

Submit your assignment

[Help](#)

After you have completed the assignment, please save, scan, or take photos of your work and upload your files to the questions below. Crowdmark accepts PDF, JPG, and PNG file formats.

Q1 (10 points)

In a behavioral study of the fruitfly *Drosophila melanogaster*, a biologist measured for individual flies, the total time spent preening during a six-minute observation period. The following are the preening times (in seconds) for 99 flies:

32 24 10 16 52 76 33 31 46

`flies<-c(32, 24, 10, 16, 52, 76, 33, 31, 46)`

(a) (2 pts) What are the lower and upper fences? Are there any outliers?

Quartiles

$$Q_1 = 20, \quad Q_3 = 49$$

$$IQR = Q_3 - Q_1 = 29$$

$$Lower\ Fence = Q_1 - 1.5 \times IQR = 20 - 1.5 \times 29 = -23.5$$

$$Upper\ Fence = Q_3 + 1.5 \times IQR = 49 + 1.5 \times 29 = 92.5$$

There are no outliers, as all the values are between Lower and upper fence.

(b) (2 pts) Construct a 95% confidence interval for the population mean.

Mean

$$\bar{x} = \frac{\sum x}{n} = \frac{320}{9} = 35.56$$

Standard deviation

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}} = \sqrt{\frac{3224.22}{8}} = 20.08$$

For 95% , $z_{0.05} = 1.96$

The 95% confidence for the population mean is

$$C.I. = \left(\bar{x} \pm \frac{z_{\alpha} \times s}{\sqrt{n}} \right) = 35.56 \pm \frac{1.96 \times 20.08}{\sqrt{9}}$$

$$C.I. = 35.56 \pm 13.12$$

$$C.I. = (22.44, 48.68)$$

(c) (2 pts) Interpret the confidence level 95% through a hypothetical scenario in which a large number of samples, each of size 9, could be taken from the population.

We are 95% confidence that the true population mean lies between 22.44 and 48.68.

(d) (2 pts) What is the sample proportion of flies whose preening times exceed 40 seconds?

Sample size whose preening times exceeds 40 seconds is 3,

as there 3 out of 9 values exceed 40

(e) (2 pts) Suppose that we are interested in p, the true proportion of flies in the population whose preening times exceed 40 seconds. How large should the sample size n of another larger scale study be, such that the margin of error does not exceed 2% at the 95% level of confidence? (It is your option to use or not use the sample proportion in the previous part).

We are given that margin of error <2%

Margin of error is given by

$$\text{Margin of error} = z_{\alpha} \times \sqrt{\frac{p(1-p)}{n}}$$

In our case n = 3 , for sample size exceeding preening time exceed 40 seconds

Putting values,

$$\frac{2}{100} = 1.96 \sqrt{\frac{p(1-p)}{3}}$$

Solving, we get

$$P = 0.99$$

Q2 (8 points)

Following up the last question, assume now that the preening time (during a six-minute observation period) in the population follows a normal distribution with mean $\mu=30$ seconds and standard deviation $\sigma=16$ seconds.

(a) (2 pts) What is the probability that the preening time of a fly, randomly selected from the population, is between 28 seconds and 32 seconds?

Since $\mu = 30, \sigma = 16$

$$P(28 < X < 32) = P\left(\frac{28 - 30}{16} < \frac{X - \mu}{\sigma} < \frac{32 - 30}{16}\right)$$

$$\rightarrow P(-0.13 < Z < 0.13)$$

Using standard normal table

$$P(28 < X < 32) = 0.1034$$

(b) (2 pts) If we take another sample of size $n=16$, what is the probability that the average preening time of the sample is between 28 seconds and 32 seconds?

$$\mu = 30, \sigma = \frac{16}{\sqrt{16}} = 4$$

$$P(28 < X < 32) = P\left(\frac{28 - 30}{4} < \frac{X - \mu}{\sigma} < \frac{32 - 30}{4}\right)$$

$$\rightarrow P(-0.5 < Z < 0.5)$$

Using standard normal table

$$P(28 < X < 32) = 0.383$$

(c) (2 pts) If we take another sample of size $n=10$, what is the probability that **at least** 5 flies have a preening time longer than 30 seconds?

