**STATISTICS 411 MIDTERM EXAM 1** Name KEY

*Fall 2017, 200 pts*

1. **TRUE/FALSE**: (70 pts, 5 pts each) For each of the problems below, circle **T** for True or **F** for False.

   (a) **T** / **F** If $[20, 30]$ is a 99% confidence interval for $\mu$, then $\mu$ will be between 20 and 30 in 99% of all random samples.

   (b) **T** / **F** The $t$ distribution and the standard normal distribution are both symmetric, bell-shaped, and centered at 0.

   (c) **T** / **F** If a 95% confidence interval for $\mu_1 - \mu_2$ is [-21.2, -17.3], then, with 95% confidence, $\mu_2$ is larger than $\mu_1$ by between 17.3 and 21.2.

   (d) **T** / **F** You should use an un-pooled two-sample $t$-test when you are not sure if both populations have the same variance.

   (e) **T** / **F** If the sample size is small and if the data are normal then one should consider a Box-Cox transform of the data before performing a $t$-test.

   (f) **T** / **F** A statistician determines that the correct Box-Cox power transform for a data set corresponds to $\lambda = \frac{1}{2}$. The correct transform to use is $Y = X^{-\frac{1}{2}}$.

   (g) **T** / **F** In a normal probability plot, when performing the correlation test for normality, a small test correlation value is evidence that the data is not normal.

   (h) **T** / **F** If an investigator fails to reject $H_0$, then the data provide strong support for $H_0$.

   (i) **T** / **F** Vitamin D is measured in 9 individuals. After a treatment, these 9 individuals have Vitamin D measured again. A matched pairs $t$-test is appropriate for determining if Vitamin D is higher, on the average, after treatment.

   (j) **T** / **F** A random effects ANOVA is robust to deviations of the constant variance assumption.

   (k) **T** / **F** Kruskal Wallis ANOVA is resistant to the influence of outliers.

   (l) **T** / **F** The $p$-value for the $F$-test in ANOVA is two times the area to the right of the calculated F statistic under the appropriate F distribution curve.

   (m) **T** / **F** A statistician $\log_{10}$-transforms a data set of insect counts to a symmetric distribution and determines there is a difference in the means of the $\log_{10}$-transformed counts between two different species ($p$-value = 0.012). This is the same $p$-value = 0.012 for testing for a difference in the median counts of the two species.

   (n) **T** / **F** A statistician $\log_{10}$-transforms a data set of insect counts to a symmetric distribution and finds a 95% CI = [2.2, 3.1] for the difference in the means of the log-transformed counts between two different species. Back-transforming, $[10^{2.2}, 10^{3.1}]$, gives a 95% CI for the difference in the median counts of the two species.

2. A toothpaste manufacturer randomly assigns 60 guinea pigs to ~~three~~ receive one of 3 different Vitamin C doses (0.5, 1 or 2mg/day) The tooth length (mm) of each guinea pig was measured at the end of the study. The main hypothesis of interest was to determine whether teeth are longer on average for the guinea pigs that receive the highest dose compared to the lowest dose. Most of the R output:

```
Mean=tapply(ToothGrowth$len,ToothGrowth$dose,mean)
SD=tapply(ToothGrowth$len,ToothGrowth$dose,sd)
n=tapply(ToothGrowth$len,ToothGrowth$dose,length)
cbind(Mean,SD,n)
##       Mean       SD  n
##0.5 ______ 4.499763 20
##1   ______ 4.415436 20
##2   ______ 3.774150 20

m.tooth = lm(len ~ as.factor(dose),data=ToothGrowth)
anova(m.tooth)
##                Df Sum Sq Mean Sq F value    Pr(>F)
##as.factor(dose) __ 2426.4  1213.2  ______ 9.533e-16 ***
##Residuals       __ 1025.8    18.0

summary(m.tooth)
##                 Estimate Std. Error t value Pr(>|t|)
##(Intercept)       10.6050     0.9486  11.180 5.39e-16 ***
##as.factor(dose)1   9.1300     1.3415   6.806 6.70e-09 ***
##as.factor(dose)2  15.4950     1.3415  11.551  < 2e-16 ***

TukeyHSD(aov(len ~ as.factor(dose),data=ToothGrowth))
##         diff       lwr       upr    p adj
##1-0.5  9.130  5.901805 12.358195 0.00e+00
##2-0.5 15.495 12.266805 18.723195 0.00e+00
##2-1    6.365  3.136805  9.593195 4.25e-05
```

(a) (8 pts) Calculate the value of the $F$-statistic that is missing from the ANOVA table above. SHOW YOUR WORK!

$$F = \frac{EMS}{MSE} = \frac{1213.2}{18} = 67.4$$

(b) (8 pts) Calculate the degrees of freedom of the "full model" that is missing from the ANOVA table above. SHOW YOUR WORK!

$$DFF = n - I = 60 - 3 = 57$$

(c) (28 pts) Assess the assumptions for this ANOVA as well as you can from the output above.

RS from each group - unlikely - these guinea pigs likely all were procured from the same company or lab.

