《统计学》

图书基本信息

书名：《统计学》

13位ISBN编号：9787300187498

出版时间：2014-2

作者：吴喜之

页数：164

《统计学》

内容概要

本书通过大量例子，用简单明了的语言介绍了传统统计学的所有基本概念及方法。 书中还专门用一章的内容来介绍非常重要而实用的机器学习的回归分及类方法。本书采用的计算机语言是多年来在国际上使用排名第一的免费开源软件—r。读者在学完本书之后，能够准确理解统计最重要的基本概念,并能用计算机处理各种数据。

《统计学》

作者简介

吴喜之，北京大学数学力学系学士，美国北卡罗来纳大学教堂山分校(UNC-Chapel Hill)统计系博士。中国人民大学统计学院教授，博士生导师。曾在美国加利福尼亚大学戴维斯分校(UC-Davis)、北卡罗来纳大学教堂山分校(UNC-Chapel Hill)、北卡罗来纳大学夏洛特分校(UNC-Charlotte)、加利福尼亚大学伯克利分校(UC-Berkeley)、南开大学、中国人民大学、北京大学、中山大学、四川大学等十余所著名学府执教。

《统计学》

书籍目录

《统计学》

章节试读

1、《统计学》的笔记-第 2 章 数据及其模式

```
      ## -------------------------- System Setting Start --------------------------
## 01. Choose Directory AND Set As Working Directory
#path1 &lt;- "D:/06-Training_SelfPromotion/002_00-Data_Analysis_Presentation/"
#path2 &lt;- "06-Statistics_WuXiZhi/Code_Data_Notes/"
#path3 &lt;- "Data/Chapter02/"
#WorkingDirectory &lt;- paste(path1, path2, path3, sep = "")
WorkingDirectory &lt;- choose.dir()
setwd(WorkingDirectory)
## FilePath &lt;- file.choose(WorkingDirectory)
## FilePath &lt;- file.choose(getwd())
## dir(getwd())
## 02. loading R Script and input data
source("MyRCode.R")
#rawData &lt;- read.csv(file.choose(WorkingDirectory), header = F, sep = ',')
rawData &lt;- read.csv(file.choose(getwd()), header = F, sep = ',')
## 03. Record Current Date Time
DateTime &lt;- format(Sys.time(), "%Y-%m-%d %H-%M-%S")
## -------------------------- System Setting Ended --------------------------

## ------------------------- File Name Module Start -------------------------
# get the file name from the choosen file
FilePath &lt;- file.choose(WorkingDirectory)
# [1] "F:\\JMP-DOE-Statistics\\Statistics-R\\Code_Data_Notes\\
# Chapter01-Code_Answer_Notes.r"
# Split File Path 2 List By 2 Backslash \\
FilePathList &lt;- strsplit(FilePath,"\\\\")
# get the number of elements from list File Path List
FilePathListLength &lt;- length(FilePathList[[1]])
# get the last element that is the file name
FileName &lt;- FilePathList[[1]][FilePathListLength]
sprintf("The File Name is: ")
sprintf("%s", FileName)

## ------------------------- File Name Module Ended -------------------------


# e.g. 2.1  names.txt
#   姓名 性别  教育  籍贯 年龄 观点
#   王芳  女  大学  北京 62  是
# 李泽娜  女  大学  山东 57  否
#   刘伟  男  中学 河北 19  是
#   刘东  男  大学  河北 29  是
#   李锐  男  中学  北京  15  否
# 张节福  男  研究生 北京 41  是
# 赵思雨  男  研究生  山东 48  否
```

```
# 唐慧聪  女  中学  山东 47  否
# 熊爱珊  女 大学   山东 23  是
#   王冰  男  大学 河北  23  否

w=read.table("Data/Chapter02/names.txt",header=T)
# filter variable 1, 4, 5: 姓名 籍贯 年龄
v=w[,-c(1,4,5)]
tt=table(v)
# flat table
ftable(tt)
# 观点 否 是
# 性别 教育
# 男  大学      1  1
# 研究生    1  1
# 中学      1  1
# 女  大学      1  2
# 研究生    0  0
# 中学      1  0

ftable(tt,col.vars=c(1,3))
# equal statement to ftable(tt,col.vars=c(1,3))
ftable(tt,col.vars=c("性别","观点"))
#      性别 男   女
#      观点 否 是 否 是
# 教育
# 大学      1  1  1  2
# 研究生    1  1  0  0
# 中学      1  1  1  0

