1. (5 points) Smoke or Coffee?

Two research reports are published about a disease risk of living habits. The first report shows that the smoker proportion is higher in the diseased persons than in the health persons, with a p-value of 0.011. The second report shows that the coffee drinker proportion is higher in the diseased persons than in the health persons, with a p-value of 0.021 The first study uses a sample of 500 diseased persons and 300 healthy persons. The second study uses a sample of 1324 diseased persons and 2850 healthy persons.

Based on these two reports, which factor do you believe that caused the disease: smoking or drinking coffee?

Solution: The P-value technique is used mostly to identify the element that caused the sickness, such as smoking or coffee use.

From the information provided, it can be deduced that the proportion of smokers is greater in the sick than in the healthy, with a p-value of 0.011; similarly, the proportion of coffee drinkers is higher in the sick than in the healthy, with a p-value of 0.021.

The hypotheses for the first report can be given as follows:

H_0: There is no significance evidence that the proportion of smokers is higher in sick people than in healthy people.

H_a: There is significance evidence that the proportion of smokers is higher among sick people than in healthy people.

Decision about the null hypothesis

The null hypothesis should be rejected if the p-value is less than the level of significance; otherwise, fail to reject the null hypothesis.

The p-value in this case is lower than the level of significance.

This equal $0.011 < 0.05$.

Thus, the null hypothesis ought to be rejected.

So, at the 5% level of significance, there is enough data to draw the conclusion that the proportion of smokers is larger in those with illness than in those in good health.

The hypotheses for the second report can be given as follows:

H_0: There is no significant evidence that the proportion of coffee drinkers is higher in sick people than in healthy people.

H_a: There is significant evidence that the proportion of coffee drinkers is higher among sick people than in healthy people.

Decision about the null hypothesis

The null hypothesis should be rejected if the p-value is less than the level of significance; otherwise, fail to reject the null hypothesis.

The p-value in this case is lower than the level of significance.

This equals $0.021 < 0.05$.

Thus, the null hypothesis ought to be rejected.

So, at a 5% level of significance, there is enough data to draw the conclusion that the proportion of coffee drinkers is larger in sick people than in healthy people.

In this case, the first report's P-value is lower than the second report's P-value.

This equals $0.011 < 0.021$.

Hence the sickness was a result of smoking.

2. (5 points)

If there are only two groups, would the results of a one-way ANOVA analysis be the same as those of a two-sample t-test?

Solution: No, even if there are only two groups, the outcomes of a one-way ANOVA analysis and a two-sample t-test might not be the same.
If there are significant differences between the means of three or more groups, a one-way ANOVA is used to analyze the data; if there are significant differences between the means of two groups, a two-sample t-test is used to analyze the data.
The outcomes of a one-way ANOVA and a two-sample t-test, however, can be equivalent when there are two groups. This is so because the t-test in a two-sample t-test and the F-test in an ANOVA both rely on the ratio of variation within groups to variance between groups.
Hence, the F-value and t-value will be the same, and the outcomes of both tests will be the same, if the variances between the two groups are equal. The F-value and t-value, as well as the outcomes of the two tests, will differ if the variances in the two groups are, however, different.
So, even though a two-sample t-test is adequate for comparing the means of two groups, situations where there are more than two groups or when the two-sample t-test's assumption of equal variances is not met may still call for the implementation of a one-way ANOVA.

3. (15 points) Do exercise12.4.8.

One of the goals of the Edinburgh Artery Study was to investigate the risk factors for peripheral arterial disease among persons 55 to 74 years of age. You wish to compare mean LDL cholesterol levels, measured in mmol/liter, among four different populations of subjects: patients with intermittent claudication or interruptions in movement, those with major asymptomatic disease, those with minor asymptomatic disease, and those with no evidence of disease at all. Samples are selected from each population; summary statistics are provided below [218].

	n	$\bar{x}$	s

Intermittent claudication	73	6.22	1.62
Major asymptomatic disease	105	5.81	1.43
Minor asymptomatic disease	240	5.77	1.24
No disease	1080	5.47	1.31

(a) At the 0.05 level of significance, test the null hypothesis that the mean LDL cholesterol levels are the same for the four populations. What are the degrees of freedom associated with this test?

