STAT 571B Exam 1

Amy Fan

2 Mar 2024

(0) Academic Honesty Statement I prepared my solutions to this exam in accordance with the guidelines on the exam. I did not consult directly with anyone else, nor did I post any questions anywhere for others to see. Signature: Amy Fan Date: 3 Mar 2024

I. First section

(1) [2 pts] Which statistical design technique is used to protect against inferential bias introduced by unmeasured, potentially confounding variables? *The factorial principle*

(a) Randomization
(b) Blocking
(c) Replication
(d) **Factorial principle**

(2) [2 pts] Which statistical design technique is used to control for the effects of nuisance factors to increase the precision of estimates of treatment effects?

(a) Randomization
(b) **Blocking**
(c) Replication
(d) Factorial principle

(3) [6 pts] A mathematician studying the quality of various brands of chalk would like to know which one produces the quietest sound during use. She plans to compare three brands (A, B, and C) by measuring the volume of the chalk using a sensitive microphone. She also wants to control for the nuisance effect of which individual person in her department is using the chalk, so she selects three random mathematicians (Drs. X, Y, and Z) from her department to agree to test each brand of chalk.

(a) [3 pts] What type of design would be appropriate for this experiment? Why is it appropriate? What benefits does your suggested design offer?

Compare brands (A,B,C) while blocking by professor (X,Y,Z) - We can do a Randomized Complete Blocking Design. This gives us maximum power because of balanced design and orthogonal predictors.

Block X	Block Y	Block Z
Chalk A	Chalk C	Chalk B
Chalk B	Chalk A	Chalk C
Chalk C	Chalk B	Chalk A

(b) [3 pts] Now suppose the experimenter only has two pieces of chalk for each brand for the study (six total). What type of design would be appropriate under this constraint? Provide a detailed description of the design.

A Balanced Incomplete Block Design would be the most appropriate with these constraints. Since we only have two pieces of each chalk, we cannot do a randomized complete block design. However, we can still keep the design balanced and allocate 2 chalk to 2 professors for each comparison. a=3, b=3, r=2, k=2, λ=1.

table	Block X	Block Y	Block Z
A	A	-	A
B	B	B	-
C	-	C	C

(4) [9 pts] Each figure below (A–C) corresponds to a diagnostic plot for the model

$$y_{ij} = \mu + \tau_i + \beta_j + \epsilon_{ij}, \epsilon_{ij}\ N(0, \sigma^2),$$

where τ_i is the effect of the ith level of a treatment factor and β_j is the effect of a non-ignorable nuisance factor, fit to data. For each figure, explain which modeling assumption is shown to be violated and make a suggestion for remediation.

A) There appears to be a possible interaction between group 3 and the treatment. We can address this by adding an interaction term between the group and treatment in the model. e.g. $y_{ij} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \epsilon_{ij}, \epsilon_{ij}\ N(0, \sigma^2)$ where $(\tau\beta)_{ij}$ is an interaction.

B) There is nonnormality of residuals in the model. We can remedy this by performing a transform on the data such as a log transform on the response: $log(y_{ij}) = \mu + \tau_i + \beta_j + \epsilon_{ij}, \epsilon_{ij}\ N(0, \sigma^2)$

C) Heteroskedasticity in the residuals can also be remedied by a transform to the response: $log(y_{ij}) = \mu + \tau_i + \beta_j + \epsilon_{ij}, \epsilon_{ij}\ N(0, \sigma^2)$

II. Second section

(5) [5 pts] Henry is determined to find the best cup of coffee near campus. After some preliminary investigation, he narrows his search to two cafes and now he wants to determine the number of samples he'll need from each cafe so that he will be able to detect a difference in mean rating of 0.25 with 95% confidence and power 0.9. He has pilot data from two other cafes to help. Conduct a power analysis based on the pilot data to estimate how many cups of coffee Henry will need to test from each cafe.

