R language and data analysis: summary statistics

Qiang Shen

Dec 26, 2017

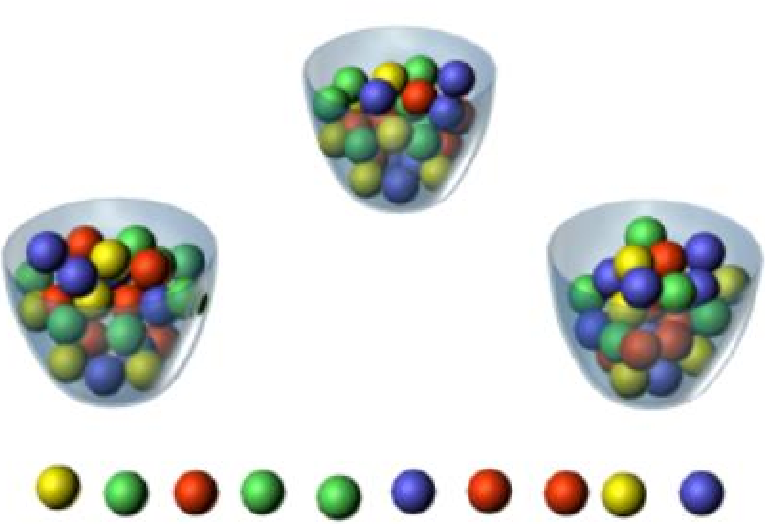
outline

- discriptive statistics
- Frequency and contingency tables
- correlation
- t test

outline

- **discriptive statistics**
- Frequency and contingency tables
- correlation
- t test

sample



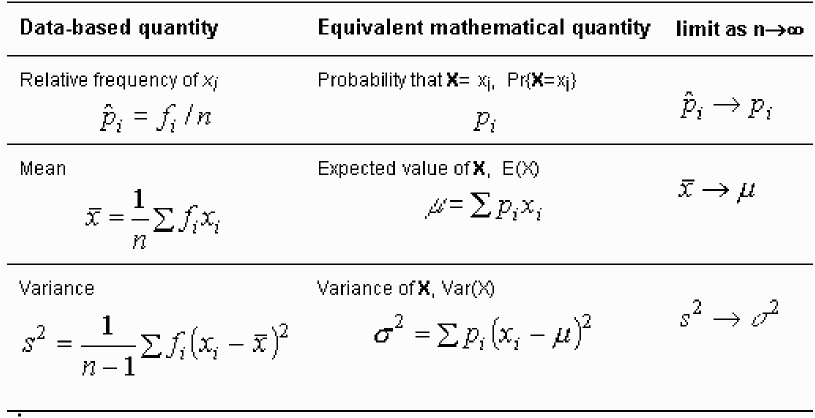
sample

```
str(sample)
n1=100;n2=20
str(sample)
set.seed(1234)
y<-round(100*runif(20))
x<-sample(1:n1,size=n2,replace=F)## change to T for display
 ## apply
x;y
sort(x)
x[order(x)] ##
z=c(10,3,8,1);order(z)
```

average summation (with missing value)

```r
set.seed(1234)
str(sample)
x<-sample(1:100,replace=F)
mean(x)
sum(x)/length(x)
mean(x)
y<-x
n=100
y[sample(1:n,size=20)]<-NA
y
mean(y,na.rm=T)
```

weighted summation



Data-based quantity	Equivalent mathematical quantity	limit as n→∞
Relative frequency of x_j $\hat{p}_i = f_i / n$	Probability that **X**= x_j, Pr(**X**=x_j) p_i	$\hat{p}_i \rightarrow p_i$
Mean $\bar{x} = \frac{1}{n} \sum f_i x_i$	Expected value of **X**, E(X) $\mu = \sum p_i x_i$	$\bar{x} \rightarrow \mu$
Variance $s^2 = \frac{1}{n-1} \sum f_i (x_i - \bar{x})^2$	Variance of **X**, Var(X) $\sigma^2 = \sum p_i (x_i - \mu)^2$	$s^2 \rightarrow \sigma^2$

```r
grades<-c(95,72,87,66)
weights<-c(1/2,1/4,1/8,1/8)
mean(grades)
weighted.mean(grades,weights) ## same as the expectation of
sum(grades*weights)
```

