# What is Statistics?

# Simple Summaries and Plots

Brian D. Ripley

# Statistics

The name derives from state-istics, quantification of the state, an area now know as *official statistics*. Health statistics have long been important, one early and influential example being Florence Nightingale's work in the Crimean war.

Modern statistics can be paraphrased as

*Understanding uncertainty*

Mainly (but not always, e.g. simulations) statistics is concerned with looking at data and producing informative summaries, building models for the data-generating mechanism and making predictions or decisions.

# Where do the Data come from?

NB: ‘data’ is the plural of ‘datum’, and pedants do care.

- Experiments
- Observational studies
- Surveys: more precisely sample surveys.

Sometimes it is sufficient to draw conclusions about the dataset available, but normally one is interested in some population from which that dataset was drawn. To do so we need to consider ‘*what might have happened but did not*’.

Causality *vs* association.

Data mining *vs* data dredging.

# Types of data

The possible types of data collected are limited only by your imagination, and people analyze maps, images, audio streams (e.g. speech), manuscripts of medieval music, .... But ultimately we will have (possibly many) variables of a few basic types.

- Numerical
- Counts
- Ordinal
- Categorical

Most of this course is about numerical variables as they are most important in science (whereas categorical data are very important in social science).

# Numerical Summaries

Suppose we have a set of numbers $x_1, \ldots, x_n$.

The *mean* is the average $\overline{x} = \sum x_i/n$.

The *standard deviation* is the square root of

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

The *median* is the middle observation if $n$ is odd, the average of the middle two if $n$ is even.

*Quantiles* are defined for $0 \leqslant \alpha \leqslant 1$ so that proportion $\alpha$ of the data are less than $q_\alpha$ and proportion $1 - \alpha$ is greater than $q_\alpha$. There are many (at least 8) different definitions if $\alpha n$ is not an integer.

The *quartiles* are the quantiles $q_{1/4}$, $q_{3/4}$, and their difference is the *inter-quartile range*, the IQR.

For one definition of quantile, the quartiles are known as *hinges*.

# Stem-and-leaf Plots

These provide a visual overview of the numbers, here measurements of serum IgM in young children.

```
N = 298   Median = 0.7
Quartiles = 0.5, 1

Decimal point is 1 place to the left of the colon

    1 : 000
    2 : 0000000
    3 : 000000000000000000
    4 : 00000000000000000000000000
    5 : 0000000000000000000000000000000000
    6 : 000000000000000000000000000000000000
    7 : 0000000000000000000000000000000000000000
    8 : 0000000000000000000000000000000000000000
    9 : 00000000000000000000
   10 : 00000000000000000
   11 : 00000000000000000
   12 : 000000
   13 : 0000000
   14 : 000000000
   15 : 000000
   16 : 00
   17 : 000
   18 : 000
   19 :
   20 : 000

High: 2.1 2.1 2.2 2.5 2.7 4.5
```

# Histogram

Unfortunately, Americans do not conform to international usage here!

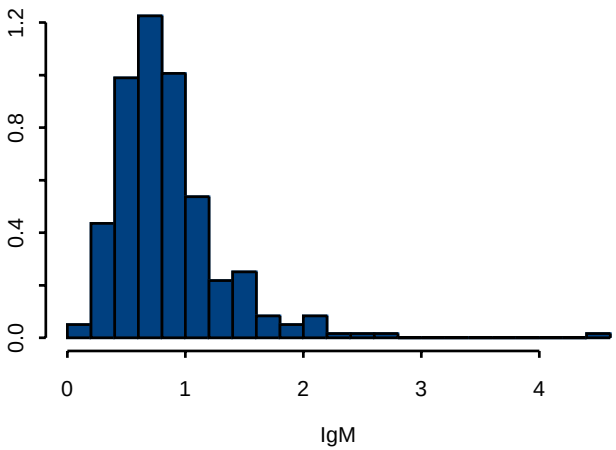
We can choose a set of breakpoints, say $0, 0.2, 0.4, \ldots, 4.6$ covering the data, and count how many points fall into each interval. Americans plot the counts or the proportions or percentages.

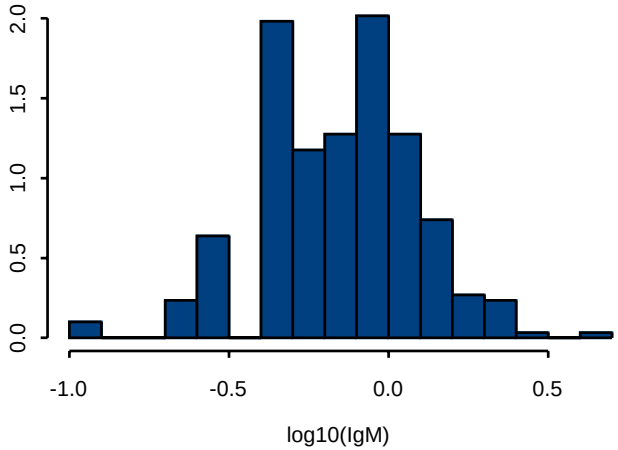
A true histogram has the *area* of each bar proportional to the count, and total area one. This matters if the breaks are not equally spaced, as in the example of people's ages when admitted to hospital following an accident.

