Statistical Methods in R: End-Of-Class Test

TIME ALLOWED: 3 Hours, 20 Minutes

INSTRUCTIONS TO CANDIDATES

Answer any three of the following problems. If more than three problems are attempted, only the first three will be marked. The question you choose not to answer should be left completely blank.

The test is fully open-book, open-notes, and open-web. However, you may not consult or communicate with anyone else at any point, with the sole exception that you may seek clarification from the invigilator. You may not use any chat or messaging software during the test.

Solutions should be prepared only in R / Rstudio without the use of any other software. An RMarkdown template is available and use of this to complete the questions is required. At the end of the test, you should submit your code and a compiled PDF with your responses. The content of the code and the PDF should be consistent. Be sure to give yourself plenty of time at the end to make sure that your document knits correctly.

If the question requests a specific number or group of numbers, make sure your answer is unambiguous: either by making it the sole output of a block of R code, or (if the answer appears in a large summary-type output) by explicitly adding a one-sentence written remark or formatted data table below the code chunk identifying the answer or answers. Marks will not be awarded if multiple potential "answers" could be inferred from your output and the correct one is not clearly distinguished by a remark.

A blank R code chunk is provided for every question; however, some questions may not require any R code. Verbal responses (in place of or in addition to a code chunk) should be written as text outside the code chunk, not as comments within the code chunk. A blank "setup" chunk is provided at the start of each problem for initial operations (e.g. loading the data file).

Comment out any extraneous outputs (e.g. exploratory plots) before compiling and submitting, and please keep additional explanations to a minimum. Do *not* comment out or hide code or calculations used to answer a question. If a question is answered incorrectly, partial credit will be awarded based only on what is performed in the code. Some questions are not eligible for partial credit.

The results of hypothesis tests should specify p and/or α, regardless of whether the question specifically asks for this number. Model parameter estimates should include confidence intervals (with the associated confidence level), unless otherwise stated. Calculations of descriptive statistics do *not* require standard errors, unless otherwise stated.

Mark values for each question are indicated in brackets, e.g. [**2**]. Each complete problem is worth a maximum of 20 marks. The maximum mark for the test is 60.

The page limit for this test is **16 pages** including this cover page. Please ensure that your output file is within the page limit before submitting and remove unnecessary outputs as needed. Do *not* delete any of the instructions or questions. Marks will be deducted for submissions in excess of the page limit.

The time limit for the test is 3 hours and 20 minutes. A late penalty of 2 marks will be applied to submissions up to 12 minutes late. Submissions more than 12 minutes late will be penalized at 2 additional marks per additional minute late. Submissions more than 30 minutes late will not be accepted.

P1: 17 / 20

Problem 1. Practical Probabilities

The lengths of members of a prized fish species are thought to be normally distributed. The distribution of lengths has a mean of 20cm with a standard deviation of 3cm.

```
mu = 20
deviation = 3
```

1. What is the probability that a randomly-chosen fish of this species has a length between 20 cm and 27 cm? [**2**]

2 / 2

```
#cdf(27) - cdf(20)
ans = pnorm(27,mean=20,sd=3) - pnorm(20,mean=20,sd=3)
ans
```

```
## [1] 0.4901847
```

2. Two fish of this species are caught independently. What is the probability that both of them are shorter than 16cm? [**2**]

2 / 2

```
#cdf(16) * cdf(16)
ans = pnorm(16,mean=20,sd=3)*pnorm(16,mean=20,sd=3)
ans
```

```
## [1] 0.008319487
```

3. Forty percent of all fish of this species have lengths greater than 20cm but less than X cm. What is X? [**2**]

2 / 2

```
#cdf(X) - cdf(20) = 0.4
#cdf(X) = cdf(20) + 0.4 = 0.9
ans = qnorm(0.9, mean=20, sd=3)
ans
```

```
## [1] 23.84465
```

4. In a commercial catch of 10000 fish, how many (on average) do you expect to have a length greater than 30 cm? (You do not need to supply a confidence interval.) [**2**]

2 / 2

```
#p = 1-cdf(30)
#expected value = 10000*p
prob = 1 - pnorm(30, mean=20,sd=3)
expected_val = 10000*prob
expected_val
```

```
## [1] 4.290603
```

From historical records stretching back 1000 years, a volcano is known to have erupted on sixteen occasions. The eruption times are thought to be completely random and unpredictable, and the average frequency of eruptions is not changing with time.

