

Statistical Methods:

- Summaries and Plots – Simple exploratory data analysis
- Estimation and Hypothesis Testing – Choosing between models
- Logistic and log-linear models – Trying to explain associations between variables

Statistics - the science of learning from data

Statistics deals with the collection, analysis, interpretation and presentation of data. This data can come from many sources, for example:

- Medical Experiments — Trialing a new drug
- Finance — Price of an asset
- Sample surveys — Opinion Polls

As statisticians we are often concerned with looking at such data and attempting to provide informative summaries, building models for the underlying process and trying to make predictions.

Types of data

The data we collect can come in several basic types:

- Numerical — Price of a stock market asset
- Counts — Number of trips to hospital
- Ordinal — The finishing position of a runner in a race
- Categorical — Hair colour e.g. Black/Blonde

In this course we will mainly be concerned with numerical variables as they tend to occur most frequently in science.

How to present data

A very important part of statistics is presenting the conclusions of your analysis in a clear format. Suppose we have a set of numbers x_1, \dots, x_n which we wish to display simply. We have a number of ways in which we could try and summarize this data for a reader.

Numerical Summaries

Some common numerical summaries we might use to help are

- **Mean** — the average

$$\bar{x} = \sum x_i/n$$

- **Standard Deviation** — the square root of

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

- **Median**

$$\text{Median} = \begin{cases} \text{the middle observation} & \text{if } n \text{ is odd} \\ \text{average of middle two} & \text{if } n \text{ is even} \end{cases}$$

- **Quantile** — a quantile q_α is defined for $0 \leq \alpha \leq 1$ so that proportion α of the data are less than q_α and proportion $1 - \alpha$ is greater than q_α . Note: there are a lot of different definitions if αn is not an integer.
- **Quartiles** (or hinges) — the quantiles $q_{1/4}$ and $q_{3/4}$. Their difference is known as the **interquartile range** or **IQR**.

Plotting data

A very simple and effective way to present data is via plots. There are many types of plots we might use depending on what we think displays the data best.

Stem-and-leaf plots - stem()

These give a visual overview of the numbers. Suppose we have given 10 people a test and received the following scores,

41 94 65 73 84 92 88 67 82 78

We can show this data with a stem and leaf plot

<i>stem</i>	<i>leaf</i>
4	1
5	
6	5 7
7	3 8
8	2 4 8
9	2 4

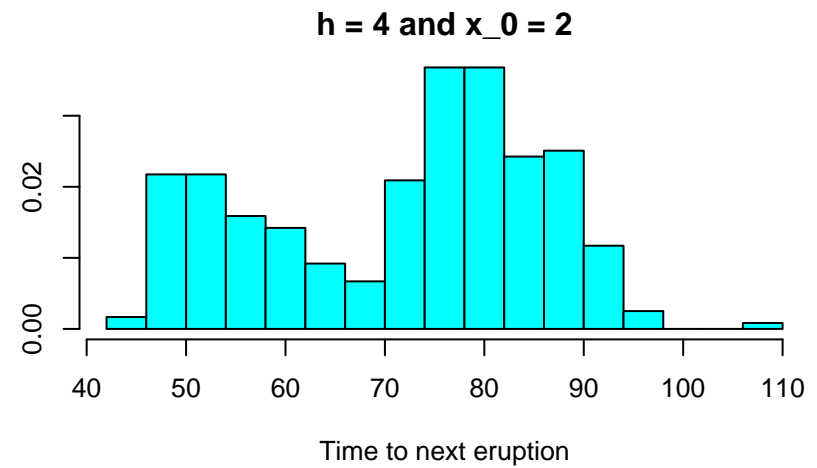
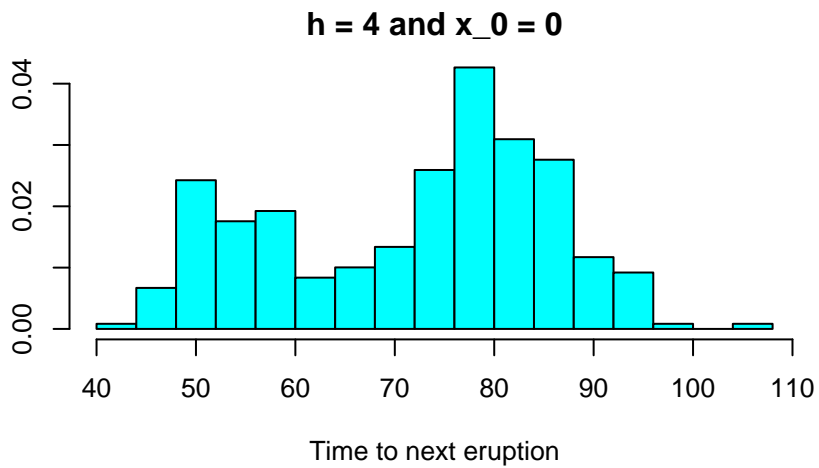
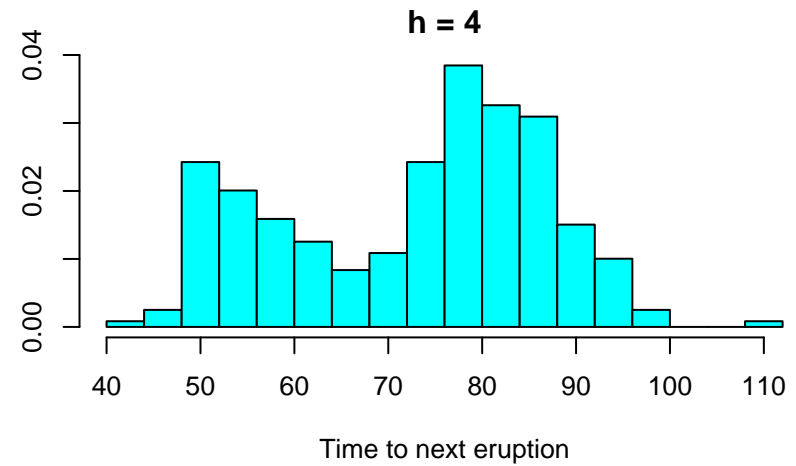
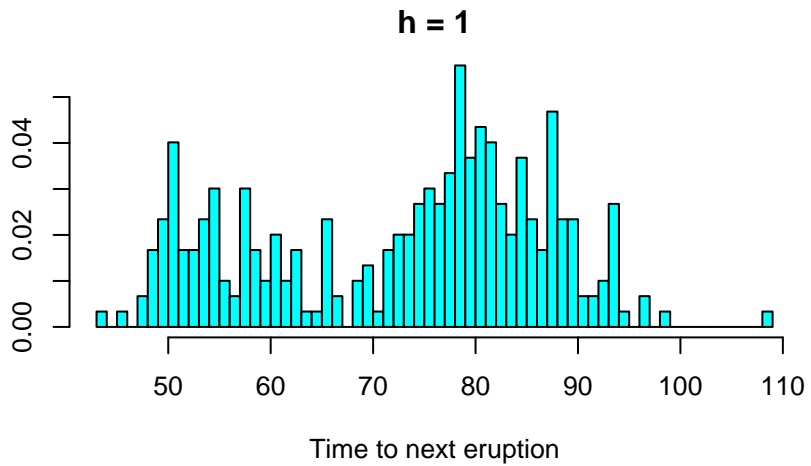
The "leaf" is normally the last digit of the score with the other digits to the left forming the "stem". For example 41 is represented as 4|1.

