توصیف داده‌ها با R

۱ مقدمه

استفاده از روش‌های مناسب توصیف داده‌ها به مقیاس آن‌ها وابسته است. بنابراین، ابتدا لازم است تا انواع مقیاس داده‌ها را بشناسیم.

۱.۱ اندازه‌گیری و مقیاس داده‌ها

نسبت دادن یک عدد حقیقی به یک ویژگی اعضای یک جامعه را، بر اساس ضابطه‌ای مشخص، اندازه‌گیری می‌نامند. روشن است که ویژگی‌های مورد نظر می‌توانند ماهیت‌های مختلفی داشته باشند و در نتیجه، نوع اعدادی که به آن‌ها نسبت داده می‌شوند نیز متفاوت هستند. به‌عنوان نمونه، جامعه را سیب‌های یک باغ در نظر بگیرید. ویژگی مورد نظر می‌تواند وزن سیب‌ها یا رنگ آن‌ها باشد. نوع اعدادی که به وزن سیب‌های باغ نسبت داده می‌شود با اعدادی که برای رنگ آن‌ها تعیین می‌شوند، متفاوت است.

نوع مقیاس اعداد، نماینده‌ای از ماهیت ویژگی مورد علاقه است. منطقی است که انتظار داشته باشیم روش‌های توصیفی مناسب اعداد به مقیاس آن‌ها وابسته باشند. در یک رده‌بندی کلی، چهار نوع مقیاس اندازه‌گیری مطرح شده‌اند:

۱. اسمی[۱]: هر گاه عدد مورد نظر (که معمولا یک عدد طبیعی است) برای شناسایی یا رده‌بندی افراد، اشیاء، مکان‌ها یا ویژگی‌ها به‌کار رود، آن را با مقیاس اسمی می‌نامند. مثلا اعداد ۱، ۲ و ۳ که به سه رنگ مختلف عروسک‌های یک فروشگاه نسبت داده می‌شوند، تنها مشخص می‌کنند که هر عروسک دارای چه رنگی است. اعداد با مقیاس اسمی را نمی‌توان برای مقایسه با چهار عمل اصلی حساب به‌کار برد. همچنین طبیعی است که نمی‌توان مدعی شد عروسکی که دارای رنگ ۲ است، دو برابر عروسکی که دارای رنگ ۱ است، رنگین است.

۲. ترتیبی[۲]: هر گاه مقیاس مورد نظر (که یک عدد حقیقی است) برتری را بیان کند، مقیاس ترتیبی نامیده می‌شود. در این مقیاس، اعداد نسبت‌داده شده به افراد، آن‌ها را از بیشترین به کمترین مقدار ویژگی مورد نظر مرتب می‌کنند. به‌عنوان مثال، اگر مهندس یک کارخانه کارگران را از نظر مهارت با اعداد ۱ تا ۴ درجه‌بندی کند، کارگر با مهارت ۴ از کارگر با مهارت ۲، ماهرتر است. اما نمی‌توان گفت مهارت او دو برابر دیگری است. برای اعداد این مقیاس نیز نمی‌توان چهار عمل اصلی را به‌کار برد.

۳. فاصله‌ای[۳]: هر گاه مقیاس اعداد (که حقیقی هستند)، نسبت دو تفاضل یا دو فاصله را ثابت نگه دارد، آن را یک مقیاس فاصله‌ای می‌نامند. برای توضیح، فرض کنید x_1، x_2، x_3 و x_4 چهار عدد با مقیاس فاصله‌ای باشند. در این

[۱] Nominal

[۲] Ordinal

[۳] Interval

صورت نسبت

$$\frac{x_۴ - x_۳}{x_۲ - x_۱}$$

ثابت می‌ماند و به واحد اندازه‌گیری وابسته نیست. مثلا مقیاس اندازه‌گیری دمای اجسام از نوع فاصله‌ای است. در این مقیاس، صفر به معنی پوچ و هیچ نیست. مثلا دمای صفر با واحد سانتیگراد، معادل ۳۲ با واحد فارنهایت است. بر روی این مقیاس هر چهار عمل اصلی قابل استفاده هستند.

