

3. Examining data

Examining data	2
Univariate data	3
QQ-plot	4
Bivariate data	5
Multivariate data	6
General remarks	7

Examining data

- Graphical examination of data is important in all stages of data analysis (see example on overhead)
- Examining data is a good way to get started with R

2 / 7

Univariate data

- Basic univariate displays:
 - ◆ Stem-and-leaf diagram - `stem()`
Good for small data sets.
 - ◆ Histogram - `hist()`
Good for larger data sets.
 - ◆ Density estimation - `plot(density())`
Smoothed version of the histogram.
- To examine whether data comes from a certain distribution:
 - ◆ Histogram/density estimation
 - ◆ Better: QQ-plot = quantile quantile plot
`qqplot()`, `qqnorm()`, `qqline()`
- To summarize main characteristics:
 - ◆ Boxplot - `boxplot()`. Good for outliers, asymmetry, and to compare various distributions.

3 / 7

QQ-plot

- See script section 1.7.1
- Graphical tool to determine whether a sample is consistent with a certain theoretical distribution (usual the normal distribution)
- p^{th} quantile of a distribution:
point x such that $P(X \leq x) = p$ (draw picture).
- p^{th} quantile of a sample:
point x such that $\frac{\#\text{observations} \leq x}{n} \approx p$.
- Each point in a qq-plot corresponds to a probability p :
 - ◆ x -coordinate: p^{th} quantile of theoretical distribution
 - ◆ y -coordinate: p^{th} quantile of sample
- If the sample comes from the theoretical distribution, then the sample and theoretical quantiles are approximately equal. Hence the x and y -coordinates are approximately equal. The qq-plot looks like the line $y = x$. See overheads.

4 / 7

Bivariate data

- Scatterplot - `plot(x,y)`
 - ◆ To show trend:
 - Add nonparametric regression
`lines(loess.smooth(x,y))`
 - ◆ If many points overlap:
 - Jitter the points if they overlap
`jitter()` or add random noise by hand

5 / 7

Multivariate data

- In case of three variables:
 - ◆ 3-d scatterplot
Useful if you can interactively turn the plot around
- In case of more variables:
 - ◆ Scatterplot matrix - `pairs()`

6 / 7

General remarks

- Add informative titles and axis labels
`main`, `xlab`, `ylab`
- Pay attention to the range of the axes
`xlim=c(a,b)`, `ylim=c(a,b)`
- Add a legend when appropriate
`legend`
- Try to optimize the information/ink ratio

7 / 7