Statistical Methods for Data Science (25/7/2017)

Students cannot use teaching materials, smartphones or computers. They can only use a scientific calculator. Duration of the written exam is 2h.

Exercise 1 (6 points). Let X and Y be two independent $Ber(\frac{1}{2})$ random variables. Define the random variables U and V by

U=X+Y V=|X-Y|

a) Determine the joint and marginal distributions of U and V.
b) Find out whether U and V are dependent or independent
c) Determine the covariance Cov(U,V) and the correlation coefficient Cor(U,V)

Exercise 2 (6 points). Suppose that x_1, x_2, …, x_n is a dataset, which is a realization of a random sample from a Rayleigh distribution, which is a continuous distribution with probability density function:

$$f_\theta = \frac{x}{\theta^2} e^{-\frac{x^2}{\theta^2}} \quad \text{for } x \geq 0.$$

In this case what is the maximum likelihood estimate of θ?

Exercise 3 (6 points). One is given a number t, which is the realization of a random variable T with an N(μ,1) distribution. To test H_0:= $\mu = 0$ against H_1: $\mu \neq 0$ one uses T as the test statistic. One decides to reject H_0 in favor of H_1 if $|t| \leq 2$.

a) Compute the probability of committing a type I error.
b) Compute the probability of committing a type II error if the true value of μ is 1.

Use the Table B.1 (attached) from the textbook for the tail probability of normal random variables.

Exercise 4 (6 points). Consider a data frame of students:

```
d <- data.frame(id, enrolledYear, cfu0, cfu1, cfu2, cfu3, cfu4, cfu5, cfu6, cfu7, cfu8, cfu9)
```

where `id` is the student ID (matricola), `enrolledYear` is the year of enrollment (eg., 2014), `cfu0` is the number of credits given in the year of enrollment (eg., 30 in 2014), `cfu1` is the number of credits given in the year of enrollment+1 (eg., 24 in 2015), etc. Write an R function that transforms `d` into a data frame of the form:

```
d1 <- data.frame(id, year, cfu, inc)
```

where a row contains `id` the student ID (matricola), a `year` (eg., 2015), `cfu` is the number of credits (if non-zero) given in that year (eg., 24), and `inc` is the difference of credits wrt the previous year (eg., -6).

Exercise 5 (6 points). Code in R bootstrapped confidence intervals for the test of mean.

Statistical Methods for Data Science - Solutions - (25/7/2017)

Exercise 1. (a) see solution of Ex. 9.6 (a) at page 437 of [B1].
(b) Dependent, because for e.g.,

$$P(U=0,V=0)=\frac{1}{4}\neq P(U=0)P(V=0)=\frac{1}{4}\frac{1}{2}=\frac{1}{8}$$

(c) We have

$$E[U]=0\frac{1}{4}+1\frac{1}{2}+2\frac{1}{4}=1$$

$$E[V]=0\frac{1}{2}+1\frac{1}{2}=\frac{1}{2}$$

$$E[UV]=0P(U=0\vee V=0)+1P(U=1,V=1)+2P(U=1,V=2)=\frac{1}{2}.$$

Therefore

$$Cov(U,V)=E[UV]-E[U]E[V]=\frac{1}{2}-\frac{1}{2}=0$$

and then $Cor(U,V)=0$ as well.

Exercise 2. The likelihood is

$$L=\prod_{i=1}^{n}\frac{x_i}{\theta^2}e^{-\frac{x_i^2}{\theta^2}}$$

and then the log-likelihood is

$$logL=\sum_{i=1}^{n}(logx_i-2log\theta-\frac{1}{2\theta^2}x_i^2)$$

The (log-)likelihood has maximum when the derivative (w.r.t. θ) is zero, i.e., when

$$\frac{\partial logL}{\partial\theta}=-\frac{2n}{\theta}+\frac{1}{\theta^3}\sum_{i=1}^{n}x_i^2=0$$

which occurs for

$$\theta=\sqrt{\frac{1}{2n}\sum_{i=1}^{n}x_i^2}$$

Exercise 3

(a) A type I error is done when H0 is true but it is rejected. Hence it is the probability $P_{H0}(|t|>2) = 2\ P_{H0}(t>2) = 2\ 0.0228 = 0.0456$\$, where \$$P_{H0}(t>2) = 1 - \mathbf{\Phi}(2) = 0.0228$\$ accordingly to the distribution of N(0, 1).
(b) A type II error is done if the true value of μ is 1 but H0 is not rejected. Thus it is $P_{H1}(|t|\leq 2) = P_{H1}(|z+1|\leq 2) = P_{H1}(-3\leq z\leq 1)$ where Z = T-1 ~ N(0, 1). Therefore, $P_{H1}(-3\leq z\leq 1) = P_{H1}(z\geq 3) + (1 - P_{H1}(z\geq 1)) = 0.0013 + (1-0.1587) = 0.84$.

Exercise 4

```
transform = function(d) {
  # calculate increments
  incs = d[,3:12]
  incs = incs - cbind(0, incs[,-10])
  names(incs) = c("inc0", "inc1", "inc2", "inc3", "inc4", "inc5", "inc6", "inc7", "inc8", "inc9")
  d0 = cbind(d, incs)
  # init result
  d1 <- data.frame()
  for(i in 0:9) {
    # select a cfu and inc
    tmp= d0[c("id", "enrolledYear", paste("cfu", i, sep=""), paste("inc", i, sep=""))]
    # rename cols
    names(tmp) = c("id", "year", "cfu", "inc")
    # calculate time
    tmp$year = tmp$year+i
    # append
    d1= rbind(d1, tmp)
  }
  return(d1)
}
```

Exercise 5

```
data=rnorm(20) # example input
# actual answer
library(boot)
b <- boot(data, function(x,d) mean(x[d]), R = 1000)
quantile(b$t, c(0.025, 0.975))
# or
boot.ci(b, type = "norm")
```

References

[B1] F.M. Dekking C. Kraaikamp, H.P. Lopuha, L.E. Meester. A Modern Introduction to Probability and Statistics. Springer, 2005.

[B2] P. Dalgaard. Introductory Statistics with R. 2nd edition, Springer, 2008.