Practice Questions
Set 1

Initialize R by entering the following commands at the prompt. Type the commands exactly as shown.

```r
options(contrasts=c("contr.sum","contr.poly") )
load(file=url("http://psycserv.mcmaster.ca/bennett/psy710/p1/iq-data.Rdata") )
aov.data <- read.csv(file=url("http://psycserv.mcmaster.ca/bennett/psy710/p1/q1Data.csv") )
```

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group</td>
<td>3</td>
<td>578</td>
<td>192.68</td>
<td>2.524</td>
</tr>
<tr>
<td>Residuals</td>
<td>56</td>
<td>4276</td>
<td>76.35</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Use this ANOVA table to answer question 1.

1. Table 1 shows the results of an ANOVA performed on an experiment that examined differences among groups of subjects who were assigned randomly to several treatments. Use Table 1 to answer all of the following parts of this question.

(a) A one-way ANOVA can be conceptualized as a comparison of a pair of nested models, and parts of the ANOVA table can be thought of as showing the results of that comparison. In this framework, what do the values of 578 and 4276 represent?

**Answer:** The value of of 4276 represents the sum of the squared residuals of the full model which includes three free parameters (i.e., an intercept and 2 α’s) and 578 is the difference between the goodness of fit (i.e., the sum of squared residuals) in the full and reduced models. More specifically, it represents the increase in the sum of squared residuals when we compare the full model (in which the treatment effects (αs) are varied to yield the best least-squares fit to the data) to the reduced model (in which the treatment effects are set to zero).

(b) The interpretation of the p-value for our F test requires us to make some assumptions about the data. Briefly describe these assumptions.

**Answer:** We will discuss this answer in class.

(c) An underlying assumption of ANOVA is that the observed dependent variable (yij) for subject i in group j is the sum of a constant (μ), a group effect αj, and error (e): yij = μ + αj + eij. Based on the information in Table 1, what is the value of σ2?

**Answer:** MSResiduals, 76.35, is an estimate of the population error variance. [N.B. The value for MSGroup is another estimate of population error but only if you are willing to assume that the treatment effects are all zero.]

(d) What hypothesis is being evaluated by the F test in the first row of the table?

**Answer:** We will discuss this answer in class.
(e) What does the $p$ value in the table mean?

**Answer:** We will discuss this answer in class.

2. Verbal IQ scores were obtained from 20 children. The data are stored in the variable `iq.dat`, which was placed into R with the `load` command above. Use the data in `iq.dat` to answer the following questions.

(a) Use a $t$ test to evaluate the null hypothesis that the data were sampled from a population with a mean of 100. Make sure answer includes the values of $t$ value, $p$, and the degrees of freedom, and a clear description of the null and alternative hypotheses being evaluated.

```r
 t.test(iq.dat, mu=100)
```

```
##
## One Sample t-test
##
## data: iq.dat
## t = -0.7272, df = 29, p-value = 0.4729
## alternative hypothesis: true mean is not equal to 100
## 95 percent confidence interval:
## 92.76059 103.44165
## sample estimates:
## mean of x
## 98.10112
```

**Answer:** My null hypothesis is that $\mu = 100$ and my alternative hypothesis is that $\mu \neq 100$. The $t$ test was not significant ($t = -0.727$, df = 29, $p = 0.47$). Therefore I choose to not reject the null hypothesis. (N.B. In real life I would examine the data for outliers, deviations from normality, etc.).

(b) What does the $p$ value from the previous $t$ test mean?

**Answer:** It is the probability of obtaining a $t$ value at least as large as the observed value ($t < -0.727$ or $t > 0.727$) when the null hypothesis is true.

(c) Suppose that you were interested in determining if the data were sampled from a population with a mean that was equivalent to 100, where equivalent is defined as being within 100 ± 5. Conduct two one-tailed $t$ tests to determine if it is reasonable to conclude that $95 < \mu < 105$. Make sure your answer includes the values of $t$, $p$, and the degrees of freedom, and a clear description of the null and alternative hypotheses being evaluated.

```r
 t.test(iq.dat, mu=105, alternative="less")
```

```
##
## One Sample t-test
##
## data: iq.dat
## t = -2.642, df = 29, p-value = 0.006572
## alternative hypothesis: true mean is less than 105
## 95 percent confidence interval:
## -Inf 102.5379
## sample estimates:
## mean of x
## 98.10112
```
Answer: My null hypothesis is that $\mu \geq 105$ and my alternative hypothesis is that $\mu < 105$. The $t$ test was significant ($t = -2.642$, df = 29, $p = 0.0066$). Therefore I reject the null hypothesis in favour of the alternative ($\mu \leq 105$).

```r
t.test(iq.dat,mu=95,alternative="greater")
##
## One Sample t-test
##
## data: iq.dat
## t = -2.642, df = 29, p-value = 0.0066
## alternative hypothesis: true mean is greater than 95
## 95 percent confidence interval:
## 93.66434 Inf
## sample estimates:
## mean of x
## 98.10112
```

Answer: My null hypothesis is that $\mu \leq 95$ and my alternative hypothesis is that $\mu > 95$. The $t$ test was not significant ($t = 1.187$, df = 29, $p = 0.122$). Therefore I do not reject the null hypothesis ($\mu \leq 95$).

Answer: In order to accept the hypothesis that $95 < \mu < 105$, both of my one-tailed $t$ tests would have to be significant. In other words, I would have to accept the hypothesis that $\mu < 105$ (from the first $t$ test) and $95 < \mu$ (from the second $t$ test). The second $t$ test was not statistically significant, and therefore I do not have sufficient evidence to conclude that the observed mean is within the range $100 \pm 5$.

