HW 6 Solutions (26 pts)

1. (4 pts, 2 for correct CI, 2 for correct 1-sided test)

   We found (in HW #5) that an ANOVA fit to the Incomes across the 5 education levels did not satisfy the constant variance assumption. One way to address this is to $\log_{10}$ transform the incomes. A residual plot is shown in the Appendix that suggests that this assumption is not violated with the $\log_{10}$-transformed incomes across education groups. Based on this ANOVA, the data suggest that there is some difference in the mean $\log_{10}$ transformed incomes based on education level ($F = 62.9$, p-value $< 2.2\times10^{-16}$). In other words, the data suggest that there is some difference in the median incomes based on education level ($F = 62.9$, p-value $< 2.2\times10^{-16}$)

   Assuming that the researchers *planned* before the study was conducted to test that the median income level for workers with a high school education is less than the median income of workers with a bachelor's degree, a follow-up t-test was performed with degrees of freedom equal to $DFE = 2579$ from the ANOVA, and with $MSE = 0.1632$. Hence, the evidence suggests that the median education level of folks with a bachelor's degree was larger than the the median income level for folks with a high school education ($t = 10.44$, p-value < 0.0001). A 95% CI for the difference (Bachelor's - HS) in the $\log_{10}$ transformed incomes is [0.20, 0.29]. Back transforming this CI gives a 95% CI for the ratio in medians, $[10^{0.20},\ 10^{0.29}] = [1.59,\ 1.97]$. Therefore, we are 95% confident that the median income for folks with a bachelor's degree is 59% to 97% larger than the median income for folks with only a high school education. See R-code and R-output in the Appendix.

2. (10 pts total) The $\log_{10}$ transformed incomes were also analyzed by a permutation test using the variance of the mean $\log_{10}$ transformed incomes as the test statistic to test whether there was any difference in median incomes across the 5 educational levels.

   (a) (4 pts) The assumptions of a permutation test of these data are:

   - Same shape One assumption of this permutation test is that the groups have the same shape (but not necessarily normal) and hence also the same variance! Hence, we must apply the permutation test to the $\log_{10}$ transformed incomes which do not appear to violate the assumption of constant variance appear to have constant variance across the groups as indicated by the residual vs. fits plot in the Appendix.
   - Independence within groups: As these data were obtained from a random sample of individuals in the National Longitudinal Youth Study each individual is independent of the others.
   - Independence between groups: Similarly, because of the random sample of individuals the independence between educational levels does not seem to be violated. As one person cannot be sampled multiple times (in different groups).

   (b) (3 pts) Results There are few different test statistics that you can use (SSG, MSG, variance of the means, F) with the permutation test. Here, the variance of the means is used as the test statistic. The approximate permutation distribution of the variance of the means test statistic over $10,000$ simulations is shown in Figure 1 in the Appendix.

   (c) (1 pt)

   p-value: The variance in means for the actual data set is 0.032 - this is the test statistic. There was not a single permuted data set that attained or exceeded this value; hence, p-value $< \frac{1}{10000}$.

   (d) (2 pts) Compare ANOVA and PERMANOVA As reported in #1 above, based on a (parametric) ANOVA of these same data, the data suggest that there is some difference in the median incomes based on education level ($F = 62.9$, p-value $< 2.2\times10^{-16}$). The PERMANOVA performed here also shows a tiny p-value; hence, from the PERMANOVA, we also conclude that the data suggest that there is some difference in the median incomes based on education level (variance in means $= 0.032$, p-value $< 2.2\times10^{-16}$).

3. (2 pts) Kruskal-Wallis non-parametric ANOVA was applied to incomes and also to $\log_{10}$-transformed incomes (see Appendix) to test whether the median income is different for the educational levels. We get *exactly* the same test statistic value ($\chi^2 = 349.5$) and p-value ($< 2.2 \times 10^{-16}$) because Kruskal-Wallis analyzes the ranks of the data and log transforms do not re-order the data or the ranks. Hence, the ranks of the incomes and the ranks of $\log_{10}$-transformed incomes are the same. These results lead to a p-value that is indistinguishable from that generated by ANOVA and PERMANOVA.

