STAT 3011
Spring 2022
Final Exam (A)
Time Limit: 120 Minutes

Name (Print): SOLUTION

Student ID: _____

Instructions:

- Do *not* begin or turn this page until you are instructed.
- Enter all requested information on the top and bottom of this page, and **put your initials on the top of every page**, in case the pages become separated.
- This exam contains 17 pages (including this cover page and the multiple choice answer sheet). Check to see if any pages are missing. There are 21 multiple choice questions including extra point question and 3 short answer problems with sub-parts.
- The exam is closed book. You may *not* use your books, or any wireless device on this exam.
- You may use a calculator and two sheets of paper (size A4 or 8.5" by 11") with formulas or other notes on both sides. You may *not* share calculators or notes!
- Show all your work on each problem for full credit except multiple choice problems. The following rules apply:
  - *Organize your work*, in a reasonably neat and coherent way, in the space provided. Work scattered all over the page without a clear ordering will receive very little credit.
  - *Mysterious or unsupported answers will not receive full credit* for short answer problems. A correct answer, unsupported by calculations, explanation, or algebraic work will not receive full credit; an incorrect answer supported by substantially correct calculations and explanation may still receive partial credit.
  - If you need more space, use the back of the pages; clearly indicate when you have done this.

Honesty Statement and Pledge:

I have not given or received any aid or assistance to or from any other student in this course during the exam period. Everything I have written on this exam represents my own work and knowledge. I sign this knowing that infringements on the University's Academic Honest policy may result in failure or expulsion.

Signed By: _____ **Date:** _____

Problem 1. (61 points) **Multiple Choice**

Choose the ONLY ONE correct answer for each question. Circle your answers to all questions in the answer sheet provided. (NO explanation is needed).

1. (3 points) Which of the following is not quantitative (numeric) variable?
   (A) Exact age of a person
   (B) Number of cars parked at a parking lot.
*** (C) **Zip code of a randomly selected house**
   (D) All of the above are quantitative variable.

2. (3 points) For a family with two children, let A denote first child is male, and let B denote both children are male. What is the probability of $P(B|A)$?
   (A) 0
   (B) 0.25
*** (C) **0.5**
   (D) 1

3. (3 points) Let $P(A) = 0.4$ be the probability that event A occurs, and $P(B) = 0.4$ be the probability that event B occurs. Let $P1 = P(A \cup B)$ be the probability if A and B are disjoint. Let $P2 = P(A \cup B)$ be the probability if A and B are independent. What is the relationship between P1 and P2?
*** (A) $P1 > P2$
   (B) $P1 < P2$
   (C) $P1 = P2$
   (D) None of the above

4. (3 points) Which of the following random variable X is binomial distribution?
   (A) Toss a coin until you get 5 heads, and let X be the number of rolls it takes.
   (B) When waiting for a bus, let X be the waiting time.
*** (C) **Toss a coin 5 times, and let X be the number of tails.**
   (D) Suppose a bag contains 3 red marbles and 3 blue marbles. Choose 2 marbles without replacement and let X be the number that are blue.

5. (3 points) Which one of the following statements about a normal distribution $N(\mu, \sigma)$ is true?
*** (A) **The mean of the normal distribution is same as its median**
   (B) The total area under a normal curve depends on its mean and standard deviation.
   (C) If we decrease μ, the normal curve will be stretched horizontally.
   (D) The IQR of this normal distribution is 2σ

6. (3 points) Which of the following is correct?
   (A) If sampling distribution of sample means is approximately normally distributed, then the population is normally distributed too.
   (B) Sampling distributions can only be constructed for sample mean and sample standard deviation.
   (C) Mean of all possible sample means is smaller than the population mean.
*** (D) **Sampling distribution describes how the values of a statistic vary in all possible samples of same size n from a population.**

