PH142 Spring 2019 Final Exam SOLUTIONS

Please read the following statement and sign below to indicate that you understand the policies for this exam.

You can use the back of each page as scratch paper, however no points will be given for answers on the back pages.

Cellphones and computers must be stored and an silent during this exam. You may use a non-graphing calculator and one double sided page of handwritten notes or notes printed at 10 point font or larger.

You must show your student ID when you submit your test.

For several of the questions we have added an answer box for your final answer. This helps us to grade quickly. You may receive partial credit if you show your work, but to receive full credit your final answer must be recorded in the answer box.

For questions with calculations, please express your answer to 2 decimal places.

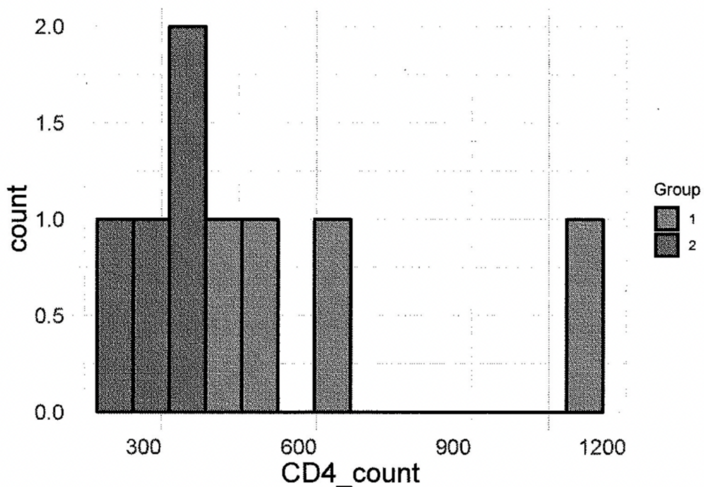
Please sign below to indicate your understanding of the academic integrity statement. I understand the University of California policy on academic integrity. I will not collaborate on this exam with my classmates or any other individuals. I will not use any resources to complete this exam other than those noted above. I understand that if I violate any of these policies, I will receive a O on this exam.

Signature: ____

Question 1 [8 points total]

A CD4 test counts the number of working white blood cells a person has in their body to help them fight infections. Below is a dataframe with the CD4 counts of 8 patients with HIV from two study groups.

```
##   Group CD4_count
## 1     1       524
## 2     1       610
## 3     2       340
## 4     1      1100
## 5     2       220
## 6     2       340
## 7     1       400
## 8     2       290
```



a) [1 point] What statistical test should be used to compare these two groups?

```
### SOLUTION: Wilcoxon Ranksum test/ Mann-Whitney test
```

b) [1 point] State the null and alternative hypotheses for this test (as one sentence each).

```
### SOLUTION:
# $H_0$: The two distributions are equal
# $H_A$: The two distributions are not equal
```

c) [1 point] Perform this test by hand. Report the value of the test statistic.

SOLUTION:

$\mu_w = \frac{4*(4+4+1)}{2} = 18$

$\sigma_w = \sqrt{frac4.4 * (4+4+1)12} = 3.464$

$Z_w = \frac{18-10}{3.464} = -2.309$

d) [3 points] Fill in the blanks. The p-value for this test statistic is 0.0294.

Based on this you would have evidence to __________________ the null hypothesis.

This value represents the __________________ of observing __________________ or __________________ if the __________________ is __________________.

```
### SOLUTION:
# reject
# probability
# a z-score of -2.309
# more extreme
# null hypothesis
# true
```

e) [2 points] Fill in the blanks of the line of code that you would run to perform this test in R.

```
____________________(_________________~________________, ______________ = FALSE)
```

```
### SOLUTION:
# wilcox.test
# CD4_count
# Group
# paired
```

Question 2 [3 points total]

It is often cited that women make 78 cents for every dollar that a man makes in the United States. To investigate this, you collect data on weekly earnings from 25 large companies who employ software engineers within a few years of graduation from college. In each company you randomly select a male and female software engineer who have been with the company for 2 years and who were hired directly after college graduation.

Fill in the blanks.