۴. نسبتی[۴]: هر گاه مقیاس اندازه‌گیری، نسبت دو عدد را حفظ کند، آن را نسبتی می‌نامند. یعنی برای هر دو عدد $x_۱$ و $x_۲$ باید در این مقیاس نسبت $\frac{x_۱}{x_۲}$ به واحد اندازه‌گیری وابسته نباشد و ثابت بماند. مثلا وزن اشیاء، نمره درسی دانشجویان، و سرعت وزش باد از این جمله‌اند.

این مقیاس برترین مقیاس اندازه‌گیری است، همه ویژگی‌های مقیاس‌های قبلی را داراست و صفر مطلق در آن تعریف‌شدنی است.

در دنیای واقعی، اعداد با هر چهار مقیاس ذکرشده به فراوانی وجود دارند. کافی است به اطراف خود نگاه کنید تا آن‌ها را ببینید: سردی یا گرمی هوا (مقیاس اسمی)، رضایت‌مندی شما از یک برنامه تلویزیونی (مقیاس ترتیبی)، دمای هوا (مقیاس فاصله‌ای)، تعداد افراد خانواده (مقیاس نسبتی).

۲.۱ متغیرها

ویژگی مورد مطالعه در جامعه، برای اعضای مختلف جامعه یکسان نیست و از عضوی به عضو دیگر تغییر می‌کند. به چنین ویژگی یک متغیر می‌گوییم. به‌عنوان نمونه، گروه

[۴] Ratio

خونی، هوش، مهارت یا وزن افراد یک شهر می‌توانند متغیره‌های مورد نظر باشند. با توجه به مقیاس متغیرها، دو رده متغیر قابل تعریف هستند:

۱. متغیر کیفی[۵] یا گروهی: متغیرهایی که اعداد آن‌ها دارای مقیاس اسمی یا ترتیبی هستند و به نوعی اعضای جامعه را گروه‌بندی می‌کنند. مثل گروه خونی یا جنسیت.

۲. متغیر کمی[۶]: متغیرهایی که اعداد آن‌ها دارای مقیاس‌های فاصله‌ای یا نسبتی هستند. مانند نمره درسی، وزن، و هزینه یک سفر درون‌شهری.

اکنون می‌توان داده‌ها و انواع آن‌ها را تعریف کنیم.

۳.۱ داده‌ها

فرض کنید بخواهیم یک (یا چند) ویژگی یک جامعه را (که یک متغیر است) مطالعه کنیم. اگر این ویژگی (متغیر) را برای تمام اعضای جامعه یا نمونه‌ای از آن (با مقیاسی مناسب) اندازه‌گیری کنیم، یک مجموعه از اعداد حاصل می‌شود که آن را داده‌ها[۷] می‌نامند.

داده‌ها را به دو دسته تقسیم می‌کنند:

۱. داده‌های گسسته[۸]: اعدادی که برای متغیرهای کیفی یا کمی شمارشی اندازه‌گیری می‌شوند را گسسته می‌گویند. مثل تعداد تصادف‌های جاده‌ای در یک سال، و سالم یا بیمار بودن پسرهای یک خانواده.

---

[۵] Qualitative variable

[۶] Quantitative variable

[۷] Data

[۸] Discrete

۲. داده‌های پیوسته[۹]: اعدادی که برای متغیرهای کمی (غیرشمارشی) اندازه‌گیری می‌شوند. برای این داده‌ها، متغیرها هر عدد حقیقی را می‌توانند اختیار کنند. مثل وزن، طول، و دما.

حال که انواع داده‌ها را شناختیم، باید بتوانیم آن‌ها را توصیف کنیم. بسته به نوع داده‌ها، روش‌های توصیفی مناسب آن‌ها متفاوت هستند. اما چرا توصیف؟

داده‌های خام، به داده‌هایی گفته می‌شود که در یک یا چند ستون از اعداد فهرست می‌شوند. این انبوه داده‌ها می‌توانند حاوی اطلاعات ارزشمندی باشند. اما

- این اطلاعات با نگاه کردن به آن‌ها قابل استخراج نیستند.
- ماهیت تغییرپذیری داده‌ها و ویژگی‌های آن‌ها قابل دستیابی نیستند.

