PH142 Fall 2019 Final Examination SOLUTIONS

Academic Integrity Pledge

I, _____, affirm that I will not plagarize answers from those around me or seek to communicate with others either verbally or non-verbally about the examination.

- Students continuing to write after the examination has finished will recieve a grade of zero.
- Students must keep their student IDs on their desks and must show them to the professor or GSI when submitting their test.

1. [1 point] The only additional information needed to run a one sample z-test instead of a one sample t-test is the population standard deviation.

◯ True
◯ False

```
### SOLUTION: True
```

2. [1 point] Fill in the bubble(s) next to the appropriate null hypothesis for a chi-square test for goodness of fit. More than one may be applicable.

◯ The distribution of weights of the population is the same as in a 1970 study.
◯ Heart disease is independent of anger levels.
◯The mean difference in a subject's weight under Diet 1 or Diet 2 is 0.
◯ $p_1 = p_2 \ldots = p_k$

```
### SOLUTION: (a), (d)
```

3. [2 points] Circle the TWO most helpful figures for determining if my sample is approximately normally distributed (no credit if more than two answers are selected):

◯ Histogram
◯ Pie Chart
◯ Boxplot
◯ Q-Q Plot

```
### SOLUTION: Histogram and Q-Q Plot.
```

4. [1 point] Permutation test and bootstrap sampling require the data to come from a simple random sample with a large sample size.

◯ True
◯ False

```
### SOLUTION: False
```

5. [1 point] The statistic from the Chi-square test for independence follows the Chi-square distribution with $k-1$ degrees of freedom where k is the number of categories.

◯ True
◯ False

```
### SOLUTION: False. df = (number of rows - 1)*(number of cols - 1)]
```

6. [1 point] One condition for running a two sample t-test for the difference in population means is that the standard deviations of both populations are the same.

◯ True
◯ False

```
### SOLUTION: False.
```

7. [1 point] To put bounds on our best guess of the mean response y for a given x value from a line of best fit we calculate a prediction interval.

◯ True
◯ False

```
### SOLUTION: False.
# The bounds of mean response is given by the confidence interval,
# while for a single observation we use the prediction interval.
```

8. [1 point] If a paired t-test has 30 observations in one group and 30 observations in the second group, then the number of degrees of freedom for running the paired t-test is 59.

◯ True
◯ False

```
### SOLUTION: False. The degrees of freedom should be 30-1 = 29
```

9. [1 point] ANOVA's test statistic follows the F distribution with $N_{total} - 1$ in the numerator, and $N_{total} - k_{group}$ in the denominator.

◯ True
◯ False

```
### SOLUTION: False.
```

10. [1 point] Using the same random sample from a population, you compute both a 95% confidence interval and a 99% confidence interval for a population proportion. The 99% confidence interval will always be wider than the 95% confidence interval.

◯ True
◯ False

```
### SOLUTION: True
```

11. [1 point] The underlying distribution of a population has to be Normal for the sampling distribution of the mean to become Normal as sample size becomes larger.

◯ True
◯ False

```
### SOLUTION: False
```

12. [1 point] We can get a p-value of 0 from a permutation test.

◯ True
◯ False

```
### SOLUTION: True
```

13. [1 point] This code will produce the correct p-value from the chi-square distribution for a specified `test_stat` and `degf`:

```
pchisq(q = test_stat, df = degf)
```

◯ True
◯ False

```
### SOLUTION: False. needs lower.tail = F
```

14. [1 point] You want to find the appropriate sample size to obtain a confidence interval with your desired confidence level and margin of error. To do this, you have to input a guessed value of $\hat{p}$, which we label p^*. If we have no best guess for our sample proportion, what is the most conservative guess you can make for p^*?

◯ p* = 0
◯ p* = 0.5
◯ p* = 1
◯ It depends on the desired confidence level, z*
◯ It depends on the desired margin of error, m

```
### SOLUTION: p* = 0.5
```

15. [1 point] Which of the following are true about Tukey's HSD test?

◯ It maintains a 5% *experimentwise* or *"family"* error rate no matter how many tests you conduct.
◯ It overcomes the issue of multiple testing.
◯ It compares all possible pairs of means.
◯ All of the above.

```
### SOLUTION: All of the above
```

Question 16 [3 points total]

It has been claimed that beetroot juice reduces blood pressure. Scientists investigate this question using three different experimental designs. For each design, fill in the circle next to the most appropriate test.

a) [1 point] A group of randomly sampled individuals are randomly assigned to two groups. One group is given beetroot juice everyday for 4 weeks; the other group is given a red drink everyday for 4 weeks. After 4 weeks, the blood pressure for both groups is measured.

◯ one-sample proportion
◯ one-sample t-test
◯ ANOVA
◯ paired t-test
◯ independent two-sample t-test

```
### SOLUTION: independent two-sample t-test
```

b) [1 point] The blood pressure of a group of randomly sampled individuals was measured before the study. Then every day for four weeks they were given beetroot juice and their blood pressure was measured again after the study, the difference was analyzed.

◯ one-sample proportion
◯ one-sample t-test
◯ ANOVA
◯ paired t-test
◯ independent two-sample t-test

```
### SOLUTION: paired t-test
```

c) [1 point] A group of randomly sampled individuals with known high blood pressure was given beetroot juice everyday for 4 weeks. After 4 weeks their blood pressure was measured. The average blood pressure at the end of the study is compared with 140/90mmHg (which is the level where high blood pressure is diagnosed).