We would need a sample size of n=14 from each cafe to detect a difference in mean rating of 0.25 with 95% confidence and power 0.9.

```
# pilot data
coffee <- data.frame("cafe 1" = c(3.1, 3.5, 3.2, 3.2),
                     "cafe 2" = c(4.5, 4.4, 4.0, 4.2))
set.seed(100)
n <- 14 # CHANGE
sigma <- sqrt( ( (n-1)*sd(coffee$cafe.1)^2 + (n-1)*sd(coffee$cafe.2)^2 ) / (2*n-2) )
alpha <- 0.05
typeII <- sapply(1:2000, function(i){
  mus <- c(mu1 = 3.25, mu2 = 3.5)
  sim_data <- data.frame(cafe = as.factor(rep(c(1,2), rep(n, 2))),
                         rating = c(sapply(mus, function(mu) rnorm(n, mu, sd = sigma))))
  sim_aov <- aov(rating ~ as.factor(cafe), data = sim_data)
  summary(sim_aov)[[1]]$`Pr(>F)`[1] < alpha
})
mean(typeII) # Needs to be ~ 0.90
```

```
## [1] 0.9005
```

```
## [1] 0.9005
```

(6) [8 pts] The owner of a local pizza restaurant called Saucetistics is interested in growing his own tomatoes to make sauce for the pizzas. He conducted an experiment to determine which of four varieties of tomato yielded the most fruit in the hot desert spring weather. To control for the possibility that the soil is different in his field Close to or Far from the road, he chose an appropriate randomized design and recorded where each test plant grew. A total of 32 plants were used in the experiment, and the number of fruit produced by each plant ("yield") recorded in the file tomatoes.csv.

(a) [3 pts] Examine the data. What type of design was used? How many replications are there?

This experiment uses 4X2 Factorial Design. There are 4 varieties (A-D) and 2 locations (far/close). There are 4 replicates.

```
tomatoes <- read.csv("./tomatoes.csv")
tomatoes
```

```
##    variety location yield
## 1        A    Close    10
## 2        A    Close     8
## 3        A    Close     6
## 4        A    Close     6
## 5        B    Close    15
## 6        B    Close    16
## 7        B    Close    13
## 8        B    Close    10
## 9        C    Close     8
## 10       C    Close     8
## 11       C    Close     6
## 12       C    Close    14
## 13       D    Close     7
## 14       D    Close     5
## 15       D    Close     0
## 16       D    Close     3
## 17       A      Far     6
## 18       A      Far    11
## 19       A      Far     3
## 20       A      Far     1
## 21       B      Far    18
## 22       B      Far    15
## 23       B      Far    18
## 24       B      Far    13
## 25       C      Far    17
## 26       C      Far      9
## 27       C      Far     15
## 28       C      Far     18
## 29       D      Far      0
## 30       D      Far      6
## 31       D      Far      1
## 32       D      Far      5
```

(b) [2 pts] Test the hypothesis that all varieties of tomato plant produce the same expected number of fruit while controlling for location. Report the p-value and conclusion. **Using anova, we have a p-value of 6.99e-07 which means that with an alpha of 0.05 not all varieties of tomato plant produce the same expected number of fruit while controlling for location.**

```
tomatoes_aov <- aov(yield ~ location*variety, data = tomatoes)
summary(tomatoes_aov)
```