Histograms - `hist()` or `truehist()` (in MASS library)

Having chosen a set of breakpoints which cover the data (for example 40, 45, 50, ..., 110) a histogram counts how many points fall into each interval.

A true histogram is an estimate of the underlying probability density function and is hence required to have area 1. The area of each rectangle is then the fraction of the data points falling in that cell. *This is not the default using R (you have to use `true.hist()` within the MASS library instead).*

A big problem with histograms is in how we choose the breakpoints. Different breakpoints can lead to very different looking histograms. We illustrate this with a histogram of the waiting time between eruptions of the “Old Faithful” geyser in Yellowstone National Park.



The effect of various bin widths and origins.

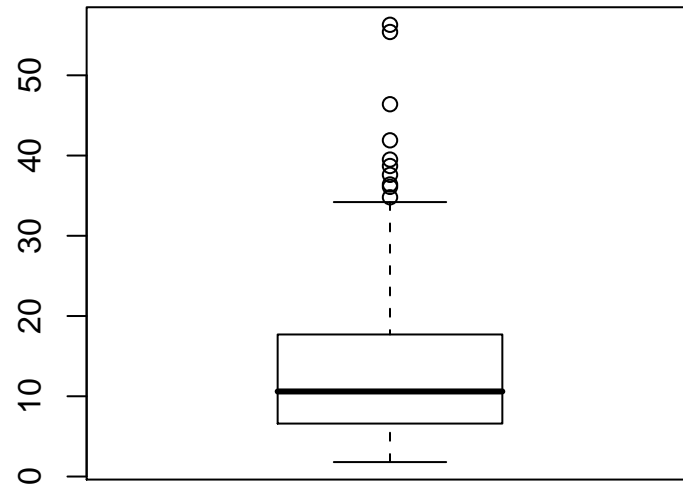
Box plots - `boxplot()`

Box plots give another simple method for representing how data are distributed.

The main box extends from the lower quartile to the upper quartile (hinges) with the central line denoting the median.

The whiskers run to the observation that is nearest to $1.5 \times$ the size of the box (IQR) from the nearest hinge.

