Part III

Data Summaries

3.1.1 Measures of average: The sample mean

One of the most important and widely used measures of average is the mean:

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

So if our data set was $\{0, 3, 2, 0\}$, then $n = 4$.Hence,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{0+3+2+0}{4} = 1.25.$$

The mean can be distorted if our sample contains outliers.

3.1.1 Measures of average: The sample mean

One of the most important and widely used measures of average is the mean:

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

So if our data set was $\{0, 3, 2, 0\}$, then $n = 4$. Hence,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{0+3+2+0}{4} = 1.25.$$

The mean can be distorted if our sample contains outliers.

3.1.1 Measures of average: The sample mean

One of the most important and widely used measures of average is the mean:

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

So if our data set was $\{0, 3, 2, 0\}$, then $n = 4$.Hence,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{0+3+2+0}{4} = 1.25.$$

The mean can be distorted if our sample contains outliers.

3.1.1 Measures of average: The sample mean

One of the most important and widely used measures of average is the mean:

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

So if our data set was $\{0, 3, 2, 0\}$, then $n = 4$.Hence,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{0 + 3 + 2 + 0}{4} = 1.25.$$

The mean can be distorted if our sample contains outliers.

3.1.1 Measures of average: The sample mean

One of the most important and widely used measures of average is the mean:

$$\begin{aligned}\bar{x} &= \frac{x_1 + x_2 + \ldots + x_n}{n} \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i \ .\end{aligned}$$

So if our data set was $\{0, 3, 2, 0\}$, then $n = 4$.Hence,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{0 + 3 + 2 + 0}{4} = 1.25.$$

The mean can be distorted if our sample contains outliers.

3.1.1 Measures of average: The sample median

- The sample median is the middle observation when the data is ranked in increasing order.
- Denote the ranked observations as $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$.
- The sample median is defined as:

$$\text{Sample median} = \begin{cases} x_{(n+1)/2}, & n \text{ odd;} \\ \frac{1}{2}x_{(n/2)} + \frac{1}{2}x_{(n/2+1)}, & n \text{ even} . \end{cases}$$

- The median is more robust than the sample mean, but has less useful mathematical properties.

3.1.1 Measures of average: The sample median

- The sample median is the middle observation when the data is ranked in increasing order.
- Denote the ranked observations as $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$.
- The sample median is defined as:

$$\text{Sample median} = \begin{cases} x_{(n+1)/2}, & n \text{ odd;} \\ \frac{1}{2}x_{(n/2)} + \frac{1}{2}x_{(n/2+1)}, & n \text{ even} \end{cases}.$$

- The median is more robust than the sample mean, but has less useful mathematical properties.

3.1.1 Measures of average: The sample median

- The sample median is the middle observation when the data is ranked in increasing order.
- Denote the ranked observations as $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$.
- The sample median is defined as:

$$\text{Sample median} = \begin{cases} x_{(n+1)/2}, & n \text{ odd;} \\ \frac{1}{2}x_{(n/2)} + \frac{1}{2}x_{(n/2+1)}, & n \text{ even} . \end{cases}$$

- The median is more robust than the sample mean, but has less useful mathematical properties.

3.1.1 Measures of average: The sample median

- For our simple data set $\{0, 3, 2, 0\}$, to calculate the median we re-order it to: $\{0, 0, 2, 3\}$.
- Then take the average of the middle two observations to get 1.

3.1.1 Measures of average: The sample mode

- The mode is the value which occurs with the greatest frequency.
- Only makes sense to calculate or use it with discrete data.

3.1.2 Measures of spread

As well as knowing the location statistics of a data set, we also need to know how variable or ‘spread-out’ our data are.

3.1.2 Measures of spread: The range

The range is easy to calculate.

It is simply the largest minus the smallest:

$$\text{Range} = x_{(n)} - x_{(1)} .$$

- So for our data set of $\{0, 3, 2, 0\}$, the range is $3 - 0 = 3$.
- It is very useful for data checking purposes, but in general it's not very robust.

3.1.2 Measures of spread: The range

The range is easy to calculate.

It is simply the largest minus the smallest:

$$\text{Range } = x_{(n)} - x_{(1)} \ .$$

- So for our data set of $\{0, 3, 2, 0\}$, the range is $3 - 0 = 3$.
- It is very useful for data checking purposes, but in general it's not very robust.

3.1.2 Measures of spread: The range

The range is easy to calculate.

It is simply the largest minus the smallest:

$$\text{Range } = x_{(n)} - x_{(1)} \ .$$

- So for our data set of $\{0, 3, 2, 0\}$, the range is $3 - 0 = 3$.
- It is very useful for data checking purposes, but in general it's not very robust.

3.1.2 Measures of spread: Variance/standard deviation

The sample variance, s^2, is defined as

$$\begin{aligned} s^2 &= \\ &= \end{aligned}$$

The second formula is easier for calculations.

3.1.2 Measures of spread: Variance/standard deviation

The sample variance, s^2, is defined as

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

The second formula is easier for calculations.

3.1.2 Measures of spread: Variance/standard deviation

The sample variance, s^2, is defined as

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= \frac{1}{(n-1)}\left\{\left(\sum_{i=1}^{n} x_i^2\right) - n\bar{x}^2\right\} . \end{aligned}$$

The second formula is easier for calculations.

3.1.2 Measures of spread: Variance/standard deviation

So for our data set $\{0, 3, 2, 0\}$, we have

$$\sum_{i=1}^{4} x_i^2 = 0^2 + 3^2 + 2^2 + 0^2 = 13$$

So

$$s^2 = \frac{1}{n-1}\left\{\left(\sum_{i=1}^{n} x_i^2\right) - n\bar{x}^2\right\} = \frac{1}{3}\left(13 - 4 \times 1.25^2\right) = 2.25.$$

The sample standard deviation, s, is the square root of the sample variance, i.e. for our toy example $s = \sqrt{2.25} = 1.5$.