```{r }
# 2 dimension flat table
ftable(tt,col.vars=3,row.vars=2)
#      观点 否 是
# 教育
# 大学      2  3
# 研究生    1  1
# 中学      2  1

ftable(tt, row.vars=3, col.vars=2)
#    教育 大学 研究生 中学
# 观点
# 否        2    1   2
# 是        3    1   1

```

# xtabs() cross table
xtabs(~., v)   # same as tt, same as table(v)
# , , 观点 = 否
```

```
#
#    教育
# 性别 大学 研究生 中学
#  男   1    1   1
#  女   1    0   1
#
# , , 观点 = 是
#
#    教育
# 性别 大学 研究生 中学
#  男   1    1   1
#  女   2    0   0
xtabs(~性别+观点, v)
#    观点
# 性别 否 是
#  男 3 3
#  女 2 2


# e.g. 2.2
# read data
w=read.csv("Data/Chapter02/Rich.csv",header=T)
#print head 20 lines
w[1:20, ]
head(w, 20)
# print 3x3
w[1:3, 1:3]
# names(w), write muti-statement in one line
names(w);
# [1] "Rank"     "Name"     "Net.Worth" "Age"      "Source"   "Residency"
summary(w);str(w)
#sort by Residency and retrieve top 10
v=rev(sort(table(w[,6])))[1:10]
#sort by Source and retrieve top 10
u=rev(sort(table(w[,5])))[1:10]
# setting margin
rep(0, 4)
op &lt;- par(mar = rep(0, 4))
#op
# $mar
# [1] 5.1 4.1 4.1 2.1
plot.new()
par(op) #setting margin
# side-by-side 1 Row X 2 Column
par(mfrow=c(1,2))
# pie chart Top 10 by residency
pie(v,cex.names=.8,main="by residency")
# pie chart Top 10 by source
pie(u,cex.names=.8,main="by source")
```

```
# paralleling 2 Row X 1 Column
par(mfrow=c(2,1))
barplot(v,cex.names=.8,main="by residency")
barplot(u,cex.names=.8,main="by source")
#set back to default 1 x 1
par(mfrow=c(1,1))


#e.g. 2.3 global top 2000 Company
w=read.csv("Data/Chapter02/Forbes2000.csv",header=T)
names(w);summary(w)
# [1] "Rank"        "Company"     "Country"     "Sales"
# [5] "Profits"     "Assets"      "Market.Value"

#draw 4 histogram, data do log transform
par(mfrow=c(2,2))
for(i in 4:7){
  hist(log(w[,i]),main=paste("log",names(w)[i]),xlab="")
  rug(log(w[,i]))
}

# box plot, box-and-whisker plot
# draw china companies market value, horizontal positioned
par(mfrow=c(1,1))
boxplot(w[w[,3]=="China",7],horizontal=T,
       main="market value")
rug(w[w[,3]=="China",7])

stem(w[w[,3]=="China",7])

v=w[1:100,]
plot(v$Assets,v$Sales,pch=1,col=1,
    xlim=c(-500,3000),ylim=c(0,600),
    cex=sqrt(v$Profits))
identify(v$Assets,v$Sales,labels=v$Company)

C=w[w[,3]=="China",]
G=w[w[,3]=="Germany",]
par(mfrow=c(1,2))
hist(C$Market.Value,20,main="histogram of market value(China)",