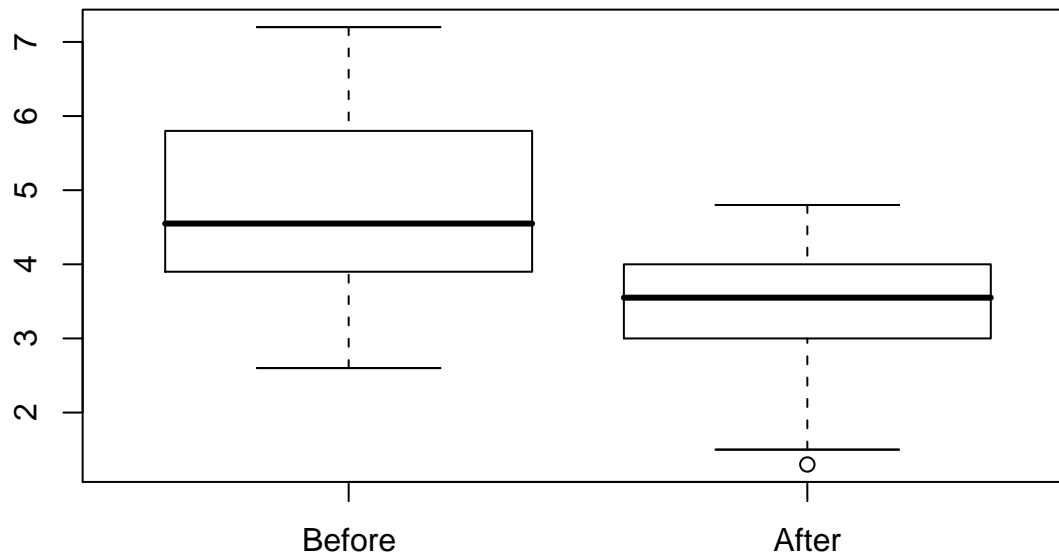
Observations that are more extreme are shown separately.



Concentration of GAG in urine of 314 children

From the plot, we can observe considerable skewness in the data with very heavy tails (and a couple of possible outliers). Overall however, you can see that most children have a concentration between 6 and 18 with a median of around 10.

The real use of boxplots though are for comparison between data sets. As shown by this example of gas consumption by houses before and after insulation is installed.



Empirical Distribution Functions (ECDF's) - `ecdf()`

If X_1, \dots, X_n are IID from a distribution function F , the empirical distribution function is

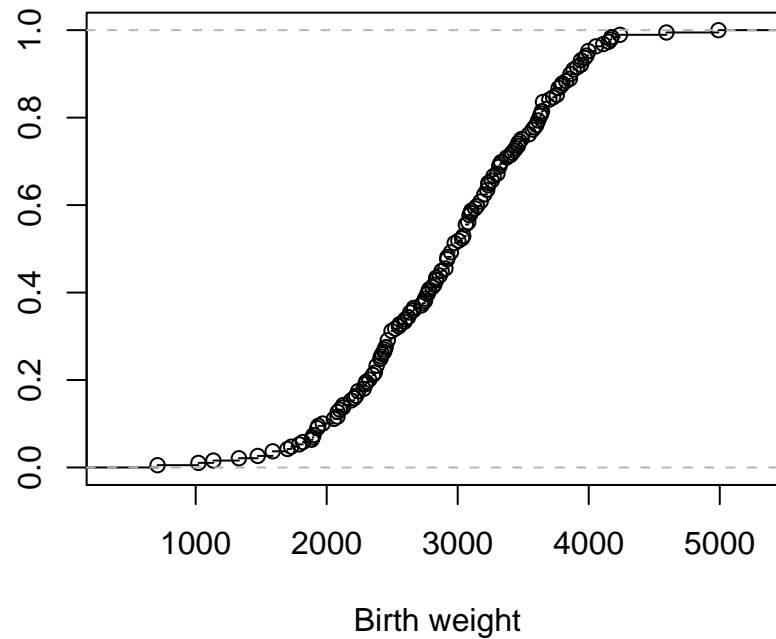
$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{(-\infty, x]}(X_i) = \#\{i : X_i \leq x\}/n$$

it jumps by $1/n$ at each of the observations. By the SLLN we have that $\hat{F}_n(x) \xrightarrow{a.s.} F(x)$.

For a uniform distribution, it would be close to a straight line.

It is also useful since it is the basis for several goodness of fit tests which we cover later.

Birth weight of babies in US hospital — Hosner & Lemeshow (1989)



When curve is steep we have lots of observations, when flat we have fewer.

Quantile plots - `qqnorm()`, `qqline()` e.t.c.

These are used to examine whether a data set might come from a distribution with cdf $F(x)$. Most commonly used to investigate whether a data set is plausibly normal.

To justify how q-q plots work we need the following two results

Result 1 If X has continuous distribution function F , then $F(X) \sim U(0, 1)$.

Results 2 If $Y_{(1)}, Y_{(2)}, \dots, Y_{(n)}$ are the order statistics of a random sample from a uniform $U(0, 1)$ then

$$E\left(Y_{(k)}\right) = \frac{k}{n+1} = P\left(Y_i \leq \frac{k}{n+1}\right).$$

