Base graphics

Jumping Rivers

Base Graphics

The standard plot

The standard plotting command in R is `plot()`. It takes two coordinate arguments: `plot(x, y)`.[1] For example,

```
data(movies, package = "ggplot2movies")
sub_movies = movies[1:500, ]
## Plot first 500 movies
plot(sub_movies$length, sub_movies$rating)
```

gives figure @ref(fig:moviesscatter) a. The basic `plot` command has multiple arguments:

- `main`, `xlab` and `ylab` control the axis labels.
- `xlim` and `ylim` control the axis ranges.
- `lwd` controls the thickness of the lines.
- `lty` specifies the line type, e.g. dotted lines.
- `type = 'p'` - draws points.
- `type = 'l'` - draws lines.
- `type = 'b'` - draws lines and points.
- `type = 'o'` - draws lines and points over-plotted.
- `log = 'y'` - use the log scale on the `y` (or `x`) axis.

Using the default settings is fine for investigating data, but we often have to set limits and colours explicitly:

```
plot(sub_movies$length, sub_movies$rating, ylab = "Rating",
    xlab = "Movie Length", ylim = c(0, 10), xlim = c(0,
        250), pch = 21, bg = sub_movies$Action +
        1, panel.first = grid(), cex = 0.7)
```

to get figure @ref(fig:moviesscatter) b. To add lines, we use the `lines()` function. Horizontal and vertical lines can be added using the convenience function `abline()`, viz.

```
## Add a horizontal line at the origin
abline(h = 0)
## Vertical line
abline(v = 1)
```

and plain text is added using the `text()` function:

```
## Adds '(b)' to (0, 9.3) on the plot
text(0, 9.3, "(b)")
```

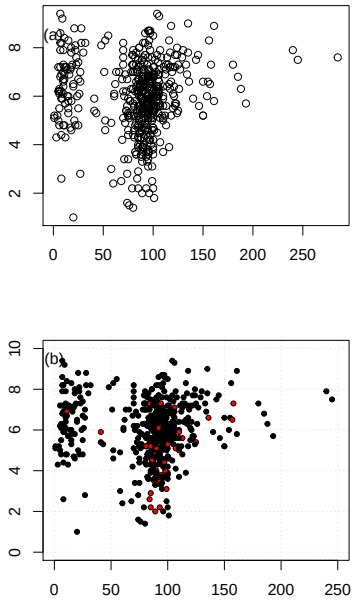


Figure 1: A scatter plot of movie length against budget.

[1] Actually, you can just type `plot(y)` and it will generate a default for the x-axis

Bar charts

The most useful way to display qualitative data is usually with a bar chart. The length of each bar is equal to the frequency of the corresponding value of the variable in the sample of data. Note that the widths of the bars should be equal to avoid giving a false impression.

Figure @ref(fig:moviesbarchart) shows the breakdown in MPAA ratings for the movie data set. To create a bar chart in R we use the following commands:

```
barplot(table(movies$mpaa), xlab = "MPAA Rating",
    ylab = "Frequency", border = "black")
```

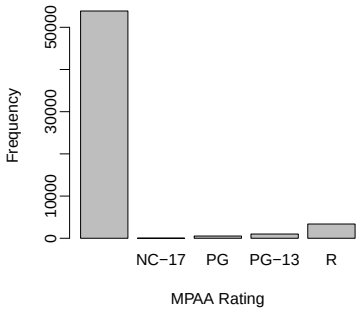


Figure 2: Barchart of the MPAA ratings for films.

Histograms

```
hist(movies$rating, col = "grey", main = "User rating",
    freq = TRUE, xlim = c(1, 10), xlab = "Rating")
hist(movies$budget, col = "grey", main = "Film budget",
    freq = FALSE, xlab = "Budget")
```

To represent the distribution of a sample of values of a continuous variable we can use a histogram. The range of values of the variable is divided into intervals, known as *classes*, and the frequencies in classes are represented by columns. As the variable is continuous, there are no gaps between neighbouring columns, unlike a bar chart. Note also that, strictly speaking, it is the *area* of the column which is equal to the frequency, not the height. The reason for this is that columns need not be of the same width. Computer software tends to use columns of the same width. However this default can be overridden in R if you really want to. Figure @ref(fig:movieshist) shows histograms of film ratings and budgets.

To generate Figure @ref(fig:movieshist) in R we use the following commands:

```
hist(movies$rating, col = "grey", main = "User rating",
    freq = TRUE, xlim = c(1, 10), xlab = "Rating")
hist(movies$budget, col = "grey", main = "Film budget",
    freq = FALSE, xlab = "Budget")
```

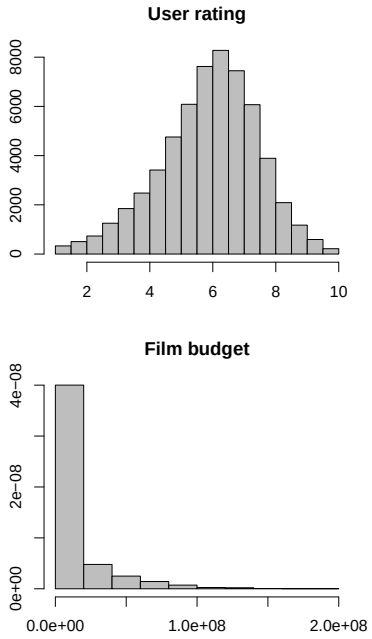
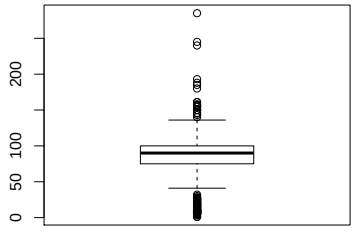
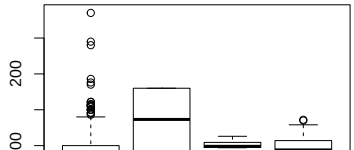


Figure 3: Histograms of the film ratings and budgets.

Box and whisker plots

A box and whisker plot, sometimes called simply a boxplot, is another way to represent continuous data. This kind of plot is particularly useful for comparing two or more groups, by placing the box-plots side-by side. Figure @ref(fig:moviesbox1) shows box-plots of film length for different categories.





The central bar in the "box" is the sample *median*. The top and bottom of the box represent the upper and lower sample *quartiles*. Just as the median represents the 50% point of the data, the lower and upper quartiles represent the 25% and 75% points respectively.

The lower whisker is drawn from the lower end of the box to the smallest value that is no smaller than 1.5IQR below the lower quartile. Similarly, the upper whisker is drawn from the middle of the upper end of the box to the largest value that is no larger than 1.5IQR above the upper quartile. Points outside the whiskers are classified as outliers.

To draw box and whisker plots in R, we use the `boxplot()` function. So to construct a boxplot of the movie lengths @ref(fig:moviesbox1) a, we have

```
boxplot(sub_movies$length, ylab = "Film length")
```

The `boxplot()` function also accepts formula notation[2]. For example,

[2] Formula notation in R is of the form `response_variable ~ explanatory_variable`. It is a common construct for specifying model formulae in statistical routines with R but can also be used for Base graphics.

```
## Formula pulling out columns from the data
## individually
boxplot(sub_movies$length ~ sub_movies$mpaa)
## Formula and data arguments, looks a bit
## cleaner
boxplot(length ~ mpaa, data = sub_movies)
```

constructs box and whisker plots for each MPAA rating while

```
boxplot(length ~ Action + Romance, data = sub_movies)
```

separates length conditional whether it is a romantic and/or an action movie.