به‌منظور شناخت اولیه ویژگی‌های پایه‌ای داده‌ها و دسترسی (جزیی) به اطلاعات ارزشمند درون آن‌ها، که اولین گام در پردازش داده‌ها محسوب می‌شود، توصیف و خلاصه‌سازی آن‌ها در قالب

الف) جدول‌ها

ب) نمودارهای بصری

ج) یک یا چند خلاصه عددی

ضروری است. مطالبی که در این نوشته برای توصیف داده‌ها مطرح می‌شوند، بخشی از زمینه‌ای است که جان توکی آن را بنیان نهاد و به تحلیل اکتشافی داده‌ها[۱۰] (EDA) معروف شد. در ادامه، ابتدا مقدمات لازم برای استفاده از نرم‌افزار R و ورود داده‌ها به آن را مطرح می‌کنیم.

---

[۹] Continuous

[۱۰] Exploratory Data Analysis

۲ مقدمه‌ای بر R

۱.۲ خلاصه کردن داده‌های کمی:

توابعی برای خلاصه کردن داده‌ها موجود می‌باشد.از جمله توابع زیر:

min(x): کوچکترین مقدار x

max(x): بزرگترین مقدار x range(x): فاصله کوچکترین و بزرگترین مقدار x

IQR(x): برد میان چارکی مقادیر x

mean(x): میانگین مقادیر x

median(x): میانه مقادیر x

var(x): واریانس مقادیر x

sd(x): انحراف معیار مقادیر x

cor(x,y): همبستگی بین x و y

quantaile(x,p): کواریانس بین x و y

مثال: دستورات زیر نحوه محاسبه میانگین، انحراف معیار، واریانس و میانه را برای ۱۰۰ داده تصادفی از توزیع نرمال استاندارد نشان می‌دهد.

```
> set.seed(10)
> x <- rnorm(100)
> mean(x)
[1] -0.1365489
> sd(x)
[1] 0.9412359
> var(x)
[1] 0.885925
```

```
> median(x)
[1] -0.1933164
```

- **نکته:** اگر در داده‌ها مقدار گمشده موجود باشد، نمی‌توان خلاصه‌های آماری را برای آن محاسبه کرد. برای حل این مشکل می‌توان با شناسه na.rm=T مقادیر گمشده را نادیده گرفت و آماره مورد نظر را محاسبه کرد.

```
> y <- scan()
1: 1 50 6 10 85 NA
7:
Read 6 items
> mean(y)
[1] NA
> mean(y , na.rm=T)
[1] 30.4
```

با استفاده از تابع quantaile(x،p) می‌توان چندک‌های مهم را بدست آورد.

```
> quantile(x)
0%         25%         50%         75%        100%
-2.1852868 -0.8290776 -0.1933164  0.5933604  2.2205197
```

- **نکته:** تابع quantaile(x،p) به‌طور پیش‌فرض مقادیر مینیمم، ماکزیمم، چارک اول و دوم و سوم را محاسبه می‌کند. از تابع ()quantaile می‌توان برای محاسبه

هر چندک دلخواه استفاده کرد. مثلا برای محاسبه دهک‌های متغیرx از دستور زیر استفاده می‌کنیم:

```
> set.seed(10)
>  x <- rnorm(100)
> a <- seq(0,1,0.1)
> quantile(x,a)
```

```
0%         10%         20%         30%         40%         50%
-2.32101703 -1.36779025 -1.10005315 -0.68135560 -0.40052066  0.06175308
 60%         70%         80%         90%        100%
 0.31773630  0.62061624  0.77836759  1.07208676  2.06820961
```

تفاضل چارک سوم و چارک اول، برد میان چارکی نامیده می‌شود و توسط تابع IQR() قابل محاسبه است.

```
> IQR(x)
[1] 1.422438
```