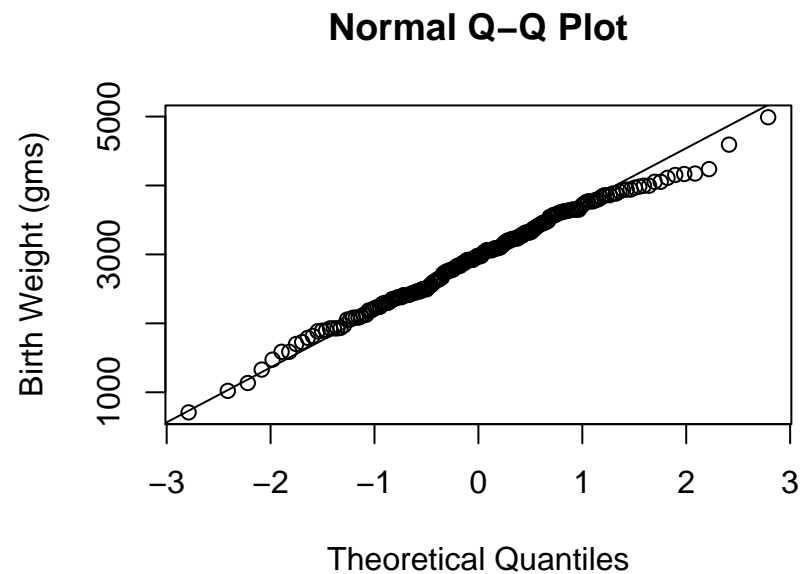
We can now create the following method

Does a data set come from a distribution with cdf $F(x)$?

1. Order observed data to create sequence $X_{(1)}, \dots, X_{(n)}$ — if data does come from a distribution with cdf $F(x)$ then $F(X_{(1)}), \dots, F(X_{(n)})$ are realizations of order statistics from $U(0, 1)$.
2. Solve $F(z_{(k)}) = \mathbb{E} [F(X_{(k)})] = \frac{k}{n+1}$ for $z_{(k)}$.
3. Plot the ordered pairs $(x_{(k)}, z_{(k)})$.

If the data does come from $F(x)$ then the points should form approximately a straight line. Departures from this straight line indicate departures from the specified $F(x)$.

Returning to our birth weight example, does this come from a normal distribution?



The smaller spread of the extreme quantiles is indicative of a lighter tailed distribution but overall reasonably plausible fit.

Density estimation

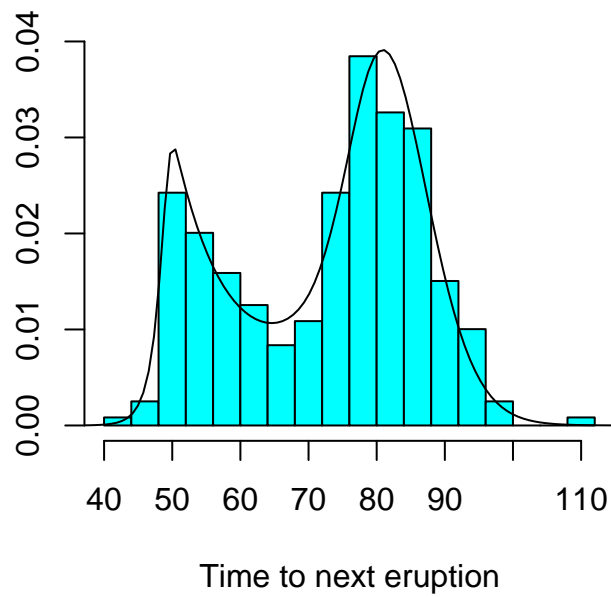
Histograms are only one way of density estimation and can be substantially improved upon. These alternative techniques include:

- Kernel density estimation — `density()` in R
- Log-spline density estimation — `logspline()` in R's `logspline` package

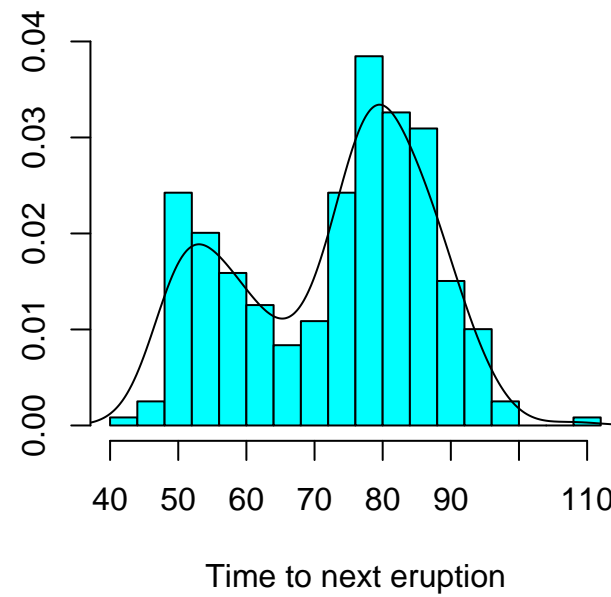
These will be explained in the Computer Intensive Statistics course.

Returning to our geyser data,

Log-spline estimation



Kernel estimation



Transformations

It is often useful to perform a transformation on the data to make it easier to visualize. These transformations can have several aims:

1. Make the distribution less skewed
2. Make a scatter plot more linear
3. Reduce heteroscedasticity

(Random variables are heteroscedastic if they have different variances)

While they often work well in these aims, caution should still be applied in their usage due to the increased difficulty of interpretation.

