STAT 3011 **Name (Print):** _____
Fall 2022
Final Exam (A) **Student ID:** _____
Time Limit: 120 Minutes

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 15 pages (including this cover page and the multiple choice answer sheet). Check to see if any pages are missing. There are 17 multiple-choice problems and 3 short-answer problems.
- The exam is closed book. **Do not** use your books, or any electronic devices 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. **Do 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 order 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 explanations 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. (50 points) **Multiple Choice**

Choose ONLY ONE answer for each question. Circle your answers to all questions in the answer sheet provided on page 15. (NO explanation is needed).

1. (3 points) Which of the following is a continuous variable when the measurements are as precise as possible?
   (A) Number of text messages received in a day
   (B) Length of forearm from elbow to wrist
   (C) Population size of a city
   (D) All of the above

2. (3 points) Which of the following is resistant to outliers?
   (A) Median
   (B) Sampling mean
   (C) Average of the largest and smallest value
   (D) Standard deviation

3. (3 points) Toss a fair coin and roll a fair dice once each. What is the probability that the {(coin lands on a head), or (the dice comes up with a 5 or higher), or both}?
   (A) 1 / 6
   (B) 2 / 3
   (C) 7 / 12
   (D) 5 / 6

4. (3 points) Let A and B be any events. Under what condition does the equation hold?

$$P(A \cup B) = P(A) + P(B)$$

   (A) This equation holds when A and B are independent.
   (B) This equation holds when A and B are disjoint.
   (C) This equation never holds.
   (D) This equation always holds.

5. (3 points) We select 5 balls with replacement from a box, which contains 20 red balls, 30 blue balls, and 50 yellow balls. Let A be the event that the first selected ball is red, B be the event that the first selected ball is not blue. Let X be the number of times that the selected ball is not yellow. Select the claim that is correct.
   (A) $P(A \cap B) = 0.14$
   (B) $P(A \cap B) = 0.2$
   (C) $X \sim Bin(100, 0.5)$
   (D) $X \sim Bin(5, 0.7)$

6. (3 points) Suppose we have a random variable $X \sim Bin(3, 0.8)$, what is the probability that X is smaller than 1?
   (A) 0.096
   (B) 0.104
   (C) 0.008
   (D) 0.024

7. (3 points) Let p be the population proportion and $\hat{p}$ be the sample proportion from a random sample of size n. Then the sampling distribution of the sample proportion:
   (A) is approximately normally distributed if $np \geq 15$ and $n(1-p) \geq 15$.
   (B) has its mean equal to $\frac{p}{\sqrt{n}}$.
   (C) becomes very close to normal distribution when n is very small.
   (D) has its standard deviation $\sqrt{np(1-p)}$.

8. (3 points) In a two-sided hypothesis test with $H_a : p \neq 0.3$ and 64 observations, the p-value is 0.3. Which of the following R command produces the plausible value of the test statistic?
   (A) `qt(0.85, df = 63)`
   (B) `qnorm(0.85)`
   (C) `qt(0.3, df = 63)`
   (D) `qnorm(0.3)`

9. (3 points) Which of the following statements is NOT correct?
   (A) If we reject the null hypothesis at significance level 0.05, then we will also reject the null hypothesis at significance level 0.1.
   (B) We should always set up the hypotheses before collecting data.
   (C) When a point estimate and the hypothesized value differ by 0.0001, their difference can still be statistically significant.
   (D) When p-value is greater than the significance level, we accept H_0 and conclude that the null hypothesis is true.

10. (3 points) A study compares the population means sleep hours for freshmen μ_1 and for seniors μ_2 using a 95% confidence interval for $\mu_1 - \mu_2$.
    Choose the best correct answer.
    (A) If the confidence interval is $(0.5, 1.2)$, then it is plausible that $\mu_1 > \mu_2$.
    (B) If the confidence interval is $(-0.5, 1.2)$, then the test of $H_a : \mu_1 \neq \mu_2$ rejects H_0 with significance level of 0.05.
    (C) Both A and B are true.
    (D) Neither A nor B are true.

11. (3 points) Emma measured the weights of 20 people in a program to quit smoking. For each person, she collected the weight at the start and the end of the program. She wants to test if the mean weight change is zero or not.
What is the distribution of the test statistic? Assume that the distribution of difference is approximately normal and its standard deviation is unknown.

