Exam Statistics and Probability

October 27, 2023

Download this Markdown (Rmd) file, and add your R-code to answer the questions. When you are finished, render the document either to pdf or html and upload to Brightspace. If the rendering fails, upload the Markdown (Rmd) file with your R code, or upload your R code in any format you want. If the uploading fails, email your work to E.W.van_Zwet@lumc.nl.

Some familiar R commands and formulas are available at the end of this exam.

Problem 1.

Use the following `R` code to generate a small dataset.

```
set.seed(123)
x=rnorm(15,2,3)
```

Suppose we observe these 15 numbers. We will assume that they are a sample from a normal distribution, but we do not know μ and σ. We want to estimate the probability that a sample from this distribution is larger than 1. We can use `1 - pnorm(1,m,s)` as an estimator where `m` is the sample average $\frac{1}{n}\sum_{i=1}^n X_i$ and `s` is sample standard deviation $\sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2}$.

(a) Use the parametric bootstrap to compute the mean squared error of this estimator.

An alternative estimator is the proportion of the 15 observations that is larger than 1.

(b) Use the parametric bootstrap to compute the mean squared error of this alternative estimator.

```
# add your R code here
set.seed(123)
x=rnorm(15,2,3)
m = mean(x)
s = sd(x)
```

(a)

```
k = 1e5
B = numeric(k)
for (i in 1:k){
  x_boot = rnorm(15,m,s)
```

```
  m_boot = mean(x_boot)
  s_boot = sd(x_boot)
  B[i] = (1-pnorm(1, m_boot, s_boot)) - (1 - pnorm(1,m,s)) # B = T-theta
}
mse = mean(B**2)
mse # This is the mse.
```

```
## [1] 0.009015988
```

(b)

```
B1 = numeric(k)
for (i in 1:k){
  x_boot = rnorm(15,m,s)
  B1[i] = mean(x_boot > 1) - (1 - pnorm(1,m,s))  # B = T-theta
}
mse1 = mean(B1**2)
mse1 # This is the mse.
```

```
## [1] 0.01336893
```

Problem 2.

Suppose we have an i.i.d. sample $X_1, X_2, \ldots, X_{10}$ from a distribution with density

$$f_\theta(x) = \frac{2\theta}{\pi} \exp\left(-\frac{x^2\theta^2}{\pi}\right), \qquad x \geq 0.$$

where θ is an unknown, non-negative parameter. The sample is

```
x=c(1.20, 1.12, 1.60, 0.48, 0.04, 0.35, 0.67, 0.51, 4.28, 0.64)
```

(a) Plot the loglikelihood function for values of θ between 0 and 3.
(b) Compute the maximum likelihood estimator of θ numerically.
(c) Use pen and paper to show that the maximum likelihood estimator of θ is given by

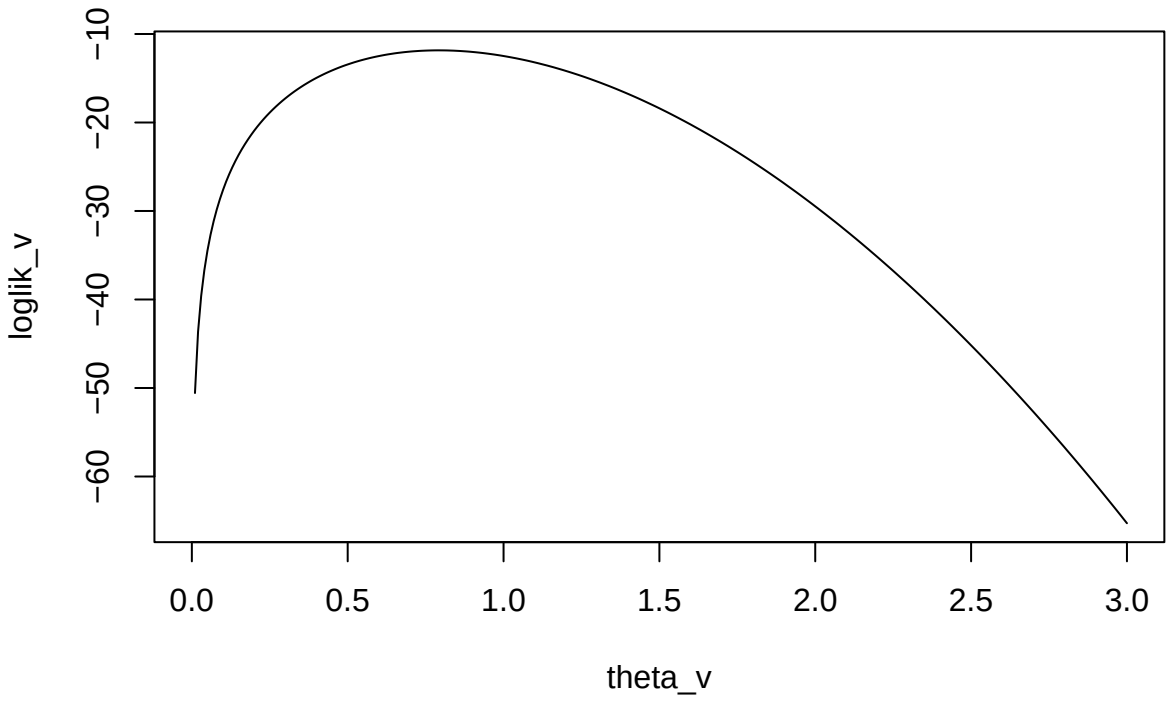
$$\hat{\theta} = \sqrt{\frac{n\pi}{2\sum_{i=1}^n x_i^2}}.$$

