STAT511 – Exam 2
Fall 2018

Honor Pledge: I have not given, received, or used any unauthorized assistance on this exam.

Signature:____

Printed Name:____

Instructions:

- **Open book, open notes, calculator required. No computers or cell phones.**
- **Time limit is 1 hour 50 minutes - strictly enforced!**
- If an answer is in the computer output, use it; don't calculate it by hand.
- Show your work where appropriate. Put your final answer in the box (if provided).
- Make explanations brief and legible.
- All questions are worth 4 points except where noted. Maximum score is 100.
- Computer input/output is provided at the end of the exam.
- The exam contains a total of 9 pages (including computer input/output).
- If you run out of space, you may use your own paper.

Questions 1 through 3: Consider two analysis scenarios. Both Scenarios have **n = 6** observations per trt and s_W^2 = **MSResid = 3**.
Scenario A: t = 3 treatments, dfResid = 15
Scenario B: t = 8 treatments, dfResid = 40.

1. **Complete the table below by calculating the missing HSD and LSD values.** (**8 pts**) Recall that the LSD value is a 95% unadjusted ME for pairwise comparisons of means. The HSD value is the corresponding 95% Tukey adjusted ME.

	$LSD_{0.05}$	$HSD_{0.05}$
Scenario A t = 3, dfResid = 15	2.13	
Scenario B t = 8, dfResid = 40		3.20

2. Based on your completed table above, using the LSD (unadjusted) method which scenario has higher power? Circle one answer, no need to justify. (**2 pts**)

   Scenario A Scenario B

3. Based on your completed table above, using the HSD (Tukey) method which scenario has higher power? Circle one answer, no need to justify. (**2 pts**)

   Scenario A Scenario B

<u>Questions 4 through 7 (Fish):</u> An investigator is planning a study where **n = 6** fish will be grouped into each tank. Prior to grouping into tanks, a large number of fish are held in a large holding vessel (consider this the population). Assume that in this holding vessel (population), **half of the fish are female.** It is not possible to tell the sex of the fish before grouping them into tanks. Let Y be the random variable representing the number of females in a tank (with a total of n = 6 fish). A partial table of probabilities is shown.

k	0	1	2	3	4	5	6
P(Y = k)	0.016	0.094	0.234	?	?	0.094	0.016

4. Y has a binomial distribution. Give the values of n (size) and π (prob). (**2 pts**)

n:
π:

5. Give the mean of Y. (**2 pts**)

6. What is the probability that exactly 3 fish in a tank will be female? In other words, find P(Y = 3).
NOTE: 0!=1, 1!=1, 2!=2, 3!=6 , 4!=24, 5!=120, 6! = 720.

7. What is the probability that <u>at least</u> 2 of the fish in a tank will be female? In other words, find $P(Y \geq 2)$.

<u>Questions 8 through 9 (Virus1):</u> A study was done to compare two different preparation methods (Prep1, Prep2) for a certain virus. There are n = 11 independent preparations using each method. Let the response (Y) represent the viral abundance.

```
> str(Virus1)
'data.frame':      22 obs. of  2 variables:
 $ Method: Factor w/ 2 levels "Prep1","Prep2": 1 1 1 ..
 $ Y     : int  18 20 9 14 38 26 15 10 25 7 ...
> var.test(Y ~ Method, data = Virus1)

      F test to compare two variances
data:  Y by Method
F = 0.82025, num df = 10, denom df = 10, p-value = 0.7601
alternative hypothesis: true ratio of variances is not eq
ual to 1
95 percent confidence interval:
 0.2206879 3.0487019
sample estimates:
ratio of variances
         0.8202509
```

Virus1 continued…..

8. State the hypotheses for this test. Be specific.

   H0:

   HA:

9. The test above requires the assumption of normality. (**2 pts**)

   TRUE FALSE

__**Questions 10 through 13 (Virus2):**__ Similar to above, suppose a study was done to compare two different preparation methods (Prep1, Prep2) for a certain virus. For this group of questions suppose the methods were compared by applying the virus to leaves of tobacco plants and counting the number of virus spots. A total of n = 22 leaves were used. Let the response (Y) represent the number of viral spots. The R input and output are labeled **Virus2**.

