BIOINFORMATICA

STATISTIEK IN R_

Week 6

HAN_UNIVERSITY
OF APPLIED SCIENCES

ONDERWERPEN

- Inlezen bestanden
- Grafieken
- Correlatie
- Statistische testen: cor.test()

INLEZEN BESTANDEN

?read.table

```
read.table(file, header = FALSE, sep = "", quote = "\"'", dec = ".", numerals = c("allow.loss", "warn.loss", "no.loss"),
           row.names, col.names, as.is = !stringsAsFactors, na.strings = "NA", colClasses = NA, nrows = -1,
           skip = 0, check.names = TRUE, fill = !blank.lines.skip,  strip.white = FALSE, blank.lines.skip = TRUE,
           comment.char = "#",  allowEscapes = FALSE, flush = FALSE,
           stringsAsFactors = default.stringsAsFactors(), fileEncoding = "", encoding = "unknown", text, skipNul = FALSE)

read.csv(file, header = TRUE, sep = ",", quote = "\"",
         dec = ".", fill = TRUE, comment.char = "", ...)

read.csv2(file, header = TRUE, sep = ";", quote = "\"",
          dec = ",", fill = TRUE, comment.char = "", ...)

read.delim(file, header = TRUE, sep = "\t", quote = "\"",
           dec = ".", fill = TRUE, comment.char = "", ...)

read.delim2(file, header = TRUE, sep = "\t", quote = "\"",
            dec = ",", fill = TRUE, comment.char = "", ...)
```

INLEZEN BESTANDEN

- read.table("C:\\Users\\Gonny Henkes-Veleman\\OneDrive - HAN\\R-statistiek\\Longcapaciteit.txt", header=T, sep="\t")

- read.table(file.choose(), header=T, sep="\t")

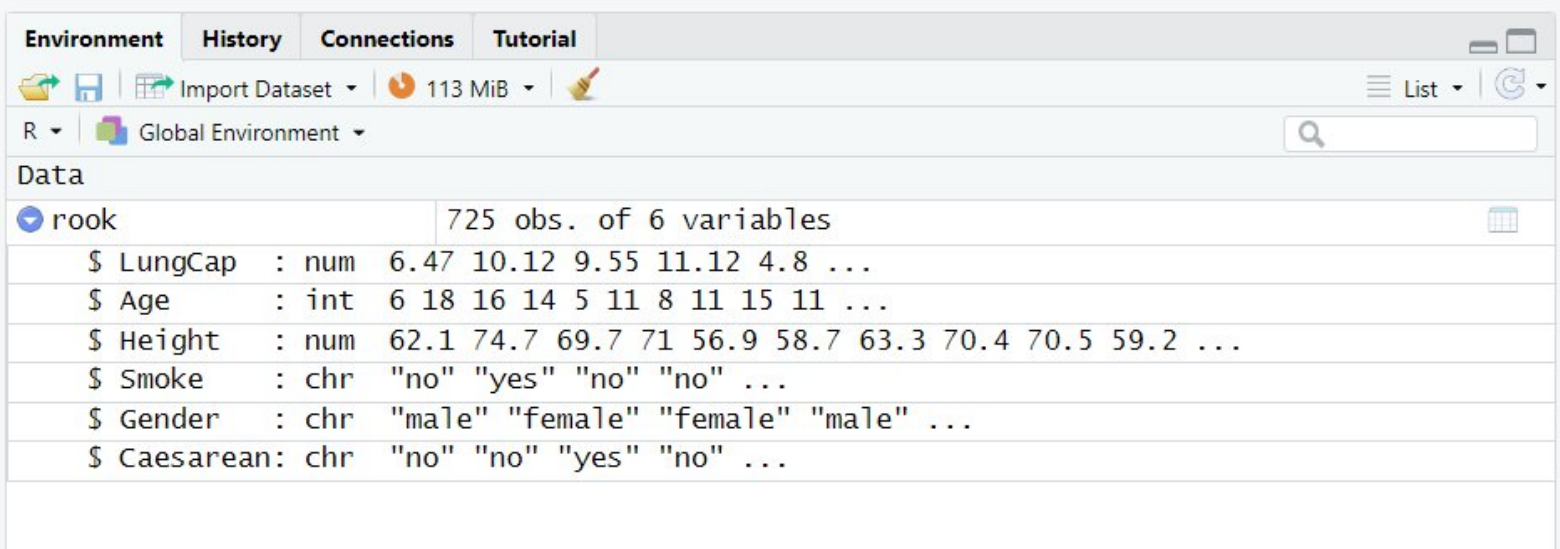
- rook <- read.table(file.choose(), header=T, sep="\t")

INGELEZEN BESTAND

head(rook)

