

MATH 204 - ASSIGNMENT 1: SOLUTIONS

1. (a)
 - Experimental units: the adults recruited for the study
 - The factor: viewer group as defined by the type of show to which the viewer was exposed.
 - Factor levels: V, E and N (or the equivalent names)
 - Response: number of names recalled
 - Design: Completely randomized design

5 Marks

(b) Computing the ANOVA table in R

```
flink<-'https://www.math.mcgill.ca/dstephens/204/Assignments/A1/A1Q1.csv'
Ads<-read.csv(flink)
print(summary(aov(Number~Code,data=Ads)),digits=6)

##              Df  Sum Sq Mean Sq F value    Pr(>F)
## Code           2 123.265  61.6327  20.4518 4.3641e-09 ***
## Residuals    321 967.352   3.0136
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

atab<-anova(aov(Number~Code,data=Ads))
```

Here the p-value is 4.3640595×10^{-9} , so the test is highly significant; the null hypothesis of equal group means is rejected.

5 Marks

The full ANOVA table with rounding to three d.p. is

SOURCE	DF	SS	MS	F	p
Code	2	123.265	61.633	20.452	0.000
ERROR	321	967.352	3.014		
TOTAL	323	1090.617			

To complete the test, we would compare against the appropriate F-distribution, that is, the $F_{2,321}$ distribution. The appropriate critical value for a test at the $\alpha = 0.05$ level is 3.02.

```
qf(1-0.05,atab[1,1],atab[2,1])
## [1] 3.023865
```

(c) We use `tapply` in R to compute the sample means.

```
mvec<-tapply(Ads$Number,Ads$Code,mean) #Sample means
vvec<-tapply(Ads$Number,Ads$Code,var)  #Sample Variances
mvec

##           E           N           V
## 1.712963 3.166667 2.083333
```

Therefore the means are (1.713, 3.167, 2.083) for E, N and V respectively. For the population variance pooled estimate, this is simply the MSE from the ANOVA table, 3.014. We can check this by direct computation of the pooled sample variance

```
table(Ads$Code)

##
##      E      N      V
## 108 108 108

nvec<-as.numeric(table(Ads$Code))
sp.sq<-sum((nvec-1)*vvec)/sum(nvec-1)
sp.sq

## [1] 3.013557
```

5 Marks

This is the correct and intended answer; we only use the term 'pooled estimate' to mean this computed quantity. However, the wording is perhaps ambiguous, and could have been taken to mean the overall variance of the sample, ignoring the group structure. This quantity is

$$\frac{SS}{n-1} = \frac{1090.617}{323} = 3.377.$$

```
var(Ads$Number)

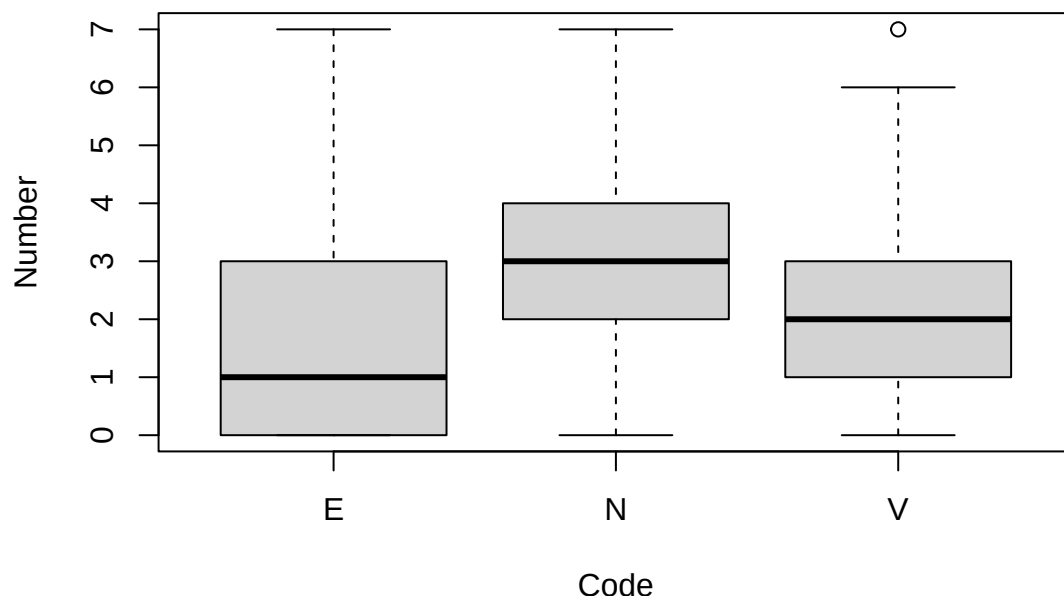
## [1] 3.376524
```

Given the potential ambiguity, this variance answer is acceptable.