7. (3 points) Let Y = the number of hours a student studies for the first 3011 midterm. It is known that $Y \sim N(6, 1.5)$. Which of the following R commands gives the probability that for 100 randomly chosen students, sample mean study time will be less than 6.3 hours?
   (A) `pnorm(6.3,mean=6, sd=1.5)`
*** (B) `pnorm(6.3,mean=6, sd=0.15)`
   (C) `qnorm(6.3,mean=6, sd=1.5)`
   (D) `qnorm(6.3,mean=6, sd=0.15)`

8. (3 points) Suppose a researcher wishes to decrease the width of a confidence interval for μ. To accomplish this goal, they should do:
   i Decrease the sample size
   ii Increase the confidence level
   (A) Both i and ii
   (B) i only
   (C) ii only
*** (D) **Neither i nor ii**

9. (3 points) Which of the following is a correct way to state null hypothesis and alternative hypothesis?
*** (A) $H_0 : p = 0.75$ **against** $H_a : p < 0.75$
   (B) $H_0 : p = 0.75$ against $H_a : p \neq 0.65$
   (C) $H_0 : \hat{p} = 0.75$ against $H_a : \hat{p} < 0.75$
   (D) $H_0 : \hat{p} = 0.75$ against $H_a : \hat{p} \neq 0.65$

10. (3 points) Suppose an independent two-sample t-test with $H_a : \mu_1 > \mu_2$ yields a P-value of 0.026. Which of the following statements is true?
   (A) The test statistic is normal distributed.
   (B) The test statistic is t-distribution with degree-of-freedom $= n_1 + n_2 - 1$
   (C) For the same two samples, if we want to test $H_a : \mu_1 \neq \mu_2$, we will reject null hypothesis at significance level $\alpha = 0.05$.
*** (D) **For the same two samples, if we want to test $H_a : \mu_1 \neq \mu_2$, we will fail to reject null hypothesis at significance level $\alpha = 0.05$.**

11. (3 points) Below is R output to test $H_a : \mu \neq 3.3$ where μ represents population mean. Using the same sample, a 99% confidence interval for μ is found to be (2.67, 3.28). Use the equivalence between hypothesis testing and confidence interval to find the hidden p-value (marked as ******).

```
> t.test(y, mu=3.3, alternative="two.sided", conf.level=0.99)

One Sample t-test

data:  y
t =____, df =____, p-value = ******
alternative hypothesis: true mean is not equal to 3.3
99 percent confidence interval:
 2.67     3.28
sample estimates:
mean of x
 2.975
```

(A) 2.98
(B) 0.99
(C) 0.04
*** (D) **0.008**

12. (3 points) A test to screen for a serious but curable disease is similar to a hypothesis testing, with a null hypothesis of no disease and an alternative hypothesis of disease. If the null hypothesis is rejected, treatment will be given. Otherwise, it will not. Assuming the treatment does not have serious side effects, in this scenario it is better to increase the probability of:

*** (A) **making a Type 1 error, providing treatment when it is not needed.**
(B) making a Type 1 error, not providing treatment when it is needed.
(C) making a Type 2 error, providing treatment when it is not needed.
(D) making a Type 2 error, not providing treatment when it is needed.

13. (3 points) If a instructor wants to test whether the scores from exam 2 improve from exam 1 for all students in intro statistics course. She decide to use randomly select students and collected their exam 1 and 2 scores and conduct a hypothesis test. Assume that everyone took both exams. Which of the following is the appropriate test?

(A) Two sided independent two-sample z-test
*** (B) **One sided matched-pair t-test**.
(C) One sided independent two-sample t-test
(D) Two sided matched-pair t-test

14. (3 points) The following is the incomplete output for an ANOVA table. What is the total sample size N?

```
> summary(aov1)
            Df    Sum Sq    Mean Sq    F value    Pr(>F)
x           ***   60        20         4          ***
Residuals   ***   340       ***
```

(A) 60
(B) 68
*** (C) **72**
(D) 76

15. (3 points) Suppose one ANOVA test yields a P-value of 0.035. Select the correct statement.