A general family of transformations is named after Box & Cox (1964),

$$y = \frac{x^\lambda - 1}{\lambda}$$

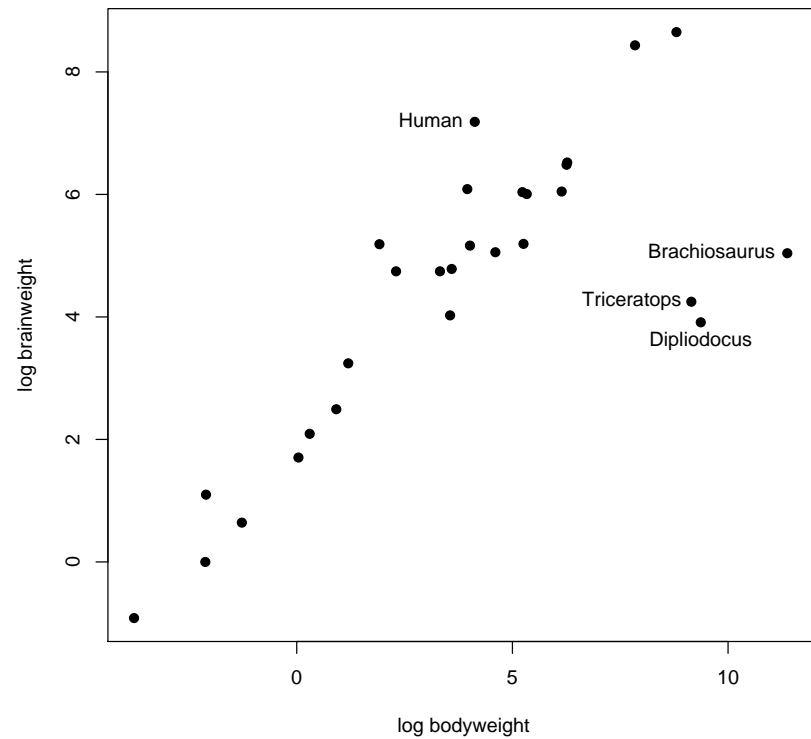
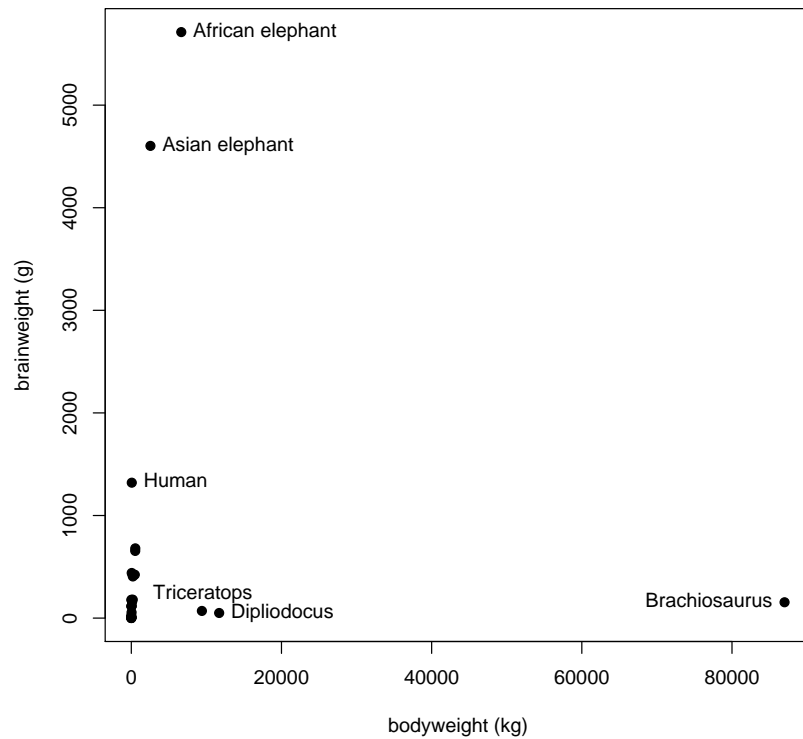
with the case $\lambda = 0$ being filled in by continuity as $y = \log_e x$.

(**Note** - a log transformation is natural for positive quantities)

In practice we use $y = x^\lambda$ for a few convenient values such as $-2, -1, -0.5, 0, 0.5, 1, 2$.

There are other transformations e.g. arcsin and logit (both used for proportions)

Example 1 - The average brain and body weights for 28 species of land animals



Discrete Data

We often have data sets containing a few ordinal or categorical variables and we are interested in finding out about how they vary together. Lets consider an example on the caffeine consumption of women in a maternity ward by marital status.

