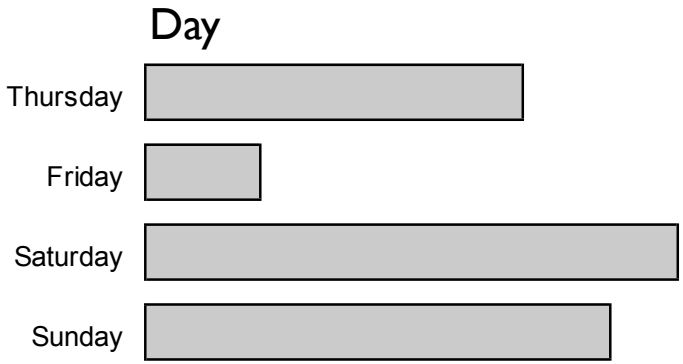
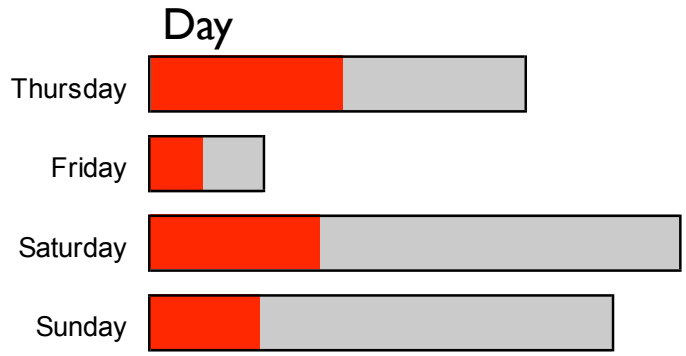
Chapter 2

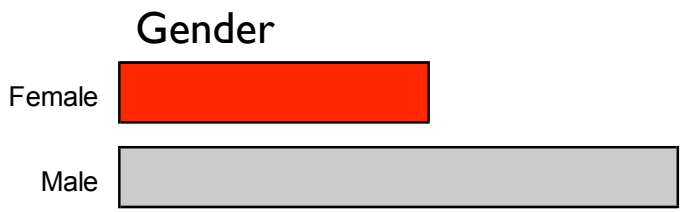
Examining a Single Variable

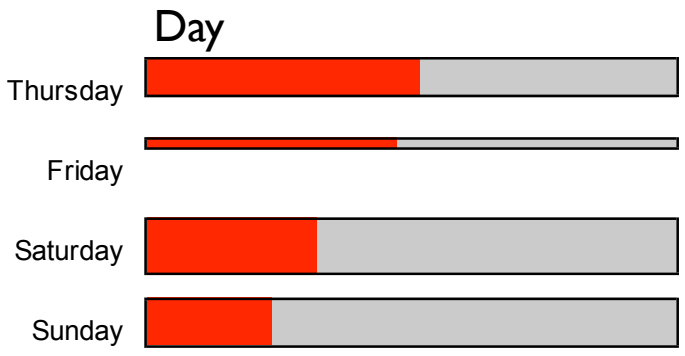
Barcharts and Spineplots

- Barcharts are one of the most simple, yet effective graphical displays, though the benefit of barcharts only arises when selection and highlighting is possible
- Selecting one of more categories allows to trace these groups through all plots
- Highlighting shows the proportion of the selected group in each category
- Spineplots exchange proportionality of height and width to facilitate comparison





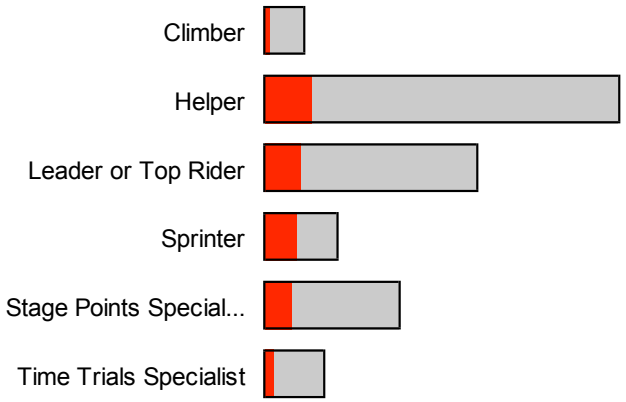


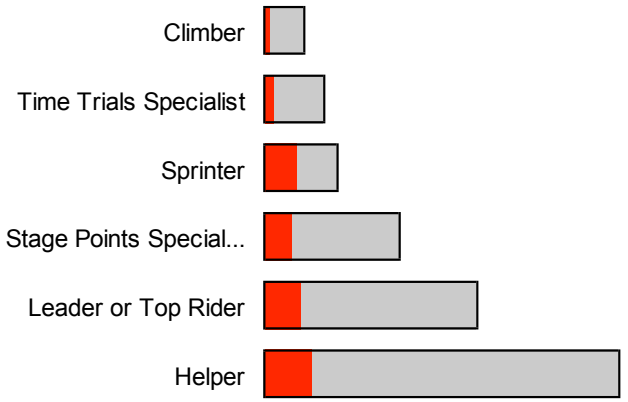


Barcharts: Sorting

- Sorting can make a big difference for the interpretation of barcharts
- Example: Tour de France 2005; drop-outs selected

