## [1] 6
```

```
var(x)
```

```
## [1] 1843.5
```

```
sd(x)
```

```
## [1] 42.936
```

Pendekatan yang lebih kuat adalah membagi distribusi data (yang dipesan) menjadi empat, dan menemukan poin-poin di bawah ini yaitu 25%, 50% dan 75% dari distribusi. Ini dikenal sebagai kuartil (median adalah kuartil kedua).

Contoh: 6 9 9 10 9 10 3 12 7 6 6 4 8 8 3 8 6 4 11 11 Sampel yang dipesan 3 3 4 4 6 6 6 6 7 8 8 8 9 9 9 9 10 11 11 12 Sampel yang dipesan dibagi menjadi 4 bagian: 3 3 4 4 6 | 6 6 6 7 8 | 8 8 9 9 9 | 9 10 11 11 12 Kuartil: titik potong: 6, 8 (median) dan 9.

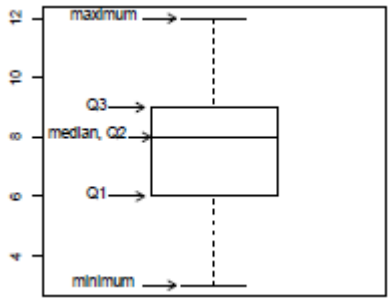
R Di R, Anda dapat menggunakan fungsi kuantil untuk mendapatkan median dan kuartil, atau Anda juga dapat menggunakan ringkasan fungsi te, untuk mendapatkan juga rata-rata:

```
z<-scan()
summary(z)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##
```

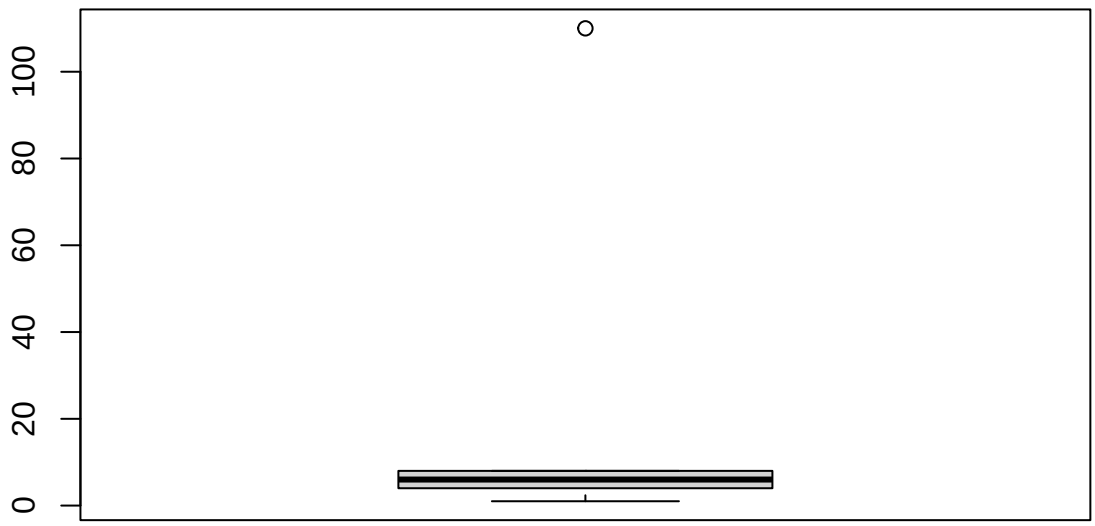
Variasi data dapat diringkas dalam rentang interquartile (IQR), jarak antara kuartil pertama dan ketiga (di sini: IQR= 9-6 = 3).

Secara umum persentil pth adalah titik potong p%- dari data yang dipesan (dari yang terkecil ke yang terbesar). Terkadang dalam statistik resmi, desil digunakan – persentil ke-10, ke-20, dll. Boxplot adalah representasi grafis dari median dan kuartil:



R:

```
boxplot(x)
```

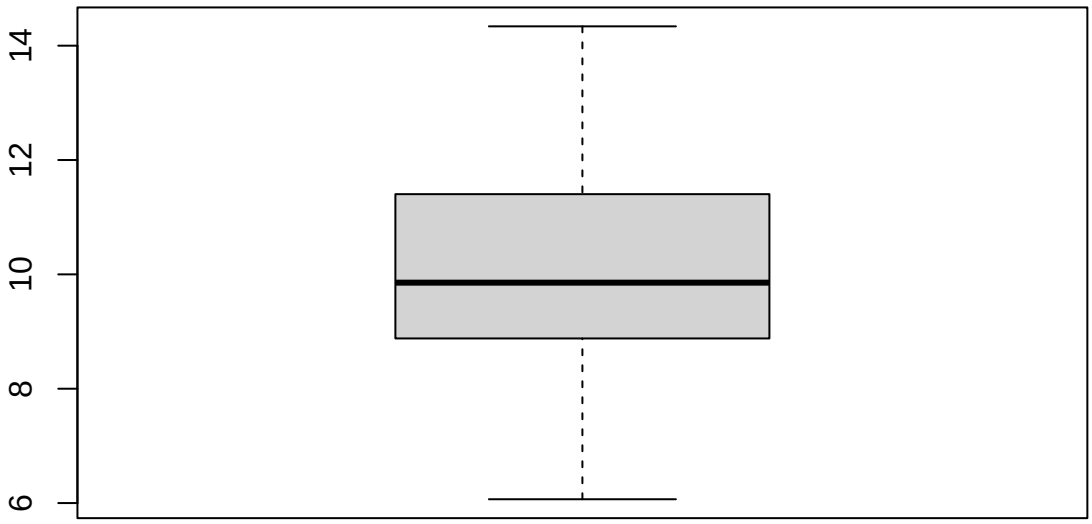


Boxplot memberikan gambaran umum tentang distribusi data. Ini sering digunakan untuk membandingkan data di berbagai kelompok.

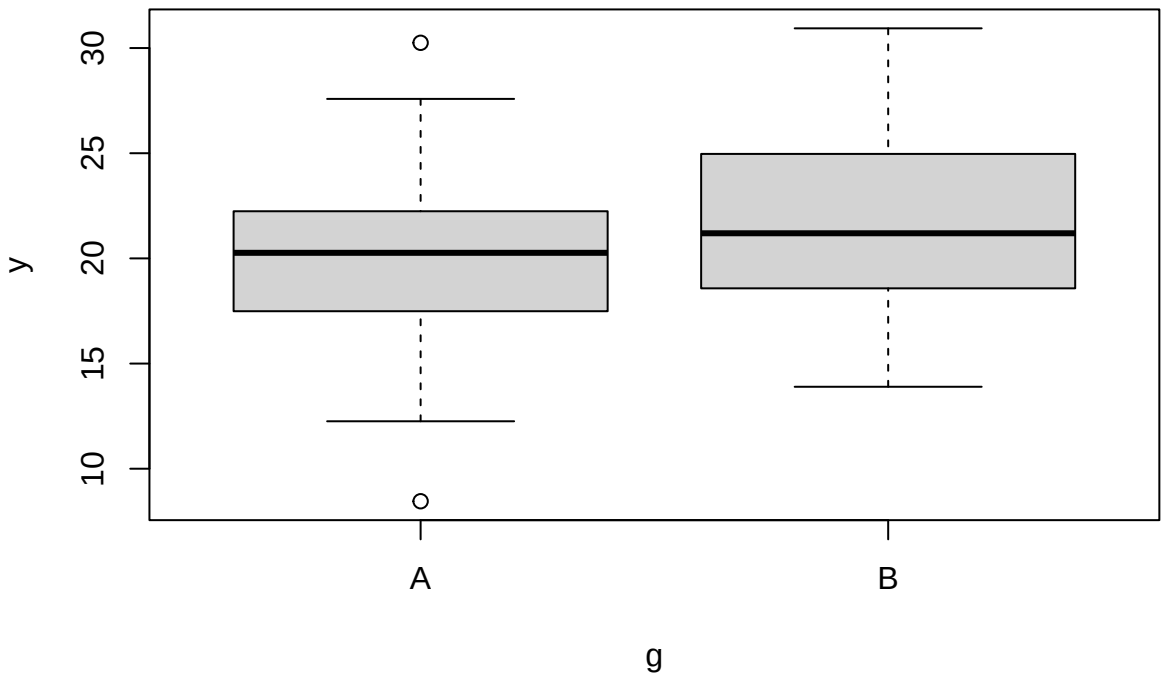
```
#boxplot(Y~g) # g adalah variabel kategoris dengan nilai 1, 2 dan 3 (genotipe)
```

R:

```
set.seed(123) # Untuk menghasilkan angka yang sama setiap kali dijalankan

# Variabel numerik c
x <- rnorm(50, mean = 10, sd = 2)

# Variabel numerik y
y <- rnorm(50, mean = 20, sd = 5)

# Variabel faktor g
g <- rep(c("A", "B"), each = 25)

 boxplot(x)
```




```
boxplot(y~g)
```

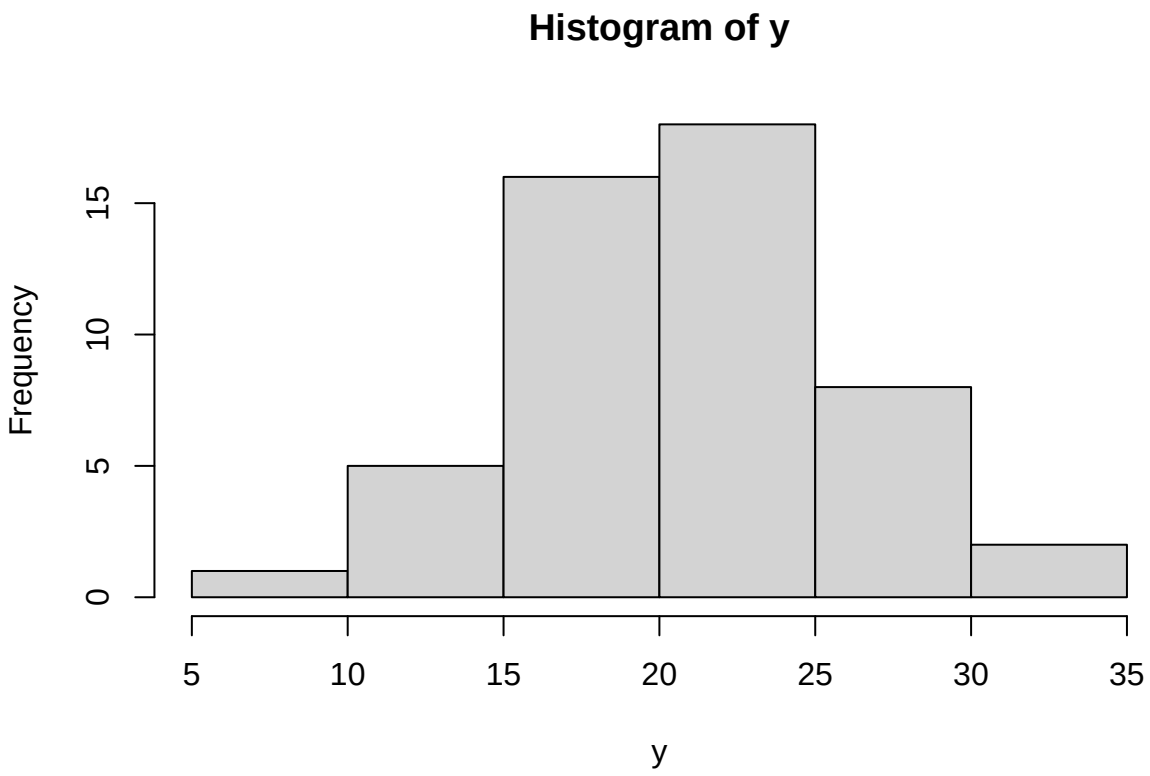


Kedua perintah tersebut adalah perintah di bahasa R untuk membuat boxplot.

Perintah pertama **`boxplot(x)`** akan membuat boxplot untuk satu variabel numerik x. Boxplot ini akan menunjukkan median (garis tengah kotak), kuartil pertama dan ketiga (batas atas dan bawah kotak), dan rentang interkuartil (jarak antara kuartil pertama dan ketiga). Boxplot juga akan menunjukkan pencilan (outliers) jika ada.

Perintah kedua **`boxplot(Y~g)`** akan membuat boxplot untuk variabel numerik Y dengan membaginya berdasarkan variabel faktor g. Dalam hal ini, boxplot akan menunjukkan distribusi Y untuk setiap kategori yang ada di variabel g. Boxplot ini juga akan menunjukkan median, kuartil pertama dan ketiga, rentang interkuartil, serta pencilan untuk setiap kategori. Boxplot ini sangat berguna untuk membandingkan distribusi Y antar kategori g.

```
hist(y)
```



Untuk data nominal, rata-rata dan standar deviasi tidak masuk akal; begitu pula median dan persentil. Untuk melihat distribusi data, lihat tabel frekuensi.