variance and standard deviation

$$\text{Sample Variance} = s^2 = \frac{\Sigma(X - \overline{X})^2}{n - 1}$$

```
x <- c(1, 2, 3, 4, 5, 6, 7, 8)
var(x)
sd(x)  ##sample standard deviation.

n <- length(x)
meanx <- sum(x)/n
css <- sum((x - meanx)**2)
sdx <- sqrt(css / (n-1)) #adjustment of degree of freedom.
sdx
```

basic function

```r
x <- sample(1:100,8,replace=T)
var(x,na.rm=T)
sd(x,na.rm=T)
min(x)
max(x)
median(x)
quantile(x)
quantile(x,c(0.1,0.9))
```

descriptive stats via summary

```
vars <- c("mpg", "hp", "wt")
head(mtcars[vars])
summary(mtcars[vars])
sapply(mtcars[vars],median)
sapply(mtcars[vars],function(x) quantile(x,0.5))
sapply(mtcars[vars],quantile,0.5)
```

descriptive stats via sapply()

```r
mystats <- function(x,na.omit = FALSE) {
   if (na.omit)
  x <- x[!is.na(x)]
  m <- mean(x)
  n <- length(x)
  s <- sd(x)
  skew <- sum((x - m)^3/s^3)/n
  kurt <- sum((x - m)^4/s^4)/n - 3
  return(c(n = n, mean = m, stdev = s, skew = skew, kurtos
}
vars <- c("mpg", "hp", "wt")
sapply(mtcars[vars], mystats,na.omit=T)
mtcars$mpg[2]<-NA;mtcars$hp[c(3,4)]<-NA;mtcars$wt[c(6:9)]<-

# sapply(mtcars[vars], mystats)
# sapply(mtcars[vars], mystats,na.omit=T)
```

function explanation

```
a<-c(1,2,3,NA)
mystats <- function(x,na.rm=F) {
  if (na.rm) x <- x[!is.na(x)]
  return(x)
}
mystats(a)
```

```
[1]  1  2  3 NA
```

```
mystats(a,na.rm=T)
```

```
[1] 1 2 3
```

Descriptive statistics (psych package)

- The SD quantifies scatter — how much the values vary from one another.
- The SE quantifies how precisely you know the true mean of the population. It takes into account both the value of the SD and the sample size.

```r
library(psych)
vars <- c("mpg", "hp", "wt")
describe(mtcars[vars])
sd(mtcars$hp)/sqrt(dim(mtcars)[1])
```

Descriptive statistics by group with aggregate()

```
vars <- c("mpg", "hp", "wt")
aggregate(mtcars[vars], by = list(am = mtcars$am), mean)
aggregate(mtcars[vars], by = list(am = mtcars$am), sd)
```

Summary statistics by group (psych package)

```r
library(psych)
vars <- c("mpg", "hp", "wt")
describeBy(mtcars[vars], mtcars$am)
```

outline

- discriptive statistics
- **Frequency and contingency tables**
- correlation
- t test

Frequency table

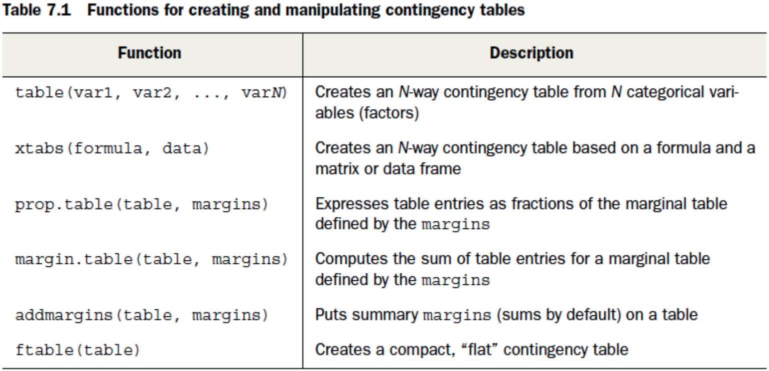


Table 7.1 Functions for creating and manipulating contingency tables

Function	Description
`table(var1, var2, ..., varN)`	Creates an *N*-way contingency table from *N* categorical variables (factors)
`xtabs(formula, data)`	Creates an *N*-way contingency table based on a formula and a matrix or data frame
`prop.table(table, margins)`	Expresses table entries as fractions of the marginal table defined by the `margins`
`margin.table(table, margins)`	Computes the sum of table entries for a marginal table defined by the `margins`
`addmargins(table, margins)`	Puts summary `margins` (sums by default) on a table
`ftable(table)`	Creates a compact, "flat" contingency table