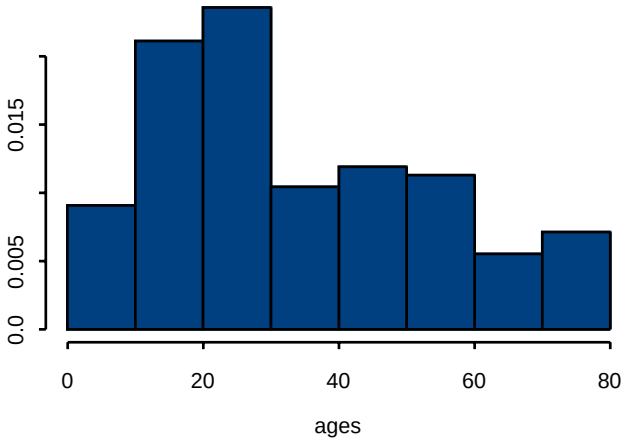
How do we choose the number and position of the breaks?

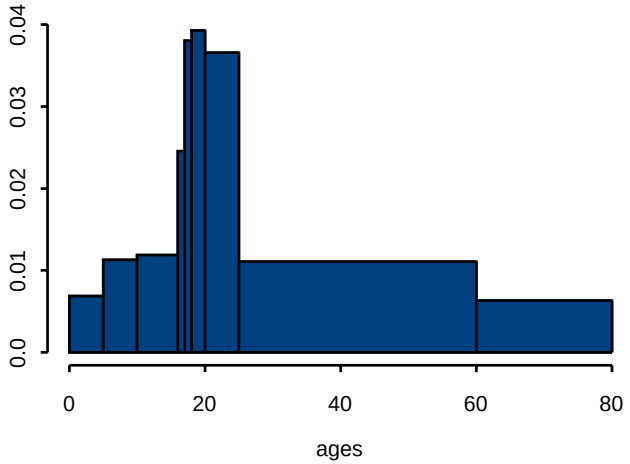
Normally best left to the software designer.