	0	1 – 150	151 – 300	> 300
<i>Married</i>	652	1537	598	242
<i>Prev.Married</i>	36	46	38	21
<i>Single</i>	218	327	106	67

Barplot representation - `barplot()`

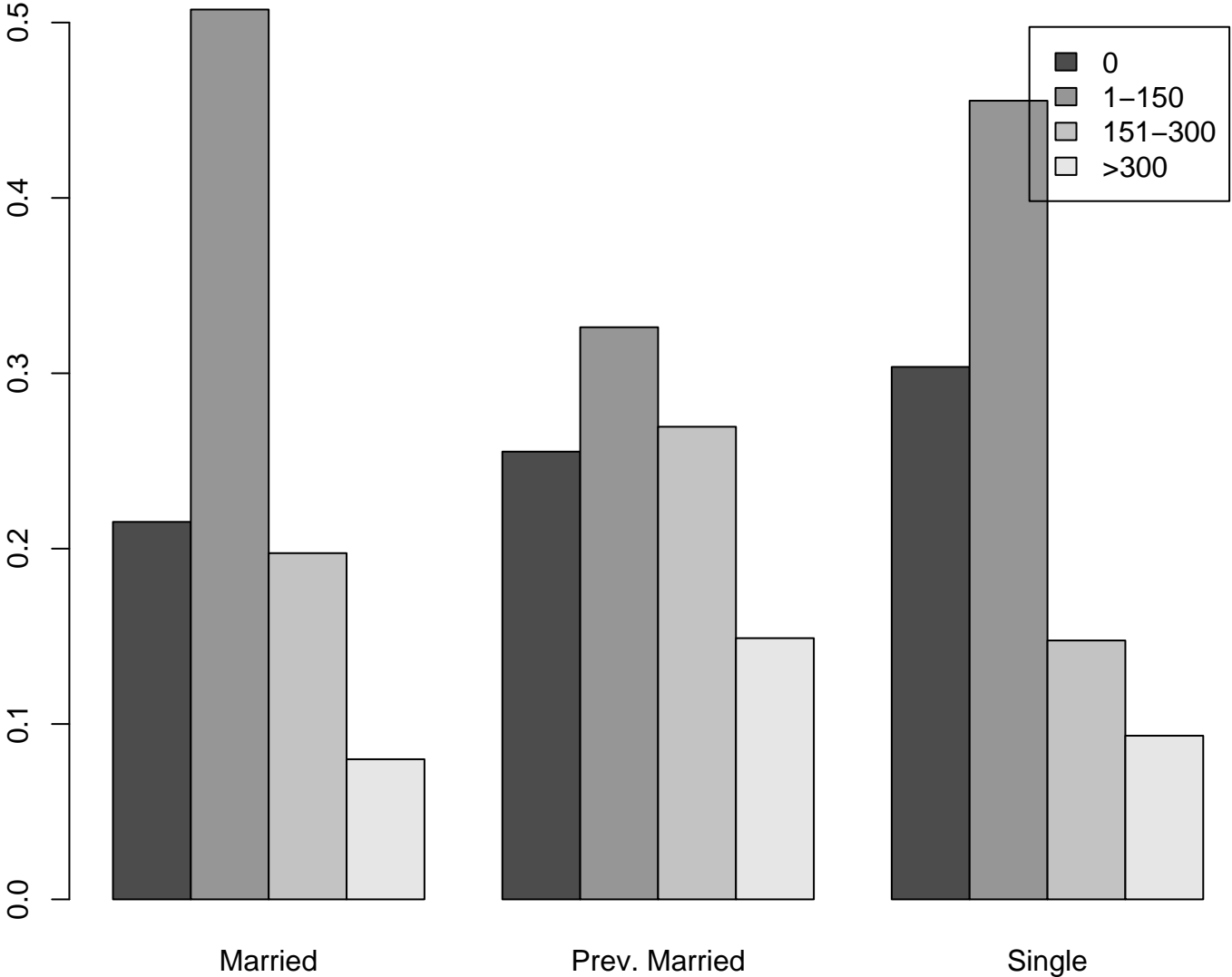
This shows for each category (marital status), the proportion of women with each level of caffeine consumption.

Again the default for R is to plot the number of women and not proportion.

Mosaic plot - `mosaicplot()`

This creates a unit square consisting of separate cells (corresponding to the cells in the contingency table) with the size of each cell proportional to the corresponding entry.

Barplot of relationship between martial status and caffeine consumption



Mosaic plot of caffeine consumption



Multidimensional data

Parallel Coordinate Plots `parcoord()`

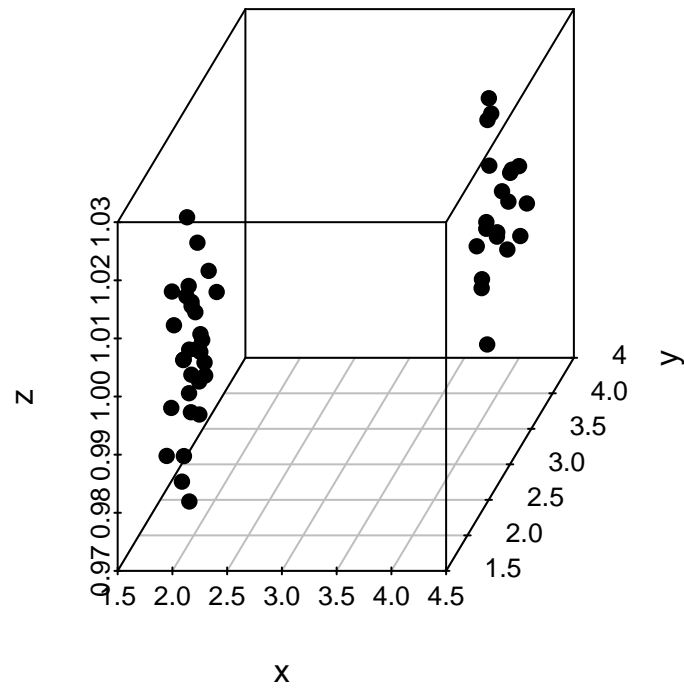
Alternatively, we may have high dimensional data. Since we are used to living in a 3-d world, this is difficult to visualize. One way to try and get around this is a parallel coordinate plot.