4. (10 pts total) Three experiments were performed at MSU's Center for Biofilm Engineering in January, February and October of 2017. In each experiment, a biofilm that represents those found in chronic wounds was grown on three different surfaces for a total of $n = 9$ biofilms. The density of bacteria in the biofilm was recorded as number of bacteria per cm^2. It is important to grow up a similar biofilm in multiple experiments so that any results regarding topical antibiotics or anti-microbials are with respect to the same bio-challenge.

   (a) (1 pt) Displaying these data using an individual value plot (Figure 2) is preferable to a boxplot because there are only $n_i = 3$ data per experiment! A boxplot uses a five number summary to generate the plot, why would you ever use a five number summary of less than 5 data points!

   (b) (1 pt) It would be appropriate to present a 95% CI of the true mean $\log_{10}$-transformed densities using a 1-sample t-CI with 8 df *if* the $n = 9$ biofilms were independent and not clustered by experiments. A random effects ANOVA accounts for the clustering of the biofilm samples in each experiment. This model assumes that these three experiments are representative of **all** possible experiments and uses the results of these experiments to estimate the distribution of **all** possible experiments. The random effects ANOVA uses only $I - 1 = 2$ degrees of freedom to build a 95% CI for these data (as for problem #4f below).

   (c) (2 pts) A random effects ANOVA was fit to these data with a random effect for experiment. Three assumptions need to be checked, Normality of the residuals, constant variance, and independence across Dates. Independence across dates is met if experiments performed on different days are independent of one another. Residual plots (Figure 3) do not suggest that the Normality assumption is violated because the points in the normal probability plot are fairly close to the line. Due to the small sample size of this study, we cannot utilize the Central Limit Theorem which gives us Normality of the sampling distribution of the mean. The Residual vs. Fitted plot, however, suggests that the constant variance assumption across experiments may be violated because of the larger spread in residuals for the two dates with higher bacterial densities compared to the single date with a slightly smaller bacterial density.

   (d) (1 pt) The $\log_{10}$-transformed densities that were produced from the same experiment are not independent. The correlation among these biofilm $\log_{10}$ densities is estimated to be $r = \frac{0.015163}{0.015163+0.006933} = 0.686$.

   (e) (1 pt) The proportion of variance (out of the total variance) due to experiment-to-experiment sources is $r = 0.686$.

   (f) (3 pts, 1 for the CI, 2 for the interpretation) A 95% CI for the true mean $\log_{10}$-transformed densities of the chronic wound biofilms is [6.4, 7.1]. We are 95% confident that the true mean log_{10}-density of biofilm in these chronic wound experiments is between 6.4 and 7.1.

   (g) (1 pt) Back transforming we get $[10^{6.4},\ 10^{7.1}] = [2.7 \times 10^6,\ 1.2 \times 10^7]$. We are 95% confident that the true median density of the chronic wound biofilms is between 2.7×10^6 and 1.2×10^7 bacteria per cm^2.

Appendix

Housekeeping

```
source("http://www.math.montana.edu/parker/courses/STAT411/diagANOVA.r")  # residual plots
library(Sleuth3)   # to get datasets
library(ggplot2)   # to use qplot() in #3
library(lme4)      # to use lmer() in #4
```

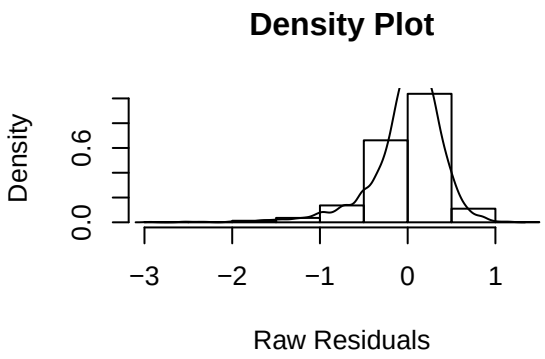
```
## Loading required package: Matrix
```