تابع summary() یک نمایش خلاصه از متغیرهای عددی ارائه می‌کند. برای مثال مجموعه داده Loblolly مربوط به رشد درختان کاج استفاده می‌کنیم.

```
> summary(Loblolly)
```

```
height           age             Seed
```

```
Min.   : 3.46   Min.   : 3.0   329    : 6
1st Qu.:10.47   1st Qu.: 5.0   327    : 6
Median :34.00   Median :12.5   325    : 6
Mean   :32.36   Mean   :13.0   307    : 6
3rd Qu.:51.36   3rd Qu.:20.0   331    : 6
Max.   :64.10   Max.   :25.0   311    : 6
(Other):48
```

می‌بینیم که نتیجه اجرای تابع ()summary مقادیر مینیمم، چارک اول، میانه، میانگین، چارک سوم و ماکزیمم است.

- **نکته:** ملاحظه می‌کنید که متغیر seed مربوط به نوع بذر است و از آنجایی که یک متغیر اسمی است، تنها فراوانی مربوط به هر سطح آن مشخص می‌شود.

همچنین می‌توان از تابع ()summary برای متغیر عددی height که مربوط به مجموعه داده Loblolly است استفاده کرد.

```
> attach(Loblolly)
> names(Loblolly)
[1] "height" "age" "Seed"
> summary(height)
Min. 1st Qu. Median Mean 3rd Qu. Max.
3.46 10.47 34.00 32.36 51.36 64.10
```

۱.۱.۲ نمودار شاخه و برگ:

روش‌های مختلفی برای نمایش گرافیکی داده‌ها وجود دارد. اگر مجموعه داده‌ها نسبتا کوچک باشد نمودار شاخه و برگ برای مشاهده شکل توزیع بسیار کارآمد است. دستور لازم برای رسم نمودار شاخه و برگ در **R**، stem(x, scale = 1, width = 80) می‌باشد. که x داده‌های ما، scale برای کنترل طول نمودار و width عرض نمودار می‌باشد. فرض کنید اعداد زیر امتیازات کسب شده توسط هر بازیکن در تیم بسکتبال باشد.

1 , 2 , 8 , 4 , 5 , 14 , 31 , 26 , 0 , 0 , 0 , 32 , 12 , 9 , 15 , 17 , 20 , 50 , 3 , 62

داده‌ها را با دستور scan() به صورت زیر وارد می‌کنیم.

```
> scores <- scan()
1: 1 2 8 4 5 14 31 26 0 0 0 32 12 9 15 17 20 50 3 62
21:
Read 20 items
```

و با تابع stem() نمودار شاخه و برگ را رسم می‌کنیم.

```
> stem(scores)

The decimal point is 1 digit(s) to the right of the |

0 | 00012345892457
```

```
2 | 0612
4 | 0
6 | 2
```

عدد سمت چپ خط را به‌عنوان شاخه و عدد سمت راست را به عنوان برگ در نظر می‌گیرند.

مثال‌های دیگر با استفاده از مجموعه داده‌های موجود در R:

```
> stem(islands)

The decimal point is 3 digit(s) to the right of the |

0 | 000000000000000000000000000000111111222338
2 | 07
4 | 5
6 | 8
8 | 4
10 | 5
12 |
14 |
16 | 0

> stem(log10(islands))

The decimal point is at the |

1 | 1111112222233444
1 | 5555556666667899999
2 | 3344
2 | 59
```

```
3 |
3 | 5678
4 | 012
```

- **نکته:** در دستور ()stem مقادیر گم‌شده و بی‌نهایت حذف می‌شوند.

۲.۱.۲ نمودار هیستوگرام(بافت‌نگار):

با رسم هیستوگرام یک متغیر، می‌توان شکل تقریبی توزیع آن را مشاهده کرد. که صورت کلی آن به شرح زیر است.

hist(x, breaks="Sturges", prob=FALSE)

آرگومان breaks تعداد دسته‌ها را نشان می‌دهد. تعداد دسته‌های خیلی زیاد یا خیلی کم می‌تواند شکل نامناسبی را ایجاد کند. در صورت پیش فرض زبان R از فرمولStrurges استفاده می‌کند که رابطه آن به شکل زیر است.