# Examples of histograms



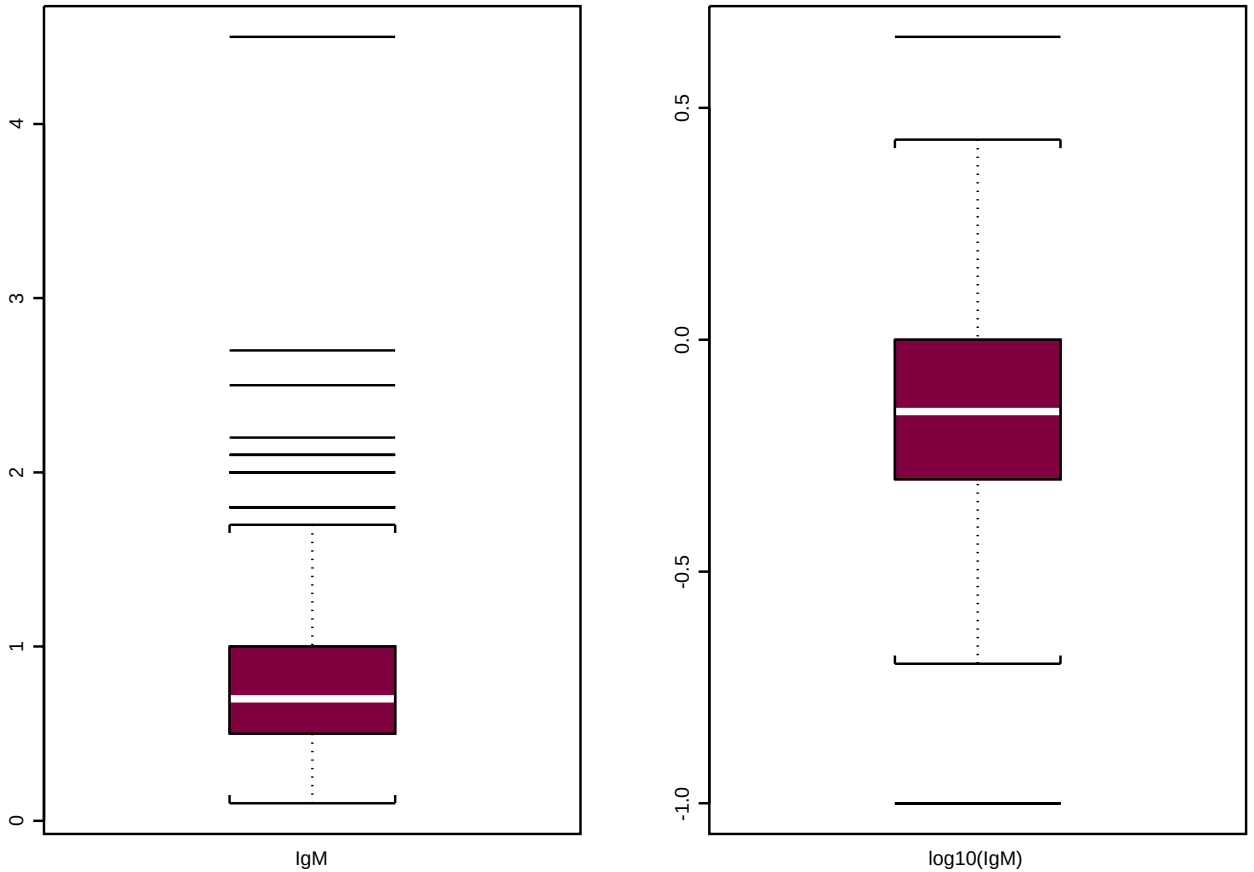
# Box-and-whisker plots

or just *box plots*. The central box extends from one hinge to the other, and the central line in the box is the median. The whiskers run to the observation that is nearest to $1.5\times$ the size of the box from the nearest hinge. Observations that are more extreme are show separately.

There are about as many variations as software designers.

*Parallel* box plots are often useful to show the differences between sub-groups of the data.

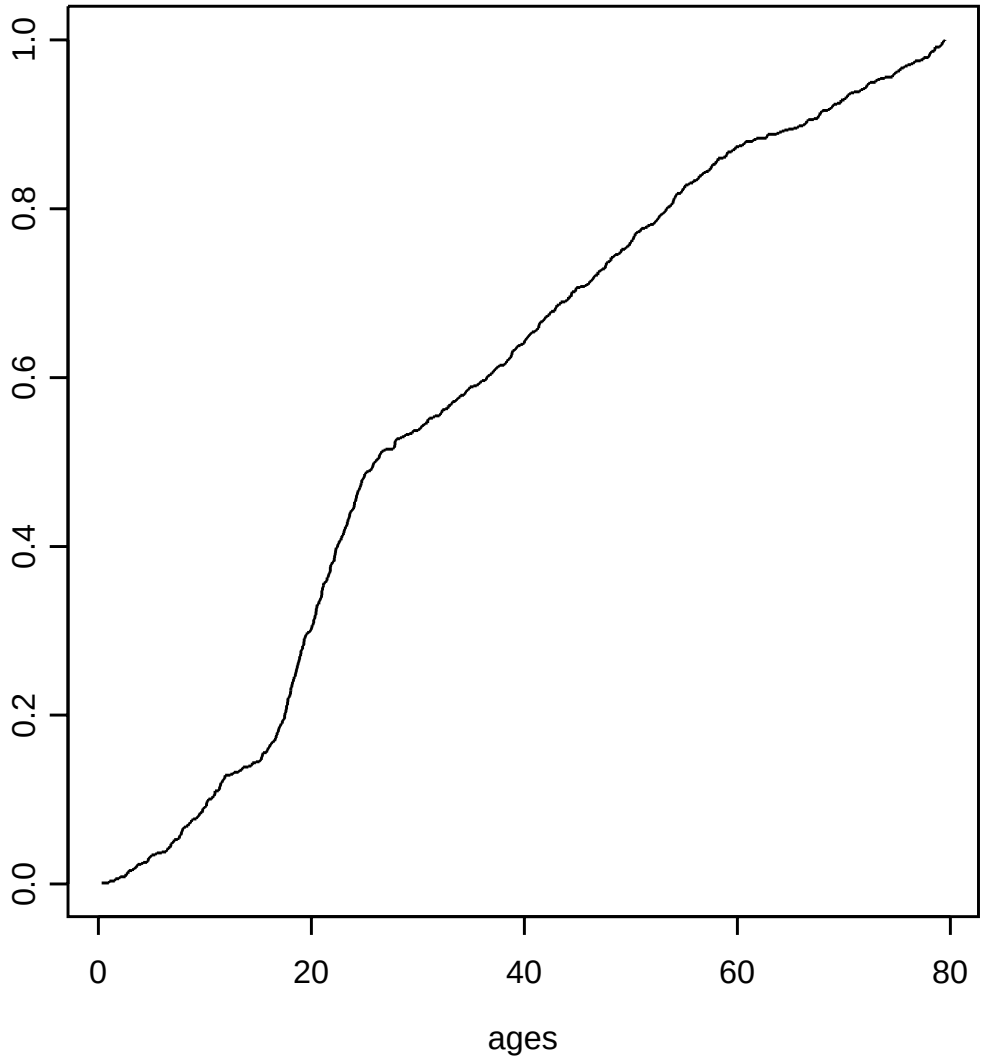
# Example boxplots



# Distribution functions

The empirical distribution function (ECDF) jumps $1/n$ at each of the observations ($r/n$ if there is an $r$-way tie). This would close to a straight line for a uniform distribution, an ogive for a unimodal distribution, ....