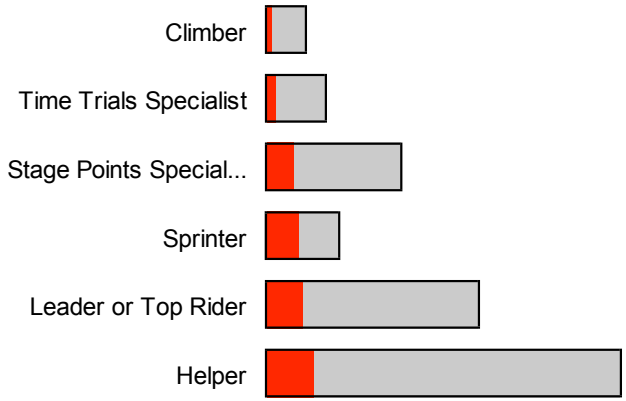
Default: **Lexicographic**

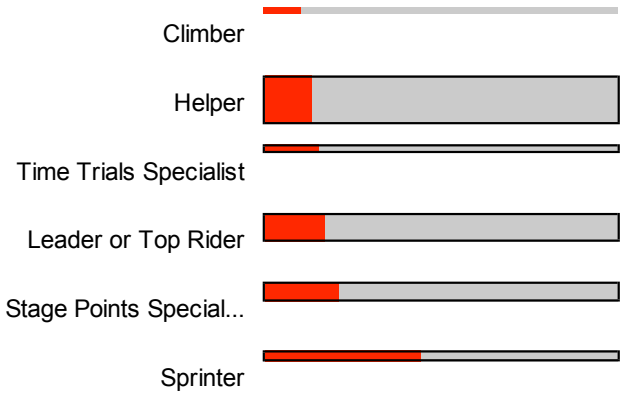




Sorted by category **size**

Sorted by **absolute** highlighting

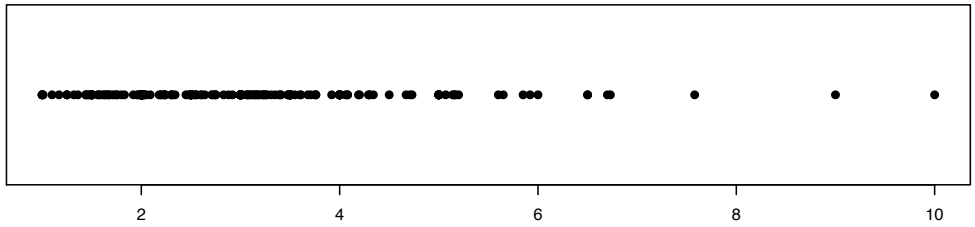


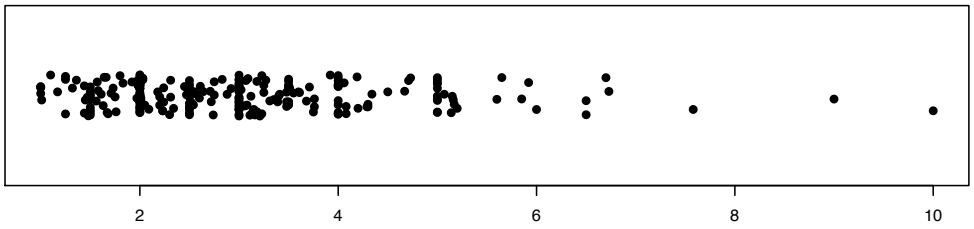


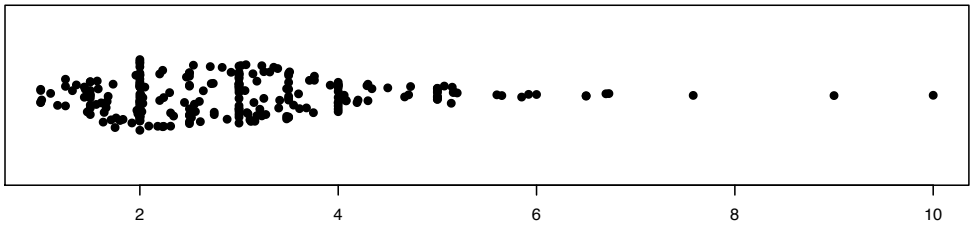
Sorted by **relative** highlighting

Dotplots

- As long as a sample is really small, dotplots often work well
- Overplotting becomes a serious issue for larger samples
- As a cure, jittering can be used in different flavors
  - Standard, no jittering
  - Uniform
  - Proportional to density