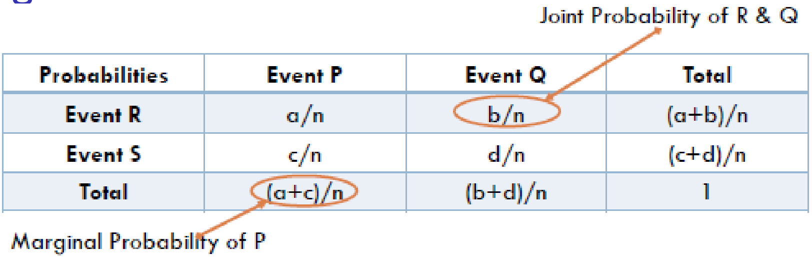
One-way table

```r
library(vcd)
mytable <- with(Arthritis, table(Improved))
mytable
prop.table(mytable)
prop.table(mytable)*100
```

Two-way table

```
table(Arthritis$Treatment,Arthritis$Improved)
mytable<-xtabs(~ Treatment+Improved, data=Arthritis)
mytable
```

margin.table



Joint Probability of R & Q

Probabilities	Event P	Event Q	Total
Event R	a/n	b/n	(a+b)/n
Event S	c/n	d/n	(c+d)/n
Total	(a+c)/n	(b+d)/n	1

Marginal Probability of P

$$P(Q) = P(R,Q) + P(S,Q) = P[(R \cap Q) \cup P(S \cap Q)]$$

```
library(vcd)
mytable<-xtabs(~ Treatment+Improved, data=Arthritis)
prop.table(mytable)
```

```
         Improved
Treatment       None       Some     Marked
  Placebo 0.34523810 0.08333333 0.08333333
  Treated 0.15476190 0.08333333 0.25000000
```

equation

$$p(A|B) = \frac{P(AB)}{P(B)}$$

margin.table

```r
library(vcd)
mytable<-xtabs(~ Treatment+Improved, data=Arthritis)
## marginal probability by row
mytable
margin.table(mytable, 1)
prop.table(mytable, 1) # conditional probability
## marginal probability by column
margin.table(mytable, 2)
prop.table(mytable, 2)
#joint probability
prop.table(mytable)
addmargins(mytable)
##contingency table
addmargins(prop.table(mytable))
#conditional probability
0.34523810/0.51190476; 0.34523810/0.50000000
## conditional probability.
```

xtabs

```r
library(datasets)
UCB.df<-as.data.frame(UCBAdmissions)
UCB.df
```

```
      Admit Gender Dept Freq
1  Admitted   Male    A  512
2  Rejected   Male    A  313
3  Admitted Female    A   89
4  Rejected Female    A   19
5  Admitted   Male    B  353
6  Rejected   Male    B  207
7  Admitted Female    B   17
8  Rejected Female    B    8
9  Admitted   Male    C  120
10 Rejected   Male    C  205
11 Admitted Female    C  202
12 Rejected Female    C  391
```

Three-way contingency table

```
mytable <- xtabs(~ Treatment+Sex+Improved, data=Arthritis)
mytable
ftable(mytable)
# margin.table(mytable, 1)
# margin.table(mytable, 2)
# margin.table(mytable, 3)
# margin.table(mytable, c(1,3))
# ftable(prop.table(mytable, c(1, 2)))
# ftable(addmargins(prop.table(mytable, c(1, 2)), 3))
# ftable(addmargins(prop.table(mytable, c(1, 2)), 3)) * 10
```

Frequency and contingency tables

- table()
- prop.table()
- xtabs()
- margin.table()
- addmargin()

converting a table into a flat file via table2flat

```r
table2flat <- function(mytable) {
  df <- as.data.frame(mytable)
  rows <- dim(df)[1]
  cols <- dim(df)[2]
  x <- NULL
  for (i in 1:rows) {
    for (j in 1:df$Freq[i]) {
      row <- df[i, c(1:(cols - 1))]
      x <- rbind(x, row)
    }
  }
  row.names(x) <- c(1:dim(x)[1])
  return(x)
}
```

