STA305/1004 - L0101 Term Test 2 SOLUTIONS

March 14, 2018, 11:10-12:40

Name:

Student Number:

Instructions: Answer all four questions in the space provided within 90 minutes. Work written on the back of pages will not be graded.

Aids allowed: One 8.5'x11' sheet with writing on both sides, and a non-programmable calculator.

The table below shows the value of each question.

Question	Marks
1	30
2	20
3	25
4	25
Total	100

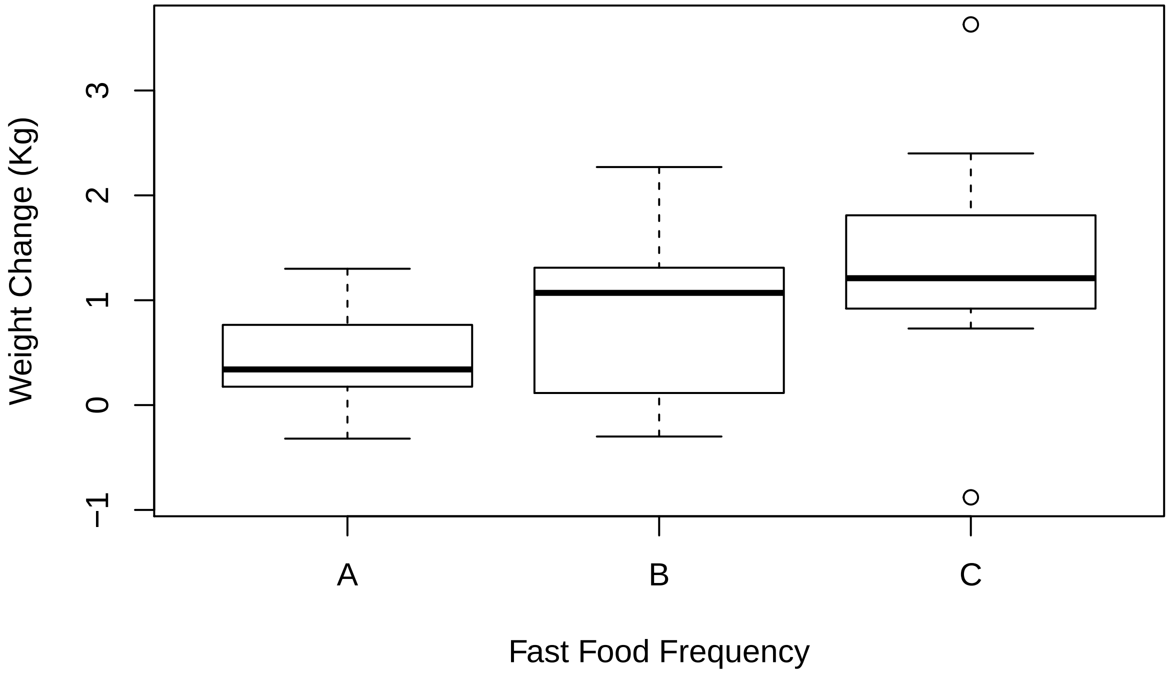
This test has 15 pages including this page.

1. (30 marks) A study at UofT recruited twenty-one students to complete a thirty minute survey on their diet and eating habits at the end of an academic year. Students were paid $10 to complete the survey and answer a few questions. The data below shows their weight gain from September to April classified by the frequency that students ate fast food. In group A students reported eating fast food once per month; the students in group B reported eating fast food twice per month; and the students in group C reported eating fast food four times per month.

	A	B	C
	1.02	1.44	0.73
	-0.32	0.40	1.11
	0.27	-0.30	3.63
	0.08	2.27	-0.88
	0.51	-0.17	1.21
	0.34	1.07	1.22
	1.30	1.18	2.40
Treatment Average	0.46	0.84	1.35
Treatment SD	0.55	0.92	1.40

The researchers analyzed the data using R.