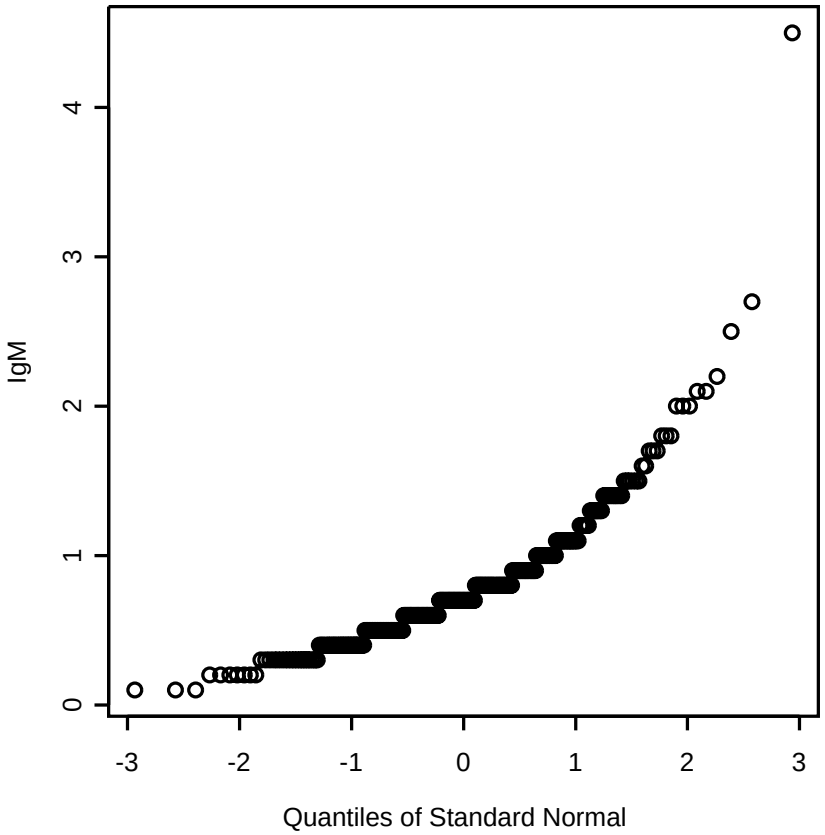
# ECDF of ages

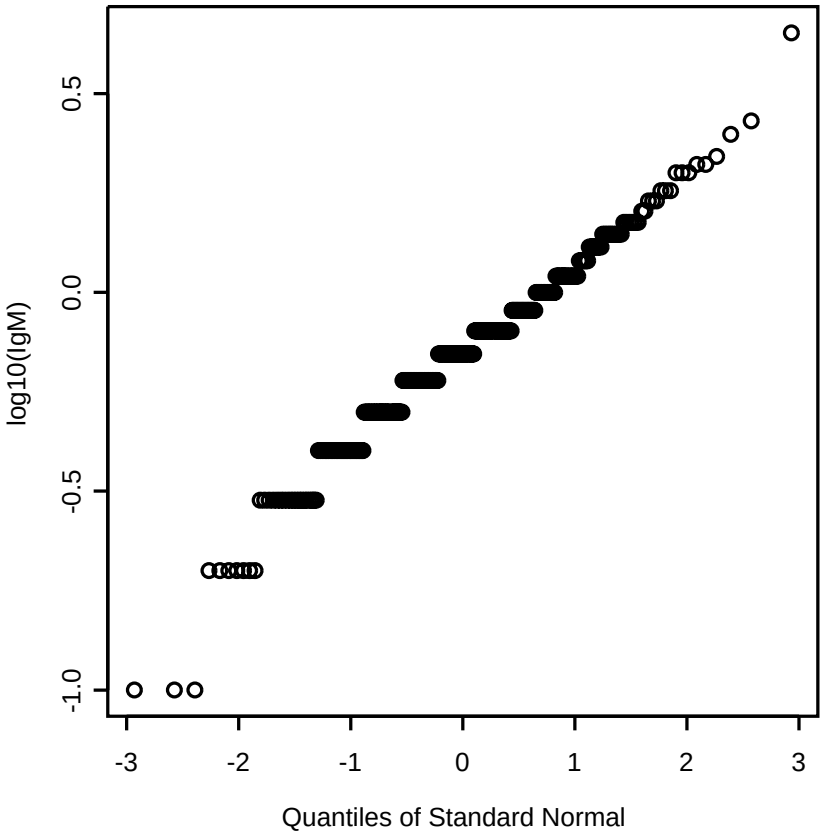


# Quantile plots

The quantiles can be read off by ‘inverting’ an ecdf plot. Q-Q plots compare two sets of data by plotting the quantiles of one against the other. Perhaps more commonly one set of data is replaced by the quantiles of a theoretical distribution.

