STAT3011 **Name (Print):** SOLUTION
Fall 2021
Final Exam (B) **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 17 pages (including this cover page and the multiple choice answer sheet). Check to see if any pages are missing. There are 19 multiple choice questions and 3 short answer problems.
- 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. (55 points) **Multiple Choice**

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

1. (3 points) A study was done to see if GPA is related to where students prefer to sit in a classroom. Students were asked where they prefer to sit in a large class (front, middle, back) and to report their GPA. It was found that the mean GPA was highest for students who prefer to sit in the front of the room and lowest for students who prefer to sit in the back of the room.

   Select true statement(s).

   i) The study is an experimental study.
   ii) Researchers can conclude that sitting in the front of the room causes students to have higher GPA.
   iii) The explanatory variable of this study is where a student prefers to sit.

   (A) Only (i) and (ii) are true
   (B) Only i) is true
   *** (C) **Only iii) is true**
   (D) Only (ii) and (iii) are true

2. (3 points) Which of the following statements is always TRUE if events A and B are disjoint? (Assume $P(A) \neq 0$ and $P(B) \neq 0$.)

   *** (A) $P(A) = P(A \cup B) - P(B)$
   (B) $P(A|B) = P(A)$
   (C) $P(B|A) = P(B)$
   (D) $P(A) > P(B)$

3. (3 points) Consider rolling a fair 6-sided die. Which of the following statements are correct?

   (i) The probability that it does NOT land on a 3 is $\frac{5}{6}$.
   (ii) If we roll this die a large number of times, then for about $\frac{1}{6}$ of the time, it will land on a 4.
   (iii) Suppose the results of the first five rolls were 1, 6, 3, 4, and 2, then the sixth roll must land on a 5, because it is the only outcome that hasn't occurred yet.

   (A) i only
   (B) ii only
   *** (C) **i and ii**
   (D) i, ii and iii

4. (3 points) Suppose that two events A and B are independent and that event B occurs with probability 0.25. The probability that A and B both occur is 0.05. Find the probability that A or B (or both) occur.

(A) 0.7
(B) 0.6
(C) 0.5
*** (D) **0.4**

$P(A \cap B) = P(A)P(B)$ **because of independence.** $P(B) = 0.25$ **and** $P(A \cap B) = 0.05$
Hence $P(A) = 0.2$
Use the general addition rule to find $P(A \cup B)$

5. (3 points) Suppose X follows a binomial distribution $\text{Bin}(n, p)$, and we know that the expected value $E(X) = 300$ and the standard deviation $\sigma = 10$. What is the value of the parameter p?

(A) 1/4
(B) 1/3
(C) 1/2
*** (D) **2/3**

$X \sim binom(n, p)$ **where** $E(X) = np$ **and** $\sigma = \sqrt{np(1-p)}$

6. (3 points) The sampling distribution of a sample mean for a random sample size of 100 describes :

(A) How observations tend to vary from person to person in a random sample of size 100
*** (B) **How sample means tend to vary form random sample to random sample of size 100**
(C) How the data (sample) distribution looks like the population distribution when the sample size is larger than 30
(D) How the standard deviation varies among samples of size 100.

7. (3 points) Suppose X follows a discrete distribution with $x = \{4, 5, 6\}$, the distribution of x is :

What is the standard deviation of the sampling distribution of $\bar{x}$ with sample size $n = 1$?

*** (A) $\sqrt{\frac{2}{3}}$

x	4	5	6
$P(X=x)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

(B) $\frac{2}{3}$
(C) 1
(D) $\frac{\sqrt{2}}{3}$

use $\sigma = \sqrt{\sum (x-\mu)^2 P(X=x)}$ **and** $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$

8. (3 points) Suppose an 95% confidence interval for the mean US family annual income is (\$ 45000, \$ 55000). Which of the following statement is correct?

(A) 95% of all US families have annual income between \$ 45000 and \$ 55000.

(B) 95% of sampled US families have annual income between \$ 45000 and \$ 55000.

(C) 95% of all samples with the same sample size drawn from the same population have their sample means between \$ 45000 and \$ 55000.