```
eyecol<-c(1,2,1,2,2,2,3,3,1,4,2,2,2,3,1,4,3,2,1,1,1)
table(eyecol)
```

```
## eyecol
## 1 2 3 4
## 7 8 4 2
```

```
eyecol<-factor(eyecol, labels=c("blue","grey","brown","green"))
table(eyecol)
```

```
## eyecol
##  blue  grey brown green
##     7     8     4     2
```

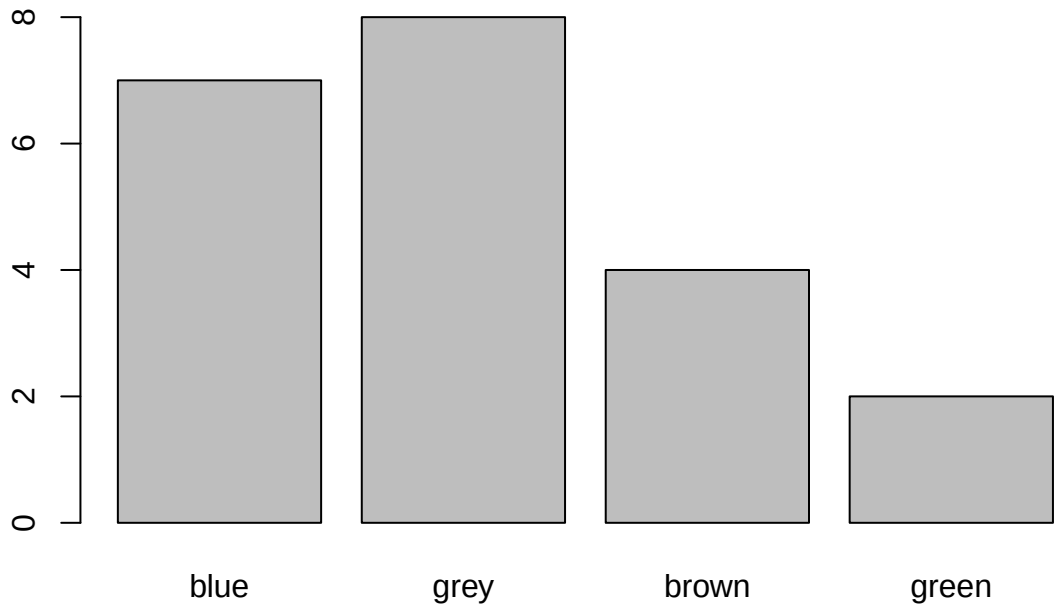
```
prop.table(table(eyecol))
```

```
## eyecol
##      blue      grey     brown     green
## 0.3333333 0.3809524 0.1904762 0.0952381
```

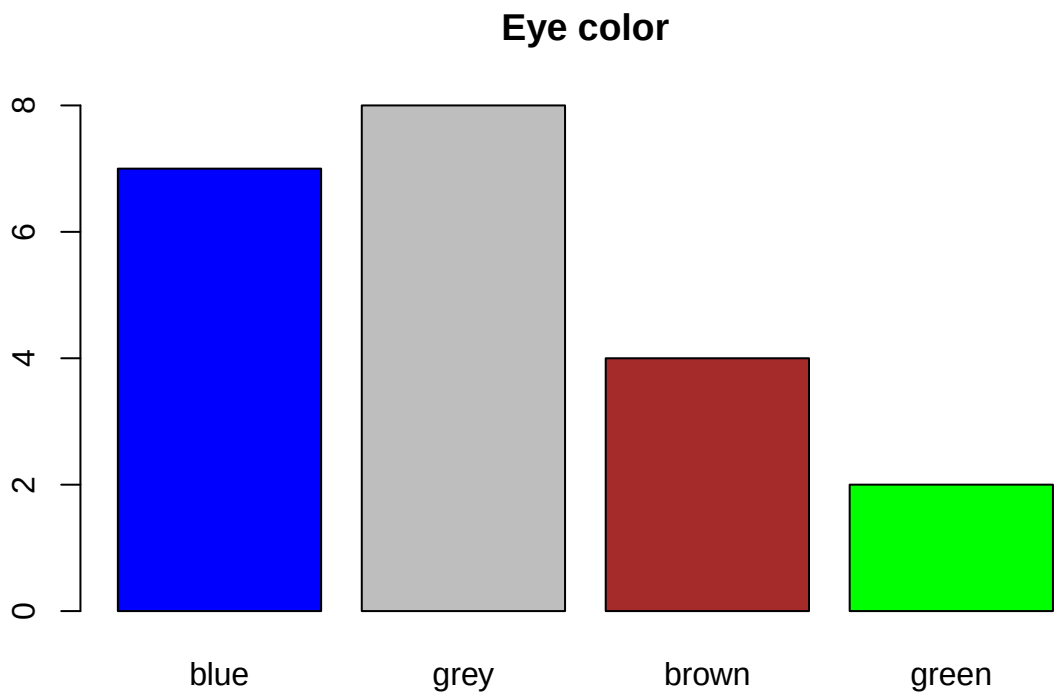
```
round(100*prop.table(table(eyecol)),1)
```