◯ one-sample proportion
◯ one-sample t-test
◯ ANOVA
◯ paired t-test
◯ independent two-sample t-test

```
### SOLUTION: one-sample t-test
```

17. [1 point] Fill in the blanks with the type of variable. ANOVA test involves one _____ explanatory variable, and one _____ response variable.

```
### SOLUTION: categorical or grouping; continuous
```

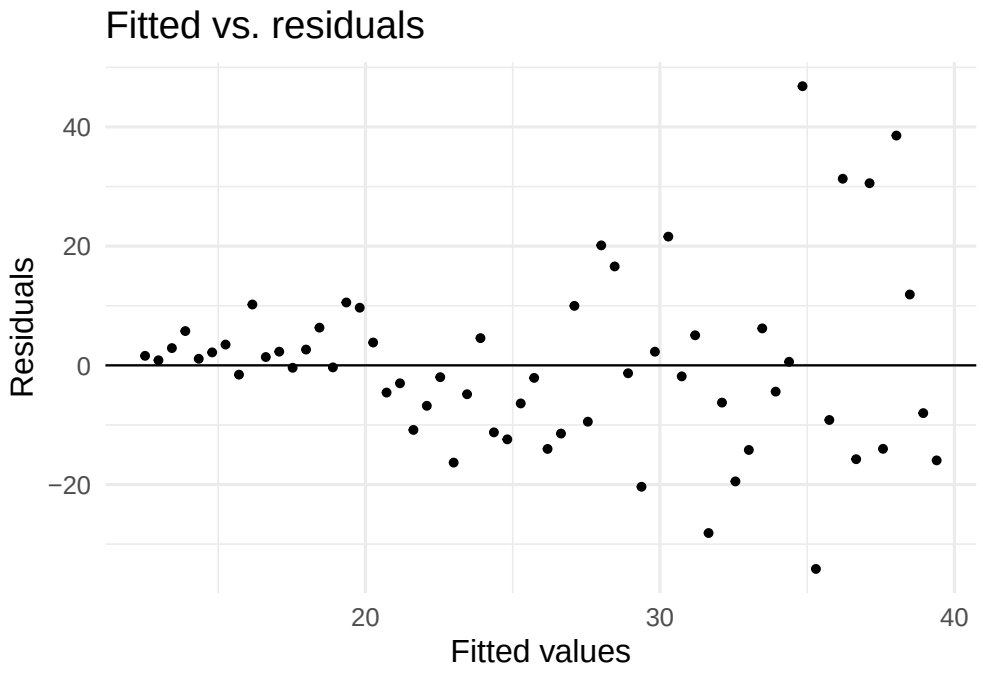
Question 18 [5 points total]

Consider the line of best fit between x and y.

a) What do a 95% confidence interval for the predicted value for a given x and the 95% prediction interval for an individual value given the same x have in common? Which interval is wider? Why? [3 points]

```
### SOLUTION: They share the same center.
# The predication interval is wider.
# Because averages are less variable than individual responses.
```

b) [2 points] Suppose you were presented with the following plot based on a linear regression. What is the pattern in the plot known as? Which assumption of linear regression is most clearly violated from the plot?




```
### SOLUTION: The pattern is known as "fanning out".
# The constant variance assumption about residuals has been violated.
```

Question 19 [5 points total]

A researcher wants to assess the effects of watching news during mealtime on amount of food intake. A random sample of 50 individuals was assigned to the study: on day one, they had to eat a pasta meal while watching a 30-minute news segment; on day two, they had to eat the same pasta meal while watching nothing. Pasta consumption (in grams) during dinner on each day was collected.

a) [1 point] The p-value that is obtained from running an independent, two-sample t-test on this data will most likely be _____ than that of the paired t-test p-value.

◯ equal
◯ smaller than
◯ greater than
◯ cannot say without additional information

```
### SOLUTION: greater than
```

b) [1 point] In one sentence, justify your above response.

```
### SOLUTION: The paired t-test only uses variation within a subject
# and does not use variation between subjects.
```

c) [1 point] This study design can be improved. Provide one change that can be made in order to improve the design. (Do not undo the pairing in the design.)

```
### SOLUTION: Options include: Randomize the exposure given on day one
# instead of giving everybody the same exposure.
# Increase the wash-out period to diminish carry over effects
# (perhaps there was a particularly intense piece of news on day one).
```

d) [1 point] Suppose the mean difference in food consumption between days one and two was -30g and the standard deviation of the differences is 43 g. Calculate the appropriate test statistic. Round your answer to two decimal places.

```
### SOLUTION: $t = \frac{\bar{x_d} - 0}{s_d / \sqrt{n}}
# = \frac{-30 - 0}{43 / \sqrt{50}} = -4.93$
```

