Practice Exam Statistics and Probability

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. If the uploading fails, email to E.W.van_Zwet@lumc.nl.

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

Problem 1.

Suppose we have a sample of size 10 from the exponential distribution with unkown rate λ. Generate a small dataset by running the following `R` code:

```
set.seed(123)
x=rexp(10,rate=2)
```

As you know, the mean of the exponential distribution is $1/\lambda$ and the variance is $1/\lambda^2$. We decide to estimate the variance with the square of the sample average. Use the bootstrap to answer the following questions.

(a.) What is the bias of the estimator?

(b.) What is the variance of the estimator?

(c.) What is the mean squared error of the estimator?

```
set.seed(123)
x=rexp(10,rate=2)

var.hat=mean(x)^2           # estimate of the variance
rate.hat=sqrt(1/var.hat)    # same as rate.hat=1/mean(x)

k=10^5
B=numeric(k)
for (i in 1:k){
  x.boot=rexp(n=10,rate=rate.hat) # simulate from the estimated distribution
  var.boot=mean(x.boot)^2
  B[i]=var.boot-var.hat
}
# a.
mean(B)    # bias
```

```
## [1] 0.01010321
```

```
# b.
var(B)     # variance
```

```
## [1] 0.005192095
```

```r
# c.
mean(B^2)  # MSE
```

```
## [1] 0.005294118
```

Problem 2

We assume that $X = (1.02, 3.35, 6.83, 4.60, 1.80)$ are distributed according to a Rayleigh distribution with parameter θ^2. The Rayleigh distribution only takes positive values and its density for $x > 0$ is

$$f(x) = \frac{x}{\theta^2} e^{-x^2/(2\theta^2)}.$$

The expectation of the Rayleigh distribution is $\theta\sqrt{\pi/2}$.

(a.) Calculate a method-of-moments estimate of θ^2.

Answer: We have $\bar{X} \approx \mathrm{E}(X) = \theta\sqrt{\pi/2}$, so $\hat{\theta} = \bar{X}/\sqrt{\pi/2}$, so $\hat{\theta}^2 = 2\bar{X}^2/\pi$

```
X = c(1.02, 3.35, 6.83, 4.60, 1.80)
hat_theta2 = 2*mean(X)^2/pi
hat_theta2
```

```
## [1] 7.887974
```

(b.) Find the Maximum likelihood estimator of θ^2 analytically (by pen and paper).

Answer:

$$\log f(x) = \log x - \log \theta^2 - x^2/(2\theta^2)$$

so the likelihood is

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

so the score is

$$\mathrm{loglik}'(\theta^2) = -\frac{n}{\theta^2} + \frac{1}{2\theta^4} \sum_{i=1}^{n} x_i^2.$$

We solve

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

so

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

(c.) Find the maximum likelihood estimator of θ^2 numerically using optimize.

Answer:

```
X = c(1.02, 3.35, 6.83, 4.60, 1.80)
loglik_fun = function(theta2) sum(log(X/theta2 * exp(-X^2/(2*theta2))))
optimize(loglik_fun, c(1,20), maximum = TRUE)$maximum
```

```
## [1] 8.331182
```

Problem 3

Suppose we have a *single* observation X from the Poisson distribution with unknown λ. We want to perform a one-sided test

$$H_0 : \lambda = 10 \quad \text{versus} \quad A : \lambda > 10$$

We reject the null hypothesis if $X > 13$. Use the `R` function `ppois` to answer the following questions.

(a) Determine the probability of a type I error.

(b) Determine the probability of a type II error when $\lambda = 12$.

(c) Draw the power curve for λ between 1 and 20.

(d) What is the smallest critical value such that the test has level of significance $\alpha \leq 0.05$.