To approach this problem with a parametric test you would use a __________________ to

assess the null hypothesis that __________________ = __________________.

If you graphed the outcome and had evidence to suggest that the underlying

distribution was not __________________ and thus violated an important assumption of your

test, you might instead choose to test this hypothesis with a __________________ test (type of test).

In this case the __________________ (specific test) would be a good choice.

```
### SOLUTION:
# paired t-test
# mu_(female weekly earnings)
# mu_(male weekly earnings)
# Normal
# non-parametric
# Wilcox Sign-Rank
```

Question 3 [2 points total]

To understand the workings of the placebo effect on patients with Parkinson's disease, scientists measure the level of activity at a key location in the brain when patients receive a placebo that they think is an active drug and also when no treatment is given. They measure the brain response activity under two conditions for each patient. The 47 patients are randomized to either active treatment and then placebo, or placebo and then active treatment. The data are slightly skewed.

a) [1 point] State the null hypothesis.

```
### SOLUTION:
# $H_0$: difference in mu for each patient under different conditions = 0
```

b) [1 point] What statistical test should you use to test the hypotheses? Explain.

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

Question 4 [3 points total]

You have heard that the grading scale is harsher at UC Berkeley than at other California universities. You want to test this rumor with data. You have data from a random sample of 100 transcripts from students at Berkeley who took PH142 and data on the letter grade distribution for undergraduate statistic courses in general from a California wide survey.

a) [2 points] State the null and alternative hypotheses.

```
### SOLUTION:
# $H_0$: UC Berkeley distribution of letter grades is equal to
# the California wide distribution

# $H_A$: UC Berkeley distribution of letter grades is not equal to
# the California wide distribution
```

b) [1 point] What statistical test should you use to test the hypotheses?

```
### SOLUTION:
# chi-squared goodness of fit
```

Question 5 [3 points total]

Your group is interested in mosquito control and resistance to insecticides used in the spraying of houses in areas where malaria is endemic. You have set up a study with test structures in an area where there are a lot of mosquitoes. You have treated each structure with a different amount of insecticide and set up a mosquito trap in each structure. You then measure the number of mosquitoes trapped in each structure.

a) [2 points] State the null and alternative hypotheses.

```
### SOLUTION:
# $H_0$: the effect of insecticides is 0, beta = 0
# $H_A$: the effect of insecticides is positive, beta does not = 0
```

b) [1 point] What statistical test should you use to test the hypotheses? Explain.

```
### SOLUTION:
# t-test for regression coefficient, since we aim to test whether
# the coefficient beta is significantly not 0
```

Question 6 [2 points total]

Name two ways to increase the power for a hypothesis test in a study.

```
### SOLUTION:
# Increase the sample size.
# Increase the effect size (difference between the null and alternative).
```

Question 7 [9 points total]

The 2018 NBA final champions were the Golden State Warriors. Klay Thompson, Stephen Curry, and Draymond Green are three of the basketball players instrumental to this championship win. Below is a table summarizing how many points the player(s) made by a 3-point shot versus points made by all other types of shots during their final four games.

	Points made by 3 point shot	All other points made	Total
Klay Thompson	36	28	64
Draymond Green	9	28	37
Stephen Curry	66	A	110
Total	B	C	D

a) [5 points] You want to test the null hypothesis that there is no relationship between player and probability that a point was made by a 3 point shot. Fill in the table below with the expected values under the null hypothesis of no association.

	Points made by 3 point shot	All other points made	Total
Klay Thompson			
Draymond Green			
Stephen Curry			
Total			

i) First find the values of A, B, C, D:

```
### SOLUTION:
# A: 44
# B: 111
# C: 100
# D: 211
```

ii) Then calculate the expected values using the total points made by each player:

```
### SOLUTION:
# |                |  Points made by 3 point shot | All other points made | Total|
# |----------------|------------------------------|-----------------------|------|
# | Klay Thompson  |            33.67             |         30.33         |  64  |
# | Draymond Green |            19.46             |         17.54         |  37  |
# | Stephen Curry  |            57.87             |         52.13         |  110 |
# | Total          |             111              |          100          |  211 |
```

b) [2 points] Calculate the appropriate test statistic by hand (round to two decimal places).

