STAT511 – Exam 1
Fall 2019

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

Signature: KEY

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 the blank area on page 6 or extra paper.

Questions 1 through 4 (Fish 1): It is known that for a particular species of fish, the males have lengths that are normally distributed with mean = 54 mm and standard deviation = 6 mm. In other words, let Y be a random variable representing male fish length and assume $Y \sim N(\mu = 54, \sigma = 6)$. Give your answers to 2 decimal places. **Note:** There is no R output corresponding to this group of questions.

1. What proportion of male fish have lengths of 45 mm or less? In other words, find $P(Y \leq 45)$.

$P(Y \leq 45) = P(Z \leq -1.5) = 0.0668$

$\boxed{0.07}$

2. What proportion of male fish have lengths of greater than 58 mm? In other words, find $P(Y > 58)$.

$P(Y > 58) = 1 - P(Z \leq +0.67)$
$= 1 - 0.7486$
$= 0.2514$ 0.2546 OK

$\boxed{0.25}$

3. What proportion of male fish have lengths between 45 and 58 mm? In other words, find $P(45 < Y \leq 58)$.

-2 for 0.18

$P(45 < Y \leq 58)$
$= P(Z \leq +0.67) - P(Z \leq -1.5)$
$= 0.7486 - 0.0668 = 0.6818$ 0.6786 OK

$\boxed{0.68}$

4. Suppose a random sample of n = 9 male fish is selected from the population. What is the probability that the sample mean is 50 mm or less? In other words, find $P(\bar{Y} \leq 50)$.

-4 for 0.75
-2 for 0.25

Sampling Dist'n of $\bar{Y}$
$\bar{Y} \sim N(\mu = 54, \sigma = 6/\sqrt{9} = 2)$

$P(\bar{Y} \leq 50)$
$= P(Z \leq -2)$
$= 0.0228$

$\boxed{0.02}$

-2 for 0.98

16

Questions 5 through 10 (Fish 2): A CSU researcher is interested in determining whether length can be used to distinguish between male and female fish of a particular species. (It is very difficult to determine the sex of a live fish.) For this study they started with a total of n = 55 randomly selected fish of unknown sex and then measured the **Length** (in mm) and determined the **Sex** (F or M) for each individual. This process identified n = 30 females and n = 25 males. Use $\alpha = 0.05$. The R input and output are labeled **Fish 2.**

5. Using the t.test output, test H_0: $\mu_F - \mu_M = 0$ vs H_A: $\mu_F - \mu_M \neq 0$. Briefly justify your response.

   Conclusion: Reject H_0

   Justification: p-value $= 0.002 < \alpha = 0.05$
   CI does not include zero.

6. A histogram and Shapiro-Wilks test are provided in the output. But, what mistake did the analysist make when evaluating the assumption of normality for the two-sample t-test? ~~Note~~: I am NOT looking "they should have looked at qqplot".

-2 for "data not normal" Checked normality of the combined data.
-4 for violation of equal var, independence, sample size. Instead should check normality of each group separately.

7. The current output shows the Welch-Satterthwaite t-test. But would it have been reasonable to use the pooled t-test? Justify your response based on specific output.

   Yes, $6.47/6.39 = 1.01 < 2$

   This supports the assumption of equal variances.

8. Regardless of your answer to the previous question, we will consider using pooled t-test to test H_0: $\mu_F - \mu_M \geq 0$ vs H_A: $\mu_F - \mu_M < 0$.

   A. (**4 pts**) Calculate the test statistic. Use $s_p = 6.43$. Need to show your work to get credit.

   $$t = \frac{(\bar{y}_F - \bar{y}_M)}{s_p \cdot \sqrt{\frac{1}{n_F} + \frac{1}{n_M}}} = \frac{(48.63 - 54.12)}{6.43\sqrt{1/30 + 1/25}} = -3.153$$

   -2 for -0.85
   -1 for +3.15

   B. (**2 pts**) Calculate df.

   $df = 30 + 25 - 2 = 53$

   C. (**4 pts**) Define the rejection region. Give the numerical table value for full credit. If exact df does not appear in the table, use the closest value.

   -2 for +1.676

   Reject H_0 if $t < -t_\alpha = -1.676$ (using $df = 50$)

   D. (**2 pts**) Make a conclusion for this test.

   Reject H_0.

24

Fish 2 questions continued….

9. Recall that the goal of the study is to determine whether "length can be used to distinguish between male and female fish". Discuss whether the current analysis provides evidence that we can predict sex based on length. Hint: Consider the boxplots.