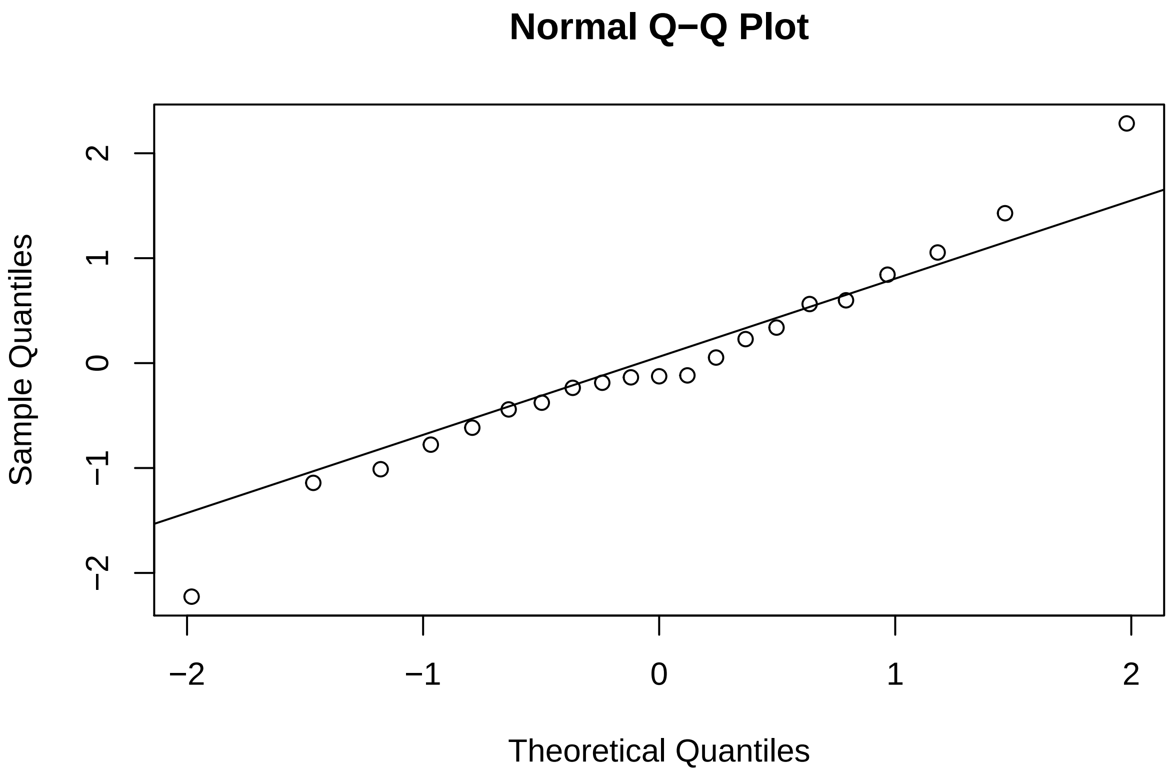
```
boxplot(wtchange~grp,data = surveydat,ylab="Weight Change (Kg)", xlab="Fast Food Frequency")
```