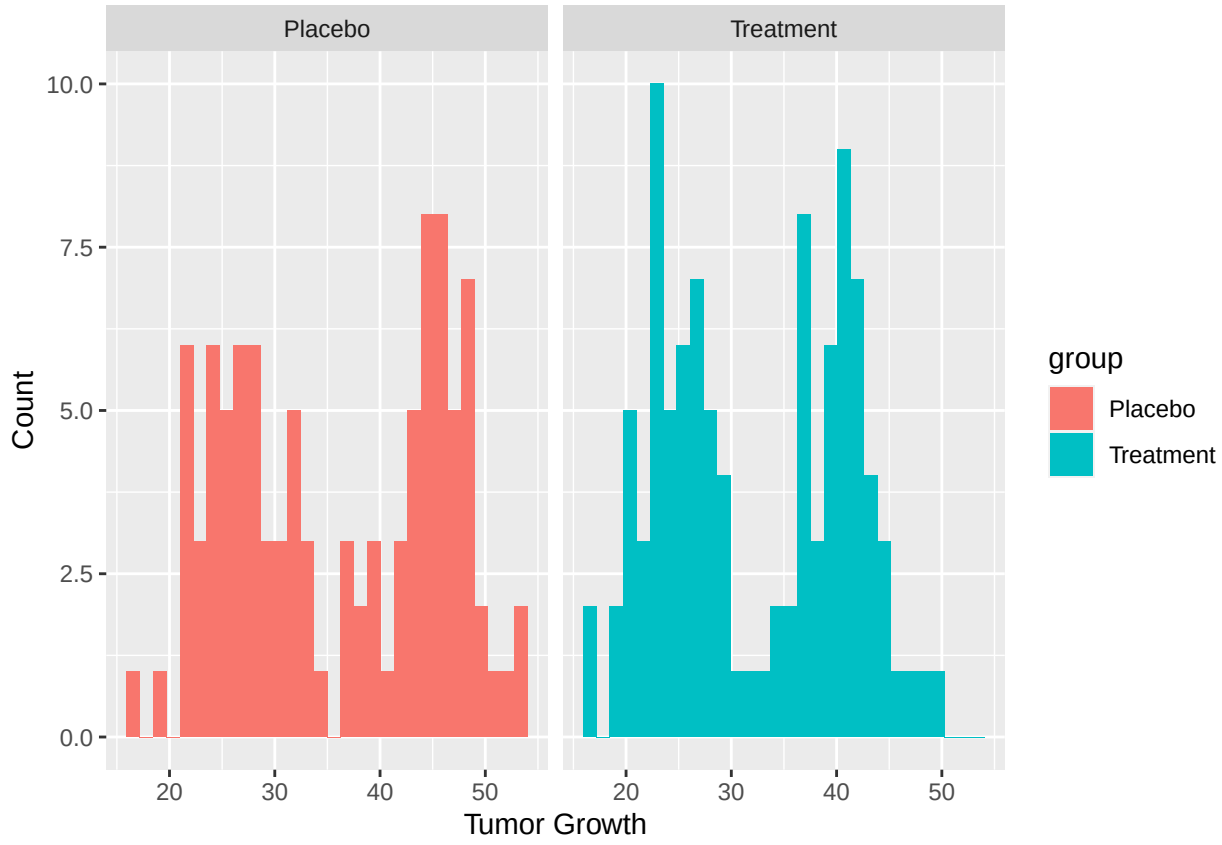
e) [1 point] Write the R code to calculate the two-sided p-value using the test statistic you calculated above.

```
### SOLUTION: 2 * pt(q = -4.93, df = 49, lower.tail = TRUE)
```

Question 20 [9.5 points total]

Pharmaceutical company JoeSchmovartis has developed a drug to slow lung tumor growth. In order to demonstrate the drug's efficacy to the FDA, they run a randomized trial on 200 lung cancer patients, randomly assigning 100 to take a placebo and 100 to test the new drug. Growth in the tumor (in mms) is recorded after two months.

Below is a plot of tumor growth for the placebo group and the treatment group:



a) [1 point] The distribution of tumor growth is:

◯ Approximately normal
◯ Skewed Right
◯ Skewed Left
◯ Bimodal

```
### SOLUTION: Bimodal
```

b) [1 point] Fill in the blank. Even though we have evidence that the population distribution of tumor growth is not normal, since our sample sizes are large, we apply the Central Limit Theorem to conclude the _____ for our two sample t-test is approximately normal.

```
### SOLUTION: sampling distribution
```

We proceed with a two sample t-test for difference of group means to see if the mean tumor growth for the placebo group is less than the mean tumor growth for patients on the drug.

The sample means and standard deviations for both groups are reported below, as well as the degrees of freedom:

```
## # A tibble: 2 x 3
##   group     sample_mean sample_sd
##   <chr>           <dbl>     <dbl>
## 1 Placebo          36.0     10.0
## 2 Treatment        32.2      8.67
```

```
## There are 193.9652 degrees of freedom
```

c) [1.5 points] Write the R code to calculate the critical value for a 99% confidence interval for the difference in group means.

```
### SOLUTION:
# qt(0.995, df = 193.9652)
# or
# qt(0.005, df = 193.9652)
```

The critical value you calculated in c) is:

```
## [1] 2.6
```

d) [3 points] Compute a 99% confidence interval for the difference between the two means (treatment - placebo)

```
### SOLUTION:
# lower = (32.25 - 35.97) - 2.6 * sqrt(8.67^2/100 + 10.02^2/100)
# upper = (32.25 - 35.97) + 2.6 * sqrt(8.67^2/100 + 10.02^2/100)
# c(lower, upper)
```

[Hint for this part and future parts: the confidence interval does not include 0]

e) [1 point] Would you reject a two-sided $\alpha = 1\%$ null hypothesis that the difference between the two means are the same? Why or why not?

```
### SOLUTION: Yes, I would reject, because the 99% CI does not include 0.
```

f) [1 point] Based on your confidence interval, place an upper or lower bound on the p-value of the t-statistic you would calculate if you carried out the test suggested in 10(f).

```
### SOLUTION: The p-value would be less than 0.01.
```

BONUS [1 point] True or False. Based on these results, I know that I would reject a one-sided test that the treatment mean tumor growth is **less** than the placebo mean tumor growth.