From the t-test, we have evidence of a difference between means. But this is much more lenient than being able to predict sex. looking at boxplots, there is lots of overlap.

10. The boxplots shows several outliers for the Female fish (with lengths between 60 – 65). Consider (just for this question) if these outliers were removed from the analysis. Just circle one answer for each question, no need to justify.

    A. The (sample) mean for females would be:

    Lower   Higher

    B. The (sample) standard deviation for females would be:

    Lower   Higher

Questions 11 through 18 (Fish 3): For this group of questions we focus just on the n = 25 male fish from the study. Consider this a random sample of male fish. The investigators want to estimate the mean length of male fish from this species. Use $\alpha = 0.05$. The R input and output are labeled **Fish 3**.

11. A p-value (2.2e-16) is shown in the t.test output. What hypotheses are being tested? Be specific!

    H_0: $\mu = 0$

    H_A: $\mu \neq 0$

12. Thinking about your answer to the previous question and considering the study goal, would the t-test results (shown in the output) be of interest to the investigator? Discuss. This question is referring specifically to the hypothesis test results (test statistic, and p-value, ~~etc~~).

NOT of interest.

- Investigator interested in estimation, not testing
- All fish have positive length. Mean will obviously be above zero.

13. Looking at the 95% CI (51.48, 56.76), a colleague expresses concern about this interval "because there are several observations that fall outside of this interval". (This can be seen from the summary statistics or the histogram.) Explain why this is not a problem.

CI provides information about mean.
It does not provide information about individual values.
Hence do not expect CI to include all observations or even 95% of observations.
(I don't really see evidence of "outliers" for Males)

20

Fish 3 questions continued….

14. Briefly explain how you could construct an (approximate) interval that contains 95% of observations. Note: you do not actually have to construct the interval to get full credit.

Empirical Rule ($\bar{y} \pm 2s$)
Chebyshev's Rule
Tolerance Interval
Prediction Interval (OK)

15. We say a confidence interval is "valid" if method assumptions are satisfied. A colleague claims that the 95% CI (51.48, 56.76) is valid because it contains the sample mean (54.12). Is this an appropriate check? Briefly discuss.

NO! CI will always include sample mean ($\bar{y} \pm ME$).

16. Using the histogram and qqplot, discuss whether the assumption of normality is satisfied.

Discussion more important than firm conclusion.
Both histogram and qqplot show evidence of skew, but not too bad.

17. Using the Shapiro-Wilk test, discuss whether the assumption of normality is satisfied.

"Large" p-value supports normality
SW $p = 0.264 > \alpha = 0.05$

18. Give the name of a method that could be used to construct a confidence interval for mean length without assuming normality. Just state the method; no need to justify. **(2 pts)**

Bootstrap.

Questions continue on the next page…..

20

Questions 19 through 23 (Mice): Suppose that for this study, mice were randomly assigned to have blood drawn using one of two procedures. There were n = 9 mice who had blood drawn using procedure A and n = 9 mice who had blood drawn using procedure B. The goal of the study is to test for a difference between means for blood parameters (ex: granulocytes, lymphocytes) for the two procedures. **Note:** There is no R output corresponding to this group of questions.

Questions 19 through 21 (Mice 1): For **granulocytes** (10^3/uL), the following summary statistics were obtained. We will use this information for a power calculation for a larger study comparing the same two procedures. Like the original study, mice will be randomly assigned to one of two groups and the goal is to compare means.

	Procedure A (n = 9)	Procedure B (n = 9)
mean (SE)	1.60 (0.32)	2.30 (0.34)

19. Using the information above, fill in the power.t.test code to calculate power corresponding to a sample size of n = 30 per group (**8 pts**):

```
power.t.test(
n = 30,
sig.level = 0.05,
```

$SE = s/\sqrt{n}$

$s = SE \cdot \sqrt{n}$

(A) `delta =` 2.30 - 1.60 = 0.7

(B) `sd =` ~ 1

(C) `alternative =` (Circle one answer)

`one.sided` `two.sided`

(D) `type=` (Circle one answer)

`one.sample` `two.sample`

20. In the investigator planned to use n = 20 mice per group (instead of n = 30), would the resulting power be higher or lower? (**2 pts**)

Lower Higher

21. If the investigator wanted to power the study to detect a "meaningful" difference of 2 (10^3/uL), would the resulting power be higher or lower? (**2 pts**)

Lower Higher

Questions continue on the next page…..