(d) The three assumptions

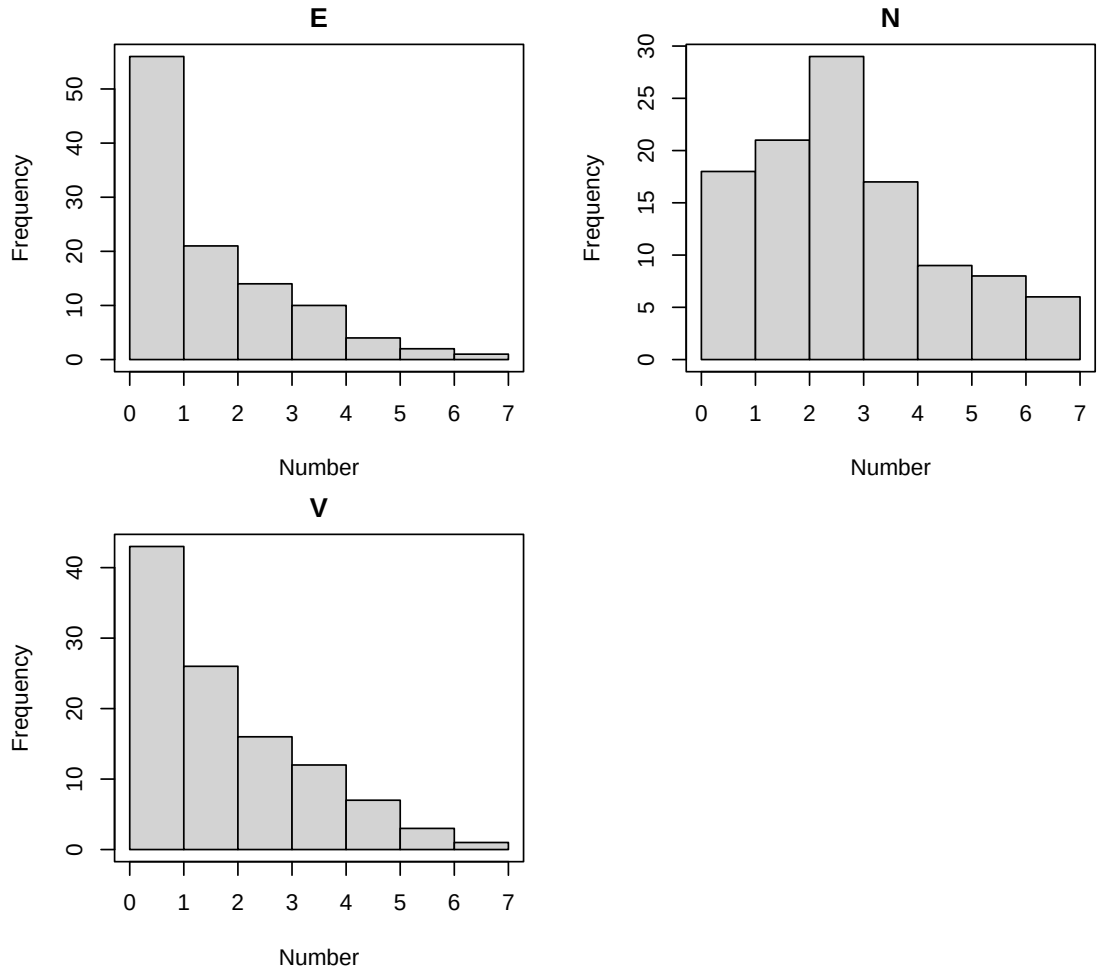
- **Independence:** this can be assumed due to the design being used.
- **Normality:** we can check this using a boxplot, for example

```
par(mar=c(4,4,2,2))
boxplot(Number~Code,data=Ads)
```



The boxplot seems to indicate the Normality assumption is being met, although perhaps there is some skewness in the first group. We therefore also check histograms

```
par(mar=c(4,4,2,2),mfrow=c(2,2))
hist(Ads$Number[Ads$Code=='E'], xlab='Number', main='E');box()
hist(Ads$Number[Ads$Code=='N'], xlab='Number', main='N');box()
hist(Ads$Number[Ads$Code=='V'], xlab='Number', main='V');box()
```



These histograms indicate that the data are **not** Normally distributed.