(b) What do you conclude?

(c) What assumptions about the data must be true in order to use the one-way analysis of variance technique?

(d) Before carrying out the study, we expected that the cholesterol levels for people with no diseases (4th group) should be lower than people with disease. Therefore, to check if this is true, we conduct a hypothesis test for $H_0: \mu_1 + \mu_2 + \mu_3 - 3\mu_4 = 0$ by t-test. Does the conclusion changes with Bonferroni correction? Does it change with Schéffe's method? What procedure should be used here (Bonferroni, Schéffe, or no correction at all)?

Add a part (e)

(e) It is clear from the data that the group with intermittent claudication has much higher LDL cholesterol levels than other groups. Therefore, we want to test $H_0: \mu_1 - (\mu_2 + \mu_3 + \mu_4)/3 = 0$ by t-test. Carry out the test. Is any multiple testing adjustment procedure needed? If so, which one? If not, why not?

Solution:

(a) $S_W^2 = \frac{(73-1)1.62^2+(105-1)1.43^2+(240-1)1.24^2+(1080-1)1.31^2}{73+105+240+1080-4}$

$$S_w^2 \approx 1.75$$

$$\overline{x} = \frac{(73)(6.22) + (105)(5.81) + (240)(5.77) + (1080)(5.47)}{73 + 105 + 240 + 1080}$$

$$\overline{x} \approx 5.58$$

$$S_B^2 = \frac{73(6.22 - 5.58)^2 + 105(5.81 - 5.58)^2 + 240(5.77 - 5.58)^2 + 1080(5.47 - 5.58)^2}{4 - 1}$$

$$S_B^2 \approx 19.06$$

$$F = \frac{S_B^2}{S_w^2} = \frac{19.06}{1.75} = 10.89$$

$K - 1 = 4 - 1 = 3$
$n - k = 1498 - 4 = 1494$

(b) $F_{3,\,1494} \approx 2 \cdot 61\ ,\ p = 0.05$

$F_{3,\,1494} \approx 5 \cdot 43\ ,\ p = 0.001$

$=>\ p < 0.001$

Therefore, reject H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$
=> Mean changes in LDL cholesterol are different among the four different populations

(c) Assumptions:
1) Each sample is an independent random sample.
2) Samples are drawn from normal populations.
3) Variances of populations are equal.

(d) H_0: $\mu_1 + \mu_2 + \mu_3 - 3\mu_4 = 0$

$\vec{L} = \overrightarrow{x_1} + \overrightarrow{x_2} + \overrightarrow{x_3} - 3\overrightarrow{x_4} = 1.39$
$var(\vec{L}) = 1.754 \cdot \left(\frac{1}{73} + \frac{1}{105} + \frac{1}{240} + \frac{(-3)^2}{1080}\right)$
$var(\vec{L}) \approx 0.0627$
$t = \left|\frac{\vec{L}}{\sqrt{var(\vec{L})}}\right| = \left|\frac{1.39}{\sqrt{0.8627}}\right| = 5.55$
$t_{1494,\,0.0975} = 1.96 < 5.55$

Reject H_0, cholesterol levels for people with no disease(s) is different from the cholesterol levels of people with disease(s).
Bonferroni:
The conclusion is the same Bonferroni because the alpha level is the same.

Schette's Method:

$$T_{abs} = 5.55 > \sqrt{(k-1)F_{k-2,\,n-k,\,\alpha}}$$
$$T_{obs} = \sqrt{3 \cdot 2.61} = 2.798$$
$$T_{obs} > 2.388$$

Reject H_0. There is no correlation required therefore it is the same conclusion between Bonferroni and Schette.

(e) H_0: $\mu_1 - \left(\frac{\mu_2+\mu_3+\mu_4}{3}\right) = 0$

$\vec{L} = \overrightarrow{x_1} - \left(\frac{\overrightarrow{x_2} + \overrightarrow{x_3} + \overrightarrow{x_4}}{3}\right) = 0.5367$

$var(\vec{L}) = 0.0268$

$t_{obs} = \left|\frac{0.5367}{\sqrt{0.0268}}\right| = 3.27$

Because 3.27 > 1.96, we reject H_0.