To show a set of points in n -dimensional space we draw n parallel axes, typically vertical and equally spaced. Each n -dimensional point is then represented by a polyline that intersects the axes the parallel coordinate axes at the coordinate values of the data point.

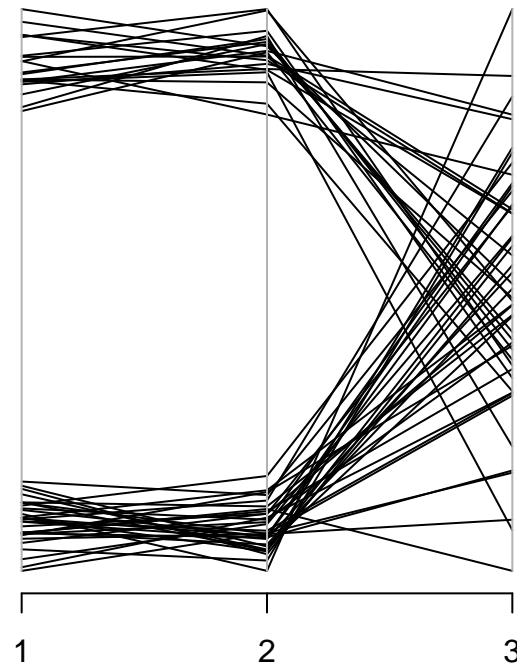
While we do lose some information about the data set, we can still recognize many spatial patterns if we know what they look like projected to parallel coordinates e.g. Cluster, Lines, Outliers

Clusters in Parallel Projection - Centred on $(4, 4, 1)$ and $(2, 2, 1)$

Two clusters

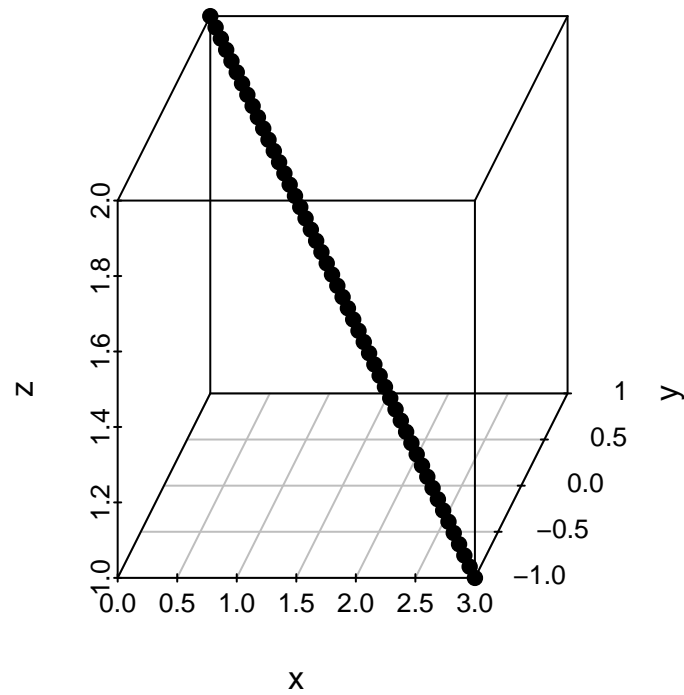


Parallel Representation

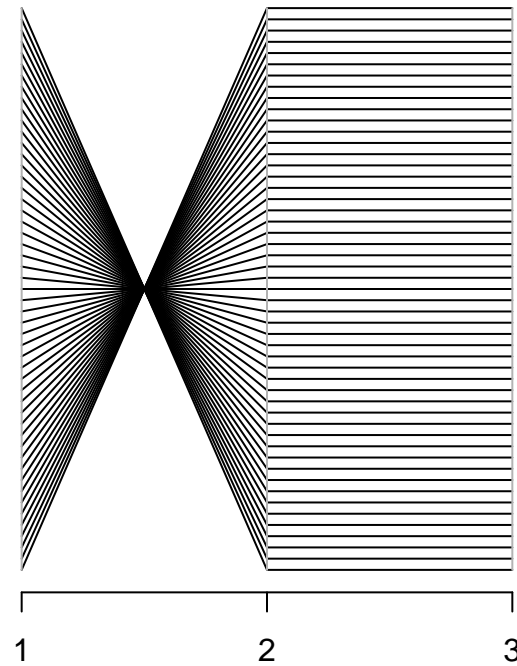


Line in parallel projection - $x = 3t, y = 1 - 2t, z = 2 - t$

Line



Parallel Representation



Scatter Plots - `pairs()`

Scatter plots also allow an insight into the relationship between variables. They allow us to see:

- Pairwise relationships between variables
- Outliers
- Clusters

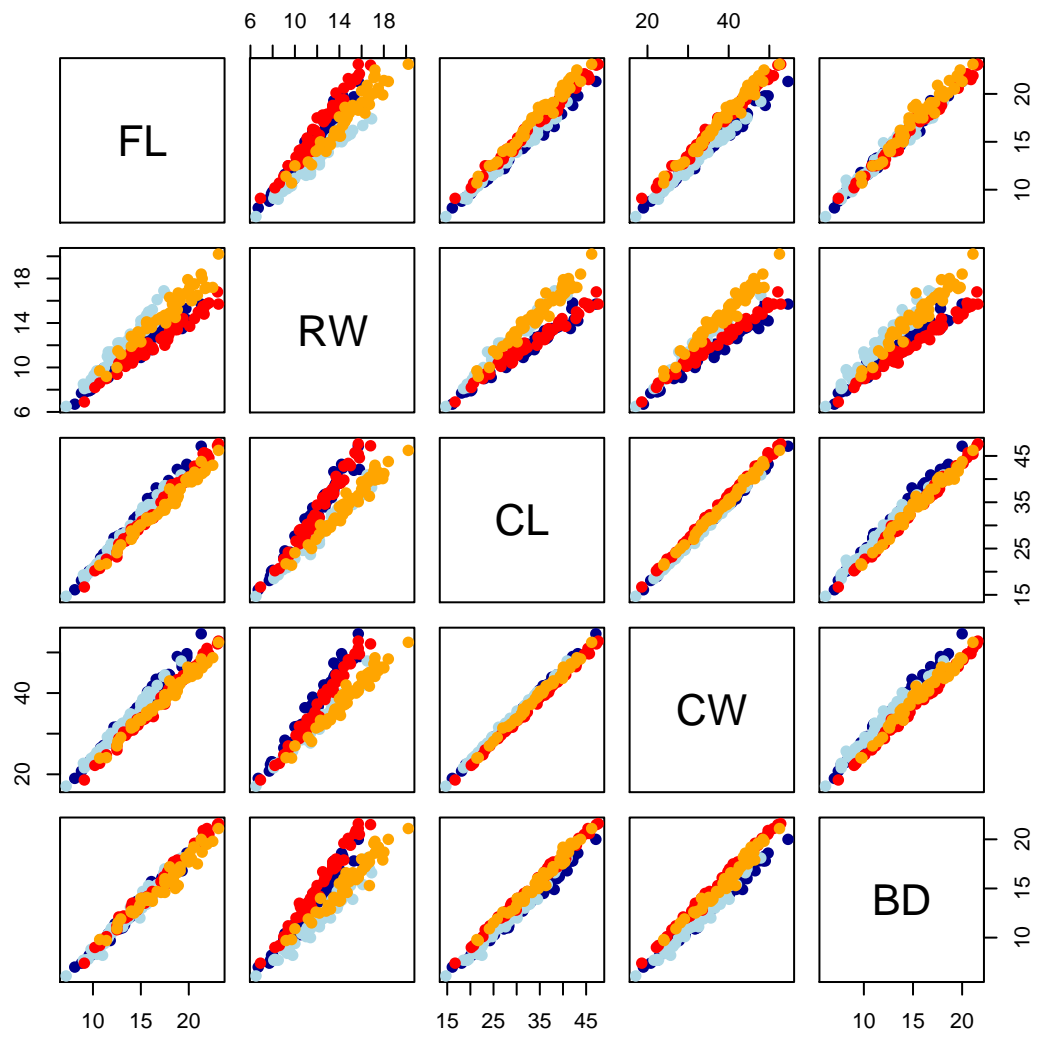
However, as we are only looking orthogonal to coordinate directions it is still difficult get a complete visualization of the data set.

In the following example, data was recorded on 200 *Leptograpsus variegatus* crabs from Western Australia. They occur in two colour forms, blue and orange. Mahon collected 50 of each form of each sex and made five measurements. Carapace (shell) length CL and width CW, the size of the frontal lobe FL, rear width RW and body depth BD.

Are the colour forms are different species?

In the next plot we show the scatterplot with **Male Blue Crabs** , **Female Blue Crabs**, **Male Orange Crabs** , **Female Orange Crabs**.

We can see the data are highly correlated and there is a suggestion that they are separate species. Need to perform further analysis e.g. principal components and cluster analysis



Strip Plots - `stripchart()` in R's library `lattice`

Plot the data along a line with each data point represented by a box. Often we do this in a trellis pattern so that we can analyse the data by type and observe any differences.

We consider the example of 214 glass fragments collected in forensic work by B. German. The glass is grouped as window float glass (WinF), window non-float glass (WinNF), vehicle window glass (Veh), containers (Con), tableware (Tabl) and headlamps (Head).

Every fragment has a measured refractive index (RI) and composition (weight percentage of oxides of NA, MG, Al, Si, K, Ca, Ba and Fe). Can we identify a new glass fragment based upon these measurements?

