W1-2-60-1-6

JOMO KENYATTA UNIVERSITY

OF

AGRICULTURE AND TECHNOLOGY

UNIVERSITY EXAMINATIONS 2017/2018

SECOND YEAR FIRST SEMESTER EXAMINATIONS FOR THE DEGREES OF

BACHELOR OF SCIENCE IN FINANCIAL ENGINEERING

BACHELOR OF SCIENCE IN ACTUARIAL SCIENCE

BACHELOR OF SCIENCE IN BIOSTATISTICS

BACHELOR OF SCIENCE IN STATISTICS

&

BACHELOR OF SCIENCE IN OPERATIONS RESEARCH

STA 2202: COMPUTER INTERACTIVE STATISTICS

DATE: JANUARY 2018 TIME: 2 HOURS

INSTRUCTIONS TO CANDIDATES:

1. *Answer question ONE (section A) and any other two questions in section B.*
2. *Be neat and show all your workings*
3. *All questions except question one carry equal marks*

This paper consists of 6 printed pages.

SECTION A (30 MARKS)

1. (a) State the difference between a statistic and a parameter? [2 mark]

(b) Jane is interested in the Nairobi county gubernatorial seat. She wants to find out if she has any chances of winning the TEA party primaries so as to be selected as the party candidate for the position. In order to examine the voters' opinions, she hires the services of SYNOVATE, a polling agency. Polling is conducted among 500 registered voters from the TEA party. One of the questions asked by the pollster is on to the voters' willingness to vote for a female candidate. 42% of the respondents say they prefer to have a women running for the job. 38% think the candidate's gender is irrelevant. The rest prefer a male candidate. From this literature, identify the following:

42
38
20

20%

(i) The population [1 mark]

(ii) The sample. [1 mark]

(iii) A parameter. [1 mark]

(iv) A statistic. [1 mark]

42/100 × 500

4.2 × 5 = 210

38 × 5 = 190

(c) If you executed the following commands in R, what would the output be?

(i)
```
> (5 > 7) & (6*7 == 42)
```
[1 mark]

(ii)
```
> (5 > 7) | (6*7 == 42)
```
[1 mark]

(iii)
```
>  round(7) == 7
```
[1 mark]

(iv)
```
> weight <- c(60,72,57,90,95,72)
> height <- c(1.75,1.80,1.65,1.90,1.74,1.91)
> sex <- rep(c("female", "male"), each = 3)
> smoking <- rep(c("TRUE","FALSE"), 3)
> data <- data.frame(weight, height, sex, smoking)
>data
```
[3 marks]

(d) Consider the following data with mean =155.00 and standard deviation = 24.68.

```
> distance <- c(148,182,173,166,109,141,166)
```

Write a function in R called mean.and.sd that will give the output. [3 marks]

```
> mean.and.sd(distance)
    mean        SD
155.00000  24.68468
```

(e) The Fibonacci Sequence is the series of numbers: 0, 1, 1, 2, 3, 5, 8, 13, 21,... where, the next number is found by adding up the two numbers before it. Write a function in

R that returns the first n Fibonacci numbers. The function should take n as an input argument. [4 marks]

(f) Write a code in R will plot the normal pdf $p(t) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(t-\mu)^2}{2\sigma 2}}$ over the interval $-3 \leq t \leq 3$ for $\mu = 0$ and for the values $\sigma = 1.5$, $\sigma = 1$ and $\sigma = 0.5$, on a single graph and colours the graphs black, red and blue respectively. [4 marks]

(g) The following are 14 samples from a normal population with unknown mean and unknown standard deviation:

`2.59 2.67 2.16 1.95 2.61 1.11 2.62 2.06 2.06 1.66 2.16 3.35 2.46 2.55`

(i) Write a code in R that will estimate the mean μ, the standard deviation σ, and the variance σ^2 from this sample. [3 marks]-

(ii) If the estimation in (i) above yields $\mu \approx 2.286429$, $\sigma \approx 0.5342228$ and $\sigma^2 \approx 0.2853940$, write a code in R that will test the hypothesis $H_0 : \mu = 2$, using the significance level $\alpha = 0.05$. If this test yields a `p-value=0.06611`, what decision do you make? [2 marks]

(iii) Write a code in R that will test the hypothesis $H_0 : \mu \leq 2$, using the significance level $\alpha = 0.05$. If this test yields a `p-value=0.03306`, what decision do you make? [2 marks]

SECTION B (20 MARKS EACH)