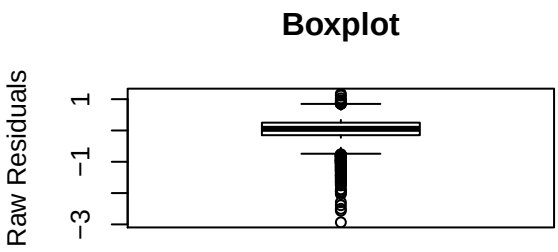
Problem 1

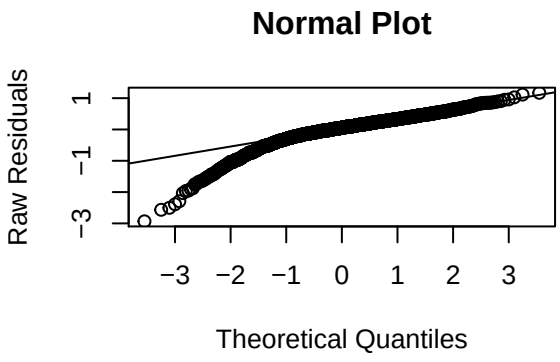
```
# Import the data
library(Sleuth3)
d = ex0525
d$Educ = factor(as.character(d$Educ),levels = c("<12","12","13-15","16",">16"))
summary(d)
```

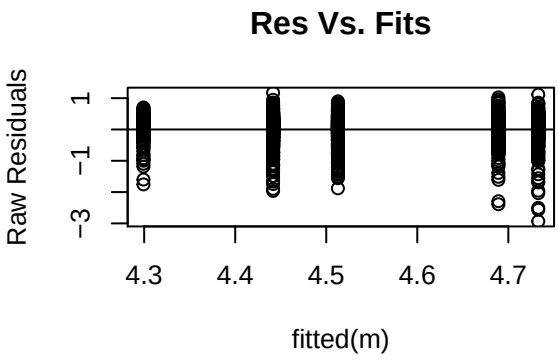
```
##     Subject          Educ        Income2005
##  Min.   :    2   <12  : 136   Min.   :    63
##  1st Qu.: 1586   12   :1020   1st Qu.: 23000
##  Median : 3108   13-15: 648   Median : 38231
##  Mean   : 3494   16   : 406   Mean   : 49417
##  3rd Qu.: 4636   >16  : 374   3rd Qu.: 61000
##  Max.   :12140                Max.   :703637
```

```
# Conduct ANOVA on log-transformed data
m = lm(log10(Income2005) ~ Educ,data=d)
diagANOVA(m)   # Assess the residuals
```










```
anova(m)  # ANOVA table
```

```
## Analysis of Variance Table
##
## Response: log10(Income2005)
##             Df Sum Sq Mean Sq F value    Pr(>F)
## Educ         4  41.05 10.2630   62.87 < 2.2e-16 ***
## Residuals 2579 421.00  0.1632
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
DFE = 2579
MSE = 0.1632

# Get means and sample sizes
Mean = tapply(log10(d$Income2005), d$Educ, mean)
SD = tapply(log10(d$Income2005), d$Educ, sd)
N = tapply(log10(d$Income2005), d$Educ, length)
data.frame(Mean,SD,N)
```

```
##           Mean        SD    N
## <12   4.299229 0.4338084  136
## 12    4.441623 0.3708812 1020
## 13-15 4.512845 0.4033802  648
## 16    4.689115 0.4160997  406
## >16   4.732899 0.4632146  374
```

```r
# Perform the follow-up t-test
# Test-stat
(4.689115 - 4.441623)/sqrt(0.1632*(1/1020 + 1/406))
```

```
## [1] 10.44009
```

```r
# Upper one-sided p-value
1-pt(10.44,DFE)
```

```
## [1] 0
```

```r
# 2-sided 95% confidence interval
(4.689115 - 4.441623) + c(-1,1)*qt(.975,DFE)*sqrt(0.1632*(1/1020 + 1/406))
```

```
## [1] 0.2010074 0.2939766
```

```r
10^((4.689115 - 4.441623) + c(-1,1)*qt(.975,DFE)*sqrt(0.1632*(1/1020 + 1/406)))
```

```
## [1] 1.588574 1.967780
```