Schette's method:

$$t_{obs} = 3.27 > 2.798$$

Again, we reject H_0. The group with intermittent claudication has a higher LDL cholesterol.

4. (10 points)

For the data set airquality contained in the R base package (there is no need to input it, see the example in Lab1), we want to know for which pairs of months the mean Ozone measurements are different. (You can use "?airquality" to check the description of the data set.)

(a) Do the multiple pairwise comparison test at 0.05 level, using Bonferroni, Tukey, and fdr adjustments respectively. What are your conclusions?

(b) Are the results for those three correction methods the same? If not, which you think is the more appropriate conclusion here? Why?

Solution:

(a) Bonferroni R Code:

```
> attach(airquality)
> airquality[is.na(airquality)] = 0
>
>
> pairwise.t.test(Ozone, Month, p.adjust.method = "bonf")

   Pairwise comparisons using t tests with pooled SD

data:  Ozone and Month

  5       6       7       8
6 1.00000 -       -       -
7 0.00029 0.10225 -       -
8 0.00019 0.08312 1.00000 -
9 1.00000 1.00000 0.00697 0.00485

P value adjustment method: bonferroni
> pairwise.t.test(Ozone, Month, p.adjust.method = "fdr")
```

FDR R Code:

```
  Pairwise comparisons using t tests with pooled SD

data:  Ozone and Month

  5       6       7       8
6 0.76096 -       -       -
7 0.00015 0.01704 -       -
8 0.00015 0.01662 0.91744 -
9 0.46493 0.91744 0.00174 0.00162

P value adjustment method: fdr
>
```

Tukey R Code:

```
> airquality$Month<-as.factor(airquality$Month)
> p<-aov(Ozone~Month,data=airquality)
> TukeyHSD(p,conf.level = 0.95)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Ozone ~ Month, data = airquality)

$Month
           diff          lwr       upr     p adj
6-5 -10.9731183 -32.27095900 10.324722 0.6139469
7-5  29.7741935   8.65164668 50.896740 0.0013894
8-5  30.4838710   9.36132410 51.606418 0.0009868
9-5  10.5935484 -10.70429233 31.891389 0.6454439
7-6  40.7473118  19.44947111 62.045153 0.0000044
8-6  41.4569892  20.15914853 62.754830 0.0000029
9-6  21.5666667   0.09496314 43.038370 0.0484120
8-7   0.7096774 -20.41286945 21.832224 0.9999830
9-7 -19.1806452 -40.47848588  2.117196 0.0990957
9-8 -19.8903226 -41.18816330  1.407518 0.0795001
```

The following mean ozone levels are different
Bonferroni: May-July, May-August, July-September, July-September, August-September
FDR: May-July, May-August, July-September, August-September, June-July, June-August
Tukey: May-July, May-August, July-September, August-September

(b) Bonferroni and Tukey yield the same results. FDR yields different results. In this case, Tukey is more appropriate that Bonferroni as the sample size is larger.

5. (5 points)

For the data set *lowbwt* used before, systolic blood pressure measurements for 100 low birth weight infants are saved under variable sbp, gender is saved under variable sex and Preeclampsia (formerly called toxemia, a complication of pregnancy for the child's mother) is saved under the variable "preeclampsia". (For the lowbwt.csv from the website of the 3rd edition of the textbook, the variable is coded as "preeclampsia". If you use the data file from previous edition, that variable is coded as "tox".)

We wish to see whether mean systolic blood pressure is the same for low-birth-weight boys and girls.

(a) Use R to produce the ANOVA tables: one with preeclampsia status as the blocking variable and one without any blocking.

(b) What is the p-value for gender effect on systolic blood pressure with the blocking? What is the p-value for gender effect on systolic blood pressure without the blocking?