2. The height and sharpness of the peak relative to the rest of the data are measured by a number called kurtosis. If you have data for only a sample, you have to compute the sample excess kurtosis using the following formula:

$$G_2 = \frac{n-1}{(n-2)(n-3)}[(n+1)g_2 + 6]$$

where

$$g_2 = a_4 - 3$$

$$a_4 = \frac{M_4}{M_2^2}$$

$$M_4 = \sum \frac{(X - \bar{X})^4}{n}$$

$$M_2 = \sum \frac{(X - \bar{X})^2}{n}$$

A statistical consultant wants to compute the confidence interval for the population variance of credit card balances of Kenyan consumers. The $100(1-\alpha)\%$ confidence interval for the population variance is defined as:

$$S^2 * exp\left(-Z_{(1-\alpha/2)} * \sqrt{\frac{\hat{\gamma}-1}{n}}\right) \leq \sigma^2 \leq S^2 * exp\left(Z_{(1-\alpha/2)} * \sqrt{\frac{\hat{\gamma}-1}{n}}\right)$$

where $\hat{\gamma}$ is the estimator of kurtosis defined as:

$$\hat{\gamma} = G_2 + 3$$

The balances (in Euros) in table (1) are at his disposal. Required: Write a well commented program in R that does the following:

295	3147	283	569	1141	788	1255	2038	978	548
1133	1641	959	816	955	1473	702	459	1844	

Table 1: Credit Card Balances

Reads in the data in table (1). [3 marks]

(ii) Computes the sample variance, S^2, for the data. [3 marks]

(iii) Computes the sample excess kurtosis, G_2, of the data. [6 marks]

(iv) Computes the 90% confidence interval for the population variance, σ^2, of credit card balances of Kenyan consumers. [8 marks].

Hint: The R function exp(X) computes the exponent of X. e.g. exp(10) returns 22026.47.

3. In a quality control operation, we examine the paint of n brand-new cars taken at random among those produced by a certain company. Let $X_k = 1$ if the paint of the k^{th} car examined has at least one flaw, and $X_k = 0$ otherwise, for $k = 1, 2, ..., n$. We assume that the random variables X_k are independent and that the probability that the paint of a brand-new car is flawless is equal to 0.75. Write an R program that does the following:

   (a) Generates the random variables $X_k, k = 1, 2, ..., 40$. [5 marks]

   (b) Computes the following sums from *(a)* above:

   (i) $S_1 = |\sum_{k=1}^{40} X_k - 10|$. [5 marks]

   (ii) $S_2 = \sum_{k=1}^{40} X_k$. [4 marks]

   (c) Summarizes the results from *(i)* above using an appropriate and properly labelled graph. [6 marks]

4. (a) A researcher carefully computes the correlation coefficient between two variables and gets $r = 1.23$. What does this value mean? [1 mark]

   (b) Write a command in R that will print the following output. [2 marks]

```
[1] "square of 1 = 1"
[1] "square of 2 = 4"
[1] "square of 3 = 9"
[1] "square of 4 = 16"
[1] "square of 5 = 25"
```

(c) Consider the function

$$f(x) = 2x^4 - 0.9x - 1.$$

Write a function in R that will compute the values of $f(0)$ and $f(1.5)$ [2 marks]

(d) Define hypothesis testing and state the four steps of hypothesis testing. [5 marks]

(e) A marine biologist measured the length (in millimetres) and the weight (in grams) of 10 fish that where collected in one of her expeditions. The results are summarized in the table below.

Length (x)	4.5	3.7	1.8	1.3	3.2	3.8	2.5	4.5	4.1	1.1
Weight (y)	9.5	8.2	4.9	6.7	12.9	14.1	5.6	8.0	12.6	7.2

(i) Write a code in R that will capture this data and display it in a scatter plot and add the line of best fit. Consider y as the response variable. [4 marks]

(ii) After keying in this data and regressing y on x, the following output is obtained.

```
> fit <- lm(y~x)
> fit

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)              x
4.616          1.427

> summary(fit)

Call:
lm(formula = y ~ x)

Residuals:
Min        1Q  Median      3Q      Max
-3.0397 -2.1388 -0.6559  1.8518  4.0595

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.6165     2.3653   1.952   0.0868 .
x             1.4274     0.7195   1.984   0.0826 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.791 on 8 degrees of freedom
```

```
Multiple R-squared:  0.3297,     Adjusted R-squared:  0.246
F-statistic: 3.936 on 1 and 8 DF,  p-value: 0.08255
```

(A) What is the aim of fitting the regression line to data? [1 mar

(B) What is the null hypothesis for each of the coefficients, and bases on the compute $p-$ values of the coefficients, what is our decision with respect to these hypotheses

[2 mark

(C) Give the command that will give the 95% confidence interval for the parameters of the linear regression model. [1 mark]

(D) Comment on the goodness of fit of the regression model on the data. [2 marks]