When 5 flies

$$\mu = 30, \sigma = \frac{16}{\sqrt{5}} = 7.16$$

$$P(X > 30) = P\left(\frac{X - \mu}{\sigma} > \frac{30 - 30}{7.16}\right)$$

$$\rightarrow P(Z > 0)$$

Using standard normal table

$$P(X > 30) = 0.5$$

(d) (2 pts) If 10 random samples are taken from this population, how many of them will yield a sample mean preening time longer than 30 seconds?

$$\mu = 30, \sigma = \frac{16}{\sqrt{10}} = 5.06$$

$$P(X > 30) = P\left(\frac{X - \mu}{\sigma} > \frac{30 - 30}{5.06}\right)$$

$$\rightarrow P(Z > 0)$$

Using standard normal table

$$P(X > 30) = 0.5$$

Q3 (6 points)

A laboratory blood test is 99% effective in detecting a certain disease when it is, in fact, present. However, the test also yields a "false positive" result for 2% of healthy persons tested. Assume that 5% of the population actually has the disease.

$$P(\text{has disease}) = 0.05$$

$$P(\text{tested positive}) = 0.95$$

$$P(\text{false positive}) = 0.02$$

(a) (2 pts) If a person is randomly selected from this population, then what is the probability that this person tests positive?

$$P(\text{tested positive}) = P(\text{has disease and tested positive}) + P(\text{does not have disease and tested positive})$$

$$P(\text{tested positive}) = [(0.05 \cdot 0.95) + ((1 - 0.05) \cdot 0.02)] = 0.0665$$

(b) (2 pts) What is the probability that this person does not have the disease given that the test result is positive?

$$P(\text{does not have disease and tested positive}) = (1-0.05) \times 0.02 = 0.019$$

(c) (2 pts) For five random selected people from the population, what is the probability that all of them test **negative**?

$${}^5C_5 P(\text{tested negative}) = 1 \times 0.05 = 0.05$$

Q4 (6 points)

Are wine drinkers being misled by bottle labels? The *London Times* (August 5, 2001) reported the results of a study that compared the actual alcohol content of wine to the alcohol content printed on the label. laboratory analysis was used to determine the alcohol content (percentage) of different brands and varieties selected at random from wines produced in 1999 and 2000. A representative subset (six bottles) of the resulting data is given in the following table:

Actual<-c(14.2, 14.5, 14.0, 14.9, 13.6, 12.6)

Label<-c(14.0, 14.0, 13.5, 15.0, 13.0, 12.5)

(a) (3 pts) What is the best test, paired t-test, or two-sample pooled t-test, or two-sample Welch's t-test, to use for answering the question?

Since all the data are not independent and the samples are related so we need to apply paired t test.

(b) (3 pts) Carry out the appropriate test. State the null and alternative hypotheses, specify the test statistic value and/or the P-value and state your conclusion in the context. Use $\alpha=0.05$.

Let μ_1 and μ_2 indicate the actual and label alcohol contents of wine, respectively. μ_d represents the difference $\mu_d = \mu_1 - \mu_2$

Null Hypothesis $H_0: \mu_d = 0$

Alternative Hypothesis $H_1: \mu_d > 0$

Paired T for actual - label

	N	Mean	StDev	SE Mean
actual	6	13.967	0.802	0.327
label	6	13.667	0.876	0.357
Difference	6	0.300	0.276	0.113

95% lower bound for mean difference: 0.073

T-Test of mean difference = 0 (vs > 0): T-Value = 2.67 P-Value = 0.022

Since the $p\text{-value} < 0.05$, We reject the null hypothesis. we conclude that the label underestimates the actual alcohol content significantly.

Q5 (6 points)

Kuru is a prion disease of the Fore people of highland New Guinea. Using archived tissue samples, Mead et al. (2009) investigated genetic variants that might confer resistance to kuru. The data in the following table are genotypes at codon 129 of the prion protein gene of young and elderly individuals all having the disease.

Age	MM	MV	VV
Elderly	13	77	14
Young	22	12	14

```
> kuru<-as.table(matrix(c(13,22,77,12,14,14), 2, 3))  
> chisq.test(kuru)
```

Pearson's Chi-squared test

```
data: kuru
```

```
X-squared = 33.733, df = ?, p-value = 4.73e-08
```

(a) (3 pts) Test whether genotype frequencies differ between the two age groups. State clearly the null and alternative hypotheses, specify the P-value, and state your conclusion in the context. Use $\alpha=0.05$.

Null hypothesis is that the genotype frequencies same between the two age groups.