(A) We fail to reject null hypothesis at 0.01 level, and we need to use Tukey HSD multiple comparison method to find out which pairs are different.
(B) The p-value is some upper tail probability of an F distribution with degree-of-freedom $= N - 1$.
(C) The null hypothesis for this test is: all groups have the same mean. The alternative hypothesis is: all groups have different means.
*** (D) **We reject null hypothesis at 0.05 level, and we conclude that at least two of the groups have different means.**

16. (3 points) Select the correct statement about the association between two variables.

*** (A) **In the relationship between types of fertilizer and plant growth rate, the response variable is plant growth rate, and the explanatory variable is fertilizer type.**
(B) In the relationship between one's gender and their political party affiliation, the response variable is gender, and the explanatory variable is party affiliation.
(C) In the relationship between weather and COVID spread rate, the response variable is weather, and the explanatory variable is COVID spread rate.
(D) None of the above.

17. (3 points) A small P-value in the Chi-squared test means:

(A) Strong dependence/association between two categorical variables.
(B) Weak dependence/association between two categorical variables.
*** (C) **Existence of statistically significant dependence/association between two categorical variables.**
(D) None of the above

18. (3 points) A Chi-squared test is conducted to test whether there is an association between marital status(single, married, divorced, widowed, or separated)and smoking (smoke or not) status. Suppose the test statistic is 25.1. Which one of the following R command calculates the correct p-value?

*** (A) **pchisq(25.1, df = 4, lower.tail = FALSE)**
(B) pchisq(25.1, df = 4, lower.tail = TRUE)
(C) pchisq(25.1, df = 10, lower.tail = TRUE)
(D) pchisq(25.1, df = 1, lower.tail = FALSE)

For Problem 19 and 20 : Use the following description and R output.

We are interested in exploring the relationship between box office sales and the cost of production. Suppose the assumptions for linear regression are satisfied. We use R to fit a linear regression model:

$$\text{Box} = \alpha + \beta \text{Prod} + \epsilon.$$

```
> myfit <- lm(Box ~ Prod)
> summary(myfit)
Coefficients:
            Estimate    Std. Error    t value    Pr(>|t|)
(Intercept) 15.513      11.603        1.337      0.217989
Prod        7.978       *****         6.522      0.000184

Residual standard error: 14.26 on 8 degrees of freedom
Multiple R-squared: 0.8417, Adjusted R-squared: 0.8219
```

19. (3 points) What is the sample size n?

(A) 7
(B) 8
(C) 9
*** (D) **10**

20. (3 points) What is the standard error of the estimator of β?

*** (A) **1.22**
(B) 1.46
(C) 1.94
(D) 4.88

21. (1 point) Did you circle your multiple choice answers on page 17?

*** (A) **No, but I will now.**
*** (B) **Yes.**

Problem 2. (12 points) Be sure to show all work for full credit.

A college student wants to investigate how effective washing with soap is in eliminating bacteria. To do so, she tested four different methods - (i) washing with water only, (ii) washing with regular soap, (iii) washing with antibacterial soap (ABS), and (iv) spraying hands with antibacterial spray (AS) (containing 65% ethanol as an active ingredient).
She suspected that the number of bacteria on her hands before washing might vary considerably from day to day. To help even out the effects of those changes, she generated random numbers to determine each method to use to wash her hands.
Each morning, she washed her hands according to the treatment randomly chosen. Then she placed her right hand on a sterile media plate designed to encourage bacteria growth. She incubated each plate for 2 days at 36°C, after which she counted the bacteria colonies. She replicated this procedure 8 times for each of the four treatments.