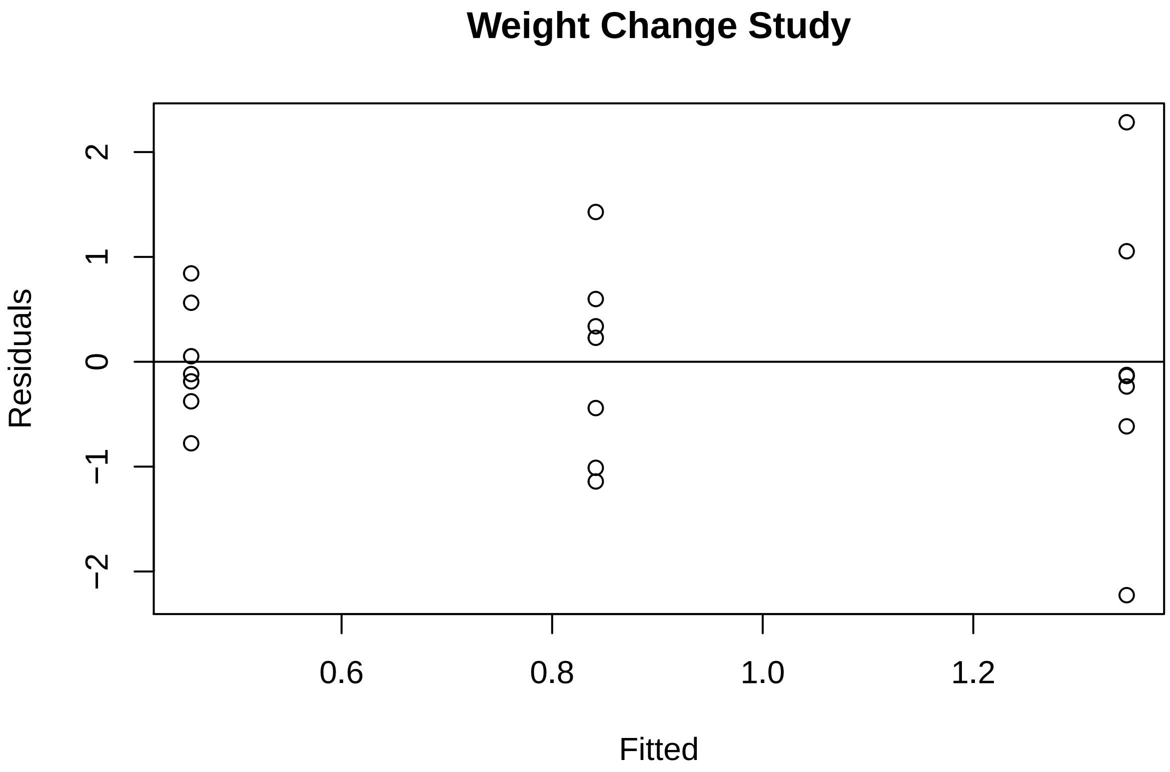
```
aovsurvey <- aov(wtchange~grp,data=surveydat)
summary(aovsurvey)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## grp          2   2.78   1.390   1.341  0.287
## Residuals   18  18.66   1.037
```

```
qqnorm(aovsurvey$residuals);qqline(aovsurvey$residuals)
```



```
plot(aovsurvey$fitted.values, aovsurvey$residuals,ylab="Residuals",
     xlab="Fitted",main="Weight Change Study")
abline(h=0)
```



(a) (5 marks) Is this study an experiment or observational study? What are the treatments? Briefly explain.

This is an observational study since the treatment assignment mechanism is unknown. The treatment has three levels: eating fast food once per month; eating fast food twice per month; and eating fast food four times per month.

(b) (5 marks) Would it have been feasible for the researcher to randomized students to the treatments? What randomization scheme (assigning the subjects to the treatments) could the researcher use to accomplish the randomization?

A randomized study is not feasible since the treatment is the frequency of eating fast food. Nevertheless, a hypothetical randomization scheme could be constructed by labeling all the students using the numbers 1 to 21 then obtaining a random permutation of these numbers. Assign the first 7 students to diet A, the next 7 to diet B, etc.

(c) (5 marks) What are the null and alternative hypotheses that the researchers are testing in the data analysis? Is there evidence to reject the null hypothesis? Explain.

$$H_0 : \mu_1 = \mu_2 = \mu_3 \text{ vs. } H_1 : \mu_i \neq \mu_j, i \neq j.$$

There is no evidence to reject H_0 since the p-value of 0.287 is large.

(d) (10 marks) What are the statistical assumptions behind the data analysis? Which tools can be used to check the assumptions? Are the assumptions satisfied? Explain.

The three assumptions are outlined below.

i. Additive model: $y_{ij} = \mu + \tau_i + \epsilon_{ij}$.

   This seems plausible since change in weight from each diet can be viewed as the sum of a common mean plus a random error term.

ii. The errors ϵ_{ij} are independent and identically distributed (iid) with common variance $Var(\epsilon_{ij}) = \sigma^2$, for all i, j

   - A plot of the residuals versus fitted values can be used to investigate the assumption that the residuals are randomly distributed and have constant variance.
   - In this case the points don't fall randomly on both sides of 0. The residuals are increasing as the fitted values increase. This is an indication that the common variance assumption is not satisfied. This can also be seen from the standard deviations in each treatment group: the largest (1.4) is approximately three times as large as the smallest (0.55).
   - We are not given any information to confirm that that observations are independent. For example, if some of the students in the sample roommates then their weight gains and fast food consumption may not be independent.

iii. $\epsilon_{ij} \sim N(0, \sigma^2)$.