Using table2flat with published data

```r
treatment <- rep(c("Placebo", "Treated"), 3)
improved <- rep(c("None", "Some", "Marked"), each = 2)
Freq <- c(29, 13, 7, 7, 7, 21)
mytable <- as.data.frame(cbind(treatment, improved, Freq))
mytable
mydata <- table2flat(mytable)
mydata
```

outline

- discriptive statistics
- Frequency and contingency tables
- **correlation**
- t test

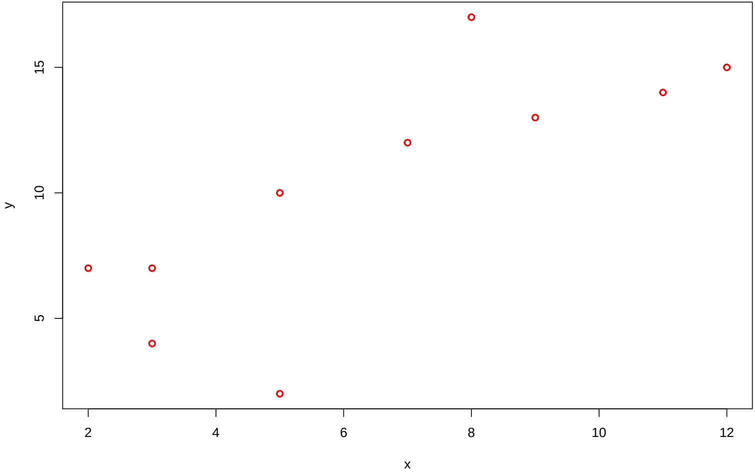
Correlation

- what is correlation?
- correlation coefficient
- statistical inference
- correlation vs. regression
- correlation coeffcient extended
- correlation table and visualization
- missing values
- output of correlation table
- correlation vs. casuality