The alternative hypothesis is that the genotype frequencies differ between the two age groups

From the result shown above
 $p\text{-value} = 4.73e-08 \sim 0.00$.

Since $p\text{-value} < 0.005$, we reject the null hypothesis. we conclude that the genotype frequencies differ between two age groups significantly.

(b) (1 pt) What are the degrees of freedom for the χ^2 -test?

Degrees of freedom = $(3-1) * (2-1) = 2$

(c) (2 pts) Construct a 90% confidence interval for the true proportion of genotype MM individuals among elderly patients.

Q6 (7 points)

Research was carried out to investigate the effect of storage time on the calcium content of wheat. Four different storage times were considered. The data are shown below.

Storage time	Observations						
0 month	58.75	57.94	58.91	56.85	55.21	57.30	
1 month	58.87	56.43	56.51	57.67	59.75	58.48	
2 months	59.13	60.38	58.01	59.95	59.51	60.34	
4 months	62.32	58.76	60.03	59.36	59.61	61.95	

```
> aovobj<-aov(calcium~storageTime, data=Calcium)
```

```
> summary(aovobj)
```

```

              Df Sum Sq Mean Sq F value    Pr(>F)
storageTime (A)  32.14      (B)      (C) 0.00298 **
Residuals   (D)  32.90    1.645
---
```

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

```
> TukeyHSD(aovobj, "storageTime")
```

```
Tukey multiple comparisons of means
```

```
95% family-wise confidence level
```

```
Fit: aov(formula = calcium ~ storageTime, data = Calcium)
```

```
$'storageTime'
```

```

diff          lwr          upr          p adj
1 month-0 months  0.4583333 -1.61429836 2.530965 0.9248149
2 months-0 months 2.0600000 -0.01263169 4.132632 0.0517708
4 months-0 months 2.8450000  0.77236831 4.917632 0.0051726
2 months-1 month  1.6016667 -0.47096503 3.674298 0.1679537
4 months-1 month  2.3866667  0.31403497 4.459298 0.0204414
4 months-2 months 0.7850000 -1.28763169 2.857632 0.7169106
```

(a) (2pts) Calculate the missing values labeled (A), (B), (C) and (D) in the ANOVA table.

A = 3

B = 10.71

C = 6.51

D = 20

(b) (2 pts) Based on the R output, what is the P-value of the F test? State your conclusion of the test in the context of the research. Use $\alpha=0.01$

The conclusion is that p-value is less than $\alpha=0.01$

So the null hypothesis is rejected. There is sufficient evidence to indicate that the mean calcium content is not the same for the four different storage times. The result is statistically significant.

(c) (1 pt) In making your decision in the test, you may have made an error. Is it a type I or a type II error?

Type I error, it occurs when we reject a true null hypothesis

(d) (2 pts) Identify the pair of means that are significantly different at $\alpha=0.05$

Q7 (8 points)

You might think that increasing the nutrient resources available would elevate the number of plant species that an area could support, but the evidence suggests otherwise. The data in the accompanying table are from the Park Grass Experiment at Rothamsted Experimental Station in the U.K., where grassland field plots have been fertilized annually for the past 150 years (collated by Harpole and Tilman 2007). The number of plant species recorded in 10 plots is given in response to the number of different nutrient types added in the fertilizer treatment (nutrient types include nitrogen, phosphorus, potassium, and so on).

	Number of	Number of
PLot	Nutrients added	plant species
1	0	36
2	0	36
3	0	32
4	1	34
5	2	33
6	3	30
7	1	20
8	3	23
9	4	21
10	4	16

```
> NumberNutrients<-c(0,0,0,1,2,3,1,3,4,4)
> NumberSpecies<-c(36,36,32,34,33,30,20,23,21,16)
> summary(lm(NumberSpecies~NumberNutrients))
```

Call:

```
lm(formula = NumberSpecies ~ NumberNutrients)
```

Residuals:

```
Min      1Q  Median      3Q      Max
-10.771  -1.856   1.068   2.894   5.907
```

Coefficients:

```
Estimate Std. Error t value Pr(>|t|)
(Intercept)      34.110      2.599   13.12 1.08e-06 ***
NumberNutrients  -3.339      1.098   -3.04  0.0161 *
```

