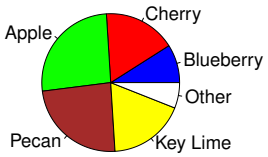
20.3 Categorical data

A categorical data value can take on one of a limited number of potential values. A data set that consists of categorical data values is known as a *categorical data set*. Although statisticians generally prefer more sophisticated presentation tools for categorical data sets, one of the simplest graphics that can be generated is a *pie chart*, which can be created with the `pie` function.

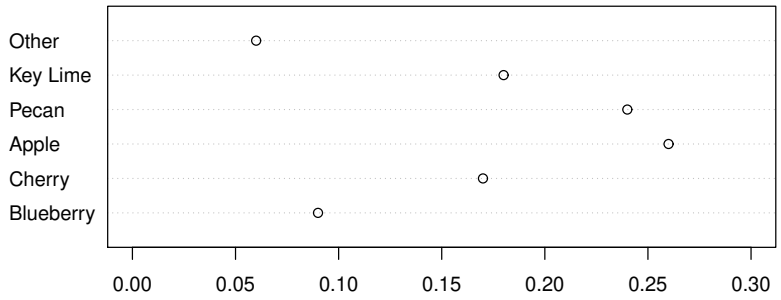
```
> pie.sales = c(0.09, 0.17, 0.26, 0.24, 0.18, 0.06)
> names(pie.sales) = c("Blueberry", "Cherry", "Apple", "Pecan", "Key Lime",
+                      "Other")
> pie.sales                                  # display proportion of pie sales
Blueberry    Cherry     Apple     Pecan  Key Lime     Other
     0.09      0.17      0.26      0.24      0.18      0.06
> colors = c("blue", "red", "green", "brown", "yellow", "white")
> pie(pie.sales, col = colors)               # draw pie chart
```



The `names` function appears on the left-hand side of the assignment operator in order to set the names for `pie.sales`. As pointed out earlier, the `names` function can be used to both get (retrieve) and set (assign) names.

The preferred mode of displaying categorical data is with a horizontal dot chart or bar chart. The pie sales data can be displayed in a dot chart with the command

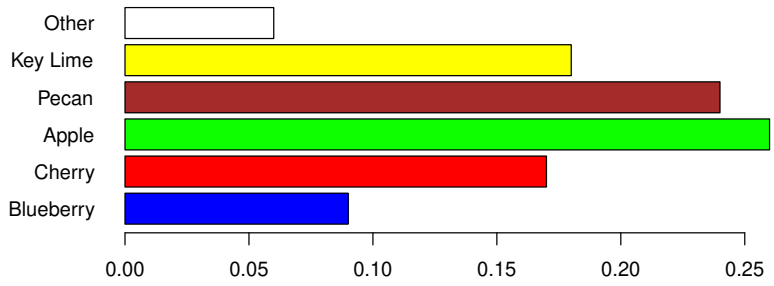
```
> dotchart(pie.sales, xlim = c(0, 0.3))   # force axis to include zero
```



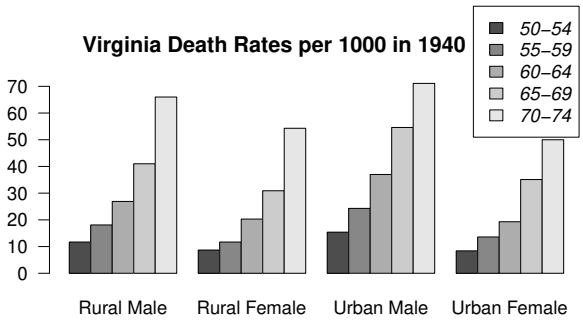
The `xlim` (*x* limit) argument forces the horizontal axis to run from 0 to 0.3. Similarly, a horizontal bar chart for the pie sales data can be graphed with the command

```
> barplot(pie.sales, col = colors, horiz = TRUE, las = 1)
```

The `VADeaths` data set contains the annual death rates per 1000 persons in Virginia in 1940, broken down by age, gender, and urban/rural dwelling status. A bar chart for this more complicated data set can be generated with the R command




```
> barplot(VADeaths, beside = TRUE, legend = TRUE,
+         main = "Virginia Death Rates per 1000 in 1940")
```



Lots of information has been packed into this graphic, which is generated with just a single R command. Death rates are consistently increasing with age; death rates are higher for men than for women; there might be a slight advantage for older women who live in the city. All of these conclusions, however, would have to be verified by conducting an appropriate statistical test.

This ends the discussion of graphics for categorical data. The last section concerns high-level graphics functions for time-series data.

20.4 Time-series data

Many data sets consist of observations that are collected over time. One of the primary graphical exploratory data analysis tools for a time series is simply a plot of the time-series values over time. The `plot.ts` function can be used to plot a time series, such as the built-in `AirPassengers` data set, which contains the monthly number of airline passengers (in thousands) between 1949 and 1960:

```
> plot.ts(AirPassengers)                # international airline passengers
```

The time-series plot reveals that there is a growth (perhaps nonlinear) in the number of passengers over time, a pronounced cyclic variation that appears to be seasonal, and an amplitude to the seasonal component that increases with time. A branch of statistics known as *time-series analysis* considers the development of probability models for complex time series such as this one. A second time series named `presidents` contains the quarterly U.S. presidential approval ratings from 1945 to 1974. This time series also contains several `NA` values associated with times in which such polls were not taken.