The standard deviation is often preferred as it is in the original units of the data.

3.1.2 Measures of spread: Variance/standard deviation

So for our data set $\{0, 3, 2, 0\}$, we have

$$\sum_{i=1}^{4} x_i^2 = 0^2 + 3^2 + 2^2 + 0^2 = 13$$

So

$$s^2 = \frac{1}{n-1}\left\{\left(\sum_{i=1}^{n} x_i^2\right) - n\bar{x}^2\right\} = \frac{1}{3}\left(13 - 4 \times 1.25^2\right) = 2.25.$$

The sample standard deviation, s, is the square root of the sample variance, i.e. for our toy example $s = \sqrt{2.25} = 1.5$.

The standard deviation is often preferred as it is in the original units of the data.

3.1.2 Measures of spread: Variance/standard deviation

So for our data set $\{0, 3, 2, 0\}$, we have

$$\sum_{i=1}^{4} x_i^2 = 0^2 + 3^2 + 2^2 + 0^2 = 13$$

So

$$s^2 = \frac{1}{n-1}\left\{\left(\sum_{i=1}^{n} x_i^2\right) - n\bar{x}^2\right\} = \frac{1}{3}\left(13 - 4 \times 1.25^2\right) = 2.25.$$

The sample standard deviation, s, is the square root of the sample variance, i.e. for our toy example $s = \sqrt{2.25} = 1.5$.

The standard deviation is often preferred as it is in the original units of the data.

3.1.2 Measures of spread: Variance/standard deviation

So for our data set $\{0, 3, 2, 0\}$, we have

$$\sum_{i=1}^{4} x_i^2 = 0^2 + 3^2 + 2^2 + 0^2 = 13$$

So

$$s^2 = \frac{1}{n-1}\left\{\left(\sum_{i=1}^{n} x_i^2\right) - n\bar{x}^2\right\} = \frac{1}{3}\left(13 - 4 \times 1.25^2\right) = 2.25.$$

The sample standard deviation, s, is the square root of the sample variance, i.e. for our toy example $s = \sqrt{2.25} = 1.5$.

The standard deviation is often preferred as it is in the original units of the data.

3.1.2 Measures of spread: Variance/standard deviation

So for our data set $\{0, 3, 2, 0\}$, we have

$$\sum_{i=1}^{4} x_i^2 = 0^2 + 3^2 + 2^2 + 0^2 = 13$$

So

$$s^2 = \frac{1}{n-1}\left\{\left(\sum_{i=1}^{n} x_i^2\right) - n\bar{x}^2\right\} = \frac{1}{3}\left(13 - 4 \times 1.25^2\right) = 2.25.$$

The sample standard deviation, s, is the square root of the sample variance, i.e. for our toy example $s = \sqrt{2.25} = 1.5$.

The standard deviation is often preferred as it is in the original units of the data.

3.1.2 Measures of spread: Variance/standard deviation

So for our data set $\{0, 3, 2, 0\}$, we have

$$\sum_{i=1}^{4} x_i^2 = 0^2 + 3^2 + 2^2 + 0^2 = 13$$

So

$$s^2 = \frac{1}{n-1}\left\{\left(\sum_{i=1}^{n} x_i^2\right) - n\bar{x}^2\right\} = \frac{1}{3}\left(13 - 4 \times 1.25^2\right) = 2.25.$$

The sample standard deviation, s, is the square root of the sample variance, i.e. for our toy example $s = \sqrt{2.25} = 1.5$.

The standard deviation is often preferred as it is in the original units of the data.

3.1.2 Measures of spread: Quartiles and the IQR

The upper and lower quartiles are defined as follows:

- $Q1$: Lower quartile $= (n+1)/4^{\text{th}}$ smallest observation.
- $Q3$: Upper quartile $= 3(n+1)/4^{\text{th}}$ smallest observation.

We can linearly interpolate between adjacent observations if necessary.

The interquartile range is the difference between the third and first quartile, i.e.

$$IQR = Q3 - Q1.$$

3.1.2 Measures of spread: Quartiles and the IQR

The upper and lower quartiles are defined as follows:

- $Q1$: Lower quartile $= (n+1)/4^{\text{th}}$ smallest observation.
- $Q3$: Upper quartile $= 3(n+1)/4^{\text{th}}$ smallest observation.

We can linearly interpolate between adjacent observations if necessary.

The interquartile range is the difference between the third and first quartile, i.e.

$$IQR = Q3 - Q1.$$

3.1.3 Example

The number of earth tremours recorded for five randomly chosen locations in Iceland, close to the *Mid-Atlantic ridge*, is recorded below:

Location	1	2	3	4	5
# tremours	7	1	14	13	20

- Calculate the mean and median number of earth tremours for this sample.

  Solution:

  The mean is:

$$\bar{x} = \frac{1}{5}(1 + 7 + 13 + 14 + 20) = 11$$

  The median is the middle value, so we get 13.

3.1.3 Example

The number of earth tremours recorded for five randomly chosen locations in Iceland, close to the *Mid-Atlantic ridge*, is recorded below:

Location	1	2	3	4	5
# tremours	7	1	14	13	20

- Calculate the mean and median number of earth tremours for this sample.

  Solution:

  The mean is:

$$\bar{x} = \frac{1}{5}(1 + 7 + 13 + 14 + 20) = 11$$

  The median is the middle value, so we get 13.

3.1.3 Example

- There was a recording error for location 5, which is closest to the Mid-Atlantic Ridge; this observation *should have* been 200, and not 20. Find new values for the mean and median, and comment.