SOLUTION:

$$\chi^2 = \frac{(36-33.664)^2}{33.664} + \frac{(28-30.336)^2}{30.336} + \frac{(9-19.462)^2}{19.462} + \frac{(28-17.538)^2}{17.538} + \frac{(66-57.86)^2}{57.86} + \frac{(44-52.14)^2}{52.14} = 14.62$$

c) [2 points] Write the line of code in R that you would use to obtain the p-value for this test statistic.

```
____________(____________, df = ____________, ____________  = ____________ )
```

```
### SOLUTION:
# pchisq
# q = 14.61
# 2
# lower.tail
# FALSE
```

Question 8 [1 point total]

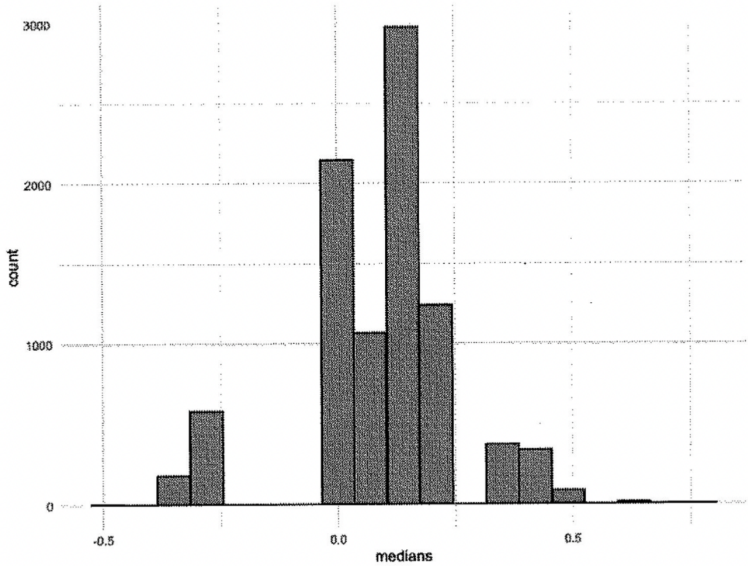
The bootstrap method allows us to create confidence intervals based off of a ____________________.

a) simulated null distribution
b) sampling distribution
c) simulated sampling distribution
d) normal distribution

```
### SOLUTION: c) simulated sampling distribution
```

Question 9 [2 points total]

Given the information and distribution below, what would the 95% bootstrap confidence interval be? (round to 2 decimal places).




```
simulated_sampling_dist %>% summarize(a=quantile(medians, 0.025),
                                      b=quantile(medians, 0.05),
                                      c=quantile(medians, 0.10),
                                      d=quantile(medians, 0.9),
                                      e=quantile(medians, 0.95),
                                      f=quantile(medians, 0.975))

##            a         b c         d         e         f
## 1 -0.3051921 -0.288562 0 0.1979838 0.3511125 0.4223171
```

```
### SOLUTION:
# lower bound: -0.31
# upper bound: 0.42
```

Question 10 [6 points total]

To determine if there is an association between human immunodeficiency virus type 1 (HIV-1) infection and Trypanosoma brucei gambiense sleeping sickness, all incident cases of trypanosomiasis and a control group of blood donors presenting to the same rural hospital in Zaire were tested for anti-human immunodeficiency virus type 1 (anti-HIV-1) antibodies. There was no significant difference in the prevalence of HIV-1 infection between the two groups (7 of 220 [3.2%] for the incident cases and 8 of 388 [2.1%] for the blood donors). Among the three HIV-1 seropositive incident cases of trypanosomiasis treated with difluoromethylornithine, two (67%) relapsed after treatment compared with 4 of 39 (10%) HIV-1 seronegative incident cases treated with the same drug.

a) [2 points] Fill in the following two-by-two table based on the abstract.