5. What is the long-term average rate of eruptions, in units of eruptions per century (per 100 years)? Provide an estimated value and an “exact” confidence interval. [**4**]

```
#16 eruptions in 1000 years = 1.6 eruptions in 100 years
average_rate_of_eruptions_per_year = 16/1000
average_rate_of_eruptions_per_century = average_rate_of_eruptions_per_year*100
average_rate_of_eruptions_per_century # this is the estimated value
```

```
## [1] 1.6
```

3 / 4

```
confintlower = qchisq(0.025,2*average_rate_of_eruptions_per_century)/2
confintupper = qchisq(0.975,2*(average_rate_of_eruptions_per_century+1))/2
CI = c(confintlower,confintupper)
CI
```

The value in the CI calculation should always be the actual count

```
## [1] 0.1310581 6.5804601
```

6. Assume the estimated average rate you calculated in part 5 is correct. What is the probability that the next 100 years will pass without an eruption from this volcano? [**2**]

2 / 2

```
#the following problem can be modeled as a poisson distribution
#for a century, average eruptions = 1.6 (part 5)
#probability of no eruption in 100 years = probability of event occurring 0 times, with an average 1.6
ans = dpois(0,1.6)
ans
```

```
## [1] 0.2018965
```

7. Again assuming the average rate from part 4, what is the probability that the volcano will erupt *two or more* times in the next *200* years? [**2**]

2 / 2

```
#average eruptions per century = 1.6 (part 5)
#average eruptions per 200 years = 2*average eruptions per century = 3.2
# prob = 1 - prob of 0 eruptions in 200 years - prob of 1 eruption in 200 years
# prob = 1 - p(0 events, avg of 3.2) - p(1 event, avg of 3.2)
ans = 1 - dpois(0,3.2) - dpois(1,3.2)
ans
```

```
## [1] 0.8287987
```

A university classroom has 17 students. Six of these students are frequent readers of mystery novels. According to a national poll, 10% of the population are frequent readers of mystery novels.

8. Can you rule out the hypothesis that students in the classroom are randomly drawn from the national population, as far as interest in mystery novels is concerned? Provide a p-value. (Use an exact method to calculate the p-value for full credit, or an approximate method for partial credit.) [**4**]

2 / 4

```
#Problem can be modeled as a binomial distribution with n = 17, p = 0.1
#p-value = P(X=6)
p = dbinom(6,17,0.1)
p
```

dbinom() gives you the density, not a p-value.Fo

```
## [1] 0.00388372
```

Since 0.003 < p-value < 0.05; the hypothesis is marginally significant. Accept null hypothesis.

0.0038 is a very low p-value: you would almost certainly rule out the null hypothesis

P2: 12 / 20

Problem 2. Genetic screening

In a large medical study, 30000 individuals have their genomes sequenced and the presence or absence of 26 specific genetic mutations (each labeled by a letter from A-Z) is extracted. Of these individuals, a small number developed a certain type of cancer. Mutation “A” is already known to be associated with a higher risk of this cancer, but the others are still under study.

The file `cancer.csv` contains the study results. The first column is the participant ID number. Each of the letter columns indicates whether that specific mutation is present in the individual (0=no, 1=yes); the rightmost column indicates whether the individual developed the cancer (0=no, 1=yes).

```
df = read.csv('cancer.csv')
dim(df)
```

```
## [1] 30000    28
```

```
#show(df)
```

1. Produce a simple table (1x26, plus row/column labels) giving the number of individuals in the study with each of the 26 mutations. Which mutation is the most common? [**3**]

```
df_new = df[,2:27]
number_of_individuals = colSums(df_new)
number_of_individuals
```

3 / 3

```
##    A    B    C    D    E    F    G    H    I    J    K    L    M    N    O    P 
## 1559  583  804  458  470  131   94  317  214   61   26 1134   16  451  345 1124 
##    Q    R    S    T    U    V    W    X    Y    Z 
##  483  107   36  438   34   40  281  250  791   61 
```

```
number_of_individuals[order(number_of_individuals, decreasing=TRUE)[1]]
```

```
##    A 
## 1559 
```

2. How many individuals in the study developed this cancer? [**1**]

1 / 1

```
cancer_total = sum(df$cancer=='1')
cancer_total
```

```
## [1] 105
```

1 / 1

3. What proportion of individuals in the study developed this cancer? [**1**]

```
cancer_total/30000*100
```

```
## [1] 0.35
```

0.35% of the 30,000 individuals developed this cancer.