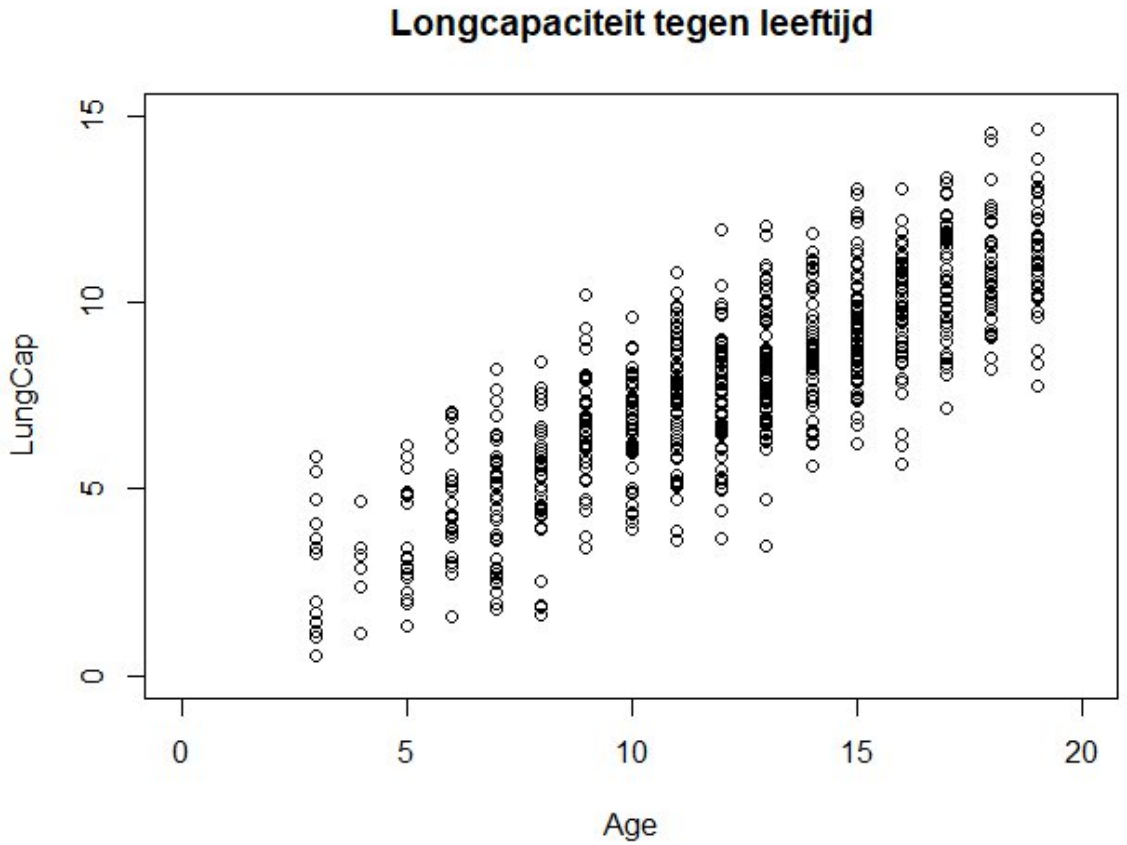
	LungCap	Age	Height	Smoke	Gender	Caesarean
1	6.475	6	62.1	no	male	no
2	10.125	18	74.7	yes	female	no
3	9.550	16	69.7	no	female	yes
4	11.125	14	71.0	no	male	no
5	4.800	5	56.9	no	male	no
6	6.225	11	58.7	no	female	no

attach(rook)



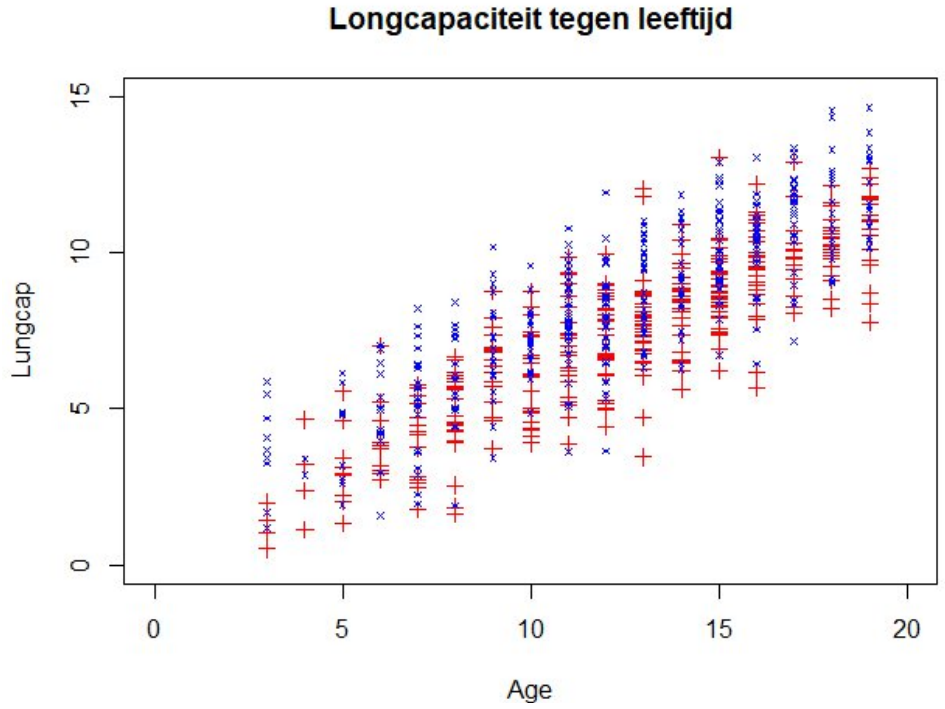
PLOT(LUNGCAP~AGE)