(d) Compute the observed Fisher Information at the maximum likelihood estimator. You may either use pen and paper or numerical methods.

```
# add your R code here.
x=c(1.20, 1.12, 1.60, 0.48, 0.04, 0.35, 0.67, 0.51, 4.28, 0.64)
```

(a)

```
loglik = function(theta){
  n = length(x)
  result = n*(log(2*theta/pi))-(theta**2)/pi*sum(x**2)
  return(result)
}
theta_v = seq(0, 3, 0.01)
loglik_v = sapply(theta_v, loglik)
plot(theta_v, loglik_v, type = 'l')
```



(b)

```
opt = optimize(loglik, c(0, 3), maximum = TRUE)
theta_mle = opt$maximum
theta_mle # mle
```

```
## [1] 0.7919372
```

(d)

```
library(numDeriv)
fisher_obs = -hessian(loglik, theta_mle)
fisher_obs # observed Fisher Information
```

```
##          [,1]
## [1,] 31.88949
```

Problem 3.

Fisher's exact test is designed to assess whether two binary random variables are independent of each other. Ronald Fisher invented the test for the following experiment, called "The Lady Tasting Tea". A lady named Muriel Bristol claimed that she could detect, in a cup of tea with milk, whether the milk was poured first or whether the tea was poured first.

Fisher prepared 8 cups, 4 of each kind, and asked Muriel to indicate which 4 cups were of the tea-first type. The null hypothesis is that her choices are independent of the true state, i.e., that her guesses were random. He used the test statistic X, defined as the number of correct guesses (between 0 and 4). Under the null hypothesis, X has the hypergeometric distribution with parameters $m = 4$, $n = 4$ and $k = 4$. In `R`, the hypergeometric distribution is implemented as `p/r/q/dhyper`.

(a) Find the critical value of this test for $\alpha = 0.05$.

(b) What is the Type I error of the test that rejects when $X \geq 3$?

(c) Muriel Bristol guessed all four cups correctly. Calculate the p-value that corresponds to this result.

(a)

Based on the null hypothesis and test statistics, I use two-side test.

```
alpha = 0.05
X_cv1 = qhyper(alpha/2, m = 4, n = 4, k = 4)
X_cv2 = qhyper(1-alpha/2, m = 4, n = 4, k = 4)
c(X_cv1, X_cv2)
```

```
## [1] 1 3
```

In conclusion, when $X > 3$ or $X < 1$, we reject null hypothesis for $\alpha = 0.05$.

(b)

```
type1 = 1 - phyper(3-1, m = 4, n = 4, k = 4)
type1 # type I error
```

```
## [1] 0.2428571
```

(c)

```
X_obs = 4
p_value = 2*(1 - phyper(X_obs, m = 4, n = 4, k = 4)) # two-side test
p_value # p-value
```

```
## [1] 0
```

Problem 4.