Solution:

The mean is:

$$\bar{x} = \frac{1}{5}(1 + 7 + 13 + 14 + 200) = 47$$

The median is the middle value, so we still get 13. However, we can clearly see distortion in the mean here.

3.1.3 Example

- There was a recording error for location 5, which is closest to the Mid-Atlantic Ridge; this observation *should have* been 200, and not 20. Find new values for the mean and median, and comment.

  Solution:

  The mean is:

$$\bar{x} = \frac{1}{5}(1 + 7 + 13 + 14 + 200) = 47$$

  The median is the middle value, so we still get 13. However, we can clearly see distortion in the mean here.

3.1.3 Example

- Find the standard deviation and IQR for the earth tremours data, and comment.

Solution:

We have

$$\sum x^2 = 1^2 + 7^2 + \ldots + 200^2 = 40415.$$

Thus,

$$s = \sqrt{\frac{1}{4}\left\{40415 - 5 \times 47^2\right\}} = 85.69.$$

We also have that $Q1$ is in position 1.5 and $Q3$ position 4.5, giving:

$$Q1 = \frac{1+7}{2} = 4 \quad \text{and} \quad Q3 = \frac{14+200}{2} = 107,$$

giving

$$IQR = 107 = 4 = 103.$$

3.1.3 Example

- Find the standard deviation and IQR for the earth tremours data, and comment.

Solution:

We have

$$\sum x^2 = 1^2 + 7^2 + \ldots + 200^2 = 40415.$$

Thus,

$$s = \sqrt{\frac{1}{4}\left\{40415 - 5 \times 47^2\right\}} = 85.69.$$

We also have that $Q1$ is in position 1.5 and $Q3$ position 4.5, giving:

$$Q1 = \frac{1+7}{2} = 4 \quad \text{and} \quad Q3 = \frac{14+200}{2} = 107,$$

giving

$$IQR = 107 = 4 = 103.$$

3.1.3 Example

- Find the standard deviation and IQR for the earth tremours data, and comment.

Solution:

We have

$$\sum x^2 = 1^2 + 7^2 + \ldots + 200^2 = 40415.$$

Thus,

$$s = \sqrt{\frac{1}{4}\left\{40415 - 5 \times 47^2\right\}} = 85.69.$$

We also have that $Q1$ is in position 1.5 and $Q3$ position 4.5, giving:

$$Q1 = \frac{1+7}{2} = 4 \quad \text{and} \quad Q3 = \frac{14+200}{2} = 107,$$

giving

$$IQR = 107 = 4 = 103.$$

3.2 Using R

Command	Comment	Example
`mean`	Calculates the mean of a vector	`mean(x)`
`sd`	Calculates the standard deviation of a vector	`sd(x)`
`var`	Calculates the variance of a vector	`var(x)`
`quantile`	The vector quartiles. Make sure you use `type=6`	`quantile(x, type=6)`
`range`	Calculates the vector range	`range(x)`
`summary`	Calculates the quartiles **However,** it doesn't use `type=6` quartiles.	`summary(x)`

Table: Summary of R commands in this chapter.

Command	Comment	Example
`mean`	Calculates the mean of a vector	`mean(x)`
`sd`	Calculates the standard deviation of a vector	`sd(x)`
`var`	Calculates the variance of a vector	`var(x)`
`quantile`	The vector quartiles. Make sure you use `type=6`	`quantile(x, type=6)`
`range`	Calculates the vector range	`range(x)`
`summary`	Calculates the quartiles **However,** it doesn't use `type=6` quartiles.	`summary(x)`

Table: Summary of R commands in this chapter.

3.3 Graphical summaries

- Graphical displays of data can be very useful in showing the main features of a dataset
- The appropriate form of graph depends on the nature of the variables being displayed and what aspects are to be shown
- However it should always be borne in mind that the object is to provide a clear and truthful representation of the data, not to distort and not to impress with unnecessary “fancy” features

3.3 Graphical summaries

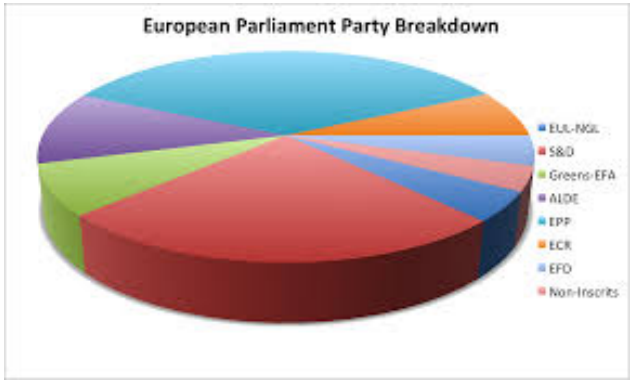
- Graphical displays of data can be very useful in showing the main features of a dataset
- The appropriate form of graph depends on the nature of the variables being displayed and what aspects are to be shown
- However it should always be borne in mind that the object is to provide a clear and truthful representation of the data, not to distort and not to impress with unnecessary “fancy” features