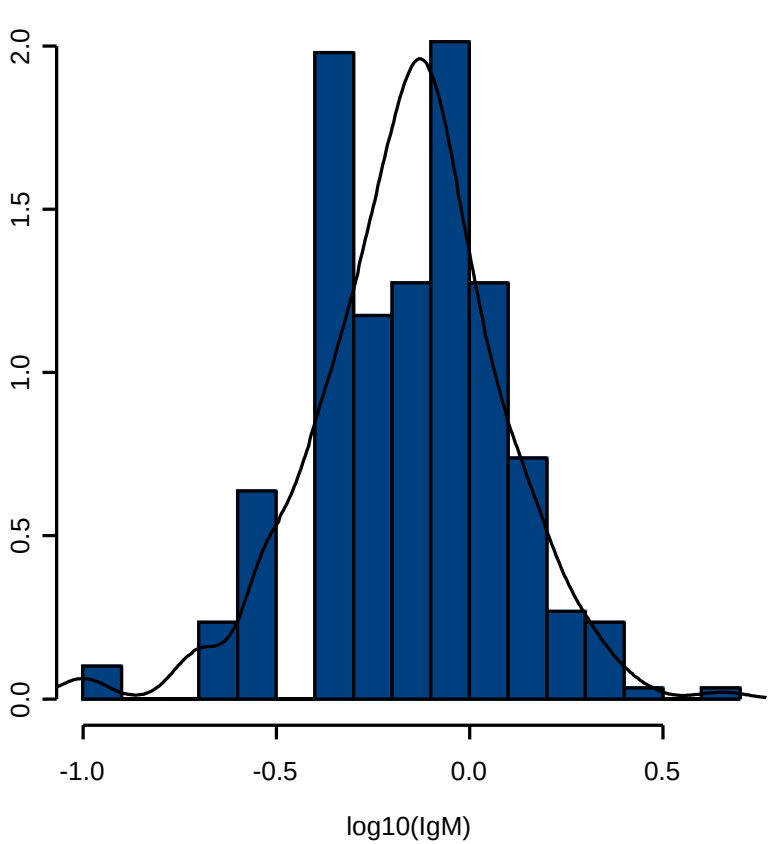
Small point: perhaps not the quantiles but the expected values of the sample quantiles from that distribution.