10. State the hypotheses for **Test1**. Be specific.

    H0:

    HA:

11. The hypotheses for **Test2** are the same as the previous question. (**2 pts**)

    TRUE FALSE

12. Suppose that the study had been conducted using n = 11 plants. For each plant, one leaf has Prep1 applied and another leaf has Prep 2 applied. Which test is appropriate for this study design? (**2 pts**)

    Test1 Test2

13. Beyond being appropriate for the design, what is <u>another benefit</u> of using the test that you selected in the previous question?

<u>Questions 14 through 20 (Wheat):</u> A study was conducted to look at the effect of fertilizer on the yield of wheat. Four different treatments (t = 4) were considered: control (C), nitrogen only (N), phosphorous only (P) and both nitrogen and phosphorus (B). Treatments were applied to a total of 16 equally sized plots of wheat (four randomly selected plots per treatment). At the end of the experiment the yield was recorded. The R input and output are labeled **Wheat**. Let μ_i represent the population mean for treatment i. Let σ_i^2 represent the population variance for treatment i.

14. Test H0: $\mu_B = \mu_C = \mu_N = \mu_P$. Provide the test statistic, p-value and conclusion.

    Test Statistic:

    p-value:

    Conclusion:

15. The test above requires the assumption of normality (of the data or residuals). (**2 pts**)

    TRUE FALSE

16. Test H0: $\sigma_B^2 = \sigma_C^2 = \sigma_N^2 = \sigma_P^2$. Provide a p-value and conclusion.

    p-value:

    Conclusion:

17. The test above requires the assumption of normality. (**2 pts**)

    TRUE FALSE

18. The investigator is interested in estimating and testing H_0: $\mu_C - \mu_N = 0$. **2 pts** per question.

    A. Give an estimate of the difference.

    B. Without adjusting for multiple testing, give the p-value for the comparison.

    C. Using Tukey's method, give the p-value for the comparison.

    D. Now suppose Bonferoni's method had been used. What can be said about the Bonferoni adjusted p-value for this comparison?

    $p < 0.05$ $p \geq 0.05$

19. Construct a **lines or CLD display** based on **Tukey's method**. Recall that a CLD display uses lines (or numbers) to indicate which pairs of means are (or are not) significantly different from each other. Hint: Use information from the R output to construct the CLD display.

20. Before the experiment was started, the investigator was interested in testing whether the combined treatment (B) yielded higher mean yield than the average of the two single treatments means (N and P). Provide appropriate contrast coefficients.

B	C	N	P

Questions 21 through 23: Consider an analysis where an investigator is comparing t = 7 treatment means using one-way ANOVA. There are n = 7 observations per treatment, for a total of 7*7 = 49 observations.

21. Suppose that the investigator is interested in testing <u>all pairwise comparisons</u> of means. Name <u>two methods</u> that could be used to control the <u>experiment wise error rate</u> (EER). Just name two methods, no need to justify.

    Method 1:

    Method 2:

22. Considering your answer to the previous question, which of the two methods has <u>higher power</u> (while controlling EER)? No need to justify. (**2 pts**)

23. Suppose that ANOVA assumptions had NOT been satisfied. Give an alternate approach that could be used for the analysis.

Questions 24 and 25: An investigator is interested in <u>planning a study</u> to estimate the proportion of CO voters who support Amendment 74. They will be making inference for a single population proportion using the large sample normal approximation.

24. They would like the 95% ME (margin of error) to be 15% or less (95%ME ≤ 0.15). They conjecture that 30% of voters support Amendment 74. What sample size is required? Give the sample size as an integer value.

25. Suppose they used a conjectured value of 50% (instead of 30%). Would the required sample size be larger or smaller than your answer from the previous question? Just circle one answer, no need to justify. (**2 pts**).