```
## eyecol
##  blue  grey brown green
##  33.3  38.1  19.0   9.5
```

```
barplot(table(eyecol))
```




```
barplot(table(eyecol),col=c("blue","grey","brown","green"),main="Eye color")
```



- Fungsi factor() digunakan untuk mengubah vektor numerik eyecol menjadi faktor, di mana setiap level faktor direpresentasikan oleh label yang telah ditentukan.
- Fungsi table(eyecol) digunakan untuk menghitung frekuensi kemunculan setiap level faktor dalam vektor eyecol.
- Fungsi prop.table(table(eyecol)) digunakan untuk menghitung proporsi dari setiap level faktor dalam vektor eyecol.
- Fungsi round() digunakan untuk membulatkan nilai proporsi menjadi 1 angka desimal.
- Hasil terakhir menunjukkan bahwa 33.3% orang memiliki mata biru, 47.6% orang memiliki mata abu-abu, 14.3% orang memiliki mata coklat, dan 9.5% orang memiliki mata hijau.
- Baris terakhir adalah perintah untuk membuat diagram batang (barplot) dari distribusi frekuensi. Di sini, fungsi barplot() digunakan untuk membuat diagram batang dan opsi col=c(“blue”,“grey”,“brown”,“green”) digunakan untuk menentukan warna

2.5 R: Statistik deskriptif menurut kelompok, tabel 2 dimensi

Misalkan Anda ingin membandingkan cara atau statistik deskriptif lainnya dalam subkelompok yang berbeda dari sampel Anda. Di R, Anda dapat menggunakan fungsi tapply untuk itu. Fungsi ini mengambil 3 argumen: variabel numerik, variabel pengelompokan kategoris dan fungsi yang akan diterapkan.

Untuk memahami, cara kerjanya, coba contoh berikut:

```r
weight <- c(56, 67, 65, 78, 49, 87, 55, 63, 70, 72, 79, 52, 60, 78, 90)
sex <- c(1,1,1,2,1,2,1,1,1,2,1,1,1,2,2)
tapply(weight,sex,mean)
```

```
##    1    2
## 61.6 81.0
```

```r
tapply(weight,sex,summary)
```

```
## $`1`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   49.00   55.25   61.50   61.60   66.50   79.00
##
## $`2`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      72      78      78      81      87      90
```

```r
eyecol<-c(1,2,1,2,2,2,3,3,1,4,2,2,2,3,1)
table(sex,eyecol)
```

```
##    eyecol
## sex 1 2 3 4
##   1 3 5 2 0
##   2 1 2 1 1
```

```r
prop.table(table(sex,eyecol),1)
```

```
##    eyecol
## sex   1   2   3   4
##   1 0.3 0.5 0.2 0.0
##   2 0.2 0.4 0.2 0.2
```

```r
prop.table(table(sex,eyecol),2)
```

```
##    eyecol
## sex         1         2         3         4
##   1 0.7500000 0.7142857 0.6666667 0.0000000
##   2 0.2500000 0.2857143 0.3333333 1.0000000
```

Baris pertama adalah perintah di bahasa R untuk membuat vektor numerik **`weight`**, yang berisi data berat badan, dan vektor numerik **`sex`**, yang berisi data jenis kelamin (1: perempuan, 2: laki-laki).

Perintah **`tapply()`** kemudian digunakan untuk menghitung statistik ringkasan (mean dan summary) dari data berat badan berdasarkan jenis kelamin.

Penjelasan:

- Fungsi **`tapply()`** digunakan untuk menghitung statistik ringkasan dari data **`weight`** berdasarkan **`sex`**.
- Pada output pertama, nilai mean dari data berat badan dihitung untuk setiap jenis kelamin.

- Pada output kedua, nilai statistik ringkasan (min, max, median, dll.) dihitung untuk setiap jenis kelamin.

Baris berikutnya adalah perintah untuk membuat tabel frekuensi dari kombinasi data **`sex`** dan **`eyecol`**. Kemudian, fungsi **`prop.table()`** digunakan untuk menghitung proporsi dari setiap level faktor dalam tabel frekuensi, baik dalam baris maupun kolom.