- plot(LungCap~Age, ylim = c(0,15), xlim=c(0,20), main="Longcapaciteit tegen leeftijd")



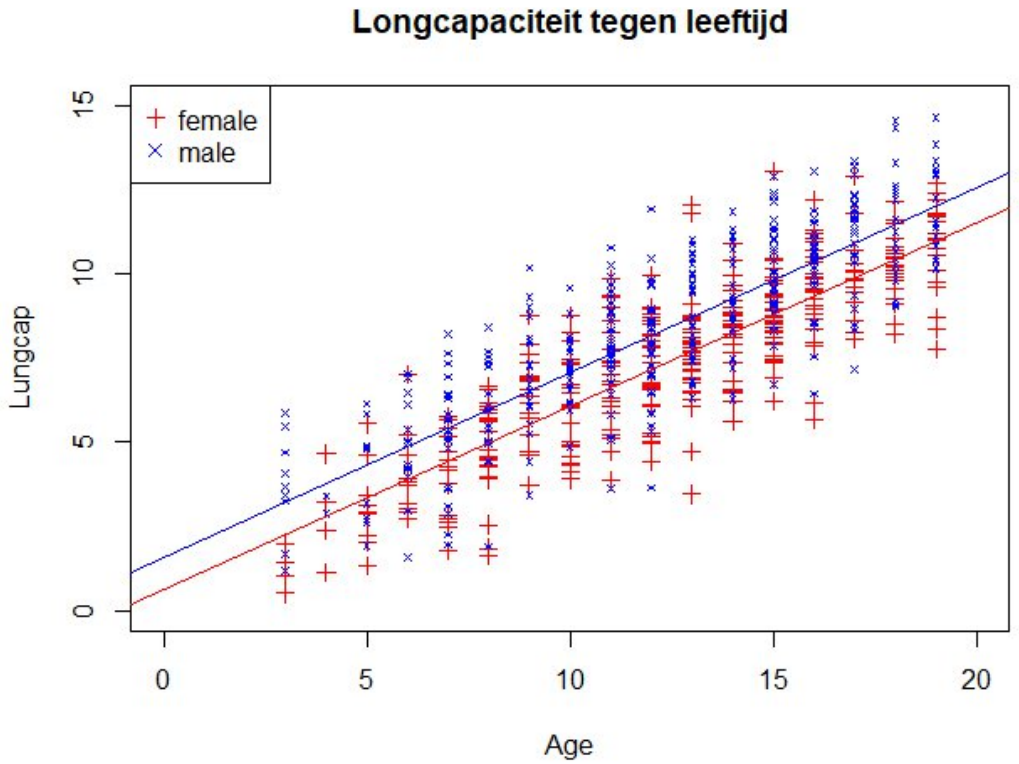
UITPLOTTEN VAN 2 GROEPEN IN 1 GRAFIEK (MANNEN VS VROUWEN)

- plot(LungCap[Gender=="female"]~Age[Gender=="female"], ylim = c(0,15), xlim=c(0,20), main="Longcapaciteit tegen leeftijd", col=2, pch=3, xlab ="Age", ylab="Lungcap")
- points(LungCap[Gender=="male"]~Age[Gender=="male"], ylim = c(0,15), xlim=c(0,20), main="Longcapaciteit tegen leeftijd", col=4, pch=4, xlab ="Age", ylab="Lungcap", cex = 0.5)