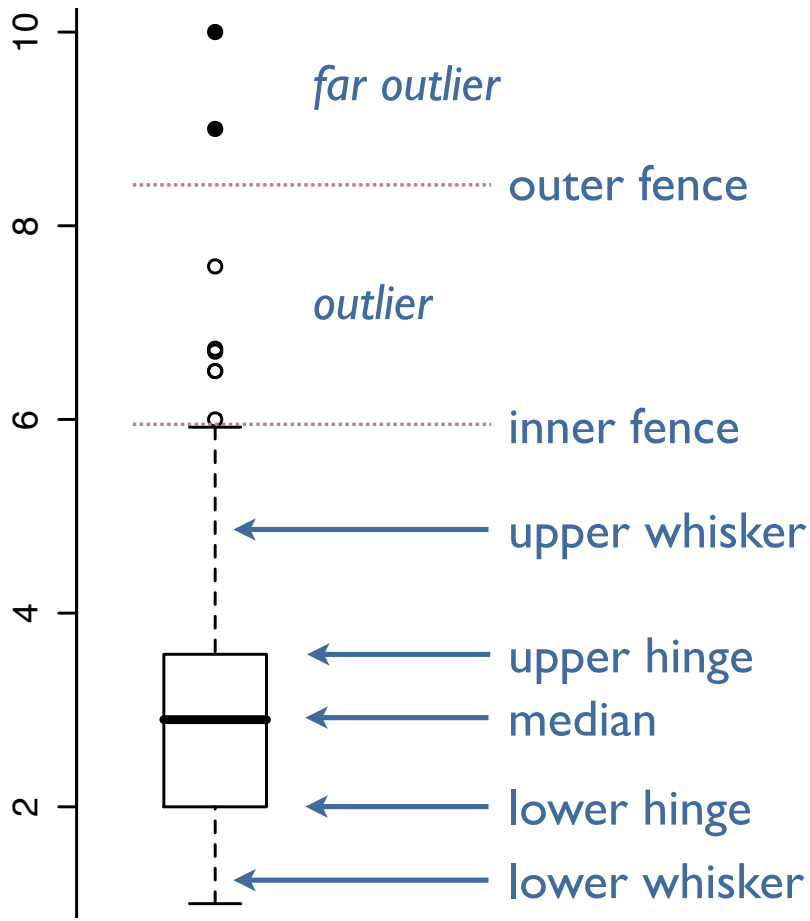




- No matter how big a sample is, or how severe the over-plotting might be, we always can find **gaps** (where no data at all) in the dataset

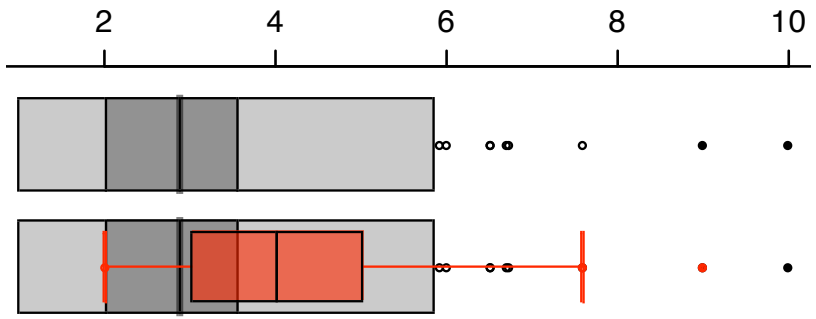
Boxplots

- Boxplots are somehow the “round tag in the square hole”, but extremely efficient when comparing distributions
- Based upon (robust) statistics, the boxplot has the following properties
  - 50% of the data in the center of the distribution fall within the box
  - The probability for a value to be classified as outlier is just below 1%
- Often boxplots are drawn not according to Tukey’s definition, but follow other definitions which are often less robust.
- Highlighting is not straight forward …