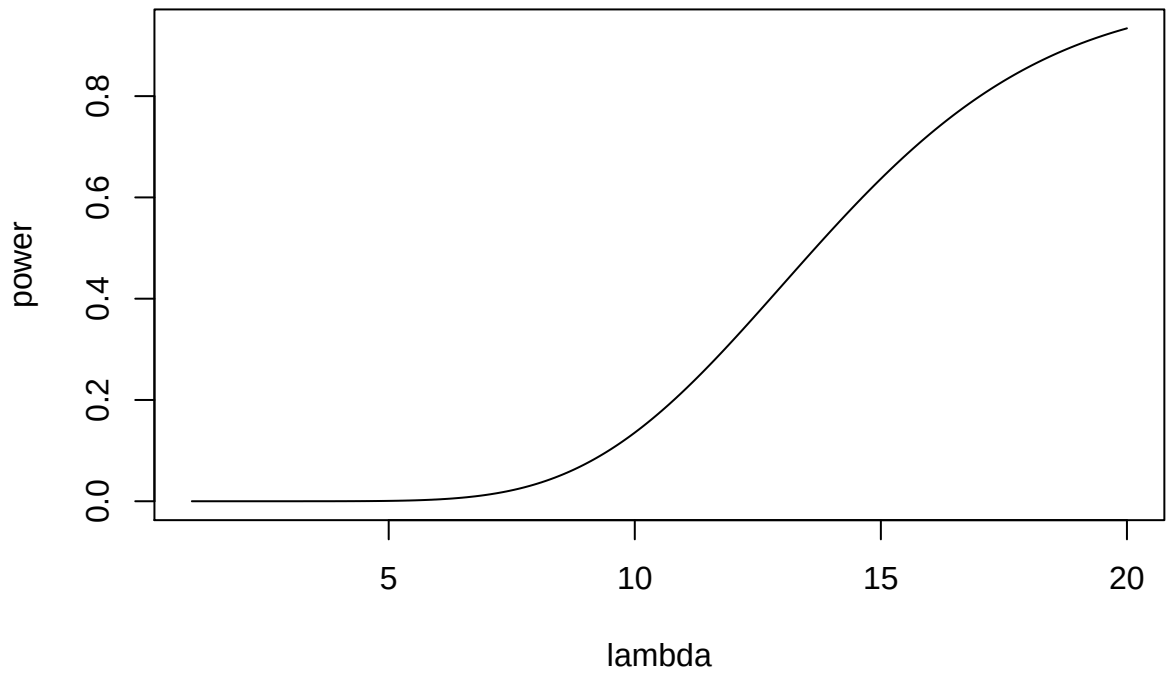
```
# a
1-ppois(13,lambda=10)    # P(X > 13) = 0.14 when lambda=10
```

```
## [1] 0.1355356
```

```
# b
ppois(13,lambda=12)      # P(X <= 13) = 0.68 when lambda=12
```

```
## [1] 0.6815356
```

```
# c
lambda=seq(1,20,0.01)
plot(lambda,1-ppois(13,lambda),type='l',xlab='lambda',ylab='power')
```




```
# d. Reject if X > 15 is a level 5% test. There are few different
# ways to find the critical value:
# method 1 Just try out for different critical values. You find that
#
#P(X>15)=1-ppois(15,lambda=10) is just below 5%, while
#
#P(X>14)=1-ppois(14,lambda=10) is just above 5%.
#
# method 2 Use simulation. Sample the test statsitic under the
# null hypothesis, and compute the 95% quantile:
x=rpois(10^6,lambda=10)
quantile(x,0.95)
```

```
## 95%
##  15
```

```
# method 3 The distribution of the test statistic under the null
# hypothesis is very easy. The quantile function is available
# directly in R:
qpois(0.95,lambda=10)
```

```
## [1] 15
```

Problem 4.

The hypergeometric distribution is used for random draws *without replacement*. It is implemented in R in the d/r/p/qhyper function. Have a look at the help file of this function in R. Suppose we know that an urn contains $m + n = 100$ balls, but we do not know how many white balls m or black balls n it contains. We formulate the null hypothesis $H_0 : m = n = 50$. Drawing $k = 10$ balls without replacement, we observe $x = 2$ white balls.

(a.) Show that the likelihood estimate for m is $\hat{m} = 20$ numerically. Hint: use which.max instead of optimize.

Answer:

```
loglik_fun = function(m) dhyper(x=2, m=m, n=100-m, k=10, log=TRUE)
m_values = 0:100
loglik = sapply(m_values, loglik_fun)
m_values[which.max(loglik)]
```

```
## [1] 20
```

(b.) Calculate the test statistic and the critical value of the likelihood ratio test statistic for H_0 at $\alpha = 0.05$ numerically. Use the asymptotic distribution of the likelihood ratio test. What is your conclusion about H_0?

Answer:

```
loglik_fun = function(m) dhyper(x=2, m=m, n=100-m, k=10, log=TRUE)
hat_m = 20
T_LRT = loglik_fun(20) - loglik_fun(50)
T_LRT
```

```
## [1] 2.125177
```

```
critical_value = qchisq(0.95, df=1)/2
critical_value
```

```
## [1] 1.920729
```

Since the test statistic is larger than the critical value, we reject H_0.

(c.) Calculate the p-value corresponding to the test of exercise b.

Answer:

```
p_value = pchisq(2*T_LRT, df=1, lower.tail = FALSE)
p_value
```

```
## [1] 0.03924214
```

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.