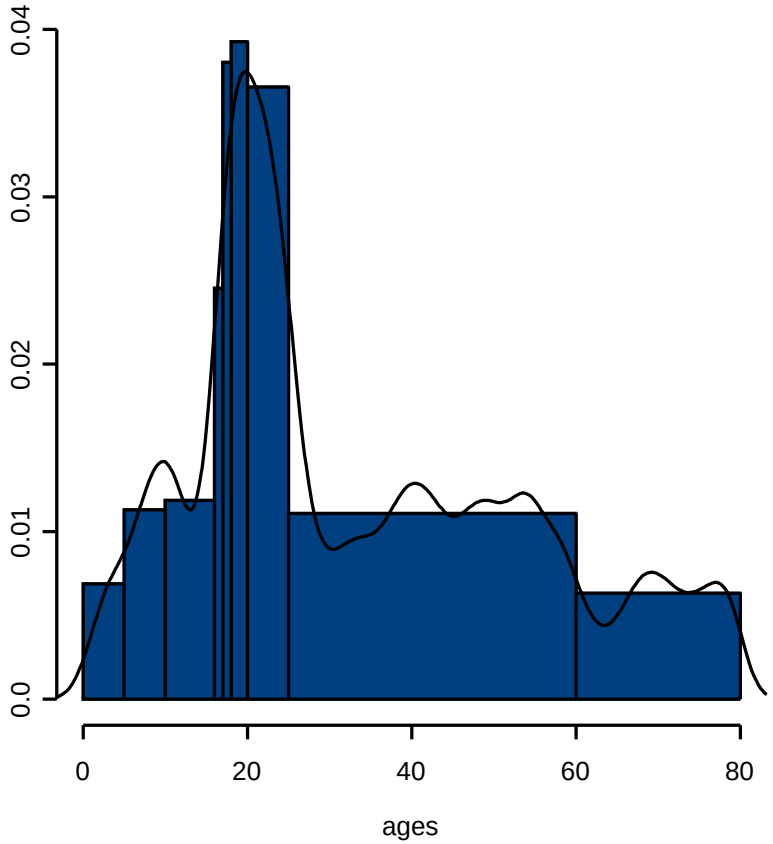

IgM
0
1
2
3
4
-3
-2
-1
0
1
2
3
Quantiles of Standard Normal


log10(IgM)
-1.0
-0.5
0.0
0.5
-3
-2
-1
0
1
2
3
Quantiles of Standard Normal

# Density estimates

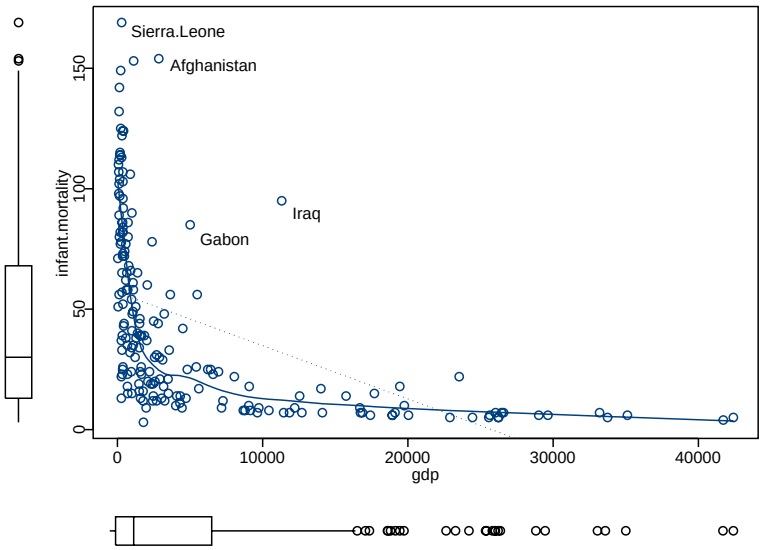
The true histogram estimates an underlying probability density function, which is why it has area one. We can find better estimates of the density function (assumed to be smooth). How we do it is highly technical, and reliable methods are from the last decade.

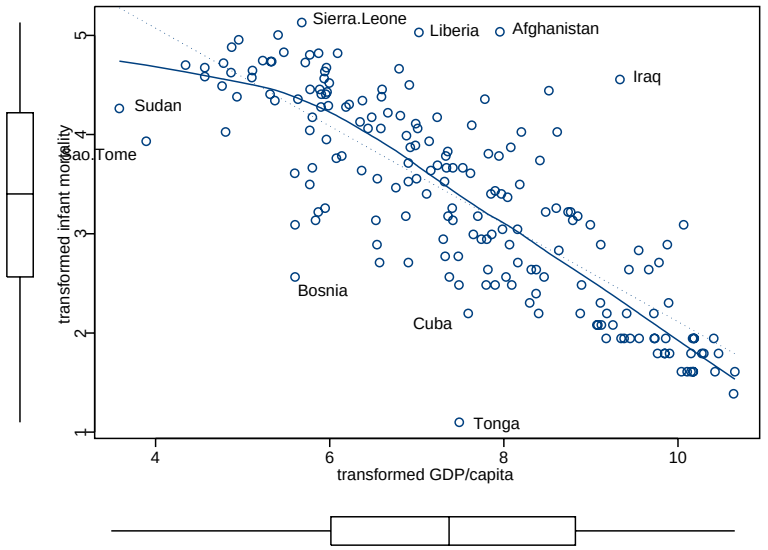

0.0
0.5
1.0
1.5
2.0
-1.0
-0.5
0.0
0.5
log10(IgM)


0.0
0.01
0.02
0.03
0.04
0
20
40
60
80
ages

# Transformations

We have already seen that working on log scale can make data easier to visualize, as well as making sense for positive quantities. Here is another example, from UN data on countries.



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

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

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

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

Transforming the data can have several aims:

- to make the distribution less skewed
- to make a scatterplot more linear
- to reduce heteroscedasticity

There are other transformations, e.g. arcsin and logit for proportions.