◯ True
◯ False

```
### SOLUTION: True. One-sided hypotheses are easier to reject than two-sided.
```

Question 21 [13 points total]

Suppose that you're attending your maternal and child health seminar and learned about differences in birth weight among babies born to mothers of different race/ethnicities. In your internship, you have access to a random sample of births at the UCSF Benioff Children's Hospital in Oakland and want to know if babies born at this hospital also exhibit these differences by race/ethnicity.

You first run these commands on your random sample of births that you stored in the data frame called `birth_data`:

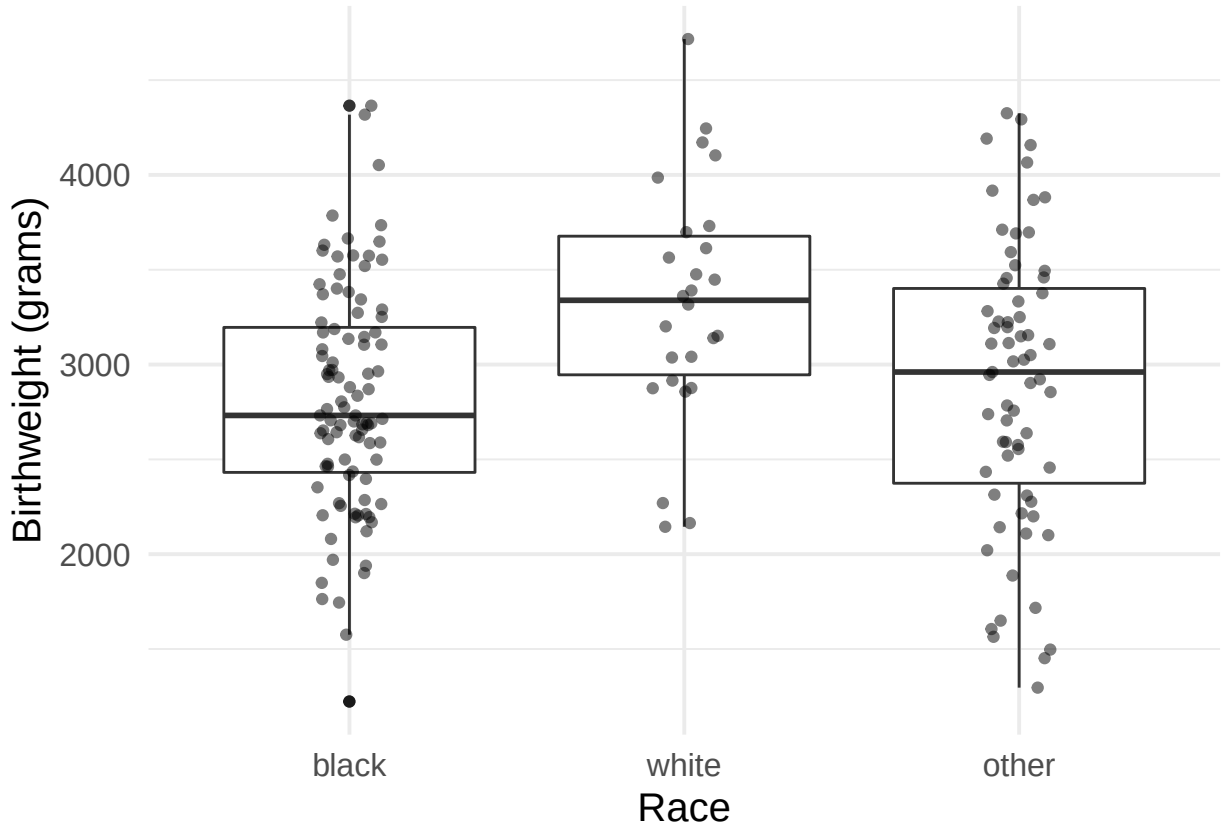
```
dim(birth_data)
head(birth_data)
str(birth_data)
```