   The normal quantile plots indicates that this assumption is satisfied since the points fall along the straight line.

It is not the case that all the assumptions are satisfied since the non-constant variance assumption may not be true. In addition, it's difficult to confirm if the data are independent. Therefore, it might be the case that the p-value is not accurate (e.g., the p-value might actually be smaller.)

(e) (5 marks) The researcher is convinced that the results of the study would have provided strong evidence that eating fast food four times per month causes students to gain weight, if the sample size in each group was larger. Is this a valid statement? Explain.

This is not a valid statement. Consider any one of the following points:

- There is no comparison group where students did not eat fast food so it's not possible to calculate the causal effect of fast food versus no fast food on weight gain. It is possible to calculate the non-causal effect of eating less fast food versus more fast food.
- Subjects assigned the “treatment" to themselves so we don't know if there are differences in the types of students (e.g., age, sex, history of being overweight) that selected themselves to be in the groups.
- If the sample size in each group was larger then the power would increase. Although, even if the study was designed to have high power and the analysis yielded a small p-value then this still wouldn't fix the way treatment was assigned. So, we wouldn't know if the differences are due to the treatment or due to differences between the types of students in the groups.

2. (20 marks) Suppose that the means of 4 new treatments, namely Treatments A, B, C and D, are being compared to that of a standard (Control) treatment. We can conduct several pairwise t tests to do multiple comparisons among the 5 treatments. Assume that the pairwise tests are independent of each other. Answer the following questions.

   (a) (1 mark) What is the maximum number of two-sample t tests that can be done?

   The maximum number of two-sample t tests that can be done among the five treatments is $\binom{5}{2} = 10$.

   (b) (2 marks) State the null and alternative hypotheses of one of the two-sample t- tests.

$$H_0 : \mu_i = \mu_j$$
$$H_a : \mu_i \neq \mu_j$$

   for any $i \neq j$, $i, j = 1, 2, \ldots, 5$

   (c) (3 marks) If the two-sample t tests are conducted at the 0.10 level, and all the null hypotheses are true, what is the distribution of the total number of tests that will be significant? Completely specify the distribution.

   Let Y represent the number f tests that will be significant.
   A suitable model for Y is $Binomial(10, 0.10)$.

   (d) (1 mark) If the two-sample t tests are conducted at the 0.10 level, and all the null hypotheses are true, how many tests do you expect would produce a significant result? Show your work.

   We would expect $10(0.10) = 1$ test to produce a significant result.

(e) (3 marks) If the two-sample t tests are conducted at the 0.10 level, and all the null hypotheses are true, what is the probability that at least one null hypothesis is rejected? Show your work. Answer to two decimal places.

We want

$$\begin{aligned} P(Y \geq 1) &= 1 - P(Y = 0) \\ &= 1 - (1 - 0.10)^{10} \\ &= 1 - 0.9^{10} \\ &= 0.65 \end{aligned}$$

Hence, if the 10 independent two-sample t tests are conducted at the 0.10 level, and all the null hypotheses are true, the probability that at least one null hypothesis is rejected is 0.65.

(f) (5 marks) Under this scenario, compare the pairwise error rate to the experiment-wise error rate. Does it make sense to consider all pairs of treatment means given that there is a control treatment? Explain.

The pairwise error rate (PWER) is 0.10 while the experiment-wise error rate (FWER) is 0.65. The FWER is more than 6 times larger than the PWER. Based on this, it does not make sense to consider all pairs of treatment means. The FWER rate increases with the the number of comparisons made. Since there is a control we can consider comparing the other treatments to it, which would lead to 4 comparisons, rather than doing all 10 comparisons.

(g) (5 marks) Describe one method to control the experiment-wise error rate.

The Bonferroni and the Tukey methods were methods, studied in this course, to control the FWER. Describe ONE.