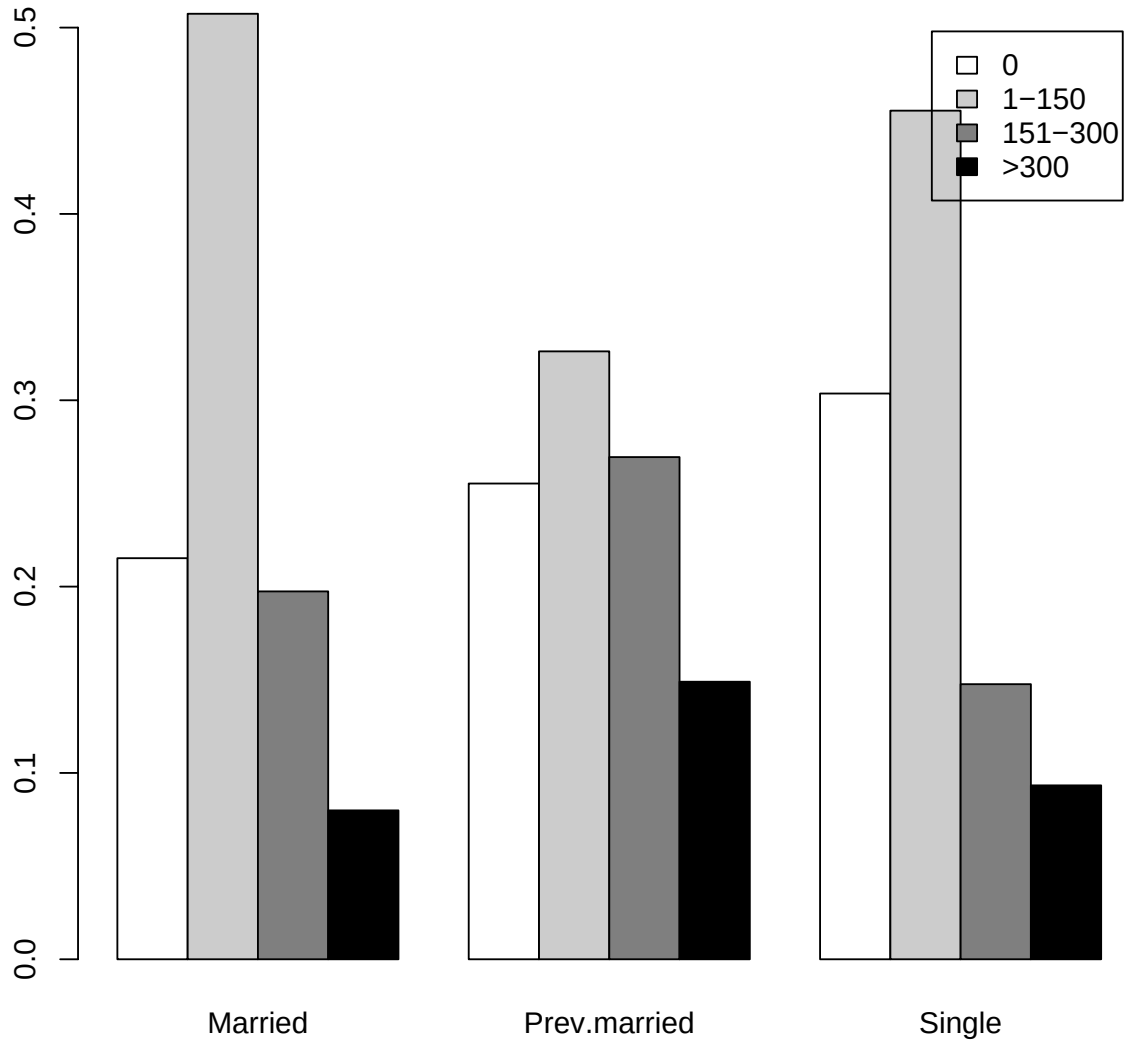
# Discrete data

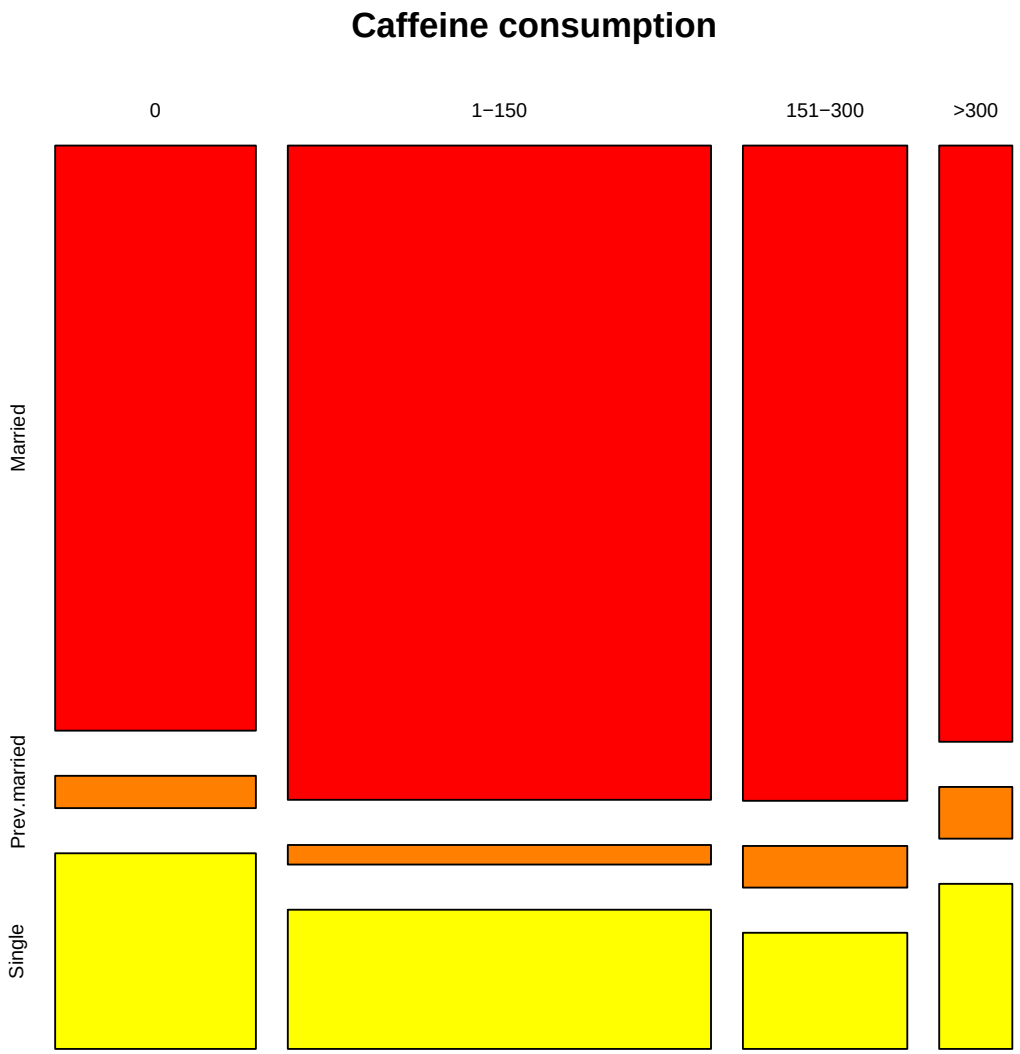
A set of observations on a single categorical variable is not at all interesting.

