MAST90044 Thinking and Reasoning with Data
Semester 1 2019
Assignment 2 Solutions

Q1

(a)
```
> binom.test(x = 183, n = 400, p = 0.42, alternative = "greater", conf.level = 0.95)

Exact binomial test

data:  183 and 400
number of successes = 183, number of trials = 400, p-value = 0.0713
alternative hypothesis: true probability of success is greater than 0.42
95 percent confidence interval:
0.4156136 1.0000000
sample estimates:
probability of success
0.4575
```

The p-value is 0.0713. We do not have enough evidence to reject the null hypothesis. I.e., based on our sample we do not have enough evidence to claim that the frequency of the contaminant has increased.

```
> prop.test(183,400,0.42, alternative ="greater")

1-sample proportions test with continuity correction

data:  183 out of 400, null probability 0.42
X-squared = 2.1577, df = 1, p-value = 0.07093
alternative hypothesis: true p is greater than 0.42
95 percent confidence interval:
0.4157171 1.0000000
sample estimates:
p
0.4575
```

The p-value is 0.07093. Same conclusion as exact test.

Total 6 pt: 1pt for using binom.text; 1pt for using propr.test; 1pt for each correct p-value; 0.5pt for "alternative = "less"" in binom.test and 0.5pt for "alternative = "less"" in prop.test; 1pt for correct conclusion at each of the tests.

(b) Type I error - conclude that there is a difference when there is no real difference between the historical level of contaminant and the current one. In this case a large type I error

may lead to an unnecessary financial investment in water cleaning or a ban on the water use.

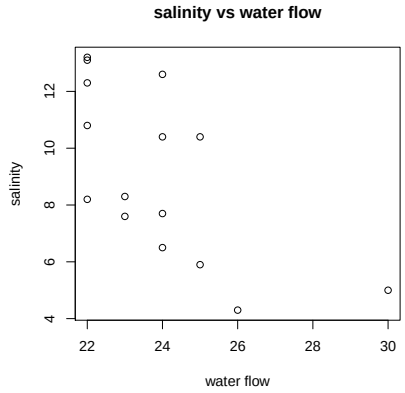
Type II error - conclude that there is no any difference when actually the current contaminant level is higher than the historical one. In this case a large type II error may put the water consumers in danger.

Total 4 pt: 1pt for each correct definition and 1 pt for each correct interpretation of the error in the context.

[6 + 4 = 10 marks]

Q2

(a)
```
> x <- c(23, 24, 26, 25, 30, 24, 23, 22, 22, 24, 25, 22, 22, 22, 24)
> y <- c(7.6, 7.7, 4.3, 5.9, 5, 6.5, 8.3, 8.2, 13.2, 12.6, 10.4, 10.8,
       + 13.1, 12.3, 10.4)
> plot(y ~ x, xlab = "water flow", ylab = "salinity", main = "salinity vs water flow
```



1pt for using using scatter plot. 1pt for including title, 1pt for axes titles

(b) $y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, \ldots, 15, \quad e_i \sim \mathrm{N}(0, \sigma)$

```
> salinity.lm <- lm(y ~ x)
> summary(salinity.lm)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8444 -2.2153  0.0138  1.9138  3.6347

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  30.8163     7.0248   4.387 0.000735 ***
x            -0.9105     0.2932  -3.105 0.008369 **
---
```

