STAT 1201 – Summer Semester, 2022

Short answers for STAT1201 – Semester 2 2020 Final Exam

Scenario 1 - Tuatara Mass

Question 1

H_0: μ_A=μ_B vs H_1: μ_A≠μ_B

Question 2

Difference = 617.7- 585.8 = 31.9

Question 3

The standard error of difference between sample means; $se(\bar{x}_A - \bar{x}_B)$ = $\sqrt{\frac{38.67^2}{9} + \frac{23.39^2}{14}}$ = 14.32585

$se(\bar{x}_A - \bar{x}_B) = 14.33\ g$ for two decimal places.

Question 4

$$t_{stat} = \frac{(\bar{x}_A - \bar{x}_B) - 0}{se(\bar{x}_A - \bar{x}_B)}$$

$$t_{stat} = \frac{(617.7 - 585.8) - 0}{14.32585}$$

$t_{stat} = 2.226744.$ (For 3 decimal places, the answer is 2.227)

Question 5

The degrees of freedom are; min(14-1, 9-1) = 8

Question 6

p-value = 2*pt(-2.226744, 8) = 0.057

0.05 < p-value < 0.1 ------ Weak evidence to suggest that there is a difference in mean tuatara mass between the two locations.

Question 7

Margin of error in a 95% Confidence Interval is given by

$t^* \times se(\bar{x}_A - \bar{x}_B)$ where $t^* = qt(0.975, df = 8) = 2.306004$

MOE = 2.306004 $*$14.32585 = 33.03547

MOE = 33.0 g (for one decimal place)

Question 8

Refresh the assumptions of two-sample t-test.

Answer is Normal variability

Question 9

IQR = 621.6 – 597.8 = 23.8

Q1 – 1.5*IQR = 597.8-1.5*23.8 = 562.1

Q1 + 1.5*IQR = 621.6+1.5*23.8 = 657.3

714.5 > 657.3.

Yes, at least one to the right.

Question 10

Location A has 9 Tautara

Location B has 14 Tautara

Define W = sum of the ranks for location A

E(W) = $\frac{9*(9+14+1)}{2} = 108$

Question 11

SD(W) = $\sqrt{\frac{9*14*(9+14+1)}{12}}$ =15.87451

SD(W) = 15.87 (for 2 decimal places)

Question 12

$$z_{stat} = \frac{145 - 108}{15.87451} = 2.330781$$

p-value = 2*pnorm(-2.330781)= 0.01976491

p-value < 0.05. That is, moderate evidence to conclude that tuatara mass distributions are different in two locations.

Scenario 2 – River Pollution

First read the data file into RStudio

river = read.csv("River.csv")

Question 1

Using median(river$Elevation) in RStudio, median = 8.45m

Question 2

Draw a boxplot in R.

boxplot(river$Elevation)

Skewed to the left.

Question 3

aggregate(LogCadmium~LandUse, data=river, mean) gives the table of mean LogCadmium by land use.

-0.0243 (for 4 decimal places) for Fruit tress

Question 4

summary(aov(LogCadmium~LandUse, data=river)) gives the ANOVA table. Consider the column ‘Df’.

Question 5

Consider the ‘Sum Sq’ column of the ANOVA table.

SST = SSG + SSR

SST = 4.688 + 16.896 = 21.584

Question 6

Consider the ‘Pr(>F)’ column of the ANOVA table.

p-value = 0.000274 < 0.01. Strong evidence ….

Question 7

Consider the ‘Sum Sq’ column of the ANOVA table.

$$R^2 = \frac{SSG}{SST} = \frac{4.688}{21.584} = 0.2171979$$

$R^2 = 0.2172$ (for 2 decimal places)

Question 8

3 flooding levels. Degrees of freedom = (3-1) =2

Question 9

First transform Flooding a factor variable.

river$Flooding=factor(river$Flooding)

Then run ANOVA.

summary(aov(LogCadmium~Flooding, data=river))

Consider the ‘F value’ column of the ANOVA table.

F_{stat} = 51.72

Question 10

R^2 for the LogCadmium and Flooding model is 13.099/21.584 = 0.6069

Flooding gives a larger R^2 than LandUse.

Question 11

This is a simple linear regression question.

Dependent variable – LogCadmium

Independent variable – Distance

summary(lm(LogCadmium ~ Distance, data=river)) gives the table of regression coefficient estimates.

Considering ‘Estimate’ column; b_1= -0.00161

Question 12

Consider ‘Pr(>|t|)’ column in the table from Question 11.

p-value corresponding to Distance is very small (<0.01). Strong evidence to conclude there is an association between LogCadmium and Distance.

Question 13

$MOE = t^* \times se(b_1)$

t^* can be found using R: qt(0.975, 68) = 1.995469.

se(b_1) can be found from ‘Std. Error’ column of the Question 11 R output.

MOE = 1.995469 x 0.0002054 = 0.00041

Question 14

Use predict() or substitute Distance = 270 in the estimated regression equation in Question 11.

predict(lm(LogCadmium ~ Distance, data=river), newdata=data.frame(Distance = 270)) gives 0.2557189.

The answer for 3 decimal places is 0.256.

Question 15

From Question 14, $\log_{10}$(Cadmium) = 0.2557189

Estimated Cadmium concentration = $10^{(0.2557189)}$ = 1.8 (for 1 decimal place)

Question 16

This is a multiple linear regression question.

Dependent variable – LogCadmium

Independent variables – Distance and Elevation

summary(lm(LogCadmium ~ Distance+Elevation, data=river)) gives regression coefficient estimates, their standard errors, t-statistics and corresponding p-values.