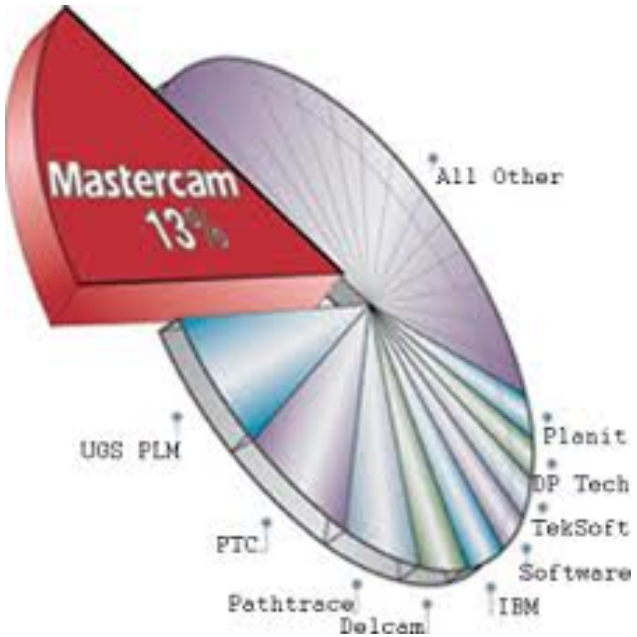
3.3 Graphical summaries

- Graphical displays of data can be very useful in showing the main features of a dataset
- The appropriate form of graph depends on the nature of the variables being displayed and what aspects are to be shown
- However it should always be borne in mind that the object is to provide a clear and truthful representation of the data, not to distort and not to impress with unnecessary “fancy” features

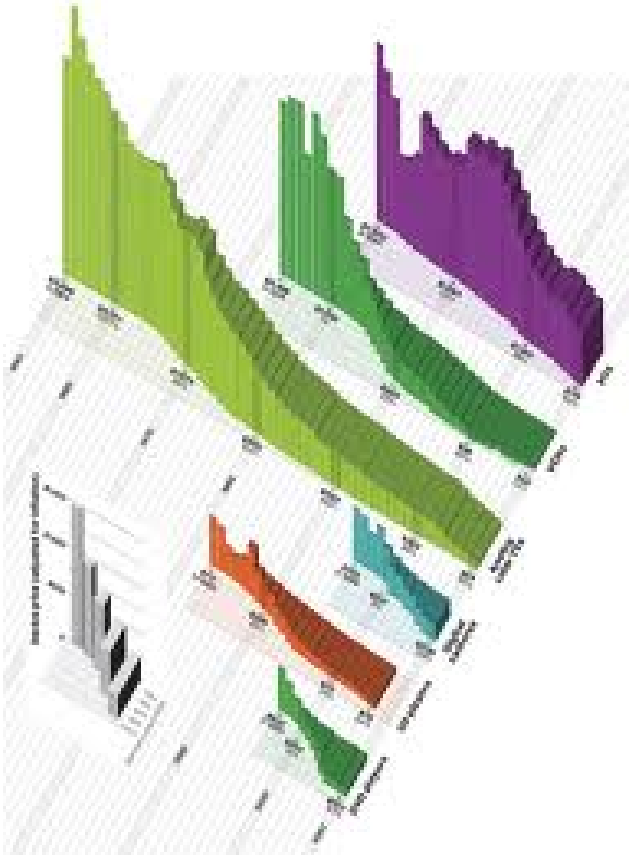
3.3 Graphical summaries

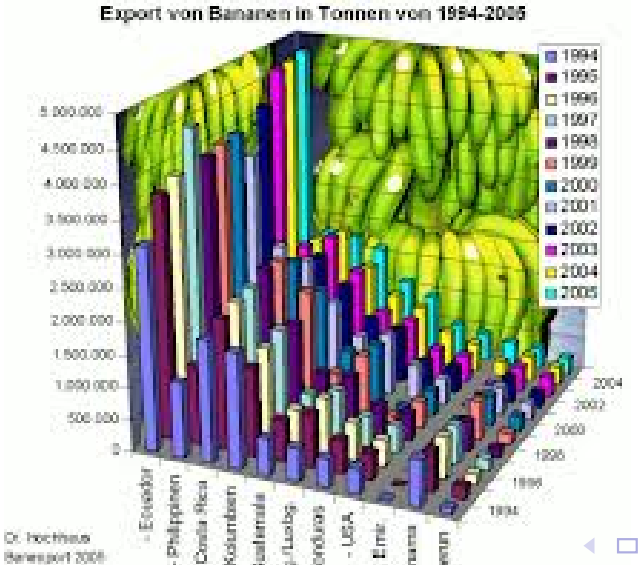
- Graphical displays of data can be very useful in showing the main features of a dataset
- The appropriate form of graph depends on the nature of the variables being displayed and what aspects are to be shown
- However it should always be borne in mind that the object is to provide a clear and truthful representation of the data, not to distort and not to impress with unnecessary “fancy” features