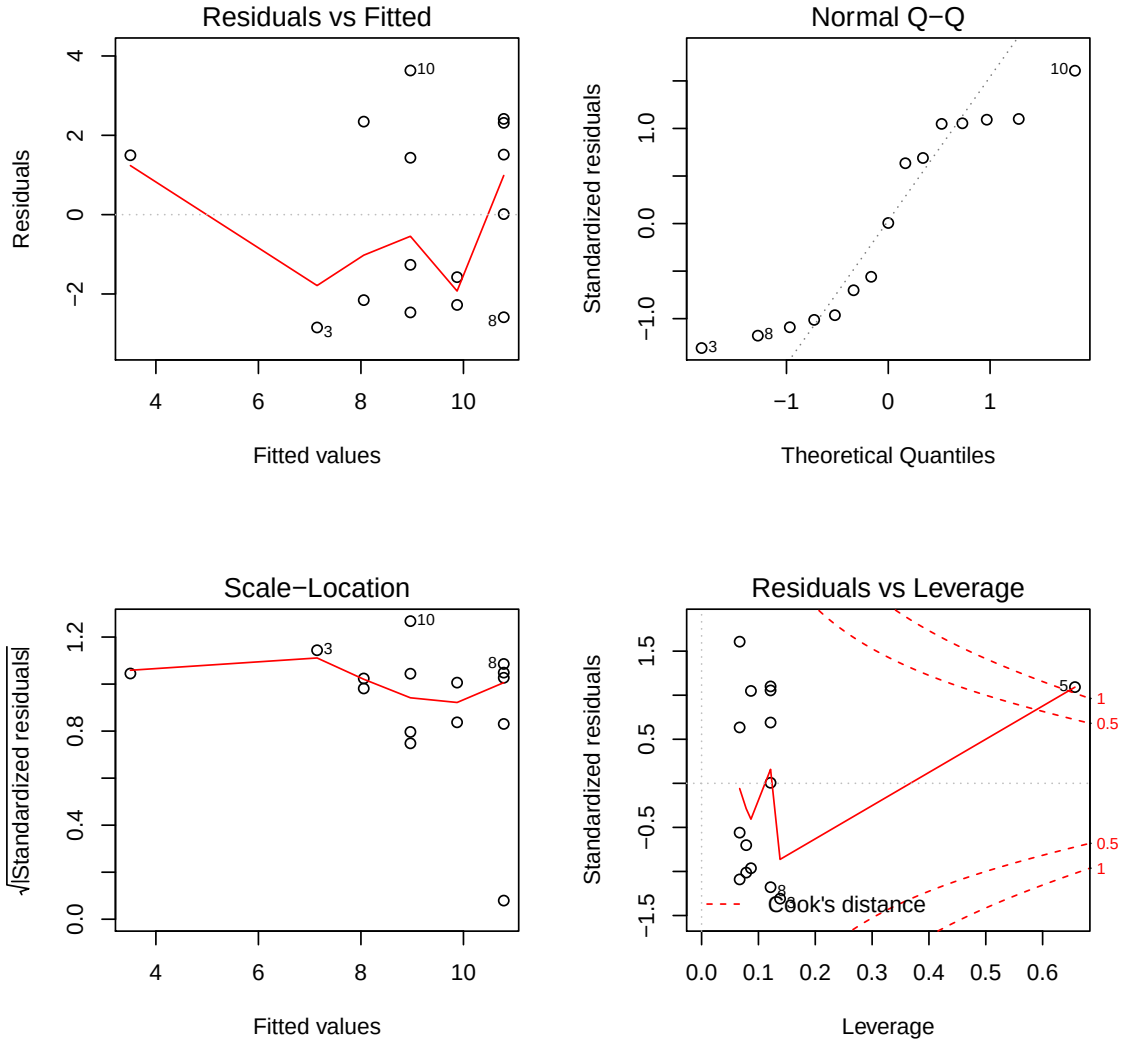
```
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.341 on 13 degrees of freedom
Multiple R-squared:  0.4258,Adjusted R-squared:  0.3816
F-statistic:  9.64 on 1 and 13 DF,  p-value: 0.008369
```

$r = -\sqrt{R^2} = -\sqrt{0.4258} = -0.653.$

~~1pt~~ for running regression; 1pt for running summary; 2pt for calculating r

(c)
```
> par(mfrow=c(2,2))
> plot(salinity.lm)
```



The QQ plot challenges the assumption of normality of the errors. The fifth observation $(30, 5.0)$ is a point of high leverage.

1 pt for correct plots; 2 pt for recognizing that normality assumptions not hold; 2pt saying that the 5th obs. has high leverage.

(d)
```
> confint(salinity.lm, level = 0.99)
                 0.5 %     99.5 %
(Intercept)  9.655786 51.9768501
x           -1.793786 -0.0271341
```

The estimated slope (-0.91) is useful because it indicates the expected decrease in salinity for each additional unit of water flow. The estimated intercept is not useful because it is the predicted salinity when there is no water flow (which appears unlikely).