4. Estimate the probability of developing this cancer for an individual *with* mutation A. Also estimate the probability for an individual *without* mutation A. (You do not need to supply a confidence interval.) [**2**]

```
df_with_A = df[df$A==1,]
cancer_with_A = colSums(df_with_A)[28]
people_with_A = nrow(df_with_A)
probability_cancer_A = cancer_with_A/people_with_A
probability_cancer_A
```

2 / 2

```
##     cancer
## 0.01026299
```

```
df_without_A = df[df$A==0,]
cancer_without_A = colSums(df_without_A)[28]
people_without_A = nrow(df_without_A)
probability_cancer_without_A = cancer_without_A/people_without_A
probability_cancer_without_A
```

```
##      cancer
## 0.003129285
```

Therefore, probability of developing cancer for an individual with mutation A is 0.01026 and probability of developing cancer for an individual without mutation A is 0.00313.

0 / 4

5. Confirm that individuals with mutation A have a significantly higher cancer risk than those without mutation A using a categorical contingency test, and provide a p-value. [**4**]

6. Using a regression-like method with error family appropriate for this data type, search for any other genetic mutations (besides mutation A) which may also be associated with higher cancer risk. List these clearly in your answer, accounting for the many hypotheses under consideration with an appropriate post-hoc adjustment such as Bonferroni's correction. Do not consider interactions between mutations. [**9**]

5 / 9

```
regression_like = glm(df$cancer~. , data = df, family = 'binomial')
```