- **Equal variances:** For this we may use Levene's test

```
library(car)
## Loading required package: carData
leveneTest(Number~as.factor(Code),data=Ads, centering='mean')
## Levene's Test for Homogeneity of Variance (center = median: "mean")
##      Df F value Pr(>F)
## group  2  0.0162 0.9839
##      321
leveneTest(Number~as.factor(Code),data=Ads, centering='median')
## Levene's Test for Homogeneity of Variance (center = median: "median")
##      Df F value Pr(>F)
## group  2  0.0162 0.9839
##      321
```

These results indicate that the equal variance assumption seems appropriate, as the p-values are not below the usual $\alpha = 0.05$ level.

5 Marks

2. (a) Computing the ANOVA table in R

```
flink<-'https://www.math.mcgill.ca/dstephens/204/Assignments/A1/A1Q2.csv'
Jeans<-read.csv(flink)
print(summary(aov(AbsRate~Week+Day,data=Jeans)),digits=6)

##              Df  Sum Sq Mean Sq F value    Pr(>F)
## Week          8 575.228  71.9035  6.10417 8.7408e-05 ***
## Day           4  94.199  23.5498  1.99923  0.11834
## Residuals    32 376.941  11.7794
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

atab<-anova(aov(AbsRate~Week+Day,data=Jeans))
```

10 Marks

The full ANOVA table with rounding to three d.p. is

SOURCE	DF	SS	MS	F	p
Week	8	575.228	71.903	6.104	0.000
Day	4	94.199	23.550	1.999	0.118
ERROR	32	376.941	11.779		
TOTAL	44	1046.368			

Note that here the interaction cannot be fitted, as there is only one replicate per Week/Day combination.

(b) There are two principal conclusions

- The p-value for Week is very small (0.000 to 3 d.p.), indicating that there is a significant effect of the blocking factor, as anticipated in the design. Absentee rates do appear to differ significantly by week.
- The p-value for Day is not small enough (0.118) to indicate a significant difference by day of the week.

Thus the difference in absentee rates is not notably different by day of the week if Week is taken into account.

5 Marks

Note that if the blocking factor is ignored, and the data are treated as if they arose from a CRD, the following ANOVA table is obtained.

```
print(summary(aov(AbsRate~Day,data=Jeans)),digits=6)

##              Df  Sum Sq Mean Sq F value    Pr(>F)
## Day           4  94.199  23.5498  0.98931  0.42445
## Residuals    40 952.169  23.8042
```

- (c) Using the data, construct a 95% confidence interval for the difference in absentee rates between Mondays and Fridays. Explain your calculation.

There are two legitimate ways to answer this question:

- The simplest analysis simply takes the data from the two days, and performs a two sample analysis. There are several ways to do this in R: first using `lm`, subsetting the data to only include the Monday and Friday data

```
f0<-lm(AbsRate~Day,data=Jeans, subset=(Day == 'M' | Day == 'F'))
summary(f0)$coef
##              Estimate Std. Error   t value    Pr(>|t|)
## (Intercept)  6.377778    2.003292  3.1836485 0.005772656
## DayM         2.466667    2.833083  0.8706652 0.396812512
summary(f0)$sigma
## [1] 6.009876
confint(f0)
##              2.5 %    97.5 %
## (Intercept)  2.130988 10.624567
## DayM        -3.539201  8.472534
```

In this analysis, the baseline level of the Day factor is 'F' (the first in alphabetical order), so these results tell us that the confidence interval for the 'Monday - Friday' difference is (-3.539, 8.473), and therefore the confidence interval for the difference of interest is

$$(-8.473, 3.539)$$

As this interval contains zero, the difference is not significant. We can obtain the same conclusion using the two sample t-test functionality.

```
x0<-Jeans$AbsRate[Jeans$Day=='F']
x1<-Jeans$AbsRate[Jeans$Day=='M']
mean(x0)-mean(x1)
## [1] -2.466667
t.test(x0,x1, var.equal=TRUE)
##
## Two Sample t-test
##
## data:  x0 and x1
## t = -0.87067, df = 16, p-value = 0.3968
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -8.472534  3.539201
## sample estimates:
## mean of x mean of y
## 6.377778  8.844444
```

This analysis is made under the assumption of equal residual variances for the two days: given the small sample sizes, this assumption is necessary. The estimate of the difference is simply the difference in the sample means.

```
mean(x0)-mean(x1)
## [1] -2.466667
```

so the absentee rate is slightly lower on Friday than on Monday, but the difference is not significant. From the model summary above, the estimate of the residual error variance is $6.010^2 = 36.119$.

