Correlations

Data Analysis Using R (2017)

Wan Nor Arifin (wnarifin@usm.my), Universiti Sains Malaysia
Website: sites.google.com/site/wnarifin

Contents

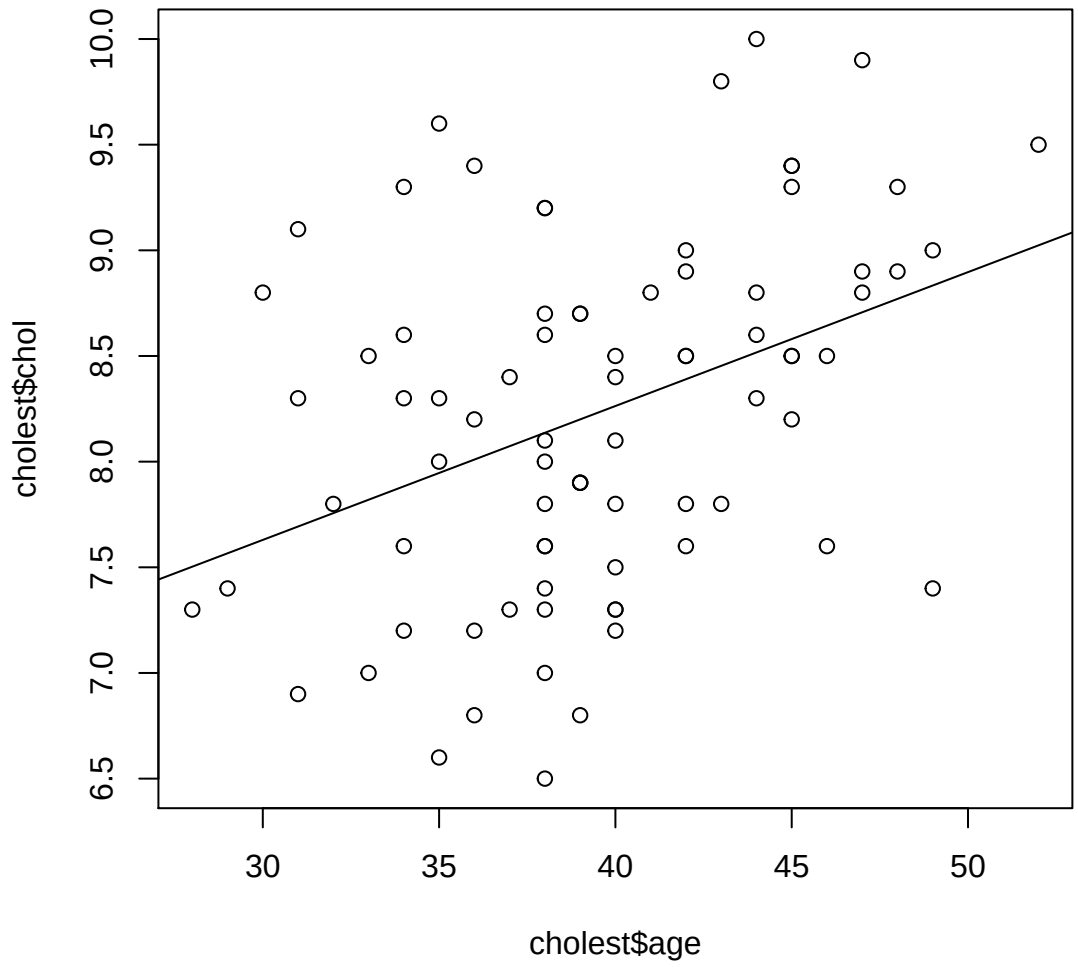
1 Pearson's correlation

- numerical

```
library(foreign)
cholest = read.spss("cholest.sav", to.data.frame = T)
str(cholest)
```

```
## 'data.frame':    80 obs. of  5 variables:
##  $ chol    : num  6.5 6.6 6.8 6.8 6.9 7 7 7.2 7.2 7.2 ...
##  $ age     : num  38 35 39 36 31 38 33 36 40 34 ...
##  $ exercise: num  6 5 6 5 4 4 5 5 4 6 ...
##  $ sex     : Factor w/ 2 levels "female","male": 2 2 2 2 2 2 2 2 2 2 ...
##  $ categ   : Factor w/ 3 levels "Grp A","Grp B",..: 1 1 1 1 1 1 1 1 1 1 ...
##  - attr(*, "variable.labels")= Named chr  "cholesterol in mmol/L" "age in year" "duration of exercis
##   ..- attr(*, "names")= chr  "chol" "age" "exercise" "sex" ...
##  - attr(*, "codepage")= int 65001
```

```
plot(cholest$chol ~ cholest$age)
abline(lm(chol ~ age, cholest))
```




```
cor(cholest$chol, cholest$age)
```

```
## [1] 0.3874574
```

```
cor.test(cholest$chol, cholest$age)
```

```
##
##  Pearson's product-moment correlation
##
## data:  cholest$chol and cholest$age
## t = 3.7119, df = 78, p-value = 0.0003841
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1833492 0.5595401
## sample estimates:
##       cor
## 0.3874574
```

```r
# you can also explore what is offered by psych package
```

2 Spearman's correlation

- ranked

```r
cor(cholest$chol, cholest$age, method = "spearman")
```

```
## [1] 0.3771289
```

```r
cor.test(cholest$chol, cholest$age, method = "spearman")
```

```
## Warning in cor.test.default(cholest$chol, cholest$age, method = "spearman"): Cannot
## compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  cholest$chol and cholest$age
## S = 53143, p-value = 0.0005641
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.3771289
```

3 Others

3.1 phi (binary x binary)

```r
lung = read.csv("lung.csv")
str(lung)
```

```
