



Introduction to R: Statistics

Aron C. Eklund

Center for Biological Sequence Analysis
Technical University of Denmark

October 7, 2008

Outline

Descriptive statistics

Continuous data

Categorical data

Comparing groups

Statistical inference

Correlation

Statistical models

Linear regression

Miscellaneous

Random sampling

Survival analysis

Exercise

Outline

Descriptive statistics

Continuous data

Categorical data

Comparing groups

Statistical inference

Correlation

Statistical models

Linear regression

Miscellaneous

Random sampling

Survival analysis

Exercise

The basics – describing a bunch of numbers

Functions

- `mean(x)`, `median(x)`
- `var(x)`, `sd(x)`
- `summary(x)`, `quantile(x)`
- `min(x)`, `max(x)`, `range(x)`



Describing a bunch of numbers using pictures

Functions

- `hist(x)`
- `density(x)`
- `qqnorm(x)`
- `ecdf(x)`



Outline

Descriptive statistics

Continuous data

Categorical data

Comparing groups

Statistical inference

Correlation

Statistical models

Linear regression

Miscellaneous

Random sampling

Survival analysis

Exercise



Categorical data (factors)

Functions

- `table(x)` – contingency table
- `plot(table(x))`

Outline

Descriptive statistics

Continuous data

Categorical data

Comparing groups

Statistical inference

Correlation

Statistical models

Linear regression

Miscellaneous

Random sampling

Survival analysis

Exercise

Student's t test

What is it?

- Test whether the mean of x is different from the mean of y

Here is one way to do it:

- `x <- swiss$Fertility[swiss$Education > 10]`
- `y <- swiss$Fertility[swiss$Education <= 10]`
- `t.test(x, y)`

Student's t test: shortcuts

The formula interface

- `t.test(swiss$Fertility ~ swiss$Education > 10)`

The data parameter

- `t.test(Fertility ~ (Education > 10), data = swiss)`

If we only want the P value:

- `res <- t.test(Fertility ~ (Education > 10), data = swiss)`
- `res$p.value`

Some other statistical tests

Just to get an idea:

- `grep("test", ls("package:stats"), value = TRUE)`

Outline

Descriptive statistics

Continuous data

Categorical data

Comparing groups

Statistical inference

Correlation

Statistical models

Linear regression

Miscellaneous

Random sampling

Survival analysis

Exercise

Correlation

The `cor` function

- `cor(x, y, method)`
- *method* can be "pearson", "spearman", or "kendall"

Example

- `cor(swiss$Fertility, swiss$Education)`
→ -0.6637889

Many correlations

- *x* and *y* can be matrices

Outline

Descriptive statistics

Continuous data

Categorical data

Comparing groups

Statistical inference

Correlation

Statistical models

Linear regression

Miscellaneous

Random sampling

Survival analysis

Exercise

Linear regression

The `lm` function

- `m <- lm(Fertility ~ Education, data = swiss)`
- `m`

Working with `lm` objects

- `summary(m)` – a nice summary
- `coef(m)` – coefficients of regression
- `residuals(m)` – residuals from the fit

Visualizing the regression

- `plot(Fertility ~ Education, data = swiss)`
- `abline(m, col = "red")`

Multivariate linear regression

Basically the same as before:

- `m2 <- lm(Fertility ~ Education + Catholic, data = swiss)`
- `m2`

Access is as before:

- `coef(m2)`, `summary(m2)`, **etc.**

We might want to do ANOVA:

- `anova(m2)`
- `aov(m2)`

Outline

Descriptive statistics

Continuous data

Categorical data

Comparing groups

Statistical inference

Correlation

Statistical models

Linear regression

Miscellaneous

Random sampling

Survival analysis

Exercise

Random numbers in R

- `set.seed` – leading to reproducible but random numbers
`set.seed(2008)`
- `rnorm` – generate random numbers from the normal distribution
`runif` – generate random numbers from a uniform distribution
- `sample` – randomly choose elements from a vector

Outline

Descriptive statistics

Continuous data

Categorical data

Comparing groups

Statistical inference

Correlation

Statistical models

Linear regression

Miscellaneous

Random sampling

Survival analysis

Exercise

Survival analysis

- `library(survival)`

Statistics Exercise

1. Calculate the Pearson correlation between all possible pairs of columns in `swiss`. The result should be in the form of a matrix.
2. Does the mean *petal length* differ between the various species of `iris`?