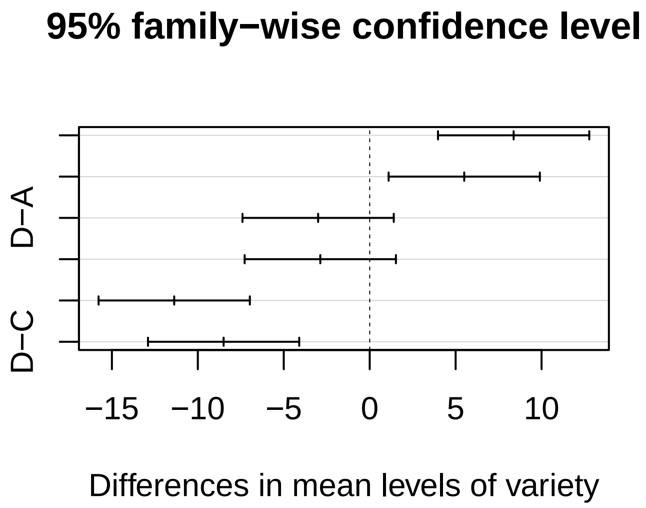
```
##                   Df Sum Sq Mean Sq F value   Pr(>F)
## location           1   13.8   13.78   1.354   0.2560
## variety            3  638.6  212.86  20.916 6.99e-07 ***
## location:variety   3   76.1   25.36   2.492   0.0843 .
## Residuals         24  244.2   10.18
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(c) [3 pts] Use Tukey's Test to compare all possible pairs of tomato variety. Which variety(ies) appear to produce the most fruit? Explain your reasoning. **With a threshold of 0.05, a Tukey's test reveals that A/B, D/B, D/C, and A/C are significantly different. This means we can create two groups: (B,C) varieties appear to produce more fruit than (A, D) varieties.**

```
tomatoes_hsd <- TukeyHSD(tomatoes_aov, which = "variety")
tomatoes_hsd
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = yield ~ location * variety, data = tomatoes)
##
## $variety
##        diff        lwr       upr     p adj
## B-A   8.375   3.974806 12.775194 0.0001226
## C-A   5.500   1.099806  9.900194 0.0105226
## D-A  -3.000  -7.400194  1.400194 0.2625497
## C-B  -2.875  -7.275194  1.525194 0.2967529
## D-B -11.375 -15.775194 -6.974806 0.0000013
## D-C  -8.500 -12.900194 -4.099806 0.0001008
```

```
plot(tomatoes_hsd)
```



(7) [8 pts] An audiophile who goes by the name Samba Squares has decided to conduct an experiment to see which of four manufacturers (denoted I-IV) of over-the-ear headphones have the best sound quality. Samba wants the findings to be broadly generalizable to various types of music in many settings, so they decide to test the audio quality of each headphone across four different songs and four different background conditions. The design is shown in the table below, and the data are available in headphones.csv.

(a) [2 pts] What kind of design is this? **This is a Randomized Complete Block Design with 2 blocking variables (Latin Square).**