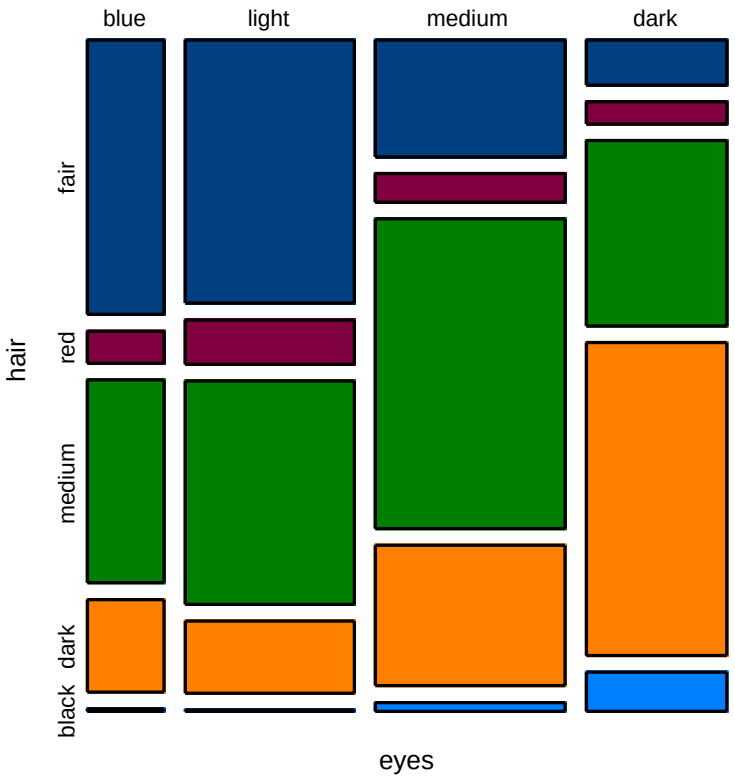
Suppose we have a few ordinal or categorical variables: the interest questions are then how they vary together. Here is a cross-tabulation on the caffeine consumption of women in a maternity ward by marital status.

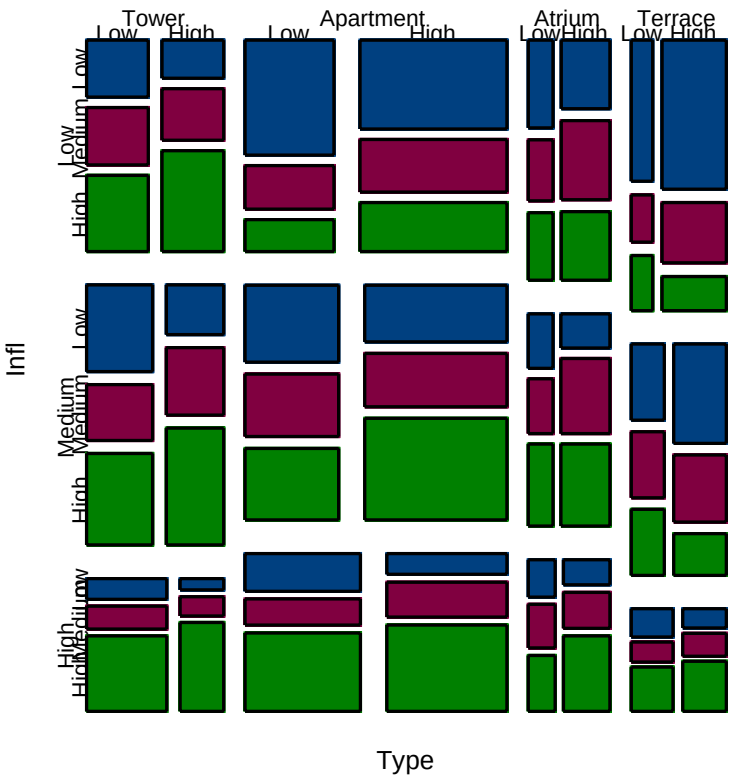
| | 0 | 1–150 | 151–300 | >300 |
|---|---|---|---|---|
| Married | 652 | 1537 | 598 | 242 |
| Prev.married | 36 | 46 | 38 | 21 |
| Single | 218 | 327 | 106 | 67 |

The next two slides show graphical representations, a set of barplots, and a *mosaic plot*.


0.0
0.1
0.2
0.3
0.4
0.5
0
1−150
151−300
>300
Married
Prev.married
Single


Caffeine consumption
0
1−150
151−300
>300
Married
Prev.married
Single


People in Caithness
blue
light
medium
dark
fair
red
medium
dark
black
hair
eyes


Copenhagen housing survey
Tower
Apartment
Atrium
Terrace
Low
High
Low
High
Low
High
Low
High
Low
Medium
High
Low
Medium
High
Low
Medium
High
Infl
Type