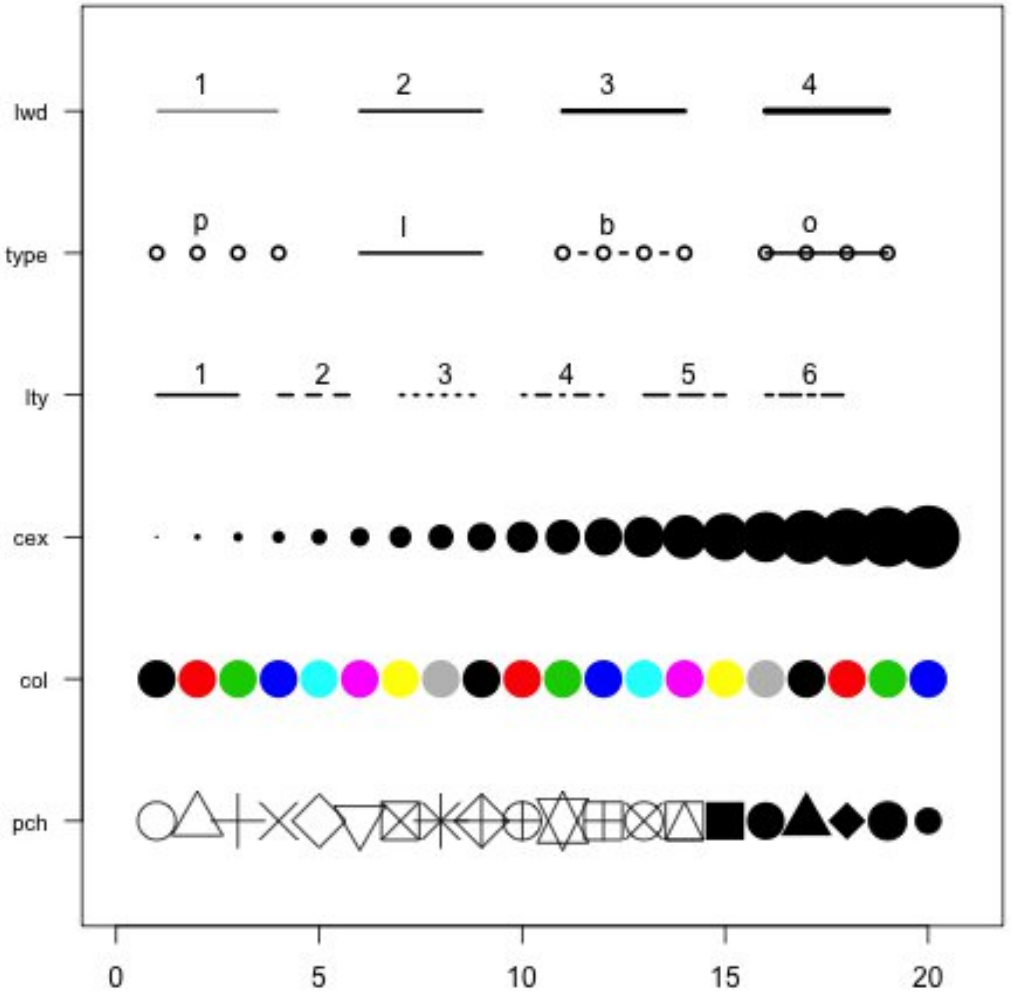


TRENDLIJNEN TOEVOEGEN

- abline(lm(LungCap[Gender=="male"]~Age[Gender=="male"]), col=4)
- abline(lm(LungCap[Gender=="female"]~Age[Gender=="female"]), col=2)
- legend("topleft",legend = c("female","male"), pch = c(3,4), col = c(2,4))