12

Questions 22 and 23 (Mice 2): For **lymphocytes** (10^3/uL), the following summary statistics were obtained.

	Procedure A (n = 9)	Procedure B (n = 9)
mean (SE)	2.56 (0.63)	4.21 (0.35)
min – max	0.9 – 10.2	2.2 – 6.1

Empirical Rule
$\hat{\sigma} = \frac{10.2 - 0.9}{6} = 1.55$

22. Considering the table above, a colleague comments that the data is "likely not normally distributed". Focusing on Procedure A, what does your colleague see that makes them suspicious about normality? Briefly discuss.

Mean (2.56) not centered between Min, Max (0.9, 10.2)

This is evidence of skew or outlier

-2 for wide range/large SE.

23. Without assuming normality, what method could be used to compare the lymphocyte distributions for the two procedures? Just name the method. Be specific, but no need to justify. **(2 pts)**

Wilcoxon Rank Sum test.

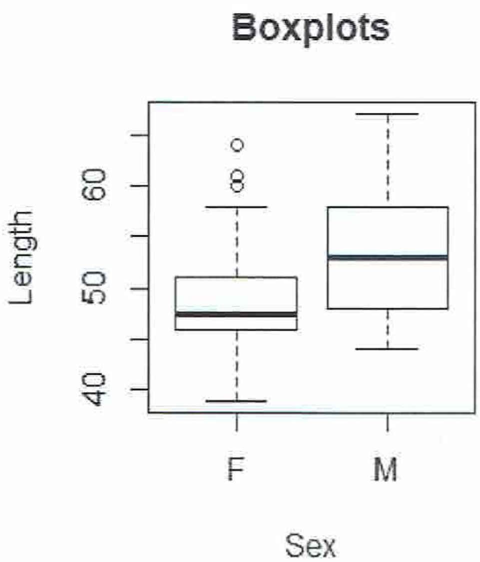
8

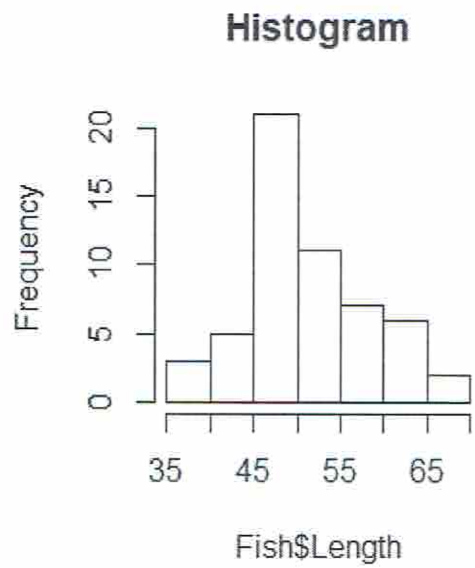
Fish 2 (Questions 5 - 10)

```
library(tidyverse)
str(Fish)

## 'data.frame':    55 obs. of  2 variables:
##  $ Length: int  56 61 48 67 58 58 44 48 64 60 ...
##  $ Sex   : Factor w/ 2 levels "F","M": 2 2 2 2 1 1 2 2 2 1 ...

par(mfrow = c(1, 2))
boxplot(Length ~ Sex, data = Fish, main = "Boxplots")
hist(Fish$Length, main = "Histogram")
```






```
shapiro.test(Fish$Length)

##
##  Shapiro-Wilk normality test
##
## data:  Fish$Length
## W = 0.9541, p-value = 0.03501

SumStats <- summarise(group_by(Fish, Sex),
                      n = n(),
                      mean = mean(Length),
                      sd = sd(Length),
                      min = min(Length),
                      max = max(Length))
SumStats

## # A tibble: 2 x 6
##   Sex       n  mean    sd   min   max
##   <fct> <int> <dbl> <dbl> <int> <int>
## 1 F        30  48.6  6.47    39    64
## 2 M        25  54.1  6.39    44    67
```

Fish 2 continued (Questions 5 - 10)

```
t.test(Length ~ Sex, data = Fish)
```

```
##
##  Welch Two Sample t-test
##
## data:  Length by Sex
## t = -3.1527, df = 51.435, p-value = 0.002699
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.979730 -1.993603
## sample estimates:
## mean in group F mean in group M
##        48.63333        54.12000
```