	Relapse	No Relapse
HIV seropositive & treated		
HIV seronegative & treated		

```
### SOLUTION:

# |                             |  Relapse     |  No Relapse  |
# |-----------------------------|--------------|--------------|
# | HIV seropositive & treated  |      2       |       1      |
# | HIV seronegative & treated  |      4       |       35     |
```

b) [1 point] Would you feel comfortable testing the association between relapse and HIV serostatus with a χ^2 test? Why or why not?

```
### SOLUTION:
# No - not all cells have an expected value > 5
# (requirement for a 2x2 table in order to use chi-square)
# too small of sample size
```

c) [2 points] Create a 95% confidence interval around the estimate of HIV serostatus among cases of trypanosomiasis. When the proportion is close to .5 and the sample is large, we use a large sample CI. We studied one manual method that allows you to appropriately adjust your estimation to account for situations where the sample is small or the proportion is further from .5. Use this method to construct your confidence interval.

```
### SOLUTION:
# lower bound: 0.0145
# upper bound: 0.0659
```

d) [1 point] Interpret this confidence interval in 1-2 sentences.

```
### SOLUTION:
# If we were to repeat this procedure many times, the 95% CI we calculate
# would include the true proportion 95% of the time
# We treat our [0.0145, 0.0659] CI as just one of these potential many CI
```

Question 11 [6 points total]

We wish to explore the association between glucose exposure and plasma glucose levels in patients with diabetes. The following two outputs in R assess the association between the response variable: plasma glucose levels (FPG, in mg/mL) and explanatory variable: natural logged glucose exposure (ln(HbA), in %). Use the output below to answer the questions.

```
library(broom)

glucose_lm <- lm(fpg ~ ln_hba)

tidy(glucose_lm)

## # A tibble: 2 x 5
##   term        estimate std.error statistic p.value
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)    69.5      32.6       2.13  0.0500
## 2 ln_hba          8.92      3.33      2.68  0.0173


glance(glucose_lm)

## # A tibble: 1 x 11
##   r.squared adj.r.squared sigma statistic p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>   <dbl> <int>  <dbl> <dbl> <dbl>
## 1     0.323         0.278  44.7      7.16  0.0173     2  -87.7  181.  184.
## # ... with 2 more variables: deviance <dbl>, df.residual <int>
```

a) [1 point] Write the equation for the line of best fit. Specify what the variables in the context of this question.

```
### SOLUTION:
# $\hat y$ = 8.92*x + 49.5
# y is plasma glucose levels, x is natural logged glucose exposure
```

b) [1 point] How much variation in the outcome variable is determined by the predictor variable?

```
### SOLUTION:
# R^2 = 0.323
```

c) [1 point] What do residuals represent?

```
### SOLUTION:
# The difference between the actual Y values and the predicted Y values.
```

d) [1 point] What command (from the `broom` package in R) would you use to calculate the residuals?

```
### SOLUTION:
augment(glucose_lm)
```

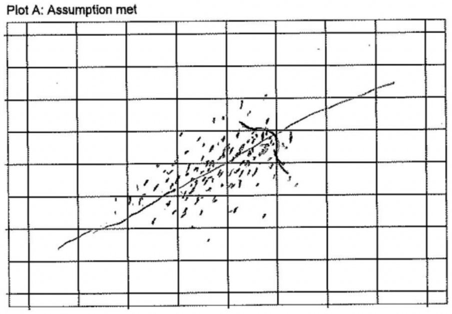
e) [1 point] Name one assumption of the linear model that can be assessed using residuals and identify a plot that you would use to assess that assumption.

```
### SOLUTION:
# Linearity assumption: the relationship of x and y is linear in the population.
# Use the scatterplot with fitted line to assess this assumption.
```

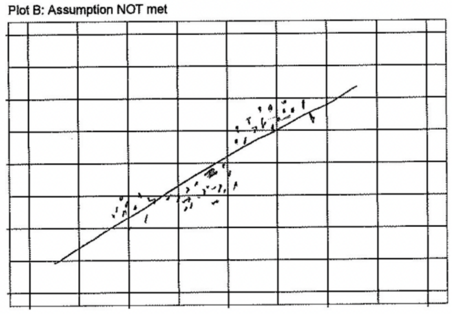
f) [1 point] For the assumption you named, roughly sketch 2 versions of the plot you would use to assess the assumption. One plot should roughly represent what the plot would look like if the assumption was met and the other plot should roughly represent what it would look like if the assumption was *violated*.

```
### SOLUTION:
# assumption was met: points are located around the line.
# assumption was violated: points are located like a nonlinear pattern.
```