(A) Standard normal distribution
(B) T-distribution with degrees of freedom 9
(C) T-distribution with degrees of freedom 10
(D) T-distribution with degrees of freedom 19

12. (3 points) In the ANOVA analysis, the greater the value of the F test statistic,

(A) the larger the P-value is.
(B) the larger the total variance.
(C) the less the sample distributions (i.e. box plots) overlap.
(D) the larger the within-group-variance in comparison to the between-group-variance.

13. (3 points) One of the big factors of an unhappy marriage, Harvard sociology professor Alexandra Killewald found, is the husband's employment status. For the past four decades, she discovered that the estimated relative risk of divorce among those who aren't employed full-time versus those who are employed full-time is 1.32.

Select the correct interpretation of the relative risk of 1.32.

(A) The researcher estimates that husbands' part-time employment status causes divorce.
(B) The researcher estimates that 1.32% of husbands divorce regardless of their employment status in any given year.
(C) The researcher estimates that those husbands who aren't employed full time are 1.32 times more likely to divorce than those who are employed full time.
(D) The researcher estimates that the difference between the risks of divorce for husbands who aren't employed full time and who are employed full time is 1.32%.

14. (3 points) Which of the following is true?

(A) Mean of residuals from a least-squares regression line is always 0.
(B) If there is a very strong positive association between two quantitative variables, then the correlation between them is greater than 1.
(C) Switching the role and x and y doesn't change the estimated least-square regression equation.
(D) A least-squares regression line maximizes the sum of the squared residual values.

15. (3 points) r^2 of a regression model is found to be 0.9. This indicates that:
    (A) a weak association between x and y.
    (B) For each unit increase in x, the predicted y increases by 0.9.
    (C) A strong positive linear association between x and y.
    (D) 90% of the variability in y can be explained by its linear relationship with x.

16. (3 points) Suppose the population regression model is $\mu_x = \alpha + \beta x$ where μ_x is the mean of y given x. Use the following R outputs and choose the statement that is NOT correct.

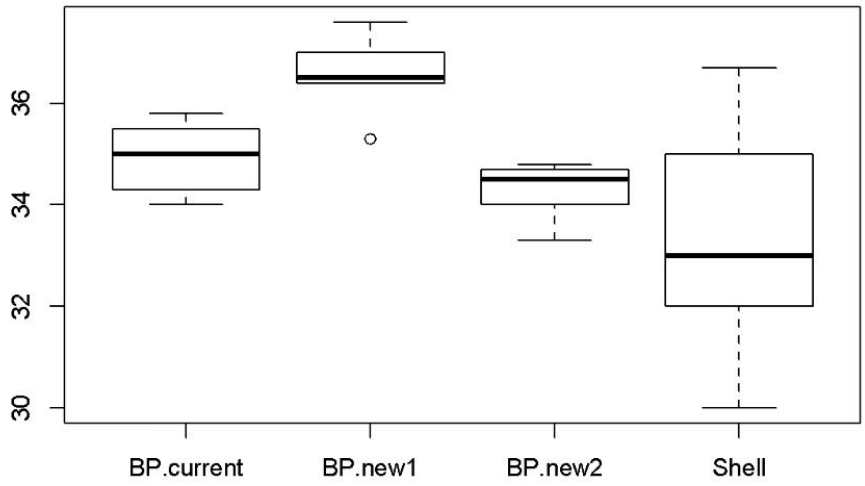
```
> summary(lm(y ~ x))

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1348 -1.5496  0.2734  1.1090  4.5593

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.4674     0.9224  -0.507   0.6150
x             0.5935     0.2658   2.233   0.0308 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.116 on 43 degrees of freedom
Multiple R-squared:  0.1039,Adjusted R-squared:  0.08304
F-statistic: 4.985 on 1 and 43 DF,  p-value: 0.03083
------------------------------------------------------------
> sd(x)
[1] 1.2
```

    (A) Suppose we want to use the data to test $H_0 : \beta = 0$ vs. $H_a : \beta < 0$, then the p-value is 0.9846.
    (B) Let $t_{0.025,43}$ be the t-multiplier used in a 95% confidence interval with degrees of freedom 43, then we have $t_{0.025,43} < 2.233$.
    (C) The sample standard deviation of y is 2.742477.
    (D) Suppose we want to test whether the population mean of x is greater than 0, then the test statistic follows a t distribution with degrees of freedom 44.

17. (2 points) Did you circle multiple choice answers on page 15?
    (A) Yes, I did.
    (B) I will now.

Problem 2. (20 points) Be sure to show all work for full credit.