```r
headphones <- read.csv("./headphones.csv")
headphones
```

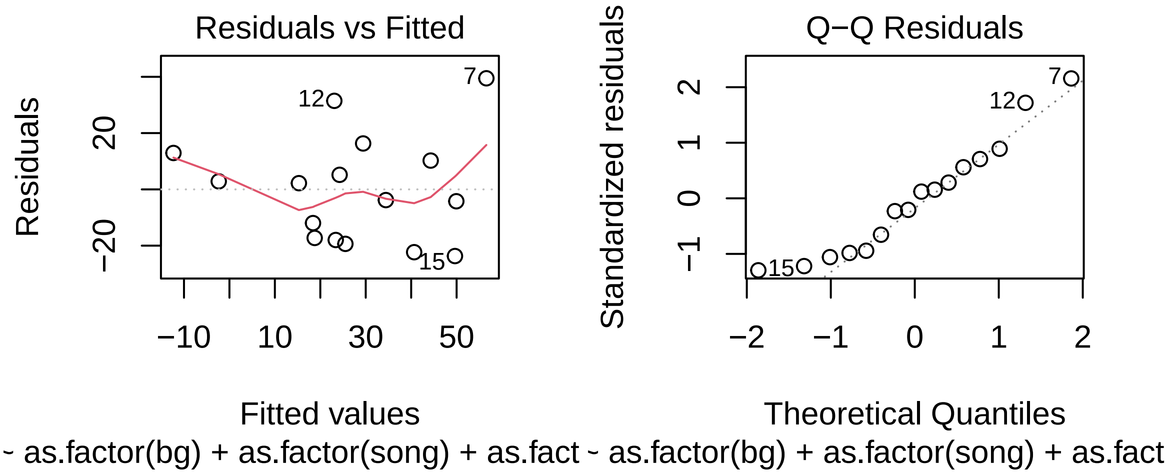
```
##    manufacturer song bg    quality
## 1             I    1  1  0.6138319
## 2            II    2  1 54.5278408
## 3           III    3  1 30.6313820
## 4            IV    4  1  6.1540980
## 5            II    1  2  5.4091655
## 6           III    2  2 45.7115598
## 7            IV    3  2 96.0283554
## 8             I    4  2  1.4962495
## 9           III    1  3  0.4991348
## 10           IV    2  3 18.3319432
## 11            I    3  3  6.4113595
## 12           II    4  3 54.5389050
## 13           IV    1  4 17.4997444
## 14            I    2  4 45.7420486
## 15           II    3  4 25.9380382
## 16          III    4  4 29.4262055
```

(b) [2 pts] Fit an additive model for audio quality as a function of manufacturer. Create diagnostic plots and assess the degree to which each model assumption is met: (i) normality; (ii) constant variance. **The QQ plot shows nonnormal residuals, especially in the smaller values. There appears to be heteroskedasticity in the fitted v. residuals plot. There also appears to be nonconstant variance in the predictor v. residuals plots: manufacturer III has a low variance, and song 3 has an abnormally high variance.**

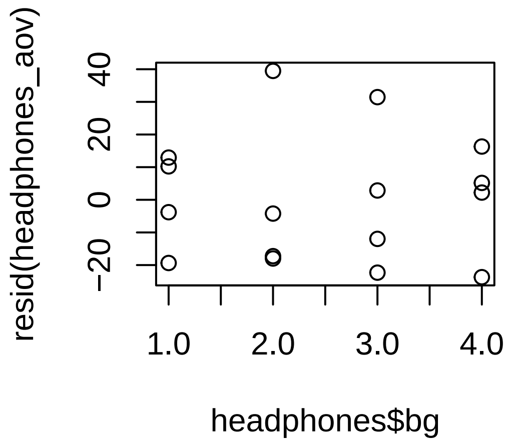
```r
headphones_aov <- aov(quality ~ as.factor(bg) + as.factor(song) + as.factor(manufacturer), data = headph
summary(headphones_aov)
```

```
##                        Df Sum Sq Mean Sq F value Pr(>F)
## as.factor(bg)           3    702   233.9   0.262  0.851
## as.factor(song)         3   3270  1090.1   1.221  0.381
## as.factor(manufacturer) 3   1208   402.5   0.451  0.726
## Residuals               6   5359   893.1
```

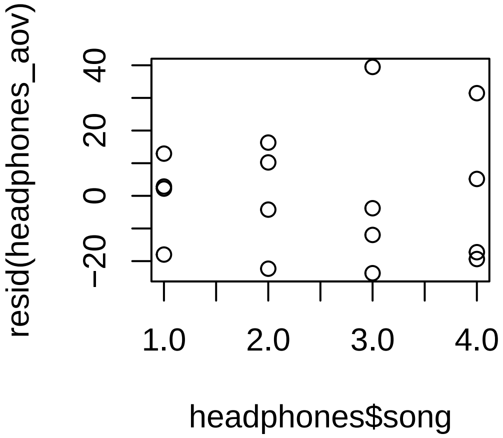
```r
plot(headphones_aov, which = c(2,1))
```



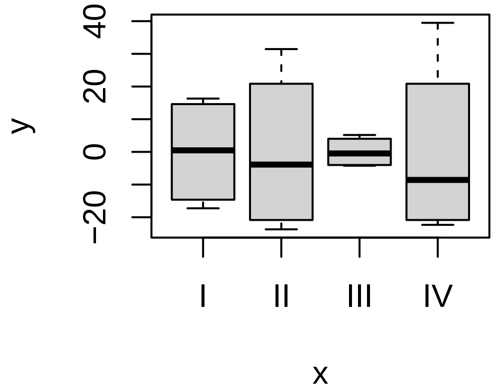
```r
plot(headphones$bg, resid(headphones_aov))
```