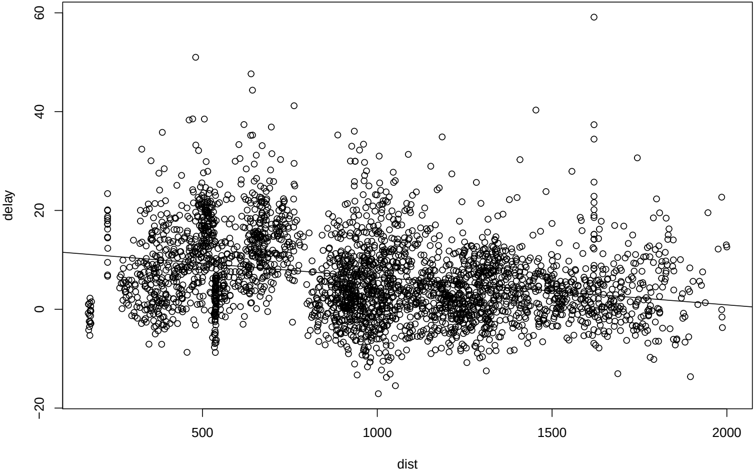
correlation

- **co- relation**

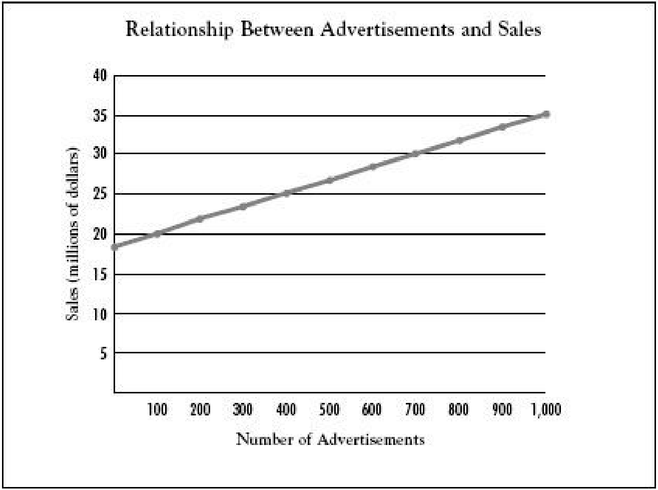
correlation example 1



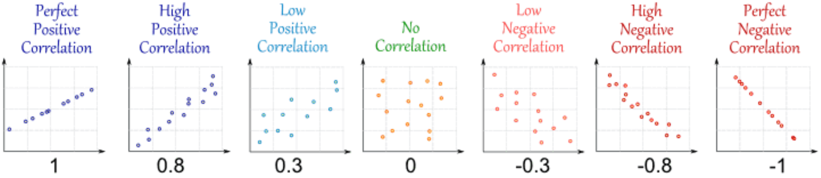
correlation example 2



correlation example 3



correlation



equation

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

```r
a<-c(2,3,5,8,9,11,5,3,12,7)
b<-c(7,4,10,17,13,14,2,7,15,12)
n<-length(a)
mean_a<-mean(a);mean_b<-mean(b)
sum((a-mean_a)*(b-mean_b))/(n-1)/sqrt(var(a)*var(b))
## covariance vs. correlation
sum((a-mean_a)*(b-mean_b))/(n-1);cov(a,b)
cov(a,b)/(sd(a)*sd(b))
cor(a,b)
```

statistical test: cor.test

```r
a<-c(2,3,5,8,9,11,5,3,12,7)
b<-c(7,4,10,17,13,14,2,7,15,12)
cor.test(a,b,method = 'spearman')
```

```

    Spearman's rank correlation rho

data:  a and b
S = 32.293, p-value = 0.005031
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.8042851
```

correlation vs. regression

correlation vs. regression

- 1. direction.

correlation vs. regression

- 1. direction.

correlation vs. regression

- 1. direction.
- 2. qualitative or quantitative.

correlation vs. regression

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

correlation vs. regression

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

correlation vs. regression

$$\hat{\beta}_1 = r\frac{S_y}{S_x}$$

```
x<-c(2,3,5,8,9,11,5,3,12,7)
y<-c(7,4,10,17,13,14,2,7,15,12)
summary(lm(y~x))$coe
```

```
            Estimate Std. Error  t value    Pr(>|t|)
(Intercept) 2.821198  2.2782106 1.238340 0.250694326
x           1.119816  0.3126415 3.581789 0.007169592
```

```
cor(x,y)*sd(y)/sd(x)
```

```
[1] 1.119816
```

correlation coeffcient extended

```
str(cor)
a<-c(2,3,5,8,9,11,5,3,12,7)
b<-c(7,4,10,17,13,14,2,7,15,12)
cor(a,b,method='pearson')
cor(a,b,method='spearman')
cor(a,b,method='kendall')
```

example

```
require(ggplot2)
head(economics)
?economics
# pce:personal consumption expenditures
# pop:total population, in thousands
# psavert:personal savings rate
# unemploy:number of unemployed in thousands
# median duration of unemployment, in week
```

example

```r
library(ggplot2)
with(economics,cor(pce,psavert)) #significance
```

```
[1] -0.837069
```

from correlation table to visualization

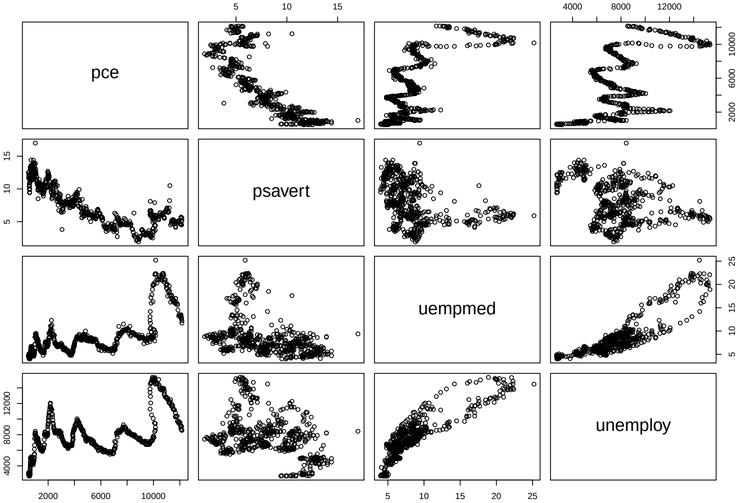
- single correlation—>correlation table

```r
library(ggplot2)
cor(economics[,c(2,4:6)])
```

```
                pce    psavert    uempmed   unemploy
pce       1.0000000 -0.8370690  0.7273492  0.6139997
psavert  -0.8370690  1.0000000 -0.3874159 -0.3540073
uempmed   0.7273492 -0.3874159  1.0000000  0.8694063
unemploy  0.6139997 -0.3540073  0.8694063  1.0000000
```