$$[log_۲(n) + ۱]$$

روش‌های دیگری نیز وجود دارد که پهنای دسته را معین می‌کند، مثل فرمول-Freedman Diaconis(FD) که بر اساس محدوده بین چارکی (iqr) قرار دارد. رابطه آن به شکل زیر است.

$$۲ * iqr * n^{-(۱/۳)}$$

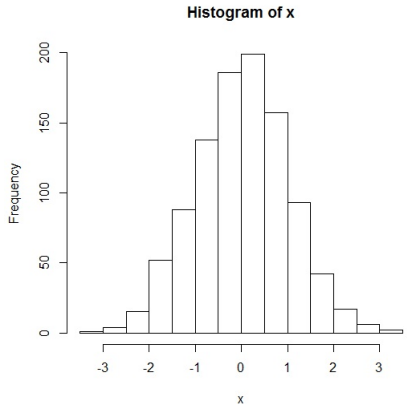
فرمول دیگری نیز وجود دارد و توسط Soott پیشنهاد شده است.

$$۳/۵ * s * n^{-(۱/۳)}$$

که در آن s انحراف معیار است.

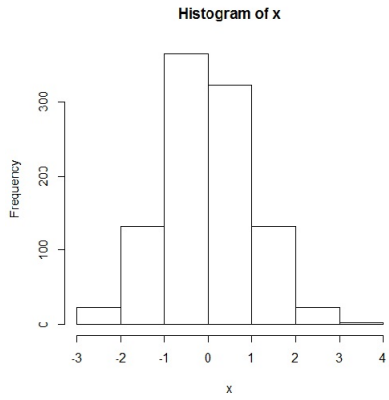
مثال:

```
> x=rnorm(1000)
> hist(x)
```



این تابع تعداد دسته‌ها را به طور خودکار انتخاب می‌کند. این انتخاب به‌گونه‌ای است که ضمن حفظ اطلاعات جزئی، هموارترین نمودار ممکن نیز رسم می‌شود.
اگر بخواهیم کنترل کاملا در اختیار ما باشد، می‌توانیم از آرگومان nclass=num. of class استفاده کنیم. به عنوان مثال nclass=7 برای هیستوگرام مورد نظر ۷ دسته ایجاد می‌کند.

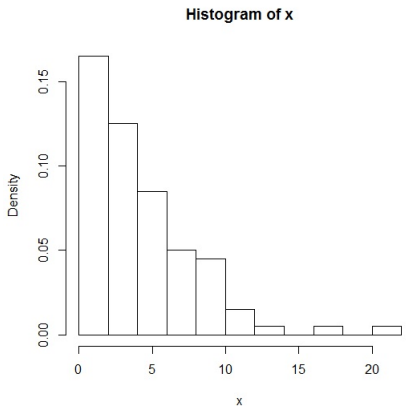
```
> hist(x,nclass=7)
```



در دستور prob=T و یا freq=F فراوانی غیر فعال می‌شود و در واقع چگالی ظاهر می‌گردد و شکل به گونه‌ای تنظیم می‌گردد که مجموع مساحت‌ها در هیستوگرام برابر واحد شود.

مثال:

```
> x <- rchisq(100, df = 4)
> hist(x, freq = FALSE)
```

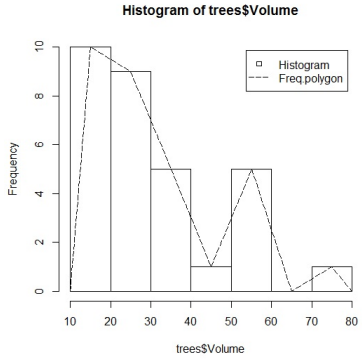


۳.۱.۲ چندضلعی فراوانی(چند بر فراوانی):

همراه با نمودار هیستوگرام نمودار دیگری را می‌توان رسم نمود که چندضلعی فراوانی نامیده می‌شود و از شهرت به سزایی در آمار توصیفی برخودار است. برای ترسیم نمودار گفته شده به مثال زیر توجه کنید.