Problem 2

```r
var.test.stat=var(tapply(log10(d$Income2005),d$Educ,mean))
var.test.stat
```

```
## [1] 0.03192832
```

```r
num_sim = 10000   # Draw num_sim randomizations
var.mean<-numeric(num_sim) # storage vector


# generate num_sim random assignments and calculate the variance in means
for(i in 1:num_sim)
    {
    grp<-sample(d$Educ,2584,replace=FALSE)
    Mean.sim = tapply(log10(d$Income2005),grp,mean)
    var.mean[i]<- var(Mean.sim)
    }

# Get the p-value wrt var.test.stat
sum(var.mean>=var.test.stat)/num_sim
```

```
## [1] 0
```

```r
hist(var.mean,prob=T,main="Hist. of Var(means)",
     xlab="Var(means)") # a density histogram
abline(v=var.test.stat,lwd=3) # puts a vertical line at the observed variance
```

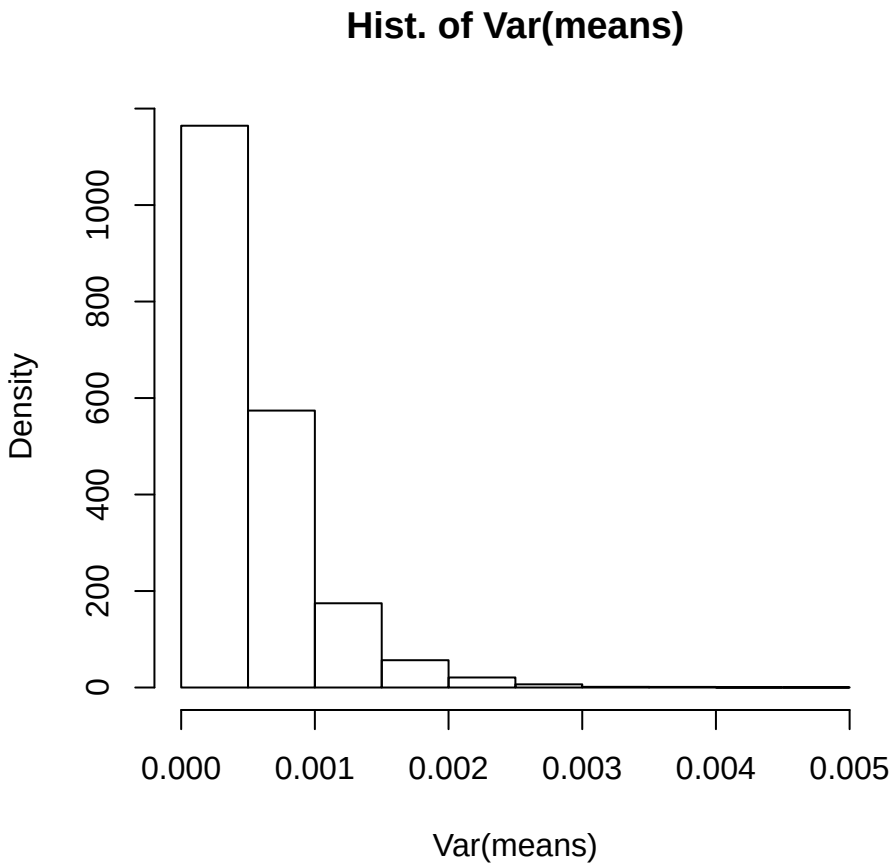


Figure 1: Simulation results: an approximation to the permutation distribution for the test statistic equal to the variance of the means.