from correlation table to visualization

- single correlation—>correlation table–>data visualization

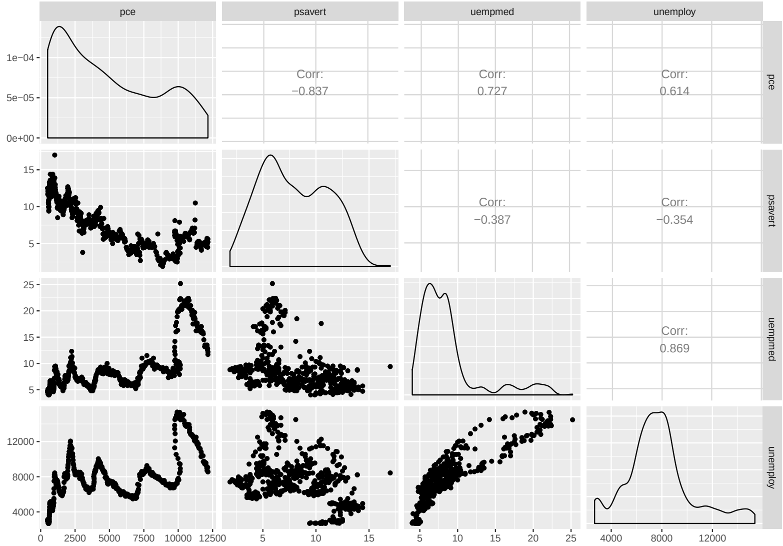


from correlation table to visualization

- single correlation—>correlation table–>data visualization

```r
library(ggplot2)
pairs(economics[,c(2,4:6)], pch = 21) ##base
```

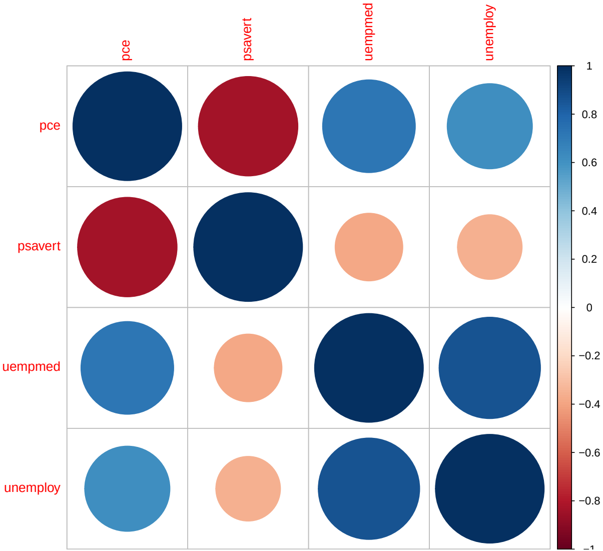
Plot: GGally


pce
psavert
uempmed
unemploy
1e−04
5e−05
0e+00
Corr:
−0.837
Corr:
0.727
Corr:
0.614
Corr:
−0.387
Corr:
−0.354
Corr:
0.869
15
10
5
25
20
15
10
5
12000
8000
4000
0
2500
5000
7500
10000
12500
5
10
15
5
10
15
20
25
4000
8000
12000
pce
psavert
uempmed
unemploy

Plot: GGally

```r
library(ggplot2)
library(GGally,quietly=T)
# GGally::ggpairs(economics[,c(2,4:6)],params=list(labelSi
ggpairs(economics[,c(2,4:6)])
with(economics,plot(pce,psavert))
# data(tips,package='reshape2')
# head(tips)
# GGally::ggpairs(tips)
```