Coefficient estimate for Elevation = -0.2554673 = -0.25547 (for 5 decimal places)

Question 17

Use predict() in R or substitute Distance = 270 and Elevation = 6.2 in the estimated regression equation is Question 16.

predict(lm(LogCadmium ~ Distance+Elevation, data=river), newdata=data.frame(Distance=270, Elevation=6.2)) gives the estimated logCadmium = 0.7510874

Estimated logCadmium = 0.751 (for 3 decimal places)

Question 18

From the data, the LogCadmium when Distance = 270 and Elevation = 6.2 is 0.462

From Qestion 17, the estimated LogCadmium level when Distance = 270 and Elevation = 0.751

Residual = 0.462 – 0.751 = -0.289

Question 19

Consider 'Pr(>|t|)' column in the table from Question 16, p-value corresponding to the variable Elevation is less than 0.01.

Thus, strong evidence to suggest that there is an association between log-transformed cadmium concentration and site elevation, after taking into account the distance from the river.

Question 20

Higher closer to the river and decreases with higher elevations.

One way to find calculate the fitted values of Cadmium concentration.

river$Fitted = 10^predict(lm(LogCadmium ~ Distance+Elevation, data=river))

View(river) and look at the Fitted column

Scenario 3 – Migraine Headache

Question 1

Comparative experiment

The study is designed to compare the differences in the effects of two different treatments.

Read the data file into RStudio.

migraine = read.csv("Migraine.csv")

Question 2

mean(migraine$Days0) gives 9.2125

9.21 for 2 decimal places.

Question 3

Number of observations (n) = 160

95% confidence interval for the mean number of headache days of participants in the first month of the study is given by

$\bar{x} \pm t^{*} se(\bar{x})$

$se(\bar{x}) = \frac{s}{\sqrt{n}}$

Using sd(migraine$Days0) in R, *s*= 2.999764

*t** = qt(0.975, 159) = 1.974996

95% CI

$9.2125 \pm 1.974996 * \frac{2.999764}{\sqrt{160}}$

9.2125 ± 0.4683747

(8.744125, 9.680875)

(8.74, 9.68) for 2 decimal places

Question 4

Using Welch t-test

t.test(Age~Group, data=migraine) gives p-value = 0.6203

Thus, no evidence to suggest that there is a difference in mean age between the BWL and ME groups

Question 5

Using table() function in R; table(migraine$Education)

Participants with a Masters degree = 28

Question 6

Using prop.table() function in R; prop.table(table(migraine$Education))

Proportion of participants with a Masters degree = 0.175

Question 7

First find the observed numbers in each cell of the contingency table created from the two variables of Group and Education.

created from two variables of Group and Education.

addmargins(table(migraine$Group, migraine$Education))

	Bachelors	HighSchool	Masters	Sum
BWL	37	33	10	80
ME	19	43	18	80
Sum	56	76	28	160

Expected count for subjects in the BWL group having had a Masters degree = 80 x 28/160 =14

Alternatively, we can use Chi-squre test to find the expected number.

chisq = chisq.test(table(migraine$Group, migraine$Education))

chisq$expected

Question 8

chisq$statistic gives 9.387218

Chi-square stat = 9.39 for 2 decimal places

Question 9

chisq$p.value gives 0.009153591

p-value < 0.01

Thus, strong evidence to suggest that there is an association between education level and treatment group.

Question 10

The difference between means of Days0 and Days4 variables.

mean(migraine$Days0)-mean(migraine$Days4) gives 3.375.

Mean reduction = 3.38 (for 2 decimal places)

Question 11

First, create a new variable called Change = Days0 – Days4

Define μ as the population mean reduction in the number of headache days for subjects in the study between the first month and the fourth month.

We need to test

H_0: $\mu_{BWL} = \mu_{ME}$ Vs H_1: $\mu_{BWL} > \mu_{ME}$

Using t.test() in R; t.test(Change~Group, data=migraine, alternative="greater")

t_{stat} = 1.5908

t_{stat} = 1.591 (for 3 decimal places)

Question 12

From the results in Question 11, p-value = 0.05686, which is between 0.05 and 0.1.

Thus, weak evidence to suggest that the behavioural weight loss (BWL) intervention is more effective than the migraine education control at reducing migraine headache days.

Scenario 4 – Infection

Question 1

E(X) = (0 x 0.25) + (1 x 0.20) + (2 x 0.30) + (3 x 0.25) = 1.55

Question 2

Var(X) = $\sum_x P(X = x)(x - E(X))^2$

Var(X) =$0.25(0 - 1.55)^2 + 0.20(1 - 1.55)^2 + 0.30(2 - 1.55)^2 + 3(3 - 1.55)^2$

Var(X) = 1.2475

SD(X) = $\sqrt{1.2475}$

SD(X) = 1.116915 = 1.117 (for 3 decimal places)

Question 3

$Y = X_1 + X_2 + \cdots . + X_{60}$

$E(Y) = 60 * E(X)$

$E(Y) = 60 * 1.55 = 93$

Question 4

$SD(Y) = \sqrt{60} \times SD(X)$

$SD(Y) = \sqrt{60} \times 1.116915 = 8.651586$

SD(Y) = 8.652 (for 3 decimal places)

Question 5

We need to find $P(Y \geq 104)$

Assuming normal approximation $Y \sim N(\mu = 93, \sigma = 8.652)$

$$P(Y \geq 104) = P(Z \geq \frac{104 - 93}{8.651586})$$
$$P(Y \geq 104) = P(Z \geq 1.271443)$$

In R: 1-pnorm(1.271443) = 0.1017855

$P(Y \geq 104)$ = 0.102 (for 2 decimal places)

Question 6

The count of the total people infected has a Binomial distribution.