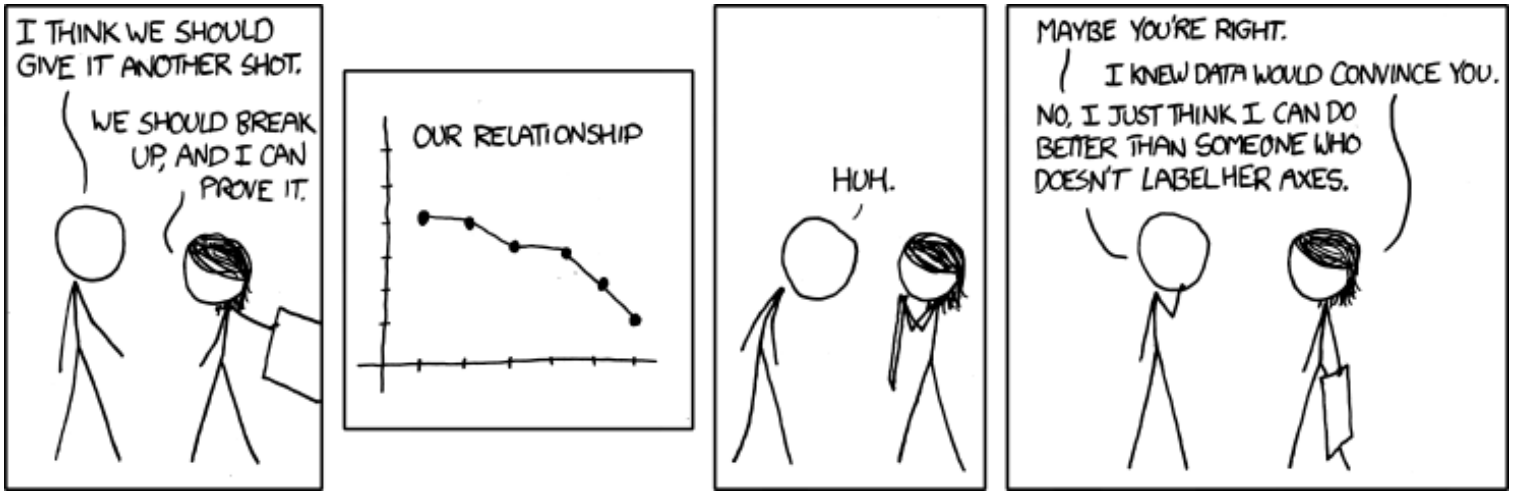


http://www.fusioncharts.com/explore/pie-doughnut-charts



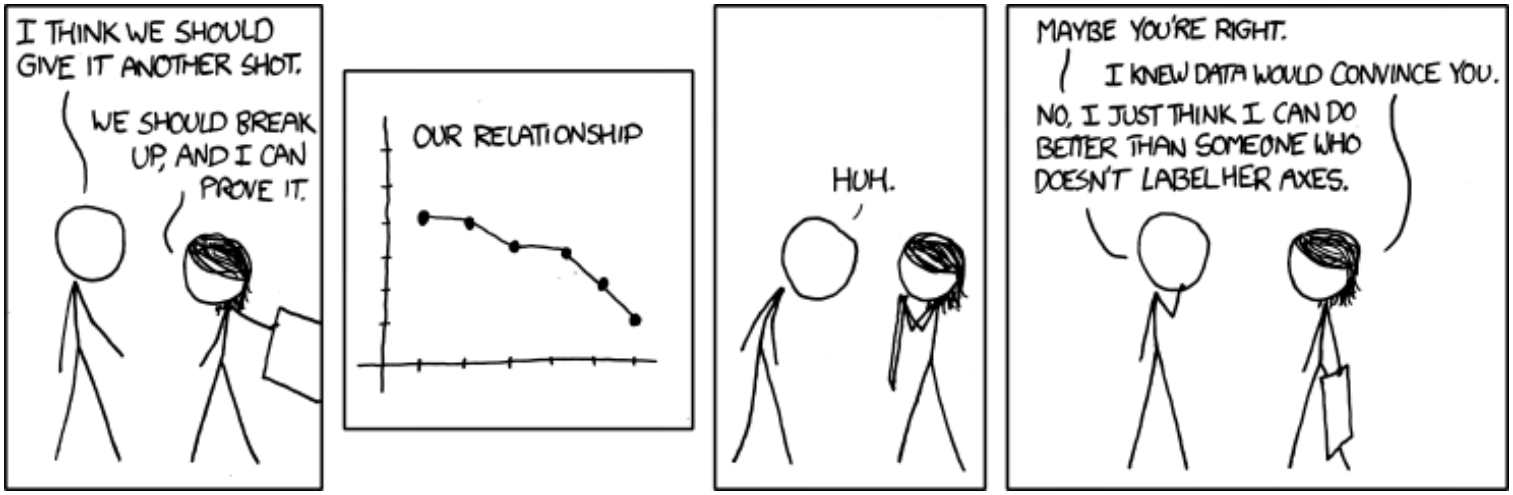

Export von Bananen in Tonnen von 1994-2005
5.000.000
4.500.000
4.000.000
3.500.000
3.000.000
2.500.000
2.000.000
1.500.000
1.000.000
500.000
0
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
Ecuador
Philippinen
Costa Rica
Kolumbien
Guatemala
Honduras
USA
Dr. Hochhaus
Bananenport 2005

3.3 Label your %#?* axes!



Rant over!

3.3 Label your %#?* axes!



Rant over!

3.3.2 Qualitative data: bar charts

- The most useful way to display qualitative data is usually with a bar chart
- The length of each bar is proportional 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

3.3.2 MPAA breakdown

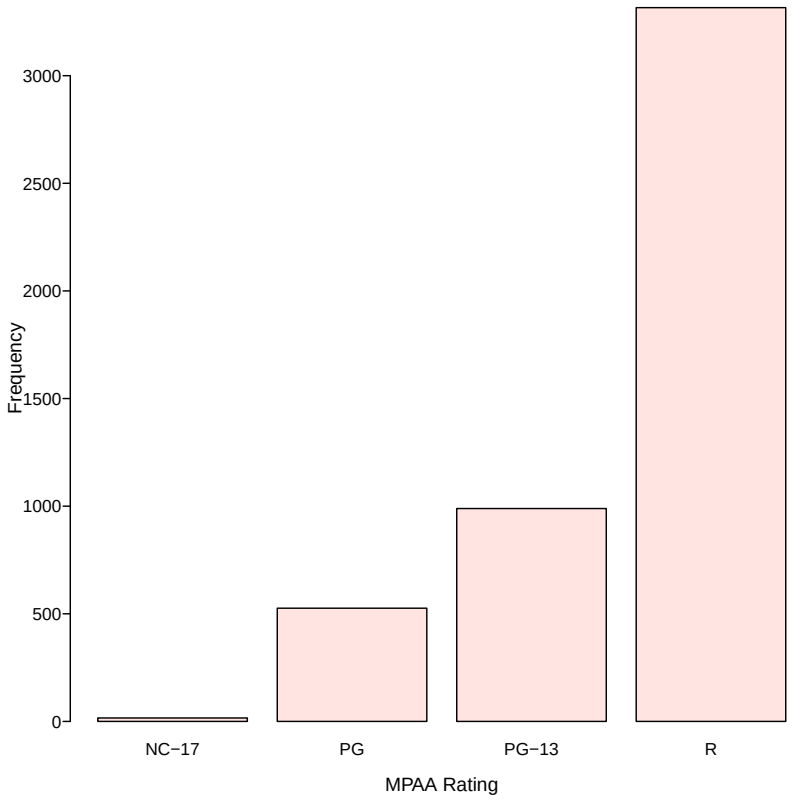


Figure: Barchart of the mpaa ratings for 4847 films.