Plot: corrplot



Plot: corrplot

```r
library('corrplot')
econCor<-cor(economics[,c(2,4:6)])
corrplot(econCor, method = "circle") #plot matrix
```

visualization summary

- base::pairs
- GGally::ggpairs
- corrplot::corrplot

missing value

```
str(cor)
m<-c(9,9,NA,3,NA,5,8,1,10,4)
n<-c(2,NA,1,6,6,4,1,1,6,7)
p<-c(8,4,3,9,10,NA,3,NA,9,9)
q<-c(10,10,7,8,4,2,8,5,5,2)
r<-c(1,9,7,6,5,6,2,7,9,10)
theMat<-cbind(m,n,p,q,r)
theMat
cor(theMat)
cor(theMat,use='everything')
cor(theMat,use='all.obs')
cor(theMat,use='complete.obs')
cor(theMat,use='na.or.complete')
data<-na.omit(theMat)
data
cor(data)
class(data)
```

missing value

```r
m<-c(9,9,NA,3,NA,5,8,1,10,4)
n<-c(2,NA,1,6,6,4,1,1,6,7)
p<-c(8,4,3,9,10,NA,3,NA,9,9)
q<-c(10,10,7,8,4,2,8,5,5,2)
r<-c(1,9,7,6,5,6,2,7,9,10)
theMat<-cbind(m,n,p,q,r)
cor(theMat,use="pairwise.complete.obs")
cor(theMat[,c('m','n')],use='complete.obs')
cor(theMat[,c('m','p')],use='complete.obs')
```

statistical inference

- single correlation—>correlation table

```
Call:corr.test(x = iris[, 1:4], use = "complete")
Correlation matrix
             Sepal.Length Sepal.Width Petal.Length Petal.Wi
Sepal.Length         1.00       -0.12         0.87        0
Sepal.Width         -0.12        1.00        -0.43       -0
Petal.Length         0.87       -0.43         1.00        0
Petal.Width          0.82       -0.37         0.96        1
Sample Size
[1] 150
Probability values (Entries above the diagonal are adjusted
             Sepal.Length Sepal.Width Petal.Length Petal.Wi
Sepal.Length         0.00        0.15            0
Sepal.Width          0.15        0.00            0
Petal.Length         0.00        0.00            0
Petal.Width          0.00        0.00            0
```

statistical inference

- single correlation—>correlation table

```r
# cor.test(iris[,1:4])
library(psych)
corr.test(iris[,1:4], use = "complete")
```

output

```r
source("corstartl.R")
corstarsl(economics[,c(2,4:6)])
```

```
              pce  psavert  uempmed
pce
psavert  -0.84***
uempmed   0.73*** -0.39***
unemploy  0.61*** -0.35***  0.87***
```

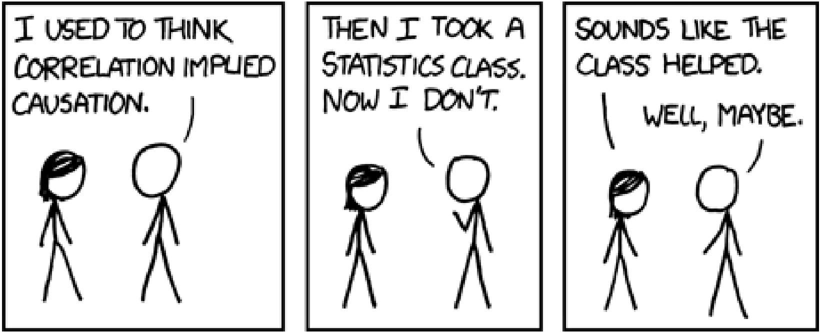
```r
library(xtable);kable(corstarsl(swiss[,1:4]))
```

	Fertility	Agriculture	Examination
Fertility			
Agriculture	0.35*		
Examination	-0.65***	-0.69***	
Education	-0.66***	-0.64***	0.70***

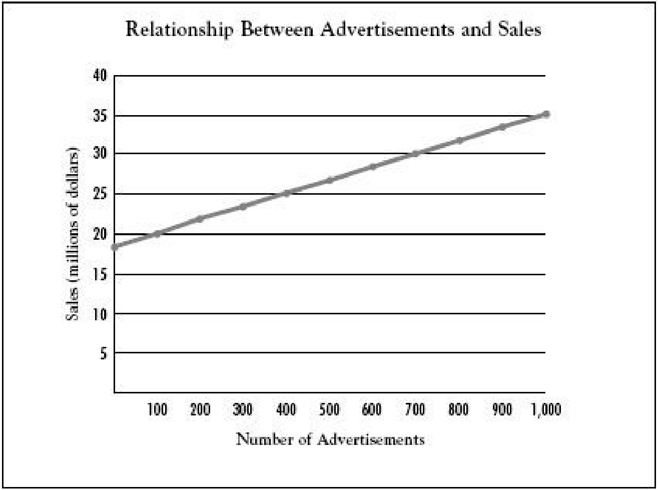
output

```r
source("corstartl.R");corstarsl(economics[,c(2,4:6)])
xtable(corstarsl(swiss[,1:4]))
```