Solution:

a) R Code:

```
> data <- read.csv(file="lowbwt.csv", header = TRUE, sep = ",")
> data$sex<-as.factor(data$sex)
> data$tox<-as.factor(data$tox)
> summary(aov(sbp~sex+tox, data=data))
            Df Sum Sq Mean Sq F value Pr(>F)
sex          1     48   48.25   0.367  0.546
tox          1     67   66.76   0.508  0.478
Residuals   97  12758  131.53
> summary(aov(sbp~sex,data=data))
            Df Sum Sq Mean Sq F value Pr(>F)
sex          1     48   48.25   0.369  0.545
Residuals   98  12825  130.87
>
```

b) p-value for gender with blocking: 0.546
p-value for gender without blocking: 0.545

6. (10 points) Mini-Project: measuring your response time.

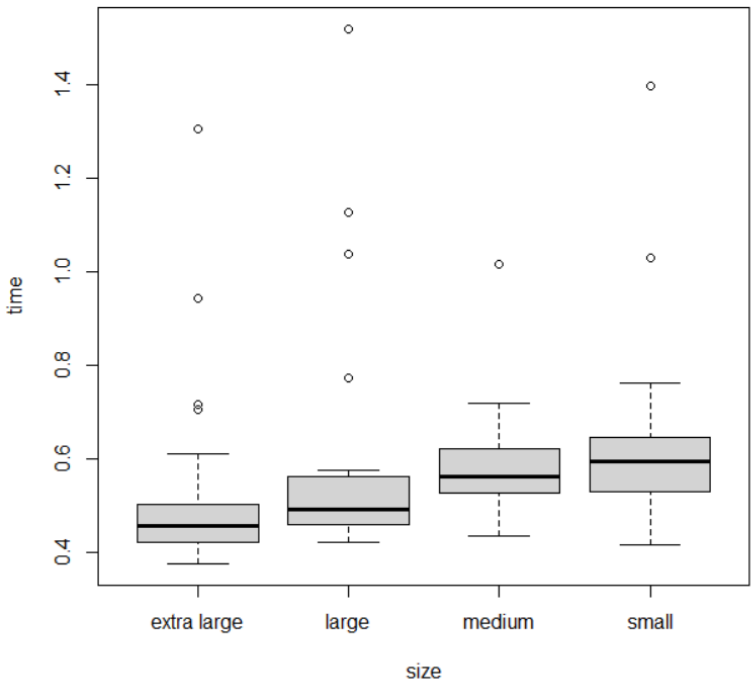
Using the webpage in Lab 1 to collect 30 response times each with all four sizes of boxes: "small", "medium", "large" and "xlarge".

(a) Conduct an ANOVA to see if there is a difference between your response times for different box sizes.

(b) Conduct the pairwise comparisons between groups using HSD and LSD.

Solution:

```
library(agricolae)
rt <- read.csv("response_times.csv", header = TRUE, sep = ",")
boxplot(time~size, data = rt)
```



(a)

```
> res.aov <- aov(time ~ size, data = rt)
> summary(res.aov)
             Df Sum Sq Mean Sq F value Pr(>F)
size          3  0.207 0.06888   1.952  0.125
Residuals   116  4.094 0.03529
```

(b)

```
> TukeyHSD(res.aov, confidence.level=0.95)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = time ~ size, data = rt)

$size
                          diff          lwr       upr     p adj
large-extra large  0.06020000 -0.066232839 0.1866328 0.6020936
medium-extra large 0.06580000 -0.060632839 0.1922328 0.5290668
small-extra large  0.11706667 -0.009366172 0.2434995 0.0801337
medium-large       0.00560000 -0.120832839 0.1320328 0.9994465
small-large        0.05686667 -0.069566172 0.1832995 0.6454087
small-medium       0.05126667 -0.075166172 0.1776995 0.7162516

>
> pairwise.t.test(rt$time, g=rt$size, p.adjust.method = 'none')

        Pairwise comparisons using t tests with pooled SD

data:  rt$time and rt$size

       extra large large medium
large  0.217       -     -
medium 0.178       0.908 -
small  0.017       0.243 0.293

P value adjustment method: none
```

We fail to reject null hypothesis in both cases.