Larger Smaller

<u>Questions 26 through 29:</u> An investigator is interested in making inference for a single proportion (π). Specifically, they want to test H0: $\pi \geq 0.2$ vs HA: $\pi < 0.2$. They are working from a random sample of size n = 40.

```
> prop.test(6, 40, p = 0.2, alternative = "less", correct = FALSE)

      1-sample proportions test without continuity correction
data:  6 out of 40, null probability 0.2
X-squared = 0.625, df = 1, p-value = [redacted]
alternative hypothesis: true p is less than 0.2
95 percent confidence interval:
 0.0000000 0.2647439
sample estimates:
   p
0.15
```

26. Calculate the Z test statistic. Note: This is different from the X-squared statistic in the output. Watch out for the sign! Give your answer to two decimal places.

27. Give the (one-sided) p-value.

28. The <u>sample size is large enough</u> for the test above to be <u>valid</u>. Note: 3*SE = 0.17. (**2 pts**)

TRUE FALSE

29. Regardless of your answer to the previous question, if the sample size was NOT large enough, give an alternative method that could be used. You can either give the name of the method <u>or</u> the name of the R function.

Virus 2 (Questions 10 through 13)

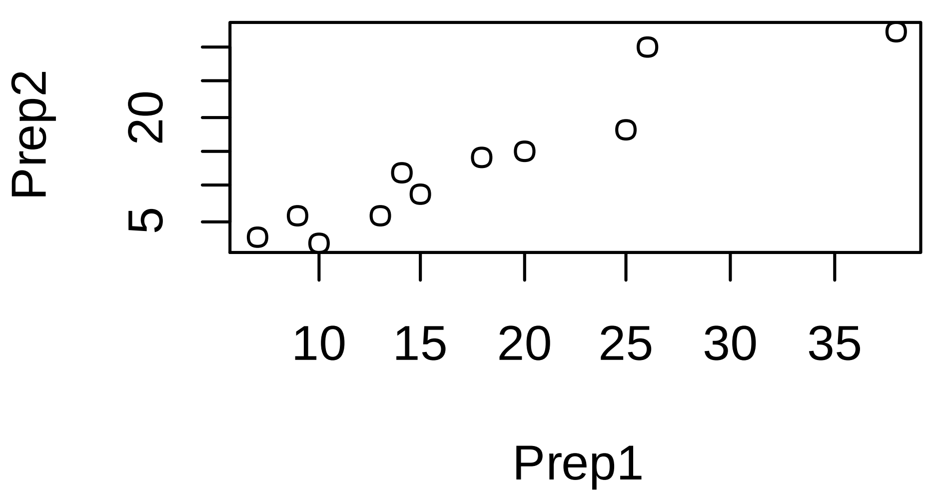
```
> str(Virus2)
'data.frame':	11 obs. of  2 variables:
 $ Prep1: int  18 20 9 14 38 26 15 10 25 7 ...
 $ Prep2: int  14 15 6 12 32 30 9 2 18 3 ...
> #Test1
> t.test(Virus2$Prep1, Virus2$Prep2)

	Welch Two Sample t-test
data:  Virus2$Prep1 and Virus2$Prep2
t = 1.065, df = 19.807, p-value = 0.2997
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.188671 12.915944
sample estimates:
mean of x mean of y
 17.72727  13.36364

> #Test2
> t.test(Virus2$Prep1, Virus2$Prep2, paired = TRUE)

	Paired t-test
data:  Virus2$Prep1 and Virus2$Prep2
t = 4.3529, df = 10, p-value = 0.001437
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 2.129980 6.597293
sample estimates:
mean of the differences
               4.363636

> plot(Prep2 ~ Prep1, data = Virus2)
```