Plot A: Assumption met



Plot B: Assumption NOT met



Question 12 [4 points total]

In a city far away, a car enthusiast is interested in whether Honda and Toyota sales were the same. We are interested in testing these hypotheses:

$H_0 : \mu_{Honda} = \mu_{Toyota}$

$H_A : \mu_{Honda} < \mu_{Toyota}$

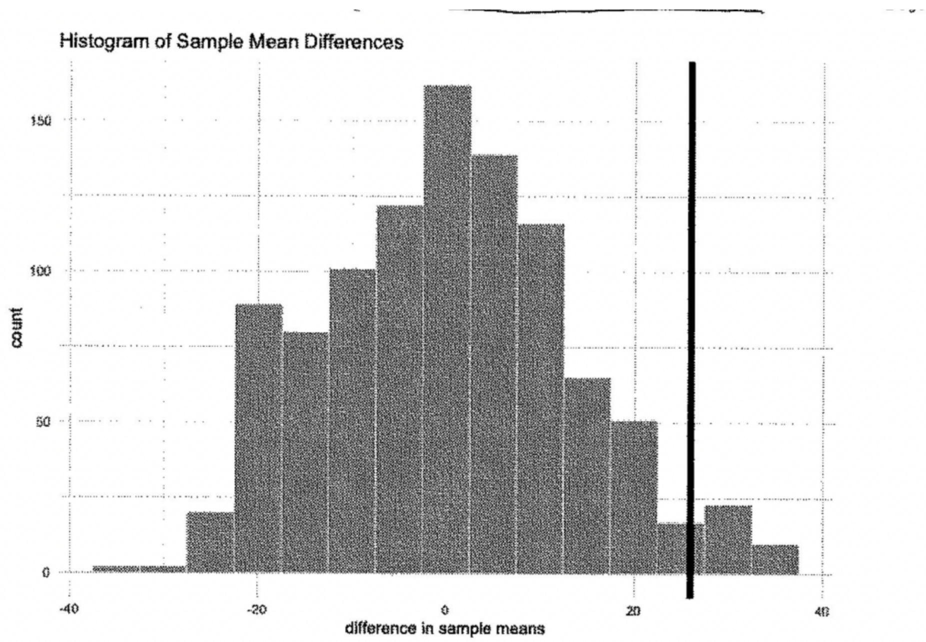
We do not know anything about the true underlying distributions of Honda and Toyota sales. We are also working with a small dataset (n = 7 for both groups).

a) [2 points] Why would we be interested in testing this data with a non-parametric test instead of a t-test? (1-2 sentences).

```
### SOLUTION:
# We would have to satisfy the following conditions if we were to use a t-test:
# The data should be collected from a simple random sample (SRS)
# The distribution of the data should be roughly Normal (with no outliers or strong skew).

# Since we do not know anything about the true underlying distributions
# and the fact that the small # sample sizes (n = 7) makes it difficult for us to
# establish any type of distribution, testing with
# a non-parametric test would be better than t-testing.
```

b) [2 points] The black line is our observed difference in sample means. The plot shows our simulated null distribution based on reshuffling our labels on our data. 3.6% of our data lie above our observed difference of $\bar{x}_{Toyota} - \bar{x}_{Honda} = 26$.



Based on this information and the given distribution, if we are interested in testing at an $\alpha = 0.05$ significance level, what would we conclude?

```
### SOLUTION:
# Our p-value is 0.036; in other words, there is a 3.6% chance
# that we observe a difference in sample means of 26 or more extreme
# given the null condition that the mean sales of Hondas and Toyotas are the same.

# Given our alpha of 0.05, because our p-value is less than
# the significance level (0.036 < 0.05), we have sufficient evidence
# to reject the null hypothesis.
# There is sufficient evidence to determine a greater mean sale of Toyotas compared to Hondas.

# Reject H0
```