Chemical engineers at BP wanted to examine the performance of four different types of gasoline. They developed several new additives that they incorporated into the current gasoline to create two new gasolines. They wanted to compare their performance to the current gasoline and one of their competitor's (Shell). The number of miles per gallon (m.p.g.) was used to evaluate each gasoline's performance. Each gasoline type was tested in five different cars. Test at $\alpha = 0.05$ whether there is any difference among the four types of gasoline.



1. (3 points) State the assumptions of ANOVA F-test. Determine whether each assumption is met or not. Briefly explain.

2. (5 points) State the null and alternative hypotheses. Define parameters of interest.

3. (4 points) Fill in the ANOVA table below.

Source	df	SS	MS	F
Delay	(a)	(b)	9.219	(c)
Error	(d)	(e)	(f)	–
Total	(g)	61.54	–	–

4. (4 points) Use the following R command to (i) determine whether the p-value is greater than $\alpha = 0.05$ or less than 0.05. (ii) Briefly explain why. (iii) Draw a conclusion, and (iv) interpret it in context.
(Degrees of freedom in R commands are removed intentionally).

```
 > qf(0.95, df1=_, df2=__)
[1] 3.238872
> qf(0.05, df1=_, df2=__, lower.tail=FALSE)
[1] 3.238872
```

5. (4 points) Determine which pair(s) of means are statistically different, if any, at $\alpha = 0.05$ using Tukey's HSD multiple comparisons. Interpret its confidence interval.

	diff	lwr	upr	p adj
BP.new1-BP.current	1.64	-0.9932264	4.2732264	0.3172654
BP.new2-BP.current	-0.66	-3.2932264	1.9732264	0.8888541
Shell-BP.current	-1.58	-4.2132264	1.0532264	0.3476317
BP.new2-BP.new1	-2.30	-4.9332264	0.3332264	0.0982307
Shell-BP.new1	-3.22	-5.8532264	-0.5867736	0.0141837
Shell-BP.new2	-0.92	-3.5532264	1.7132264	0.7517968

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

A survey by the International Ice Cream Association (ICA) in 2013 was performed to study whether the distribution of favorite ice cream flavors was dependent on gender. No one at ICA knew how to analyze the data, so they sought you out to answer their question since they heard that you just completed Statistics 3011. You decided to answer their question through a chi-squared test at the significance level $\alpha = 0.01$.

The null and alternative hypotheses of the chi-squared test are:
H_0 : Ice cream preference and gender are independent
H_a : Ice cream preference and gender are associated

The contingency table of the top three flavors of ice cream preference and gender is as following:

Observed cell counts		Gender		
		Male	Female	Total
Ice Cream Flavor	Chocolate	304	4,029	4,333
	Vanilla	998	2,471	3469
	Cookie Dough	99	3,069	3,168
	Total	1,401	9,569	10,970

1. (2 points) The partial table of the expected cell counts is below:
   Calculate a) and b). Round your answers to the first decimal place.

Expected cell counts		Gender		
		Male	Female	Total
Ice Cream Flavor	Chocolate	553.4		4,333
	Vanilla	$a)$_____	$b)$_____	3469
	Cookie Dough			3,168
	Total	1,401	9,569	10,970

Copy of tables from page 9.

Observed cell counts		Gender		
		Male	Female	Total
Ice Cream Flavor	Chocolate	304	4,029	4,333
	Vanilla	998	2,471	3469
	Cookie Dough	99	3,069	3,168
	Total	1,401	9,569	10,970

Expected cell counts		Gender		
		Male	Female	Total
Ice Cream Flavor	Chocolate	553.4		4,333
	Vanilla			3469
	Cookie Dough			3,168
	Total	1,401	9,569	10,970

2. (2 points) Write down the formula for computing the test statistic of the chi-squared test from the observed and expected cell counts. Fill in this formula with data from **the single cell, chocolate, and Male, only.** No calculation is needed.

3. (2 points) What is the distribution of the test statistic if H_0 is true?

4. (2 points) The test statistic of this test is 1190.48. What is the p-value for the test statistic? Draw a conclusion for this test and interpret it in the context of the problem.

   The following code may be helpful. (Degrees of freedom are removed intentionally.)

   ```
   pchisq(1190.48,df=*,lower.tail=FALSE) = 3.09e-259
   pf(1190.48,df1=**,df2=**,lower.tail=FALSE) = 0.0204
   ```

Problem 4. (22 points) Be sure to show all work for full credit.

Lego is a line of construction plastic toys that are manufactured by The Lego Group, in Denmark. Lego bricks are joined together by studs on the top, and holes in the bottom of the brick. (Copied from Wikipedia, Simple English).