    xlab="market value",ylab="density",
    col=3,prob=T,ylim=c(0,0.07))
lines(density(C$Market.Value),lwd=2)
hist(G$Market.Value,20,main="histogram of market value(Germany)",
    xlab="market value",ylab="density",
    col=3,prob=T,ylim=c(0,0.07))
lines(density(G$Market.Value),lwd=2)
par(mfrow=c(1,1))
```

```
w=scan("soi.txt")
w=ts(w,start=1950,frequency=12)
plot(w,ylab="SOI")
abline(h=0,lty=2)
title("the southern oscillation index 1950-1995")

w=read.table("Data/Chapter02/USIP.txt",header=T)
v=ts(w[,-c(1,2)],start=c(1947,1),frequency=12)
ts.plot(v,lty=1:8,col=1:8,ylab="indices",xlab="time")#若用plot，得出8个图而不是放在一起的一个图
title("US indices of industrial production")
legend("topleft",legend=names(w)[3:10],lty=1:8,col=1:8)

w=read.csv("Data/Chapter02/Salinity.csv",header=T)
attach(w)
plot(Waterflow,Salinity)
title=("Salinity")
identify(Waterflow,Salinity,labels=rownames(w))
detach(w)

x=scan("Data/Chapter02/income.txt")#若用read.table函数，则x是数据框，不能直接求平均，需将x向量化
mean(x)
median(x)

w=read.table("Data/Chapter02/Acorn.txt",sep=",",header=T)#数据是一逗号隔开的
mean(w[,4]);median(w[,4])
hist(w[,4],prob=T,
   xlab="Acorn size",main="Acorn size")
rug(w[,4])
arrows(5,0.2,median(w[,4]),0)
arrows(10,0.2,mean(w[,4]),0)
text(locator(2),c("meadian=1.8","mean=3.34"))

C=w[w[,2]=="California",4]
A=w[w[,2]=="Atlantic",4]
summary(C);summary(A)
fivenum(C);fivenum(A)
boxplot(Acorn_size~Region,w,horizontal=T)
par(mfrow=c(1,2))
hist(C,12,prob=T,
   xlab="Acorn size",main="California",
   xlim=c(0,18),ylim=c(0,0.5))
rug(C);lines(density(C))
hist(A,12,prob=T,
   xlab="Acorn size",main="Atlantic",
   xlim=c(0,18),ylim=c(0,0.5))
rug(A);lines(density(A))
par(mfrow=c(1,1))
```

```
#习题
#1
w=read.csv("Data/Chapter02/Old.csv",header=T)
dim(w)
nrow(w)-nrow(na.omit(w))
names(w)
hist(w[,5]);boxplot(w[,5],horizontal=T)
table(w[,5])
barplot(table(w[,5]))
pie(table(w[,5]))
stem(w[,5])
v=na.omit(w)
v[v[,5]&gt;=15,1]
v[order(v[,5],decreasing=T),]
v[v[,1]=="China",]
median(v[,5])
mean(v[,5])

#3
w=read.table("Data/Chapter02/chismoke.dat",header=T)
names(w)
x0=xtabs(Count~.,w);x0;dim(x0)
attributes(x0)
x1=xtabs(Count~.,w[,2:4]);x1
x2=xtabs(Count~.,w[,-2]);x2
#实验数据

#4
(164-157)/qnorm(0.9)
pnorm(175,167.8,5.61819)
1-pnorm(226,167.8,5.61819)
1-pnorm(160,167.8,5.61819)
pnorm(162,157,5.462129)
1-pnorm(180,167.8,5.61819);1-pnorm(175,157,5.462129)
```