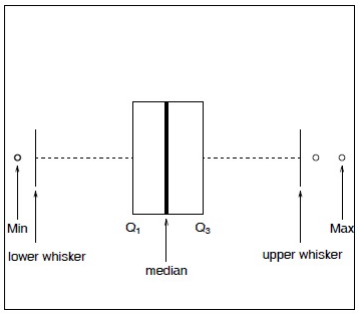
```
> tmp <- hist(trees$Volume)
>  lines(c(min(tmp$breaks),tmp$mids,max(tmp$breaks)),
+ c(0,tmp$counts,0),lty=5,lwd=1.75)
>  legend("topright",c("Histogram","Freq.polygon"),
+ lty=c(NA,5),lwd=c(NA,1.75),pch=c(22,NA),inset=0.05)
```

در کدهای فوق ابتدا تمام مشخصات هیستوگرام در شی tmp قرار می‌گیرد. سپس در تابع lines() از آرگومان‌های هیستوگرام استفاده می‌شود. اولین آرگومان breaks است که دسته‌ها را نشان می‌دهد. دومین آرگومان mids است که بردار نقاط وسط دسته‌ها را نشان می‌دهد. سومین آرگومان counts است که تعداد فراوانی هر دسته نشان می‌دهد.



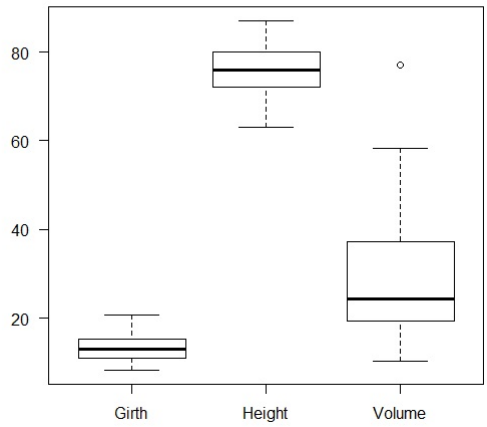
۴.۱.۲ نمودار جعبه‌ای:

یکی از نمودارهای مشهور در آمار boxplot است. این نمودار به خوبی توزیع داده‌ها حول میانه را نشان می‌دهد و همچنین داده‌های چوله و داده‌های پرت را به دست می‌دهد. برای توضیح عناصر این نمودار به شکل زیر توجه کنید.



Q_1, Q_3 به ترتیب چارک اول و سوم است. نقاطی که بیرون از خط عمودی در طرفین باکس قرار دارند به نقاط پرت مشهورند.

برای رسم نمودارجعبه‌ای از دستور boxplot() استفاده میکنیم که دارای آرگومان منطقی horizontal است و در حالت پیش فرض ،نادرست(FALSE) میباشد. و شکل نمودار جعبه‌ای افقی به حالت عمودی درمی‌آید. اکنون به مثالی توجه کنید که دارای چندگروه داده باشد. برای انجام این کار از داده trees که جزء داده‌های R است استفاده می‌گردد.



اما چگونه می‌توان مشخصات عددی boxplot را به دست آورد. در تابع گرافیکی آرگومان منطقی plot وجود دارد که مقدار پیش‌فرض آن درست(TRUE) است. اگر مقدار این آرگومان نادرست شود، آنگاه مقادیر عددی ظاهر می‌شود.

```
> boxplot(trees,plot=FALSE)
$stats
[,1] [,2] [,3]
[1,]  8.30   63 10.2
[2,] 11.05   72 19.4
[3,] 12.90   76 24.2
[4,] 15.25   80 37.3
[5,] 20.60   87 58.3

$n
[1] 31 31 31
```

```
$conf
[,1]     [,2]    [,3]
[1,] 11.70814 73.72979 19.1204
[2,] 14.09186 78.27021 29.2796

$out
[1] 77

$group
[1] 3

$names
[1] "Girth"  "Height" "Volume"
```

همانطور که ملاحظه می‌شود مقادیر فوق به‌صورت یک لیست هستند که مولفه‌های آن به شرح زیر است.

stats: ماتریسی است که پنج عنصر هر گروه که شامل whisker پایین، چارک اول، میانه، چارک سوم و whisker بالا است را نشان می‌دهد.

n : برداری است که تعداد عناصر هر گروه را مشخص می‌کند.

conf: ماتریسی است که مقادیر پایین و بالا notch هر گروه را نشان می‌دهد.

out: مقادیر خارج از whiskerها را نشان می‌دهد که در واقع داده‌های پرت هستند.

group: تعداد گروه‌های یک داده را نشان می‌دهد.

names: نام هر گروه یک داده را نشان می‌دهد.