For either method, we can calculate the test statistic,

$$t_{ij} = \frac{\bar{y}_{j\cdot} - \bar{y}_{i\cdot}}{\hat{\sigma}\sqrt{\frac{1}{n_j} + \frac{1}{n_i}}}$$

to test $H_0 : \mu_i = \mu_j$ vs $H_a : \mu_i \neq \mu_j$, for $i \neq j$ at level α.

Bonferroni says that the pair of means are different from each other if

$$|t_{ij}| > t_{N-k,\alpha/2c}$$

where c is the total number of pairs compared.

Tukey's method says that a pair of means are different from each other if

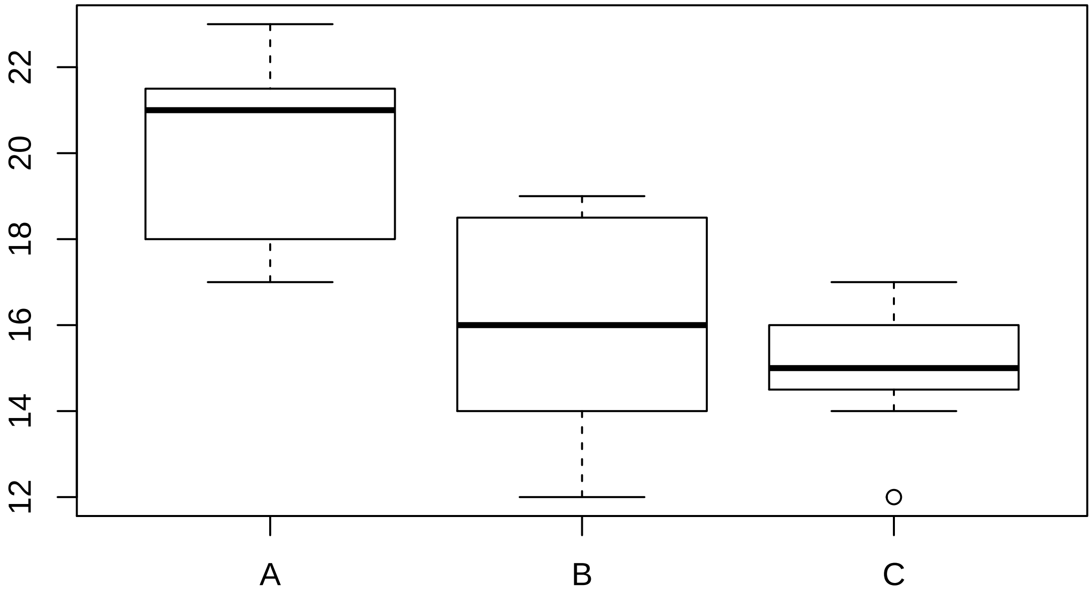
$$|t_{ij}| > \frac{1}{\sqrt{2}} q_{k,N-k,\alpha}$$

where $q_{k,N-k,\alpha}$ is from the critical value from the Studentized Range distribution.

3. (25 marks) (Adapted from Box, Hunter and Hunter) Each of 21 student athletes, randomly grouped into three teams, A, B, and C, attempts to successfully toss a basketball through a hoop within a fixed time period. The number of successes per team is given in the table below. The main objective of this study is to compare the three teams. In this question, use a significance level of 5%.

	Team A	Team B	Team C
	21	13	15
	19	16	16
	17	15	14
	21	12	15
	22	19	16
	23	19	12
	17	18	17
Team Average	20	16	15
Team SD	2.38	2.83	1.63

(a) (5 marks) Consider the side-by-side boxplots of distribution of the number of successes per team. Do the distributions look similar or different? Explain.



The distributions look roughly symmetrical. There are no obvious outliers and values range between 12 and 23 buckets per athlete. The standard deviations similar; they are within a 2:1 ratio of each (for example, 2.83/1.63 = 1.74 < 2).

However, the mean of Team A (20) (and the median that is about 21) is bigger than the of the means (and corresponding medians) of the other two teams. Teams B and C have similar means, that is 16 and 15 respectively.