Use the following R outputs to answer questions.

```
> names(hand.washing)
[1] "Bacterial.Counts" "Method"
> summary(aov(Bacterial.Counts~Method, data=hand.washing))
            Df Sum Sq Mean Sq F value  Pr(>F)
Method       3  29882    9961   7.064 0.00111 **
Residuals   28  39484    1410
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

1. (2 points) State assumptions for ANOVA F-test.

random independent sample, normal population distribution , Constant variance assumptions

2. (2 points) State the null and alternative hypothesis of ANOVA test using statistical notations. Define parameters of interest.

0.5 pts $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$
1 point H_1 **: at least two means are different. (or not all means are the same) where**
0.5 pts μ_1 **: population mean bacterial counts for washing hands with water only,** μ_2 **for using regular soap,** μ_3 **for antibacterial soap,** μ_4 **antibacterial spray.**

Copy of R output from the previous page

```
> names(hand.washing)
[1] "Bacterial.Counts" "Method"
> summary(aov(Bacterial.Counts~Method, data=hand.washing))
            Df Sum Sq Mean Sq F value  Pr(>F)
Method       3  29882    9961   7.064 0.00111 **
Residuals   28  39484    1410
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

3. (4 points) Identify test statistic and P-value. Draw conclusion with $\alpha = 0.05$ in context of the problem.

test statistics : 7.064
P-value is 0.0011
conclusion: P-value is less than α, so we reject the null hypothesis. At least two mean bacterial counts are different depending on hand washing method.

4. (3 points) Below is the Tukey's Honest Significant difference of the same data set.

```
   > TukeyHSD(aov(Bacterial.Counts~Method, data=hand))
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Bacterial.Counts ~ Method, data = hand)

$Method
                                    diff        lwr       upr     p adj
Antibacterial Soap-Alcohol Spray 55.0   3.735849 106.26415 0.0319648
Soap-Alcohol Spray               68.5  17.235849 119.76415 0.0055672
Water-Alcohol Spray              79.5  28.235849 130.76415 0.0012122
Soap-Antibacterial Soap          13.5 -37.764151  64.76415 0.8886944
Water-Antibacterial Soap         24.5 -26.764151  75.76415 0.5675942
Water-Soap                       11.0 -40.264151  62.26415 0.9355196
```

   a) According to this Tukey's Honest Significant difference method, which pairs of methods of hand washing are significantly different at $\alpha = 0.05$?

   b) Pick ONE pair with significant difference, and interpret its confidence interval in context of the problem.

2 points for identifying all three pairs
a) Antibacterial Soup (ABS) and Alcohol Spray (AP), Soap and Alcohol Spray, Water and Alcohol Spray have significant difference in the number of bacterial counts.
b) 1 pt for interpretation one For ABS - AP : (3.7, 106.3). We are confident that when we use Antibacterial soup to wash our hands, the average number of bacteria counts is between 3.7 and 106 more than when we use Alcohol Spray.
For Soap - Alcohol Spray (AP) : We are confident that when we use AP, the average number of bacterial counts is between 12.2 and 119.8 less than when we use regular soap.
For Water- Alcohol Spray (AP): We are confident that when we use AP, the average number of bacterial counts is between 28 and 130 less than when we use water to wash hands.

5. (1 point) Based on the previous question part b), which method is the most effective in minimizing bacteria counts? Do you think you can draw a cause-and-effect relationship from this study? Explain.

Alcohol Spray is the most effective. Yes, we can draw a casual relationship because it was a randomized experiment.

Problem 3. (8 points) Be sure to show all work for full credit.

A study looked at the severity of side effects after getting a flu shot in the initial trial stage. The following table shows the counts for the subjects randomized to the group that received the actual shot and the placebo group.

	Severity			
Group	Mild	Moderate	Severe	Total
Active	60	30	10	100
Placebo	50	20	5	75
Total	110	50	15	175

1. (2 points) Calculate the 4 missing expected counts for (1)-(4) below. Round your answers to the nearest integer.

	Severity		
Group	**Mild**	**Moderate**	**Severe**
Active	**63**	**28**	**9**
Placebo	**47**	**22**	**6**

2. (2 points) a) Based on your answer from 1, write down the formula for calculating the test statistic with correct values plugged in. You don't need to calculate the final value but just need to set up the equation.
b) State the degrees of freedom of test statistic.