3. An experiment was conducted to assess the effects of four treatments on a dependent variable, $y$. The experiment measured $y$ on 32 subjects assigned randomly to the four treatments ($n = 8$ per treatment), and the data are stored in the data frame `aov.data`. Use `aov.data` to answer all of the following questions.

(a) Calculate the mean and standard deviation of $y$ for each treatment group.

```r
names(aov.data)
## [1] "y"  "treatment"
with(aov.data,tapply(y,treatment,mean))
## t1  t2  t3  t4
## 86.375 101.375 93.250 100.500
with(aov.data,tapply(y,treatment,sd))
## t1  t2  t3  t4
## 9.840695 8.895223 11.744300 7.559289
```

(b) Conduct an analysis of variance to evaluate the effect of `treatment` on $y$. Record the results of your ANOVA (i.e., write the ANOVA table).

```r
q1.lm.01 <- lm(y~treatment,data=aov.data)
anova(q1.lm.01)
```
(c) Our ANOVA makes assumptions about the distribution of the errors: specifically, that the errors are distributed normally and have constant variance across groups. Use statistical tests to evaluate these assumptions.

```r
ggplot(data = q1.lm.01, aes(x = treatment, y = y)) + geom_boxplot()
```

```r
shapiro.test(residuals(q1.lm.01)) # evaluate normality assumption
```

```r
## Shapiro-Wilk normality test
## data: residuals(q1.lm.01)
## W = 0.97432, p-value = 0.626
```

```r
bartlett.test(y~treatment,data=aov.data) # evaluate equal variance assumption
```

```r
## Bartlett test of homogeneity of variances
## data: y by treatment
## Bartlett's K-squared = 1.3454, df = 3, p-value = 0.7184
```

(d) Explain what is meant by Type I and Type II errors and how they relate to your analysis.

**Answer:** We will discuss this answer in class.

(e) What does Cohen’s $f$ represent? Calculate Cohen’s $f$ for **treatment**.

```r
q1.lm.01 <- lm(y~treatment,data=aov.data)
adj.r.squared <- summary(q1.lm.01)$adj.r.squared

( cohens.f <- sqrt( adj.r.squared / (1-adj.r.squared)) )
```

```r
## [1] 0.5606852
```

**Answer:** Cohen’s $f$ is calculated from adjusted-R-squared, which can be found in the summary table for an `lm` object. For our data, Cohen’s $f$ is 0.5606. We will discuss what Cohen’s $f$ represents in class.

(f) Evaluate a linear comparison between the mean of treatment 1 and the mean of the other three treatments.

**Answer 1:** You could answer this question using several methods. For example, you could use the `linear.comparison` command:

```r
# method 1:
source(url("http://psycserv.mcmaster.ca/bennett/psy710/Rscripts/linear_contrast_v2.R"))
```
Answer 2: Or you could use R’s `aov` or `lm` commands. First, you would need to associate the contrast weights with the factor, `g`:

```r
# method 2:
contrasts(g) <- w
```

Then you would use `aov` to perform the ANOVA, list the ANOVA table, and use the split command to break the sum-of-squares for `g` into separate pieces. The contrast is significant ($F(1, 28) = 9.314, p = 0.0049$).

```r
q1.aov.01 <- aov(aov.data$y ~ g)
summary(q1.aov.01, split=list(g=list(myContrast=1, others=2:3)))
```

Instead of listing the ANOVA table, you could list the regression table for an `lm` object. The first line reports the results of a two-tailed $t$ test on our linear contrast. The contrast is significant ($t(28) = -3.052, p = 0.0049$).

```r
q1.lm.01 <- lm(aov.data$y ~ g)
summary(q1.lm.01)
```
Answer 3: Or you could use the `lsmeans` and `contrast` command in the `lsmeans` package:

```r
# method 3:
# install.packages("lsmeans")
library(lsmeans)
ls.q1 <- lsmeans(q1.lm.01,"g")
contrast(ls.q1,list(G1vsG234=w))
```

```
## contrast estimate SE df t.ratio p.value
## G1vsG234 -36 11.8 28 -3.052 0.0049
```

Answer 4: Finally, you could calculate your contrast directly from the group means:

```r
# method 3:
m <- with(aov.data,tapply(y,treatment,mean)) # group means
n <- with(aov.data,tapply(y,treatment,length)) # group n
MS.w <- 92.76 # from ANOVA table
psi <- sum(w * m) # value of PSI
# from Eq 6 in chapter 4 notes:
SS.contrast <- (psi^2) / sum( (w^2)/n)
(F <- SS.contrast / MS.w )
```

```
## [1] 9.31436
```

```r
N <- sum(n) # total N
a <- length(m) # number of groups
df1 <- 1 # numerator df
 ( df2 <- N-a ) # denominator df
```

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

```
(p.value <- 1 - pf(F,df1,df2) ) # our p value [remember to use 1 - pf ]
```

```
## [1] 0.004937986
```

```
(p.value <- pf(F,df1,df2,lower.tail=FALSE) ) # [or set lower.tail to FALSE]
```

```
## [1] 0.004937986
```

(g) List the weights for two new contrasts that are orthogonal to each other and to the contrast performed in the previous question. State the null hypothesis that is tested by each new set of weights.

```r
w1 <- c(3, -1, -1, -1) # original contrast
w2 <- c(0, 2, -1, -1) # HO: mean of treatment 2 = mean of treatments 3 and 4
w3 <- c(0, 0, 1, -1) # HO: mean of treatment 3 = mean of treatment 4
sum(w1*w2)
```

```
## [1] 0
```
\texttt{sum(w1*w3)}

## [1] 0

\texttt{sum(w2*w3)}

## [1] 0