In this problem, we want to construct a regression model to predict the listed sale price using the number of pieces in a set.
The data set `LEGO` contains two variables `pieces` and `prices` for randomly selected Lego sets. Use the following R outputs and plots to answer questions.

```
> summary(lm(LEGO$price~LEGO$pieces))

Call:
lm(formula = LEGO$price ~ LEGO$pieces)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 40.01025   13.28102   3.013   0.0167 *
    pieces   0.05987    0.01238   4.835   0.0013 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22.75 on 8 degrees of freedom
Multiple R-squared:  0.7451,Adjusted R-squared:  0.7132
```

1. (2 points) Based on the R outputs provided above, what is the sample size? Explain.

2. (3 points) Based on the R outputs provided above, write down the estimate regression equation. Remember to use correct notations.

3. (4 points) Find r and interpret in context of the problem.

4. (4 points) Construct a 95% confidence interval for β. Interpret the result in the context of the problem.
   Use the following R command if needed.

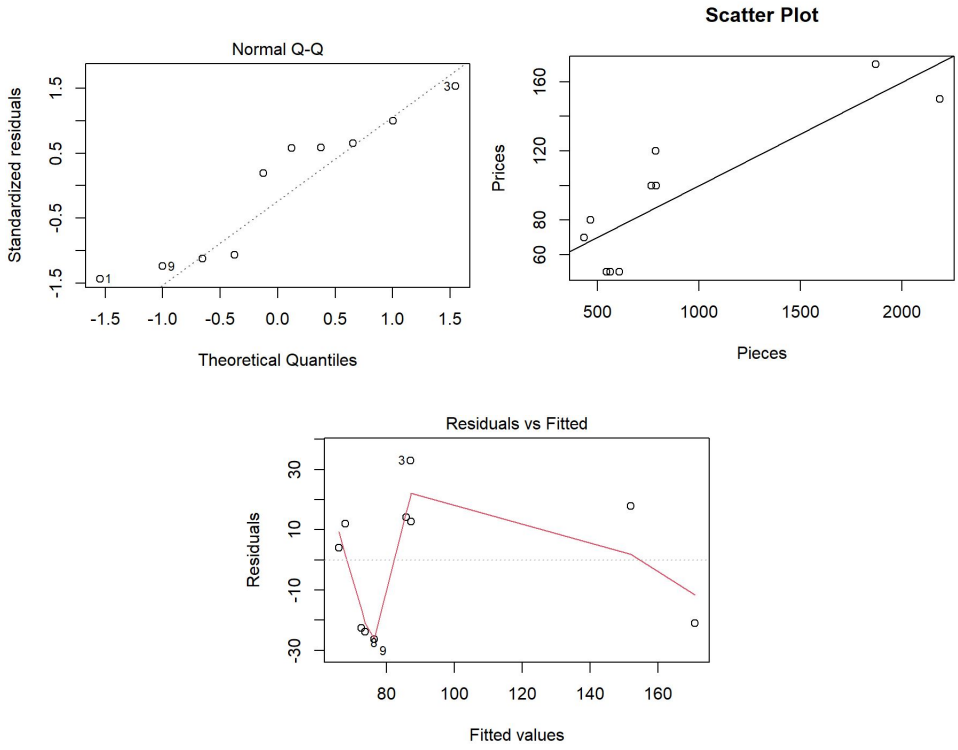
```
> qt(0.025, df=8)
[1] -2.3
```

5. (6 points) Conduct five-step hypothesis test for $\beta \neq 0$. Use diagnostic plots provided on the following page to check assumptions. Use $\alpha = 0.05$.

   - Assumptions :

   - Hypotheses :
   - Test statistic :
   - P-value :
   - Conclusion and interpretation :

Diagnostic plots for Problem 4 Part 5



6. (3 points) A Star Wars-themed Lego set has 10,000 pieces and its listed sale price is $79.99. A student in STAT 3011 calculated the residual of this set based on the sample regression from this problem. She is worried that her prediction suffers from extrapolation.
   (i) Calculate the residual of this set.
   (ii) Comment on whether you agree or not with this student and explain.

Name: _____

Lecture Section:	001	006	0011	016	021
Lecture time:	9:05 am	8:00 am	10:10 am	11:15 am	12:20 pm
(**Circle One**)	Zhang	Yang	Park	Park	Park

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

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

Question	I	II	III	IV	100
Score					
Total					100