```
plot(headphones$song, resid(headphones_aov))
```



```
plot(as.factor(headphones$manufacturer), resid(headphones_aov))
```

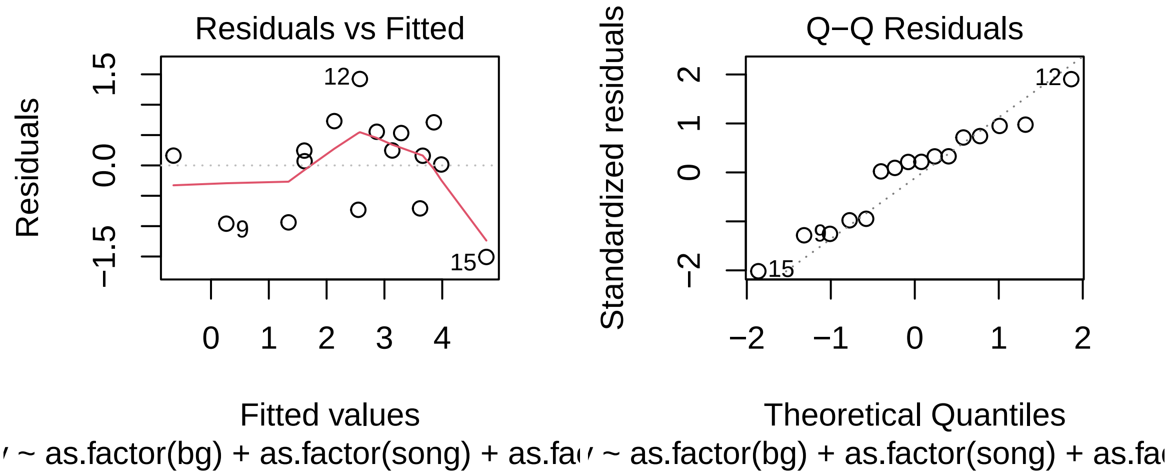


(c) [2 pts] Find a transformation to the response vartiable that results in a model for which (i) and (ii) are better satisfied. Provide your transformation. **The log transform on the response (quality) produces normal residuals (QQ plot) and better satisfies the constant variance assumption (predictor v residuals and fitted v residuals plots all appear to have constant variance).**

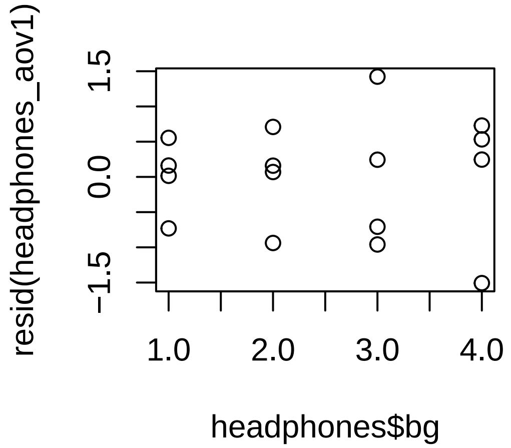
```
headphones$logquality = log(headphones$quality)

headphones_aov1 <- aov(logquality ~ as.factor(bg) + as.factor(song) + as.factor(manufacturer),
                       data = headphones)
summary(headphones_aov1)
```

```
##                            Df Sum Sq Mean Sq F value Pr(>F)
## as.factor(bg)               3  4.114   1.371   0.921  0.486
## as.factor(song)             3 18.599   6.200   4.162  0.065 .
## as.factor(manufacturer)  3  8.147   2.716   1.823  0.243
## Residuals                   6  8.938   1.490
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