VOORBEELDEN VAN PLOT OPTIES

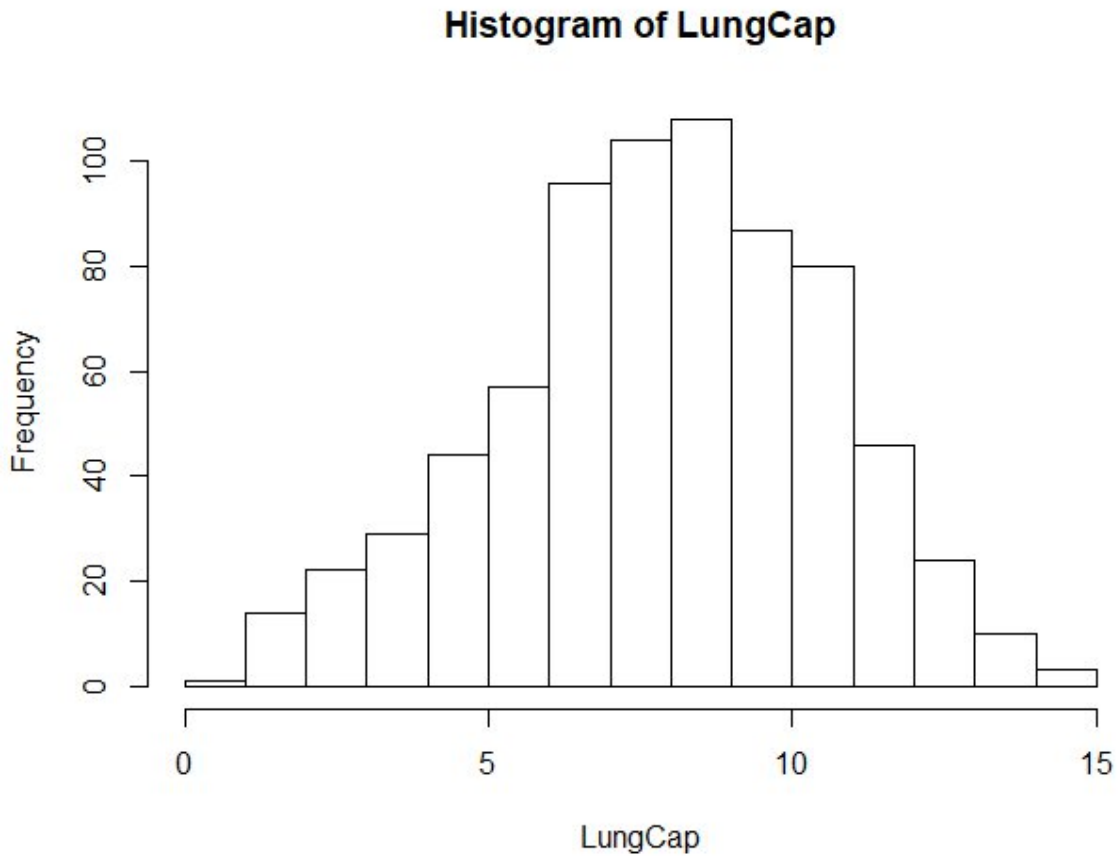


Bron afbeelding: http://www.r-graph-gallery.com/wp-content/uploads/2015/09/6_graph_param_reminder1.png geraadpleegd op 15-03-2019

Legenda plaatsen

x- en y-coördinaten of:
“bottomright”,
“bottom”,
“bottomleft”,
“left”,
“topleft”,
“top”,
“topright”,
“right”,
“center”

HIST(LUNGCAP)