۲.۲ خلاصه کردن داده‌های رسته‌ای:

اغلب داده‌های رسته‌ای را در قالب جداول نمایش می‌دهند. این داده‌ها را می‌توان با نمودار میله‌ای یا دایره‌ای نیز نمایش داد.

۱.۲.۲ استفاده از جداول:

با استفاده از تابع table() می‌توان داده‌های رسته را به صورت جدولی مشاهده کرد. فرم ساده این دستور به صورت table(x) است که در آن x متغیر رسته‌ای است.

مثال: در یک تحقیق که درمورد استعمال سیگار از افراد سوال شده است، داده‌های زیر بدست آمده است:

بله، خیر، خیر، بله، بله

داده‌ها را با دستور c() وارد و با دستور table() خلاصه می‌کنیم:

```
> x=c("Yes","No","No","Yes","Yes")
>  table(x)
x
No Yes
2   3
```

درواقع دستور table() فراوانی هر دسته را مشخص می‌کند.

۲.۲.۲ نمودار میله‌ای:

ارتفاع میله‌ها در یک نمودار میله‌ای، مقادیر متناظر با فراوانی رسته‌ها را مشخص می‌کند، به طوری که بلندترین میله نشانگر رسته است که بیشترین مقدار(فراوانی) را دارد. نمودار میله‌ای را می‌توان براساس فراوانی مطلق و نیز فراوانی نسبی (درصد) رسم کرد.

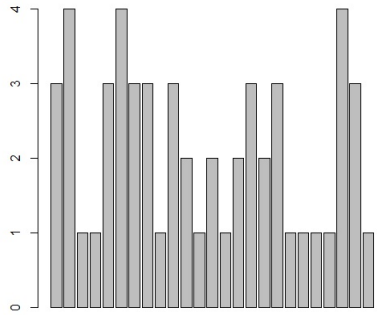
فرض کنید در مورد نحوه آمدن به دانشگاه از ۲۵ دانشجو سوال کرده‌ایم. رسته ها عبارتند از:

۱: اتوبوس، ۲:تاکسی، ۳:اتومبیل شخصی، ۴:پیاده. داده‌های زیر بدست آمده است:

۳ ، ۴ ، ۱ ، ۱ ، ۳ ، ۴ ، ۳ ، ۳ ، ۱ ، ۳ ، ۲ ، ۲ ، ۱، ۲ ، ۱ ، ۱ ، ۲ ، ۳ ، ۲ ، ۳ ، ۱ ، ۱ ،
۱ ، ۱ ، ۴ ، ۳ ، ۰ ۱

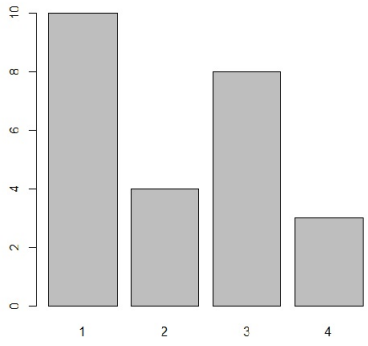
حال نمودار میله‌ای مربوط به فراوانی مطلق و فراوانی نسبی داده‌ها را رسم می‌کنیم.

```
> uni=scan()
1: 3 4 1 1 3 4 3 3 1 3 2 1 2 1 2 3 2 3 1 1 1 1 4 3 1
26:
Read 25 items
> barplot(uni)
```