3.3.2 MPAA breakdown

```
> table(movies$mpaa)
NC-17    PG PG-13     R
   16   526   989  3316
```

Inside the `barplot` function:

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

Remember to load the data first!

```
> library(mas1343)
> data(movies)
```

3.3.2 MPAA breakdown

```
> table(movies$mpaa)
NC-17    PG PG-13     R
   16   526   989  3316
```

Inside the `barplot` function:

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

Remember to load the data first!

```
> library(mas1343)
> data(movies)
```

3.3.2 MPAA breakdown

```
> table(movies$mpaa)
NC-17    PG PG-13     R
   16   526   989  3316
```

Inside the `barplot` function:

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

Remember to load the data first!

```
> library(mas1343)
> data(movies)
```

3.3.3 Histograms

- 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* or bins, 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 proportional 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, meaning that we can just use frequencies on thr y-axis (giving "frequency histograms")
- However this default can be overridden in R if you really want to

3.3.3 Histograms

- 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* or bins, 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 proportional 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, meaning that we can just use frequencies on thr y-axis (giving "frequency histograms")
- However this default can be overridden in R if you really want to

3.3.3 Histograms

- 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* or bins, 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 proportional 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, meaning that we can just use frequencies on thr y-axis (giving "frequency histograms")
- However this default can be overridden in R if you really want to

3.3.3 Histograms

- 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* or bins, 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 proportional 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, meaning that we can just use frequencies on thr y-axis (giving "frequency histograms")
- However this default can be overridden in R if you really want to

3.3.3 Histograms

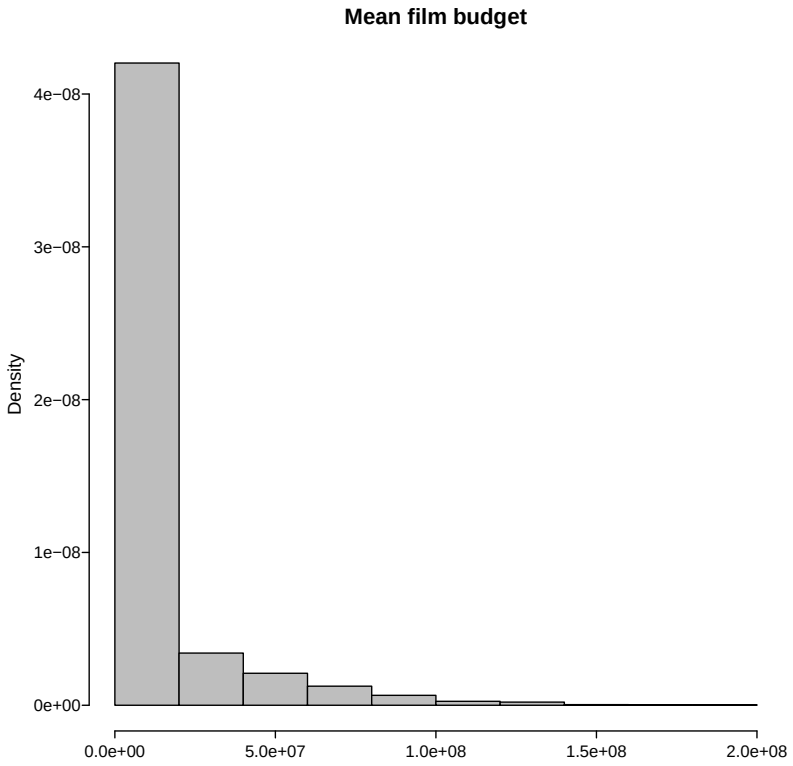


Figure: Histogram of movie budgets.

3.3.3 Histograms

- Figure 3.2 shows a histogram of the movie budgets
- When dealing with *relative* frequency (proportions, often termed "density histograms"), we can easily work out the height of each bin using this formula:

$$\text{Height} = \frac{\text{frequency}}{n \times \text{ Bin-width}}$$

- When the y-axis is labelled with density or relative frequencies, the area under the histogram is one
- Bin widths should be chosen so that you get a good idea of the distribution of the data, without being swamped by random variation

3.3.3 Histograms

- Figure 3.2 shows a histogram of the movie budgets
- When dealing with *relative* frequency (proportions, often termed "density histograms"), we can easily work out the height of each bin using this formula:

$$\text{Height} = \frac{\text{frequency}}{n \times \text{ Bin-width}}$$

- When the y-axis is labelled with density or relative frequencies, the area under the histogram is one
- Bin widths should be chosen so that you get a good idea of the distribution of the data, without being swamped by random variation

3.3.3 Histograms

To generate Figure 3.2 in R we use the following commands:

```
> hist(movies$Budget, col="grey",
+      main="Mean film budget", freq=FALSE,
+      xlab="Budget ($)")
```

3.3.3 How many bins should we have?

First we will define the notation we will use:

- n: the sample size;
- k: the number of bins in the histogram;
- h: the bin-width.