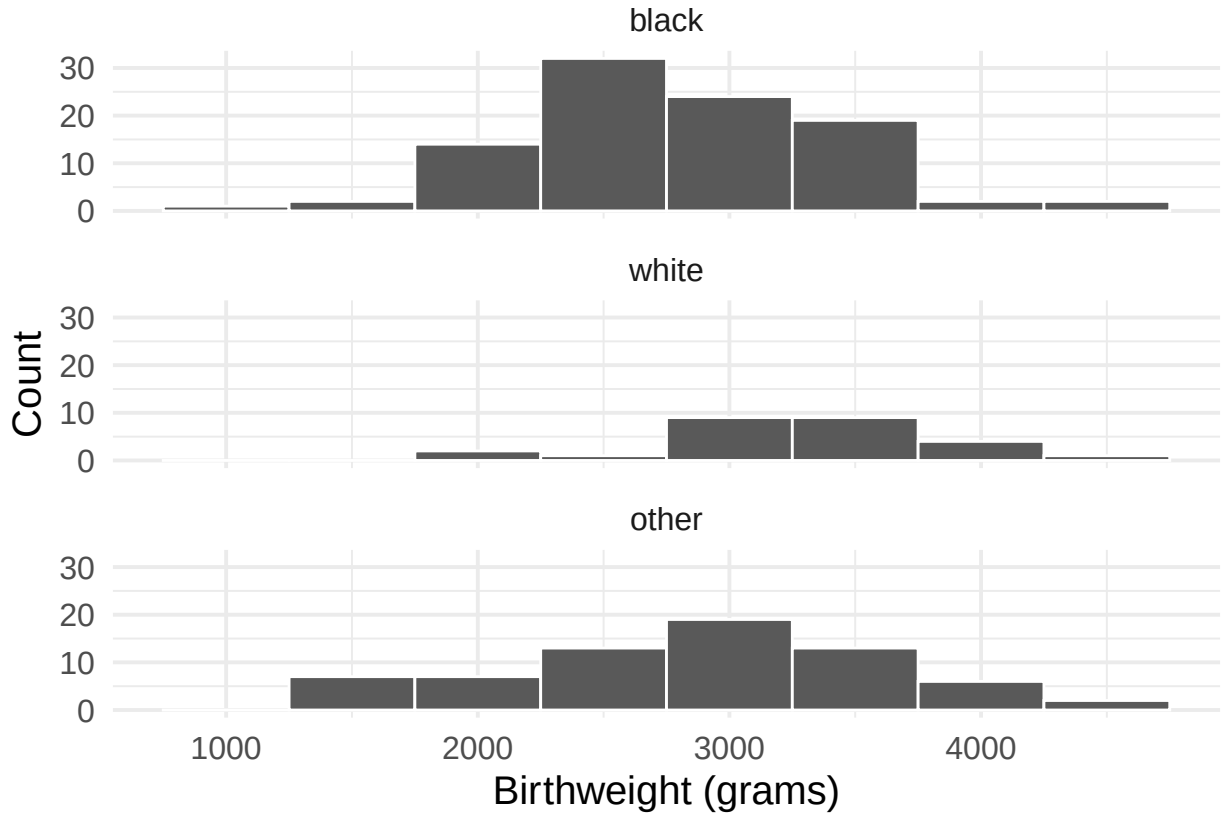
You then use `dplyr` to examine characteristics of a random sample of births at the hospital:

```
birth_data %>%
  group_by(race_order) %>%
  summarize(n = n(),
            mean = round(mean(birthweight), 1),
            sd = round(sd(birthweight), 1))
```

```
## # A tibble: 3 x 4
##   race_order     n  mean    sd
##   <fct>      <int> <dbl> <dbl>
## 1 black         96 2813   594.
## 2 white         26 3327.  630.
## 3 other         67 2893.  747.
```

You also make the following descriptive plots to visually explore the relationship between race/ethnicity and birthweight.


Birthweight (grams)
4000
3000
2000
black
white
other
Race



a) [3 points] Write the code required to produce the output below. Load any package that may be required.

```
## # A tibble: 2 x 6
##   term      df        sumsq   meansq statistic  p.value
##   <chr>     <lgl>     <dbl>    <dbl>     <dbl>    <dbl>
## 1 race      NA     5434359. 2717180.      6.30  0.00226
## 2 Residuals NA    80277279.  431598.     NA    NA
```

```
### SOLUTION:
# library(broom)
# aov_bw <- aov(birthweight ~ race, data = birth_data)
# aov_table <- tidy(aov_bw)
```

b) [2 points] The above output corresponded to a hypothesis test with a specific null and alternative hypothesis. Write the null and alternative hypotheses.

```
### SOLUTION:
# H_0: No difference between the underlying population mean birth weights
# for babies born to mothers with race/ethnicity of white, black, and other.
#
# Or:
#
# $H_0: \mu_{white} = \mu_{black} = \mu_{other}$
#
# H_a: At least one of the underlying population means is different from the others.
```

c) [3 points] Based on the table above, report the value of the test statistic and state its distribution. What are the degrees of freedom for this distribution? (Hint: some of the earlier report information will be helpful!)