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

Residual standard error: 5.336 on 8 degrees of freedom

Multiple R-squared: 0.536, Adjusted R-squared: 0.478

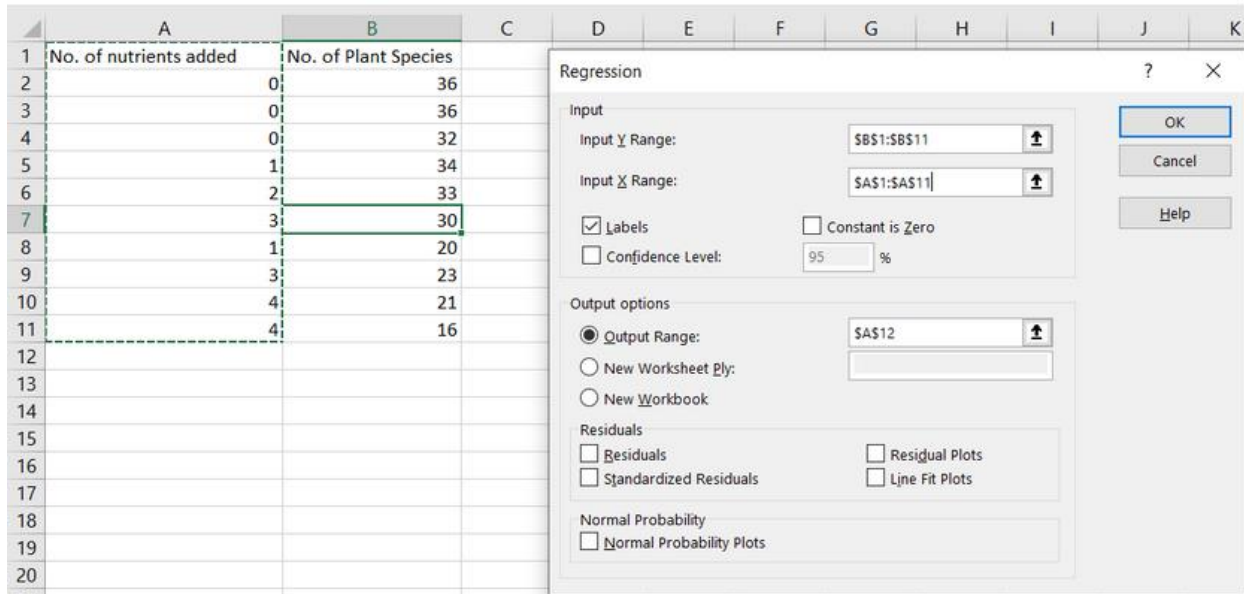
F-statistic: 9.241 on 1 and 8 DF, p-value: 0.01607

(a) (2 pts) Calculate a 90% confidence interval for the slope of the population regression.

Regression line is $Y = a + bX$

We need to find the slope and its standard error.

We need to do regression. For performing regression in excel, we click on to Data-> Data analysis -> Regression



After clicking OK, we get the regression output as shown below

12	SUMMARY OUTPUT								
13									
14	Regression Statistics								
15	Multiple R	0.732105551							
16	R Square	0.535978538							
17	Adjusted R Square	0.477975855							
18	Standard Error	5.336058185							
19	Observations	10							
20									
21	ANOVA								
22		df	SS	MS	F	Significance F			
23	Regression	1	263.1118644	263.1119	9.240582	0.016067			
24	Residual	8	227.7881356	28.47352					
25	Total	9	490.9						
26									
27		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
28	Intercept	34.11016949	2.599312938	13.12276	1.08E-06	28.11614	40.1042	28.11614	40.1042
29	No. of nutrients added	-3.338983051	1.098410195	-3.03983	0.016067	-5.87192	-0.80604	-5.87192	-0.80604

So, the rate of change in the number of plant species supported by per nutrient type addition is, the value of the slope, -3.339

(b) (3 pts) Test the null hypothesis of no treatment effect on the number of plant species. Use $\alpha=0.05$. Write down the pair of hypotheses, specify the P-value and state your conclusion of the test.

Null Hypothesis $H_0: \beta_1 = 0$

Alternative Hypothesis $H_1: \beta_1 \neq 0$

Since the p-value (0.0161) is less than the significance level (0.05), we can reject the null hypothesis.

Therefore, we can conclude that there is a treatment effect on the number of plant species (c) (1 pt)
What fraction of the variation in the number of plant species is "explained" by the number of nutrients added?

(d) (2 pts) What is the correlation between the two variables?

$$R^2 = 0.536$$

$$R = \sqrt{0.536} = 0.732$$

Correlation is 0.732