Problem 3

```
kruskal.test(log10(Income2005) ~ Educ,data=d)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  log10(Income2005) by Educ
## Kruskal-Wallis chi-squared = 349.45, df = 4, p-value < 2.2e-16
```

```
kruskal.test(Income2005 ~ Educ,data=d)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Income2005 by Educ
## Kruskal-Wallis chi-squared = 349.45, df = 4, p-value < 2.2e-16
```

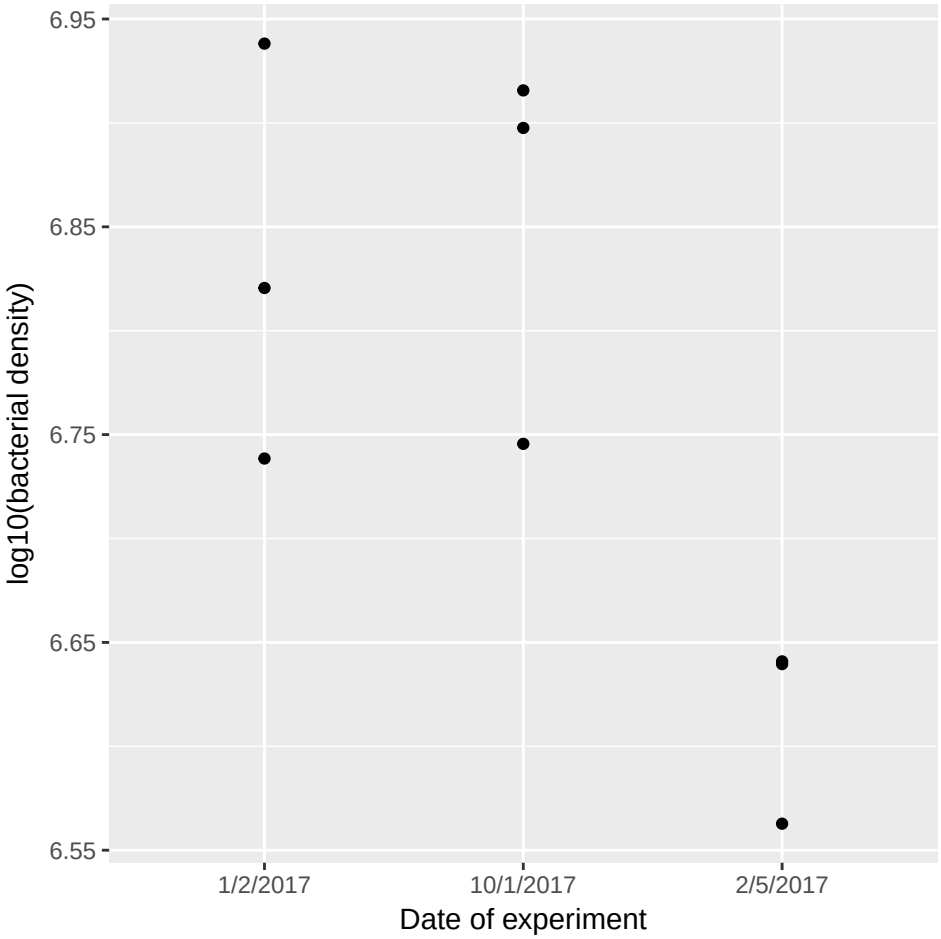


Figure 2: An individual value plot of log-transformed bacterial biofilm densities across 3 experiments.