Then the number of bins we will use to construct a histogram is:

$$k = \left\lceil \frac{\max(x) - \min(x)}{h} \right\rceil \tag{3.1}$$

where $\lceil \cdot \rceil$ is the ceiling function.

Table 3.2: Sturges' rule

- $k_{ST} = \lceil \log_2 n + 1 \rceil$
- Default
- Tends not to be very good for $n > 30$.

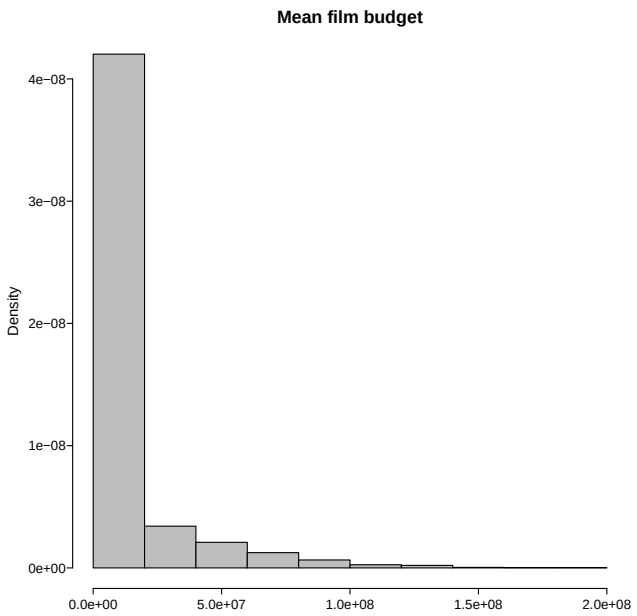


Table 3.2: Scotts' rule

- $h_{SC} = 3.49 \times s \times n^{-1/3}$
- breaks = "Scott"

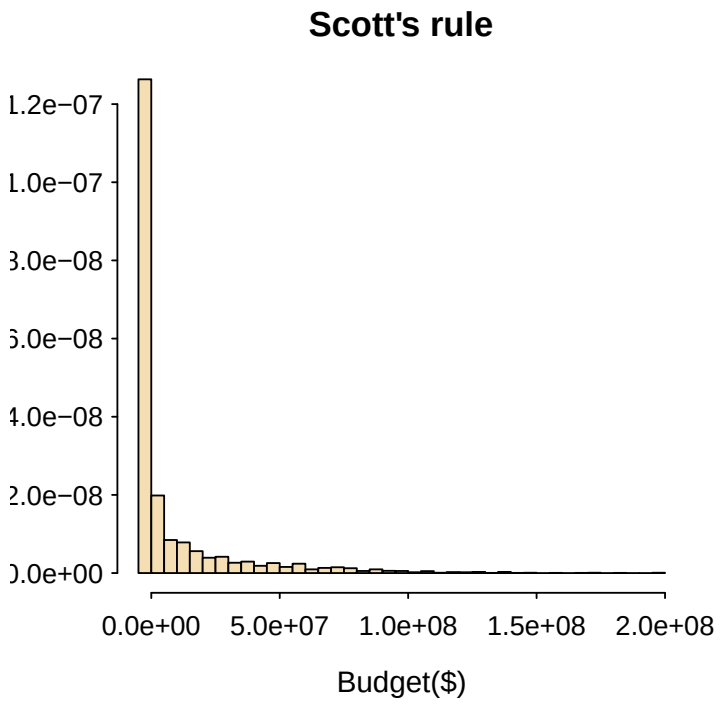
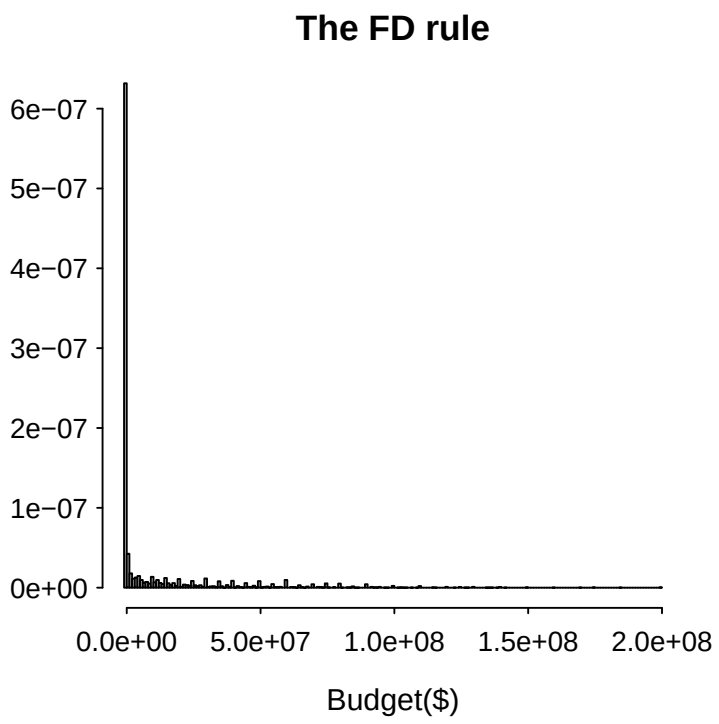


Table 3.2: Freedman-Diaconis' rule

- $h_{FR} = 2 \times IQR(x) \times n^{-1/3}$
- breaks = "FD"
- When the distribution is symmetric, this is very similar to Scott's rule.



3.3.4 Box and whisker plots

A box and whisker plot, often referred to simply as 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 boxplots side-by-side.