## 'data.frame':    200 obs. of  2 variables:
##  $ Smoking: Factor w/ 2 levels "no smoking","smoking": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Cancer : Factor w/ 2 levels "cancer","no cancer": 1 1 1 1 1 1 1 1 1 1 ...
```

```r
table(lung)
```

```
##             Cancer
## Smoking      cancer no cancer
##   no smoking     55       113
##   smoking        20        12
```

```r
levels(lung$Smoking)
```

```
## [1] "no smoking" "smoking"
```

```r
lung$Smoking = relevel(lung$Smoking, ref = "smoking")
levels(lung$Smoking)
```

```
## [1] "smoking"    "no smoking"
```

```r
table(lung)
```

```
##             Cancer
## Smoking      cancer no cancer
##   smoking        20        12
##   no smoking     55       113
```

```
library(psych)
phi(table(lung))
```

```
## [1] 0.23
```

3.2 tetrachoric (binary x binary)

- code as 0 1 smoking & cancer as smoking/cancer=1 & no smoking/cancer=0

```
lung1 = NULL
lung1$Smoking = rep(c(1, 0, 1, 0), c(20, 55, 12, 113))
lung1$Cancer = rep(c(1, 1, 0, 0), c(20, 55, 12, 113))
lung1 = as.data.frame(lung1)
table(lung1)
```

```
##        Cancer
## Smoking   0   1
##       0 113  55
##       1  12  20
```

```
tetrachoric(lung1)
```

```
## Call: tetrachoric(x = lung1)
## tetrachoric correlation
##         Smkng Cancr
## Smoking 1.00
## Cancer  0.41  1.00
##
##  with tau of
## Smoking  Cancer
##    0.99    0.32
```

3.3 polychoric (ordinal x ordinal)

- code as ordinal e.g. 1 2 3 rating by two doctors, mild=1 moderate=2 severe=3

```
doc = read.csv("doc.csv")
str(doc)
```

```
## 'data.frame':    121 obs. of  2 variables:
##  $ doc1: int  1 1 1 1 1 1 1 1 1 1 ...
##  $ doc2: int  1 1 1 1 1 1 1 1 1 1 ...
```

```
head(doc)
```

```
##   doc1 doc2
## 1    1    1
## 2    1    1
## 3    1    1
## 4    1    1
## 5    1    1
```

```
## 6    1    1
```

```
table(doc)
```

```
##     doc2
## doc1  1  2  3
##    1 44  4  1
##    2  5 38  5
##    3  1  2 21
```

```
polychoric(doc)
```

```
## Call: polychoric(x = doc)
## Polychoric correlations
##      doc1 doc2
## doc1 1.00
## doc2 0.91 1.00
##
##  with tau of
##          1    2
## doc1 -0.24 0.85
## doc2 -0.22 0.76
```