Problem 4

```
b = read.csv("http://www.math.montana.edu/parker/courses/STAT411/HW6_biofilmdata.csv")
```

```
qplot(b$experiment,log10(b$Number),xlab="Date of experiment",ylab="log10(bacterial density)")
```

```
m.bio = lmer(log10(Number) ~ (1|experiment),data=b)
summary(m.bio)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: log10(Number) ~ (1 | experiment)
##    Data: b
##
## REML criterion at convergence: -10.8
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.15224 -0.86187  0.06062  0.67340  1.37469
##
## Random effects:
```

```
##  Groups     Name        Variance Std.Dev.
##  experiment (Intercept) 0.015163 0.12314
##  Residual               0.006933 0.08326
## Number of obs: 9, groups:  experiment, 3
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  6.76658    0.07632   88.66
```

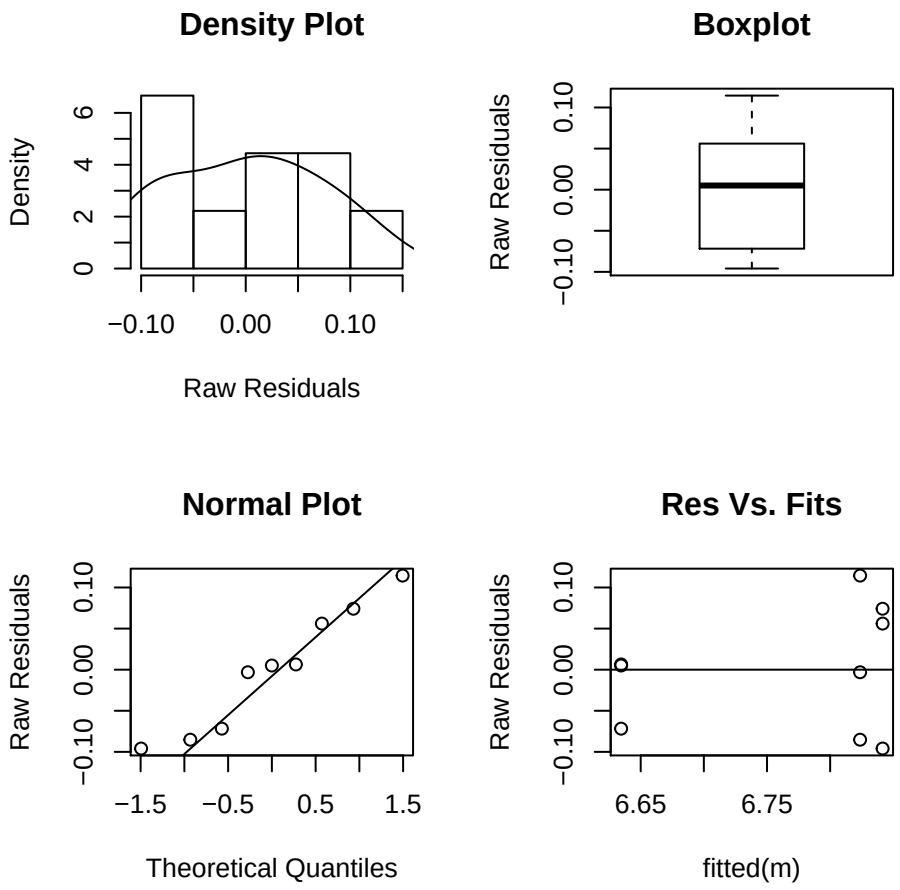


Figure 3: Residal plots for the random effects ANOVA fit to the biofilm data

```
diagANOVA(m.bio)
```

```
# Correlation
0.015163/(0.015163 + 0.006933)
```

```
## [1] 0.6862328
```

```
# 95% CI for true mean log density
6.76658 + c(-1,1)*qt(.975,2)*0.07632
```

```
## [1] 6.438202 7.094958
```

```
# 95% CI for true median density
10^(6.76658 + c(-1,1)*qt(.975,2)*0.07632)
```

```
## [1]  2742847 12443956
```