(b) (5 marks) Using linear regression, an ANOVA table was obtained. However, some values went missing. Complete the ANOVA table by finding (A), (B), (C), (D) and (E).

```
anova(lm(y~as.factor(team),data = bbdata))
```

```
## Analysis of Variance Table
##
## Response: y
##                  Df Sum Sq Mean Sq F value   Pr(>F)
## as.factor(team) (A)   98    (B)      (E)   0.001953 **
## Residuals       (C)   98    (D)
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(A) $3 - 1 = 2$

(B) $\dfrac{98}{2} = 49$

(C) $21 - 3 = 18$

(D) $\dfrac{49}{9} = 5.44$

(E) $\dfrac{49}{49/9} = 9$

(c) (5 marks) What is the appropriate null hypothesis? What do you conclude from the ANOVA table?

The null hypothesis is

$$H_0 : \mu_A = \mu_B = \mu_C$$

where μ represents the true mean score of a team.

From the ANOVA table, there is strong evidence that the teams different in terms of their averages since the p-value of 0.0019 is very small.

(d) (5 marks) Given the R codes and output that follow, define the underlying statistical model in terms of dummy variables. Explicitly state the dummy variables. What do the parameters represent?

```
bbdata$team <- as.factor(bbdata$team)
contrasts(bbdata$team) <- contr.treatment(n = 3,base = 2)
summary(lm(y~team, data = bbdata))
```

```
##
## Call:
## lm(formula = y ~ team, data = bbdata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##     -4     -1      0      2      3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  16.0000     0.8819  18.142 5.15e-13 ***
## team1         4.0000     1.2472   3.207  0.00489 **
## team3        -1.0000     1.2472  -0.802  0.43314
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.333 on 18 degrees of freedom
## Multiple R-squared:    0.5,  Adjusted R-squared:  0.4444
## F-statistic:     9 on 2 and 18 DF,  p-value: 0.001953
```

The underlying model is

$$y_{ij} = \tau_0 + \tau_1 I_{1j} + \tau_2 I_{2j} + \epsilon_{ij}$$

where the dummy variables are

$$I_{1j} = \begin{cases} 1, \text{if jth athlete is in Team A} \\ 0, \text{if jth athlete is not in Team A} \end{cases}$$

$$I_{2j} = \begin{cases} 1, \text{if jth athlete is in Team C} \\ 0, \text{if jth athlete is not in Team C} \end{cases}$$

and the parameters,

τ_0 represents the mean score of Team B,
τ_1 represents the difference in mean scores between Team A and Team B, and
τ_2 represents the difference in mean scores between Team C and Team B.

(e) (5 marks) Given the unadjusted p-values for all pairwise comparisons of the three teams, which pairs of teams have a statistically significant difference? Do you expect that your results will change if you adjust for multiple comparisons using Bonferroni or Tukey method? If so, explicitly describe the changes in terms of the p-values.

```r
pairwise.t.test(bbdata$y,as.factor(bbdata$team),p.adjust.method = "none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  bbdata$y and as.factor(bbdata$team)
##
##   A       B
## B 0.00489 -
## C 0.00082 0.43314
##
## P value adjustment method: none
```

Both teams B and C differ from Team A.

If we use Bonferroni or Turkey's method, the results will not change. However, the p-values will be about three times larger than those of the unadjusted method.

Specifically, under Bonferroni's method, the p-values would be 0.015, 0.0025 and 1 respectively. The p-values under Tukey's method would be more similar to those of Bonferroni's than of the unadjusted procedure.

4. (25 marks) A psychologist is designing an experiment to investigate the effects of four different learning methods on short term memory. Subjects will be shown a series of 20 words after undergoing some training in the learning method that they were assigned. The outcome of the experiment is the total number of words that a subject is able to recall after being trained in one of the learning methods. An equal number of subjects will be randomly assigned to each learning method.

   Based on previous research the psychologist estimates that the mean and standard deviation for each method are:

Learning Method	Mean	Standard deviation
1	15	6
2	14.5	4
3	12.5	3.5
4	15.3	3

   The psychologist would like to know how many subjects she will require so that her study has 80% power at the 5% significance level. She used R to obtain the following output. Answer the questions that follow.