*** (D) **If the samples are repeatedly selected and we use the same method for constructing CIs, then around 95% of the resulting CIs will contain the true mean US family annual income.**

9. (3 points) Consider the test $H_0 : \mu = 10$ vs. $H_a : \mu \neq 10$, where the t-statistic, t = 2.46, and the sample size was 43. Then using R, which of the below would compute the correct P-value?

(A) `pt(2.46, df=42)`

(B) `2*pt(2.46, df=42)`

(C) `1-pt(2.46, df=42)`

*** (D) `2*(1-pt(2.46, df=42))`

10. (3 points) A student in STAT 3011 is interested in the academic performance differences between individuals who wake up before 7AM every morning versus those who do not. If the student finds a significant difference, when in fact one does NOT exists in the population, then

*** (A) **a Type I error has been made**

(B) a Type II error has been made

(C) both Type I and Type II errors have been made

(D) neither Type I nor Type II error has been made

11. (3 points) If two-sided confidence intervals that include only positive numbers for $p_1 - p_2$

*** (A) **It is plausible that** $p_1 > p_2$

(B) It is plausible that $p_1 < p_2$

(C) It is plausible that $p_1 = p_2$

(D) All of the above are correct

12. (3 points) What feature(s) about a relationship can be obtained from examining a scatter plot of two quantitative variables x and y?

(A) Whether the strength of the association is strong, moderate or weak.

(B) Whether the relationship is positive, negative or not clear.

(C) Whether the relationship is linear, curved or not clear.

*** (D) **All of the above.**

13. (3 points) Verbal SAT scores were recorded for independent samples of students who intend to major in engineering and students who intend to major in literature. Suppose histograms of both samples show no strong skewness and no outliers. From the data we calculate:

Intended Major	Sample Size	Sample Mean	Sample St. Dev.
(1) Engineering	100	508	12
(2) Literature	100	534	5

Compute the test statistic used to test the null hypothesis $H_0 : \mu_1 = \mu_2$ vs. $H_a : \mu_1 > \mu_2$

*** (A) **-20**

(B) 20

(C) -2

(D) 2

For the next two questions, consider the following information:
The Gunning fog index is a measure of reading difficulty based on the average number of words per sentence and the percentage of words with three or more syllables. High values of the fog index are associated with difficult reading levels. Independent random samples of six advertisements were taken from three different magazines (Scientific American (SA), Fortune (F) and Entertainment Weekly (EW)), and fog indexes were computed for each advertisement. The data are entered into R, you can see the one-way ANOVA table for the analysis. For this problem, you may assume the fog indexes for the ads in each magazine are approximately normal. Use $\alpha = 0.05$

```
            Df Sum Sq Mean Sq F value Pr(>F)
X.magazine   2  48.53  24.264    6.97 0.00723 **
Residuals   15  52.22   3.481
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

14. (3 points) What is the Null hypothesis for this one-way ANOVA problem?

*** (A) $H_0 : \mu_1 = \mu_2 = \mu_3$**, where the μ's represents the fog index means for the three magazines.**

(B) $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$,where the μ_i represents the fog index means for the i-th advertisement for each magazine.

(C) $H_0 : X_1 = X_2 = X_3$, where the X's represents the magazines.

(D) None of above.

15. (3 points) What is the correct conclusion based on the p-value for this ANOVA test?

(A) Fail to reject the null hypothesis, there is no difference.

(B) Reject the null hypothesis, conclude that the means of fog index are different for all magazines.

*** (C) **Reject the null hypothesis, conclude that the means of fog index score is different for at least 2 of the magazines.**

(D) None of above.

16. (3 points) From the following R output, what is the correlation between the response and the explanatory variable?