2、《统计学》的笔记-通过来学统计很容易

```
    #
# Copyright (c) 2010-~ siqin.hou All rights reserved.
#
# This source code is released for free distribution under the terms of the
# GNU General Public License
#
# ---Author: siqin&lt;siqin.hou@gmail.com&gt;
# Date-Time: 2014-11-30 23:24:09
# File-Name: Chapter01-Notes.r
# --Version: 1.0
# -Function: Statistics-Application base on R---XiZhi_Wu
```

```
#

# install.packages("package_name","dir")
# package_name:package name, take care upper case, lower case.
# dir:the directory of the package, default directory is ../library.
# change the dir parameter, can choose the directory which install the package.
# e.g. package mvtnorm install to D:/DM/r/R-2.15.2/library/
# install.packages("mvtnorm","D:/DM/r/R-2.15.2/library/")

# load package
# library("package_name"),
# require("package_name"),
# library(fields)
# library(akima)

#system setting code here
DateTime &lt;- format(Sys.time(), "%Y-%m-%d %H-%M-%S")
path1 &lt;- "D:/06-Training_SelfPromotion/002_00-Data_Analysis_Presentation/06-Statistics_WuXiZhi/"
path2 &lt;- ""
WorkingDirectory &lt;- paste(path1, path2, sep = "")
setwd(WorkingDirectory)
#getwd()

#data input
rawData &lt;- read.csv(file.choose(WorkingDirectory), header = F, sep = ',')

#output jpeg file
#imageFileName &lt;- paste(TestPoints, " Points Test Map_", DateTime, ".jpeg",sep = "")
#print(imageFileName)
#outPutimageFilePath &lt;- paste(WorkingDirectory, imageFileName, sep = "")
#jpeg(outPutimageFilePath, width = 800, height = 800)

#process data code here

#eg. 1.1
((7324+388.674i)^2*(-245-0.245i)+(1785-43.4i)^3)/((236-0.0076i)^2*(43-299i)+(3.46-54i)*(344-23i)^5)
#[1] 1.417506e-05-2.531173e-05i

#eg. 1.2
#input matrix equation coefficient
x=scan()
528  446   -334 -143 571
-197 -304 414   -863 771
-502 722  302   789      -737
643  -995 -137  20 484

# form a matrix
x &lt;- matrix(x, 4, 5, b = T)
```

```
# solve the matrix equation
solve(x[, -5], x[,  5])
#[1]  1.4401909  0.2608429  1.2261890 -0.7258072

solve(x[, -5])
#           [,1]        [,2]       [,3]        [,4]
#[1,]  0.0017271256  0.0009252856 0.0012889296  0.0014267521
#[2,]  0.0009599372  0.0003027787 0.0005126416 -0.0002952618
#[3,]  0.0011047884  0.0020733899 0.0024269509  0.0016227977
#[4,] -0.0002024128 -0.0004819728 0.0006894502  0.0005568107

#eg. 1.3 solve root x
p(x) = a0+a1*x+a2*x^2+...+an*x^n
p(x)  = 2 + 3*x^2  -&gt; 7*x^3 + 0.8*x^4 -5*x^6 + 2.5*x^7
#coefficient (a0, a1, a2, ... an)
coefficient &lt;- c(2, 0, 3, -7, 0.8, 0, -5, 2.5)
polyroot(coefficient)
polyroot(c(2, 0, 3, -7, 0.8, 0, -5, 2.5))


#eg. 1.4 set.seed draw lots
#sampling: random take 10 from 100 people, not put back
set.seed(10)
sample(1:100, 10)

#andom take 10 from 100 people, put back, rep = T: replace set true
set.seed(0)
sample(1:100, 10, rep = T)



# http://cran.r-project.org/

#save save my current workspace in .RData and Save the Commands History
RDataFile &lt;- paste(WorkingDirectory, "R_in_Action.RData", sep = "")
save.image(RDataFile)


#save .RData
RDataFileName &lt;- paste(DateTime, ".RData", sep = "")
outPutRDataFilePath &lt;- paste(WorkingDirectory, RDataFileName, sep = "")
save.image(outPutRDataFilePath)
#unlink(outPutRDataFilePath), #Delete Files and Directories
print(paste(RDataFileName, "are successfully outputed at: ",WorkingDirectory))

#save .Rhistory
#loadhistory(file = ".Rhistory")
outPutRhistoryPath &lt;- paste(WorkingDirectory, ".Rhistory", sep = "")
savehistory(file = outPutRhistoryPath)
```

```
#images plot device off and close
dev.off()

#sessionInfo()
sessionInfo()
```

3、《统计学》的笔记-目录

统计学-基于R的应用 本书目录

统计学-基于R的应用 作者介绍
吴喜之，北京大学数学力学系学士，美国北卡罗来纳大学教堂山分校(UNC-Chapel Hill)统计系博士。中国人民大学统计学院教授，博士生导师。曾在美国加利福尼亚大学戴维斯分校(UC-Davis)、北卡罗来纳大学教堂山分校(UNC-Chapel Hill)、北卡罗来纳大学夏洛特分校(UNC-Charlotte)、加利福尼亚大学伯克利分校(UC-Berkeley)、南开大学、中国人民大学、北京大学、中山大学、四川大学等十余所著名学府执教。

《统计学》

版权说明