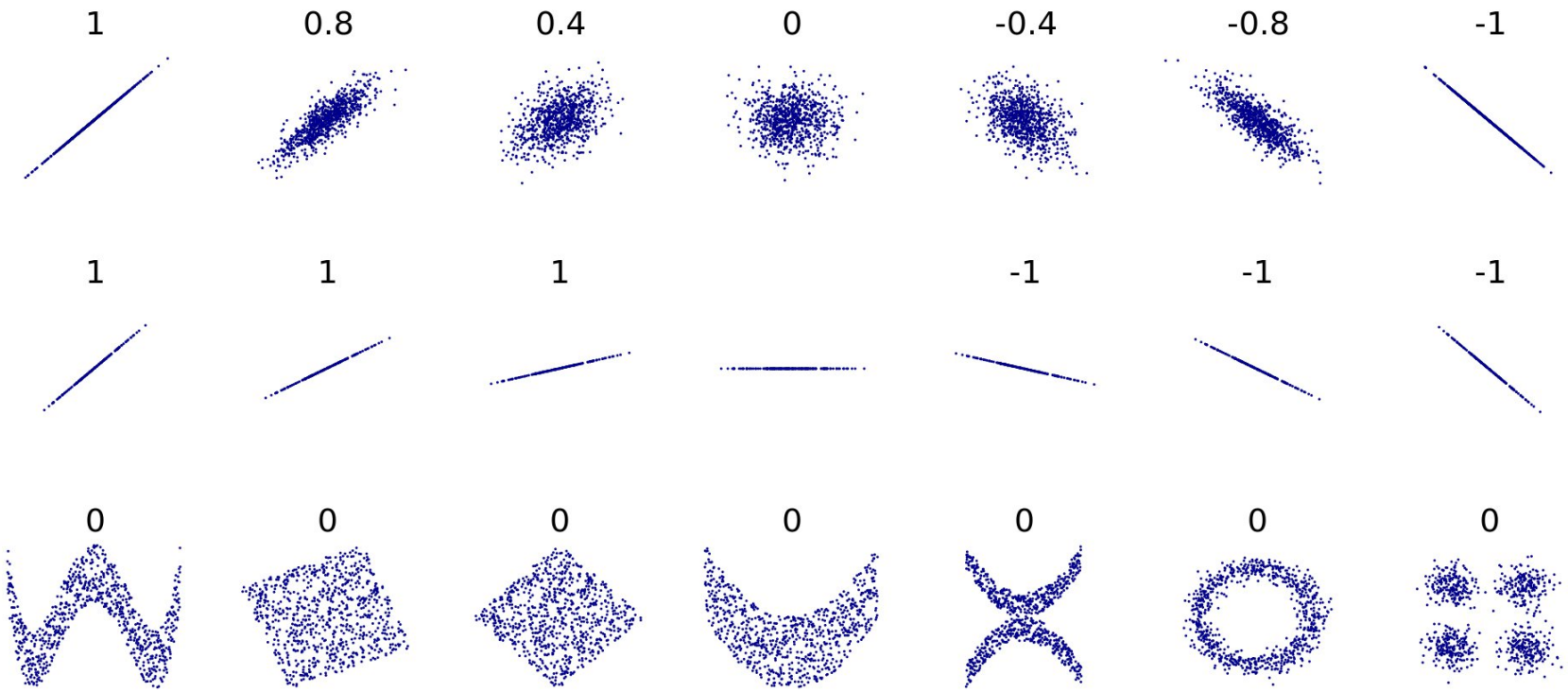


CORRELATIE

- Correlatie is de statistische samenhang tussen twee grootheden.
- De mate van correlatie tussen twee variabelen wordt uitgedrukt in de correlatiecoëfficiënt.
- De waarde daarvan kan variëren tussen –1 en +1.
- Daarbij betekent 0 : geen lineaire samenhang,
  +1 : een perfecte positieve lineaire samenhang
  en –1 : een perfecte negatieve lineaire samenhang.
- Hoe verder de correlatiecoëfficiënt verwijderd is van 0, hoe sterker de correlatie.

- ?cor

CORRELATIE

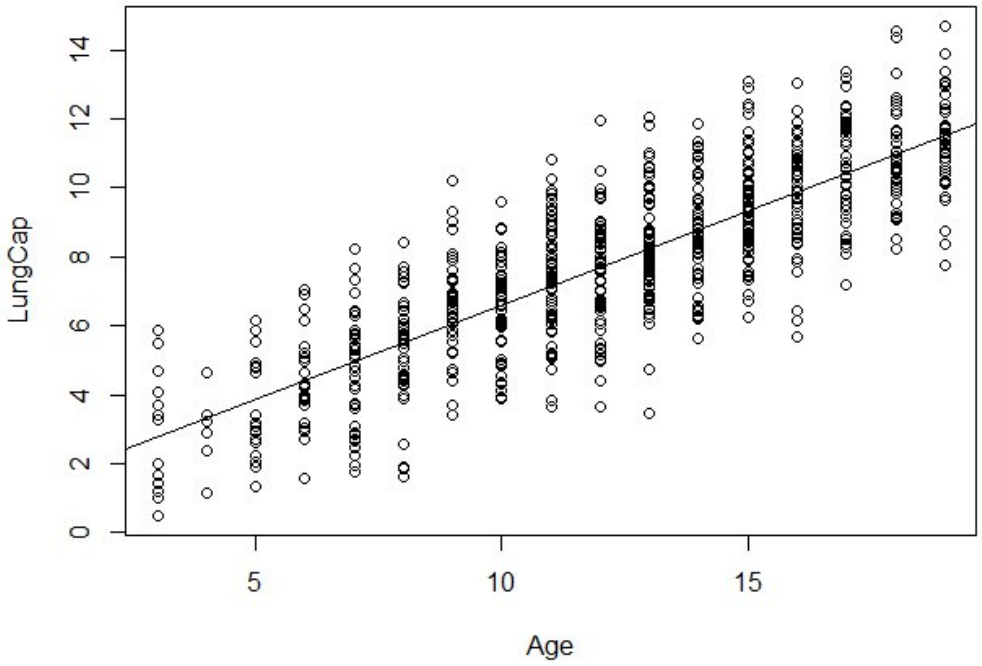


Bron afbeelding: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient#/media/File:Correlation_examples2.svg geraadpleegd 15-03-2019

CORRELATIE

- cor(LungCap,Age)

- plot(LungCap~Age)
- abline(lm(LungCap~Age))



COR.TEST()

Description
Test for association between paired samples, using one of Pearson's product moment correlation coefficient, Kendall's tau or Spearman's rho.