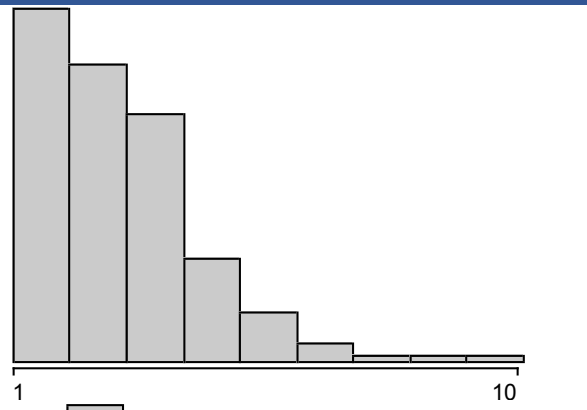
Highlighting in Boxplots

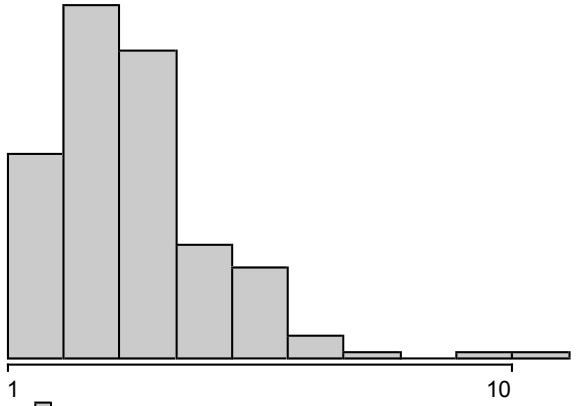
- Because of the definition of the boxplot (neither area base nor glyph based) it is not directly obvious how highlighting should be implemented.
- Highlighting can not be easily possible, because we need to compare statistics of distributions
- When we modify the layout of the (unhighlighted) boxplot (replace the whiskers with light gray boxes) we can superpose the highlighted boxplot in its standard layout
- Making the highlighted boxes semi-transparent further enhances the readability of the highlighted plot.

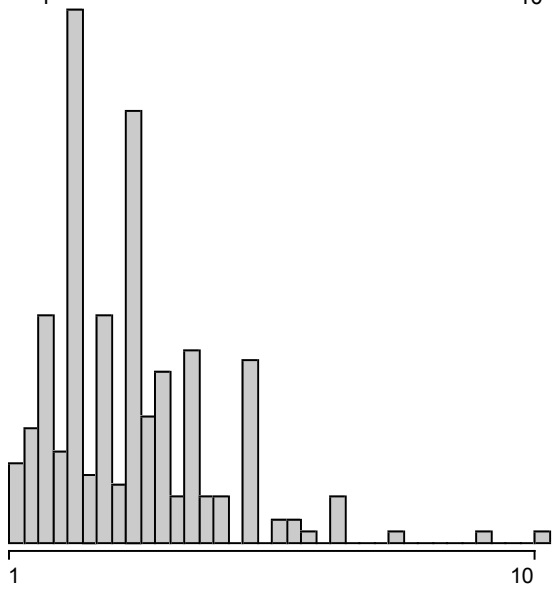


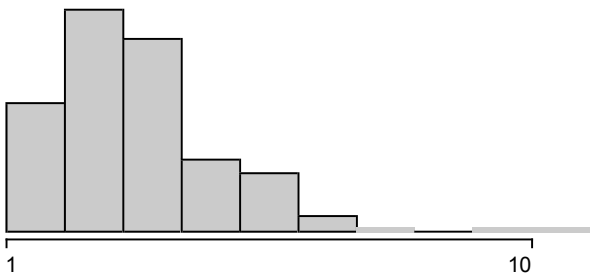
Histograms

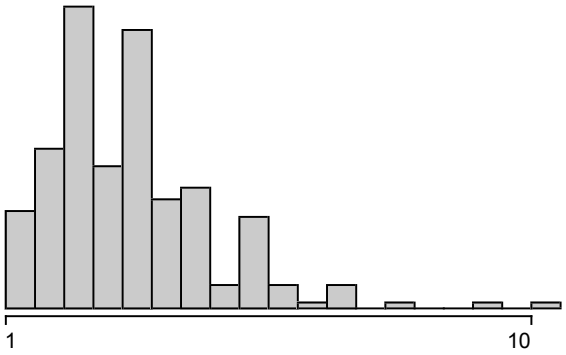
- Histograms are most often used to display the shape of the distribution of samples.
- Without a careful choice of the anchor point and the bin width, this is often not successful
- The modification of these plot parameters must be done interactively to be effective
- In many cases there are naturals breaks we should seek in order to make the interpretation of the histogram (and highlighting) easier











Spinograms

- When we want to interpret the highlighting in a histogram, we face a hard problem: how to compare the highlighted proportions?
- This is a task we can't perform without guidance by the software, and we would always start to interpret results towards our expectation.
- A **spinogram** is area-proportional just like the histogram, but allows a non-linear x-axis and thus can make all boxes of equal height.
- The comparison is then trivial
- Both views should be used in conjunction