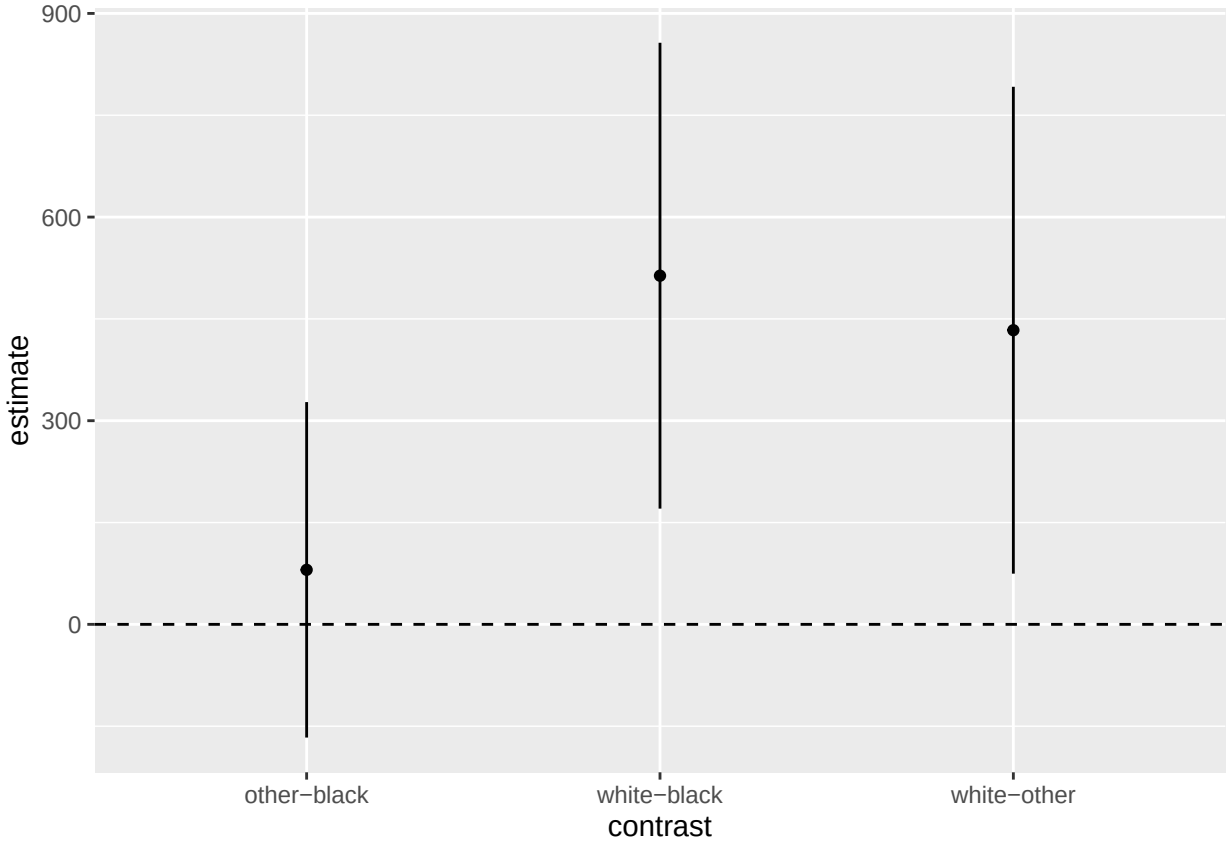
```
### SOLUTION: F-statistic is 6.295622 follows an F distribution
# with 2 degrees of freedom in the numerator and
# 186 degrees of freedom in the denominator
```

d) [1 point] Interpret the p-value corresponding to the test statistic.

```
### SOLUTION: A p-value of 0.002 provided strong evidence against the null hypothesis
# in favor of the alternative hypothesis that at least one of the group means differs
# from the other two.
```

e) [2 points] Because the p-value is so small, you decide to run the following code and make the following plot:

```
## # A tibble: 3 x 7
##   term  contrast    null.value estimate conf.low conf.high adj.p.value
##   <chr> <chr>            <dbl>    <dbl>    <dbl>     <dbl>       <dbl>
## 1 race  other-black          0     80.3   -167.       327.     0.723
## 2 race  white-black          0    514.     170.       857.     0.00148
## 3 race  white-other          0    433.      74.7      792.     0.0133
```



Based on the results, comment on which groups appear the most different.

```
### SOLUTION:
# white vs. black is very different [1 point]
# white vs. other is very different [1 point]
# other vs. black - no evidence for difference
```

f) [2 points] In the table produced from the `TukeyHSD` code, `adj.p.value` stands for adjusted p-value. What is it adjusted for? Will it be higher or lower than an unadjusted p-value?

```
### SOLUTION:
# Adjusted p-values are adjusted for multiple testing/ Adjusted p-values fix the
# family-wise error rate to 5% [1 point]
# Adjusted p-values are higher than unadjusted p-values [1 point]
```

Question 22 [10 points total]

In a 2015 study examining the use of tobacco among students in California, students were asked about their tobacco usage and the number of school days they missed in the past thirty days. Students were not asked why they missed school. We are interested in looking at the relationship between missing school and tobacco usage. The two-way table below summarizes the findings of the study.

	Currently Using	Not Currently Using	Total
0 days	1,421	14,364	15,785
1-5 days	2,863	17,584	20,447
6+ days	1,346	3,678	5,024
Total	5,630	35,626	41,256

a) [2 points] State which test you would like to conduct followed by your null and alternative hypotheses (in words) in the context of this question.

Test:

Null hypothesis:

Alternative hypothesis:

```
### SOLUTION:
# Test: Chi-squared test for independence

# Null: Missing school is independent of tobacco usage

# Alternative: Missing school is dependent on tobacco usage
```

b) [1 point] Restate your hypotheses using probability notation.