```r
library(pwr)
mu1 <- 15; mu2 <- 14.5;mu3 <- 12.5; mu4 <- 15.3
sigma1 <- 6; sigma2 <- 4; sigma3 <- 3.5;sigma4 <- 3;
mug <- sum(mu1,mu2,mu3,mu4)/4
mui <- c(mu1,mu2,mu3,mu4)
f1 <- sqrt(sum((mui-mug)^2)/4)/sigma1
pwr.anova.test(k = 4,f = f1,n=15,sig.level = 0.05)
```

```
##
##      Balanced one-way analysis of variance power calculation
##
##               k = 4
##               n = 15
##               f = 0.181955
##       sig.level = 0.05
##           power = 0.1805942
##
## NOTE: n is number in each group
```

```r
f2 <- sqrt(sum((mui-mug)^2)/4)/sigma2
pwr.anova.test(k = 4,f = f2,n=15,sig.level = 0.05)
```

```
##
##      Balanced one-way analysis of variance power calculation
##
##               k = 4
##               n = 15
##               f = 0.2729326
##       sig.level = 0.05
##           power = 0.3727171
##
## NOTE: n is number in each group
```

```
f3 <- sqrt(sum((mui-mug)^2)/4)/sigma3
pwr.anova.test(k = 4,f = f3,n=15,sig.level = 0.05)
```

```
##
##      Balanced one-way analysis of variance power calculation
##
##               k = 4
##               n = 15
##               f = 0.3119229
##       sig.level = 0.05
##           power = 0.475779
##
## NOTE: n is number in each group
```

```
f4 <- sqrt(sum((mui-mug)^2)/4)/sigma4
pwr.anova.test(k = 4,f = f4,n=15,sig.level = 0.05)
```

```
##
##      Balanced one-way analysis of variance power calculation
##
##               k = 4
##               n = 15
##               f = 0.3639101
##       sig.level = 0.05
##           power = 0.6170057
##
## NOTE: n is number in each group
```

```
NSIM <- 10000
res <- numeric(NSIM)
mu1 <- 15; mu2 <- 14.5;mu3 <- 12.5; mu4 <- 15.3
sigma1 <- 6; sigma2 <- 4; sigma3 <- 3.5;sigma4 <- 3;
n <- 15
for (i in 1:NSIM){
y1 <- rnorm(n,mu1,sigma1)
y2 <- rnorm(n,mu2,sigma2)
y3 <- rnorm(n,mu3,sigma3)
y4 <- rnorm(n,mu4,sigma4)
y <- c(y1,y2,y3,y4)
trt <- as.factor(c(rep(1,n),rep(2,n),rep(3,n),rep(4,n)))
m <- lm(y~trt)
res[i] <- anova(m)[1,5] # p-value of F test
}
sum(res<=0.05)/NSIM
```

```
## [1] 0.3263
```

(a) (5 marks) The formula for effect size is

$$f = \sqrt{\frac{\sum_{i=1}^{k}(\mu_i - \bar{\mu})^2/k}{\sigma^2}}$$

where $\bar{\mu} = \sum_{i=1}^{k} \mu_i/k$ and σ^2 is the within group error variance. What are the effect sizes that the psychologist can detect if she uses the different variances of the different learning methods? Briefly explain what effect size represents.

The effect sizes are 0.18, 0.27, 0.31 and 0.36.

Effect size represents a ratio of the standard deviation of the between group population means to the standard deviation of the within group population means. Larger values correspond to stronger evidence that between group variation is larger than within group variation.

(b) (5 marks) Assuming that she can enrol 15 subjects per group, what is the power (to 2 decimal places) to detect each effect size at the 5% level? Explain how power changes with effect size.

The power is 0.18, 0.37, 0.48 or 0.62 respectively for each effect size.

Power increases as effect size increases.

(c) (10 marks) Based on the computer simulations, what is the power of the study using 15 subjects per group assuming that the standard deviations for the four methods are not equal, but are as shown in the table above, and that the distribution of observations in each group is normal? Is this power level desirable? Explain what power means in this context.

The power is 0.33. This power is low since it is closer to 0 than 1.

In this context, power is the probability to detect that the learning methods are different when indeed at least 2 methods differ.

(d) (5 marks) What do parts (b) and (c) show about the assumption of a common within group variance in calculating power for an ANOVA experiment? Explain.

When the assumption of equal variance is not satisfied, the ANOVA test would have lower power.