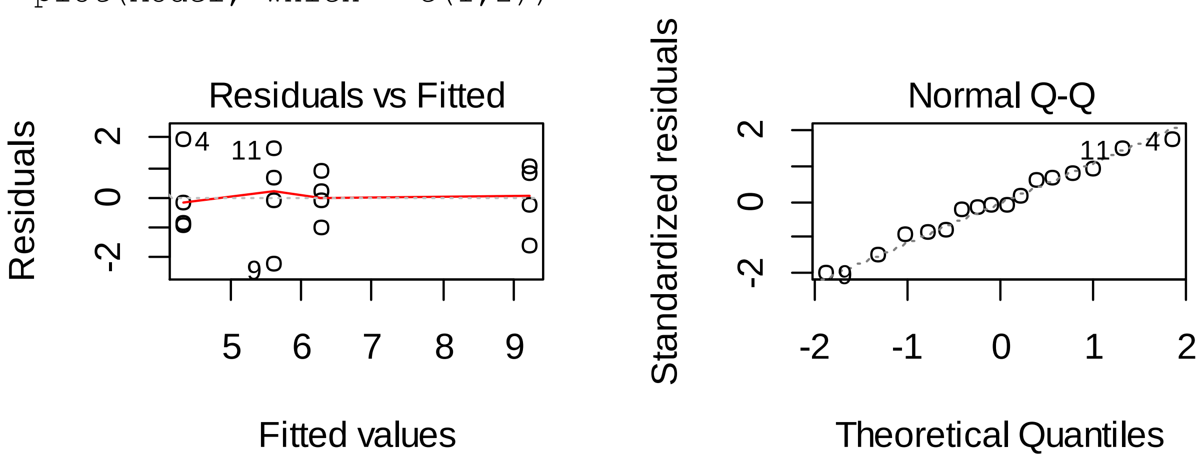


Wheat (Questions 14 through 20)

```
> library(dplyr)
> library(car)
> library(emmeans)
> str(Wheat)
'data.frame':	16 obs. of  2 variables:
 $ Trt  : Factor w/ 4 levels "B","C","N","P": 2 2 2 2 3 3 3 3 4 4 ...
 $ Yield: num  3.5 4.2 3.4 6.3 6.2 5.3 6.5 7.2 3.4 6.3 ...

> SumStats <- summarise(group_by(Wheat, Trt),
+                       n = n(),
+                       mean = mean(Yield),
+                       sd = sd(Yield),
+                       SE = sd/sqrt(n))
> SumStats
# A tibble: 4 x 5
  Trt       n  mean    sd    SE
  <fct> <int> <dbl> <dbl> <dbl>
1 B         4  9.22 1.22  0.609
2 C         4  4.35 1.35  0.674
3 N         4  6.3  0.787 0.394
4 P         4  5.62 1.66  0.828
> Model <- lm(Yield ~ Trt, data = Wheat)
> anova(Model)

Analysis of Variance Table
Response: Yield
          Df Sum Sq Mean Sq F value   Pr(>F)
Trt        3 51.165 17.0550  10.241 0.001255 **
Residuals 12 19.985  1.6654
> par(mfrow = c(1,2))
> plot(Model, which = c(1,2))
```



Wheat continued (Questions 14 through 20)

```
> leveneTest(Yield ~ Trt, data = Wheat)

Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  3  0.3863 0.7649
      12

> shapiro.test(Model$residuals)

      Shapiro-Wilk normality test
data:  Model$residuals
W = 0.98008, p-value = 0.9642

> emout <- emmeans(Model, "Trt")
> emout
 Trt emmean        SE df lower.CL upper.CL
 B    9.225 0.6452551 12  7.81911 10.63089
 C    4.350 0.6452551 12  2.94411  5.75589
 N    6.300 0.6452551 12  4.89411  7.70589
 P    5.625 0.6452551 12  4.21911  7.03089

Confidence level used: 0.95
> pairs(emout, adjust = "none")
 contrast estimate        SE df t.ratio p.value
 B - C       4.875 0.9125285 12   5.342  0.0002
 B - N       2.925 0.9125285 12   3.205  0.0076
 B - P       3.600 0.9125285 12   3.945  0.0019
 C - N      [redacted] 0.9125285 12  -2.137  0.0539
 C - P      -1.275 0.9125285 12  -1.397  0.1876
 N - P       0.675 0.9125285 12   0.740  0.4737

> pairs(emout)
 contrast estimate        SE df t.ratio p.value
 B - C       4.875 0.9125285 12   5.342  0.0009
 B - N       2.925 0.9125285 12   3.205  0.0332
 B - P       3.600 0.9125285 12   3.945  0.0091
 C - N      [redacted] 0.9125285 12  -2.137  0.1965
 C - P      -1.275 0.9125285 12  -1.397  0.5242
 N - P       0.675 0.9125285 12   0.740  0.8792

P value adjustment: tukey method for comparing a family of 4 estimates
```