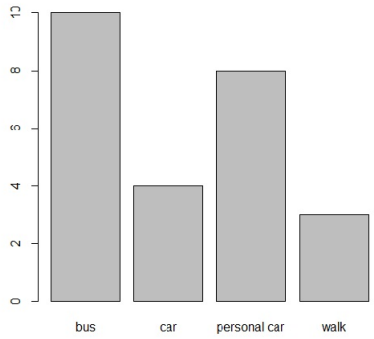
این نمودار صحیح نیست زیرا ما ۴ تا رده داشتیم نه ۲۵ تا.

```
> tuni=table(uni)
> barplot(tuni)
```



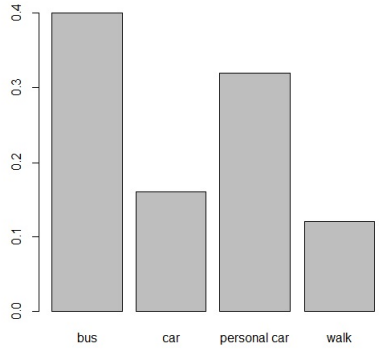
برای اینکه به جای شماره رسته‌ها نام هر رسته را روی نمودار مشاهده کنیم با تابع names() اسامی رسته‌ها را مشخص می‌کنیم.

```
> names(tuni)=c("bus","car","personal car","walk")
> barplot(tuni)
```



برای رسم نمودار میله‌ای بر حسب فراوانی نسبی(درصد) داده‌ها از دستور زیر استفاده می‌کنیم.

```
barplot(tuni/length(uni))
```



دستور زیر جدول فراوانی نسبی داده‌ها را ایجاد می‌کند.

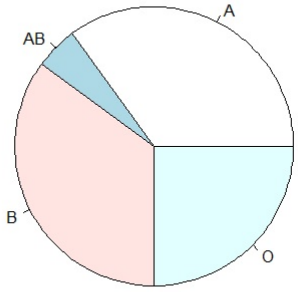
```
> tuni/length(uni)
bus          car personal car         walk
0.40         0.16         0.32         0.12
```

۳.۲.۲ نمودار دایره‌ای:

داده‌های رسته‌ای را می‌توان با نمودار دایره‌ای نیز نشان داد. برای رسم یک نمودار دایره‌ای از تابع pie() استفاده می‌شود.

مثال: نمونه‌ای ۲۰ تایی از گروه خون مراجعه کنندگان به یک درمانگاه در دست داریم. با دستورات زیر نمودار دایره‌ای گروه خونی این افراد رسم می‌شود.

```
> blood=sample(c("A" , "B" , "AB" , "O"),20,replace=T)
> blood
[1] "B"  "A"  "O"  "A"  "O"  "B"  "A"  "O"  "B"  "A"  "B"  "B"  "B"  "A"  "AB"
 [16] "O"  "B"  "O"  "A"  "A"
> tblood=table(blood)
> tblood
blood
A AB  B  O
7  1  7  5
> blood
[1] "B"  "A"  "O"  "A"  "O"  "B"  "A"  "O"  "B"  "A"  "B"  "B"  "B"  "A"  "AB"
 [16] "O"  "B"  "O"  "A"  "A"
> pie(tblood)
```



در این مثال داده‌های خود را به صورت یک نمونه تصادفی در نظر گرفتیم. برای تولید نمونه تصادفی از تابع sample() استفاده می‌شود. پارامتر اول این تابع، جامعه‌ای است که می‌خواهیم از آن نمونه بگیریم که در این مثال بردار اسامی ۴ نوع گروه خونی موجود است. پارامتر دوم حجم نمونه مورد نیاز است. شناسه replace را برابر TRUE قرار می‌دهیم تا نمونه‌گیری تصادفی ساده با جایگذاری انجام شود. مانند دستور رسم نمودار میله‌ای در اینجا نیز نمودار را برای جدول فراوانی رسم می‌کنیم.

۳ ترکیبات:

با توجه به فرمول ترکیبات

$$C(n,r) = \frac{n!}{r!(n-r)!}$$

این مقدار در R با تابع chooes() محاسبه میشود. **مثال:**

```
> choose(5,4)
[1] 5
> choose(5,3)
[1] 10
> choose(5,2)
[1] 10
> choose(5,1)
[1] 5
```