3.4 biserial (binary x numerical)

```
str(cholest)
```

```
## 'data.frame':    80 obs. of  5 variables:
##  $ chol    : num  6.5 6.6 6.8 6.8 6.9 7 7 7.2 7.2 7.2 ...
##  $ age     : num  38 35 39 36 31 38 33 36 40 34 ...
##  $ exercise: num  6 5 6 5 4 4 5 5 4 6 ...
##  $ sex     : Factor w/ 2 levels "female","male": 2 2 2 2 2 2 2 2 2 2 ...
##  $ categ   : Factor w/ 3 levels "Grp A","Grp B",..: 1 1 1 1 1 1 1 1 1 1 ...
##  - attr(*, "variable.labels")= Named chr  "cholesterol in mmol/L" "age in year" "duration of exercis
##   ..- attr(*, "names")= chr  "chol" "age" "exercise" "sex" ...
##  - attr(*, "codepage")= int 65001
```

```
cholest$sex1 = as.numeric(cholest$sex) - 1  # convert factored sex to 1/0
biserial(cholest$age, cholest$sex1)  # x = cont, y = binary
```

```
##            [,1]
## [1,] -0.4459835
```

```
biserial(cholest$chol, cholest$sex1)
```

```
## Warning in biserialc(x[, j], y[, i], j, i): For x = 1 y = 1 x seems to be dichotomous, not
## continuous
```

```
##      [,1]
## [1,]   -1
```

```
# cannot biserial chol ~ sex1 -> error, not suitable for analysis
```

3.5 polyserial (ordinal x numerical)

```
str(cholest)
```

```
## 'data.frame':    80 obs. of  6 variables:
##  $ chol    : num  6.5 6.6 6.8 6.8 6.9 7 7 7.2 7.2 7.2 ...
##  $ age     : num  38 35 39 36 31 38 33 36 40 34 ...
##  $ exercise: num  6 5 6 5 4 4 5 5 4 6 ...
##  $ sex     : Factor w/ 2 levels "female","male": 2 2 2 2 2 2 2 2 2 2 ...
##  $ categ   : Factor w/ 3 levels "Grp A","Grp B",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ sex1    : num  1 1 1 1 1 1 1 1 1 1 ...
##  - attr(*, "variable.labels")= Named chr  "cholesterol in mmol/L" "age in year" "duration of exercis
##   ..- attr(*, "names")= chr  "chol" "age" "exercise" "sex" ...
##  - attr(*, "codepage")= int 65001
```

```
cholest$categ1 = as.numeric(cholest$categ)
# polyserial(cholest$age, cholest$categ1) cannot perform complex polyserial with psych
```

3.6 Using `polychor`

```
library(polycor)
```

```
##
## Attaching package: 'polycor'
```

```
## The following object is masked from 'package:psych':
##
##     polyserial
```

```
polychor(lung1$Smoking, lung1$Cancer)  # tetrachoric
```

```
## [1] 0.4056244
```

```
polychor(doc$doc1, doc$doc2)  # polychoric
```

```
## [1] 0.9068393
```

```
polyserial(cholest$age, cholest$sex1)  # biserial
```

```
## [1] -0.4487973
```

```
polyserial(cholest$age, cholest$categ1)  # polyserial
```

```
## [1] 0.404529
```

References

Fox, J. (2016). *Polycor: Polychoric and polyserial correlations.* Retrieved from https://CRAN.R-project.org/package=polycor

Revelle, W. (2017). *Psych: Procedures for psychological, psychometric, and personality research.* Retrieved from https://CRAN.R-project.org/package=psych