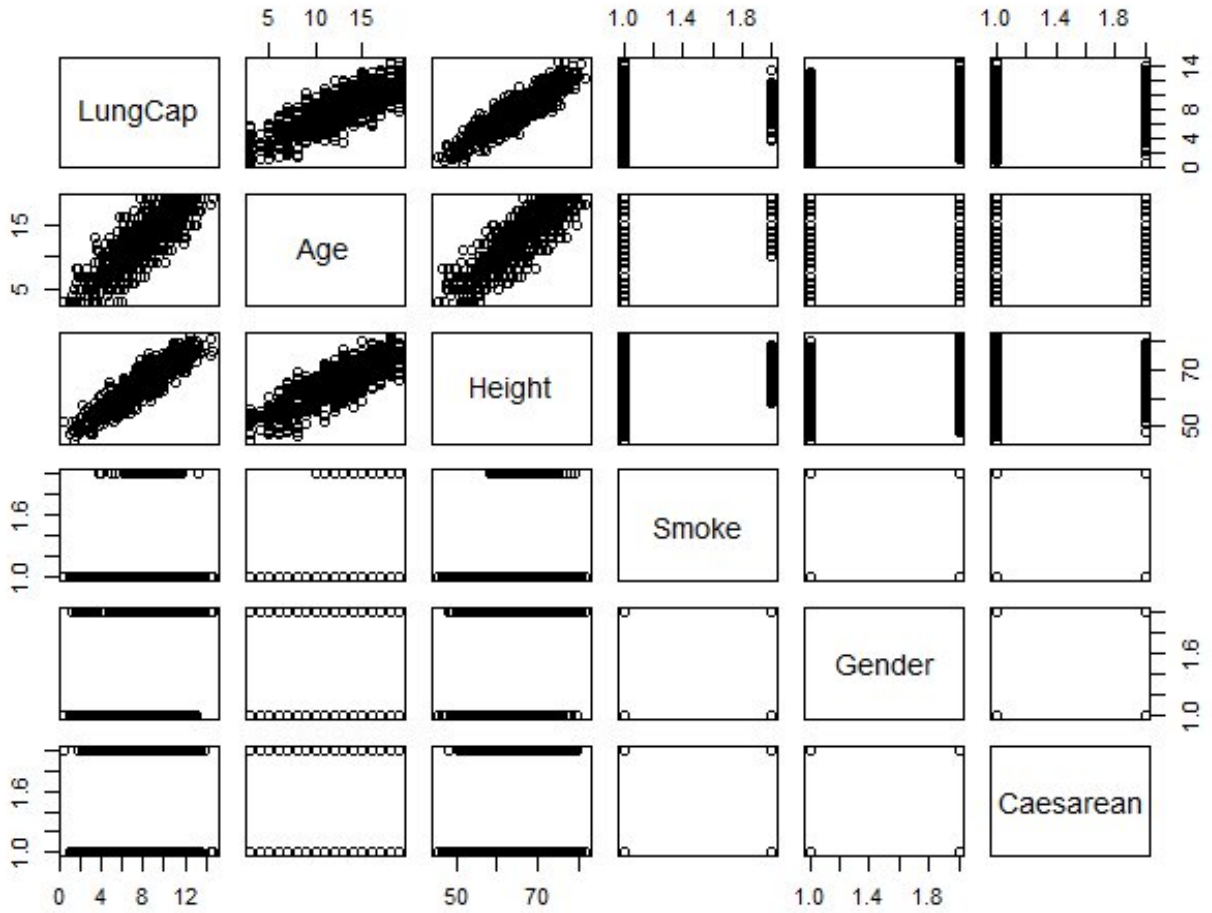
Usage
```
cor.test(x, ...)

## Default S3 method:
cor.test(x, y,
        alternative = c("two.sided", "less", "greater"),
        method = c("pearson", "kendall", "spearman"),
        exact = NULL, conf.level = 0.95, continuity = FALSE, ...)

## S3 method for class 'formula'
cor.test(formula, data, subset, na.action, ...)
```
Arguments

x, y	numeric vectors of data values. x and y must have the same length.
Alternative	indicates the alternative hypothesis and must be one of "two.sided", "greater" or "less". You can specify just the initial letter. "greater" corresponds to positive association, "less" to negative association.
method	a character string indicating which correlation coefficient is to be used for the test. One of "pearson", "kendall", or "spearman", can be abbreviated.