Test statistic is $(60-63)^2/63 + (30-29)^2/29 + (10-9)^2/9 + (50-47)^2/47 + (20-22)^2/22 + (5-6)^2/6 = 0.828$**, df = 2.**

3. (1 point) State the null and alternative hypothesis in the context of the problem for constructing a chi-squared test for independence to see if the severity of side effects is the same in both active and placebo groups.

H_0: Group and severity of side effects are independent.
H_a: Group and severity of side effects are dependent.

4. (3 points) Find the relative risk of having severe side effects among the active group and placebo group. **Interpret** this value based on the problem context.
Below is the copy of table from the previous page.

	Severity			
Group	Mild	Moderate	Severe	Total
Active	60	30	10	100
Placebo	50	20	5	75
Total	110	50	15	175

(10/100)/(5/75) = 1.5
Those who are in the active group are 1.5 times more likely to have severe side effects than those who are in the placebo group.
OR
(5/75)/(10/100) =0.67
Those who are in placebo are 0.67 times likely to have severe side effect than those who are in the active group.

Problem 4. (20 points) Be sure to show all work for full credit.

Data set `cellphone` contains various specs of a random sample of cell phones collected from 2014. Engineers would like to analyze how the weight (measured in grams) of a phone depends on the size of the battery, the heaviest component of a cell phone. Here, the size is measured by the capacity of the battery, which is the amount of energy it can supply on a full charge (measured in milliampere-hours, mAh).

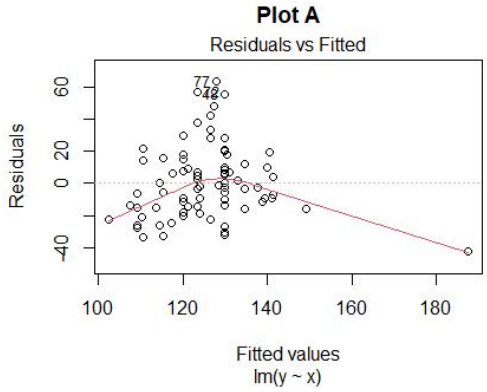
A data set `cellphone` contains the following variable names :

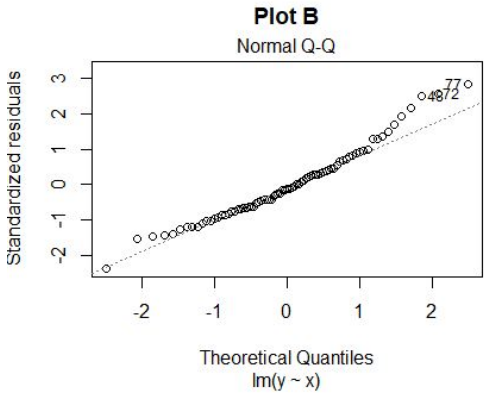
- `Weight.g` : Weight of a cell phone
- `Battery.Capacity.mAh` : Size of battery

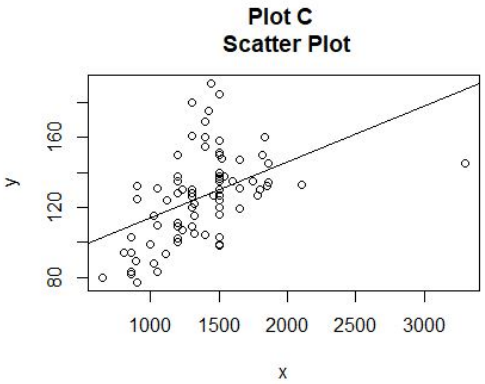
1. (2 points) Identify the explanatory variable and response variable.

1 pt: explanatory variable = Battery size
1 pt: response variable = weight of cell phone

Use Plot A-C to answer sub-problem 5 on page 14.