Figures 3.4 and 3.5 shows boxplots of film length for different categories of film

- Central bar in the box: median
- Top/bottom of the box: Q3/Q1
- Whisker: smallest value within $1.5IQR$ of $Q1$ and largest value within $1.5IQR$ of $Q3$
- outliers: asterisks - points outside the whiskers

3.3.4 Box and whisker plots

A box and whisker plot, often referred to simply as 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 boxplots side-by-side.

Figures 3.4 and 3.5 shows boxplots of film length for different categories of film

- Central bar in the box: median
- Top/bottom of the box: Q3/Q1
- Whisker: smallest value within $1.5IQR$ of $Q1$ and largest value within $1.5IQR$ of $Q3$
- outliers: asterisks - points outside the whiskers

3.3.4 Box and whisker plots

A box and whisker plot, often referred to simply as 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 boxplots side-by-side.

Figures 3.4 and 3.5 shows boxplots of film length for different categories of film

- Central bar in the box: median
- Top/bottom of the box: Q3/Q1
- Whisker: smallest value within $1.5IQR$ of $Q1$ and largest value within $1.5IQR$ of $Q3$
- outliers: asterisks - points outside the whiskers

3.3.4 Box and whisker plots

A box and whisker plot, often referred to simply as 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 boxplots side-by-side.

Figures 3.4 and 3.5 shows boxplots of film length for different categories of film

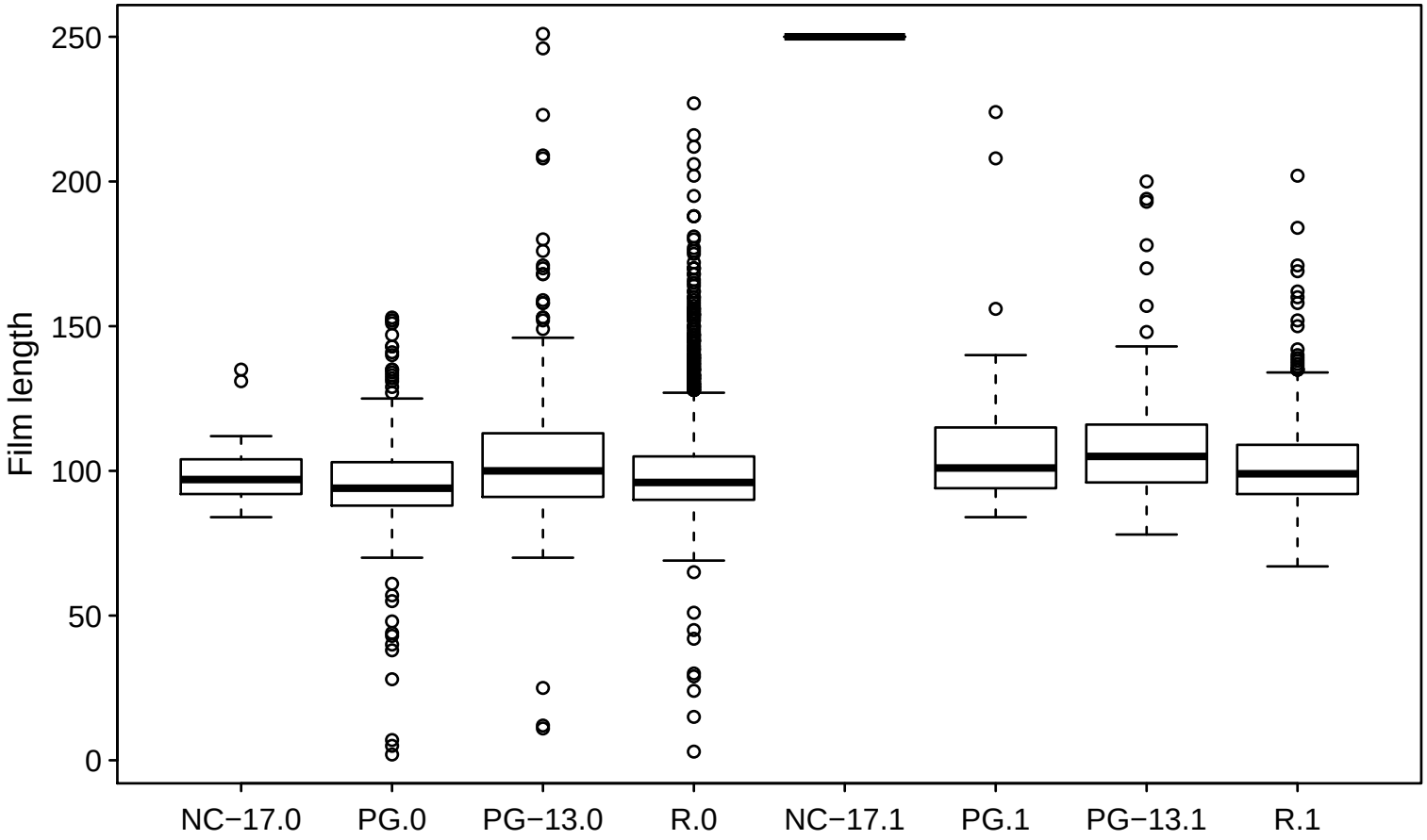
- Central bar in the box: median
- Top/bottom of the box: Q3/Q1
- Whisker: smallest value within $1.5IQR$ of $Q1$ and largest value within $1.5IQR$ of $Q3$
- outliers: asterisks - points outside the whiskers

3.3.4 Box and whisker plots

To do this in R we use the following commands:

```
> par(mfrow=c(1, 2))
> boxplot(movies$Length, ylab="Film length",
+         col="bisque")
> boxplot(movies$Length ~ movies$mpaa,
+         ylab="Film length",
+         col="bisque")
```

3.3.4 Box and whisker plots




```
> boxplot(movies$Length ~ movies$mpaa + movies$Romance)
```