Null:

Alternative:

```
### SOLUTION:
# Null: P(Missing School | Tobacco Use) = P(Missing School) (1/2 point)
# Or: Null:P(Missing School | Tobacco Use) = P(Missing School | No Tobacco Use)
# Alternative: P(Missing School | Tobacco Use) $\neq$ P(Missing School) (1/2 point)
# Or: Alternative:
# P(Missing School | Tobacco Use) $\neq$ P(Missing School | No Tobacoo Use) (1/2 point)
```

c) [3 points] Calculate the expected values for this test by creating a table with 6 cells and filling it in with your 6 calculated values.

```
###  SOLUTION: 1/2 point for each correct interior box
#
# |        | Currently Using | Not Currently Using | Total  |
# |--------|-----------------|---------------------|--------|
# |0 days  |      2,154.1    |       13,630.9      | 15,785 |
# |1-5 days|      2,790.3    |       17,656.7      | 20,447 |
# |6+ days |       685.6     |        4,338.4      |  5,024 |
# |--------|-----------------|---------------------|--------|
# |Total   |      5,630      |       35,626        | 41,256 |
```

d) [2 points] Write the formula for calculating the test statistic. From your values in part c), write the first term (based on the upper left cell of your table) of the test statistic. You do not need to calculate the test statistic completely.

```
### SOLUTION:
# Formula: (1 point) $\sum\frac{(O_i - E_i)^2}{E_i}$
# value of first term of test statistic: 249.49
```

e) [1 point] The test statistic equals 1027.769. Write a line of code that you could use to calculate your p-value. Make sure to specify your degrees of freedom.

```
### SOLUTION: P-value: (1 point) pchisq(1027.769, df = 2, lower.tail = F)]
```

f) [1 point] Running your code, R returns a p-value for this test of 6.649627e-224. Interpret the p-value.

```
### SOLUTION:
# Because p-value is less than $\alpha$, it is significant at the .05 confidence level.
# We have sufficient evidence to reject the null hypothesis and we can conclude
# that the probability of missing more school days per month is dependent
# on tobacco use among teens
```

Question 23 [5 points total]

Two cities are concerned that wheezing may be related to pollution from traffic. The residents of each city had their wheezing symptoms compared after a traffic bypass was constructed in one city in 2018 to remove congestion. Data was collected after the bypass was constructed in order to assess the impact of reduction in air pollution on wheezing. Residents reported wheezing symptoms experienced in 2017 and 2018. In the city that had a bypass constructed, 45 out of 282 people reported an improvement in wheezing symptoms. In the city that did not have a bypass constructed, 21 out of 162 people reported an improvement. Assume these represent random samples.

a) Perform a two-sided test (but not a chi-square test) to conclude whether the bypass city has a statistically different proportion of improvement in wheezing symptoms relative to the city that did not get a bypass. Assume the significance level is 5%. Make sure to:

- State your hypotheses,
- Name and calculate your test statistic, and
- Write the code to find the p-value (you do not have to calculate the actual p-value)

```
### SOLUTION: H_0: p_1 = p_2
# (where p_1 represents the proportion of individuals who had
# improved wheezing symptoms where a bypass was constructed)

### H_A: p_1 does not equal p_2
# We can use the large sample method for two proportions because we have more than
# 10 failures or successes in each group

# observed_difference <- 45/282 - 21/162
# p_hat <- (45 + 21) / (282 + 162)
# se <-  sqrt(p_hat * (1 / 282 + 1 / 162))
# test_stat <- observed_difference / se
# p_val <- pnorm(test_stat, lower.tail = F) * 2
```

b) Assume the p-value comes out to be 0.393. Interpret this value. Is this evidence consistent with the null hypothesis or in favor of the alternative?

```
### SOLUTION: The evidence is consistent with the null hypothesis,
# so we fail to reject the null hypothesis at any conventional significance level
# (0.01, 0.05, or 0.1). There is a 39.3% change of observing a sample statistic
# as extreme or more extreme than what we observed.
```

Question 24 [6 points total]

A Pew Research Center survey found that 73% of 2,378 adult American internet users interviewed said that the internet has helped them in some way in obtaining health information.

a) [3 points] Use the Plus Four Method to find the 95% confidence interval for the proportion of all adult American internet users who feel the same way.

```
### SOLUTION: p_tilde = ((0.73*2378) + 2)/2378 = 0.73084
### z = 1.96
### se = sqrt((p_tilde)*(1-p_tilde)/(2378+4)) = 0.0090875
### lower_bound = p_tilde - z*se = 0.713 or 71.3%
### upper_bound = p_tilde - z*se = 0.749 or 74.9%
```

b) [2 points] If you were told that Pew randomly surveyed adult American internet users, but only 30% of those asked actually participated, which assumption of the method you used to create the confidence interval might be violated due to non-response and why?