```
Call:
lm(formula = y ~ x)
Residuals:
Min          1Q      Median      3Q      Max
-29.069     -9.525  -2.272       9.215   43.201
Coefficients:
             Estimate    Std. Error  t value     Pr(>|t|)
(Intercept) ######       6.7584      -2.601      0.0123 *
x            -3.9324     0.4155      -9.464       1.49e-12 ***
---
Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared: 0.6511, Adjusted R-squared: 0.6438
F-statistic: 89.57 on 1 and 48 DF, p-value: 1.49e-12
```

(A) 0.6511
(B) 0.8069
(C) 0.4155
*** (D) **-0.8069**

17. (3 points) The regression of y on x has a prediction equation $\hat{y} = -2.0 + 5.0x$ and a correlation of 0.3. After changing x and y around, the new regression equation

*** (A) **also has a correlation of 0.3**
(B) has a prediction equation $\hat{y} = \frac{1}{5} + \frac{1}{0.2}x$.
(C) All of above.
(D) None of above.

18. (3 points) Which of the following distributions are non-negative?
i. t-distribution with 6 degrees of freedom
ii. Chi-squared distribution with 6 degrees of freedom
iii. F-distribution with 10 and 45 degrees of freedom

(A) i
(B) ii
*** (C) **ii and iii**
(D) i, ii and iii

19. (1 point) Did you circle your multiple choice answers on page 17?

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

Problem 2. (15 points) Be sure to show all work for full credit.

The leaves of certain plants in the genus Albizzia will fold and unfold in various light conditions. We have taken 15 different leaves and subjected them to red light for 3 minutes. The leaves were divided into three groups of five at random. The leaflet angles were then measured 30,45, and 60 minutes after light exposure in the three groups.

Delay (minutes)	Angle (degrees)				
30	140	138	140	138	142
45	140	150	120	128	130
60	118	130	128	118	118

Do mean leaflet angles differ according to how long the delay time is? We will conduct an ANOVA F-test to answer this question. Use $\alpha = 0.05$.

1. (3 points) What were the assumptions on which the ANOVA was based?

Random sample, normal distribution, equal variance.

2. (2 points) Identifying notation, state the **null hypotheses** for conducting an ANOVA with this dataset. Remember to define your notations.

Let μ_1 : mean leaflet angle with 30 min delay, μ_2 : mean leaflet angle with 45 min delay, μ_3 : mean leaflet angle with 60 min delay.
$H_0 : \mu_1 = \mu_2 = \mu_3$

3. (3 points) The incomplete summary of the ANOVA is given in the following:

```
> summary(aov1)
            Df Sum Sq Mean Sq F value Pr(>F)
delay        *  *****   381.1   ***** 0.0119
Residuals   **  697.6    ****
```

Fill in the (a)-(e) of the following ANOVA table.
For (c): calculate to three decimal places.
For (e): calculate to one decimal place.

Source	**df**	**SS**	**MS**	F	p-**value**
Delay	(*a*) 2	(*b*) 762.2	381.1	(*c*) 6.559	0.0119
Error	(*d*) 12	697.6	(*e*) 58.1	–	–
Total	14	1459.8	–	–	–

4. (5 points) **True/False Question**
Researchers performed Tukey HSD at the 0.05 significance level. The R output is given below.

```
      Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = angle ~ delay)

$delay
       diff       lwr       upr     p adj
45-30  -6.0    -18.86489  6.864895 0.4513105
60-30 -17.2    -30.06489 -4.335105 0.0100207
60-45 -11.2    -24.06489  1.664895 0.0908232
```

Determine if the following statements are true or false at $\alpha = 0.05$ level. No explanation needed. (Circle T or F)

a) This Turkey's Honest Significant multiple comparison is unnecessary. (T, (F))

b) Only one pair population means are significantly different. ((T), F)

c) The group with 60 min delay has the largest mean. (T, (F))

d) All the means are significantly different from one another. (T, (F))

e) The delay after exposure does not affect leaflet angle. (T, (F))

F,T, F, F, F

5. (2 points) Do you think you can draw a 'cause-and-effect' relationship from this data? Why or why not?

Yes, because this data is collected through randomized experiment (rather than observational study), we can draw a causal relationship between the delay after exposure and leaflet angle.

Problem 3. (10 points) Be sure to show all work for full credit.

Are the hair and eye color of children's dolls related in any way? To test this using the significance level 0.05, 100 dolls' hair and eye color are observed and summarized below.

		eye color			
		brown	blue	grey	Total
Hair color	light	13	18	9	40
	dark	37	12	11	60
Total		50	30	20	100

1. (2 points) State the null and alternative hypotheses for the test.

H_0 **Eye color and Hair color of children's dolls are independent.**
H_a **Eye color and hair color of children's dolls are NOT independent (or associated/ related)**

2. (1 point) What is the degrees of freedom for this test?

df=(# of rows - 1)(# of column-1) = (2-1)(3-1)=2

3. (3 points) The test statistic for this table equals to 9.29. Use the following R command to find the correct p-value. Draw conclusion and interpret in context of the problem. Remember to use $\alpha = 0.05$.