~~1 pt~~ for calculating CI; 1pt for recognizing the slop estimate; 1pt for saying useful plus a sensible explanation; 1pt for recognizing the intercept estimate; 1pt for saying useful plus a sensible explanation;

(e)
```
> predict(salinity.lm, newdata = data.frame(x = 21), interval = "prediction")
       fit      lwr      upr
1 11.69665 6.166569 17.22674
```

If the water flow was 21, we would be 95% confident that the salinity would be between 6.2 1nd 17.2.

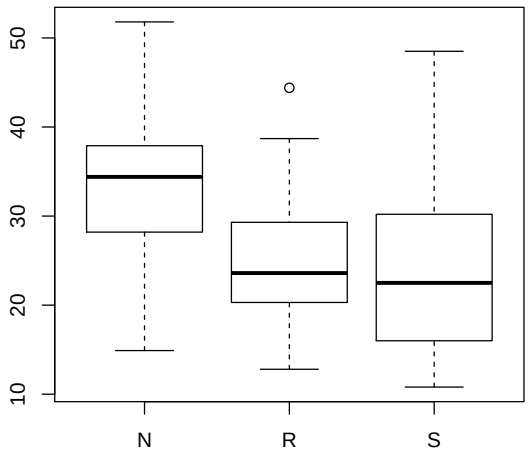
2pt for correct usage of "predict", 1pt for correct confidence interval.

$[3 + 5 + 5 + 6 + 3 =$ 22 marks$]$

Q3

```
> fruitfly <- read.csv("fruitfly.csv")
```

(a)
```
> boxplot(fecundity~line,data=fruitfly)
```



Fecundity is generally higher for the control than for the two selected lines. R and S appear to be similar in location but S has greater spread.

1pt for the plot; for saying that "Fecundity is generally higher" 1pt , similar location for R and S 1pt, spread of N is larger 1pt.

(b) Model: $y_{ij} = \mu_i + e_{ij}$, $e_{ij} \sim N(0, \sigma)$ where
y_{ij} denotes fecundity, $i = 1, 2, 3$ denotes genetic line, $j = 1, \ldots, 25$ denotes female.
The null hypothesis is $H_0 : \mu_1 = \mu_2 = \mu_3$.

2 pt for the model, 1pt for specifying the error distribution, 1pt for definitions of the variables and the index. 2pt for the correct H_0

(c)
```
> fecundity.lm <- lm(fecundity ~ line, data = fruitfly)
> anova(fecundity.lm)
Analysis of Variance Table

Response: fecundity
          Df Sum Sq Mean Sq F value    Pr(>F)
line       2 1362.2  681.11  8.6657 0.0004244 ***
Residuals 72 5659.0   78.60
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

or

```
> summary(fecundity.lm)

Call:
lm(formula = fecundity ~ line, data = fruitfly)

Residuals:
Min      1Q  Median      3Q     Max
-18.472  -5.764  -0.728   4.436  24.872

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)   33.372      1.773  18.821  < 2e-16 ***
lineR         -8.116      2.508  -3.237 0.001829 **
lineS         -9.744      2.508  -3.886 0.000224 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 8.866 on 72 degrees of freedom
Multiple R-squared:  0.194,Adjusted R-squared:  0.1716
F-statistic: 8.666 on 2 and 72 DF,  p-value: 0.0004244
```

The null hypothesis is rejected, from which it can be concluded that the mean fecundity is not the same for the three lines.

1pt for regression code, 1pt for anova code or summary code plus mentioning the F-statistic and the p-value $< 5\%$, 1pt for saying "reject", 1pt the interpretation

(d)
```
> fruitfly.RS <- fruitfly[1:50, ]
> t.test(fecundity ~ line, data = fruitfly.RS)
```

```
Welch Two Sample t-test

data:  fecundity by line
t = 0.6521, df = 45.693, p-value = 0.5176
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.39843  6.65443
sample estimates:
mean in group R mean in group S
         25.256          23.628
```

The p-value of the t-test is 0.5176, the confidence interval includes zero. Based on the p-value of the t-statistic and the confidence interval we can not reject the null hypothesis. I.e. we do not have enough evidence to claim that the means of the two lines are significantly different.

1pt for taking sub-sample, 1pt for the correct t.test, 1pt for correct P-value. 1pt fpr saying "reject", 2pt for explanation that we have not enough evidence.

$[4 + 6 + 4 + 6 = \quad 20 \text{ marks}]$