Homework Two

Josh Virene

Problem 1

Problem 1(a)

Plot A exhibits the most 'within group' variation out of all three plots and, will result in a large SS_{error} because the individual observations vary most significantly from the mean. MS_{error} is calculated as SS_{error}/DF_{error} so because plot A has the largest SS_{error}, it will have the largest MS_{error}.

Problem 1(b)

Plot C exhibits the most 'between group' variation of the three plots, and will result in a large $SS_{treatment}$ because the sample means for each group differ the most from the overall mean across all groups. $MS_{treatment}$ is equal to $SS_{treatment}/DF_{treatment}$; because plot C has the largest $SS_{treatment}$ of the three, it will have the largest $MS_{treatment}$.

Problem 1(c)

The plot that will have the largest F-statistic is plot C. The formula for the F_{test} is $F_{test} = MS_{treatment}/MS_{error}$. Plot C has the largest MS_{error}, plot B has small values for MS_{error} and $MS_{treatment}$, and plot A has small $MS_{treatment}$, but large MS_{error}. Based on the formula for the F statistic, plot C will have the largest F_{test}

Problem 1(d)

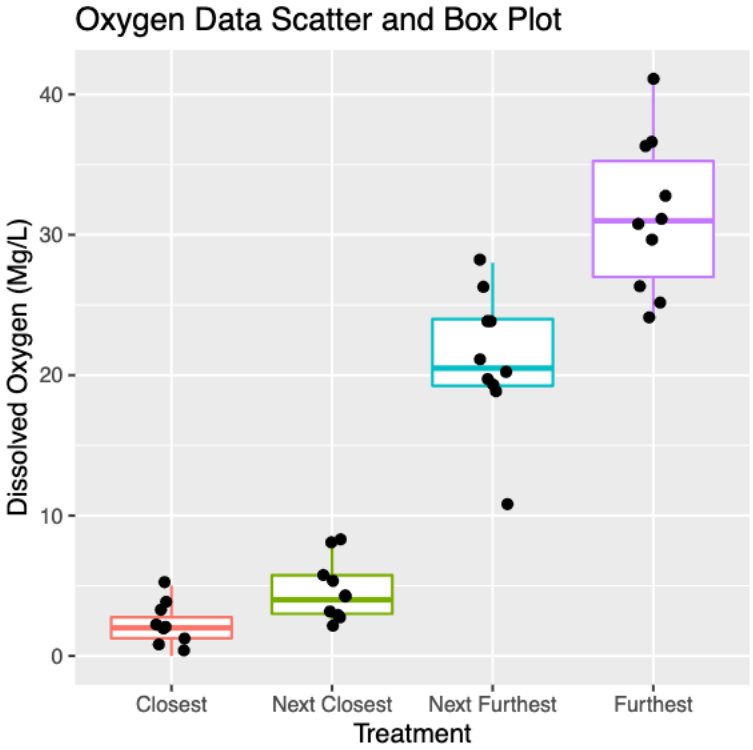
The experiment that resulted in the largest p-value will be plot B. For these ANOVA plots, the null hypothesis H_0 states that $\mu_1 = \mu_2 = \mu_3 = \mu$, the population means are equal to each other. Looking at these three plots, the means for the three groups in plot B are very close together and there is very little between group variation. The p-value will be large for this group because out of all the plots, in plot B we have the lowest confidence in rejecting the null hypothesis H_0 saying that these groups are equal based on the small $MS_{treatment}$, how close these means are to each other.

Problem 2

This researcher does not see the group means because he did not convert treatment to a factor. R is treating treatment as a numeric variable instead of a factor and there is a slope coefficient for treatment 'trt' in the R output as a result. In order to fix this issue, the researcher should convert treatment to a factor, as is done in the code below: (also written in the appendix) data(dollarsign)factor = as.factor(data$treatment)

Problem 3

Problem 3(a)



Problem 3(b)

distance	n	sample_mean	sample_sd	se
1	10	2.2	1.475730	0.4666667
2	10	4.6	2.118700	0.6699917
3	10	21.2	4.732864	1.4966630
4	10	31.4	5.521674	1.7461068