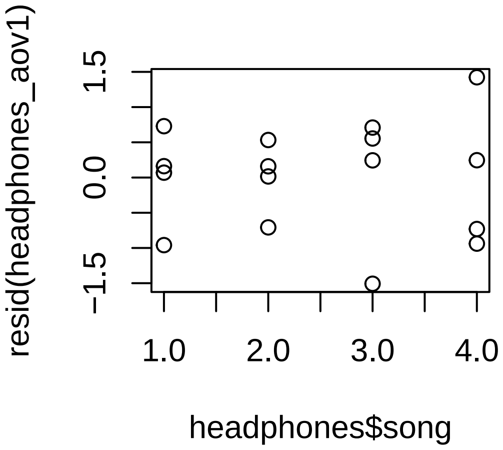
```
plot(headphones_aov1, which = c(2,1))
```



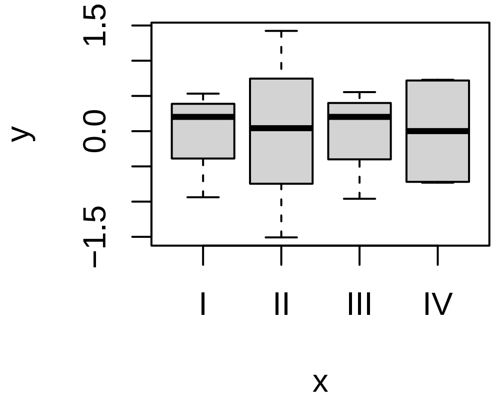
```
plot(headphones$bg, resid(headphones_aov1))
```



```
plot(headphones$song, resid(headphones_aov1))
```



```
plot(as.factor(headphones$manufacturer), resid(headphones_aov1))
```



(d) [2 pts] Based on your transformed response in (b), which factor appears to have more evidence of a non-zero effect on audio quality between song and background condition? **With a p-value of 0.065, the song variable seems to have the most likely affect on the sound quality.**

```
summary(headphones_aov1)
```

```
##                          Df Sum Sq Mean Sq F value Pr(>F)
## as.factor(bg)             3  4.114   1.371   0.921  0.486
## as.factor(song)           3 18.599   6.200   4.162  0.065 .
## as.factor(manufacturer)   3  8.147   2.716   1.823  0.243
## Residuals                 6  8.938   1.490
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
TukeyHSD(headphones_aov1)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = logquality ~ as.factor(bg) + as.factor(song) + as.factor(manufacturer), data = hea
##
## $‘as.factor(bg)‘
##           diff       lwr      upr     p adj
## 2-1  0.4320577 -2.555526 3.419642 0.9559896
## 3-1 -0.1697672 -3.157351 2.817817 0.9970056
## 4-1  1.1432450 -1.844339 4.130829 0.5818311
## 3-2 -0.6018249 -3.589409 2.385759 0.8945510
## 4-2  0.7111873 -2.276397 3.698771 0.8415147
## 4-3  1.3130122 -1.674572 4.300596 0.4814347
##
## $‘as.factor(song)‘
##           diff        lwr      upr     p adj
## 2-1  2.7963395 -0.1912445 5.783923 0.0647257
## 3-1  2.4332707 -0.5543133 5.420855 0.1068448
## 4-1  1.5583780 -1.4292059 4.545962 0.3555039
## 3-2 -0.3630688 -3.3506527 2.624515 0.9728422
## 4-2 -1.2379615 -4.2255454 1.749622 0.5247599
## 4-3 -0.8748927 -3.8624766 2.112691 0.7482534
##
## $‘as.factor(manufacturer)‘
##              diff       lwr      upr     p adj
## II-I    1.8363536 -1.151230 4.823938 0.2453867
## III-I   1.0838415 -1.903742 4.071425 0.6185772
## IV-I    1.6391441 -1.348440 4.626728 0.3199350
## III-II -0.7525121 -3.740096 2.235072 0.8192726
## IV-II  -0.1972095 -3.184793 2.790374 0.9953413
## IV-III  0.5553026 -2.432281 3.542886 0.9141115
```