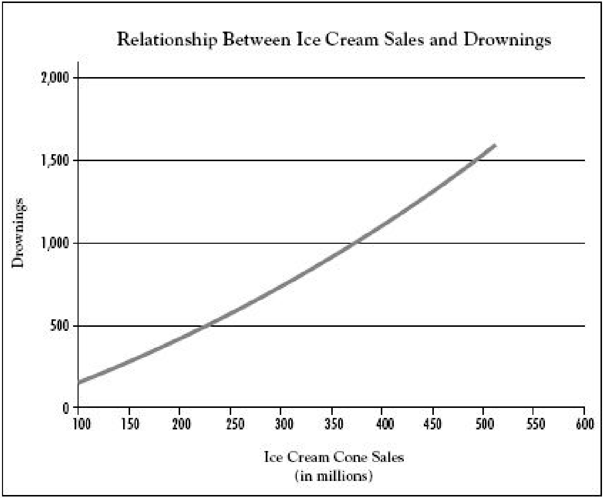
correlation vs. causality.


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

correlation example 3



correlation vs. causality.



correlation vs. causality.

"A zillion things can correlate with each other, depending on how you structure the data and what you compare. To discern meaningful correlations from meaningless ones, you often have to rely on some causal hypothesis about what is leading to what. You wind up back in the land of human theorizing." –David Brooks

equation

- cor(a,b)
- cov(a,b)
- pairs
- cor.test(a,b) vs.psych::corr.test()

correlation

- what is correlation?
- correlation coefficient
- statistical inference
- correlation vs. regression
- correlation coeffcient extended
- correlation table and visualization
- missing values
- output of correlation table
- correlation vs. casuality

outline

- discriptive statistics
- Frequency and contingency tables
- correlation
- **t test**

t test

- single sample t test
- independent sample t test
- paired t test

sample data

```
  extra group ID
1   0.7     1  1
2  -1.6     1  2
3  -0.2     1  3
4  -1.2     1  4
5  -0.1     1  5
6   3.4     1  6
```

```
'data.frame':   20 obs. of  3 variables:
 $ extra: num  0.7 -1.6 -0.2 -1.2 -0.1 3.4 3.7 0.8 0 2 ...
 $ group: Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 1 1 1
 $ ID   : Factor w/ 10 levels "1","2","3","4",..: 1 2 3 4 5
```

one-sample t test

- t.test

```r
head(sleep)
```

```
  extra group ID
1   0.7     1  1
2  -1.6     1  2
3  -0.2     1  3
4  -1.2     1  4
5  -0.1     1  5
6   3.4     1  6
```

```r
t.test(sleep$extra,mu=0)
```

```

    One Sample t-test

data:  sleep$extra
```

independent two-sample t-test

```r
# library(tidyr);sleep_wide<-spread(sleep,group,extra)
library(reshape2)
sleep_wide<-dcast(sleep,  ID~group, value.var = "extra")
names(sleep_wide)<-c('ID','group1','group2')
## Welch t-test
 #long format
t.test(extra ~ group, sleep) # tilde
 #wide format
t.test(sleep_wide$group1, sleep_wide$group2)

# Student t-test
t.test(extra ~ group, sleep, var.equal=TRUE)
```

Paired-sample t-test

```
# wide format
# library(tidyr);sleep_wide<-spread(sleep,group,extra)
library(reshape2)
sleep_wide<-dcast(sleep,  ID~group, value.var = "extra")
names(sleep_wide)<-c('ID','group1','group2')
t.test(sleep_wide$group1, sleep_wide$group2, paired=TRUE)

# long format
# Sort by group then ID
sleep <- sleep[order(sleep$group, sleep$ID), ]
sleep
# Paired t-test
t.test(extra ~ group, sleep, paired=TRUE)

## equivalent to testing whether difference between
## each pair of observations has a population mean of 0.
t.test(sleep_wide$group1 - sleep_wide$group2, mu=0,
```

t test

- single sample t test
- independent sample t test
- paired t test

summary

- discriptive statistics
- Frequency and contingency tables
- correlation
- t test