Ind among groups - yes, guinea pigs were randomly assigned ⇒ ✓

constant variance - (largest SD = 4.5)/(3.8 = smallest SD) < 2 ✓

normality - cannot check from output above

(d) (20 pts) Test the primary research question are teeth longer on average for the guinea pigs that receive the highest dose compared to the lowest dose? Cite the test statistic, the degrees of freedom, and the $p$-value from the relevant R-output and state a conclusion in terms of the problem.

The evidence suggests that the Vitamin C caused tooth lengths to be longer, on average (by 15.5mm, $t = 11.6$, DFF = 57, $p < 2 \times 10^{-16}$) compared to the low Vitamin C group.

(e) (20 pts) Give a *Scope of Inference* for this problem.

cause-and-effect conclusion OK because this was a randomized experiment.

Not clear what population of guinea pigs this can be generalized to due to unclear sampling plan.

(f) (10 pts) Which Vitamin C dose, if any, results in the longest teeth on average compared to any other dose? Cite the relevant R-output.

"2 > 1" (6.4mm, $p = 4.3 \times 10^{-5}$) and "2 > ½" (15.5mm, $p = 0$)

⇒ 2mg Vit. C group had longest teeth on average

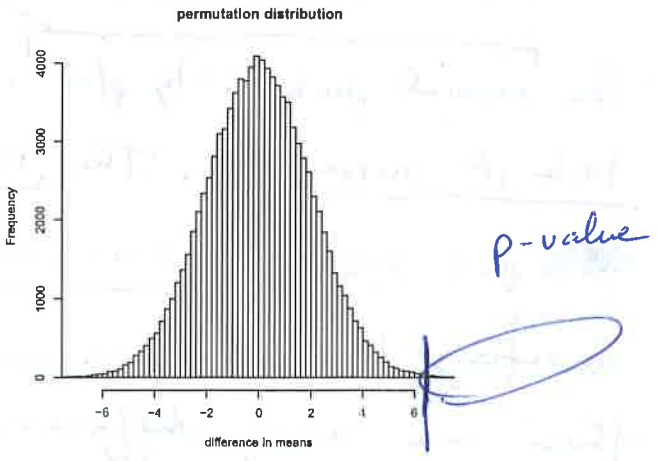
(g) (18 pts) You use a randomization test to assess whether the group with a dose of 2mg/day has longer teeth on average compared to the group with a dose of 1mg/day. The approximate randomization distribution of the test statistic, equal to the mean for the 2mg/day minus the mean of the 1mg/day group, is shown below.

i. Use the R output on the previous page to give the value of the test statistic for this test.

$t = 6.365$

ii. Indicate the value of the test statistic in the plot. ✓

iii. Circle the part of the plot that corresponds to the $p$-value for this test. ✓

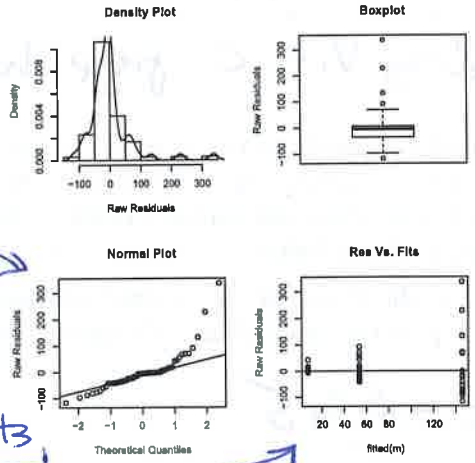


3. (18 pts) A researcher conducts a 6 month experiment growing independent samples of fungi using three different nutrient conditions (A, B and C). The hydrocarbon output of these fungi are measured as parts per million. The researcher performs the following steps in R:

```
d.HC = read.csv("HydrocarbonData.csv")
m.HC = lm(Hydrocarbon ~ Nutrient,data=d.HC)
anova(m.HC)
##Analysis of Variance Table
##
##               Df Sum Sq Mean Sq F value   Pr(>F)
##Nutrient        2 199897   99949  19.874 2.83e-07 ***
##Residuals      57 286665    5029

summary(m.HC)
##Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
##(Intercept)       10.07      15.86   0.635    0.528
##NutrientB         39.70      22.43   1.770    0.082 .
##NutrientC        137.37      22.43   6.125 8.96e-08 ***
##
##Residual standard error: 70.92 on 57 degrees of freedom
##Multiple R-squared:  0.4108,    Adjusted R-squared:  0.3902
##F-statistic: 19.87 on 2 and 57 DF,  p-value: 2.83e-07

diagANOVA(m.HC)
```




The researcher brings these outputs to you and asks for your help.

What do you conclude?

What are the next steps for data analysis?

The normal probability plot indicates lack of normality. The res vs. fits plot indicates non-constant variance. With these deviations to ANOVA assumptions, we fail to conclude there are any differences between mean hydrocarbon outputs because we cannot trust the ANOVA outcomes. Next steps: try a log or other Box-Cox transform of the data then fit ANOVA to transformed data to test for differences in medians.