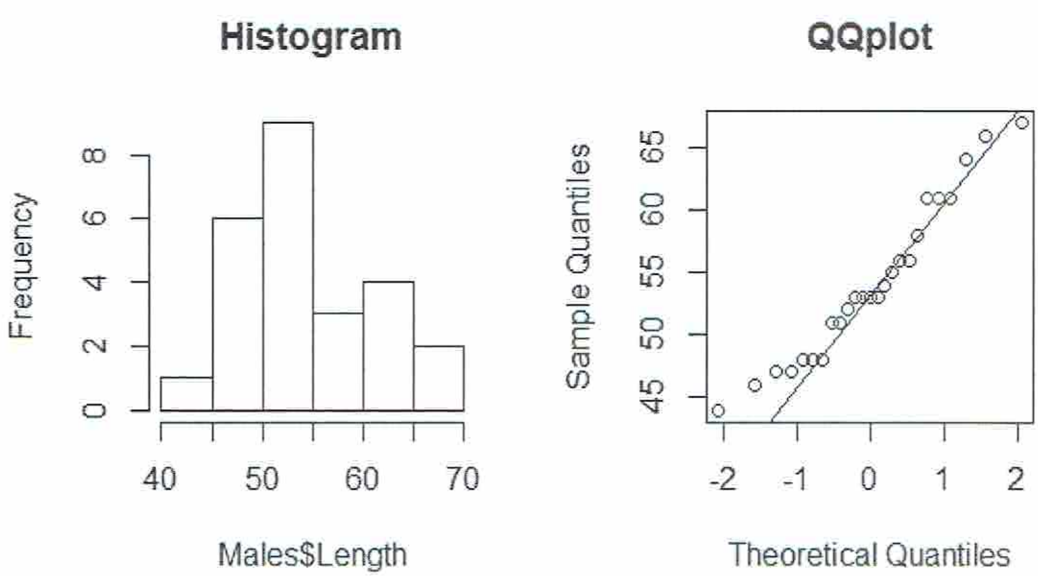
Fish 3 (Questions 11 - 18)

```
str(Males)

## 'data.frame':    25 obs. of  2 variables:
##  $ Length: int  56 61 48 67 44 48 64 56 66 46 ...
##  $ Sex   : Factor w/ 2 levels "F","M": 2 2 2 2 2 2 2 2 2 2 ...
summary(Males$Length)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   44.00   48.00   53.00   54.12   58.00   67.00

hist(Males$Length, main = "Histogram")
qqnorm(Males$Length, main = "QQplot");qqline(Males$Length)
```




```
shapiro.test(Males$Length)

## 
##  Shapiro-Wilk normality test
## 
## data:  Males$Length
## W = 0.951, p-value = 0.264

t.test(Males$Length)

## 
##  One Sample t-test
## 
## data:  Males$Length
## t = 42.333, df = 24, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  51.48144 56.75856
## sample estimates:
## mean of x 
##     54.12
```

Exam 1 Extra Results

This information was NOT provided during the original exam!

#1-4

```
> pnorm(45, mean = 54, sd = 6)
[1] 0.0668072
> 1-pnorm(58, mean = 54, sd = 6)
[1] 0.2524925
> pnorm(58, mean = 54, sd = 6) - pnorm(45, mean = 54, sd = 6)
[1] 0.6807003
> pnorm(50, mean = 54, sd = 6/sqrt(9))
[1] 0.02275013
```

#8

```
> t.test(Length ~ Sex, var.equal = TRUE, data = Fish)

	Two Sample t-test

data:  Length by Sex
t = -3.1493, df = 53, p-value = 0.002688
alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
 -8.981030 -1.992303
sample estimates:
mean in group F mean in group M
       48.63333        54.12000
> qt(0.05, df = 53)
[1] -1.674116
```

#10

```
> Females <- subset(Fish, Sex == "F")
> sort(Females$Length)
 [1] 39 39 39 41 42 45 45 46 46 46 46 46 47 47 47 48 48 48 48 48
[21] 49 50 51 51 56 58 58 60 61 64
> FemaleSS <- subset(Females, Length <= 59)
> mean(FemaleSS$Length)
[1] 47.18519
> sd(FemaleSS$Length)
[1] 4.953919
```

```
#16-18
power.t.test(
   n = 30,
   sig.level = 0.05,
   delta = 0.70,
   alternative = "two.sided",
   type = "two.sample")

power = 0.7599031

power.t.test(
   n = 20,
   sig.level = 0.05,
   delta = 0.70,
   alternative = "two.sided",
   type = "two.sample")

power = 0.5782714

power.t.test(
   n = 30,
   sig.level = 0.05,
   delta = 2,
   alternative = "two.sided",
   type = "two.sample")

power = 1
```