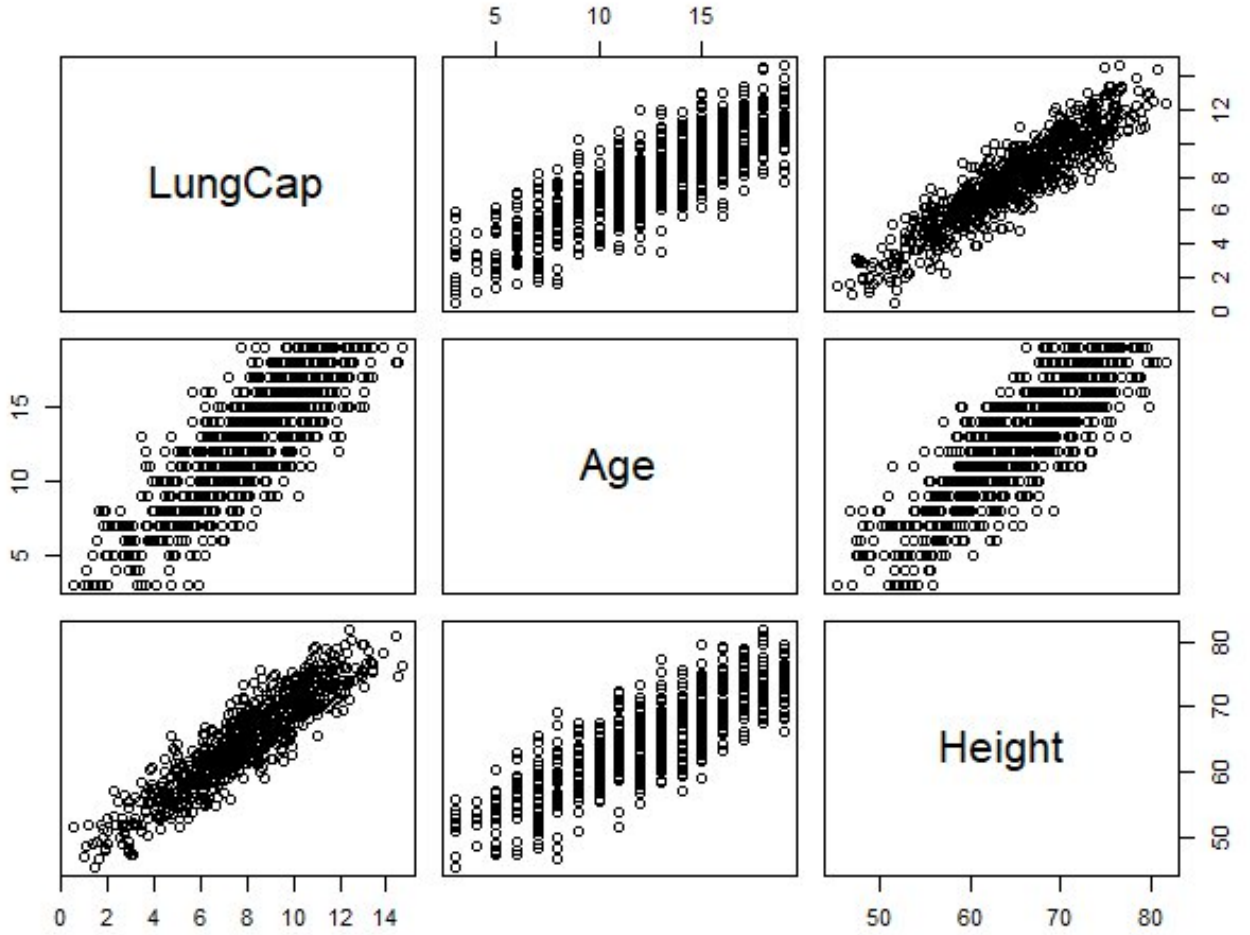
COR.TEST(LUNGCAP, AGE)

```
        Pearson's product-moment correlation

data:  LungCap and Age
t = 38.476, df = 723, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:
 0.7942660 0.8422217

sample estimates:
      cor
0.8196749
```

PLOT(ROOK)



PLOT(ROOK[,1:3])



COR(ROOK[,1:3])

	LungCap	Age	Height
LungCap	1.0000000	0.8196749	0.9121873
Age	0.8196749	1.0000000	0.8357368
Height	0.9121873	0.8357368	1.0000000

OPDRACHTEN

1. Onderzoek met behulp van correlatie of er een lineaire samenhang kan zijn tussen de windsnelheid en de hoeveelheid ozon in de lucht in de dataset airquality.
2. a. Voer een chikwadraattoets voor onafhankelijkheid uit op de variabelen Smoke en Caesarean in de dataset longcapaciteit.
   b. Wat is je H_0?
   c. Geef de bijbehorende p-waarde
   d. met een α van 0.05, H_0 aannemen of verwerpen?
3. Maak een plaatje waarin Height en LungCap tegen elkaar uit gezet zijn. Maak hierin onderscheid tussen Smoke of niet.

BONUSOPDRACHT

- Lees het bestand van weektaak 5, course 2 in (21213.hcdiffs.txt)
- Filter deze zoals beschreven in weektaak 5.

- Selecteer op nog niet bekende non-synonymous varianten en nog niet bekende splice-site varianten. Selecteer alleen kwalitatief goede varianten door alleen varianten te nemen die minstens 5 variant reads hebben en minimaal 20% variatie reads.
  Dit doe je door te filteren op de kolommen SNP state (moet leeg zijn om varianten te selecteren die niet in dbSNP staan), Gene component, Synonymous, variation reads, % variation

VERANTWOORDING

- In deze uitgave is géén auteursrechtelijk beschermd werk opgenomen
- Alle teksten © Gonny Henkes tenzij expliciet externe bronnen zijn aangegeven
- Screenshots op basis van eigen werk auteur en/of vernoemde websites
- Eventuele images zijn opgenomen met vermelding van bron