good (but, should have exclude

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(regression_like)
```

```
##
## Call:
## glm(formula = df$cancer ~ ., family = "binomial", data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4260  -0.0802  -0.0769  -0.0737   3.6409
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.708e+00  2.035e-01 -28.051  < 2e-16 ***
## id          -8.175e-06  1.134e-05  -0.721   0.4710  <-- ID number is not an explanatory variable
## A            1.197e+00  2.734e-01   4.379 1.19e-05 ***
## B           -7.019e-01  1.006e+00  -0.697   0.4856
## C           -3.439e-01  7.157e-01  -0.480   0.6309
## D            6.270e-01  5.890e-01   1.065   0.2871
## E            1.187e-01  7.190e-01   0.165   0.8688
## F            7.326e-01  1.013e+00   0.723   0.4695
## G           -1.381e+01  1.091e+03  -0.013   0.9899
## H           -1.532e-01  1.008e+00  -0.152   0.8792
## I           -1.384e+01  7.200e+02  -0.019   0.9847
## J           -1.396e+01  1.348e+03  -0.010   0.9917
## K           -1.385e+01  2.095e+03  -0.007   0.9947
## L            4.215e-01  4.226e-01   0.997   0.3185
## M           -1.363e+01  2.660e+03  -0.005   0.9959
## N            2.259e-01  7.166e-01   0.315   0.7526
## O            5.268e-01  7.176e-01   0.734   0.4629
## P           -7.080e-01  7.151e-01  -0.990   0.3222
## Q            1.631e+00  3.722e-01   4.381 1.18e-05 ***
## R           -1.381e+01  1.025e+03  -0.013   0.9892
## S           -1.375e+01  1.753e+03  -0.008   0.9937
## T           -4.205e-01  1.007e+00  -0.418   0.6763
## U            2.258e+00  1.026e+00   2.201   0.0277 *
## V           -1.378e+01  1.675e+03  -0.008   0.9934
## W            1.015e-01  1.008e+00   0.101   0.9198
## X           -1.392e+01  6.627e+02  -0.021   0.9832
## Y            3.875e-01  5.126e-01   0.756   0.4497
## Z           -1.373e+01  1.366e+03  -0.010   0.9920
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1397.2  on 29999  degrees of freedom
## Residual deviance: 1354.7  on 29972  degrees of freedom
## AIC: 1410.7
##
## Number of Fisher Scoring iterations: 18
```

No post-hoc adjustment applied (U i

It becomes evident from the results that besides mutation A, mutation Q is also associated with high cancer risk. On the other hand, mutation U shows some association with cancer risk as well.

Note: You will likely want to suppress any very long output sequences in your PDF to avoid exceeding the page limit. Don’t suppress any calculations or results critical to your conclusions.

12 / 20

Problem 3. Hormone Suppression

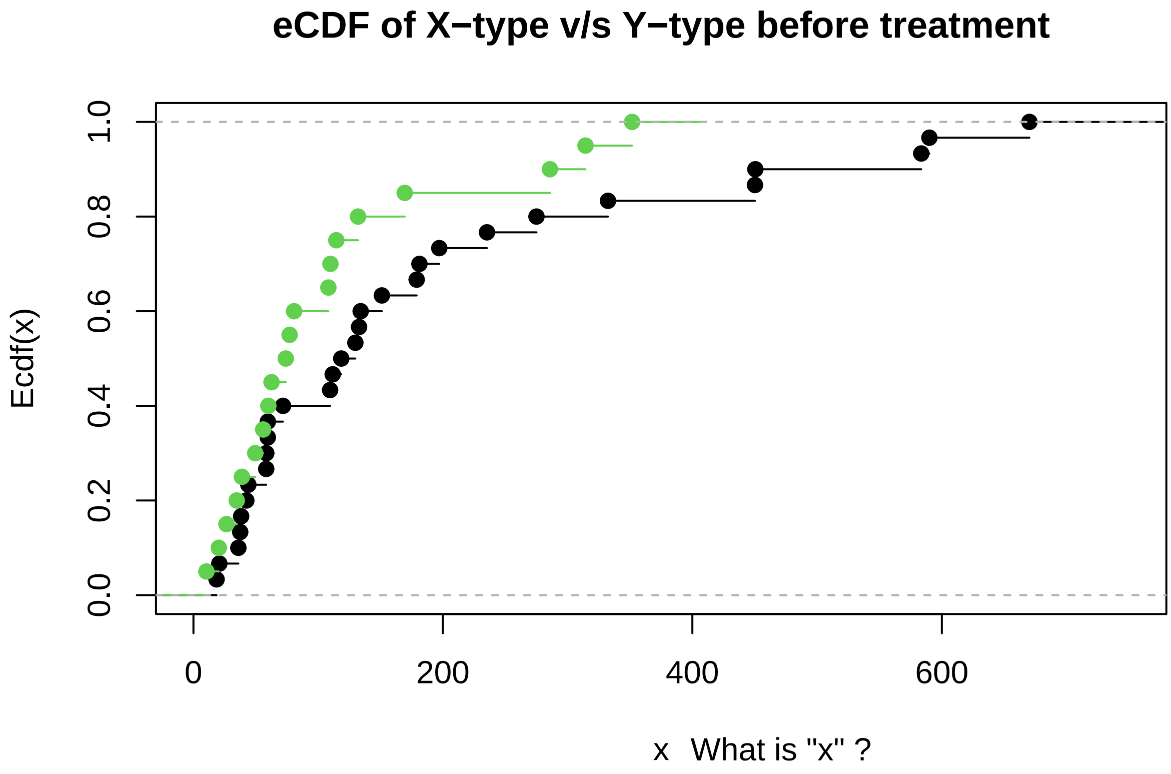
A laboratory study is examining the effectiveness of a new drug for reducing thyroid hormone production. The hormone levels in fifty cell cultures are measured once before treatment, and again after treatment. There are two types of cell cultures, type X and type Y. Data (including sample ID number, culture type, and hormone levels) are stored in `thyroid.csv`.

```
df = read.csv('thyroid.csv')
df_x = subset(df,df$type=='X') # dim(df_x) to verify the new subset df
df_y = subset(df,df$type=='Y') # Similarly, dim(df_y)
```

1. Plot the empirical cumulative distribution function of the pre-treatment hormone values for X-type cultures. Overplot the empirical cumulative distribution function of the same quantity for Y-type cultures as a different colour. The X-type curve should be black, the Y-type curve should be a different colour and/or style, and the axes should be labeled informatively. [**4**]

```
plot(ecdf(df_x$hormone.before),xlim=c(0,750), main='eCDF of X-type v/s Y-type before treatment' , ylab=
lines(ecdf(df_y$hormone.after), col=c(3))
```

3 / 4



What is "x" ?

2. Use a nonparametric test to examine whether X and Y type cultures have different distributions of hormone levels (pre-treatment). State your conclusion (with p-value). [**2**]

2 / 2