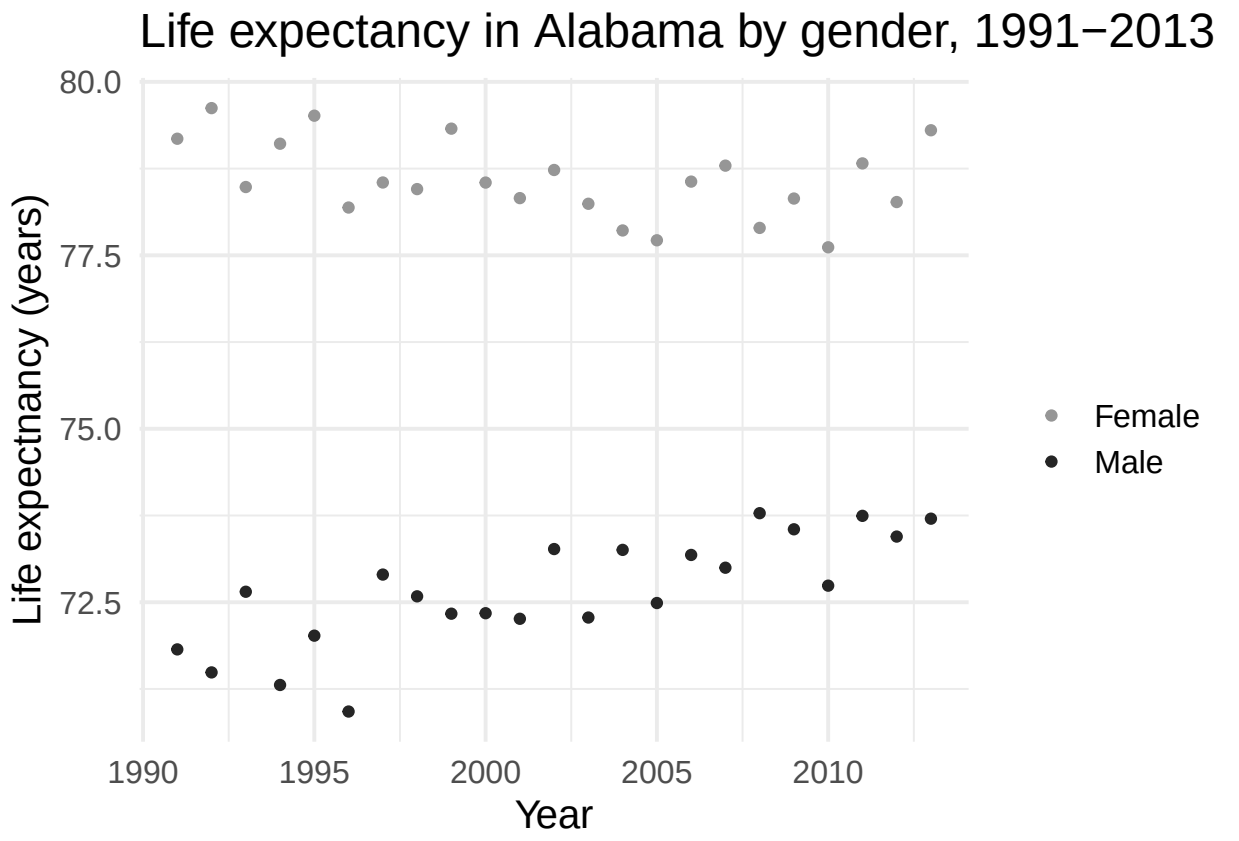
```
### SOLUTION: This means that the nonresponse rate of the study would be 70%.
# When we increase the nonresponse rate of a study, we decrease the likelihood
# that the sample is truly representative of the population we are interested in
# making inference about.
# So, the assumption we made about taking a simple random sample (SRS)
# could arguably be violated since the nonresponse rate is quite high.
```

c) [1 point] State one change you can make to the question to make this confidence interval narrower.

```
### SOLUTION: Increase the number of people (n) that you sample.
### OR Increase the significance level alpha to decrease the confidence level.
# For example, a 90% CI (with alpha = 0.10) would have a smaller width
# because it will have a smaller critical value.
```

Question 25 [7 points total]

Suppose that you had data on life expectancy for white men and women in Alabama between 1990 and 2013, where such data was based on a random sample of Alabama white residents across those years. You first decided to graph these data as shown below:



You decide to run two models, one for women and the other for men, looking at the relationship between year and life expectancy. Here are the model outputs:

Model for women:

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept) 149.       32.8         4.55 0.000174
## 2 year         -0.0354    0.0164     -2.16 0.0427
```

Model for men:

```
## # A tibble: 2 x 5
##   term         estimate std.error statistic    p.value
##   <chr>           <dbl>     <dbl>     <dbl>      <dbl>
## 1 (Intercept) -115.       31.1        -3.68 0.00138
## 2 year           0.0935    0.0155      6.02 0.00000565
```

a) [3 points] You are interested in knowing whether life expectancy changed linearly over the time period examined for both men and women. Fill in the following table.

Group	Estimate of the slope	Test statistic	p-value
Women			
Men			

```
### SOLUTION:
# | Group  | Estimate of the slope | Test statistic | p-value     |
# |--------|-----------------------|----------------|-------------|
# | Women  |        -0.0354        |     -2.16      | 0.0427      |
# | Men    |         0.0935        |      6.02      | 0.00000565  |
```

b) [1 point] For which group (men, women, or both) do you have evidence against the null hypothesis in favor of the alternative? (assume a 5% significance level).

```
### SOLUTION: both men and women
```

c) [3 points] Interpret the slope coefficient for men in the context of the question. Create a 95% confidence interval for the slope coefficient for men (Hint: the output provided below will be helpful).

```
## [1] 2.079614
```

```
### SOLUTION:
# For every 1 year increase from 1990 to 2013, white men in Alabama
# experienced an average of a 0.0935 year increase in life expectancy.

# Estimate +/- t_star*std.error
# lower: 0.09354491 - (2.079614*0.0155401)
# upper: 0.09354491 + (2.079614*0.0155401)
# 95% CI: [0.0612275, 0.1258623]
```

d) [BONUS point] The intercept estimate is about -115 for men. What does this mean?

```
### SOLUTION: If you were to extend the line back to Year = 0, the estimate life
# expectancy is -115. This is nonsensical, and reflects extreme extrapolation
# way outside the range of the data.
```