```
pchisq(9.29, df=*)
[1] 0.9903905
```

p-value is 1-0.99=0.01 is less than $\alpha = 0.05$. We reject the null hypothesis and conclude that children's dolls' eye color and hair color are not independent.

Do not use any information above.

A study was conducted to investigate whether vitamin C has a therapeutic value for treating the common cold. During a 2 week period, each subject from a sample of 279 was randomly assigned to take Vitamin C or a placebo. In 2 weeks, the researcher counted how many have a cold or not.

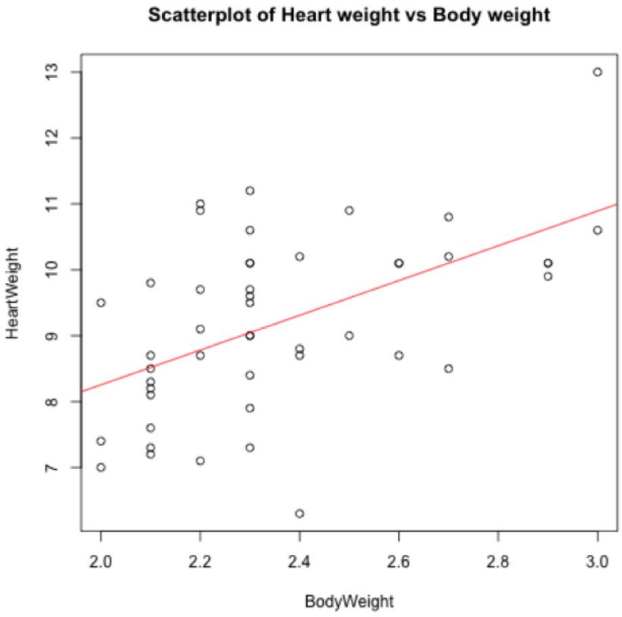
	Cold	No cold	Totals
Placebo	31	109	140
Vitamin C	17	122	139
Totals	48	231	279

4. (4 points) (i) Find the risk of getting a cold for those receive a placebo.
(ii) find the risk of getting a cold for those received Vitamin C. Then,
(iii) calculate and interpret the relative risk of getting a cold for those with no Vitamin C intake (placebo group) and those taking Vitamin C.

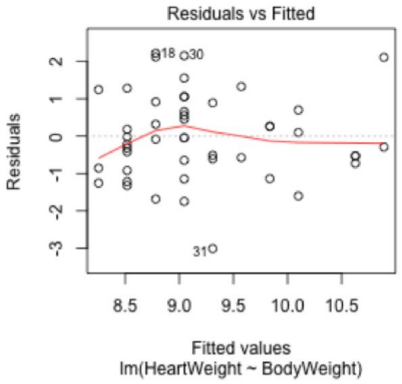
risk of getting a cold for placebo group = 31/140 = 0.22
risk of getting a cold for vitamin c group = 17 / 139 = 0.12
Relative risk = risk of getting a cold for placebo group / risk of getting a cold for vitamin c group = 0.22/ 0.12 = 1.8
Those who don't take vitamin C is 1.8 times more likely to catch a cold than those taking vitamin C. OR
Relative risk = risk of getting a cold for Vitamin C group / risk of getting a cold for placebo group = 0.12 / 0.22 = 0.545
Those who take vitamin C has risk of a cold 0.54 times of the risk for those who don't take vitamin C.

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

A.



B.



C.

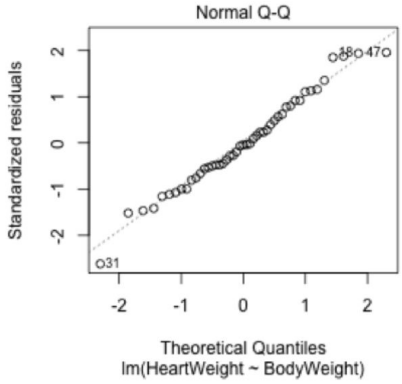


Figure 1: Use to answer Problem 4 Part 4

People analyzed the relationship between the body weight and the heart weight of domestic cats through a linear regression model. There are a total of 47 cats in the sample. For each cat in the sample, the following measurements has been recorded:

- BodyWeight - Body weight in kilograms, and
- HeartWeight - Heart weight in grams.

Part of the data is shown below. We assume that this data represents a random sample from the population of all domestic cats.

Cat#	Body Weight (kg)	Heart Weight (g)
1	2.0	7.0
2	2.0	9.4
...	...	...
47	3.0	13.0
Sample Mean	$\bar{x} = 2.36$	$\bar{y} = 9.20$
Sample Standard Deviation	$s_x = 0.274$	$s_y = 1.358$

The summary of the linear regression model of HeartWeight on BodyWeight is given in the following:

```
Call:
lm(formula = HeartWeight ~ BodyWeight, data = cats)
Residuals:
Min         1Q          Median      3Q      Max
-3.00871    -0.68599    -0.04506    0.79583 2.21858
Coefficients:

            Estimate    Std. Error  t value     Pr(>|t|)
(Intercept) 2.9813      1.4855      2.007       0.050785 .
BodyWeight  2.6364      0.6254      4.215       0.000119 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 1.162 on 45 degrees of freedom
Multiple R-squared: 0.2831,Adjusted R-squared: 0.2671
F-statistic: 17.77 on 1 and 45 DF, p-value: 0.0001186
```

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

Explanatory variable: body weight. Response: heart weight.

2. (3 points) Write down the estimated regression line equation *and* interpret the slope in the context.

Regression line: $\hat{y} = 2.9813 + 2.6364x$**. On average, as the body weight increases 1kg, the heart weight will increase 2.6364g.**

3. (5 points) The body weight of cat #21 was 2.3 kg. Predict the heart weight for this cat. If the actual heart weight of cat #21 was 8.4 g. Find the residual for this cat.

$\hat{y} = 2.9813 + 2.6364 * 2.3 = 9.04$**. Residual:** $y - \hat{y} = -0.64$**g.**

4. (3 points) Use the plots Figure 4 A – C on page 11 to determine if the data meets the linear regression model assumptions. You only need to state and assess ONE of the three assumptions for complete credit.

Linearity: satisfied as the points in the scatter plot A roughly lie along a straight line.
Normality: satisfied as the points in the QQ normal plot C of the residuals lie along a straight line.
Constant variance(standard deviation): satisfied as the magnitudes of the residuals (plot B) do not change with x.

5. (7 points) Is there any evidence to support that the heart weight and the body weight of domestic cats have a **positive** association, assuming all linear regression model assumptions are met? Test it at $\alpha = 0.01$ level.

1. Assumptions: linearity, normality and constant variances.
2. Hypotheses: $H_0 : \beta = 0$ **vs.** $H_a : \beta > 0$**.**
3. test statistic: 4.215.
4. pvalue: $0.000119/2 = 0.00006$**.**
5. conclusion: since the pvalue is less than 0.01, we have evidence that the heart weight and the body weight of domestic cats have a positive association.

Name: _____

Lecture Section:	001	005	009	013	017
Lecture time:	9:05 am	8:00 am	12:20 pm	10:10 am	11:15 am
(**Circle One**)	Zhao	Shen	Xu	Park	Song

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)

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

Question	I	II	III	IV	Total
Score					
Out of	55	15	10	20	100