The gamma-distribution is given in `R` by `p/r/q/dgamma`. It has a shape parameter k and a rate parameter θ. For $k = 4$, the distribution function is given by

$$f(x) = \frac{1}{6}\theta^4 x^3 \exp(-x\theta).$$

Suppose we have i.i.d. data $(X_1, \ldots, X_n) = (21.3, 4.1, 18.5, 15.2, 9.3)$ from a gamma distribution with known $k = 4$ and unknown θ.

(a) Show by pen-and-paper that the maximum likelihood estimator for θ is $4/\bar{X}$.

(b) Find numerically the (non-asymptotic) critical value for $\alpha = 0.05$ of the likelihood ratio test for $H_0 : \theta = 1$.

(c) Is $\theta = 1$ contained in the asymptotic 95%-likelihood confidence interval for θ?

(b)

```
X = c(21.3, 4.1, 18.5, 15.2, 9.3)
n = length(X)
loglik = function(theta){
  result = n*(-1*log(6)+4*log(theta))+3*sum(log(X))-theta*sum(X)
  return(result)
}
k = 1e5
lrt = numeric(k)
theta0 = 1
for (i in 1:k){
  X1 = rgamma(n, shape = 4, rate = theta0)
  theta_hat = 4/mean(X1)
  lrt[i] = sum(dgamma(X1, shape = 4, rate = theta_hat, log = TRUE)) - sum(dgamma(X1, shape = 4, rate = 
}
lrt_cv = quantile(lrt, 0.95)
lrt_cv # numerical critical value
```

```
##      95%
## 1.939771
```

(c)

```
lrt_cv_asy = qchisq(0.95, df = 1)/2 # asymptotic critical value
theta_hat = 4/mean(X)
theta0 = 1
lrt = loglik(theta_hat) - loglik(theta0) # observed likelihood ratio test
lrt <= lrt_cv_asy
```

```
## [1] FALSE
```

In conclusion, $\theta = 1$ is not contained in the asymptotic 95%-likelihood confidence interval.

some R functions

help: help

calculator: + - * / abs x^2 sqrt log exp

vectors: c seq rep 1:10

operations on vectors: sum prod length max which.max

plot: plot points lines hist boxplot

boolean variables: which & |

sub-setting with square brackets: x=c(5,2,6,1,2); x[3]

boolean variables and sub-setting: sum(x<4); x[x<4]

input/output: cat load

control structures: for loop and if statement

probability distributions: d/p/q/rnorm *unif *binom *exp *geom *pois

descriptive statistics: summary mean median var sd

hypothesis testing: t.test, prop.test, chisq.test

some formulas

Suppose X and Y are random variables and a and b are scalars (constants, numbers).

$$\begin{aligned} E(aX+b) &= aE(X)+b \\ E(X+Y) &= E(X)+E(Y) \\ \mathrm{Var}(X) &= E(X^2)-E(X)^2 \\ \mathrm{Var}(aX+b) &= a^2\mathrm{Var}(X) \\ \mathrm{Var}(X+Y) &= \mathrm{Var}(X)+\mathrm{Var}(Y)+2\mathrm{Cov}(X,Y) \end{aligned}$$

Moreover, the mean squared error of an estimator $\hat{\theta}$ of a parameter θ is

$$\mathrm{MSE}(\hat{\theta}) = E[(\hat{\theta}-\theta)^2] = \mathrm{Var}(\hat{\theta}) + E(\hat{\theta}-\theta)^2$$

In other words, the MSE is the variance plus the square of the bias.

An overview of the Wald, score and likelihood ratio tests and their asymptotic distributions:

Statistic	Definition	Distribution
Wald	$\hat{\theta}-\theta_0$	$\mathcal{N}\{0, 1/\mathcal{I}(\hat{\theta})\}$
Score	$\mathrm{loglikelihood}'(\theta_0)$	$\mathcal{N}\{0, \mathcal{I}(\theta_0)\}$
Likelihood ratio	$\mathrm{loglikelihood}(\hat{\theta}) - \mathrm{loglikelihood}(\theta_0)$	$\frac{1}{2}\chi^2_{\mathrm{df}=1}$