Problem 3(c)

```
## Analysis of Variance Table
##
## Response: oxygen
##           Df Sum Sq Mean Sq F value    Pr(>F)
## distance   3 5793.1 1931.03   129.7 < 2.2e-16 ***
## Residuals 36  536.0   14.89
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Problem 3(d)

The decision for the F-test based on this ANOVA table output will be to reject the null hypothesis that $\mu_1 = \mu_2 = \mu_3 = \mu_k$. The F-statistic measures the ratio of the variability of between group means in this case $MS_{distance}$ to that of within group variability MS_{error} for each sample. Larger f values mean that the between group variability is larger than that within groups. From this we conclude that not all of the population group means are equal. In terms of this experiment, this means that the dissolved oxygen content will vary depending on the distance from the mouth of the Mississippi River.

Problem 4

Problem 4(a)

$\bar{Y_1}\cdot = 50.3$ $\bar{Y_2}\cdot = 50.3 + 10.8 = 61.1$ $\bar{Y_3}\cdot = 50.3 - 12.0 = 38.3$

Problem 4(b)

$\bar{Y_1}\cdot = 50.3$ $\bar{Y_2}\cdot = 61.1$ $\bar{Y_3}\cdot = 38.3$

$\bar{Y}\cdot\cdot = (\Sigma_{i=1}^{k}\Sigma_{j=1}^{n}Y_{ij})/nk$ $\bar{Y}\cdot\cdot = (\bar{Y_1}\cdot + \bar{Y_2}\cdot + \bar{Y_3}\cdot)/3$ $\bar{Y}\cdot\cdot = 49.9$

$SS_{treatment} = n\Sigma_{i=1}^{k}(\bar{Y}\cdot - \bar{Y}\cdot\cdot)^2 = 3121.92$

Problem 5

Calculating the values in the ANOVA table (with steps)

A. $df_{treatmnet} = df_{total} - df_{error}$ $df_{treatment} = 4$

B. $SS_{total} = SS_{error} + SS_{treatment}$ $1174.24 = 186.52 + SS_{treatment}$ $SS_{treatment} = 1174.24 - 186.52$ $SS_{treatment} = 987.71$

C. $MS_{error} = SS_{error}/df_{error}$ $MS_{error} = 186.53/25$ $MS_{error} = 7.4612$

D. $F_{stat} = MS_{treatment}/MS_{error}$ $F_{stat} = 33.095$

Appendix

```
library(knitr)
# install the tidyverse library (do this once) install.packages('tidyverse')
library(tidyverse)
knitr::opts_chunk$set(echo = FALSE, message = FALSE, warning = FALSE, fig.width = 4,
    fig.height = 4, tidy = TRUE)
# code for prob1a code for prob1b

# code for prob1a

# code for prob1d

# code for prob2 data$factor = as.factor(data$treatment) code for prob3a
```

```
## Setup items: load dataset
oxygendata <- read_csv("oxygen.csv", show_col_types = FALSE)
attach(oxygendata)
# convert distance into a factor, right now it is a number
oxygendata$distance = as.factor(oxygendata$distance)

## 3a code
ggplot(oxygendata, aes(x = distance, y = oxygen)) + geom_boxplot(aes(color = distance),
    outlier.colour = NA, position = "dodge") + geom_jitter(position = position_jitter(width = 0.1)) +
    ylab("Dissolved Oxygen (Mg/L)") + ggtitle("Oxygen Data Scatter and Box Plot") +
    scale_x_discrete(name = "Treatment", breaks = c("1", "2", "3", "4"), labels = c("Closest",
        "Next Closest", "Next Furthest", "Furthest")) + theme_gray(base_size = 10) +
    theme(legend.position = "none")
# code for prob3b
oxygen_summary2 <- oxygendata %>%
    group_by(distance) %>%
    summarize(n = length(oxygen), sample_mean = mean(oxygen), sample_sd = sd(oxygen),
        se = sample_sd/sqrt(n))
kable(oxygen_summary2)
# code for prob3c
lm_oxygen <- lm(oxygen ~ distance, data = oxygendata)
anova(lm_oxygen)
# code for prob3d

# code for prob4a code for prob4b

# code for prob5
```