In the following R output y represents the response variable, and x represents the explanatory variable.

```
> summary(lm(y~x))

Call:
lm(formula = y ~ x)


Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 81.664281   9.632127   8.478 1.22e-12 ***
x            0.032083   0.006745   4.757 9.00e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22.32 on 77 degrees of freedom
Multiple R-squared:  0.2271,Adjusted R-squared:  0.2171
F-statistic: 22.63 on 1 and 77 DF,  p-value: 9.004e-06
```

2. (3 points) Based on R output provided, write down the estimate regression equation. Remember to use correct statistical notations. Interpret the slope *in context of the problem.*

$\hat{y} = 81.66 + 0.032x$
For each unit increase in battery size, the *predicted* weight of a cell phone increase by 0.032 grams.
Or we *estimate* the weight of a cell phone increase by 0.032 grams *on average* when battery size increases by 1 mAh.

3. (4 points) Apple iPhone 5 released in September 2013, has its weight 112 grams and battery size of 1560 mAh. According to the estimated regression equation from the previous question, what is the residual of iPhone 5?

Predicted weight is $81.66 + 0.032(1560) = 131.58$ **(grams)**
Residual is $y - \hat{y} = 112 - 131.58 = -19.58$

Copy of R output from the previous page.

```
> summary(lm(y~x))

Call:
lm(formula = y ~ x)


Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 81.664281   9.632127   8.478 1.22e-12 ***
x            0.032083   0.006745   4.757 9.00e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22.32 on 77 degrees of freedom
Multiple R-squared:  0.2271,Adjusted R-squared:  0.2171
F-statistic: 22.63 on 1 and 77 DF,  p-value: 9.004e-06
```

4. (3 points) Find r^2 and interpret in context of the problem. Round your answer to the 3rd decimal points.

$r^2 = 0.227$ **About 23% of the variability in cell phones weight can be explained its linear relationship with battery size.**

5. (3 points) State all assumptions for linear regression model. Use the plots Figure A - C on page 12 to determine if the data meets one of the assumptions. State which plot you based your answer on.

Linearity : scatter plot: Plot C. Met
Normality : Q-Q plot Plot B. Approximately normal
Constant variance : Residual and fitted plot, Plot A. overall no clear pattern

Copy of R output from the previous page.

```
> summary(lm(y~x))

Call:
lm(formula = y ~ x)


Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 81.664281   9.632127   8.478 1.22e-12 ***
x            0.032083   0.006745   4.757 9.00e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22.32 on 77 degrees of freedom
Multiple R-squared:  0.2271,Adjusted R-squared:  0.2171
F-statistic: 22.63 on 1 and 77 DF,  p-value: 9.004e-06
```

6. (5 points) Use the R output above to test $H_0 : \beta = 0$ against $H_a : \beta \neq 0$ at $\alpha = 0.05$.

   - Test statistic :
   - P-value :
   - Conclusion :

   - **Test statistic = 4.757**
   - **P-value is approximately 0**
   - **We reject the null hypothesis and conclude that there exists linear association between x (battery size) and y (cell phone weight). or population slope of the regression model is not 0.**

This page is left blank intentionally.

Name: _____

Lecture Section:	001	006	011	016
Lecture time:	11:15 am	8:00 am	12:20 pm	10:10 am
(**Circle One**)	Park	Yang	Xu	Shen

Question	Answer			
1	A	B	(C)	D
2	A	B	(C)	D
3	(A)	B	C	D
4	A	B	(C)	D
5	(A)	B	C	D
6	A	B	C	(D)
7	A	(B)	C	D
8	A	B	C	(D)
9	(A)	B	C	D
10	A	B	C	(D)
11	A	B	C	(D)
12	(A)	B	C	D
13	A	(B)	C	D
14	A	B	(C)	D
15	A	B	C	(D)
16	(A)	B	C	D
17	A	B	(C)	D
18	(A)	B	C	D
19	A	B	C	(D)
20	(A)	B	C	D
21	(A)	(B)	(C)	(D)

Please do NOT write in the following table. This is for grading purpose only!

Question	1	2	3	4	Total
Score					
Total	61	12	8	20	101