```
ans = wilcox.test(df_x$hormone.before,df_y$hormone.before)
ans$p.value
```

```
## [1] 0.4855889
```

(Do not reject)

P-value: 0.4856; therefore, there is no significant difference. Accept null hypothesis.

For questions 3 and onward, consider type X and type Y cultures together.

3. Use a normality test to determine whether the hormone levels in the pre-treatment cell cultures are consistent with being distributed normally, and state your conclusion (with p-value). [**2**]

```
s = shapiro.test(df$hormone.before)
s$p.value
```

```
## [1] 4.292817e-07
```

2 / 2

```
s$p.value < 0.05
```

```
## [1] TRUE
```

p-value = 4.293e-07; Assuming Alpha is 0.05, since the p-value is less than alpha it can be concluded that the data is not normally distributed.

4. Apply a logarithm transform to the pre-treatment hormone levels. Perform another normality test on these transformed values, and state your conclusion (with p-value). [**2**]

2 / 2

```
lop_pre = log(df$hormone.before)
s_log_pre = shapiro.test(lop_pre)
s_log_pre$p.value
```

```
## [1] 0.6808698
```

p-value: 0.6808698; Since p-value is greater than 0.05(Alpha), the transformed values form a normal distribution.

5. Examine whether the treatment is associated with a significant change in average log-transformed thyroid levels. State your conclusion (with p-value). [**6**]

2 / 6

```
ans = lm(df$hormone.after~ lop_pre)
summary(ans)
```

This is correlating the log-transformed treatment values a

```
##
## Call:
## lm(formula = df$hormone.after ~ lop_pre)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -83.80 -49.24 -18.29  27.16 212.00
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -489.12      50.34  -9.717 6.43e-13 ***
## lop_pre       135.03      10.58  12.766  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 70.36 on 48 degrees of freedom
## Multiple R-squared:  0.7725, Adjusted R-squared:  0.7677
## F-statistic:   163 on 1 and 48 DF,  p-value: < 2.2e-16
```

Treatment is highly significantly associated with a change in average log-transformed thyroid levels. p-value: $<$ 2e-16 (Assuming alpha=0.05).

6. By what *percent* (on average) did the treatment increase or reduce thyroid hormone levels? State a value and a confidence interval. [**4**]

```
t.test(df$hormone.after,lop_pre, paired = TRUE)
```

Should have done this for Q5 (but with transforme

1 / 4

```
##
##  Paired t-test
##
## data:  df$hormone.after and lop_pre
## t = 6.6314, df = 49, p-value = 2.471e-08
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##   94.87793 177.38408
## sample estimates:
## mean difference
##         136.131
```

mean difference decrease in thyroid hormone level: 136.131 ; Confidence interval: 94.878 to 177.384.

P4: NA

Problem 4. V for Viscosity

You are trying to calibrate an instrument that measures the viscosity of a substance. Your instrument produces raw measurement readings R, which you want to use to predict the actual viscosity V. The instrument is trialed on seven different substances of known viscosity V and the results are as follows:

V	R
1.04	0.63
1.39	0.89
1.44	0.93
2.08	1.38
2.61	1.73
3.63	2.34

You have reason to believe that the equation relating V and R is one of the following:

Model 1: $V(R) = \sqrt{I + A * R^2}$
Model 2: $V(R) = \sqrt{I + A * R^2 + B * R^4}$
Model 3: $V(R) = \sqrt{I + A * R^2 + B * R^4 + C * R^6}$

1. Perform a mathematical transform on V and/or R to transform the nonlinear relationships above into linear ones on the transformed variables. Write down the new, linearized equations in terms of the transformed variables (use latex if possible to write the new equations, or just write them as R code). [**5**]
2. Use a linear regression to fit each of the above three models to the data. Then output the best-fit parameter values by filling in the table below the code chunk. [**5**]

model	I	A	B	C
1			N/A	N/A
2				N/A
3				

3. Determine which of the three possible models best explains the data. State explicitly what this model is. [**5**]
4. Make a plot of the data (in terms of the original, un-transformed variables: R versus V), overplotting your best-fit model as a well-sampled curve. [**3**]
5. Following calibration, you obtain a reading of a new substance with unknown viscocity: it is R = 3.0. Estimate the value of V for this substance (with prediction interval). [**2**]