- A more sophisticated analysis accounts for the systematic difference between weeks. It is evident from the ANOVA analysis that blocking factor Week is significant, so it should be accounted for in analysis. Here, as we are forced to use the model with no interaction due to the lack of replication, it the case that the two factors are assumed to act additively; that is, the effect on absentee rate of changing Day from Monday to Friday is the **same** in each week, even if the Weeks have systematically different means. Therefore, if we estimate the contrast parameter for the difference between Monday and Friday, this is an estimate of the quantity required.

This approach has the advantage of using an estimate of the residual error variance, σ^2 , that is obtained from the entire data set.

```
fJeans<-lm(AbsRate~Week+Day,data=Jeans)
summary(fJeans)$coef
```

##		Estimate	Std. Error	t value	Pr(> t)
##	(Intercept)	2.3244444	1.844706	1.2600626	2.167588e-01
##	WeekW2	3.0200000	2.170659	1.3912825	1.737377e-01
##	WeekW3	-0.3200000	2.170659	-0.1474207	8.837249e-01
##	WeekW4	2.2200000	2.170659	1.0227308	3.141050e-01
##	WeekW5	12.2800000	2.170659	5.6572678	2.934221e-06
##	WeekW6	4.3600000	2.170659	2.0086065	5.308006e-02
##	WeekW7	3.4000000	2.170659	1.5663445	1.271048e-01
##	WeekW8	6.4800000	2.170659	2.9852683	5.393580e-03
##	WeekW9	5.0400000	2.170659	2.3218754	2.676231e-02
##	DayM	2.4666667	1.617914	1.5245971	1.371829e-01
##	DayR	-1.0222222	1.617914	-0.6318150	5.319959e-01
##	DayT	-1.7777778	1.617914	-1.0988087	2.800502e-01
##	DayW	-0.5888889	1.617914	-0.3639804	7.182666e-01

```
confint(fJeans)
```

##		2.5 %	97.5 %
##	(Intercept)	-1.43309781	6.081987
##	WeekW2	-1.40148798	7.441488
##	WeekW3	-4.74148798	4.101488
##	WeekW4	-2.20148798	6.641488
##	WeekW5	7.85851202	16.701488
##	WeekW6	-0.06148798	8.781488
##	WeekW7	-1.02148798	7.821488
##	WeekW8	2.05851202	10.901488
##	WeekW9	0.61851202	9.461488
##	DayM	-0.82891590	5.762249
##	DayR	-4.31780479	2.273360
##	DayT	-5.07336034	1.517805
##	DayW	-3.88447145	2.706694

Therefore the estimate of the mean difference 'Friday - Monday' as before is -2.4666667, but the confidence interval is

$$(-5.762, 0.829)$$

which is much narrower than the previous one. This is due to the fact that the estimate of the residual error variance from this analysis is $3.432^2 = 11.779$, as given in the ANOVA table as the MSE quantity.

Either of these approaches is acceptable here, provided the analysis is correctly explained.

5 Marks

3. (a) From the ANOVA tables, we may read off from the DF column that $a - 1 = 5$, so $a = 6$, $b - 1 = 4$, so $b = 5$. Finally, we know that $n = abr$ for balanced and complete replication, so as from the ANOVA table we know that for the ERROR DF we have $n - ab = 30$, and with $ab = 30$, we must have that $n = 60$, and so $r = n/(ab) = 2$. 3 Marks

- (b) The difference is that for the second analysis the interaction term between the two factors has been removed. That is, the first analysis uses the model $A + B + A : B$, and the second analysis uses the model $A + B$. 2 Marks

- (c) This is simply because the F-statistic is computed for Factor A

$$F = \frac{MSA}{MSE}$$

and even though MSA does not change, MSE **does** change from 0.838 to 0.748. The same results holds for Factor B.

This change in MSE value shows that even though adding in all the interaction parameters makes the SSE smaller, it does not decrease the MSE.

4 Marks

- (d) It is clear from inspecting the p-values in the two tables that there is a significant effect of Factor A and a significant effect of Factor B, but no significant interaction; the p-values in the first case are not greater than 0.00005 (we have rounding to 4 d.p., but the p-value for the interaction is 0.7672, which is not below the usual $\alpha = 0.05$ level. 5 Marks

- (e) Given that the no interaction model is being preferred, the estimate (that is, the MSE) from that analysis, 0.748, is arguably more appropriate. However, it could also be argued that the interaction model-based estimate, 0.838, is more suitable, as even though the interaction is not significant, there is still a small contribution to the overall variability that is due to presence of interaction terms.

Either of these arguments could be made here.

3 Marks

- (f) It would be advisable to check Normality (via a boxplot or histogram), and equal (residual) variances across the factor level combinations. If either of these does not hold, the conclusions from the ANOVA analyses may not be correct. 3 Marks